Neueste Entwicklungen bei IBM im High Performance Computing Umfeld

Dr. Achim Bömelburg
IBM Stuttgart
Systems & Technology Group
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

---

Autonomic Computing Enhancements

- Simultaneous multi-threading (SMT)
- Micro-partitioning
- Dynamic firmware updates
- Enhanced scalability, parallelism
- High throughput performance
- Enhanced cache/memory subsystem

* All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
Modifications to POWER4 to create POWER5

- Better systems scalability
- Reduced L3 latency
- Faster access to memory

POWER5 MCM
IBM ~ p5: Simultaneous multi-threading

### POWER5 (simultaneous multi-threading)

- 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

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

System throughput

- ST: Single Thread
- SMT: Simultaneous Multi-Threading

### POWER4 / POWER5 Unterschiede

<table>
<thead>
<tr>
<th></th>
<th>POWER4 Design</th>
<th>POWER5 Design</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Cache</td>
<td>2-way associative</td>
<td>4-way associative</td>
<td>Improved L1 cache performance</td>
</tr>
<tr>
<td>L2 cache</td>
<td>8-way associative</td>
<td>10-way associative</td>
<td>Fewer L2 cache misses Better performance</td>
</tr>
<tr>
<td>L3 Cache</td>
<td>32MB</td>
<td>36MB</td>
<td>Better cache performance 40% improvement</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>4GB / sec / chip</td>
<td>~16GB / sec / chip</td>
<td>4X improvement Faster memory access</td>
</tr>
<tr>
<td>Simultaneous Multi-Threading</td>
<td>No</td>
<td>Yes</td>
<td>Better processor utilization 40% System improvement</td>
</tr>
<tr>
<td>Processor Addressing</td>
<td>1 processor</td>
<td>1/10 of processor</td>
<td>Better usage of processor resources</td>
</tr>
<tr>
<td>Chip Interconnect Type</td>
<td>Distributed switch</td>
<td>Enhanced dist. switch Processor speed</td>
<td>Better systems throughput Better performance</td>
</tr>
<tr>
<td></td>
<td>½ Proc. speed</td>
<td>½ Proc. speed Processor speed</td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>412mm</td>
<td>389mm</td>
<td>50% more transistors in the same space</td>
</tr>
</tbody>
</table>

Distributed switch

Enhanced dist. switch Processor speed

Faster memory access

POWER4 Design

POWER5 Design

Benefit

Improved L1 cache performance

Fewer L2 cache misses Better performance

Better cache performance 40% improvement

4X improvement Faster memory access

Better processor utilization 40% System improvement

Better usage of processor resources
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 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

**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)

---

**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

**Software support**
- AIX 5L V5.2 and AIX 5L V5.3
- RHEL AS 3
- SLES 9

---

© 2004 Copyright by DYNAmore GmbH
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)

**Software support**
- AIX 5L V5.2 and AIX 5L V5.3
- SLES 9
- RHEL AS 3

**Modules**
- Primary + 3 additional

---

**IBM ~ OpenPower Family**

Linux only POWER servers

<table>
<thead>
<tr>
<th>Form Factor</th>
<th>Planned HV1</th>
<th>Planned HV2</th>
<th>Model 720 (9124-720)</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Processors</td>
<td>1 U Rack</td>
<td>2U Rack</td>
<td>4U Rack Deskside</td>
</tr>
<tr>
<td>Processor</td>
<td>POWER5</td>
<td>POWER5</td>
<td>POWER5</td>
</tr>
<tr>
<td>Memory</td>
<td>512MB – 16GB</td>
<td>512MB – 32GB</td>
<td>1GB – 64GB</td>
</tr>
<tr>
<td>PCI-X Slots</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Integrated</td>
<td>Dual 10/100/1000</td>
<td>Dual 10/100/1000</td>
<td>Dual 10/100/1000</td>
</tr>
<tr>
<td>Disk Bays</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>USB / Serial</td>
<td>3 / 2</td>
<td>2 / 2</td>
<td>2 / 2</td>
</tr>
<tr>
<td>CuD</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>LPAR</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>OS</td>
<td>Red Hat SUSE</td>
<td>Red Hat SUSE</td>
<td>Red Hat SUSE</td>
</tr>
</tbody>
</table>
### P520, 1.65 GHz POWER5

<table>
<thead>
<tr>
<th></th>
<th>p520,1.65 GHz</th>
<th>p630,1.45 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPECint_base2000</td>
<td>1201</td>
<td>884</td>
</tr>
<tr>
<td>SPECint_rate_base2000</td>
<td>30.3</td>
<td>35.8</td>
</tr>
<tr>
<td>SPECfp_base2000</td>
<td>2034</td>
<td>1097</td>
</tr>
<tr>
<td>SPECfp_rate_base2000</td>
<td>41.5</td>
<td>38.1</td>
</tr>
</tbody>
</table>
### P550, 1.65 GHz POWER5

<table>
<thead>
<tr>
<th></th>
<th>p550, 1.65 GHz</th>
<th>p650, 1.45 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-way POWER5</td>
<td>1200</td>
<td>909</td>
</tr>
<tr>
<td>SPECint_base2000</td>
<td>60.4</td>
<td>72.7</td>
</tr>
<tr>
<td>SPECint_rate_base2000</td>
<td>1221</td>
<td>79.7</td>
</tr>
<tr>
<td>8-way POWER4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPECfp_base2000</td>
<td>82.1</td>
<td>79.7</td>
</tr>
<tr>
<td>SPECfp_rate_base2000</td>
<td>2121</td>
<td>1221</td>
</tr>
</tbody>
</table>

### P570, 1.9 GHz POWER5

<table>
<thead>
<tr>
<th></th>
<th>p570, 1.9 GHz</th>
<th>p655, 1.7 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-way POWER5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPECint_base2000</td>
<td>1398</td>
<td>1064</td>
</tr>
<tr>
<td>SPECint_rate_base2000</td>
<td>74.4 (4x)</td>
<td>47.7 (4x)</td>
</tr>
<tr>
<td></td>
<td>141 (8x)</td>
<td>83.5 (8x)</td>
</tr>
<tr>
<td></td>
<td>273 (16x)</td>
<td></td>
</tr>
<tr>
<td>8-way POWER4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPECfp_base2000</td>
<td>2576</td>
<td>1642</td>
</tr>
<tr>
<td>SPECfp_rate_base2000</td>
<td>125 (4x)</td>
<td>66.5 (4x)</td>
</tr>
<tr>
<td></td>
<td>241 (8x)</td>
<td>103 (8x)</td>
</tr>
<tr>
<td></td>
<td>438 (16x)</td>
<td></td>
</tr>
</tbody>
</table>
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)*

<table>
<thead>
<tr>
<th>Platforms</th>
<th>Memory GB</th>
<th>CPUs</th>
<th>1</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM p5-570 1.9 GHz POWER5</td>
<td>64</td>
<td>16</td>
<td>1459</td>
<td>1137</td>
<td>935</td>
</tr>
<tr>
<td>IBM p655 1.7GHz POWER4™</td>
<td>16</td>
<td>4</td>
<td>1750</td>
<td>1348</td>
<td>1111</td>
</tr>
<tr>
<td>HP rx5670 1.5 GHz Itanium 2</td>
<td>24</td>
<td>4</td>
<td>1851</td>
<td>1454</td>
<td>1212</td>
</tr>
</tbody>
</table>

- 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: select “Hardware Support Database”, then benchmarks.

---

**ANSYS V7.1**

*Sum of 12 standard ANSYS runs (Elapsed Time in sec)*

![Graph showing the comparison of elapsed time for different platforms](image)

- 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](http://www.ansys.com/services/hardware_support/index.htm) select “Hardware Support Database”, then benchmarks.
AIX System Unterstützung

- AIX 5.1
  Unterstützt alle derzeit verfügbaren und in Wartung befindlichen pSeries Systeme
  AIX 5.1 ist die Mindestvoraussetzung für POWER4 basierte Systeme
  Keine Unterstützung für zukünftige POWER5 basierte Systeme

- AIX 5.2
  Unterstützt alle POWER basierten Systeme, die von AIX 5.1 unterstützt werden mit Ausnahme von Microchannel und CHRP/PR6P Systemen.
  Neu: Unterstützung von POWER Blades und POWER5 Systemen (ohne SMT) über neuen Maintenance Level

- AIX 5.3
  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/CuD
- 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
- Enterprise Distributed Storage
- JFS2 - quotas, shrinkfs, >16TB
- Large storage objects (fs, lvm, devices)
- NFSv4 - strong security & performance
- Networking and I/O updates
- 4X Infiniband adapter and protocol support
- TCP/IP offload engine support
- IP over FC, iSCSI
- Affinity/Application Enabling
- POSIX Real Time
- JAVA 1.4.2 support
- PAM Extensions, SVR4 Loader/Linker opts.
- System Management, Usability, Stds
- HA & Secure NIM Server
- GNOME as a supported desktop
- UNIX03 compliance

AIX 5L support for IBM ~ BladeCenter™ JS20

- 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
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

- IBM p5-520, p5-550, p5-570

More information ...

- IBM @server p5 520 Technical Overview and Introduction
- IBM @server p5 550 Technical Overview and Introduction
- IBM @server p5 570 Technical Overview and Introduction
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
- 48-bit virtual address space
  - Nocona (2Q04) 36-bit 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

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

© 2004 Copyright by DYNAmore GmbH
I/O Busses

- **PCI Express**
  - Serial I/O 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

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

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