COL862: Low Power Computing

MEANTIME: Achieving Both Minimal Energy and Timeliness with Approximate Computing

Authors: Anne Farrell and Henry Hoffmann, University of Chicago. 2016 USENIX Annual TechnicalConference.

– Presented By: Rajesh Kedia, Radhika D.

Experimental Setup

- ODROID-XU3 (from HardKernel)
  - ARM® big.LITTLE™ technology with HMP solution. Has Big Cortex-A15 2.0Ghz quad core and the LITTLE Cortex-A7 1.2 Ghz quad core CPUs
  - Provides embedded INA-231 power sensors provide power data for the big.LITTLE CPUs, DRAM and GPU.
  - Supports 5 configurable resources that can be used to modify performance and power tradeoff
  - Extremely low-power idle state (around 0.1 Watt) and max power consumption is just under 6 Watts

Experimental Setup

- ***taskset*** utility is used to manage processor core assignment
- ***cpufrequtils*** is used for managing DVFS settings

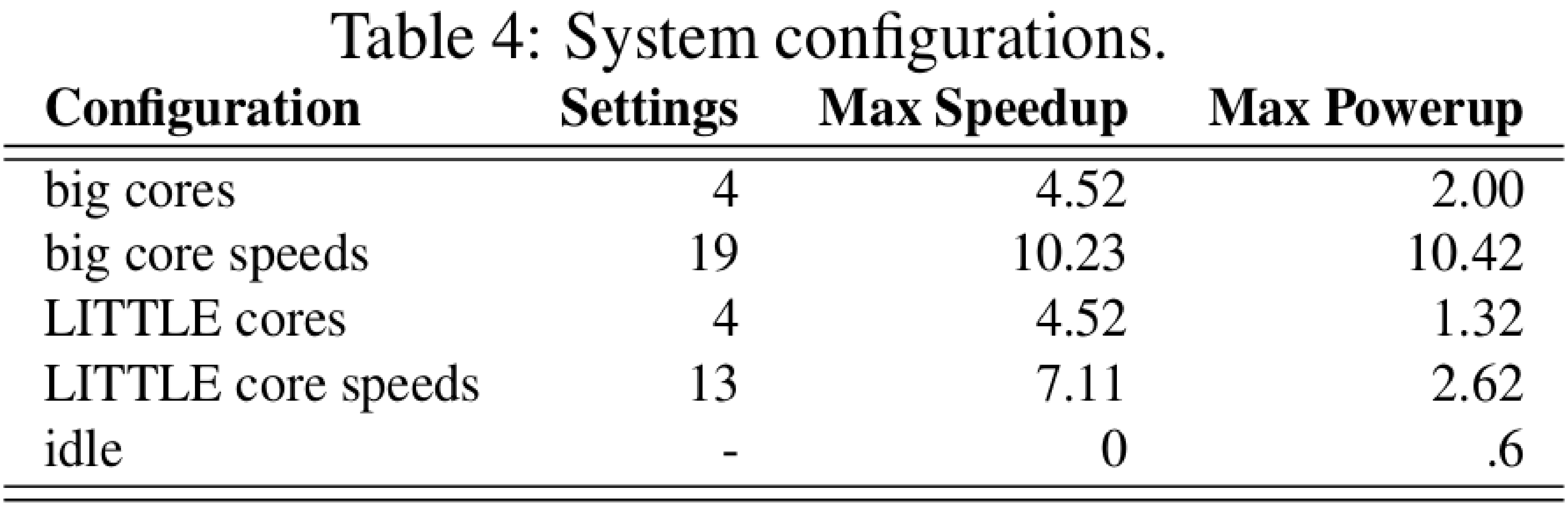


Table 4: System configurations.

Configuration	Settings	Max Speedup	Max Powerup
big cores	4	4.52	2.00
big core speeds	19	10.23	10.42
LITTLE cores	4	4.52	1.32
LITTLE core speeds	13	7.11	2.62
idle	-	0	.6

Benchmarks Used

- Used 6 benchmarks used
  - X264
  - Bodytrack
  - Swaptions
  - Ferret
  - Streamcluster
  - Radar
- PowerDial Framework is used to modify all six benchmarks to support dynamic approximation

Benchmarks

- X264:
  - video encoder compresses a raw input as per the H.264 standard.
  - It can decrease the frame latency at a cost of increased noise.
  - Accuracy is measured by the PSNR and encoded bitrate.

Benchmarks

- X264:
  - video encoder compresses a raw input as per the H.264 standard.
  - It can decrease the frame latency at a cost of increased noise.
  - Accuracy is measured by the PSNR and encoded bitrate.
- Bodytrack:
  - This application uses an annealed particle filter to track a human moving through a space.
  - The filter parameters trade the track’s quality and the frame latency.

Benchmarks

- X264:
  - video encoder compresses a raw input as per the H.264 standard.
  - It can decrease the frame latency at a cost of increased noise.
  - Accuracy is measured by the PSNR and encoded bitrate.
- Bodytrack:
  - This application uses an annealed particle filter to track a human moving through a space.
  - The filter parameters trade the track’s quality and the frame latency.
- Swaptions:
  - This is a financial analysis application that uses Monte Carlo simulation to price a portfolio of swaptions.
  - This application can reduce accuracy in the swaption price for decreased pricing latency.

Benchmarks

- Ferret:
  - This application performs similarity search for images
  - The search accuracy is evaluated using F-measure, the harmonic mean of precision and recall.

Benchmarks

- Ferret:
  - This application performs similarity search for images
  - The search accuracy is evaluated using F-measure, the harmonic mean of precision and recall.
- Streamcluster:
  - This application is an online approximation of the k-means clustering algorithm.
  - Tradeoff between cluster accuracy and latency is done by either changing number of iterations used in the approximation or by changing the distance metric used to assign a sample to a cluster

Benchmarks

- Ferret:
  - This application performs similarity search for images
  - The search accuracy is evaluated using F-measure, the harmonic mean of precision and recall.
- Streamcluster:
  - This application is an online approximation of the k-means clustering algorithm.
  - Tradeoff between cluster accuracy and latency is done by either changing number of iterations used in the approximation or by changing the distance metric used to assign a sample to a cluster
- Radar:
  - This application is the front-end of a radar signal processor and it turns raw antenna data into a target list.
  - Four Parameters this application uses to tradeoff SNR are:
    - First two change the strength of the low-pass filter.
    - Third changes the number of distinct directions the phased array antenna can “look.”
    - Fourth parameter changes the range resolution.

Benchmarks

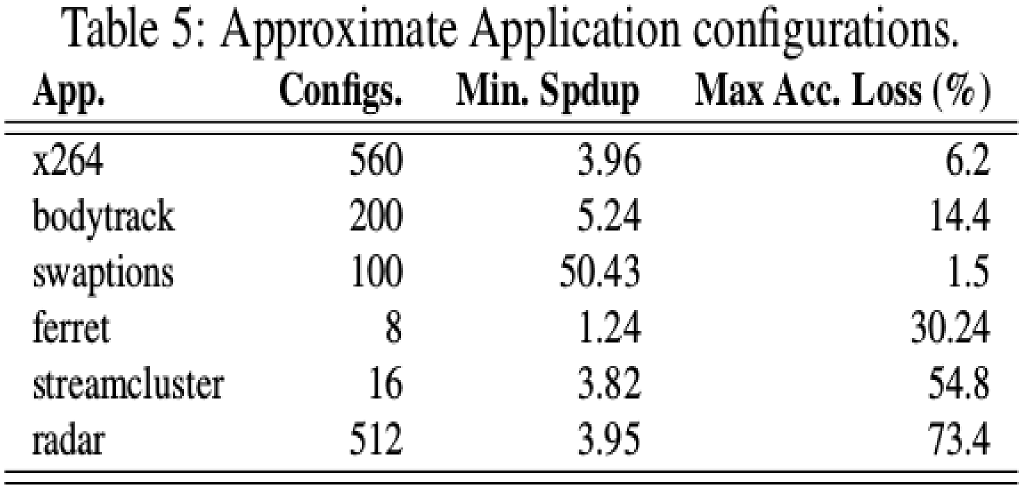


Table 5: Approximate Application configurations.

App.	Configs.	Min. Spdup	Max Acc. Loss (%)
x264	560	3.96	6.2
bodytrack	200	5.24	14.4
swaptions	100	50.43	1.5
ferret	8	1.24	30.24
streamcluster	16	3.82	54.8
radar	512	3.95	73.4

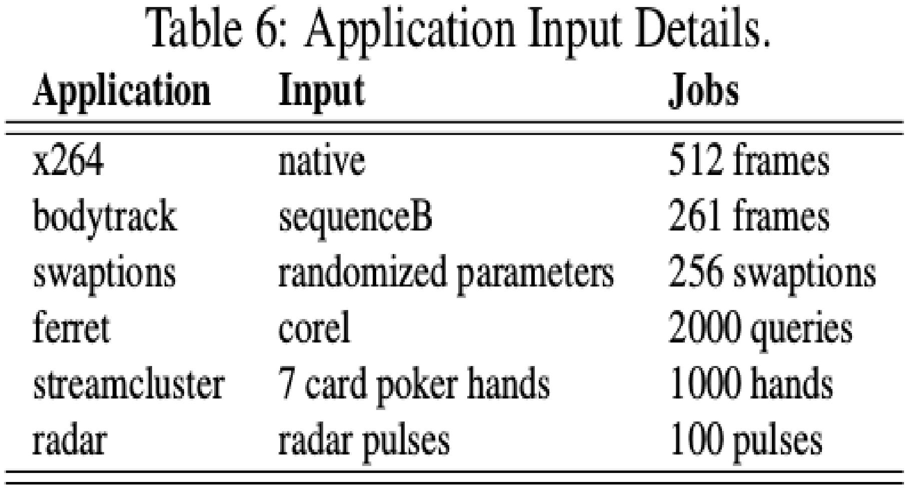


Table 6: Application Input Details.

Application	Input	Jobs
x264	native	512 frames
bodytrack	sequenceB	261 frames
swaptions	randomized parameters	256 swaptions
ferret	corel	2000 queries
streamcluster	7 card poker hands	1000 hands
radar	radar pulses	100 pulses

Benchmarks

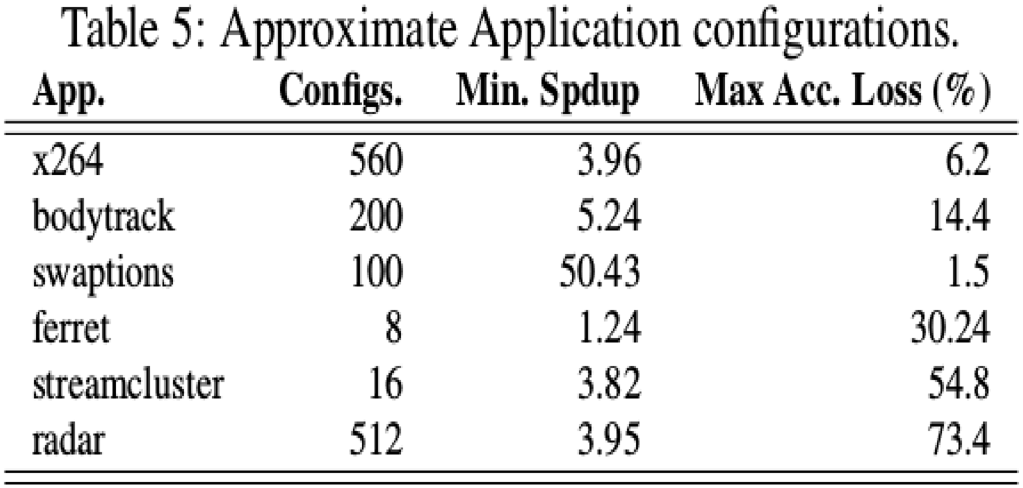


Table 5: Approximate Application configurations.

App.	Configs.	Min. Spdup	Max Acc. Loss (%)
x264	560	3.96	6.2
bodytrack	200	5.24	14.4
swaptions	100	50.43	1.5
ferret	8	1.24	30.24
streamcluster	16	3.82	54.8
radar	512	3.95	73.4

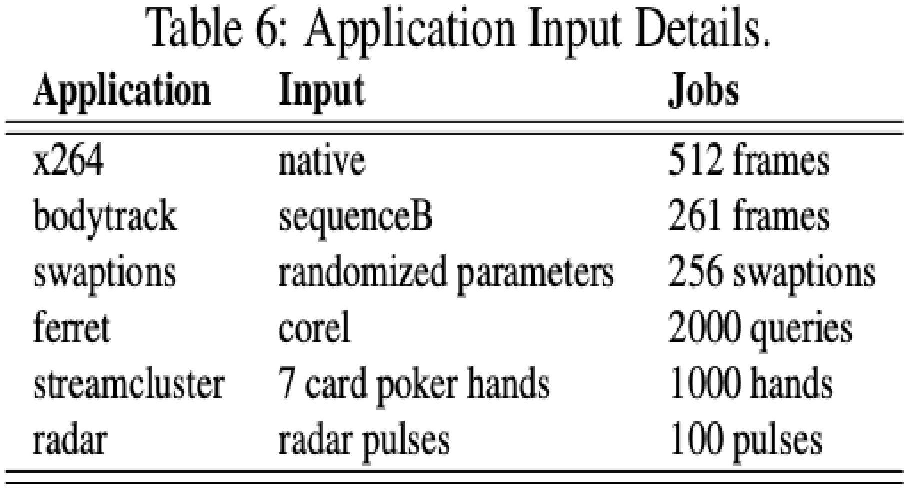


Table 6: Application Input Details.

Application	Input	Jobs
x264	native	512 frames
bodytrack	sequenceB	261 frames
swaptions	randomized parameters	256 swaptions
ferret	corel	2000 queries
streamcluster	7 card poker hands	1000 hands
radar	radar pulses	100 pulses

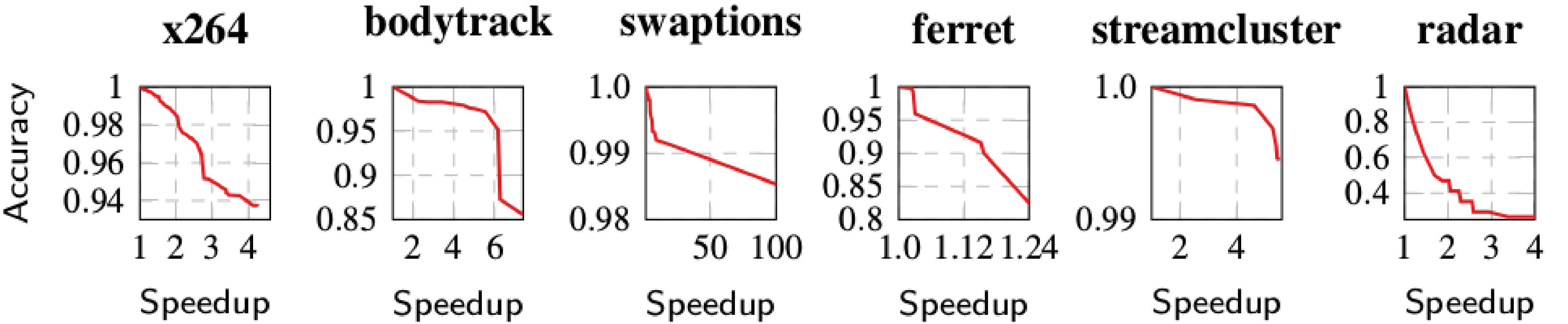
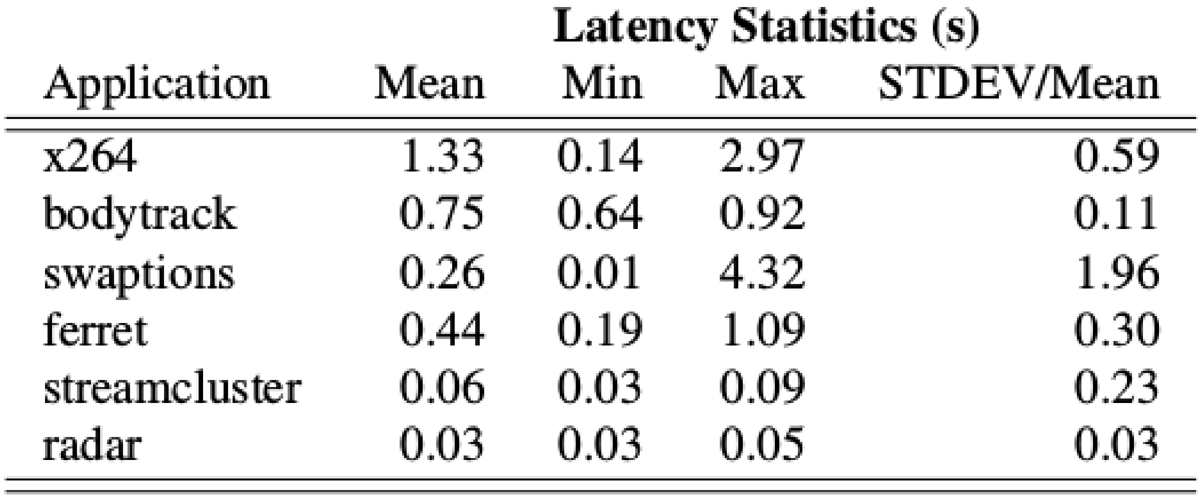


Figure 4: Speedup and accuracy tradeoffs for test applications.

Benchmarks

- None of these benchmark were intended to be run with hard-timing constraints.
- To quantify this inherent unpredictability application performance changes are measured as speedup (factor by which latency decreases when moving from the nominal setting)
- Then mean, minimum, maximum, and standard deviation over mean for all jobs in a benchmark are calculated.

Table 7: Application Timing Statistics.



	Latency Statistics (s)			
Application	Mean	Min	Max	STDEV/Mean
x264	1.33	0.14	2.97	0.59
bodytrack	0.75	0.64	0.92	0.11
swaptions	0.26	0.01	4.32	1.96
ferret	0.44	0.19	1.09	0.30
streamcluster	0.06	0.03	0.09	0.23
radar	0.03	0.03	0.05	0.03

Experiment Evalutation

- MEANTIME is compared with following approaches
  - wcet
  - PowerDial
  - cross
  - Energy-aware
  - optimal

Experiment Evalutation

- MEANTIME is compared with following approaches
  - **wcet**
    - *Meets hard real-time constraints by reserving resources sufficient for worst case latency*
    - *Then sleeps if job completes early (racing to idle)*
    - *Used worst observed latency*
  - PowerDial
  - cross
  - Energy-aware
  - optimal

Experiment Evalutation

- MEANTIME is compared with following approaches
  - **wcet**
  - **PowerDial**
    - *This uses control theory to adjust the application level parameters and performance to achive maximum accuracy*
    - *Since this does not adjust system level resource usage it cannot lower system energy consumption*
  - cross
  - Energy-aware
  - optimal

Experiment Evalutation

- MEANTIME is compared with following approaches
  - **wcet**
  - **PowerDial**
  - **cross**
    - *Combines application accuracy and system resource management*
    - *Do not provide hard guarantees*
    - *For experiments they are considering worst case timing for both the application accuracy and system resource usage*
  - Energy-aware
  - optimal

Experiment Evalutation

- MEANTIME is compared with following approaches
  - **wcet**
  - **PowerDial**
  - **cross**
  - **Energy-aware**
    - *uses state-of-the-art control and optimization techniques to allocate minimal energy resources for the current job*
  - optimal

Experiment Evalutation

- MEANTIME is compared with following approaches
  - **wcet**
  - **PowerDial**
  - **cross**
  - **Energy-aware**
  - **Optimal**
    - *To get optimal numbers each application was run in each system configuration and logged latency for each job within the application.*
    - *Then post proceed these logs to determine the minimal energy configuration for each job that would have met the latency goal.*

Experiment Results:Timing Properties

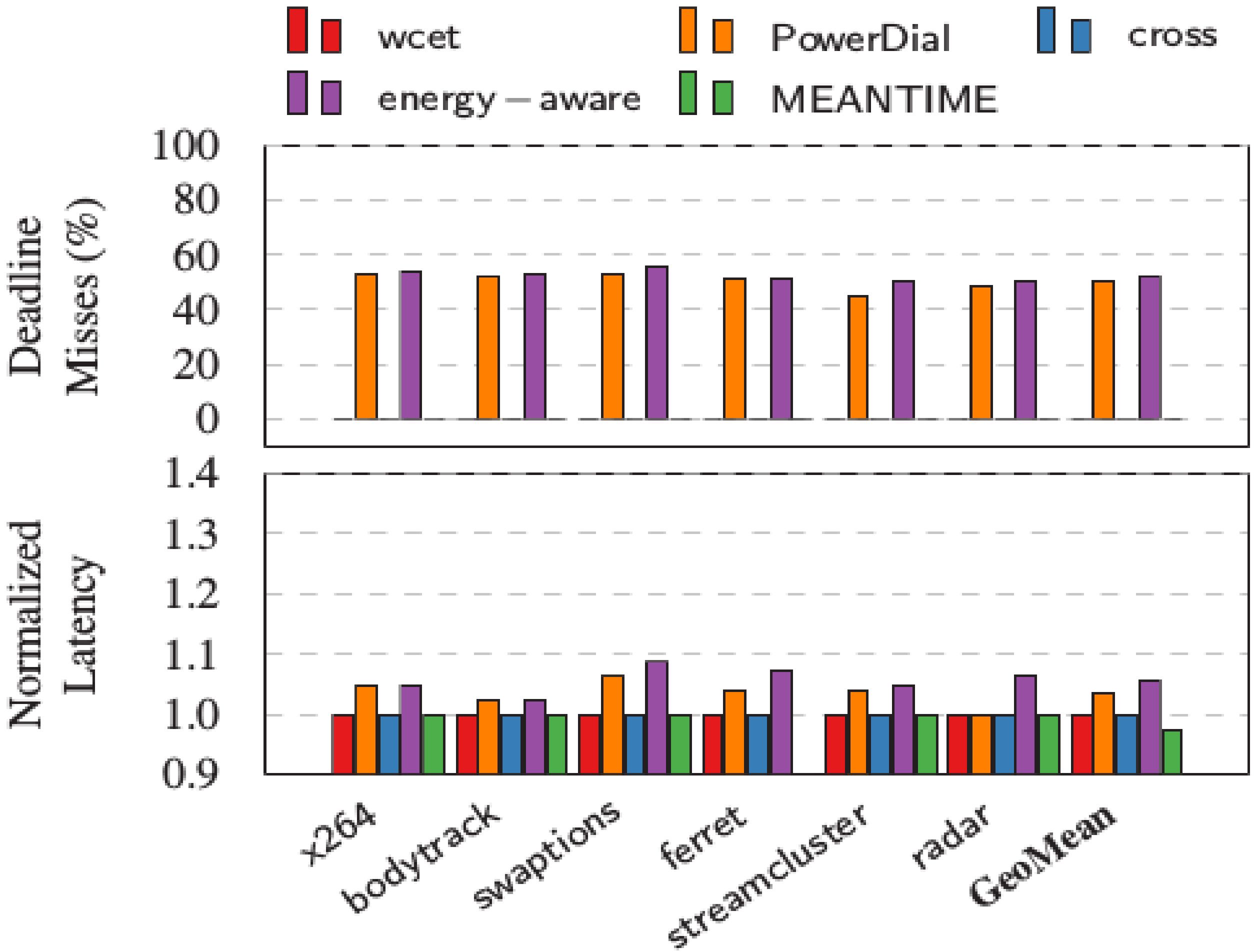


Figure 5: Deadline misses (top) and normalized latency (bottom) for different resource techniques.

Experiment Results: Energy

- Two factors that predict the enrgy savings are
  - Greater the variance in timing, the greater the energy saving potential
  - If an application’s approximate configurations do not provide much speedup, then the potential for energy savings is also reduced

Experiment Results: Energy

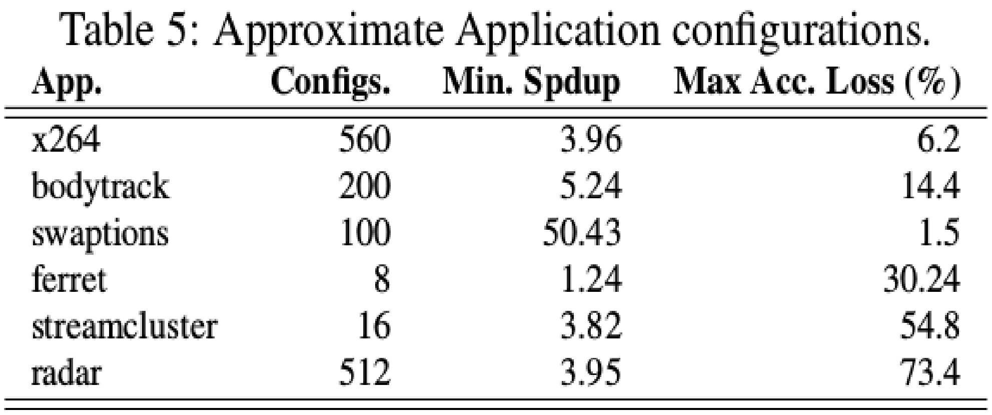


Table 5: Approximate Application configurations.

App.	Configs.	Min. Spdup	Max Acc. Loss (%)
x264	560	3.96	6.2
bodytrack	200	5.24	14.4
swaptions	100	50.43	1.5
ferret	8	1.24	30.24
streamcluster	16	3.82	54.8
radar	512	3.95	73.4

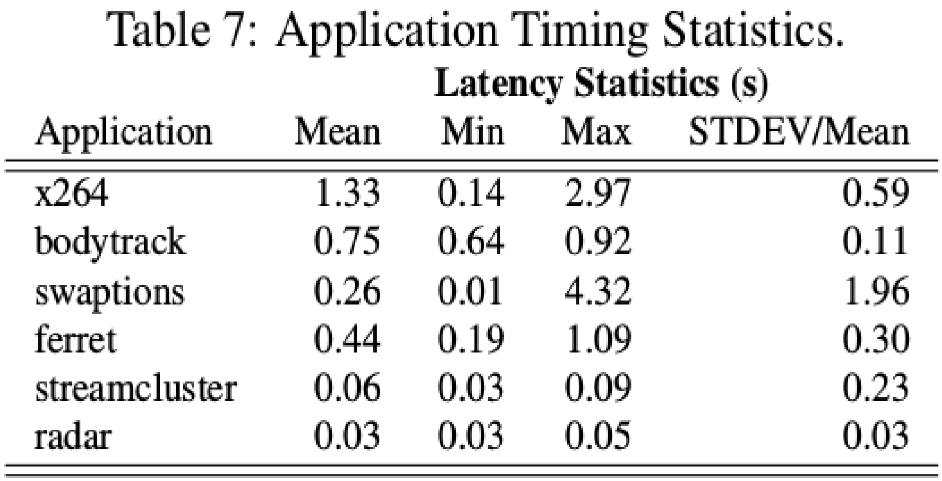


Table 7: Application Timing Statistics.

	Latency Statistics (s)			
Application	Mean	Min	Max	STDEV/Mean
x264	1.33	0.14	2.97	0.59
bodytrack	0.75	0.64	0.92	0.11
swaptions	0.26	0.01	4.32	1.96
ferret	0.44	0.19	1.09	0.30
streamcluster	0.06	0.03	0.09	0.23
radar	0.03	0.03	0.05	0.03

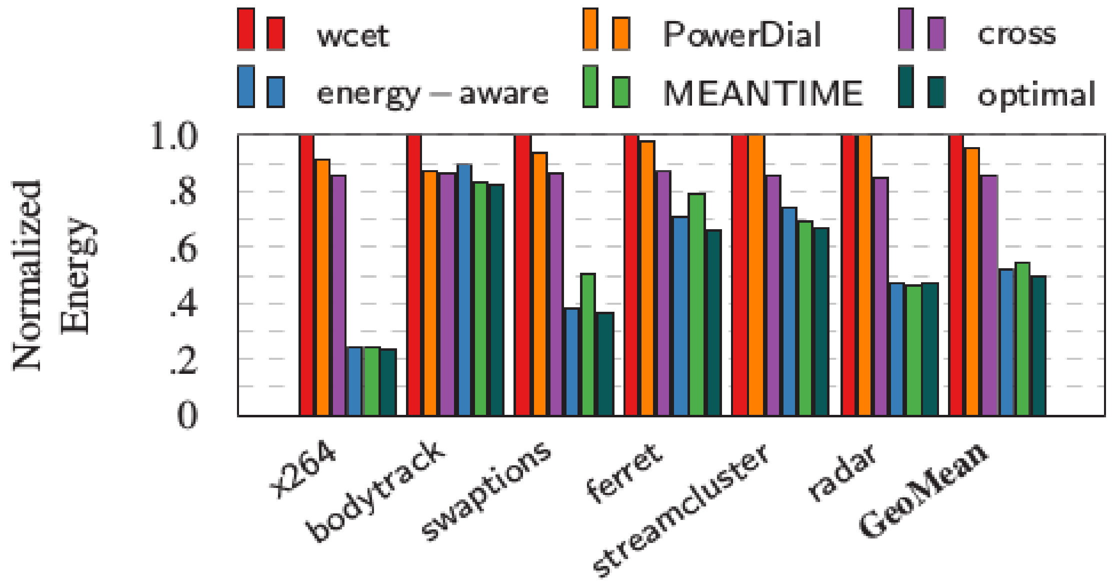


Figure 6: Energy consumption normalized to wcet. Lower numbers represent reduced energy consumption.

Experiment Results: Accuracy

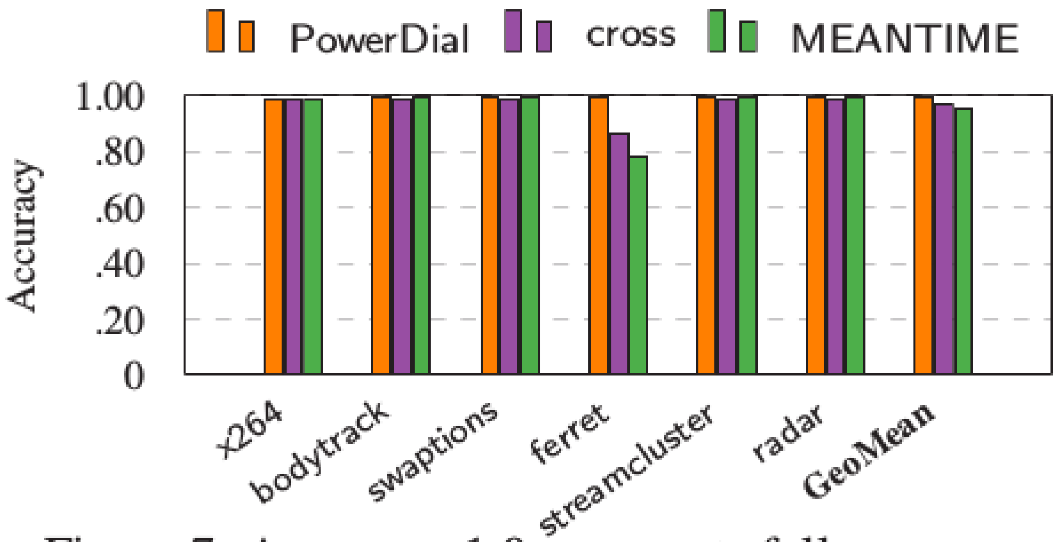
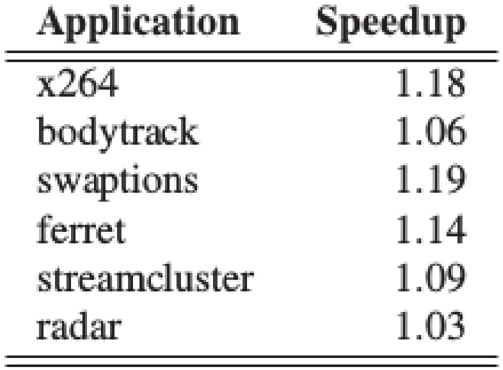


Figure 7: Accuracy, 1.0 represents full accuracy.

Table 8: Required Speedup.



Application	Speedup
x264	1.18
bodytrack	1.06
swaptions	1.19
ferret	1.14
streamcluster	1.09
radar	1.03

Adapting to Changing Requirements

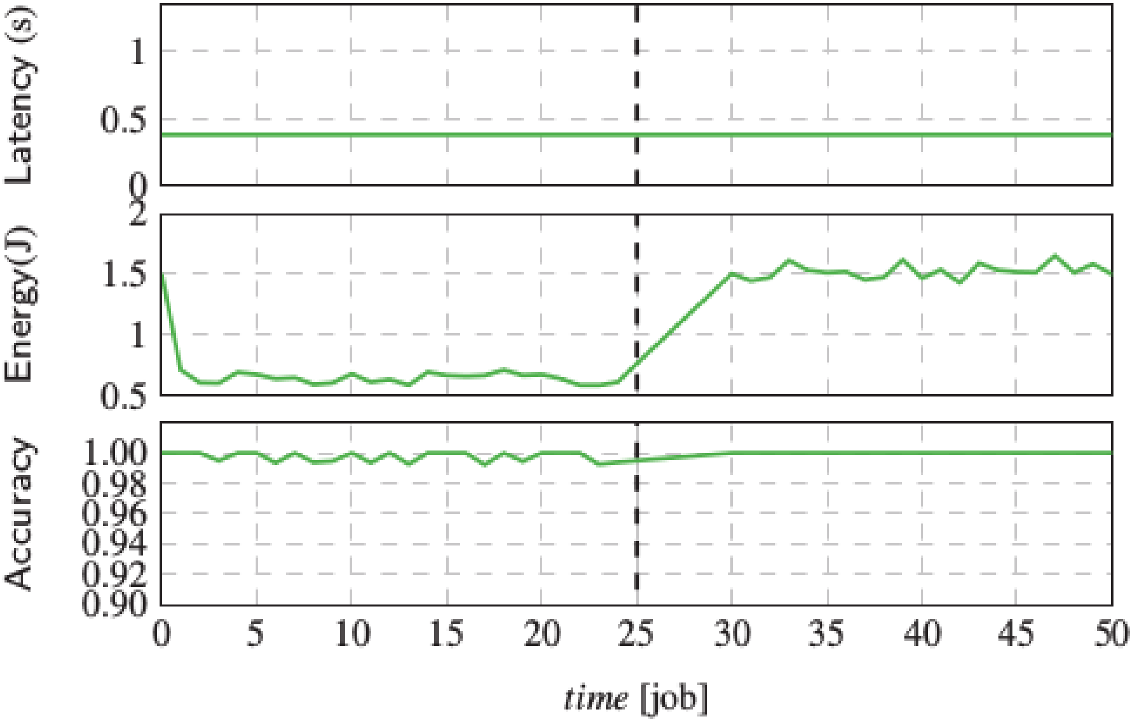


Figure 8: MEANTIME reacting to changing radar accuracy goals.

Adapting to Phases

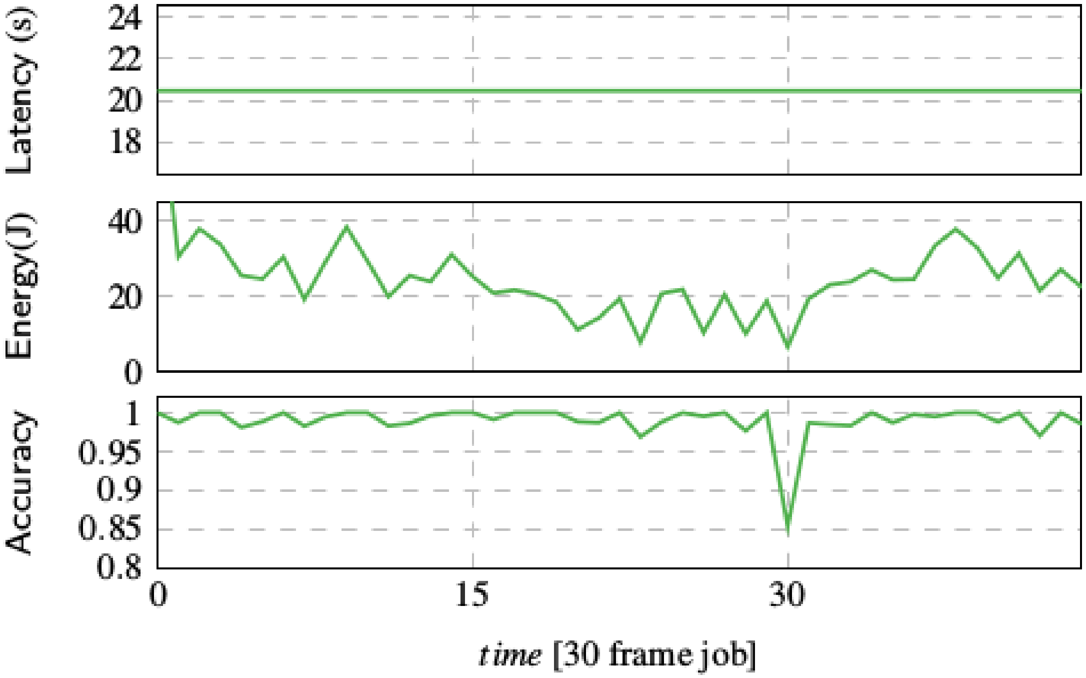


Figure 9: MEANTIME reacting to phases in x264.

Framework Overhead

- Computational complexity of the methodology is of O(1)
- To calculate worst case latency, MEANTIME is run with no applications to manage and dummy resources are allocated through same system calls which do not change the timing (latency of the system call should not be considered).
- This was executed for 1000 iterations.
- Worst case latency was measured as ~100μs
- This is used as switching overhead in the equations.

Conclusion

- MEANTIME's contribution is using application approximation to provide both hard real-time guarantees and energy efficiency.
- This methodology is applicable for applications with hard real-timings which can trade-off accuracy for energy efficiency.

Queries ????

Thank you