AI算力跟踪深度：
辨析Scale Out与Scale Up——
AEC在光铜互联夹缝中挤出市场的What、Why、How

证券分析师：张良卫
执业证书编号：S0600516070001
联系邮箱：zhanglw@dwzq.com.cn
联系电话：021-60199793

研究助理：李博韦
执业证书编号：S0600123070070
联系邮箱：libw@dwzq.com.cn

2025年1月6日

核心观点

我们认为AEC是AI计算时代Scale Up需求被放大后的新兴技术方向，与Scale Out光互联并不构成需求的“零和游戏”，后续有望在柜间、柜内、ToR层互联中继续渗透：

1、绪论：如何辨析Scale Out与Scale Up网络？Scale Out网络实现集群内（Cluster，如万卡、十万卡集群）所有GPU卡互联，亮点在于网络内连接GPU数量大，与传统数据中心网络类似，Scale Up网络实现超节点内（SuperPod，如NVL 72）所有GPU卡互联，亮点在网络内单卡通信带宽高，**为AI算力场景下并行计算、内存墙等瓶颈催生出的新兴需求；**

2、What：DAC、AEC、AOC是什么？1）DAC、AEC都是铜连接，DAC无源（没有信号处理芯片）、AEC有源（有信号处理芯片），AOC是有源光连接；2）信号传输的核心部件与原理不同导致三类连接方式的功耗、距离、成本成倍递增；

3、Why：为什么AEC在DAC、AOC的夹缝中挤出空间？1）光进铜退已经发生于Scale Out网络：由于传输速率、距离均不断提升，光几乎已占据Scale Out所有互联场景；**2）能用铜的场景就只会用铜不会用光：**当前铜在10m以内高速连接仍可使用，因此光模块、CPO尚无法替代此场景；**3）Scale Up互联GPU数量少距离近，10m以内铜连接或可全覆盖，并不构成对光互联空间的侵蚀；4）距离、尺寸等差距导致铜缆内部有源（AEC）进无源（DAC）退；**

4、How：AEC在算力网络侧如何部署、前景如何？1）目前AEC主要用在Scale Up的柜间连接，如目前亚马逊Trn2-Ultra64使用 AEC柜间互联，ASIC芯片与AEC配比为1:1；**2）AEC与ASIC两者的兴起有相关性而非因果性，其底层逻辑是计算与通信的再解耦：云厂使用ASIC或英伟达HGX等，而非英伟达DGX方案时，完全来自英伟达的计算+通信方案也随之解耦，云厂便可以自主选择使用AEC；3）AEC还可以向柜内与ToR层渗透：**假如英伟达GB200 NVL72/8柜内换用AEC，一枚B200对应4.5支等效1.6T AEC，假如亚马逊Trn2-Ultra64柜内换用AEC，一枚Trainium2对应约3支800G AEC，**决定配比的关键因素仍为单卡带宽及交换机层数；**假如AEC参与ToR层连接，和算力卡配比为1:1；4）与DAC产业链中连接器品牌方是最核心环节不同，Retimer芯片供应商+品牌方变为AEC产业链中主导方；

投资建议：1）AEC有望在Scale Up兴起的趋势下获得越来越多的市场空间：关注兆龙互连，博创科技，推荐中际旭创，关注澜起科技；2）Scale Up有望带来新的交换机需求：推荐盛科通信，关注锐捷网络，紫光股份，中兴通讯；3）**“光退铜进”并未发生，光模块市场需求基本未被动摇：**推荐中际旭创，天孚通信，关注新易盛。

风险提示：算力互联需求不及预期；客户开拓与份额不及预期；产品研发落地不及预期；行业竞争加剧。

目录

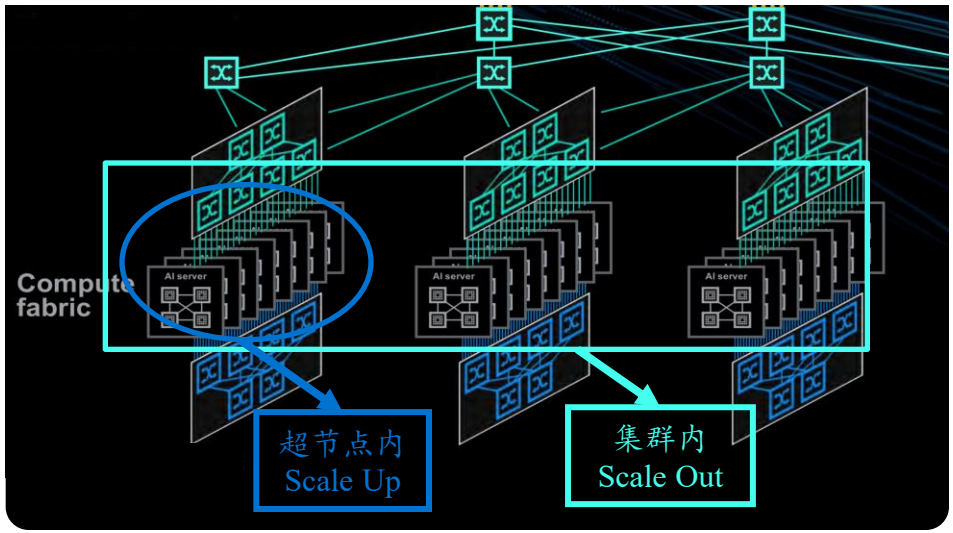
1. 绪论：如何辨析Scale Out与Scale Up网络？

若干超节点（SuperPod，如NVL 72）组成集群（Cluster，如万卡、十万卡集群）；

- Scale Out网络实现集群内所有GPU卡互联，亮点在于网络内连接GPU数量大，与传统数据中心网络类似；
- Scale Up网络实现超节点内所有GPU卡互联，亮点在于网络内单卡通信带宽高，为AI算力场景下新兴的网络架构。

（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

Scale Out网络与Scale Up网络



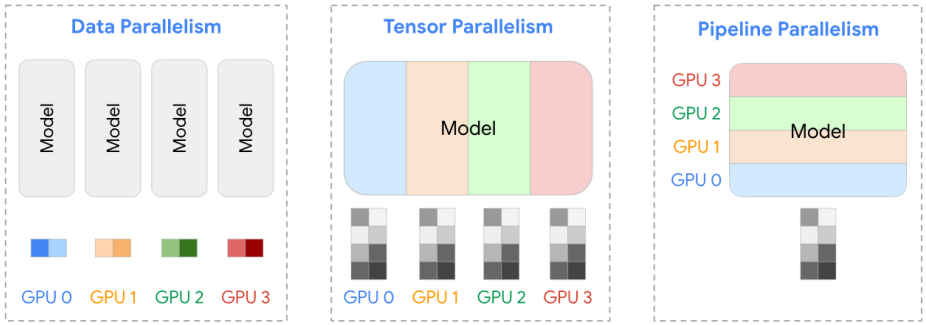
Scale Out与Scale Up网络对比（NVL72+CX-8网卡+三层Quantum-X800 IB网络）

	最大GPU数（张）	单卡带宽（Gb/s）
Scale Out	746496	800
Scale Up	72	7200

- AI训推需要分布式并行计算，基于对计算效率不断提升的追求，并行计算方式有数据并行（Data Parallelism）、流水线并行（Pipeline Parallelism）及张量并行（Tensor Parallelism）。
- **数据并行：**将输入数据分配给各个负载，各负载上基于不同数据进行同一模型的训练/推理；
- **流水线并行：**将模型分为若干层分配给各个负载，各负载分别进行不同层的计算；
- **张量并行：**将模型参数运算的矩阵拆分至各个负载，各负载分别进行不同的矩阵运算。

（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

数据并行（左），张量并行（中），流水线并行（右）计算原理图

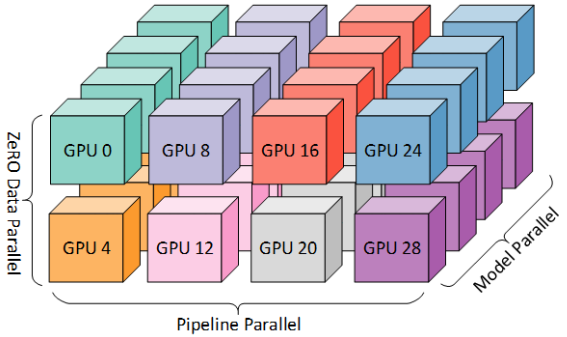


并行计算催生Scale Up网络需求：

- 几类并行计算方法各有优劣，大模型训练采用集合了多种并行方式的混合并行计算，如3D并行计算；
- 与数据并行、流水线并行相比，张量并行矩阵运算后需要同步，因此需要更高频、更低延时的数据传输，传输数据量也高出一到两个数量级；
- 通常数据并行、流水线并行基于容纳卡数更高的Scale Out网络，张量并行基于单卡带宽更高的Scale Up网络。

（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

3D并行计算



张量并行需要传输的数据量多出一到两个数量级
（GPT-3B模型基于32个GPU训练数据）

Traffic type	Volume	Number of messages	Message size
TP	~85 GB	680	125 MB
PP	~1 GB	16	125 MB
DP	741 MB	1	741 MB
EmbTableSyn	96 MB	1	96 MB

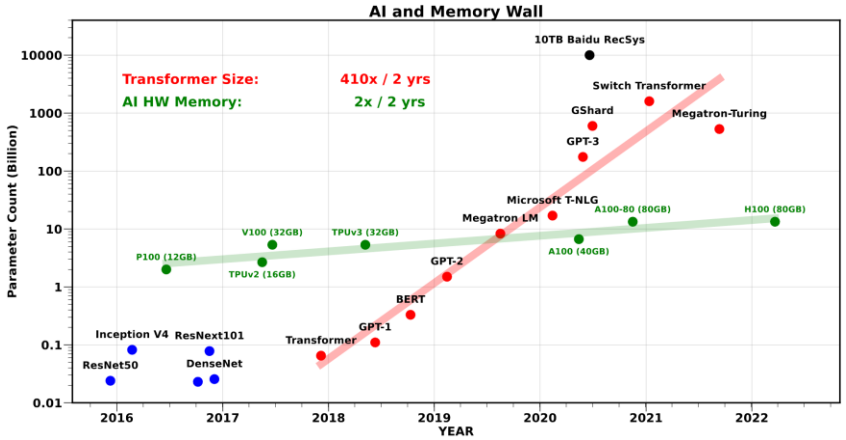
资料来源：《3D parallelism: Scaling to trillion-parameter models》，《Understanding Communication Characteristics of Distributed Training》，东吴证券研究所

训推计算的“内存墙”催生出通过Scale Up网络将显存池化的需求：

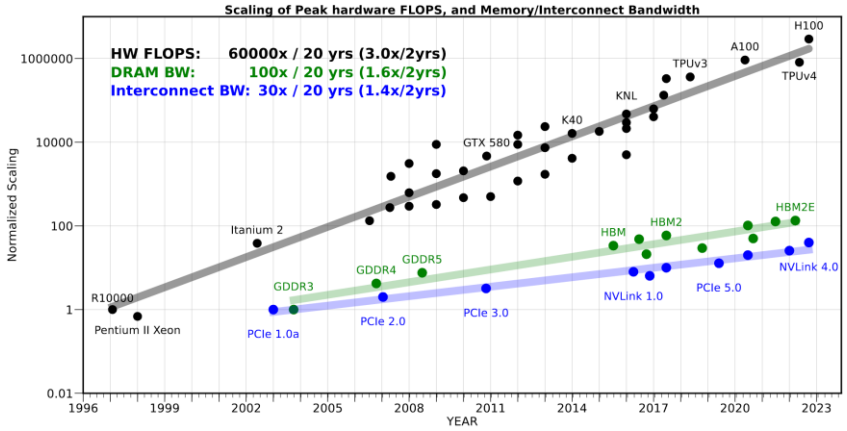
- 单一大模型的参数量与单卡显存的差距（即模型内存墙）、单卡算力与单卡显存间的差距（即算力内存墙）均逐代放大；
- 除模型参数外，推理计算生成的KV Cache（关键中间值的缓存，用于简化计算）占用显存大小也可达模型的50%甚至以上；
- 因此单卡运算时需从多张卡的显存读取所需参数、数据，为了尽可能减少数据传输时延，目前产业化应用最优解是使用Scale Up网络将显存池化，如NVL72。

（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

模型内存墙逐代放大



算力内存墙逐代放大



以一个通俗的例子辨析Scale Out与Scale Up：

- 上海市有加工厂A，B，C，……，Y，Z，AA，……（对应GPU），各工厂均配有自己的仓库a，b，c，……，y，z，aa，……（对应配套显存）；
- 所有工厂组成一个市内集群（Cluster），每三个工厂组成一个超节点（SuperPod），之前上海市集群内所有工厂都通过市内高架、快速路连接（即Scale Up网络）；
- 现在超节点内工厂做完每一个加工步骤，都需要把中间品汇总再分发至各个工厂进行下一步加工（即张量并行计算），同时开工工厂用到的原料、中间料大小超出自身配套仓库容量（即内存墙）；

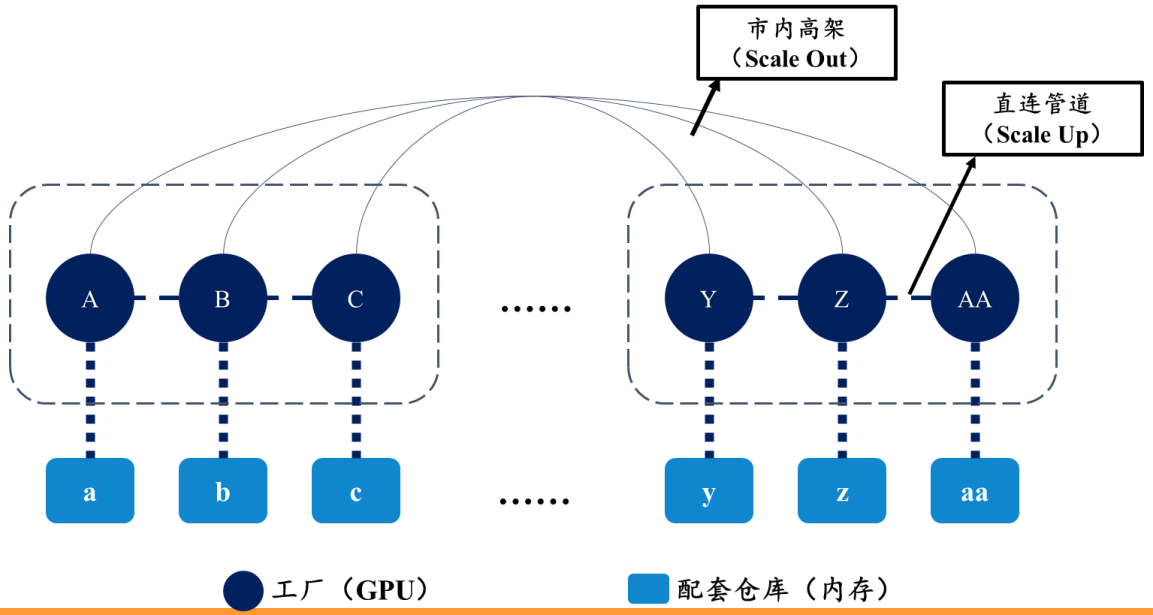
（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

以一个通俗的例子辨析Scale Out与Scale Up：

- 因此除了市内高架、快速路外（延安高架都时不时堵车...），各节点内3个工厂需要更加高速的直连方案（Scale Up），如地下直连管道
- 上海市集群内所有工厂通过市内高架互联（Scale Out），每组超节点内部工厂通过挖通的地下直连管道互联（Scale Up）
- 线条粗细代表信道传输速率

（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

市内工厂和配套仓库通过两类网络连接

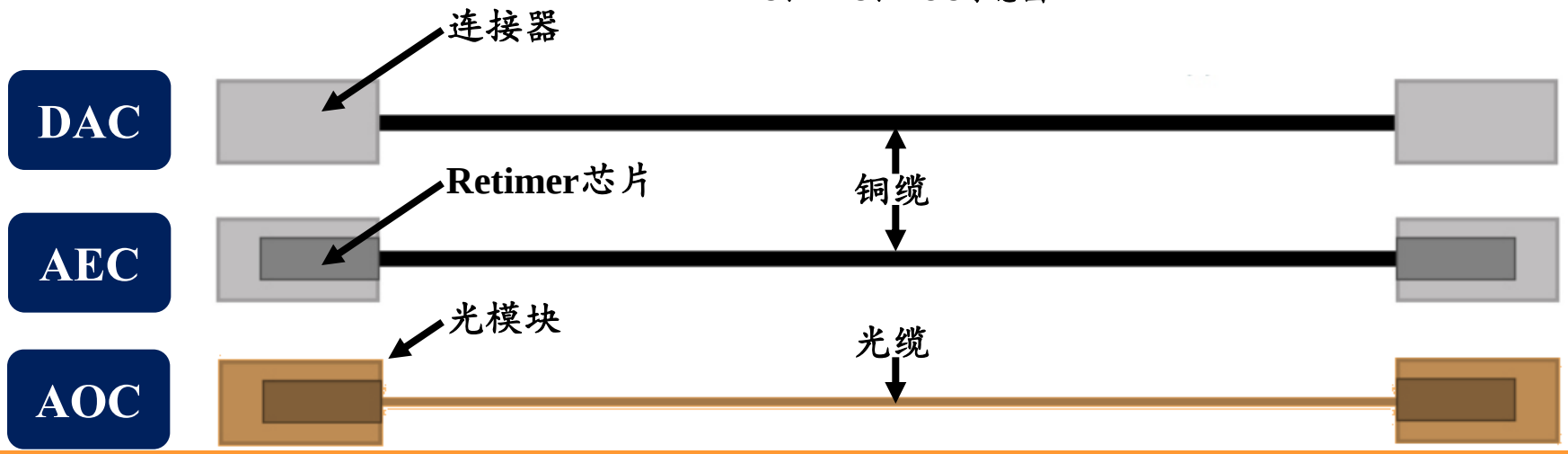


2. What：DAC、AEC、AOC是什么？

DAC、AEC都是铜连接，DAC无源（没有信号处理芯片）、AEC有源（有信号处理芯片），AOC是有源光连接：（目前ACC实用性不高本文暂不介绍）

- DAC（Direct Attach Cable）采用铜线将两端的连接器端口组装起来，不包含任何主动组件；
- AEC（**Active** Electrical Cable）含铜缆、连接器、Retimer芯片、PCB等，Retimer芯片可消除噪声并非线性放大信号，从而延长铜缆连接距离；
- AOC（**Active** Optical Cable）由两端光模块和光纤集成，通过光缆传输高速信号。

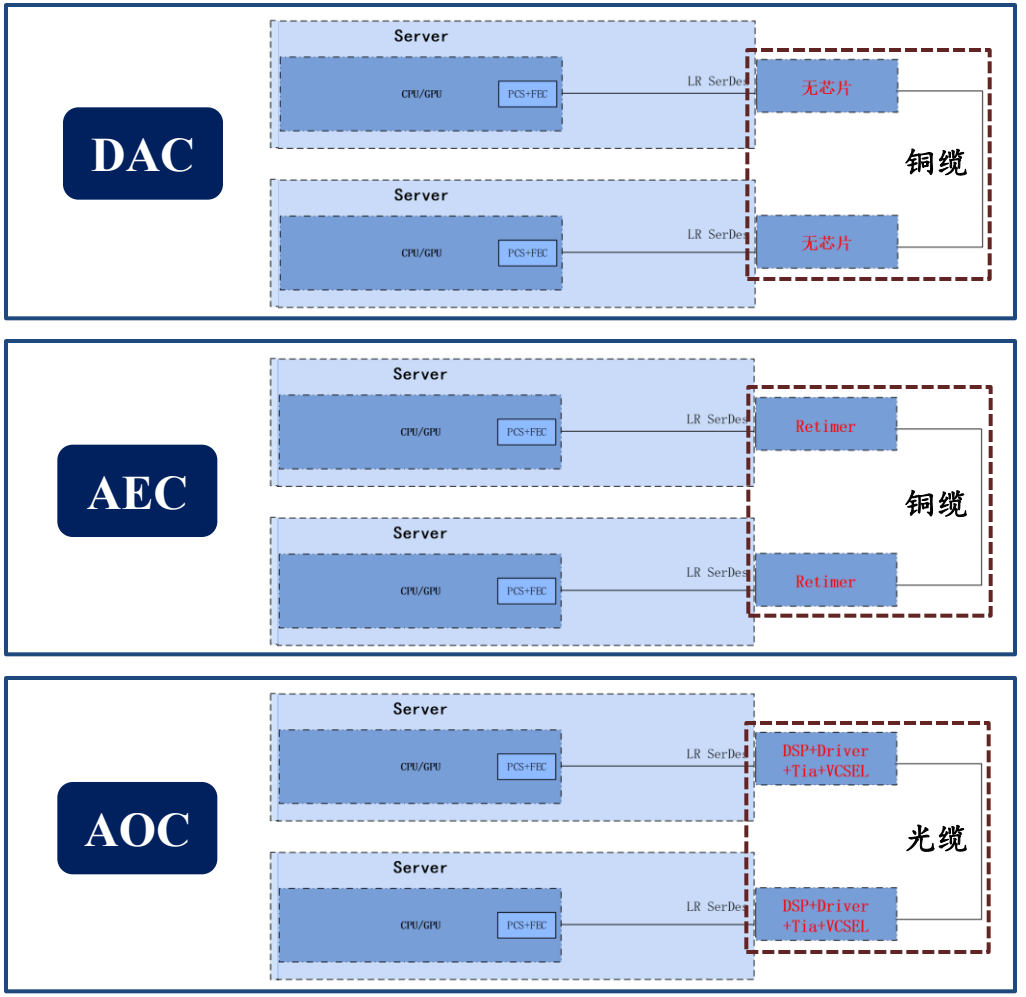
DAC、AEC、AOC示意图



信号传输的核心部件与原理不同导致三类连接方式的功耗、距离、成本成倍递增：

- DAC没有信号处理芯片，没有时延噪声消除、信号恢复等功能，直接通过铜缆传输信息；
- AEC中Retimer芯片将时延噪声消除、信号恢复，再通过铜缆传输信息；
- AOC中DSP、Driver、Tia芯片将时延噪声消除、信号恢复，再利用VCSEL等光芯片将电信号调制为光信号后通过光缆传输信息。

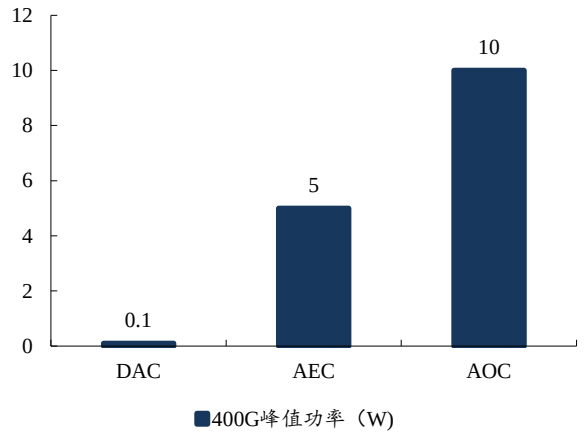
DAC、AEC、AOC的核心部件及原理图

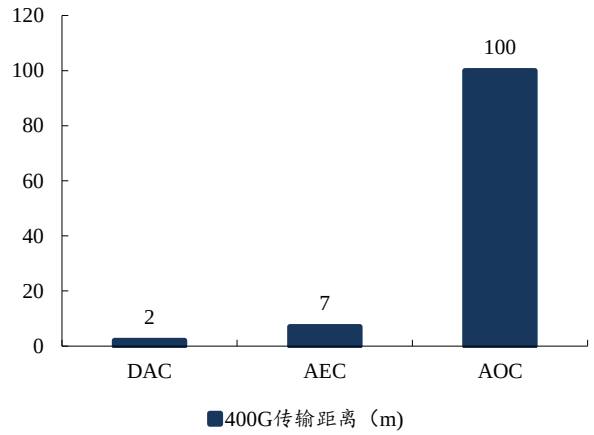


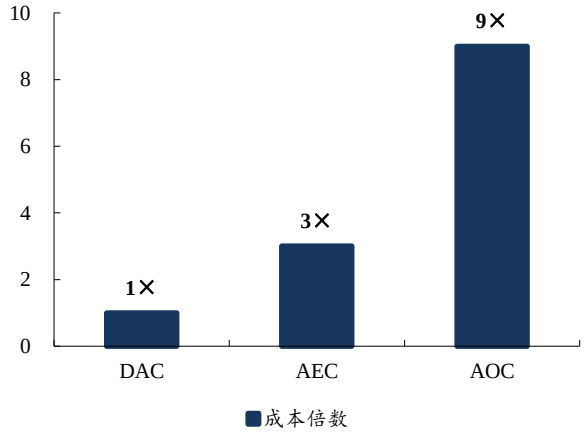
信号传输的核心部件与原理不同导致三类连接方式的功耗、距离、成本成倍递增：

- **功耗：**DAC、AEC、AOC中有源芯片复杂度逐渐增加，因此功耗也逐级提升，以400G速率为例，三者功耗分别为0.1、5、10W；
- **传输距离：** DAC、AEC、AOC对信号处理能力逐渐提升，因此有效距离也逐渐提升，以400G速率为例，三者传输距离分别为2、7、100米；
- **成本：** DAC、AEC、AOC中有源芯片复杂度逐渐增加，因此成本也逐级提升，以400G速率为例，AEC、AOC的成本分别为DAC的3倍、9倍。

DAC、AEC、AOC的功耗、传输距离、成本对比







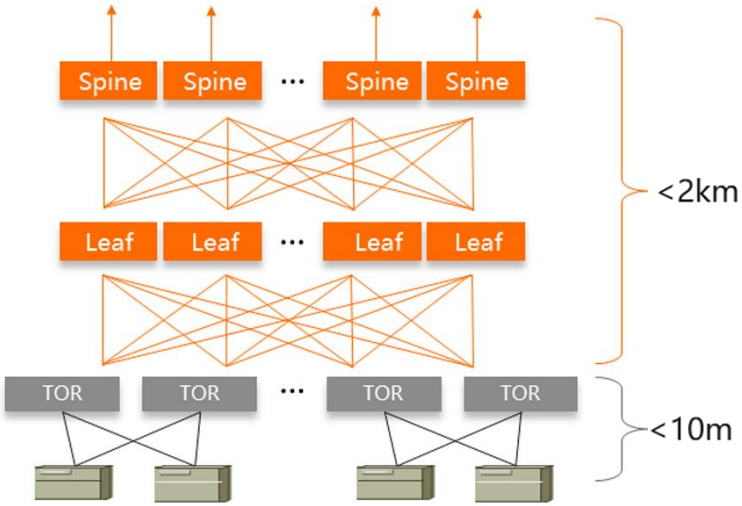
资料来源：Precision，东吴证券研究所

3. Why：为什么AEC在DAC、AOC的夹缝中挤出空间？

由于传输速率、距离均不断提升，光几乎已占据Scale Out所有互联场景：

- 目前AIDC内Scale Out网络的主流端口速率为400G、800G，在英伟达CX-8网卡及Quantum-X800交换机投入使用后更会高达1.6T；
- 同时在用于Scale Out的3层CLOS网络中，自上到下各层距离分别在千米级、百米级，服务器到ToR交换机的距离在10米以内；
- 前面已经提到，DAC、AEC等电互联在400G及以上速率的有效距离均在10米以内，因此在Scale Out场景光是主角。

用于Scale Out的典型3层CLOS网络（各层命名方式可能存在差异）



3. 光进铜退已经发生，但能用铜的场景就只会用铜不会用光

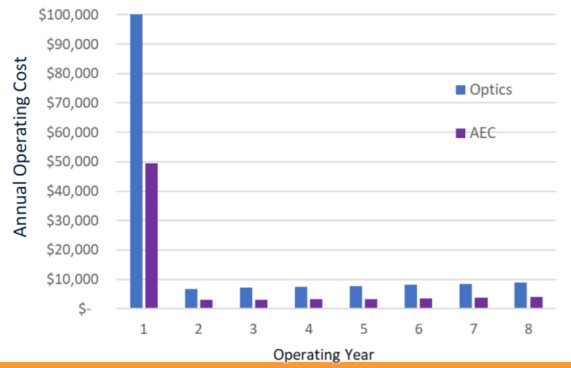
当前主流速率下铜在10m以内仍可使用，光模块、CPO尚无法替代此场景：

- 和铜连接相比，光连接最显著的优势是有效距离长，特别是在单通道速率不断提升的趋势下，以单通道100G的800G端口连接为例，AEC的有效范围在10m以内，而AOC可达百米，分立式光模块则更远；
- 和光连接相比，铜连接芯片复杂度低，在成本、功耗、稳定性上更有优势；
- **在铜连接有效的距离区间内（如<10m的800G传输），铜就是第一选择；**
- 随着单通道速率不断提升，铜的有效距离将不断减小，在铜连接有效距离无法覆盖的场景，**光进铜退已经且仍将继续发生，但后续需考虑单通道速率继续提升的难度及所需时间**（篇幅有限此处未详细展开，具体细节欢迎进一步交流）。

AOC与AEC性能比较

	主要芯片	传输距离	功耗	成本	稳定性
AOC	DSP，Driver，Tia，VCSEL	100m	高	高	低
AEC	Retimer	10 米以内	低	低	高

8年运营时间内AOC与AEC历年成本比较

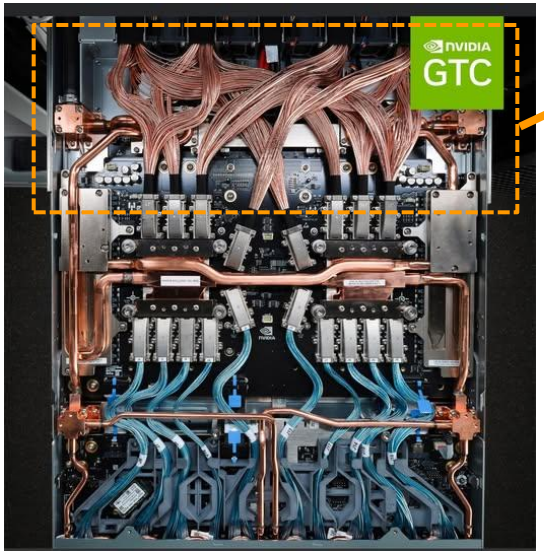


资料来源：Credo，东吴证券研究所

Scale Up互联GPU数量在数十、数百级别，10m以内铜连接或可全覆盖：

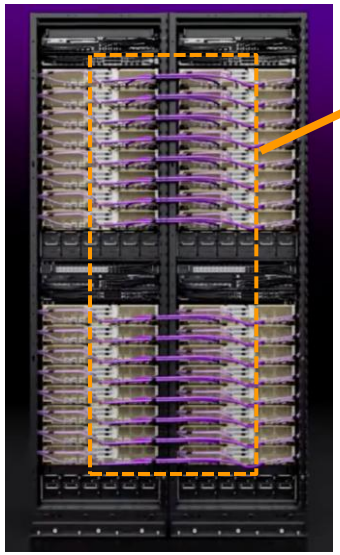
- 如第一章所述，并行计算、内存墙等瓶颈推动AI计算中涌现出Scale Up需求，**这类需求是增量需求，不构成对Scale Out网络中光互联需求的侵蚀**；
- 以英伟达GB200 NVL72及亚马逊Trn2-Ultra64超级服务器（超节点）为例，一个超节点内需要Scaling Up互联的算力卡在同一或相邻服务器内，连接距离<10m，因此都采用铜缆来分别实现柜内和柜外Scaling Up（具体分析详见下一章）。

GB200 NVL72卡间互联（柜内）



铜缆

Trn2-Ultra64超级服务器连接四个16卡服务器（柜间）

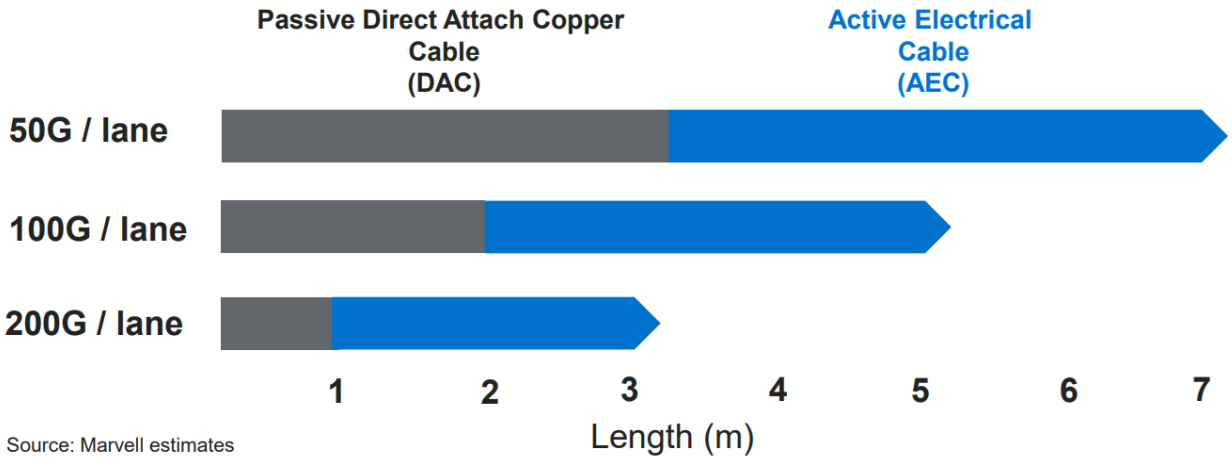


铜缆

与光进铜退逻辑类似，距离、尺寸等差距导致铜缆内部有源（AEC）进无源（DAC）退：

- 由于多出Retimer（广义上也算DSP），AEC与DAC相比在有效距离上更远；
- 根据Marvell，单通道200G（一般对应1.6T）DAC最远为1m，单通道100G（一般对应1.6T）DAC最远2m，而Marvell/Credo的1.6T AEC最远分别为3m/2.75m，800G AEC最远分别为5m/7m。

AEC与DAC有效距离对比



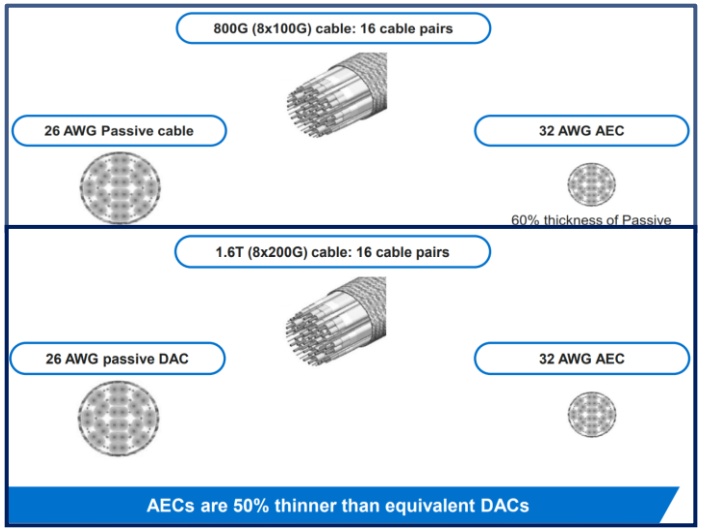
资料来源：Marvell，Credo，东吴证券研究所

与光进铜退逻辑类似，距离、尺寸等差距导致铜缆内部有源（AEC）进无源（DAC）退：

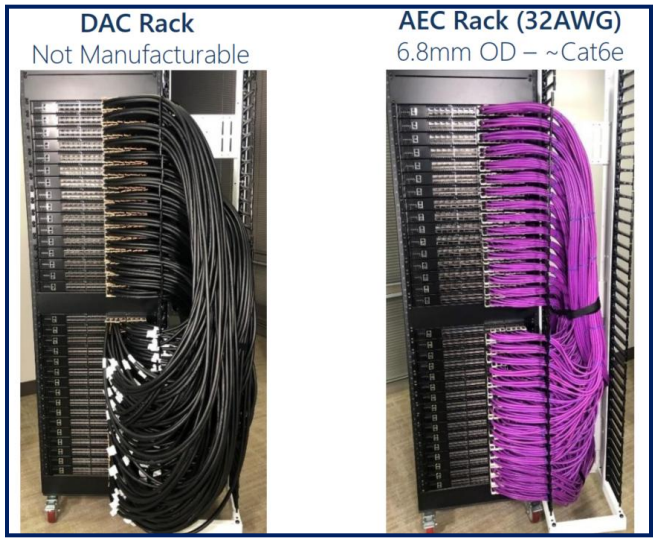
- 铜缆中的介质层可以减少双绞线内两根铜缆的电磁干扰，当速率提升时，干扰增大，为了减少干扰需要将介质层体积增大；
- DAC没有Retimer，需要将介质层做得更大，因此线缆更粗；
- 如单通道100G、200G DAC线径一般需为26 AWG（对应0.4mm，AWG为线径标准，值越小代表线缆越粗），而AEC可以做到32AWG（对应0.2mm）。线缆越粗，做高密度连接的弯折、排线越难（右图可作为对比参考）。

（由于篇幅有限本文未就技术原理做详细阐述，具体细节欢迎进一步交流）

AEC与DAC线径对比



AEC与DAC排线对比



资料来源：Marvell，Credo，东吴证券研究所

4. How：AEC在算力网络侧如何部署、前景如何？

目前亚马逊Trn2-Ultra64超级服务器使用DAC实现柜内互联、AEC实现柜间互联，大体算力结构如下：

- 每个Compute Tray（计算托盘）中有两枚Trainium2芯片；
- 8个计算托盘共16枚Trainium2芯片组成Trn2服务器的算力硬件部分，每枚芯片通过AEC与所在服务器相邻两个服务器内各一个芯片互联；
- **4台Trn2在两个机位上2×2叠放组成Trn2-Ultra64超级服务器。**

每个计算托盘有两个Trainium2芯片

8个托盘组成16卡Trn2服务器算力部分

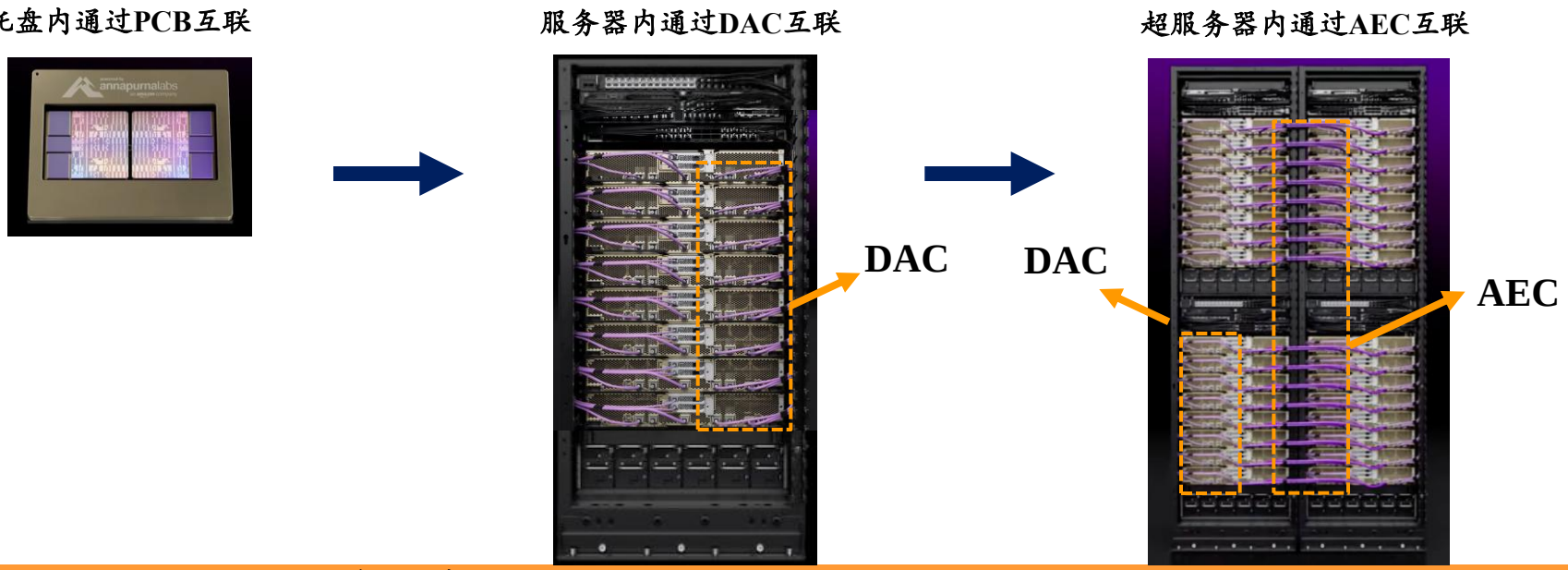
4台服务器组成Trn2-Ultra64超级服务器（柜间）

4. AEC目前主要用在Scale Up的柜间连接

目前亚马逊Trn2-Ultra64超级服务器使用AEC做柜间互联，DAC做柜内互联，单个芯片基于NeuronLinkv3（对标NVLink）可实现640GB/s，即5120Gb/s带宽的Scale Up网络，与超服务器内6枚芯片实现互联，大体算力结构如下：

- **托盘内：**每枚芯片与同一托盘内另一芯片通过PCB互联；
- **服务器内：**每枚芯片通过DAC和服务器内另三个托盘内各一枚芯片互联；
- 每枚芯片通过AEC与所在服务器相邻两台服务器内各一枚芯片互联。

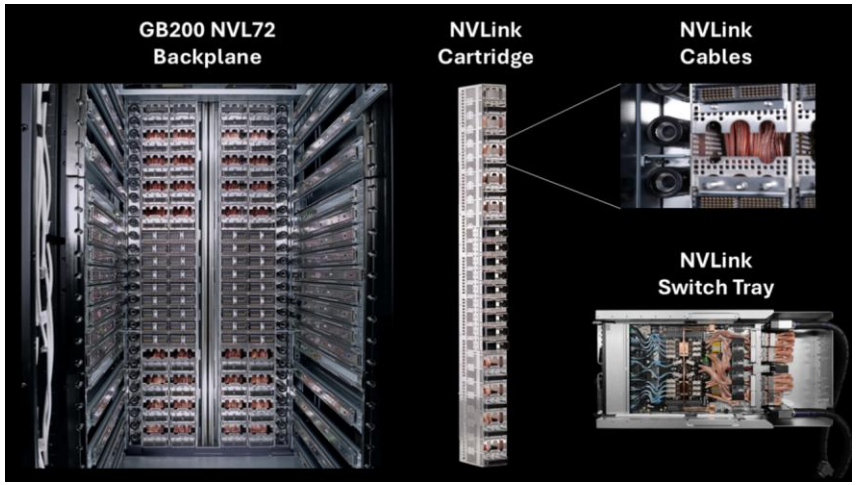
因此目前Trn2-Ultra64的Scale Up网络内Trainium2芯片与AEC比例为1:1。



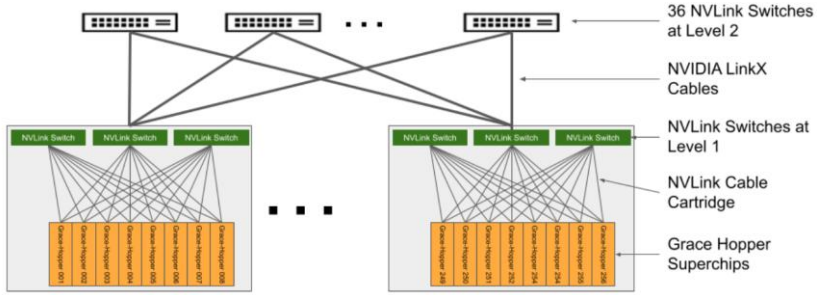
4. AEC与ASIC两者的兴起有相关性而非因果性，其底层逻辑是计算与通信的再解耦

- **目前英伟达DGX产品将算力与通信产品耦合，Scale Up主要使用DAC或光连接：**GB200 NVL72中主要使用DAC连接，GH200 NVL256主要使用光连接；
- 当云厂使用自研ASIC时，计算与通信是天然解耦的，云厂可以自行选择搭建网络的方式，包括Scale Up网络；
- 同时，**当云厂可以拿到充足的HGX等算力模块时，之前完全来自英伟达的计算+通信一体方案也随之解耦，云厂也可以选择可能更节省精力、成本的AEC而非DAC、光连接；**
- **计算与通信的解耦不仅仅会发生在ASIC上。**

GB200 NVL72使用DAC做Scale Up连接



GH200 NVL256主要使用光连接做Scale Up连接



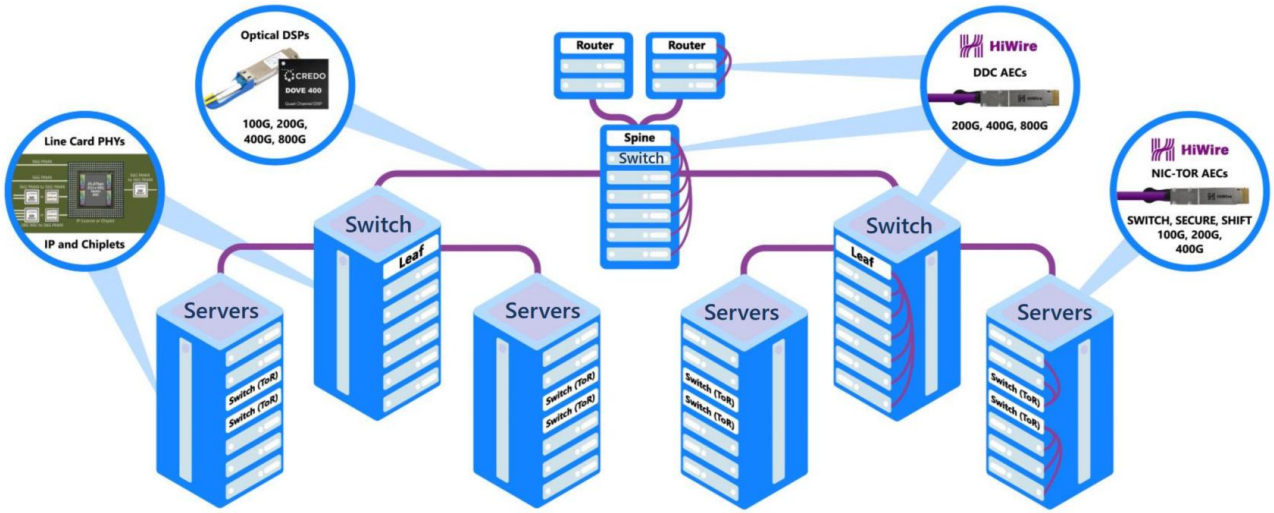
1）从第三章AEC与DAC的比较可以看出，**后续AEC有望在柜内场景替代DAC，但需注意时延和功耗上的权衡：**

- **假如英伟达柜内换用AEC，算力卡和AEC配比如何？**以GB200 NVL72/8为例，单枚B200 NVLink带宽为单向900GB/s，即7200Gb/s，以等效1.6T计，假设仍有NVSwitch，则**一枚B200需等效7200/1600=4.5支1.6T AEC**，如果柜内卡数减少，转向类似GH200 NVL256的单层交换机架构，比例仍有望不变；
- **假如亚马逊柜内换用AEC，算力卡和AEC配比如何？**以Trn2-Ultra64为例，单枚Trainium2带宽为单向640GB/s，即5120Gb/s，以等效800G计，假设仍是直连无交换机，则**一枚Trainium2需等效5120/800/2≈3支800G AEC。**

决定配比的关键因素与Scale out中一样——单卡带宽及交换机层数。

2）从第三章AEC与AOC的比较可以看出，**后续AEC有望在<10m的服务器网卡到ToR交换机连接场景中渗透，和算力卡配比为1:1（速率取决于选用网卡）。**

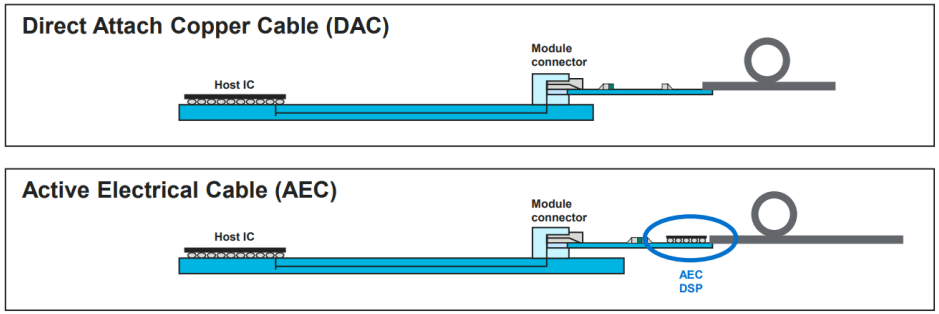
AEC潜在应用场景



Retimer芯片供应商+品牌方变为AEC产业链中主导方：

- 正是相比DAC多出的Retimer芯片使得AEC在有效长度、尺寸等性能上优于DAC；
- 与DAC产业链中连接器+品牌方是最核心环节不同，AEC产业链中Retimer芯片+品牌方成为了最核心环节，我们认为是否拥有Retimer自供能力将是决定品牌方AEC生产成本的最关键因素，因此后续AEC产业链的主导方将是Retimer芯片商。

AEC与DAC构成对比



AEC与DAC主要物料环节代表厂商（标红代表为产业链主导环节）

主要物料环节	DAC	AEC
Retimer芯片	-	**Credo，博通，Marvell**
连接器	**安费诺，泰科，莫仕**	安费诺，泰科，莫仕，瑞可达
铜缆	沃尔核材	兆龙互连，沃尔核材

5. 投资建议

我们认为AEC有望在Scale Up兴起的趋势下获得越来越多的市场空间：

- 关注兆龙互连，博创科技，推荐中际旭创，关注澜起科技；

我们认为Scale Up有望带来新的交换机需求：

- 推荐盛科通信，关注锐捷网络，紫光股份，中兴通讯；

我们认为“光退铜进”并未发生，光模块市场需求基本未被动摇：

- 推荐中际旭创，天孚通信，关注新易盛。

6. 风险提示

6. 风险提示

- **算力互联需求不及预期：**若后续下游客户算力建设投入未达预期，或AEC在AI算力网络架构中的使用数量未达本报告中预期情况，各客户对于AEC等产品的需求也将不及预期，相关公司业绩表现将受到影响；

- **客户开拓与份额不及预期：**如果相关公司未如预期开拓潜在客户，或在客户处份额低于预期，公司业绩将受到影响；

- **产品研发落地不及预期：**如果相关公司在具有潜在应用前景的产品研发及量产应用上未达预期，将对公司业绩的表现造成影响；

- **行业竞争加剧：**如果行业竞争持续加剧，相关公司产品份额存在下降的可能。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后6至12个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证50指数），具体如下：

公司投资评级：

买入：预期未来6个月个股涨跌幅相对基准在15%以上；

增持：预期未来6个月个股涨跌幅相对基准介于5%与15%之间；

中性：预期未来 6个月个股涨跌幅相对基准介于-5%与5%之间；

减持：预期未来 6个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来6个月内，行业指数相对强于基准5%以上；

中性： 预期未来6个月内，行业指数相对基准-5%与5%；

减持： 预期未来6个月内，行业指数相对弱于基准5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街5号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn

东吴证券 SOOCHOW SECURITIES

东吴证券 财富家园