HP-MPI – a universal approach

EuroPVM/MPI’06 -- Bonn, September 19th, 2006

Henry Strauss -- Strategic Technical Consultant HPC
HP-MPI – a universal approach

• The Universal MPI for Linux, HP-UX, Tru64 and MS Windows
  − Same interface and run-time environment across all platforms
    • Same launch, header files and API for different OSs
  − Supports HP 9000 servers, HP Integrity servers, and HP ProLiant servers
  − Transparent support for multiple interconnects
    • TCP/IP; InfiniBand (all vendors); Myrinet; Quadrics, Level 5 Networks
  − Enables single executable for each OS (HP-UX, Linux, Tru64, Windows)
  − HP-UX version used by major ISVs for over a decade; now also a leader in Linux

• Greatly reduces the complexity to develop and deploy MPI applications
  − HP-MPI is the only MPI you need to support a wide choice of industry-standard processors, interconnects, and OSs
  − single qualification for each platform (CPU/OS), independent of interconnect

• MPICH Object Compatibility
  − HP-MPI is object compatible with MPICH V1.2.5 and later
  − build as shared library, dynamically link
HP-MPI Value Propositions

<table>
<thead>
<tr>
<th>Value Propositions</th>
<th>ISV &amp; End User Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portability</td>
<td>Application independence from switch, OS, CPU type ...</td>
</tr>
<tr>
<td>Robustness</td>
<td>Bulletproof run time execution; backward compatibility</td>
</tr>
<tr>
<td>Performance</td>
<td>Parity or performance gain over alternatives</td>
</tr>
<tr>
<td>Support by HP</td>
<td>Superior support to public domain or other commercial MPI libraries</td>
</tr>
<tr>
<td>Applications</td>
<td>Broad adoption ensures application availability on widest choice of platforms</td>
</tr>
</tbody>
</table>
HP-MPI Portability for developers

• Debug and Profiling Tools
  – Supports TotalView and Intel Trace Collector/Analyzer (Vampir)
  – Unique built-in diagnostic library
  – Advanced run-time error checking and analysis
    • Message signature analysis to detect type mismatches in MPI calls
    • MPI object-space corruption detection
    • Multiple buffer writes detect whether the data type specified in a receive or gather operation causes MPI to write to a user buffer more than once

• Full MPI-2 functionality
  – Dynamic processes
  – One-sided communications
  – Extended collectives
  – Thread safe
  – Current ROMIO (a portable implementation of MPI I/O)
## HP-MPI support matrix

<table>
<thead>
<tr>
<th>OS:</th>
<th>Linux</th>
<th>Windows CCS</th>
<th>HP-UX 11i &amp; 11i V2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Switch:</strong></td>
<td>Quadrics</td>
<td>Myrinet</td>
<td>Infini-Band</td>
</tr>
<tr>
<td><strong>Protocol:</strong></td>
<td>Elan3</td>
<td>Elan4</td>
<td>GM2.1</td>
</tr>
<tr>
<td>Integrity (Itanium)</td>
<td>V2.1.1 V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
</tr>
<tr>
<td>ProLiant (Opteron)</td>
<td>V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
</tr>
<tr>
<td>ProLiant (Xeon32 &amp; EM64T)</td>
<td>V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
</tr>
<tr>
<td>HP9000 (PA-RISC)</td>
<td>V2.2 V2.2.5</td>
<td>V2.2 V2.2.5</td>
<td>V2.2</td>
</tr>
</tbody>
</table>

### Notes:
- Windows CCE port will add InfiniBand and other switches
- MPI V2.0 is also available for Tru64 UNIX
- v2.2 supports 10GigE switches for Itanium/HP-UX
- ** with v2.2, all 4 IB vendor’s hardware is now supported: Voltaire, Cisco/TopSpin, SilverStorm, Mellanox

*Announced at SC’05*
HP-MPI Object Compatibility

A compatibility is documented in the MPI V2.1 & later Release Note
HP-MPI Robustness for production deployment

- HP-MPI supports important reliability features
  - resource cleanup
    - HP-MPI implements a mechanism to automatically clean-up resource when your MPI job exits (successfully or not)
  - signal propagation
    - signals sent to mpirun are propagated to all ranks
  - stdio processing
    - HP-MPI provides optional stdio processing features
      - stdin can be targeted to a particular process, or can be broadcast to every process.
      - stdout processing includes buffer control, pre-pending MPI rank numbers, and combining repeated output
Built-In Profiling

Instrumentation
- Lightweight method for cumulative runtime statistics
- Profiles for applications linked with standard HP-MPI library
- Profiles for applications linked with the thread-compliant library

Diagnostic Library
- Advanced run time error checking and analysis
- Message signature analysis detects type mismatches
- Object-space corruption detects attempts to write into objects
- Detects operations that causes MPI to write to a user buffer more than once
HP-MPI Reliability

- Extensive testing
  - Over 1000 tests in general test suite
  - Over 1000 tests in error suite
  - Application testing
  - Resource cleanup
  - Signal propagation
  - stdio processing
  - High availability
  - Exit path testing

- Built & Tested on Kernel 2.4, 2.6
- Intel Fortran 8.1, 9.0, 9.1
- Intel C 8.1., 9.1
- PathScale 2.0, 2.2, 2.3, 2.4
- Portland Group 5.2, 6.05, 6.1
- GNU C 3.2
- glibc 2.2, 2.3
HP-MPI Performance for scalability

- Performance optimization
  - shared memory optimization for intra-host communication
    - take advantage of shared memory for lower latency and higher bandwidth
  - native implementations on high performance interconnects
    - low level APIs are used to take full advantage of high performance interconnects
      - Quadrics Elan3/4
      - InfiniBand VAPI, uDAPL, OpenFabrics, IT-API for HP-UX
      - Myrinet GM/GM2, MX
  - collective routines are optimized
Applications: ISVs standardizing on HP-MPI

“One of the top reasons that we went with HP-MPI is that we’ve had a great working relationship with HP. It was a win-win for ANSYS, HP and our customers - in terms of cost, interconnects, support and performance compared to other message passing interfaces for Linux and Unix. In addition, I’ve always had great turnaround from HP in response to hardware and software issues.”
Lisa Fordanich, Senior Systems Specialist, ANSYS

“HP-MPI is an absolute godsend,” notes Keith Glassford, director of the Materials Science division at San Diego, CA-based Accelrys Software Inc. “It allows us to focus our energy and resources on doing what we’re good at, which is developing scientific and engineering software to solve customer problems.”
QA advantage of bundling HP-MPI with ISV applications!

- Chaos of MPI is caused by thinking of this library as part of the Operating Environment:

  ![Diagram showing MPI Library and applications](image)

  Appl 1 might break if System Administrator upgrades MPI library due to requirements of new version of Appl 2

- Thinking of MPI as part of the application allows ISVs to deliver a fully tested cluster-capable application:

  ![Diagram showing bundled applications](image)

  Bundling allows full version locking of what was tested by ISV and what is executed by customer
HP-MPI Roadmap

**Focus**

- **Linux/Platform/OS:**
  - XC, Itanium, X86, X86-64, RedHat V3.0, SLES 9.0, SLES 9.1, RH4 2.6 kernel

- **HP-UX/Platform/OS:**
  - Integrity Servers, PA-RISC, HP-UX 11i V2(11.23), HP-UX 11i V1(11.11)

- **Windows/Platform/OS:**
  - HP Cluster Platform 3000 & 4000, Windows CCS

**Q1 CY2005**
- Linux: V2.1.1
- HP-UX: V2.1.1
- Windows: IB Support

**Q2 CY2005**
- Linux: V2.2
- HP-UX: V2.2
- Windows: Scalability

**Q3 CY2005**
- Linux: V2.2.5
- HP-UX: V2.2.5
- Windows: Performance

**Q4 CY2005**
- Linux: V2.3
- HP-UX: V2.3
- Windows: OpenIB Partitioning

**Q1 CY2006**
- Linux: V1.0
- HP-UX: V1.0
- Windows: CCS

**Q2 CY2006**
- Linux: TCP/IP, Elan3/4, Myrinet: GM, IB
- HP-UX: PA-RISC: HF2, Integrity: IB
- Windows: TCP/IP, IB

**Q3 CY2006**
- HP-UX: Integrity: IB
- Windows: TCP/IP, IB

**Q4 CY2006**
- Linux: OpenIB
- HP-UX: IB Partitioning
- Windows: CCS

**Q4 CY2007**
- Linux: V2.2.5
- HP-UX: V2.2.5
- Windows: V2.3

**Q1 CY2007**
- Linux: V2.3
- HP-UX: V2.3
- Windows: V2.0

**Q2 CY2007**
- Linux: V2.4
- HP-UX: V2.4
- Windows: V2.1.1

**Q3 CY2007**
- Linux: V2.5
- HP-UX: V2.5
- Windows: V2.2

**Q4 CY2007**
- Linux: V2.6
- HP-UX: V2.6
- Windows: V2.3

Legend:
- **Released**
- **Committed**
- **Planned**
HP-MPI V2.2.5  10/06

• Enhanced usability of interconnect environment variables
• MPI-2 name publishing functions
• IB partitioning
• QLogic’s InfiniPath™ support
• Myrinet MX support
• Expansion of Signal Propagation
• New mpirun option for intra-host performance tuning
• Fast One-sided lock/unlock under InfiniBand VAPI
• OpenFabric support
HP-MPI for Windows V1.0

- HP-MPI for Linux port to Microsoft Windows CCS
- Capabilities of HP-MPI for Linux
  - exception: dynamic process component of MPI-2
- IBAL for IB available
  - superior performance over wsd
- Beta tested by HP-MPI ISV
- Same license/pricing structure as HP-MPI for Linux