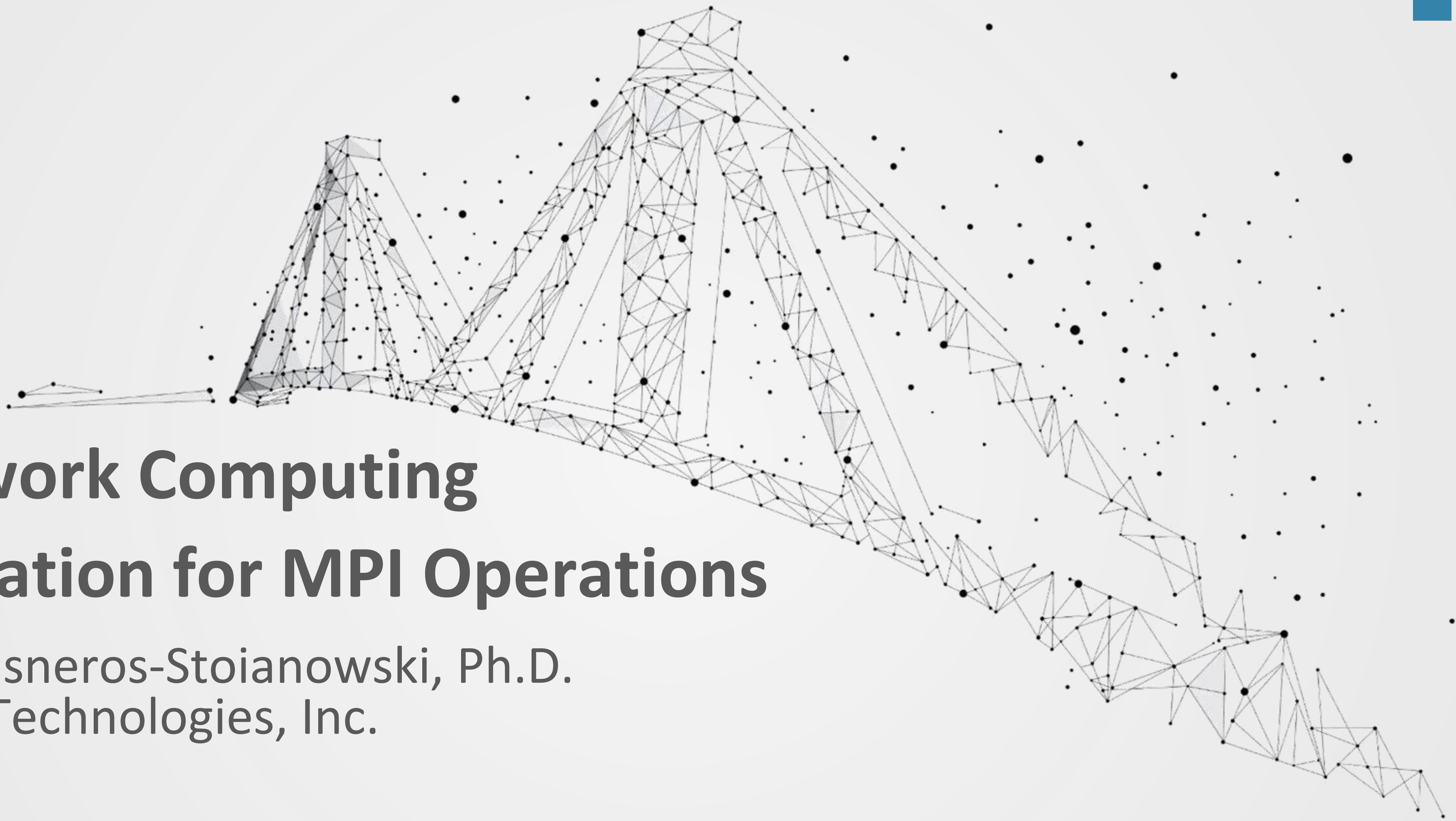


In-Network Computing Acceleration for MPI Operations

Gerardo Cisneros-Stoianowski, Ph.D.
Mellanox Technologies, Inc.

September2018

Mellanox Accelerates Leading HPC and AI Systems

World’s Top 3 Supercomputers

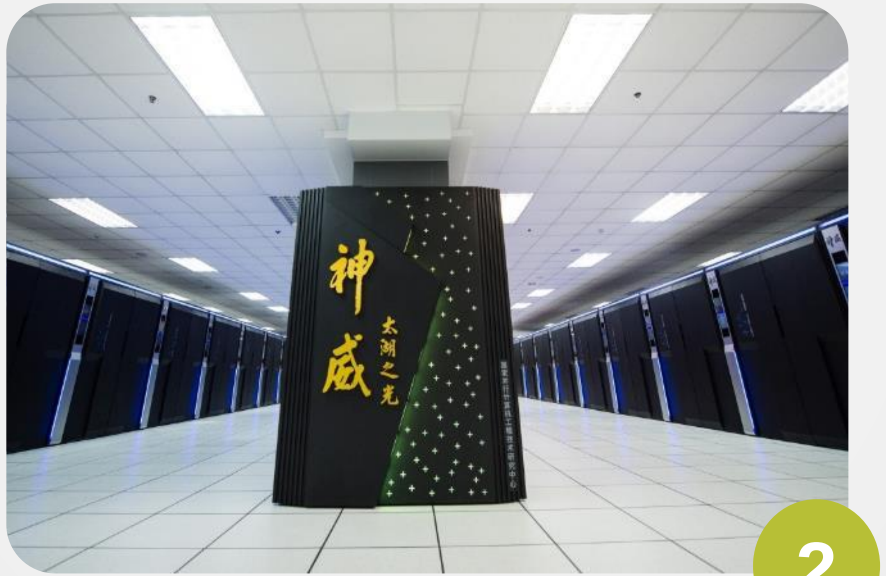
Oak Ridge National Laboratory
U.S. Department of Energy

1

Summit CORAL System
World’s Fastest HPC / AI System
9.2K InfiniBand Nodes

TOP 500 The List.

国家超级计算无锡中心
National Supercomputing Center in Wuxi



2

Wuxi Supercomputing Center
Fastest Supercomputer in China
41K InfiniBand Nodes

TOP 500 The List.

Lawrence Livermore National Laboratory
U.S. Department of Energy

3

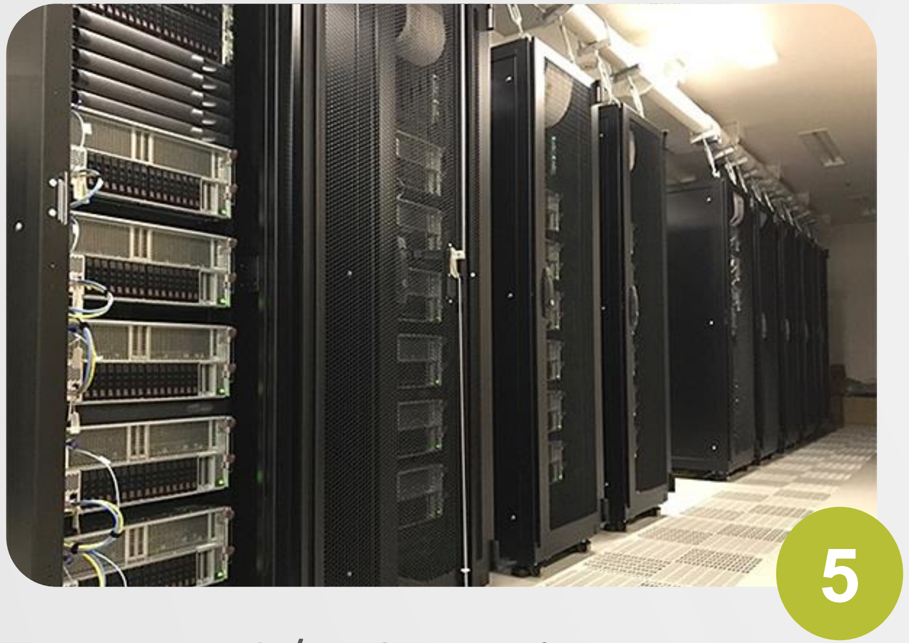
Sierra CORAL System
#2 USA Supercomputer
8.6K InfiniBand Nodes

TOP 500 The List.

Mellanox Accelerates Leading HPC and AI Systems

(Examples)

NATIONAL INSTITUTE OF
ADVANCED INDUSTRIAL SCIENCE
AND TECHNOLOGY (AIST)



Fastest HPC / AI System in Japan
1.1K InfiniBand Nodes

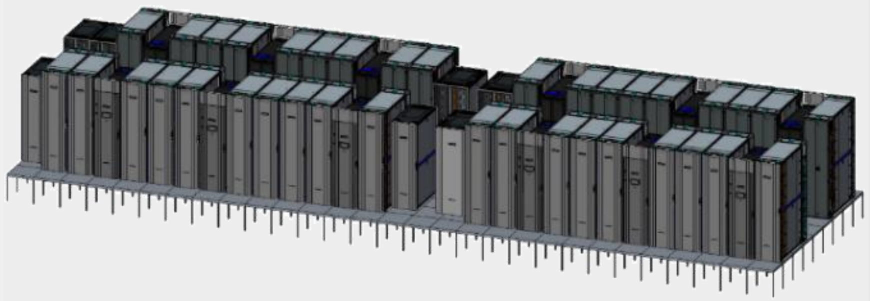
TOP 500 The List.

The world's Fastest Industry Supercomputer
1.6K InfiniBand Nodes

TOP 500 The List.

Sandia National Laboratories

NNSA National Nuclear Security Administration



'Astra' Arm-Based Supercomputer
NNSA Vanguard Program
2.6K InfiniBand Nodes

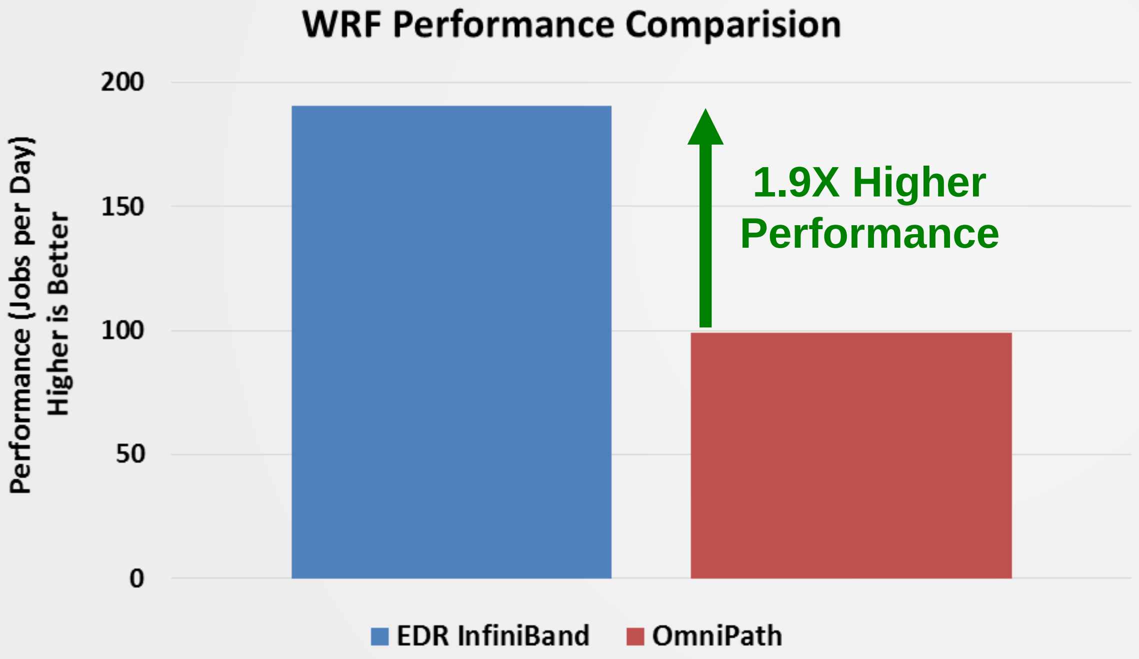
To be Listed Nov'18 (TOP100)

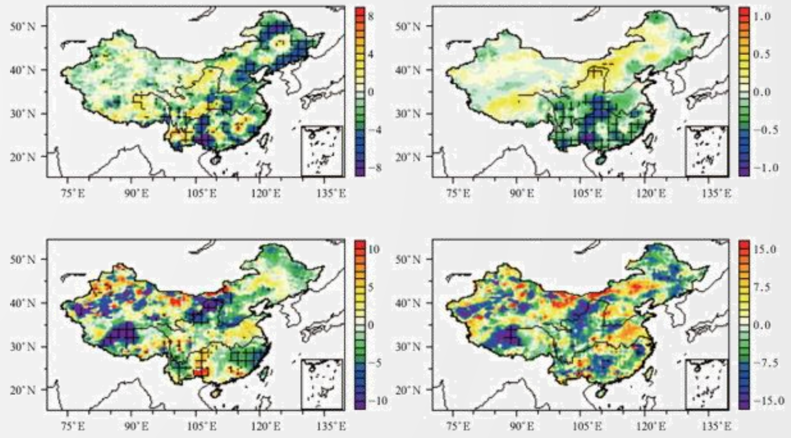
Chinese Weather Forecast Organization

1.9X Higher Performance

InfiniBand over OmniPath

Customer replaced OmniPath with InfiniBand



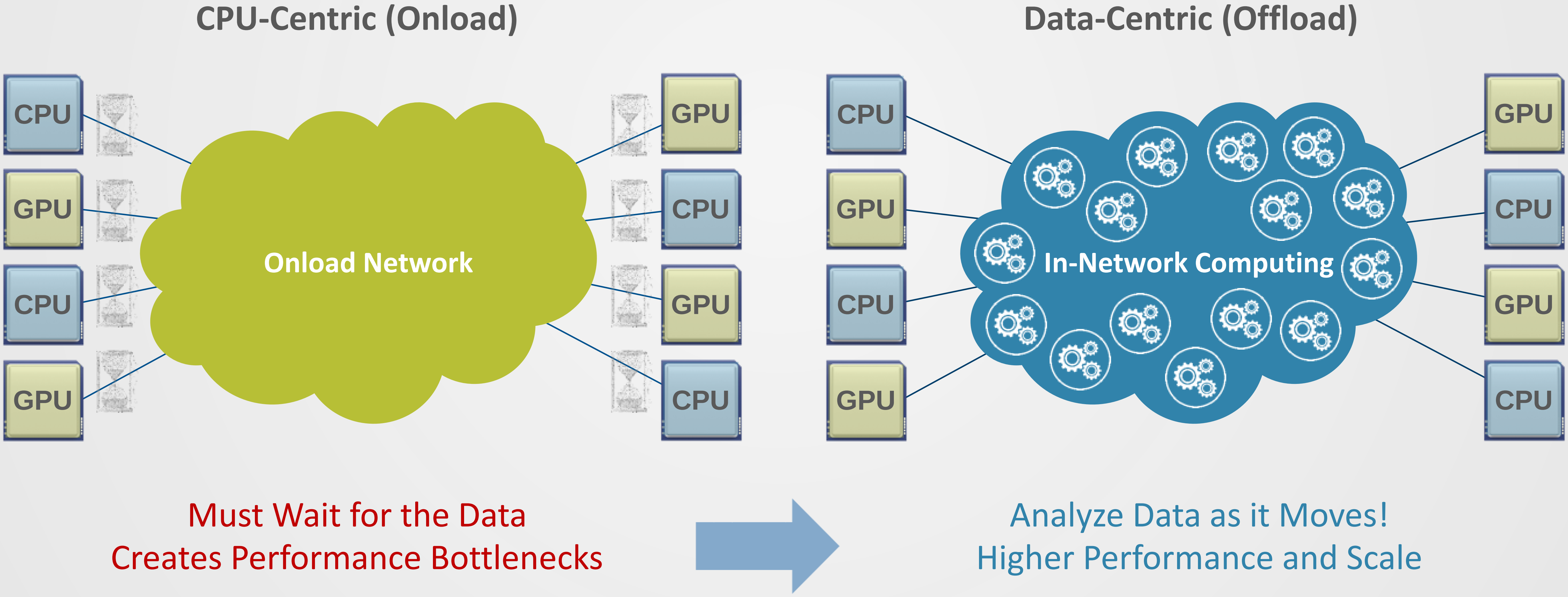


- Chinese weather forecast institute benchmarked InfiniBand and OmniPath
- For their customized WRF application, InfiniBand provides 92% higher performance
- As a result, the institute replaced its OmniPath connectivity with InfiniBand EDR

In Network Computing

The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale



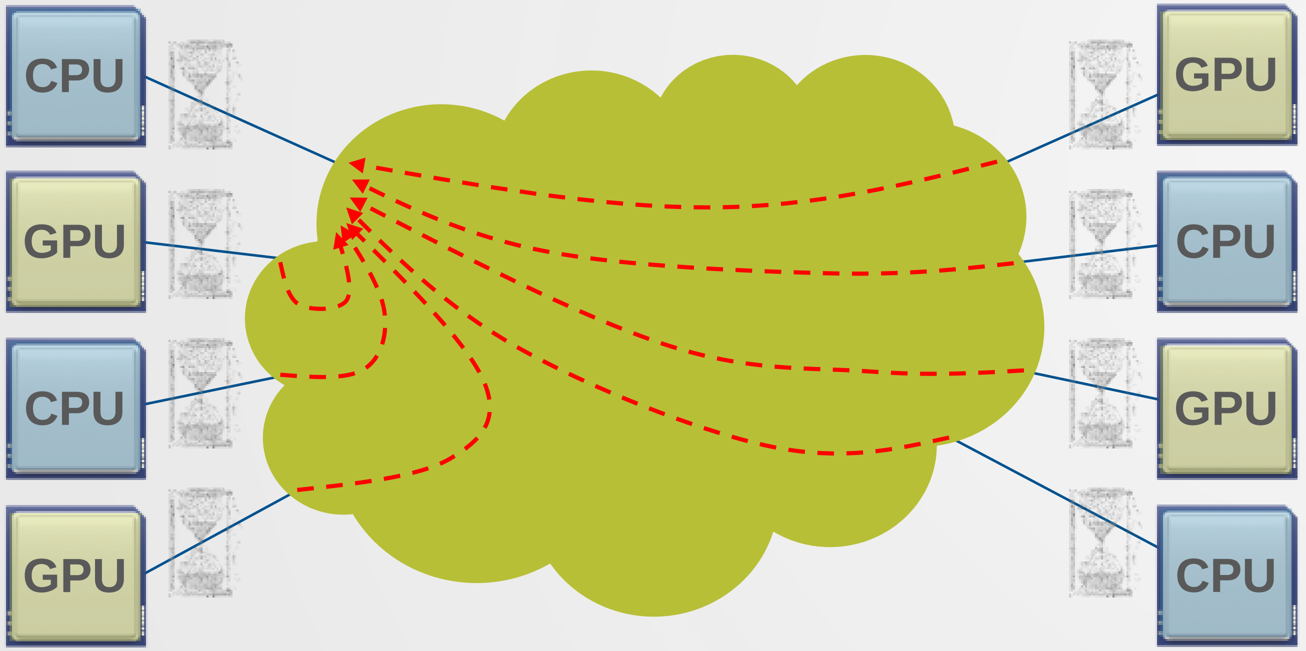
Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

Data Centric Architecture to Overcome Latency Bottlenecks

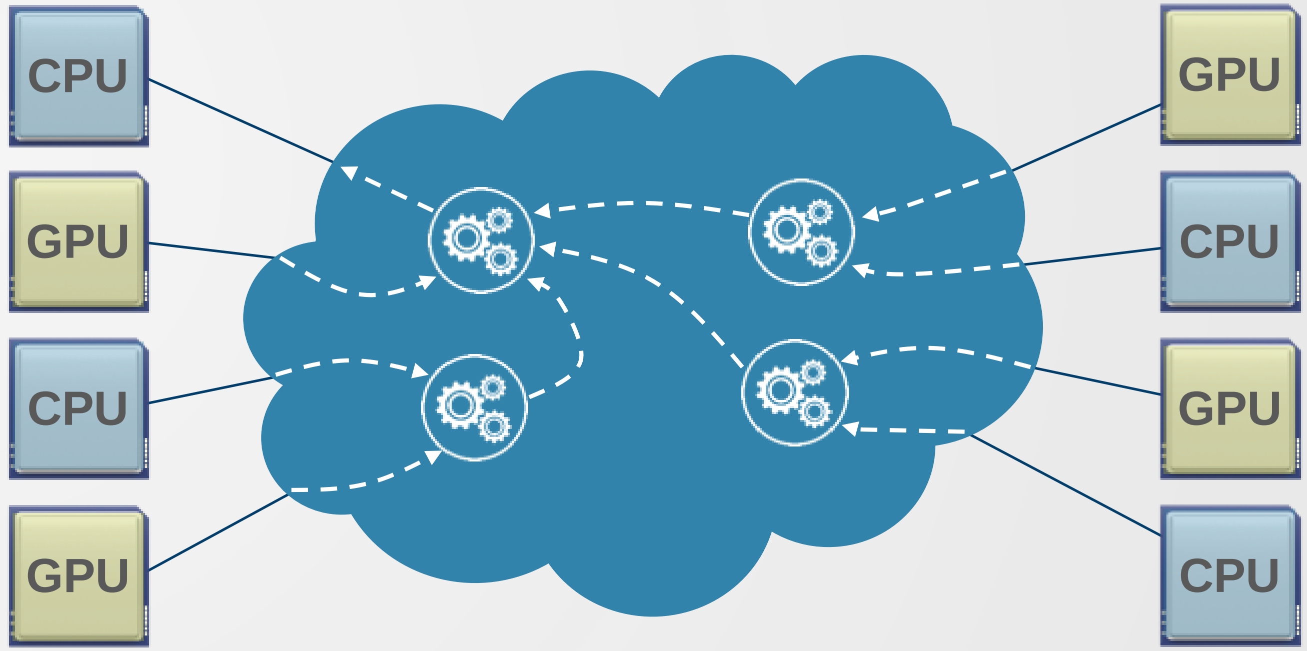
Intelligent Interconnect Paves the Road to Exascale Performance

CPU-Centric (Onload)



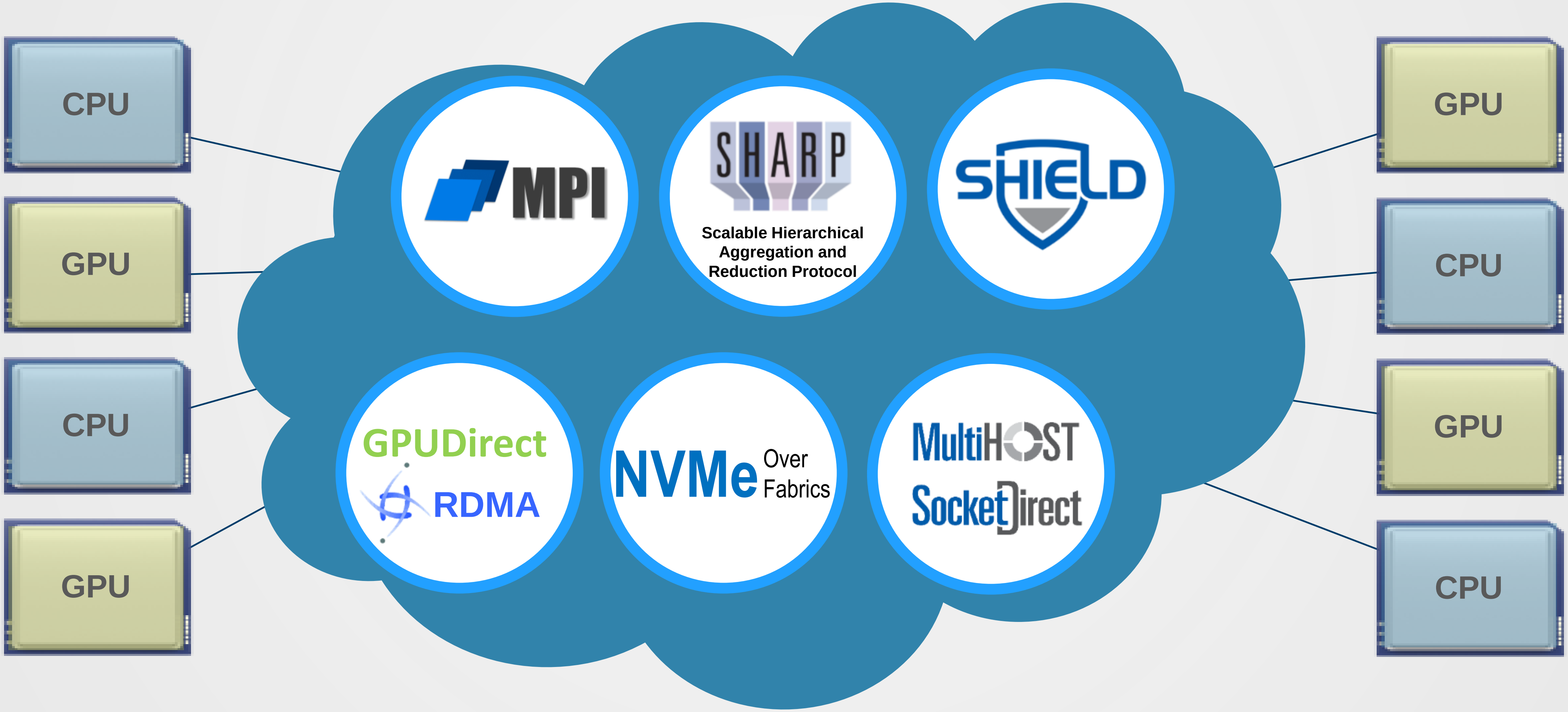
Communications Latencies
of 30-40us

Data-Centric (Offload)



Communications Latencies
of 3-4us

In-Network Computing to Enable Data-Centric Data Centers



HPC-X

Mellanox HPC-X™ Scalable HPC Software Toolkit

HPC-X™

- Complete MPI and OpenSHMEM package
- Optimal application performance
- For commercial and open source applications
- Best out of the box experience
- Can be downloaded from http://www.mellanox.com/products/hpcx/

Mellanox HPC-X - Package Contents

- HPC-X – Mellanox Scalable HPC Toolkit
- Allow fast and simple deployment of HPC libraries
  - Both Stable & Latest Beta are bundled
  - All libraries are pre-compiled
  - Includes scripts/modulefiles to ease deployment

- Package Includes
  - OpenMPI and OpenSHMEM
  - UCX (Point-to-point communications)
  - MXM (Point-to-point communications – it is being replaced by UCX)
  - HCOLL (Collectives)
  - Profiling Tools
    - IPM
  - Standard Benchmarks
    - OSU
    - IMB

UCX

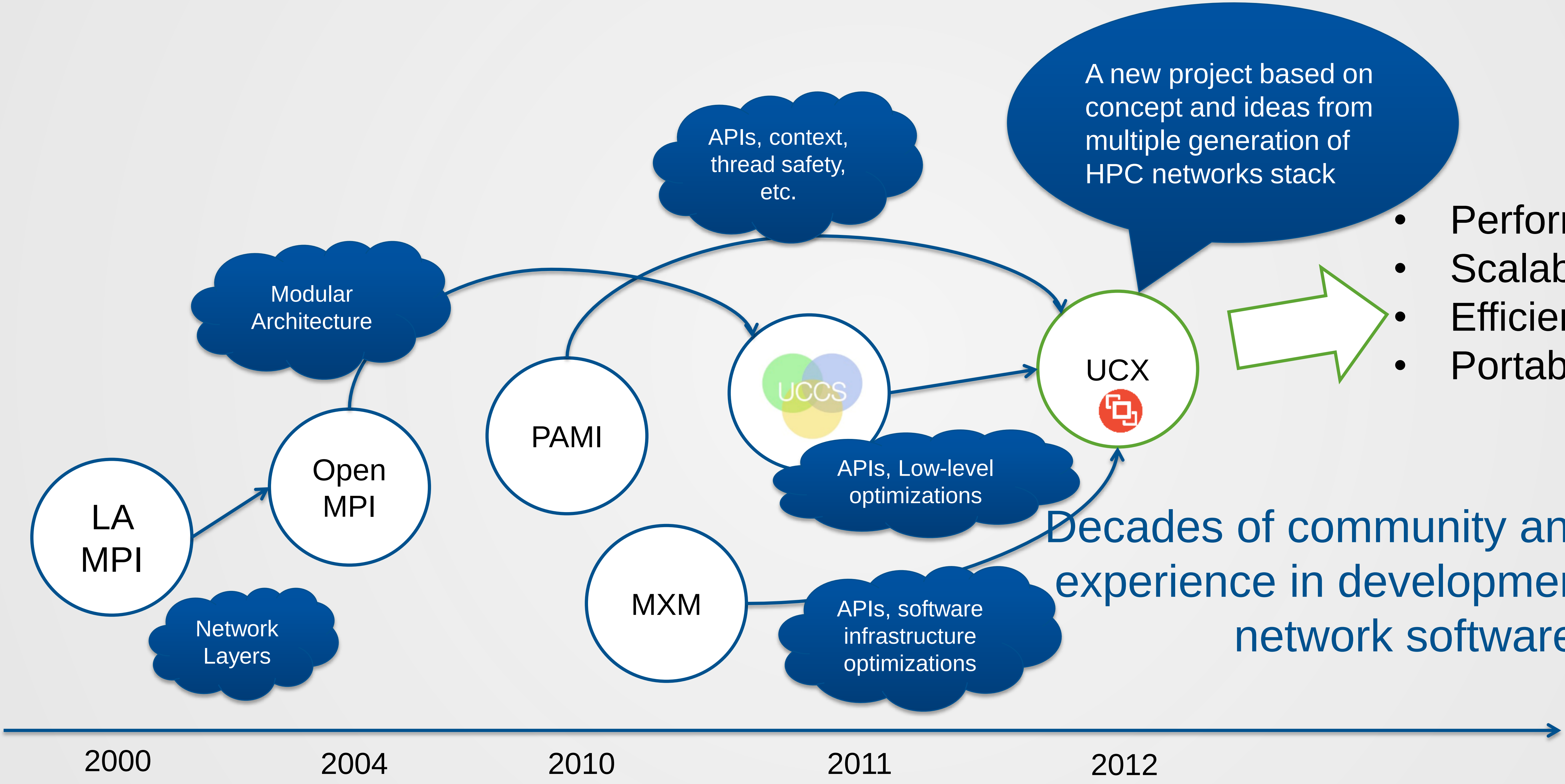
UCF Consortium

- Mission:
  - Collaboration between industry, laboratories, and academia to create production grade communication frameworks and open standards for data centric and high-performance applications

- Projects
  - UCX – Unified Communication X
  - Open RDMA

- Board members
  - **Jeff Kuehn**, UCF Chairman (Los Alamos National Laboratory)
  - **Gilad Shainer**, UCF President (Mellanox Technologies)
  - **Pavel Shamis**, UCF treasurer (ARM)
  - **Brad Benton**, Board Member (AMD)
  - **Duncan Poole**, Board Member (Nvidia)
  - **Pavan Balaji**, Board Member (Argonne National Laboratory)
  - **Sameh Sharkawi**, Board Member (IBM)
  - **Dhabaleswar K. (DK) Panda**, Board Member (Ohio State University)
  - **Steve Poole**, Board Member (Open Source Software Solutions)

UCX - History


A new project based on concept and ideas from multiple generation of HPC networks stack
APIs, context, thread safety, etc.
Modular Architecture
UCCS
UCX
PAMI
Open MPI
LA MPI
APIs, Low-level optimizations
MXM
Network Layers
APIs, software infrastructure optimizations
2000
2004
2010
2011
2012

- Performance
- Scalability
- Efficiency
- Portability

Decades of community and industry experience in development of HPC network software

UCX Framework Mission

- Collaboration between industry, laboratories, government (DoD, DoE), and academia
- Create open-source production grade communication framework for HPC applications
- Enable the highest performance through co-design of software-hardware interfaces

API

Exposes broad semantics that target data centric and HPC programming models and applications

Performance oriented

Optimization for low-software overheads in communication path allows near native-level performance

Production quality

Developed, maintained, tested, and used by industry and researcher community

Community driven

Collaboration between industry, laboratories, and academia

Research

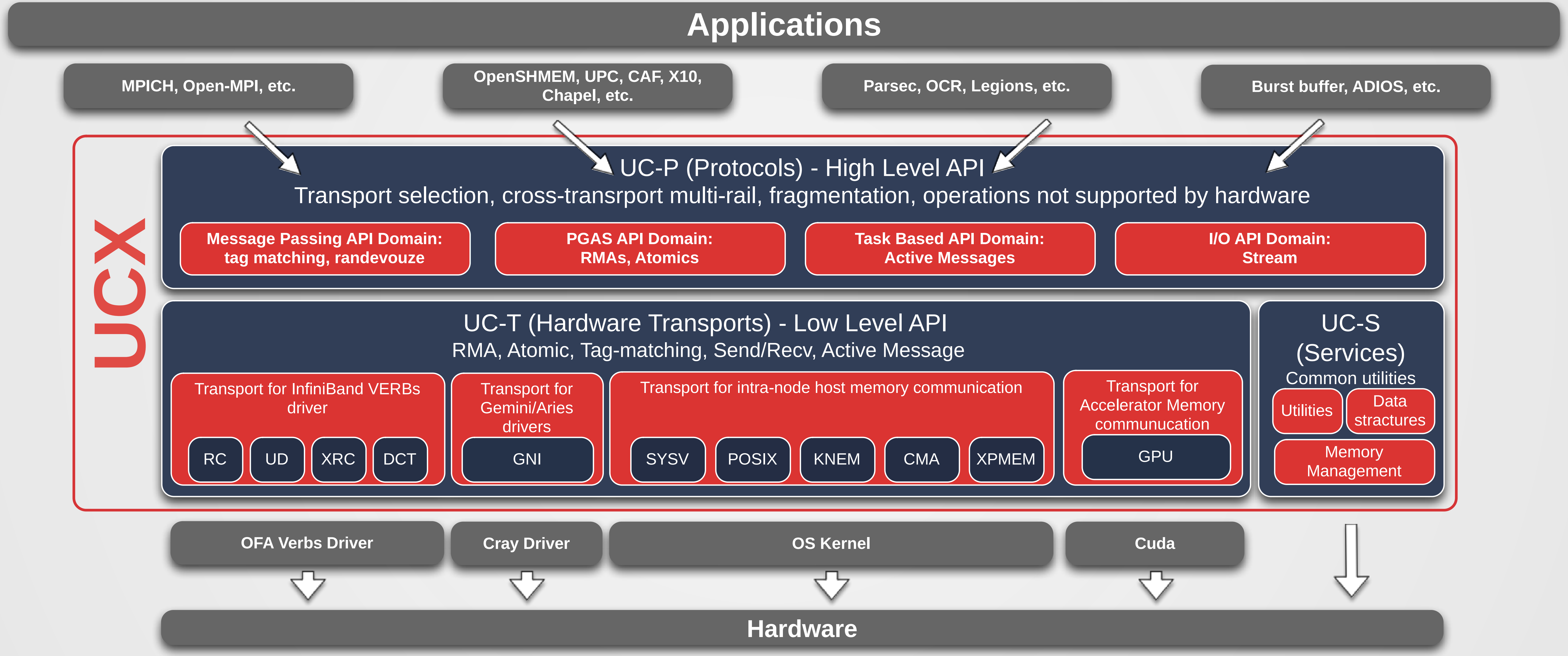
The framework concepts and ideas are driven by research in academia, laboratories, and industry

Cross platform

Support for Infiniband, Cray, various shared memory (x86-64, Power, ARMv8), GPUs

Co-design of Exascale Network APIs

UCX High-level Overview



Applications

MPICH, Open-MPI, etc.

OpenSHMEM, UPC, CAF, X10, Chapel, etc.

Parsec, OCR, Legions, etc.

Burst buffer, ADIOS, etc.

UCX

UC-P (Protocols) - High Level API

Transport selection, cross-transrport multi-rail, fragmentation, operations not supported by hardware

Message Passing API Domain: tag matching, randevouze

PGAS API Domain: RMAs, Atomics

Task Based API Domain: Active Messages

I/O API Domain: Stream

UC-T (Hardware Transports) - Low Level API

RMA, Atomic, Tag-matching, Send/Recv, Active Message

Transport for InfiniBand VERBs driver

RC

UD

XRC

DCT

Transport for Gemini/Aries drivers

GNI

Transport for intra-node host memory communication

SYSV

POSIX

KNEM

CMA

XPMEM

Transport for Accelerator Memory communucation

GPU

UC-S (Services)

Common utilities

Utilities

Data structures

Memory Management

OFA Verbs Driver

Cray Driver

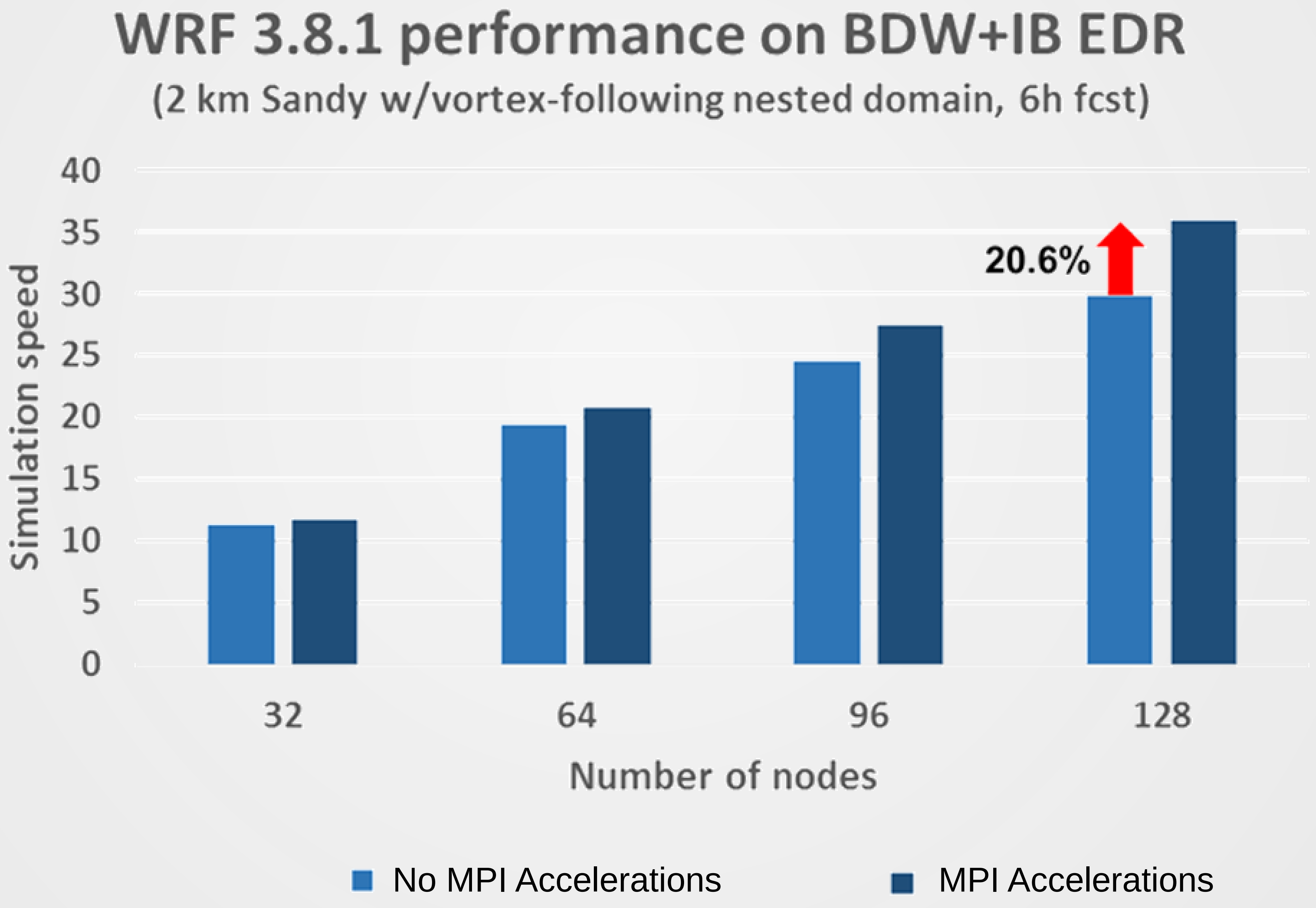
OS Kernel

Cuda

Hardware

WRF

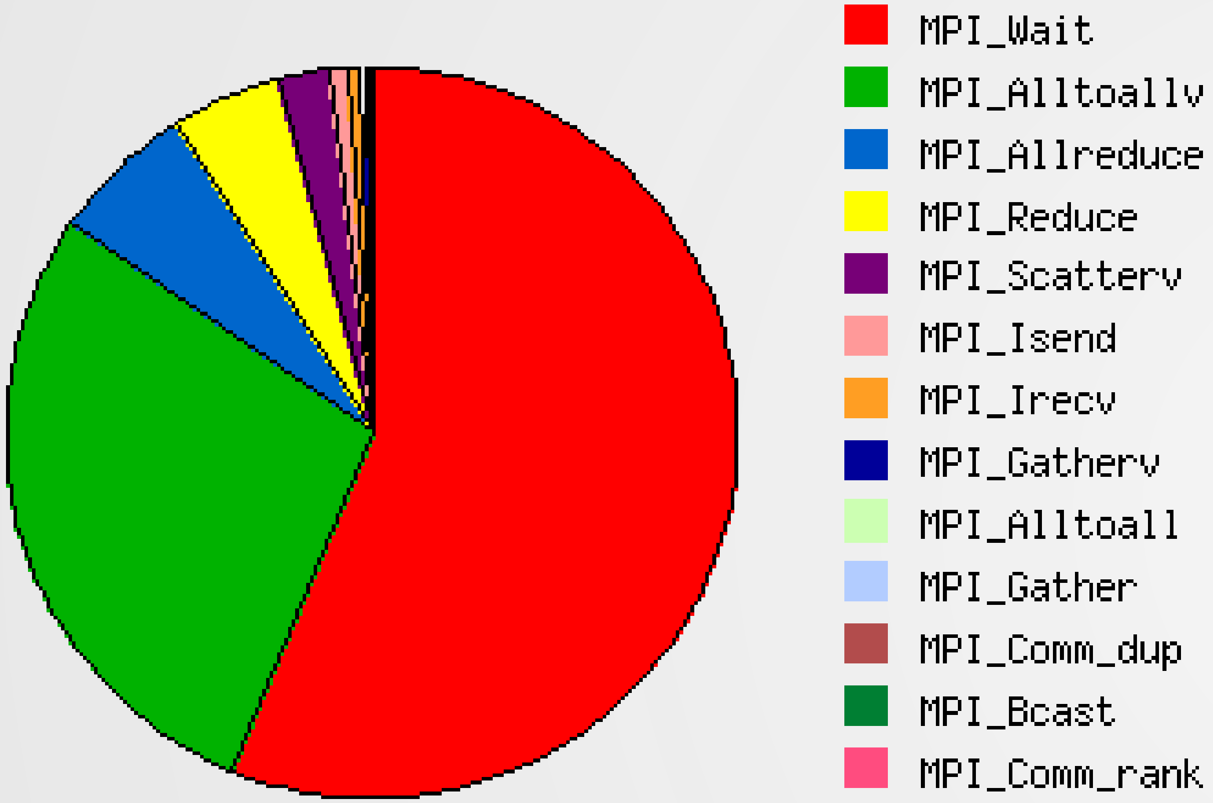
WRF with moving nested domain/2km Sandy

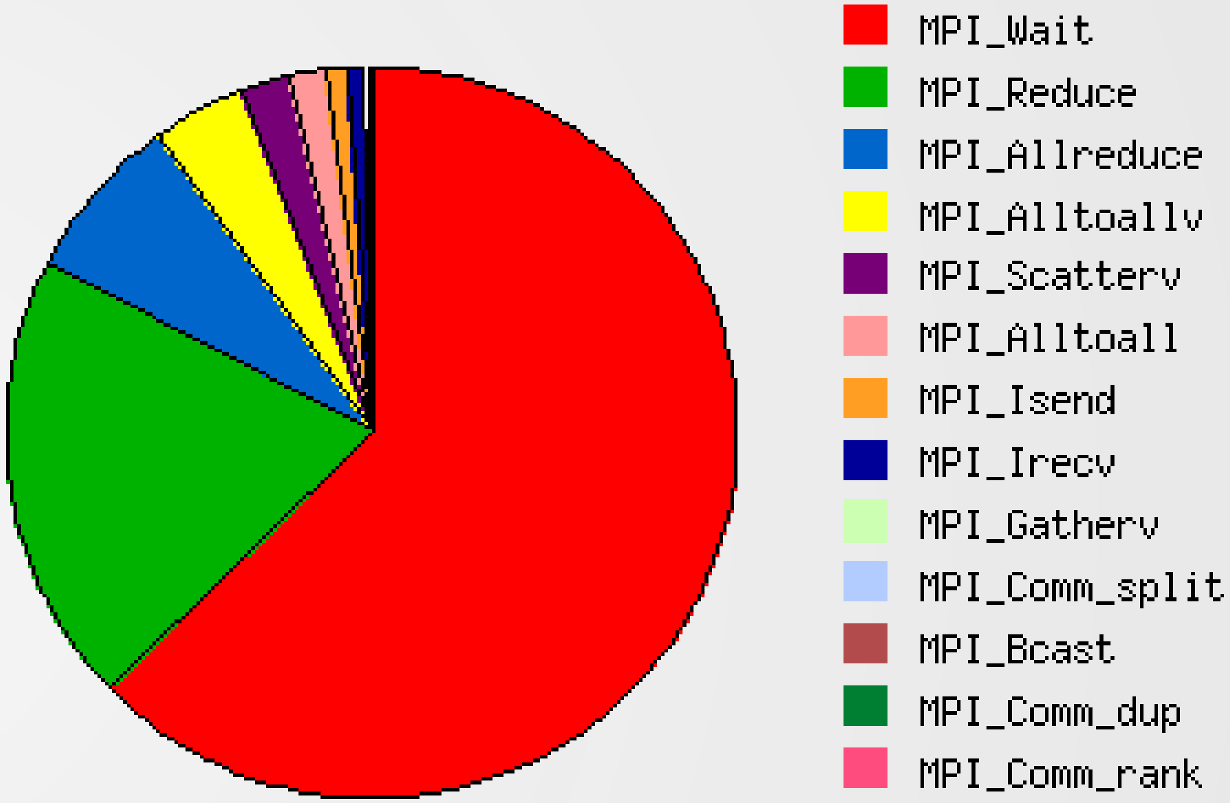


WRF 3.8.1 performance on BDW+IB EDR

(2 km Sandy w/vortex-following nested domain, 6h fcst)

WRF with moving nested domain/2km Sandy



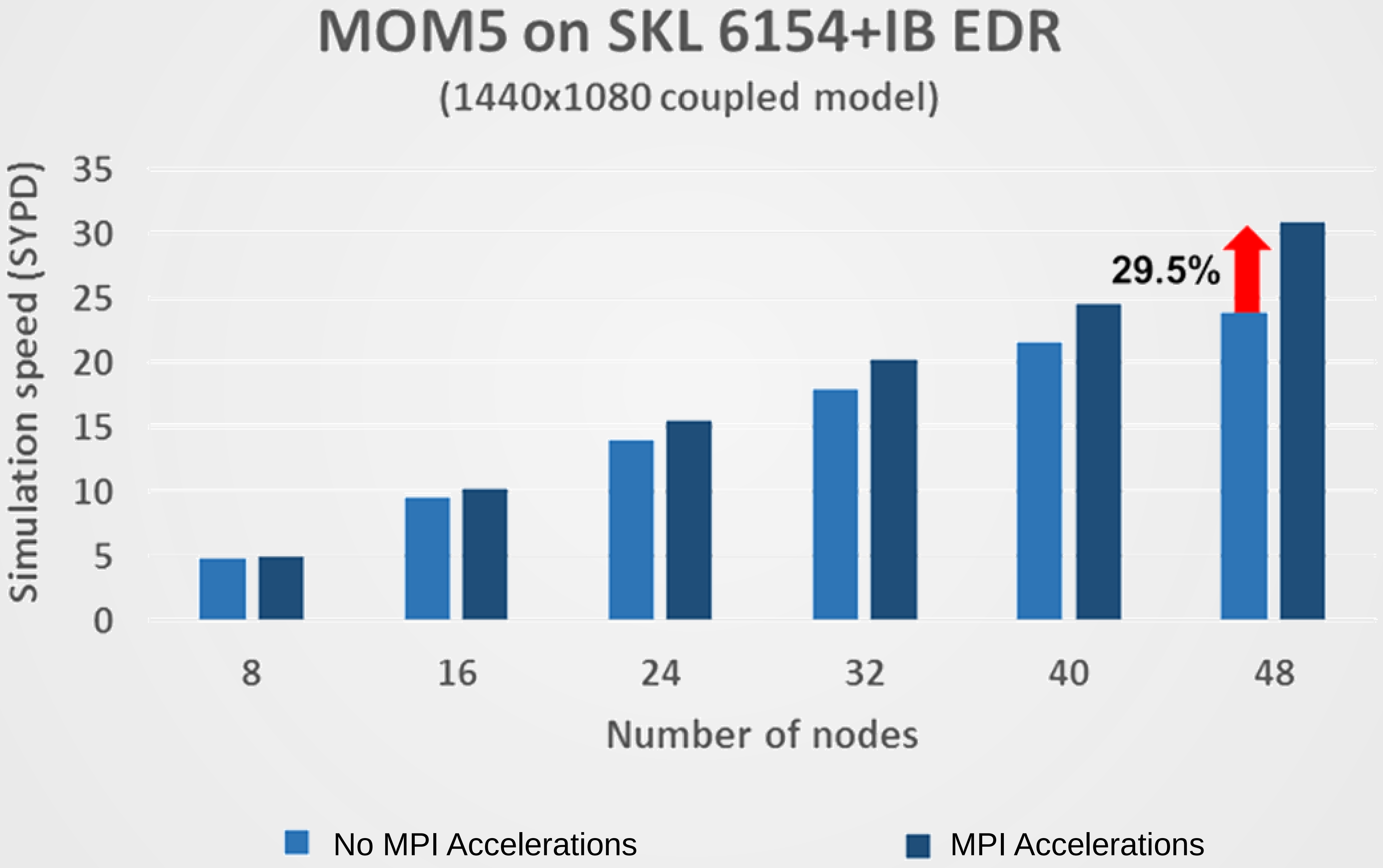


No MPI Accelerations
MPI ~27.8% of total wall time (2740s)
MPI_Alltoallv ~28% of total MPI
(32 nodes, 1152 SKL cores)

MPI Accelerations
MPI ~26.9% of total wall time (2502s)
MPI_Alltoallv ~4.2% of total MPI
(32 nodes, 1152 SKL cores)

MOM5

MOM5/SIS


MOM5 on SKL 6154+IB EDR
(1440x1080 coupled model)
Simulation speed (SYPD)
35
30
25
20
15
10
5
0
29.5%
8
16
24
32
40
48
Number of nodes
No MPI Accelerations
MPI Accelerations

Thank You