学术观点

新型存算分离架构技术展望

舒继武[1, 2]
[1] 清华大学
[2] 厦门大学

关键词：存算分离　分离式存储系统　存储模组　资源池化　云和互联网场景

数字化、信息化的不断发展是推动计算、存储、网络等 IT 基础设施演化和进步的重要动力。随着全球范围内企业数字化转型的快速发展，各行各业产生的数据呈现海量增长趋势。云和互联网行业构建了我国最大的 IT 基础设施平台，其存储和处理的数据量占比最大。同时，我国东数西算工程的持续推进，对数据中心走向绿色集约、基础设施自主可控提出了更高要求。

实际应用中，大数据存储解决方案一般有存算融合和存算分离两种部署形态。存算融合以基于服务器的超融合系统为代表，它将服务器的计算、存储、网络等资源进行统一管理和调度，具有弹性的横向扩展能力；但是当实际业务对计算和存储需求不同时，该方案存在资源扩展不灵活、利用率低下等问题。存算分离将存储资源和计算资源拆分为独立的模块进行建设，在资源利用率、存储资源高效共享、多场景灵活部署、网存算协同等方面具有显著优势。存算分离架构当前已经在许多场景得到应用：金融及电信等行业核心交易系统、数据库等关键应用通常采用小型机结合高端存储的方式；在企业办公场景下，为满足跨平台企业应用的数据共享诉求，通常采用通用服务器结合文件共享存储（Network Attached Storage，NAS）的方式；在大数据场景下，为实现多业务数据共享，通常采用分析计算服务结合数据湖的形式。多样化的存算分离实践为存储系统带来了数据共享、灵活伸缩等优势。

存算分离技术发展分析

新的业务挑战

从云和互联网的业务场景来看，其存储域主要采用基于服务器部署分布式存储服务的融合方式，它面临如下挑战：

1. 数据保存周期与服务器更新周期不匹配。大数据、人工智能等新兴业务催生出海量数据，大量数据需按照其生命周期策略（例如 8~10 年）进行保存。而在当前存储域中，基于服务器的存储系统换代周期由处理器的升级周期（例如 3~5 年）[1] 决定。这种数据生命周期与服务器更新周期之间巨大的差异导致系统资源被大量浪费，比如，存储域中服务器组件都随 CPU 升级而淘汰，为此须进行相应的数据迁移等 [2]。

2. 性能可靠与资源利用率难以兼得。支撑业务的分布式存储系统大致可以分为性能型存储和容量型存储，它们均无法同时实现高性能可靠与高资源利用率。具体地，性能型存储主要运行数据库、虚拟化等关键业务，通常采用三副本或两副本并配合独立冗余磁盘阵列（Redundant Array of Independent Disks，RAID）卡模式；这类方案虽兼顾了性能和可靠性，但其大约 30% 的空间利用率却是对存储资源的极大浪费。容量型系统为了提升空间利用率，采用纠删码（Erasure Code，EC）方式，然而，EC

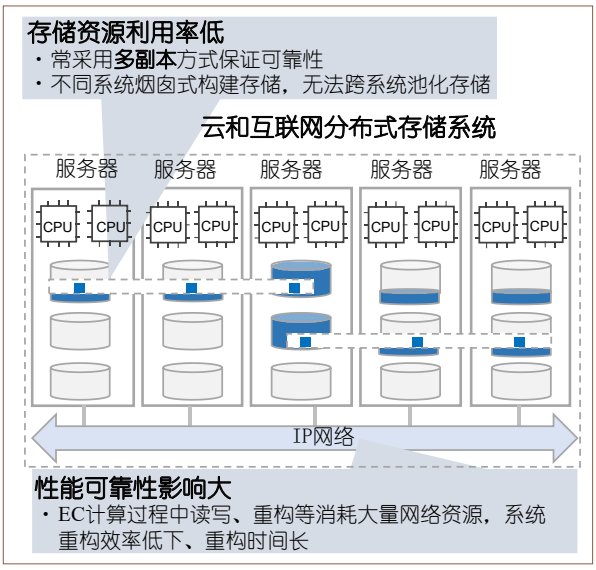

图1　分布式存储资源利用率

计算过程中的读写、重构等会消耗大量网络资源，导致系统重构效率低下、重构时间长，给系统可靠性带来风险（如图 1 所示）。

3. 新型分布式应用的极简高效共享存储诉求。以无服务器（serverless）应用为代表的新型分布式应用在近些年涌现，这类应用从无状态化向有状态化扩展，比如数据库、消息总线等组件纷纷容器化，数据共享访问的诉求不断增多。与此同时，人工智能和机器学习等应用需要大量异构算力协同，甚至产生共享内存访问的诉求，它们关注高带宽、低时延的访问能力，仅需要轻量、便捷的共享存储系统即可，不需要搭载具有复杂企业特性的传统存储。

4. 数据中心税导致数据密集型应用效率低下。面向数据密集型场景，在基于以 CPU 为中心的服务器架构下，应用为获取数据所缴纳的“数据中心税”（datacenter tax）日益加重。例如，服务器内的 CPU 为处理网络及存储 IO 请求，需要消耗高达 30% 的算力[3]；此外，由于通用 CPU 并不擅长数据处理运算，导致其能效比低下。

传统存算分离架构将算力资源和存储资源（机械硬盘、固态硬盘等）分离至彼此独立的计算域和存储域，并通过以太网或专用存储网络（例如光纤通道）将二者互连，实现了存储资源的灵活扩展和高效共享（如图 2 左侧所示）；该架构主要为复杂的传统企业特性设计，难以应对上述挑战，为了让云和互联网存储域服务兼顾资源利用率、可靠性、性能、效率等众多诉求，亟须基于新型软硬件技术构建新型存算分离架构。

硬件技术趋势

面对数据中心在容量利用率、存力效率等方面的挑战，近年来，专用数据处理器、新型网络等技术快速发展，为数据中心基础设施的重构提供了技术基础。

首先，为取代服务器本地盘，很多厂商推出以太网闪存簇（Ethernet Bunch of Flash，EBOF）高性能盘框（例如，近期陆续发布的西数 OpenFlex、Vast Data Ceres 高性能盘框等）。这类盘框不再具有复杂企业特性，而是注重采用新型的数据访问标准，比如支持 NoF（NVM Express over Fabric）等接口，以提供高性能存储实现对本地盘的替换。NoF 协议由 NVM Express（NVMe）标准组织在 2016 年发布，提供了 NVMe 命令到多种网络传输协议的映射，使一台计算机能够访问另一台计算机的块存储设备。同时，一些研究机构进一步探索远程内存池化技术，例如，韩国 KAIST 实验室实现了基于 FPGA 的 CXL（Compute Express Link）互连协议[4]；CXL 为英特尔于 2019 年 3 月在 Interconnect Day 2019 上推出的一种开放性互联协议，能够让 CPU 与 GPU、FPGA 或其他加速器之间实现高速高效互联，从而满足高性能异构计算的要求。

其次，业界涌现出越来越多的数据处理单元（Data Processing Unit，DPU）和基础设施处理单元（Infrastructure Processing Unit，IPU）专用芯片，在数据流处理路径上取代通用处理器，提升算力能效比。同时，基于可编程交换机的网存协同也是研究热点，例如在网数据缓存的 NetCache[5]、KV-Direct[6]，在网数据协调的 NetLock[7]、Concordia[8]、SwitchTx[9]，在网数据聚合的 SwitchML[10]、NetEC[11]，在网数据调度的 FLAIR[12]、AlNiCo[13] 等

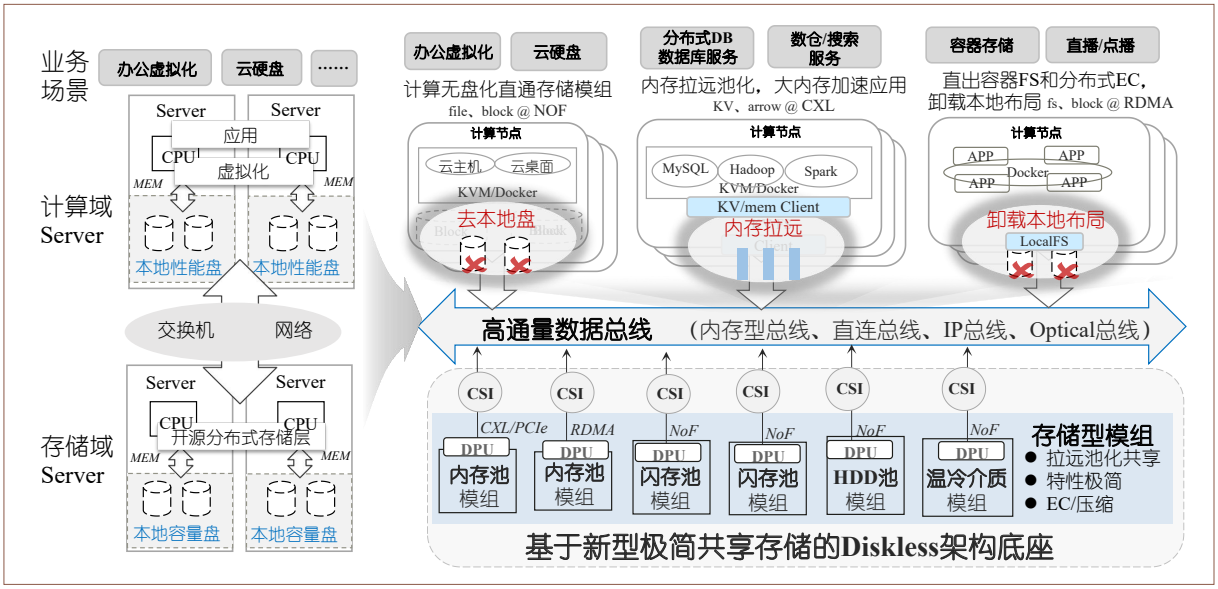

图2 传统存算分离架构与新型存算分离架构对比

相关论文纷纷在主流会议期刊发表。

最后，数据访问网络标准也在持续增强，比如CXL协议新版本加强了内存池化方向的技术特性，同时吸收了Gen-Z[14]（由AMD、ARM、HPE等公司发起定义的面向内存语义的技术）、OpenCAPI[15]（Open Coherent Accelerator Processor Interface，最早由IBM提出的异构计算接口）等技术的成果，正逐步成为业界主流高速互联标准。NVMe 2.0也在向着语义统一、Fabric统一和介质统一方向演进。

这些新型存储、计算和网络硬件为构建面向云和互联网场景的新型存算分离架构带来了诸多机遇，譬如使用DPU等专用芯片能够打破传统以CPU为中心的服务器架构，由此提升数据密集型应用的效率。

新型存算分离架构的特征

随着远程直接内存访问（Remote Direct Memory Access，RDMA）、CXL、可编程网络设备、高性能NVMe SSD、持久性内存等新型硬件技术的发展，需要构建新型存算分离架构，以确保云和互联网存储域服务能够兼顾资源利用率、可靠性、性能、效率等众多诉求。相较于传统架构，新型存算分离架构最为显著的区别在于：(1) 更为彻底的存算解耦，该架构不再局限于将CPU和外存解耦，而是彻底打破各类存算硬件资源的边界，将其组建为彼此独立的硬件资源池（例如处理器池、内存池、机械硬盘（HDD）/固态硬盘（SSD）池等），从真正意义上实现各类硬件的独立扩展及灵活共享；(2) 更为细粒度的处理分工，即打破了传统以通用CPU为中心的处理逻辑，使数据处理、聚合等原本CPU不擅长的任务被专用加速器、DPU等替代，从全局角度实现硬件资源的最优组合，进而提供极致的能效比（如图2右侧所示）。

总结来说，新型存算分离架构具有如下特征：

1. 无盘化的服务器。新型存算分离架构将服务器本地盘拉远构成无盘化（diskless）服务器和远端存储池，同时还通过远程内存池扩展本地内存，实现了真正意义上的存算解耦，可极大提升存储资源利用率。业务使用时，可根据应用需求选择配置不同性能、容量的虚拟盘及池化内存空间，这样一方面可以避免由于不同服务器本地存储空间利用率过低导致超配造成的浪费；另一方面，当服务器出现故障或者更新换代时，也不影响数据的保存，不需要额外的数据迁移。

2. 多样化的网络协议。连接计算和存储间的网络协议从当前的IP或光纤通道（Fibre Channel，

FC）协议扩展到 CXL+NoF+IP 协议组合。CXL 协议使得网络时延降低到亚微秒级别，有助于内存型介质的池化 ；NoF 协议加速 SSD 池化 ；IP 协议可满足 HDD 等慢速介质访问诉求。通过这几类协议组合构建的高通量网络，满足了多种场景池化接入诉求。

3. 专用化的数据处理器。数据存储、访问等操作不再由通用处理器负责，而是卸载到专用数据处理器。此外，特定的数据操作可由专用硬件加速器进行进一步加速，如纠删码、加密压缩、网络通信等。通过专用数据处理器，可以释放通用处理器算力，用于服务更适合的场景，显著提升系统整体能效比。

4. 极高存力密度的存储系统。分离式存储系统（disaggregate storage）是新型架构的重要组件，作为持久化数据的底座，在存储介质的集约化管理基础上，结合芯片、介质的深度协同设计，整合当前系统、盘两级的空间管理，通过大比例纠删码算法减少冗余资源开销比例。此外，还可通过基于芯片加速的场景化数据缩减技术提供更多的数据可用空间。

面向云和互联网场景的存算分离架构及关键技术

面向云和互联网场景的存算分离架构

新型存算分离架构意在解决前文所提的当前架构面临的几大痛点挑战，通过将原有架构的多级分层资源进行彻底解耦池化和重组整合，形成新的三大简化分层 ：存储模组、总线网络和算力模组，从而提供服务器本地存储拉远池化、新型网络灵活组装、以数据为中心的多元处理、高容量极简盘框等几大新兴能力。

存储模组

面向云和互联网数据中心，需要以更专业的存储能力重新定义云和互联网的存储架构。新型存算分离架构中，存储型模组主要以 EBOF、以太网内存簇（Ethernet Bunch of Memory，EBOM）、以太网磁盘簇（Ethernet Bunch of Disk，EBOD）等新型盘框形态存在，RAID/EC/ 压缩等传统存储能力下沉到新型盘框中，构成“ 盘即存储 ”的大盘技术，对外通过 NoF 等高速共享网络提供块、文件等标准存储服务。这一类新型盘框将传统磁盘阵列的冗余池化技术和数据缩减技术进行了高度集约化和小型化。

从存储模组内部架构来看，其介质层可由标准硬盘组成，也可由晶圆工艺直接整合的颗粒大板组成，盘与框的边界融合有助于实现极致成本的创新。在介质层之上，存储模组需构建类似传统存储阵列的池化子系统，基于 RAID、EC 等可靠冗余技术实现本地介质的池化，结合重删压缩等算法技术对数据容量进行大比例缩减，进一步实现可得容量的提升。为了支撑新型存算分离架构的高通量数据调度，存储模组需要提供更加高效的数据吞吐能力，通常基于硬件直通等技术构建极简的快速数据访问路径。和传统阵列相比，存储模组在 IO 处理上尽量避免用户数据和控制数据（元数据等）的低效交织，尽量减少传统存储阵列的某些复杂特性处理（复制、双活等容灾特性），尽量减少子系统分层进而缩短 IO 处理的路径，最终实现高吞吐、低时延的极致性能体验。最后，通过硬盘亚健康管理，例如慢 IO 快速返回、慢盘隔离等能力，实现毫秒级稳定时延。

云和互联网的多样业务主要分为三种典型的应用场景（如图 3 所示）。第一种场景是针对虚拟化业务，直接将数据中心存储域服务器的本地盘拉远，对分布式开源存储集群的物理硬盘层形成替代。第二种场景是为数据库、大数据服务等需要极热数据处理的业务提供大内存、键 – 值（Key-Value，KV）接口，加速数据处理效率 ；第三种场景是针对容器等新业务场景，为 Ceph、Lustre 等分布式应用直接提供文件语义，卸载本地数据布局，并支持将温热数据分级到更冷的 EBOD 等机械硬盘或磁带型存储模组中，提升整系统资源使用效率。

存储模组作为一种存力集约化、紧凑化、极致化的新型存储形态，加速了服务器的无盘化发展趋势，将服务器的本地盘、内存等拉远进行池化共享，

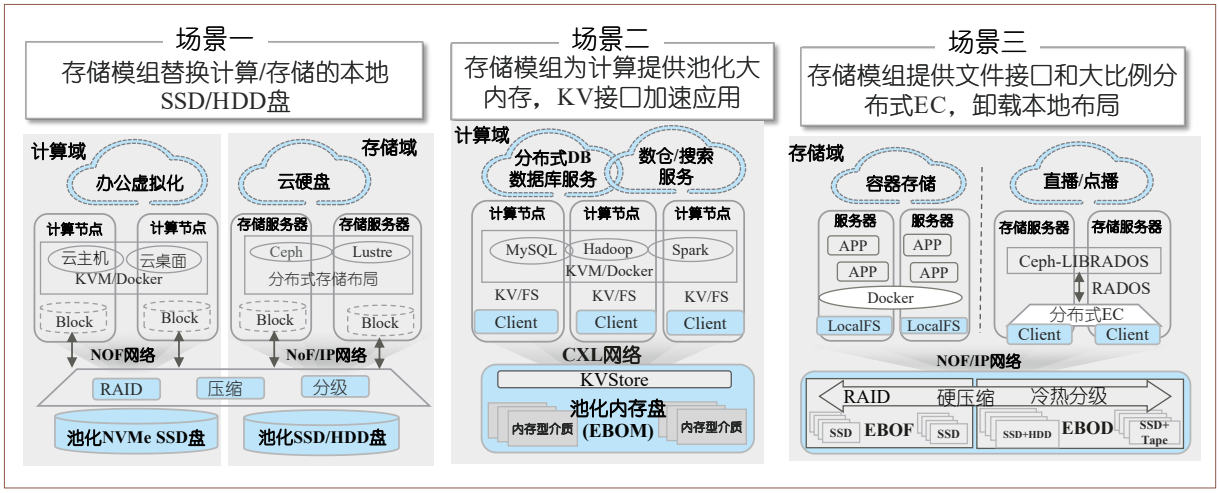

图3　存储模组的三类典型应用场景

有效支撑了传统数据中心架构朝极简分层的新型存算分离架构演进。

算力模组

当前，摩尔定律演进变缓，只有采用专用处理器才能进一步以异构方式发挥出下一阶段的算力。引入专用处理器后，算力池化是必然选择；否则，如果为每台服务器配置异构算力卡，不仅使整机功耗巨大，还会导致资源利用率十分低下。以DPU为代表的专业数据处理器具备成本更低、功耗更低、即插即用、即换即用等独特优势，并且在运行状态下不与业务应用发生资源争抢，保证用户业务正常运行的同时也保障了基础设施的服务质量。

高通量数据总线

存算分离架构中，网络技术非常重要，它决定了系统的响应速度以及吞吐能力，也决定了系统资源池化的能力范围。过去10年，万兆IP网络促使HDD池化，基于IP网络发展了支持块、文件、对象共享的访问协议。当前，面向热数据处理，NVMe/RoCE（RDMA over Converged Ethernet, RDMA融合以太网）促使SSD池化；并且，NVMe协议快速发展使其开始收编烟囱式协议规范。下一步，面向极热数据处理，内存型网络（例如CXL Fabric）将促使内存资源池化，为业务提供更大的共享内存空间（如图4所示）。

关键技术

新型存算分离架构改变了各类硬件资源的组合形式，其近远关系、松紧耦合的变化催生了一系列围绕该架构的关键技术，例如场景化数据缩减、高通量超融合网络、网存协同、盘芯协同等。

场景化数据缩减

在新型存算分离架构下，数据缩减能力将下沉到存储模组中，通过前后台缩减任务配合，可有效减少对性能的影响，同时提升数据缩减率。此外，针对不同场景的数据特征，可使用不同的缩减技术。例如，针对基因、医疗等场景，可通过多帧图片聚合压缩、多波段数据合并压缩等实现更高缩减率；在数据保护场景，可通过变长或相似性重删获得更高缩减率；在视频、媒资场景，可通过前景提取、

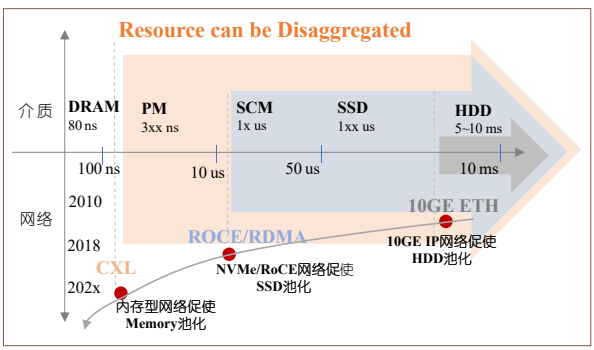

图4　网络技术发展时间线

码率控制等技术实现更高缩减率。

高通量超融合网络

根据部署场景及多样化网络敏捷和自适应性的业务需求，存算模组间的网络连接可以选用基于 CXL Fabric、NoF、IP 的组合进行组网，在网络协议上，有以下关键技术需要考虑 ：

首先，网络连接可以选用直连模式或是池化模式。直连模式下，网卡资源被设备独占使用 ；池化模式下，网卡资源池化，被多个设备共享使用，可以提供更经济的使用效率。池化模式下，需要解决网络资源细粒度动态分配能力、安全隔离能力等，从而保证多个设备对资源的公平使用。

其次，跨机架通信通常采用 RDMA 机制。传统 RDMA 连接数受限，需要解决大规模互联的扩展性问题。例如，可使用无连接等技术，解耦连接状态和网络应用，做到支持数万的连接规模 ；同时通过软硬协同，硬件确保请求及时完成，提供快速的操作失败通知，而软件负责操作重试和故障恢复，两者配合实现高性能可靠连接。此外，还可考虑使用多路径技术，通过支持数据包乱序收发能力、异步 ACK 机制，解决单条网络带宽有限等问题。

网存协同

智能网卡和 DPU 是服务器的数据出入口，可编程交换机是服务器、存储之间的数据交换中枢，它们在系统中占据特殊的位置。因此，结合其可编程能力，可以实现高效的数据协同处理。

首先，智能网卡和 DPU 可以实现任务卸载，包括 NoF 加速、压缩及解压加速、安全算法（AES、RSA、ECC、Diffie 等）卸载、正则表达式卸载等 ；其次，利用其可编程能力，通过精细化的流水线并行技术，可将存储的文件服务、块服务、内存与 KV 服务等卸载到智能网卡和 DPU 里面，缩短 IO 访问的响应时间 ；最后，面向特定场景加速（例如，虚拟化场景的虚机直通、大数据场景算子下推、shuffle 协同等），可极大地提升系统运行效率。总之，充分利用好智能网卡和 DPU 的硬件加速资源，协同好主机和 DPU 间的任务调度，有助于降低主机数据处理开销，提升 IO 访问效率。

可编程交换机具备自定义网络协议和网络包转发的能力，并且其上配置了小块片上内存用于存储数据，这些可编程能力结合交换机的中心化和高性能的优势，可以实现在网数据处理加速。例如，将消息转发和并发控制卸载至交换机，降低分布式协调开销 ；或者，将元数据、热点键值对缓存在交换机的内存中，并利用交换机中的算力执行相关插入、排序、查找、删除等操作加速元数据响应。此外，还可利用交换机广播、组播能力，实现数据副本传输，降低主机开销。

盘芯协同

通过介质和控制芯片深度协同可获得端到端最佳总体拥有成本（Total Cost Ownership，TCO）和效率。以数据冗余设计为例，原有的体系架构中存在多层数据冗余 ：即 SSD 盘内第一层的介质层 EC 冗余，在硬盘之上基于 RAID 或副本技术形成第二层冗余，每一层冗余设计相互独立且无法协同。新型存储型模组直接集成介质颗粒，仅在框这一级构建一层大比例 EC 的池化空间，辅助专有芯片对算法进行卸载加速，最终简化了原有的多层冗余设计，有效改善端到端的资源利用率。

此外，还可基于专用芯片设计实现对磨损均衡、垃圾回收、多流等特定介质高级能力的深度管理，与上层应用协同垂直优化，实现场景优化。

最后，新型存储模组基于专有芯片除了提供传统 IO 接口外，还可以提供基于控制器内存缓冲区（Controller Memory Buffer, CMB）的旁路接口加速，这有助于系统元数据路径绕开厚重的 IO 栈，使用远程内存访问方式来提升系统访问并行能力。

面向云和互联网场景的存算分离的技术挑战与机遇

技术挑战

面向云和互联网场景的新型存算分离架构受网络、算力等技术驱动，顺应未来数据中心可组合式架构（composable infrastructure）趋势。然而，构建

这类系统并充分发挥其潜在效率，也面临众多技术挑战，需要产业界、学术界专家共同探索解决。

首先，计算和存储之间的数据访问接口及标准主要采用“主－从”请求响应模式，并以传输块存储语义为主。然而，随着内存盘、计算型盘、智能网卡异构算力的快速发展，内存访问语义、计算协同存储语义等方面的表现能力出现不足。此外，当前国内对于业界主流数据访问接口的定义，如NVMe标准等，缺乏自主可控能力，实现我国自主定义存算间新型数据访问标准意义十分重大。

其次，如何与已有生态应用结合，发挥出基于新型存算分离架构的基础设施的潜力仍需深入探索。例如，新的数据处理器、全局共享存储系统的引入，计算、存储的独立弹性扩展等，都为新型应用提供了较好的基础设施能力，但如何最大限度地将这些基础设施潜力发挥出来，如何设计更高效的应用服务框架，如何与上层应用协同等，都是一个长期而艰巨的任务。

机遇展望

根据我国“十四五”规划，为助推社会经济高质量发展、加强数字政府建设、激活数据要素潜能以及为各行各业的数字化转型注入新动能，国家在多个地区构建智算中心，并基于一体化大数据中心构建东数西算工程。预计到2025年，我国将具备300 EFLOPS（Exa FLoating-point Operations Per Second）的算力，数据量将达到48.6 ZB[16]。这些规划都为未来基础设施的发展带来极大的挑战，例如，在如此高速度的算力增长下，如何避免因为存储、网络等性能限制导致算力长期处于空闲状态？在介质产能有限的情况下，如何保存这些海量数据？此外，未来数据中心的发展将严重受限于其功耗预算和碳排放配额，如何提供绿色节能的基础设施系统？这些问题既是挑战，也是机会，相信新型存算分离以其灵活的架构、精细化的资源利用率、绿色低碳的能耗比等优势，在我国宏大数字化历史进程下，将迎来最佳的历史发展机遇。 ■

舒继武

CCF会士、信息存储专委会主任。清华大学教授。厦门大学信息学院院长。主要研究方向为信息存储系统、并行分布式系统、边缘数据存储系统等。

shujw@tsinghua.edu.cn

（本文责任编委：**郭得科**）

参考文献

[1] Bozman J S, Broderick K. IDC: Server Refresh: Meeting the Changing Needs of Enterprise IT with Hardware/Software Optimization[OL]. https://www.oracle.com/us/corporate/analystreports/corporate/idc-server-refresh-359223.pdf, July 2010.

[2] Zhang T, Zuck A, E.Porter D, et al. Flash Drive Lifespan is a Problem[C]// *Proc. of the 16th Workshop on Hot Topics in Operating Systems* (HotOS 2017), 2017:42-49.

[3] Kanev S, Darago J, Hazelwood K, et al. Profiling a warehouse-scale computer[C]//*The 42nd International Symposium on Computer Architecture* (ISCA 2015), 2015:158-169.

[4] Gouk D, Gouk D, Lee S, et al. Direct Access, High-Performance Memory Disaggregation with DIRECTCXL[C]//*USENIX Annual Technical Conference* (USENIX ATC 2022), 2022:287-294.

[5] Jin X, Li X, Zhang H, et al. NetCache: Balancing Key-Value Stores with Fast In-Network Caching[C]// *The 26th ACM Symposium on Operating Systems Principles* (SOSP 2017), 2017:121-136.

[6] Li B, Ruan Z, Xiao W, et al. KV-Direct: High-Performance In-Memory Key-Value Store with Programmable NIC[C]// *The 26th ACM Symposium on Operating Systems Principles* (SOSP 2017), 2017:137-152.

[7] Yu Z, Zhang Y, Braverman V, et al. NetLock: Fast, Centralized Lock Management Using Programmable Switches[C]// *The Conference of the ACM Special Interest Group on Data Communication* (SIGCOMM 2020), 2020:126-138.

[8] Wang Q, Lu Y, Xu E, et al. Concordia: Distributed Shared Memory with In-Network Cache Coherence, *The 19th USENIX Conference on File and Storage Technologies* (FAST 2021), 2021:277-292.

[9] Junru Li, Youyou Lu, Yiming Zhang, et al. SwitchTx: Scalable In-Network Coordination for Distributed Transaction Processing[C]. *48th International Conference on Very Large Data Bases* (VLDB 2022),

2022:2881-2894.
[10]Amedeo Sapio, Marco Canini, and Chen-Yu Ho, et al. Scaling Distributed Machine Learning with In-Network Aggregation[C], The 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2021), 2021:785-808.
[11]Yi Qiao, Xiao Kong, Menghao Zhang, et al, Towards In-network Acceleration of Erasure Coding, The ACM SIGCOMM Symposium on SDN Research (SOSR 2020), 2020:41-47.
[12]Hatem Takruri, Ibrahim Kettaneh, Ahmed Alquraan, et al, FLAIR: Accelerating Reads with Consistency-Aware Network Routing, The 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2020), 2020:723-737.
[13]Junru Li, Youyou Lu, Qing Wang, et al. AlNiCo: SmartNIC-accelerated Contention-aware Request Scheduling for Transaction Processing. USENIX Annual Technical Conference (USENIX ATC 2022), 2022:951-966.
[14]Gen-Z, Consortiumhttps://genzconsortiumorg, 2022
[15]OpenCAPI, https://opencapi.org, 2022
[16]David Reinsel, 武连峰, John F.Gantz, John Rydning, IDC: 2025 年中国将拥有全球最大的数据圈[OL], http://www.d1net.com/uploadfi le/2019/0214/20190214023650515.pdf, 2019年1月