Scala Computing

Drivenet 模拟外部报告

修订记录：V1 - 2023 年 11 月 14 日

概述

通过介绍一组网络模拟，本文件旨在探索如何在大规模机器学习训练集群中使用基于 DNX 的解耦解决方案。这些模拟使用二级网络来演示 NIC 背压导致的“扰邻”耦合效应。工作负载中使用了全对全集合。

模拟有助于了解 DNX 和传统的分组交换机解决方案：

- 基于博通 DNX ASIC 的解决方案与传统多级分组交换机在 ROCE 工作负载上进行机器学习训练的实际对比。
- 精确识别和量化 DNX 与工作负载行为相比的优势。
- 数千个端点的不同集群规模的结果。
- 在集群规模较小、终端速度较慢的情况下，与物理 POC 测量结果具有合理的相关性。

Scala 平台可帮助客户仔细审查结果，或者重新运行拓扑或工作负载的各种变化。大多数拓扑变化和部分工作负载变化都可以通过 Scala 图形用户界面进行控制，而无需进行任何模型开发或编码。

明确的模拟目标

1) 进行评估

评估机器学习训练集群在特定整体结构下（尤其是博通 DNX）的性能。

2) RDMA 传输模型的精度

使用 ROCE RDMA 传输模型代表具有实地经验的实际商业市场设备，以便模型捕捉网络、队列、调度和主机效应。

3) 纳入模型并行和数据并行工作负载属性

使用考虑并捕捉与模型并行和数据并行阶段相关的不同模式和行为的机器学习训练工作负载模型。

4) 重点关注可分区集群，即多个作业

模型、工作负载和指标将支持在更大物理集群的分区中执行多个独立作业。

5) 评估部署亲和性及缺乏亲和性的影响

鉴于不受约束的作业部署高度理想化，分析作业部署的物理亲和性（本地性）和缺乏亲和性对单个作业指标的影响。

6) 对网络损伤进行模拟和量化

考虑并模拟现实生活中可能存在的网络损伤。量化每一种损伤对作业指标的影响，以及多种损伤的复合影响。我们认为应包括以下损伤才能成功：

- 多路径上不均匀的流量分布
- In-cast 模式
- 覆盖端点 I/O 子系统的 NIC 背压。

7) 在流量和布局模型中考虑并模拟 GPU 分组

异构机器学习训练终端系统倾向于将 GPU（一般为 8 个一组）与其自身的本地链接（例如英伟达 NVlinks）组合在一起，本地互连会吸收分组内的本地流量需求。模型中的部署亲和性会尊重这种分组，而专门针对集合的流量需求必须模拟这种带宽分层，其中以太网整体结构用于跨组流量。

项目方法

- **测试结果基于绝对时间和时间之间的比率**
- **测试结果可见性包括带宽、缓冲区使用情况、控制报文、链路阻塞以及总体、单个或最小/最大级别的分布均匀性**

测试结果

工作负载：所有性能测试均采用机器学习训练工作负载，并测量完成工作负载所有数据传输所需的时间。我们将需要交换的数据总量称为“集合规模”，虽然交换并不总是以集合的形式进行。大多数测试都是在集群各分片之间交换嵌入表结果。我们重点关注工作负载的这个部分，因为交换的数据随集群规模的平方而增长，因此构成限制性能的关键因素。

指标：我们没有报告完成的绝对时间，而是将其归一化为模拟的绝对时间与通过和网络端口速度相同的全网状点对点链路完成相同数据传输所需的理想时间之间的**比率**。例如，数值越接近 1 就表示结果越好。数值越大意味着训练时间越长，而且可能无法根据集群规模对训练时间进行线性调整，因为网络时间可能会主导训练时间。

这**一理想比率**以尾分布的形式记录在 Scala 分析平台中，具体包括平均值、P50、P99、P99.9 和 P100。运营商可能会关心 P99.9，或者绝对最坏情况下的 P100。为清晰起见，本报告的表格只显示平均值一栏和 P100 一栏。

基于比率的指标与拓扑、速度和数据大小无关。

传输：所有数据传输都使用 ROCE RC（可靠连接）服务，工作负载将 WQE（工作队列条目）排入发送队列。ROCE RDMA 状态机、PCIe 主机逻辑、主机端调度程序和链路端调度程序都扩展到了 800G 的速度。使用的 ROCE 有效载荷为 4 千字节。

所有 ROCE 流量都通过无损级以太网在所有网络链路上运行，并依靠 PFC 进行背压。连接设置包含在模型中。

512 节点集群测试拓扑

T1) Jericho3/Ramon3 DDC - 512 x 400G，1.12 倍速。

16 台 Jericho3 Leaf 节点交换机，32 个 400G 以太网端口连接 GPU，144 个 100G 整体结构链路连接 Spine 交换机。

9 台 Ramon3 Spine 交换机，每台 256 条 100G 通道。

T2) TH5 - 512 x 400G

8 台 TH5 Leaf 节点交换机，64 个 400G 端口连接 GPU，32 个 800G 链路连接 Spine 交换机。

4 台 TH5 Spine 交换机，64 个 800G 端口。

测试案例 - 交错全对全工作负载

TH5 对比 DNX，512 节点，512MB 集合大小，无损伤，400G

测试：

测试 1) 512 节点 TH5，ECMP，无损伤，拓扑 T2（2 级）。

测试 3) 512 节点 DNX，无损伤，拓扑 T1（2 级）。

测试结果

这是一次基线测试，因为不存在任何类型的损伤，我们看到了预期中的流畅性能。

TH5 512 节点无损伤 - 2 级				
测试#	交换机	EcmpHash	配置	JCT
1	TH5	ECMP	512 节点 TH5，ECMP，无损伤	1.039

DNX 512 节点无损伤 - 2 级				
测试#	交换机	EcmpHash	配置	JCT
3	DNX	不适用	512 节点 DNX，无损伤	1.038

摘要：

在小型拓扑结构中，如果没有 NIC 造成的损伤，DNX 和分组交换机的性能相当，并且非常接近理想状态（相差 4% 以内）。全对全工作负载能够缓解可能出现的 ECMP 热点，而从不穿越上行链路的部分本地流量则为 ECMP 路径提供了富余的带宽。

TH5 对比 DNX，512 节点，512MB 集合大小，RX NIC 损伤

测试：

测试 5) 512 节点 TH5，RX 损伤为 PCIe 峰值带宽的 50% - 拓扑 T2

测试 7) 512 节点 DNX，RX 损伤为 PCIe 峰值带宽的 50% - 拓扑 T1

这两次测试在小型集群中引入了 NIC 损伤。众所周知，ROCE NIC 一旦跟不上到达速率就会向 Leaf 节点交换机发送 PFC。原因包括 NIC 瓶颈和主机平台 Root Complex 瓶颈，还可能因为 GPU 直接配置中使用的 PCIe 交换机瓶颈。在 GPU 直接配置中，NIC 通过相同的二级 PCIe 总线将流量导向 GPU 和主机。

我们将 NIC 损伤模拟为 NIC DMA *写入方向*的带宽速率限制。损伤 X% 意味着写入带宽被限制在 PCIe 峰值带宽的 X%。例如，PCIe gen5 损伤 50% 意味着写入速率被限制在 0.5 * 500 Gbps=250 Gbps。通过交换机产生的背压刚好导致 NIC 的传输速率超过其他类似 NIC 的接收速率。**RX 损伤为 PCIe 峰值带宽的 50% 在最佳情况下会将网络带宽限制为网络速率的 62.5%。**

测试结果探讨了指标下降是否与损伤导致的带宽损失相匹配。我们还报告了实现最佳结果是否需要按工作负载进行调整，因为通常情况下 DNX 无需调整即可获得最佳结果。

我们预计会先从 NIC 收到 PFC，然后才会从 Spine 交换机收到 PFC，这样才与 NIC 拥塞传播受损的情况一致。

TH5 512 节点单个作业，NIC 受损 - 2 级				
	交换机	EcmpHash	配置	JCT
5	TH5	不适用	512 节点 TH5，PCIe RX 损伤 导致网络速率至少降低 37%。	1.97

注：报告的结果始终与理想和无损伤的情况相比较。

DNX 512 节点单个作业，NIC 受损 - 2 级				
	交换机	EcmpHash	配置	JCT
7	DNX	不适用	512 节点 DNX，PCIe RX 损伤 导致网络速率至少降低 37%。	1.74

注：报告的结果始终与理想和无损伤的情况相比较。

摘要：

将 NIC 损伤设置为 50%，从而将 PCIe 写入带宽限制为 250G PCIe 峰值带宽的一半。

XON 时间 = 0.5 * 500 G / 400G = 0.625　　XOFF 时间 = 1 - XON 时间 = 0.375

因此，网络 HOL 受阻于 50% NIC 损伤时的预计 XOFF 时间为 0.375。

由于向所有目的地的传输都是完全交错进行的，我们预计 HOL 阻塞率较低。也就是说，在一条路径上受到早期拥塞背压节流的 NIC 会降低所有路径上的传输速率，从而减少拥塞的扩散（考虑到模式的规律性）。

模拟结果表明会先从 NIC 收到 PFC，然后才会从 Spine 交换机收到 PFC，这与 NIC 拥塞传播受损的情况一致。

2K 节点集群测试拓扑

这些拓扑的目的是分两个阶段重复更大规模的测试。

T7) Jericho3/Ramon3 DDC - 2k x 400G

> **64 台 Jericho3 Leaf 节点交换机，32 个 400G 以太网端口连接 GPU，144 个 100G 通道连接整体结构 FE。**
>
> **36 台 Ramon3 整体结构交换机，每个方向 256 个 100G 通道。**

T8) TH5 - 2k x 400G

> **32 台 TH5 Leaf 节点交换机，64 个 400G 端口连接 GPU，32 个 800G 通道连接 Spine 交换机。**
>
> **16 台 TH5 Spine 交换机，64 个 800G 端口。**

测试案例 - 400G 交错全对全工作负载

2k 节点 TH5 对比 2k 节点 J3 DNX，4 个作业，每个作业集合大小 2GB，无损伤，400G

测试：

> **测试 12) 2k 节点 TH5，ECMP，随机放置，无损伤 - 拓扑 T8**
>
> **测试 14) 2k 节点 J3 DNX，随机放置，无损伤 - 拓扑 T7**

在这两次测试中，我们将集群规模扩大到 2k 个节点。

多机架配置被划分为多个独立的训练作业。出于将这些大型集群与单机箱集群进行对等比较的实际考虑，我们决定呈现将集群划分为四个作业的结果，每个作业 512 个节点。

在操作上，最好不要对作业使用的节点施加任何物理亲和性或邻近性限制。我们称其为**随机放置**，并将其用于本报告中显示的结果。*请注意，随机放置表示节点的放置是随机的，但每 8 个为一组。*

我们测量并报告每个作业的理想比率，不仅衡量该数值与 1 的接近程度，还衡量各个作业的结果均衡性，以及是否存在“扰邻”类型的耦合效应。

以下是在没有任何损伤的情况下得出的结果：

TH5 (ECMP) 2k 节点，多个作业 (4x512)，随机放置， 无损伤 - 2 级					
测试#	交换机	EcmpHash	配置	作业	JCT
12	TH5	ECMP	2k 节点 TH5，ECMP，多个作业（4x512），随机放置，无损伤	1	1.148
				2	1.147
				3	1.148
				4	1.132

DNX J3 2k 节点，多个作业 (4x512)，随机放置， 无损伤 - 2 级					
测试#	交换机	EcmpHash	配置	作业	JCT
14	DNX	不适用	2k 节点 DNX，多个作业（4x512），随机放置，无损伤	1	1.039
				2	1.039
				3	1.04
				4	1.04

摘要：

结果表明，我们获得了均衡的**理想比率**值。DNX 的指标与单个机箱集群的指标类似，证明将 DNX 整合到更大的集群中具备很高的可行性，而且相较于使用 DNX 专用基础设施不会造成任何损失。至于 TH5 ECMP，即使在没有损伤的情况下，指标也会下降。我们认为原因在于使用上行 ECMP 链路的流量比例较高，因此富余的带宽较少。由于 2k 节点是 2 级拓扑结构的极限，因此 TH5 设置无法提供更多的上行链路。

2k 节点 TH5 对比 J3 DNX，多个作业（4），随机放置，每个作业集合大小 2GB，RX NIC 损伤，400G

测试：

测试 16e) 2k 节点 TH5，ECMP，RX 损伤为 PCIe 峰值带宽的 50%，单个作业损伤 - 拓扑 T8。

测试 18) 2k 节点 J3 DNX，RX 损伤为 PCIe 峰值带宽的 50%，单个作业损伤 - 拓扑 T7。

在这两次测试中，我们测量并报告每个作业的理想比率，其中现实生活中的 NIC 损伤仅适用于作业 #1。我们再次衡量该数值与 1 的接近程度，并衡量各个作业的结果均衡性，以及是否存在“扰邻”类型的耦合效应。

我们使用 PCIe gen5 进行 400G 测试，并显示 NIC 接收路径的带宽损伤 50% 的结果。

本次测试中受损的作业为 #1。

TH5 (ECMP) 2k 节点，随机放置，一个 Nic 受损 - 2 级					
测试#	交换机	EcmpHash	配置	作业	JCT
16e	TH5	ECMP	2k 节点 TH5，ECMP，4x512，随机，PCIe RX 损伤导致网络速率至少降低 37%，单个作业损伤。	1	1.949
				2	1.703
				3	1.713
				4	1.691

本次测试中受损的作业为 #1。

DNX J3 2k 节点，随机放置，一个 Nic 受损 - 2 级					
测试#	交换机	EcmpHash	配置	作业	JCT
18a	DNX	不适用	2k 节点 DNX，4x512，随机，PCIe RX 损伤导致网络速率至少降低 37%，单个作业损伤。	1	1.742
				2	1.039
				3	1.039
				4	1.04

摘要：

在作业不受损的情况下，DNX 的理想比率值与不受损的单个机箱集群的指标类似，证明将 DNX 整合到更大的集群中具备很高的可行性，而且相较于使用专用基础设施即使在“扰邻”的情况下也不会造成任何损失。

读者可能会问，我们为什么要为单个作业引入损伤（鉴于整个集群使用相同的 NIC 硬件和 I/O 总线技术，而不区分具体的作业）。其逻辑是，NIC 对网络的背压是工作负载的函数（在多少 QP 上以何种顺序、何种报文和事务大小移动多少数据），也是根复合体上的工作负载和应用压力的函数。我们衡量的是一个作业是否受到可能对网络 I/O 子系统构成更大压力的另一个作业的影响。如果没有隔离就会产生“扰邻”效应，从而降低其他相邻作业的完成时间。

TH5 的结果表明，由于内部链路上的 HOL 阻塞以及与未受损作业的明确耦合，NIC 损伤会产生更大的影响。