Supercomputing 2017 Doctoral Showcase

Using Runtime Optimizations to improve Energy Efficiency in High Performance Computing

Sridutt Bhalachandra

Department of Computer Science, University of North Carolina at Chapel Hill

Advisors: Dr. Allan Porterfield & Prof. Jan Prins
Mentors: Dr. Stephen Olivier, Sandia National Laboratories & Dr. Rob Fowler

November 14, 2017

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci

To Exascale and Beyond...

System/Site	Performance (PFLOPS)	Power (MW)	Energy Efficiency (GFLOPS/W)
Exascale	1000	20	50
Taihulight	93	15	6
Tianhe 2	34	18	2
Piz Daint	20	2	9

To Exascale and Beyond...

System/Site	Performance (PFLOPS)	Power (MW)	Energy Efficiency (GFLOPS/W)
Exascale	1000	20	50
Taihulight	93	15	6
Tianhe 2	34	18	2
Piz Daint	20	2	9
TSUBAME 3.0	2	0.14	14
kukai	0.46	0.03	14
AIST AI Cloud	0.96	0.08	13

5x - 10x improvement in energy efficiency required

Previous work and our approach

- Many related works involve static analysis of program

  (Feng et al. 2005, Kamil et al. 2008, Huang and Feng 2009...)

  - Use a runtime approach
- Many solutions require code changes

  (Ge et al. 2005, Wang et al. 2015...)

  - Make solution transparent
- Many results are simulated or use NAS parallel benchmarks

  (Hsu and Feng 2005, Kandalla et al. 2010, Rountree et al. 2012, Livingston et al. 2014...)

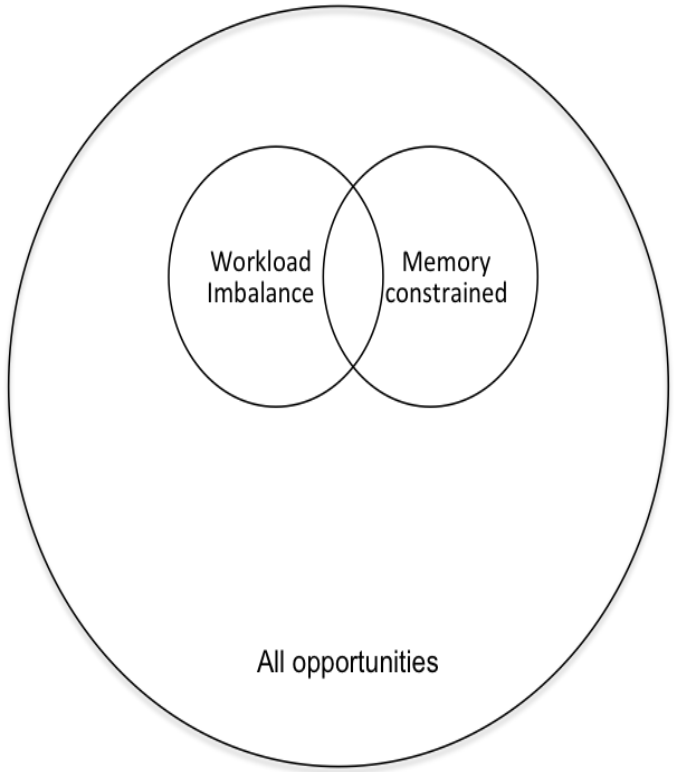
  - Show results on mini-apps and applications
- Socket-wide Dynamic Voltage Frequency Scaling (DVFS)

  (Kappiah et al. 2005, Rountree et al. 2009, Ge et al. 2005, Freeh and Lowenthal 2005...)

  - Explore other, core-specific power controls

Scope

- Targets applications exhibiting workload imbalance and/or memory constrained
- Pure MPI applications with one rank per core (No MPI+X)
- Considers only Intel processor architecture (No accelerators)



Computational workload balance

- Inherent to application nature or configuration-specific
- Increasing due to system heterogeneity
  - Will further increase in power-limited/over-provisioned systems

Published work

1. An Adaptive Core-specific Runtime for Energy Efficiency (IPDPS 2017)
2. Using Dynamic Duty Cycle Modulation to improve energy efficiency in High Performance Computing (HPPAC 2015)

Core-specific perspective to Energy Efficiency

Socket-wide power control $\longrightarrow$ can slow critical core

- trade performance for power reduction (save energy)

Core-specific perspective to Energy Efficiency

Socket-wide power control $\longrightarrow$ can slow critical core

- trade performance for power reduction (save energy)

Core-specific control

- match a core's duty cycle to its workload

$$Duty\ cycle = \frac{Time\ core\ in\ active\ state}{Total\ time\ (clock\ cycles)}$$

$*$Change core active time using DDCM or clock cycles using DVFS

Core-specific perspective to Energy Efficiency

Socket-wide power control $\longrightarrow$ can slow critical core

- trade performance for power reduction (save energy)

Core-specific control

- match a core's duty cycle to its workload

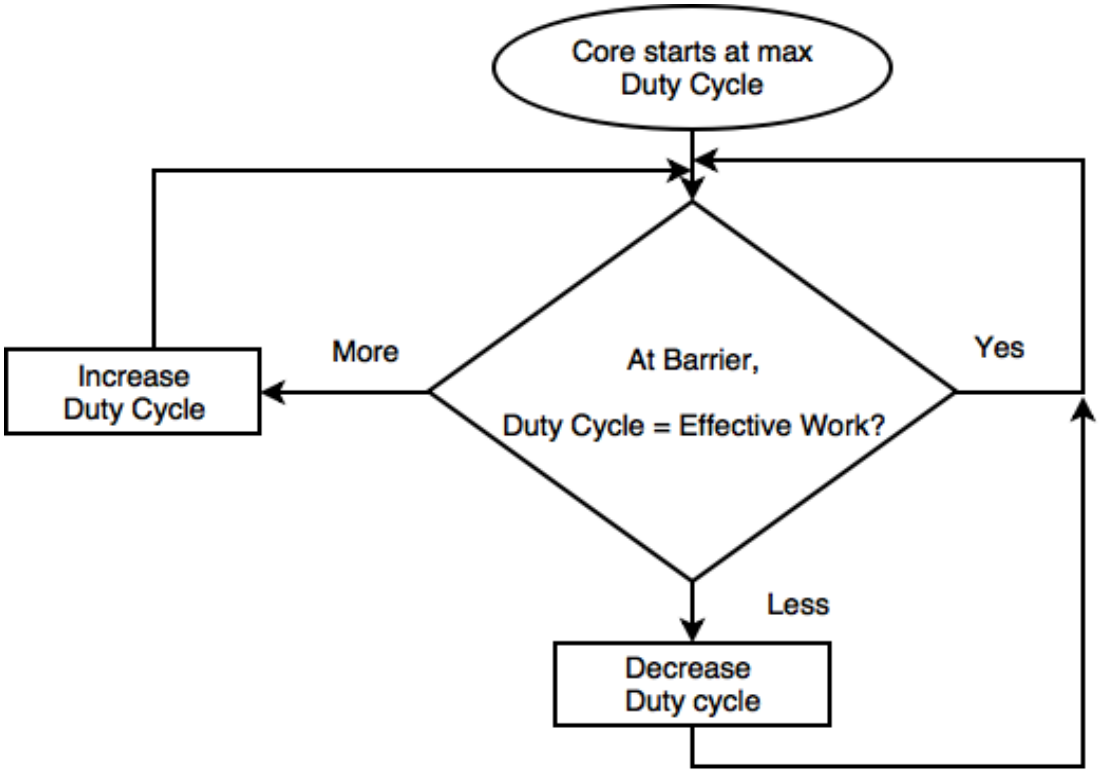
$$Duty\ cycle = \frac{Time\ core\ in\ active\ state}{Total\ time\ (clock\ cycles)}$$

$*$Change core active time using DDCM or clock cycles using DVFS

$$Work = \frac{Compute\ time}{Compute\ time + Idle\ time}\ (constant\ frequency)$$

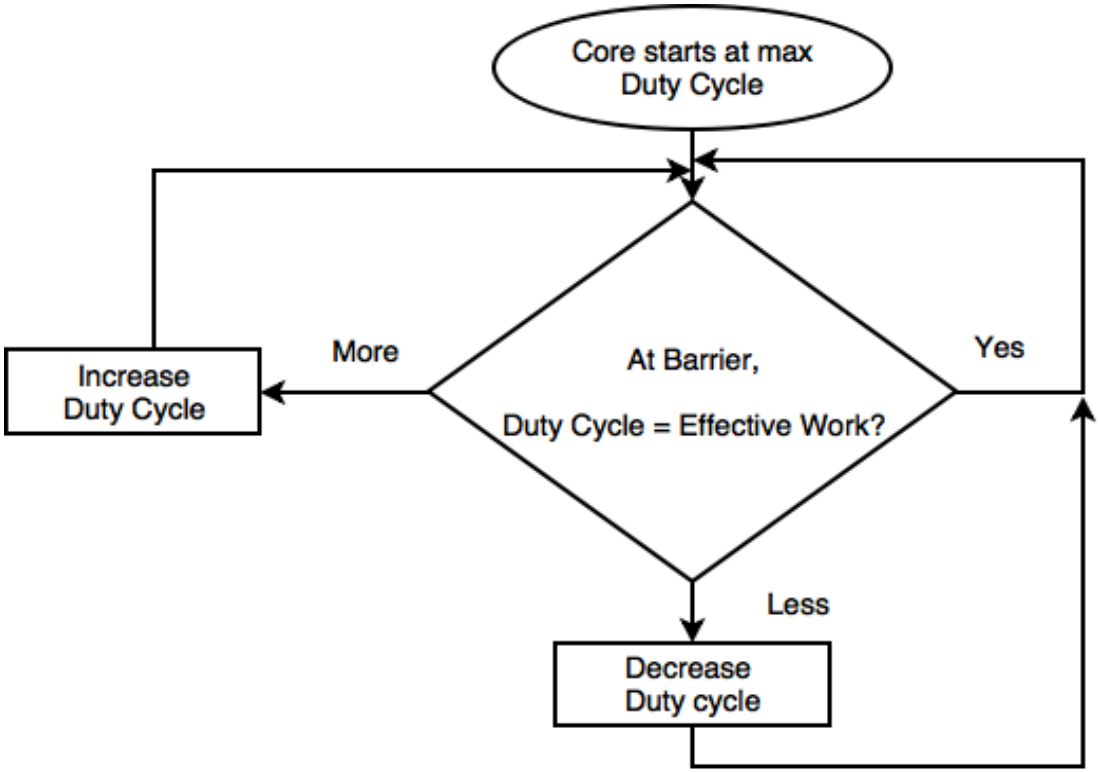
$$Effective\ Work = \frac{Compute\ time}{Compute\ time + Idle\ time} * \frac{Max\ frequency}{Current\ frequency}$$

Runtime policy



- Assumes temporal behavior across phases (basic solution to follow)
- Policy calculation local to core, no communication

Runtime policy

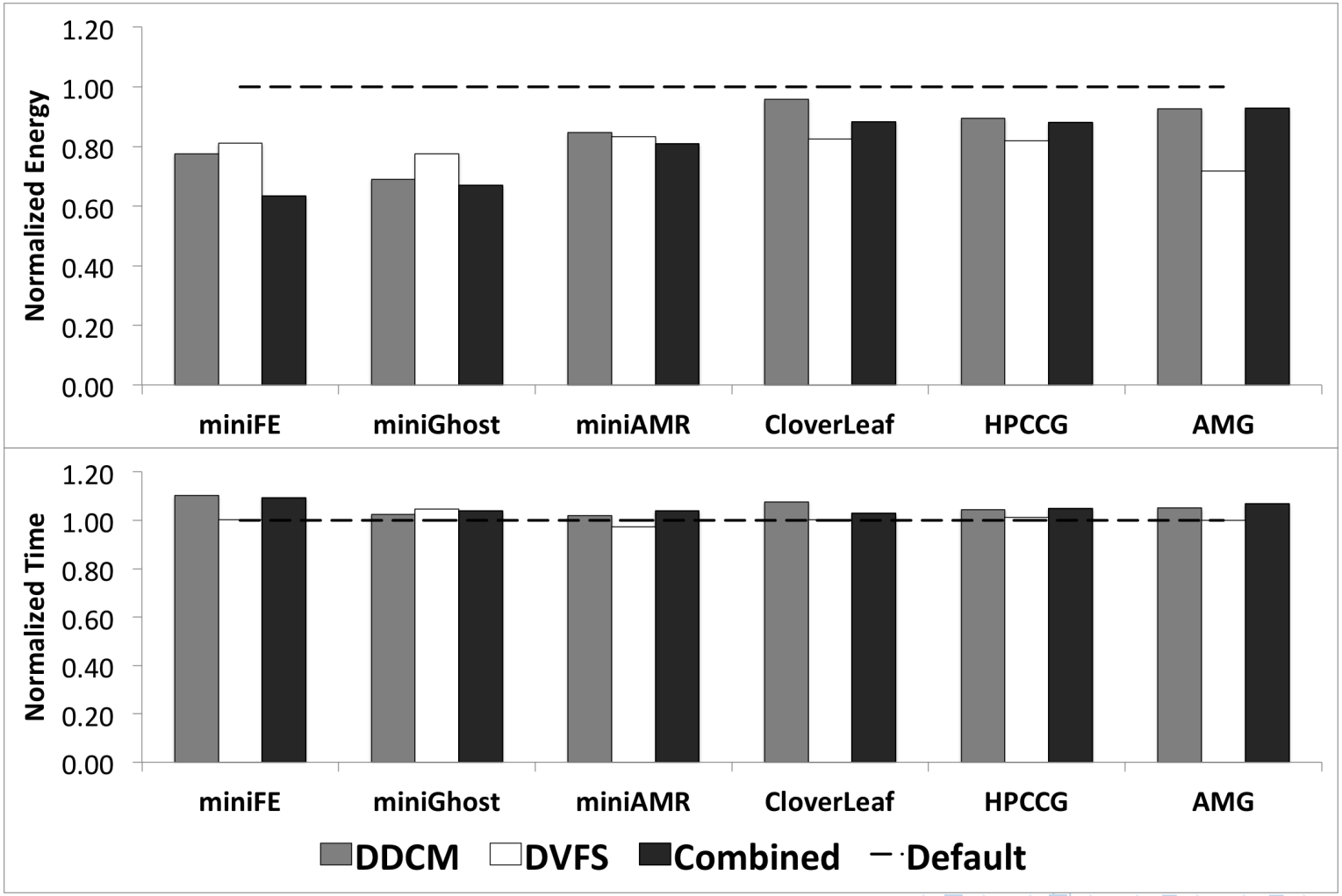


- Assumes temporal behavior across phases (basic solution to follow)
- Policy calculation local to core, no communication

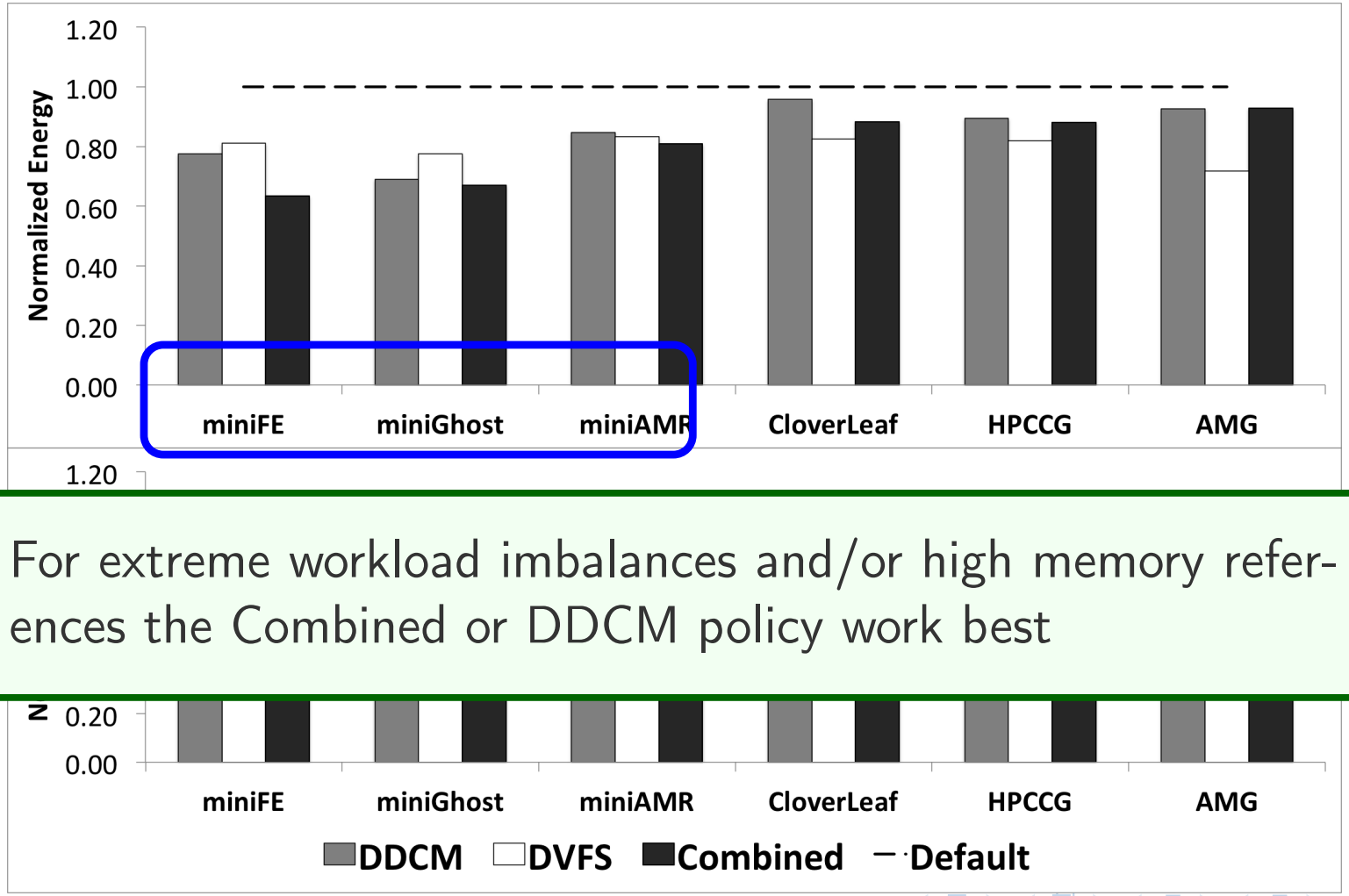
Combined policy
($Power_{DVFS} < Power_{DDCM}$)

- Use DVFS policy until lowest frequency reached
- Thereafter, use DDCM policy

Results for mini-apps

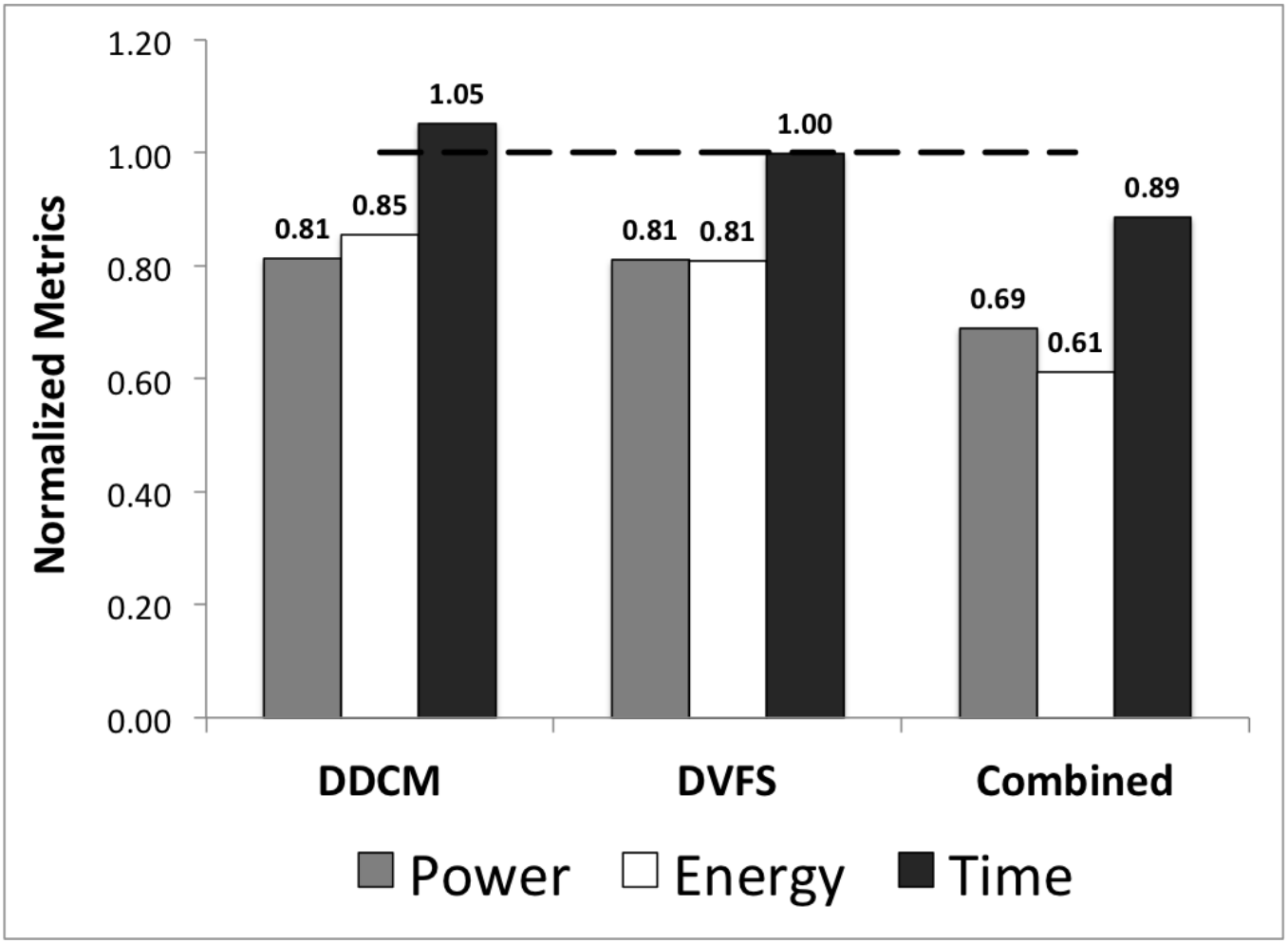


Results for mini-apps

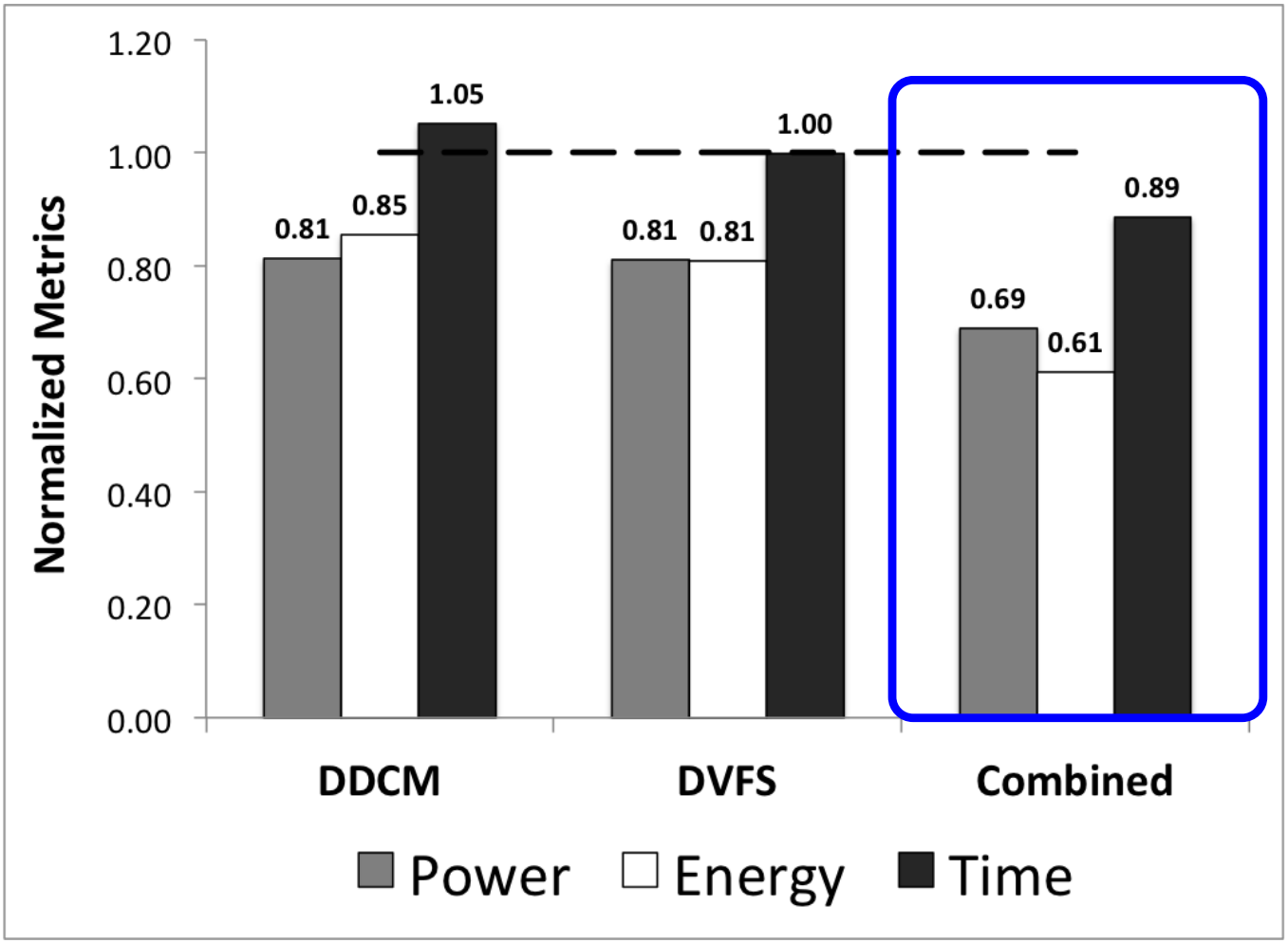


For extreme workload imbalances and/or high memory references the Combined or DDCM policy work best

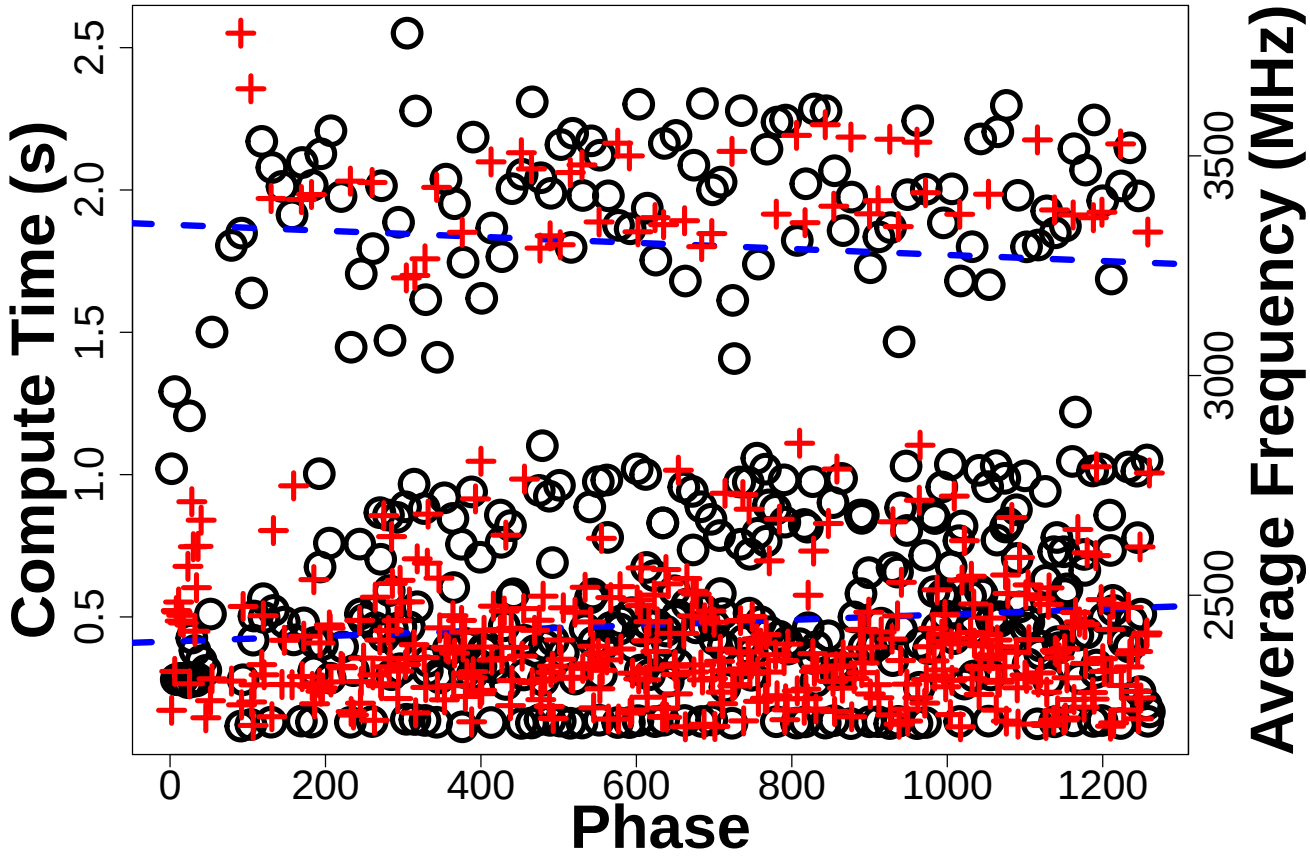
ParaDis results



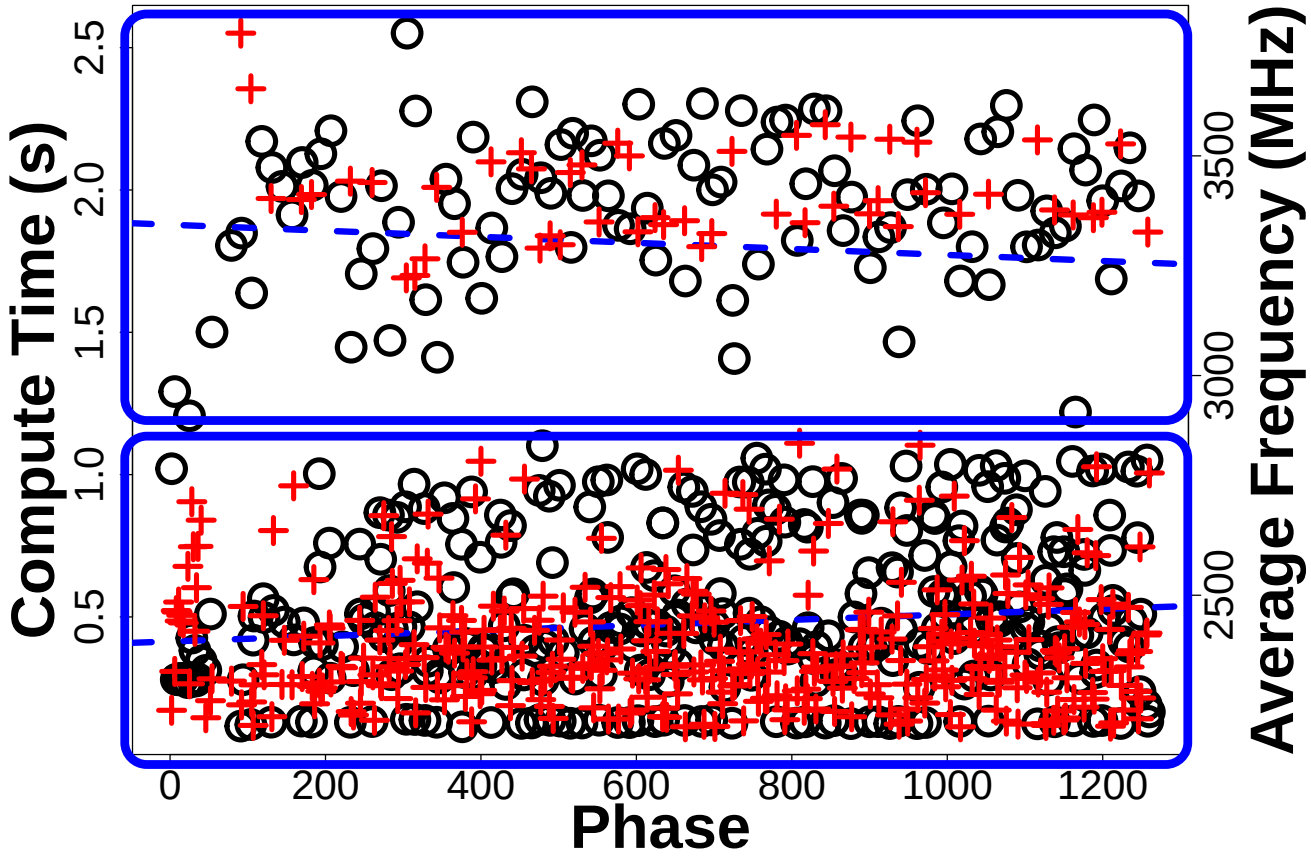
ParaDis results



ParaDis critical path on 24 nodes (768 cores) - Default

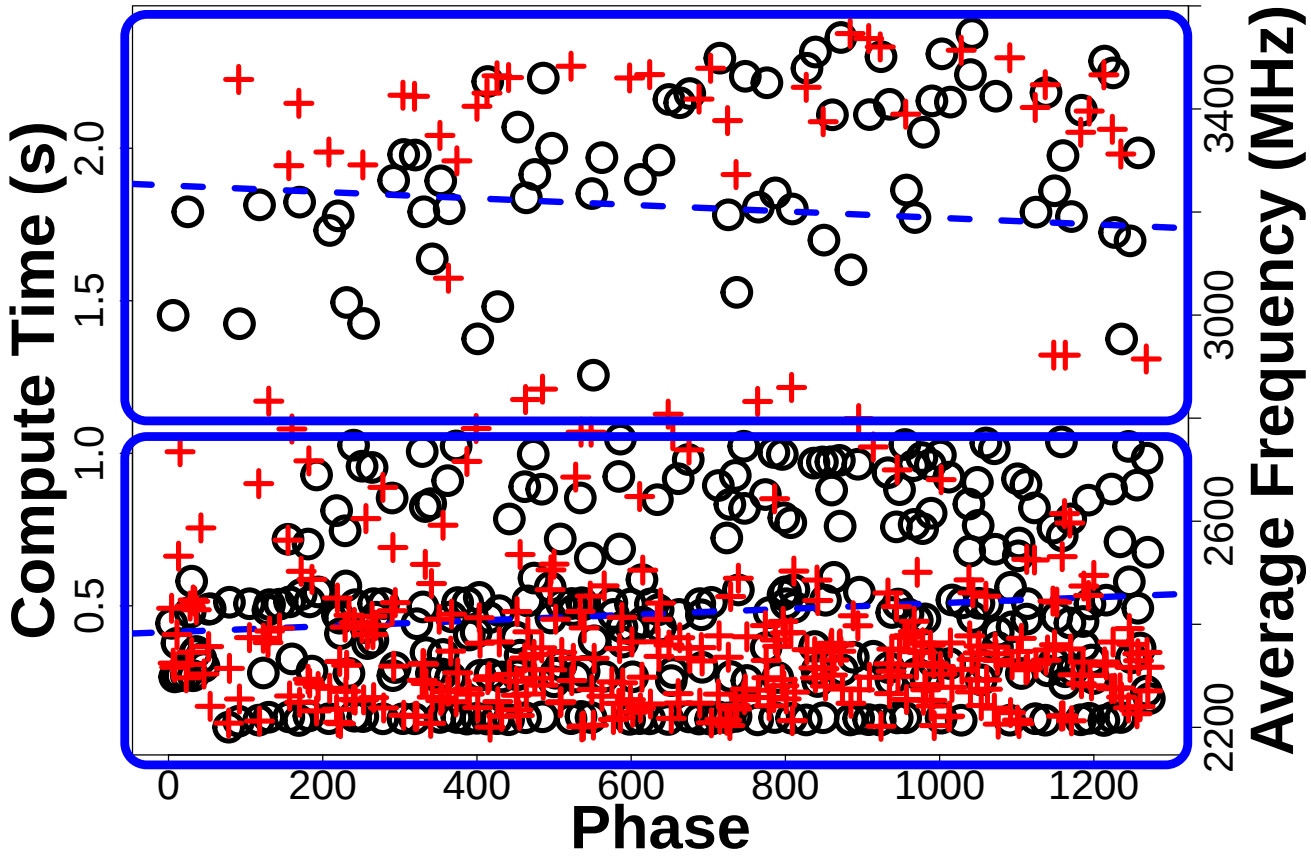


ParaDis critical path on 24 nodes (768 cores) - Default



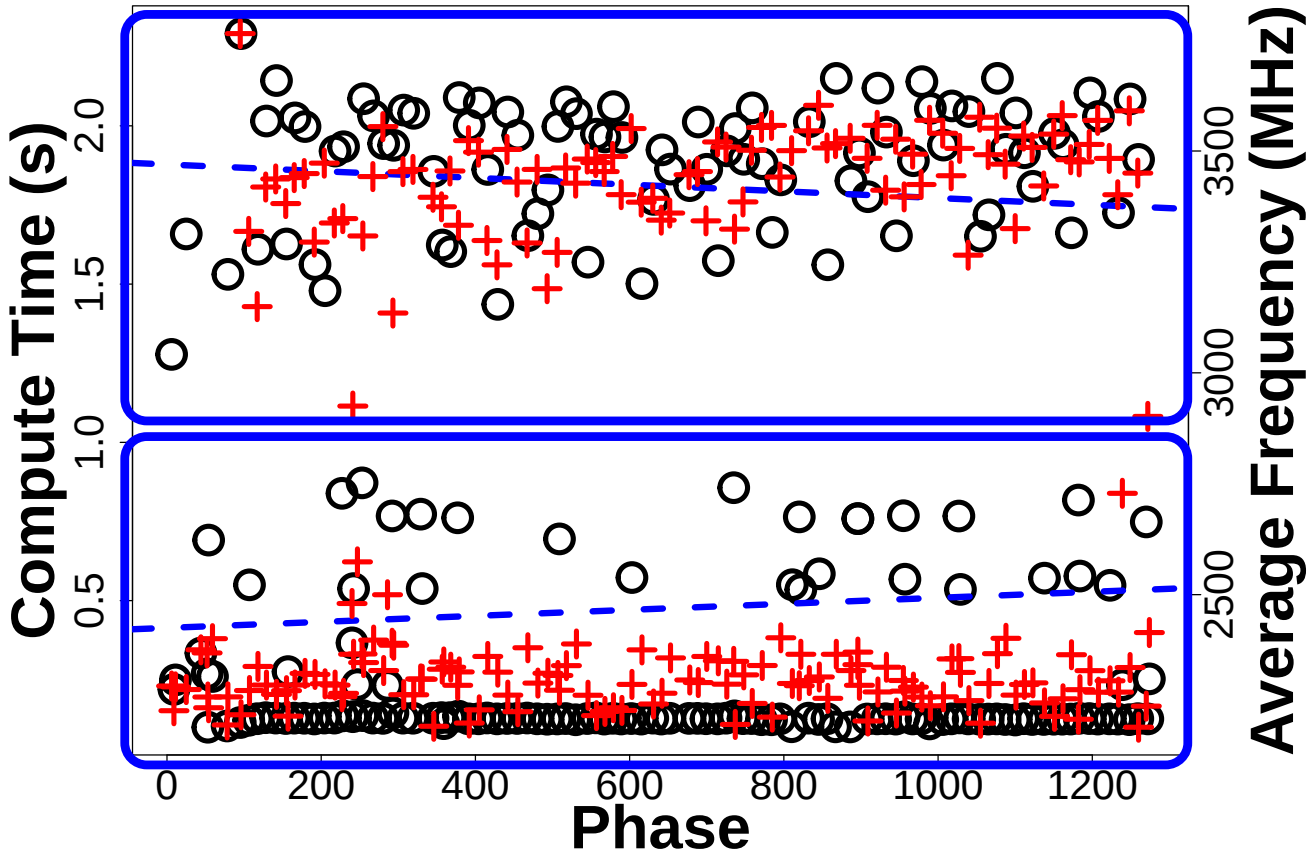
- Bimodal distribution of critical path times < 1.0s and > 1.0s
- Successive phases are similar, with only occasional jumps
- Average critical path frequency (Default) = **2507.4MHz**

ParaDis critical path on 24 nodes (768 cores) - DVFS



- Average critical path frequency (Default) = **2467.3MHz**

ParaDis critical path on 24 nodes (768 cores) - DDCM



- Very low frequency on non-critical cores for prolonged periods **reduces variation**, and **increases available thermal headroom** for critical cores
- Average critical path frequency (Default) = **2784.8MHz**

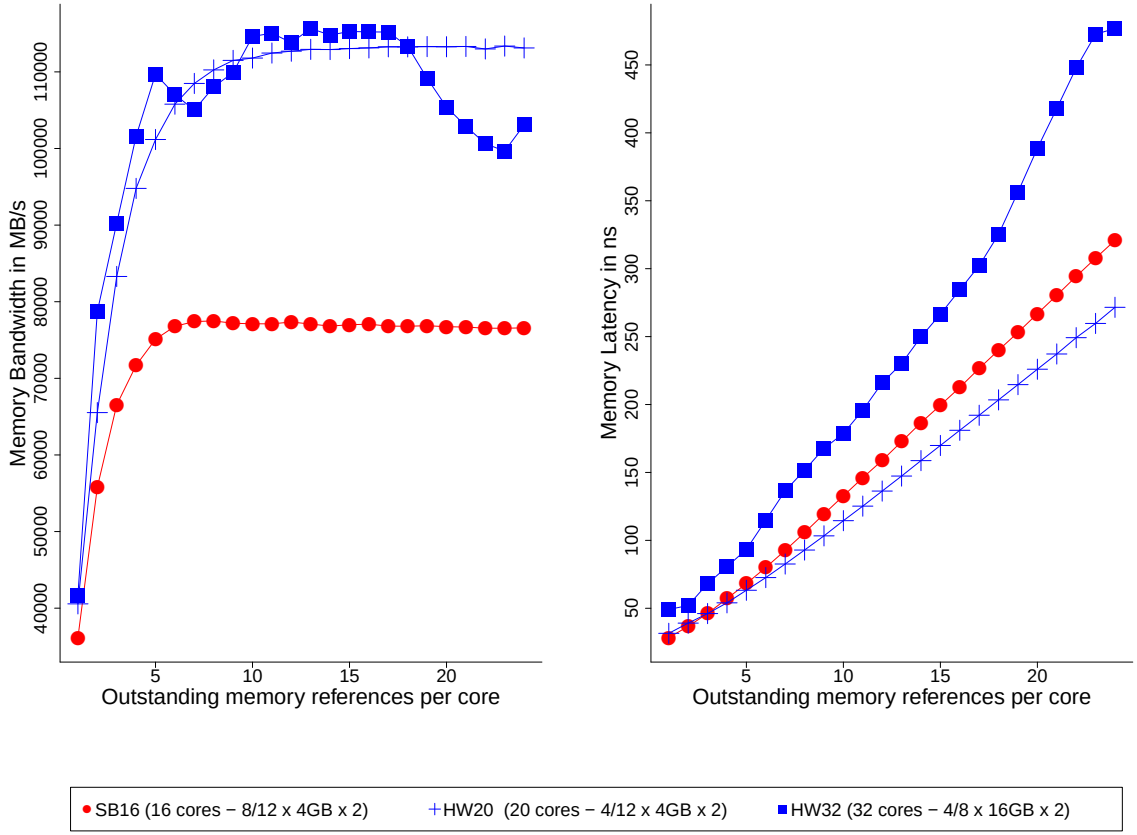
When applications are memory constrained

- Many HPC applications access memory often
- Memory operations are seldom explicit
  - Power wasted in CPU while waiting on memory

Proposed solutions

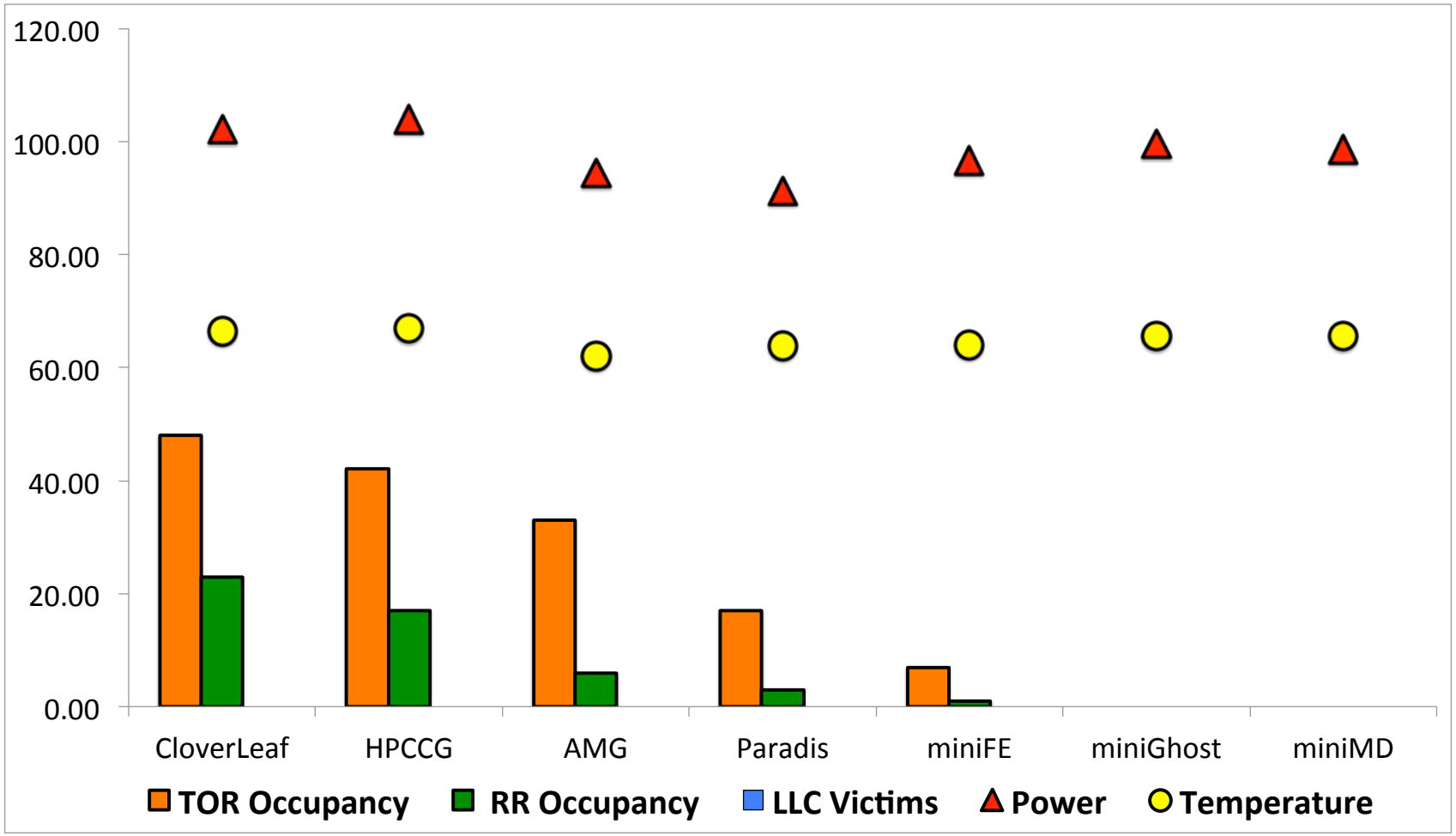
1. Improving Energy Efficiency in Memory-constrained Applications Using Core-specific Power Control (E2SC 2017)

Bandwidth and Latency analysis on modern hardware

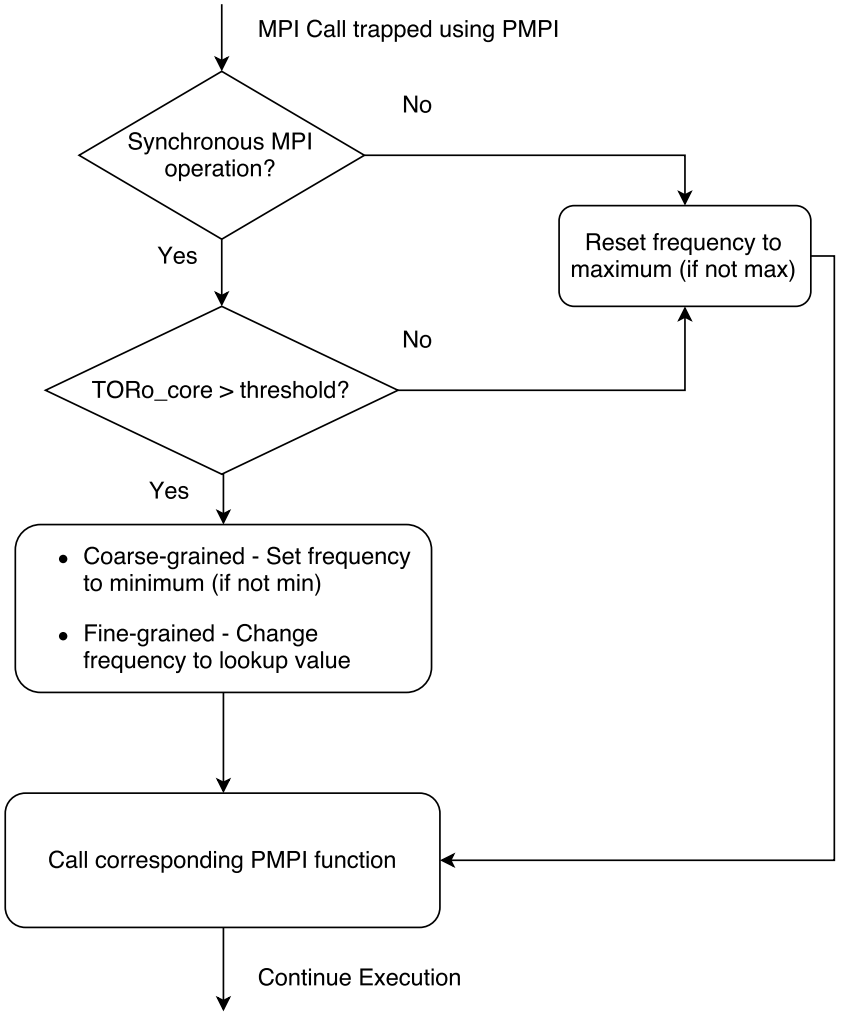


- pChase originally developed by Doug Pase consists of pointer chasing loops aimed to defeat the prefetcher.
  - works well for low level memory studies

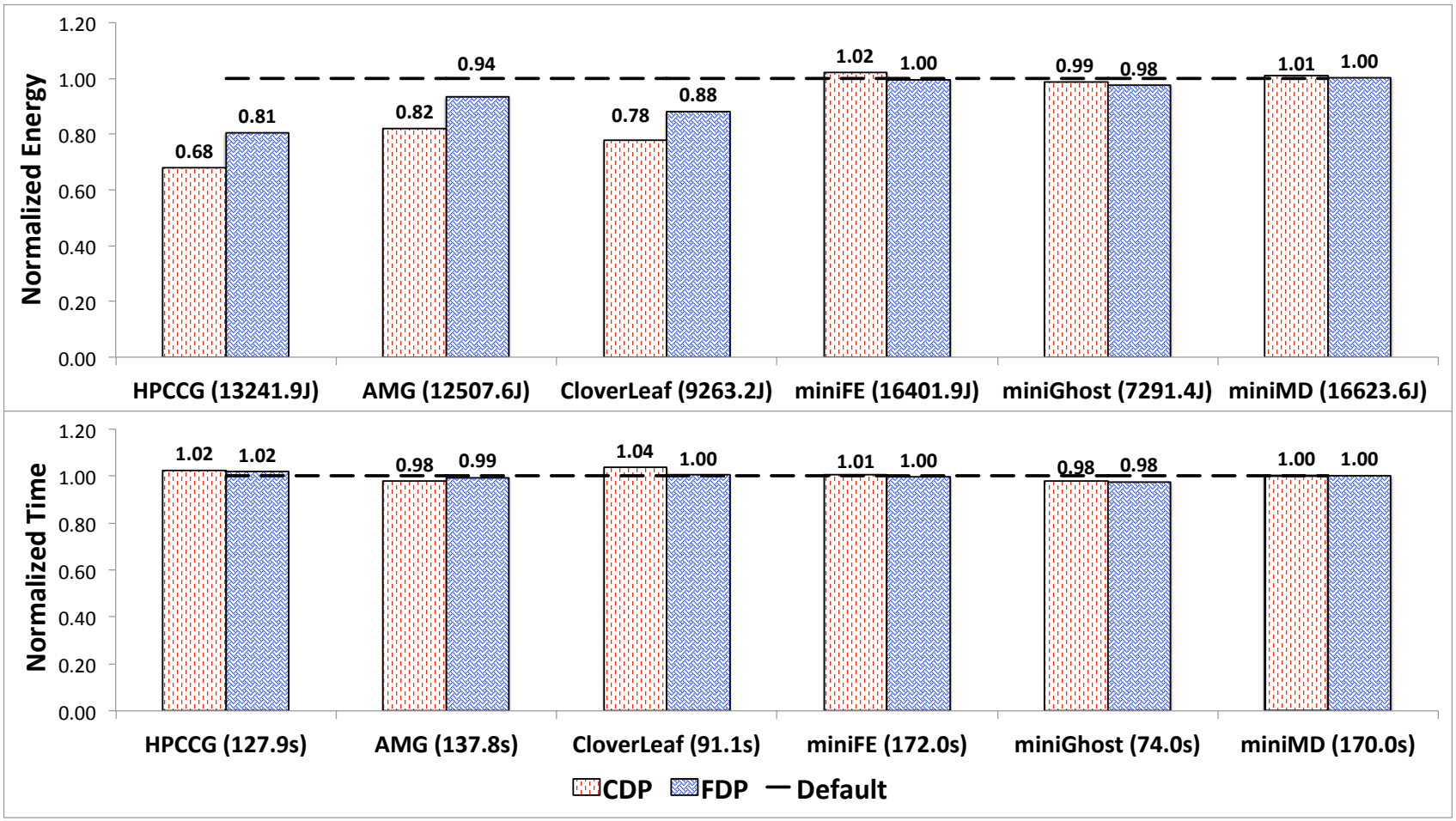
Characterizing Memory



Memory Policy



Results for mini-apps



Conclusion

- **Computational Workload Imbalance**
  - DDCM as an alternative to save energy
  - ACR allows processors with per-core specific power control to reduce power with little performance impact
  - MPI library transparent to an application and allows use of multiple power levers to improve energy efficiency
  - Power optimization can be a performance optimization

- **Memory constrained**
  - New metrics proposed conform to the memory behavior exhibited
  - The metrics are useful in constructing runtime policies
  - Quality of the new metrics not only allow memory characterization of the application, but facilitate their stand-alone use in policies without the need to monitor other metrics like power among others

Acknowledgements

- Anirban Mandal, RENCI
- Nathan Gauntt, SNL
- U.S. DOE: XPRESS project & SciDAC SUPER Institute

Questions?

sriduttb@cs.unc.edu