OPEN POSSIBILITIES.

Introducing Cloud-Native Supercomputing: Bare-Metal, Secured Supercomputing Architecture

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Introducing Cloud-Native Supercomputing: Bare-Metal, Secured Supercomputing Architecture

Gilad Shainer, SVP Networking, NVIDIA
Dhabaleswar Panda, Professor and University Distinguished Scholar, Ohio State University

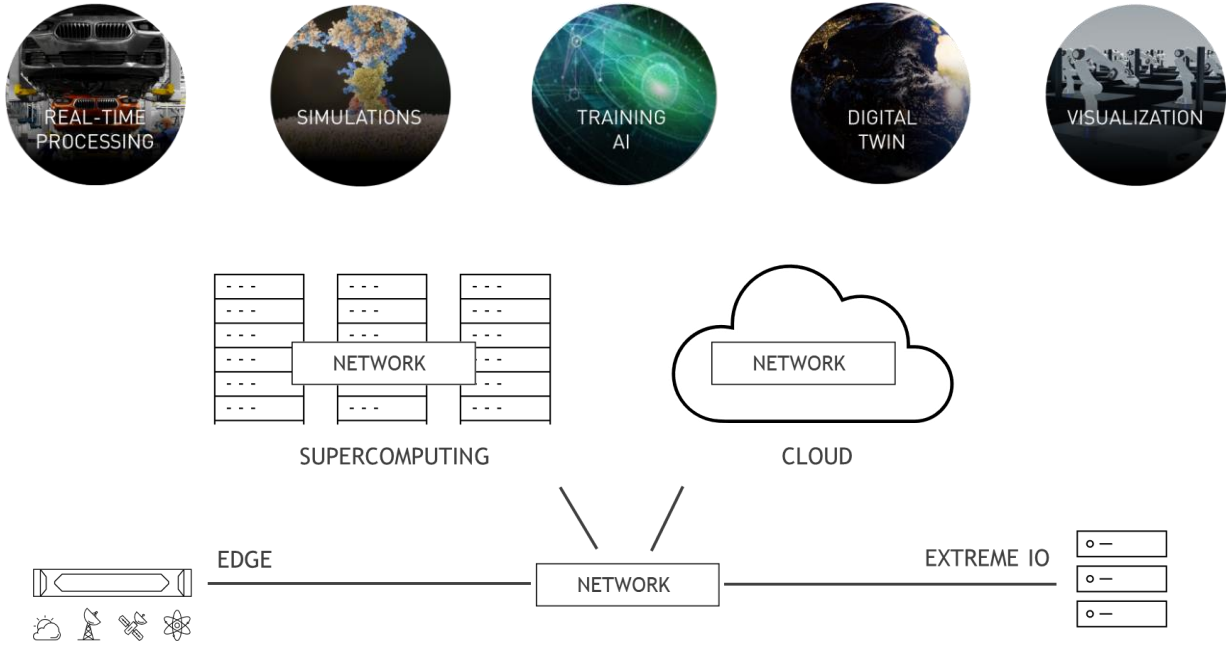
OPEN PLATINUM™

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Expanding Universe of High Performance Computing

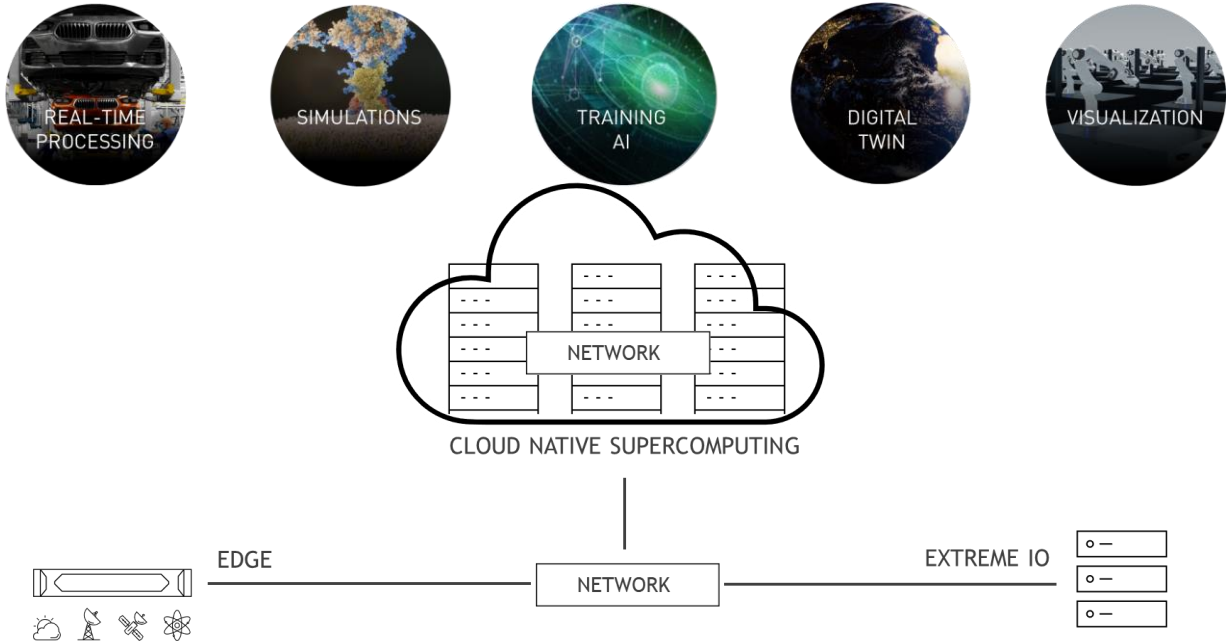
NETWORKING



OPEN POSSIBILITIES.

Expanding Universe of High Performance Computing

NETWORKING

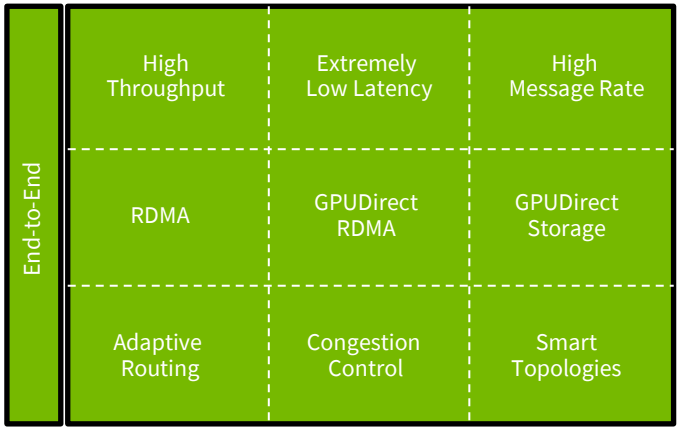


OPEN POSSIBILITIES.

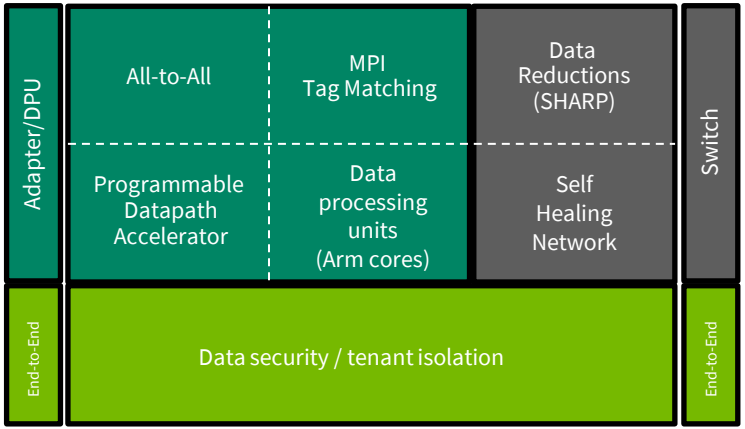
In-Network Computing Accelerated Supercomputing
Software-Defined, Hardware-Accelerated, InfiniBand Network

NETWORKING

Networking Services



In-Network Computing



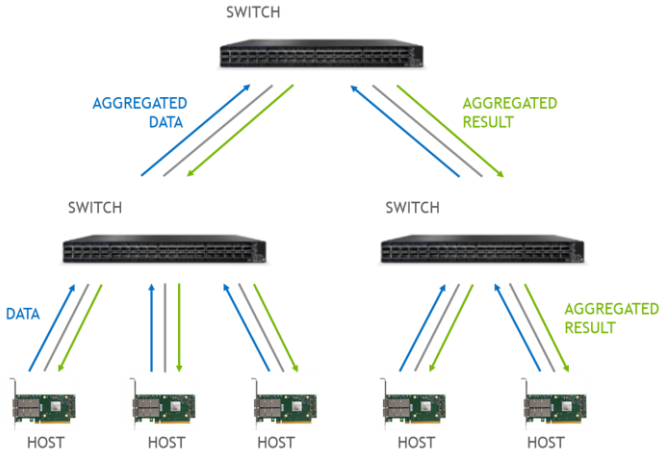
OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

NETWORKING

- In-network Tree Based Aggregation Mechanism
- Multiple Simultaneous Outstanding Operations
- Small Message and Large Message Reduction
- Barrier, Reduce, All-Reduce, Broadcast and More
- Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
- Integer and Floating-Point, 16/32/64 bits



Cloud-Native Supercomputing

NETWORKING

- Bare-metal Secured Infrastructure
- Higher Application Performance
- From the Edge to the Main Data Center

BARE-METAL PERFORMANCE

MULTI TENANCY

EDGE COMPUTING

BLUEFIELD DPU

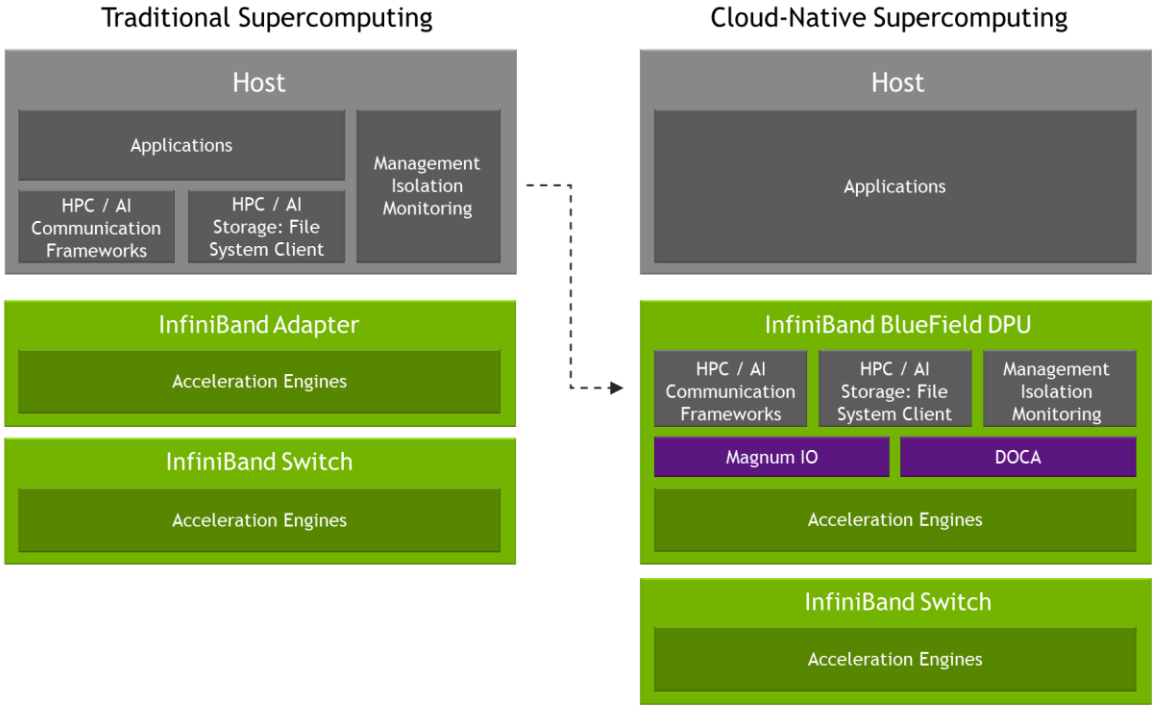
OPEN INSPIRED™

OPEN Compute Project SOLUTION PROVIDER®

OCP GLOBAL SUMMIT NOVEMBER 9-10, 2021

OPEN POSSIBILITIES.

Cloud-Native Supercomputing Infrastructure

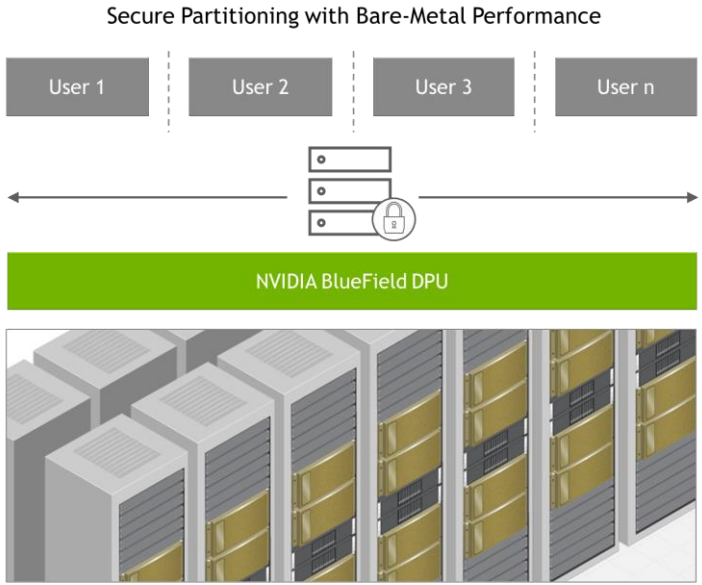


NETWORKING

Multi-Tenant Isolation - Zero-Trust Architecture

NETWORKING

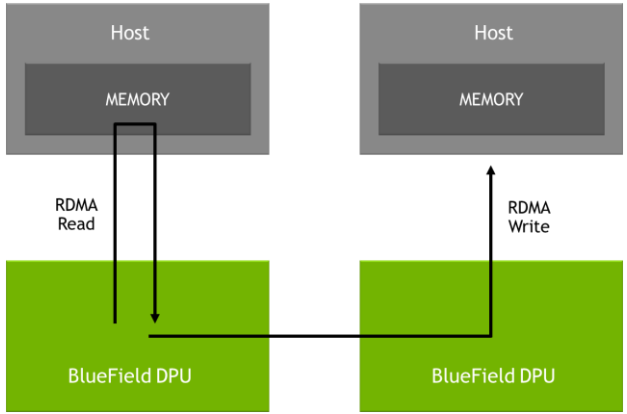
- Secured Network Infrastructure and Configuration
- Storage Virtualization
- Tenant Service Level Agreement (SLA)
- 32K Concurrent Isolated Users on Single Subnet



Higher Applications Performance

NETWORKING

- Parallel applications
- Collective Offloads
- Active Messages
- Smart MPI Progression
- Data Compression
- User-defined Algorithms



Requirements for Next-Generation MPI Libraries

NETWORKING

- Message Passing Interface (MPI) libraries are used for HPC and AI applications
- Requirements for a high-performance and scalable MPI library:
  - Low latency communication
  - High bandwidth communication
  - Minimum contention for host CPU resources to progress non-blocking collectives
  - High overlap of computation with communication
- CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation

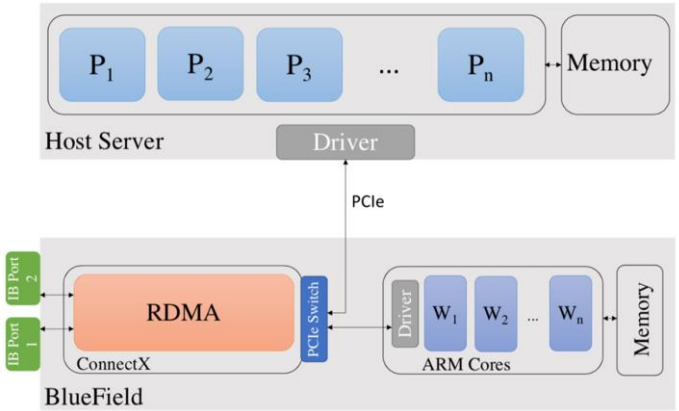
Can MPI Functions be Offloaded?

- The area of network offloading of MPI primitives is still nascent and cannot be used as a universal solution
- State-of-the-art BlueField DPUs bring more compute power into the network
- Can we exploit additional compute capabilities of modern BlueField DPUs into existing MPI middleware to extract
  - Peak pure communication performance
  - Overlap of communication and computation
- For dense non-blocking collective communications?

Overview of BlueField-2 DPU

NETWORKING

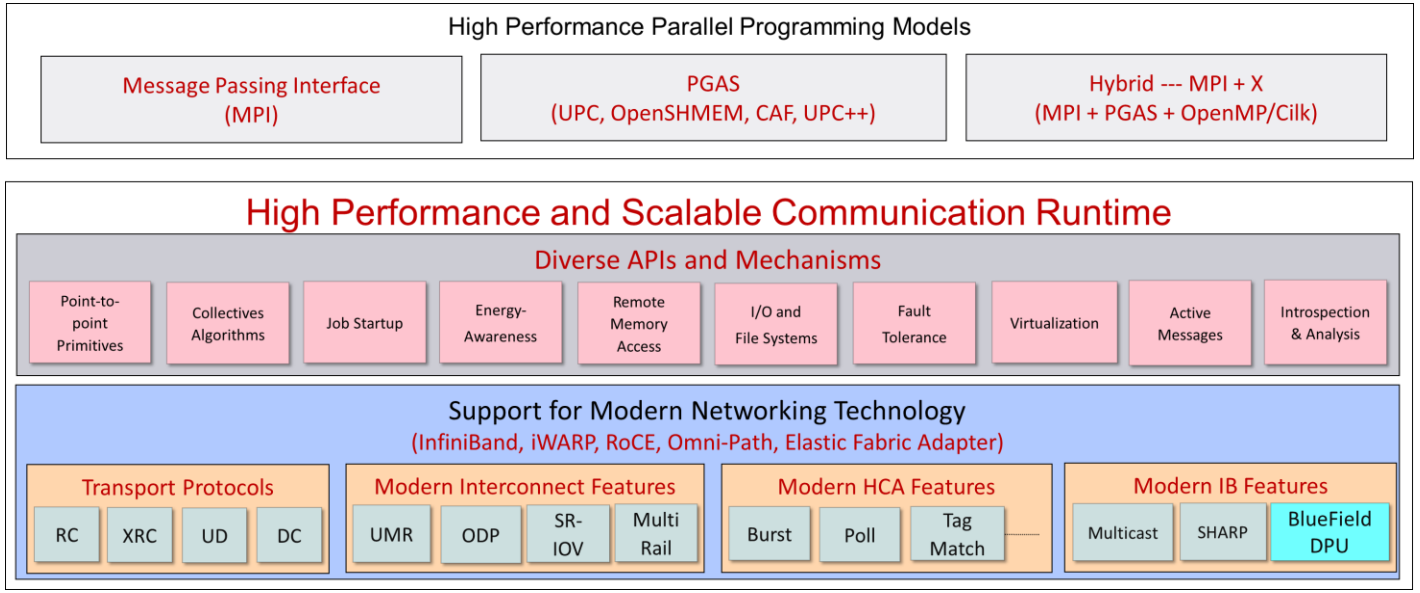
- ConnectX-6 network adapter with 200Gbps InfiniBand
- System-on-chip containing eight 64-bit ARMv8 A72 cores with 2.75 GHz each
- 16 GB of memory for the ARM cores
- How can one re-design an MPI library to take advantage of DPUs and accelerate scientific applications?



OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

MVAPICH2 Software Architecture with DPU



NETWORKING

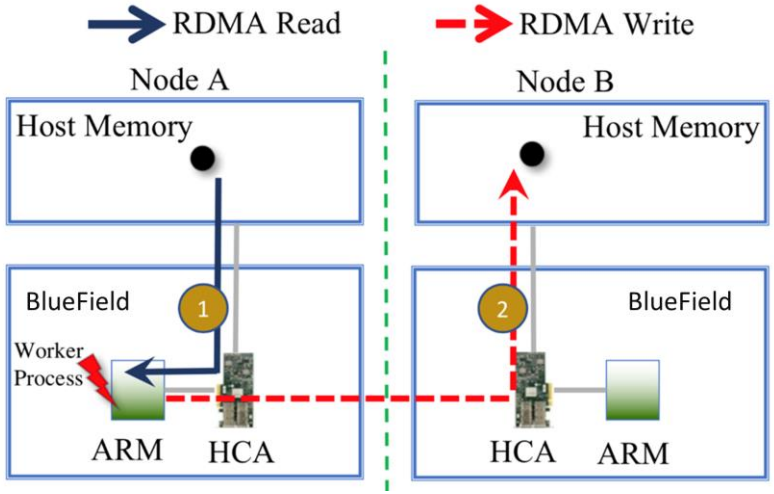
Proposed Offload Framework

- Non-blocking collective operations are offloaded to a set of “worker processes”
- BlueField is set to separated host mode
- Worker processes are spawned to the ARM cores of BlueField
- Once the application calls a collective, host processes prepare a set of metadata and provide it to the Worker processes
- Using these metadata, worker processes can access host memory through RDMA
- Worker processes progress the collective on behalf of the host processes
- Once message exchanges are completed, worker processes notify the host processes about the completion of the non-blocking operation

Proposed Non-blocking Alltoall Design

NETWORKING

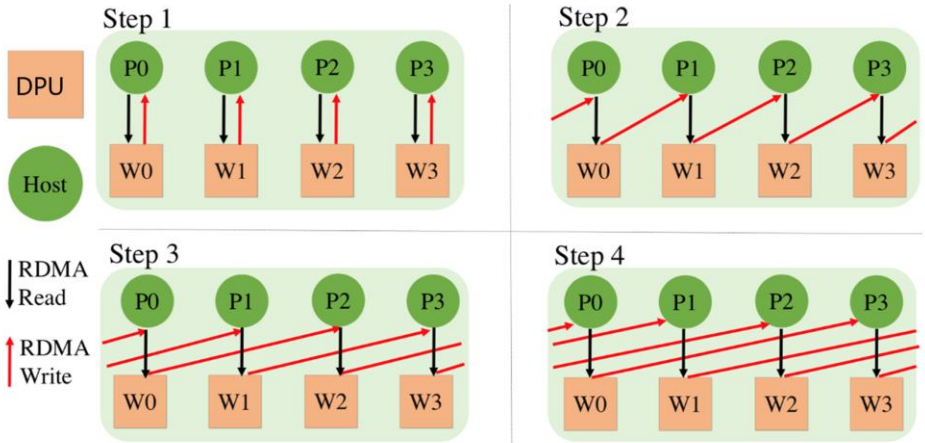
- Worker process performs RDMA Read to receive the data chunk from host main memory
- Once data is available in the ARM memory, worker process performs RDMA Write to the remote host memory



Proposed Non-blocking Alltoall Design

NETWORKING

- Example: Scatter Destination Algorithm
- Focus is on medium and large messages
- Message chunking and pipelining is utilized to reduce the overheads of staging



Experimental Setup for Performance Evaluation

- HPC Advisory Council High-Performance Computing Center
- Cluster has 32 compute-node with Broadwell series of Xeon dual-socket, 16-core processors operating at 2.60 GHz with 128 GB RAM
- NVIDIA BlueField-2 adapters are equipped with 8 ARM cores operating at 2.0 GHz with 16 GB RAM
- Based on the MVAPICH2-DPU MPI library
- OSU Micro Benchmark for nonblocking Alltoall and P3DFFT Application

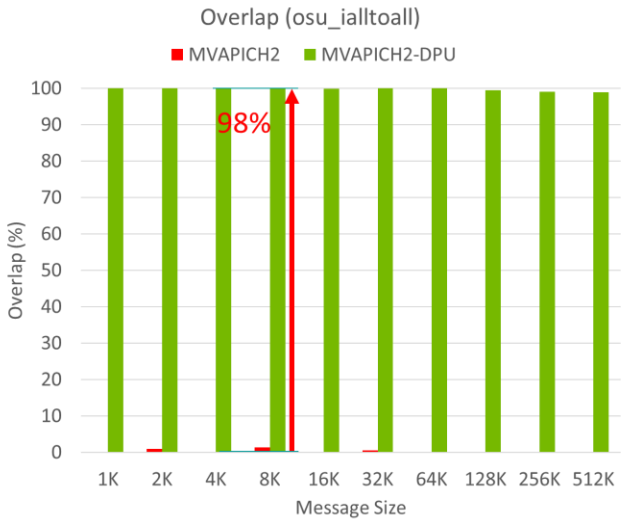
OSU Micro benchmark ialltoall

NETWORKING

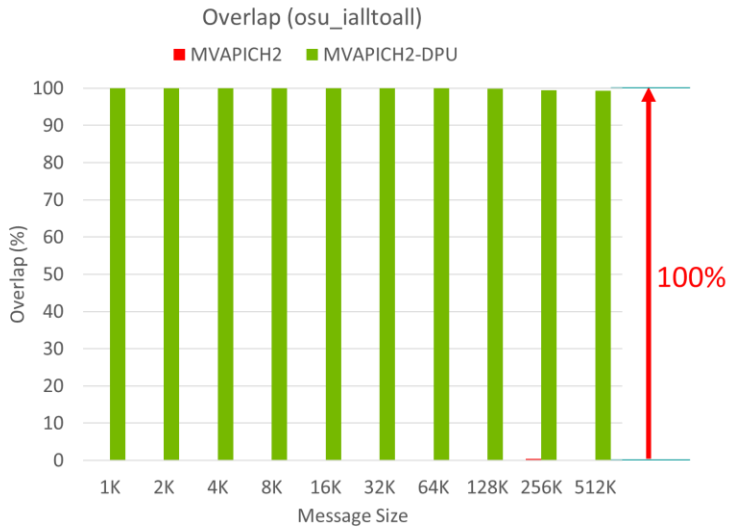
- osu_ialltoall benchmark metrics
- Pure communication time
  - Latency t is measured by calling MPI_Ialltoall followed by MPI_Wait
- Total execution time
  - Total T = MPI_Ialltoall + synthetic compute + MPI_Wait
- Overlap
  - Benchmark creates a synthetic computation block that takes t microsecond to finish. Before starting compute, MPI_Ialltoall is called and after that MPI_Wait. Overlap is calculated based on total execution time and compute time.
- Part of the standard OSU Micro-Benchmark

Overlap of Communication and Computation with osu_Ialltoall (32 nodes)

NETWORKING



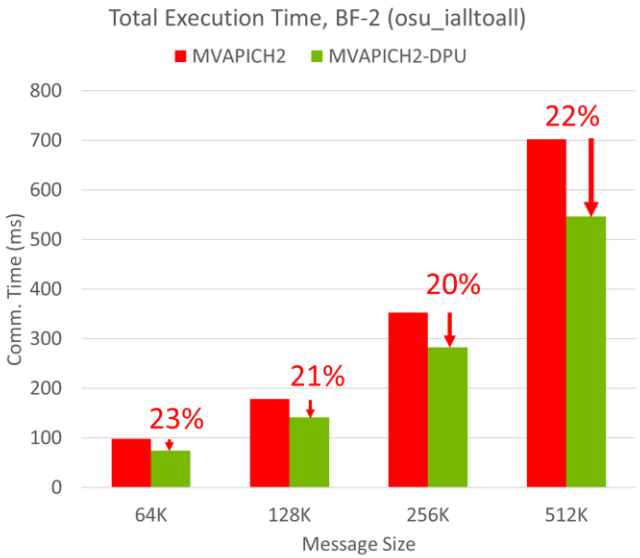
32 Nodes, 16 PPN



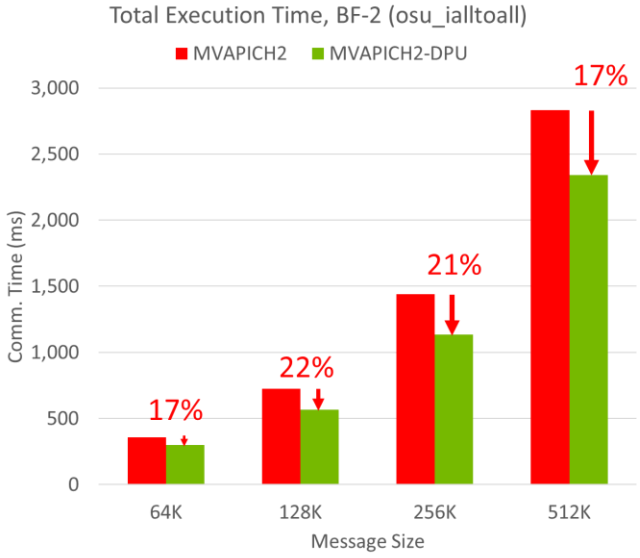
32 Nodes, 32 PPN

Delivers peak overlap

Total Execution Time with osu_Ialltoall (32 nodes)



32 Nodes, 16 PPN



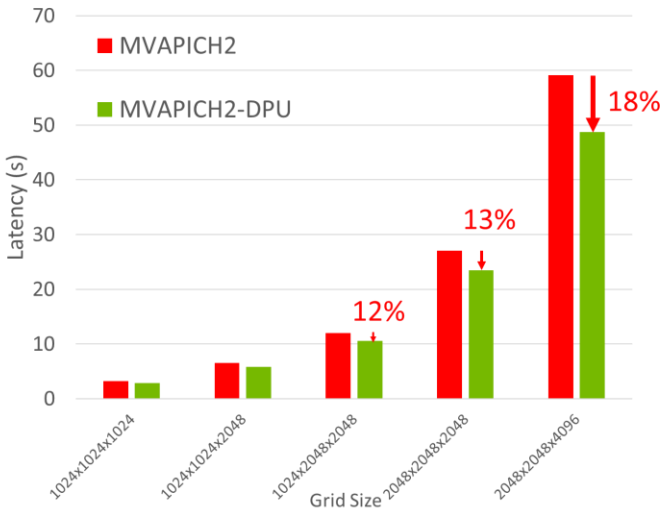
32 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

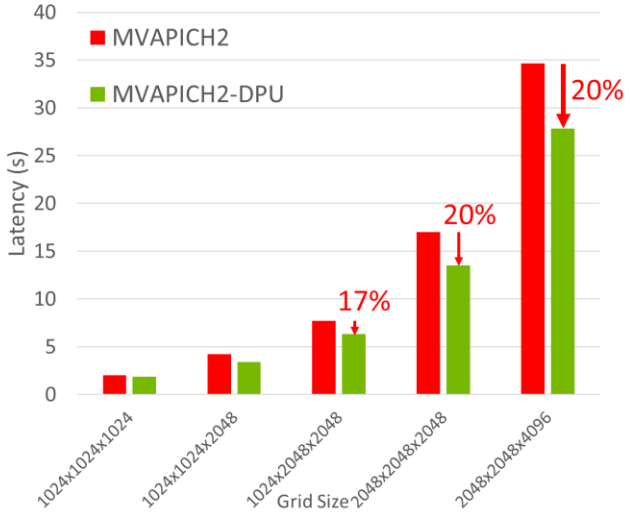
NETWORKING

P3DFFT Application Execution Time (16 nodes)

NETWORKING



16 Nodes, 16 PPN



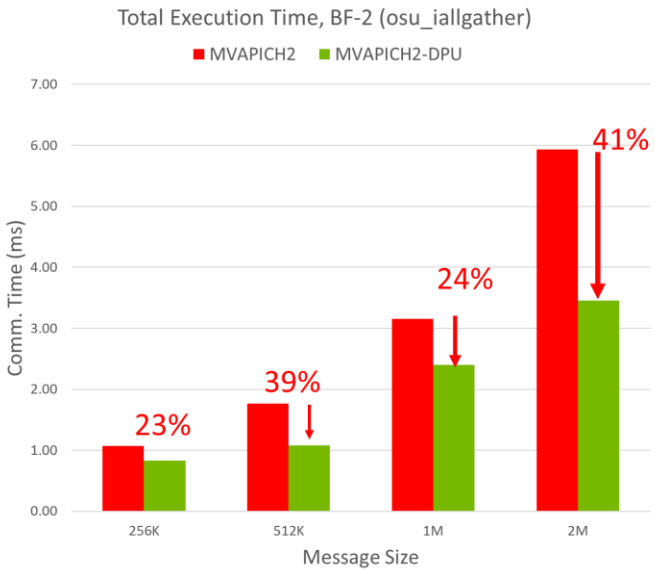
16 Nodes, 32 PPN

Benefits in application-level execution time

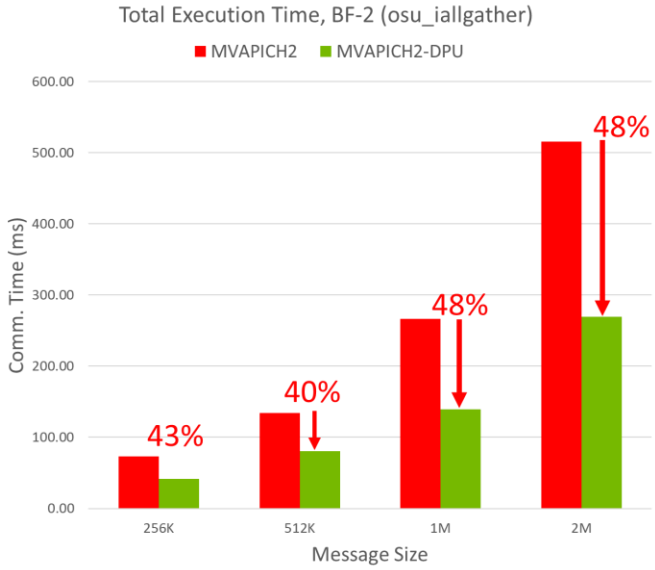
OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Total Execution Time with osu_Iallgather (16 nodes)



16 Nodes, 1 PPN



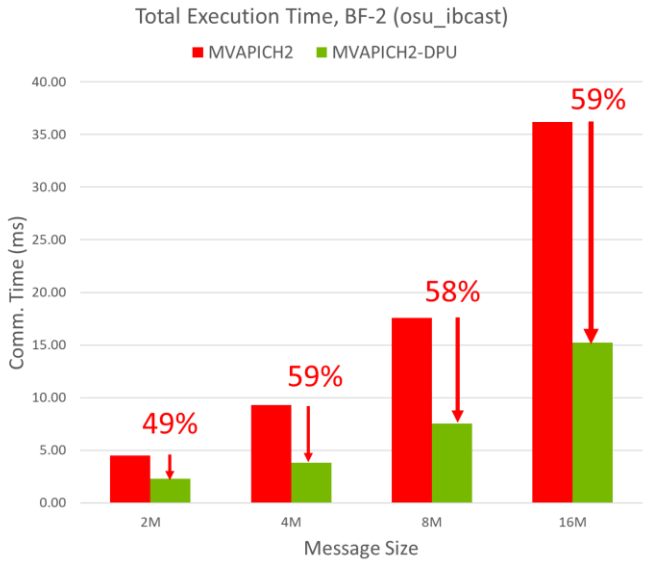
16 Nodes, 16 PPN

NETWORKING

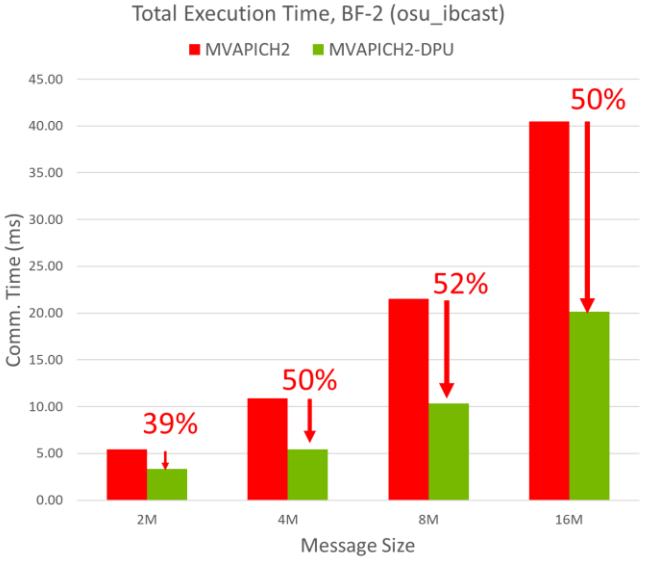
OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Total Execution Time with osu_Ibcast (16 nodes)



16 Nodes, 16 PPN



16 Nodes, 32 PPN

NETWORKING

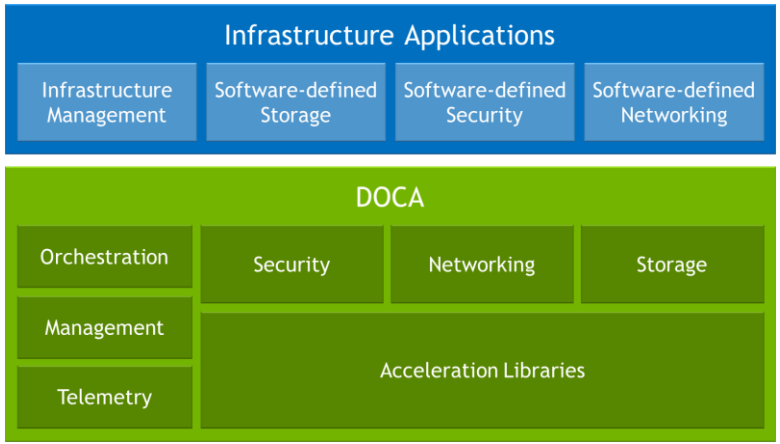
OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

NVIDIA DOCA SDK

NETWORKING

- Software Application Framework for BlueField DPUs
- DOCA is for DPUs What CUDA is for GPUs
- Protects Developer Investment for Future DPUs
- Certified Reference Applications, APIs & Partner Solutions
- Rich Partner Ecosystem Across Industries and Workloads
- https://developer.nvidia.com/networking/doca



OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Call to Action – More Information

- Technical overview - https://nvdam.widen.net/s/plhzlwmtrg/tech-overview-infiniband-cloud-native-supercomputing-web
- Community open source software development - https://ucfconsortium.org/projects/opensnapi/
- The MVAPICH2-DPU MPI library - https://x-scalesolutions.com/mvapich2-dpu/
- News announcements and more links - https://www.hpcwire.com/2021/04/14/gtc21-dell-building-cloud-native-supercomputers-at-u-cambridge-and-durham/

Thank you!