Real Parallel Computers
Modular data centers
Background Information

Recent trends in the marketplace of high performance computing
Strohmaier, Dongarra, Meuer, Simon
Parallel Computing 2005
Short history of parallel machines

• 1970s: vector computers
• 1990s: Massively Parallel Processors (MPPs)
  – Standard microprocessors, special network and I/O
• 2000s:
  – Cluster computers (using standard PCs)
  – Advanced architectures (BlueGene)
  – Comeback of vector computer (Japanese Earth Simulator)
  – IBM Cell/BE
• 2010s:
  – Multi-cores, GPUs, Intel Phi
  – Cloud data centers
Performance development and predictions

Performance Development

Projected Performance Development
Clusters

• Cluster computing
  – Standard PCs/workstations connected by fast network
  – Good price/performance ratio
  – Exploit existing (idle) machines or use (new) dedicated machines

• Cluster computers vs. supercomputers (MPPs)
  – Processing power similar: based on microprocessors
  – Communication performance was the key difference
  – Modern networks have bridged this gap
    • Infiniband, 10G Ethernet, Myrinet
Literature

History of Cluster Computing

Exploratory Period: Before 1980


Classical Period: 1992-2004

Advanced Period: 2005-now
History of Cluster Computing

• **Exploratory Period: Before 1980**
  – Intel X86, Ethernet (Xerox Parc), TCP, Unix, CSP model

• **Enabling Period: 1980-1992**
  – 16 and 32 bit processors @ 100 MHz
  – 10 Mbit Ethernet
  – BSD Unix (Berkeley), with virtual memory and networking
  – PVM (message passing library)
  – Condor: match-making scheduler
Classical Period: 1992-2004

- Two large research projects:
  - UCB: NOW (Network of Workstations)
    - High-end workstations & networks, proprietary software
  - NASA: Beowulf
    - Low-cost PCs, commodity networks, open source, <$50K
- 1994: MPI message passing standard
- 1995: Myrinet network (Myricom): expensive high-speed network that can be plugged into PCs
- 1997: First cluster in top-500
- Supercomputer world was skeptical about whole concept
- 2004: 50% of top-500 were clusters
Advanced Period: 2005-now

- Clusters with multi-core & GPU nodes
- Variety of new programming systems: CUDA, OpenCL, OpenMP, OpenACC, Cilk, TBB, …
- Sometimes called "MPI + X"
  - MPI for message passing, something unknown (X) for accelerators
- Infiniband network: low latency, inexpensive
  - 2011: 36% of top-500 was Infiniband, 50%: Ethernet
- Clusters have >80% of HPC market
- Supercomputers (IBM Blue Gene, Cray XT5) for high end of the market
VU clusters

• Cluster computers at our department
  – DAS-1: 128-node Pentium-Pro / Myrinet cluster
  – DAS-2: 72-node dual-Pentium-III / Myrinet-2000 cluster
  – DAS-3: 85-node dual-core dual Opteron / Myrinet-10G
  – DAS-4: 72-node cluster with accelerators (GPUs etc.)
  – DAS-5: expected Spring 2015

• Part of a wide-area system:
  – Distributed ASCI Supercomputer

Going Dutch: How to Share a Dedicated Distributed Infrastructure for Computer Science Research, keynote lecture at Euro-Par 2014 (Porto, 28 August 2014):
**Distributed ASCI Supercomputer**

- Distributed common infrastructure for Dutch Computer Science
  - Distributed: multiple (4-6) clusters at different locations
  - Common: single formal owner (ASCI), single design team
    - Users have access to entire system
  - Dedicated to CS experiments (like Grid’5000)
    - Interactive (distributed) experiments, *low* resource utilization
    - Able to modify/break the hardware and systems software
- Dutch: small scale
DAS generations: visions

- DAS-1: Wide-area computing (1997)
  - Homogeneous hardware and software
- DAS-2: Grid computing (2002)
  - Globus middleware
  - Dedicated 10 Gb/s optical links between all sites
- DAS-4: Clouds, diversity, green IT (2010)
  - Hardware virtualization, accelerators, energy measurements
  - Wide variety of accelerators, larger memories and disks
**DAS-1 (1997-2002)**

A homogeneous wide-area system

- **VU** (128 nodes)
- **Amsterdam** (24 nodes)
- **Leiden** (24 nodes)
- **Delft** (24 nodes)

- 200 MHz Pentium Pro
- Myrinet interconnect
- BSDI → Redhat Linux
- Built by Parsytec

6 Mb/s ATM
DAS-2 (2002-2006) a Computer Science Grid

two 1 GHz Pentium-3s
Myrinet interconnect
Redhat Enterprise Linux
Globus 3.2
PBS ➔ Sun Grid Engine
Built by IBM
DAS-3 (2006-2010)
An optical grid

Dual AMD Opterons
2.2-2.6 GHz
Single/dual core nodes
Myrinet-10G
Scientific Linux 4
Globus, SGE
Built by ClusterVision

VU (85)

TU Delft (68)

Leiden (32)

UvA/MultimediaN (40/46)

SURFnet6
10 Gb/s
DAS-4 (2011) Testbed for Clouds, diversity, green IT

- Dual quad-core Xeon E5620
- Infiniband
- Various accelerators
- Scientific Linux
- Bright Cluster Manager
- Built by ClusterVision

TU Delft (32)
ASTRON (23)
VU (74)
SURFnet6
UvA/MultimediaN (16/36)
Leiden (16)

10 Gb/s
DAS-4 (2011) Testbed for Clouds, diversity, green IT

- Dual quad-core Xeon E5620
- Infiniband
- Various accelerators
- Scientific Linux
- Bright Cluster Manager
- Built by ClusterVision

VU (74)
ASTRON (23)
TU Delft (32)
Leiden (16)

UvA/MultimediaN (16/36)
SURFnet6
10 Gb/s
# 4 DAS generations

<table>
<thead>
<tr>
<th></th>
<th>DAS-1</th>
<th>DAS-2</th>
<th>DAS-3</th>
<th>DAS-4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>Pentium Pro</td>
<td>Dual Pentium3</td>
<td>Dual Opteron</td>
<td>Dual 4-core Xeon</td>
</tr>
<tr>
<td><strong>LAN</strong></td>
<td>Myrinet</td>
<td>Myrinet</td>
<td>Myrinet 10G</td>
<td>Infini Band</td>
</tr>
<tr>
<td><strong># CPU cores</strong></td>
<td>200</td>
<td>400</td>
<td>792</td>
<td>1600</td>
</tr>
<tr>
<td><strong>SPEC CPU2000 INT (base)</strong></td>
<td>78.5</td>
<td>454</td>
<td>1445</td>
<td>4620</td>
</tr>
<tr>
<td><strong>SPEC CPU2000 FP (base)</strong></td>
<td>69.0</td>
<td>329</td>
<td>1858</td>
<td>6160</td>
</tr>
<tr>
<td><strong>1-way latency MPI (µs)</strong></td>
<td>21.7</td>
<td>11.2</td>
<td>2.7</td>
<td>1.9</td>
</tr>
<tr>
<td><strong>Max. throughput (MB/s)</strong></td>
<td>75</td>
<td>160</td>
<td>950</td>
<td>2700</td>
</tr>
<tr>
<td><strong>Wide-area bandwidth (Mb/s)</strong></td>
<td>6</td>
<td>1000</td>
<td>40000</td>
<td>40000</td>
</tr>
</tbody>
</table>
Blue Gene/L Supercomputer
Blue Gene/L

System
64 Racks, 64x32x32

180/360 TF/s
32 TB

2.8/5.6 TF/s
512 GB

90/180 GF/s
16 GB

5.6/11.2 GF/s
1.0 GB

90/180 GF/s
16 GB

2 chips, 1x2x1

2.8/5.6 GF/s

2 processors
Blue Gene/L Networks

3 Dimensional Torus
- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- 1 µs latency between nearest neighbors, 5 µs to the farthest
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 68TB/s total bandwidth

Global Collective
- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link
- Latency of one way traversal 2.5 µs
- Interconnects all compute and I/O nodes (1024)

Low Latency Global Barrier and Interrupt
- Latency of round trip 1.3 µs

Ethernet
- Incorporated into every node ASIC
- Active in the I/O nodes (1:8-64)
- All external comm. (file I/O, control, user interaction, etc.)

Control Network