IOPathTune：面向并行文件系统 I/O 路径的自适应在线参数调优

Md. Hasanur Rashid
mrashid2@uncc.edu 北卡罗来纳大学夏洛特分校 美国北卡罗来纳州夏洛特

Youbiao He
yh54@iastate.edu 爱荷华州立大学 美国艾奥瓦州艾姆斯

Forrest Sheng Bao
forrest.bao@gmail.com 爱荷华州立大学 美国艾奥瓦州艾姆斯

Dong Dai ddai@uncc.edu 北卡罗来纳大学夏洛特分校 美国北卡罗来纳州夏洛特

arXiv:2301.06622v1 [cs.DC] 16 Jan 2023

摘要

研究动机。并行文件系统（PFS）在高性能计算（HPC）平台中发挥着基础性作用，为用户和应用程序提供满足所需速度和容量的数据访问能力。确保并行文件系统能够为高性能计算系统中运行的种类繁多的科学应用提供极致性能至关重要。并行文件系统以其漫长而复杂的 I/O 路径而闻名。具体而言，在并行文件系统中，I/O 请求由计算节点或 I/O 节点通过并行文件系统客户端库发起，随后经由高速网络上的 RPC 机制传送至存储服务器。最后，这些请求会缓存在存储服务器的 I/O 队列中，等待内部 I/O 调度器处理数据。高效且理想的 I/O 路径意味着数据能够从客户端流向存储服务器，而不会在任何环节受阻，这也是并行文件系统实现优化 I/O 性能的关键。要实现最佳 I/O 性能，并行文件系统必须合理设置多个控制 I/O 路径中数据流动方式的参数。尽管默认设置经过了精心挑选，但面对多样且不断变化的工作负载时，它们往往无法实现最佳性能。

为了有效调节并行文件系统的 I/O 路径相关参数，我们关注几个关键特性：自适应、及时和灵活。调节框架应在工作负载发生变化时自适应地调整参数，并对运行时变量（如 I/O 争用）作出响应。由于工作负载具有短时突发的特性，调节决策必须快速下发。可调参数也应能够动态修改。该框架必须灵活地为不同客户端采用不同的调节方式。考虑到性能调节可用的 I/O 、计算和通信资源有限，该框架还应避免对 HPC 环境进行开销高昂的探测或分析。

IOPathTune 的设计。在本研究中，我们提出了 IOPathTune，这是一种新的调优框架，旨在从客户端自适应地在线调优 Lustre [1] I/O 路径。经过对 Lustre I/O 路径的深入考察，我们重点关注两个参数：max_pages_per_rpc（单个 RPC 中的最大页面数）和 max_rpcs_in_flight（正在处理中的 RPC 最大数量）。这两个参数均属于 Lustre 门户 RPC（PtlRPC）子系统。它们可以动态调节、立即生效，并控制每个 Lustre 客户端的 I/O 流。这些重要特性与我们的目标高度契合。借助这两个参数，RPC 会以结构化的方式形成并传输，从而帮助我们避免对工作负载进行特征分析。

与这两个参数相关的统计数据也很容易获取，无需额外探测系统。

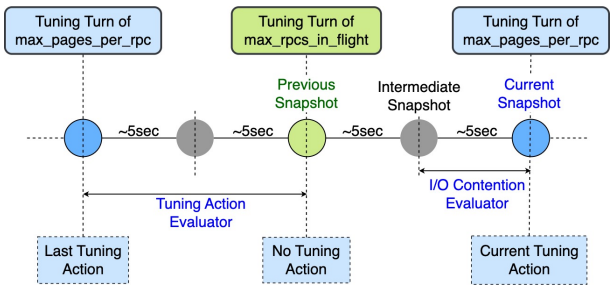


图 1：启发式调节方法

IOPathTune 的一项关键设计选择是不探测存储服务器或其他计算节点。我们将其设计为仅依赖并行文件系统客户端库收集的统计数据。我们的调节算法每十秒执行一次调节，并在每一轮中交替选择两个参数之一进行调节，如图 1 所示。每次调节时，参数值要么乘以二，要么除以二，这与 TCP拥塞控制的主要方法十分相似。

具体而言，在调优过程中，我们会沿 I/O 路径计算不同的指标。我们得出的指标包括：脏页缓存中的数据量、页面缓存的速度、客户端生成 RPC 的速度，以及客户端在最近一个观测周期内传输 RPC 时达到的速度。通过比较这些指标，如果我们发现上一次调优操作提升了带宽，就会重复上一次调优操作；否则，就采取与上一次调优操作相反的措施。我们还会根据这些指标判断是否正在出现 I/O 争用。出现 I/O 争用时，我们会采取保守的做法：将问题归咎于上一次调优操作，并将其撤销。

评估。为了测试我们的框架 IOPathTune 的有效性，我们利用 CloudLab [2] 开展评估。我们的高性能计算集群由一台 MGS/MDS 机器、四台 OSS 机器和五台客户端机器组成（运行 CAPES 时额外使用一台机器）。每台机器均配有两颗主频为 2.20 GHz 的 10 核英特尔至强银牌 4114 CPU、192GB ECC DDR4-2666内存，以及一块 480 GB 6G SATA 接口的英特尔 DC S3500 固态硬盘。每台 OSS 配有两个 200GB 对象存储目标

表 1：独立工作负载执行

工作负载 Name	8KB I/O 请求	1MB I/O 请求	16MB I/O请求
	I/O 吞吐量提升*(%)*	*I/O* 吞吐量提升*(%)*	*I/O* 吞吐量提升*(%)*
随机写入	7.82	**22.97**	10.93
五流随机写入	**64.46**	**231.98**	**43.44**
随机读写	-7.46	5.57	-2.91
顺序写入	-4.39	-0.73	7.56
Fivestream 顺序写入	-7.29	3.75	-7.59
顺序读写	4.03	**113.19**	**72.6**
整文件写入			**86.45**
整文件读写			**96.58**

表 2：多客户端执行

工作负载 Name	客户端 Name	默认执行 *I/O* 带宽（*MB/s*）	CAPES 执行 *I/O* 带宽（*MB/s*）	启发式执行 *I/O* 带宽（*MB/s*）
五流随机写入	节点1	385.4	237	**2627.9**
随机写入	节点2	95.2	101.4	**206.3**
随机读写	节点3	2127.6	**4209.3**	3199.8
顺序读写	节点4	639.2	630.8	**1134.6**
整文件读写	节点5	1682.3	784.3	**4135**
多客户端带宽总计（MB/s）		4929.7	5962.8	**11303.6**

并挂载在固态硬盘上。网络由双端口 Intel X520-DA2 10Gb 网络接口卡提供支持。我们将 CentOS 7 设置为操作系统，并将 Lustre 2.12.8 作为高性能计算集群的主文件系统。我们测试了 20 种不同的 Filebench [4]工作负载，其 I/O 模式（随机或顺序）、I/O 操作（读或写）、I/O 请求大小、进程数和线程数等各不相同。我们进行了三项不同的评估，检查 IOPathTune 如何应对不同场景。

第一项测试是在单个客户端上独立执行工作负载。我们观察到，在所有工作负载中，性能要么与默认配置持平或略有下降，要么有所提升。表 1 展示了其中一些显著提升：Filebench fivestreamwriternd、整文件读写和顺序读写工作负载分别提升了 231%、113% 和 96%。第二类测试是动态测试，即每运行 300 秒后便持续更换工作负载。我们在单次运行中更换了六次工作负载，并进行了五次运行，每次运行均采用六种不同的工作负载组合。我们观察到，IOPathTune 持续带来了性能提升，这与独立性能评估的结果相似，表明我们的算法能够快速适应并收敛到更优配置。

我们的最后一类测试是在同一时间从不同客户端运行不同的工作负载。我们将 IOPathTune 的性能与默认配置和 CAPES [3]执行进行了比较。如表 2 所示，与默认执行相比，IOPathTune 将集群总带宽提升了 129.30%；与 CAPES 执行相比，提升了 89.57%。研究结果证明，我们提出的 IOPathTune 框架能够独立调优不同客户端的参数，以获得更好的性能。

结论。我们的研究表明，无需对工作负载进行特征分析，无需进行昂贵的性能分析，也无需与其他机器通信，就能实现自适应在线参数调优。我们希望这一方法能引起更多关注，推动系统研究领域探索成本低廉却非常有效的解决方案。未来，我们希望扩展算法，按照这一启发式方法调优更多参数。我们也希望在真实的 HPC 设施中进行测试，观察该方案在 I/O 性能方面能带来多大提升。

致谢。本研究部分由美国国家科学基金会项目 CNS-1817089 和 CNS-2008265 资助。

CCS 概念

• **计算方法** → 共享内存算法。

参考文献

[1] Peter Braam。2019。《Lustre 存储架构》。*arXiv* 预印本 *arXiv:1903.01955*（2019）。[2] Dmitry Duplyakin、Robert Ricci、Aleksander Maricq、Gary Wong、Jonathon Duerig、Eric Eide、Leigh Stoller、Mike Hibler、David Johnson、Kirk Webb 等。2019。《{CloudLab} 的设计与运行》。载于 *2019* 年 *USENIX* 年度技术大会（*USENIX ATC 19*）。1–14。[3] Yan Li、Kenneth Chang、Oceane Bel、Ethan L Miller 和 Darrell DE Long。2017。CAPES：使用基于神经网络的深度强化学习进行无监督存储性能调优。载于 国际高性能计算、网络、存储与分析大会论文集。1–14。[4] Vasily Tarasov、Erez Zadok 和 Spencer Shepler。2016。Filebench：灵活的文件系统基准测试框架。*USENIX*；*login* 41，1（2016），6–12。