Interconnect Agnostic Checkpoint/Restart in Open MPI

Joshua Hursey, Timothy I. Mattox, Andrew Lumsdaine
Indiana University
Open Systems Laboratory
{jjhursey,timattox,lums}@osl.iu.edu

HPDC - June 2009

Fault Tolerance

As HPC applications *run longer* and/or *scale further* the incorporation of fault tolerance techniques becomes a necessity.

MPI is the most popular parallel programming model for HPC

Checkpoint/Restart is the most popular fault tolerance technique

Checkpoint/Restart Components

Checkpoint/Restart System (CRS)

Capture the state of a running ***process***

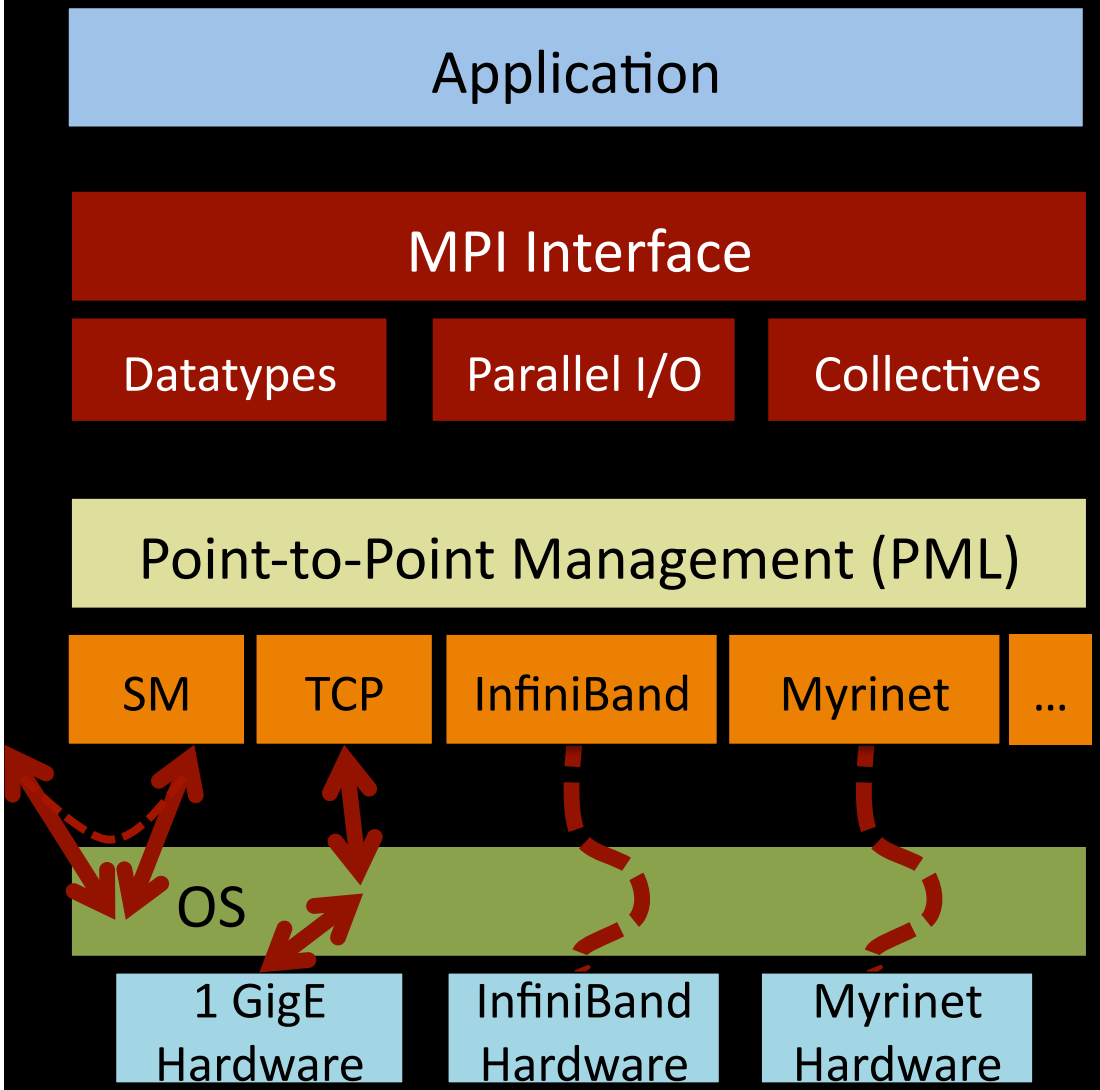
E.g., BLCR, Condor, Chpox, Libckpt, …

Coordination Protocol (CRCP)

Capture the state of the ***network connections***

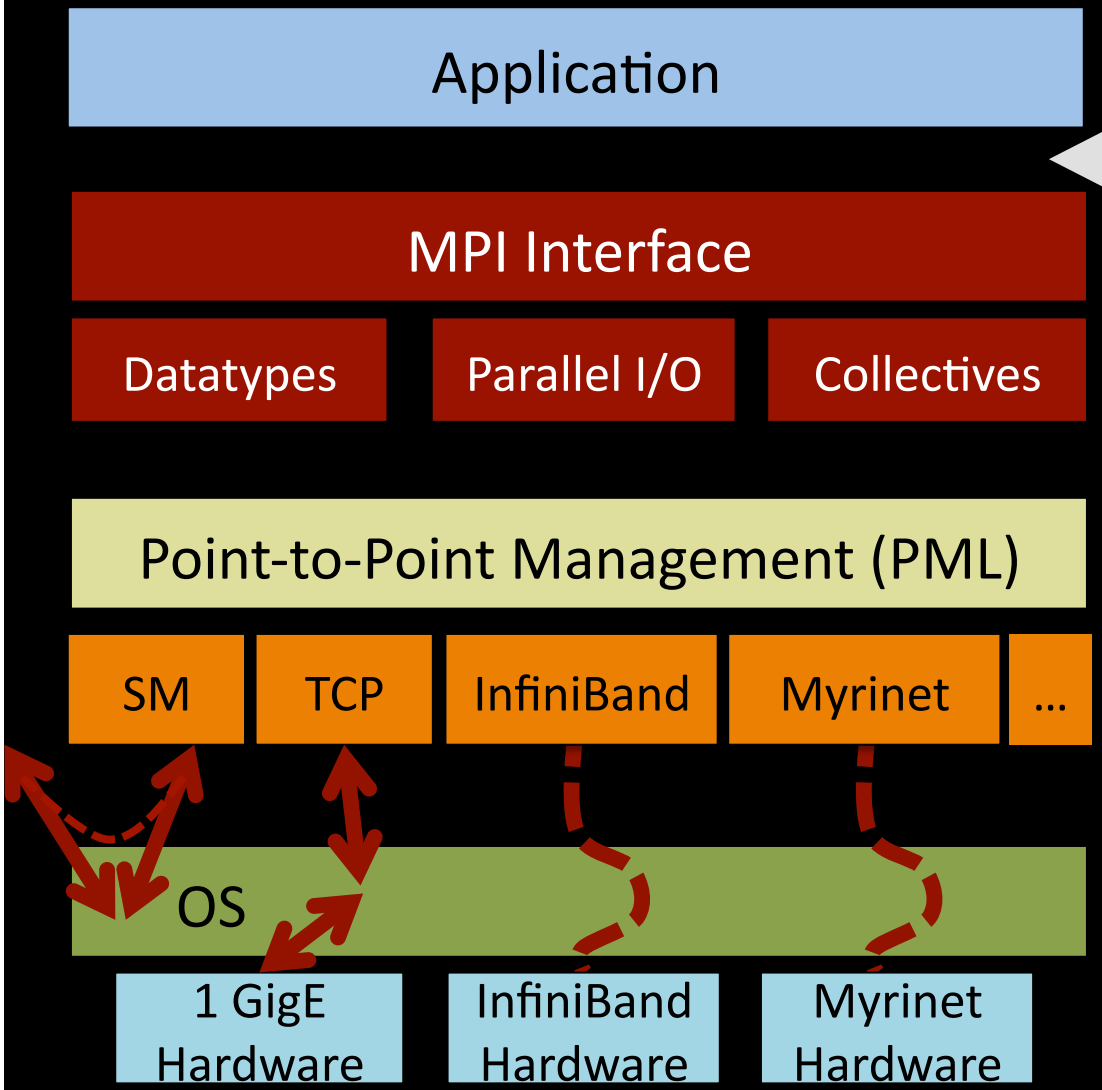
E.g., Coordinated, Uncoordinated, Msg. Induced

MPI Stack



	Latency	Bandwidth
Native IB	**5.3 µs**	**4713 Mbps**
IP over IB	25.7 µs	2905 Mbps
Native MX	**3.9 µs**	**9319 Mbps**
IP over MX	22.3 µs	4576 Mbps
Ethernet	**49.9 µs**	**893 Mbps**

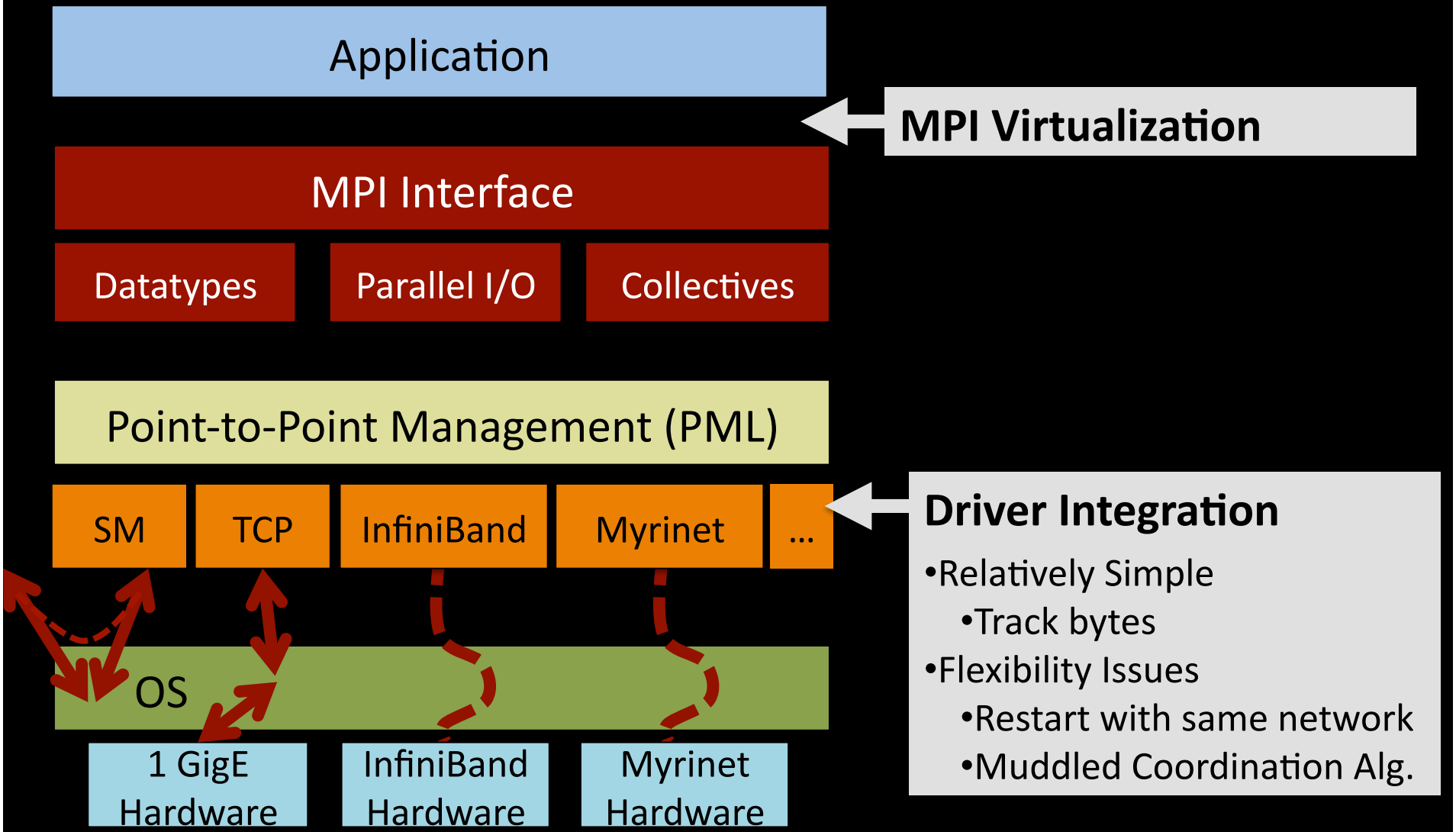
Coordination Protocol Integration



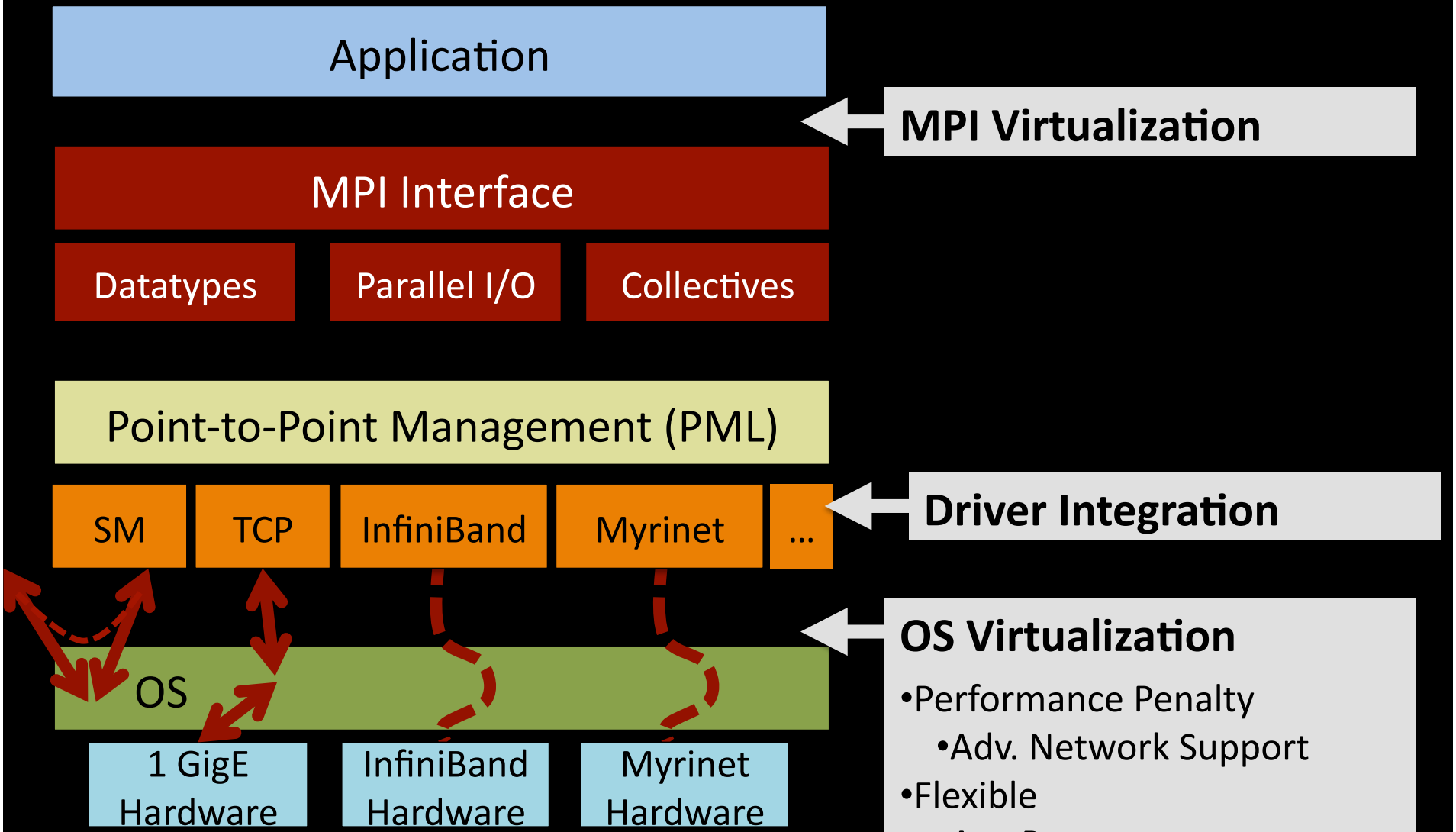
MPI Virtualization

- Complex
  - ~300 functions
- Flexible
  - Any Network & MPI

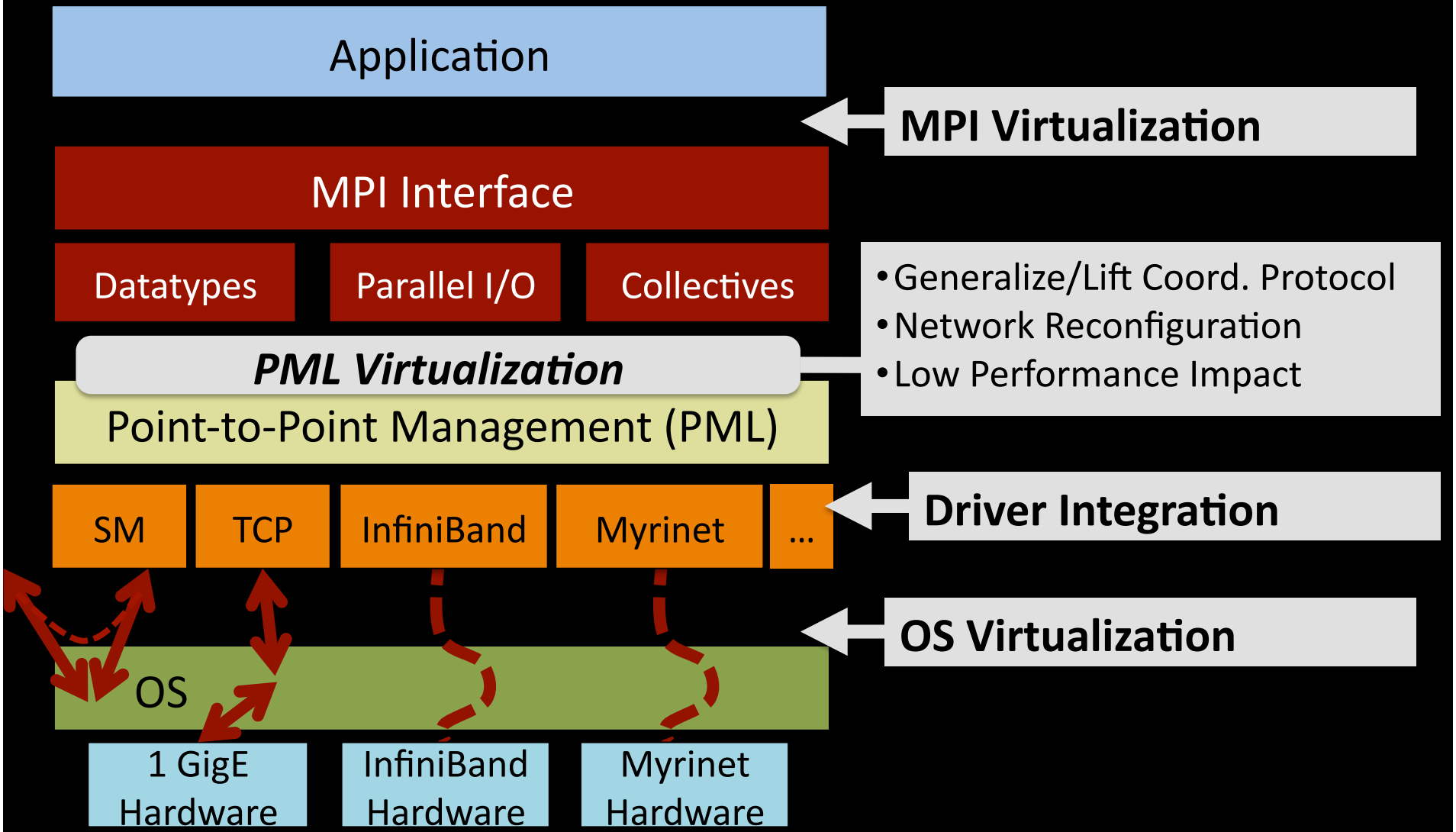
Coordination Protocol Integration



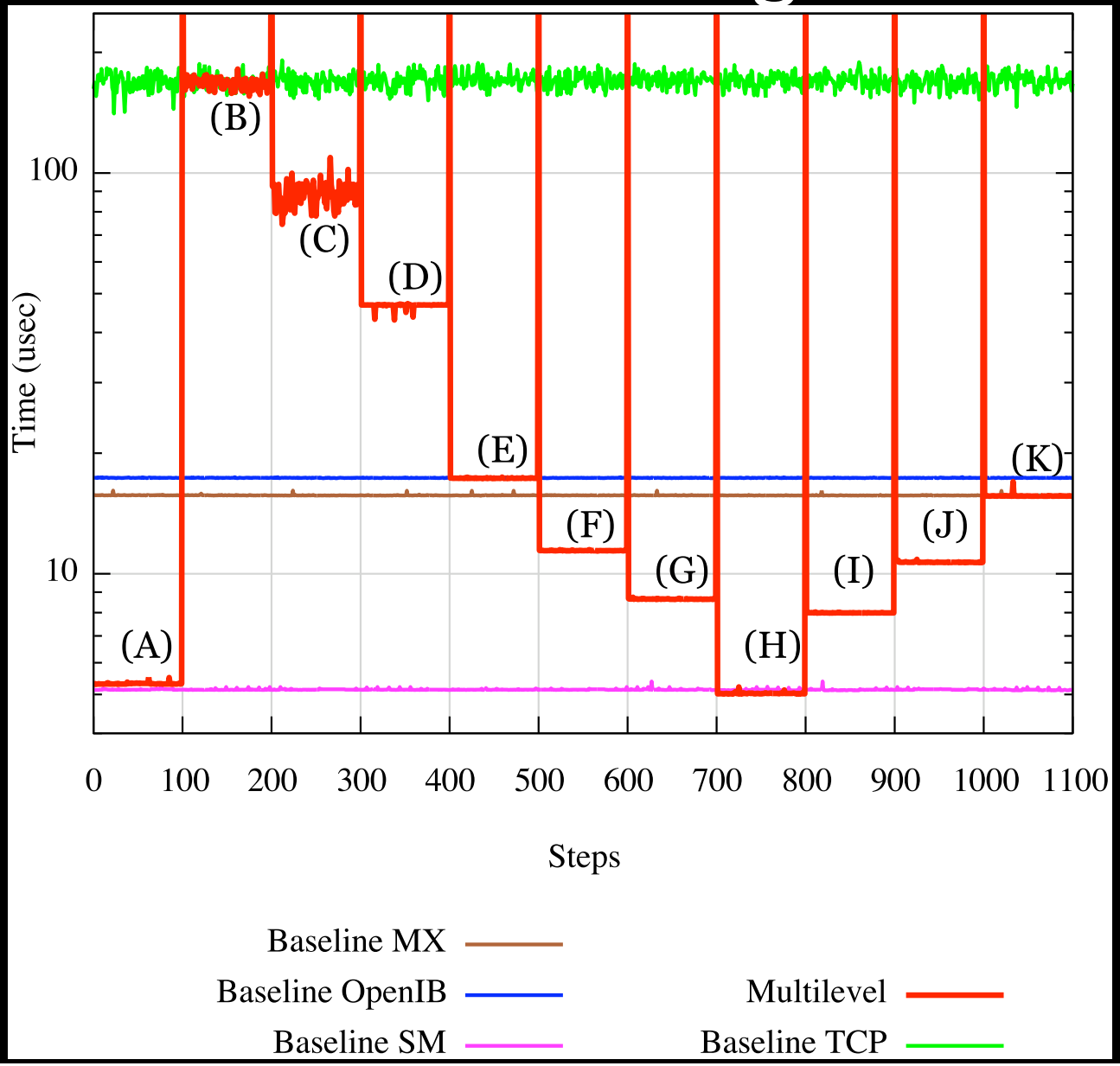
Coordination Protocol Integration



Coordination Protocol Integration



Network Reconfiguration



Low Performance Impact

Latency

Interconnect	No C/R	With C/R	% Overhead
Ethernet (TCP)	49.92 µs	50.01 µs	0.2 %
InfiniBand	8.25 µs	8.78 µs	6.4 %
Myrinet MX	4.23 µs	4.81 µs	13.7 %
Shared Memory	1.84 µs	2.15 µs	16.8 %

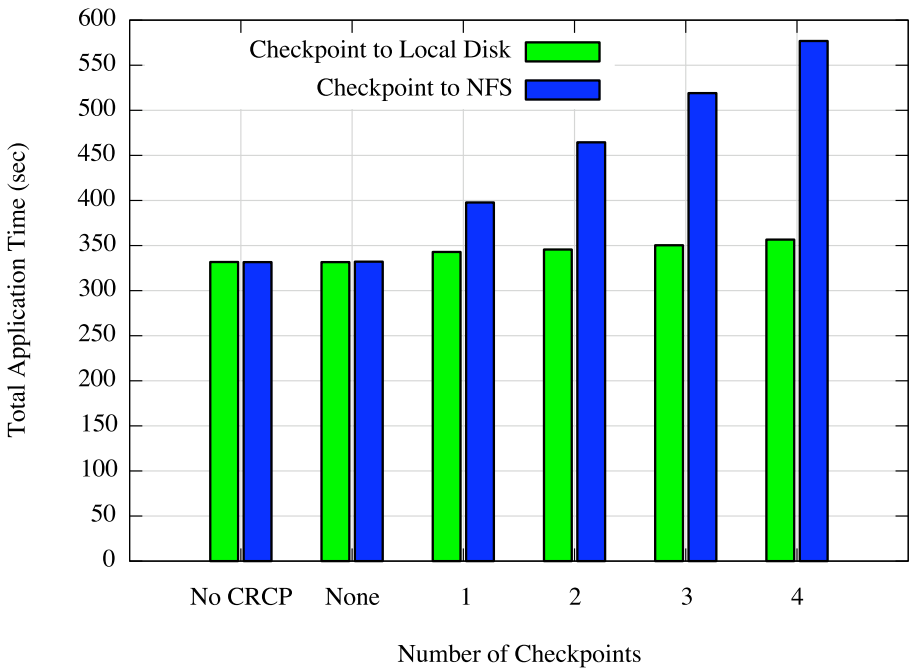
Bandwidth

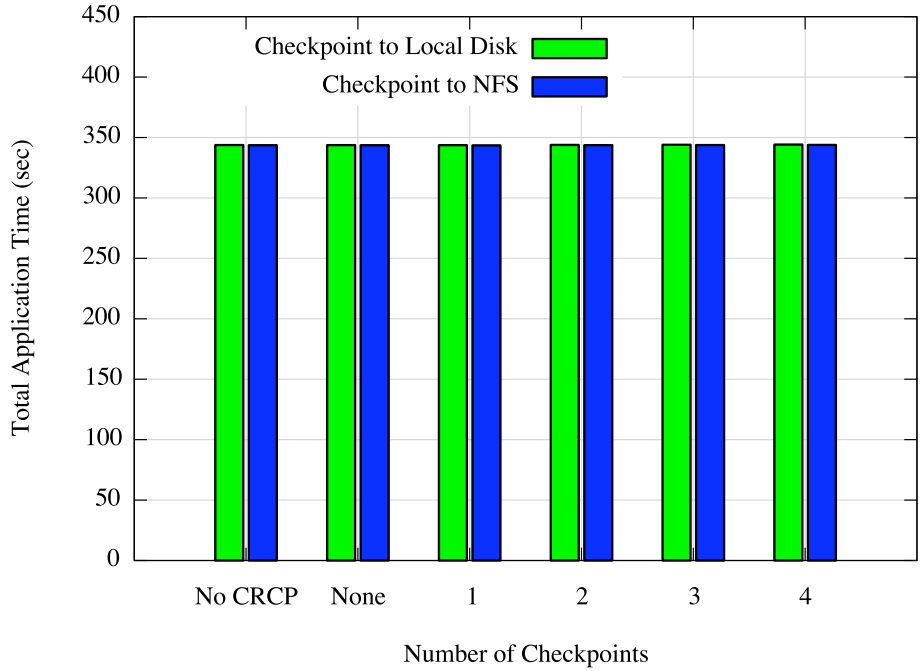
Interconnect	No C/R	With C/R	% Overhead
Ethernet (TCP)	738 Mbps	738 Mbps	0.0 %
InfiniBand	4703 Mbps	4703 Mbps	0.0 %
Myrinet MX	8000 Mbps	7985 Mbps	0.2 %
Shared Memory	5266 Mbps	5258 Mbps	0.2 %

NASA Parallel Benchmarks: 0 – 0.6 %

Gromacs (DPPC): 0%

Checkpoint Overhead

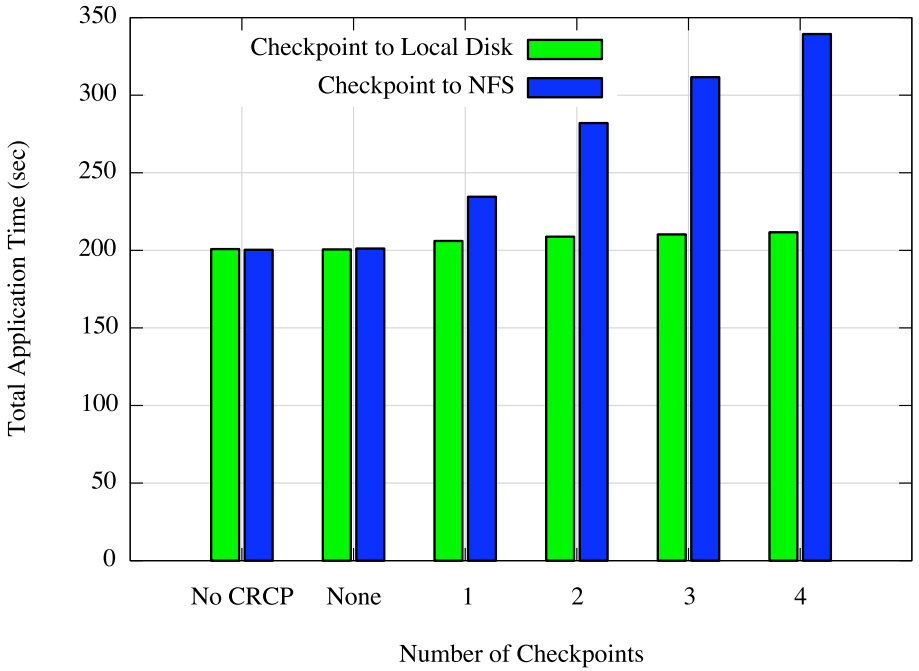


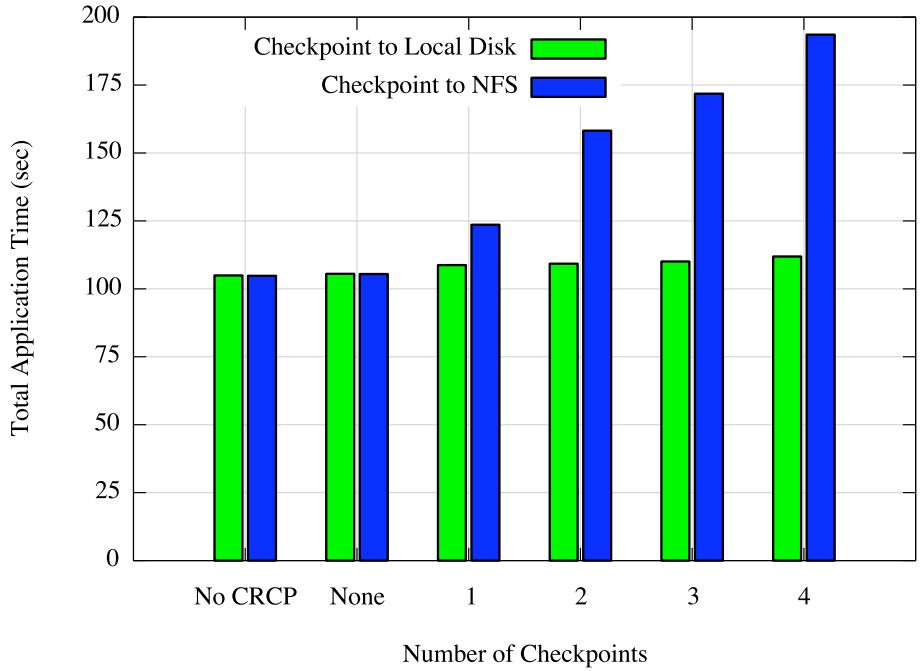


BT Class C 36 Procs
4.2 GB/120 MB

EP Class D 32 Procs
102 MB/3.2 MB

Checkpoint Overhead

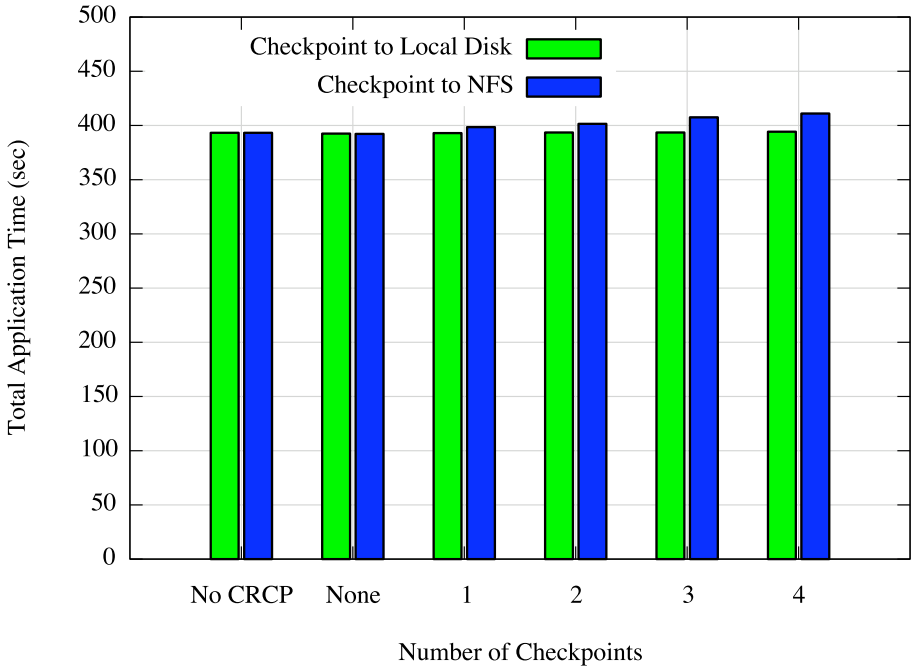


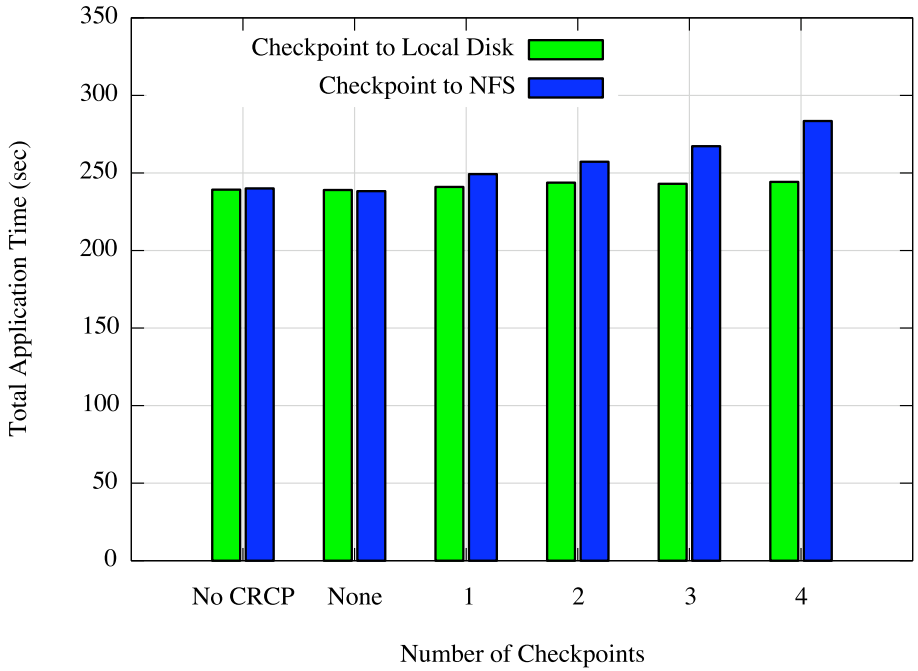


SP Class C 36 Procs
1.9 GB/54 MB

LU Class C 32 Procs
1 GB/32 MB

Checkpoint Overhead



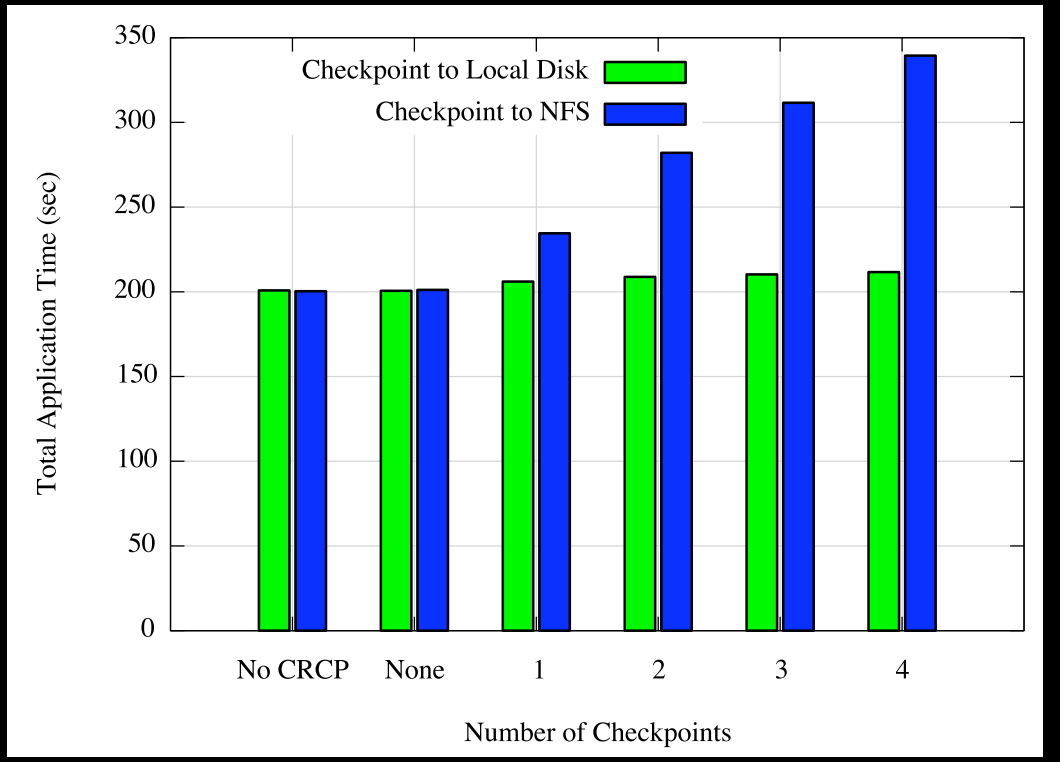


Gromacs (DPPC) 8 Procs
267 MB/33 MB

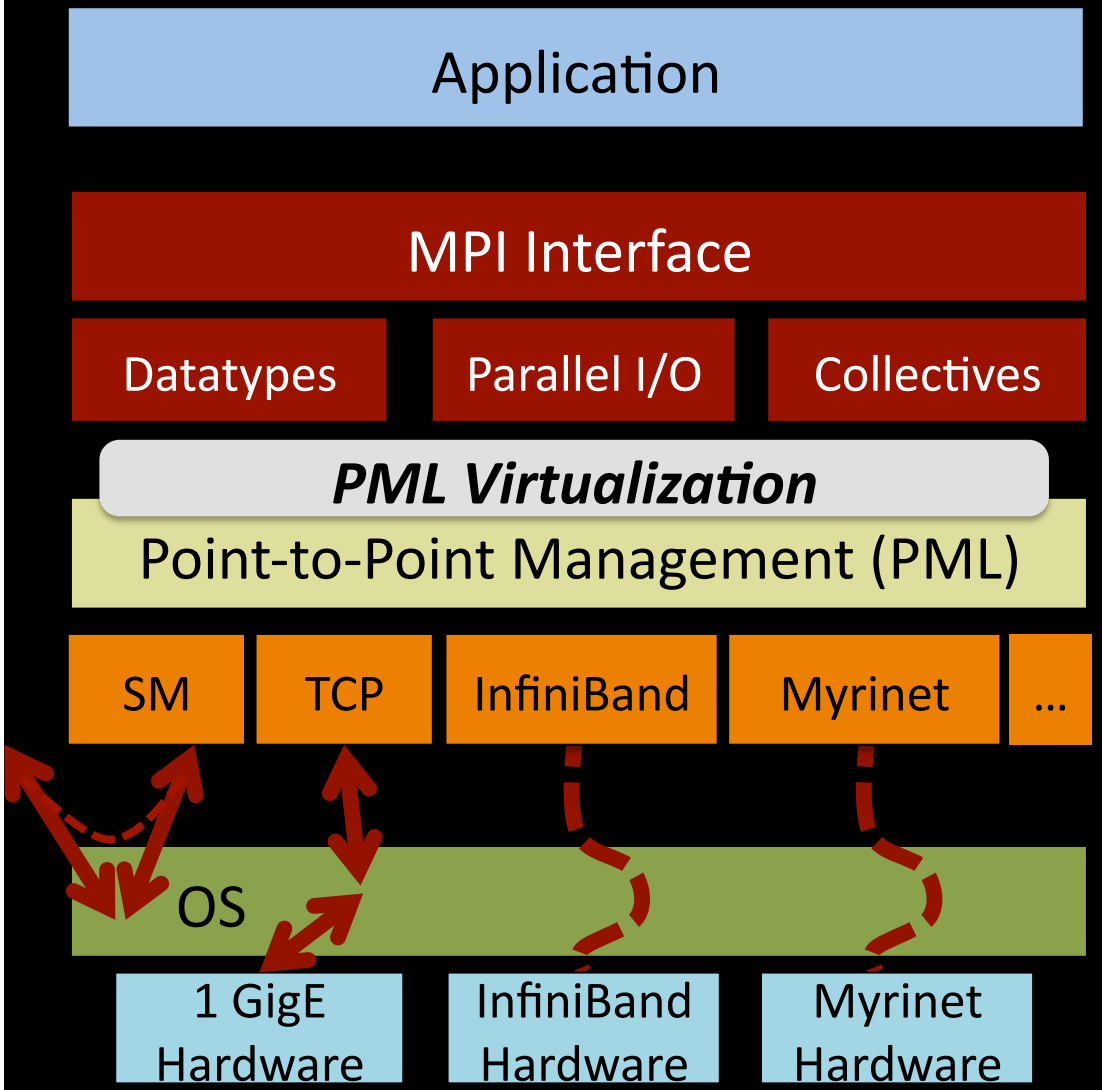
Gromacs (DPPC) 16 Procs
473 MB/30 MB

Checkpoint Bottlenecks

98.8% File I/O
0.7% Modex
0.3% Coord. Protocol
0.2% Internal Coord.



Conclusions



- Generality
  - Lift coordination protocol away from interconnect details
- Flexibility
  - Always choose best available network(s)
- Adaptability
  - Switch Failure
  - Load Balance
  - Cluster Outages
- Low Performance Impact

Future Directions

Bottlenecks

98.8% File I/O

0.7% Modex

0.3% Coord. Protocol

0.2% Internal Coord.

- Stable Storage
  - Multi-stage, distributed store
  - Checkpoint aggregation
- Modex
  - Better interconnect drivers
- Coordination Protocols
  - Scalable, semi-coordinated
- Application Interaction
- Live Process Migration

Questions & Comments

OPEN MPI

INDIANA UNIVERSITY

PERVASIVE TECHNOLOGY INSTITUTE