MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiBD
High-Performance Big Data

HiDL
High-Performance Deep Learning

Designing Secured MPI for HPC: Opportunities and Challenges

Talk at the 4th High-Performance Computing Security Workshop (May '24)

by

Follow us on

https://twitter.com/mvapich

Dhabaleswar K. (DK) Panda

The Ohio State University

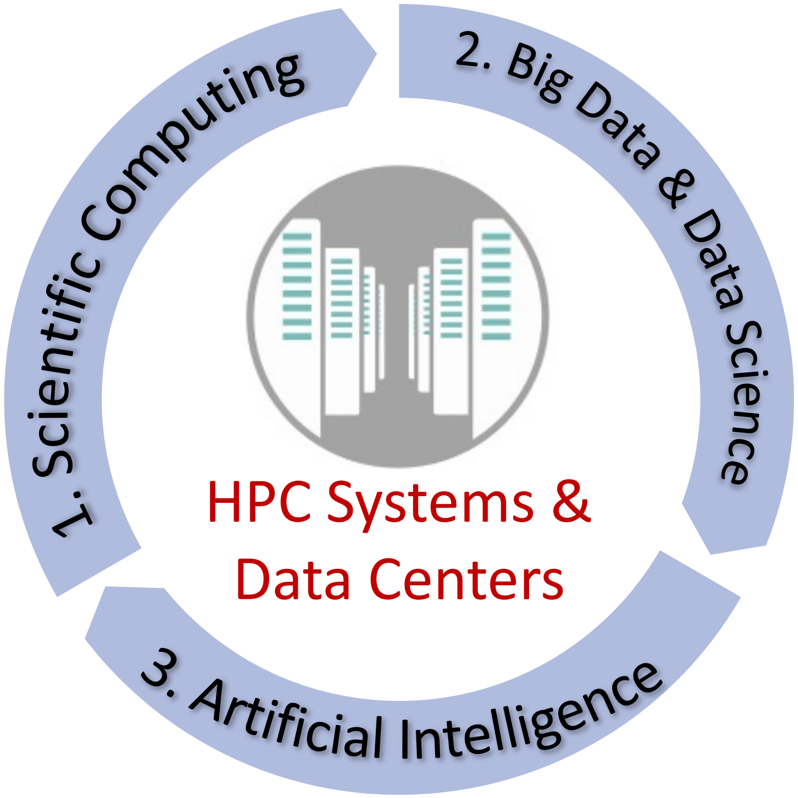
E-mail: panda@cse.ohio-state.edu

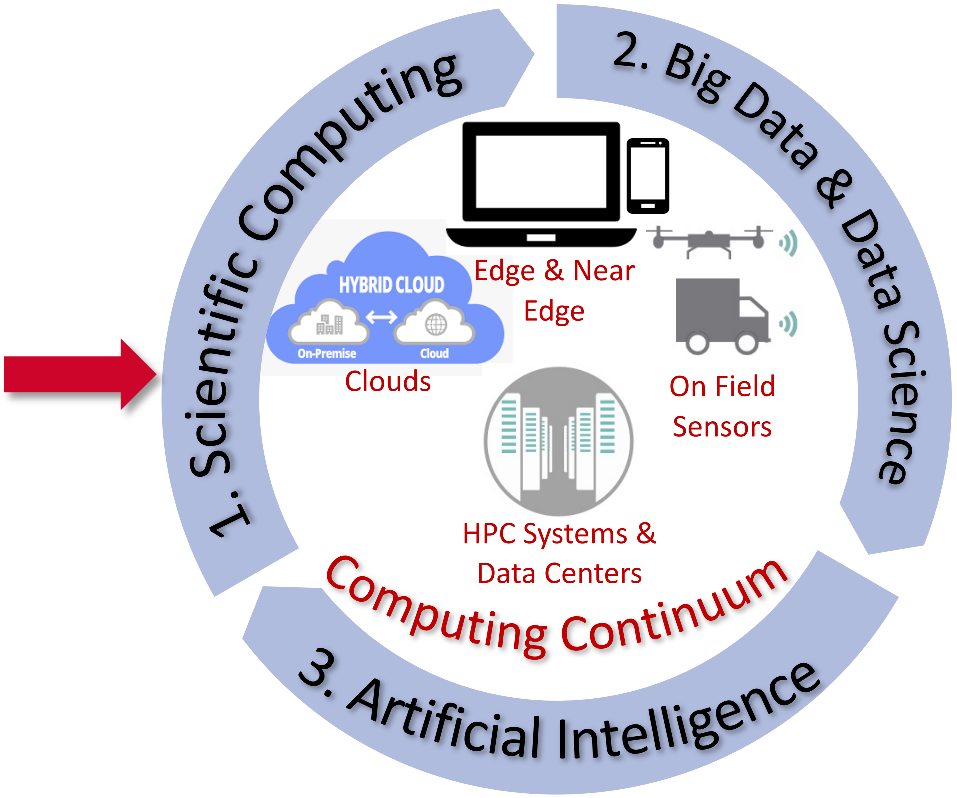
http://www.cse.ohio-state.edu/~panda

Computing has been evolving over the last three decades with multiple **phases:**

- **Phase 1 (1975-): Scientific Computing/HPC**
- **Phase 2 (2000-): HPC + Big Data Analytics**
- **Phase 3: (2010-): HPC + AI (Machine Learning/Deep Learning)**

Emergence of the Computing Continuum



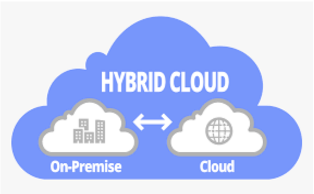


Data Movement and Control in Computing Continuum

Emerging Computing Continuum

On Field
Sensors

Edge & Near Edge

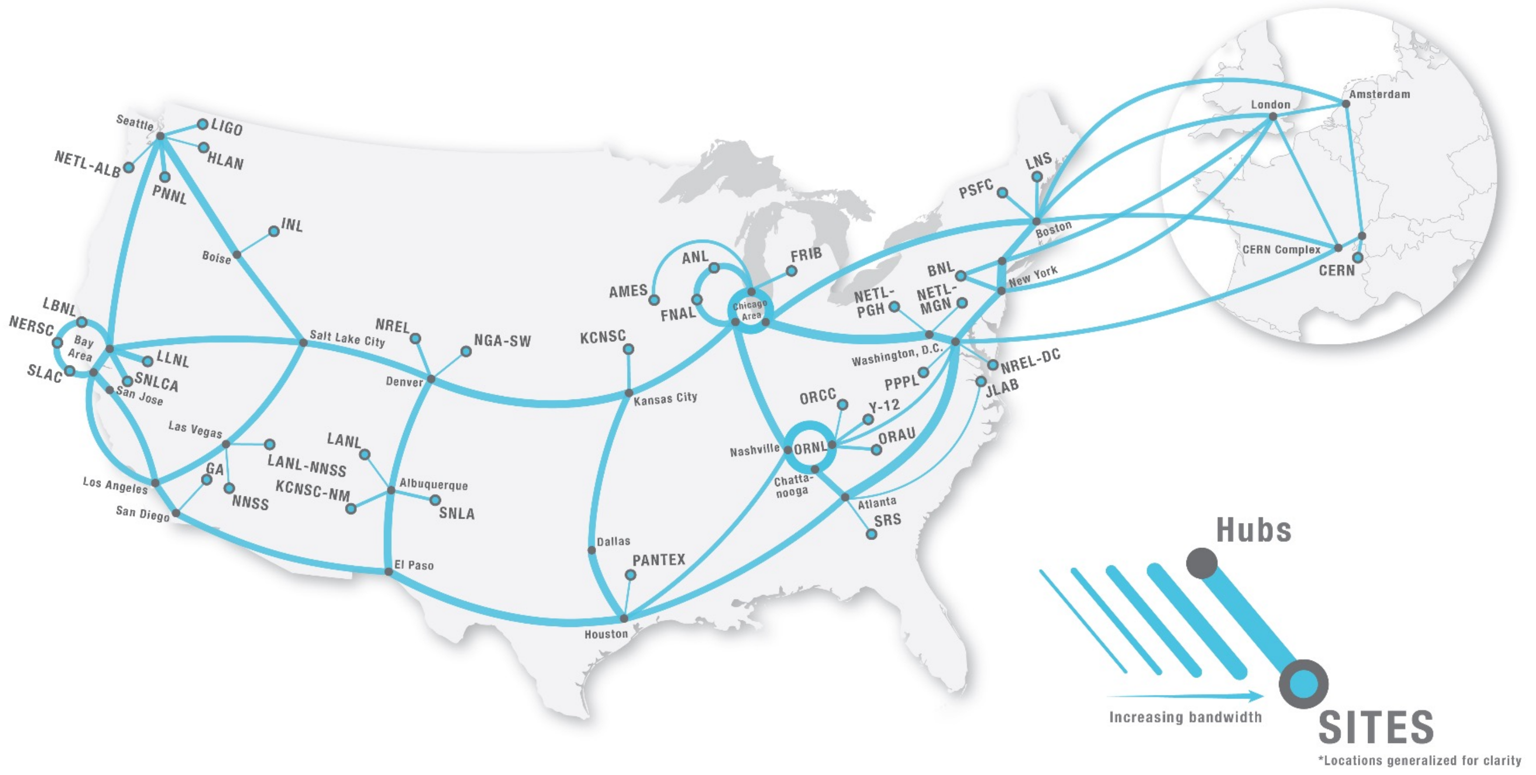


Clouds

HPC Systems & Data
Centers

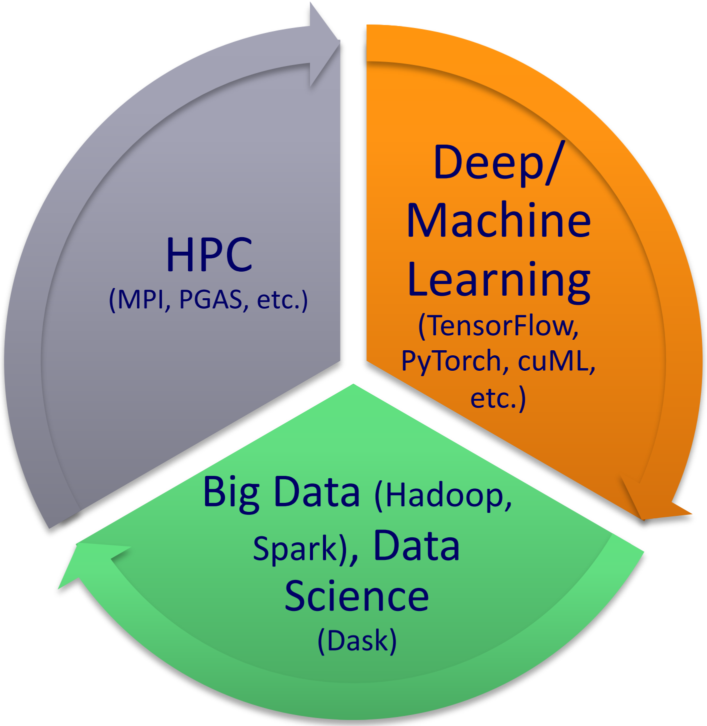
Data Movement and Control

ESnet6



Seattle
LIGO
HLAN
NETL-ALB
PNNL
INL
Boise
LBNL
NERSC
Bay Area
SLAC
LLNL
SNLCA
San Jose
Salt Lake City
NREL
NGA-SW
Denver
Las Vegas
LANL
LANL-NNSS
GA
NNSS
KCNSC-NM
Albuquerque
SNLA
Los Angeles
San Diego
El Paso
KCNSC
Kansas City
AMES
ANL
FNAL
Chicago Area
FRIB
Dallas
PANTEX
Houston
NETL-PGH
NETL-MGN
BNL
New York
Boston
PSFC
LNS
Washington, D.C.
NREL-DC
JLAB
PPPL
ORCC
Y-12
ORAU
Nashville
ORNL
Chatta-nooga
Atlanta
SRS
London
Amsterdam
CERN Complex
CERN
Hubs
Increasing bandwidth
SITES
*Locations generalized for clarity

Increasing Usage of HPC, AI, and Data Science in multiple Disciplines with Computing Continuum



Convergence of HPC, Deep/Machine Learning, and Data Science!

Increasing Need to Run these applications on the Cloud!!

MPI-Driven Middleware increasingly being used for all three domains

Many Examples

- Digital Agriculture
- Smart Cities
- Smart Manufacturing
- Smart Transportation
- Real-time Surveillance
- Computational Medicine (Pathology, Radiology, ..)

Designing Intelligent Cyberinfrastructure for Computing Continuum

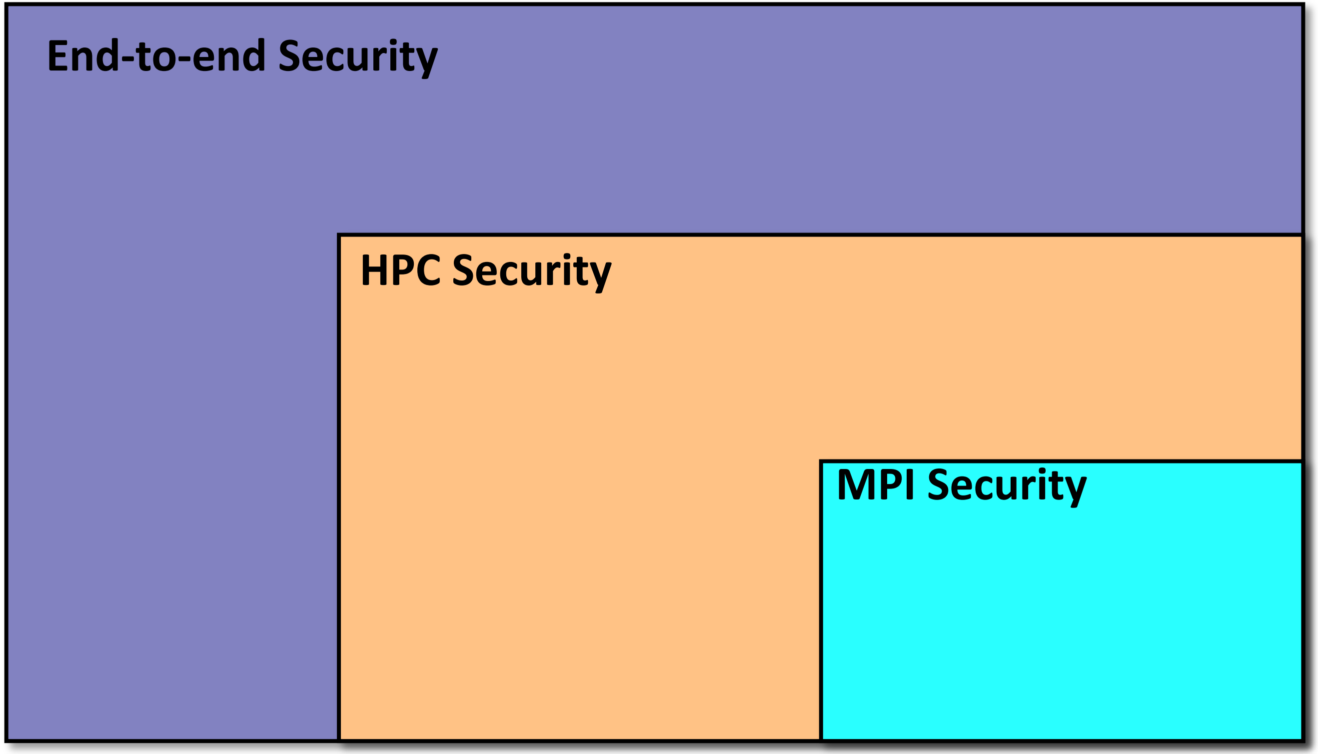
NSF-AI Institute ICICLE (icicle.ai)

Keynote Talk on Wednesday morning (8:30-9:15 am)

Broad Challenge:

How to provide end-to-end security for these emerging applications on computing continuum with HPC systems?

Three Levels of Security Support:



Presentation Overview

- **Challenges in HPC Security**
- Challenges in MPI Security
- Overview of the MVAPICH MPI Library Project
- Examples of MPI Security Solutions
- Conclusions

HPC Overview

Check for updates

NIST Special Publication 800
NIST SP 800-223

High-Performance Computing Security

Architecture, Threat Analysis, and Security Posture

Yang Guo
Ramaswamy Chandramouli
Lowell Wofford
Rickey Gregg
Gary Key
Antwan Clark
Catherine Hinton
Andrew Prout
Albert Reuther
Ryan Adamson
Aron Warren
Purushotham Bangalore
Erik Deumens
Csilla Farkas

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-223

NIST | NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

HPC Security – Multiple Zones

NIST SP 800-223
February 2024

High-Performance Computing Security

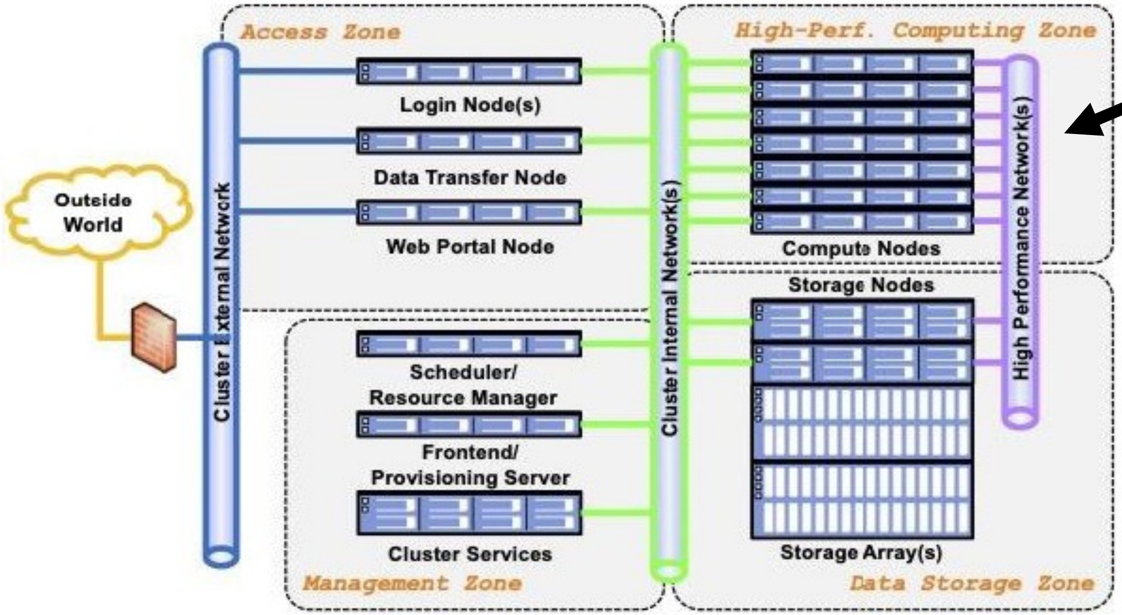


Fig. 1. HPC System Reference Architecture

MPI Security

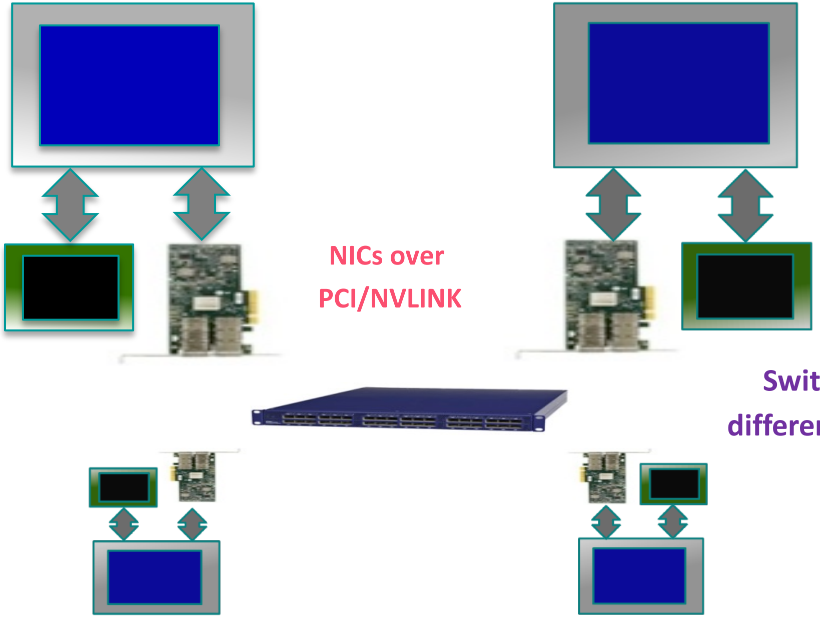
Presentation Overview

- Challenges in HPC Security
- **Challenges in MPI Security**
- Overview of the MVAPICH MPI Library Project
- Examples of MPI Security Solutions
- Conclusions

A Typical Cluster Configuration (On-Premise or Cloud)

CPUs/Cores connected with shared memory

GPUs connected with NVLink



Switches with different topologies

User Sharing Modes on a Cluster (On-Premise or Cloud)

- It is changing over the years
- Single user/VM per node in the past, getting obsolete
- Nodes are having
  - Large number of CPU cores
  - Large number of GPUs
  - Multiple NICs
- To increase system utilization, schedulers are allowing multiple users (VMs) to share a single node
  - Sharing of shared-memory
  - Sharing of GPUs
  - Sharing of NICs
  - Sharing of the switches
- Security threats from concurrently running MPI Jobs

MPI library has many Primitives

- Point-to-point
  - Inter-node
    - CPU-CPU
    - GPU-GPU
  - Intra-node
    - CPU-CPU
    - GPU-GPU
    - CPU_GPU
  - Operations could be blocking or non-blocking
  - Two-sided vs. one-sided (RMAs)
- Collectives (Broadcast, Alltoall, Allreduce, ….)
  - CPU-based or GPU-based
  - Algorithms involve inter-node and intra-node communication steps
  - May also incorporate in-network computing support of the switches
- Dynamic process management
- Many more primitives

Security support for all possible Communication primitives inside an MPI library

Delivering High-Performance

HPC community does not want Low-Performance while providing security support

Presentation Overview

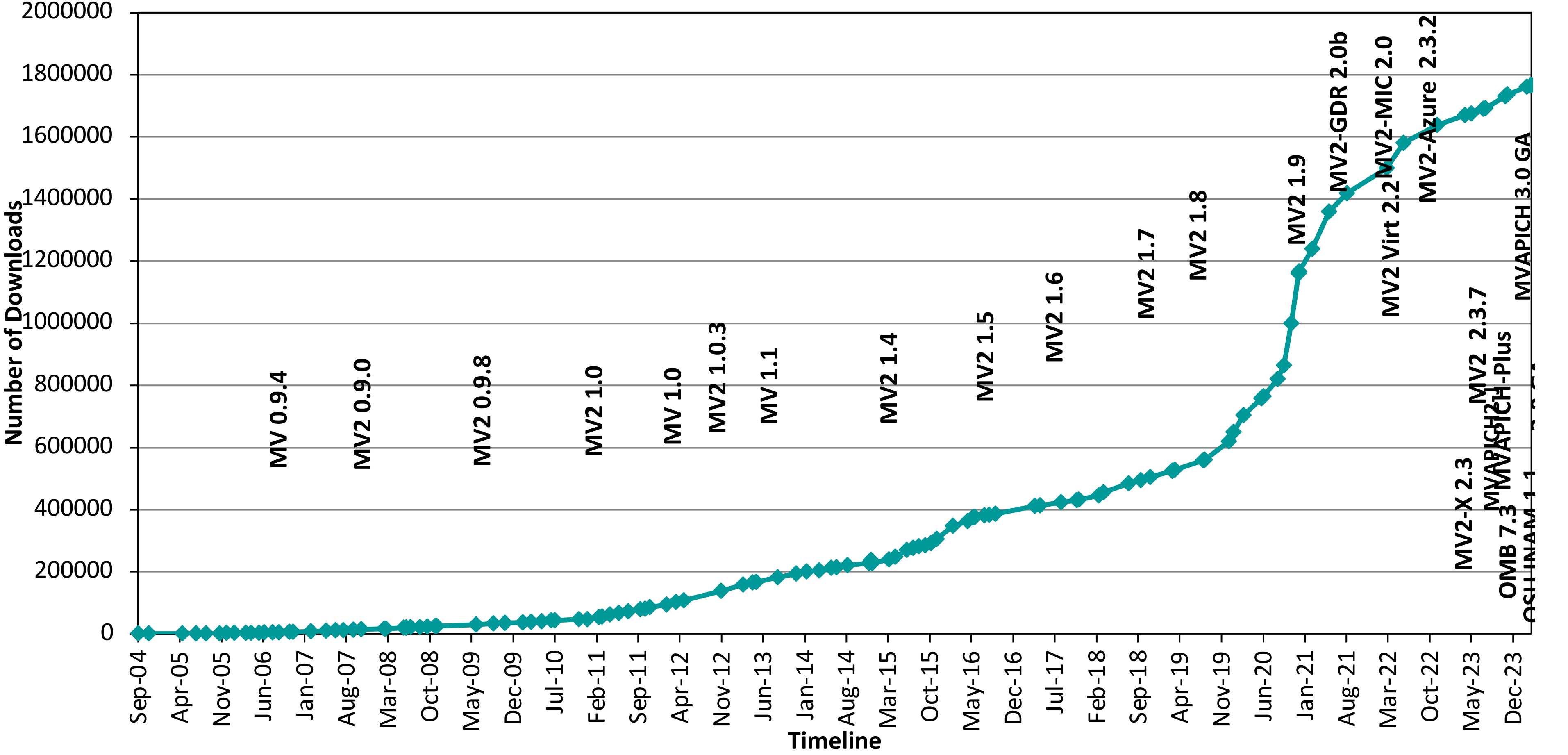
- Challenges in HPC Security
- Challenges in MPI Security
- **Overview of the MVAPICH MPI Library Project**
- Examples of MPI Security Solutions
- Conclusions

Overview of the MVAPICH Project

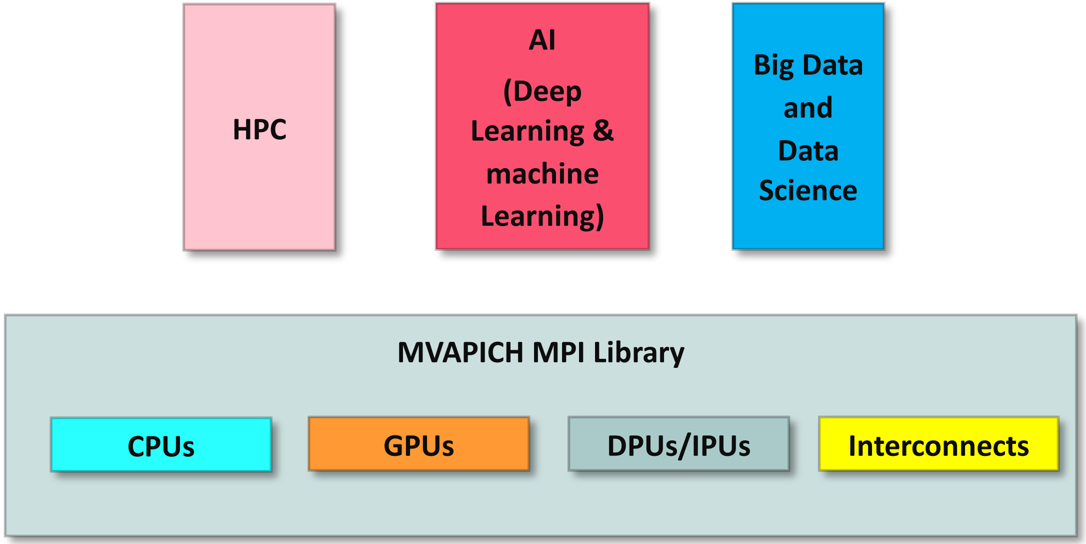
- High Performance open-source MPI Library
- Support for multiple interconnects
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, OPX, Broadcom RoCE, Intel Ethernet, Rockport Networks, Slingshot 10/11
- Support for multiple platforms
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD)
- Started in 2001, first open-source version demonstrated at SC '02
- Supports the latest MPI-3.1 standard
- http://mvapich.cse.ohio-state.edu
- Additional optimized versions for different systems/environments:
  - MVAPICH2-X (Advanced MPI + PGAS), since 2011
  - MVAPICH2-GDR with support for NVIDIA (since 2014) and AMD (since 2020) GPUs
  - MVAPICH2-MIC with support for Intel Xeon-Phi, since 2014
  - MVAPICH2-Virt with virtualization support, since 2015
  - MVAPICH2-EA with support for Energy-Awareness, since 2015
  - MVAPICH2-Azure for Azure HPC IB instances, since 2019
  - MVAPICH2-X-AWS for AWS HPC+EFA instances, since 2019
- Tools:
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

- **Used by more than 3,400 organizations in 92 countries**
- **More than 1.78 Million downloads from the OSU site directly**
- Empowering many TOP500 clusters (May '24 ranking)
  - 13th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 33rd , 448, 448 cores (Frontera) at TACC
  - 57th, 288,288 cores (Lassen) at LLNL and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 33rd ranked TACC Frontera system
- **Empowering Top500 systems for more than 18 years**

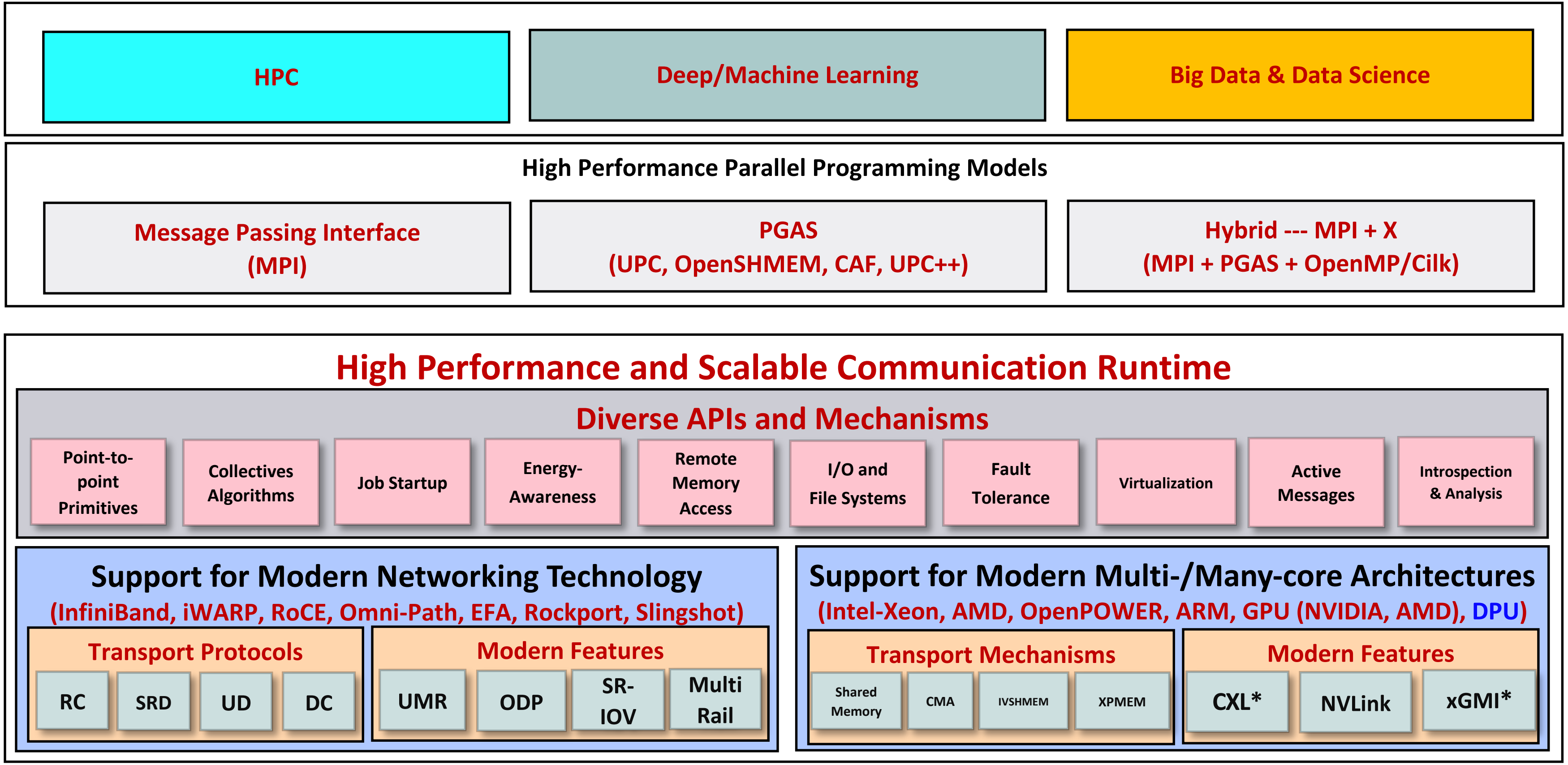
MVAPICH Release Timeline and Downloads



MPI (MVAPICH)-driven Converged Software Stack for HPC, AI, Big Data, and Data Science



MVAPICH Architecture (HPC, DL/ML, Big Data, & Data Science)



* Upcoming

Presentation Overview

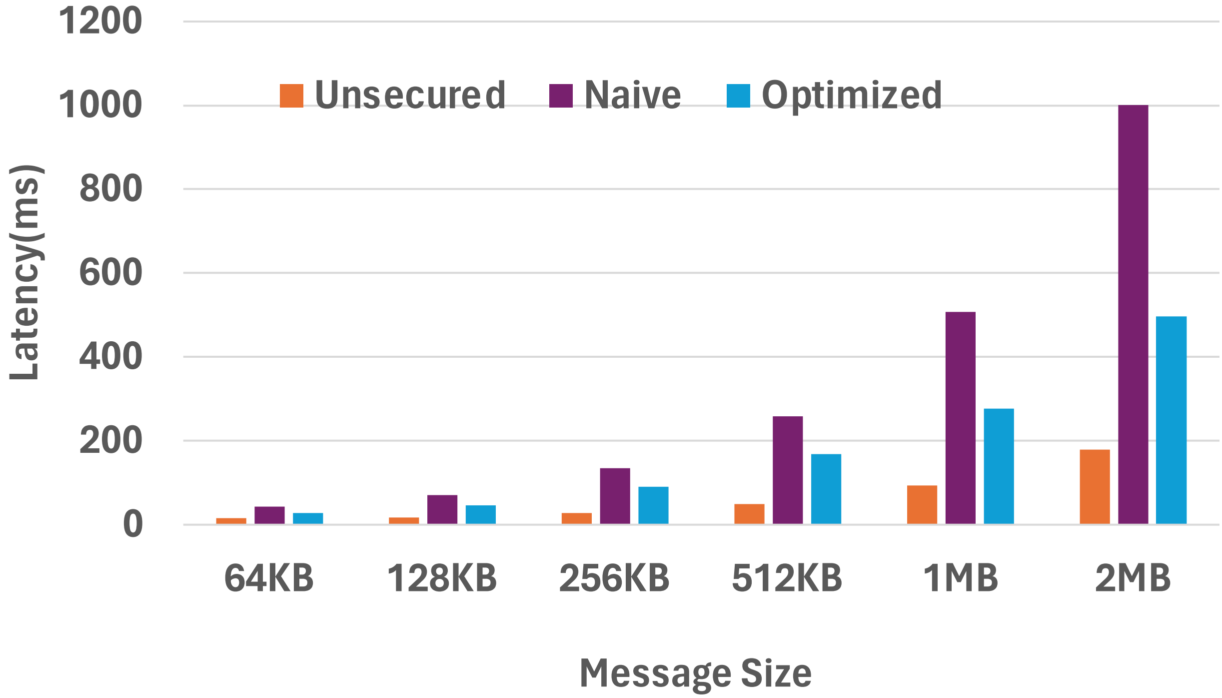
- Challenges in HPC Security
- Challenges in MPI Security
- Overview of the MVAPICH MPI Library Project
- **Examples of MPI Security Solutions**
- Conclusions

Secured MPI – An Example

- Jointly done with X-ScaleSolutions (x-scalesolutions.com) under a DOE SBIR grant
  - An industry sponsor of this workshop
- Scalable solutions of secure communication middleware based on the OSU MVAPICH2 library
- Flexible Support for multiple cryptographic libraries and encryption schemes, configurable per user request
- Supports SSL/TLS encryption protocol
- Supports secured point-to-point communication operations (blocking and non-blocking) for inter-node communication
- Supports collective operations including broadcast, alltoall, and allgather
- Tested with MPI micro-benchmarks and MPI applications up to 1,024 ranks

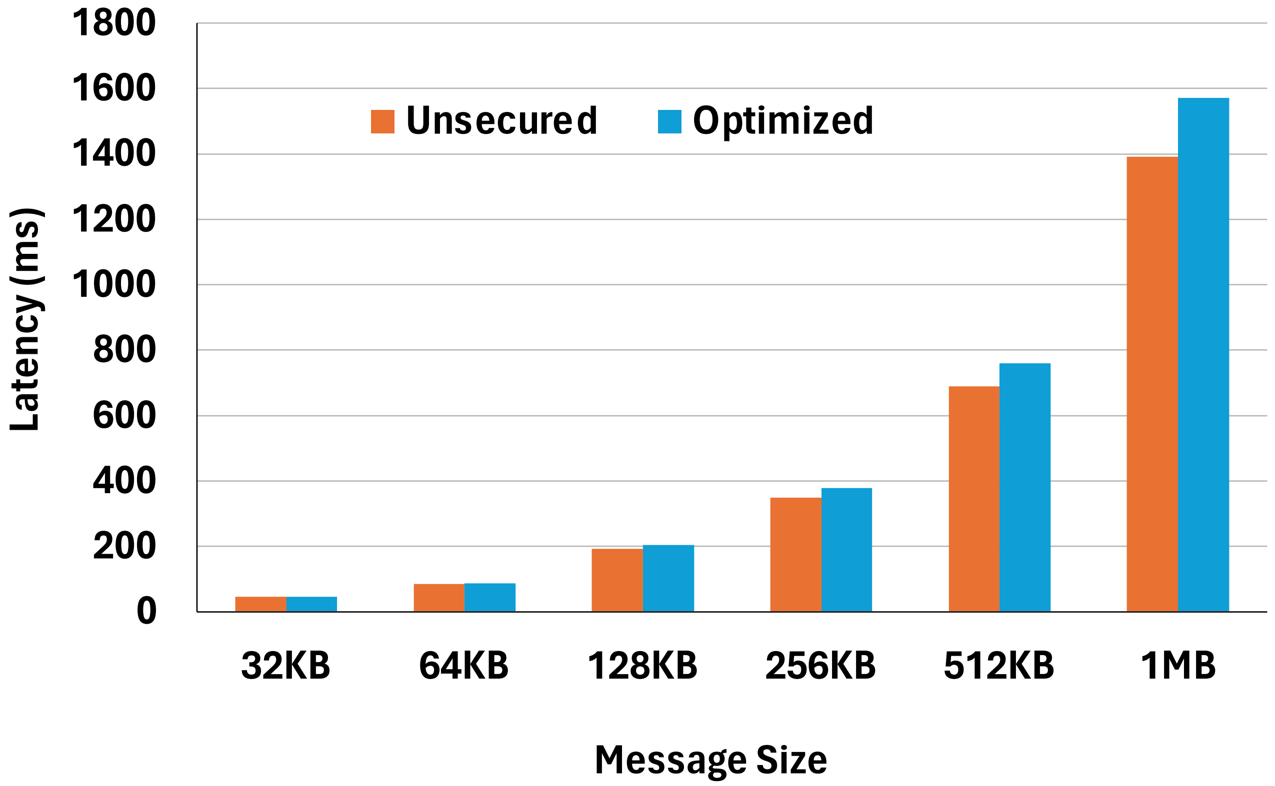
Secured MPI Performance: OSU_Latency Micro-benchmark

- Blocking point-to-point send/recv
- 2 nodes, 1 ppn



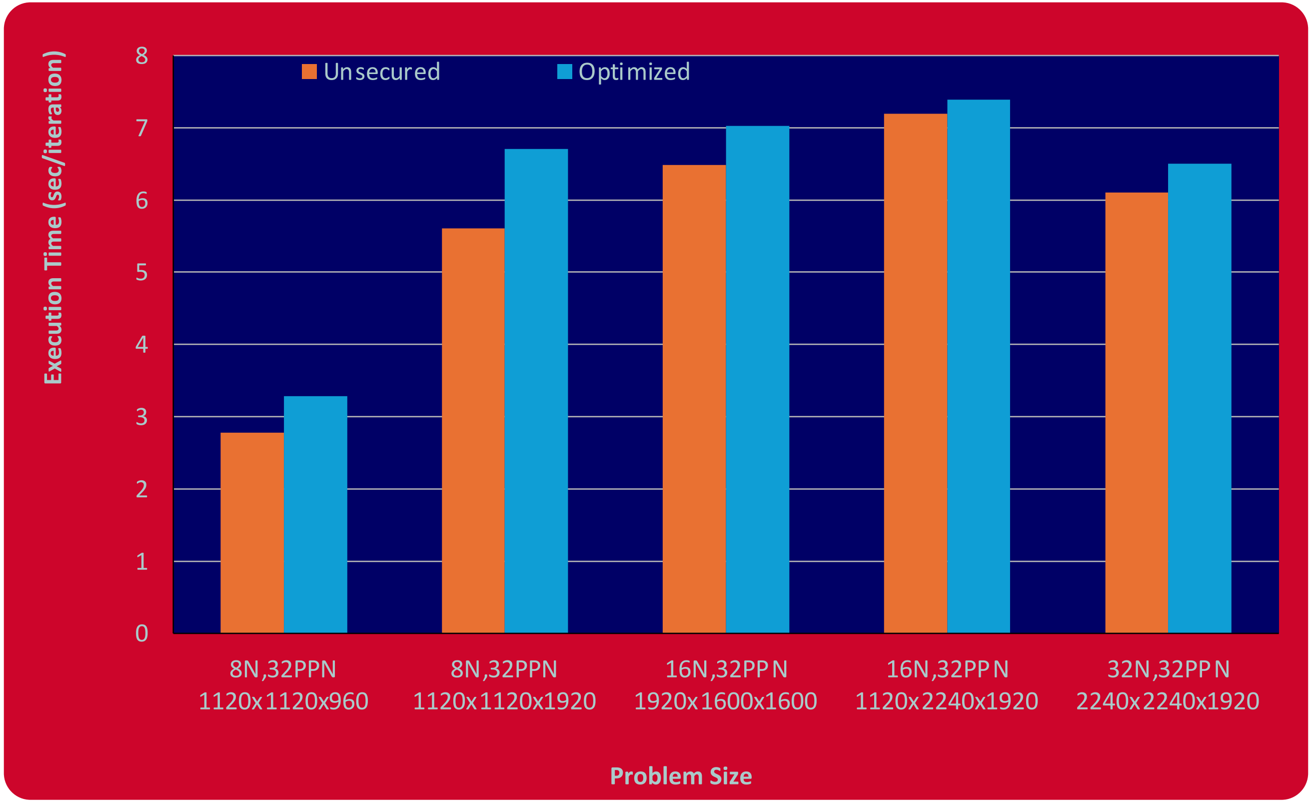
Secured MPI Performance: OSU_Alltoall Micro-benchmark

- Blocking alltoall operation
- 16 nodes,32 ppn
- 1-13% overhead based on message size



Secured MPI Performance: P3DFFT Application Kernel

- Parallel 3D FFT application kernel with various problem sizes
- Up to 32 nodes,32 ppn (1,024 processes). Includes both internode and intranode communication
- 6-20% overhead



Presentation Overview

- Challenges in HPC Security
- Challenges in MPI Security
- Overview of the MVAPICH MPI Library Project
- Examples of MPI Security Solutions
- **Conclusions**

Concluding Remarks

- Upcoming Exascale systems and Cloud are being designed with a holistic view of HPC, Big Data, Deep/Machine Learning, Data Science, and computing continuum
- Presented an overview of opportunities and challenges in providing MPI-level security for these systems
- Presented an example of secured MPI design with sample performance numbers
- The results demonstrate that MPI-level security can be supported with high-performance
- Such designs will lead to providing HPC-level security and end-to-end-security for next-generation systems and applications

Thank You!

panda@cse.ohio-state.edu

Network-Based Computing Laboratory

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

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/