



中国移动
China Mobile



移动云

开放解构超节点： 筑牢可持续发展的AI算力基石

徐小虎
中国移动云公司

Contents

01 传统超节点架构面临的挑战

02 开放解构超节点系统架构

03 超节点业界发展动态

04 总结

网络协议私有封闭



硬件架构紧耦合



功耗散热挑战大



系统规模扩展受限



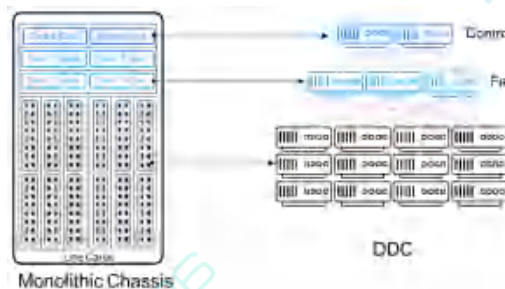
互联可维护性差



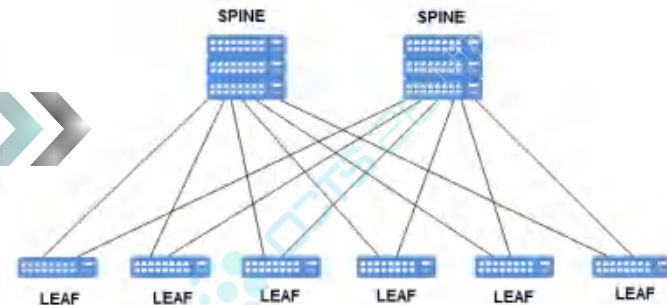
框式路由器/交换机 集群路由器/交换机



分布式解构机框



开放CLOS架构



刀片服务器



整机柜服务器



机架式服务器



Contents

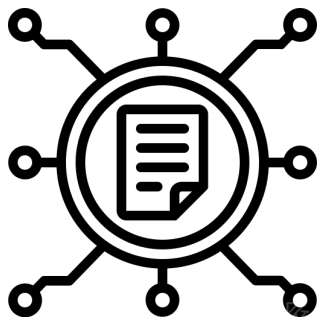
01 传统超节点架构面临的挑战

02 开放解构超节点系统架构

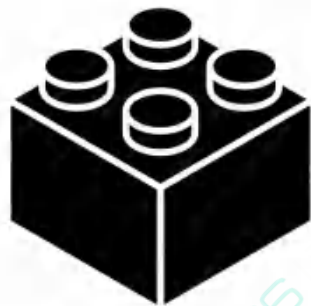
03 超节点业界发展动态

04 总结

网络协议完全开放



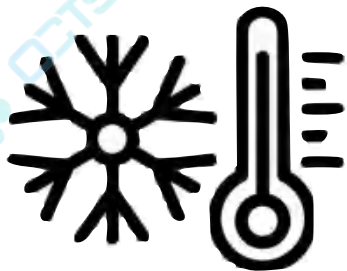
硬件架构全面解耦



系统规模弹性扩展



散热方案灵活选择



互联效率全面优化





4U

计算节点
(含4x GPU)

16+

800G OSFP
/计算节点

3.2T+

Scale-up带宽
/GPU

灵活的连接

光(FRO, LPO, LRO),
铜(AEC, DAC)

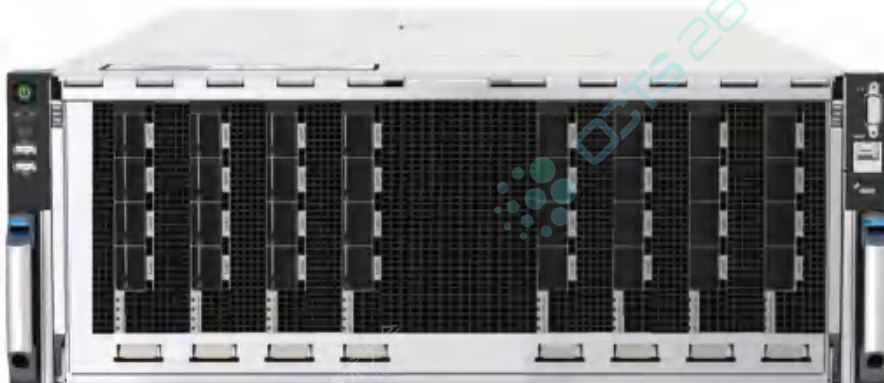
简化的设计

风冷
标准PSU供电

开放的协议

支持标准以太网和FARE-in-
SUN*协议（规划中）

*FARE-in-SUN: Fully Adaptive Routing Ethernet in Scale-Up Networks
IETF标准草案: draft-xu-rtgwg-fare-in-sun



4U

计算节点
(含8x GPU)

32

800G OSFP
/计算节点

3.2T

Scale-up带宽
/GPU

灵活的连接

光(FRO, LPO, LRO),
铜(AEC, DAC)

便捷的设计

液冷盲插
供电盲插

开放的协议

支持标准以太网和FARE-in-
SUN*协议（规划中）

*FARE-in-SUN: Fully Adaptive Routing Ethernet in Scale-Up Networks
IETF标准草案: draft-xu-rtgwg-fare-in-sun



4U

以太网交换机

128

400G QSFP112
或800G OSFP

一机多用

适用于Scale-up、Scale-out以及前端网络

灵活的连接

光(FRO, LPO, LRO)
铜(AEC, DAC)

简化的设计

风冷
标准PSU供电

开放的协议

支持标准RoCE和FARE*
协议

*FARE: Fully Adaptive Routing Ethernet (全自适应路由以太网)
IETF标准草案: draft-xu-idr-fare、draft-xu-lsr-fare

64卡（风冷）



128卡（液冷）



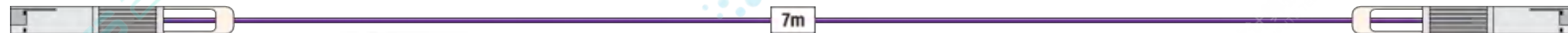
128卡（风冷）/256卡（液冷）



400G DAC



400G AEC



功耗降低~50%



成本降低~50%

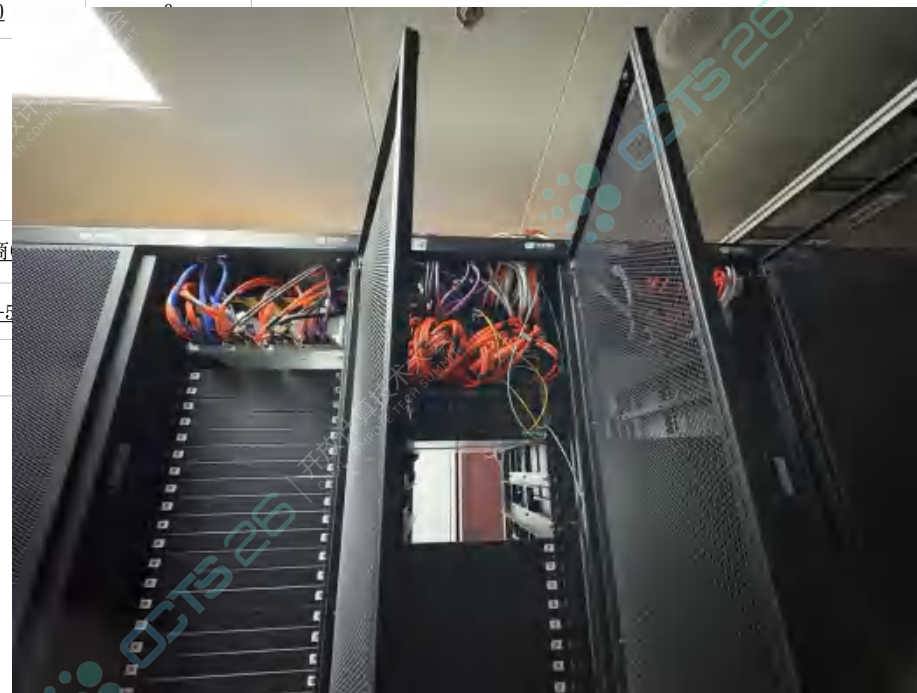


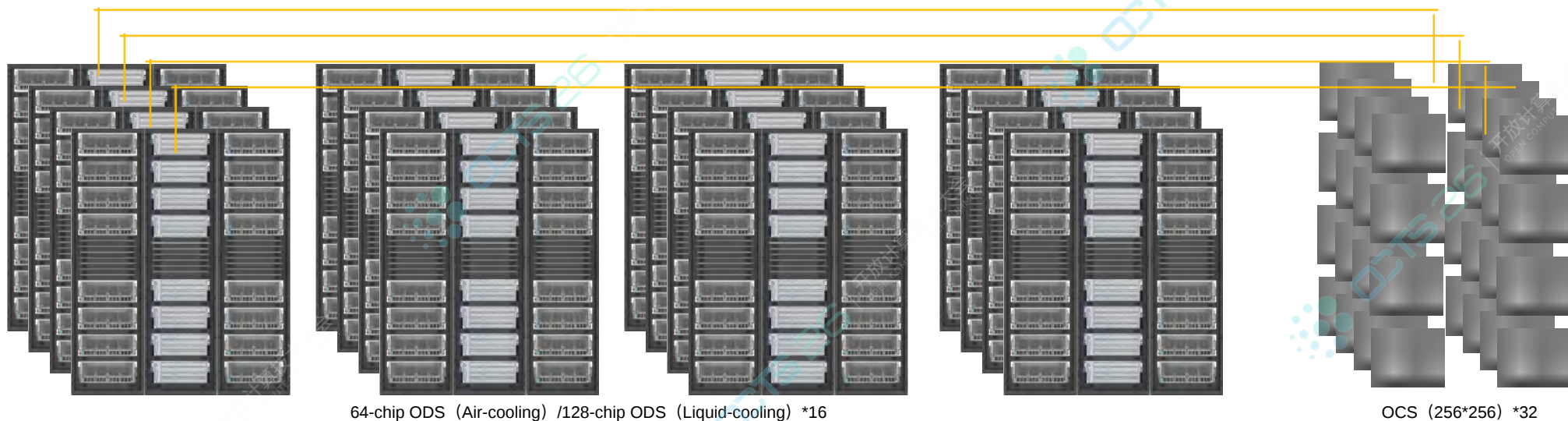
可靠性提高两个数量级

3M 800G AEC	厂商B	厂商D	5米 800G AEC	厂商A	厂商B	厂商C
pre_FEC_BER	5e-8	N/A	pre_FEC_BER	1e-10	5e-10	9e-11
post_FEC_BER	0	0	post_FEC_BER	0	0	0
7M 800G AEC	厂商A	厂商B	9米 800G AEC	厂商A		
pre_FEC_BER	5e-10	1e-10	pre_FEC_BER	5e-8		
post_FEC_BER	0	0	post_FEC_BER	0		
2M 800G DAC	厂商A		4米 800G ACC	厂商A	厂商B	厂商C
pre_FEC_BER	7e-7		pre_FEC_BER	1e-6	3e-5	3e-5
post_FEC_BER	0		post_FEC_BER	0	0	0

测试结论：

AEC、DAC、ACC均符合SI要求。





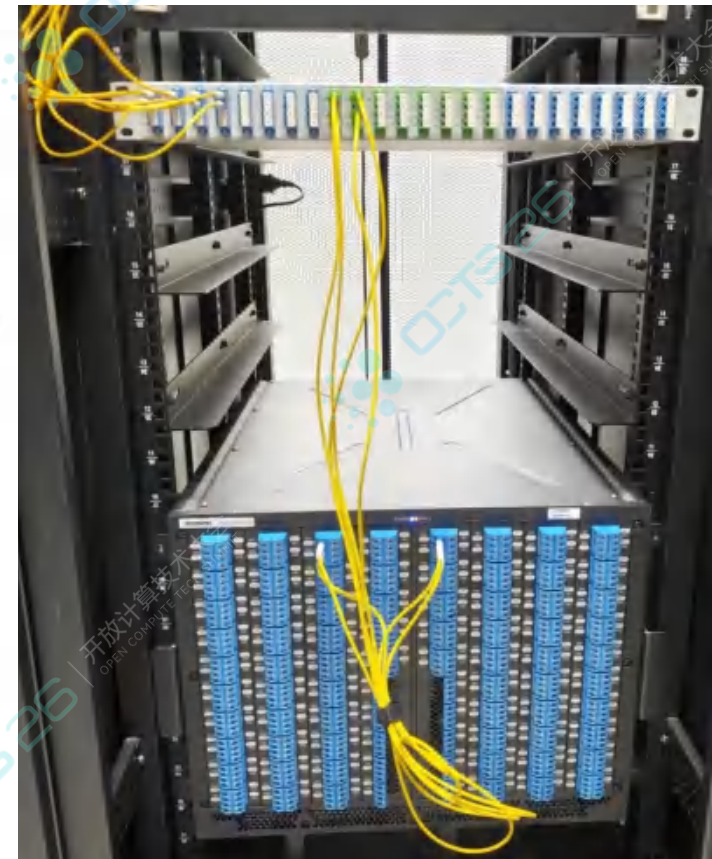
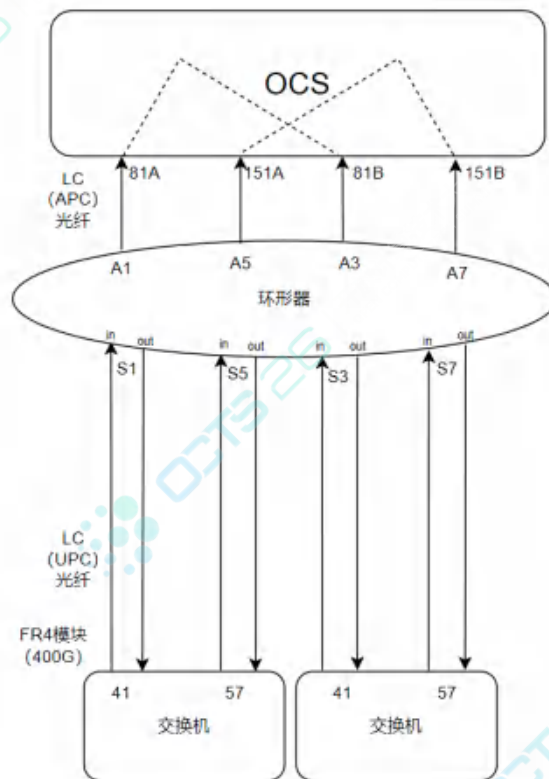
功耗降低~90%

成本降低~60%

99.999%
互联更可靠



300*300 OCS		厂商B
插入损耗		0.94~1.3dB
重复切换误差		0.02dB
FR4光模块		厂商A
pre_FEC_BER		2.5e-9
post_FEC_BER		0



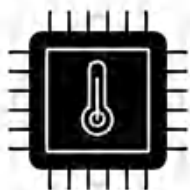
测试结论:

- 1、OCS满足插损 $\leq 3\text{dB}$ 的指标要求;
- 2、OCS满足重复切换误差 $\leq 0.2\text{dB}$ 的指标要求。

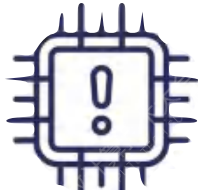
CPO



难维护



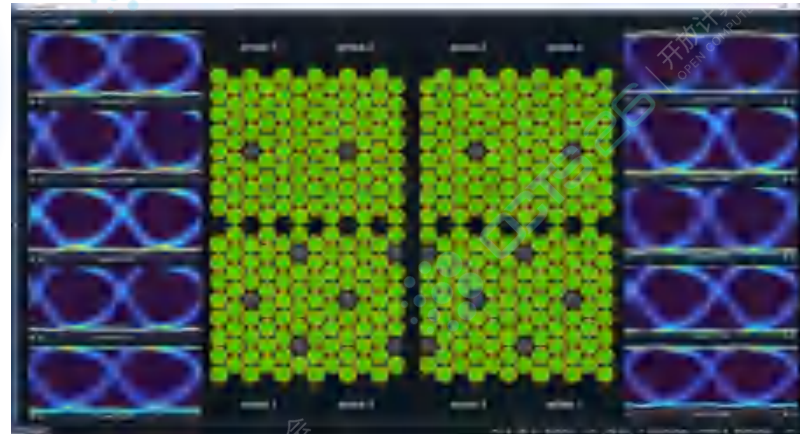
难散热



低良率

VS

MicroLED



易维护



低功耗



高可靠



Contents

01 传统超节点架构面临的挑战

02 开放解构超节点系统架构

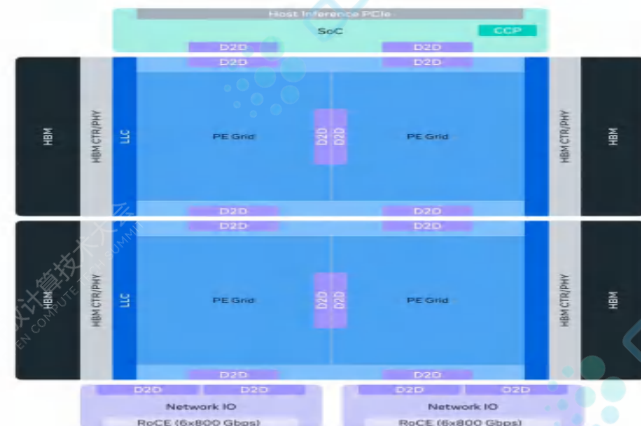
03 超节点业界发展动态

04 总结

微软Maia200



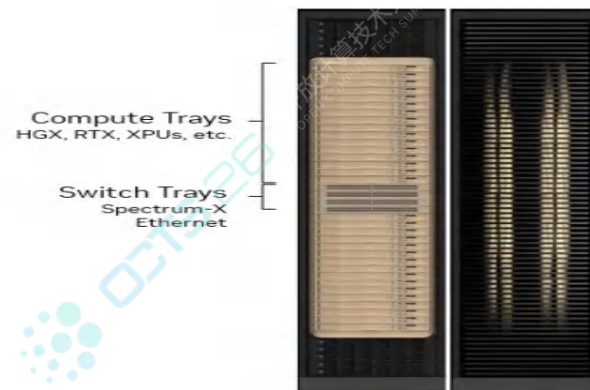
Meta MTIA200/300/400



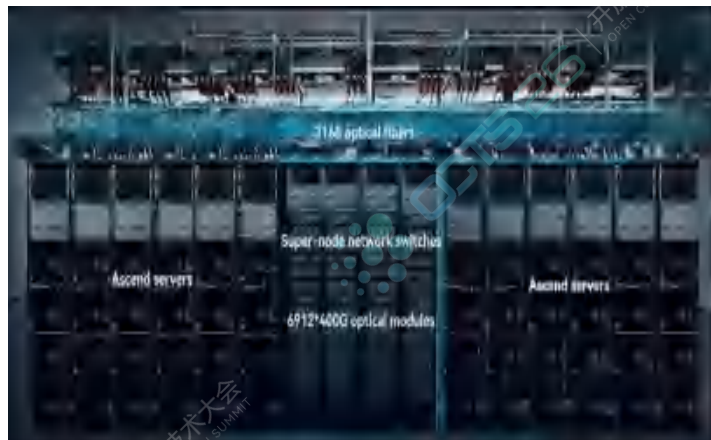
AMD MI455 Helios (UALoE)



Nvidia MGX ETL (Spectrum-X)



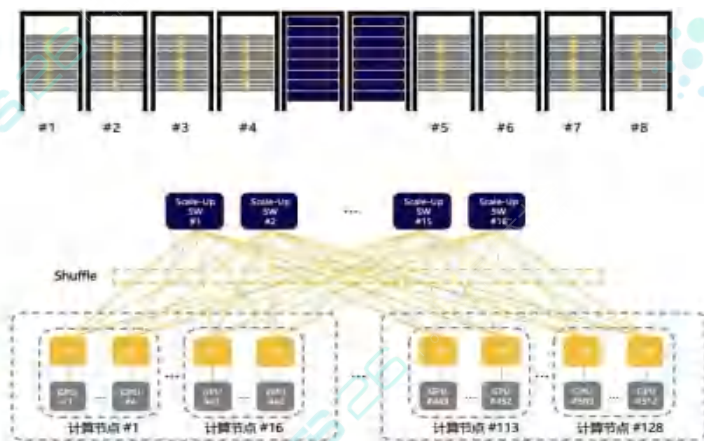
华为CM384



AMD MI500 UAL256



腾讯Eth-X Ultra



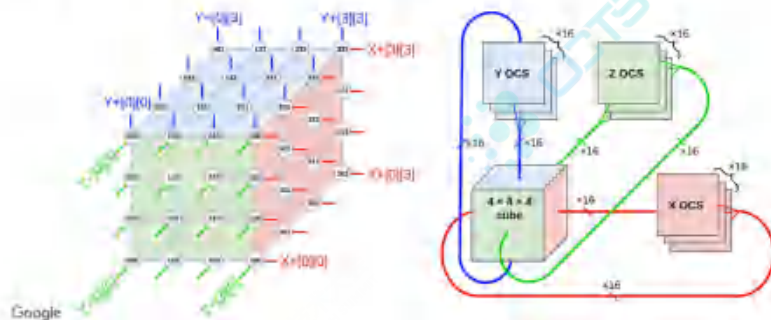
阿里UPN512



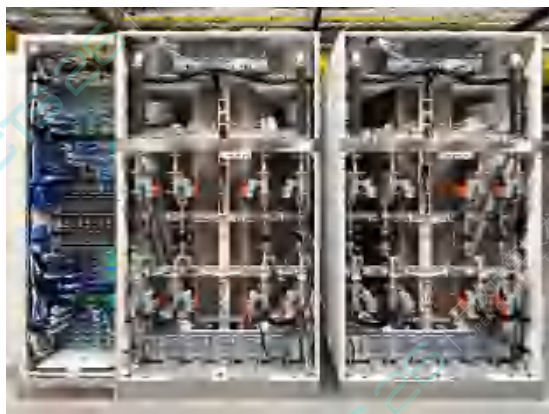
谷歌TPUv7 Ironwood (9216卡)

Racks Are Connected With Optical Circuit Switches (OCS)

- Different ranks of OCS connect different dimensions and indices



微软Maia200超节点(6144卡)



英伟达NVL576



英伟达NVL1152





中国移动
China Mobile



移动云

Contents

01 传统超节点架构面临的挑战

02 开放解构超节点系统架构

03 超节点业界发展动态

04 总结

• 牵头单位

- 中国移动云公司



• 参与单位（排名不分先后）

- 英特尔、博通、云尖信息、锐捷网络、浪潮信息、木犀智能、默升科技、立讯技术、燧原、壁仞、海光、中科曙光、盛科、光迅、万德漆、兆龙、摩尔线程、天数智芯、昆仑芯、烽火



AI Infra
旧范式

SU私有协议栈

封闭超节点架构

SO/SU割裂技术栈

建设运维成本高

AI Infra
新范式

SU开放协议栈

开放解构超节点架构

SO/SU统一技术栈

建设运维成本低



中国移动
China Mobile



移动云

Thanks!



中国移动
China Mobile



移动云

Thanks!



中国移动
China Mobile



移动云

Thanks!