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.”
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)

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*)

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

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

MODEL-BASED RTM: HETEROGENEITY

Heterogeneous Platforms

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

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

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

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

ENERGY RTM ON HPC SYSTEMS

- 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

MODEL-BASED RTM

MODEL-BASED RTM

Runtime Management

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.

![Comparison Graph](comparison-graph-data.csv)

**Training**: Small set of 20 workloads

**Testing**: Full set of 60 workloads

---


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]
