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

Dr Geoff Merrett

Adaptive Many-Core Architectures and Systems workshop
13-15 June 2018 | York, UK

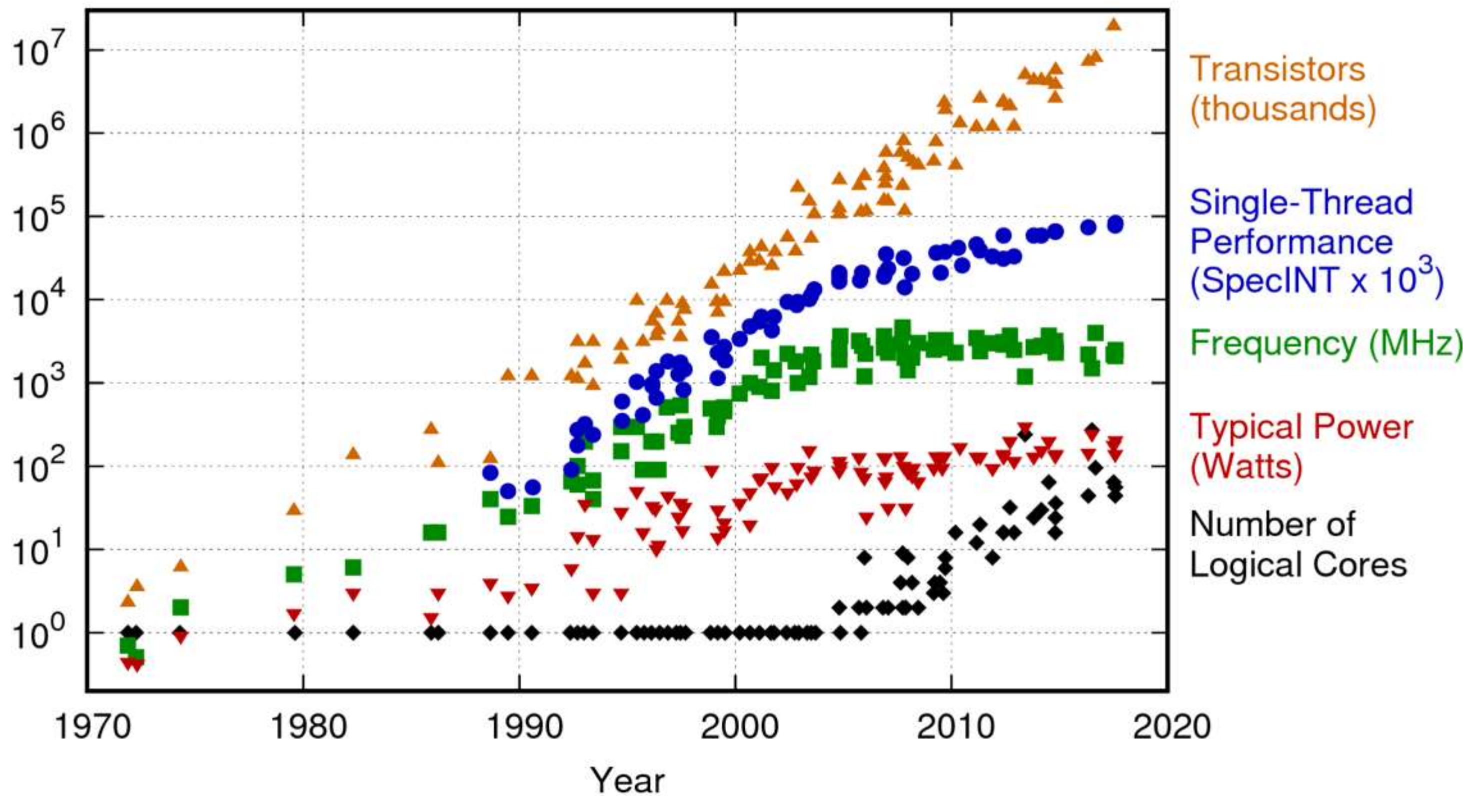
THE PRiME PROJECT

"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."

www.prime-project.org

WE ARE MANY-CORE

42 Years of Microprocessor Trend Data

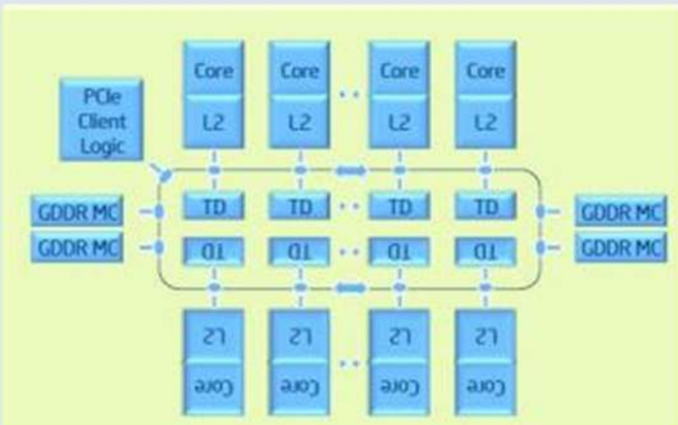


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

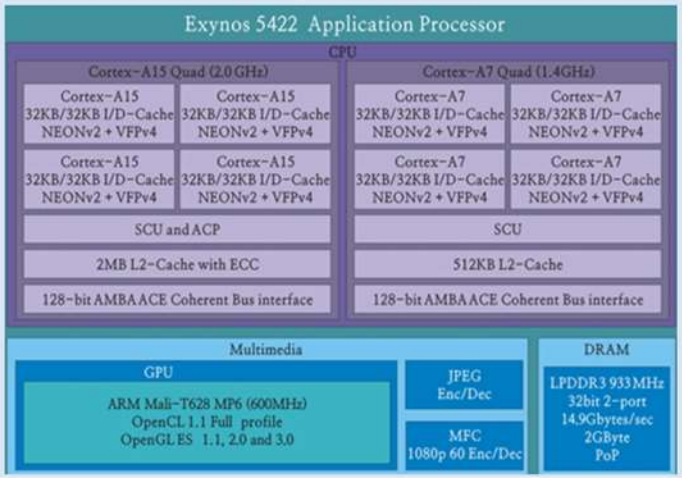
Image from www.karlrupp.net/

MANY-CORE PLATFORMS

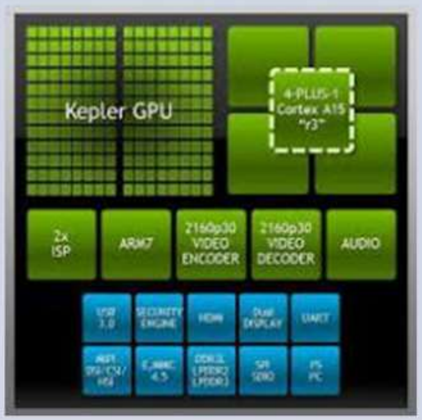
Intel Xeon Phi - **Homogeneous 61 Cores**



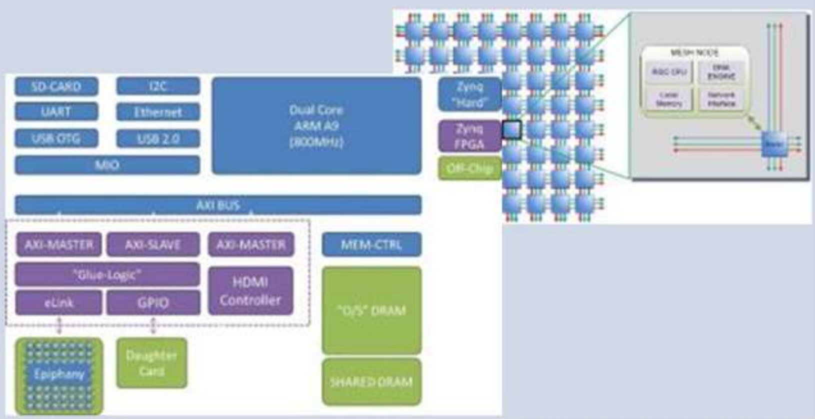
ODROID XU3 – **8 core big.LITTLE CPU + 6 cores GPU**



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

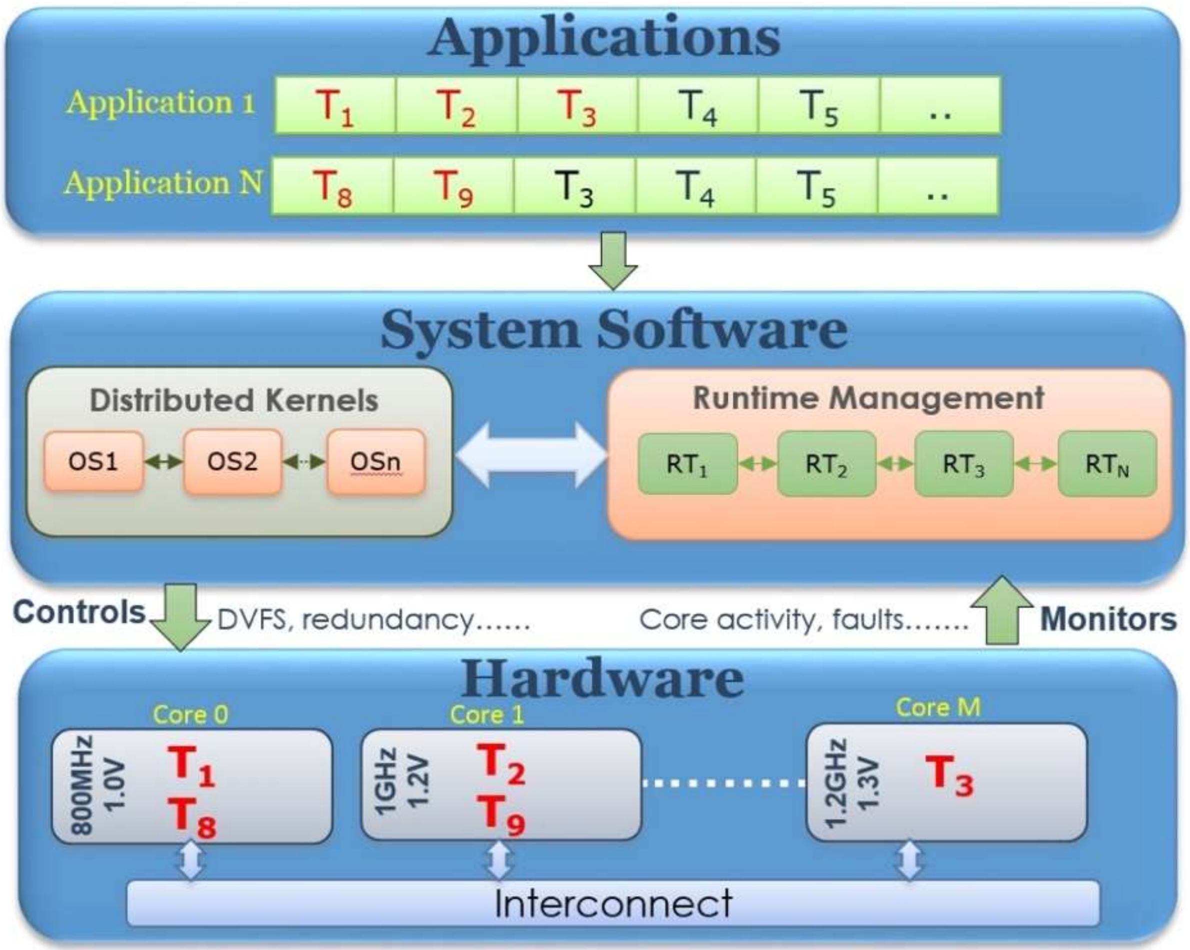


Parallella - **Dual core CPU + FPGA + 16 cores NoC**

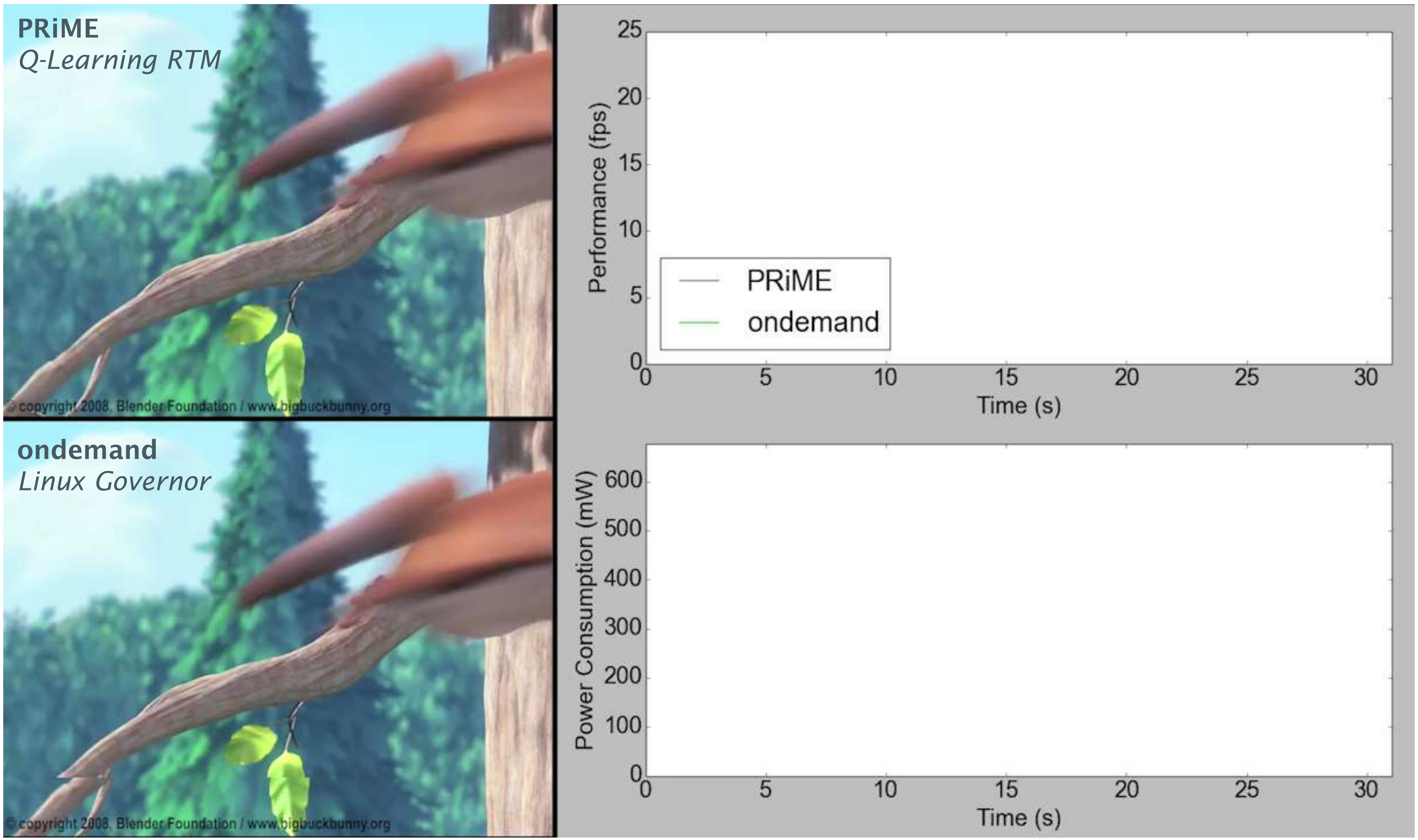


THE PRiME PROJECT

www.prime-project.org



RUNTIME POWER MANAGEMENT

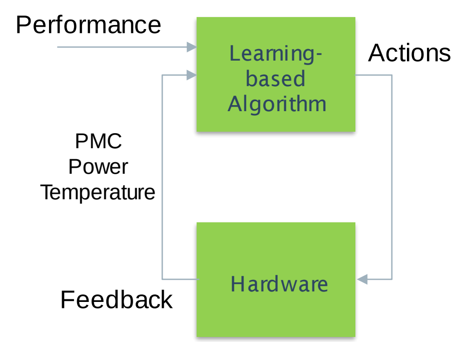


Maeda-Nunez, Luis Alfonso, Anup K. Das, Rishad A. Shafik, Geoff V. Merrett, and Bashir Al-Hashimi. "*PoGo: an application-specific adaptive energy minimisation approach for embedded systems*." **HiPEAC Workshop on Energy Efficiency with Heterogenous Computing**, 2015

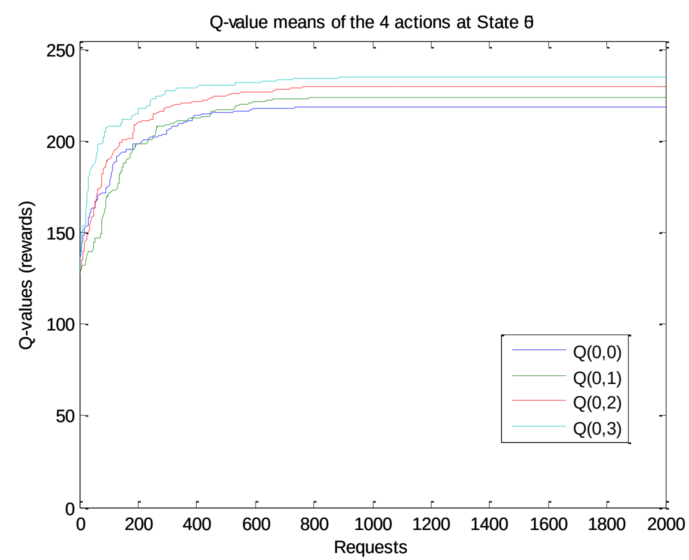
LEARNING OPTIMAL DVFS CHOICES

Reinforcement Learning

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

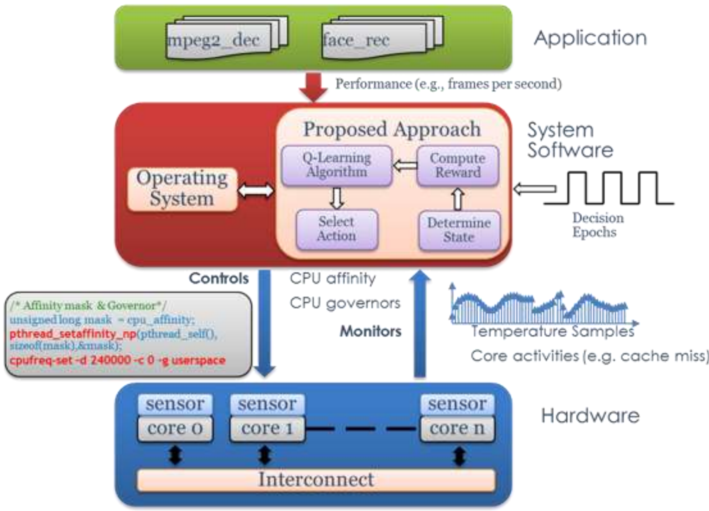


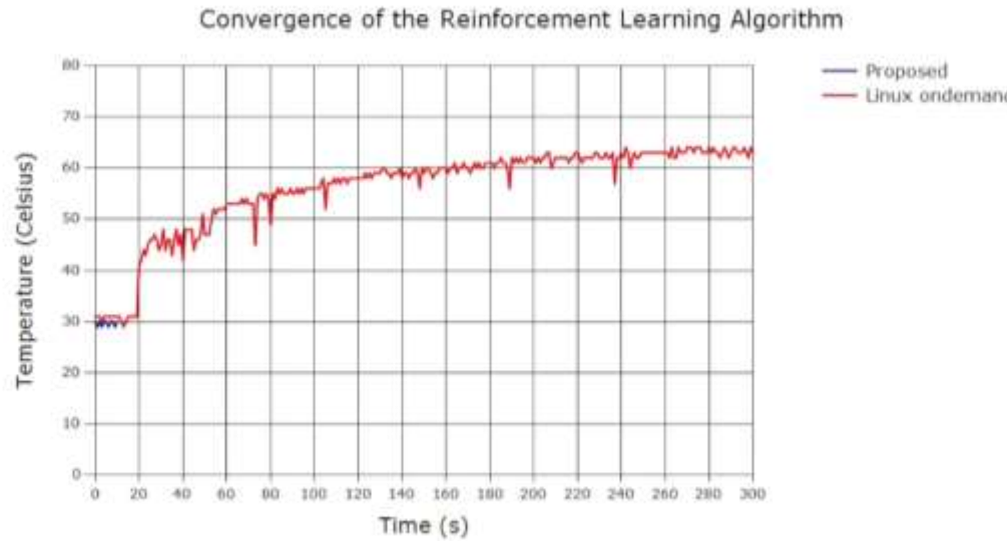
	ACTIONS (Power Modes)			
STATES (Tasks)	P0	P1	P2	P3
WD0	128	128	128	128
WD1	128	128	128	128
WD2	128	128	128	128
WD3	128	128	128	128
WD4	128	128	128	128
WD5	128	128	128	128



Shafik, Rishad Ahmed, Das, Anup K., Maeda-Nunez, Luis Alfonso, Yang, Sheng, Merrett, Geoff V. and Al-Hashimi, Bashir (2015) *Learning transfer-based adaptive energy minimization in embedded systems*. **IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems**, 1-14.

MANAGING THERMAL (LIFETIME) RELIABILITY





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)

Das, Anup, Al-Hashimi, Bashir and Merrett, Geoff (2015) *Adaptive and hierarchical run-time manager for energy-aware thermal management of embedded systems.* **ACM Transactions on Embedded Computing Systems**, 1-25.

PRiME Graceful

UNIVERSITY OF Southampton

OVERVIEW

Applications

- From single > sequential > concurrent execution

Offline Characterisation

- Can we improve RTM through offline characterisation?

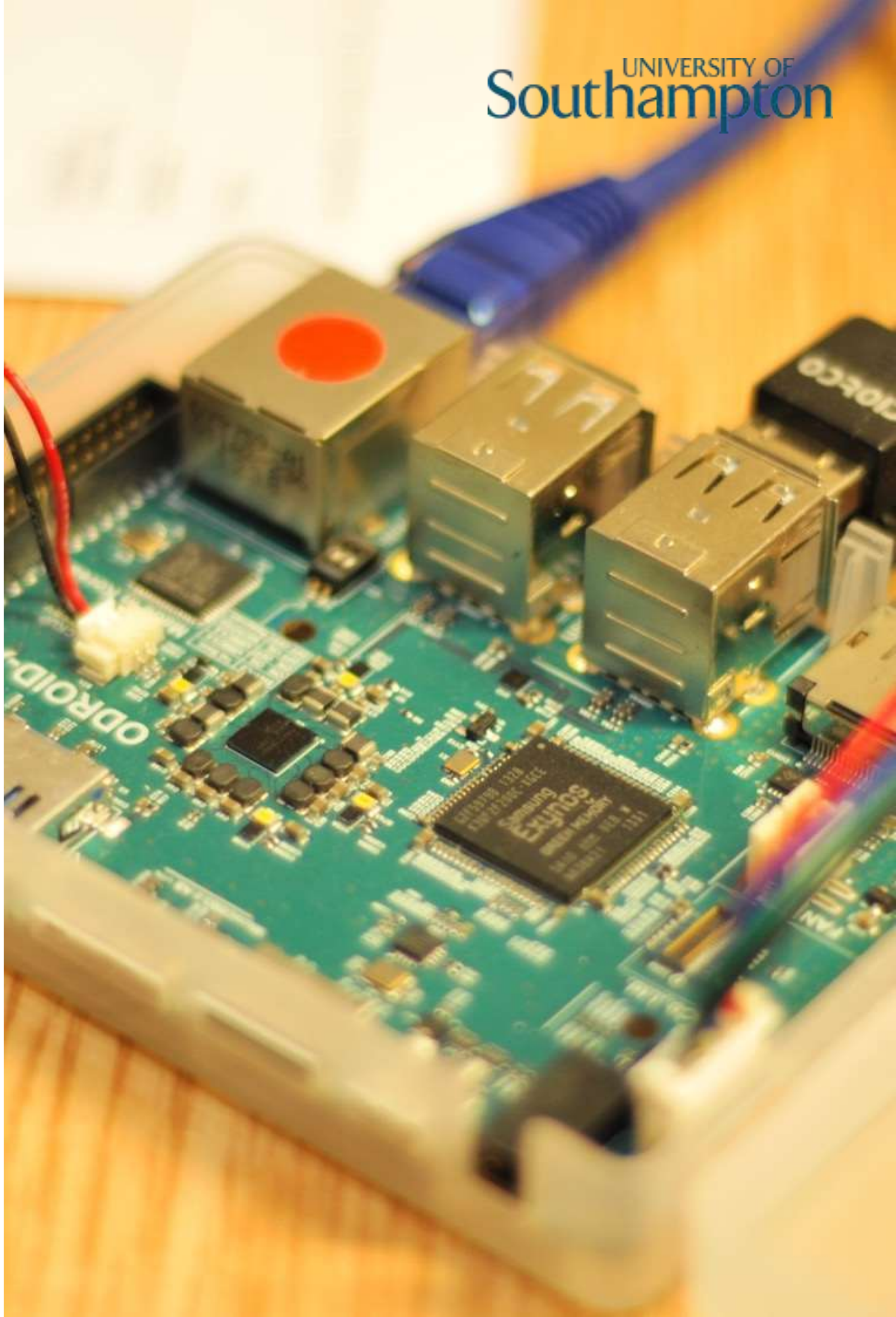
Towards Many-Core

- How do RTM approaches scale with number of cores?

Novel Platforms

- Can our RTM approaches be applied to novel platforms?

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



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)

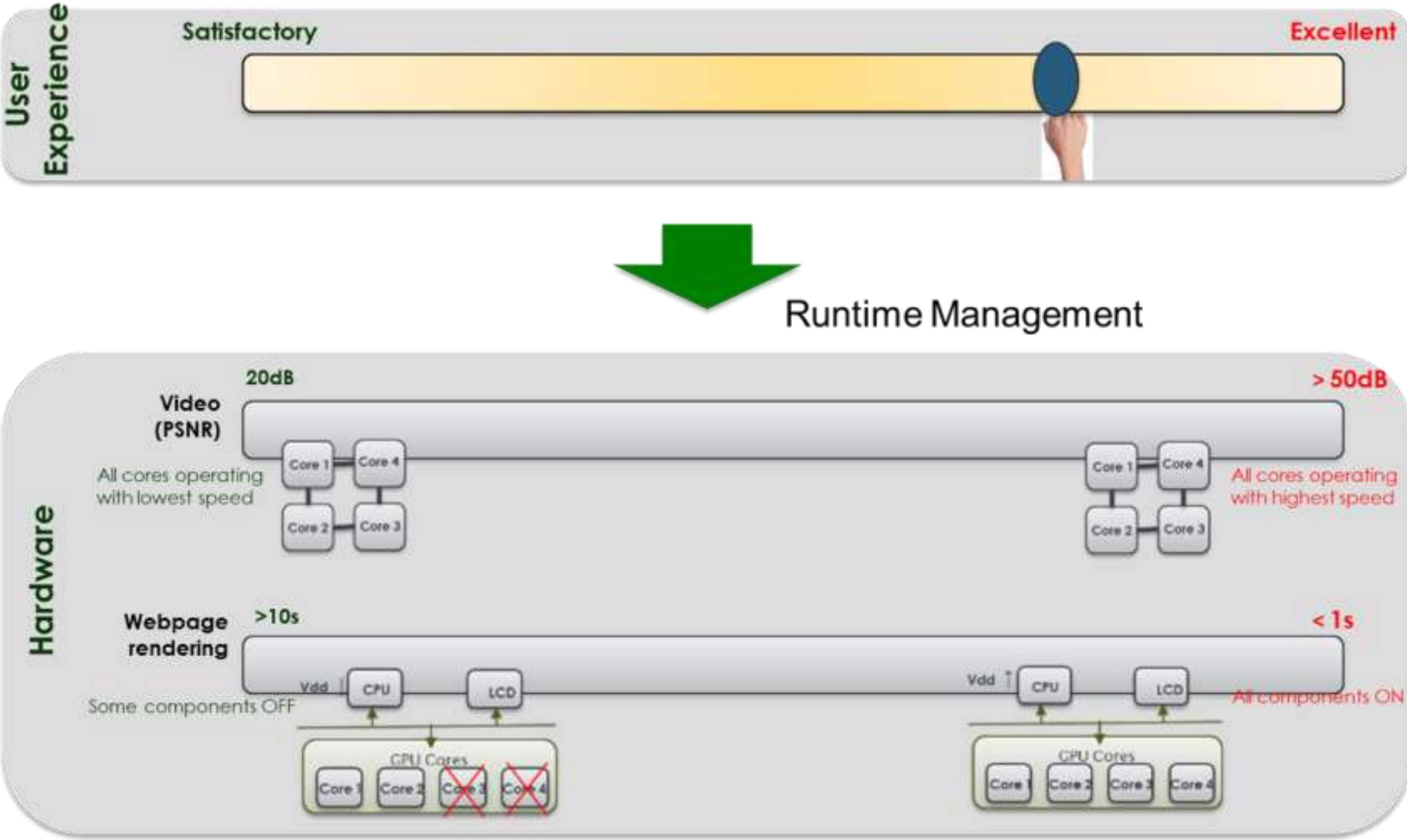
- Therefore, optimise for **quality of user experience** (QoE)
  - "*good-enough*" performance
  - Minimum energy usage

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



QUALITY OF EXPERIENCE

Workload Classification

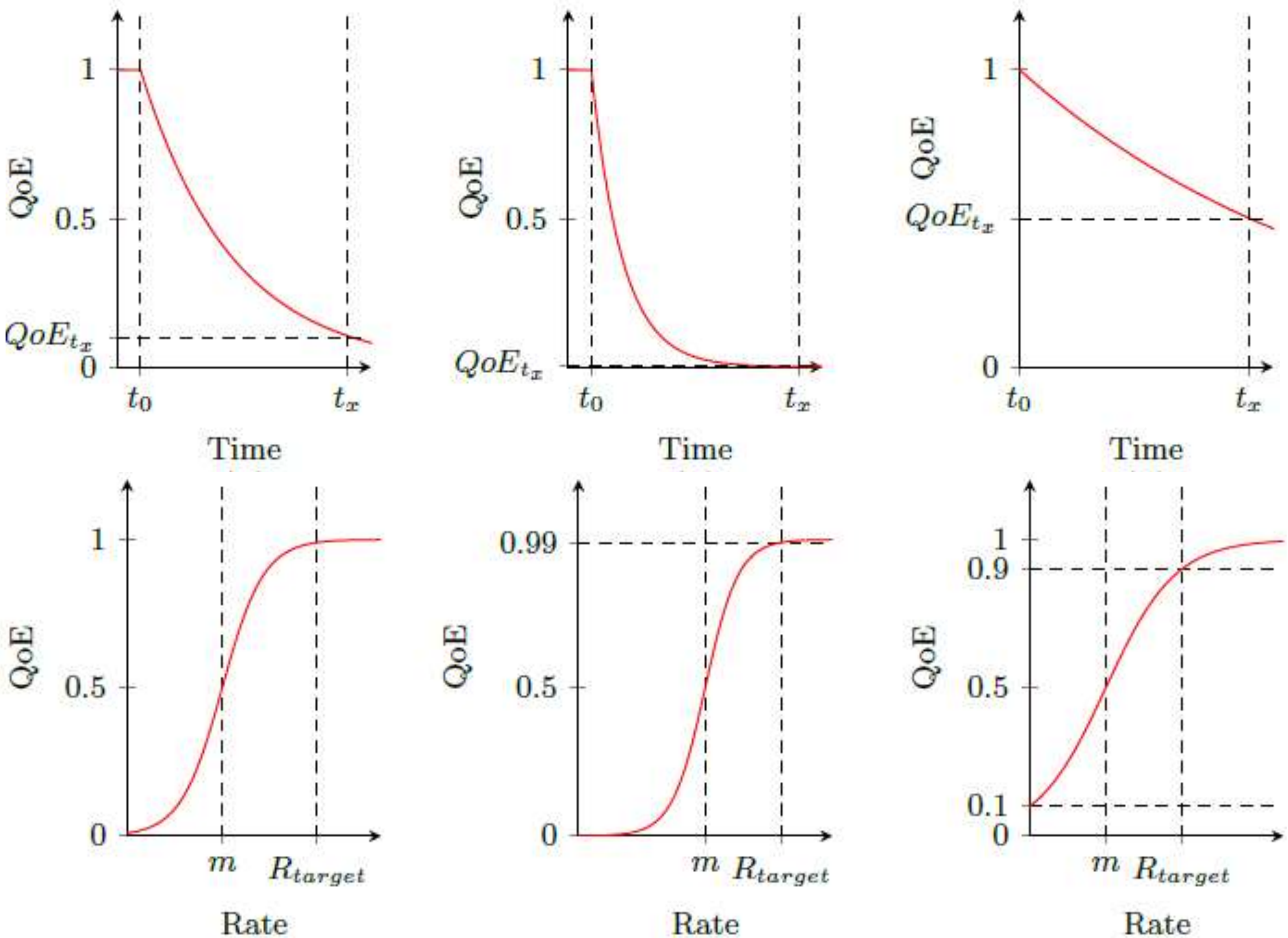
Applications	Type of QoE
Audio	Throughput
Video	Throughput
Application Loading	Latency
Web Page Loading	Latency
Downloading a File	Latency
3D Gaming	Throughput
Word Processing	Latency

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

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.

QoE CHARACTERISTICS



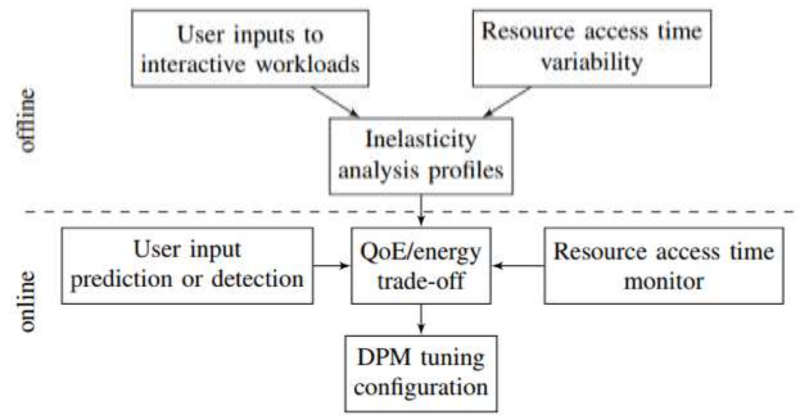
Latency Sensitive

Inverse exponential

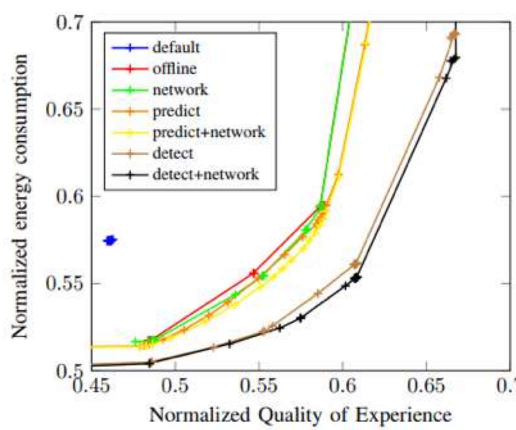
Throughput Sensitive

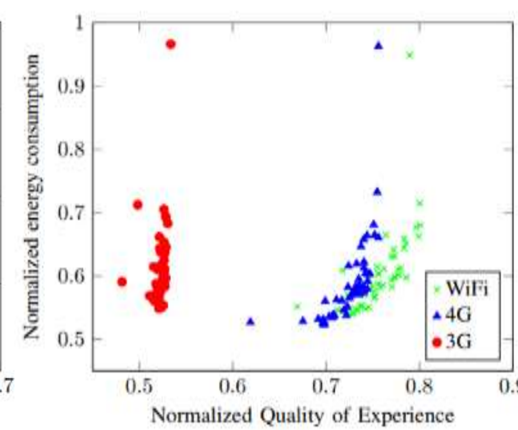
Sigmoid function

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



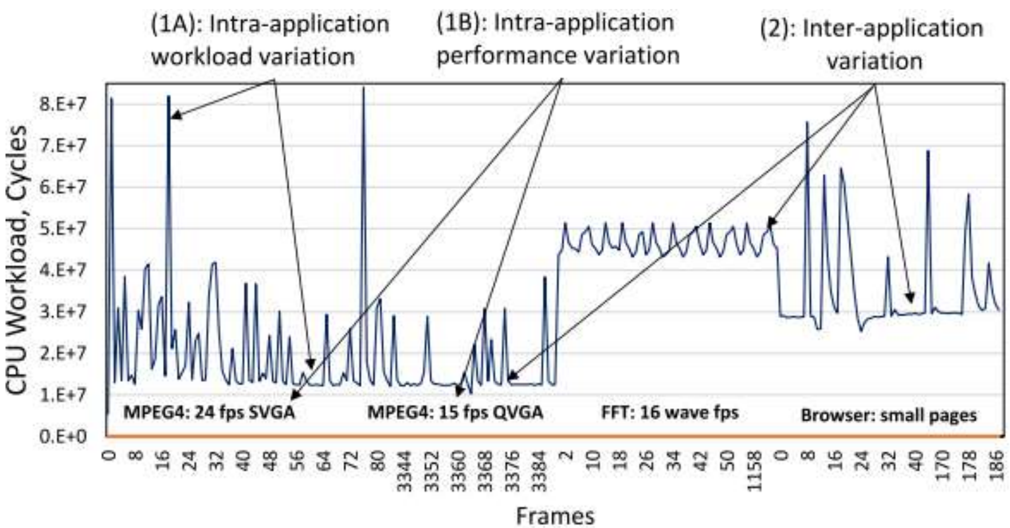


Exynos-5422 A15/A7, Android 6.0
Google Chrome browser workloads
Touch input emulation
Network throttling (UL, DL, RTT latency)

Bantock, James, Robert Benjamin, Tenentes, Vasileios, Al-Hashimi, Bashir and Merrett, Geoffrey (2017) *Online tuning of Dynamic Power Management for efficient execution of interactive workloads* **International Symposium on Low Power Electronics and Design**. IEEE. 6 pp.

EXECUTING MULTIPLE APPLICATIONS

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

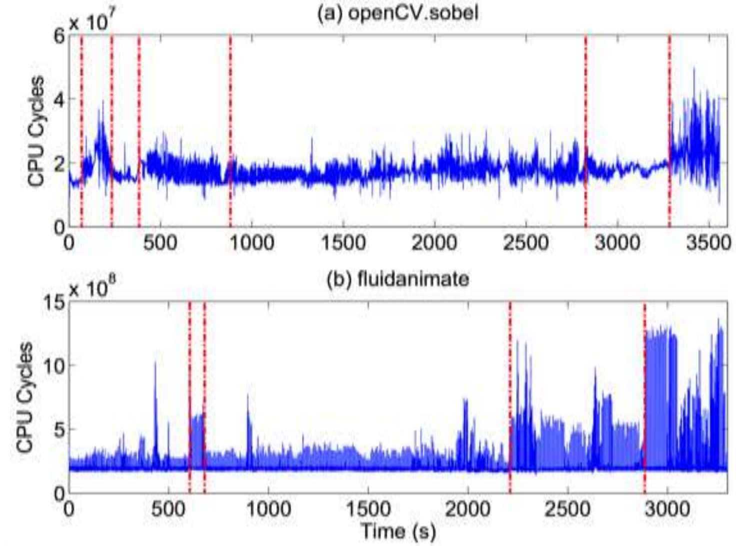


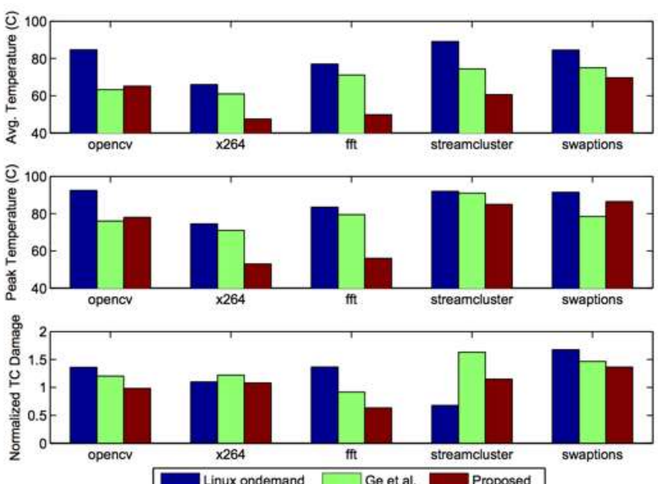
  - Overlapping applications (*concurrent execution*)

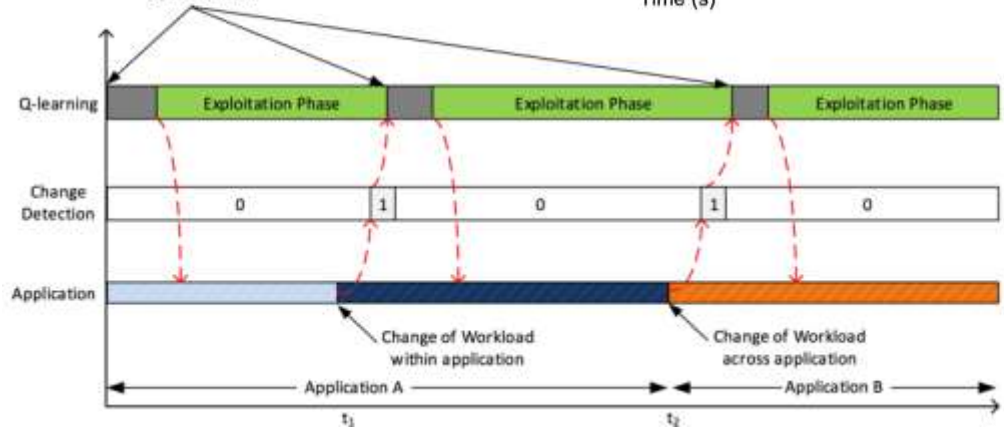
Shafik, Rishad, Das, Anup, Maeda-Nunez, Luis, Yang, Sheng, Merrett, Geoff and Al-Hashimi, Bashir (2015) *Learning transfer-based adaptive energy minimization in embedded systems*. **IEEE TCAD**.

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)

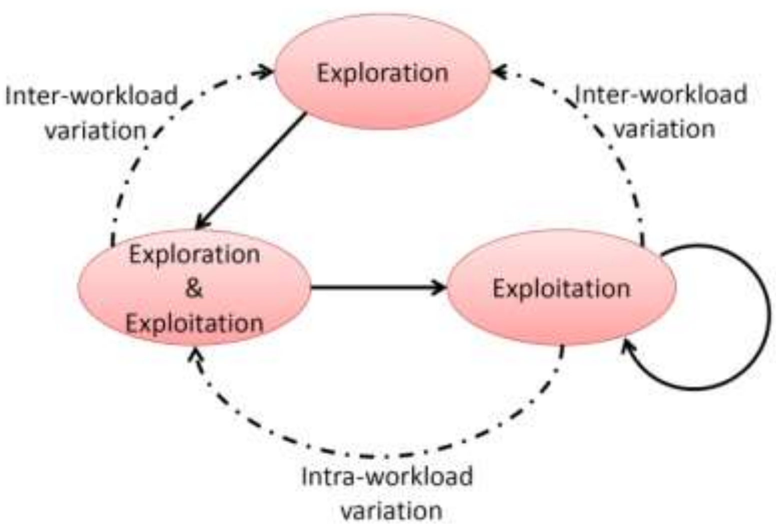


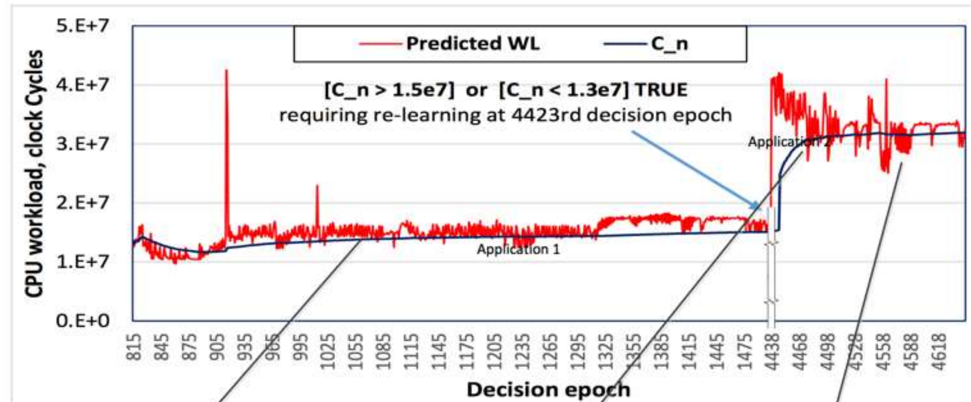




TRANSFERING LEARNING

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





(a)

Q-table during exploration in the initial learning step with base workload, $C_b = 1.5 \times 10^7$ cycles

Exploitation Phase

WL		(V1, F1)	(V2, F2)	(V3, F3)	(V4, F4)
WL: 1.5×10^7 to 1.65×10^7	$L_t > 15\%$	0	-58	-13	102
	$15\% < L_t <= 5\%$	-703	-102	15.2	-35.6
	$5\% < L_t <= -5\%$	-30.2	58.1	-35	-112.1
	$-5\% < L_t <= 15\%$	-341	126.2	78.1	0
	$Lt < -15\%$	-32	-117	-137	0
WL: 1.35×10^7 to 1.5×10^7	$L_t > 15\%$	0	0	0	-16
	$15\% < L_t <= 5\%$	0	0	0	0
	$5\% < L_t <= -5\%$	0	-14	0	0
	$-5\% < L_t <= 15\%$	0	0	0	0
	$Lt < -15\%$	-12	-19	0	0

(b)

Q-table after Q-value & action scaling in the learning transfer step with base workload, $C_b = 3.0 \times 10^7$ cycles

Q-value scaling

WL		(V1, F1)	(V2, F2)	(V3, F3)	(V4, F4)
WL: 3.0×10^7 to 3.3×10^7	$L_t > 15\%$	0	-10.4	-2.3	18.3
	$15\% < L_t <= 5\%$	-5.4	-0.8	0.1	-0.3
	$5\% < L_t <= -5\%$	-0.4	0.7	-0.4	-1.4
	$-5\% < L_t <= 15\%$	-2.6	1.0	0.6	0
	$Lt < -15\%$	-6.3	-27.1	-38.5	0
WL: 2.7×10^7 to 3.0×10^7	$L_t > 15\%$	0	0	0	-2.9
	$15\% < L_t <= 5\%$	0	0	0	0
	$5\% < L_t <= -5\%$	0	-2.4	0	0
	$-5\% < L_t <= 15\%$	0	0	0	0
	$Lt < -15\%$	-2.1	-2.7	0	0

(c)

Q-table after minimal re-exploration in learning transfer step with base workload, $C_b = 3.0 \times 10^7$ cycles

After Learning Transfer and re-exploration

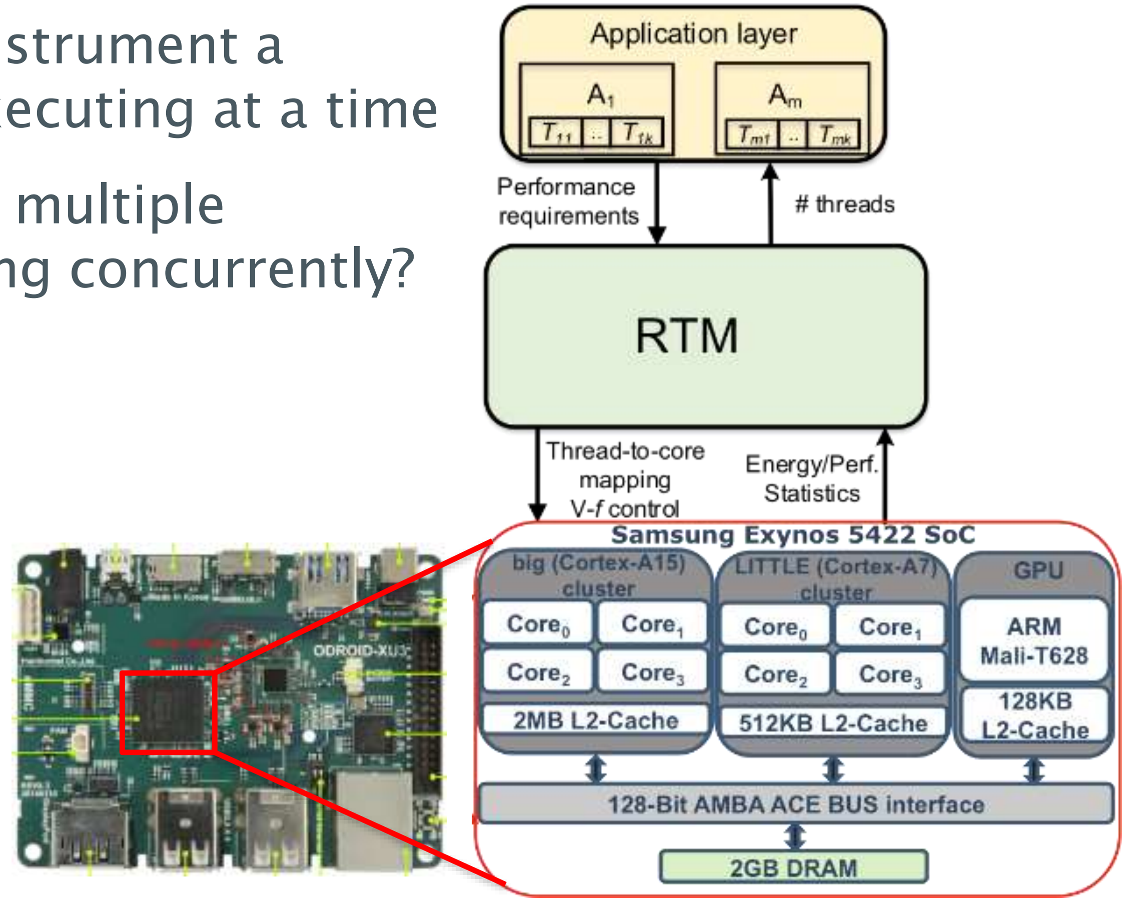
WL		(V1, F1)	(V2, F2)	(V3, F3)	(V4, F4)
WL: 3.0×10^7 to 3.3×10^7	$L_t > 15\%$	0	-12.5	14.8	22.0
	$15\% < L_t <= 5\%$	-1.7	-0.9	0.3	-13.2
	$5\% < L_t <= -5\%$	10.5	-5.2	-6.1	-19.2
	$-5\% < L_t <= 15\%$	7.4	11.5	-15.5	0
	$Lt < -15\%$	-29.5	-6.9	-17.0	0
WL: 2.7×10^7 to 3.0×10^7	$L_t > 15\%$	0	0	0	-2.9
	$15\% < L_t <= 5\%$	0	0	0	0
	$5\% < L_t <= -5\%$	0	-2.4	0	0
	$-5\% < L_t <= 15\%$	0	0	0	0
	$Lt < -15\%$	-2.9	-6.9	-12	0

(d)

Shafik, Rishad A., et al. "*Learning transfer-based adaptive energy minimization in embedded systems*", **IEEE TCAD**, pp 1-15, Oct 2015

RTM FOR CONCURRENT EXECUTION

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



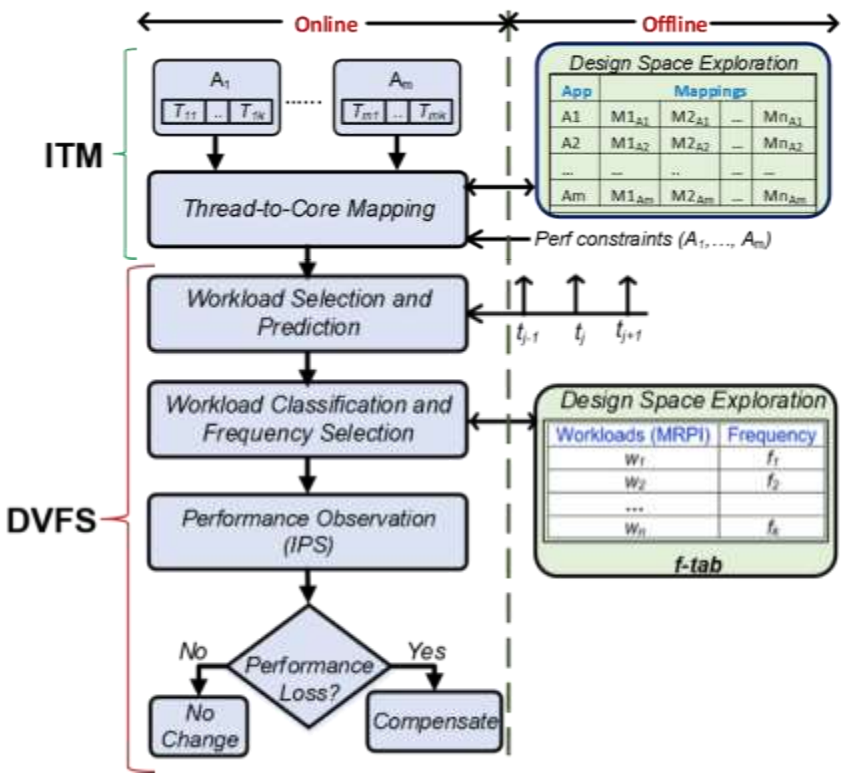
Reddy, Basireddy Karunakar, Singh, Amit, Biswas, Dwaipayan, Merrett, Geoff and Al-Hashimi, Bashir (2017) *Inter-cluster thread-to-core mapping and DVFS on heterogeneous multi-cores* **IEEE Transactions on Multiscale Computing Systems**, pp. 1-14.

Online vs Offline

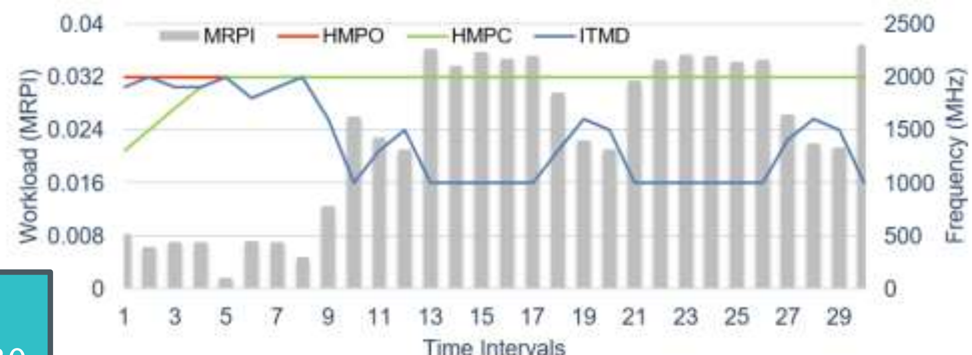
Can we improve RTM through offline characterisation?

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

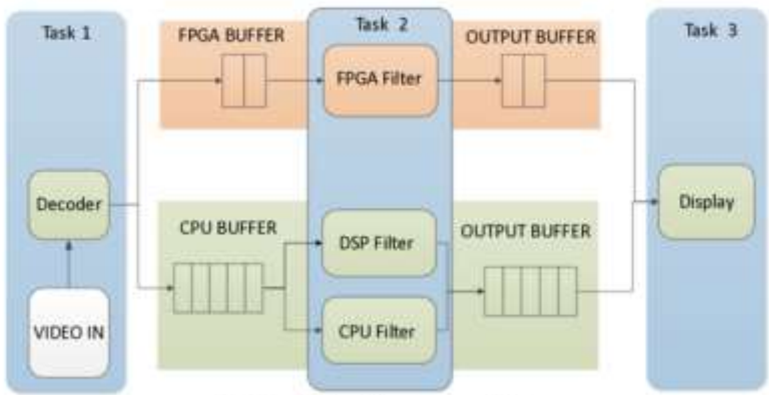
Reddy, Basireddy Karunakar, Singh, Amit, Biswas, Dwaipayan, Merrett, Geoff and Al-Hashimi, Bashir (2017) *Inter-cluster thread-to-core mapping and DVFS on heterogeneous multi-cores* **IEEE Transactions on Multiscale Computing Systems**, pp. 1-14.

MODEL-BASED RTM: HETEROGENEITY

Heterogeneous Platforms

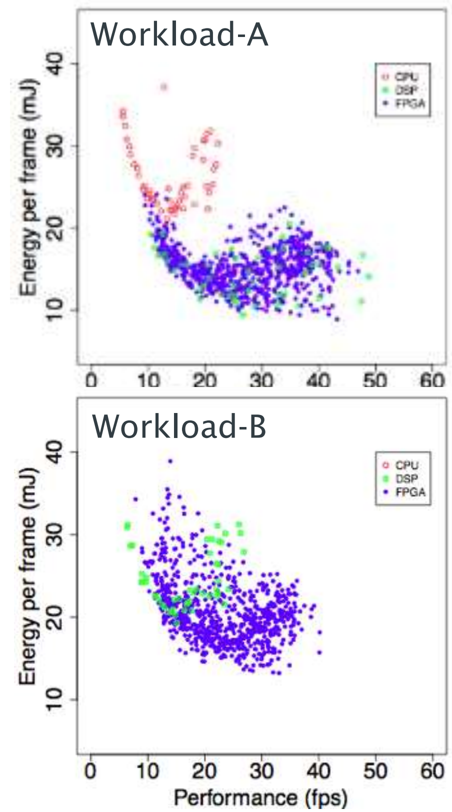
Run-time changes in:

- Performance requirements
- Application workload changes



(a) Convolution filter implementation

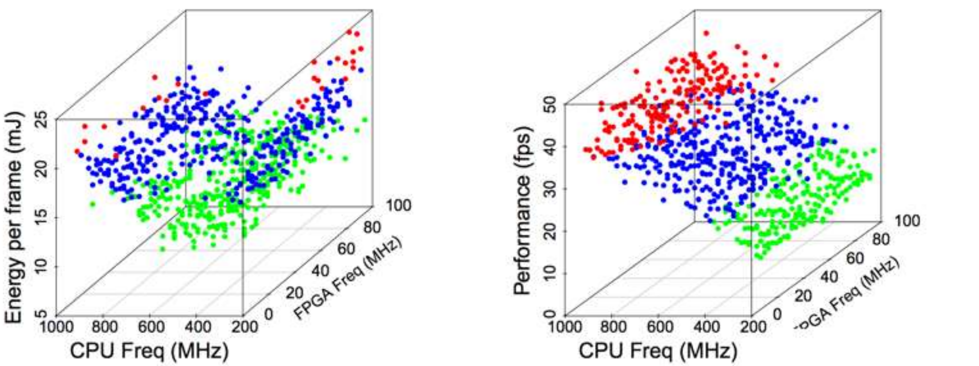
(b) Original (c) Edge detected (d) Blurred



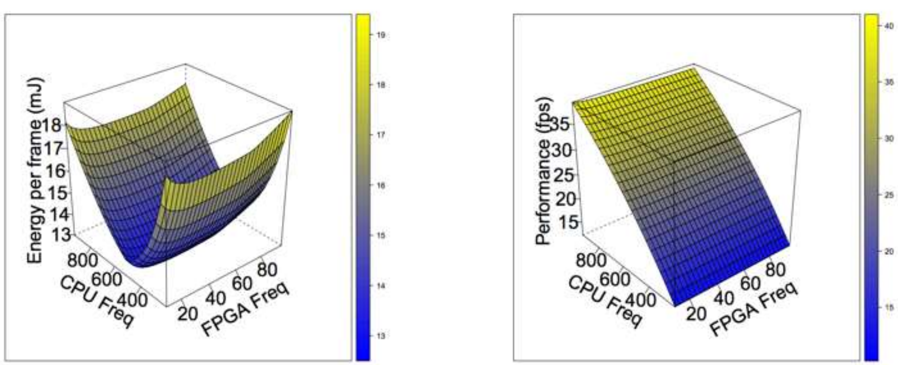
Yang, Sheng, Shafik, Rishad Ahmed, Merrett, Geoff V., Stott, Edward, Levine, Joshua, Davis, James and Al-Hashimi, Bashir (2015) *Adaptive energy minimization of embedded heterogeneous system using regression-based learning.* **PATMOS 2015**, Salvador, BR, 01 - 04 Sep 2015. 8pp.

MODEL-BASED RTM: HETEROGENEITY

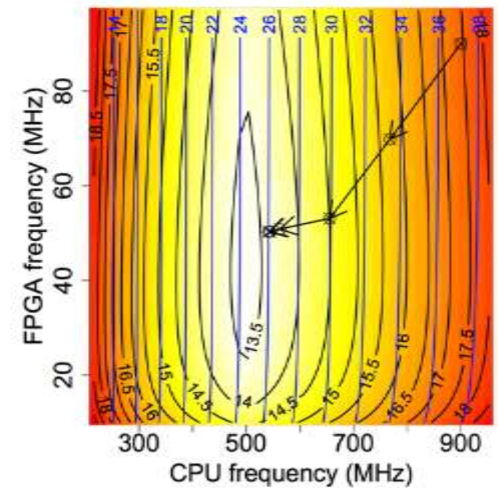
Heterogeneous Platforms



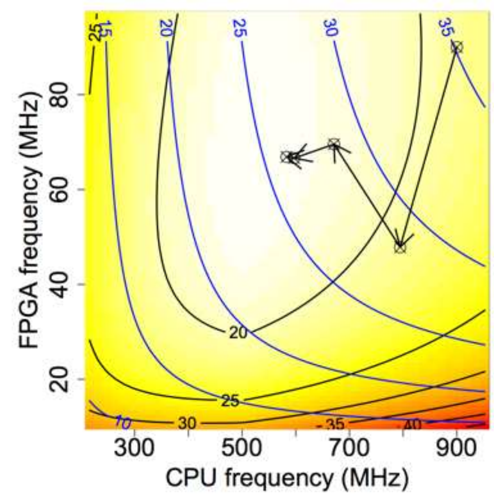
(a) FPGA measured energy (b) FPGA measured performance



(c) FPGA modeled energy (d) FPGA modeled performance



Edge Detection

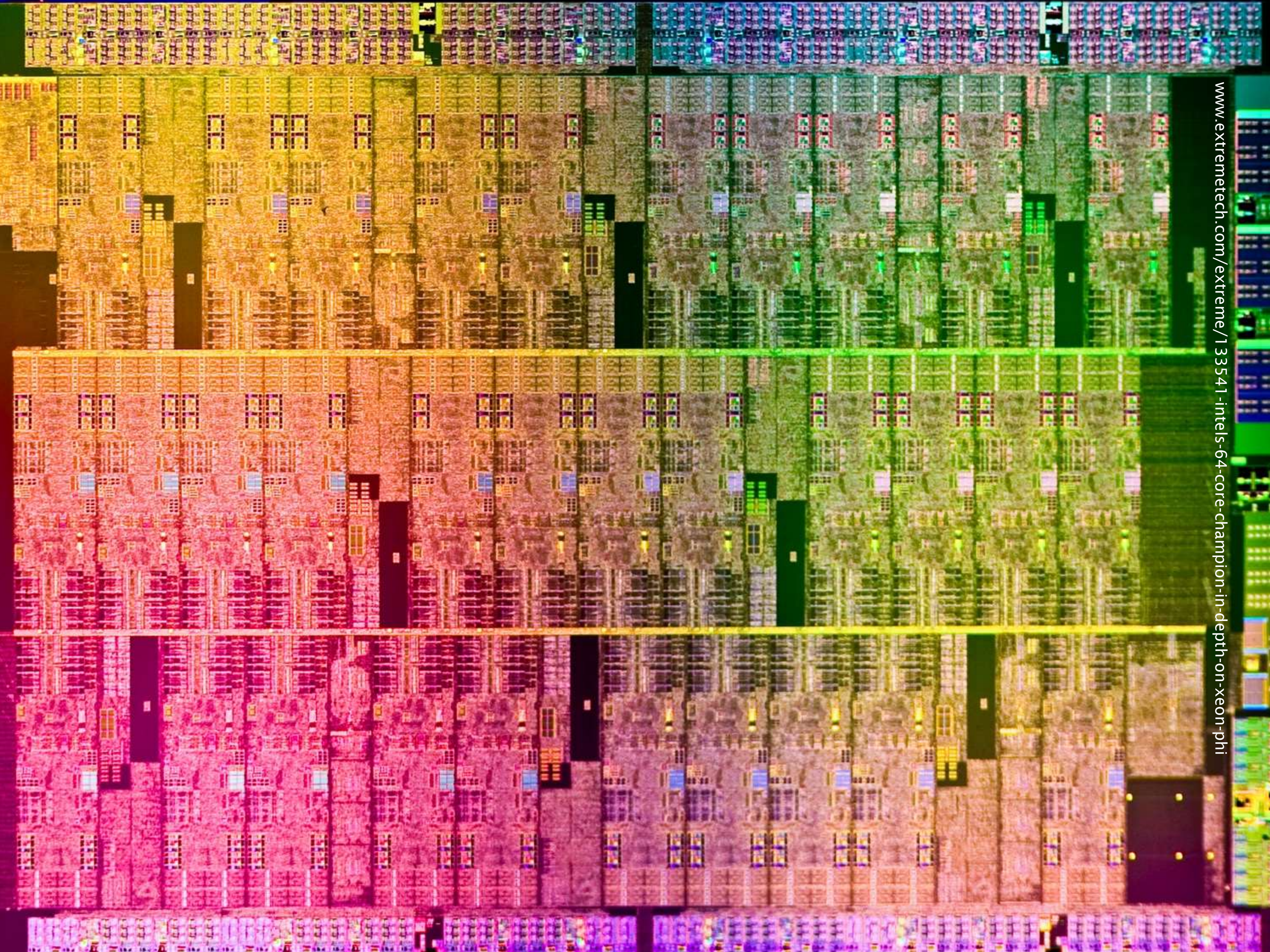


Multi-Pass Blurring

Yang, Sheng, Shafik, Rishad Ahmed, Merrett, Geoff V., Stott, Edward, Levine, Joshua, Davis, James and Al-Hashimi, Bashir (2015) *Adaptive energy minimization of embedded heterogeneous system using regression-based learning*. **PATMOS 2015**, Salvador, BR, 01 - 04 Sep 2015. 8pp.

Towards Many-Core

How do RTM approaches scale with number of cores?

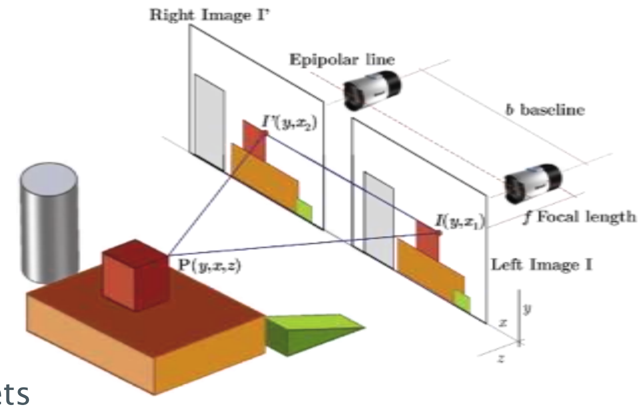


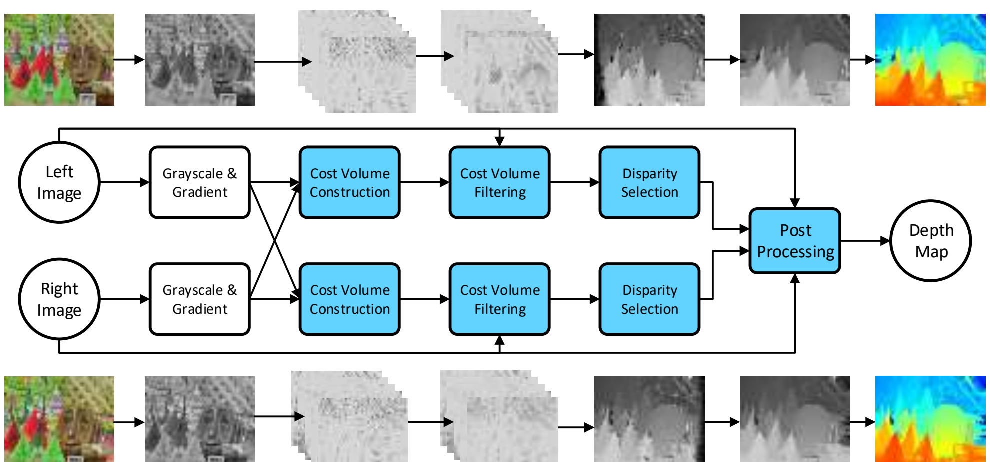
www.extremetech.com/extreme/133541-intels-64-core-champion-in-depth-on-xeon-phi

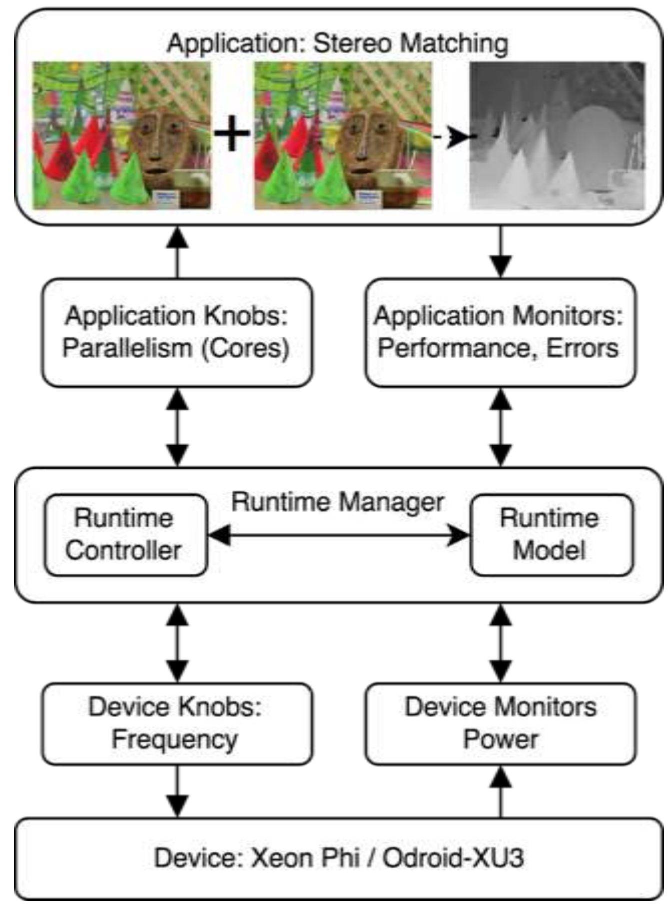
MODEL-BASED RTM

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

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



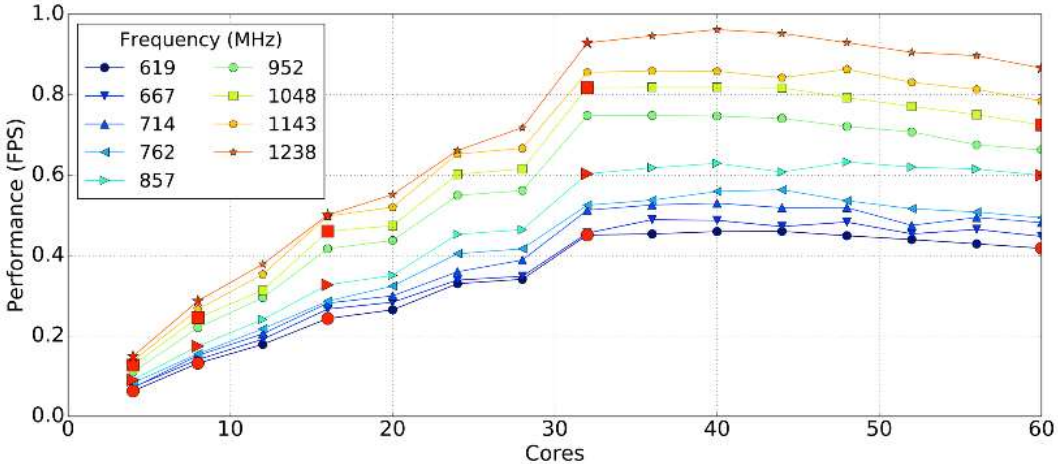


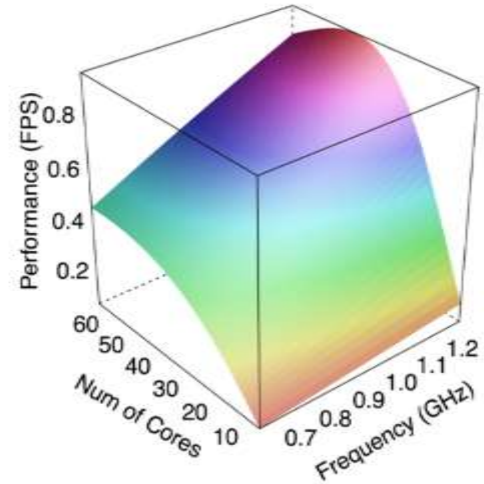


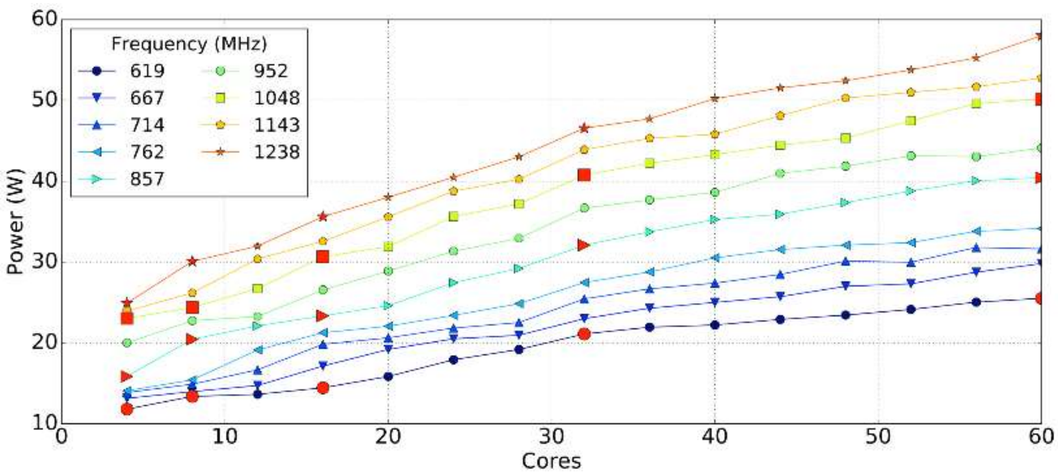
Leech, Charles, Vala, Charan Kumar, Acharyya, Amit, Yang, Sheng, Merrett, Geoffrey and Al-Hashimi, Bashir (2017) *Run-time performance and power optimization of parallel disparity estimation on many-core platforms* **ACM Transactions on Embedded Computing Systems**

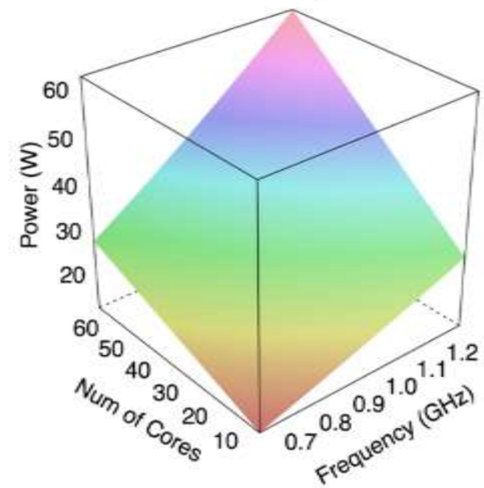
MODEL-BASED RTM

Model Building





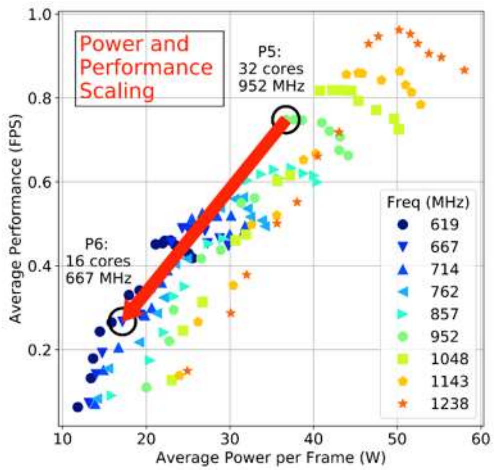


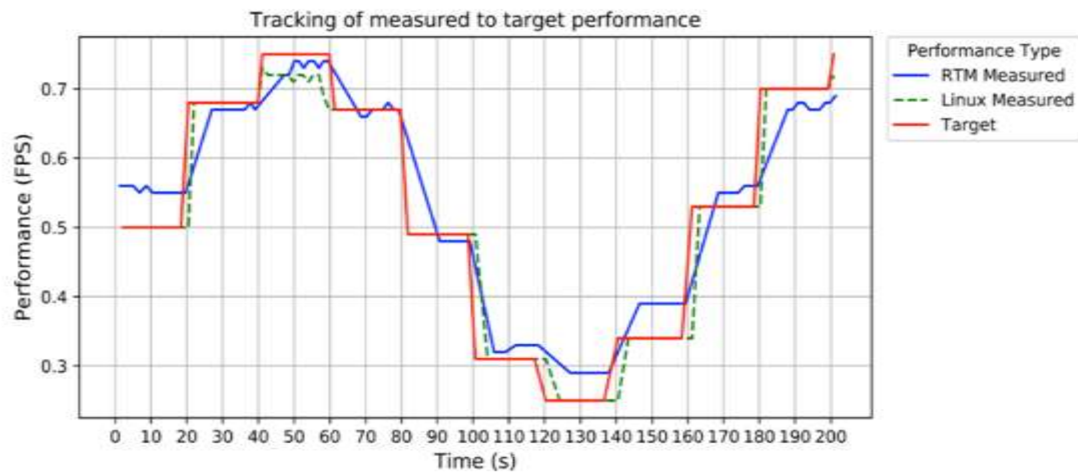


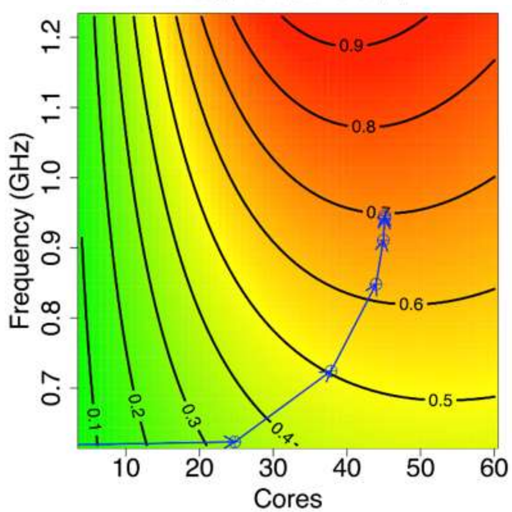
Leech, Charles, Vala, Charan Kumar, Acharyya, Amit, Yang, Sheng, Merrett, Geoffrey and Al-Hashimi, Bashir (2017) *Run-time performance and power optimization of parallel disparity estimation on many-core platforms* **ACM Transactions on Embedded Computing Systems**

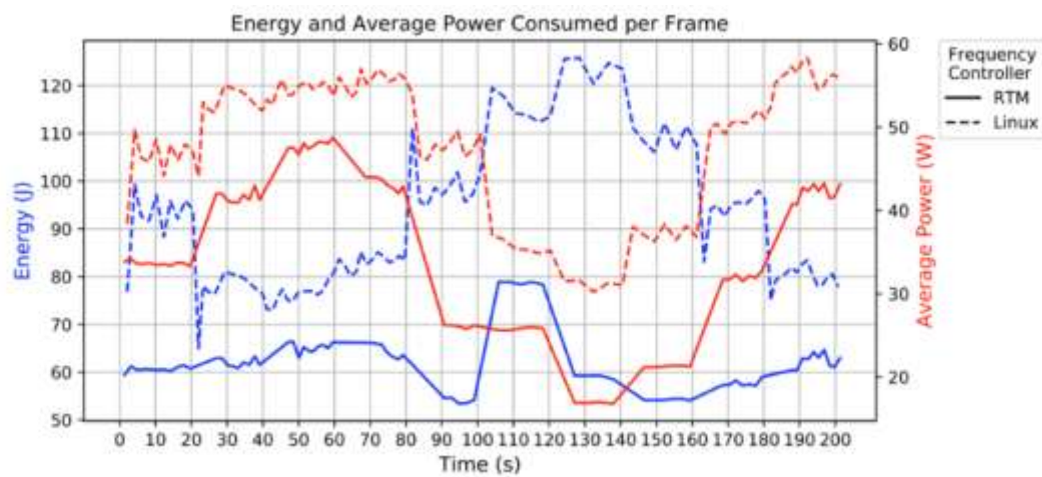
MODEL-BASED RTM

Runtime Management









Leech, Charles, Vala, Charan Kumar, Acharyya, Amit, Yang, Sheng, Merrett, Geoffrey and Al-Hashimi, Bashir (2017) *Run-time performance and power optimization of parallel disparity estimation on many-core platforms* **ACM Transactions on Embedded Computing Systems**

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

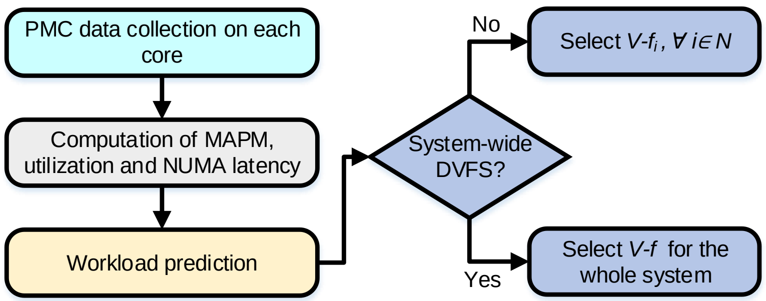
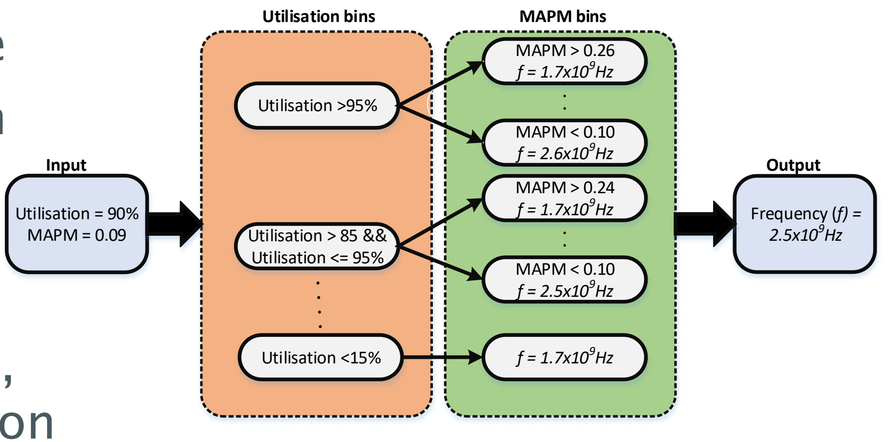


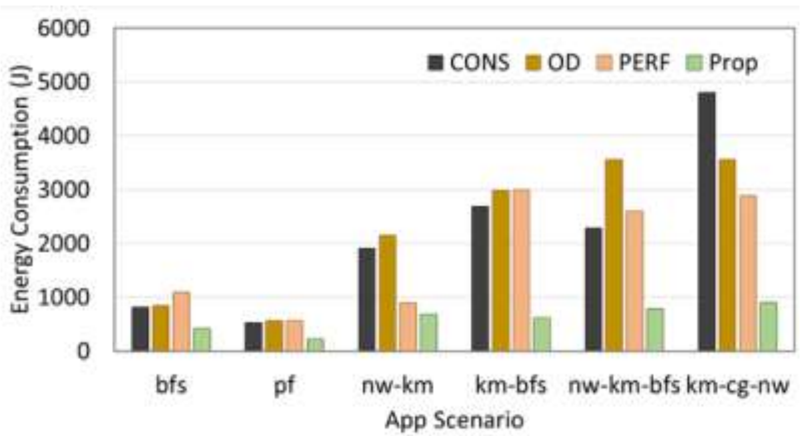
Illustration of various steps in the proposed approach



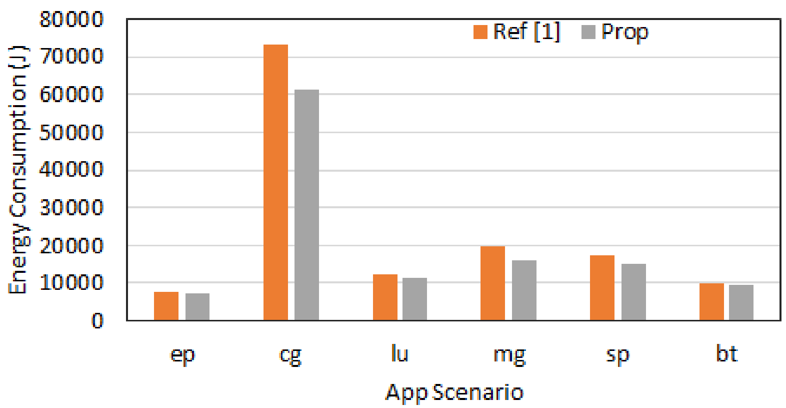
An example of *V-f* setting selection using binning-based approach

Basireddy, Karunakar Reddy, Wachter, Eduardo W., Al-Hashimi, Bashir M. and Merrett, Geoff V. (2018) *Workload-aware runtime energy management for HPC systems*. In **Int'l Workshop Optimization of Energy Efficient HPC & Distributed Systems (OPTIM'18)**. (In Press)

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

Basireddy, Karunakar Reddy, Wachter, Eduardo W., Al-Hashimi, Bashir M. and Merrett, Geoff V. (2018) *Workload-aware runtime energy management for HPC systems*. In **Int'l Workshop Optimization of Energy Efficient HPC & Distributed Systems (OPTIM'18)**. (In Press)

RTM of Novel Platforms

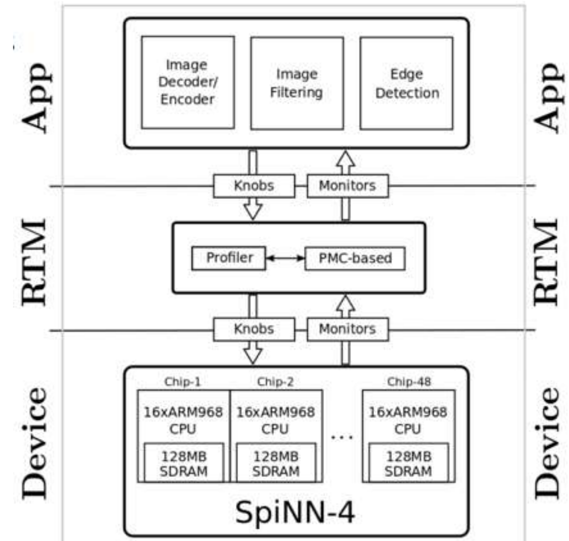
Can our RTM approaches be applied to novel platforms?

Southampton UNIVERSITY OF

RTM ON SPINNAKER

Keynote 5 + S2.3
Thursday 11:10

- 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

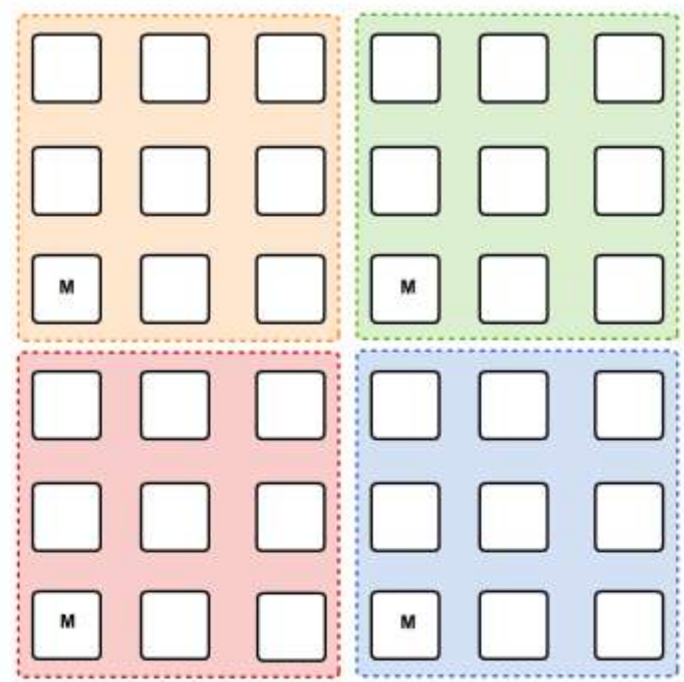
Energy Consumption (J)

Indar Sugiarto, Stephen Furber, Delong Shang, Amit Singh, Bassem Ouni, Geoff Merrett, and Bashir Al-Hashimi, (2018) *Software-defined PMC for runtime power management of a many-core neuromorphic platform*. In **12th Int’l Conf. Computer Engineering and Systems (ICCES 2017)**.

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



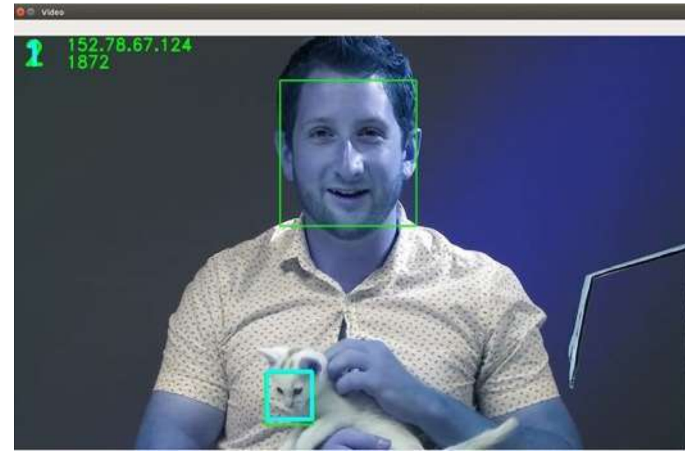
- ***See our (early) demonstration of this***

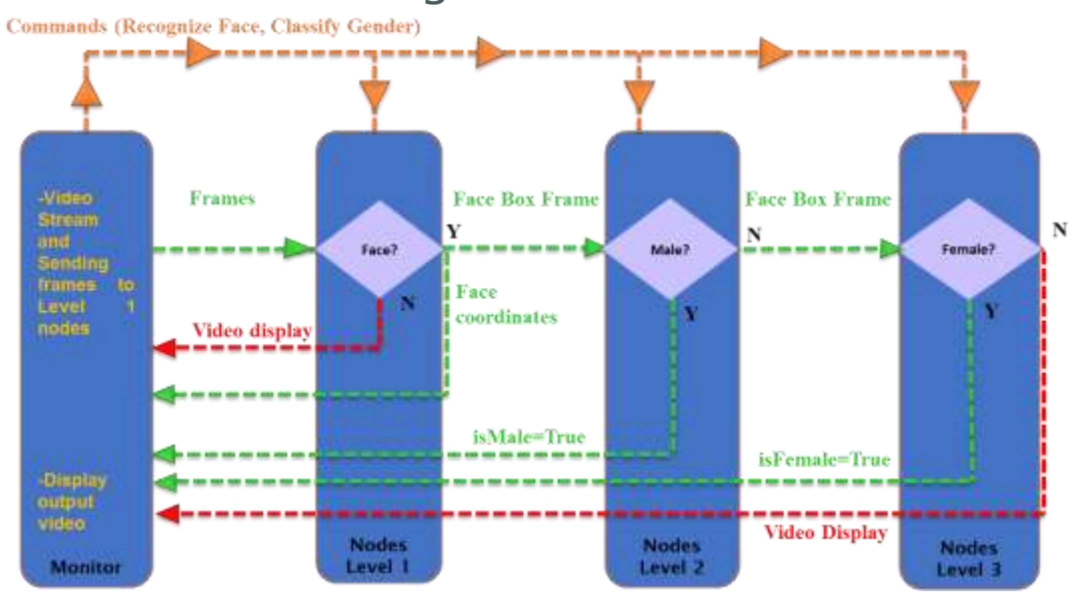
Poster and Demo Session
Thursday 14:30

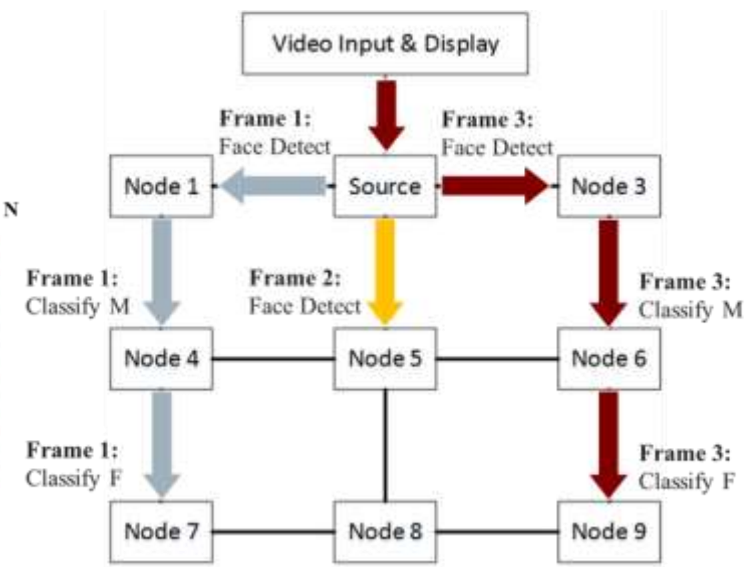
RTM ON THE GRACEFUL PLATFORM

Example Application

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





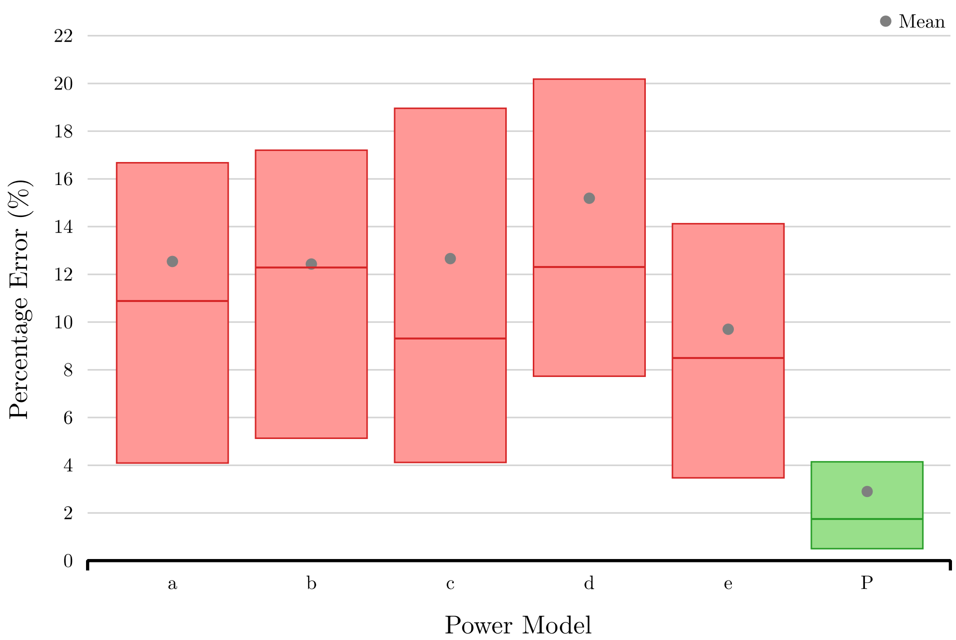


OPEN SOURCE TOOLS

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.



Training: Small set of 20 workloads

Testing: Full set of 60 workloads

[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.

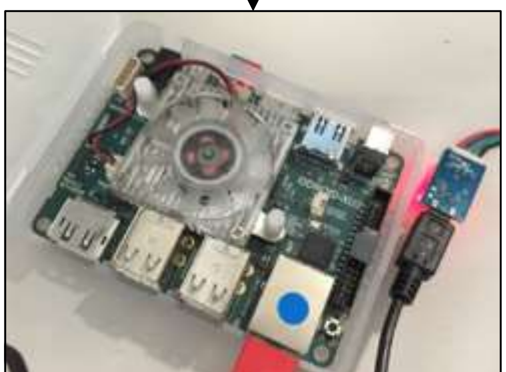
POWMON: METHODOLOGY

www.powmon.ecs.soton.ac.uk

S3.1 Friday 09:50

1. Run workloads
@ different DVFS levels

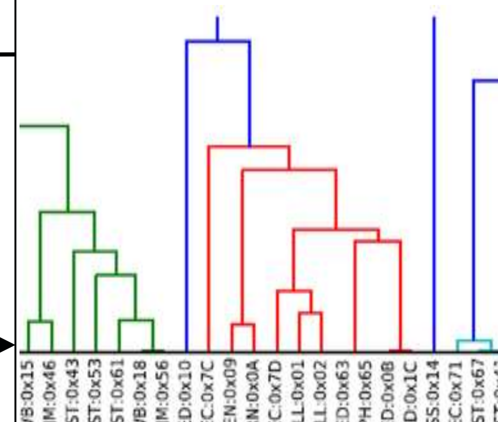
39 workloads used: MiBench, LMBench, Roy Longbottom, ParMiBench and ALPBench



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

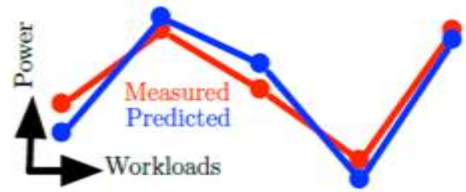
2. Record
- PMCs
- Power, Voltage, Temperature, etc.

3. Choose PMCs
Hierarchical cluster analysis, Correlation matrix analysis, Exhaustive search, etc.



4. Build Model
- OLS multiple regression
- Considers collinearity and heteroscedasticity
- “sensible” equation

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



6. Uses
- OS Run-time management
- Reference for research
- gem5 add-on

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.

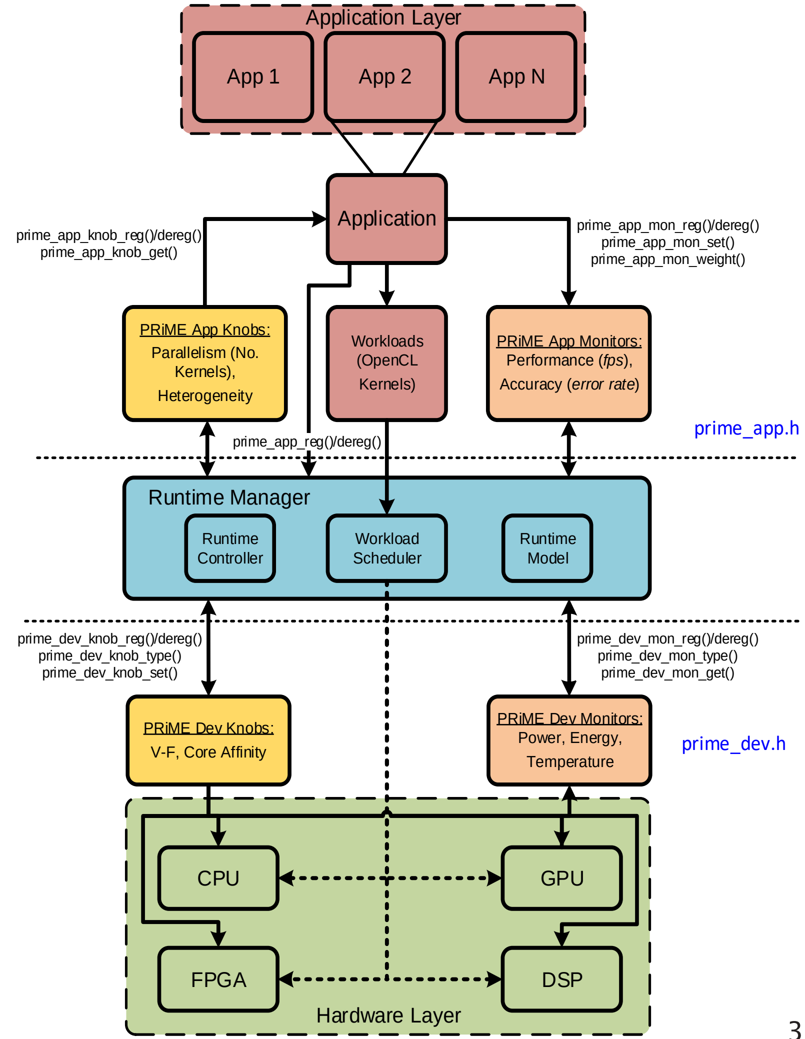
POWMON: METHODOLOGY

Increasing RTM Usability/Comparative Evaluation

S2.2 Thursday 10:15

Plat.	Const.	Space	Type	For	No.
Odroid-XU3	knob	disc	GOVERNER	A7 cluster	1
		disc	GOVERNER	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
		disc	PMC	A15 cores	24
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	$\mathbb{N} \in [1, \infty)$
	Data type	knob	disc	{float, double}
	Device type	knob	disc	{CPU, GPU/FPGA}
	Throughput	mon	cont	$\mathbb{R} \in [10, \infty)$
	Error	mon	cont	$\mathbb{R} \in (-\infty, 1e^{-12}]$
Video decoder	Throughput	mon	cont	$\mathbb{R} \in [25, \infty)$
Whetstone	Threads	knob	disc	$\mathbb{N} \in [1, \infty)$
	Throughput	mon	cont	$\mathbb{R} \in [2.5, \infty)$



CONCLUSIONS

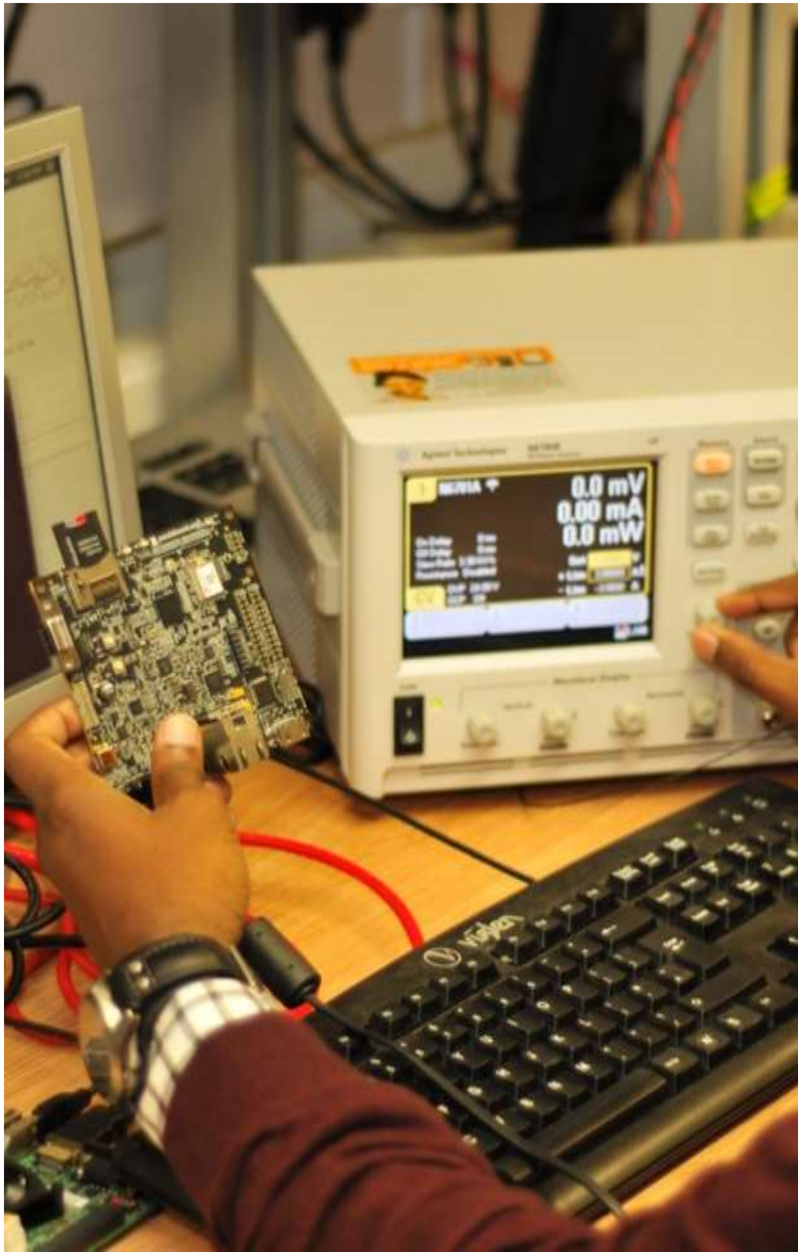
Runtime Power Management

- 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?

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