High performance libraries for Windows 2000: from a developer standpoint

Laurent Lefèvre, Roland Westrelin

Supported by:

Microsoft®
Basic Interface for Parallelism: BIP

Developed by:
- Patrick Geoffray, Loïc Prylli, Bernard Tourancheau and Roland Westrelin
- at ENS Lyon and University of Lyon (France)

Goals:
- Maximization of the application level performance (close to the hardware maximum)
- Providing legacy programming interfaces: MPI (MPI-BIP based on MPICH) and IP

Initial targets: Myrinet, Linux

Composed of: libraries (BIP, MPI-BIP), an OS module, a firmware (architecture and OS independent), a basic runtime
From Linux to Win2000: the OS module

- A driver
- An IP driver
- Provide direct access to the Network Interface and Memory management

The idea is to use the GM OS module (available on several architecture, in 2 layers)

- Easier
- Incremental
- Trick: use winNT driver (not win2k)
From Linux to Win2000: the libraries

- Rely on the cygwin porting layer: emulate UNIX libc calls with WIN32 calls
- Makes maintenance easier
- gcc!
- Open-source, widely used, quickly improving
- No syscall in the critical path of communication in BIP (memory registration is cached)
- Still have to rewrite part of the library in native code (driver calls, synchronization)
- MPI: `./configure -arch=LINUX; make`
From Linux to Win2000: the runtime environment

- Perl scripts + ssh
- Route discovery, setting configuration, launching jobs
- Cygwin provides perl and ssh!
- Private key authentication (no password) only works with local accounts, NOT with domain controller
- Disk area shared with a linux samba server (no password authentication)
Managing the cluster

- We chose the brute-force method: replicate raw partition over the network (account management, software installation)
- SCSI disk performance: 30-50MB/s; use IP/Myrinet
- Broadcast the partition in a ring fashion (pipeline)
- Replicating the 5GB partition on 8 machines takes 2mn30s (≈ a local copy)
- Scale quite well up to 100 machines (theoretically)
- Performed under linux
- IP address managed by DHCP
- Hostname: edit the windows registry under linux
Performance: latency

Latency of BIP and MPI-BIP

One-way latency (microseconds)

Message size (bytes)
Performance: bandwidth

Bandwidth of BIP and MPI-BIP

Bandwidth (Bytes/s)

Message size (bytes)

BIP (Windows)
MPI-BIP (Windows)
BIP (Linux)
MPI-BIP (Linux)
### Performance: NAS application benchmarks

<table>
<thead>
<tr>
<th></th>
<th>Sequential</th>
<th>4 processes</th>
<th>8 processes</th>
<th>16 processes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Win.</td>
<td>Linux</td>
<td>Win.</td>
</tr>
<tr>
<td>IS (class A)</td>
<td>9.47</td>
<td>9.46</td>
<td>2.52</td>
<td>2.52</td>
</tr>
<tr>
<td>(class A)</td>
<td></td>
<td>(3.6)</td>
<td>(3.8)</td>
<td>(6.2)</td>
</tr>
<tr>
<td>IS (class B)</td>
<td>38</td>
<td>38</td>
<td>10.70</td>
<td>10.31</td>
</tr>
<tr>
<td>(class B)</td>
<td></td>
<td>(3.6)</td>
<td>(3.7)</td>
<td>(6.3)</td>
</tr>
<tr>
<td>LU (class A)</td>
<td>1597</td>
<td>1230</td>
<td>398</td>
<td>309</td>
</tr>
<tr>
<td>(class A)</td>
<td></td>
<td>(4)</td>
<td>(4)</td>
<td>(7.9)</td>
</tr>
<tr>
<td>LU (class B)</td>
<td>5646</td>
<td>1647</td>
<td>1419</td>
<td>862</td>
</tr>
</tbody>
</table>
Conclusions

- Same level of performance under Windows and Linux,
- Using cygwin provides a UNIX-like environment for the runtime,
- Replicating the disk efficiently allows an easy management of the cluster,
- Clustering under Windows is quite feasible even though Linux remains easier.