2021 中国智能网卡研讨会

CHINA SMARTNIC WORKSHOP

混合态异构高性能计算平台网络发展的趋势和挑战

演讲：奥工科技

目录

混合态异构高性能计算简介

高性能计算基本概念

高性能计算(High performance computing， 缩写 HPC) 指通常使用很多处理器（作为单 个机器的一部分）或者某一集群中组织的几台计算机（作为单个计 算资源操作）的计算系统和环境。有许多类型的 HPC 系统，其范围从标准计算机的大型集群，到高度专用的硬件。 大多数基于集群的 HPC 系统使用高性能网络互连，比如那些来自 InfiniBand 或 Myrinet 的 网络互连。基本的网络拓扑和组织可以使用一个简单的总线拓扑，在性能很高的环境中，网状网络系统在主机之间提供较短的潜伏期，所以可改善总体网络性能和传输速率。

并行计算

高端计算

超算

超级计算

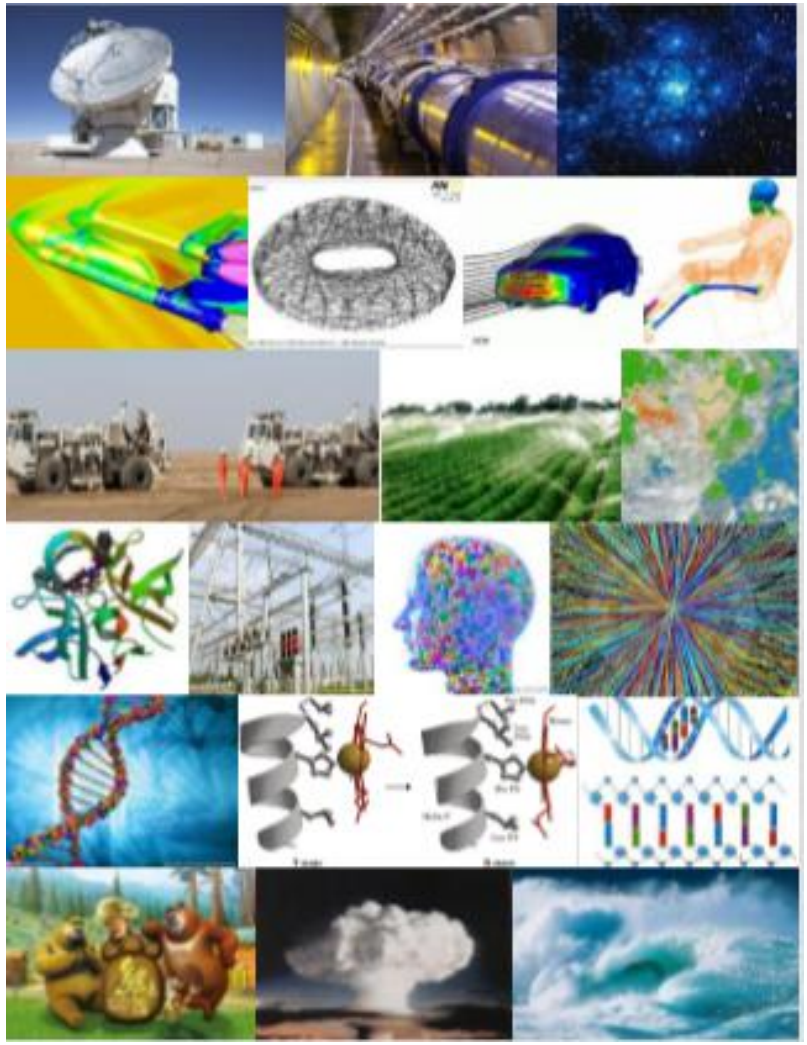
高性能计算的重要性

高性能计算已经深入到科学研究、国民生产、人民生活，作用和重要性越来越大，超算需求面逐步扩大、需求量迅速增长

高性能计算技术和应用在美国、欧洲、日本等发达国家的光柱度很高，被视为国家实力的象征

国家正在持续加大对**高性能计算的投入**，推动国家/区域/行业超级计算中心的建设

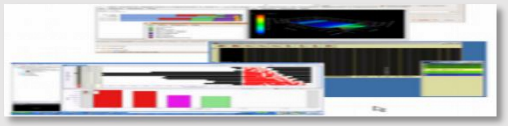
超级计算中心作为支撑和孵化平台，能推动科技发展、促进社会经济、服务社会民生



混合态异构高性能计算平台架构

应用环境





各类科学应用

编译环境

C,C++,Fortran OpenMP | docker 容器调度 | tensorflow、cafe深度学习 | 门户系统 | MPI | VNC

平台环境

OGSP奥工超算服务平台

作业调度 | 用户管理 | 平台计费 | 安全管理

系统管理 | 平台监控

OGFS高性能并行文件系统

硬件环境

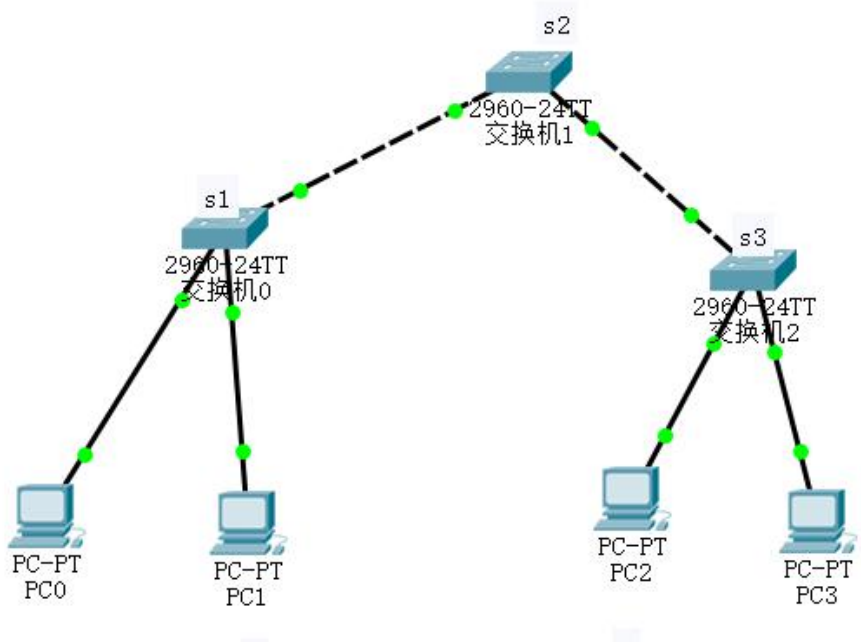
CPU资源池 | GPU资源池 | 存储资源池 | 网络资源池

云资源池

混合态异构高性能计算平台网络

高性能集群常用管理网络—以太网络

以太网是现实世界中最普遍的一种计算机网络。以太网有两类：第一类是经典以太网，第二类是交换式以太网，使用了一种称为交换机的设备连接不同的计算机。经典以太网是以太网的原始形式，运行速度从3~10 Mbps不等；而交换式以太网正是广泛应用的以太网，可运行在100、1000和10000Mbps那样的高速率，分别以快速以太网、千兆以太网和万兆以太网的形式呈现。



高性能集群常用网络—以太网络

存在的问题:

极易引起广播风暴

占用大量的网路带宽
甚至网络彻底瘫痪

数据内存拷贝
大幅增加延迟

中断处理消耗CPU资源
极大影响性能

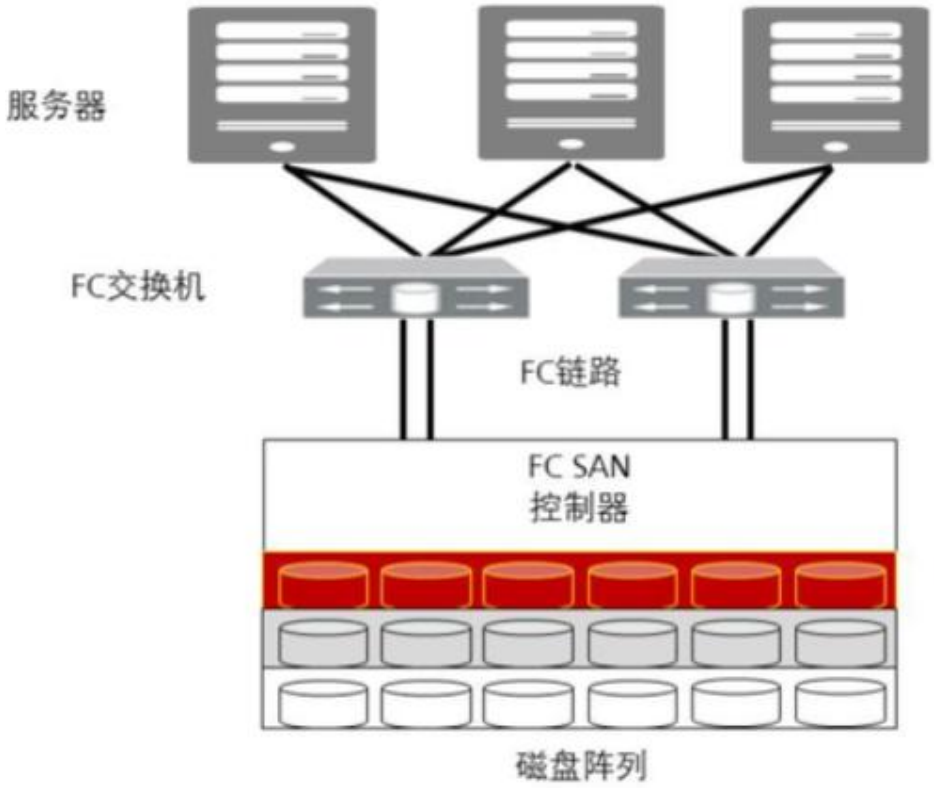
软件协议栈处理
性能低下

作为存储网络(ISCSI)
性能提升有巨大瓶颈

高性能集群常用网络—FC网络

FC是由美国标准化委员会（ANSI）的X3T11小组于1988年提出的高速串行传输总线，解决了并行总线SCSI遇到的技术瓶颈，并在同一大的协议平台框架下可以映射更多FC-4上层协议。

通道和网络双重优势，具备高带宽、高可靠性、高稳定性，抵抗电磁干扰等优点，能够提供非常稳定可靠的光纤连接，容易构建大型的数据传输和通信网络，支持1x、2x、4x和8x的带宽连接速率，随着技术的不断发展该带宽还在不断进行扩展，以满足更高带宽数据传输的技术性能要求。



高性能集群常用网络—FC网络

优势：

低CPU占用率

协议无丢包，能够有效保障存储数据的可靠性

存在的问题:

作为专用网络无法向计算网路延伸

如何发挥集群整体性能需要更高性能网络?

高带宽

- 实现节点间数据大吞吐
- 数据装卸阶段(存储IO)

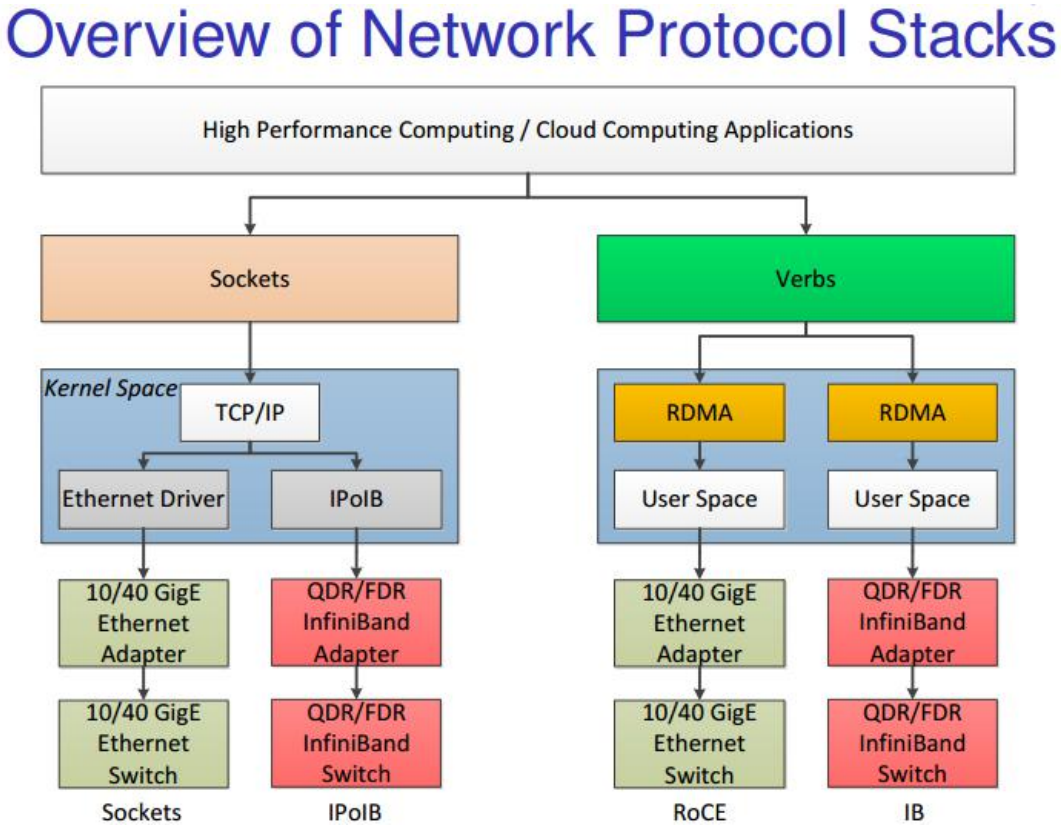
低延迟

- 节点间应用信息快速交换
- 节点信息交互阶段

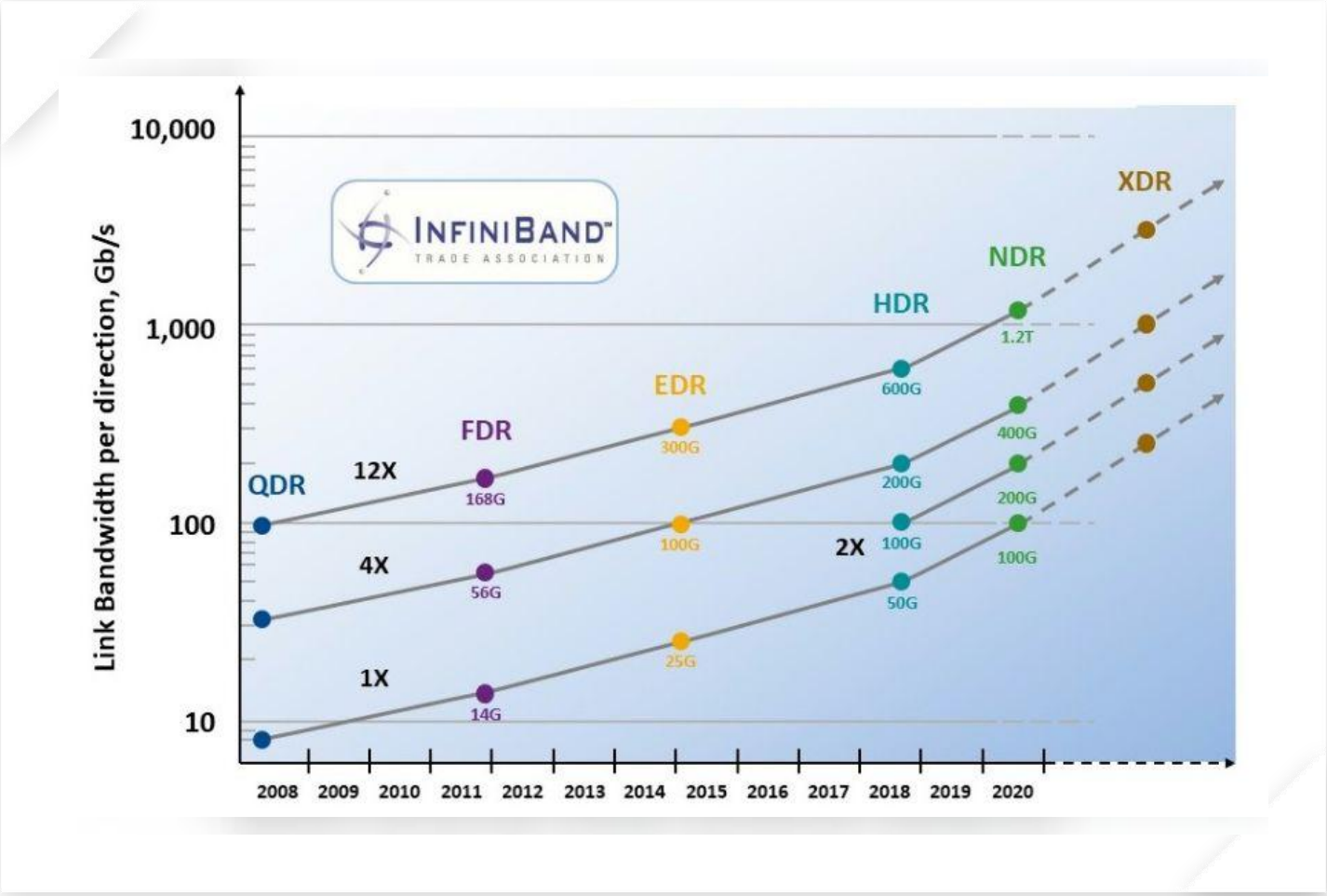
高性能集群常用网络—Infiniband网络

InfiniBand “无限带宽”技术是一个用于超算的计算机网络通信标准，它具有极高的吞吐量和极低的延迟，用于计算机与计算机之间的数据互连。

InfiniBand技术不是用于一般网络连接的，它的主要设计目的是针对服务器端的连接问题的。因此，InfiniBand技术将会被应用于服务器与服务器（比如复制，分布式工作等），服务器和存储设备（比如SAN和直接存储附件）以及服务器和网络之间（比如LAN， WANs和the Internet）的通信。

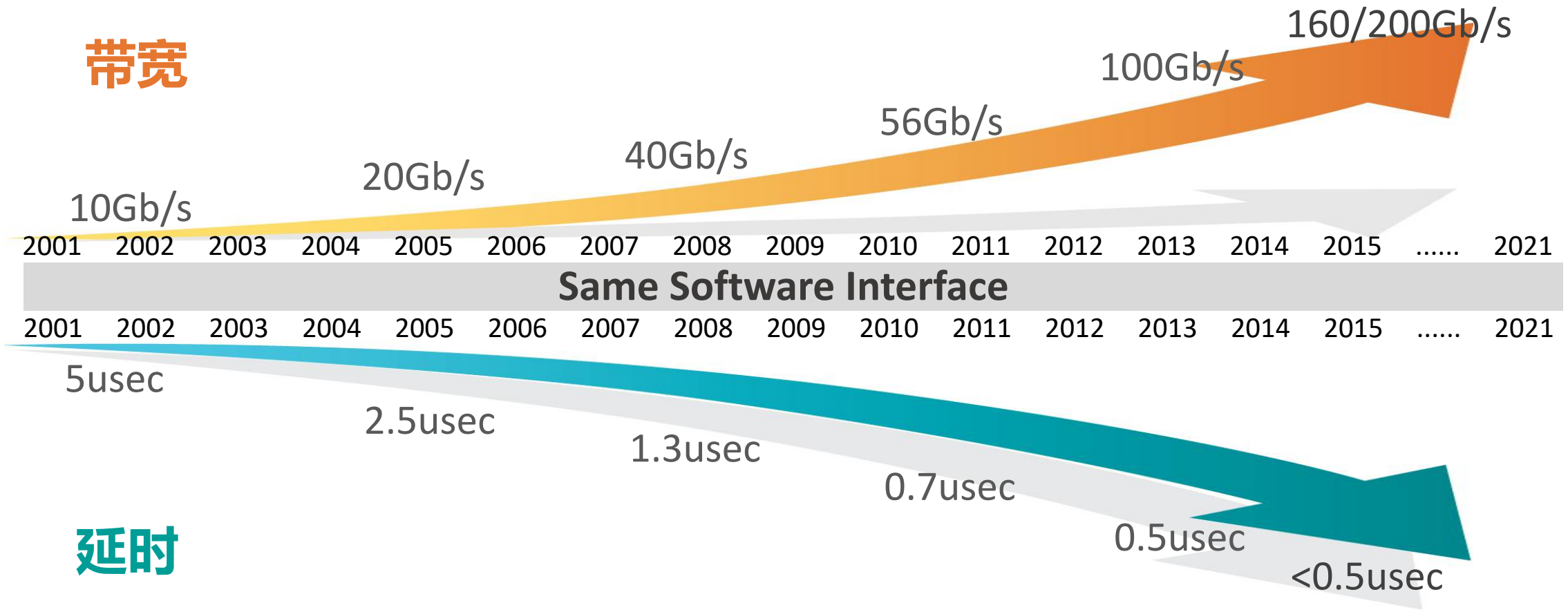


Infiniband网络设计路线



从SDR 的10Gbos——HDR 200Gbps发展到今天的NDR 400Gbps，受益于RDMA技术，CPU没有因为速率的大幅提升而牺牲更多的资源用于网络处理而拖慢整个HPC性能的发展，未来将会推出800Gbps的XDR。

Infiniband网络发展趋势



Infiniband网络开放协议

InfiniBand 软件主要由 OpenFabrics 开源联盟所开发

http://www.openfabrics.org/index.html

InfiniBand 标准由 InfiniBand贸易协会制定维护

http://www.infinibandta.org/home

计算网络技术特点

PCI串行高带宽连接

- DDR: 20Gb/s HCA连接
- QDR: 40Gb/s HCA连接
- FDR: 56Gb/s HCA连接
- EDR: 100Gb/s HCA连接
- HDR: 200Gb/s HCA连接

极低的延迟

- 低于1 微妙（纳妙级）的应用延迟

可靠、无损、自主管理的网络

- 基于链路层的流控机制
- 先进的拥塞控制机制可以防止阻塞

完全的CPU卸载功能

- 基于硬件的传输协议
- 可靠的传输
- 内核旁路技术

内存可提供远程节点访问

- RDMA读和RDMA写

服务质量控（QoS）

- Mellanox网卡: 提供多个独立的I/O通道
- 在链路层提供多条虚拟通道

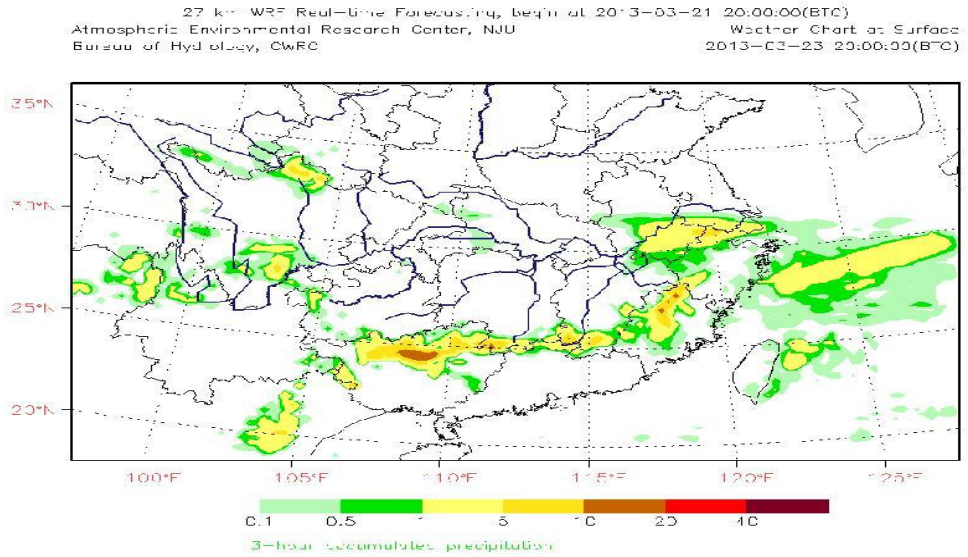
集群可扩展性和灵活性

- 一个子网可支持48,000个节点，一个网络可支持2128 个节点
- 在节点之间为平行路由
- 提供多种集群拓扑方式:胖树，3D

简化集群管理

- 集中路由管理
- 支持带内网络诊断和升级

高性能网络集群测试对比



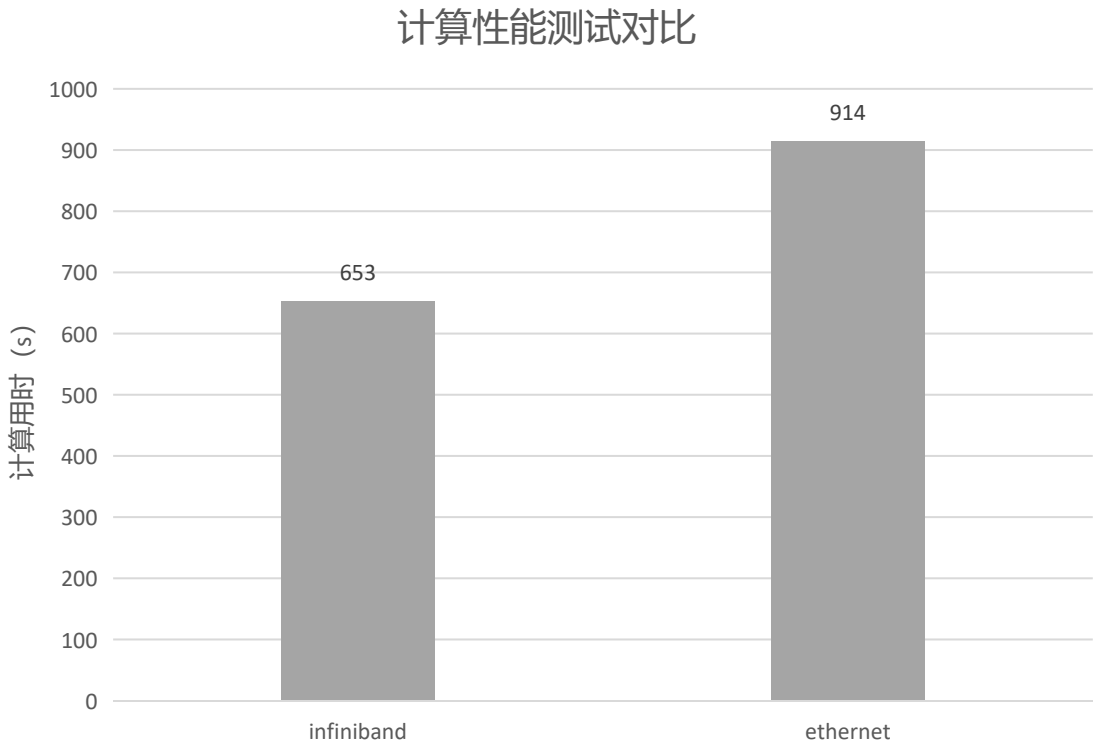
测试算例：三层嵌套区域，

第一层区域网格数182*172，空间分辨率是45 km；

第二层嵌套区域网格数是243*273，空间分辨率是15 km；

第三层嵌套区域网格数是300*249，空间分辨率是5 km。

三层嵌套区域垂直层数都是30层，预报时长12h。



测试平台：

***计算资源1: 2*AMD EPYC 7742 2.25GHz 64c**, 512GB DDR4, Intel compiler 2018u1, Intelmpi 2018u1, WRF-3.8.1,100Gb infiniband;*

***计算资源2: 2*AMD EPYC 7742 2.25GHz 64c**, 512GB DDR4, Intel compiler 2018u1, Intelmpi 2018u1, WRF-3.8.1,100Gb ethernet;*

网络在高性能计算中的重要性

Infinband在高性能领域的重要性

Infinband

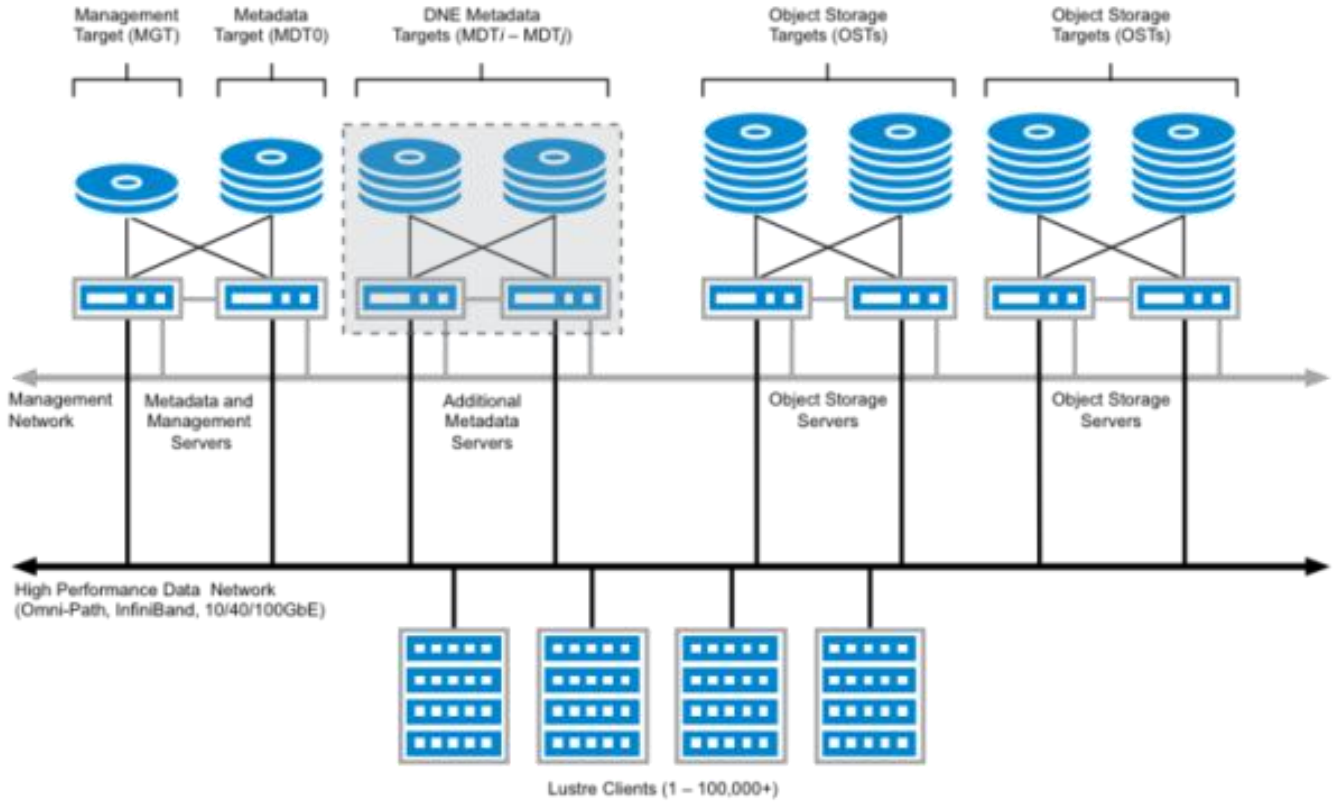
2012年，Intel完成对Qlogic Infiniband业务的收购，之后推出自研网络Omni-path用于超算领域。

2019年，Nvidia完成对Mellanox的收购以扩大其在超算及人工智能领域的绝对性地位。

存储的发展对网络的需求—lustre为例

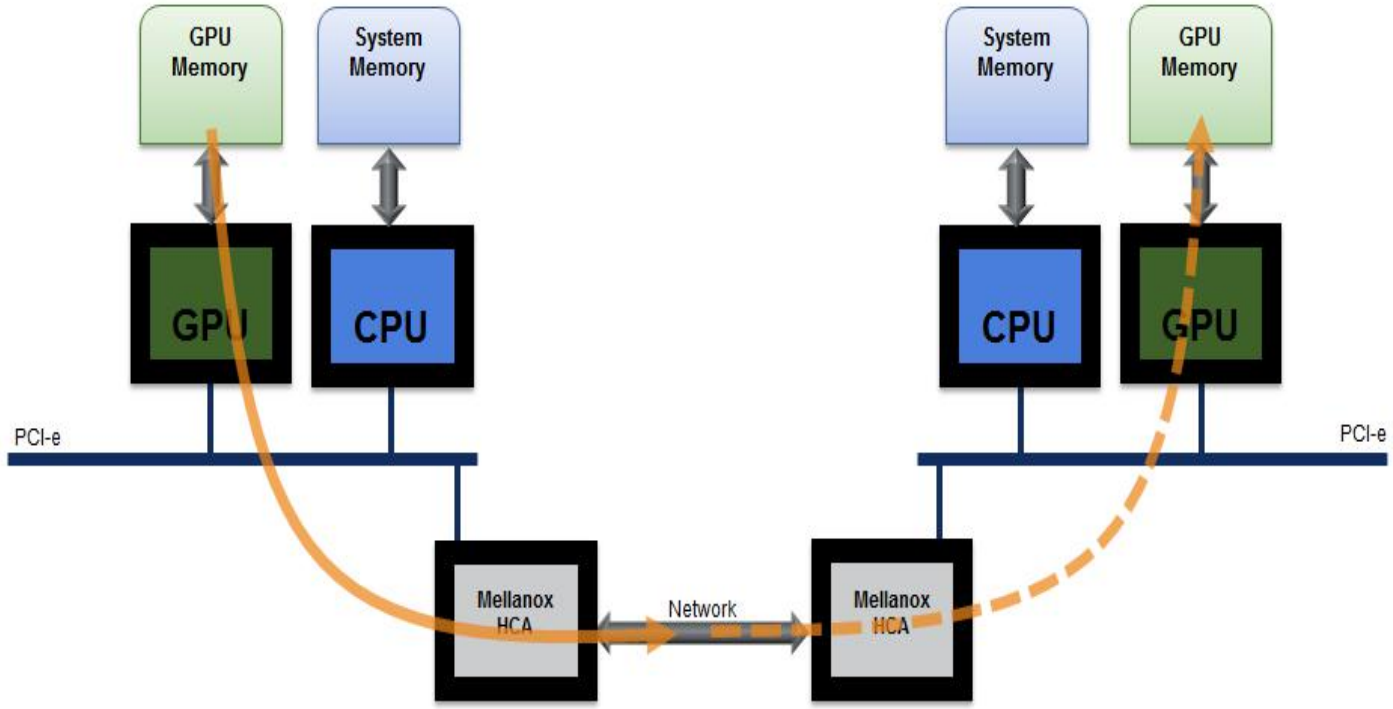
各角色之间通过高速网络互联

- 元数据服务器(MDS)
- 对象存储服务器(OSS)
- 管理服务器(MGS)



GPU Direct 技术

- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPI SendRecv efficiency between GPUs in remote nodes
- Based on PeerDirect technology



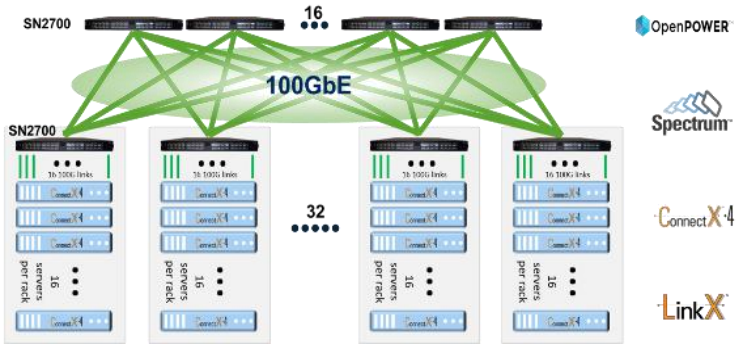
智能计算网络未来展望

智能计算网络未来的发展

And More Use Cases

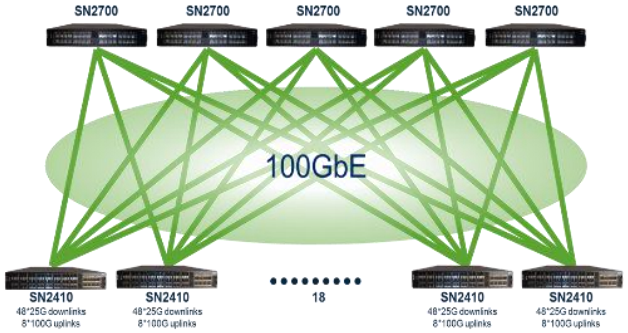
World’ s Fastest Real-time Big Data Analytic

实时大数据分析



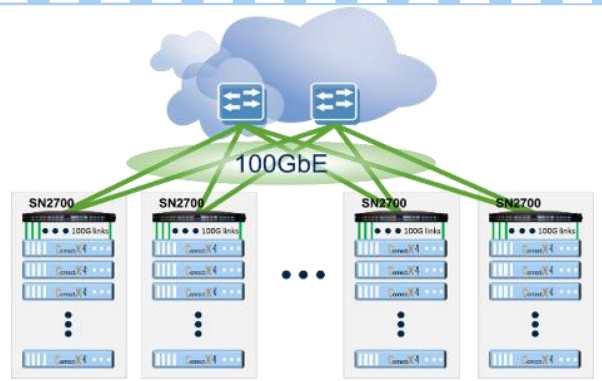
Intelligent Speech Information Services Platform

实时智能语音平台



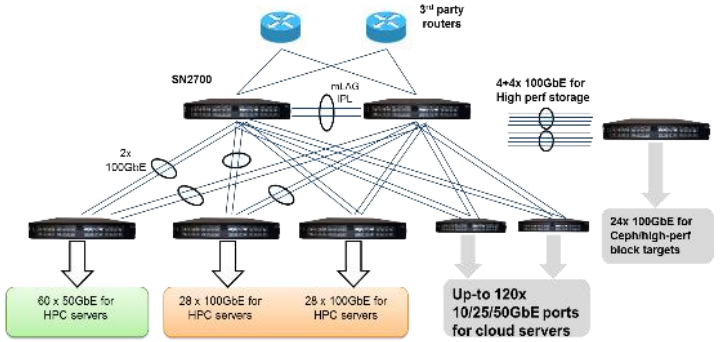
100GbE Deep Learning Platform

人工智能深度学习平台



High-Performance Cloud

超算云



THANKS