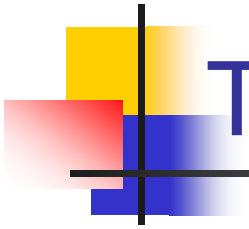
PxP: Performance x Power Optimizations for Sparse Scientific Computing

Padma Raghavan, Mary Jane Irwin, Mahmut Kandemir
Suzanne Shontz, and Jia Li

Sarah Conner, Yang Ding, and Konrad Malkowski (Ph.D.)
Department of Computer Science Engineering
The Pennsylvania State University

Collaborations
Argonne National Lab (McInnes and Norris), Lawrence Berkeley National Lab (Ng, Shalf and Simon)

http://www.cse.psu.edu/~raghavan
Supported by NSF –STHEC, other ..

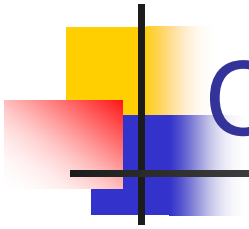


Trends at the Petascale

- Gordon Moore, 1966: 2 X # transistors in 18 months, ..., Patrick Gelsinger, 2004: *power is the only real limiter...'* DAC Keynote

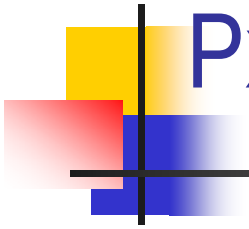
- HPC/SC Scaling = Parallelism at multiple levels
  - ILP, CMP, CMPxMPPs+Network, algorithm, aplication
  - H/W scaling= ½ ~1 Petaops peak 1/2 years

- Power at Petascale = 4 ... 20 Megawatts
  - Energy Bill @$.1 KWH = $10M+/year



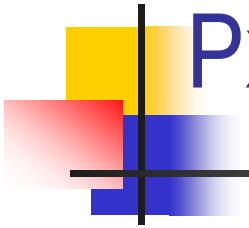
Ops/S vs Ops/J

- Ops/S Peak ~ CPU frequency (f), observed for dense LAPACK, TOP500
- Factor 10 to 1000 gap between peak and sustained rates on real workloads (Loft at NCAR, Simon at NERSC, NSA, ..)
- Most apps are memory, network, I/O bound
  - Low ops/data load typical of sparse codes with O(N) computational costs
- **PxP goal: Multilayer adaptivity for energy-aware supercomputing**



PxP Supercomputing

- Characterizing *power reductions* and *performance improvements* *x* quality x cost tradeoffs
  - Utilizing power control modes of the CPU, memory, network
  - Developing/utilizing optimizations to improve performance
  - Leveraging interactions between code tuning/phases & h/w
  - Utilizing application/algorithm/implementation trade-offs for quality and performance
- **QxPxP** from single processor, to CMPs, to MPPs
- Tools and environments for adaptive feature/method/mode tuning for QxPxP optimizations



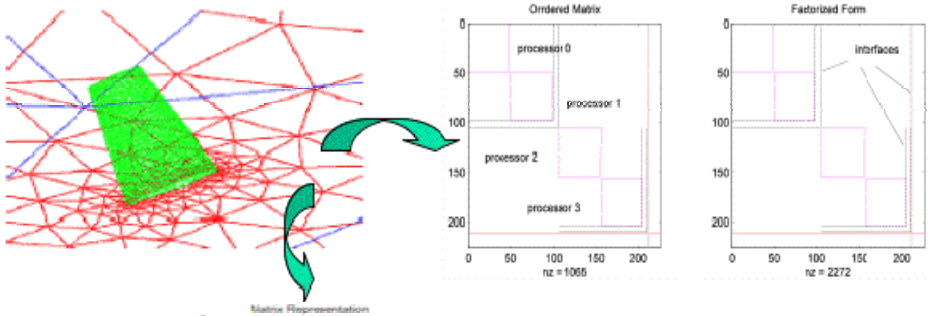
PxP For Sparse Scientific Codes

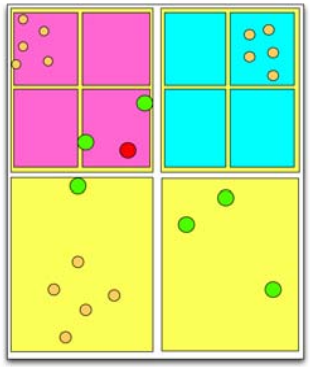
- Sparse codes (irregular meshes, matrices, graphs), unlike tuned dense codes, do not operate at peak rates (despite tuning)
- Sparse codes represent scalable formulations for many applications but ...
  - Limited data locality, poor data re-use
  - Memory and network latency bound
  - Load imbalances despite partitioning/re-partitioning
  - Multiple algorithms, implementations with different quality/performance trade-offs
- Present many opportunities for adaptive QualityxPowerxPerformance tuning

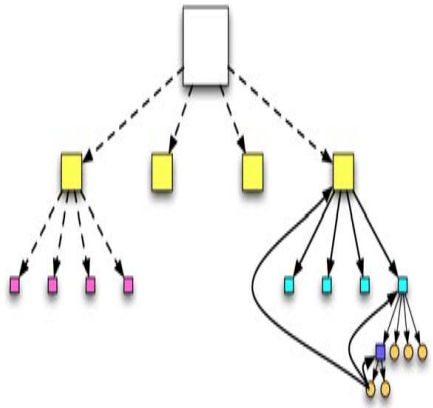
Sparse Codes and Data

Interoperable, Sparse Data Structures and Transformations

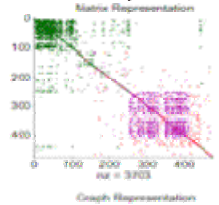
Discretizing with adaptive meshes



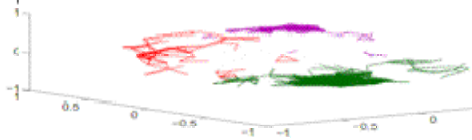




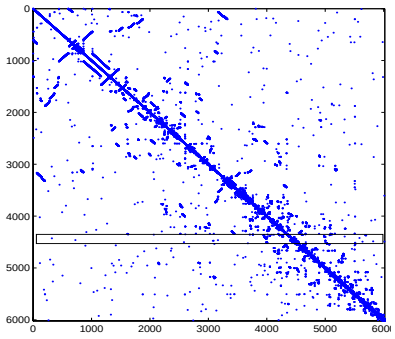
Graphs, networks, clusters in low dimensional space

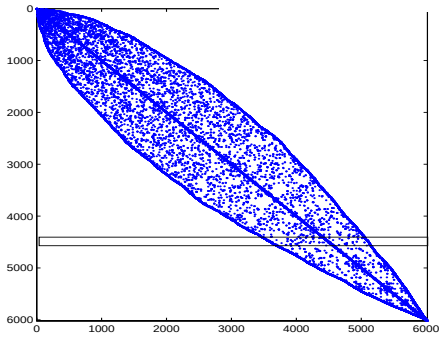


Sparse matrices, automatically ordered to reveal hierarchical structure

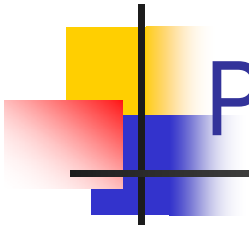


Sparse structures enable linear scaling of computational cost with problem size, parallelism,..



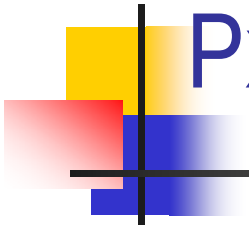


Application requiremets-> algorithm selection + tuning -> H/W, S/W adaptivity



PxP Recent Results

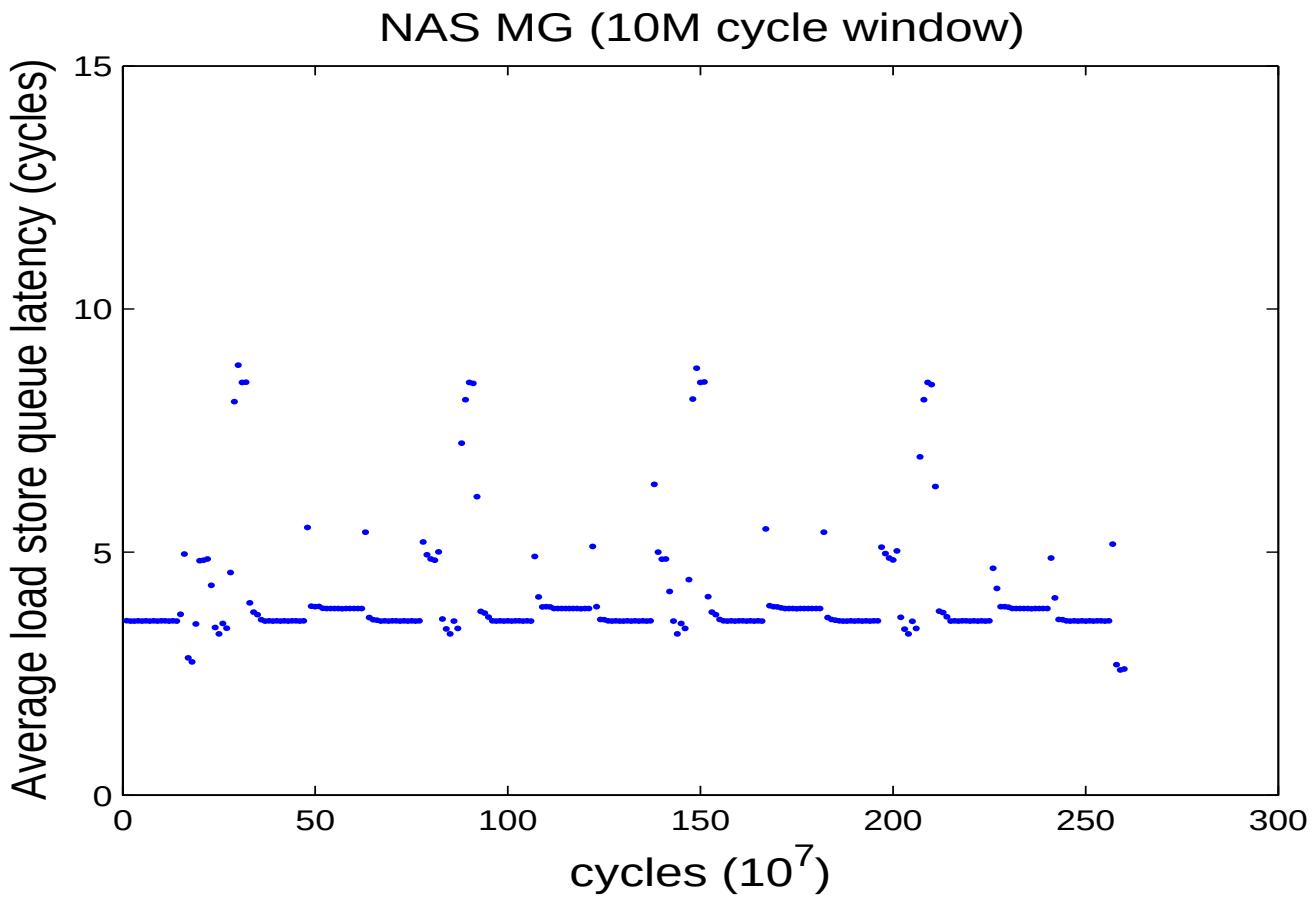
- PxP through adaptivity
  - Single CPU s/w phase-aware h/w adaptivity
  - MPP network link shutdown adaptivity in collective communications
  - CMP: adapting to reduced processor availability
- Methodology: Simulation based including, Simplescalar, Wattch and Cacti, SIMICS + new tool TorusSim



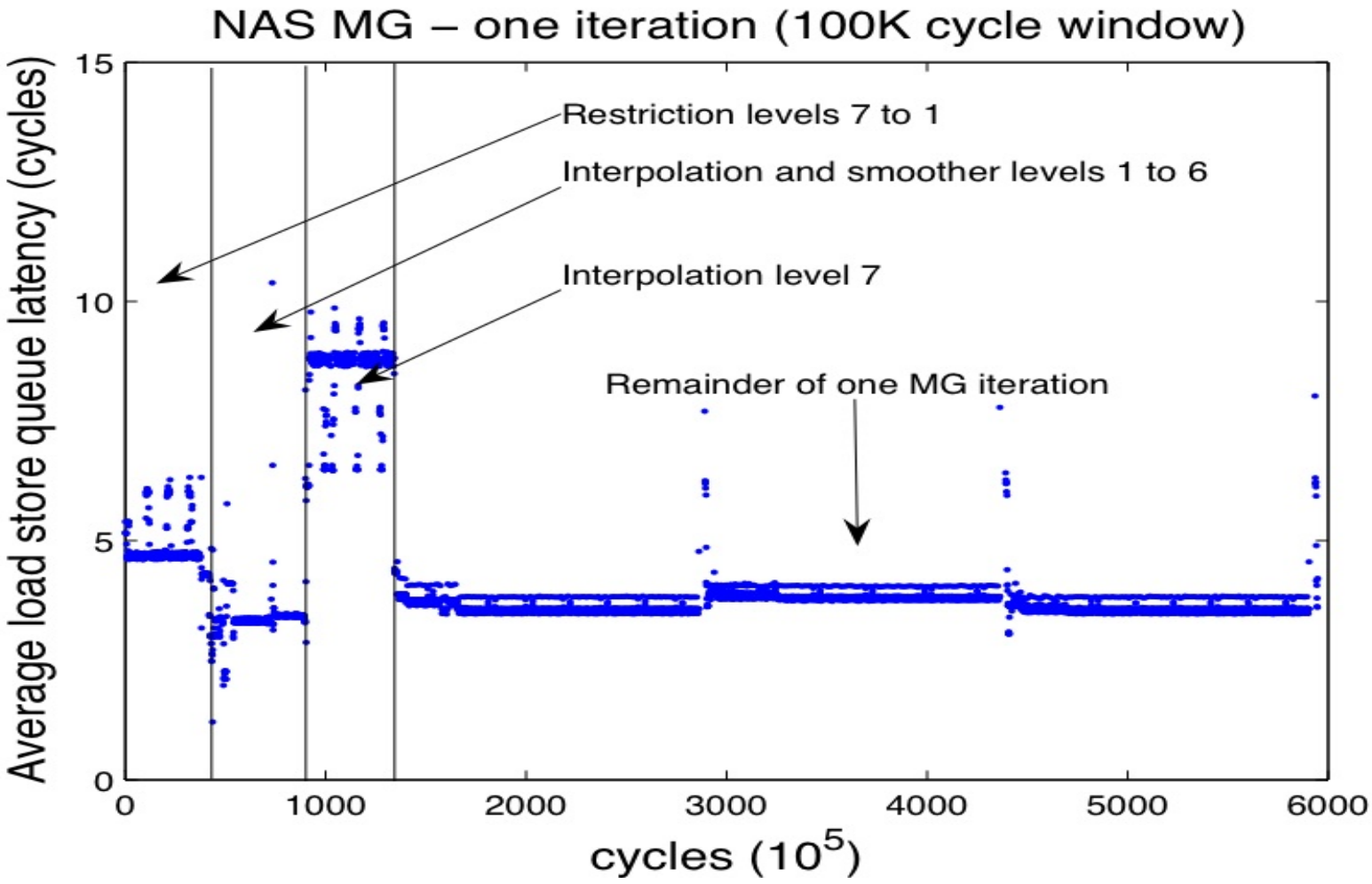
PxP Results – I: CPU+Memory

- Different S/W phases can benefit from different H/W features
- Challenges:
  - How do known s/w phases correspond to h/w detectable phases?
  - What H/W metric can be used to detect phase change? (lightweight)
- Goals:
  - Reduce power subject to performance constraint
  - Reduce time subject to power constraint

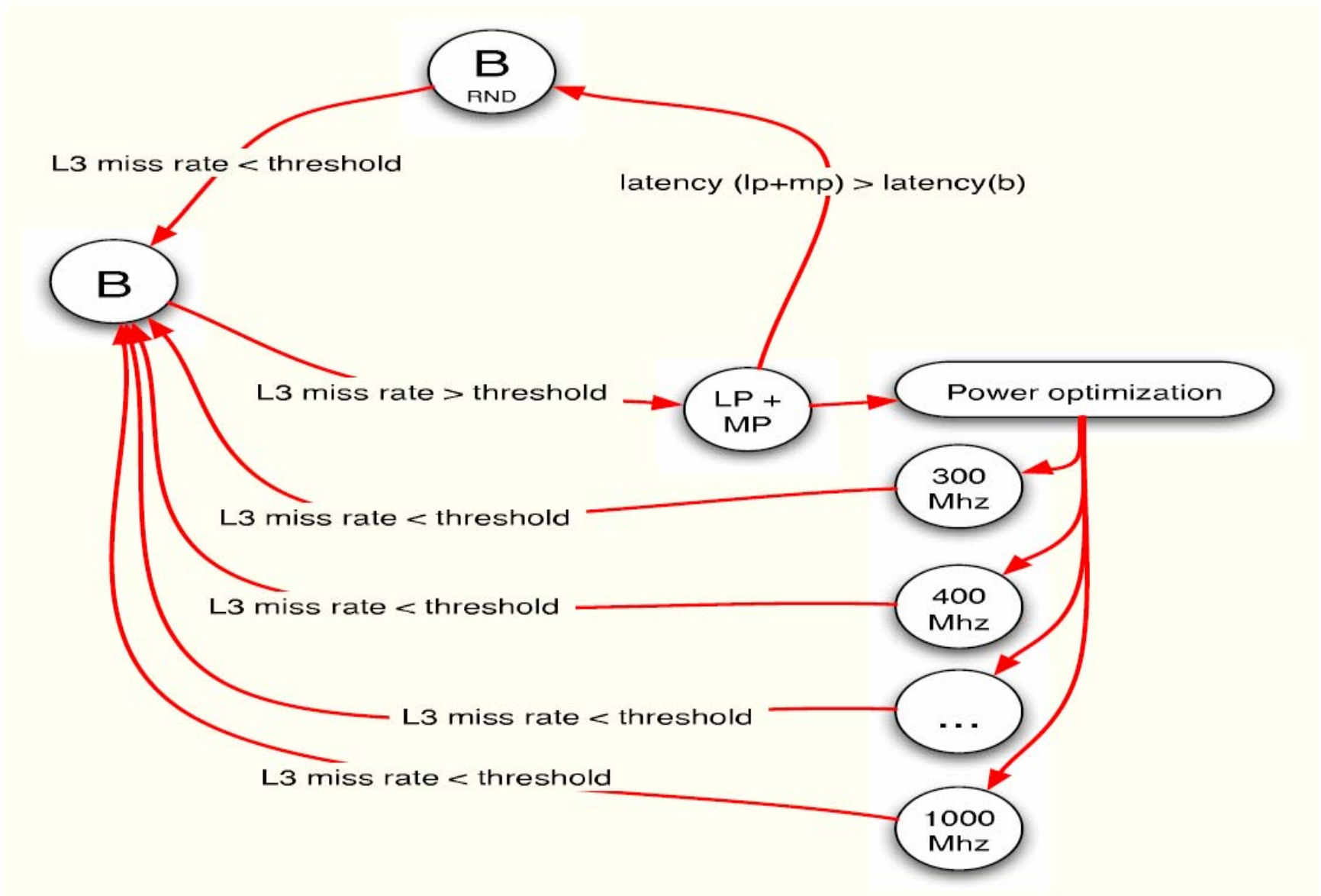
NAS MG: LSQ and 10M cycle window



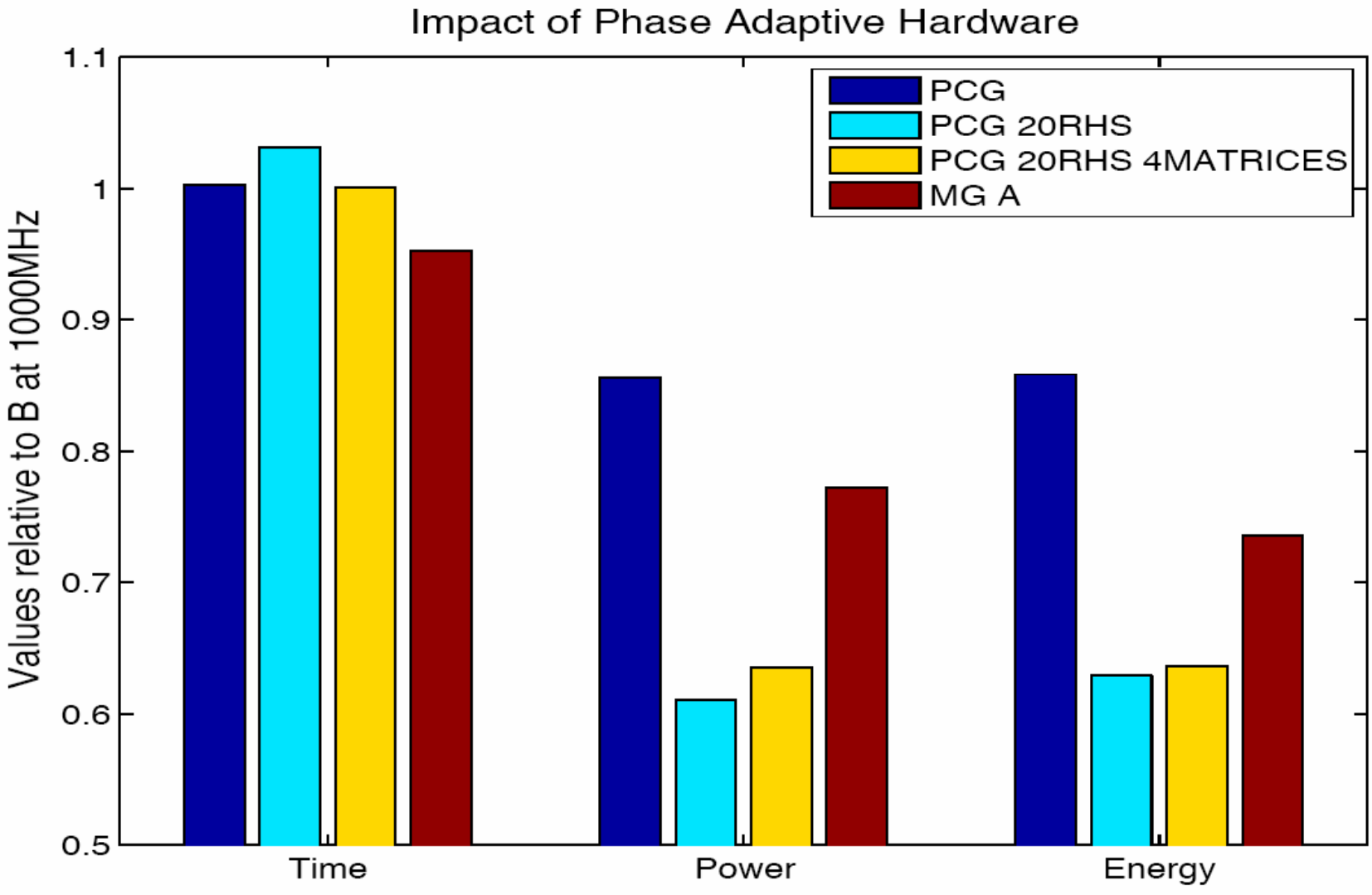
NAS MG: LSQ and 100K cycle window



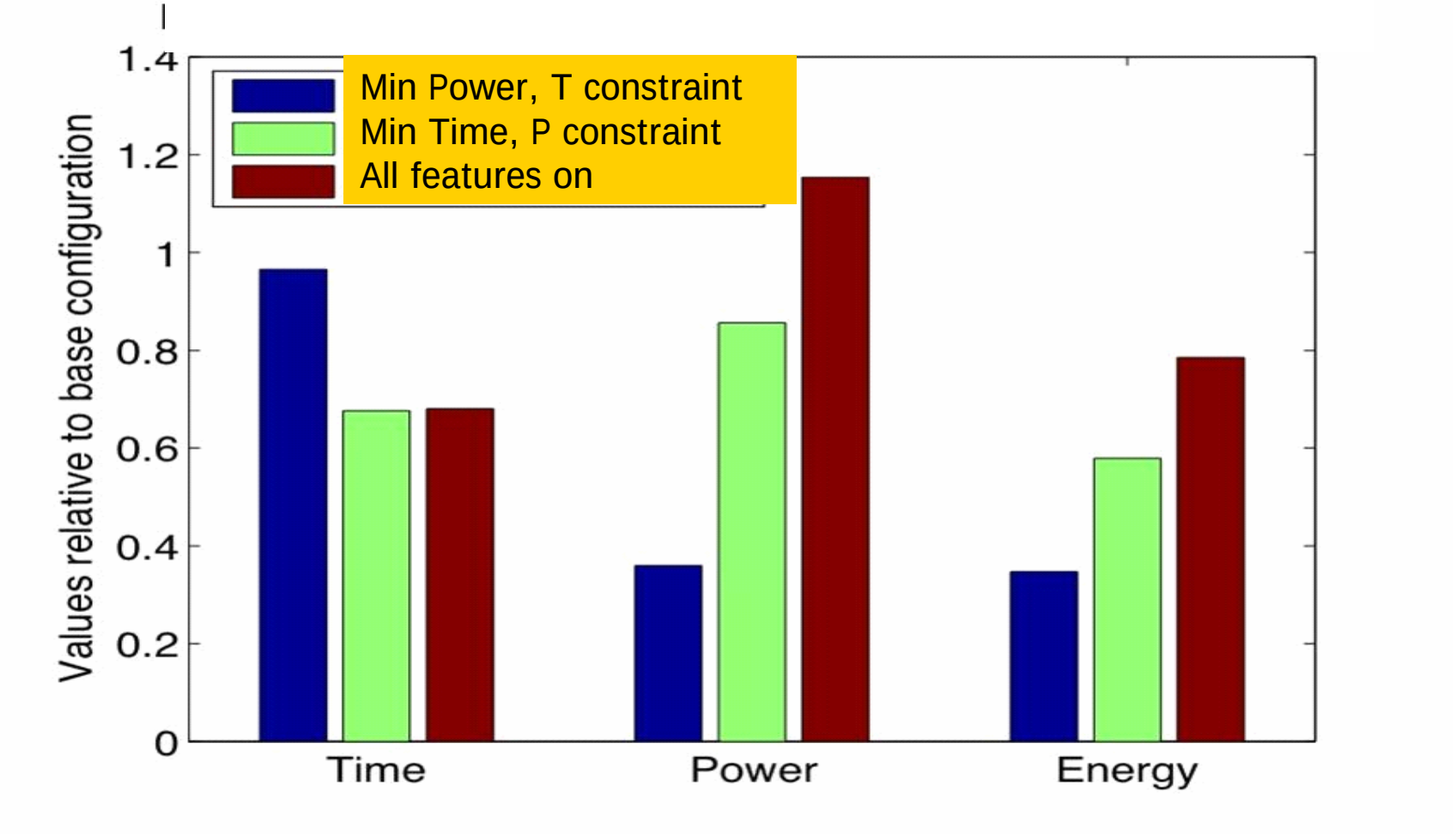
FSM For H/W Adaptivity



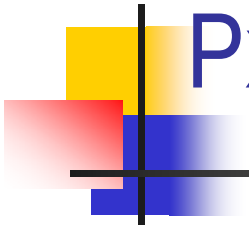
Impact of Adaptivity: T constraint



All vs Adaptive summary



PxP Results – II: MPP Networks



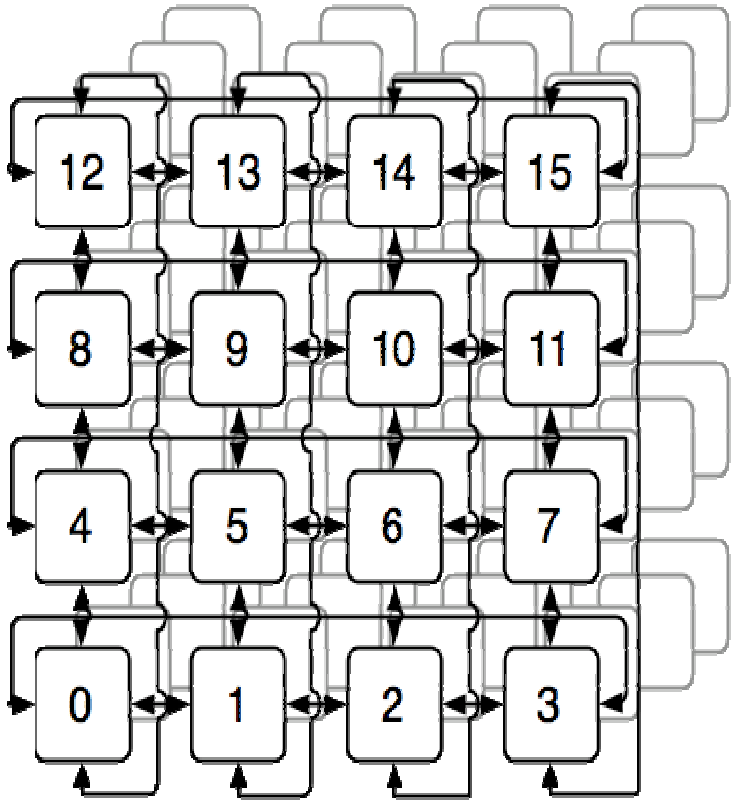
- HPC codes: compute, compute, compute, communicate; compute, compute, ...
- Network link shutdown can save energy during compute phase
- Can network link shutdown save energy even during communication phase, e.g., for collective communication?

TorusSim

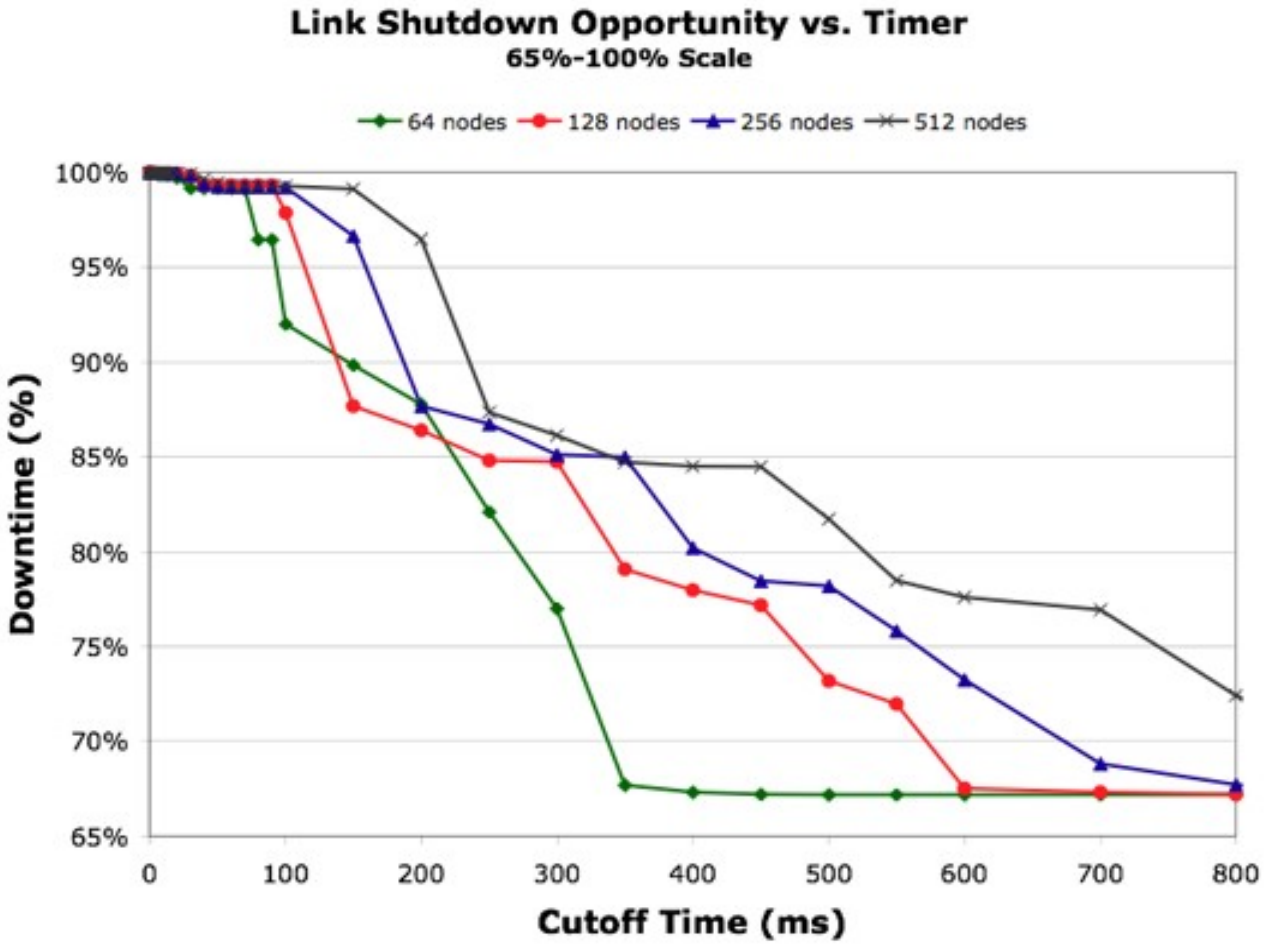
•TorusSim: Simulator models network energy and performance
•Tracks performance and energy statistics
• Simulates large nets — (2563) 1-, 2-, and3-D toruses and meshes, like BlueGene/L

• Identifies link shutdown opportunities with a cut off timer

• Simulates in minutes for real traces, and is deadlock-free under realistic network conditions.

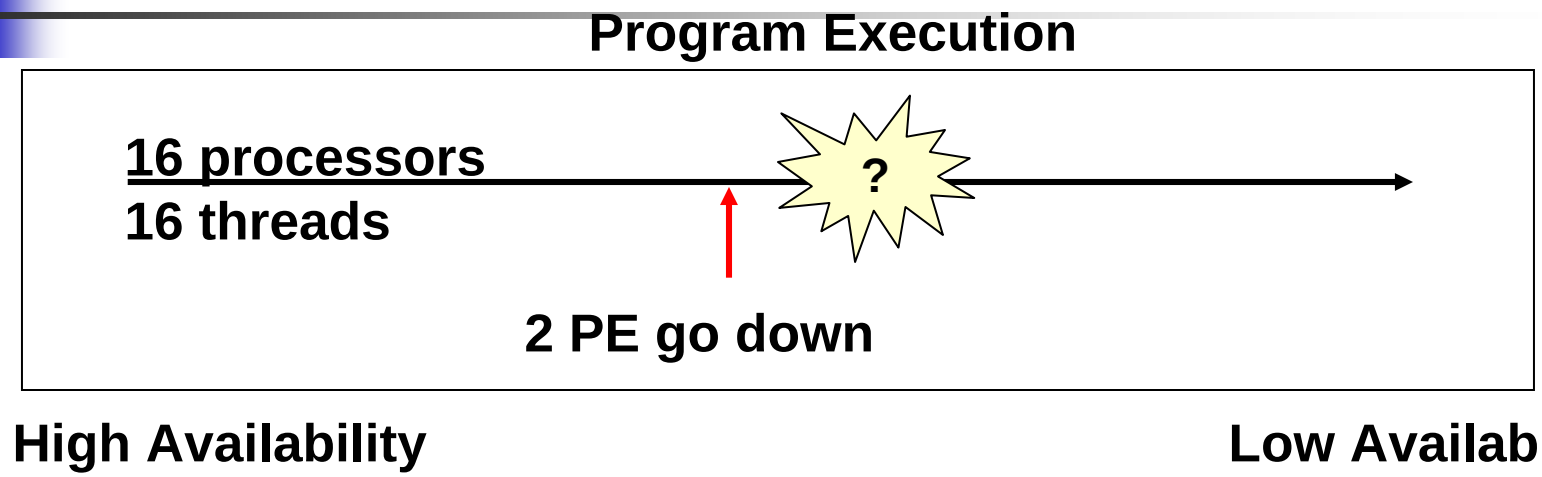


Link Shutdown vs Cutoff Time (Reduce)



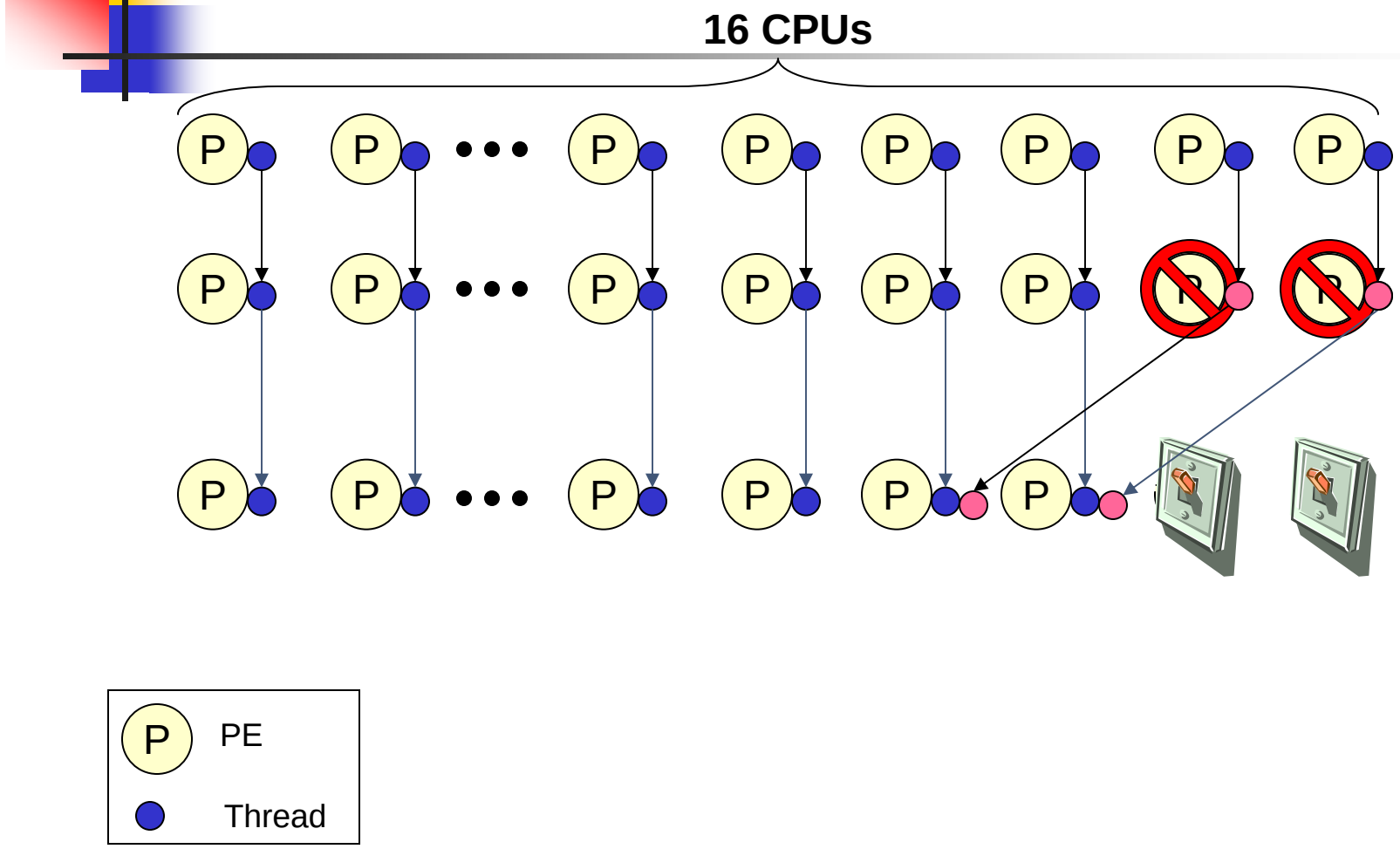
- Many links remain unused. For reduce, it's 66%.
- Implement simple link shutdown (LS) hardware in the net
  - MPI library X inf LS hardware can permit optimal collective communication shutdown exploitation

PxP Results – III: CMPs

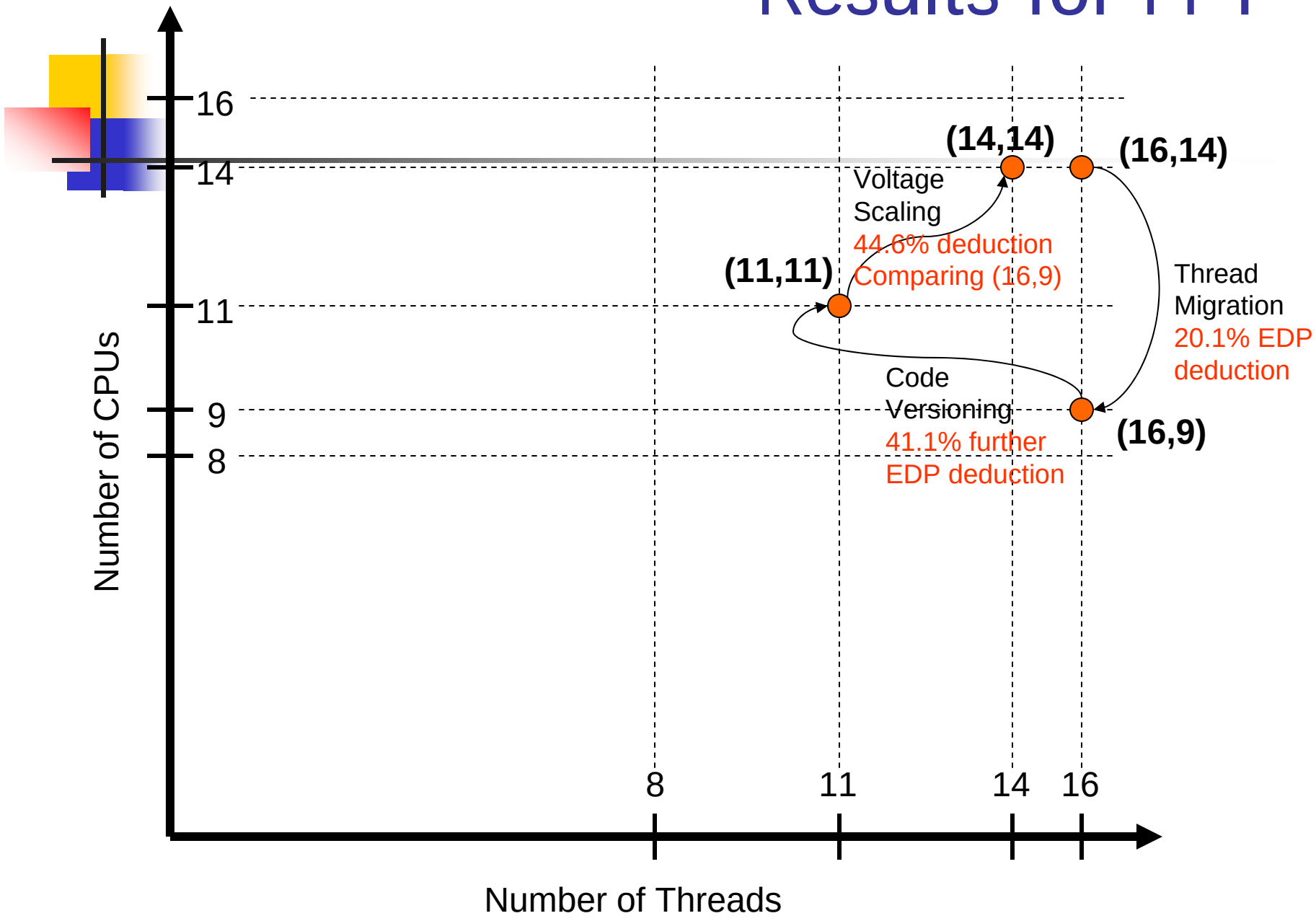


How to allocate processors and maps threads to handle runtime processor availability changes for PxP?

Adapting to Low Availability



Results for FFT



Summary and Challenges

- Large potential for HPC QxPxP optimizations

- AESxSMA challenge: Need for formalisms and optimization methodology and their incorporation into runtime systems
  - **Applications, algorithms <- libraries**, language/compiler **-> architecture**
    - **Applications, algorithms: models** of quality, parallelism, scaling as f(N,P,Err)
    - **Architecture: models** of multiple components, at different scales, hybrid eg.x CMP+MPP, PxP features
    - **Libraries,languages**: at many levels, for many functions/goals
  - **Methodology:** multi-objective **stochastic** optimization in high dimensional QxPxP parameter space