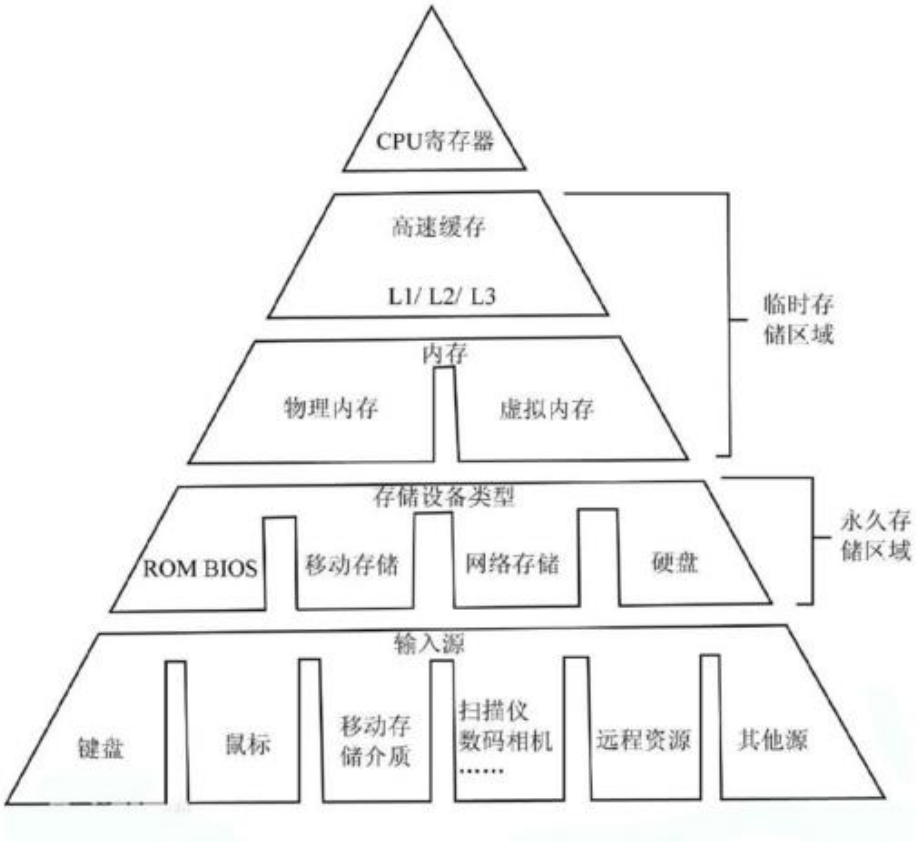
面向海量数据的高效存储技术

武汉大学测绘遥感信息工程国家重点实验室 2017 级 徐贝妮

班级：2017 级 硕士三班　　学号：2017286190090

1. 计算机存储系统的发展

存储器（Memory）计算机系统中的记忆设备，用来存放程序和数据。计算机中的全部信息，包括输入的原始数据、计算机程序、中间运行结果和最终运行结果都保存在存储器中。它根据控制器指定的位置存入和取出信息。如下图所示，在计算机系统中存储层次可分为高速缓冲存储器、主存储器、辅助存储器三级。高速缓冲存储器用来改善主存储器与中央处理器的速度匹配问题。辅助存储器用于扩大存储空间。



自世界上第一台计算机问世以来，计算机的存储器件也在不断的发展更新，最初采用串行的延迟线存储器，不久后采用磁鼓存储器。50 年代中期，主要使用磁芯存储器作为主存。60 年代中期以后，半导体存储器已取代磁芯存储器。在逻辑结构上，并行存储和从属存储器技术的采用提高了主存的供数速度，缓和了主存和高速的中央处理器速度不匹配的矛盾。1968 年 IBM-360/85 最早采用了高速缓冲存储器——主存储器结构。高速缓冲存储器的存取周期与中央处理器主频周期一样，由硬件自动调度高速缓冲存储器与主存储器之间信息的传递，使中央处

理器对主存储器的绝大部分存取操作，可以在中央处理器和高速缓冲存储器之间进行。1970 年，美国 RCA 公司研究成功虚拟存储器系统。IBM 公司于 1972 年在 IBM370 系统上全面采用了虚拟存储技术。从一开始的汞延迟线，磁带，磁鼓，磁芯，到现在的半导体存储器，磁盘，光盘，纳米存储，虚拟存储等，无不体现着科学技术的快速发展。而面对当前信息爆炸的大数据时代，海量的数据计算需求对存储系统的发展又提出了新的挑战。

2. 当前存储系统的研究

1）多元存储介质的研究

近年来，随着半导体存储技术的发展，对存储系统的研发的热点逐渐转移到新型存储设备的研发中，并且这些新设备一些科学或者商业应用中得到了较为良好的应用效果，存储介质当前已经向多元化的方向发展。例如非易失存储器（Non-Volatile Memory，NVM）和设备类非易失内存 SCM（Storage-class memory）相继问世，主要有微电子机械系统（Micro Electromechanical System，MEMS）、带后备电池的内存存储器（DRAM）、闪存型固态盘（Flash based Solid State Disk，Flash-SSD）、相变存储器（PCM，PRAM 或者 PCRAM）、磁性存储器（Magneto-resistive Random Access Memory，MRAM）、铁电存储（Fe RAM）、电阻型随机存储器(Resister Random Access Memory, RRAM)与和自旋转移器(spin-transfer torque memories，STTM)。由于硬盘等受到的数据寻道时间等机械特性限制，采用新型存储介质存储数据，传统磁介质已经不是研究的重点。新型的存储介质以其具有随机访问能力高、功耗低以及存储密度的可提升潜力大以及非易失等综合特性而得到当前研究者的注意。

2）存储体系结构的研究

千万亿次超级计算系统对存储系统提出了大容量、可扩展的存储服务等多种 I/O 需求：要求存储系统需要支持用户数据的按需存储，需要提供可扩展的聚合 I/O 能力，需要满足科学计算模拟和数据处理应用的多样化 I/O 需求，能够有效支持海量存储空间共享和高性能 I/O 访问。这就存在一个大规模并行存储系统中的平衡设计问题。美国能源部（DOE）早在 2001 年就对 I/O 带宽均衡设计提出了明确的要求。国外的 IBM，SGI；国内的国防科大，曙光等超级计算机研制单位，在存储系统的设计时，都非常注重存储与通信，计算的均衡设计问题。

随着存储规模的扩大，并行存储系统中存储系统的层次也在变多，导致同一存储层次内部以及不同的存储层次之间的存在过多的数据副本数量，导致缓存的数据管理以及存储数据一致性维护日益困难，那么结合应用，重新审视和设计存储体系结构是当前的必然发展趋势。

3）存储性能优化的研究

并行存储系统 I/O 性能是影响超级计算实际应用性能的关键要素。从过去的研究来看，提高存储系统性能的方法主要有缓存、预取、并行 I/O 指令调度等。缓存技术可以使存储系统响应时间减少大约 90% ；当前也有很多研究如何利用远程内存，来减少磁盘读写的缓存技术，从而有效地减少 I/O 的延迟。存储系统优化的一个重要研究方向还包括将预取与缓存紧密结合。预取和缓存的优化效果与应用 I/O 访问模型是密切相关的，超级计算下存储优化的研究方向还有结合访问模型的预取策略和根据访问模式的自适应缓存策略。但是这些预取方法仅仅是根据用户的访问模型来进行设计，还没有考虑存储介质的特性以及文件本身的特性。随着当前多种非易失性存储介质的出现，根据存储介质的特征以及文件特性进行 I/O 优化成为当前的一个研究重点。

4）存储功耗优化的研究

随着计算规模和性能的不断提升，功耗问题日益严重，倍受重视。低功耗的微电子设计方法以及系统节能技术一直是超级计算领域的研究热点。近几年超级计算更加注重绿色环保概念，推出了全球 Green 500 超级计算机排名。过高的功耗给超算系统设计和运营带来严重的问题。

随着存储系统规模的扩大，功耗也正在制约着存储系统的发展，以 64 盘位的盘阵为例，功耗大约 2500 瓦，带宽 3G 左右，如果要构建一个带宽为 1T 的存储系统，需要 330 台这样的盘阵。不考虑系统中网络扩展和服务器带来的功耗开销，仅盘阵本身的功耗大约 800KW，运行一天需要的电量近 2 万度。而存储系统的功耗还包括数据服务器和网络传输等各个环节的功耗。这些功耗并不比数据的存储功耗低，因此存储系统的低功耗研究越来越受到重视和关注。

3. 海量数据存储应用背景

随着大数据时代的来临，科学研究、工业应用和网络服务等领域的数据规模正爆炸式增长。当前，超级计算机在科学研究和国民经济建设中发挥着非常重要

的作用,应用领域十分广泛,包括卫星遥感数据处理、金融数据分析、气象预报和气候预测、海洋环境数值模拟、土木工程设计、新材料研发、基础科学研究等。然而,随着数据规模不断扩大,超级计算机在处理大数据应用时仍暴露出一些突出的问题。超级计算机多采用集中共享式存储系统，计算节点通过内部高速互连网络与存储系统相连。以 Lustre 文件系统为例,通常,每个 Lustre 系统的容量为 1-4PB,实际运行情况表明:当 Lustre 文件系统容量使用率大于 70%时,存储系统会出现不稳定现象,使得超级计算机系统的稳定性和可用性降低。现在,多个典型的大数据应用均对超级计算中心的大数据存储提出了越来越高的要求,如要求总的存储容量达到 10PB 甚至以上,能与超级计算机的 Lustre 系统有机融合等。

存储系统是超级计算机系统中负责数据保存以及支撑应用程序进行数据输入输出的软硬件系统。存储系统的数据存取效率是影响超级计算机系统实际应用性能的关键要素。大数据时代很多超级计算应用已成为 I/O 密集型或数据密集型应用，需要非常庞大的数据存储、处理能力以及 I/O 带宽，对存储系统的 I/O 性能提出了日益苛刻的应用需求。

一方面计算性能，访存性能和网络通信性能快速发展，而另一方面构建存储系统的磁盘性能停滞不前，严重阻碍了千万亿次超级计算能力的有效发挥。所以,我们需要研究新的存储技术和存储结构,有效解决 10PB 到 100PB 级海量数据存储的重大技术挑战,为越来越多的大数据应用提供存储和处理服务。在研读相关论文后，本文将介绍几种面向海量数据的存储技术的思考与创新，包括以下几个方面:(1) 并行存储系统性能分析模型与优化策略；(2) 元数据服务系统优化技术；(3) 基于闪存的存储优化；(4) 存储功耗模型与优化技术。

4. 并行存储系统性能分析模型与优化策略

大规模数据处理是一个非常耗时的计算过程,使得传统的单机系统远远无法满足大数据对计算性能的要求。因此,需要研究提供高效的并行化大数据计算技术方法与系统。大数据的有效分析利用通常涉及到对大规模数据的分析挖掘,而巨大的数据量使得传统的单机机器学习和数据挖掘算法都难以在可接受时间内完成计算,导致算法失效。因此,需要研究提供有效的并行化大数据机器学习与分析挖掘算法和大数据机器学习系统。大数据处理不同于传统的计算与信息处理技

术的另一个重要特点是,它是一项涉及计算与信息处理技术众多方面的综合性技术,具有显著的技术综合性和交叉性特征,以任何一个单一和隔离的技术层面和技术方法,都难以有效完成大数据的处理。因此,大数据的有效处理需要将存储、计算与分析层面的技术紧密结合、交叉综合,以形成一种完整的大数据处理技术栈,构成一体化的大数据处理系统平台。

针对存储系统性能优化当前有很多研究工作,从存储体系结构的角度看主要包括几个方面，有的研究如何通过提升 Cache 的命中率提高存储系统的性能，有的研究如何提升元数据服务器的管理能力,如降低元数据服务器的运算量或提高 I/O 处理能力等,有的研究如何利用本地存储等。这些性能优化策略和方法，一般是面向某一类问题或者某些应用提出的改进策略,在超级计算机并行存储系统中应用时，往往解决了某一些问题，但是又引起其它新的问题。同时这些优化策略缺乏量化指标，很难就优化的性价比给出评判。而且超级计算机存储系统十分庞大，对系统的优化升级往往需要数日才能完成，很多性能优化方法缺乏可行性。

在此介绍一种一种可定义的存储架构,可解决前文所提到的问题。可定义存储系统分为 I/O 客户端层（主要由计算节点构成，还包括登陆，管理节点等)、数据加速层和共享数据存储层。数据加速层，由多个可统一编址的 I/O 服务加速节点（ION）构成。I/O 服务加速节点挂接大容量高速闪存阵列，用于满足了应用程序瞬时 I/O 操作的性能需求，同时连接高速互连网和 IB 存储网。N 个 ION 节点构成一个应用需求的数据加速层。共享数据存储层由 2 个互为热备的元数据服务器和多个连接 200T 容量盘阵的数据服务器构成，为应用程序提供了足够的存储空间和持续数据读写性能。在新的存储结构中，计算节点需要通过数据加速层才能访问到数据。数据加速层利用 I/O 服务节点处理或转发计算节点的 I/O 请求，使并行文件系统的客户端数目得以适度调控，提供可均衡扩展的快速 I/O 能力；虽然在数据加速层，由于数据转发等会增加数据通路上的时间，但是根据“规则”底层的共享层性能会大幅提高，从而总时间有可能减小，同时由于数据加速层的缓存策略,数据在加速层命中后,就不再访问共享数据层,同样可以减少数据的访问时间。

在存储系统层次式架构的基础上，文件系统将 ION 节点的本地存储和共享

数据存储层的地址统一管理起来，构成一个完整的命名空间。每个 ION 配置 SSD 存储设备作为本地存储，构成一个数据处理单元（Data Processing Unit，DPU）。在 Y2FS 文件系统中，由多个 DPU 联合提供的统一地址的虚存储空间，被称为近端存储（Local space），因为从任务运行的角度看，这些存储空间就像是计算节点的本地空间一样快。在路径上，DPU 距离计算节点更近，一般情况下每 N 个 DPU 单元对应 M 个计算节点进行服务。N 和 M 的比值是由应用需求的性能指定的。一般情况下为了避免木桶效应的产生，N 个 DPU 应该选用距离 M 个计算节点最近的，且最好是距离一样远的。利用多个 DPU 提供的 Flash 存储空间，可以为应用程序的局部性访问提供更好的支持，这个工作是由虚目录空间服务器负责的，同时管理员也可以通过手动的配置来选择 ION 的数量和位置。DPU 的使用满足了大规模数据密集型应用的扩展性要求，缓解了几万个计算节点同时访问服务器带来的 I/O 压力。

当用户需要提交规模为 M 的作业在超级计算机上运行时（M 表示需要占用的节点个数），首先作业管理系统通过访问虚目录空间服务器。虚目录空间服务器根据用户的 IO 带宽的需求或者作业的规模（默认情况根据作业规模），从数据加速层的 ION 服务器中选择可以使用的 N 个 DPU 构成任务的近端存储，并为作业运行提供虚存储空间。当作业运行时，DPU 从共享数据存储层的数据分区中获得资源数据送给计算节点，计算节点将计算产生的数据，发送给近端存储的虚存储空间。此时 DPU 可以提供两种策略转发模式（Forward mode）和异步模式（Asynchronous mode）。前者 DPU 在将数据存入到本地空间时，还会将数据发送到共享数据层。后者产生的数据将一直保留在虚空间中，直到任务完成，空间被收回时，才会根据需要将需要的数据写回到后端存储。

文件系统将 DPU 存储空间和全局存储空间进行统一编址管理，充分利用 DPU 的可扩展 I/O 能力和全局存储空间的可靠性，构造高效、可靠、统一的融合虚拟存储空间，可以有效提高大规模应用条件下系统整体的 I/O 性能。客户端产生数据可以分布在 DPU 上，也可以分布在全局存储空间上。在统一目录视图中，数据文件可以单独存在于局部存储空间或全局存储空间，也可以同时存在于局部和全局存储空间形成副本。文件的元数据信息在局部存储空间和全局存储空间共同维护，从而实现了对全局存储空间的一致性、透明性访问。

文件系统不仅支持传统的 POSIX 访问接口，还提供了布局接口，策略接口以及对 HDF/MPI-IO 的支持。布局接口指导应用程序数据在命名空间中的分布，数据可以直接保存在全局存储空间中，也可以保存在 DPU 的局部存储空间中。DPU 局部存储空间可以一直保存应用程序数据，也可以将数据刷新到全局存储，混合层次式文件系统还提供了相应的指导命令。策略接口用于设定调度策略，层次管理策略，数据保存和迁移策略。这些策略使得应用程序更便于使用混合层次式文件系统。除此以外，混合层次式文件系统还提供了对 HDF/MPI-IO 的支持，并面向这两类接口进行了有针对性的优化。上述扩展接口采用库的方式实现，应用程序在编译时直接连接库即可调用相关的接口。

4. 元数据服务系统优化技术

由“设计 4 规则”，并行存储系统的存储带宽与元数据服务器的处理能力成正比，提升元数据服务器的处理能力可以提升系统的带宽并能有效降低延迟。因此构建大规模并行存储系统时必须对元数据服务器进行优化，使其在所属的存储系统中发挥最大的能力。

元数据服务器在大规模存储系统中主要有两个工作：逻辑上负责整个系统的目录结构，权限控制和文件命名等；物理上需要有高吞吐率的 I/O 处理能力，对元数据进行读写。其数据处理能力取决于 CPU 的频率和访存带宽，访问存储介质的延迟开销和带宽。CPU 对内存的访存带宽可以通过更多的内存和处理器来实现，对存储介质的访问带宽可以通过放置更多的存储设备来实现，但是降低存储介质的延迟开销，在过去很难真正实现数量级的提升。

针对商用的服务器不是受限于计算能力、就是受限于 I/O 能力，难以有效管理规模急剧扩展的存储设备；而基于多服务器的分布式元数据管理技术受限于一致性开销过大、协议过于复杂，在工程实现中难以实用化的问题。基于“设计 4 规则”定理 4，按照分区并行、集中管理的设计思路，提出了基于闪存存储阵列和大 SMP 节点服务平台的元数据分域处理技术，在单个元数据服务器上支持高并发的多分区元数据服务，大幅提高了元数据服务系统的吞吐能力，从而显著提升整个存储系统的 I/O 性能。

5.基于闪存的存储优化

新型层次式可定义存储架构利用 I/O 服务节点 ION 的本地闪存阵列作为数据服务加速层，提升数据服务器的 I/O 带宽，减少从计算节点到数据服务器的数据传输距离，满足应用程序瞬时 I/O 操作的高速 I/O 带宽需求。

系统访问作为采用存储介质的闪存颗粒，有两个主要的方法：一通过闪存的 FTL 转换层，将闪存介质当作一个物理磁盘使用。这种方法就是将闪存存储虚拟成一个磁盘，所以在文件接口，访问模式方面都是与当前的磁盘可以保持一致。另外一种模式重新设计底层文件接口。后一种方式的到的效率会更好。但是在超级计算领域由于受各种商业软件和文件系统等限制，采用前者的设计策略更为成熟。

一个好的 FTL 可以提升 Flash 的读写性能，减少写放大，提升 Flash 的生命周期。将 Flash 介质用到超级计算中，面临更多的数据写入问题，为了有效提升 Flash 的读写性能，减小写放到对 Flash 生命周期的影响。特别针对 FTL 表做了优化，以期能够提升 Flash 的读写性能，提升服务器的 I/O 处理能力。

使用基于闪存的固态存储盘构建 RAID 5 阵列，可以有效提升系统的 I/O 带宽和吞吐率，但是由于写更新引发的写放大会在这种架构中更加明显，从而导致 Flash 生命耗尽，甚至有可能连续多块介质损坏，导致数据丢失。将这种架构用到元数据服务器时，由于元数据服务器的量小粒度的写使得这个问题更加严重。针对这个问题提出了一种基于缓存的可重构 RAID 策略。一方面通过重组策略，使得校验码的更新次数明显减少，增加了介质的生命周期；另外，采用缓存策略后写盘和读盘的通道可以并行执行，提升了存储设备的 I/O 能力。

6. 存储功耗模型与优化技术

对于超大规模的计算存储系统，显然一个好的任务分布可以节省大量的存储传输功耗，但同时计算的需求决定了这个过程必须是秒级，即对作业分布的时间提出严格的要求，当前的算法暂时还没有考虑到如果最短时间内来解决一个分布问题。在此介绍一种基于通信连接特性的快速映射策略（Communications-Cluster based Simulated Annealing,CCSA）。根据各 IP 间数据传输的通信特性将节点进行分类，其搜索空间远远小于原来，大大减小了问题的规模。而且在同一个通信类内部搜索的时候，搜索算法由于受通信类的约束，会大大减小了

搜索的范围，加快了算法的执行。基于通信类的模拟退火算法，利用通信类将问题规模迅速缩小，通过通信约束减少搜索空间的方法，采用模拟退火算法实现快速的作业调度策略，实验表明此种策略较传统任务分配策略时间更短，节能效果更好。

7.总结与展望

大数据时代的存储复杂性是一个多样化、多元化的技术问题，存储系统需要满足的应用需求正在变得异常丰富。数据巨大的数量和复杂的数据类型要求海量数据存储系统具有相应等级的扩展能力。数据量的快速增长和数据分级存储的需求，是当前海量数据存储和处理面临的主要技术挑战。

但是正如科技是不断发展的一样，存储技术也在不断的更新中，展望未来，不管是方兴未艾的云存储技术，还是当前引领热点的融合与超融合存储，都拥有无限的发展空间。目前超融合存储已经可以从技术上替代传统磁盘阵列了，至于会给我们带来多大的惊喜，就拭目以待吧。