A Transparent Process Migration Framework for Open MPI

Joshua Hursey
Open Systems Lab.
jjhursey@osl.iu.edu

INDIANA UNIVERSITY
PERVASIVE TECHNOLOGY INSTITUTE

Process Migration

The movement of a set of processes from one machine to another without residual dependencies

- Proactive process migration
  - Migrate when asked by a predictor (CIFTS FTB, RAS, …)
- Cluster management
  - Migrate when asked by an end user
- Load balancing
  - Migrate when a load imbalance is detected

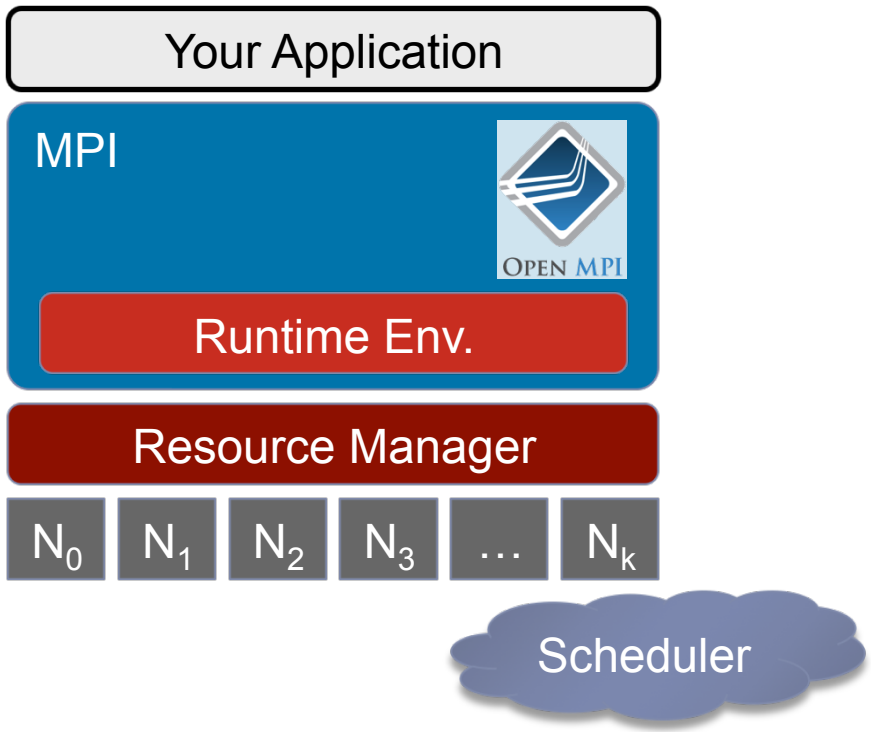
Process Migration Implementation

- Builds upon a checkpoint/restart infrastructure
  - State saved on one machine,
  - Transferred to another machine,
  - Restarted and rejoined to the computation
- Many types of process copy techniques available
  - **Eager**
  - Pre-copy
  - Lazy
  - Post-copy

Hursey, J., et. al., *The design and implementation of checkpoint/restart process fault tolerance for Open MPI*. IEEE IPDPS, 2007.

Open MPI Integration

- MCA Frameworks
  - Runtime pluggable components
- Checkpoint/Restart
  - Transparently available in Open MPI
  - Supports a wide variety of interconnects
- Process Migration
  - Added to the Runtime Env.



Recovery Service (RecoS) Framework

Policy enforcement for runtime recovery and preventative actions

Supports MPI application fault tolerance policy

- Abort:
  Terminate job
- Ignore:
  Stabilize and run without the failed process
- Migrate:
  Preventatively move processes between resources
- Restart:
  Automatically restart from the last available checkpoint

Targeting End Users

Terminal 1

```
shell$ mpirun –np 16 –am ft-enable-cr my-app
```

Terminal 2

```
shell$ ompi-migrate --off node01 123

shell$ ompi-migrate -v -x node01 --onto node02,node03 123
[localhost:01300] [  0.00 /   0.00]  Requested     - ...
[localhost:01300] [  0.00 /   0.00]  Running       - ...
[localhost:01300] [  0.00 /   0.00]  Checkpointing - ...
[localhost:01300] [  1.10 /   1.10]  Restarting    - ...
[localhost:01300] [  1.08 /   2.18]  Finished      - ...
```

Availability & Future Work

- Availability
  - Checkpoint/Restart: Available in the current v1.3
  - Process Migration - Currently under development
    - Public release - Spring 2010 (v1.5 series)
- Future work
  - Automatic recovery
  - Improved file handling
  - Alternative process copy techniques (e.g., pre-copy)
  - MPI application fault tolerance policies

Questions

Joshua Hursey
jjhursey@osl.iu.edu

www.cs.indiana.edu/~jjhursey

osl.iu.edu/research/ft

www.open-mpi.org

INDIANA UNIVERSITY
PERVASIVE TECHNOLOGY INSTITUTE

Also @ SC09

CIFTS FTB BOF - Tues., 12:15 – 1:15

CIFTS FTB Demos
Argonne Booth
Tues., 2:00 – 3:00
Wed., 2:00 – 3:00

A Resilient Runtime Environment for HPC and Internet Core Router Systems
Poster Session
Tues., 5:15 – 7:00

Open MPI BOF - Wed., 12:15 – 1:15

MPI Forum BOF - Wed., 5:30 – 7:00

Open MPI Tutorial
Indiana University Booth
Thurs., 10:00 – 12:00

OPEN MPI