面向超节点的新一代网卡技术发展趋势

姓名：武晓军

灵达科技 产品经理

面向超节点的新一代网卡技术发展趋势

姓名：武晓军

灵达科技 产品经理

CONTENTS

目录

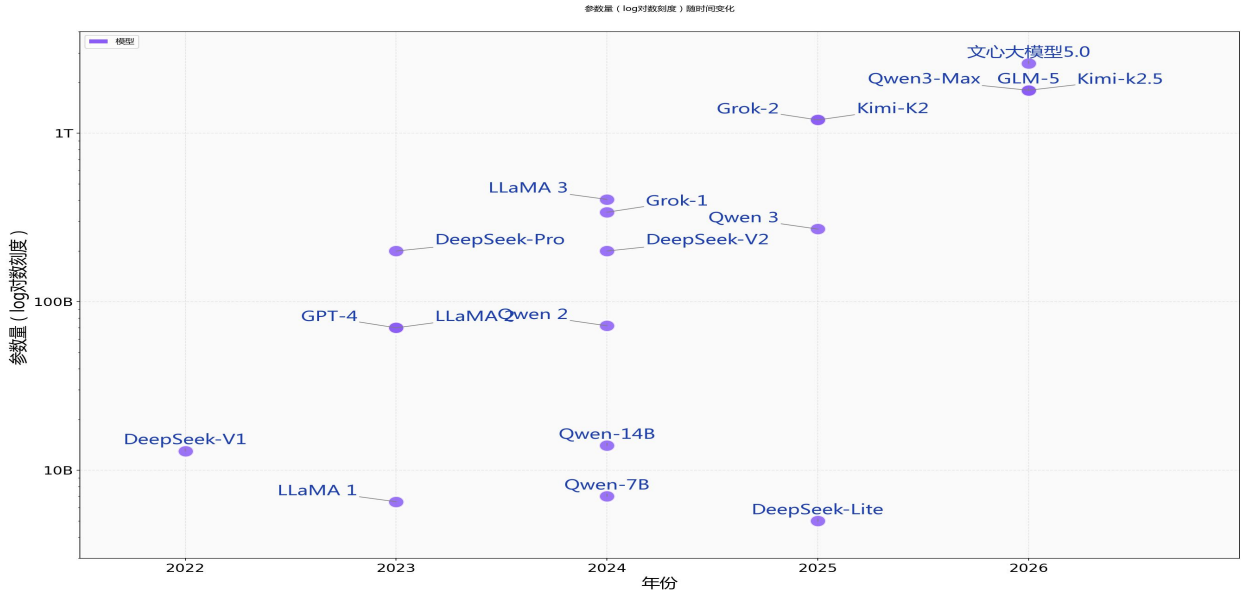
AI发展趋势

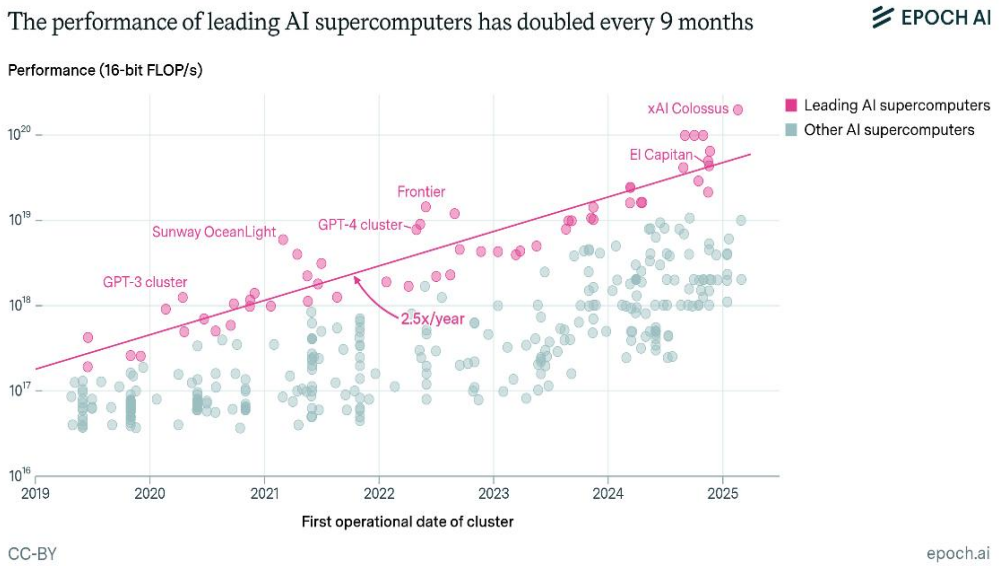
Open AI Infra
A Chapter of GCC

生成式AI → 推理AI → 代理式AI → 世界模型 → 物理AI

内容生成　深度思考　任务执行　环境模拟　现实交互

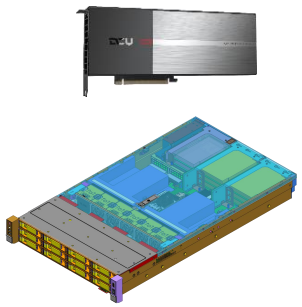
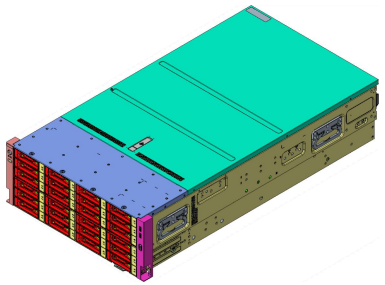
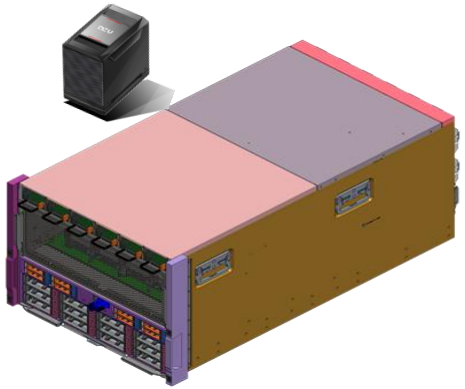
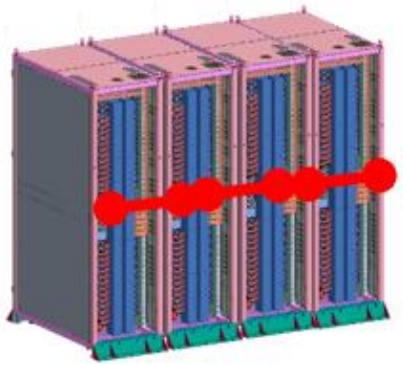
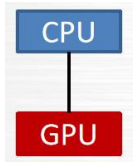
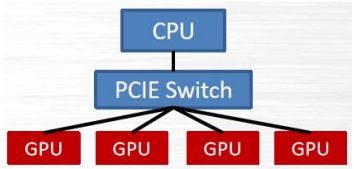
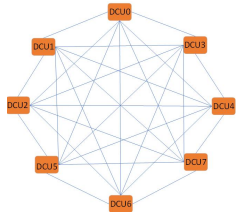
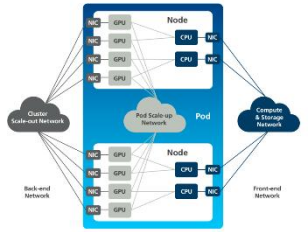
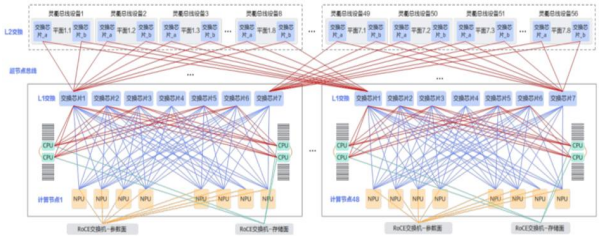
- **训练走向推理**：随着Agent和推理模型普及，推理阶段的算力消耗增速已超过训练，数据中心正从“训练中心”转向“推理工厂”
- **互联至关重要**：万亿参数MoE模型下，单个Agent上下文可达百万级token，KV cache大小呈指数级膨胀，多智能体协同、多轮持续推理对AI基础设施提出严峻挑战，互联已从辅助基础设施升级为Agent AI的“神经中枢”





AI基础设施演进

Open AI Infra
A Chapter of GCC

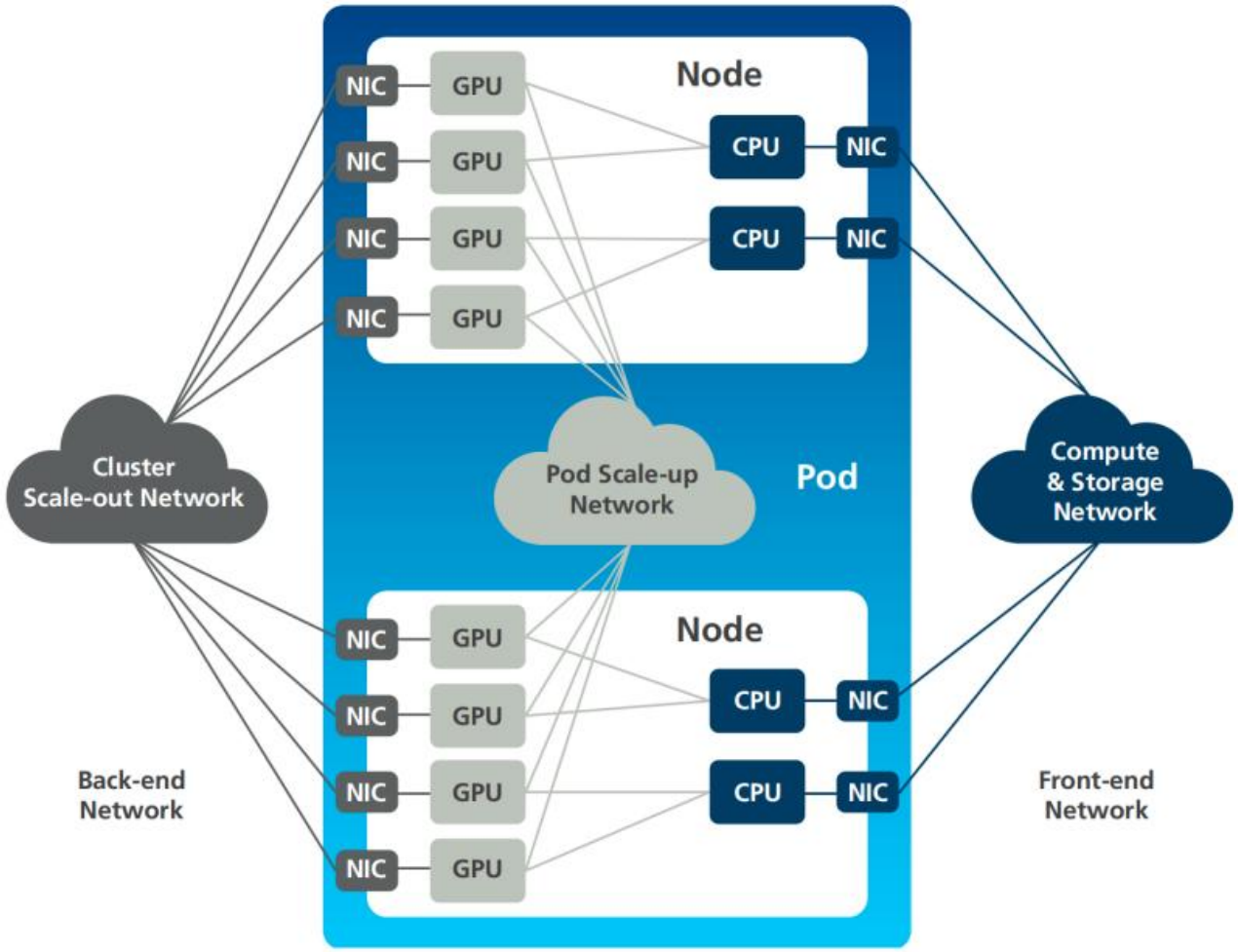
单卡	多卡直连	OAM(8卡)	超节点(32-百卡)	集群(千卡-万卡-10万卡)
				
2U4卡	4U8卡PCIe	8卡OAM	NVL36/72	智算集群
				
单卡-模型放不下	单机-互联性能不足(带宽/延迟)	单机全互联-扩展性不足	ScaleUp/ScaleOut 并行发展：系统复杂，演进新的互联协议、光互联	

网络速率需求 25G 100G 200G 400G 800G 1600G

Scale Up(PCIE、NVlink、HSL、UALink、OISA、UB、SUE、EthLink...) 百花齐放

Scale Out(ROCE、IB、UEC、增强型RoCE) 趋于收敛

超节点网络拓扑



Scale Up网络:

- 用于超节点/HBD 域GPU和GPU之间互联，特点是极高带宽、极低时延、Load store 语义、内存共享，但GPU互联规模有限，当前典型为32~128卡，未来演进规模一般在1~2K以内
- 业务流量主要是大模型中的张量并行和专家并行
- GPU直出网络，距离较短，以电为主，未来电光混合

Scale Out网络:

- 用于超节点更大规模GPU间互联，特点是高带宽、低时延、Message 语义、远程内存直接访问，GPU互联规模很大，可以扩展到万卡、10万卡集群，走RoCE或IB，需要强壮的拥塞控制、负载均衡策略
- 业务流量主要是大模型中的数据并行、流水线并行和专家并行
- AI NIC，以NV CX7、CX8为典型代表，距离较长，以光为主

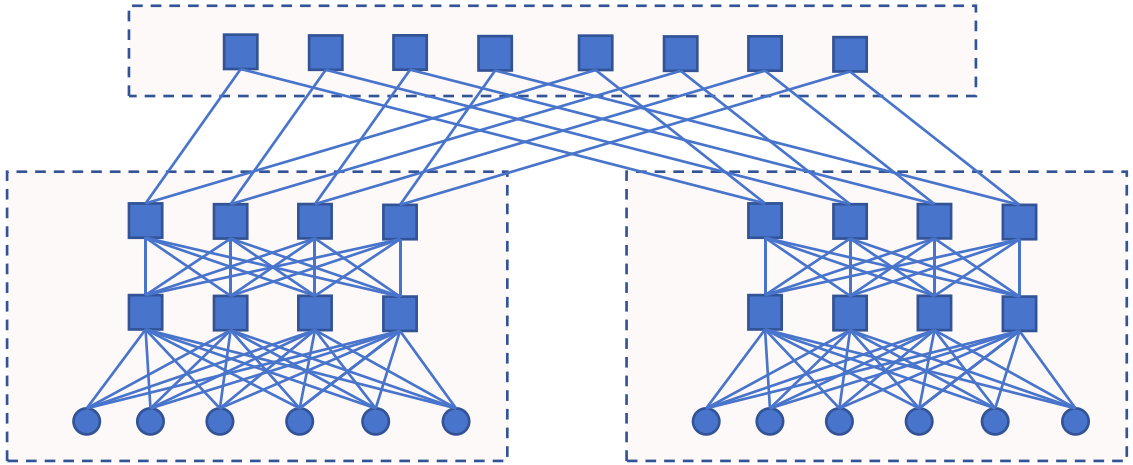
Frond End 网络:

- 存储/业务网络，训练/推理数据拉取，Checkpoint 读写，特点是高带宽、低时延、Message 语义，走以太网或RoCE，虚拟化、多租户隔离，传统VPC网络
- DPU或标准网卡，算力卸载，距离较长，以光为主

Scale out 网络面临挑战

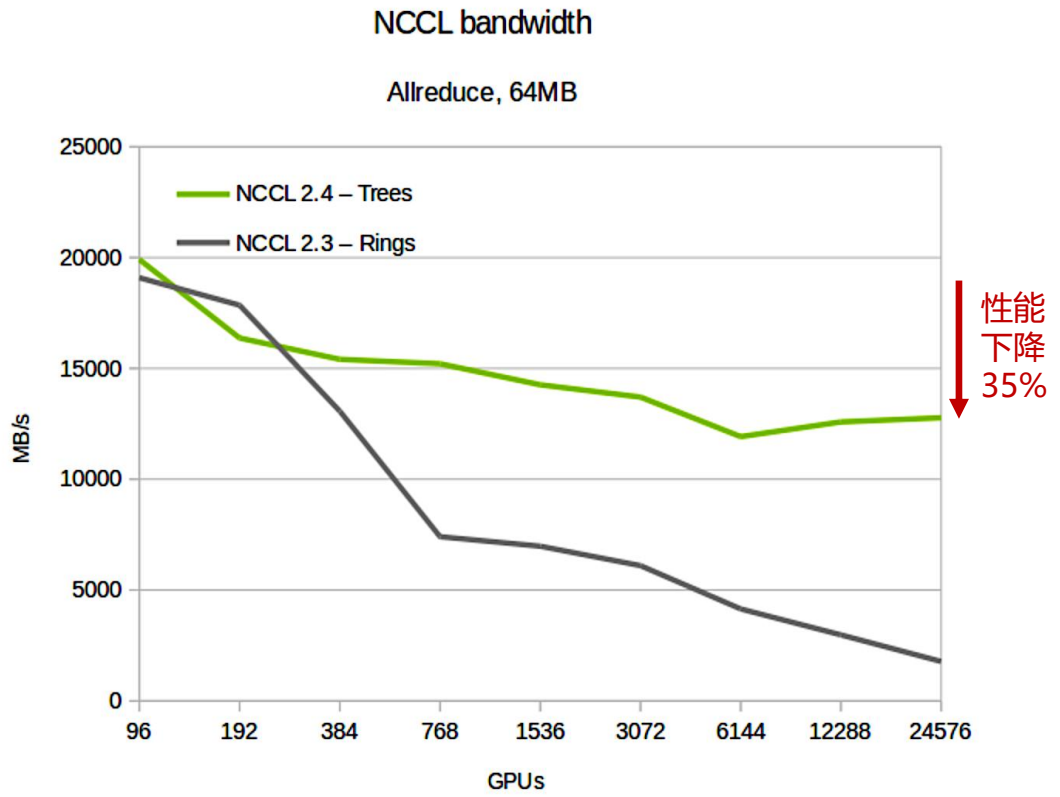
挑战1：网络层级多导致GPU集群通信性能下降

GPU集群规模受限于交换芯片吞吐、设备最大端口数



- **性能**：受多级负载均衡不均、多级网络拥塞等影响，集合通信性能下降
- **成本**：交换机、光模块数量增加
- **运营**：受多级PFC反压影响，网络运营风险增大

大规模组网场景下，性能下降35%



Scale out 网络面临挑战

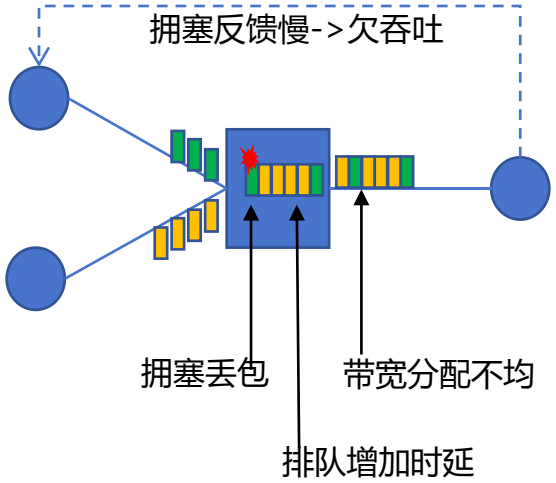
Open AI Infra
A Chapter of GCC

挑战2：网络协议制约GPU运行效率

网络协议

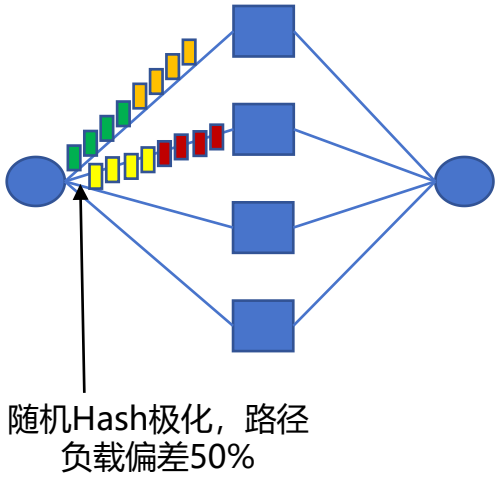
拥塞控制

拥塞：利用率低、延时大、丢包



多路径负载均衡

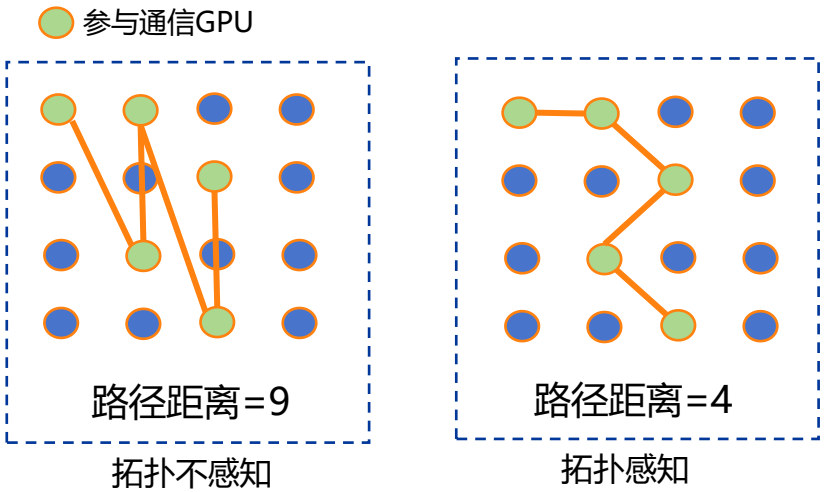
随机选路：极化、利用率低



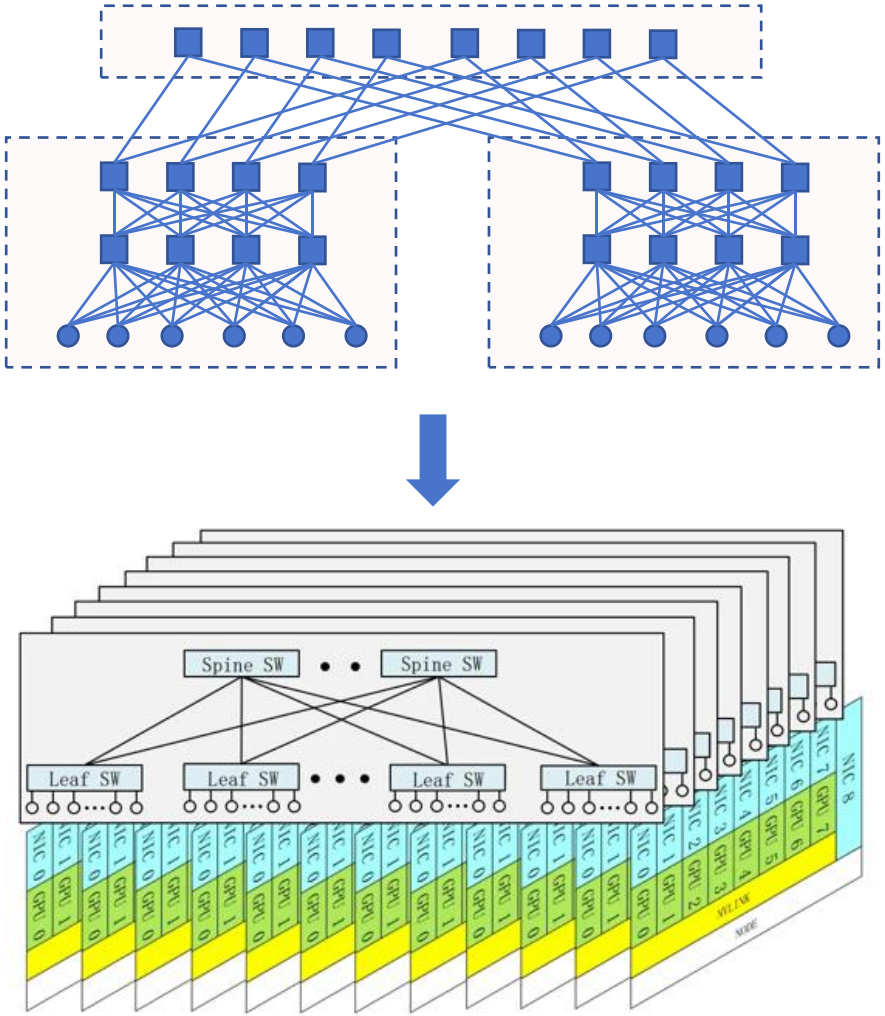
端侧集合通信库

集合通信优化

拓扑不感知：绕路、次优路径



Scale out 网络应对方案



网络拓扑优化

三层到二层：

- 网络拓扑从传统Leaf->Spine->Core 演进到Leaf->Spine大二层拓扑
- GPU to GPU 通信经过交换路径从5跳下降到3跳，时延降低40%
- 接入相同GPU节点数，交换机用量下降40%，光模块用量下降40%
- 交换机芯片需要更高Radix，端口数量加倍，GPU节点数增至4倍

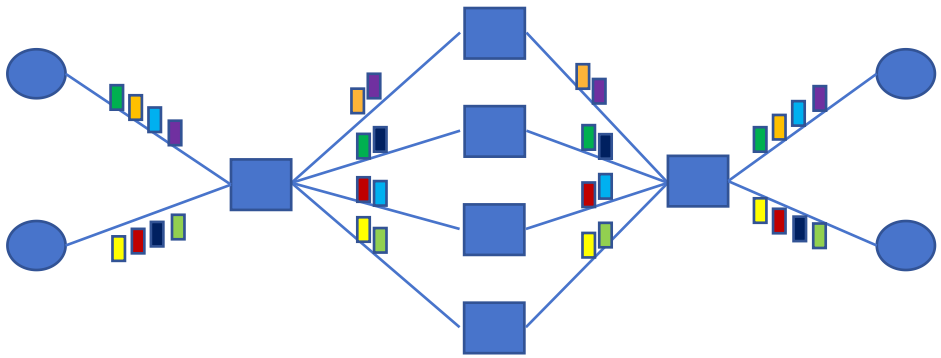
单平面到多平面：

- 网卡应具有多个物理端口，每个端口分别接入单独的网络平面，通过端口绑定，以单个逻辑端口向上层用户呈现
- 从业务视角看，单个或多个队列对 (QP) 喷洒到多个网络端口，通过多个网络平面到达目的的，路径更多、负载更均衡。但来自QP的数据包可能会穿越不同的网络路径，并以无序方式到达接收方，需要接收网卡支持乱序重组，以保证消息一致性并保留正确的排序语义
- 多平面架构通过增加网卡端口数量+降低网卡单端口速率，使用相同带宽的交换机，相比单平面架构可以大大扩展二层网络的节点数量，使得更大规模GPU通过二层网络即可互联

Scale out 网络应对方案

Open AI Infra
A Chapter of GCC

包喷洒、自适应路由

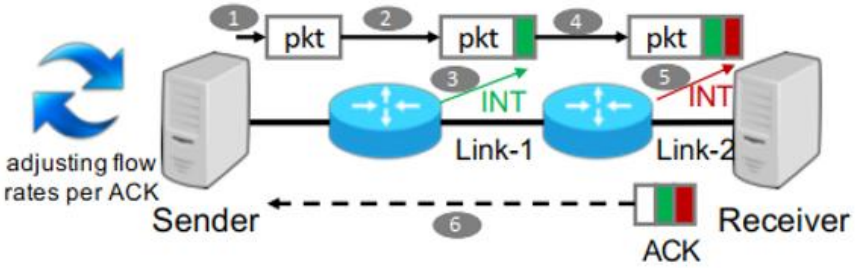


方式一：端侧包喷洒+传统网侧ECMP

方式二：传统端侧+网侧自适应路由

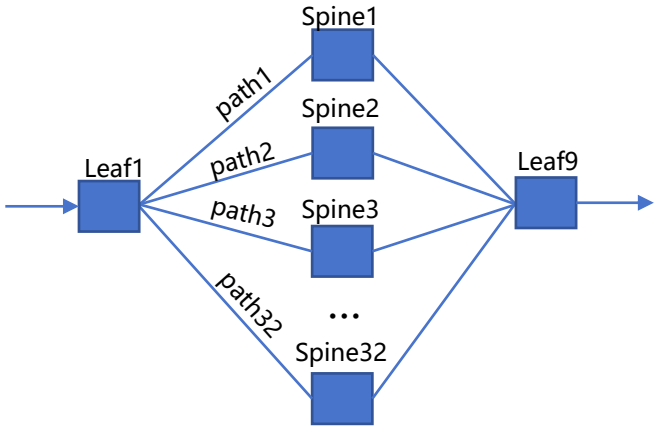
消除Hash不均，提升带宽负载均衡

端网协同拥塞控制



HPCC

- 交换机沿路插入元数据
- 接收端通过ACK返回元数据
- 发送端通过元数据信息调速



确定性路径控制

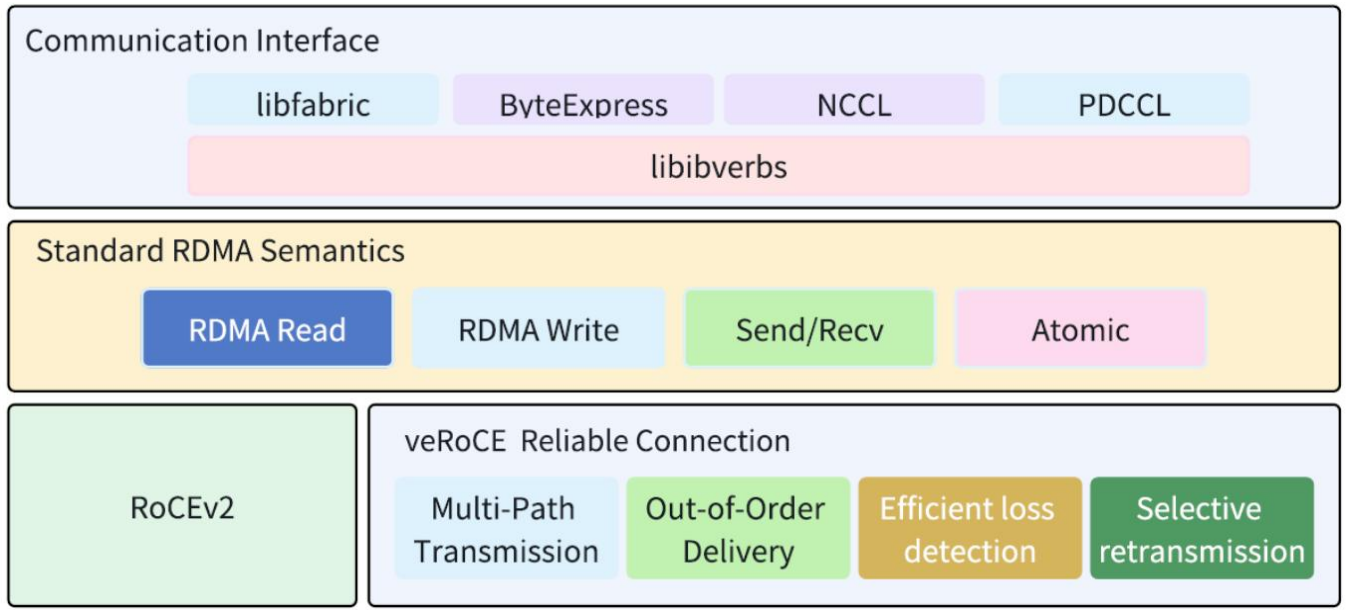
- 发送端报文填充PathID或Sport ID 信息
- 交换机根据五元组、PathID或Sport ID Hash选路
- 拥塞后改变PathID或Sport ID切路

端对端拥塞控制，All2All通信效率提升

字节veRoCE方案

Open AI Infra
A Chapter of GCC

12月18日火山引擎 Force 大会，字节跳动正式发布 veRoCE——高性能自研 RDMA 传输协议



原生支持全语义多路径传输、DDP乱序接收

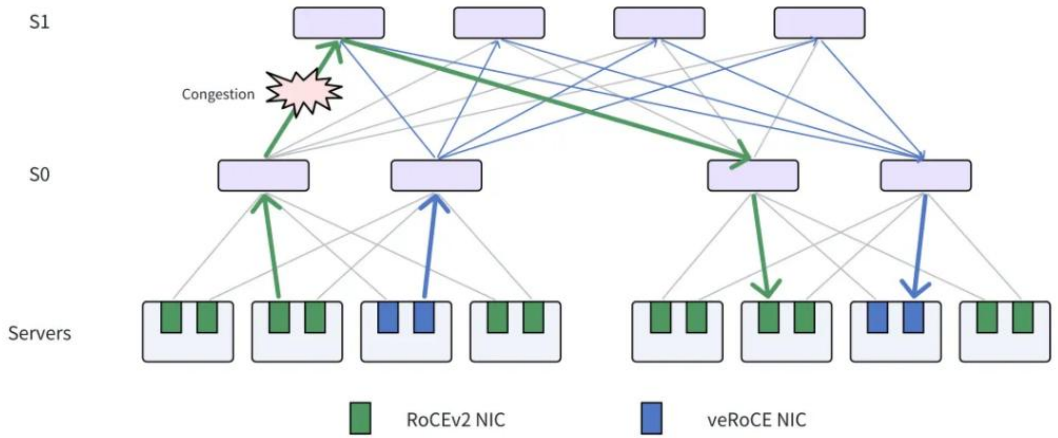
丢包快速检测、快速传递、选择性重传

多路径拥塞控制

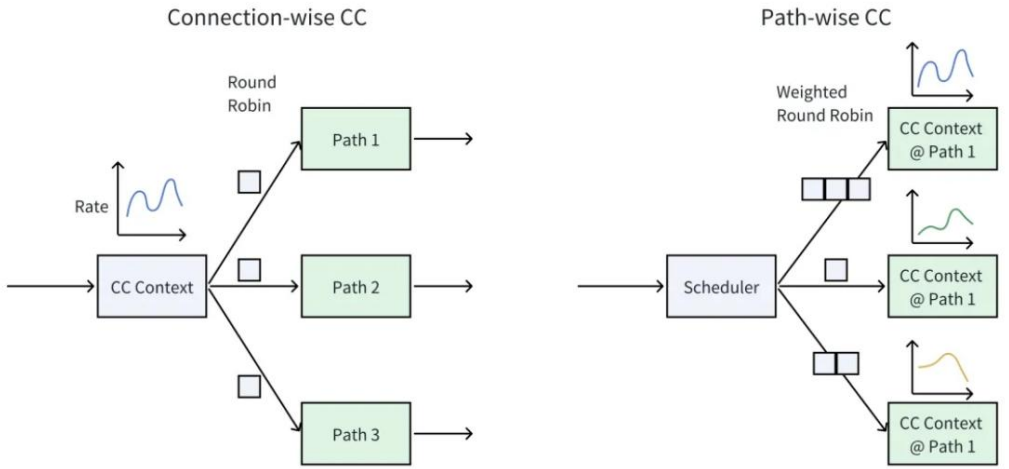
RoCE v2协议兼容互通

veRoCE 关键特性

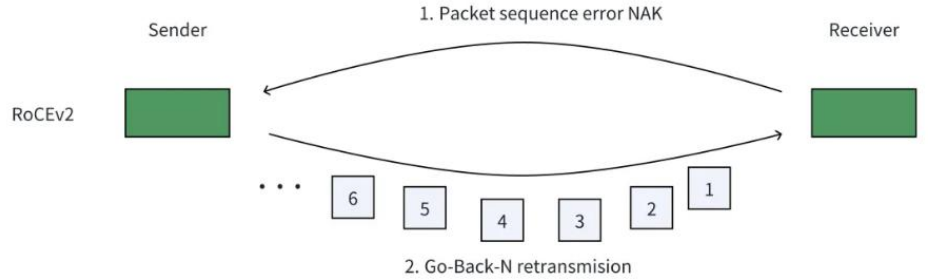
多路径与乱序优化： 源端修改熵值+交换机报文喷洒，DDP乱序重组

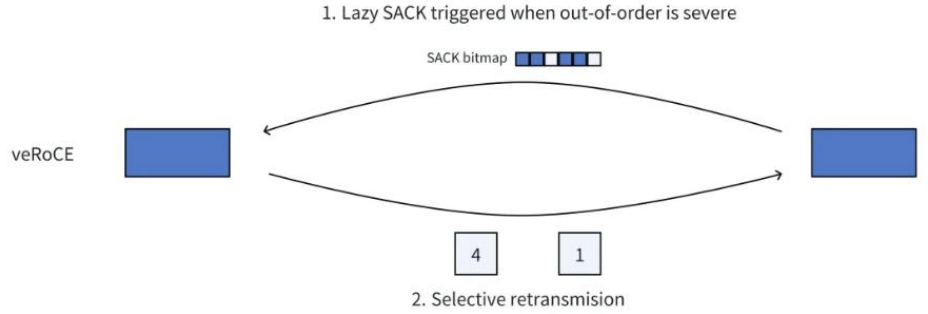


多路径拥塞控制： 支持路径粒度和连接粒度两种拥塞控制模式

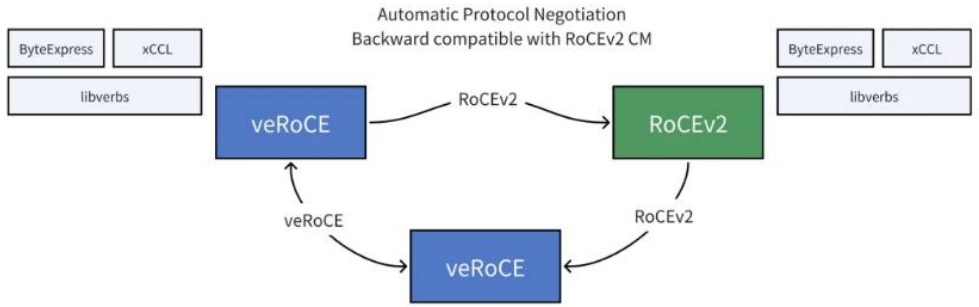


高效重传机制： 支持基于SACK/Lazy SACK的选择性重传策略





兼容性和易用性： 支持通用Verbs接口，兼容RoCE v2，协议自协商



veRoCE 和RoCE v2对比

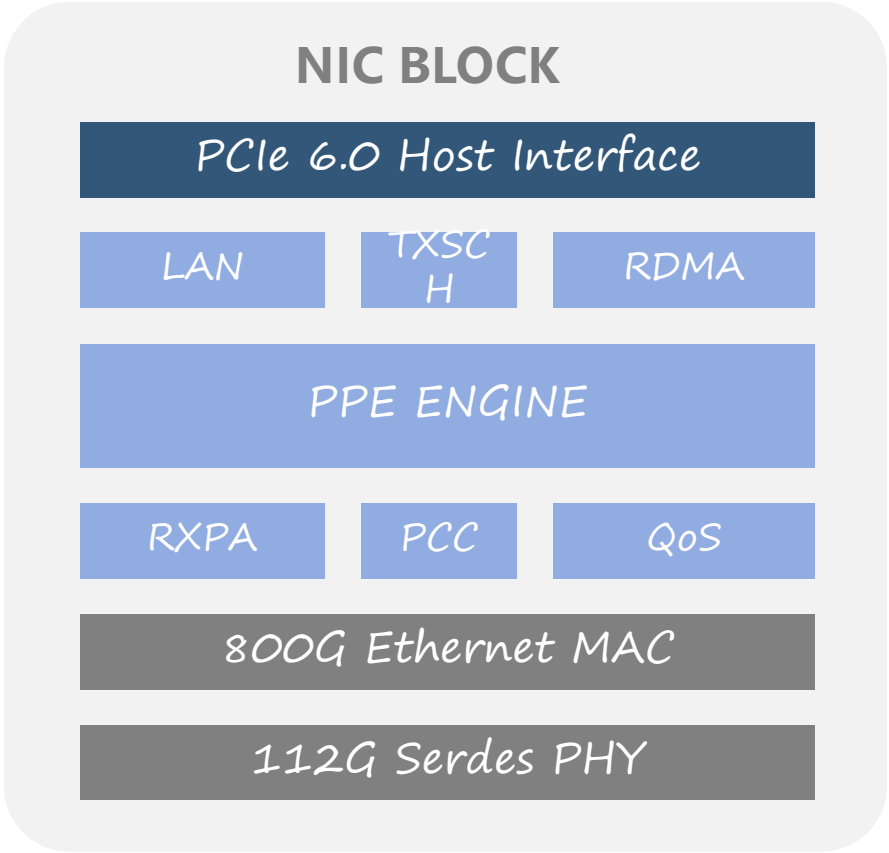
特性	RoCE v2	veRoCE
多路径传输	依赖ECMP，但传输层无显式多路径设计，多路径利用率不足，负载均衡效果不佳	原生支持多路径传输，可通过数据包熵值或交换机AR以报文粒度喷洒到多路径，负载均衡效果好
乱序处理	基于连接的报文顺序到达模型，不支持乱序接收	原生支持乱序交付，每个报文带 DDP 扩展头，Responder 可直接处理乱序包
PSN 空间设计	Read Response 与 Send/Write 共用 PSN，机制不统一	为Read Response分配独立PSN空间，统一各类报文的丢包检测与重传
丢包检测	基于 ACK + 超时机制，丢包定位慢、精度低	引入Lazy SACK+短超时，精确识别已收/丢失报文，丢包检测更快更准
重传机制	通常为 Go-Back-N 式重传，易重传大量已达报文	硬件友好的选择性重传，只重传丢失报文，带宽利用率更高
拥塞控制	ECN+DCQCN机制，调参难度大且无通用性，依赖PFC	RTT+慢路径遥测，全面搜集路径拥塞信息，智能调控；去PFC设计，避免PFC死锁和风暴

下一代AI NIC发展趋势

Open AI Infra
A Chapter of GCC

网络端口	**400G->800G/1.6T，单/双端口->多端口** • 网络带宽从400G逐渐向800G、1.6T演进，带宽密度不断攀升 • 网络端口从单端口、双端口逐渐向多端口演进，配合多平面网络拓扑，实现二层万卡、十万卡集群 • 随着带宽、功耗不断升高，网络光模块从传统DSP方案转向LPO/LRO，最终到NPO/CPO演进
网络协议	**网络多协议->统一标准，从封闭走向开放** • IB、RoCE、UEC、高通量、ETH-X 百花齐放走向统一标准，兼容互通、繁荣生态 • 芯片传统ASIC硬化走向灵活可编程，应对AI网络高速发展变化需求
功能性能	**全功能->功能聚焦、简化，全场景->分布式计算互联场景，性能极致追求** • 应用场景相对固定，业务流量确定性较强，删除不需要特性，优化业务痛点特性，做精做强 • 多路径、拥塞控制加强，虚拟化、云卸载特性减弱，把芯片面积留给业务痛点特性，发挥更大价值 • 高带宽、低时延极致追求，让高速互联不成为瓶颈，让xPU算力利用到极致
产品形态	**PCIe 标卡/OCP标卡->主板集成/板载->XPU IO Die/Chip** • 当前以标准PCIe HHHL形态为主，随着超节点集成度、功耗的不断提高，网卡可能会以板载形式存在 • Scale up 和Scale out网络未来可能走向融合，AI NIC传统网卡形态可能消失，将以XPU IO Die形式存在

灵达800G AI NIC



核心竞争力

集成PCIe Gen6 Switch 48 Lane
适配超节点Compute Tray下一代架构

支持1x800G、2*400G、4*200G
Scale out多平面

RoCE v2、VeRoCE协议卸载
硬件级逐包喷洒、乱序重组、选择性重传、可编程CC

高带宽、低时延
单QP 800G线速，1us 级时延

灵达IO互联全栈产品布局

Open AI Infra
A Chapter of GCC

国产唯一SAS/SATA/NVMe 全场景企业存储扩展解决方案

RAID/HBA卡

SAS EXP芯片

芯片、软件、硬件全栈自研，性能对标行业标杆产品

- Tri-Mode HBA/RAID控制芯片支持 RAID0/1/5/6 JBOD等多种模式，独创8+2/16+2下联拓扑架构，打破存储边界，百变存储一卡掌控
- SAS Expander扩展芯片可支持SAS/SATA硬盘类型，接口速率高达12Gb/s，助力24/36/60及更高密度存储扩展

10G到800G全系列高速网卡

板卡

芯片

芯片、软件、硬件全栈自研，契合通用计算、云计算与智能计算的高带宽、低延迟的ScaleOut组网方案

- 双口万兆/25G网卡性能媲美行业标杆，超低功耗设计，助力企业绿色算力战略落地
- 400G/800G AI NIC 助力客户构建千卡、万卡、十万卡超智融合集群

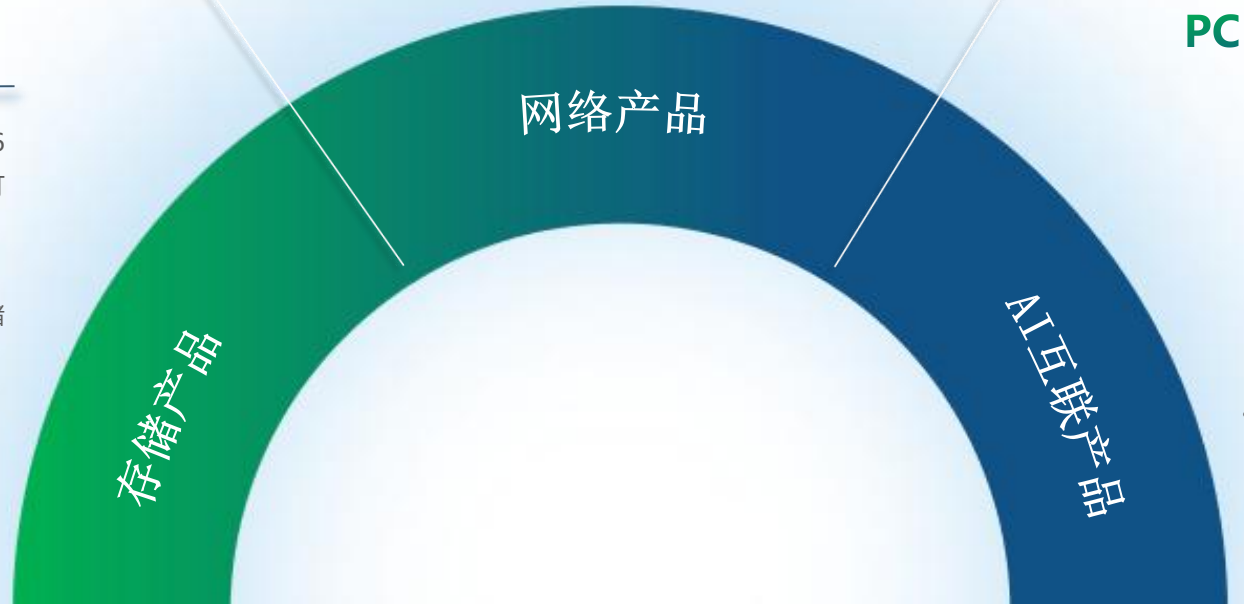
高速扩展，行业客户实测认可，国产唯一批量供货

PCIe SW芯片

ScaleUp SW芯片

高性能、低延迟、可扩展、经济高效的PCIe扩展与定制ScaleUp互联方案

- PCIe4.0 Switch支持48/104 Lane 商用、工业级别，在图站、NVMe闪存扩展场景广泛应用
- PCIe5.0 Switch支持AI服务器全场景GPU互联拓扑，已导入国内主流厂商数十款机型和头部互联网客户
- 112G/224G 高速网络互联芯片将成为国内智算数据中心高速接口不可或缺的关键组件



开放合作，繁荣生态

开放合作，繁荣生态

开放合作，繁荣生态