Network Cloud-AI – 从理论到实践

来源：https://drivenets.com/blog/network-cloud-ai-from-theory-to-practice/

不久前，我们对用作人工智能集群后端网络整体结构的两种不同架构进行了独立的实验室测试。此次测试由 Scala Computing 负责实施，专门比较了 DriveNets Network Cloud DDC（Distributed Disaggregated Chassis）架构与另一种基于以太网架构的替代方案 Ethernet Clos 的性能。

现将测试的三个主要成果总结如下：

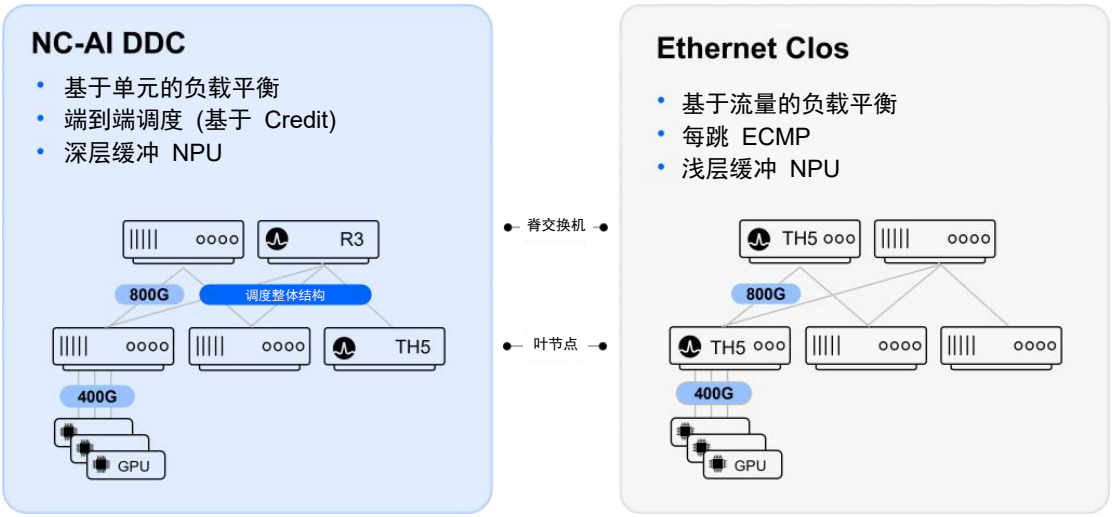
1. 在稳定状态下（即无损伤），与 Ethernet Clos 架构相比，DDC 架构的作业完成时间（JCT）提高了 10%。
2. 在 DDC 架构中，当端点（网络接口卡，NIC）受损时，性能下降只会影响受损的工作负载；而在 Ethernet Clos 架构中，服务性能下降会影响共用该基础架构的所有工作负载（由于"扰邻"现象）。
3. 使用 400Gbps 整体结构（相较于 800Gbps 整体结构）时，Ethernet Clos 的性能会降低，而 DDC 架构的性能则与整体结构的速度无关。

下面详细介绍测试方法以及三个不同阶段的结果。

网络架构测试方法

以下架构接受了测试：

1. **Network Cloud-AI DDC（NC-AI DDC）：**2000x400Gbps 端口整体结构架构。该架构采用 Accton 在新型白盒中使用的博通 DNX 芯片组（具体包括用于叶节点的 Jericho3-AI 和用于脊交换机的 Ramon3），每个白盒支持 18 个 800Gbps 客户端接口。
2. **Ethernet Clos：**2000x400Gbps 端口以太网整体结构架构（叶节点和脊交换机均采用博通 Tomahawk 5）。选择 Tomahawk 5（TH5）芯片组作为 ASIC 高端第二层交换机的示例。类似的 ASIC 也可从其他平台获得。



测试采用 2000 个 GPU 的设置，每个 GPU 都连接到一个 400Gbps 以太网端口。

测试使用的工作负载是机器学习训练工作负载。具体而言，接受测试的流程是通过聚合以太网的远程直接内存访问（RDMA）和 NCCL（英伟达集体通信库）双 B 树在 RoCE 上进行集体缩减。

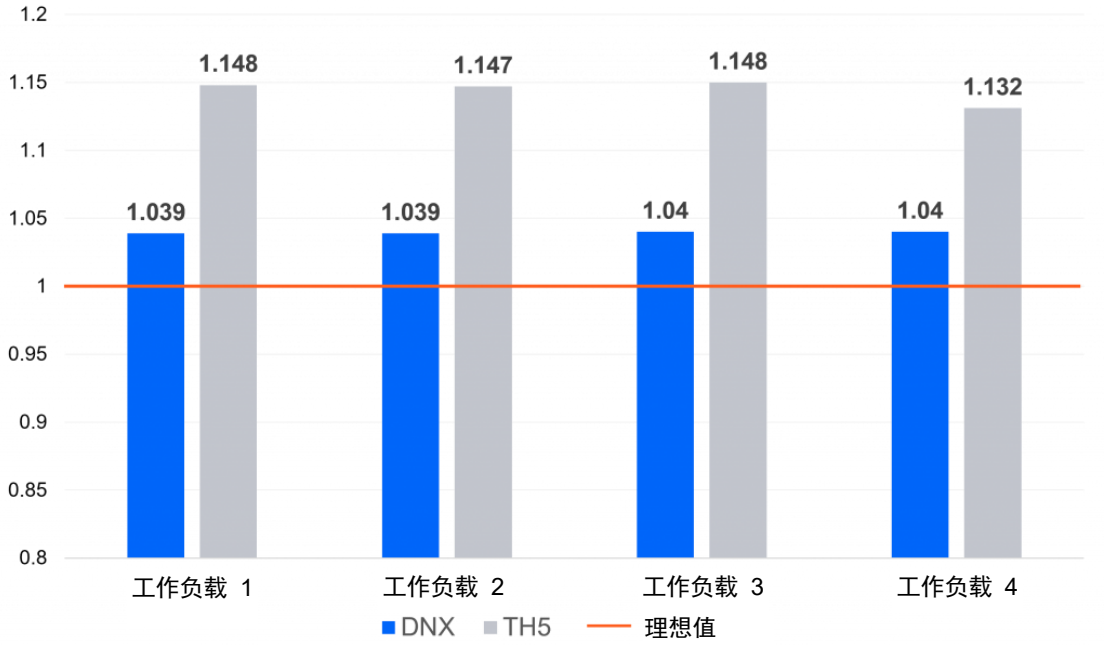
四个工作负载并行运行并在 2000 个 GPU 之间平均分配，以模拟人工智能云架构等多租户环境。此类环境需要在一个平台上支持多个客户的独特工作负载，并维护数据主权、安全性等多租户准则。

将测量结果与理想的理论结果进行比较，后者被设定为参考平面，其值为“1”。这个理想的结果是全网状、无阻塞架构的预测结果，其中所有 GPU 直接连接到所有 GPU，线路上不存在任何延迟或性能下降。（这是一种理论情景，用于设定基准。）

测试分为三个阶段：

1. **稳定状态阶段**，即在设置中不引入任何损伤，使用 2000 个 GPU 之间的全对全集合测量架构的 JCT 性能。
2. **受损阶段**，即将与四个工作负载之一相关的网络接口卡（NIC）的吞吐量减少 50%。这一阶段测试的目的是模拟“扰邻”环境，并测量每个工作负载的性能受到的影响。
3. **整体结构速度阶段**，即与类似设置的性能进行比较，其中整体结构线路速度从 800Gbps 变为 400Gbps。这一阶段测试的目的是测量 Ethernet CLOS 解决方案中跨 ECMP 组散列对 JCT 性能的影响，并与基于 DDC 单元的分布进行比较。

稳定状态阶段测试结果



对两种架构的工作负载 JCT 进行测试并测得以下结果：DDC/DNX 和 Clos/TH5

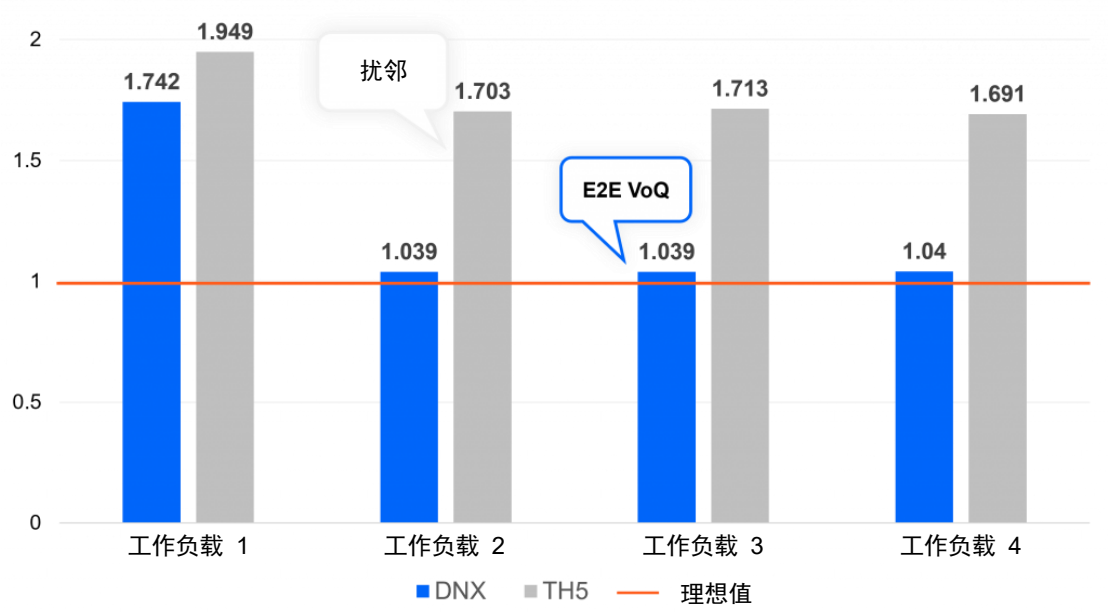
由于采用了无损、可预测和基于单元的整体结构，基于 DNX 的 DDC 架构显示出接近理想水平的 JCT 数据。而基于 TH5 的 Ethernet Clos 架构显示的数据比 DDC 高出约 10%，比理想性能数据高出约 14%。

这个结果意义重大；在较长的时间内（即稳定状态），JCT 数据提高 10% 意味着少用大约 10% 的 GPU 即可实现相同的结果。这一点有助于大大降低大规模人工智能集群的构建成本，因为网络成本约占整个设置的 10%，因此几乎能够“收回成本”。

受损阶段测试结果

这个阶段测量的是扰邻效应对两个架构的影响。这个指标对于多租户环境非常重要，例如多个企业工作负载共用一个 GPU 集群和互连整体结构进行训练或推理的情况。

测试表明，将一个工作负载（共四个）的网络接口卡（NIC）吞吐量减少 50% 以后，Ethernet Clos 架构对共用相同基础设施的所有工作负载都会产生线性影响。其原因在于 PFC（基于优先级的流量控制）机制的工作方式。当一个 NIC 出现拥塞时，它会向与之连接的叶节点交换机发出“停止”信号。该交换机随后会将这一措施反向传播到整个网络，从而导致吞吐量下降，进而影响网络中的所有连接设备。



另一方面，得益于端到端虚拟输出队列（VOQ）机制，DDC 架构的性能下降仅限于受损的连接。

当一个或多个 GPU 的性能受到同一节点上其他 GPU 活动的负面影响时，就会出现 GPU 扰邻现象。导致这种情况的原因很多，例如网络资源争用或 NIC 性能下降。

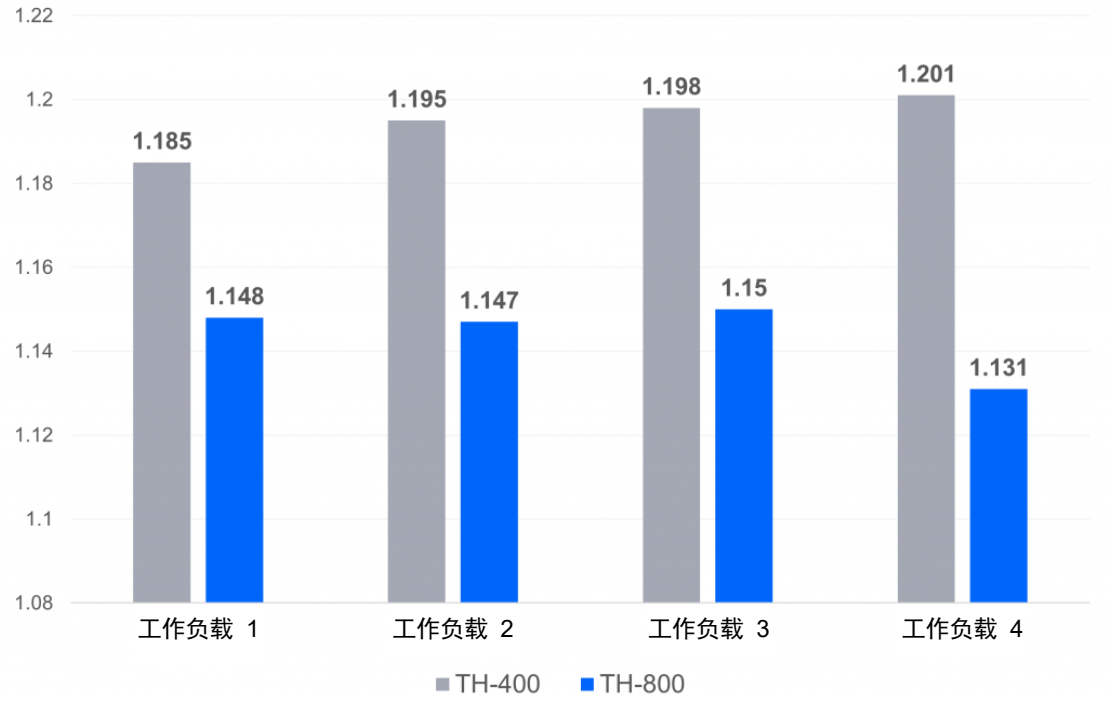
测试结果表明，Network Cloud-AI 针对 GPU 扰邻场景进行了优化，能够确保并行运行的其他人工智能作业不会受到性能影响。

整体结构速度阶段测试结果

最后，整体结构速度在 400Gbps 和 800Gbps 之间变化。

由于 Network Cloud-AI DDC 架构采用基于单元的整体结构，因此整体结构线路速度不会影响整体结构性能，客户可根据自己的偏好选择 400Gbps 或 800Gbps 整体结构。

而在 Ethernet Clos 架构中，400Gbps 整体结构需要两倍的端口数，因此需要两倍的 ECMP（等价多路径路由）组规模，从而导致性能下降，如下图所示。



Network Cloud-AI 速度更快、更可靠也更具成本效益

本次独立测试再次证实了多个实地部署的数据。根据这些数据，Network Cloud-AI 解决方案可获得最高的 JCT 性能，从而更快地训练和部署大规模人工智能模型。这个结论适用于稳定状态和受损环境，能够确保在多租户环境中实现最高性能。灵活的整体结构速度还有助于实现经性能验证的 400Gbps 至 800Gbps 迁移过程。