Dealing with Disaster: Fault Tolerance in Open MPI

Josh Hursey
Open Systems Laboratory
Indiana University
jjhursey@open-mpi.org
What is there to worry about?

- **HPC Systems**
  - Growing in size and complexity
  - Increased frequency of component failure

- **HPC Applications**
  - Running longer as data sets become more complex
  - Scaling to higher degrees due to algorithmic advances.
What can we do?

- Lazy Optimism
  - Do nothing, hope for the best.
  - Scale back job submissions

- It is the Systems fault!
  The System should deal with it for me!
  - Wait for it to provide a stable, transparently fault tolerant solution.
What can we “really” do?

- Do it yourself failure handling
  - Step 1: Take legacy code base
  - Step 2: Learn about fault tolerance techniques
  - Step 3: Adapt code base for a set of fault scenarios
  - Step 4: Test and debug...

- Depend upon fault tolerance libraries and support services
  - Link with checkpoint/restart libraries
  - Use fault tolerant communication libraries
Fault Tolerance in MPI

- MPI is the *de facto* standard message passing environment for HPC applications.
  - Many implementations available

- MPI positioned to have unique knowledge of the distributed job state
  - Manage all inter-process communication
    - Must be a good steward of all data communicated
  - Detect and respond to process and node failures
  - Usually contain a distributed runtime environment
Open MPI

- Next generation MPI implementation
  Combine **best practices** from previous MPI implementations into a single **open source**, **production quality**, **MPI-2 compliant** MPI implementation.
Fault Tolerance Techniques

- Network failover & data reliability

- Rollback recovery
  - Checkpoint & restart
  - Message logging

- Replication
Fault Tolerance in Open MPI

- Network failover & data reliability
  - LA-MPI
- Rollback recovery
  - Checkpoint & restart
    - LAM/MPI
  - Message logging
- Replication
- Interactive
  - FT-MPI
Checkpoint/Restart

2 Processes using MPI

Checkpoint during failure-free execution

P₀ Fails
Restart it from last checkpoint

Rollback P₁ for consistency
Checkpoint/Restart in Open MPI

```
mpirun
```

Diagram of processes and communication in Open MPI for checkpoint/restart.
Checkpoint/Restart in Open MPI
Checkpoint/Restart in Open MPI

Checkpoint/restart system

File Management

mpirun
Checkpoint/Restart in Open MPI

Checkpoint/restart system

File Management

Snapshot Coordination
Checkpoint/Restart in Open MPI

- Checkpoint/restart system
- File Management
- Snapshot Coordination
- Intra-process Coordination
Checkpoint/Restart in Open MPI

- **Checkpoint/restart system**
- **File Management**
- **Snapshot Coordination**
- **Intra-process Coordination**
- **Inter-process Coordination**

Diagram:

- Processes: P₀, P₁, P₂, P₃, P₄, P₅
- mpirun
- Arrows indicate communication and coordination between processes.
Checkpoint/Restart in Open MPI

Checkpoint/restart system
- File Management
- Snapshot Coordination

Intra-process Coordination

Inter-process Coordination
# Checkpoint/Restart in Open MPI

<table>
<thead>
<tr>
<th>Process Type</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single process checkpoint/restart system</td>
<td>(e.g., BLCR, libckpt, Condor, ‘self’)</td>
<td>OPAL, CRS</td>
</tr>
<tr>
<td>File management &amp; movement</td>
<td>(e.g., Unix, RSH/SSH, Out-of-band comm.)</td>
<td>ORTE, FileM</td>
</tr>
<tr>
<td>Snapshot Coordinator</td>
<td>(e.g., Centralized, Replicated checkpoint servers)</td>
<td>ORTE, SnapC</td>
</tr>
<tr>
<td>Intra-process Coordinator</td>
<td>(e.g., resolve network addresses)</td>
<td>INCs</td>
</tr>
<tr>
<td>Inter-process Coordinator</td>
<td>(e.g., Coordinated, Uncoordinated, Msg. Induced)</td>
<td>OMPI, CRCP</td>
</tr>
</tbody>
</table>

---

**References**


What does this mean to me?

- **Fault Tolerance Researcher:**
  - Frameworks provide isolation
  - Benefit from progress in other areas
  - Focus on the experiment not MPI development
  - Apples-to-apples comparison of algorithms

- **Application Developer:**
  - Provide transparent fault tolerance solutions by default
  - Not required to know algorithmic details
  - Development hooks available for more fine grained control

- **Application User:**
  - Renewed focus on usable fault tolerance solutions
  - Seamless benefit from fault tolerance advancements
Demonstration

$ mpirun -np 2 --mca ft-enable cr my-app
At phase 1...
At phase 2...
At phase 3...
Demonstration

$ mpirun -np 2 --mca ft-enable cr my-app
At phase 1...
At phase 2...
At phase 3...

Slight pause in execution

$ omni-checkpoint 1234
Demonstration

$ mpirun -np 2 --mca ft-enable cr my-app
At phase 1...
At phase 2...
At phase 3...
At phase 4...

Resume execution

$ omni-checkpoint 1234
Ref: 0 global-snapshot-1234
$


Demonstration

$ mpirun -np 2 --mca ft-enable cr my-app
At phase 1...
At phase 2...
At phase 3...
At phase 4...
At phase 5...
$

Termination requested

$ omni-checkpoint 1234
Ref: 0 global-snapshot-1234
$ omni-checkpoint --term 1234
Ref: 1 global-snapshot-1234
Demonstration

```bash
$ mpirun -np 2 --mca ft-enable cr my-app
At phase 1...
At phase 2...
At phase 3...
At phase 4...
At phase 5...
$  
$ ompi-restart global-snapshot-1234
At phase 6...
At phase 7...
At phase 8...
At phase 9...
$  
$ ompi-checkpoint 1234
Ref: 0 global-snapshot-1234
$ ompi-checkpoint --term 1234
Ref: 1 global-snapshot-1234
```
Conclusions

- HPC applications must be prepared to handle system failure.
- MPI libraries are well positioned to provide (semi-)transparent fault tolerance solutions to HPC applications.
- Open MPI provides many fault tolerance solutions for modern HPC applications.
Wow! Where can I find this?

- **Network Failover & Data Reliability**
  - Scheduled to be released in **v1.2**

- **Rollback Recovery: Checkpoint/Restart**
  - Scheduled to be released in **v1.3**
  - First release will support:
    - MPI-1 standard point-to-point operations
    - Collective implementations layered over point-to-point operations
    - LAM/MPI-like coordinated checkpoint/restart
    - Asynchronous checkpoint/restart commands

- Watch the Open MPI mailing lists for updates:
  
  [http://www.open-mpi.org](http://www.open-mpi.org)
Questions
Extra Slides