Hartree Centre
Science & Technology Facilities Council

An update from the Hartree Centre on ARM based activities

Neil Morgan

neil.morgan@stfc.ac.uk

14th November 2016 – ARM HPC User Group Meeting

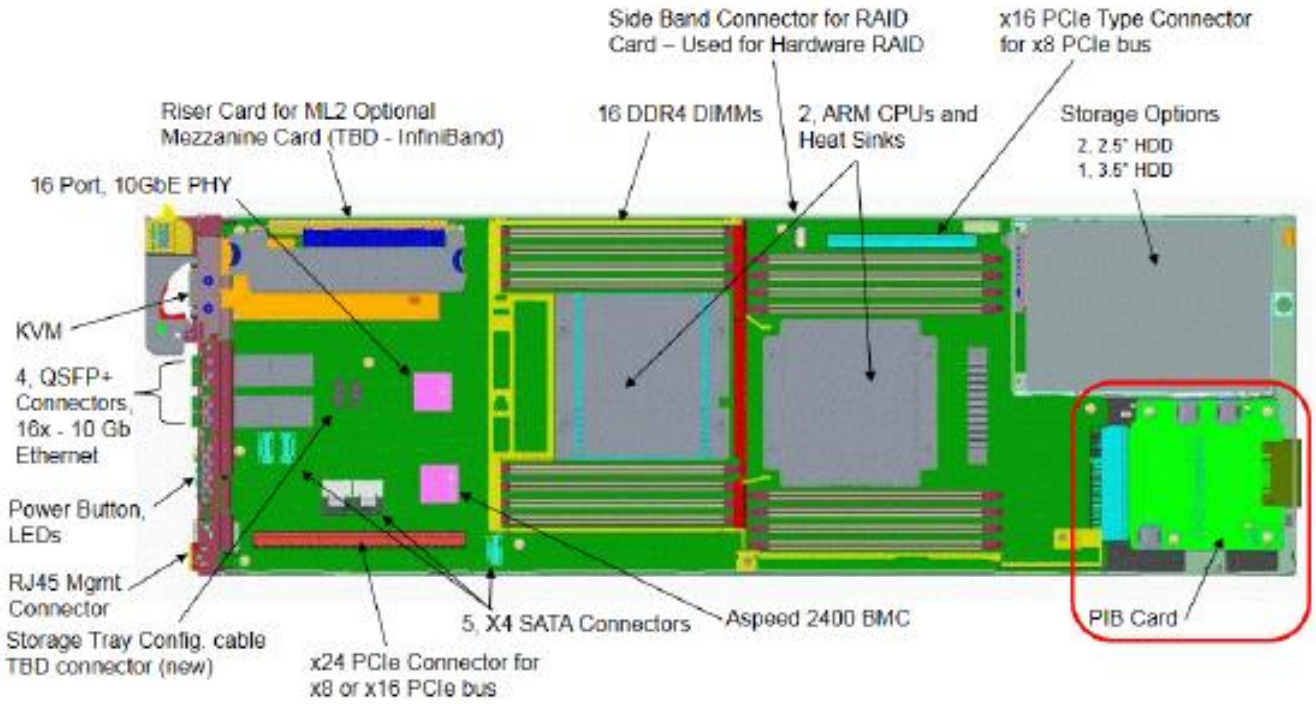
Hartree Centre: An Update

- On going collaboration with Lenovo, Cavium, Mellanox, RedHat, UoL, UoC has resulted in 2 presented works:

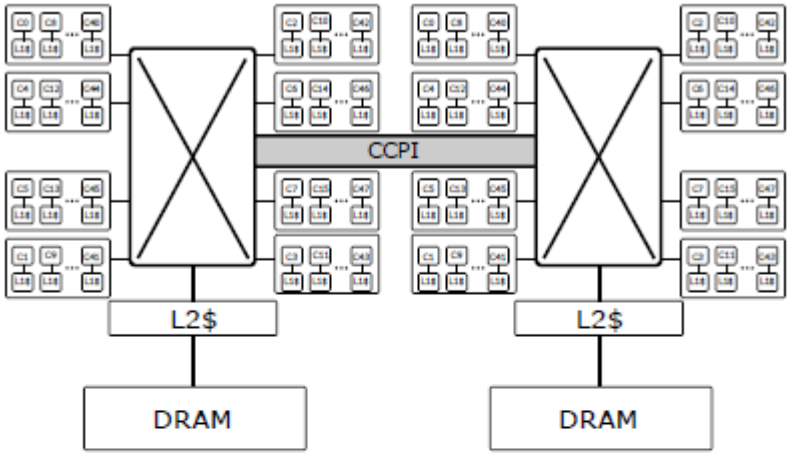
- Quantifying Energy Use in Dense Shared Memory HPC Node - Milos Puzovic, Srilatha Manne, Shay Galon and Makoto Ono. E2SC Workshop @ SC16

- Guided Scheduling Using Accurate Per-Core Performance Hardware Counters. Milos Puzovic, Jeyan Thiyagalingam, David Greaves. ARM Research Summit.

Quantifying energy use in a dense shared memory HPC node – 'Asian Cat'



Cavium ThunderX SoC



L1D-Cache	Policy	Write-through
	Type	Private
	Size	32Kb
	Associativity	32-way
	Block	128-bytes
I-Cache	Size	78Kb
	Associativity	39-way
	Block size	128-bytes
L2-Cache	Policy	Write-back
	Type	Shared
	Size	16MB
	Associativity	16-way
	Block	128-bytes

Power Metering and Accuracy

- Via a hot-swap controller (HSC)
- Board Management Controller (BMC)
  - Queries and Translates
- Intelligent Power Management Interface
  - Power, Performance, Thermal
- Interval between consecutive measurements 3ms

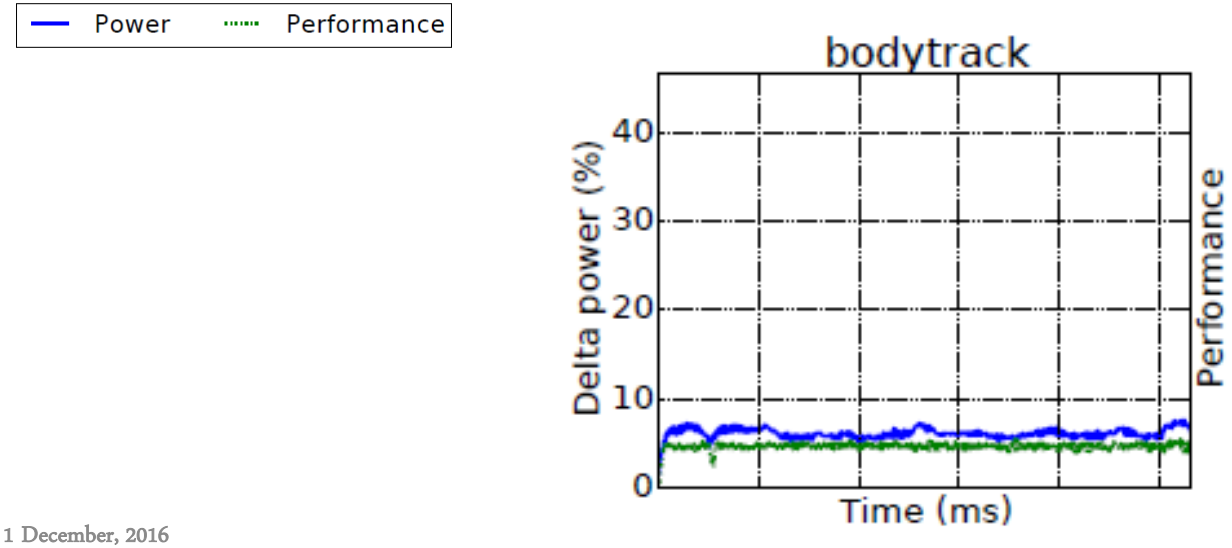
- Good accuracy (+/- 3%) can be maintained down to approximately 2A current draw.

Methodology

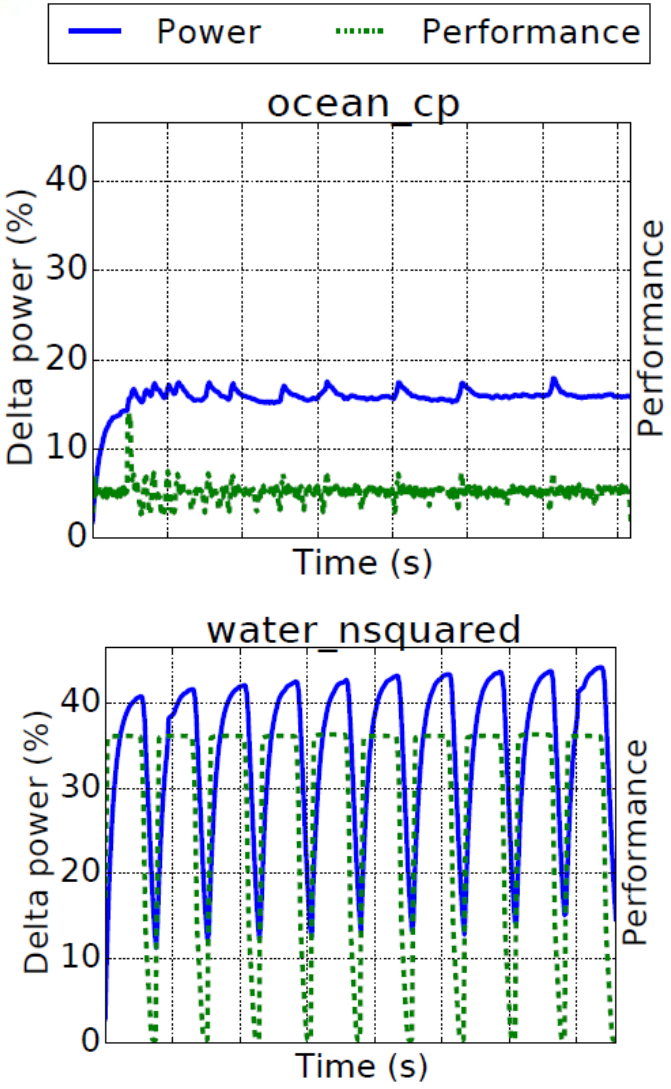
- Benchmarks: PARSEC and SPLASH-2
- Power measurement running independently of benchmarks.
- Daemon running on a separate machine queries via SSH over Ethernet.
- Invokes benchmark and starts the energy monitor.
- We capture: Delta power, performance, data movement e.g. percentage of data accesses that require a cache line to be moved.
- Focus of measurement is on the parallel region of the benchmark runs.

Power Trace Categorisation

- Using Manhattan distance Hierarchical clustering we obtained the following 3 power consumption clustering characteristics.
- Constant power consumption characteristic
  - Similar number of instructions per thread
  - Work in the benchmarks separated evenly
  - All threads at any time have enough data to keep working



Power Trace Categorisation

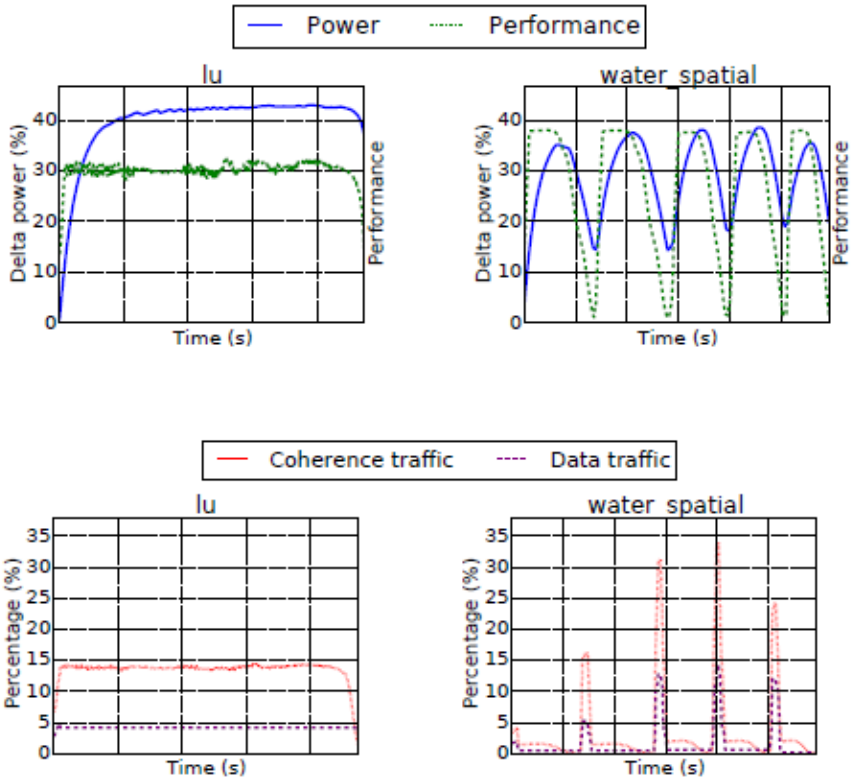


- Small amplitude – a small number of threads have slightly more work to do
- Large amplitude – a small number of threads have significantly more work to do .

Memory Traffic

Dependency between CCPI bandwidth Usage and Power Consumption

- CCPI link keeps L2 caches coherent between the 2 THunderX SoC
- Observations
  - No functional dependency between keeping caches coherent and power
  - Memory traffic impact on the magnitude of power consumption
- Impact on Power Magnitude
  - Accesses (especially power hungry stores) to L2 or main memory in absolute terms
  - Floating point operations more expensive than integer operations



Summary

- Built infrastructure to measure
  - Instantaneous power consumption
  - System performance using hardware performance counters
- Identified patterns in power traces and categorised them into groups
- No correlation between the amount of traffic across the CCPI link when used to keep the caches coherent and power consumed.
- Showed qualitatively that to estimate the magnitude of the power consumption it is not only sufficient to look at performance but also memory access and traffic must be considered.

Guided Scheduling Using Accurate Per-Core Performance Hardware Counters.

Optimising for Energy

Optimising for Energy

- Data movement requires powering bus lines
- This costs energy and transfer times
- This is particularly important on NUMA or multi – and/or many core, multi package, multi-level cache systems

Multithreaded Application Communications

- Even on shared memory systems communication between threads cannot be avoided
- These communications are primarily via cache memories

New Performance Counters

New Performance Counters

- Thread level communication can be reasonably be modelled by cache to cache (L1 to L1) data movements
- By modifying the cache coherency protocol (double snooping) we can gain this information.

Tracing Communication

- Need to trace two state changes:
  - When a cache line changes state from exclusive (E) to (S) shared, and
  - When a cache line changes state from modified (M) to shared (S)

	M	E	S	I
M	PR *or* PW		BR	BW
E	PW	PR	BR	BW
S		PW	PR *or* BR	BW
I		PR/$\overline{\text{S}}$ *or* PW	PR/S	BR *or* BW

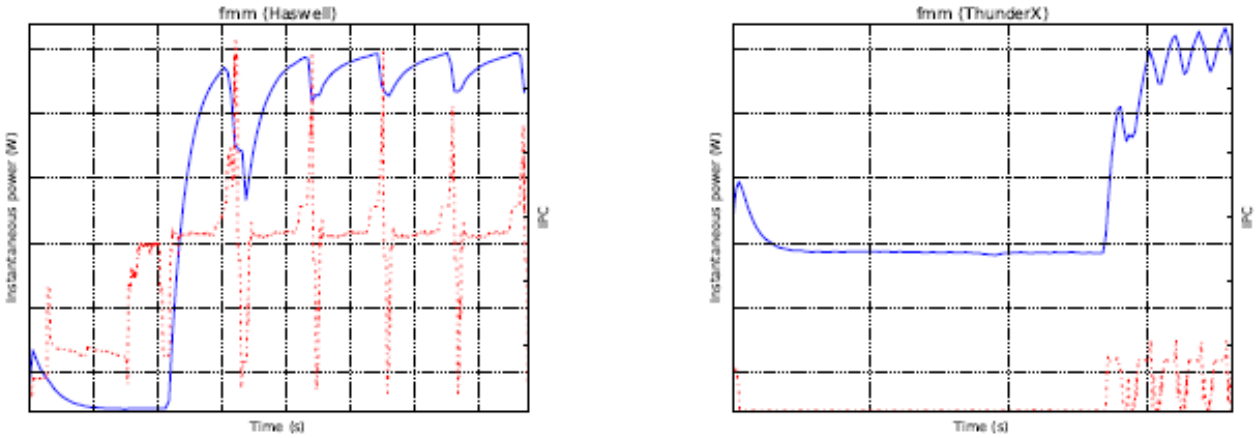
Evaluation Platform

- Very high-level, full system simulator codenamed PRAZOR
  - Capable of rendering simultaneously complete performance and energy estimates from real workloads with high level of accuracy.
  - Binary compatible with the real hardware meaning it can run the same operating system and use the same disk images as the real hardware board that is being simulated.
  - Target hardware (microarchitecture and peripherals) can be easily built by combining highly modular and extensible virtual prototyping blocks.

Simulator uses transaction level modelling from SystemC library

Future Work

Micro architectural Comparison and Modelling

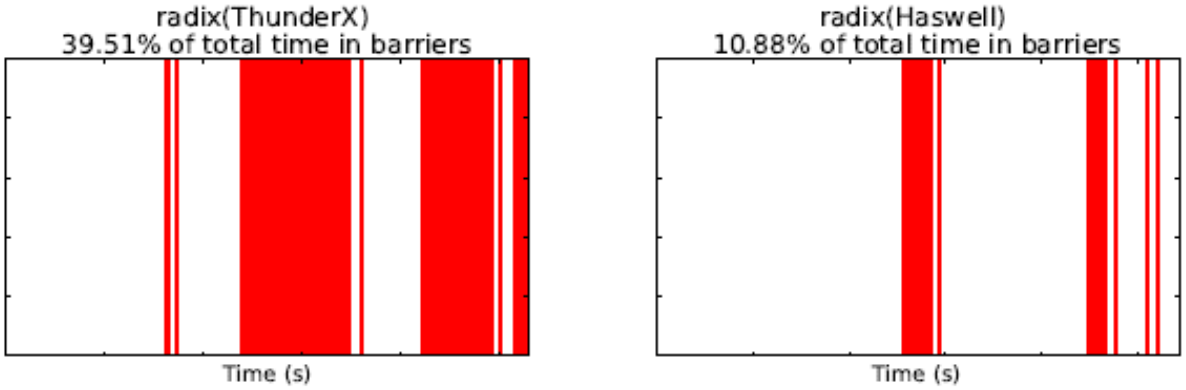


- Quantitative and qualitative comparison between different implementations of microarchitectures, such as x86 Haswell and ARM ThunderX:
  - what design decisions influence the shape of the power (blue) curve?
  - where does the difference in the shape of performance (red) curve come from?
- Cross-architectural Power and Performance mathematical models
  - estimate the **magnitude** of the curves
  - use data that is easy to measure but *correlate* to what to estimate

Future Work

Scheduling and Concurrency

- Energy aware scheduling:
  - exploiting periodicity of power and performance curves to predict when cores are underutilised,
  - use this information to *park* cores so that less energy is used.
- Concurrency



- ThunderX spent more time for synchronisation than Haswell,
- locking libraries *not* optimised for weak consistency model,
- need to quickly *delay* or *back-off* when failing to obtain lock.

Acknowledgments

- Milos Puzovic – The Hartree Centre, STFC.
- Jeyan Thiyagalingam – Electrical Engineering and Electronics, The University of Liverpool.
- David Greaves – Computer Laboratory, The University of Cambridge.
- Srilathe Manne – Cavium.
- Shay GalOn – Cavium.
- Makoto Ono – Lenovo.

Thank you. Any Questions?

Neil.morgan@stfc.ac.uk

http://community.hartree.stfc.ac.uk
http://www.stfc.ac.uk/hartree

hartree@stfc.ac.uk

Hartree Centre
Science & Technology Facilities Council