ACMP: An Architecture to Handle Amdahl's Law

M. Aater Suleman
Advisor: Yale Patt
HPS Research Group

Acknowledgements

Eric Sprangle, Intel

Anwar Rohillah, Intel

Anwar Ghuloum, Intel

Doug Carmean, Intel

Background

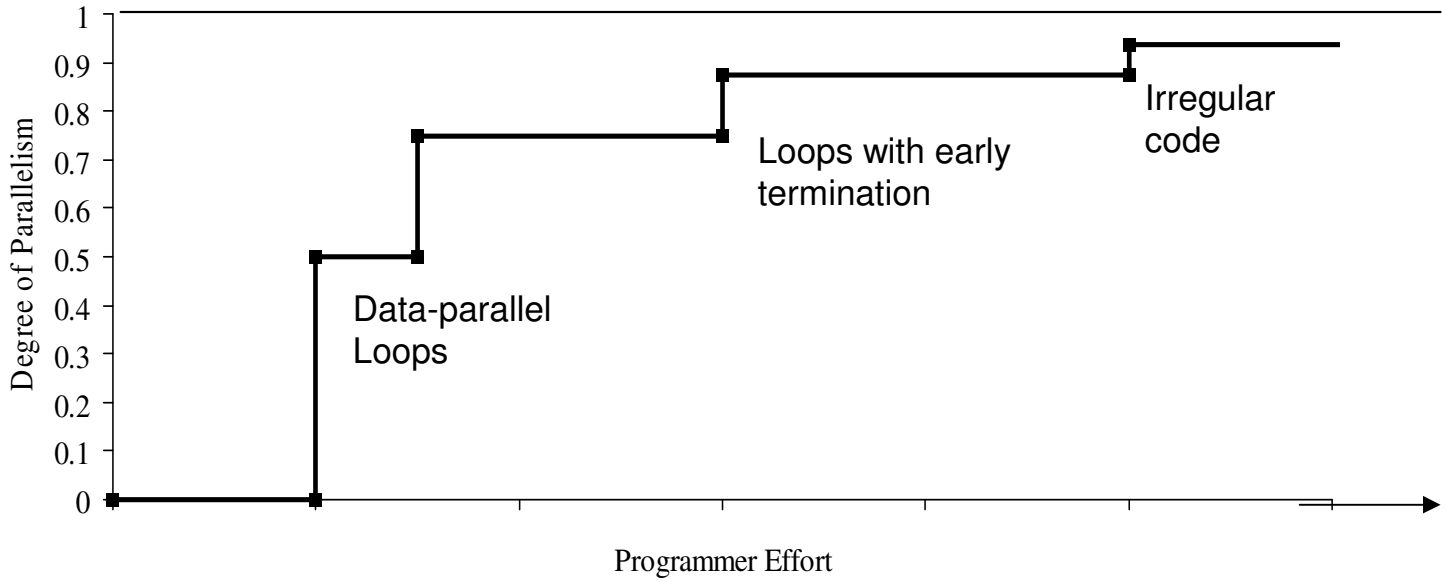
- Single-thread performance is power constrained
- To leverage CMPs for a single application, it must be parallelized
- Many kernels cannot be parallelized completely
- Applications likely include both serial and parallel portions
- Amdahl's law is more applicable now than ever

Serial Bottlenecks

- Inherently serial kernels

```
For I = 1 to N
        A[I] = (A[I-1] + A[I])/2
```

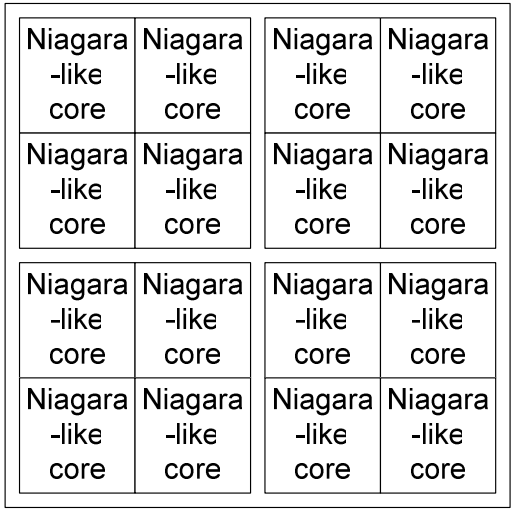
- Parallelization requires effort



CMP Architectures

- Tile small cores e.g. Sun Niagara, Intel Larrabee
  - High throughput on the parallel part
  - Low serial thread performance
  - Highest performance for completely parallelized applications
- Tile large cores e.g. Intel Core2Duo, AMD Barcelona, and IBM Power 5.
  - High serial thread performance
  - Lower throughput than Niagara

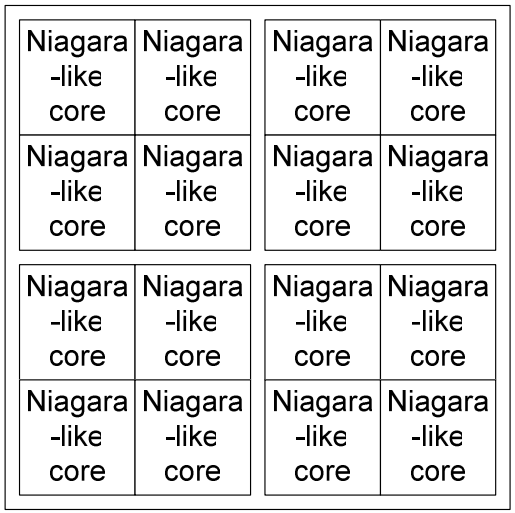
ACMP



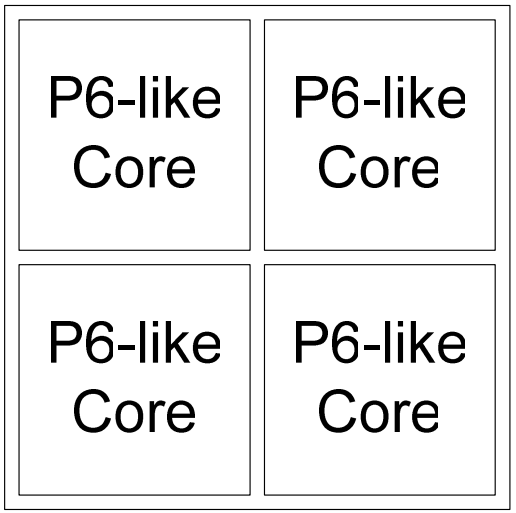
"Niagara" Approach

- Run serial thread on the large core to extract ILP
- Run parallel threads on small cores

ACMP



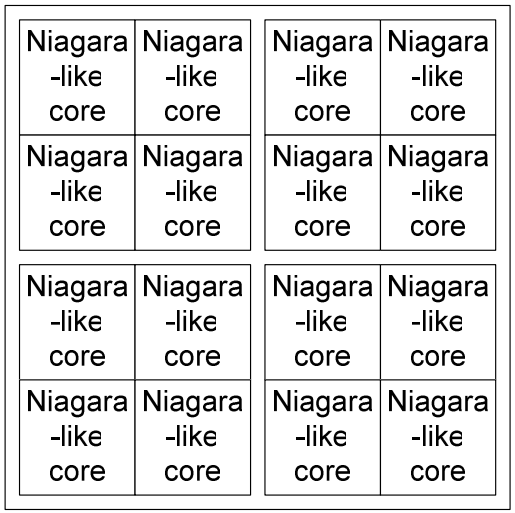
"Niagara" Approach



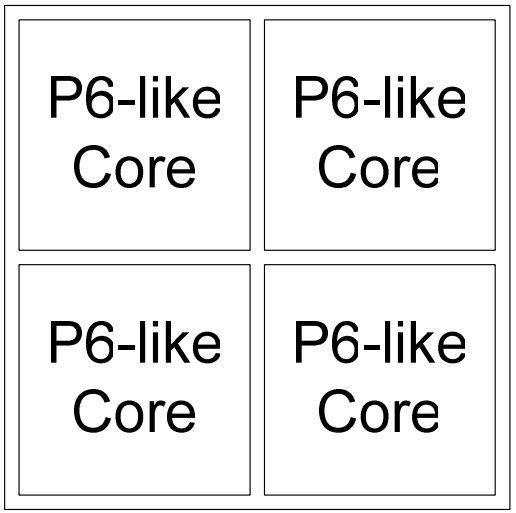
"Tiled-P6" Approach

- Run serial thread on the large core to extract ILP
- Run parallel threads on small cores

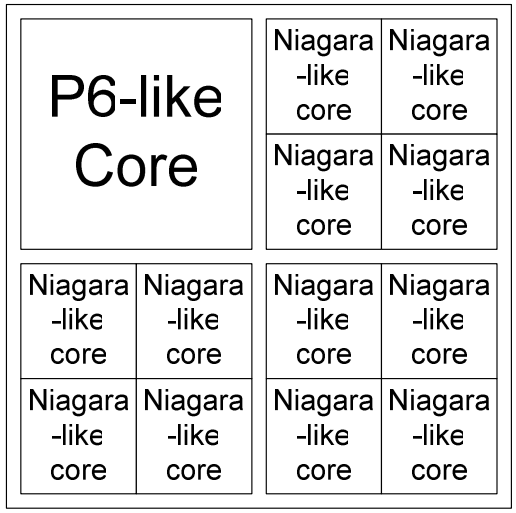
ACMP



"Niagara" Approach



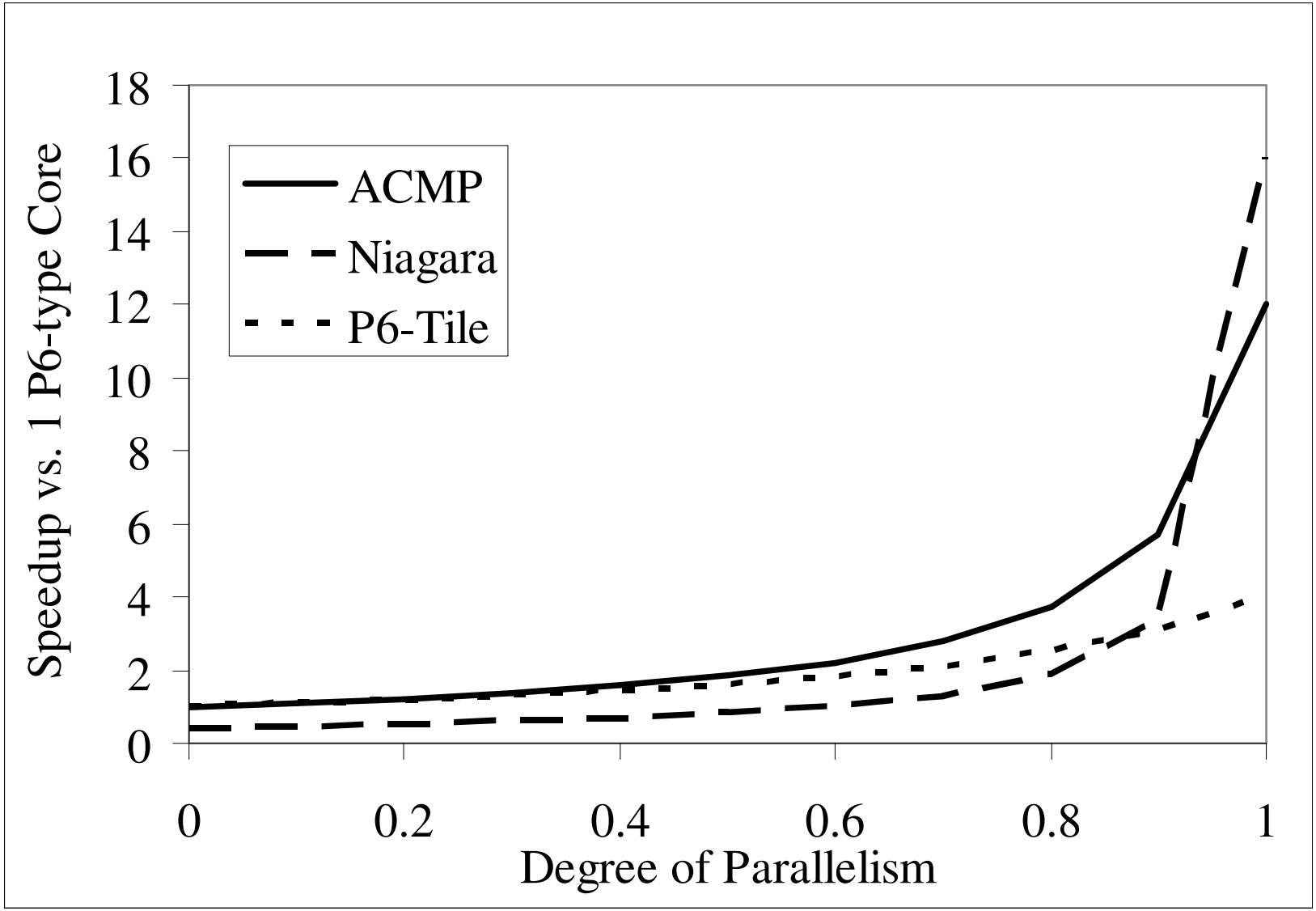
"Tiled-P6" Approach



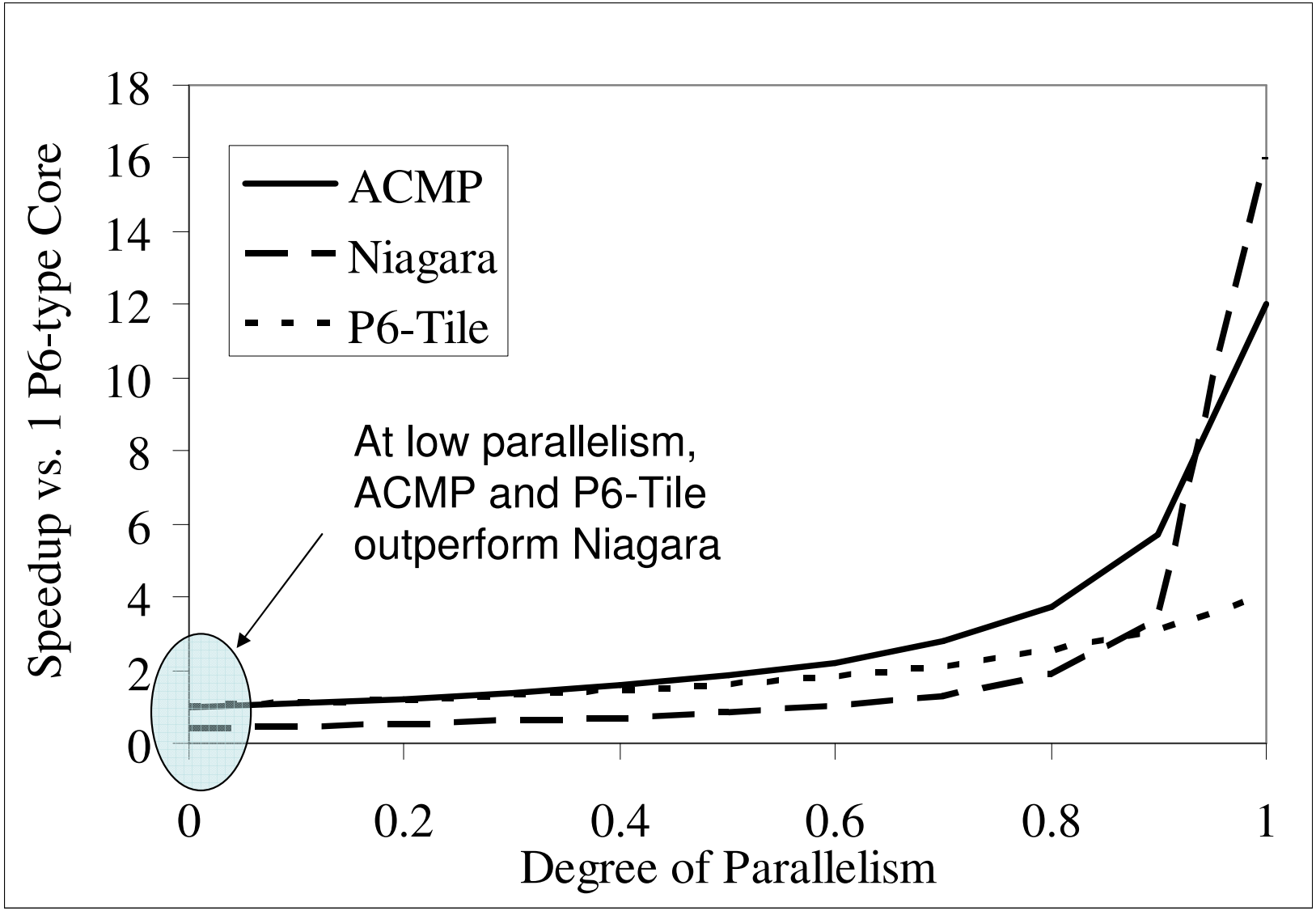
ACMP Approach

- Run serial thread on the large core to extract ILP
- Run parallel threads on small cores

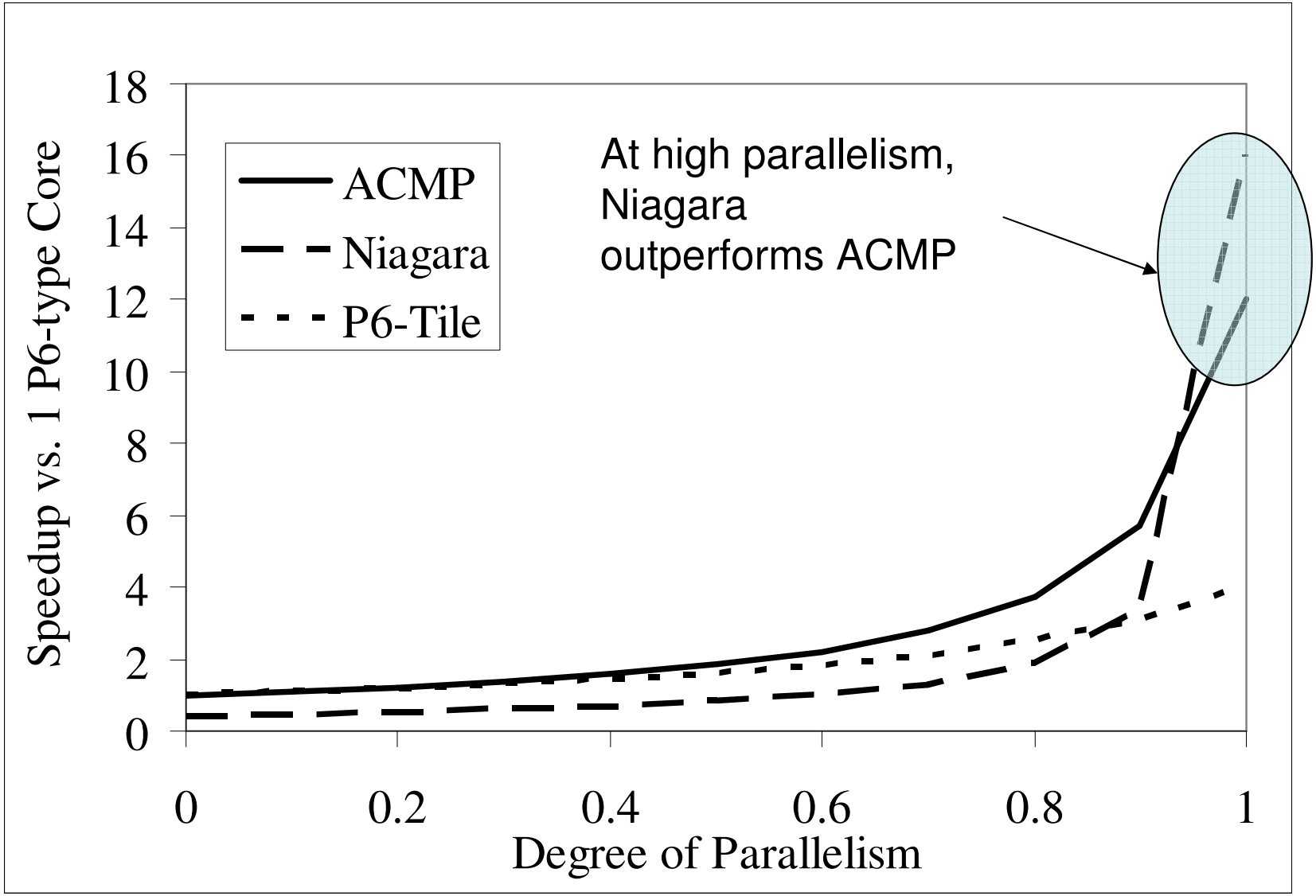
Performance vs. Parallelism



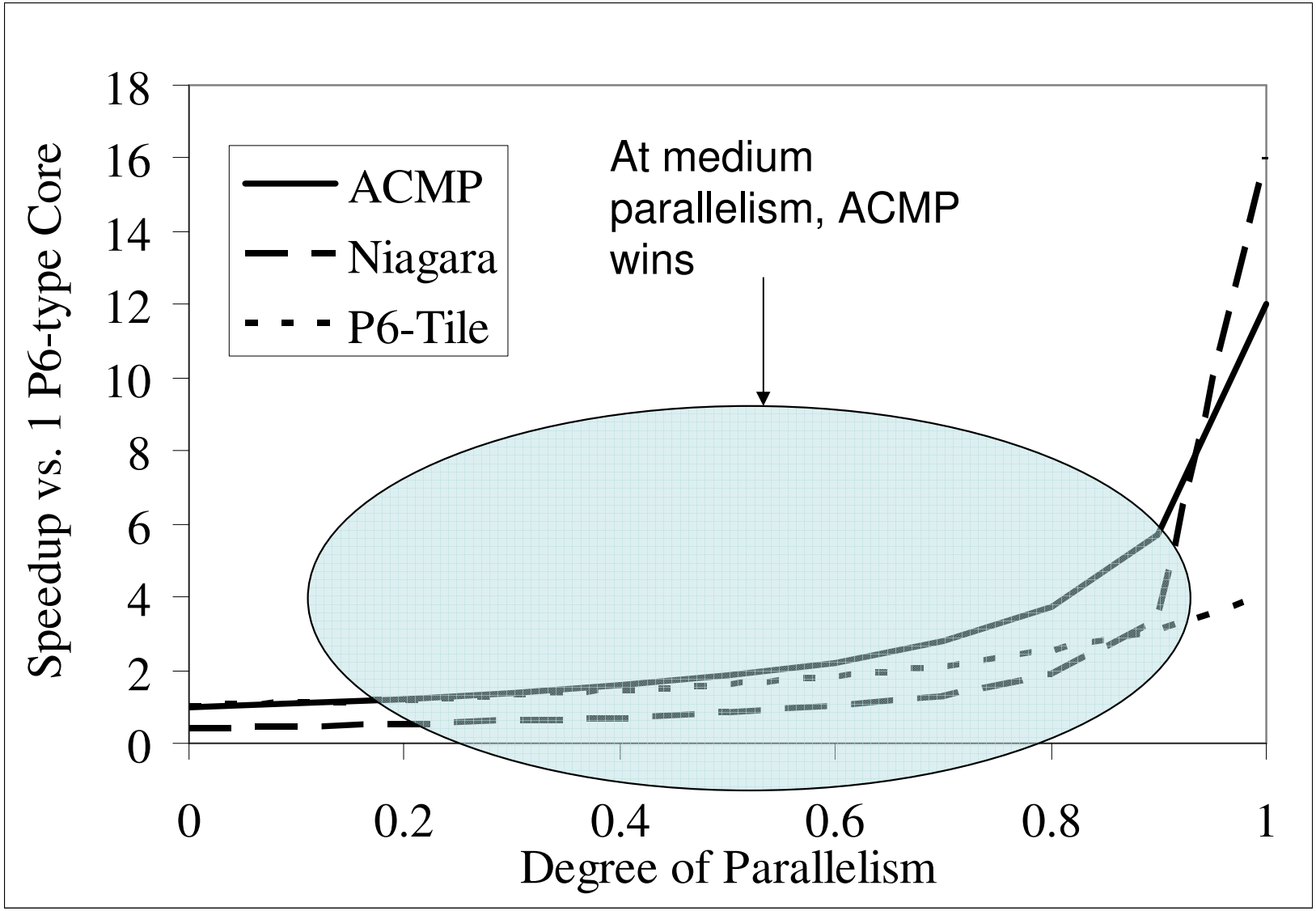
Performance vs. Parallelism



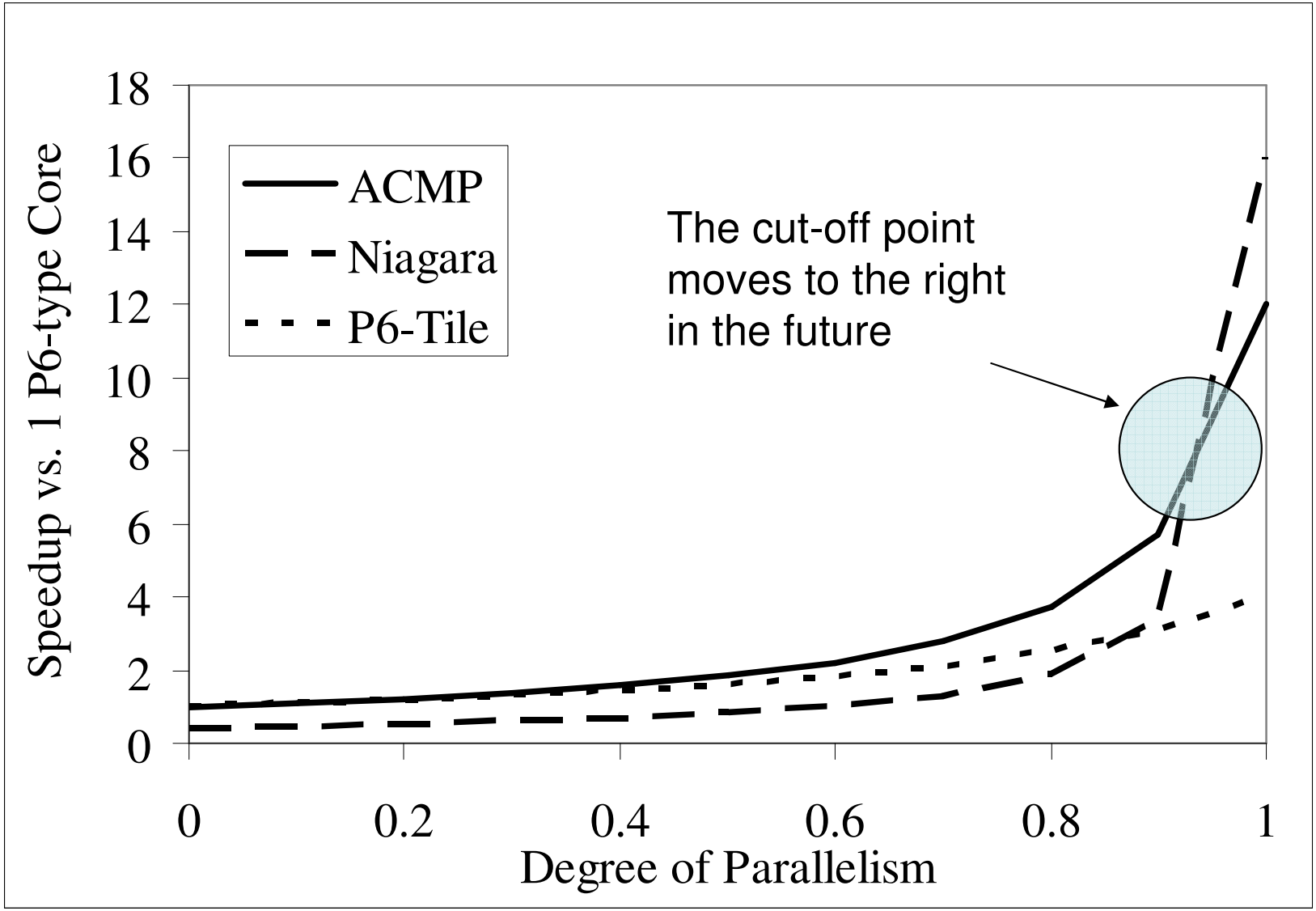
Performance vs. Parallelism



Performance vs. Parallelism



Performance vs. Parallelism

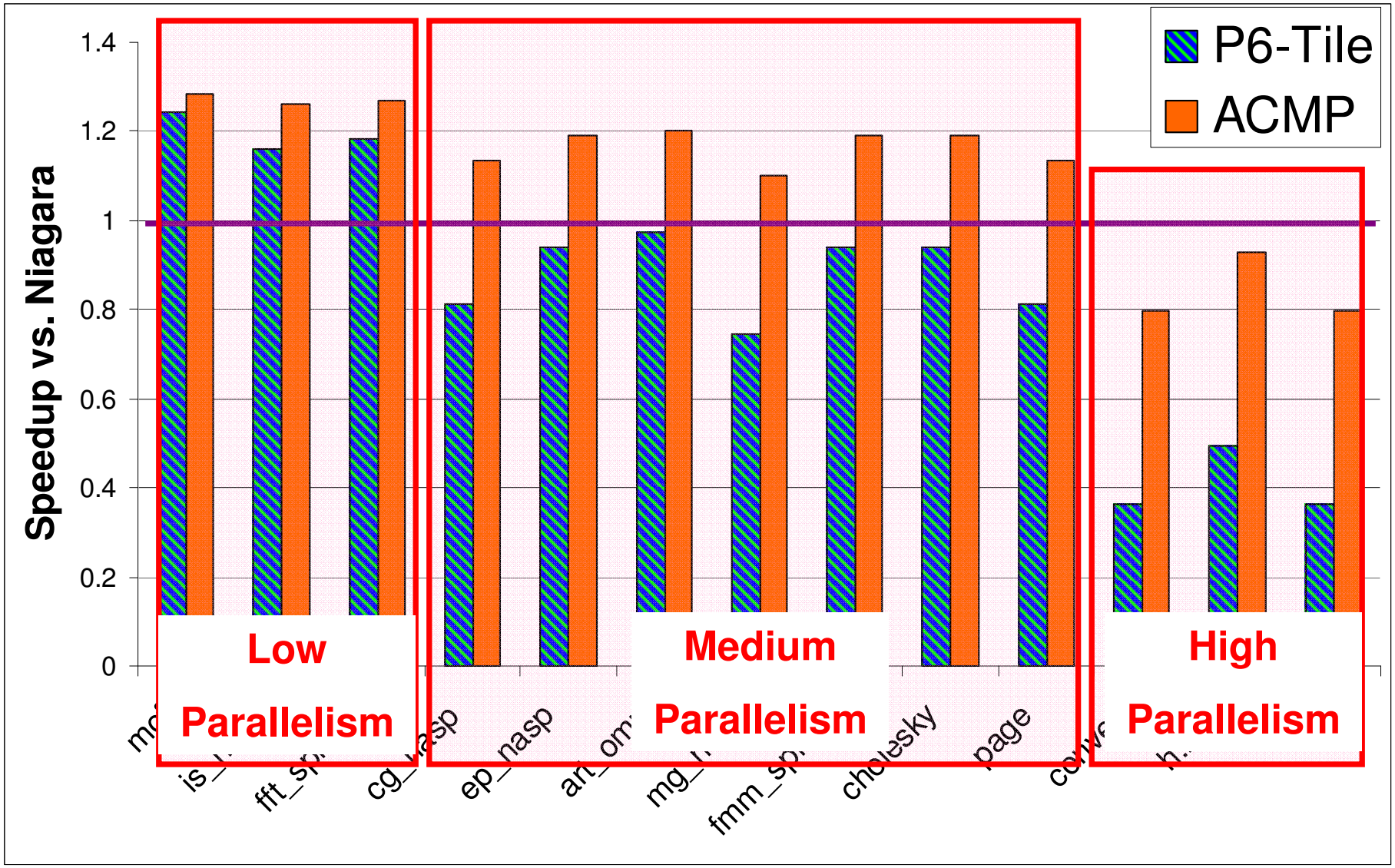


Experimental Methodology

- Large core: Out-of-order (similar to P6)
- Small Core: 2-wide, In-order
- Configuration:
  - Niagara: 16 small cores
  - P6-Tile: 4 large cores
  - ACMP: 1 Large core, 12 small cores
- Single ISA, shared memory, private L1 and L2 caches, bi-directional ring interconnect

- Simulated existing multi-threaded applications without modification
- ACMP Thread Scheduling
  - Master thread → large core
  - All additional threads → small cores

Performance Results



Summary

- ACMP trades peak parallel performance for serial performance
- Improves performance for a wide range of applications
- Performance is less dependent on length of serial portion
- Improves programmer efficiency
  - Programmers can only parallelize easier-to-parallelize kernels

Future Work

- Enhanced ACMP scheduling
  - Accelerate execution of finer-grain serial portions (critical sections) using the large core
  - Requires compiler support and minimal hardware
- Improved threading decision based on run-time feedback

Thank you