# MadFS：高性能超算缓存文件系统

## *MadFS: a high performance burst buffer file system*

**陈康**（1976- ），男，清华大学计算机科学与技术系研究员，主要研究方向为分布式系统、存储系统等。

**武永卫**（1974- ），男，清华大学计算机科学与技术系教授，中国计算机学会（CCF）高级会员，主要研究方向为并行和分布式处理、云计算和存储等。

**郑纬民**（1946- ），男，中国工程院院士，清华大学计算机系教授，CCF原理事长，何梁何利基金科学与技术进步奖获得者，中国存储终身成就奖获得者，《大数据》期刊主编。长期从事计算机系统结构、大规模数据存储、高性能计算等领域的科研教学工作。获国家科学技术进步奖一等奖1次，获国家科学技术进步奖二等奖2次，获国家技术发明奖二等奖1次。

中图分类号：TP316.4　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.20210031

**通信作者：** 陈康，chenkang@mail.tsinghua.edu.cn

对于存储系统来说，信息资源的爆炸性增长在I/O支持应用的性能以及数据可用性等方面提出了越来越高的要求。可以预见，人工智能、大数据和图计算等新型计算模式对存储系统的I/O性能更是提出了极致要求。从技术发展趋势上来看，新型的网络传输硬件及使用模式、新型存储硬件都提供了极高的访问带宽和极低的访问时延，这一发展趋势导致现有的存储软件成为性能瓶颈。特别是在提供低时延访问上，存储软件的结构需要进行革新。

传统的分布式文件系统结构按照扩展的方式来看，主要有两个发展思路。一个是先对磁盘进行扩展，之后在扩展的磁盘基础上建立文件系统，提供服务。这个方面的典型是Petal磁盘扩展服务以及Frangipani文件系统。在高性能文件系统中，这个方面的典型是IBM公司的通用并行文件系统（general parallel file system，GPFS）。另外一个发展思路是直接对文件系统进行扩展，由一个或者少数几个节点来保存元数据，记录文件数据的分布情况，其他的节点用来保存数据。这方面的典型是Google文件系统（Google file system）以及衍生的Hadoop分布式文件系统（Hadoop distributed file system，HDFS）。在高性能文件系统中，采用该思路的是大部分高性能计算机标配的Lustre文件系统。传统的高性能文件系统在构造时大部分将磁盘作为数据的存储介质。但是，现有的文件系统不能满足新一代的人工智能、大数据、机器学习等的应用，对于新的存储体系结构、新的网络体系结构带来的高带宽、低时延的性能优势也缺乏考虑。

在当前数据密集型计算普及发展的时代，存储软件的访问性能直接制约了数据密集型计算的性能。下一代的存储系统刚刚开始起步，包括Intel分布式异步对象存储（distributed asynchronous object storage，DAOS）在内的新型存储结构与系统正在形成。为了适应这种趋势，清华大学计算机系的E级计算机系统结构研究团队构建了下一代的分布式存储系统——MadFS，从分布式文件系统软件的架构上进行革新，消除现有存储架构的系统性问题，充分释放硬件的性能，满足下一代应用对数据快速处理的需求。MadFS的设计以性能为第一原则，利用高速远程直接内存访问（remote direct memory access，RDMA）网络和NVMe SSD存储设备，将数据快速分散到存储节点上进行持久化，达到高吞吐、低时延、高性能的特性。

MadFS的系统架构设计遵循了以下3个关键的设计原则。

- 数据块和元数据的全分散存储：传统并行或者分布式文件系统一般使用少量节点管理元数据，导致元数据节点成为整个系统的性能瓶颈。下一代分布式存储系统MadFS将元数据分散到全部节点上，以避免元数据的性能瓶颈，同时数据块也需要分散在全部节点上。
- 建立内核旁路，避免操作系统切换开销：为了提高系统的性能，MadFS采用避免应用程序频繁进入操作系统内核的方式来降低上下文切换的开销。随着I/O设备性能的不断提高，操作系统进出内核切换的开销日益突出，MadFS使用用户态驱动、协议栈等方式直接控制设备。在对应用程序的支持上使用系统调用截获技术，直接在用户态处理应用的I/O请求，避免其进入内核。
- 语言级协程机制与零拷贝序列化：MadFS利用Rust语言内建的异步协程机制、零拷贝序列化技术实现了极低开销的任务切换和远程函数调用。高性能存储系统常采用异步的方式处理I/O请求，这会给编程引入很大的复杂性。而新型编程语

言Rust提供了利用协程处理异步逻辑的语言机制，可以极大地降低异步编程的复杂性，同时保持极低的任务切换开销，保证整体的高性能。

2020年11月19日，由清华大学计算机科学与技术系存储系统研发团队研发的超算缓存文件系统MadFS在鹏城实验室“鹏城云脑Ⅱ”的IO500测试中，分别以7 043.99分和1 129.75分同时获得全球IO500总榜第一名与10节点榜单第一名，这是国内科研机构首次夺得该排行榜榜首。“鹏城云脑Ⅱ”是一台基于华为鲲鹏920架构的高性能计算系统，于2020年10月开始试运行。本次“鹏城云脑Ⅱ”的存储系统基于MadFS，针对“鹏城云脑Ⅱ”的硬件特征，采用了基于Rust的高可扩展并发访问、大粒度数据缓存/旁路访问、数据访问/落盘流水化、零拷贝极速远程过程调用（remote procedure call，RPC）处理技术等创新优化方法。

IO500是高性能计算领域针对存储性能评测的全球排行榜，是高性能计算领域权威的榜单之一。IO500测试包括数据带宽BW（GiB/s）和元数据性能MD（kIOPS）两大部分，各项分数取几何平均后得到总分。在高性能计算领域，不仅CPU算力非常重要，I/O系统的数据传输更是瓶颈。自2017年11月开始，每年IO500榜单会在高性能计算领域的会议——全球超级计算大会（SC）和国际超级计算大会（International Supercomputing Conference）上发布。 □