# 新应用和硬件驱动下的存储技术创新
## ——第十三期CCF秀湖会议报告

整理：周　可　何水兵　王　桦 等

2024年5月10~12日，第十三期CCF秀湖会议在苏州CCF业务总部&学术交流中心举办，以“新应用和硬件驱动下的存储技术创新”为主题进行深入交流和研讨。来自学术界与工业界的30多位专家全程参与，围绕存储介质与芯片、存储器与设备、存储系统、存储服务与应用等方面进行探讨，并形成如下报告。

## 背景与意义

随着信息技术的不断发展，全球数据总量持续高速增长，数据成为国家性战略资源，数据中心成为国家未来重点发展的新基建项目。目前，全球存储市场规模已超千亿美元，预计在未来五年呈现稳定增长态势。面对如此强劲的存储需求，存储的重要性也日益凸显。

近年来，随着数据量的指数级增长，新应用和新硬件为存储技术创新带来了机遇与挑战。一方面，智能计算、图计算、大数据处理等新应用的涌现，对数据存储的高效性、安全性、可靠性、可扩展性等方面提出了更高需求和全新挑战。另一方面，新型存储硬件（HBM、PM、NVMe SSD、ZNS等）和新型网络硬件（CXL、RDMA等）的采用，为存储系统的设计提供了新路径。如何结合新应用需求和新硬件特性推动存储技术创新仍然是一个开放性问题。

本次秀湖会议分析了新应用与新硬件驱动下存储面临的机遇与挑战，探讨了存储的物理基础、技术应用、科技创新与发展瓶颈，以期共同推进中国存储技术的创新与发展。

## 观点集萃

### 特邀嘉宾观点

存储系统发展面临机遇与挑战。当前全球数据总量持续高速增长，实时性数据占比愈发显著。在应用需求趋于多样化的背景下，存储硬件呈现出性能急剧飙升、接口多样化和硬件能力发生变化等三大演进趋势，存储软件面临着软件效率制约整体性能、硬件接口变化颠覆I/O软件栈设计、有限硬件能力/特性难以满足复杂应用需求等挑战。在架构方面，传统存储系统架构面临服务器架构与以CPU为中心的（CPU-centric）架构两方面的问题，存储系统架构在向分离式数据中心架构与网存算融合架构两个不同方向发展，前者解决了服务器架构硬件资源利用率低、弹性差的问题，后者解决了CPU-centric架构的数据移动多、通信开销大的问题。

进一步来说，可根据数据面与控制面划分将二者做到统一共存，具体做法是，数据面采用分离式数据中心架构，控制面采用网存算融合架构。新一代AI在规模成本、访问效率、可靠容错方面对存储提出更高要求，并带来数据集存储、采样预处理、模型训练、模型推理等方面的技术挑战。

未来数据中心对存储平台技术提出了新的要求。全球芯片与IT市场给存储带来了新机遇，高安全、高可靠、节能化、智能化、服务化成为数据中心的演进方向。未来数据中心xDC以数据为中心，跨空间、透明流动、高效共享，跨时间、安全可靠持久存储，融合了端、边、云，数据安全高效地收集、汇聚、访问、流动。未来数据中心的“六性”，即业务实时性、计算多样性、数据流动性与高效性、系统健壮性与管理便捷性，对存储提出了“六度”的新需求，即存储性能容量的深度、多样化负载的宽度、数据畅通的广度、节能增效的密度、安全可靠的强度与资源调度的柔度。未来数据中心存储将发展“两体三层三面”的第五种架构[1]，基于新协议互连、全品类介质、存算分离、弹性共享的超大规模分布式融合智能存储平台是未来数据中心存储的演进方向。

存储的重要性与作用需要得到重视与强调。（1）存储具有相对独立性，存储也需要保持相对独立性。由于我国了解和从事存储行业的人并不多，人们对存储的相对独立性存在疑问，常将存储作为计算的附属品。然而，不论从数据独立性，还是从技术发展和产业市场的角度，存储都表现出其日益增强的相对独立性。从数据独立性的角度看，存储作为数据的载体，本质上不依附于计算而存在；从技术发展的角度看，从以计算为中心到以数据为中心的技术演进，使存算分离架构成为当前数据中心的主流；从产业市场的角度看，传统的将存储捆绑在计算内的市场销售模式，越来越不适应数据量指数级增长导致的存储容量不断扩容的现实需求。因此，存储的相对独立性应该受到重视，而不必太过拘泥于以往的传统认识。（2）存储在AI大模型领域起着重要作用。AI大模型参数量庞大，动辄以千亿、万亿计，训练非常耗时。在AI大模型的生命周期内，许多过程与存储密切相关。例如，训练大模型时，数据预处理占用大量时间，而GPU则有相当一部分时间处于闲置空转状态；由于大模型训练故障频繁，需要高频次存储检查点（checkpoints）以防止模型信息丢失；大模型处理的数据多为小文件、多模态文件，对存储要求较高。如果借助合适的存储技术，则有望提升GPU利用率，减少GPU资源浪费，提升模型训练的稳定性与安全性。目前的研究对大模型参数量、显卡的算力关注较多，而对有助于解决AI大模型诸多问题的存储关注还不够。存储界需要阐明存储对AI大模型的作用，并讲清楚支撑AI大模型的存储系统应该呈现什么样的形态。（3）推动存储指标进入系统性能评估体系。当前系统评估缺少存储指标，有些系统的算存比高达数百，存储的作用完全被忽视，而在系统投入实际环境运行时，又常常遇到存储资源迅速消耗殆尽的情况，因此，建议系统评估时考虑算存平衡，也可以考虑推出存储TOP 500。

## 存储介质与芯片专题观点

新材料、新结构、新机理实现颠覆性的闪存器件。存储芯片是集成电路产业第二细分市场，其规模仅次于逻辑芯片。随着人工智能、物联网、汽车电子等领域的快速发展，存储芯片的市场需求仍在持续增长。在存储芯片市场中，闪存作为主流的非易失存储器，占据40%以上的市场份额，具有成本低、容量大、功耗低的优势，已服务于超过99%的非易失存储器应用。然而，目前闪存面临着性能和寿命方面的问题。首先，闪存可实现超过10年的非易失存储，但性能较慢，仅能达到10~100 μs的编程速度；其次，闪存的耐久性只有$10^4$~$10^5$次，

[1] 第五存储架构的两体指存储平台本体和管理编排体，三层涵盖硬件资源层、微服务化功能层和开放使能层，三面包括安全性、可靠性和可扩展性。

远低于DRAM等存储器。利用新材料、开发新结构、探索新机理是实现颠覆性闪存器件可能的技术路线。探索高集成度、高速非易失存储器对占领AI芯片高地极具战略意义。

忆阻器等新存储介质推动存算一体新范式发展。目前计算机发展面临三方面问题。在架构方面，存算分离带来的存储墙瓶颈制约了计算性能的提升；在能耗方面，单位面积功耗增大，散热问题日趋严峻，功耗墙问题凸显；在智能方面，数据驱动的神经网络算法依赖海量样本与密集计算，性能提升面临诸多挑战。以忆阻器为代表的新兴存储介质具备良好的多态性（>5 bits）、耐久性（$>10^9$）与非易失性（>10年），不但可以作为下一代非易失存储介质的候选器件，也为在存储阵列中实现矩阵运算提供可能。基于新器件的存算一体计算架构可以有效打破存储墙等问题，并通过挖掘器件丰富的物理特性，推动类脑计算等新型计算范式的高效硬件实现，为人工智能系统变革性发展提供新技术。

发展绿色节能的下一代光信息存储成为国际趋势。据《数据存储2030》白皮书预测，到2030年，全球每年产生的数据总量将超过1 YB（$2^{80}$字节）。面对如此大规模的数据，基于SSD和磁盘等设备的传统磁电存储面临着高能耗、高碳排放、数据无法长程保存的问题。而光信息存储技术具有绿色、低能耗、长寿命的优势，相比传统磁电存储节能85%。华录、微软、松下、索尼等国内外公司均大力推进下一代光信息存储的研究，目标是研发绿色节能的长寿命光存储技术，为海量数据的长期存储提供技术支撑。

DNA用于数据长程存储具有良好前景。DNA作为信息存储介质有着存储密度高、抗电磁稳定、适应极端环境等优势。每克DNA分子理论上可存储455 EB（1 EB=$2^{60}$字节）数据，相比现有存储介质提高6~7个数量级。DNA半衰期为521年，能够抵抗物理冲击和电磁波干扰。DNA在–80~95 °C环境下具有较强的稳定性，能够保持分子序列信息。这些优点使得DNA可用于海量冷数据存储，提供EB/g的数据存储密度以及超过100年的断电后数据保存时间；可用于有着海量存储需求、信息安全需求、极端环境需求的数据密集型军事应用。

## 存储器与设备专题观点

持久性内存有望成为未来内存系统的新趋势。（1）持久内存打破传统存储架构。持久性内存设备同DRAM一样位于CPU内存总线上，CPU可以通过Load/Store指令直接对其进行字节寻址访问，带来低延迟、高带宽的数据读写。持久内存还具备外存设备（如HDD或SSD）的非易失性，可用作数据持久化存储，打破了传统内外存存储架构，弥补了内外存设备间的巨大性能差距。（2）持久内存在加速数据密集型应用方面展现出了巨大的潜力。例如，在图计算、数据库和KV（Key-Value）存储等领域，持久内存的高吞吐量和低延迟特性，使得这些应用能够更快地处理和访问大量持久化数据，从而提升计算效率和响应速度，减少传统存储设备的I/O瓶颈。（3）下一代持久内存研发是重要机遇。随着英特尔（Intel）结束傲腾持久内存设备的开发和支持，未来的持久内存产品发展趋势尚不明朗。基于CXL（Compute eXpress Link）互联协议的持久内存设备可能会成为傲腾持久内存的替代品。国内头部科技公司已经开始研发相关持久内存介质和设备，致力于打造自主内存产品，避免技术受制于人。

高密度闪存赋能未来存储。（1）高密度闪存需要适配的控制管理技术支持。高密度闪存通过3D堆叠和单元比特数量增加的方式在过去10年得到了长足的发展，单元密度达到了最新的28.5 Gb/mm$^2$。然而伴随高密度存储设备的是更低的可磨损次数、更差的访问性能和复杂的可靠性特征。这些问题需通过适配的控制管理技术实现介质的可用、性能的保障和数据的可靠存储。（2）深度软硬件协同的高密度闪存。当前闪存存储设备进入到四层式存储单元（Quad-Level Cell, QLC）甚至五层式存储单元（Penta-Level Cell, PLC）阶段，为新一代计算系统带来了超大存储容量和高性能的存储访问。目前高密度闪存设计沿着与系统深度结合的方向高速发展，包括基于多流的固态存储系统实现存储设备垃

圾回收的优化，分区命名空间固态硬盘（ZNS SSD）进一步实现了极致放大优化和性能优化，灵活数据放置（Flexible Data Placement, FDP）则支持良好的生态发展。（3）大模型时代的高密度闪存。大模型是目前各类场景的关键应用，通过大模型训练和推理，能够为行业提供高质量的计算服务。目前大模型的发展还处在初级阶段，当前对计算的依赖依然较高。但是随着大模型的不断深化，未来提供大模型的存储系统将变得至关重要。构建针对大模型需求的高密度闪存存储系统也将得到前所未有的重视。

基于数据处理单元（DPU）实现网存算协同设计。随着网络带宽和连接数的增加，数据传输的通路变得更宽和更密，而通用CPU的性能增长率与数据量增长率出现了显著的“剪刀差”现象，导致CPU负担过重，造成任务堆积，无法直接应对网络带宽和数据量的增速，DPU芯片就是在这样的趋势下提出和发展起来的。作为CPU/GPU之外新兴的第三处理器技术，各大云厂商及创业公司纷纷布局DPU相关技术。DPU作为计算卸载的引擎，直接效果就是给CPU减负，它通过提供专门的硬件加速和优化的数据处理能力，有效地处理数据中心内的数据移动和处理任务，从早期的网络协议处理卸载，到后续的网络、存储、虚拟化卸载都有应用，这在实现云规模计算、提高网络性能以及满足现代应用程序的需求方面发挥着至关重要的作用。DPU的多功能加速能力使其适用于处理复杂的网络工作负载，如虚拟化、加密、流量整形等，并可以实现更高效的数据传输和处理，同时减轻CPU和GPU的负担，提高整体系统的吞吐量和响应速度。

基于高速总线的内存池是破墙之道。（1）高速互连总线原生支持存算分离架构。传统数据中心的服务器节点往往采用内存和计算紧耦合的胖服务器架构，各类计算节点都配置了大量的本地内存资源，节点间通过高速网络如RDMA进行通信，这种架构不利于计算和内存等不同维度硬件资源的动态扩展和高效共享，容易造成硬件资源浪费或应用性能不佳。未来的分离式数据中心将采用存算分离的资源池化架构，有利于计算资源和内存资源的动态扩展和全局共享，实现资源按需供给，提高资源利用率，并提供细粒度的容错能力。高带宽、低延迟的互连总线CXL技术为存算分离的数据中心架构扫清了互连技术的障碍，支持CXL接口的大容量内存扩展设备也为内存资源池化提供了关键的硬件支撑，使得计算－内存解耦的服务器架构逐渐具身化，推动了传统数据中心向存算分离的分离式资源池化架构演化。（2）CXL作为一项创新性的互连技术，对未来计算架构和并行应用的影响是颠覆性的。首先，CXL支持CPU、加速器及I/O设备之间的缓存一致性以及节点间内存语义的高速互连，将推动以CPU为中心的计算模式向以数据为中心的异构对等计算模式转变。其次，CXL将改变未来数据中心的通信模式，使并行应用任务间大量的数据交换由传统值传递转变为引用传递，实现数据零拷贝，大幅降低数据中心网络负载。最后，基于CXL的分离式内存池支持分布式共享内存的编程模型，将推动并行与分布式应用广泛采用的消息传递编程模型向分布式共享内存的编程模型转变。

## 存储系统专题观点

构建资源全局共享的编码存储架构。编码存储能够以低冗余开销提高数据存储的可靠性，但是编码存储面临着数据恢复速度慢、部分条带写性能低以及快速设备下性能不稳定等问题。可以考虑通过存储架构创新，系统性地解决上述问题。具体来说，可以通过构建资源全局共享的编码存储架构，实现多用户对全局计算资源、网络资源和存储资源的共享使用，从而提升数据恢复过程中数据读取、传输、恢复计算和写入的速度；使用更多资源缓存条带写；为长尾延迟的I/O分配更多资源。

AI for Storage从点到体是未来发展趋势。存储系统随着各类应用兴起而不断演进，例如面向PC应用的直连存储架构、面向互联网应用的集中式或对等架构、面向移动互联网应用的三方架构等。如今AI应用的兴起正在推动存储系统进一步演进：从假设驱动的AI for Storage开始，简单应用机器学习方法，拓展可实现的功能边界，辅助进行存储系

统的优化；演进到数据驱动的 AI for Storage，利用存储系统运行过程中产生的海量数据，对其进行分析，洞察其中的规律并发现可优化的问题点，从而实现复杂存储系统的优化；接下来，如何从存储系统的点状智能优化发展到存储智能体，是极具挑战性且需要深入研究的问题。

AI 的兴起带来三大机会。首先，迫切需要数据高效预处理，为 AI 提供高质量知识数据，提升模型训练质量和推理精度；其次，AI 重塑存储价值曲线，激活数据价值，数据热度提升，需要构建高性能存储介质与系统；最后，AI 耗尽全量数据，促使数据留存比例提升，需要新型存储介质支持高倍率数据留存。AI 为存储系统提出新的量纲，即极致性能、数据韧性、全新数据范式、高扩展性、绿色节能、数据编织。未来需要联合定义新一代 AI 存储，针对应用 AI 场景，创新架构、接口、协议及介质。

立足评测技术创新，构建国内自主存储能力评测体系。随着应用场景不断拓展，特别是互联网应用、业务智能化等的发展，对存储系统的需求也不断变化，导致存储系统的综合负载压力快速增长。传统的性能、可靠性、扩展性等衡量指标难以覆盖存储全部能力评估，迫切需要开发新的规范和工具以衡量存储系统在数据保护、数据安全、协议兼容性、能效、业务相关优化、质量控制等方面的能力。

## 存储服务与应用专题观点

现实问题驱动云计算服务模式变革。当前云服务市场逐渐发展形成众多云服务商分割竞争市场的局面，供需双方问题频出。云用户面临的平台锁定问题日益尖锐，云服务商面临资源费效比持续升高的问题。相较而言，云用户更关注用户体验，追求云平台的稳定性、数据的高可用性、读写访问效率的高效性；而云服务商更关注系统效益，追求低存储成本、良好的传输效率等。

新型应用以及对单比特存储成本的极致追求需要新型的分布式文件系统。现有分布式文件系统的数据冗余策略带来的存储成本仍比较高昂。随着人工智能等新型应用的数据量不断增加，现有分布式文件系统管理百拍字节（PB）以上的数据困难重重，数据在多个节点之间传输，需要有效的数据管理和资源调度。数据驱动的新应用范式要求分布式文件系统满足存储访问高并发、高吞吐量和快速扩展的需求，同时能够处理多源异构数据的复杂交汇。交叉学科研究的不断深入和大模型的飞速发展带来各种多模态功能集成，分布式文件系统数据检索不仅需要支持关键字检索，还需要提供多模态检索功能。不同主体之间的数据高效实时协作处理，需要分布式文件系统支持数据确权、溯源、隐私保护等功能，在元数据、接口规范、数据汇聚协议等方面提升数据的互操作性。

以 CXL 为代表的统一内存总线技术将引发大规模计算系统体系结构的巨大变革。统一内存总线支持计算节点间的内存共享，从而为共享内存池的构建带来契机；跨节点的内存语义数据访问，可为节点间的消息通信提供多路径支持；内存语义接口的固态盘则会促进存储形态的全面革新。面对上述多方面变化，超算存储系统赢得前所未有的发展空间。从层次架构看，可新增一个距计算节点更近的内存池存储层；从 I/O 路径看，可借助多路径网络通信实现优化的 I/O 调度；从软件栈看，可重构基于全内存语义的 I/O 软件栈，实现 I/O 性能的显著提升。

数据中心存储采用 ZNS SSD，推进分区存储（zoned storage）技术的发展，提升存储效能。在数据中心存储场景，为简化分布式协议设计复杂度，底层分布式存储系统抽象的存储资源具有“追加写”特点，可以很好地使用 ZNS SSD，不再依赖 SSD 盘提供的块设备接口语义。从全栈设计角度来看，去除 SSD 提供的块设备抽象后，减少了抽象层次，在存储系统内部可以直接实现抽象逻辑资源到分区存储资源的映射。将 SSD 内部的数据布局任务交给存储软件，存储软件可以根据应用特点很好地实现数据布局以及垃圾回收（garbage collection）机制控制。总的来讲，ZNS 对数据中心存储带来的优势主要有：去除 SSD 提供的块设备抽象，减少资源抽象层；将数据布局任务交给存储软件，存储软件可

以根据业务特点灵活进行数据布局，配置不同的垃圾回收策略，控制写放大系数；SSD 只负责 NAND Flash 物理资源管理，并抽象成分区存储资源，可以通过存储软件定义分区资源类型（例如 TLC Zone 与 pSLC Zone 等），实现软件定义闪存，增强了数据中心存储的灵活性。

# 面临的挑战

## 存储介质与芯片面临的挑战

超快闪存电路设计与工艺集成。基于互补金属氧化物半导体（CMOS）加新型二维材料有望实现超快闪存，但面临三方面的技术挑战。首先，在电路设计方面，CMOS 电路产生脉冲的幅值和脉宽会受到电路最大电流和负载电容的限制，需要设计满足闪存超快擦写需求的外部电路；其次，在规模化 2D 超快闪存工艺方面，需要一套可扩展的超快闪存制造工艺，实现满足良率和一致性要求的超快闪存大规模阵列；最后，在 CMOS 与 2D 超快闪存集成的工艺兼容性方面，需要综合考虑 CMOS 和超快闪存工艺特点，优化集成工艺，满足集成工艺热预算要求。

存算一体系统集成与应用。存算一体新介质的技术发展取得了长足进步，但仍须解决系统集成与应用生态等方面的挑战。一方面，存算性能的发挥需要器件各方面性能同时达标，如读写精度 / 速度、读写功耗、保持特性等，全面性能优化难，而从独立器件到大规模阵列的集成也需要解决相关工艺问题。另一方面，模拟存算与数字系统存在硬件差异，在基础软件、神经网络算法设计等方面缺乏完善的生态建设，相关技术的落地应用仍面临挑战。

多维与超分辨光信息存储设备构建与产品化。基于多维与超分辨技术路径的光信息存储的相关研究已在超高容量、超低能耗、超高安全性、超长寿命等方面取得一定的成果。但是，仍须解决构建存储设备时面临的技术成熟度较低、读写速度低、介质材料大规模制备等问题。此外，目前基于多维与超分辨技术路径的光信息存储的光路系统规模大且关键元器件对国外依赖比较严重，使得如何小型化和国产化成为亟待解决的重要挑战。

DNA 存储信息写入受到限制。通过合成 DNA 分子写入信息面临两方面的挑战。在技术层面，DNA 分子信息串行写入，存在串行合成繁琐、反应复杂耗时、序列长度受限、写入成本高昂等问题；在知识产权层面，现有很多“从头合成”DNA 信息写入技术的相关核心专利在国外已被提前布局，使得 DNA 存储写入技术的国产化面临挑战。

## 存储器与设备面临的挑战

全新的持久内存数据管理机制。（1）持久内存系统的数据一致性保障涉及崩溃一致性与并发一致性两方面。首先，持久内存系统需要确保在发生故障重启后，仍然能访问到正确的数据，以保持崩溃前后数据的一致性。其次，对于多核或分布式系统，持久内存系统需要保证每次数据更新可以及时在其他节点可见，使不同节点的每次访问都能获取到最新数据。（2）相比于传统易失性 DRAM 内存系统，持久内存系统在成本、容量和持久化存储等方面具有显著优势。然而，持久内存也表现出了一些固有的硬件特征，如读写不对称、读写放大、高缓存刷新延迟以及非一致性内存访问（NUMA）远端访问速度慢等。如果在构建持久内存系统时未充分考虑这些硬件特征，可能会无法完全释放持久内存设备的硬件潜力，导致系统的性能表现低效。（3）持久内存系统的持久化存储特性带来了新的数据安全威胁，不仅增加了数据完整性被破坏的风险，也提高了数据泄露的可能性。攻击者可以通过软件漏洞恶意修改持久内存中的数据，这些被篡改的数据不会因系统重启而消失，并且在系统重启时会重新加载，为二次攻击提供便利。此外，如果持久内存设备缺乏数据加密机制，一旦设备被盗，攻击者将能直接访问其中的敏感数据。

高密度闪存赋能大模型具有局限性。（1）高密度闪存可靠性挑战大，目前高密度闪存的发展受寿命、可靠性和性能的严重约束。随着密度的进一步

增长，可磨损次数甚至不到千次，而可靠性则受制于制程差异等，导致数据丢失。另外由于电子逸出、温度、湿度的影响而产生的数据访问性能问题，将会导致闪存的发展严重受阻。（2）系统与介质的结合松散。经典的设备和系统的发展采用的是分离式发展方式。系统的设计人员假设基于一个典型的存储设备设计系统，而介质设备的开发人员则为提供一个能够满足系统基本需求的设备而努力。然而随着应用的需求提升以及存储设备的发展，传统的松耦合方式已经难以满足系统和存储的需求。（3）支撑大模型的高密度闪存的挑战进一步放大。大模型有不同于传统的应用特征，它不仅需要存储海量的数据，同时在模型的训练和推理过程中存在密集的数据操作。这种带宽需求和存储需求基于传统的存储系统设计已经发生了根本变化。

DPU产品生态和存储任务卸载。（1）在技术路线上，基于网卡的DPU一般分为两类，第一类是NIC与FPGA或CPU核心依靠PCIe等技术相连的混合解决方案，第二类是NIC和处理器核心高度集成在一个片上系统（SoC）的方案。由于DPU相关的标准和生态不一致，导致DPU厂商的相关产品还是百家争鸣的状态。例如，英特尔的相关产品采用FPGA/ASIC的加速引擎和ARM处理芯片设计，云豹自主研发的DPU芯片采用创新的层级化可编程设计并融合自主研发的RISC-V指令集进行设计。DPU产品的供货商如英伟达、AMD、英特尔、Marvell、云豹智能和中科驭数等都在凭借着更高的性能和AI服务器的热潮，不断推出创新的DPU方案设计。（2）存储任务卸载的收益不确定。DPU算力不及节点CPU的算力，DPU上的FPGA和RISC-V等计算加速引擎的计算能力有限，片上ARM众核参与计算也会带来内存拷贝开销。目前企业界和学术界已有相关研究利用DPU和智能网卡等设备卸载键值存储系统中的压实操作，以减轻主机CPU负担。例如，2021年阿里云发布了第四代神龙架构MoC卡/神龙芯片CIPU，实现了网络和存储操作的完全硬件卸载，进一步增强了网络I/O性能，并率先支持大规模的弹性RDMA加速，应用在盘古云存储系统中。目前通过测试已验证了采用DPU直通NVMe（非易失性内存主机控制器接口规范）设备，进而实现数据面旁路CPU的技术方案可以有效减少数据移动开销，提高系统数据吞吐率。但是，在基于DPU存储任务的卸载方面，特别是数据重删、压缩和编解码等存储任务的卸载方面是否会带来收益还有待进一步研究和探索。

基于CXL总线的内存池是关键挑战。（1）高速互连总线CXL的硬件研发远远滞后于技术规范。CXL互连总线技术对硬件生态的影响非常深远，不仅需要CPU、加速器设备支持CXL接口，资源池化更需要CXL交换机的支持。CXL3.1技术规范白皮书已经于2023年11月推出，但目前的硬件设备仅仅可支持CXL1.1，严重制约了基于CXL总线的系统和软件的研发，支持CXL2.0分布式内存池化的系统尚不多见，阻碍了需要大容量内存的AI大模型、大数据处理等应用层面的研究。（2）高速互连总线CXL给操作系统的内存管理带来巨大冲击。传统操作系统仅仅需要管理本地单一形态的内存资源，但高速互连总线CXL将使得操作系统需要管理本地和远端的多层级池化内存资源。由于同一计算节点上的应用对于池化内存资源的视图不统一，而且存在不同的访存接口和访存特性，使得操作系统对内存资源的管理比较困难。此外，由于池化内存容量急剧增大，而计算节点CPU的转译后备缓冲区（TLB）不具有可扩展性，TLB覆盖率下降使得地址转换开销变大。（3）高速互连总线CXL给并行编程模型带来巨大冲击。传统的并行与分布式应用往往采用消息传递模式实现并行任务间的通信，造成大量的数据交换，对网络资源的消耗较大。基于高速互连总线CXL的分布式共享内存池系统天然地支持分布式共享内存的编程模型，因此需要重新审视传统的并行与分布式应用编程模型以适配CXL环境。

## 存储系统面临的挑战

构建资源全局共享的编码存储系统存在三点主要挑战。第一是数据布局管理的挑战，即如何在多用户细粒度共享物理资源池的同时保证系统的低元

数据存储开销；如何在静态数据布局基础上适应 I/O 负载的动态变化，进而实现 I/O 的动态调度。第二是恢复任务调度的挑战，在大规模编码存储系统中，负载特征复杂多变，存算节点数目众多，恢复任务总量巨大，要求恢复任务调度器能够以较低的时间开销找到较优的恢复调度策略。第三是如何保证中位延迟不增加的前提下，优化长尾延迟和系统总体 I/O 的性能与稳定性。

根据多 AI 智能体和人类认知系统双系统联动的启发，在上层不同计算模式的牵引下，通过在垂直领域定制相对完整的多个智能子系统，并在智能体内进行协同联动与自我进化，是突破当前数据驱动的 AI for Storage 的点状智能、实现存储智能体的可能途径。但是，如何对不同计算模式的共性进行提炼和表达，划分通用性与个性化，以及如何探究 AI 的静态和动态双重可解释性，是当前面临的巨大挑战。

新一代 AI 为存储系统带来新痛点。为提升模型训练和推理精度，需要高效预处理数据，但是目前低质量数据占比高，多模态数据增长快，数据清洗流程复杂，导致数据预处理性能无法满足要求。在大规模集群场景中，盘故障、节点故障是常态，一块盘故障能导致整个训练集群等待，本地盘 / 缓存加速方案难以满足训练集群性能、可靠性等诉求，急需高性能存储集群。大模型必须结合私域知识库存储，才能满足推理准确性、实时性、安全性诉求。

国内 AI、CPU、整机评测榜单已陆续发布，但存储领域依然处于空白状态。当前国外存储排行榜多以考核性能为主，其评测维度和评价方法无法充分涵盖实际业务场景对存储产品实际能力、特性的关注点。如何针对多种应用场景制定评测规范、开发可定制可扩展的评测工具成为难题。

## 存储服务与应用面临的挑战

算力网云际存储面临诸多数据流通壁垒，需要解决地域分布广、主体归属多造成的庞大数据流通难、不好用的问题，数据归属主体不同导致的不敢、不愿流通的问题，存储载体和服务异构导致的数据出入难的问题，数据跨地域分散无全局视角导致的数据访问效率低的问题，数据“上网”导致的数据安全可靠性低的问题，以及数据跨域流通导致的数据传输效率低的问题。

分布式文件系统面临诸多挑战。现有的分布式文件系统一般采用统一的数据格式管理数据，而数据多源异构导致数据格式不一致将是未来应用的常态，特别是开放科学应用场景。对于数据所有权属不同的主体，现有的分布式文件系统缺乏细粒度的存储访问控制手段，缺乏高效的数据确权、数据索源和不可篡改等手段。对于百 PB 级别的海量数据，现有分布式文件系统在系统按需扩展、数据全局管理方面的不足，导致数据迁移开销大、单比特存储成本高等问题。

统一内存总线技术推动大规模计算系统 I/O 软件栈的全面重构。在统一内存总线技术的支持下，存储系统的层次特性将发生显著变化，灵活的网络拓扑将为 I/O 请求提供很大的调度空间。然而现有的 I/O 软件栈并不能适配新的体系结构，发挥统一内存总线技术的优势需要重构当前的 I/O 软件栈，包括重新定义数据访问接口，对应用屏蔽系统的底层细节，在多层次存储中为多样化的应用定制数据空间、提供相适应的 I/O 请求执行路径等。

对于后端分布式存储系统来讲，目前面临的最大网络问题是互连延迟过高，并且随着每秒读写操作次数（IOPS）的提升，网络延迟会急剧增大，已经成为整个 I/O 链路延迟的大头。如何进一步提升网络资源利用率、提升网络性能和服务级别协议（SLA）、降低延迟成为存储网络关注的重点。各家公司和研究机构针对该问题提出了高性能网络协议，延伸和发展了现有 RoCE（RDMA over Converged Ethernet，一种基于以太网的远程直接内存访问协议）技术。国际组织超以太网联盟（Ultra Ethernet Consortium, UEC）于 2023 年成立，旨在共同推动超高性能以太网络技术发展，解决 AI、存储等场景面临的高性能互连问题。

# 发展建议

## 存储介质与芯片发展建议

已报道的超快闪存技术尚处于单器件级原型验证阶段，而超快闪存的规模集成需要可靠的外围电路来满足读取和擦写需求。现有 CMOS 技术成熟稳定，适合为超快闪存集成提供外围电路支撑。因此，需要推进 CMOS 技术与超快闪存技术的集成，支撑新型闪存存储技术迈向规模化与系统化。

探索存算一体技术与类脑计算技术的结合。存算一体的技术特点非常适合类脑计算的硬件实现。新兴存储介质不但可以高效模拟神经元和突触单元，存算一体阵列的运行方式也可以更为自然地体现并行计算、事件驱动等类脑计算特性，比 CMOS 类脑计算的实现方式更有优势。探索存算一体技术与类脑计算技术的结合有望推动相关技术领域的重要突破。

持续投入多维与超分辨光信息存储技术研究。目前国内的多维与超分辨光信息存储技术仍处于原理样机阶段，距离未来产品化还有一段路要走。需要对相关研究进行持续投入，期望在“十五五”规划时研制出下一代光信息存储原型样机。

加强 DNA 存储关键技术攻关。DNA 存储虽然有着诸多优势，但是目前仍面临着读写速度慢、成本高等问题，距离广泛推广应用还有诸多关键技术需要攻关。作为国家科技战略布局，仍需要围绕 DNA 存储关键技术，如并行酶催分子合成（信息写入）、新型分子测序（信息读取）、DNA 信息编解码和纠错、多介质 DNA 存储，不断朝着 DNA 存储实用性的方向迈出坚实的一步。

## 存储器与设备发展建议

加强持久内存研究并推动应用。（1）推动持久内存在分离式内存架构下的研究。分离式内存架构将单个服务器中的计算和内存资源解耦成独立的计算和内存池。其中每个计算和内存池都可以灵活部署和扩展，从而提高资源利用率和故障隔离性。持久内存技术不仅能够扩展内存池容量和带宽，还具备持久性，避免了内存池数据写入后端存储的开销。因此，推动持久内存在分离式内存架构下的研究，可以为实现高性能、可弹性扩展的内存资源供应提供更多解决方案。（2）推动持久内存在大模型应用中的研究。大模型应用需要存储大规模的参数、模型以及训练数据，要求存储系统具备大容量的存储能力。同时，由于大模型应用通常涉及长时间的模型训练过程，存储系统还须确保数据的一致性和可靠性，以避免数据丢失或损坏带来的重训练开销。持久内存具有高速、持久、容量大等优势，为解决大模型应用的存储需求提供了新的可能性。通过深入研究持久内存在大模型应用数据管理等方面的问题，可以有效提升大模型应用的存储效率和性能。

推动高密度闪存的垂直设计及与大模型的融合。在推动高密度闪存的垂直设计方面，我们迫切需要开展从高密度闪存介质到系统的深入融合。随着未来各类场景对存储容量、寿命和性能要求的上升，我们要鼓励闪存芯片团队、闪存控制器团队和系统设计团队的深度融合，构建联合团队，推动从芯片设计、控制器设计到系统设计的垂直建设，从而为新一代存储需求场景提供质量更高的存储服务。基于国产的高密度闪存存储系统，开展针对大模型场景的存储系统设计，通过对大模型计算、访问、存储需求的分析，构建专用的存储计算功能、访问路径设计以及存储需求控制管理技术，从而为新一代大模型发展提供关键的技术底座。

探索基于 DPU 的网存算协同和存储池化。当前 DPU 在数据中心的虚拟化和计算加速等方面有着相对明确的需求和体系架构，但在存储场景下如何减少数据移动和加速数据处理值得深入探索。未来发展方向主要包括两个方面：（1）网存算协同。充分利用 DPU 的硬件卸载、数据重删、压缩和编解码等存储任务加速能力，协同好主机 CPU 和 DPU 间的任务调度，降低主机数据处理开销，提升 I/O 效率。（2）存储池化。随着 DPU 等新型网络硬件和 CXL 协议等高速互连技术的快速发展和应用，计算和存储资源逐渐解耦，进一步推动了存储设备的池化架

构，实现存储资源的按需横向扩展，提高存储资源的利用率。

加强高速互连总线建设。（1）学术界和产业界要加大对高速互连总线研究的投入。尽管高速总线互连技术CXL为未来计算描绘了令人振奋的蓝图，但基于CXL总线的分离式内存系统仍面临体系结构、操作系统、编程模型等诸多挑战。此外，CXL设备研发远滞后于技术白皮书，这给基于CXL的分离式内存系统研发和应用带来了诸多困难。当前，国际上对总线互连技术CXL/UB的相关研究才刚刚兴起，我国科研机构和硬件设备提供商应加大研发投入，并大力发展具有自主可控能力的国产总线标准，才能在计算/内存架构的国际竞争中占据前沿高地，为我国构建新一代数据中心架构的技术体系和标准打下坚实的基础。（2）推动高速互连总线应用生态的建设。AI大模型、大数据处理等新兴应用需要往往需要TP级甚至PB级的内存资源才能保证应用的高效执行，使内存墙问题成为制约这些应用性能提升的瓶颈，因此迫切需要大内存系统。CXL/UB等高速互连总线为分布式内存扩展提供了新的机遇。尽早布局和开展基于CXL/UB的分布式共享内存的应用优化研究，消除应用移植和生态建设的壁垒，将有助于推动高速互连总线软硬件的推广应用。

## 存储系统发展建议

建议研究资源全局共享的编码存储系统的设计方法，通过设计静态布局函数、基于中继节点的恢复任务调度、静态布局函数和动态映射表相结合的数据布局等方式，解决编码存储中存在的挑战。

围绕如何实现存储系统体状智能的研究目标，研究任务驱动的AI for Storage，构建多AI存储智能体，实现智能体内多系统协同联动与自我进化。需要重点突破任务驱动的智能存储系统架构、自适应的资源配置、智能存储系统的软硬协同、智能存储系统的任务感知与进化等关键技术点。

随着传统存储到AI存储的演进，数据生产模式由以人为主变成以AI为主，数据交互主体由人变为AI，交互方式和内容从低维数据共享转变为高维语义检索。这些变化对定义新一代AI存储范式推动存储架构、接口、协议、介质创新提出迫切需求。

建立专业存储测试规范。首先，需要制定一套规范，既描述通用的指导原则，又具备面向场景和指标的扩展能力，能够针对多种应用场景开发相应的评测工具、评测流程、负载模型。可扩展至AI、高性能计算、数据库（SQL数据库、NoSQL数据库、向量数据库、图数据库等）、大数据处理（Hadoop/HDFS等）、云计算（虚拟化/对象存储）等场景/领域。其次，需要开发一套可定制、可扩展的工具，能够针对业务场景提供Trace采集、存储、分析、重放等综合服务，支持多用户、多任务、多实例的任意模式多级组合的负载播放和仿真测试。从性能（任意读写模式组合的带宽、延迟、IOPS等）、可靠性（数据冗余、热备、恢复能力）、生态协议、数据保护、扩展性、能效等多角度实施一体化综合测试。探索并实现在性能/读写测试中集成存储功能/特性的测试和验证的模型、流程和方法。

## 存储服务与应用发展建议

加快推动算力网云际存储多领域共同建设：算力是基础，加快推动布局全国一体化算力网络国家枢纽节点,加快实施“东数西算”工程;数据有价值，促进数据要素高效流通以挖掘潜在数据效益；能耗是关键，通过数算联动实现绿色低碳的数据存储管理和分析处理。针对跨多管理域、多算力中心的数据如何流得动和用得好的问题，须突破兼容多种异构存储服务的高可靠云际数据存储技术，构建支撑云际数据集约化汇聚和对等协作处理的高效能算网融合机制，形成数据存储、传输、处理于一体的云际数据协作关键技术体系，为服务数字经济、东数西算等国家战略提供理论和技术储备。

统一分布式文件系统，为所有数据实现一个统一的存储访问，支持跨区域、跨单位的数据安全访问。调研显示，统一存储是解决企业存储架构面临的成本高、可拓展性差等诸多痛点的重要途径，被认为是数据存储与管理的未来方向，而通过研发能处理各种类型数据的统一分布式文件系统有助于实

现统一存储。

在统一内存总线技术支持下，内存、网络（基于CXL的网卡）、I/O（基于CXL的SSD）均可通过内存语义来访问。在此背景下，可以研发全内存语义存储系统，在数据缓存、传输、持久化过程中完全基于内存语义实现，规避块级数据访问缺陷，支持任意粒度的数据访问，降低不同内存语义和块接口语义之间的转换开销。

研究高性能存储网络的协议及控制器芯片、提升网络资源利用率，研究内容包括网络层分布式流控技术，基于报文（message）传输协议，拥塞控制协议，乱序传输机制，多路径、灵活可编程的网络控制器等内容。通过网络协议、高性能网络控制器的研究，解决网络延迟高、网络突发流量、性能长尾、资源利用率低等问题。 ■

**整理：**周　可　何水兵　王　桦　陈志广　魏学亮　洪志明

**特邀嘉宾：**

郑纬民　孙凝晖　金　海　舒继武　李　辉

**参会嘉宾**（按姓氏拼音排序）：

陈仁海　陈志广　何水兵　蒋德钧　林　芃　刘海坤　刘　俊　毛　波　石　亮　孙　斌　王　桦　王意洁　武永卫　吴忠杰　曾令仿　翟季冬　张　成　张　弓　张广艳　张启明　周　鹏

**秀湖会议学术委员会主席：**胡事民

**本次会议执行主席：**周　可

**本次会议联合主席：**冯　丹　肖　侬

**本次会议工作人员：**魏学亮　洪志明

（本文责任编委：**朱弘恣**）

# CCF第十三届常务理事会第三次会议在京召开

2025年1月19日，CCF第十三届常务理事会第三次会议在北京召开，常务理事出席会议，监事会、执行机构负责人、奖励委员会主席列席会议。会议由CCF理事长**孙凝晖**主持。

CCF秘书长**唐卫清**报告了CCF近期工作进展；副理事长**胡事民**报告了亚太计算机学会联合会筹备进展报告和党委工作情况；副理事长**金海**报告了学术出版提升计划；奖励委员会主席**李晓明**报告了2024年度奖励工作；监事长**金芝**作了监事会工作报告，并对学会及常务理事会工作提出建议。经过表决，本次会议通过了计算经济学专业组申请升级专业委员会和成立CCF教学成果推荐委员会的动议，《CCF奖励条例》和《中国计算机学会财务管理条例》的修订动议，以及制定《CCF关联交易管理办法》的动议，以进一步加强对CCF的内控管理。会议还通过了其他若干动议。

会议围绕提升学术会议的质量和国际化水平、推进国际学术合作、加强学术出版工作、计算机博物馆建设、教学成果推荐等方面展开热烈讨论。孙凝晖强调，未来CCF将不断提升会员服务质量、增强会员获得感，让更多会员在CCF的平台上获得发展。

下次常务理事会将于2025年7月5~6日在成都召开。