IBM High Performance Computing

Neueste Entwicklungen bei IBM im High Performance Computing Umfeld

Dr. Achim Bömelburg

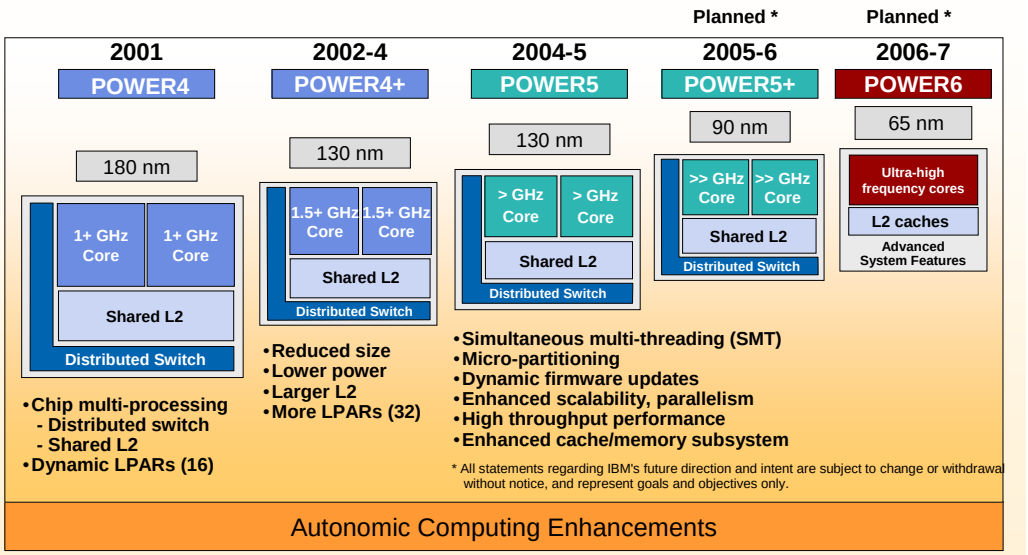
IBM Stuttgart
Systems & Technology Group

ON DEMAND BUSINESS™

Bamberg | 14.-15.10.2004

Processors

IBM POWER technology roadmap for pSeries

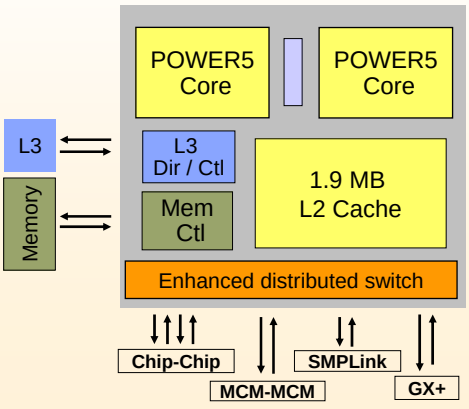


POWER5 architecture

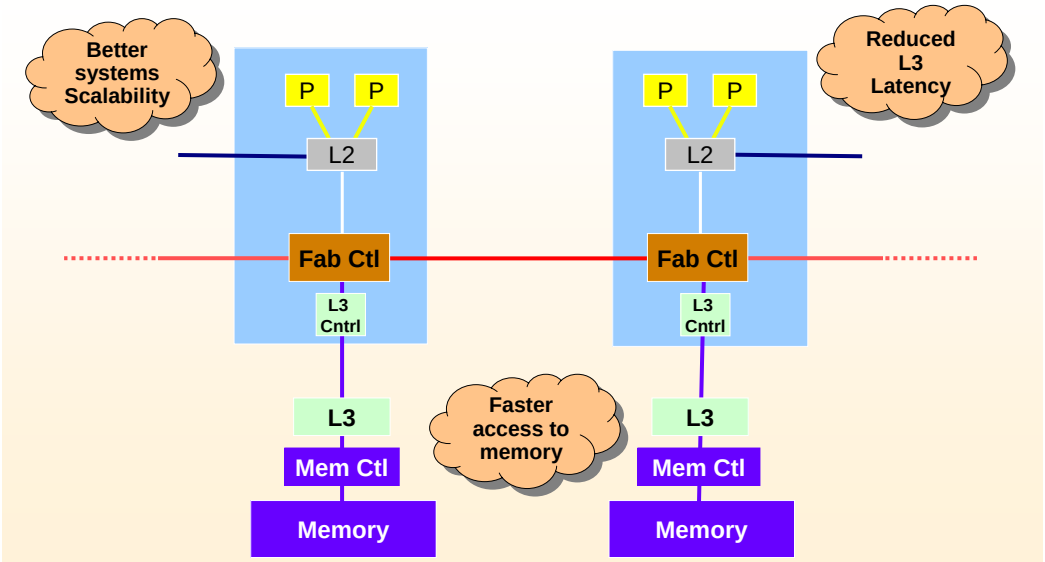
POWER5 design	1.5, 1.65 and 1.9 GHz	276M transistors	.13 micron

POWER5 enhancements

- Simultaneous multi-threading
- Hardware support for Micro-Partitioning
  Sub-processor allocation
- Enhanced distributed switch
- Enhanced memory subsystem
  Larger L3 cache: 36MB
  Memory controller on-chip
- Improved High Performance Computing
- Dynamic power saving

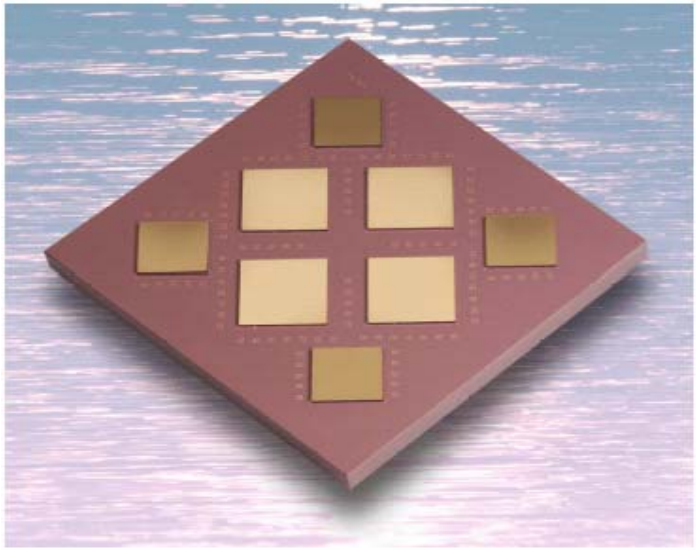


Modifications to POWER4 to create POWER5



5 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

POWER5 MCM

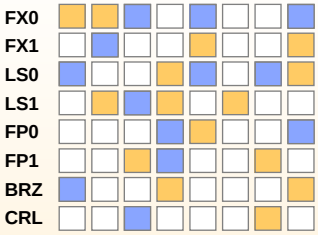


6 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

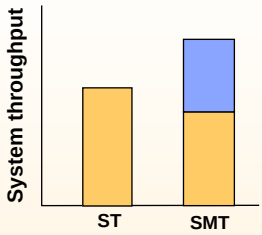
IBM eServer p5 and pSeries

IBM ~ p5: Simultaneous multi-threading

POWER5 (simultaneous multi-threading)



Appears as 4 CPUs per chip to the operating system (AIX 5L V5.3 and Linux)



- Utilizes unused execution unit cycles
- Presents symmetric multiprocessing (SMP) programming model to software
- Natural fit with superscalar out-of-order execution core
- Dispatch two threads per processor: *"It's like doubling the number of processors."*
- Net result:
  - ***Better performance***
  - ***Better processor utilization***

POWER4 / POWER5 Unterschiede

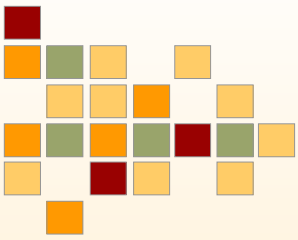
	POWER4 Design	POWER5 Design	Benefit
L1 Cache	2-way associative FIFO	4-way associative LRU	Improved L1 cache performance
L2 cache	8-way associative 1.44MB	10-way associative 1.9MB	Fewer L2 cache misses Better performance
L3 Cache	32MB 8-way associative 118 clock cycles	36MB 12-way associative ~80 clock cycles	Better cache performance 40% improvement
Memory Bandwidth	4GB / sec / chip	~16GB / sec / chip	4X improvement Faster memory access
Simultaneous Multi-Threading	No	Yes	Better processor utilization 40% System improvement
Processor Addressing	1 processor	1/10 of processor	Better usage of processor resources
Chip Interconnect Type Intra MCM data bus Inter MCM data bus	Distributed switch ½ Proc. speed ½ Proc. speed	Enhanced dist. switch Processor speed ½ Proc. speed	Better systems throughput Better performance
Size	412mm	389mm	50% more transistors in the same space

POWER4 und POWER5

- Simultaneous Multithreading
- Increased floating point rename registers
  - → DGEMM ~ 95% of peak performance, LINPACK ~ 90% of peak perf.
- Doubled store buffers
- Memory controller: reduce cost and latency
- Memory bandwidth and latency improved

  upto 3x better sustainable bandwidth than POWER4

  latency reduced upto 50% in MCM systems



IBM ~ p5 hardware

IBM ~ p5 520

Entry-level system	▪ 2-way systems @ 1.65 GHz ▪ 4U rack mount or desk-side system
Functions supported	▪ Dynamic LPAR ▪ IBM Advanced POWER™ Virtualization option Micro-Partitioning support (1/10th processor granularity) Virtual networking and storage support Partition Load Manager
Features	▪ **Up to 32 GB memory** ▪ **6 PCI-X slots** ▪ **Service processor** ▪ **Dual 10/100/1000** ▪ **USB: 2 HMC: 2** ▪ **Up to 8 (4+4) disk drive bays** ▪ **Up to 4 RIO-2 drawers** ▪ **Redundant cooling and optional redundant power**
Software support	▪ AIX 5L V5.2 and AIX 5L V5.3 ▪ Red Hat Enterprise Linux AS 3 for POWER (RHEL AS 3) ▪ SUSE LINUX Enterprise Server 9 for POWER (SLES 9)

11 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

IBM ~ p5 550

Entry-level 4-way system	▪ 2-way to 4-way systems @ 1.65 GHz ▪ 4U rack mount or desk-side system
Functions supported	▪ Dynamic LPAR ▪ IBM Advanced POWER Virtualization option Micro-Partitioning support (1/10th processor granularity) Virtual networking and storage support Partition Load Manager ▪ CoD options
Features	▪ **Up to 64 GB memory** ▪ **5 PCI-X slots** ▪ **Service processor** ▪ **Dual 10/100/1000** ▪ **USB: 2 HMC: 2** ▪ **Up to 8 (4+4) disk drive bays** ▪ **Up to 8 RIO-2 drawers** ▪ **Redundant cooling and optional redundant power**
Software support	▪ AIX 5L V5.2 and AIX 5L V5.3 ▪ RHEL AS 3 ▪ SLES 9

12 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

IBM ~ p5 570: *"Pay as you grow" modular architecture*

New POWER5 mid-range system	▪ 2-way, 4-way, 8-way, 12-way and 16-way systems ▪ Processor speeds: 1.65 GHz and 1.9 GHz ▪ Dynamic LPAR
Functions supported/ base system	▪ IBM Advanced POWER Virtualization option Micro-Partitioning support (1/10th processor granularity) Virtual networking and storage support Partition Load Manager ▪ CoD options
Features/ primary module	▪ **Up to 128 GB memory** ▪ **6 PCI-X slots** ▪ **Service Processor** ▪ **Dual 10/100/1000** ▪ **USB: 2 HMC: 2 (max per system)** ▪ **Up to 6 (3+3) disk drive bays** ▪ **Up to 8 RIO-2 drawers** ▪ **Redundant cooling and power**
Software support	▪ AIX 5L V5.2 and AIX 5L V5.3 ▪ RHEL AS 3 ▪ SLES 9
Modules	▪ Primary + 3 additional

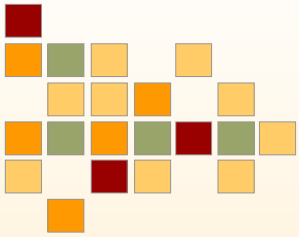
13 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

IBM ^ OpenPower Family

Linux only POWER servers

	Planned HV1	Planned HV2	Model 720 (9124-720)
Form Factor	1U Rack	2U Rack	4U Rack Deskside
# of Processors	1 / 2	1 / 2	1 / 2 / 4
Processor	POWER5	POWER5	POWER5
Memory	512MB – 16GB	512MB – 32GB	1GB – 64GB
PCI-X Slots	2	3	6
Integrated	Dual 10/100/1000	Dual 10/100/1000	Dual 10/100/1000
Disk Bays	2	4	8
USB / Serial	3 / 2	2 / 2	2 / 2
CuD	No	Yes	Yes
LPAR	Yes	Yes	Yes
OS	redhat suse	redhat suse	redhat suse

14 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS



Performance

P520, 1.65 GHz POWER5

	p520,1.65 GHz 2-way POWER5	p630,1.45 GHz 4-way POWER4
SPECint_base2000	1201	884
SPECint_rate_base2000	30.3	35.8
SPECfp_base2000	2034	1097
SPECfp_rate_base2000	41.5	38.1

P550, 1.65 GHz POWER5

	p550,1.65 GHz 4-way POWER5	p650,1.45 GHz 8-way POWER4
SPECint_base2000	1200	909
SPECint_rate_base2000	60.4	72.7
SPECfp_base2000	2121	1221
SPECfp_rate_base2000	82.1	79.7

P570, 1.9 GHz POWER5

	p570,1.9 GHz 16-way POWER5	p655,1.7 GHz 8-way POWER4
SPECint_base2000	1398	1064
SPECint_rate_base2000	74.4 (4x) 141 (8x) 273 (16x)	47.7 (4x) 83.5 (8x)
SPECfp_base2000	2576	1642
SPECfp_rate_base2000	125 (4x) 241 (8x) 438 (16x)	66.5 (4x) 103 (8x)

ISV Code: ANSYS

1.9 GHz p5-570 Performance: Test Machine Specifications

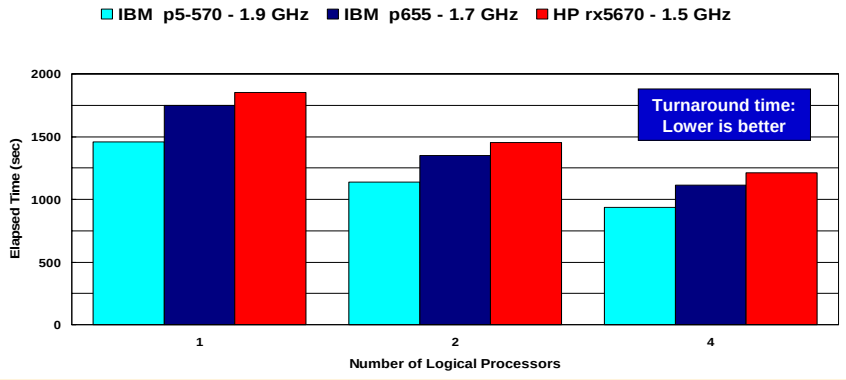
- 16 1.9 GHz POWER5™ processors
- Memory configurations and control
  - 64GB RAM
    - 533 MHz DDR2 memory
  - Memory Affinity enabled
  - 32 to 48GB allocated to Large Technical Page
  - Remaining memory used with 4KB pages
- Simultaneous multi-threading
  - Used for selected tests
- System software
  - AIX 5L™ V5.3
    - Includes C run time environment
  - XLF V 9.1 Run Time Environment
  - PE V4.1.1

ANSYS V7.1
Sum of 12 standard ANSYS runs
(Elapsed Time in sec)

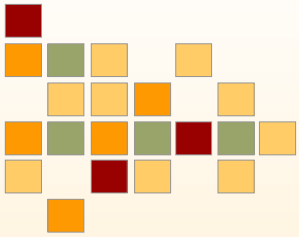
	Installed		Logical Processors		
Platforms	Memory GB	CPUs	1	2	4
IBM p5-570 1.9 GHz POWER5	64	16	1459	1137	935
IBM p655 1.7GHz POWER4+™	16	4	1750	1348	1111
HP rx5670 1.5 GHz Itanium 2	24	4	1851	1454	1212

- Simultaneous multi-threading used on the p5-570 where appropriate.
- IBM data current as of 7/13/04. Other data current as of 6/28/04.
- Source: http://www.ansys.com/services/hardware_support/index.htm select "Hardware Support Database", then benchmarks.

ANSYS V7.1
Sum of 12 standard ANSYS runs
(Elapsed Time in sec)

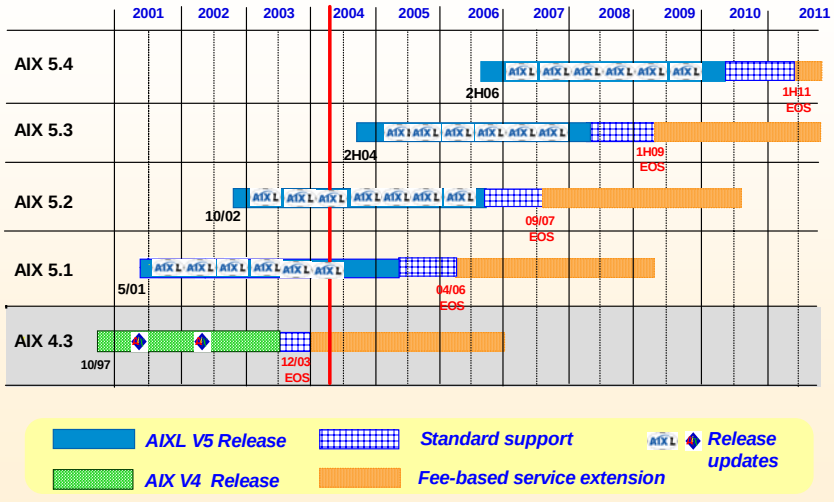


- Simultaneous multi-threading used on the p5-570 where appropriate.
- IBM data current as of 7/13/04. Other data current as of 6/28/04.
- Source: http://www.ansys.com/services/hardware_support/index.htm select "Hardware Support Database", then benchmarks.



Software

AIX Release Plan 2001-2011



AIX System Unterstützung

- AIX 5.1
  Unterstützt alle derzeit verfügbaren und in Wartung befindlichen pSeries Systeme
  AIX5.1 ist die Mindestvoraussetzung für POWER4 basierte Systeme
  Keine Unterstützung für zukünftige POWER5 basierte Systeme
- AIX5.2
  Unterstützt alle POWER basierten Systeme, die von AIX 5.1 unterstützt werden mit Ausnahme von Microchannel und CHRP/PReP Systemen.
  Neu: Unterstützung von POWER Blades und POWER5 Systemen (ohne SMT) über neuen Maintenance Level
- AIX5.3
  Unterstützung von allen POWER basierten Systemen, die von AIX 5.2 unterstützt werden
  Unterstützung von POWER Blades und allen POWER5 Systemen, incl. SMT Unterstützung
  Erweiterte Skalierbarkeit durch Unterstützung von 64-wege POWER5 Systemen
  Zusätzliche Funktionen und Einhaltung von zusätzlichen Software Standards

AIX 5L Version 5.2 Überblick (aktuelle AIX Version)

- Unterstützung von 32-bit und 64-bit Anwendungen
  32-bit Binär Kompatibilität für alle AIX 4 und AIX 5L Versionen
  64-bit Binär Kompatibilität für alle AIX 5L Versionen
- Erweiterte Skalierbarkeit, Bedienungsfreundlichkeit, Sicherheit
  32-way SMP, **1TB memory, Dynamic LPAR/CUoD**
  Autonomic computing support, self-managing features
  High perf. Journaling Filesystem **(JFS2 -16TB capacity), Native MPIO**
  AIX Workload Mgr, IBM LDAP Directory, Kerberos Auth. server
  Linux interoperability and AIX Toolbox for Linux Applications
  Integrated SVR4 Affinity services
  Formal security certification (Common Criteria CAPP/EAL4+)
- Zusätzliche optionale Software
  HACMP Version 5.1 for system and application failover
  Cluster Systems Manager (CSM) for AIX and Linux
  Grid Toolkit, ...

AIX 5L 5.3 Erweiterungen

- 64-way SMP/SMT — Höchste Skalierbarkeit
- Enterprise Distributed Storage — Daten Zugriff, Sicherheit, Kapazität
  - JFS2 - quotas, shrinkfs, >16TB spt
  - Large storage objects (fs, lvm, devices)
  - NFSV4 - strong security & performance
- Networking and I/O updates — Connectivity und Performance
  - 4X Infiniband adapter and protocol support
  - TCP/IP offload engine support
  - IP over FC, iSCSI
- Affinity/Application Enabling — Erweiterte Komaptibilität
  - POSIX Real Time
  - JAVA 1.4.2 support
  - PAM Extensions, SVR4 Loader/Linker opts.
- System Management, Usability, Stds — Verbessertes Management
  - HA & Secure NIM Server
  - GNOME as a supported desktop
  - UNIX03 compliance

27 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

AIX 5L support for IBM ~ BladeCenter™ JS20

AIX 5L V5.2 for JS20 blades*

- Planned Availability: August 20, 2004
  Linux currently available

JS20 blade

- Processor: POWERPC 970
  - 2-way SMP: 1.6 GHz
- Memory: Up to 4GB
- DASD: Dual drive support
- Networking: Dual 10/100/1000 MBps Ethernet
- Fibre: Optional

28 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

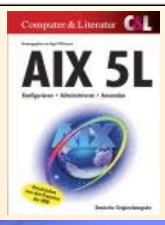
Linux on POWER5

SUSE LINUX	▪ SLES 9 ▪ 64-bit kernel enablement ▪ POWER5 support ▪ Dynamic LPAR
Red Hat	▪ RHEL AS 3 ▪ 64-bit kernel enablement ▪ POWER5 support
IBM Advanced POWER Virtualization Option	▪ Micro-Partitioning ▪ Virtual networking, I/O

p5-520, p5-550, p5-570

29 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

More information ...



30 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

xSeries

Intel announcement on Feb 17th

- Intel will introduce 64bit addressing extensions to their IA-32 (x86) processors
- This is based on a “secret” internal project codenamed Yamhill and then later Clackamass Technologies (CT)
- Intel is referring to these extensions as EM64T (was IA-32e)
- These extensions are compatible with the AMD64 Opteron instruction set
- EM64T, IA-32e, AMD64, x86-64, Yamhill ... all refer to 64 bit extensions to the current x86 32-bit instruction set and are basically the same thing.
- Intel plans to introduce EM64T processors throughout 2004 and 2005 – first one will be Nocona in Summer 04

Xeon with 64 bit Extensions and Opteron Comparison

Intel Xeon with 64-bit Extensions

- Convert 32-bit to 64-bit pointers & registers
- New registers
  - 8 general purpose registers, 16 total
  - Adds 8 new SSE registers
- New instructions
  - Adds 3 new instructions
  - Extends 9 more instructions
- 48-bit virtual address space
  - Nocona (2Q04) 36-bits physical addressing
- Unique attributes
  - CPUID = Genuine Intel
  - SSE3

AMD Opteron

- 64-bit pointers & registers
- 16 64-bit general purpose registers
- 16 128-bit SSE/SSE2 registers
- 48-bit virtual address space
  - 40-bit physical address space
- Unique attributes
  - CPUID = Authentic AMD
  - 3DNow Technology

33 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

Memory Technology

- **Ideally systems should have the memory bandwidth to match the processor bus speed**
  - 800MHz FSB wants 800MHz memory
  - DDR2 memory is lower power, higher density than DDR,higher performance
  - More pins in the socket (240 vs 184), Lower voltage (1.8v vs 2.5v)
  - But more complex to manufacture – means higher cost
- **Complexities in design mean interleaving is common**
  - Most servers are 2 way interleaved
- **The faster the memory bus, the fewer chips you can put on it**
  - DDR 1 400MHz limited to 2 DIMMs per channel, DDR2 to 4 DIMMs per channel
- **Solution is to put buffers on the DIMMs**
  - These increase latency, and add cost
  - Currently most server modules are registered
  - For the 1067MHz memory bus, expect to see Fully Buffered DIMMs

34 IBM HPC Update| 14.-15.10.2004 ON DEMAND BUSINESS

I/O Busses

- **PCI Express**
  - Serial IO architecture designed by Intel
  - 16x will replace AGP on desktops
  - Will get used for high speed devices (10G Ethernet and Fibre)
  - Is the main interface for connecting Intel's chipset components together
  - Three speeds of connector – 4x, 8x, 16x
- **PCI-X 2.0**
  - Evolution of PCI-X
  - 1.5 and 3.3v signalling (no 5v support)
  - In 2005 will provide 266MHz slots – with option for 533MHz (1.5v only)
  - Not implemented in the Intel chipsets for Nocona
- **Infiniband**
  - Probably now just an external interconnect - for clustering

Specifications Compared to Predecessor

x335		x336
Up to 2 Prestonia 533MHz processors	→	Up to 2 Nocona 800MHz processors
4 DIMM slots - 8GB Max DDR Memory	→	8 DIMM slots - 16GB Max DDR-II Memory
Up to 2 HS SCSI HDD (3.5 inch)	→	Choice of 4x 2.5 inch or 2x 3.5 inch HS SCSI
Up to 2 fixed IDE HDD	→	Up to 2 SATA HDD
Single channel Ultra 320 SCSI controller	→	
RAID 1 integrated	→	RAID 0, 1 and 1E, Optional RAID 5
Dual Gigabit Ethernet	→	
Two 100 MHz PCI-X slots	→	Two slots - 133MHz, opt. PCI-Express x8
Internal Light Path Diagnostics panel	→	Drop down Light Path Diagnostic panel
C2T Interconnect	→	Remote KVM over IP via optional Peregrine
Hawk Integrated System Management	→	Vulture Integrated System Management
Integrated CD and FDD	→	No internal FDD. Optional internal DVD
No redundant/hot swap power supplies	→	Redundant/hot swap power supplies
No redundant/hot swap fans	→	Redundant/hot swap fans

Specifications Compared to Predecessor

x345	x346
Up to 2 Prestonia 533MHz processors	Up to 2 Nocona 800 MHz processors
4 DIMM slots - 8GB Max Memory	8 DIMM slots - 16 GB Max Memory
Up to 6 HS SCSI HDD (3.5 inch)	
No Internal Tape Option	Optional DDS5 Internal Tape
Dual channel Ultra 320 SCSI controller	
RAID 0 and 1 integrated, Optional RAID 5	ROMB
Dual Gigabit Ethernet	
Five PCI-X slots	Four PCI-X, Optional PCI-E Riser Card
Light Path Diagnostics	Light Path Diagnostics/Drop Down Panel
Act Cabling	
Hawk Integrated System Management	Vulture Integrated System Management
Optional RSA 1/RSA II	RSA II Daughter Card, no slot used
Integrated CD and FDD	
Redundant/hot swap power supplies	
Redundant/hot swap fans	

Thank You!