Evaluating DVFS and Concurrency Throttling on IBM's Power8 Architecture

Wei Wang (University of Delaware), Edgar A. Leon (Lawrence Livermore National Laboratory)

Abstract

LLNL's next supercomputer Sierra will be based on IBM's Power Architecture. Early insights of how applications perform on this architecture are beneficial. In this work, we study application performance under different CPU frequency and concurrency settings on a Power8 system. DVFS and Concurrency Throttling improve both throughput and energy efficiency.

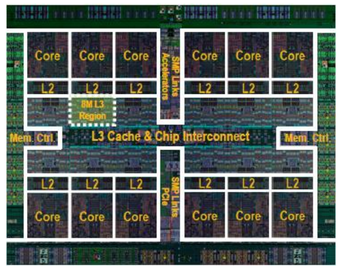
IBM's Power8 Architecture

❖4 cores, 32 threads total

❖Supports SMT1 to SMT8

❖69 DVFS frequencies

❖What is the impact on application performance?



IBM Power8 Architecture: 12-cores (SMT8)

Methodology

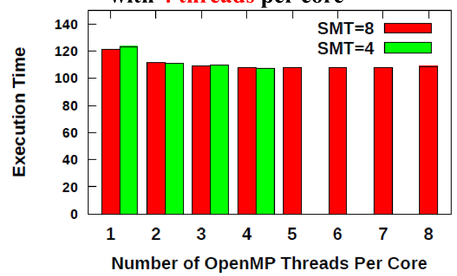
- ❖ Measurement of application performance
  - ➢ Varying frequencies
  - ➢ Different number of threads per core
  - ➢ SMT8 vs. SMT4
- ❖ Measurement granularity
  - ➢ OpenMP parallel loop
- ❖ Runtime frequency change
  - ➢ Energy control API
- ❖ Benchmarks
  - ➢ Graph500 – Breadth-First Search
  - ➢ LULESH 2.0 – Shock Hydrodynamics
  - ➢ miniFE – Finite Element Code

New Insights with DVFS and Concurrency Throttling

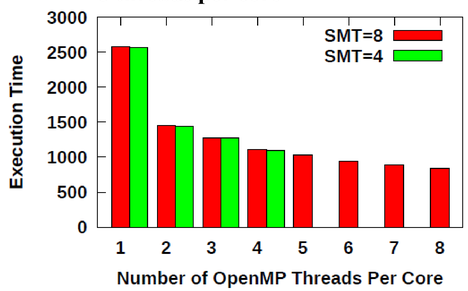
❖Concurrency Throttling

- ➢ Applications do not always need 8 threads per core
  - ❑ miniFE and LULESH perform best with 4 threads per core
  - ❑ Graph500 performs best with 8 threads per core
- ➢ SMT4 performs as well as SMT8

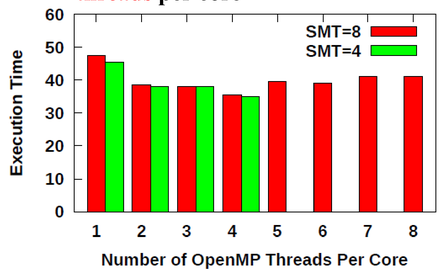
miniFE performs the best with 4 threads per core



Graph500 performs the best with 8 threads per core



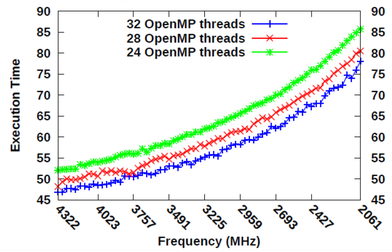
LULESH performs the best with 4 threads per core



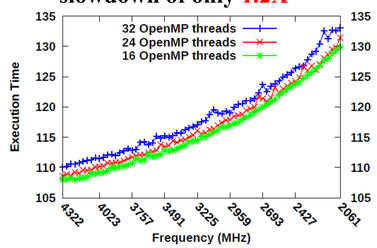
❖DVFS

- ➢ Lower frequency leads to longer execution time for all benchmarks
- ➢ However, applications have quite different slowdown when reducing the frequency
- ➢ Memory bound applications like miniFE benefit from DVFS

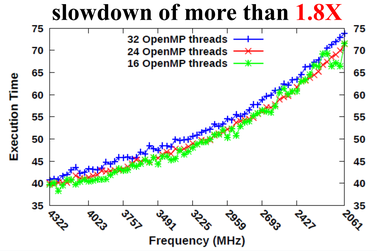
Graph500: performance slowdown of more than 1.6X



miniFE: performance slowdown of only 1.2X



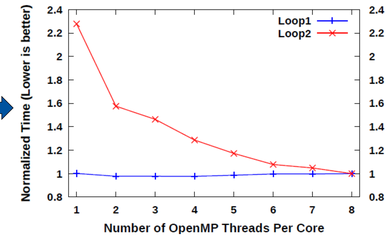
LULESH: performance slowdown of more than 1.8X

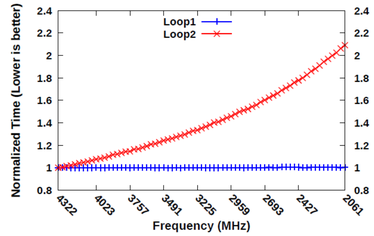


Fine-grain Analysis of Benchmarks

❖miniFE Loops respond differently when applying concurrency throttling and DVFS

Loop2's performance gets worse but Loop1's performance slightly improves. Loop1 dominates the execution time.





Loop2's performance gets worse, but Loop1's performance remains the same. Loop1 dominates the execution time.

❖Applying DVFS and Concurrency Throttling improves performance and energy

Executables	Total Number of Threads	DVFS Frequency	Execution Time
miniFE baseline	32	4.322GHz	108.68 seconds
miniFE with DVFS & Concurrency Throttling	16	4.322GHz & 2.061GHz	107.49 seconds

Running miniFE with reduced number of threads (for the entire application) and reduced frequency (for a loop)

Conclusions

❖Applications benefit from using concurrency throttling

❖Some application regions are insensitive to frequency change, providing energy savings

❖DVFS and Concurrency Throttling improve performance and energy efficiency

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

IM Review and Release Number: LLNL-POST-675609