Compute Clusters at the Chrysler Group: Battle Scars and Victory Parades

June 24, 2003

John Picklo

Manager, Mainframes and High Performance Computing

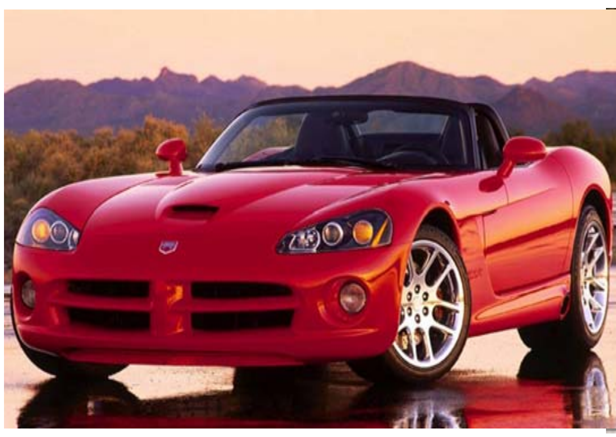
Agenda

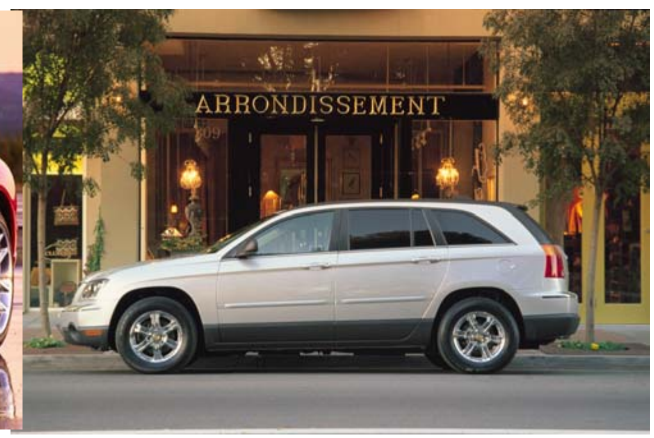
- Introductory comments
- History of clusters at the Chrysler Group
- Implementation issues
- Lessons learned

Chrysler Group Product Development

- Design and engineering for Chrysler Group vehicles

DODGE CHRYSLER Jeep





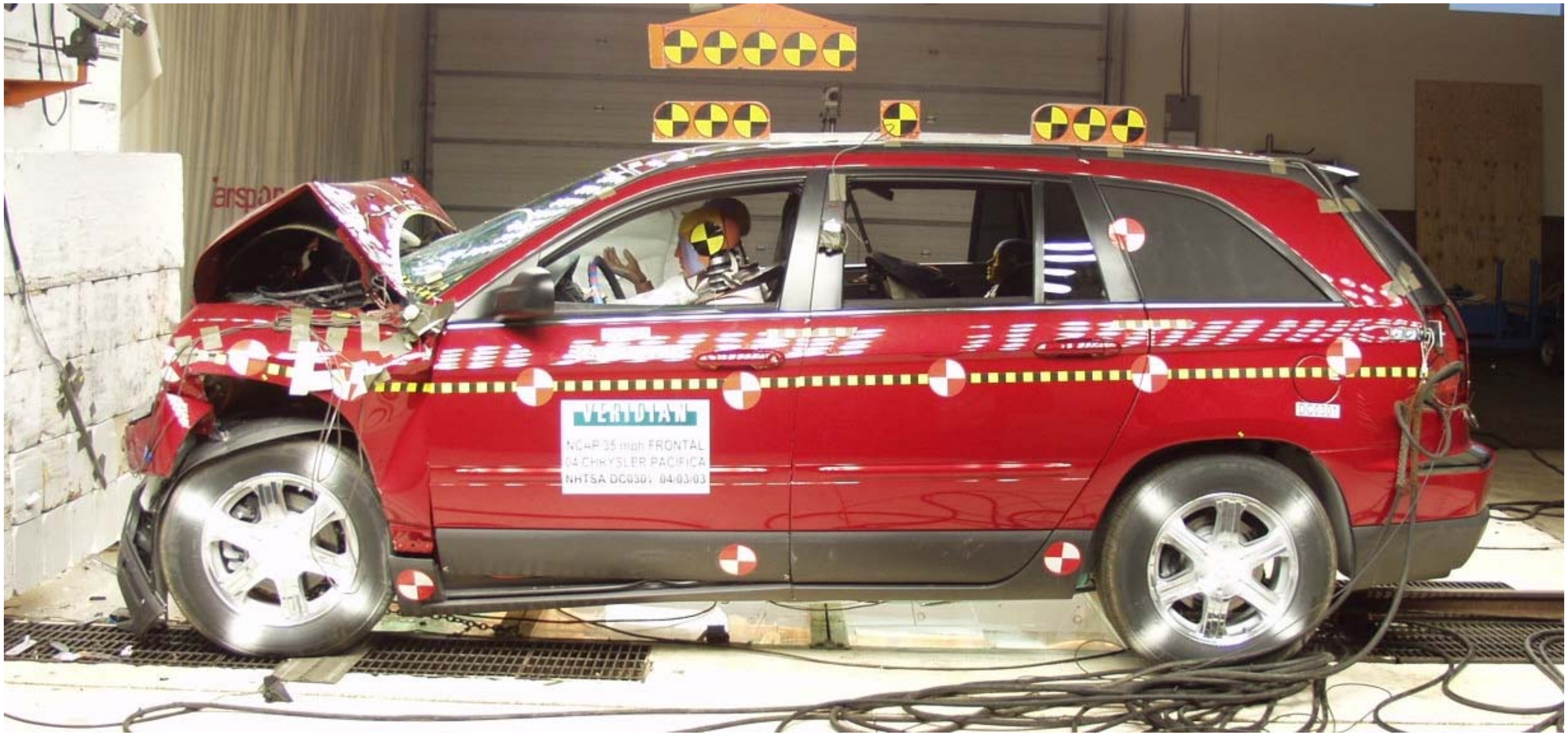
High Performance Computing

- Computer Aided Engineering
- Vehicle simulation
  - Fluid Dynamics
  - Impact Events
  - Noise, Vibration, Harshness

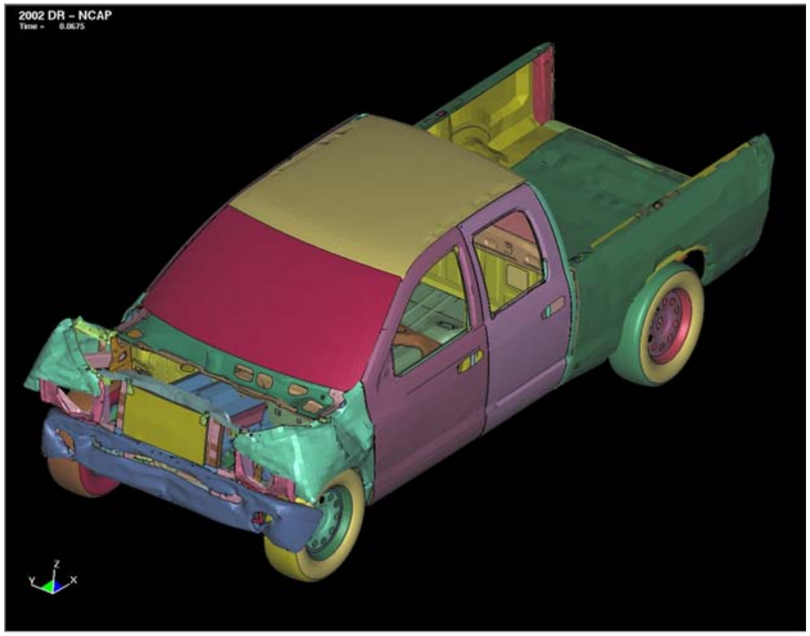
CAE Simulation Process

- Pre-process
  - Create meshed model
  - Elements and properties
  - > 700,000 elements
- Simulation
  - Batch process
  - Compute and memory intensive
  - Duration: hours to days
- Post-process
  - Graph results
  - Visualize
  - Animate

Simulation Examples



Simulation Examples



Workload Characteristics

- Multi-cpu jobs: 1-24
- Processor, memory, and I/O intensive
- Not data centric
- Heterogeneous systems
  - Multi-vendor
  - Evolutionary introduction of technology
  - Jobs modified to match environment
- LSF provides workload management (Impact, NVH)
  - Job scheduling interface
  - Load leveling across systems
  - Maximize utilization and minimize queue time

HPC Goals

- Reduce simulation cycle time
  - Shorten the design timeline
  - Evaluate more design alternatives
- Reduce costs
  - Engineer productivity
  - Physical tests
  - Cost of computing
- Increased accuracy
- Improved vehicles

HPC Systems

➤ HP:	6-Superdomes	(352 CPUs)
	96 Node Itanium Cluster	(192 CPUs)
➤ SGI:	12 Origin 300	(384 CPUs)
	4 Origin 3000	(256 CPUs)
	2 Origin 2000	(128 CPUs)
➤ IBM	172 Node Pentium Cluster	(344 CPUs)
	64 Node Pentium Cluster	(128 CPUs)

Cluster History

Chrysler group's first cluster

- Installed 1998
- First entry into MPI and clusters
- **Retired 2003**

➤ HP N-Class

- 56 Compute nodes
- Impact and NVH
- 4-8 cpus per node
- 400 mhz PA8500
- 8-12 cpus per job
- Gigabit and hiperfabric backbone
- 11tb scratch disk
- 5 Front end L-class data servers with fail-over software

Impact Pentium Cluster

- IBM Intellistations
  - 172 Nodes, 344 CPUs
  - Xeon 2.2ghz, 2.8ghz
  - Gigabit internal network
  - 12 nodes per switch
  - RedHat Linux V7.2
  - Installed July 2002
  - Management node w XCAT software
  - Storage node w 1.8tb shared disk

Itanium NVH Cluster

- 96 Nodes, 192 CPUs
- HP ZX6000
- HP/UX V11.22
- MSC/Nastran
- 675gb shared storage
- 288gb scratch space per node
- Gigabit internal network

CFD Hybrid Clusters

- 2 SGI SMP front-end systems
  - Geographically separated for disaster recovery purposes
  - 64 CPUs, 64gb
  - 64 bit applications, job scheduling, and file serving
- Pentium compute cluster behind each front-end
  - 32 nodes each (64 cpus)
  - IBM x335, 2.8ghz
  - 32 bit applications
- PBSpro
  - Load level between site and front/back end
  - Match workload to available licenses

Implementation Issues

Cluster Enablers

- ISV's
  - Implement Message Passing Interface
  - Port to Linux
  - Quality assurance
- Intel
  - 1.8ghz Xeon performance exceeds Risc
  - Itanium 2 performance
- Integrators
  - Industrial grade solutions
    - Ø Bag-of-wires
  - Turnkey implementation services
  - Hardware and software support
  - Cost effective designs
    - Integrator provides research and development
    - Node speed and size
    - Back-end network
  - ClusterWare

Internal Network

- Speed vs. efficiency
  - Proprietary (Myrinet, Scali, etc.)
  - 10/100/1000 Ethernet
  - New technology (Infiniband)
- Switch size
  - Flat is expensive
  - Hierarchical limits job size
    - Sub clusters
    - Maintaining locality of traffic for jobs
- Multiple clusters
  - Network of private networks
  - Separating public and private traffic
- Traffic
  - Data
  - MPI
  - Management

Workload Management

- Workload management
  - More than job scheduling
    - Clean-up after kill / failure
    - Accounting
    - Visibility to back-end processes
  - Sub-clusters
    - Maintaining locality of traffic
  - Heterogeneous systems
    - Unequal nodes
      - Number of CPUs
      - CPU speed
    - Mixed environment (SMP and clusters)

Data

- Internal disks
  - Do you need / want internal disks?
  - Is there activity that should be targeted to the internal disks?

- Shared disk
  - SAN
    - Costly and complex
    - High performance
  - NFS
    - Less expensive and simpler
    - Moderate performance

Lessons Learned

Operating System

- Corporate Culture
  - Corporate standards for open source
  - Niche implementations vs. corporate direction
  - Can you fly below the radar?
- Which OS is right for your cluster?
  - Linux concerns
  - What about Windows?
  - HP/UX for Itanium

Resilience

- Likelihood of failure
  - Are commodity components more or less likely to fail?
- Impact of failure
  - Cost of lost node
  - HW support
    - Back end nodes vs. shared components
    - Off-hours level of diagnosis and ability to repair

DCX Failure experiences (2002)

- SMP : 0.10% jobs lost to system failure
- Cluster: 0.05% jobs lost to system failure
- Cost savings for mixed mode hardware support

ClusterWare

- Number of offerings
- Need for standardization
- Interface to other standard tools
- Open source vs. proprietary

Choosing Your Integrator

- Roll your own: “bag of wires”
  - Who you gonna call?
  - How much did you really save?
  - When will you find out if you made a configuration mistake?
- Application expertise is dominant factor
- HW and SW support
- Let your integrators do your research and development
- There are plenty of players. When in doubt, bid it out.

Futures

- Internal networks
  - Ethernet improvements
  - Infiniband
  - Proprietary: Myrinet, Quadrics, etc.
    - Cost
    - Future
    - Duplicate network

Futures

- Multiple clusters
  - Homogeneous interface to heterogeneous clusters
  - Multiple operating systems
  - Hybrid hardware
    - 32bit and 64bit applications
  - ClusterWare and workload management
- Cluster experience will benefit Grid implementation

Closing Remarks

Benefits

- Cost
  - Metrics?
  - Transactions (jobs) per dollar spent
- Performance
  - Metrics?
  - Turnaround vs. throughput
- Visibility

Benefits

- Pentium Cluster
  - 20% performance improvement over equivalent number of RISC processors
  - 40% cost benefit
- Itanium Cluster
  - 50% performance improvement over RISC

Enablers

- It is all about the application
- Know your workload
  - How many bells and whistles will provide benefits?
  - Run benchmarks
  - Use application savvy integrators

Cluster Strategy

- Good for only the next installation
  - Technology is changing too fast
  - When does it make sense to expand an existing cluster
- Niche solutions
  - Don’t have to transform the enterprise
- Corporate resistance
  - The only thing more notorious than a successful implementation is an unsuccessful one

Perspective

Remember, the primary reason to embrace clusters is to reduce costs. It is easy to lose sight of this goal and succumb to the temptation to over-configure.

Questions?