AI大模型的技术挑战和解决方案

Dec. 2023

LUXSHARE TECH
立讯技术

AI--创新涌现，拥抱智能时代

人工智能的“涌现”时刻即将出现，人类社会也将迎来一个波澜壮阔的智能时代。 迈入智能时代，

◆ 最大的需求是**算力**

◆ 最关键的基础设施是**数据中心**。

根据华为《智能世界 2030》报告预测，2030 年，人类将迎来 YB 数据时代，对比 2020 年，

通用算力将增长 **10 倍**

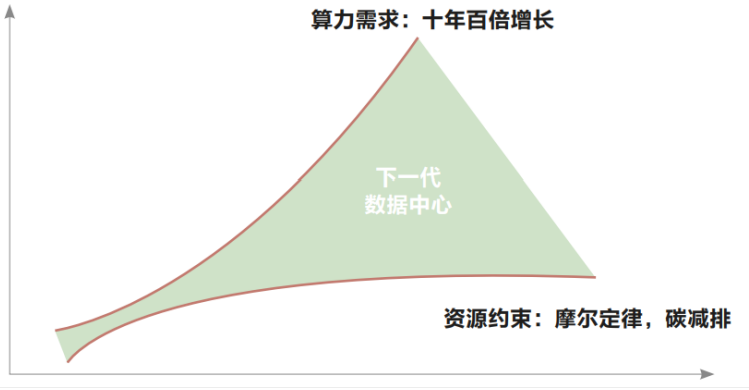
人工智能算力 增长 **500 倍**

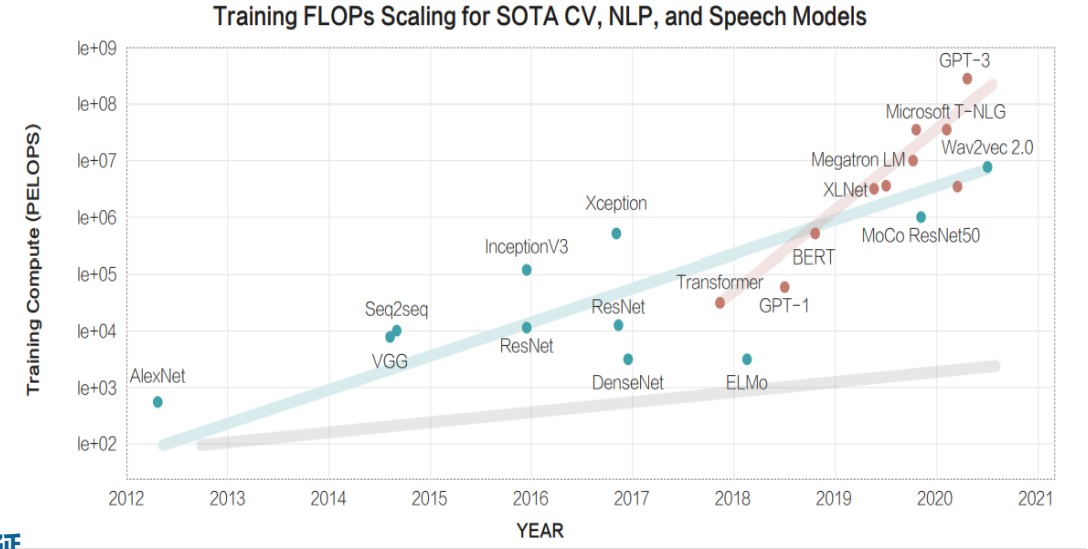
算力需求十年百倍的增长将成为常态

在 AI 大模型训练过程中，当模型大到一定规模之后，性能会发生突变，开始呈现指数级快速增长，科学界称这个现象为“涌现”。

数据中心的发展趋势及挑战

一边是算力需求以远超摩尔定律的陡峭增长，而另一边却是多重的资源约束。单芯片摩尔定律的失效、 以及全球可持续发展目标下对于碳减排的要求，将迫使未来的数据中心必须在更优的计算架构、以及更低的能耗下产生更大的算力。



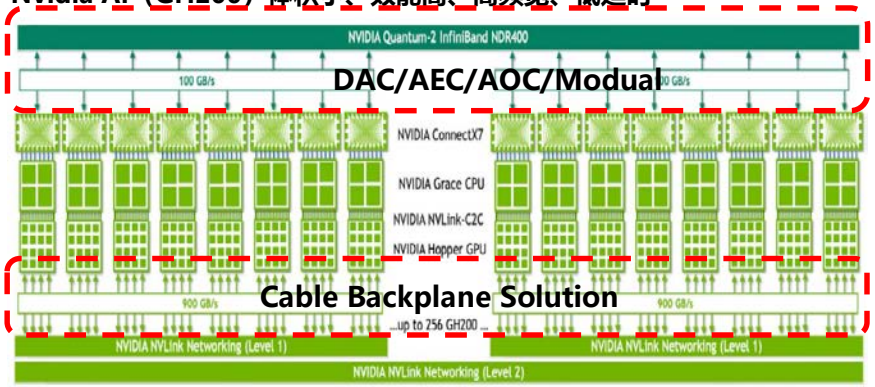


数据中心的发展趋势及挑战

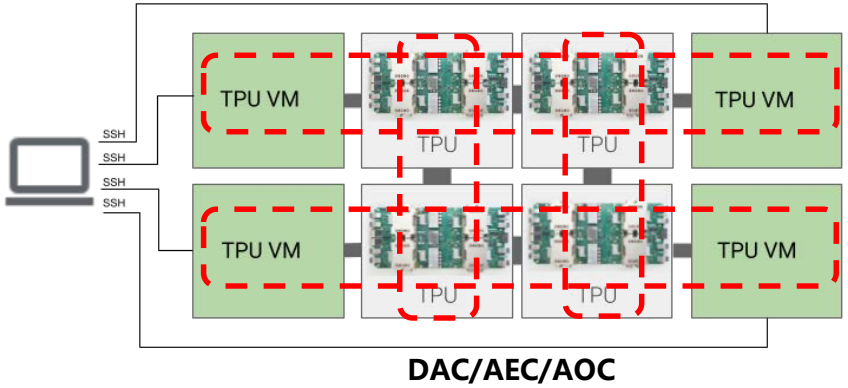
- **绿色化-------极致PUE-----液冷技术**
- **智能化-------AI 大模型计算与推理------高带宽&低时延**
- **大型化+集群化--------统一的互连协议-------整机柜交付+池化**

AI带来数据中心架构的革新： Nvlink,CXL,UEC超级以太网,各类新型协议带来互连架构的革新

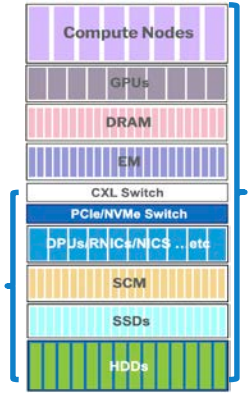
Nvidia AI（GH200）体积小、效能高、高频宽、低延时

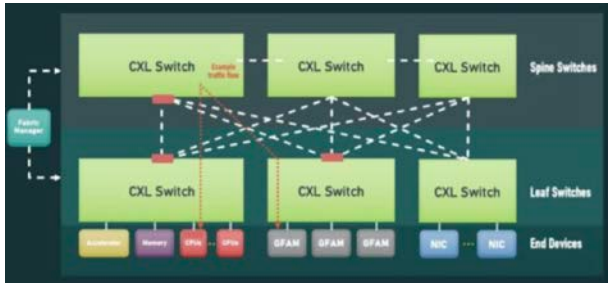


TPU AI独有架构效能高、高频宽、低延时



Intel CXL 池化方案：池化架构，弹性高、灵活扩展、适用通用架构





PCIE CXL/Ethernet DAC/AEC/AOC

超级以太网加速互连技术以更大规模，更高带宽，更低延迟的网络架构发展

超以太网联盟 (Ultra Ethernet Consortium，UEC) 正式成立, 人工智能和高性能计算给网络带来了新的挑战，需要更大规模、更高带宽密度、多路径、低延迟的网络技术, UEC将提供基于以太网的开放、可互操作、高性能的全通信堆栈架构，以满足大规模人工智能和高性能计算不断增长的网络需求。

Ultra Ethernet Consortium

AMD ARISTA BROADCOM cisco EVIDEN an atos business

Hewlett Packard Enterprise intel Meta Microsoft

LUXSHARE TECH 立讯技术

AI数据大模型算力中心网络架构：NVIDIA DGX GH200 全互连全景图

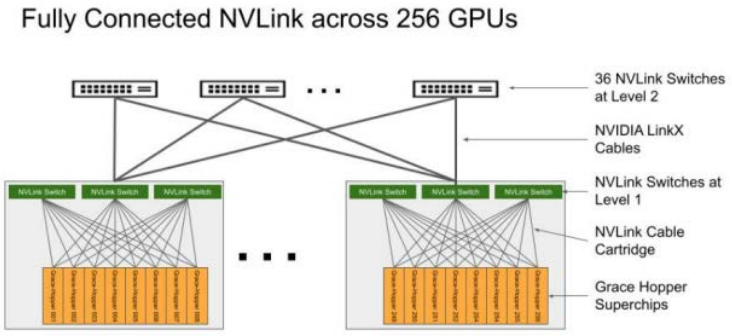


Figure 2. Topology of a fully connected NVIDIA NVLink Switch System across NVIDIA DGX GH200 consisting of 256 GPUs

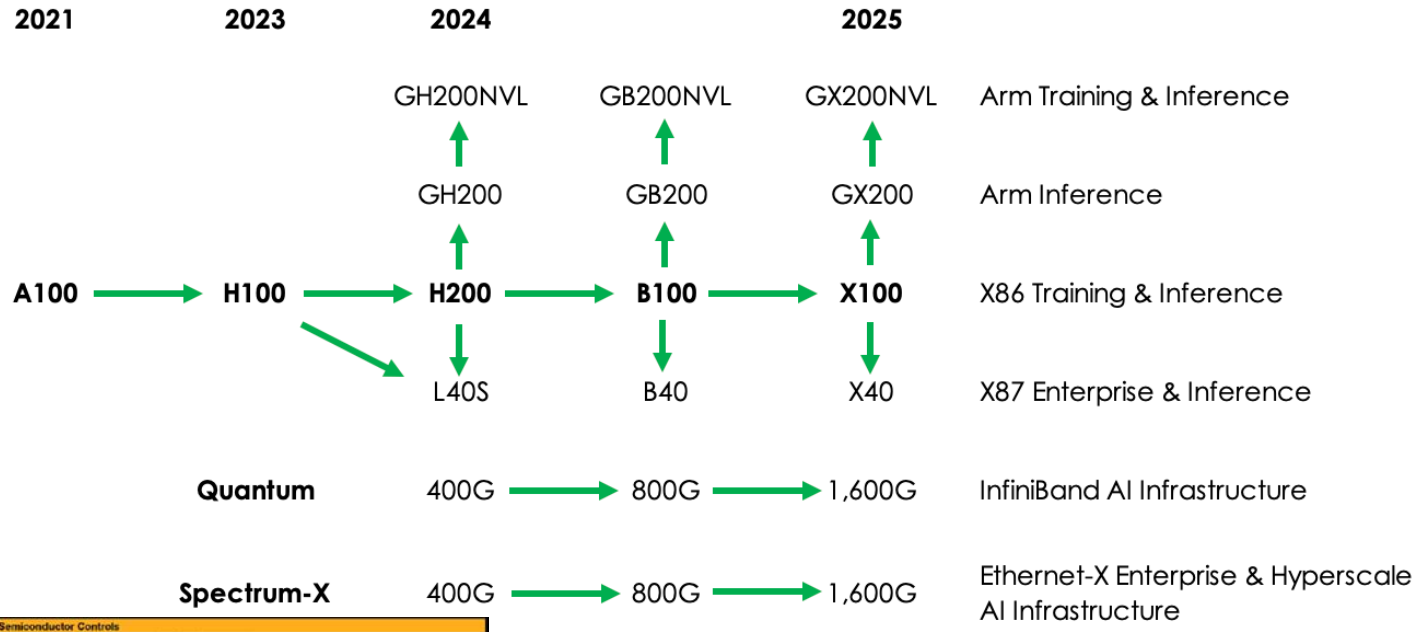
1. 柜间互连解决方案
2. 柜内系统解决方案
3. 服务器整机解决方案
4. 交换机器整机解决方案-Level1 Nvlink
5. 交换机器整机解决方案-Level2 Nvlink
6. 交换机器整机解决方案-Infiniband/Ethernet
7. 交换机器整机解决方案-管理

Server + Switch：

- ✓ 256台GPU服务器--256张NIC卡（OSFP112），256张DPU卡（2xQSFP112）
- ✓ Nvlink L1层交换机1U 96台 （叶交换机） （OSFP112）
- ✓ Nvlink L2层交换机1U 36台（脊交换机） （OSFP112）
- ✓ I/B交换机1U 24台（核心交换机） （OSFP112）
- ✓ Ethernet交换机1U 22台（叶&脊交换机） （QSFP56）
- ✓ Management Ethernet交换机1U 20台（RJ45+QSFP28）

高速互连（柜内互连采用铜缆，柜外互连采用光模块）

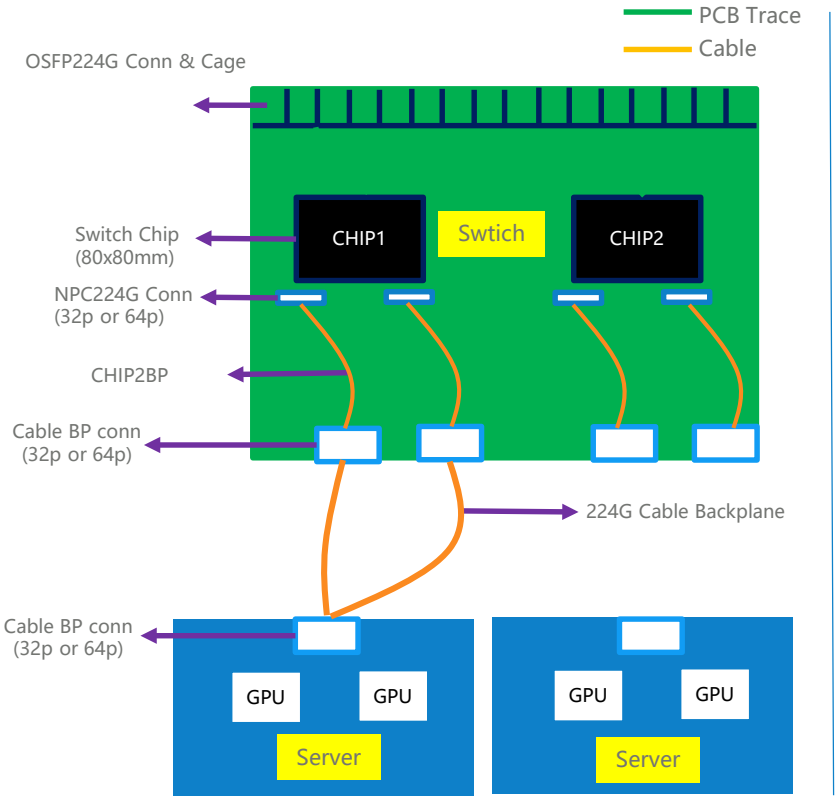
AI数据大模型算力中心网络架构：NV GPU发展路标-单通道速率向224G演进

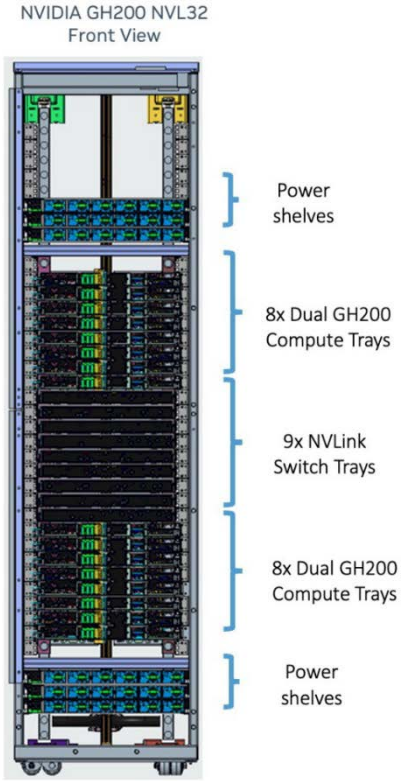


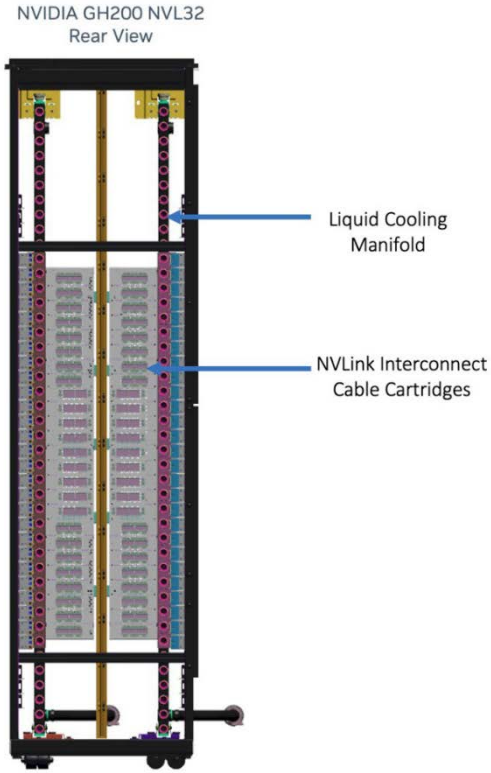
US AI Semiconductor Controls									
GPU	Memory Capacity (GB)	Memory Bandwidth (Tbps)	TeraFLOPs	Bitlength	TPP (TeraFLOPS x Bitlength)	Die size (mm^2)	Performance density (TPP / Die size)	Rule 3A090.a	Rule 3A090.b
H100 SXM	80	3.4	1,979	8	15,832	814	19.4	APPLIES	DOESN'T APPLY
H20 SXM	96	4.0	296	8	2,368	814	2.9	DOESN'T APPLY	DOESN'T APPLY
L40S	48	0.9	733	8	5,864	608	9.6	APPLIES	DOESN'T APPLY
L40	48	0.9	362	8	2,896	608	4.8	DOESN'T APPLY	APPLIES
L20	48	0.9	239	8	1,912	608	3.1	DOESN'T APPLY	DOESN'T APPLY
L4	24	0.3	242	8	1,936	295	6.6	APPLIES	DOESN'T APPLY
L2	24	0.3	193	8	1,544	295	5.2	DOESN'T APPLY	DOESN'T APPLY
A100 SXM	40	1.6	312	16	4,992	826	6.0	APPLIES	DOESN'T APPLY
V100 SXM	16	0.9	125	16	2,000	815	2.5	DOESN'T APPLY	DOESN'T APPLY
RTX 4090[1]	24	1.0	661	8	5,285	609	8.7	APPLIES	DOESN'T APPLY
RTX 4080[1]	16	0.7	320	8	2,560	379	6.8	APPLIES	DOESN'T APPLY
AMD MI210	64	1.6	181	16	2,896	770	3.8	DOESN'T APPLY	APPLIES
AMD MI250X	128	3.2	383	16	6,128	1,540	4.0	APPLIES	DOESN'T APPLY
AMD MI300X[2]	192	5.6	2,400	8	19,200	2,381	8.1	APPLIES	DOESN'T APPLY
Intel Gaudi2[2]	96	2.5	700	8	5,600	826	6.8	APPLIES	DOESN'T APPLY

1. Not "designed" for datacenter
2. No official specs, estimated

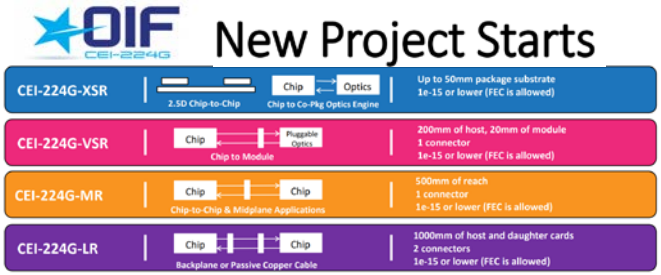
AI数据大模型算力中心网络架构：NV 架构发展趋势

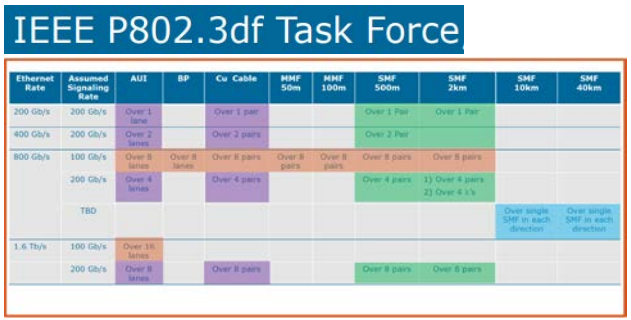






AI数据大模型算力中心网络架构：224G互连技术的挑战





2020年6月份OIF率先启动了面向下一代的CEI-224G项目

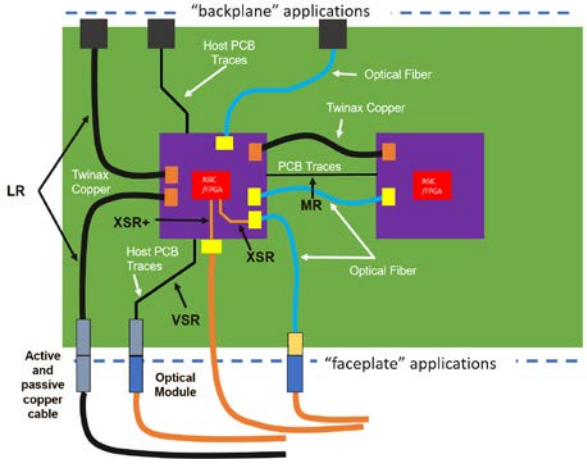
IEEE启动IEEE802.3df项目，开始修订224Gbps速率的下一代以太网标准。

224G的技术参数：

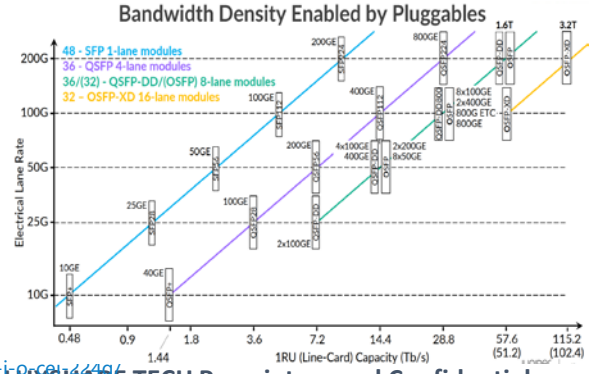
- 基频：56G
- 调制方式：PAM4
- 链路损耗~36dB

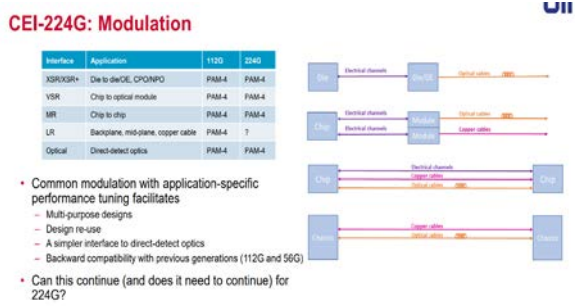
面对的挑战：

- 芯片封装Loss,PCB Loss,线缆，BGA/VIA设计，连接器，端接，制程稳定性，及散热，电源设计都存在较大的挑战

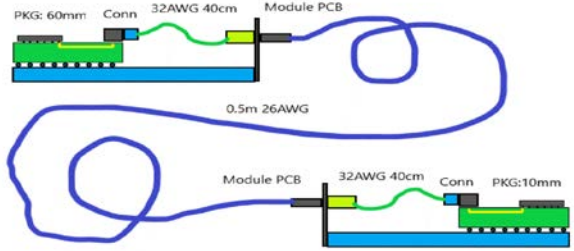


224G/通道IO口可以通过增加通道数或提升单通道数率来实现，升单通道速率将比增加通道数各有优势。





OIF-224MR,VSR,XSR 将采用PAM-4编码，背板，及铜缆同样采用PAM-4较高，中频预计在53.125GHz,整体的产品设计，制造良率，工艺技术难度挑战性更强，可能会有更多的接口产生。



互连结构将发生变化，C2IO的应用会在224G上应用，C2M,C2C,背板及铜缆的应用仍持续存在，但链路将会变短。对于连接器的设计特别是接口及端接的设计会是一个较大的挑战。DAC预计到2m，软有源及LPO预计到5m机柜内互连.

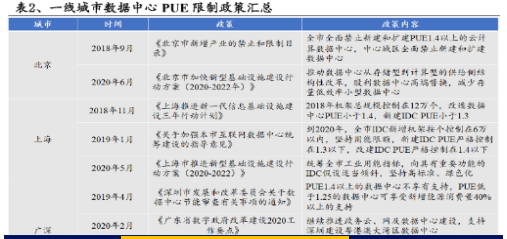
Source：https://grouper.ieee.org/groups/802/3/B400G/
https://www.oiforum.com/technical-work/hot-topics/common-electrical-i-o-cei-224g/
DesignCon Slide:SLIDES_Track09_OIFElectricalIOSpecificationProgressOnCEI_Tracy

LUXSHARETECH 立讯技术

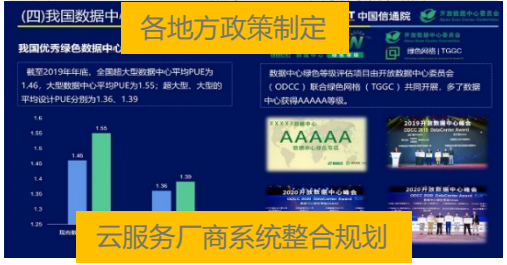
AI大模型数据中心互连解决方案：轻有源

- **低功耗的诉求：**数据中心的整个运营过程中，由于能耗占比超过了50%，降低数据中心的PUE一直是业内关心的重点
- **低成本的诉求：**随着高速网络，AI，大数据，5G/6G网络的持续发展，高速物理网络单通道速率迈向224Gbps，服务器内部互连PCIe信号向PCIE7.0(单通道128Gbps)发展，整体数据中心的成本也在不断的上涨，而面对如此高速率的挑战下，单纯被动铜缆（DAC）面临传输链路较短，无法满足布线要求，急需可以延展整体线缆长度但是又具备较低成本的方案

工信部关于绿色数据中心建设发文



各地方政策制定

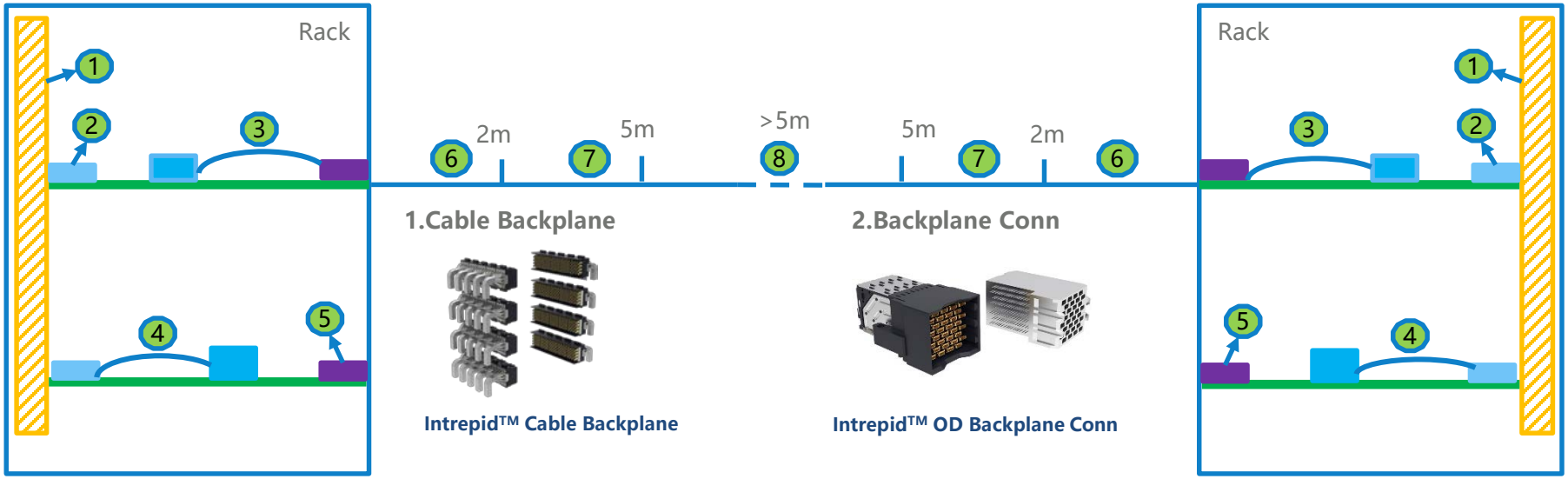


云服务厂商系统整合规划

“轻有源”原理：基于高速裸线数据库及光电转换技术对高速信号采用波束整形，时间重构，芯片直驱的方式，对高速传输链路的信号质量进行恢复达到降低功耗，延长传输距离，降低成本的目的。

- “有源（Active）”是指的主动器件在高速信号传输过程中的补偿，通过模拟或者数字电路对高速信号的传输进行处理，解决信号传输数据丢失或者失真的问题；
- “轻”是指的应用方案带来的低成本和低功耗（相对于传统的方案），达到延长高速电信号传输距离（变长），或同等距离下减小线径（变细），同等光传输功耗降低及时延缩短，实现整体降低成本和功耗的目的；
- “轻有源”的芯片目前主要指铜互连采用线性波束整形及信号重构的方案，光互连采用光芯片物理层讯号Direct-Drive的方案，去除功耗较高的数字芯片，对信号进行线性放大，驱动，均衡及时钟重构等方式做损耗补偿，进而实现低成本，低功耗，低时延，或延长传输距离的目的。

AI大模型数据中心互连解决方案：高密、高速、高算力



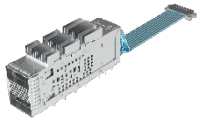
1.Cable Backplane

Intrepid™ Cable Backplane

2.Backplane Conn

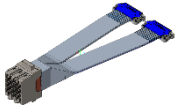
Intrepid™ OD Backplane Conn

3/4.Connector Module



CHIP2IO

OmniEdge™ CRE



CHIP2BP

5.High Speed I/O

800G OSFP Conn & Cage

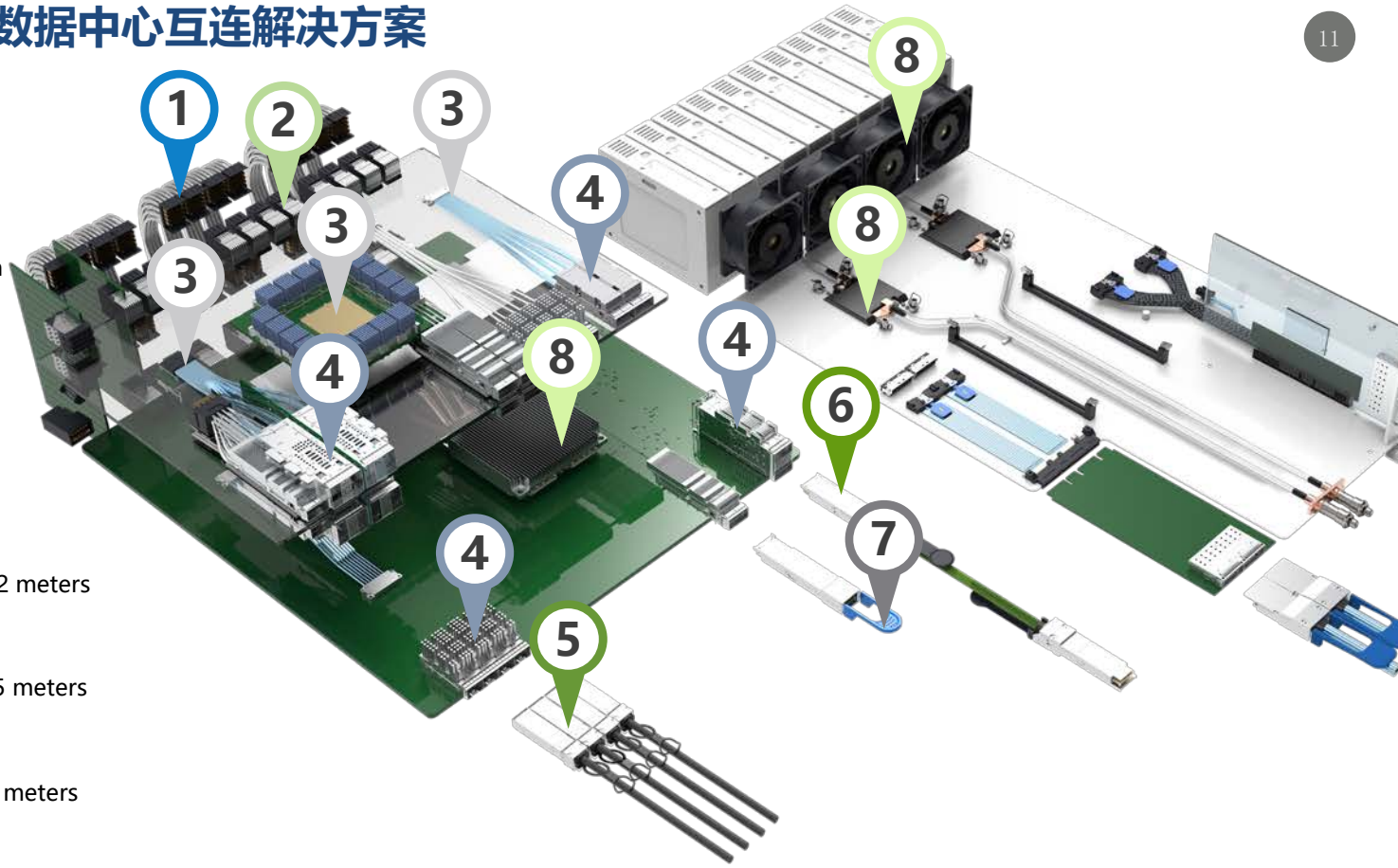
6.DAC & 7.AEC & 8.AOC/ transceiver

800G OSFP DAC &AEC

800G OSFP Silicon Photonic DR4

立讯技术---AI大模型数据中心互连解决方案

1. Intrepid™ Cable Backplane Cable Tray
2. Intrepid™ OD or internal cable solution
3. KOOLIO224 CPC / NPC OmniEdge™ CRE 112G
4. OSFP/QSFP-DD/QSFP Chip2IO
5. External transmission up to 2 meters
6. External transmission Up to 5 meters
7. External transmission Over 5 meters
8. Cold-plate, heat-sinks and fans



立讯技术---AI大模型数据中心热管理解决方案

1.Manifold

Customized Manifold

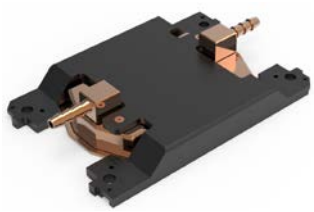
2.Fan

40 & 60 & 80 Axial flow fans

3.Heatsink

900W Switch Heatsink

4.Cold Plate



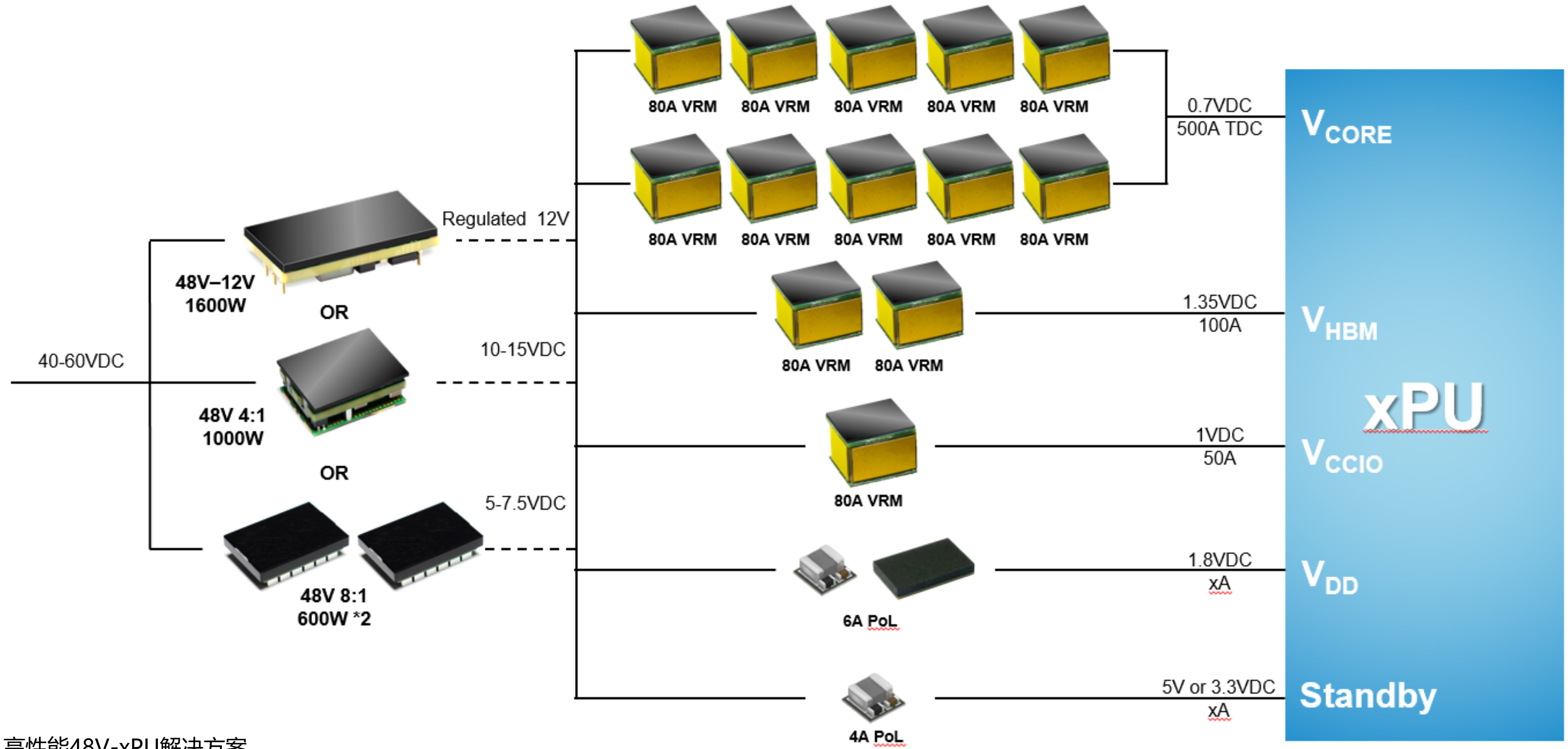
560W Cold Plate

5.CDU

In Rack CUD

LUXSHARETECH 立讯技术

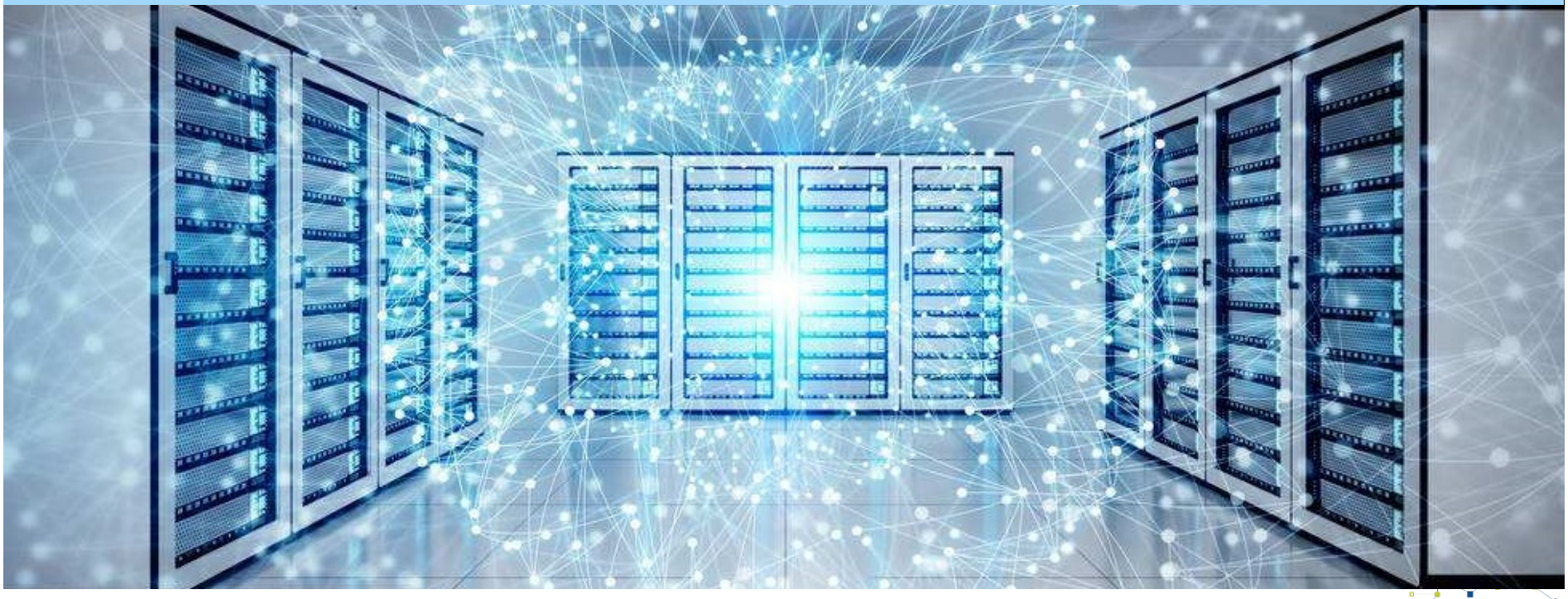
立讯技术---AI大模型数据中心电源解决方案



- 高性能48V-xPU解决方案
- 高效率，高密度，扩展性强

结语：从万物互联到万物智能、万智互联----智能世界，未来已至

未来十年，以大数据、AI、元宇宙等为代表的智能世界将加速到来，在迈向更美好的智能世界的征程中， 数据中心基础设施的建设和发展将越发重要。从多样性算力的融合，到算力的互联互通，都需要全行业的持续探索与创新，让我们大家齐心协力，助力全球计算产业的高质量发展和人类社会数字经济的腾飞



THANK YOU