IBM's POWER5 Microprocessor Design and Methodology

Ron Kalla
IBM Systems Group

Outline

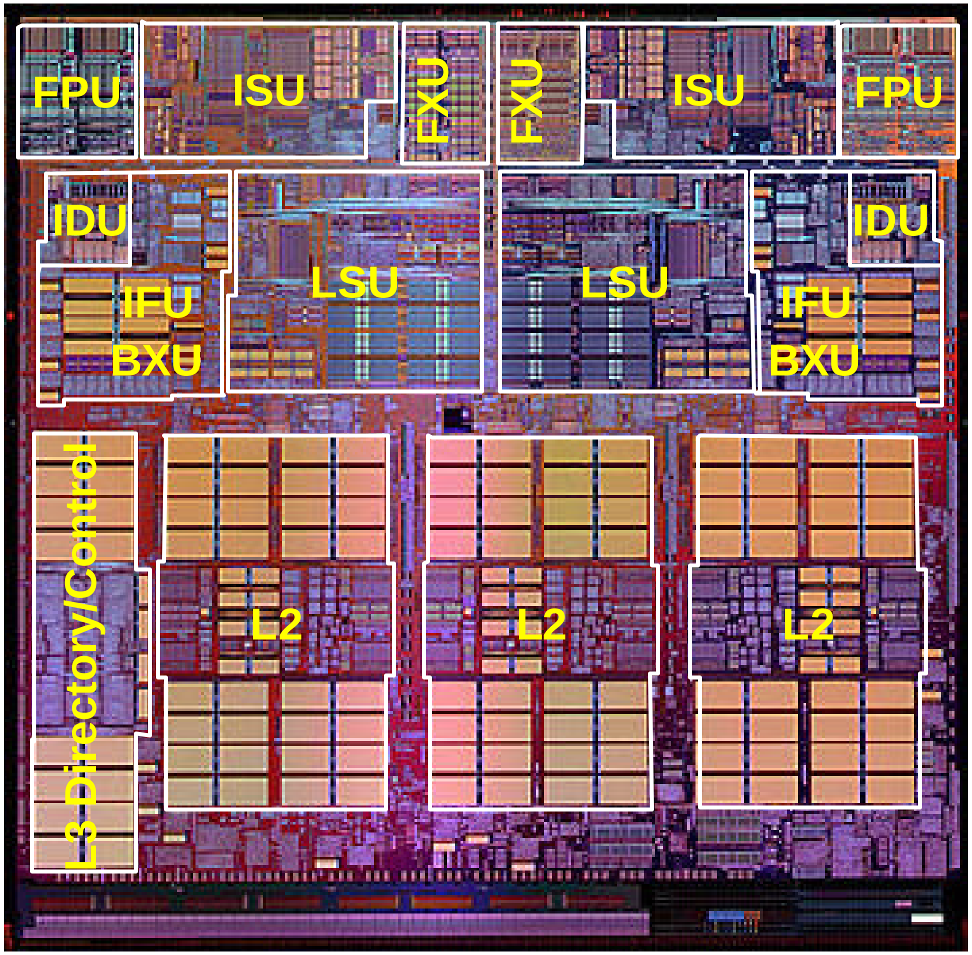
- Motivation
- Background
- Threading Fundamentals
- Enhanced SMT Implementation in POWER5
- Memory Subsystem Enhancements
- Additional SMT Considerations
- Summary

Microprocessor Design Optimization Focus Areas

- Memory latency
  - Increased processor speeds make memory appear further away
  - Longer stalls possible
- Branch processing
  - Mispredict more costly as pipeline depth increases resulting in stalls and wasted power
  - Predication drives increased power and larger chip area
- Execution Unit Utilization
  - Currently 20-25% execution unit utilization common
- Simultaneous multi-threading (SMT) and POWER architecture address these areas

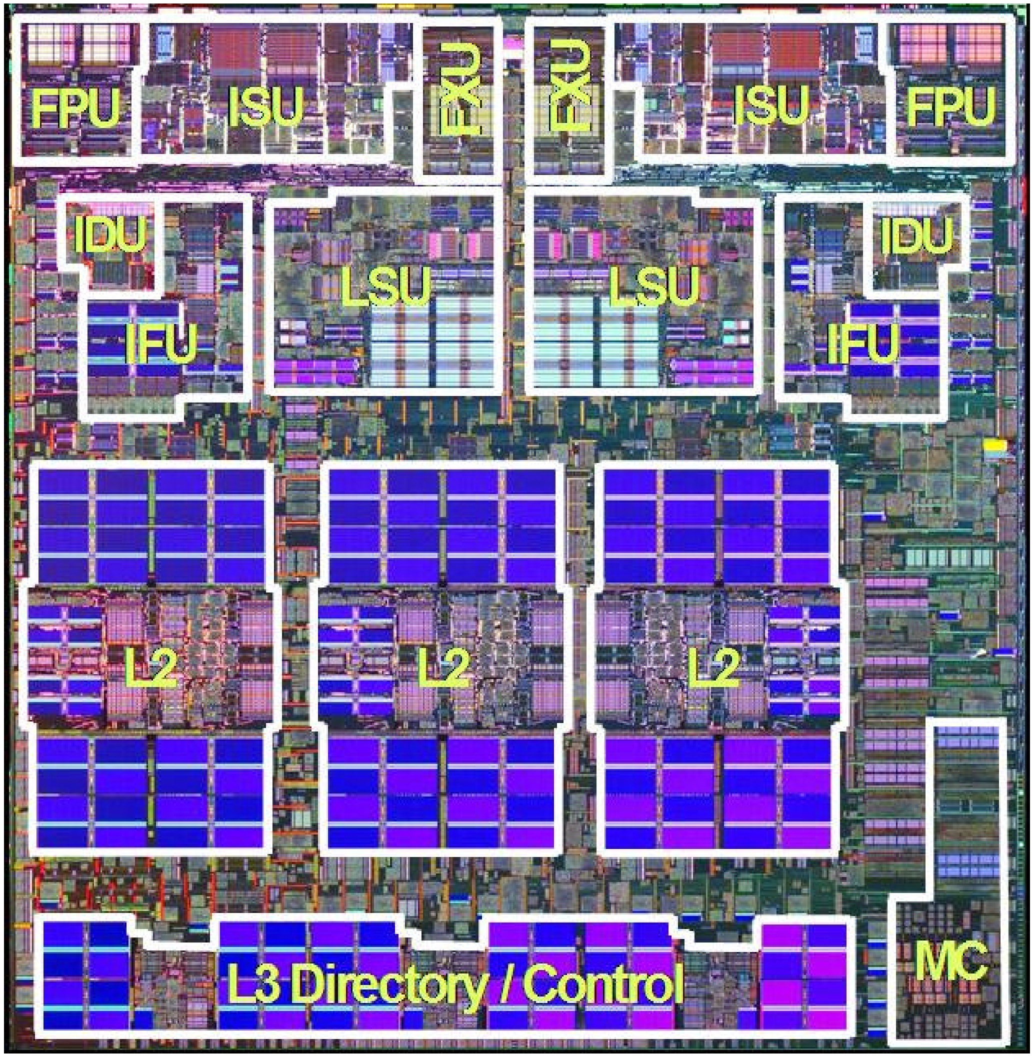
POWER4 --- Shipped in Systems December 2001

- Technology: 180nm lithography, Cu, SOI
  - POWER4+ shipping in 130nm today
  - 267mm^2 185M transistors
- 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 --- The Next Step

- Technology: 130nm lithography, Cu, SOI
- 389mm^2 276M Transistors
- Dual processor core
- 8-way superscalar
- Simultaneous multithreaded (SMT) core
  - Up to 2 virtual processors per real processor
  - Natural extension to POWER4 design



Multi-threading Evolution

Single Thread

FX0	🟦	⬜	⬜	🟦	⬜	⬜	⬜	⬜
FX1	⬜	🟦	⬜	⬜	⬜	⬜	🟦	⬜
FP0	⬜	⬜	⬜	⬜	⬜	⬜	🟦	⬜
FP1	⬜	⬜	🟦	⬜	🟦	⬜	⬜	⬜
LS0	🟦	⬜	⬜	⬜	⬜	⬜	🟦	🟦
LS1	⬜	⬜	⬜	🟦	⬜	⬜	⬜	⬜
BRX	⬜	🟦	⬜	⬜	⬜	🟦	⬜	🟦
CRL	⬜	⬜	⬜	🟦	⬜	⬜	⬜	⬜

Coarse Grain Threading

FX0	🟦	⬜	⬜	⬜	⬜	⬜	⬜	⬜
FX1	⬜	🟦	⬜	⬜	⬜	⬜	🟥	⬜
FP0	⬜	⬜	⬜	⬜	⬜	⬜	🟥	⬜
FP1	⬜	⬜	⬜	⬜	🟥	⬜	⬜	⬜
LS0	🟦	🟦	⬜	⬜	⬜	⬜	🟥	⬜
LS1	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
BRX	⬜	🟦	⬜	⬜	⬜	🟥	⬜	⬜
CRL	⬜	⬜	⬜	⬜	🟥	⬜	⬜	⬜

Fine Grain Threading

FX0	🟦	⬜	⬜	🟥	⬜	⬜	⬜	⬜
FX1	⬜	🟥	⬜	⬜	⬜	⬜	🟦	🟥
FP0	⬜	⬜	⬜	⬜	⬜	🟥	🟦	⬜
FP1	⬜	⬜	🟦	⬜	🟦	⬜	⬜	⬜
LS0	🟦	⬜	⬜	⬜	⬜	⬜	🟦	🟥
LS1	⬜	⬜	⬜	🟥	🟦	⬜	⬜	⬜
BRX	⬜	🟥	🟦	⬜	⬜	🟥	⬜	🟥
CRL	⬜	⬜	⬜	🟥	⬜	⬜	⬜	⬜

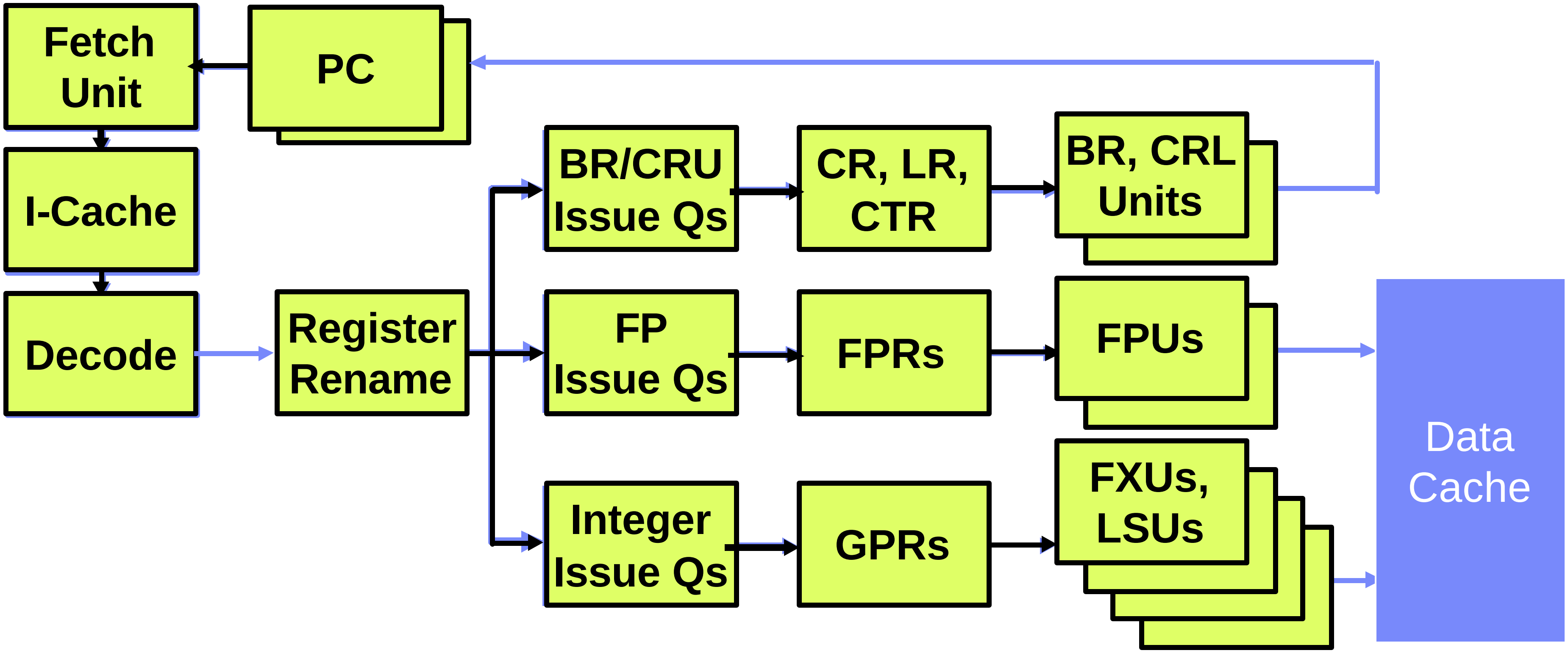
Simultaneous Multi-Threading

FX0	🟦	⬜	⬜	🟥	⬜	⬜	⬜	🟦
FX1	⬜	🟥	⬜	⬜	⬜	⬜	🟥	⬜
FP0	⬜	🟦	⬜	⬜	⬜	🟥	🟦	⬜
FP1	⬜	⬜	🟦	⬜	🟥	⬜	⬜	⬜
LS0	🟦	⬜	⬜	⬜	⬜	🟦	🟦	🟥
LS1	🟥	⬜	⬜	🟥	🟦	⬜	⬜	⬜
BRX	⬜	🟦	🟦	⬜	⬜	🟥	🟥	🟥
CRL	⬜	⬜	⬜	🟥	⬜	⬜	⬜	⬜

🟦 **Thread 0 Executing** 🟥 **Thread 1 Executing** ⬜ **No Thread Executing**

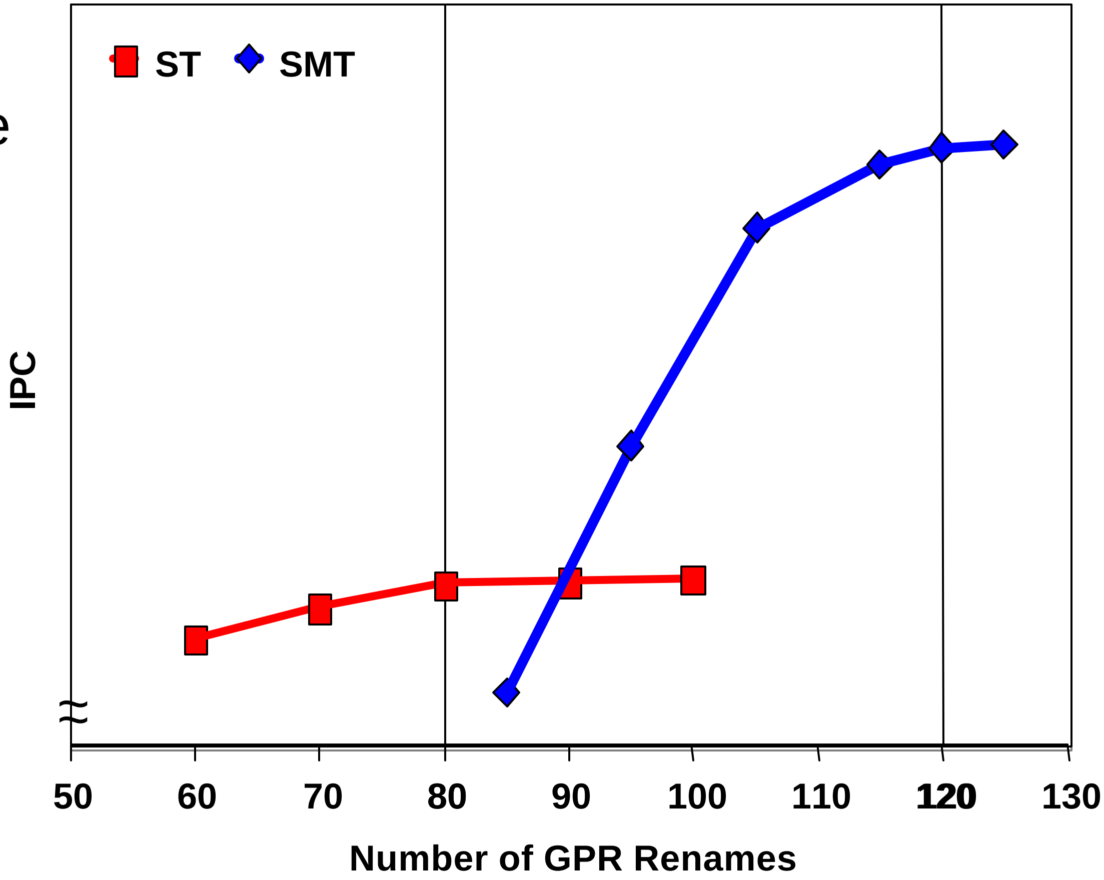
Changes Going From ST to SMT Core

- SMT easily added to Superscalar Micro-architecture
  - Second Program Counter (PC) added to share I-fetch bandwidth
  - GPR/FPR rename mapper expanded to map second set of registers (High order address bit indicates thread)
  - Completion logic replicated to track two threads
  - Thread bit added to most address/tag buses



Resource Sizes

- Analysis done to optimize every micro-architectural resource size
  - GPR/FPR rename pool size
  - I-fetch buffers
  - Reservation Station
  - SLB/TLB/ERAT
  - I-cache/D-cache
- Many Workloads examined
- Associativity also examined



Results based on simulation of an online transaction processing application

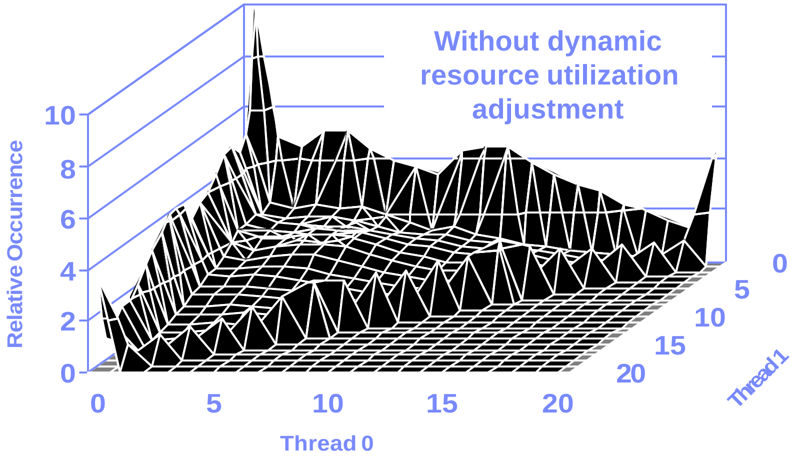
Vertical axis does not originate at 0

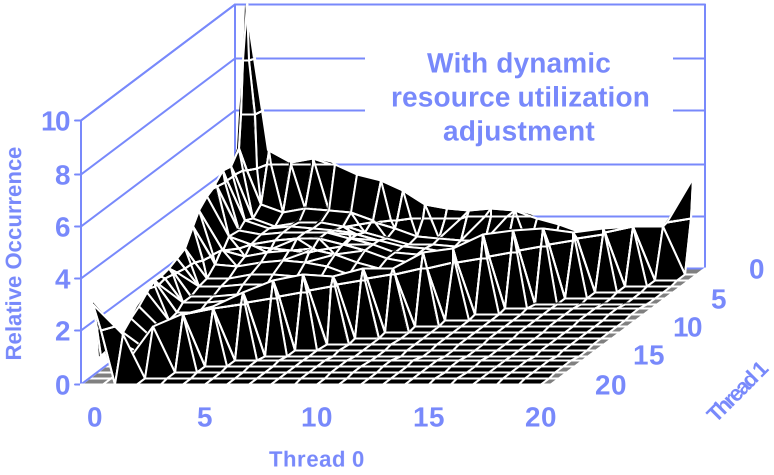
POWER5 Resources Size Enhancements

- Enhanced caches and translation resources
  - I-cache: 64 KB, 2-way set associative, LRU
  - D-cache: 32 KB, 4-way set associative, LRU
  - First level Data Translation: 128 entries, fully associative, LRU
  - L2 Cache: 1.92 MB, 10-way set associative, LRU
- Larger resource pools
  - Rename registers: GPRs, FPRs increased to 120 each
  - L2 cache coherency engines: increased by 100%
- Enhanced data stream prefetching
- Memory controller moved on chip

Resource Sharing

Global Completion Table Occupancy



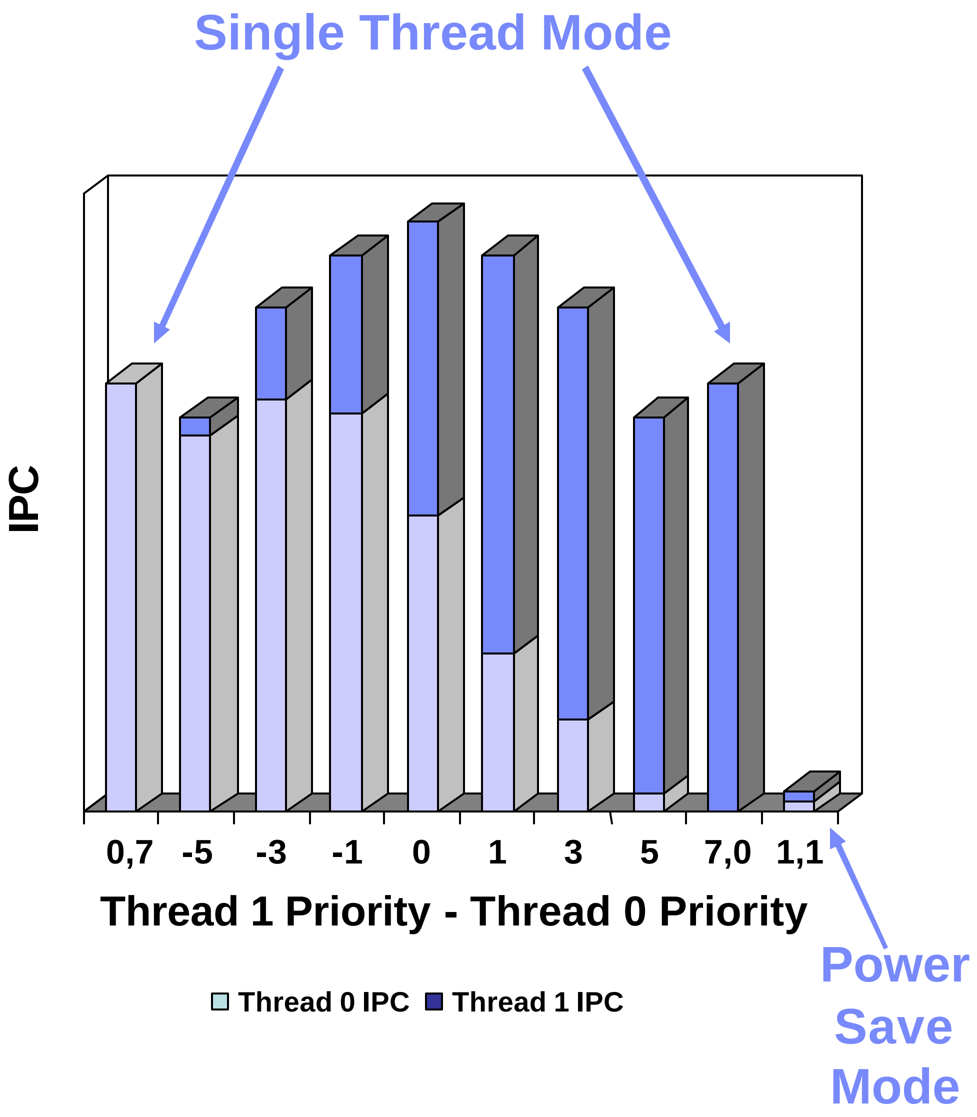


Results based on simulation of an online transaction processing application

- Threads share many resources
  - Global Completion Table, BHT, TLB, . . .
- Higher performance realized when resources balanced across threads
  - Tendency to drift toward extremes accompanied by reduced performance
- Solution: Dynamically adjust resource utilization

Thread Priority

- Instances when unbalanced execution desirable
  - No work for opposite thread
  - Thread waiting on lock
  - Software determined non uniform balance
  - Power management
  - …
- Solution: Control instruction decode rate
  - Software/hardware controls 8 priority levels for each thread



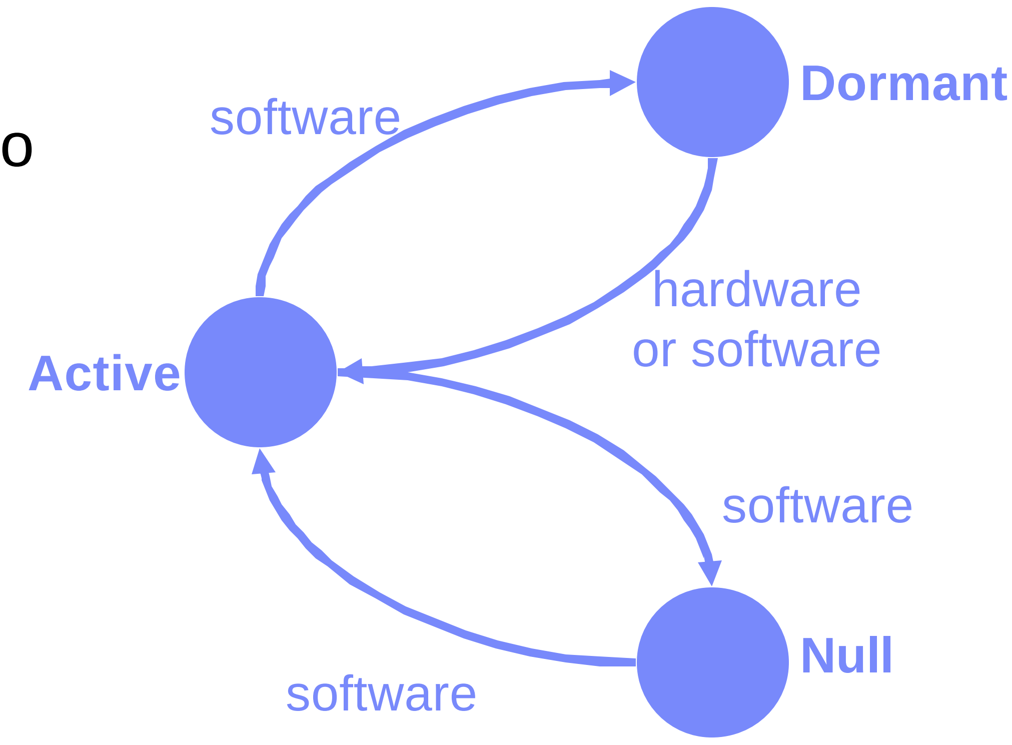
Single Thread Mode

Power Save Mode

Dynamic Thread Switching

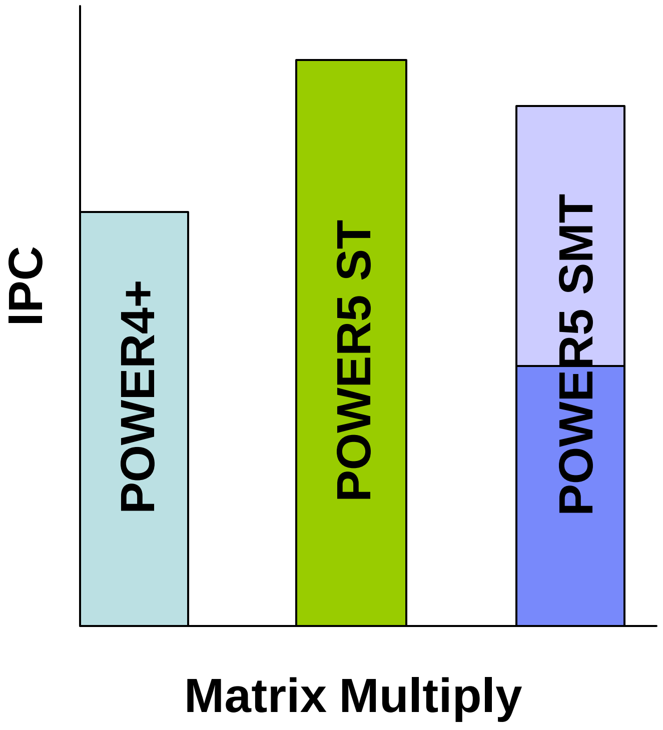
- Used if no task ready for second thread to run
- Allocates all machine resources to one thread
- Initiated by software
- Dormant thread wakes up on:
  - External interrupt
  - Decrementer interrupt
  - Special instruction from active thread

Thread States

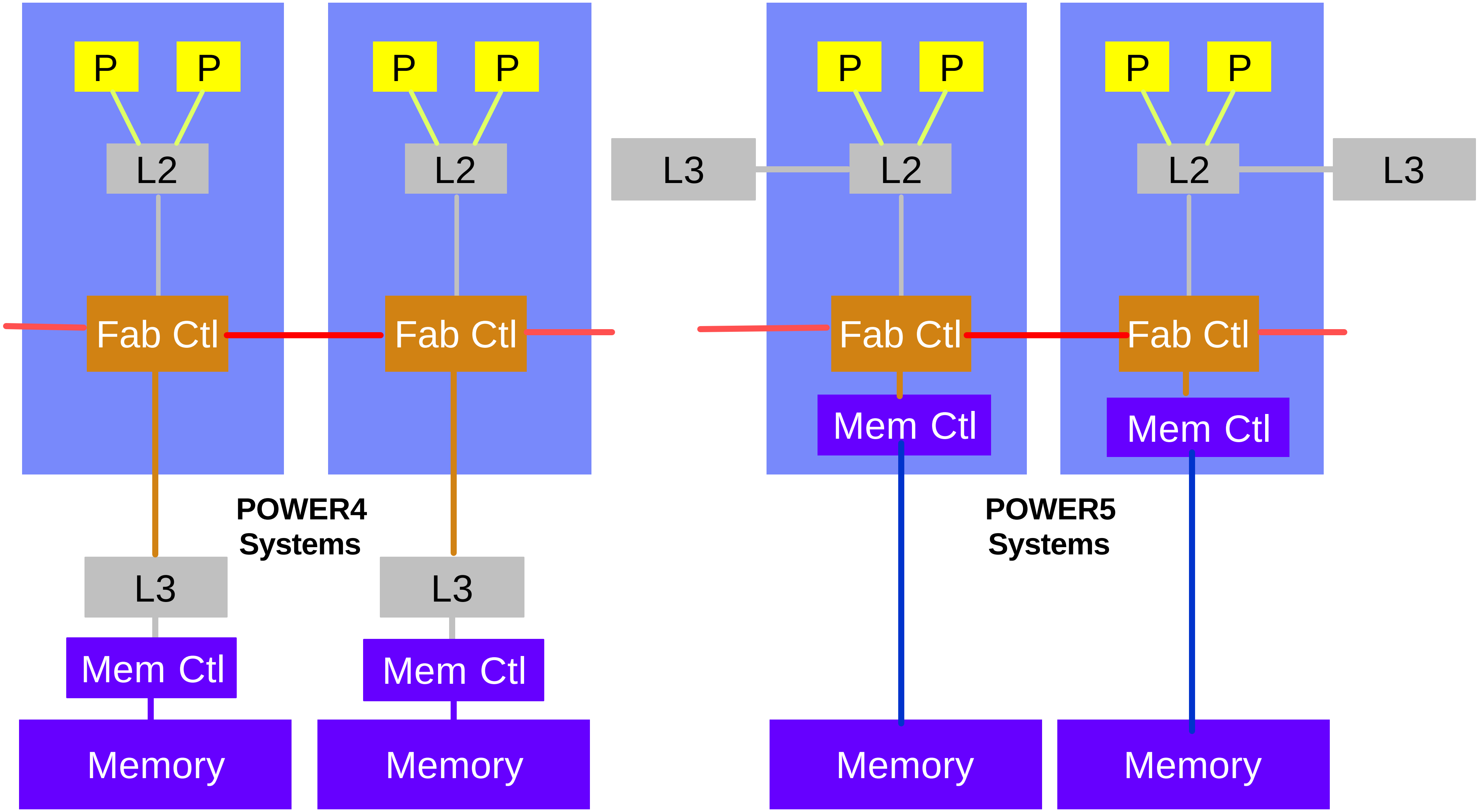


Single Thread Operation

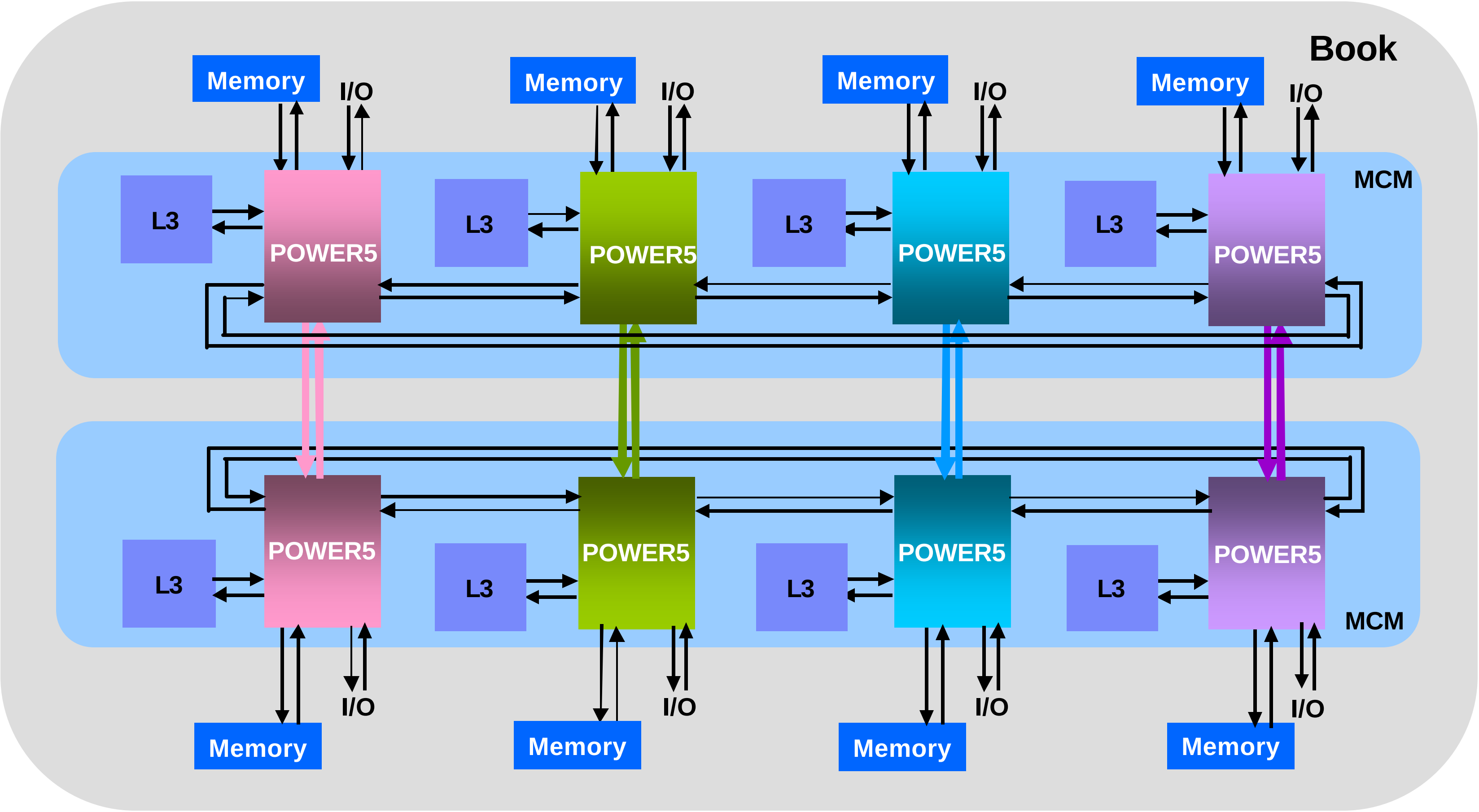
- Advantageous for execution unit limited applications
  - Floating or fixed point intensive workloads
- Execution unit limited applications provide minimal performance leverage for SMT
  - Extra resources necessary for SMT provide higher performance benefit when dedicated to single thread
- Determined dynamically on a per processor basis



Modifications to POWER4 System Structure

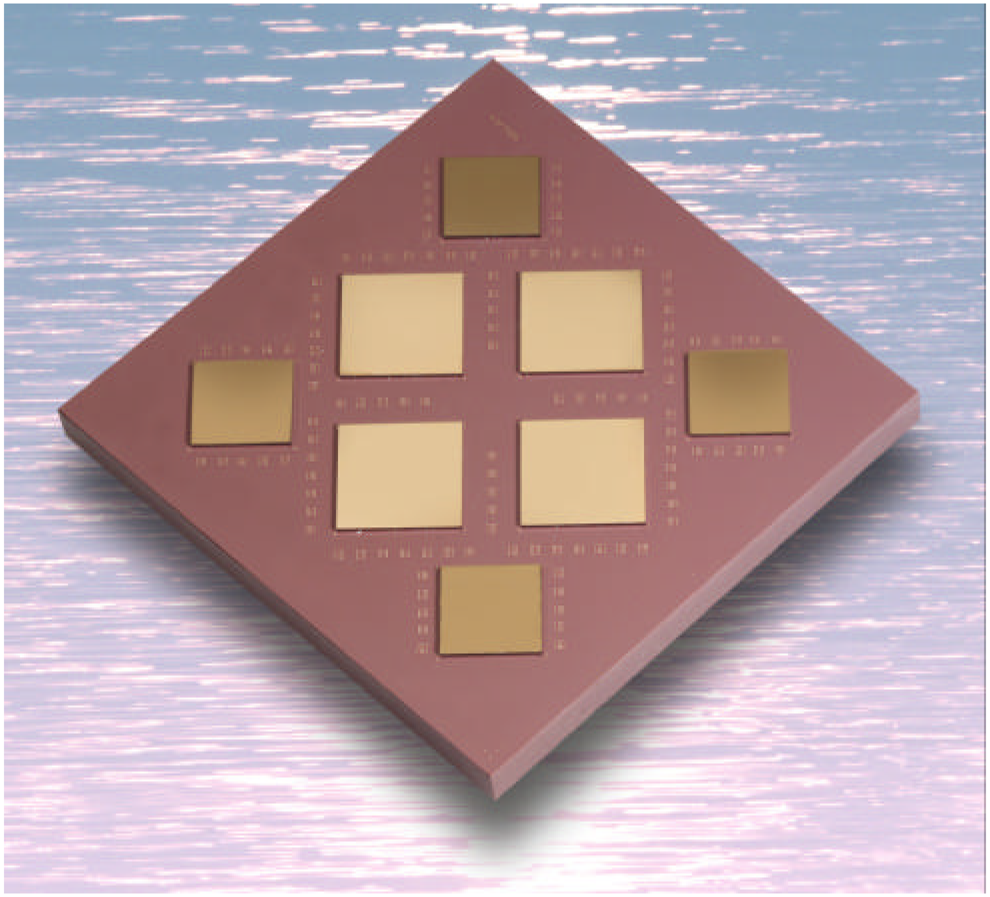


16-way Building Block

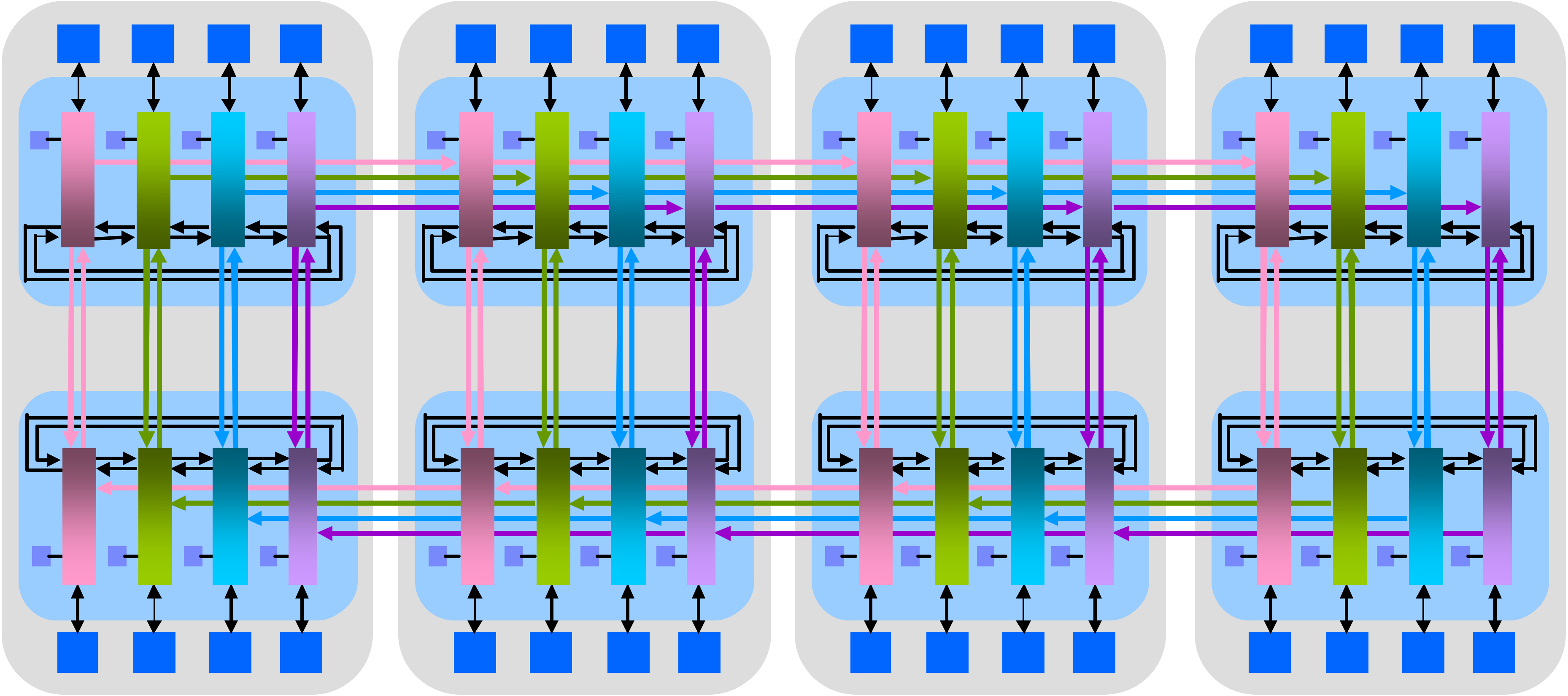


POWER5 Multi-chip Module

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



64-way SMP Interconnection



Interconnection exploits *enhanced distributed switch*

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

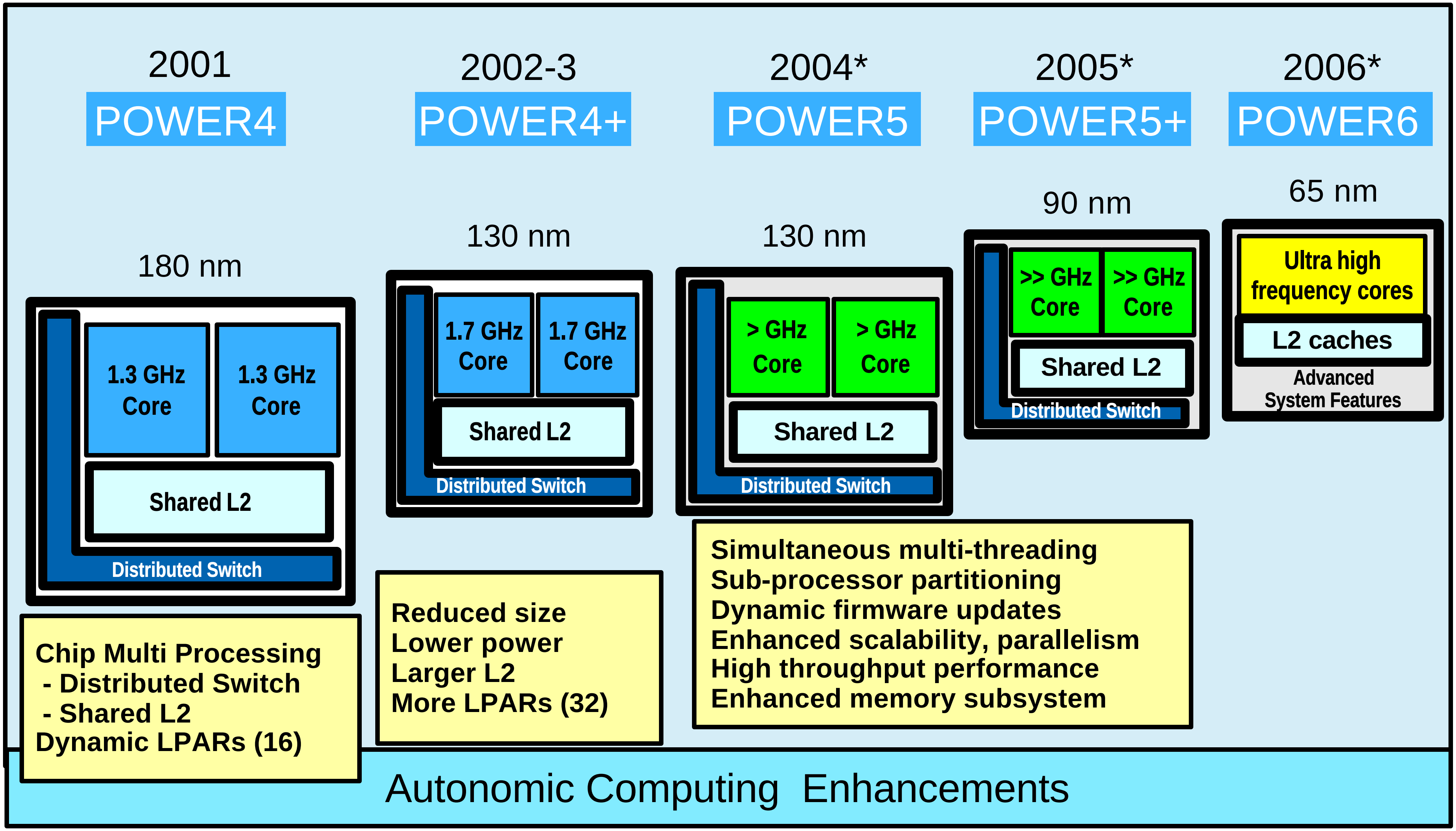
POWER4 and POWER5 Storage Hierarchy

	POWER4	POWER5
L2 Cache		
Capacity, line size	1.44 MB, 128 B line	1.92 MB, 128 B line
Associativity, replacement	8-way, LRU	10-way, LRU
Off-chip L3 Cache		
Capacity, line size	32 MB, 512 B line	36 MB, 256 B 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	½ processor speed	½ processor speed
Memory	512 GB maximum	1024 GB (1 TB) maximum

Other SMT Considerations

- Power Management
  - SMT Increases execution unit utilization
  - Dynamic power management does not impact performance
- Debug tools / Lab bring-up
  - Instruction tracing
  - Hang detection
  - Forward progress monitor
- Performance Monitoring
- Serviceability

POWER Server Roadmap



2001

POWER4

180 nm

1.3 GHz Core

1.3 GHz Core

Shared L2

Distributed Switch

2002-3

POWER4+

130 nm

1.7 GHz Core

1.7 GHz Core

Shared L2

Distributed Switch

2004*

POWER5

130 nm

> GHz Core

> GHz Core

Shared L2

Distributed Switch

2005*

POWER5+

90 nm

>> GHz Core

>> GHz Core

Shared L2

Distributed Switch

2006*

POWER6

65 nm

Ultra high frequency cores

L2 caches

Advanced System Features

Chip Multi Processing
- Distributed Switch
- Shared L2

Dynamic LPARs (16)

Reduced size
Lower power
Larger L2
More LPARs (32)

Simultaneous multi-threading
Sub-processor partitioning
Dynamic firmware updates
Enhanced scalability, parallelism
High throughput performance
Enhanced memory subsystem

Autonomic Computing Enhancements

*Planned to be offered by IBM. All statements about IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

Summary

- POWER5 SMT implementation is more than SMT
  - Good ROI for silicon area: Performance gain > Area increase
  - Resource sizes optimized
  - Dynamic feedback enhances instruction throughput
  - Software controlled priority exploits machine architecture
  - Dynamic ST to/from SMT mode capability optimizes system resources
- SMT impacts pervasive throughout chip
- Storage subsystem scalable to 64 Processor/ 128 Threads
- Operating in laboratory
  - AIX, Linux and OS/400 booted and running