

RUN-TIME POWER AND ENERGY MANAGEMENT OF MULTI- AND MANY- CORE SYSTEMS

Dr Geoff Merrett

Tutorial on the Spectrum of Run-time Management for Modern and Next Generation Multi/Many-core Systems
30 September 2018 | ESWeek, Torino, Italy

ECS/SOUTHAMPTON

University of Southampton

- ~25,000 students
- Top 100 universities worldwide (QS'19)
- Founding member of UK's Russell Group

School of ECS

- ~1,500 students
- ~300 PhD research students
- ~110 academics/faculty
- Top 3 in UK for EEE
- 14 research groups/centres



THE PRiME PROJECT

“Enable the sustainability of **many-core scaling** by preventing the uncontrolled increase in **energy consumption** and **unreliability** through a step change in holistic design methods and **cross-layer** system optimisation.”

UNIVERSITY OF
Southampton

Imperial College
London

MANCHESTER
1824

Newcastle
University

arm

Imagination

intel

Microsoft Research

NXP

EPSRC
Engineering and Physical Sciences
Research Council

nmi
Semiconductors
to Systems

Innovate UK
Knowledge Transfer Network

www.prime-project.org

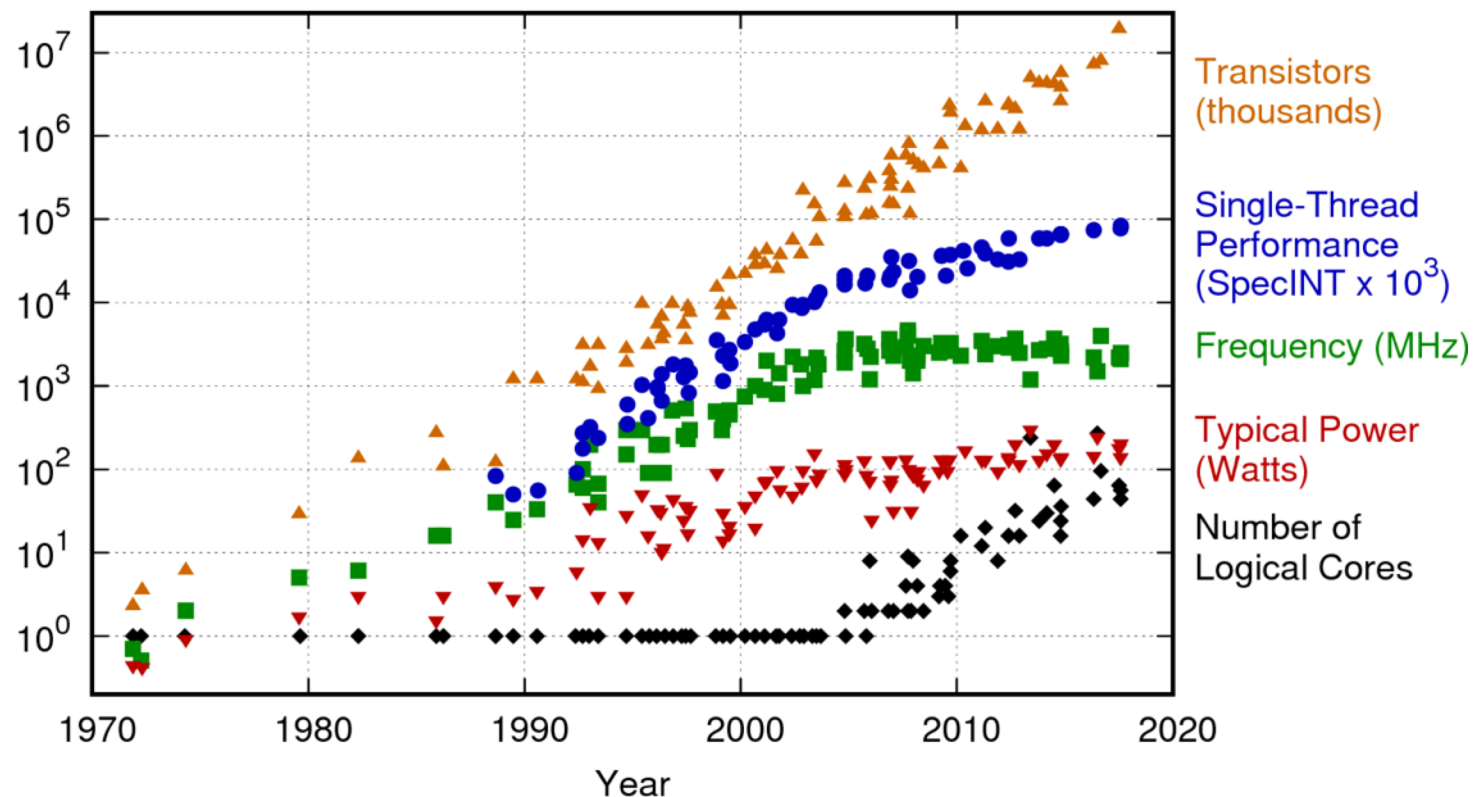
MORE COMPUTE FOR THE SAME POWER



A “*supercomputer in your hands*”, running 100s of cores with a battery lasting for a day

WE ARE MANY-CORE

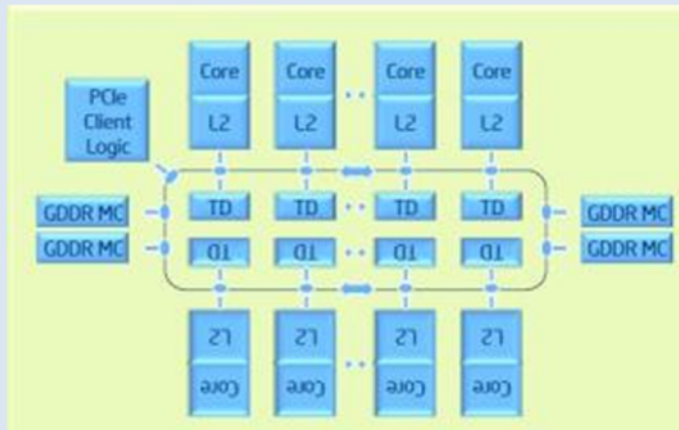
42 Years of Microprocessor Trend Data



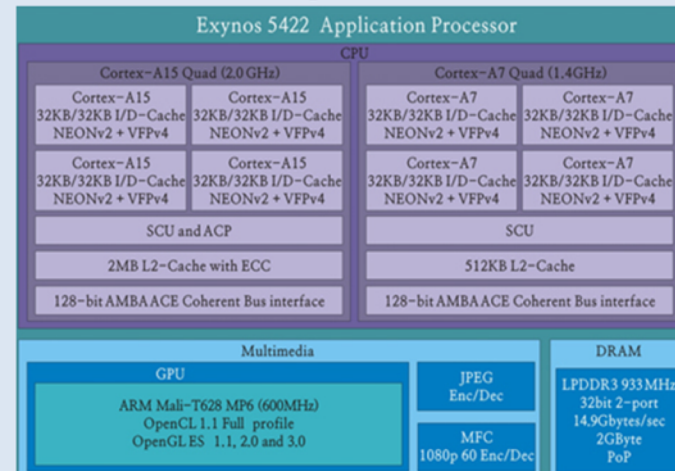
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

MANY-CORE PLATFORMS

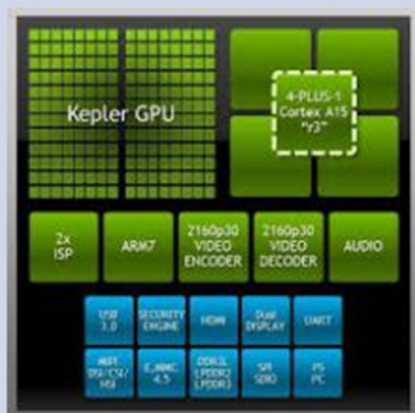
Intel Xeon Phi - **Homogeneous 61 Cores**



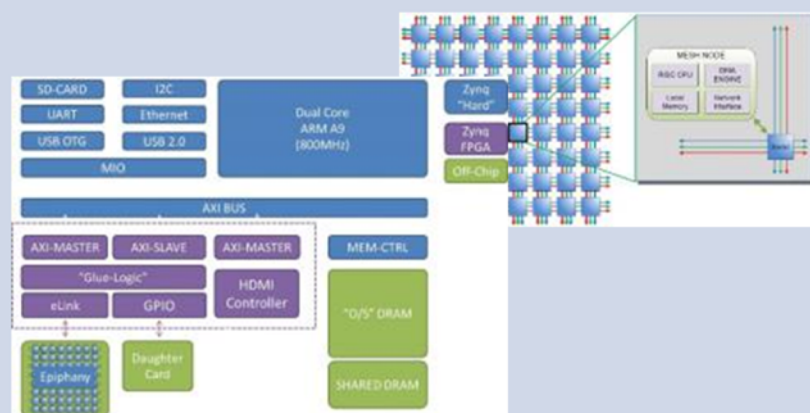
ODROID XU3 – **8 core big.LITTLE CPU + 6 cores GPU**



Nvidia Jetson TK1 - **Quad core CPU + 192 cores GPU**

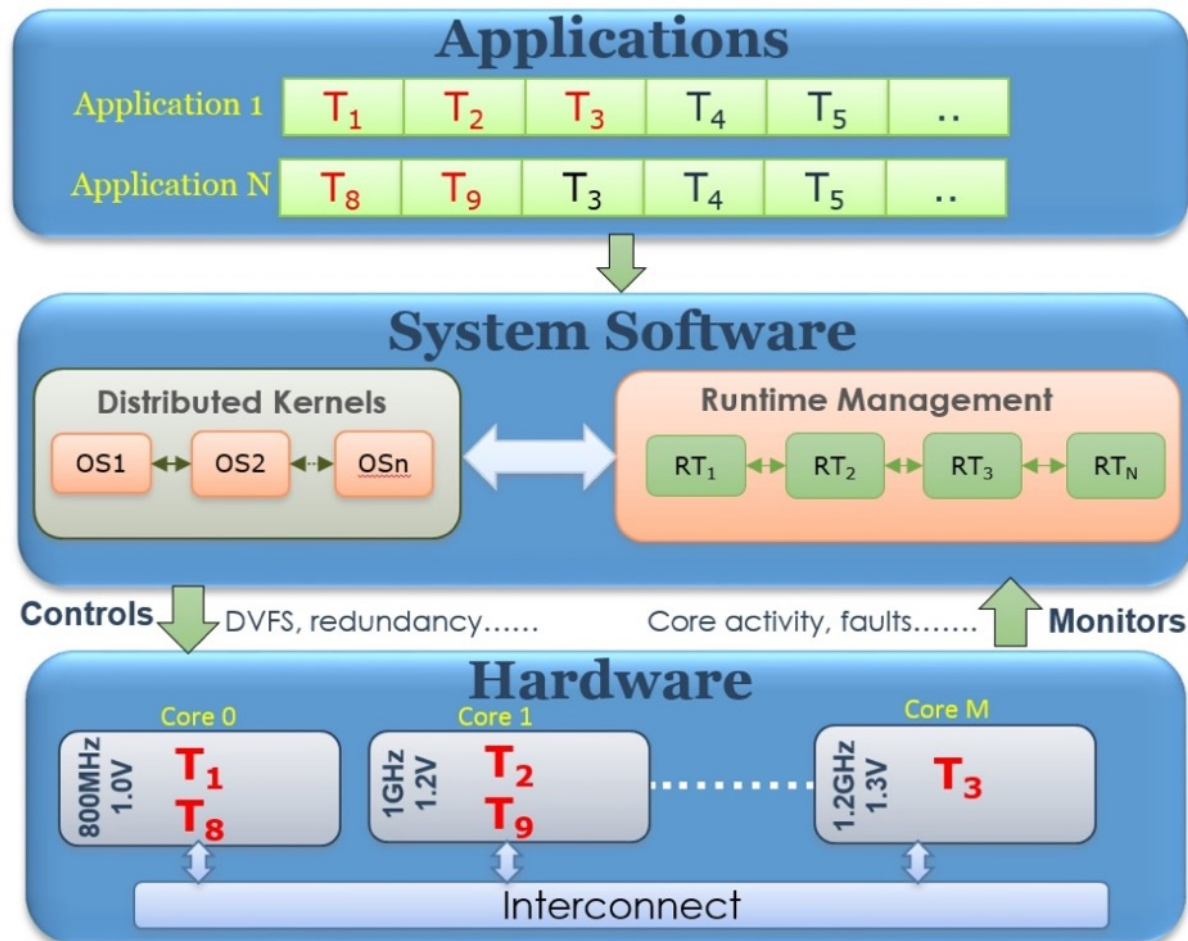


Parallella - **Dual core CPU + FPGA + 16 cores NoC**

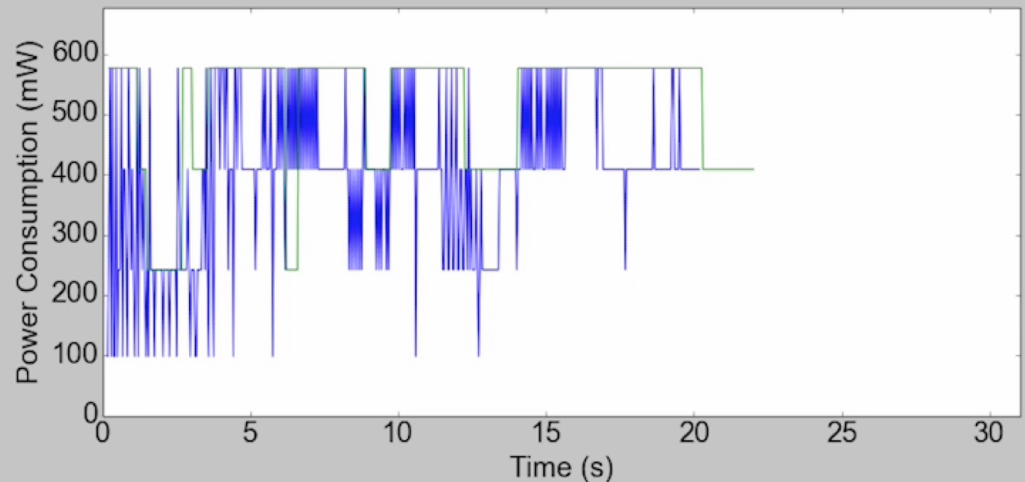
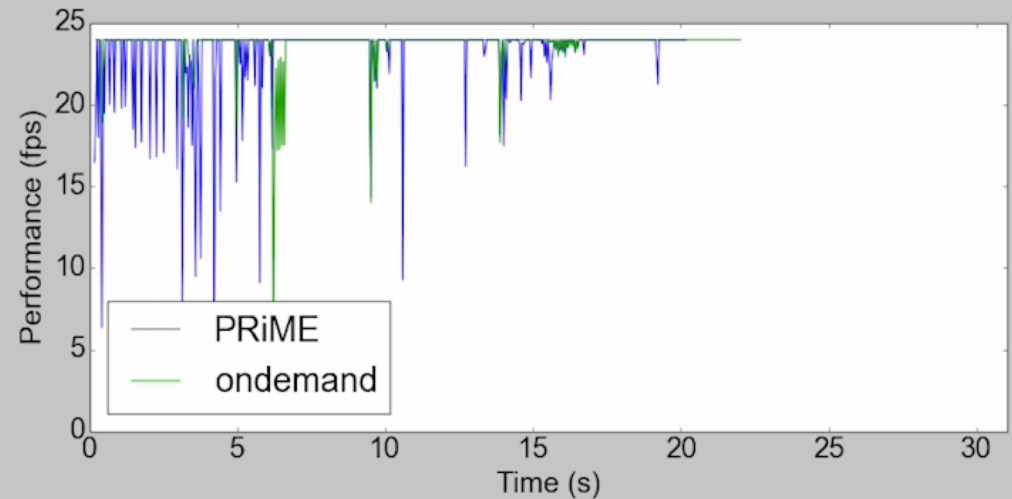


THE PRiME PROJECT

www.prime-project.org



RUNTIME POWER MANAGEMENT

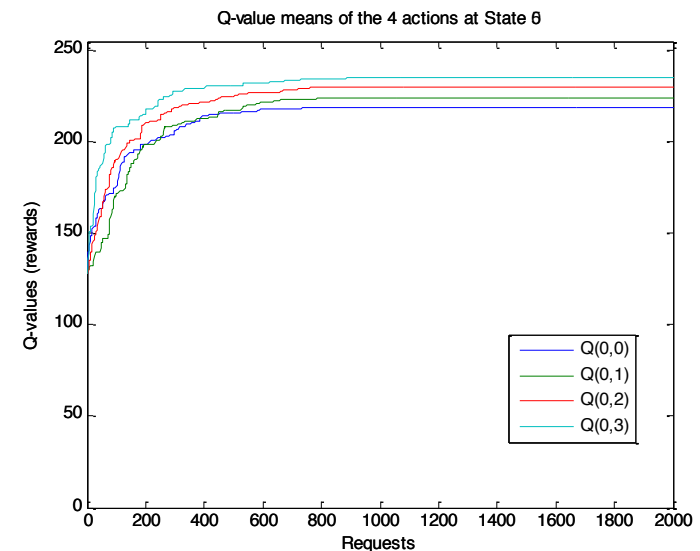
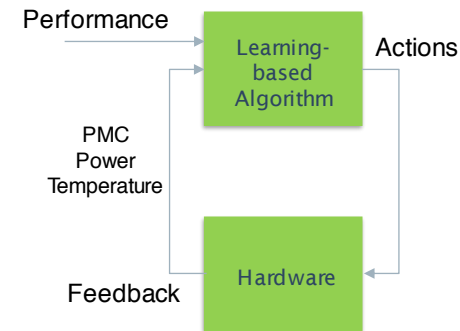


LEARNING OPTIMAL DVFS CHOICES

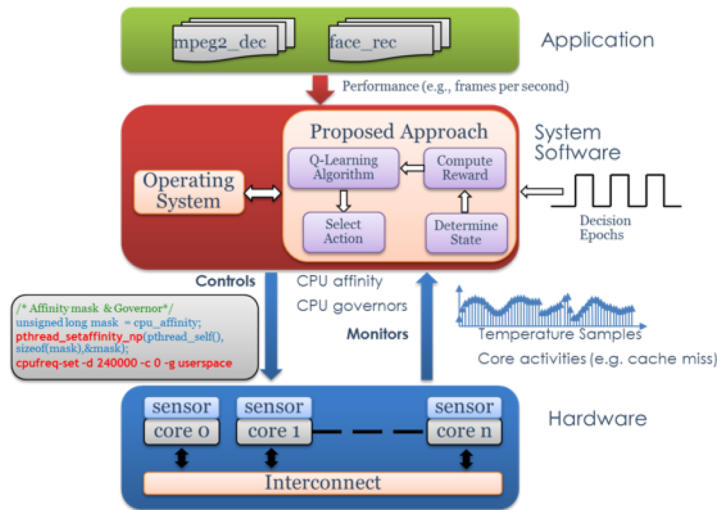
Reinforcement Learning

- Observes the current system state
- Selects an action (V-F pairs)
- Changes the state (workload)
- Leads to a payoff (reward/penalty)

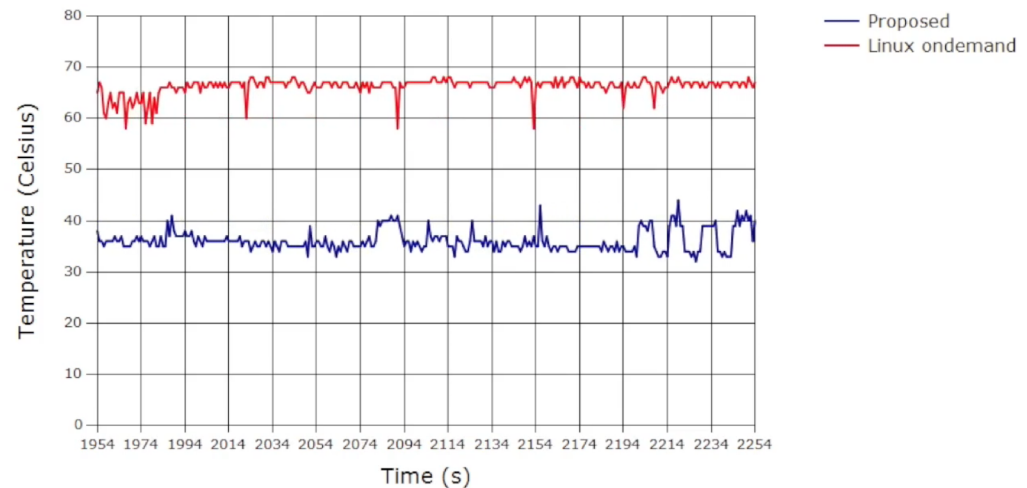
	ACTIONS (Power Modes)			
STATES (Tasks)	P0	P1	P2	P3
WD0	128	128	128	128
WD1	128	128	128	128
WD2	128	128	128	128
WD3	128	128	128	128
WD4	128	128	128	128
WD5	128	128	128	128



MANAGING THERMAL (LIFETIME) RELIABILITY



Convergence of the Reinforcement Learning Algorithm



Application	Data Set	Average Temperature (Celcius)			Peak Temperature (Celcius)		
		Linux	Ge et al.	Proposed	Linux	Ge et al.	Proposed
tachyon	set 1	69.2	52.6	38.6	71.5	63	60
	set 2	50.5	44.5	43.8	57.3	56.3	52
	set 3	50.8	44.7	41.6	57.8	54.5	48.8
mpeg2_dec	clip 1	36	34	34.2	42.7	41.3	39
	clip 2	35.6	34.4	34.2	42.3	42	39.3
	clip 3	34.3	34.4	34	43	39.7	44.3

Average MTTF improvements: 5x (thermal aging); 4x (thermal cycling)

OVERVIEW

Applications

- From single > sequential > concurrent execution

Offline Characterisation

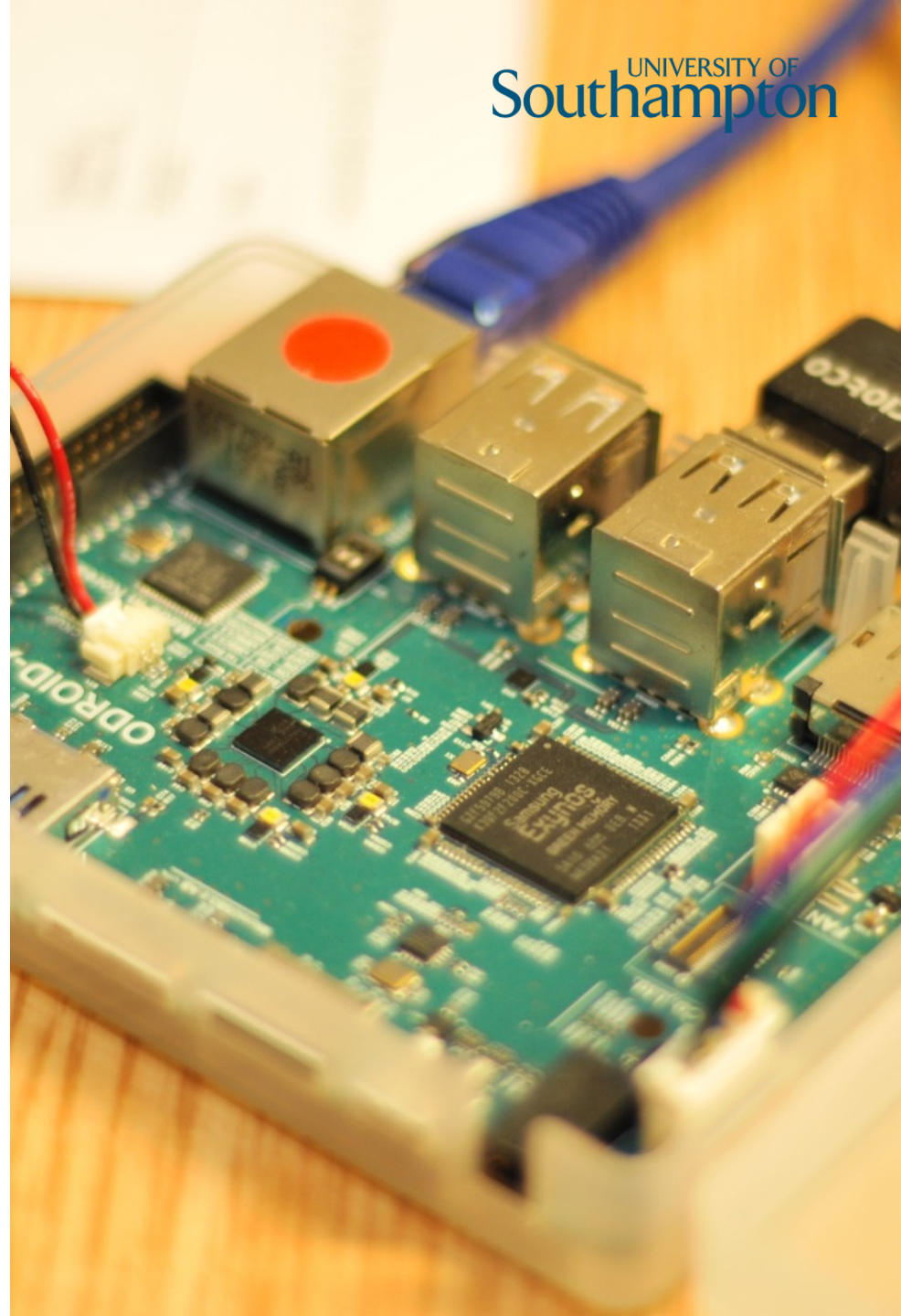
- Can we improve RTM through offline characterisation?

Towards Many-Core

- How do RTM approaches scale with number of cores?

Novel Platforms

- Can our RTM approaches be applied to novel platforms?



RTMs and Application Workloads

From single > sequential > concurrent execution

QUALITY OF EXPERIENCE

- User cares about **observable performance**
 - Responsiveness, battery life, consistency, uninterrupted service
 - Doesn't really care about FLOPS, FPS, bandwidth, latency (QoS)
- Therefore, optimise for **quality of user experience** (QoE)
 - “*good-enough*” performance
 - Minimum energy usage

Bischoff, Alexander S. (2016) *User-experience-aware system optimisation for mobile systems*, University of Southampton, Electronics and Computer Science, Doctoral Thesis , 199pp.

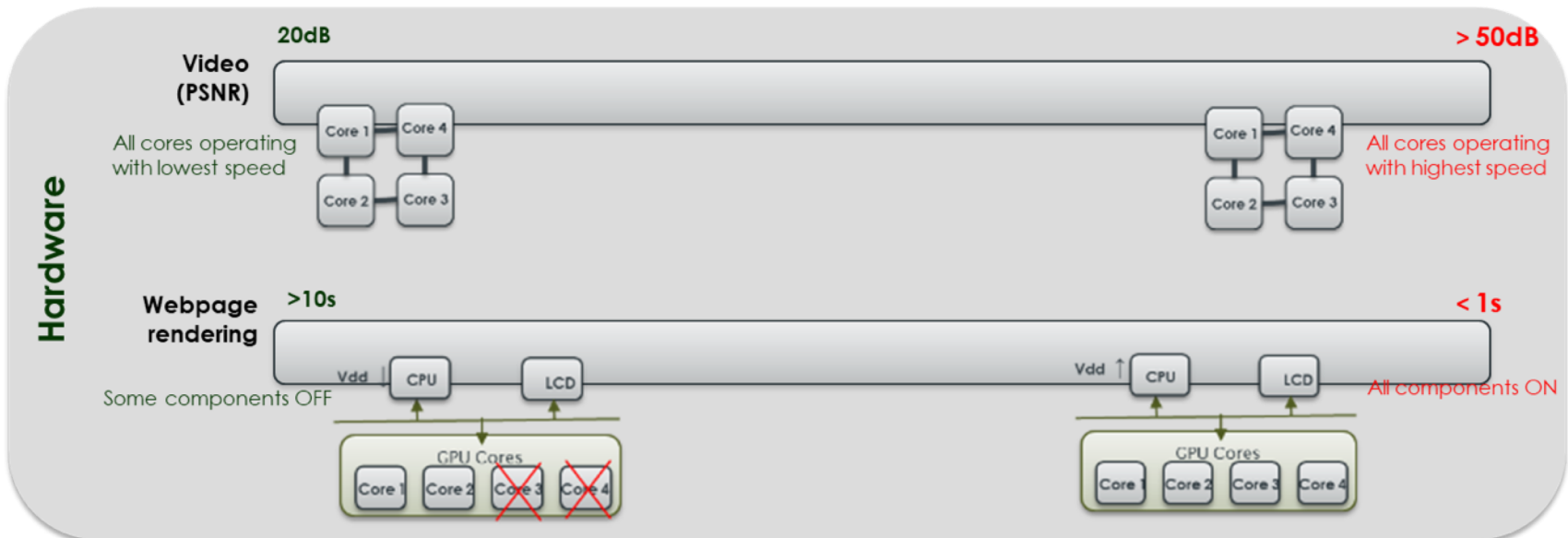
Bischoff S, Hansson A and Al-Hashimi BM. *Applying of Quality of Experience to System Optimisation*. International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS), Germany, 2013.

QUALITY OF EXPERIENCE

Example Scenario



Runtime Management



QUALITY OF EXPERIENCE

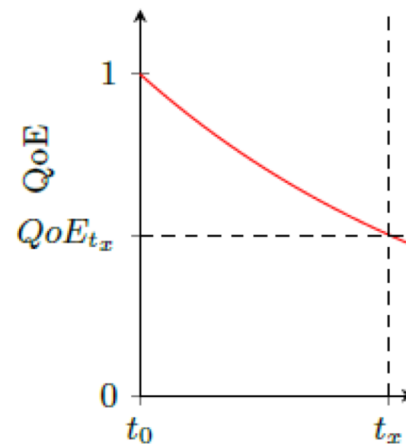
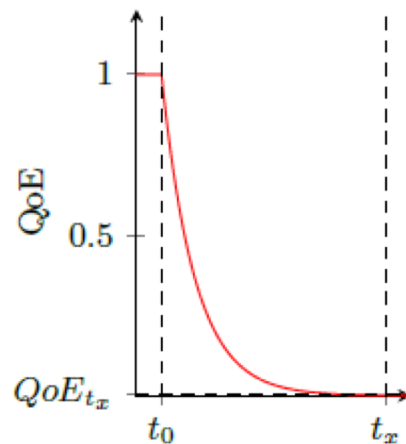
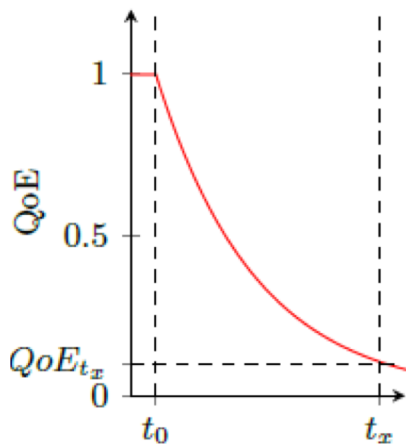
Workload Classification

Applications	Type of QoE
Audio	Throughput
Video	Throughput
Application Loading	Latency
Web Page Loading	Latency
Downloading a File	Latency
3D Gaming	Throughput
Word Processing	Latency

Delay of
<0.1 s appears instant,
1 s becomes noticeable
10 s become disruptive'

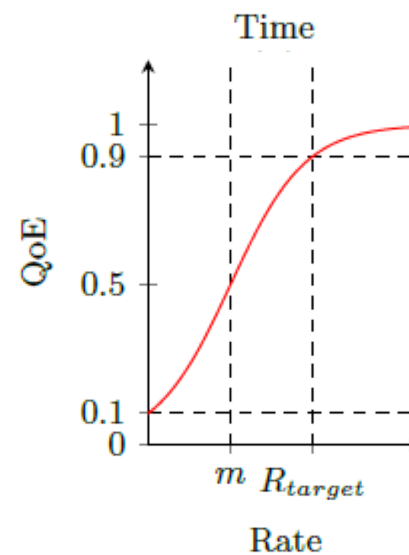
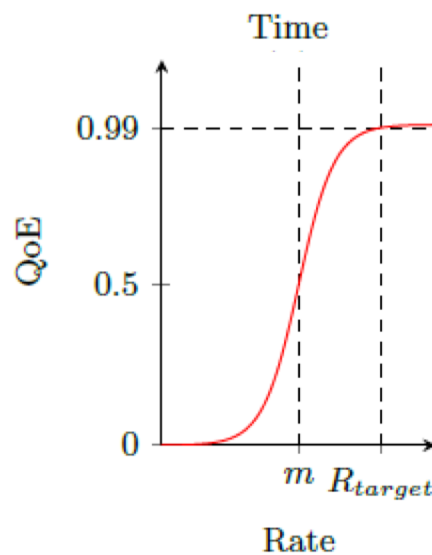
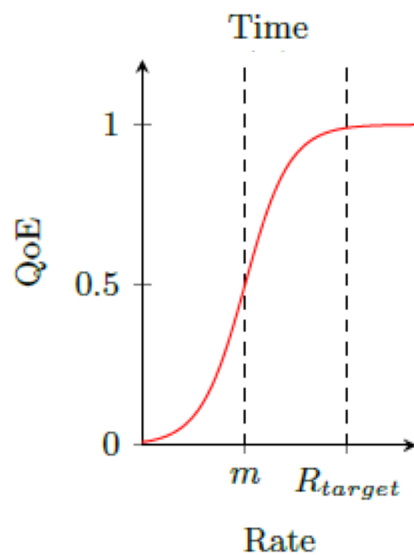
- Types of QoE:
 - Latency sensitive - complete workload in short time period
 - Throughput sensitive - complete at minimum rate

QoE CHARACTERISTICS



Latency
Sensitive

Inverse exponential



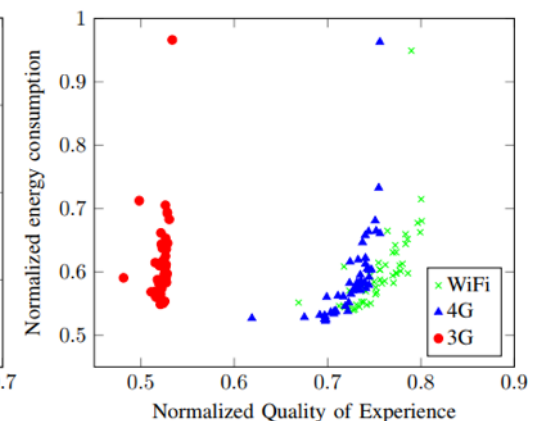
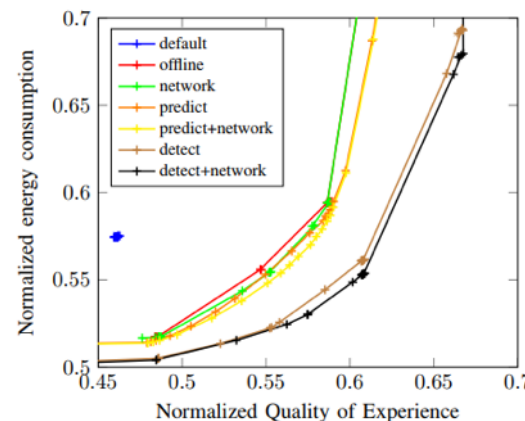
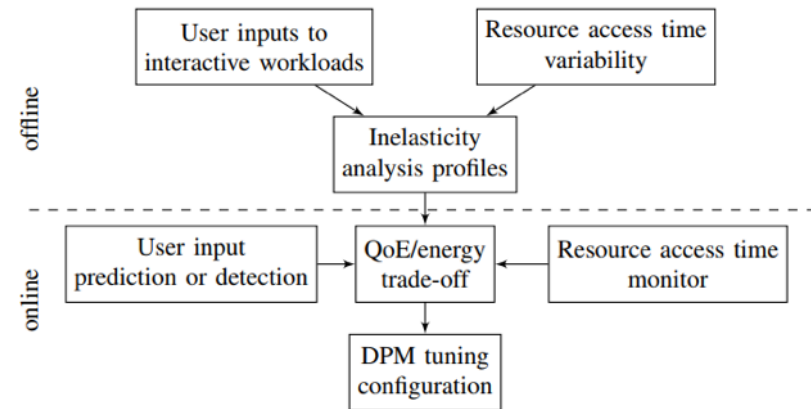
Throughput
Sensitive

Sigmoid function

TUNING DPM/RTM PARAMETERS

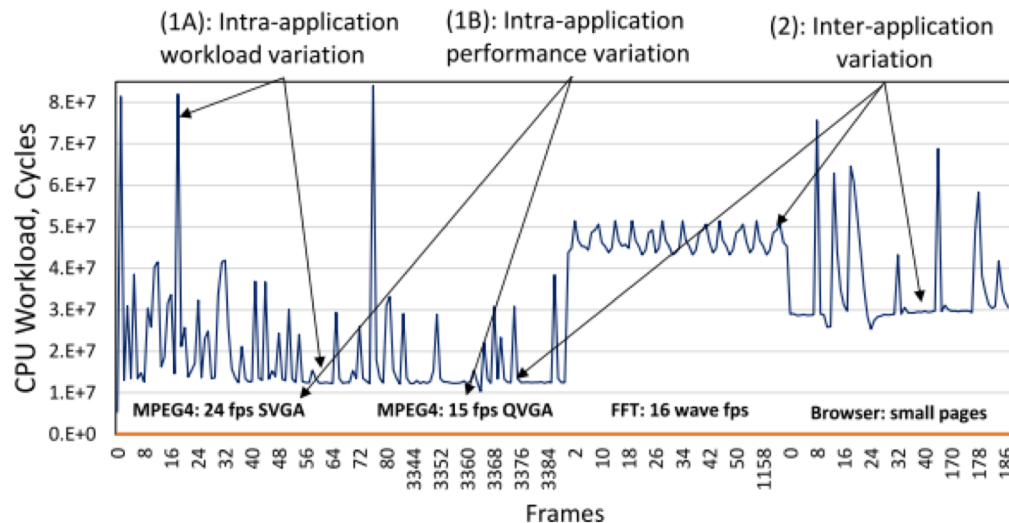
- Tune governor parameters for the executing (interactive) workload
- Account for variability in access times and user input
- Prediction/detection dependent
- Energy saving/QoE improvement compared to ‘default’, e.g.
 - 13% energy saving
 - 27% QoE improvement
 - 9% energy + 15% QoE

Exynos-5422 A15/A7, Android 6.0
Google Chrome browser workloads
Touch input emulation
Network throttling (UL, DL, RTT latency)



EXECUTING MULTIPLE APPLICATIONS

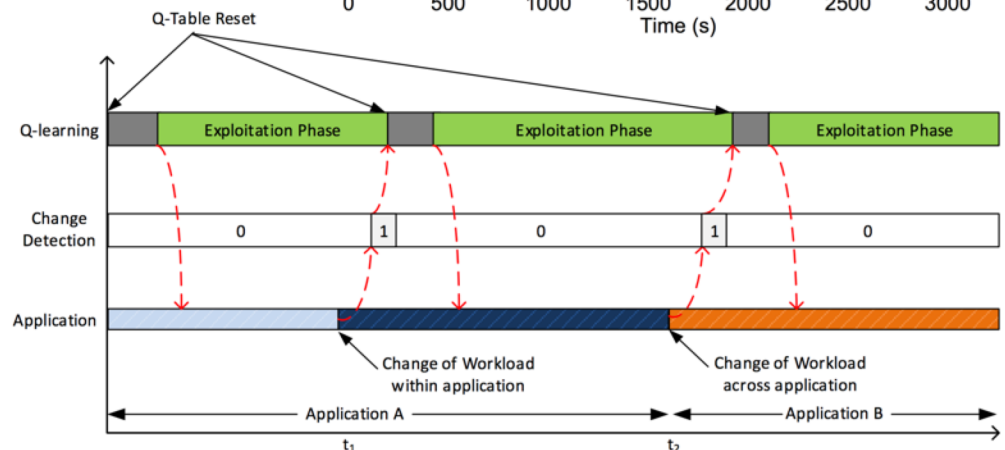
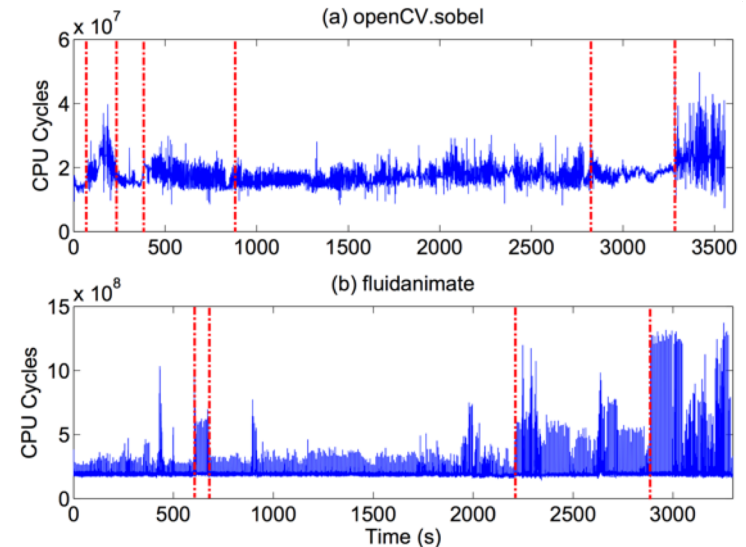
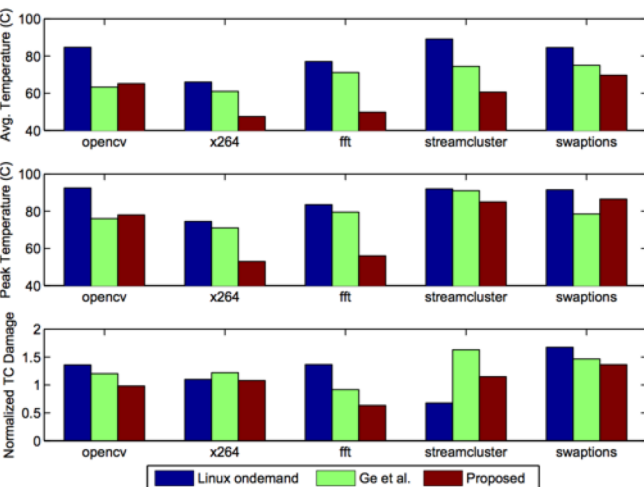
- Workload and performance variation due to:
 - Changes within an application
 - Changing applications (*sequential execution*)



- Overlapping applications (*concurrent execution*)

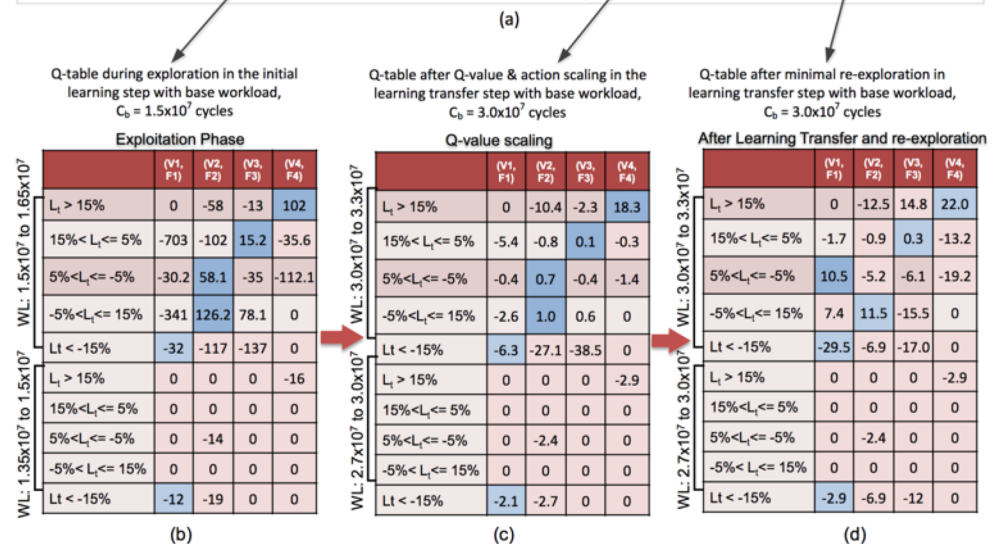
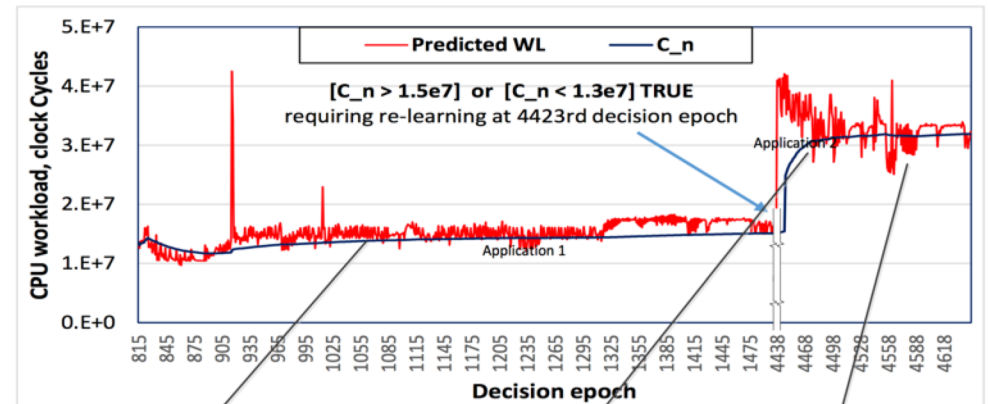
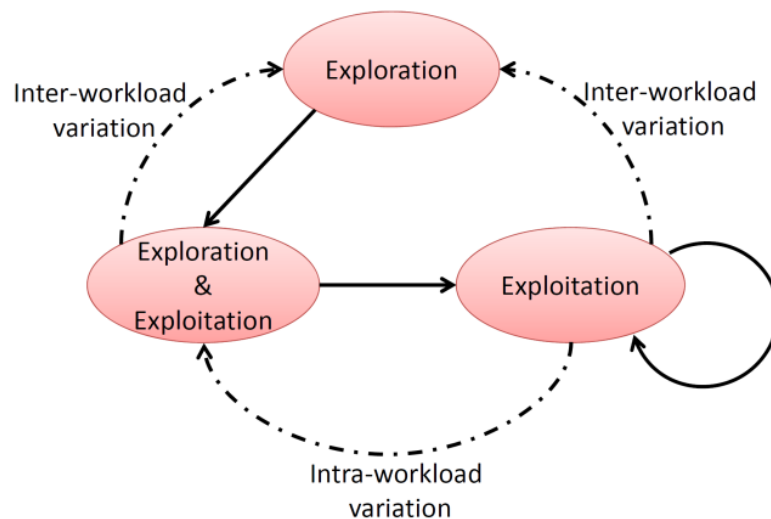
DETECTING WORKLOAD CHANGES

- Density ratio-based statistical divergence between overlapping sliding windows of CPU cycles
- Use this information to clear learning table (i.e. start afresh)



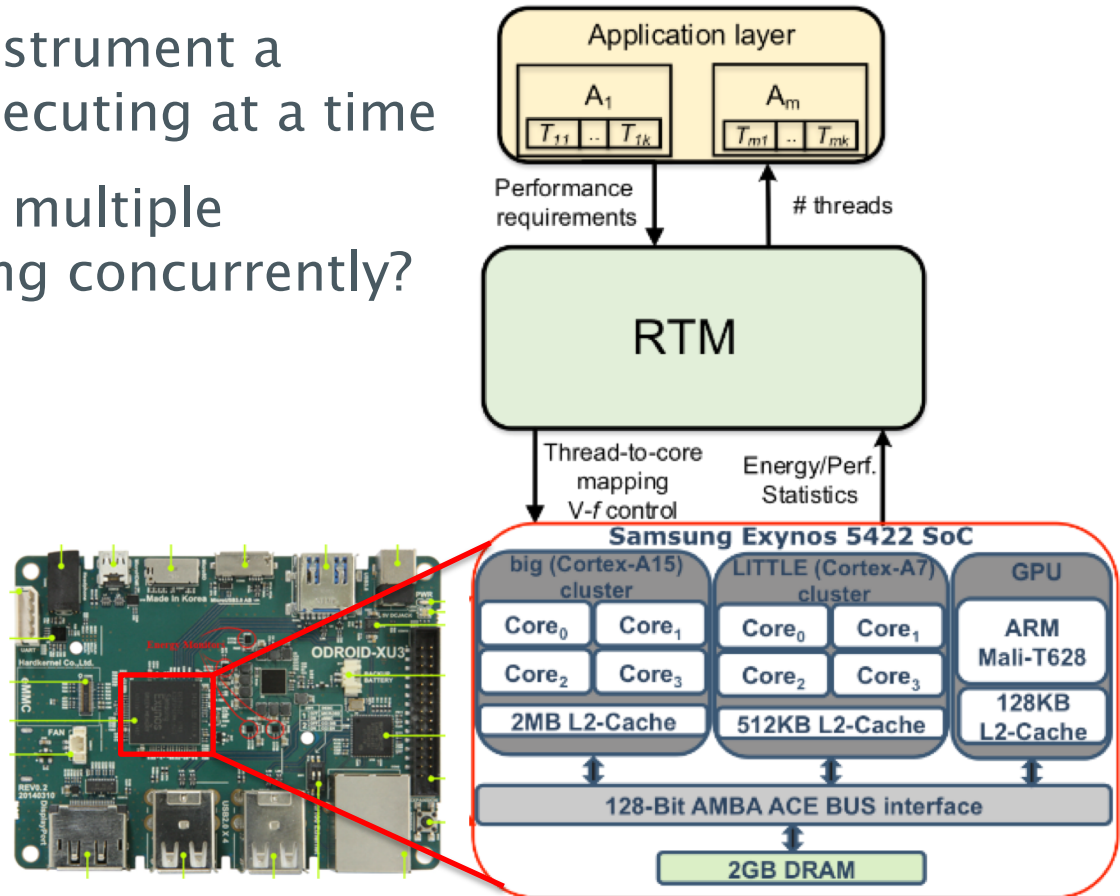
TRANSFERRING LEARNING

- Detect workload changes
- Transfer knowledge where possible
- Learn again fresh when not



RTM FOR CONCURRENT EXECUTION

- Approaches so far instrument a single application executing at a time
- How can we manage multiple applications executing concurrently?

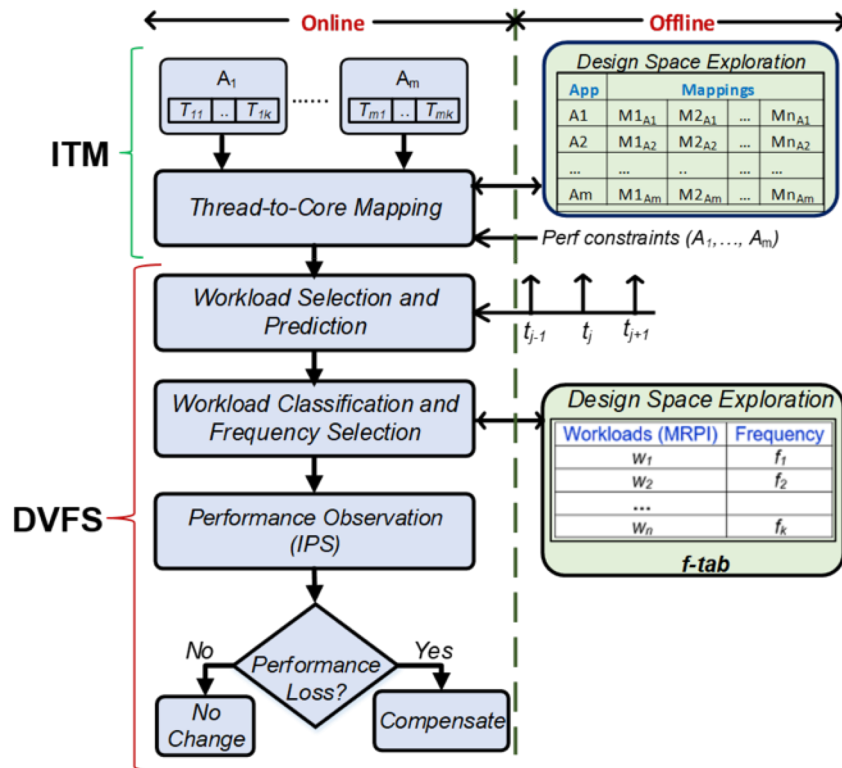


Online vs Offline

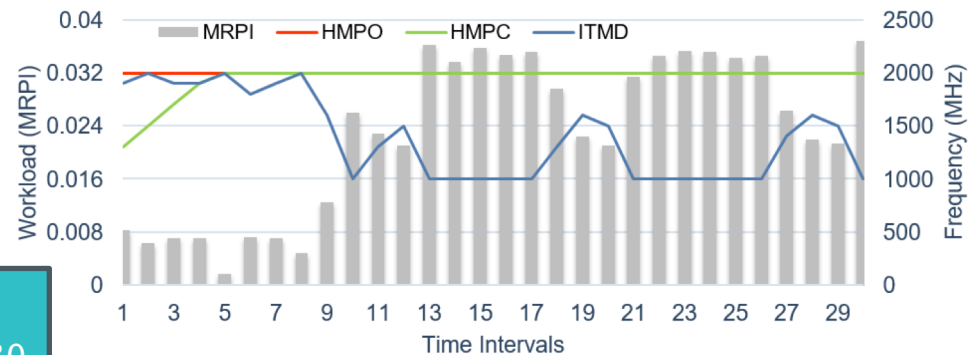
Can we improve RTM through offline characterisation?

RTM FOR CONCURRENT EXECUTION

MRPI (Memory Reads Per Instruction)



- Supports concurrent execution of applications
- Inter-cluster Thread-to-core Mapping (ITM).
- MRPI informs DVFS control



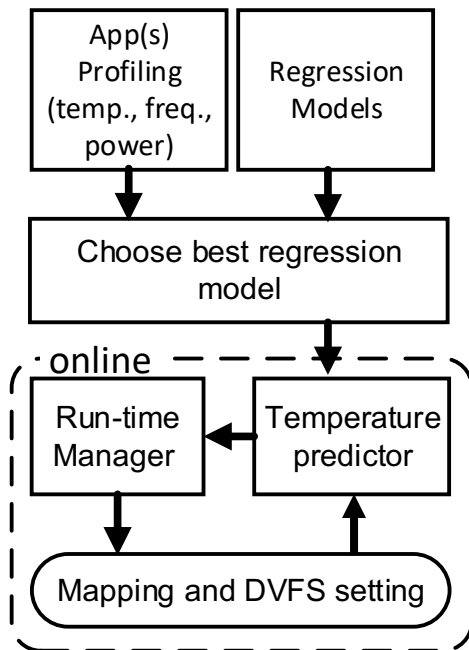
S1.6 Wednesday 17:15

Poster and Demo
Session Thursday 14:30

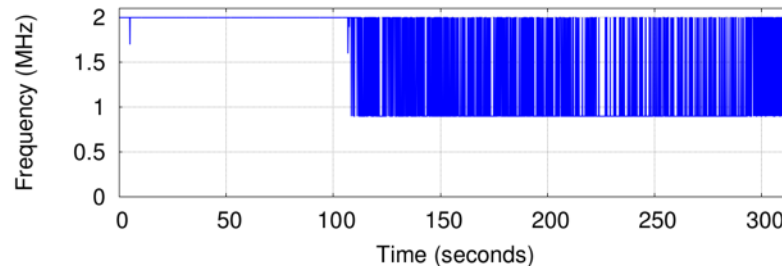
MANAGING TEMPERATURE

MRPI (Memory Reads Per Instruction)

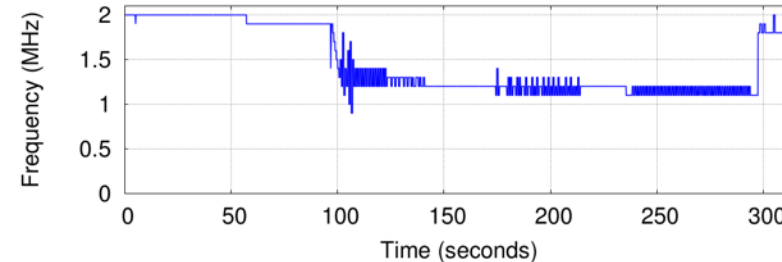
- Aims to avoid frequency throttling at temperature threshold.
- Predicts temperature using a regression-based model



Core frequency without prediction



Core frequency with prediction



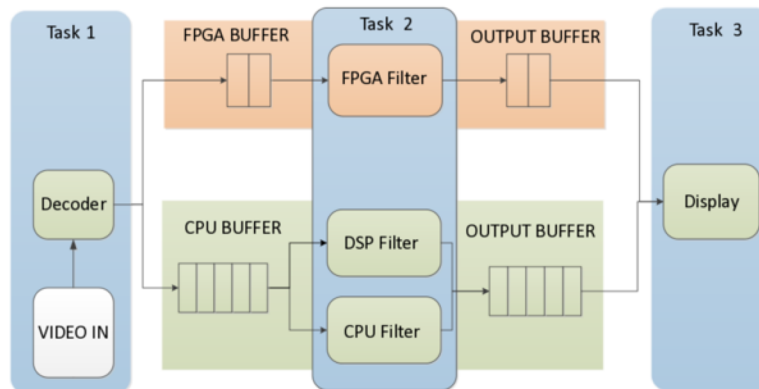
- Can achieve a 10% improvement in energy **and** performance

MODEL-BASED RTM: HETEROGENEITY

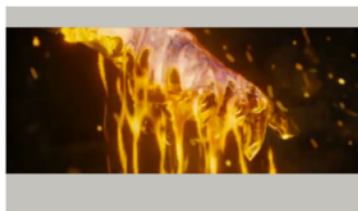
Heterogeneous Platforms

Run-time changes in:

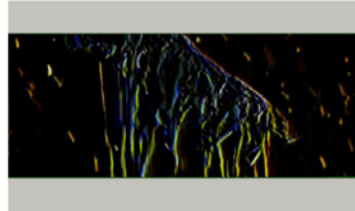
- Performance requirements
- Application workload changes



(a) Convolution filter implementation



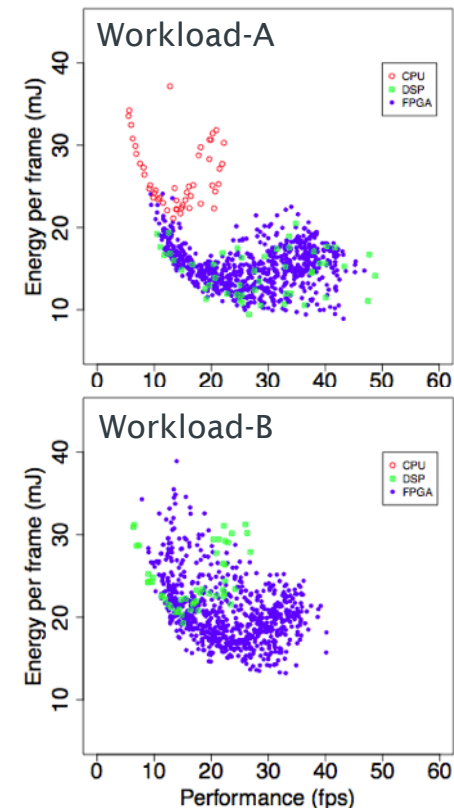
(b) Original



(c) Edge detected

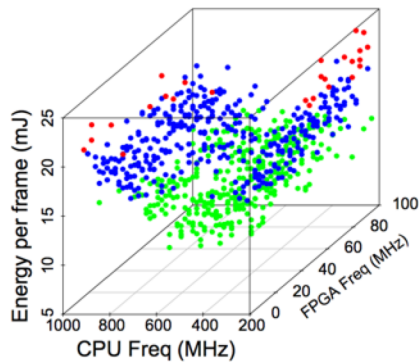


(d) Blurred

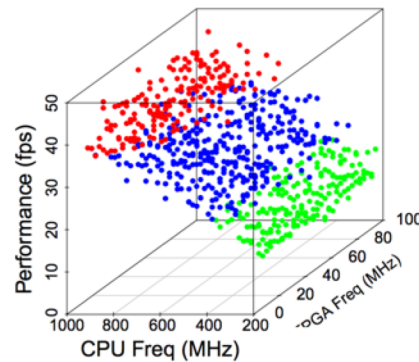


MODEL-BASED RTM: HETEROGENEITY

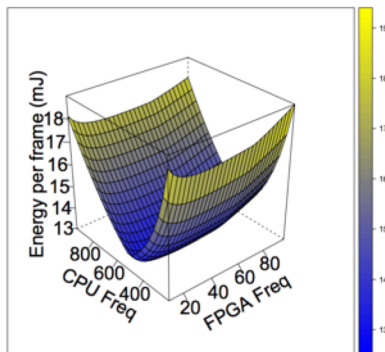
Heterogeneous Platforms



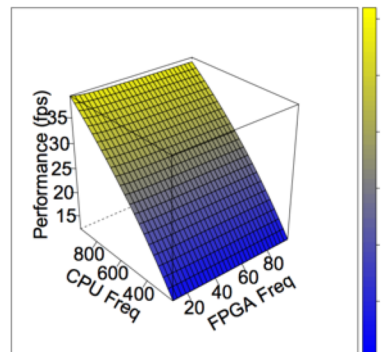
(a) FPGA measured energy



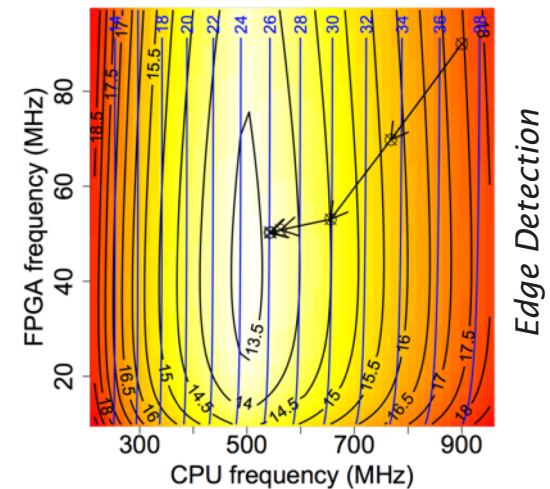
(b) FPGA measured performance



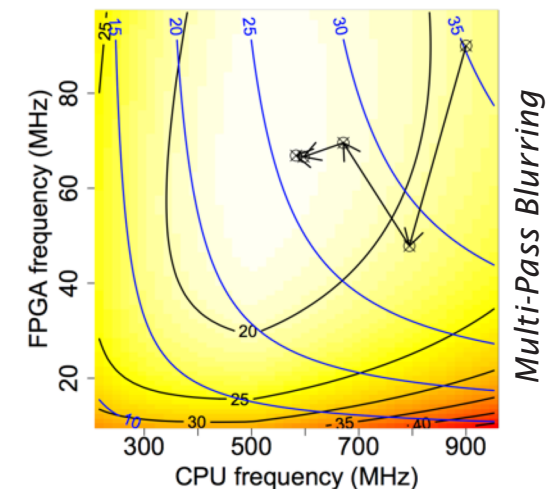
(c) FPGA modeled energy



(d) FPGA modeled performance



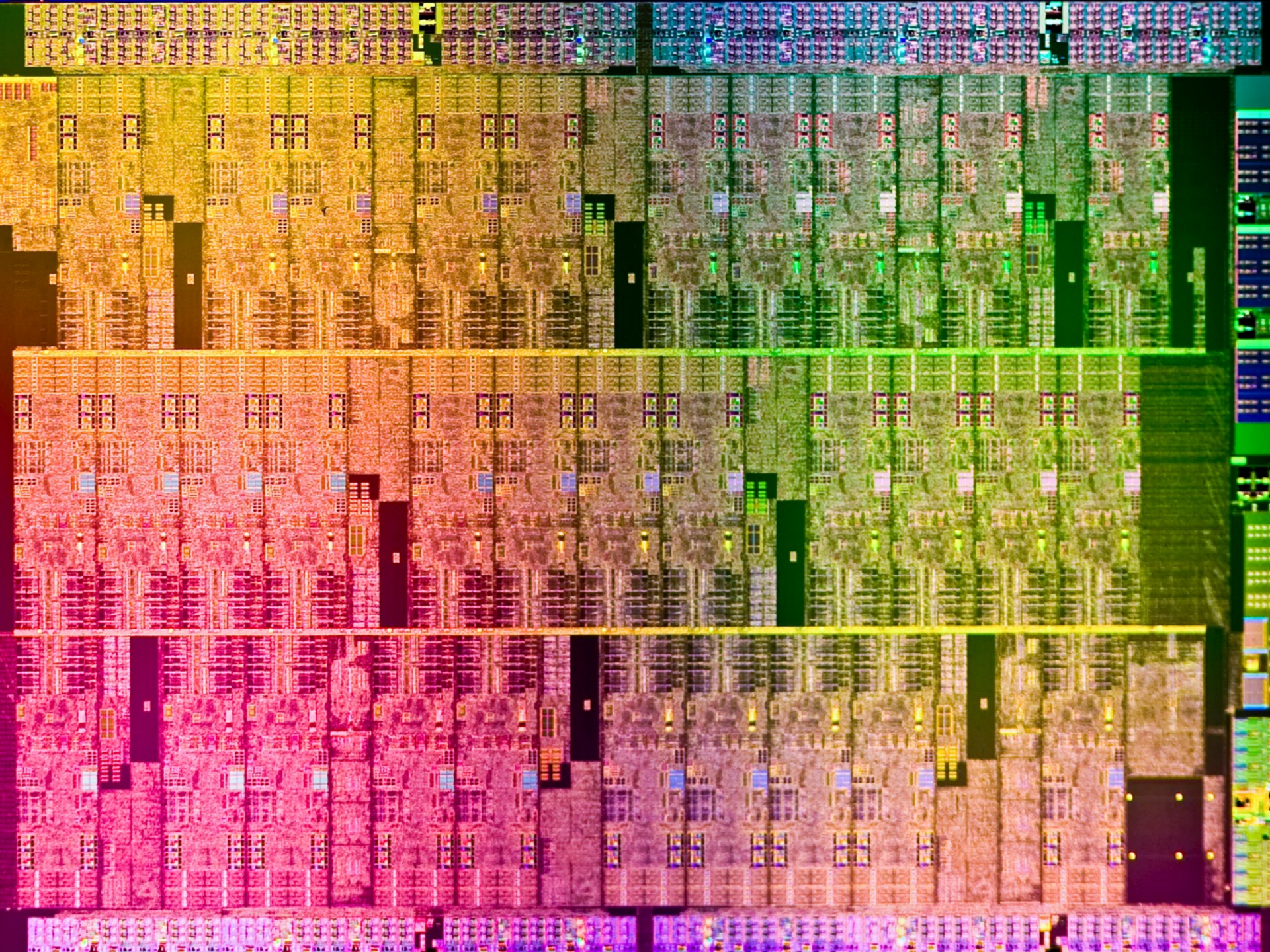
Edge Detection



Multi-Pass Blurring

Towards Many-Core

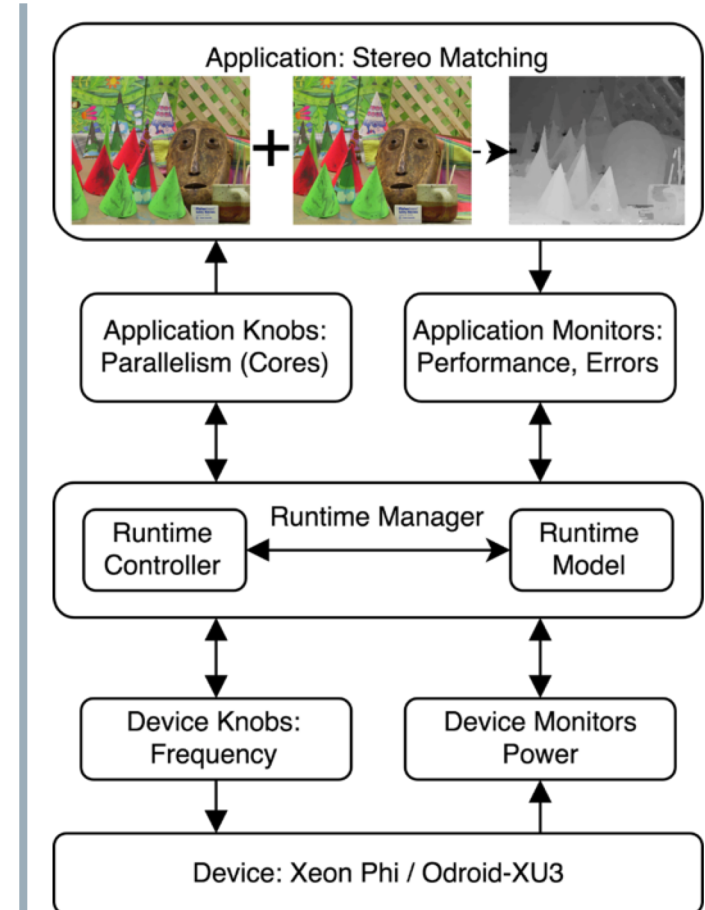
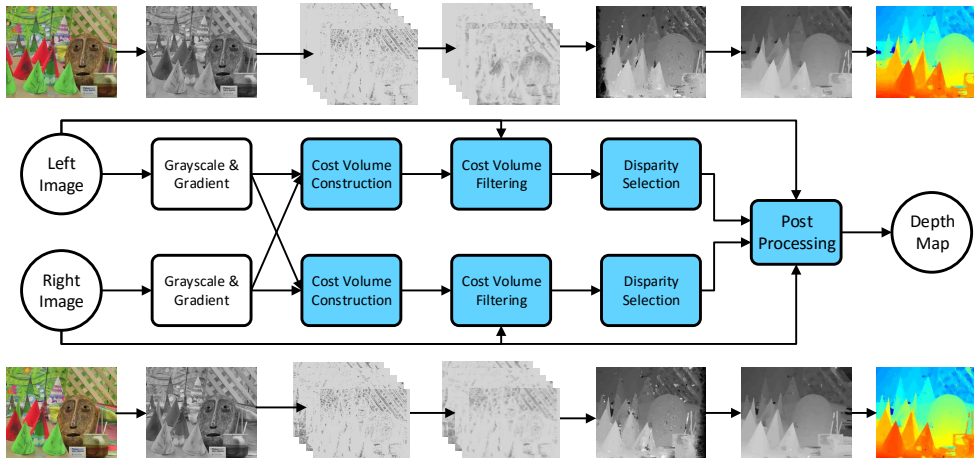
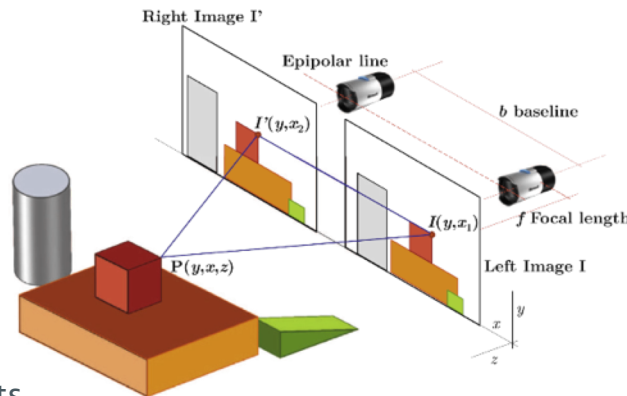
How do RTM approaches scale
with number of cores?



MODEL-BASED RTM

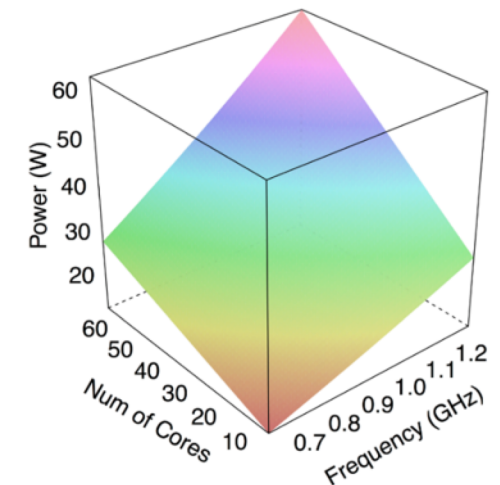
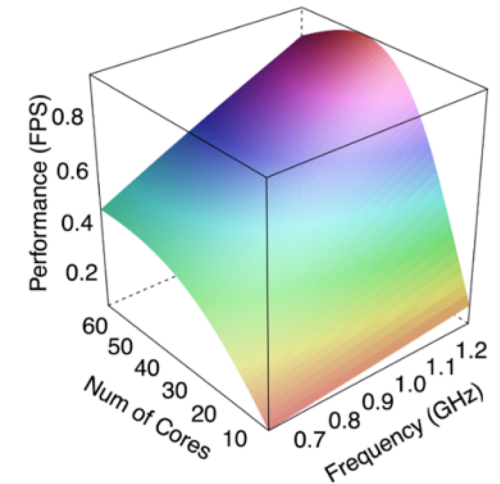
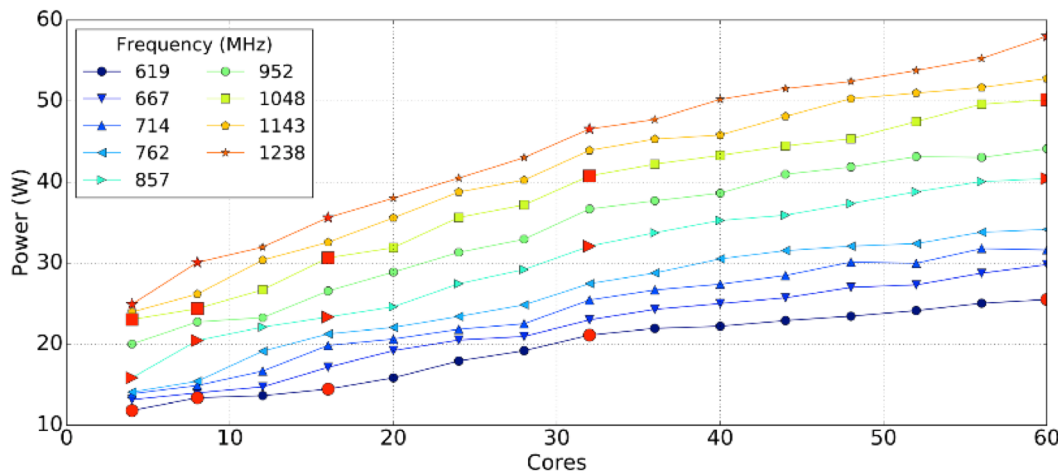
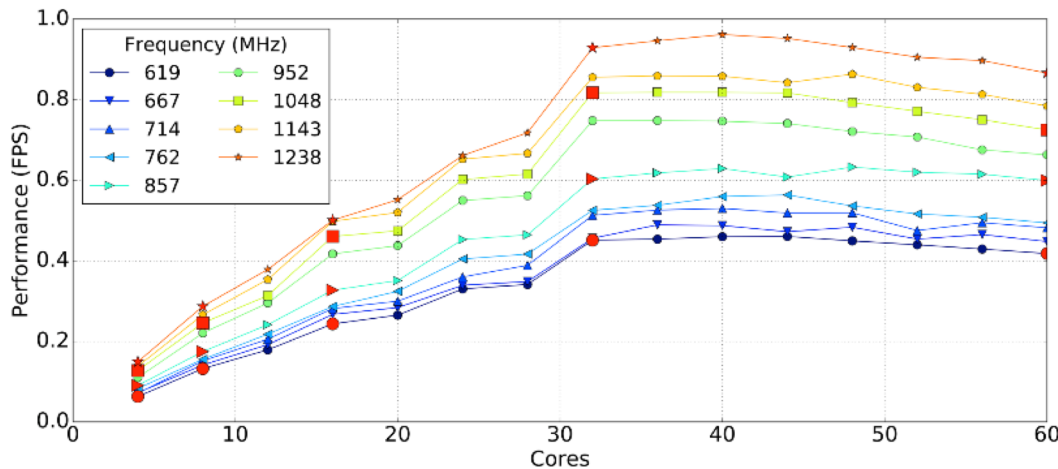
Stereo Matching Application: <http://github.com/PRiME-project/PRiMEStereoMatch>

- Processes still images, video or a camera feed
- OpenCL supported
- Includes test datasets



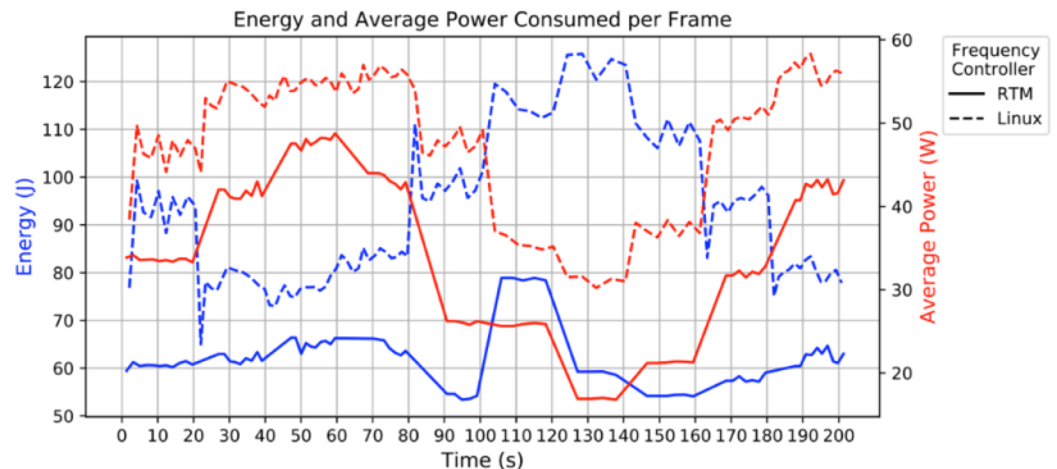
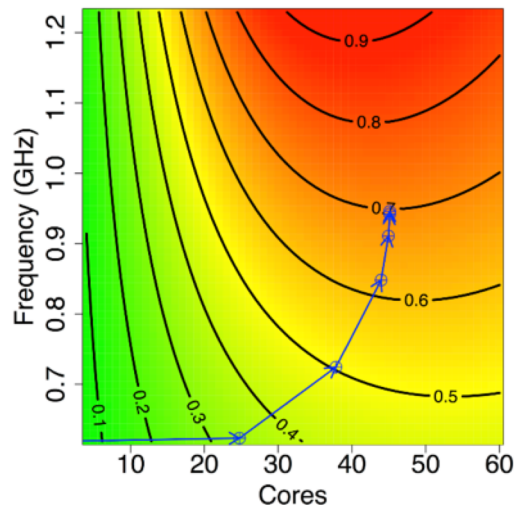
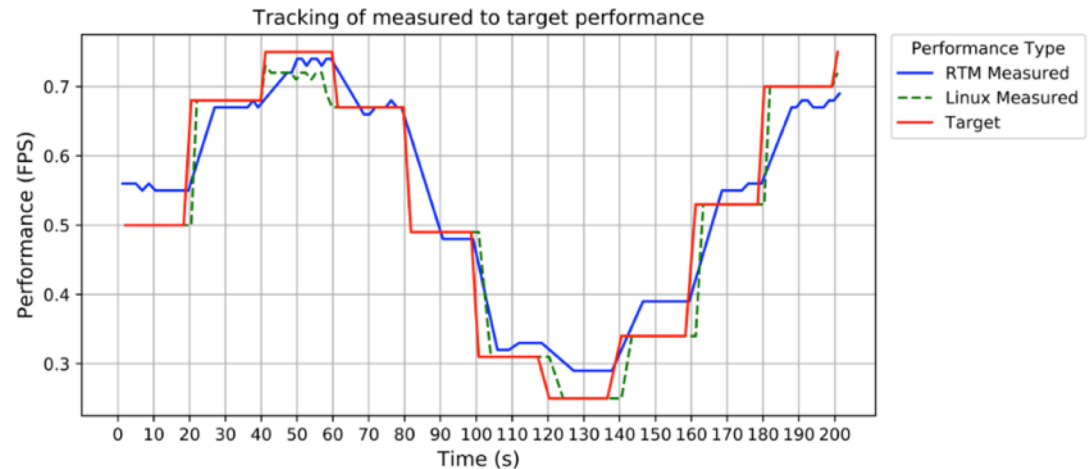
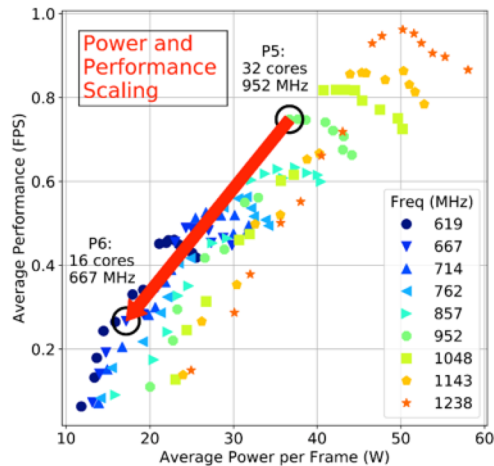
MODEL-BASED RTM

Model Building



MODEL-BASED RTM

Runtime Management



ENERGY RTM ON HPC SYSTEMS

- Applications targeted for HPC are usually multi-threaded
- Modern HPC often based on Non-Uniform Memory Access (NUMA) architecture
- Our Approach:
 - Platform characterized offline
 - Workload estimated based on memory-intensity, thread synchronization contention, NUMA latency
 - V - f determined using binning, while accounting for contention due to concurrent execution

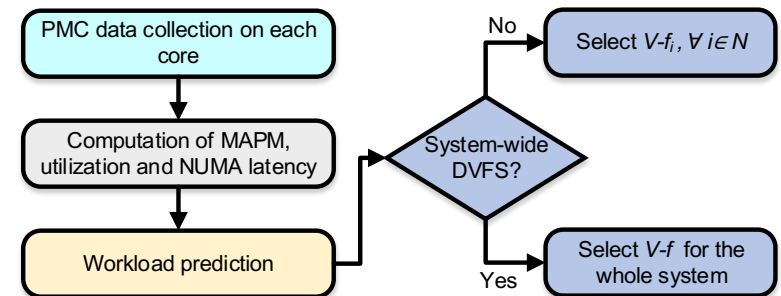
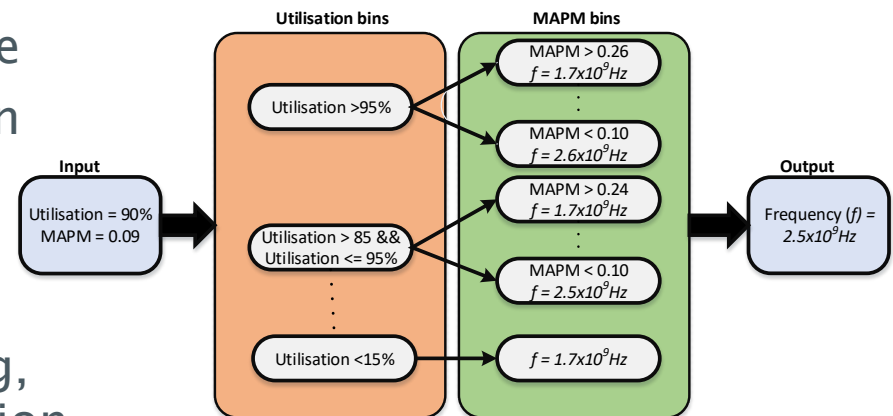
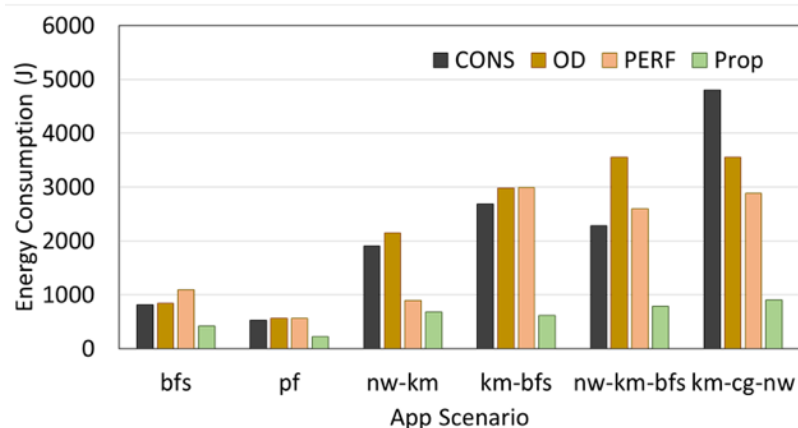


Illustration of various steps in the proposed approach

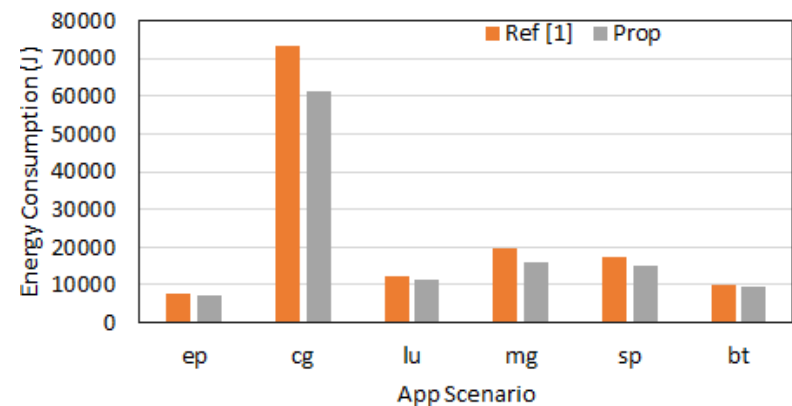


An example of V - f setting selection using binning-based approach

ENERGY RTM ON HPC SYSTEMS



Energy consumption of different approaches



Comparison of presented approach with Sundriyal *et al*

- Xeon E5-2630 (12 cores, 24 threads) and Xeon Phi 7620P (61 cores, 244 threads); NAS and Rodinia benchmarks
- Proposed (Prop) approach achieves energy savings of up to 81% (Xeon) and 61% (Phi) compared to Linux's governors
- Outperforms Sundriyal *et al.* by 10% in energy efficiency and 3.7% in performance

RTM of Novel Platforms

Can our RTM approaches be applied to novel platforms?

RTM ON “SPINNAKER”

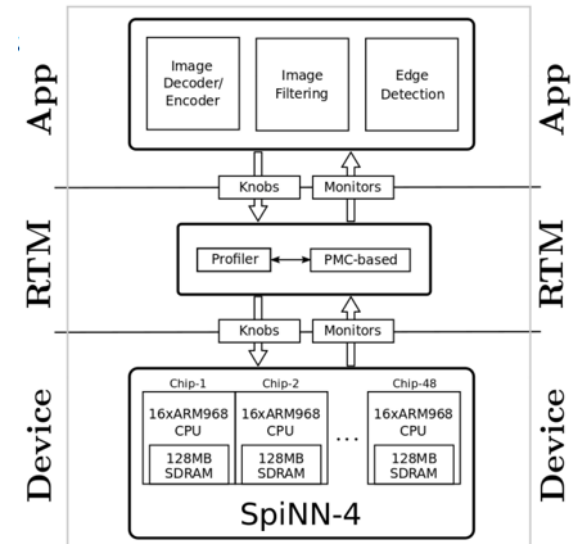
- Implemented 4 RTMs
 - User (G1):** user-defined static f
 - On-demand (G2):** Highest f when CPU load is high, lowest when it's low
 - Conservative (G3):** Increase or decrease f by fixed step according to load.
 - Proposed (G4):** As G3, but using a non-linear f step

App.	Res.	Governor			
		G1	G2	G3	G4
A1	vga	955	976	976	975
	svga	1490	1522	1522	1523
	xga	2444	2498	2498	2498
A2	vga	2670	3080	3080	3080
	svga	4408	4737	4737	4737
	xga	7114	7342	7342	7342
A3	vga	437	454	454	451
	svga	674	696	696	697
	xga	1111	1150	1150	1150

Timing (ms)

App.	Res.	Governor			
		G1	G2	G3	G4
A1	vga	2.76	1.98	2.11	2.27
	svga	6.40	5.06	5.05	5.12
	xga	17.79	13.74	13.82	13.74
A2	vga	8.24	6.84	7.16	7.06
	svga	22.20	16.46	17.02	16.13
	xga	58.72	39.29	40.95	39.29
A3	vga	9.41	7.16	7.17	6.62
	svga	17.07	13.01	13.08	11.90
	xga	48.30	37.19	36.87	33.92

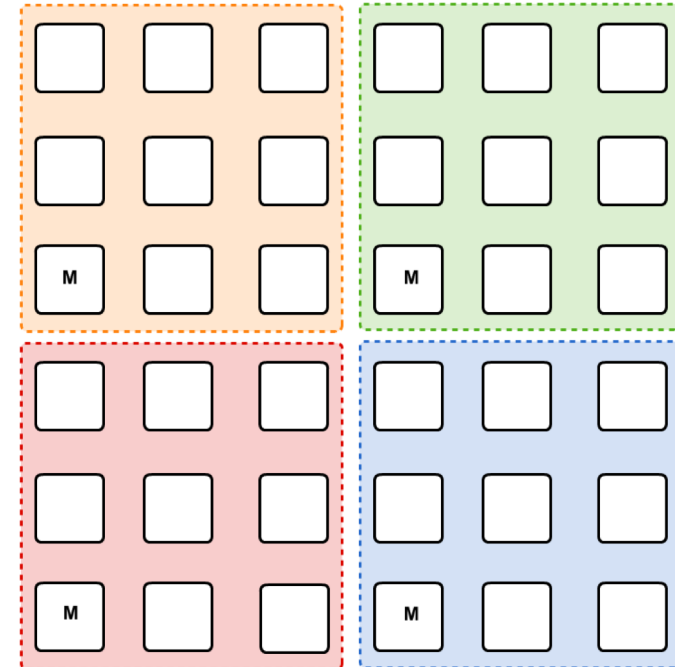
Energy Consumption (J)



RTM ON THE “GRACEFUL” PLATFORM

Approach

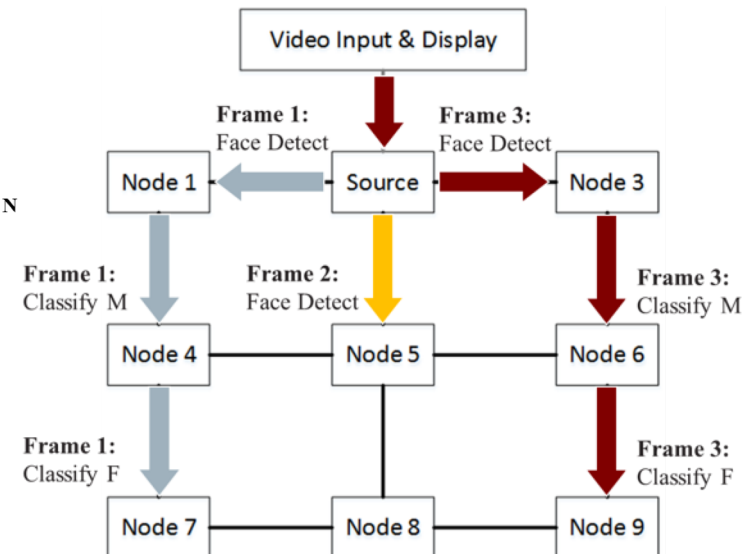
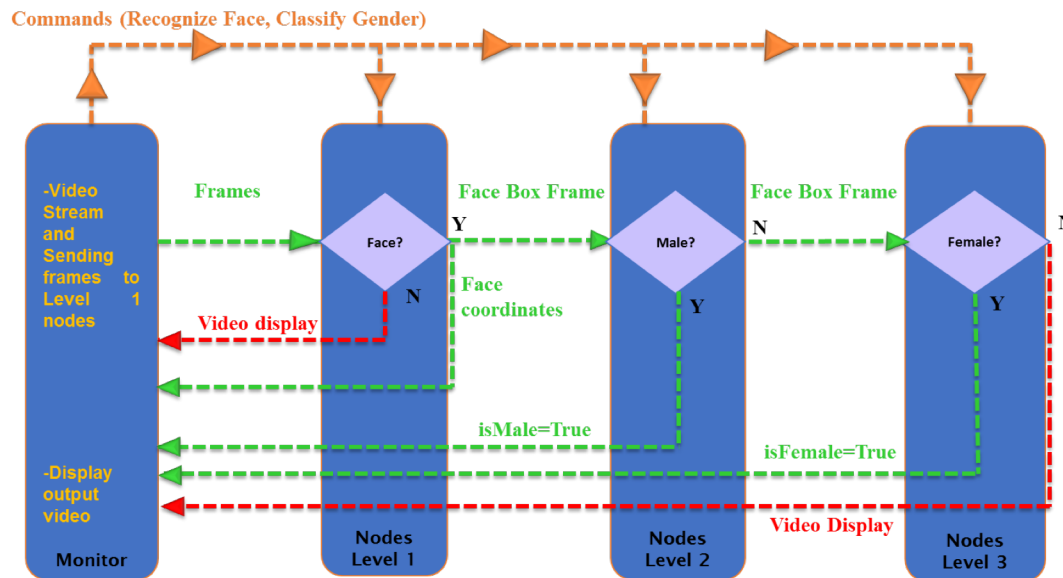
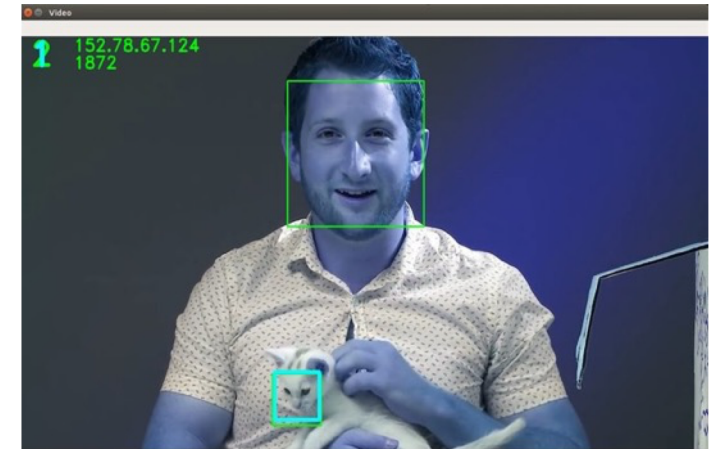
- Opportunity for Hierarchical RTM
 - Local RTM (DVFS, local mapping etc) on each node
 - Higher level ‘strategic’ RTM (mapping within cluster, migration, load balancing etc) in clusters
 - Potential for a third level negotiating between clusters

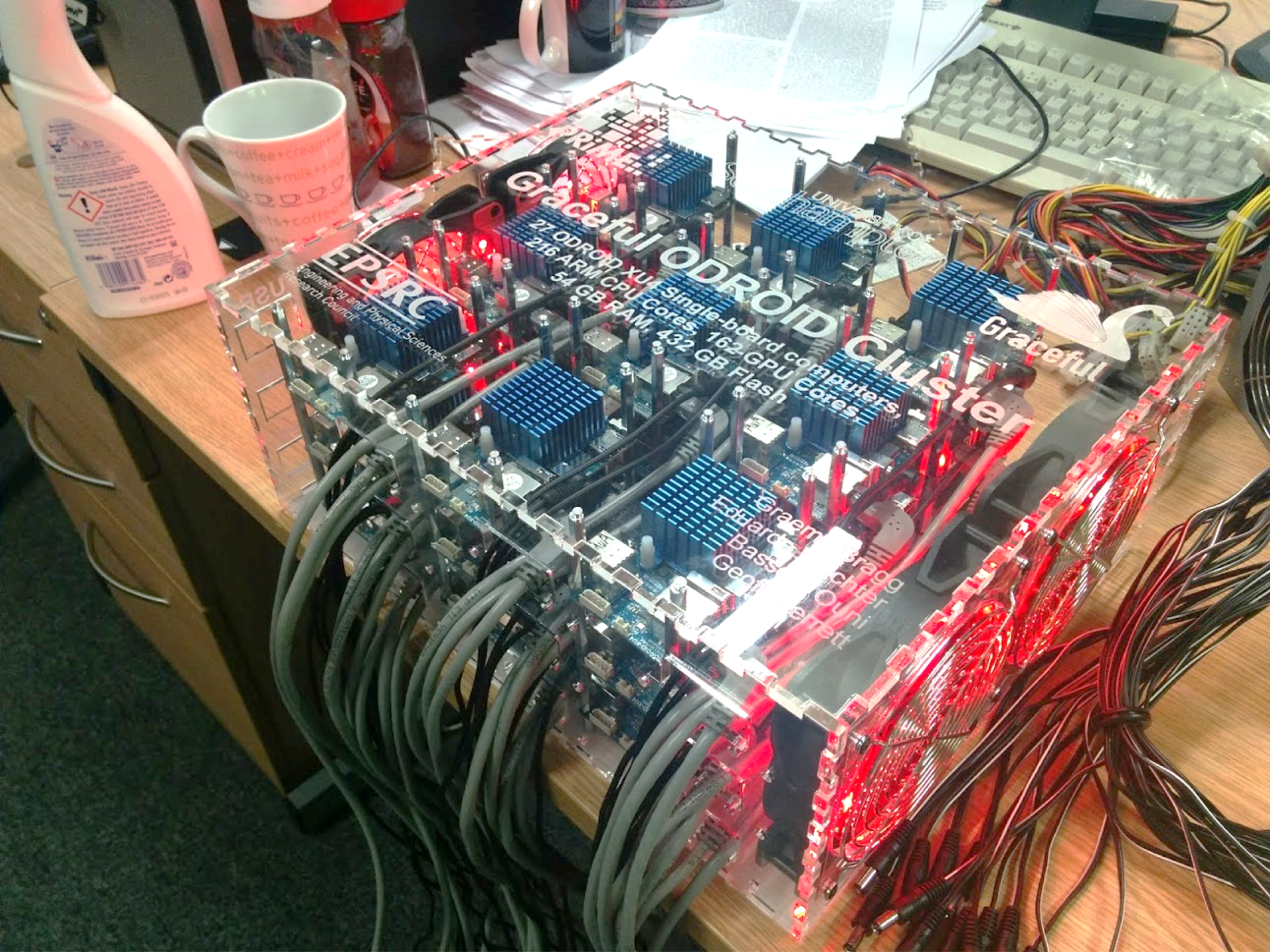


RTM ON THE “GRACEFUL” PLATFORM

Example Application

- Face/Object Detection/Classification
- Uses OpenCV classifiers
 - Detect faces/animals/objects
 - Classify gender
 - Estimate age





Graceful ODR0ID Cluster

EPSRC
Engineering and Physical Sciences
Research Council

22 ODR0ID xul single-board computers,
216 ARM CPU Cores, 162 GPU Cores,
54 GB RAM, 432 GB Flash

Graem Edwards
Bassam Ghaleb
George Bennett

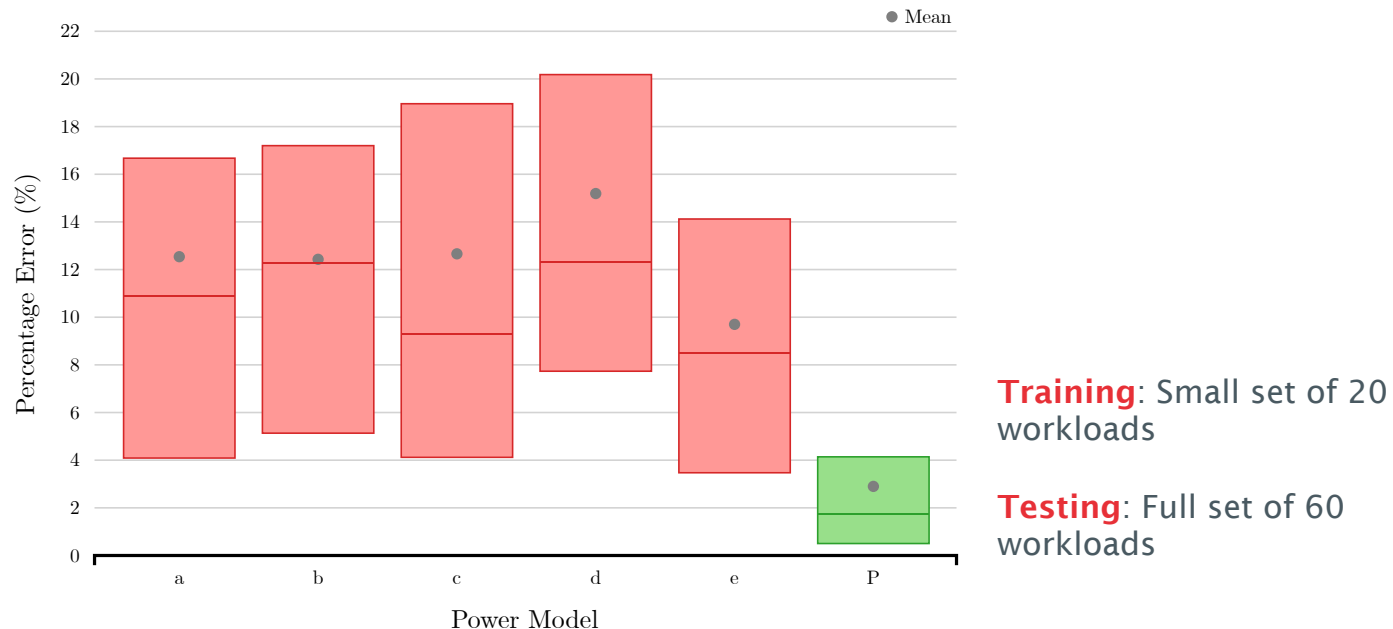
Graceful ODR0ID Cluster

OPEN SOURCE TOOLS

POWMON: STABLE POWER MODELLING

www.powmon.ecs.soton.ac.uk

Our stable approach achieves a low average error and narrow error distribution compared to existing techniques.



- [a] M. Pricopi, T. S. Muthukaruppan, V. Venkataramani, T. Mitra, and S. Vishin, "Power-performance modeling on asymmetric multi-cores," CASES '13.
- [b] M. Walker et al., "Run-time power estimation for mobile and embedded asymmetric multi-core cpus," HIPEAC Workshop Energy Efficiency with Hetero. Comp. 2015
- [c] S. K. Rethinagiri et al., "System-level power estimation tool for embedded processor based platforms," RAPIDO '14. New York, 2014.
- [d], [e] R. Rodrigues et al, "A study on the use of performance counters to estimate power in microprocessors," IEEE TCAS II, vol. 60, no. 12, pp. 882-886, Dec 2013.

M. J. Walker et al., "Accurate and Stable Run-Time Power Modeling for Mobile and Embedded CPUs," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 1, pp. 106-119, Jan. 2017.

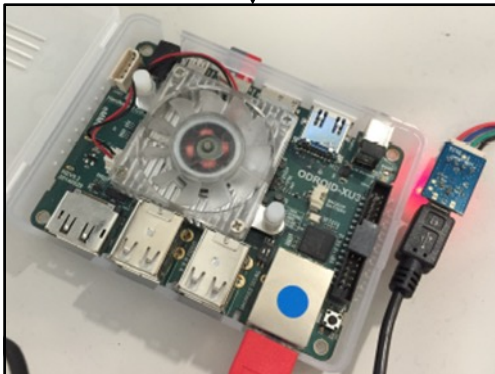
POWMON: METHODOLOGY

www.powmon.ecs.soton.ac.uk

1. Run workloads

@ different DVFS levels

39 workloads used: MiBench, LMBench, Roy Longbottom, ParMiBench and ALPBench



ODROID-XU3
Exynos-5422
4x Cortex-A7
4x Cortex-A5

5. Validate

- K-fold cross validation
- $R^2 : > 0.99$
- Error: 2.8 – 3.7%



4. Build Model

- OLS multiple regression
- Considers collinearity and heteroscedasticity
- “sensible” equation

2. Record

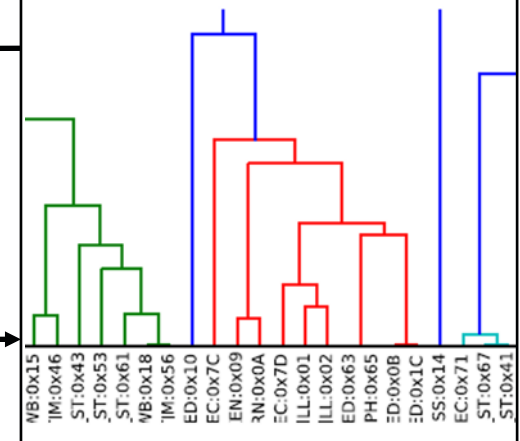
- PMCs
- Power, Voltage, Temperature, etc.

6. Uses

- OS Run-time management
- Reference for research
- gem5 add-on

3. Choose PMCs

Hierarchical cluster analysis,
Correlation matrix analysis,
Exhaustive search, etc.

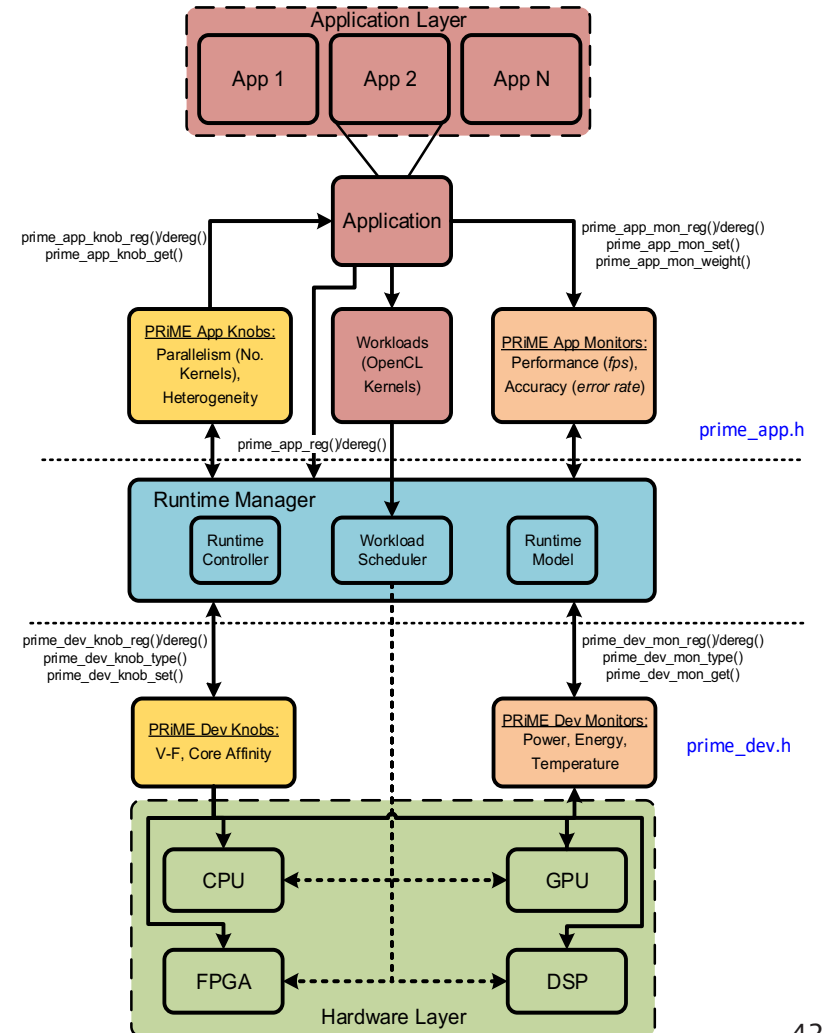


PRIME RTM FRAMEWORK

<https://github.com/PRiME-project/PRiME-Framework>

Plat.	Const.	Space	Type	For	No.
Odroid-XU3	knob	disc	GOVERNOR	A7 cluster	1
		disc	GOVERNOR	A15 cluster	1
		disc	FREQ	A7 cluster	1
		disc	FREQ	A15 cluster	1
		disc	FREQ_EN	GPU DVFS	1
		disc	FREQ	GPU	1
		disc	PMC_CTRL	A7 cores	16
		disc	PMC_CTRL	A15 cores	24
	mon	cont	POW	Clusters, RAM, GPU, SoC	5
		cont	TEMP	A15 cores	4
		cont	TEMP	GPU	1
		disc	CYCLE	A7 cores	4
		disc	CYCLE	A15 cores	4
		disc	PMC	A7 cores	16
Cyclone V	knob	cont	VOLT	A9 cluster, peripherals	4
		cont	VOLT	FPGA, peripherals	3
	mon	cont	POW	A9 cluster, peripherals	5
		cont	POW	FPGA, peripherals	4
		cont	POW	SoC	1

Application	Name	Const.	Space	Allowed/target values
Jacobi	Iterations	knob	disc	$N \in [1, \infty)$
	Data type	knob	disc	{float, double}
	Device type	knob	disc	{CPU, GPU/FPGA}
	Throughput	mon	cont	$R \in [10, \infty)$
	Error	mon	cont	$R \in (-\infty, 1e^{-12}]$
Video decoder	Throughput	mon	cont	$R \in [25, \infty)$
Whetstone	Threads	knob	disc	$N \in [1, \infty)$
	Throughput	mon	cont	$R \in [2.5, \infty)$



CONCLUSIONS

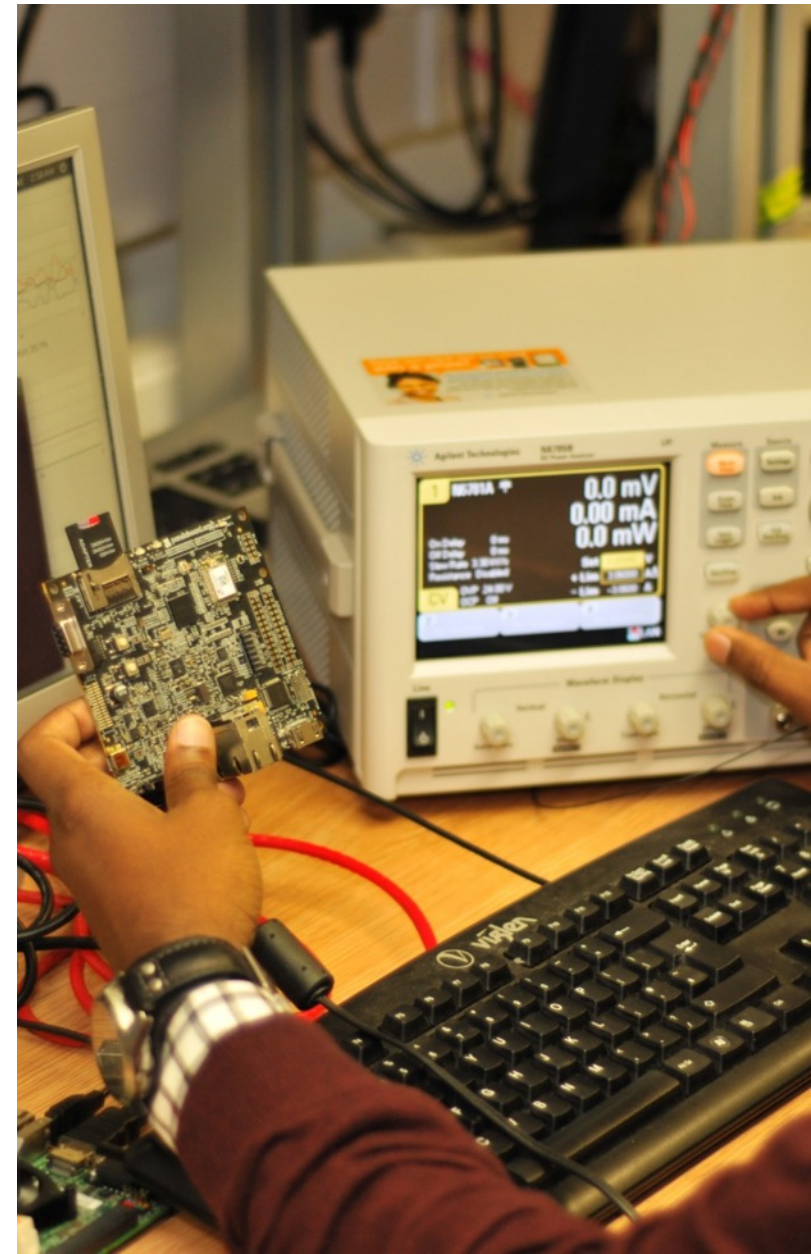
Runtime Power Management

- Single > multiple > concurrent applications
- Online *vs* offline+online approaches
- >> Number of cores
- COTS > Novel multi-/many-core platforms
- Homogeneous *vs* Heterogeneous platforms

Tools and Support www.prime-project.org

- PowMon power estimation
www.powmon.ecs.soton.ac.uk
www.gemstone.ecs.soton.ac.uk
- PRiME RTM Framework
github.com/PRiME-project/PRiME-Framework
- PRiMEStereoMatch application
github.com/PRiME-project/PRiMEStereoMatch

<http://www.prime-project.org/>





Any Questions?

UNIVERSITY OF
Southampton

Dr Geoff V Merrett

Associate Professor | Head of Centre

Centre for IoT and Pervasive Systems

Tel: +44 (0)23 8059 2775

Email: gvm@ecs.soton.ac.uk | www.geoffmerrett.co.uk

Highfield Campus, Southampton, SO17 1BJ UK