Approaches for Parallel Applications Fault Tolerance

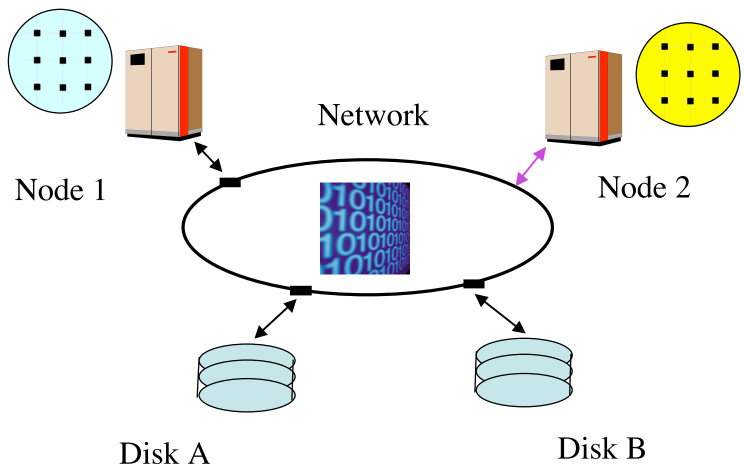
Richard L. Graham
Advanced Computing Laboratory
Los Alamos National Laboratory
LA-UR-06-6526

Overview

- Problem definition
- Introduction to the Open MPI collaboration
- Fault Recovery
  - Data transmission errors
  - Network failures
  - Process failure
- Future work

UNCLASSIFIED

Problem definition



Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

UNCLASSIFIED

Guiding Principles

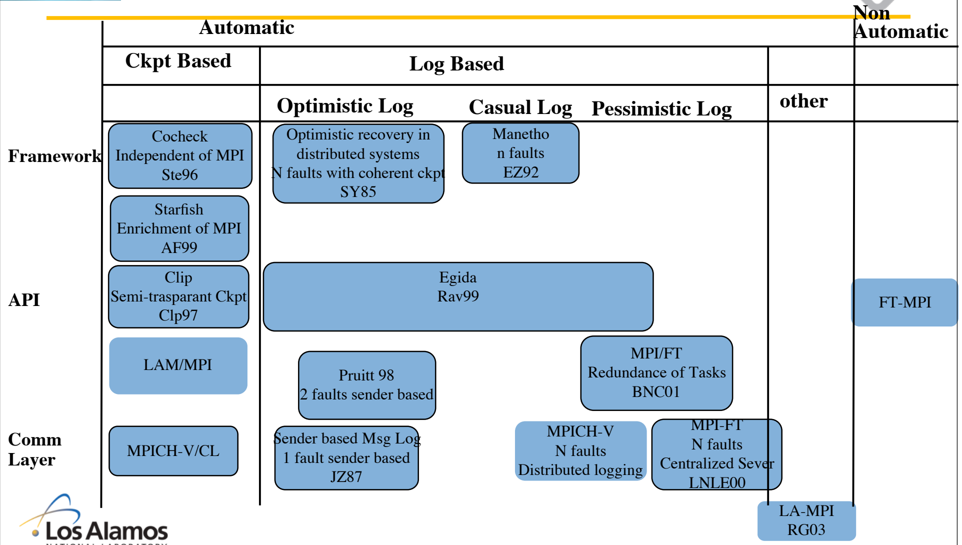
- End goal: Increase application MTBF
- Automation is desirable - more likely to be used
- No One-Solution-Fits-All
  - Hardware characteristics
  - Software characteristics
  - System complexity
  - System resources available for fault recovery
  - Performance impact on application
  - Fault characteristics of application

Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

UNCLASSIFIED

Related Work



UNCLASSIFIED

UNCLASSIFIED

Open MPI Collaboration

- The University of Tennessee
- Indiana University
- HLRS
- The University of Huston
- Sandia National Laboratory
- LANL

- Cisco
- Mellanox
- Voltaire
- Sun Microsystems
- Myricom
- IBM
- QLogic

UNCLASSIFIED

Contributors to this talk

- Tim Woodall
- Galen Shipman
- Brian Barrett
- Ralph Castain
- Jeff Squyres
- Josh Hursey
- Mitch Sukalski
- Graham Fagg
- George Bosilca

UNCLASSIFIED

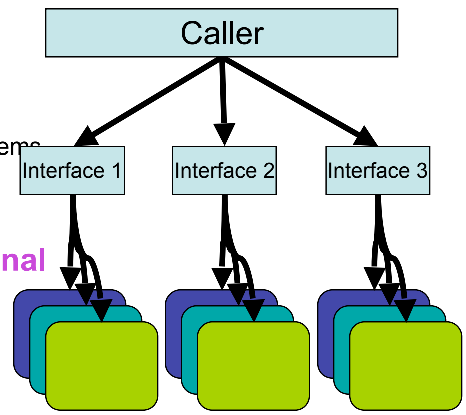
Design Features Assisting in Fault Tolerance

LA-UR-???

UNCLASSIFIED

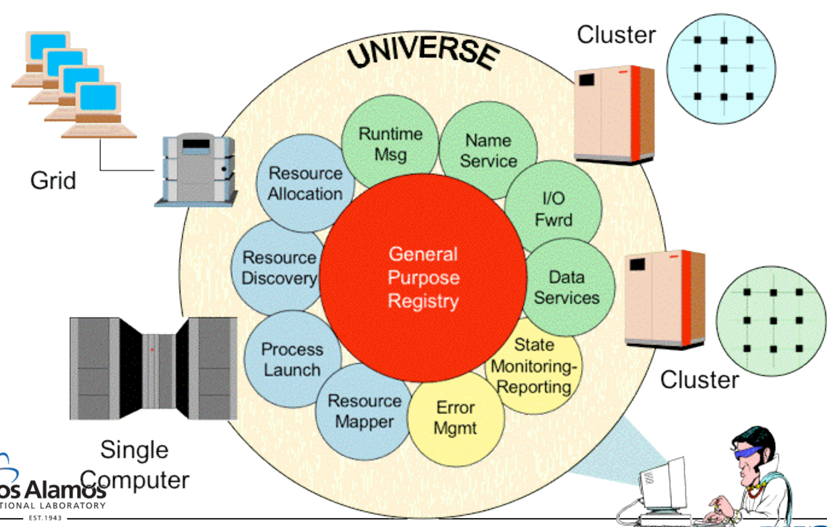
Components

- Formalized interfaces
  - Specifies “black box” implementation
  - Different implementations available at run-time
  - Can compose different systems on the fly



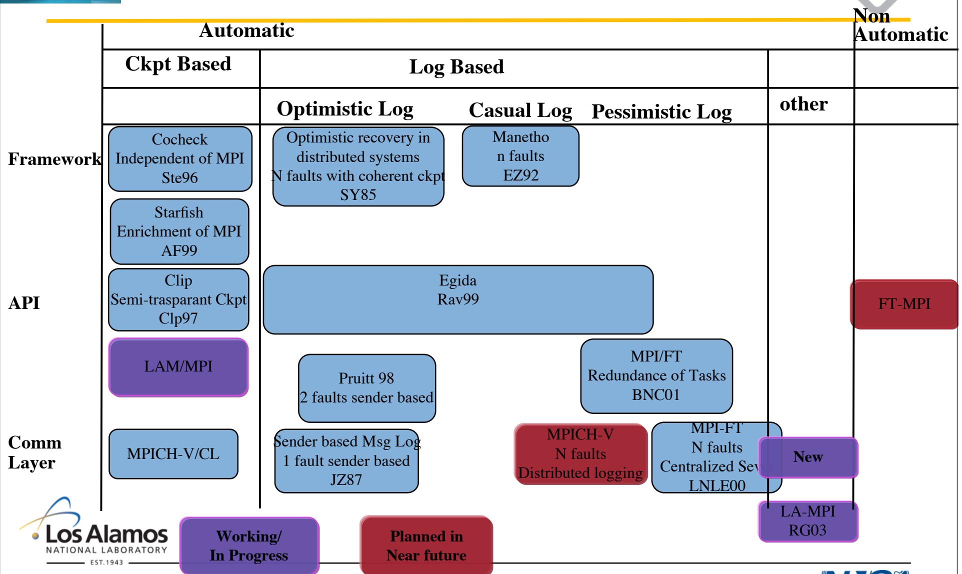
Fault Tolerance Is Optional

OpenRTE Architecture



UNCLASSIFIED

Open MPI



UNCLASSIFIED

UNCLASSIFIED

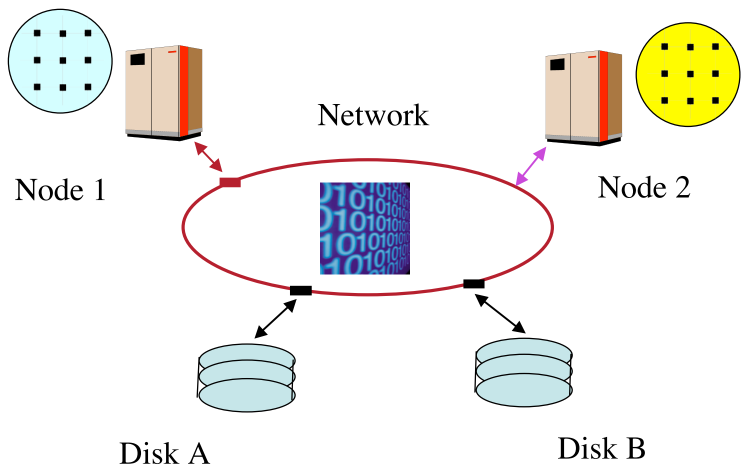
Data Transmission Errors

LA-UR-???

UNCLASSIFIED

UNCLASSIFIED

Data Transmission Errors



Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

UNCLASSIFIED

Errors Handled

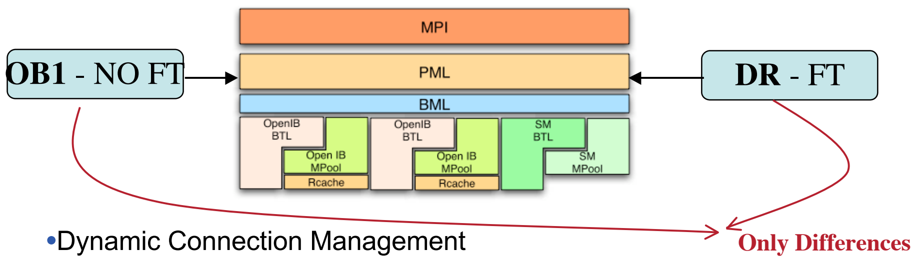
- Data corruption on the network
- Data corruption between the NIC and main-memory
- Dropped Packets

Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

General Point-To-Point Design

- Component Architecture:
  - "Plug-ins" for different capabilities (e.g. different networks)
  - Tunable run-time parameters



- Dynamic Connection Management
  - On initial send to peer connection is established via OOB connection
  - Resources allocated upon connection
- Shared Resource Allocation

Los Alamos NATIONAL LABORATORY EST. 1943

Implementation features

- Refinement of the LA-MPI implementation
- Main-memory to Main memory Checksum/CRC
  - Ack/Nack
  - Retransmit Corrupt packets
- Small packets
- Watch-dog timers
  - Retransmit timed-out packets (duplicate packet detection)
- User level protocol
  - Unpredictable time slice w/o progress thread

Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

Performance Impact: GM Ping-Pong Latency (usec)

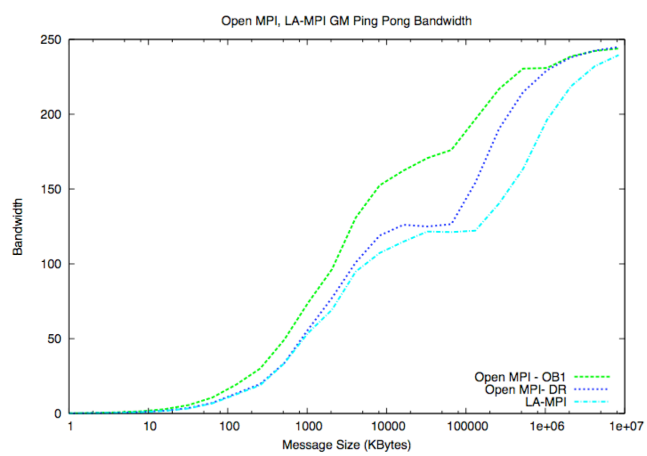
Data Size	Open MPI - OB1	Open MPI - FT	LA-MPI - FT
0 Byte	5.24	8.65	9.2
8 Byte	5.50	8.67	9.26
64 Byte	6.00	9.07	9.45
256 Byte	8.52	13.01	13.54

Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

UNCLASSIFIED

Performance Impact: GM Ping-Pong Bandwidth (MB/sec)



Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

UNCLASSIFIED

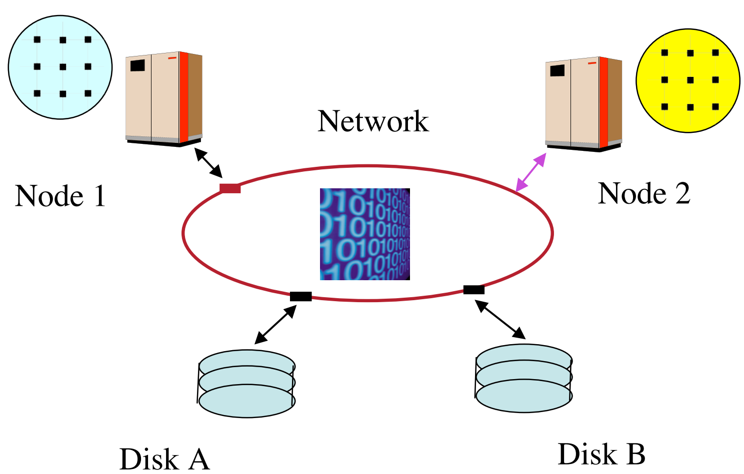
Network Failover

LA-UR-???

UNCLASSIFIED

UNCLASSIFIED

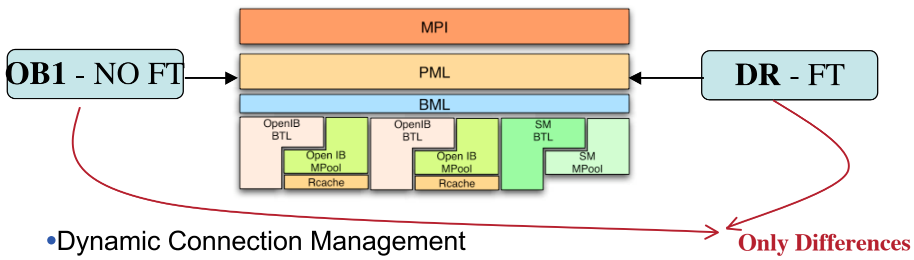
Network Device Failover



UNCLASSIFIED

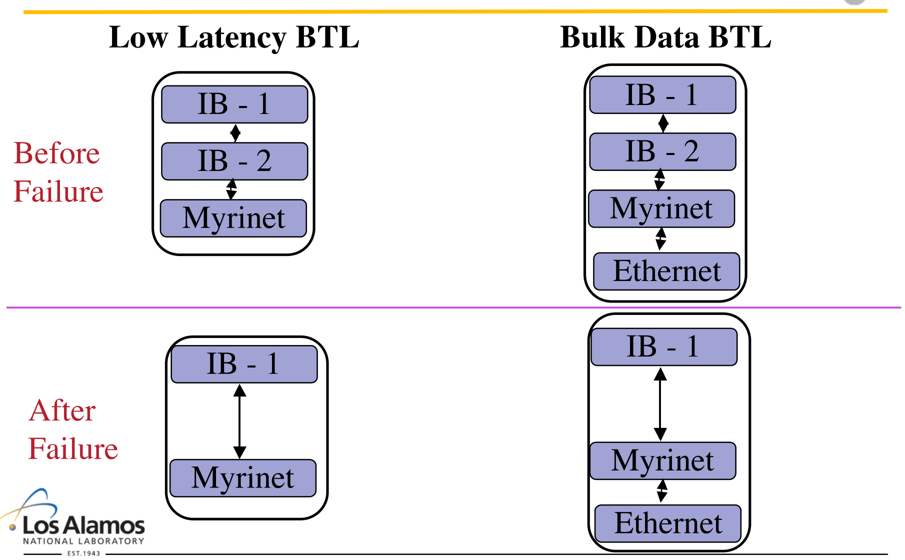
General Point-To-Point Design

- Component Architecture:
  - "Plug-ins" for different capabilities (e.g. different networks)
  - Tunable run-time parameters



- Dynamic Connection Management
  - On initial send to peer connection is established via OOB connection
  - Resources allocated upon connection
- Shared Resource Allocation

IB-2 Failure



UNCLASSIFIED

Implementation Features

- Requires error detection - more expensive
- Error detection
  - ORTE
  - Watchdog timers
- Reconnect
- Remove NIC from list of available resources

Los Alamos NATIONAL LABORATORY EST.1943

UNCLASSIFIED

UNCLASSIFIED

Device failover

gshipman@boxtop1:~/ompi-test/simple/ping

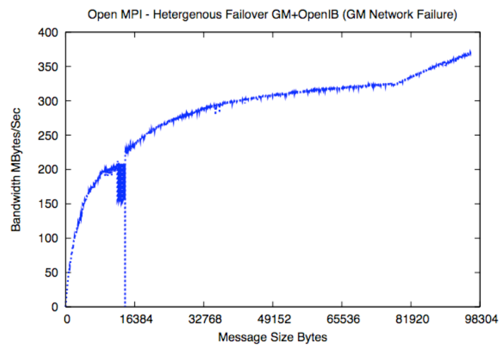
```
  0 pinged   1:    15264 bytes     96.08 uSec   158.87 MB/s
  0 pinged   1:    15296 bytes     96.22 uSec   158.97 MB/s
  0 pinged   1:    15328 bytes     72.16 uSec   212.42 MB/s
  0 pinged   1:    15360 bytes     96.77 uSec   158.72 MB/s
  0 pinged   1:    15392 bytes     96.67 uSec   159.23 MB/s
  0 pinged   1:    15424 bytes     72.76 uSec   211.98 MB/s
[boxtop2.lanl.gov:03305] ../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:83:mca_pml_dr_vfrag_wdog_timeout: failing BTL: gm
[boxtop2.lanl.gov:03305] ../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:167:mca_pml_dr_vfrag_reset: selected new BTL: openib
[boxtop1.lanl.gov:03148] ../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:83:mca_pml_dr_vfrag_wdog_timeout: failing BTL: gm
[boxtop1.lanl.gov:03148] ../../../../../ompi_europvm/ompi/mca/pml/dr/pml_dr_v
frag.c:167:mca_pml_dr_vfrag_reset: selected new BTL: openib
  0 pinged   1:    15456 bytes  52295.24 uSec     0.30 MB/s
  0 pinged   1:    15488 bytes     64.81 uSec   238.97 MB/s
  0 pinged   1:    15520 bytes     64.50 uSec   240.62 MB/s
  0 pinged   1:    15552 bytes     64.31 uSec   241.83 MB/s
  0 pinged   1:    15584 bytes     64.69 uSec   240.90 MB/s
  0 pinged   1:    15616 bytes     64.54 uSec   241.98 MB/s
  0 pinged   1:    15648 bytes     64.72 uSec   241.78 MB/s
  0 pinged   1:    15680 bytes     64.62 uSec   242.63 MB/s
  0 pinged   1:    15712 bytes     64.78 uSec   242.53 MB/s
  0 pinged   1:    15744 bytes     64.74 uSec   243.17 MB/s
```

Los Alamos NATIONAL LABORATORY EST.1943

UNCLASSIFIED

UNCLASSIFIED

NIC Failover: Ping-Pong (MB/sec)



Los Alamos NATIONAL LABORATORY EST. 1943

UNCLASSIFIED

NNSA

UNCLASSIFIED

Checkpoint/Restart

Los Alamos NATIONAL LABORATORY EST. 1943

LA-UR-???

UNCLASSIFIED

NNSA

UNCLASSIFIED

Goals

- Support a variety of checkpoint/restart protocols
  - Coordinated [First implementation]
  - Uncoordinated
- Support a variety of checkpoint/restart systems
  - Berkeley Labs Checkpoint/Restart (BLCR) [First implementation]
  - User level checkpoint/restart (self) [First implementation]
  - Others (Condor, libckpt, …)
- Internal and external checkpoint/restart request mechanisms
  - Command line tools
  - API
- Support process migration

Los Alamos NATIONAL LABORATORY EST.1943

UNCLASSIFIED

UNCLASSIFIED

Goals

- Designed to support fault tolerance research
  - Extensible set of MCA frameworks with clearly defined interfaces
- Improved interconnect support
  - tcp, self, Infiniband, Myrinet, …
- Checkpoint/restart system heterogeneity
  - The use of more than one checkpoint/restart system to form a consistent global checkpoint of an application.
- Improved user interface to support transparency and reduce complexity
  - User does not need to know which checkpoint/restart systems or protocols are being used to checkpoint or restart an application
- Attention paid to performance and scalability

Los Alamos NATIONAL LABORATORY EST.1943

UNCLASSIFIED

Architecture

- OPAL Checkpoint/Restart Service (CRS)
  - Single process checkpoint/restart system interface
- ORTE Snapshot Coordinator (SnapC)
  - Launch and monitor a distributed checkpoint/restart
  - Support checkpoint server architecture
- ORTE File Manager (FileM)
  - Distributed file management
- OMPI Checkpoint/Restart Coordination Protocol (CRCP)
  - Distributed checkpoint/restart coordination protocol interface
  - Support at least Coordinated and Uncoordinated protocols

Architecture

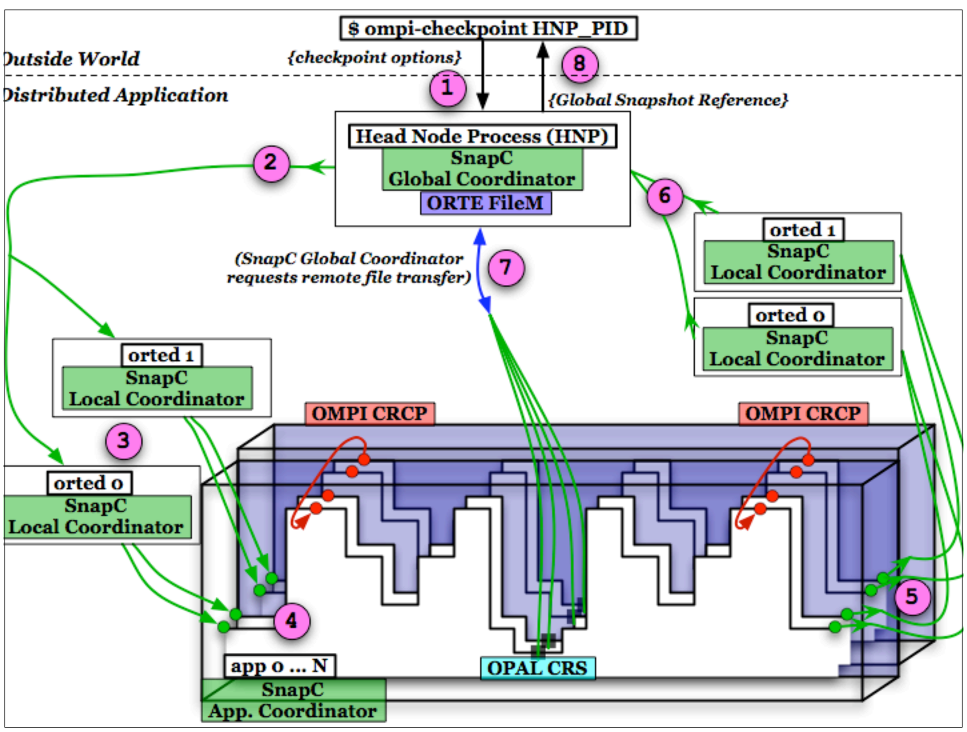
- Multilevel notification mechanism
  - Allows all layers in Open MPI to take action around a checkpoint/restart request
- MCA framework design allows for minimal changes to the Open MPI core
- Many mechanisms available for an application to choose (not) to use checkpoint/restart fault tolerance
  - Compiler option
  - Runtime option(s)

Implementation status

- OMPI CRCP framework still in development
- Checkpoint/restart protocol support:
  - Coordinated
- Checkpoint/restart system support:
  - BLCR, self
- Interconnects:
  - self
  - tcp
  - Others as time permits
- Command line tools:
  - ompi-checkpoint, ompi-restart, ompi-ps
- Current development on a branch, with plans to merge to trunk soon

Los Alamos
NATIONAL LABORATORY
EST. 1943

NNSA



Future Directions

- Refine implementations
  - Optimization
  - Vendor specific optimizations
- Process Fault Tolerance
  - Not a solved problem
  - No One-Solution-Fits-All in the small cluster to Peta-Scale systems