构建支持大模型训练的计算机系统需要考虑的4个问题

郑纬民
清华大学计算机科学与技术系，北京 100084

摘要
支持大模型训练的计算机系统有3种类型，其中基于国产AI芯片系统的生态系统不是很好，要想改变这个局面，需要做好AI编译器、并行加速等10个关键软件；基于超级计算机的系统需要做好软硬件协同设计，从而更好地服务于大模型训练。针对如何搭建大模型的基础设施，提出4点平衡设计，以确保系统的性能、可靠性和可扩展性。

关键词
大模型训练；计算机系统；超算系统；大模型基础设施

中图分类号：TP319 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024016

Four issues to consider in building a computer system supporting large model training

ZHENG Weimin
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract
There are three types of computer systems that support large model training, among which the ecosystem based on domestic AI chip systems is not very good. To change this situation, it is necessary to develop 10 key software such as AI compilers and parallel acceleration. Moreover, systems based on supercomputers require good software and hardware collaborative design to better serve large model training. This article proposes a 4-point balanced design for building the infrastructure of a large model to ensure system performance, reliability, and scalability.

Key words
large model training, computer system, supercomputing system, large model infrastructure

1 三大算力系统

目前，有3种支持大模型训练的算力系统，分别为基于英伟达公司GPU的系统、基于国产AI芯片的系统，以及基于超级计算机的系统，3种算力系统各有优劣。

基于英伟达公司GPU的系统的硬件性能好、编程生态好，但由于中美博弈引发的禁售风险，目前我国无法购买相关高端芯片。此外，英伟达GPU产品价格高，价格暴涨的同时，一卡难求。

当然，这也给国产AI芯片带来了一个难得的发展机会。基于国产AI芯片的系统已经取得了很多成果，但是其生态系统不好。要改变生态系统不好的局面，需要在编程框架、并行加速、通信库、算子库、AI编译器、编程语言、调度器、内存分配系统、容错系统、存储系统10个关键软件上下功夫。

除了GPU这种算力形态，超级计算机也可以用于支持大模型训练。部分超算中心的算力利用率并不饱和，可将空闲算力用于大模型训练。基于超级计算机的系统更要注重软硬件协同设计，从而更好地服务于大模型训练。

2 国产AI芯片的破局之路

2.1 芯片的生态系统

芯片的生态系统是指与该芯片相关的一系列硬件、软件、工具和支持服务，这些共同构成了一个完整的生态系统。如何评价芯片生态系统的优劣？以下是一些常见的评价标准。

- 软硬件支持：丰富的软硬件支持能够提高芯片的可用性和灵活性。
- 兼容性和标准：具备广泛的兼容性，支持多种硬件和软件标准，有助于降低开发成本，提高整个系统的集成性。
- 健壮性：具备强大的适应性，能够及时响应市场需求和技术变革，保持长期的竞争力。
- 安全性：具备成熟的安全性措施，能够防范各类威胁，特别是在涉及敏感数据或关键基础设施的情况下。
- 供应链：具备健全、可靠的供应链，在市场需求波动时能够保持稳定的供应。
- 开发者社区：能够提供技术支持、经验分享和创新性的解决方案，促使生态系统不断发展和优化。

芯片的生态系统涉及多个方面，将上述评价标准落实到具体场景中，评价起来也不难。例如，在基于英伟达公司GPU的系统上训练的模型，若能便捷地移植到基于国产AI芯片的系统上，那就可以说其生态系统好；如果移植过程中碰到了很多问题，需要耗费一年、两年，甚至更长的时间，那就是生态系统不好。

2.2 国产AI芯片需要做好的10个软件

目前国内已经有30多家公司推出了国产芯片，并取得了很多优秀的成果，进步很大。但用户不喜欢用国产AI芯片，最主要的原因是国产AI芯片的生态系统不好。要改变国产AI芯片生态系统不好的局面，需要做好10个关键软件，分别是：编程框架、并行加速、通信库、算子库、AI编译器、编程语言、调度器、内存分配系统、容错系统、存储系统。

(1) 编程框架

编程框架需要支持常见的大模型，如TensorFlow和PyTorch，使开发者能够无缝迁移其现有的代码。

(2)并行加速

其需要提供高效的并行计算库和工具，充分利用硬件的并行处理能力，优化系统性能。

(3)通信库

通信库要支持在多个芯片或多机之间进行数据传输和协同计算，提高数据传输效率。

(4)算子库

算子库需要提供常见的神经网络层的各种运算支持，简化开发工作，提高开发效率。

(5)AI编译器

AI编译器需要提高对主流深度学习框架的支持和优化，在异构处理器上对人工智能程序、算子库不能提供的操作等生成高效的目标代码。XLA、TVM在机器学习编译方面表现突出。

(6)编程语言

编程语言需要满足以下3个要求：一是能够支持在异构处理器上方便编写并行程序；二是能够描述底层硬件功能，与底层硬件紧密配合，充分发挥硬件性能；三是能够编写人工智能模型的基本算子(operator)，支持AI开发。英伟达的CUDA和Intel的oneAPI在这一方面提供了很好的学习范本。

(7)调度器

调度器需要设计高效的调度算法，合理分配芯片上的计算资源，进一步提高集群资源的利用率，如Kubernetes(K8S)、华为ModelArts。调度器应该在大规模系统上高效调度人工智能任务。

(8)内存分配系统

针对人工智能应用的特点，需要提供高效的内存分配策略，充分利用芯片的内存层次结构，提高数据读写效率，减少数据传输延迟。

(9)容错系统

引入容错机制，确保在模型训练过程中面临硬件故障或其他异常情况时，系统能够快速恢复模型训练，保持稳定运行。这对可靠性要求较高的应用领域非常重要。

(10)存储系统

存储系统需要支持训练过程中高效的数据读写，如检查点、训练数据等。可以考虑采用先进的存储技术，如快速闪存、持久内存等，以提高整体的存储效率。

目前，上述10个软件都有国产的，但仍存在不足，比如种类做得不够齐全、性能不够好等。这也导致用户不喜欢用国产AI芯片。针对大多数任务，用户不会因为芯片性能只有60%而有明显感知，用户感觉不好用还是生态系统存在问题。如果国产AI芯片硬件性能达到国外芯片性能的60%，只要把软件和生态做好，用户也会满意的。

3 做好软硬件协同设计，超算也可以支持大模型训练

超级计算机是计算机中功能最强、运算速度最快、存储容量最大的一类计算机，多用于国家高科技领域和尖端技术研究。目前，我国超算水平已经处于国际第一梯队，有14个国家一级超算中心，另外还有不少由地方和行业建设运营的超算中心。这些超算中心在科学计算上做得很不错，发挥了很大的作用。但是部分超算中心的算力利用率并不饱和，可将空闲算力用于大模型训练。

3.1 新一代神威超级计算机

新一代神威超级计算机采用全新的申威SW26010-Pro处理器，性能比上一代提高了4倍。它配备了96 000个处理器，共计37 440 000个核，半精度性能达到5.308 EFLOPS。此外，它的内存空间超过9 PB，

聚合带宽高达23 Pbit/s。

神威高性能计算机体系结构如**图1**所示。最左边的结构是核心组，每个核心组包含8×8个核，再加上1个主核，即65个核；每个核心组都有自己的内存控制器，配备16 GB内存，理论带宽达 307.2 Gbit/s。而中间的结构是SW26010-Pro处理器，它包含6个核心组，即配备390个核、96 GB内存，相比于SW26010有了显著提升。超节点由256个CPU组成，超节点内部的CPU之间均有直接通路，因此超节点内部的通信速度非常快；超节点之间采用裁剪网络连接，因此超节点间的通信性能比超节点内部低。

3.2 利用超级计算机做大模型训练面临的3个问题

大规模算力为扩展预训练模型提供了一个绝佳的机会。利用超级计算机做大模型训练，仍面临以下3个问题。

（1）新型硬件层出不穷

随着应用程序对算力需求的日益提升和摩尔定律的逐渐放缓，新型计算硬件，特别是异构加速处理器，成为高性能计算发展的主流。为了打破存储墙对应用程序计算性能的桎梏，新型存储器件（如固态存储设备SSD等）也被广泛应用于高性能计算机。过去，一台机器的基本配备是“CPU+存储器+硬盘”，如今，除了CPU，还涌现了GPU、TPU（张量处理单元）、NPU（神经网络处理器）、SSD等新硬件。

由于近几年美国的芯片限令，国内正在积极探索硬件的国产化之路。在高性能计算机方面，新一代的神威、天河等超级计算机均采用了国产的处理器和加速器，如**图2**、**图3**所示。

（2）新型应用程序快速发展

近年来，新型的应用程序也在快速发展，如基于张量计算的人工智能应用、基于图数据的图计算应用、面向大规模数据的大数据应用等，如**图4**所示。新型硬件的使用对软件系统的设计也是一个巨大的挑战。如何在新型硬件系统上设计与优化新型应用程序，是一个亟待解决的关键问题。

（3）整体系统工程化的挑战

软硬件都在快速发展，整体系统工程化仍面临一定的挑战。如**图5**所示，在硬件层面，新型异构高性能计算机的体系结构在计算、网络、存储等方面存在硬件复杂性；在软件层面，不规则应用程序导致节点

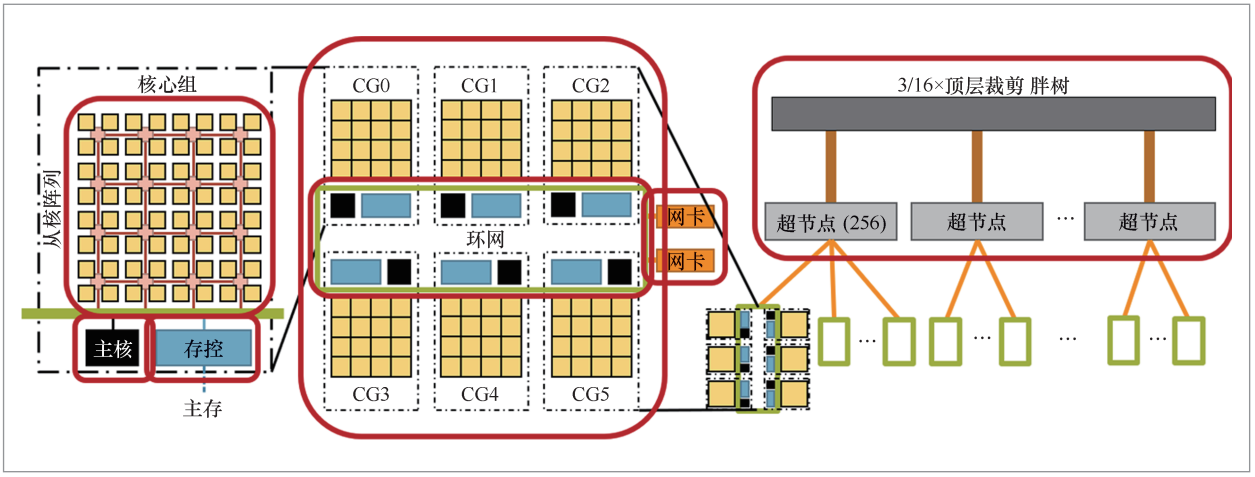

图1 神威高性能计算机体系结构

间负载不均衡、并行扩展难。因此，面临的主要挑战是如何有效适配应用程序到复杂的硬件系统。

3.3 软硬件协同设计与优化

“八卦炉”是清华大学联合北京智源人工智能研究院、阿里达摩院等开发的超大规模预训练模型，模型参数接近于人脑中的突触数量（人脑有大约1 000万亿个突触），比GPT3多1 000倍。要训练如此多的模型参数，需要一台性能更强的机器提供算力支持。“八卦炉”的模型训练是在青岛超算中心完成的。本节以“八卦炉”大模型的训练为例，介绍大规模预训练模型系统如何进行软硬件协同设计与优化。

（1）拓扑感知的混合并行策略

大模型的训练包括单节点训练和并行式训练，其中单节点训练受限于计算性能与内存，模型难以扩展，而分布式训练通过不同的并行模式，能够扩展模型规模与吞吐量。分布式训练能够加速深度学习模型的训练过程，是大模型常用的训练方式。

当前分布式训练并行策略各异，各自具有特定的通信模式。如**图6**所示，数据并行采用All-Reduce通信，能够扩展吞吐量；模型并行采用All-Gather等通信，能够扩展模型规模；专家并行采用All-to-All通信，能够同时扩展模型规模和吞吐量。简单使用单一并行策略，可能导致严重

图 2 基于申威系列处理器的神威超级计算机

图 3 基于飞腾处理器和迈创系列加速器的天河超级计算机

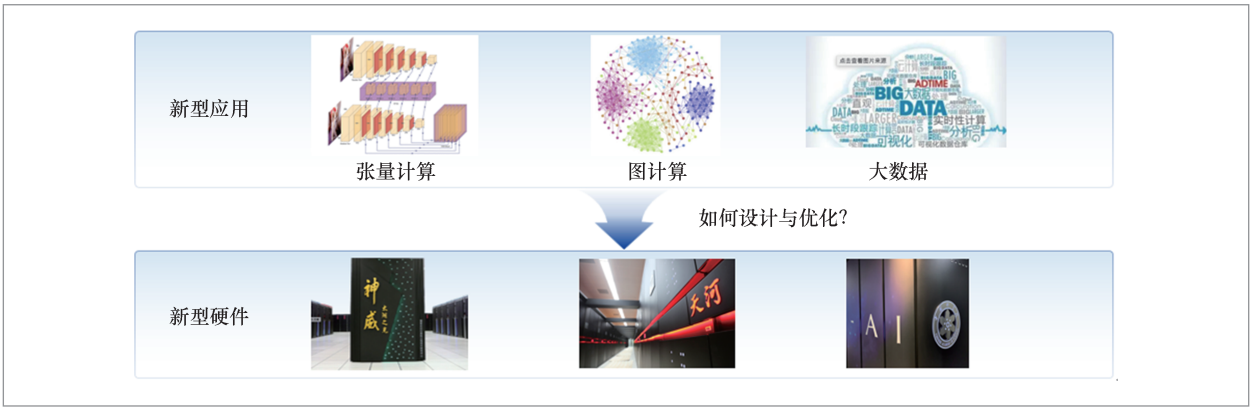

图 4 新型应用与新型硬件

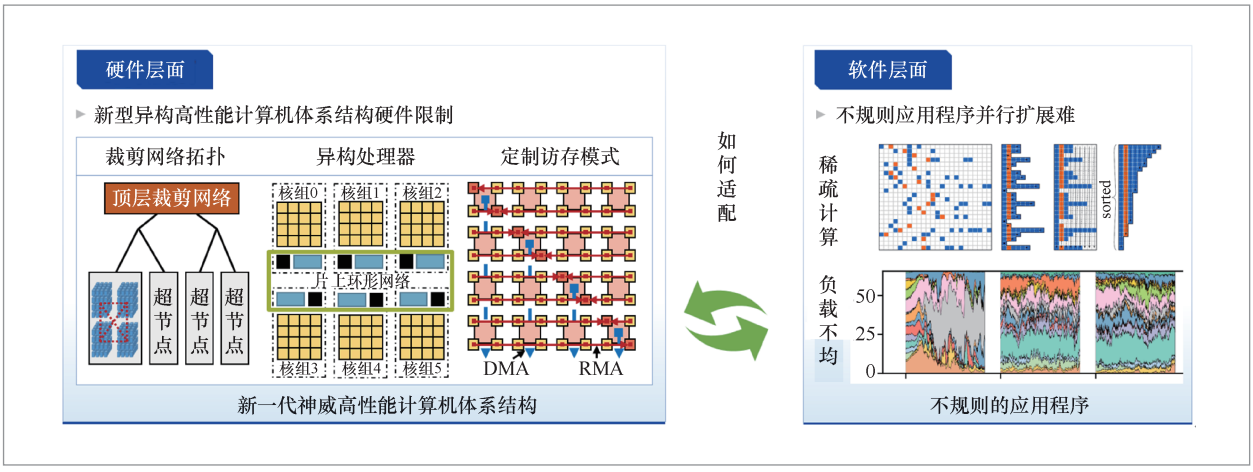

图 5 整体系统工程化的主要挑战

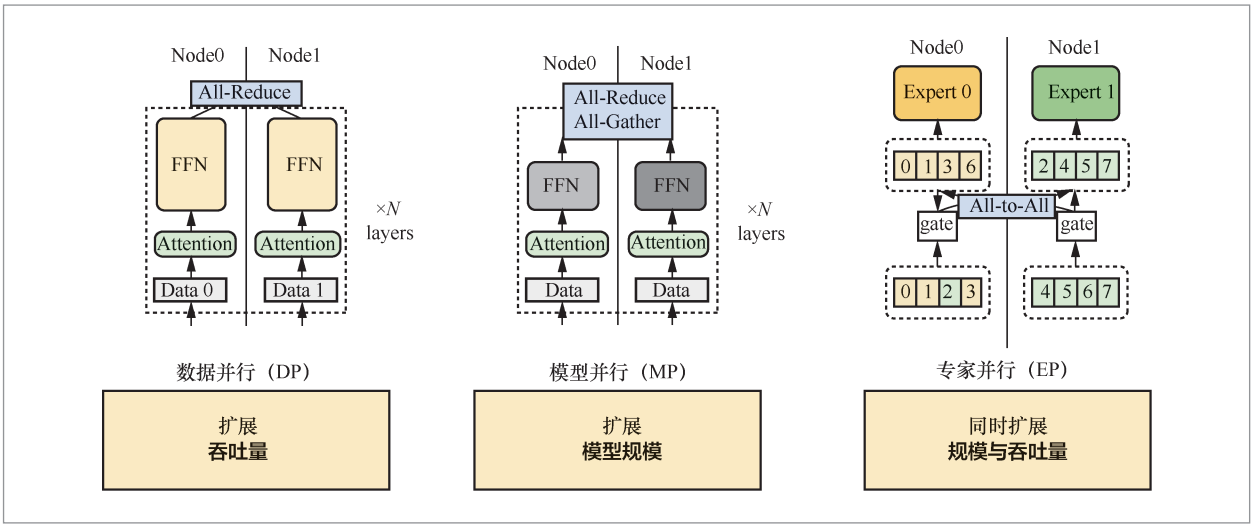

图 6 3 种典型的大模型并行训练方式

的性能降级，比如网络裁剪的影响。

如何在新一代神威的网络上高效训练模型？拓扑感知的混合并行策略是一个好的解决方法，它面向网络拓扑结构选择合适的并行模式，将通信需求更高的任务分配到高带宽网络。我们采用了一种混合数据并行与专家并行的模式。由于数据并行的通信量大、带宽高，将其分配到超节点内部；由于专家并行的通信量小、带宽低，将其分配到超节点之间。相比对称策略，拓扑感知的混合并行策略的性能提升可达到1.6倍。

（2）体系结构感知的访存性能优化

如**图7**所示，申威26010 Pro异构处理器包含6个核组，每个核组有64个计算核心和1个计算控制核心。申威26010 Pro异构处理器配备了6个存储控制器，每个存储控制器管理16 GB的内存空间。核组和存储控制器通过片上环形网络互连，访存操作经环形网络到存储控制器进行处理。环形网络和存储控制器是影响访存操作性能的主要瓶颈。

环形网络的一个潜在问题是网络拥

塞。如图8所示，当CPU的390个核同时访问内存时，大量请求被提交到环形网络，环形网络无法及时处理请求。从而导致网络拥塞，吞吐量大大降低，应用程序的访存性能会明显下降。

如何优化访存性能，可以参考以下方法：一是选择合适的拓扑结构，优化环网上的路由算法，使访存请求能够以最短路径到达目标节点，降低节点之间的通信时延；二是采用高效的请求调度算法，缩短访存请求排队和等待的时间；三是设计有效的流量控制策略，合理分配网络带宽；四是通过合并请求、压缩数据等方式减小通信负担。

（3）大规模检查点存储性能优化

大模型所需的训练时间较长，可能花费1个月、1年，甚至更长时间。机器在执行大规模训练任务时负载重，发生错误的概率高。训练过程中，经常会发生一些硬件错误和软件错误。针对训练时间较长的任务，若每3小时出错一次，出错后又重新训练，那么训练任务根本无法完成。

写容错检查点是一个很好的解决方法，在大规模训练中至关重要。容错检查点是如何发挥作用的？它在故障发生前将当前的软硬件状态存储到硬件中，存储完成后继续训练。当机器发生故障时，可直接取出硬盘中存储的软硬件状态，从这个状态继续训练。但是大模型的训练参数非常庞大，检查点需要存储大量的数据，会耗费很长时间。未经优化时，写一次检查点可能需要耗费多个小时。

如何优化检查点存储，是大模型训练面临的重要问题。一个方法是使用高性能的存储硬件，如SSD、高速磁盘；另一方法是优化文件系统算法，利用并行I/O技术，将检查点数据划分为多个块，并通过多个存储节点并行写入。经过优化，我们在青岛的超级计算机上写一次容错检查点只需20 min。

大规模预训练模型系统将以上优化方法整合到神威平台，并且做了以下调整：在算子库方面，完善了swTensor，支持混合精度算子；在深度学习框架方面，深度优化了swPyTorch，支持混合并行模式；在预训练模型方面，实现分层混合精度策略，支持负载均衡方法。

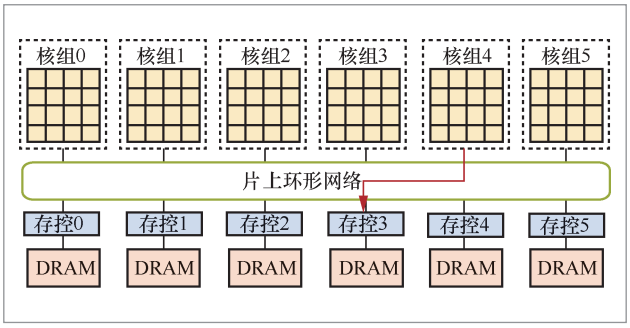

图 7　申威 26010 Pro 异构处理器

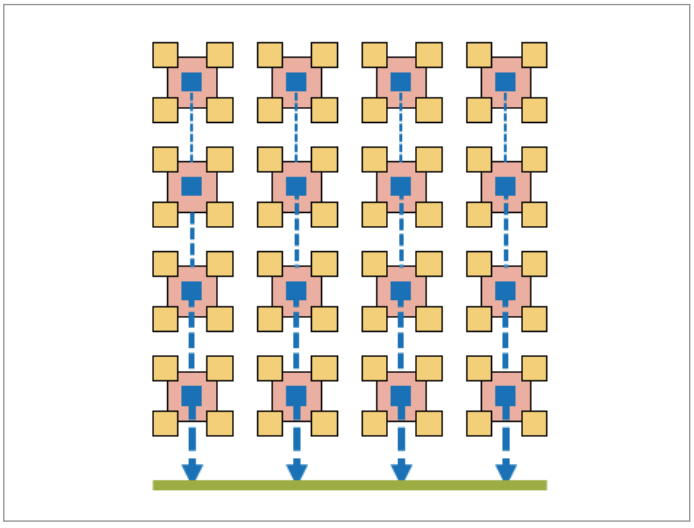

图 8　环形网络的网络拥塞

“八卦炉”大模型的训练被扩展到新一代神威超级计算机的全机规模，这是首次在国产超算平台上支持完整的预训练流程，首次支持高达百万亿规模的模型训练。通过软硬件协同设计与优化，依托超算算力资源进行大模型训练完全可以达到英伟达GPU平台的性能。除此之外，我们团队目前已经在青岛超算中心成功适配了LLaMA大模型和国内的百川大模型，半精

度和全精度的训练效果可以与英伟达GPU平台完全对齐，而成本只需要英伟达平台的六分之一。

4 大模型基础设施的4个问题

大模型基础设施通常是指用于训练和部署大规模深度学习模型的庞大计算和存储系统。这些基础设施在训练复杂模型、处理大量数据和提供高性能推理服务方面发挥了关键作用。在搭建大模型基础设施时，需要考虑多个方面的问题，以确保系统的性能、可靠性和可扩展性。

（1）半精度运算性能与双精度运算性能的平衡

设计中不仅要考虑半精度运算性能，还要考虑双精度运算能力。根据团队经验，双精度与半精度运算性能之比为1:100~1:50比较合适。根据科学智能（AI for science）和大模型训练的发展趋势，我们提出了变精度平衡设计的思想，为适应科学计算和更广泛的AI算法和应用提供了保障。

（2）网络平衡设计

网络设计不能只针对CNN类算法，还需考虑大规模预训练模型对系统的需求。大规模预训练模型需要高带宽、低时延的网络，这个网络要支持数据并行、模型并行和专家平行等模式。

假设你买了1万块卡，这些卡如何连接？最好的办法是每块卡和其他9 999块卡都建设一条直连通路。如此便需要大量的连接卡，连接卡的花费甚至高于AI卡，并且每块卡无法提供如此多的插槽用于连接。一种解决方法是每128台机器保持两两互连，以此支持数据并行、模型并行和专家平行模式。不同的并行方式需要的连接方式也不一样。

（3）内存平衡设计

大量访问内存的请求会导致网络拥塞，吞吐量大大降低，应用程序的访存性能会显著下降。多个访问内存的请求可能访问同一存控对应的内存空间，负载不均，存控需要顺序处理访存请求。因此，要做好体系结构感知的内存平衡设计。

（4）I/O子系统平衡设计

系统的本地NVMe SSD仅通过本地文件系统访问，这限制了其应用范围，可将每台服务器上的快速本地NVMe整合成应用可见的全局分布式文件系统。此外，I/O子系统可以通过增加SSD来支持检查点。

英伟达GPU的利用率为50%左右，而国产卡很多只能达到20%。为什么只有20%？就是这4个问题没解决好。大模型基础设施的4点平衡设计做得好，别人要用1万块卡，我们可能用9 000块卡就可以了。

作者简介

郑纬民（1946- ），男，中国工程院院士，清华大学计算机科学与技术系教授，中国计算机学会第十届理事长，数博会专家咨询委员会委员，《大数据》主编，何梁何利科学与技术进步奖获得者，中国存储终身成就奖获得者，享受国务院政府特殊津贴。获北京市优秀教师奖和北京市教学名师称号，获国家科技进步奖一等奖 1 项、二等奖 2 项，国家技术发明奖二等奖 1 项，2016 年获 ACM 戈登·贝尔奖。2019 年当选中国工程院院士。主要研究方向为网络存储系统，长期从事网络存储系统科学研究、工程建设和人才培养，在存储系统扩展性、可靠性和集约性等科学问题和工程技术方面，取得了国内外同行认可的创新性成果；研制的网络存储系统、容灾系统和自维护存储系统在多个重大工程中发挥了重要作用。教学方面长期讲授计算机系统结构课程，2008 年被评为国家级精品课程；已编写和出版计算机系统结构教材和专著 10 本，与合作者一起发表论文 530 余篇。

收稿日期： 2023-12-25