Designing Fault Resilient and Fault Tolerant Systems with InfiniBand

Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: panda@cse.ohio-state.edu

http://www.cse.ohio-state.edu/~panda

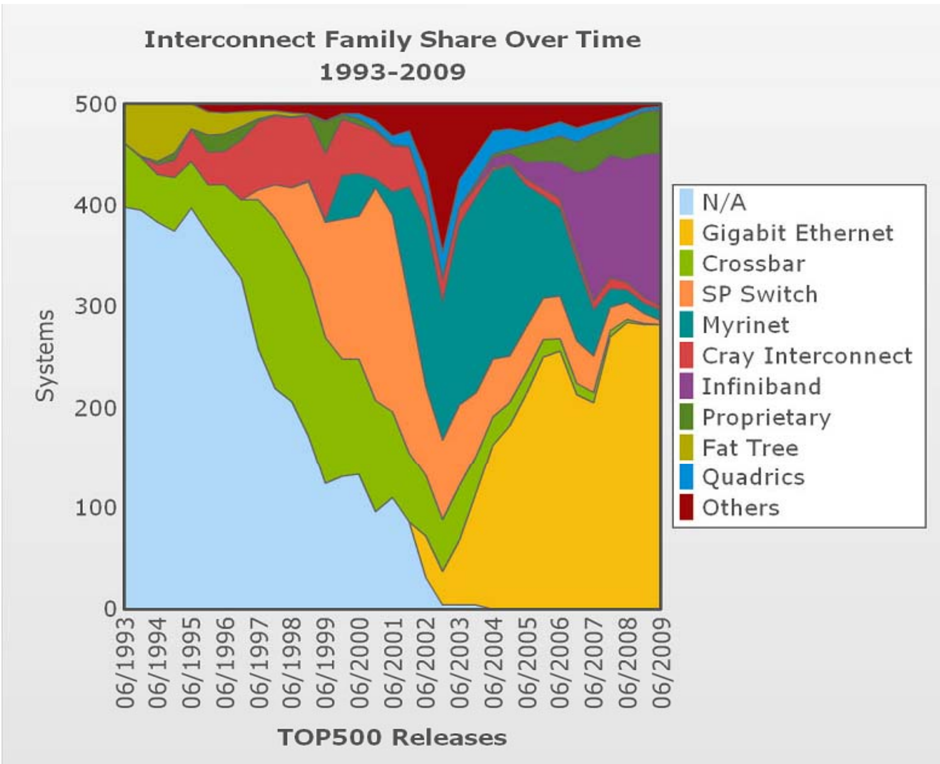
Trends for Computing Clusters in the Top 500 List

- Top 500 list of Supercomputers (www.top500.org)

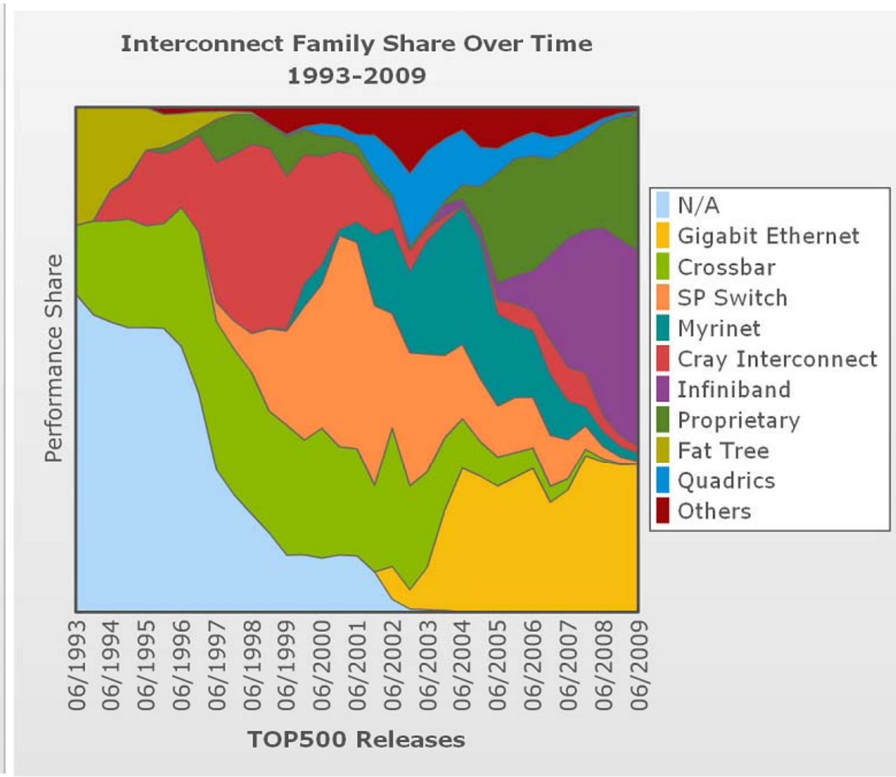
Jun. 2001: 33/500 (6.6%)	Nov. 2005: 360/500 (72.0%)
Nov. 2001: 43/500 (8.6%)	Jun. 2006: 364/500 (72.8%)
Jun. 2002: 80/500 (16%)	Nov. 2006: 361/500 (72.2%)
Nov. 2002: 93/500 (18.6%)	Jun. 2007: 373/500 (74.6%)
Jun. 2003: 149/500 (29.8%)	Nov. 2007: 406/500 (81.2%)
Nov. 2003: 208/500 (41.6%)	Jun. 2008: 400/500 (80.0%)
Jun. 2004: 291/500 (58.2%)	Nov. 2008: 410/500 (82.0%)
Nov. 2004: 294/500 (58.8%)	Jun. 2009: 410/500 (82.0%)
Jun. 2005: 304/500 (60.8%)	Nov. 2009: To be announced

InfiniBand in the Top500

Systems



Performance



Percentage share of InfiniBand is steadily increasing

Large-scale InfiniBand Installations

- 151 IB clusters (30.2%) in the June '09 TOP500 list (www.top500.org)
- Installations in the Top 30 (15 of them):

129,600 cores (RoadRunner) at LANL (1^{st})	12,288 cores at GENCI-CINES, France (20^{th})
51,200 cores (Pleiades) at NASA Ames (4^{th})	8,320 cores in UK (25^{th})
62,976 cores (Ranger) at TACC (8^{th})	8,320 cores in UK (26^{th})
26,304 cores (Juropa) at TACC (10^{th})	8,064 cores (DKRZ) in Germany (27^{th})
30,720 cores (Dawning) at Shanghai (15^{th})	12,032 cores at JAXA, Japan (28^{th})
14,336 cores at New Mexico (17^{th})	10,240 cores at TEP, France (29^{th})
14,384 cores at Tata CRL, India (18^{th})	13,728 cores in Sweden (30^{th})
18,224 cores at LLNL (19^{th})	***More are getting installed !***

MVAPICH/MVAPICH2 Software

- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 975 organizations in 51 countries
  - More than 34,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 8^{th} ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
  - http://mvapich.cse.ohio-state.edu/

Presentation Overview

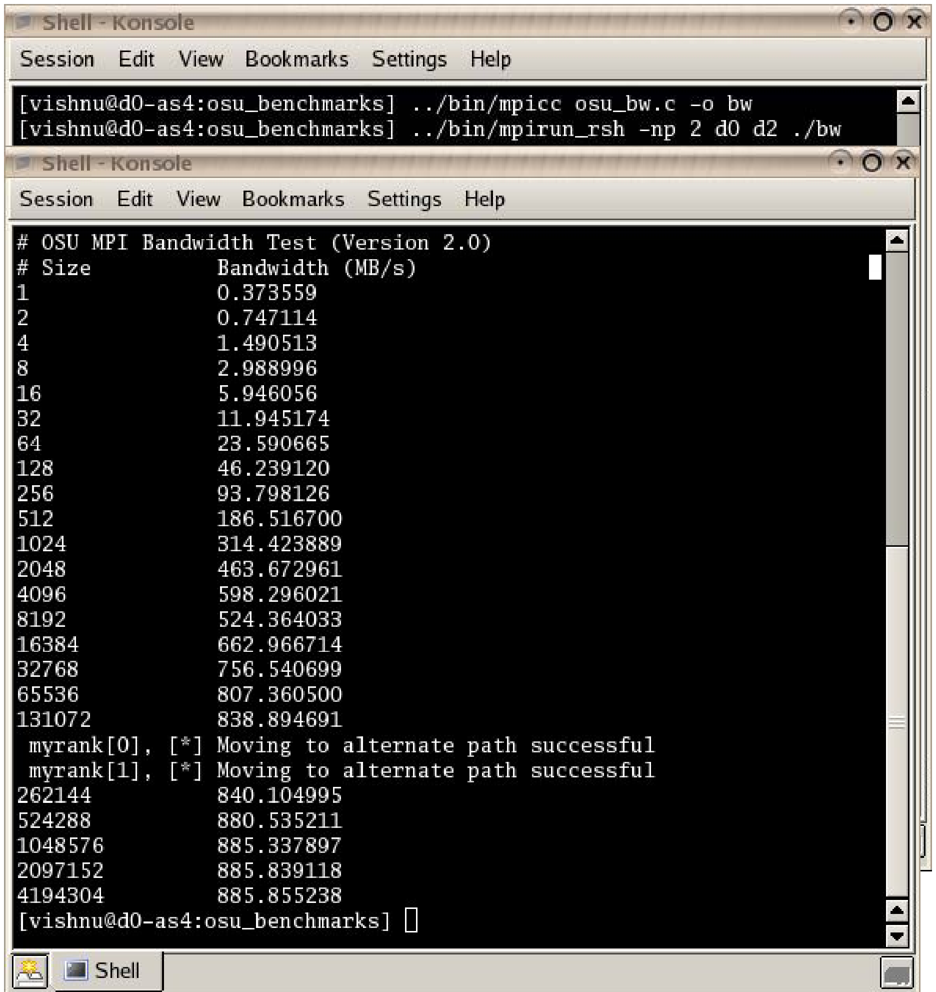
- Network-Level Fault Tolerance/Resiliency in MVAPICH/MVAPICH2
  - Automatic Path Migration (APM)
  - Mem-to-Mem Reliability
  - Resiliency to Network Failures
- Process-Level Fault Tolerance in MVAPICH2
  - BLCR-based systems-level Checkpoint-Restart (CR)
  - Enhancing CR Performance with I/O Aggregation
  - Fault-Tolerant Backplane (FTB) over InfiniBand
  - Pro-active Migration with Job-Suspend and Resume
- Virtualization and Fast Migration with InfiniBand
- Conclusion and Q&A

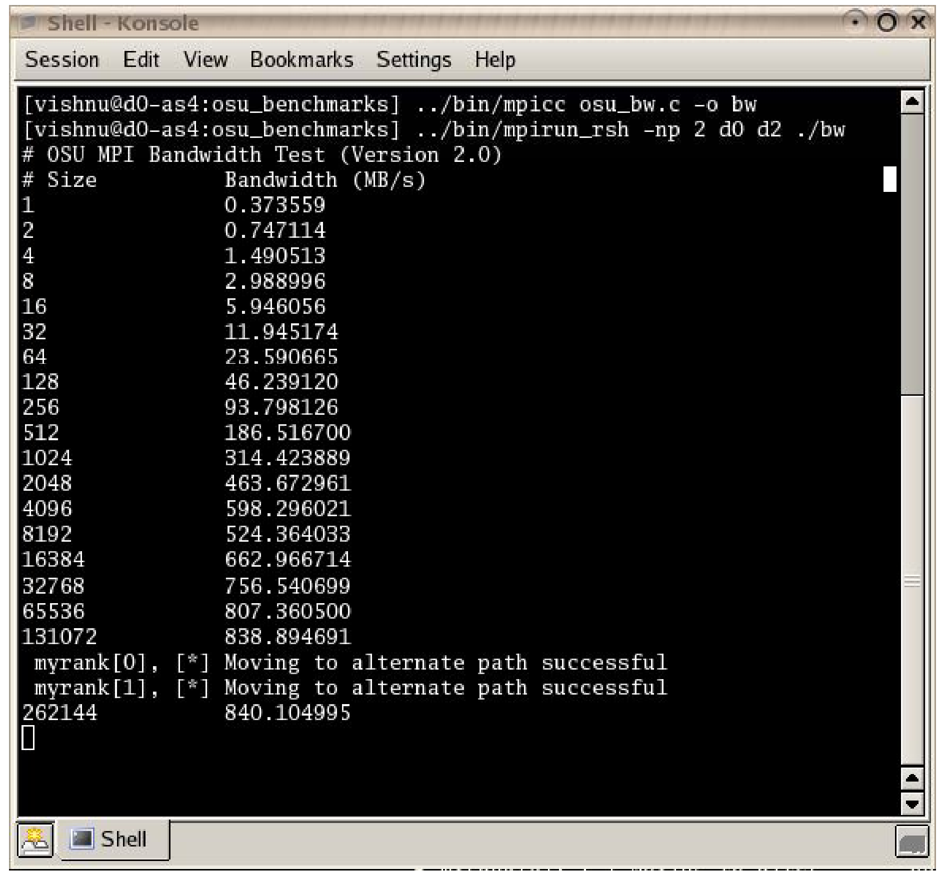
Network-Level Fault Tolerance with Automatic Path Migration (APM)

- Utilizes Redundant Communication Paths
  - Multiple Ports
  - LMC (LID Mask Control)
- Enables migrating connections to a different path
- Reliability guarantees for Service Type Maintained during Migration
- Support in both MVAPICH and MVAPICH2

A. Vishnu, A. Mamidala, S. Narravula and D. K. Panda, Automatic Path Migration over InfiniBand: Early Experiences, Third International Workshop on System Management Techniques, Processes, and Services, held in conjunction with IPDPS '07, March 2007.

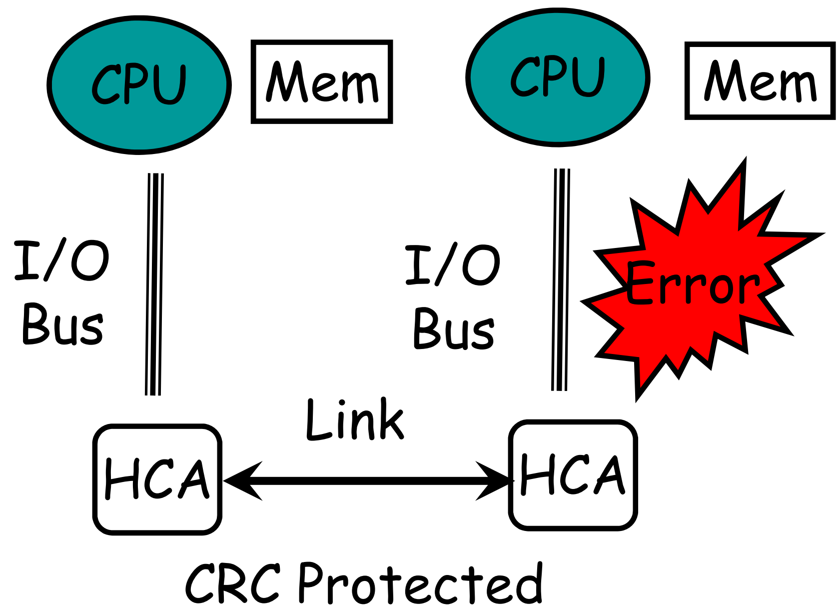
Screenshots: APM with OSU Bandwidth test





Memory-to-Memory Reliability

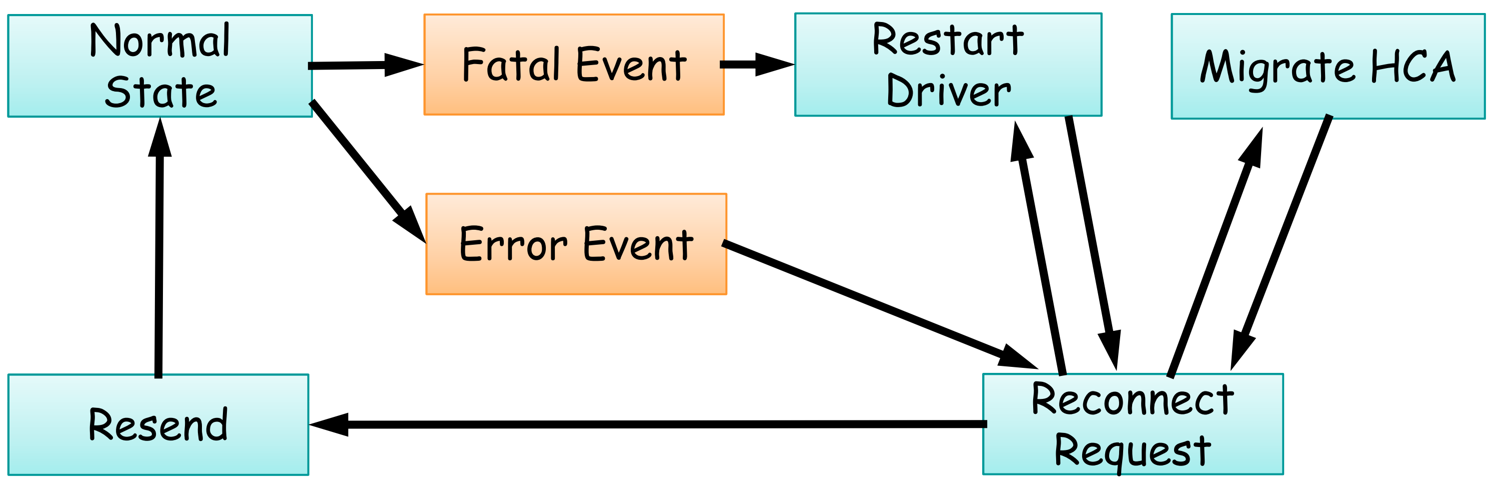
- InfiniBand enforces HCA to HCA reliability using CRC
- No check to see if data is transmitted reliably over I/O Bus
- In different situations (high-altitudes or in hotter climates), error rate increases sharply
- MVAPICH uses CRC-32 bit algorithm to ensure safe message delivery



Network-Level Resiliency

- Protection against various network failures
  - Switch reboot/failure
  - HCA failure
  - Severe congestion
- Can we stall a job instead of aborting it while the failed component is fixed
- Being designed and developed together with Mellanox
- Will be available in MVAPICH 1.2

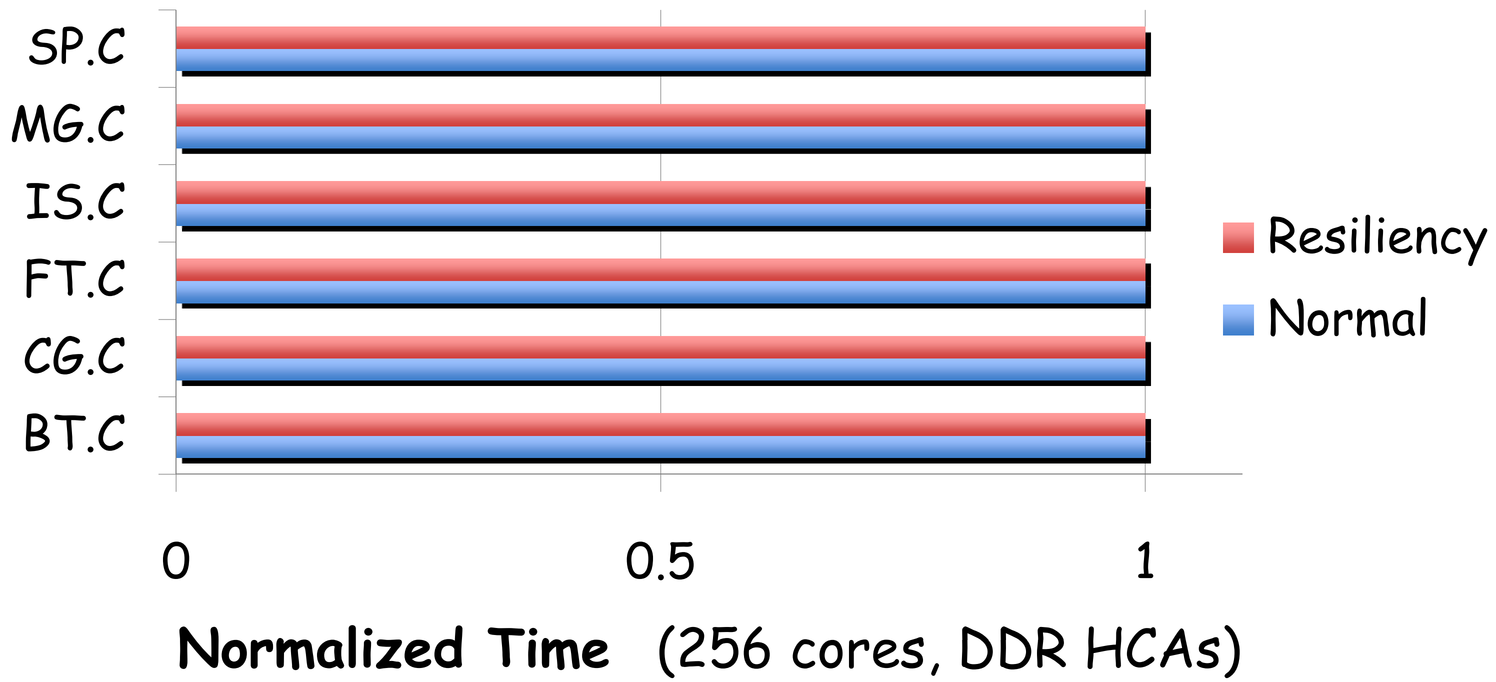
Network-Level Resiliency Flow



- Recover from a fatal HCA failure (first restart, then migrate)
- Recover from errors (intermittent switch failure, etc)
- Configurable retry settings

This differs from Automatic Path Migration (APM) which can only recover from a single error event (non-fatal) and cannot wait for a specified time to retry

Performance Impact



No performance change for application kernels

Presentation Overview

- Network-Level Fault Tolerance/Resiliency in MVAPICH/MVAPICH2
  - Automatic Path Migration (APM)
  - Mem-to-Mem Reliability
  - Resiliency to Network Failures
- Process-Level Fault Tolerance in MVAPICH2
  - BLCR-based systems-level Checkpoint-Restart (CR)
  - Enhancing CR Performance with I/O Aggregation
  - Fault-Tolerant Backplane (FTB) over InfiniBand
  - Pro-active Migration with Job-Suspend and Resume
- Virtualization and Fast Migration with InfiniBand
- Conclusion and Q&A

Checkpoint/Restart Support for MVAPICH2

- Process-level Fault Tolerance
  - User-transparent, system-level checkpointing
  - Based on BLCR from LBNL to take coordinated checkpoints of entire program, including front end and individual processes
  - Designed novel schemes to
    - Coordinate all MPI processes to drain all in flight messages in IB connections
    - Store communication state and buffers, etc. while taking checkpoint
    - Restarting from the checkpoint
- Available for the last two years with MVAPICH2 and is being used by many organizations
- Systems-level checkpoint can also be initiated from the application

Enhancing CR Performance

- Checkpoint time is dominated by writing the files to storage
- Multi-core systems are emerging
  - 8/16-cores per node
  - a lot of data needs to be written
  - affects scalability
- Can we reduce checkpoint time with I/O aggregation of short messages?

Profiled Results

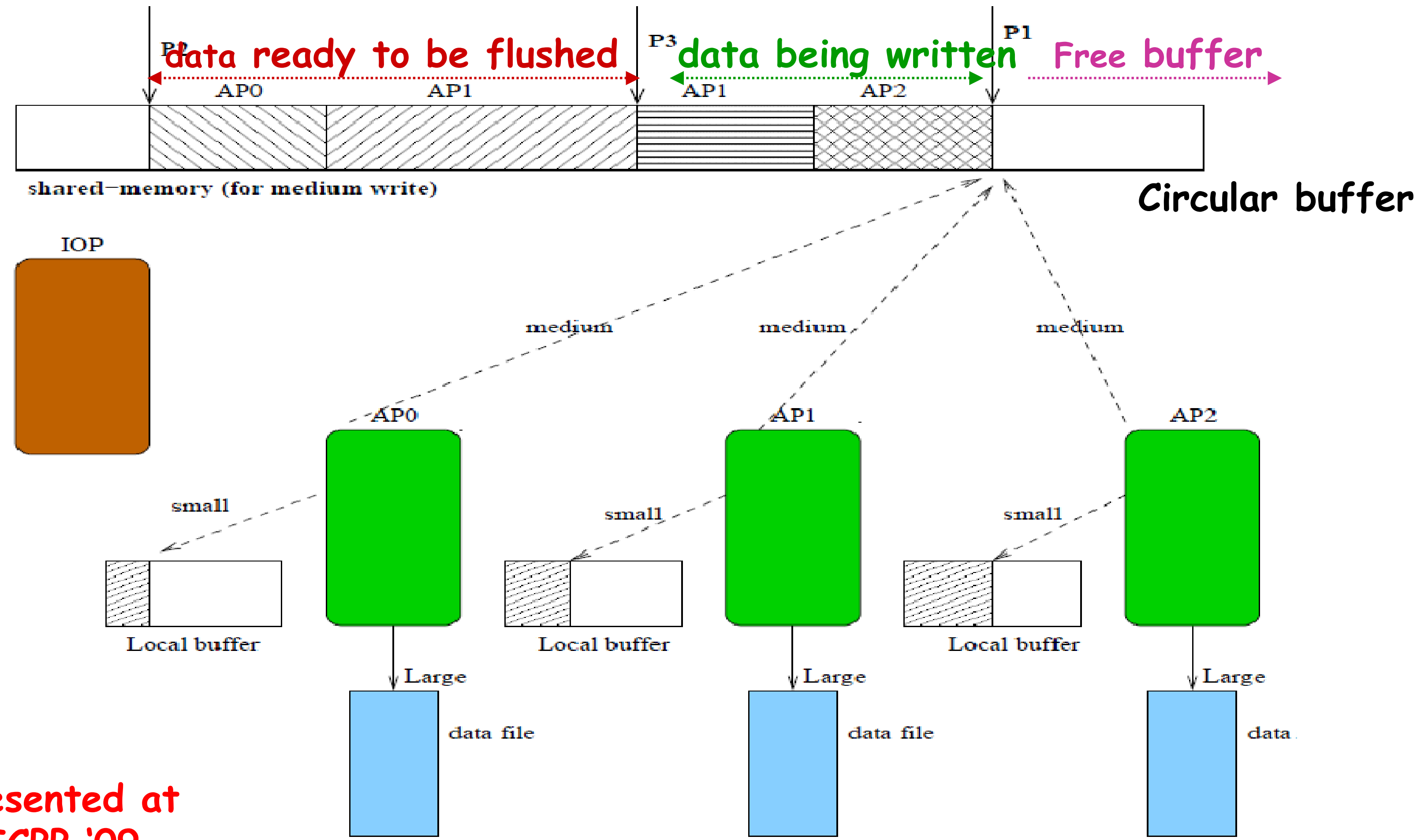
Basic checkpoint writing information
(class C, 64 processes, 8 processes/node)

	LU	BT	SP	CG
Time for one check-point(seconds)	7.6	11.3	10.3	7.1
Total data size(MB) per node	184.0	320.0	316.0	163.2
Number of VFS write per process	975	1057	1367	820
Total number of VFS writes per node	7800	8456	10936	6560

Checkpoint Writing Profile for LU.C.64

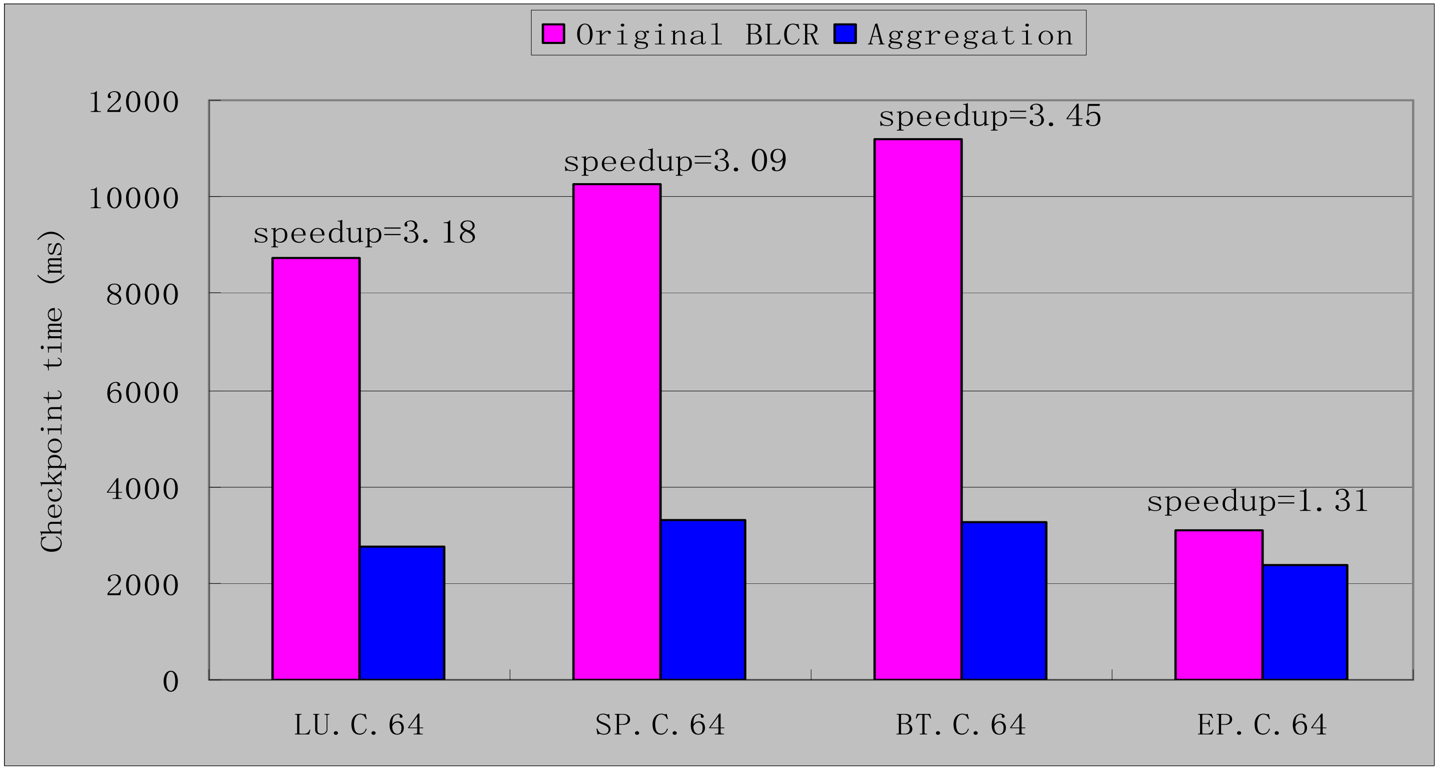
	% of Writes	% of Data	% of Time
0-64	50.86	0.04	0.17
64-256	0.61	0.00	0.00
256-1K	0.25	0.01	0.00
1K-4K	9.46	1.53	0.01
4K-16K	36.49	11.36	44.66
16K-64K	0.74	0.77	6.55
64K-256K	0.49	3.79	11.80
256K-512K	0.25	3.58	1.75
512K-1M	0.61	17.72	14.72
> 1M	0.25	61.21	20.35

Write-Aggregation Design



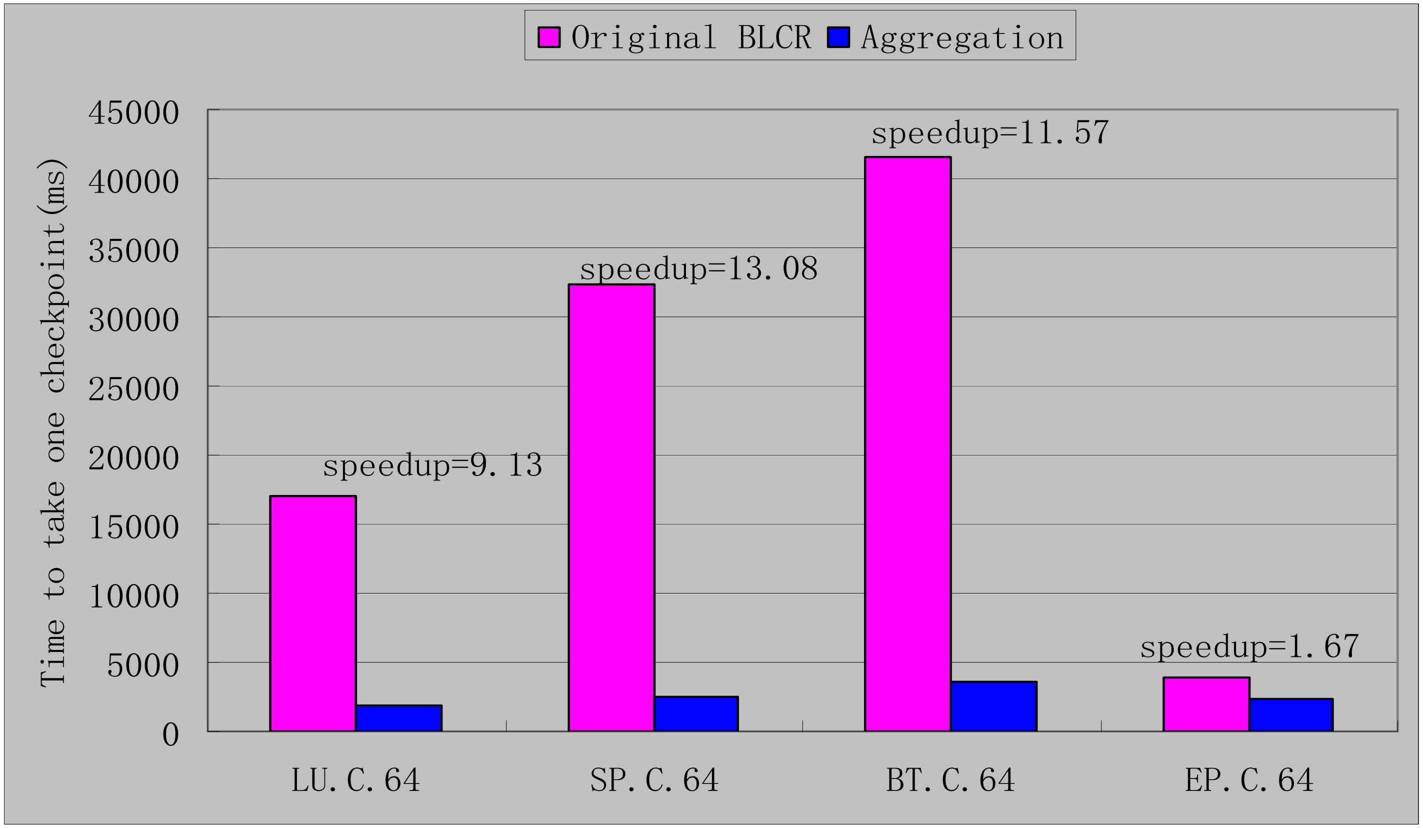
Presented at ICPP '09

Time to Take One Checkpoint – 64 processes (8 nodes with 8 cores)



- 64 MPI processes on 8 nodes, 8 processes/node
- Checkpoint data is written to local disk files

Time to Take One Checkpoint – 64 processes (4 nodes with 16 cores)



- 64 MPI processes on 4 nodes, 16 processes/node
- Checkpoint data is written to local disk files

Will be available in the Next MVAPICH2 Release

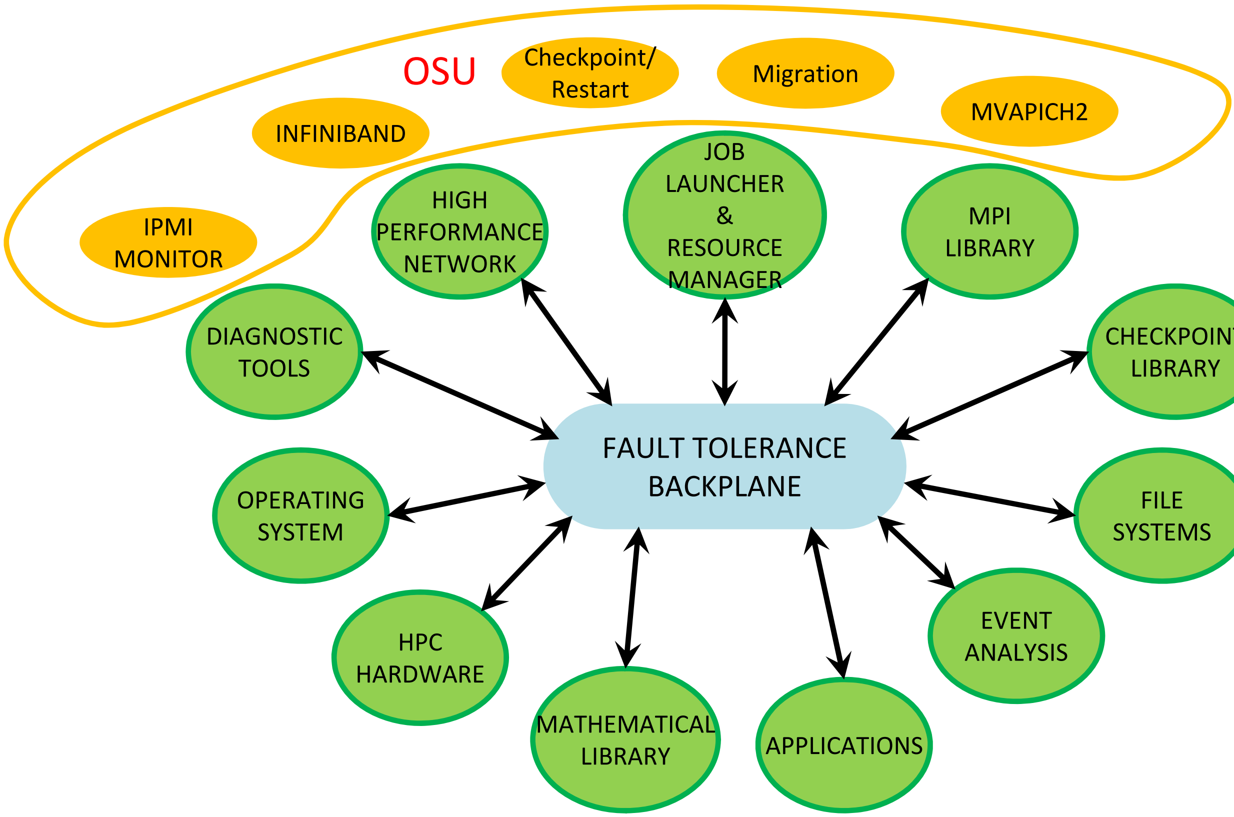
Presentation Overview

- Network-Level Fault Tolerance/Resiliency in MVAPICH/MVAPICH2
  - Automatic Path Migration (APM)
  - Mem-to-Mem Reliability
  - Resiliency to Network Failures
- Process-Level Fault Tolerance in MVAPICH2
  - BLCR-based systems-level Checkpoint-Restart (CR)
  - Enhancing CR Performance with I/O Aggregation
  - Fault-Tolerant Backplane (FTB) over InfiniBand
  - Pro-active Migration with Job-Suspend and Resume
- Virtualization and Fast Migration with InfiniBand
- Conclusion and Q&A

Coordinated Infrastructure for Fault Tolerant Systems (CIFTS) Framework

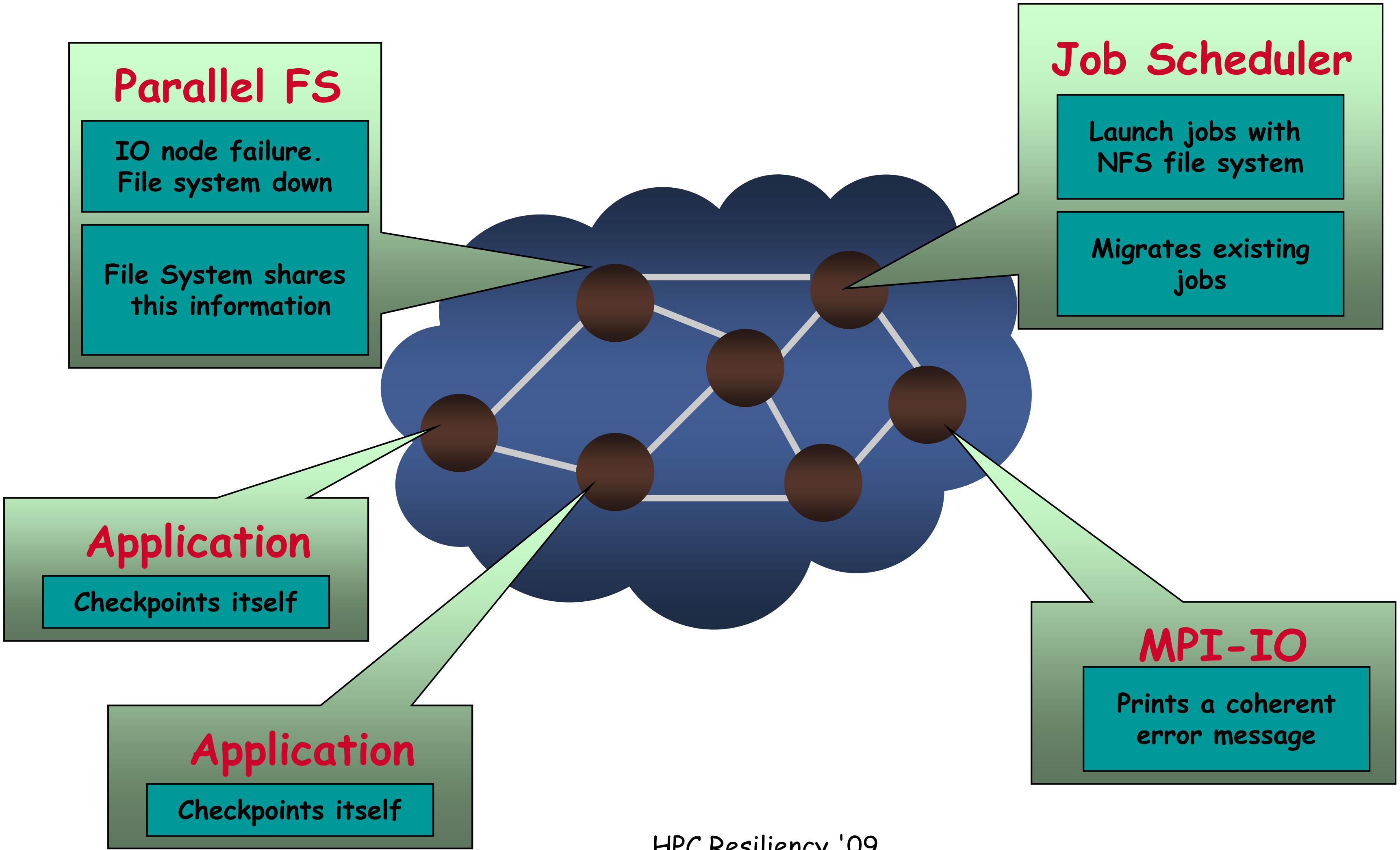
Collaborators

- ANL
- OSU
- ORNL
- LBNL
- IU
- UT

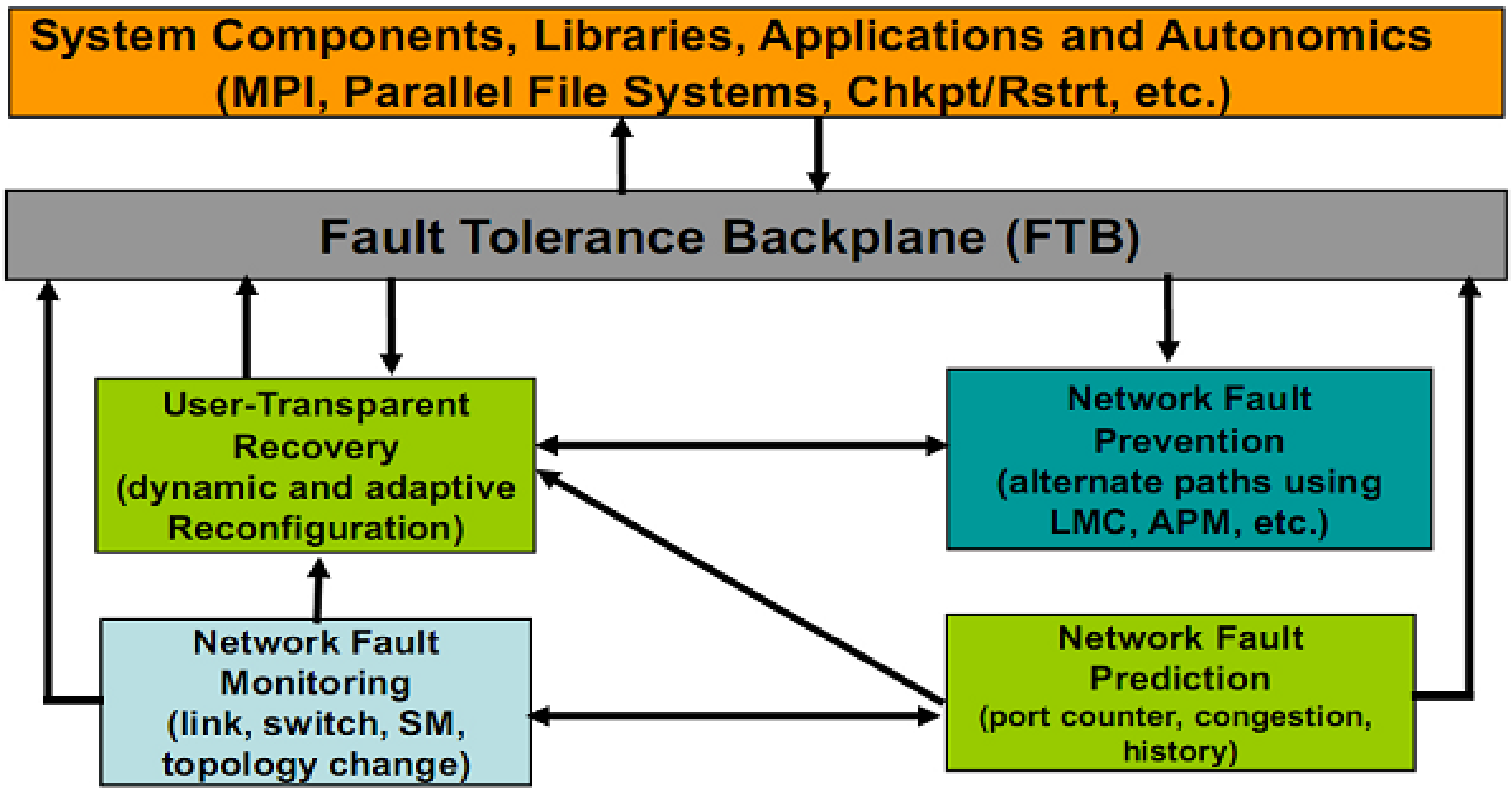


http://www.mcs.anl.gov/research/cifts/

CIFTS - Usage Scenario

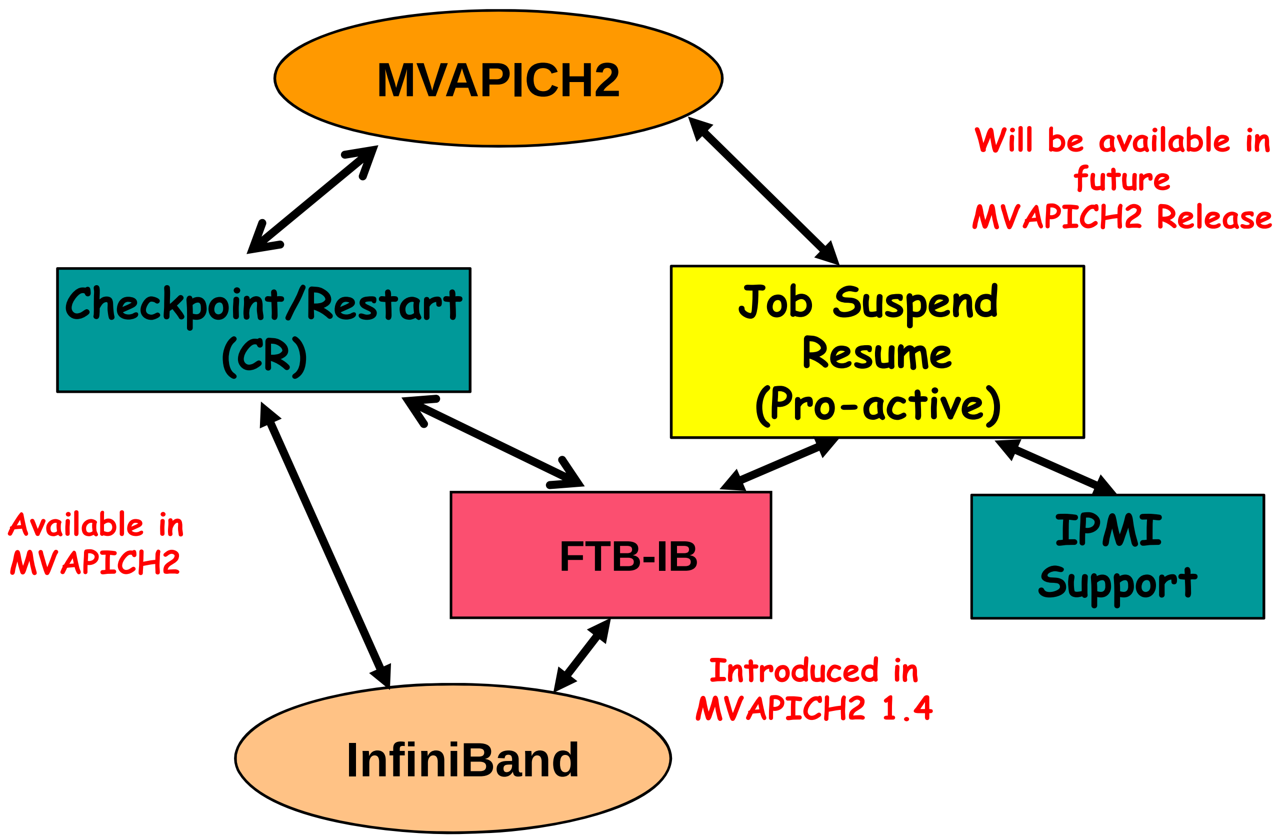


FTB-IB (Overall Plan)



FTB-IB 1.0 version is available for download:
http://nowlab.cse.ohio-state.edu/projects/ftb-ib/index.html

Comprehensive Solution (Putting All Components Together)

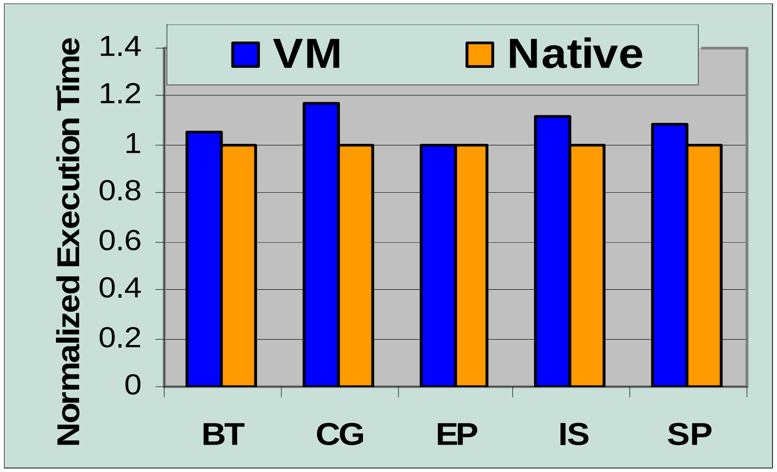


Presentation Overview

- Network-Level Fault Tolerance/Resiliency in MVAPICH/MVAPICH2
  - Automatic Path Migration (APM)
  - Mem-to-Mem Reliability
  - Resiliency to Network Failures
- Process-Level Fault Tolerance in MVAPICH2
  - BLCR-based systems-level Checkpoint-Restart (CR)
  - Enhancing CR Performance with I/O Aggregation
  - Fault-Tolerant Backplane (FTB) over InfiniBand
  - Pro-active Migration with Job-Suspend and Resume
- Virtualization and Fast Migration with InfiniBand
- Conclusion and Q&A

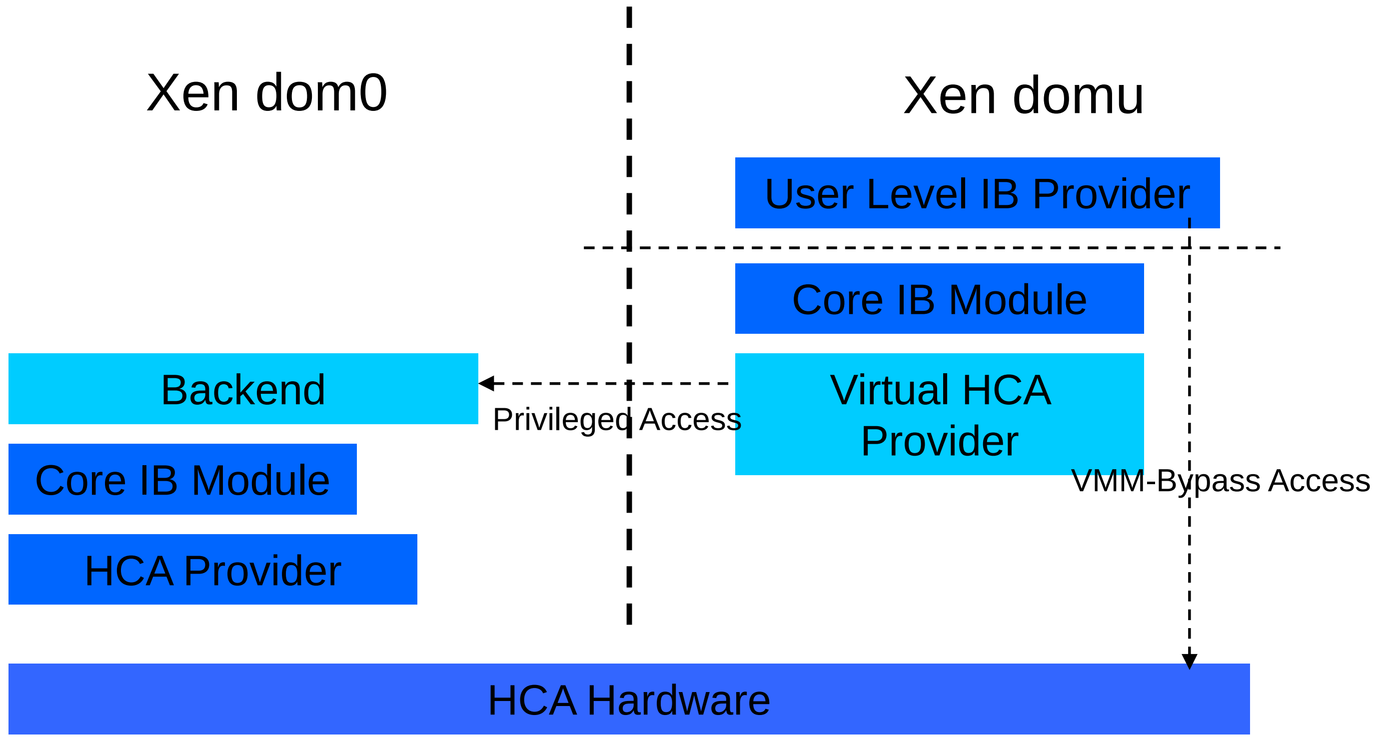
Problem with Current I/O Virtualization

- Performance
  - Every I/O operation involves the VMM and/or another VM
  - VMM may become a performance bottleneck
  - Using a special VM results in expensive context switches between different VMs
  - Undesirable for high end systems, especially those used in high performance computing (HPC)



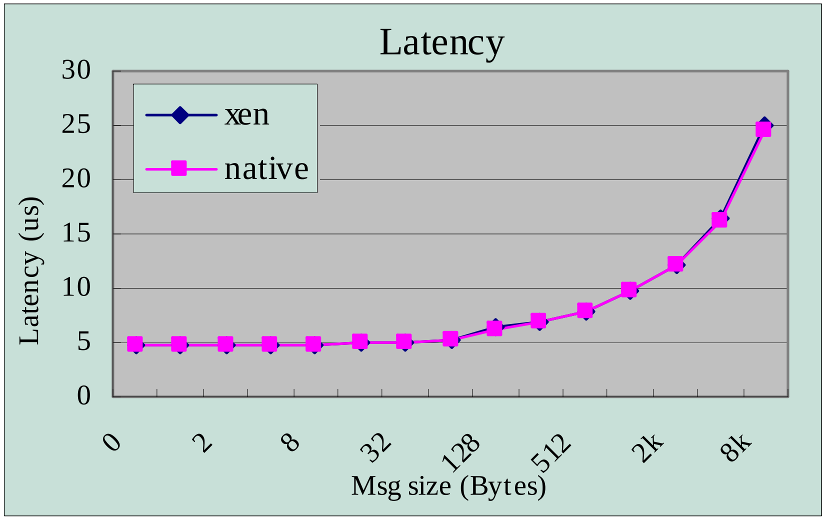
	Dom0	VMM	DomU
CG	16.6%	10.7%	72.7%
IS	18.1%	13.1%	68.8%
EP	00.6%	00.3%	99.0%
BT	06.1%	04.0%	89.9%
SP	09.7%	06.5%	83.8%

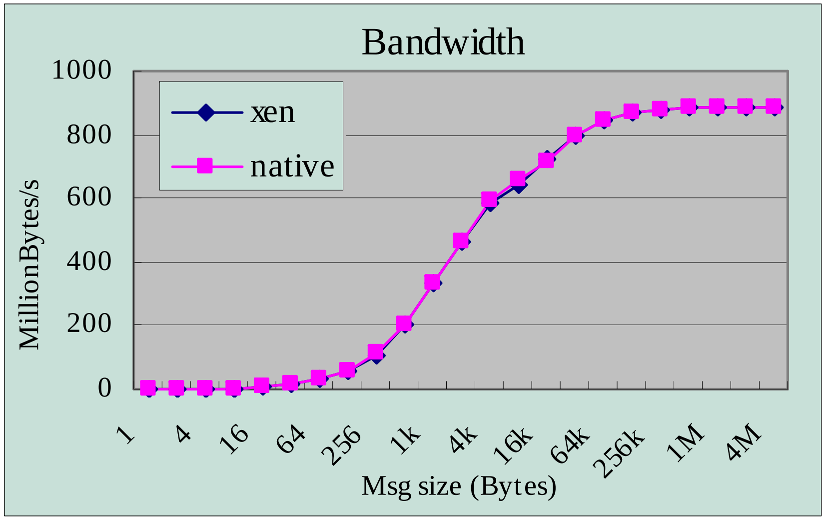
Xen-IB and VMM-Bypass



J. Liu, W. Huang, B. Abali, D. K. Panda. High Performance VMM-Bypass I/O in Virtual Machines, *USENIX Annual Technical Conference (USENIX'06)*, May, 2006

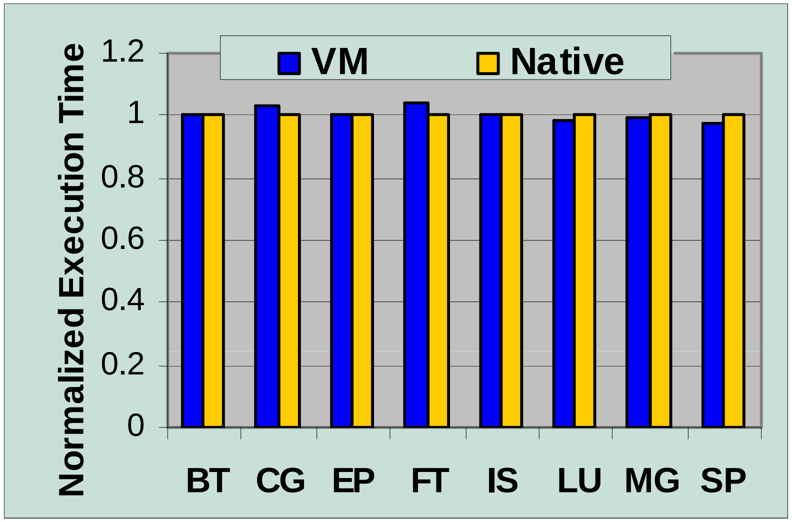
MPI Latency and Bandwidth (MVAPICH)





- Only VMM Bypass operations are used
- Xen-IB performs similar to native InfiniBand
- Numbers taken with MVAPICH

HPC Benchmarks (NAS)



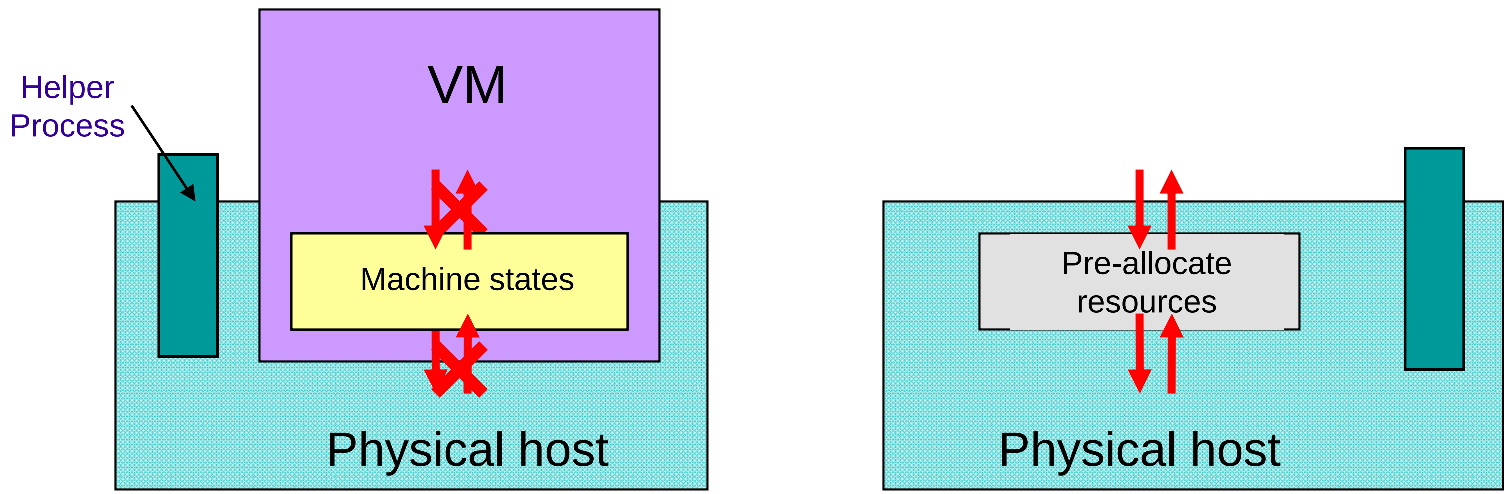
	Dom0	VMM	DomU
BT	0.4%	0.2%	99.4%
CG	0.6%	0.3%	99.0%
EP	0.6%	0.3%	99.3%
FT	1.6%	0.5%	97.9%
IS	3.6%	1.9%	94.5%
LU	0.6%	0.3%	99.0%
MG	1.8%	1.0%	97.3%
SP	0.3%	0.1%	99.6%

- NAS Parallel Benchmarks achieve similar performance in VM and native environment (8x2)

–J. Liu, W. Huang, B. Abali, D. K. Panda. High Performance VMM-Bypass I/O in Virtual Machines, *USENIX Annual Technical Conference (USENIX'06)*, May, 2006

–W. Huang, J. Liu, B. Abali, D. K. Panda. A Case for High Performance Computing with Virtual Machines, *ACM International Conference on Supercomputing (ICS '06)*, June, 2006

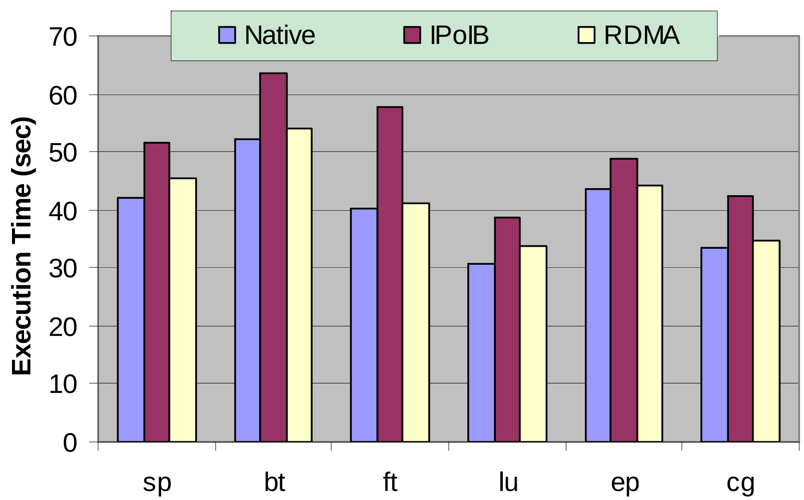
Optimizing VM migration through RDMA

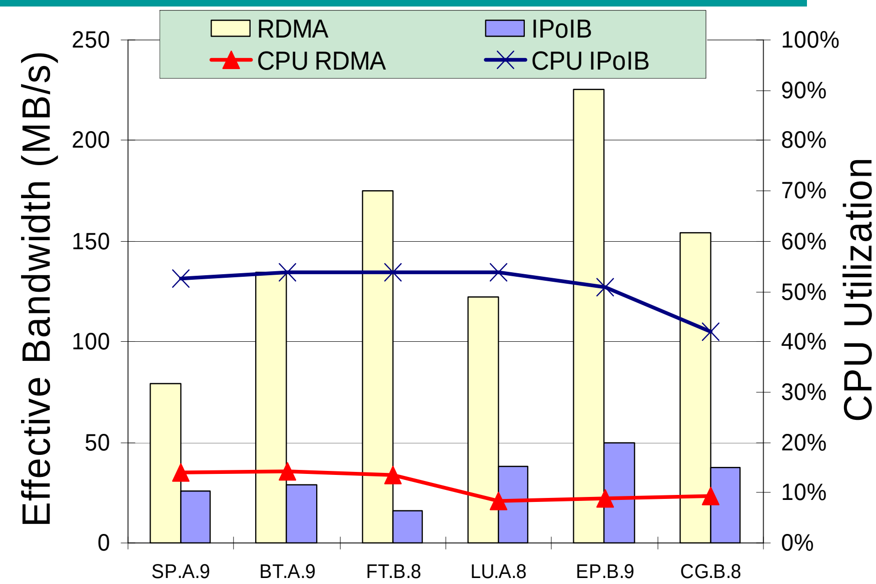


Live VM migration:

- Step 1: Pre-allocate resource on target host
- Step 2: Pre-copy machine states for multiple iterations
- Step 3: Suspend VM and copy the latest updates to machine states
- Step 4: Restart VM on the new host

Fast Migration over RDMA





- Migration overhead with IPoIB drastically increases
- RDMA achieves higher migration performance with less CPU usage

W. Huang, Q. Gao, J. Liu, D. K. Panda. High Performance Virtual Machine Migration with RDMA over Modern Interconnects. *IEEE Conference on Cluster Computing (Cluster'07),* September 2007 (Best Paper Award)

Xen-IB Software

- Initially designed jointly with IBM
- Taken up by Novell later on
- Available from OFED and Mellanox sites
- Integration with MVAPICH2 and other components are planned in future

Summary and Conclusions

- Fault-tolerance and resiliency issues are becoming extremely critical for next generation Exascale systems
- InfiniBand is an emerging interconnect which provides basic functionalities for fault-tolerance at the network-level
- Presented how InfiniBand features can be used at the MPI layer to provide fault-tolerance and resilience
- Presented expanded solutions using virtualization
- Many open research challenges needing novel solutions for fault resiliency and fault tolerance in next generation Exascale systems

Web Pointers

MVAPICH

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu/

E-mail: panda@cse.ohio-state.edu