Towards Highly Scalable Clusters for Crash

Dr. Achim Bömelburg,

IBM Germany

IBM Deep Computing Team

Towards Highly Scalable Clusters for Crash

Dr. Achim Bömelburg, IBM Germany

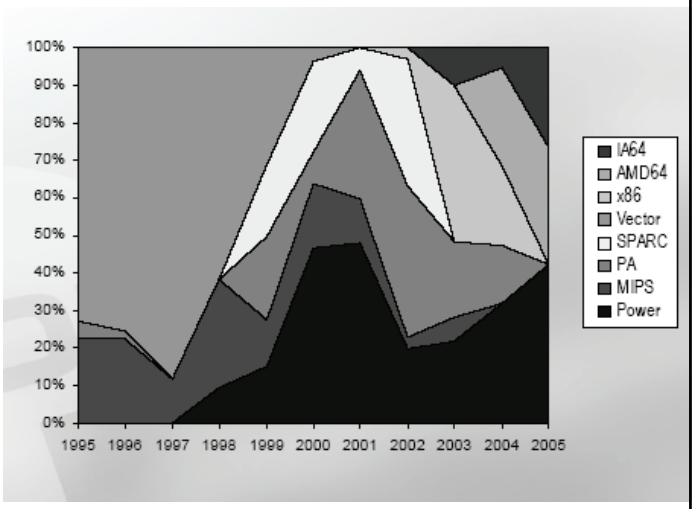
Spetember 2007

Contents

IBM Deep Computing

Automotive (CAE) Segment in the "Top500" Processor Market Share

Clusters, POWER5 and Linux, dominate the market



Ref.: "Top20 Auto survey…"
Christian Tanasescu

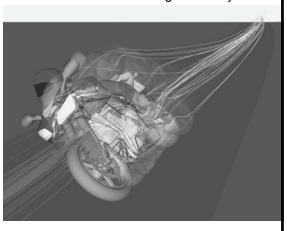
3 | 2007 Q1 Auto Aero statusCAE Overview | 9/16/2007 |

IBM Deep Computing

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

Image courtesy of CEI



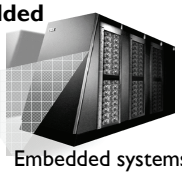
IBM is a driver … others are passengers

4 | 2007 Q1 Auto Aero statusCAE Overview | 9/16/2007 |

IBM Deep Computing Team

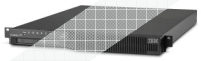
Evolution of HPC Hardware

Embedded
(now)



Embedded systems show new perspectives for CAE

Clusters
(~2001)

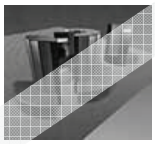


Cluster architecture (Unix & Linux) now dominate crash and CFD environment

RISC SMPs
(~1994)

SMP architecture was often first introduced in the CFD department and helped push parallel computing.

Vectors
(~1983)



Beginning in 1986 crash simulation drove CAE compute requirements

MainFrames
(~1979)

Mostly MSC.Nastran

5 2007 Q1 Auto Aero status 9/16/2007 © 2006 IBM Corporation

IBM Deep Computing Team

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

6 2007 Q1 Auto Aero status 9/16/2007 © 2006 IBM Corporation

IBM Deep Computing Team

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**

7 2007 Q1 Auto Aero status 9/16/2007

IBM Deep Computing Team

Challenges of Clusters

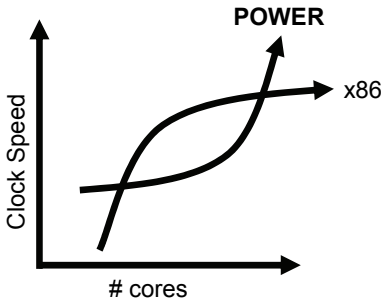
- **Applications that do not scale**
  - IBM value: POWER5/6 processor and “fast” x86 MSC.Nastran
- **Cluster nodes have weak I/O (compared to large SMP)**
  - IBM value: MIO for Linux (fast IO libraries).
- **Parallel I/O across the cluster**
  - IBM value: GPFS (General Parallel File System)
- **Lots of processors generate lots of heat.**
  - IBM value: “blades”, “Power Executive”, water cooled rack.
- **Lots of processors generate lots of data**
  - IBM value: working with ISVs to promote simulation data management

8 2007 Q1 Auto Aero status 9/16/2007

IBM Deep Computing Team

PROCESSORS

- **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.



9 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

IBM Deep Computing Team

POWER

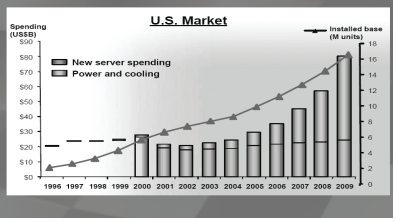
- **Compute center economics paradigm shift**
  - –worldwide demand for energy is increasing faster the supply
  - –as a result, energy consideration with become increasingly important factor in providing CAE server solutions
  - –while BG/L currently has limited applicability within CAE, its technology is pushing the envelope of energy efficiency, which will play a crucial role for future servers

▪ Power and cooling spend will exceed new server spending (Gartner 2006)

2000 – Raw processing "horsepower" is the primary goal, while the infrastructure to support it is assumed ready

2006 – Raw processing "horsepower" is a given, but the infrastructure to support deployment is a limiting factor

Three Cooling Challenges
1. The System
2. The Rack
3. The Data Center



10 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

IBM Deep Computing Team

- **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.

Frames
↓
Racks
↓
Blades
↓
Bricks

PACKAGING

11 2007 Q1 Auto Aero status 9/16/2007 © 2006 IBM Corporation

IBM Deep Computing Team

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**

12 2007 Q1 Auto Aero status 9/16/2007 © 2006 IBM Corporation

IBM Deep Computing Team

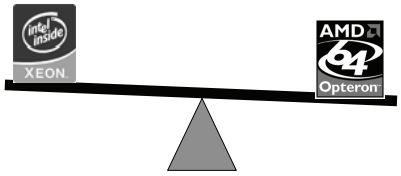
Contents

- 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

IBM Deep Computing Team

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



IBM Deep Computing Team

New System x™ and BladeCenter® Servers: Intel Xeon processors

Position

x3550	x3650	X3850/3950	*Ultimate scale-out integration* HS21
			
Low cost HPC compute node	*Highly available application server*	*Mid-Market, Large Enterprise HPC*	*Enterprise class scalable 2-socket blades for front and mid tier applications*

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	▪ **Four socket XEON per node** ▪ **16 2/64GB of DDR2 per node** ▪ **Up to 8 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 SIO blade

15 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

IBM Deep Computing Team

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

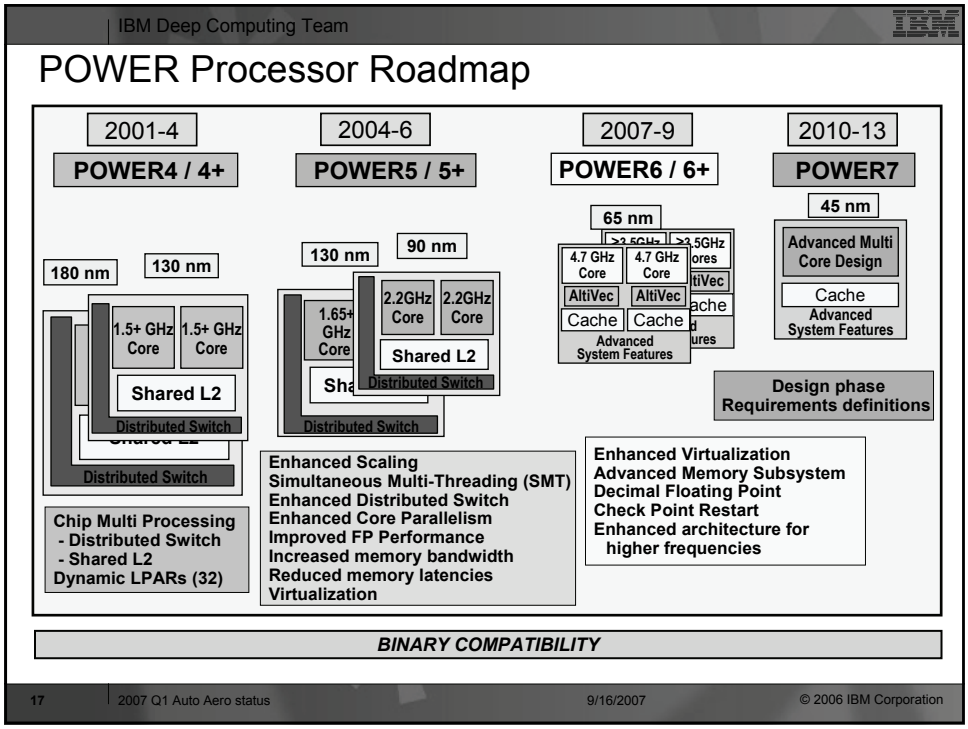
Position

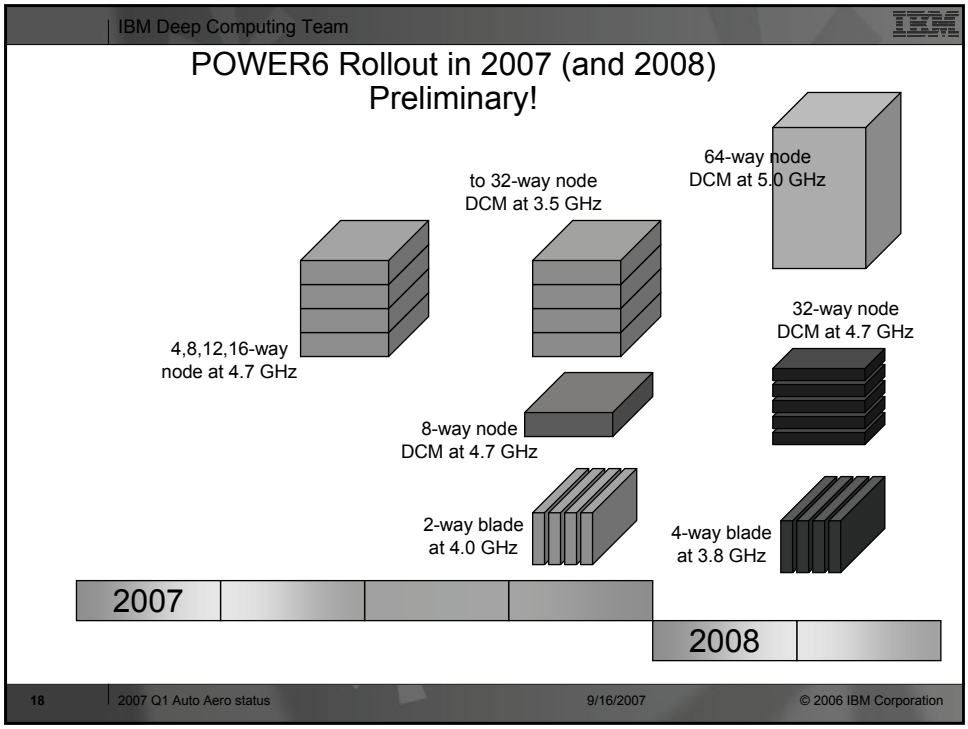
x3455	x3655	x3755	*Ultimate scale-out integration* LS21 LS41
			 
Low cost HPC compute node	*Highly available application server*	*Mid-Market, Large Enterprise HPC*	*Enterprise class scalable 2-4 socket blade for front and mid tier applications*

Key Features

▪ **Dual socket Opteron processors** ▪ **48GB of DDR2 memory** ▪ 3.5" Fixed SATA ▪ Leadership I/O with PCI-E, and HTx	▪ **Dual socket Opteron processors** ▪ **64GB of DDR2 memory** ▪ 2.5"and 3.5" internal storage and tape ▪ Ready RAID and Ready RSA ▪ Trusted Platform Module ▪ Standard TOE	▪ **Four 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 SIO blade	▪ **Four socket Opteron processors** ▪ **64GB of DDR2 memory** ▪ 2 SAS HDDs and RAID ▪ TOE NIC solution ▪ High speed enablement ▪ Supports the new SIO blade ▪ 4 ethernet ports

16 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation



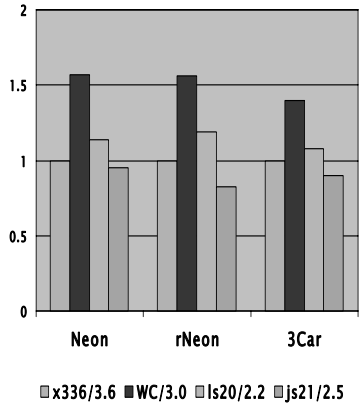


IBM Deep Computing Team

LSDYNA

- Performance characteristics
  - benchmarks:
    - Neon: 230K; 20ms; 4-way
    - rNeon: 550K; 30ms; 4-way
    - 3Car: 1.2M; 10ms; 4-way
  - standard benchmarks: *www.topcrunch.org*
  - cache friendly (follows SPECfp)
    - similar performance on POWER5, Opteron and XEON
    - JS21 offers potential for excellent price/performance for AIX customers
  - scales well with clusters
    - high performance network now the norm
- 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 p5 575+; *1.9GHz 16-way nodes; 16GBmem; 2 internal drives; HPS*
  - when price/performance is most important factor
    - System x HS21; *3.00 GHz 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*

Performance *(relative to x336, higher is better)*

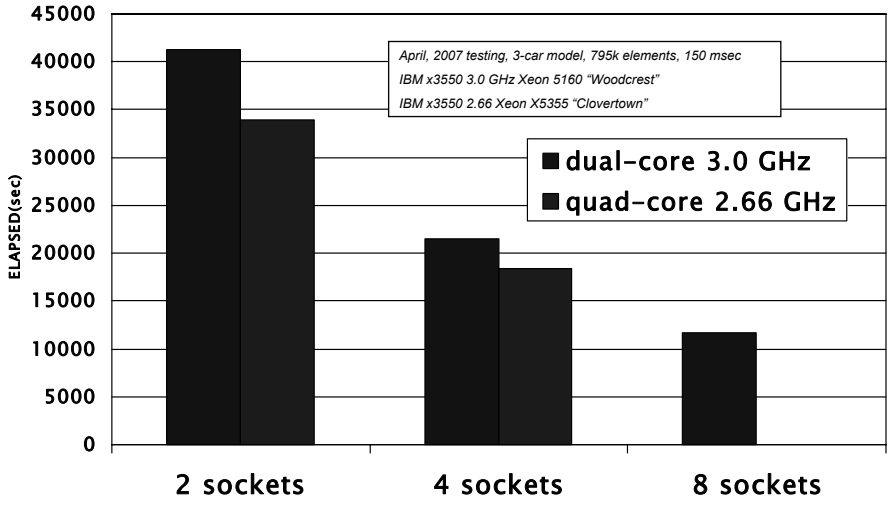


LSTC
Livermore Software Technology Corp.

19 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

IBM Deep Computing Team

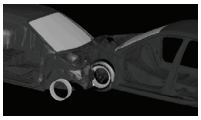
LS-DYNA comparison: dual-core vs. quad-core



20 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

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

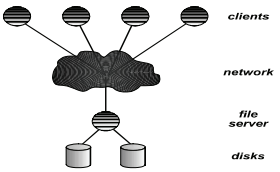
Contents

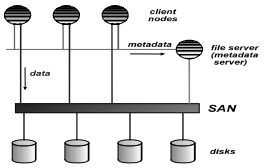
- Industry Trends
  - Auto/HPC trends
- HPC Product Update
  - XEON 5100 series
  - Opteron “rev F”
- IBM “value add”
  - GPFS
  - vMIO
- General Discussion

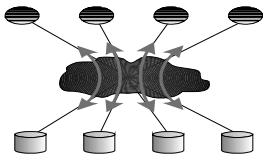
IBM Deep Computing Team

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 LANor 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







23 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

IBM Deep Computing Team

Members of IBM's CAE Team (1)

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

24 | 2007 Q1 Auto Aero status | 9/16/2007 | © 2006 IBM Corporation

IBM Deep Computing Team

Members of IBM's CAE Team (2)

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

25 2007 Q1 Auto Aero status 9/16/2007 © 2006 IBM Corporation

IBM Deep Computing Team

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.

26 2007 Q1 Auto Aero status 9/16/2007 © 2006 IBM Corporation