Modeling Energy Optimization of Parallel Workloads on unreliable Hardware

Chhaya Trehan, Hans Vandierendonck, Georgios Karakonstantis, Dimitris S. Nikolopoulos

Queen's University of Belfast

January 19, 2016

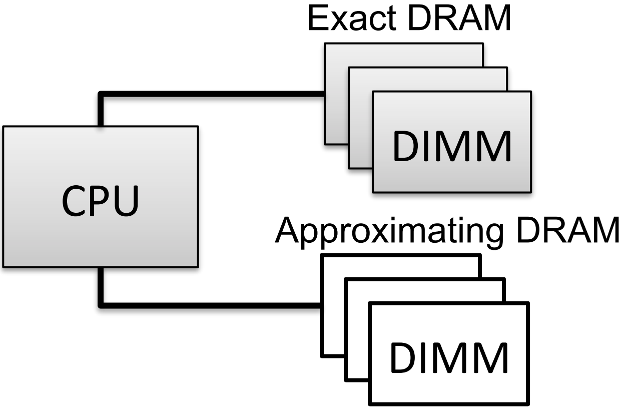
Introduction

Motivation

- Existing memory systems refresh for the worst case retention time to guarantee 100% accuracy.
- Do all applications need 100% accurancy?

Introduction

Hybrid Memories: A vision for the future



Energy optimization using global DVFS

Problem description

- Model the energy consumed by a parallel application represented by a task dependency graph on a multicore system .
- Determine the frequencies used during execution that minimize the overall energy consumption subject to a performance budget.

Insights from the Existing Model

The existing Analytical Model by Gerard et al. [1] has following key points:

- Models power consumption as a convex function of frequency.
- Uses Global DVFS for optimizing energy consumption.
- A single frequency throughout the entire duration of a parallel application with dependencies is not optimal.
- Varying the frequency according to the changes in the amount of parallelism (number of active cores) leads to energy savings.
- Does not account for the energy and time consumed on memory accesses.

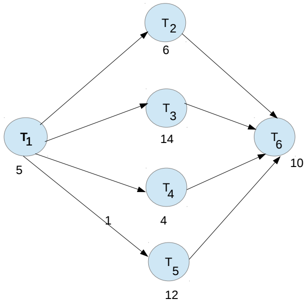
Our Model: An extension of [1]

The key points of our model are:

- Accounts for the CPU energy consumed during memory accesses.
- The frequency selection is not solely a function of amount of parallelism, takes into account other factors such as memory to CPU workload of tasks, average memory access delay.
- Can be easily extended to a mix of exact and approximate memory accesses.

Application and computing platform

An example Application representation



Node labels (below each node): compute workload

Problem Formulation

Energy and Performance

$$p_{CoreDynamic} = c_1 f^{\alpha} \tag{1}$$

$$p_{CoreStatic} = c_2 f + c_3 \tag{2}$$

Power when m cores are active:

$$p_m = m \times P_{CoreDynamic} + P_{CoreStatic} \tag{3}$$

Energy per CPU cycle ($\frac{p_m}{f}$):

$$\overline{p}_m = m \times c_1 f^{\alpha-1} + c_2 + \frac{c_3}{f} \tag{4}$$

Problem Formulation

Energy and Performance in terms of parallelism

Energy per CPU cycle when m cores running at a frequency f_m:

$$\overline{p}_m = m \times c_1 f_m^{\alpha-1} + c_2 + \frac{c_3}{f_m} \quad (5)$$

For a given schedule, parallelism is defined as vector $[w_1, w_2, \ldots w_m \ldots w_M]$ where w_m is the total number of CPU cycles for which exactly m cores are active.

Total Energy consumption in terms of parallelism is thus:

$$E(f_1, f_2, \ldots\ldots f_m) = \sum_{m=1}^{M} [\overline{p}_m(f_m) w_m] \quad (6)$$

Problem Formulation

Energy and Performance in terms of parallelism

Total Energy consumption in terms of parallelism is:

$$E(f_1, f_2, ...f_m) = \sum_{m=1}^{M} [\overline{p}_m(f_m) w_m] \tag{7}$$

Time to completion in terms of parallelism:

$$t_{completion} = \sum_{m=1}^{M} [\frac{w_m}{f_m}] \tag{8}$$

Problem Formulation

Accounting for memory accesses; reliable and unreliable

- d: the application wide data to CPU workload ratio.
- r: the application wide reliability quotient, ratio of reliable to unreliable memory accesses.
- t_{rel}: the access time of reliable portion of memory.
- t_{unrel}: the access time of unreliable portion of memory.

Effect of memory accesses on parallelism

- A given amount of parallelism w_m has $w_m \times d$ memory accesses.
- Out of these $w_m \times d$ memory accesses, $w_m \times d \times r$ and reliable and $w_m \times d \times (1-r)$ are unreliable.
- The extra cycles expended by CPU for a given parallelism w_m while waiting for memroy acceses:
  $\{w_m \times d \times r \times t_{rel} + w_m \times d \times (1-r) \times t_{unrel}\} f_m$
- Parallelism w_m can be replaced by
  $w_m + (w_m \times d \times r \times t_{rel}) f_m + (w_m \times d \times (1-r) \times t_{unrel}) f_m$

Problem Formulation

Optimization Problem The energy optimization problem after accounting for CPU cycles expended during memory accesses:

$$\underset{f_1, f_2, \ldots\ldots f_m}{\text{minimize}} \quad \sum_{m=1}^{M} [\overline{p}_m(f_m)\{w_m\{1 + d(rt_{rel} + (1-r)t_{unrel})f_m)\}\}]$$

$$\text{subject to} \quad \sum_{m=1}^{M} [\frac{w_m\{1 + d(rt_{rel} + (1-r)t_{unrel})f_m)}{f_m}] \leq t_{budget}$$

Frequency selection and parallelism

Our results compared to [1]

Let f_m and f_n be optimal frequencies for $E_{(n,m)}$

The relation between f_m and f_n is:

- $\sqrt[\alpha]{m} f_m = \sqrt[\alpha]{n} f_n$, not accounting for memory accesses
- $\sqrt[\alpha]{m(1 + dt_a \alpha f_m)} f_m = \sqrt[\alpha]{n(1 + dt_a \alpha f_n)} f_n$, accounting for memory accesses

Questions?

Please send an email to the authors: Email: {c.trehan, h.vandierendonck, g.karakonstantis, d.nikolopoulos}@qub.ac.uk

Johann L. Hurink Marco E.T. Gerards and Jan Kuper.
On the interplay between global dvfs and scheduling tasks with precedence constraints.
IEEE TRANSACTIONS ON COMPUTERS, 64(06), 2015.