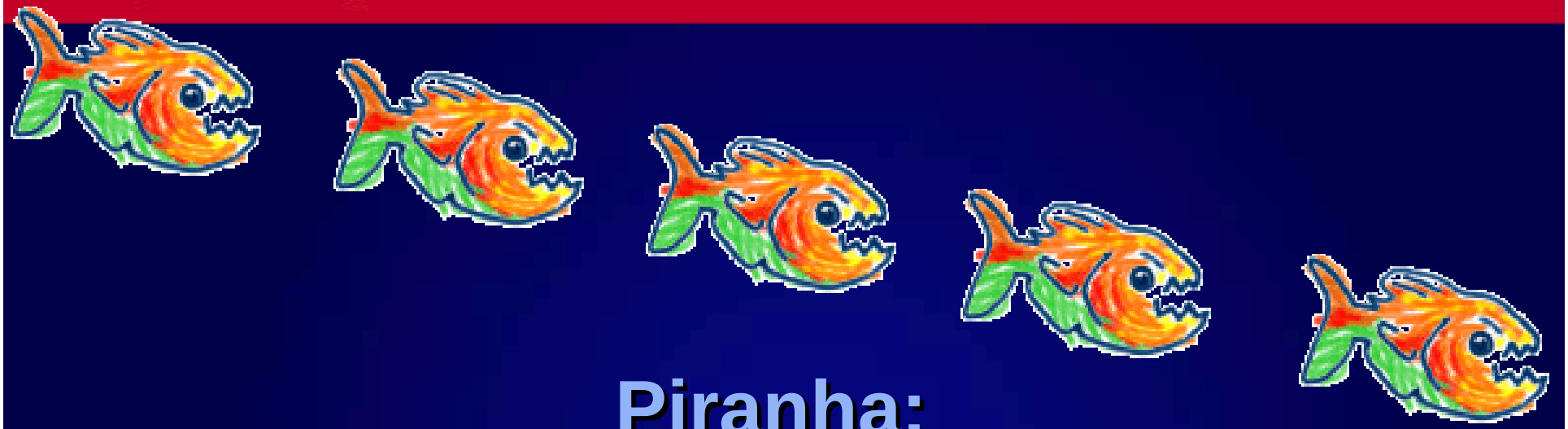
COMPAQ
Inspiration Technology



Piranha:

Designing a Scalable CMP-based System for Commercial Workloads

Luiz André Barroso
Western Research Laboratory

COMPAQ

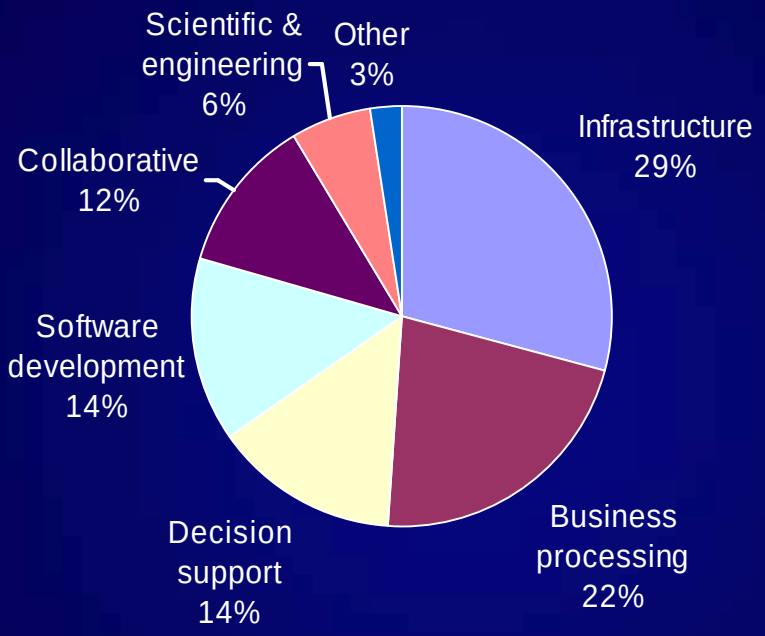
What is Piranha?

- A scalable shared memory architecture based on chip multiprocessing (CMP) and targeted at commercial workloads
- A research prototype under development by Compaq Research and Compaq NonStop Hardware Development Group
- A departure from ever increasing processor complexity and system design/verification cycles

COMPAQ

Importance of Commercial Applications

Worldwide Server Customer Spending (IDC 1999)

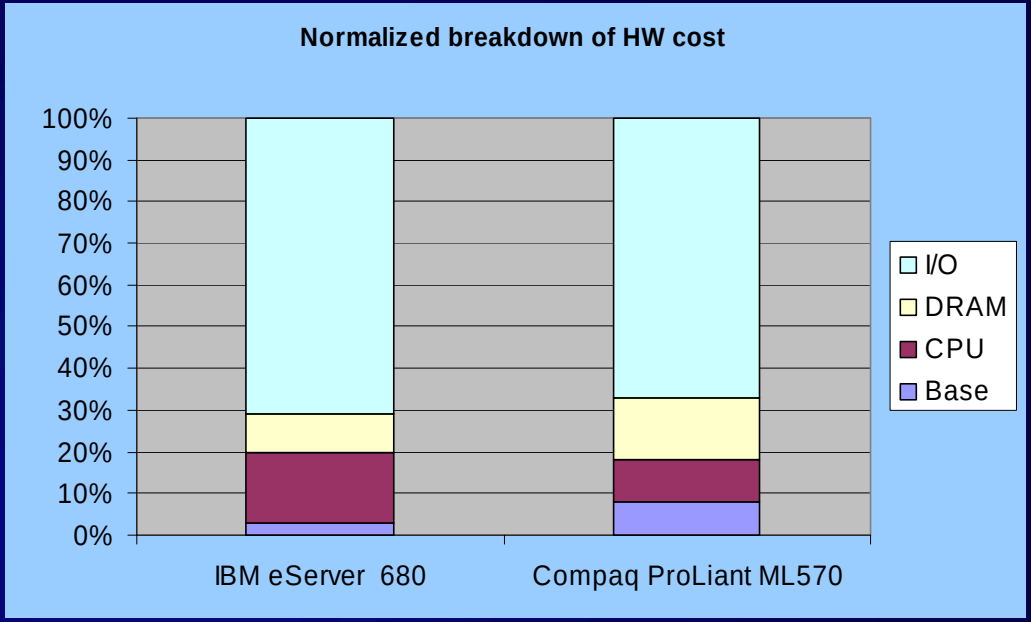


- Total server market size in 1999: ~$55-60B
  - technical applications: less than $6B
  - commercial applications: ~$40B

COMPAQ

Price Structure of Servers

- IBM eServer 680
  (220KtpmC; $43/tpmC)
  - 24 CPUs
  - 96GB DRAM, 18 TB Disk
  - $9M price tag
- Compaq ProLiant ML370
  (32KtpmC; $12/tpmC)
  - 4 CPUs
  - 8GB DRAM, 2TB Disk
  - $240K price tag



System	Price per component		
	$/CPU	$/MB DRAM	$/GB Disk
IBM eServer 680	$65,417	$9	$359
Compaq ProLiant ML570	$6,048	$4	$64

- Storage prices dominate (50%-70% in customer installations)
- Software maintenance/management costs even higher (up to $100M)
- Price of expensive CPUs/memory system amortized

COMPAQ

Outline

- Importance of Commercial Workloads
- Commercial Workload Requirements
- Trends in Processor Design
- Piranha
- Design Methodology
- Summary

COMPAQ

Studies of Commercial Workloads

- Collaboration with Kourosh Gharachorloo (Compaq WRL)
  - ISCA'98: Memory System Characterization of Commercial Workloads (with E. Bugnion)
  - ISCA'98: An Analysis of Database Workload Performance on Simultaneous Multithreaded Processors (with J. Lo, S. Eggers, H. Levy, and S. Parekh)
  - ASPLOS'98: Performance of Database Workloads on Shared-Memory Systems with Out-of-Order Processors (with P. Ranganathan and S. Adve)
  - HPCA'00: Impact of Chip-Level Integration on Performance of OLTP Workloads (with A. Nowatzyk and B. Verghese)
  - ISCA'01: Code Layout Optimizations for Transaction Processing Workloads (with A. Ramirez, R. Cohn, J. Larriba-Pey, G. Lowney, and M. Valero)

COMPAQ

Studies of Commercial Workloads: summary

- Memory system is the main bottleneck
  - astronomically high CPI
  - dominated by memory stall times
  - instruction stalls as important as data stalls
  - fast/large L2 caches are critical
- Very poor Instruction Level Parallelism (ILP)
  - frequent hard-to-predict branches
  - large L1 miss ratios
  - Ld-Ld dependencies
  - disappointing gains from wide-issue out-of-order techniques!

COMPAQ

Outline

- Importance of Commercial Workloads
- Commercial Workload Requirements
- Trends in Processor Design
- Piranha
- Design Methodology
- Summary

COMPAQ

Increasing Complexity of Processor Designs

- Pushing limits of instruction-level parallelism
  - multiple instruction issue
  - speculative out-of-order (OOO) execution
- Driven by applications such as SPEC
- Increasing design time and team size

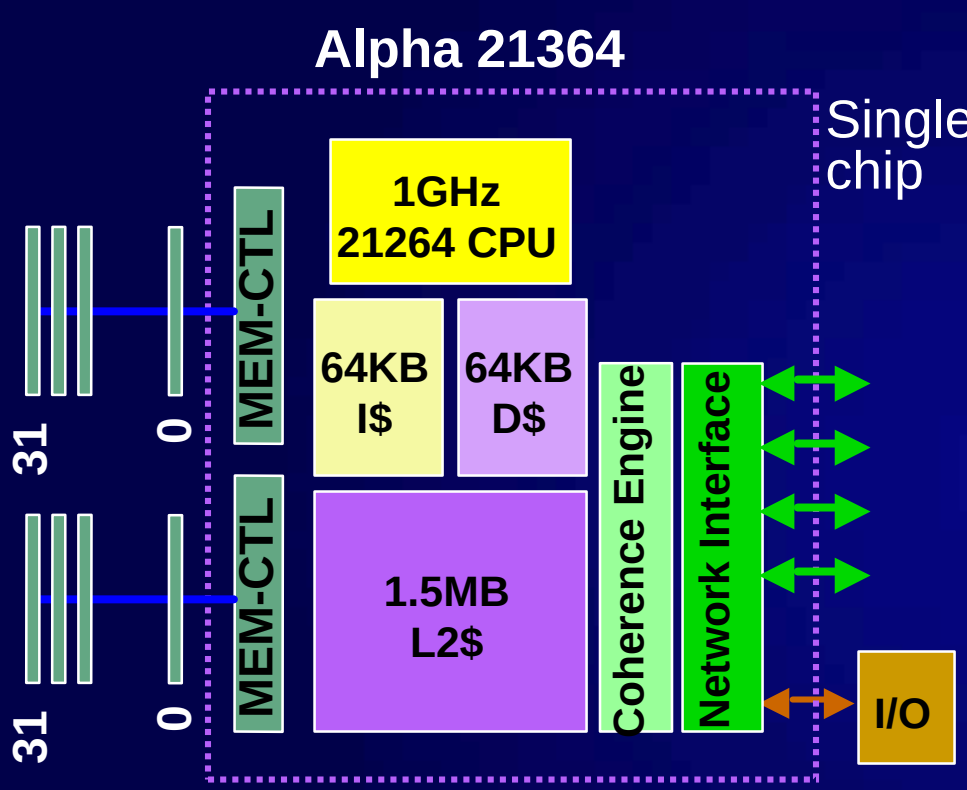
Processor (SGI MIPS)	Year Shipped	Transistor Count (millions)	Design Team Size	Design Time (months)	Verification Team Size (% of total)
R2000	1985	0.10	20	15	15%
R4000	1991	1.40	55	24	20%
R10000	1996	6.80	>100	36	>35%

courtesy: John Hennessy, IEEE Computer, 32(8)

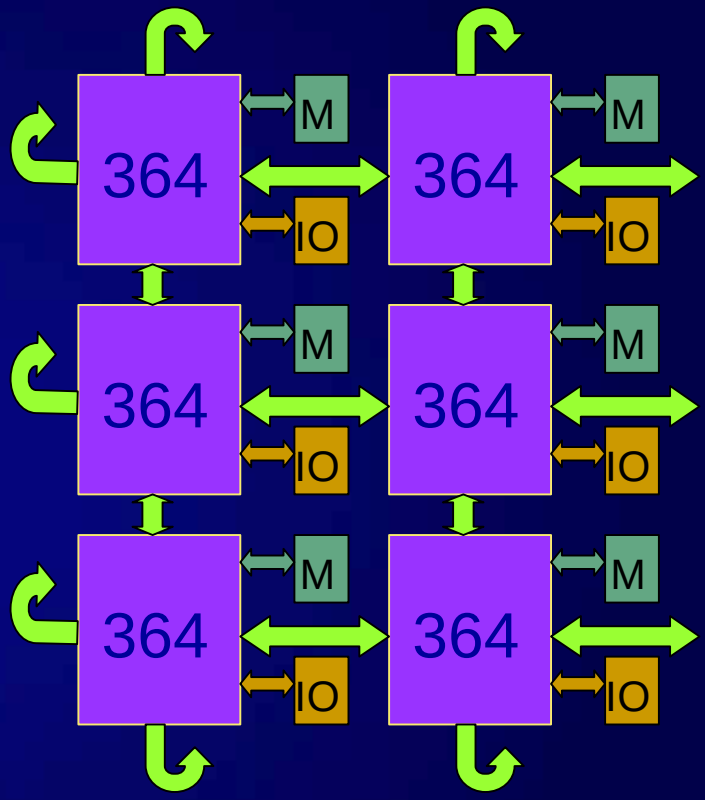
- Yielding diminishing returns in performance

COMPAQ

Exploiting Higher Levels of Integration



- lower latency, higher bandwidth
- reuse of existing CPU core addresses complexity issues

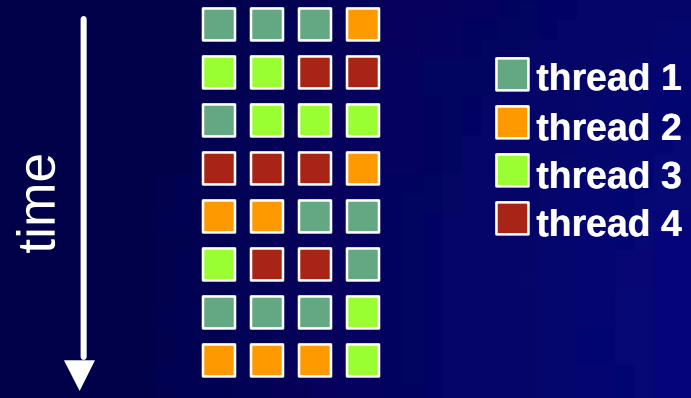


- incrementally scalable glueless multiprocessing

COMPAQ

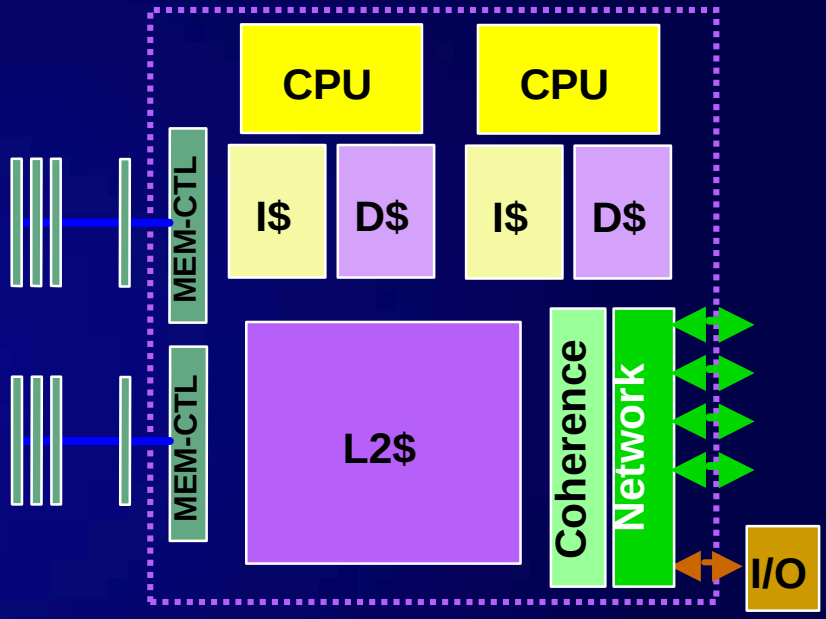
Exploiting Parallelism in Commercial Apps

Simultaneous Multithreading (SMT)



Example: Alpha 21464

Chip Multiprocessing (CMP)



Example: IBM Power4

- SMT superior in single-thread performance
- CMP addresses complexity by using simpler cores

COMPAQ

Outline

- Importance of Commercial Workloads
- Commercial Workload Requirements
- Trends in Processor Design
- Piranha
  - Architecture
  - Performance
- Design Methodology
- Summary

COMPAQ

Piranha Project

- Explore chip multiprocessing for scalable servers
- Focus on parallel commercial workloads
- Small team, modest investment, short design time
- Address complexity by using:
  - simple processor cores
  - standard ASIC methodology

Give up on ILP, embrace TLP

COMPAQ

Piranha Team Members

Research

- Luiz André Barroso (WRL)
- Kourosh Gharachorloo (WRL)
- David Lowell (WRL)
- Joel McCormack (WRL)
- Mosur Ravishankar (WRL)
- Rob Stets (WRL)
- Yuan Yu (SRC)

NonStop Hardware Development
ASIC Design Center

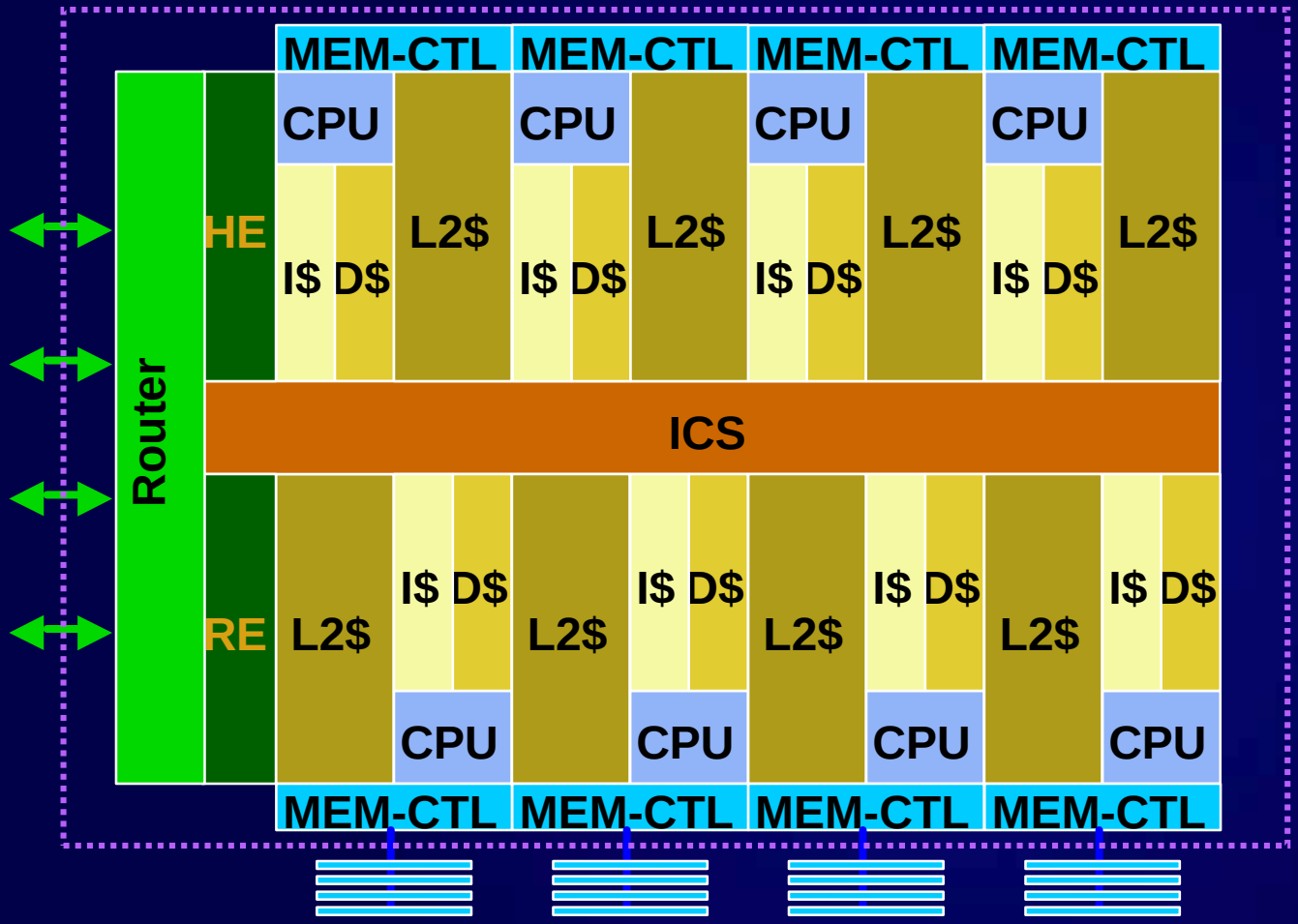
- Tom Heynemann
- Dan Joyce
- Harland Maxwell
- Harold Miller
- Sanjay Singh
- Scott Smith
- Jeff Sprouse
- … several contractors

Former Contributors

Robert McNamara
Basem Nayfeh
Andreas Nowatzyk
Joan Pendleton
Shaz Qadeer
Brian Robinson
Barton Sano
Daniel Scales
Ben Verghese

COMPAQ

Piranha Processing Node



Alpha core:
- 1-issue, in-order, 500MHz

L1 caches:
- I&D, 64KB, 2-way

Intra-chip switch (ICS)
- 32GB/sec, 1-cycle delay

L2 cache:
- shared, 1MB, 8-way

Memory Controller (MC)
- RDRAM, 12.8GB/sec

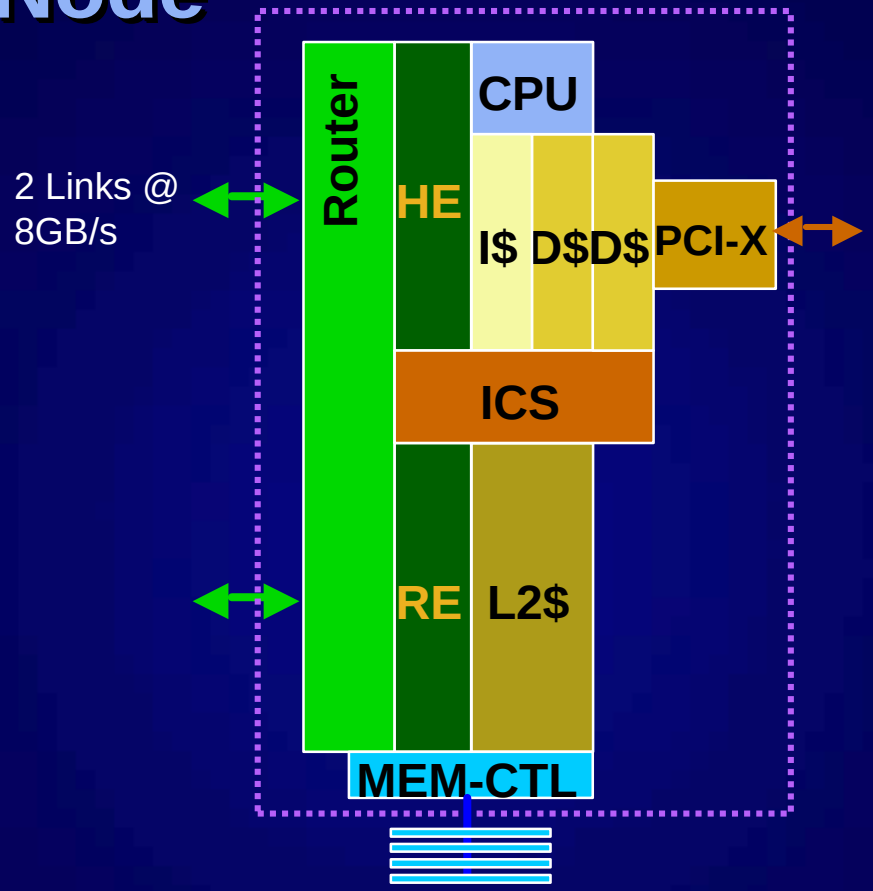
Protocol Engines (HE & RE):
- µprog., 1K µinstr., even/odd interleaving

System Interconnect:
- 4-port Xbar router
- topology independent
- 32GB/sec total bandwidth

Single Chip

COMPAQ

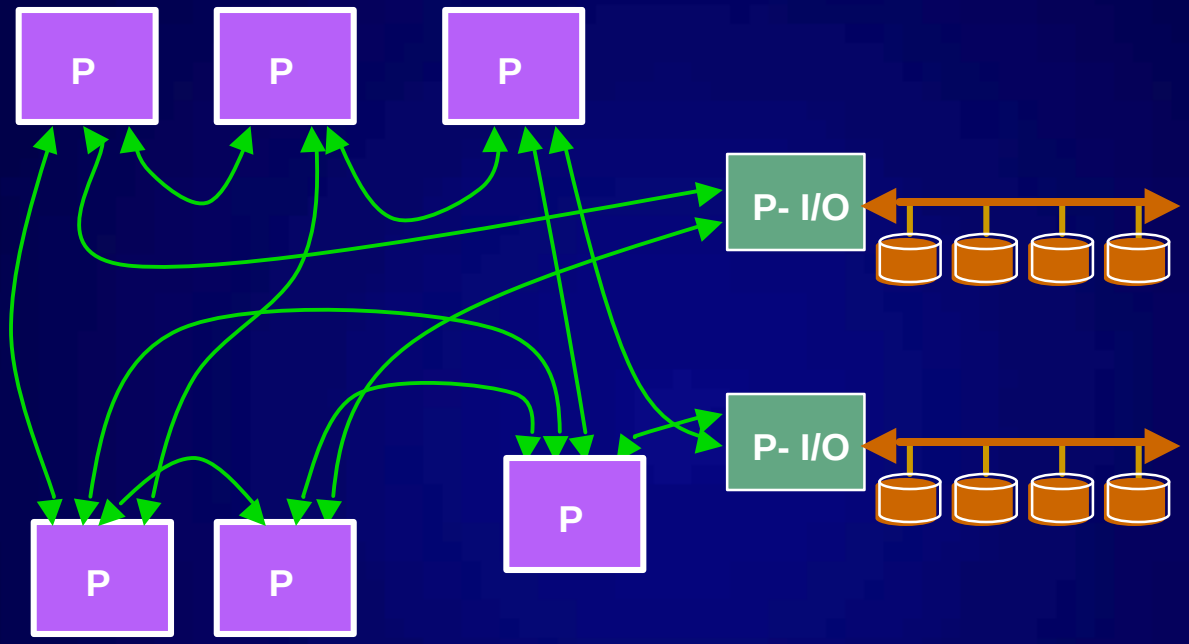
Piranha I/O Node



- I/O node is a full-fledged member of system interconnect
  - CPU indistinguishable from Processing Node CPUs
  - participates in global coherence protocol

COMPAQ

Example Configuration



- Arbitrary topologies
- Match ratio of Processing to I/O nodes to application requirements

COMPAQ

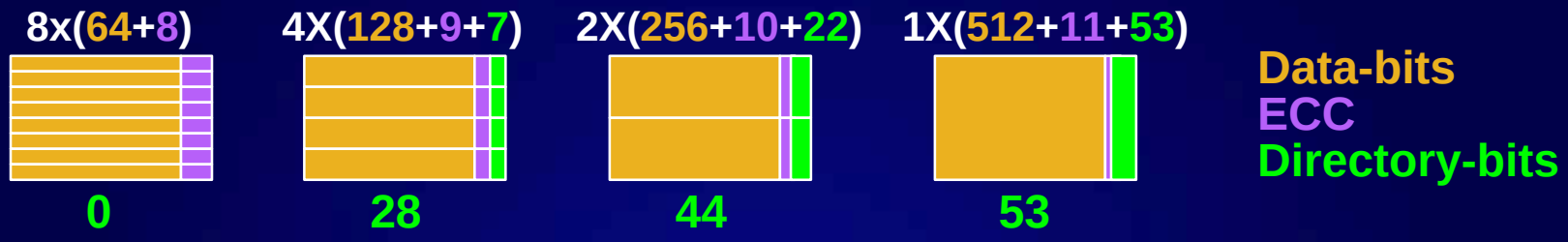
L2 Cache and Intra-Node Coherence

- No inclusion between L1s and L2 cache
  - total L1 capacity equals L2 capacity
  - L2 misses go directly to L1
  - L2 filled by L1 replacements
- L2 keeps track of all lines in the chip
  - sends Invalidates, Forwards
  - orchestrates L1-to-L2 write-backs to maximize chip-memory utilization
  - cooperates with Protocol Engines to enforce system-wide coherence

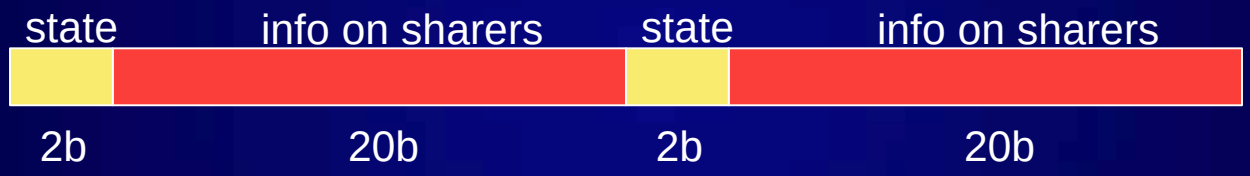
COMPAQ

Inter-Node Coherence Protocol

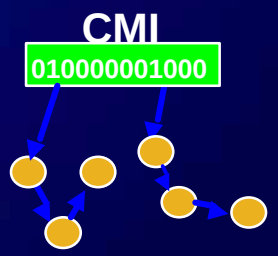
- ‘Stealing’ ECC bits for memory directory



- Directory (2b state + 40b sharing info)



- Dual representation: limited pointer + coarse vector
- “Cruise Missile” Invalidations (CMI)
  - limit fan-out/fan-in serialization with CV
- Several new protocol optimizations



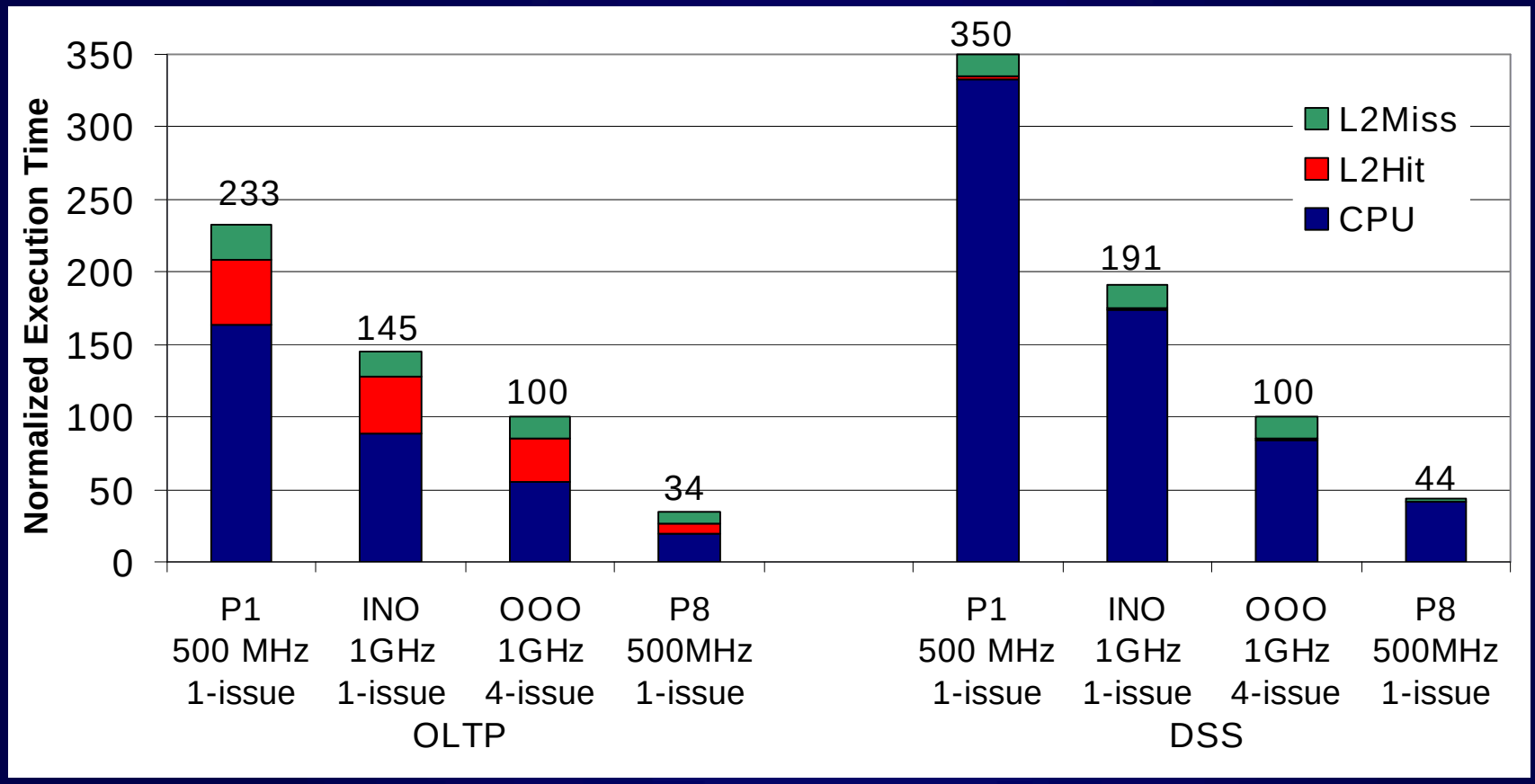
COMPAQ

Simulated Architectures

Parameter	Piranha (P8)	Next-Generation Microprocessor (OOO)	Full-Custom Piranha (P8F)
Processor Speed	500 MHz	1 GHz	1.25 GHz
Type	in-order	out-of-order	in-order
Issue Width	1	4	1
Instruction Window Size	-	64	-
Cache Line Size	64 bytes	64 bytes	64 bytes
L1 Cache Size	64 KB	64 KB	64 KB
L1 Cache Associativity	2-way	2-way	2-way
L2 Cache Size	1 MB	1.5 MB	1.5 MB
L2 Cache Associativity	8-way	6-way	6-way
L2 Hit / L2 Fwd Latency	16 ns / 24 ns	12 ns / NA	12 ns / 16 ns
Local Memory Latency	80 ns	80 ns	80 ns
Remote Memory Latency	120 ns	120 ns	120 ns
Remote Dirty Latency	180 ns	180 ns	180 ns

COMPAQ

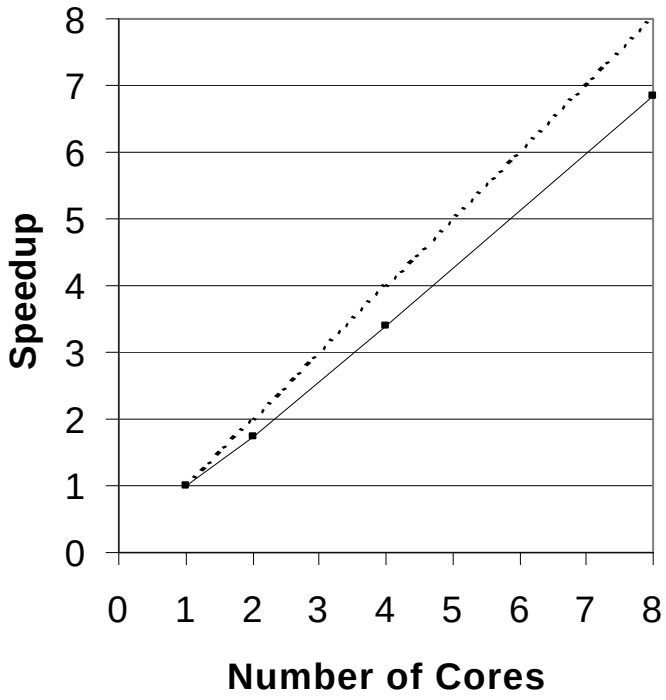
Single-Chip Piranha Performance

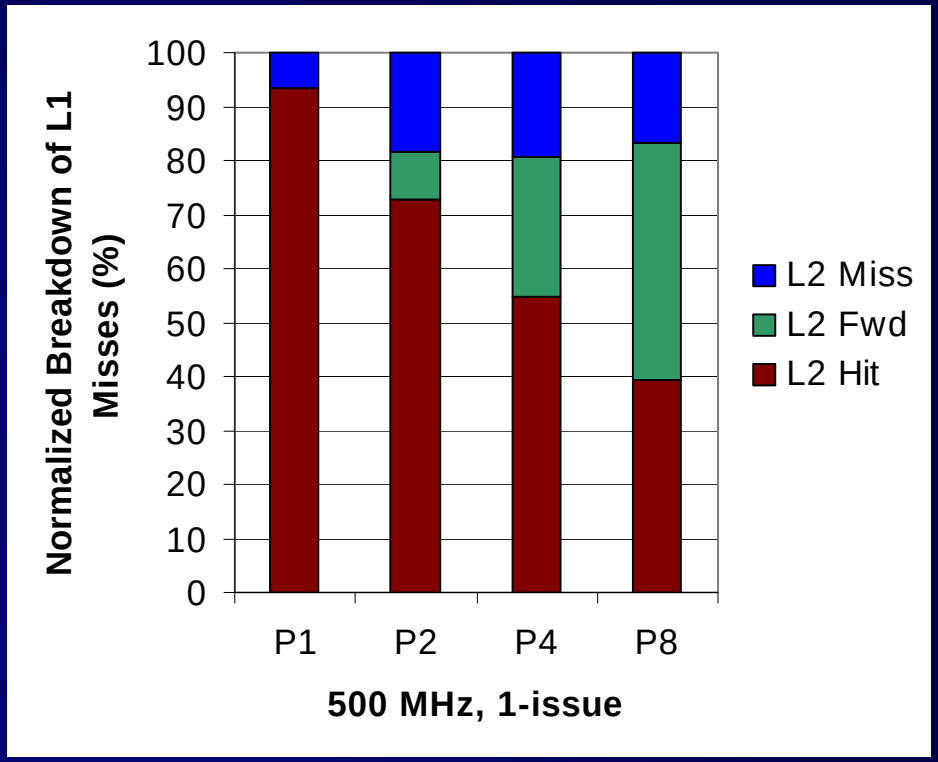


- Piranha's performance margin 3x for OLTP and 2.2x for DSS
- Piranha has more outstanding misses ➡ better utilizes memory system

COMPAQ

Single-Chip Performance (Cont.)

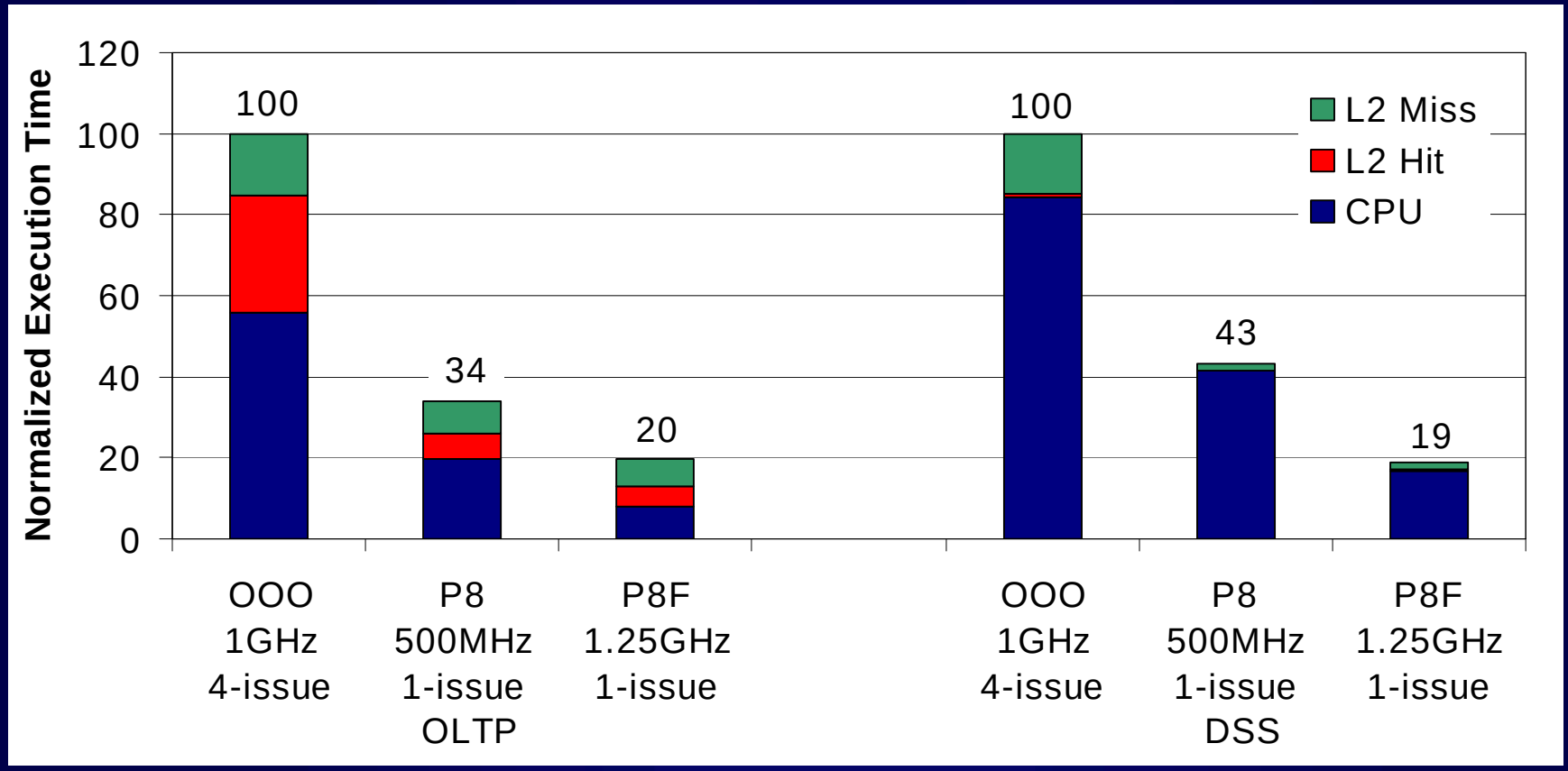




- Near-linear scalability
  - low memory latencies
  - effectiveness of highly associative L2 and non-inclusive caching

COMPAQ

Potential of a Full-Custom Piranha



- 5x margin over OOO for OLTP and DSS
- Full-custom design benefits substantially from boost in core speed

COMPAQ

Outline

- Importance of Commercial Workloads
- Commercial Workload Requirements
- Trends in Processor Design
- Piranha
- Design Methodology
- Summary

COMPAQ

Managing Complexity in the Architecture

- Use of many simpler logic modules
  - shorter design
  - easier verification
  - only short wires*
  - faster synthesis
  - simpler chip-level layout
- Simplify intra-chip communication
  - all traffic goes through ICS (no backdoors)
- Use of microprogrammed protocol engines
- Adoption of large VM pages
- Implement sub-set of Alpha ISA
  - no VAX floating point, no multimedia instructions, etc.

COMPAQ

Methodology Challenges

- Isolated sub-module testing
  - need to create robust bus functional models (BFM)
  - sub-modules' behavior highly inter-dependent
  - not feasible with a small team
- System-level (integrated) testing
  - much easier to create tests
  - only one BFM at the processor interface
  - simpler to assert correct operation
  - Verilog simulation is too slow for comprehensive testing

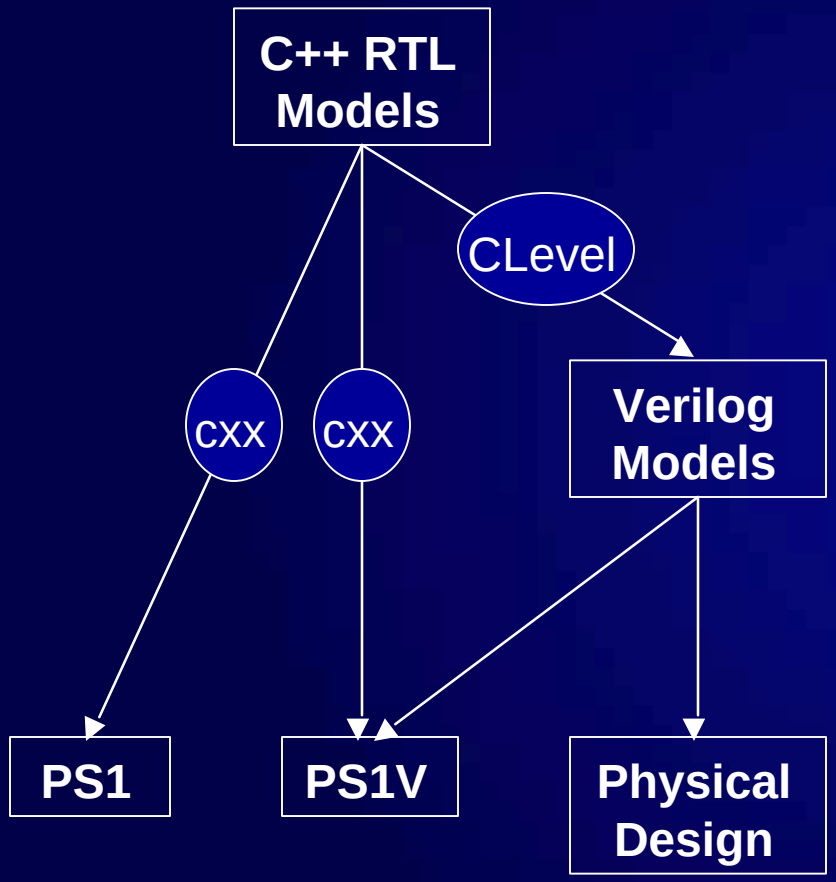
COMPAQ

Our Approach:

- Design in stylized C++ (synthesizable RTL level)
  - use mostly system-level, semi-random testing
  - simulations in C++ (faster & cheaper than Verilog)
    - simulation speed ~1000 clocks/second
  - employ directed tests to fill test coverage gaps
- Automatic C++ to Verilog translation
  - single design database
  - reduce translation errors
  - faster turnaround of design changes
  - risk: untested methodology
- Using industry-standard synthesis tools
- IBM ASIC process (Cu11)

COMPAQ

Piranha Methodology: Overview



cxx: C++ compiler

CLevel: C++-to-Verilog Translator

C++ RTL Models: Cycle accurate and "synthesizeable"

PS1: Fast (C++) Logic Simulator

Verilog Models: Machine translated from C++ models

Physical Design: leverages industry standard Verilog-based tools

PS1V: Can "co-simulate" C++ and Verilog module versions and check correspondence

COMPAQ

Summary

- CMP architectures are inevitable in the near future
- Piranha investigates an extreme point in CMP design
  - many simple cores
- Piranha has a large architectural advantage over complex single-core designs (> 3x) for database applications
- Piranha methodology enables faster design turnaround
- Key to Piranha is application focus:
  - One-size-fits-all solutions may soon be infeasible

COMPAQ

Reference

- Papers on commercial workload performance & Piranha

  research.compaq.com/wrl/projects/Database


COMPAQ