浪潮信息

X400超级AI以太网 释放AI算力潜能

浪潮信息网络研发部系统架构师 郭巍松

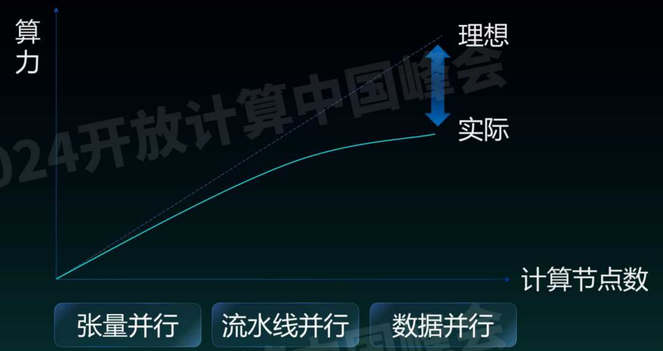
大模型应用爆发，亟需新一代高性能网络

浪潮信息

大模型训练需要高性能网络

大集群≠大算力

大规模　高带宽低时延　无阻塞零丢包



大模型训推融合迈向云化

新一代高性能网络的核心需求

超大规模：
数十万卡到百万卡

超高性能：
400G→1.6T互联
高带宽低时延

多租户：
性能隔离
多任务并行无干扰

开放：
开放操作系统
开放网络协议

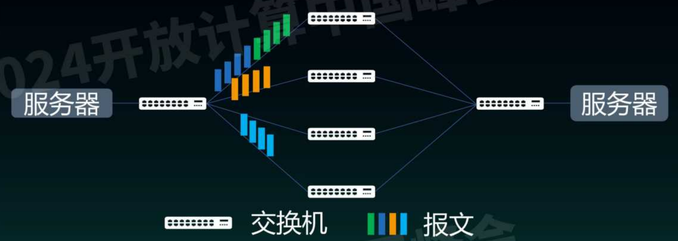
传统RoCE难应对AI独特的工作负载

浪潮信息

负载不均，抑制算力释放

ECMP逐流不均

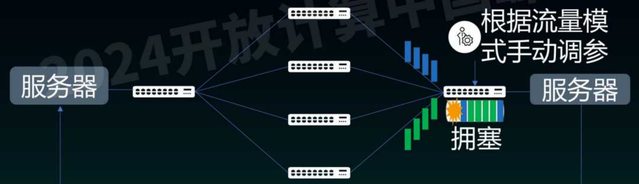
带宽利用率低，尾延时高，FCT时间大



部署缓慢，阻碍业务上线

拥塞控制部署复杂

DCQCN参数需要针对性调优

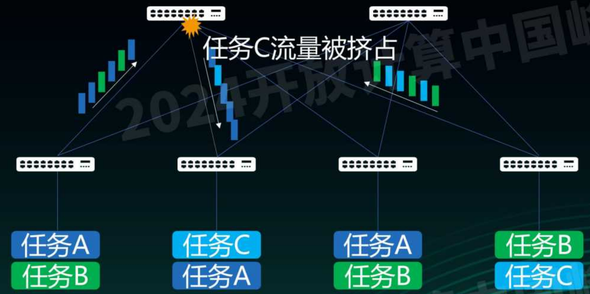


拥塞反馈，从源端调速

多任务混跑，抢占网络资源

多租户性能下降

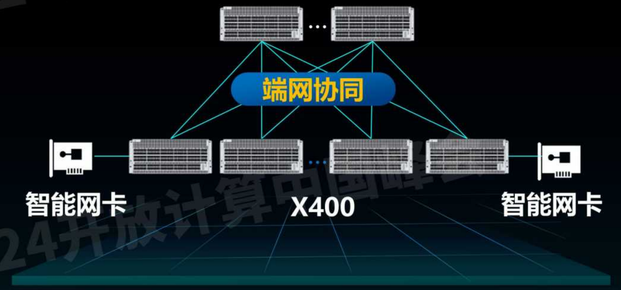
多租户多任务训练，互相影响



端网协同突破以太网性能极限 充分释放算力潜能

浪潮信息

X400 AI Fabric方案



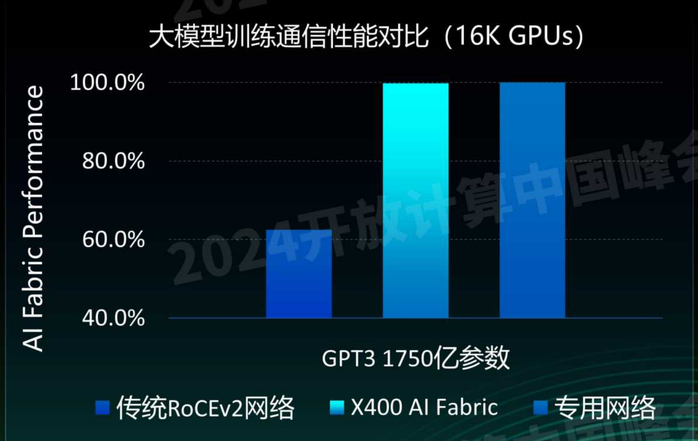
万卡集群
支持**十万卡级**

传输0丢包
端到端**400G**

有效带宽
高达**95%**

超级AI以太=X400+智能网卡

性能领先传统RoCE 1.6倍



AR+保序服务构建无阻塞通道 使能算力线性增长

浪潮信息

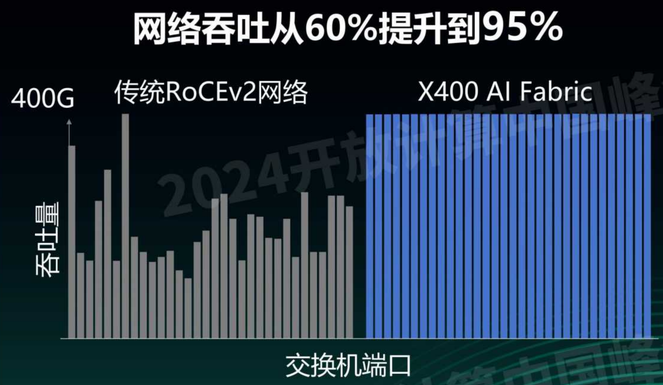
基于包的负载均衡

基于队列深度和链路负载逐包喷洒

业务无感知的保序服务

基于DDP（Direct Data Placement）技术的报文保序

有效带宽率高达95%



AUTO ECN缩短30%FCT 加速AI大模型训练

浪潮信息

基于GNN的模型训练

状态信息传递 = 消息聚合 + 状态更新

AI助力，ECN自动调参

采集多时刻状态信息

$Port\_utilization(t_0 \sim t_n)$

$Buffer\_occupancy(t_0 \sim t_n)$

$ECN\_marking\_rate(t_0 \sim t_n)$

输入神经网络

更新参数

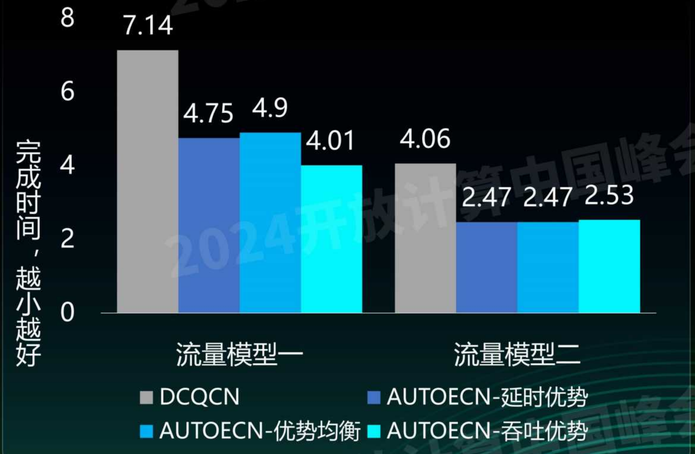
$K_{min}(t_n)$

$K_{max}(t_n)$

$P_{max}(t_n)$

自动调优 模型泛化 精准控速

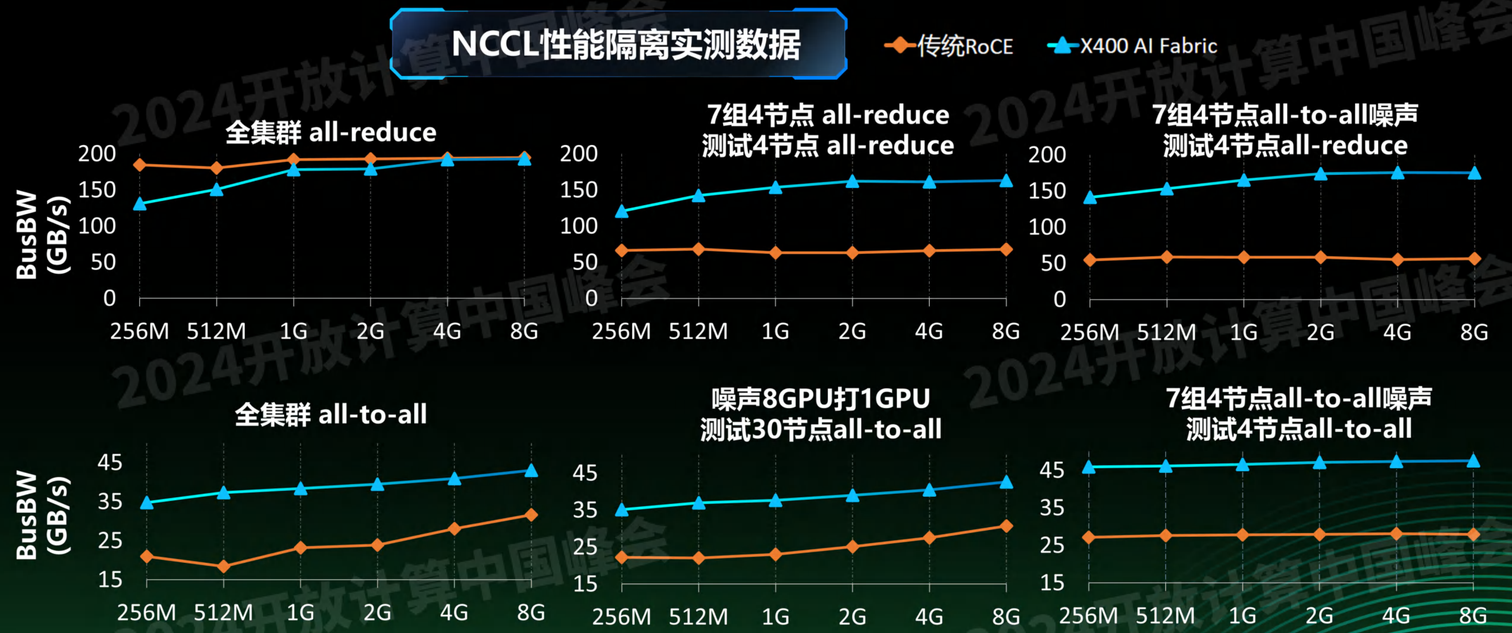
FCT缩短30%



NCCL性能隔离实测，X400 AI Fabric远超传统以太

NCCL性能测试拓扑

NCCL性能隔离实测数据



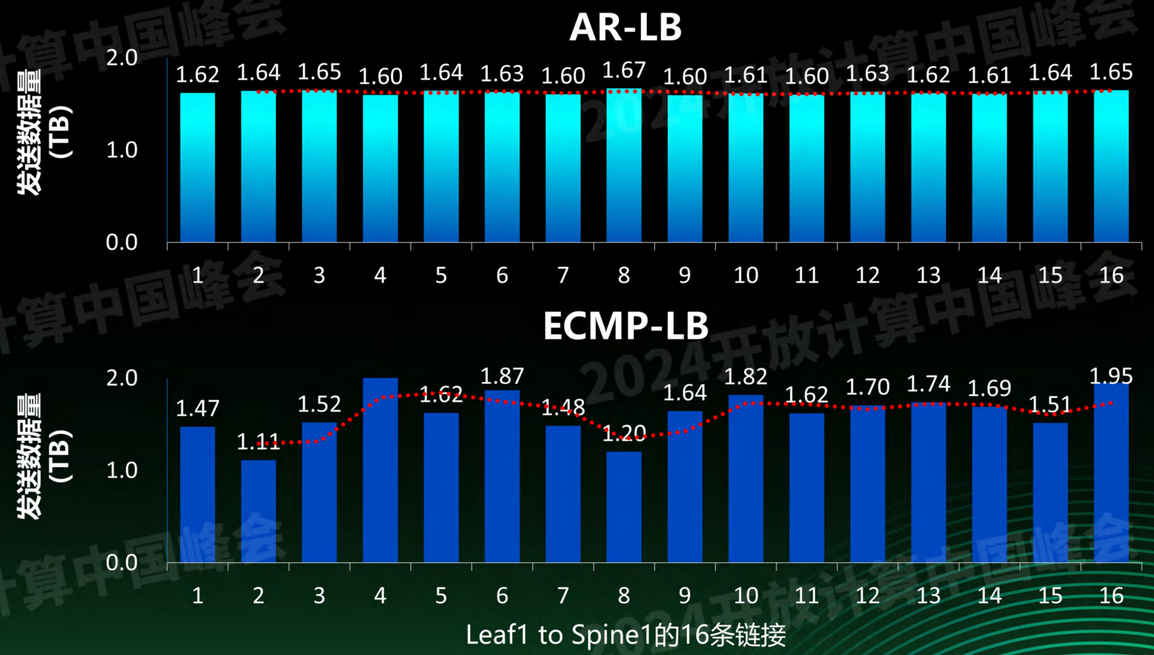
测试结果分析---负载不均衡率对比

全集群all-to-all测试下的负载不均衡率=
链路组发送数据量标准差/链路组发送数据量均值

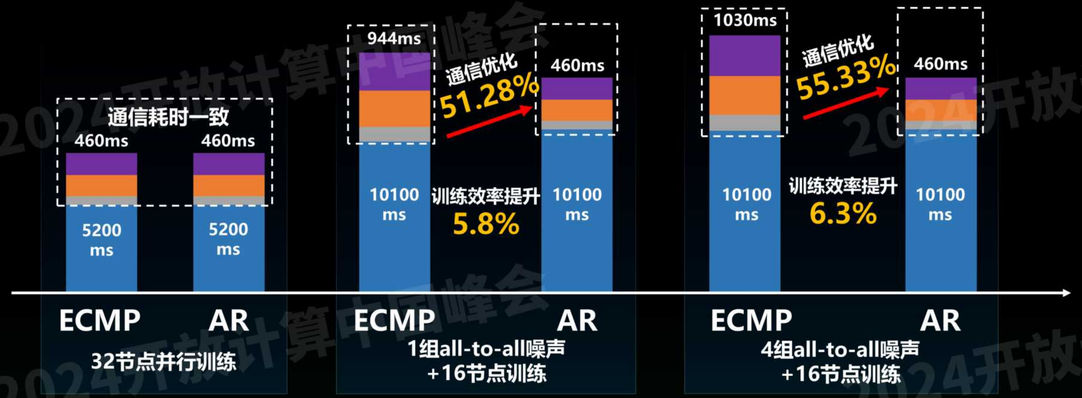
配置方案	Leaf1→Spine	Leaf2→Spine	Leaf3→Spine	Leaf4→Spine
ECMP	16.80%	14.57%	21.95%	19.63%
AR	0.66%	0.64%	0.58%	0.80%

配置方案	Spine1→Leaf1	Spine1→Leaf2	Spine1→Leaf3	Spine1→Leaf4
ECMP	17.10%	18.59%	21.15%	16.93%
AR	0.89%	0.62%	0.56%	0.66%

配置方案	Spine2→Leaf1	Spine2→Leaf2	Spine2→Leaf3	Spine2→Leaf4
ECMP	18.98%	21.51%	17.03%	12.64%
AR	0.68%	0.75%	0.65%	0.85%



大模型训练实测

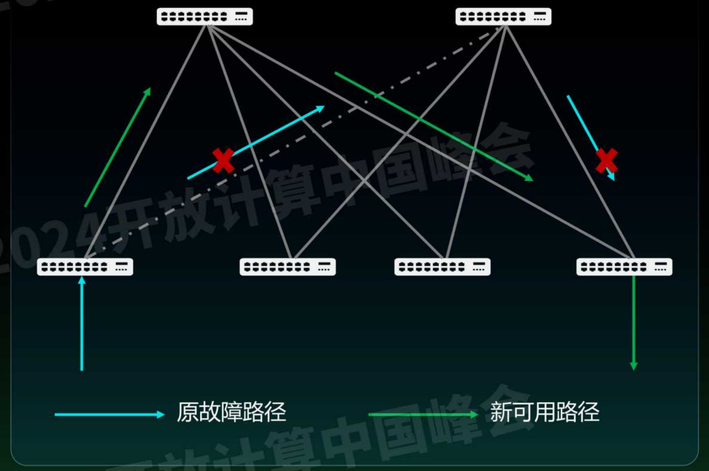


训练场景	网卡：DCQCN+PFC 交换机：ECMP+ECN+PFC	网卡：RTT+PFC 交换机：AR+ECN+PFC	性能提升
32 节点测试	22min10s	22min10s	-
1组all-to-all噪声 + 16节点测试	45min11s	42min35s	5.75%
4组all-to-all噪声 + 16节点测试	45min2s	42min12s	6.3%

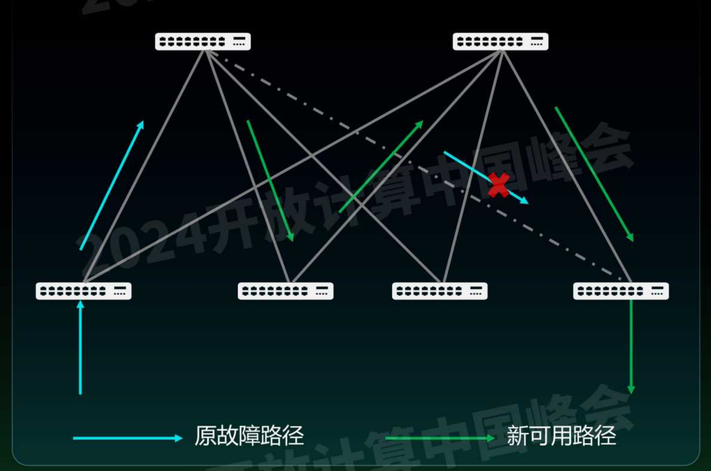
多重可靠性技术 保障AI训练业务持续可用

浪潮信息

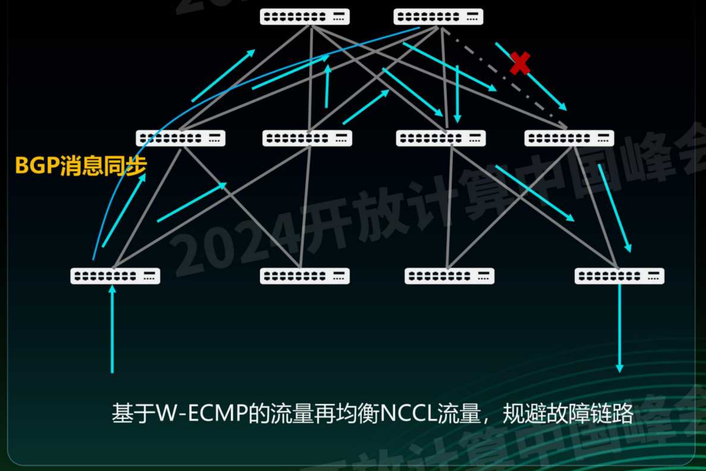
本地上行链路故障自动切换



本地下行链路故障自动切换

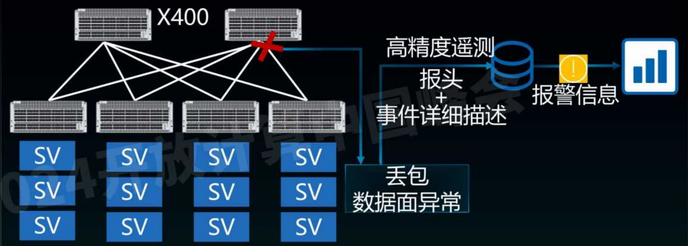


远端链路故障负载再均衡



全面可视，掌控AI节点间通信状态

高精度遥测技术



芯片级、系统级、链路级监控

- 底层链路
- Congestion
- L2/L3/Overlay
- Latency
- Buffer
- RoCE
- ACLs
- Load Balance

高精度遥测核心信息

价值体现

保障业务连续性

整网业务数据可视保障了关键业务可用性

提升运维效率

基于因果树的大数据分析可以快速找到故障根因

网络维护计划

提前预测故障模块，从而制定网络维护计划
减少对业务影响

UXOS赋能开放智能的AI网络

基于SONiC构建自研UXOS

开放智能可编程加速业务创新

- 敏捷运维
- 开放标准接口
- 网络特性灵活定制
- 云原生，容器化架构
- 可编程数据平面

拥抱开放回馈社区

稳定优化	25
组件改善	15
简化运维	35
特性贡献	1

202311分支发布说明中专门感谢了浪潮信息团队贡献

浪潮信息

浪潮信息提供全系列1G~400G交换机

400GE

X400（128*400G）

SC8661SL(32*400G，满配400G接口卡)

SC7650EL-32D(32*400G)

200GE

SC7650EL-24H8D(24*200G，8*400G)

100GE

SC8661SL(128*100G，满配100G接口卡)

SC8650EL-128C(128*100G)

SC6630EL-32C(32*100G)

SC6650EL-48L8D（48*100G，8*400G)

25GE

SC5631EL-48Y8C(48*25G,8*100G)

CN5610EL-48Y8C(48*25G,8*100G)

1GE

CN2610EL-48T4X2Q(48*1G电,4*10G，2*40G)

CN2610EA-48S4X (48*1G光,4*10G)

CN2610EA-48T4X(48*1G电,4*10G)

浪潮信息

THANKS