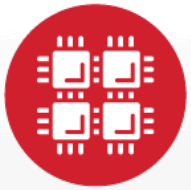


Ohio Supercomputer Center

An OH·TECH Consortium Member

April 2014 HPC Tech Talk

Agenda

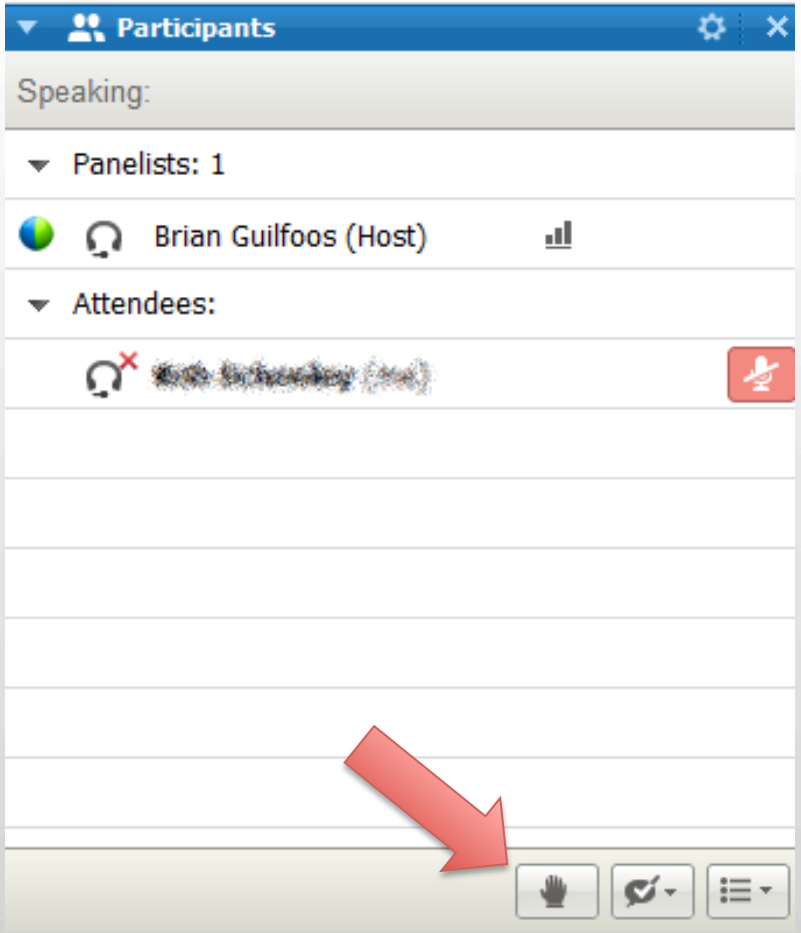
- Mission
- WebEX tips
- Overview of service updates
- Attendee-driven discussion
- "Tech Notes" (30 minutes) – MPI3 Invited Talk
- Slides are available at
  - http://www.osc.edu/tech_talks

Mission

- We want to better engage the daily users of the system
- Provide another avenue for the community to raise issues and talk about unmet needs.
- This event is for you! Please ask questions, make comments, and provide feedback as to how these can be improved to better serve you.
- https://www.surveymonkey.com/s/TV6H5VN

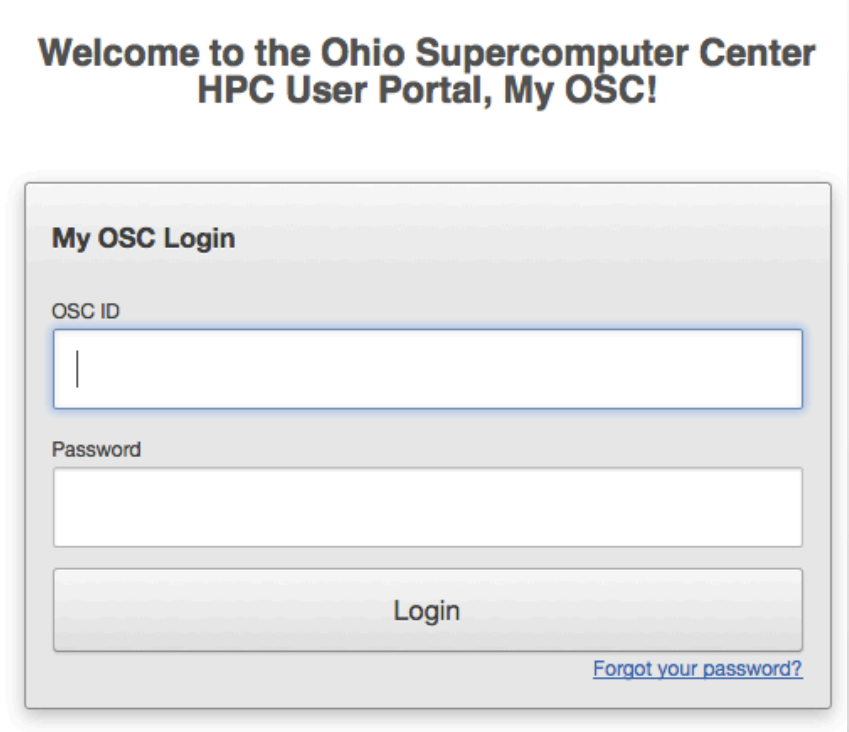
WebEX tips

- You can use the “Raise hand” icon to ask a question or make a comment (this will notify us so that we can acknowledge you)
- You may also use the Q&A section to ask questions.
- Please mute your microphones when not talking, to avoid feedback and interference noises.



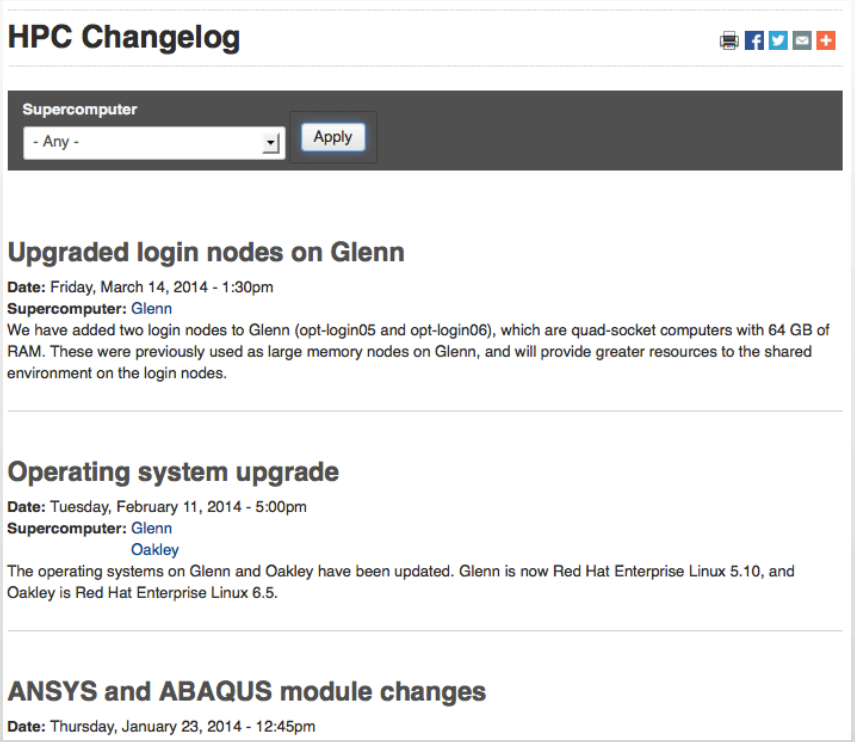
Introducing MyOSC!

- We have launched a new service called MyOSC
  - https://my.osc.edu
- Currently allows changing password, email address, and shell.
- Can regain access to your account if you forgot your password.
- More features in development!
- These functions have been removed from ARMSTRONG



“Changelog” now available on osc.edu

- Significant configuration changes will be reported here
  - http://www.osc.edu/supercomputing/changelog
- Can view individual years (“…changelog/2014”) or filter by HPC.
- Will be embedding the changelog in the sidebar on some pages.



SUG Software Focus Groups

- We'd like your input on decisions regarding software installed for general use (via modules) on our systems
- To indicate interest, please visit http://goo.gl/Dfsfqp
  - Use the access code "OSCSUG"
- Establishing several initial focus groups:
  - Bio informatics/Bio Sciences
  - Fluid Dynamics
  - Structural mechanics
  - Quantum Chemistry/ Materials
  - Physics
  - Atmospheric and Environmental Modeling

AweSim Apps

- We are looking for ideas for the AweSim app store, and for developers for beta testing.
- Please contact us if you are interested.
- http://www.awesim.org/

Glenn is still in production!

- Long waits on Oakley, but often no wait at all on Glenn
- We can provide assistance in migrating your jobs to Glenn, or helping you decide if you will benefit from switching.
- If there is anything preventing you from using Glenn (for example, missing software), please let OSC Help know.

Ruby cluster available for general use this summer

- Ruby is a small 8 node cluster that has been used for research purposes
- We will be buying ~240 nodes for the cluster this summer (4800 cores)
- Ruby will much larger than Glenn and nearly as large as Oakley (in peak performance)
- Large number of newer NVIDIA accelerators and Intel Xeon Phi accelerators
- We will be retiring a portion of Glenn to free up the power necessary to expand Ruby.

Charging policy change under consideration

- Proposal to charge serial jobs on Oakley a number of cores proportional to requested memory
  - 4GB of RAM per core on Oakley compute nodes
  - nodes=1:ppn=1:mem=12GB would be charged for 3 cores
- Memory use will be limited to the amount requested, or the implied amount (4GB * ppn)
- No impact for whole-node jobs (including parallel jobs)
- More details: http://www.osc.edu/memcharging
- Public comment period is now open

Known Issues

- You cannot check your quota on GPFS. No workaround; the quota numbers reported when you log in are calculated once per day.
- If you block popups on Safari, some OnDemand features will silently fail. We are working on both a workaround and a long-term fix.
- Look at the bottom of http://www.osc.edu/supercomputing for up to date issue reporting.

Upcoming Training

- VSCSE Advanced MPI Tutorial, May 6-7
  - http://www.vscse.org
- XSEDE Monthly Workshop, May 7-8
  - MPI
  - https://www.osc.edu/calendar/events/xsede_hpc_monthly_workshop_mpi
- XSEDE Bootcamp, June 24-26
  - Official announcement coming soon.

Discussion period

- Floor open to questions of presented material, or for the community to raise other issues to discuss.

MPI-3 Support in MVAPICH2

An HPC Tech Talk at OSC, April'14

by

Hari Subramoni

The Ohio State University

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

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

Khaled Hamidouche

The Ohio State University

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

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

Presentation Outline

- **Overview of MVAPICH2 and MVAPICH2-X**
- Optimizing and Tuning MPI-3 Non-blocking Collectives
- Efficient use of MPI-3 Remote Memory Access
- Support for MPI-3 Tools : MPI-T Interface : Usage and Benefits

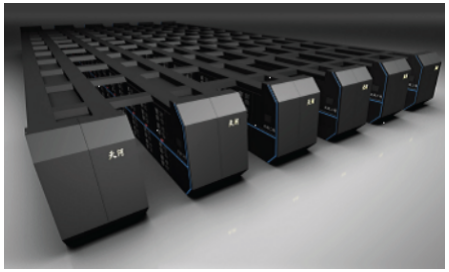
Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, >100Gbps Bandwidth

Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip

- Multi-core processors are ubiquitous
- InfiniBand is very popular in HPC clusters
- Accelerators/Coprocessors are becoming common in high-end systems
- Pushing the envelope for Exascale computing



Tianhe – 2 (1)

Titan (2)

Stampede (6)

Tianhe – 1A (10)

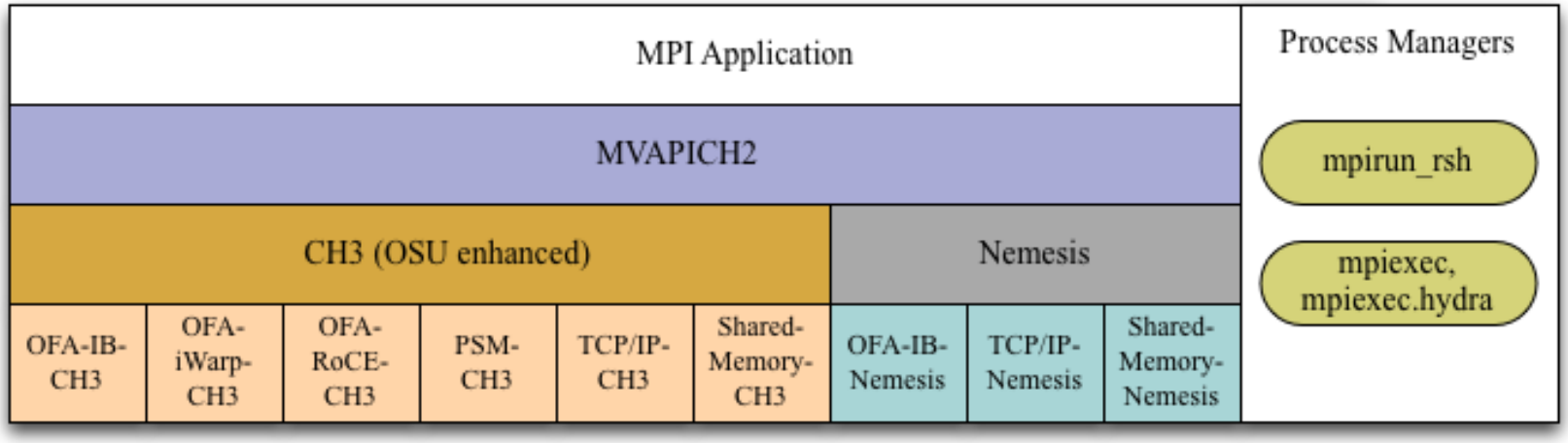
MVAPICH2/MVAPICH2-X Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2012
  - Support for GPGPUs and MIC
  - **Used by more than 2,150 organizations in 72 countries**
  - **More than 210,000 downloads from OSU site directly**
  - Empowering many TOP500 clusters
    - 7th ranked 462,462-core cluster (Stampede) at TACC
    - 11th ranked 74,358-core cluster (Tsubame 2.5) at Tokyo Institute of Technology
    - 16th ranked 96,192-core cluster (Pleiades) at NASA
    - 75th ranked 16,896-core cluster (Keenland) at GaTech and many others . . .
  - Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Partner in the U.S. NSF-TACC Stampede System

Major Features in MVAPICH2/MVAPICH2X for Multi-Petaflop and Exaflop Systems

- Scalability for large number of processes
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Extremely minimum memory footprint
- Scalable Job Startup
- Support for Efficient Process Mapping and Multi-threading
- High-performance Inter-node / Intra-node Point-to-point Communication
- Support for Multiple IB Transport Protocols for Scalable Communication
- Support for Multi-rail Clusters and 3D Torus Networks
- QoS support for Communication and I/O
- Scalable Collective Communication
- Support for GPGPUs and Accelerators
- Hybrid Programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, …)
- Enhanced Debugging System
- *and many more…*

MVAPICH2 Architecture (Latest Release 2.0rc1)



All Different PCI, PCI-Ex interfaces

Major Computing Platforms: IA-32, Ivybridge, Nehalem, Westmere, Sandybridge, Opteron, Magny, ..

MVAPICH2 2.0RC1 and MVAPICH2-X 2.0RC1

- Released on 03/24/14
- Major Features and Enhancements
  - Based on MPICH-3.1
  - Improved performance for MPI_Put and MPI_Get operations in CH3 channel
  - Enabled MPI-3 RMA support in PSM channel
  - Enabled multi-rail support for UD-Hybrid channel
  - Optimized architecture based tuning for blocking and non-blocking collectives
  - Optimized Bcast and Reduce collectives designs
  - Improved hierarchical job startup time
  - Optimization for sub-array data-type processing for GPU-to-GPU communication
  - Updated hwloc to version 1.8
  - Enhanced build system to avoid separate builds for different networks/interfaces
    - Updated compiler wrappers (example: mpicc) to avoid adding dependencies on network and other libraries
- MVAPICH2-X 2.0RC1 supports hybrid MPI + PGAS (UPC and OpenSHMEM) programming models
  - Based on MVAPICH2 2.0RC1 including MPI-3 features; Compliant with UPC 2.18.0 and OpenSHMEM v1.0f
  - Improved intra-node performance using Shared memory and Cross Memory Attach (CMA)
  - Optimized UPC collectives

MVAPICH2-2.0b GPU Direct RDMA (GDR) Release

- MVAPICH2-2.0b with GDR support can be downloaded from
  https://mvapich.cse.ohio-state.edu/download/mvapich2gdr/
- System software requirements
  - Mellanox OFED 2.1
  - NVIDIA Driver 331.20 or later
  - NVIDIA CUDA Toolkit 5.5
  - Plugin for GPUDirect RDMA
    (http://www.mellanox.com/page/products_dyn?product_family=116)
- Has optimized designs for point-to-point communication using GDR
- Work under progress for optimizing collective and one-sided communication
- Contact MVAPICH help list with any questions related to the package
  mvapich-help@cse.ohio-state.edu
- MVAPICH2-GDR-RC1 with additional optimizations coming soon!!

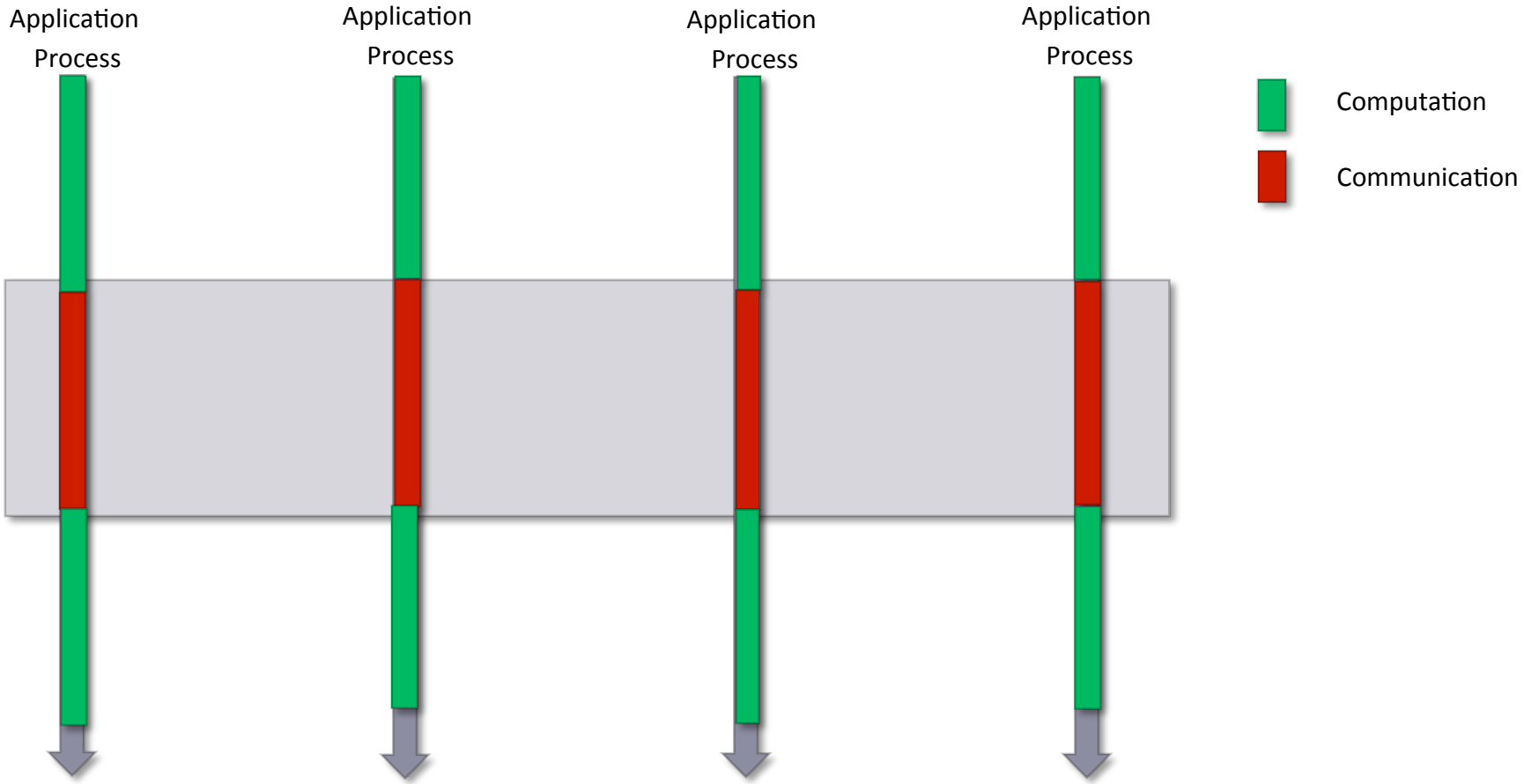
Major New Features in MPI-3

- Major features
  - Non-blocking Collectives
  - Improved One-Sided (RMA) Model
  - MPI Tools Interface
- Specification is available from
  - http://www.mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf

Presentation Outline

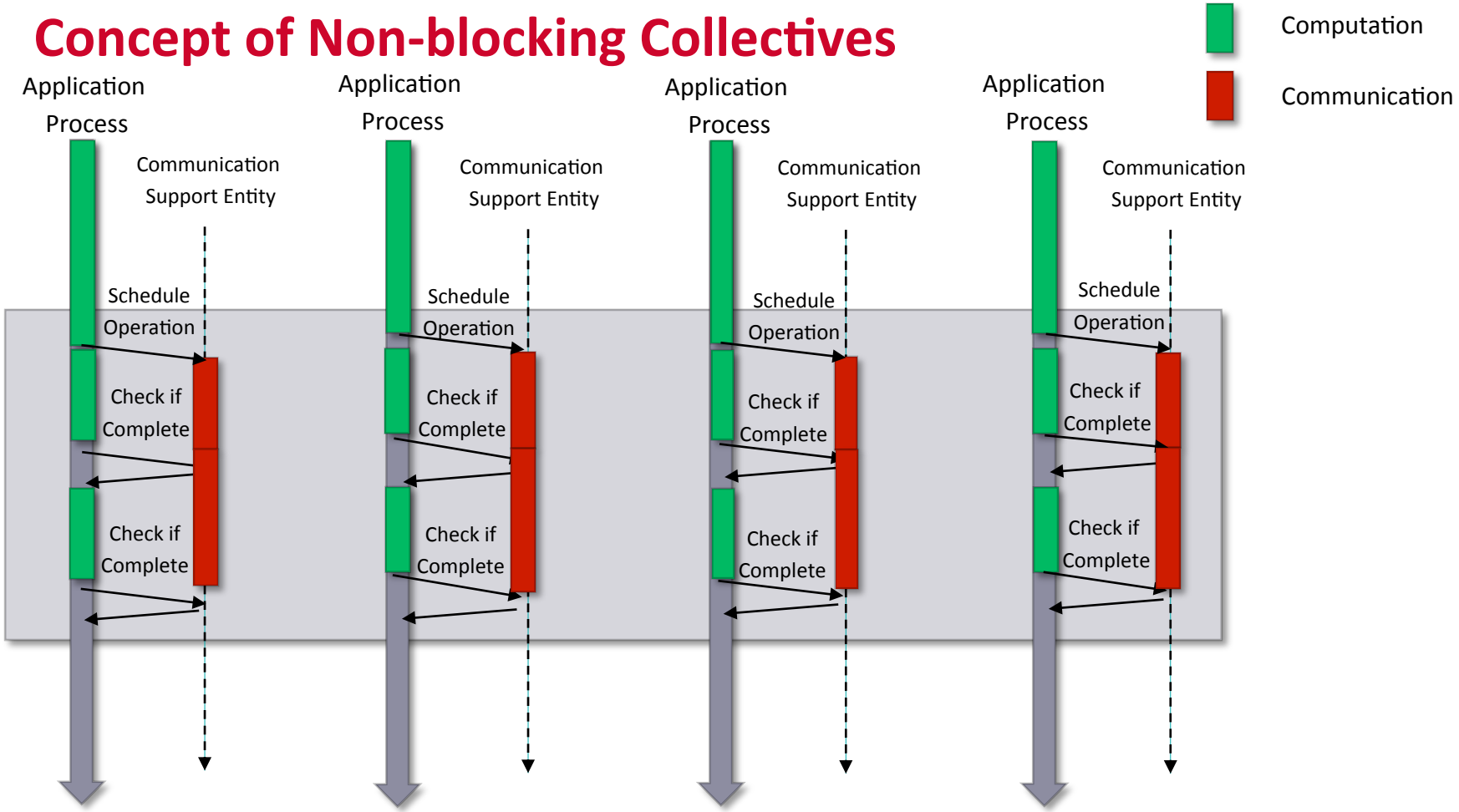
- Overview of MVAPICH2 and MVAPICH2-X
- **Optimizing and Tuning MPI-3 Non-blocking Collectives**
- Efficient use of MPI-3 Remote Memory Access
- Support for MPI-3 Tools : MPI-T Interface : Usage and Benefits

Problems with Blocking Collective Operations



- Communication time cannot be used for compute
  - No overlap of computation and communication
  - Inefficient

Concept of Non-blocking Collectives



- Application processes schedule collective operation
- Check periodically if operation is complete
- **Overlap of computation and communication => Better Performance**
- *Catch: Who will progress communication*

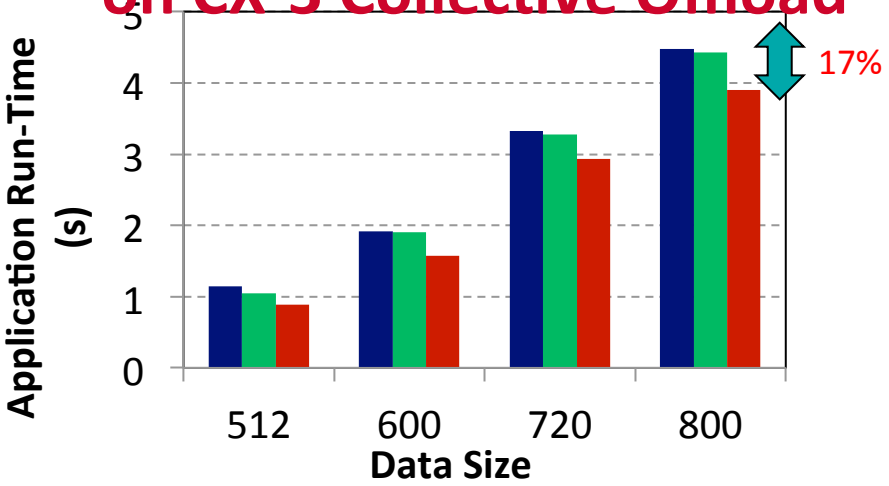
Non-blocking Collective (NBC) Operations

- Enables overlap of computation with communication
- Non-blocking calls do not match blocking collective calls
  - MPI may use different algorithms for blocking and non-blocking collectives
  - Blocking collectives: Optimized for latency
  - Non-blocking collectives: Optimized for overlap
- A process calling a NBC operation
  - Schedules collective operation and immediately returns
  - Executes application computation code
  - Waits for the end of the collective
- The communication progress by
  - Application code through MPI_Test
  - Network adapter (HCA) with hardware support
  - Dedicated processes / thread in MPI library
- There is a non-blocking equivalent for each blocking operation
  - Has an "I" in the name
    - MPI_Bcast -> MPI_Ibcast; MPI_Reduce -> MPI_Ireduce

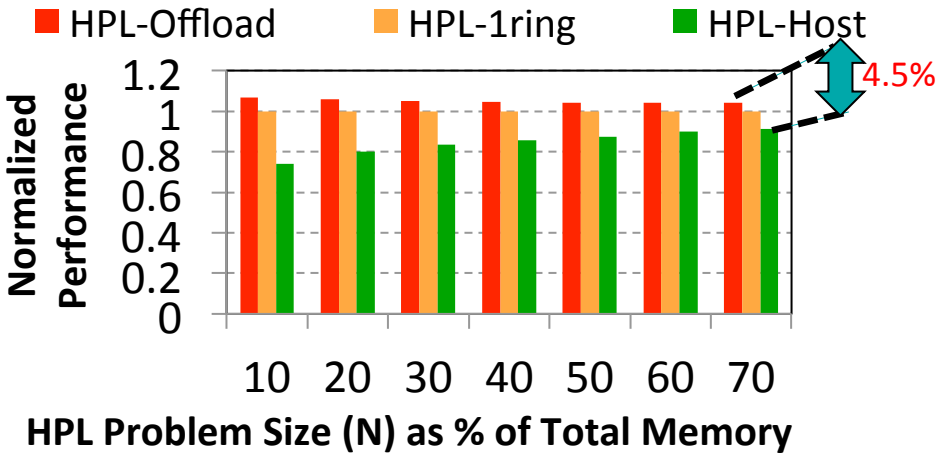
How do I write applications with NBC?

```
void main()
{
    MPI_Init()
    …..
    MPI_Ialltoall(…)
    Computation that does not depend on result of Alltoall
    MPI_Test(for Ialltoall) /* Check if complete (non-blocking) */
    Computation that does not depend on result of Alltoall
    MPI_Wait(for Ialltoall) /* Wait till complete (Blocking) */
    …
    MPI_Finalize()
}
```

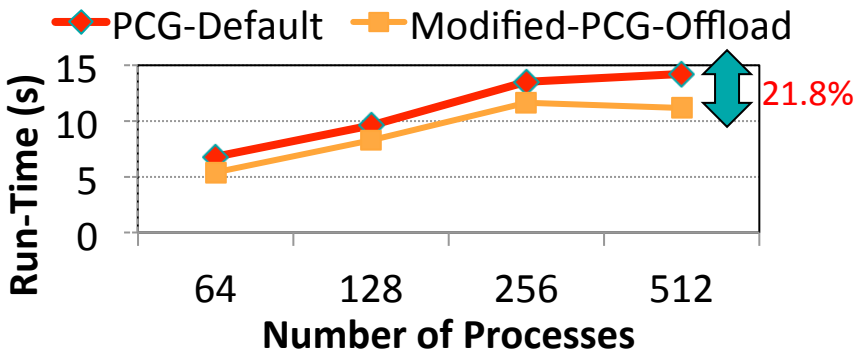
Application benefits with Non-Blocking Collectives based on CX-3 Collective Offload



Modified P3DFFT with Offload-Alltoall does up to 17% better than default version (128 Processes)



Modified HPL with Offload-Bcast does up to 4.5% better than default version (512 Processes)



Modified Pre-Conjugate Gradient Solver with Offload-Allreduce does up to 21.8% better than default version

K. Kandalla, et. al.. High-Performance and Scalable Non-Blocking All-to-All with Collective Offload on InfiniBand Clusters: A Study with Parallel 3D FFT, ISC 2011

K. Kandalla, et. al, Designing Non-blocking Broadcast with Collective Offload on InfiniBand Clusters: A Case Study with HPL, HotI 2011

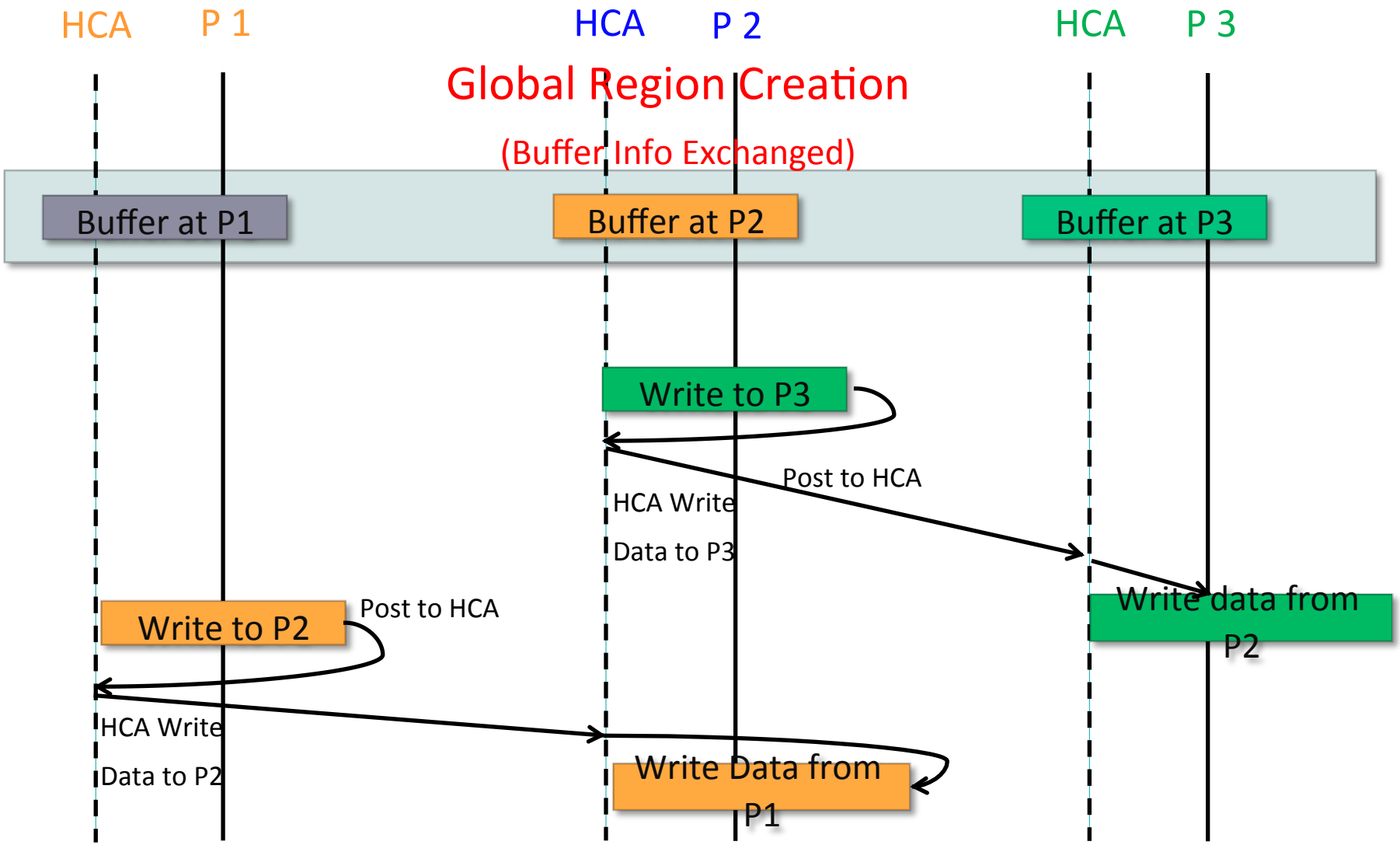
K. Kandalla, et. al., Designing Non-blocking Allreduce with Collective Offload on InfiniBand Clusters: A Case Study with Conjugate Gradient Solvers, IPDPS '12

Can Network-Offload based Non-Blocking Neighborhood MPI Collectives Improve Communication Overheads of Irregular Graph Algorithms? K. Kandalla, A. Buluc, H. Subramoni, K. Tomko, J. Vienne, L. Oliker, and D. K. Panda, IWPAPS' 12

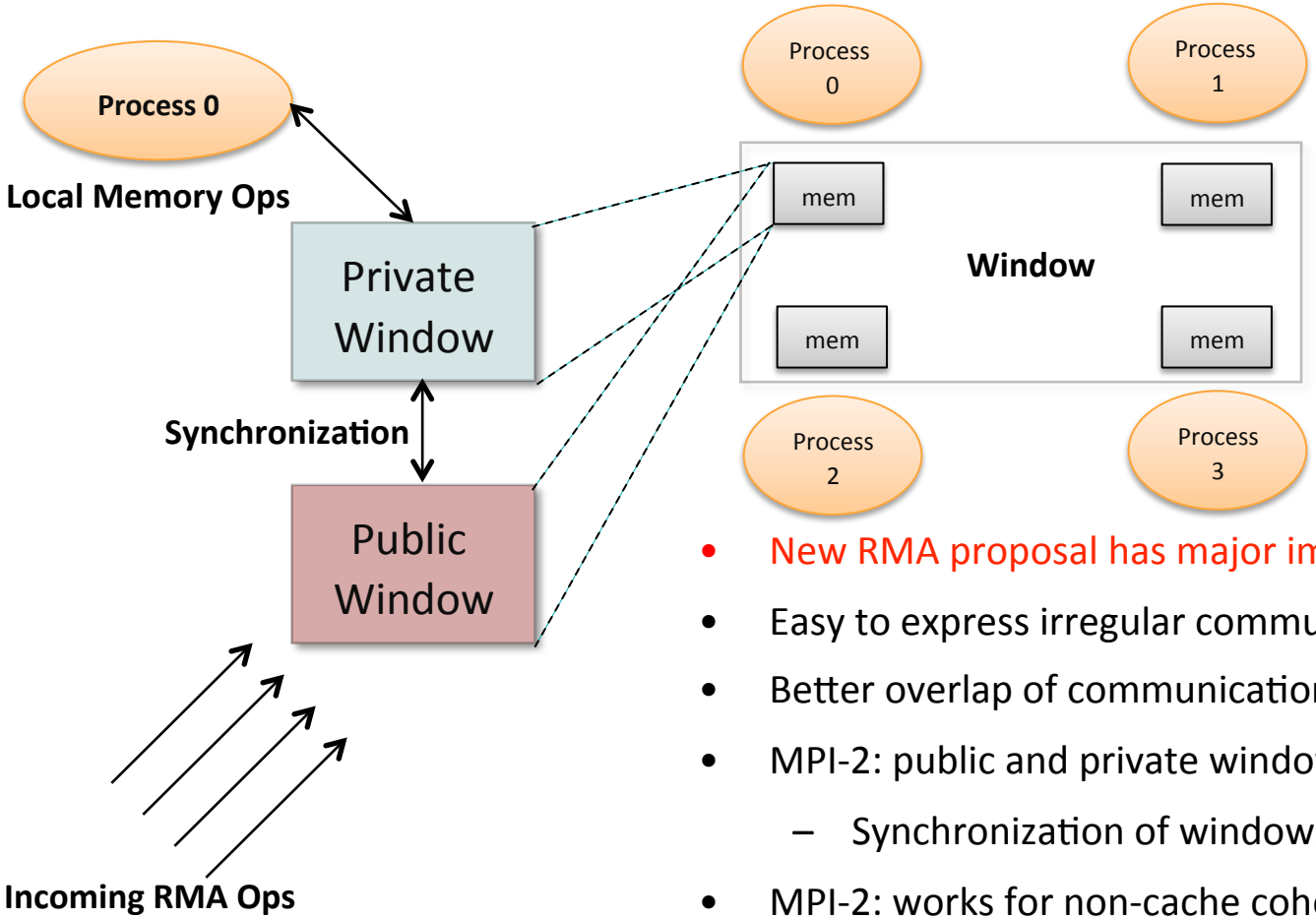
Presentation Outline

- Overview of MVAPICH2 and MVAPICH2-X
- Optimizing and Tuning MPI-3 Non-blocking Collectives
- **Efficient use of MPI-3 Remote Memory Access**
- Support for MPI-3 Tools : MPI-T Interface : Usage and Benefits

One-sided Communication Model



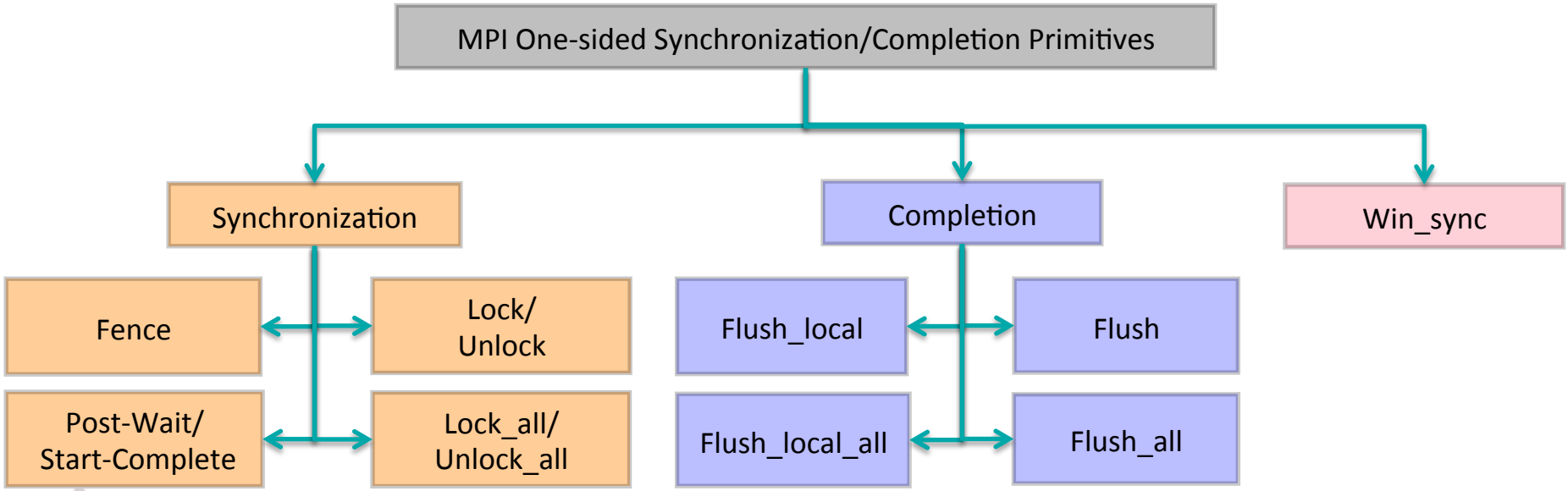
Improved One-sided (RMA) Model



- New RMA proposal has major improvements
- Easy to express irregular communication pattern
- Better overlap of communication & computation
- MPI-2: public and private windows
  - Synchronization of windows explicit
- MPI-2: works for non-cache coherent systems
- MPI-3: two types of windows
  - Unified and Separate
  - Unified window leverages hardware cache coherence

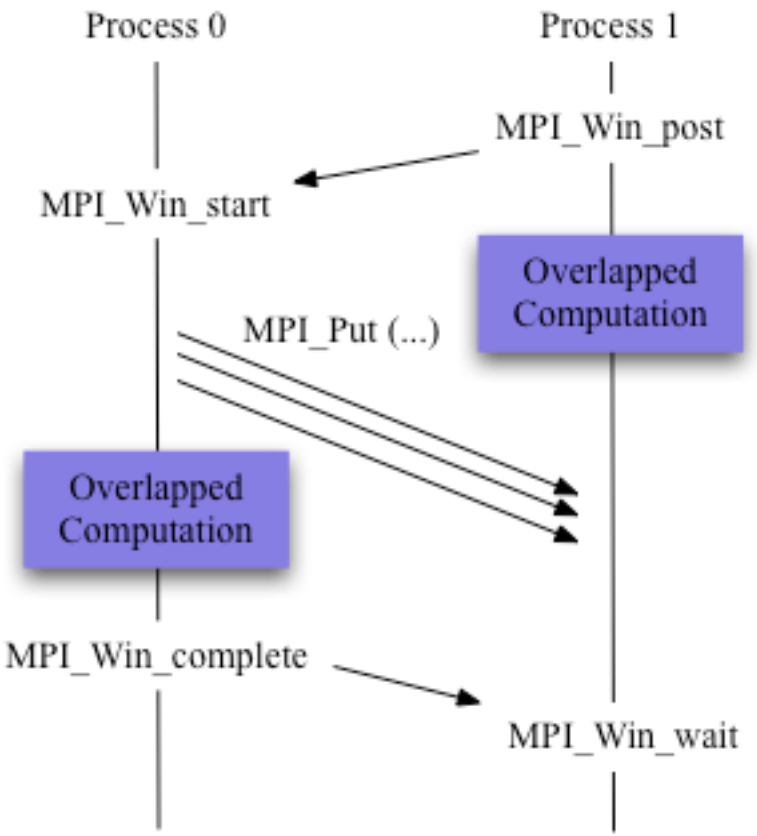
MPI-3 One-Sided Primitives

- Non-blocking one-sided communication routines
  - Put, Get
  - Accumulate, Get_accumulate
  - Atomics
- Flexible synchronization operations to control initiation and completion

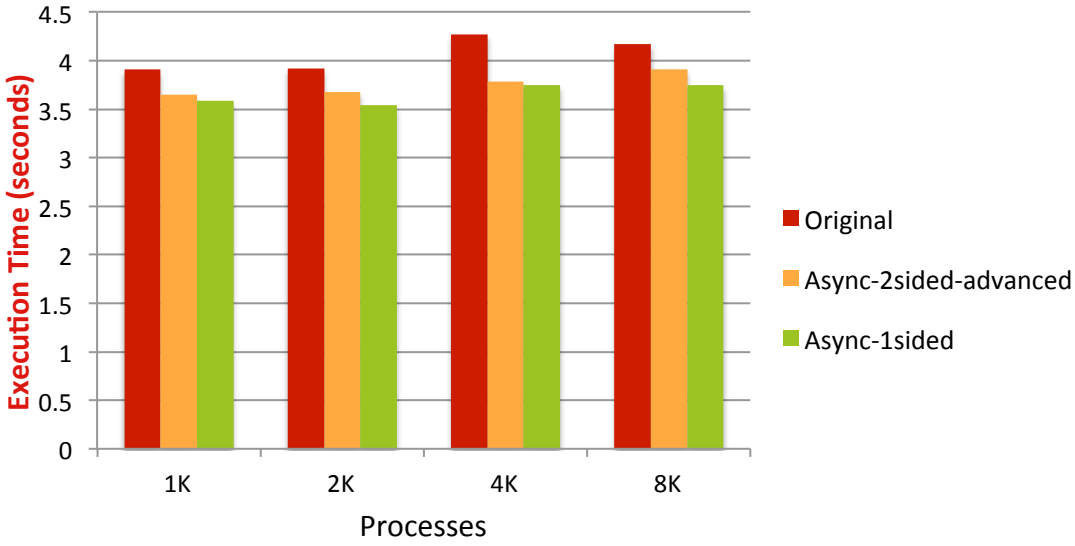


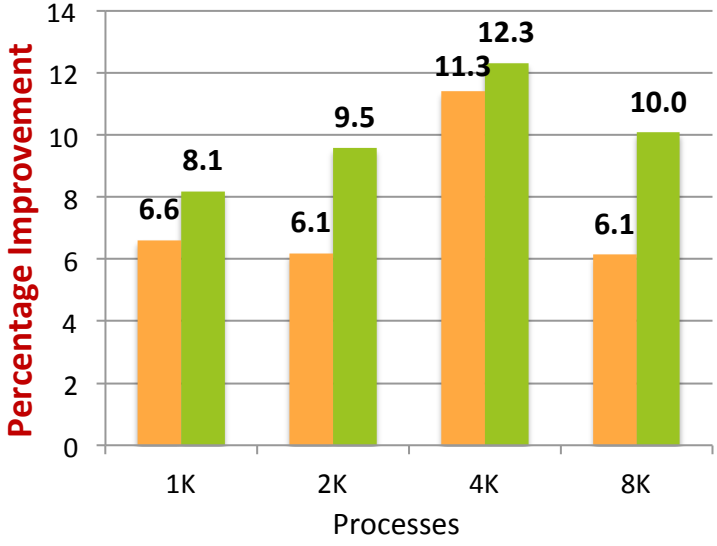
Overlapping Communication with MPI-3-RMA

- Network adapters can provide RDMA feature that doesn't require software involvement at remote side
- As long as puts/gets are executed as soon as they are issued, overlap can be achieved
- RDMA-based implementations do just that



Performance of AWP-ODC using MPI-3-RMA





- Experiments on TACC Ranger cluster 64x64x64 data grid per process – 25 iterations – 32KB messages
- On 4K processes
  - 8% with 2sided basic, 11% with 2sided advanced, 12% with RMA
- On 8K processes
  - 2% with 2sided basic, 6% with 2sided advanced, 10% with RMA

S. Potluri, P. Lai, K. Tomko, S. Sur, Y. Cui, M. Tatineni,K. Schulz, W. Barth, A. Majumdar and D. K. Panda, Quantifying Performance Benefits of Overlap using MPI-2 in a Seismic Modeling Application, ICS '10.

Presentation Outline

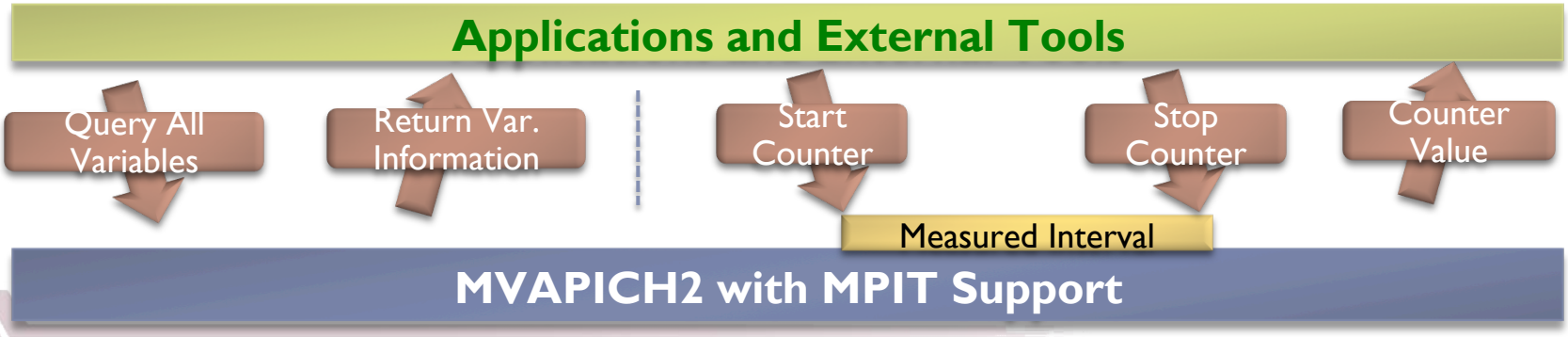
- Overview of MVAPICH2 and MVAPICH2-X
- Optimizing and Tuning MPI-3 Non-blocking Collectives
- Efficient use of MPI-3 Remote Memory Access
- **Support for MPI-3 Tools : MPI-T Interface : Usage and Benefits**

MPI Tools Interface

- Introduced in MPI 3.0 standard to expose internals of MPI to tools and applications
- Generalized interface – no defined variables in the standard
- Variables can differ between
  - MPI implementations
  - Compilations of same MPI library (production vs debug)
  - Executions of the same application/MPI library
  - There could be no variables provided
- Two types of variables supported
  - Control Variables (CVARS)
    - Typically used to configure and tune MPI internals
    - Environment variables, configuration parameters and toggles
  - Performance Variables (PVARS)
    - Insights into performance of an MPI library
    - Highly-implementation specific
    - Memory consumption, timing information, resource-usage, data transmission info.
    - Per-call basis or an entire MPI job
- More about the interface: *mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf*

Who should use MPI-T and How?

- Who???
  - Interface intended for tool developers
    - Generally will do *anything* to get the data
    - Are willing to support the many possible variations
- How???
  - Can be called from user code
  - Useful for setting control variables for performance
  - Documenting settings for understanding performance
  - Care must be taken to avoid code that is not portable
  - Several workflows based on role: End Users / Performance Tuners / MPI Implementers
    - Two main workflows
      - Transparently using MPIT-Aware external tools
      - Co-designing applications and MPI-libraries using MPI-T

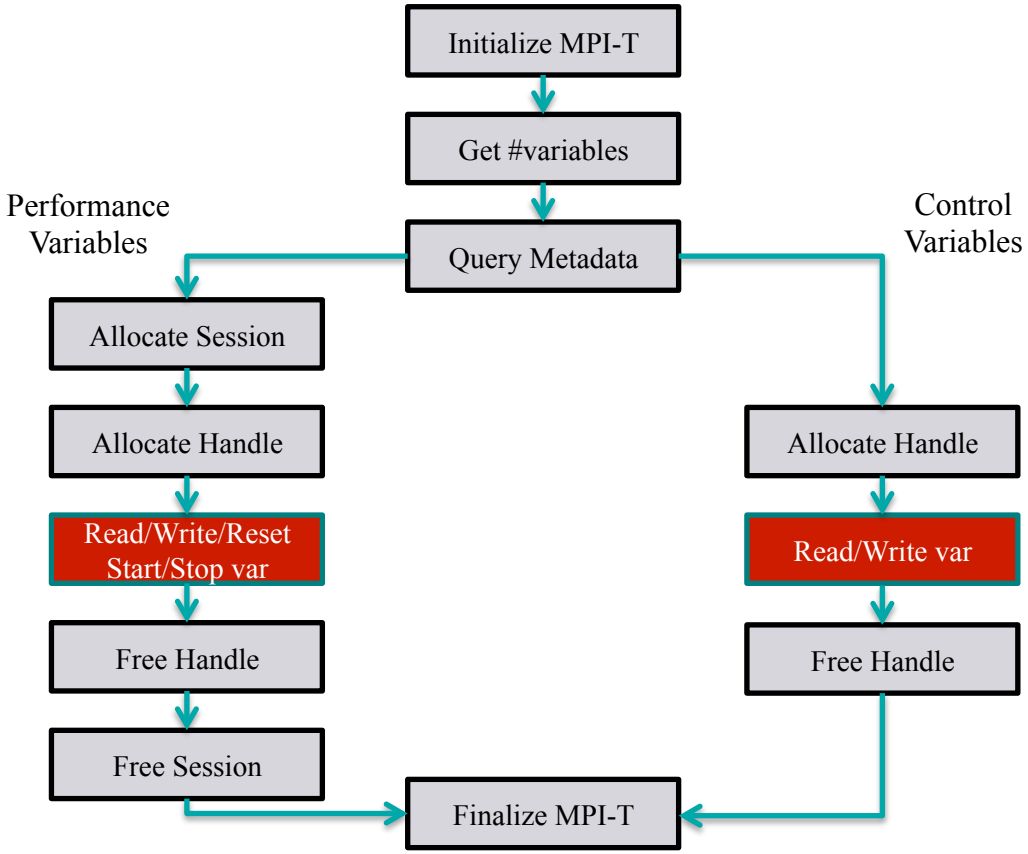


MPI-T Support in MVAPICH2

Memory Usage: - current level - maximum watermark	InfiniBand N/W: - #control packets - #out-of-order packets	Pt-to-pt messages: - unexpected queue length - unexp. match attempts - recvq. length
Registration cache: - hits - misses	Shared-memory: - limic/ CMA - buffer pool size & usage	Collective ops: - comm. creation - #algorithm invocations [Bcast – 8; Gather – 10] …

- Initial focus on performance variables
- Variables to track different components
  - MPI library's internal memory usage
  - Unexpected receive queue
  - Registration cache
  - VBUF allocation
  - Shared-memory communication
  - Collective operation algorithms
  - IB channel packet transmission
  - Many more in progress..

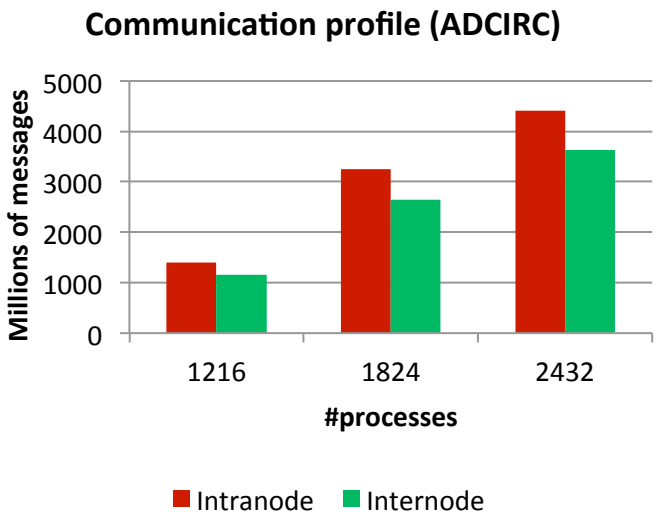
Co-designing Applications to use MPI-T

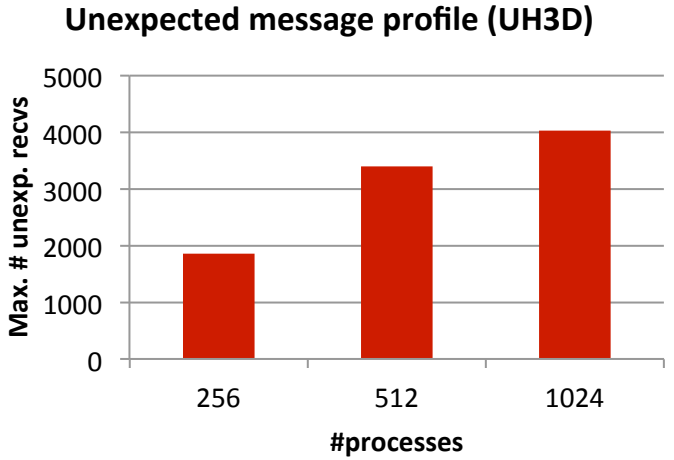


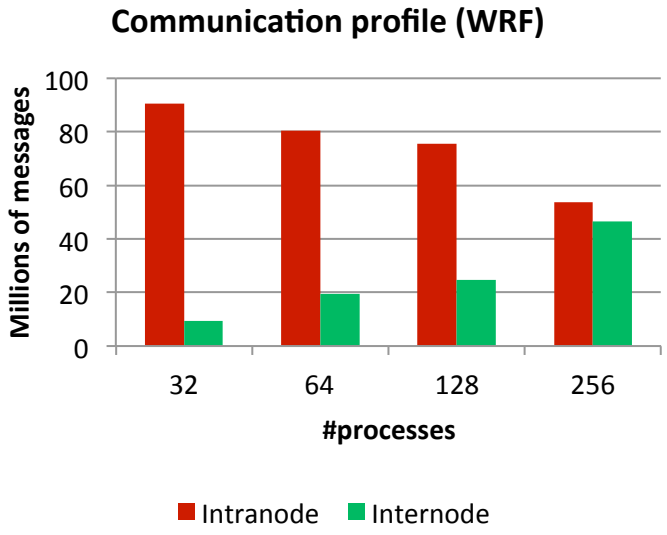
Example: Optimizing the eager limit dynamically ->

```
MPI_T_init_thread()
MPI_T_cvar_get_info(MV2_EAGER_THRESHOLD)
if (msg_size < MV2_EAGER_THRESHOLD + 1KB)
        MPI_T_cvar_write(MV2_EAGER_THRESHOLD, +1024)
MPI_Send(..)
MPI_T_finalize()
```

Evaluating MPI-T with Applications







- Users can gain insights into application communication characteristics!

User Resources

- MVAPIVH2 Quick Start Guide
- MVAPICH2 User Guide
  - Long and very detailed
  - FAQ
- MVAPICH2 Web-Site
  - Overview and Features
  - Reference performance
  - Publications
- Mailing List Support
  - mvapich-discuss@cse.ohio-state.edu
- Mailing List Archives
- All above resources accessible from: http://mvapich.cse.ohio-state.edu/

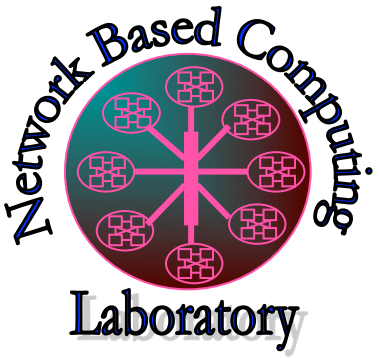
MVAPICH2/MVPICH2-X – Plans for Exascale

- Performance and Memory scalability toward 500K-1M cores
  - Dynamically Connected Transport (DCT) service with Connect-IB
- Enhanced Optimization for GPGPU and Coprocessor Support
  - Extending the GPGPU support (GPU-Direct RDMA) with CUDA 6.0 and Beyond
  - Support for Intel MIC (Knight Landing)
- Taking advantage of Collective Offload framework
  - Including support for non-blocking collectives (MPI 3.0)
- RMA support (as in MPI 3.0)
- Extended topology-aware collectives
- Power-aware collectives
- Extended support for MPI Tools Interface (as in MPI 3.0)
- Checkpoint-Restart and migration support with in-memory checkpointing
- Hybrid MPI+PGAS programming support with GPGPUs and Accelertors

Concluding Remarks

- Provided an overview of advanced MPI-3 support in MVAPICH2
- Presented in-depth details on configuration and runtime parameters, optimizations and their impacts
- MVAPICH2 has many more features not covered here
  - Fault tolerance, Dynamic Process Management etc
  - Please visit http://mvapich.cse.ohio-state.edu for details
- More information about optimizing / tuning MVAPICH2 / MVAPICH2-X available at MVAPICH User Group Meeting (MUG) 2013 website
  - http://mug.mvapich.cse.ohio-state.edu

Pointers



http://nowlab.cse.ohio-state.edu

http://mvapich.cse.ohio-state.edu

subramon@cse.ohio-state.edu
hamidouc@cse.ohio-state.edu

Exit survey

- Please complete a quick survey to help us improve these Tech Talks
- https://www.surveymonkey.com/s/TV6H5VN