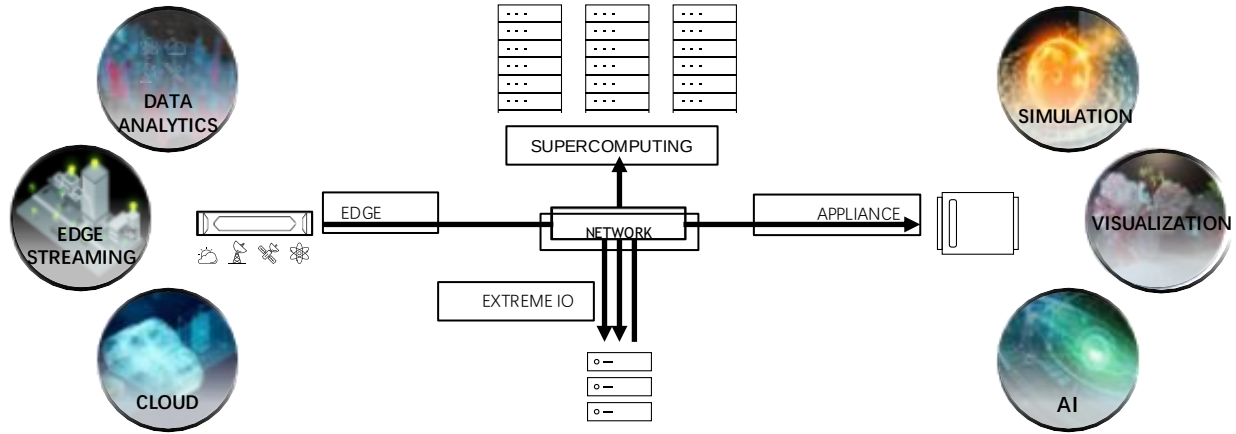
AMD | Lenovo NetApp | NVIDIA

联想助力互联网智能转型

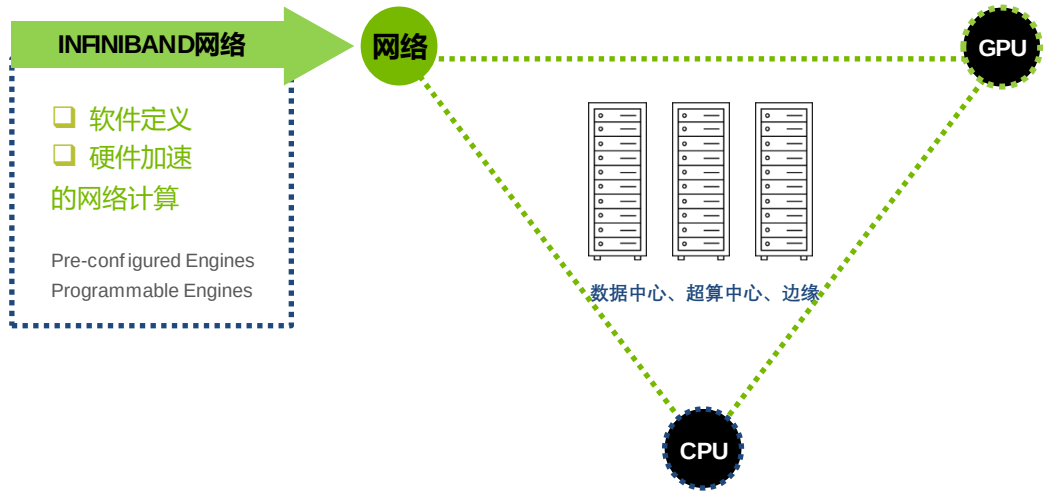
网络计算及DPU在数据中心和边缘云上的应用

NVIDIA, Jan 8, 2021

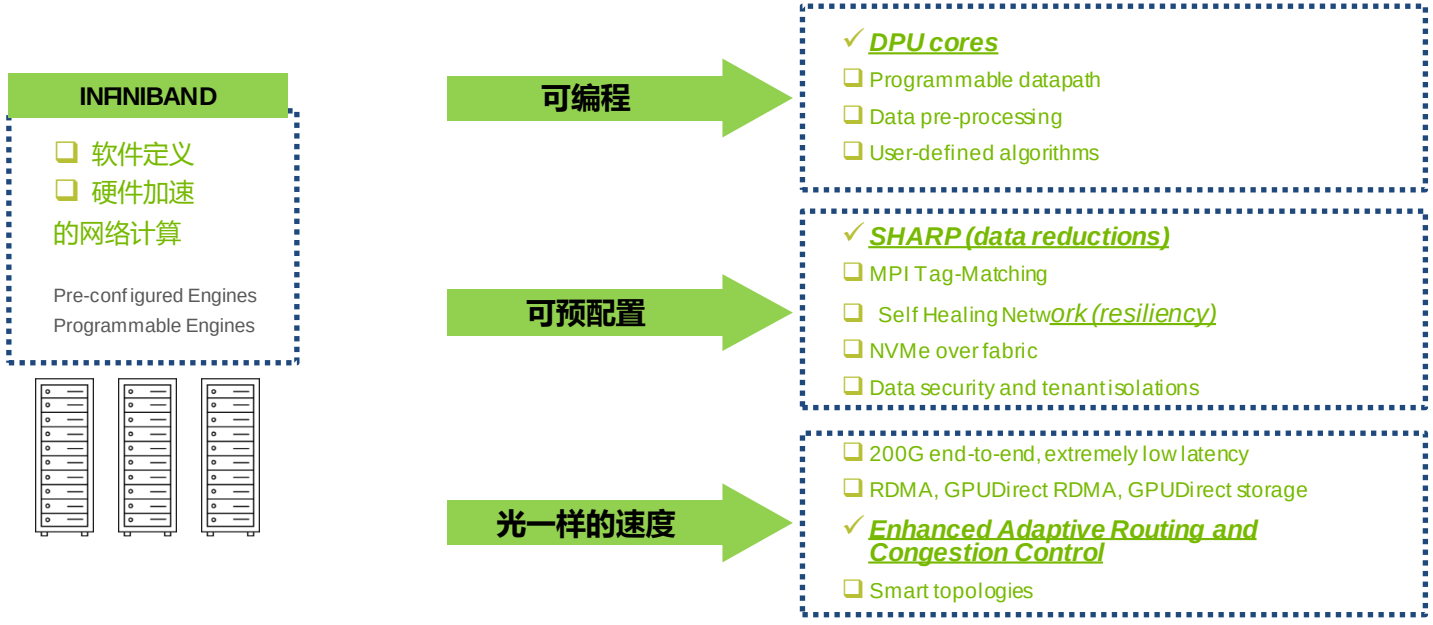
计算无处不在，数据成为中心



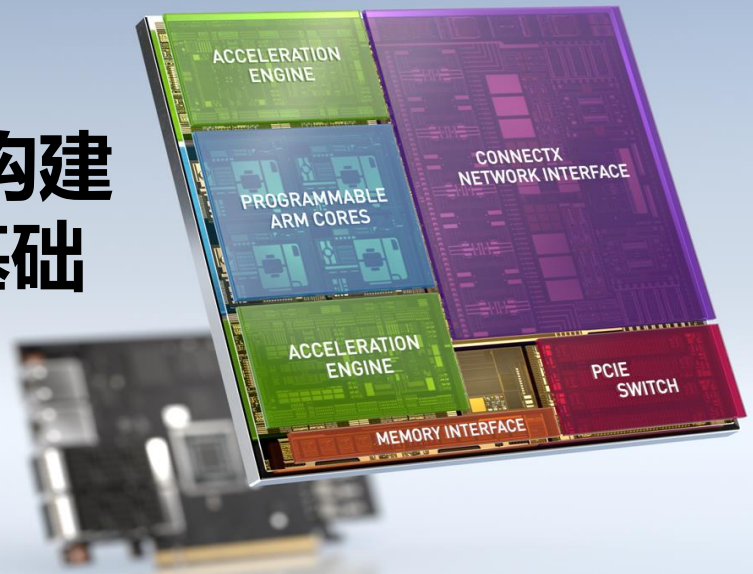
网络计算成为数据中心的三大计算支柱之一



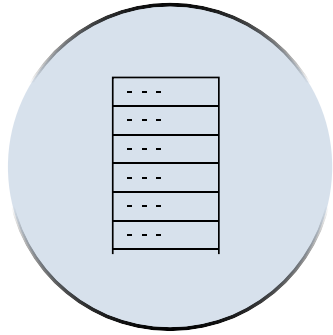
网络计算加速互联网数据中心和HPC云



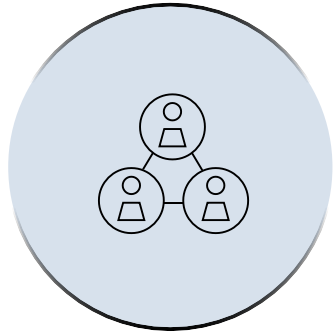
InfiniBand DPU - 构建云原生数据中心的基础



高性能云原生数据中心的需求



BARE-METAL 的性能

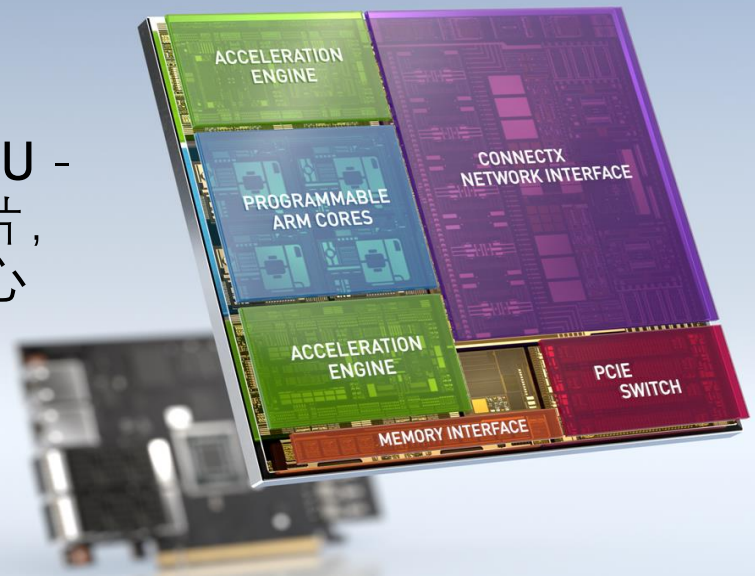


支持多租户

安全可靠

服务类型可配置

BLUEFIELD INFINIBAND DPU -
集数据中心基础架构于芯片，面向高性能云原生数据中心



NVIDIA | Lenovo NetApp

INFRASTRUCTURE 应用

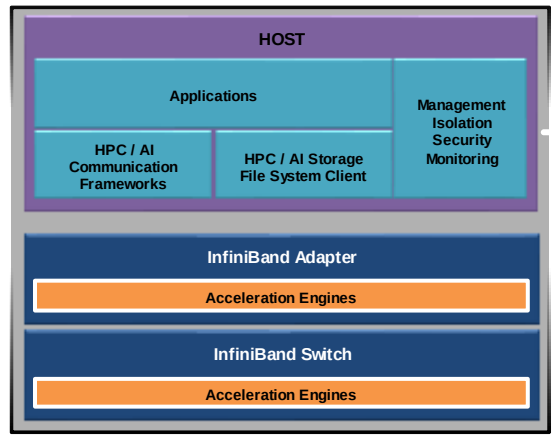
Infrastructure Management	Software-defined Storage	Software-defined Security	Software-defined Networking

DOCA SDK开发包

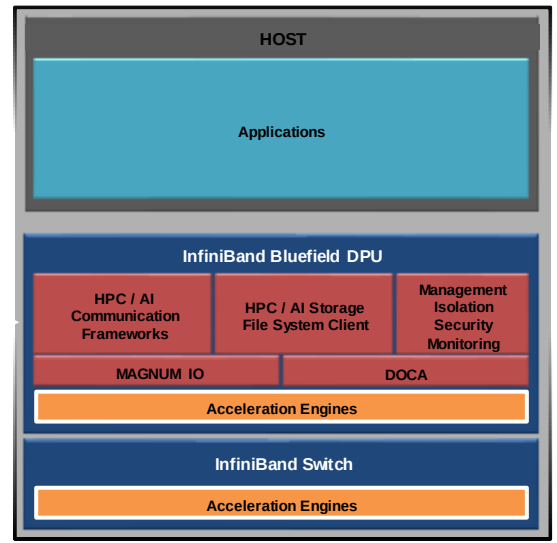
Telemetry	Management	Storage SPDK	Security DPDK	Networking DPDK / P4
		SNAP	ASAP²	CRYPTO	RoT	RDMA

Bluefield DPU -面向高性能云原生数据中心

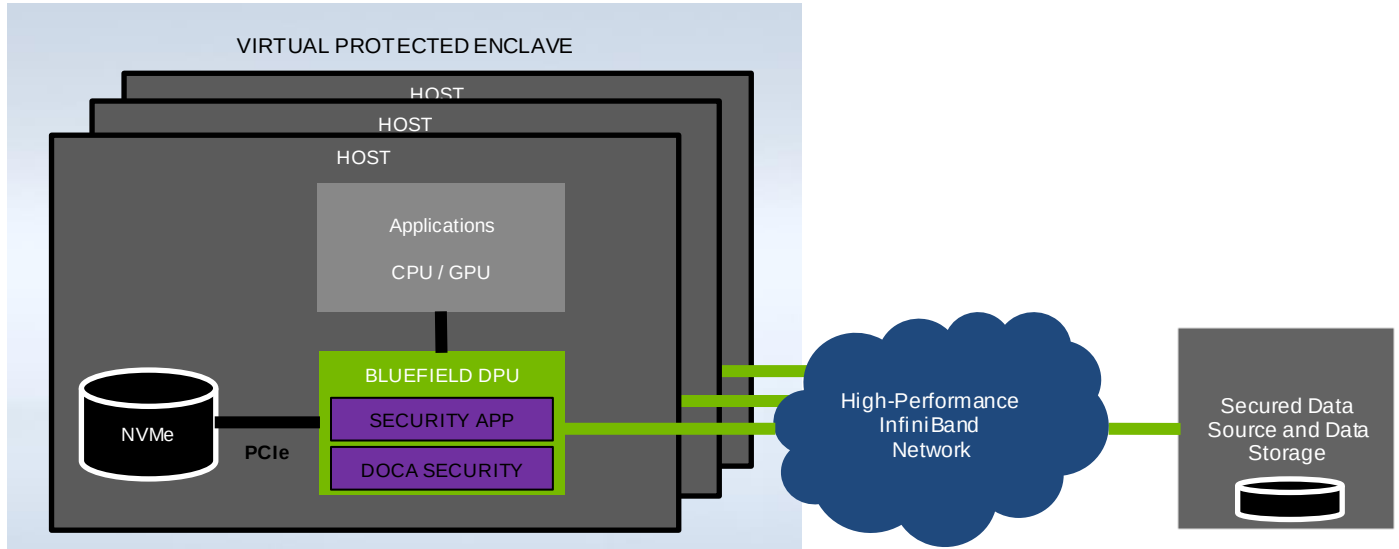
传统的超级计算中心/高性能数据中心



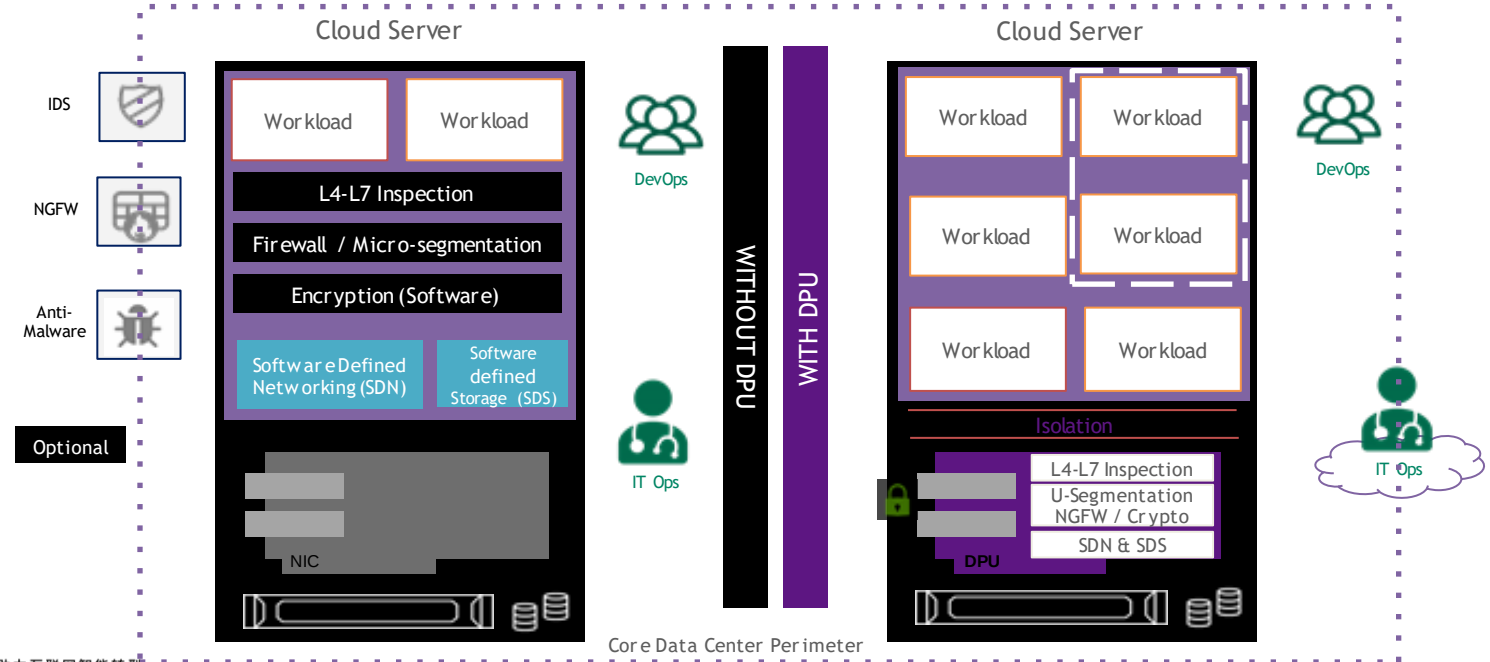
CLOUD NATIVE 云原生高性能数据中心



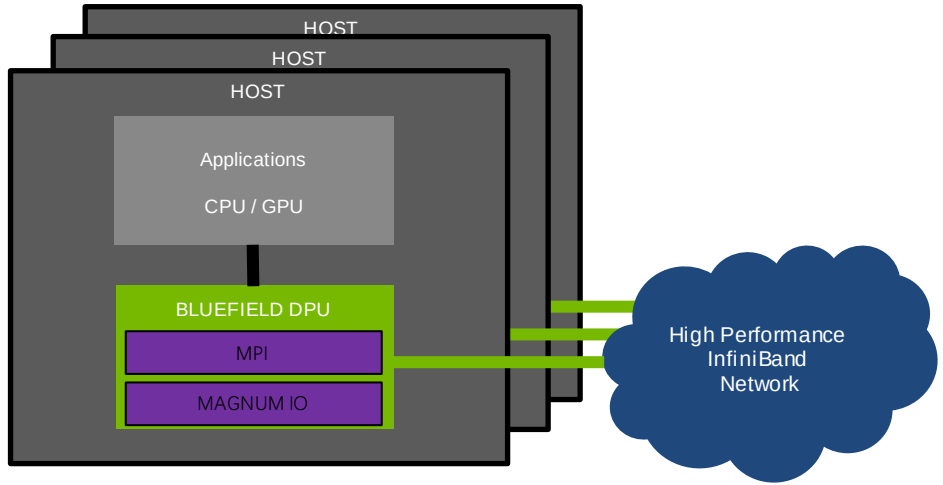
Bluefield DPU – 兼顾用户数据安全和计算性能



云的安全防御由外围转向服务器内部

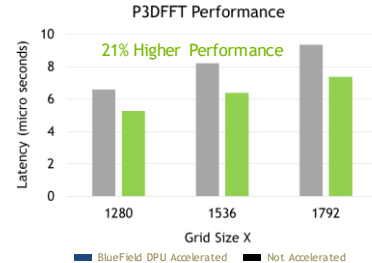


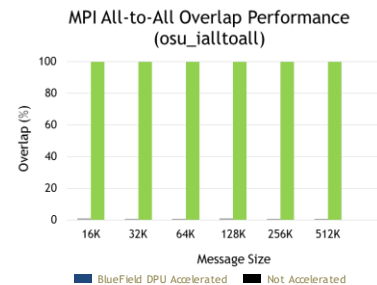
Bluefield DPU - HPC 和 AI 通信卸载



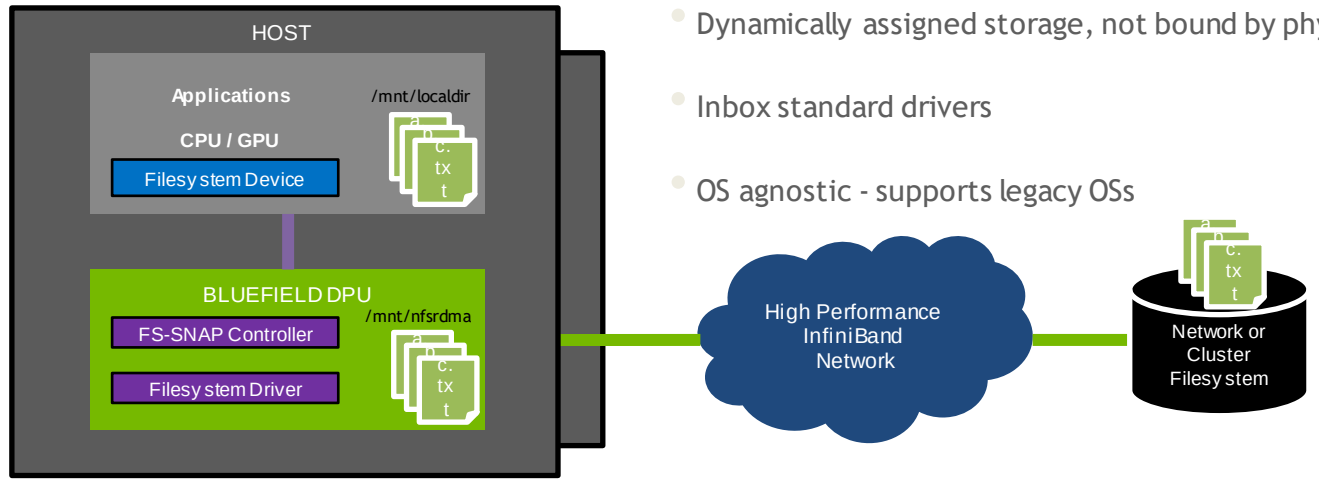
Eight servers, Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz (32 processes per node), NVIDIA BlueField-2 HDR100 DPUs and ConnectX-6 HDR100 adapters, NVIDIA Mellanox HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand, 256GB DDR4 2400MHz RDIMMs memory and 1TB 7.2K RPM SATA 2.5" hard drive per node.

Courtesy of Ohio State University MVAPICH team and X-ScaleSolutions





Bluefield DPU - HPC & AI 高性能存储池化及卸载



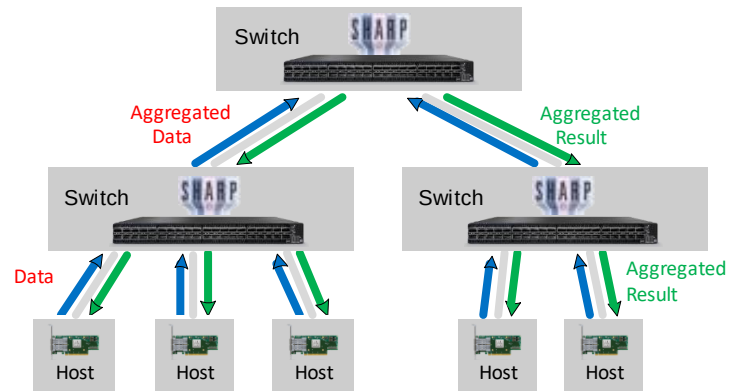
- Emulates remote storage to appear as local to the host OS
- Dynamically assigned storage, not bound by physical capacity
- Inbox standard drivers
- OS agnostic - supports legacy OSs

InfiniBand 网络
会计算的SDN网络

交换机计算的核心 - SHARP技术

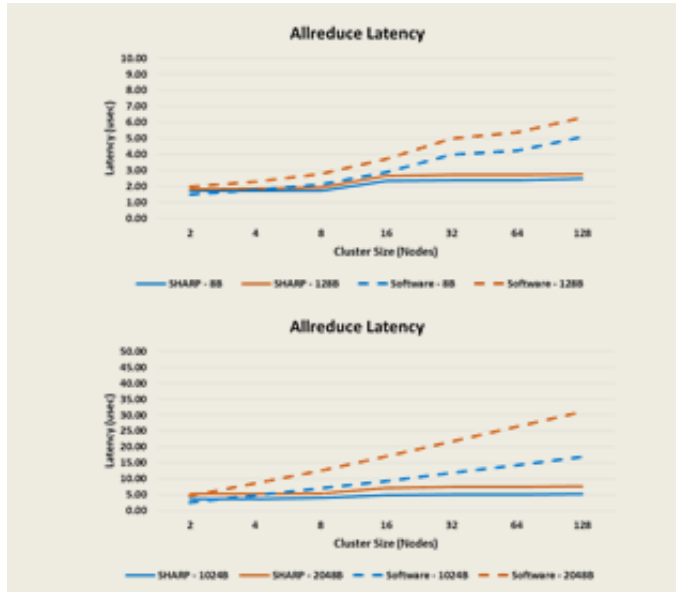
Scalable Hierarchical Aggregation and Reduction Protocol

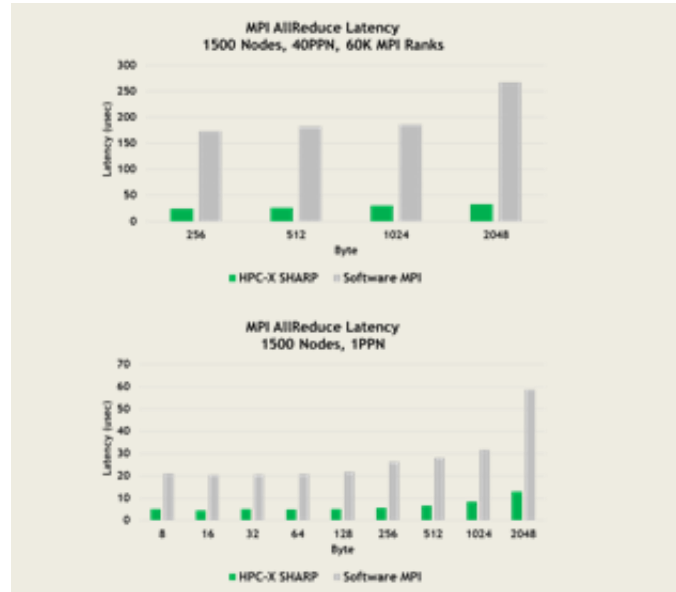
- 支持多个操作并发进行
- 支持应用: HPC (MPI / SHMEM) 和分布式机器学习等
- 支持操作: Barrier, Reduce, All-Reduce, Broadcast and more
- 支持计算: Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
- 支持数据: 整型和16/32/64 bits 浮点数据



SHARP AllReduce 性能提升

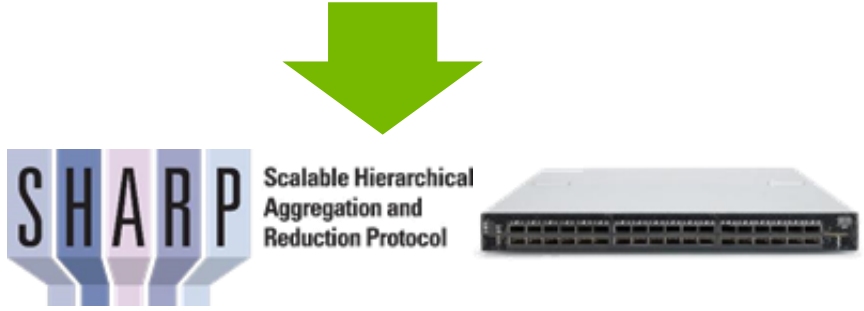
- 提供稳定的低延时，7倍的性能提升





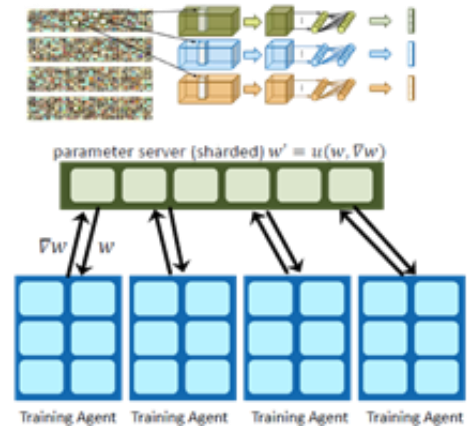
交换机取代了AI训练的参数服务器

传统方案：参数服务器上的 CPU 成为训练的瓶颈



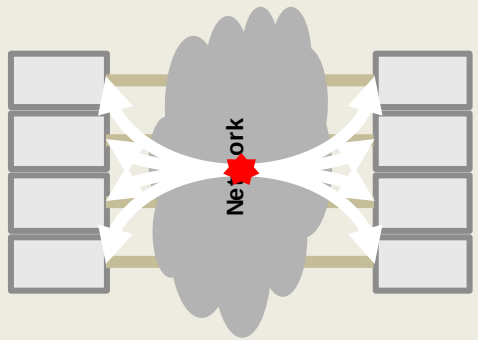
Performs the Gradient Averaging
Replaces all physical parameter servers
Accelerate AI Performance

优化方案：交换机成为参数服务器



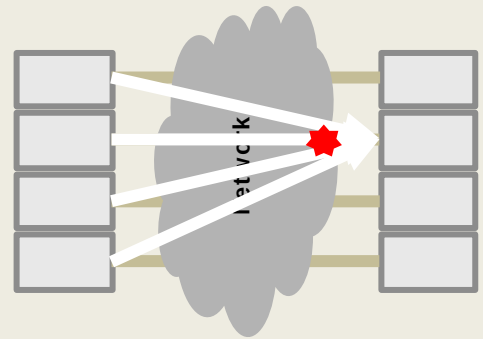
网络拥塞成为数据中心的最大挑战之一

网络中的In-Network拥塞



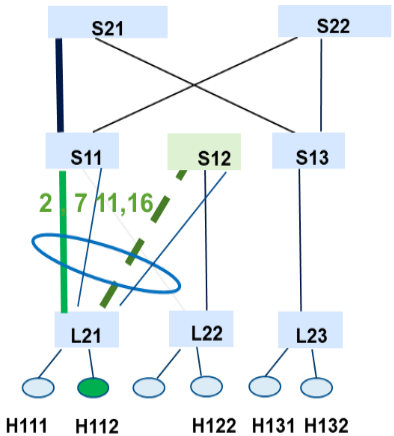
解决方案: 动态路由

交换机内的In-cast拥塞

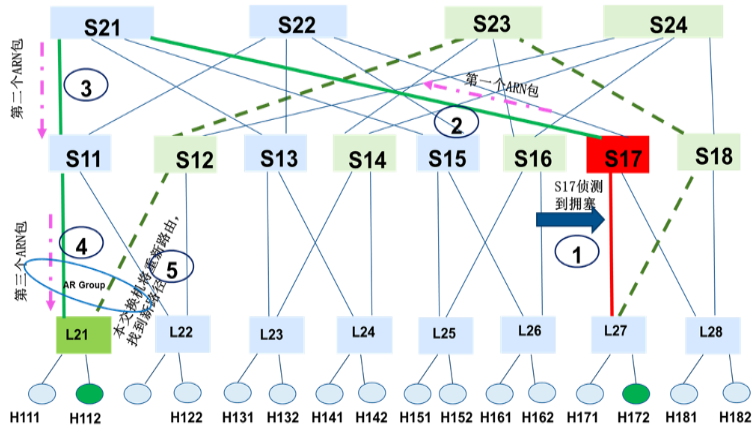


解决方案: 网络拥塞控制

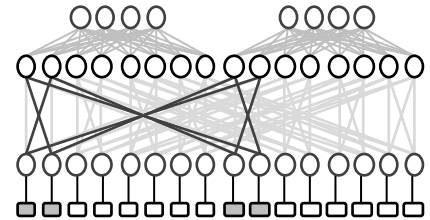
动态路由解决了网络In-Network拥塞问题



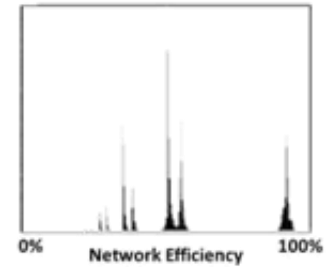
交换机上行端口拥塞



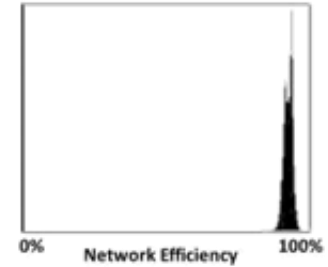
交换机下行端口拥塞



MPIGraph: 静态和动态路由对照



静态路由

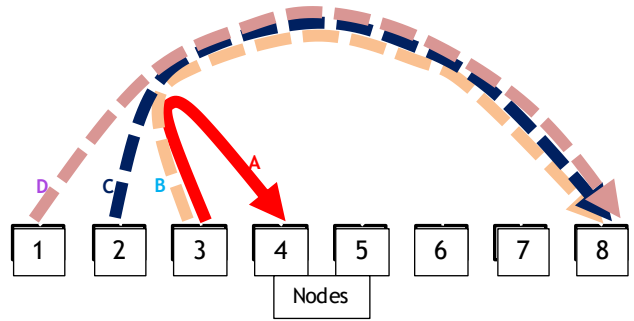


动态路由

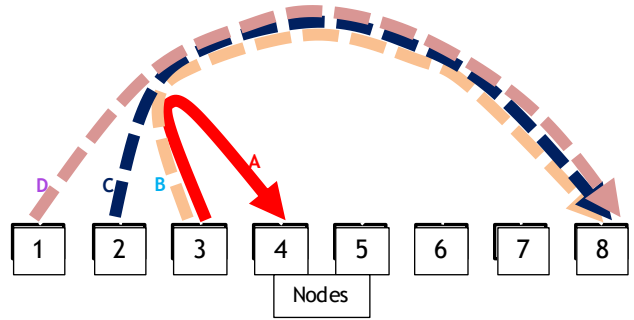
The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems

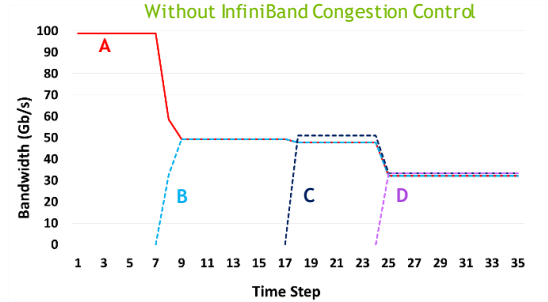
Sudharshan S. Vazhkudai[1], Bronis R. de Supinski[2], Arthur S. Bland[1], Al Geist[1], James Sexton[3], Jim Kahle[3], Christopher J. Zimmer[1], Scott Atchley[1], Sarp Oral[1], Don E. Maxwell[1], Veronica G. Vergara Larrea[1], Adam Bertsch[2], Robin Goldstone[2], Wayne Joubert[1], Chris Chambreau[2], David Appelhans[3], Robert Blackmore[3], Ben Casses[3], George Chochia[3], Gene Davison[3], Matthew A. Ehlerl[1], Tom Gooding[3], Rka Gorenorowski[3], Leopold Grinberg[3], Bill Hanson[3], Bill Hartner[3], Ian Karlin[2], Matthew L. Leininger[2], Dustin Leverman[2], Chris Marroquin[3], Adam Moody[2], Martin Ohmacht[3], Ramesh Pankajakshan[2], Fernando Pizzano[3], James H. Rogers[1], Bryan Rosenburg[3], Drew Schmidt[2], Mallikarjun Shankar[1], Feiyi Wang[1], Py Watson[3], Bob Walkup[3], Lance D. Weems[2], Jesup Yin[2]
[1] Oak Ridge National Laboratory, [2] Lawrence Livermore National Laboratory, [3] IBM
[vazhkudai@ornl.gov, bronis@llnl.gov]

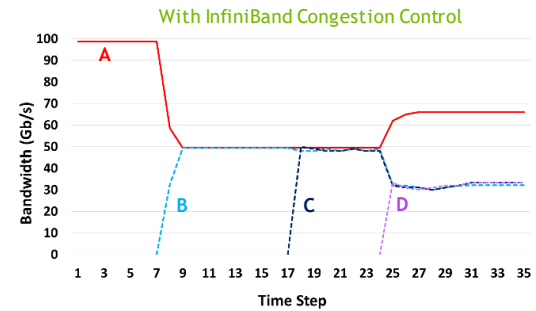
HDR InfiniBand 拥塞控制解决了IN-CAST 拥塞问题



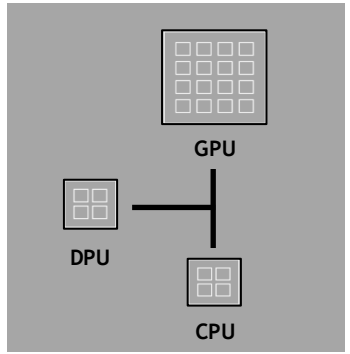
HDR InfiniBand 拥塞控制解决了IN-CAST 拥塞问题





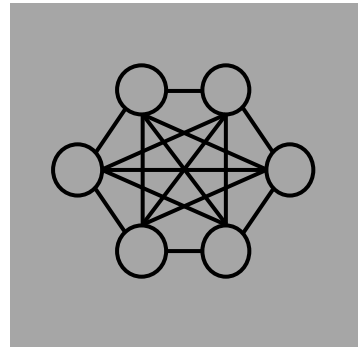


InfiniBand网络总结 – 会计算的SDN网络，面向高性能云原生数据中心



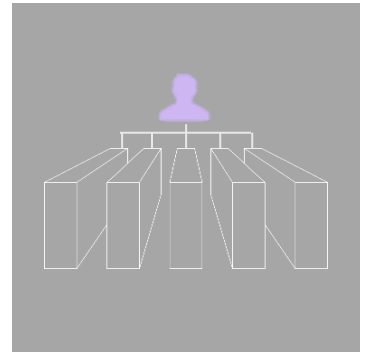
服务器三大核心计算单元之一

面向以数据为中心的计算



会计算的交换机，无限可扩展

面向E级及更大规模数据中心



集中管理，安全高效，天然SDN

面向云原生数据中心

标准、开放、向前向后兼容

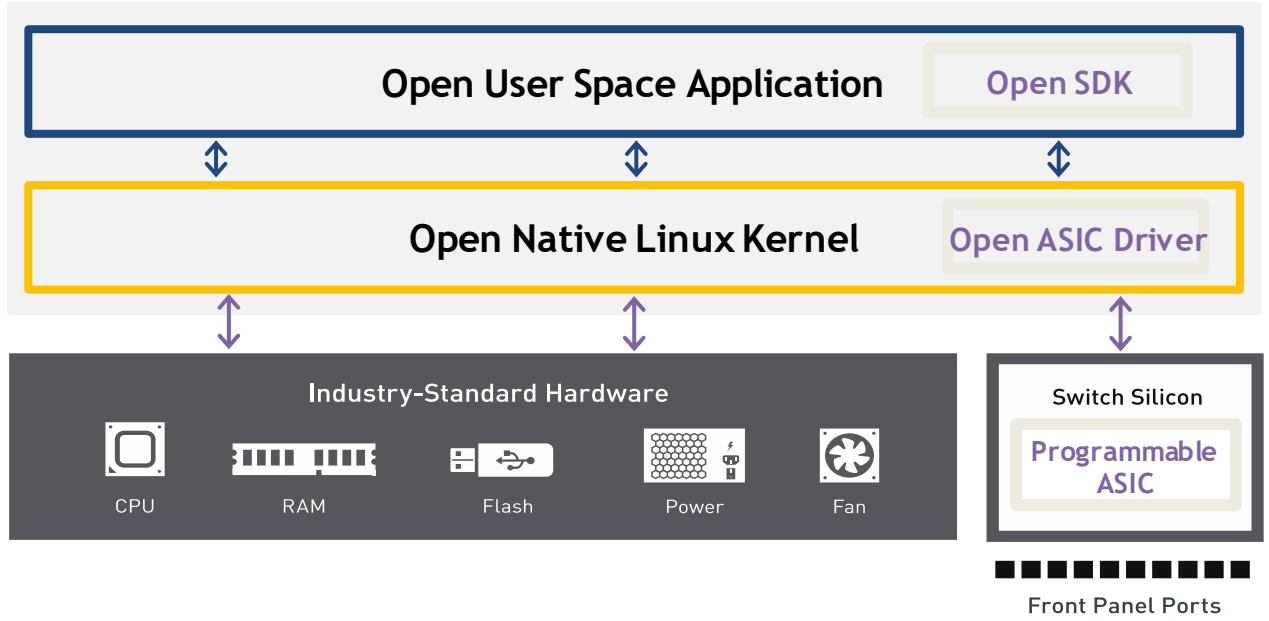
面向未来，兼容过去

开放以太网
解锁软硬件、供应商

NVIDIA | Lenovo NetApp

数据中心以太网的演进

NVIDIA 推动以太网走向开放



NVIDIA 开放以太网让用户灵活选择网络操作系统

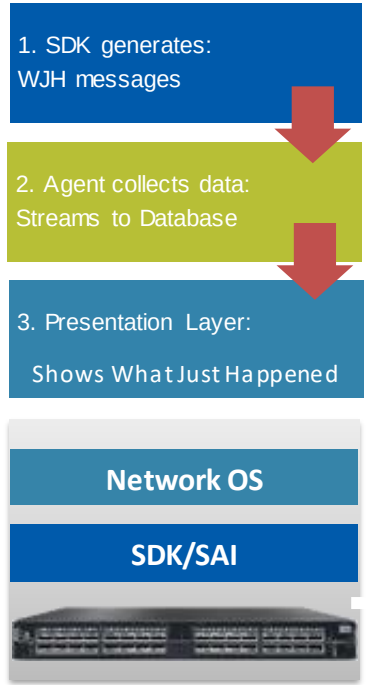
- 适合小规模部署
- 一键RoCE部署

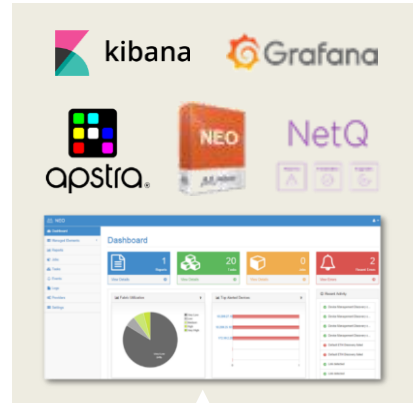
- 优异的 VXLAN 性能
- 基于 Linux 网络操作系统
- 用Linux的方法管理主机和网络

- 开源以太网操作系统
- 不必受限于网络厂商
- 免费

开放以太网的健康保障 – WJH （What Just Happened?）

Telemetry
- 轻量
- 可部署
- 事件驱





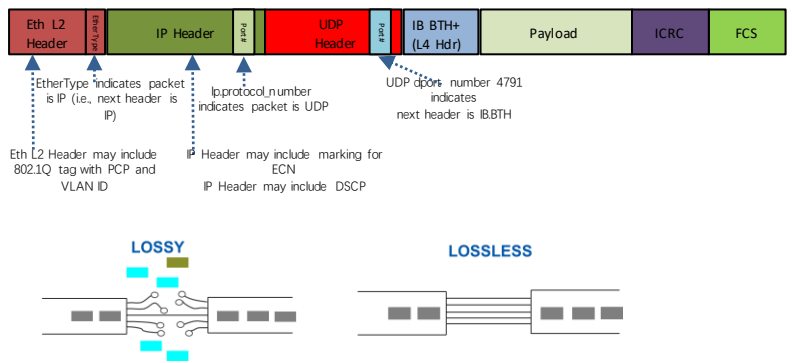
The Important Questions

- ✓ WHO is being impacted
- ✓ WHEN it happened
- ✓ WHAT is causing the problem
- ✓ WHERE is the problem
- ✓ WHY it is happening

Root Cause + how to fix it

以太网的性能保障 - RoCE

- 一键 RDMA部署
  - CLI “RoCE” vs 26+ commands in other NOS
- 支持RDMA的最佳硬件设计
  - 低转发时延和优秀的共享缓存设计
- NEO网管软件端到端管理



- Lossless、Semi-Lossless、Lossy多种RDMA部署模式
- RDMA和TCP混合部署
- RoCE over VxLAN
- Fast ECN

Parameters	Lossy	Semi-lossless	Lossless
Port trust mode L3	✓	✓	✓
Port sw-prio-TC mapping • **sw-prio 3**—TC 3 (RoCE) • **sw-prio 6**—TC 6 (CNP) • other sw-prio—TC 0	✓	✓	✓
Port ETS • **TC 6 (CNP)**—strict • **TC 3 (RoCE)**—WWR 50% • **TC 0 (other traffic)**—WWR 50%	✓	✓	✓
Port ECN absolute threshold 150-1500 **TC 3 (RoCE)**	✓	✓	✓
LLDP + Application TLV (RoCE) (UDP, Protocol: 4791, Priority 3)	✓	✓	✓
Enable PFC on sw-prio 3 (RoCE)		✓	✓
Prio 3 to roce lossless traffic pool			✓

NVIDIA | Lenovo NetApp