UiT

THE ARCTIC UNIVERSITY OF NORWAY

Power Models Supporting Energy-Efficient Co-Design on Ultra-Low Power Embedded Systems

Vi Ngoc-Nha Tran[1], Brendan Barry[2], Phuong Ha[1]

[1] Department of Computer Science, UiT The Arctic University of Norway

[2] Movidius Ltd., Ireland

SAMOS XVI, Samos, Greece (July 18-21, 2016)

UNIVERSITY OF TROMSØ

What are energy/power models for?

- Predict how much energy a computing system consumes
- Provide the understanding how a computing system consumes energy/power
- Give hints on designing and implementing algorithms/ platforms to improve energy efficiency

Why do we need new power models for ULP systems?

- Ultra-low power (ULP) embedded systems
  - Have Different architectures from the high-performance systems (e.g., CPU and GPU)
  - Have low energy per instruction and require more accurate fine-grained modelling approaches
  - Have low static power, do not support DVFS but can turn on/off individual core

However

There is no available power model that provides insights into how a given application running on an ULP embedded system consumes power

Contributions

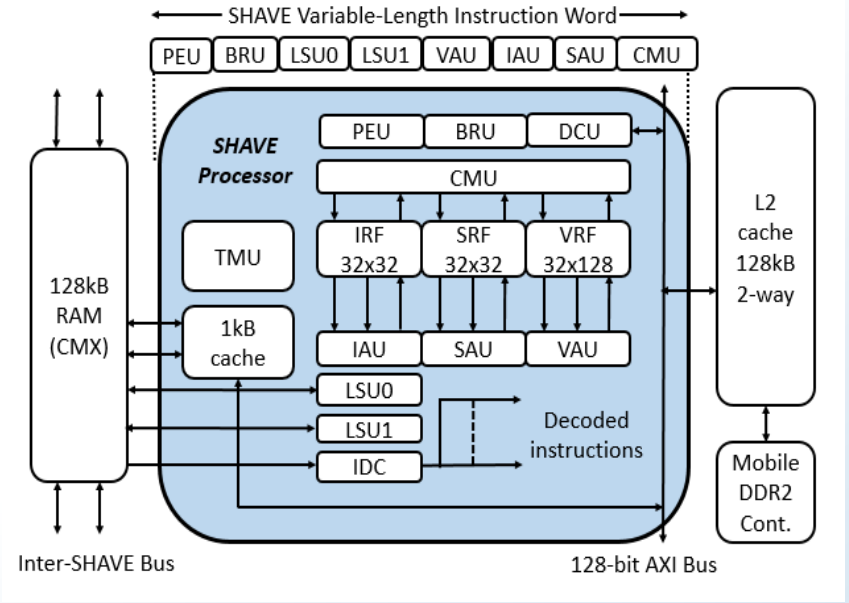
We propose RTHpower models that:

- Support co-design on ULP systems by considering:
  - platform properties,
  - application properties (e.g., operational intensity and scalability)
  - execution settings (e.g., the number of cores executing a given application)
- Built and validated with
  - Movidius platform
  - Application kernels (i.e., Matmul, SpMV and BFS)
  - Accuracy 8.5% for micro-benchmarks and 12% for application kernels
- Support predicting race-to-halt (RTH) effect for a given application

Outline

- Motivations
- Contributions
- Movidius Myriad – an ULP embedded system
- RTHpower models
- Model validation
- Predicting RTH effect
- Conclusion

Movidius Myriad – an ULP Embedded System



- Different architecture from the general-purpose architectures
- Energy per instruction as low as a few pJ
- Not support DVFS features, power on/off individual cores
- Difficult to program

RTHpower Models

- RTHpower model for Myriad platform
- RTHpower model for applications
  - Longer computation time than data transfer time
  - Shorter computation time than data transfer time

RTHpower Model for Myriad Platform

$$P^{units} = P^{sta} + n \times \left(P^{act} + \sum_i P_i^{dyn}(op) \right)$$

$$P^{sta} = 62.125 \text{ mW}$$

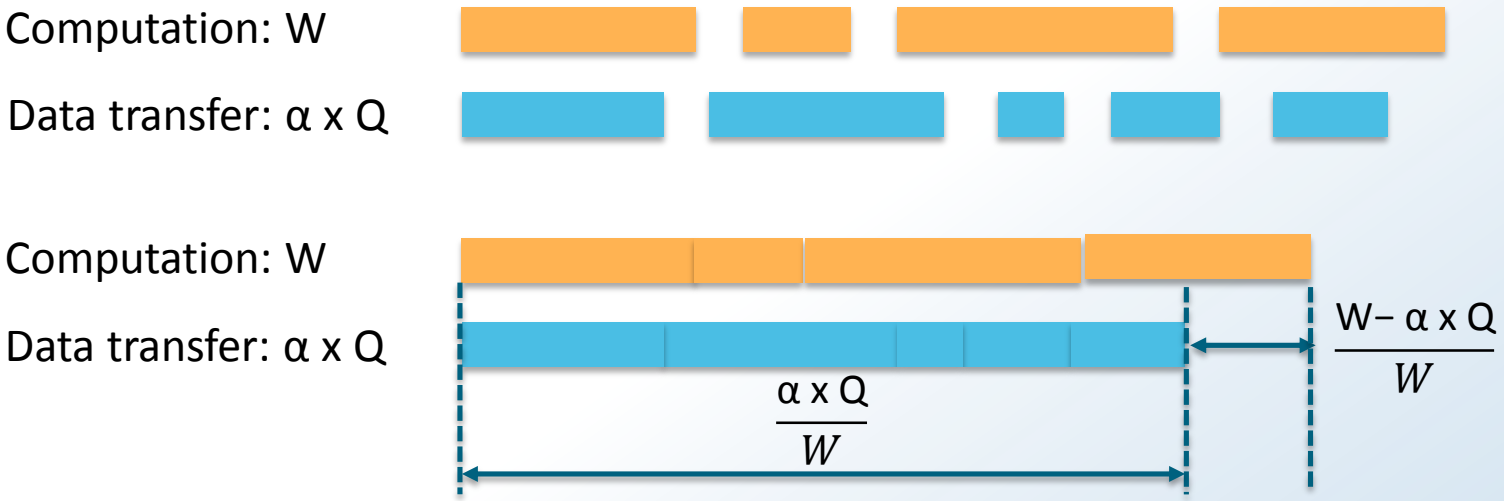
$$P^{act} = 30 \text{ mW}$$

Operation	Description	P^{dyn} (mW)
SAUXOR	Perform bitwise exclusive-OR on scalar	15
SAUMUL	Perform scalar multiplication	18
VAUXOR	Perform bitwise exclusive-OR on vector	35.6
VAUMUL	Perform vector multiplication	52.6
IAUXOR	Perform bitwise exclusive-OR on integer	15
IAUMUL	Perform integer multiplication	21
CMUCPSS	Copy scalar to scalar	20
CMUCPIVR	Copy integer to vector	13
LSULOAD	Load from a memory address to a register	28
LSUSTORE	Store from a register to a memory address	37

RTHpower Power Model for Applications

- When computation time is **longer** than data transfer time

α: time ratio of data transfer to computation

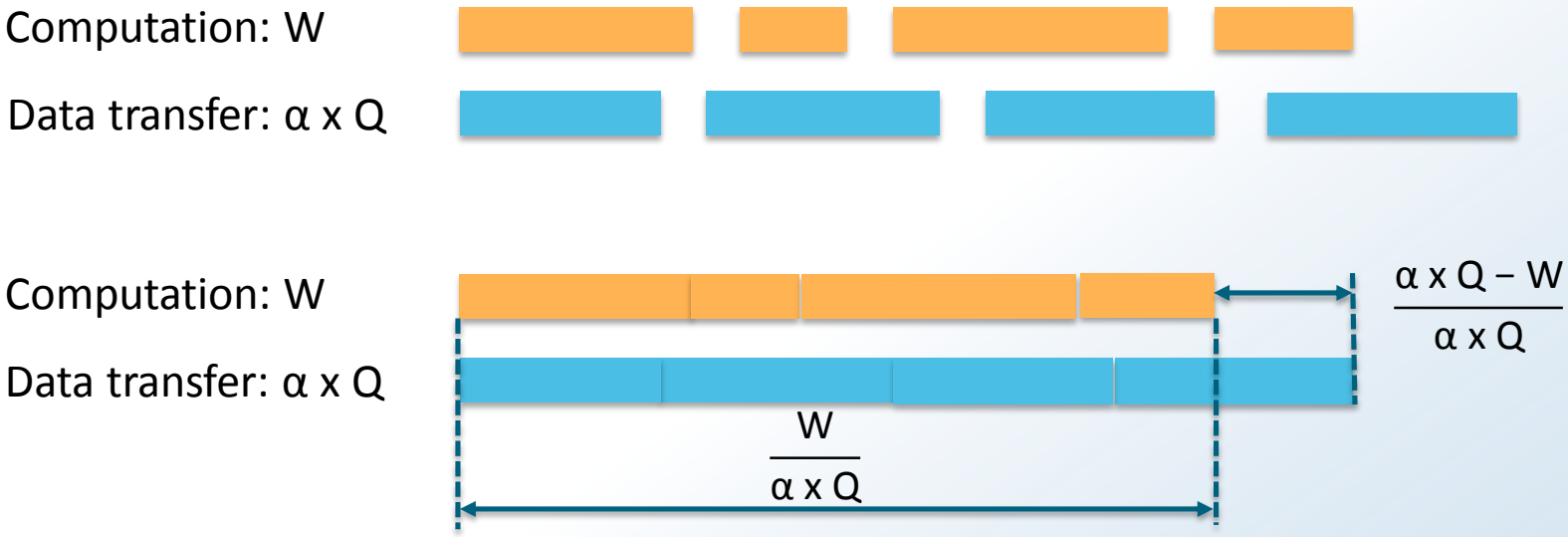


- The power model when computation time is longer

$$P = P^{comp||data} \times (\frac{\alpha \times Q}{W}) + P^{comp} \times (\frac{W - \alpha \times Q}{W})$$

RTHpower Power Model for Applications

- When computation time is **shorter** than data transfer time



- The power model when computation time is shorter

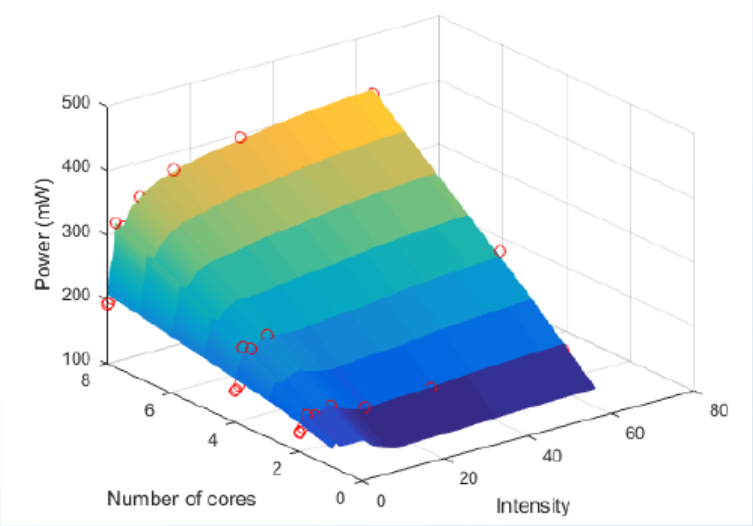
$$P = P^{comp||data} \times (\frac{W}{\alpha \times Q}) + P^{data} \times (\frac{\alpha \times Q - W}{\alpha \times Q})$$

RTHpower Power Model for Applications

- With operational intensity $I = \frac{W}{Q}$ [1] , the models are derived as

$$P = P^{comp||data} \times (\frac{I}{\alpha}) + P^{data} \times (\frac{\alpha - I}{\alpha})$$

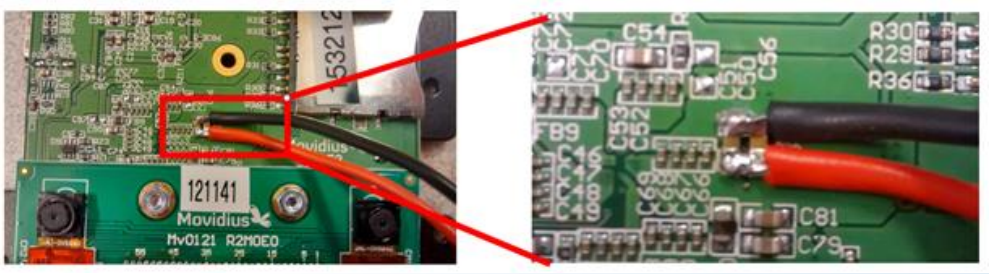
$$P = P^{comp||data} \times (\frac{\alpha}{I}) + P^{comp} \times (\frac{I - \alpha}{I})$$



[1] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (April 2009), 65-76.

Experimental Study

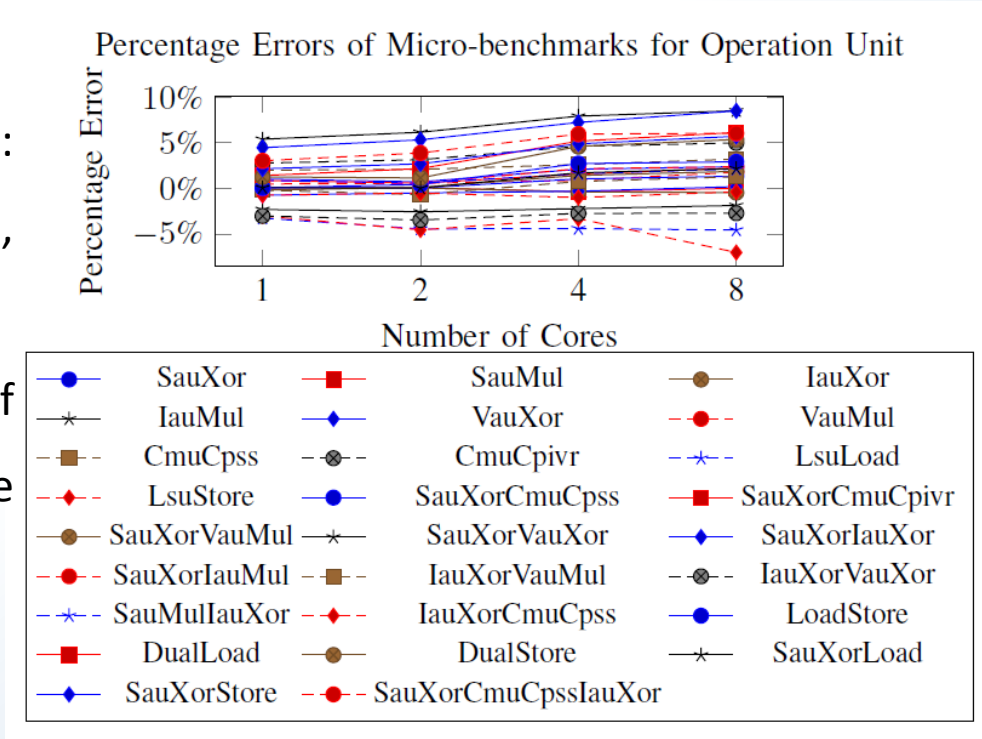
- Design 35 micro-benchmarks (i.e., operation-unit suite (26) and intensity-based suite (9))
- Use external multi-meters to measure the power consumption of the Movidius Myriad platform



- Train the model with measured power data from running micro-benchmarks with 1, 2 cores and validate with data from 4, 8 cores

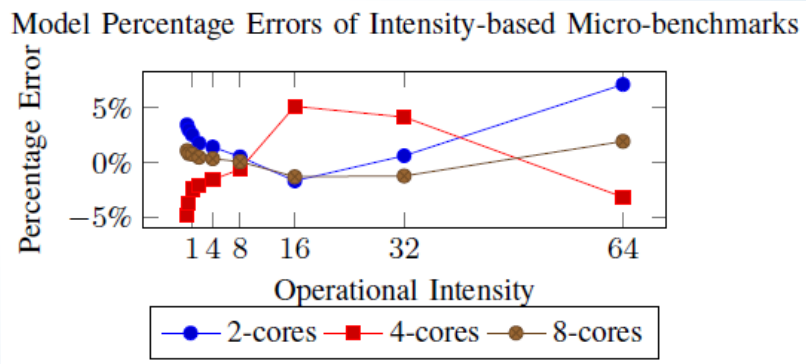
RTHpower Model for Myriad Platform

- Operation-unit micro-benchmarks: execute only operation units (e.g., SAU, IAU, VAU)
- The absolute percentage errors of unit-suite micro-benchmarks are at most 8.5%



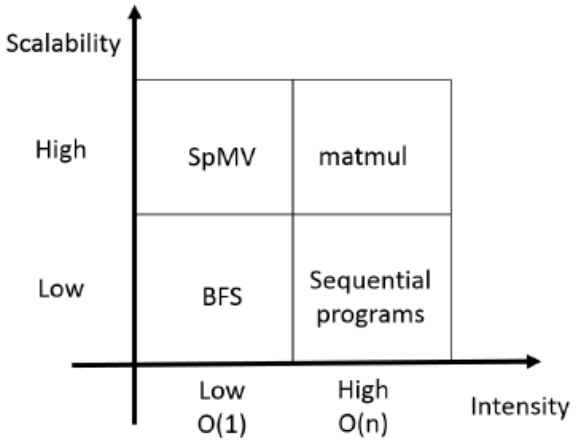
RTHpower Model for Applications – Micro-benchmarks

- 9 Intensity-based micro-benchmarks: execute both arithmetic units (e.g., SAU) and data transfer units (e.g., LSU)
  - Operational Intensity: operations per bye [1]
  - The ratio of the number of SAU isntructions to the number of LSU instructions define intensity value
- The absolute percentage errors of model fitting for intensity-suite are at most 7%

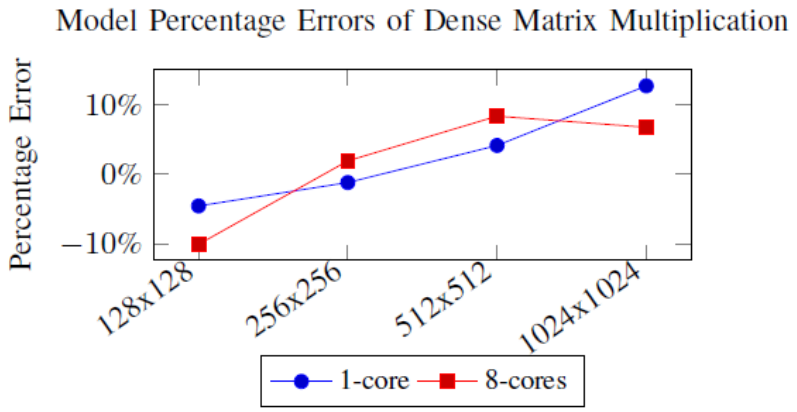


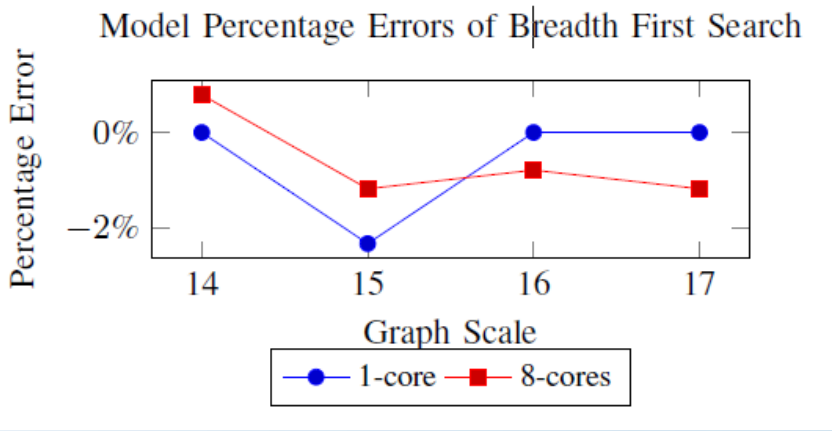
[1] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (April 2009), 65-76.

RTHpower Model for Applications - Application Benchmarks



Kernel	Error
SpMV	4%
Matmul	12%
BFS	3%



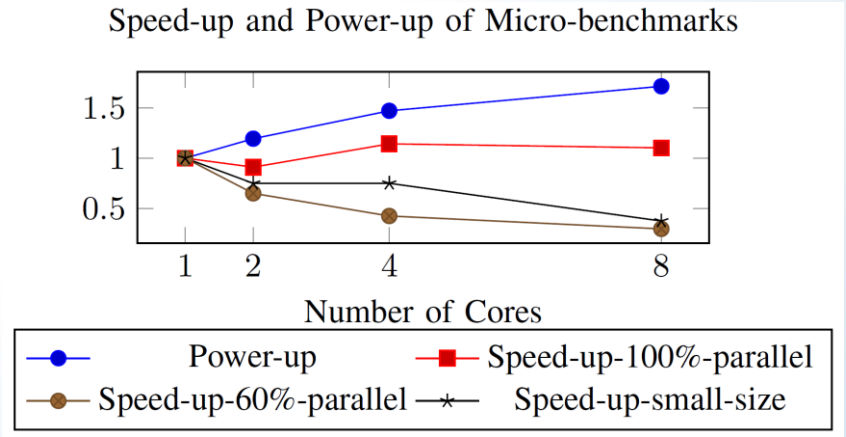


Outline

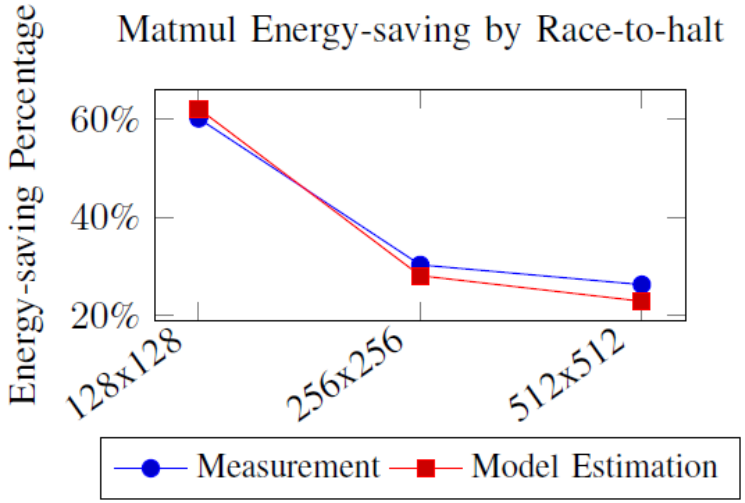
- Motivations
- Contributions
- Movidius Myriad – an ULP embedded system
- RTHpower models
- Model validation
- Predicting RTH effect
- Conclusion

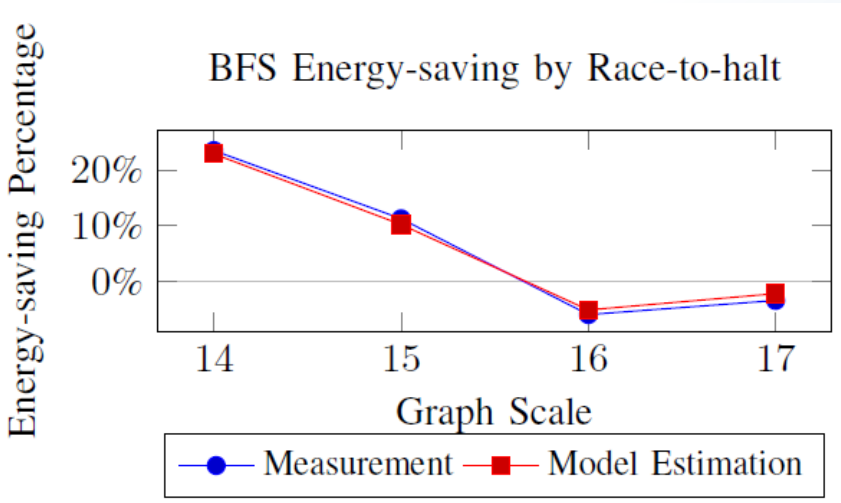
Predicting RTH Effect – Micro-benchmarks

- Three micro-benchmarks with intensity I=0.25
  - 100% parallel: loop 1000000 times for 1 core and loop 125000 times for 8 cores
  - 60% parallel: loop 1000000 times for 1 core and 475000 times for 8 cores
  - Small-size: high overhead
- They have speed-up less than platform power-up
- RTH is not an energy-saving strategy for these micro-benchmarks



Predicting RTH Effect - Applications





Kernel	Energy-saving
SpMV	Up to 61% by using RTH
Matmul	Up to 59% by using RTH
BFS	Up to 23% by using RTH and 5% by not using RTH

Conclusion

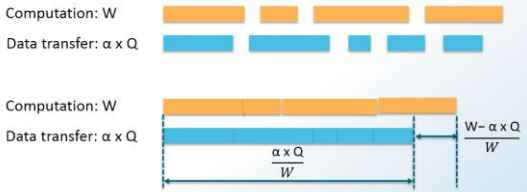
- RTHpower models provide insights into how an application consumes energy when executing on an ultra-low power (ULP) embedded system.
- RTHpower models support architecture-application co-design by considering platform, setting and application properties.
- Race-to-halt strategy is not always true on ULP systems and RTHpower models support predicting RTH effect for a given application.

Q&A

Thank you!

RTHpower Power Model for Applications

- When computation time is **longer** than data transfer time



- The power model when computation time is longer

$$P = P^{comp||data} \times (\frac{\alpha \times Q}{W}) + P^{comp} \times (\frac{W - \alpha \times Q}{W})$$

RTHpower Power Model for Applications

- If computation time is longer than data transfer time

$$P = P^{comp||data} \times (\frac{\alpha \times Q}{W}) + P^{comp} \times (\frac{W - \alpha \times Q}{W})$$

- If computation time is shorter than data transfer time

$$P = P^{comp||data} \times (\frac{W}{\alpha \times Q}) + P^{data} \times (\frac{\alpha \times Q - W}{\alpha \times Q})$$

- With $I = \frac{W}{Q}$ [1] , the models are derived as

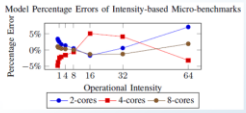
$$P = P^{comp||data} \times (\frac{I}{\alpha}) + P^{data} \times (\frac{\alpha - I}{\alpha})$$

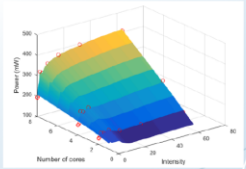
$$P = P^{comp||data} \times (\frac{\alpha}{I}) + P^{comp} \times (\frac{I - \alpha}{I})$$

[1] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (April 2009), 65-76.

Model Validation - RTHpower Power Model for Applications

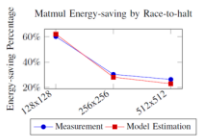
- Operational Intensity: operations per bye [1]
- Intensity-based microbenchmarks: execute both arithmetic units (e.g., SAU) and data transfer units (e.g., LSU)
- The ratio of the number of SAU isntructions to the number of LSU instructions define intensity value
- The absolute percentage errors of model fitting for intensity-suite are at most 7%

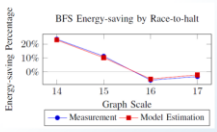




[1] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (April 2009), 65-76.

Predicting RTH Effect - Applications





Kernel	Energy-saving
SpMV	Up to 61% using RTH
Matmul	Up to 59% using RTH
BFS	Up to 23% using RTH and 5% by not using RTH

Contact information:
Vi Ngoc-Nha Tran
vi.tran@uit.no