Outline

• This lecture
  – Introduction
    • What are “scalable clusters”?
    • Why should we care?
    • Topics to be covered
  – Administrative Stuff
    • Course organization
    • Projects
What are Scalable Clusters?

Use of commodity PCs or workstations connected together with commodity low-latency, high-bandwidth networks as a viable alternative to custom parallel machines

• “Commodity” is influenced by several factors
  – price: high-volume products (sold in tens of thousands)
  – intrusiveness of interfaces: whether standard or not, etc.
• “Viable alternative to custom parallel machines”
  – focus on reliable, predictable performance
  – target applications include those that are communication-intensive

Scalable Clusters: Why bother?

• Potential to finally realize the dream of general-purpose parallel computing
• Hardware: cost/performance advantages
  – (potential for) comparable performance levels
  – at significantly reduced cost (economy of scale)
  – scalability and availability advantages
• Software: build upon pre-existing PC software base
• Huge interest in industry
  – Microsoft has a Windows/NT cluster API: WolfPack
  • targeted towards the database server market
  – Intel/Microsoft/Compaq have come together with a clustering interconnect standard (Virtual Interface Architecture)
Clusters: The Performance Argument

<table>
<thead>
<tr>
<th>PC Cluster w/ Myrinet</th>
<th>SGI Origin 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Node</strong></td>
<td><strong>Node</strong></td>
</tr>
<tr>
<td>- Intel Pentium II</td>
<td>- MIPS R10000</td>
</tr>
<tr>
<td>300 MHz</td>
<td>195 MHz</td>
</tr>
<tr>
<td>11.7 SpecInt95</td>
<td>11.5 SpecInt95</td>
</tr>
<tr>
<td>8.5 SpecFp95</td>
<td>19.1 SpecFp95</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td><strong>Network</strong></td>
</tr>
<tr>
<td>- Myrinet from Myricom, Inc.</td>
<td>- Dedicated controller on memory bus + SGI CrayLink</td>
</tr>
<tr>
<td>- PCI node interface</td>
<td>- latency: 500 ns</td>
</tr>
<tr>
<td>- latency: 1µs</td>
<td>- bandwidth: 125 MB/s</td>
</tr>
<tr>
<td>- bandwidth: 125 MB/s</td>
<td></td>
</tr>
</tbody>
</table>

Commodity components can deliver performance levels comparable to custom parallel machines

1/21/98

Clusters: The Cost Argument

<table>
<thead>
<tr>
<th>PC Cluster w/ Myrinet</th>
<th>SGI Origin 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PC</strong>: $2526.00 (Dell Optiplex)</td>
<td><strong>Node</strong></td>
</tr>
<tr>
<td>- 300 MHz Pentium II</td>
<td>- 195 MHz</td>
</tr>
<tr>
<td>- 512K L2 cache</td>
<td>- 64 MB memory</td>
</tr>
<tr>
<td>- 64 MB memory</td>
<td>- 4.5 GB disk</td>
</tr>
<tr>
<td>- 6.4 GB disk</td>
<td></td>
</tr>
<tr>
<td><strong>Myrinet</strong>: $1500.00 - $2000.00</td>
<td><strong>SGI CrayLink interconnect</strong></td>
</tr>
<tr>
<td>- PCI cards ($1400.00)</td>
<td>- $28,000.00 - $30,000.00 per node</td>
</tr>
<tr>
<td>- 8-port switch ($1500.00)</td>
<td></td>
</tr>
<tr>
<td>- cables ($100.00)</td>
<td></td>
</tr>
<tr>
<td><strong>$4500.00 per node</strong></td>
<td></td>
</tr>
</tbody>
</table>

Significant cost advantages!

1/21/98
Clusters: The Challenges

• Communication performance
  – user-level to user-level latency: 10µs (cluster) vs. 2µs (Origin)
  – node bandwidth: 125MB/s (cluster), 1250MB/s (Origin)
• Reliability
  – custom parallel machines exhibit high levels of availability/redundancy
• Lack of a single-system image
  – harder to make decisions about global resource management
• Central theme of the course
  – how to overcome these challenges to reliably/predictably achieve custom parallel machine-like performance on clusters?

Topics to be covered

• Technology trends (Lecture 2)
  – why is now the right time for clusters?
• Architecture of clusters (Lectures 3-5)
  – in an ideal world (Lecture 3)
  – present-day parallel machines (Lecture 4)
  – decoupled processor-network interfaces (Lecture 5)
• Middleware (Lectures 6-11)
  – communication:
    • messaging (Lectures 6-7)
    • distributed shared memory (Lectures 8-9)
  – fault-tolerance and availability (Lecture 10)
  – resource coordination (Lecture 11)
• Case studies (Lecture 12)
What this course excludes

- Loosely-coupled clusters
  - typically interesting only for compute-intensive applications

- Programming models
  - message passing: sequential node programs + MPI/PVM
  - data parallel: HPF
  - shared memory: threads + synchronization, Java
  - distributed objects: CORBA/DCOM, Java RMI

- I/O issues
  - commodity hard disks serve as a cheap I/O bandwidth resource

- Metacomputing issues
  - resource discovery and administration
  - Milan (NYU), Javelin (UCSB), Legion (UVa)

Course Organization

- Structure
  - lectures
    - background material
    - common framework which relates the individual papers
  - paper presentations and discussion
    - details about a particular system/technique and its contributions

- Grading
  - presentations
  - participation in discussion
  - project
  - no exams!
Assumed Background

- **Computer Architecture**
  - processor: ALU, registers, data paths, RISC
  - memory: caches, DRAMs, cache/memory controllers, *cache-coherence protocols*
  - system: memory- and I/O-mapped devices

- **Operating Systems**
  - virtual memory and protection domains
  - device drivers
  - process scheduling

Projects

- **Guidelines**
  - in groups of 2-3
  - any topic that explores the research issues discussed in the course
  - one-page writeup due March 3, final report due May 5
  - project presentations: April 21, and April 28

- **Topic suggestions**
  - application sensitivity to performance/structural differences between clusters and parallel machines
  - develop/evaluate
    - alternate cluster processor-network interfaces
    - cluster middleware (messaging, distributed shared-memory, others)
    - models for resource coordination
    - applications which are well-matched to clusters
Next Lecture

• Technology trends
  – CPU
  – Memory
  – Communication networks
  – Why is now the right time?

• Readings
  – *A Case for NOW (Networks of Workstations)*
    The Berkeley NOW Group
  – *Myrinet: A Gigabit-per-second Local-Area Network*
    C. Seitz et al.