Towards Highly Scalable Clusters for Crash

Dr. Achim Bömelburg,

IBM Germany
Towards Highly Scalable Clusters for Crash

Dr. Achim Bömelburg, IBM Germany

Contents

- Industry Trends
  - Auto/HPC trends
- HPC Product Update
  - XEON 5100 series
  - Opteron "rev F"
- IBM “value add”
  - GPFS
  - vMIO
  - 10M element crash
- General Discussion
Automotive (CAE) Segment in the “Top500” Processor Market Share

Clusters, POWER5 and Linux, dominate the market

Ref.: “Top20 Auto survey...”
Christian Tanasescu

Leadership is a result of good decisions and solid execution

- Investment in POWER processor line (1999)
  - Increased commitment to HPC market
  - Leverage technology across HPC and commercial
- Introduction of POWER4 (2001)
  - First “dual-core” processor
  - 2x performance advantage over competition
- Early investment/commitment in Linux (2001)
  - Expertise in place for Auto industry transition.
- First “tier 1” vendor with Opteron (2004)
  - Workstation, servers and blades
  - Did not invest in Itanium
- Investment in “blade” technology
  - Offerings for POWER, Xeon, and Opteron

IBM is a driver … others are passengers

Image courtesy of CEI
**Evolution of HPC Hardware**

- **MainFrames (~1979)**: Mostly MSC.Nastran
- **Vectors (~1983)**: Beginning in 1986 crash simulation drove CAE compute requirements
- **RISC SMPs (~1994)**: SMP architecture was often first introduced in the CFD department and helped push parallel computing.
- **Clusters (~2001)**: Embedded systems show new perspectives for CAE
- **Embedded (now)**: Cluster architecture (Unix & Linux) now dominate crash and CFD environment

**CAE System Architecture: AIX and/or Linux**

- IBM business is about 50/50 Linux clusters and AIX clusters
- Linux systems tend to be special purpose
- AIX/Power systems are preferred for “implicit structures” and general purpose systems
Challenges of Clusters

- Applications that do not scale
- Cluster nodes have weak I/O (compared to large SMP)
- Parallel I/O across the cluster
- Lots of processors generate lots of heat.
- Lots of processors generate lots of data

- IBM value: POWER5/6 processor and “fast” x86 MSC.Nastran
- IBM value: MIO for Linux (fast I/O libraries).
- IBM value: GPFS (General Parallel File System)
- IBM value: working with ISVs to promote simulation data management
**Processor landscape becoming ‘simplified’**
- down to x86 and POWER

**POWER**
- first to dual-core
- POWER6 to push clock
- increasing reliance of SMT to maximize performance
- POWER family (i.e. embedded, gaming) influence on future

**x86**
- initial push to higher clocks
- thermal problems push direction to multi-core
- increasing reliance on SSE for HPC
- memory future is blurry (FBdimm vs DDR?)

**Commonality**
- ultimately the measure of performance will be dictated by the speed and number of threads per socket.

**Compute center economics paradigm shift**
- worldwide demand for energy is increasing faster than the supply
- as a result, energy consideration will become increasingly important factor in providing CAE server solutions
- while BG/L currently has limited applicability within CAE, its technology is pushing the envelope of energy efficiency, which will play a crucial role for future servers
• Costs have become an overriding consideration in packaging
  – product cycles changing from 36 to 6 months!
  – reuse of components is a must
• Environmentals (space and energy) pushing packaging technology forward
• Utility mentality emerging
  – shift from homogenous computer floor to constant upgrade of resource grid.

**HPC Hardware Value**

*It is now much more than $/MFLOPS*

Total Cost of Ownership (TCO) is now more complicated.

• ISV application cost
• Power and Cooling
• Engineer productivity
• Data center floor space
Contents

- Industry Trends
  - Auto/HPC trends
  - the 5 P's
- HPC Product Update
  - XEON 5100 series
  - Opteron "rev F"
- CAE server solutions
  - structures
  - impact analysis
  - CFD
  - Interconnects
- IBM "value add"
  - PowerExecutive, GPFS
  - MD Nastran Tuning, vMIO, Accuracy
- General Discussion

XEON vs. Opteron Product Positioning

- The benchmark wars are in full swing
  - Intel 5160 "Woodcrest 3.0 GHz"
  - AMD/Opteron "Rev. F" 2.8 GHz
  - dual-core chips with comparable performance
- It is often difficult to identify optimal product to deploy
- There are several key things to understand about each solution that help us identify which is optimal for a given workload
- But remember, the areas were there are overwhelming and compelling differences between the two (Xeon and Opteron) are usually easily identified
  - and in many cases it boils down to customer preference
New System x™ and BladeCenter® Servers: Intel Xeon processors

Position

- **x3550**
  - Low cost HPC compute node
  - Highly available application server

- **x3650**
  - High availability

- **X3850/3950**
  - Ultimate scale-out integration

Key Features

- Dual socket XEON 5100 series processors
- 1/32GB of FBD memory
- 2(3.5”) or 4(2.5”) SAS internal storage

- Dual socket XEON 5100 series processors
- 1/48GB of FBD memory
- 8(2.5”) or and 6(3.5”) SAS + tape internal storage

- Dual socket XEON per node
- Up to 6 nodes per system
- 6(2.5”) SAS internal storage per node

- Dual socket XEON 5100 series processors
- 1/32GB of FBD memory
- 2(2.5”) SAS internal storage + 3(2.5”) with optional iO blade

New System x™ and BladeCenter® Servers: AMD Opteron™ Rev F

Position

- **x3455**
  - Low cost HPC compute node

- **x3655**
  - Highly available application server

- **x3755**
  - Mid-Market, Large Enterprise HPC

Key Features

- Dual socket Opteron processors
- 4GB of DDR2 memory
- 3.5” Fixed SATA
- Leadership I/O with PCI-E, and HTx

- Dual socket Opteron processors
- 64GB of DDR2 memory
- 3.5” and 3.5” internal storage and tape
- Ready RAID and Ready RSA
- Trusted Platform Module
- Standard TOE

- Dual socket Opteron processors
- 128GB of DDR2 memory
- 3.5” SAS internal HDD
- Ready RAID and Ready RSA
- Trusted Platform Module
- Standard TOE

- Dual socket Opteron processors
- 32 GB of DDR2 memory
- SAS HDD technology
- TOE NIC solution
- High speed enablement
- Supports the new iO blade

- Dual socket Opteron processors
- 64GB of DDR2 memory
- 2 SAS HDDs and RAID
- TOE NIC solution
- High speed enablement
- Supports the new iO blade
- 4 ethernet ports
POWER Processor Roadmap

- 2001-4
  - POWER4 / 4+
    - 180 nm
    - 1.5 GHz Core
    - Shared L2
    - Chip Multi Processing
    - Distributed Switch
    - Advanced Multi Core Design
    - Advanced System Features

- 2004-6
  - POWER5 / 5+
    - 130 nm
    - 1.5 GHz Core
    - Shared L2
    - Distributed Switch
    - Shared L2
    - Enhanced Virtualization
    - Advanced Memory Subsystem
    - Decimal Floating Point
    - Check Point Restart
    - Enhanced architecture for higher frequencies

- 2007-9
  - POWER6 / 6+
    - 90 nm
    - 2 GHz Core
    - 2.2 GHz Core
    - Distributed Switch
    - Advanced Features
    - Virtualization

- 2010-13
  - POWER7
    - 65 nm
    - 1.5 GHz Core
    - 1.5 GHz Core
    - Distributed Switch
    - Advanced System Features

BINARY COMPATIBILITY

POWER6 Rollout in 2007 (and 2008)

- Preliminary!
- 4, 8, 12, 16-way node at 4.7 GHz
- 8-way node DCM at 4.7 GHz
- 2-way blade at 4.0 GHz
- 4-way blade at 3.8 GHz
- 64-way node DCM at 5.0 GHz
- 32-way node DCM at 4.7 GHz
- to 32-way node DCM at 3.5 GHz

2007

- 32-way node

2008
Performance characteristics

- benchmarks:
  - Neon: 230K; 20ms; 4-way
  - rNeon: 550K; 30ms; 4-way
  - 3Car: 1.2M; 10ms; 4-way
- standard benchmarks: www.topcrunch.org
- similar performance on POWER5, Opteron and XEON
- JS21 offers potential for excellent price/performance for AIX customers
- scales well with clusters
- cache friendly (follows SPECfp)

IBM solutions

- when AIX is the most important factor:
  - System p JS21: 2.5GHz 4-way nodes; 8GBmem; 1 internal drive; Myrinet preferred
  - System p 5.75*: 1.9GHz 16-way nodes; 16GBmem; 2 internal drives; HPS
- when price/performance is most important factor:
  - System x H521: 3.0GHz 4-way blades; 8GBmem, 1 internal drive; Myrinet or Ib
  - System x 3550: 3.00 GHz 4-way nodes; 8GBmem, 1-2 internal drives; Myrinet or Ib

LS-DYNA comparison: dual-core vs. quad-core

- April, 2007 testing: 3-car model, 795k elements, 150 msec
- IBM x3550 3.0 GHz Xeon 5160 "Woodcrest"
- IBM x3550 2.66 Xeon X5355 "Clovertown"
CAE Server Solutions

- No dominant server choice for all CAE applications
  - System p
    - strength of AIX
    - industry leading performance for many problems
    - well balanced performance for wide variety of simulation
  - System x
    - economics and flexibilities of open standards
    - extensive application portfolio
    - typically excellent price/performance

- No dominant server strategy for CAE customers
  - General purpose CAE servers
    - System p typically offers best performance for variety of applications
    - System x typically offer best price/performance
  - Application specific CAE servers
    - complex landscape which is always in flux

Contents

- Industry Trends
  - Auto/HPC trends

- HPC Product Update
  - XEON 5100 series
  - Opteron “rev F”

- IBM “value add”
  - GPFS
  - vMIO

- General Discussion
### Scalable Parallel I/O: General Parallel File System (GPFS)

- **NFS**
  - Client-server file systems have server bottleneck and protocol overhead

- **SAN**
  - SAN with a single metadata server have potential bottleneck

- **GPFS**
  - **General Purpose**
    - Any node can read from or write to any of the disks
    - The entire cluster can be administered from a single node
    - Supports Linux, AIX and mixed clusters
  - **High Performance**
    - Has provided 15GB/s to a single node and 100GB/s against a single file
    - GPFS is not a client-server file system and has much lower protocol overhead
    - All system data & metadata is equally accessible from all nodes
    - All data & metadata flows between the disks and nodes in parallel
  - **Scalability**
    - Currently supports 100s of nodes and 200+TB of storage over LAN or HPS
    - (more by special bid)
  - **Reliability**
    - Parallel operation means no single point of failure
    - One large research customer reported 100% uptime for GPFS for an entire year

---

### Members of IBM’s CAE Team (1)

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization</th>
<th>Experience Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nick Alsopp</td>
<td>ABAQUS, CFX</td>
<td>Many years experience in HPC and with HKS</td>
</tr>
<tr>
<td>Balaji Atyam</td>
<td>ANSYS, LMS, Madymo</td>
<td>Application Support USA</td>
</tr>
<tr>
<td>John Bauer</td>
<td>HPC I/O Libraries</td>
<td>Original developer of EIEIO libraries, 19 years HPC experience</td>
</tr>
<tr>
<td>Steve Behling</td>
<td>STAR-CD</td>
<td>13 years experience with STAR-CD source code</td>
</tr>
<tr>
<td>Achim Bömelburg</td>
<td>Permas, CAE Team</td>
<td>16 years experience with automotive customers</td>
</tr>
<tr>
<td>David Wei Chen</td>
<td>Detroit CAE Team</td>
<td>9 years working with automotive users in Detroit</td>
</tr>
<tr>
<td>Greg Clifford</td>
<td>Leader CAE Practice</td>
<td>20 years working with CAE customers and ISVs</td>
</tr>
<tr>
<td>Martin Beyereisen</td>
<td>Pam-Crash, LS-Dyna</td>
<td>Considered a member of the ESI development team</td>
</tr>
<tr>
<td>John Hague</td>
<td>VECTIS</td>
<td>ACTC Team UK</td>
</tr>
</tbody>
</table>

© 2007 Copyright by DYNAmore GmbH
Members of IBM’s CAE Team (2)

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization/Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holger Holthoff</td>
<td>RADIOSS, AVL 11 years experience in parallel computing on IBM platforms</td>
</tr>
<tr>
<td>Nobuhiro Kudanami</td>
<td>Tokyo CAE Team Working with automotive users in Tokyo</td>
</tr>
<tr>
<td>Guangye Li</td>
<td>LS-DYNA Extensive experience with LS-DYNA on Linux</td>
</tr>
<tr>
<td>Doug Petesch</td>
<td>NASTRAN, AMLS 15 years experience working with MSC and customers</td>
</tr>
<tr>
<td>Hari Reddy</td>
<td>FLUENT, PowerFLOW 6 years experience with FLUENT, experience with various CAE codes</td>
</tr>
<tr>
<td>A. Sugavanam</td>
<td>PowerFLOW, CEM Many years experience with NASA and CAE codes</td>
</tr>
<tr>
<td>Erling Weibust</td>
<td>CAE Team Sweden 19 years working with technical customers in Nordic region</td>
</tr>
<tr>
<td>Jeff Zais</td>
<td>Leader CAE technical team Key player in the 1999 success of MPP-DYNA</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

The IBM Value

- Experienced HPC applications team
- Worldwide customers
  - Longstanding relations with key application vendors
- Full range of computing solutions
  - POWER6 to Linux Clusters
  - Storage Solutions
- Presence of IBM
  - Stable company, growing in technical computing
  - Able to offer complementary solutions for storage and the desktop.