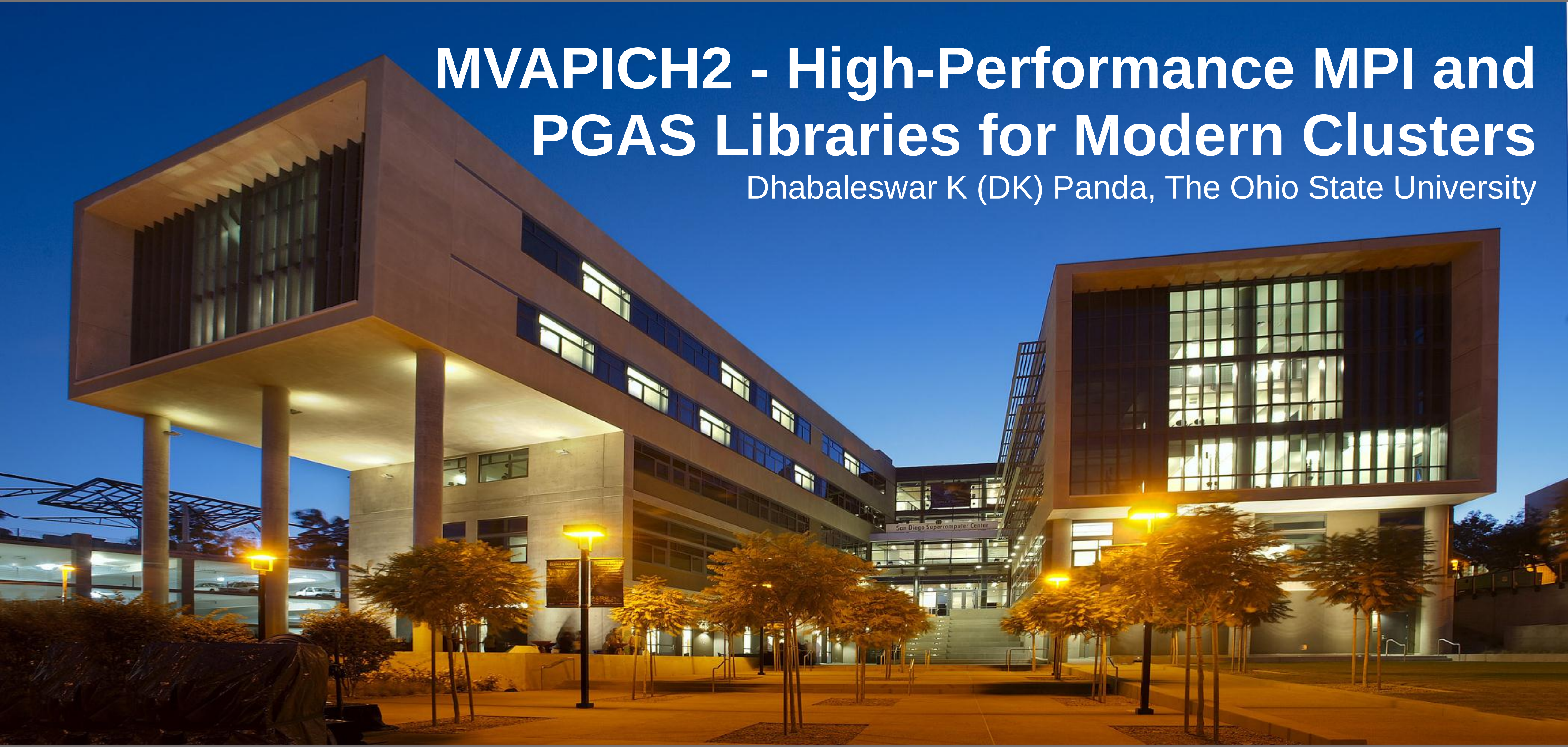
MVAPICH2 - High-Performance MPI and PGAS Libraries for Modern Clusters

Dhabaleswar K (DK) Panda, The Ohio State University

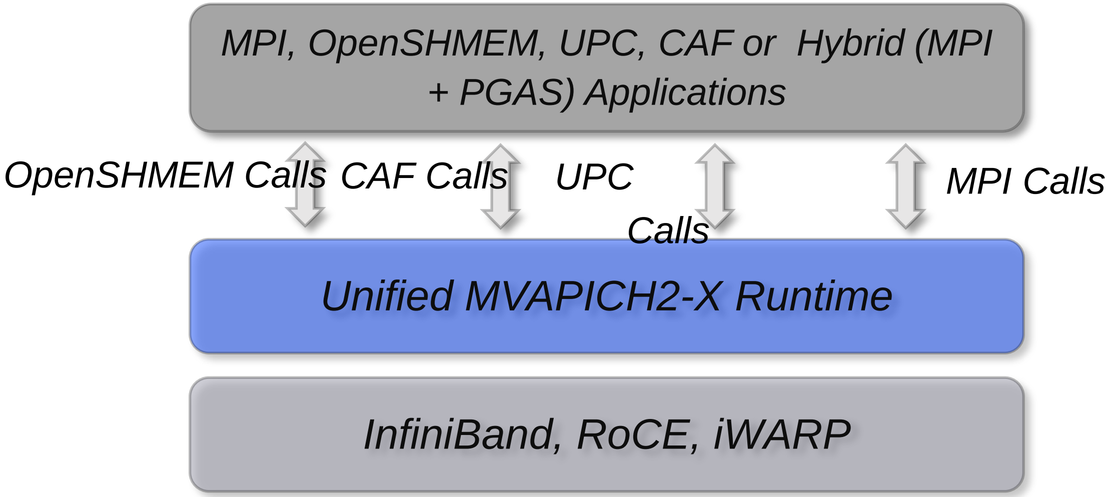


SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

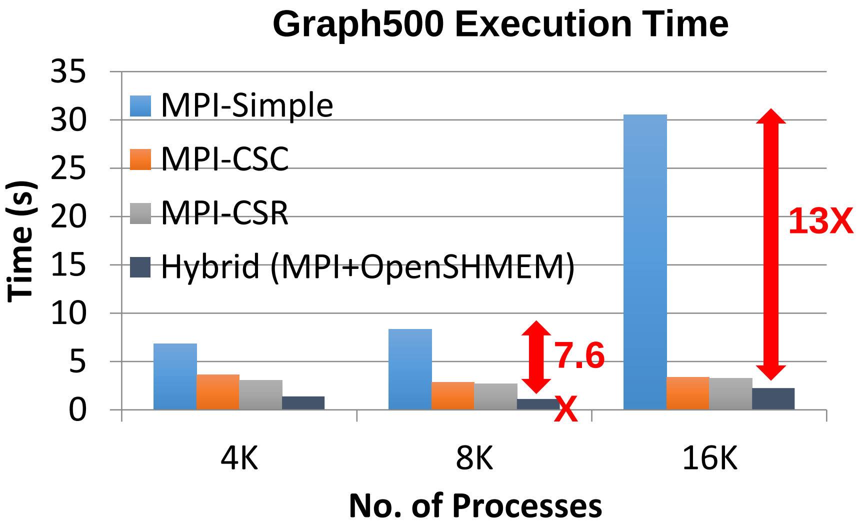
Overview of MVAPICH2 for MPI and PGAS

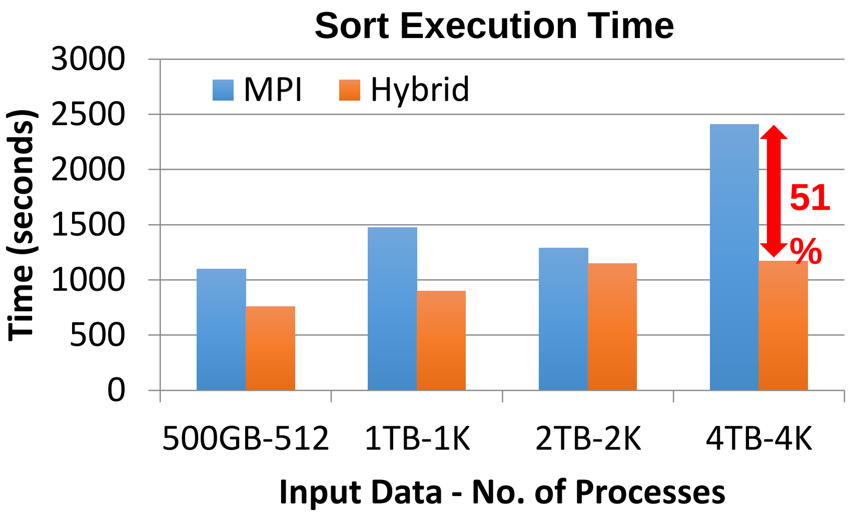
- High performance open-source MPI for InfiniBand, 10-40Gig/iWARP, and RoCE
  - **MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002**
  - **MVAPICH2-X (MPI + PGAS), Available since 2011**
  - **Support for GPGPUs (MVAPICH2-GDR) , Available since 2014**
  - **Support for MIC (MVAPICH2-MIC), Available since 2014**
  - **Support for Virtualization (MVAPICH2-Virt), Available since 2015**
    - **To be Used for Comet@SDSC**
  - **Support for Energy-Awareness (MVAPICH2-EA), Available since 2015**
  - **Used by more than 2,475 organizations in 76 countries**
  - **More than 307,000 downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov'15 ranking)
    - 10th ranked 519,640-core cluster (Stampede) at TACC
    - 13th ranked 185,344-core cluster (Pleiades) at NASA
    - 25th ranked 76,032-core cluster (Tsubame 2.5) at Tokyo Institute of Technology
    - 350th ranked 16,160-core cluster (Gordon) at SDSC
  - http://mvapich.cse.ohio-state.edu
- Empowering Top500 systems for over a decade


MPI, OpenSHMEM, UPC, CAF or Hybrid (MPI + PGAS) Applications
OpenSHMEM Calls
CAF Calls
UPC Calls
MPI Calls
Unified MVAPICH2-X Runtime
InfiniBand, RoCE, iWARP

- Unified communication runtime for MPI, UPC, OpenSHMEM, CAF
- Available with MVAPICH2-X 1.9 (2012) onwards!
  - http://mvapich.cse.ohio-state.edu
- Feature Highlights
  - Supports MPI(+OpenMP), OpenSHMEM, UPC, CAF, MPI(+OpenMP) + OpenSHMEM, MPI(+OpenMP) + UPC + CAF
  - MPI-3 compliant, OpenSHMEM v1.0h standard compliant, UPC v1.2 standard compliant (with initial support for UPC 1.3), CAF 2008 standard (OpenUH)
  - Scalable Inter-node and intra-node communication – point-to-point and collectives

Hybrid MPI+PGAS Performance with Graph500 and Sort





- *Performance of Hybrid (MPI+ OpenSHMEM) Graph500 Design*
  - *8,192 processes*
    - *- 2.4X improvement over MPI-CSR*
    - *- 7.6X improvement over MPI-Simple*
  - *16,384 processes*
    - *- 1.5X improvement over MPI-CSR*
    - *- 13X improvement over MPI-Simple*

- *Performance of Hybrid (MPI+OpenSHMEM) Sort Application*
  - *4,096 processes, 4 TB Input Size*
    - *- MPI – 2408 sec; 0.16 TB/min*
    - *- Hybrid – 1172 sec; 0.36 TB/min*
    - *- 51% improvement over MPI-design*

J. Jose, K. Kandalla, S. Potluri, J. Zhang and D. K. Panda, Optimizing Collective Communication in OpenSHMEM, Int'l Conference on Partitioned Global Address Space Programming Models (PGAS '13), October 2013.

J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models, International Supercomputing Conference (ISC'13), June 2013

J. Jose, K. Kandalla, M. Luo and D. K. Panda, Supporting Hybrid MPI and OpenSHMEM over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP '12), September 2012