高速互联技术在AI超节点的应用实践

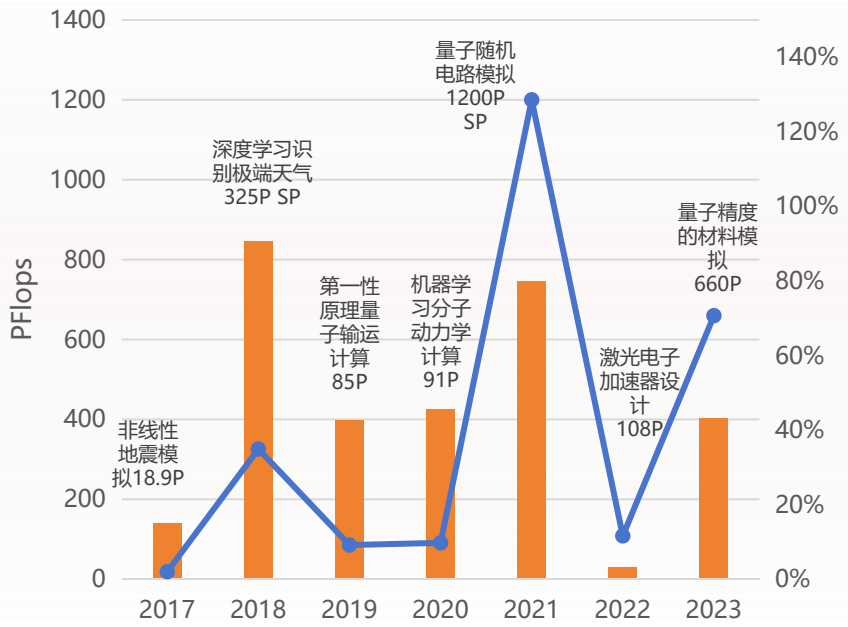
罗 宾

曙光信息产业（北京）有限公司

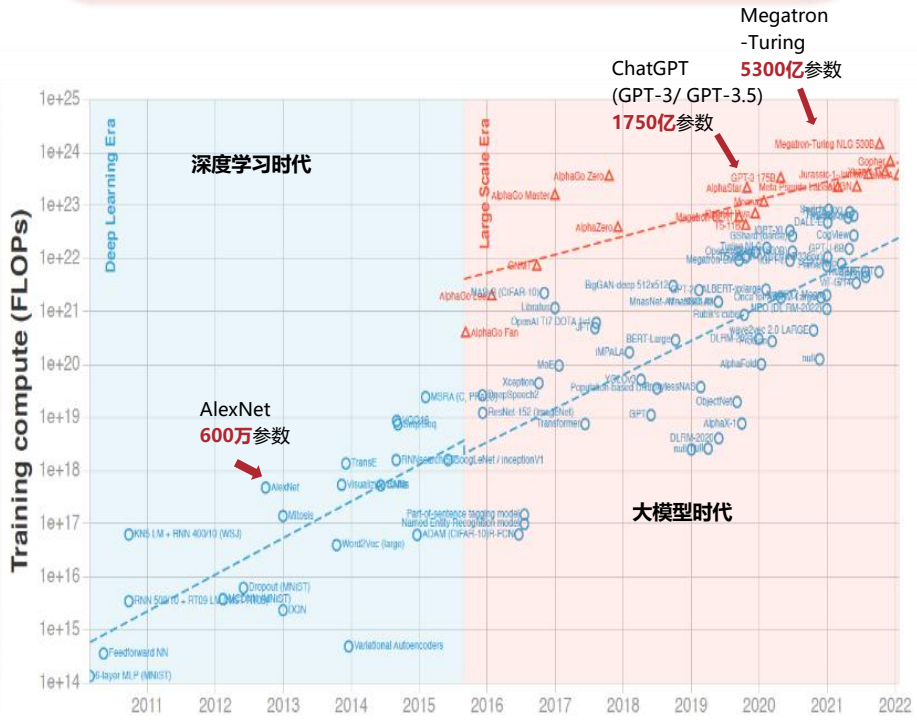
AI大模型发展驱动计算架构持续演进

Open AI Infra
A Chapter of GCC

超算：算力及应用已发展到E级



智算：模型参数万亿级，十万卡算力规模



系统正由传统节点向高密度AI超节点方向演进

算力竞赛进入“网络决胜期”

Open AI Infra
A Chapter of GCC

MOE模型

通信时间：40%-60%

稠密模型

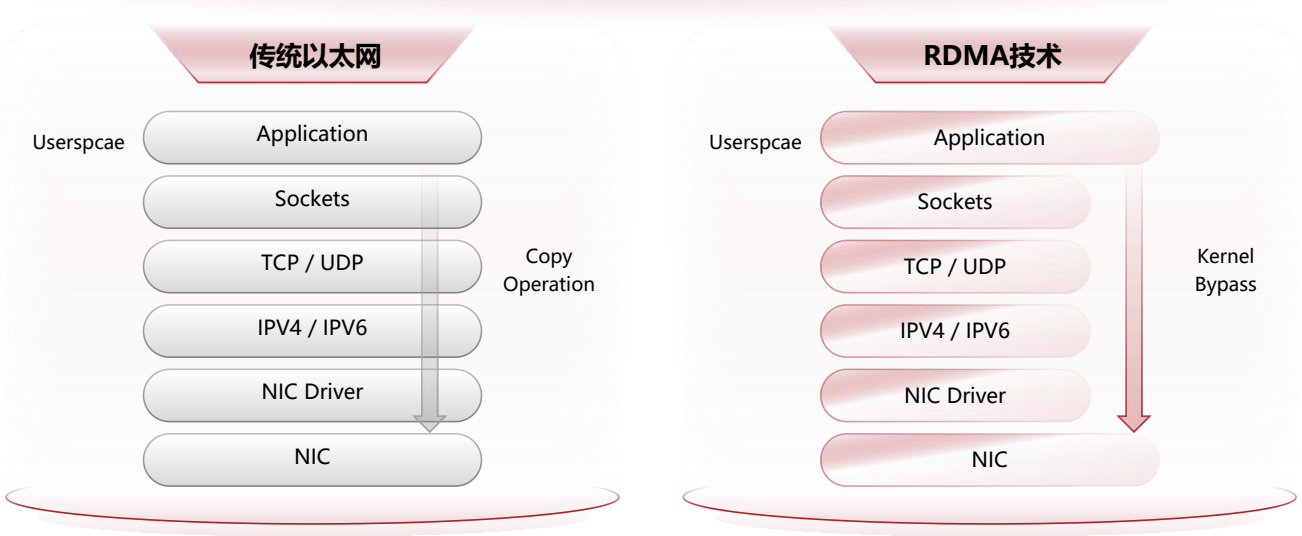
通信时间：10%-20%

HPC应用

通信时间：20%-50%

RDMA技术成为计算集群必备标准

实现**零丢包、高带宽、低延迟、高扩展和高容错**，极大提升通信效率成为关键。



若想富，先修路

算力设施趋势——从单卡算力提升转向系统效能提升

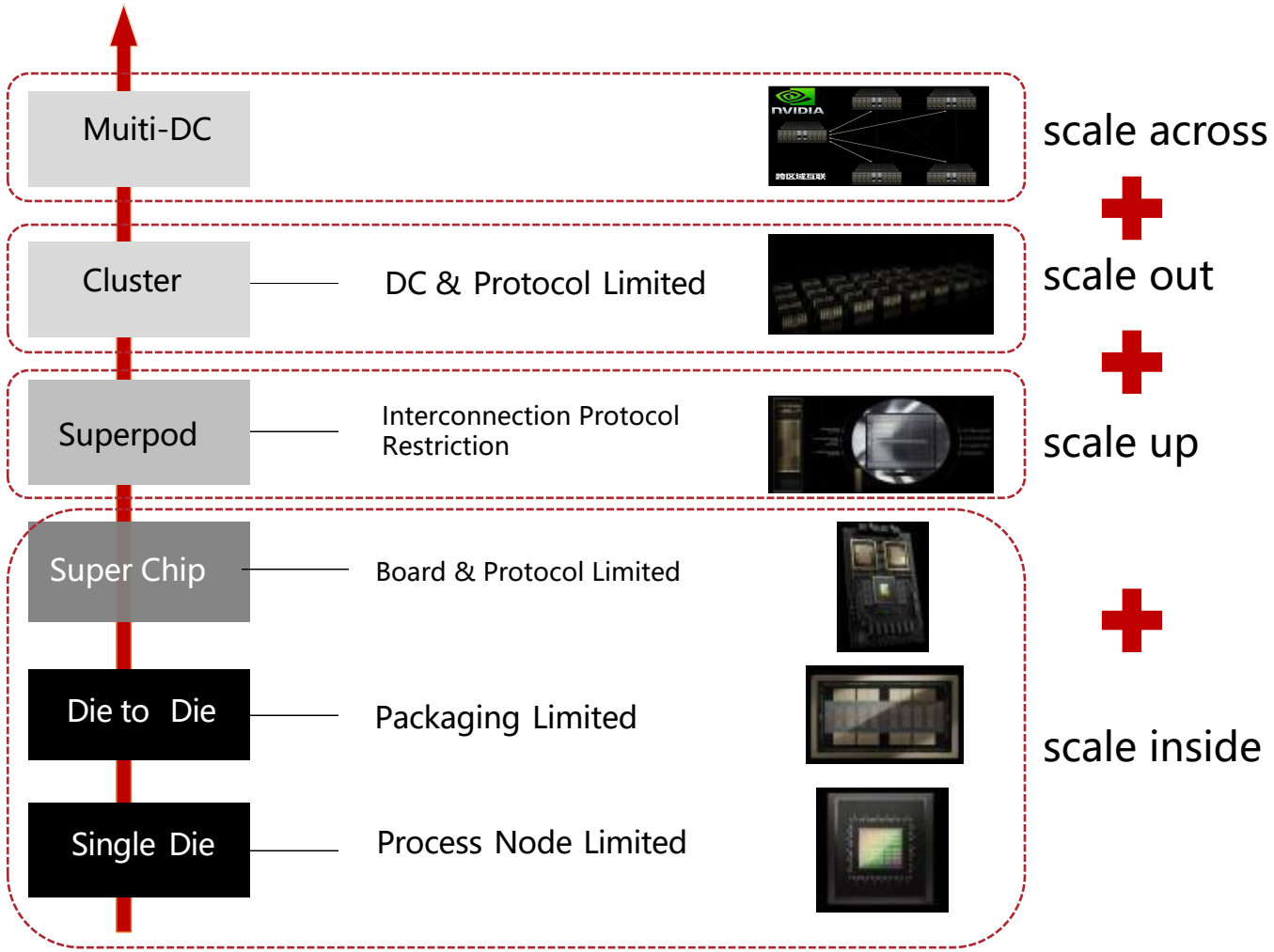
Open AI Infra
A Chapter of GCC

技术层面

- 受制程和功耗等制约，单卡算力增长放缓
- 通信开销占比增加，算力利用率困境
- 万亿参数时代需系统优化保障高可用

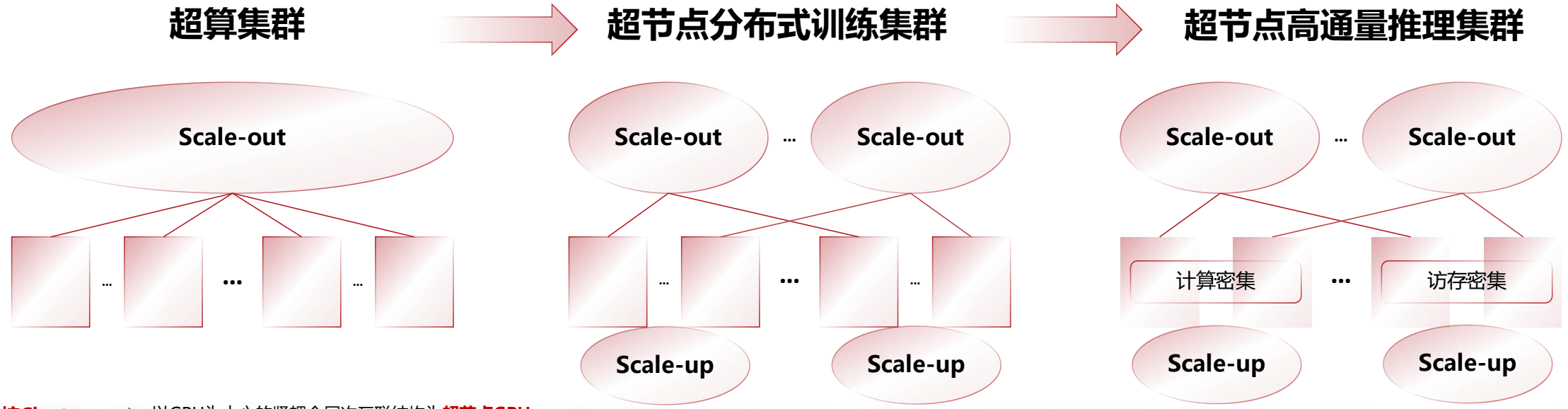
产业共识

- 2026 年 1 月ODCC 超节点大会：重构AI算力竞争逻辑，从“单点极致”到“系统致胜”
- 头部厂商布局超节点+集群扩展
- 单卡高成本不可持续



超节点架构：Scale-Up + Scale-Out协同演进

Open AI Infra
A Chapter of GCC



异构众核Cluster超智融合时代
2016年至今
曙光6000、7000、8000

- 以GPU为中心的紧耦合层次互联结构为**超节点GPU紧耦合**设计奠定技术基础。
- **异构并行编程模型、通信库/数学库优化**为超节点提供借鉴。

Cluser时代
2000年至今
曙光2000、3000、4000、5000

- 多机并行、规模扩展为超节点**集群扩展**提供依托。

SMP、MPP和DSM时代
1993年-2016年
曙光一号

- SMP多处理器并行、MPP多机并行提升算力的思路与超节点**多卡聚合为“大卡”**的理念契合。
- DSM分布式共享内存为超节点**内存统一编址、池化访问**奠定技术基础。

向量机时代
1964-1993年
Cray CDC6600

- 向量数据表示、SIMD技术、流水并行的设计逻辑成为**AI并行优化**的思想源头。

节点间同构
局域 Scale-up + 多轨 Scale-out
(多层级分布式并行算法)

X

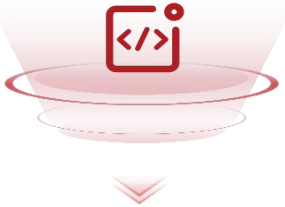
节点/单元间异构融合
(PD分离、AF分离...)
Scale-up + Scale-out

AI超节点本质：节点内Scale-Up + 节点间Scale-Out协同

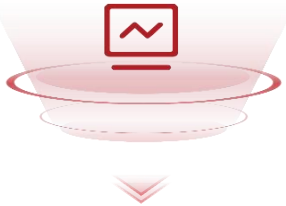
曙光scaleFabric高速网络介绍

当前用户面临四大痛点

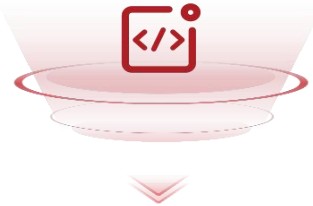
组网成本高



供应不稳定



服务能力受限



国外厂商垄断

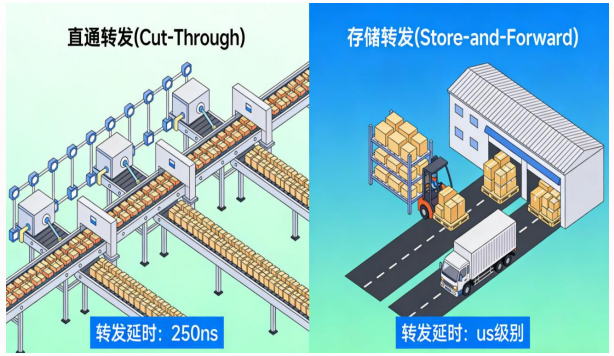
原生RDMA技术在AI/HPC集群中不可替代的优势

Open AI Infra
A Chapter of GCC

RDMA性能发挥
对网络Lossless特性有极高要求

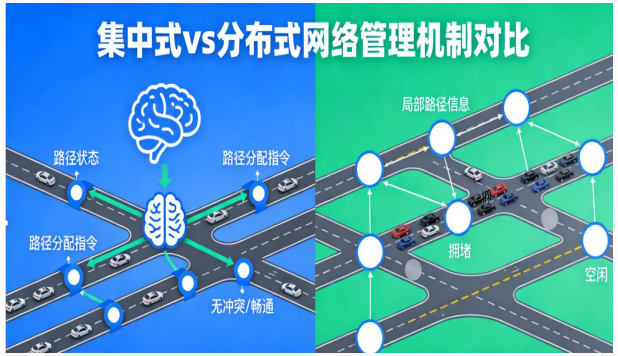
基于Credit-base流控+链路层重传技术可以做到真正的无损

通信延时
对应用性能有显著影响



基于VCT(虚切换)及Cut-Through的交换机制相比存储转发机制有更低的时延

需把整个集群
作为单一系统进行管理



基于SDN的集中式控制让路由收敛更快，容错能力更强，问题定位更方便

当前各种热点的scale-out网络协议，本质上都是把以太网改得越来越像原生RDMA技术

首款国产原生无损RDMA高速网络芯片

Open AI Infra
A Chapter of GCC

首款国产400G原生RDMA高速网络系统

双芯国产·无损极速·智算基石

scaleFabric400 网卡芯片

- 400Gb/s 高带宽
- 低至0.93us 端到端传输延迟
- 自研RDMA引擎
- 采用自研112G PAM4 SerdesIP
- 链路层、传输层双层重传保障

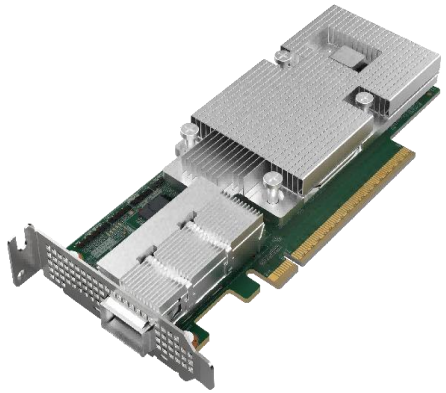
scaleFabric400 交换芯片

- 40 x 800G 或 80 x 400G
- VCT机制，转发延迟低至260ns
- 64Tbps　高交换容量(双向)
- 采用自研112G PAM4 SerdesIP

曙光scaleFabric400全系列网络产品

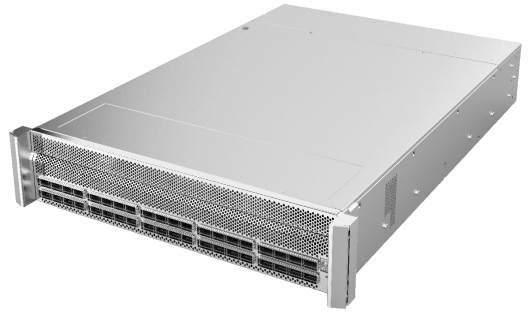
Open AI Infra
A Chapter of GCC

scaleFabric400
单口 标准网卡



scaleFabric400 1U
80口 液冷交换机

scaleFabric400 2U
80口 风冷交换机



曙光scaleFabric400技术规格全面对标英伟达IB NDR

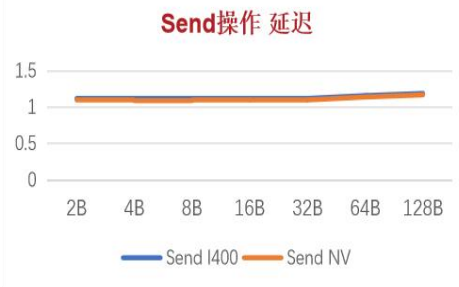
Open AI Infra
A Chapter of GCC

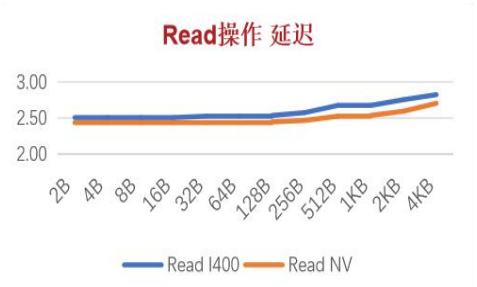
	scaleFabric 400	InfiniBand NDR	主流RoCE网络
端口速率	400Gb/s	400Gb/s	200|400Gb/s
协议类型	原生RDMA	原生RDMA	以太网
无损特性	Credit based流控+LLR	Credit based流控	PFC流控+ECN
交换端口密度	80*400Gb/s	64*400Gb/s	64*400Gb/s
交换容量	64Tb	51.2Tb	51.2Tb
交换延时	~260ns	~200ns	>450ns
网卡通信延迟	<1us	<1us	>1us
网卡最大QP数	256K	128K	128K
网络管理	即插即用	即插即用	水线优化

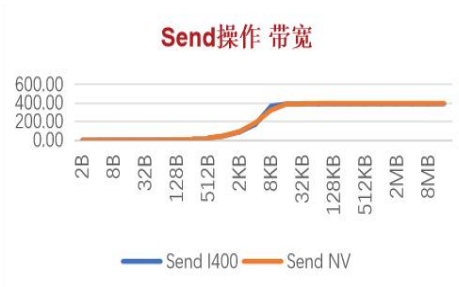
高性能-基于国产IB网络的应用实践验证

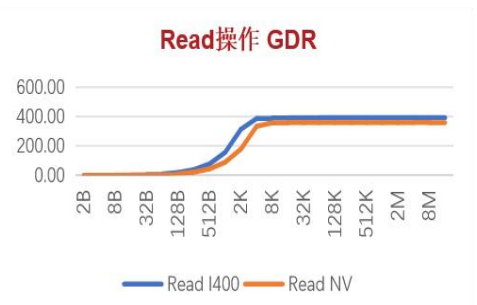
Open AI Infra
A Chapter of GCC

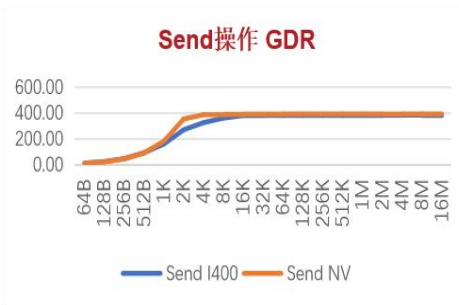
390Gb/s+传输带宽,0.93us端到端时延



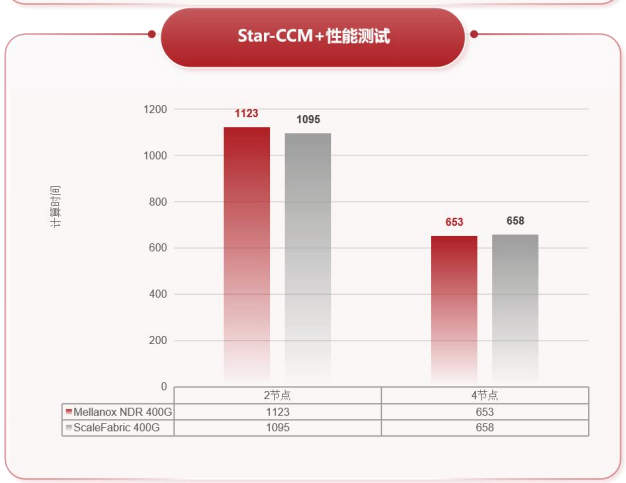








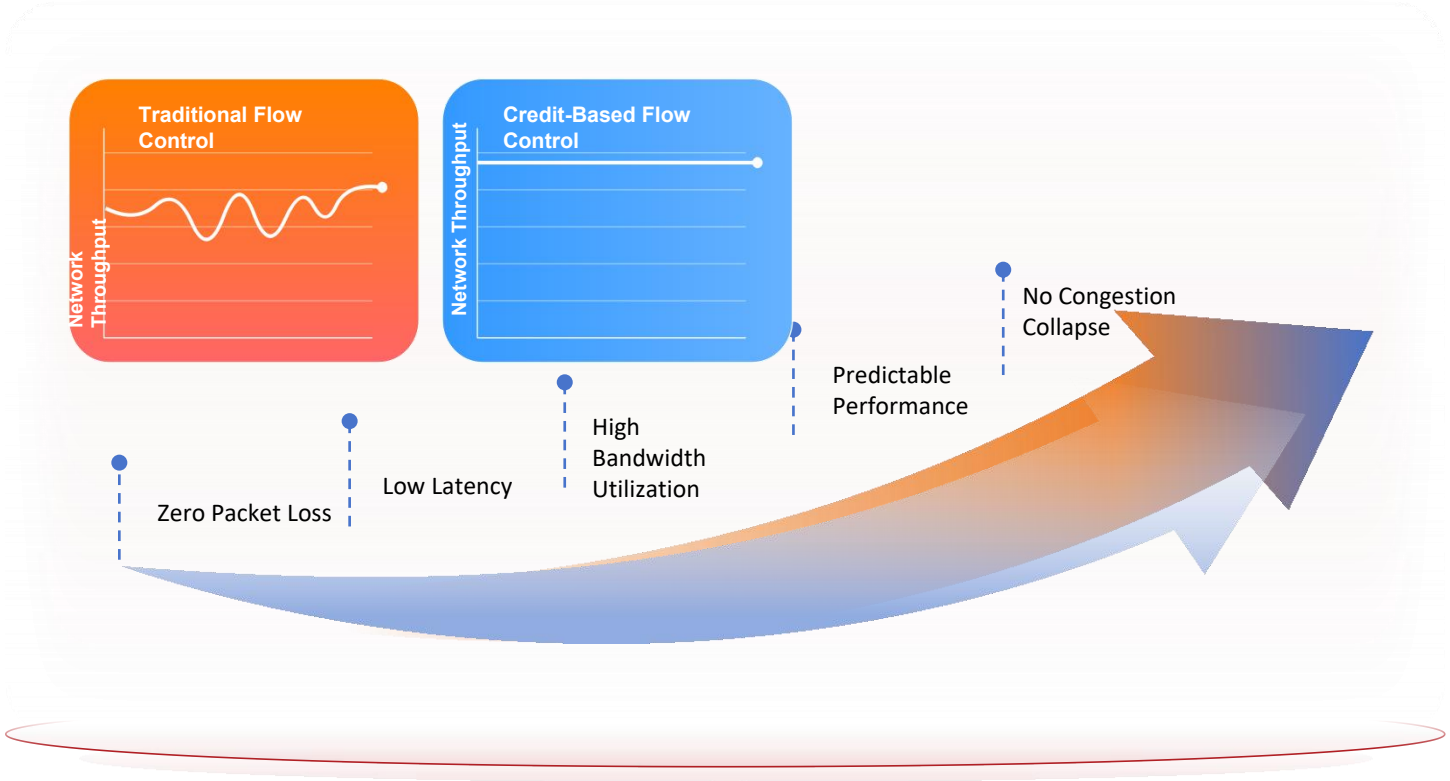
- Fluent和Star-CCM+运行命令参数原生支持
- 性能可达NVIDIA InfiniBand NDR 400G的96%-105%



真无损-即插即用的最优网络

Open AI Infra
A Chapter of GCC

基于信用的流控+链路层重传机制



真正的无损机制
天然不需要网络参数优化来获得网络的稳定性

无需调优零配置，36小时完成3万卡规模集群部署交付

400G以后时代拥塞控制调试愈发困难

高可靠-自主研发112G PAM4高速Serdes IP

Open AI Infra
A Chapter of GCC

较差链路质量对计算集群的影响

集群效率低

0.1%的丢包率计算集群效率降低50%

影响范围大

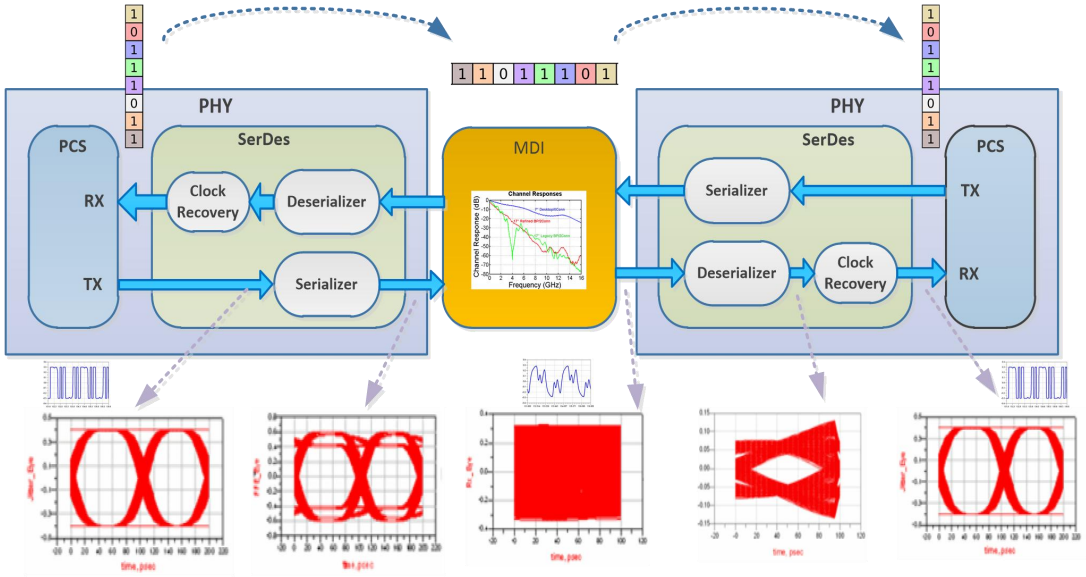
链路质量问题影响范围与并行节点数有关。

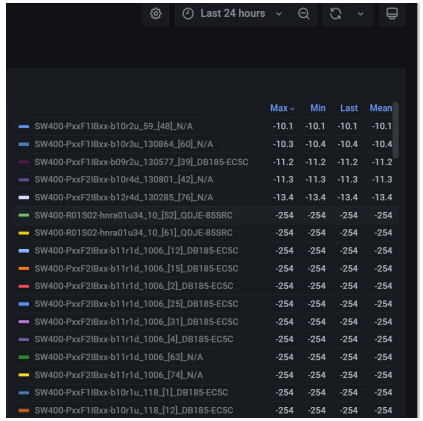
应用性能低

重传过多，增加通信延迟，影响应用性能

用户体验差

链路问题最终会传导至用户端，损害用户体验





L2-Act端口数	34447
L3-Act端口数	12140
L1-Act端口数	39325
HCA-Act端口数	16939

业内领先的驱动能力(>42dB@BER=Xe-6)

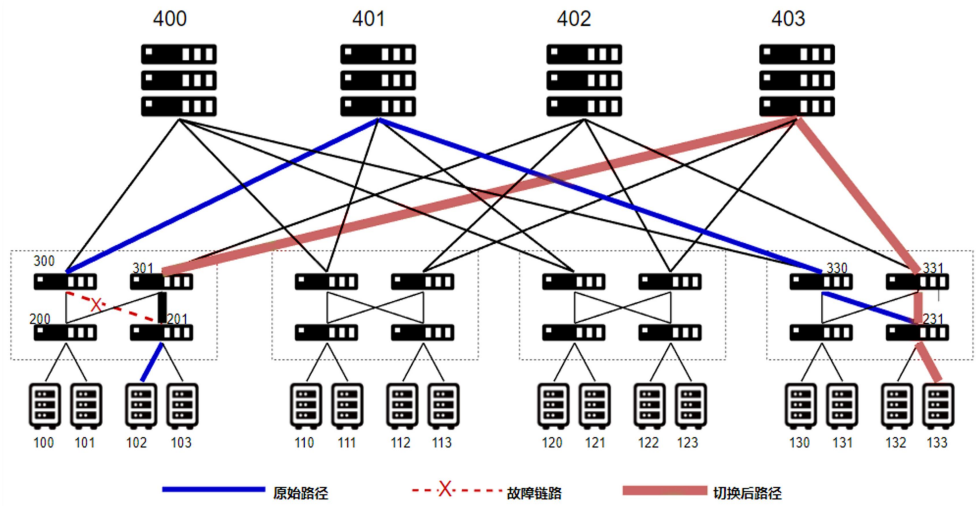
Serdes作用

高速Serdes自主研发团队，是链路稳定性保障的底气

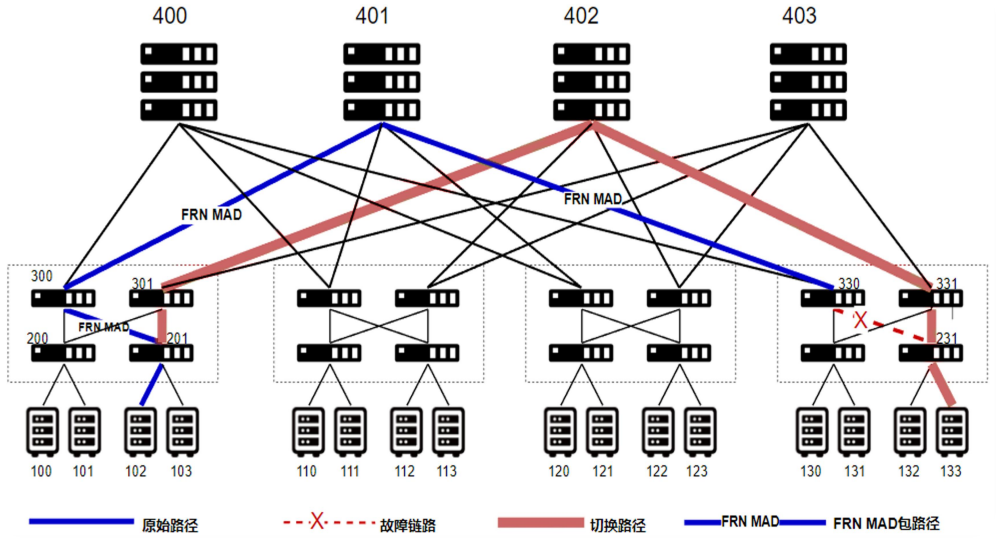
高容错-链路故障路由快速恢复技术

Open AI Infra
A Chapter of GCC

毫秒级链路故障路由恢复时间，且不随网络规模增长而增长



上行链路故障

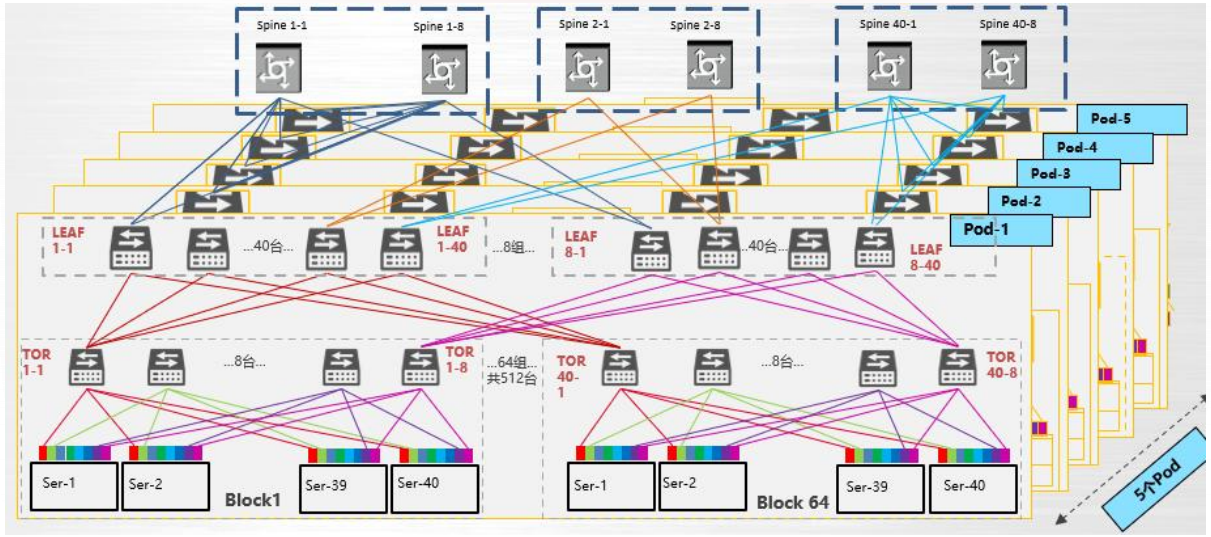


下行链路故障

高扩展-支撑10万卡集群规模建设需求

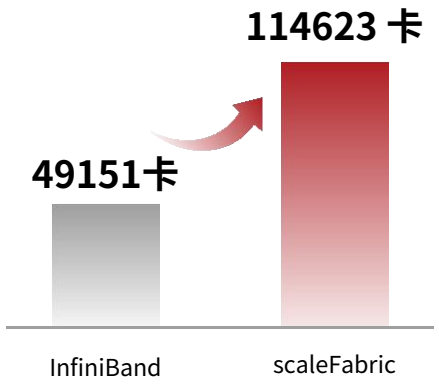
Open AI Infra
A Chapter of GCC

基于scaleFabric网络的10万卡超集群组网



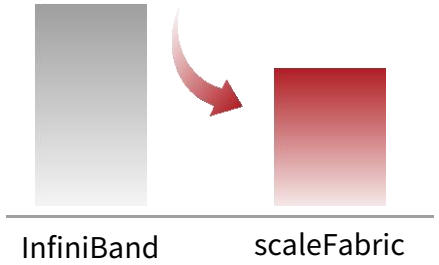
2.33倍

单一网络互连规模



50%

网络总体成本降低



生态优势-全面兼容原生IB生态

Open AI Infra
A Chapter of GCC

提供原生的Verbs接口及全面的通信库适配，无缝兼容各种HPC/AI应用

网络管理和维护方式均符合InfiniBand用户的使用习惯

Fluent Validation | WRF | CFX | OpenFOAM
GROMACS FAST. FLEXIBLE. FREE. | STAR-CCM+ | ROMS | LS-DYNA
LAMMPS | VASP | MATLAB | gaussian

KIMI | Llama | ultralytics YOLOv5 | Mask RCNN | Qwen
deepseek | Tencent Hunyuan | 百川智能 BAICHUAN AI | MISTRAL AI_ | ChatGLM

SGL | TensorFlow | DeepSpeed | vLLM | PyTorch | PaddlePaddle | Megatron-LM

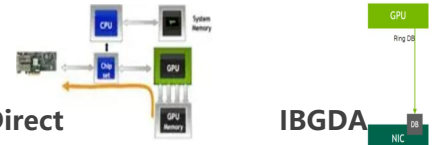
支持主流通信库接口 | RCCL | NCCL | Open MPI | Intel MPI | MPICH | UCX Unified Communication X | Verbs API

scaleFabric

主流OS和容器 | CentOS | Ubuntu | 中科方德 | kubernetes | docker

国产处理器与GPU加速卡

各种通信架构 | GPUDirect RDMA | IBGDA



曙光scaleX40超节点介绍

曙光40卡超高密度AI超节点整体架构

Open AI Infra
A Chapter of GCC

标准19英寸

兼容主流机柜

16U

部署密度是8卡机的2.5倍

40张
GPU

> 28FLOPS
FP8算力

> 5TB
HBM总显存

> 80TB/s
访存总带宽

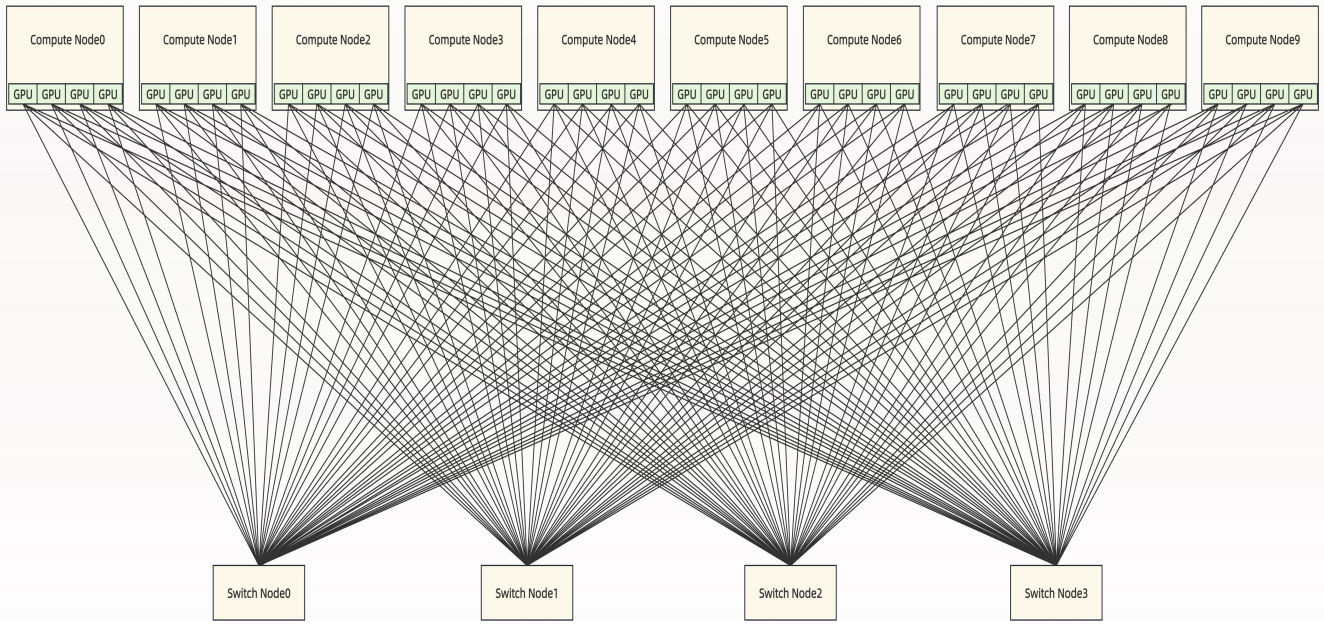
> 17TB/s
Scale-up聚合带宽

- 一层组网，通信更高效
- 解耦设计，部署门槛低
- 正交架构，更稳更可靠
- 全栈集成，开箱即可用

节点内部高速互联（Scale-Up）

Open AI Infra
A Chapter of GCC

40卡通过一层Scale-up总线技术实现全互连，更快更稳



高速总线

- **~10倍** 比NDR IB组网提高
- 支持Load/Store等内存语义操作
- 支持显存统一编址

一层组网

- **百ns级** 芯片P2P单向通信时延
- **30-100%** 相比二层组网时延降低
- **30-50%** 相比二层组网故障率降低

曙光40卡超节点带来的价值

Open AI Infra
A Chapter of GCC

强

GPU算力强

参数强，算力更强

144GB HBM3　　**FP8原生支持**

大显存统一编制，模型参数任意访问

快

Clos全互联低时延

一级互联，超高带宽、超低时延

>80TB/s	>17TB/s	百ns级
单PoD 访存总带宽	单PoD 片间互连总带宽	芯片P2P单向 通信时延

计算不用等通信

稳

非光互联运行稳

高可靠，易运维

Scale-Up域无光互联　　**计算、交换节点直接对插**

部署快、运行稳、成本低

通

全栈方案硬软通

全系芯片、软件均为同厂商方案

硬	软
国产 CPU	HSL开放协议
国产 GPU	
HySwitch	DAS

全栈方案，全栈服务

易

完备生态使用易

兼容CUDA/ROCm双生态

>800个	>95%
AI模型全面适配	极高的API接口覆盖度

新模型“拿来即用”

开

开放解耦落地易

软硬件、IT与机房环境可解耦

支持风液CDU解耦机房环境
支持机框与机柜解耦
支持模型与芯片解耦

开放解耦，快速部署

超大规AI模智算集群实践

Open AI Infra
A Chapter of GCC

全国产

国产CPU+GPU+各类SW芯片
全栈国产IT软硬件

多功能

芯片全精度覆盖
超智融合算力服务千行百业

大生态

国产x86+国产GPGPU生态
软硬协同打通应用落地
产学研用协同

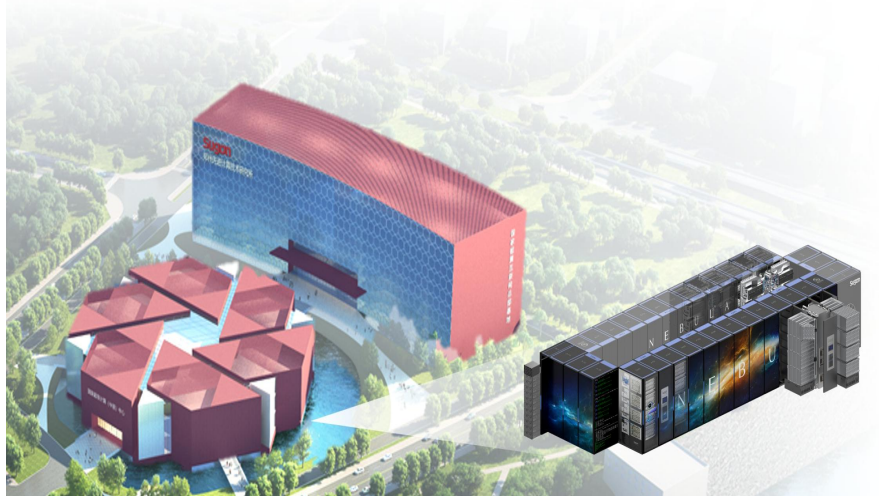
超扩展

Scale out构建超大规模集群
Scale up提升单节点算力密度

高能效

浸没式相变液冷
高压直流配电
PUE≤1.04

十万卡智算集群



郑州4E@FP64+50E@FP16算力集群

20+算力中心建设

寻求性能、成本和能效的最优解

100+细分行业
1000+场景落地

超智融合支持多功能产业发展

AIGC/大模型　气象海洋预报　生物制药　燃烧流体　新材料研发　…… ……

千行百业算力综合体实践

AI正从“基础资源”升级为“智能引擎”，将深度重构各行业的价值体系

深言科技 DeepLang AI　小米汽车　Moonshot AI　智谱

中科星图　智元机器人　BAAI 智源研究院　中国移动 China Mobile　百川智能 BAICHUAN AI

天津移动智算中心
商汤大模型/人工智能/生物医药
智慧城市/具身智能/智能制造

北京人工智能公共算力平台
智源大模型/三代视频大模型
月之暗面大模型/奇绩大模型

WA怀来液冷智算中心
阿里百炼平台/曙光液冷方案
行业大模型训练、推理、知识库

中科院地球系统数值模拟装置
地理环境模拟/大气预报大模型
海洋环境预报/地球表面温度预测

山东魏桥国科智算中心
智能制造/新能源汽车/科研实验
先进材料与装备/无人机/仿真设计

THANK YOU