

The HPCx building block

Dual-core POWER5 in the p5-575

Jonathan Follows

IBM United Kingdom Limited

jonathan_follows@uk.ibm.com

Requirements

- Floating point performance
- Price/performance
- Power
- Memory bandwidth
- Packaging
- Binary compatibility
- Reliability
- Structural compatibility

Two aspects of the solution

- POWER5
 - Evolution from POWER4
 - SMT
 - Reduced memory and cache latency
 - Increased memory bandwidth
 - p5-575
 - Aggregate memory bandwidth
 - Packaging density
 - RAS

POWER5

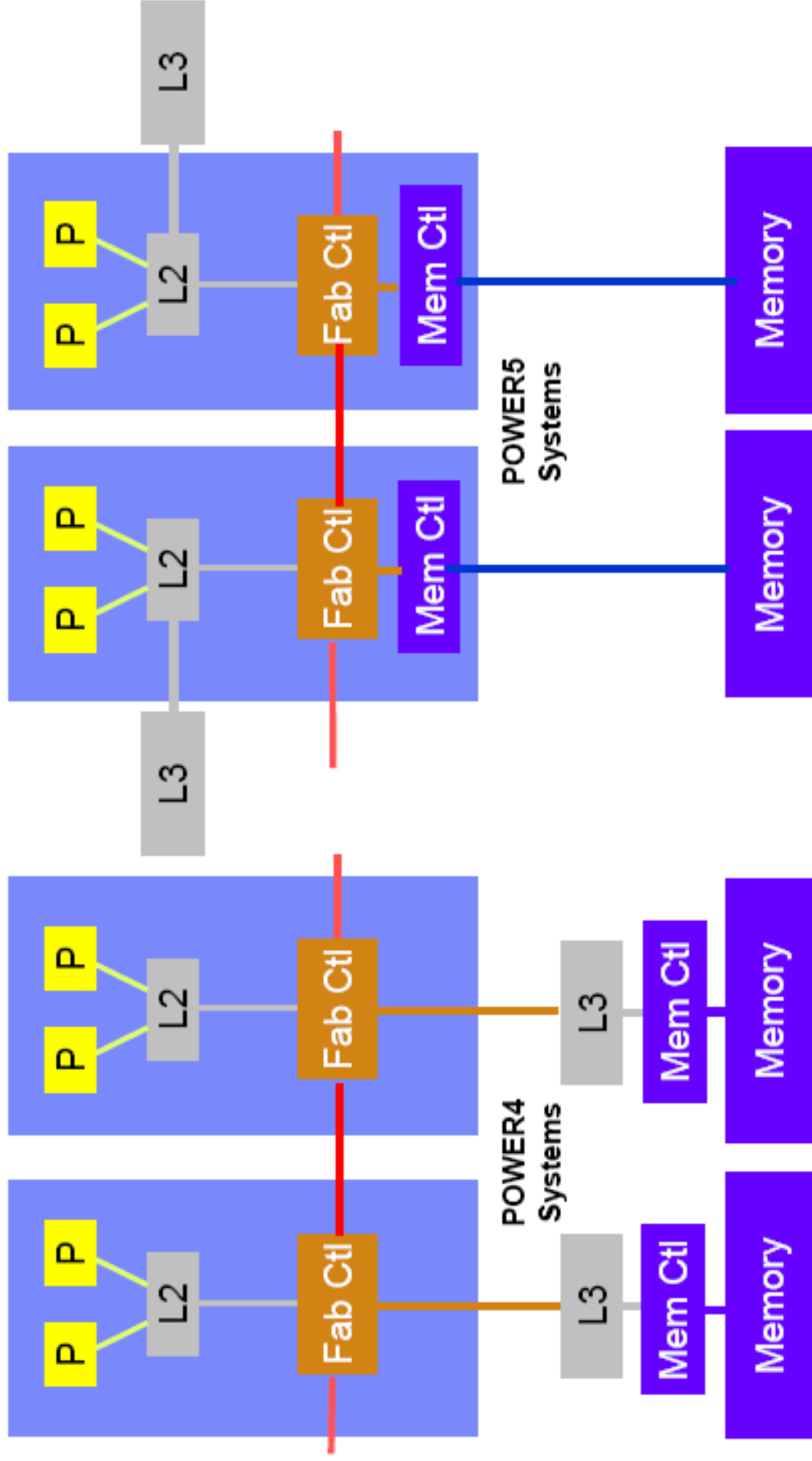
Terminology

- Just to be clear, when IBM introduced the POWER4 processor in 2001, and since then also for POWER5:
 - Dual-core CHIP
 - Single L2 cache
 - One or two PROCESSORS per chip
- Other manufacturers use different terminology, beware of making false comparisons either deliberately or inadvertently

POWER5

- 64-bit PowerPC architecture
- 2 independent FPU per core
- 1.875MB L2 per chip
- Binary compatibility with POWER4 and earlier (PPC601 50MHz even)
- Structural compatibility for code previously optimised for POWER4

POWER4 System Evolution to POWER5 Systems



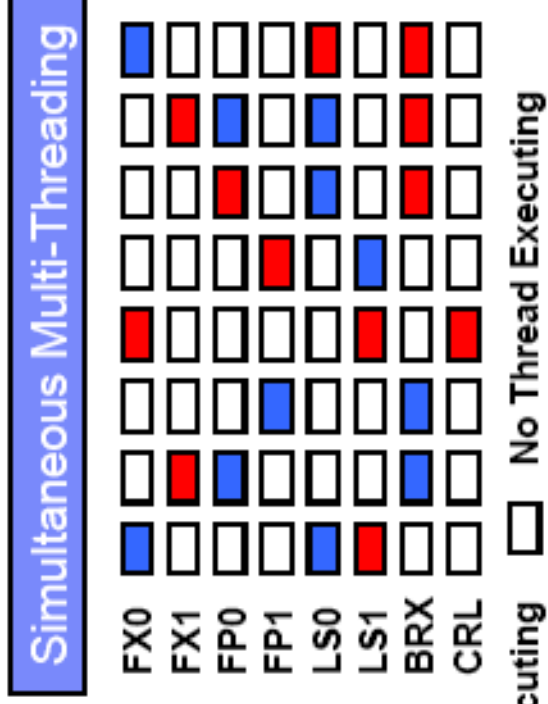
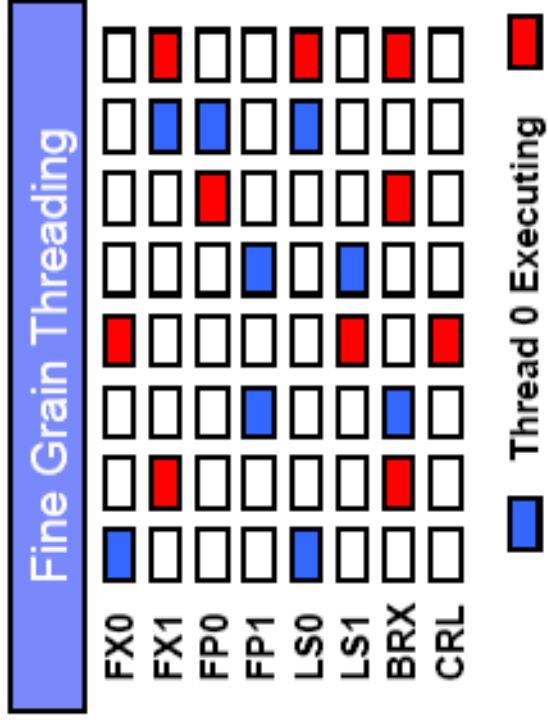
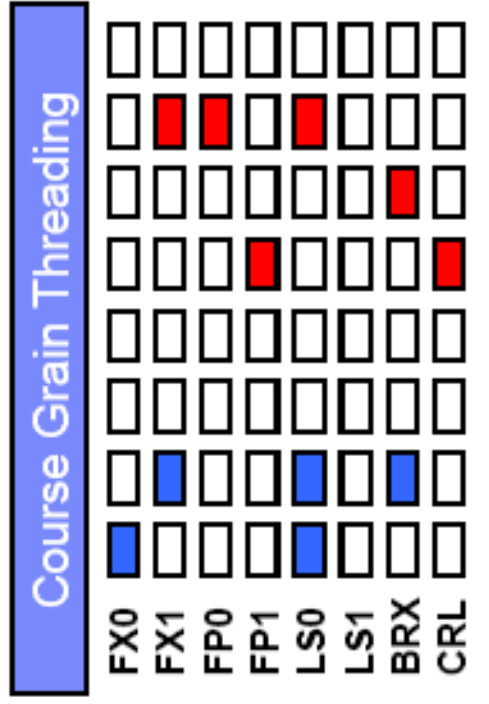
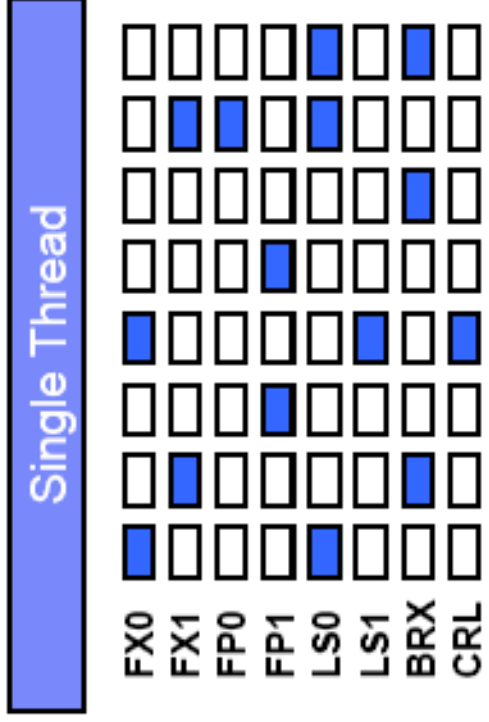
POWER5 processor

- 130nm process, the same as used for POWER4+
- 267mm² -> 389mm²
- 184 million transistors -> 276 million
- +10% size for SMT core
- +11% size for L2 changes
- +7% for L3 directory changes
- POWER5+ is 90nm process (240mm²)

Simultaneous MultiThreading

- Compare with
 - Coarse-grain multithreading
 - One thread executes at a given instant
 - Context switch at long-latency event
 - IBM previous implementation RS64 chip
 - Fine-grain multithreading
 - Switch between threads every cycle
 - Allows overlap of short pipeline latencies

Multi-threading Evolution



█ Thread 0 Executing
 █ Thread 1 Executing
 No Thread Executing

Reduced memory latency

- L3 cache no longer in the path between processor and memory controller
- POWER4 -> POWER5:
 - L3 latency 123 cycles -> 87 cycles
 - Memory latency 351 -> 235
- POWER4 1.7GHz -> POWER5 1.9GHz:
 - SPECint +30%, SPECfp +61%, clock +12%

Increased memory bandwidth

- Implementation depends on number of Synchronous Memory Interface chips
- Each POWER5 chip connects to 2 or 4 SMI chips, behind which up to 4 DIMMs can connect
- 12.7GBps per chip on p5-575, which uses 4 SMI per POWER5 chip

p5-575

System description

- 2-U, 24"x48", up to 12x per rack
 - Up to 16x per rack in future (POWER5+)
- 8-way or 16-way, 8-chip SMP server
- 8xDDR1 slots per chip, 64 slots in total
- Peak memory bandwidth 99.7GBps
- 8-way 1.9GHz, 16-way 1.5GHz POWER5
 - Power \propto frequency³
 - 16-way has lower power consumption
- 16-way 1.9GHz POWER5+ “preview”

Attributes/Features

- ▶ 2U, 24" X 50 " Deep, Full Drawer
- ▶ FC5793 Rack w/Dual 350V Bulk Power
- ▶ Squadrons Hbased 350V Power Subsystem
 - 200 to 480 AC Input
- ▶ 1 - 12 SQ-IH+ Max Configed Nodes per Rack
 - ▶ 96 Single Core up / 192 Dual Core up
 - ▶ 14 Nodes w / <=32GB Memory (512MB DIMM x 64) per Node
 - ▶ 112 Single Core up / 224 Dual Core up
 - ▶ 16 Nodes w / <=16GB Memory (512MB DIMM x 32) per Node
 - ▶ 128 Single Core up / 256 Dual Core up
- 0 - 5 IO Drawers per Rack
- ▶ 0 - 2 Switch Drawers/Rack
- ▶ Squadrons L4+ based Logic Topology
- ▶ LPAR Partitions: 10/Processor
 - 80/8Way, 160/16W

Core Electronics

- Power5+ SMP
- ▶ 8W 2.2GHz GS -Trimaran DCM (10Ke)
- ▶ 16W 1.9GHz GS - Trimaran DCM (10Ke)
- ▶ 144 MB ECC L3 Cache (Trimaran)
- ▶ 4 SMI-II Memory Bridges / DCM
- 8 DIMM (2-32 GB) / DCM DDR II

Integrated Features

- ▶ 4X 10/100/1000 Ethernet (Duval)
- ▶ Gemstone Dual Port Ultra 3 LVD SCSI Controller
- ▶ Virtual SES

Storage Bays

- ▶ 2 DASD Hot Swap (36.4 / 73.4 / 146.8 / 300 GB)
- ▶ 2 Buses of 1 DASD

Standard Expansion Slots

- ▶ 2 Dual GX+ Bus Adapter Slots

Featured Expansion Slots

- ▶ 4 fullsize PCI-X 133 MHz, 64b, Blindsw ap Slots
- ▶ 2 External RIO-G ports on Base Planner

I/O Expansion

- 0, 1/2, 1 Bonnie&Clyde-XG IO Drawer
- ▶ 0, 1/2, 1, 2 Bonnie&Clyde-IB IO Drawer
- ▶ 0 - 4 Pearl - SCSIIO Drawer

Supported Gx Adapters

- ▶ 1x Dual Port Cu Federation Adapter (Sulu)
- ▶ 2x Dual Port Optical Federation Adapter (TBD)
- ▶ 2x Dual Port Galaxy IBT 12X (TaIIisto)
- ▶ 2x Dual Port Galaxy IBT 4X (Tellisto)

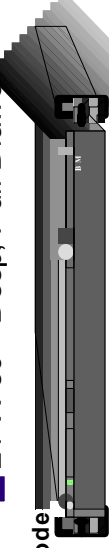
SQ-IH+ w/Mars

(Processor FRU GA - 04/28/06)

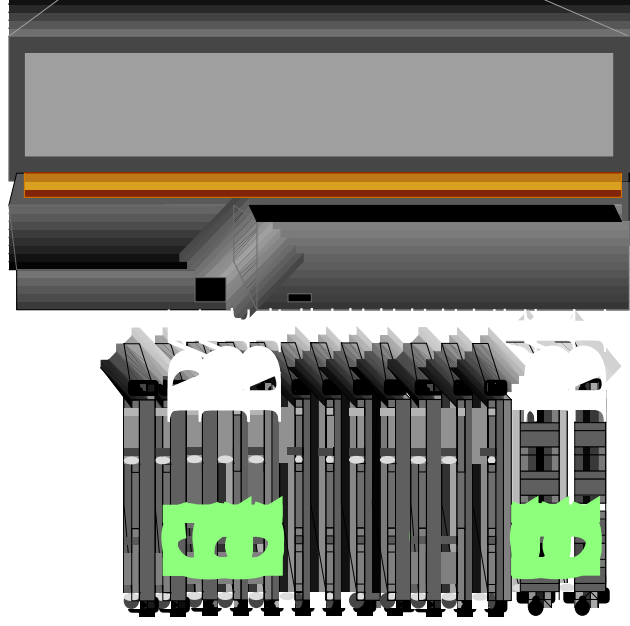
Node Form Factor

- 2U rack chassis

▶ 24" X 50 " Deep, Full Drawer



POWER5+ IH System



OS Support

- ▶ AIX 5.2M / AIX 5.3E
- ▶ Linux-Suse SLES9 SP3,
- ▶ Linux - Redhat RHEL 4 U2

Linux Cluster

- ▶ CSM 1.3.5: 64 CEC's / 128 LPAR's
- ▶ GPFS 2.3 GigEthernet
- ▶ LoadLeveler 3.2
- ▶ ESSL 4.2 / PESSL 3.1.1 Myrinet

AIX Clusters

- ▶ AIX
- ▶ CSM 1.3.5: 64 CEC's / 128 LPAR's
- ▶ GPFS 2.3
- ▶ LoadLeveler 3.2
- ▶ PE 4.1, ESSL 4.2/PESSL 3.1.1

AIX HPS Cluster Support

- ▶ CSM 1.3.6
- ▶ 4GB FED 24 " Switch
- ▶ 2-Link Sulu Adapter
- ▶ 2-Link Sulu Optical Adapter

CECs	IPARs	Sw Links	CPU's/Cluster
128	128	256	1024

RAS

- ▶ Blackwidow FSP
- ▶ Run time processor de-allocation
- ▶ ECC / Chipkill memory
- ▶ PCI-X bus parity & PCI-2.0 bus slot error recover
- ▶ Hot Swap DASD
- ▶ Blindswap PCI-2.0 adapters
- ▶ Memory DIMM FRU
- ▶ Service Focal Point

Certifications

- ▶ FCC Class "A"
- ▶ Acoustical Product Category 1A (Unattended)
- ▶ Environmental Class B Extended
 - 32C Max Ambient up to 4.25K Ft
 - 24C Max Ambient @ 10K Ft



Performance	8W @ 2.2GHz DDR2@ 533	16W @ 1.9GHz DDR2@ 533
SPECfp-rate	386	519
SPECint-rate	184	296
LINPACK HPC	59.9	103.5
TPCC	529.2	745.0
Peak Gflops	70.4	121.6
Memory BW	204.7 GB/S	204.7 GB/S

IBM Confidential

08/02/2005

HPCx

- <http://www.hpcx.ac.uk>
- Old 50x32-way 1.7GHz POWER4+
 - 6.188Tflop/s Linpack; 10.88Tflop/s “peak”
- New 96x16-way 1.5GHz POWER5
 - 7.395Tflop/s Linpack; 9.216Tflop/s “peak”
 - 12-rack footprint rather than 50
- pSeries High Performance Switch
- AIX 5.3 operating system (formerly AIX 5.2)

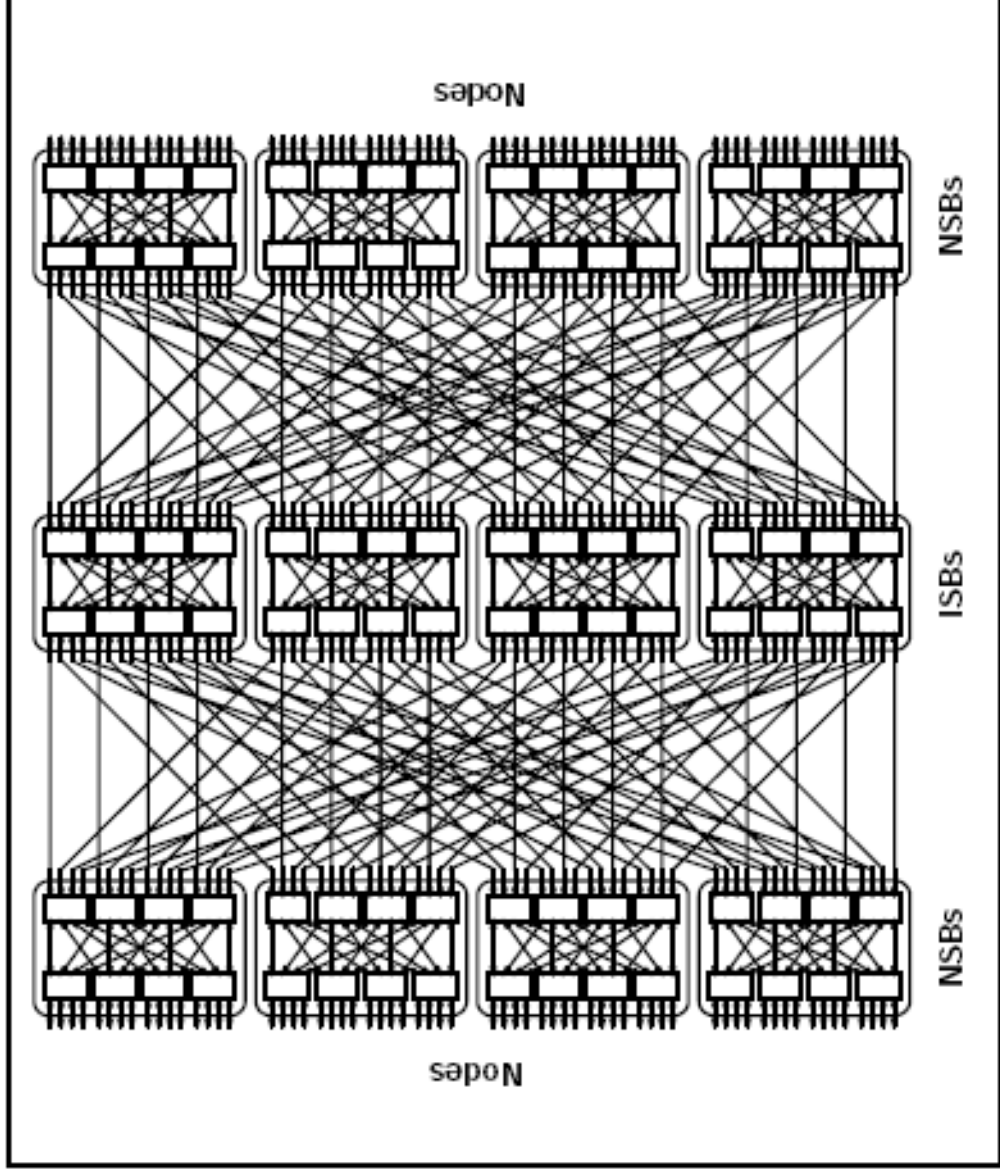
System Description

- Login node – used for compilation and job submission
- TSM node – used for backup to tape
 - Also standby/failover for login node
- VSD nodes – drive the parallel file system
- Compute nodes: 96x16-way 1.5GHz POWER5 with 32GB of memory
- Rack contains 8 nodes and “Federation” switch

Pseries High Performance Switch

- Dual-network configuration
- Each p5-575 server has two links (sn0 and sn1)
- Each link connects to a different Node Switch Board (NSB) in a different switch network
- Worst case – 3 switch hops
- RDMA for large packet transfers

Switch topology: Two identical independent switch networks



Switch usage

- MPI using IP or “user space” protocol
- LAPI using IP or “user space” protocol
- GPFS using IP (invisible to users)
- MPI will use shared memory by default instead of the switch if it can
- $<5\mu\text{s}$ switch latency, single link bandwidth up to $\sim 2\text{Gbps}$

Parallel File System

- GPFS – General Parallel File System version 2.3
(was GPFS 2.2)
- Includes user home directories
- Plus “scratch” file system (not backed up, don't use /tmp)
- “Striped” across multiple RAID controllers -> FC disk
- Aggregate performance ~2GBps

Some figures

- 1.152Tflop/s “theoretical peak”/rack for 1.5GHz,
1.459Tflop/s for 1.9GHz
- STREAM triad “standard” (untuned)
 - 42,632MBps 16-way
 - 41,585MBps 8-way
- Linpack HPL
 - 56.78Gflop/s 8-way 1.9GHz
 - 87.3Gflop/s 16-way 1.5GHz
 - 7.395Tflop/s 96x16-way (80% of “peak”)

Acceptance test results

- Single MPI task bandwidth 5.3GBps across two switch links
- 512-task all-to-all in 50% to 75% of the time of the Phase2 system
- 4.76 microsecond latency in production environment
- 16-way STREAM TRIAD 33% greater than 32-way p690 (so 32-way 166% greater)

Future plans

- Upgrade to 12Tflop/s Linpack in 2006
- Enable the use of SMT on POWER5
- User-initiated RDMA, now available, to be incorporated into Global Arrays, performance boost for NWCHEM and GAMESS-UK
- Next release of compilers “imminent”
- New release of ESSL in 2006 – more tuning on standard libraries