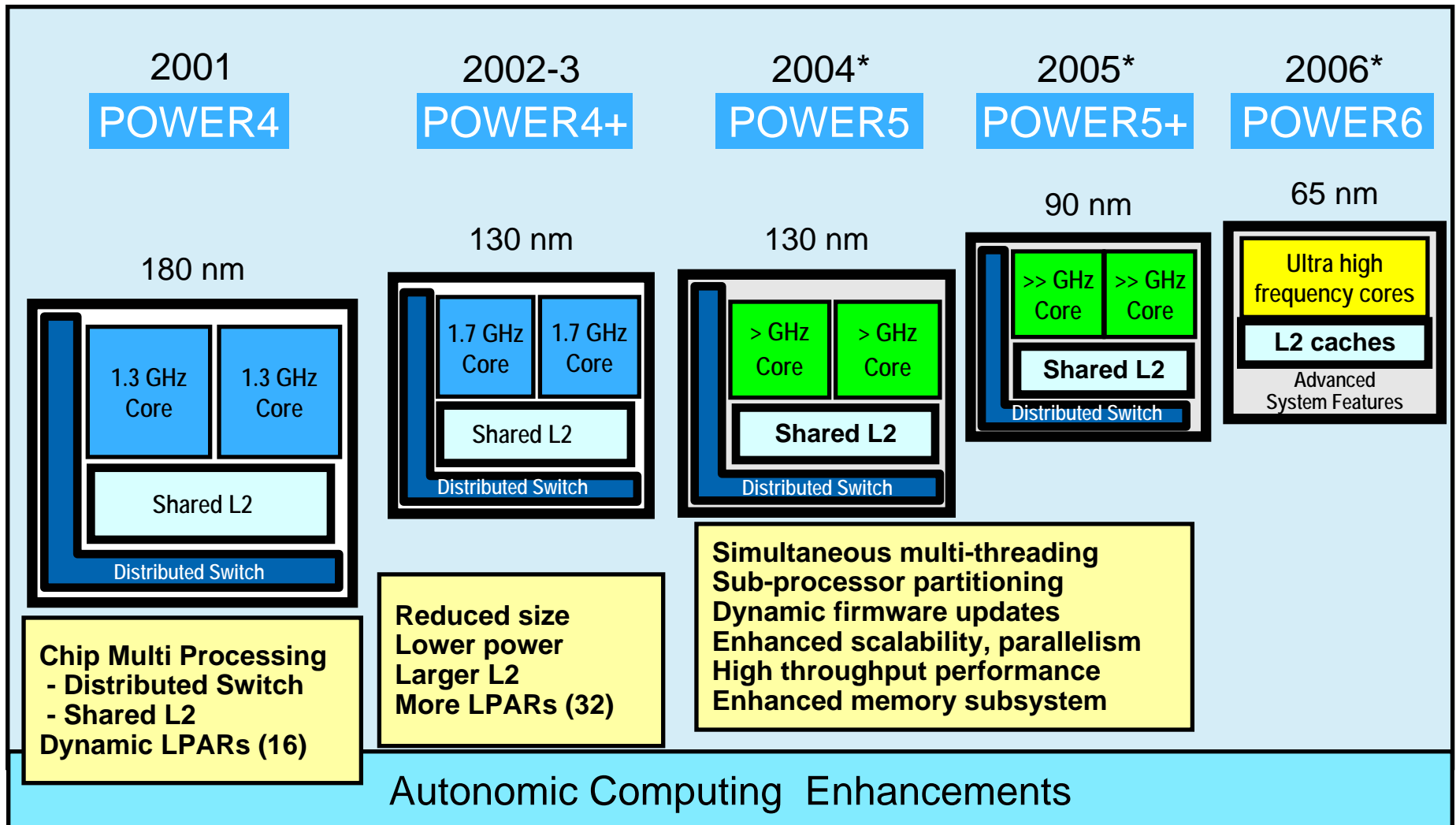# IBM's POWER5 Micro Processor Design and Methodology

Ron Kalla
IBM Systems Group

Outline

- POWER5 Overview
- Design Process
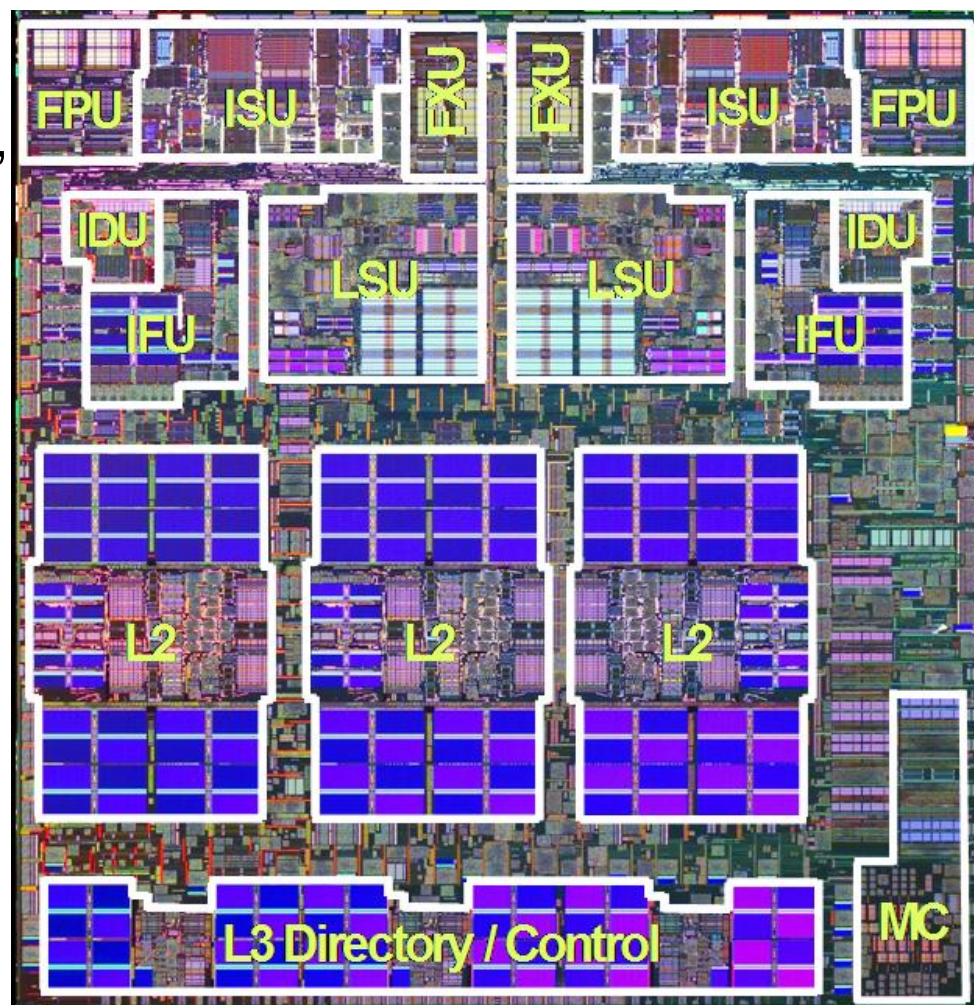- Power

POWER Server Roadmap

2001	2002-3	2004*	2005*	2006*
POWER4	POWER4+	POWER5	POWER5+	POWER6

2006* 65 nm — Ultra high frequency cores / L2 caches / Advanced System Features

2005* 90 nm — >> GHz Core, >> GHz Core / Shared L2 / Distributed Switch

2004* 130 nm — > GHz Core, > GHz Core / Shared L2 / Distributed Switch

2002-3 130 nm — 1.7 GHz Core, 1.7 GHz Core / Shared L2 / Distributed Switch

2001 180 nm — 1.3 GHz Core, 1.3 GHz Core / Shared L2 / Distributed Switch

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.
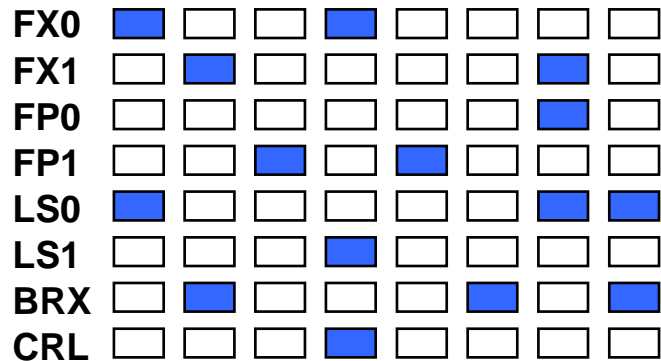
POWER5

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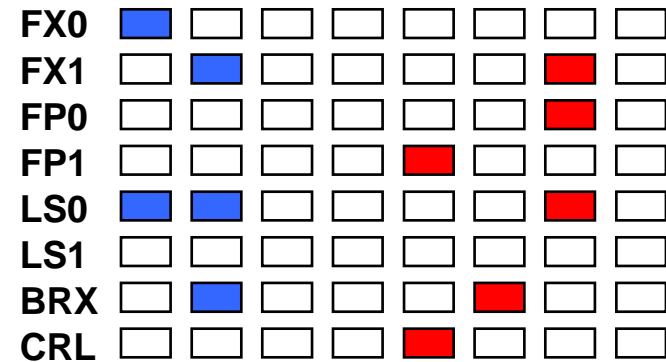 ▸ Up to 2 virtual processors per real processor
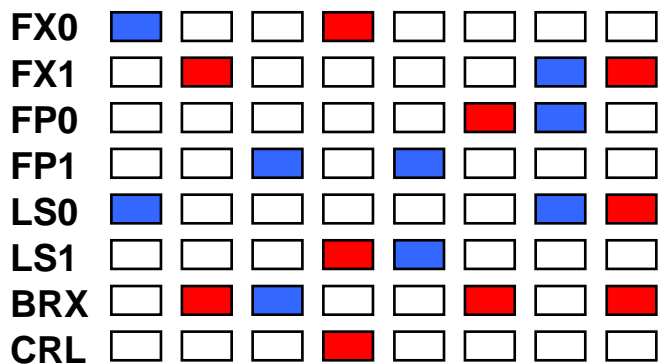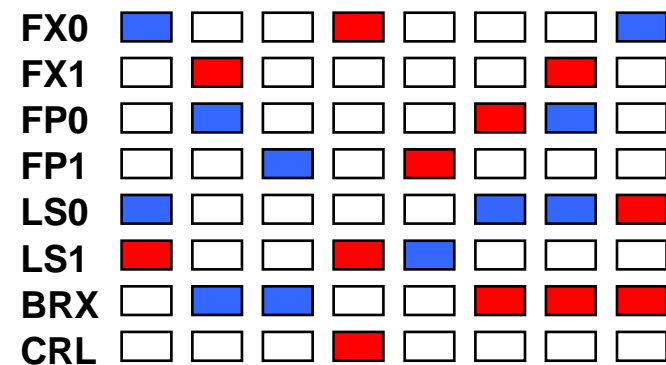
Multi-threading Evolution

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**
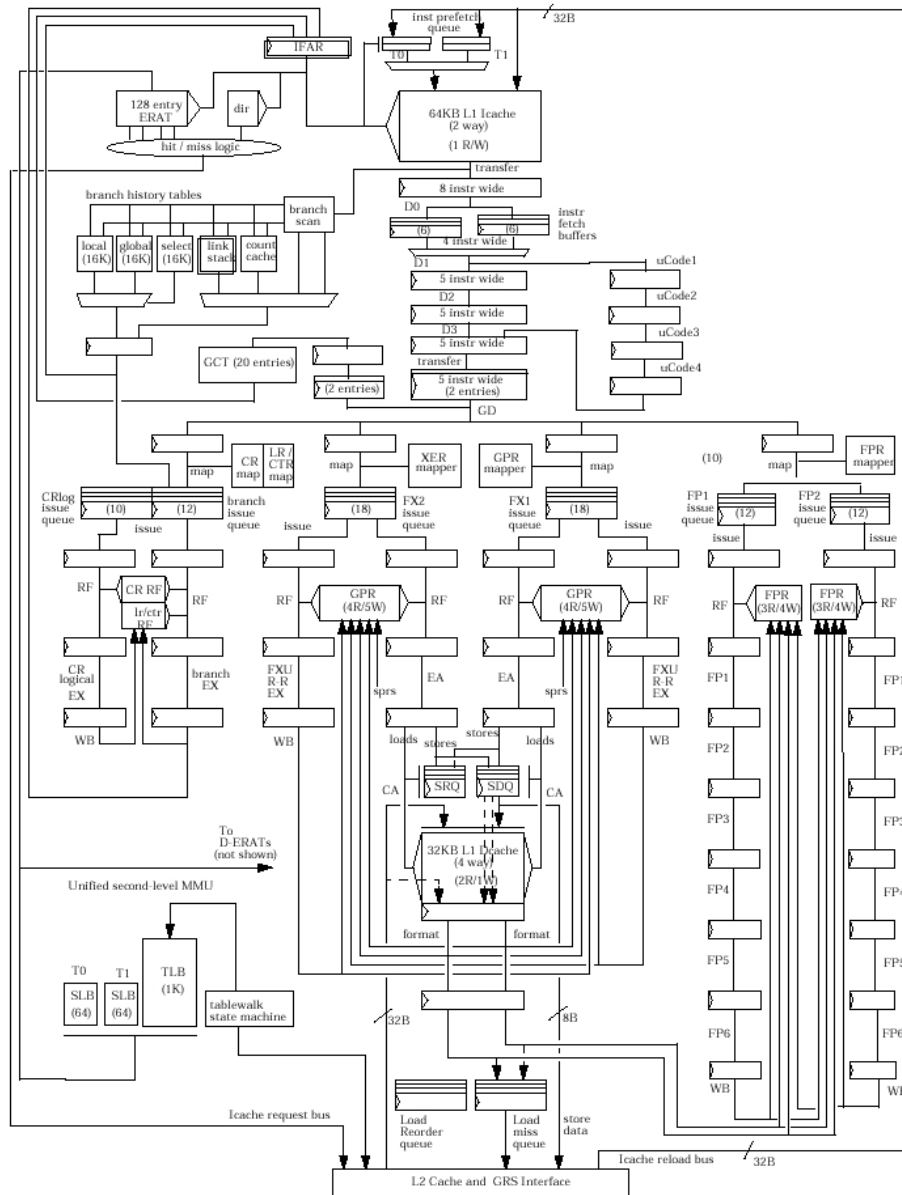
Thread Priority

- **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**
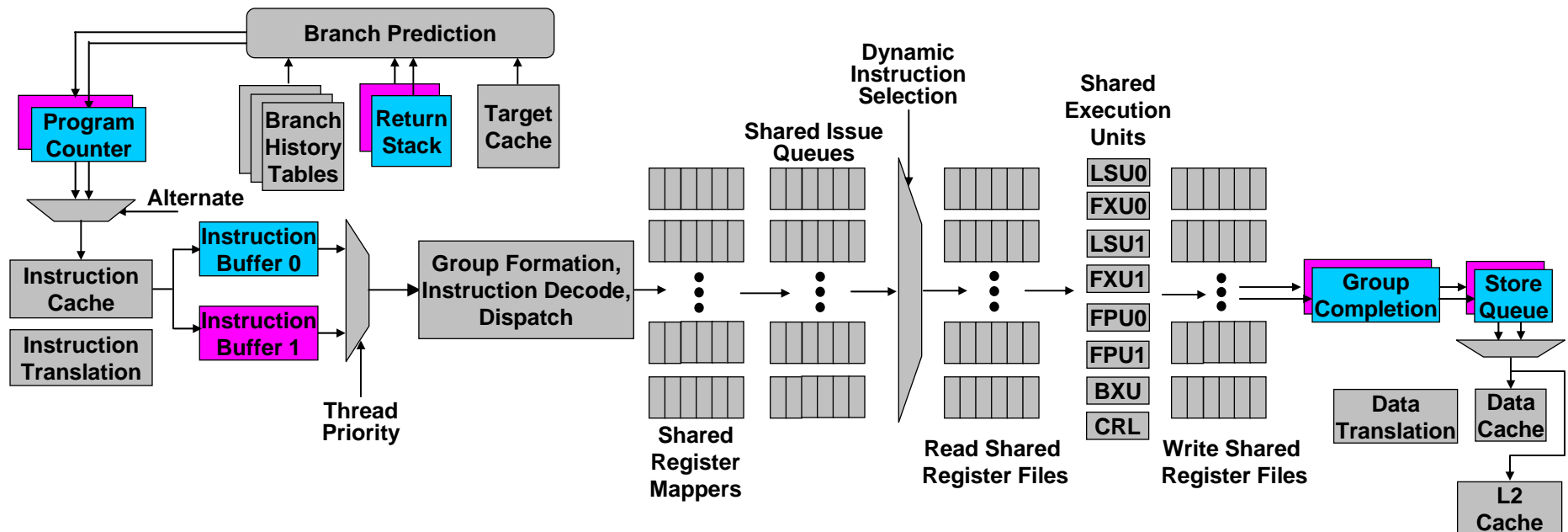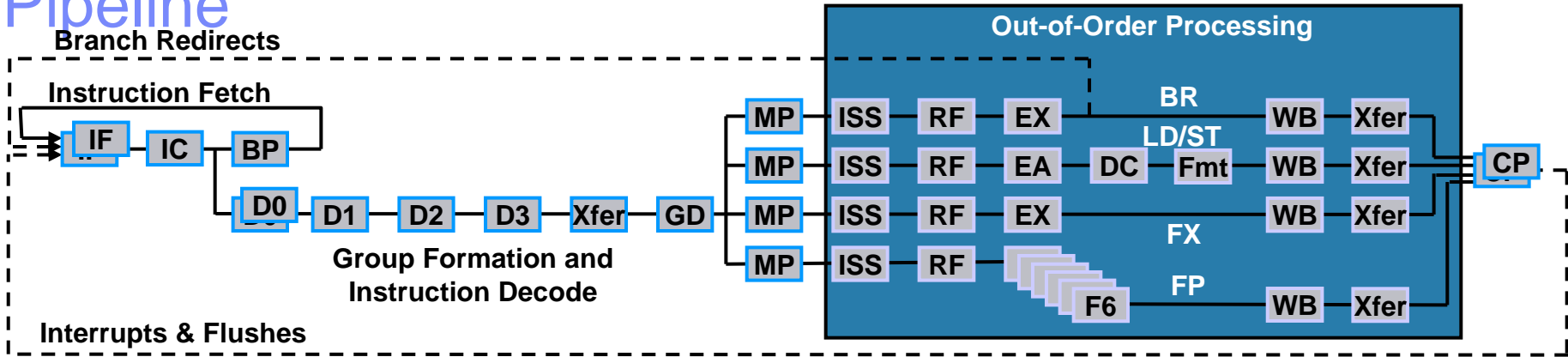   - ▸ Software/hardware controls 8 priority levels for each thread

Single Thread Mode

IPC

| 0,7 | -5 | -3 | -1 | 0 | 1 | 3 | 5 | 7,0 | 1,1 |

Thread 1 Priority - Thread 0 Priority

□ Thread 0 IPC ■ Thread 1 IPC

Power Save Mode
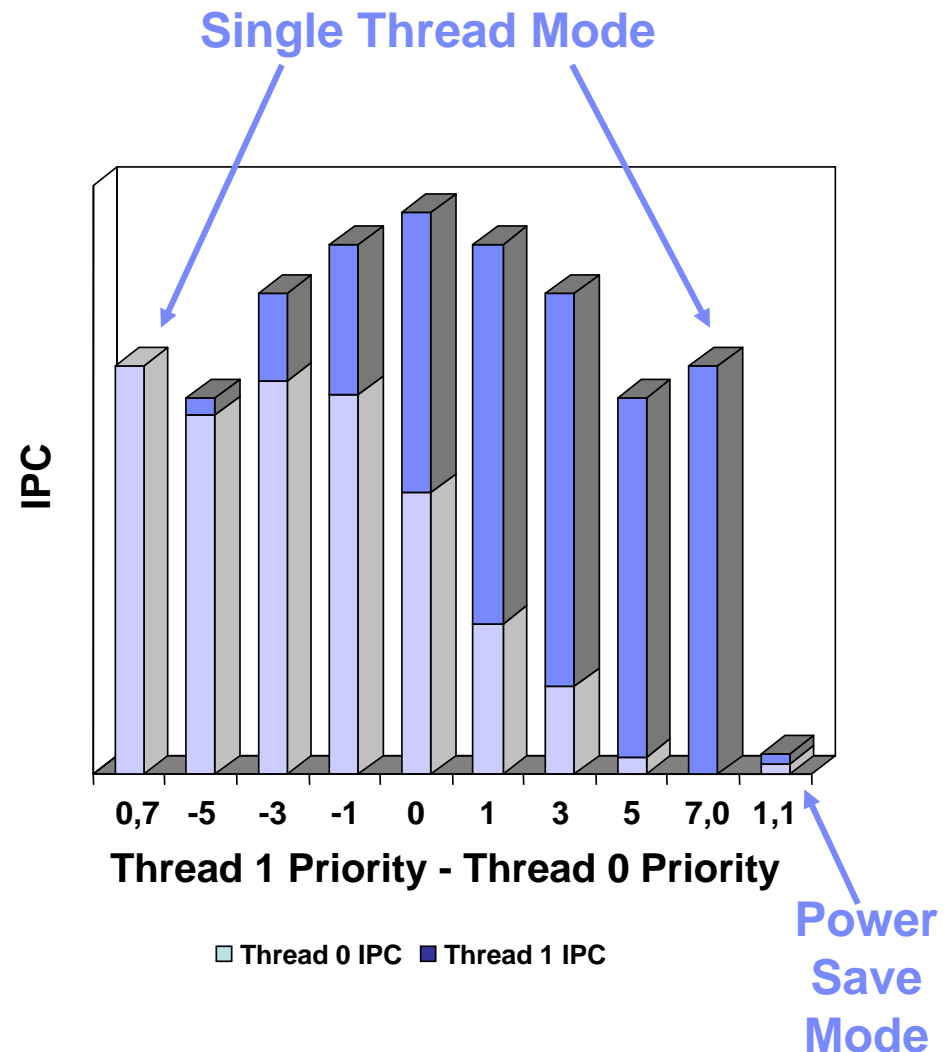
Terminology

- **PowerPC Addresses**
  - ▸ Effective>(SLB)>Virtual>(Page Table)>Real>(LPAR)>Physical
- **Instruction Execution**
  - ▸ I-fetch
  - ▸ Decode
  - ▸ Dispatch
 - ▸ Issue
  - ▸ Finish
  - ▸ Complete

Multithreaded Instruction Flow in Processor Pipeline

Out-of-Order Processing

Branch Redirects

Instruction Fetch

IF — IC — BP

D0 — D1 — D2 — D3 — Xfer — GD

Group Formation and Instruction Decode

Interrupts & Flushes

MP — ISS — RF — EX — **BR** — WB — Xfer

MP — ISS — RF — EA — DC — Fmt — **LD/ST** — WB — Xfer

MP — ISS — RF — EX — **FX** — WB — Xfer

MP — ISS — RF — F6 — **FP** — WB — Xfer

CP

Branch Prediction

Program Counter

Branch History Tables   **Return Stack**   **Target Cache**

Alternate

Instruction Cache

Instruction Translation

Instruction Buffer 0

Instruction Buffer 1

Thread Priority

Group Formation, Instruction Decode, Dispatch

Shared Issue Queues

Dynamic Instruction Selection

Shared Execution Units

LSU0
FXU0
LSU1
FXU1
FPU0
FPU1
BXU
CRL

Group Completion

Store Queue

Shared Register Mappers

Read Shared Register Files

Write Shared Register Files

Data Translation

Data Cache

L2 Cache

■ Shared by two threads   ■ Resource used by thread 0   ■ Resource used by thread 1

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**
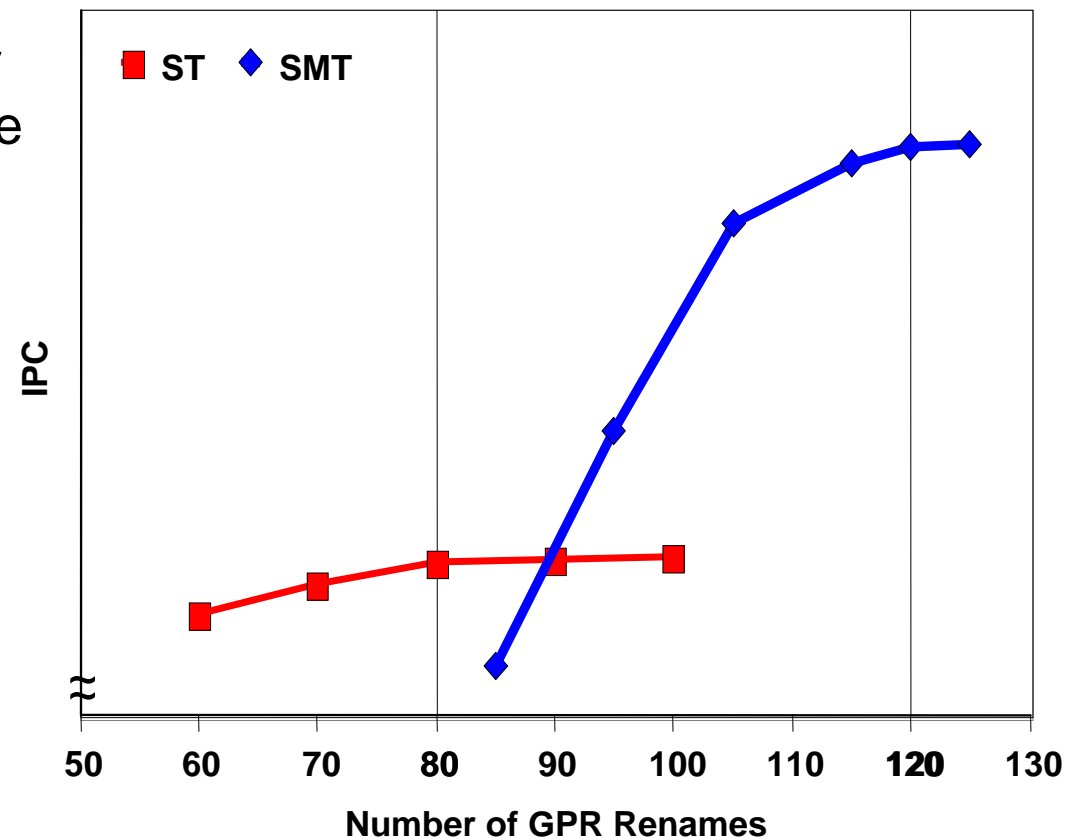
Results based on simulation of an online transaction processing application

Vertical axis does not originate at 0

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

IPC

POWER4+   POWER5 ST   POWER5 SMT

Matrix Multiply
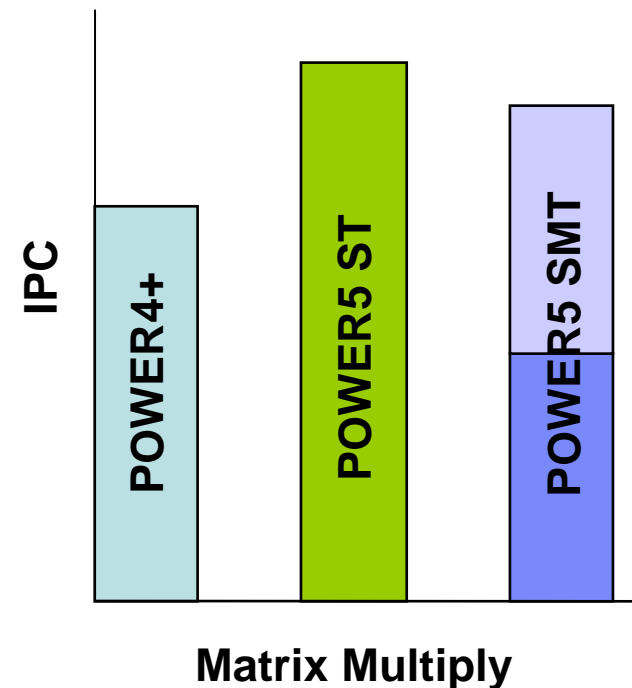
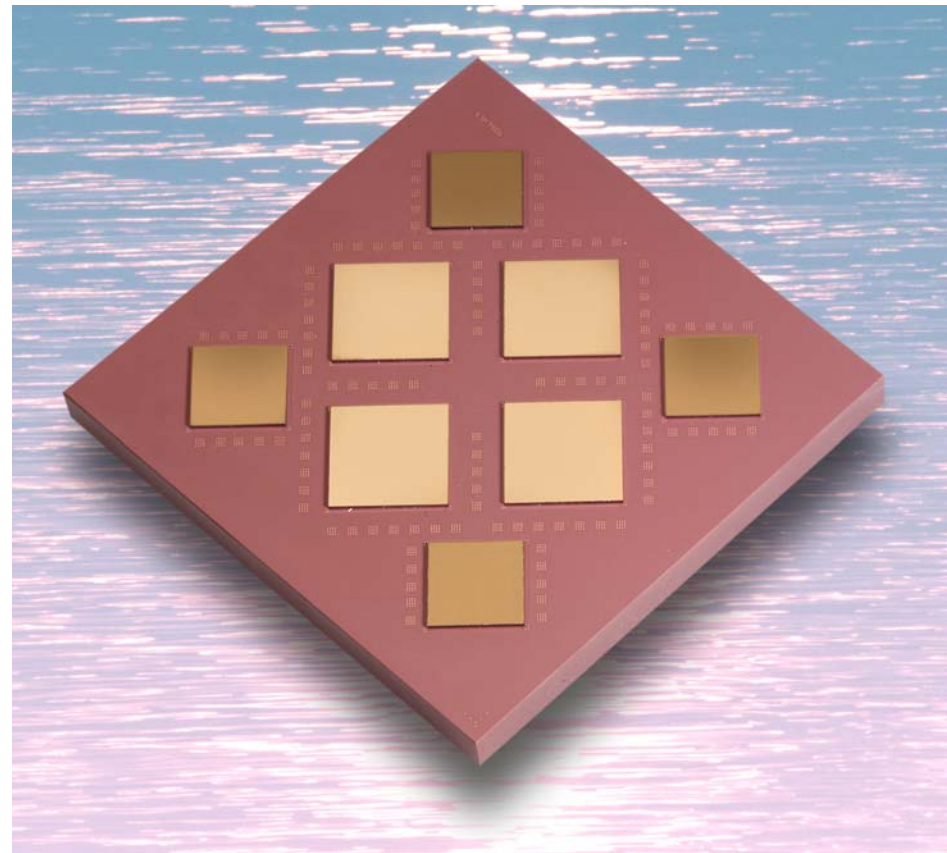# Modifications to POWER4 System Structure
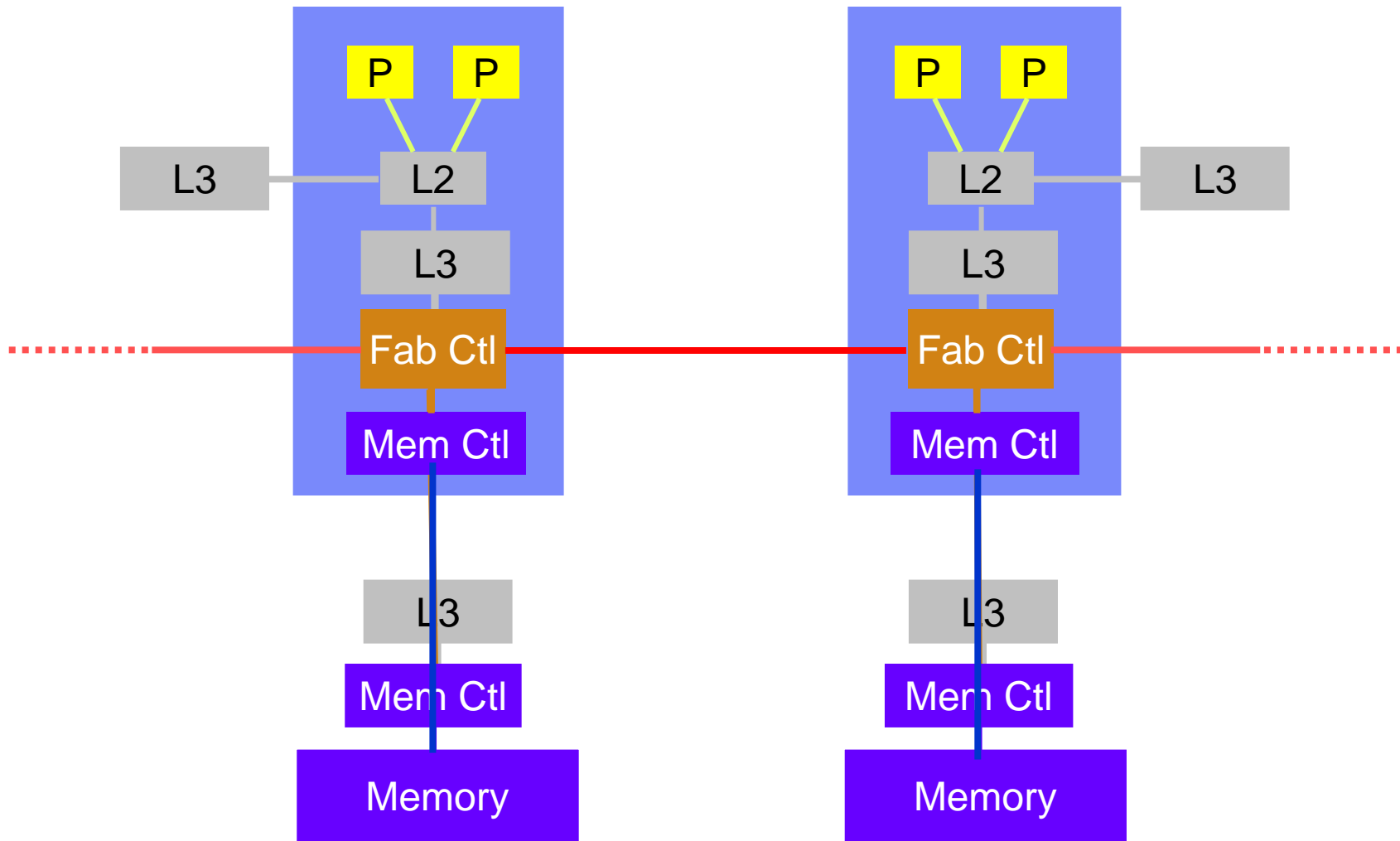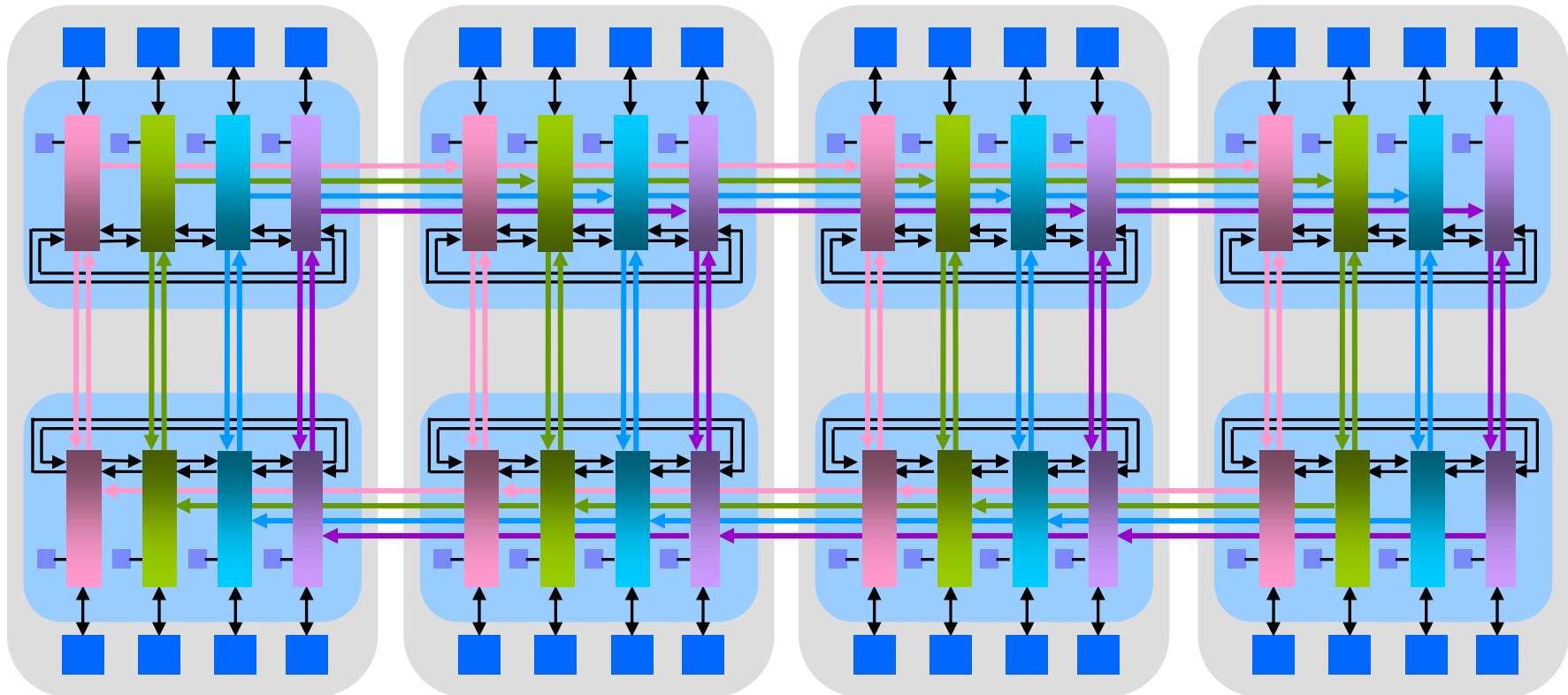
16-way Building Block
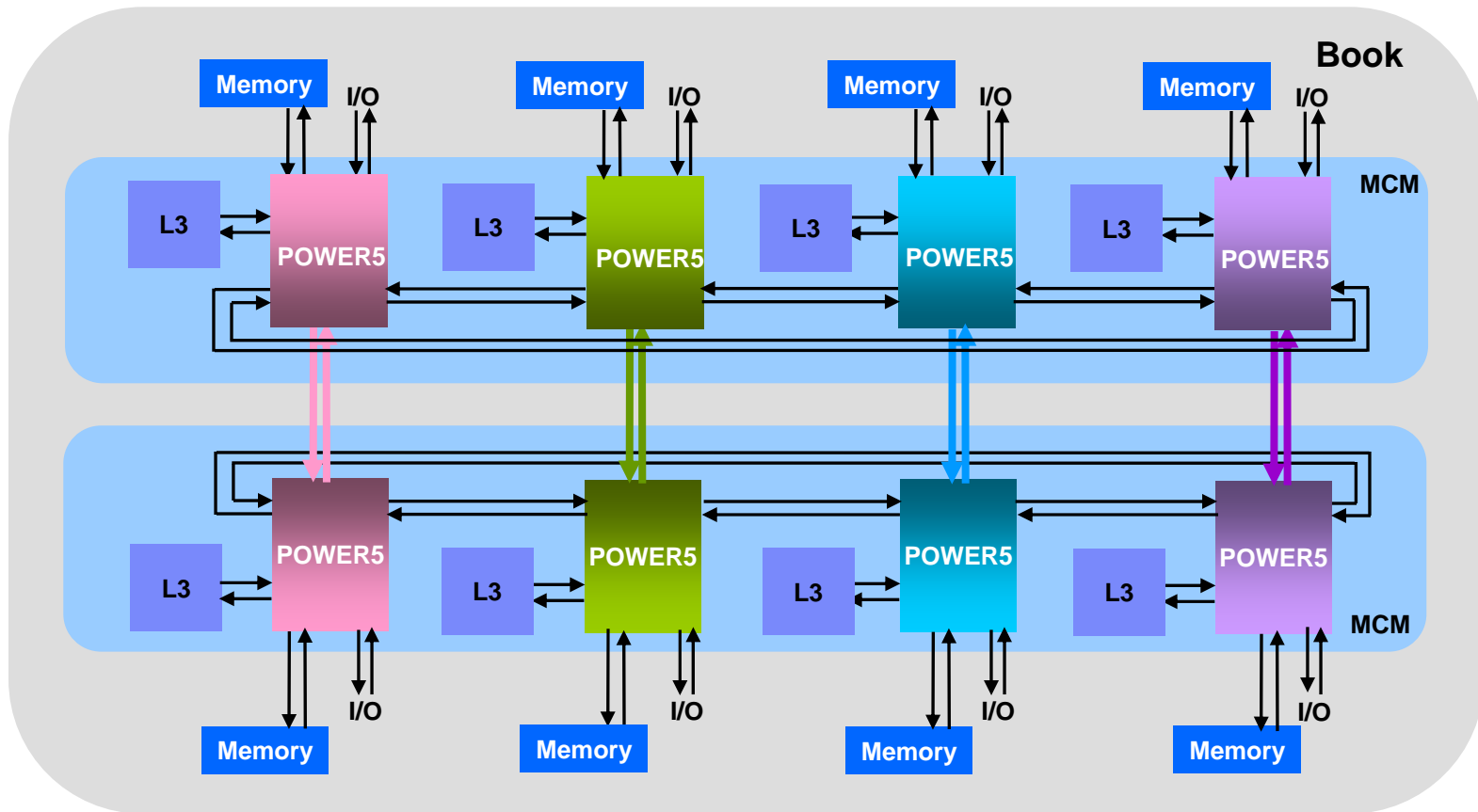
POWER5 Multi-chip Module

- 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

Design Process

- Concept Phase   (~10 People/4 months)

  ▸ Competitive assessment

  ▸ Customer requirements

  ▸ Technology assessment

    – Try to match technology introduction with products

  ▸ AREA/POWER estimates

    – Cost

  ▸ Schedule

  ▸ Staffing

  ▸ Lots or executive/technical reviews

Design Process (cont.)

- **High Level Design Phase   (~50 People/6 months)**
  - ▸ Micro-architecture
  - ▸ Cycle Time closed
    - – Cross section of critical paths
    - – Power closed
  - ▸ Area/Power budgets established
  - ▸ Arrays architected
  - ▸ All Unit I/O defined timing contract in place
  - ▸ Performance model meets concept phase targets
  - ▸ HLD Exit review.
  - ▸ Everything is now in place to effectively engage full team

Design Process (cont.)

- **Implementation Phase (~200 People/12-18 months)**
  - ▶ Write VHDL
  - ▶ Schematic entry for circuits
  - ▶ Simulation
    - – Block>unit>core>chip>system
  - ▶ Integration process begins
    - – RLM and Custom macros > unit > core >chip
  - ▶ Design rule checking, (can't trust human's)
  - ▶ Bulk simulation
 - ▶ Check recheck and check again
- **RIT (Release Information Tape)**

Bring-up 100's People 12-18 Months

- Starts at Wafer test.
- AVPs
- Random test programs
- Boot operating system

  ▸ OS based testing

- Formal testing (new groups of people)

  ▸ POWER (Will discuss later)

  ▸ Performance testing

- Additional RITs if necessary
- Release to mfg
- GA   (Hurray)
- Field Support

Power

- **Micro-Processor Power Complex equation.**
  - ▸ DC Power (IDDQ)
    - – Voltage
    - – Temperature
    - – Speed of Chip (PSRO) Processor speed Ring Oscillator.
  - ▸ AC Power
    - – Function of CV^2f X Switching Factor.
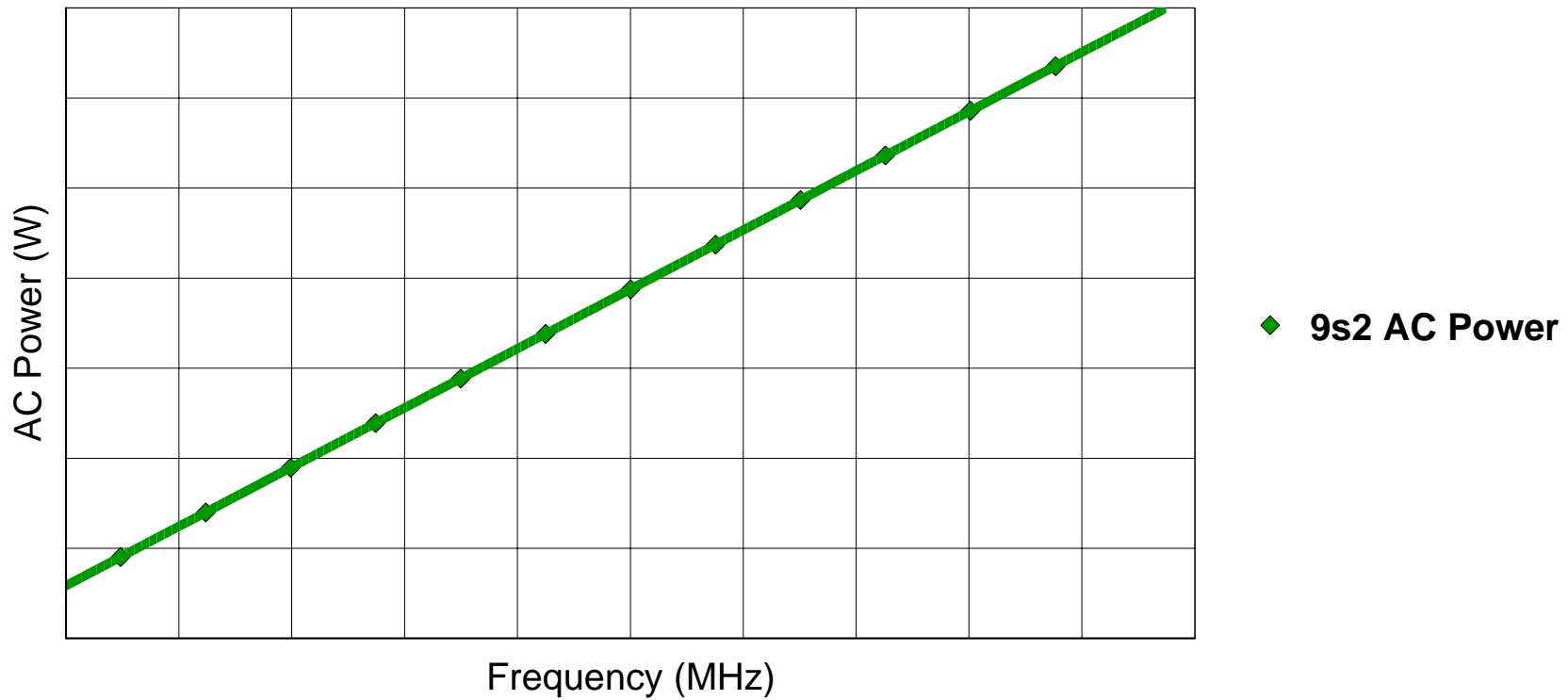   - – Workload Dependent.

GR DD2.0 Fmax vs. 1.2v, 25C PSRO
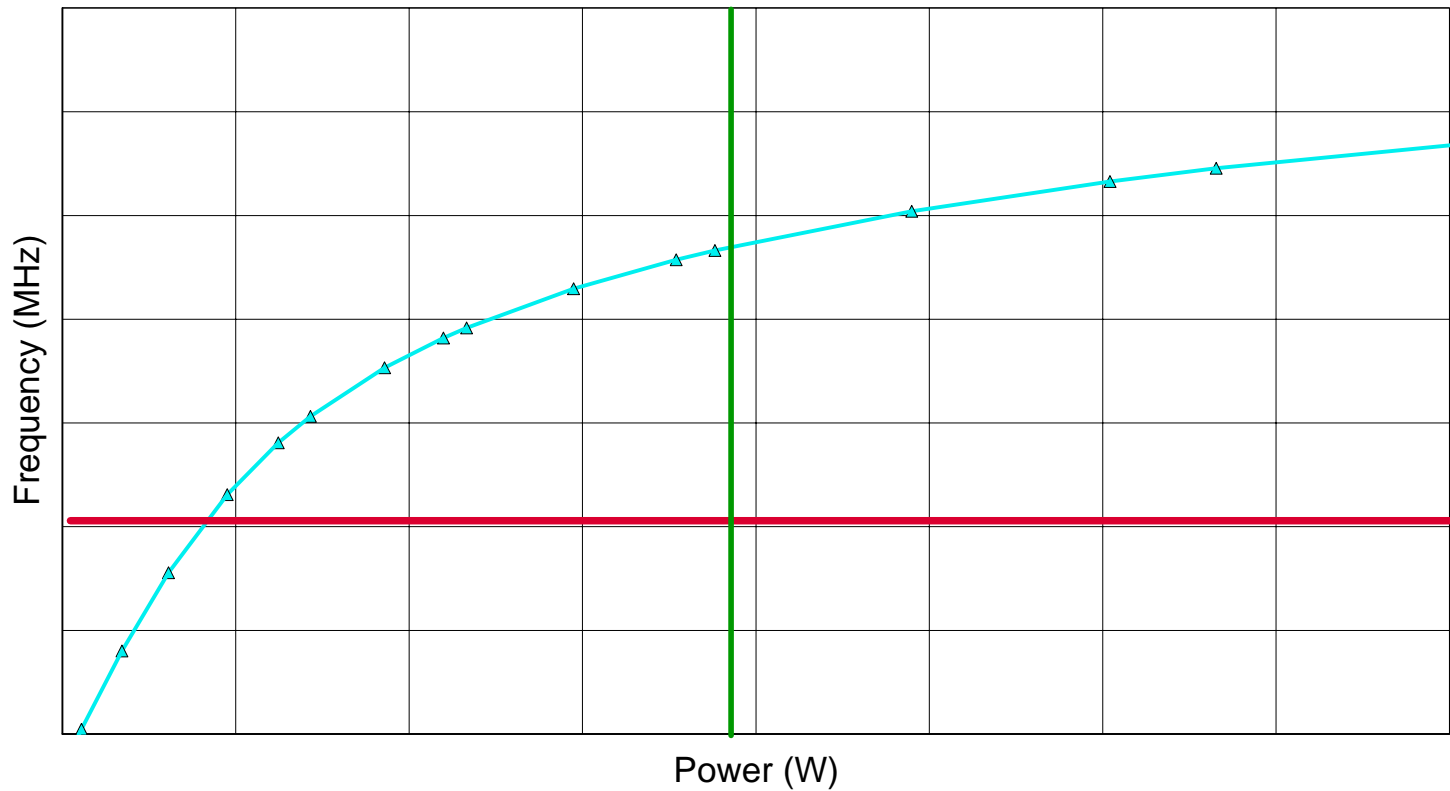## Unguardbanded, 1.3v, 85C Fmax

Frequency (MHz)

PSRO (ps/stage)
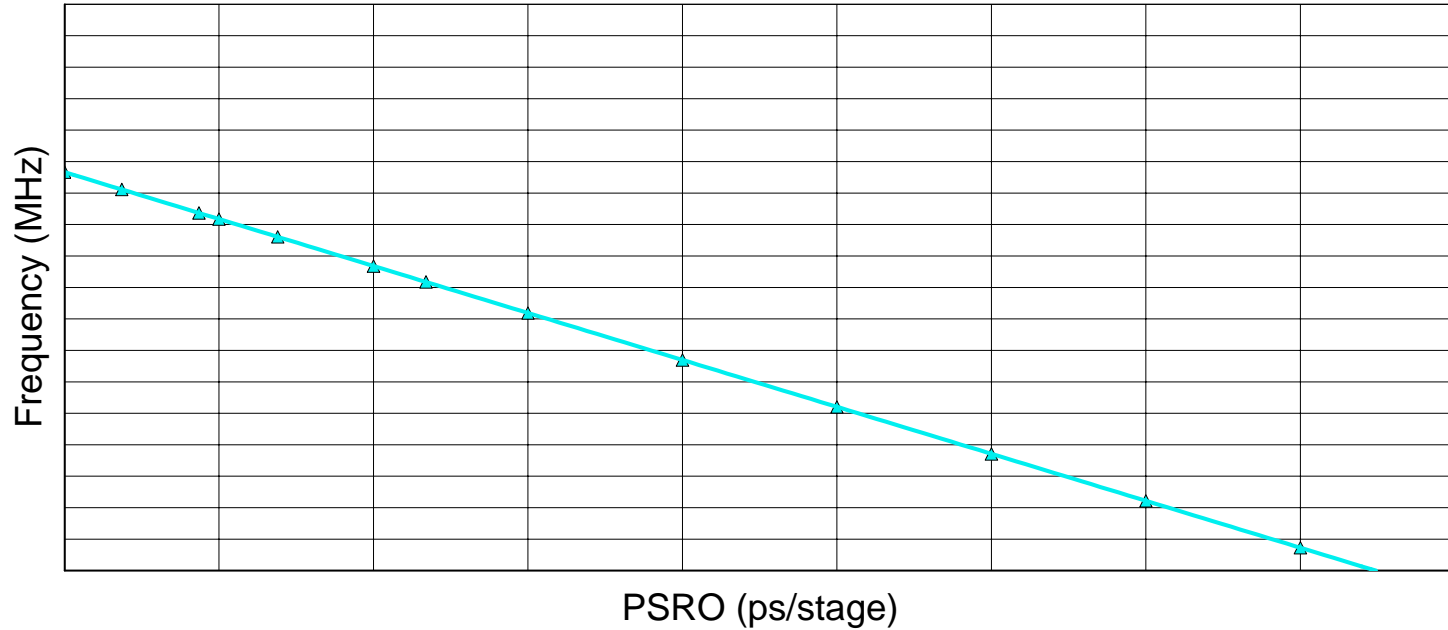
Power 5 Leakage Power vs. PSRO
1.3 V and 85C

9s2
Leakage
Power

Leakage Power (W)

PSRO (ps/stage)
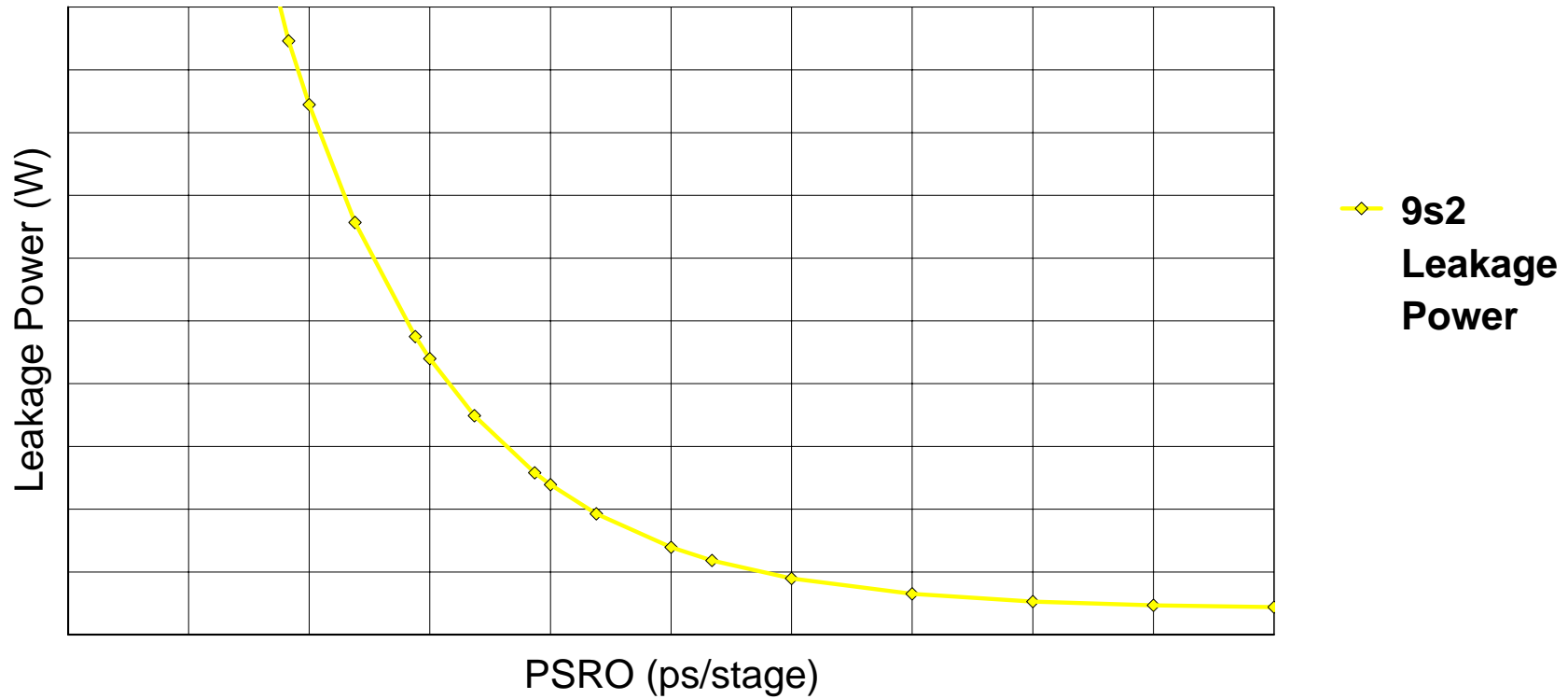
Power 5 AC Power vs. Frequency
1.3 V and 85 C

Chip Power vs Frequency 1.3V 85C

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