



# POWER5, Federation and Linux: System Architecture

Charles Grassl  
IBM  
August, 2004

© 2004 IBM



## Agenda

- **Architecture**
  - Processors
    - POWER4
    - POWER5
  - System
    - Nodes
    - Systems
- **High Performance Switch**

## Processor Design

### POWER4

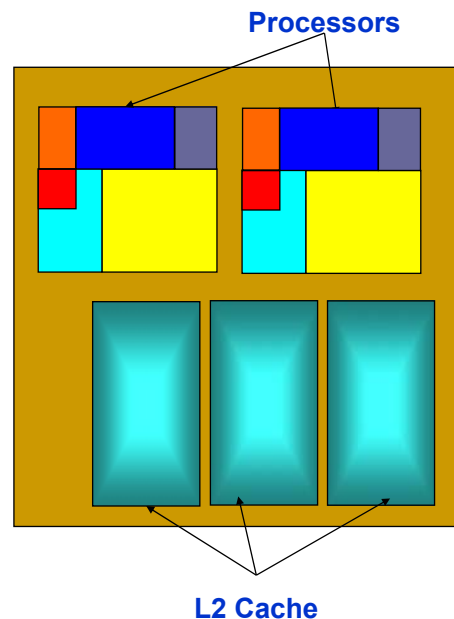
- Two processors per chip
- L1:
  - 32 kbyte data
  - 64 kbyte instructions
- L2:
  - 1.45 Mbyte per processor pair
- L3:
  - 32 Mbyte per processor pair
  - Shared by all processors

### POWER5

- Two processors per chip
- L1:
  - 32 kbyte data
  - 64 kbyte instructions
- L2:
  - 1.9 Mbyte per processor pair
- L3:
  - 36 Mbyte per processor pair
  - Shared by processor pair
  - Extension of L2 cache
- Simultaneous multithreading
- Enhanced memory loads and stores
- Additional rename registers

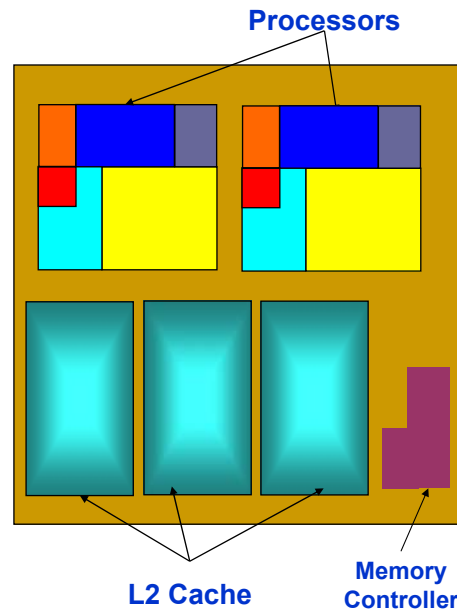
## POWER4: Multi-processor Chip

- Each processor:
  - L1 caches
  - Registers
  - Functional units
- Each chip (shared):
  - L2 cache
  - L3 cache
    - Shared across node
  - Path to memory



## POWER5: Multi-processor Chip

- **Each processor:**
  - L1 caches
  - Registers
  - Functional units
- **Each chip (shared):**
  - L2 cache
  - L3 cache
  - **NOT shared**
  - Path to memory
  - **Memory controller**



## POWER4: Features

- **Multi-processor chip**
- **High clock rate**
  - High computation burst rate
- **Three cache levels**
  - Bandwidth
  - Latency hiding
- **Shared Memory**
  - Large memory size

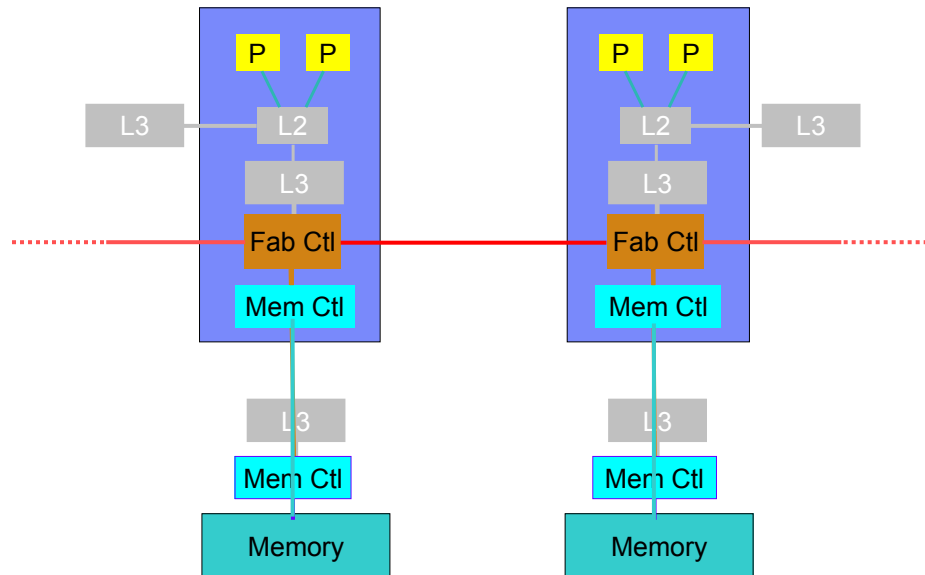
## POWER5 Features

- **Performance**
  - Registers
    - Additional rename registers
  - Cache improvements
    - L3 runs at ½ freq. (POWER4: 1/3 freq.)
  - Simultaneous Multi Threading (SMT)
  - Memory Bandwidth
- **Virtualization**
  - Partitioning
- **Dynamic power management**

## POWER4, POWER5 Differences

	POWER4 Design	POWER5 Design	Benefit
<b>L1 Cache</b>	2-way Associative FIFO	4-way Associative LRU	Improved L1 Cache performance
<b>L2 cache</b>	8-way Associative 1.44MB	10-way Associative 1.9MB	Fewer L2 Cache misses Better performance
<b>L3 Cache</b>	32MB 8-way Associative 118 Clock Cycles	36MB 12-way Associative ~80 Clock Cycles	Better Cache performance
<b>Memory Bandwidth</b>	4GByte / sec / Chip	~16Gbyte / sec / Chip	4X improvement Faster memory access
<b>Simultaneous Multi- Threading</b>	No	Yes	Better processor utilization
<b>Processor Addressing</b>	1 processor	1/10 of processor	Better usage of processor resources
<b>Size</b>	412mm	389mm	50% more transistors in the same space

## Modifications to POWER4 System Structure



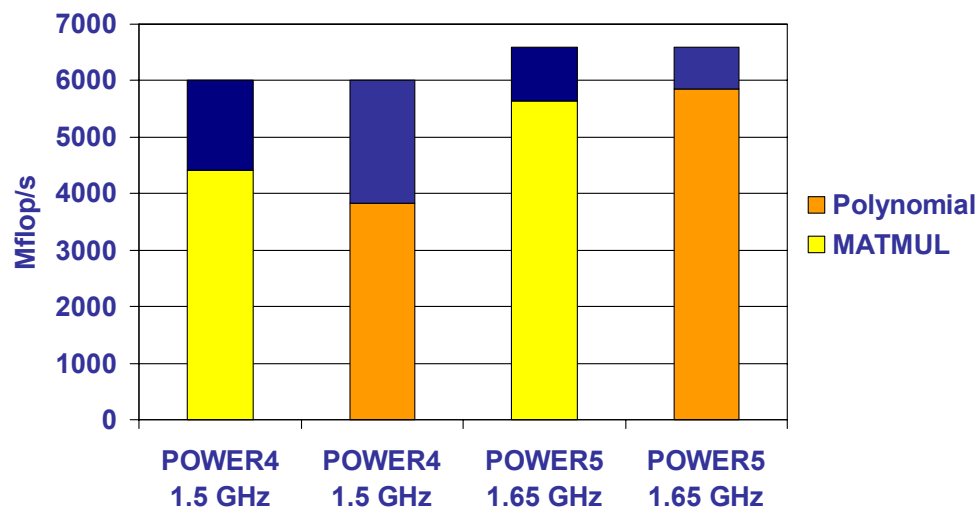
## POWER5 Challenges

- **Scaling from 32-way to logical 128-way**
  - Large system effects, 2Tbyte of memory
- **1.9MB L2 shared across 4 threads (SMT)**
  - L2 miss rates
- **36 Mbyte private L3 versus 128 Mbyte shared L3**
  - Will have to watch multi-programming level
  - Read-only data will get replicated
- **Local/Remote memory latencies**
  - Worst case remote memory latency is 60% higher than local

## Effect of Features: Registers

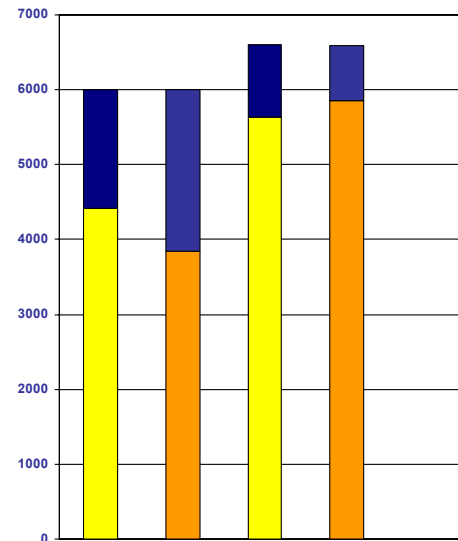
- **Additional rename registers**
  - POWER4: 72 total registers
  - POWER5: 110 total registers
- **Matrix multiply:**
  - POWER4: 65% of burst rate
  - POWER5: 90% of burst rate
- **Enhances performance of computationally intensive kernels**

## Effect of Rename Registers

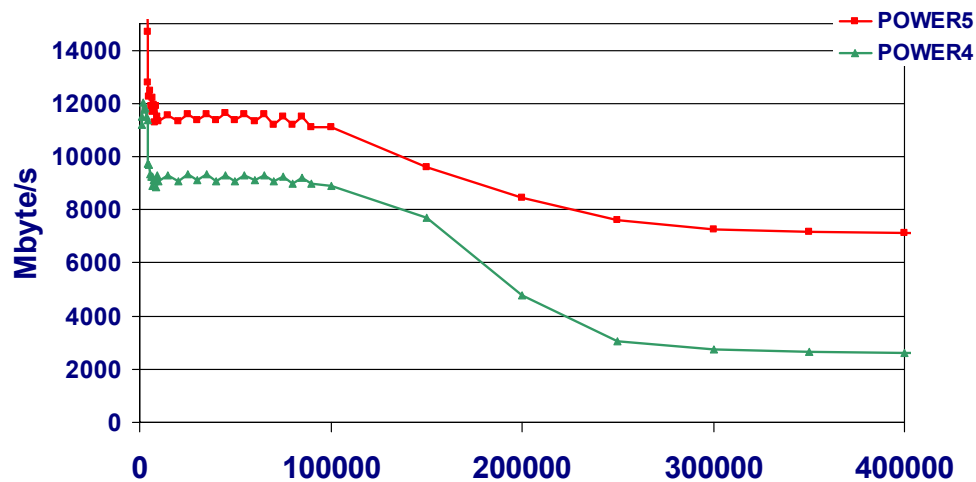


## Effect of Rename Registers

- “Tight” algorithms were previously register bound
- Increased floating point efficiency
- Examples
  - Matrix multiply
  - LINPACK
  - Polynomial calculation
  - Horner’s Rule
  - FFTs

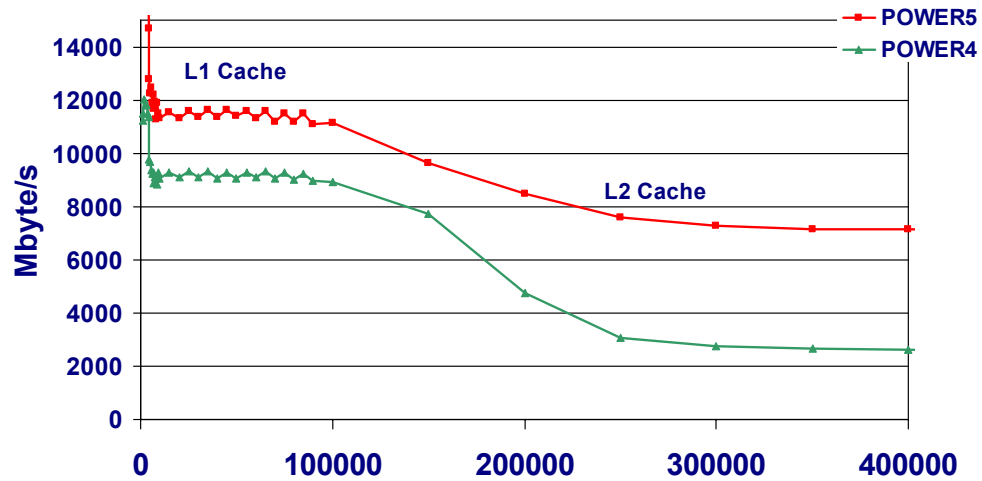


## Memory Bandwidth

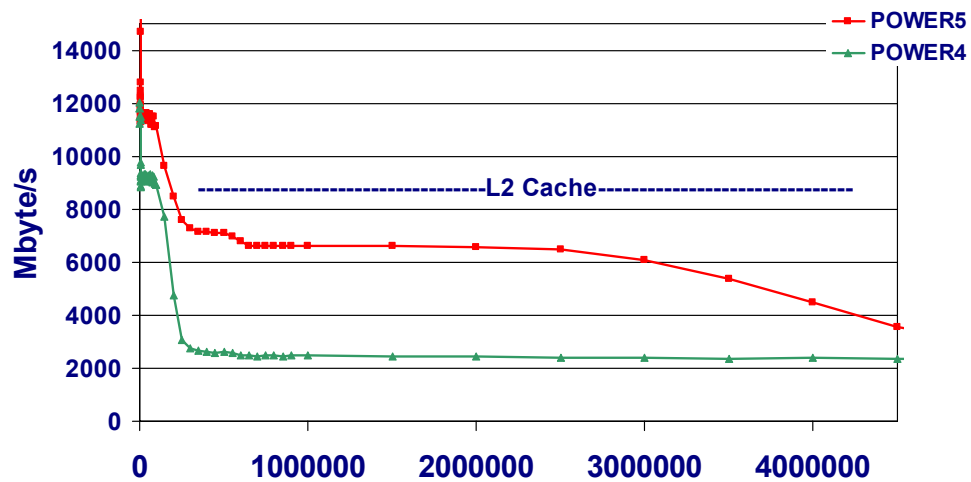


$s=s+A(i)$   
 1.65 GHz POWER5  
 1.5 GHz POWER4

## Memory Bandwidth: L1 – L2 Cache Regime

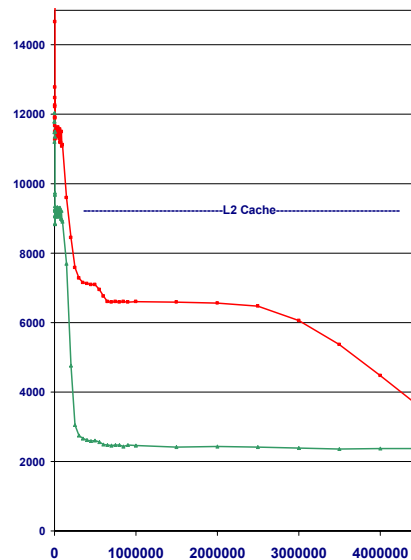


## Memory Bandwidth: L2 Cache Regime



## L2 and L3 Cache Improvement

- L3 cache enhances bandwidth
  - Previously only hid latency
  - New L3 cache is like a “2.5 cache”
- Larger blocking sizes

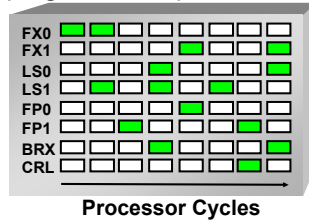


## Simultaneous Multi-Threading

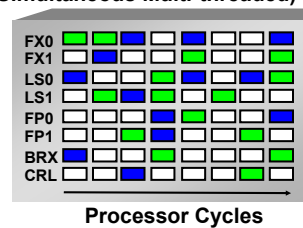
- Two threads per processor
  - Threads execute concurrently
- Threads share:
  - Caches
  - Registers
  - Functional units

# POWER5 Simultaneous Multi-Threading (SMT)

**POWER4  
(Single Threaded)**



**POWER5  
(Simultaneous Multi-threaded)**

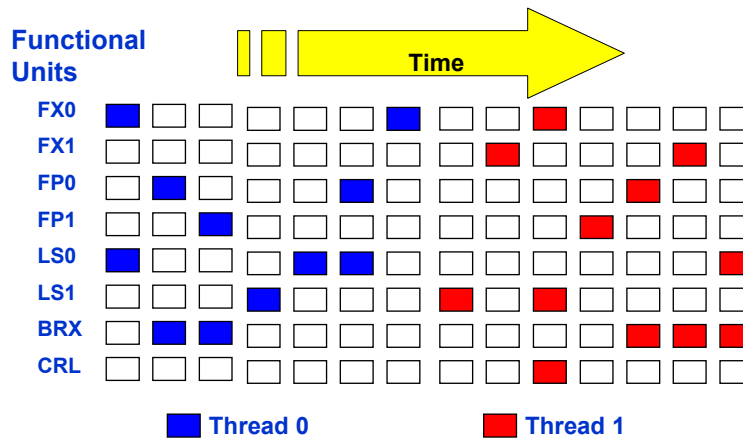


**Legend**

- █ Thread0 active
- No Thread active
- █ Thread1 active

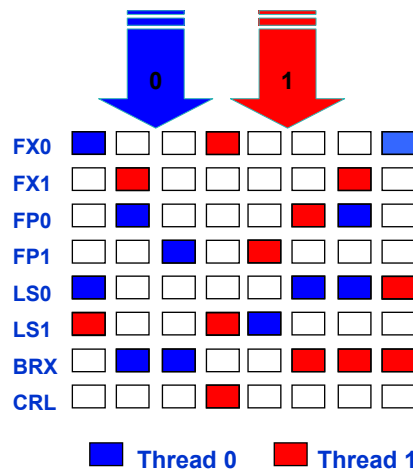
- Presents SMP programming model to software
- Natural fit with superscalar out-of-order execution core

# Conventional Multi-Threading



- Threads alternate
- Nothing shared

## Simultaneous Multi-Threading



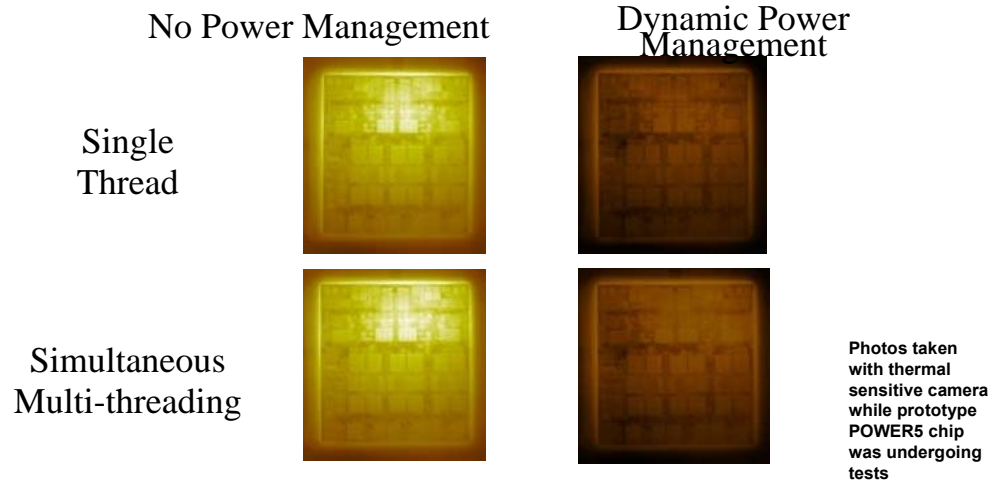
- **Simultaneous execution**
  - Shared registers
  - Shared functional units

## Manipulating SMT

- Administrator function
- Dynamic
- System-wide

Function	Command
View	<code>/usr/sbin/smtctl</code>
Off	<code>/usr/sbin/smtctl -m off now</code>
On	<code>/usr/sbin/smtctl -m on now</code>

## Dynamic Power Management

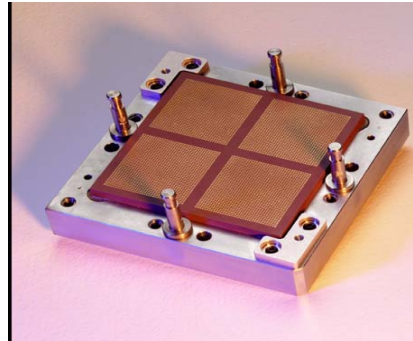


## Multi-Chip Modules (MCMs)

- **Four processor chips are mounted on a module**
  - Unit of manufacture
  - Smallest freestanding system
- **Multiple modules**
  - Node
- **Also available: Single Chip Modules**
  - Lower cost
  - Lower bandwidth

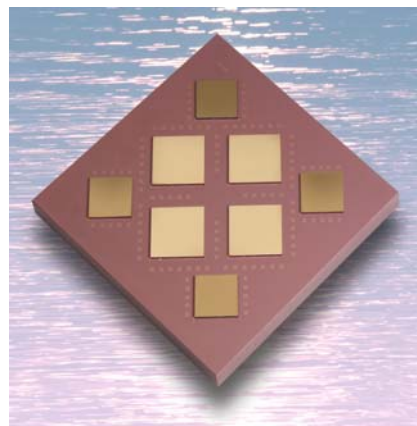
## POWER4 Multi-Chip Module (MCM)

- 4 chips
  - 8 processors
  - 2 processors/chip
- >35 Gbyte/s in each chip-to-chip connection
- 2:1 MHz chip-to-chip busses
- 6 connections per MCM
- Logically shared L2's, L3's
- Data Bus (Memory, inter-module):
- 3:1 MHz MCM to MCM



## POWER5 Multi-Chip Module (MCM)

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



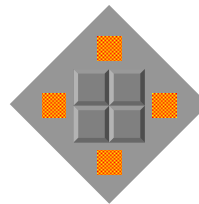
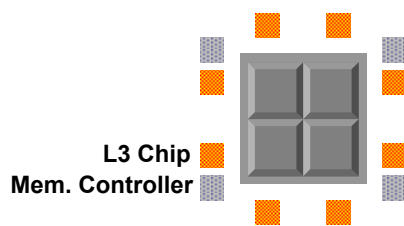
## Multi Chip Module (MCM) Architecture

### 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:

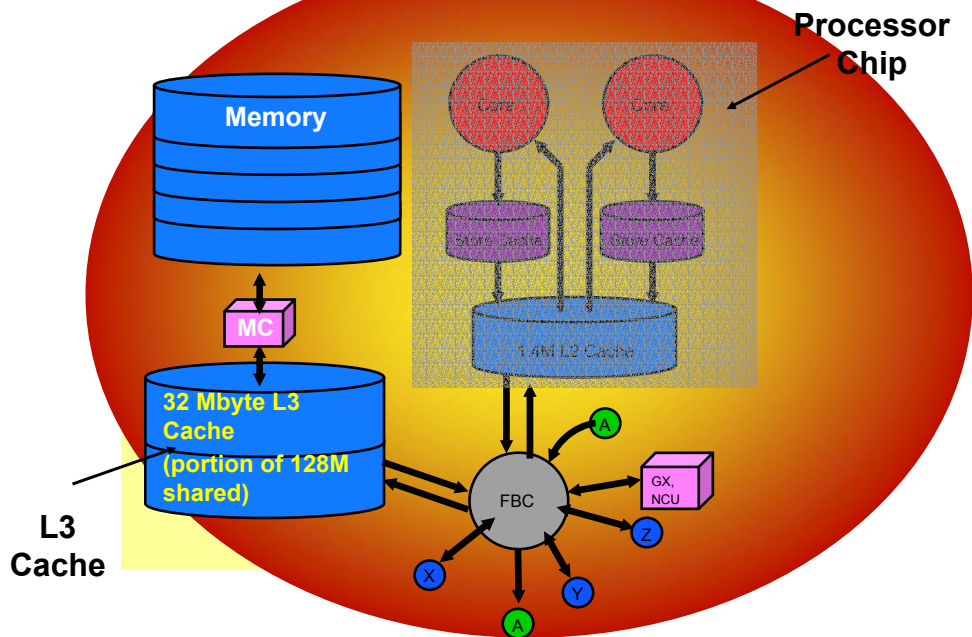
- 4 processor chips
    - 2 processors per chip
  - 4 L3 cache chips
    - L3 cache is used by processor pair
    - "Extension" of L2
- 
- 8 chips



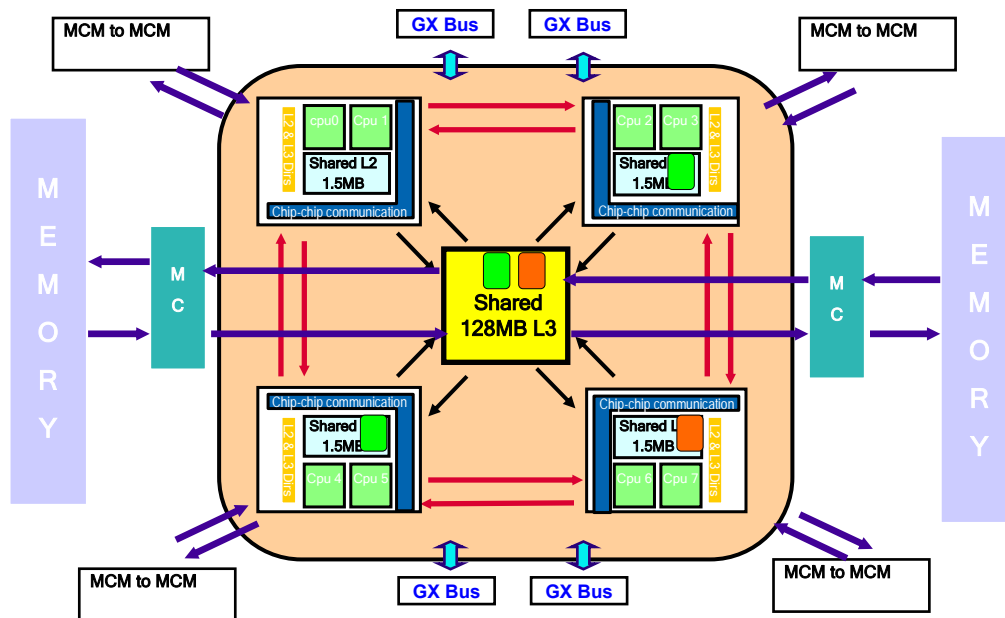
## System Design with POWER4 and with POWER5

- Multi Chip Module (MCM)
  - Multiple MCMs per system node
- Other issues:
  - L3 cache sharing
  - Memory controller

### POWER4

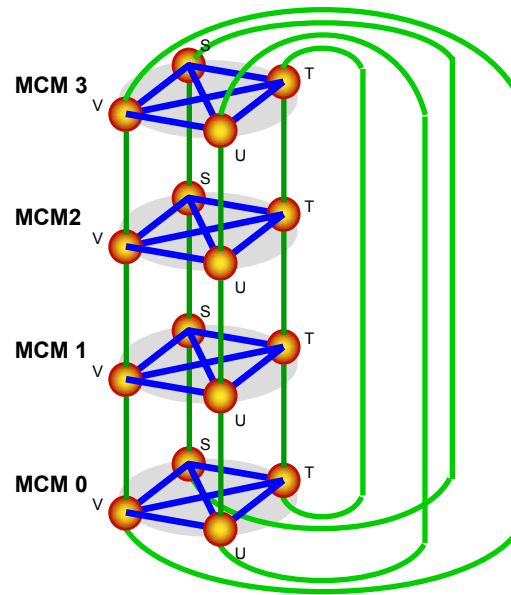


### POWER4 Multi-chip Module

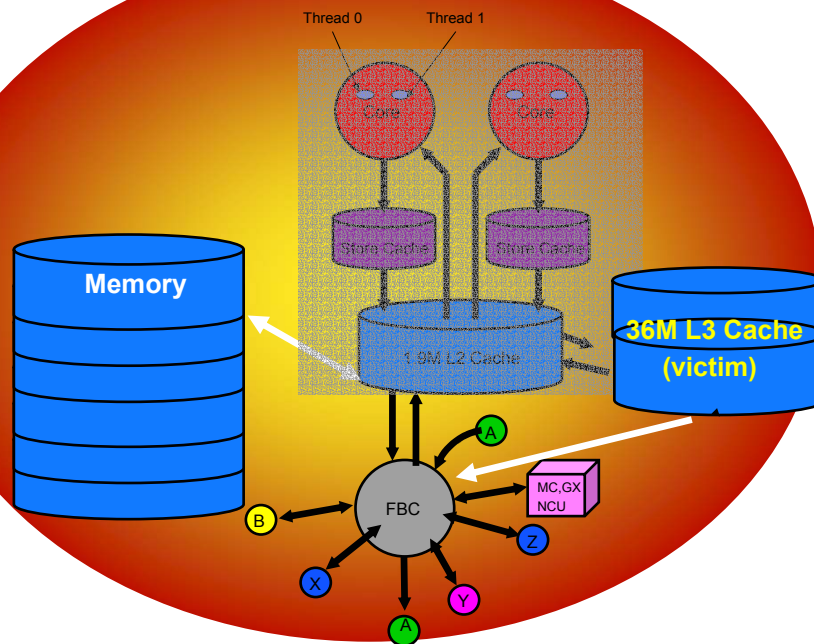


## POWER4 32-way Plane Topology

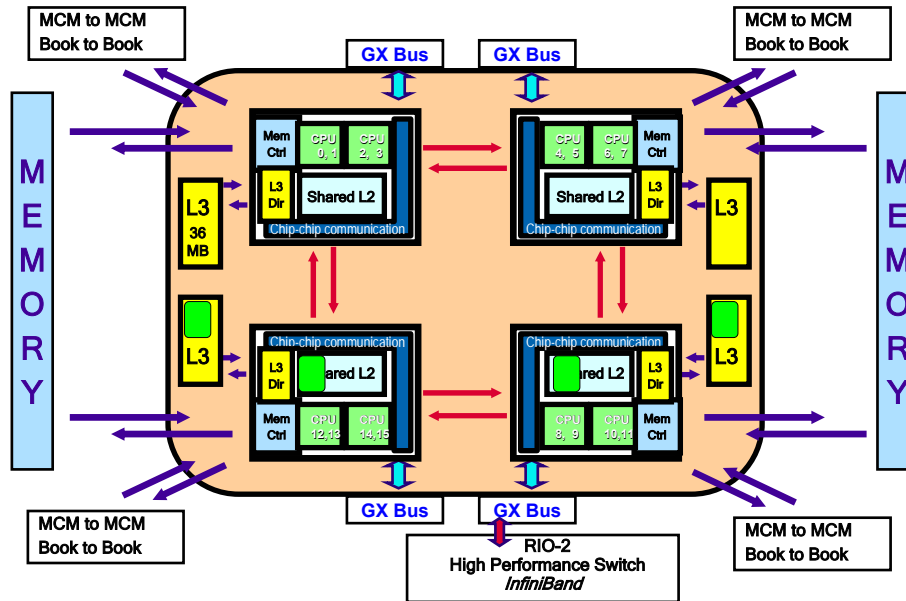
- Uni-directional links
- 2-beat Addr @ 850Mhz = 425M A/s
- 8B Data @ 850Mhz (2:1) = 7 GB/s



## POWER5



## POWER5 Multi Chip Module

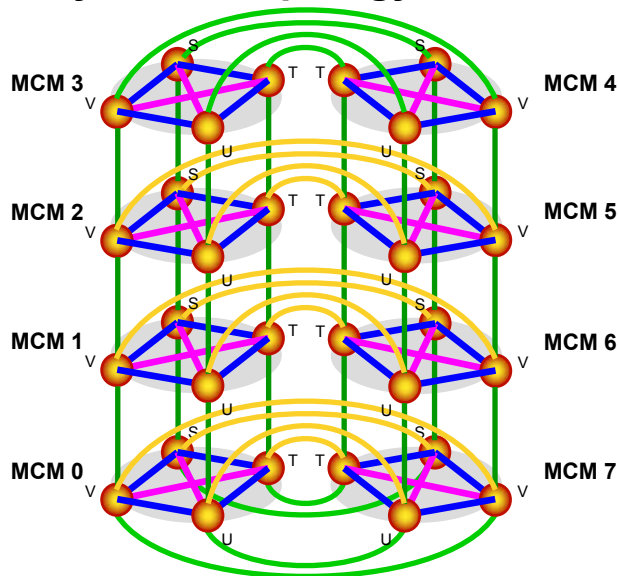


## POWER5 64-way Plane Topology

Vertical node links  
Used to reduce latency  
and increase bandwidth  
for data transfers.

Used as alternate route  
for address operations  
during dynamic ring  
re-configuration (e.g.  
concur repair/upgrade).

Uni-directional links  
2-beat Addr @ 1GHz  
= 500M A/s  
8B Data @ 1 GHz (2:1)  
= 8 GB/s



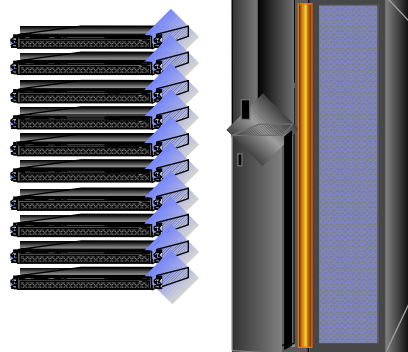
## Upcoming POWER5 Systems

- **Early introductions are iSeries**
  - Commercial
- **Next:**
  - Small nodes
  - 2,4,8 processors
  - SF, ML
- **Next, next**
  - iH
  - H
  - Up to 64 processors

## Planned POWER5 iH



- 2/4/6/8 – way POWER5
- 2U rack Chassis
- Rack: 24" x 43" deep



## Auxiliary

## Switch Technology

- **Internal network**
  - In lieu of, e.g. Gig Ethernet
- **Multiple links per node**
  - Match number of links to number of processors

Generation	Processors
HPS Switch	POWER2
SP Switch	POWER2 → POWER3
SP Switch 2 (Colony)	POWER3 → POWER4
HPS (Federation)	POWER4 → POWER5

## HPS Software

- **MPI-LAPI (PE V4.1)**
  - Uses LAPI as the reliable transport
  - Library uses threads, not signals for async activities
- **Existing applications binary compatible**
- **New performance characteristics**
  - Improved collective communication

## High Performance Switch (HPS)

- **Also Known As “Federation”**
- **Follow on to SP Switch2**
  - Also known as “Colony”
- **Specifications:**
  - 2 Gbyte/s (bidirectional)
  - 5 microsecond latency
- **Configuration:**
  - Four adaptors per node
    - 2 links per adaptor
    - 16 Gbyte/s per node

## HPS Specifications



	Latency [microsec.]	Bandwidth, single [Mbyte/s]	Bandwidth, multiple [Mbyte/s]
SP Switch 2 ("Colony")	15	350	550
HPS ("Federation")	5	1700	200

## HPS Packaging



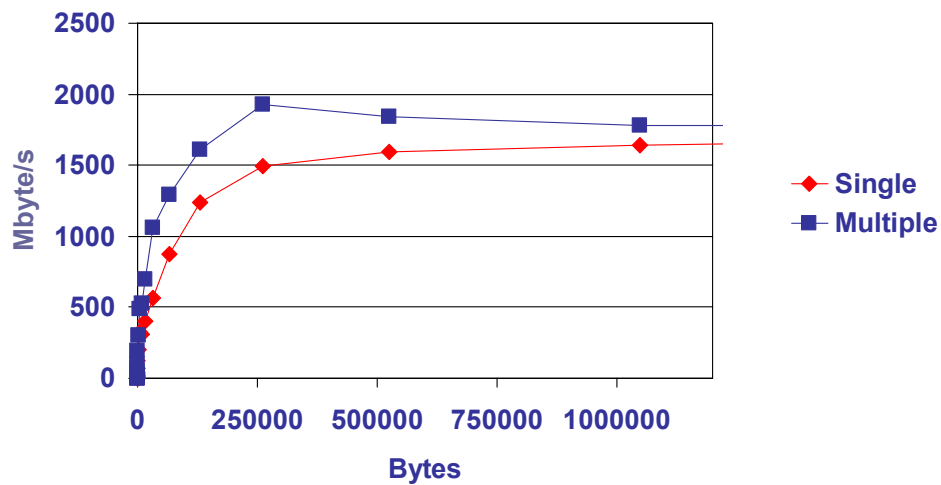
- **4U, 24-inch drawer**
  - Fits in the bottom position of p655 or p690 frame
- **Switch only frame (7040-W42 frame)**
  - Up to eight HPS switches
- **16 ports for server-to-switch**
- **16 ports for switch-to-switch connections**
- **Host attachment directly to server bus via 2-link or 4-link Switch Network Interface (SNI)**
  - Up to two links per pSeries 655
  - Up to eight links per pSeries 690

## HPS Scalability

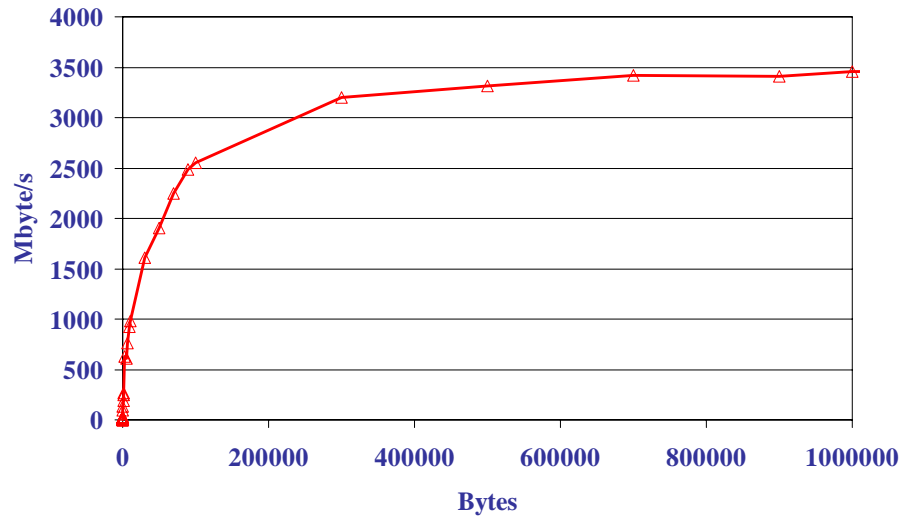


- Supports up to 16 p690 or p655 servers and 32 links at GA
- Increased support for up to 64 servers
  - 32 can be pSeries 690 servers
  - 128 links planned for July, 2004
- Higher scalability limits available by special order

## HPS Performance: Single and Multiple Messages



## HPS Performance: POWER5 Shared Memory MPI



## Summary

- **Processor technology:**
  - POWER4 → POWER5
  - Multiprocessor chips
  - Enhanced caches
  - SMT
- **Cluster systems**
  - P655
  - P690