RDMA 技术在数据中心中的应用研究

涂晓军[1] 孙 权[1,2] 蔡立志[3]

[1](中国银联股份有限公司 上海 201201)

[2](复旦大学 上海 200433)

[3](上海计算机软件技术开发中心上海市计算机软件评测重点实验室 上海 201112)

摘 要 随着云计算向数据化智能化的方向演进,数据的流转与有效利用将为业务带来核心价值。大规模深度学习、机器训练等应用是极其依赖算力的,大量的信息交互对网络提出了很高的要求,由此需要一个低时延、无丢包、高吞吐的算力网络。考察 RDMA[1] 技术在数据中心中的应用,并分析其对于未来云数据中心高性能集群计算的影响。

关键词 RDMA 低延时 大算力网络

中图分类号 TP3 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2021.03.004

APPLICATION OF RDMA TECHNIQUE IN DATA CENTER

Tu Xiaojun[1] Sun Quan[1,2] Cai Lizhi[3]

[1](*China UnionPay Co.,Ltd.*, *Shanghai 201201*, *China*)

[2](*Fudan University*, *Shanghai 200433*, *China*)

[3](*Shanghai Key Laboratory of Computer Software Testing & Evaluating*, *Shanghai Development Center of Computer Software Technology*, *Shanghai 201112*, *China*)

Abstract As cloud computing evolves towards data and intelligence, the flow and effective use of data will bring core value to the business. Applications such as large-scale deep learning and machine training are extremely dependent on computing power. A large number of information interactions place high demands on the network, which requires a low latency, no packet loss, and high throughput computing network. This paper mainly examined the application of RDMA[1] technology in data centers and analyzed its impact on future high-performance cluster computing in cloud data centers.

Keywords RDMA Low latency High throughput computing network

0 引 言

RDMA 即远程 DMA,是最早脱胎于 InfiniBand 网络[1-3] 的技术,主要应用于高性能科学计算中。随着云计算的兴起,RDMA 技术也逐渐被应用到云数据中心具有高性能要求的场景中。

当前 RDMA 的使用案例主要集中于互联网企业,尤其是高性能计算领域,相应的标准组织也在积极推动,同时也出现了提供 RDMA 专用技术服务的公司。

国际上,微软是在数据中心大规模部署 RDMA 的第一家超大规模公司[4]。FaceBook 主导的 OCP(Open Computing Project)对于推动网络的开放解耦与 RDMA 的应用也做了许多的工作。

国内来看,从 2016 年开始,阿里巴巴就投入专项研究,以改造 RDMA,提高传输性能,从网卡底层开始设计满足大规模应用的网络,基于 RDMA 网络技术的云存储和电商数据库服务器可以从容地应对峰值流量考验[5]。百度在 2014 年前后开始引入 RDMA 网络,先后部署了 Infinband 集群和 RoCEv1 集群。2015 年,百度分别在 SZWG 机房和 YQ01 机房大规模部署了 RoCEv2 集群,分别承载了深度学习、语音识别和自然

收稿日期:2020-06-09。上海市优秀学术/技术带头人计划项目(19XD1433700)。涂晓军,高工,主研领域:云计算,大数据,移动支付,项目管理。孙权,教授级高工。蔡立志,教授级高工。

语言处理等相关的机器学习任务。目前 RDMA 集群总体规模为 600 台左右，这是国内比较大的一个 RoCEv2 网络。京东人工智能研发团队在分布式的模型训练场景中，也使用了 RDMA 技术，针对模型文件的高性能传输，满足了分布式训练的需求。

RDMA 对于端到端的网络传输做了多重的优化，其定位是高性能的网络技术，所带来的效果主要体现在如下几个方面:

(1) 减轻 CPU 负荷: 通过主机侧内核旁路零拷贝以及网络对于传输控制协议的卸载，可以极大地解放主机的 CPU，从侧面提升计算效率。

(2) 拥塞快速处理: 除了端侧外，网络侧也直接参与拥塞处理，可以第一时间检测到报文的拥塞堆积，并且及时有效地进行反馈，避免报文大规模的重传。

(3) 低延时: 低延时是 RDMA 最显著的特征，主机侧的精简处理以及网络侧的拥塞及时反馈，可以有效确保时延的可预期性，提升通信的效率。

1　数据中心 RDMA 应用场景

RDMA 在网络技术中主要解决的是拥塞控制的问题，在主机侧采用了内核旁路与网卡卸载等方法降低网络通信的开销。以下是其适用的数据中心场景:

(1) 高性能 MPI 计算。并行编程结构 MPI 计算是最早使用 RDMA 的应用场景，通常应用在大型科研机构的超算中心。其程序通常使用 MPI 框架进行开发，MPI 的底层调用 RDMA 的 API 进行网络通信。MPI 计算通常在天文气象、流体力学等科学计算领域有大量的使用，但在企业级市场中普及度不高。

(2) 大数据/AI 类的应用。大数据/AI 类的应用通常涉及到海量数据的搬运与交互，兼具计算密集、网络密集与 I/O 密集的特征，因此非常适合通过 RDMA 技术进行集群优化。大数据/AI 领域常见的项目，例如 Hadoop、Spark[6]、TensorFlow[7]、Pythorch 等都已经加入了对于 RDMA 通信接口的支持。

(3) 分布式存储/数据库。分布式存储或者分布式数据库也是高吞吐数据密集型的应用。随着 SSD 以及 NVMe 技术的引入，I/O 的速度也大幅增长。Samba 文件共享系统、Ceph 分布式存储等都加入了对于 RDMA 通信接口的支持。在分布式数据库领域，有些数据库采用了计算存储分离的设计，在存储部分也会用 RDMA 进行加速，例如阿里巴巴的 PolarDB。内存数据库方面，也有相应的采用 RDMA 进行网络设计的研究，甚至是对 Redis 的 RDMA 化改造。

2　高性能 RDMA 集群架构优化

2.1　网络架构设计-流控与选路的结合

在网络架构上，SDN 通常能够实现更灵活的选路控制，而 RDMA 主要处理网络的拥塞流控，两者此前一般都独立存在。RDMA 把端到端的网络通信做到了极致，如果能结合上整体的网络视图，两者就能够实现更好的网络优化，如图 1 所示。

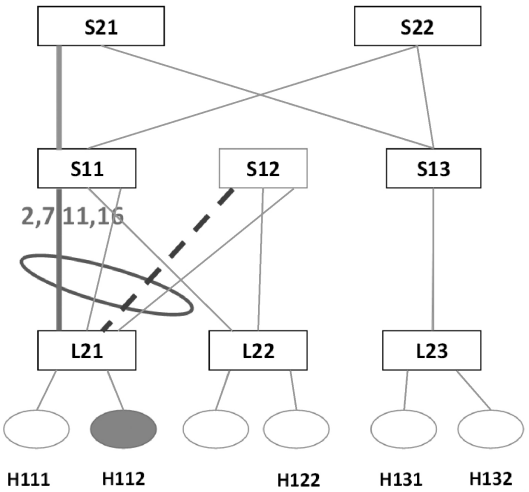

图 1　SDN 与 RDMA 的联合优化

在图 1 的 Leaf-Spine 的网络通信场景中，当一个 Spine 的出端口出现拥塞时有两种解决方法: 一种是通过拥塞控制降低 Leaf 服务的发送速率; 另一种是通过 SDN 控制器流量调度，将部分的 Leaf 流量切换到另一个空闲的 Spine 中，实现整个网络的吞吐最优。

2.2　集群应用的通信结构设计

如果以更高的视野来看，除了网络侧的高性能网络优化外，在应用方面也需要从源头对集群应用的通信结构进行优化。

例如在当前的分布式学习计算中，比较常用的是如图 2 所示的 PS-Worker 式的汇总型通信[8]。这种通信结构会出现多对一的网络流，对 PS 节点将会造成一定的压力，即使网络层面能够很好地处理拥塞，但整体的吞吐量仍然会有所限制。另外一种是 Horovod 的分布式训练，将通信结构更改为环状的通信，以避免出现多对一的网络流，造成单一节点拥塞瓶颈的出现，但会出现单点失效或者跳数增多与时延加长的性能损失。分布式的 AllReduce 的方式则是结合了汇总型与环状处理两者各自的优势，对模型进行了综合的优化[9]，如图 3 所示。

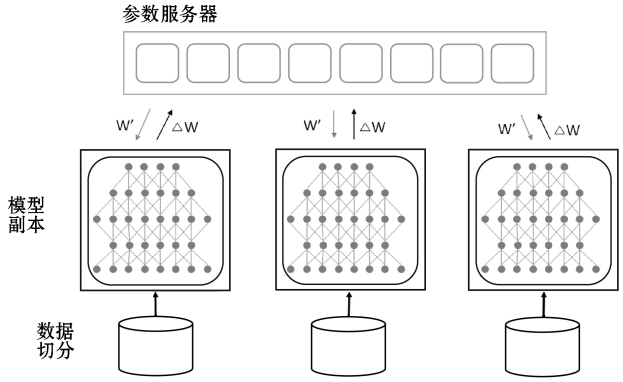

图 2 分布式 AI 任务的通信结构

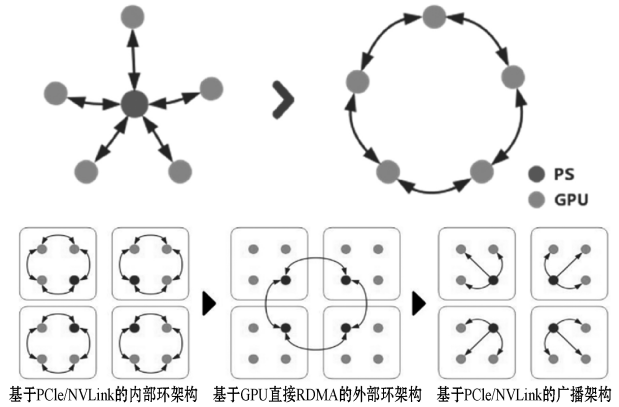

图 3 集中与环状通信结构的融合

2.3 基于 GPU 与 RDMA 的大算力集群设计

为了更好地将 RDMA 网络技术应用于金融人工智能的场景，本文设计并构建了基于 GPU 虚拟化与 RDMA 加速的云原生大算力网络集群。该系统的底层基于异构的 GPU 芯片进行算力加速，以及 RDMA 高性能网络进行数据端到端的低延时高吞吐传输。虚拟化层对 GPU 以及 RDMA 网卡的资源进行池化与虚拟化。平台层采用 Kubernetes 容器云平台，提供轻量级弹性的资源编排，为租户动态分配所需的 GPU 算力以及 RDMA 网络资源。框架层集成常用的人工智能框架与分布式的模型通信，为上层应用提供高效的建模支撑。整个平台结合了轻量级云原生平台、GPU 算力虚拟化、高性能 RDMA 网络，实现了一个面向多租户超高性能的大数据算力集群，如图 4 所示。

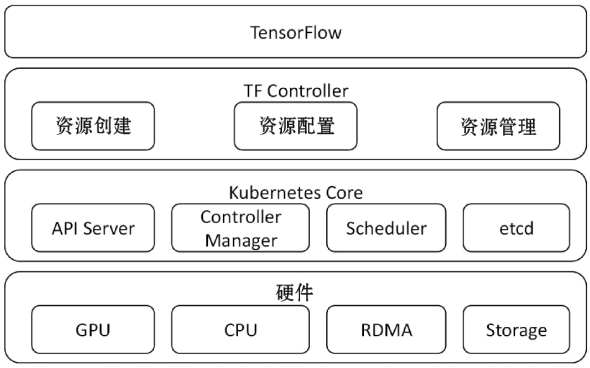

图 4 基于 GPU 与 RDMA 的大算力集群

3 RDMA 技术应用

3.1 性能调优

RDMA 对网络的流量与拥塞控制主要采用 PFC 以及 ECN 两种机制。

PFC 主要是流量控制，当交换机的入口队列出现拥塞时，它会向上游的端口发送 PFC 帧，以短暂地阻塞上游端口的发送；ECN 则是拥塞控制，在交换机的出口端打上标记，这样当接收端收到带 ECN 标记的报文时，就可以向源节点反向发送控制报文，以调整源端的发送速率。相关的关键参数为：

1) PFC 的触发阈值以及收到 PFC 帧后端口暂停的时间间隔。

2) ECN 的触发与恢复门限值，标记概率。

3) 此外为了保证非 RDMA 流量的吞吐，也需要对交换芯片的缓存分配设置合适的比例。

上述参数的设置，将会对整个网络产生的影响。通常门限值设置低，时延低，吞吐低；门限设置高，时延高，吞吐高。

对于这些参数的设置，通常需要一些经验式的参数调优，比如 Mellanox 的推荐设置便是一些静态的参数，而华为则采用了一种叫作 AIECN 或者动态 ECN 的方式，动态地根据当前的流量状况，对这些参数进行调优。

3.2 RDMA 无损网络的支持

RDMA 的目标是实现高性能的网络传输，其中需要避免因为网络丢包而引起的大规模重传，重传会导致很严重的性能损耗与开销。因此为了配合 RDMA，通常在网络侧需要做到无损不丢包。在 RoCE 具体的实现中，主要是利用了 PFC 的流量控制机制和 ECN 的拥塞控制机制。

但实现网络的无损与网络的高吞吐存在一定的矛盾，尤其是 PFC 的机制，一旦出现拥塞，它会短暂地阻塞端口的传输，并且在稍微大一些的网络中还会出现 PFC 的头端死锁，大大降低网络的吞吐。

在允许少量的丢包的实际网络中，则可以发挥网络的高吞吐能力。另一方面，在重传的机制上以及拥测控制的机制上的改进，能够尽量减少丢包，并且当丢包出现的时候，也能够以很小的代价对网络进行恢复。

即使在无损网络的实现中，通常也是尽量减少 PFC 的出现，以防止其对网络的吞吐造成阻塞，这对于 PFC 与 ECN 的参数设置就提出了很高的要求。更多

的是希望当网络出现拥塞时，ECN 的机制首先发挥作用，而 PFC 只是作为紧急情况下的一个补充手段。

3.3　大规模组网限制

RDMA 的网络技术具有超低延时、高吞吐的特性。但 RDMA 在支持大规模可扩展网络上存在一些瓶颈。其最主要的原因是 RDMA 将所有的传输层逻辑都卸载到硬件网卡上进行维护，从而大大降低主机 CPU 的处理负担，并降低了时延。但是硬件 RDMA 有连接数量的限制，通常在千量级，这对整个 RDMA 网络的可扩展规模有较大的影响，很难与大型数据中心上万量级的服务器规模相匹配。这也是为何当前最大规模的 RDMA 集群也只能在 1 000 台物理服务器左右，通常适用于机器学习训练以及分布式存储等专用集群。与之相对应，TCP 的传输状态通常都是靠 CPU 和内存进行维系的，因此只要内存足够大，其连接数可以扩展到百万级别，非常适合大规模云数据中心场景。

3.4　RDMA 实现方式选择

RDMA 对于网络侧的主要要求便是流量与拥塞的控制。从控制论的角度而言，无非就是一个负反馈的控制系统。而在具体实现中，主要涉及到三点:(1) 如何有效地检测到拥塞。拥塞的检测通常发生在交换机芯片的缓存管理中，交换芯片一般都能够有效通过当前队列中的报文数量来判断是否有拥塞发生。(2) 如何将拥塞信息有效地传播到上下游的链路上。一种方法是发送标准的 PFC 以及 ECN 的帧。另外像阿里的 HPCC 还采用了当前比较流行的 INT 技术来对当前的拥塞状态进行记录与传输，除了简单的标记拥塞事件外，INT 所携带的信息还可以包括拥塞的比例以及来回 RTT 的测量等，以便于上下游节点做更好的决策。(3) 各端点收到拥塞信息后如何进行调整，并且能最好地做到全局最优。根据所测量到的拥塞信息，能够及时有效地调整源端的发送速率，最理想的情况是调整到全局最优的速率，既不造成丢包，也不浪费带宽，另外还要能够保证流量的公平性，尤其是在多对一的网络场景中。

4　测试验证

本文采用了业界常用的 benchmark 模型 AlexNet 和 VGG16 作为测试用例，在基于 GPU 与 RDMA 的大算力集群平台之上，结合两种主流的分布式参数聚合策略 PS-worker 及 Ring Allreduce，对传统的 TCP 以及 RDMA 通信模式下的模型训练进行了相关性能测试。按照测试组合，分别构建 8 个镜像，根据不同的测试场景，利用构建好的 8 个镜像发起多次训练任务，训练过程中查看集群 pod 的创建情况以及 GPU 的利用率。参与测试的深度学习模型和参数聚合方法组合如表 1 所示。

表 1　不同的 AI 模型与参数聚合方法组合

参数名称	批量大小	模型名称	变量更新模式	远程过程调用方式
所选参数	256	VGG16	parameter_server	grpc
	256	VGG16	parameter_server	grpc + gdr
	1 024	AlexNet	parameter_server	grpc
	1 024	AlexNet	parameter_server	grpc + gdr
	256	VGG16	ring_all_reduce	grpc
	256	VGG16	ring_all_reduce	grpc + gdr
	1 024	AlexNet	ring_all_reduce	grpc
	1 024	AlexNet	ring_all_reduce	grpc + gdr

图 5 为不同网络模式下的机器学习训练速度的对比。对于 RDMA，本文主要对比 g 与 d 两组实验结果(g 代表普通的 grpc，d 代表 gdr，即 GPU Direct RDMA，它在 GPU 与 RDMA 层面做了联合的优化，数据从一个 GPU 的显存直接 RDMA 到另一个节点的 GPU 显存上)。对于 AlexNet 模型，1acd 与 1apd 的处理帧速均比 1acg 和 1apg 提升了 2 倍以上; 对于 Vgg 模型，GDR 的加速也有一定的提升效果，2vcd 比 2vcg 提升了 2 倍以上，2vpd 比 2vpg 提升了 0.3 ~ 0.4 倍。

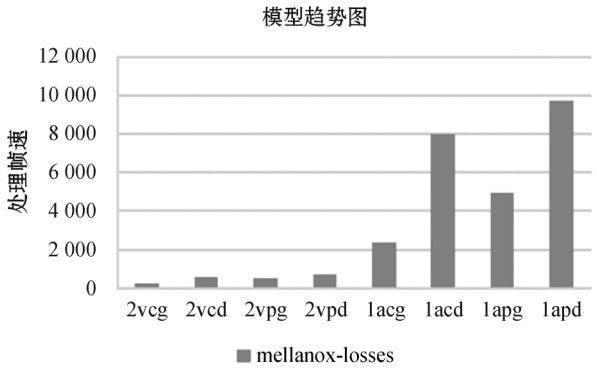

图 5　TCP 与 RDMA 通信模式下的性能对比

5　结　语

RDMA 将端到端的网络通信做到了极致，目标上是为了实现快速的远程数据传输，技术上是多重优化的结合体(涉及到主机侧的内核旁路、传输层网卡卸载、网络侧的拥塞流控)，达到的效果是低时延、高吞吐、低 CPU 损耗。同时，当前 RDMA 的实现也存在组网规模受限、配置与改造难度大等局限性。

(下转第 45 页)

4 结　语

为了实现资源联盟的总体利益最大化,本文提出了一种云联盟形成博弈算法。利用高效的联盟合并与分裂机制,使云资源提供者能以稳定的联盟结构形态满足用户需求。此外,还设计了一种基于标准化估计 Banshaf 值法的总体利益在个体成员间的收益分割机制,不仅可以保证利益分配的公平性,而且还可以使得联盟成员不会产生脱离联盟结构的动机。

参 考 文 献

[1] Varghese B, Buyya R. Next generation cloud computing: new trends and research directions [J]. Future Generation Computer Systems, 2017,79(3):849 - 861.

[2] Lu G, Zeng W H. Cloud computing survey [J]. Applied Mechanics and Materials,2014,530 - 531:650 - 661.

[3] Rochweger B, Breitgand D, Levy E, et al. The reservoir model and architecture for open federated cloud computing [J]. IBM Journal of Research and Development, 2009,53(4):1 - 4.

[4] Rochweger B, Breitgand D, Levy E, et al. Reservoir-when one cloud is not enougy [J]. Computer,2011,44(3):44 - 51.

[5] Celesti A, Tusa F, Villari M, et al. How to enhance cloud architectures to enable cross-federation [C]//2010 IEEE 3rd International Conference on Cloud Computing. IEEE, 2010: 337 - 345.

[6] Goii I, Guitart J, Torres J. Characterizing cloud federation for enhancing providers' profit [C]//2010 IEEE 3rd International Conference on Cloud Computing. IEEE,2010: 123 - 130.

[7] Bossche R V D, Vanmechelen K, Broeckhove J. Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads [C]//2010 IEEE 3rd International Conference on Cloud Computing. IEEE,2010:228 - 235.

[8] 朱匆,刘元君,彭自然,等. 移动云计算中基于协作式博弈模型的资源分配方案 [J]. 计算机应用研究,2014,31(3): 912 - 916.

[9] 李卫平,武海燕,杨杰. 基于效益博弈的云计算资源动态可协调分配策略研究 [J]. 计算机工程与科学,2016,38(1): 57 - 61.

[10] Subrata R, Zomaya A Y, Landfeldt B. A cooperative game framework for QoS guided job allocation schemes in grids [J]. IEEE Transactions on Computers,2008,57(10):1413 - 1422.

[11] Hajdukova J. Coalition formation games: A survey [J]. International Game Theory Review,2006,8(4):613 - 641.

[12] Miranda E, Montes I. Shapley and banzhaf values as probability transformations [J]. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems,2018, 26(6): 917 - 947.

[13] Ballester C. NP-completeness in hedonic games [J]. Games and Economic Behavior,2004,49(1):1 - 30.

[14] Burbidge J B, DePater J A, Myers G M, et al. A coalition-formation approach to equilibrium federations and trading blocs [J]. American Economic Review,1997,87(5): 940 - 956.

[15] Calheiros R N, Ranjan R, Beloglazov A, et al. CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms [J]. Software-Practice and Experience,2011,41(1):23 - 50.

(上接第 25 页)

随着数据中心数据量的巨量增长与算力密集度的提升,RDMA 流量在数据中心中的比重将逐步上升。它的出现具有重大意义,也对高性能的计算集群的演进具有一定的启发性。它是大数据与智能计算大规模普及的必然结果,也将成为数据智能时代的网络利器。RDMA 技术实现方面,技术复杂度与配置便捷性仍有可改进的空间。

参 考 文 献

[1] Remote direct memory access [OL]. https://en.wikipedia.org/wiki/Remote_direct_memory_access.

[2] InfiniBand [OL]. https://en.wikipedia.org/wiki/InfiniBand.

[3] RDMA over Converged Ethernet [OL]. https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet.

[4] DCQCN: Data Center QCN [EB/OL]. https://www.microsoft.com/en-us/research/project/dcqcn-data-center-qcn/.

[5] PolarDB [EB/OL]. https://www.aliyun.com/product/polardb.

[6] Mellanox accelerates apache spark performance with RDMA and RoCE technologies [EB/OL]. (2018 - 12 - 05). https://blog.mellanox.com/2018/12/mellanox-accelerates-apache-spark-rdma-and-roce-technologies/.

[7] Accelerating TensorFlow with RDMA for High-Performance Deep Learning [OL]. https://insidehpc.com/2019/03/accelerating-tensorflow-with-rdma-for-high-performance-deep-learning/.

[8] Jia C, Liu J S, Jin X, et al. Improving the performance of distributed tensorflow with RDMA [J]. International Journal of Parallel Programming,2018,46:674 - 685.

[9] Baidu's ´ring allreduce´ library increases machine learning efficiency across many GPU nodes [EB/OL]. (2017 - 02 - 21). https://www.tomshardware.com/news/baidu-svail-ring-allreduce-library,33691.html.