国联证券 GUOLIAN SECURITIES

证券研究报告

行业研究|行业专题研究|通信

百万卡算力之路：多DC分布式训练和DCI需求增长

请务必阅读报告末页的重要声明

2024年10月07日

报告要点

高能耗的算力集群将迫使 AI 模型训练从单计算中心训练，走向多 DC 协同训练，远距离异步协同训练将成为主流。Meta 和 Google 已经开始了多 DC 分布式训练，其中 Google 的 Gemini 1 Ultra 就是通过多 DC 的分布式训练实现的，OpenAI 和微软计划将各个超大型园区互连在一起起来，并在全国范围内进行大规模的分布式训练。多 DC 协同训练给网络带来挑战。400G ZR 相干技术优势明显，ZR 光模块需求有望增长。我们认为 AI 算力对网络的需求正在向 DCI 场景扩散，有望带动 DCI 市场的高速增长。建议关注 DCI 产业链和 400G/800G ZR 光模块供应商。

分析师及联系人


张宁　　张建宇

SAC：S0590523120003　SAC：S0590524050003

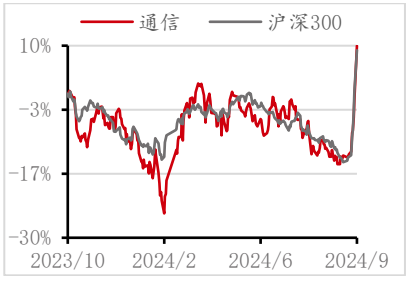
通信

百万卡算力之路：多 DC 分布式训练和 DCI 需求增长

投资建议：强于大市（维持）

上次建议：强于大市

相对大盘走势



相关报告

1、《通信：华为领航，AI 和国产算力产业持续蓬勃发展》2024.09.22

2、《通信：CIOE2024:聚焦 AI，关注 1.6T 和 DCI 新变量》2024.09.16

扫码查看更多

➢ 海外科技巨头积极布局多 DC 分布式训练

关于 AI 大模型训练在什么阶段需要 DCI 联接，需要多少 DCI 带宽，我们认为不同的互联网公司，因为 IDC 资源不同、业务模型不同，会有较大的配置差异。但是 Meta 和 Google 已经开始了多 DC 分布式训练，其中 Google 的 Gemini 1 Ultra 就是通过多 DC 的分布式训练实现的。谷歌目前有两个主要的多数据中心区域，分别位于俄亥俄州和爱荷华州/内布拉斯加州。OpenAI 和微软更加雄心勃勃，计划将各个超大型园区互连在一起起来，并在全国范围内进行大规模的分布式训练。

➢ 分布式训练给网络带来挑战

AI 训练步入十万卡时代，跨 DC 协同训练对网络带来挑战。（1）AI 训练对网络丢包的敏感度高。（2）大象流会导致网络中的传统基于五元组的负载分担方法失效，链路负载不均衡，降低网络使用率。（3）在万卡集群中，极端情况下流量瞬时并发可达上千 Tbps。目前，十公里的跨机楼并行训练算效损失可低于 5%，具备可行性，未来百公里级、千公里级的跨地域并行训练欲将损失控制在 10%以下，除需建设长距离超宽 DCI 网络之外，还涉及模型切分策略、集合通信算法、无损网络技术等。

➢ 400G ZR 相干技术优势明显，ZR 光模块需求有望增长

400G ZR 相干光学技术有望在 DCI 中取代传统的波分复用（WDM）系统。相比于传统的 WDM 系统，400G ZR 系统更加简洁，主要有 MUX/DEMUX，并采用可调谐激光器的相干光模块，直接放在客户侧的交换机/路由器上。根据 LightCounting 的预测，2024-2028 年，400G ZR，ZR+的光模块保持增长。产品价值量方面，根据 LightCounting 预测，2023 年 400G ZR 的价格为 3230 美元，2024 年 800G ZR 的价格为 4800 美元。

➢ 建议关注 DCI 产业链和 400G/800G ZR 供应商

海外科技巨头积极布局多 DC 分布式训练，我们认为 AI 算力部署对网络的需求正在向 DCI 场景扩散，有望带动 DCI 市场的高速增长。我们建议关注：国内 OTN 厂商：中兴通讯、烽火通信、光迅科技；有 400G/800G ZR 产品布局的德科立、中际旭创、新易盛、华工科技；铌酸锂调制供应商：光库科技。

风险提示：AI 产业发展不及预期风险、算力需求不及预期风险、技术发展不及预期风险、市场竞争加剧风险。

正文目录

图表目录

1. 多 DC 协同训练，算力竞争下半场

1.1 海外科技巨头积极布局多 DC 分布式训练

海外科技巨头积极布局多 DC 分布式训练。关于 AI 大模型训练在什么阶段需要 DCI 联接，需要多少 DCI 带宽，我们认为不同的互联网公司，因为 IDC 资源不同、业务模型不同，会有较大的配置差异。但是我们可以清楚的看到 Meta 和 Google 已经开始了多 DC 分布式训练，其中 Google 的 Gemini 1 Ultra 就是通过多 DC 的分布式训练实现的。

图表1：Meta 的分布式训练架构

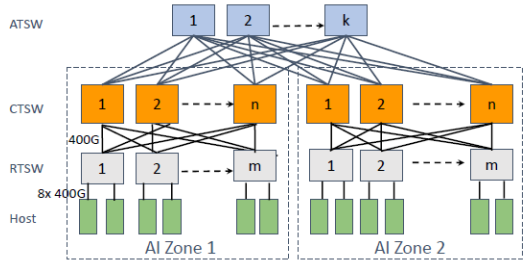
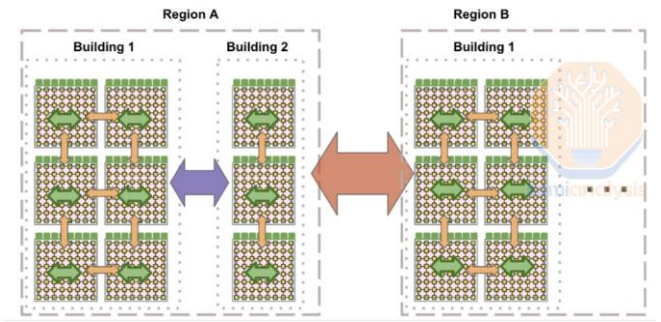


Figure 6: Backend Network Topology

资料来源：《RDMA over Ethernet for Distributed AI Training at Meta Scale》Adithya Gangidi 等，国联证券研究所

图表2：Google 的大规模训练结构图

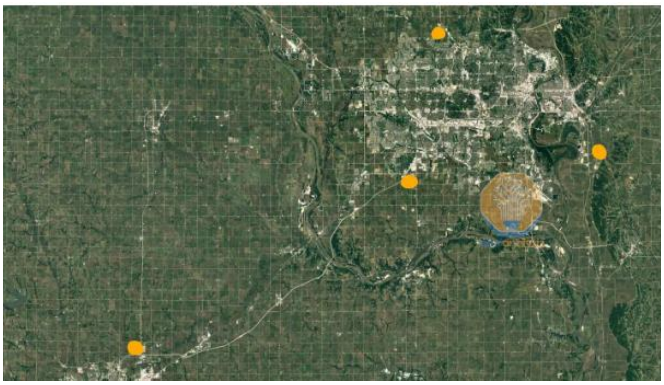


资料来源：semianalysis，Google，国联证券研究所

谷歌积极布局多 DC 分布式训练。谷歌有两个主要的多数据中心区域，分别位于俄亥俄州和爱荷华州/内布拉斯加州。康瑟尔布拉夫斯周围的区域正在积极扩展，容量将超过现有容量的两倍。除了上述园区外，谷歌还在该地区拥有另外三个正在建设中的站点，这些站点都在升级高带宽的网络。

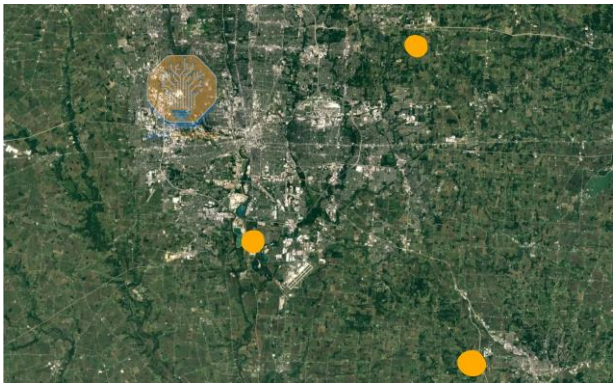
有三个站点彼此相距约 15 英里（康瑟尔布拉夫斯、奥马哈和爱荷华州帕皮隆），另一个站点距离约 50 英里，位于内布拉斯加州林肯市。预计到 2026 年，四个园区的结合将形成一个 GW 级的人工智能训练集群，其中林肯数据中心将是谷歌最大的单个站点。

图表3：Google 的 IDC 集群（位于康瑟尔布拉夫斯、奥马哈、爱荷华州帕皮隆和内布拉斯加州林肯市）



资料来源：semianalysis，国联证券研究所

图表4：Google 的 IDC 集群（位于俄亥俄州哥伦布市附近）



资料来源：semianalysis，国联证券研究所

OpenAI 和微软更加雄心勃勃，计划将各个超大型园区互连在一起起来，并在全国范围内进行大规模的分布式训练。

图表5：微软在凤凰城区域的 IDC 园区位置

资料来源：《Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure》Gigawatt Clusters 等，国联证券研究所

图表6：微软在德克萨斯州的 IDC 园区位置

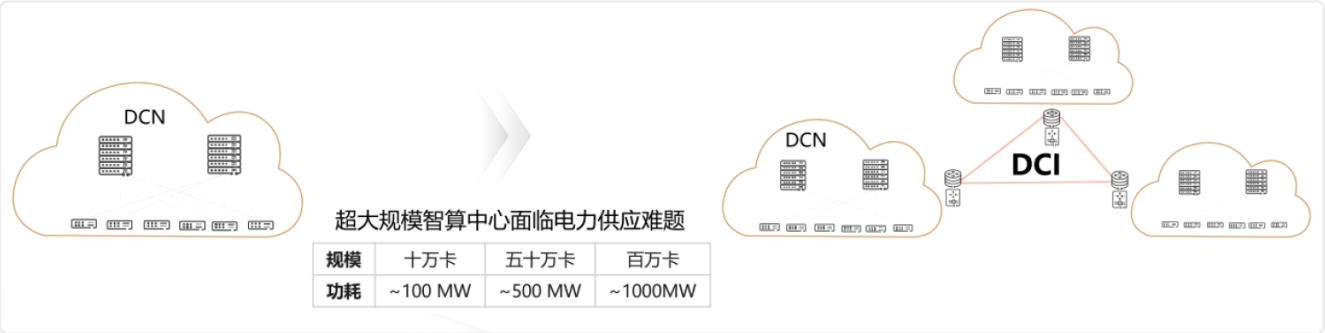
资料来源：《Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure》Gigawatt Clusters 等，国联证券研究所

1.2 分布式训练给网络带来挑战

AI 训练步入十万卡时代，跨 DC 协同训练对网络带来挑战。（1）AI 训练对网络丢包的敏感度高，即使是 0.1%的丢包率也可能导致训练效率降低 50%，严重影响协同训练效果。（2）大象流会导致网络中的传统基于五元组的负载分担方法失效，链路负载不均衡，降低网络使用率。（3）在万卡集群中，由于业务高突发和高并发，极端情况下流量瞬时并发可达上千 Tbps。

目前，十公里的跨机楼并行训练算效损失可低于 5%，具备可行性，未来百公里级、千公里级的跨地域并行训练欲将损失控制在 10%以下，除需建设长距离超宽 DCI 网络之外，还涉及模型切分策略、集合通信算法、无损网络技术等。

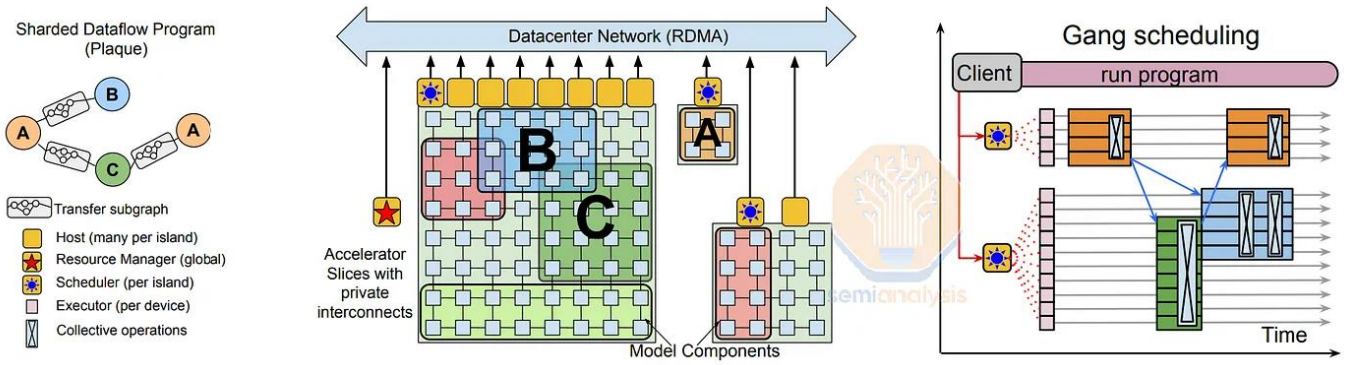
图表7：跨 DC 协同训练给网络带来挑战



资料来源：华为《迈向智能世界白皮书 2024》，国联证券研究所

为了实现多园区训练，Google 目前使用功能强大的分片工具 MegaScaler，它能够使用 Pathways 的同步训练将一个园区内的多个 pod 和一个区域内的多个校区进行分区。在扩大单个训练工作负载所需的芯片数量时，MegaScaler 为 Google 在稳定性和可靠性方面提供了强大优势。

图表8：谷歌 Pathways 训练系统



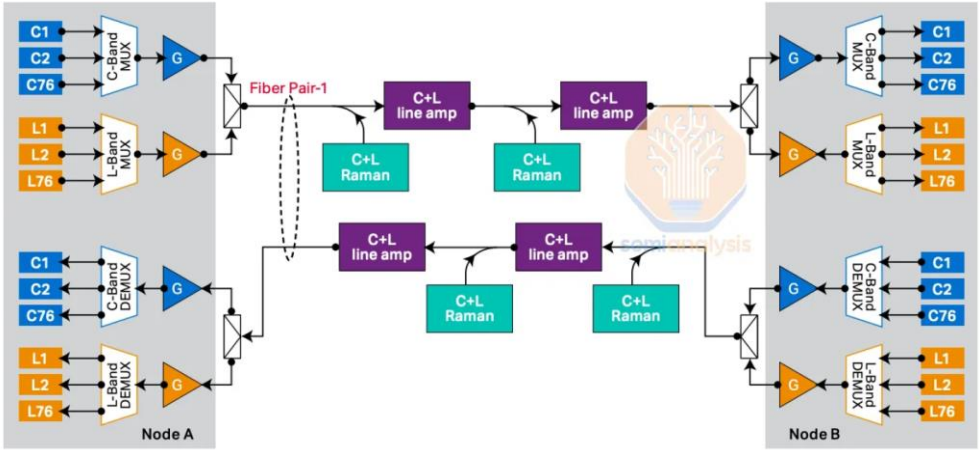
资料来源：semianalysis ,Google,国联证券研究所

未来在多园区、多区域集群上训练的模型将达到 100T+的数量级。在不久的将来，我们认为，一个区域内的园区站点间的带宽增长到 5Pbit/s 左右是较为合理的假设，而不同区域之间的合理带宽是 1Pbit/s。

1.3DCI 互联方案和市场空间分析

更大的带宽可以通过更高阶的调制方式或者采用 DWDM（密集波分复用）来实现。与使用 PAM4 的强度调制直接检测方案（IMDD）相比，DP-16QAM 的带宽增加了 8 倍。长距离传输仍然存在光纤限制，DWDM 将多种波长的光聚合到同一根光纤上，也可以用来实现更高的带宽。在下面示例中，C 波段（1530nm 到 1565nm）上的 76 个波长和 L 波段(1565nm 到 1625nm)上的 76 个波长被复用到同一根光纤上。

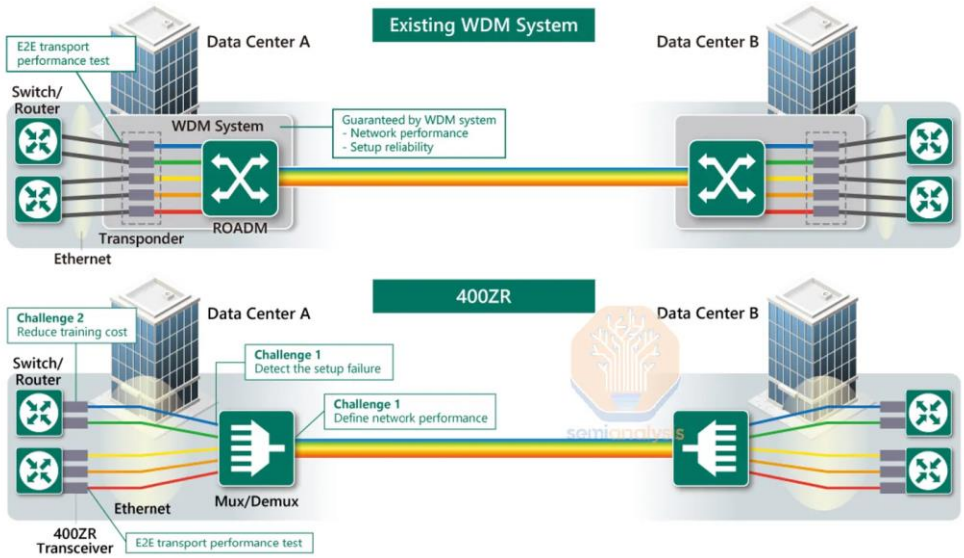
图表9：DWDM 工作原理



资料来源：semianalysis，Ciena，国联证券研究所

400G ZR 相干光学技术有望在 DCI 中取代传统的波分复用（WDM）系统。相比于传统的 WDM 系统，400G ZR 系统更加简洁，主要有 MUX/DEMUX，并采用可调谐激光器的相干光模块，直接放在客户侧的交换机/路由器上。

图表10：400G ZR 和传统波分复用系统（WDM）的对比



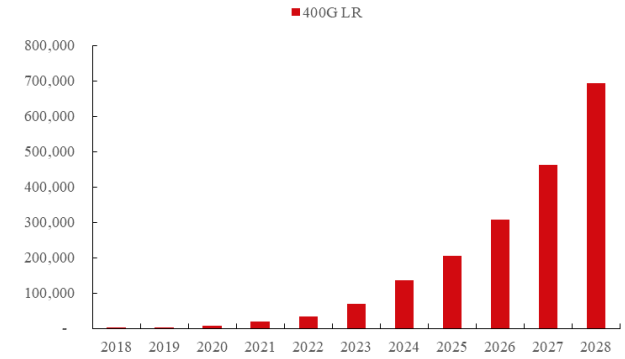
资料来源：semianalysis，anritsu，国联证券研究所

根据通信距离的不同，DCI 场景也会选择不同的产品。（1）在 IDC 园区内部，多个不同的 DC 之间互连，一般会优先选择在楼宇间布放大量光缆+LR 光模块的方式。（2）跨园区的 DCI 互联，一般选择 DWDM+ZR 光模块的方案。

根据 LightCounting 的预测，2024-2028 年，400G LR 的光模块保持增长。产品价值

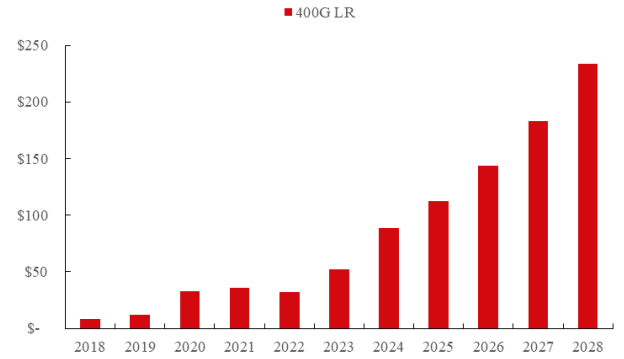
量方面，根据 LightCounting 预测，2023 年 400G LR 的价格为 760 美元，2024 年 400G ZR 的价格为 646 美元。

图表11：全球 400G LR 光模块出货量预测



资料来源：LightCounting，国联证券研究所

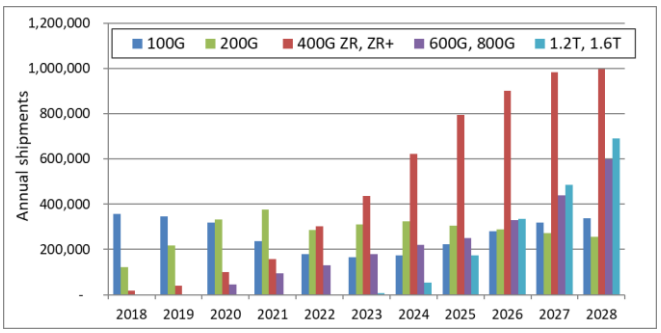
图表12：全球 400G LR 光模块市场规模预测（百万美元）



资料来源：LightCounting，国联证券研究所

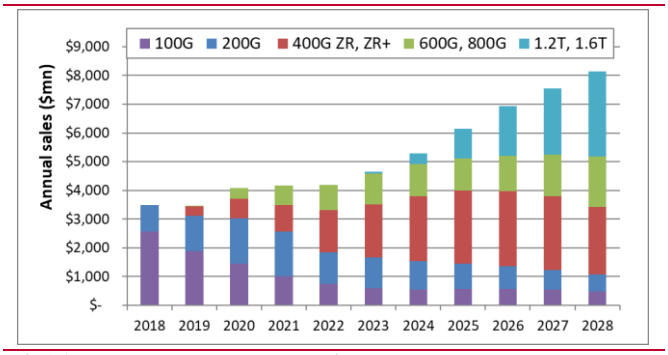
根据 LightCounting 的预测，2024-2028 年，400G ZR，ZR+、600G、800G、1.2T、1.6T 的光模块保持增长。产品价值量方面，根据 LightCounting 预测，2023 年 400G ZR 的价格为 3230 美元，2024 年 800G ZR 的价格为 4800 美元。

图表13：全球 WDM 光模块出货量预测



资料来源：LightCounting，国联证券研究所

图表14：全球 WDM 光模块市场规模预测（百万美元）



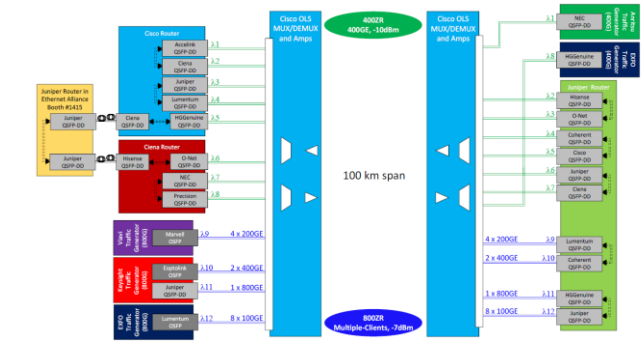
资料来源：LightCounting，国联证券研究所

2. 投资建议：优先看海外 DCI，长期看国内 DCI

2.1 国内主要的 DCI 厂家

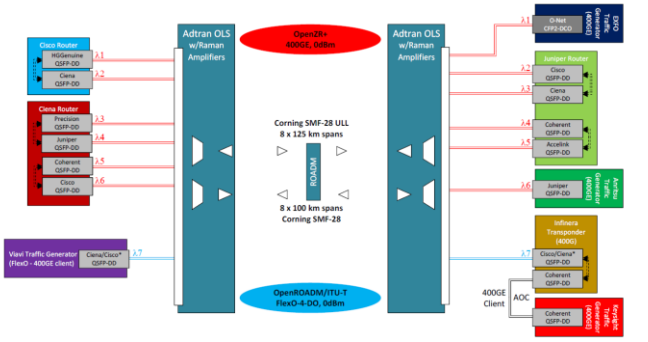
在 OFC 2024 上，OIF 组织了 400G/800G ZR 的演示。（1）在单条直连的 DCI 场景，国内参与的厂商有：海信宽带、光迅科技、华工科技、新易盛。（2）多跳长距离直连的场景，国内参与的厂商有：光迅科技、华工科技。

图表15：OFC 2024 上 OIF 的单跳 400G/800G ZR demo



资料来源：《400ZR, OpenZR+, 800ZR, OpenROADM/ITU-T Interoperability Demo OFC 2024》，国联证券研究所

图表16：OFC 2024 上 OIF 的多跳 400G/800G ZR demo



资料来源：《400ZR, OpenZR+, 800ZR, OpenROADM/ITU-T Interoperability Demo OFC 2024》，国联证券研究所

同时，中际旭创在 OFC 2023 现场演示了 400G ZR 和 400D ZR+ QSFP-DD 相干光模块。德科立作为 Ciena、Nokia、Infinera 的供应商，在 2023 年完成 400G 长距离相干模块的研发并给客户送样。

2.2 建议关注 DCI 产业链和 400G/800G ZR 供应商

海外科技巨头积极布局多 DC 分布式训练，我们认为 AI 算力部署对网络的需求正在向 DCI 场景扩散，有望带动 DCI 市场的高速增长。我们建议关注：国内 OTN 厂商：中兴通讯、烽火通信、光迅科技；有 400G/800G ZR 产品布局的德科立、中际旭创、新易盛、华工科技；铌酸锂调制供应商：光库科技。

3. 风险提示

AI 产业发展不及预期风险。若 AI 发展不及预期，则可能影响 DCI 建设的需求，进而影响相关设备和相干光模块的采购。

算力需求不及预期风险。若算力需求不及预期，则可能影响 DCI 市场的增速。

技术发展不及预期风险。若 DCI 技术发展不及预期，则可能影响 DCI 建设的需求。

市场竞争加剧风险。若 DCI 市场竞争加剧，则可能影响相关厂商的利润率。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

评级说明

投资建议的评级标准		评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即：以报告发布日后的 6 到 12 个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A 股市场以沪深 300 指数为基准，北交所市场以北证 50 指数为基准；香港市场以摩根士丹利中国指数为基准；美国市场以纳斯达克综合指数或标普 500 指数为基准；韩国市场以柯斯达克指数或韩国综合股价指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅大于 10%
		增持	相对同期相关证券市场代表性指数涨幅在 5%～10%之间
		持有	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
		卖出	相对同期相关证券市场代表性指数涨幅小于-5%
	行业评级	强于大市	相对表现优于同期相关证券市场代表性指数
		中性	相对表现与同期相关证券市场代表性指数持平
		弱于大市	相对表现弱于同期相关证券市场代表性指数

一般声明

特别声明

版权声明

联系我们

北京：北京市东城区安外大街 208 号玖安广场 A 座 4 层
上海：上海市虹口区杨树浦路 188 号星立方大厦 8 层
无锡：江苏省无锡市金融一街 8 号国联金融大厦 16 楼
深圳：广东省深圳市福田区益田路 4068 号卓越时代广场 1 期 13 楼