数据中心算力评估:现状与机遇

郭亮[1] 吴美希[1] 王峰[2] 龚敏[3]

(1. 中国信息通信研究院云计算与大数据研究所,北京 100191;
2. 中国电信股份有限公司北京研究院,北京 102200;
3. 英特尔数据平台事业群,深圳 518000)

摘要:对我国算力及算力能效的研究现状进行了综述,包括 SPEC CPU、SPEC Power、MLperf 等数据中心单服务器的算力,TOP500 和 Green500 等超算的算力,以及电能使用效率(PUE)。通过分析,提出一种数据中心算力和算效的衡量方式,并据此测算出我国当前的数据中心算力和算效的水平。

关键词:数据中心;算力;算效

中图分类号:TP393 **文献标识码**:A

引用格式:郭亮,吴美希,王峰,等. 数据中心算力评估:现状与机遇[J]. 信息通信技术与政策, 2021,47(2):79-86.

doi: 10.12267/j.issn.2096-5931.2021.02.014

0 引言

2020 年 3 月 4 日,中共中央政治局常务委员会召开会议,明确指出“加快 5G 网络、数据中心等新型基础设施建设进度”,将数据中心纳入“新基建”范畴。2020 年 4 月 20 日,国家发展和改革委员会(简称“国家发改委”)明确新型基础设施的范围,数据中心作为算力基础设施成为信息基础设施的重要组成部分[1]。

人工智能、云计算、大数据的发展离不开网络和数据中心,5G 和工业互联网的发展也离不开数据中心,甚至对数据中心的依赖程度会更高[2]。数据中心算力水平的提升将会带动全社会总体算力的提升,满足各行业的算力需求。对数据中心算力及算效进行衡量与评估将为数据中心产业发展提供重要的指导,数据中心监管部门、运营商及相关从业人员能够根据数据中心算力和算效情况判断行业发展趋势。同时,为数据中心未来算力规划和部署提供思路。

1 研究现状

以往的算力研究更加关注对超算及常规服务器算力的测试及评估,对数据中心的算力测试及评估研究则相对较少。

1.1 超算算力评估

在超算性能评价方面,普遍用计算速度,即浮点运算速度(FLOPS)来衡量超算的算力性能。国际知名排行榜 TOP 500,主要以超算系统运行 LINPACK 基准测试所能达到的最高性能对 500 个超算系统进行排名,TOP 500 排行榜每年 6 月和 11 月更新一次[3]。同时,超算的能耗问题也受到了广泛的关注。2007 年,Green 500 榜单发布,该榜单以用电效率为评估指标对 500 个超算进行排名[4]。从 TOP 500 到 Green 500,超算算力评价指标逐渐从以运算速度为主转变为运算速度和用电效率兼顾,这充分说明世界各国在先进算力竞争中从一味追求运算速度向追求算力能效进行理性转变。

1.2 常规服务器算力评估

1.2.1 SPEC CPU

SPEC CPU 是一套行业标准的针对常规服务器的 CPU 密集型基准测试套件,该测试套件由全球权威性能评估机构“标准性能评估机构”(Standard Performance Evaluation Corporation,SPEC)推出[5]。最

新版本 SPEC CPU 2017[6] 主要通过 4 个套件的 43 个测试项目,对 CPU 整点运算能力、浮点运算能力、整型并发速率和浮点并发速率进行测试。SPEC CPU 套件将会根据测试结果为 CPU 整数运算及浮点运算能力进行打分,用户能够通过打分结果直观地看出不同 CPU 的性能差异。

1.2.2 SPEC Power

SPEC 早在 2006 年就成立了 SPEC Power 工作组,目标是研究和开发可用的能源效率基准测试工具。2007 年,SPECpower_ssj2008[7] 在美国环保总署和能源使用效率协会赞助下推出。SPEC Power 委员会在 2013 年正式发布的服务器效率评级工具[8](Server Efficiency Rating Tool, SERT),由数十个被称为 Worklet 的负载组件组成,在运行时分别对服务器的 CPU、内存、存储组件进行测试。

1.2.3 MLPerf

MLPerf[9] 起源于 2018 年,是业内首套测量机器学习软硬件性能的基准套件。该基准套件囊括了一组关键的机器学习训练和推理的工作负载,代表了重要的生产级别用例。对于训练,它涵盖了图像和自然语言处理以及推荐系统和强化学习共 7 个测试项目[10];对于推理,它涵盖了图像、自然语言处理 2 种计算任务在 4 个应用场景下的测试项目。截止到 2020 年 4 月,MLPerf 已经发布了两轮训练(Training)测试结果以及一轮推理(Inference)测试结果。2020 年 7 月,MLPerf 发布了第三个版本 MLPerf Training v0.7 基准测试[11]。

1.2.4 服务器能效规范

开放数据中心委员会[12](Open Data Center Committee,ODCC)于 2019 年发布了《服务器能效评测规范》[13],该测试规范将服务器能效定义为服务器计算性能与功耗的比值,并将服务器综合能效视为电源模块效率、服务器空闲能效及服务器工作能效的加权平均数。在服务器空闲及工作能效测试过程中,该测试规范将服务器性能测试划分为 CPU、内存及存储 3 个部分,利用 Benchmark 软件对服务器各部分性能及功耗值进行记录,在不同负载条件下得到服务器空闲和工作状态功耗。

1.3 电能利用效率评估

电能利用效率(Power Usage Effectiveness, PUE)[14] 是绿色网格(the Green Grid,TGG)发布的一项用于评价数据中心能效的指标,该指标已经得到了业界的广泛认可。PUE 在数值上等于数据中心总耗电与 IT 设备耗电的比值,在整个数据中心中,IT 设备是对外提供服务的主体设备,是产生算力的主要源泉。PUE 值越小表明数据中心 IT 设备能耗占比越高,有更多电能被用于产生算力资源。尽管数据中心能效与算力具有关联,但这并不意味着提升数据中心能效水平就一定能够提升数据中心算力能效,数据中心算力能效除了与电能供给有关,还与 IT 设备的硬件性能、虚拟化技术的应用等因素有关。

数据中心算力评估与超算、常规服务器算力评估有很大不同,数据中心算力水平不仅取决于服务器的算力,同时受到存储及网络设备算力水平的影响,计算、存储及网络传输能力相互协同能够促使数据中心算力水平的提升。单独讨论服务器能力并不能反映数据中心的实际算力水平。目前,尚无针对数据中心算力评估的完整体系,构建一套算力及算效评估体系将成为当前数据中心算力研究的重点。

2 算力及算效指标

2.1 算力的定义

数据中心算力是数据中心的服务器通过对数据进行处理后实现结果输出的一种能力。在服务器主板上,数据传输的顺序依次为 CPU、内存、硬盘和网卡,若针对图形则需要 GPU。所以,从广义上讲,数据中心算力是一个包含计算、存储、传输(网络)等多个内涵的综合概念,是衡量数据中心计算能力的一个综合指标。

数据中心算力由数据处理能力、数据存储能力和数据流通能力 3 项指标决定。其中,数据处理能力又可以区分为以 CPU 为代表的通用计算能力和以 GPU 为代表的高性能计算能力。综上,数据中心算力指标包含 4 大核心要素,即通用计算能力、高性能计算能力、存储能力、网络能力。

2.1.1 通用计算能力

CPU 作为通用处理器,偏重支持控制流数据。CPU 每个物理核中大部分的硬件资源被做成了控制电路和缓存,用来提高指令兼容性和效率,只有小部分是用来做计算的逻辑运算单元(ALU)。在没有 AI 或其他高计算力要求时,CPU 可以应付得绰绰有余,在

AI 或高计算力要求时,CPU 在异构系统当中扮演和发挥重要的指挥统筹,控制核心的功能。CPU 的芯片分为多种架构,主要包含 x86、ARM 等。其中,x86 为主流架构,几乎占据全部市场份额。

2.1.2 高性能计算能力

随着近年来硅芯片逼近物理的极限和经济成本高升,摩尔定律已趋近失效,单纯使用通用处理器无法满足人工智能等新型数字化技术对高性能计算的需求。因此,GPU、FPGA、ASIC 或其他加速器支撑的高并行、高密集计算能力的异构高性能计算成为未来更复杂 AI 应用的必然选择。

(1)GPU

截至目前,全球人工智能的计算力主要是以 GPU 芯片为主,GPU 能够提供强大而高效的并行计算能力。对于海量训练数据,GPU 训练深度神经网络所使用的训练集更大,所耗费的时间更短,占用的数据中心基础设施也更少。此外,GPU 还被广泛用于云端进行分类、预测和推理,从而在耗费功率更低、占用基础设施更少的情况下能够支持远比从前更大的数据量和并发吞吐量。

(2)FPGA

现场可编程逻辑门阵列(Field Programmable Gate Array,FPGA),作为一种高性能、低功耗的可编程芯片,可以根据客户定制来做针对性的算法设计。FPGA 灵活性介于 CPU、GPU 等通用处理器和专用集成电路 ASIC 之间,在硬件固定的前提下,允许使用者灵活使用软件进行编程。近年来,随着深度学习等计算密集型业务的发展,FPGA 由于并行计算方面的优秀特性受到了互联网企业越来越多的关注,并开始研究如何在数据中心中发挥 FPGA 的优势。

(3)ASIC

特殊应用专用集成电路(Application Specific Integrated Circuit,ASIC)是为了某种特定的需求而专门定制的芯片。ASIC 芯片的计算能力和计算效率都可以根据算法需要进行定制,所以 ASIC 具有以下几个方面的优越性:体积小、功耗低、计算性能高、计算效率高、芯片出货量越大成本越低。但是缺点也很明显:算法是固定的,一旦算法变化就可能无法使用。

2.1.3 存储能力

目前,数据中心的特点是数据量爆炸性增长,数据总量呈指数上升,传得快、无篡改是存储关心的问题,亦是算力关心的问题。数据存储能力由存储容量、存储性能、存储安全三方面共同决定。数据中心存储系统不仅要有大量的现实容量,还应该具有良好的可扩展性,能根据数据量的增长提供无缝的、不停机的容量扩充。数据是具有时效性的,及时获得所需数据非常关键,对于 ICP 而言,较高的访问速度是服务质量的重要指标。对于宽带应用,存储系统的带宽要与网络带宽相适应。因此,存储系统的响应速度和吞吐率对于数据中心存储系统的整体性能非常关键。数据中心存储系统存储了企业大量的关键数据,必须保证这些数据始终是安全可用的,在任何情况下数据都不能丢失。系统应具有快速故障恢复能力,保证数据始终保持完整性和一致性。

2.1.4 网络能力

在数据中心中,网络起着承上启下的作用,将计算和存储资源连接在一起,并以服务的形式对内部及外部提供数据访问能力。带宽、延迟、丢包率都是数据中心网络关注的重点。带宽越高意味着数据中心可以具有更强的处理能力,可以完成更多的业务应用。网络延迟也是体现数据中心网络性能的重要参数,网络延迟和网络延迟的抖动越小,网络性能越好。数据在网络中是以数据包为单位传输的,丢包率是数据包丢失部分与所传数据包总数的比值,丢包率越低,网络性能越好。

2.2 数据中心算力模型(CP)

2.2.1 方法

目前,数据中心内部的服务器芯片类型以 CPU 和 GPU 这两个类型为主。前者主要用作执行一般任务,后者主要承担图形显示、大数据分析[14]、信号处理、人工智能和物理模拟等计算密集型任务。FLOPS 为每秒执行的浮点运算次数,是对计算机性能的一种衡量方式。

在计算机系统的发展过程中,曾经提出过多种方法表示计算能力,目前为止使用最广泛的是“浮点运算次数表示法”。FLOPS 的概念最早由 Frank H. McMahon[15]在其报告中提出。国内外不少文献以及服务器产品参数都采用浮点运算次数对算力进行描述,例如 Yifan Sun[16]等使用 FLOPS 作为度量标准,以评估 CPU 和 GPU 的单精度和双精度计算能力。

“浮点运算次数表示法”利用科学计数法来表达,

包含 3 种常见类型。

(1)双精度浮点数(FP64):采用 64 位二进制来表达一个数字,常用于处理的数字范围大而且需要精确计算的科学计算。

(2)单精度浮点数(FP32):采用 32 位二进制来表达一个数字,常用于多媒体和图形处理计算。

(3)半精度浮点数(FP16):采用 16 位二进制来表达一个数字,适合在深度学习中应用。

本文使用“每秒浮点运算次数”(Floating-point Operations Per Second,FLOPS)来评估数据中心的通用算力和高性能算力。同时,与 Linpack 仅关心双精度的浮点计算(FP64)能力不同,将给出双精度(FP64)和单精度(FP32)浮点计算能力算法,以便更加清晰地辅助判断数据中心适合的计算场景:用双精度浮点计算能力评估数据中心的高性能计算能力;用单精度浮点数计算能力评估数据中心的通用计算能力。

除了双精度(FP64)和单精度(FP32)之外,其他的计算精度也越来越广泛地被用于计算领域。对于人工智能来说,半精度(FP16)大有后来居上的趋势。主流的 AI 芯片和 AI 软件都已经支持半精度(FP16)用于深度学习训练。同时,INT8 也越来越多用于深度学习推理领域。在本文中,目前仅采用双精度(FP64)和单精度(FP32)两种精度衡量数据中心算力和算效,未来考虑加入更多的精度以更加全面地衡量数据中心的算力。

2.2.2 模型

数据中心算力(Computational Power,CP)的模型如下。

CP=f(通用算力,高性能算力,存储能力,网络能力) (1)

(1)通用算力计算方法

通用算力= ∑(某型号 CPU 服务器存数×该型号服务器 CPU 算力) (2)

以 Intel 主流 CPU 型号为例,理论计算能力如表 1 所示。

表 1 Intel 主流 CPU 服务器算力[17]

序号	型号	FP32
1	Intel© Xeon© Processor E7 Family	1.8 TFLOPS
2	Intel© Xeon© Processor E5 Family	1.5 TFLOPS
3	Intel© Xeon© D Processors	1.8 TFLOPS
4	Intel© Xeon© W Processors	2.4 TFLOPS
5	Intel© Xeon© Scalable Processors	3.2 TFLOPS

(2)高性能算力计算方法

高性能算力 = ∑(某型号 GPU 服务器存数×该型号服务器 GPU 算力) (3)

以 NVIDIA 主流 GPU 型号为例,理论计算能力如表 2 所示。

(3)存储能力

固态硬盘在启动速度、读写速度、质量、抗震上相比 HDD 传统硬盘有着绝对的优势,而 HDD 发展至今,在价格、寿命和数据恢复方面的成绩也是 SSD 无法取代的[19]。SSD 硬盘由于使用了高速的闪存颗粒作为物理存储资源,并且使用 PCIe 等高速传输协议/接口作为主流数据交换的物理通道,其在 IOPS 和带宽方面远优于传统的 HDD 硬盘。以企业级 PCIe SSD 卡和企业级 SAS HDD 硬盘来比较,PCIe SSD 卡的 4K 随机读的 IOPS 为 1 M 以上,而 SAS HDD 硬盘的 IOPS 为 700 左右;带宽方面,PCIe SSD 可达到 7000 Mbit/s 以上,而 SAS HDD 仅为 200 Mbit/s 左右。

存储对算力的贡献,一方面体现在高速存储对高性能计算的支撑,另一方面体现在对海量数据的存储。

表 2 NVIDIA 主流 GPU 型号算力[18]

序号	型号	FP64	FP32	FP16
1	NVIDIA Tesla P100(PCIe)	4.7 TFLOPS	9.3 TFLOPS	18.7 TFLOPS
2	NVIDIA Tesla V100(PCIe)	7 TFLOPS	14 TFLOPS	112 TFLOPS
3	NVIDIA Tesla A100(NVLink)	9.7 TFLOPS	19.5 TFLOPS	312 TFLOPS

(4)网络能力

随着 AI 训练集群规模的增大,以及单节点算力的增长,分布式 AI 集群系统已经逐渐从计算约束转换为网络通信约束。一方面,AI 计算量每年增长 10 倍[20],而数据中心网络接口过去 5 年从 1000 M 网口升级到了 10 G 或者 25 G,仅增长 10 多倍;另一方面,当前的 AI 集群系统中,当 GPU 集群达到一定规模以后,随着计算节点数的增加,由于分布式 AI 集群节点之间的通信代价的增加,可能导致集群每秒训练的图片数量不增反减。网络将成为数据中心计算、存储能力能否充分发挥的重要支撑。

2.3 数据中心算效模型(CE)

受摩尔定律的影响,CPU 的算力提升方法通常有两种,一是增加“数量”,即增加核心的数量;二是提高“质量”,即提高单核心的运算效率,即提高主频。但主频的提高并不是无限制的,会受到功耗的制约。所以,数据中心算力功耗也是一个非常重要的方面。将算力与功耗结合来看,单位功耗的算力是评价数据中心计算效果更为准确的一个指标。

本文定义数据中心算效(Computational Efficiency,CE)为数据中心算力与 IT 设备功耗的比值,即“数据中心每瓦功耗所产生的算力”(单位:FLOPS/W),这是同时考虑数据中心计算性能与功耗的一种效率,其计算公式如下。

$$CE=\frac{CP}{\sum IT\ 设备功率} \tag{4}$$

3 数据中心情况分析

3.1 近几年我国机架总体情况

数据中心总体算力水平与数据中心机架规模密切相关,数据中心机架上承载着各类服务器、存储设备及网络设备,这些设备共同构成了数据中心的算力基础。在保证上架率的情况下,数据中心机架规模越大表明数据中心能够提供的理论算力越高,算力资源供给更为充足。

图 1 反映了近几年全国数据中心在用机架规模和大型规模以上机架的变化情况,从 2016 年到 2019 年,我国数据中心在用机架规模以 30%左右的比例逐年增长[21]。机架规模的增长充分表明我国企业及用户对数据中心算力的潜在需求较为旺盛,与此同时,这种高速增长的算力需求进一步推动了我国机架规模的增长,逐年增加的算力资源将为云计算、人工智能、物联网等应用服务的开展提供重要保障。

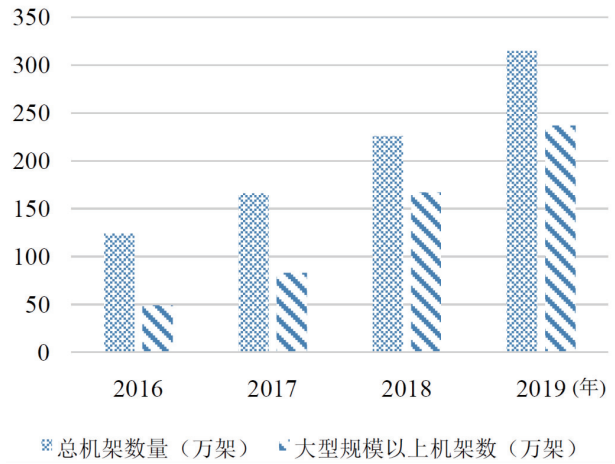

图 1 近 4 年我国数据中心机架规模

3.2 我国数据中心上架率

数据中心总体算力水平及算力能效不仅与机架总体规模有关,同时还会受到数据中心在用上架率的影响,数据中心在用上架率能够反映当前数据中心的实际算力水平及算力能耗,在评价社会总体或某地区数据中心实际在用算力时应充分考虑到机架规模及上架率。

截止到 2019 年年底,国内数据中心总体平均上架率为 53.2%。其中,北京、上海及广东平均上架率近 70%,远高于全国平均水平,核心区域大型以上数据中心上架率超过 85%。东部发达地区及一些自然资源较为充足的中西部省份上架率相对较高,东部发达地区对时效性较高的“热数据”需求较多,提升上架率有助于进一步满足这种实时的算力需求。中西部等自然资源较为充足的地区在建设能效导向型数据中心方面具有一定优势,也逐渐受到资本加持,一些实效性要求不高的“冷数据”通常可以在这些地区进行远端部署。

3.3 服务器出货量

3.3.1 CPU 服务器

根据 Gartner 的数据[23],中国(不包括港澳台地区)的 CPU 架构服务器出货量在 2015—2019 年基本呈现上升趋势,5 年复合增长率近 8%,2019 年的出货量为 340 万台左右,其中 x86 架构在 CPU 市场的占比都在 99%以上(见图 2)。

在厂商市场份额方面，Intel 市场基础庞大，在 CPU 市场市占率基本维持近 95%左右的水平。以 2019 年第 4 季度为例，根据 IDC 数据[24]显示，Intel 在全球数据中心 CPU 微处理器市场份额的占比为 93.6%，其次为 AMD 为 4.9%。

3.3.2　GPU 服务器

根据 IDC 的数据显示，全球 GPU 的出货量呈现上升趋势，在 2019 年达到了 840 万 Unit（见图 3）。

在市场份额方面，NVIDIA 为行业龙头企业。以 2019 年第 4 季度为例，根据 IDC[24]数据显示，NVIDIA 在全球数据中心 GPU 服务器市场份额的占比为 94.4%，其余为 AMD，占比 5.6%。

4　我国数据中心算力情况

4.1　算力分析

（1）在通用算力方面。CPU 类型的服务器几乎部署在所有的数据中心中，根据前文的数据可知 x86 服务几乎占据了 CPU 服务器的全部市场。经过测算，截止到 2019 年年底，我国数据中心通用计算能力为 71.96 EFLOPS（FP32）。

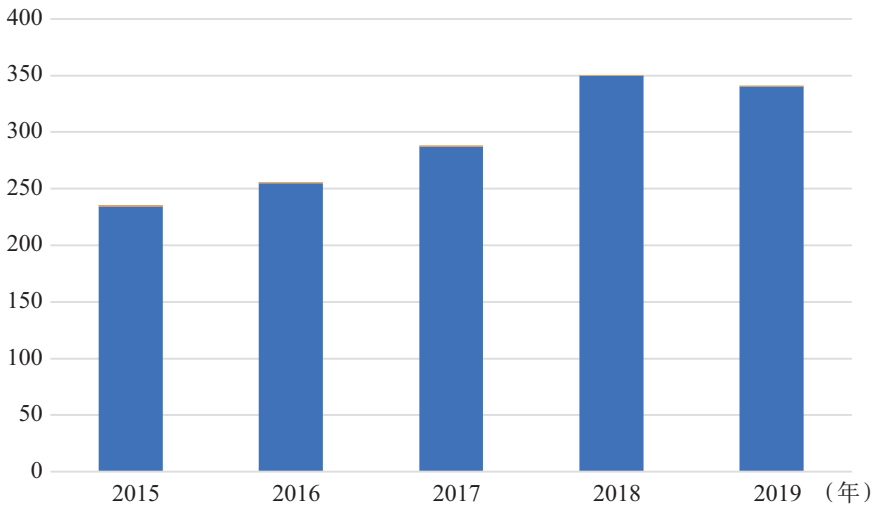

图 2　2015—2019 年我国 CPU 架构服务器出货量（单元：万台）

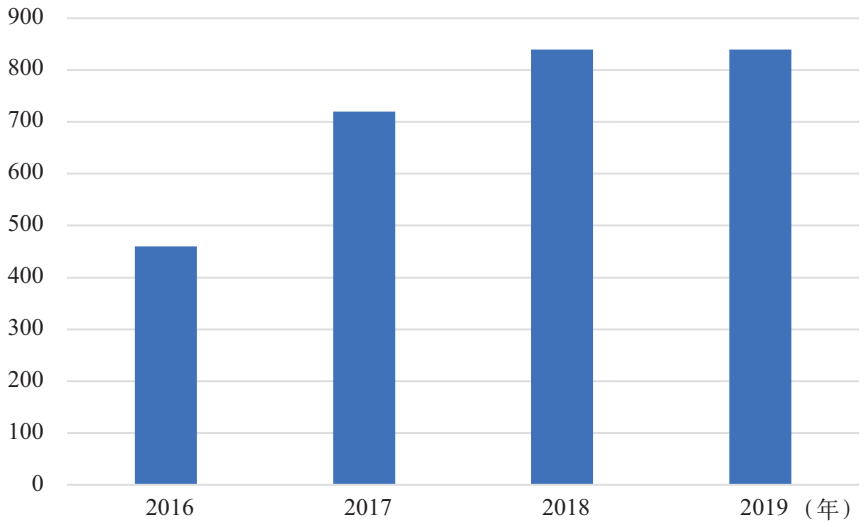

图 3　2016—2019 年全球 GPU 出货量（单位：万 Unit）

(2)在高性能算力方面。GPU 更多地使用在 AI 等应用场景中,部署于部分数据中心中且规模较小。经过测算,截止到 2019 年年底,我国数据中心高性能计算能力为 3.90 EFLOPS(FP64),折算为单精度浮点算力为 7.78 EFLOPS(FP32)。

综上,截止到 2019 年年底,我国数据中心总算力(含通用算力和高性能算力)即 CP 为 79.74 EFLOPS(FP32)。

4.2 算效分析

截止到 2019 年年底,我国数据中心的通用计算能力的算效为 15.7 GFLOPS/W (FP32);高性能计算能力的算效为 22.8 GFLOPS/W(FP64),折算为单精度浮点的算效为 45.5 GFLOPS/W(FP32)。

综合通用计算能力和高性能计算能力的算效,全国数据中心的总体算效达到 18.16 GFLOPS/W(FP32)。

5 结束语

数据中心作为新一代信息通信技术的重要载体,是算力输出的底座。未来,异构加速计算的需求日益旺盛,高性能计算能力将大有可为,并逐渐成为数据中心算力的主要力量。计算、存储、网络的深度融合,更加丰富了“算力”的内涵。同时,未来“大型+边缘”的双向发展对算力提出了多样性的要求。未来算力发展的挑战将来自于功耗,应大力推动“绿色算力”的发展,在提高算力的同时降低数据中心的能耗,使得能源在数据中心的利用效益最大化。

本文为数据中心算力提供了一种行之有效的衡量方法,后续开放数据中心委员会(ODCC)将依托中国信息通信研究院云计算与大数据研究所开展数据中心算力评估的相关业务。通过评估,各数据中心可以明确自身的“算力”和“算效”,这将有助于精细化明确数据中心的计算能力以及能耗真正的利用情况,使得数据中心在不断调优 PUE 的同时,通过不断调优“算效”,从而进一步将数据中心作为新型基础设施的杠杆作用发挥到极致。同时,下一步将不断扩充算力研究的内涵以及多样性,在把已有研究细化深化之后,继续将算力的成本、经济效益、社会影响等方面也纳入相应的考虑,使得研究的体系更加完善。

参考文献

[1] 国家发展和改革委员会. 国家发改委举行 4 月份例行新闻发布会[EB/OL]. (2020-04-20)[2020-12-29]. http://www.scio.gov.cn/xwfbh/gbwxwfbh/xwfbh/fzggw/Document/1677563/1677563.htm.

[2] 郭亮. 边缘数据中心关键技术和发展趋势[J]. 信息通信技术与政策, 2019(12):55-58.

[3] Top500. TOP500 November 2020[EB/OL]. (2020-11-25)[2020-12-29]. https://www.top500.org/lists/top500/2020/11/.

[4] Top500. GREEN500 November 2020[EB/OL]. (2020-11-25)[2020-12-29]. https://www.top500.org/lists/green500/2020/11.

[5] Standard Performance Evaluation Corporation[EB/OL]. (2020-12-15)[2020-12-29]. http://www.spec.org.

[6] Standard Performance Evaluation Corporation. SPEC CPU® 2017[EB/OL]. (2020-11-25)[2020-12-29]. http://www.spec.org/cpu2017/.

[7] Standard Performance Evaluation Corporation. SPEC Power®[EB/OL]. (2020-12-16)[2020-12-29]. http://www.spec.org/power_ssj2008/.

[8] Standard Performance Evaluation Corporation. SPEC SERT® Suite[EB/OL]. (2020-11)[2020-12-29]. http://www.spec.org/sert/.

[9] MLPerf. MLPerf Inference[EB/OL]. (2020-08-25)[2020-12-29]. https://www.mlperf.org/inference-overview/.

[10] MLCommons. October 21, 2020—Inference: Datacenter v0.7 Results[EB/OL]. (2020-12)[2020-12-29]. https://mlcommons.org/en/inference-datacenter-07/.

[11] MLCommons. 07.29.2020—Mountain View, CAMLPerf Training v0.7 results[EB/OL]. (2020-12)[2020-12-29]. https://mlcommons.org/en/news/mlperf-training-v07/.

[12] 开放数据中心委员会[EB/OL]. (2020-12)[2020-12-29]. http://www.odcc.org.cn/index.html.

[13] 开放数据中心委员会. 服务器能效评测规范[EB/OL]. (2020-12)[2020-12-29]. http://www.odcc.org.cn/download/p-1169551993955831810.html.

[14] The Green Grid. WP#49 PUE™: a comprehensive examination of the metric[R], 2011.

[15] John Nickolls, William J Dally. The gpu computing era[J]. IEEE micro, 2010,30(2):56-69.

[16] McMahon, F H. The livermore fortran kernels: a computer test of the numerical performance range[J]. United States: N, 1986, 179.
[17] Yifan Sun. Summarizing CPU and GPU design trends with product data[Z], 2019.
[18] Intel. Intel product specifications[EB/OL]. (2020-12)[2020-12-29]. https://ark.intel.com/.
[19] Nvidia. 适用于服务器的 TESLA 数据中心 GPU[EB/OL]. (2020-12)[2020-12-29]. https://www.nvidia.cn/data-center/tesla/.
[20] 郭亮. 面向云计算的企业级硬盘基准测试[J]. 电信网技术, 2016(10):1-4.
[21] OpenAI. AI and Compute[EB/OL]. (2019-11-07)[2020-12-29]. https://openai.com/blog/ai-and-compute/.
[22] 中国信息通信研究院, 开放数据中心委员会. 数据中心白皮书(2020)[R], 2020.
[23] 工业和信息化部信息通信发展司. 全国数据中心应用发展指引(2019)[M]. 人民邮电出版社, 2020.
[24] Gartner. Servers marketshare WW country 2019Q4[R], 2019.
[25] IDC. Datacenter processing, 4Q19: server CPU, GPU, FPGA, AI ASICs, and ASSPs[R], 2019.

作者简介:

郭亮 中国信息通信研究院云计算与大数据研究所副总工程师,主要从事数据中心产业咨询、标准制定等工作,主要研究领域为数据中心网络、服务器等创新技术

吴美希 中国信息通信研究院云计算与大数据研究所高级业务主管,主要从事数据中心政府支撑、产业咨询、技术研究和标准制定等工作,主要研究方向为边缘数据中心、绿色数据中心、数据中心算力等

王峰 中国电信股份有限公司研究院教授级高级工程师,长期从事云计算、大数据、人工智能等新兴信息技术领域的技术研发和产品创新工作

龚敏 英特尔(中国)有限公司高级技术经理,主要从事数据中心事业群平台应用、推广及生态建设工作,主要研究领域为通信及边缘计算

Research on evaluation of computing power and efficiency in data center: status and opportunities

GUO Liang[1], WU Meixi[1], WANG Feng[2], GONG Min[3]

(1. Institute of Cloud Computing and Big Data, China Academy of Information and Communications Technology, Beijing 100191, China;
2. Beijing Research Institute, China Telecom, Beijing 102200, China;
3. Data Platform Business Group, Intel, Shenzhen 518000, China)

Abstract: This paper reviews the research status of computing power and efficiency, including SPEC CPU, SPEC power, MLperf, etc., which describe the computing power of a single server in the data center, while TOP 500 and green 500 describe the supercomputing power, while PUE describes the power utilization efficiency. And there is no scientific methods to describe the computing power and efficiency of the data center. Based on the analysis, this paper puts forward a method to measure the computing power and efficiency of data centers, and calculates the current level of computing power and efficiency of data centers in China.

Keywords: data center; computational power; computational efficiency

(收稿日期:2020-12-20)