

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

University of Southampton
Faculty of Physical Sciences and Engineering
Electronics and Computer Science

Cost-Effective 3D-IC Design using Near-Field Inter-Tier
Wireless Communication

by

Benjamin Fletcher

A thesis submitted for the degree of
Doctor of Philosophy

November, 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

ELECTRONICS AND COMPUTER SCIENCE

Thesis for the degree of Doctor of Philosophy

COST-EFFECTIVE 3D-IC DESIGN USING NEAR-FIELD INTER-TIER WIRELESS COMMUNICATION

by Benjamin Fletcher

Modern Internet of Things (IoT) devices are becoming increasingly complex, often incorporating a range of different components (sensors/processing/memory/logic) fabricated using a variety of process technologies. To integrate these disparate elements in a low-cost, small and power-efficient way, research has looked to ‘3D integration’ where several tiers are stacked and interconnected vertically within a single chip. Most research into 3D integration assumes the use of Through Silicon Vias (TSVs) to interconnect stacked tiers; however, TSVs are presently expensive to manufacture and only available in leading-edge process nodes, making them poorly suited to cost-sensitive IoT applications.

In this thesis, wireless Inductive Coupling Links (ICLs) are investigated as an alternative to TSVs for vertical communication (and power delivery) within a 3D-IC. The motivation for focusing on wireless links is primarily cost-driven, as ICLs do not require 3D-specific fabrication processes and can facilitate simple pick-and-place assembly using only adhesive. Specifically, this work explores the design challenges associated with such ICLs, aiming to establish a standard interface that can be used for IoT-style 3D stacking applications. The key novel contributions include: (i) A low-energy ICL transceiver that uses time-domain encoding to reduce the number of transmit pulses, and hence overall energy, by over 13% when compared to existing solutions. (ii) A CAD tool for automated ICL inductor optimisation that significantly reduces the design time (by over 6 orders-of-magnitude) when compared with finite element tools, whilst maintaining an average accuracy within 7.8%. (iii) A near-field wireless clock link for many-tier clock synchronisation that achieves low-skew clock distribution across a wide range of frequencies (results show less than 61ps of clock skew across five tiers when operating between 50MHz and 2.3GHz). (iv) A hybrid ICL transceiver for concurrent wireless data and power transmission. The proposed transceiver can achieve wireless power transfer of up-to 2.0mW/link whilst simultaneously transferring 1.4Gbps of data using a BPSK scheme.

These four contributions are also validated through two 3D-stacked silicon test-chip demonstrators, the first fabricated in 0.35 μm CMOS technology (showcasing the low-energy ICL transceiver), and the second fabricated in 65nm CMOS technology (showcasing wireless data, power and clock transmission as part of a 3D stacked Arm Cortex M0 SoC). Overall, this work represents an exciting step towards a new era in VLSI where IC designers can ‘pick-and-mix’ the functional circuit blocks and technologies within a given chip (in the form of separate semiconductor dies) and stack them together in a low-cost way using ICLs.

Table of Contents

Abstract	iii
Declaration of Authorship	ix
Acknowledgements	xi
Acronyms/Abbreviations	xiii
Nomenclature	xvii
1 Introduction	1
1.1 3D Integration Approaches	2
1.1.1 Through Silicon Vias	3
1.1.2 3D SiPs using Wire Bonding	4
1.1.3 Monolithic 3D Integration	5
1.2 Wireless 3D Integration	6
1.3 Research Justification	8
1.4 Research Context	10
1.5 Research Questions	12
1.6 Research Contributions	13
1.7 Publications and Patents	14
1.7.1 Peer-Reviewed Publications	14
1.7.2 Patents	16
1.7.3 Other Research Engagement Activities	16
1.8 Thesis Outline	16
2 Wireless Three-Dimensional (3D) Integrated Circuits	19
2.1 Capacitive Coupling Links	19
2.2 Inductive Coupling Links	23
2.2.1 Data Encoding Schemes	25
2.2.2 Transceiver Designs	28
2.2.3 Inductor Layout for ICLs	32
2.3 Clock Delivery in Wireless 3D-ICs	33
2.3.1 Wireless Clock Delivery	34
2.4 Power Delivery in Wireless 3D-ICs	35
2.4.1 Wireless Power Transfer	36
2.5 Summary	39
3 Low-Energy Transceiver Design using Spike-Latency Encoding	41
3.1 Background and Related Work	42

3.2	Proposed Spike-Latency Encoding Modulation Scheme	44
3.2.1	Mathematical Modelling	45
3.3	Architecture Design and Hardware Implementation	47
3.3.1	Encoding/Decoding Logic	47
3.3.2	Tuneable Current Driver	49
3.3.3	Inductive Channel	51
3.3.4	Sense Amplifier	51
3.3.5	Clock Synchronisation	52
3.4	Experimental Validation and Results	53
3.4.1	ICL Layout Parameter Selection	53
3.4.2	Validation using Mathematical Models	54
3.4.3	Experimental Validation using SPICE	55
3.5	Case Study: Test-Chip Demonstration	62
3.5.1	Tuneable Current Driver Evaluation	62
3.5.2	Timing Margin Evaluation	63
3.5.3	Energy-per-Bit Evaluation	63
3.6	Discussion	67
3.7	Summary	68
4	Design and Optimisation of Inductive Coupling Channels	71
4.1	Background and Related Work	72
4.2	Modelling and Analysis of ICLs	75
4.2.1	Inductive Coupling Data Links	75
4.2.2	Inductive Coupling Power Links	76
4.2.3	Objective Functions	77
4.2.4	Planar Spiral Inductors	78
4.3	ICL Layout Optimisation (COIL-3D)	80
4.3.1	Scalable Inductor Model	81
4.3.2	Parameter Evaluation	81
4.3.3	Optimisation Approach	84
4.3.4	Software Implementation	86
4.4	Experimental Results and Evaluation	87
4.4.1	Inductor Topology Evaluation	88
4.4.2	Lumped Model Accuracy Evaluation	89
4.4.3	Empirical Expression Evaluation	90
4.4.4	Optimisation Flow Evaluation	92
4.4.5	Overhead Evaluation	92
4.5	COIL-3D Example Usage Application	94
4.6	Summary	95
5	Wireless Inter-Tier Clock Distribution Using Inductive Links	97
5.1	Background and Related Work	99
5.2	WiSync Design and Implementation	100
5.2.1	Dual-Mode Transmitter Design	101
5.2.2	Inductor Layout Selection	103
5.2.3	Receiver Design	104
5.3	Experimental Validation and Results	105
5.3.1	Device Sizing and Area Evaluation	106

5.3.2	Energy per Cycle Evaluation	108
5.3.3	Cycle Error Rate (CER) and Maximum Stack Height Evaluation . .	109
5.3.4	Clock Skew Evaluation	111
5.4	Test-Chip Validation	112
5.4.1	Parameter Tuning	114
5.4.2	Energy Measurement	114
5.4.3	Jitter Measurement	116
5.4.4	Tolerance to Misalignment	116
5.4.5	Discussion	118
5.5	Summary	119
6	Concurrent Wireless Data and Power Transmission	121
6.1	Background and Related Work	122
6.2	Concurrent Power and Data Delivery Architecture	123
6.2.1	Bi-Phase Transmitter Design	124
6.2.2	Inductive Channel and Tuning Circuit Design	126
6.2.3	Rectifier and Low Drop-Out Regulator Design	128
6.2.4	Bi-Phase Demodulator Design	129
6.3	Results and Evaluation	130
6.3.1	Channel Inductor Optimisation	131
6.3.2	Area Evaluation	132
6.3.3	Start-Up and Transient Performance	133
6.3.4	Power Delivery Performance	134
6.3.5	Data Delivery Performance	136
6.4	Case Study: A 3D stacked Arm Cortex M0 SoC (Silicon Evaluation)	137
6.4.1	ICL AHB-Lite Bus Integration	139
6.4.2	System Design	142
6.4.3	Experimental Results	143
6.5	Summary	149
7	Conclusions and Future Work	151
7.1	Research Questions	153
7.2	Future Work	156
7.2.1	Security for ICL-based 3D-ICs	156
7.2.2	Thermal Management for ICL-based 3D-ICs	158
7.2.3	Interference and PVT Variation Effects in ICLs	159
7.2.4	Interposed Silicon Usage in ICL Channels	160
	Appendix	vii
A	Iso-Area Bandwidth and Energy Comparison of ICLs and TSVs	165
A.1	Interface using Inductive Coupling Links	166
A.2	Interface using Through Silicon Vias	168
A.3	Comparison Results	171
A.4	Discussion	173
B	Dual-Dirac Model for Low-BER Jitter Extrapolation	175
B.1	Overview	175

B.2	Mathematical Modelling	176
B.2.1	BER Extrapolation	177
C	Test Chip 1 Fabrication & Assembly Details	179
C.1	Layout and Floorplan	179
C.2	Thinning and Stacking	181
C.3	Bonding and Packaging	181
D	Test Chip 2 Fabrication & Assembly Details	183
D.1	Layout and Floorplan	183
D.2	Thinning and Stacking	184
D.3	Bonding and Packaging	185
References		187

Declaration of Authorship

I, Benjamin Fletcher, declare that this thesis entitled *Cost-Effective 3D-IC Design using Near-Field Inter-Tier Wireless Communication* and the work presented in it are my own and have been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Publications reported directly as contributions (listed in reverse chronological order):

- B. J. Fletcher, S. Das, T. Mak, "A Spike-Latency Transceiver with Tuneable Pulse Control for Low-Energy Wireless 3D Integration", *IEEE Journal on Solid State Circuits (JSSC)* 55(9) pp. 2414-28 (2020).
- B. J. Fletcher, T. Mak and S. Das, "A 3D-Stacked Cortex-M0 SoC with 20.3Gbps/mm² 7.1mW/mm² Simultaneous Wireless Inter-Tier Data and Power Transfer," *IEEE Symposium on VLSI Circuits*, Honolulu, Hawaii, 2020 pp. 1-2.
- B. J. Fletcher, S. Das and T. Mak, "A 10.8pJ/bit Pulse-Position Inductive Transceiver for Low-Energy Wireless 3D Integration," *IEEE European Solid State Circuits Conference (ESSCIRC)*, Cracow, Poland, 2019, pp. 121-4.
- B. J. Fletcher, S. Das, T. Mak, "A Low-Energy Inductive Transceiver using Spike-Latency Encoding for Wireless 3D Integration," *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Lausanne, Switzerland, 2019, pp. 1-6.
- B. J. Fletcher, S. Das and T. Mak, "Design and Optimization of Inductive-Coupling Links for 3-D-ICs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27(3) pp. 711-23 (2019).
- B. J. Fletcher, S. Das, T. Mak, "CoDAPT: a concurrent data and power transceiver for fully wireless 3D-ICs," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Florence, Italy, 2019, pp. 1343-48.
- B. J. Fletcher, S. Das, T. Mak, "Cost-effective 3D integration using inductive coupling links: Can we make stacking silicon as easy as stacking Lego?," *Arm Research Summit 2018*, Cambridge, United Kingdom. 17 - 19 Sep 2018.

- B. J. Fletcher, S. Das, T. Mak, “A High-Speed Design Methodology for Inductive Coupling Links in 3D-ICs”, *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, Germany, 2018, pp. 497-502.
- B. J. Fletcher, S. Das, T. Mak, “Low-Power 3D Integration using Inductive Coupling Links for Neurotechnology Applications,”, *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, Germany, 2018, pp. 497-502.

Patents directly reported as contributions:

- B. J. Fletcher, J. Myers, S. Das and T. Mak, “*A Pseudo System-on-Chip Architecture Incorporating Wirelessly Connected Bus Slaves.*”, U.S. Patent 16/685,090, Nov 2019 (pending).
- S. Gamage, B. Fletcher, and S. Das, “*Adaptive Coding for Wireless Communication.*”, U.S. Patent 16/656,937, Oct 2019 (pending).

Other publications completed during my PhD candidature:

- D. Balsamo, B. J. Fletcher, A. J. Weddell, G. Karatzias, B. Al-Hashimi and G. V. Merrett, “Power neutral performance scaling with intrinsic MPPT for energy harvesting computing systems,” *ACM Transactions on Embedded Computing Systems*, 17(6), pp. 1-25 (2019).
- B. J. Fletcher, D. Balsamo and G. V. Merrett, “Power-neutral performance scaling for self-powered multicore computing systems,”, *Adaptive Many-Core Architectures and Systems Workshop*, York, United Kingdom, 13 Jun 2018.
- Q. Ding, B. J. Fletcher, T. Mak, “Globally Wireless Locally Wired: A Clock Distribution Network for Many-Core Systems,” *IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, Italy, 2018, pp. 1-5.
- B. J. Fletcher, D. Balsamo and G. V. Merrett, “Power neutral performance scaling for energy harvesting MP-SoCs,”, *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, Lausanne, Switzerland, 2017, pp. 1516-21.

Signed:

Date:

Acknowledgements

The work presented in this thesis would not have been possible without the wonderful people that I have had the fortune of working with over the past four years, and so I must take this opportunity to extend my thanks to them.

First and foremost, I would like to express my sincere thanks and appreciation to my academic supervisor, Terrence Mak. I am very grateful to Terrence for his constant encouragement, enthusiasm and incredible creativity, which inspired me to pursue such an exciting and novel research topic. Your positivity has been contagious and has made the PhD experience truly enjoyable. I am also sincerely thankful to Shidhartha Das, my industrial supervisor at Arm Research. Sid's support and friendship have been invaluable, and I have thoroughly enjoyed working under his guidance; it has been a great privilege learn from someone with such broad technical ability.

My special thanks also go to James Myers and the rest of the Devices Circuits and Systems (DCS) group at Arm Research (including, but not limited to, Fernando, Prannay, Benoit, Sahan, Philex, Andy and Graham) for hosting and mentoring me. It has been fantastic to have the opportunity to work with, and learn from, such a talented group of researchers and our discussions have inspired much of the work presented in this thesis. I must also thank my colleagues from the University of Southampton's Cyber Physical Systems research group, including my secondary academic supervisor Geoff Merrett, as well as Qian and Domenico. It has been a joy to work with you all and our collaborations have been a great support over the past four years.

I also wish to thank Arm Research, the Engineering and Physical Sciences Research Council (EPSRC), and the University of Southampton for their financial support with tuition fees, conference attendances and chip tape-outs, without whom this work would not have been possible.

Finally, I must also take the opportunity to thank my family, particularly my parents and my wife Heather for her continued love, support and encouragement throughout the process.

Acronyms/Abbreviations

APR	Automatic Place and Route
ASK	Amplitude Shift Keying
BEOL	Back End Of Line
BER	Bit Error Rate
BPM	Bi-Phase Modulation
BPSK	Bi-Phase Shift Keying
CCL	Capacitive Coupling Link
CDN	Clock Distribution Network
CER	Cycle Error Rate
CMOS	Complementary Metal Oxide Semiconductor
CMP	Chemical-Mechanical Polishing
CoDAPT	Concurrent Data and Power Transfer
DLL	Delay Locked Loop
DMA	Direct Memory Access
DoS	Denial of Service
DP	Dynamic Programming
DRC	Design Rule Check
ECC	Error Correction Code
EDA	Electronic Design Automation
EH	Energy Harvesting
EM	Electro-Magnetic
F2B	Face-to-Back
F2F	Face-to-Face
FDM	Frequency-Division Multiplexing
FEM	Finite Element Method

FEOL	Front End Of Line
FIFO	First-In-First-Out
FLL	Frequency Locked Loop
FPGA	Field Programmable Gate Array
HDSV	Highly-Doped Silicon Via
HF	High Frequency
IC	Integrated Circuit
ICL	Inductive Coupling Link
IoT	Internet of Things
IP	Intellectual Property
LDO	Low Drop Out
LNA	Low-Noise Amplifier
LSB	Least Significant Bit
M3D	Monolithic 3D Integration
MCM	Multi-Chip Module
MDLL	Multiplying Delay Locked Loop
MEMS	Micro-Electromechanical Systems
MI	Mutual Inductance
MIM	Metal Insulator Metal
MIV	Monolithic Inter-tier Via
MSO	Mixed Signal Oscilloscope
MUX	Multiplexor
NF	Near Field
NFC	Near-Field Communication
NRZ	Non-Return to Zero
NVM	Non-Volatile Memory
PAM	Pulse Amplitude Modulation
PCB	Printed Circuit Board
PCM	Phase Change Memory
PLL	Phase Locked Loop
PPM	Pulse Position Modulation
PRBS	Pseudo-Random Binary sequence

PVT	Process Voltage Temperature
QFN	Quad Flat No-leads
QFP	Quad Flat Package
RC	Resistor-Capacitor
RF	Radio Frequency
RFID	Radio-Frequency Identification
RIE	Reactive Ion Etching
RO	Ring Oscillator
RRAM	Resistive Random Access Memory
RX	Receiver
SA	Sense Amplifier
SAFF	Sense Amplifier Flip Flop
SET	Spike-latency Encoding Transceiver
SiP	System in Package
SMU	Source Meter Unit
SoC	System-on-Chip
SPM	Single-Phase Modulation
SR	Set-Reset
STT-RAM	Spin-Transfer-Torque Random Access Memory
TSV	Through Silicon Via
TX	Transmitter
VLSI	Very Large Scale Integration
WBI	Wireless Bus Interface
WPT	Wireless Power Transfer

Nomenclature

χ_s	Track spacing graduation coefficient
χ_w	Track width graduation coefficient
$C_{i,j,k}$	Capacitance of segment k , of turn j , of inductor i [F]
t_c	Inductor metal thickness [m]
X	Communication distance [m]
k	Magnetic coupling coefficient
δ	TX pulse duration [s]
d	Inner diameter of inductor [m]
D	Outer diameter of inductor [m]
ℓ	Length of an inductor coil segment [m]
E_{pb}	Energy per bit [J]
η_{dat}	Efficiency of data ICL
η_{pow}	Efficiency of power ICL
f_{COUNT}	Counter frequency [Hz]
f_{clk}	Clock frequency [Hz]
f_{DAT}	Data frequency [Hz]
f_{hf}	BPSK carrier Signal Frequency [Hz]
ϕ	Inductor fill-factor
f	Link operating frequency [Hz]
f_{sr}	Self-resonant frequency [Hz]
g	Minimum technology grid unit [m]
$L_{i,j,k}$	Inductance of segment k , of turn j , of inductor i [H]
R_L	Load resistance [Ω]
M	Mutual inductance [H]
ω	Angular frequency [rad/s]

N	Number of bits per TX pulse
I_p	TX pulse amplitude [A]
$R_{i,j,k}$	Resistance of segment k , of turn j , of inductor i [Ω]
I_{RX}	Receiver current [A]
L_{RX}	Inductance of RX inductor [H]
V_{RX}	Received voltage [V]
s	Spacing between inductor tracks [m]
V_{DD}	Supply voltage [V]
I_{SL}	Current of transceiver supporting logic [A]
I_{TX}	Transmitter current [A]
L_{TX}	Inductance of TX inductor [H]
V_{TX}	Transmit voltage [V]
λ	TX Signal wavelength [m]

Chapter 1

Introduction

For over 40 years Moore’s Law has accurately predicted, and underpinned, the biennial doubling of device density within Integrated Circuits (ICs) [1]. However, as conventional channel length scaling progresses beyond the 10nm technology node, power and performance returns of traditional scaling are becoming incommensurate. Smaller gates require denser finer-pitch interconnect, and hence global Resistor-Capacitor (RC) interconnect delay has become the limiting factor for the performance of Complementary Metal Oxide Semiconductor (CMOS) ICs [2]. To overcome these obstacles, the semiconductor industry has explored a diversity of more-than-Moore technologies, each of which augment the performance of conventional ICs, to prolong the yearly incremental performance improvements expected by consumers. One such more-than-Moore technology is three-dimensional (3D) integration [3]. In 3D integrated circuits, active devices are not restricted to a single plane, moreover multiple planes (or *tiers*) of semiconductor dies are stacked and interconnected vertically. This has the effect of increasing transistor density and reducing long global interconnect traces to shorter vertical hops, improving performance [4].

Another, often overlooked, benefit of 3D integration is the facility that it offers for combining multiple different process technologies within a single *heterogeneous* IC. Heterogeneous 3D integration has the potential to unlock a new era in Very Large Scale Integration (VLSI) where designers can merge Micro-Electromechanical Systems (MEMS), low power CMOS, Radio Frequency (RF) Bi-CMOS and novel Non-Volatile Memory (NVM) technologies (such as Phase Change Memory (PCM), Spin-Transfer-Torque Random Access Memory (STT-RAM), and Resistive Random Access Memory (RRAM)) together within a single IC. This makes it a promising enabling platform for Internet of Things (IoT) devices which must often bring together these disparate elements within strict power and size budgets [5].

Unfortunately, however, the ideal of using 3D integration to construct highly-integrated heterogeneous IoT chips (illustrated by Figure 1.1) has not yet been realised. The commercial use of 3D integration is presently limited to homogeneous memory stacking applications (for example stacked DRAM [6–8] or Flash [9]) and examples of technologically heterogeneous 3D-ICs are sparse, even within the research community. Whilst this is in-part due to

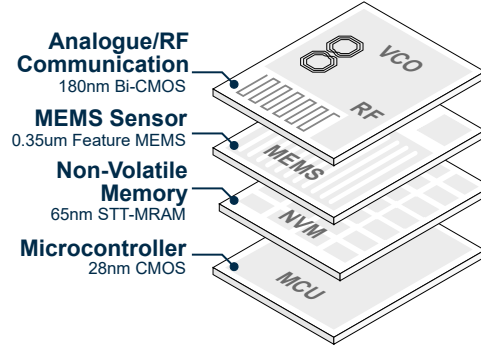


Figure 1.1: Conceptual illustration of a heterogeneous stacked 3D-IC suited for IoT sensing and compute-at-the-edge applications.

the novelty of the 3D integration concept (in addition to the lack of EDA support for designing 3D chips), when considering IoT applications (where minimising manufacturing costs is of paramount importance), it is primarily due to the high costs of designing and fabricating 3D-ICs, in addition to the limited number of TSV-enabled process nodes enabled by foundries¹.

To address this challenge, this thesis explores a *low-cost* alternative to existing 3D integration methodologies, achievable at *any* process technology node: using wireless vertical links to communicate data and power between tiers. The use of wireless communication means that an existing 2D fabrication processes can be used, without any alteration, making development straightforward, and possible across all technologies. In addition to this, the use of wireless communication means that the lateral placement tolerance (between stacked dies) is significantly relaxed, facilitating low-cost pick-and-place 3D assembly. Considering these benefits, the main appeal of Inductive Coupling Link (ICL)-based 3D integration is in applications that require low cost, heterogeneous stacking such as the IoT applications discussed above.

1.1 3D Integration Approaches

Broadly, three dimensional integrated circuits are defined as: “*Integrated circuits containing multiples layers of interconnected active devices*” [11]. Whilst *stacking* multiple tiers of active devices is a relatively straightforward task, establishing a method of *connecting* these vertically integrated tiers poses a more significant challenge. Early research in the field of 3D integration proposed innovative and controversial approaches to interconnect silicon dies, with some of the very first publications suggesting vertical communication between using “*photonic light beams passing through the silicon wafer*” (1987 [12]) or using “*mechanical spring clips*” (1990 [13]). However, as time has progressed, the research community has

¹For example, TSMC’s ‘Memory Cube’, their only *true* 3D fabrication process (using TSVs to stack several tiers) is only available at nodes <28nm [10].

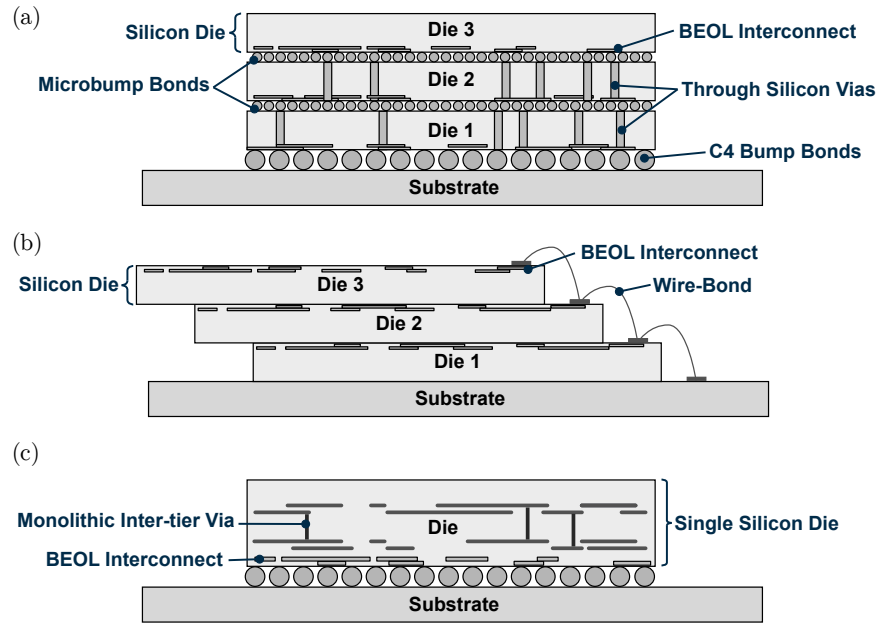


Figure 1.2: Conceptual illustration of existing 3D integration approaches including (a) a 3-tier 3D-IC assembled using TSVs to provide vertical connectivity between dies, (b) a 3 tier 3D stacked SiP using wire-bonds to interconnect stacked tiers, and (c) a monolithic 3D-IC with three sequentially fabricated active silicon tiers.

become increasingly focussed on a limited number of 3D assembly approaches which are outlined in the following sub-sections.

1.1.1 Through Silicon Vias

Over recent years, Through Silicon Vias (TSVs) have become synonymous with 3D integration. Within the context of 3D-IC design, the number of TSV-related publication eclipse those of competing approaches with approximately three-quarters of published works assuming TSV-based stacking. TSVs are electrical connections that pass entirely through the silicon substrate from front to back [14], allowing several dies to be stacked using microbumping in a fashion reminiscent of multi-layer Printed Circuit Boards (PCBs). Figure 1.2 (a) illustrates conceptually a 3D-IC assembled using through silicon vias to interconnect the Back End Of Line (BEOL) layers of each stacked die. TSVs can be fabricated in a number of ways including *via-first* (before Front End Of Line (FEOL) processing [15]), *via-middle* (after FEOL processing, but before BEOL processing [16]), or *via-last* (after BEOL processing [17]), however broadly, incorporating Through Silicon Vias (TSVs) necessitates deep Reactive Ion Etching (RIE) of the substrate, deposition and plating (of the conductive via material, typically polysilicon, copper, or tungsten), and aggressive Chemical-Mechanical Polishing (CMP) for thinning the wafer [18]. Compared to other vertical interconnect technologies, TSVs are a popular choice as they provide high density, high bandwidth connectivity between, potentially tens of stacked dies [19].

One drawback when using TSVs, however, is that they are presently only available at the latest foundry nodes, and are difficult to retrospectively include (using via-last fabrication) due to the high temperatures required for the annealing process [20]. This limits the facility for building the heterogeneous 3D-ICs discussed above. TSV-based 3D-ICs are also often also reported to suffer from yield and reliability challenges [21], arising from TSV/micro-bump deformation if precise alignment is not achieved between layers of the stack [22, 23].

Another significant drawback, particularly when considering IoT devices (which are highly cost-sensitive), is the high cost of TSV-enabled manufacturing processes [24]. Due to the additional fabrication and assembly stages associated with TSV-based 3D integration, it is estimated that, presently, the fabrication of a TSV-based 3D-IC will cost $1.4\text{--}2\times$ that of a comparable system in 2D [24].

1.1.2 3D SiPs using Wire Bonding

To address this, and facilitate 3D integration with low-cost (even at non-TSV-supporting process nodes), some researchers are exploring 3D assembly using wire-bonds between tiers [25]. One notable example where this is well applied in the IoT domain is the Michigan Micromote project [25–27], where multiple heterogeneous silicon tiers are stacked and integrated together for IoT sensing applications [28, 29]. Figure 1.2 (b) illustrates a stacked 3D-SiP assembled using wire-bonding. Here, each tier is placed upon the previous with a small lateral offset to allow access to the pads placed at the edge of the die. Wire-bonds are then used to interconnect the dies in the same way as the chip-to-package bonding process as illustrated in Figure 1.2 (b). This is, by far, the most mature (and hence cost-effective) technologies when considering 3D integration [30], but typically classed as three-dimensional *System in Package* solution. The distinction between 3D *packaging* (3D System in Package (SiP)) and 3D *Integrated Circuits* (3D-ICs) is widely debated however generally, 3D SiPs consist of multiple dies packaged together with only sparse package-level interconnect between them [19].

Whilst 3D-SiP assembly using wire-bonding can enable highly heterogeneous integration, the complexity of the required stacking arrangements, and the physical limits of wire-bonding (including loop-height, pitch, *etc.*) mean that assembling dies in this way quickly reaches practical limits. The ICs fabricated for the Micromote project ([25–29], discussed above) require custom, *manual* wire-bonding in an assembly process that cannot be scaled to mass production. Further to this, interconnects in wire-bonded SiPs must be placed on the periphery of each die, meaning that the achievable integration bandwidth is low and, due to their length, interconnect parasitics (capacitance and inductance) are high when using wire-bonds, limiting the die-to-die signalling speed [19].

Another 3D packaging approach which addresses some of these issues is flip-chip bonding, whereby fine pitch solder bumps are used to bond adjacent layers. When using flip-chip

bonding, landing pads are created on the surface of the recipient wafer or die which are then deposited with solder paste. The donor chip (which is to be bonded) is flipped and aligned with the recipient, before being heated to re-melt the solder bumps, resulting in an electrical bond between the two die. Whilst flip-chip bonding is an alternative SiP solution that can scale well in mass production (and can offer relatively high inter-tier interconnect bandwidths with achievable pitches between 50 μm and 200 μm [31]), one significant drawback of flip-chip bonding is that support is limited to only two tiers as dies are bonded in a face-to-face fashion [31].

1.1.3 Monolithic 3D Integration

Another approach to 3D integration involves sequentially fabricating multiple silicon layers, one after another, known as Monolithic 3D Integration (M3D). Of all the 3D integration approaches, M3D is the closest to the conceptual ideal of seamless 3D integration. In M3D, transistors are fabricated in sequential layers which are interconnected using fine-pitched Monolithic Inter-tier Vias (MIVs) [32]. Compared to other 3D-IC approaches which bring together each of the layers post-fabrication, the sequential M3D approach facilitates much higher connection density, allowing even for gate level vertical connectivity [33]. This is illustrated in Figure 1.2 (c).

The ability to fabricate several layers of active devices within an IC is clearly desirable. However presently, M3D is in very early stages of development, and is not a viable option for commercial fabrication. As an indication of the progress in this area, some of the leading research into monolithic 3D integration at the time of writing is being performed by the Taiwan Semiconductor Research Institute (TSRI). Their work, however, has only successfully managed to fabricate a handful of transistors (< 10) in the upper layer [34].

The biggest challenges facing M3D are the thermal budget constraints for sequential fabrication. Presently, it is not possible to utilise high processing temperatures ($> 400\text{-}500^\circ\text{C}$) after forming the copper BEOL interconnect of the ‘lower’ die [32, 35], however, such temperatures are essential for silicon epitaxy and reliable low-resistance gate formation [36]. One of the leading device-based research groups seeking to overcome this challenge is CAE-Leti, as part of their CoolCubeTM technology. Recent published work by CAE-Leti demonstrates the very first SRAM bit cells formed at low-temperatures ($< 500^\circ\text{C}$), in addition to a functional 81 stage ring oscillator fabricated end-to-end using process temperatures not exceeding 500°C [37]. This is achieved by: (1) performing in-situ dopant activation using an excimer laser (to lower gate resistance), and (2) replacing the traditional epitaxial pre-bake stage (HF-Last cleaning and 650°C bake) with an in-situ Siconi NH_3/NF_3 remote plasma process that can be followed by a 500°C H_2 bake [38].

Whilst these recent breakthroughs have not been demonstrated as part of a monolithic 3D chip, the ability to fabricate functional and yielding devices within the temperature budgets

imposed by M3D represents a significant milestone, meaning that is only a matter of time before two-layer M3D is possible [39]. It is, however, important to note that M3D is a long way from being commercially viable at scale (due to the challenges related to yield). Further to this, because all processing stages in a monolithic 3D-IC must be agreeable with previous layers (including temperature, materials and mechanical stresses), it is unlikely that M3D will ever be an enabling technology for the truly *heterogeneous* 3D integration discussed above (to combine several different technologies in the same chip).

1.2 Wireless 3D Integration

One final 3D integration approach (which forms the focus of this thesis) is using wireless inter-tier links. Here, instead of physically connecting the dies within the 3D stack, data to be communicated between layers is encoded as an Electro-Magnetic (EM) field to allow contactless communication, in a way similar to Near-Field Communication (NFC)/Radio-Frequency Identification (RFID) tags used in contactless smart cards.

To achieve reasonable power efficiency wireless inter-tier links typically operate in the near-field region of the electromagnetic (EM) spectrum [40] (where communication occurs across a distance less than the Fraunhofer distance, given by $2D^2/\lambda$, where λ is the wavelength of the communicated signal and D is the diameter of the antenna [41]). Approaches to wireless 3D integration in this near-field region can broadly be sub-divided into two main categories: communication through *inductive* coupling (which relies upon the modulation of a magnetic, **H**-Field) and communication through *capacitive* coupling (which relies upon the modulation of an electric, **E**-Field). As capacitive coupling is *voltage driven* (meaning that the communication distance increases proportionally to the drive voltage, $V_{RX} \propto C dV_{TX}/dt$) and inductive coupling is *current driven* (meaning that the communication distance increases proportionally to the drive current, $V_{RX} \propto L dI_{TX}/dt$), ICLs are usually favoured. This is because it is much easier to increase the drive current on-chip (simply by increasing the channel width (or number) of driver transistors), than it is to increase the drive voltage beyond VDD (which requires complex structures such as charge pumps *etc.* [42])². This means that ICLs can offer significantly enhanced communication distances when compared with Capacitive Coupling Links (CCLs) which are typically limited to Face-to-Face (F2F) stacking arrangements.

When performing contactless 3D integration using ICLs, data is transmitted via electromagnetic (EM) coupling between transmit (TX) and receive (RX) coils fabricated in the upper BEOL interconnect layers of each die as illustrated in Figure 1.3. Typically, to maximise dI_{TX}/dt , whilst minimising energy dissipation, pulse-based modulation schemes are used, where the data is encoded into a series of short, discrete current bursts. As these current

²A more complete comparison of each of these approaches (wireless 3D integration using capacitive coupling and wireless 3D integration using inductive coupling) is provided in Chapter 2.

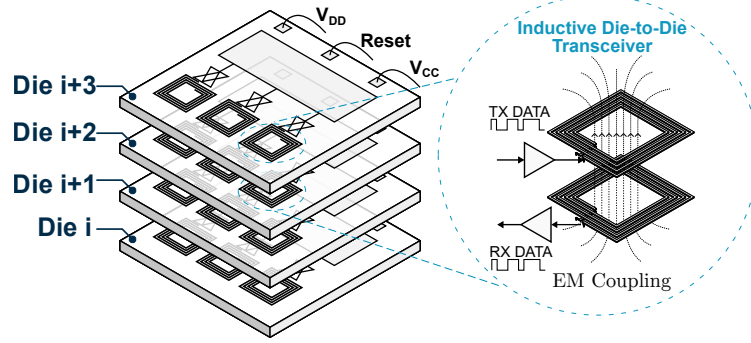


Figure 1.3: Illustration of a multi-tier wireless 3D-IC vertical inter-die implemented using near-field inductive coupling links (ICLs).

pulses flow through the Transmitter (TX) coil, a magnetic field is formed [43]. Provided that the Receiver (RX) coil intersects this magnetic field, a corresponding current (and hence voltage) will be induced, according to Faraday’s law [44]. This induced voltage can be detected and decoded in the recipient die to recover the transmitted data. This process allows communication between several Face-to-Back (F2B) stacked dies [45] without the need for 3D-specific nanotechnology, or precise alignment between stacked tiers.

When compared to the use of TSVs, this *wireless* approach to 3D integration promises several potential benefits, one of which being significantly reduced fabrication costs. The inductors required for ICLs can easily be fabricated in the BEOL interconnect layers of existing CMOS processes with no alteration. These standard processes are highly refined, and hence fabrication using this method is much less expensive.

The economic benefits of using ICLs are also evident at the assembly/packaging stage. Whilst 3D-ICs assembled using TSVs require sub-micron pick and place accuracy to ensure reliable operation [22] (in addition to wafer bumping and then micro-bump bonding to interconnect layers), 3D-ICs assembled using ICLs can simply be picked and attached using adhesive, with a relatively relaxed placement tolerance (in the order of 10’s of micro meters [24]). This means that, overall, manufacturing is significantly cheaper. ICLs can also be integrated at almost *any* process technology (unlike TSVs which are presently only available at the most advanced foundry nodes, as discussed previously). The ability to go off-menu with respect to foundry process technologies, thereby allowing each separate functional element in the 3D system to be fabricated in the most cost-effective way, makes ICLs well suited for low-cost heterogeneous integration applications such as the IoT stack illustrated conceptually in Figure 1.1 [46].

Another benefit of performing 3D integration using ICLs, particularly when considering IoT-style applications is the fact that they offer a high level of customisability. For example, disparate sensor/processing/memory/logic dies (each of which manufactured by different vendors), could be designed to communicate using the same wireless ICL interface (even

offering the possibility of intrinsic wireless voltage level conversion between dies operating at different supply voltages [46]), thereby allowing system-level designers to ‘pick-and-mix’ the circuit blocks that they require. These blocks could then be stacked in a customised 3D-IC, tailored for a specific application, with minimal cost and effort. This is presently not possible using TSVs as the physical specifications of via placement/pitch/diameter *etc.* vary between vendors.

One final benefit of using ICLs is that design cycle times (which is another significant concern for IoT devices) can be significantly reduced compared with TSVs, due to the fact that ICLs can be designed and signed off using existing *planar* (2D) Electronic Design Automation (EDA) tools. Designing with TSVs introduces a whole range of new requirements for partitioning, Automatic Place and Route (APR) and Design Rule Checks (DRCs), *etc.* all of which are immature at the moment.

For these reasons, wireless 3D integration using inductive coupling links (ICLs) has been identified as a compelling alternative to TSVs, to enable the heterogeneous computing ideals discussed above, with little additional cost when compared to standard planar ICs.

1.3 Research Justification

Despite these potential benefits, wireless 3D integration does have several drawbacks. One of the biggest criticisms of wireless 3D integration is its inferior power efficiency when compared with traditional contact-based approaches [47]. For markets like the IoT (where devices typically operate from limited battery, or Energy Harvesting (EH) sources), maintaining low power consumption is essential [48]. Because of this, the use of wireless 3D integration has not yet been considered, despite its potential economic benefits. Most works presenting ICL transceivers prioritise performance (in terms of link-bandwidth) above energy-efficiency [43, 49, 50]. One focus of this thesis, therefore, will be exploring energy-efficient ICL transceiver design in order to reduce the energy consumption of existing ICL transceivers to an acceptable level for Internet of Things devices.

Extending this further, when considering the sources of power consumption within an ICL transceiver, by far the most energy inefficient component is the inductive channel itself [51], with a significant proportion of the energy used to form the **H**-Field often being wasted [52]. To maximise the energy efficiency, therefore, it is essential that the layout/geometry of the inductors used to form the ICL channel (for example shape, track-width, track-spacing, number of turns *etc.*) is optimised [51]. Presently, this involves using Finite Element Method (FEM) tools for EM analysis, and then converting the system’s EM characteristics into equivalent circuit models that can be handled by electrical simulators (*e.g.* SPICE) [52]. The layout can then be manually adjusted, and the process repeated until a satisfactory solution is found. Solvers using FEM, however, often take several hours to converge, even whilst

analysing a single geometry [53]. Due to this, determining coil pairs with optimised geometries (which typically necessitates analysing thousands of layouts) is extremely computationally expensive, if not impossible. Another focus of the research in this thesis will, therefore, be the ICL inductor design process, with the target of achieving rapid inductor layout optimisation to maximise power efficiency.

Whilst there are a range of prior works exploring the use of wireless *data* communication between tiers of a 3D stack [43, 49, 50, 54–56], to make the stacking process truly wireless (and avoid the need for, and hence cost of, die-to-die wire bonding completely) it is also necessary to perform wireless *power* delivery between tiers. Presently, most 3D-ICs using ICLs rely on wire-bonded power and ground connections to supply power to each die in the stack [43, 55, 57, 58]. Whilst this is an adequate solution that circumvents the use of TSVs, the addition of wire-bonds to each tier of the 3D stack undermines many of the benefits associated with *wireless* 3D integration and inflates the assembly cost and complexity. Because of this, a handful of works also explore using Wireless Power Transfer (WPT) within the chip [59–61]. All of these works, however, separate wireless power transfer and data communication across two inductive channels, meaning that a minimum of four inductors are required for a given system [59–61]. One often-faced criticism of using ICLs for 3D integration is their high silicon area usage (due to the relatively large footprint of the on-chip inductors) [62]. Doubling the number of channels (to perform power *and* data transfer) exacerbates this problem, especially considering that WPT links typically require larger channels (in some cases consuming up to as much as 4mm^2 of silicon area [61]). To address this issue, the third research theme of this thesis is to explore new approaches for wireless power delivery in contactless 3D-ICs, with a focus on area efficiency.

In a similar vein, to achieve fully wireless 3D integration, especially when implementing a coherent data transceiver, it is necessary to achieve precise clock synchronisation between the transmitting and receiving dies. In many prior works, clock delivery is performed externally using wire-bonding [19]. Although this is an adequate solution, the addition of wire-bonds suffers from the same issues discussed above and is incompatible with fully-wireless 3D integration. Further to this, the parasitic overheads of the pad drivers (and *RLC* parasitics of each bond-wire) mean that the clock frequency that can be achieved is limited. The use of coupled resonators has been proposed to address this issue, delivering the clock wirelessly between tiers [43, 63–65]. Here, *LC* tanks with a resonant frequency corresponding to the clock frequency are used, where the inductive (*L*) component is part of a coupled inductive link [63]. The use of coupled resonators is a promising solution that allows the clock to be wirelessly transmitted between dies with low jitter and skew (due to the natural harmonics of the link), however such links must either be very high-frequency or very large in terms of area (as the resonant *LC* decreases with diameter, *D*, of the inductor, *L*). To achieve clock distribution at frequencies in the order of 100’s of Mega Hertz using this approach

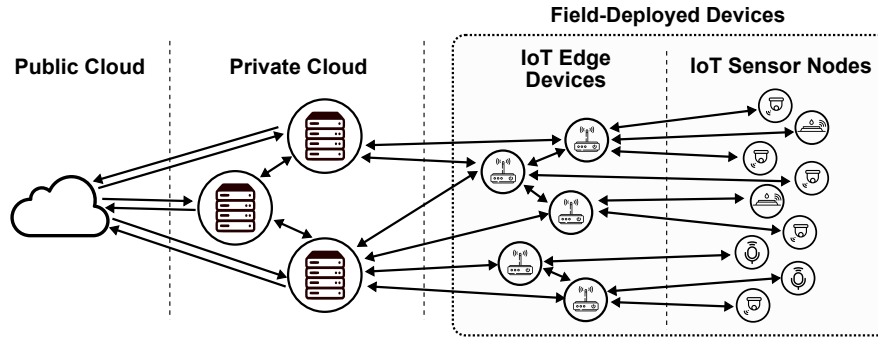


Figure 1.4: A typical Internet of Things (IoT) fog network (reproduced from [67]).

(which is typical for IoT applications) often requires inductors with diameter $> 300\text{ }\mu\text{m}$ [64]. Motivated by this, the fourth theme of this research will be clock distribution in wireless inductive 3D-ICs, again focussing on area efficiency.

Finally, although there are already several impressive works demonstrating many-tier vertical interconnection using ICLs, the number of publications demonstrating such links as part of a fully-functional system are limited [51] and even when full system integration is performed (for example [66]), custom non-standard communication protocols are used. For IoT applications, short time-to-market is essential and hence it is desirable to integrate functional blocks in a standardised way. The final research focus of this thesis, therefore, will be developing a standardised ICL interface to enable ‘pick-and-mix’ stacking of different circuit blocks (in the form of separate semiconductor dies) with short design times and low-cost.

1.4 Research Context

In addition to outlining the *justification* for this research, it is also important to discuss the *context* of the research and the Power, Performance and Area (PPA) constraints for Integrated Circuits (ICs) used in low-cost Internet of Things (IoT) devices. Figure 1.4 (reproduced from [67]) illustrates the topology of a typical IoT fog network [68]. As shown, a hierarchy of devices exists, with data initially being collected by small embedded systems or sensors, collated by larger field-deployed edge devices and then being processed on cloud-based servers [67, 68]. It is estimated that, by 2030, there will be 24.1 billion field-deployed IoT devices and so establishing ways of manufacturing the ICs in these devices, with low-cost, is of paramount importance [69] (hence forming the motivation for this work).

As shown in Figure 1.4, field-deployed IoT devices can broadly be categorised in two main classes:

- **Wireless Sensor Nodes** - Small sensor devices that gather data about their surroundings and communicate it through the network (e.g. body-worn sensors for healthcare applications [70], traffic sensors for city-management [71], or asset tracking devices

IoT Sensors			
Metric	Importance	Typical Specification	Comments
IC Die Size	High	0.5-5mm ² [25, 27, 74]	IC size is important for IoT sensors which are often deployed constrained locations (such as labels [72], implanted devices [75] or other in-situ sensing/monitoring devices [76]).
IC Power Consumption	Very High	<50mW [25, 27, 74, 75]	Due to the volume constraints of IoT sensors, energy storage is typically limited. This means that minimising power consumption is of very high importance, to avoid the need for frequent battery replacement/recharging [77].
IC Clock Frequency	Low	<10MHz [25, 27, 74, 78]	Most data is processed away from the sensor, so high-frequency processing is not required within IoT sensor ICs [75].
Heterogeneity / Functional Components	High	MEMS, Wireless Communications, Non-Volatile Memory, Energy Storage, Energy Harvesters [5]	IoT sensors typically require a high level of heterogeneous integration to combine sensing dies (e.g. using MEMS), memory and power management/storage circuits [5].

IoT Edge Devices			
Metric	Importance	Typical Specification	Comments
IC Die Size	Medium	5-20mm ² [79–81]	Edge devices must be low-cost and remotely deployable which means that typical IC die sizes are limited (when compared with, for example, desktop or server-scale compute dies) [67].
IC Power Consumption	High	50-200mW [79–81]	These devices are typically battery powered (often with energy harvesting top-ups), meaning that power efficiency is of high importance. However, as edge devices are slightly larger than typical sensors, their power budgets are usually more generous [77].
IC Clock Frequency	Low	200-800MHz [79–81]	IC Clock frequencies for edge devices must be sufficient to perform useful local data processing and networking. However, typical frequencies are much lower than most consumer electronic devices to conserve energy [79].
Heterogeneity / Functional Components	Medium	CPU, DRAM, Networking, Wireless Communications	Edge devices require integration of networking, processing and memory components which may each be fabricated in different technologies to reduce the overall device cost and/or energy consumption [82].

Table 1.1: Tabular outline of the typical PPA constraints for Integrated Circuits (ICs) used in (i) IoT *edge* device ICs and (ii) IoT *sensor* device ICs.

[72]).

- **Edge Devices** - Field deployed hubs to collate and process data from several sensors before forwarding it to web-servers in the cloud [67, 68, 73].

The typical requirements of the ICs used in these two classes of IoT device are discussed in the paragraphs below, with a concluding tabular summary presented in Table 1.1.

Wireless sensor nodes must often adhere to very strict size/volume constraints, particularly when integrated within small objects (e.g. in labels used for asset tracking [72]) or when deployed in confined spaces (e.g. in healthcare applications, where devices must be implanted into the human body [70]). Such ICs must often also combine a range of different heterogeneous elements and process technologies such as MEMS (for performing sensing), NVM (for

storing sensor data), and wireless networking. As discussed above, this makes 3D integration a highly promising enabling technology for constructing wireless sensor node ICs [5]. Due to their stringent volume constraints however, energy storage, and hence power budgets for these devices are typically very limited, meaning that minimising energy consumption is often of paramount importance. State-of-the-art battery technologies can provide power densities between 670mWh/cm^3 (Li coin cell/alkaline battery [77]) and 760mWh/cm^3 (Li-ion 18650 battery [77]) which means that, as an example, a small $\sim 5\text{cm}^3$ sensor which operates intermittently for a total of 2 hours per day is limited to a power budget of around 5mW (assuming it must last for at least one year without battery replacement). As wireless sensor nodes do not typically perform any processing of the gathered data, operating frequency is a low priority for these ICs. Existing implementations of 3D stacked IoT sensors operate with very low clock frequencies (73kHz [27] - 500kHz [78]), sometimes with short high-frequency bursts to handle wireless communication via protocols like Zigbee or Bluetooth [71].

In contrast with this, IoT edge devices (which do perform data processing) typically operate with clock frequencies in the order of 100's of Mega Hertz [79–81]. Although not bound by the same stringent volume constraints as many IoT sensors, IoT edge devices are typically battery powered and often supplement their charge using energy harvesting. This results in slightly larger power budgets than discussed previously for IoT sensor nodes (in the region of $50\text{--}200\text{mW}$ across previously published works [79–81]), but still means that energy efficiency is a very important design factor. Although edge devices do not need to incorporate sensor dies, they can still benefit from heterogeneous 3D integration for bringing together wireless communication, non-volatile memory, and digital processing elements with high energy efficiency and low cost.

Table 1.1 provides a summary of these requirements (for both IoT sensor and edge devices). Overall, it is clear that alongside manufacturing cost, energy consumption is the most important constraint when considering integrated circuit design for IoT devices. Low-energy ICL implementation will therefore form a key focus of the research presented in this thesis. Secondary to cost and energy efficiency, silicon area is also a significant factor, particularly in IoT sensor devices which are deployed in constrained environments. Therefore, ways of improving area-efficiency when designing inductive coupling links will also be explored as part of this work.

1.5 Research Questions

The above discussion motivates the following six research questions which will be answered in this thesis:

1. How is it possible to reduce the energy consumption of existing ICL transceivers for use in IoT devices?

2. How can the search process for finding optimised inductor geometries for inductive coupling link applications be automated? Subsequently, what techniques can be used to evaluate a given inductor geometry faster than using finite element modelling?
3. Is it practical to design an ICL transceiver for use in 3D stacked ICs that performs both wireless data transfer and wireless power transfer?
4. How should clock distribution be performed in a many-tier, wirelessly stacked 3D-IC?
5. What is the sensitivity of inductive coupling links to die-to-die stacking misalignment during the packaging process?
6. For the IoT, customisability is important. Is it practical to design ICLs in a standard way to allow interchangeable stacking with a range of different memory/sensor/logic dies?

1.6 Research Contributions

Towards addressing these research questions, the major contributions of this thesis include:

- **A Low-Energy Inductive Transceiver Using Spike-Latency Encoding**

The first contribution of this thesis, to address [Research Question 1](#) is a novel, low-energy inductive transceiver that uses a time-domain encoding technique to represent frames of data in terms of the *latency* between sequential pulses. This reduces the number of transmit pulses required to send given bit stream, hence reducing the overall system energy consumption. The presented transceiver also includes a tuneable pulse-based transmitter which allows the transmit current to be precisely adjusted, post-fabrication, to compensate for stacking assembly defects (such as uneven die-thinning, or stacking misalignment), also addressing [Research Question 5](#).

- **COIL-3D**

The second contribution of this thesis is COIL-3D, A CAD tool for the Optimisation of Inductive Links in 3D ICs. For power efficient design of inductive coupling links, it is essential that the layouts of the coupled inductors (forming the EM channel) are optimised. COIL-3D is a software tool³ that utilises a rapid solver based upon semi-empirical expressions to quickly and accurately characterise a given link, in conjunction with a high-speed refined optimisation flow (also developed as part of this work) to find optimal inductor geometries to maximise efficiency in ICL channels. COIL-3D is developed to address [Research Question 2](#), related to optimisation of ICL channels.

- **CoDAPT**

³Available for open-source download at <https://github.com/bjflg13/coil-3d>.

The third contribution presented in this thesis, is a novel ICL transceiver for Concurrent wireless Data And Power Transfer (CoDAPT). To address [Research Question 3](#), the CoDAPT transceiver is designed to perform simultaneous wireless data and power transmission, through a single inductive channel, using a Bi-Phase Shift Keying (BPSK) scheme. WPT is achieved using the high-frequency BPSK carrier signal, whilst data is encoded by modulating and sampling the phase. Combining wireless power and data delivery in a single inductive channel results in significant area savings when compared with prior works that use separate links for data and power transmission.

In addition to this, to address [Research Question 6](#) (focussed on maintaining a high level of customisability), the CoDAPT link is implemented as part of a standard AHB-lite bus to realise a pseudo-System-on-Chip (SoC) architecture⁴. This allows, for example, a range of different standard AHB sensor/memory/processor Intellectual Property (IP) blocks to be vertically interconnected in a standardised way, with the CoDAPT ICL forming the main system bus.

- **WiSync**

The final contribution of this thesis is WiSync, a low-area, inductive link for Wireless clock Synchronisation in many-tier 3D-ICs. The WiSync transceiver directly addresses [Research Question 4](#), facilitating low-skew clock distribution across several stacked silicon tiers. The practical implementation of the WiSync transceiver is also used for a study exploring the effects of die-to-die stacking alignment on link performance, thereby also addressing [Research Question 5](#) which is focussed on quantifying the sensitivity of ICLs to errors/variations during the packaging process.

1.7 Publications and Patents

Some of the research presented in this thesis has also published and patented, as outlined in the following sub-sections. The first part of Section 1.7.1 lists the peer-reviewed (conference and journal) publications completed during my PhD candidature that are directly reported as contributions in this thesis, whilst the second half of the list outlined peer-reviewed publications that were completed during my PhD candidature, but are not directly reported as contributions in this thesis. Section 1.7.2 outlines the patents submitted as a direct result of this work and finally, Section 1.7.3 list other academic engagement activities stemming from the research presented in this thesis such as awards.

1.7.1 Peer-Reviewed Publications

1. B. J. Fletcher, S. Das, T. Mak, “A Spike-Latency Transceiver with Tuneable Pulse Control for Low-Energy Wireless 3D Integration”, *IEEE Journal on Solid State Circuits (JSSC)* 55(9) pp.

⁴The ‘pseudo’-SoC refers to the fact that, *architecturally*, the components form a standard System-on-Chip, but *physically*, the SoC elements exist in separate wirelessly connected dies.

- 2414-28 (2020). [83]
2. B. J. Fletcher, T. Mak and S. Das, "A 3D-Stacked Cortex-M0 SoC with 20.3Gbps/mm² 7.1mW/mm² Simultaneous Wireless Inter-Tier Data and Power Transfer," *IEEE Symposium on VLSI Circuits*, Honolulu, Hawaii, 2020 pp. 1-2. [84]
 3. B. J. Fletcher, S. Das and T. Mak, "A 10.8pJ/bit Pulse-Position Inductive Transceiver for Low-Energy Wireless 3D Integration," *IEEE European Solid State Circuits Conference (ESSCIRC)*, Cracow, Poland, 2019, pp. 121-4. [85]
 4. B. J. Fletcher, S. Das, T. Mak, "A Low-Energy Inductive Transceiver using Spike-Latency Encoding for Wireless 3D Integration," *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Lausanne, Switzerland, 2019, pp. 1-6. [86] (**Best-Paper Award Winner**)
 5. B. J. Fletcher, S. Das and T. Mak, "Design and Optimization of Inductive-Coupling Links for 3-D-ICs," *IEEE Transactions on Transactions on Very Large Scale Integration (VLSI) Systems* 27(3) pp. 711-23 (2019). [87]
 6. B. J. Fletcher, S. Das, T. Mak, "CoDAPT: a concurrent data and power transceiver for fully wireless 3D-ICs," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Florence, Italy, 2019, pp. 1343-48. [88] (**Best-Paper Award Nominee**)
 7. B. J. Fletcher, S. Das, T. Mak, "Cost-effective 3D integration using inductive coupling links: Can we make stacking silicon as easy as stacking Lego?," *Arm Research Summit 2018*, Cambridge, United Kingdom. 17 - 19 Sep 2018. [89]
 8. B. J. Fletcher, S. Das, T. Mak, "A High-Speed Design Methodology for Inductive Coupling Links in 3D-ICs," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, Germany, 2018, pp. 497-502. [90]
 9. B. J. Fletcher, S. Das, T. Mak, "Low-Power 3D Integration using Inductive Coupling Links for Neurotechnology Applications," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, Germany, 2018, pp. 497-502. [91]

During the time of my PhD candidature, I have also authored the following peer-reviewed papers, however these are not reported as contributions within this thesis:

11. D. Balsamo, B. J. Fletcher, A. J. Weddell, G. Karatzias, B. Al-Hashimi and G. V. Merrett, "Power neutral performance scaling with intrinsic MPPT for energy harvesting computing systems," *ACM Transactions on Embedded Computing Systems*, 17(6), 1-25 (2019). [92]
12. B. J. Fletcher, D. Balsamo and G. V. Merrett, "Power-neutral performance scaling for self-powered multicore computing systems," *Adaptive Many-Core Architectures and Systems Workshop*, York, United Kingdom, 13-15 Jun 2018. [93]
13. Q. Ding, B. J. Fletcher, T. Mak, "Globally Wireless Locally Wired: A Clock Distribution Network for Many-Core Systems," *IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, Italy, 2018, pp. 1-5. [94]
14. B. J. Fletcher, D. Balsamo and G. V. Merrett, "Power neutral performance scaling for energy harvesting MP-SoCs," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, Lausanne, Switzerland, 2017, pp. 1516-21. [95]

1.7.2 Patents

In addition to the publications listed below, the work presented in this thesis has also led to the filling of two patents:

15. B. J. Fletcher, J. Myers, S. Das and T. Mak, “*A Pseudo System-on-Chip Architecture Incorporating Wirelessly Connected Bus Slaves.*”, U.S. Patent 16/685,090, Nov 2019 (pending). [96]
16. S. Gamage, B. Fletcher, and S. Das, “*Adaptive Coding for Wireless Communication.*”, U.S. Patent 16/656,937, Oct 2019 (pending). [97]

1.7.3 Other Research Engagement Activities

Finally, the work presented in this thesis has directly resulted in the following awards.

17. B. J. Fletcher, S. Das, and T. Mak “3D Integration Using Wireless Inductive Links,” *Science Technology and Mathematics for Britain Exhibition*, London, Mar. 2020. (**IEEE ComSoc Award Winner**)
18. B. J. Fletcher, “Can we make stacking silicon as easy as stacking Lego?,” The Institution of Engineering and Technology (IET) Postgraduate Prize, 2018. (**Winner**)

1.8 Thesis Outline

The remaining content of this thesis is organised as follows. Chapter 2 first provides background information on near-field wireless 3D integration, summarising the prior art using capacitive coupling transceivers (Section 2.1) and inductive coupling transceivers (Section 2.2). Chapter 2 also discusses ICL channel layout optimisation (Section 2.2.3), wireless power transfer in 3D-ICs (Section 2.4) and near-field clock delivery (Section 2.3).

Following this, Chapter 3 focusses on low-energy transceiver design, presenting a time-domain modulation scheme (spike-latency encoding) to facilitate low-energy data consumption in wireless 3D-ICs. This is the first instance of time-domain coding in the context of die-to-die communication, resulting in significant energy savings when the approach is evaluated (Section 3.4). A tuneable pulse driver circuit is also presented in Chapter 3, which allows the transmit pulse energy to be tuned to an absolute minimum (post-assembly), accounting for any variations in the manufacturing or stacking process.

Having explored the energy optimisation in the transceiver circuits in Chapter 3, Chapter 4 then focusses on energy optimisation of the inductive channel itself. A set of strictly solvable mathematical expressions are presented for evaluating a link’s performance directly from its layout parameters, alongside a refined optimisation flow (Section 4.3.2) as a high-speed alternative to manual geometry optimisation using FEM. Comparison of different inductor

shapes (square, circle, octagon) in terms of bandwidth and energy efficiency is also presented in this chapter.

Chapters 5 and 6 focus on enabling fully wireless 3D integration, exploring wireless clock distribution and power delivery respectively. The wireless many-tier clock link presented in Chapter 5 uses a dual-mode transmitter to conserve energy, whilst operating in the non-resonant portion of the link's frequency spectrum. As a result of this, the design can operate across a wide range of frequencies (making it suitable for use with a variety of different coherent data ICL designs, such as the spike-latency encoding transceiver from Chapter 3 and the CoDAPT transceiver in Chapter 6). Operating away from resonance also means that the silicon area can be significantly reduced, and as such the design presented in Chapter 5 achieves the lowest silicon area overhead ever reported for an inductive clock link. Section 5.4.4 of Chapter 3 also presents a practical study of die-to-die misalignment on the link's performance.

Wireless power delivery is then explored in Chapter 6. To minimise the ICL area overhead, this chapter presents the first instance of concurrent data and power delivery (through a single ICL channel). The presented transceiver uses a BPSK modulation scheme (where power is recovered from the *amplitude* of the carrier signal and the data is decoded from its *phase*) and is implemented as part of a full 3D-stacked Cortex M0 SoC with stacked microprocessor and memory. This represents the first integration of a wireless link as part of a SoC bus, and the smallest ever reported wireless power link.

Finally, Chapter 7 concludes the thesis and outlines potential avenues for exploration in future work (Section 7.2).

Chapter 2

Wireless Three-Dimensional (3D) Integrated Circuits

As discussed in the introduction, wireless 3D integration is a promising *low-cost* alternative to Through Silicon Vias (TSVs) for interconnecting stacked silicon dies [24] which avoids the need for post-fabrication processing (significantly reducing the manufacturing cost) [98], or 3D-specific EDA tools. Further to this, once manufactured, dies can be simply picked and stacked using adhesive, greatly reducing the assembly cost. This makes them an attractive option for Internet of Things applications, which are driven by cost and design-time, rather than performance [82].

To achieve reasonable power efficiency, wireless links for 3D integration typically operate in the near-field region of the electromagnetic (EM) spectrum (where the communication distance, X is less than $\lambda/2\pi$, and λ is the wavelength of the transmitted signal [41]). Broadly speaking, there are two near-field communication mechanisms, these are *capacitive*, and *inductive* coupling. Capacitive coupling relies on communicating data by modulating an *electric* field formed between the transmitter and receiver (similar to a parallel plate Metal Insulator Metal (MIM) capacitor). Conversely, inductive coupling relies on modulating a *magnetic* field between the transmitter and receiver (similar to a magnetic transformer). Both capacitive and inductive coupling links have been demonstrated for the purposes of 3D integration in prior-art [51, 99] and each has its own advantages and disadvantages. A comparison of these two approaches is provided in the following sections alongside a summary of the prior art.

2.1 Capacitive Coupling Links

Capacitive coupling links (CCLs) rely on communicating through a time-varying electric field between the transmitter and receiver [100]. This electric field is formed across two conducting plates, typically fabricated in the top-most Back End Of Line (BEOL) interconnect layer [99, 101, 102] (where an electric potential exists between the Transmitter (TX) plate and the

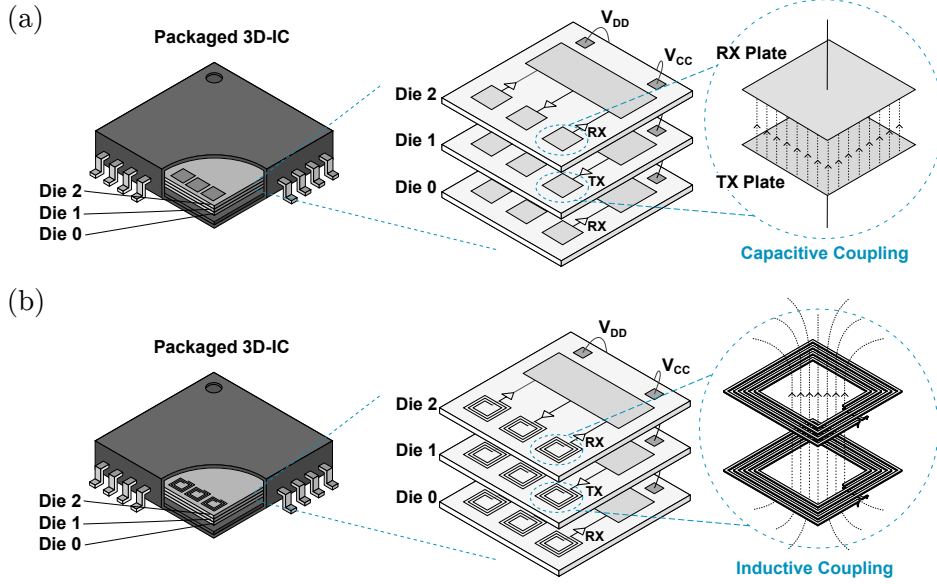


Figure 2.1: Figurative illustration of 3-tier 3D-IC assembled using: (a) Capacitive Coupling Links (CCLs) to communicate data between tiers (as discussed in Section 2.1), and (b) Inductive Coupling Links (ICLs) to communicate data between tiers (as discussed in Section 2.2).

Receiver (RX) plate, as shown in Figure 2.1 (a)). Between the two plates, a layer of dielectric adhesive is typically used to attach the dies in a face-to-face (F2F) arrangement [99], but also to provide the ‘insulator’ layer in the MIM capacitor structure. To communicate data across a CCL, the electric potential across the plates can be varied in accordance with the data stream. This will in turn modulate the electric field strength, which can be sensed and used to decode the transmitted data [100].


The efficiency of a capacitive coupling link (V_{RX}/V_{TX}), η_{CCL} , is given by Equation 2.1 below:

$$\eta_{CCL} = \frac{V_{RX}}{V_{TX}} = \frac{C_{Link}}{C_{Link} + C_{RX}} \quad (2.1)$$

where C_{Link} is the intentional capacitance of the CCL (formed by the upper and lower conducting plates) and C_{RX} is the parasitic capacitance of the receiving plate [99]. Assuming a plate size of $20\ \mu\text{m} \times 20\ \mu\text{m}$ (in line with prior publications in this domain [103]) and a load capacitance of 6fF [99], the efficiency can be calculated using Equation 2.1 in conjunction with the parallel plate capacitor Equation 2.2 [104] for approximating C_{Link} , shown below.

$$C_{Link} = \frac{\epsilon_0 \epsilon_{glue} A}{X} \quad (2.2)$$

Using this approximation yields an efficiency of 0.124×10^{-3} assuming a communication distance, X , of $10\ \mu\text{m}$. This means that, for $V_{DD}=3.3\text{V}$, the RX voltage amplitude would be in the region of $\sim 400\ \mu\text{V}$.



	R. J. Drost <i>et al.</i> [105] ('04)	A. Fazzi <i>et al.</i> [99] ('07)	M. Aung <i>et al.</i> [101] ('14)	K. Kanda <i>et al.</i> [103] ('03)	L. Luo <i>et al.</i> [106] ('03)
Communication Distance	8 μ m	1 μ m	2.5 μ m	1-2 μ m	<1 μ m
Process Technology	0.35 μ m	0.13 μ m	65nm	0.35 μ m	180nm
Energy Per Bit	3.90pJ	0.14pJ	0.05pJ	2.36pJ	5.00pJ
Max. Bandwidth	1.35 Gbps	0.93 Gbps	3.00 Gbps	1.27 Gbps	3.00 Gbps

Table 2.1: Comparison of prior works in the domain of contactless 3D integration using capacitive coupling.

Because of the low amplitude of the RX voltage in such systems, an RX-side amplifier is typically required. The amplification stage can be either implemented synchronously (*e.g.* using a sense-amplifier) or asynchronously (*e.g.* using a standard Low-Noise Amplifier (LNA) design). The synchronous amplification approach boasts better noise immunity (as the opportunity for noise to upset the receiver output is limited to the time window in which the transmit pulse is being sent), however requires precise clock synchronisation between the two stacked dies (typically achieved by transmitting the clock through a separate link).

A range of previous publications explore the use of CCLs for 3D integration (and Multi-Chip Module (MCM) integration), starting with the seminal work by D. Saltzman in 1994 [107]. Table 2.1 provides a summary of these prior publications summarizing their key performance figures. Work [105], by Drost *et al.* implements a bank of eight parallel asynchronous (or non-coherent) CCLs between two 0.35 μ m technology chips. The authors use a very simple transceiver implementation where the TX pad is driven by a standard-cell inverter ($2\times$ drive strength). The RX side also uses a standard-cell inverter as a voltage amplifier followed by a mirrored-inverter latch for stability in the output [105]. In this work, each chip is connected to an adjustable Vernier mount allowing the alignment and separation (between the two stacked dies) to be finely tuned. Using this approach, Drost *et al.* demonstrate successful communication across distances up to 20 μ m with this simple (all-digital, non-coherent) transceiver [105].

More recent works ([99], [103] and [106]) in the domain of wireless 3D integration using CCLs all implement synchronous (or coherent) links, where knowledge of the transmit clock is required in the receiver. In [99] and [106] the authors demonstrate Capacitive Coupling Links (CCLs) channels between two stacked chips (in 0.13 μ m CMOS and 180nm CMOS technologies respectively) using the same inverter-based transmitter discussed above, but this time incorporating a synchronous Sense Amplifier (SA) based receiver in the RX chip. The use of a *synchronous* receiver means that much smaller RX voltage amplitudes can be

accurately detected without interference from transient on-chip noise. This means that the transmit energy can be significantly reduced (for example, [99] achieves an energy-per-bit of 0.14pJ, a significant reduction compared to the 3.9pJ/bit achieved by the non-coherent scheme in [105] that was discussed previously). This energy saving in the *data-links*, however, comes at the expense of additional implementation complexity, as these designs rely upon accurate clock synchronization between the TX and RX dies. This can be achieved using external wire-bonding, or through a separate link where the clock signal is wirelessly transmitted in a similar manner to the data. With both approaches, clock recovery circuits (DLL, PLL etc.) are also typically required to ensure precise clock synchronisation and hence achieve high performance (for example using the multi-stage DLL adopted in [106])¹.

Motivated by the significant area overhead of the inductive plates needed to form the CCL channel (four of which are required for bi-directional communication; two for the uplink, and two for the downlink), in [101], M. Aung *et al.* presented a *bidirectional* CCL transceiver in a 65nm CMOS technology test-chip. Here, the same capacitive coupling channel is used for the uplink and downlink, meaning that the area overhead of the whole communication system is much reduced. To achieve the bi-directionality, in addition to a high data-rate per link (3Gbps), a four level Pulse Amplitude Modulation (PAM) voltage signalling scheme is used (with approximately 130mV between each signalling level). Here, two signalling levels are reserved and controlled by the up-link, and two signalling levels are reserved for, and controlled by, the down-link [101]. The transceiver is not tested in a stacked system, but ‘emulated’ by using two of the BEOL metal layers within a standard planar IC, and achieves high performance operating up to 3.0Gbps/channel [101].

As can be observed from Table 2.1, this is at the higher end of the data rates that can be achieved using CCLs; capacitive links reported in previous works are able to achieve typical data-rates in the range of 1.2 Gbps [103, 108] to 3.0Gbps [101] with a relatively small footprint (typically in the order of $300\mu\text{m}^2$). This makes them an attractive option in terms of area, particularly when compared with inductive coupling links which often require a much larger areas (often in excess of 0.02mm^2)². Another feature of CCLs is that they are voltage driven (the electric field across the TX and RX plates is proportional to the *voltage* applied across them³). Conventional digital electronic systems are also voltage driven (with signal levels usually being mapped to discrete voltages, rather than currents), meaning that transceiver design for such CCLs is very straightforward, and the capacitive plates can often be simply driven by standard logic cells [105].

Despite these advantages, the use of CCLs for wireless integration does have several drawbacks, most notably the communication distance. Whilst the voltage-driven nature of CCLs makes

¹A survey of existing literature focused on clock distribution in wirelessly stacked 3D-ICs is provided later, in Section 2.3.1.

²See Section 2.2 for further comparison.

³This contrasts with ICLs, where the magnetic field is proportional to the *current*.

them easy to integrate as part of a system, it also means that increasing the transmit power requires increasing the transmit voltage (V_{TX}). Although it is theoretically possible to increase V_{TX} beyond the nominal supply voltage of the chip using charge-pumps, such structures consume significant area, and hence prior practical implementations limit V_{TX} to the chip supply voltage (V_{DD}). This means that the range of such capacitive links is typically limited to a number of microns, facilitating only face-to-face stacking [109]. As a result, the number of tiers supported by this technology is typically limited to two, making it unsuitable for the heterogeneous many-tier stacking discussed in Chapter 1. Additionally, this two-tier stacking limit means that capacitive coupling 3D-ICs hold little advantage over, for example, flip-chip bonded SiPs which achieve a similar interconnect density [110] with relatively low fabrication cost.

2.2 Inductive Coupling Links

The use of inductive coupling (illustrated in Figure 2.1 (b)) overcomes this problem, typically allowing communication over a much greater range, facilitating face-to-back die stacking [43, 44, 49, 51, 111]. In contrast to CCLs, Inductive Coupling Links (ICLs) rely on a time-varying *magnetic* field to communicate between the transmitter and receiver, which is current driven (therefore easier to modulate on-chip, and not limited by the supply voltage). This field is formed across two coupled inductors (normally fabricated in the upper-most BEOL metal layers of the die [51]⁴) and modulated by varying the current flowing through the TX inductor (I_{TX}). As I_{TX} varies, a magnetic field will be formed; provided that the RX inductor intersects this field, a corresponding current (and hence voltage) will be induced in the RX coil, according to Faraday's law [44]. This voltage can then be amplified and/or sensed, and therefore used to decode the transmitted data stream [51].

The area footprint of ICLs varies depending on the communication distance, however inductors used in Face-to-Back (F2B) stacking arrangements are typically between $80\mu\text{m} \times 80\mu\text{m}$ [111] and $500\mu\text{m} \times 500\mu\text{m}$ [112] in size. As inductive coupling is current driven, an intrinsic trade-off between the inductor's area and the power consumption of the system exists. Communication bandwidths between 200kbps and 3Gbps (per link) [111, 113] have been reported using this technology, with power consumptions in the order of 0.5mW to 50mW per link [113, 114]. Usually, previous works demonstrate many links combined to form a high-bandwidth interface.

As ICLs are, on average, at least one order-of-magnitude larger in area than TSVs (which are typically around $15\mu\text{m}$ in diameter [115]), the achievable bandwidth per-unit-area using ICL-based 3D integration is significantly less than that which can be achieved using TSVs (a detailed iso-area analysis of TSV and ICL bandwidth is provided in Appendix A, concluding

⁴Due to the fact that this is typically the thickest metal, and hence the parasitic resistance of inductors fabricated in this layer is lower, improving the Q-factor.

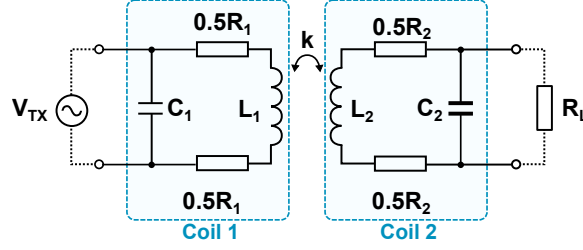


Figure 2.2: Simplified equivalent circuit model of an inductive coupling link (ICL) [51].

that TSV-based designs can achieve bandwidths up to $39\times$ higher than state-of-the-art ICL transceivers, for the same area). As discussed above, however, the use of ICLs has several economic advantages when compared to TSVs and can also offer the possibility of intrinsic wireless voltage level conversion⁵.

Figure 2.2 shows a simplified equivalent circuit diagram of an ICL where L_i is the inductance of each of each channel inductor i , R_i is the self-resistance of each channel inductor i , C_i is the self-capacitance of each channel inductor i , and k is the magnetic coupling coefficient. Analysing this circuit, the transfer function (V_{R_L}/V_{TX}) is given by Equation 2.3 below.

$$\eta_{ICL} = \frac{V_{R_L}}{V_{TX}} = \frac{1}{(1 + j\omega R_2 C_2)} \times j\omega k \sqrt{L_1 L_2} \times \frac{1}{R_L (1 - \omega^2 L_1 C_1) + R_1 + j\omega (C_1 R_1 R_L + L_1)} \quad (2.3)$$

Considering Equation 2.2, two important observations can be made about the operation of the inductive link. Firstly, the gain of the inductive link is proportional to the term $k\sqrt{L_1 L_2}$. This means that the link's efficiency (η_{ICL}) depends on the inductance of the channel inductors and the coupling coefficient k between them [51]. The coupling coefficient depends on a range of factors but is predominantly defined by the geometric parameters of the inductors used to form the link (such as size, shape, track width, track spacing, and number of turns) and the distance across which they communicate, X ⁶. In order to improve the efficiency of the link, therefore, it is necessary to carefully select these parameters, to maximise L and k whilst remaining within the given link power or area budget [52].

The second observation that can be made from the transfer function in Equation 2.3 is that the link can be modelled as a band-pass filter. The first term, related to the receiver, shows that the RX side behaves as a low-pass filter with a cut-off frequency of $1/\sqrt{2\pi C_2 R_2}$. The second term, related to the transmitter, shows that the TX side behaves like a second order high-pass filter with a self-resonant frequency of $1/2\pi\sqrt{L_1 C_1}$. The intrinsic band-pass

⁵The analysis presented in Appendix A found that in some cases (where a large difference in supply voltage exists between the layers of the stack), the ability to perform intrinsic level conversion using ICLs can also result in enhanced *energy* efficiency, when compared with TSVs.

⁶The magnetic permittivity of materials interposed within the link will also affect k .

nature of the ICL is another benefit of *inductive* near-field coupling (as opposed to *capacitive* coupling); the natural characteristics of the link mean that power and ground noise are filtered out. This does, however, also mean that the physical layout of each inductor in the channel must be carefully selected to ensure maximum effective gain across the channel, with respect to the operating frequency.

For typical ICL implementations in prior publications, the coils used for forming the channel achieve inductances in the order of $\sim 14\text{nH}$ with a coupling factor, k , of around 0.2 to be expected across a $60\text{ }\mu\text{m}$ communication distance [51]. Applying these numbers (along with typical published values for R_i , C_i , and R_L) to the transfer function in Equation 2.3 yields an efficiency of around 1%. For a transmit voltage, V_{DD} , of 3.0V , this corresponds to a V_{RX} of 30mV across the output load. Whilst this is larger than the typical RX voltage observed when using capacitive coupling (*c.f.* Section 2.1), it is still too low to interface directly with nominal-voltage RX circuitry without intermediate amplification. As with the capacitive coupling links, RX-side ICL amplifiers may be implemented synchronously (which offers favourable noise immunity, but requires the presence of an external synchronous clock in the RX die), or asynchronously (which has worse noise immunity, but does not require a synchronous clock source), and a more in-depth comparison of these two transceiver implementations is presented later in Section 2.2.2.

2.2.1 Data Encoding Schemes

As discussed above, the inductive link is *current driven*, such that, for a given transmit current I_{TX} , the voltage V_{RX} is observed at the receiver is given by Equation 2.4:

$$V_{\text{RX}} = k\sqrt{L_1 L_2} \cdot \frac{dI_{\text{TX}}}{dt} \quad (2.4)$$

In order to minimise the power consumption of the transceiver, whilst maximising dI_{TX}/dt (and hence the magnetic flux linkage within the die stack and V_{RX}), most ICL transceivers use *pulse-based* modulation schemes, where the flow of transmit current (I_{TX}) is limited to a short duration [5, 49, 51, 54, 112]. The following sections outline some of the most popular pulse-based encoding schemes (used to map data sequences to I_{TX} pulse patterns) and their relative advantages and disadvantages.

Bi-Phase Modulation (BPM)

One of the most intuitive solutions to the challenge of mapping a binary data stream to a series of current pulses is to use Bi-Phase Modulation (BPM). As the name suggests, BPM is a bi-phase encoding scheme where ‘1’s’ are mapped to pulses with positive polarity and ‘0’s’ are mapped to pulses with a negative polarity (or vice-versa) [43], as shown in Figure 2.3 below.

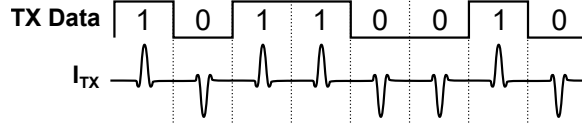


Figure 2.3: Illustration of inductive Bi-Phase Modulation (BPM) [43].

BPM is a simple and robust approach that can be implemented in an ICL using only a handful of logic gates to generate the appropriate pulse polarity. On the RX-side, BPM relies upon correct detection of the *phase* of each pulse, which is typically more straightforward than amplitude detection (as the amplitude is more susceptible to transient noise) [56]. BPM transceivers are typically used when high-bandwidth connectivity is required between tiers as the favourable noise immunity of BPM means that higher data-rates can be achieved [54] and works such as [54], [116] and [117] all use BPM in conjunction with other circuit techniques (pulse shaping, charge recycling and dual-coil transmission respectively) to leverage these high data-rates whilst reducing energy. When using BPM, if the energy of a single I_{TX} pulse is given by E_p , the energy-per-bit is also E_p , as one pulse is required per transmitted bit.

Single Phase Modulation (SPM)

One alternative to the use of BPM is Single-Phase Modulation (SPM), proposed in [56]. The concept of Single-Phase Modulation (SPM) illustrated in Figure 2.4 below. Here, a binary value ‘1’ is represented by the *presence* of an I_{TX} current pulse, whereas the binary value of ‘0’ is represented by the *absence* of an I_{TX} pulse.

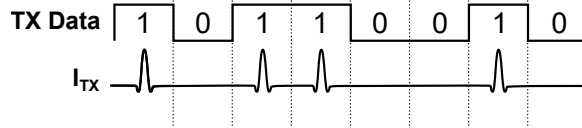


Figure 2.4: Illustration of inductive Single-Phase Modulation (SPM) [56].

The advantage of the SPM scheme is that, when considering the transmission of an equiprobable Pseudo-Random Binary sequence (PRBS), the energy-per-bit is reduced from E_p to $E_p/2$. This reduction, however, comes at the cost of more sensitive amplification requirements at the RX-side (or a greater I_{TX} pulse amplitude) to achieve the same BER, due to the fact that, in SPM, the phase margin is reduced from 180° to 90° [118]. Zhang *et al.* explore this trade-off in more detail in [118], and conclude that significant energy savings can still be realised by switching from BPM to SPM, even with the additional transmit pulse amplitude required for competitive noise immunity [118].

Non Return to Zero (NRZ) Modulation

One other alternative modulation scheme which overcomes this issue, whilst maintaining the power benefits of SPM, is inductive non-return to zero (NRZ) signalling, proposed in

[51] by Miura *et al.*. This approach is the most popular encoding scheme for inductive links within the context of 3D integration, adopted in almost all of the existing prior art. Figure 2.5 below illustrates its operation; Non-Return to Zero (NRZ) encoding operates in a similar way to BPM, however here, rising and falling data *edges* are mapped to pulses with positive and negative polarity (for example a positive current pulse is sent when the data stream transitions from ‘0’ to ‘1’ and a negative current pulse is sent when the data stream transitions from ‘1’ to ‘0’ [51]).

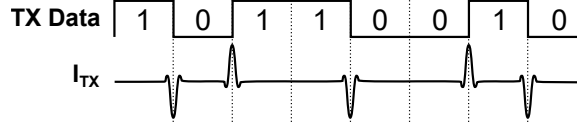


Figure 2.5: Illustration of inductive Non-Return-to-Zero (NRZ) encoding proposed in [51] by Miura *et al.*.

The reason that this approach is favoured when compared to those previously outlined, is that it achieves the same energy-per-bit as SPM (on average $E_p/2$), whilst also maintaining 180° of phase margin between symbols. Another advantage of the NRZ encoding scheme is that it can be implemented using extremely simple hardware. Inductive channels are typically driven using an H-Bridge driver [51]⁷ which maps naturally to NRZ encoding; if one H-Bridge branch is slightly delayed (by a time δ) compared with the other branch, this will result in a potential difference across the two branches for a time δ , at each data edge, with polarity corresponding to the edge direction. The NRZ scheme can also be decoded using simple circuits in the receiver, such as a Set-Reset (SR) latch where the set and reset inputs correspond the positive and negative pulse detection terminals of the RX amplifier [119].

Pulse Amplitude Modulation (PAM)

Whilst phase modulation is by far the most popular approach for use in ICLs (due to its favourable noise immunity), other works such as [120] have also explored Pulse Amplitude Modulation (PAM). In PAM schemes, it is the *amplitude* of the current pulse that denotes its value, rather than its phase or polarity. Whilst amplitude modulation schemes are common in conventional wireless communication systems, amplitude modulation is very difficult to achieve within the context of a 3D-IC due to the low energy and area budgets that are typically available; as discussed previously, the received voltage amplitude when using an inductive coupling channel is typically in the order of 30mV [51] and this voltage will vary slightly depending on the quality of the stacked die assembly (*e.g.* die-to-die stacking alignment, die thickness, adhesive thickness, etc.). This makes accurate RX-side pulse amplitude detection very difficult.

⁷A more detailed discussion of ICL transceiver hardware architectures is provided in Section 2.2.2.

Nevertheless, inductive transceivers using PAM have been demonstrated, typically using 4PAM [120]. Here two binary bits are mapped to one TX pulse which is signalled using one of four *combinations* of phase and amplitude (*e.g.* high-amplitude-positive-phase, low-amplitude-positive-phase, high-amplitude-negative-phase, low-amplitude-negative-phase). These mappings are calculated using Gray coding, to ensure that the effect of incorrect detection is minimised in terms of bit errors.

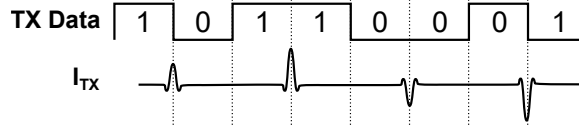


Figure 2.6: Illustration of inductive 4-Pulse-Amplitude Modulation (PAM) [120].

Figure 2.6 illustrates the 4PAM scheme used in [120]. Using this approach, the *number* of I_{TX} transmit pulses is reduced, however, assuming the minimum sense amplifier sensitivity is V_{SA} , the *amplitude* of some pulses has now increased from V_{SA} to $2V_{SA}$. For an N -ary PAM scheme, therefore, the energy per bit is given by $(N + \sqrt{N})E_{pb}/2N$ making it a competitive scheme in terms of energy, but the most susceptible to on-chip noise.

2.2.2 Transceiver Designs

Having summarised the main data encoding schemes used in prior literature, this section explores the transceiver hardware architectures used to implement them. Broadly, these transceivers can be classified in two main categories: (1) *coherent* (or synchronous) transceivers, which require the presence of a clock source that is phase-locked between TX and RX dies, and (2) *non-coherent* (or asynchronous) transceivers, where the data can be recovered without the need for a clock source. The specific implementations of these receivers and their relative advantages and disadvantages are discussed in the following paragraphs.

Coherent Transceiver Circuits

Coherent transceivers are favoured in systems where noise and crosstalk immunity is important [121]. In coherent systems, the received voltage is only sampled during a short window where the pulse is *expected* to be. For this reason, the opportunity, and hence probability, of the output being upset by transient noise is significantly reduced. Figure 2.7 illustrates a circuit-level implementation of a coherent ICL transceiver using NRZ encoding (outlined in Section 2.2.1). On the RX-side, a sense-amplifier (SA) based receiver is used (adopted in a number of prior works [5, 51, 54, 112]). On the TX-side, data is driven through the transmitting inductor using the delayed H-bridge topology discussed earlier. As the right-arm of the H-bridge is delayed when compared with the left-arm, a short current pulse will be allowed to flow clockwise through the inductor at each falling data edge, and counter-clockwise through the inductor at each rising data edge. The length of the current

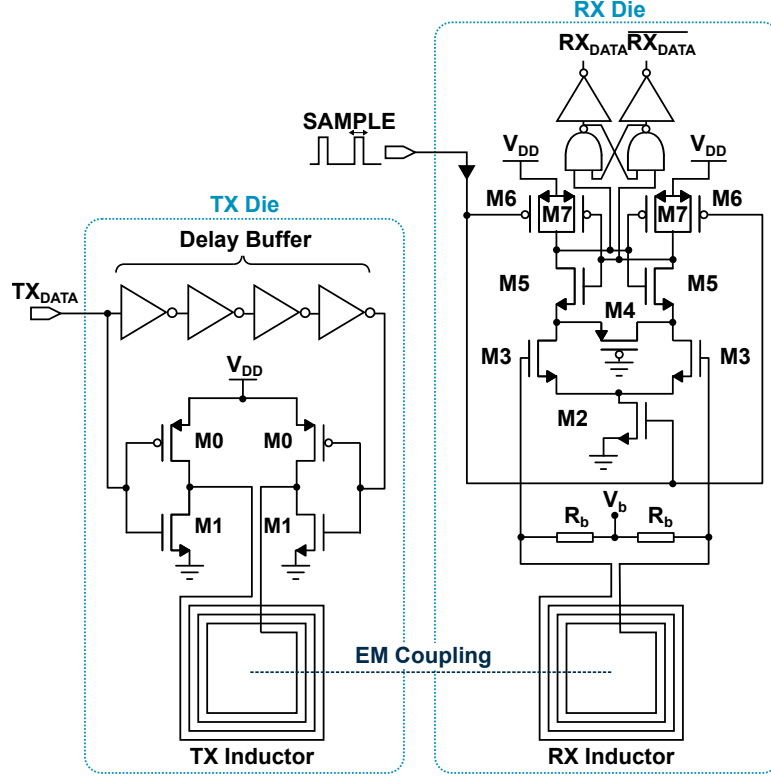


Figure 2.7: Illustration of coherent ICL transceiver using NRZ encoding [51].

pulse will be given by the propagation delay of the delay buffer⁸ and the amplitude will be controlled by the width of transistors M0 and M1. These transistors should be sized to ensure that dI_{TX}/dt is large enough to cause sufficient flux linkage, such that the pulse can be correctly detected in the receiver.

On the RX side, an SA is used which is en/disabled using the SAMPLE signal. When the SAMPLE signal is low, transistors M6 are switched on, placing the SA in the pre-charge phase [122]. In this phase, the nodes at the drain of M5 charge to VDD. When the RX pulse is expected, the SAMPLE signal goes high, this disables M6 but switches on M2, putting the SA in the evaluate phase [122]. In the evaluate phase, the transistors M3 amplify the received voltage signal (which is biased in the saturation region using Rb). Depending on the polarity of the received voltage pulse, a positive or negative potential difference will exist across M3. This will be amplified and will then cause an inverted-pulse at either the set or reset input of the SR latch at the SA output. The SA latch then effectively recovers the data such that the bit stream is present at RX_{DATA}.

Whilst this is a robust solution with high noise immunity, the disadvantage of using this coherent approach is that it requires a low-jitter synchronous clock (phase-synchronised

⁸In the example shown in Figure 2.7, this will be $4T_{PINV}$, where T_{PINV} is the propagation delay of a single inverter.

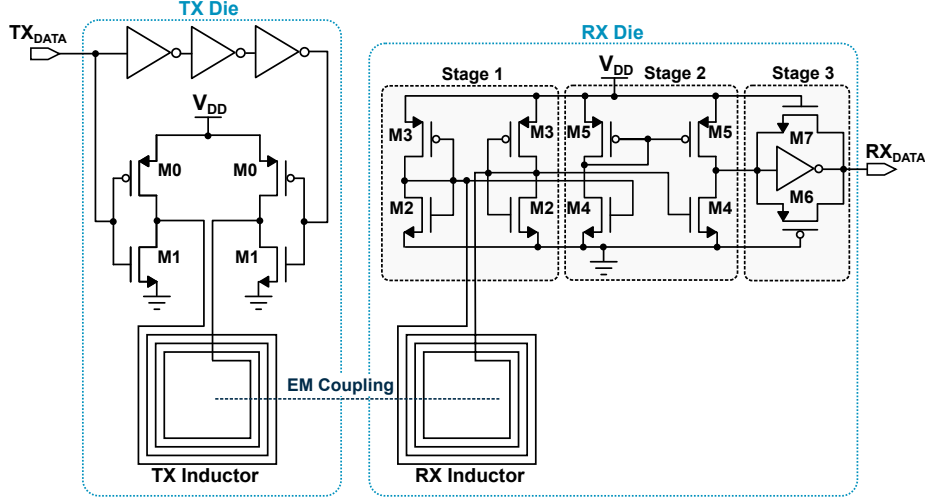


Figure 2.8: Illustration of non-coherent inductive transceiver implementing inductive Bi-Phase Modulation (BPM) signalling [113].

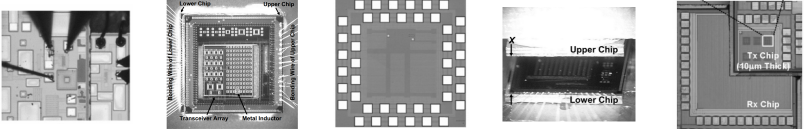
between TX and RX dies), to precisely generate the **SAMPLE** signal. A discussion of how this TX-RX clock synchronisation can be performed is presented later in Section 2.3.

Asynchronous/Non-Coherent Transceiver Circuits

ICL receivers can also be implemented *asynchronously* using non-coherent demodulator circuits such as that shown in Figure 2.8, presented in [113]. Here, instead of sampling the RX voltage within a particular time window, the attenuated RX signal is continuously amplified to reflect the transmitted data signal. Various implementations of RX-side amplifiers have been explored in prior publications, but the majority are similar to the architecture shown in Figure 2.8, consisting of 3 stages. Here, the first stage (formed from transistors M2 and M3) amplifies the RX pulses into a spiking voltage signal that corresponds to the data stream. The second stage (formed from transistors M4 and M5) is then sized to amplify this spiking signal until rail-to-rail saturation occurs, hence recovering the square data wave [111]. Finally, stage 3 usually consists of a latch to stabilise the data output.

In [113], Xu *et al.* implement an asynchronous ICL in 0.35 μm technology and demonstrate the ability to communicate at data-rates up to 2.8Gbps/channel, indicating that very competitive bit-rates can be achieved using this approach (especially in light of the utilised technology node). The energy per transmitted bit reported by Xu *et al.*, however is 17pJ/bit [113]; significantly larger than that which is typically reported for equivalent *synchronous* ICL designs. Lee *et al.* also present similar findings [111], this time implementing an asynchronous ICL in 50nm CMOS technology.

The increased power consumption of asynchronous transceiver designs is primarily due to the fact that the RX-side amplifier is *always* operational. This is in contrast with synchronous



	J. Xu <i>et al.</i> [113]	D. Mizoguchi <i>et al.</i> [123]	S. Kawa <i>et al.</i> [120]	N. Miura <i>et al.</i> [51]	N. Miura <i>et al.</i> [54]
Communication Distance	90um	300um	500um	60um	10um
Coherent/Non-Coherent	Non-Coherent	Coherent	Coherent	Coherent	Coherent
Modulation Scheme	NRZ	NRZ	PAM	NRZ	BPM
Process Technology	0.35um	0.35um	65nm	0.35um	90nm
Energy Per Bit	17.0pJ	54.6pJ	6.0pJ	36.8pJ	0.14pJ
Max. Bandwidth	2.80Gbps	1.20 Gbps	2.50 Gbps	1.25 Gbps	1.00 Gbps

Table 2.2: Table showing comparison of selected key prior works in the domain of contactless 3D integration using inductive coupling.

designs (where the RX-side amplifier is selectively enabled/disabled depending on the clock) which typically consumes less power, but requires additional supporting circuits for clock distribution.

Overall Summary

Table 2.2 provides a summary of some of the key previously reported works performing wireless 3D integration using ICLs (both synchronous and asynchronous across a variety of modulation schemes) [51, 54, 113, 120, 123]. As shown on the table, the majority of works target high-bandwidths with all of the links operating at frequencies $> 1.0\text{Gbps}$. This is driven by the *applications* that have motivated the use of ICLs in prior art, such as image sensors [124, 125] and stacked memory [43, 49]. Targeting high operating bit-rates, however, often comes at the expense of additional energy consumption (even when normalised in terms of energy-per-bit). Research into inductive links for low power IoT devices (that do not typically require gigabit operating frequencies) is very sparse and therefore, exploring low-energy ICL transceiver design for these applications will form one of the focusses of this thesis.

Another interesting point that can be noted from the results shown in the top row of Table 2.2, is that the communication distance, X ($X = \text{chip-thickness} + \text{glue-thickness}$) varies significantly between each of the previously reported works (between $500\mu\text{m}$ in [120], and $10\mu\text{m}$ in [54]). Most inductive links operate in the ‘square’ field region⁹, which means that a linear decrease in X results in a squared increase in efficiency [121]. It is, therefore, important to appreciate the results in terms of their reported X . To contextualise these numbers, the initial thickness of a raw 200mm silicon wafer is $725\mu\text{m}$ [126]. As standard

⁹A discussion of the effects of inductor diameter and communication distance is provided in the following section, 2.2.3.

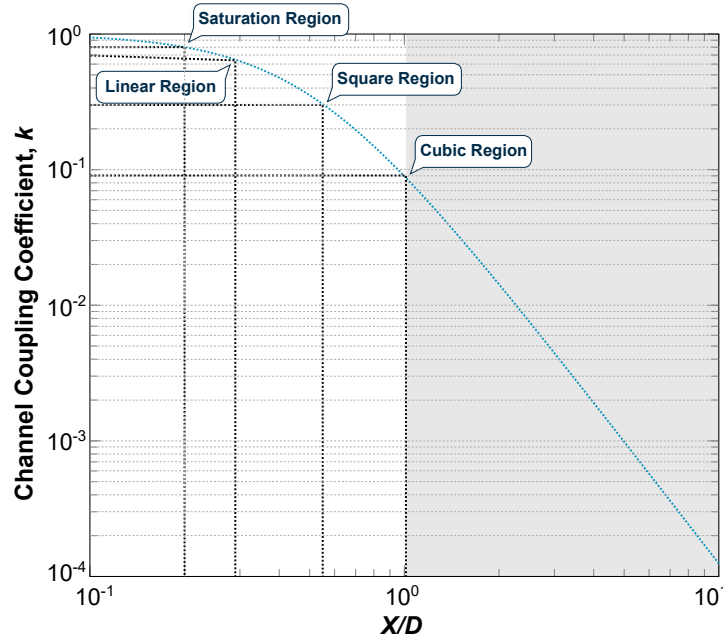


Figure 2.9: Illustration showing approximately how the coupling coefficient, k varies as a function of the communication distance X , normalised with respect to the channel inductor diameter D . (Reproduced from [128]).

procedure, wafer back lapping is then performed as a final fabrication stage in the foundry which usually takes the final thickness to between $150\mu\text{m}$ and $500\mu\text{m}$ [126]. Optional, low-cost, die or wafer level thinning can then also be performed on top of this which takes the dies to a final thicknesses between $50\mu\text{m}$ and $100\mu\text{m}$ [127]. Beyond this, additional thinning can often result in cracking, and wafer bending and so requires very expensive specialist polishing processes. Although possible to achieve, this means that thicknesses $<50\mu\text{m}$ presently demand very high costs, and hence undermine many of the economic benefits of using wireless 3D integration.

2.2.3 Inductor Layout for ICLs

Having provided an overview of different ICL modulation schemes and transceiver circuit implementations in the previous sub-sections, this sub-section discusses the design of the inductive channel itself (consisting of two coupled inductors). Whilst the transceiver design is important in determining the energy of a given link, it is the inductive channel itself that governs the link's maximum fundamental efficiency. As outlined in the previous sections, the link efficiency, η_{ICL} , is approximately proportional to the Electro-Magnetic (EM) coupling coefficient, k , between the TX and RX channel inductors.

Figure 2.9 (reproduced from [128]) shows how the channel coupling coefficient k varies as a function of X/D (the communication distance, X , normalised with respect to the inductor diameter, D). The graph, plotted on a log-log scale, shows four distinct regions [128, 129]:

the saturation region (where $D/X > 5$) which is typically used for ICLs performing wireless power transmission¹⁰, the linear region, the square region (where $X < D < 3X$) which is generally used for implementing ICL data transceivers and the cubic region (where $D < X$), in which the EM coupling is generally too low for use in ICL applications [128, 129].

The majority of ICL implementations operate within the ‘square region’ where the coupling coefficient, k is proportional to $(X/D)^2$. Typically, when designing an ICL transceiver however, X and D are fixed; the communication distance X is governed by the substrate thickness of the dies within the stack (and the height of the adhesive used for the die-attach process), and D depends on the maximum silicon area that is available for the channel inductor layout. It is therefore desirable to maximise the coupling coefficient k by optimising the *physical geometry* of the inductors used for forming the channel. This refers to, for example, the inductor’s shape (square, circle, hexagon, octagon *etc.*), the number of turns in the inductor and width and spacing of each turn.

To find geometries for the ICL channel inductors, existing works typically use a manual optimisation flow [52]. This involves: (1) defining an initial inductor layout with arbitrary parameters, (2) importing this layout into a full-wave field-solver for extraction of the system’s S-Parameters, (3) manually extracting (using curve fitting) a SPICE model of the link, and (4) analysing the overall link performance using SPICE. The process can be repeated, adjusting the layout parameters slightly on each iteration, until adequate layouts are found [52]. Full-wave simulation is typically performed using comprehensive FEM software packages such as CST Studio or Ansys HFSS. These solvers provide high accuracy, however, often take hours to converge at a solution. Further to this, the search for *best-performing* inductors requires analysis of thousands of layouts, making this flow extremely time consuming.

Other, more rapid solvers include application specific tools such as SPIRAL and ASITIC [130], developed for on-chip inductor analysis. These use electrostatic and magnetostatic approximations to provide much faster modelling, however, lack the ability to analyse mutual inductance between vertically stacked inductors, required for wireless 3D integration applications. As a result, physical optimisation of the ICL channel inductors (in terms of their geometry) poses a significant challenge and will form another of the focusses of this thesis.

2.3 Clock Delivery in Wireless 3D-ICs

As outlined in Section 2.2.2, the vast majority of near-field transceivers outlined in prior-art (both capacitive and inductive) use *synchronous* modulation schemes [5, 49, 51, 54, 99, 103, 106, 112] due to their superior noise immunity and energy efficiency. To implement these

¹⁰A survey of existing literature related to wireless power delivery in 3D-ICs is provided in Section 2.4.

coherent receivers however, precise clock synchronisation is required between the TX and RX dies. The typical timing margins measured in prior works can be as low as 10's of pico-seconds [50], and as such, establishing a low-jitter clock link is important for minimising bit errors.

In many prior works (both inductive and capacitive), clock delivery is performed externally using wire-bonding [19, 103, 105]. Here, the clock signal is generated in one tier, and distributed to each of the other dies within the 3D stack using wire-bonded links [123]. Whilst this is an adequate solution, the addition of wire-bonds to each tier undermines many of the cost-saving benefits associated with *wireless* 3D integration and, as discussed in Section 1.1.2 (Chapter 1), may not always be scalable to mass production. In addition to this, the parasitic overheads of the pad drivers, and *RLC* parasitics of each bond-wire, mean that the clock frequency (and minimum inter-tier clock *skew*) that can be achieved using this approach is limited. As a result, the tight timing margins discussed above are difficult to achieve and typically, 3D-ICs using this approach must also include large timing control circuits (such as Phase Locked Loops (PLLs)/Delay Locked Loops (DLLs)/Frequency Locked Loops (FLLs), or similar) on each die [123]. These circuits can have significant area footprints (in some cases larger than the wireless channels themselves [131]), and also consume a large amount of power.

2.3.1 Wireless Clock Delivery

To address these challenges, some other works have explored performing clock distribution wirelessly. Wireless clock distribution is reported in several prior publications ([60, 61, 123, 132, 133]) and offers the ability to achieve much higher clock rates than the wire-bonded approach (which is typically limited to around 400MHz due to the *RLC* bond-wire parasitics and pad driver, as discussed above). To provide a stable clock source, particularly when operating at high frequencies, most works performing wireless clock distribution still include locked-loop recovery circuits, but the ability to transmit the clock in the same way as the data means that the clock and data are very tightly coupled.

To minimise RX jitter when transmitting the clock, the most popular way of achieving wireless clock distribution in prior art is using coupled resonators [63, 133]. In these works, *LC* tanks, with a resonant frequency corresponding to the clock frequency, are formed between the layers of the 3D stack, most commonly in ICLs, where the inductive (*L*) component is part of a coupled inductive link [63]. This is a promising solution that allows the clock to be wirelessly transmitted between dies with very low jitter and skew (due to the natural harmonics of the link).

However, such clock links based on resonant *LC* tanks do suffer from several drawbacks. Firstly, the links must either be very high-frequency or very large in terms of area (as the resonant *LC* frequency decreases with diameter, *D*, of the inductor, *L* [134]). To achieve

clock distribution at frequencies in the order of 100's of Mega Hertz using this approach (which is typical for IoT applications) often requires inductors with diameter $> 300\text{ }\mu\text{m}$ [64]. The second challenge with such links is their susceptibility to variation. As coupled-resonator clock links rely on the voltage amplitude boost at a specific frequency (typically within a margin $< \pm 5\%$ [63]), any slight variations in either the generated clock frequency, or the link's inductance (including mutual inductance contribution) can result in significant performance deviations or, in the worst case, the whole system ceasing to operate. These shifts can arise from Process Voltage Temperature (PVT) variations or packaging variations (*e.g.* different die-to-die stacking alignments or adhesive thicknesses between chips) but can also occur naturally if using a low precision clock source. To overcome these challenges, such works usually require large, high-precision (and high power) clock generators and timing control circuits [63], which are not typically included in IoT devices. Further to this, the third drawback is that such links only operate at a single frequency [133]. It is often desirable to reuse the same Intellectual Property (IP) block across multiple designs. Using a *resonant* clock link limits the ability to do this, as porting to other frequencies requires fundamental re-design of the transceiver and layout. Motivated by these challenges, this thesis will also investigate clock distribution in wirelessly stacked inductive 3D-ICs, focussing on area constrained IoT devices.

2.4 Power Delivery in Wireless 3D-ICs

Aside from clock and data transmission, the final element that must be considered to establish fully-wireless 3D assembly is power delivery. The majority works outlined in the sections above opt to provide power delivery using wire-bonded VDD and GND connections [5, 49, 51, 54, 99, 112]. Whilst this is an adequate solution, that circumvents the use of TSVs, the addition of wire-bonds to each die in the stack undermines many of the benefits associated with *wireless* 3D integration.

To address this, some research has suggested the use of Highly-Doped Silicon Vias (HDSVs) [135] to deliver power *contactlessly* through the 3D stack. HDSVs are vertical power delivery channels formed from highly doped wells to conduct charge between dies after aggressive thinning (the use of doping here avoids the introduction of metal, required when using TSVs). It has been suggested that this approach will not require bump or wire bonding, and that a resistance of less than $3\text{ m}\Omega$ could be achieved within the channel [135]. Whilst this is a promising future technology, HDSVs are yet to be practically realised and will require a substrate thickness less than $5\text{ }\mu\text{m}$ [135]. As discussed earlier, thinning this aggressive introduces a plethora of physical challenges (considering that the standard thickness of an eight-inch wafer is $725\text{ }\mu\text{m}$ [126]) in addition to significantly increasing manufacturing costs.

2.4.1 Wireless Power Transfer

Another, more established method of delivering power between tiers of a 3D-IC, without using wire-bonding, is through Wireless Power Transfer (WPT) [59, 61, 136]. This section, therefore, provides a summary of the state-of-the-art works in this area. As with wireless data links (which have been discussed above), the works exploring die-to-die Wireless Power Transfer (WPT) can also be broadly classified as *inductive* (relying on a time-varying magnetic field [61]) or *capacitive* (relying on a time-varying electric field [137]) approaches. A discussion of the prior art in each of these categories, and their relative advantages and disadvantages is provided below.

Inductively-Coupled

In all prior art to-date, wireless power delivery is treated separately from wireless data transmission, being divided either *spatially* (for example, when the WPT and data ICLs are placed in different physical locations on the chip [136]) or *temporally* (for example, where communication of data and transmission of power occur at different times [61]). One of the first published works proposing the use of wireless power delivery in stacked 3D-ICs is that by Onizuka *et al.* [136] which presents a WPT link for spatially separated wireless stacking applications. Here, wireless power transfer of up to 2.5mW per link is achieved across a 20 μm distance using 700 $\mu\text{m} \times 700 \mu\text{m}$ octagonal inductors. This work (implemented in 0.35 μm CMOS technology) uses a full-wave active rectifier operating at a switching frequency of 330MHz.

Other similar works include [138] and [59]. These achieve WPT densities (power/area) of between 5.1mW/mm² and 49.2mW/mm² using similar schemes in conjunction with full-wave and half-wave rectifiers (respectively). These works use carrier frequencies in the order of 100's of Mega-Hertz [59, 138] and Yuan *et al.* [59] propose the use of an innovative time-interleaved transmission scheme, to reduce the ripple in the RX voltage signal across multiple staggered transmitter-receiver pairs.

Work by Han *et al.* improves upon the power efficiencies achieved in these previous works by leveraging *resonant* inductive coupling where the link is excited at its own self resonant frequency [139]. Using this approach, the RX voltage is significantly boosted (through the same mechanisms discussed in Section 2.3 earlier) resulting in a much enhanced WPT efficiency. Han *et al.* report WPT densities of up to 610mW/mm² using the proposed resonant scheme across a total communication distance of 50 μm . This represents an order of magnitude improvement compared to other prior art, clearly highlighting the benefits of resonant operation for WPT.

One final work that uses spatially separated wireless power and data transmission is [60] by Rackecki *et al.*. To reduce the area overhead of placing power and data links side-by-side

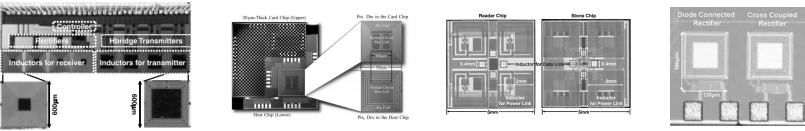
on the chip, Rackecki *et al.* propose the use of *nested* power and data coils. Here a larger, 750 μm inductor is used to deliver power between the two stacked dies, and a smaller 300 μm clover coil arrangement (placed within the 750 μm power coil) is used for data transmission [60]. This does mean that two separate sites are not required for data and power delivery, however the total bounding area of the channel used for this chip is still very large (0.49mm^2). Using these nested coils, Rackecki *et al.* report a power delivery density of $20.8\text{mW}/\text{mm}^2$ across a distance of 20 μm (in a 65nm CMOS technology).

In contrast to these works which use *spatially* separated wireless power and data transmission, in [61] Yuxiang *et al.* propose the use of *temporally* separated power and data transmission. Although separate physical inductors are used to transmit data and power in this work, to avoid interference, power and data transmission never occur simultaneously. First, power is transferred to the receiver using the WPT link (which is 2mm in diameter) and used to charge an on-chip capacitor. Once enough power has been transferred and accumulated in the capacitor, the power transmission is stopped. At this point bi-directional data transmission is enabled across a separate set of data links (which are 400 μm in diameter). This bi-directional communication continues until the energy has been depleted, at which point the communication ends and the power delivery is re-enabled to re-charge the receiver [61].

Whilst this is a good method for operating robustly (as temporal separation means that there is no interference between the power delivery phase and the data delivery phase), it has several drawbacks. Firstly, the fact that data transmission must be periodically interrupted to re-charge the energy storage in the receiver means that overall data-rates are relatively low (in the order of 150Mbps [61]). Secondly, the recovered energy must be stored for use at a later point in time. This requires the addition of on-chip capacitors such as MIM capacitors (or even supercapacitors) which consume significant silicon area [140] and also inherently reduce the efficiency of the system (due to self-discharge caused by leakage currents [141]). Finally, the proposed approach still requires two separate inductor geometries (for the contrasting requirements of *power* and *data* delivery) which, in total, consume over 4mm^2 in the 180nm CMOS technology adopted in this paper [61].

Capacitively-Coupled

Although the majority of works exploring wireless power transfer operate using *inductive* coupling (due to the fact that it is current driven), Culurciello *et al.* [137, 142] also explore the feasibility of performing wireless power transfer using *capacitive* coupling. Here, power is transferred through a time-varying electric field, which is provided by driving a metal plate using an all-digital ring-oscillator circuit in the TX die. The advantage of using capacitively-coupled WPT, is that the silicon area overhead is much reduced; Culurciello *et al.* use $90\mu\text{m} \times 90\mu\text{m}$ capacitive pads (which represents an order of magnitude area



	Yuan <i>et al.</i> [59] (‘08)	Radecki <i>et al.</i> [60] (‘12)	Yuxiang <i>et al.</i> [61] (‘09)	Han <i>et al.</i> [139] (‘12)
Communication Distance	10um	20um	200um	50um
Technology	180nm	65nm	180nm	65nm
Area (per link)	0.36mm ²	0.49mm ²	4mm ²	0.014mm ²
Power Delivery (per link)	3mW	42mW	13.5mW	15mW
Power Delivery Density	8.3mW/mm ²	20.8mW/mm ²	3.37mW/mm ²	607mW/mm ²

Table 2.3: Comparison of key prior works in the domain of WPT within 3D-ICs.

reduction when compared to the works using inductive power transfer discussed in Section 2.4.1).

Using capacitively-coupled wireless power transfer does, however, have several challenges. As the drain-source voltage of the transistors used on the driver side is limited (and capacitive coupling is *voltage-driven*), the amplitude of the received voltage signal is very low [142]. Culurciello *et al.* address this by using a five stage Dixon charge pump to ‘pump up’ the voltage on the receiver side. This is an adequate solution, which allows around 1.65mW of power to be recovered [137], however the maximum WPT efficiencies that can be achieved using such circuits are limited. The second drawback is that capacitive WPT approaches can only be used for Face-to-Face (F2F) stacking applications, with the maximum WPT distance reported in [137] being limited to less than 1 μm .

Overall Summary

Table 2.3 provides an overall summary of the key works explored in this section discussing wireless power transfer. For each work, the power delivery density (power delivered per unit area) has been added to the table as a metric for comparison. As can be observed, useful levels of power delivery can be achieved via WPT, however this comes at a significant silicon area cost (with all works using inductors larger than $500\mu\text{m} \times 500\mu\text{m}$ [59–61]). Additionally, the WPT links proposed in these previous works are separate to those used to transmit data. This is due to the conflicting layout constraints of the channel inductors; efficient power delivery requires inductors with very strong EM coupling (at the expense of high parasitic capacitance), whereas efficient data delivery requires low-parasitic layouts to maximise bandwidth. This means that the *overall* footprint of the wireless interface (including power and data transmission) can often reach over 1mm^2 [61]. For devices that are volume/area constrained, this level of silicon overhead is clearly undesirable. Further to this, the power efficiencies that can be achieved by such WPT links are often very low (in the order of 10% or less [60]). For IoT applications that typically operate on stringent power budgets, this also poses a significant challenge. Motivated by these two challenges,

the final technical contribution of this thesis will therefore explore wireless power delivery in contactless 3D-ICs with a focus on area efficiency.

2.5 Summary

Previous work on wireless 3D integration has explored both capacitive and inductive links to perform chip-to-chip communication within stacked 3D-ICs. Generally speaking, CCLs offer better energy and area efficiency, but have a limited communication distance because they are *voltage*-driven (and hence usually limited by the supply voltage of the chip). Current-driven ICLs address this issue, facilitating much larger communication distances (and hence also allowing F2B stacking). This chapter has reviewed a range of ICL modulation schemes and transceiver designs, concluding that coherent modulators can offer better noise immunity and lower-energy operation overall. The literature survey has also highlighted that the majority of works to-date focus on high-bandwidth applications (such as stacked memory and image sensors), often sacrificing on energy efficiency (which is of paramount importance when considering the context of IoT devices).

To implement coherent demodulation, wireless clock synchronisation is also required. Previous works have typically focussed on using coupled resonators for forwarding the clock between dies. Whilst this has several benefits (including lower jitter and inter-die skew), such links require large areas and are very sensitive to variations in the generated clock frequency and channel quality (for example in terms of die-to-die alignment and communication distance). Coupled-resonators are also generally not portable between different designs as they are strongly layout dependant and only operate at a fixed frequency.

The literature review has also highlighted that most prior works focus only on wireless *data* communication. To remove the need for all post-fabrication processing (in the form of wire-bonding or flip-chip bonding), and hence realise truly low-cost 3D assembly, it also is necessary to establish a method of delivering *power* wirelessly within the 3D stack. A range of prior works exploring wireless power delivery have been discussed, however the main drawback with these existing works is their large area overhead. Prior implementations all use inductors larger than $500\text{ }\mu\text{m} \times 500\text{ }\mu\text{m}$ in size which are separate to the inductors used for data transmission. This means that the overall footprint of the wireless power/data interface can often reach over 1mm^2 .

The following technical chapters of this thesis will seek to address some of these research challenges, with a focus on answering the six research questions outlined in Chapter 1.

Chapter 3

Low-Energy Transceiver Design using Spike-Latency Encoding

As discussed in the literature review from Chapter 2, 3D-ICs constructed using ICLs have been demonstrated in several publications [50, 54, 123], however one of the reported drawbacks of using ICLs is their inferior energy efficiency [54], which is of paramount importance when considering IoT sensor devices that typically have total power budgets $<50\text{mW}$, owing to their battery powered nature (as discussed in Section 1.4, Chapter 1). This chapter focusses on improving the encoding approaches used by existing Inductive Coupling Link (ICL) designs to address this challenge. The chapter initially provides analysis of the encoding approaches used in prior art (such as Bi-Phase Modulation (BPM), Single-Phase Modulation (SPM), or Non-Return to Zero (NRZ) encoding) demonstrating that, when using these schemes, by far the largest energy contribution is from the transmit pulse energy.

To reduce the transmit pulse energy, and hence the energy-overhead of the ICL as a whole, this chapter then proposes a novel ICL transceiver that uses a *time-domain* encoding approach. The approach uses the latency *between* pulses to encode frames of data, thereby reducing the number of Transmitter (TX) current pulses and overall power consumption. This encoding scheme is also combined with a tuneable current driver to minimise the transmit current for a given integration scenario. The main novel contributions of this chapter can therefore be summarised as:

- Proposal of a low-energy inductive transceiver that applies a time-domain encoding approach (spike-latency encoding) in the context of intra-chip communication for transmitting data between tiers of a 3D-IC. The approach uses the latency between sequential pulses to represent data, hence reducing the required transmit energy.
- Mathematical modelling of the proposed transceiver design for evaluating best-performing algorithm parameters across a range of 3D integration scenarios.
- Presentation of a tuneable current driver circuit, to precisely control the TX energy (within 0.25pJ) depending on the channel quality (and hence compensate for up to

40 μm of die-to-die stacking misalignment in both x and y directions by post-assembly tuning).

- Experimental validation of the proposed transceiver using post-layout SPICE simulations in 0.35 μm , 65nm, and 28nm technologies, demonstrating an energy consumption as low as 0.26pJ/bit across a 110 μm channel (a 28% improvement compared with the state-of-the-art)¹.
- Silicon validation of the proposed transceiver on a 2-tier 3D stacked test-chip in 0.35 μm CMOS technology, demonstrating a 13% reduction in energy-per-bit when compared with state-of-the art transceivers.

The remainder of the chapter is organised as follows: Section 3.1 presents a summary of background work related to ICLs and their modulation schemes, Section 3.2 outlines the spike-latency encoding scheme proposed in this chapter (including mathematical modelling in Section 3.2.1). Section 3.3 outlines the hardware implementation of the transceiver, including the tuneable pulse driver circuit, before experimental validation is performed using SPICE in Section 3.4, and in a silicon test-chip in Section 3.5. Finally, a discussion and the chapter’s conclusion are presented in Sections 3.6 and 3.7 respectively.

3.1 Background and Related Work

As discussed in Chapter 2, Section 2.2.1, the most energy efficient way of encoding data in inductive coupling links (ICLs) is using current pulse encoding, where the data stream is mapped to a series of TX current pulses; using short current pulses maximises dI_{TX}/dt , and hence the magnetic flux linkage within the die stack, resulting in better efficiency. Prior works have explored a range of pulse encoding schemes, including Bi-Phase Modulation (BPM) [143], Single Phase Modulation (SPM) [144] and Non-Return-to-Zero (NRZ) encoding [54] (the most popular approach in prior art).

Figures 3.1 (a) - (c) provide a recap of how these different schemes operate (a full explanation, highlighting their relative advantages and disadvantages was presented in Section 2.2.1 of the previous chapter). As can be seen from the figure, BPM (Figure 3.1(a)) has the highest intrinsic energy consumption, using one TX pulse per transmitted bit. BPM maps ‘1’s to positive I_{TX} pulses, and ‘0’s to negative I_{TX} (or vice-versa), which offers favourable noise immunity, but high energy consumption [45, 143, 145].

SPM offers lower energy consumption, on average using only one TX pulse for every two transmitted bits. In SPM, ‘1’s are represented by the presence of an I_{TX} pulse, and ‘0’s are represented by the absence of an I_{TX} pulse. Whilst this does reduce the intrinsic energy

¹Simulation results assuming a 28nm CMOS technology. Simulated energy in 65nm CMOS is 0.66pJ/bit, and silicon measurements in 0.35 μm CMOS show an energy of 7.6pJ/bit.

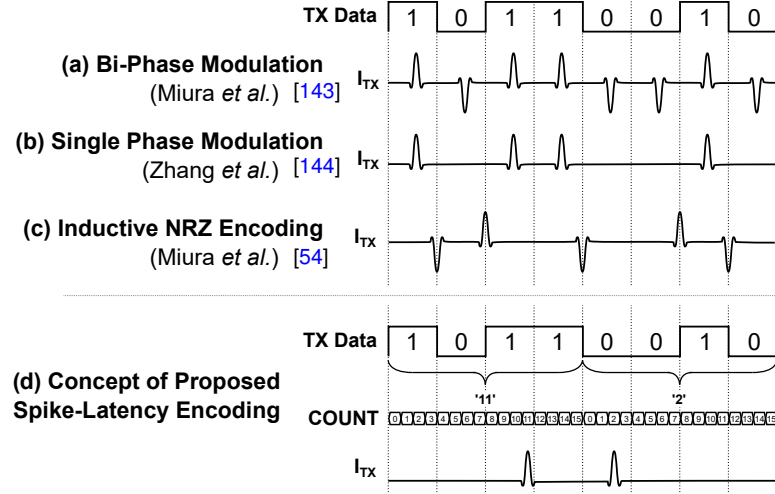


Figure 3.1: Illustration of: (a) Bi-Phase Modulation (BPM) [143], (b) Single Phase Modulation (SPM) [56, 144] (c) Inductive Non-Return-to-Zero (NRZ) line code (proposed by Miura *et al.*, this is the state-of-the art approach for inductively coupled communication within a 3D-SiP/3D-IC context) [50, 51, 54, 123]), and (d) The spike-latency encoding scheme proposed in this chapter.

consumption of the scheme, SPM suffers from reduced noise immunity because there is no *phase* difference between symbols.

To overcome this issue, whilst maintaining the power benefits of SPM, most works exploring wireless 3D integration use the inductive non-return to zero (NRZ) signalling scheme, proposed in [51] by Miura *et al.*. This approach is illustrated by the waveforms in Figure 3.1 (c). Here, each rising/falling data *edge* is encoded as a current pulse with corresponding positive/negative polarity. This is a robust solution that allows data to be simply encoded using a delay buffer, and decoded using just a Sense Amplifier (SA) and Set-Reset (SR) latch [50, 51, 54, 123, 146]. Assuming that the TX data is an equiprobable Pseudo-Random Binary sequence (PRBS), the NRZ scheme still uses, on average, only one I_{TX} pulse per two transmitted bits, however the 180-degree phase difference (between symbols) is maintained. This makes it particularly favourable for robust low-energy communication in previously published works.

When considering a typical ICL architecture, three main sources of power consumption exist:

1. Power consumption derived from the analogue transmit current (I_{TX}) through the driver circuits and TX inductor to form the magnetic field.
2. Power resulting from the analogue receiver current draw (I_{RX}), consumed by the Receiver (RX) amplifier detecting the induced RX voltage.
3. The dynamic power consumption of the supporting digital logic including the data encoding/decoding circuits (derived from the current consumed by the supporting

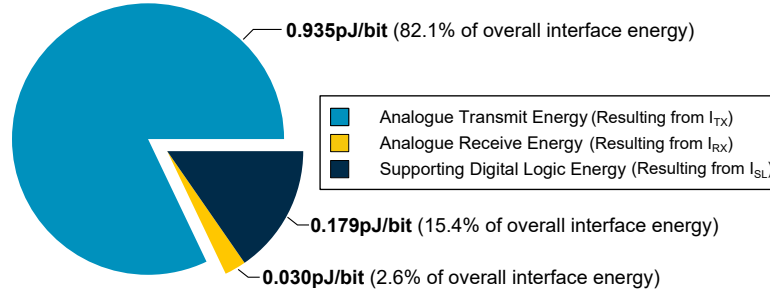


Figure 3.2: Energy breakdown of Non-Return-to-Zero (NRZ) inductive transceiver [51] when implemented in 65nm technology, communicating over a distance of 100 μm .

logic block (I_{SL}) at each data edge).

The pie chart in Figure 3.2 illustrates the breakdown of these three components in terms of energy, for state-of-the-art transceivers using NRZ encoding [54, 123]. It can be observed that, by far, the largest contribution is the I_{TX} analogue transmit current (over 80%) as this scheme requires, on average, one pulse for every two transmitted bits. In typical ICLs, each I_{TX} pulse is very expensive in terms of energy consumption, especially when communicating over large distances (such as multiple stacked dies) meaning that these transceivers are still power hungry when compared to conventional 3D integration approaches such as TSVs [19]. To address this challenge, this chapter proposes the use of an alternative data encoding scheme to reduce the high I_{TX} energy consumption in favour of additional digital processing. The design of this transceiver is outlined in Section 3.2 below.

3.2 Proposed Spike-Latency Encoding Modulation Scheme

To address the high I_{TX} power consumption of existing ICL transceivers, this chapter proposes the use of *spike-latency* encoding to encode data frames in the time domain. Under the proposed scheme, values are not represented directly by current pulse patterns, but by the latency *between* the start of the frame, and the transmit current pulse (also known as Pulse Position Modulation (PPM)). Figure 3.1 (d) illustrates this concept. Here, N bits (in this case $N=4$) are translated into a decimal value which is represented by a single I_{TX} pulse. The value of these encoded bits is encoded in terms of *the latency with which the pulse is transmitted*. In this example, ‘b1011’ is denoted by transceiving a pulse when the RX/TX counter (COUNT) is at time value 11, and ‘b0010’ is denoted by transceiving a pulse when the RX/TX counter (COUNT) is at time value 2. This scheme is only possible provided that precise counter synchronisation is available between the transmitter and receiver making it well suited to 3D-IC/3D SiP applications where the channel is fixed and communication is over a short distance. As with the other encoding schemes discussed above, the data-bit to TX pulse ratio can also be increased by encoding one bit in terms of phase.

When using the proposed Spike-latency Encoding Transceiver (SET), as N increases, the

Table 3.1: Nomenclature for the equations outlined in this section.

Parameter	Description
N	Algorithmic parameter representing the number of binary bits mapped to a single TX pulse
I_{TX}	Transmitter current
I_{RX}	Receiver current
I_{SL}	Digital supporting logic current
$E_{pbSET}, E_{pbBPM}, E_{pbSPM}, E_{pbNRZ}$	The energy per transmitted bit when using the proposed Spike Encoding Transceiver (SET) Modulation / Bi-Phase Modulation (BPM) / Single Phase Modulation (SPM) / Non Return to Zero (NRZ) Modulation
V_{DD}	Supply voltage
f_{DAT}	TX/RX data frequency
f_{COUNT}	Counter frequency (when using the proposed SET scheme)
L_{TX} or L_{RX}	Inductance of TX or RX inductor
k	Electromagnetic coupling coefficient between the TX and RX inductors
I_p	TX Pulse current
δ	Pulse duration
$I_{DFF}, I_{XOR}, I_{AND}$	Current consumed by a single DFF / XOR / AND cell

number of pulses-per-transmitted-bit decreases linearly, allowing for significant I_{TX} energy savings. However, as N increases, the COUNT frequency (and hence supporting digital logic energy required to maintain the existing data rate) increases proportionally to 2^N when using single phase encoding, or 2^{N-1} when using bi-phase encoding (as considered in this chapter). To find the most energy efficient implementation of the proposed modulation scheme therefore, the parameter N must be carefully selected to best-exploit the trade-off between I_{TX} and I_{SL} by considering the transceiver design *as a whole*. Section 3.2.1 provides mathematical modelling to explore this trade-off in more detail.

3.2.1 Mathematical Modelling

As discussed above, when using the proposed spike-latency encoding scheme, it is important to select an appropriate value for the parameter N which trades off the linear reduction in the number of transmit pulses against the additional digital processing energy. This section provides more in-depth modelling of this trade-off.

Similarly to the NRZ scheme, discussed in Section 3.1, the power consumption of the proposed transceiver can be categorised in three main sources: (1) The analogue transmit current, I_{TX} , (2) The analogue receive current, I_{RX} , (consumed by the RX amplifier detecting the induced RX voltage) and (3) the current consumed by the supporting digital logic, I_{SL} , (including the data encoding/decoding circuits). For the proposed Spike-latency Encoding

Transceiver (SET) scheme, the energy-per-bit ($E_{pb_{\text{SET}}}$) is therefore given by:

$$E_{pb_{\text{SET}}} = \frac{V_{DD}}{N} \int_0^{\frac{1}{f_{\text{COUNT}}}} I_{\text{TX}}(t) dt + \frac{2^{N-1} V_{DD}}{N} \int_0^{\frac{1}{f_{\text{COUNT}}}} I_{\text{RX}}(t) + I_{\text{SL}}(t, N) dt \quad (3.1)$$

where V_{DD} is the supply voltage and f_{COUNT} is the link counter frequency (which will be equivalent to $f_{\text{DAT}}/2^{N-1}$, where f_{DAT} is the data frequency). Here, the first term represents the transmit pulse current, which will decrease by $1/N$ as N increases (as more bits are encoded using a single pulse). The second term represents the current consumed by the sense amplifier; as N increases, the number of sense operations increases by 2^{N-1} and hence this term² is proportional to 2^{N-1} . The final component represents the supporting logic. The number of clock *edges* in the supporting logic to maintain a given data-rate will also increase proportionally to 2^{N-1} ; however, the number of *gates* also depends on N , so I_{SL} is also a function of N (see below).

To approximate $E_{pb_{\text{SET}}}$ using Equation 3.1, these three elements (I_{TX} , I_{RX} , and I_{SL}) can be approximated as follows. The transmit pulse current (I_{TX}) can be modelled mathematically by a Gaussian pulse [51]:

$$I_{\text{TX}}(t) = I_p \cdot \exp \left[- \left(\frac{t\pi}{\delta} \right)^2 \right] \quad (3.2)$$

Where I_p is the peak amplitude of the current pulse required to ensure error-free detection in the receiver, and δ is the minimum TX pulse width, a technology dependent parameter. Given a wireless channel, with coupling coefficient k , using inductors with inductance L_{TX} and L_{RX} , the voltage pulse amplitude induced in the RX coil is given by:

$$V_{\text{RX}} = k \sqrt{L_{\text{TX}} L_{\text{RX}}} \cdot \frac{dI_{\text{TX}}}{dt} \quad (3.3)$$

For transmission to be robust, V_{RX} must be greater than the minimum receiver sensitivity threshold V_{St} , a technology-dependent parameter indicating the minimum RX voltage fluctuation that can be accurately distinguished by the SA. I_p can therefore be obtained using Eqn. 3.4 below:

$$V_{\text{St}} + V_{\text{noise}} < \max \left\{ \frac{2\pi^2 I_p t}{\delta^2} \cdot \exp \left[- \left(\frac{t\pi}{\delta} \right)^2 \right] \right\}_0^t \quad (3.4)$$

Where $0 > t > 1/f_{\text{COUNT}}$, $t \in \mathbb{R}^+$, and V_{noise} is the maximum amplitude of transient noise in the SA supply (*e.g.* any substrate noise, supply droop etc.). Once I_p has been obtained, Eqn. 3.2 can be used to find I_{TX} .

The receiver current (I_{RX}) consumed in the sense amplifier can be modelled statically,

²Here the ‘-1’ term corresponds to the additional bit that can be encoded using phase.

because the average current required for a single sense operation will remain constant. However, the number of supporting logic gates in the data encoder/decoder will depend on N , as discussed above. Approximately, $I_{SL}(N)$ can be modelled by:

$$I_{SL}(N) \approx 2NI_{DFF} + NI_{XOR} + (N + 2)I_{AND} \quad (3.5)$$

where I_{DFF} , I_{XOR} , and I_{AND} represent the dynamic current consumption of a flip-flop, XOR and AND gate respectively³(justification for this is provided later, in Section 3.3.1).

Analysing the equations presented above, the advantages of the proposed Spike-Latency scheme (in terms of reducing the I_{TX} current) can be observed. To transmit i bits using BPM requires i , pulses. To transmit i bits using SPM or NRZ requires, on average, $i/2$ pulses, but to transmit i bits using the proposed SET scheme requires only i/N pulses. Increasing N , however, comes at the cost of increasing I_{RX} and I_{SL} and so N must be carefully selected. Section 3.4.2 evaluates this trade-off mathematically using databook logic gates parameters for 0.35 μm , 65nm, and 28nm technologies, in conjunction with the equations above, to find the optimal value of N for a given channel quality.

3.3 Architecture Design and Hardware Implementation

Following the theoretical modelling of the proposed spike-latency encoding scheme, this section explores how it can be implemented in hardware. Figure 3.3 shows the architecture of the low-energy inter-tier link proposed in this chapter, consisting of five key components: (1) the spike-latency encoding logic, to implement the modulation scheme discussed in the previous section (outlined in Section 3.3.1), (2) a tuneable current driver, to adaptively control the transmit current such that it is absolutely minimised depending on the integration scenario, (3) the inductive channel itself, consisting of two coupled planar inductors, (4) a sense amplifier, to amplify the received voltage signal, and (5) the demodulation logic to recover the transmitted data stream. The following sub-sections outline the design of each of these five components in more detail, before Section 3.3.5 discusses the TX/RX clock synchronisation infrastructure adopted in this chapter.

3.3.1 Encoding/Decoding Logic

The first, and most important element of the proposed transceiver design is the encoding/decoding logic. Figure 3.3(a) illustrates a practical implementation of the en/decoding logic consisting of an $N - 1$ bit counter (that generates the **COUNT** signal) and XOR-based match logic which compares the parallel TX data bits with the incrementing **COUNT** signal. This generated signal is then fed through a final multiplexer stage, controlled by the MSB of

³For this basic mathematical modelling, static power consumption is considered negligible and hence ignored.

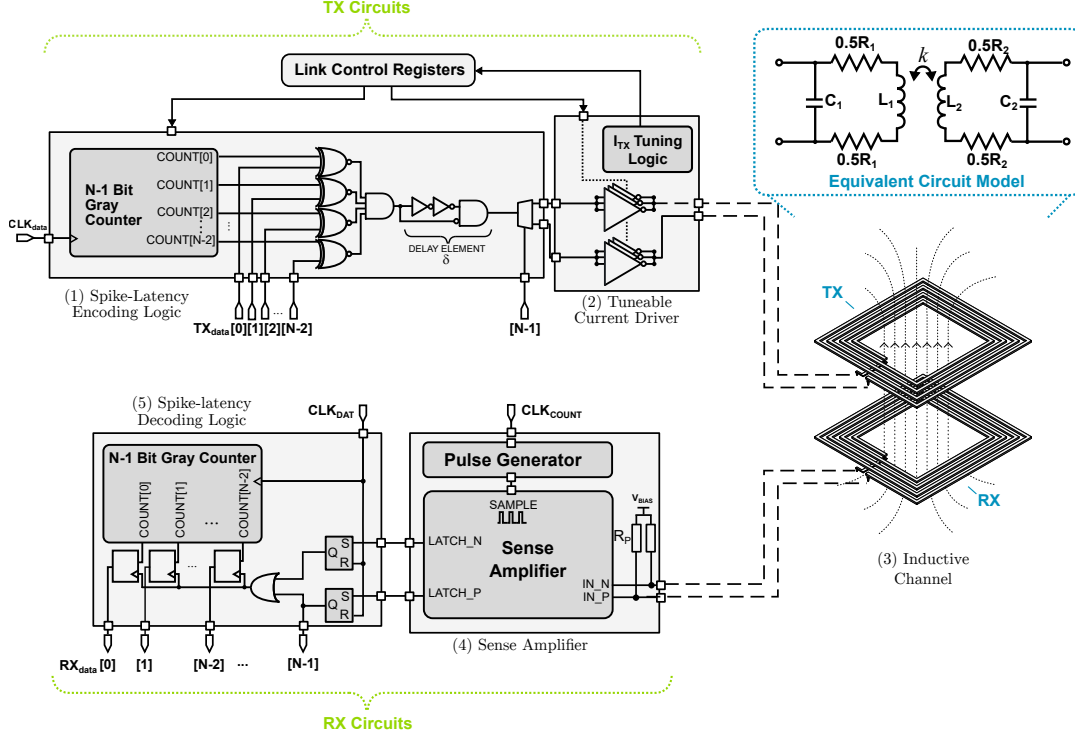


Figure 3.3: Schematic diagram showing architecture of proposed low-energy inductive coupling link consisting of: (1) spike latency encoding logic, (2) an tuneable current driver, to minimise the transmit current depending on the assembly quality, (3) the inductive coupling channel, (4) a sense amplifier to sample the received voltage signal, and (5) the spike latency decoding logic.

Binary	Decimal	Pulse Code			
		0	1	2	3
000	0				
001	1				
010	2				
011	3				

Binary	Decimal	Pulse Code			
		0	1	2	3
110	6				
111	7				
100	4				
101	5				

Figure 3.4: Example code-book using SET with parameter $N=3$. Incorrect phase/position decisions result in only one bit error.

the data which selects the *phase*. Here, the impact of increasing N on the logic size can be observed. Not only will a higher N result in a faster clock frequency, as N increases, one additional flip-flop will be required in the counter (in addition to extra match logic). To minimise the power consumption of the system, the width of the I_{TX} pulse is limited by a delay element with length δ , as shown on Figure 3.3. This is analogous to the δ delay used for modulation in the benchmark NRZ scheme.

To maximise the BER of the system, the $COUNT$ signal is implemented using a *Gray-coded*

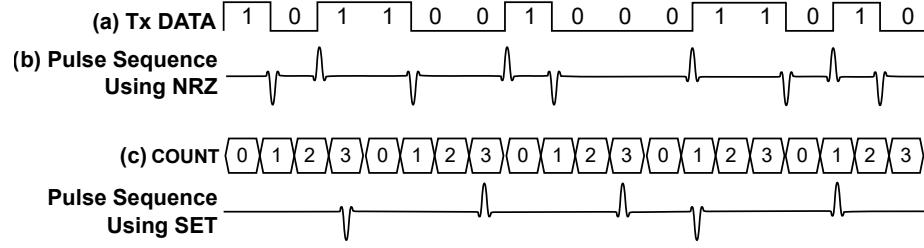


Figure 3.5: Illustration of the bit-stream to I_{TX} pulse mapping for the TX Data shown in (a) when using (b) the existing NRZ encoding benchmark approach [50, 51, 54, 123], and (c) the proposed SET scheme with the codebook shown in Table 3.4 ($N=3$).

counter, as shown. The use of a Gray-coded counter means that, if a pulse is detected in the wrong sub-window, the effect of the incorrect detection on the data frame is minimised (*e.g.* incorrect detection of the RX pulse at the $N \pm 1^{\text{th}}$ COUNT value only results in 1 bit of error in the whole frame). Additionally, the multiplexer stage means that an incorrect detection of *phase* results in only a single bit error in the output. An example code-book for these bit-to-code mappings for $N = 3$ is shown in Table 3.4. Here the first two binary bits are the Gray-coded counter value, and the final bit is the phase-based decision bit. Figure 3.5 illustrates how this works in practice when compared with the benchmark NRZ scheme. Here, using the benchmark NRZ scheme to transmit the data stream 0x591A (shown in Figure 3.5 (a)) results in 9 I_{TX} pulses (Figure 3.5 (b)), whilst using the proposed SET scheme, with the bit-to-code mappings from Table 3.4, requires only 5 I_{TX} pulses (Figure 3.5 (c)).

To minimise the power consumption of the en/de-coding logic, these digital blocks (counter, match logic etc.) are implemented in a separate supply domain where near-threshold voltage scaling is applied.

3.3.2 Tuneable Current Driver

The second element of the proposed low-energy transceiver is the tuneable current driver circuit. One of the benefits of using *wireless* 3D integration as opposed to traditional approaches, such as TSVs, is the relaxed assembly requirement when stacking each of the individual dies, making ICL-based 3D-ICs is ideally suited to low-cost applications. Low-cost assembly, however, means that variation from chip-to-chip is typically significant.

Figure 3.6 illustrates different variation mechanisms, introduced at assembly time, that can affect the channel coupling quality: Figure 3.6 (a) shows variation in quality between channels ① and ② due to adhesive thickness, Figure 3.6 (b) shows variation in quality due to lateral die-to-die stacking misalignment, Figure 3.6 (c) shows variation in quality between channels ① and ② due to substrate thickness, and Figure 3.6 (d) shows variation in quality between channels ① and ③ due to interference from other neighbouring links (②).

Because of these variations, ICLs must typically be designed to meet the *worst-case* assembly

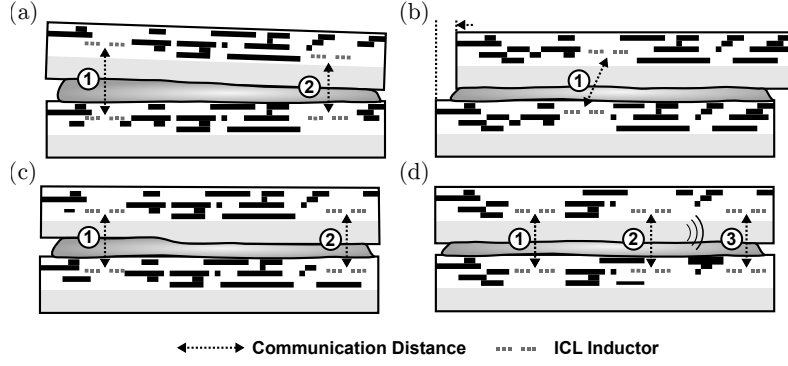


Figure 3.6: Illustration of channel quality variation (across ICL channels ①, ②, and ③) in end devices due to (a) uneven adhesive thickness, (b) laterally misaligned die attach, (c) uneven substrate thinning, and (d) communication over different numbers of dies.

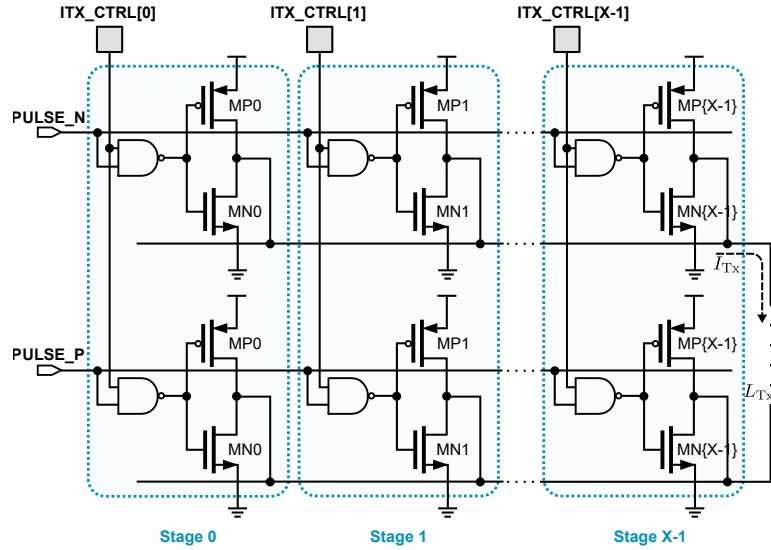


Figure 3.7: Schematic diagram showing structure of the proposed differential adaptive current driver circuit.

specification ($\text{Min}(k)$) meaning that often, the TX pulse current, I_{TX} , is much larger than needed for robust operation. To address the need for this over provisioning, in this work, the tuneable current driver architecture, shown in Figure 3.7, is proposed. The design uses a multi-stage differential driver, as shown in the figure. Each stage in the driver circuit (0 to $X - 1$) can be individually en/disabled according to the appropriate bit of the register ITX_CTRL . Each stage (0 to $X - 1$) is also comprised of inverters with differing transistor widths $w_{MN0} > w_{MN1} > w_{MN2} \dots$ etc. to allow precise control of I_{TX} using the control register. Using this approach, the dies can be stacked and then I_{TX} can be tuned (post-stacking) to compensate for the assembly defects shown in Figure 3.6, without using an unnecessarily large transmit current.

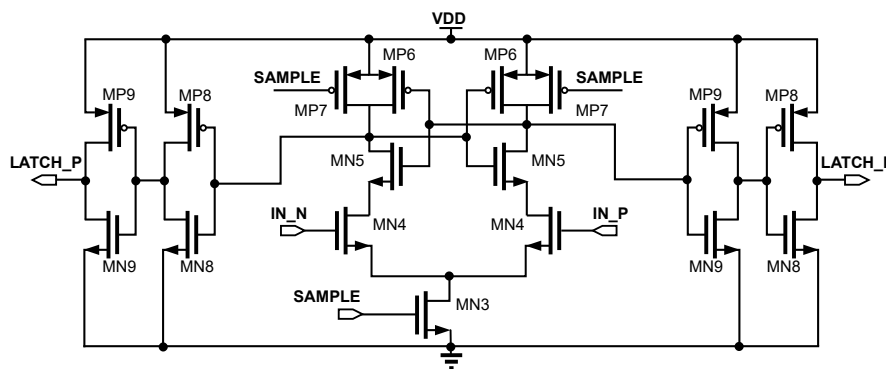


Figure 3.8: Buffered sense amplifier used in the proposed transceiver. The **SAMPLE** signal is a short, pulsed signal (with duration δ , generated at each rising edge of the **COUNT** signal. The buffered outputs **LATCH_N** and **LATCH_P** are passed to an SR latch (as shown in Figure 3.3) and then used as inputs to the SET decoder.

3.3.3 Inductive Channel

One other important element of the ICL is the inductive channel itself, consisting of two coupled planar metal inductors. To maximise the performance of the system, it is desirable to maximise the EM coupling, k (*c.f.* Figure 3.3) between the TX and RX inductors, such that the minimum I_{TX} pulse has maximum effect, as observed by the receiver. The coupling coefficient depends on a range of factors, however most notably the physical layout parameters of the inductor [134]. These are the inductor diameter (D), track width (w), track spacing (s), and number of turns (n). In order to determine best-performing parameters for these physical values and map them to an electrical link model, a manual optimisation flow was used where a range of different geometries were simulated using Finite Element Method (FEM) to evaluate the trade-off between silicon area and Electro-Magnetic (EM) coupling. The results from these simulations are presented in Section 3.4.1⁴.

3.3.4 Sense Amplifier

Figure 3.8 shows the sense-amplifier adopted in the proposed transceiver. The design is similar to that used to implement the NRZ scheme and operates on the basis that, whilst **SAMPLE** is high, the RX signal is amplified by the NMOS pair MN4. This causes a negative pulse at the drain of MN5 based on the differential potential IN_P - IN_N which is latched to avoid glitching, and then used to copy the RX *N*-bit counter to the output, as shown in Figure 3.3. To minimise the effects of process-variation induced bias in the SA, the layout of each corresponding device pair (MN4, MN5, MP6 and MP7) was performed using a common centroid topology, and the channel lengths were increased as far as possible (whilst maintaining the channel width:length ratio required for sampling at the target frequency).

⁴Chapter 4 of this thesis focusses on inductor layout optimisation for ICL channels, and so a more in-depth discussion of this design flow can be found in Section 4.2.

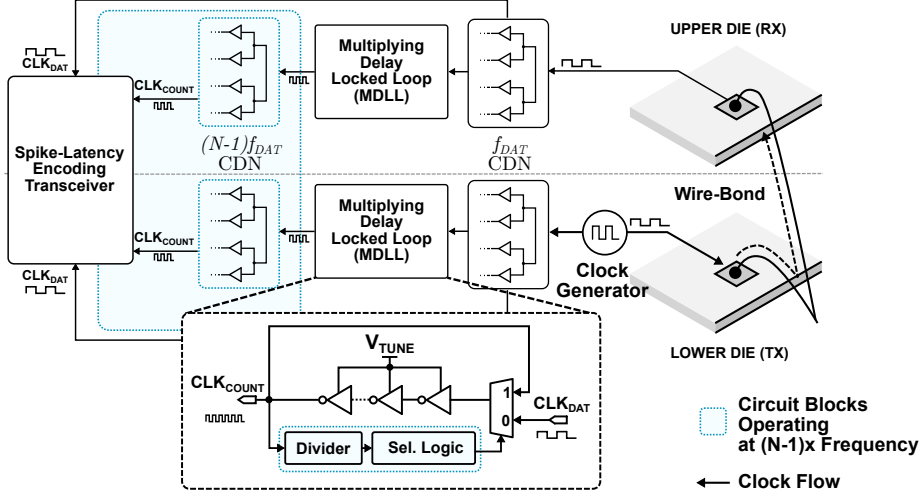


Figure 3.9: Illustration of the clock generation and synchronisation infrastructure in the presented test-chip. The data clock (with frequency f_{DAT}) is generated in the lower (TX) die and delivered to the upper (RX) die through a wire-bonded link. In each die, a multiplying delay locked loop (MDLL) is then used to generate the higher frequency f_{COUNT} clock.

In this implementation, the **SAMPLE** signal is generated by a synthesised programmable pulse generator block (incorporated in the SET logic⁵), as shown in Figure 3.3.

3.3.5 Clock Synchronisation

Although external to the transceiver circuits themselves, one other important consideration for implementing the transceiver is the TX/RX clock synchronisation infrastructure. To provide TX/RX clock synchronisation in this chapter, the clock architecture shown in Figure 3.9 is used. Here, the data clock (with frequency f_{DAT}) is generated in the lower (TX) die and delivered through a wire-bonded link to the upper (RX) die. To minimise jitter, this low-frequency (f_{DAT}) clock is then passed to a Multiplying Delay Locked Loop (MDLL) in each die which also generates the higher frequency **COUNT** clock ($f_{\text{COUNT}} = (N - 1)f_{\text{DAT}}$). Compared with the existing NRZ benchmark scheme, the areas that operate at higher frequency when using the SET scheme (and hence incur additional energy overheads) are: the pulse generator, the high frequency CDN ($(N - 1)f_{\text{DAT}}$), and the MDLL control logic. These elements are highlighted in blue on Figure 3.9, and their energy overheads are evaluated in Section 3.4.3.

In general, however, it is often more convenient to transmit the clock wirelessly using a separate ICL channel (this is explored later in this thesis, in Chapter 5). The SET approach proposed in this chapter could be combined with such a scheme (which would result in a different set of energy trade-offs), however wireless clock synchronization is beyond the scope

⁵For this reason its energy contribution is accounted for within the SET logic for the results presented in Sections 3.4 and 3.5.

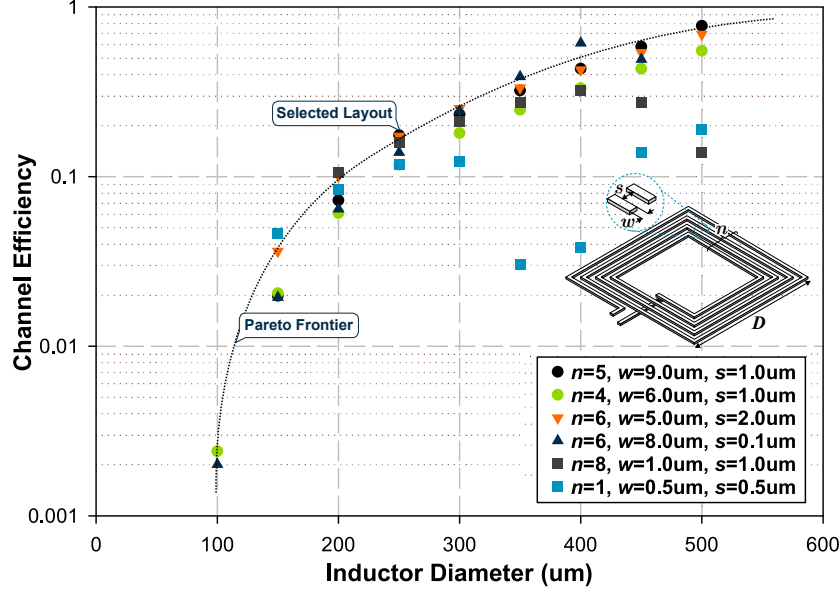


Figure 3.10: Scatter plot showing the simulated efficiency vs. area trade-off, including pareto-optimal frontier. The $250\mu\text{m} \times 250\mu\text{m}$ square geometry used evaluation in this section is highlighted.

of work in this chapter (which is focussed on the low-energy *data* transceiver design).

3.4 Experimental Validation and Results

This section presents experimental validation of the proposed low-energy inductive transceiver outlined previously (through simulation, practical validation is provided later in Section 3.5). Initially, Section 3.4.1 determines the geometric layout parameters for the ICL channel inductors. Following this, Section 3.4.2 evaluates the spike-latency encoding concept using the mathematical modelling presented in Section 3.2.1, and Section 3.4.3 performs post-layout simulation of the transceiver using SPICE.

3.4.1 ICL Layout Parameter Selection

As outlined in Section 3.3.3, the geometric parameters of the inductive channel (which largely determine the EM coupling coefficient, k) were selected manually by generating a range of layouts, and selecting the best-performing using FEM⁶. Figure 3.10 shows a scatter plot of channel efficiency (V_{RX}/V_{TX}) vs. diameter for a selection of these trialled geometries. As can be observed from the figure, a strong trade-off between efficiency and area exists and therefore, the $250\mu\text{m} \times 250\mu\text{m}$ layout on the ‘knee’ of the pareto curve was selected for use. Whilst this may seem like a significant overhead, prior research by Niitsu *et al.* [147] has

⁶Chapter 4 focusses on improving this design process, presenting a tool for automated ICL inductor layout generation.

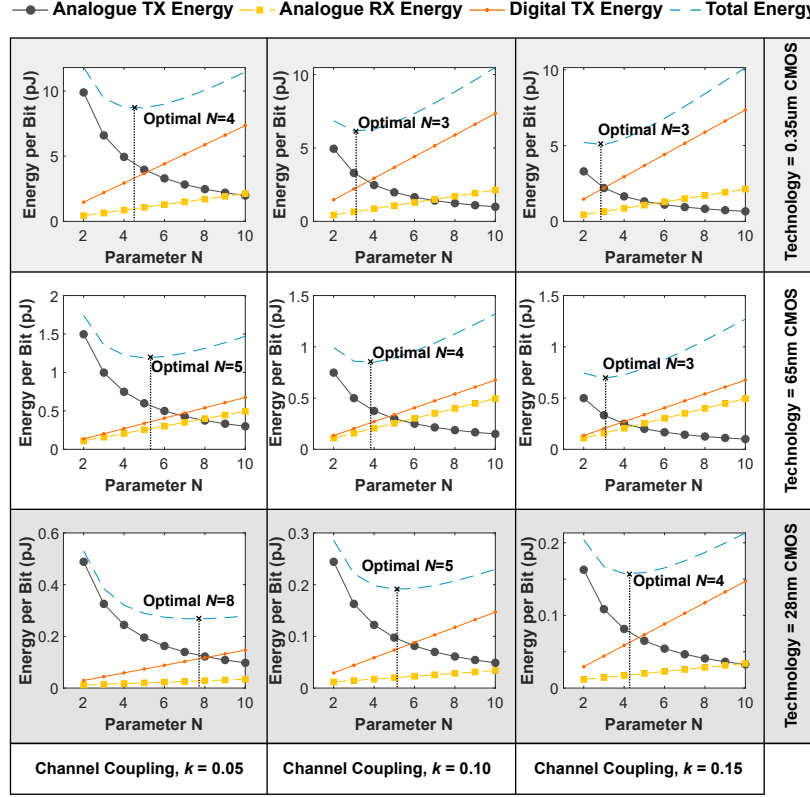


Figure 3.11: Mathematical modelling results showing how the transceiver energy (and optimal N value) varies as a function of N across three different technology nodes (28nm, 65nm and 0.35µm and three different channel coupling strengths ($k=0.05$, $k=0.10$, and $k=0.15$).

demonstrated that SRAM cells can be placed within the channel area without significant performance degradation, and that standard logic cells (automatic place and route) can be placed within the channel area with only a minimal performance impact (which can be overcome by increasing the TX power by around 9%) [147]. This implies that for certain applications (digital logic/memory) the area overhead of the ICL inductors is limited to the coil tracks themselves (typically in a high metal layer), and the interposed silicon can still be utilized⁷. The selected design has physical parameters $D = 250\text{ }\mu\text{m}$, $w = 9\text{ }\mu\text{m}$, $w = 1\text{ }\mu\text{m}$ and $n = 5$, corresponding to a channel efficiency ≈ 0.13 .

3.4.2 Validation using Mathematical Models

Having established the approximate coupling coefficient k that can be achieved within the $250\text{ }\mu\text{m} \times 250\text{ }\mu\text{m}$ area, this section evaluates the energy breakdown of the proposed scheme using the equations from Section 3.2.1 in conjunction with databook logic gate parameters for 0.35µm, 65nm, and 28nm technologies across a range of values for parameter N . Figure

⁷Exploring utilisation of the silicon area interposed by the ICL channel is beyond the scope of the work presented in this thesis, however is discussed as a future work item in Section 7.2.4, Chapter 7.

3.11 shows the results of these simulations, plotting the projected energy breakdown of the presented design, as N varies, for a range of values of k . As predicted, a trade-off between E_{pb} and N can be observed. In each case, energy savings (compared to BPM and NRZ schemes) are projected for every value of N between 2 and 10, and an optimal point (typically around $N=4$) exists where a good balance between I_{TX} and I_{SL} is established. At the less advanced process technology nodes the predicted optimal N value is lower (between 3-4) because the digital logic is expensive in terms of energy. As the process technology scales down to 28nm, the digital logic energy decreases and hence the predicted optimal N shifts to the right, increasing up-to a maximum of 8.

The results presented in Figure 3.11 also illustrate how the optimum value of N varies as the coupling strength k between dies changes. As the EM coupling deteriorates, k reduces, the TX current required for robust operation increases, and hence the best-performing value of N increases. For the inductor layout determined in the previous section, k is in the order of 0.13 and hence the modelling predicts that the optimal value of N will be around $N=3-4$, depending on the technology.

3.4.3 Experimental Validation using SPICE

Following the theoretical modelling of the proposed spike-latency encoding scheme, the presented transceiver was implemented in 0.35 μm , 65nm and 28nm CMOS technologies for validation using SPICE, alongside the existing state-of-the-art schemes (BPM [143], SPM [144] and NRZ [54]) to provide a benchmark for comparison. 0.35 μm , 65nm and 28nm CMOS technologies were selected to represent the full spectrum heterogeneity that would likely be found in IoT devices (the context of this work). For each chip, a substrate thickness of 100 μm was assumed (in line with presently available low-cost wafer lapping technologies) and an adhesive thickness of 10 μm was assumed for die attach.

Ansys HFSS was used for EM modelling of the inductive coupling channel, with the simulation setups shown in Figure 3.12a-3.12c. To include the effects of inter-link noise for BER evaluation, three ICL channels were included in the simulation environment, as shown in Figure 3.12a. Only measurements from the central channel (port S(0) \rightarrow port S(1)) were used for analysis, whilst the neighbouring channels (N(0) and N(1)) were assumed to transmit a PRBS for generating interference. Figure 3.12c shows a cross-sectional view of the three channels presented in Figure 3.12a, overlaid with the simulated magnetic field strength from Ansys HFSS. Here, it can be observed that a **H-Field** strength of around 1kA/m is achieved within each channel, with an inter-channel fringe coupling field strength of around 0.5kA/m. For finite element modelling, the Back End Of Line (BEOL) metal thicknesses shown in Figure 3.12b were used for each technology ((i) 0.35 μm - (iii) 28nm). The analogue circuit blocks (discussed above) were each sized for their respective technologies with the circuit architecture remaining the same between simulations. The only notable difference was that,

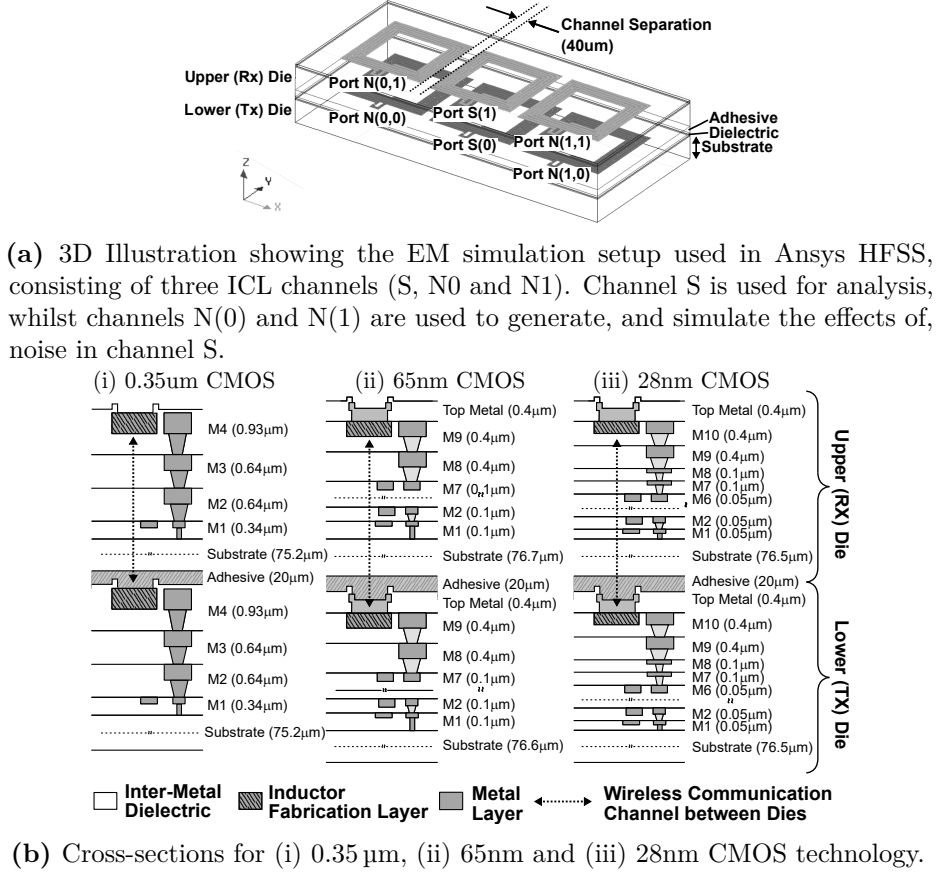


Figure 3.12: Figures showing the EM simulation setup used for evaluation in this section.

in the 28nm node, a level shifter was inserted between the encoding logic and the driver circuits, allowing the driver to be implemented using thick-oxide transistors to meet the $\text{Min}(I_{\text{TX}})$ requirements. A number of different comparisons were performed and the results are documented in the following subsections.

Area Evaluation

Figure 3.13 (a) shows the layout of the proposed low-energy transceiver in 65nm CMOS technology consisting of the TX/RX inductor ($250 \mu\text{m} \times 250 \mu\text{m}$), the sense amplifier ($15.4 \mu\text{m} \times 43.7 \mu\text{m}$), and the tuneable TX driver circuit ($36.0 \mu\text{m} \times 22.0 \mu\text{m}$). When compared to the

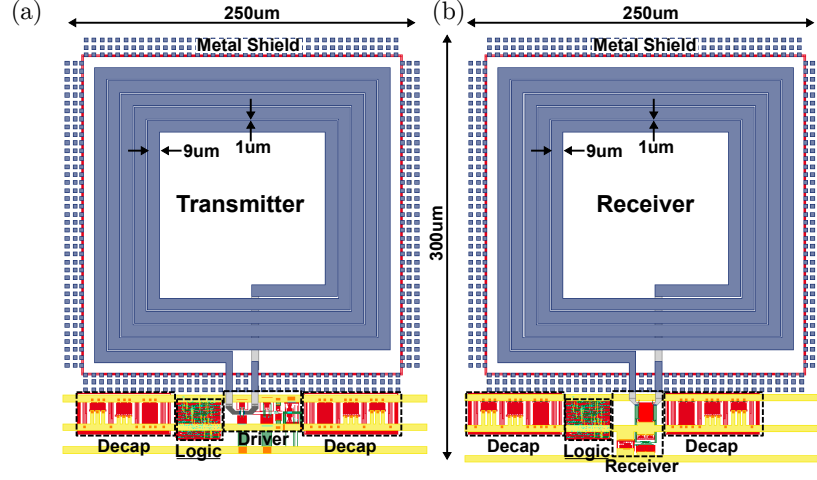


Figure 3.13: Layout of proposed low-energy transceiver in 65nm CMOS technology including (a) transmitter, and (b) receiver highlighting the additional digital coding logic area [underlined] (the only additional silicon area overhead when compared with the state-of-the-art scheme).

existing state-of-the-art transceivers using BPM, SPM or NRZ-encoding, the only additional area overhead is derived from the supporting digital logic which is highlighted on the figure ($13.6\mu\text{m} \times 17.8\mu\text{m}$ at the pictured 65nm technology node). As shown, the additional SET control logic does not add significant overhead to the footprint of the transceiver, in fact only contributing 0.4% of the overall area. The required digital logic was also synthesised in 28nm and $0.35\mu\text{m}$ technologies and found to measure $6.7\mu\text{m} \times 10.8\mu\text{m}$ and $54.2\mu\text{m} \times 92.0\mu\text{m}$ respectively, still only contributing a small overhead, less than 14% (in the case of $0.35\mu\text{m}$).

Bit Error Rate (BER) Evaluation

The BER of the proposed scheme was then evaluated in each technology using the channel model generated by Ansys HFSS. As shown in Figure 3.12a, the EM setup includes 3 channels, each of which transmits an equiprobable random bit stream. This generates noise in the channel of interest (as shown in Figure 3.12c), facilitating estimation of BER through Monte-Carlo simulation. The results of these simulations are presented in Table 3.2 alongside comparison to simulations implementing the existing BPM, SPM and NRZ benchmark schemes. For parity, BER analysis was performed at the same data-rate across each of the four designs, corresponding to the maximum operating frequency of the slowest transceiver in each technology (*i.e.* 800Mbps in 28nm CMOS technology, 1.05Gbps in 65nm CMOS technology, and 300Mbps in $0.35\mu\text{m}$ CMOS technology). As shown in the table, the measured BER when using the proposed transceiver is approximately equal to the BER achieved when using the BPM or NRZ approaches (and better than that achieved using SPM). This is due to the combination of using Gray-coded pulse mappings and phase-coding

Table 3.2: Simulated performance of the proposed low-energy transceiver (with optimal parameter N), compared to Bi-Phase Modulation (BPM) [143], Single Phase Modulation (SPM)[144], and Non-Return to Zero (NRZ) [50, 51, 54, 91, 123] transceivers across three technology nodes.

	Performance Metric	Bi-Phase Modulation (BPM) [143] †	Single Phase Modulation (SPM) [144] †	Existing State-of-The-Art (NRZ) [54] †	Proposed Approach
28nm	Total Footprint	0.064mm ²	0.064mm ²	0.064mm ²	0.064mm ²
	Max. Bandwidth	2.4Gbps	2.4Gbps	2.4Gbps	800Mbps
	BER (at 800Mbps)	9.8E-7	9.1E-5	2.2E-6	2.8E-6
	Energy-per-bit (at 800Mbps)	0.70pJ	0.36pJ	0.36pJ	0.26pJ (28.1% Reduction)
65nm	Total Footprint	0.066mm ²	0.066mm ²	0.066mm ²	0.066mm ²
	Max. Bandwidth	1.6Gbps	1.6Gbps	1.6Gbps	1.05Gbps
	BER (at 1.05Gbps)	9.2E-7	8E-5	1.15E-6	2.0E-6
	Energy-per-bit (at 1.05Gbps)	1.60pJ	0.93pJ	0.84pJ	0.66pJ (21.4% Reduction)
0.35µm	Total Footprint	0.075mm ²	0.075mm ²	0.075mm ²	0.0855mm ²
	Max. Bandwidth	450Mbps	450Mbps	450Mbps	300Mbps
	BER (at 300Mbps)	8.0E-7	6.3E-5	2.3E-6	1.2E-6
	Energy-per-bit (at 300Mbps)	14.96pJ	8.80pJ	8.50pJ	7.56pJ (11.1% Reduction)

† BPM, SPM and NRZ results are based on custom, manual SPICE implementations of the circuits presented in [143], [144] and [54] respectively, using each process technology.

(in which 180 degrees of phase shift exist between MSB ‘1’ and ‘0’ values). The latency when using the SET approach is, however, greater than that when using the existing NRZ approach as the full data frame must be present before transmission; when using SET, the latency is N clock cycles, rather just a single cycle.

Clock Jitter Sensitivity Evaluation

The sensitivity of the proposed approach to TX/RX clock jitter was also evaluated. Figure 3.14 shows the results of these simulations (performed using Monte-Carlo SPICE simulation in conjunction with the dual-Dirac model for low-BER extrapolation, described in Appendix B) across three technology nodes, (a) 0.35µm, (b) 65nm and (c) 28nm CMOS. For each node, the timing sensitivity was evaluated by inducing jitter in the RX clock signal and simulating the Bit Error Rate (BER). The grey bathtub curves show the timing sensitivity of the proposed SET scheme, and the black bathtub curves show the timing sensitivity of the benchmark NRZ scheme for the same f_{DAT} frequency. As can be observed, the proposed scheme is more sensitive to RX clock jitter (by between $3.6\times$ and $8.3\times$, depending on the technology node) due to the increased SA sample frequency. Whilst this does not limit the *BER performance* for the presented data rates (as the timing margin is greater than the maximum expected COUNT clock jitter, shown by the shaded area), it does have the effect of limiting the maximum transceiver *bandwidth*, as shown in Table 3.2. This bandwidth reduction represents the most significant trade-off for the additional energy gains achievable

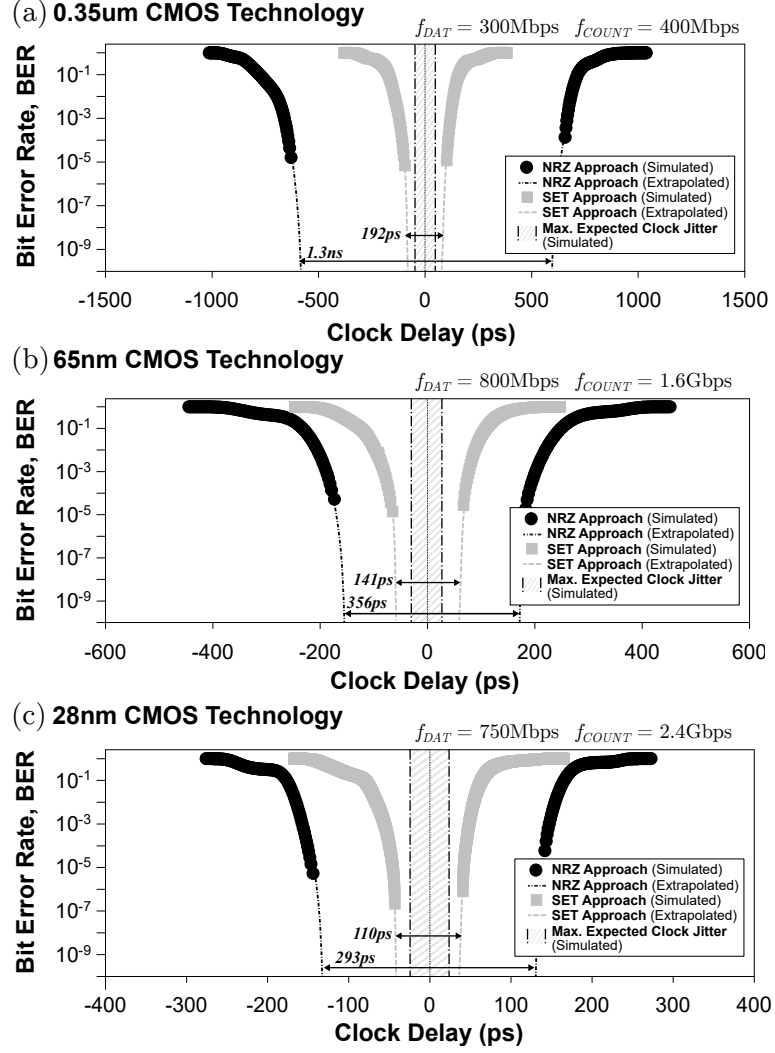


Figure 3.14: Bathtub curves showing the SAMPLE signal timing margin when using the proposed SET approach when compared with the inductive NRZ benchmark approach across 3 technology nodes: (a) 0.35 μm CMOS, (b) 65nm CMOS, and (c) 28nm CMOS. Silicon measurement of this timing margin in the 0.35 μm technology is presented later in Section 3.5.2.

using SET.

Energy Evaluation

The effectiveness of the proposed transceiver in reducing energy consumption (the primary motivation for this study) was then evaluated. The energy-per-bit of the proposed approach was measured for a range of values of N and compared with BPM, SPM and NRZ transceiver designs. As in Section 3.4.3 energy per bit was evaluated on an iso-data rate basis, for each technology node at a bit-rate corresponding to the maximum data rate of the slowest transceiver design. Figure 3.15 shows the energy required to transmit a single bit for each

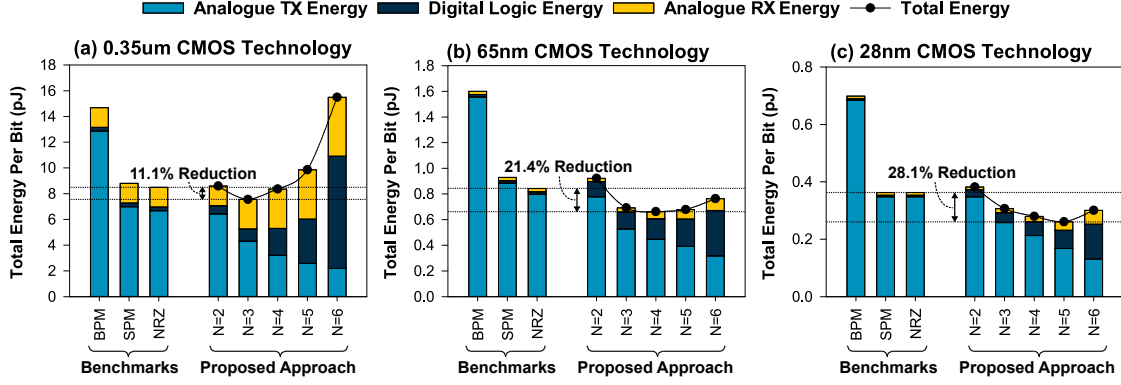


Figure 3.15: Simulated energy of the proposed transceiver (for various values of N) when compared to the compared to Bi-Phase Modulation (BPM) [143], Single Phase Modulation (SPM) [144], and Non-Return to Zero (NRZ) modulation [50, 51, 54, 123] benchmark designs at three different technology nodes. Results show improvements between 11.1% and 28.1% using the proposed scheme.

Table 3.3: Table showing the energy overhead associated with the higher frequency clock infrastructure, ΔE_{CI} .

Source of Energy Dissipation	Additional Energy Contribution (per Bit)		
	0.35um	65nm	28nm
(1) ΔE_{CDN} (per bit)	0.11pJ	0.019pJ	0.007pJ
(2) ΔE_{DLL} (per bit)	0.36pJ	0.034pJ	0.032pJ
Total Clock Infrastructure Overhead, ΔE_{CI} (per bit)	0.47pJ	0.053pJ	0.039pJ

of these cases across the three technology nodes. As can be observed from the figure, the proposed transceiver is successful in reducing the energy consumption by up to 62.7% when compared with previously published BPM transceivers, and 28.1% compared to the existing state-of-the-art in low-energy modulation, NRZ encoding. Figure 3.15 also validates the mathematical modelling in Section 3.2.1, demonstrating that $N=3-5$ performs optimally across the range of technologies considered.

Additional Clocking Overhead Evaluation

Although the additional dynamic energy associated with the $(N - 1) \times$ faster clock and SAMPLE pulse generation is accounted for in the simulation of the SET transceiver block, consideration should also be given to the additional energy overheads associated with the *implementation* of a faster clock (as discussed in Section 3.3.5). These additional energy overheads are derived from two main sources: (1) The additional energy consumed by the Clock Distribution Network (CDN) from the output of the MDLL to the sink node in the SET modulator, and (2) the additional energy consumed by the MDLL control logic to maintain a higher output frequency.

To evaluate the energy overhead of (1) and (2) in the presented design, the relevant modules

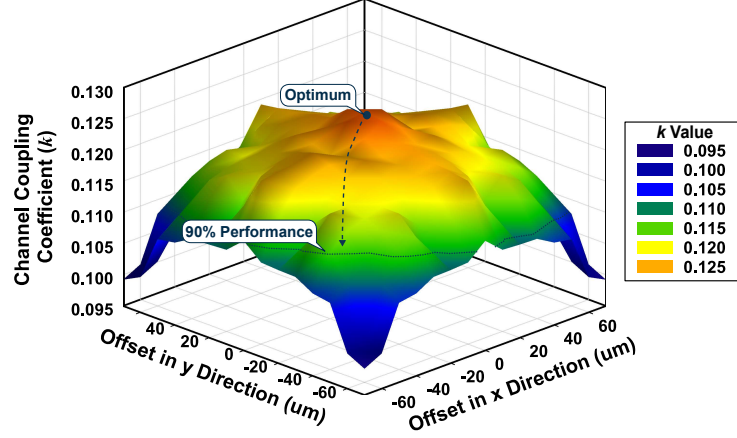


Figure 3.16: Simulated channel performance with respect to x and y die-to-die stacking misalignment (in terms of coupling coefficient, k).

of the clock distribution architecture were simulated in each of the three technology nodes. Simulations were performed at the data frequency (f_{DAT}) which is representative of the benchmark NRZ/SPM approaches, and at $(N - 1) \times f_{\text{DAT}}$, representative of the SET approach. Table 3.3 shows the energy difference (per bit) resulting from using the SET scheme for the CDN (ΔE_{CDN}) and the MDLL (ΔE_{DLL}). As can be observed, these energy overheads are small in comparison to the overall energy per bit (between 6-15% depending on the technology node), but still important to consider when designing such a system. It should also be noted, however, that these additional energy contributions are implementation dependent (*e.g.* would change if a wireless clock link was used) and can often be amortised when multiple parallel data links are present on the same chip.

The summary table (Table 3.5) on page 67 calculates the energy benefits of the proposed scheme (compared with the NRZ benchmark approach) taking into account this additional penalty, assuming 1 clock link (wire-bonded) per data link. Even considering these additional energy overheads, the proposed SET link still offers competitive energy reductions between 7.4% and 16.9% depending on the technology.

Misalignment Tolerance Evaluation

Finally, the tolerance of the proposed transceiver to lateral die-to-die stacking misalignment was also explored. As discussed in Section 3.3.2, one of the benefits of using wireless 3D integration is that it avoids the need for precise (and hence expensive) pick-and-place accuracy when performing the die stacking. To assess the tolerance of the channel to lateral placement misalignment, the channel coupling coefficient (k) was evaluated for various levels of offset using Finite Element Modelling (FEM). For each x and y misalignment value, the channel's performance was simulated using Ansys HFSS assuming the same 65nm EM set-up presented in Figure 3.12 (Section 3.4.3), but with a lateral offset introduced between the

upper and lower dies. The generated S-Parameters were then used to extract the mutual inductance, and hence coupling coefficient, at a representative frequency of 1.0GHz.

Figure 3.16 presents the results of these simulations, illustrating the effect of alignment accuracy on k . As shown, results suggest that the channel will tolerate up-to 40 μm of die-to-die misalignment in both x and y directions (a total diagonal offset of 56 μm) whilst maintaining performance within 10% of the optimum (representative of that which can be tolerated by tuning the ITX_CTRL register). When compared to 3D assembly using TSVs, which typically mandates sub-micron placement accuracy [22], this represents an approximately 100 \times improvement.

3.5 Case Study: Test-Chip Demonstration

Following the success of the proposed low-energy transceiver in SPICE, the design was implemented on a 2-tier 3D stacked silicon test-chip in 0.35 μm CMOS technology for practical performance evaluation. Figure 3.17 (a) shows a photograph of the assembled 2-tier test-chip with the upper (RX) and lower (TX) dies highlighted. Before stacking, each die was thinned to a height of 100 μm and attached using epoxy adhesive with 10 μm thickness as shown in Figure 3.17 (b). The dies were stacked in a Face-to-Back (F2B) arrangement resulting in a total communication distance of 110 μm through the silicon substrate, BEOL, and adhesive layers⁸. Silicon measurement results of the proposed SET transceiver's performance are outlined in the following sections.

3.5.1 Tuneable Current Driver Evaluation

Initially, the transmit pulse amplitude I_{TX} was selected using the tuneable current driver circuit. To find the optimal value of the ITX_CRL register (and hence I_{TX} amplitude), the BER of the link (missed pulses vs. total pulses, without the spike-latency modulation scheme) was measured whilst gradually sweeping the ITX_CRL register from 0 to 32. Figure 3.18 shows the results of this sweep for two separate test-chips: Chip **A** (which is assembled with perfect alignment in the inductive channel), and Chip **B** (which is assembled with an offset of 20 μm in the inductive channel, to demonstrate the effects of stacking misalignment during assembly). At the smallest settings (1,2,3) the TX current is low, and hence the pulses are not detected. As the ITX_CTRL register is incremented further, the link begins to operate. Eventually, both chips reach the target threshold BER ($1\text{E-}5$) at different tuning register values ($\text{ITX_CTRL} = 16$ in Chip **A**, and $\text{ITX_CTRL} = 26$ in Chip **B**, due to the assembly offset). This demonstrates that the proposed tuneable driver circuit is successful in overcoming significant packaging variations whilst maintaining performance within the specified bounds. At its tuned value, Chip **A** achieves a BER in the order of $1\text{E-}5$ with a pulse energy of 12.6pJ.

⁸Further details related to the design and manufacture of the test-chip are presented in Appendix C.

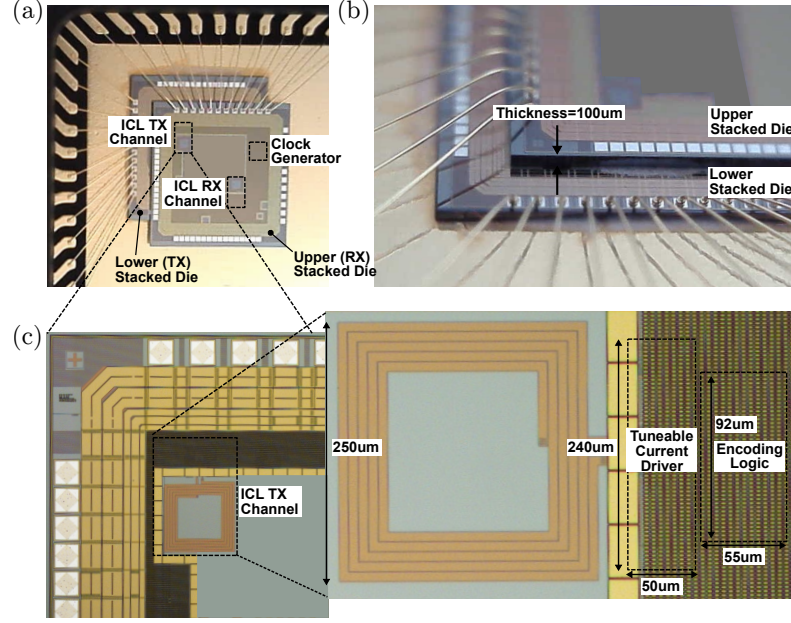


Figure 3.17: Micrograph of (a) the 2-tier stacked IC with wire-bonded power, reset and debug pins. (b) Side elevation showing vertical die stacking arrangement and communication distance. (c) A single die layout, showing the dimensions of the proposed transceiver and the 250 μm square channel used for evaluation.

3.5.2 Timing Margin Evaluation

Following this, the transceiver’s tolerance to variations in TX/RX clock delay (evaluated through simulation in Section 3.4.3) was empirically measured. Figure 3.19 revisits the bathtub curves presented in Section 3.4.3, this time comparing the *simulated* bathtub timing curve with the *measured* curve (varied by adjusting V_{TUNE} (*c.f.* Figure 3.9)). As shown on the figure, the measured timing margin is very close to the margin predicted by SPICE with the small variation likely attributed to on-chip noise in the V_{TUNE} supply. These silicon measurements also show that, whilst the sample margin is reduced when using the proposed scheme compared with the benchmark NRZ scheme (by approximately 90%), the transceiver can still operate within this margin, achieving a low BER, $< 10^{-5}$.

3.5.3 Energy-per-Bit Evaluation

The energy of the proposed transceiver (the primary motivation for this work) was then evaluated for a range of N values between 2 and 6 at the tuned `ITX_CTRL` register value. Energy was measured using knowledge of the transmit frequency combined with power measurements, taken with a Keysight B2900A Source Meter Unit (SMU). Figure 3.20 shows the results of these experiments highlighting the energy split between the TX and RX dies when compared to the benchmark approaches. Here it can be observed that the optimal parameter of $N=3$ yields a 13% energy reduction compared to the state-of-the-art NRZ

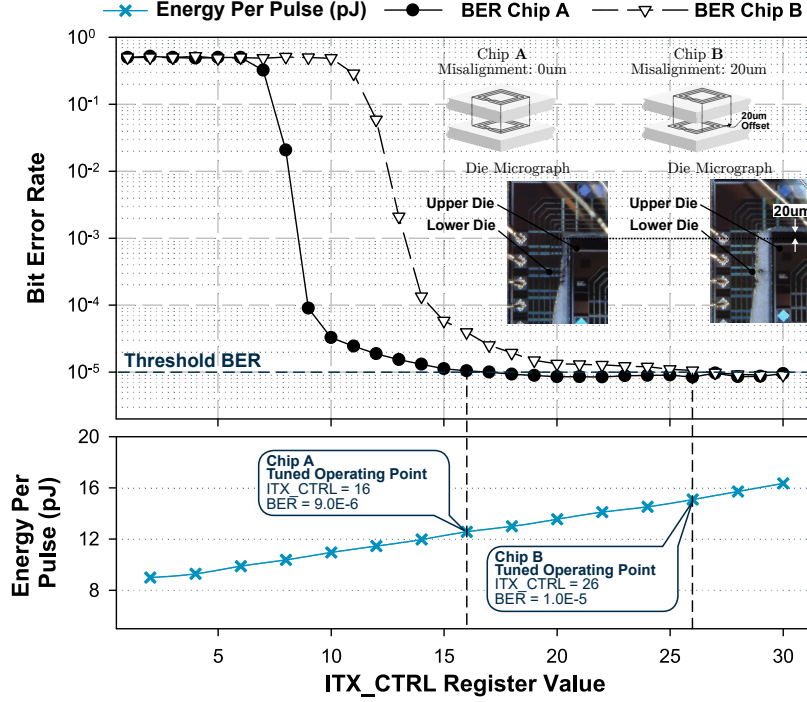


Figure 3.18: Measured link BER and energy-per-pulse as I_{TX} control register (ITX_CTRL) varies for two test chips: Chip **A** assembled with perfect die-to-die stacking alignment, and Chip **B** assembled with a significant 20 μm stacking offset (equating to almost 10% of the channel size) to explore the effects of die-to-die misalignment in the stacking process.

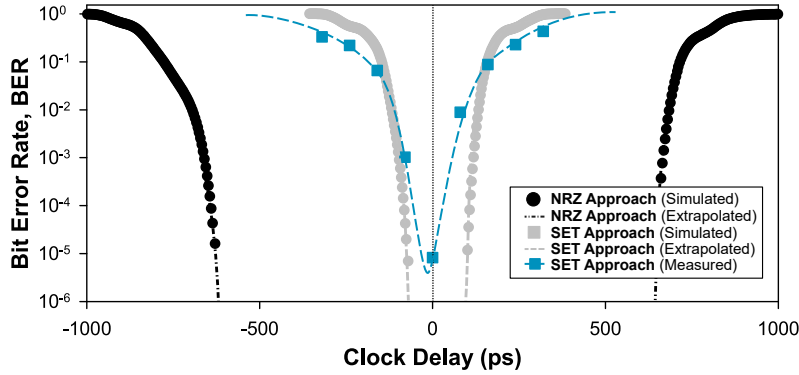


Figure 3.19: Bathtub curves showing the measured **SAMPLE** signal timing margin when compared to the simulated margin for the proposed SET approach and NRZ benchmark approach in 0.35 μm CMOS. Silicon measurement results are taken from Chip **A** with $f_{DAT}=300MHz$ and $f_{COUNT}=400MHz$.

encoding benchmark *with* I_{TX} tuning, representing a significant overall energy reduction when using the SET scheme. It can also be observed that the results closely match the simulation predictions (with the measured energy-per-bit being 7.4pJ, and the simulated energy-per-bit being 7.6pJ), indicating high confidence in the SPICE-based energy results presented in Section 3.4.3.

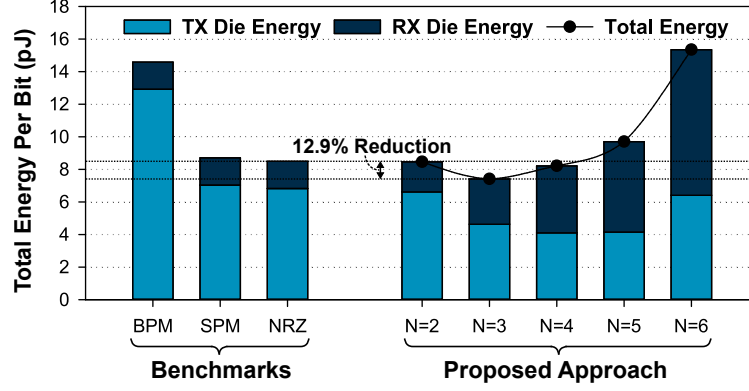


Figure 3.20: ICL energy consumption in TX and RX dies, as N varies, compared with Bi-Phase Modulation (BPM) [143], Single Phase Modulation (SPM) [144], and Non-Return to Zero (NRZ) modulation [50, 51, 54, 123] benchmarks (measured silicon results in $0.35\ \mu\text{m}$ technology).

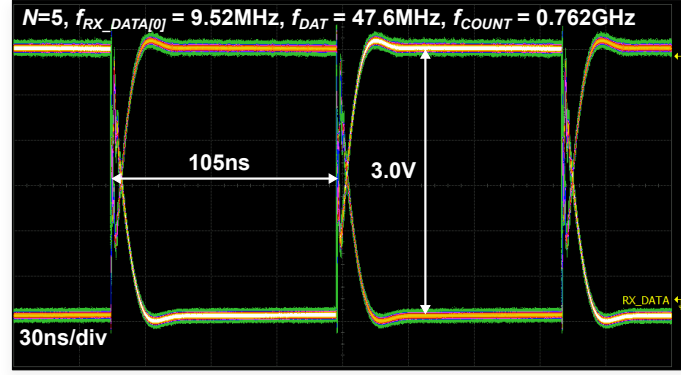
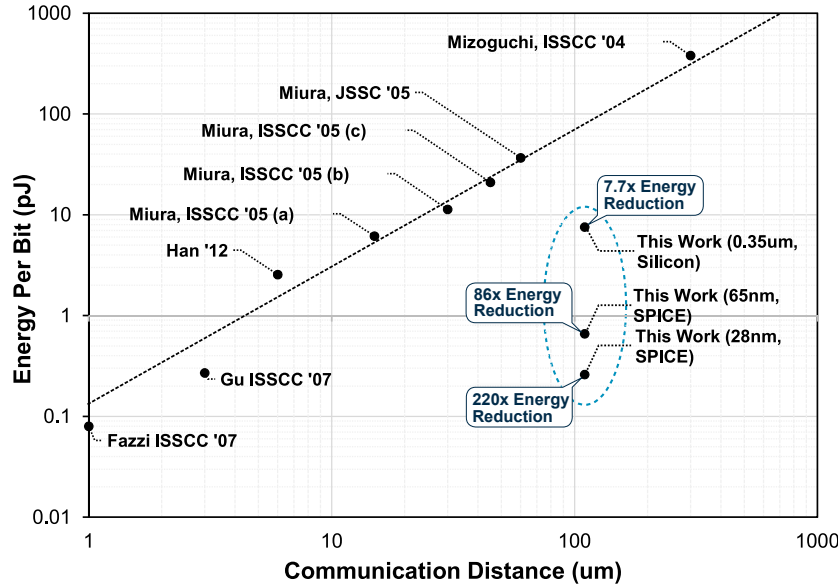


Figure 3.21: Measured eye diagram showing RX_DATA[0] (the LSB of the data output) from the proposed transceiver implemented on the $0.35\ \mu\text{m}$ 2-tier test-chip. $f_{COUNT} = 0.76\text{GHz}$, $N = 5$, $f_{DAT} = 47.6\text{MHz}$, $f_{RX_DATA[0]} = 9.52\text{MHz}$.

Figure 3.21 shows an eye diagram of the Least Significant Bit (LSB) of the RX data output when using the proposed transceiver with parameters $ITX_CTRL=16$ and $N=5$ at the maximum operation frequency, $f_{DAT}=47.6\text{MHz}$ (for $N=5$). Although the eye opening in the RX_DATA signal at this frequency is still wide, in order to meet the operating frequency of 47.6MHz with parameter $N=5$ requires a COUNT frequency, $f_{COUNT}=0.762\text{GHz}$ which represents the upper-bound when considering the sense-amplifier timing margin (discussed in Section 3.4.3). This has the effect of limiting the maximum frequency of the transceiver. For the algorithmic parameter $N=3$ (corresponding to the optimal *energy* efficiency), the maximum data rate was measured to be 266Mbps . Although this is a reduction when compared to the NRZ scheme, the 266Mbps data-rate is ample for most IoT applications (which form the motivation for this work). Table 3.4 summarises the test-chip measurements presented in this section.

Table 3.4: Measured performance of the proposed inductive transceiver (compared to simulated results from Section 3.4.3).

Evaluation Metric	Simulated Performance	Measured Performance
Technology	2-tier stacked 0.35 μ m CMOS	
Communication Distance	110 μ m (100 μ m chip + 10 μ m adhesive)	
Average Energy Per Bit	7.6pJ/bit	7.4pJ/bit
Average Bit Error Rate	1.2E-6	9.0E-6
Channel Area	250 μ m \times 250 μ m (0.063mm ²)	
Transceiver Circuits Area	TX:0.0225mm ² , RX:0.0264mm ²	
Maximum Data Rate	300Mbps/channel	266Mbps/channel

**Figure 3.22:** Energy-per-bit versus communication distance comparison of the proposed transceiver with numbers reported by previously published state-of-the-art works (*Han* [40], *Gu* [149], *Fazzi* [148], *Miura* [49, 51], and *Mizoguchi* [123]). [This comparison is directly based on published figures of merit, and therefore does not account for variations in process technology between data-points].

To demonstrate the benefits achieved by combining this approach with the tuneable pulse driver circuit, Figure 3.22 compares this proposed design with leading published research. Works [148] and [149] implement near-field *capacitive* communication, and [40, 49, 51, 123] use *inductive* communication (as adopted in this chapter). Figure 3.22 plots the energy-per-bit against communication distance for each approach. When compared to prior-art, results indicate that the proposed transceiver achieves a $7.7\times$ reduction in energy consumption (normalised with respect to the 110 μ m channel distance) when implemented in 0.35 μ m technology. Additionally, simulation results show even more significant improvements in 65nm and 28nm technologies (of $86\times$ and $220\times$ respectively).

Table 3.5: Overall comparison of proposed transceiver with the existing state-of-the-art (Inductive NRZ encoding [50, 51, 54, 91, 123]).

	Technology = 28nm CMOS		Technology = 65nm CMOS		Technology = 0.35um CMOS		
Evaluation Metric	State-of-the-Art (NRZ) [54] [†]	Proposed Approach	State-of-the-Art (NRZ) [54] [†]	Proposed Approach	State-of-the-Art (NRZ) [54] [†]	Proposed Approach	Proposed Approach
	(SPICE)		(SPICE)		(SPICE)		(Silicon)
Transceiver Circuits Area	1152um ²	1230um ²	1685um ²	1949um ²	24917um ²	32497um ²	
Total Area	0.064mm ²	0.064mm ²	0.066mm ²	0.066mm ²	0.075mm ²	0.086mm ²	0.086mm ²
Latency	1 cycle	5 cycles	1 cycle	4 cycles	1 cycle	3 cycles	3 cycles
Energy Per Bit	0.36pJ	0.26pJ (28.1% Reduction)	0.84pJ	0.66pJ (21.4% Reduction)	8.5pJ	7.6pJ (11.1% Reduction)	7.4pJ (13% Reduction)
ΔE_{CI}	0pJ	0.039pJ	0pJ	0.053pJ	0pJ	0.47pJ	
Energy Per Bit inc. ΔE_{CI}	0.36pJ	0.30pJ (16.9% Reduction)	0.84pJ	0.71pJ (15.1% Reduction)	0.85pJ	0.79pJ (7.4% Reduction)	
Energy Break-down							

[†] NRZ results are based on a custom, manual SPICE implementation of the circuits presented in [54] for each process technology node.

3.6 Discussion

Having validated the proposed transceiver through simulation and physical test-chip measurements, this chapter has demonstrated that significant energy savings (>28%) can be achieved through using the proposed Spike-latency Encoding Transceiver (SET). Table 3.5 shows an overall comparison of SET, and the existing state-of-the-art in terms of energy efficiency, NRZ encoding, combining physical test-chip results from Section 3.5 and SPICE results from Section 3.4.3. As can be observed from the table, the proposed approach outperforms prior-art across all test-cases (in terms of energy) by between 11% and 28%, depending on the technology node.

Whilst the proposed approach minimises *energy* (which was the goal of this work, motivated by the requirements of IoT devices), this chapter also highlights the importance of tailoring the modulation approach to suit the target application/integration scenario. Applications requiring high-bandwidths with low error-rates may favour Bi-Phase Modulation, (BPM), however this is energy-expensive as one TX pulse is required per transmitted bit. Conversely, the proposed SET scheme is ideally suited for low-energy applications where latency and bandwidth are less important.

A similar trade-off can be observed in the selection of the spike-latency algorithm's parameters. As the aim of the proposed transceiver design is to reduce the number of TX pulses required to transmit a given data stream, the modelling presented in Section 3.2.1 suggests that the energy reductions achieved using SET will be even more pronounced when communicating across greater distances (as the energy required for each TX pulse will be even larger). By the same reasoning, in systems where the communication distance is reduced (for example if *face-to-face* die stacking is performed) the NRZ benchmark approach may provide superior energy efficiency. Transient noise will also influence this trade-off. One advantage of using the proposed scheme in favour of existing approaches is that the algorithmic parameter N (and the tuneable current driver strength) can be dynamically tuned at runtime to compensate for channel noise. For example, dynamically increasing the drive current to counteract noise caused by an on-chip radio, and simultaneously increasing N to compensate and maintain a constant energy consumption.

Finally, as IoT devices are becoming increasingly heterogeneous, another important factor is evaluating how the proposed approach will perform at more advanced process nodes. As SET trades-off expensive analogue transmit pulses (which map to the \mathbf{H} -field strength, and hence will not scale with process technology) in favour of additional digital processing (which will reduce in energy as process technology scales), the results presented in Section 3.4.3 indicate that the energy of the proposed approach will scale at a faster rate than existing schemes with process technology size. To illustrate this, Figure 3.23 shows a plot of technology node vs. projected transceiver energy consumption on a logarithmic scale. The marked points show the three technology nodes explored in this chapter (28nm, 65nm and 0.35 μm) and the dashed-line illustrates the expected trend with technology scaling⁹. Whilst the maximum energy savings compared to the state-of-the-art demonstrated in this chapter are in the order of 28%, following the trend to the 3nm node and beyond suggests that the proposed approach has potential to offer improvements of over 80% when compared to BPM transceivers and 35% when compared to the existing state-of-the-art NRZ inductive transceiver designs.

3.7 Summary

This chapter presented a low-energy inductive transceiver for ICL-based 3D-ICs. The proposed transceiver combines: (1) a novel modulation scheme (spike-latency encoding) to perform time-domain data coding, and (2) a tuneable current driver circuit to adjust the transmit current to a minimum, depending on the 3D stacking quality. The proposed transceiver was modelled mathematically, simulated in 0.35 μm , 65nm and 28nm CMOS technologies, and experimentally validated in a 2-tier 3D stacked silicon test-chip. Silicon evaluation of the proposed transceiver demonstrates an energy of 7.4pJ/bit, representing a

⁹This trend is extrapolated based upon the results presented in this chapter.

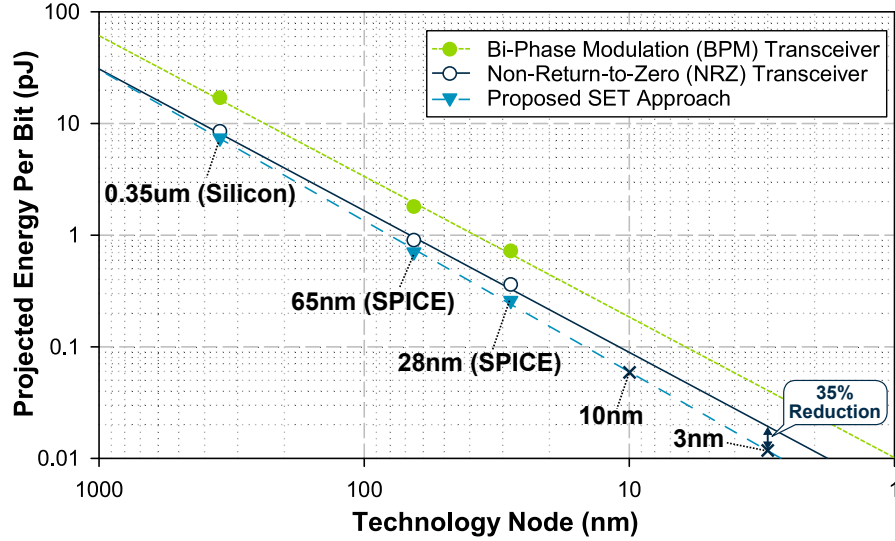


Figure 3.23: Projected energy savings when using the proposed spike-latency encoding scheme, when compared with the BPM [143] and NRZ benchmark designs [50, 51, 54, 123], as process technology scales.

reduction $>13\%$ when compared to previously reported schemes (or 7.4% when considering the additional energy overheads of peripheral clock timing control circuits). Simulated results show even greater energy savings (up to 28%) at more advanced technology nodes. Combined with the adaptive current driver, this equates to a $7.7\times$ improvement in energy-per-bit compared to state-of-the-art implementations. Whilst these gains come at the cost of a slight decrease in maximum data-rate, the transceiver proposed in this chapter shows strong promise for use in low-power, low-cost IoT devices which do not require gigabit operating bandwidths.

Chapter 4

Design and Optimisation of Inductive Coupling Channels

As discussed in the previous chapter, inductive coupling links are often criticised for their inferior power efficiency when compared with TSVs and therefore, when adopting ICLs in 3D-ICs, it is essential that the utilised channel inductor geometries are optimised. Presently, this involves simulating the coil layout using finite element method (FEM) analysis tools and then converting the system's EM characteristics into equivalent circuit models that can be handled by electrical simulators (*e.g.* SPICE), as was performed in Section 3.4.1 of Chapter 3. The layout can then be manually adjusted, and the process repeated until a satisfactory solution is found. Although this is an adequate method, solvers using FEM, can often take several hours to converge, even whilst analysing a single geometry [53]. Due to this, determining coil pairs with *optimised* geometries (which typically necessitates analysing thousands of subtly different inductor layouts) is extremely computationally expensive, if not impossible.

To partially reduce this complexity, all previous works surrounding 3D system integration using ICLs utilise *uniform* spiral inductors (where the track width and spacing remains constant between turns of the inductor) [114, 135, 150]. Whilst this reduces the design complexity of the system, in this chapter, it is demonstrated that non-uniform inductor layouts (with variable width and spacing) are often more efficient. This chapter also addresses the challenge of ICL design and optimisation, performing detailed analysis of ICL layout requirements, styles and topologies (Section 4.2), proposing a method of rapidly determining optimal coil layouts for ICLs (Section 4.3). This work is brought together in a *CAD-tool for Optimisation of Inductive coupling Links for 3D-ICs* (COIL-3D) which is a software tool for integration with inductive link 3D-IC design flows.

The main novel contributions of this chapter include:

- Detailed modelling and analysis of ICL requirements considering typical transceiver architectures.

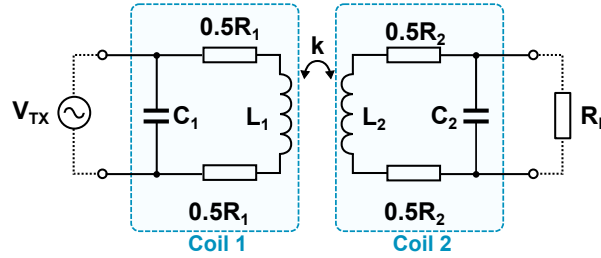


Figure 4.1: Equivalent circuit model of an ICL channel [51] which assumes that each coil, i , can be accurately modelled by its resistance (R_i), capacitance (C_i), inductance (L_i).

- Proposal of a graduated width-spacing inductor layout to improve ICL performance by up to 53.7 %.
- A comprehensive scalable inductor model for simulating the performance of ICL layouts, in addition to mathematical expressions for determining the scalable model parameters that achieve an average accuracy within 7.8% of FEM tools whilst reducing the computational overhead by four orders-of-magnitude.
- A refined optimisation flow for determining optimised ICL geometries in 3D-ICs (where both *data* and *power* ICLs are considered) that reduces the number of trial iterations by 3 orders-of-magnitude.

The remainder of this chapter is organised as follows. Background and work related to ICL implementation and optimisation is presented in Section 4.1, modelling and analysis of ICL requirements (for both *power* and *data* ICLs) is presented in Section 4.2, which concludes with formulations for the optimisation problems which this work solves. Following this, Section 4.3 outlines the proposed analysis and optimisation approach, before evaluation (Section 4.4) and a use-case example (Section 4.5). Finally, the chapter is summarised in Section 4.4.

4.1 Background and Related Work

When using ICLs to deliver either power or data between tiers of a 3D-IC, the physical layout of the inductive channel is typically designed to maximise the energy efficiency of the ICL, given a specific area constraint [135, 151]. To achieve this, requires the ability to evaluate a given layout, and hence model the *electrical* performance of the link given a set of *physical* design parameters

A range of lumped equivalent electrical models exist for simulating the performance of planar on-chip spiral inductors, typically using a π topology [152–154]. Lumped π models have been demonstrated to exhibit high accuracy when modelling on chip RF inductors, however it has been proposed that when considering vertically stacked on-chip inductors such as those in ICL-based 3D-ICs, a simpler RLMC model can be used [51]. This simpler model is

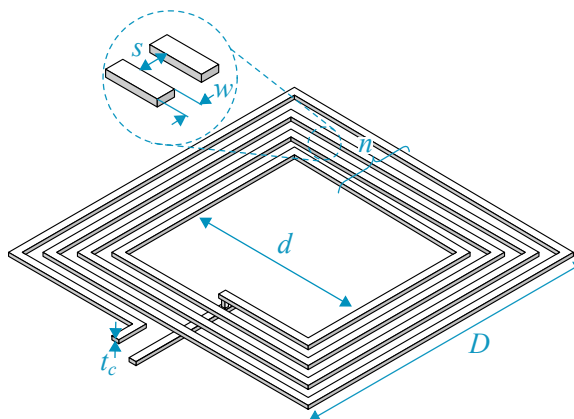


Figure 4.2: Geometric parameters of a square planar coil (outer and inner dimensions D and d , number of turns n , thickness t_c , trace width w and spacing s).

shown in Figure 4.1, and assumes that an ICL pair can be modelled by the resistance (R_i), capacitance (C_i) and self-inductance (L_i) of each coil, in addition to the EM coupling (k) that exists between the two coils. In Figure 4.1, R_L represents the load impedance, whilst V_{TX} represents the transmitted voltage signal. This model is widely reported to exhibit sufficient accuracy for evaluating the performance of an ICL, and hence will also be adopted in the work presented in this chapter¹.

As discussed above, previous works proposing and considering the use of ICLs for 3D integration use a manual simulation flow for evaluating the performance of inductor layouts where the layout parameters (shown in Figure 4.2, consisting of outer and inner dimensions (D and d), number of turns (n), thickness (t_c), trace width (w) and trace spacing (s)) are arbitrarily defined and evaluated using a full-wave field-solver. Once the channel's performance has been analysed, the layout parameters are slightly adjusted, and the process repeated until two adequate inductors are realised. Full-wave simulation is typically performed using comprehensive FEM software packages such as CST Studio or Ansys HFSS which provide accurate solutions by subdividing the 3D problem space into a mesh of smaller 'finite-elements'. In each finite element, the tool iteratively searches for approximate solutions to represent the magnetic field (in terms of flux density, magnetic scalar potential, and magnetic vector potential [155]) that then get combined and interpolated to model the overall system. Although FEM is still an approximate approach, it offers very high accuracy, however this comes at the expense of long runtimes (with such solvers often taking hours of compute to converge at a single solution, making them too slow for use in optimisation [53]). Other approaches for converting physical ICL inductor parameters to electrical models include using the Partial Element Equivalent Circuit (PEEC) method as the authors of [156] do. The PEEC method is based on the Electric Field Integral Equation (EFIE), and facilitates electrical modelling

¹The error introduced by these modelling assumptions (when compared with wide-band fitted models) is assessed later in Section 4.4.2.

of EM systems by mapping electric field interactions to capacitances and magnetic field interactions to inductances in order to form a circuit model. Like in FEM, the 3D inductor structure is discretised when using the PEEC method, such that the overall system is the superposition of many elements [156]. The PEEC method has the benefit of mapping designs directly to the electrical domain (unlike FEM where manual fitting of an electrical model is required), however is still a compute intensive process [156].

More rapid CAD-based alternative approaches include application specific tools such as SPIRAL and ASITIC [130], developed for on-chip inductor analysis. These use electrostatic and magnetostatic approximations to provide much faster modelling, however, are focussed on RF design and hence lack the ability to analyse Mutual Inductance (MI) between vertically stacked inductors. This MI evaluation is, however, the most challenging part of the analysis, and is highly important for ICL applications. Aside from using general numerical methods such as FEM to model MI, several approximate expressions have been proposed, starting with Maxwell’s MI formulae for circular loops with intersecting axes [157]. Based on Maxwell’s formulae, in 1916, Butterworth presented expressions for the mutual inductance of circular loops with parallel axes [158], which led to Grover’s filament method proposal in 1922 (that still remains one of the most popular methods for MI modelling) [159]. The filament method can be used to express the MI of inclined circular filaments, in any position, as a single integral [160]. A range of works exploring ICL design and implementation have used distilled versions of this model for MI modelling including [134], where a set of semi-empirical expressions for deriving the power efficiency of an inductive coupling *power* link are presented. These papers, however, typically focus on larger inductive coupling links (*e.g.* for bio-medical implants) and as a result, many of the approximations used do not hold true for the micron-scale coils used in 3D-ICs.

Work [52] by Hsu *et al.* is the only related publication proposing an automated optimisation flow for inductive coupling *data* links. Hsu *et al.* use a greedy linear optimisation algorithm that optimises each coil in the link separately to reduce the time complexity of the approach. Whilst this allows the algorithm to complete in a reasonable time, considering the two inductors separately means that the overall link efficiency may not be optimal. The methodology presented in [52] is also geared towards NRZ encoded data transmission using H-bridge transceivers and would require modification for analysis of alternative inductive coupling transceivers or wireless power delivery.

This chapter augments these previous works proposing: (1) a rapid solver for evaluating inductor layouts quickly and accurately, and (2) a refined optimisation flow to rapidly identify *best-performing* inductor layouts. These two contributions are provided together in a software tool, ‘COIL-3D’.

Table 4.1: Parameter notation reference for the mathematical modelling presented in this chapter.

Layout Parameters	
Parameter	Description
D	Coil outer side length
d	Coil inner length
w	Coil track width
s	Coil track spacing
g	Minimum technology grid unit
X	Communication distance
t_c	Coil metal thickness
ϕ	Coil fill-factor
ℓ	Length of single coil segment
χ (χ_w or χ_s)	Graduation coefficient (turn-to-turn graduation in track <i>width</i> or <i>spacing</i>)

Electrical Parameters	
Parameter	Description
L	Coil self-inductance (between terminals)
R	Coil resistance (between terminals)
C	Coil capacitance (between terminals)
M	Mutual inductance (between coils)
η_{dat}	Data delivery efficiency
η_{pow}	Power delivery efficiency
f_{sr}	Coil self-resonant frequency
σ	Coil metal conductivity
R_L	Receiver load resistance
f	Frequency of ICL operation

4.2 Modelling and Analysis of ICLs

Before presenting an ICL optimisation flow, the requirements of ICL inductors must first be classified. Due to the two types of ICL which exist (namely *power* and *data* ICLs), two separate optimisation objectives will be defined in Sections 4.2.1 and 4.2.2 below. To add clarity to the explanations presented in the following sections, Table 4.1 summarises the parameter notation adopted throughout the remainder of the chapter. Each parameter is referred to using the notation from the table, and where applicable, subscripts may be added in the following order: the coil number within the stack (*e.g.* 1 refers to the bottom coil, 2 the one above it *etc.*), the turn number, and the segment number (within the turn). As an example, w_{ijk} refers to the width of segment k of turn j in coil number i .

4.2.1 Inductive Coupling Data Links

As discussed in the previous chapter, a typical ICL data link uses an H-bridge transmitter, driven by the TX data signal where one side is slightly delayed by a time δ compared with the other (assuming the use of inductive NRZ encoding [54]). This has the effect of transmitting short current pulses corresponding to the data edges. These current pulses induce a small voltage in the receiver coil which can be detected using a Sense Amplifier Flip Flop (SAFF) arrangement [123].

If a maximum silicon area constraint is defined, the optimisation target is typically to minimise the power consumption of the system whilst communicating data at the required operating frequency. The power consumption of the ICL transmitter, over a period T (assuming transmission of an equiprobable random binary stream) can be calculated by:

$$P = \frac{1}{T} \int_0^T V_{TX} I_{TX}(t) dt + P_{circuit} \quad (4.1)$$

where I_{TX} is the transmit current and $P_{circuit}$ is the power consumed by the supporting transmitter circuitry. As established through the analysis in Chapter 3, the first term constitutes the majority of the power consumption whilst the latter is negligible in comparison (for the purpose of channel optimisation). The voltage induced in the RX (secondary) coil is given by equation 4.2 below.

$$V_{RX} = k\sqrt{L_{TX}L_{RX}} \cdot \frac{dI_{TX}}{dt} \quad (4.2)$$

where I_{TX} is the transmitted current, L_{TX} and L_{RX} are the inductances of the TX and RX coils respectively, and k is their coupling coefficient. For a given receiver design V_{RX} will be constrained by a minimum value (the smallest voltage at which the SAFF will correctly detect pulses), and therefore the optimisation target is to minimise I_{TX} for a given V_{RX} ; in other words, to maximise the function V_{RL}/V_{TX} (η_{dat}). By considering the ICL equivalent circuit model presented in Figure 4.1, an expression for this efficiency can be defined as:

$$\eta_{dat} = \frac{V_{RL}}{V_{TX}} = \frac{1}{(1 + j\omega R_2 C_2)} \cdot j\omega k\sqrt{L_1 L_2} \cdot \frac{1}{R_L (1 - \omega^2 L_1 C_1) + R_1 + j\omega (C_1 R_1 R_L + L_1)}$$

Broadly, considering this expression, it can be observed that ICL data efficiency (η_{dat}) is optimised when k is maximised provided the parasitic capacitances of each coil (C_1 , C_2) do not limit the bandwidth.

4.2.2 Inductive Coupling Power Links

Inductive coupling power links operate in a similar manner to data links, however typically transmit a sinusoidal signal to induce an alternating voltage in the receiver coil. The voltage in the recipient die is first rectified, and then stored in a capacitor acting as an energy buffer. Following this, voltage regulation is performed to regulate the power supply for the Integrated Circuit (IC).

This chapter focuses solely on the optimisation of the ICL channel, however optimising the channel quality will inevitably lead to improved overall power efficiency. Considering the equivalent circuit shown in Figure 4.1, an expression for the power transfer efficiency of the system (η_{pow}) can be derived, given by Eqn. 4.3 below.

$$\eta_{pow} = P_{out}/P_{in} \quad (4.3)$$

where P_{in} is given by:

$$P_{in} = \frac{R_i + j\omega L_2 + R_L / (1 + R_i j\omega C_2)}{\left[(1 + R_1 C_1 j\omega - \omega^2 L_1 C_1) / (R_1 + j\omega L_1) \right] [R_2 + j\omega L_2 + \gamma] + \omega^2 M^2} \quad (4.4)$$

and $\gamma = R_L / (1 + j\omega C_2 R_L)$, whilst P_{out} is given by:

$$P_{out} = \frac{\omega^2 M^2 R_L}{\left\{ \left[(1 + R_1 C_1 j\omega - \omega^2 L_1 C_1) / (R_1 + j\omega L_1) \right] [R_2 + j\omega L_2 + \gamma] + \omega^2 M^2 \right\}^2} \times \frac{1}{1 + j\omega C R_L} \quad (4.5)$$

This holds true, subject to the condition that each inductor is excited below its self-resonant frequency (f_{sr}):

$$f < f_{sr}, \text{ where } f_{sr} = \frac{1}{2\pi\sqrt{LC}}$$

Broadly, it could therefore be summarised that ICL power delivery efficiency (η_{pow}) is optimised when M is maximised, C_2 is minimised and f is as close as possible to the resonant frequency.

4.2.3 Objective Functions

In addition to the optimisation targets for power and data ICLs presented above, the full optimisation problem formulation must include several constraints. These are outlined below. The first constraint is that the inductors should be physically realisable without self-intersection. Mathematically, this imposes that:

$$D_i > 2 \left[\sum_{j=1}^n 2(w_{ij}) + \sum_{j=1}^{n-1} 2(s_{ij}) \right] \quad (4.6)$$

Additionally, the self-resonant frequency (f_{sr}) of each inductor in the link must be greater than the link's operating frequency. Whilst full-wave modelling is a reasonably accurate method of determining the performance of a given layout when fabricated in a specific technology, due to process variations and physical factors (such as uneven etching *etc.*) disparities will always exist between the simulated results and practical measurements of fabricated layouts. It is therefore sensible to include a marginal tolerance factor k_t . Therefore, the following constraint is added:

$$\frac{1}{2\pi\sqrt{LC}} < f(1 - k_t) \quad (4.7)$$

Bringing these details together, the optimisation problem formulation for inductive coupling *power* links can be expressed as:

$$\begin{aligned}
& \max && \eta_{pow} \\
& \text{subject to} && w_{ijk} > w_{min}, s_{ijk} > s_{min} \forall ijk, \\
& && D_i > 2 \left[\sum_{j=1}^n 2(w_{i,j}) + \sum_{j=1}^{n-1} 2(s_{i,j}) \right], \\
& && 1/2\pi\sqrt{L_i C_i} < f(1 - k_t) \\
& \text{where} && n_1, n_2 \in \mathbb{Z}+ \\
& && \text{and } w_{ijk}, s_{ijk} \forall ijk \in \mathbb{R}+
\end{aligned} \tag{4.8}$$

and the optimisation problem formation for inductive coupling *data* links can be expressed as:

$$\begin{aligned}
& \max && \eta_{dat} \\
& \text{subject to} && w_{ijk} > w_{min}, s_{ijk} > s_{min} \forall ijk, \\
& && D_i > 2 \left[\sum_{j=1}^n 2(w_{i,j}) + \sum_{j=1}^{n-1} 2(s_{i,j}) \right], \\
& && 1/2\pi\sqrt{L_i C_i} < f(1 - k_t) \\
& \text{where} && n_1, n_2 \in \mathbb{Z}+ \\
& && \text{and } w_{ijk}, s_{ijk} \forall ijk \in \mathbb{R}+
\end{aligned} \tag{4.9}$$

4.2.4 Planar Spiral Inductors

Having established the optimisation targets for both *power* and *data* ICLs, consideration should be given to the specific physical inductor layouts that maximise these expressions. Due to the requirement of achieving EM coupling between the vertically stacked inductors, layouts for ICL inductors should clearly be planar, however a plethora of planar inductor patterns and shapes exist. The following subsections review a range of these topologies, evaluating the performance of each, for use in ICL-based 3D-ICs.

Square vs. Octagon vs. Circle

Three of the most common planar inductor shapes used in VLSI are square, octagon and circle. As the axial length of circular spirals is much less than square spirals of the same area, it is widely reported that circular planar inductors offer higher Q-factors compared to their square counterparts (due to their reduced track length, and hence decreased resistance) [161]. However, due to their efficient area usage, square inductors can offer higher inductance per

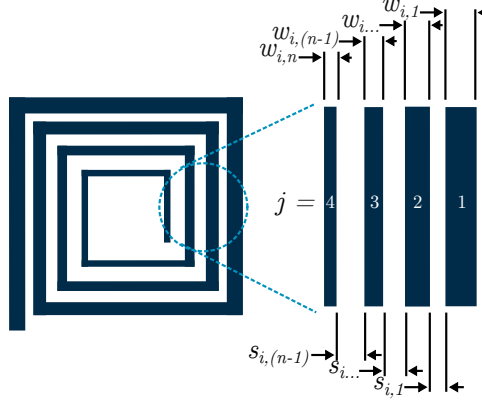


Figure 4.3: Illustration of coil layout with graduated spacing ($\chi_s = 1$, and $\chi_w = 0.4$).

unit area [162, 163]. As finding optimised inductor layouts for both *power* and *data* ICLs requires maximising L and minimising C for a specific area constraint, square inductors should theoretically, therefore, outperform their circular and octagonal counterparts. Based on this assumption, the COIL-3D tool presented in Section 4.3 will consider inductors of this shape. Empirical evidence to support this assumption is provided in Section 4.4.1 where results comparing square, octagonal and circular inductors for both power delivery and data transmission are presented.

Uniform vs. Non-Uniform

All reported previous works investigating inductive coupling based 3D-ICs use inductor layouts where the width and spacing of each turn in the coil is uniform [114, 135, 150]. This chapter explores the possibility of enhancing the efficiency of ICLs using graduated (or non-uniform) width and spacing parameters. In non-uniform planar spiral inductors, the width and/or spacing between each turn of the coil is different to the last.

To investigate non-uniform coil layouts, two linear *graduation coefficients* χ_w (for width graduation) and χ_s (for spacing graduation) are introduced. These graduation coefficients describe the linear scaling of track width and track spacing between each turn of the coil, and are calculated by:

$$\chi_{wi} = \frac{w_{in} - w_{i1}}{n} \text{ and } \chi_{si} = \frac{s_{i(n-1)} - s_{i1}}{n} \quad (4.10)$$

An illustration of a non-uniform coil, and its graduation coefficients is shown in Figure 4.3. The coil in this figure has parameters $\chi_s = 1$, and $\chi_w = 0.4$.

To meet the requirements outlined in Section 4.2, it is necessary to maximise the inductance L of a coil, whilst minimising parasitic capacitance C and resistance R . When considering non-uniform inductors, the variation in both width and spacing between turns can be

carefully exploited to meet these requirements. Examining the simple spiral inductance equations presented by Mohan *et al.* (shown below in Eqn. 4.11) [162], it can be observed that the inductance will increase as function of D and d (provided that the other parameters remain constant).

$$L = \frac{1.27\mu n^2 (D + d)}{4} \left[\ln\left(\frac{2.07}{\phi}\right) + 0.18\phi + 0.13\phi^2 \right] \quad (4.11)$$

where ϕ is the fill factor given by $\phi = (D - d)/(D + d)$.

To decrease coil resistance, and hence improve the Q-factor, tracks should be made as wide as possible. Widening the tracks however will decrease d and hence be detrimental to the coil's inductance. As the outer turns are longer than the inner turns, however, it is sensible to increase their width, whilst decreasing the width of the inner turns, maintain a constant D and d . Conversely, a similar technique can be applied to the coil spacing. As the influence of magnetically induced losses is much more significant within the inner turns of the spiral, it is sensible to increase the spacing toward the centre of the inductor [164]. Again, empirical evidence to support these claims is provided later in Section 4.4.1, where results illustrating the performance benefits that can be achieved using non-uniform inductor layouts are presented.

4.3 ICL Layout Optimisation (COIL-3D)

Having established the optimisation targets for power and data ICLs, in addition to the best performing inductor topologies, this section presents *a CAD-tool for Optimisation of Inductive coupling Links for 3D-ICs* (COIL-3D). COIL-3D combines four components in order to quickly and accurately determine best performing inductor layouts for ICLs, in addition to generating associated electrical models for simulation. These are:

- A comprehensive scalable inductor model for accurately approximating the performance of multi-turn non-uniform inductors (Section 4.3.1).
- A set of mathematical expressions for quickly and accurately determining the scalable model parameters (Section 4.3.2).
- A high-speed optimisation flow for identifying best-performing layouts (Section 4.3.3).
- An efficient software implementation of the above two elements which integrates with existing CAD flows (Section 4.3.3).

These four contributions are elaborated in the following corresponding sections.

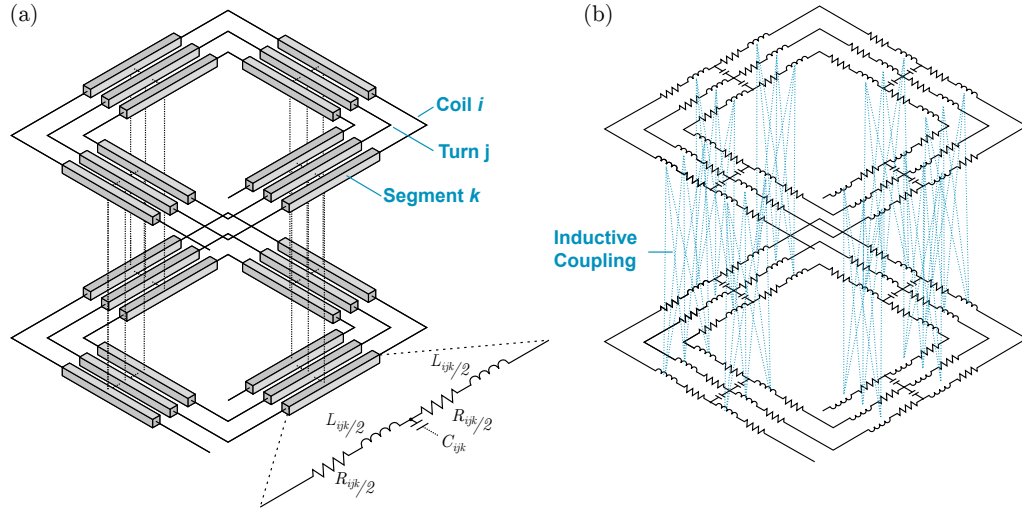


Figure 4.4: Illustration of the segmented scalable spiral inductor model presented in this chapter with 2 stacked inductors, including: (a) the overall segmented coil concept, and (b) the full equivalent circuit model.

4.3.1 Scalable Inductor Model

To quickly and accurately evaluate these parameters, the use of a scalable inductor model (based upon that presented in [164]) is proposed in this chapter, where each turn is considered as a separate segment. Using this segmented approach, the principal of superposition may then be applied in order to distil the model into its simplified lumped equivalent (shown in Figure 4.1). Figure 4.4 illustrates this concept more clearly. Here, two monolithic square spiral inductors are stacked vertically, where each turn of every coil is a single segment which exhibits resistance and inductance whilst sharing capacitance and mutual inductance with other segments. Considering the inductor in this way facilitates more accurate evaluation than the use of single expressions and allows for accurate evaluation of the non-uniform inductors proposed in this thesis.

4.3.2 Parameter Evaluation

To determine *optimised* coil layouts for power-delivery or data-transmission, it is important to establish a method of quickly and accurately evaluating the scalable model parameters for a given layout. Using the expressions for η_{dat} and η_{pow} from Section 4.2, this evaluation can simply be performed with knowledge of each segment's inductance (L) and resistance (R), in addition to the capacitance (C) and mutual inductance (M) between segments. To allow for optimisation in a reasonable time, a set of strictly solvable expressions for evaluating R , L , M and C are presented in the following sub-sections.

Coil Segment Resistance ($R_{i,k,k}$)

Other works propose a variety of methods for estimating the resistance of rectangular conductors, however the most commonly used model is the resistivity equation incorporating high-frequency conduction loss [134, 165]. Whilst this provides a reasonable approximation, when considering micron scale coils (used in 3D-IC) the yielded values are typically too low. This is due to the proximity effect; close inter-turn proximity drawing electrons to the edges of traces, hence increasing the apparent resistance by a factor k_p known as the *proximity factor*. In depth work, deriving differential equations for calculating k_p is available [166], however these expressions are not strictly solvable, making evaluation in software very computationally expensive. As such, in COIL-3D, values of k_p are empirically pre-determined and stored in a lookup table for use at runtime. Using these values, the resistance of each coil segment (R_{ijk}) is determined by Eqn. 4.12.

$$R_{ijk} = k_p \left(s_{i(j-1)k} \right) \frac{1}{2w_{ijk}t} \cdot \sqrt{\frac{\pi f \mu}{\sigma}} \quad (4.12)$$

Using the principal of superposition, the total resistance of each coil is the linear (series) summation of each of these line segments.

Coil Segment Self-Inductance ($L_{i,k,k}$)

In addition to the resistance of each line segment, it is also necessary to calculate the self-inductance of each segment within the coil (L_{ijk}). For calculating L_{ijk} , the expression shown in Eqn. 4.13 [167] is used:

$$L_{ijk} = \frac{\gamma \mu_0}{\pi w_{ijk}^2} \left[3w_{ijk}^2 \ell_{i,j,k} \ln \left(\frac{\ell_{ijk} + \sqrt{\ell_{ijk}^2 + w_{ijk}^2}}{w_{ijk}} \right) - (\ell_{ijk}^2 + w_{ijk}^2)^{\frac{3}{2}} + \ell_{ijk}^3 + w_{ijk}^3 + 3w_{ijk}\ell_{ijk}^2 \cdot \ln \left(\frac{w_{ijk} + \sqrt{\ell_{ijk}^2 + w_{ijk}^2}}{\ell_{ijk}} \right) \right] \quad (4.13)$$

where γ is an empirically determined constant. Again, using the principle of superposition, the total coil inductance is the linear (series) summation of each of these line segments.

Coil Segment Capacitance ($C_{i,k,k}$)

For calculating the capacitance between segments, the expression shown in Equation 4.14 is used:

$$C_{ijk} = k_c \frac{\pi \epsilon_0 \epsilon_r \ell_{ijk}}{\ln(4[w_{ijk} + s_{ijk}]/w_{ijk})} \quad (4.14)$$

Here, as the number of spaces between segments is equal to $n - 1$, the total capacitance is the linear (parallel) summation of each of these capacitances from $i = 1$ to $i = n - 1$ (as the

capacitance across the centre is negligible if d is sufficiently large).

Mutual Inductance Between Coils

Finally, for calculating M , an expression can be derived from Maxwell's equation for the mutual inductance between two air-cored loops. If an assumption is made that the two communicating coils are perfectly vertically aligned, the mutual inductance between two loops over a distance X is given by:

$$M_{a,b,X} = \frac{2\mu_0}{\alpha} \sqrt{ab} \left[\left(1 - \frac{\alpha^2}{2}\right) K(\alpha) - E(\alpha) \right] \quad (4.15)$$

where a and b are the radii of the two loops and $\alpha = 2\sqrt{ab/[(a+b)^2 + X^2]}$. Here $K(\alpha)$ and $E(\alpha)$ are the complete elliptic integrals of the first and second kind respectively. As the structure of a planar spiral inductor is not a single loop, moreover a set of n concentric interconnected segments, the approximation is often made that the total mutual inductance is the cumulative summation of mutual inductance between each segment of the TX coil and every segment of the RX coil [168], as illustrated in Figure 4.4 (b), leading to Eqn. 4.16.

$$M_{tot} = g \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} M(a_i, b_j, X) \quad (4.16)$$

To make this approximation more accurate, several previous works [134, 165, 169] suggest the introduction of a correction factor, g , (as shown in Equation 4.16) which takes the value $g \approx 1.1$. Although practical validation found this model to be reasonably accurate for coils with fewer than 10 turns, when considering inductors with $n > 10$, the model accuracy deteriorates. This is because the assumption of equal coupling between each turn of every coil, introduces increasing levels of error as the total number of turns (n) increases. In COIL-3D, this degradation in coupling is incorporated by a scaling factor, $r_{i,j}$ corresponding to the Pythagorean distance between turns, normalised with respect to a pair in perfect vertical alignment, such that:

$$r_{i,j} = \frac{1}{X} \left([(i-j) \cdot (w_{ij} + s_{ij})]^2 + X^2 \right)^{1/2} \quad (4.17)$$

The expression for M_{tot} therefore becomes:

$$M_{tot} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\frac{1}{r_{i,j}} \right)^{k_f} M(a_i, b_j, X) \quad (4.18)$$

where k_f is an empirical constant.

4.3.3 Optimisation Approach

Having presented the optimisation objectives for both *power* and *data* ICLs in addition to the methodology for evaluating a given inductor layout, this section presents an optimisation flow for determining best performing ICL layouts to replace the manual evaluation and adjustment cycle adopted in previous works [170].

Applying exhaustive linear optimisation to the problems outlined in Equation 4.8 and Equation 4.9 results in an extremely high time complexity $O(n^8)^2$. To address this and speed up the search process, an additional parameter, the fill factor (ϕ) is introduced. Optimised inductor layouts typically have a fill factor around 0.4 [134] and, using the equations from Section 4.3.2, it was noted that the optimal fill-factor can be pre-predicted with high accuracy. Centring the search around a fixed pre-determined fill-factor avoids the extra computational overhead incurred whilst evaluating probabilistically non-optimal designs *e.g.* where $\phi = 0.9$. By adding this constraint, the solution space can be refined, and the time complexity reduced to $O(n^6)$.

To speed-up the algorithm whilst searching for inductive coupling *data* links, optimisation is further divided into two discrete stages: RX coil optimisation, and TX coil optimisation, as shown in Figure 4.5b. From the ICL transfer equation, it can be observed that η_{dat} will be maximised when L_2 is maximised, provided that the time constant R_2C_2 (discussed earlier) in the denominator of the first term is constrained. Therefore, in the COIL-3D algorithm, the RX coil is optimised first, to provide maximum L_2 within the imposed bandwidth constraints. Following this, the TX coil is optimised separately to maximise η_{dat} considering the mutual inductance M , based on the geometry of the previously optimised RX coil. Dividing the flow in this way (performing series optimisation of the RX coil followed by the TX coil) means that it is not necessary to trial every set of TX coil layout parameters in conjunction with every set of RX coil layout parameters, thereby reducing the time complexity of the search from $O(n^6)$ to $O(n^3)$ and providing significant runtime savings without compromising on accuracy.

To reduce the time complexity of the algorithm whilst searching for inductors for Wireless Power Transfer (WPT) applications, it can be noted that η_{pow} will be maximised when $D_1 = D_2$, $w_1 = w_2$, $s_1 = s_2$ and $n_1 = n_2$. Therefore, the number of optimisation parameters, and hence time complexity, can be halved to $O(n^2)$, again without compromising on accuracy. It was also found that the optimisation of inductor *uniformity* could be considered separately to its geometric parameters. As such, to reduce the overhead of the approach, χ_s and χ_w can be determined after the key layout variables (D , n , w and s).

Combining these improvements, Algorithm 1 demonstrates the operation of the COIL-3D optimiser. First, an optimal value of ϕ is determined using the efficiency equations from

² $O(n^8)$ since four parameters w , s , n and D are considered for two coils, TX and RX.

Algorithm 1: Outline of the optimisation flow used in COIL-3D.

```

Inputs :  $D_1, D_2, f, R_L, g, w_{\min}, s_{\min}, C$ 
Constraints:  $C_{i_{\max}}, R_{i_{\max}}, D_{\max}$ 
Outputs :  $w_1, s_1, n_1, \chi_{w1}, \chi_{s1}, w_2, s_2, n_2, \chi_{w2}, \chi_{s2}$ 
/* Determine Optimal Fill-Factors */
for  $\phi = 0; \phi < 1; \phi += g$  do
     $\eta = \text{Evaluate\_}\eta()$ ;
    if  $\eta > \eta_{\max}$  then
         $\eta_{\max} = \eta$ ;
         $\phi_{\text{opt}} = \phi$ ;
    end
end
/* RX Coil Layout Optimisation */
for  $n_2 = 1; n_2 < 4wD_2/(1 + \phi_{\text{opt}}); n_2++$  do
    for  $w_2 = w_{\min}; w_2 < D_2/2; w_2 = w_2 + g$  do
         $s_2 = D_2\phi_{\text{opt}}/[n_2(1 + \phi_{\text{opt}})] - w_2$ ;
        if  $L_2(D_2, w_2, s_2, n) > L_{2_{\max}}$  then
            if Meets Constraints then
                 $L_{2_{\max}} = L_2(D_2, w_2, s_2, n)$ ;
                 $w_{2_{\text{opt}}} = w; s_{2_{\text{opt}}} = s; n_{2_{\text{opt}}} = n; D_{2_{\max}} = D_{\max}$ ;
            end
        end
    end
end
/* TX Coil Layout Optimisation */
for  $D_1 = D_{\max}; D_1 > 0; D_1 -= g$  do
    for  $n_1 = 1; n_1 < 4wD_1/(1 + \phi_{\text{opt}}); n_1++$  do
        for  $w_1 = w_{\min}; w_1 < D_1/2; w_1 = w_1 + g$  do
             $s_1 = D_1\phi_{\text{opt}}/[n_1(1 + \phi_{\text{opt}})] - w_1$ ;
             $\eta = \text{Evaluate\_}\eta()$ ;
            if  $\eta > \eta_{\text{top}}$  then
                if Meets Constraints then
                     $\eta_{\text{top}} = \eta$ ;
                     $w_{1_{\text{opt}}} = w; s_{1_{\text{opt}}} = s; n_{1_{\text{opt}}} = n; D_{1_{\max}} = D_1$ ;
                end
            end
        end
    end
end
/* Determine  $\chi_w$  and  $\chi_s$  */
for  $\chi_w = 0; n\chi_w + w_{i1} > 0; \chi_w += g$  do
    for  $\chi_s = 0; n\chi_s + s_{i1} > 0; \chi_s += g$  do
         $\eta = \text{Evaluate\_}\eta()$ ;
        if  $\eta > \eta_{\text{top}}$  then
            if Meets Constraints then
                 $\eta_{\text{top}} = \eta$ ;
                 $\chi_{w_{\text{opt}}} = \chi_w; \chi_{s_{\text{opt}}} = \chi_s$ ;
            end
        end
    end
end
end

```

Sections 4.2.2 and 4.2.1. ϕ_{opt} is then used to refine the search space and, incorporating the simplifications outlined above, the COIL-3D optimiser exhaustively searches all parameters

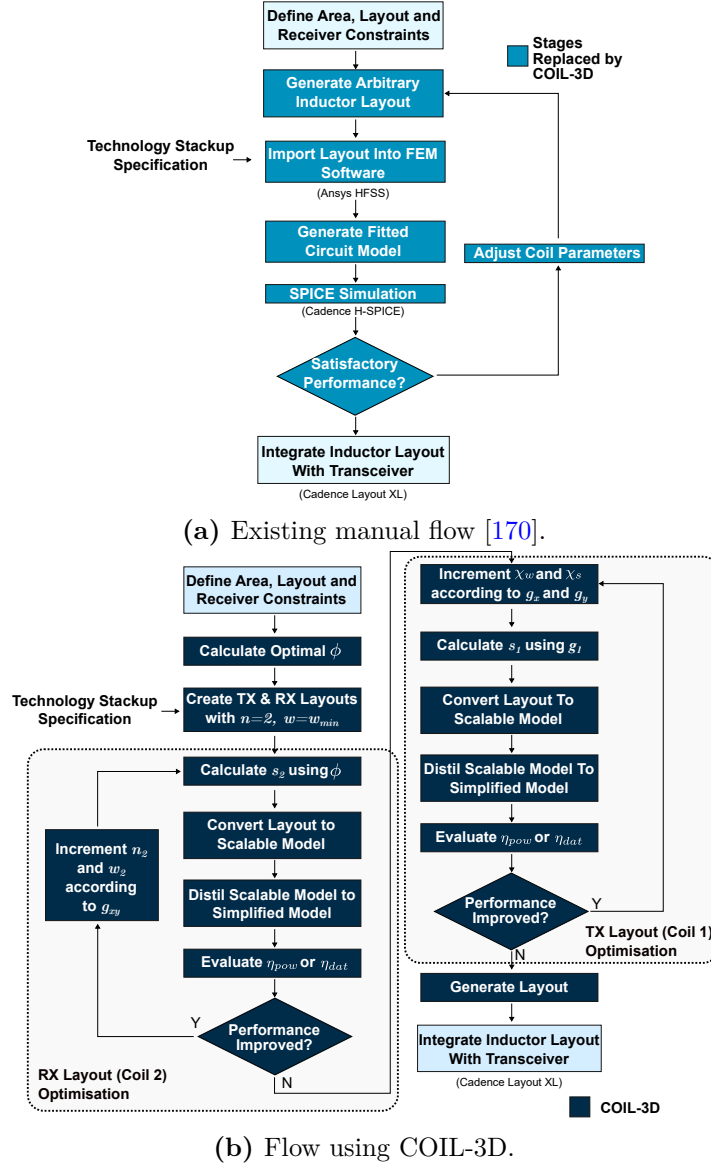


Figure 4.5: Flows for establishing inductor-pair layouts for ICL-based 3D-ICs.

to guarantee a globally optimal solution. The proposed approach is also summarised as a flow chart in Figure 4.5b for comparison with the existing manual approach, presented in Figure 4.5a.

4.3.4 Software Implementation

To minimise the runtime of COIL-3D, the solver and optimiser are both written in C, and Dynamic Programming (DP) is used where possible in the software implementation³. Using DP, large computational problems are broken down in to a collection of smaller solutions

³The C-based implementation of COIL-3D is an open-source tool that can be freely downloaded from <https://github.com/bjflg13/coil-3d>.

and/or decisions where the smaller solutions can be reused multiple times (to reduce the amount of overall computation required, and hence speed up execution) [171]; as a single coil is formed from the series superposition of many single turns (each containing 4 segments), DP can be applied here such that the solutions from previous layouts are stored and re-used. As the most compute intensive stage in the evaluation is calculating the elliptic integrals of α (consuming on average 34% of the entire runtime resource), re-use of a previous solution reduces the compute intensity of the algorithm, even when considering the look-up penalty incurred whilst locating previous useful solutions.

As an example, the mutual inductance between two coils (with layout parameters $n_1 = 4$, $n_2 = 5$, $s_1 = s_2 = 1 \mu\text{m}$, $w_1 = w_2 = 2 \mu\text{m}$, χ_w and $\chi_s = 0$) can be expressed as:

$$\begin{aligned}
 M_{tot} = & \sum_{i=4}^{n_1(=4)} \sum_{j=5}^{n_2(=5)} \left(\frac{1}{r_{i,j}} \right)^{k_f} M(a_i, b_j, X) \\
 & + \sum_{i=1}^4 \sum_{j=1}^4 \left(\frac{1}{r_{i,j}} \right)^{k_f} M(a_i, b_j, X) \\
 & + \sum_{i=1}^4 \left(\frac{1}{r_{i,j}} \right)^{k_f} M(a_i, b_5, X)
 \end{aligned} \tag{4.19}$$

where the second term is the mutual inductance between two coils ($n_1 = n_2 = 4$, $s_1 = s_2 = 1 \mu\text{m}$, $w_1 = w_2 = 2 \mu\text{m}$, χ_w and $\chi_s = 0$) which will have been calculated previously. This term can be re-used, reducing the computational overhead by 80 % in this case.

4.4 Experimental Results and Evaluation

In this section, the presented COIL-3D analysis and optimisation approaches are validated against existing commercial tools. For all simulations, the stack-up shown in Figure 4.6 was used, representative of two vertically stacked 65nm CMOS dies. Here, it is assumed that the top die has undergone die-level thinning to final a thickness of 70 μm (in line with realistic fabrication capabilities [24]) and that the dies are attached using a 20 μm thick epoxy adhesive. Ansys HFSS (an FEM tool [172]) was used as the evaluation benchmark for all tests.

Using the aforementioned experimental set-up, COIL-3D was compared against existing approaches ([134] and [169]) and FEM results. Experiments were performed to evaluate: (1) The effectiveness of the proposed non-uniform square inductor topology (Section 4.4.1), (2) The accuracy of the lumped equivalent model with respect to broadband fitted models (Section 4.4.2), (3) The accuracy of the semi-empirical expressions for modelling a particular coil layout (Section 4.4.3), (4) The effectiveness of the COIL-3D optimisation algorithm (Section 4.4.4) and (5) The runtime overheads of COIL-3D compared with existing approaches (Section 4.4.5).

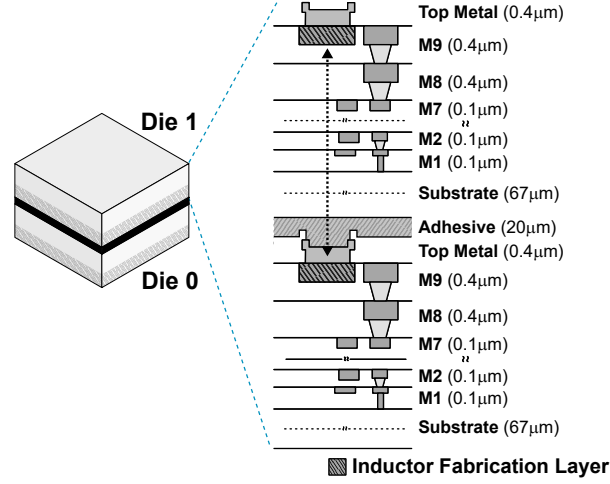


Figure 4.6: BEOL stack-up used for EM simulations (parameters representative of a typical 65nm CMOS process).

4.4.1 Inductor Topology Evaluation

Inductor Shape

In Section 4.2.4 analysis was presented suggesting that square inductors will outperform other inductor shapes (such as circular and hexagonal) for both *power* and *data* delivery between dies. Figure 4.7 presents results to support this assertion. Figures 4.7 (a) and (b) illustrate the *data* and *power* delivery efficiency (respectively) of square, octagonal and circular inductors with the same outer area ($200 \mu\text{m}$), track width ($3 \mu\text{m}$) and spacing ($1 \mu\text{m}$) as the number of turns, n varies. In both cases, it can be observed that square layouts provide better performance, due to their enhanced area utilisation efficiency⁴. Figures 4.7 (c) and (d) show similar results, this time for a range of frequencies. Again, the square topology provides the highest efficiency across all operation frequencies for a fixed area budget, upholding the earlier assumptions.

Inductor Uniformity

In addition to the inductor shape, Section 4.2.4 also presented theory to support the use of non-uniform inductors. To evaluate this theory, Figures 4.8 (a) and (b) show the simulated data and power transmission efficiency of an inductive coupling link with parameters $D = 200 \mu\text{m}$, $n = 4$, $w_1 = 3 \mu\text{m}$, $s_1 = 1 \mu\text{m}$ whilst varying χ_s and χ_w . In this case, results demonstrate that using a non-uniform layout can improve efficiency by 53.7% for the power ICL and 3.1% for the data ICL (compared with standard uniform layouts). Figure 4.8 (c) and (d) illustrate similar results, this time for a different inductor geometry ($D = 300 \mu\text{m}$, $n = 7$, $w_1 = 4 \mu\text{m}$, $s_1 = 0.5 \mu\text{m}$). Again, it can be observed that modifying χ_w and χ_s yields an efficiency improvement for both power ICLs (36.6%) and data ICLs (5.17%) when compared

⁴The same experiment was also repeated with a range of other values for s and w with concordant results.

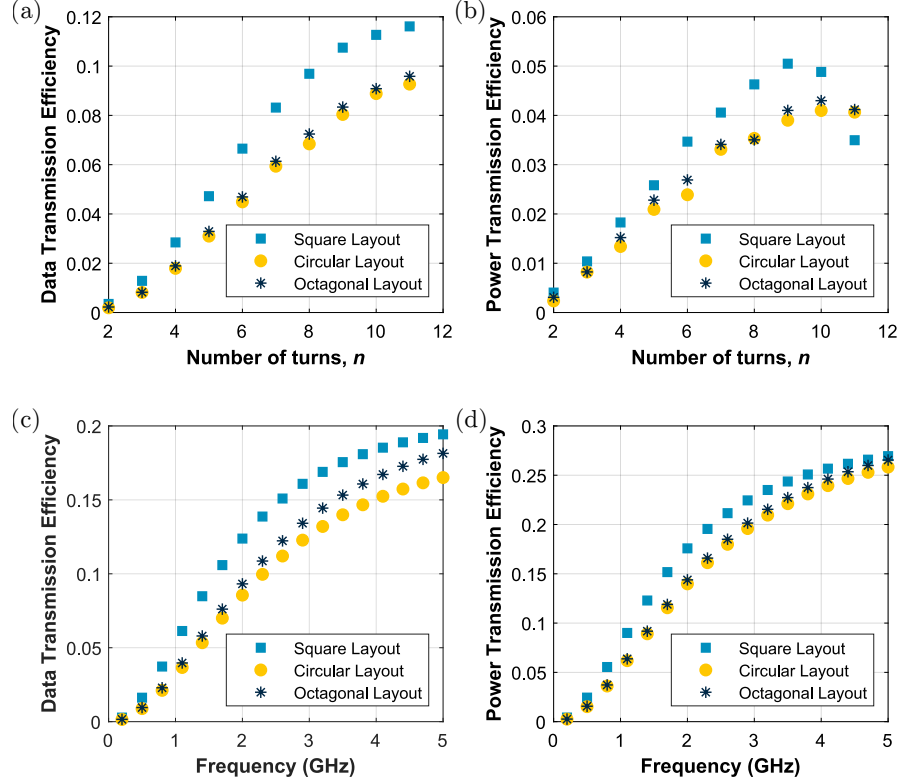


Figure 4.7: Variation of (a) η_{dat} and (b) η_{pow} with respect to n ($D = 200 \mu\text{m}$, $w=3 \mu\text{m}$ and $s = 1 \mu\text{m}$). Variation of (c) η_{dat} and (d) η_{pow} with respect to f ($D = 200 \mu\text{m}$, $w=3 \mu\text{m}$ and $s = 1 \mu\text{m}$).

with uniform layouts. These results uphold the theory that ICL efficiency can be improved by using non-uniform inductor layouts.

4.4.2 Lumped Model Accuracy Evaluation

In addition to validating the layout topology theory presented in Section 4.2, the accuracy of the simplified lumped equivalent electrical ICL channel model (presented in [123]) was also examined. Figure 4.9 shows a transient simulation of a system's performance whilst using both the simplified lumped equivalent model (with fitted R , L , M and C parameters), and a broadband SPICE model extracted using Ansys HFSS. It can be observed from the figure that the simulated V_{RX} pulse amplitude when using the simplified channel model is very similar to the simulated amplitude when using the broadband SPICE model (with a marginal average error of around 15 %), adequate for the purpose of optimisation (provided that the resulting optimal design is thoroughly validated). It is important to note that this marginal error is present in all approaches using the model from Figure 4.1 and is of comparable magnitude to the error tolerance of fabricated layouts.

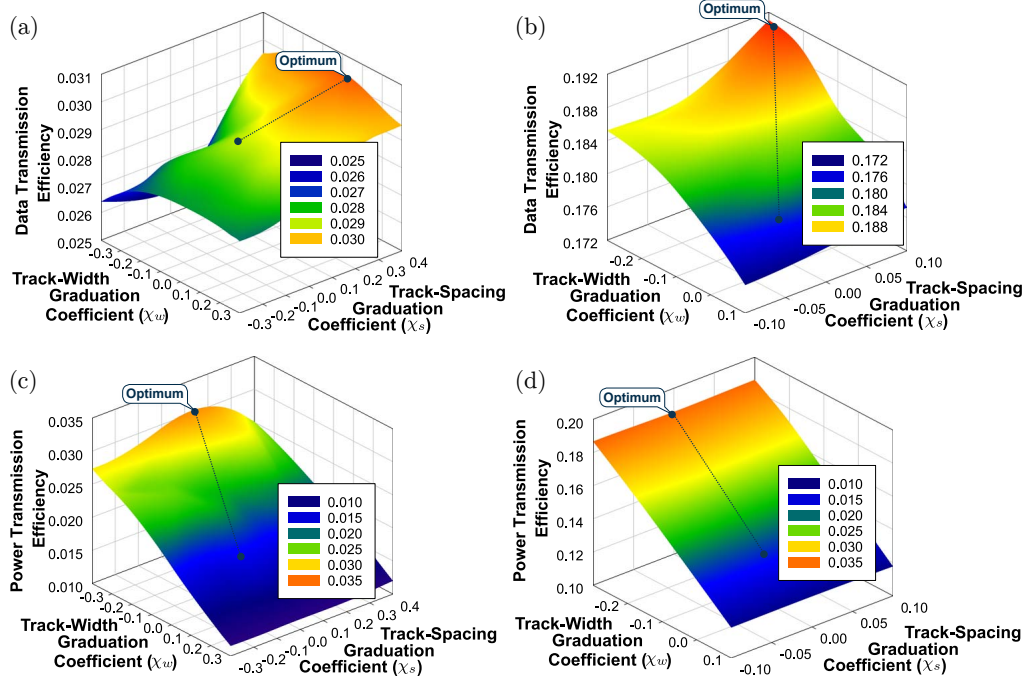


Figure 4.8: The effects of width graduation (χ_w) and spacing graduation (χ_s) on (a) data transmission efficiency and (b) power transmission efficiency when $S = 200 \mu\text{m}$ and $n = 4$. The effects of χ_w and χ_s on (c) data transmission efficiency and (d) power transmission efficiency when $D = 300 \mu\text{m}$ and $n = 7$.

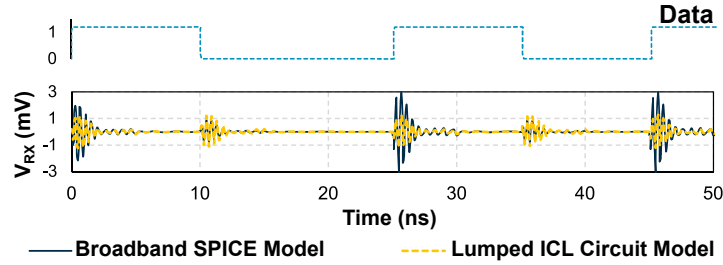


Figure 4.9: Transient simulation comparing the performance of broadband fitted SPICE channel model (generated by Ansys HFSS) and the simplified channel model (Shown in Figure 4.1 used in this work).








4.4.3 Empirical Expression Evaluation

Following this, the accuracy of the semi-empirical parameter expressions proposed in Section 4.3.2 (for R , L and C) was evaluated across a range of coil sizes. Table 4.2 shows the extraction accuracy of the R , L and C expressions for a range of seven randomly generated coils. The results also include the accuracy of approach [134] for comparison⁵.

As shown in Table 4.2, the inductance extraction accuracy of the expressions presented in this chapter is very high, exhibiting an average error of 2.5% across the generated inductor

⁵Here χ_s and χ_w are set to 0 to allow fair comparison with approach [134].

Table 4.2: Semi-empirical expression accuracy evaluation of COIL-3D for parameters L , R and C compared with existing approaches.

															
		I		II		III		IV		V		VI		VII	
Parameters	D (um)	150		250		200		250		300		150		400	
	w (um)	3.0		3.0		5.0		3.0		4.0		1.0		4.0	
	s (um)	1.0		1.0		1.0		0.5		1.0		1.0		0.8	
	n	5		3		3		4		5		5		4	
		Value	Error (%)	Value	Error (%)	Value	Error (%)	Value	Error (%)	Value	Error (%)	Value	Error (%)	Value	Error (%)
FEM	L (nH)	9.22	0.0	8.55	0.0	5.32	0.0	14.9	0.0	23.9	0.0	14.4	0.0	25.4	0.0
	R (Ω)	16.8	0.0	18.54	0.0	8.82	0.0	23.7	0.0	24.31	0.0	52.9	0.0	26.7	0.0
	C (fF)	34.2	0.0	38.0	0.0	45.6	0.0	68.2	0.0	112	0.0	76.6	0.0	86.8	0.0
COIL-3D	L (nH)	9.12	1.1	8.57	0.3	5.39	1.4	14.5	2.4	23.9	0.0	12.9	10	26.0	2.4
	R (Ω)	16.8	0.1	18.6	2.80	8.48	4.0	24.2	2.0	26.5	10	53.9	2.0	29.2	9.4
	C (fF)	45.1	32.0	41.0	7.90	33.7	26.1	66.3	2.78	98.7	11.9	38.8	49.3	102	17.6
Work [134]	L (nH)	6.82	26.0	6.36	25.6	3.99	25.0	10.7	28.1	17.2	28.0	9.43	34.5	18.1	28.7
	R (Ω)	15.6	7.31	17.6	5.07	7.82	11.3	23.4	1.06	25.4	4.48	52.3	1.13	28.4	6.36
	C (fF)	7.62	77.7	8.60	77.4	6.35	86.1	22.8	66.6	16.5	85.3	8.50	88.9	23.1	73.4

layouts. When compared to the approach outlined in [134] (which presents a set of expressions for evaluating the design of inductive coupling power links for biomedical implants), this represents an accuracy enhancement of 91% by using the expressions and scalable model presented in this chapter.

Table 4.2 also shows that the resistance extraction accuracy of the expressions presented in this chapter is very high, exhibiting an average error of only 4.3% across the examined inductor layouts. The expressions perform very well in most cases, however approach [134] performs marginally better than the proposed scalable model for inductors which have a high axial equivalent length. These slight errors are unlikely, however, to significantly affect the optimisation process, and COIL-3D still outperforms the expressions presented in [134] by an overall average of 17.5%.

The final rows of Table 4.2 document the capacitance extraction accuracy of each approach, showing an average error of 21.1% whilst calculating the capacitance of each of the seven coils using the COIL-3D expressions (compared to FEM). Whilst this may seem high, accurate capacitance evaluation is a challenging task and the expressions presented in this chapter outperform those in [134] by 3.7 \times .

Following this, Figure 4.10 illustrates the mutual inductance extraction accuracy with respect

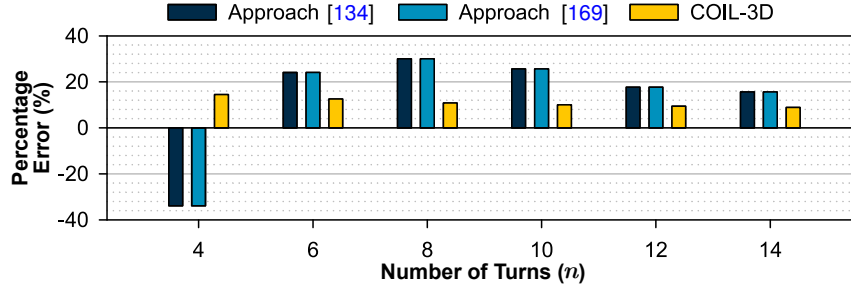


Figure 4.10: Mutual inductance extraction accuracy as n varies ($D=300\text{ }\mu\text{m}$, $w=1.5\text{ }\mu\text{m}$, $s=1\text{ }\mu\text{m}$ for both coils).

to n , using the semi-empirical expressions proposed in Section 4.3.2 (again, the accuracy of approaches [134] and [169] have been included for comparison). Here, it can be observed that the proposed mutual inductance model improves upon existing approaches (particularly in cases where $n > 10$), achieving an average error within 8.6% of FEM approaches. When combining these parameters (R , L , C and M) to evaluate efficiency (η), the overall average error was determined to be 7.8 %. This represents a 69% improvement compared to the expressions in [134].

4.4.4 Optimisation Flow Evaluation

The effectiveness of the COIL-3D optimisation algorithm was then explored and compared with both random trial-and-error approaches, and the optimisation flow outlined in [52] (the only other existing work to propose an optimisation scheme for ICL layouts). To examine the effectiveness of each approach, a layout was sought for a *data* ICL with a maximum area constraint of $200\text{ }\mu\text{m}$ assuming a grid resolution of $0.1\text{ }\mu\text{m}$. The same experimental set-up outlined in Section 4.4 was used for evaluation, and the results are shown in Figure 4.11.

Here, it can be observed that the COIL-3D optimisation flow performs the best out of the three optimisation approaches, finding an optimal solution after just 1500 iterations. The trial and error approach did reach the optimal point, however consumed approximately one billion iterations; six orders of magnitude slower than COIL-3D. Approach [52] terminated after approximately 1 million iterations at a sub-optimal solution. This is likely because mutual inductance is not considered in the optimisation flow [52] to speed up optimisation.

4.4.5 Overhead Evaluation

Finally, the execution overheads of COIL-3D were evaluated. Table 4.3 shows the average time taken to evaluate the efficiency of a single ICL using COIL-3D, approach [134] and FEM. Whilst COIL-3D is not the fastest of the three explored here, it is approximately 6 orders-of-magnitude faster than FEM, whilst maintaining high average accuracy (within 7.8%). Table 4.4 compares the total optimisation time for the COIL-3D optimiser with

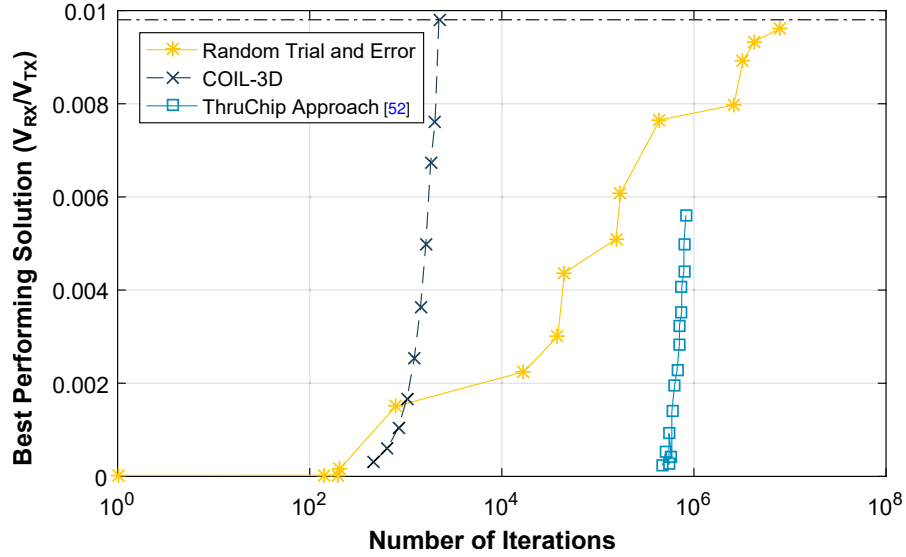


Figure 4.11: The COIL-3D optimisation approach efficiency (in terms of execution time) compared with existing approaches.

Table 4.3: Execution overheads of the proposed expressions when evaluating η_{pow} , compared with existing approaches.

Solver	Average Execution Time (per geometry)	Normalised Average Error (%)
FEM Solver	5,450 s	0%
Simplified Expressions [134]	0.008s s	22.3%
COIL-3D	0.081 s	7.8%

Table 4.4: Comparison of the total optimisation time when using COIL-3D and existing approaches.

Approach	ICL Type	Predicted [†] /Actual Execution Time
FEM solver + exhaustive linear search	Power or Data	10^{22} Years [†]
FEM solver with proposed refined search algorithm (proposed for COIL-3D)	Power or Data	518 Years [†]
Semi-empirical expressions (proposed for COIL-3D) with exhaustive linear search	Power or Data	10^{18} Years [†]
Iterative optimisation flow [134]	Power Only	124 Mins
ThruChip inductive coupling channel design optimisation flow [52]	Data Only	12.9 Mins
COIL-3D (semi-empirical solver with refined search algorithm)	Power or Data	47.1 Mins

various solver/optimiser combinations from prior art assuming a $0.1 \mu\text{m}$ grid and an area constraint of $300 \mu\text{m}$. Here, it can be observed that the COIL-3D tool arrives at optimised geometries faster than each of the alternative approaches, with the exception of approach [52] which considers the two inductors independently and hence suffers from low accuracy

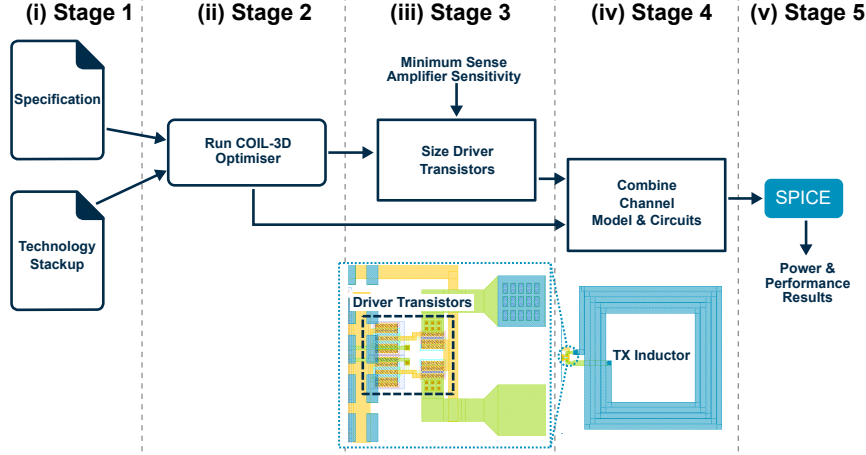


Figure 4.12: Example end-to-end (specification to GDS-II and power/performance statistics) ICL design flow when using COIL-3D.

(as discussed in Section 4.4.4).

4.5 COIL-3D Example Usage Application

To demonstrate the applications of COIL-3D, this final section presents a use-case example based upon [91], where an ICL is designed for 3D integration of digital CMOS and analogue BiCMOS dies for use in an implantable neuromodulator. The published work uses a uniform inductor layout with layout parameters $D = 200\ \mu\text{m}$, $n = 5$, $w = 9\ \mu\text{m}$ and $s = 0.72\ \mu\text{m}$. The design achieves a maximum bandwidth of 1.6GHz and requires TX current pulses with 0.77mA amplitude and 0.11ns duration to meet the minimum sense-amplifier sensitivity requirement in the receiver ($V_{RL,2} \geq 100\text{mV}$). This section provides a step-by-step overview of the ICL design process (illustrated in Figure 4.12) when using the COIL-3D tool, and the resulting power and performance benefits for this application.

The first stage of the process is to define the link specification, including the maximum coil footprint, the technology stack-up and the minimum voltage pulse threshold that can be successfully detected by the receiving sense amplifier. In this example, the maximum area is defined as $200\ \mu\text{m} \times 200\ \mu\text{m}$ (as per the benchmark work, [91]), the receiver sensitivity is defined as 100mV (again, as per the benchmark work, [91]) and the same technology stackup as that in [91] is adopted. Following this, the COIL-3D tool is run (Stage 2) to determine the best-performing layout within the specified dimensions. As discussed previously, the outputs of the tool include a SPICE model of the link and a physical GDS-II file containing the inductor layout. At this point, the SPICE model can optionally be checked using FEM by importing the GDS-II layout for subsequent analysis.

Next, in stage 3, the transmitter circuits are designed. As outlined Chapter 2 the transmit

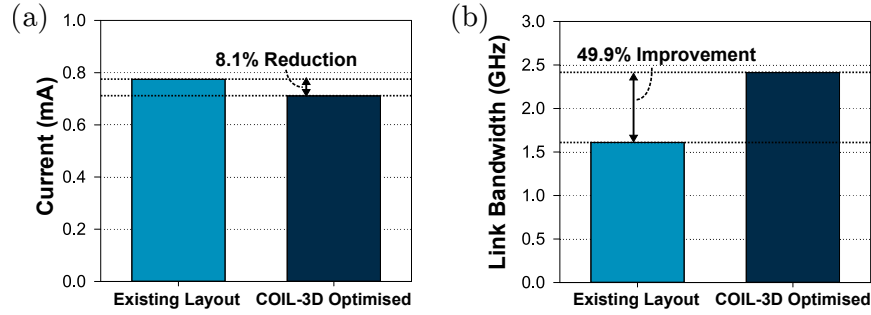


Figure 4.13: Performance (bandwidth and TX current at 1Gbps) of the inductor layout used in [91] compared with COIL-3D optimised solution.

current, I_{TX} , is controlled by the widths of the H-Bridge driver transistors. The COIL-3D tool generates a pair of inductor layouts that maximises the efficiency of the inductive coupling channel. The transmitter circuits can, therefore, be designed by gradually increasing the widths of the driver transistors until the required receiver sensitivity is met across the generated channel, resulting in the full system design (Stage 4). This full-system design can then be simulated in SPICE (Stage 5) using the inductive channel model generated by COIL-3D (or the FEM extracted model if used for checking in Stage 2) in order to obtain power and performance statistics.

Figure 4.13 illustrates the performance improvements yielded by using this design flow for the use-case example based upon [91] discussed earlier. As illustrated by the figure, for the same scenario, the maximum achievable link bandwidth is improved by 49.9% and, the required transmit pulse current is reduced by 8.1%, translating into a significant overall energy reduction through using the COIL-3D optimiser.

4.6 Summary

This chapter has focussed on design and optimisation of inductive coupling channels. In the first section, formulations for the inductor layout optimisation problems were presented for both *power* and *data* ICLs. Detailed analysis of multiple inductor topologies was performed concluding that square, non-uniform layouts can provide the highest efficiency in both settings. In the second section of the chapter, the COIL-3D tool was presented (available at <https://github.com/bjflg13/coil-3d>), which consists of: (1) a scalable comprehensive inductor model, (2) a fast mathematical solver for determining model parameters, (3) a high-speed optimisation flow, and (4) an efficient DP-based software implementation. Through a use-case example study, it is demonstrated that COIL-3D optimised inductor layouts can yield significant performance benefits when compared with manually optimised designs (achieving a 49.9 % bandwidth improvement and 8.1% power improvement in the presented example). In addition, it is demonstrated that the evaluation expressions presented in this chapter achieve an average accuracy within 7.8% of finite element tools, whilst consuming a

small fraction of the time ($1.5 \times 10^{-3}\%$), significantly reducing the design time associated with channel layout optimisation.

Chapter 5

Wireless Inter-Tier Clock Distribution Using Inductive Links

Having already covered the design of inductive *data* transceivers in Chapter 3 of this thesis, this chapter focusses on *clock* distribution within wirelessly stacked 3D-ICs. To implement coherent data transceivers (such as the low energy spike-latency encoding transceiver presented in Chapter 3), it is necessary to establish precise clock synchronisation between each of the dies within the 3D stack. As discussed in Chapter 2, however, achieving low-skew inter-die clock synchronisation in wireless 3D-ICs still poses a significant challenge; many of the previous works presenting wirelessly stacked 3D-ICs opt to perform clock distribution using external wire-bonded clock links [19, 103, 105] (often in conjunction with frequency control circuits such as Phase Locked Loops (PLLs)). Whilst this is an adequate solution (which circumvents the need for including expensive Through Silicon Vias (TSVs)), the addition of wire-bonds to each die in the stack undermines many of the benefits associated with *wireless* stacking.

Some other works have explored the use of ‘coupled resonators’ to distribute the clock wirelessly within the 3D stack [63]. The use of coupled resonators has been demonstrated to be very effective in minimising inter-tier clock skew, however the area of the channel inductors for such links is typically very large (to achieve LC resonance at the target clock frequency) [64]. Further to this, these links often require very precise frequency control circuits (as any slight frequency variations in the transmitted clock can cause the link to stop resonating [63]), the footprint and power consumption of which often contribute significantly to the overhead of the design¹ [63, 132].

To address these challenges, this chapter presents a low-overhead inductively-coupled wireless clock link for many-tier Wireless clock Synchronisation (WiSync) that operates across a wide range of frequencies (between 50MHz and 2.0GHz). Rather than using *resonant* inductive coupling, the proposed WiSync link operates in the flat portion of the frequency spectrum

¹A summary of background work related to wireless clock distribution in 3D-ICs is presented in the following section, 5.1.

to ensure high tolerance to variation without mandating the use of complex timing control circuits.

To reduce energy consumption (whilst still maintaining a high level of configurability across a wide range of operating frequencies), a novel dual-mode transmitter is proposed and experimentally validated using commercial EM and SPICE simulators. Simulation results demonstrate that the design can simultaneously broadcast the clock signal between five stacked silicon tiers, with less than 61ps of inter-layer skew. The WiSync transceiver is also implemented in a 65nm technology silicon test chip, and empirical measurements demonstrate an energy less than 16pJ/cycle (across a communication distance of 80 μm), whilst consuming only 0.0421mm² of silicon area.

The main contributions of this chapter can therefore be summarised as follows:

- Proposal of a low-power near-field wireless transceiver for inter-tier clock synchronisation in 3D-ICs. The proposed design (WiSync) enables low-latency, low-energy clock distribution between several tiers of a stacked 3D-IC (simulated results indicate a clock skew of less than 61ps across five stacked 65nm CMOS tiers).
- Experimental validation of the proposed transceiver using commercial electromagnetic and circuit simulators. The presented WiSync design operates at frequencies up to 2.0GHz whilst only consuming 0.0421mm² of silicon area (making it the smallest ever reported wireless inductive clock link).
- Silicon validation of the proposed WiSync transceiver demonstrating robust operation (cycle error-rate $< 10^{-13}$ across an 80 μm communication distance) whilst consuming, on average, 15.9pJ per clock cycle.
- A practical study analysing the effects of die-to-die stacking assembly misalignment (between 10 μm and 50 μm) on the transceiver's performance, demonstrating that the presented design can tolerate up to $\pm 10 \mu\text{m}$ of stacking misalignment (typical of that which can be achieved in existing low-cost IC packaging flows) with only a marginal ($< 10\%$) TX power increase.

The remainder of the chapter is organised as follows. Section 5.1 presents a brief overview of work related to wireless clock distribution in 3D-ICs, before Section 5.2 presents the design of the WiSync transceiver (including the dual-mode transmitter, Section 5.2.1, inductive channel, Section 5.2.2, and receiver, Section 5.2.3). Following this, Section 5.3 presents validation of the proposed transceiver design using post-layout SPICE and commercial EM simulators, before Section 5.4 presents experimental silicon validation of the proposed WiSync transceiver and evaluation of the link's tolerance to lateral die-to-die misalignment during the stacking process (Section 5.4.4). Finally, a concluding summary of the chapter is provided in Section 5.5.

5.1 Background and Related Work

In Chapter 3, and in many prior works presenting ICLs for wireless *data* communication, *clock* delivery is performed externally using wire-bonding [19, 103, 105, 123]. However, as discussed in Chapter 2, the addition of wire-bonds to each tier within the 3D stack undermines many of the benefits associated with *wireless* 3D integration, including the economic benefits of fully wireless assembly [24]. Further to this, the parasitic overheads of the pad drivers, and R, L, C parasitics of each bond-wire mean that the clock frequency and hence minimum inter-tier clock skew, that can be achieved using this approach are limited² [19]. As such, designs using this approach typically also require large Phase-Locked Loops (PLLs)/Frequency Locked Loops (FLLs)/Delay Locked Loops (DLLs) (or similar) frequency control circuits to achieve precise phase-locking between tiers.

To address this challenge, some prior art has proposed the use of coupled resonators to deliver the clock wirelessly between stacked dies [43, 63–65]. These works use LC tanks where the inductive (L) component is part of a coupled link and the resonant frequency corresponds to the clock frequency, f_{clk} [63]. This is a promising solution that allows the clock to be wirelessly transmitted between dies with very low jitter and skew (due to the natural harmonics of the link), however such resonant links have several drawbacks when considering their applicability in IoT applications.

Firstly, the link operating frequency is approximately inversely proportional to the channel inductor diameter (as the resonant frequency is proportional to $1/2\pi\sqrt{LC}$). As a result, channel inductors for coupled-resonators are typically very large, often requiring inductors with diameter $> 300\text{ }\mu\text{m}$, even at Gigahertz frequencies [64]³. For IoT edge-devices and IoT sensors which typically operate at frequencies in the order of 100’s of Mega Hertz or less (as discussed in Section 1.4, Chapter 1) this overhead would be even larger. This could be addressed by including tuning capacitors in parallel with the link to reduce the resonant frequency, however, this negatively impacts the link efficiency (particularly when performing a large frequency shift) and adds to the area footprint of the design.

Secondly, when using coupled resonators, any slight variations in the Transmitter (TX) clock frequency (for example due to Process Voltage Temperature (PVT), assembly variation, or due to natural jitter in the clock source) can result in significant performance deviations or, in the worst case, the whole system ceasing to operate [63]. To address this, previous works performing clock distribution using coupled-resonators typically also include several large timing/frequency control circuits. An example of this is can be seen in Figure 5.1, the coupled-resonator clock distribution architecture proposed by Take *et al.* in [63]. In their work, resonance is achieved using two phases: In the first phase, the frequency of the LC

²The typical maximum switching frequency of a wire-bonded IO pad is in the region of 400MHz.

³A full survey of relevant existing literature is presented in Chapter 2.

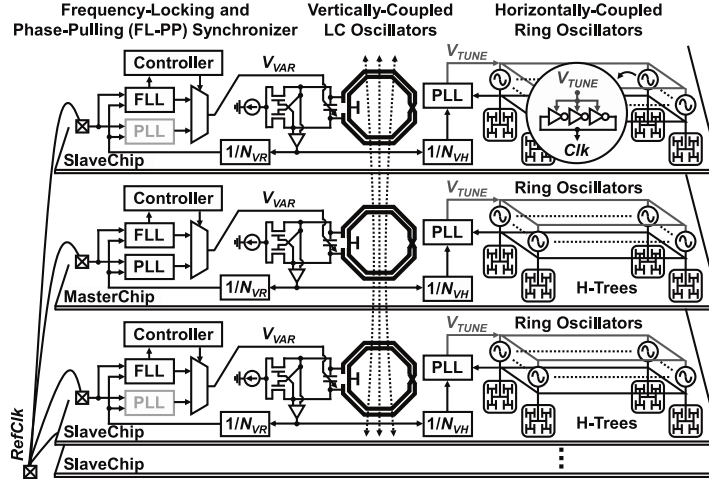


Figure 5.1: Illustration of wireless clock delivery architecture using coupled-resonators, proposed by Take *et al.* in [63] (reproduced from [63]).

oscillators is synchronised to the wire-bonded reference clock (labelled *RefClk* on Figure 5.1) on each chip, using the Frequency-Locked Loop (FLL). Following this, in the second phase, the system is switched to the PLL and the frequencies of the slave dies are pulled in to match the frequency of the master [63]. As shown on the figure, to achieve this operation, each die must contain three programmable locked-loop recovery circuits, in addition to a switched variable capacitor and a $1/N$ frequency divider (to reduce the frequency of the received clock from the channel resonant frequency, down to the operating frequency of the chip) [63]. These additional supporting circuits add a significant amount of complexity to the system, are power hungry and contribute a large silicon area overhead (approximately 0.035mm^2 for each PLL/FLL alone [63]).

One final drawback when using this approach is that it is highly layout-dependant, and hence cannot be easily ported between designs. For IoT devices, maintaining a short time-to-market is essential and it is often desirable to reuse the same Intellectual Property (IP) across multiple designs. Using a coupled-resonator clock link limits the ability to do this, as porting to other frequencies and/or technologies requires fundamental re-design of the entire transceiver.

5.2 WiSync Design and Implementation

To address these research challenges, this section outlines the design of WiSync, a low-overhead Inductive Coupling Link (ICL) for Wireless clock Synchronisation between tiers of a 3D IC. Figure 5.2 illustrates the concept of the WiSync transceiver. Here, instead of operating at resonance (like the prior art discussed above), the transceiver operates in the non-resonant portion of the frequency spectrum, to facilitate operation at Megahertz frequencies whilst also avoiding the overheads of precise timing control. As illustrated by

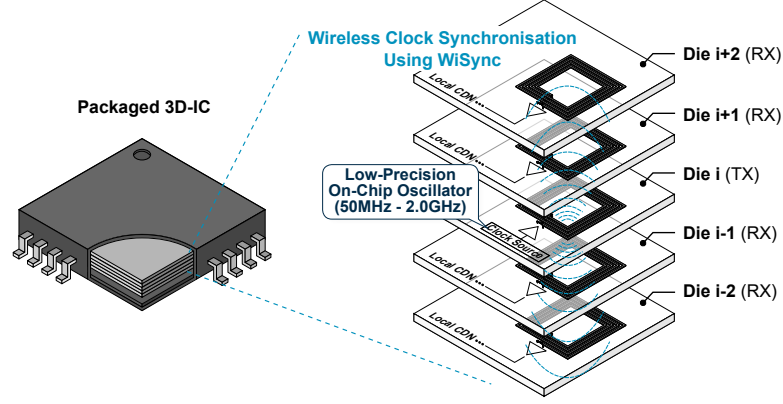


Figure 5.2: Conceptual illustration of wireless clock synchronisation between multiple stacked tiers within a 3D-IC using WiSync. Here, the clock is generated in die i which is in the centre of the stack, and broadcast in the $\pm z$ directions through near field electromagnetic coupling.

the figure, the WiSync link is also designed to broadcast the clock between several stacked tiers to achieve low skew clock distribution throughout a multi-tier 3D stack.

To achieve this, the proposed WiSync link consists of three main components: (1) the transmitter (including the encoding and driver circuits), (2) the inductive channel (consisting of the coupled system of TX/RX inductors included on each die), and (3) the receiver. The following sub-sections (Sections 5.2.1, 5.2.2 and 5.2.3) outline the design of each of these three components respectively.

5.2.1 Dual-Mode Transmitter Design

For compatibility with a wide range of present-day SoC designs, an operating frequency range of 50MHz - 2.0GHz was selected as a target design specification for the WiSync clock link. This poses a new set of design challenges when compared with prior implementations that are designed to operate at a single fixed frequency. As the \mathbf{H} -Field within the die stack is proportional to the magnetic flux linkage, which is, in-turn, proportional to dI_{TX}/dt , most high-frequency ICL transmitters use H-Bridge circuits, where one H-Bridge branch is the inverse of the other, to drive maximum current through the TX inductor. Using an inverting H-Bridge driver essentially ‘shorts’ the supply voltage through the inductor, thereby maximising dI_{TX}/dt for a given V_{DD} . This approach works well for high-frequency transmission (as the proportion of time spent in the *shorted* state, per cycle, is very low), however results in very poor energy efficiency for ICLs operating at lower clock frequencies⁴.

An alternative, more energy efficient approach for low-frequency signals, is to use level-based *pulse* encoding. Here, a short I_{TX} pulse, of fixed length, is transmitted with polarity

⁴As an example, when operating at 50MHz, it is inefficient to short the TX inductor for the full 20ns clock period.

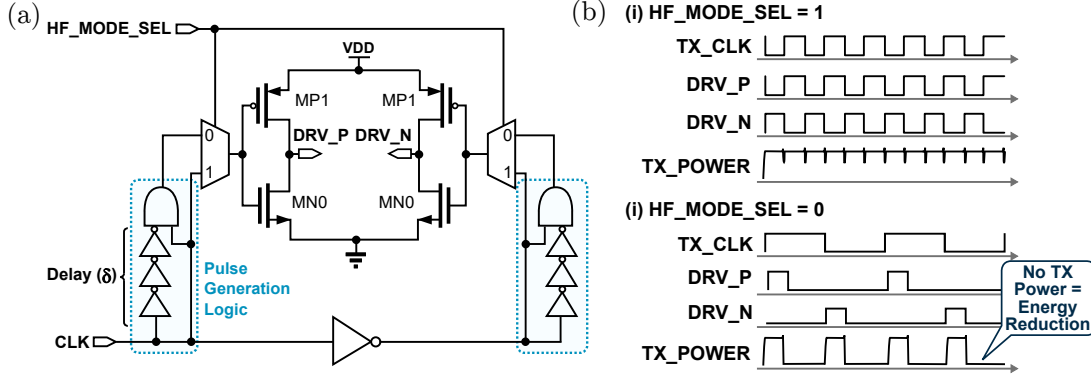


Figure 5.3: Design of the proposed WiSync clock transmitter, including (a) schematic diagram, and (b) operation of dual-mode clock transmitter.

corresponding to a particular logic level (*e.g.* a single positive pulse for a logic level of ‘1’ and a single negative pulse for a logic level of ‘0’). This has the benefit that the duration for which the supply is shorted through the TX inductor is deterministic and limited, rather than being dependent on the transmission frequency (like the inverted H-Bridge transmitter discussed above). To achieve this level-based encoding, it is typical to use a pulse-generator circuit where the signal is AND-ed with a delayed, inverted version of itself (allowing a short current pulse to flow with length corresponding to the delay). The logic required for this, however, introduces significant skew, hence limiting the maximum operating frequency of the link.

To meet the target operating frequency range, this chapter combines these two approaches, proposing the dual-mode clock transmitter shown in Figure 5.3 (a). For high-frequency operation, the pulse generation logic (highlighted in blue on Figure 5.3 (a)) is bypassed completely (using the HF_MODE_SEL signal) such that the transmitter behaves like an inverting H-Bridge. Conversely, for low frequency operation, the differential signal driven around the TX inductor (DRV_P, DRV_N) is pulse based, with a short pulse generated at each rising falling edge using the circuit shown in Figure 5.3 (b). This has the effect of limiting the current flow when the frequency is low (to reduce power consumption), whilst also supporting high-frequency operation.

The widths of transistors MN0 and MP0 control the amplitude of the I_{TX} current and hence, in this chapter, are sized to source as much current as possible (thereby maximising the communication distance) whilst still switching at the maximum 2.0GHz frequency at nominal voltage (1.2V in this case). As shown in the figure, each side of the H-Bridge driver is controlled by a two input Multiplexer (MUX) switched by the mode selector signal, HF_MODE_SEL. When HF_MODE_SEL is high, the TX inductor is driven directly by the CLK and \overline{CLK} signals. Conversely when HF_MODE_SEL is low, the TX inductor is driven by pulsed signals $CLK \cdot \delta(\overline{CLK})$ and $\overline{CLK} \cdot \delta(CLK)$, where δ is the time-domain delay added by the

delay-chain as shown on Figure 5.3 (a)⁵. the flow of current is limited, hence resulting in an overall energy reduction⁶.

5.2.2 Inductor Layout Selection

Another important consideration when designing the clock link to meet the required operating frequency target is the inductor layout. Therefore, the inductor layout used in this chapter is carefully selected using the design principles outlined Chapter 4 with some additional application-specific constraints, these are outlined below.

When considering the WiSync link design, the first constraint that must be satisfied by the inductor layout (for forming the inductive channel) is that the self-resonant frequency of the TX inductor (f_{sr}) must be much greater than the maximum operating frequency of the link, *i.e.* $f_{sr} \gg 2.0\text{GHz}$ (as 2.0GHz is the maximum target operating frequency in this case). This constraint is necessary for two reasons; firstly, it is important that the clock link's operating frequency does not exceed f_{sr} as this will result in a significant decrease in the link's coupling performance, as discussed in Chapter 4. Secondly, as the presented clock link is designed to operate across a wide frequency range, it is desirable to select a design with a high f_{sr} to avoid the exponential region that occurs as the frequency approaches resonance [173]. Avoiding this region means that $k(f)$ is as flat as possible, hence reducing the dynamic range requirements of the RX-side.

Another constraint is that the natural cut-off frequency (f_{cut}) should be less than the minimum desired operating frequency of the clock link. Although the link is designed to operate down to 50MHz, the power saving dual-mode operation (presented in the previous sub-sections) mean that the effective minimum operating frequency is equivalent to 1GHz⁷.

For practical implementation reasons (*e.g.* Design Rule Check (DRC) compliance in the TSMC 65nm technology used for fabrication), some additional *physical* constraints are also imposed upon the inductor geometry. These are $1.0\mu\text{m} < w < 12\mu\text{m}$, $\chi_w = 0$, $\chi_s = 0$ and $g = 1.0\mu\text{m}$. The maximum silicon area was also constrained to $170\mu\text{m} \times 170\mu\text{m}$ to minimise the area overhead. Considering all of these criteria, the COIL-3D optimisation flow (outlined in Chapter 4) was used to determine the best performing inductor geometries for the WiSync clock link, resulting in an inductor layout with parameters: diameter, $D = 170\mu\text{m}$, track width, $w = 5.0\mu\text{m}$, track spacing, $s = 3.0\mu\text{m}$, and number of turns, $n = 5$.

⁵Note the 3 inverter chain on Figure 5.3 (a) is purely for figurative illustration; for practical implementation much larger chains will be required (validation presented later in Section 5.3 uses a $29 \times \text{INV}$ delay chain).

⁶Empirical validation for this selection is provided later in Section 5.3.

⁷At frequencies below 1GHz, the clock rising and falling edges are mapped to individual *pulses* to conserve energy. When mapped to the frequency domain, these pulses correspond to a 1GHz response.

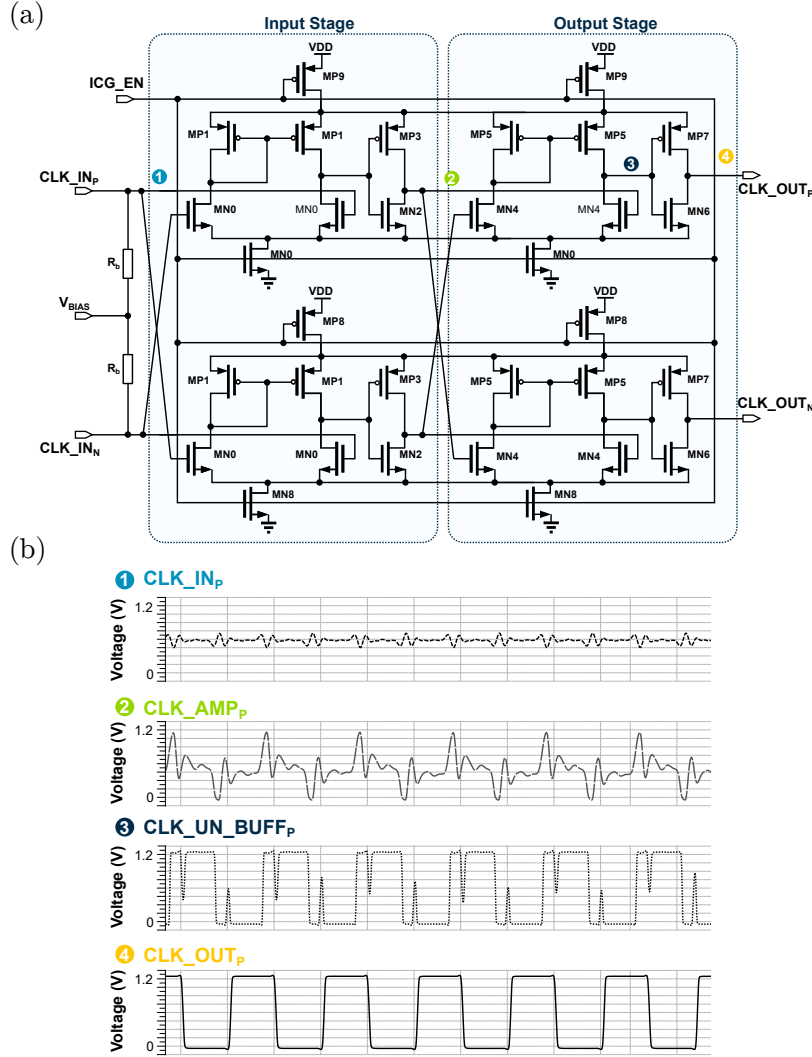


Figure 5.4: (a) Schematic diagram showing the two-stage differential WiSync clock RX amplifier, proposed in this chapter. (b) Waveforms illustrating the operation of this RX clock amplifier at the points labelled in (a).

5.2.3 Receiver Design

To detect the transmitted AC clock signal in the Receiver (RX) inductor and convert it into a logic-level clock signal, the cascaded differential amplifier shown in Figure 5.4 is used. In the first stage ('Input Stage'), the RX voltage signal (labelled ① on Figure 5.4) is amplified to a full-rail differential spiking voltage signal, as shown by trace ②. Regardless of whether the transmitter is in $\text{HF_MODE_SEL} = 0$ or $\text{HF_MODE_SEL} = 1$, the maximum dI/dt variation will be observed at data edges and so the same RX side topology can be used across both modes.

The second stage amplifies the spiking signal ② into a saturating signal where the decay-tail of the spike forms the high or low voltage level. Whilst this results in the clock single being

mostly recovered (as shown by the waveforms in ③), some glitches can still be observed where the second order harmonic has been amplified in Stage 1. These, however, are minor glitches, and so finally the signal is passed through a buffer stage, allowing the full clock signal to be recovered and providing sufficient drive strength. The clock output is then observed at CLK_OUT (④).

For heterogeneous systems comprised of multiple sensor/logic/memory dies (which form the focus of this thesis) it is often the case that some tiers operate transiently, and hence are inactive for certain periods of time (for example a radio die may only be required intermittently for hourly broadcasts). For this reason, the WiSync receiver also includes a power gating signal, ICG_EN. Whilst this signal is high, the power to the RX-side amplifier will be cut-off, meaning that the receiver will draw no current and the clock output will be disabled. For physical implementation of a stacked 3D system, this clock output can then be treated as a clock source in the RX die (with an appropriately balanced clock tree derived from CLK_OUT)⁸. As discussed, the effective operation of the receiver relies on large changes in current (dI_{RX}/dt) signalling clock *edges*. This means that the same RX-side amplifier can be used for correctly receiving the clock when using either transmission mode (pulse-based, when HF_MODE_SEL = 0, or continuous, when HF_MODE_SEL = 1).

5.3 Experimental Validation and Results

Bringing each of these three components together, this section presents experimental validation of the proposed WiSync link design using commercial EM and circuit simulators in 65nm CMOS technology (validation through a silicon prototype test-chip is also presented later in Section 5.4). Firstly, the area overheads of the proposed transceiver are evaluated in Section 5.3.1, followed by energy evaluation in Section 5.3.2. Section 5.3.3 then evaluates the Cycle Error Rate (CER) and communication distance of the proposed transceiver (in terms of number of stacked dies) and Section 5.3.4 presents the expected inter-tier clock skew results.

To simulate the performance of the EM channel, Ansys HFSS [172] was used, assuming a standard 65nm BEOL metal stack-up (shown in Figure 5.5 (c)). Two sets of experiments were performed, as illustrated by the EM stackup configurations shown in Figure 5.5 (a) and (b). Setup **A** (shown in Figure 5.5 (a)) assumes a stack height of two, and a die thickness of 80 μm , in accordance with low-cost die-thinning capabilities. Setup **B** (shown in Figure 5.5 (b)) assumes a die thickness of 30 μm , in-line with state-of-the-art competing works [64, 65] and, due to the shorter communication distance considered, explores a stack height of five dies. In both cases, it is assumed that dies are stacked using an epoxy adhesive with 10 μm thickness, taking the total communication distance between the TX inductor and top-most

⁸For chips with large area it may be necessary to combine several WiSync links (each with their own local clock tree to avoid *intra*-die skew).

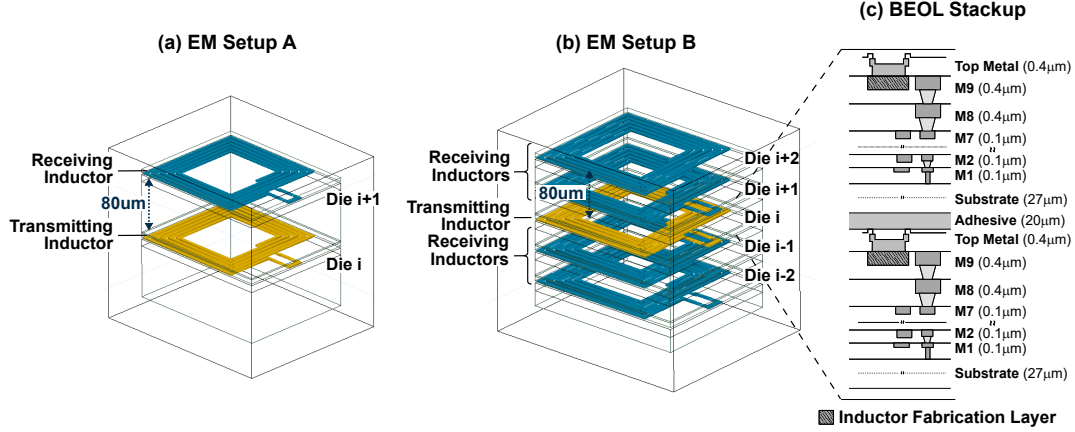


Figure 5.5: Illustration of the two stackups assumed for experiments in this Chapter. (a) Setup **A**, consisting of two 80 μm thick stacked dies, and (b) Setup **B**, consisting of five 30 μm thick stacked dies. BEOL parameters shown in (c) are the same for each setup and are representative of a standard 65nm CMOS process.

Table 5.1: Transistor sizings used for implementation of the WiSync RX amplifier.

Transistor Name (c.f. Figure 5.4)	MN0 & MN4	MP1 & MP5	MP3 & MP7	MN2 & MN6
Sized Width	1.3 μm	2.4 μm	0.8 μm	0.6 μm

RX inductor to 80 μm ($2 \times 30 \mu\text{m} + 2 \times 10 \mu\text{m}$, in Setup **B**).

As Electro-Magnetic (EM) coupling using square inductors is symmetrical about the xy -plane (square inductors form symmetrical **H**-Field lobes in the $+z$ and $-z$ directions) [139], the presented results for Setup **B** are grouped based upon their position within the stack *relative to the central transmitter* in die i . As an example, in the case shown in Figure 5.5 (b), the lowest die within the stack is labelled die $i - 2$, the one above it die $i - 1$, *etc.*. The electrical performance of the WiSync link was simulated using SPICE, based on the extracted netlist from the layout of the proposed transceiver. The channel performance was included in SPICE simulations using the S-Parameters generated by the HFSS EM modelling flow outlined above. The following sections evaluate the performance of the proposed WiSync clock link for both cases, **A** and **B**.

5.3.1 Device Sizing and Area Evaluation

Initially, the layout of the proposed WiSync transceiver was performed in 65nm CMOS technology. As discussed above, the drive transistors were sized to source the maximum current at nominal voltage (1.2V) whilst still achieving the 2.0GHz switching frequency target. The transistors in the RX-side amplifier were manually sized to provide the desired performance across the entire input frequency range and their sized values are shown in Table 5.1. The two bias resistors R_b (which bias the differential input signal such that the amplifying transistors MN0 are operating in the saturation region) were sized to 1.0 k Ω ,

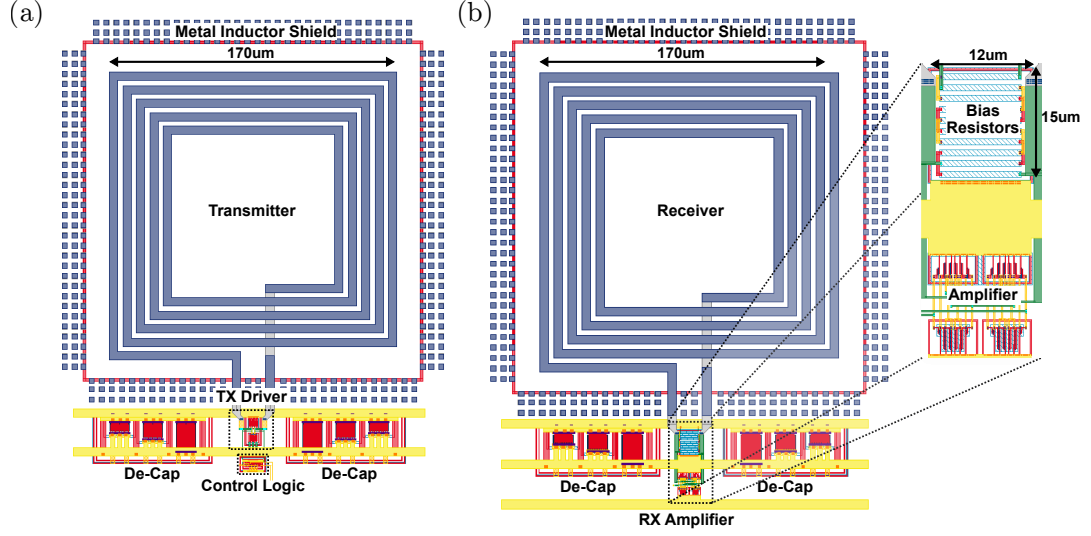


Figure 5.6: Layout and dimensions of (a) the WiSync transmitter, and (b) the WiSync receiver.

corresponding to a silicon area of $15\text{ }\mu\text{m} \times 12\text{ }\mu\text{m}$ in the 65nm CMOS technology considered.

Combining these components with the inductor layout generated in Section 5.2.2, Figure 5.6 shows the layout of the proposed WiSync transmitter (Figure 5.6 (a)) and receiver (Figure 5.6 (b)). The inductor was implemented on the highest, M9, layer and the area interposed by the ICL channel was left unused⁹. To comply with DRC regulations related to density, metal shields were placed around each of the channel inductors. These contain via stacks of all metal layers (M1-M9) to act as supports when transitioning from the low-density inductor area to the remainder of the chip (which is density filled). Using this approach, the inductors and WiSync transceiver could all be implemented whilst obeying the standard foundry technology rules, and hence can be included with no additional manufacturing processing overhead.

As highlighted on the ‘zoomed’ portion of the figure, careful consideration was given to the layout of the amplifier, to minimise the effects of process variation between the devices in each differential branch. A symmetrical interdigitated layout was adopted, where corresponding devices (for example $\text{MNO}_{\text{right}}$ and MNO_{left}) were placed in the same region of silicon to maximise fabricated threshold voltage (V_{th}) matching. This is important in differential amplifier design, as small V_{th} mismatches between branches are intensified when considering drain-source voltage, thereby introducing bias in the output. Larger channel lengths were also selected for gain-critical transistors (such as MNO and MN4) to minimise the overall influence of process/manufacturing variations. The use of a symmetrical layout (even for routing) also ensures that the RC load is matched, thereby maintaining equal sampling

⁹The interposed channel area was also defined as a non-P-Well non-N-Well region using the N_TN native layer.

latency in each arm. Aside from the gain critical devices in the receiver, the next most PVT-sensitive elements of the presented design are the $1.0\text{ k}\Omega$ bias resistors R_b (which have a nominal absolute fabrication accuracy of only $\pm 20\%$ in the target process technology). To ensure matching of these devices (and hence avoid a DC bias-point offset between each branch) an ABAB interdigitated poly-silicon layout was used, as shown on Figure 5.6. To validate the effectiveness of these layout techniques (and transistor size/width selections) the design was simulated using Monte Carlo SPICE for Slow-Slow (SS), Typical-Typical (TT) and Fast-Fast (FF) corners, between 0 and 85 degrees centigrade. The sign-off voltage margin was assumed to be 10 percent and was validated using Ansys RedHawk for dynamic IR drop analysis.

The overall bounding area of the WiSync link was found to be 0.0421 mm^2 , consisting of the channel inductor ($170\text{ }\mu\text{m} \times 170\text{ }\mu\text{m}$), the transmitter ($33\text{ }\mu\text{m} \times 14\text{ }\mu\text{m}$) and the receiver ($38\text{ }\mu\text{m} \times 16\text{ }\mu\text{m}$). This makes it the smallest ever reported inductive wireless clock link [43, 64, 65]. An itemised breakdown of these area overheads is provided in Table 5.2, and Table 5.3 (presented later in Section 5.4) provides a comparison with prior art.

5.3.2 Energy per Cycle Evaluation

Following this, the transient performance and energy consumption of the proposed design were evaluated using SPICE. As discussed in the previous sections, the dual-mode transmitter architecture is designed to minimise the energy consumption of the transmitter at low f_{clk} values, whilst still supporting high-frequency ($>1.0\text{ GHz}$) operation if required. To evaluate the effectiveness of this approach, the transceiver was simulated across a range of input frequencies in each operating mode (HF_MODE_SEL=1 and HF_MODE_SEL=0).

Figure 5.7 shows the results from these simulations for EM Setup A, plotting energy-per-cycle versus clock frequency for both modes. Here, the energy saving benefit of the proposed dual-mode scheme can be clearly observed. As the continuous encoding scheme has an energy proportional to $1/f_{\text{clk}}$, it is inefficient for use at lower frequencies, hence a significant energy saving is achieved by using pulse encoding ($> 8\times$ improvement at a frequency of 50 MHz). However, it is also important to note that the plot for HF_MODE_SEL = 0 stops at 1.0 GHz . This is because the delay when using the *pulse*-based clock encoding is too great to support operation at higher frequencies, and so the received clock becomes unreliable beyond 1.0 GHz . This motivates the *dual-mode* operation presented in this chapter.

The results indicate that, using this dual mode approach, the transceiver can operate robustly across the entire specified frequency range, with an average energy consumption of 9.8 pJ per clock cycle (assuming a single transmitter and receiver pair). When applying the WiSync transceiver to EM Setup B (where the clock is broadcast across five dies simultaneously), results forecast an average full-system energy consumption of 13.1 pJ/cycle (comprised of $4 \times 1.04\text{ pJ}$ receivers and $1 \times 8.98\text{ pJ/cycle}$ transmitter). Detailed energy evaluation across a

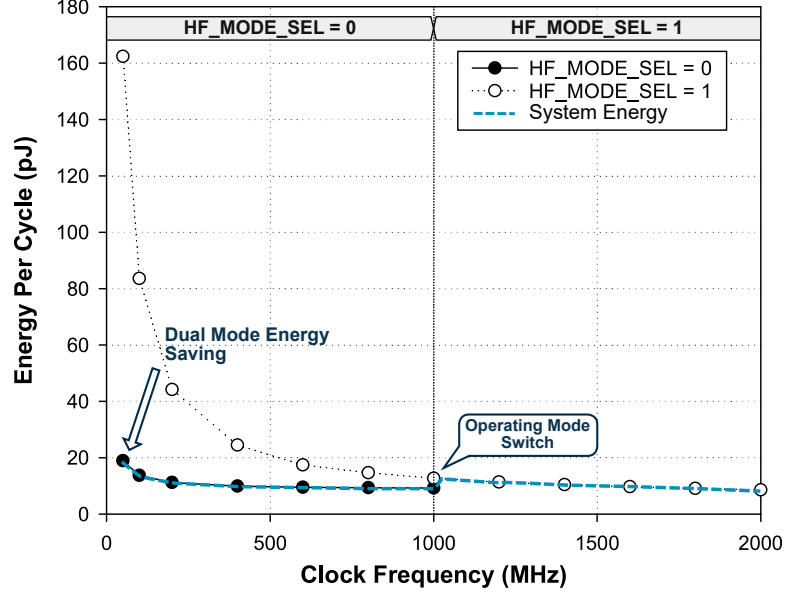


Figure 5.7: Plot showing how the energy per cycle varies with clock frequency for EM setup **A**. For high frequencies ($> 1.0\text{GHz}$), `HF_MODE_SEL` is enabled and hence the continuous transmission scheme is used. For low frequencies ($< 1.0\text{GHz}$), `HF_MODE_SEL` is disabled and hence the pulse-based transmission scheme is used. The same receiver design is used in both modes.

range of frequencies for setups **A** and **B** is also presented later in Table 5.2.

5.3.3 Cycle Error Rate (CER) and Maximum Stack Height Evaluation

The Cycle Error Rate (CER) of the proposed WiSync transceiver was then simulated by capturing the ratio of erroneous RX clock cycles to the total number of TX clock cycles, in each of the stacked dies ($i \pm 1$, $i \pm 2$ etc.). For evaluation of the CER, erroneous clock cycles were classified as those containing glitches (*i.e.* containing more edge transitions than expected within the given time period) or missing transitions (*i.e.* containing fewer edge transitions than expected within the given time period). Figure 5.8 shows the results of these simulations for Setups **A** (where communicating point-to-point across an $80\text{ }\mu\text{m}$ communication distance) and **B** (where broadcasting across four $30\text{ }\mu\text{m}$ thick dies) using Monte Carlo SPICE simulations across 100,000 clock cycles in the presence of supply noise (up to $\pm 10\%$ V_{DD} injected in the transmitter and receiver supplies).

The results demonstrate that, as would be expected, the performance decreases in each subsequent die ($i + 1$, $i + 2$ etc.) moving away from the transmitter, as the coupling strength deteriorates. For the sizing of MNO and MP0 selected in this implementation (chosen to maximise the amplitude of I_{TX} whilst still meeting the 2.0GHz switching frequency required by the specification¹⁰) the results illustrate that the maximum number of stacked tiers

¹⁰For testing in this chapter MNO and MP0 were sized to $9.0\text{ }\mu\text{m}$ width.

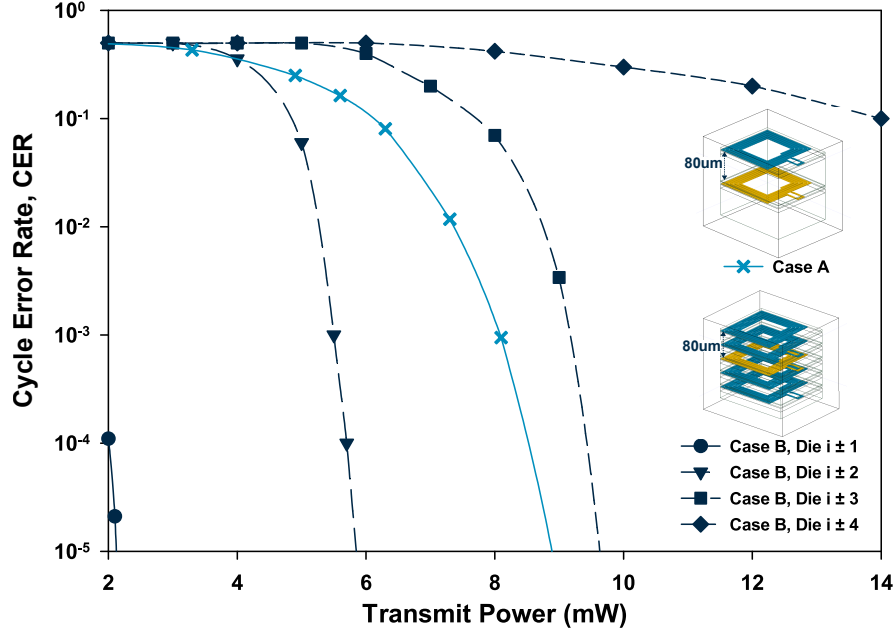


Figure 5.8: Simulated Cycle Error Rate (CER) vs. transmit power for EM Setups **A** and **B** (for Setup **B** the simulated CER is shown for each die in the stack). CER is calculated by the ratio of erroneous RX clock cycles (*e.g.* containing glitches or missing edge transitions) to the total number of TX clock cycles.

supported is five, with the CER becoming too large for robust operation beyond this height (this can be observed by the CER plot for die $i \pm 3$ which clearly would require a TX power $\gg 20\text{mW}$ to reduce the CER to an acceptable level). It is worth noting, however, that for die stacks with height greater than five stacked dies, it would be possible to implement a ‘repeater’ structure (such as that presented in [64]), where every third die contains a receiver and a transmitter to forward the clock.

The results in Figure 5.8 present CER versus *transmit* power consumption to allow for fair comparison between cases **A** and **B** (as case **B** has multiple receivers which will each contribute to the *overall* power consumption of the system). Although the communication distance in Setup **A** and die $i \pm 2$ in Setup **B** is the same ($80\mu\text{m}$), Figure 5.8 indicates that Setup **B** performs slightly worse, requiring a transmit power of approximately 9.7mW to achieve a $\text{CER} < 10^{-5}$ (whilst the same CER is reached in case **A** with a TX power of 8.8mW). This is due to the interference of the interposed channel inductor in Setup **B** (which is used for reception in die $i \pm 1$) reducing the flux linkage between die i and die $i+2$ and hence reducing the electromagnetic coupling. As shown, however, the two results are very close, indicating that using a single channel of $80\mu\text{m}$ can provide a good approximation of the performance of two stacked $40\mu\text{m}$ channel, provided the inductor layouts are congruent and aligned.

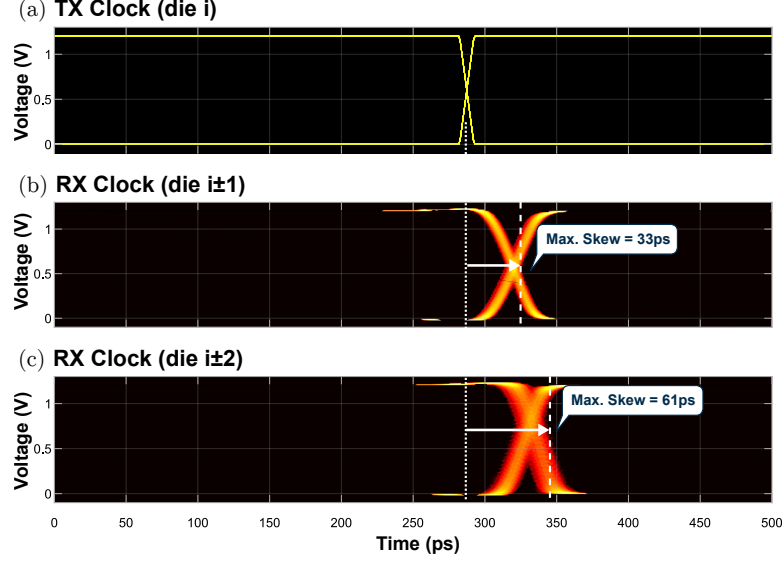


Figure 5.9: Rising/Falling edge clock skew across multiple stacked tiers when using WiSync. (a) shows the input to the WiSync Transmitter at Die i , (b) shows the output of the WiSync transceiver in die $i + 1$ (with a maximum delay of 33ps compared to the input), and (c) shows the output of the WiSync transceiver in die $i + 2$ (with a maximum delay of 61ps compared to the input).

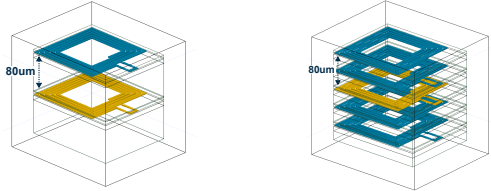
5.3.4 Clock Skew Evaluation

Following this, the inter-die clock skew when using the proposed WiSync transceiver was evaluated for Setup B¹¹. Figure 5.9 shows the simulated skew results, presenting transient simulation results of the rising clock edge as received in dies (a) i , (b) $i \pm 1$ and (c) $i \pm 2$ across 10^5 cycles. As would be expected, the clock is received first in die $i \pm 1$ (which is closer to the transmitter in die i). Compared with the rising/falling edge in the transmitting die, the maximum expected reception delay (or skew) in die $i \pm 1$ is 33ps, representing 3.3% when considering the measured operating frequency of 1.0GHz (or 6.6% when operating at the maximum 2.0GHz frequency). The maximum arrival time in dies $i \pm 2$ is 61ps, corresponding to a 6.1% clock skew at 1.0GHz.

Finally, Table 5.2 provides an overview of the simulation results presented in this section, including detailed area, power and performance statistics across a range of frequencies (50MHz, 500MHz, 1.0GHz and 2.0GHz). As shown in the table, the simulated results suggest that the WiSync design achieves its overall goal of enabling wide-band clock synchronisation across several stacked silicon tiers, operating with an average energy consumption of 11.8pJ/cycle across all the considered cases.

¹¹The maximum ‘skew’ here is defined as maximum time delay between the transmission of a rising/falling edge in the TX die, and the reception of the same rising/falling edge in the furthest RX die.

Table 5.2: Table summarising the overall simulated performance of the WiSync transceiver for EM simulation setups **A** and **B** (*c.f.* Figure 5.5).

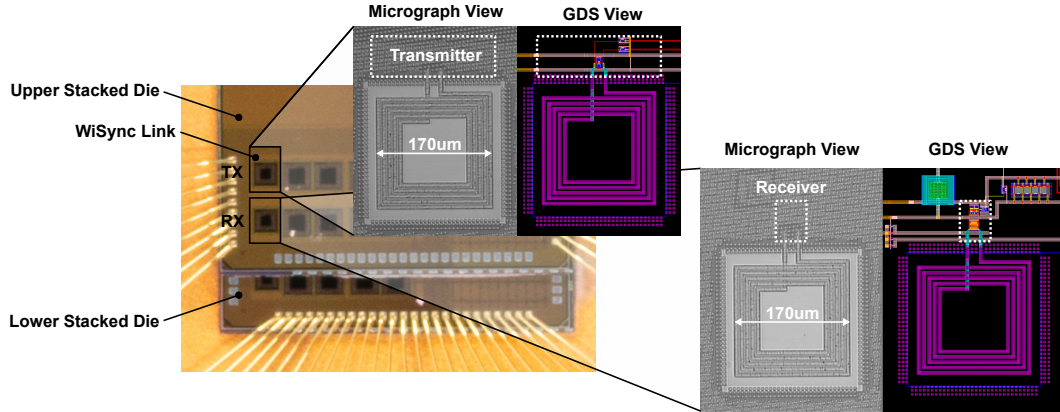


	Metric	Case A	Case B
	Channel Inductor Area	0.0289mm ²	
	Transmitter Circuits Area	462µm ²	
	Receiver Circuits Area	608µm ²	
	Total Bounding Area	0.0421mm ²	
50MHz	Transmitter Power	0.36mW	0.37mW
	Receiver Power	0.59mW	2.36mW (4 × 0.59mW)
	Total Power	0.96mW	2.73mW
	Energy Per Cycle	19.1pJ/cycle	54.6pJ/cycle
500MHz	Transmitter Power	4.23mW	4.50mW
	Receiver Power	0.67mW	2.68mW (4 × 0.67mW)
	Total Power	4.90mW	7.18mW
	Energy Per Cycle	9.8pJ/cycle	14.4pJ/cycle
1.0GHz	Transmitter Power	8.26mW	8.98mW
	Receiver Power	1.04mW	4.16mW (4 × 1.04mW)
	Total Power	9.30mW	13.1mW
	Energy Per Cycle	9.3pJ/cycle	13.1pJ/cycle
2.0GHz	Transmitter Power	16.03mW	17.44mW
	Receiver Power	1.59mW	6.36mW (4 × 1.59mW)
	Total Power	17.6mW	23.8mW
	Energy Per Cycle	8.8pJ/cycle	11.9pJ/cycle

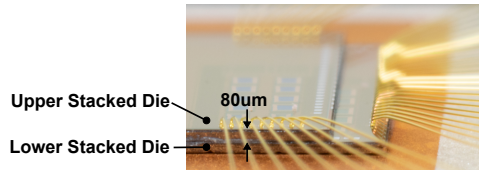
5.4 Test-Chip Validation

Following the successful evaluation of the proposed WiSync transceiver in simulation, the same design was also implemented in a 65nm CMOS technology silicon test-chip for more accurate power measurement and practical analysis. For the test-chip implementation, only two stacked tiers could be realised (hence limiting experiments to the EM setup **A**), however as discussed above in Section 5.3.4, the use of an 80 µm channel provides a good approximation of case **B** which also has a total maximum communication distance of 80 µm (2 × 30 µm chips + 2 × 10 µm glue). For physical realisation of the 3D stacked test-chip, the transmitter and receiver designs were included on the same design, 400 µm apart. Two identical dies (manufactured using the same mask-set) were then fabricated and stacked with a lateral offset of 400 µm such that the two designs (TX and RX) precisely align. Dies were attached using a non-conductive epoxy adhesive¹² and were wire-bonded for test and measurement purposes. Further details regarding the test-chip design and assembly are

¹²The adhesive used was QMI 538NB-1A1.5 [174].



(a) Top view showing a photograph of the two tier test-chip highlighting the WiSync transmitter and WiSync receiver (for which the GDS layout and detailed micrograph photo are shown).



(b) Side elevation showing the stacking and bonding arrangement used for testing.

Figure 5.10: Photographs showing the assembled 2 tier test chip used for evaluation in this section.

provided in Appendix D.

Figure 5.10a shows a photograph of the two tier stacked test chip used for evaluation, with the WiSync link highlighted¹³. The figure also shows ‘zoomed’ micrographs of the transceiver layout, highlighting the transmitter, receiver, and inductors for side-by-side comparison with the GDS-II. As can be observed from the figure (and as discussed above in Section 5.3.1), for practical evaluation, no circuits (or other metallic structures) are placed within the ICL channel area, however for area constrained designs, it has been demonstrated that it is possible to utilise this channel area with negligible performance impact when placing SRAM or standard digital place-and-route cells (a discussion of interposed channel area usage is presented in Chapter 7, Section 7.2.4) [147]. Figure 5.10b pictures the side elevation of the stacked chips, showing the lateral offset (which is intentional to ensure TX and RX channels align, as discussed earlier) and the 80 µm vertical communication distance.

The IC (shown in Figures 5.10a and 5.10b) was packaged in a Quad Flat Package (QFP) and was tested using the custom designed test Printed Circuit Board (PCB) shown in 5.11 (b). The test PCB contains connectors for each of the supplies within the design, on-board programmable regulators, an Mbed microcontroller for test configuration and a header to

¹³The other ICL channels on the chip perform wireless inter-tier data and power transfer, however the details of these links are not discussed in this chapter. Design and testing of these links is covered in Chapter 6.

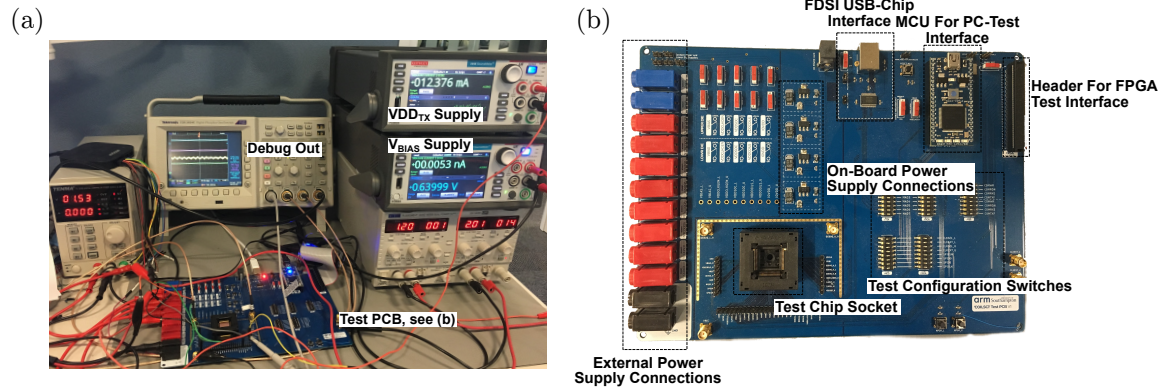


Figure 5.11: (a) Test set-up used for obtaining measurement results in this Chapter. (b) Test PCB used for evaluation in this section with key components labelled. The test-chip used for evaluation was packaged in a 100 pin QFP and placed in the test-chip socket labelled on the figure.

connect an Altera DE0 Field Programmable Gate Array (FPGA) to perform error-rate analysis. For the testing presented in the remainder of this chapter, the test set-up shown in Figure 5.11 (a) was used, where $V_{DD_{TX}}$ and V_{BIAS} are supplied externally using Keithley 2450 precision Source Meter Units (SMUs) for accurate power measurement. The clock output was measured at the I/O pin (equipped with configurable clock divider) and exported to both the FPGA and oscilloscope through a $50\ \Omega$ transmission line. Test and measurement results using this setup are outlined in the subsections below.

5.4.1 Parameter Tuning

Initially, the optimal configuration settings for the proposed WiSync link were determined, notably tuning the RX-side amplifier bias voltage (labelled V_{BIAS} on Figure 5.4). Figure 5.12 shows the results of this bias tuning process for a TX clock frequency of 1.0GHz. Here, the yellow trace (channel 1 in Figure 5.4) shows the input bias voltage which is being gradually increased over time, and the green trace (channel 2 in Figure 5.4) shows the clock signal observed at the output (with a $1/8$ clock divider enabled). The measurement results indicate that the presented RX amplifier design is tolerant of a wide range of input bias voltages, with the clock being correctly received across a 218mV bias voltage range. This makes it well suited for use in IoT devices which are often powered by batteries or energy harvesting sources, and hence exhibit high levels of supply variability (as discussed in Section 1.4, Chapter 1). For testing in the remainder of this chapter, the bias voltage at the centre of this range (corresponding to 610mV on the chip shown in Figure 5.12) was used.

5.4.2 Energy Measurement

Following this, the energy of the proposed WiSync clock transceiver was evaluated. Figure 5.13 revisits the simulated projections presented in Figure 5.13, this time comparing the

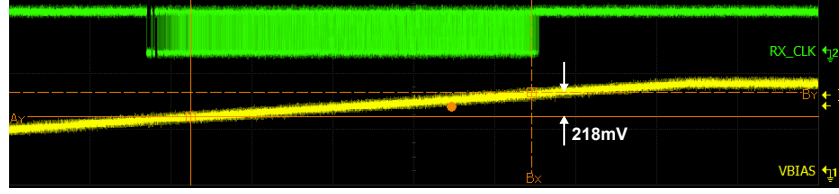


Figure 5.12: Oscilloscope capture showing the measured tolerance of the WiSync clock link to variations in bias voltage. Here, the bias voltage (shown in yellow) is being gradually ramped and the RX clock output is shown in green. The marked measurements show that the transceiver will tolerate up to 218mV of bias voltage variation.

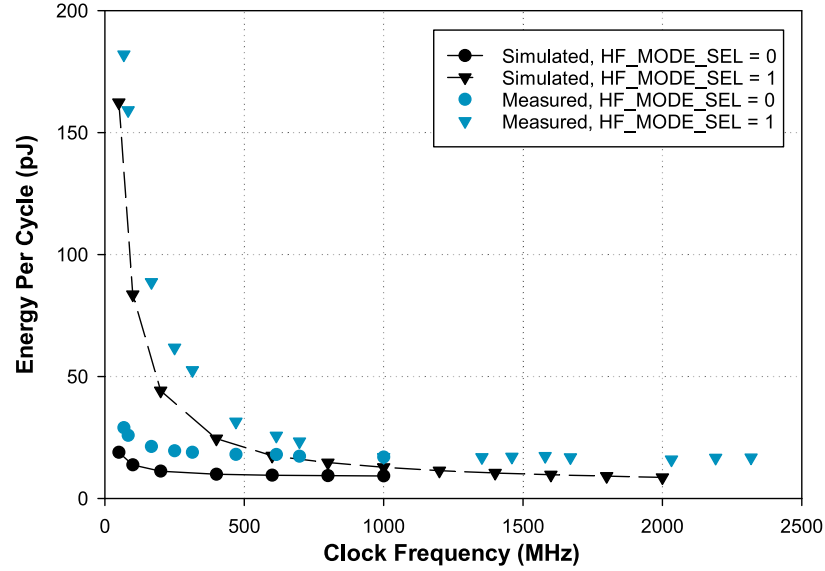


Figure 5.13: Plot showing the energy consumed by the proposed WiSync link as frequency varies. The plot compares simulated (presented in black) and silicon measurement (presented in blue) results for both high frequency ($\text{HF_MODE_SEL} = 1$) and low-frequency ($\text{HF_MODE_SEL} = 0$) modes.

measured and *simulated* energy per clock cycle for a range of TX clock frequencies. Results are presented for both the high-frequency and low-frequency transmission modes (with $\text{HF_MODE_SEL} = 1$ and 0). As shown on the figure, the measured energy consumption closely matches the simulated energy for both operating modes, indicating high confidence in the simulation setup¹⁴. The energy reduction that can be achieved through using the dual-mode transmission scheme can also be clearly seen in the results, with the energy consumed at a TX clock frequency of 50MHz being reduced from 182pJ per clock cycle with $\text{HF_MODE_SEL} = 1$ down to 27pJ per cycle with $\text{HF_MODE_SEL} = 0$ (an 85% reduction).

The link design also exceeded the operating bandwidth target, operating at frequencies as

¹⁴The slight energy offset that can be observed between the two trends here (measured and simulated) may likely be attributed to IR drop, due to the use of an external supply and the high current demand spikes in the load.

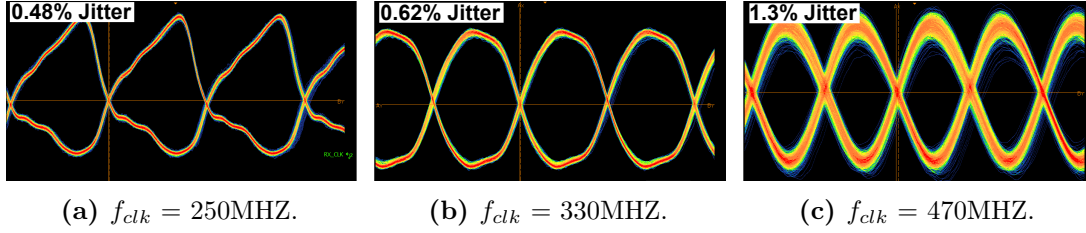


Figure 5.14: Eye diagrams showing clock jitter measurements. Clock jitter is measured externally from wire-bonded IO pins and hence represents the upper bound for each frequency.

low as 1Hz, with a maximum operating frequency of 2.4GHz. Across the target range, the maximum energy was 27pJ/cycle (at 50MHz) with a minimum energy of 15.9pJ/cycle (at 2.0GHz). The average energy across the full range of operating frequencies was 19.4pJ/cycle. Using the measurements taken from the test-chip to project the energy consumption of the WiSync link when implemented as part of a five die stack (operating with a clock frequency of 2.0GHz), this corresponds to an energy consumption of 20.1pJ/cycle, consisting of four receivers, each consuming 1.4pJ/cycle and one clock transmitter consuming 14.5pJ/cycle.

5.4.3 Jitter Measurement

Following this, the jitter in the received clock was practically evaluated across a range of frequencies. Due to the parasitics of the bond-pads and bond wires, the maximum frequency that could be exported to the oscilloscope for eye diagram measurement was 470MHz. To capture the eye diagrams an Agilent Technologies MSO9254A Mixed Signal Oscilloscope (MSO) was used with the RX clock capture triggering on the edge of the TX clock signal (exported to a separate debug pad).

Figure 5.14 presents the results of these measurements for TX clock frequencies of (a) 250MHz, (b) 330MHz and (c) 470MHz. As shown in the eye diagrams, the transceiver operates robustly with minimal jitter (less than 1.5% across all cases). It is also important to note that the clock jitter in these cases is measured using the I/O pin output so represents the upper-bound of the on-chip clock jitter (additional clock jitter will be injected into the signal as it is routed from its source, for example through the I/O pad drivers, the package bonding and PCB interconnect). The rule of thumb when considering clock distribution is that the total jitter should be less than 10% of the cycle time [175], indicating that the measured integrity of the clock is well within acceptable bounds for the frequencies measured.

5.4.4 Tolerance to Misalignment

One of the biggest appeals of wireless 3D integration is the relaxed assembly precision requirement, and hence reduced assembly *cost* when compared with galvanic approaches such as TSVs. Because of this, one final experiment was performed in this chapter to

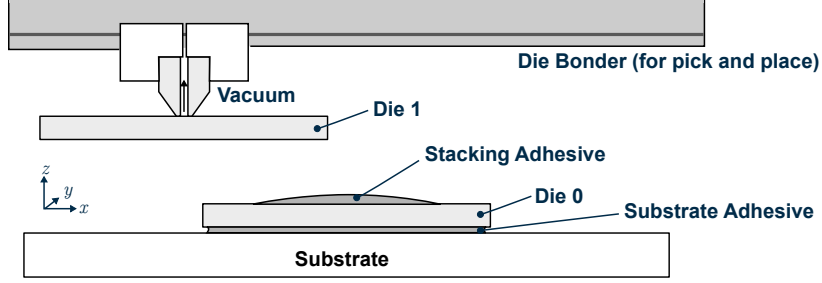


Figure 5.15: Illustration of the pick-and-place 3D stacking process using a die bonder. The lateral placement discussed in this section corresponds to an offset in the x direction.

evaluate the effects of die-to-die stacking misalignment on the performance of the proposed WiSync transceiver.

For this evaluation, rather than using test ICs where the TX and RX channel inductors are perfectly aligned, several ICs were assembled with stacking misalignments between the upper (RX) and lower (TX) tiers (beyond the deliberate $400\text{ }\mu\text{m}$ offset discussed previously to overlap the transmitter and receiver designs). Dies were stacked using a Datacon 2200 Evo die bonder [176] (as illustrated in Figure 5.15), with x -direction stacking offsets of $10\text{ }\mu\text{m}$, $25\text{ }\mu\text{m}$, and $50\text{ }\mu\text{m}$ (corresponding to a total offset of $410\text{ }\mu\text{m}$, $425\text{ }\mu\text{m}$ and $450\text{ }\mu\text{m}$ when considering the deliberate $400\text{ }\mu\text{m}$ staggering). For each experiment, the y dimension placement, substrate thickness and adhesive height were kept constant.

To evaluate the effects of stacking misalignment on performance, the Cycle Error Rate was measured (using the FPGA connected to the test board) across 10^{14} clock cycles (corresponding to approximately 28 hours of continuous operation for the utilised clock frequency of 1.0GHz). As in Section 5.3.3, the CER was defined as the ratio of erroneous RX clock cycles to the total number of TX clock cycles. This process was repeated for several values of VDD_{TX} , thereby sweeping the transmit power. Figure 5.16 shows the results of these experiments, plotting the CER (received versus transmitted clock cycles) against the total measured power consumption.

As can be observed, the link operates robustly across all the trailed stacking offsets, eventually reaching a $CER < 10^{-13}$ in each case. Expectedly, the results also show that ICs with greater misalignment in their channels perform worse than those where the TX and RX channels are aligned; assuming a target CER of 1×10^{-9} , results show that the $10\text{ }\mu\text{m}$ stacking alignment offset can be overcome by increasing the TX power by approximately 7.71%, whereas the $50\text{ }\mu\text{m}$ stacking offset requires a TX power increase of around 53.2%. To contextualise this result, typical pick-and-place machines that are used in existing packaging flows for low-cost IoT device assembly (*e.g.* used for picking the die and placing it in the package before wire-bonding) achieve a placement tolerance of $\pm 10\text{ }\mu\text{m}$. This means that, even by including a small $\pm 10\%$ TX power design margin (or by adopting the tuneable current

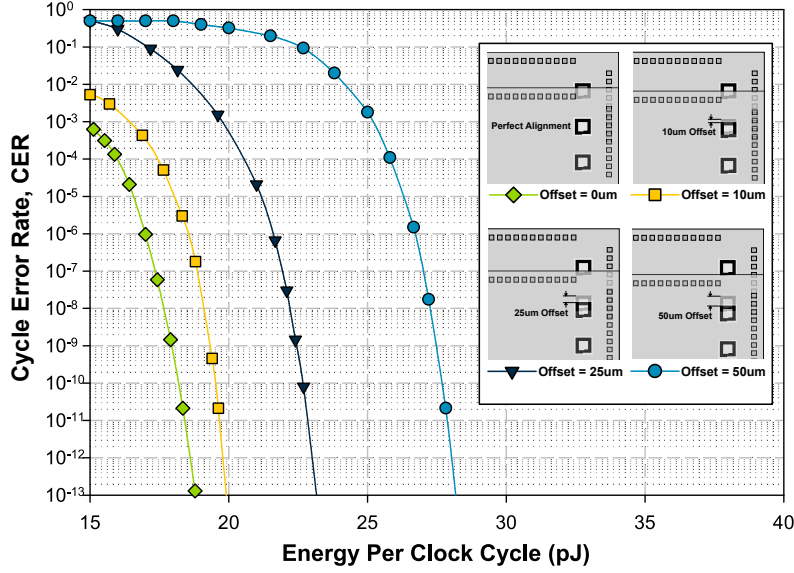


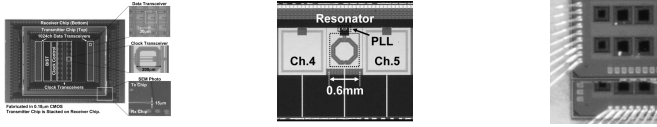
Figure 5.16: Measured Cycle Error Rate (CER) vs. energy for ICs assembled with 0 μm , 10 μm , 25 μm and 50 μm stacking assembly offsets. CER is calculated by the ratio of erroneous RX clock cycles (*e.g.* containing glitches or missing edge transitions) to the total number of TX clock cycles.

driver architecture presented in Section 3.3.2, Chapter 3) such wirelessly stacked 3D-ICs can easily be assembled using existing low-cost flows. This also represents an order of magnitude improvement when compared with the use of TSVs which typically require sub-micron placement accuracies [22].

5.4.5 Discussion

Summarising all of these measurements, Table 5.3 provides a comparison between the proposed WiSync inter-tier clock link and previously published works (Miura *et al.* ('07) [43] and Y. Take *et al.* ('15) [64]). As shown in the table, the proposed WiSync approach is the smallest reported, due to the fact that the WiSync transceiver operates in the non-resonant portion of the frequency spectrum, and so does not mandate the use of high-inductance transmit/receive coils. Although this means the energy per cycle is higher than some previous works, using non-resonant operation allows the link to operate across a wide range of frequencies (between 50MHz and 2.0GHZ) rather than a single locked frequency. It also provides much better resilience to generated clock jitter, assembly variations and PVT changes. As shown in the table, one other advantage of using the WiSync transceiver, particularly for IoT applications, is that no additional timing or frequency control circuits are required to achieve operation, unlike in [43] and [64] where large ($>0.035\text{mm}^2$) clock control circuits must also be included on the transmitting die to ensure that the link achieves resonance.

It is also important to note that energy measurements of the proposed WiSync link are

Table 5.3: Comparison of the proposed clock link design compared with existing work.


Metric	Miura <i>et al.</i> ('07) [43]	Y. Take <i>et al.</i> ('15) [64]	This Work
Approach	Coupled-Resonator	Coupled-Resonator	WiSync
Communication Distance	15um	30um	80um
Technology Node	15um	180nm	65nm
Energy Per Cycle	10pJ	0.2pJ	15.9pJ
Frequency	1GHz Only	1.1GHz Only	50MHz - 2.0GHz
Area of Inductive Link	200um×200um	600um×600um	170um×170um
Area of Additional Timing Control Circuitry (per Link)	~ 40,000um ²	~ 35,000um ²	None Required

measured across a communication distance of 80 μm , which is approximately $2\times$ the communication distance considered in previous works [64, 65]. With reference to Figure 2.9 (presented earlier in Chapter 2) the ratio of the communication distance, X , to the inductor diameter, D , in this work is ~ 0.47 placing it in the ‘square-region’ of the Near Field (NF) spectrum. This means that an x times increase in communication distance will have an x^2 impact on EM coupling strength. When normalised to consider this penalty, the energy consumption of the WiSync approach is also competitive with prior art through use of the dual-mode energy saving transmitter.

5.5 Summary

Motivated by the need for low-skew, low-energy clock synchronisation between tiers of a wirelessly stacked 3D-IC, this chapter presented WiSync, a low-overhead ICL for inter-tier clock Synchronisation. The proposed WiSync transceiver operates in the flat (non-resonant) portion of the ICL frequency spectrum to reduce silicon footprint (as resonant frequency is proportional to channel inductor size) and sensitivity to assembly, clock-source and PVT variations compared to prior-art. This chapter outlined the design of the WiSync transceiver including a dual-mode transmitter that enables the system to dynamically switch between a high-frequency operating mode (which sacrifices energy-efficiency in favour of performance), and a low-frequency operating mode (which prioritises energy efficiency), and a two stage amplifier to ensure robust clock recovery across a wide range of frequencies.

The proposed design was evaluated using commercial SPICE and EM simulators, demonstrating the ability to operate between 50MHz and 2.0GHz whilst broadcasting the clock between five stacked silicon dies (30 μm thickness) with less than 61ps of inter-tier clock skew. The WiSync link was also practically evaluated in a 2 tier 65nm CMOS 3D stacked silicon test-chip. Measurement results from this evaluation show an average energy consumption of 19.4pJ per clock cycle across an 80 μm channel, whilst consuming only 0.0421mm² of

silicon area. This makes it the smallest reported inductive wireless clock link compared with reported literature.

Finally, this chapter also presented an empirical study of CER versus packaging assembly misalignment. The results from these test show that the proposed link design can tolerate up to 50 μm of lateral pick-and-place misalignment, with only a 7.7% increase in TX power needed to overcome a 10 μm assembly fault. This represents an order-of-magnitude improvement compared with stacking tolerances when using TSVs, which typically require sub-micron die-to-die pick and place accuracies.

The WiSync link is also revisited later in Chapter 6 for use in the case study, “a 3D-stacked Cortex M0 SoC with wireless inter-tier power, data and clock delivery” (Section 6.4).

Chapter 6

Concurrent Wireless Data and Power Transmission

The previous chapters of this thesis have discussed low-energy transceiver design (Chapter 3), inductor optimisation for ICL design (Chapter 4), and wireless clock delivery (Chapter 5), however to enable *fully wireless* 3D integration, the final challenge to be addressed is delivering *power* to each die within the stack. As discussed in Chapter 2, prior works exploring ICLs for data communication typically use wire-bonding to each die to provide power (V_{DD} and GND) connections [5, 49, 51, 54, 112]. Whilst this is an adequate solution that circumvents the need for incorporating TSVs, the addition of wire-bonds to each stacked die undermines many of the benefits associated with *wireless* 3D integration.

Although there are a handful of works that do explore the possibility of wireless power transfer within the 3D stack [59, 136, 138], these works typically separate power and data delivery, meaning that the number of inductive channels (and hence area overhead) used to form the wireless interface is doubled. Further to this, the inductors required by existing Wireless Power Transfer (WPT) links are very large (typically greater than $500\text{ }\mu\text{m} \times 500\text{ }\mu\text{m}$ [59, 136, 138]) meaning that the overall footprint of the wireless power/data interface can often reach over 1mm^2 [61].

In order to address this challenge, this chapter presents an ICL that performs Concurrent Data And Power Transfer (CoDAPT), through a single *shared* inductive channel, using a Bi-Phase Shift Keying (BPSK) modulation scheme. The proposed CoDAPT transceiver is validated using commercial EM and circuit simulation tools, before being implemented in a 2-tier 3D stacked silicon test-chip in 65nm CMOS technology. A Wireless Bus Interface (WBI) is also presented to allow the CoDAPT links to interface directly with AHB-Lite peripherals.

The key the novel contributions of this chapter can be summarised as:

- Design of a Bi-Phase Shift-Keying (BPSK) ICL transceiver architecture that achieves vertical *data* and *power* delivery concurrently, through a single inductive channel.

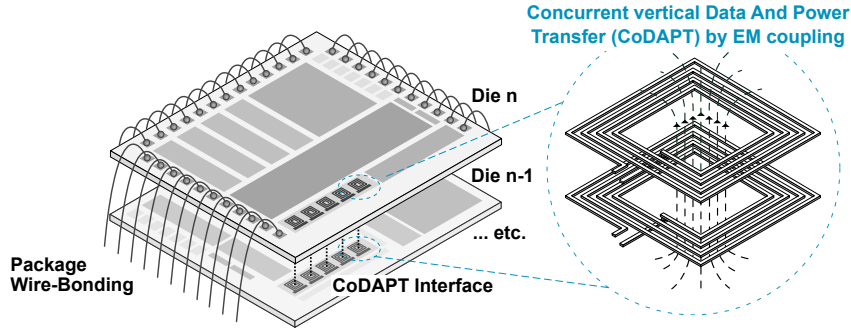


Figure 6.1: Illustration of the fully-wireless 3D integration enabled by CoDAPT. Here only the top-most die in the stack (die n) needs to be wire-bonded, and the other dies ($n - 1$, $n - 2$, etc.) can be stacked below with power and data delivered to them wirelessly.

- In-depth analysis of ICL inductor layouts focussing on the trade-off between power delivery efficiency and bandwidth, resulting in a high-coupling-coefficient, high-bandwidth design.
- Validation of the proposed transceiver, demonstrating the capability to communicate data vertically at a rate of 1.5Gbps/channel whilst simultaneously achieving power delivery of up to 2.0mW (whilst using only 0.063mm² of silicon area). This makes the presented design the smallest ever reported wireless power transfer link, and the most area efficient solution for wireless *power* and *data* delivery within a 3D-IC.
- Implementation of the proposed CoDAPT transceiver as part of a standard AHB-Lite bus (using a custom Wireless Bus Interface (WBI)) to allow straightforward, customisable stacking of various System-on-Chip (SoC) IP blocks. This marks the first ever instance of a wireless SoC bus.
- Silicon evaluation of the proposed transceiver design as part of a 2-tier 3D stacked Arm Cortex M0 SoC, demonstrating 20.3Gbps/mm² data transfer and 7.1mW/mm² power transfer simultaneously.

6.1 Background and Related Work

As discussed in Chapter 2, previous research focussed on power delivery for wirelessly stacked 3D-ICs has explored a range of different approaches to avoid the need for wire-bonded V_{DD} and GND connections to each die within the stack. In [135], Ditzel *et al.* propose the use of Highly Doped Silicon Vias (HDSVs); highly-doped wells to conduct charge between neighbouring dies once aggressive die-thinning has been performed [135]. Whilst Highly-Doped Silicon Vias (HDSVs) may present a promising future solution for power delivery in wirelessly stacked 3D-IC, they are yet to be practically realised [135] and will likely demand high fabrication costs due to the very thin ($<5\mu\text{m}$) die substrates required.

Several other works explore the possibility of using Wireless Power Transfer (WPT) in the 3D stack, most notably through inductive coupling [59, 136, 138]. A range of such works were surveyed in Section 2.4.1 (Chapter 2) concluding that these schemes can provide a useful level of power transmission, but consume a significant amount of area when implemented on-chip. Further to this, these schemes are separate to data transmission schemes, meaning that the whole wireless interface (including power and data) is typically very large.

To address this challenge, this chapter explores the possibility of performing wireless power and data transmission simultaneously through the same channel. Previous works investigating simultaneous wireless data and power transfer, outside the context of 3D integration, typically do so using Frequency-Division Multiplexing (FDM) schemes, where a low-frequency AC power-delivery signal is combined with a high-frequency modulated data signal (or vice versa) and both are transmitted concurrently through the same link [177]. Whilst this is an effective solution, in the context of 3D integration (especially for IoT devices), WPT efficiency is of high importance and therefore, when performing WPT, it is desirable to use *resonant* channels (which can offer significantly higher WPT efficiencies [139]). As FDM schemes operate across multiple frequencies, achieving resonance is very difficult [178] and so, to address this, some other prior works have explored using Amplitude Shift Keying (ASK) modulation. When using ASK, wireless power transfer occurs at a single transmission frequency and the data is signalled by modulating its amplitude [178]. Using this approach does facilitate resonant operation, however, also means that the received power is highly dependent on the transmitted data stream [178].

This chapter proposes the use of a continuous Bi-Phase Shift Keying (BPSK) scheme to address this challenge. Using BPSK, the power is recovered from the *amplitude* of the high-frequency carrier-signal, and the data is decoded from the *phase*. This means that the link uses a single operating frequency (and therefore can operate at resonance), but the dependency of the power output on the data stream is significantly reduced compared to ASK.

The remainder of the chapter is organised as follows. Section 6.2 presents the design of the CoDAPT transceiver proposed in this chapter including the transceiver circuits (Section 6.2.1) and channel inductor layout (Section 6.2.2). Following this, validation is presented through simulation in Section 6.3 using a 65nm CMOS technology, before the CoDAPT transceiver is implemented in, and experimentally validated as part of, a 3D stacked Arm Cortex M0 SoC in Section 6.4. Finally, the chapter is concluded in Section 6.5.

6.2 Concurrent Power and Data Delivery Architecture

Figure 6.2 shows an overview of the proposed CoDAPT architecture for performing concurrent data and power transmission using the BPSK scheme discussed above. The system consists

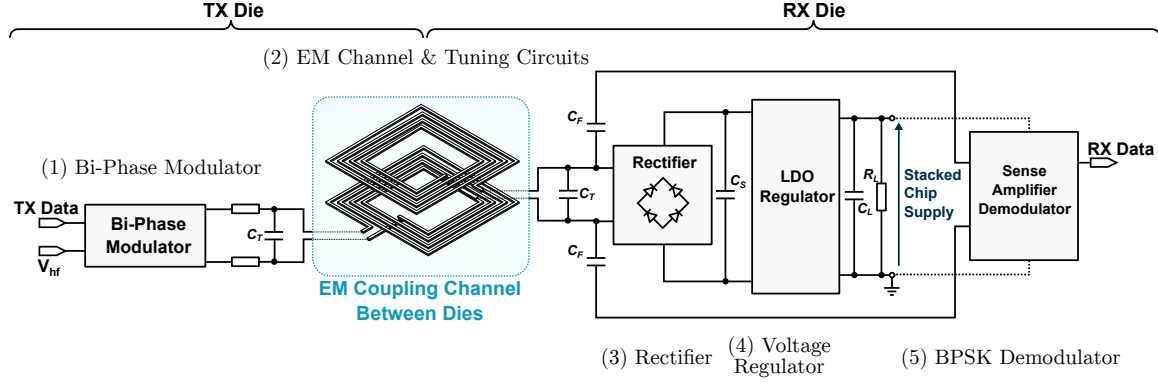


Figure 6.2: Diagram showing the end-to-end architecture of the proposed CoDAPT link including: (1) a Bi-Phase Shift Keying (BPSK) modulator, (2) the ICL channel, and tuning circuits to ensure that the system operates at resonance (to improve efficiency [40]), (3) a rectifier, (4) an LDO regulator (based upon a band-gap reference) to manage the received power supply for the recipient die, and (5) a sense-amplifier-based coherent de-modulator.

of: (1) a BPSK modulator, (2) the ICL channel and tuning circuits (to ensure that the system operates at resonance, thereby improving efficiency [40]), (3) a rectifier, (4) an LDO regulator (based upon a band-gap reference) to manage the received power supply for the recipient die, and (5) a sense-amplifier-based coherent de-modulator.

For the system to work correctly, each of these elements must be carefully designed to maximise *power* delivery efficiency, whilst still supporting *data* transmission (therefore minimising the area overhead when compared with prior art by performing both concurrently through a single inductive channel). The design of each element is documented in the sub-sections below.

6.2.1 Bi-Phase Transmitter Design

As discussed above, to perform data and power transmission concurrently, this chapter proposes the use of a BPSK modulation scheme where the carrier signal is used to deliver power between tiers, and the data signal is encoded onto it in terms of phase. The operation of the transceiver is illustrated by the waveforms in Figure 6.3 (a), which capture on the operation of the modulator, demodulator and rectifier circuits used in the CoDAPT design. Here, two high-frequency carrier signals are used by the transmitter: the in-phase carrier signal, HF_CLK (with frequency f_{hf}) and the quadrature carrier signal, HF_CLK_SHIFT (also with frequency f_{hf} , but inverted when compared with HF_CLK). The CoDAPT transmitter encodes the data stream by selectively transmitting one of these signals; when the TX_DATA signal is high, the in-phase carrier signal is transmitted, conversely, when the TX_DATA signal is low, the quadrature carrier signal is transmitted. The effect of this is shown by the black and grey waveforms in Figure 6.3 (a) (MOD_DATA_N and MOD_DATA_P) which form the differential output signals of the bi-phase modulator. By continuously selecting between

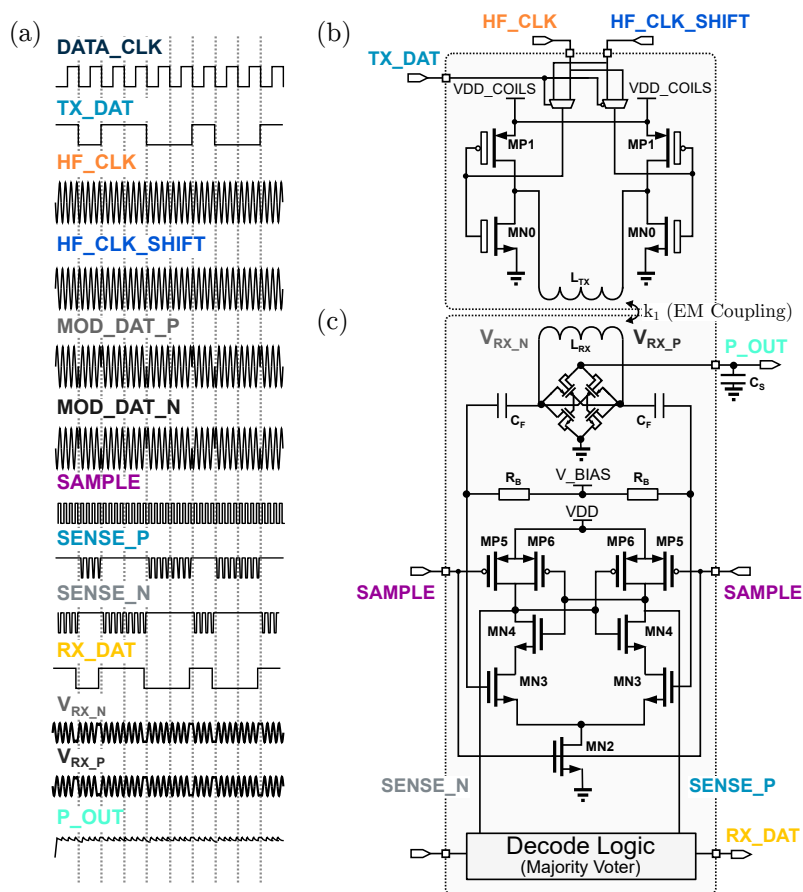


Figure 6.3: Diagrams illustrating: (a) the operation of the CoDAPT link, (b) the circuit implementation of the proposed CoDAPT BPSK modulator, and (b) the circuit implementation of the proposed CoDAPT BPSK demodulator.

the in-phase and quadrature carrier signals, the output from the CoDAPT modulator is essentially phase encoded, with 180° phase shift being observed at every TX_DAT edge.

Figure 6.3 (b) illustrates the proposed modulation circuit required to realise this transmission. Here, MN0 and MP1 are the drive transistors that source the current to form the magnetic field through the TX inductor, L_{TX} . As in previous chapters, these can be sized according to the desired transmit current (which will reflect the thickness of the stacked dies and the power delivery requirements of the application). MN0 and MP1 are fed by MUX circuits which transition between HF_CLK and HF_CLK_SHIFT, in the manner discussed above. As shown, L_{TX} is driven differentially allowing for maximum dI_{TX}/dt flux linkage, and, to further maximise the WPT achieved by the link, transistors MN0 and MP1 are implemented using thick gate-oxide transistors and driven by a separate supply (VDD_COILS). Thick oxide transistors are tolerant of higher drain-source voltages, and hence allow VDD_COILS to be increased beyond the system V_{DD} , maximising the power transmitted through the link. For the implementation in this chapter, push-pull level-shifters were used at the interface of

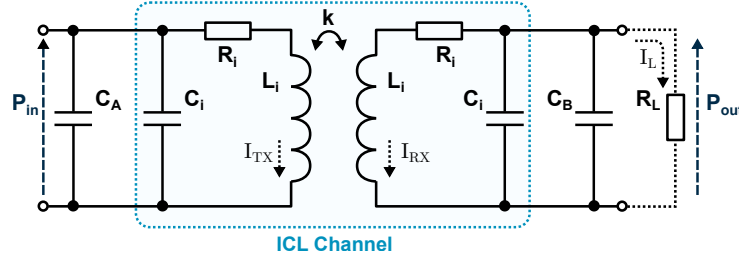


Figure 6.4: Equivalent circuit diagram on an ICL channel with parallel tuning capacitors, C_A and C_B .

the CoDAPT transmitter enable this power increase, whilst facilitating straightforward integration with the rest of the system.

6.2.2 Inductive Channel and Tuning Circuit Design

The second element in the CoDAPT architecture is the inductive coupling channel itself, consisting of two coupled planar inductors. As discussed in the previous chapter, power transmission between two stacked inductors is optimised when their layouts are congruent [179]. Therefore, two congruent square spiral inductors are adopted for the CoDAPT transceiver (as square inductors offer the highest inductance per unit area [162]).

In order to maximise the power efficiency of the CoDAPT link presented in this chapter, *resonant* inductive coupling is used (where the BPSK carrier frequency is selected to match resonant frequency of the coupled inductive channel). The proposed BPSK scheme naturally lends itself to such operation as the transmission is continuous and sinusoidal, unlike the pulse-based ICL transceivers that have been presented in the previous chapters. Resonant operation has been demonstrated to provide superior power transmission efficiency [139], but also provides a natural RX-side voltage boost making voltage regulation in the RX die more straightforward (as simple linear regulators, such as the Low Drop Out (LDO) regulator adopted in this chapter, can be used to manage the supply rather than, for example, large boost converter circuits).

Achieving resonant operation, however, introduces several new design constraints when considering the inductive channel design. As discussed in Chapter 4, due to the intrinsic resistance, capacitance and inductance of the coils used to form the channel, each pair of layout geometries has a *natural* self-resonant frequency f_{sr} and this is accounted for in the COIL-3D optimisation flow (presented in Section 4.3.3 of Chapter 4). However, when optimising based on f_{sr} alone, only a limited number of resonant operating frequencies are considered due to the fact that the resistance R_i , inductance, L_i and capacitance C_i of the coil may only be selected from combinations that correspond to physically realisable layouts.

To address this limitation in this chapter, passive tuning capacitors (C_A and C_B) are introduced to artificially adjust the resonant frequency of the channel, thereby maximising efficiency, as

shown in Figure 6.4. To incorporate the use of tuning capacitors in the optimisation flow, the expressions for the power delivery efficiency η_{pow} (presented in Section 4.3, Chapter 4) can be revisited, as outlined below.

Fundamentally, the power transmission efficiency of the link, η_{pow} is given by:

$$\eta_{pow} = P_{out}/P_{in}, \text{ where } P_{out} = I_L^2 R_L \quad (6.1)$$

Here, the current through the RX-side load, I_L is given by:

$$I_L = \frac{I_{RX}}{1/j\omega C_i + R_L} \times \frac{1}{j\omega C_i} = \frac{1}{1 + j\omega C_i R_L} \quad (6.2)$$

Applying circuit theory to the ICL equivalent circuit highlighted in blue in Figure 6.6 (excluding C_A and C_B), detailed expressions for I_{RX} (and hence P_{in} and P_{out}) can be derived for the ICL channel as follows:

$$P_{in} = \frac{R_i + j\omega L_i + R_L / (1 + R_i j\omega C_i)}{\left[(1 + R_i C_i j\omega - \omega^2 L_i C_i) / (R_i + j\omega L_i) \right] \left[R_i + j\omega L_i + \gamma \right] + \omega^2 M^2} \quad (6.3)$$

and,

$$P_{out} = \frac{\omega^2 M^2 R_L}{\left\{ \left[(1 + R_i C_i j\omega - \omega^2 L_i C_i) / (R_i + j\omega L_i) \right] \left[R_i + j\omega L_i + \gamma \right] + \omega^2 M^2 \right\}^2} \times \frac{1}{1 + j\omega C R_L} \quad (6.4)$$

where here, $\gamma = R_L / (1 + j\omega C_i R_L)$. Now, with only a slight modification, the two intentional tuning capacitors shown on Figure 6.6 can be added (C_A on the TX-side, and C_B on the RX-side) resulting in the updated expressions (for P'_{in} and P'_{out}) below:

$$P'_{in} = \frac{R_i + j\omega L_i + R_L / (1 + R_i j\omega (C_i + C_B))}{\left[(1 + R_i (C_i + C_A) j\omega - \omega^2 L_i (C_i + C_A)) / (R_i + j\omega L_i) \right] \left[R_i + j\omega L_i + \gamma \right] + \omega^2 M^2}$$

and,

$$P'_{out} = \frac{\omega^2 M^2 R_L}{\left\{ \left[(1 + R_i (C_i + C_A) j\omega - \omega^2 L_i (C_i + C_A)) / (R_i + j\omega L_i) \right] \left[R_i + j\omega L_i + \gamma \right] + \omega^2 M^2 \right\}^2} \times \frac{1}{1 + j\omega (C_i + C_B) R_L}$$

where now, $\gamma = R_L / [1 + j\omega (C_i + C_B) R_L]$. Using these new expressions (that include the

effects of the tuning capacitance) the COIL-3D optimisation flow (presented in Chapter 4) was used to find best-performing geometries for the inductive channel (that maximise η_{pow} , where $\eta_{pow} = P'_{out}/P'_{in}$ for a target clock frequency of 1.5GHz¹. Results from this optimisation process are presented later in Section 6.3.1.

6.2.3 Rectifier and Low Drop-Out Regulator Design

Revisiting the target CoDAPT architecture presented in Figure 6.2, once the TX data has been encoded and transmitted through the channel via resonant coupling, the received alternating voltage must be rectified for powering the RX die. To rectify the received signal in this chapter, a CMOS cross-coupled rectifier is used, as illustrated in Figure 6.3. Work by Han *et al.* [139] presents extensive comparison of available on-chip rectifier solutions for this style of application, concluding that a cross-coupled rectifier can provide the highest efficiency [139]; the advantage of using such a design, when compared with a diode-based rectifier (*e.g.* full-wave bridge rectifier, or half-wave-rectifier), is that the two PMOS transistors (*c.f.* Figure 6.3) mitigate the voltage threshold (V_{th}) drops that are present in the ‘ON’ state [180]. This means that cross-coupled rectifiers can operate at reasonably high efficiencies, even when the input voltage is low, making them well suited for use in the presented CoDAPT transceiver.

This low ON resistance, however, comes at the expense of slightly higher reverse leakage, especially if the input transistor does not switch quickly on each ON-OFF transition [181]. Cross-coupled rectifiers have been demonstrated in prior art at very high frequencies ($> 2.4\text{GHz}$) in the same 65nm CMOS technology adopted in this chapter [182], however, as a small amount of energy is lost in reverse leakage *at each clock edge*, an intrinsic trade-off exists between frequency and WPT efficiency. When considering the CoDAPT application discussed in this chapter, this directly translates to a trade-off between data bandwidth and WPT efficiency. To strike a balance in this work, a target f_{hf} of 1.50GHz was selected to remain competitive with prior art in terms of data-rate [51, 123], whilst avoiding the inflated reverse leakage associated with higher ($> 2.0\text{GHz}$) switching frequencies.

Once rectified, the power output is regulated for use in the RX die, as shown in Figure 6.2². To achieve this, the LDO regulator shown in Figure 6.5 is adopted, which uses a band-gap reference to create a stable output voltage, irrespective of the supply. The operation of the LDO presented in Figure 6.5 is as follows. Initially, MP7 is switched on, and the voltage at P_OUT (the output of the rectifier stage discussed above) rises as the capacitive load at P_OUT_REG (C_L) charges. As a result of this, the input to the comparator (highlighted in

¹A target frequency of 1.50GHz was selected to strike a balance between the data-rate and the rectifier efficiency; a discussion on this is provided later in Section 6.2.3.

²Where several CoDAPT links are included on the same chip (to provide higher inter-tier bandwidths and increased WPT), each link has its own local rectifier and the outputs are combined to charge the shared C_S buffer at the P_OUT node.

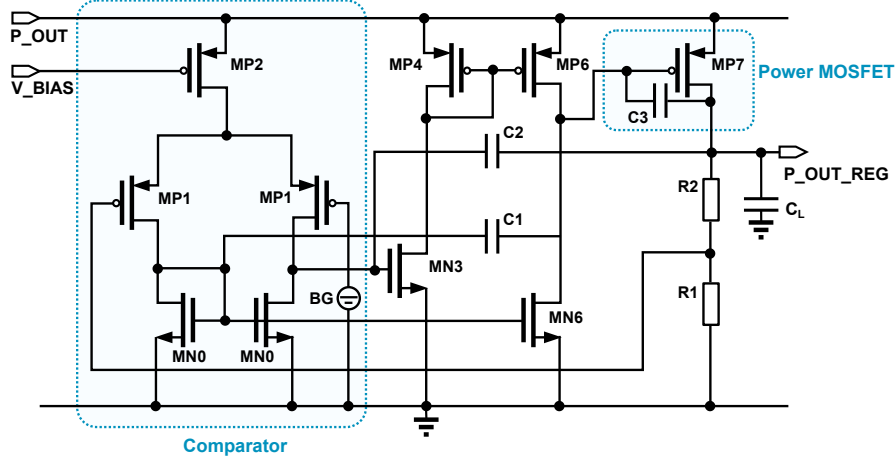


Figure 6.5: Schematic diagram of the Low Drop-Out (LDO) voltage regulator circuit used in this work with the comparator and power MOSFET components highlighted.

blue) which is set by the ratio $R1:(R1 + R2)$ will also rise. If $P_OUT_REG \times R1 / (R1 + R2)$ increases above the band-gap reference (BG) voltage, MN3 will start to conduct (due to the change in the comparator output) and MN6 will switch on. This will in-turn cause the gate voltage of MP7 to rise, limiting the current that can flow through MP7 (hence limiting the output supply P_OUT_REG). At this point, the supply will begin to discharge across the load at P_OUT_REG. Once the supply falls below the threshold $P_OUT_REG \times R1 / (R1 + R2)$ again, the inverse of this process will occur and the gate voltage of MP7 will fall. This will again re-enable source-drain conduction, and hence P_OUT will charge again. This feedback loop enables the output voltage to be regulated, provided $P_OUT > V_{DD}$ ³.

6.2.4 Bi-Phase Demodulator Design

The final element of the CoDAPT transceiver design is the *data* demodulator, for sampling the phase of the received AC voltage signal. Figure 6.3 (c) shows the proposed demodulator design, based on a StrongARM circuit [183]. The operation of this circuit is as follows. During the time where the SAMPLE signal is low⁴, the PMOS transistors (MP5/MP6) pre-charge the output nodes SENSE_N and SENSE_P to V_{DD} . Once pre-charged, at each rising data clock edge, the SAMPLE signal transitions high allowing transistors MN3 to amplify the differential signal across the RX inductor (which has been initially filtered using the filter capacitors labelled as C_F in Figure 6.3 (c) and Figure 6.2 to remove any DC offset resulting from the rectifier/regulator stages of the CoDAPT receiver, and biased through bias resistors R_B such it falls within the saturation region of M3).

Once the signal is amplified, SENSE_N and SENSE_P will be determined by the relative

³Practical evaluation of this LDO regulator is provided in Section 6.3.3.

⁴The CoDAPT demodulator is implemented coherently, and the SAMPLE signal is a short synchronous pulse generated at the *rising* edge of the TX/RX clock.

differential potential; if the in-phase carrier has been transmitted, V_{RX_P} will be greater than V_{RX_N} resulting in $SENSE_P$ being pulled low (indicating a TX_DATA value of ‘1’). Conversely, if the quadrature carrier has been transmitted, V_{RX_N} will be greater than V_{RX_P} and $SENSE_N$ will be pulled low (indicating a TX_DATA value of ‘0’). These output pulses (at $SENSE_N$ and $SENSE_P$) are then passed to a digital logic block that determines the correct output value. In the simple case that the data rate, f_{DAT} , is equal to f_{hf} only one sample will be taken per clock cycle, and hence this is used as the output. However, in typical operation $f_{hf} \gg f_{DAT}$ and so *multiple* samples are taken. In this case, the digital logic block acts as a majority voter, and passes the most frequent sample value within the data clock window to the output.

As in the previous chapters, the gate-level layout of the receiver was carefully considered to provide characteristic matching between differential components, thereby minimising the effects of PVT variations. The channel lengths of transistors M3 (which determine the gain of the receiver) were made as large as possible (within the 1.5GHz HF_CLK switching specification), and Metal Insulator Metal (MIM) decoupling capacitors were placed in the spare area around the design to minimise IR drop (and hence improve the voltage margin). Careful consideration was also given to the layout of the bias resistors, with an interdigitated layout adopted to mitigate variation effects in the polysilicon mask (which could otherwise result in an unintentional DC bias in the SA).

6.3 Results and Evaluation

Bringing together each of these five elements, this section presents evaluation of the proposed CoDAPT link using commercial EM and circuit simulators. As in previous chapters, the EM portion of the link was simulated using Ansys HFSS, assuming a standard nine metal 65nm CMOS technology BEOL stack-up. For simulation, a fitted broadband SPICE model of the channel was exported from HFSS and used for H-SPICE simulation of the link in conjunction with the above outlined CoDAPT circuits. Experiments were performed for die thicknesses of 10 μm (representative of very aggressive high-cost thinning, or Face-to-Face (F2F) stacking), 30 μm (representative of aggressive thinning in conjunction with Face-to-Back (F2B) stacking), and 70 μm (representative of low-cost die-level thinning with F2B stacking)⁵ to explore the effects of die thickness on performance. When combined with the 10 μm epoxy adhesive layer assumed for die attach in each case, this results in total channel communication distances of 20 μm , 40 μm and 80 μm respectively.

The following sections outline the experimental results, including the inductor geometry optimisation process (Section 6.3.1), in addition to area evaluation (Section 6.3.2), transient performance evaluation (Section 6.3.3), power delivery performance (Section 6.3.4), and

⁵Empirical evaluation of the CoDAPT transceiver at this thickness is explored in Section 6.4 through silicon test-chip implementation.

data delivery performance, including Bit Error Rate (BER) evaluation and tolerance to clock timing jitter (Section 6.3.5).

6.3.1 Channel Inductor Optimisation

To determine the layouts of the TX and RX channel inductors for the CoDAPT link, the COIL-3D optimisation flow was used with the adaptations outlined previously in Section 6.2.2 included, to incorporate the effects of, and find optimised values for, tuning capacitors (C_A and C_B). The objective function in this case was as shown in Equation 6.5:

$$\begin{aligned}
 &\text{maximise} \quad P'_{out}/P'_{in} \text{ (c.f. Section 6.2.2)} \\
 &\text{subject to} \quad w_{ijk} > w_{min}, s_{ijk} > s_{min} \forall ijk, \\
 &\quad D_i > 2 \left[\sum_{j=1}^n 2(w_{i,j}) + \sum_{j=1}^{n-1} 2(s_{i,j}) \right], \\
 &\quad 1/2\pi\sqrt{L_i C_i} < f(1 - k_t) \\
 &\quad 0 < C_A < 1\text{pF and } 0 < C_B < 1\text{pF} \\
 &\text{where} \quad n_1, n_2 \in \mathbb{Z}+ \text{ and } w_{ijk}, s_{ijk} \forall ijk \in \mathbb{R}+
 \end{aligned} \tag{6.5}$$

(Note here C_A and C_B have been confined to a maximum of 1pF to limit the silicon area overheads associated with their implementation).

As in Chapter 5, for practical implementation reasons (DRC compliance with the TSMC 65nm technology used for fabrication), some additional physical constraints were also imposed upon the inductor geometry. These were $\chi_w = 0$, $\chi_s = 0$, $1.0\mu\text{m} < w < 12\mu\text{m}$ and $g = 1.0\mu\text{m}$. To minimise the area usage of the link, the maximum coil diameter, D , was also limited to $250\mu\text{m}$.

Using the modelling presented in Chapter 4, along with the refined optimisation flow, an optimised layout was determined for the target frequency of 1.5GHz, along with tuning capacitor values that correspond to best-performing efficiency. In this case, the generated optimised layout parameters were: track-width (w) = $5\mu\text{m}$, track-spacing (s) = $2\mu\text{m}$, number of turns (n) = 6, $C_A = 0\text{fF}$, and $C_B = 810\text{fF}$. Using these parameters, the simulated maximum WPT efficiency of this link using COIL-3D was 7.2%, assuming the worst-case communication distance of $80\mu\text{m}$.

To illustrate the effects of using resonant tuning circuits in this work, Figure 6.6 (a) shows a plot of resonant frequency versus tuning capacitor size, assuming the inductor layout generated by COIL-3D above. Here, it can be observed that the tuning capacitance is effective in reducing the resonant frequency from its natural f_{sr} of 6.7GHz down to the target frequency of 1.5GHz at 810fF. Figure 6.6 (b) shows an AC simulation of the link efficiency

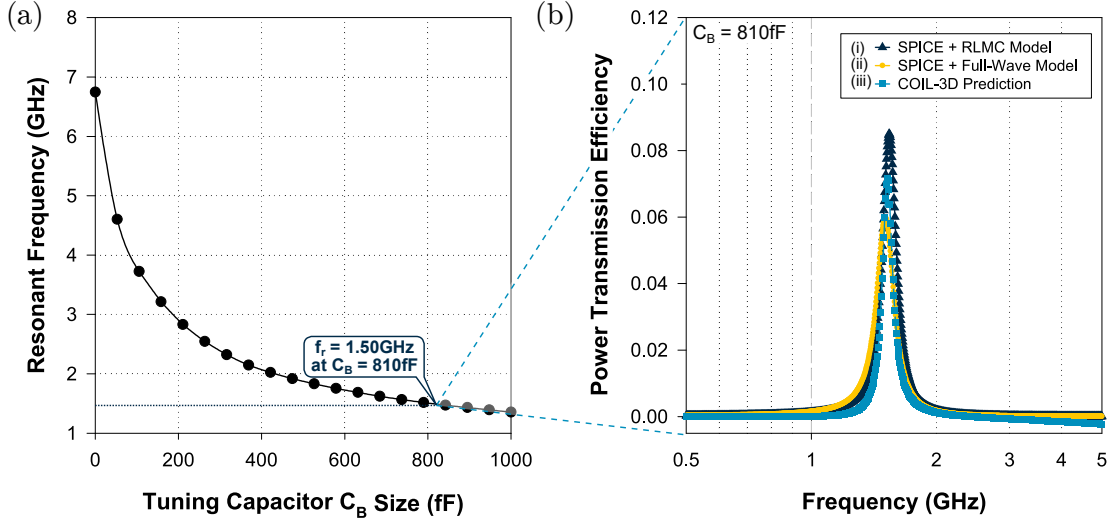


Figure 6.6: (a) Plot of simulated channel resonant frequency vs. tuning capacitor size, C_B . (b) AC simulation showing simulated power transmission efficiency vs. frequency when using (i) SPICE + RLMC model fitted using FEM (ii) SPICE + broadband model fitted using FEM, and (iii) RLMC model + COIL-3D expressions.

at this value of capacitance⁶, comparing the accuracy of the COIL-3D models (used for optimisation purposes) and the simulated performance of the link using Finite Element Methods (FEM) (in case (i) using the FEM results to generate parameters for the *RLMC* fitted model in Figure 6.6, and in case (ii) using the FEM results in HFSS to automatically generate a fitted broad-band SPICE model). As shown on the figure, the expressions (and hence frequency response used in the COIL-3D optimiser) match very closely with FEM simulations correctly determining the self-resonant frequency within 2% of the broadband FEM simulated results, indicating high confidence in the optimised layout result.

6.3.2 Area Evaluation

Bringing each of the elements outlined above together, including the optimised inductor geometry outlined in the previous section, Figure 6.7 shows the physical layout implementation of (a) the CoDAPT transmitter and (b) the CoDAPT receiver in TSMC 65nm CMOS technology, with the key components highlighted. As shown, the CoDAPT transceiver consumes approximately 0.095mm^2 of silicon area (when considering the bounding box), consisting of 0.0625mm^2 for the ICL channel inductors, 0.006mm^2 for the receiver including the tuning capacitor, rectifier and demodulator, and 0.003mm^2 for the transmitter (detailed area breakdowns are itemised later in Table 6.1).

For experiments in this chapter, it was assumed that a 1pF storage MIM capacitor (C_S) is

⁶Here, the plotted figure shows the *maximum theoretical* channel efficiency of the link across a static load, and hence does not account for losses due to the accompanying circuits; these are evaluated in the later sections.

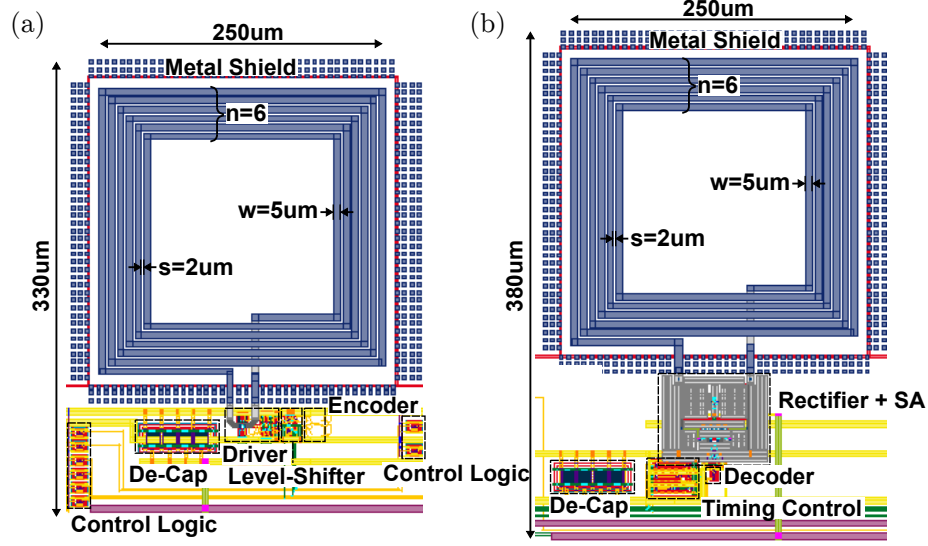


Figure 6.7: Layout of (a) the CoDAPT transmitter and (b) the CoDAPT receiver, highlighting the key components.

placed at P_OUT which is connected to the LDO regulator discussed above. The silicon areas of these components (buffer capacitor and LDO) are 0.004mm^2 and 0.012mm^2 respectively however, it is important to note that when multiple CoDAPT links are combined to form a common wireless data/power interface (as is typically required, *c.f.* the use case study in Section 6.4), these overheads are amortised across the whole interface.

6.3.3 Start-Up and Transient Performance

Following this, the start-up behaviour of the CoDAPT system was assessed. As the RX die is powered directly from the transmitted data signal when using the CoDAPT transceiver, there is a small period of time where the energy buffer must charge, this will be referred to as the *warm-up* period. Figure 6.8 shows the transient performance of the proposed system during this time, assuming a communication distance of $40\text{ }\mu\text{m}$. For this transient simulation, a constant current sink of 0.5mA (representing extraneous circuits in the RX die) is applied and, under these conditions, P_OUT_REG reaches the nominal V_{DD} (1.20V) after 24ns of warm-up. At this point, the TX die starts transmitting useful data, and the RX die successfully decodes the data using the CoDAPT demodulator. This can be observed in the RX_DAT signal on Figure 6.8. At 1.0GHz operation, one phase sample is collected per transmitted bit (as shown on the figure), and hence the measured latency is one clock cycle. For normal operation (assuming the transmission of an equiprobable Pseudo-Random Binary sequence (PRBS)) the maximum variation in the recovered supply was simulated to be 7.0% .

The summary presented in Table 6.1 (page 138) evaluates the supply stability and warm up-period in this way, across each of the considered communication distances ((i) $20\text{ }\mu\text{m}$,

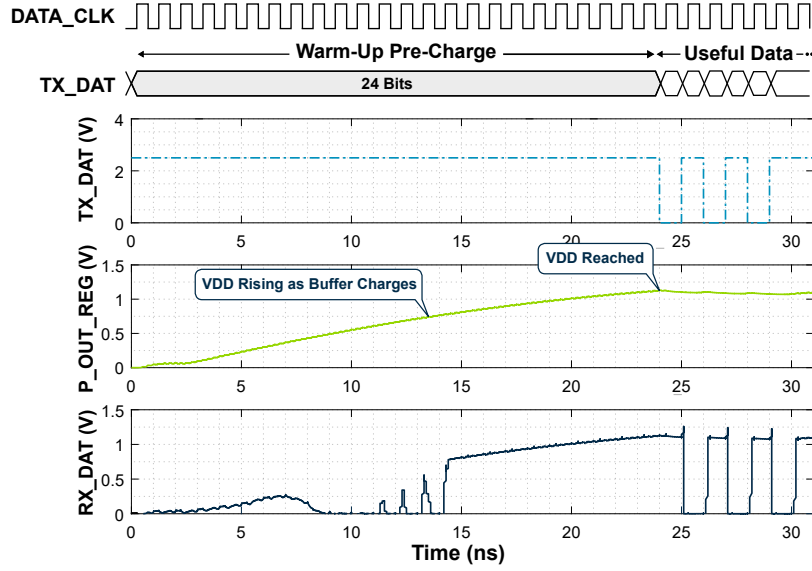


Figure 6.8: Transient simulation illustrating the warm-up period duration whilst using the CoDAPT transceiver (simulated assuming a 40 μm communication distance). The capacitive buffer at P_OUT used here is 1pF, and the load is a constant current of 500 μA .

(ii) 40 μm , and (iii) 80 μm). As shown in the table, when the communication distance is between 20 μm and 40 μm , the resonant inductive coupling channel with $V_{DD_COILS} = 2.5\text{V}$ (supported by the thick oxide drive transistors MN0 and MP1, see Figure 6.3 (a)) achieves an RX-side supply voltage greater than 1.2V, allowing the RX die to operate at nominal voltage. When communicating across 80 μm , however, the maximum output V_{DD} was simulated to be 0.9V. As such, for this case, the LDO regulator and CoDAPT receiver were sized to use a 0.8V supply voltage. Whilst this does not represent the full nominal V_{DD} , it is still sufficient to power digital logic cells at the 65nm node considered here [184] and, where required, an additional boost converter could be used to provide the extra voltage headroom.

6.3.4 Power Delivery Performance

Following this, the power delivery of CoDAPT transceiver was evaluated. As the CoDAPT transceiver performs data and power delivery concurrently, the amount of received power will, to some extent, depend on the data that is being sent. When the transceiver is transmitting data patterns that have many edge transmission, for example 0xAAAA (where a transition occurs between every bit), a phase shift will be introduced in every clock cycle. As illustrated on Figure 6.3 (a), continuously shifting the phase between every TX bit results in a reduction of dI_{TX}/dt , flux linkage, and hence power delivery. Conversely, when transmitting constant data patterns, for example 0x0000 or 0xFFFF, dI_{TX}/dt is maximised, and hence the maximum power is transferred between dies.

The results presented in Table 6.1 (page 138) summarise the *peak* and *average* power delivery

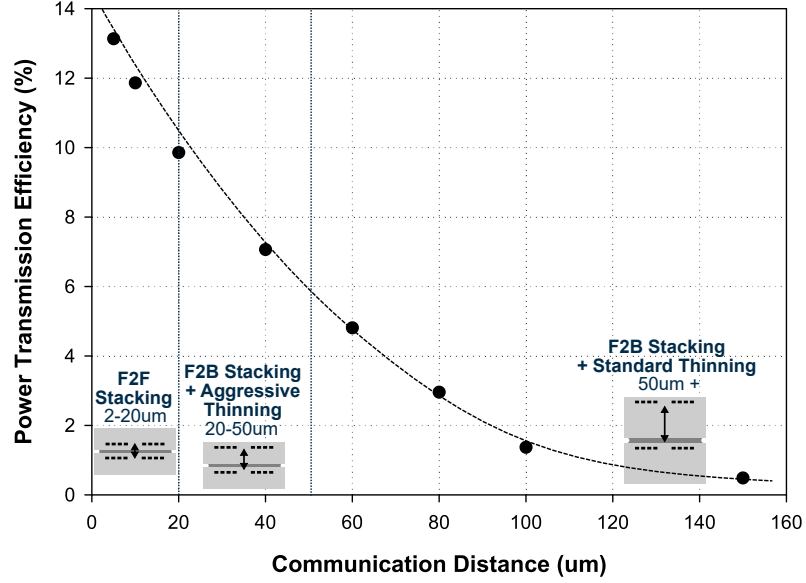


Figure 6.9: Simulated effects of die thickness on power delivery efficiency (P_{out}/P_{in}).

of the CoDAPT transceiver across the three different communication distances considered in this section⁷. Simulated results show that the CoDAPT link can perform power transmission of up to 2.01mW per link assuming best-case conditions (across a communication distance of 20 μm), or 510 μW when the communication distance is 80 μm. The channel used in this work is 250 μm in diameter, translating to a power delivery density of between 7.68mW/mm² and 24.16mW/mm². When compared to previously reported 3D-ICs performing inter-die WPT, these represent competitive figures, with previous works reporting WPT of between 3.37mW/mm² [61] and 33.3mW/mm² [59]. These previously reported works, however, *only* perform wireless power transmission, whereas the CoDAPT transceiver has the advantage of concurrently transmitting data through the same link.

Following this, a study was performed to evaluate the effects of die thickness (and hence overall communication distance) on power efficiency (P_{out}/P_{in}). Figure 6.9 shows the results of these experiments, illustrating how the efficiency varies with communication distance. Here, the plot is divided into three portions; communication distances between 2 μm and 20 μm (representative of F2F stacking), communication distances between 20 μm and 50 μm (representative of F2B stacking with aggressive, high-cost thinning) and communication distances greater than 50 μm (representative of F2B stacking with standard low-cost thinning). As shown on the figure, simulation results indicate that the maximum WPT efficiency that can be achieved using the CoDAPT transceiver in conjunction with F2B stacking and standard low-cost die thinning is around 6%.

⁷Here, the peak power delivery is measured whilst transmitting a bit sequence of 0xFFFF and the average power delivery is measured whilst transmitting a PRBS.

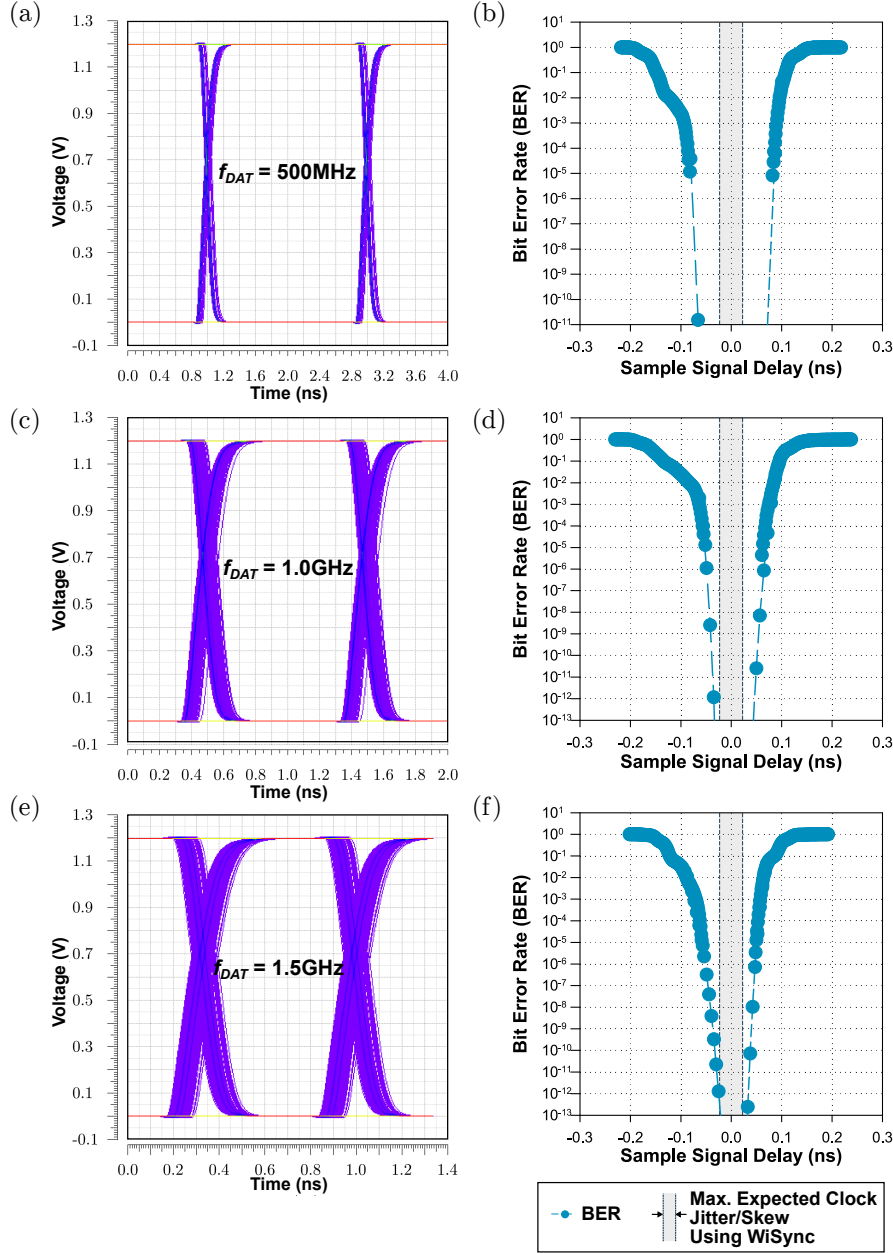


Figure 6.10: Eye diagrams and BER curves showing the simulated data transmission performance of the presented CoDAPT link whilst operating at 500MHz ((a) & (b)), 1.0GHz ((c) & (d)), and 1.50GHz ((e) & (f)).

6.3.5 Data Delivery Performance

The data communication performance of the CoDAPT link was then evaluated in terms of Bit Error Rate (BER) and tolerance to jitter in the SAMPLE signal timing using the dual-Dirac modelling approach, outlined in Appendix B. Figure 6.10 shows the simulated BER of the CoDAPT transceiver, assuming a communication distance of $80\text{ }\mu\text{m}$ (representing the worst-case EM corner for the three die thicknesses considered in this chapter) for a range

of operating frequencies. The figures on the left show the eye diagram of the received data signal (RX_DAT), and the figures on the right show BER as a function of TX/RX clock jitter.

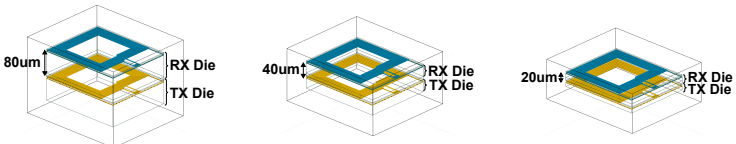
As can be observed on the figure, the WiSync link operates robustly in terms of data delivery performance across all the frequencies, with a wide eye-opening and low inter-symbol jitter. The BER curves also show that the presented transceiver is very tolerant of jitter in the clock signal which is used to selectively enable the synchronous phase-sensor in the receiver. Even at the maximum operating frequency of 1.5GHz, the simulated timing margin of the CoDAPT receiver is around 100ps. To contextualise this result, the wireless clock transceiver, WiSync, that was presented in Chapter 3 exhibits a maximum worst-case inter-tier clock skew of 61ps and jitter of 16.3ps (depicted by the grey shaded area on Figure 6.10), providing sufficient design margin to ensure robust operation when used in conjunction with the CoDAPT transceiver.

Finally, Table 6.1 presents a summary of the results from this section including the simulated area, power, BER and maximum clock jitter metrics across the three different EM simulation set-ups. As shown, the simulated results indicate high performance across each of the cases with low BER projections when used in conjunction with the WiSync clock link from Chapter 5. The projected power delivery across all cases is greater than 0.5mW, sufficient to power a sub-threshold MCU such as the Ambiq Apollo 3 [185] (a typical application for IoT systems) using just a single link.

6.4 Case Study: A 3D stacked Arm Cortex M0 SoC (Silicon Evaluation)

As discussed in the literature review, the focus of this thesis is to enable customisable stacking of disparate sensor/memory/logic dies, with low cost, for IoT-style applications. To demonstrate how this is possible using the CoDAPT links proposed in this chapter, this section presents a use-case example where two 3D-stacked Arm Cortex M0 microprocessors (representative of the class of processor used in IoT applications, the focus and motivation for this thesis) are integrated vertically (alongside 8kB of SRAM) to form a 3D SoC using CoDAPT links. Further to this, the CoDAPT links are used to form the main system AHB-Lite Bus (representing the first ever instance of a wireless SoC bus) meaning that the stacked Slave die could be easily interchanged with any other peripheral that uses the AHB-Lite protocol.

Figure 6.11 shows the architecture of the demonstrator test-chip presented in this chapter. In total, five ICLs are included in the chip: 2× Concurrent Data and Power Transfer (CoDAPT) uplinks, 2× Data Only (DO) downlinks, and 1 × WiSync clock link. The WiSync clock link ensures precise synchronisation between dies and is driven from the lower die, the CoDAPT links provide concurrent data and power transmission from the lower (Master) die to the



Metric	Simulated Performance (Communication Distance = 80µm)	Simulated Performance (Communication Distance = 40µm)	Simulated Performance (Communication Distance = 20µm)
Channel Area	0.0625mm ²	0.0625mm ²	0.0625mm ²
CoDAPT Transmitter Area	0.095mm ²	0.095mm ²	0.095mm ²
CoDAPT Receiver Area	0.095mm ²	0.095mm ²	0.095mm ²
Average Power Delivery Per Link	480uW	1.12mW	1.51mW
Peak Power Delivery Per Link	510uW	1.4mWmW	2.01mW
Average Power Delivery Per Area	7.68mW/mm ²	17.92mW/mm ²	24.16mW/mm ²
BER	<10 ⁻¹³	<10 ⁻¹³	<10 ⁻¹³
Max. Tolerated Clock Jitter [at 1.5GHz]	0.47ns	0.45ns	0.45ns
RX Supply (Voltage, % Variation)	1.2V, ± 8%	1.2V, ± 7%	0.8V, ± 7%
Warm-Up Period [500uA, 1pF load]	7ns	24ns	26ns

Table 6.1: Table summarising the simulated performance of the CoDAPT link across three different communication distances: 20 µm, 40 µm and 80 µm. Per-area metrics are calculated based on *channel* area, in accordance with prior art.

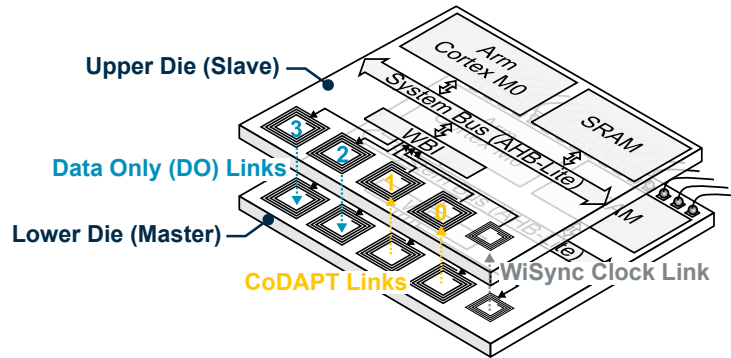


Figure 6.11: 3D illustration of the 2-tier 3D stacked Arm Cortex M0 SoC presented in this section with CoDAPT links between tiers.

upper (Slave) die, and the DO links allow the Slave die to return data back to the Master. The design of this silicon demonstrator test-chip, including the integration of the CoDAPT links as part of the main SoC bus is presented in the following sub-sections.

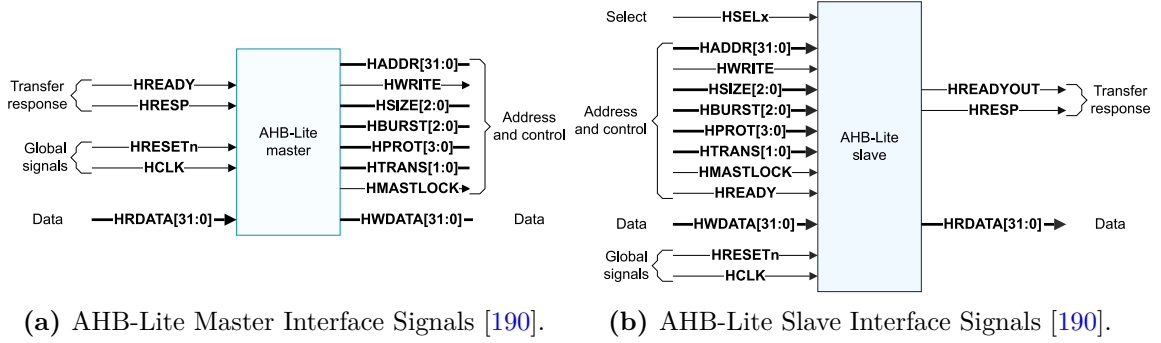


Figure 6.12: Diagrams showing the block-level interface signals for AMBA AHB-Lite Master and Slave Peripherals (Reproduced from [190]).

6.4.1 ICL AHB-Lite Bus Integration

The Advanced Microcontroller Bus Architecture (AMBA) AHB-Lite Protocol is one of the most widely used bus protocols for integrating IP block within SoCs, making it an attractive choice for adoption in this thesis. When compared with other microcontroller bus interfaces such as AXI, AHB is usually preferred for IoT applications owing to its reduced complexity (offering only single channel operation), whilst still supporting burst transfers and pipelined operation (unlike APB) [186]⁸. The Arm Cortex M0 includes an AHB-Lite bus as standard, as do a range of other designs typically used for IoT applications, such as image sensors [187], encryption accelerators [188] and memories [189]. This section presents the design of a Wireless Bus Interface (WBI) to allow the CoDAPT link to integrate seamlessly with such IP blocks through the AHB-Lite interface, thereby facilitating standardised integration of disparate elements using the CoDAPT ICL presented.

The fundamental operation of any AHB-Lite transfer occurs in two phases; the *address* phase, and the *data* phase [190]. The operation of these two phases (with reference to the AHB-Lite Master and Slave interface signals shown in Figure 6.12a and 6.12b respectively) is as follows. The HWRITE signal controls the direction of transfer, with HWRITE=1 indicating a write transfer from the Master to the Slave (such that the Master generates the data during the *data* phase), and HWRITE=0 indicating a read transaction from the Slave to the Master (such that the Slave generates the data in the *data* phase). The address of the transaction is specified in the 32-bit HADDR signal, and the HRDATA and HWDATA signals communicate the read and write data respectively (each of these signals are also 32 bits wide). The speed of a transfer is governed by HCLK, as long as HREADY is asserted. Peripherals may de-assert this signal at any time to temporarily pause the data transfer process [190].

To map this operation to the CoDAPT link, the WBI presented in this chapter ‘packetises’ each transaction using the packet structure shown in Figure 6.13. Here, the first 4 bits

⁸As only a single bus Master is required by the application in this chapter, the AHB-Lite protocol is adopted (a simplified version of the AHB specification that does not support multi-master operation).

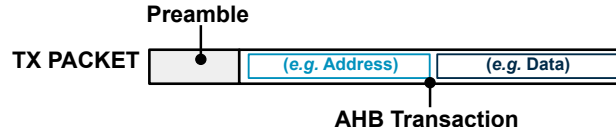


Figure 6.13: Illustration of the packet structure used in the WBI.

contain a defined preamble that indicates the start of the reception; preamble transmission is necessary when using CoDAPT as, to ensure power is being delivered, the CoDAPT link must be constantly in the transmit state. The use of a distinctive preamble, therefore, demarcates the beginning of each data transmission. Following this, the AHB transaction data is sent serially, from a First-In-First-Out (FIFO) buffer. Where multiple parallel links are included on the same chip (as in this work) the FIFO is multiplexed; for example, assuming two CoDAPT links and a four bit preamble⁹, in a 64 bit transaction, $\text{TRANS}[63:0]$, the packet sent on link 0 will be $\{\text{PREAMBLE}[0:3], \text{TRANS}[0:31]\}$ and the packet sent on link 1 will be $\{\text{PREAMBLE}[0:3], \text{TRANS}[32:63]\}$.

This clearly requires the data-rate of the inductive channel to be greater the data-rate of the bus, however, many IoT applications run tens of Megahertz frequencies, meaning that the bandwidth of the CoDAPT link presented in Section 6.2 is ample to support these applications, even with a single channel. Building upon this serialisation, each of the features of the AHB-Lite protocol can then be managed at the die-level with the WBI facilitating the local handshaking (for example de-asserting the HREADY signal when the FIFO is full and managing request and acknowledge signals). The flow diagram in Figure 6.14 illustrates this operation for a simple write transaction, which occurs as follows:

- Initially, to enable power to the Slave, the bus Master writes to the WBI control/status registers to enable power-delivery and clock-transmission to the bus Slave. This results in a sinusoid being transmitted through the CoDAPT channel which is rectified and regulated by the Slave die to provide power for the wireless peripheral.
- The Slave is now powered-on and waiting for a transaction to occur. Once in this state the bus Master can write-to the bus Slave peripheral using the standard AHB protocol. To do this, the bus Master selects the bus Slave peripheral and sends address and data bits over the AHB bus with the WBI Controller managing the handshake signals.
- The WBI Controller then assembles a wireless packet with the structure shown in Figure 6.13, by prepending the preamble to the AHB write address and write data bits. If the AHB speed is slower than the wireless link frequency, FIFOs can be used as buffers.
- Once the packet has been assembled, it is transmitted (preamble first) through the CoDAPT link using the BPSK scheme.

⁹This is the preamble length used for practical evaluation in Section 6.4.3.

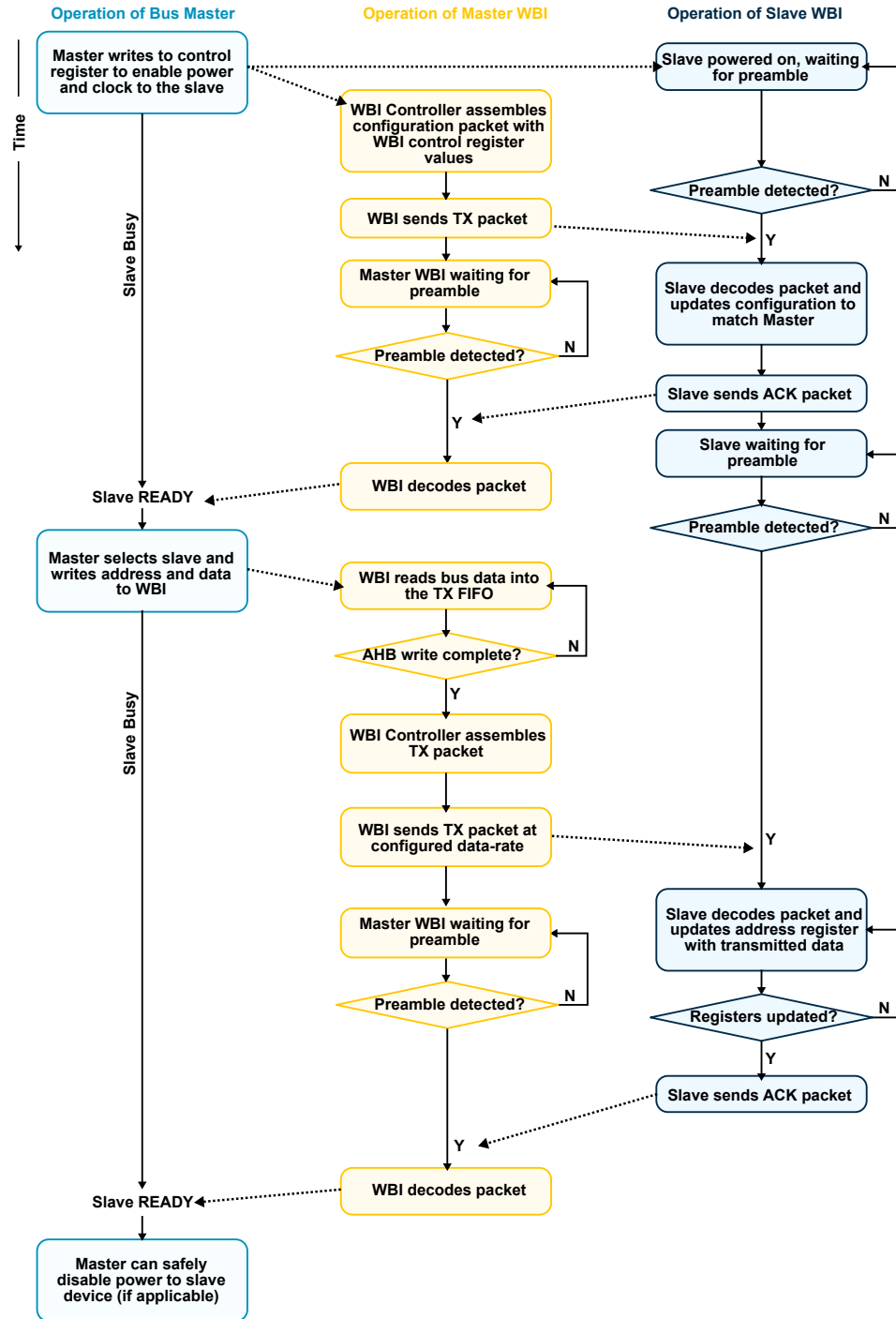


Figure 6.14: Flow diagram illustrating the operation of the proposed Wireless Bus Interface (WBI) during a Master to Slave write transaction.

- When the bus Slave recognises the phase shift pattern corresponding to the preamble in the received sinusoid (signalling the start of the transmission), the wireless bus Slave begins to store the packet in the RX FIFO and decode it. Once the full packet has been decoded, the write request is actioned in the wireless Slave peripheral, and an

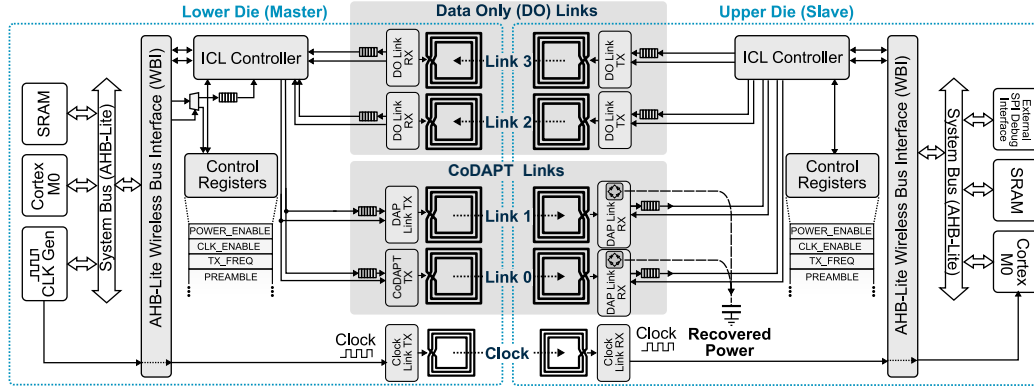


Figure 6.15: Architecture of the 2-tier 3D stacked Arm Cortex M0 SoC presented in this section with $2 \times$ CoDAPT uplinks (for data and power transmission from the lower (TX) die to the upper (RX) die), $2 \times$ Data only downlinks (for the upper die to communicate data back from the upper die to the lower die, and a clock link (for precise clock synchronisation between dies). The clock link design used here is the WiSync link presented in Chapter 5.

acknowledgement packet is transmitted back to the Master.

- Once this acknowledgement has been received by the Master, the WBI asserts the bus READY signal, indicating that it is ready for further transactions. At this point, the transaction is complete and the wireless Slave peripheral can be safely powered-off, if applicable.

6.4.2 System Design

Using this WBI design to provide an interface with the AHB-Lite bus, the detailed target system architecture for the test-chip presented in this section is shown in Figure 6.15. As discussed above in Section 6.4, the architecture contains five links: $2 \times$ CoDAPT links to transmit power and data from the lower (Master) die to the upper (Slave) die, $2 \times$ Data Only (DO) downlinks to allow the Slave communicate back to the Master (these DO links use the same design as the CoDAPT links with the omission of the power rectification and recovery circuits as the Master die has its own wire-bonded power source) and $1 \times$ clock link. For implementation of the WBI a 36-bit packet structure was used with a 4-bit programmable preamble, controlled by the WBI status registers, as shown on the figure.

To demonstrate that the CoDAPT design can operate from a low-quality (and hence low-cost) clock source, a programmable Ring Oscillator (RO) is used to generate f_{hf} in the Master die. The output of this RO is then divided using programmable clock dividers and distributed to the other elements in the SoC. In the Slave die, the Clock Distribution Network (CDN) is generated from the output of the WiSync link, meaning that the whole system (including Cortex M0 + bus) operates from the wirelessly transmitted clock.

Figure 6.16 shows the physical design and layout of this test chip, implemented in 65nm

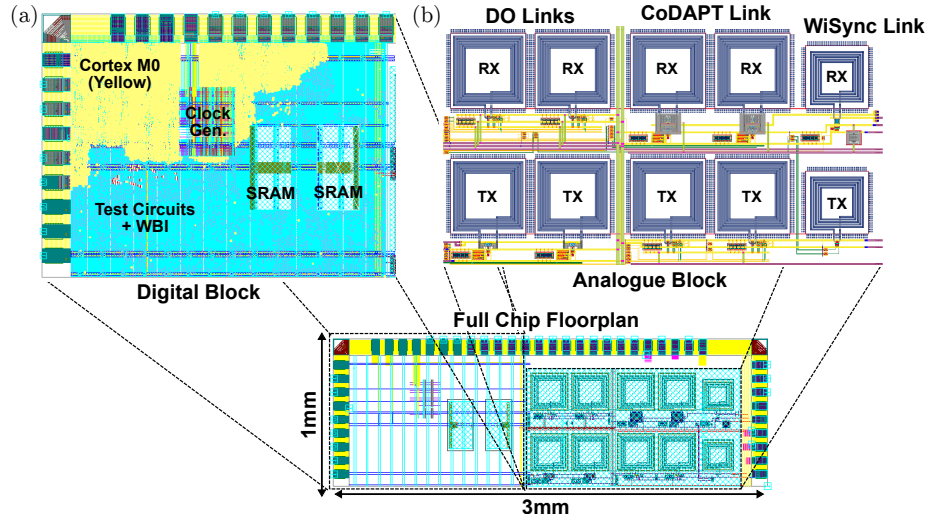


Figure 6.16: Physical layout of the proposed 3D stacked SoC design using CoDAPT inter-tier links highlighting (a) the digital block (containing the Cortex M0, SRAM and WBI), and (b) the analogue block (containing the CoDAPT, DO and WiSync links).

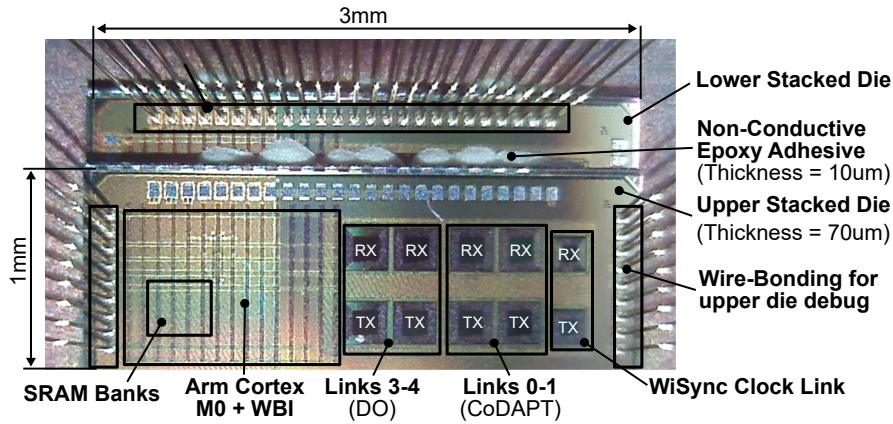


Figure 6.17: Die micrograph showing the 2-tier test chip, highlighting the key system components including the CoDAPT links, Arm Cortex M0 + WBI, and the WiSync link.

CMOS technology. As shown, the full chip (including IO pads for power and debug purposes) is 1mm × 3mm. Figure 6.16 also highlights the size of the key elements of the SoC, including the Cortex M0, SRAM banks, test circuits and the analogue block (containing the CoDAPT, DO and WiSync links). The following sections outline the practical measurement results obtained when evaluating the performance of this test-chip.

6.4.3 Experimental Results

Figure 6.17 shows a micrograph of the manufactured, stacked and packaged test IC with the key components of the system labelled (specific details regarding the fabrication and assembly of this test-chip are presented in Appendix D). As in the previous chapter, two identical dies were stacked in a face-to-back arrangement with a lateral offset of 400 μm (such

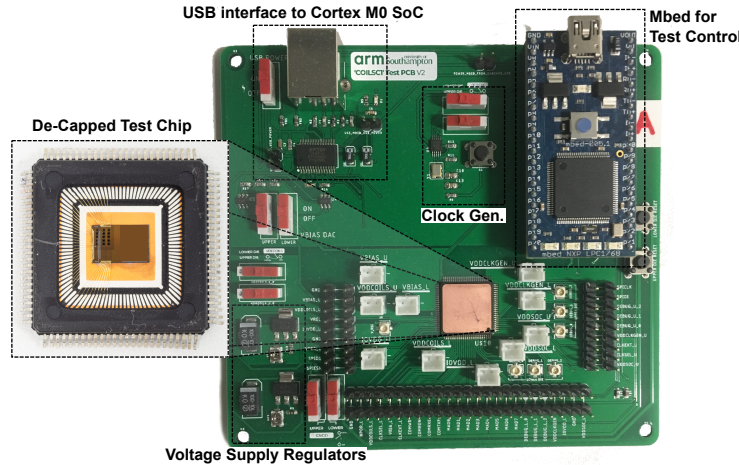


Figure 6.18: Test PCB used for evaluation of the presented 3D-stacked Cortex M0 SoC.

that the TX and RX channels align). To minimise the assembly cost of the system, only standard die-level thinning was used (allowing the upper die to be thinned to a minimum final thickness of $70\text{ }\mu\text{m}$), resulting in a total communication distance of $80\text{ }\mu\text{m}$ (including the epoxy adhesive used for assembly).

For testing purposes, the test chips were packaged in a 100 Quad-Flat Package (QFP), which was directly mounted to a custom test-PCB for evaluation, as shown in Figure 6.18. The test board contains regulators for each of the supplies within the system including the main supply voltage, V_{DD} (1.20V), and the CoDAPT link supply voltage, V_{DD_COILS} (2.5V). The PCB also contains an Mbed microcontroller for automating testing. As the presented SoC contains a Cortex M0 microcontroller, this was used for BER measurements with data patterns being generated in the MCU software and written wirelessly to addresses in the opposite die, across the AHB-Lite bus, using the WBI.

The digital system (Cortex M0 + WBI) was able to run robustly at frequencies up to 110MHz (approximately 8% of the maximum CoDAPT link frequency) and the WBI demonstrated good functionality across a range of test cases, allowing data to be written and read using the AHB-Lite interface. The following sub-sections revisit the simulated results from Section 6.3, presenting silicon evaluation of the CoDAPT links including power delivery performance (Section 6.4.3), and data delivery performance (Section 6.4.3) as part of the presented SoC.

Power Delivery Performance

Initially, the power delivery of the proposed CoDAPT links was evaluated. Measurements of power delivery in this section were taken across a $100\text{ }\Omega$ 250fF static load at P_{OUT} , and the clock frequency was varied using the programmable on-chip clock generator (with the precise frequency then being confirmed empirically using an oscilloscope). Figure 6.19 presents the results from these experiments showing how the power output of the two CoDAPT

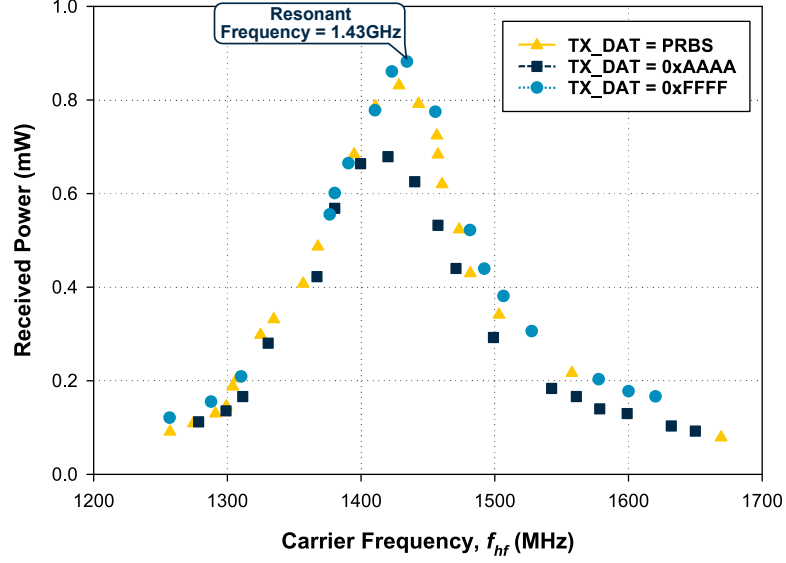


Figure 6.19: Measured power delivery performance of the CoDAPT Links (0-1) with a 100Ohm 250fF load.

links varies with frequency. As in Section 6.3.4, results are presented for three cases: (i) whilst transmitting a constant data pattern (0xFFFF) which corresponds to the best-case efficiency (as no phase-shifts are introduced in the High Frequency (HF) carrier signal), (ii) whilst transmitting a PRBS, corresponding to the average case efficiency, and (iii) whilst transmitting an alternating data pattern (0xAAAA) which corresponds to the worst-case efficiency (as phase-shifts are being introduced in the HF carrier signal for every bit).

The results clearly show the resonance of the link, with a peak power delivery occurring at a frequency of 1.43GHz. Whilst this represents a slight deviation from the theoretical target resonance of 1.50GHz (from Section 6.3), the behaviour of the design very closely matches the simulation models with only a 4.7% disparity¹⁰. At this resonant frequency, the maximum power delivery of the two CoDAPT links was found to be 0.88mW, with an average power delivery of 0.83mW whilst transmitting a PRBS.

Whilst a full comparison to the state-of-the-art works is not provided until later in Section 6.4.3, this represents a competitive figure, translating to a power delivery density of 7.1mW/mm²; prior work in this domain, which typically use much thinner substrates (and hence operate over much shorter distances) achieve WPT efficiencies of between 3.37mW/mm² [61] and 33.3mW/mm² [59]. These previously reported links, however, only perform *power* transfer, which means that, when considering the additional area that would be required for data transmission (which is included as part of the CoDAPT transceiver) the actual area utilisation is much higher. Again, the measured WPT performance agrees

¹⁰This small mismatch is likely due to slight variations in the BEOL stack-up or assembly, causing the resonant frequency to shift.

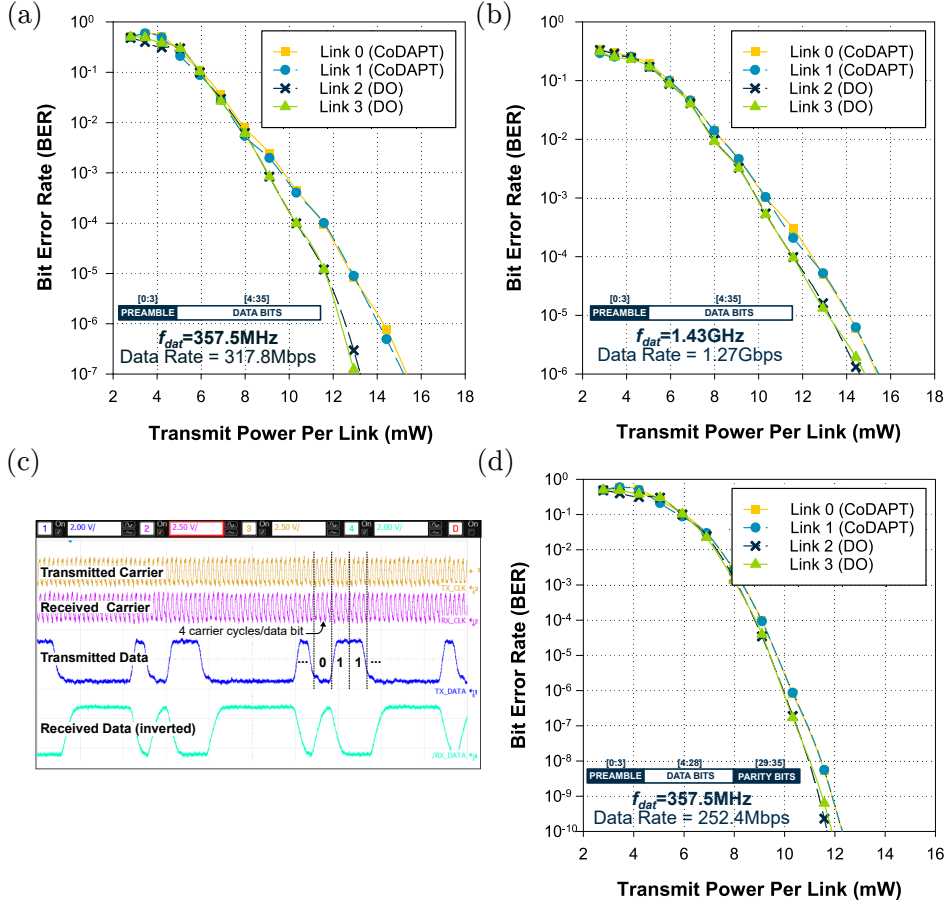


Figure 6.20: Bit Error Rate (BER) as a function of transmit power for CoDAPT links 0 & 1, and Data Only (DO) links 2 & 3 measured in (a) standard operating mode ($f_{DAT}=357\text{MHz}$), (b) high-bandwidth mode ($f_{DAT}=1.43\text{GHz}$), and (d) standard operating mode with software-based ECC. (c) Oscilloscope capture showing the operation of the CoDAPT link in standard operating mode.

very closely with the simulated results presented above in Table 6.1. This provides high confidence that with additional die thinning, WPT efficiencies up to $24.16\text{mW}/\text{mm}^2$ could be achieved (as per Table 6.1) using the same CoDAPT design.

Data Delivery Performance

Following this, the data delivery performance of the proposed CoDAPT links was evaluated on the test-chip. As 1.43GHz was found to be the resonant frequency of the system, the measurement results presented in this section were taken with $f_{hf} = 1.43\text{GHz}$. For evaluating the BER of the CoDAPT links, random data patterns were generated by the Cortex M0 microcontroller. These were then written to memory addresses in the opposite die across the CoDAPT links using the WBI. The written data was then read out directly using the SPI debug interface (*c.f.* Figure 6.15).

Figures 6.20 (a) - (d) show the results of these measurements, demonstrating that the presented CoDAPT link can achieve a $\text{BER} < 10^{-7}$ using the BPSK modulation scheme presented in this chapter, whilst operating from the main SoC supply voltage. Figure 6.20 (a) and Figure 6.20 (c) capture the operation of the CoDAPT link in the standard operating mode. As shown on the oscilloscope capture (Figure 6.20 (c)), $f_{\text{hf}} = 4 \times f_{\text{DAT}}$ and therefore four phase samples are acquired per bit (by the majority vote receiver). Figure 6.20 (b) shows the BER versus link transmit power (varied by tuning VDD_COILS) when in the high-frequency operating mode (where $f_{\text{hf}} = f_{\text{DAT}}$, and therefore one phase sample is acquired per bit). As expected, the performance of the link is slightly higher in the standard operating mode (equating to a lower BER), reaching a BER of 10^{-7} at a transmit power of 15.2mW (where the high-frequency link only reaches a BER of 10^{-6} at iso-power consumption).

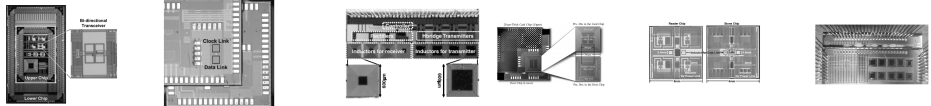
As shown on Figure 6.20, the performance of the Data Only (DO) downlinks is measured to be slightly better than the performance of the CoDAPT uplinks, despite both using the same design. The DO links, however, do not include the power rectification and measurement circuits (as the lower, Master, die has its own wire-bonded power source), and therefore the slight discrepancy is likely due to the additional noise (*e.g.* switching noise or substrate noise) introduced by this power-related circuitry.

For applications that are very sensitive to vertical read/write errors, additional error correction can be implemented in at the software level, running on the MCU. An example of this is shown in Figure 6.20 (d). Here, Hamming(32,26) SEC-DED, Error Correction Code (ECC) is being run as part of the MCU data transfer application. Using this approach, a $\text{BER} < 10^{-10}$ is achieved at a data-rate of 252Mbps per link ($f_{\text{DAT}} = 350\text{MHz}$), representing a significant (3 orders-of-magnitude) BER improvement for the additional $\sim 21\%$ data-rate overhead.

Comparison with State-of-the-Art

Summarising these results, Table 6.2 provides an overview of the CoDAPT link performance when compared to state-of-the art works in this domain. Because concurrent data and power transmission through a single link is a new concept, there are no works which can directly provide a like-for-like comparison. As such, Table 6.2 draws comparison with works [191] and [133] (leading publications demonstrating *data* delivery using ICLs), [59] (leading publications demonstrating wireless *power* transmission using ICLs), in addition to [60], [61] (which investigate wireless data and power delivery on the same chip, albeit through separate channels).

As shown in the table, CoDAPT achieves a $7.8\times$ reduction in area per link compared with existing implementations where both power and data are transmitted wirelessly on the same chip (making this the smallest ever reported WPT link). The design also achieves



	Yoshida <i>et al.</i> [191]	Miura <i>et al.</i> [133]	Yuan <i>et al.</i> [59] ('08)	Radecki <i>et al.</i> [60]	Y. Yuxiang <i>et al.</i> [61]	This Work
	Data Only Links		Power Only Links	Data And Power Links		
Approach	NRZ-Encoded Data	NRZ-Encoded Data	80MHz WPT (half-wave rectifier)	Nested Power & Data	Time-Interleaved Power & Data	Concurrent Power & Data
Die Thickness	10um (+40um glue)	20um (+2um glue)	10um	20um	200um	70um (+10um glue)
Channel Diameter	100um	110um	600um	700um	2mm	250um
Channel Bandwidth	2.0Gbps (Duplex)	1.1Gbps (Simplex)	-	6.0Gbps (Simplex)	0.15Gbps (Simplex)	1.27Gbps (Simplex)
Power/Area	-	-	8.3mW/mm ²	20.8mW/mm ²	3.37mW/mm ²	7.1mW/mm²
Data/Area	20.0Gbps/mm ²	90.9 Gbps/mm ²	-	12.2 Gbps/mm ²	0.036 Gbps/mm ²	20.32 Gbps/mm²
Technology	180nm	65nm	180nm	65nm	180nm	65nm
System Integration	Inductive Links Only	Inductive Links Only	Inductive Links Only	Inductive Links Only	ROM Interface	Full SoC

Table 6.2: Summary table comparing the measured performance of the WiSync links (in the 2 tier SoC presented in this chapter) with previously reported state-of-the-art works in this domain.

a $1.7\times$ bandwidth improvement (per unit area), whilst remaining competitive in terms of WPT/mm² compared to these prior works ([60] and [61]). This is also the first instance of concurrent wireless power and data transmission being performed in a 3D-IC and the first work to perform integration of inductive coupling links using a standard SoC bus protocol.

When compared to *combinations* of prior works which focus on wireless power transfer and inductive data communication separately, the proposed approach also still exhibits significant area savings. As an example, combining the data link presented by in Miura *et al.* in [133] and the inductive power link presented in Yuan *et al.* in [59] would result in a total silicon of area utilisation of around 0.38mm², with a power delivery of 3mW and a bandwidth of 1.1Gbps (equating to a power density of 7.9mW/mm² and a bandwidth/area of 2.9Gbps/mm²). This is without accounting for any additional power or area which may be expended mitigating crosstalk between the power and data channels.

One of the drawbacks of the presented scheme however, is that presently, it can only achieve simplex (uni-directional) data transmission (like [60] and [61]). Whilst the CoDAPT design could theoretically be extended to operate in a half-duplex mode for data communication (provided sufficient power had been previously been transferred to, and stored in, the RX die), to achieve duplex communication with the existing CoDAPT implementation requires the introduction of a separate data downlink, thereby reducing the effective *duplex*

bandwidth/area to 10.15Gbps/mm². As shown in Table 6.2, some prior art (Yoshida *et al.* [191]) has demonstrated the ability to perform duplex data communication through a single 100 μ m diameter ICL. Combining this bi-directional link presented by Yoshida *et al.* [191] and the wireless power link presented by Yuan *et al.* [59] would yield a duplex bandwidth/area of around 5.4Gbps/mm², bringing the overall area efficiency closer to that achieved by the presented scheme. Considering this, the CoDAPT design would be best suited to applications where the majority data flow is in the same direction as the power flow, for example in a peripheral which is writing data to a wirelessly powered NVM (*e.g.* through a Direct Memory Access (DMA) interface) that can be separately read by another SoC element. As energy efficiency is of very high importance for IoT sensor devices (as discussed in Section 1.4, Chapter 1) the relatively low WPT efficiency achieved by CoDAPT (2-13%, depending on die thickness) means that the presented system is also better suited to cost-constrained IoT edge devices, or for integrating peripherals that are only powered intermittently.

6.5 Summary

This chapter has presented the CoDAPT transceiver, the first ever ICL to perform wireless data and power transmission concurrently through a single channel for the purposes of 3D integration. The proposed design uses a BPSK scheme (where the *amplitude* of the carrier signal is used for wireless power transfer, whilst the data is modulated onto the carrier signal in terms of *phase*) to achieve up to 1.51mW of wireless power transfer (per 250 μ m link) whilst simultaneously communicating data at 1.5Gbps.

The CoDAPT transceiver was also experimentally validated in a 2-tier 3D stacked Arm Cortex M0 SoC, where the wireless links were used to form the main system AHB-Lite bus (through use of a custom wireless bus interface, proposed in this chapter). Evaluation of the test-chip demonstrated that the CoDAPT links operate robustly, achieving a BER < 10⁻⁷, a data rate of up to 1.27Gbps/link and a power delivery of 7.1mW/mm². When compared to previously published ICL implementations, this makes it the smallest ever inductive power link. The CoDAPT transceiver also achieves a 1.7 \times bandwidth improvement (per unit area) compared to similar prior art, whilst remaining competitive in terms of WPT/mm². Additionally, it is the first integration of any ICL as part of a 3D SoC using a standard bus protocol.

Chapter 7

Conclusions and Future Work

3D integration is expected to be a key enabling technology in the Internet of Things era and beyond, allowing disparate processing/sensor memory dies (each of which may be fabricated in different technologies) to be stacked and interconnected vertically, resulting in smaller, more efficient, heterogeneous devices [5]. However, existing approaches to 3D integration (including flip chip bonding, 3D-SiP assembly using wire-bonds, and TSV-enabled 3D-ICs) each have significant drawbacks associated with them (stack-height, cost and process availability respectively).

This thesis has explored the possibility of using ICLs to address these challenges and establish a low-cost, scalable 3D integration solution for IoT applications. The use of ICLs means that 3D integration can be performed very cheaply in terms of design and manufacture (as they do not require the additional fabrication stages associated with TSV processing, nor do they require TSV-aware Electronic Design Automation (EDA) tools). Further to this, once manufactured, dies can be simply picked and stacked with coarse placement accuracy using adhesive. This makes them an attractive option for IoT applications which are driven by cost and design-time, rather than performance.

Existing research into 3D integration using ICLs has mostly focussed on high-bandwidth applications, such as stacked DRAM [43, 49] and image sensors [124, 125]. The application of wireless integration in the IoT context, however, requires consideration of a range of new design factors which have been explored in this thesis. One primary focus when applying this technology within the IoT context is *energy* efficiency. IoT devices are often battery powered, or source energy from their environment using energy harvesting. This means that maintaining low energy consumption is of paramount importance.

Chapter 3 explored low-energy ICL transceiver design, focussing on how the encoding schemes of existing transceivers can be adapted to reduce energy consumption. A novel, time-domain coding inductive transceiver was presented in this chapter, which significantly reduced the overall ICL energy by using spike-latency encoding. The proposed transceiver was modelled mathematically, simulated in 0.35 μ m, 65nm and 28nm CMOS technologies,

and experimentally validated in a 2-tier 3D stacked silicon test chip. Silicon evaluation demonstrated an energy of 7.4pJ/bit, representing a significant reduction ($>13\%$) when compared to previously reported schemes. Simulated results show even greater energy savings ($>28\%$) at more advanced technology nodes, representing an important progression towards realising low-energy ICL transceivers suitable for use in IoT applications. This also led to the filing of the first patent based on this work, “*Adaptive Coding for Wireless Communication*” [97].

Another important focus in the IoT domain is maintaining short design cycles. IoT devices are usually application-specific and low-cost. Due to this, maintaining a quick design time is important to minimise costs and reduce time-to-market. For selecting the layout of the inductors used to form ICL channels, prior works have adopted a manual optimisation flow, in conjunction with Finite Element Modelling (FEM) (to evaluate the EM performance of each layout) [52]. However, analysis of inductor geometries using FEM can take several hours (even for a single layout), so when combined with a manual tuning and optimisation process, this makes the inductor design process very tedious and time consuming (if not computational impossible). To address this challenge, Chapter 4, presented: (1) a rapid solver for evaluating inductor layouts using strictly solvable mathematical expressions, and (2) a high-speed optimisation algorithm for determining best-performing coil pairs. These two contributions are combined as a CAD-tool for Optimisation of Inductive coupling Links for 3D-ICs (COIL-3D). Results demonstrated that COIL-3D achieves an average accuracy within 7.8% of finite element tools whilst consuming a small fraction of the time ($1.5 \times 10^{-3}\%$), thereby significantly reducing the ICL channel design time. The COIL-3D optimised inductor layouts also yielded significant performance (up to 49.9% bandwidth improvement) and power (up to 8.1% power improvement) benefits, when compared with layouts used in prior ICL implementations.

As highlighted in the literature review, to realise a truly scalable and low-cost 3D assembly approach, it is desirable to move towards *fully wireless* 3D integration. To remove galvanic connectivity between dies in the stack altogether, requires the ICLs to be extended beyond *data* transmission, to also perform wireless *power* and *clock* transmission. Towards this goal, Chapter 5 presented a wireless clock link design (WiSync) for use in 3D-stacked integrated circuits using ICLs. The proposed WiSync link was evaluated using commercial EM and circuit simulation tools and demonstrated the ability to broadcast a clock signal simultaneously between five stacked silicon tiers, with less than 61ps of clock skew. The proposed design was also implemented in a 2-tier 3D stacked silicon test chip where a study of lateral die-to-die stacking alignment vs. BER was performed. The presented WiSync design tolerated up to $\pm 10\mu\text{m}$ of stacking misalignment, without significant degradation in BER performance ($<10\%$), demonstrating that the proposed wireless assembly approach can be handled by die-bonding machines used in existing low-cost packaging flows for die

attach applications.

Chapter 6 then explored wireless *power* delivery using ICLs. To avoid the large area overheads associated with separate wireless power transfer (WPT) and data links, this chapter explored the possibility of performing data and power transmission concurrently, through a single ICL channel. The proposed Concurrent Data and Power Transfer (CoDAPT) transceiver design was experimentally validated as part of a Cortex-M0 SoC in 65nm CMOS technology, and achieved 20.3Gbps/mm² data, and 7.1mW/mm² power transfer through a 250 μ m diameter channel (sufficient to power a full commercial subthreshold MCU running at 42MHz, with only one ICL [185]). This makes it the smallest ever reported inductive data and power link. Further to this, the simulation results presented in Chapter 5 indicated that with further wafer-level thinning, additional wireless power transfer up to 2.0mW/link would be possible using the same CoDAPT design. This represents an ample power budget for most IoT applications, unlocking the ability to construct stacked 3D systems using fully wireless assembly.

Finally, for IoT applications, maintaining customisability is important. The overall goal of this project was to explore ICL-based 3D integration as a low-cost way of bringing together heterogeneous logic/memory and sensor dies. Towards this goal, Chapter 6 also presented a Wireless Bus Interface (WBI) for integrating ICLs using the commercial AHB-Lite bus protocol. This is a standard protocol used by most IP vendors, allowing straightforward ‘*plug-and-play*’ integration of disparate pre-designed SoC elements (which may be fabricated in a range of different process technologies). The proposed WBI uses a 36-bit packet structure, allowing memory mapped peripherals (that exist in separate dies) to be addressed as if they are on the same chip/bus. This is the first ever instance of an SoC with a wireless bus, and led to the filing of the second patent based on this work, “*A Pseudo System-on-Chip Architecture Incorporating Wirelessly Connected Bus Slaves*” [97].

7.1 Research Questions

This thesis aimed to address the research questions outlined in Chapter 1. Direct answers to each of these questions are enumerated below.

1. *How is it possible to reduce the energy consumption of existing ICL transceivers for use in IoT devices?*

As summarised in the literature survey in Chapter 2, existing ICL transceiver designs are typically focussed on maximising bandwidth, which is of secondary importance when compared with energy, for IoT applications. Chapter 3 demonstrated that one way of reducing energy consumption is to adapt the transceiver design to use time-domain encoding techniques (where the data is represented by the latency between TX pulses, hence reducing number of pulses and overall energy consumption). Although

the use of a time-domain coding transceiver results in a slightly reduced bandwidth, the energy per bit can be significantly reduced. This is because the TX pulse energy is greater than the digital logic energy used for implementing the encoding scheme.

2. *How can the search process for finding optimised inductor geometries for inductive coupling link applications be automated? Subsequently, what techniques can be used to evaluate a given inductor geometry faster than using finite element modelling?*

To address this research question, Chapter 4 presented an automated optimisation flow for ICL inductors that limits the search space based upon the inductor's fill-factor (the ratio of the overall inductor area to the inner gap area) to speed up execution. This technique, combined with a selection of other optimisations, reduced the number of geometries that must be considered before reaching a finalised design from several million, to around 1500 (for the example case presented in Chapter 4).

To speed up the evaluation process (when compared with using finite element methods), Chapter 4 also presented a set of strictly solvable mathematical expressions to approximate the performance of a given pair of inductors from their physical layout parameters using an R, L, M, C model. The proposed expressions take only a small fraction of the time that full-wave finite element modelling takes (1.5×10^{-3}) to execute, but still maintain an average accuracy within 7.8%, significantly speeding up the ICL design process.

3. *Is it practical to design an ICL transceiver for use in 3D stacked ICs that performs both wireless data transfer and wireless power transfer?*

This research question was explored in Chapter 6, concluding that simultaneous wireless data and power transmission can yield significant assembly cost benefits when compared with performing power delivery using wire-bonded links, and significant silicon area reductions ($7.8\times$) when compared with performing wireless power and data transfer separately. This, however, comes at the cost of reduced power efficiency (compared with wire-bonded solutions), making Concurrent Data and Power Transfer (CoDAPT) best suited for applications where cost is of paramount importance, or for integrating dies which are only powered intermittently.

To evaluate the practicality of CoDAPT, Chapter 6 presented a CoDAPT transceiver that uses a BPSK modulation scheme to deliver data and power simultaneously through a single inductive channel. This enables the possibility of *fully-wireless* stacking, with a reduced silicon footprint when compared with prior solutions (that use separate inductive channels for data and power transmission). The proposed transceiver was validated through EM and SPICE simulation in a 65nm CMOS technology and then implemented in a 2-tier 3D-stacked test-chip. Results demonstrated that the CoDAPT links operate robustly, achieving a $\text{BER} < 10^{-7}$, a data rate of up to 1.27Gbps/link

and a power delivery of $7.1\text{mW}/\text{mm}^2$.

4. *How should clock distribution be performed in a many-tier, wirelessly stacked 3D-IC?*

Chapter 5 explored clock distribution for wirelessly stacked 3D-ICs and discovered that the majority of previous work in this domain use coupled resonator circuits to synchronise the clock between tiers in the 3D stack. However, such designs typically demand high area (to achieve resonance) and are often very sensitive to PVT and assembly variations. To address these challenges, Chapter 6 proposed a wireless clock link that operates in the non-resonant portion of the frequency spectrum using a dual-mode H-bridge transmitter (that selectively transitions between continuous and pulse-based encoding).

The approach was evaluated using commercial SPICE and EM simulators, demonstrating the ability to operate between 50MHz and 2.0GHz whilst broadcasting the clock between five stacked silicon dies with less than 61ps of inter-tier clock skew. The WiSync link was also practically evaluated in a 2-tier 65nm CMOS 3D stacked silicon test-chip. Measurement results from this evaluation show an average energy consumption of 19.4pJ per clock cycle across an $80\mu\text{m}$ channel, whilst consuming only 0.0421mm^2 of silicon area.

5. *What is the sensitivity of inductive coupling links to die-to-die stacking misalignment during the packaging process?*

The research presented in Chapter 5 explored the effects of lateral stacking misalignment on channel performance (for the WiSync link discussed above). The study showed that, whilst the placement tolerance does effect BER, the presented non-coherent transceiver design (with $170\mu\text{m}$ channel diameter) could tolerate up to $\pm 10\mu\text{m}$ of stacking misalignment, without significant degradation in BER performance ($<10\%$). To compensate for stacking misalignment beyond this level, much more substantial TX power increases were required (up to a maximum of 53% for a $50\mu\text{m}$ offset).

Misalignment effects were also explored in Chapter 3 where the spike-latency encoding transceiver design was introduced. Measurement results from this separate study indicated that the proposed spike-latency transceiver (using a $250\mu\text{m}$ channel inductor in $0.35\mu\text{m}$ technology) could tolerate around $20\mu\text{m}$ of die-to-die misalignment, whilst maintaining performance within 10% of the optimum. From these experiments (and the simulation study presented in Chapter 5), the overall indication is that ICLs can tolerate misalignments of approximately 7% of their channel diameter without significant performance degradation. Assuming a pick-and-place accuracy of $\pm 10\mu\text{m}$ at the assembly stage, a 10% design margin (in terms of transmit power) appears to be sufficient to ensure robust operation across all cases.

To overcome any larger variations in assembly quality, Chapter 3 also presented a tuneable current driver circuit which allows the TX current to be finely adjusted post-assembly. Using this approach avoids the need for over-provisioning, allowing the chip to operate with ‘just-enough’ transmit power.

6. *For the IoT, customisability is important. Is it practical to design ICLs in a standard way to allow interchangeable stacking with a range of different memory/sensor/logic dies?*

As outlined in Chapter 6, one approach to interconnecting different functional units of the SoC (which may be split across different *physical* dies) is using a standard bus protocol. The advantage of using an existing bus standard is that, even if functional blocks are split across separate physical semiconductor dies, they can be addressed from software on the main system bus, as if they were fabricated in the same chip (thereby enabling straightforward, interchangeable stacking). Small SoCs, such as those designed for IoT applications, typically use the AHB-Lite bus protocol for this purpose. As IoT devices form the context of this thesis, Chapter 6 presented a Wireless Bus Interface (WBI) for integrating wireless links using AHB-Lite. The presented WBI uses a 36-bit packet structure with 4 preamble bits (representing an overhead of just 11.1%) and successfully enables transparent processing of AHB transactions (read/write *etc.*) across the inductive coupling link.

7.2 Future Work

Whilst the work presented in Chapters 3-6 represents a significant step towards enabling low-cost, fully-wireless 3D integration using ICLs, it has also presented some interesting questions that could be explored in future research. This section outlines these potential avenues for further work. In particular, four areas are discussed: (1) Security for ICL-based 3D-ICs (Section 7.2.1), (2) Thermal management for ICL-based 3D-ICs (Section 7.2.2), (3) Interference effects of ICLs (Section 7.2.3), and (4) Channel area utilisation in ICLs (Section 7.2.4).

7.2.1 Security for ICL-based 3D-ICs

The first interesting avenue for future work exploration relates the *security* implications of incorporating wireless links within an IC. Although it can be argued that including ICLs does not increase the susceptibility to attack when compared with some existing solutions (for example wire-bonded 3D SiPs, as, if the attacker has physical access to the chip, it is equally difficult to probe an ICL as it would be to probe a wire-bond), enabling wireless access to any part of an IC always raises security concerns. These fall under two main categories:

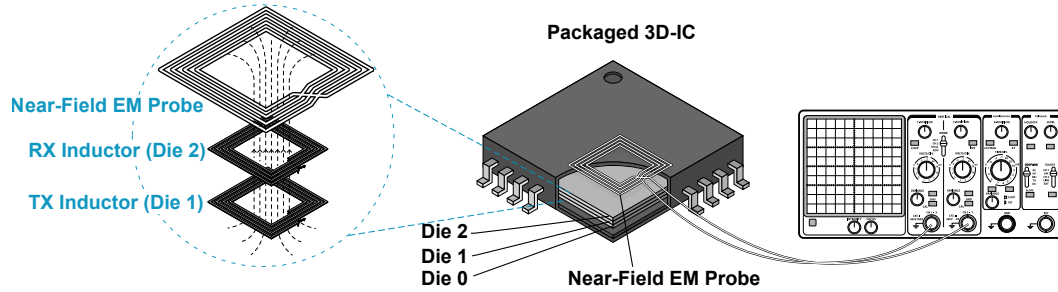


Figure 7.1: Illustration of potential security attack mechanism when using ICL-based 3D integration. Here the wireless inter-tier bus is being monitored using a near-field EM probe to sense and decode the magnetic field variations.

- Denial of Service Attacks (Externally generated interference, targeted to prevent normal operation).
- Wireless Data Sniffing (External probing of the magnetic field to gain access to data which would otherwise be privileged).

Denial of Service (DoS) attacks though contactless EM interference are a realistic threat for inductive links. Simulated results presented in Chapter 3 illustrated that typical ICL field strengths are in the order of 1kA/m (*c.f.* Figure 3.12). It has already been demonstrated (in the context of MRAM security) that fields $1.5\times$ this strength can be generated within a chip via external means [192], and therefore it is possible that a malicious third-party could deliberately generate noise within the ICL channel, resulting in bus accesses being blocked, and hence preventing the SoC from operating. As discussed above, it can be argued that the use of ICLs does not exacerbate this problem when compared with other approaches if an attacker has *physical* access to the chip, however, the use of ICLs within the chip may mean that the surface of this attack is extended to individuals who do not have physical access, but have *proximate* access (as the interference can be administered wirelessly, albeit across a short distance). Due to this, one avenue of potential future research could explore whether such DoS attacks are possible to achieve, and if so, whether they can be mitigated (either through altering the physical properties of the semiconductor dies/packageing to provide shielding (like often used for MRAM [193]), or by performing active mitigation in the circuits).

Aside from denial of service attacks, one other security implication of introducing wireless links within a chip is related to attackers probing sensitive data wirelessly via the ICL's magnetic field as shown in Figure 7.1 (*e.g.* secure keys, or data which would not otherwise be exposed). Future research in this area could, therefore, include exploring whether such an attack is possible, and how it could be avoided (for example encrypting data that is sent through the link).

Some potential further research questions on this topic could be:

1. Is it possible to extract data from an on-chip ICL using an external near-field probe?
2. Would it be possible to generate near-field EM interference using an external probe in such a way that the ICL could not operate reliably? If so, what effect would this have on the system bus?
3. How could denial-of-service attacks be detected and mitigated in ICL channels?
4. Could data encryption be used to mitigate wireless data sniffing from the ICL channel? Subsequently, what would the power/area/performance overheads of such encryption be?
5. Would a different modulation scheme (beyond the ones discussed in this thesis) offer a higher level of immunity to security vulnerabilities?

7.2.2 Thermal Management for ICL-based 3D-ICs

The second area identified for potential future research relates to the thermal implications of using wireless assembly. Thermal management is a ‘hot-topic’ within the context of 3D integration research; the ability to combine multiple layers of active silicon within the same IC brings about much higher device densities, and hence poses new thermal challenges. Another interesting avenue for exploration would, therefore, be to investigate the thermal impacts of using ICLs, when compared with existing 3D integration approaches such as TSVs, in the context of chip temperature and reliability.

Although the device density that can be achieved using ICLs is approximately the same as that using TSVs, unlike TSVs, ICLs are unable to conduct heat. This has some disadvantages, as it means that they cannot be used to provide thermal relief in the same way that dummy thermal TSVs can [194], however, in some contexts, this may also be advantageous to prevent unwanted conduction of heat between layers of the stack. It would, therefore, be interesting to explore the effects that moving from TSVs to ICLs has on the overall IC temperature. Further to this, it may be possible to improve the thermal performance of wirelessly stacked 3D-ICs by using conductive adhesive in the assembly process (to spread heat to the edges of the stack). Exploring the possibility of this, and the corresponding effects of conductive adhesive on EM channel quality, would therefore also provide interesting avenues for future experimentation.

Finally, one of the concerns associated with chips reaching high temperatures is that thermal expansion will cause cracking/shearing/deformation within the die, or the interconnect (particularly when using TSVs [22]). As *wireless* 3D integration does not rely on mechanical connections between dies, it has been suggested that ICL-based 3D-ICs will be more robust to such faults. It would, therefore, also be interesting to compare the reliability (across multiple thermal profiles) of 3D-ICs constructed using TSVs and 3D-ICs constructed using

ICLs to explore this hypothesis further.

Related to these research areas, some example future research questions could be:

1. How does the thermal profile of 3D-ICs assembled using wireless inductive coupling links compare with 3D-ICs assembled using conventional techniques, such as wire-bonded SiP assembly, or 3D stacking using TSVs?
2. Is it possible to mitigate thermal stresses by using thermally conductive adhesive at the packaging stage? If so, what are the effects of this on ICL channel performance?
3. Is it possible to model the failure rate of a given IC over time due to thermo-mechanical stress? If so, how does this failure rate compare when using TSVs to when using ICLs?

7.2.3 Interference and PVT Variation Effects in ICLs

The third potential avenue for future work exploration discussed in this chapter relates to the effects of interference and PVT variation on the performance of inductive coupling links. When using planar inductors, the main EM radiation lobe extends vertically upwards from the centre of the inductor, meaning that lateral interference from ICLs is generally minimal, and was not found to be a significant concern for the experiments presented in this thesis (with the ICLs and other digital circuit blocks operating harmoniously on the same chip, even from the same voltage supply in Chapter 6). However, radiation from the side lobes and vertically, beyond the intended RX inductor, could potentially cause unwanted EM interference within sensitive extraneous circuits.

The effects of this interference on some circuit elements has already been investigated in prior publications (for example, Papistas *et al.* present thorough modelling and evaluation of interference effects of ICLs on the Power Distribution Network (PDN) in [195], and Niitsu *et al.* evaluate the interference of ICLs on SRAM cells in [147]). However, IoT devices often incorporate a range of other, non-digital, circuit blocks (*e.g.* wireless radios, MEMS sensors *etc.*) that may be more susceptible to EM interference. Due to this, exploring the effects of interference with these devices would be an interesting avenue for future work, particularly analogue circuit blocks that operate close to the frequency of the ICL channel.

As IoT devices are often deployed in remote outdoor locations (and therefore can operate across a wide range of ambient temperatures), it is especially important that ICL transceiver designs have high immunity to Process, Voltage and Temperature (PVT) variations. Due to practical limitations, the test-chips presented in this thesis were only manufactured in small batches and hence statistical analysis of measured process and temperature variation could not be performed. However, in future work it would be interesting to conduct a full characterisation study to evaluate the reliability of the designs presented. There may also be scope to improve the PVT resilience of the designs further, for instance by replacing

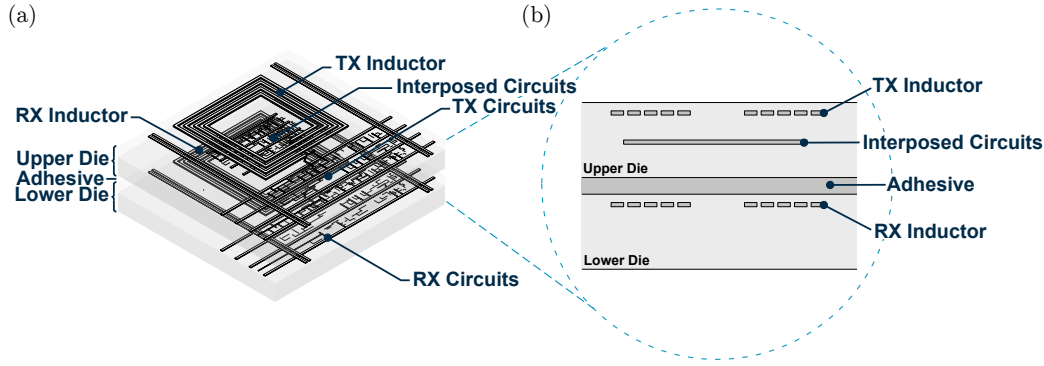


Figure 7.2: (a) Isometric and (b) cross-sectional illustration of a 2-tier 3D-IC assembled using Inductive Coupling Links (ICLs) with circuits placed in the silicon area interposed by the ICL channel.

the passive DC biasing circuits used in this thesis (which rely upon the matching of high-variability polysilicon resistors) with active variation-aware biasing circuits (*e.g.* those presented by Wang *et al.* in [196]). Some potential research questions associated with these topics could therefore include:

1. How close to the inductive coupling link can sensitive analogue circuits (for example radio transceivers, amplifiers *etc.*) be placed without interference occurring?
2. Can EM interference be limited using physical implementation techniques such as metal shielding? If so, which shielding patterns are most effective?
3. Could an interference-aware physical implementation flow be established to automatically place the inductive coupling channels to maximise reliability?
4. Is it possible to limit the coupling communication distance (*e.g.* using shielding or metallic adhesive between layers) to avoid communication with layers stacked above or below the intended recipient?
5. How robust are ICL transceivers to significant PVT variations, and how can the presented transceiver designs be adjusted to improve variation resilience, given the hostile environments which IoT devices may be deployed in?
6. Can the passive DC bias circuits adopted in this thesis (using polysilicon pull-up resistors) be replaced with active biasing circuits to minimise the effects PVT variation? If so, what are the associated area and power overheads of this?

7.2.4 Interposed Silicon Usage in ICL Channels

One other final potential avenue for future research work is exploring the extent to which the interposed silicon (between the TX and RX inductors that form the ICL channel) can be utilised for functional circuits. As the area of the inductors used for ICLs is typically

fairly significant, one criticism of using them for 3D integration is the expense of the silicon area which they occupy. For all the experiments and simulations in this thesis, it has been assumed that no extraneous circuits are placed within the ICL channel area, such that the overall footprint of the ICL is equal to its bounding-area. Although leaving the interposed silicon area empty in this way will undoubtedly result in enhanced EM coupling between each of the layers, it has been suggested that the channel area (between the TX and RX inductors) could also be utilised for active circuits, without significantly degrading the performance of the link [147]. This would significantly reduce the area *overhead* of the ICL channel. In many ways, this research avenue is linked with that discussed previously (Section 7.2.3) as it is likely that interposing active circuits within the ICL channel would result in some bilateral interference (between the circuits and the ICL, and vice-versa), however no research has been performed to quantify this interference and evaluate whether the achievable area benefits make it worthwhile.

Therefore, to explore this topic, some specific example research questions could include:

1. What is the performance impact on the ICL transceiver if digital logic cells are placed within the ICL channel?
2. Can this impact be minimised in the physical implementation flow (for example restricting routing to certain lower metal layers, or limiting routing to a certain direction?)
3. Which parts of the channel are most susceptible to interference from interposed circuits (for example, is placing circuits in the ‘eye’ of the inductor worse than under the tracks)?
4. Are interference effects static or dynamic? If dynamic, can the operating frequency selection be used to mitigate the effects?
5. Can the COIL-3D optimisation flow (presented in Chapter 4) be extended to consider interposed silicon usage?

In conclusion, the work in this thesis represents a significant step towards establishing wireless 3D integration as a low-cost 3D IC design methodology for IoT devices. It has presented: (i) a time-domain data transceiver for realising low-energy data communication (Chapter 3), (ii) an automated inductor optimisation tool to significantly reduce the design time of ICL channels (Chapter 4), (iii) a wireless clock link for inter-tier clock synchronisation (Chapter 5), and (iv) the first ever ICL transceiver to perform wireless power and data concurrently (Chapter 6). The combination of these contributions has potential to enable very low-cost, fully-wireless 3D assembly (as illustrated by the use-case example in Section 6.4 of Chapter 6) however, these advancements may also pose a new set of challenges (in terms of security, thermal management and interference) which would form interesting areas

for future research.

Appendix

Appendix A

Iso-Area Bandwidth and Energy Comparison of ICLs and TSVs

This appendix compares the energy consumption and bandwidth of Inductive Coupling Links (ICLs) and Through Silicon Vias (TSVs) in an iso-area study (for both homogeneous voltage and heterogeneous voltage applications) to support the discussion presented in Chapter 2.

As ICLs communicate data *wirelessly* through the 3D stack, they can, if required, also be designed to provide intrinsic voltage level conversion [46]. For example, when integrating dies that operate with different supply voltages, the ICL transmit and receive circuits can each operate (*generating* and *sensing* the EM field) in their own native voltage domain, without any additional voltage conversion circuitry. In contrast, TSVs rely on Ohmic electrical links between each tier to transmit data. Because of this, when performing integration of dies with different supply voltages using TSVs, additional voltage level converters must be incorporated.

This appendix compares the bandwidth and energy consumption of ICL and TSV-based vertical interfaces as part of an iso-area study. To investigate the benefits of intrinsic voltage level conversion for heterogeneous 3D integration (as discussed above), comparison is performed for: (1) Homogeneous integration of two 65nm CMOS dies (operating at 1.2V), (2) Heterogeneous integration of one 65nm CMOS die (operating at 1.2V) and one 180nm BiCMOS die (operating at 2.5 V) and (3) Heterogeneous integration of one 65nm CMOS die (operating at 1.2V) and one 0.35 μm CMOS die (operating at 3.3 V).

Section A.1 presents the architecture and modelling assumptions for the ICL-based interfaces, and Section A.2 presents the architecture and modelling assumptions when using TSVs. Comparison results for each integration scenario are presented in Section A.3, before a discussion of the results is presented in Section A.4.

Parameter	Value
n	5
w	9.0 μ m
s	0.72 μ m
D	200 μ m

Parameter	Value
R	9.1 Ω
L	48 nH
M	12.11 nH
C	389 fF

Table A.1: Physical inductor parameters. **Table A.2:** Extracted electrical parameters.

A.1 Interface using Inductive Coupling Links

ICL Channel

The most important element of the ICL-based interface is the inductive channel itself, consisting of two coupled planar metal inductors. As square-spiral coils offer the highest inductance per unit area (compared to hexagonal, octagonal and circular monolithic inductors [162]), an inductor of this type will be considered here. To ensure the accuracy of the modelling assumptions in this appendix, the silicon verified 200 μ m inductor layout from [197] is used for the ICL based design, with geometric parameters (outer diameter (D), number of turns (n), trace width (w) and spacing (s)) shown in Table A.1.

For electrical modelling, the R, L, M, C ICL channel model (proposed in [51]) is used. Here, it is assumed that each coil exhibits resistance and capacitance as well as inductance, and mutual inductance between the coils. To translate the physical coil layout into a corresponding electrical channel model, physical measurements of R and L (from the authors of [197]) and simulated values of M and C (obtained using CST MW Studio) are used, assuming a die thickness of 65 μ m (in line with the die thicknesses required for TSV-based solutions). These fitted parameters are shown in Table A.2. Using this inductor layout (for both TX and RX coils) results in a coupling coefficient, k , between 0.24 and 0.25 depending on the integration scenario.

ICL Transmitter

For the transmitter, the NRZ communication scheme discussed in Chapter 2 (presented in [51]) is assumed, with the circuits implementation shown in Figure A.1 (a). Here, as the data signal transitions from $0V \rightarrow VDD_{TX}$, a short clockwise current pulse, with duration determined by the pulse-width delay element, will flow through the coil representing a *rising* data edge. Conversely, a current pulse of the same duration will flow counter-clockwise through the coil when the data signal transitions from $VDD_{TX} \rightarrow 0V$, representing a *falling* data edge. The delay element is realised using an even (non-inverting) n -stage inverter chain, such that the current pulse duration is given by $nt_{p,inv}$, where $t_{p,inv}$ is the propagation delay of a single inverter. In order to integrate processes with different supply voltages (VDD_{TX} and VDD_{RX}) the transistors M0 and M1 are sized appropriately to ensure that the received voltage pulses are sufficiently large for detection in the receiver. The value n (corresponding to the

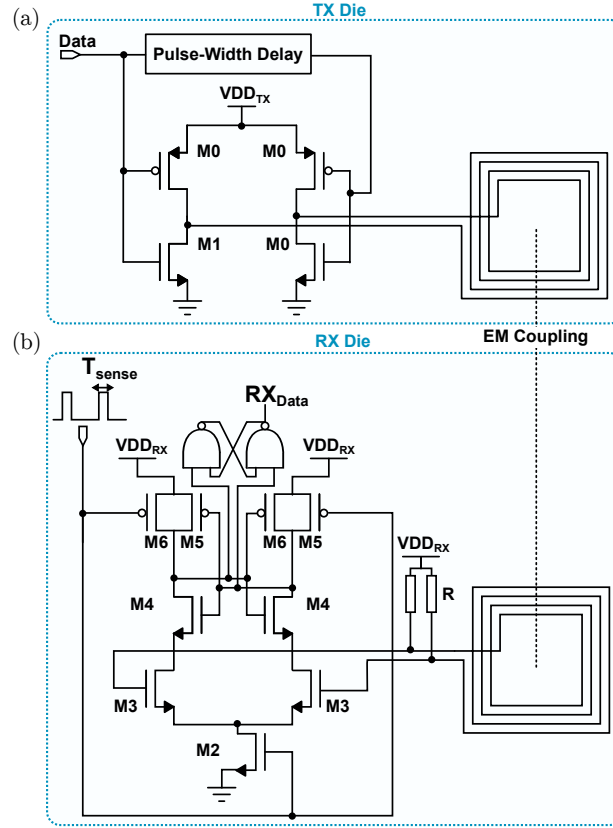


Figure A.1: Example ICL transceiver implementation used for analysis in this appendix, proposed by Miura *et al.* [51].

length of the delay buffer) is also carefully selected in correspondence with the integration scenario.

ICL Receiver

Figure A.1 (b) illustrates the utilised receiver design. Due to EM induction, transmitted current pulses will induce a corresponding voltage signal in the RX coil with a magnitude determined by the coupling coefficient (k) between the two tiers. In Section A.1, k was found to be around 0.25 and hence, from analysing the R, L, M, C ICL channel model, received voltage pulses with magnitudes of the order of 100mV can be expected. The receiver design, therefore, must incorporate amplification in order to successfully detect the received pulses. To achieve this, the low-power Sense Amplifier Flip Flop (SAFF) [198] presented in Figure A.1 (b) is adopted, consisting of a sense amplifier is coupled to a NAND SR latch. Here the transistors M3, form a differential amplifying pair determining the gain of the arrangement. The duration of the *pre-charge* and *evaluate* phases (defined by T_{sense}) are manually selected, depending on the operating frequency, to distinguish the received signal from extraneous noise.

Device	Transistor Size (Width/Length)		
	In 65nm technology	In 180nm technology	In 0.35um technology
M0	14um / 60nm	8.5um / 180nm	6.0um / 350nm
M1	10um / 60nm	4.7um / 180nm	3.5um / 350nm

Table A.3: Transistor sizing for the ICL transmitter in each technology.

Device	M2	M3	M4	M5	M6
Size (Width/Length)	0.55um / 60nm	3.9um / 60nm	0.55um / 60nm	1.2um / 60nm	1.7um / 60nm

Table A.4: Transistor sizing for the ICL receiver in 65nm CMOS technology.

For the SAFF to operate correctly, the received signal must be biased such that M3 operate in their saturation region. To achieve this, a bias voltage of $VDD_{RX}/2$ is applied through pull-up resistors (R) to maintain a high input impedance. As with the transmitter design, each of the transistors in the receiver (M2 to M6), in addition to resistors R, are sized appropriately to manage seamless voltage conversion between the two dies in the case of heterogeneous integration. Tables A.3 and A.4 document the transistor width and length selections for each of the technologies. For reasons pertaining to space in this appendix, only *heterogeneous* integrations where the higher voltage process is communicating to the lower voltage process are considered; these are most representative of real-world applications where analogue sensors communicate to digital processing dies. The methodology used here could, however, equally be applied in the opposite direction.

Combining these three elements, layout of the transceiver assumed for the ICL-based integration scenario was performed. The bounding area of the largest circuit, at the largest (0.35 μm CMOS) technology node, was found to be $200\mu\text{m} \times 260\mu\text{m}$, and so the area budget for the TSV-based comparison designs (presented below in Section A.2) was defined as 0.052mm^2 for fair comparison.

A.2 Interface using Through Silicon Vias

In addition to the ICL-based designs, to establish a benchmark for comparison, a selection of TSV-based interfaces are also considered. As TSVs have a smaller area footprint than ICLs, to maintain fair comparison, the study is performed on an iso-area basis with a budget of 0.052mm^2 (as derived above). For the TSV-based benchmarks to make use of the full area allowance, scenarios of 1:1 replacement (TSVs to ICLs) in addition to $1:n^2$ replacement, where each ICL is replaced by an $n \times n$ array of through silicon vias with the same area footprint, are investigated. For even comparison, the aggregate data rate for the interface was considered such that, in the case of an $n \times n$ array, the data rate through each via, f_{tsv} , is equivalent to f/n^2 , reducing power consumption. Three different TSV styles are considered (32 μm TSVs [199] at 180 μm pitch, 15 μm TSVs [115] at 75 μm pitch, and 15 μm

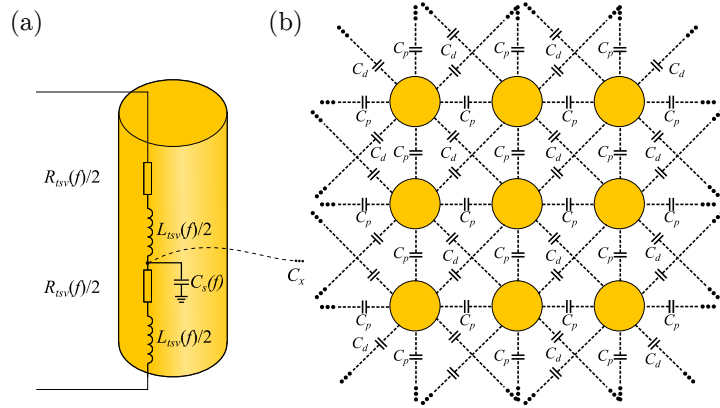


Figure A.2: (a) Electrical TSV model (where C_x denotes the coupling connections shown in (b)). (b) Parasitic coupling capacitances values in a square $n \times n$ TSV array.

Table A.5: TSV modelling parameters for each TSV array style. Extracted at 1GHz.

Style	Description	Parameters (at 1GHz)
[a]	32um TSVs [199] at 180um pitch	$R_{tsv} = 0.529 \Omega$, $L_{tsv} = 97.2 \text{ pH}$, $C_s = 76.8 \text{ fF}$, $C_{cx,y} = 19.8 \text{ fF}$, $C_{cdiag} = 17.5 \text{ fF}$
[b]	15um TSVs [115] at 75um pitch	$R_{tsv} = 0.952 \Omega$, $L_{tsv} = 1.01 \text{ pH}$, $C_s = 50.4 \text{ fF}$, $C_{cx,y} = 24.3 \text{ fF}$, $C_{cdiag} = 19.0 \text{ fF}$
[c]	15um TSVs [115] at 30um pitch	$R_{tsv} = 0.952 \Omega$, $L_{tsv} = 1.47 \text{ pH}$, $C_s = 51.8 \text{ fF}$, $C_{cx,y} = 51.8 \text{ fF}$, $C_{cdiag} = 42.6 \text{ fF}$

TSVs [115] at 30 μm pitch) in addition to three different TSV routing patterns (S-S-S, G-S-G, and isolated; outlined in Section A.2) resulting in a total of nine TSV-based designs.

TSV Modelling

In order to model TSV parasitics for subsequent analysis, the electrical model shown in Figure A.2 (a), proposed by Weerasekera *et. al.*, is used [200]. Here, it is assumed that each via exhibits resistance, $R_{tsv}(f)$, and inductance, $L_{tsv}(f)$, between its own terminals, in addition to self-capacitance, $C_s(f)$, and coupling capacitance between vias. In this chapter two coupling capacitances are considered as shown in Figure A.2 (b), namely *planar coupling* in the x or y direction, $C_p(f)$, and *diagonal coupling* $C_d(f)$. To determine representative values for each of these parameters, finite element modelling software (CST MW Studio) was used. Parameters were obtained for the various literature-based TSV sizes and pitches outlined above, which are detailed in Table A.5 in addition to the extracted RLC fitted values at a frequency of 1GHz. In this work, the driver inverters are appropriately scaled until reliable signalling is established; this ensures the minimum power consumption whilst achieving the maximum bandwidth¹.

¹This scaling is performed until the transistor size exceeds 100 \times minimum gate width. Nominal gate lengths are used for each experiment.

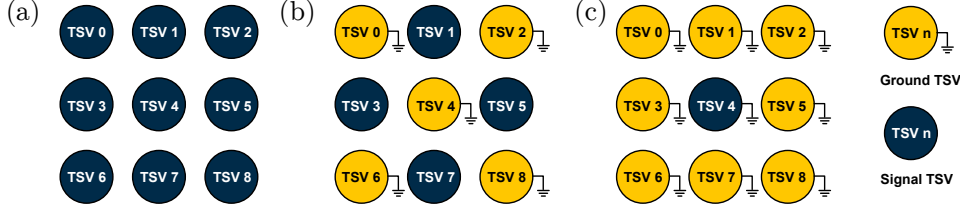


Figure A.3: (a) S-S-S, (b) G-S-G, and (c) Isolated TSV routing patterns.

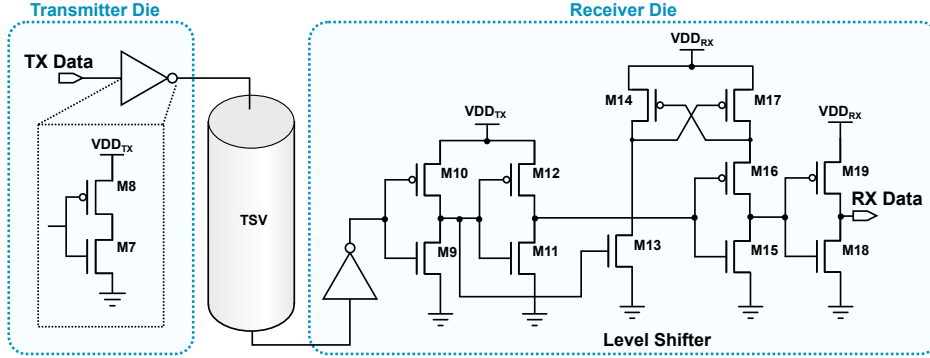


Figure A.4: TSV test configuration including drive inverters conjunction with a push-pull voltage level converter.

TSV Routing

Figure A.3 illustrates the three different TSV routing patterns which are considered in this appendix. In case (a) the full $n \times n$ TSV array is utilised for signal carrying TSVs and hence the signal is multiplexed n^2 ways such that each TSV is clocked at $1/n^2$ of the data frequency. Whilst this may seem like the most effective way to utilise the available area, and hence the most power efficient strategy, it is often advised that, to reduce coupling between neighbouring vias, especially at high frequencies, grounded vias should be inserted periodically within the array [201]. This leads to the ground-signal-ground (G-S-G) pattern shown in case (b). Finally, the case of a single signal carrying TSV (1:1 ICL to TSV replacement) is also explored, as shown in Figure A.3 (c).

Voltage Level Conversion

In the cases of heterogenous voltage integration ($2.5\text{V} \rightarrow 1.2\text{V}$ and $3.3\text{V} \rightarrow 1.2\text{V}$), additional level shifting circuitry is required when using TSVs. For experimental validation, level shifters are implemented in the 65nm (1.2V) tier, due to its favourable power efficiency and performance. To model this level conversion, the push-pull level conversion circuit presented in Figure A.4 is used.

This design is selected as it maintains very low quiescent power consumption by ensuring that there is never a constant path from either supply voltage to ground. For the implementation of this level shifter, each transistor (M0-M10) is sized with nominal length, 60nm, and width

of 200nm (in the 65nm technology) in order to manage uni-directional signal conversion for a range of VDD_{TX} values.

A.3 Comparison Results

Using the ICL and TSV interface models outlined in the previous sections, SPICE simulations were performed to evaluate the energy and maximum bandwidth of each approach for the three integration scenarios considered in this appendix (65nm \rightarrow 65nm, 180nm \rightarrow 65nm, and 0.35 μ m \rightarrow 65nm). The following sub-sections present the results of these experiments.

Bandwidth Comparison

The first two columns of Table A.6 documents the aggregate bandwidth and average transmission latency achieved by each approach. Results demonstrate that, as expected, the maximum ICL bandwidth in each case is much less than that achievable using TSVs. This is due the fact that the parasitic capacitance and resistance of the channel inductors limit the operating frequency of the link. When using TSVs, parasitics are naturally much lower and hence larger bandwidths can be achieved. When using multiple multiplexed TSVs, it can be observed that the achievable bandwidth-per-unit-area is up to $39\times$ higher than ICLs, simply due to their area efficiency. Despite this, it is important to note that, for typical low-cost IoT applications (which form the focus of this thesis), the bandwidths achievable through using ICLs are plenty sufficient [202].

Table A.6 also documents the latency of each approach. It can be observed that, in the cases of heterogeneous integration, the use of ICLs incurs approximately the same latency as the use of TSVs in conjunction with level conversion circuitry. In the case of homogeneous integration (where level conversion is not required), the latency of the wireless approach is slightly inferior due to the reduced complexity of the TSV-based designs. Whilst the latency of the ICL-based approach is higher, every case explored here results in sub-nanosecond delays which will be manifested as a single clock cycle delay when interfaced with digital logic (meaning that the latency of each approach in a practical system would likely be the same).

Energy Comparison

Finally, the energy-per-bit of each approach was also evaluated for a range of supported data-rates. The results of these simulations are presented in the final column of Table A.6. Of the TSV-based approaches, it can be observed that the larger, 32 μ m, TSVs perform better than the smaller 15 μ m TSVs in terms of energy consumption, likely due to the reduced parasitic capacitances between vias of this size. It can also be observed that the

¹Additional voltage level conversion circuitry (required for interfacing the two tiers) is incorporated in these simulation results.

Table A.6: Bandwidth, latency and energy per bit values for ICL and TSV based integration approaches.

	Integration Method	Max. Aggregate Data Rate	Average Latency	Energy Per Bit
65nm and 65nm	Inductive Coupling Link (ICL)	1.8 Gbps	0.68ns	0.86pJ/bit
	32um TSVs [199] at 180um pitch ¹			
	In Isolation	12.6 Gbps	0.06 ns	0.056pJ/bit
	G-S-G	16.0 Gbps	0.26ns	0.30pJ/bit
	S-S-S	28.0 Gbps	0.28ns	0.33pJ/bit
	15um TSVs [115] at 75um pitch ¹			
	In Isolation	4.2 Gbps	0.10ns	0.047pJ/bit
	G-S-G	15.9 Gbps	0.32ns	0.33pJ/bit
	S-S-S	32.1 Gbps	0.37ns	0.25pJ/bit
	15um TSVs [115] at 30um pitch ¹			
	In Isolation	4.2 Gbps	0.10ns	0.077pJ/bit
	G-S-G	34.2 Gbps	0.34ns	0.38pJ/bit
	S-S-S	62.2 Gbps	0.94ns	0.56pJ/bit
180nm and 65nm (2.5V \rightarrow 1.2V)	Inductive Coupling Link (ICL)	1.6 Gbps	0.72ns	0.85pJ/bit
	32um TSVs [199] at 180um pitch ¹			
	In Isolation	4.3 Gbps	0.31ns	1.03pJ/bit
	G-S-G	8.5 Gbps	0.61ns	0.94pJ/bit
	S-S-S	11.5 Gbps	0.62ns	1.05pJ/bit
	15um TSVs [115] at 75um pitch ¹			
	In Isolation	3.8 Gbps	0.46ns	0.99pJ/bit
	G-S-G	8.6 Gbps	0.76ns	0.90pJ/bit
	S-S-S	16.2 Gbps	0.78ns	1.12pJ/bit
	15um TSVs [115] at 30um pitch ¹			
	In Isolation	3.8 Gbps	0.50ns	1.29pJ/bit
	G-S-G	24.6 Gbps	0.89ns	1.25pJ/bit
	S-S-S	42.1 Gbps	0.94ns	1.40pJ/bit
0.35um and 65nm (3.3V \rightarrow 1.2V)	Inductive Coupling Link (ICL)	0.7 Gbps	0.65 ns	5.77pJ/bit
	32um TSVs [199] at 180um pitch ¹			
	In Isolation	4.1 Gbps	0.31 ns	11.2pJ/bit
	G-S-G	8.0 Gbps	0.89 ns	10.1pJ/bit
	S-S-S	11.5 Gbps	1.55 ns	1.05pJ/bit
	15umTSVs [115] at 75um pitch ¹			
	In Isolation	3.7 Gbps	0.48ns	12.0pJ/bit
	G-S-G	8.5 Gbps	0.89 ns	13.9pJ/bit
	S-S-S	16.6 Gbps	1.42 ns	21.6pJ/bit
	15um TSVs [115] at 30um pitch ¹			
	In Isolation	3.6 Gbps	0.50ns	26.1pJ/bit
	G-S-G	24.4 Gbps	1.77 ns	21.7pJ/bit
	S-S-S	41.8 Gbps	2.22 ns	24.5pJ/bit

use of G-S-G routing is broadly more energy efficient than the other routing approaches at the investigated operating frequencies (due to the trade-off that exists between the clock frequency of each via, and the amount of additional voltage conversion circuitry). Results demonstrate that, as expected, current state-of-the-art ICL designs exhibit higher energy consumption than TSVs for *homogeneous* 3D integration (by 70% on average).

However, interestingly, in each of the *heterogeneous* integration cases, the ICL-based design operates with the lowest energy-per-bit, outperforming the TSV-based approaches by an average of 36.7% in the case of 180 nm \rightarrow 65 nm integration and 67% in the case of 0.35 μ m \rightarrow 65 nm integration. This is due to the additional energy overhead introduced in voltage level conversion between the two dies when using TSVs.

A.4 Discussion

This appendix has compared the energy efficiency and bandwidth of through silicon vias and inductive coupling links in an iso-area study. Results demonstrate that for homogeneous 3D integration, the bandwidth achievable when using TSVs is much higher (up to $39\times$) for a fixed area constraint due to the ability to fit multiple TSVs in the same footprint as an inductive link. The use of TSVs was also more energy efficient, as hypothesised in the previous chapters, with the TSV-based approaches consuming, on average 70% less energy than ICLs across the explored cases.

However, when considering heterogeneous 3D integration scenarios (where a significant voltage difference exists between each tier), the ICL-based design demonstrated favourable energy efficiency when compared to TSVs. This is due to the wireless nature of ICLs, meaning that the power overheads of supplementary voltage level converters between dies can be avoided. This was demonstrated in simulation with high reliability and low latency (0.72ns) whilst reducing power consumption by 36.7% on average.

Appendix B

Dual-Dirac Model for Low-BER Jitter Extrapolation

This appendix describes the dual-Dirac model for jitter and Bit Error Rate (BER) extrapolation, used for generating the bathtub curves presented in Figure 3.14, Chapter 3 and Figure 6.10, Chapter 6.

As it would be computationally impossible to simulate millions of Monte-Carlo transmission cycles (which are required for plotting bathtub curves, such as those presented in Figures 3.14 and 6.10 of this thesis) using SPICE, the following sections outline the dual-Dirac fitting model; a widely accepted approach for estimating total jitter at a low bit error rates from a limited number of samples [203, 204]. The operation of this approach, in addition to a discussion of its applicability to ICL transceivers, is presented below in Section B.1, before the relevant mathematical models are presented in Section B.2.

B.1 Overview

In a coherent ICL transceiver, the correct reception of any given bit relies upon sampling the received voltage signal at the correct instant; sampling too early, or too late, will result in an incorrect decision, and therefore sensitivity to timing jitter is an important metric when comparing transceiver designs. The effects of timing jitter can be observed by simulating the link across many cycles whilst introducing random delay in the clock signal. Figure B.1 (a) presents a hypothetical illustration of an eye diagram where this process has been followed. Here, out of the N obtained samples, the very first rising edge (highlighted in blue) represents the earliest sample time that resulted in a successful transmission. Conversely, the final rising edge (highlighted in green) represents the latest samples time that resulted in a successful transmission. The dual-Dirac model suggests that the probability of a given sample time, x , resulting in successful transmission (and therefore appearing in the eye diagram) can be

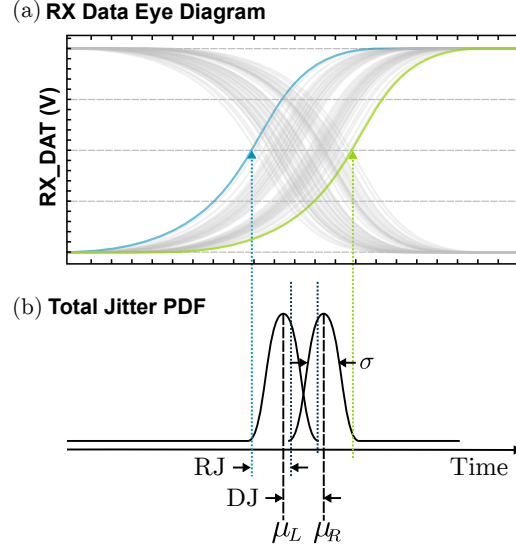


Figure B.1: (a) Illustration of a hypothetical RX data eye diagram observed when varying the delay of the ICL sample signal across several cycles. (b) The corresponding probability Density Function of the Total Jitter (TJ).

modelled by the convolution of two Gaussian distributions [204], as illustrated in Figure B.1 (b). Here the two Gaussian distributions correspond to the Random Jitter (RJ) in the received signal, and their separation corresponds to the Deterministic Jitter (DJ) [203, 204].

B.2 Mathematical Modelling

Using this model, the Probability Density Function (PDF) of the Total Jitter (TJ) is given by [203] :

$$PDF_{TJ} = PDF_{RJ} * PDF_{DJ} \quad (B.1)$$

As discussed above, RJ can be modelled by a Gaussian distribution, with mean μ and standard deviation σ . PDF_{RJ} is therefore given by:

$$PDF_{RJ} = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{x^2}{2\sigma^2} \right] \quad (B.2)$$

The Deterministic Jitter (DJ) can be modelled using the Dirac-delta function, $\delta(x - x_0)$, which takes the value of zero at every point, apart from when $x = x_0$, where it is infinite (*i.e.* a spike, centred at $x = x_0$) [203, 204]:

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases} \quad (B.3)$$

Applying these two models to Equation B.1, PDF_{TJ} can be expressed as:

$$PDF_{TJ} = [\delta(x - \mu_L) + \delta(x - \mu_R)] * \exp\left[-\frac{x^2}{2\sigma^2}\right] \quad (\text{B.4})$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \left[\exp\left(-\frac{(x - \mu_L)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x - \mu_R)^2}{2\sigma^2}\right) \right] \quad (\text{B.5})$$

where μ_L and μ_R represent the means of the left and right Gaussian distributions respectively.

The first step, therefore, when using the dual-Dirac method to perform BER extrapolation, is to plot a histogram of the RX clock jitter using the collected Monte-Carlo samples, and fit the expression in Equation B.5 to the results, in order to obtain values for parameters μ_L , μ_R and σ .

B.2.1 BER Extrapolation

This fitted model of the RX clock jitter can then be used to approximate the Bit Error Rate. The Bit Error Rate (BER) is defined as [204]:

$$BER = \lim_{N \rightarrow \infty} \frac{N_{err}(x)}{N} \quad (\text{B.6})$$

where x is the delay of the sample signal, and $N_{err}(x)$ is the number of errors that would be detected (at a sample delay of x) from a total of N transmitted bits. Assuming all bit errors arise from jitter, the BER can thereby be found by considering PDF_{TJ} . Since the BER depends on the probability of the sampling point x resulting in successful transmission, the BER for a given sample position, x , can be found by [205]:

$$BER(x) = \int_x^\infty PDF_{TJ}(x) dx \quad (\text{B.7})$$

Because fitted parameters describing $PDF_{TJ}(x)$ can be obtained from only a limited number of Monte-Carlo samples (using Equation B.5), this approach allows accurate extrapolation of BER for a full range of values of x , without the computational overhead of rigorous simulation.

Appendix C

Test Chip 1 Fabrication & Assembly Details

This appendix describes the physical design, manufacture and post-fabrication assembly processes for Test Chip 1 (discussed in Chapter 3 of this thesis), including die-level thinning, stacking and wire bonding.

C.1 Layout and Floorplan

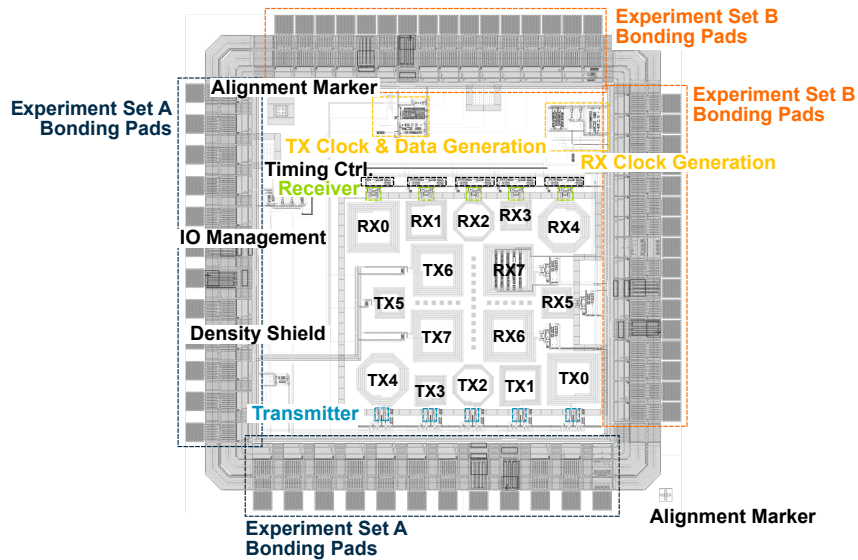


Figure C.1: Labelled floorplan of Test Chip 1 fabricated in 0.35 μm CMOS technology (discussed in Chapter 3 of this thesis).

Figure C.1 shows the floorplan of Test Chip 1 with the key components labelled. The design

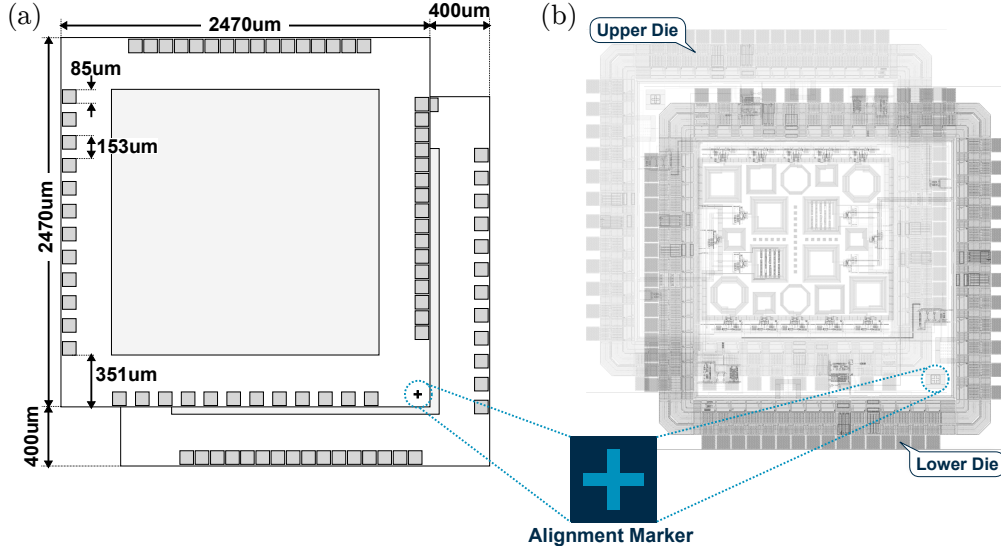


Figure C.2: Diagram showing the stacking and bonding arrangements for the first test-chip.

targets a two tier structure with one (lower) transmitter layer, and one (upper) receiver layer. In order to reduce the prototyping costs of the IC, both tiers were manufactured using the same mask-set, such that the 2-tier stacked IC contains two identical copies of the same design. To achieve this, the floorplan was designed so that the TX and RX channels (labelled on Figure C.1) align when rotated through 180 degrees, and stacked with a 400 μm offset¹ to allow wire-bond access to both dies for test and measurement purposes.

As shown on Figure C.1, the chip contains two sets of experiments: Experiment set (A) that includes TX/RX channels 0, 1, 2, 3 and 4 (which can be accessed from the bond-pads highlighted in blue), and Experiment set (B) that includes TX/RX channels 5, 6 and 7 (which can be accessed from the bond-pads highlighted in orange). The test-chip also features a clock generator block for both the TX and RX circuits (a programmable ring-oscillator), in addition to an automated test pattern generator (based upon on Linear Feedback Shift Register (LFSR)). As shown on the figure, the design contains several links, each using different inductor geometries. The transmitters in links 0-4 contain the spike-latency control circuits discussed in Chapter 3 in addition to the tuneable TX current driver. The transmitters in links 5-7 use the basic H-Bridge NRZ driver discussed in Chapter 2.

The overall area of the test chip is 2.47mm \times 2.47mm with a core area (excluding the IO ring) of 2.2mm \times 2.2mm. Several alignment markers are also included for the stacking stage (discussed below in Section C.2); most notably the ‘plus’ marker in the bottom right-hand corner, and its inverse (in the top right-hand corner of the core area).

¹An illustration of this is provided in Figure C.2 (b).

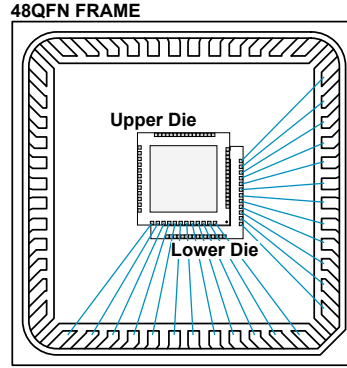


Figure C.3: Bonding diagram showing the packaging arrangements for accessing experiment Set A. Dies were packaged in a 48QFN package as shown.

C.2 Thinning and Stacking

Once fabricated, dies were thinned by an external contractor. Thinning was performed on individual dies by attaching them to a blank sacrificial substrate using a UV curable adhesive tape. Once attached to the carrier wafer, back-grinding was performed using a Disco DAG-810 automatic surface grinder, to a final thickness of 100 μm . After this had been performed, the dies were detached from the sacrificial wafer (using UV light), and then stacked in the arrangement shown in Figure C.2 (a). Here, as discussed above, the upper die is rotated through 180 degrees and stacked with a 400 μm offset when compared with the lower die. This ensures that the TX inductors align with their RX counterparts and that the plus shaped markers also align, as highlighted.

C.3 Bonding and Packaging

The stacked dies were packaged in Quad Flat No-leads (QFN) packages. As it was not possible to bond to both the lower and upper dies of the stack from the same edge in the same chip, two different bonding patterns (for the two different experiment sets (A) and (B)) were used. For experiment set (A), the bottom edge of the *upper* die, and the right edge of the *lower* die were bonded (including both of the IO pad sets highlighted in blue on Figure C.1, when accounting for the 180 degree rotation). This bonding arrangement is shown in Figure C.3. Conversely, for experiment set (B), the bottom edge of the *lower* die and the right edge of the *upper* die were bonded (including both of the IO pad sets highlighted in orange on Figure C.1, when accounting for the 180 degree rotation).

The bonding arrangement for experiment set (A)² uses 12 pads per edge, and hence was packaged in 48QFN packages as shown in Figure C.3. The bonding arrangement for experiment set (B) uses 16 pads per edge and hence would require a 64QFN package.

²Only experiments from bonding set (A), which includes the tuneable pulse driver and spike latency encoding transceiver were explored in this thesis.

Appendix D

Test Chip 2 Fabrication & Assembly Details

This appendix describes the physical design, manufacture and post-fabrication assembly processes for Test Chip 2 (discussed in this thesis in Chapters 3 and 5), including die-level thinning, stacking and wire bonding.

D.1 Layout and Floorplan

Figure D.1 shows a labelled floorplan of Test Chip 2 with the key components highlighted, including: the four CoDAPT links (0-3¹), the WiSync clock link, an Arm Cortex M0 MCU, the AHB-Lite bus + WBI, SRAM blocks, and the clock generator block (in this case, a programmable RO).

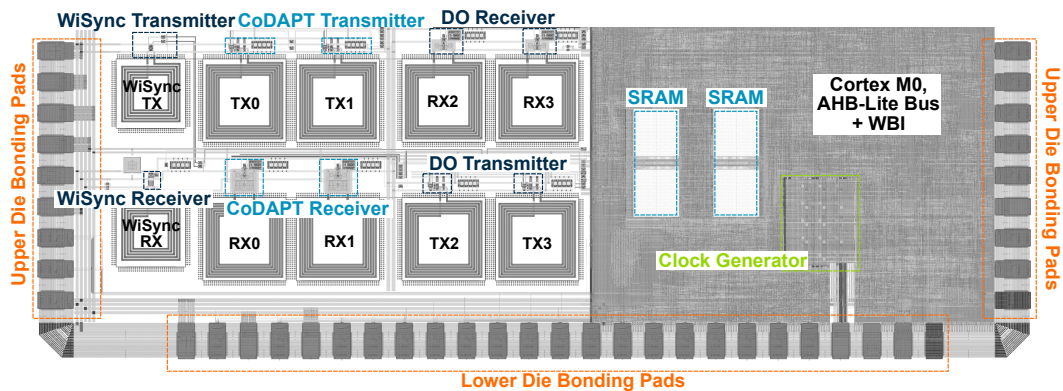


Figure D.1: Labelled floorplan of Test Chip 2, fabricated in 65nm CMOS technology (discussed in Chapters 5 and 6 of this thesis).

¹As discussed in Chapter 6 only the uplinks (0-1) include power rectification circuits.

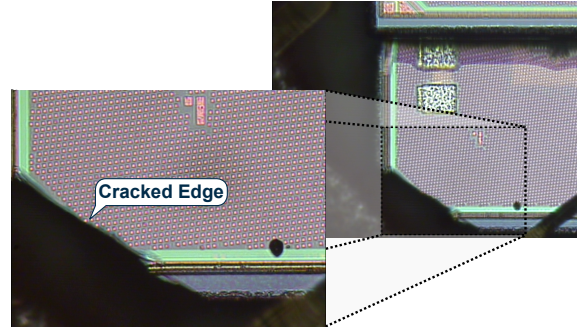


Figure D.2: Microscope view showing cracked corner when performing 70 μm individual die thinning.

In the same way as Test Chip 1 (presented in in Appendix C), Test Chip 2 (TC2) targets a two layer architecture, where the upper and lower dies are both fabricated using the same mask-set (and hence are identical copies of the same design). To achieve this in TC2, the floorplan was designed such that the RX and TX channels would align when stacked with a 400 μm lateral offset in the y direction (without need for rotation as in Appendix C). The overall area of the 65nm CMOS design is $1\text{mm} \times 3\text{mm}$, with a core area (excluding IO pads) of $0.864\text{mm} \times 2.66\text{mm}$. The design was included as part of a larger ($3\text{mm} \times 4\text{mm}$) shared shuttle, and hence has no bond-pads on the top edge in order to abut with other designs (as illustrated by Figure D.3).

D.2 Thinning and Stacking

Once manufactured, thinning was performed on each of the individual $3\text{mm} \times 4\text{mm}$ dies by an external contractor. The same die-level thinning approach outlined in Section C.2 of Appendix C was used, where dies are attached to a sacrificial wafer using UV cured tape, ground using a Disco DAG-810 machine, and then detached from the sacrificial carrier wafer using UV light. Following the success of the 100 μm back-grinding in the previous test-chip (with a 100% yield), the level of thinning used for Test Chip 2 was increased by 30 μm to a final thickness of 70 μm . Whilst this 70 μm thinning was performed successfully (with a *functional* yield of 100%), it was clear that 70 μm is approaching the physical limits of what can be achieved, as some dies were chipped during the process; an example of this can be observed in the microscope image in Figure D.2, where the corner of the lower die has cracked.

Once thinned, the dies were stacked in the arrangement shown in Figure D.3 (a). Stacking was performed using a die bonder, and QMI 538NB-1A1.5 non-conductive epoxy adhesive [174] which was thermally cured at 175 degrees centigrade. As outlined above, dies were stacked with a 400 μm offset in the y direction such that the TX and RX channel inductors align (this effect can be observed on Figure D.3 (b)). The stacking alignment of each

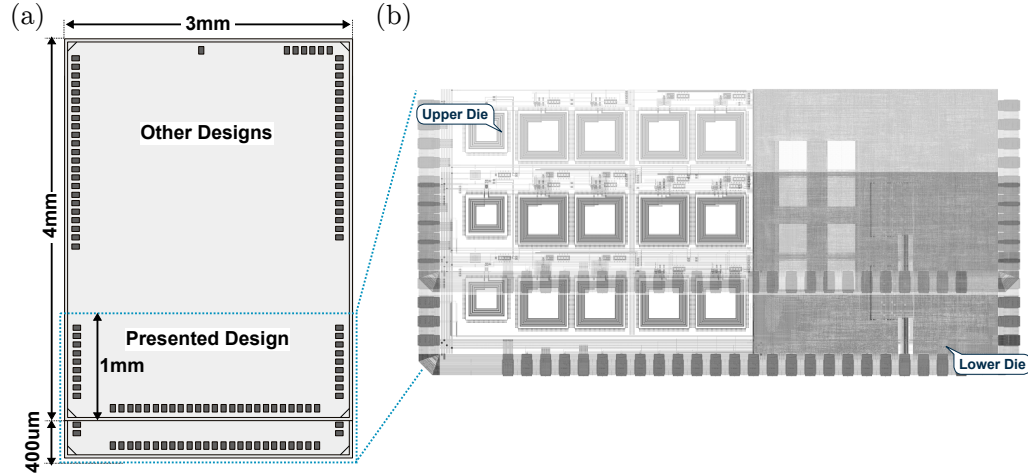


Figure D.3: Diagram showing the stacking and bonding arrangements for Test Chip 2.

individual chip was also verified through a microscope using the reference features of the chip.

D.3 Bonding and Packaging

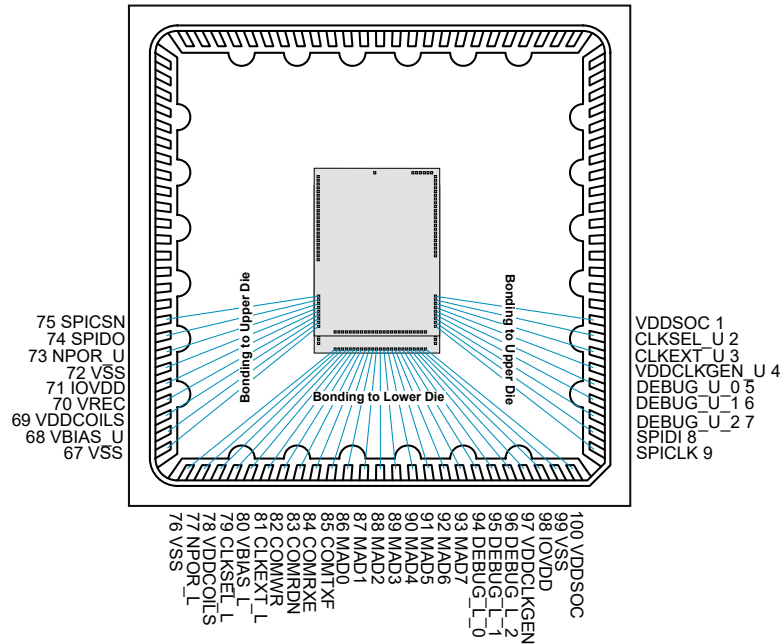


Figure D.4: Bonding diagram for Test Chip 2 using a 100QFP package.

Finally, the stacked dies were packaged in wire-bonded QFPs, according to the diagram shown in Figure D.4. Again, to avoid bond-wire shorts that could occur when attaching to the same edge of upper *and* lower dies, the floorplan was designed such that the upper die's IO pads were placed on the sides of the chip, and lower die's IO pads were placed on the

bottom edge of the chip. This meant that, once stacked, the bottom edge of the QFP frame could be bonded to the *lower* die, whilst the left and right edges of the QFP frame could be bonded to the *upper* die, as shown in Figure [D.4](#).

References

- [1] J. N. Burghartz, "Thin chips on the ITRS roadmap," in *Ultra-thin Chip Technology and Applications*. Springer, 2011, pp. 13–18.
- [2] A. Todri-Sanial and C. Tan, *Physical Design for 3D Integrated Circuits*. CRC Press, 2016.
- [3] R. Wang *et al.*, "The characterization of TSV Cu protrusion under thermal cycling," in *Int. Conf. on Electronic Packaging Technology (ICEPT)*, Aug 2015, pp. 888–890.
- [4] M. B. Kleiner *et al.*, "Performance improvement of the memory hierarchy of RISC-systems by application of 3-D technology," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 19, no. 4, pp. 709–18, Nov 1996.
- [5] M. Koyanagi, "Heterogeneous 3D integration for Internet of Things," in *IEEE Int. Conf. on Solid-State and Integrated Circuit Technology (ICSICT)*, 2014.
- [6] U. Kang *et al.*, "8 Gb 3-D DDR3 DRAM using through-silicon-via technology," *IEEE Journ. of Solid-State Circuits (JSSC)*, vol. 45, no. 1, pp. 111–119, Jan 2010.
- [7] D. H. Woo *et al.*, "An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth," in *Int. Symp. on High-Performance Computer Architecture (HPCA)*, Jan 2010, pp. 1–12.
- [8] J. S. Kim *et al.*, "A 1.2 v 12.8 GB/s 2 Gb mobile wide-I/O DRAM with 4x 128 I/Os using TSV based stacking," *IEEE Journ. of Solid-State Circuits JSSC*, vol. 47, no. 1, pp. 107–16, Jan 2012.
- [9] J. Elliott and E. Jung, "Ushering in the 3D memory era with V-NAND," in *Proc. Flash Memory Summit*, 2013.
- [10] B. Shen and W. Wu, "3D-IC system design impact, challenge and solutions," in *Proc. of the Int. Symp. on Physical Design*, ser. ISPD '14, 2014, pp. 63–64.
- [11] A. W. Topol *et al.*, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 491–506, July 2006.
- [12] L. A. Hornak and S. K. Tewksbury, "On the feasibility of through-wafer optical

- interconnects for hybrid wafer-scale-integrated architectures,” *IEEE Trans. on Electron Devices*, vol. 34, no. 7, pp. 1557–1563, Jul 1987.
- [13] M. J. Little *et al.*, “The 3-D computer,” *Journ. of VLSI Signal Processing*, vol. 2, no. 2, pp. 79–87, Oct 1990.
- [14] G. Katti *et al.*, “Electrical modeling and characterization of through silicon via for three-dimensional ICs,” *IEEE Trans. on Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan 2010.
- [15] G. Parès *et al.*, “Effects of stress in polysilicon via - first TSV technology,” in *12th Electronics Packaging Technology Conference*, 2010, pp. 333–337.
- [16] Y. Li *et al.*, “Impact of backside process on high aspect ratio via-middle Cu through silicon via reliability,” in *Int. Conf. on Electronics Packaging (ICEP)*, 2017, pp. 508–512.
- [17] X. Jing *et al.*, “Via last tsv process for wafer level packaging,” in *2016 17th Int. Conference on Electronic Packaging Technology (ICEPT)*, 2016, pp. 1216–1218.
- [18] P. S. Andry *et al.*, “Fabrication and characterization of robust through-silicon vias for silicon-carrier applications,” *IBM Journ. of Research and Development*, vol. 52, no. 6, pp. 571–581, Nov 2008.
- [19] W. R. Davis *et al.*, “Demystifying 3D ICs: the pros and cons of going vertical,” *IEEE Design Test of Computers*, vol. 22, no. 6, pp. 498–510, Nov 2005.
- [20] R. Radojicic, *More-than-Moore 2.5 D and 3D SiP Integration*. Springer, 2017.
- [21] A.-C. Hsieh and T. Hwang, “TSV redundancy: architecture and design issues in 3-D IC,” *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 4, pp. 711–722, 2012.
- [22] K. Tu, “Reliability challenges in 3D IC packaging technology,” *Microelectronics Reliability*, vol. 51, no. 3, pp. 517 – 523, 2011.
- [23] R. Johnson and Y.-L. Shen, “Analysis of TSV/micro-bump deformation due to chip misalignment and thermal processing in 3D IC packages,” vol. 9, 11 2012.
- [24] D. Ditzel, T. Kuroda, and S. Lee, “Low-cost 3D chip stacking with ThruChip wireless connections,” in *Hot Chips-A Symp. on High Performance Chips*, 2014.
- [25] G. Chen *et al.*, “A cubic-millimeter energy-autonomous wireless intraocular pressure monitor,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2011, pp. 310–312.
- [26] S. Jeong *et al.*, “A fully-integrated 71nw CMOS temperature sensor for low power

- wireless sensor nodes,” *IEEE Journ. of Solid-State Circuits (JSSC)*, vol. 49, no. 8, pp. 1682–1693, 2014.
- [27] G. Chen *et al.*, “Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2010, pp. 288–9.
- [28] Y. Lee *et al.*, “A modular 1 mm⁴ die-stacked sensing platform with low power I²C inter-die communication and multi-modal energy harvesting,” *IEEE Journ. of Solid-State Circuits*, vol. 48, no. 1, pp. 229–243, Jan 2013.
- [29] M. Fojtik *et al.*, “A millimeter-scale energy-autonomous sensor system with stacked battery and solar cells,” *IEEE Journ. of Solid-State Circuits*, vol. 48, no. 3, pp. 801–813, March 2013.
- [30] H. Nakanishi *et al.*, “Development of high density memory IC package by stacking IC chips,” in *Proc. Electronic Components and Technology Conf.*, May 1995, pp. 634–640.
- [31] H. J. Timme, K. Pressel, G. Beer, and R. Bergmann, “Interconnect technologies for system-in-package integration,” in *IEEE Electronics Packaging Technology Conf. (EPTC)*, Dec 2013, pp. 641–646.
- [32] K. Chang *et al.*, “Match-making for monolithic 3D IC: Finding the right technology node,” in *IEEE Design Automation Conf. (DAC)*, June 2016, pp. 1–6.
- [33] S. Wong *et al.*, “Monolithic 3D integrated circuits,” in *Int. Symp. on VLSI Technology, Systems and Applications (VLSI-TSA)*, Apr 2007, pp. 1–4.
- [34] S. Datta *et al.*, “Back-end-of-line compatible transistors for monolithic 3-d integration,” *IEEE Micro*, vol. 39, no. 6, pp. 8–15, Nov 2019.
- [35] M. M. Shulaker *et al.*, “Monolithic 3d integration: A path from concept to reality,” in *Design, Automation Test in Europe Conf. Exhibition (DATE)*, March 2015, pp. 1197–1202.
- [36] C. Fenouillet-Beranger *et al.*, “A review of the full 500 °C low temperature technological modules development for high performance and reliable 3d sequential integration,” in *Electron Devices Technology and Manufacturing Conf. (EDTM)*, 2019, pp. 249–51.
- [37] C. Fenouillet-Beranger *et al.*, “First demonstration of low temperature ($\leq 500^\circ\text{C}$) cmos devices featuring functional RO and SRAM bitcells toward 3D VLSI integration,” in *Int. Symp. on VLSI Circuits*, Jun 2020, pp. 1–2.
- [38] L. Brunet *et al.*, “Breakthroughs in 3D sequential technology,” in *IEEE Int. Electron Devices Meeting (IEDM)*, 2018, pp. 721–24.

- [39] S. Thuries *et al.*, “M3D-ADTCO: Monolithic 3D architecture, design and technology co-optimization for high energy efficient 3D IC,” in *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2020, pp. 1740–45.
- [40] S. W. Han, “Wireless interconnect using inductive coupling in 3D-ICs,” Ph.D. dissertation, The University of Michigan, 2012.
- [41] B. Berkowitz, *Basic microwaves*. Hayden Book Co., 1966.
- [42] *Inductive coupling thruchip interface for 3D integration*. CRC Press, Jan 2014.
- [43] N. Miura *et al.*, “A 1 Tb/s 3 W inductive-coupling transceiver for 3D-stacked inter-chip clock and data link,” *IEEE Journ. of Solid-State Circuits*, vol. 42, no. 1, pp. 111–122, Jan 2007.
- [44] P. Weidelt, “Electromagnetic induction in three-dimensional structures,” *J. Geophys*, vol. 41, no. 85, p. 109, 1975.
- [45] X. Sun *et al.*, “Inductive links for 3D stacked chip-to-chip communication,” in *IEEE Electronic Components and Tech. Conf. (ECTC)*, 2019.
- [46] I. A. Papistas and V. F. Pavlidis, “Contactless heterogeneous 3-D ICs for smart sensing systems,” *Integration*, vol. 62, pp. 329 – 340, 2018.
- [47] J. Zhao, Q. Zou, and Y. Xie, “Overview of 3-D architecture design opportunities and techniques,” *IEEE Design Test*, vol. 34, no. 4, pp. 60–68, Aug 2017.
- [48] H. Jayakumar *et al.*, “Powering the internet of things,” in *IEEE/ACM Int. Symp. on Low Power Electronics and Design (ISLPED)*, Aug 2014, pp. 375–380.
- [49] N. Miura *et al.*, “A 195Gb/s 1.2W 3D-stacked inductive inter-chip wireless superconnect with transmit power control scheme,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2005, pp. 264–597 Vol. 1.
- [50] N. Miura *et al.*, “An 11Gb/s inductive-coupling link with burst transmission,” in *IEEE Int. Solid-State Circuits Conf.*, 2008.
- [51] N. Miura *et al.*, “Analysis and design of inductive coupling and transceiver circuit for inductive inter-chip wireless superconnect,” *IEEE Journ. of Solid-State Circuits*, vol. 40, no. 4, pp. 829–837, April 2005.
- [52] L. C. Hsu *et al.*, “Analytical thruchip inductive coupling channel design optimization,” in *Asia and South Pacific Design Automation Conf. (ASP-DAC)*, 2016, pp. 731–736.
- [53] J. Park *et al.*, “Efficient calculation of inductive and capacitive coupling due to electrostatic discharge (ESD) Using PEEC Method,” *IEEE Trans. on Electromagnetic Compatibility*, vol. 57, no. 4, pp. 743–753, Aug 2015.

- [54] N. Miura *et al.*, "A 0.14pJ/b inductive-coupling inter-chip data transceiver with digitally-controlled precise pulse shaping," in *IEEE Int. Solid-State Circuits Conf.*, Feb 2007, pp. 358–608.
- [55] Y. Take, N. Miura, and T. Kuroda, "A 30 Gb/s/Link 2.2 Tb/s/mm² inductively-coupled injection-locking CDR for high-speed DRAM interface," *IEEE Journ. of Solid-State Circuits*, vol. 46, no. 11, pp. 2552–2559, Nov 2011.
- [56] L. Zhang, T. Li, B. Wang, and X. Zou, "A 50% power reduction in inductive-coupling transceiver for 3D-stacked inter-chip data link," in *IEEE Int. Nanoelectronics Conf. (INEC)*, May 2016, pp. 1–3.
- [57] Y. Shimazaki, N. Miura, and T. Kuroda, "A 5.184Gbps/ch through-chip interface and automated place-and-route design methodology for 3-D integration of 45nm CMOS processors," in *IEEE COOL Chips XV*, April 2012, pp. 1–3.
- [58] Y. Take, J. Kadomoto, and T. Kuroda, "3D integration using inductive coupling and coupled resonator (invited)," in *IEEE Int. Symp. on Radio-Frequency Integration Technology (RFIT)*, Aug 2015, pp. 46–48.
- [59] Y. Yuan *et al.*, "Chip-to-chip power delivery by inductive coupling with ripple canceling scheme," *Japanese Journ. of Applied Physics*, vol. 47, no. 4S, p. 2797, 2008.
- [60] A. Radecki *et al.*, "Simultaneous 6-gb/s data and 10-mw power transmission using nested clover coils for noncontact memory card," *IEEE Journ. of Solid-State Circuits*, vol. 47, no. 10, pp. 2484–2495, Oct 2012.
- [61] Y. Yuxiang *et al.*, "Digital rosetta stone: A sealed permanent memory with inductive-coupling power and data link," in *Symp. on VLSI Circuits*, June 2009, pp. 26–27.
- [62] I. A. Papistas and V. F. Pavlidis, "Bandwidth-to-area comparison of through silicon vias and inductive links for 3-D ICs," in *European Conf. on Circuit Theory and Design (ECCTD)*, Aug 2015, pp. 1–4.
- [63] Y. Take, N. Miura, H. Ishikuro, and T. Kuroda, "3d clock distribution using vertically/horizontally-coupled resonators," in *IEEE Int. Solid-State Circuits Conf.*, Feb 2013, pp. 258–259.
- [64] Y. Take and T. Kuroda, "Relay transmission thru chip interface with low-skew 3d clock distribution network," *IEICE Transactions on Electronics*, vol. E98C, no. 4, pp. 322–332, 4 2015.
- [65] M. Saito, N. Miura, and T. Kuroda, "A 2Gb/s 1.8pJ/b/chip inductive-coupling through-chip bus for 128-die NAND-flash memory stacking," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2010, pp. 440–441.

- [66] K. Ueyoshi *et al.*, "Quest: A 7.49TOPS multi-purpose log-quantized DNN inference engine stacked on 96MB 3D SRAM using inductive-coupling technology in 40nm CMOS," in *IEEE Int. Solid - State Circuits Conf. (ISSCC)*, 2018.
- [67] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the internet of things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–16, 2016.
- [68] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things Journ.*, vol. 3, no. 6, pp. 854–64, 2016.
- [69] Transforma Insights, May 2020. [Online]. Available: <https://transformainsights.com/news/iot-market-24-billion-usd15-trillion-revenue-2030>
- [70] T. Wu, J. Redouté, and M. R. Yuce, "A wireless implantable sensor design with subcutaneous energy harvesting for long-term IoT healthcare applications," *IEEE Access*, vol. 6, pp. 35 801–8, 2018.
- [71] I. Ud Din *et al.*, "The internet of things: A review of enabled technologies and future challenges," *IEEE Access*, vol. 7, pp. 7606–40, 2019.
- [72] P. Han *et al.*, "A 920-MHz dual-mode receiver with energy harvesting for UHF RFID tag and IoT," *Electronics*, vol. 9, p. 1042, Jun 2020.
- [73] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.
- [74] S. Paul *et al.*, "A sub-cm³ energy-harvesting stacked wireless sensor node featuring a near-threshold voltage IA-32 microcontroller in 14-nm tri-gate CMOS for always-on always-sensing applications," *IEEE Journ. of Solid-State Circuits*, vol. 52, no. 4, pp. 961–71, 2017.
- [75] X. Wu *et al.*, "A 0.04mm³ 16nW wireless and batteryless sensor system with integrated Cortex-M0+ processor and optical communication for cellular temperature measurement," in *IEEE Symp. on VLSI Circuits*, 2018, pp. 191–2.
- [76] W. Xu *et al.*, in *Int. Conf. on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*.
- [77] A. Raj and D. Steingart, "Review—power sources for the internet of things," *Journal of The Electrochemical Society*, vol. 165, pp. 3130–136, 01 2018.
- [78] B. A. Warneke and K. S. J. Pister, "An ultra-low energy microcontroller for smart dust wireless sensor networks," in *IEE Int. Solid-State Circuits Conf.*, vol. 1, 2004, pp. 316–17.
- [79] Y. Pu *et al.*, "A 9mm² ultra-low-power highly integrated 28nm CMOS SoC for internet of things," vol. 53, no. 3, pp. 936–48, 2018.

- [80] A. Pullini *et al.*, “Mr.wolf: An energy-precision scalable parallel ultra low power SoC for IoT edge processing,” *IEEE Journ. of Solid-State Circuits*, vol. 54, no. 7, pp. 1970–81, 2019.
- [81] E. Flamand *et al.*, “GAP-8: a RISC-V SoC for AI at the edge of the IoT,” in *IEEE Int. Conf. on Application-specific Systems, Architectures and Processors (ASAP)*, 2018, pp. 1–4.
- [82] I. A. Papistas and V. F. Pavlidis, “Contactless inter-tier communication for heterogeneous 3-D ICs,” in *Proc. of the IEEE Int. Symp. on Circuits and Systems*, May 2017, pp. 2585–88.
- [83] B. J. Fletcher, S. Das, and T. Mak, “A spike-latency transceiver with tunable pulse control for low-energy wireless 3D integration,” *IEEE Journ. on Solid State Systems (JSSC)*, vol. 55, no. 9, pp. 2414–28, 2020.
- [84] B. J. Fletcher, T. Mak, and S. Das, “A 3D-stacked Cortex-M0 SoC with 20.3gbps/mm² 7.1mw/mm² simultaneous wireless inter-tier data and power transferr,” in *IEEE Symp. on VLSI Circuits*, Jun 2020, pp. 1–2.
- [85] B. J. Fletcher, T. Mak, and S. Das, “A 10.8pJ/bit pulse-position inductive transceiver for low-energy wireless 3D integration,” in *IEEE European Solid State Circuits Conf. (ESSCIRC)*, 2019, pp. 121–124.
- [86] B. J. Fletcher, S. Das, and T. Mak, “A low-energy inductive transceiver using spike-latency encoding for wireless 3D integration,” in *IEEE Int. Symp. on Low Power Elec. and Design (ISLPED)*, 2019.
- [87] B. J. Fletcher, S. Das, and T. Mak, “Design and optimization of inductive-coupling links for 3-D-ICs,” *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 3, pp. 711–723, 2019.
- [88] B. J. Fletcher, S. Das, and T. Mak, “CoDAPT: A concurrent data and power transceiver for fully wireless 3D-ICs,” in *Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2019, pp. 1343–1348.
- [89] B. J. Fletcher, S. Das, and T. Mak, “Cost-effective 3D integration using inductive coupling links: Can we make stacking silicon as easy as stacking Lego?” in *Arm Research Summit 2018 (19/09/18)*, September 2018.
- [90] B. J. Fletcher, S. Das, and T. Mak, “A high-speed design methodology for inductive coupling links in 3D-ICs,” in *Design, Automation Test in Europe Conf. Exhibition (DATE)*, March 2018, pp. 497–502.
- [91] B. J. Fletcher, S. Das, C. S. Poon, and T. Mak, “Low-power 3D integration using

- inductive coupling links for neurotechnology applications,” in *Design, Automation Test in Europe Conf. Exhibition (DATE)*, March 2018, pp. 1211–1216.
- [92] D. Balsamo, B. J. Fletcher, A. S. Weddell, G. Karatzias, B. M. Al-Hashimi, and G. V. Merrett, “Momentum: Power-neutral performance scaling with intrinsic mppt for energy harvesting computing systems,” *ACM Trans. Embed. Comput. Syst.*, vol. 17, no. 6, pp. 1–25, Jan. 2019.
- [93] D. Balsamo, B. J. Fletcher, and G. Merrett, “Power-neutral performance scaling for self-powered multicore computing systems,” in *Adaptive Many-Core Architectures and Systems Workshop (15/06/18)*, June 2018.
- [94] Q. Ding, B. J. Fletcher, and T. Mak, “Globally wireless locally wired (GloWiLoW): A clock distribution network for many-core systems,” in *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.
- [95] B. J. Fletcher, D. Balsamo, and G. V. Merrett, “Power neutral performance scaling for energy harvesting mp-socs,” in *Design, Automation Test in Europe Conf. Exhibition (DATE)*, 2017, March 2017, pp. 1516–1521.
- [96] B. J. Fletcher, J. Myers, S. Das, and T. Mak, “A pseudo system-on-chip architecture incorporating wirelessly connected bus slaves,” patentus 16/685 090, Nov., 2019.
- [97] S. Gamage, B. J. Fletcher, and S. Das, “Adaptive coding for wireless communication,” patentus 16/656 937, Oct., 2019.
- [98] I. A. Papistas, V. F. Pavlidis, and D. Velenis, “Fabrication cost analysis for contactless 3-D ICs,” *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 66, no. 5, pp. 758–762, 2019.
- [99] A. Fazzi *et al.*, “3-D capacitive interconnections for wafer-level and die-level assembly,” *IEEE Journ. of Solid-State Circuits*, vol. 42, no. 10, pp. 2270–2282, Oct 2007.
- [100] R. J. Drost, R. D. Hopkins, and I. E. Sutherland, “Proximity communication,” in *Proc. of the IEEE Custom Integrated Circuits Conf.*, Sept 2003, pp. 469–472.
- [101] M. T. L. Aung *et al.*, “A 3-Gb/s/ch simultaneous bidirectional capacitive coupling transceiver for 3DICs,” *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 61, no. 9, pp. 706–710, Sept 2014.
- [102] E. Culurciello and A. G. Andreou, “Capacitive inter-chip data and power transfer for 3-D VLSI,” *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 53, no. 12, pp. 1348–1352, Dec 2006.
- [103] K. Kanda *et al.*, “1.27Gb/s/pin 3mW/pin wireless superconnect (WSC) interface scheme,” in *Int. Solid-State Circuits Conf. ISSCC.*, Feb 2003, pp. 186–487 vol.1.

- [104] K. P. P. Pillai, "Fringing field of finite parallel-plate capacitors," *Proceedings of the Institution of Electrical Engineers*, vol. 117, no. 6, pp. 1201–1204, 1970.
- [105] R. J. Drost, R. D. Hopkins, R. Ho, and I. E. Sutherland, "Proximity communication," *IEEE Journ. of Solid-State Circuits*, vol. 39, no. 9, pp. 1529–1535, Sept 2004.
- [106] Lei Luo *et al.*, "3Gb/s AC-coupled chip-to-chip communication using a low-swing pulse receiver," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2005, pp. 522–614 Vol. 1.
- [107] D. Saltzman and T. Knight, "Capacitive coupling solves the known good die problem," in *Proceedings of IEEE Multi-Chip Module Conf. (MCMC)*, March 1994, pp. 95–100.
- [108] R. Canegallo *et al.*, "3D contactless communication for IC design," in *IEEE Int. Conf. on Integrated Circuit Design and Technology and Tutorial*, June 2008, pp. 241–244.
- [109] M. T. L. Aung *et al.*, "Design review on capacitive coupling interconnect for 3D IC," in *IEEE Int. Conf. on Electron Devices and Solid-State Circuits (EDSSC)*, June 2015, pp. 245–248.
- [110] L. Bogaerts *et al.*, "Process related challenges for 3D face to face stacking test vehicles using a 40/50um pitch CuSn microbump configuration," in *IEEE Electronics Packaging Technology Conf. (EPTC)*, Dec 2012, pp. 278–282.
- [111] C. Lee *et al.*, "Transceiver with inductive coupling for wireless chip-to-chip communication using a 50-nm digital CMOS process," *Microelectronics Journ.*, vol. 44, no. 9, pp. 852–859, 2013.
- [112] N. Miura *et al.*, "A high-speed inductive-coupling link with burst transmission," *IEEE Journ. of Solid-State Circuits*, vol. 44, no. 3, pp. 947–955, March 2009.
- [113] J. Xu, J. Wilson, S. Mick, L. Luo, and P. Franzon, "2.8 Gb/s inductively coupled interconnect for 3D ICs," in *Digest of Technical Papers. 2005 Symposium on VLSI Circuits, 2005.*, June 2005, pp. 352–355.
- [114] K. Niitsu *et al.*, "An inductive-coupling link for 3D integration of a 90nm cmos processor and a 65nm CMOS SRAM," in *IEEE Int. Solid-State Circuits Conf.*, Feb 2009, pp. 480–481, 481a.
- [115] A. Yu *et al.*, "Three dimensional interconnects with high aspect ratio TSVs and fine pitch solder microbumps," in *Electronic Components and Technology Conf.*, May 2009, pp. 350–354.
- [116] Kiichi Niitsu *et al.*, "A 65 fJ/b inductive-coupling inter-chip transceiver using charge recycling technique for power-aware 3D system integration," in *IEEE Asian Solid-State Circ. Conf.*, 2008.

- [117] N. Miura *et al.*, "A 0.55V 10 fJ/bit inductive-coupling data link and 0.7V 135 fJ/Cycle clock link with dual-coil transmission scheme," *IEEE Journ. of Solid-State Circ.*, vol. 46(4), pp. 965–973, 2011.
- [118] L. Zhang *et al.*, "A single phase modulation for pulse-based inductive-coupling connection in 3d stacked chip," *IEICE Electronics Express*, 10 2017.
- [119] T. Kuroda, "Wireless proximity communications for 3D system integration," in *IEEE Int. Workshop on Radio-Frequency Integration Technology*, Dec 2007, pp. 21–25.
- [120] S. Kawai, H. Ishikuro, and T. Kuroda, "A 2.5gb/s/ch 4pam inductive-coupling transceiver for non-contact memory card," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb 2010, pp. 264–265.
- [121] I. A. Papistas, "Design methodologies for heterogeneous 3-D integrated systems," Ph.D. dissertation, The University of Manchester, 2018.
- [122] B. Nikolic *et al.*, "Improved sense-amplifier-based flip-flop: design and measurements," *IEEE Journ. of Solid-State Circuits*, vol. 35, no. 6, pp. 876–884, 2000.
- [123] D. Mizoguchi *et al.*, "A 1.2Gb/s/pin wireless superconnect based on inductive inter-chip signaling (IIS)," in *IEEE Int. Solid-State Circuits Conf.*, Feb 2004, pp. 142–517 Vol.1.
- [124] M. Takahashi *et al.*, "A 60-mw MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme," *IEEE Journ. of Solid-State Circuits*, vol. 33, no. 11, pp. 1772–1780, Nov 1998.
- [125] T. Nishikawa *et al.*, "A 60 MHz 240 mW MPEG-4 video-phone lsi with 16 Mb embedded DRAM," in *IEEE Int. Solid-State Circuits Conf.*, Feb 2000, pp. 230–231.
- [126] E. Iannone, *Labs on chip: Principles, design and technology*. CRC Press, 2018.
- [127] S. Behler, W. Teng, and A. Podpod, "Key properties for successful ultra thin die pickup," in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, 2017, pp. 95–101.
- [128] V. F. Pavlidis, I. Savidis, and E. G. Friedman, "Three-dimensional ICs with inductive links," in *Three-Dimensional Integrated Circuit Design (Second Edition)*, second edition ed. Boston: Morgan Kaufmann, 2017, pp. 137 – 161.
- [129] R. Ho and R. Drost, *Coupled Data Communication Techniques for High-Performance and Low-Power Computing*, Jan 2010.
- [130] A. M. Niknejad and R. G. Meyer, "Analysis and optimization of monolithic inductors

- and transformers for RF ICs,” in *Proc. of Custom Integrated Circuits Conf. (CICC)*, May 1997, pp. 375–378.
- [131] S. Pamarti, L. Jansson, and I. Galton, “A wideband 2.4-ghz delta-sigma fractional-npll with 1-mb/s in-loop modulation,” *IEEE Journ. of Solid-State Circuits*, vol. 39, no. 1, pp. 49–62, 2004.
- [132] N. Miura *et al.*, “A 2.7gb/s/mm² 0.9pj/b/chip 1 coil/channel thru chip interface with coupled-resonator-based cdr for nand flash memory stacking,” in *IEEE Int. Solid-State Circuits Conference*, Feb 2011, pp. 490–492.
- [133] N. Miura *et al.*, “A 0.55 v 10 fj/bit inductive-coupling data link and 07 V 135 fJ/Cycle clock link with dual-coil transmission scheme,” *IEEE Journ. of Solid-State Circuits*, vol. 46, no. 4, pp. 965–973, April 2011.
- [134] U. M. Jow and M. Ghovanloo, “Design and optimization of printed spiral coils for efficient transcutaneous inductive power transmission,” *IEEE Trans. on Biomedical Circuits and Systems*, vol. 1, no. 3, pp. 193–202, Sept 2007.
- [135] D. Ditzel and T. Kuroda, “Low-cost 3D chip stacking with ThruChip wireless connections,” in *IEEE Hot Chips 26 Symp.(HCS)*, Aug 2014.
- [136] K. Onizuka *et al.*, “Chip-to-chip inductive wireless power transmission system for sip applications,” in *IEEE Custom Integrated Circuits Conf.*, Sept 2006, pp. 575–578.
- [137] E. Culurciello and A. G. Andreou, “Capacitive inter-chip data and power transfer for 3-d vlsi,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 12, pp. 1348–1352, Dec 2006.
- [138] W. Mao *et al.*, “Analysis and design of high performance wireless power delivery using on-chip octagonal inductor in 65-nm cmos,” in *IEEE Int. System-on-Chip Conference (SOCC)*, Sept 2015, pp. 401–405.
- [139] S. Han and D. D. Wentzloff, “0.61w/mm² resonant inductively coupled power transfer for 3D-ICs,” in *Proc. of the IEEE Custom Integrated Circuits Conf.*, Sept 2012, pp. 1–4.
- [140] J. Mu *et al.*, “High-performance MIM capacitors for a secondary power supply application,” *Micromachines*, vol. 9, p. 69, 02 2018.
- [141] G. V. Merrett and A. S. Weddell, “Supercapacitor leakage in energy-harvesting sensor nodes: Fact or fiction?” in *International Conference on Networked Sensing (INSS)*, 2012, pp. 1–5.
- [142] E. Culurciello and A. G. Andreou, “Capacitive coupling of data and power for 3D

- silicon-on-insulator VLSI,” in *IEEE Int. Symp. on Circuits and Systems*, vol. 4, May 2005, pp. 4142–5.
- [143] N. Miura *et al.*, “A 1Tb/s 3W inductive-coupling transceiver for inter-chip clock and data link,” in *IEEE Int. Solid-State Circuits Conf.*, Feb 2006, pp. 1676–1685.
- [144] L. Zhang *et al.*, “A single phase modulation for pulse-based inductive-coupling connection in 3D stacked chip,” *IEICE Electronics Express*, vol. 14(20), no. 20, 2017.
- [145] N. Miura *et al.*, “A high-speed inductive-coupling link with burst transmission,” *IEEE J. of Solid-State Circ.*, vol. 44(3), pp. 947–55, 2009.
- [146] S. Gopal *et al.*, “Dual-equalization-path energy-area-efficient near field inductive coupling for contactless 3D IC,” in *IEEE MTT-S Int. Microwave Symp. (IMS)*, 2019.
- [147] Kiichi Niitsu *et al.*, “Interference from power/signal lines and to SRAM circuits in 65nm CMOS inductive-coupling link,” in *IEEE Asian Solid-State Circ. Conf.*, 2007.
- [148] A. Fazzi *et al.*, “3d capacitive interconnections with mono- and bi-directional capabilities,” in *IEEE Int. Solid-State Circuits Conf.*, Feb 2007, pp. 356–608.
- [149] Q. Gu *et al.*, “Two 10Gb/s/pin low-power interconnect methods for 3D ICs,” in *IEEE Int. Solid-State Circuits Conf.*, Feb 2007, pp. 448–614.
- [150] M. Ikebe *et al.*, “An image sensor/processor 3d stacked module featuring thru-chip interfaces,” in *Asia and South Pacific Design Automation Conf. (ASP-DAC)*, Jan 2017, pp. 7–8.
- [151] S. Han and D. D. Wentzloff, “Wireless power transfer using resonant inductive coupling for 3D integrated ICs,” in *IEEE Int. 3D Systems Integration Conference (3DIC)*, Nov 2010, pp. 1–5.
- [152] C. P. Yue and S. S. Wong, “Physical modeling of spiral inductors on silicon,” *IEEE Trans. on Electron Devices*, vol. 47, no. 3, pp. 560–568, Mar 2000.
- [153] K. B. Ashby *et al.*, “High Q inductors for wireless applications in a complementary silicon bipolar process,” *IEEE Journ. of Solid-State Circuits*, vol. 31, no. 1, pp. 4–9, Jan 1996.
- [154] N. M. Nguyen and R. G. Meyer, “Si IC-compatible inductors and LC passive filters,” *IEEE Journ. of Solid-State Circuits*, vol. 25, no. 4, pp. 1028–1031, Aug 1990.
- [155] I. E. Lager and G. Mur, “The finite element modeling of static and stationary electric and magnetic fields,” *IEEE Trans. on Magnetism*, vol. 32, no. 3, pp. 631–4, 1996.
- [156] T. Vali *et al.*, “Full-wave modeling of inductive coupling links for low-power 3D system

- integration,” in *IEEE Int. Symp. on Electromagnetic Compatibility*, Aug 2013, pp. 17–21.
- [157] J. C. Maxwell, *A treatise on electricity and magnetism*. Oxford: Clarendon Press, 1873, vol. 1.
- [158] S. Butterworth, “On the coefficients of mutual induction of eccentric coils,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journ. of Science*, vol. 31, no. 185, pp. 443–54, 1916.
- [159] S. Babic, F. Sirois, C. Akyel, and C. Girardi, “Mutual inductance calculation between circular filaments arbitrarily positioned in space: Alternative to grover’s formula,” *IEEE Transactions on Magnetics*, vol. 46, no. 9, pp. 3591–3600, 2010.
- [160] F. W. Grover, *Inductance calculations: working formulas and tables*. Courier Corporation, 2004.
- [161] G. Stojanovic and L. Zivanov, “Comparison of optimal design of different spiral inductors,” in *Int. Conf. on Microelectronics*, vol. 2, May 2004, pp. 613–616 vol.2.
- [162] S. S. Mohan *et al.*, “Simple accurate expressions for planar spiral inductances,” *IEEE Journ. of Solid-State Circuits*, vol. 34, no. 10, pp. 1419–1424, Oct 1999.
- [163] M. Bak, M. Dudek, and A. Dziedzic, “Chosen electrical and stability properties of surface and embedded planar PCB inductors,” in *Int. Spring Seminar on Electronics Technology*, May 2008, pp. 545–549.
- [164] J. R. Long and M. A. Copeland, “The modeling, characterization, and design of monolithic inductors for silicon RF IC’s,” *IEEE Journ. of Solid-State Circuits*, vol. 32, no. 3, pp. 357–369, Mar 1997.
- [165] M. Farran *et al.*, “Design, simulation and testing of planar spiral coils for the time gated interrogation of quartz resonator sensors,” 2014.
- [166] G. S. Smith, “Proximity effect in systems of parallel conductors,” *Journ. of Applied Physics*, vol. 43, no. 5, pp. 2196–2203, 1972.
- [167] Z. Piatek, B. Baron, T. Szczegielniak, D. Kusiak, and A. Pasierbek, “Self inductance of long conductor of rectangular cross section,” *Przegląd Elektrotechniczny Electr. Rev*, vol. 88, no. 9, pp. 323–326, 2012.
- [168] M. F. Chang *et al.*, “RF/wireless interconnect for inter- and intra-chip communications,” *Proc. of the IEEE*, vol. 89, no. 4, pp. 456–466, Apr 2001.
- [169] B. Noroozi and B. I. Morshed, “PSC optimization of 13.56-MHz resistive wireless analog passive sensors,” *IEEE Trans. on Microwave Theory and Techniques*, vol. PP, no. 99, pp. 1–8, 2017.

- [170] L. C. Hsu *et al.*, "Design and analysis for thurchip design for manufacturing (DFM)," in *Asia and South Pacific Design Automation Conf.*, Jan 2015, pp. 46–47.
- [171] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.
- [172] Ansys HFSS, "Ansys inc," *Canonsburg USA*, 1998.
- [173] W. Y. Yin, S. J. Pan, L. W. Li, and Y. B. Gan, "Experimental characterization of coupling effects between two on-chip neighboring square inductors," *IEEE Trans. on Electromagnetic Compatibility*, vol. 45, no. 3, pp. 557–561, Aug 2003.
- [174] Henkel Corporation, "Hysol QMI 538NB-1A1.5," <http://www.icproto.com/pdf/QMI%20538NB-1A1.5-EN.PDF>, 2011, accessed: 2018-03-27.
- [175] K. Bernstein, K. M. Carrig, C. M. Durham, P. R. Hansen, D. Hogenmiller, E. J. Nowak, and N. J. Rohrer, *High speed CMOS design styles*. Springer Science & Business Media, 1998.
- [176] *Datacon 220 Evo Plus*, Besi, 2019.
- [177] J. Wu, C. Zhao, Z. Lin, J. Du, Y. Hu, and X. He, "Wireless power and data transfer via a common inductive link using frequency division multiplexing," *IEEE Trans. on Industrial Electronics*, vol. 62, no. 12, pp. 7810–7820, 2015.
- [178] Y. Son and B. Jang, "Simultaneous data and power transmission in resonant wireless power system," in *Asia-Pacific Microwave Conference Proceedings (APMC)*, 2013, pp. 1003–1005.
- [179] B. J. Fletcher, S. Das, and T. Mak, "A high-speed design methodology for inductive coupling links in 3D-ICs," in *Design, Automation Test in Europe Conf. Exhibition (DATE)*, March 2018, pp. 497–502.
- [180] S. Haider *et al.*, "A comparative study of small voltage rectification circuits for implanted devices," in *IOP Conference Series: Materials Science and Engineering*, vol. 53, no. 1, 2013, p. 012024.
- [181] H. Cha, W. Park, and M. Je, "A CMOS rectifier with a cross-coupled latched comparator for wireless power transfer in biomedical applications," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 59, no. 7, pp. 409–413, 2012.
- [182] M. E. C. Andam *et al.*, "A design of self-biased cross coupled rectifier with integrated dual threshold voltage for rf energy harvesting application," *Procedia Computer Science*, vol. 109, pp. 384 – 391, 2017.
- [183] B. Razavi, "The strongarm latch [a circuit for all seasons]," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 2, pp. 12–17, 2015.

- [184] J. Myers *et al.*, “A subthreshold ARM Cortex-M0+ subsystem in 65nm CMOS for WSN applications with 14 power domains, 10T SRAM, and integrated voltage regulator,” *IEEE Journ. of Solid-State Circuits*, vol. 51, no. 1, pp. 31–44, 2016.
- [185] *Apollo2 Blue Datasheet*, Ambiq Micro, Nov 2017.
- [186] A. Shrivastav, G. Tomar, and A. Singh, “Performance comparison of amba bus-based system-on-chip communication protocol,” 07 2011, pp. 449 – 454.
- [187] Q. Tian and Z. Jin, “CMOS image sensor interface controller design for video surveillance,” in *2012 Int.Conference on Systems and Informatics (ICSAI2012)*, 2012, pp. 2165–2169.
- [188] P. Ceminari, A. Arelovich, and M. Federico, “AES block cipher implementations with amba-ahb interface,” in *1st Conference on PhD Research in Microelectronics and Electronics Latin America (PRIME-LA)*, 2017, pp. 1–4.
- [189] S. Sreehari and J. Jacob, “Ahb ddr sdram enhanced memory controller,” in *2013 Int.Conference on Advanced Computing and Communication Systems*, 2013, pp. 1–8.
- [190] *AMBA 3 AHB-Lite Protocol*, ARM Limited, Nov 2006.
- [191] Y. Yoshida, N. Miura, and T. Kuroda, “A 2 Gb/s bi-directional inter-chip data transceiver with differential inductors for high density inductive channel array,” *IEEE Journ. of Solid-State Circuits*, vol. 43, no. 11, pp. 2363–69, 2008.
- [192] J. Janesky, “Impact of external magnetic fields on MRAM products,” *Freescale Semiconductor Application Note AN3525 Nov*, 2007.
- [193] J. Heidecker, “MRAM technology status,” 2013.
- [194] Y. Shang *et al.*, “Thermal-reliable 3D clock-tree synthesis considering nonlinear electrical-thermal-coupled TSV,” 01 2013.
- [195] I. A. Papistas and V. F. Pavlidis, “Efficient modeling of crosstalk noise on power distribution networks for contactless 3-D ICs,” *IEEE Trans. on Circ. and Systems I*, vol. 65(8), pp. 2547–58, 2018.
- [196] D. Wang *et al.*, “A 65-nm cmos constant current source with reduced PVT variation,” *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 4, pp. 1373–85, 2017.
- [197] A. Zolfaghari, A. Chan, and B. Razavi, “Stacked inductors and transformers in CMOS technology,” *IEEE Journ. of Solid-State Circuits*, vol. 36, no. 4, pp. 620–628, Apr 2001.

- [198] B. Nikolic *et al.*, “Improved sense-amplifier-based flip-flop: design and measurements,” *IEEE Journ. of Solid-State Circuits*, vol. 35, no. 6, pp. 876–84, June 2000.
- [199] J. Kim *et al.*, “High-frequency scalable electrical model and analysis of a through silicon via (TSV),” *IEEE Trans. on Components, Packaging and Manufacturing Technology*, vol. 1, no. 2, pp. 181–195, Feb 2011.
- [200] R. Weerasekera, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, “Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs,” in *IEEE/ACM Int. Conf. on Computer-Aided Design*, 2007, pp. 212–219.
- [201] S. Lim, “Design for high performance, low power, and reliable 3D integrated circuits,” pp. 285–308, 11 2013.
- [202] P. Sethi and S. R. Sarangi, “Internet of things: architectures, protocols, and applications,” *Journ. of Electrical and Computer Engineering*, vol. 2017, 2017.
- [203] R. Stephens, “What the dual-dirac model is and what it is not,” 2006.
- [204] R. Stephens, “Jitter analysis: The dual-Dirac model, RJ/DJ, and Q-scale,” *Agilent Technical Note*, 2004.
- [205] Cadence, Virtuoso, “SpectreRF Simulation Option User Guide,” *Cadence Design Systems*, 2006.