

# Maximizing CMP Throughput with Mediocre Cores

J. D. DAVIS ET AL., STANFORD UNIVERSITY

MARISABEL GUEVARA  
JANUARY 25, 2010

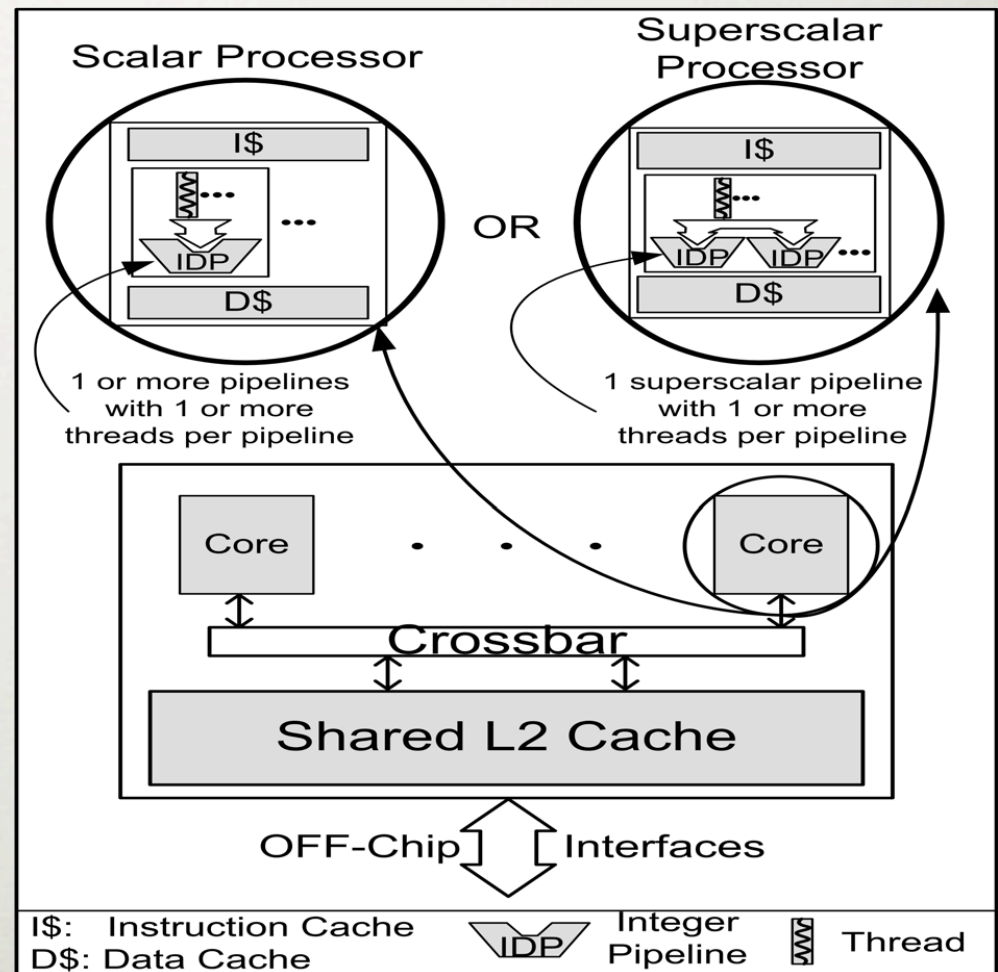
# MOTIVATION

---

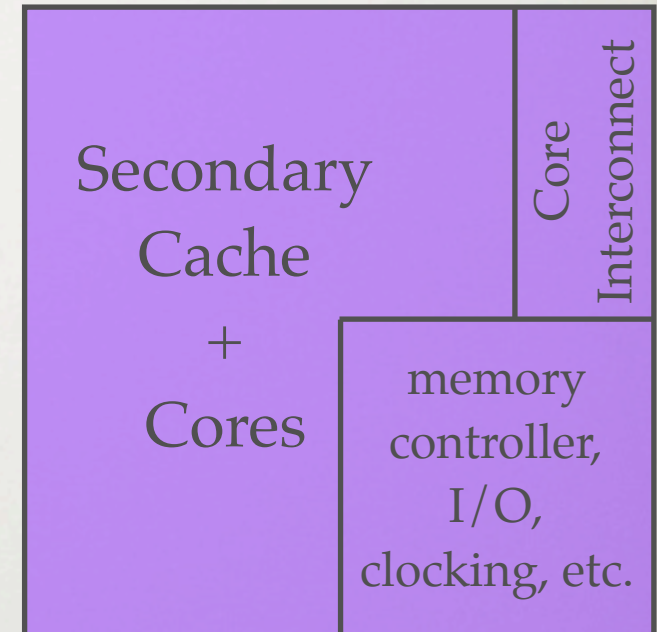
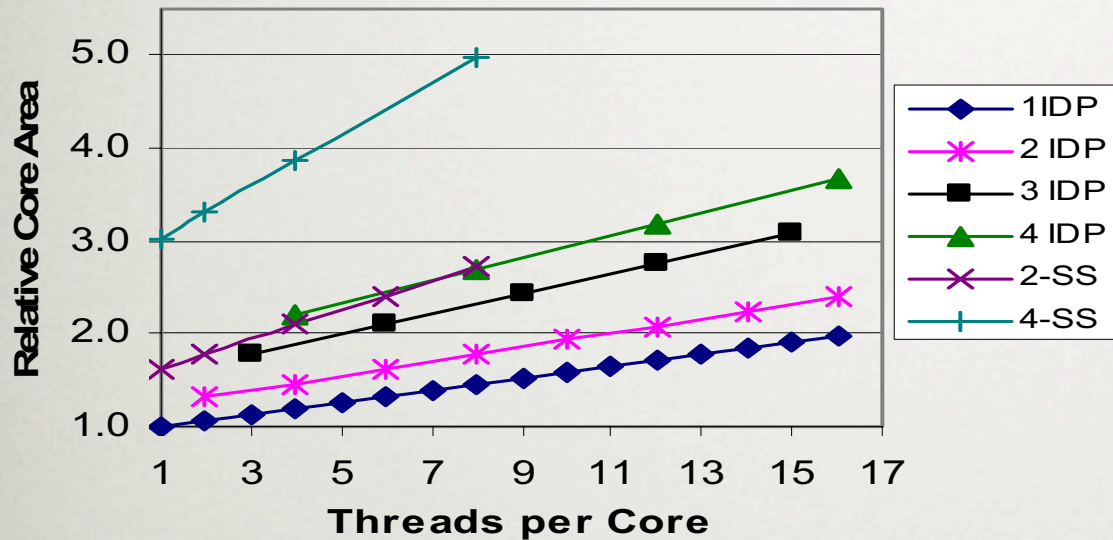
- Server applications are multithreaded, and aggregate throughput is more important than individual thread latency
- No-overhead multithreading prevents processor stalling due to low ILP or high cache miss rates
- Explore the CMT design space for equivalent area configurations

# CMT DESIGN SPACE

- Vary the number of IDPs and number of threads for each core
- Within each core configuration, vary cache size and organization
- Simulated a perfect secondary cache that provided insight into L1 cache utilization, to then guide the second simulation phase



# CORE AREA MODEL



- Developed from UltraSPARC processors (130-45 nm)
- Die area fixed at 400 mm<sup>2</sup>
- 5-6% core area increase per thread

# METHODOLOGY

---

- RASE (Rapid, Accurate Simulation Environment)
  - Built on SimCMT - cycle-based performance simulator modeling Niagara
  - execution-driven and trace-driven simulation
- Faster simulation
- No variability across test sequences
  - < 1% difference in IPC
  - < 5% difference in miss rates

2 modes:

1) Execution-driven mode: Simics issues insts and data references, and SimCMT replies with timing information.

+ Accuracy

- Long simulation time

2) Trace-driven mode: Run SimCMT with an instruction trace

+ Simulations speedup is ~20x

# RASE

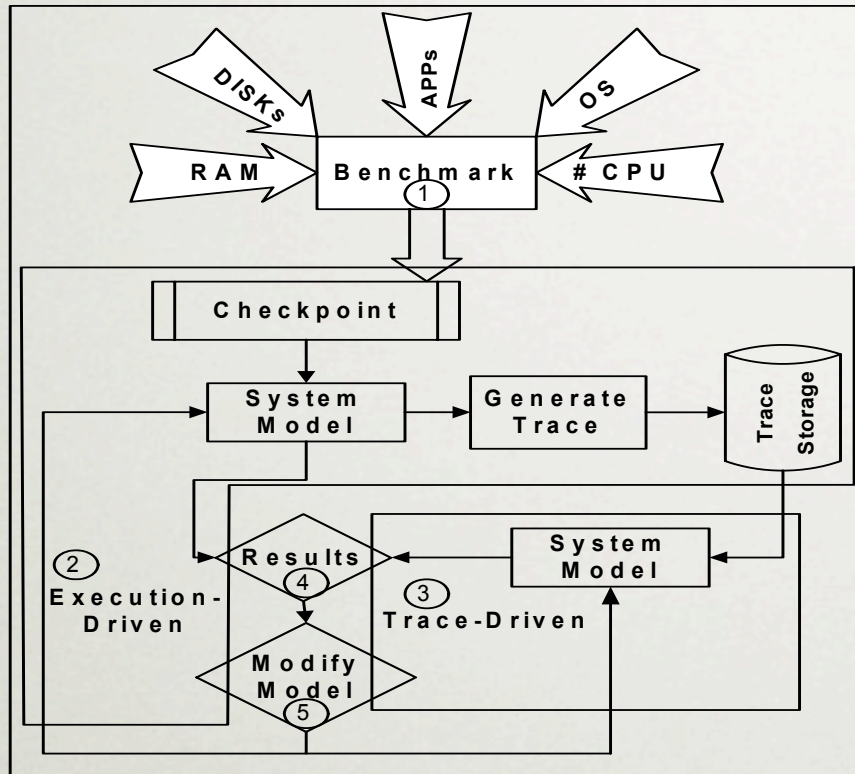


Figure 2. The steps required for trace generation and execution or trace driven simulation. (1) Tune the benchmark for the target hardware, (2) use execution-driven mode to reach the benchmark steady state, take a checkpoint, continue execution-driven simulation and/or generate trace file, (3) trace-driven simulation using input trace file and system model, (4) common result analysis framework, and (5) model modifications.

Source: Davis, J. D., Fu, C., and Laudon, J. 2005. The RASE (Rapid, Accurate Simulation Environment) for chip multiprocessors. *SIGARCH Comput. Archit. News* 33, 4 (Nov. 2005), 14-23. DOI= <http://doi.acm.org/10.1145/1105734.1105738>

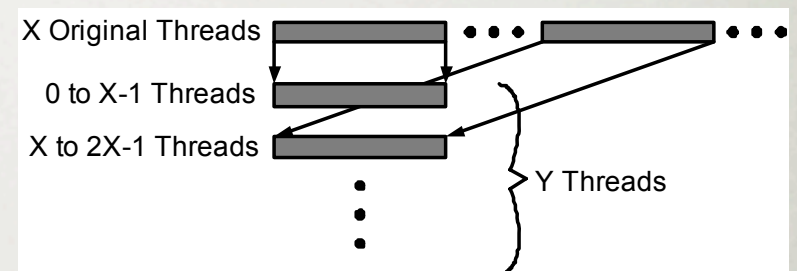


Figure 3. Thread replication from a trace with  $X$  threads for non-overlapping trace-driven execution of  $Y$  threads.

# DESIGN SPACE PARAMETERS

---

**Table 1: CMT design space parameters.**

Feature	Description
CPU	In-order scalar or superscalar
Issue Width	scalar, 2-way and 4-way superscalar
Pipeline Depth	8 stages
Integer Datapath Pipelines	1-4 IDPs or Integer ALUs
L1 D & I Cache	8KB-128KB, 16 (D) & 32 (I) Byte lines
L1 D & I Cache Set Assoc.	Direct-mapped, 2-, 4-, or 8-way
L1 D & I Cache Policies	write through, LRU-based replacement
Clock Frequency	1/3 -1/2 Maximum ITRS clock frequency [23]
Multithreading	1-32 threads/core
L2 Cache	1MB - 8MB, 128 Byte lines, banked (8 or 16), coherent, inclusive, shared, unified, critical word first, 25 cycle hit time (unloaded)
Main Memory	Fully Buffered DIMMs with 4/8/16 dual channels, 135 cycle latency (unloaded)

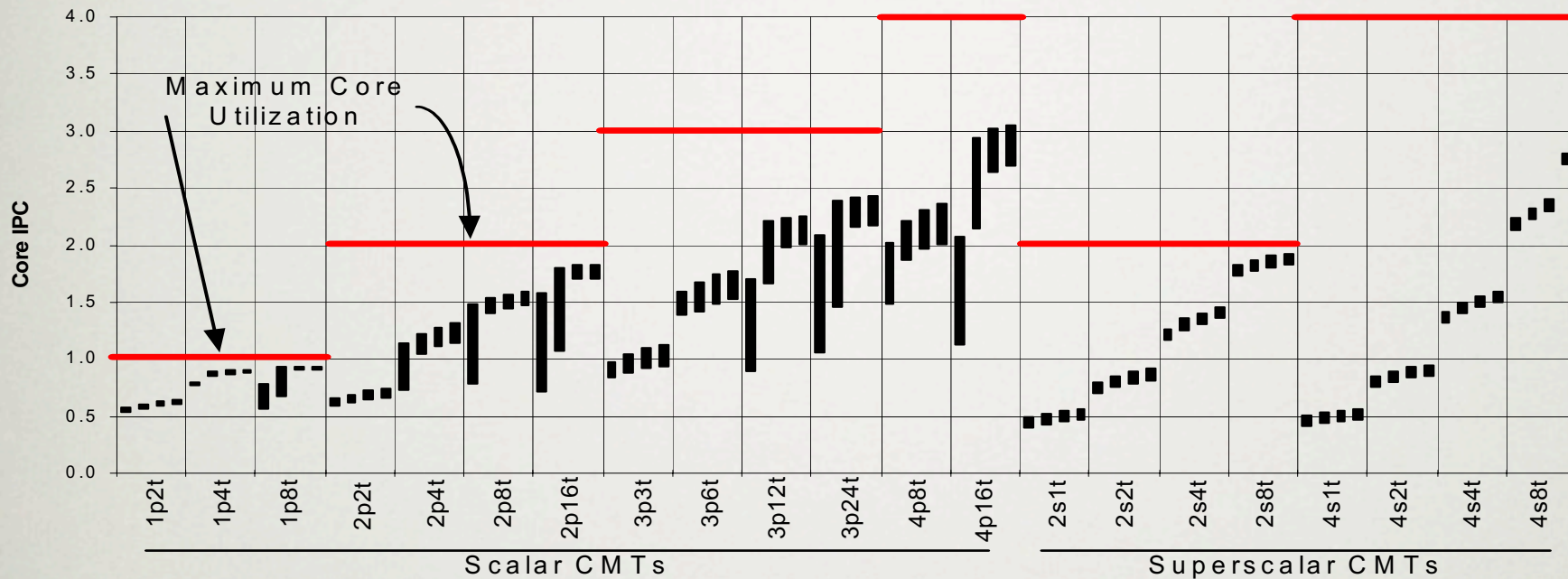
- 21 Scalar and Superscalar CMT core configurations
- Secondary caches of 25%, 40%, 60%, and 75% of CMT area

# BENCHMARKS

---

- SPEC JBB -- Java server-side performance
- TPC-C -- online transaction processing; HD, memory, and network resources are stressed
- TPC-W -- transactional web processing
- XML Test -- multithreaded XML parsing of trees

# SPEC JBB RESULTS FOR MEDIUM-SCALE CMTs



Scalar CMT cores outperform Superscalar CMT cores due to the additional scalar cores that fit the area budget

- “Overthreading” -- max IPC close to absolute peak  
Ex. 1p8t, 2p16t, 2s8t
- Insufficient secondary cache degrades performance  
Ex. 2p16t, 3p12t, 3p24t, 4p16t

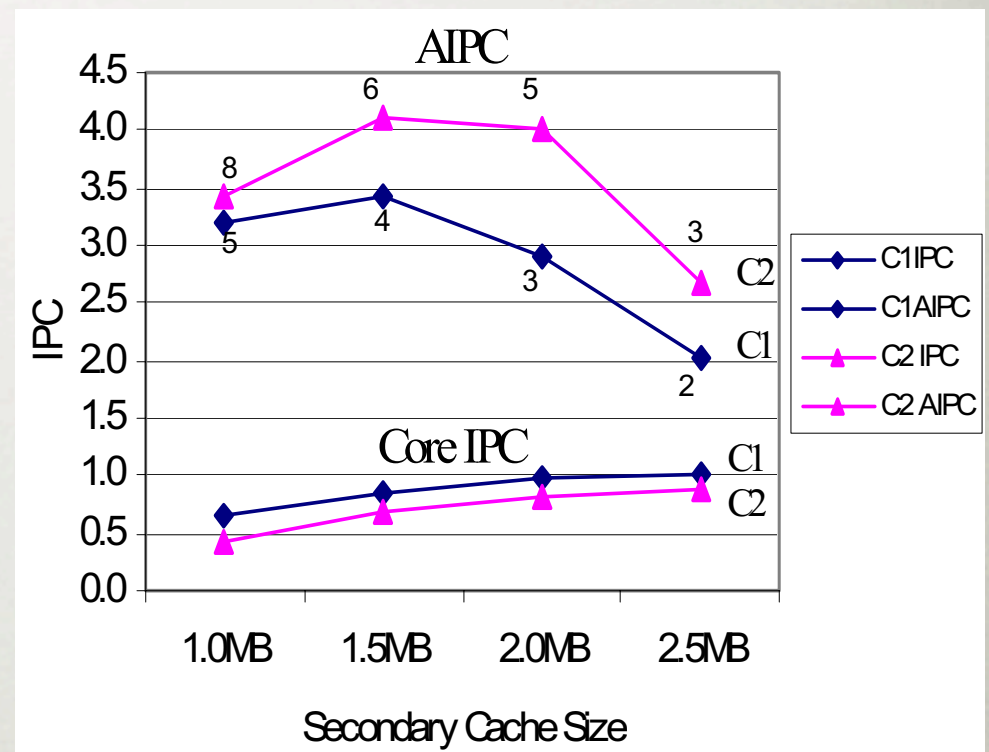
underthreading – single thread per pipeline results in no latency tolerance and low processor utilization

overthreading – too many active threads fully utilize the IDP and perf is insensitive to primary cache capacity or set associativity

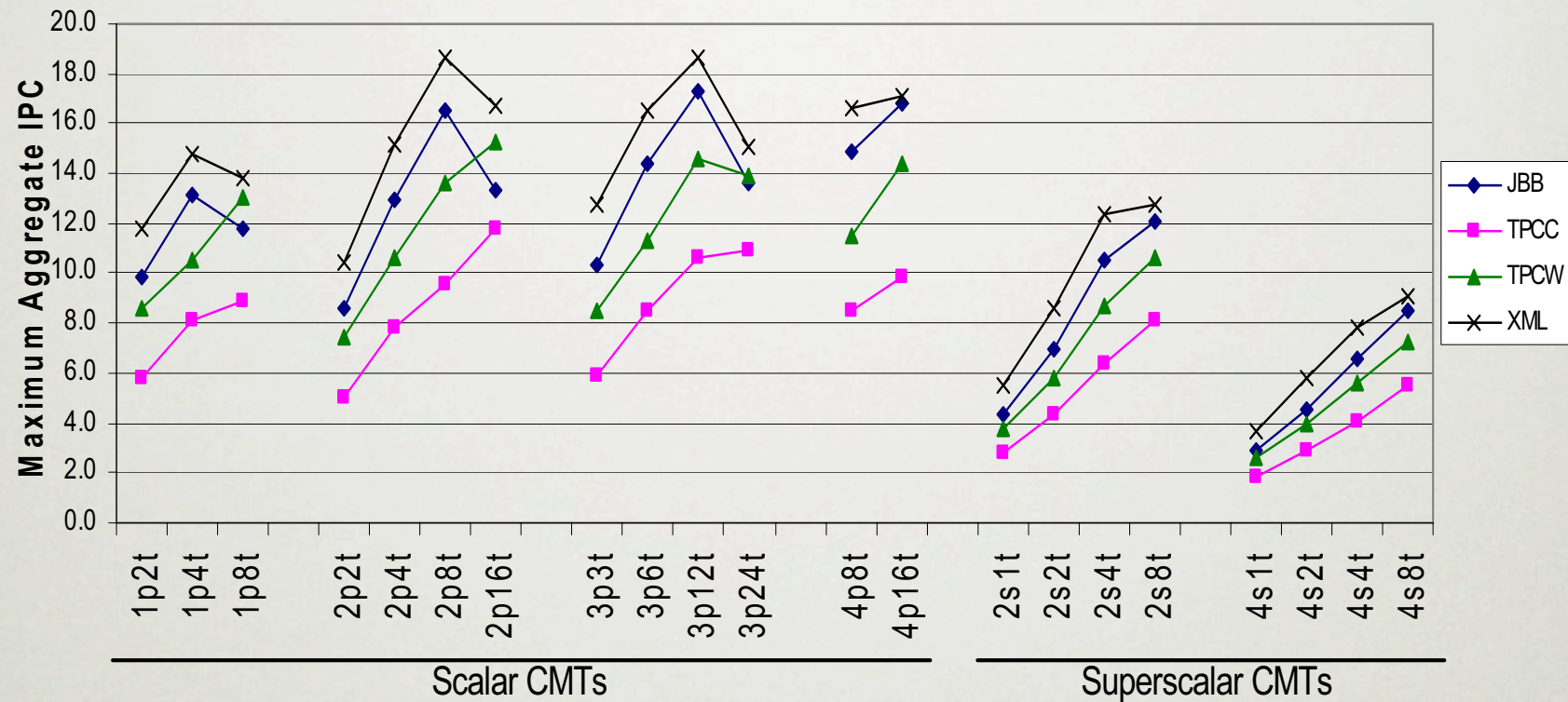
similar results for other benchmarks -- up XML Test, down TPC-W, TPC-C

# TPC-C RESULTS FOR SMALL SCALE 2P4T CMT

- AIPC underperforms due to the area limit
- C1 = best IPC core + 64KB D+I-\$
- C2 = mediocre IPC + 32MB D+I-\$
- Too many cores can degrade overall performance



# AIPC RESULTS FOR MEDIUM SCALE CMT



**Table 4: Maximum AIPC for all benchmarks.**

Benchmark	Core	Small Scale Cores, AIPC	Medium Scale Cores, AIPC	Large Scale Cores, AIPC
SPEC JBB2000	3p12t	5, 9.6	9, 17.3	15, 30.8
TPC-C	2p16t	5, 6.4	7, 11.8	12, 20.8
TPC-W	2p16t	5, 8.3	9, 15.2	15, 27.7
XML Test	3p12t	5, 11	9, 20.1	15, 35.4

Best results for each core configuration, across all benchmarks.

# MAX AIPC RESULTS FOR MEDIUM SCALE CMTs

**Table 3: Maximum AIPC for medium-scale CMTs for SPEC JBB, TPC-C, TPC-W, and XML Test.**

Core Config	SPEC JBB 2000				TPC-C				TPC-W				XML Test			
	L1	L2	Cores	AIPC	L1	L2	Cores	AIPC	L1	L2	Cores	AIPC	L1	L2	Cores	AIPC
1p2t	16/32	1.5/12	20	9.8	16/32	2.5/10	16	5.8	16/32	1.5/12	20	8.6	16/32	1.5/12	20	11.8
1p4t	16/32	1.5/12	17	13.2	16/32	2.5/10	14	8.2	16/32	1.5/12	17	10.6	16/32	1.5/12	17	14.8
1p8t	16/32	2.5/10	12	11.7	32/32	1.5/12	14	8.9	32/32	1.5/12	14	13.0	16/32	1.5/12	14	13.8
2p2t	16/32	1.5/12	16	8.6	16/32	1.5/12	16	5.1	16/32	1.5/12	16	7.5	16/32	1.5/12	16	10.5
2p4t	32/32	1.5/12	14	12.9	32/32	2.5/10	12	7.8	32/32	1.5/12	14	10.6	16/32	1.5/12	14	15.2
2p8t	16/32	1.5/12	12	16.5	32/32	2.5/10	9	9.5	32/32	1.5/12	12	13.6	32/32	1.5/12	12	18.9
2p16t	32/64	2.5/10	7	13.3	64/64	2.5/10	7	<b>11.8</b>	64/64	1.5/12	9	<b>15.2</b>	32/64	1.5/12	9	16.9
3p3t	32/32	1.5/12	13	10.3	32/32	2.5/10	10	5.9	32/32	1.5/12	13	8.5	16/32	1.5/12	13	12.7
3p6t	32/32	1.5/12	11	14.4	32/32	2.5/10	9	8.5	32/32	1.5/12	11	11.3	32/32	1.5/12	11	16.5
3p12t	32/64	1.5/12	9	<b>17.3</b>	32/64	2.5/10	7	10.7	64/64	1.5/12	9	14.6	32/64	1.5/12	9	<b>20.1</b>
3p24t	32/64	2.5/10	5	13.6	32/64	2.5/10	5	10.9	32/64	1.5/12	6	14.0	32/64	1.5/12	6	15.5
4p8t	32/32	1.5/12	9	14.9	32/32	2.5/10	7	8.5	64/64	1.5/12	9	11.5	16/32	1.5/12	9	16.6
4p16t	32/64	1.5/12	7	16.8	32/64	2.5/10	5	9.8	64/64	1.5/12	7	14.4	32/64	1.5/12	7	18.5
2s1t	64/64	1.5/12	11	4.4	64/64	1.5/12	11	2.8	64/64	1.5/12	11	3.7	64/64	1.5/12	11	5.5
2s2t	64/64	1.5/12	10	7.0	64/64	1.5/12	10	4.3	64/64	1.5/12	10	5.8	64/64	1.5/12	10	8.6
2s4t	64/64	1.5/12	9	10.5	64/64	1.5/12	9	6.4	64/64	1.5/12	9	8.7	64/64	1.5/12	9	12.4
2s8t	64/64	1.5/12	7	12.1	64/64	1.5/12	7	8.1	64/64	1.5/12	7	10.6	64/64	1.5/12	7	12.7
4s1t	64/64	1.5/12	7	2.9	64/64	1.5/12	7	1.9	64/64	1.5/12	7	2.6	64/64	1.5/12	7	3.7
4s2t	64/64	1.5/12	6	4.5	64/64	1.5/12	6	2.9	64/64	1.5/12	6	3.9	64/64	1.5/12	6	5.8
4s4t	64/64	1.5/12	5	6.6	64/64	1.5/12	5	4.1	64/64	1.5/12	5	5.6	64/64	1.5/12	5	7.8
4s8t	64/64	1.5/12	4	8.5	64/64	1.5/12	4	5.5	64/64	1.5/12	4	7.2	64/64	1.5/12	4	9.1

Note: The L1 refers to the primary data/instruction cache size. The L2 cache configuration size (MB)/set associativity (SA) are provided along with the total number of cores for that CMT configuration.

# CONTRIBUTIONS

---

- Given equivalent area, scalar CMTs with 4+ threads outperform nearly all superscalar CMTs
- Fine-grain multithreading is necessary in addition to multicore for server applications

Exceptions:

- Saturating the pipeline
- Saturating the memory bandwidth
- Best configurations require enough threads and primary cache to achieve 60-85% utilization of the pipeline

# CONTRIBUTIONS (II)

---

- CMT using lower performance cores yields better performance
- Cores with smaller primary caches are better
  - Larger I-cache than D-cache is always better
- Optimal AIPC requires only 25-40% of the area devoted to the secondary cache
- 2-way superscalar outperformed 4-way superscalar cores with the same number of threads

# SCALING CONCERNS

---

- Memory bandwidth must be sufficient to keep cores busy
- 4 dual Fully Buffered DIMM sufficient for 130 - 65 nm generations
- 4-channel DDR2 simulations show configurations with more on-chip cache are better
  - 40-60% of the area instead of 25-40%
  - Penalty for overthreading was more pronounced