POWER6™ Processor and Systems

Jim McInnes
jimm@ca.ibm.com
Compiler Optimization
IBM Canada Toronto Software Lab

Role

- I am a Technical leader in the Compiler Optimization Team
  - Focal point to the hardware development team
  - Member of the Power ISA Architecture Board
- For each new microarchitecture I
  - help direct the design toward helpful features
  - Design and deliver specific compiler optimizations to enable hardware exploitation

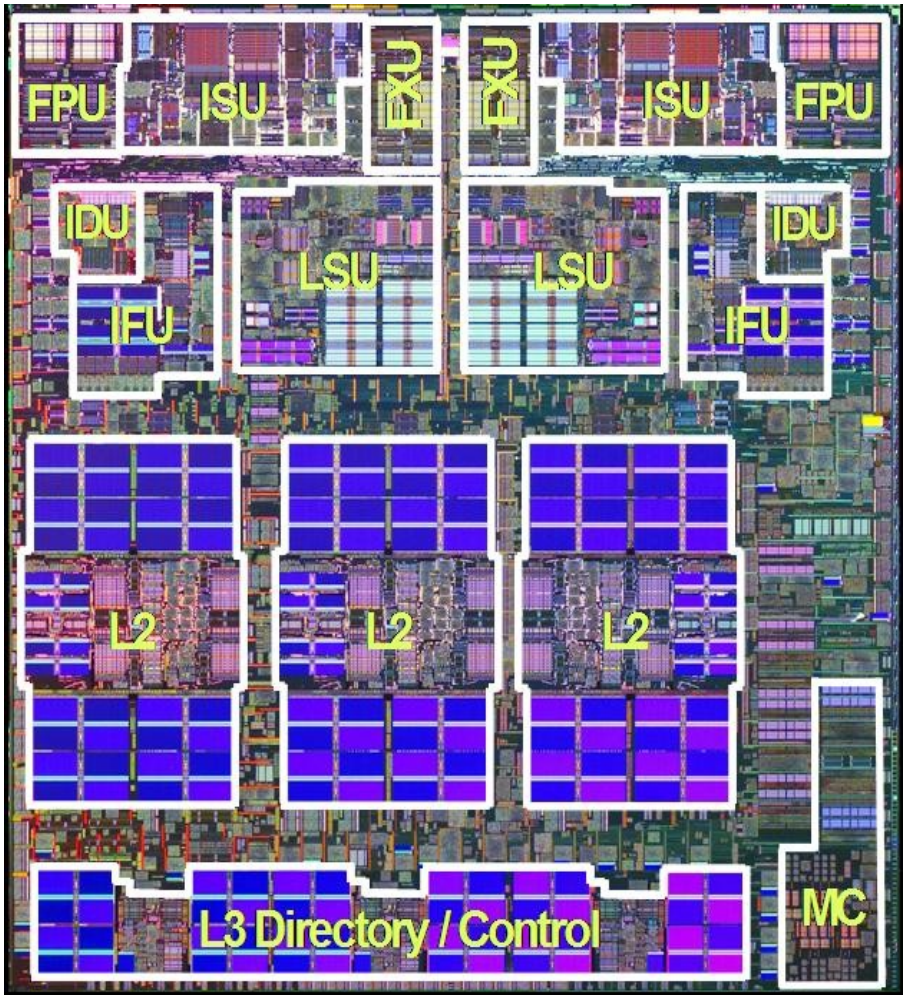
POWER5 Chip Overview

High frequency dual-core chip

- 8 execution units
  - ★ 2LS, 2FP, 2FX, 1BR, 1CR
- 1.9MB on-chip shared L2 – point of coherency, 3 slices
- On-chip L3 directory and controller
- On-chip memory controller

Technology & Chip Stats

- 130nm lithography, Cu, SOI
- 276M transistors, 389 mm^2
- I/Os: 2313 signal, 3057 Power/Gnd



POWER6 Chip Overview

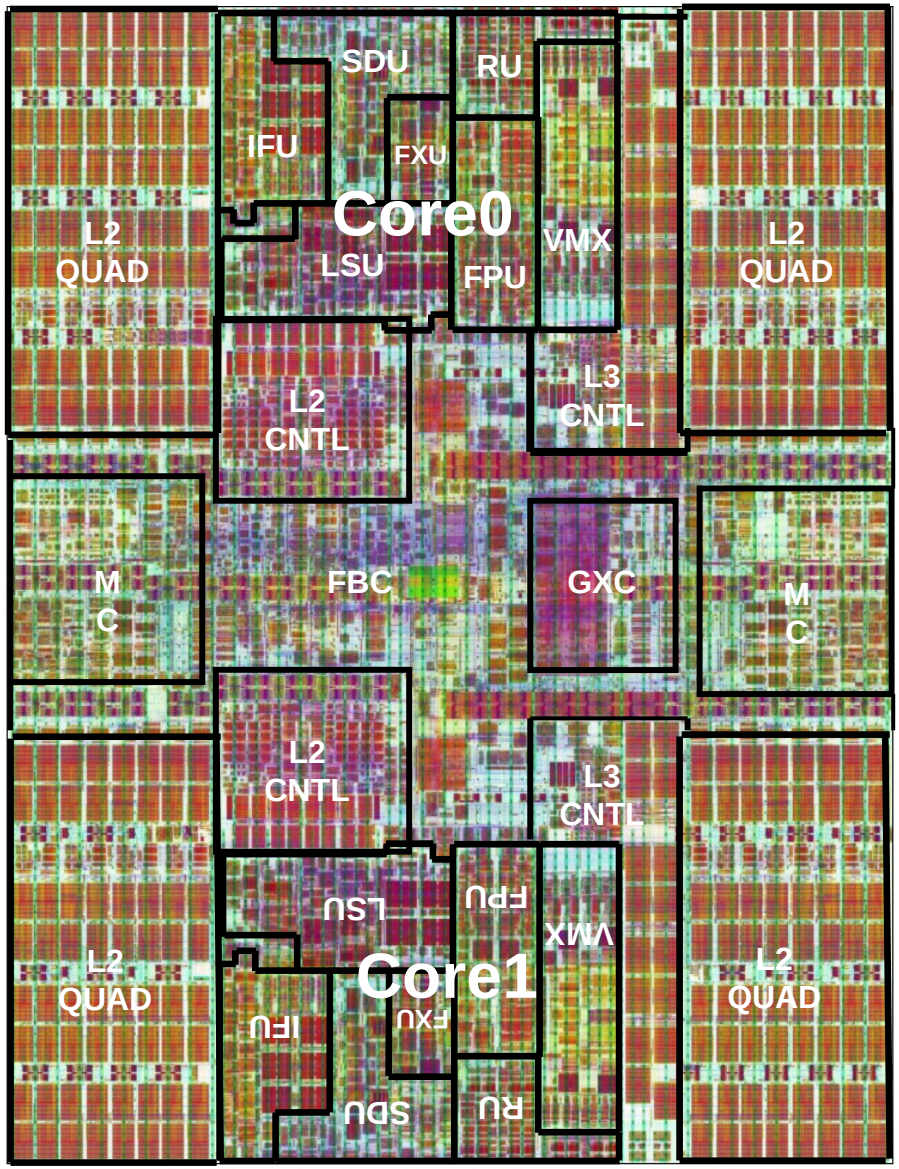
Ultra-high frequency dual-core chip

- 8 execution units
  - ★ 2LS, 2FP, 2FX, 1BR, 1VMX
- 2 x 4MB on-chip L2 – point of coherency, 4 quads
- On-chip L3 directory and controller
- Two on-chip memory controllers

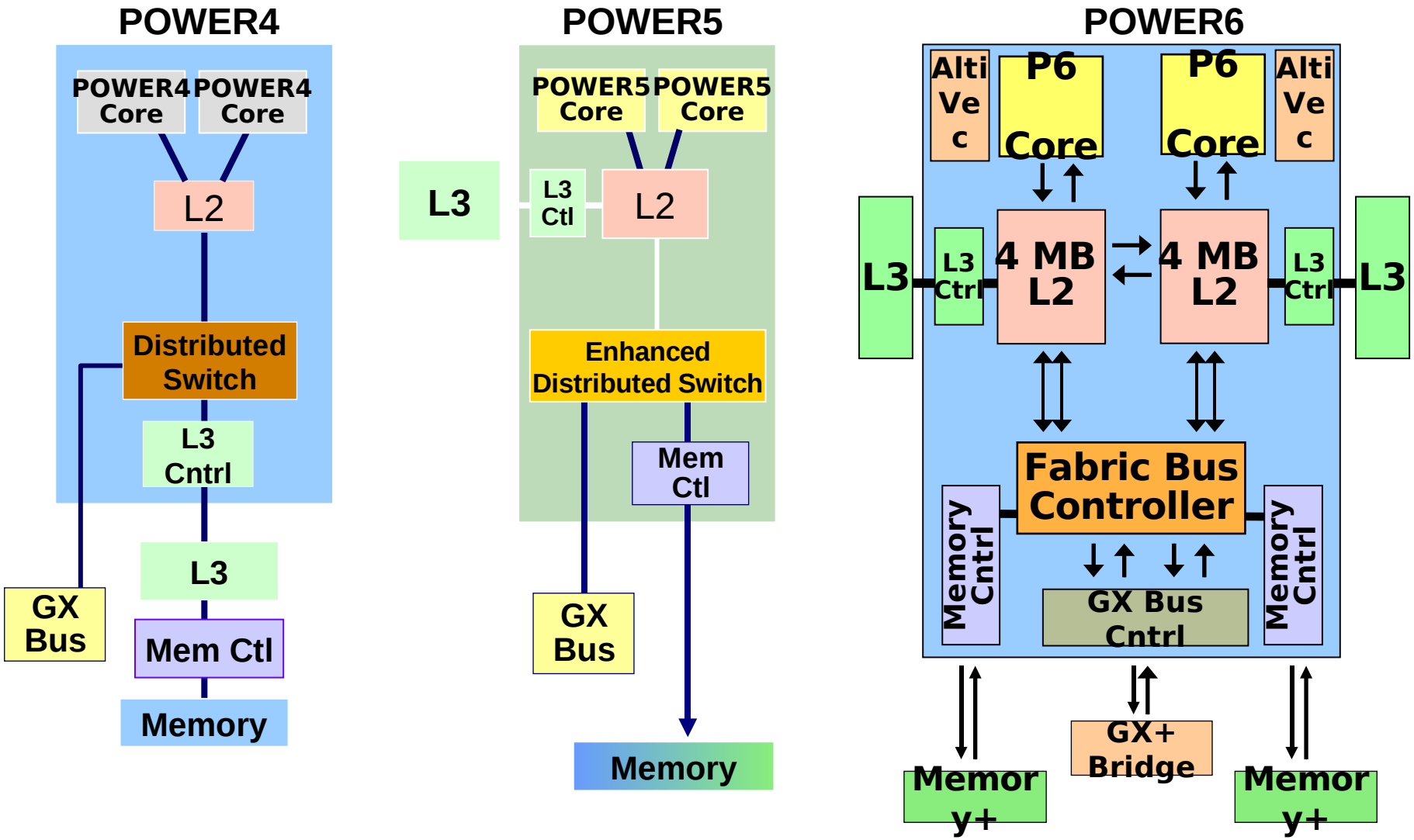
Technology & Chip stats

- CMOS 65nm lithography, SOI Cu
- 790M transistors, 341 mm^2 die
- I/Os: 1953 signal, 5399 Power/Gnd

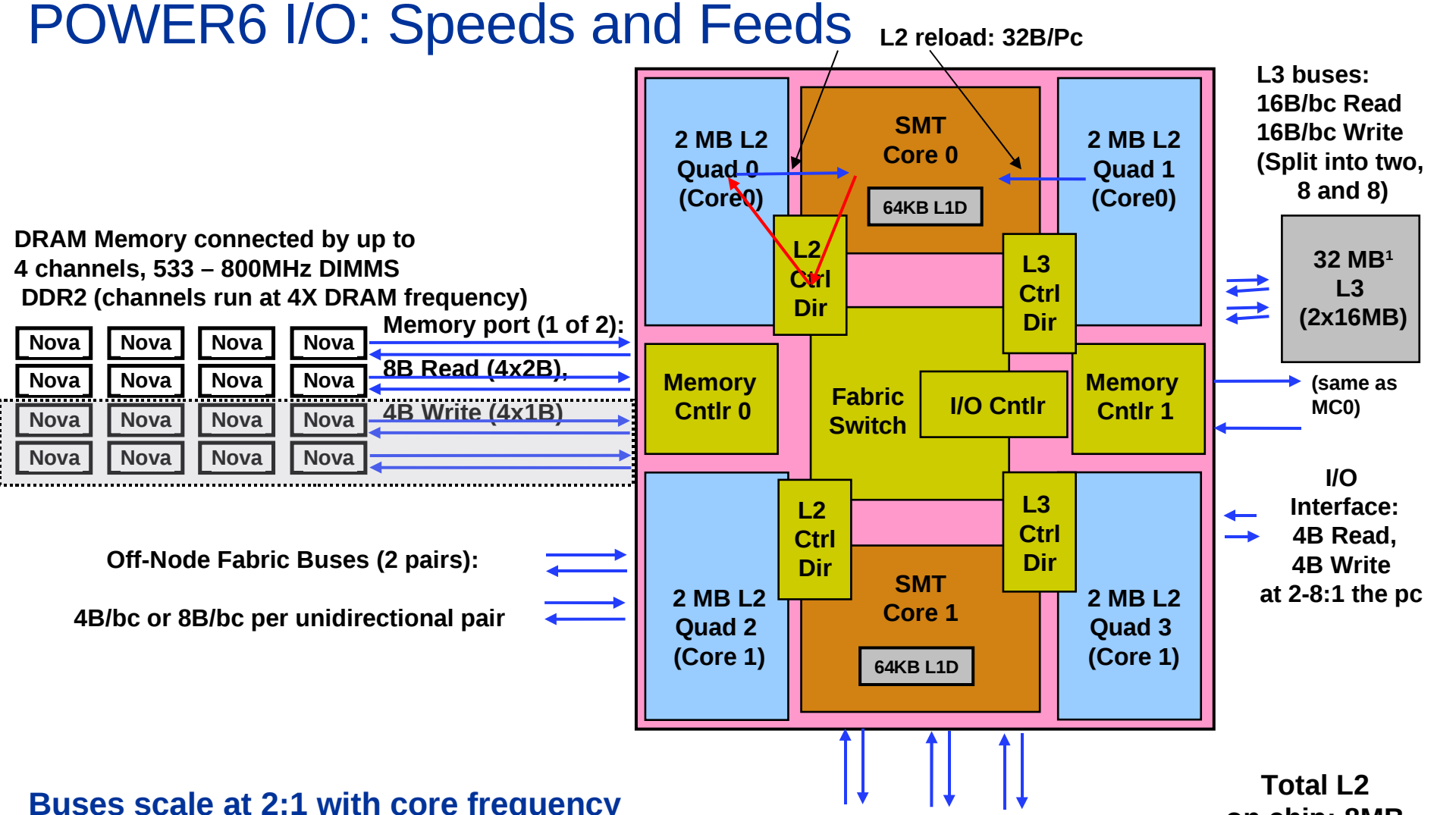
Full error checking and recovery (RU)



POWER4 / POWER5 / POWER6



POWER6 I/O: Speeds and Feeds

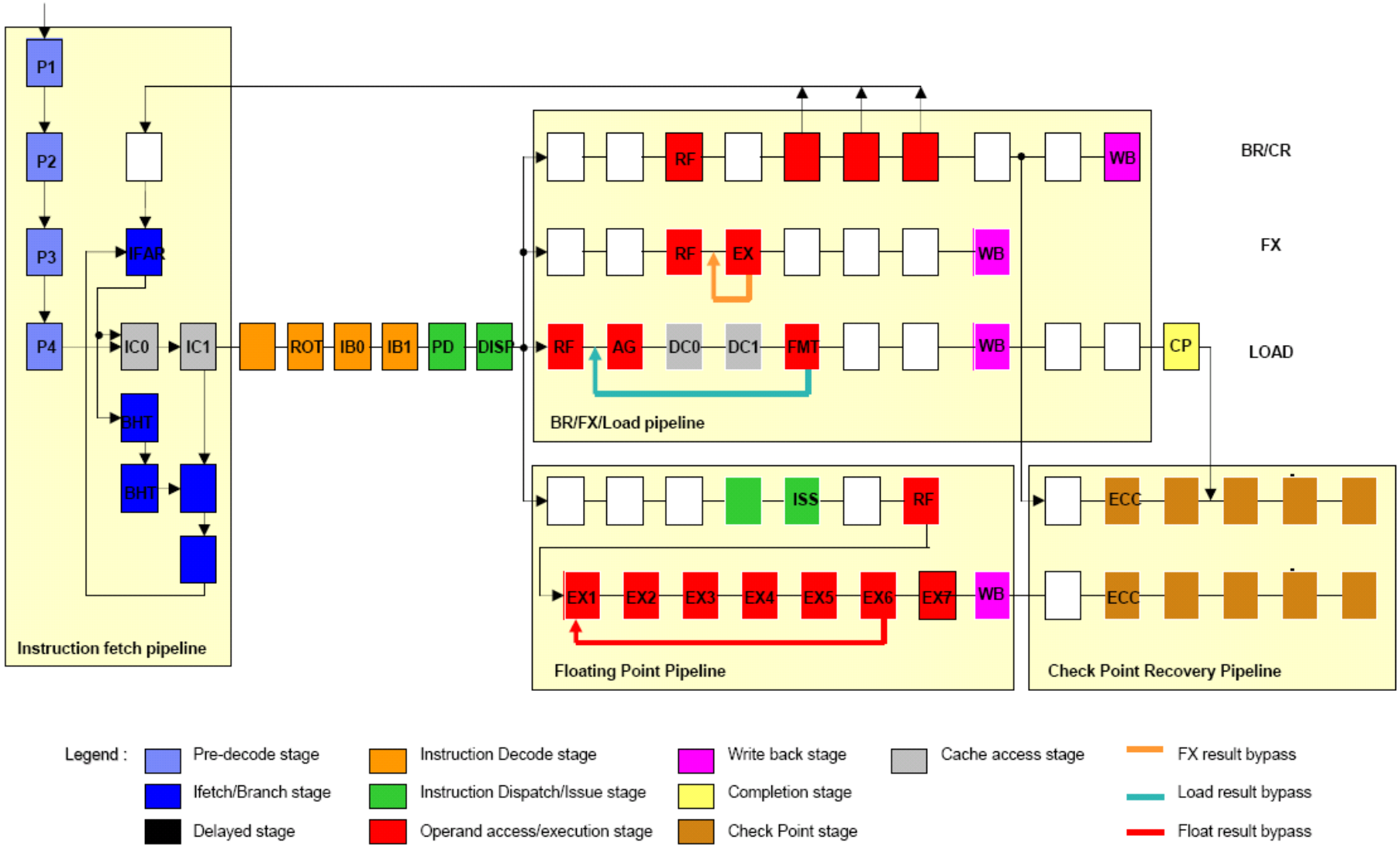


Buses scale at 2:1 with core frequency

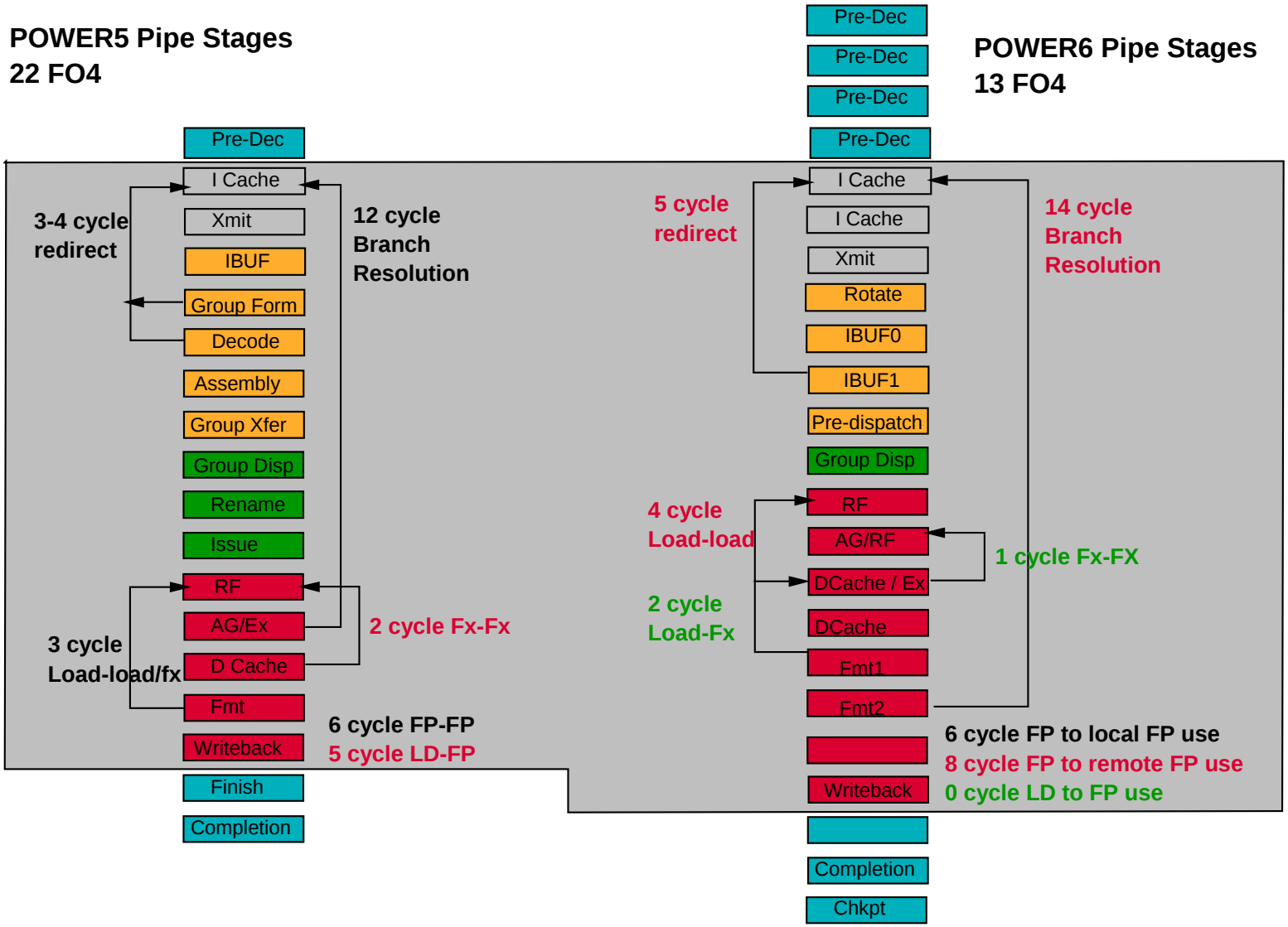
pc = processor clock
bc = bus clock
2 pc = 1 bc

[1]May be a single 32MB L3 chip with 8B buses

POWER6 Core Pipelines



POWER5 vs. POWER6 Pipeline Comparison



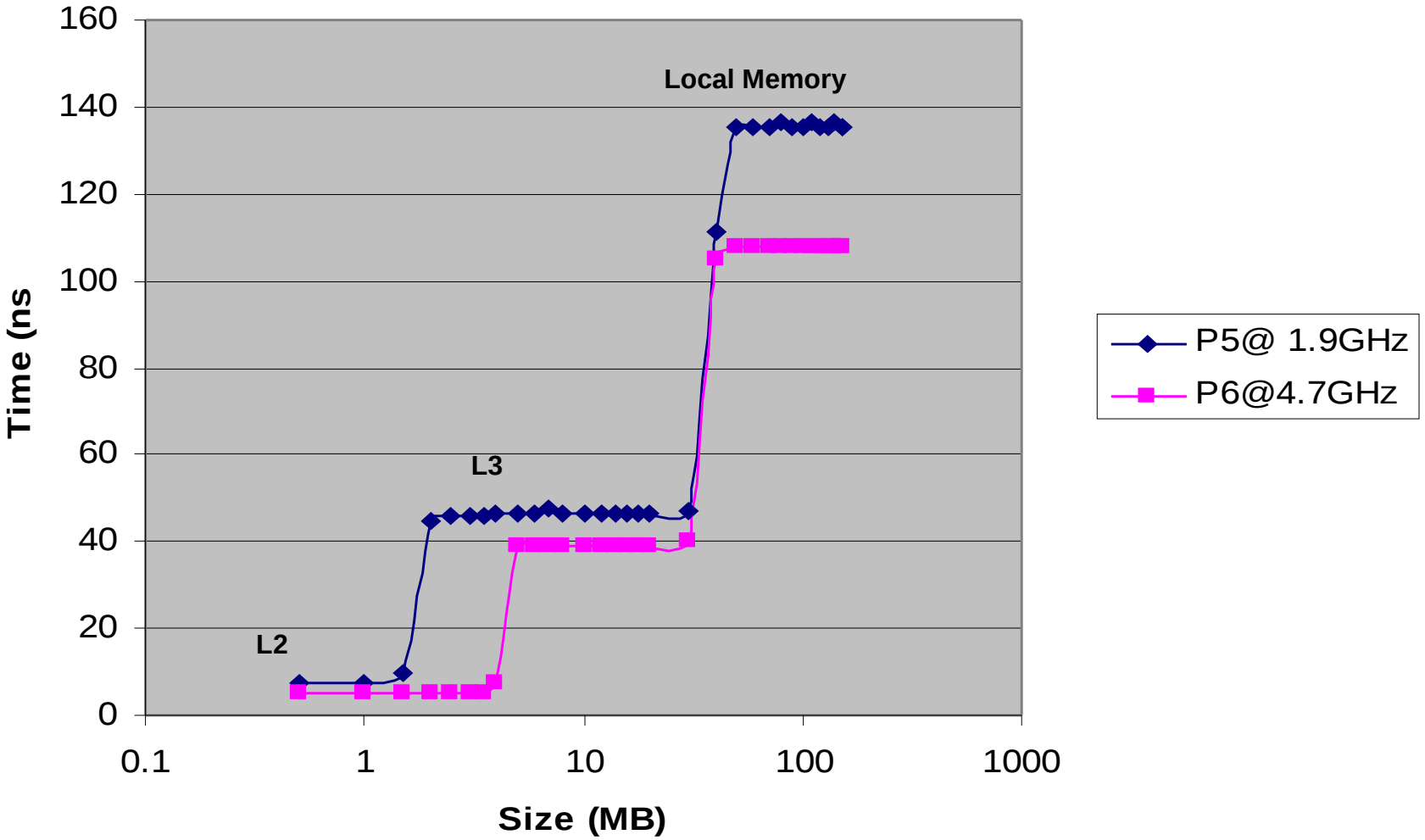
Side-by-Side comparisons

	POWER5+	POWER6
Style	General out-of-order execution	Mostly in-order with special case out-of-order execution
Units	2FX, 2LS, 2FP, 1BR, 1CR	2FX, 2LS, 2FP, 1BR/CR, 1VMX
Threading	2 SMT threads Alternate ifetch Alternate dispatch (up to 5 instructions)	2 SMT threads Priority-based dispatch Simultaneous dispatch from two threads (up to 7 instructions)

Side-by-side comparisons…

	POWER5+	POWER6
L1 Cache		
ICache capacity, associativity	64 KB, 2-way	64 KB, 4-way
DCache capacity, associativity	32 KB, 4-way	64 KB, 8-way
L2 Cache	Point of Coherency	Point of Coherency
Capacity, line size	1.9 MB, 128 B line	2 x 4 MB, 128 B line
Associativity, replacement	10-way, LRU	each 8-way, LRU
Off-chip L3 Cache		
Capacity, line size	36 MB, 256 B line	32 MB, 128 B line
Associativity, replacement	12-way, LRU	16-way, LRU
Memory **Memory bus**	4 TB maximum 2x DRAM frequency	8 TB maximum 4x DRAM frequency

Latency Profiles



Processor Core Functional Features

PowerPC AS Architecture
Simultaneous Multithreading (2 threads)
Decimal Floating Point (extension to PowerPC ISA)
48 Bit Real Address Support
Virtualization Support (1024 partitions)
Concurrent support of 4 page sizes (4K, 64K, 16M, 16G)
New storage key support
Bi-Endian Support
Dynamic Power Management
Single Bit Error Detection on all Dataflow
Robust Error Recovery (R unit)
CPU sparing support (dynamic CPU swapping)
Common Core for I, P, and Blade

Achieving High Frequency: POWER6 13FO4 Challenge Example

Circuit Design

- 1 FO4 = delay of 1 inverter that drives 4 receivers
- 1 Logical Gate = 2 FO4
- 1 cycle = Latch + function + wire
  - ★ 1 cycle = 3 FO4 + function + 4 FO4
- Function = 6 FO4 = 3 Gates

Integration

- It takes 6 cycles to send a signals across the core
- Communication between units takes 1 cycle using good wire
- Control across a 64-bit data flow takes a cycle

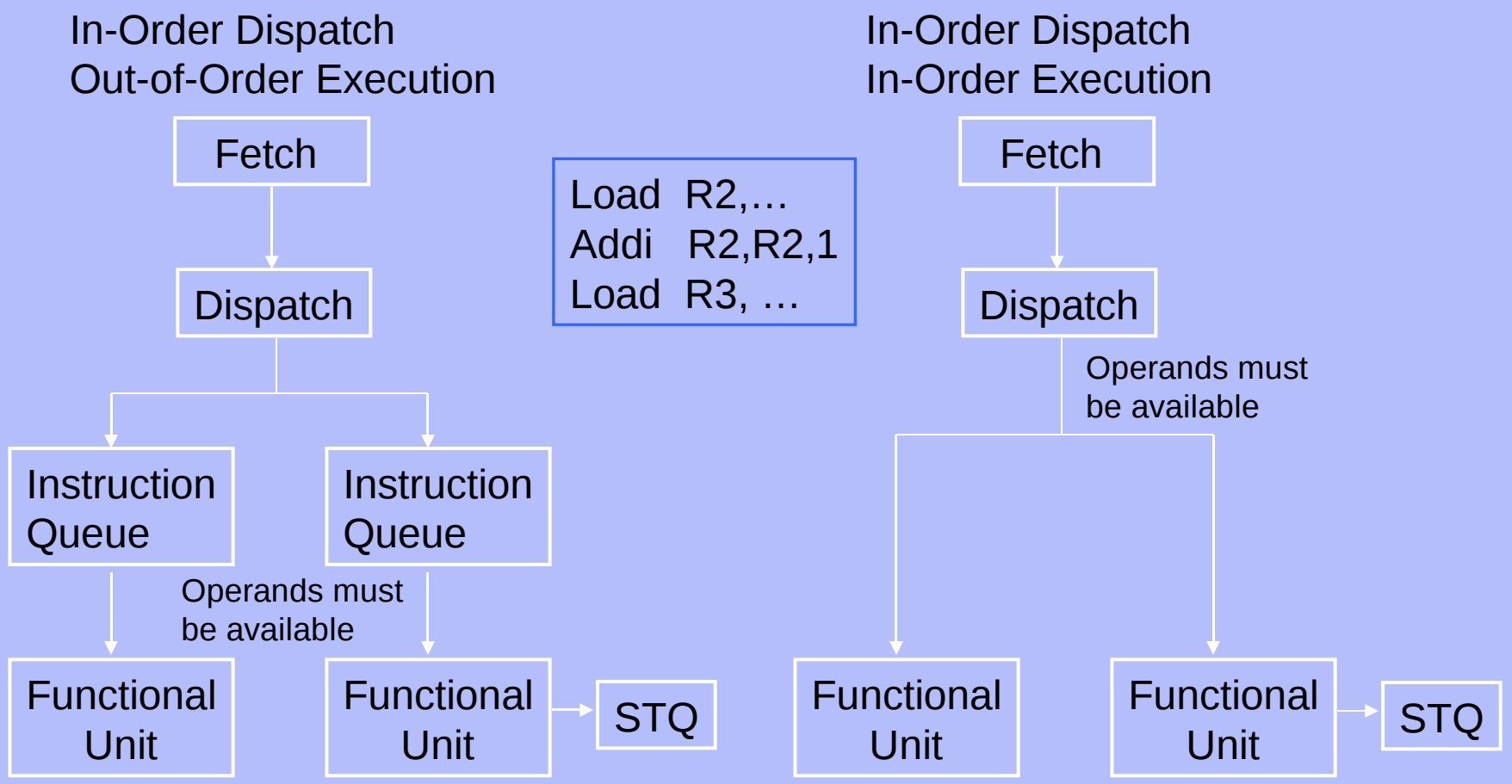
Sacrifices associated with high frequency

- In-order design instead of out-of-order
- Longer pipelines
  - Fp-compare latency
- FX multiply done in FPU
- Memory bandwidth reduced
  - Esp. Store queue draining bandwidth

Mitigating this:

- SMT implementation is improved

In-Order vs. Out of Order



P5 vs. P6 instruction groups

Both machines are “superscalar,” meaning that instructions are formed into groups and dispatched together. There are two big differences between P5 and P6:

- On P5 there are few restrictions on what types on insns and how many can be grouped together, while on P6 each group must fit in the available resources (2 fp, 2 fx, 2 ldst,..)
- On P5, insns are placed in queues to wait for operands to be ready, while on P6 dispatch must wait until all operands are ready for the whole group

→ On p6, cycles are shorter, but there are more stall cycles and more partial instruction groups

Strategic NOP insertion

ORI 1,1,0 terminates a dispatch group early

Suppose at cycle 10 the current group contains
FMA, FMA, LFL and all are ready to go.

- And the only other available instruction is a LFL ready in cycle 15

Including the LFL holds up the first three –

- Could be harmless or catastrophic
- Sometimes Nops are inserted to fix this

FP compares

Fp compare has an 11-cycle latency to its use

This is a significant delay – the best defence is for the programmer to work around it – arranging lots of work between a compare and where it is used.

Several branch free sequences are available for such code.
The compiler will in some cases try to replace fp compares with fx compares.

Store Queue optimizations

- On Power6 stores are placed into a queue to await completion.
- This queue can drain 1 store every other cycle
- If the next store in the queue is to consecutive storage with the first, they both can go that cycle.
- A 2*8 byte fp store is provided to make this more convenient

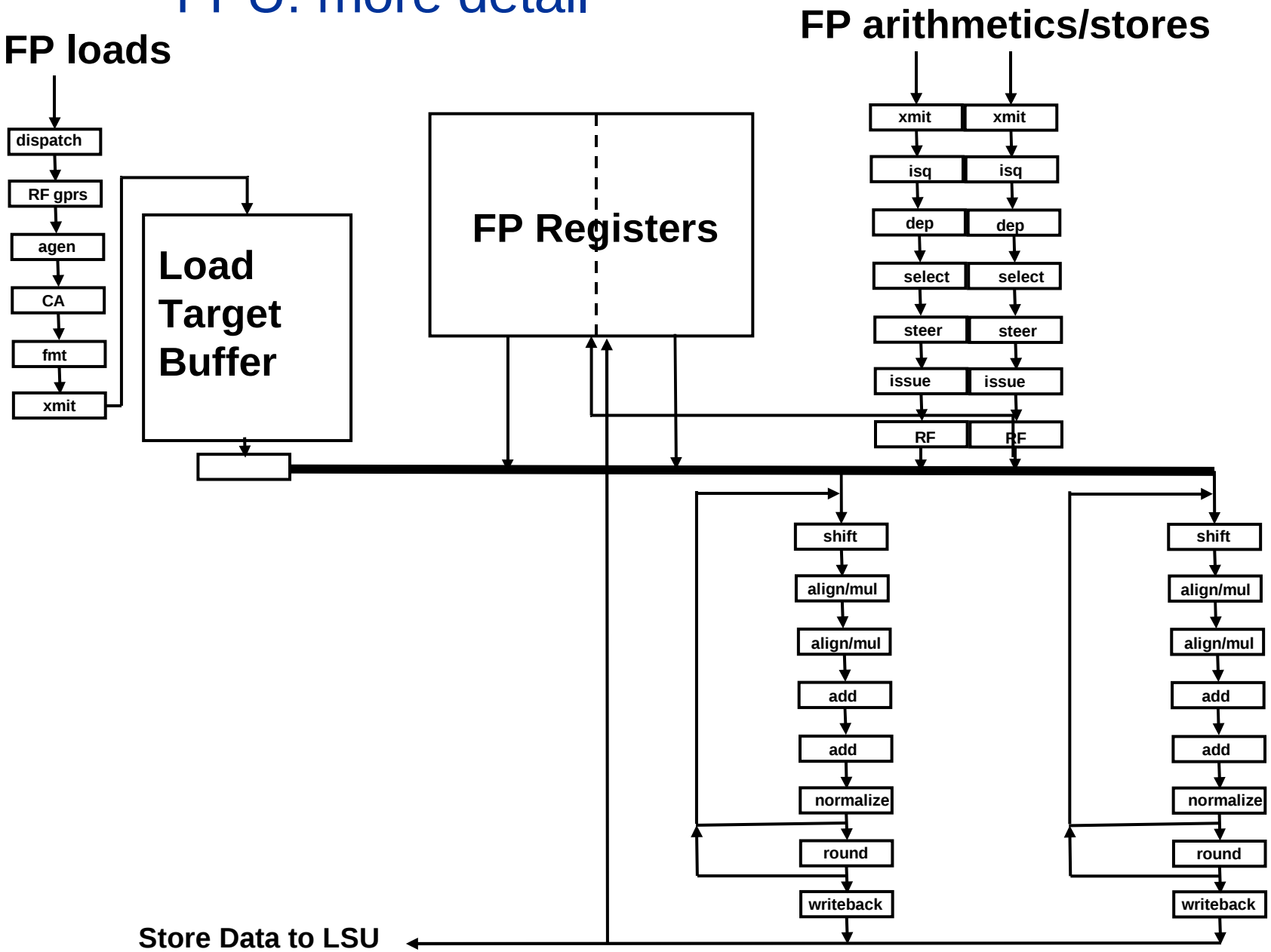
Symmetric Multithreading (SMT)

- P6 operates in two modes, ST (single thread) and SMT (multithreaded)
- In SMT mode two independent threads execute simultaneously, possibly from the same parallel program
- Instructions from both threads can dispatch in the same group, subject to unit availability
- This is highly profitable on P6 – it is a good way to fill otherwise empty machine cycles and achieve better resource usage

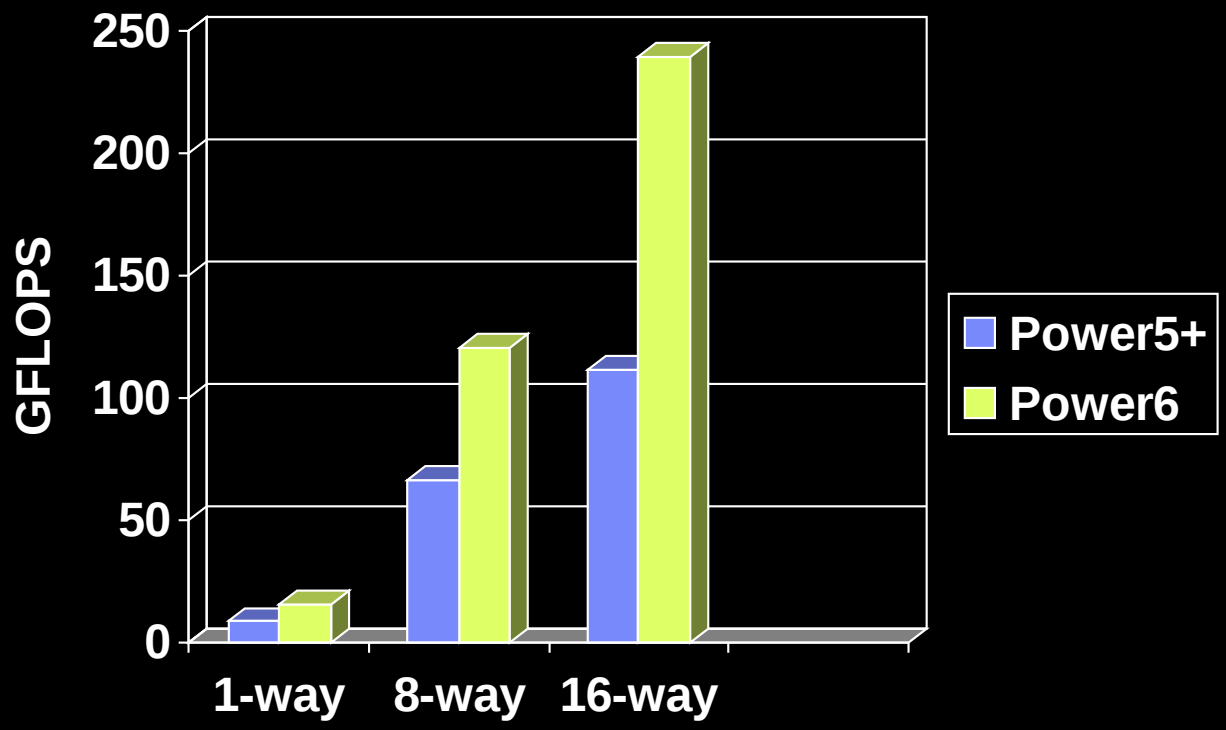
FPU details

- Two Binary Floating Point units and one Decimal Floating Point unit
- 7 stage FPU with bypass in 6 stages locally and 8 stages to other fpu
- Interleaves execution of independent arithmetic instructions with divide/sqrt
- Handles fixed point multiply/divide (pipelined)
- Always run at or behind other units
  - 10 group entry queues (group= 2ldf and 2 fp/stf instructions)
  - Allows pipelined issue of dependent store instructions
- 10 group entry load target buffer allowing speculative FP load execution
- Multiple Interrupt modes
  - Ignore, imprecise non-recoverable, imprecise recoverable, precise

FPU: more detail



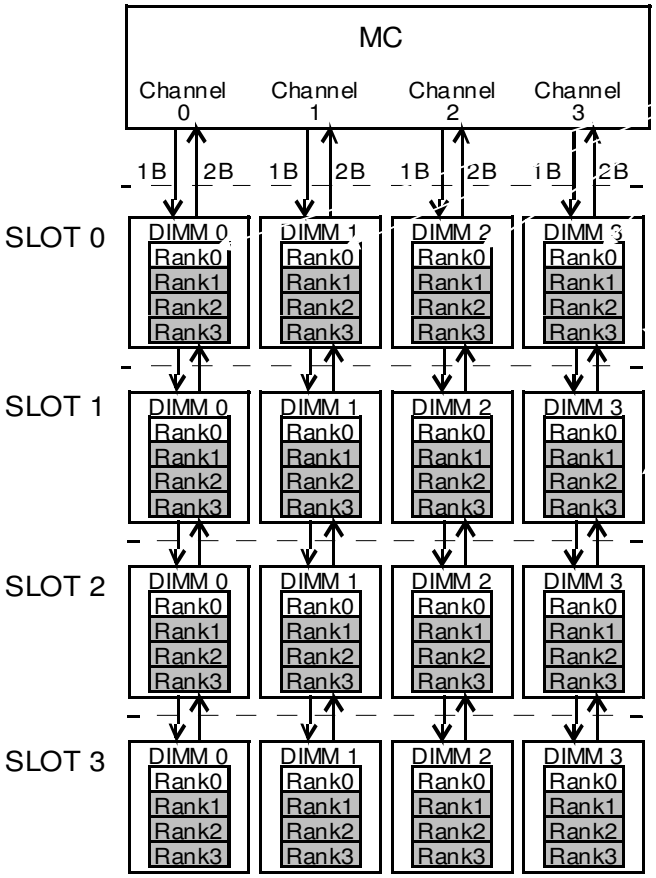
Linpack HPC Performance



Datastream Prefetching in POWER6

- Load-Look-Ahead (LLA) mode
  - Triggered on a data cache miss to hide latency
- Datastream prefetching with no filtering
  - Sixteen (16) Prefetch Request Queues (PRQs)
  - Prefetch 2 lines ahead to L1, up to 24 to L2, demand paced
  - Speculative L2 prefetch for next line on cache miss
  - Depth/aggressiveness control via SPR and dcbt instructions
    - ★ int dscr_ctl()
    - ★ dscrctl – change O/S default
  - Store prefetching into L2 (with hardware detection)
  - Full compliment of line and stream touches for both loads and stores
  - Sized for dual-thread performance (8 PRQs/thread)

Memory Physical Organization



DIMMs populated in quads

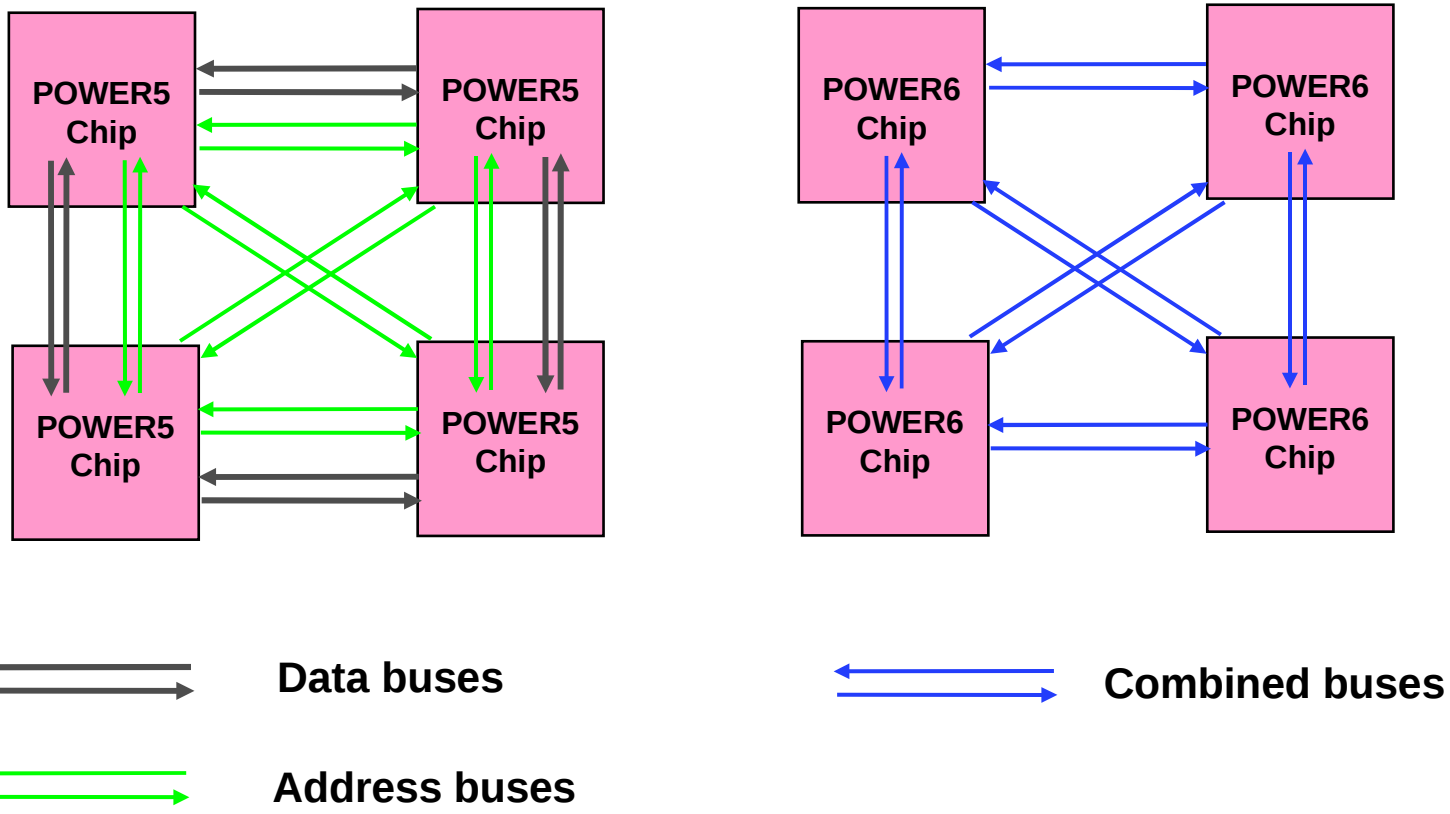
Single NOVA per DIMM (or pair of DIMMS)

DIMMs may be 1/2/4 ranks

Actual bandwidth depends on:

- DIMM speed, type
- # of ranks/DIMMs/channel
- DIMM channel b/w
- MC read/write bandwidth (from/to buffers)
- pattern of reads/writes presented to MC

POWER5 vs POWER6 bus architecture

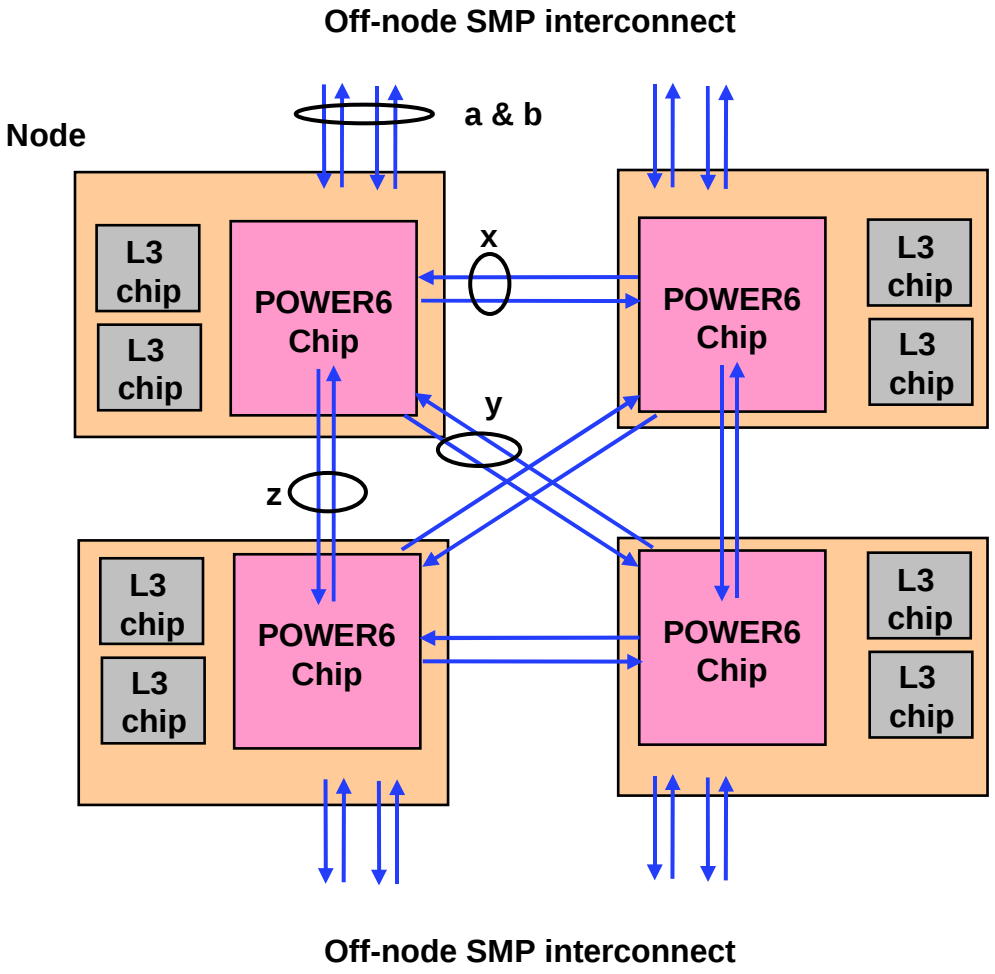


POWER6 Eight-Processor Node

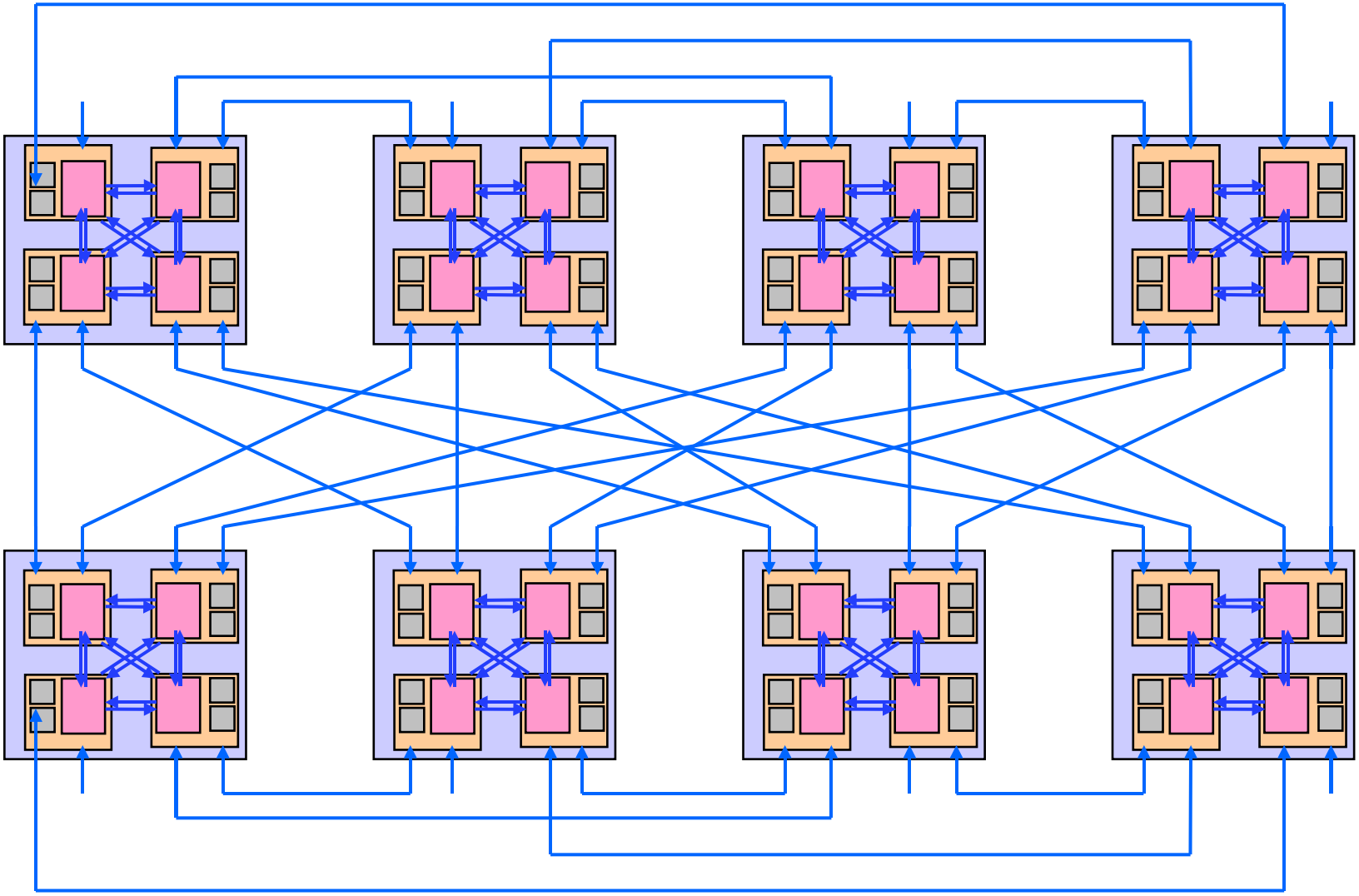
Building block for larger systems
- 64-way system has 8 nodes

Eight off-node SMP links
- Fully connects 8 nodes in a 64-way
- Additional link for RAS

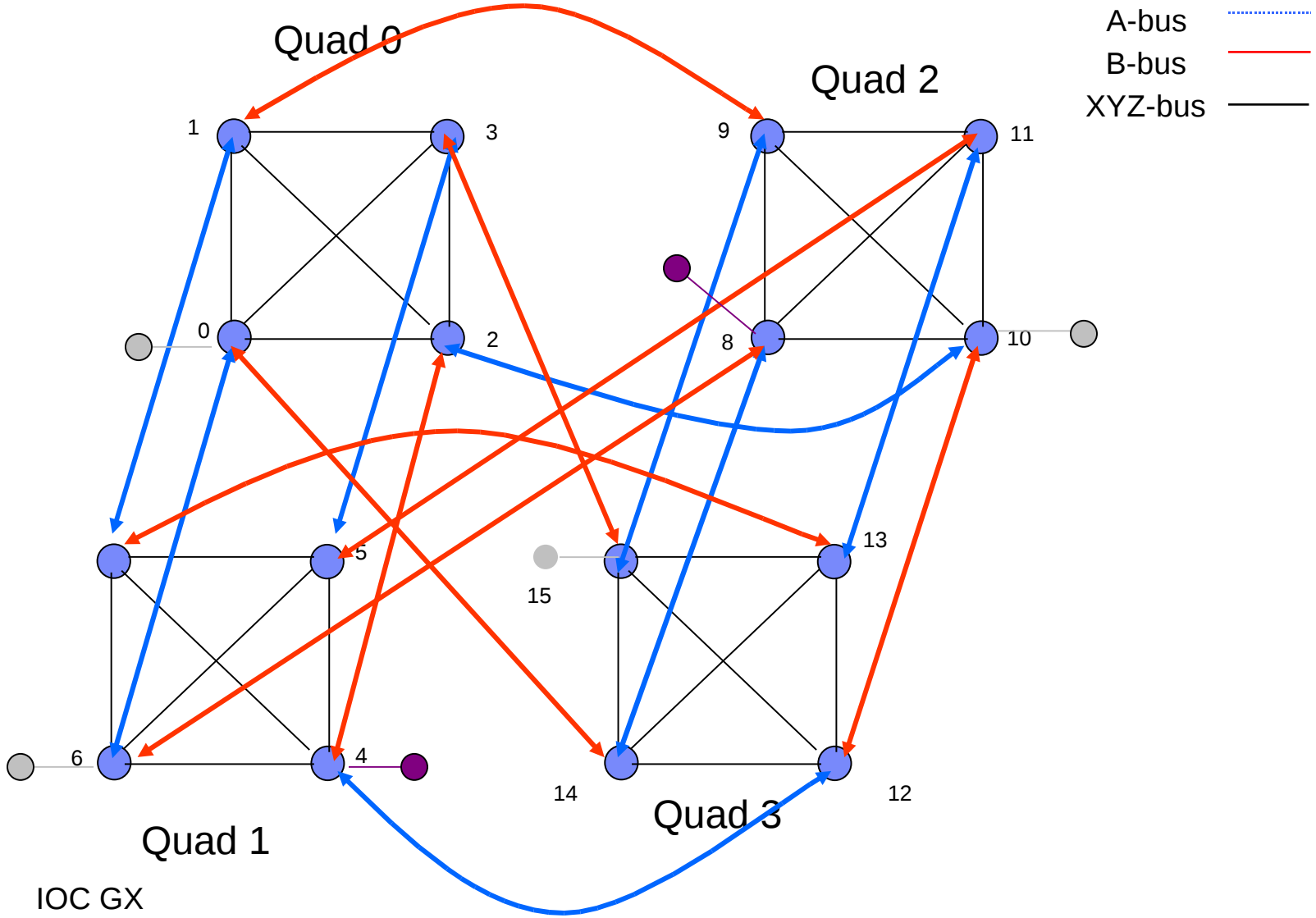
Can be used to build 128-core SMP



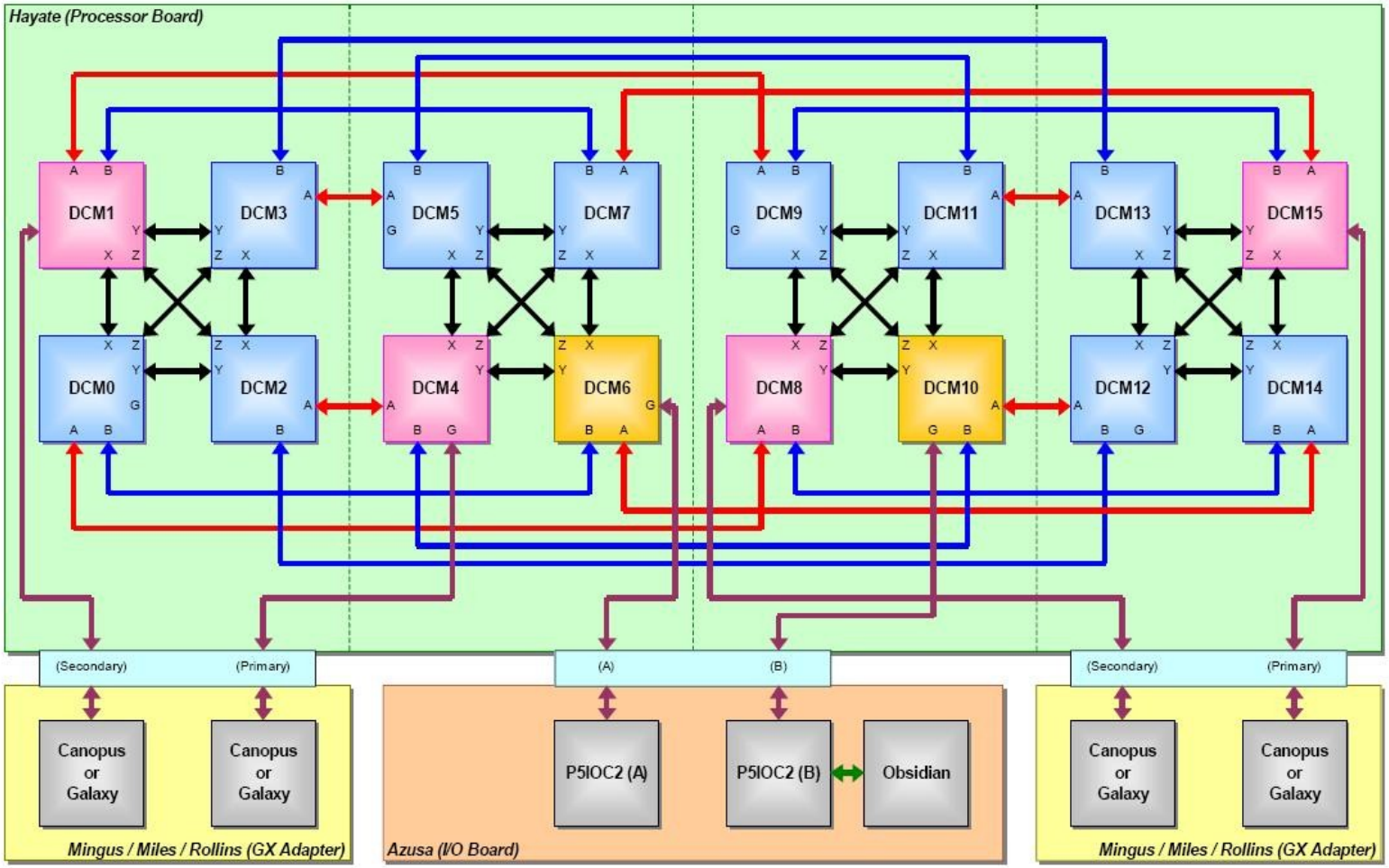
POWER6 64 way high-end system



POWER6 32-way

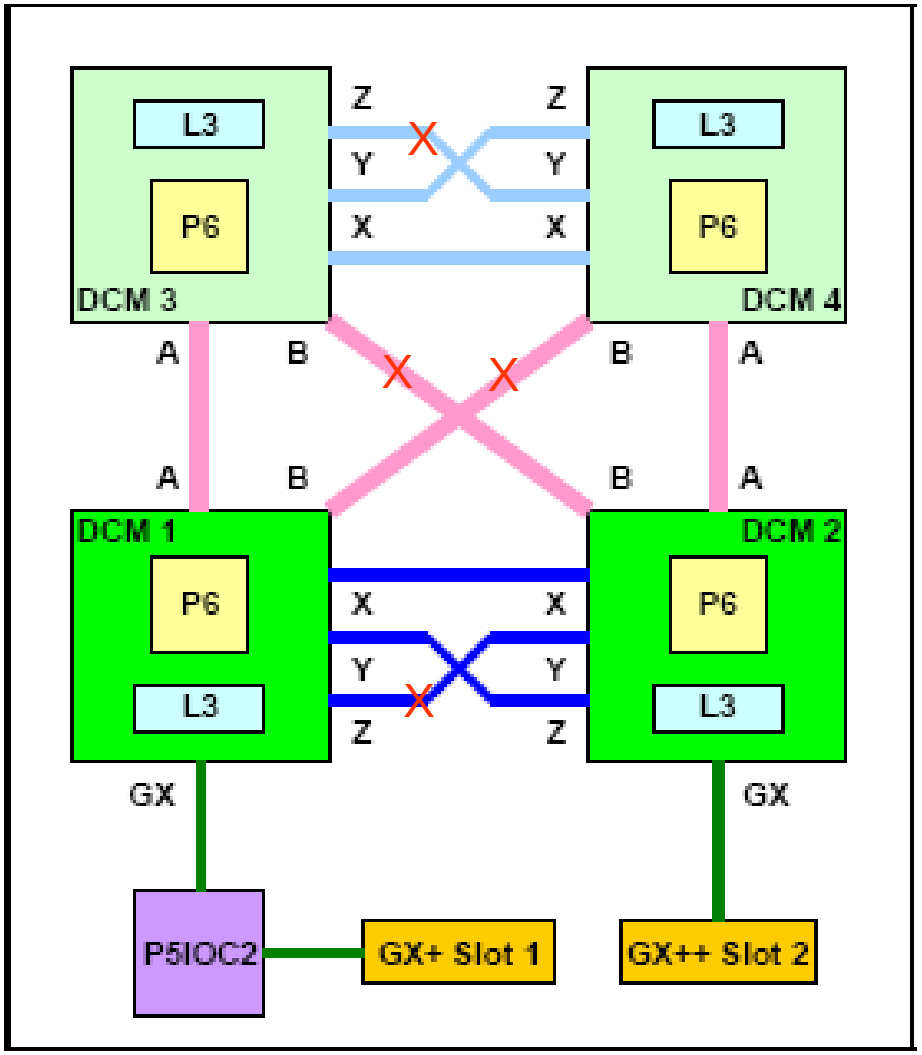


3.1.2 SMP Bus Topology



HV8

(1) DCM - 8 way maximum



X = SMP X-bus, 2 Bytes

Y = SMP Y-bus, 2 Bytes

Z = SMP Z-bus, 2 Bytes

A = SMP A-bus, 4 Bytes

B = SMP B-bus, 4 Bytes

Snoop Filtering: Broadcast Localized Adaptive-Scope Transaction Protocol (BLAST)

- Most significant microarchitectural enhancement to P6 nest
- Coherence broadcasts may be speculatively limited to subset (node or possibly chip) of SMP
  - NUMA-directory-like combination of cache states and domain state info in memory provide ability to resolve coherence without global broadcast
- Provides significant leverage for SW that exhibits node or chip level processor/memory affinity
- Several system designs significantly cost reduced by cutting bandwidth of chip interconnect

Snoop Filtering: Local/System/Double Pump

- **Send request to entire system (System Pump)**
  - Longer latency within local node
  - Shorter latency to remote nodes
  - burn address bandwidth on all nodes
- **Send request to local node only (Local Pump)**
  - better latency if data is found locally
  - uses only local XYZ address bandwidth
  - re-send to entire system if not found (Double Pump)
    - ★ Adds 100-200 pclks latency (depends on system type)
- **Local predictor uses thread ID, request type, memory address**

Tracking off-node copies

- **T/Te/Tn/Ten:**
  - “e”=not dirty / “n”=no off-node copies
- **In/Ig: on- or off-node request took line away**
  - castout Ig state to memory directory bit
- **Memory directory**
  - extra bit per cache line tells whether off node copies exist
  - MC updates when sourcing data off-node
  - allows node pump to work (MC can provide data)
  - always know from Ig or mem dir if off-node copies exist
- **Double pump**
  - No on-node cache has any info about line
  - If memory directory tells us off-node copies exist with data return (data not valid), have to do second pump

POWER6 Summary

Extends POWER leadership for both Technical and Commercial Computing

- Approx. twice the frequency of P5+, with similar instruction pipeline length and dependency
  - ★ 6 cycle FP to FP use, same as P5+
  - ★ Mostly in-order but with OOO-like features (LTB, LLA)
  - ★ 5 instruction dispatch group from a thread, up to 7 for both threads, with one branch at any position
- Significant improvement in cache-memory-latency profile
  - ★ More cache with higher associativity
  - ★ Lower latency to all levels of cache and to memory
  - ★ Enhanced datastream prefetching with adjustable depth, store-stream prefetching
  - ★ Load look-ahead data prefetching
- Advanced simultaneous multi-threading gives added dimension to chip level performance
- Predictive subspace snooping for vastly increased SMP coherence throughput

Continued technology exploitation

- CMOS 65nm, Cu SOI Technology
- All buses scale with processor frequency

Enhanced Power Management

Advanced server virtualization

Advanced RAS: robust error detection, hardware checkpoint and recovery