POWER5 Processor and System Evolution

ScicomP 11

Charles Grassl
IBM

May, 2005

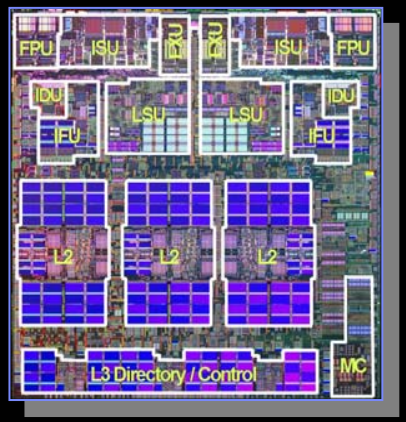
Agenda

- pSeries systems
  - Caches
  - Memory
- POWER5 Processors
  - Registers
  - Speeds
  - Design features

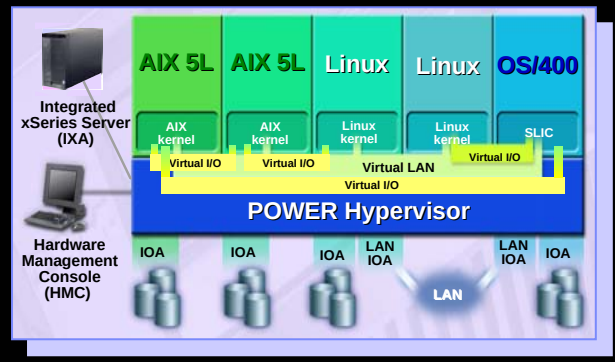
POWER5™ Design

- POWER4 base
  - Binary and structural compatibility
- Shared memory scalability
  - Up to 64 processors
  - 128 threads
- High floating point performance
- Server flexibility
  - Power efficient design
  - Utility:
    - Reliability, availability, serviceability

POWER5 Systems

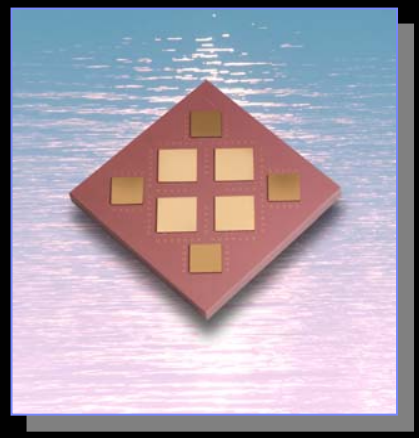


- Second generation dual core chip
- Intelligent 2-way SMT
- Power management with no performance impact
- 130 nm lithography
  - 276M transistors
  - 8 layers of metal



- Micropartitioning
  - Up to 64 physical processors,1280 virtual processors per system

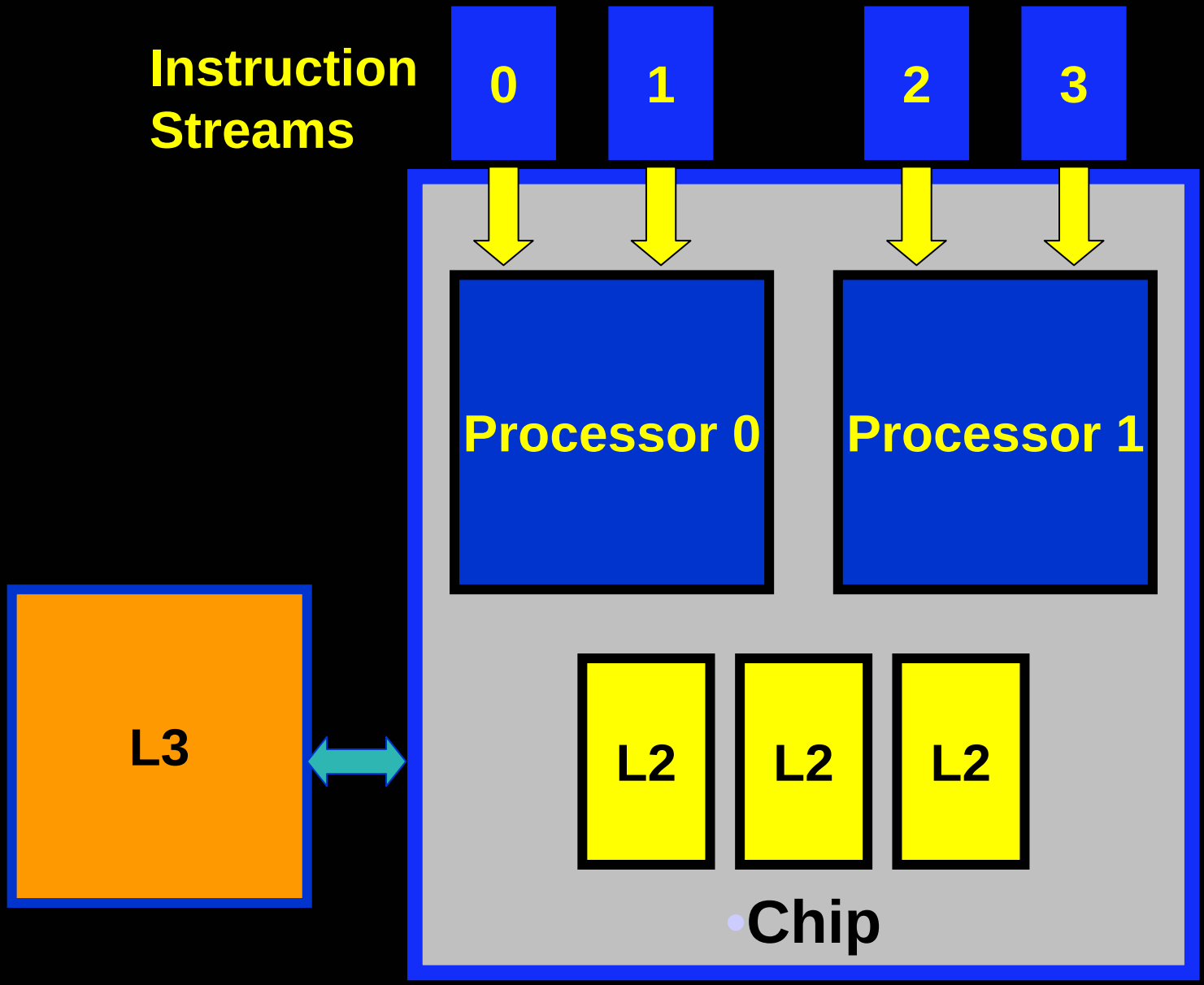
- Multi Chip Modules (MCM):
- Eight way SMP looks like 16-way to software
- 95 mm on a side



POWER5 System Features

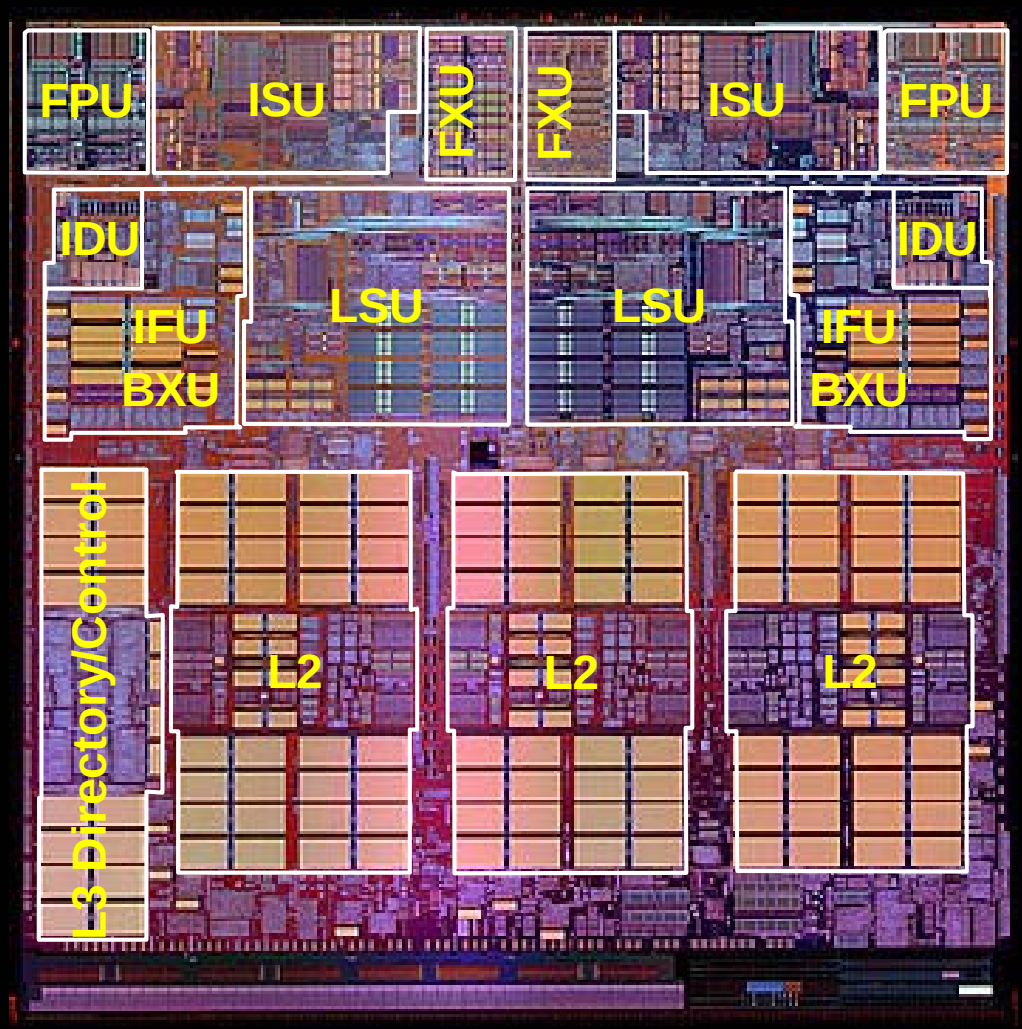
- Dual Core Chip
- Shared L2 cache
- Shared L3 cache
- Shared Memory
- Multiple Page Size support
- Simultaneous Multi Threading

POWER5 Features



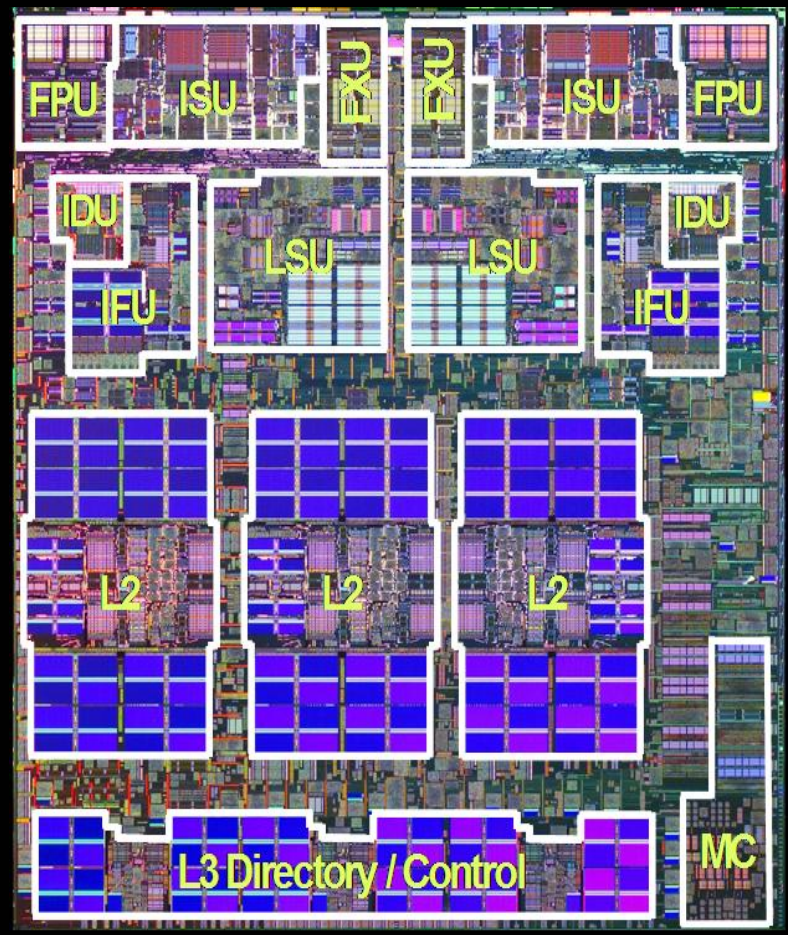
POWER4 Chip --- (December 2001)

- Technology: 180nm lithography, Cu, SOI
  - POWER4+ shipping in 130nm today
- Dual processor core
- 8-way superscalar
  - Out of Order execution
  - 2 Load / Store units
  - 2 Fixed Point units
  - 2 Floating Point units
  - Logical operations on Condition Register
  - Branch Execution unit
- > 200 instructions in flight
- Hardware instruction and data prefetch



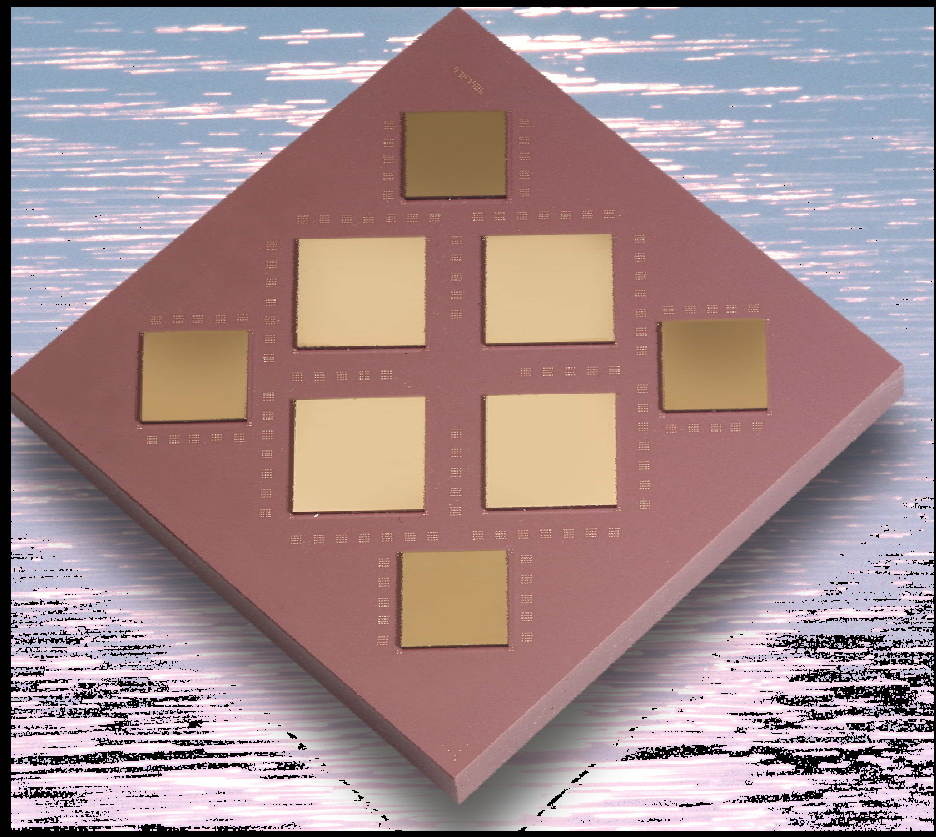
POWER5 Chip

- IBM CMOS 130nm
  - Copper and SOI
  - 8 layers of metal
- Chip
  - 389 mm^2
  - 276M transistors
  - I/Os: 2313 signal, 3057 power
  - Same technology as POWER4+



POWER5 Multi-chip Module

- 95mm × 95mm
- Four POWER5 chips
- Four cache chips
- 4,491 signal I/Os
- 89 layers of metal

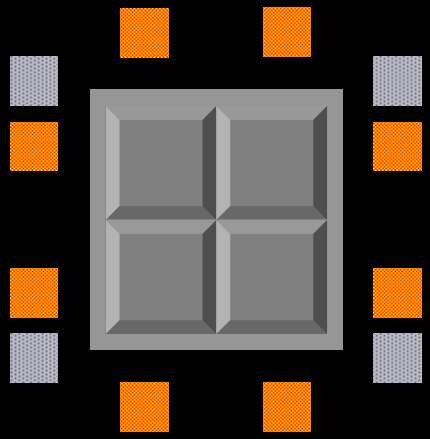


Multi Chip Module (MCM) Architecture

POWER4

- 4 processor chips
  - 2 processors per chip
- 8 off-module L3 chips
  - L3 cache is controlled by MCM and logically shared across node
- 4 Memory control chips

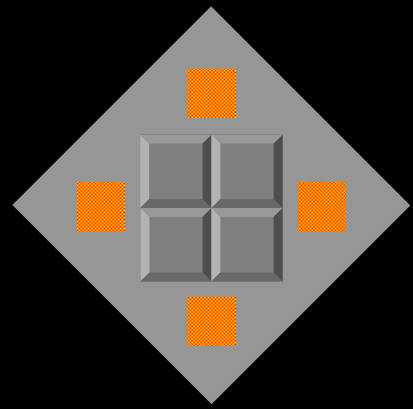
- 16 chips



POWER5

- 4 processor chips
  - 2 processors per chip
- 4 L3 cache chips
  - L3 cache is used by processor pair
  - “Extension” of L2

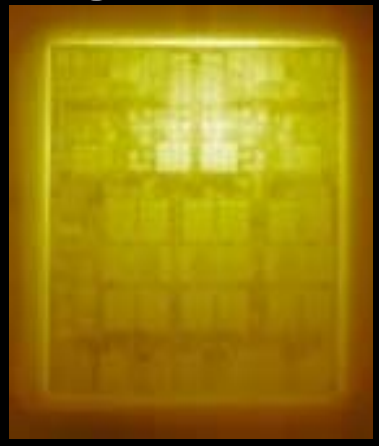
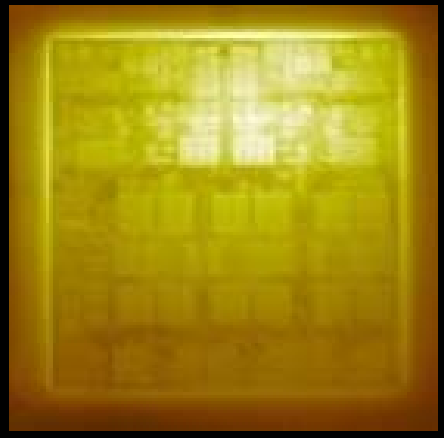
- 8 chips



Dynamic Power Management

- Two components:
  - Switching power
  - Leakage power
- Impact of SMT on power:
  - More instructions executed per cycle
- Switching power reduction:
  - Extensive fine-grain, dynamic clock-gating
- Leakage power reduction
  - Minimal use of low Vt devices
- No performance impact
- Low power mode for low priority threads

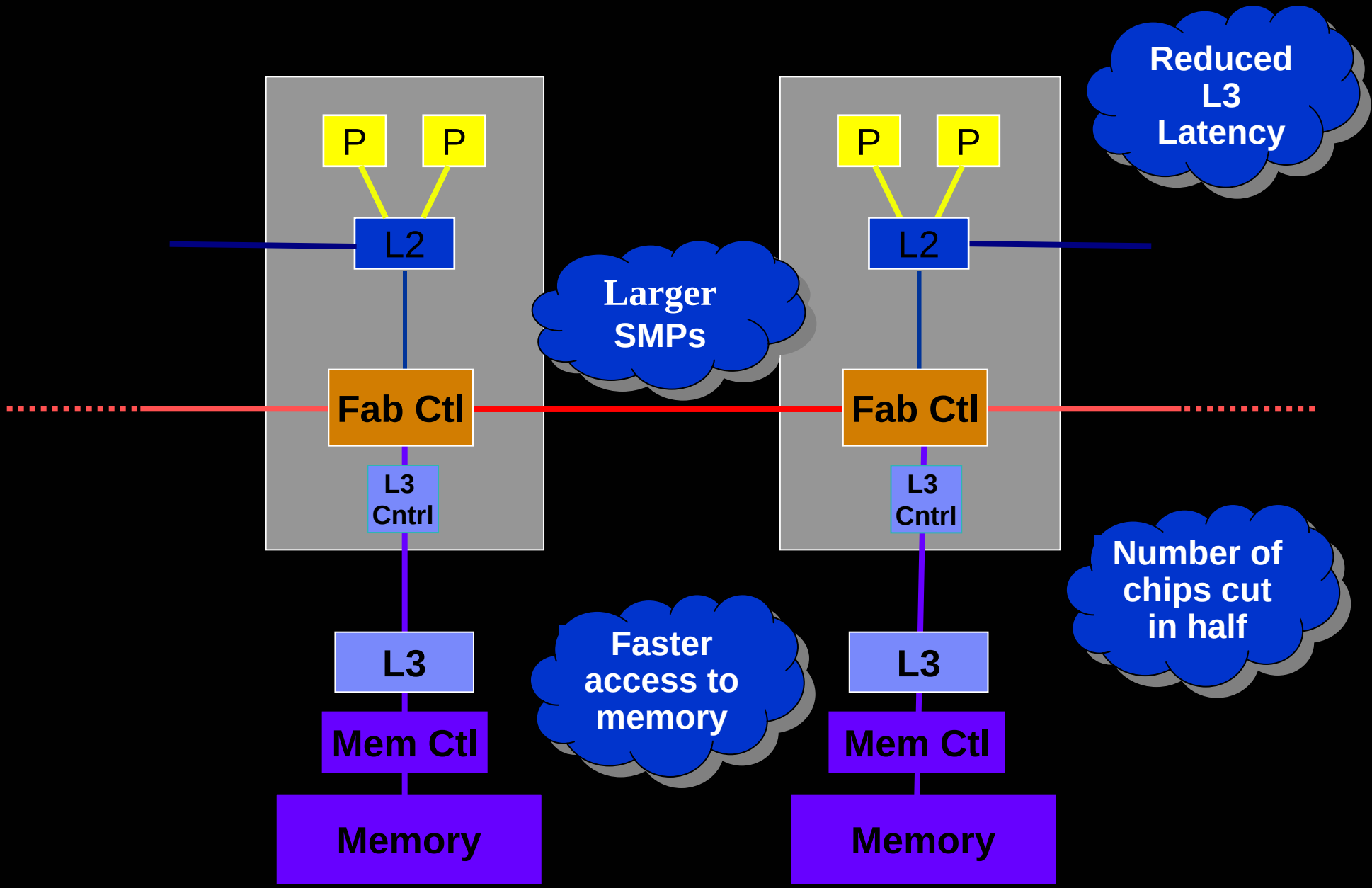
Dynamic Power Management

	No Power Management	Dynamic Power Management
Single Thread		
Simultaneous Multi-threading		

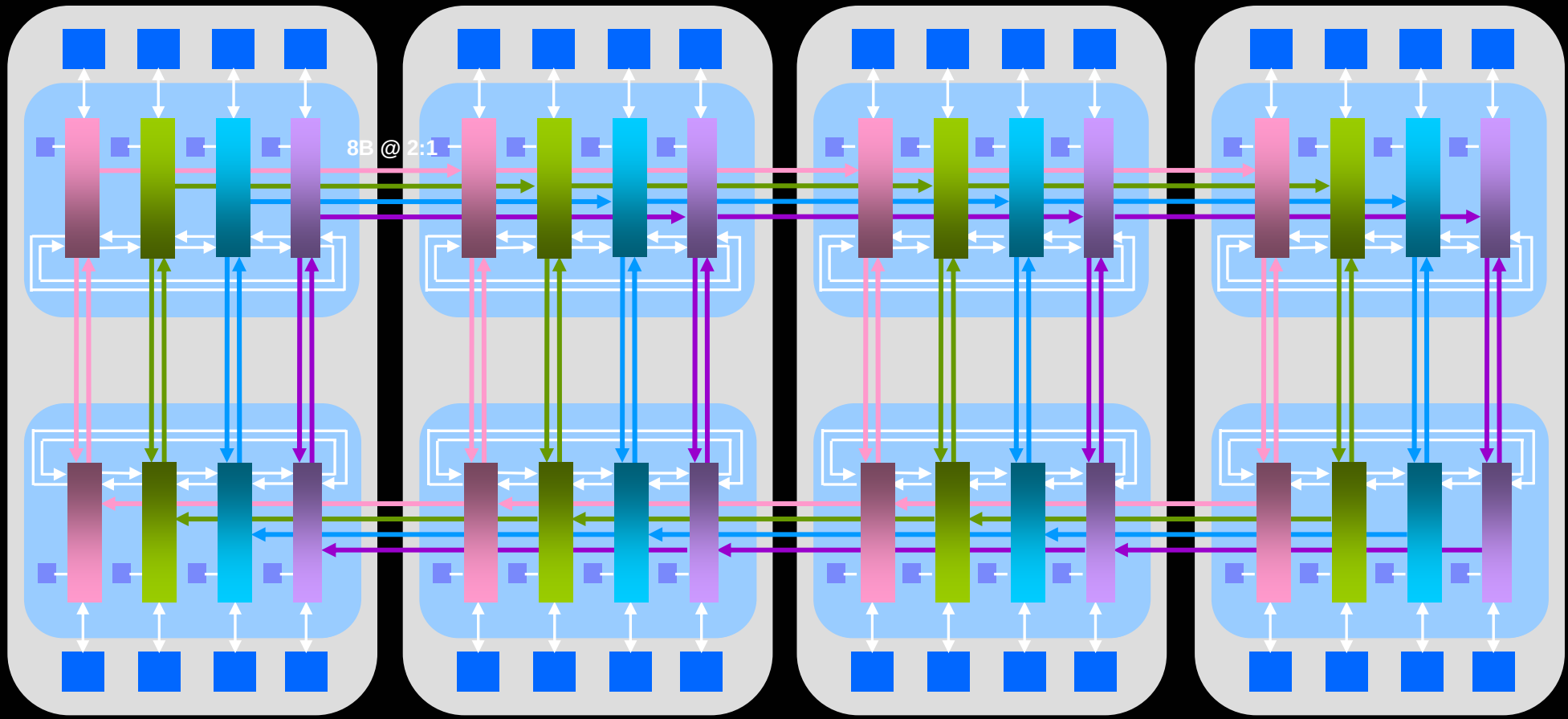
Photos taken with thermal sensitive camera while prototype POWER5 chip was undergoing tests

Simultaneous Multi-threading with dynamic power management reduces power consumption below standard, single threaded level

Modifications to POWER4 to create POWER5



64-way SMP Interconnection

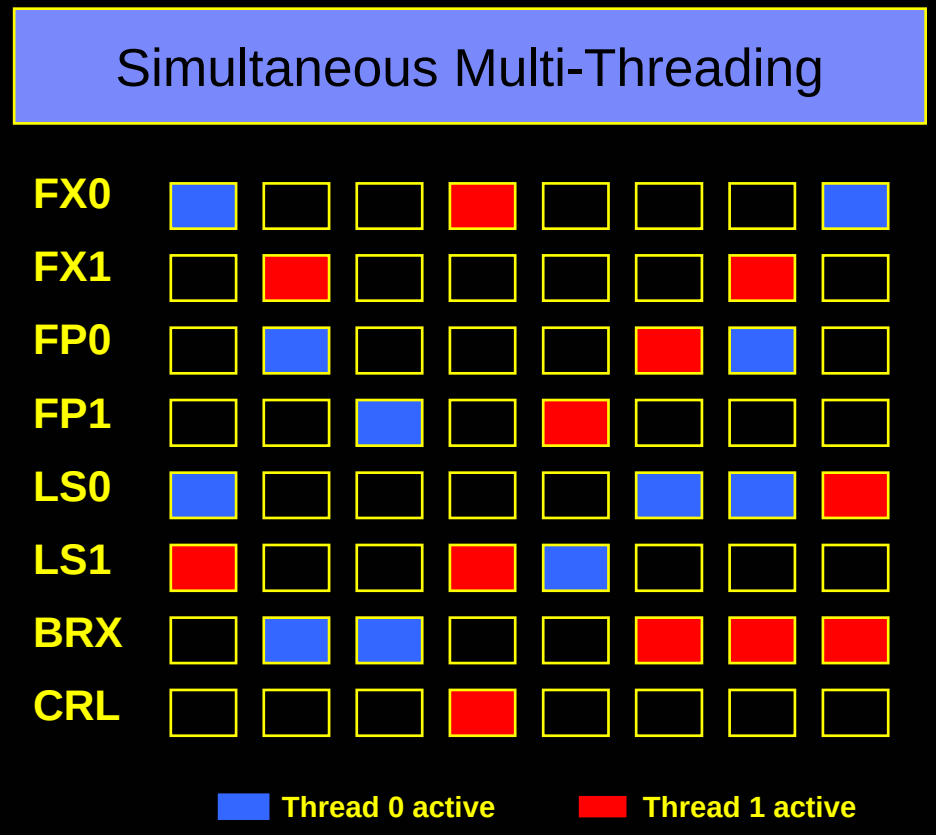


Interconnection exploits *enhanced distributed switch*

- **All chip interconnections operate at half processor frequency and scale with processor frequency**

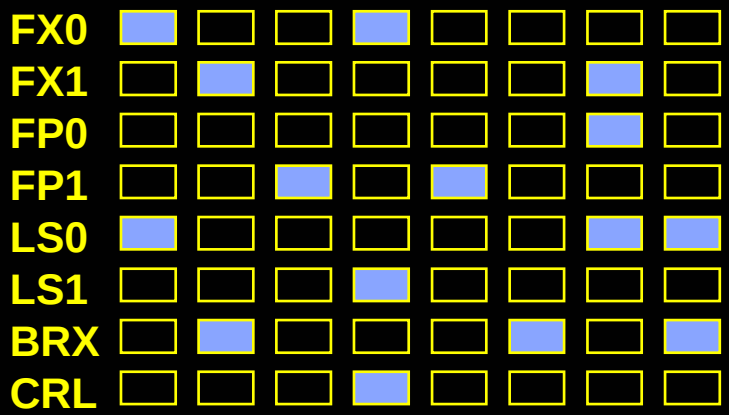
Simultaneous Multi-Threading in POWER5

- Each chip appears as a 4-way SMP to software
  - 2 processors
    - 2 threads per processor
- Processor resources optimized for enhanced SMT performance
- Software controlled thread priority
  - Dynamic feedback of runtime behavior to adjust priority
- Dynamic switching between single and multithreaded mode

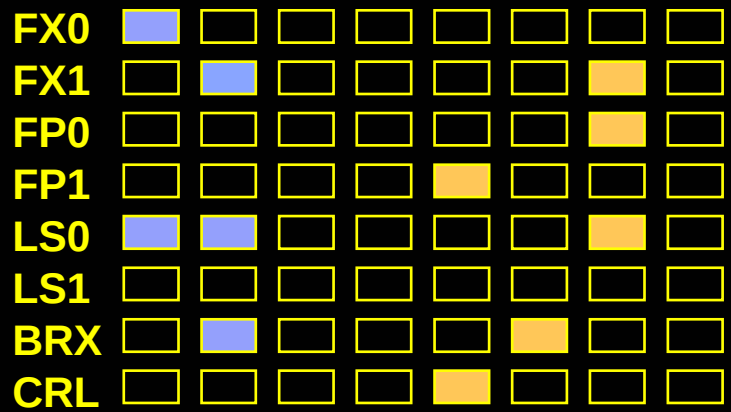


Multi-threading Evolution

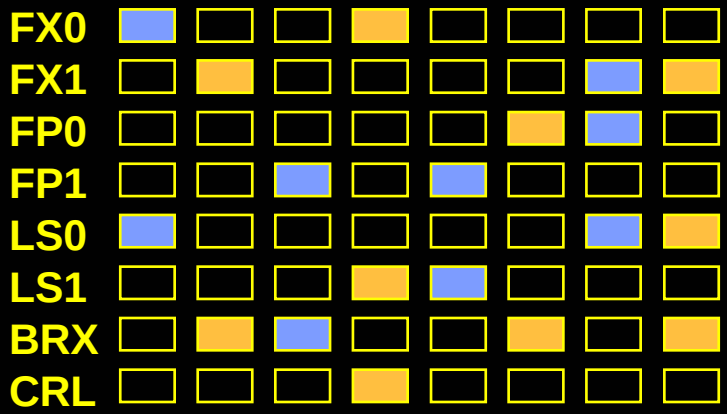
Single Thread



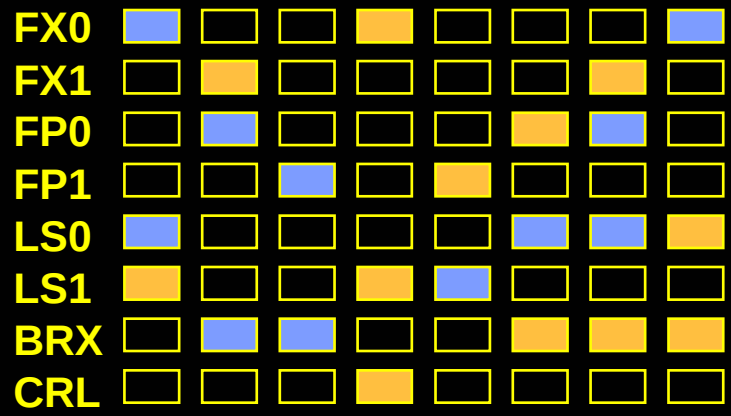
Coarse Grain Threading



Fine Grain Threading



Simultaneous Multi-Threading



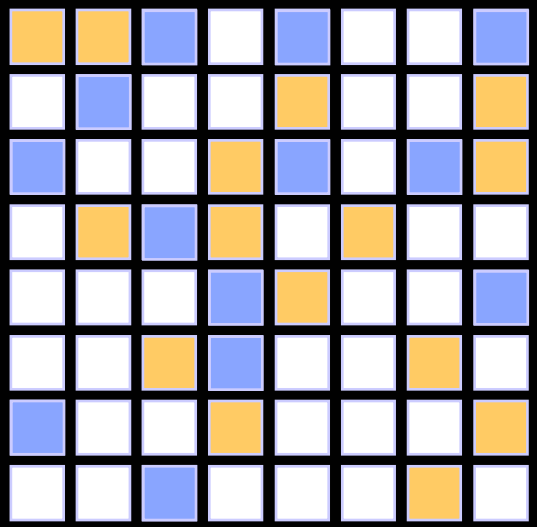
Thread 0 Executing

Thread 1 Executing

No Thread Executing

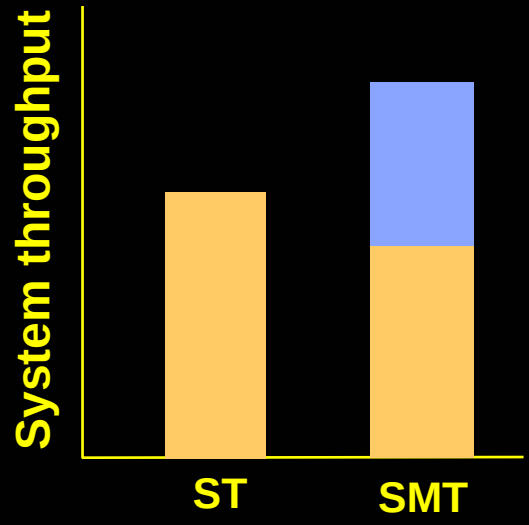
Simultaneous multi-threading

POWER5 Simultaneous Multi Threading



Thread0 active
No thread active
Thread1 active

Appears as 4 CPUs per chip to the operating system (AIX 5L V5.3 and Linux)



- Utilizes unused execution unit cycles
- Symmetric multiprocessing (SMP) programming model
- Natural fit with superscalar out-of-order execution core
- Dispatch two threads per processor. Net result:
  - Better processor utilization

POWER5 Performance Expectations

- Higher sustained-to-peak floating point rate ratio compared to POWER4
- Reduction in L3 and memory latency
  - Integrated memory controller
- Increased rename resources
  - Higher instruction level parallelism in compute intensive applications
- Fast barrier synchronization operation
- Enhanced data prefetch mechanism

Processor Architecture

- CPUs
- Caches
- Performance Features

Chip Enhancements

- Caches and translation resources
  - Larger caches
  - Enhance associativity
- Resource pools
  - Rename registers: GPRs, FPRs increased to 120 each
  - L2 cache coherency engines: increased by 100%
- Memory controller moved on chip
- Dynamic power management

New POWER5 Instructions

- Enhanced data prefetch (eDCBT)
- Floating-point:
  - Non-IEEE mode of execution for divide and square-root
  - Reciprocal estimate, double-precision
  - Reciprocal square-root estimate, single-precision
- Population count

Processor Characteristics

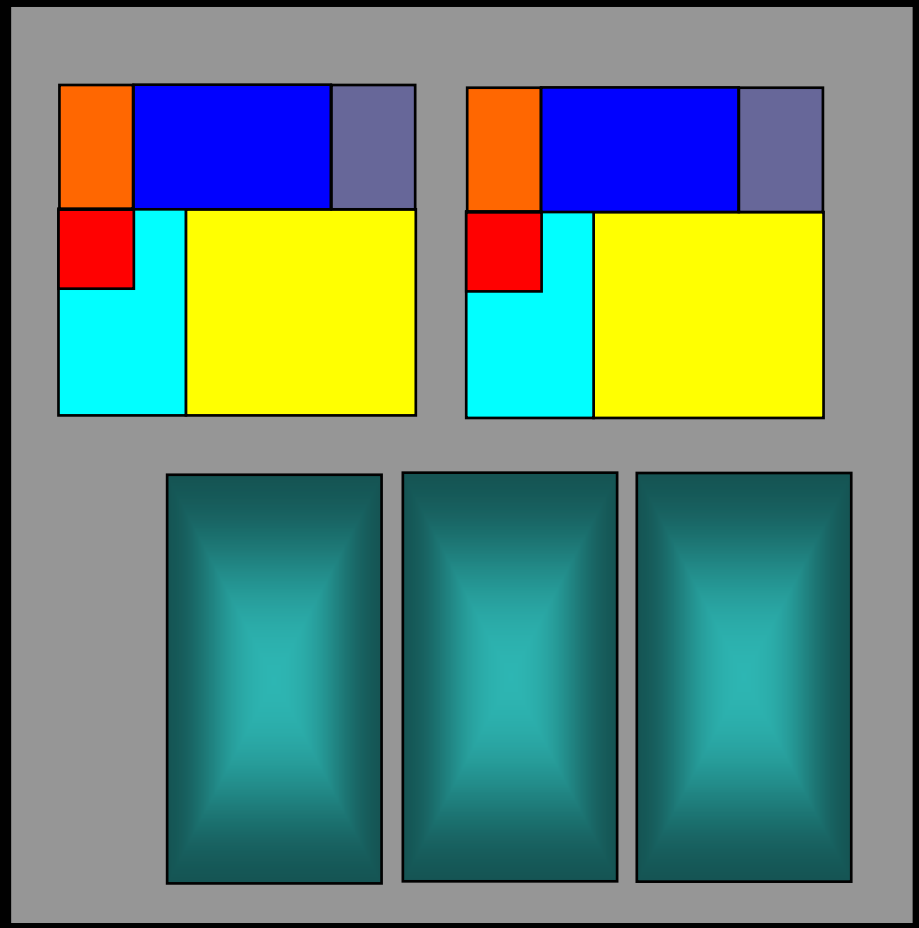
- Deep pipelines
- High frequency clocks
- High asymptotic rates
- Superscalar
- Speculative out-of-order instructions
- Up to 8 outstanding cache line misses
- Large number of instructions in flight
- Branch prediction
- Prefetching

POWER4 and POWER5 Storage Hierarchy

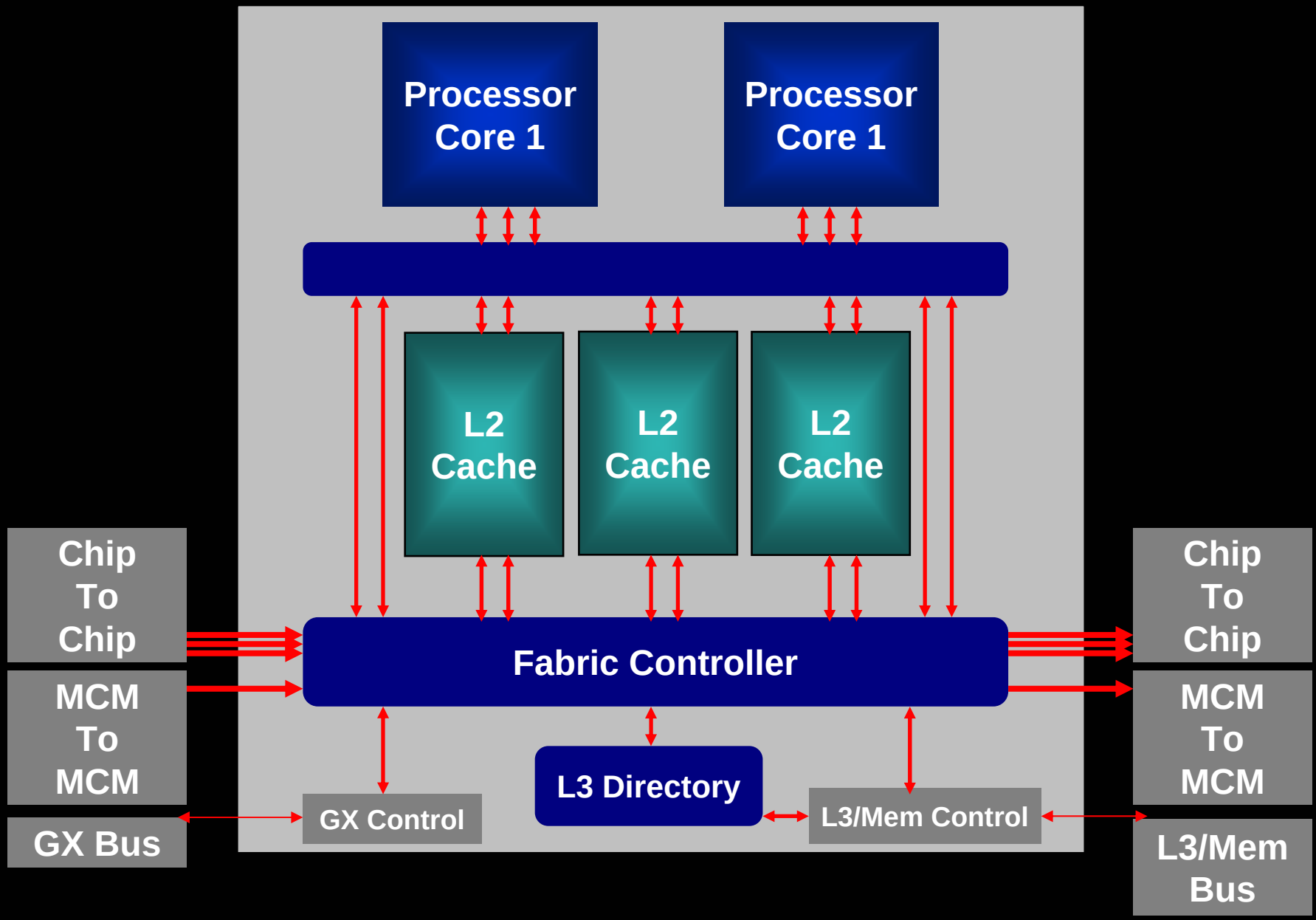
	POWER4	POWER5
L2 Cache		
Capacity, line size	1.44 Mbyte 128 byte line	1.92 Mbyte 128 byte line
Associativity, replacement	8-way, LRU	10-way, LRU
Off-chip L3 Cache		
Capacity, line size	32 Mbyte 512 byte line	36 Mbyte 256 byte line
Associativity, replacement	8-way, LRU	12-way, LRU
Chip interconnect		
Type	Distributed switch	Enhanced distributed switch
Intra-MCM data buses	½ processor speed	Processor speed
Inter-MCM data buses	1/3 processor speed	½ processor speed
Memory	1 Tbyte	2 Tbyte

Multiprocessor Chip

- 2 CPUs (processors) on one chip
- Each processor:
  - L1 cache
    - Data
    - Instruction
- Each chip:
  - Shared memory path
  - Shared L3 cache
    - 32 Mbyte
  - Shared L2 cache
    - 1.5 Mbyte



Chip Structure

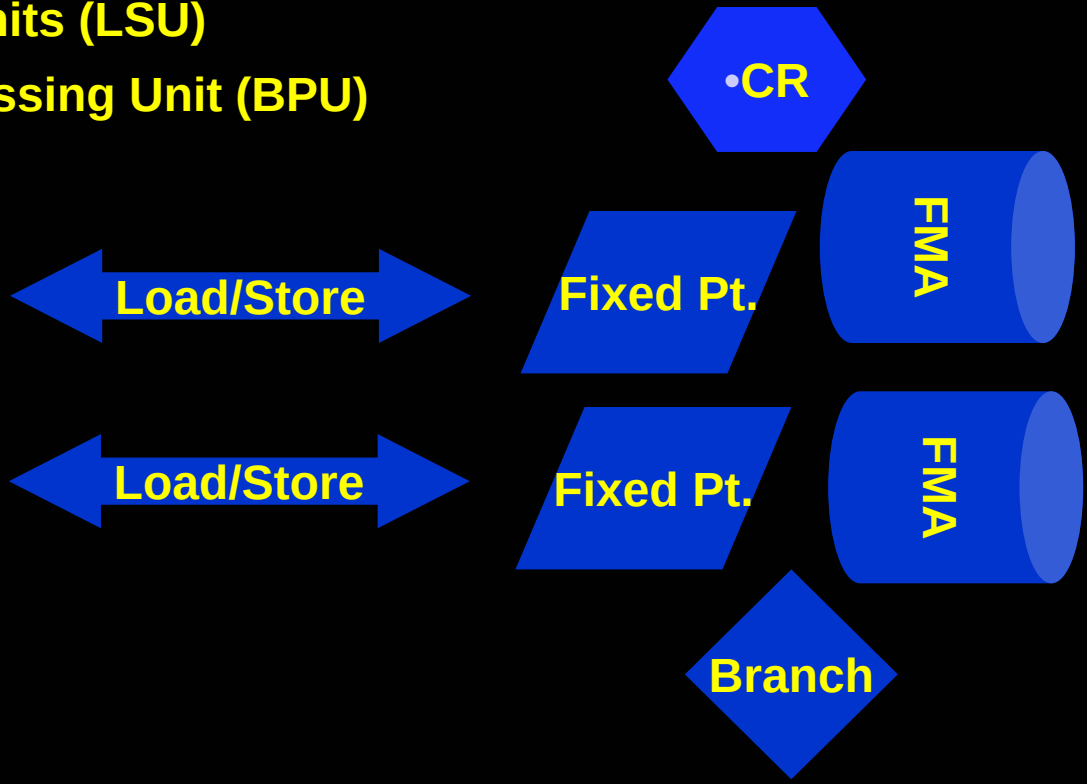


Micro Architecture

- 64-bit RISC Microprocessor
- Multiple Execution Units
- Hardware Data Prefetch
- Out-of-Order Execution
- Speculative Execution
- 8 Instructions / Cycle

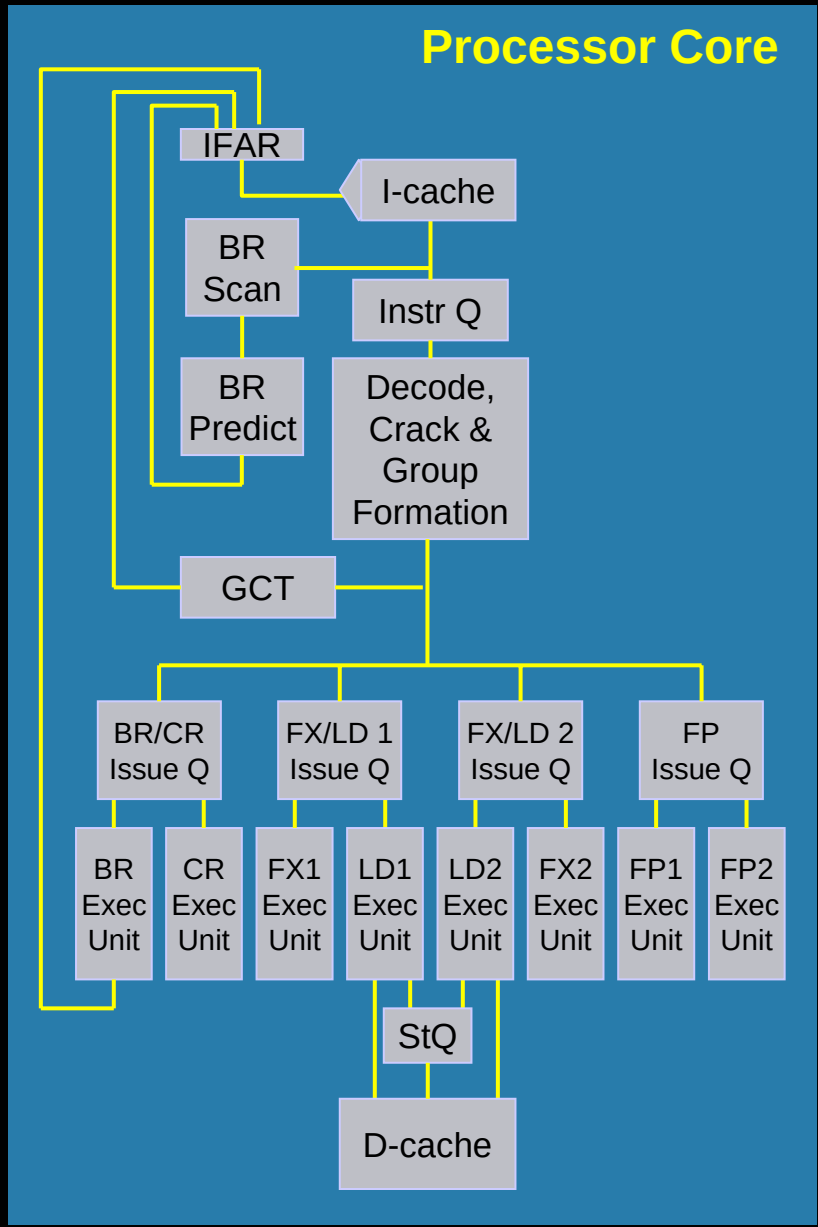
Multiple Functional Units

- Symmetric functional units
  - Two Floating Point Units (FPU)
  - Three Fixed Point Units (FXU)
    - Two Integer
    - One Control
  - Two Load/Store Units (LSU)
  - One Branch Processing Unit (BPU)



Fast Core: Instruction-level Parallelism

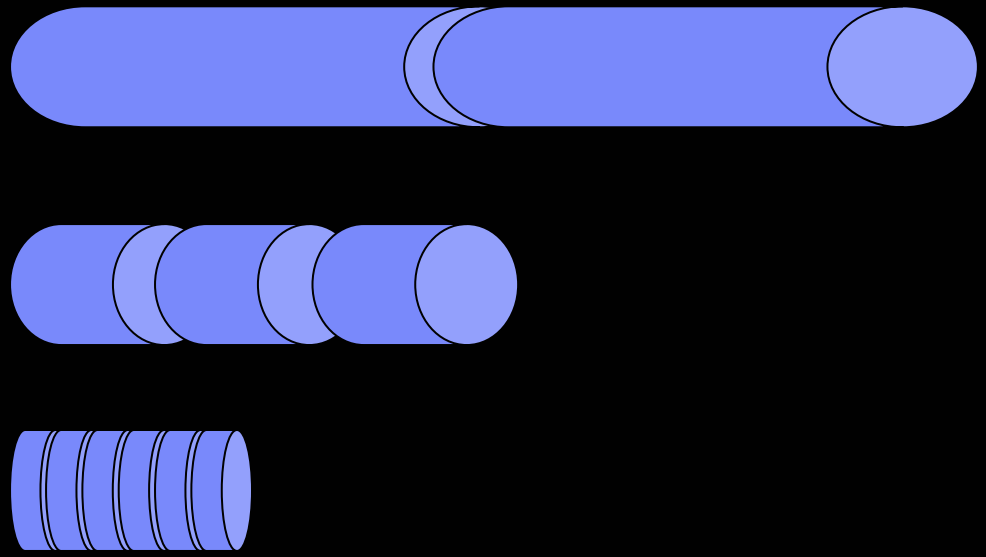
- Speculative superscalar organization
  - Out-of-Order execution
  - Large rename pools
  - 8 instruction issue, 5 instruction complete
  - Large instruction window for scheduling
- 8 Execution pipelines
  - 2 load / store units
  - 2 fixed point units
  - 2 DP multiply-add execution units
  - 1 branch resolution unit
  - 1 CR execution unit
- Aggressive branch prediction
  - Target address and outcome prediction
  - Static prediction / branch hints used
  - Fast, selective flush on branch mispredict



Registers

Resource	Logical	POWER4: Physical	POWER5: Physical
GPRs	32	80	120
FPRs	32	72	120
CRs	8 (9) 4-bit fields	32	32
Link/Count	2	16	16
FPSCR	1	20	20
XER	4 fields	24	24

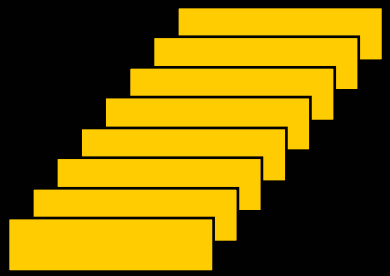
Functional Unit Progression

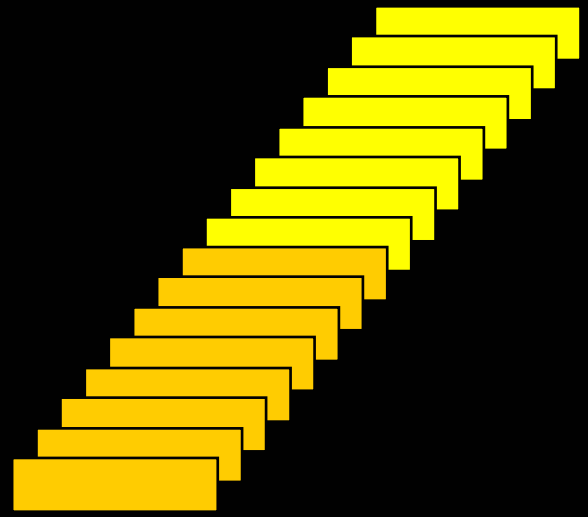


	POWER2	POWER3	POWER4	POWER5
Clock Periods	2	3	6	6
Clock Rate	125 MHz	375 MHz	1.3 GHz	1.9 GHz
Time (Nanosec.)	16	8	4.6	3.2

Registers

- CPU's point of view
  - 120 FP registers (POWER5)
- User point of view
  - 32 FP registers (architecture)
- Rename registers
  - Relieve register "pressure"





Register Renaming

- Architecture has 32 registers
  - Legacy
- Cases which require additional registers:
  - Tight loops
    - Computationally intensive
  - “Broad” loops
    - Many variables involved
  - Deep pipe lines
- Renaming registers are increasingly important with Simultaneous MultiThreading

Register Renaming: Read After Write

$R_{13} = R_{14} + R_{15}$

Nothing to be done

...

$R_{16} = R_{13} + R_{12}$

Register Renaming: Write After Write

$R_{13} = R_{14} + R_{15}$

...

$R_{13} = R_{16} + R_{17}$

$R_{19} = R_{13} + R_{18}$

$R_{13} = R_{14} + R_{15}$

...

$R_{42} = R_{16} + R_{17}$

$R_{19} = R_{42} + R_{18}$

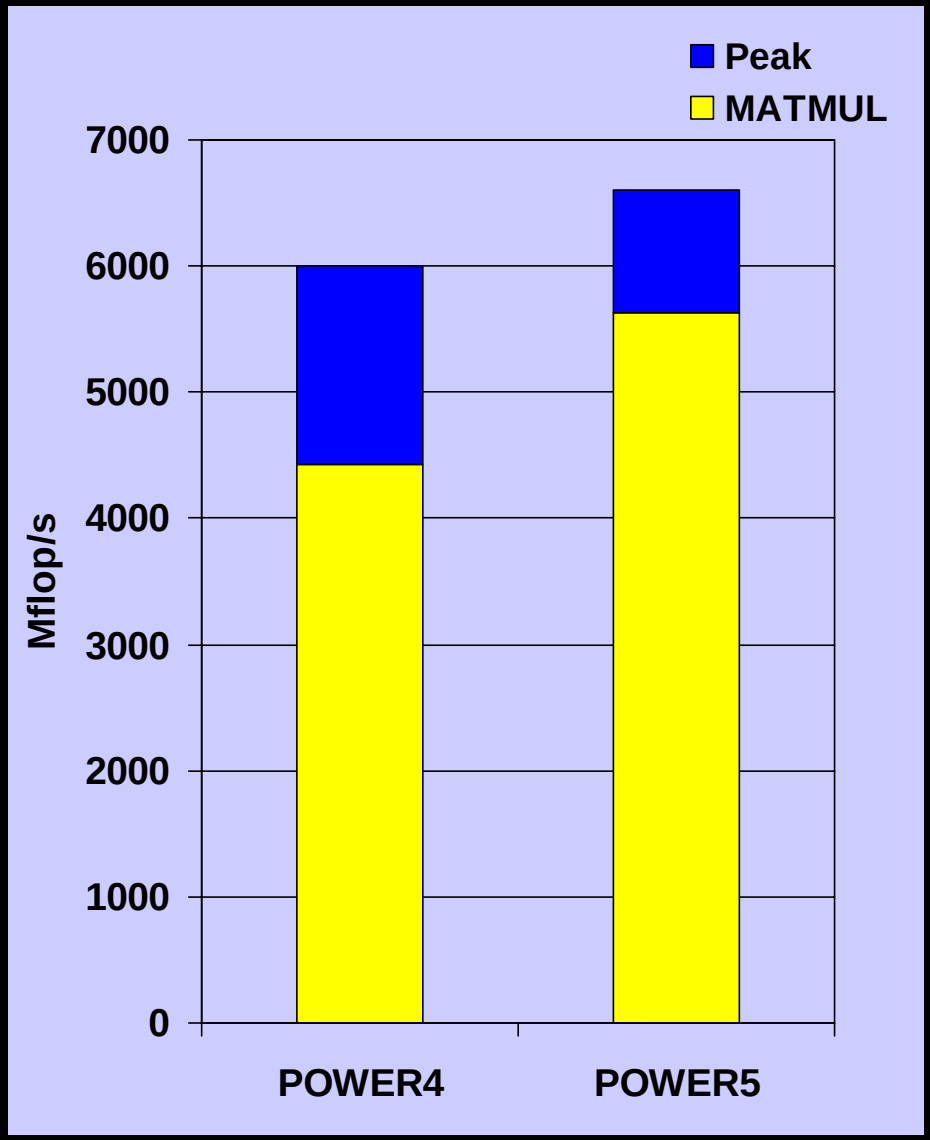
Effect of Registers

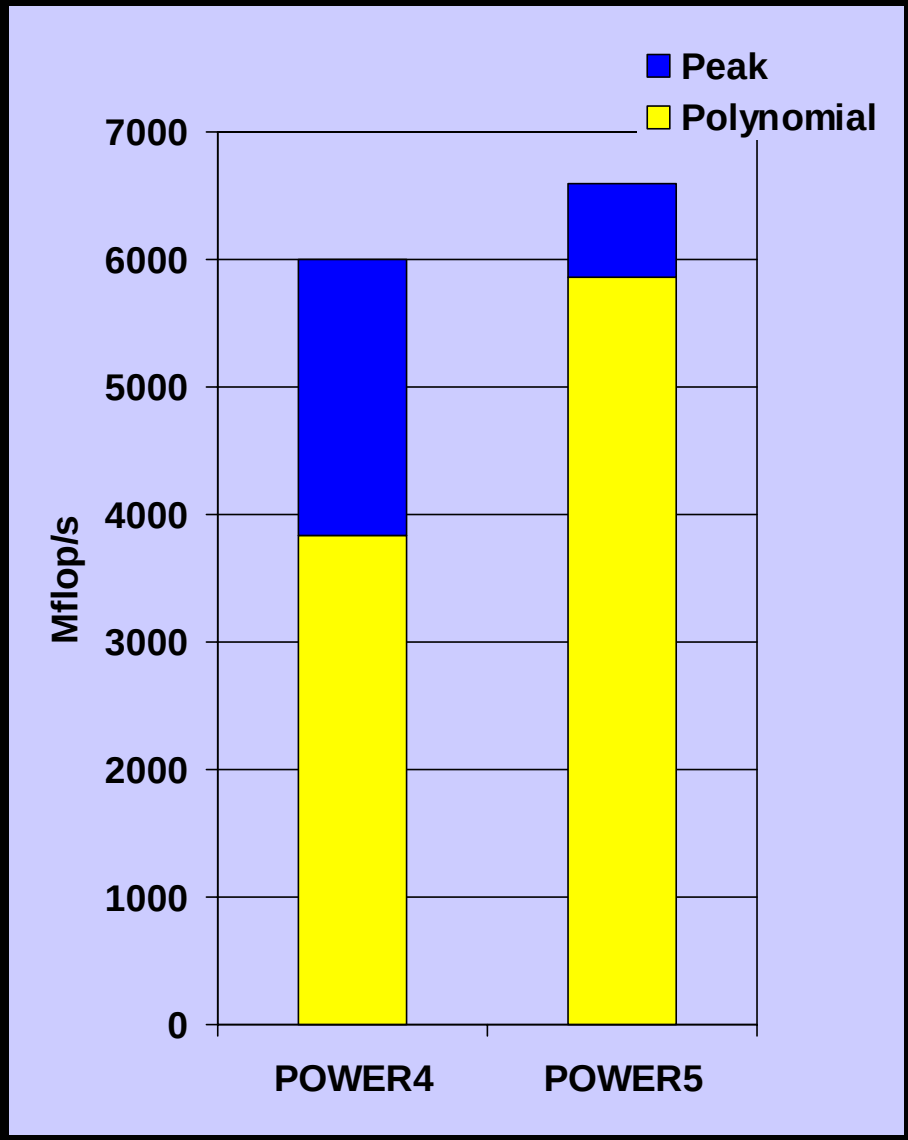
	POWER4	POWER5
GP Registers	80	120
FP Registers	72	120
DGEMM speed	60% of burst	90% of burst

Renaming Example

- Matrix multiple with 4x unrolling
  - 2 FMAs and 1 LFD per cycle
  - 3 renames per cycle
  - After 13 cycles, 39 FP renames of the 40 (POWER4, 72-32) are allocated
  - Cycles 14, 16, and 18:
    - Instruction are rejected due to lack of renames
- Result: ~70 to 75% of peak
  - Software rule of thumb:
  - approx. 13 renames available every 6 cycles
- POWER5 alleviates this with 120 FP renames

Effect of Rename Registers





POWER4 @ 1.5 GHz POWER5 @ 1.65 GHz

Floating Point Functional Units

- Two floating point execution units
  - Divide and square root sub-units
  - NOT pipelined
  - Double precision (64-bit) data path

Instruction	Single (Cycles)	Double (Cycles)
Fma	6	6
Fdiv	~25	32
Fsqrt	-	34

Floating Point Functional Units

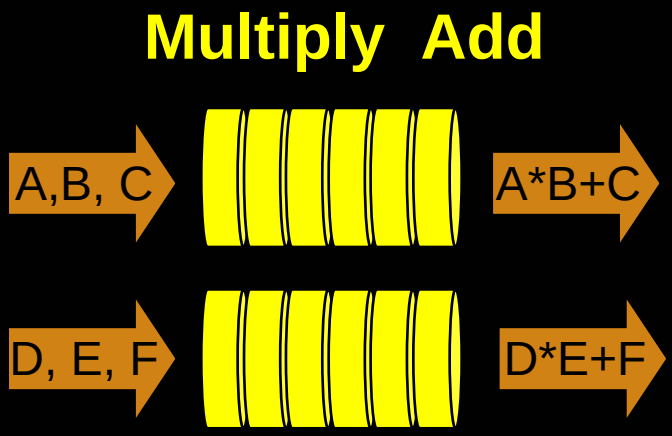
- 2 floating add-multiply (FMA) units
  - Per instruction:
    - 1 floating point add
    - 1 floating point multiply
  - 4 floating point ops per clock period
- IEEE arithmetic

Arithmetic

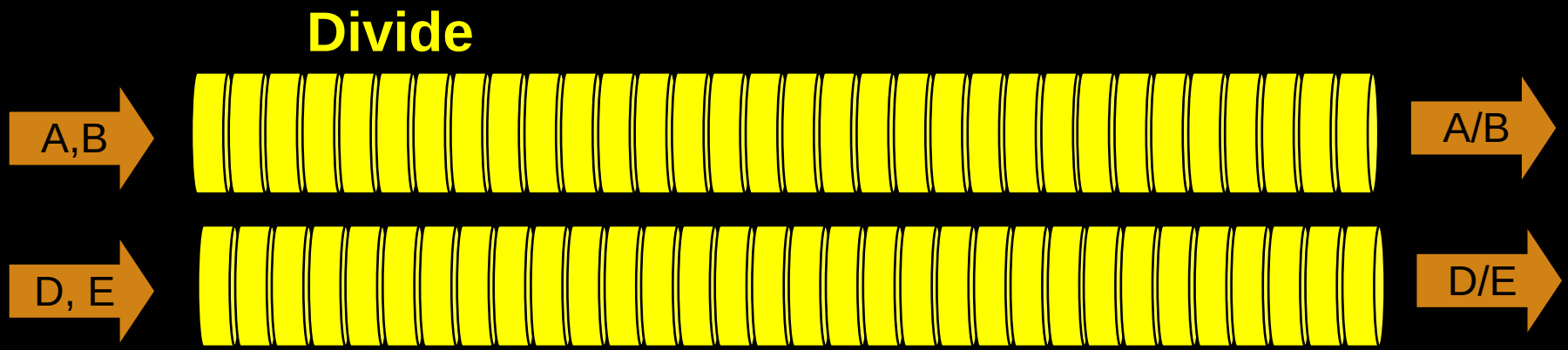
- IEEE 754 single and double floating-point
- Floating multiply-add:
  - Intermediate value is not rounded
- 64-bit integer arithmetic instructions
  - Used only in 64-bit addressing mode

Size	Integer	Floating Point
16	Yes	No
32	Yes	Yes
64	Yes (-q64)	Yes
128	No	No

Pipelined Functional Units



Multiply or Add	12 results/ 6 clock periods
Divide	2 results/ 32 clock periods
Square root	2 results/ 34 clock periods

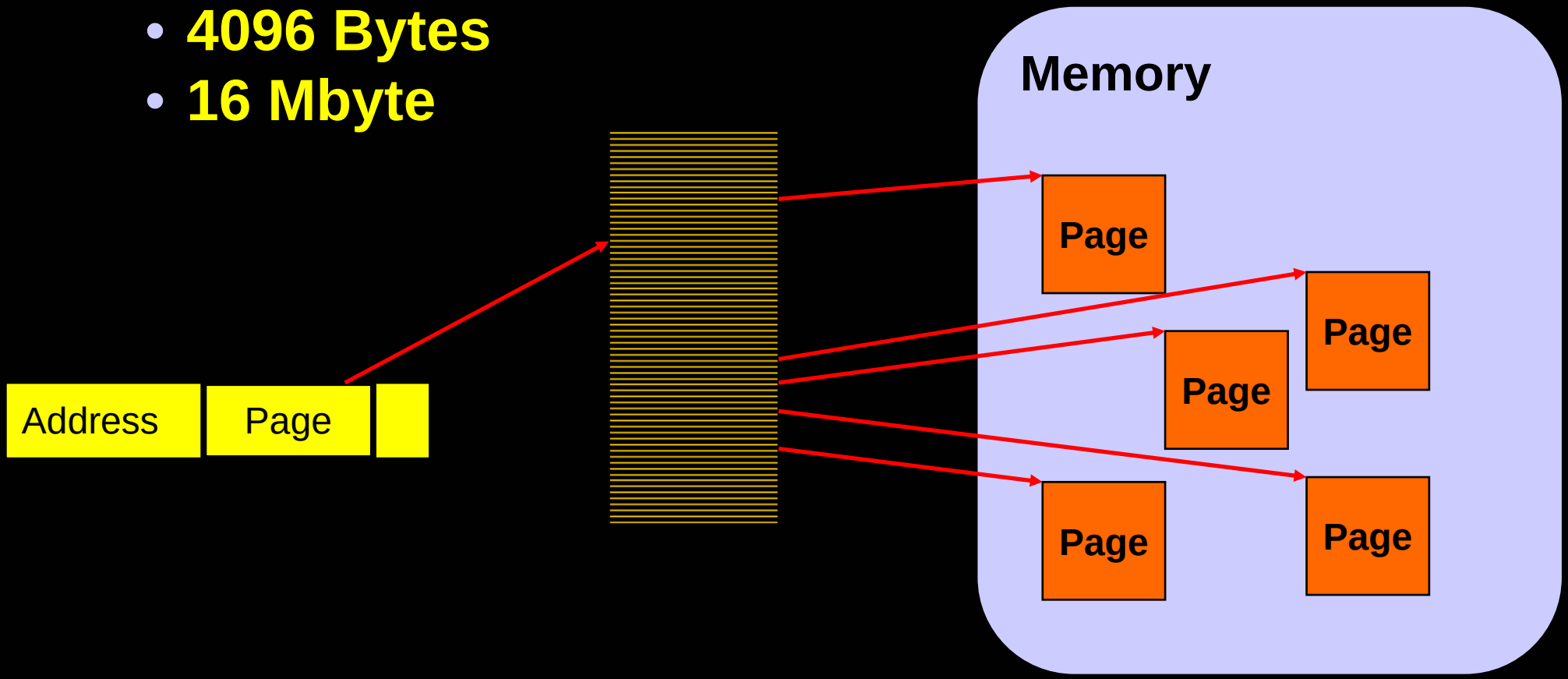


Deep Pipelines

- Operations limited by functional unit transit time:
  - Divide
  - Square root
  - Intrinsic functions
  - Recursion

Translation Lookaside Buffer (TLB)

- 1024 entry
- Page sizes:
- 4096 Bytes
- 16 Mbyte



TLB Thrashing

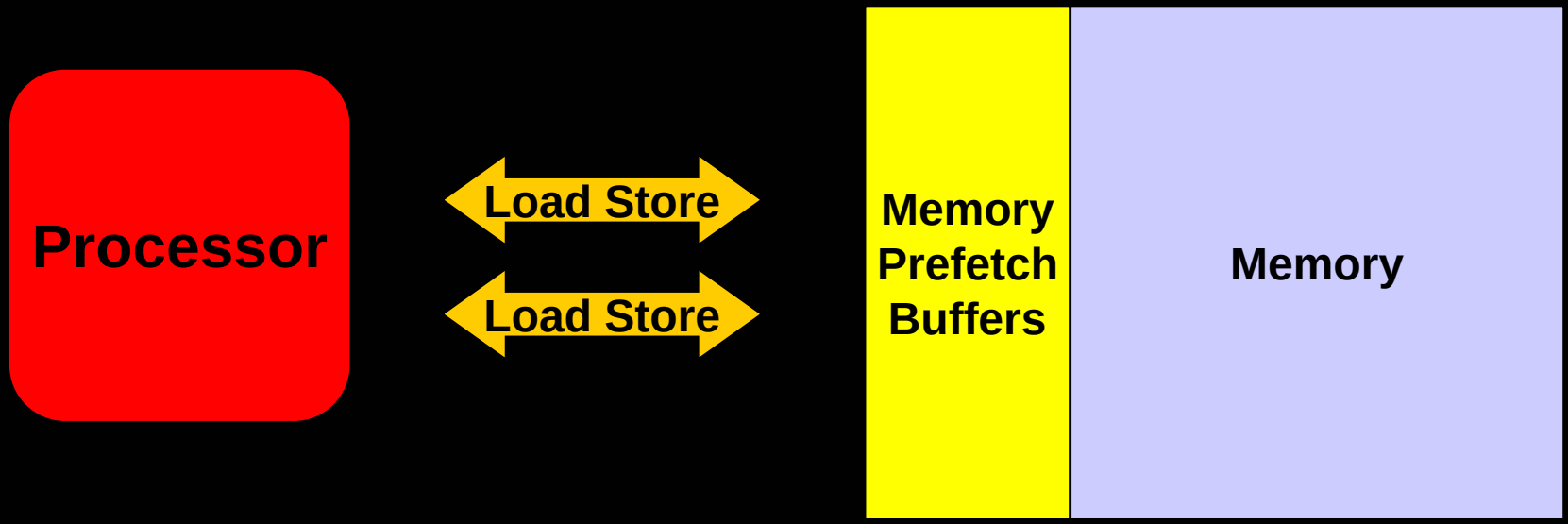
- TLB spans a small amount of memory
- Strategy:
  - Avoid large strides
  - Avoid randomly using large constructs
  - Gather and scatter are very bad
- Common problem on RISC microprocessors

Hardware Prefetch

- Detects adjacent cache line references
  - Forward and backward strides
  - Prefetches up to two lines ahead per stream
- Up to eight concurrent streams
  - Twelve prefetch filter queues
  - No prefetch on store misses
    - (when a store instruction causes a cache line miss)
- Ramped Initialization
  - L2 to L1 prefetches
  - L3 to L2 prefetches
  - Memory to L3 prefetches

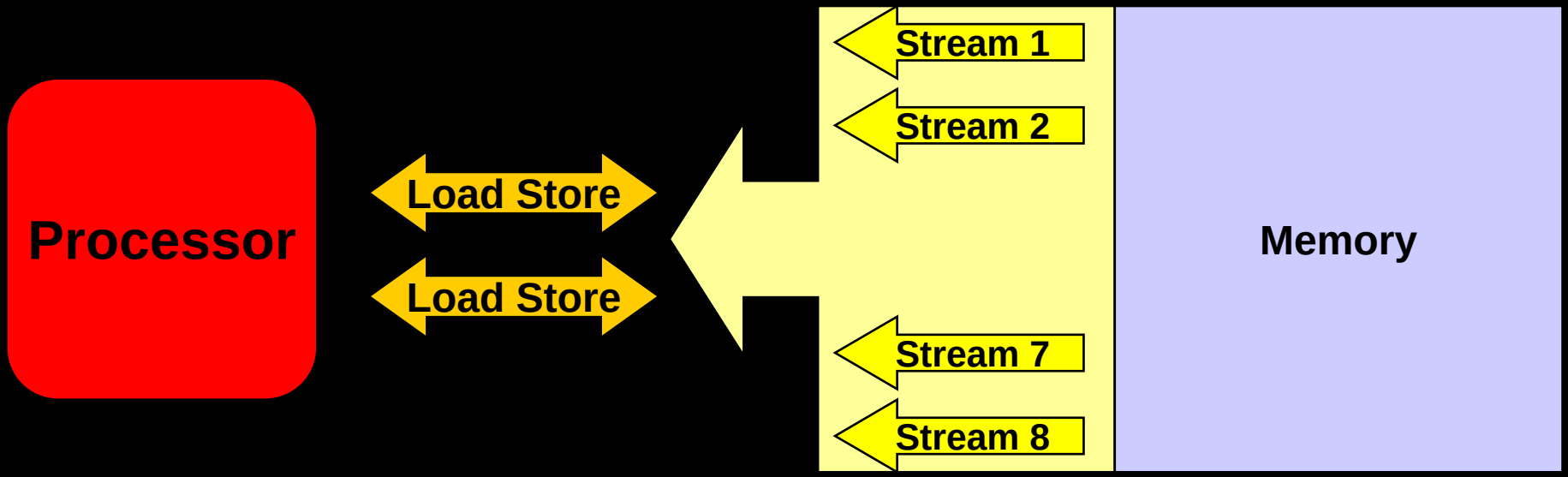
Memory Access

- Load and Store
- Two per CPU
- Connect CPU to memory



Memory Prefetching

- Eight prefetch stream buffers
- Connect CPU to memory



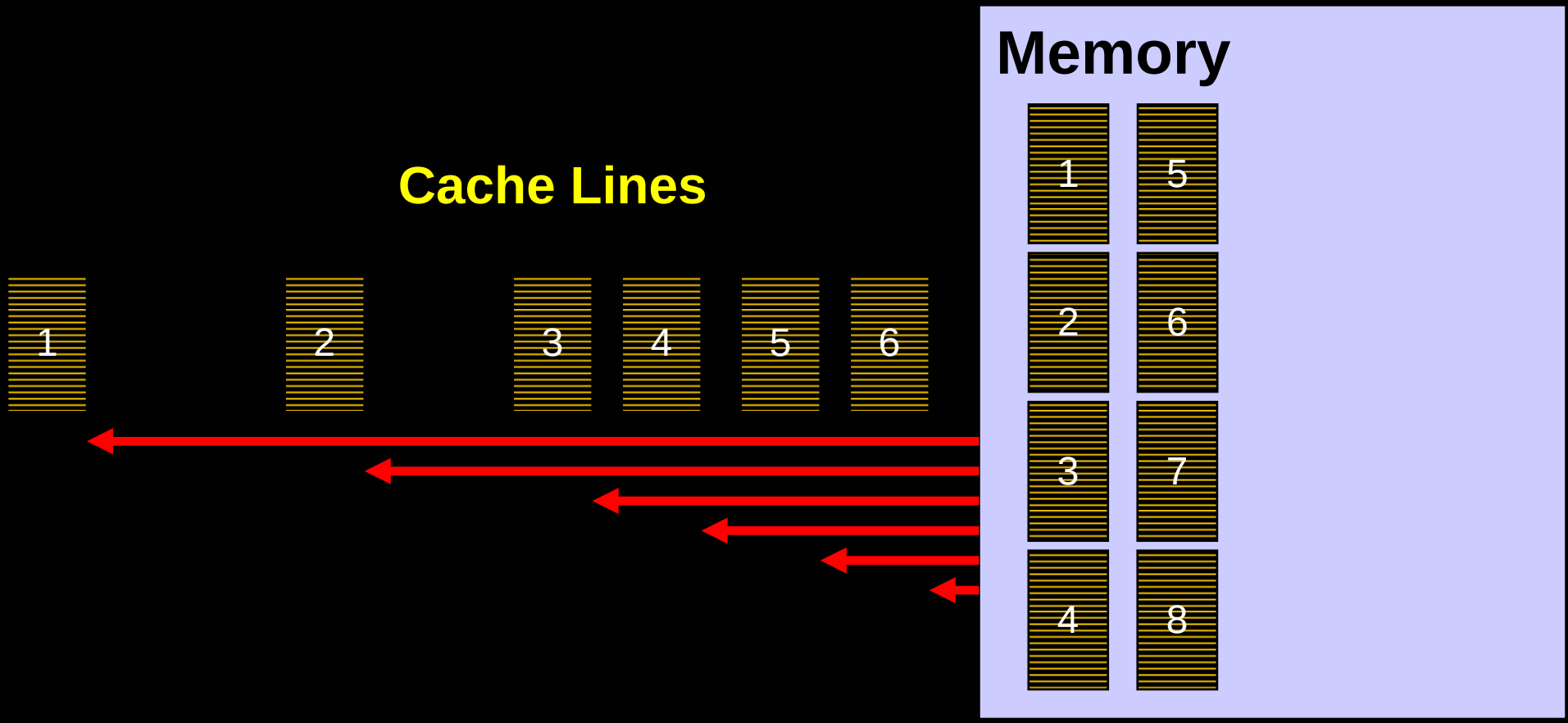
Cache Line Load

- Memory system does not detect patterns within cache line
- Detect location within first 3/4 or last 1/4 of cache line

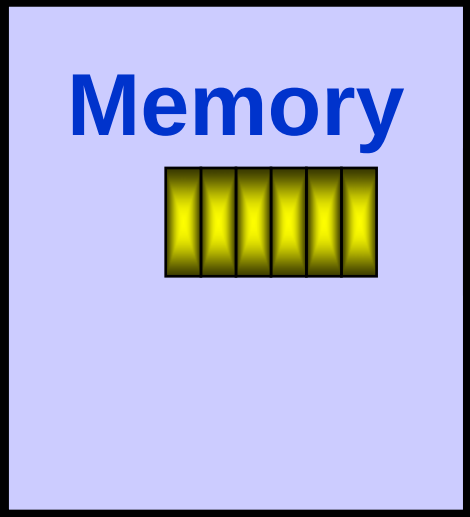
Stride Pattern Recognition

- Upon a cache miss:
  - Biased guess is made as to the direction of that stream
  - Guess is based upon where in the cache line the address associated with that miss occurred
  - If it is in the first 3/4, then the direction is guessed as ascending
  - If in the last 1/4, the direction is guessed descending

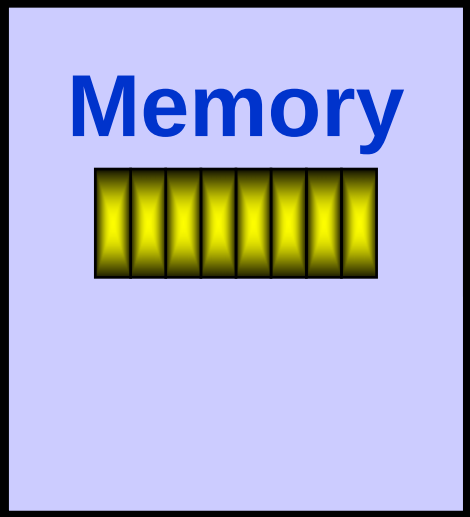
Prefetching



Streams

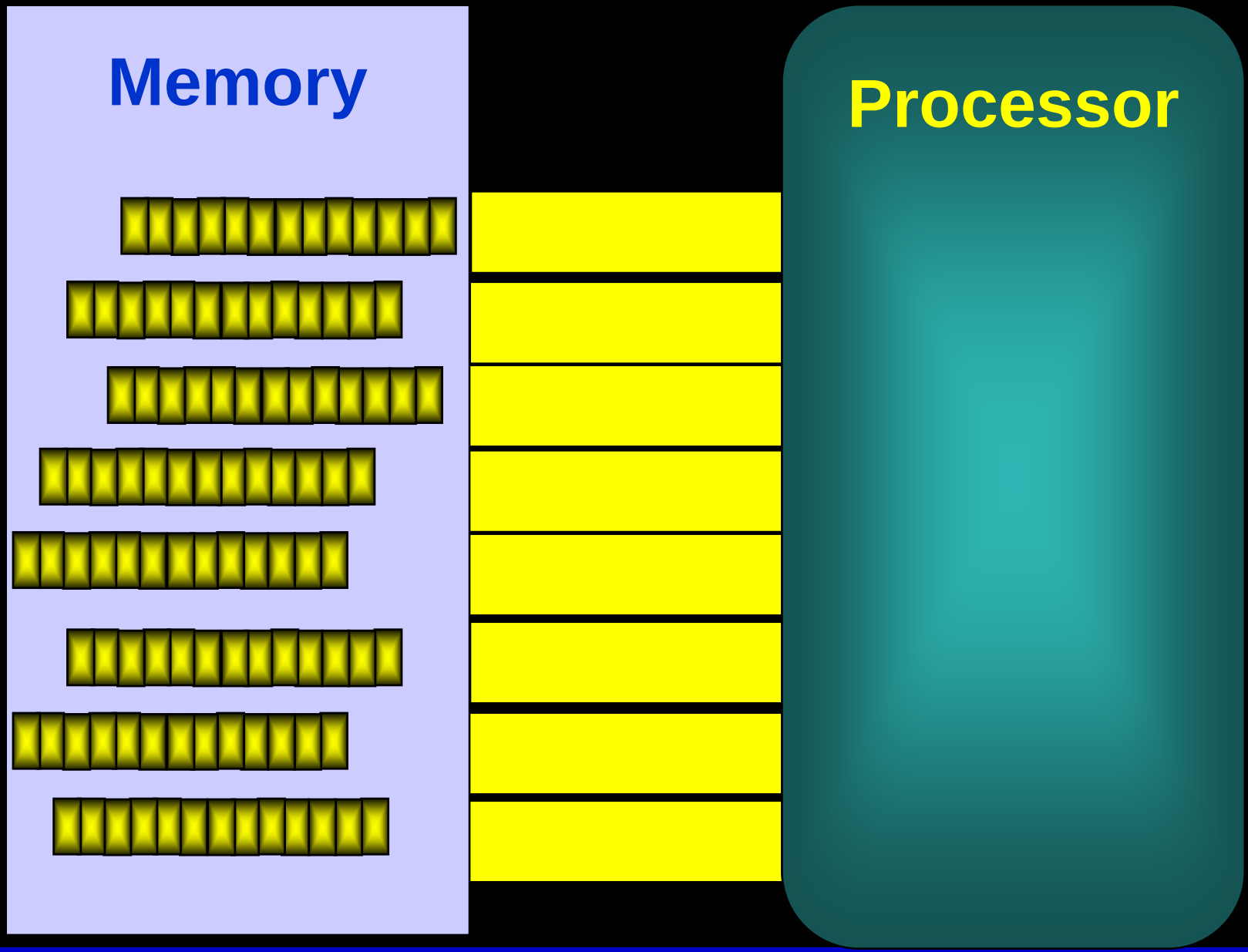


Non-Streaming



Streaming

Streams

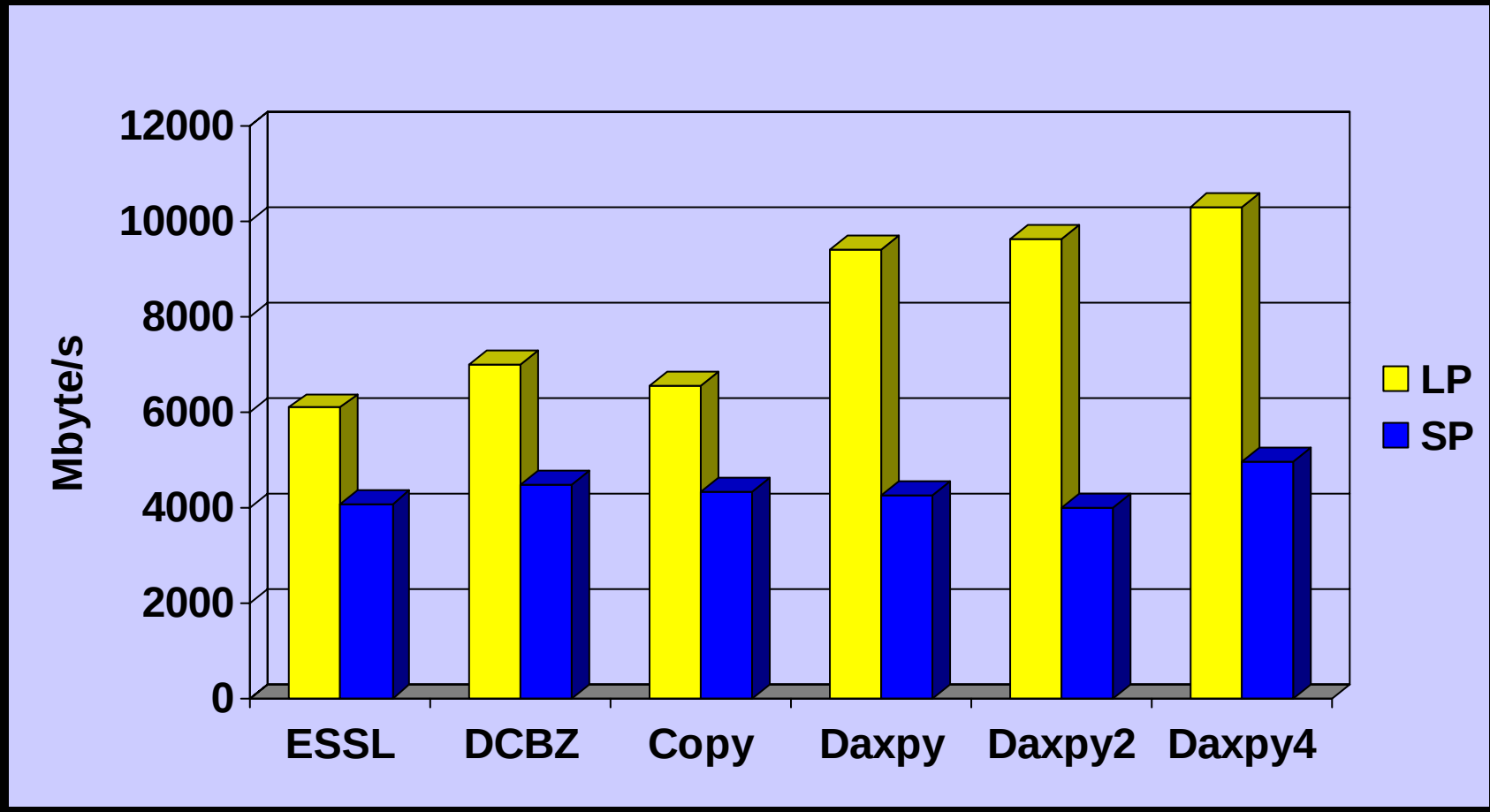


Effect of Prefetch Buffers

- Memory load overlap
- Up to 8 streams
- Variables
- Patterns

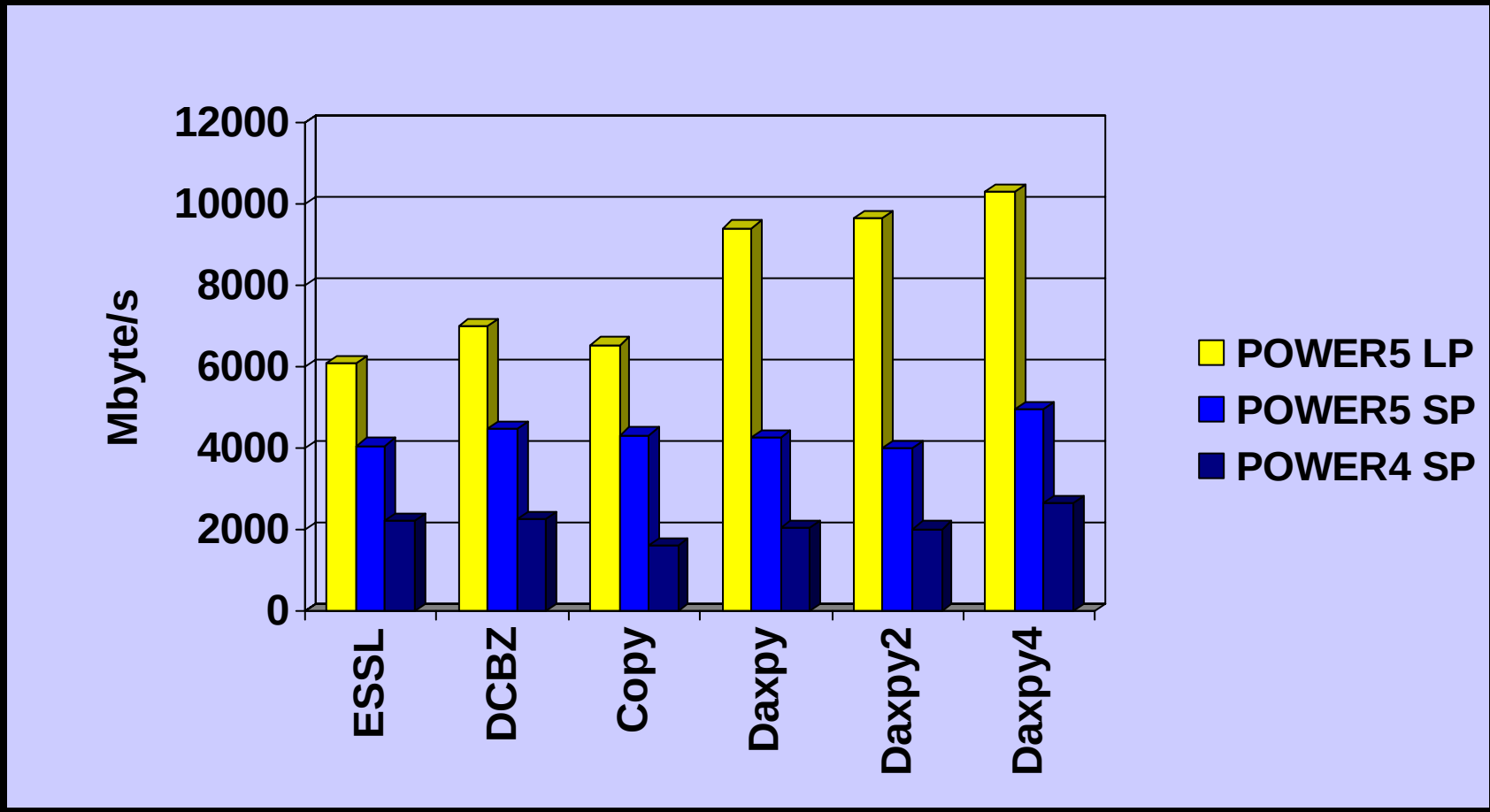
```
for (j=0;j<m;j++)
  for (i=0;i<n;i++)
    A[j][i] = A[j][i]
            +s0*B[j+0][i]
            +s1*B[j+1][i]
            +s2*B[j+2][i]
            +s3*B[j+3][i]
            +C[i]
```

Memory Bandwidth



p5-595 1.9 GHz

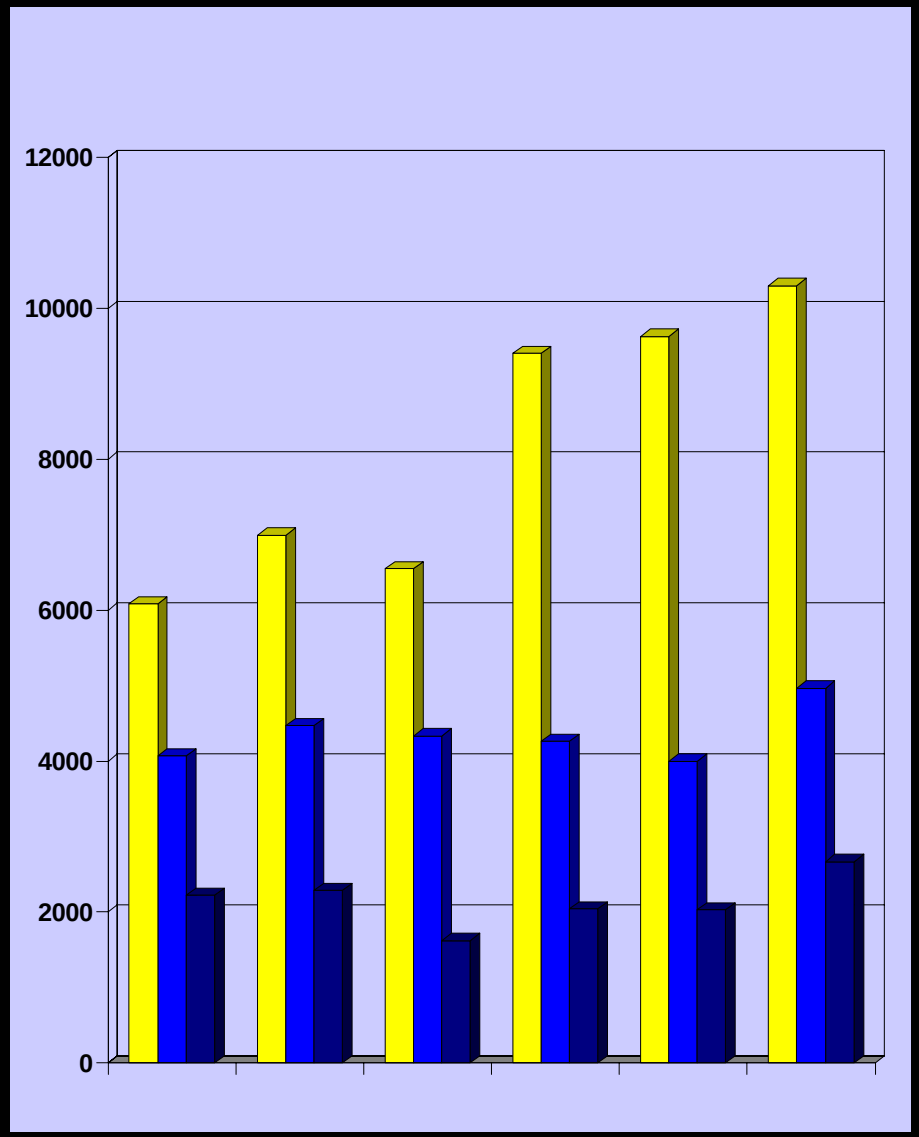
Memory Bandwidth



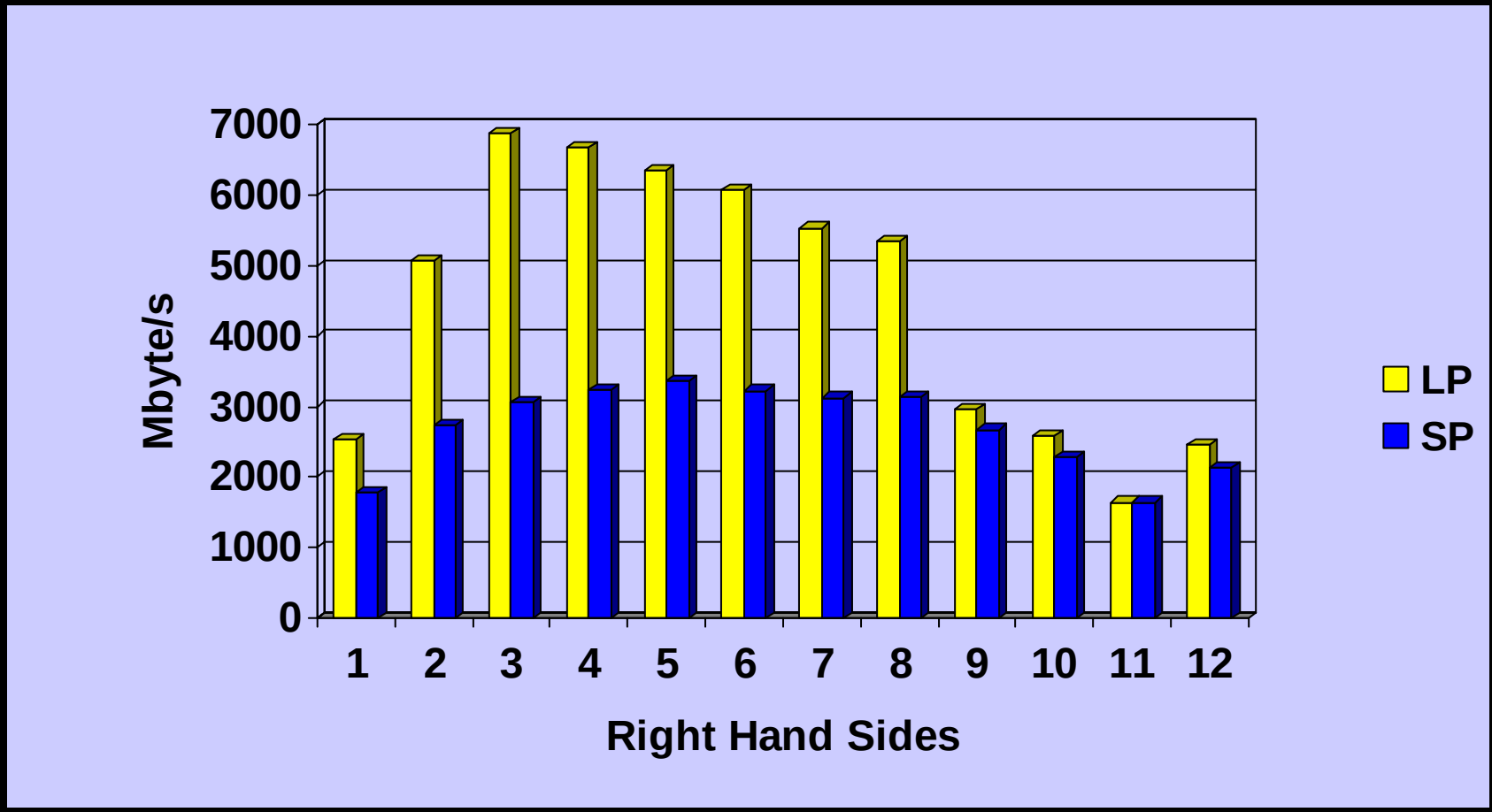
p5-595 1.9 GHz
p690 1.3 GHz

Memory Bandwidth

- Typical POWER5 bandwidth:
  - 4 Gbyte/s Small Pages (SP)
  - 8 Gbyte/s Large pages (LP)
  - Twice the bandwidth of POWER4



Memory Bandwidth: Stream Buffers



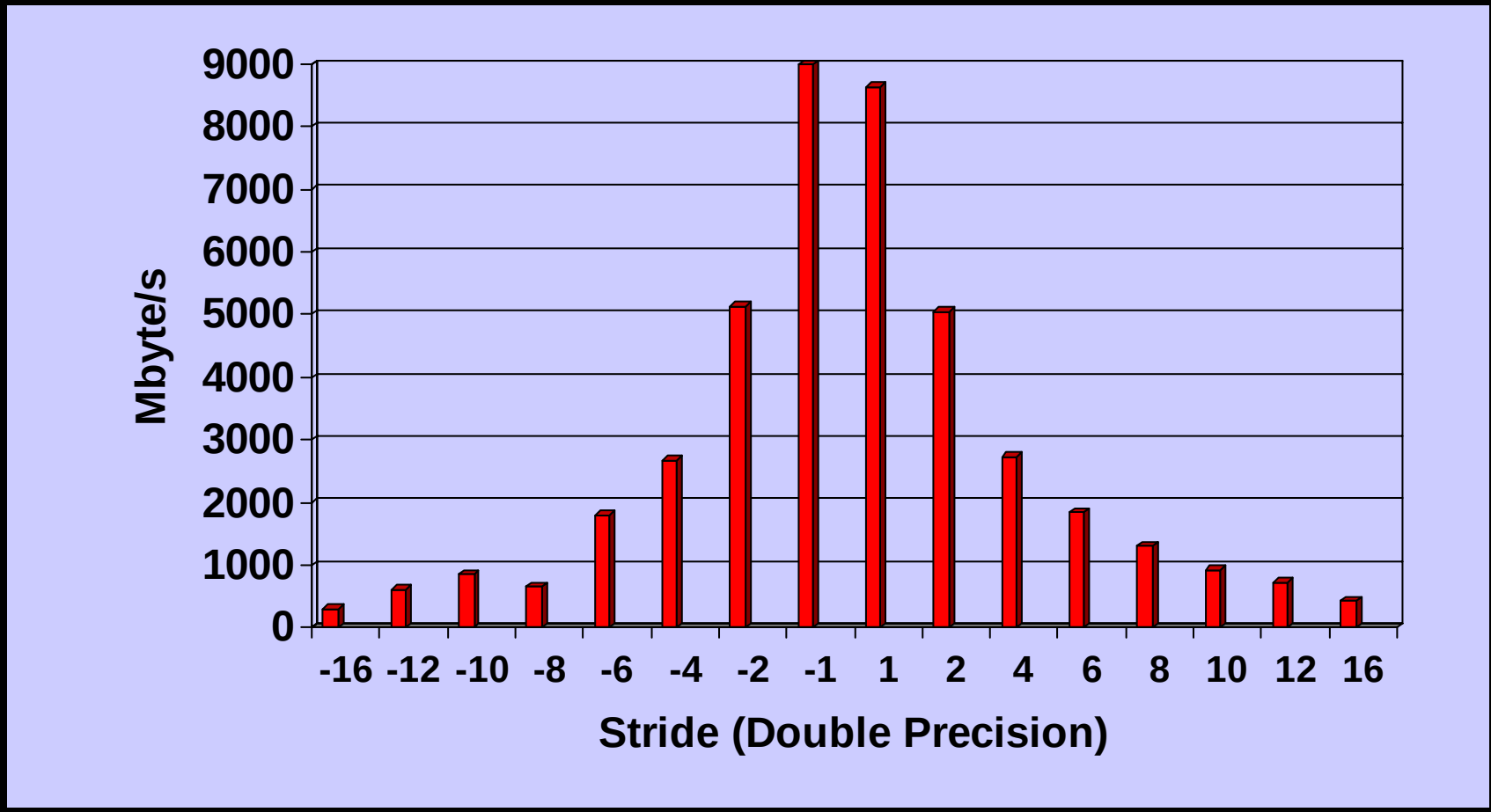
p5-595 1.9 GHz

Strides

- Cache line size is 128 bytes
  - Double precision: 16 words
  - Single precision: 32 words

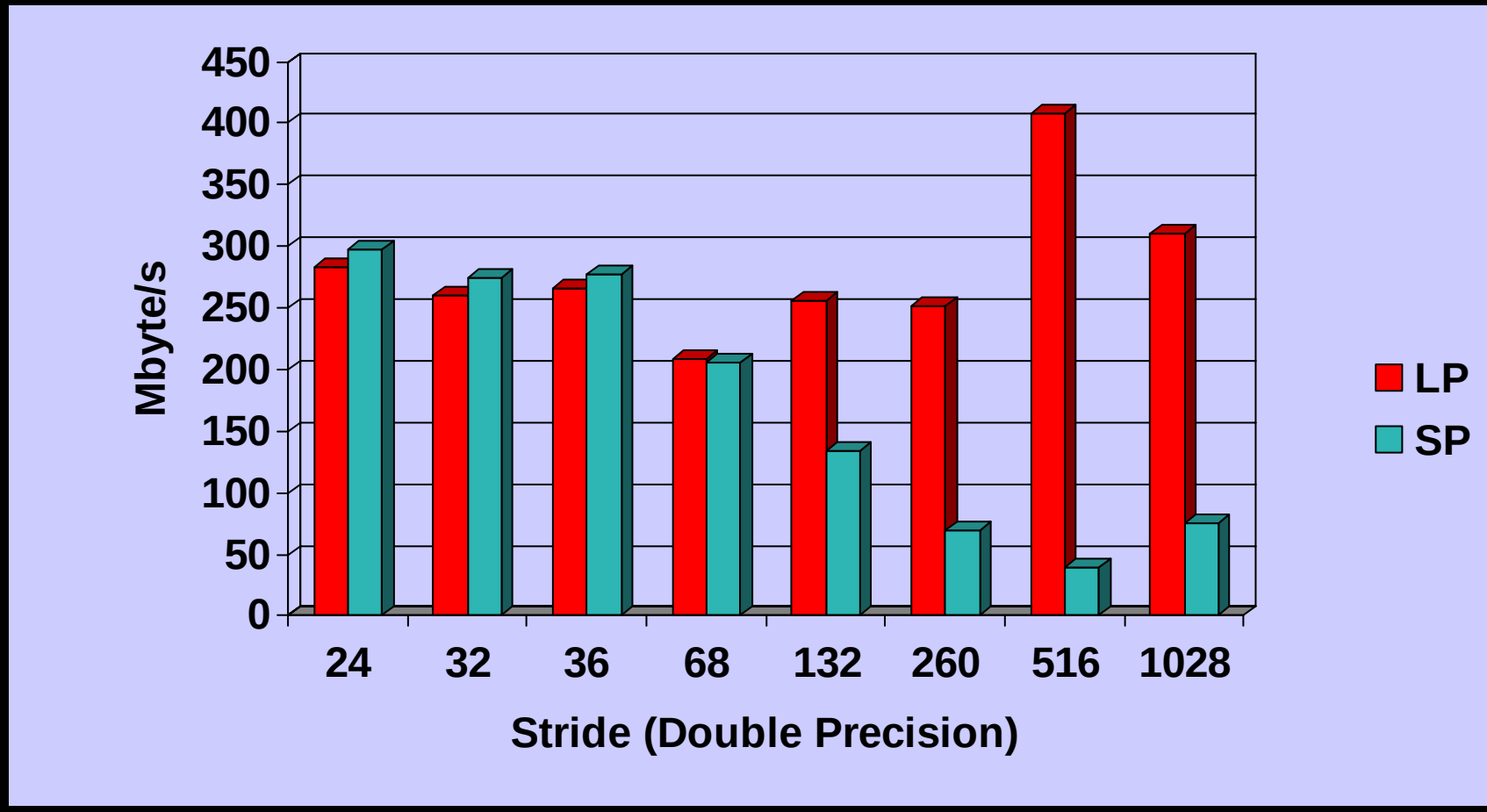
	Bandwidth Reduction	
Stride	Single	Double
1	1x	1x
2	½	½
4	¼	¼
8	1/8	1/8
16	1/16	1/16
32	1/32	1/16
64	1/32	1/16

Stride Test: Small Strides



p5-595 1.9 GHz

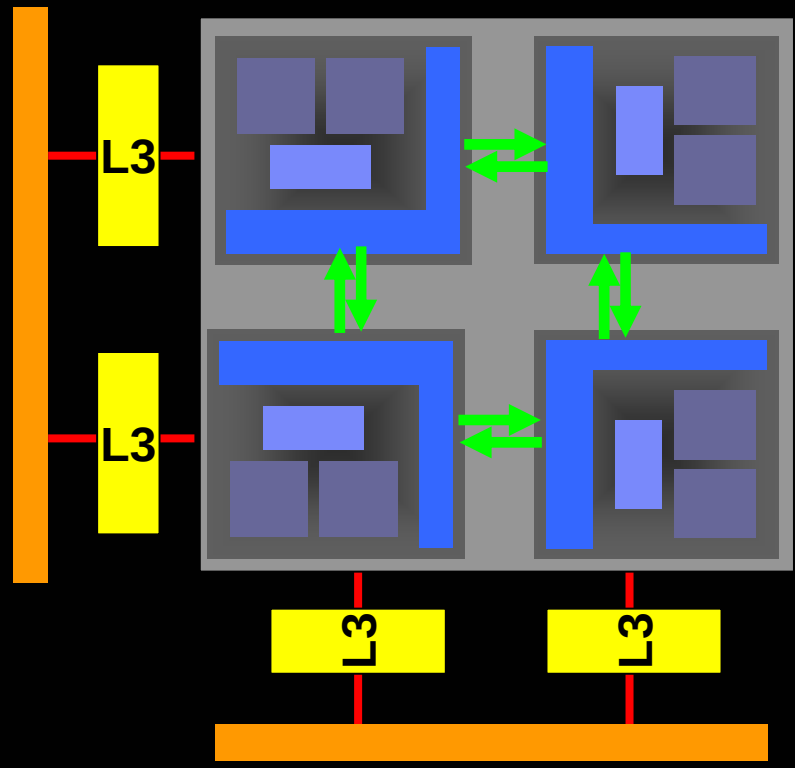
Stride Test: Large Strides



p5-595 1.9 GHz

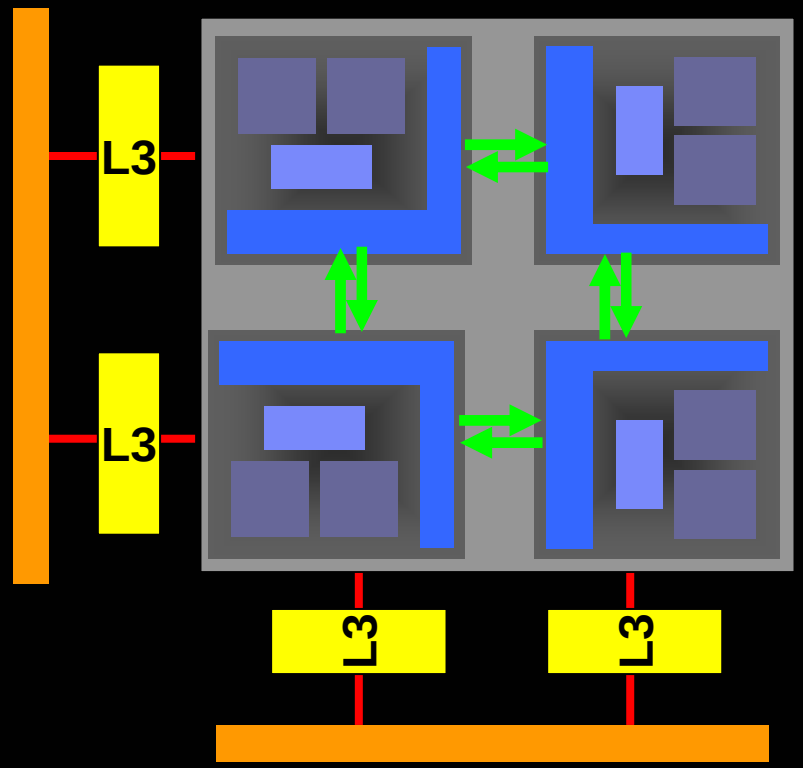
POWER5: Memory

- One or two memory cards per MCM
  - Best bandwidth with two memory cards
- 16 Gbyte/s bandwidth per chip



Interleaved Memory

- Interleaved 4 way within MCM
  - (Only if 2 memory cards match in size)
- Pages interleaved within an MCM
- Consecutive pages can be on any MCM

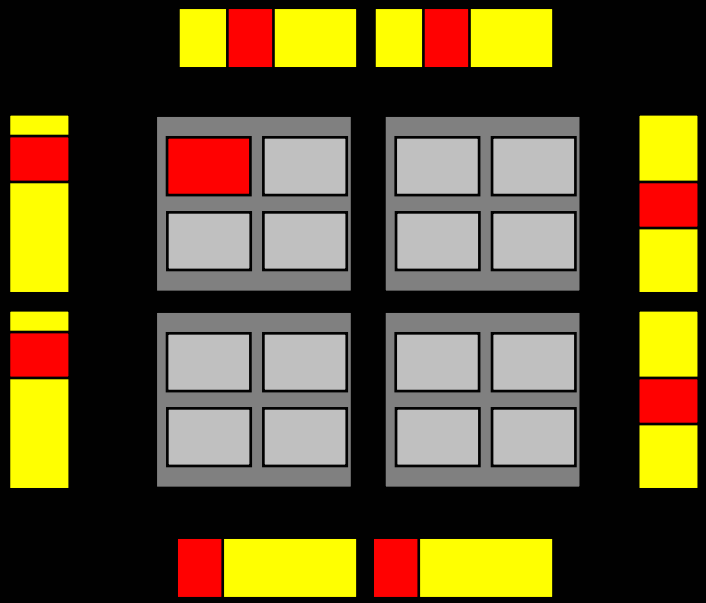


Memory Page Placement

- Default is random page placement
  - Small pages
- Local page placement optional with "first touch" policy
- "Round robin" option available with AIX 5.2
- Large pages are also available
  - Loader option or "tag" binary
  - Large pages are statically allocated
  - Placed at allocation time, not first reference

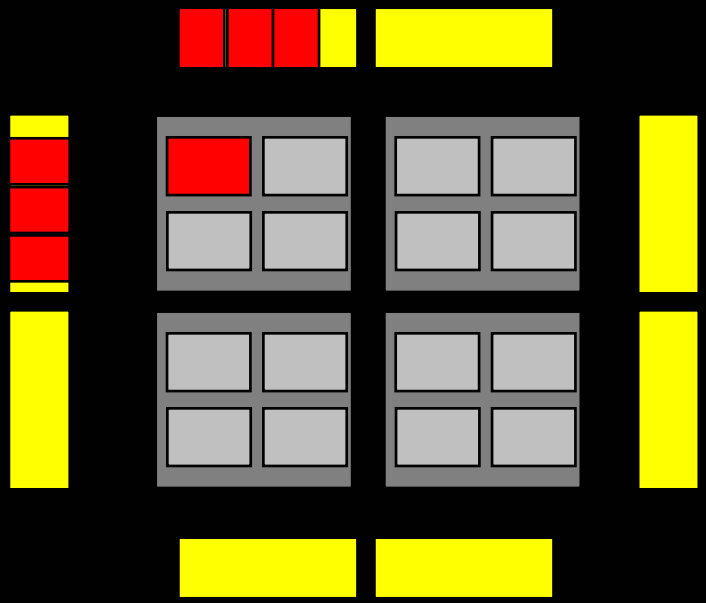
Memory Allocation

- Pages are allocated by module
- Approximately uniform distribution
- Approximately round robin

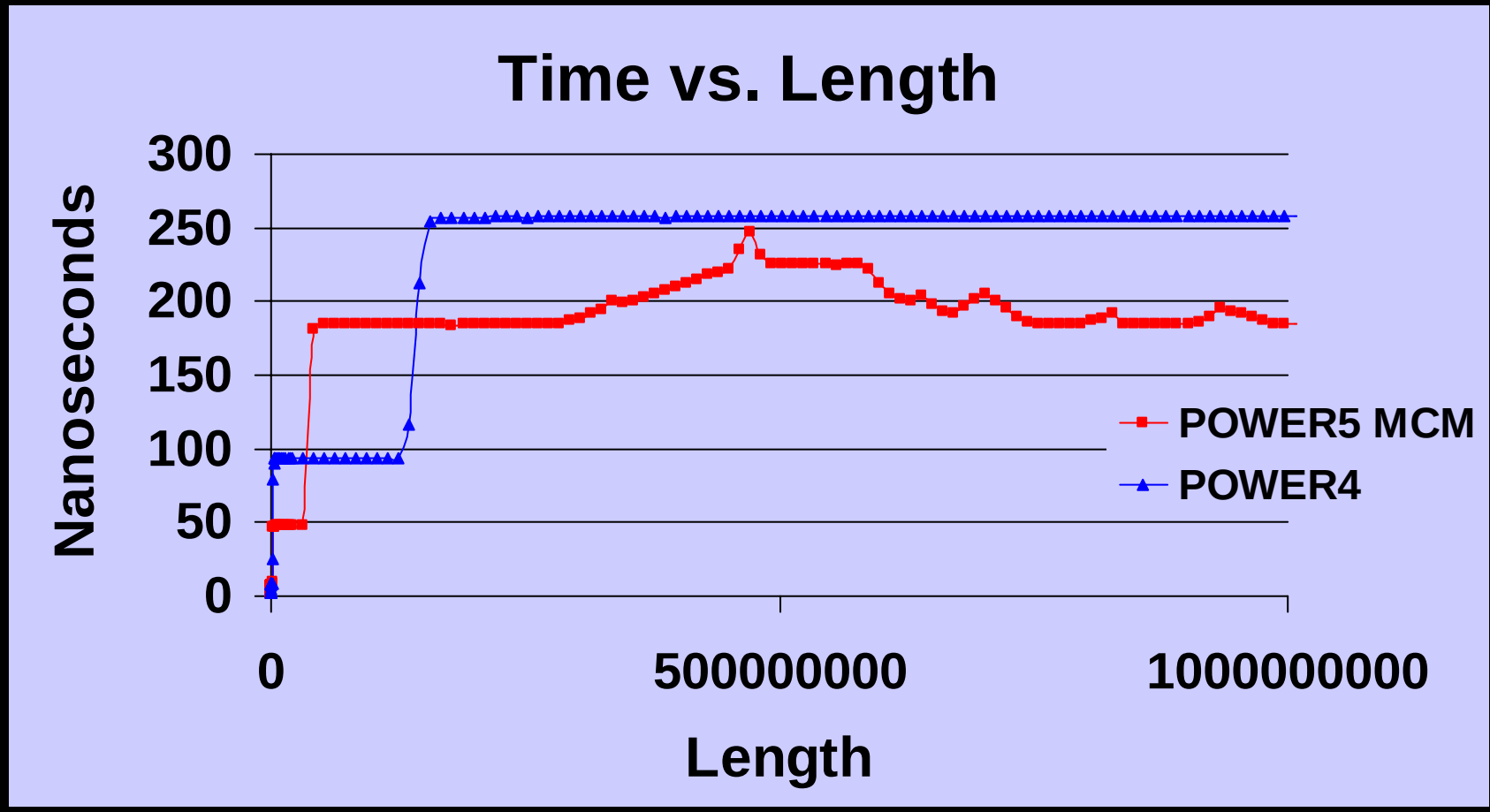


Memory Allocation

- **Memory Affinity**
  - **Allocate pages on memory local to module**

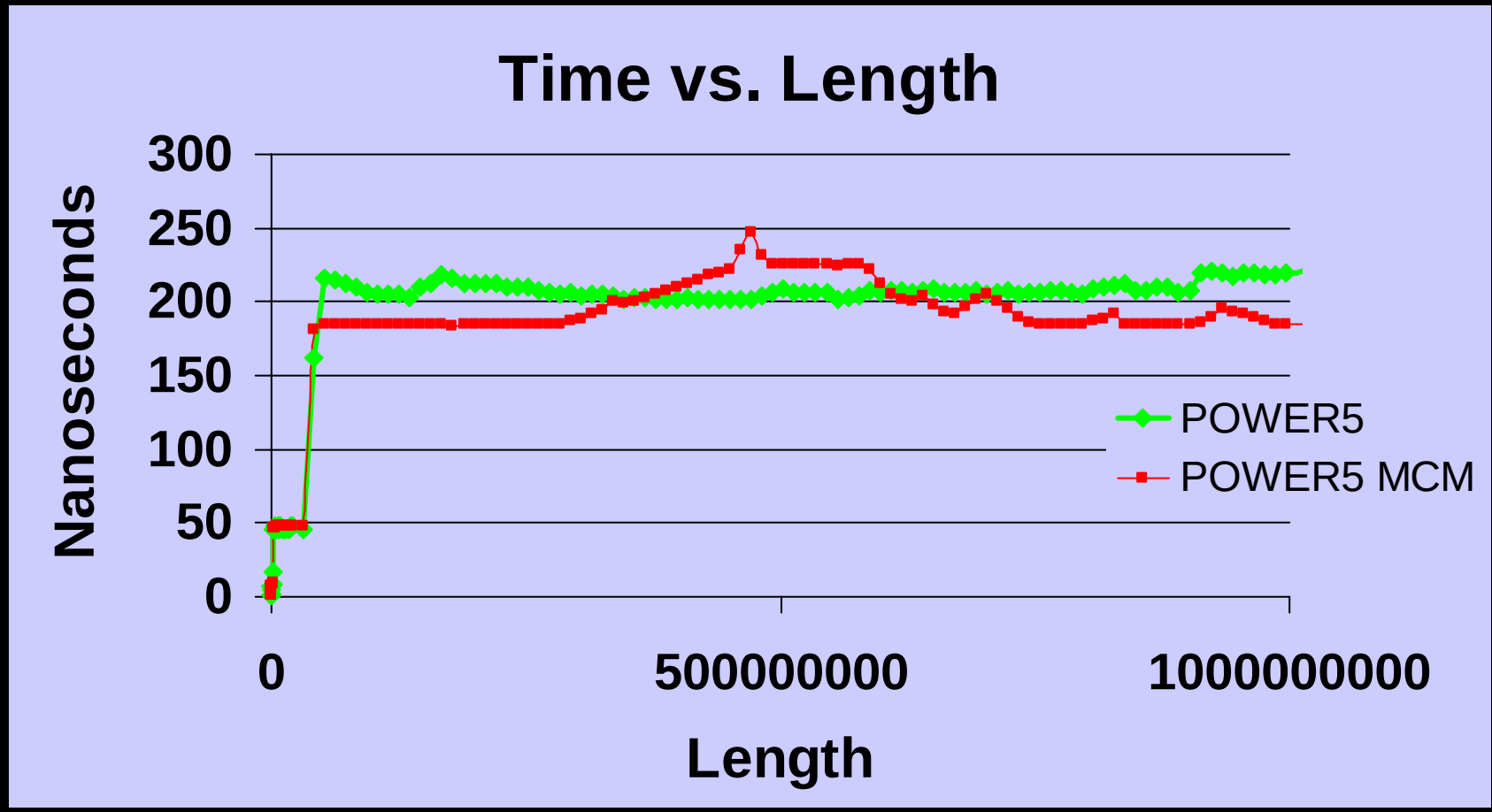


Memory Latencies: POWER4 and POWER5



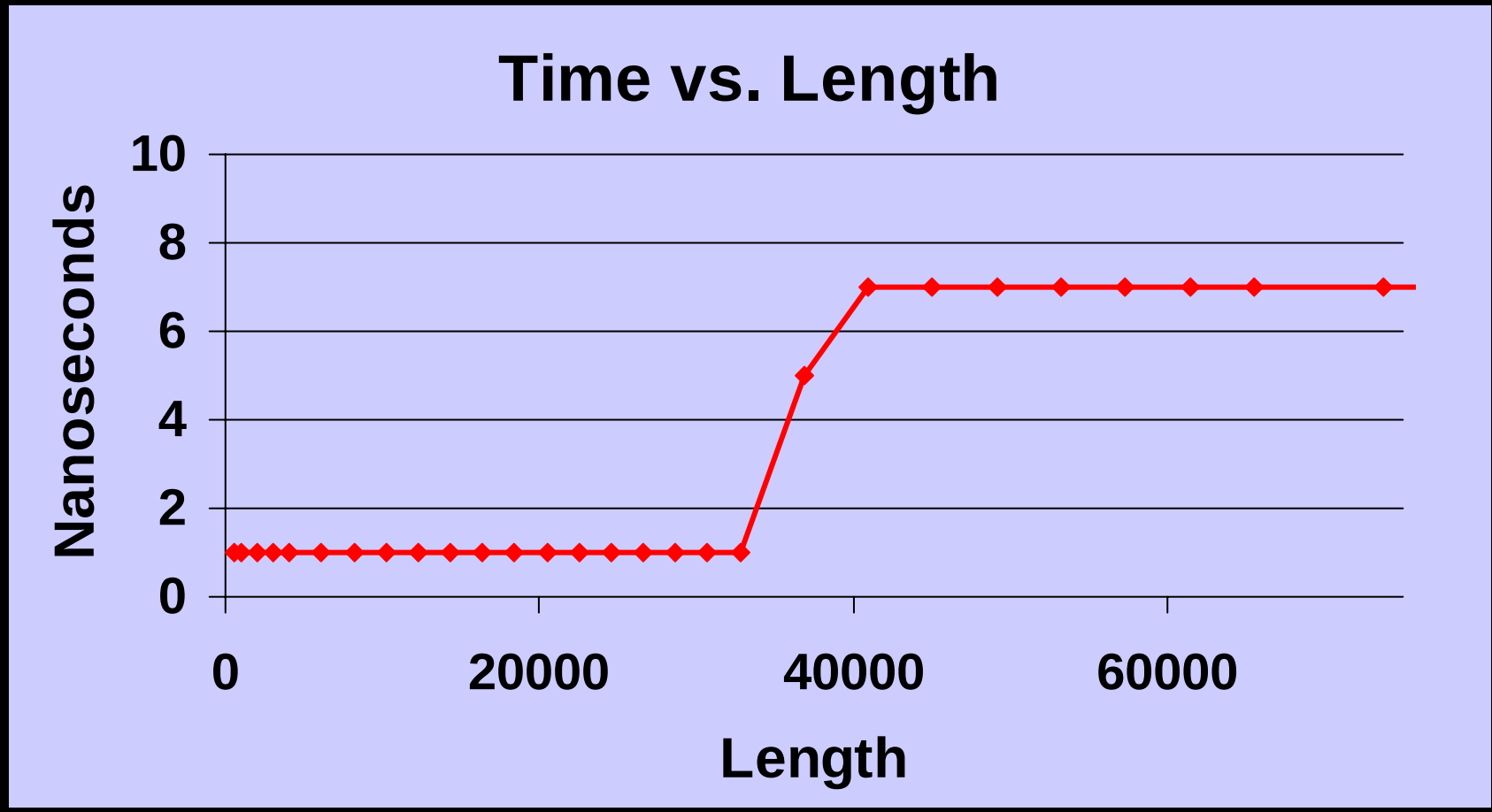
p655 1.5 GHz
p5-595 1.9 GHz

Memory Latencies: POWER5 Memory Affinity



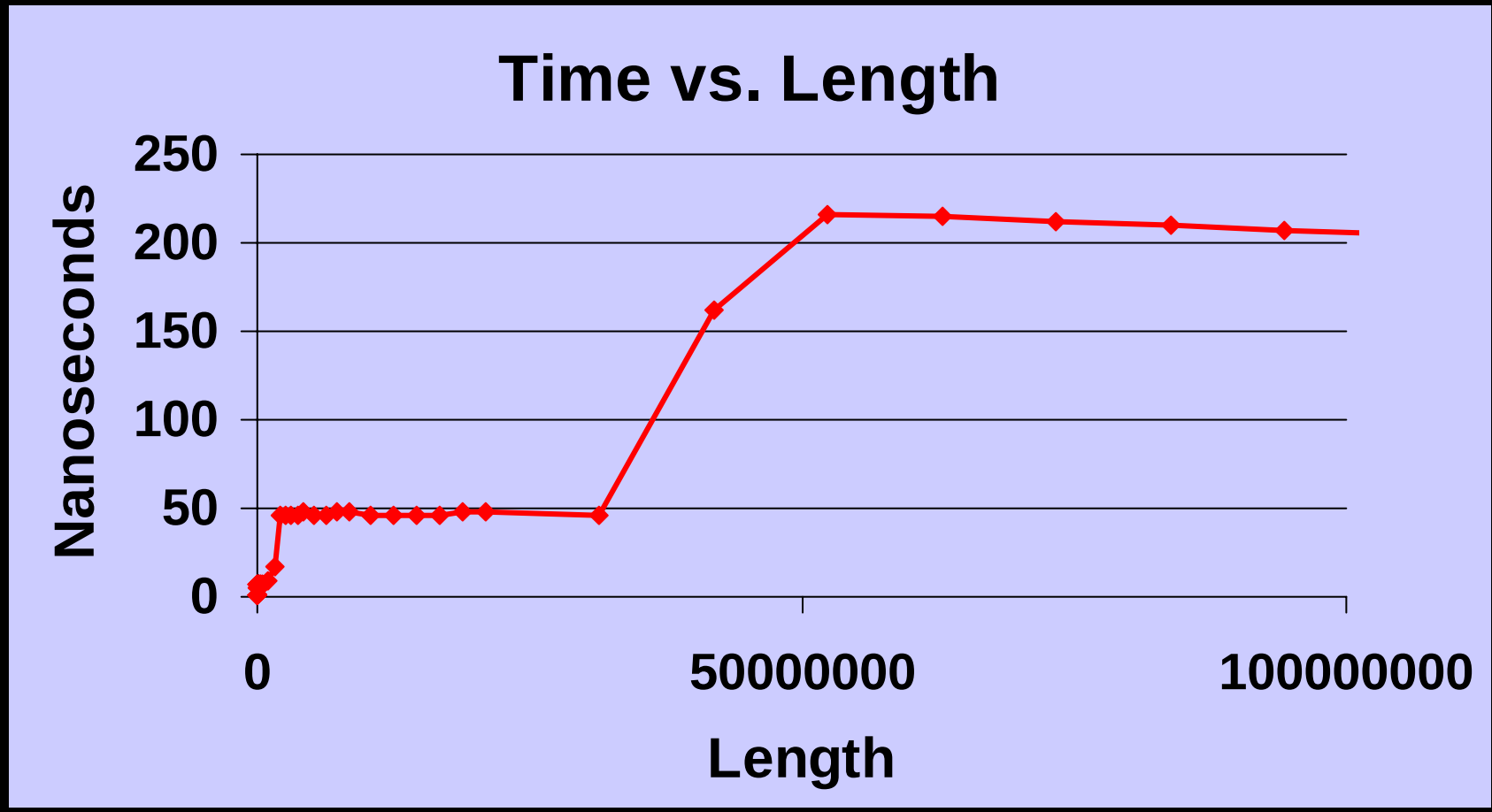
p655 1.5 GHz
p5-595 1.9 GHz

Memory Latencies: L1 Cache – L2 Cache



p5-595 1.9 GHz

Memory Latencies: L3 cache - Memory



p5-595 1.9 GHz

Memory Latencies

Region	Size (byte)	Time (nanosec.)	Clocks (1.9 GHz)
L1	32 kbyte	1	2
L2	36 Mbyte	9	32
L3	36 Mbyte	48	92
Memory	-	210	403

p5-595 1.9 GHz
Program: Lmbench

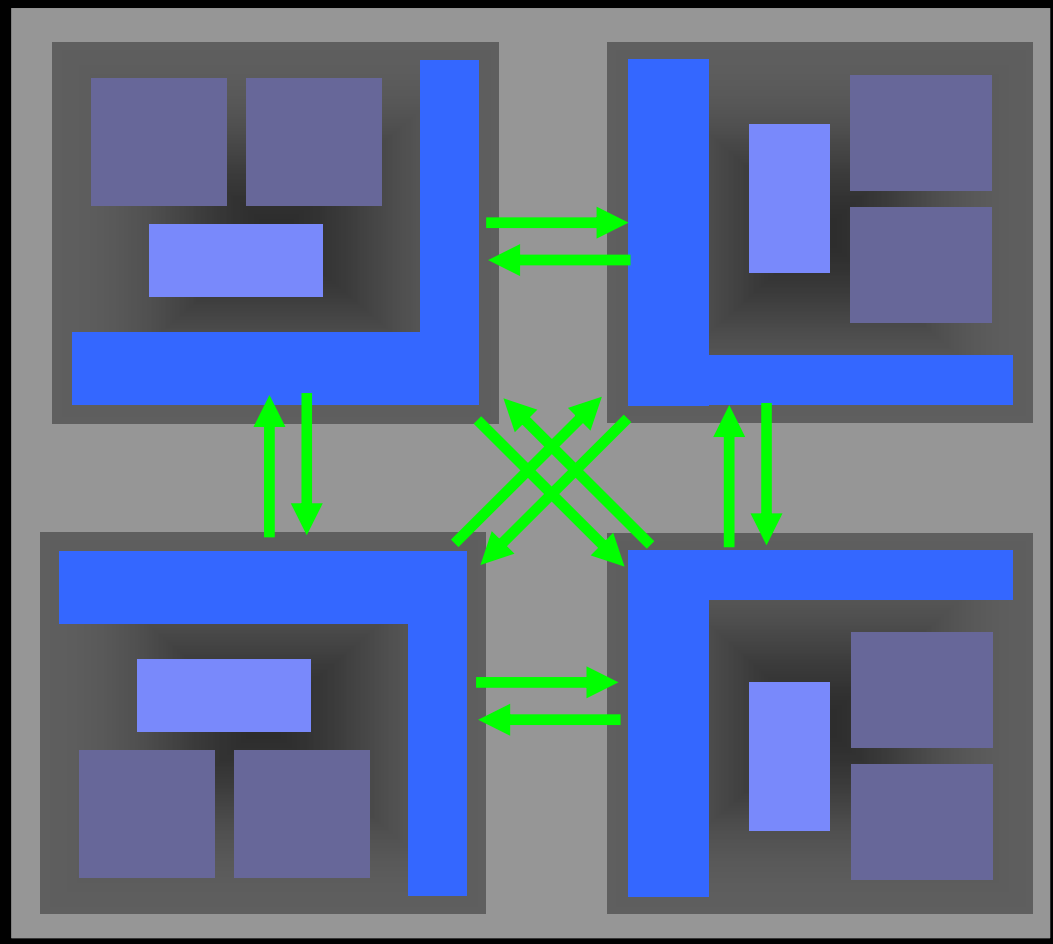
Memory Latency

Level	POWER4	POWER4+	POWER5
L1	2	2	2
L2	9	7	9
L3	95	75	48
Memory	295	255	210

Results are in nanoseconds

Chip to Chip Communications

- On chip
  - 2:1 bus frequency
- Bandwidth
- 35 Gbyte/s



Memory Performance

- **Bandwidth**
  - **3 – 8 Gbyte/s per single processor**
  - **200 Gbyte/s per p5-595**
- **Latency**
  - **~200 nanoseconds**

Bandwidth Considerations

- Affect bandwidth:
  - Right hand side streams
  - Use of 8 stream buffers
    - Overlap cache line loads
  - Page size
    - Small or large memory pages
      - 4 kbyte or 16 Mbyte

Summary

- POWER5
  - 8 Functional units
- Chip:
  - 2 processors
  - Shared L2 cache
- Module
  - Four chips
  - 8 processors
- P5-595 system
  - 8 Modules

Summary

- Registers:
  - 32 architecture
  - 88 rename
- Functional Units
  - 6 clock periods for FMA
    - Pipelined
  - 32 clock periods for FDIV
    - NOT pipelined
- 1024 entry TLB
- 4096 byte page
- 8 prefetch streams