Exploring the Design Space of Future CMPs

Jaehyuk Huh, Stephen W. Keckler, and Doug Burger,
in Proceedings of the 2001 International Conference on Parallel Architectures and Compilation Techniques (PACT 2001)

Summary

- Design space exploration of chip multiprocessors
  - Chip area and performance tradeoffs
  - how many processing cores?
  - core complexity
    - in-order or out-of-order
  - how big on-chip caches?
  - limited off-chip bandwidth will become an increasing problem
    - (#on chip transistors)/(# chip signal pins) will increase by a factor of 45 between 180 and 35 nanometer technologies.
      - need for larger caches to reduce bandwidth demand
  - applications with different access patterns require different CMP designs to maximize throughput

Introduction

- Job throughput in servers
- Future CMPs will have a larger # of cores
  - superscalar paradigm is reaching its limits
    - little use of more than 4 or 5-way issue superscalar processors
    - global wire delays will limit the area of the chip that is useful for a single conventional processing core
- Considers CMOS technology scaled to ultrasmall (35 nanometer) devices
- Power consumption not considered

Technology model

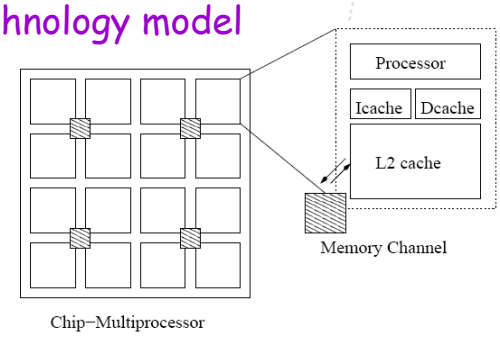


Figure 1. Chip-multiprocessor model.

Memory channel

- L2 cache connected to off-chip memory via a set of distributed memory channels
  - limited resource
    - time multiplexing
    - channel contention

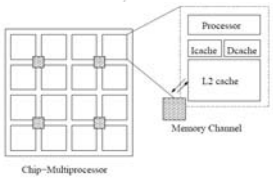


Figure 1. Chip-multiprocessor model.

Method

- throughput-oriented workloads with no sharing of data among tasks
- technology independent area models
  - found empirically
  - core area and cache area measured in cache byte equivalents (CBE)
- study the relative costs in area versus the associated performance gains --- maximize performance per unit area for future technology generations

Processor (core) models

Pin

Pout

	L2 cache size	2-way	4-way	8-way
In-order	128KB	0.20	0.21	0.21
	256KB	0.23	0.24	0.25
	512KB	0.24	0.25	0.25
	1MB	0.27	0.28	0.29
Out-of-order	128KB	0.26	0.31	0.33
	256KB	0.31	0.38	0.40
	512KB	0.32	0.39	0.41
	1MB	0.38	0.47	0.50

Table 1. Harmonic means of IPCs for six processor models.

Most area-efficient

What is harmonic mean?

- **Harmonic mean**
  - In mathematics, the harmonic mean is one of several methods of calculating an average. Typically, it is appropriate for situations when the average of rates is desired.
  - The harmonic mean (H) of the positive real numbers $a_1,\ldots,a_n$ is defined to be

$$H = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n}}.$$

[Wikipedia]

Smith, CACM oct. 1988

TABLE I. Performance of Three Computers on Two Benchmarks

Benchmark	Millions of floating pt. ops.	Computer 1 time (secs.)	Computer 2 time (secs.)	Computer 3 time (secs.)
Program 1	100	1	10	20
Program 2	100	1000	100	20
Total Time		1001	110	40

TABLE II. Performance of Benchmarks in Mflops

Benchmark	Computer 1	Computer 2	Computer 3
Program 1	100.0 mflops	10.0 mflops	5.0 mflops
Program 2	.1 mflops	1.0 mflops	5.0 mflops
Arith. Mean	50.1 mflops	5.5 mflops	5.0 mflops
Geom. Mean	3.2 mflops	3.2 mflops	5.0 mflops
Harm. Mean	.2 mflops	1.8 mflops	5.0 mflops

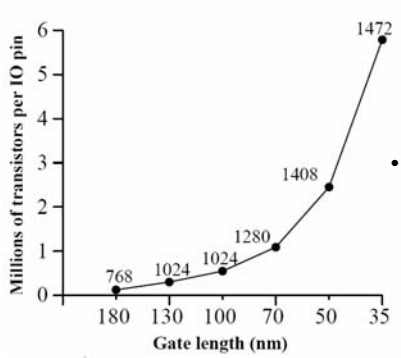
Techn. scaling and # of cores

Gate length	CBE (Megabytes)	λ^2 area	P_{IN}	P_{OUT}
100nm	7.6	1.60e+11	68	24
70nm	15.5	3.26e+11	139	50
50nm	30.5	6.40e+11	273	99
35nm	61.9	1.30e+12	556	201

Table 3. Total Chip Area.

- Assumes
  - a large fixed die of 400 *mm*^{2} (20 x 20 *mm*)
  - per core
    - 32 KB L1 I-cache and 32 KB L1 D-cache
    - no on-chip L2 cache

The I/O pin problem



- # I/O signaling
  - pins limited by physical tecnology
  - speeds have not increased at the same rate as processor clock rates
- Projections
  - from ITRS (International Technology Roadmap for Semiconductors)

Server workload

- maximize throughput
  - assume multiprogramming mode, i.e. independent threads, or several independent tasks to be executed
- P_i = performance of core i
- N_c = number of cores

$$P_{cmp} = \sum_{i=1}^{N_c} P_i$$

Application characteristics

- 10 SPEC-2000 app's + sphinx
- wide range of memory system behaviour
- three groups
  - Processor-bound
  - Cache-sensitive
  - Bandwidth-bound
- applications move among these three domains as the processor, cache, and bandwidth capacities are changed

Experiments

- Simplescalar tool set
- Effect of cache size on cache access latency given by the eCacti tool
- SimpleScalar modification to CMP
  - multiprogrammed SPEC workload
  - sharing of memory channels
  - DRAM as Rambus
  - avoid cold-start effects
    - skip first 5 billion instructions
  - then simulate 200 million instructions in detail for every application

Application memory requirements, uniprocessor

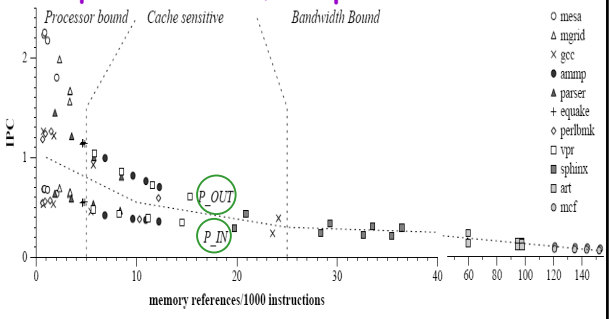
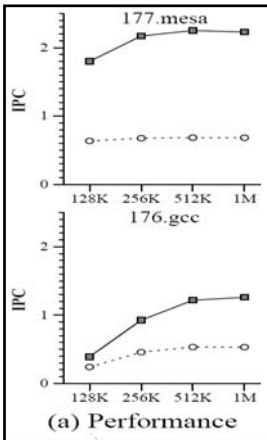


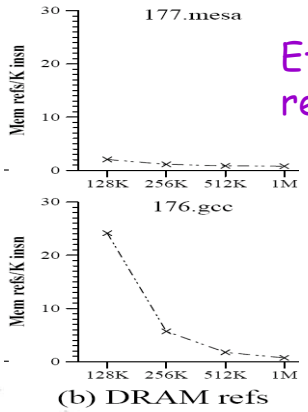
Figure 3. IPC versus rate of DRAM accesses.

Effect of varying L2 size on IPC



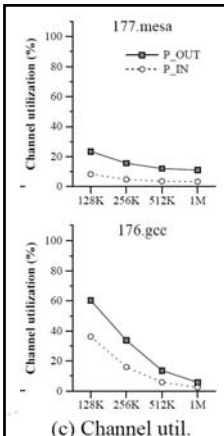
- P_out much better than P_in for processor bound apps
- Lesser difference for apps. with more frequent refs. to memory

Effect on mem ref frequency



- little effect for processor bound apps.
- large effect on cache-sensitive apps.

Effect on Channel Utilization



- cache size reduces demand
- P_out gives heavier demand on mem channels

Maximizing CMP Throughput

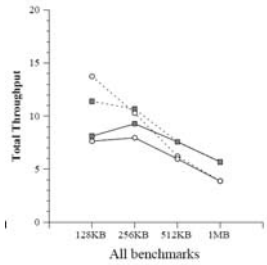
- Combine area analysis, performance simulations and technology projections
- Table 4, area analysis
  - # cores that will fit on a 400mm**2 chip in 70 nm techn.

	No. of cores		Cores/channel	
L2 cache size	P_{IN}	P_{OUT}	P_{IN}	P_{OUT}
No L2	139	50	6.6	2.4
64KB	89	41	4.2	1.9
128KB	65	35	3.1	1.7
256KB	42	27	2.0	1.3
512KB	25	19	1.2	1
1MB	13	11	1	1

Table 4. Number of cores and cores/channel.

Best configurations

- Figure 7
  - total throughput for all benchmarks
  - 256 KB L2 cache is the best overall solution

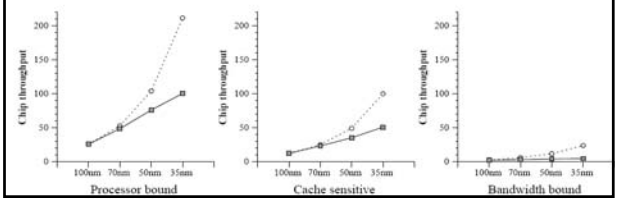


What about a shared L2 cache?

- ”the performance of a large monolithic L2 cache shared by a number of processors will sharply diminish with advances in fabrication processes and increases in clock rates, due to large cache bandwidth requirements and slow global wires”

Technology scaling

- Total chip throughput by best performance/area design at each techn.
- For proc. bound and cache sens. apps new technology gives improved performance and off-chip bandwidth becomes an increasing problem
- For bandwidth bound apps., new technology does not help!



Bandwidth constraints ☹

- ”Even for the ideal configurations, the large performance gap between limited and scaled channel organizations indicates that much of the throughput potential of future CMPs will go unrealized if solutions are not found to mitigate these bandwidth restrictions”
- “limited off-chip bandwidth will always constrain the maximum number of cores that can be placed on a chip”

Unused chip area – problem or possibility

- ”As applications become bandwidth bound, and global wire delays increase, an interesting scenario may arise. It is likely that monolithic caches cannot be grown past a certain point in 50 or 35nm technologies, since the wire delays will make them too slow. It is also likely that, **given a ceiling on cache size, off-chip bandwidth will limit the number of cores**. Thus, there may be useless area on the chip which cannot be used for cache or processing logic, and which performs no function other than as a placeholder for pin area. **That area may be useful to use for compression engines, or intelligent controllers to manage the caches and memory channels.”**
- **➔ go ahead -- innovate!**