POWER5 and HPS Programming Strategies: System Architecture

Charles Grassl
IBM
July, 2004

© 2004 IBM

Agenda

- Architecture
  - Processors
    - POWER4
    - POWER5
  - System
    - Nodes
    - Systems
  - Clusters
    - p655
    - p690
- High Performance Switch

© 2004 IBM Corporation

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

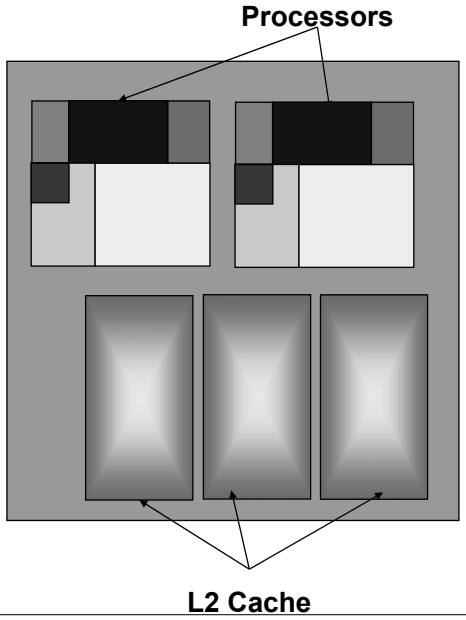
POWER5

- 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

© 2004 IBM Corporation

POWER4: Multi-processor Chip

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



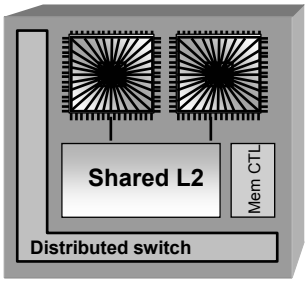
© 2004 IBM Corporation

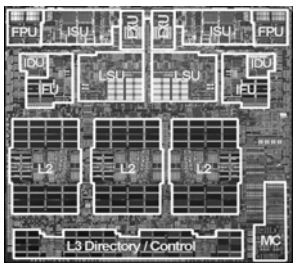
POWER4: Features

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

© 2004 IBM Corporation

POWER5 Chip





- IBM CMOS 130nm
  - Copper and SOI
  - 8 layers of metal
- Chip
  - 389 mm^2
  - 276 M transistors
  - I/Os: 2313 signal, 3057 power
- 2 processors per chip
- 1.65 – 2.0 GHz

© 2004 IBM Corporation

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

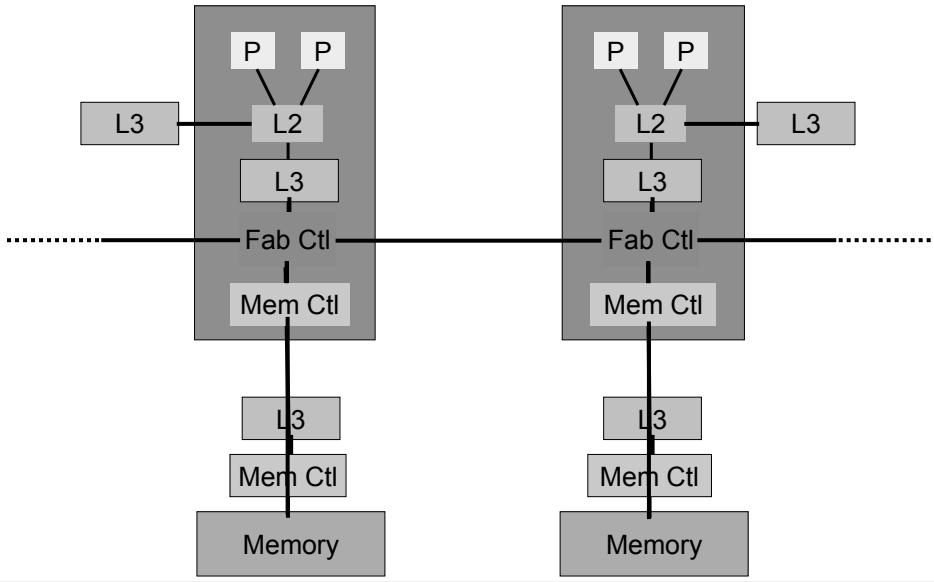
© 2004 IBM Corporation

POWER4, POWER5 Differences

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

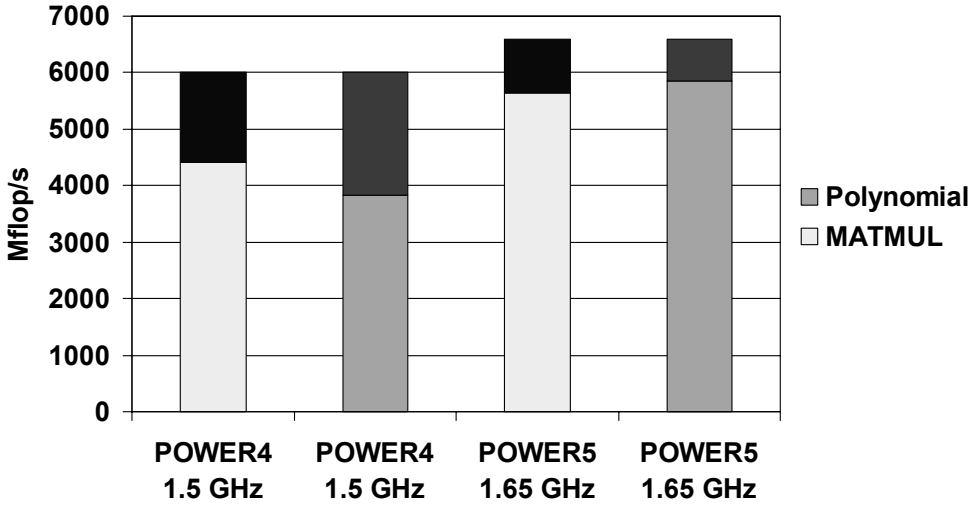
© 2004 IBM Corporation

Modifications to POWER4 System Structure



© 2004 IBM Corporation

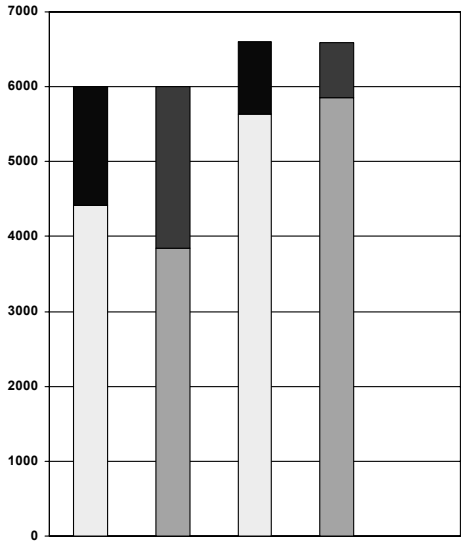
Effect of Rename Registers



© 2004 IBM Corporation

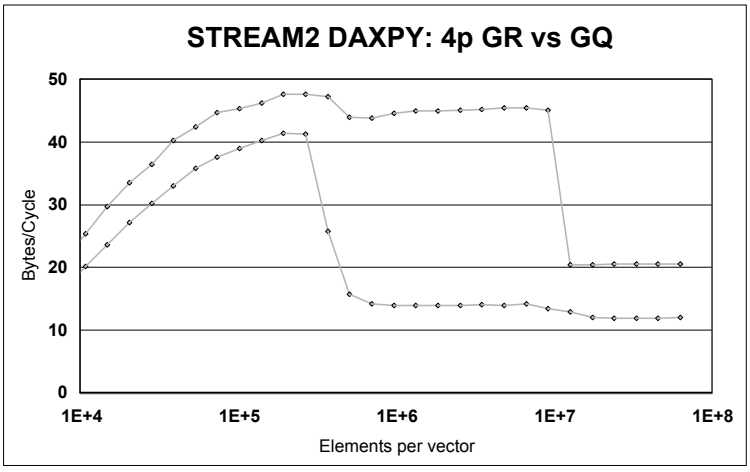
Effect of Rename Registers

- "Tight" algorithms were previously register bound
- Increased floating point efficiency
- Examples
  - Matrix multiply
    - LINPACK
  - Polynomial calculation
    - Horner's Rule



© 2004 IBM Corporation

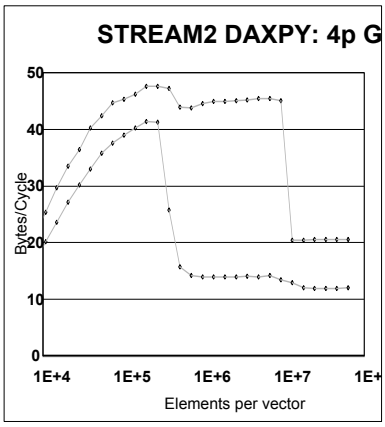
L2 and L3 Cache Improvement



© 2004 IBM Corporation

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



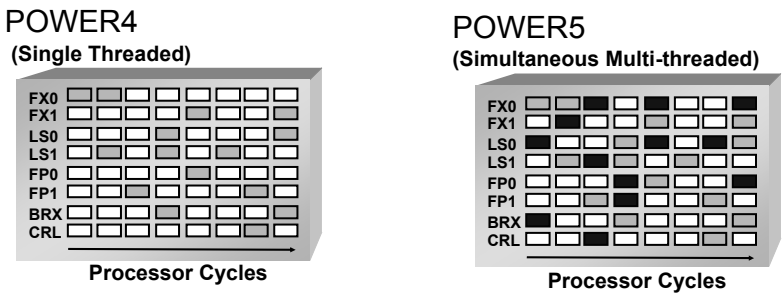
© 2004 IBM Corporation

Simultaneous Multi-Threading

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

© 2004 IBM Corporation

POWER5 Simultaneous Multi-Threading (SMT)



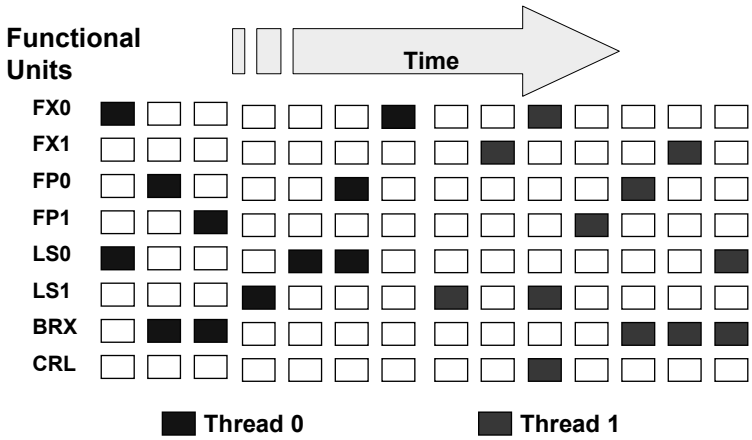
Legend

- Thread0 active
- No Thread active
- Thread1 active

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

© 2004 IBM Corporation

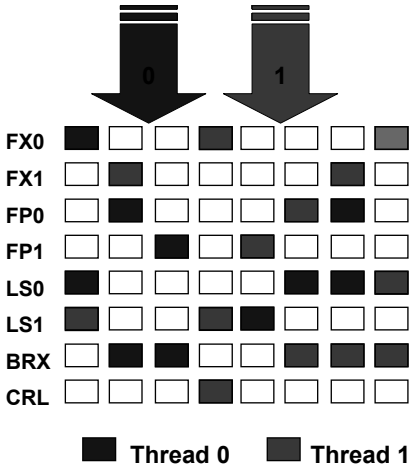
Conventional Multi-Threading



- **Threads alternate**
  - Nothing shared

© 2004 IBM Corporation

Simultaneous Multi-Threading



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

© 2004 IBM Corporation

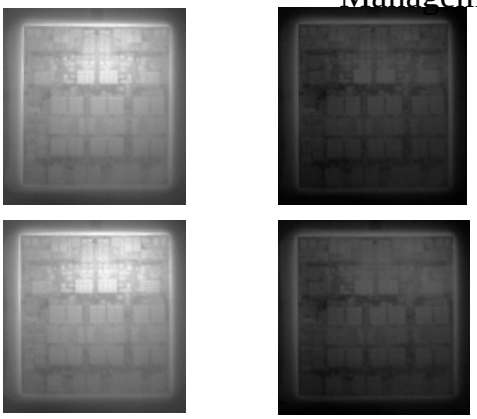
Dynamic Power Management

No Power Management

Dynamic Power Management

Single Thread

Simultaneous Multi-threading



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

© 2004 IBM Corporation

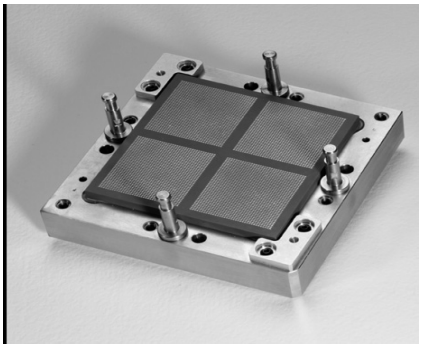
Multi-Chip Modules (MCMs)

- 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

© 2004 IBM Corporation

POWER4 Multi-Chip Module (MCM)

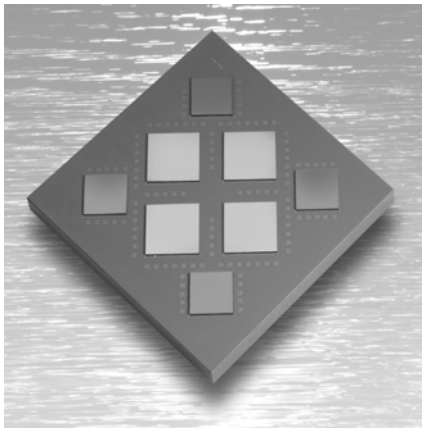
- 4 chips
  - 8 processors
    - 2 processors/chip
- >35 Gbyte/s in each chip-to-chip connection
- 2:1 MHz chip-to-chip busses
- 6 connections per MCM
- Logically shared L2's, L3's
- Data Bus (Memory, inter-module):
- 3:1 MHz MCM to MCM



© 2004 IBM Corporation

POWER5 Multi-Chip Module (MCM)

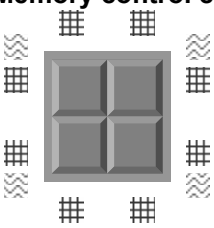
- 95mm × 95mm
- Four POWER5 chips
- Four cache chips
- 4,491 signal I/Os
- 89 layers of metal



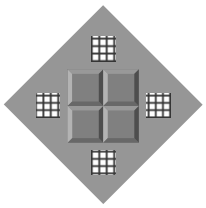
© 2004 IBM Corporation

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



- POWER5:
  - 4 processor chips
    - 2 processors per chip
  - 4 L3 cache chips
    - L3 cache is used by processor pair
    - "Extension" of L2



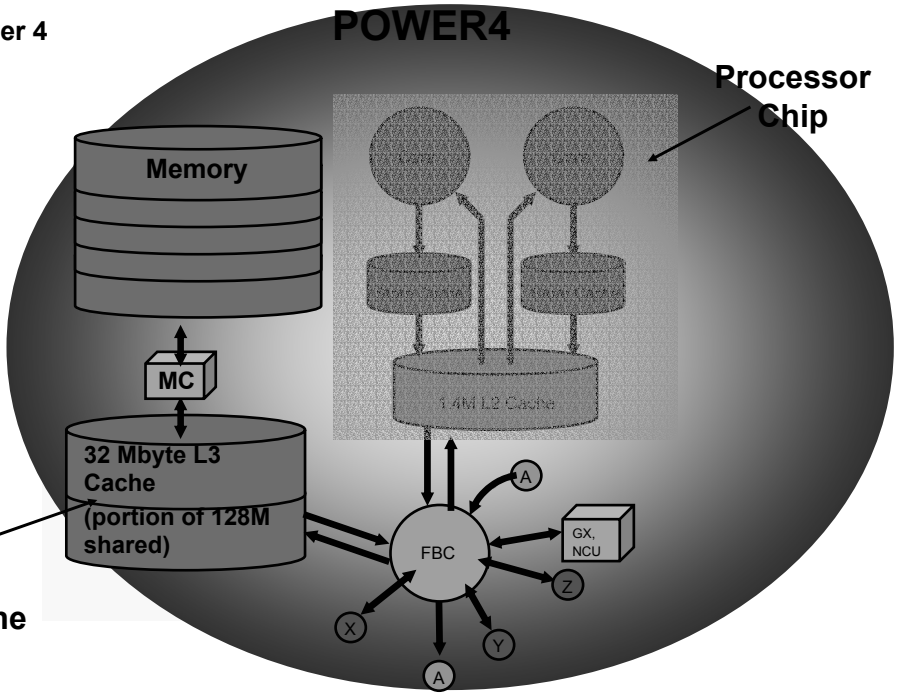
© 2004 IBM Corporation

System Design with POWER4 and with POWER5

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

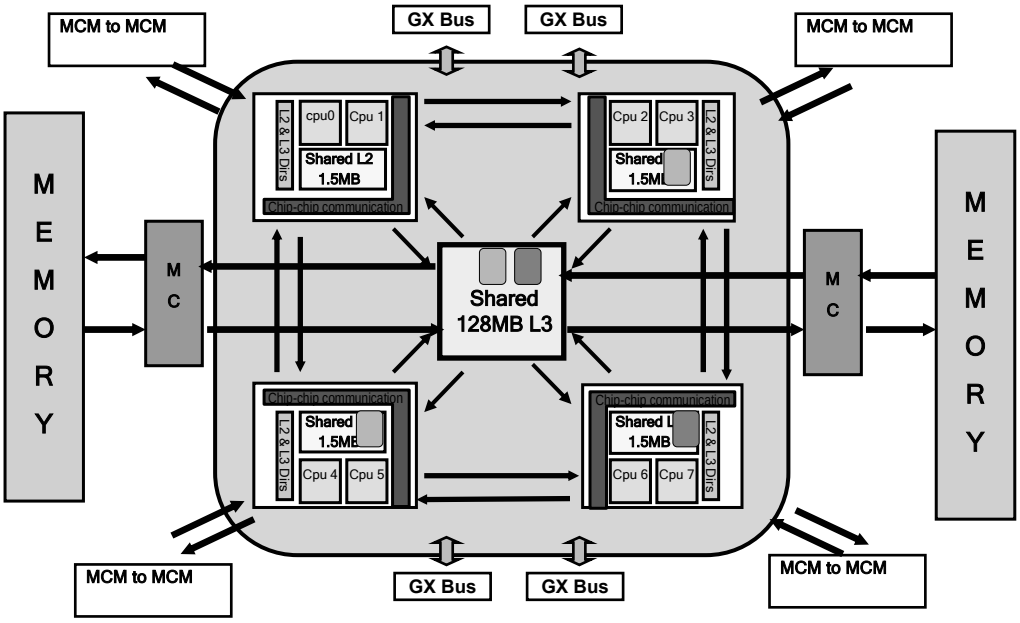
© 2004 IBM Corporation

Power 4



© 2004 IBM Corporation

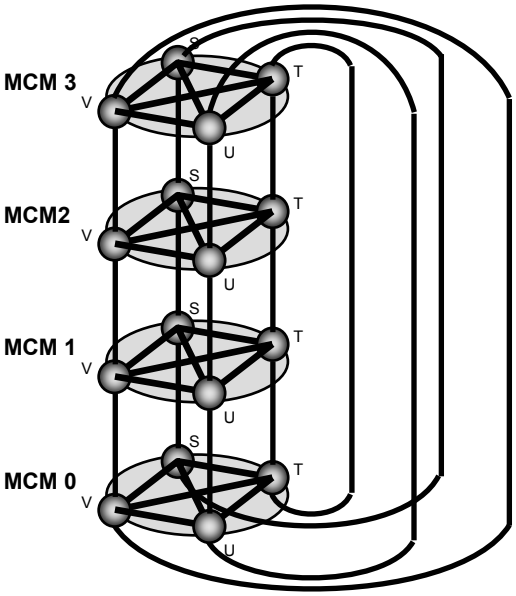
POWER4 Multi-chip Module



© 2004 IBM Corporation

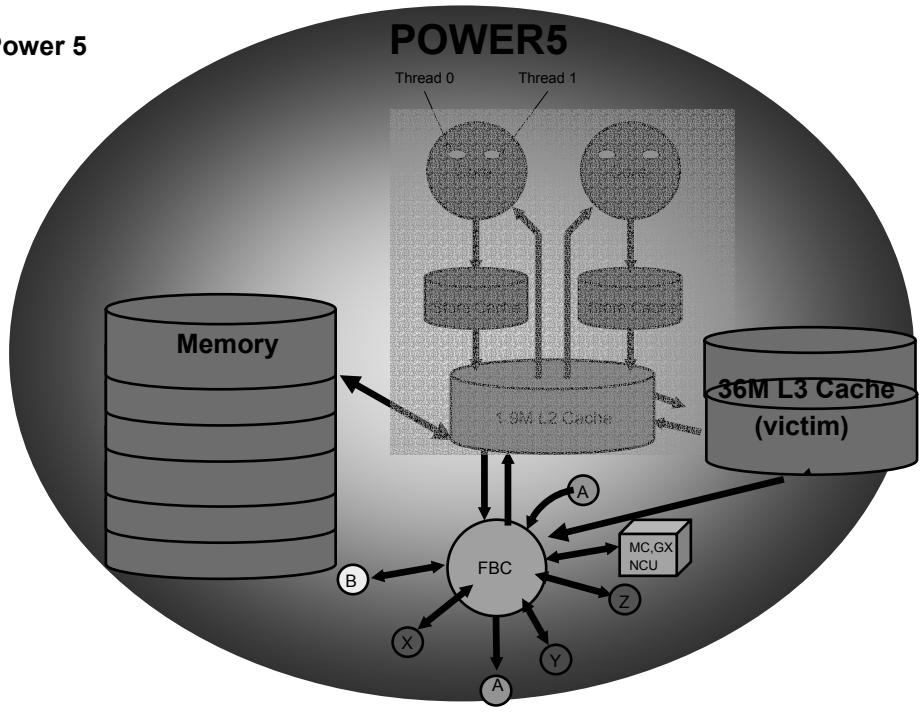
POWER4 32-way Plane Topology

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



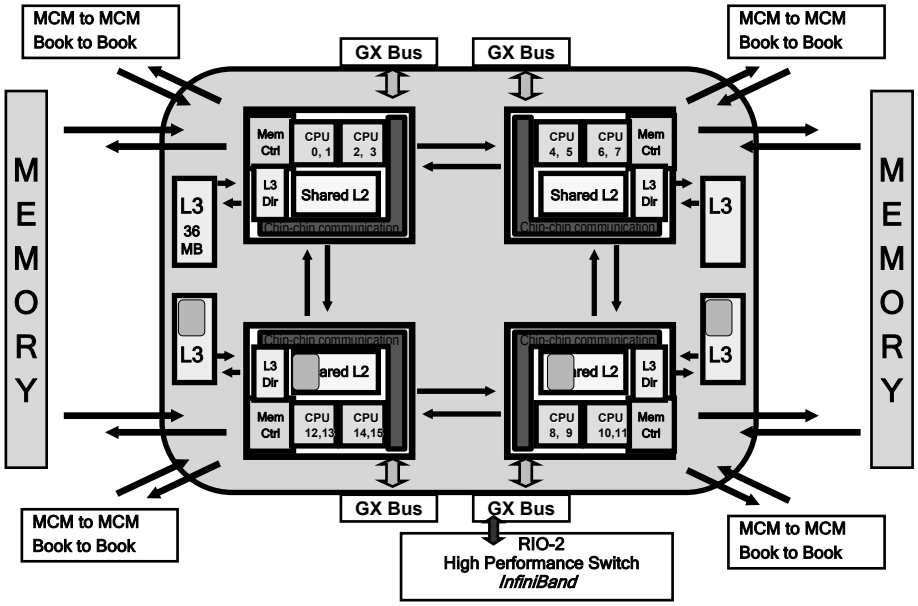
© 2004 IBM Corporation

Power 5



© 2004 IBM Corporation

POWER5 Multi Chip Module



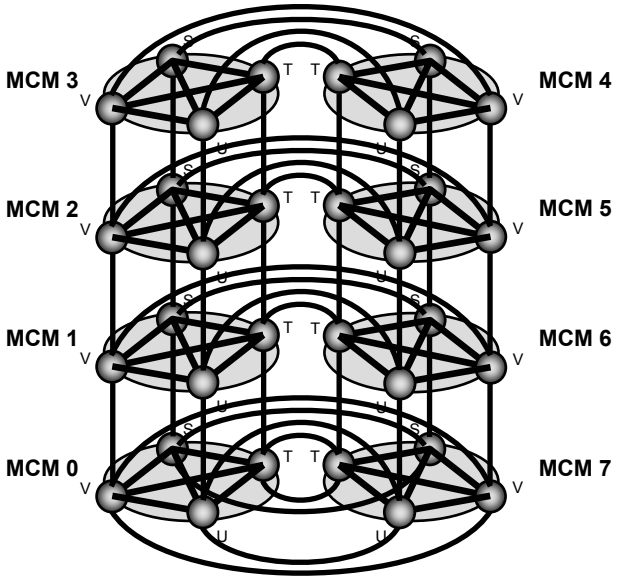
© 2004 IBM Corporation

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



© 2004 IBM Corporation

Cluster Systems

- **Build with nodes**
  - **P655**
    - **8 processors per node**
  - **P690**
    - **32 processors per node**

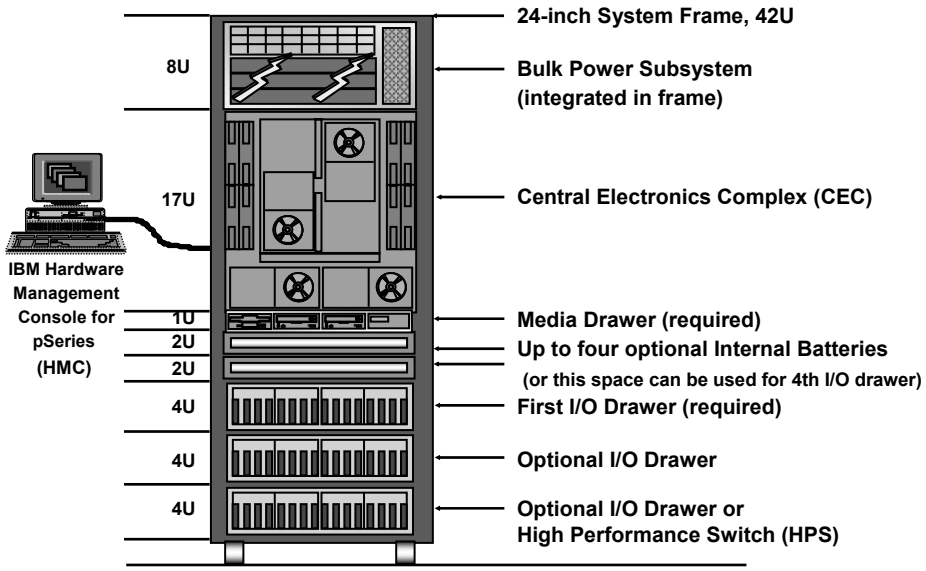
© 2004 IBM Corporation

pSeries 690

- **POWER4 Processors**
  - **Multiple Chip sorts**
    - **POWER4+**
      - **1.7GHz Standard**
    - **POWER4++**
      - **1.9GHz Turbo**
  - **8, 16, 24, 32 -way**
  - **8GB - 1 TB Memory (DDR)**
    - **Memory Cards:**
      - **4, 8, 16, 32, 64GB**

© 2004 IBM Corporation

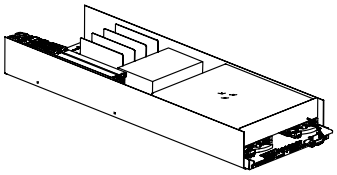
P690 Primary System Frame Organization



© 2004 IBM Corporation

P655 Systems

- **Packaging:**
  - 2 systems per drawer
  - 1-8 drawers per rack
  - Battery backup (optional)
  - RIO I/O (Optional)
  - 11 inch width x 39” depth



- **Processors**
  - POWER4 MCM packaging
  - 8 Standard Processors / System 1.5GHz
  - 4GB to 64GB / System
  - 4 Memory cards / System

© 2004 IBM Corporation

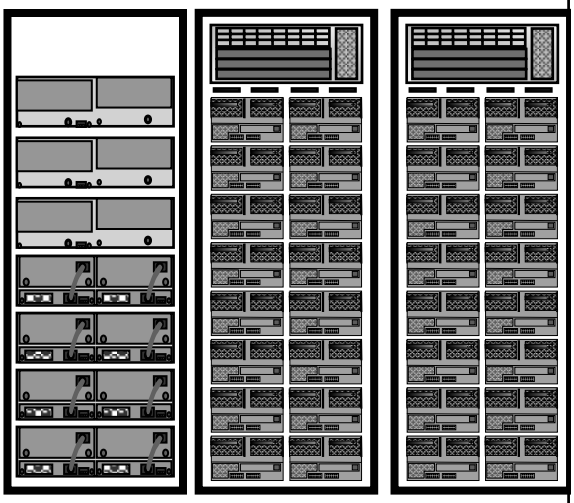
pSeries Multi-Tflop/s System

- **P655 cabinet**
  - 8 processor nodes
    - One Multi Chip Modules (MCM) per node
  - 16 8-processor nodes per cabinet
  - SP Switch2 or HPS interconnect
- **P690 cabinet**
  - 32 processor nodes
    - Four Multi Chip Modules (MCM) per node
  - 1 32-processor node per cabinet
  - SP Switch2 or HPS interconnect

© 2004 IBM Corporation

P655 System

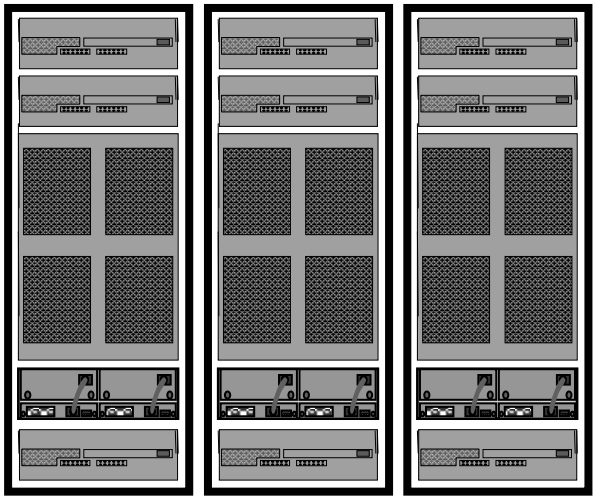
- 8 processors per node
- 16 nodes per cabinet
- Multiple cabinets
  - Connected with HPS (Federation) switch.



© 2004 IBM Corporation

P690 System

- 32 processors per node
- 1 node per cabinet
- Multiple cabinets
  - Connected with HPS (Federation) switch.



© 2004 IBM Corporation

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

© 2004 IBM Corporation

Planned POWER5 iH

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

© 2004 IBM Corporation

POWER5 IH Details

.

Core Electronics
- Power5 Processor
- 4W / 8W SMP
- 144MB / 288MB ECC L3 Cache
- DDR I: 2-32 GB / Processor
- DDRII: 2-16 GB / Processor

Attributes / Features
- 2U, 24" X 43 " Deep, Full Drawer
- W42 Rack w/Dual 350V Bulk Power
- 1 - 16 SQ-IH Nodes / Rack
- 0 - 8 RIO-2 IO Drawers / Rack
- 0 - 2 Switch Drawers / Rack

Integrated Features
- 4X 10/100/1000 Ethernet
- Ultra 3 LVD SCSI Controller

Expansion Slots
- 2 Dual GX+ Bus Slots (Federation & IBT)
- 6 PCI-X 266 MHz, PCI-Express

Storage Bays
- 2 DASD Hot Swap (36.4 / 73.4 / 146.8 GB)

I/O Expansion
- 2 External RIO-G ports
- 0,1/2,1 RIO-2 "p690" Drawer

Software Support
- AIX 5.2H
- Linux
  - Suse SLES9
  - Redhat RHEL3+
- LPAR Partitions: Qty TBD

Cluster/Attach Support
- CSM: 64 CEC's / 128 LPAR's
- HPS support

RAS
- Blackwidow FSP
- Run time processor de-allocation
- ECC / Chipkill memory
- PCI-X bus parity & PCI-X bus slot error recovery
- Hot Swap DASD
- Blindswap PCI-X adapters
- Memory DIMM FRU
- Service Focal Point

© 2004 IBM Corporation

Planned POWER5 H

	POWER5 H
Architecture	16, 32, 48, 64- way POWER5
Packaging	42U Frame
Max Memory	1TB
LPAR	254 LPARs
DASD / Expan. Slots	192 DASD (Hot Plug)
I/O Expansion	12 RIO Drawers)
Virtual Network / Storage	Yes
Max PCI-X slots	240 (Blind Swap)
RIO	PCI-X
Integrated SCSI	Four / IO drawer
Integrated Ethernet	two 10/100 (HMC)
Media Bays	7212 IO Drawer
HMC	Ethernet

© 2004 IBM Corporation

Switch Technology

- Internal network
  - In lieu of, e.g. Gig Ethernet
- Fourth generation
  - HPS Switch (POWER2 generation)
  - SP Switch (POWER2 -> POWER3)
  - SP Switch 2 (POWER3 -> POWER4)
  - HPS (POWER4 -> POWER5)
- Multiple links per node
  - Match number of links to number of processors

 © 2004 IBM Corporation

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

 © 2004 IBM Corporation

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

© 2004 IBM Corporation

HPS Specifications

	Latency [microsec.]	Bandwidth, single [Mbyte/s]	Bandwidth, multiple [Mbyte/s]
SP Switch 2	15	350	550
HPS	5	1700	1930

© 2004 IBM Corporation

HPS Packaging

- 4U, 24-inch drawer that fits in the bottom position of either the p655 frame or the p690 frame
- Up to eight HPS switches can be installed into a switch-only 7040-W42 frame
- 16 ports for server-to-switch and 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

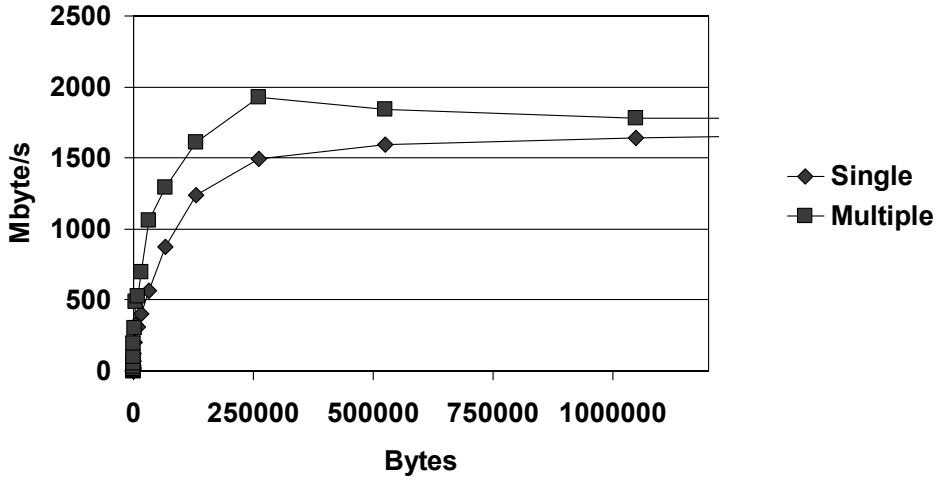
© 2004 IBM Corporation

HPS Scalability

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

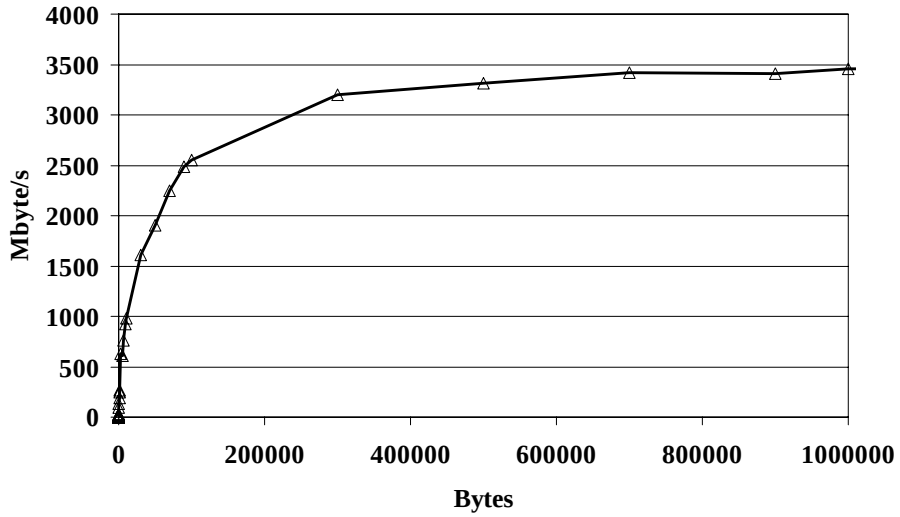
© 2004 IBM Corporation

HPS Performance: Single and Multiple Messages



© 2004 IBM Corporation

HPS Performance: POWER5 Shared Memory MPI



© 2004 IBM Corporation

Summary

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

© 2004 IBM Corporation