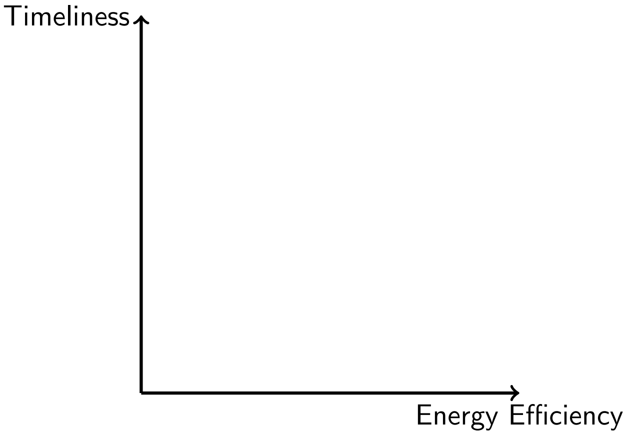
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

Anne Farrell[1] Henry Hoffmann[1]

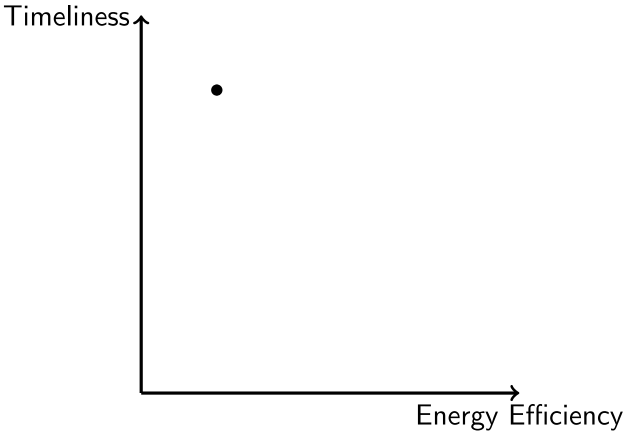
[1]Department of Computer Science
University of Chicago

USENIX ATC 2016

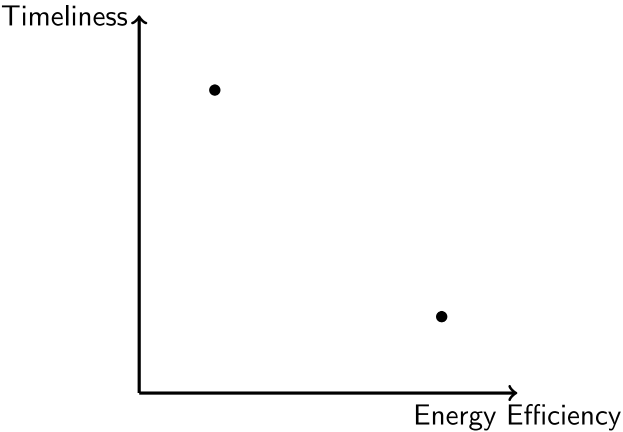
Problem



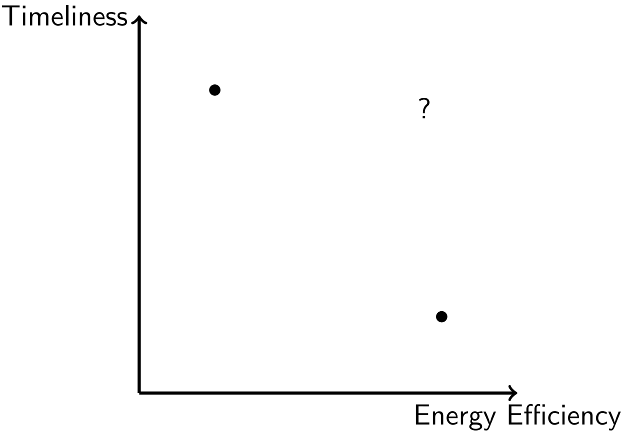
Problem



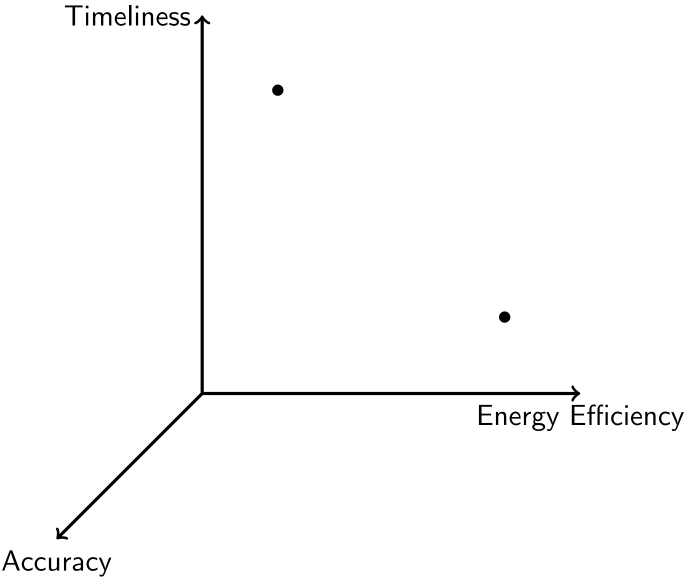
Problem



Problem



Problem



Problem

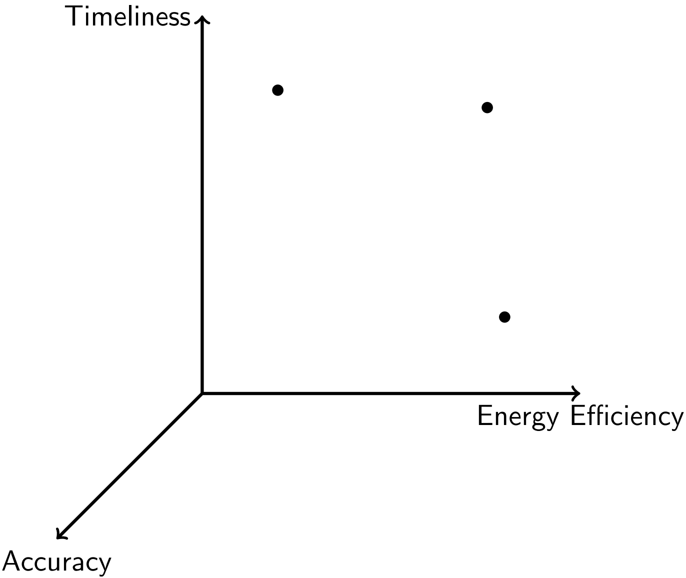
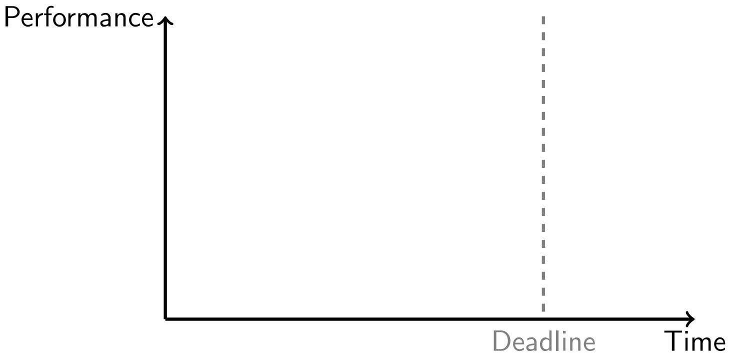
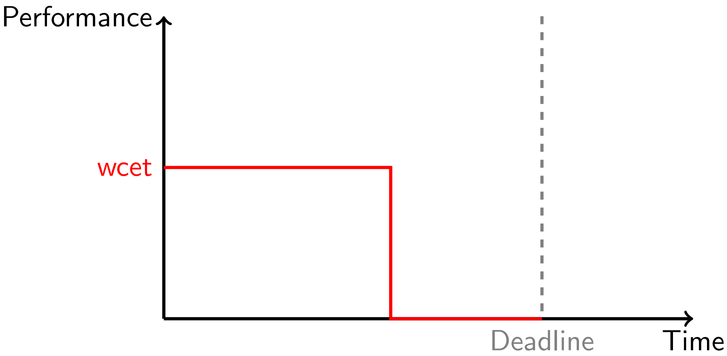


Table of Contents

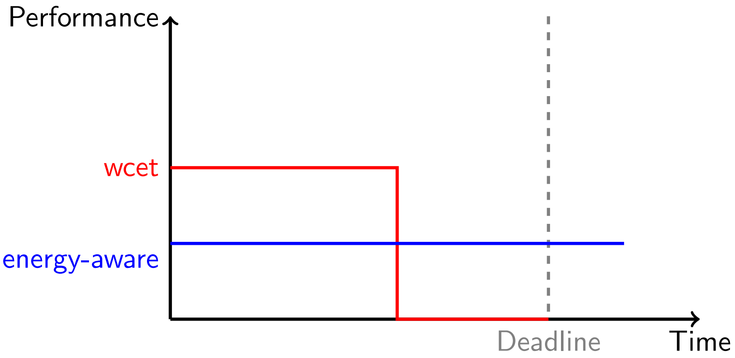
Techniques



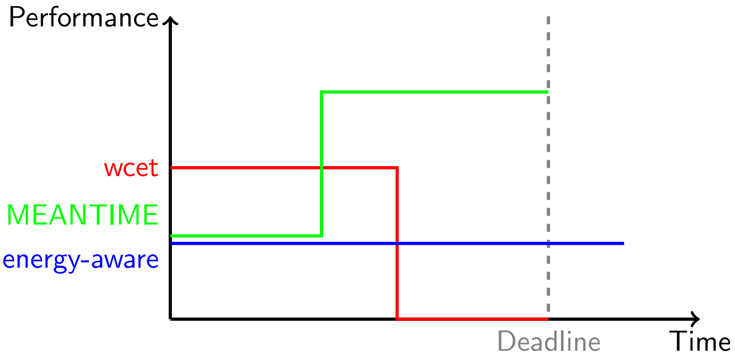
Techniques



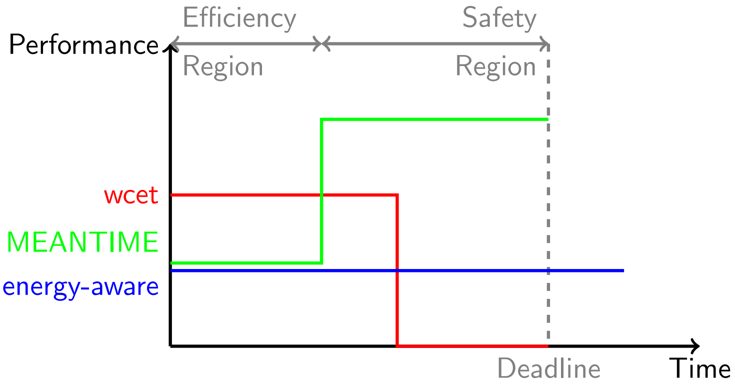
Techniques



Techniques



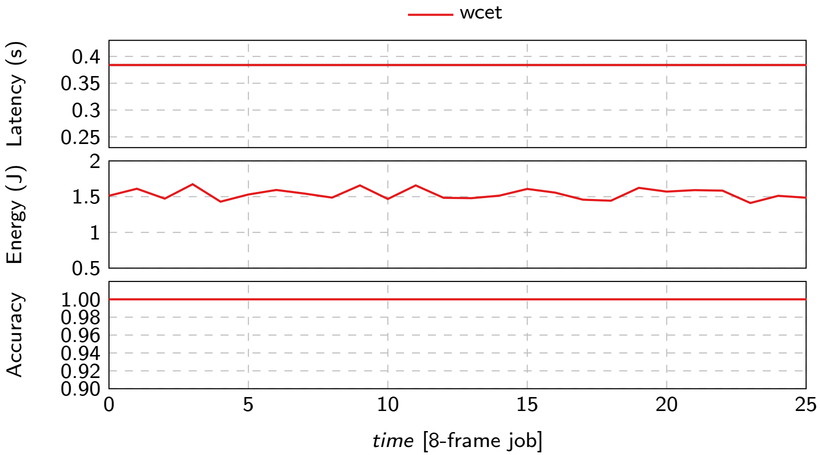
Techniques



Efficiency
Region
Safety
Region
Performance
wcet
MEANTIME
energy-aware
Deadline
Time

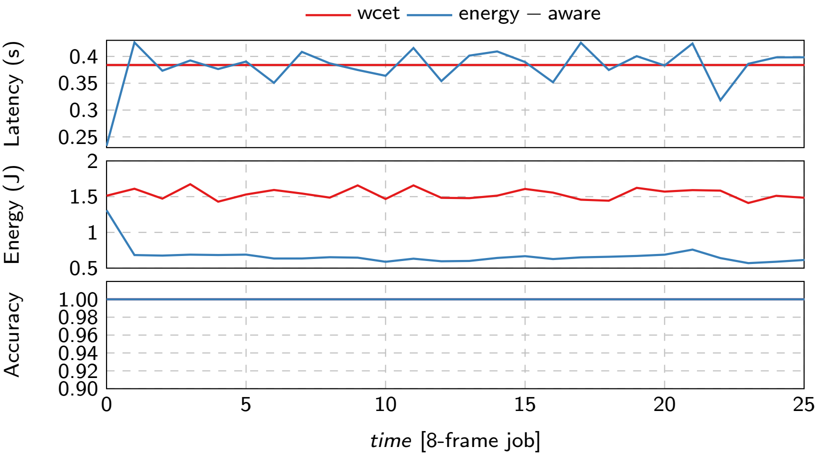
Table of Contents

Radar



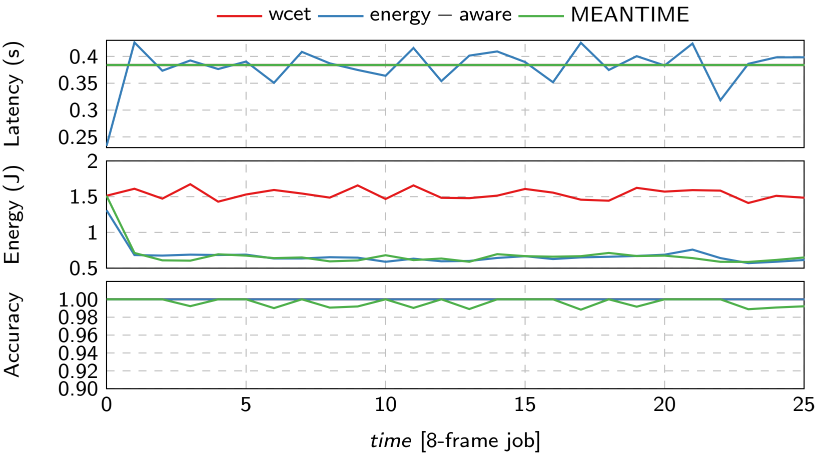
wcet
Latency (s)
0.4
0.35
0.3
0.25
Energy (J)
2
1.5
1
0.5
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0
5
10
15
20
25
time [8-frame job]

Radar



wcet
energy — aware
Latency (s)
0.4
0.35
0.3
0.25
Energy (J)
2
1.5
1
0.5
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0
5
10
15
20
25
time [8-frame job]

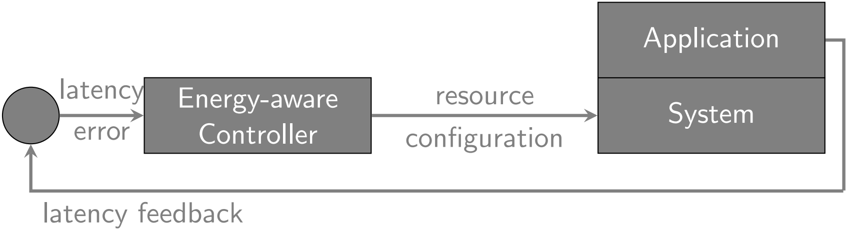
Radar



wcet
energy — aware
MEANTIME
Latency (s)
0.4
0.35
0.3
0.25
Energy (J)
2
1.5
1
0.5
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0
5
10
15
20
25
time [8-frame job]

Table of Contents

The MEANTIME Framework



The MEANTIME Framework

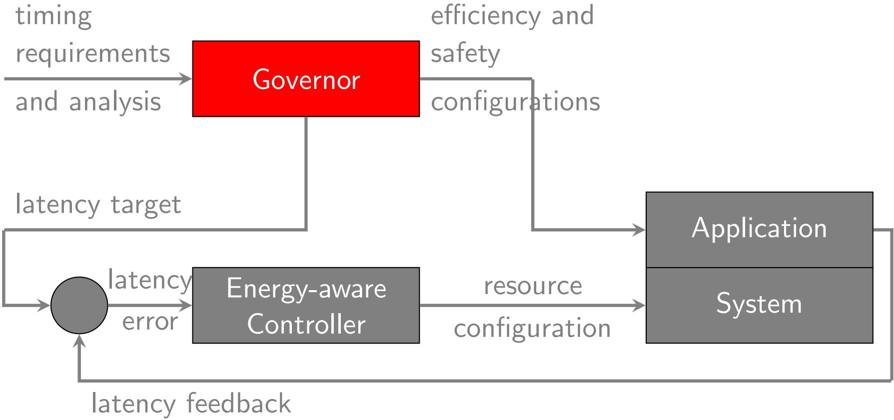
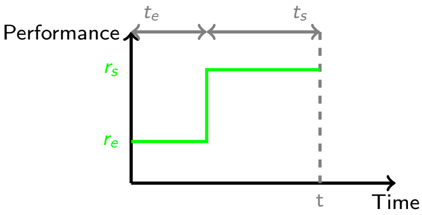


Table of Contents

Governor



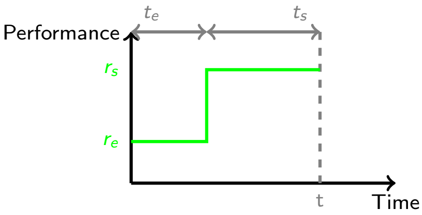
$$t_s \cdot r_s + t_e \cdot r_e = W$$

$$t_s + t_e + t_{switch} = t$$

$$t_s, t_e \geq 0$$

Variable	Meaning
W	job workload in worst case
t_{switch}	worst case time to switch app. and sys. config.
s_0	minimum speedup from approximation
t_{wc}	worst case latency

Governor



$$t_s \cdot r_s + t_e \cdot r_e = W$$

$$t_s + t_e + t_{switch} = t$$

$$t_s, t_e \geq 0$$

$$t_e = \frac{t_{wc} - s_0 \cdot (t - t_{switch})}{1 - s_0}$$

Variable	Meaning
W	job workload in worst case
t_{switch}	worst case time to switch app. and sys. config.
s_0	minimum speedup from approximation
t_{wc}	worst case latency

Table of Contents

Experimental Setup

Hardware:

- Platform
  - ODROID-XU3
- Processor
  - Samsung Exynos5 Octa
- big.LITTLE
  - 8 cores
    - 4 big A15 cores
    - 4 LITTLE A7 cores
- Configurations
  - 128

Experimental Setup

Hardware:

- Platform
  - ODROID-XU3
- Processor
  - Samsung Exynos5 Octa
- big.LITTLE
  - 8 cores
    - 4 big A15 cores
    - 4 LITTLE A7 cores
- Configurations
  - 128

Applications:

- x264
- bodytrack
- swaptions
- ferret
- streamcluster
- radar

Approaches Evaluated

- wcet
  - reserve resources for worst case and then sleep
  - hard guarantees

Approaches Evaluated

- wcet
  - reserve resources for worst case and then sleep
  - hard guarantees
- energy-aware
  - use control techniques to allocate minimize energy
  - soft guarantees
  - [POET RTAS 2015], [Maggio et al IEEE CST 2013] [3, 4]

Approaches Evaluated

- wcet
  - reserve resources for worst case and then sleep
  - hard guarantees
- energy-aware
  - use control techniques to allocate minimize energy
  - soft guarantees
  - [POET RTAS 2015], [Maggio et al IEEE CST 2013] [3, 4]
- PowerDial
  - adjust application-level parameters to meet performance goals
  - soft guarantees
  - [PowerDial ASPLOS 2011] [2]

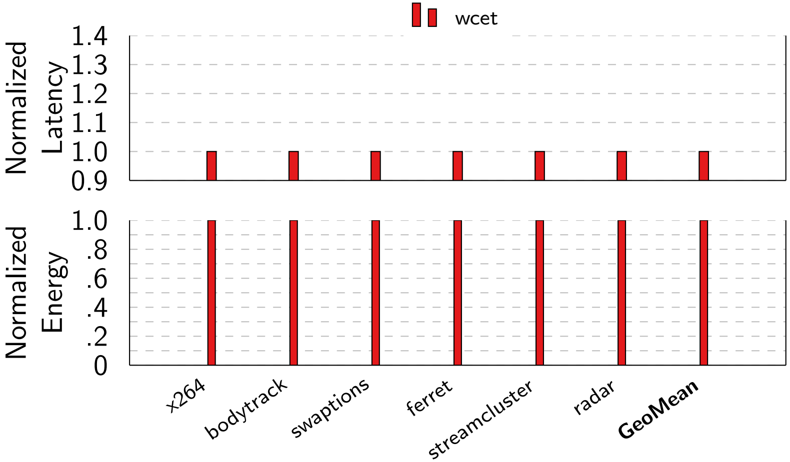
Approaches Evaluated

- wcet
  - reserve resources for worst case and then sleep
  - hard guarantees
- energy-aware
  - use control techniques to allocate minimize energy
  - soft guarantees
  - [POET RTAS 2015], [Maggio et al IEEE CST 2013] [3, 4]
- PowerDial
  - adjust application-level parameters to meet performance goals
  - soft guarantees
  - [PowerDial ASPLOS 2011] [2]
- cross
  - combine application accuracy and system resource management
  - simple extension for hard guarantees
  - [Flinn et al ACM Trans. Comp. Syst. 2004], [Yuan et al ACM SIGOPS OSR 2003] [1, 5]

Approaches Evaluated

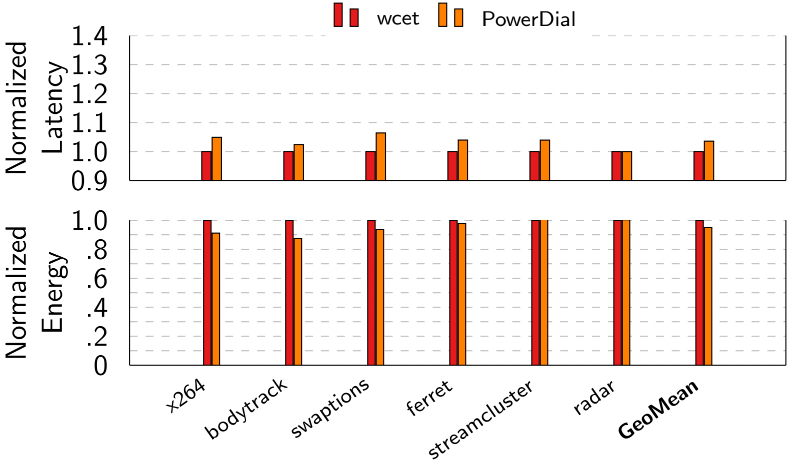
- wcet
  - reserve resources for worst case and then sleep
  - hard guarantees
- energy-aware
  - use control techniques to allocate minimize energy
  - soft guarantees
  - [POET RTAS 2015], [Maggio et al IEEE CST 2013] [3, 4]
- PowerDial
  - adjust application-level parameters to meet performance goals
  - soft guarantees
  - [PowerDial ASPLOS 2011] [2]
- cross
  - combine application accuracy and system resource management
  - simple extension for hard guarantees
  - [Flinn et al ACM Trans. Comp. Syst. 2004], [Yuan et al ACM SIGOPS OSR 2003] [1, 5]
- optimal
  - found by running every configuration

Results



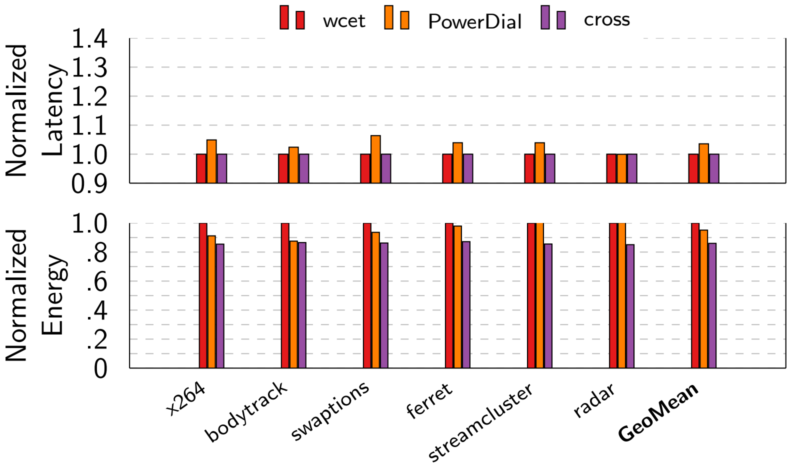
wcet
Normalized Latency
1.4
1.3
1.2
1.1
1.0
0.9
Normalized Energy
1.0
.8
.6
.4
.2
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Results



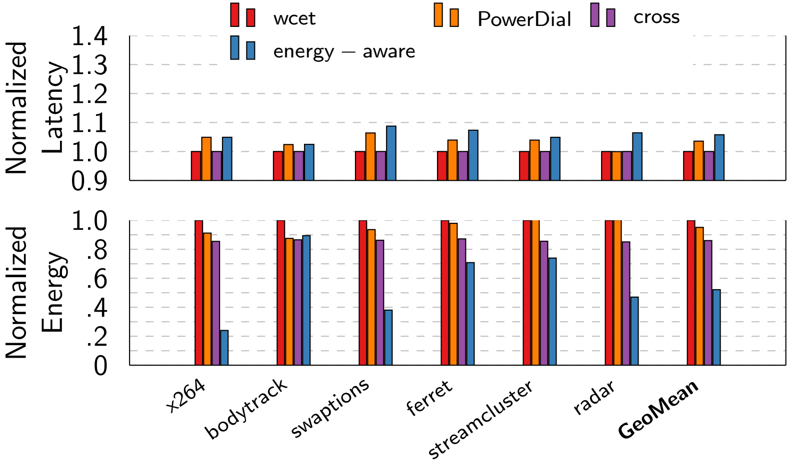
wcet
PowerDial
Normalized Latency
1.4
1.3
1.2
1.1
1.0
0.9
Normalized Energy
1.0
.8
.6
.4
.2
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Results



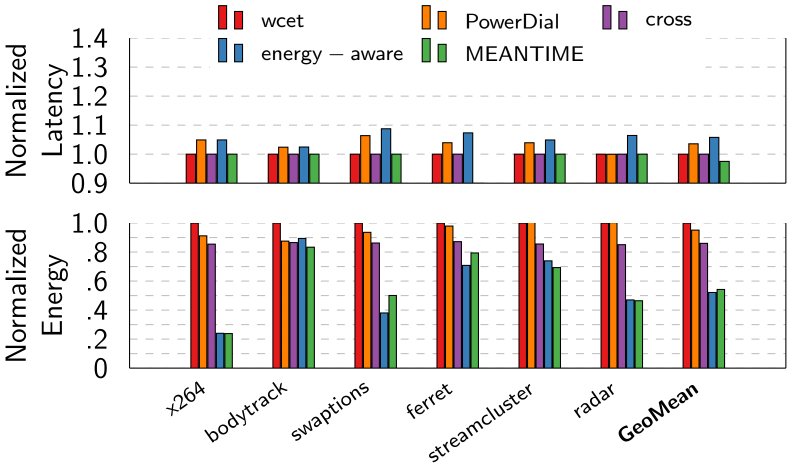
wcet
PowerDial
cross
Normalized Latency
1.4
1.3
1.2
1.1
1.0
0.9
Normalized Energy
1.0
.8
.6
.4
.2
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Results



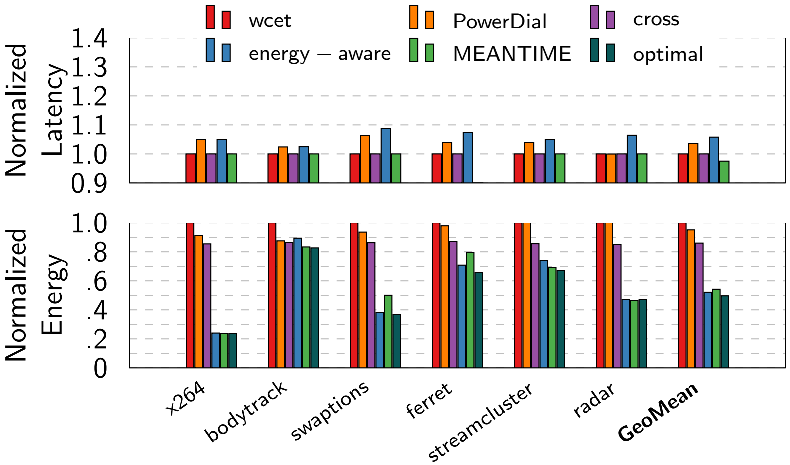
wcet
PowerDial
cross
energy − aware
Normalized Latency
1.4
1.3
1.2
1.1
1.0
0.9
Normalized Energy
1.0
.8
.6
.4
.2
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Results



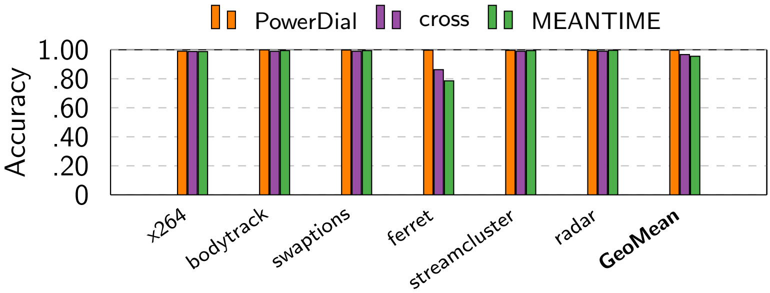
wcet
PowerDial
cross
energy — aware
MEANTIME
Normalized Latency
1.4
1.3
1.2
1.1
1.0
0.9
Normalized Energy
1.0
.8
.6
.4
.2
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Results



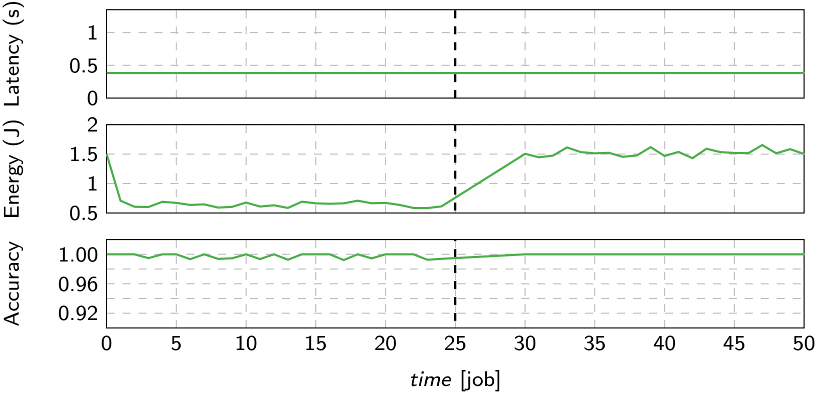
wcet
PowerDial
cross
energy − aware
MEANTIME
optimal
Normalized Latency
1.4
1.3
1.2
1.1
1.0
0.9
Normalized Energy
1.0
.8
.6
.4
.2
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Results

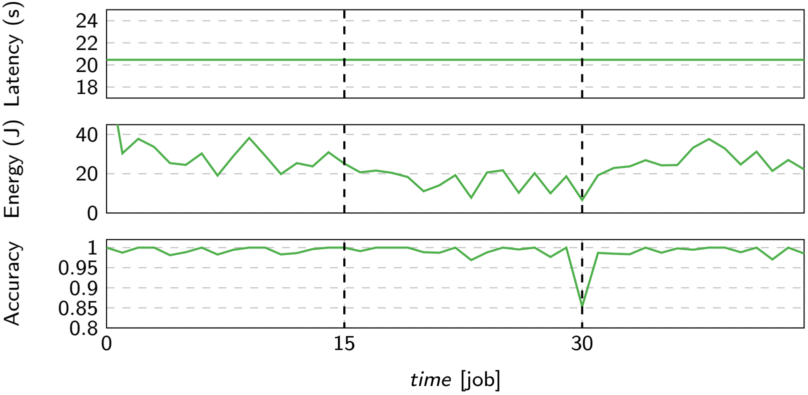


PowerDial
cross
MEANTIME
Accuracy
1.00
.80
.60
.40
.20
0
x264
bodytrack
swaptions
ferret
streamcluster
radar
GeoMean

Changing accuracy goals



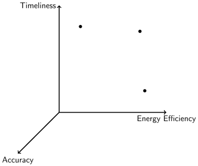
Phases



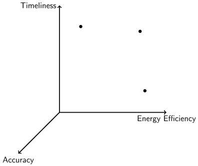
Latency (s)
24
22
20
18
Energy (J)
40
20
0
Accuracy
1
0.95
0.9
0.85
0.8
0
15
30
time [job]

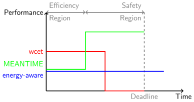
Table of Contents

MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

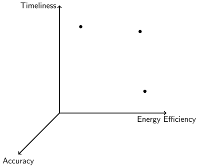


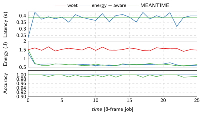
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

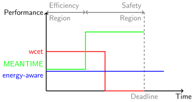




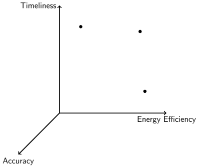
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

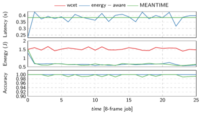


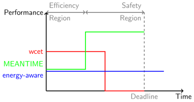


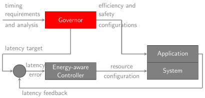


MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

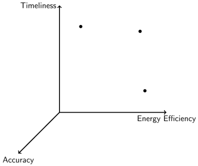


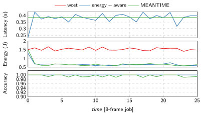


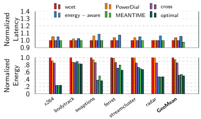


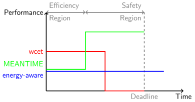


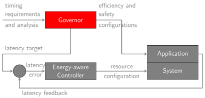
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing



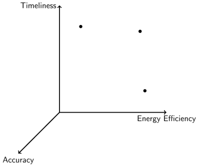


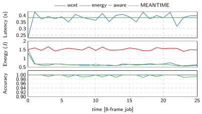


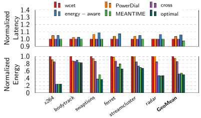


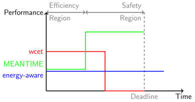


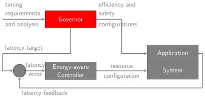
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

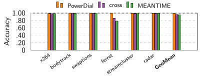




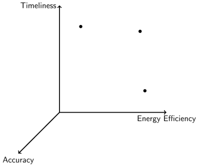


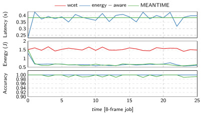


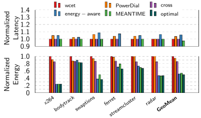


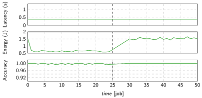


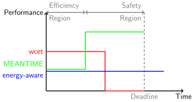
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

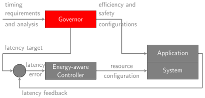


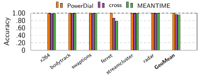




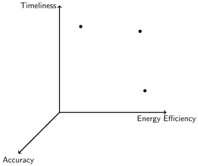


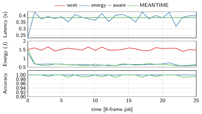


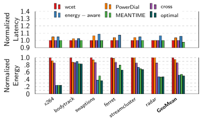


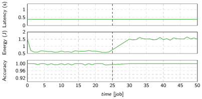


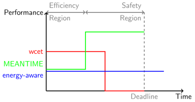
MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

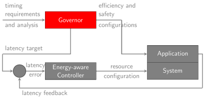


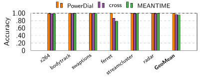


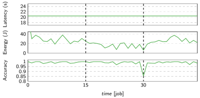












References I

J. Flinn and M. Satyanarayanan. "Managing battery lifetime with energy-aware adaptation". In: *ACM Trans. Comp. Syst.* 22.2 (May 2004).

H. Hoffmann, S. Sidiroglou, M. Carbin, S. Misailovic, A. Agarwal, and M. Rinard. "Dynamic Knobs for Responsive Power-Aware Computing". In: *ASPLOS*. 2011.

C. Imes, D. H. K. Kim, M. Maggio, and H. Hoffmann. "POET: A Portable Approach to Minimizing Energy Under Soft Real-time Constraints". In: *RTAS*. 2015.

M. Maggio, H. Hoffmann, M. D. Santambrogio, A. Agarwal, and A. Leva. "Power Optimization in Embedded Systems via Feedback Control of Resource Allocation". In: *IEEE Trans. on Control Systems Technology* 21.1 (2013).

References II

W. Yuan and K. Nahrstedt. "Energy-efficient soft real-time CPU scheduling for mobile multimedia systems". In: *ACM SIGOPS Operating Systems Review* 37.5 (2003), pp. 149–163.