A Single Chip Multiprocessor Integrated with DRAM

Tadaaki Yamauchi[1], Lance Hammond[2], and Kunle Olukotun[2]

1. ULSI Laboratory
Mitsubishi Electric Corporation
Itami, 664 Japan

2. Computer Systems Laboratory
Stanford University
Stanford, CA 94305

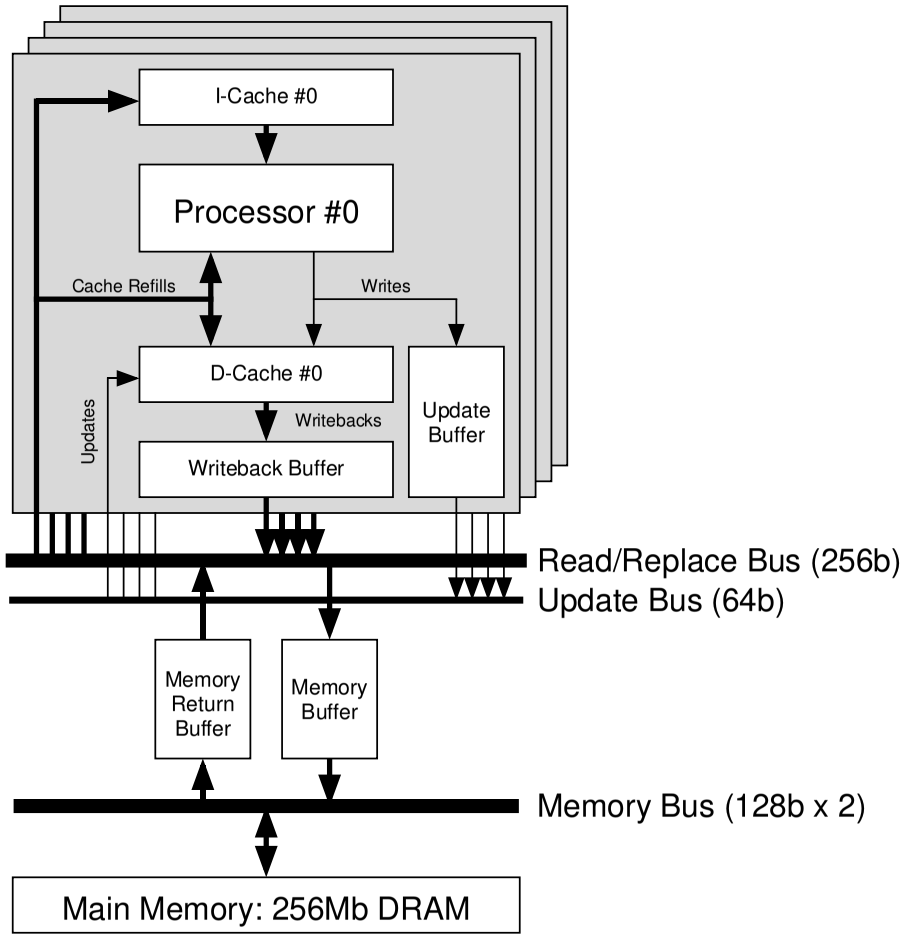
http://www-hydra.stanford.edu

Introduction

- An analysis of integrated CPUs & memories
- Two CPU architectures:
  - — 4 x 2-way superscalar MP-on-a-chip
  - — Single 2-way superscalar processor
- Three memory architectures:
  - — 32 MB DRAM on-chip main memory
  - — 2 MB & 4 MB SRAM on-chip L2 caches
- Paging effects analyzed with half-size memories
- SimOS used for detailed analysis

Overall Architecture

- CPUs & DRAMs on a chip
- Buses connect CPUs & memory
  - — Full-speed CPU buses
  - — Half-speed DRAM bus
- Update protocol gets its own bus
- Buffers between CPU & DRAM
- SRAM caches connect to CPU buses directly



Simulated Parameters

	DRAM System	SRAM Systems
CPU Configuration	4x2-way Superscalar	4x2-way Superscalar
CPU Frequency	500 MHz	500 MHz
L1 Capacity	16 KB I + 16 KB D	16 KB I + 16 KB D
L1 Access Time	1 CPU Cycle	1 CPU cycle
L2 Capacity	—	2 MB or 4 MB unified
Control Overhead	—	1 CPU cycle
L2 Access Time	—	4 CPU cycles
Memory Configuration	On-chip 32 MB DRAM	Off-chip DRAM
Number of Banks	16	4
Memory Bus Width	256 bits	64 bits
Bus Frequency	250 MHz	250 MHz
Control Overhead	5 CPU Cycles	5 CPU Cycles
DRAM Access Time	30 ns	100 ns
DRAM Cycle Time	60 ns	100 ns

Simulation Method

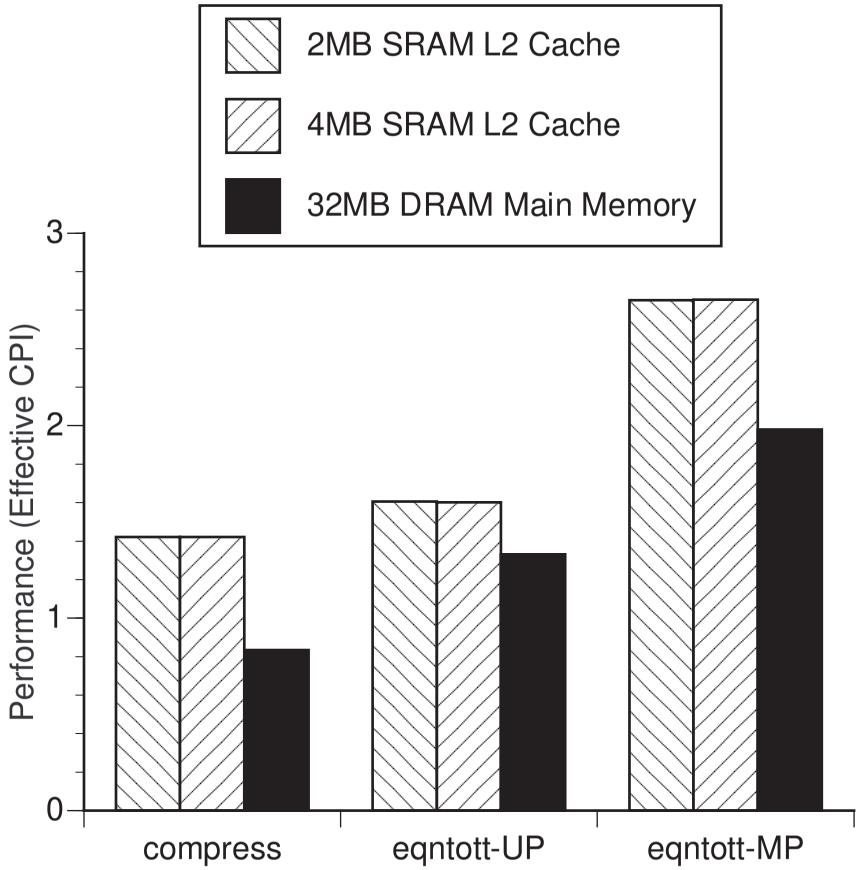
- SimOS Environment
  - — Realistic models of all components
  - — Full OS support

- Five SPEC applications:

Application	Type	Suite	Parallelization
compress	integer	SPEC95	not parallelized
eqntott	integer	SPEC92	hand parallelized
tomcatv	FP	SPEC95	SUIF-parallelized
swim	FP	SPEC95	SUIF-parallelized
applu	FP	SPEC95	SUIF-parallelized

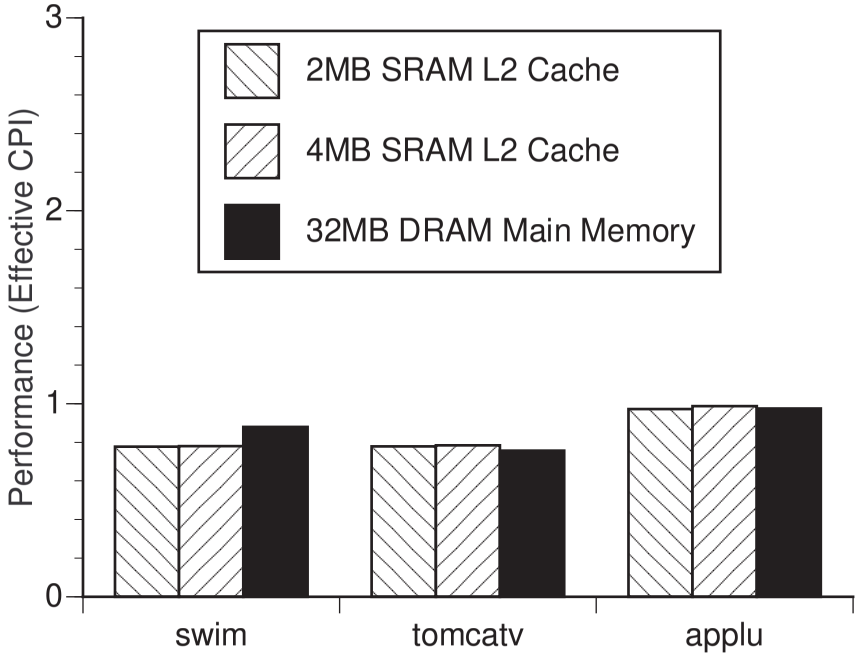
Integer Performance

- Both compress and eqntott **slow down** with on-chip DRAM
- Apps are small, and fit in caches
- On-chip memory latency is more important than bandwidth requirements
- Performance proportional to L1 cache hit rate
- UP & MP similar



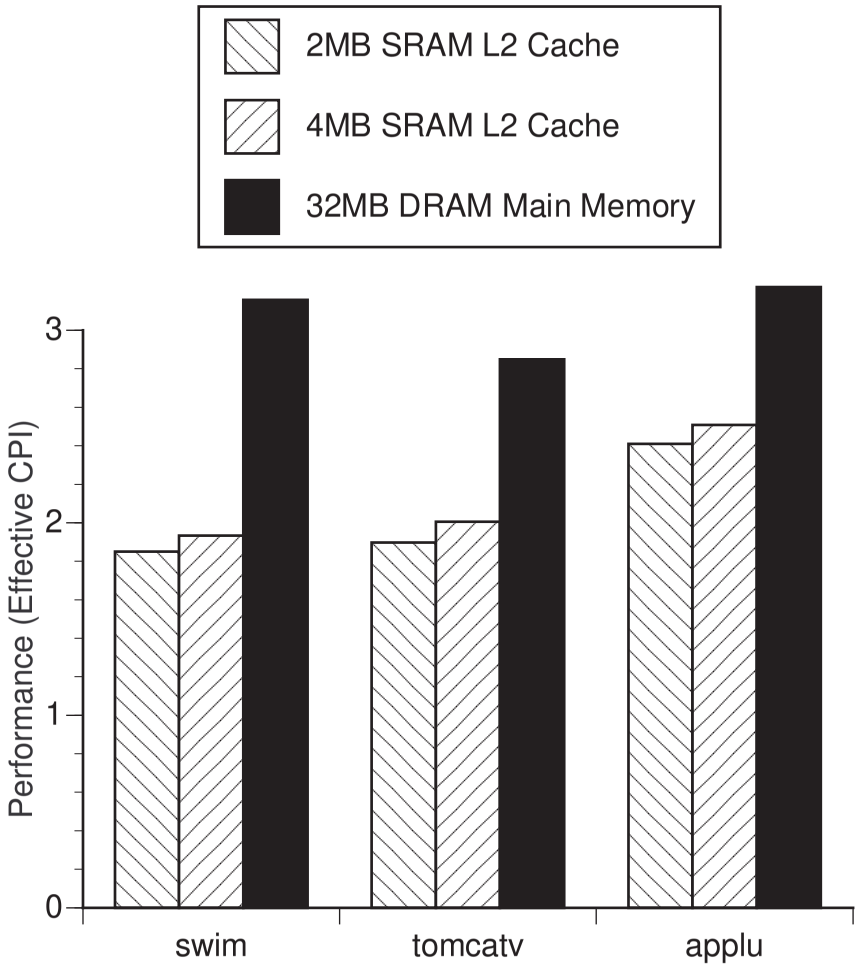
FP Performance: UP case

- Small performance variations
- 25-30% miss rates in L2 caches
- Off-chip bandwidth is sufficient



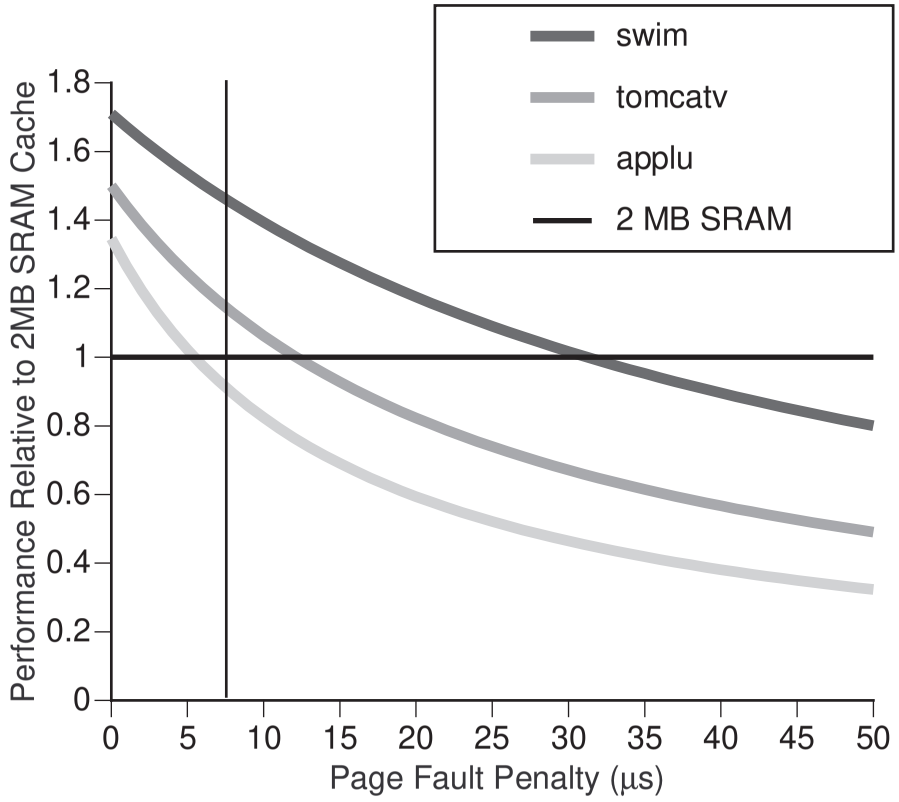
FP Performance: MP case

- DRAM much better than SRAM
- Still 25-30% miss rates in L2s
- Off-chip interface is overwhelmed
- On-chip DRAM has enough bandwidth
- Both SRAM systems offer similar performance



FP Performance: Paging

- On-chip DRAM sets hard memory size limits
- Tested by halving DRAM size
- Even μs paging times to/from off-chip DRAM can nullify on-chip DRAM performance
- Protocols to allow DRAM expansion beyond chip size are a critical research point



Conclusions

- On-chip DRAM may not help integer performance
  - — Applications are typically small
  - — We're testing larger apps now
- On-chip DRAM *can* help FP performance
  - — Slow CPUs don't benefit much
  - — Fast CPUs and MPs utilize increased bandwidth
- Applications that are larger than the on-chip DRAM can negate DRAM advantages