PRiME prime-project.org
Graceful
EPSRC Pioneering research and skills
UNIVERSITY OF Southampton

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

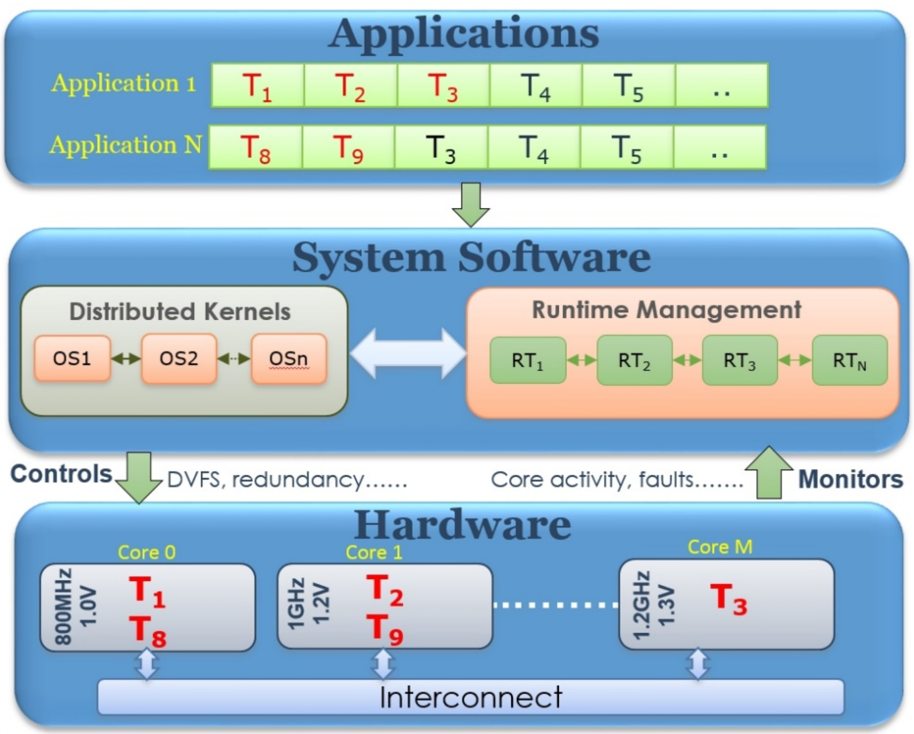
Dr Geoff Merrett

19 September 2018 | Arm Research Summit, Cambridge, UK

THE PRIME PROJECT

www.prime-project.org

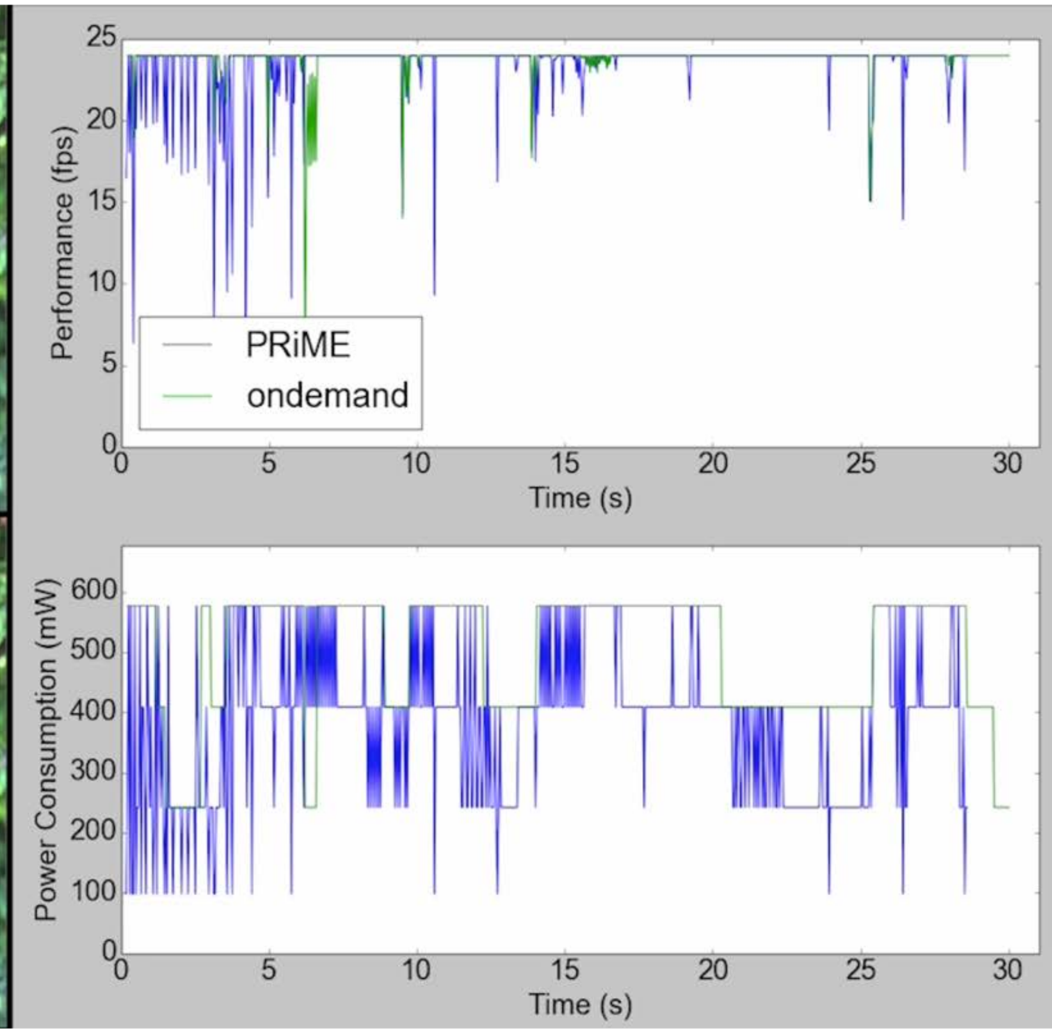
"Enable the sustainability of **many-core scaling** by preventing the uncontrolled increase in **energy consumption** and **unreliability** through a step change in holistic design methods and **cross-layer** system optimisation."



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

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

RUNTIME POWER MANAGEMENT

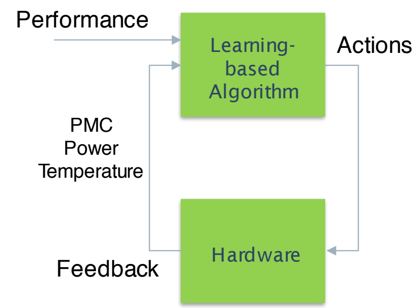


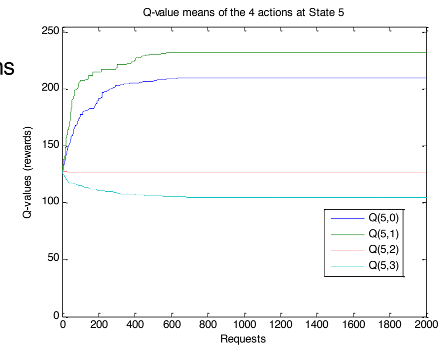
Maeda-Nunez, Luis Alfonso, Anup K. Das, Rishad A. Shafik, Geoff V. Merrett, and Bashir Al-Hashimi. "*PoGo: an application-specific adaptive energy minimisation approach for embedded systems.*" **HiPEAC Workshop on Energy Efficiency with Heterogenous Computing**, 2015

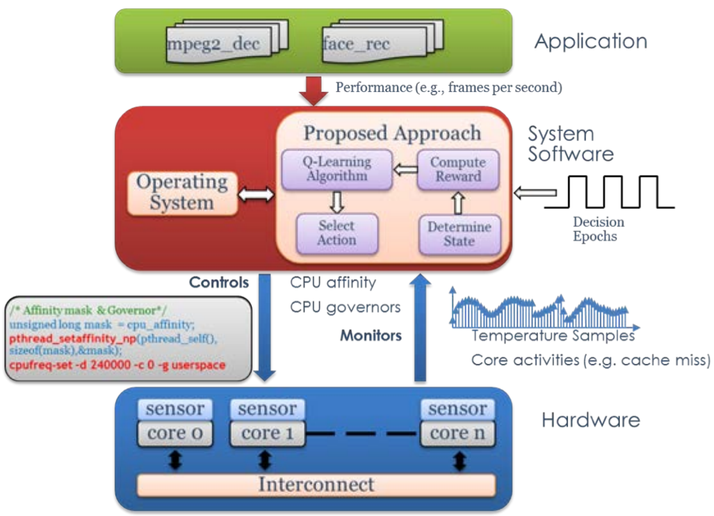
LEARNING OPTIMAL DVFS CHOICES

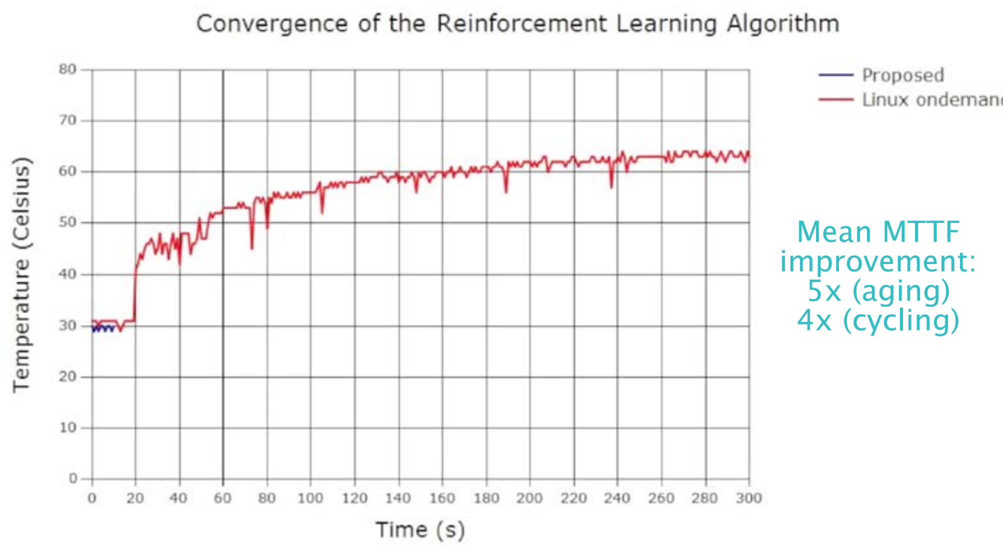
Reinforcement Learning

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









Mean MTTF improvement:
5x (aging)
4x (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?

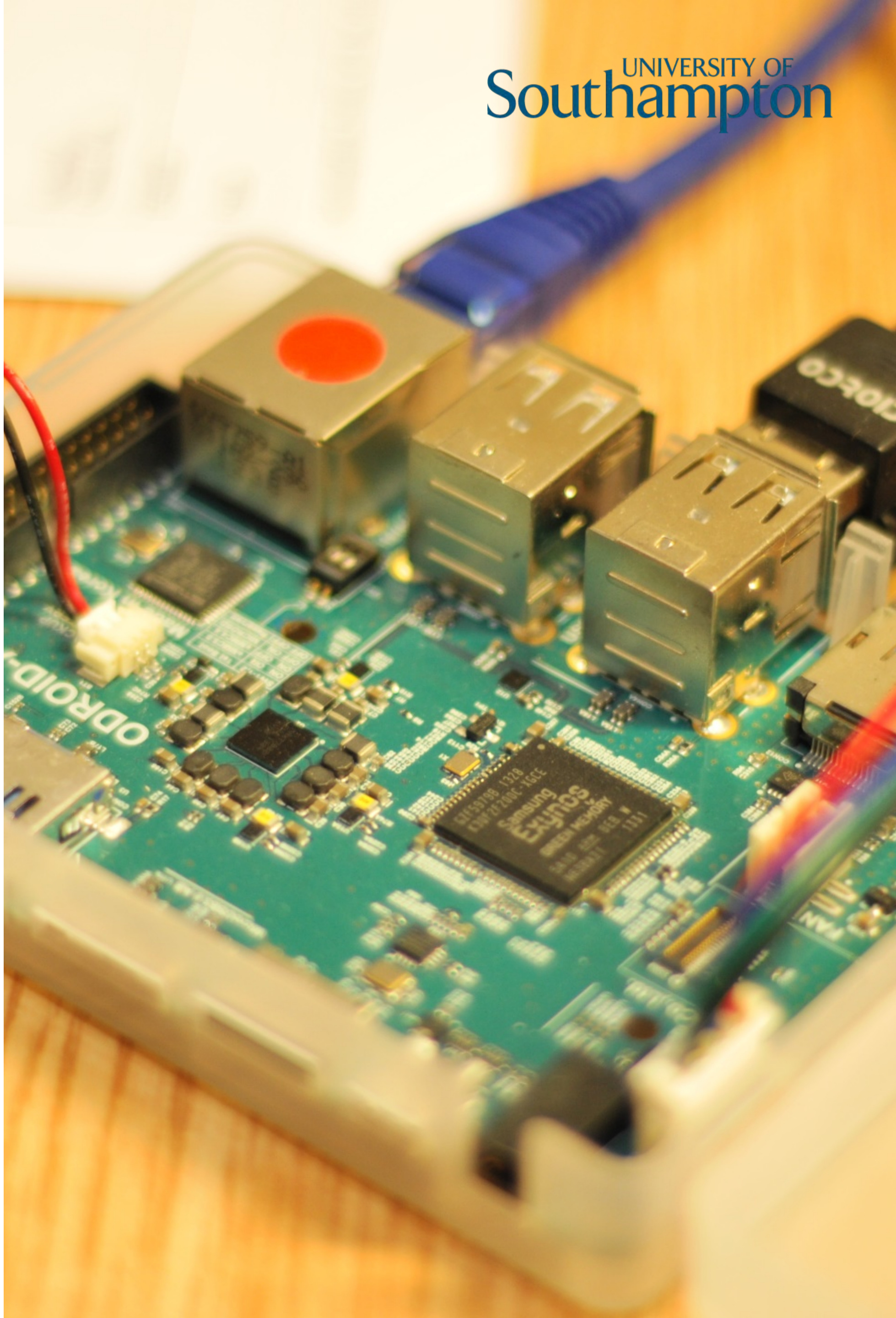
Towards Many-Core

- 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/

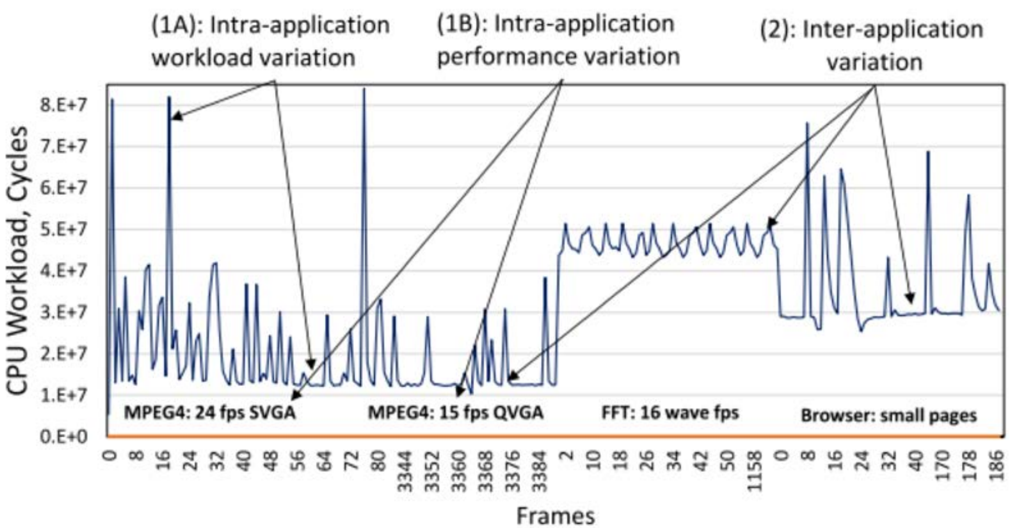


RTMs and Application Workloads

From single > sequential > concurrent execution

EXECUTING MULTIPLE APPLICATIONS

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

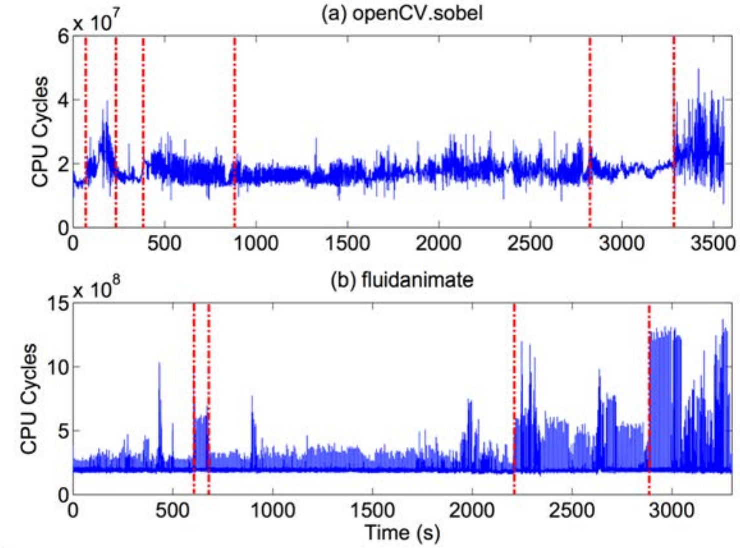


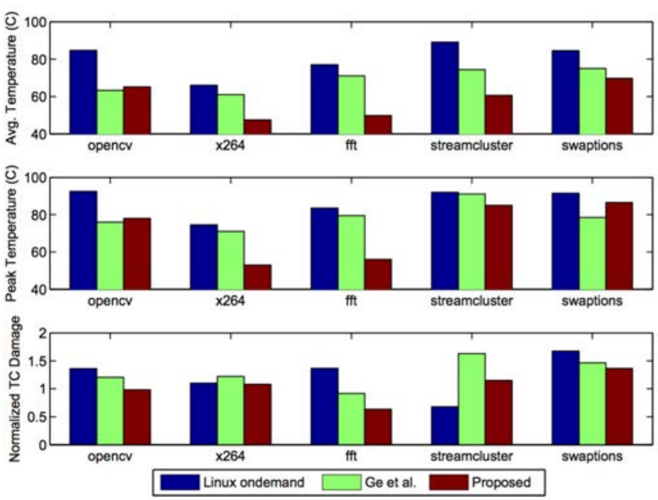
  - Overlapping applications (*concurrent execution*)

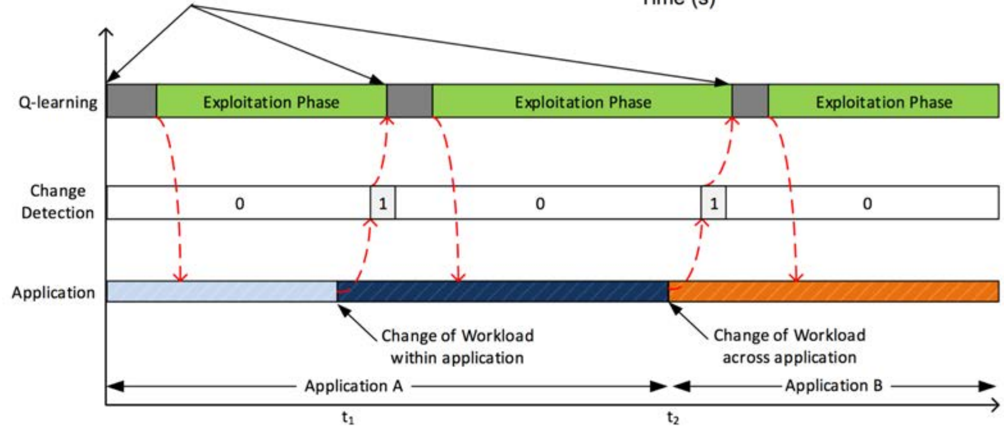
Shafik, Rishad, Das, Anup, Maeda-Nunez, Luis, Yang, Sheng, Merrett, Geoff and Al-Hashimi, Bashir (2015) *Learning transfer-based adaptive energy minimization in embedded systems*. IEEE TCAD.

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)

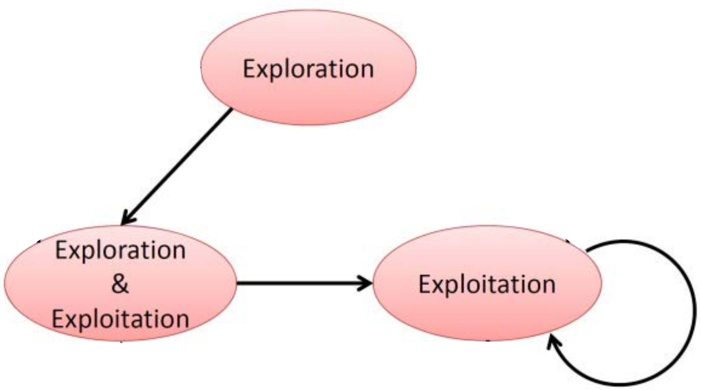


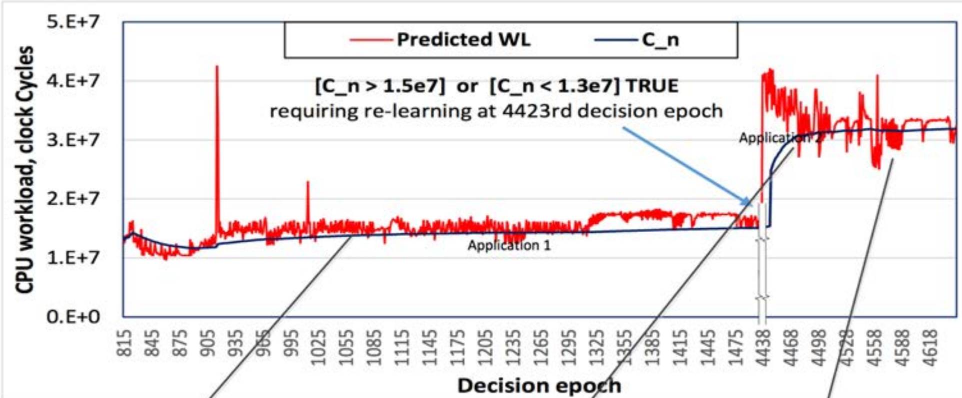




TRANSFERING LEARNING

- 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, 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**.

Online vs Offline

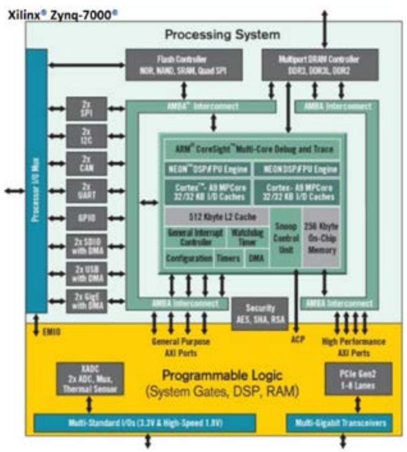
Can we improve RTM through offline characterisation?

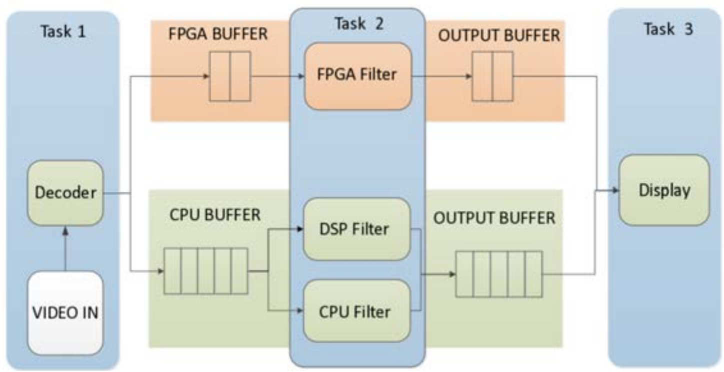
MODEL-BASED RTM: HETEROGENEITY

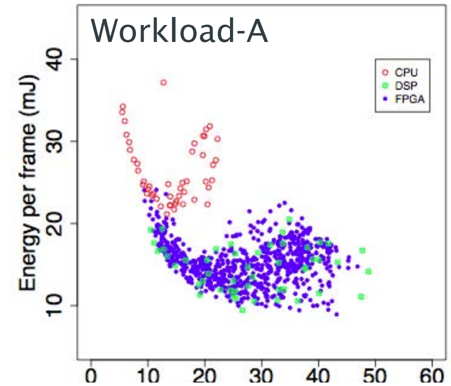
Heterogeneous Platforms

Run-time changes in:

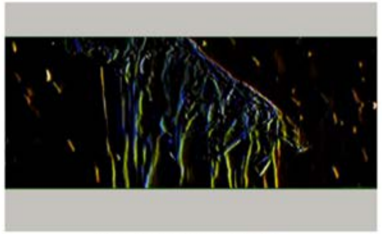
- Performance requirements
- Application workload changes





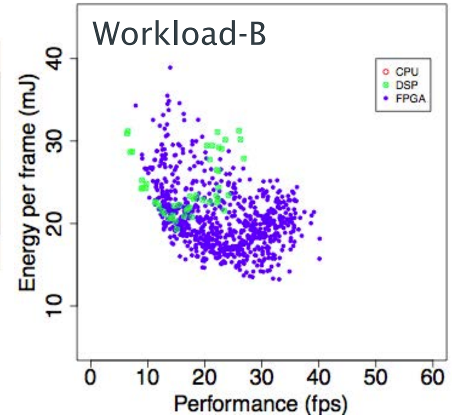


(b) Original



(c) Edge detected

(d) Blurred



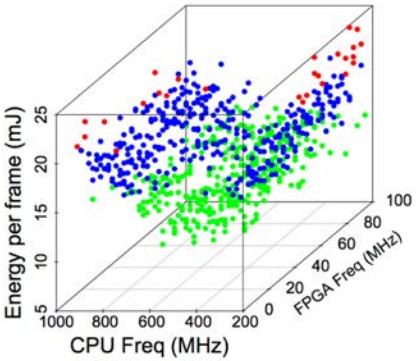
Yang, Sheng, Shafik, Rishad Ahmed, Merrett, Geoff V., Stott, Edward, Levine, Joshua, Davis, James and Al-Hashimi, Bashir (2015) *Adaptive energy minimization of embedded heterogeneous system using regression-based learning.* **PATMOS 2015**, Salvador, BR, 01 - 04 Sep 2015. 8pp.

PRiME www.prime-project.org

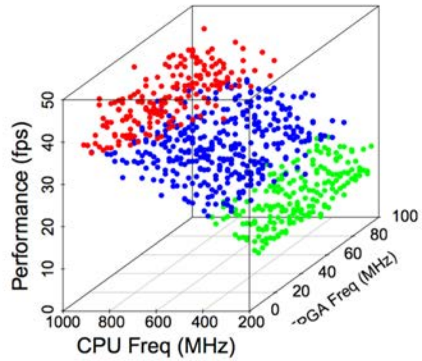
UNIVERSITY OF Southampton

MODEL-BASED RTM: HETEROGENEITY

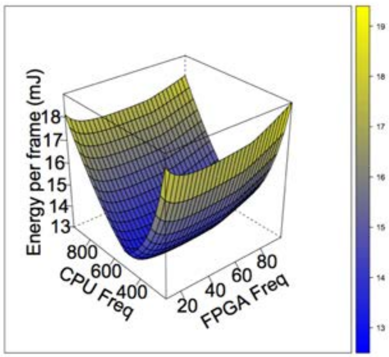
Heterogeneous Platforms



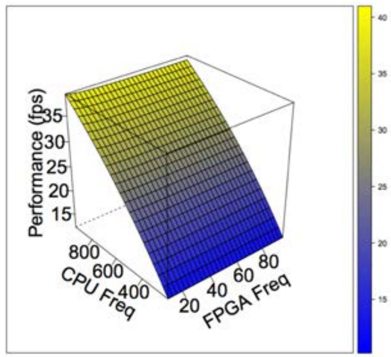
(a) FPGA measured energy



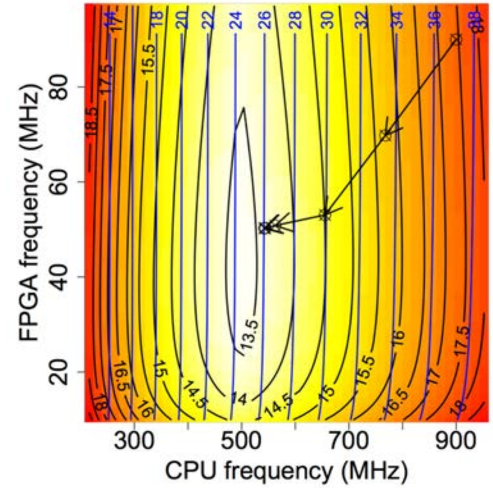
(b) FPGA measured performance



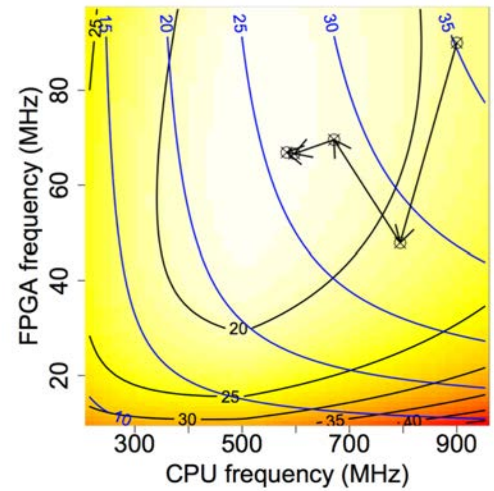
(c) FPGA modeled energy



(d) FPGA modeled performance



Edge Detection

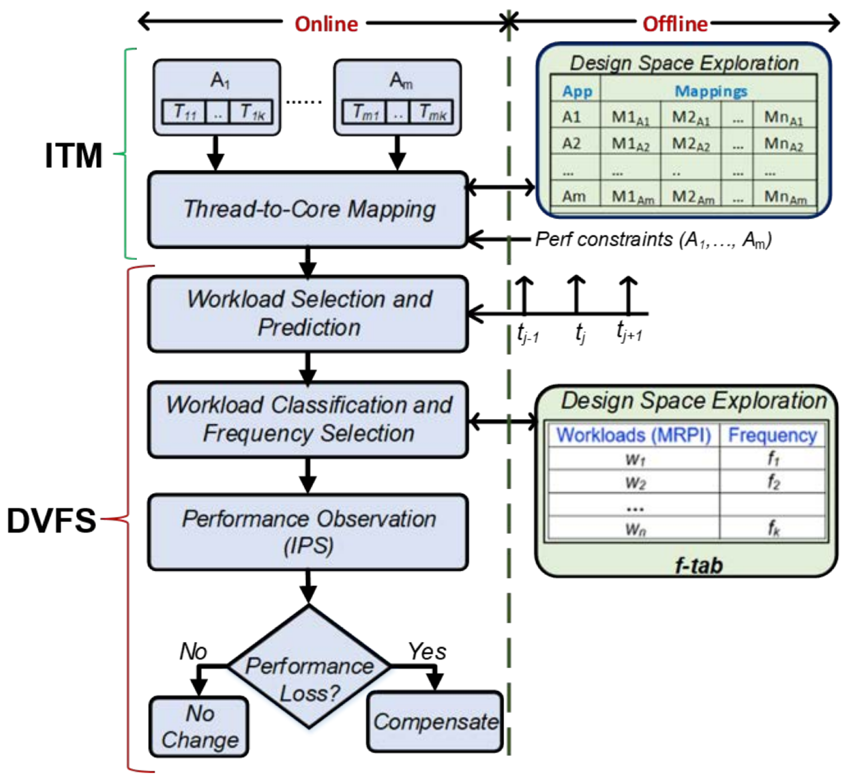


Multi-Pass Blurring

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.

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

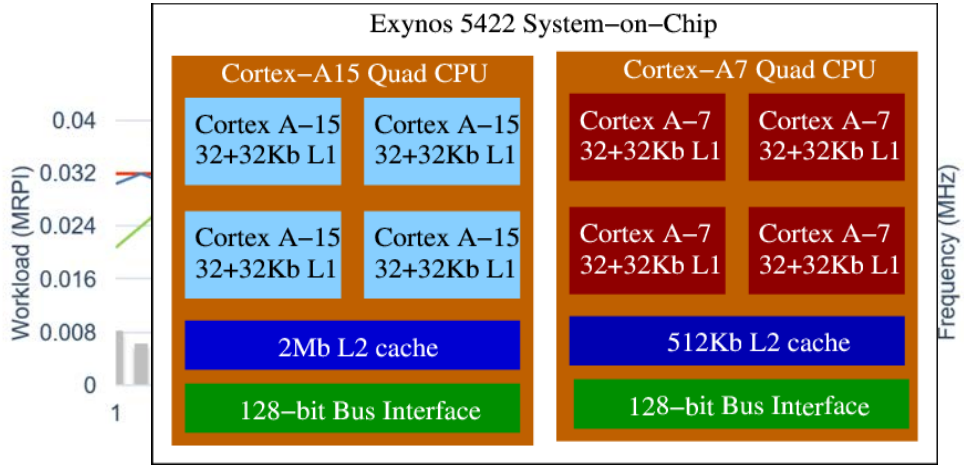


Image: Catalán et al., Performance and Energy Optimization of Matrix Multiplication on Asymmetric big.LITTLE Processors, 2015

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.

Towards Many-Core

How do RTM approaches scale
with number of cores?

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

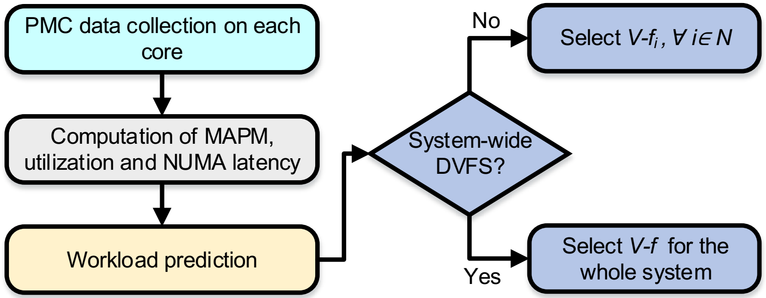
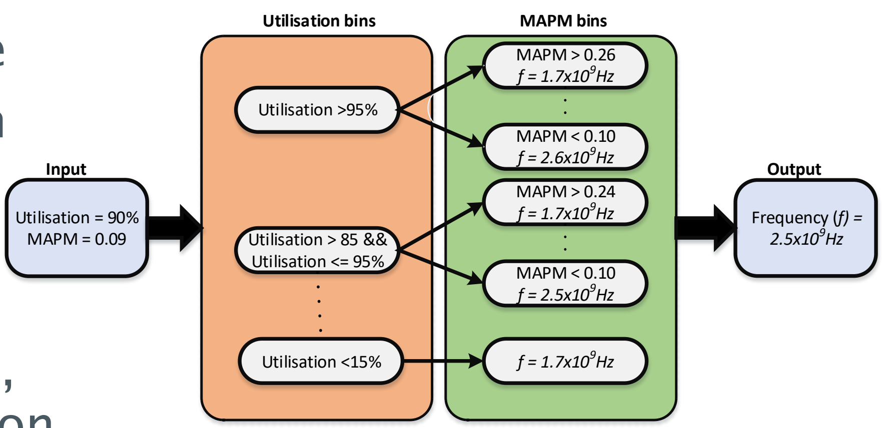


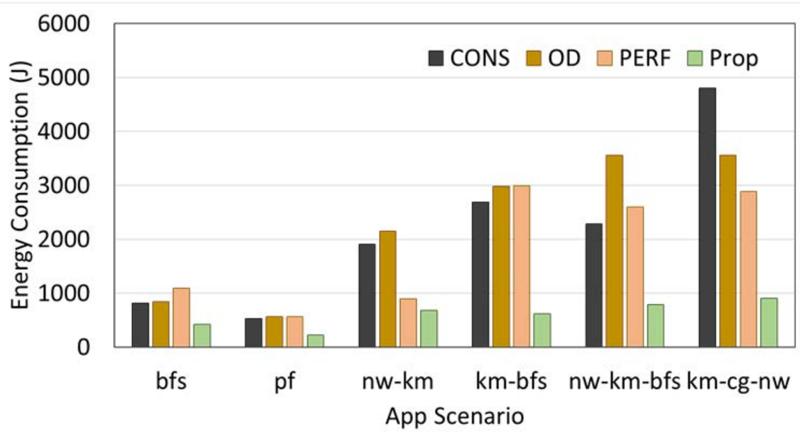
Illustration of various steps in the proposed approach



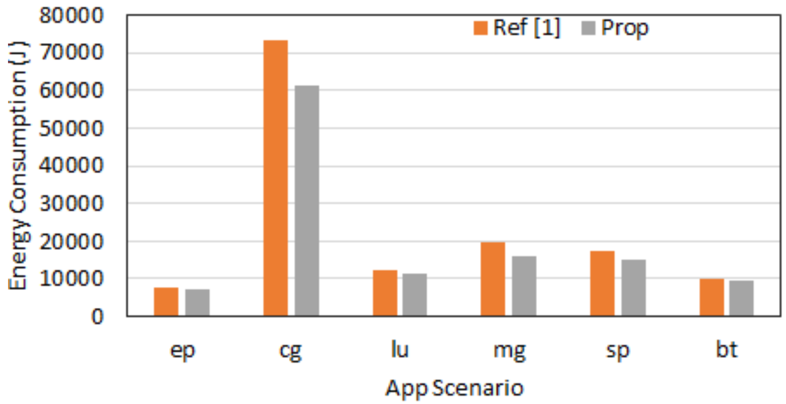
An example of *V-f* setting selection using binning-based approach

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)**.

ENERGY RTM ON HPC SYSTEMS



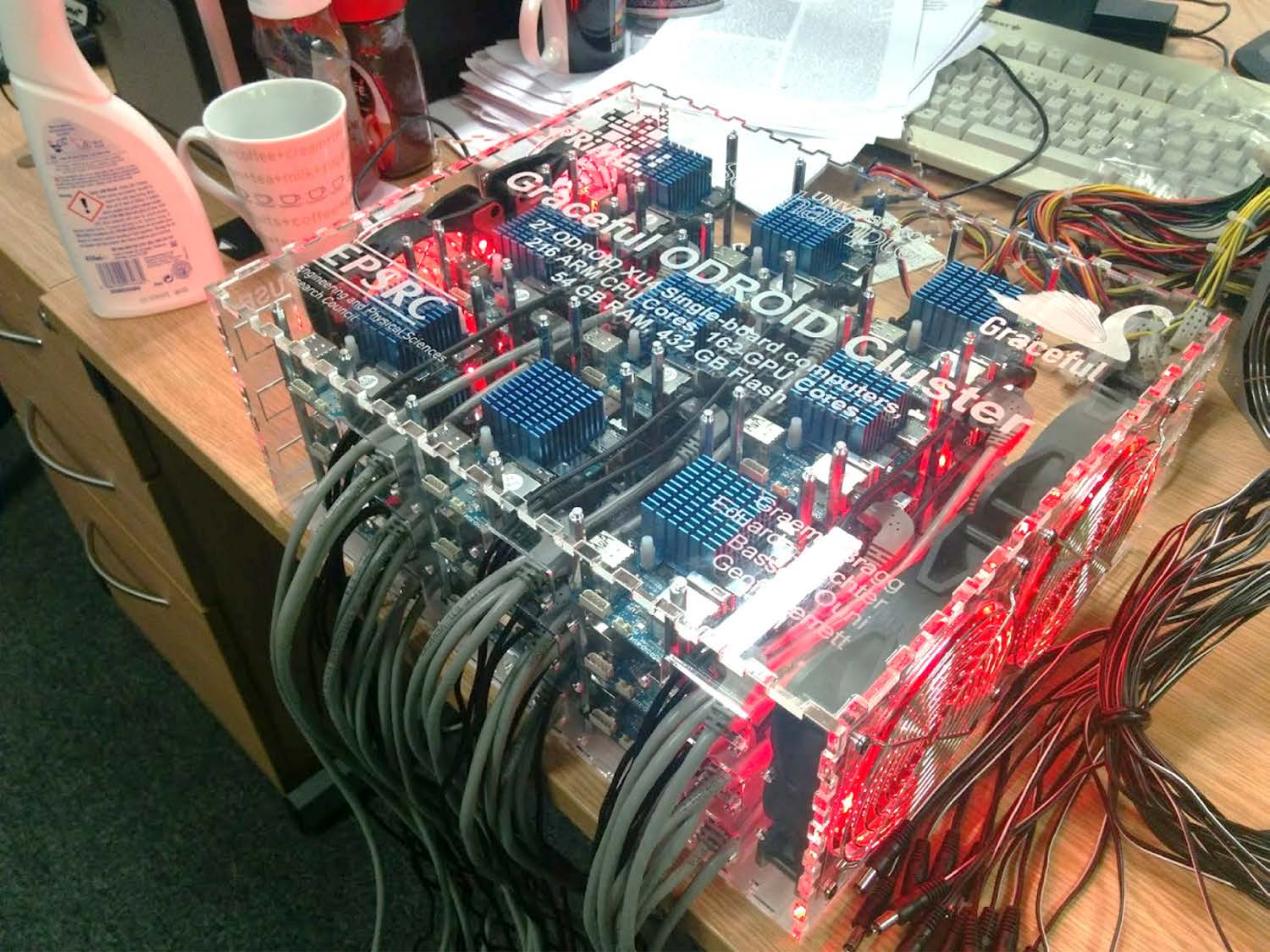
Energy consumption of different approaches



Comparison of presented approach with Sundriyal et al

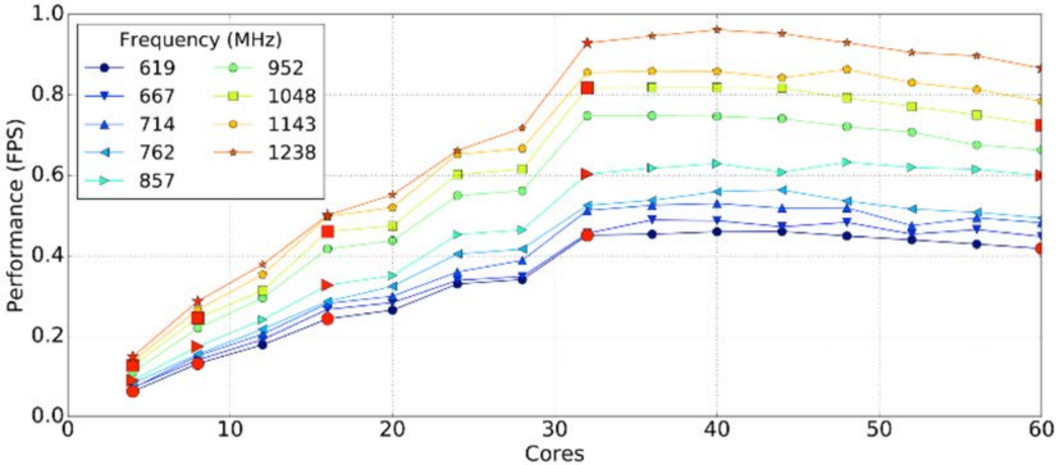
- Xeon E5-2630 (12 cores, 24 threads) and Xeon Phi 7620P (61 cores, 244 threads); NAS and Rodinia benchmarks
- Proposed (Prop) approach achieves energy savings of up to 81% (Xeon) and 61% (Phi) compared to Linux's governors
- Outperforms Sundriyal *et al.* by 10% in energy efficiency and 3.7% in performance

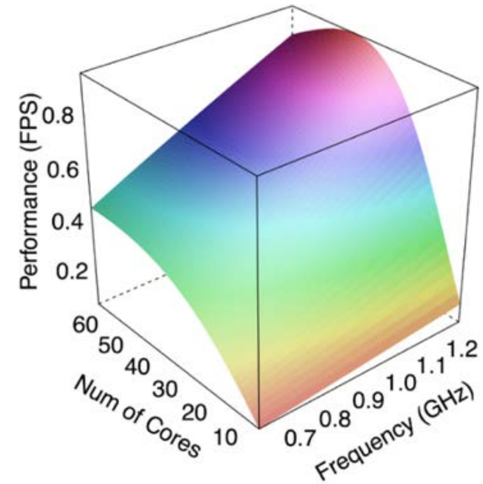
Basireddy, Karunakar Reddy, Wachter, Eduardo W., Al-Hashimi, Bashir M. and Merrett, Geoff V. (2018) *Workload-aware runtime energy management for HPC systems*. In **Int'l Workshop Optimization of Energy Efficient HPC & Distributed Systems (OPTIM'18)**.

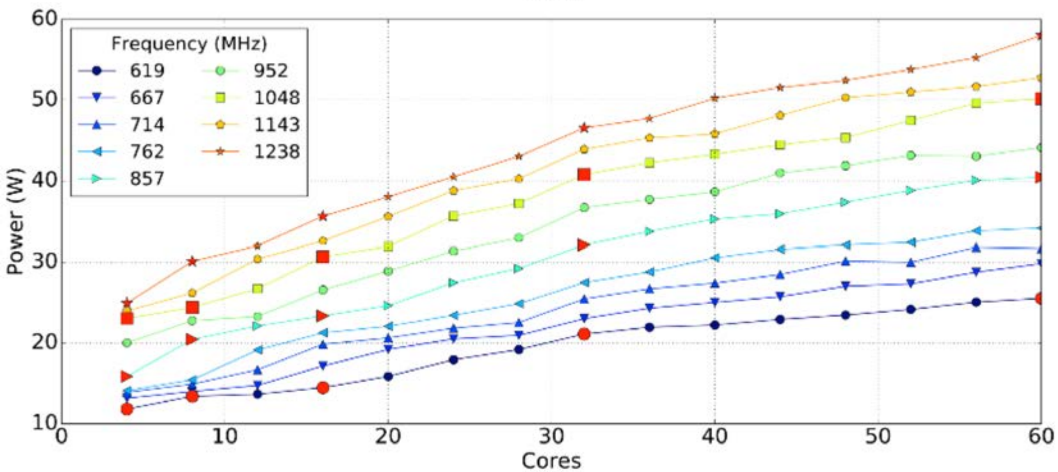

Graceful ODROID Cluster
27 ODROID XU4 Single-board computers,
216 ARM CPU Cores, 162 GPU Cores,
54 GB RAM, 432 GB Flash
EPSRC
Engineering and Physical Sciences
Graceful

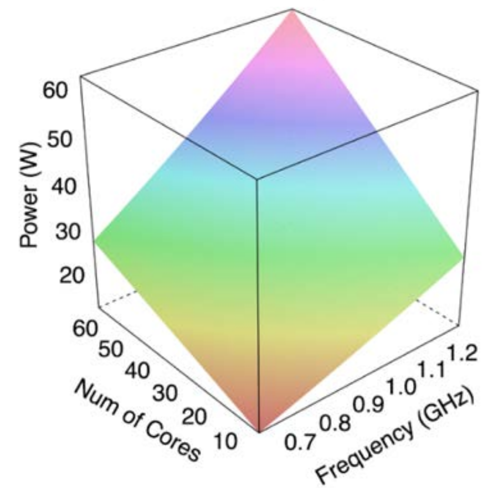
MODEL-BASED RTM

Model Building





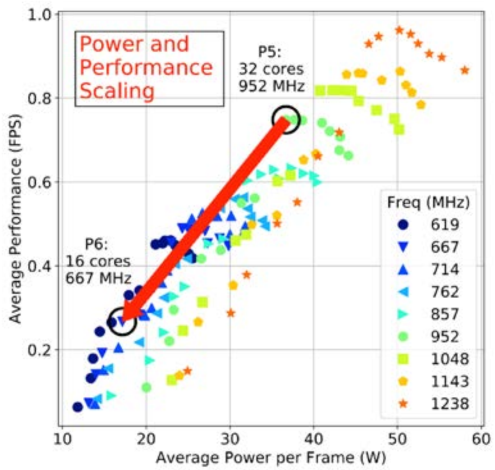


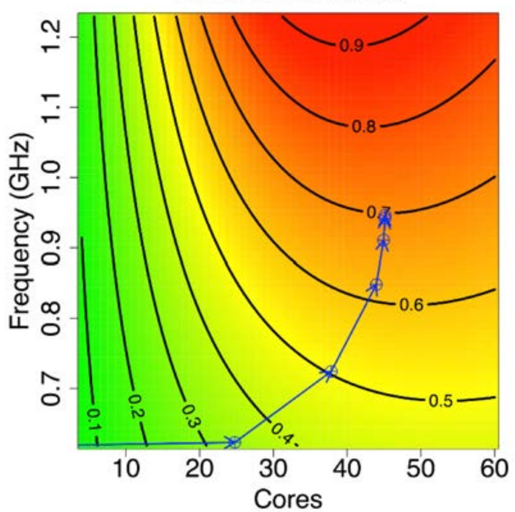


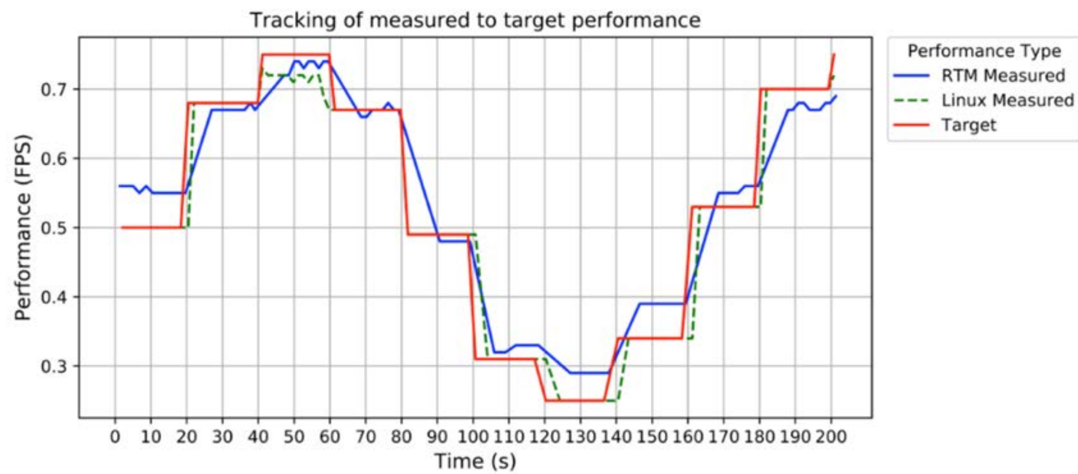
Leech, Charles, Vala, Charan Kumar, Acharyya, Amit, Yang, Sheng, Merrett, Geoffrey and Al-Hashimi, Bashir (2017) *Run-time performance and power optimization of parallel disparity estimation on many-core platforms* **ACM Transactions on Embedded Computing Systems**

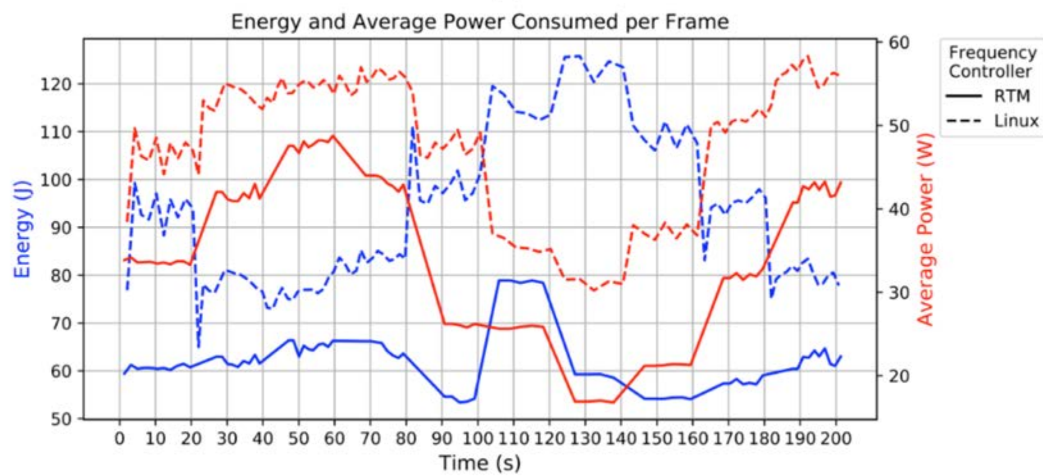
MODEL-BASED RTM

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**

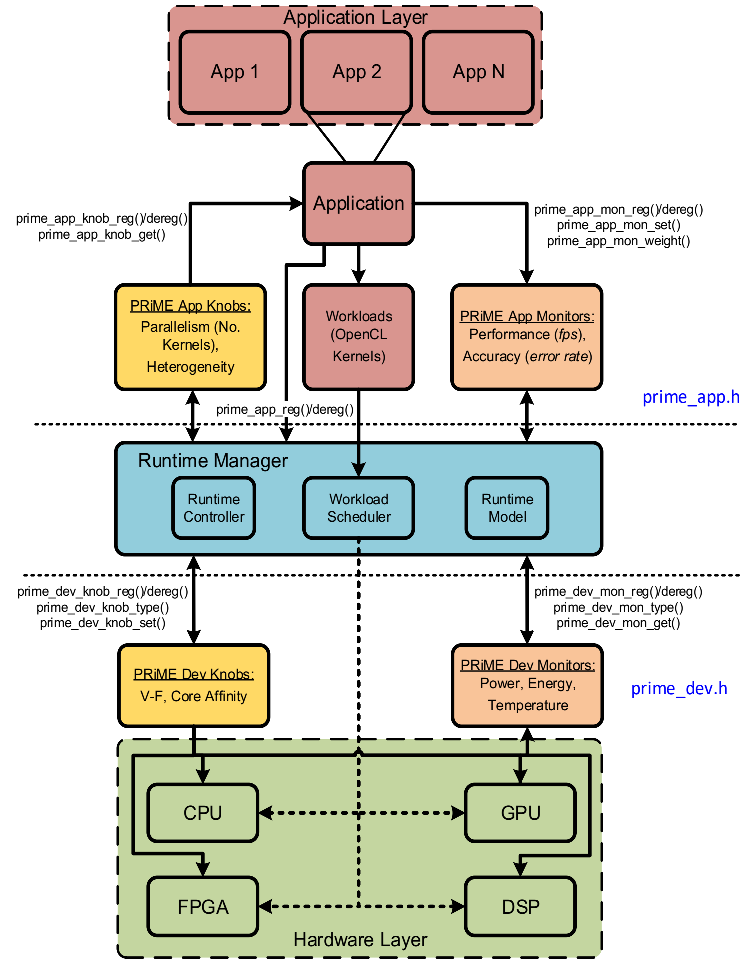
OPEN SOURCE TOOLS

PRIME RTM FRAMEWORK

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

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	$\mathbb{R} \in [2.5, \infty)$

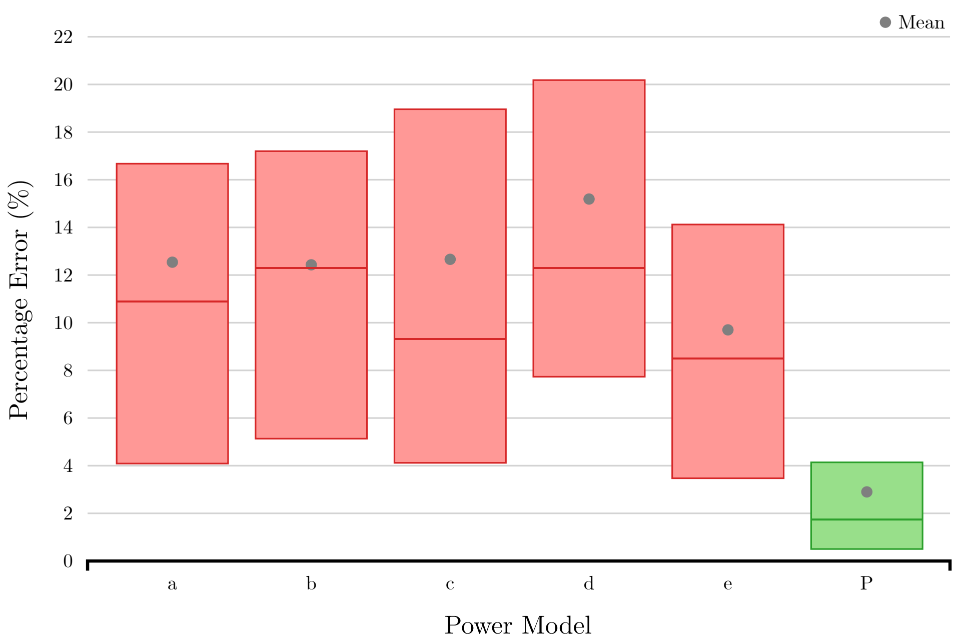


G.M. Bragg, C.R. Leech, D. Balsamo, J.J. Davis, E.W. Wachter, G.V. Merrett, G.A. Constantinides, B.M. Al-Hashimi, (2018) "*An application- and platform-agnostic control and monitoring framework for multicore systems*". **Int'l Conf. Pervasive and Embedded Computing**, Porto, Portugal. 29-30 Jul 2018 (Best Paper Award).

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.

CONCLUSIONS

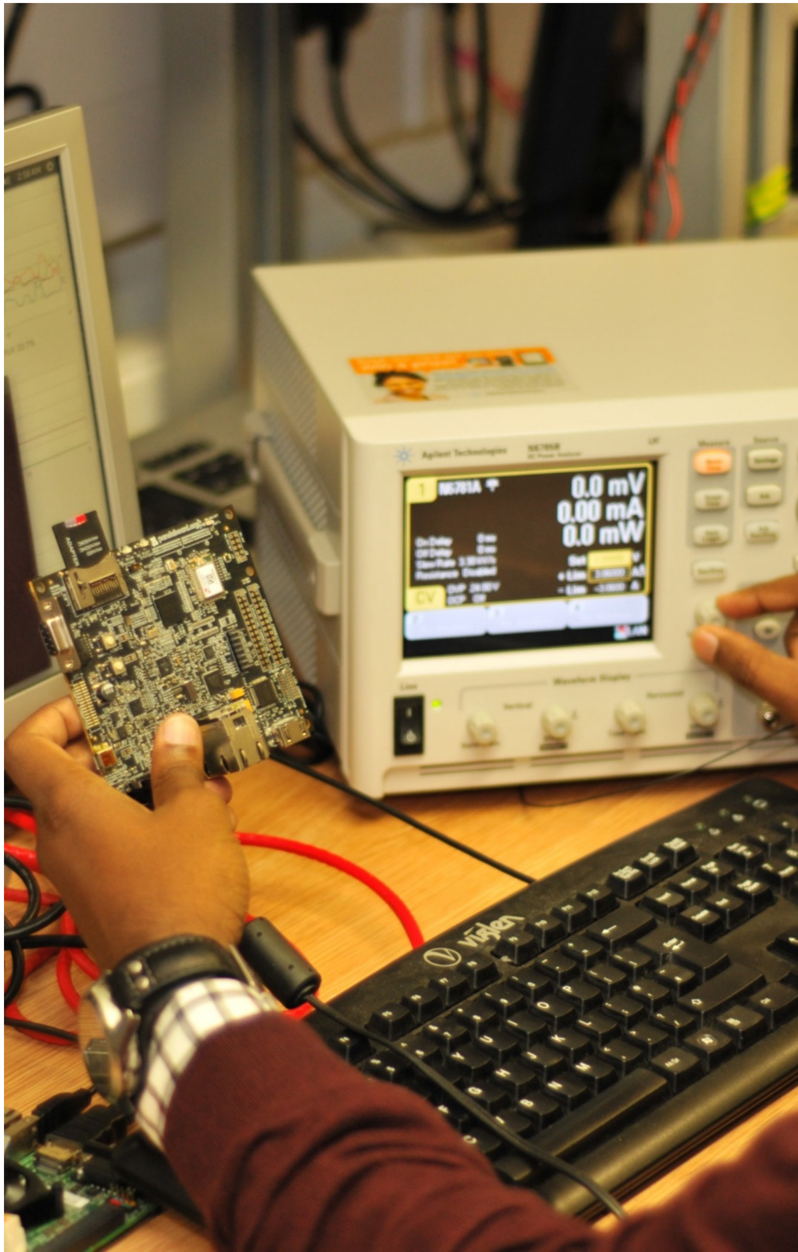
Runtime Power Management

- Single > multiple > concurrent applications
- Online vs offline+online approaches
- >> Number of cores
- 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

A. Singh, C. Leech, K. Basireddy, B.M. Al-Hashimi, G.V. Merrett, (2017) "*Learning-based run-time power and energy management of multi/many-core systems: current and future trends*". **Journal of Low Power Electronics**



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