



超大规模高性能计算机中的突发I/O

从刻画到优化

中国空气动力研究与发展中心

喻杰

目录

Overview

- 01 背景介绍
- 02 CN级突发I/O
- 03 ION级突发I/O
- 04 OST级突发I/O



■ 高性能计算机的运算性能持续增长

- Top 1高性能计算机的性能已超越E级

■ Amdahl定律

- 一个系统的整体加速比，受限于它最慢的那个部件
- 实际机器建设中，对存储系统的资金投入往往偏少
- 当超大规模高性能计算机的性能越来越高，读写海量数据成为了性能瓶颈
- 为了释放高性能计算机的全部潜能，I/O优化的需求越来越强烈

A supercomputer is a device for turning **compute-bound** problems into **I/O-bound** problems.

-- Ken Batcher



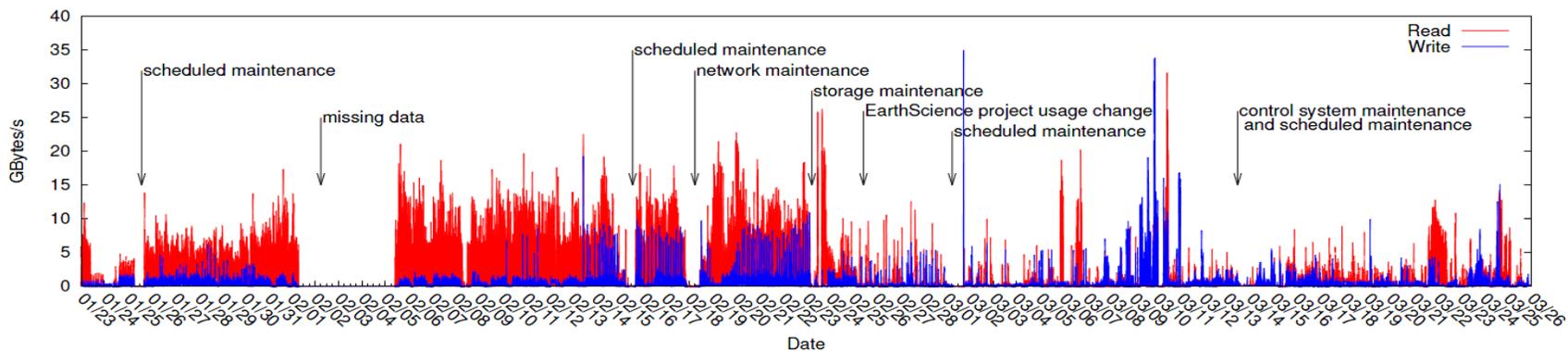


■ 理解高性能计算机的I/O行为

- 改造现实之前，先要理解现实
- 通过分析应用的I/O访问模式，并总结系统的I/O流动规律，可发现潜在的性能瓶颈，并进一步通过多种手段开展优化

■ 突发I/O

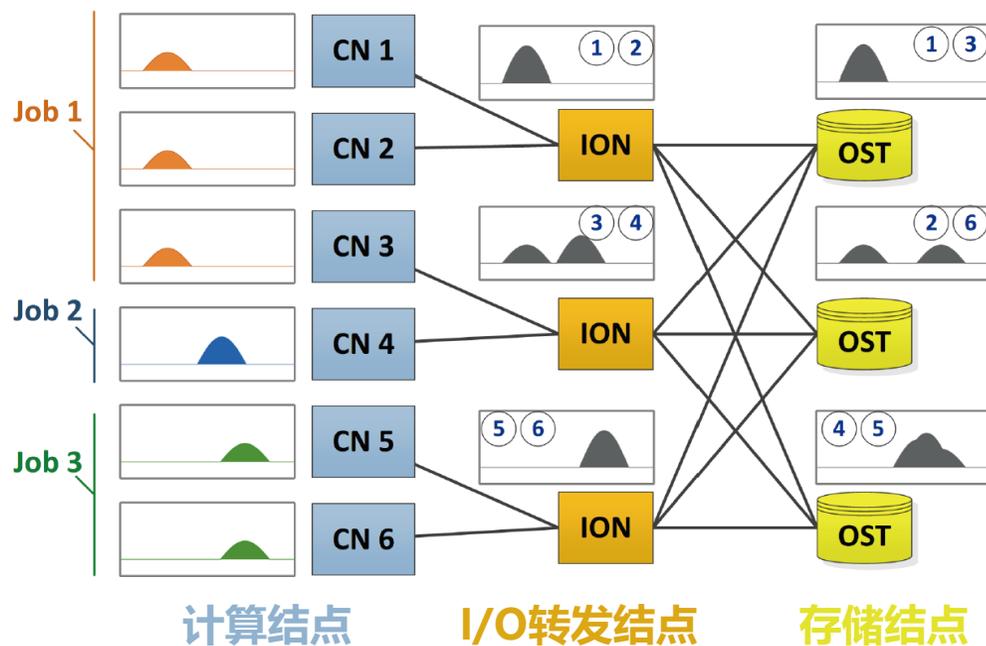
- 一种受学术界与工业界广泛认同的I/O模式
- 系统持续运行过程中，仅少部分时间会有大量I/O需求，I/O流量总是突然地、集中地到来
- 常见于应用从计算阶段转入数据访问阶段





■ 突发I/O在高性能计算机中的流动

- 超大规模高性能计算机往往采用了I/O转发层来缓解海量并发I/O压力
- 系统中往往包括数千个计算节点CN，数百个I/O转发节点ION，数百个存储节点OST
- 不同作业在其运行的CN上产生突发I/O，然后在对应的ION上汇集，最后流动到数据所在的OST上



作业产生的突发I/O会在流动过程中不断合并、分裂，在不同层次上形成不同类型的突发I/O

研究突发I/O的流动规律是优化突发I/O的**第一步**

目录

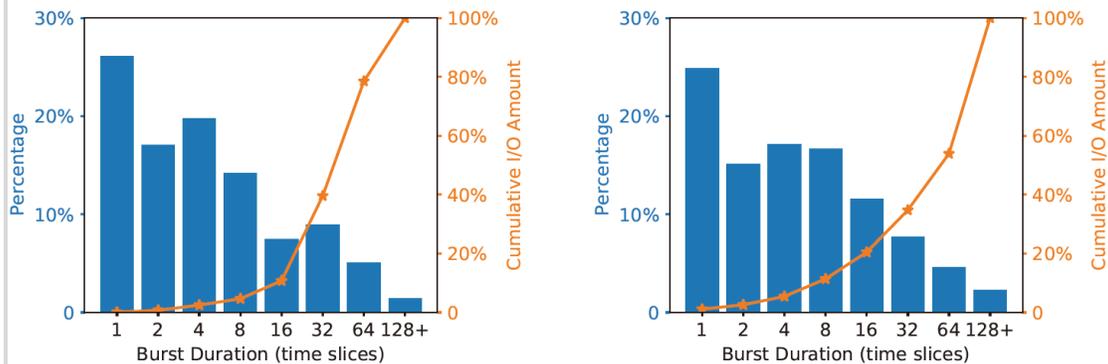
Overview

- 01 背景介绍
- 02 CN级突发I/O
- 03 ION级突发I/O
- 04 OST级突发I/O



■ 突发I/O的持续时间与间隔时间

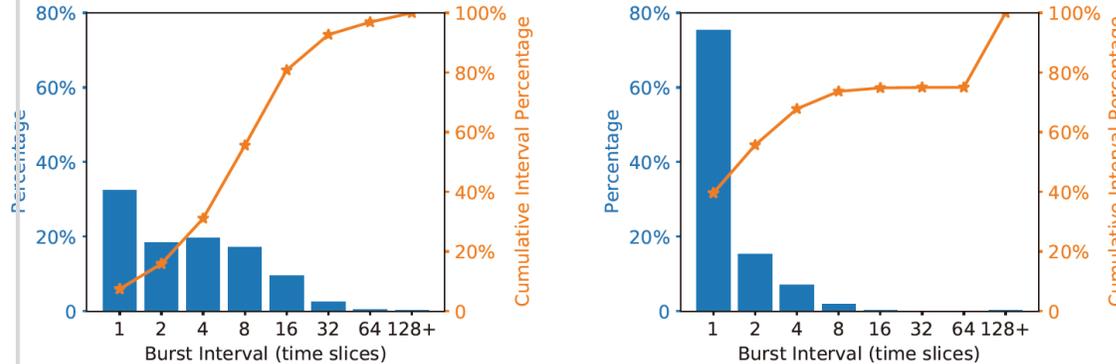
持续时间频数分布



(a) CS-19-F

(b) Tianhe-1A

间隔时间频数分布



(a) CS-19-F

(b) Tianhe-1A

- 90%的I/O量出现在约一半时间内 (39.2%与52.7%)
- 两个集群的持续时间分布类似
- 大部分突发I/O的持续时间较短
- 长突发I/O虽然数量少, 但它们的I/O量更大, 贡献了大部分I/O量 (橙线为累积I/O量)

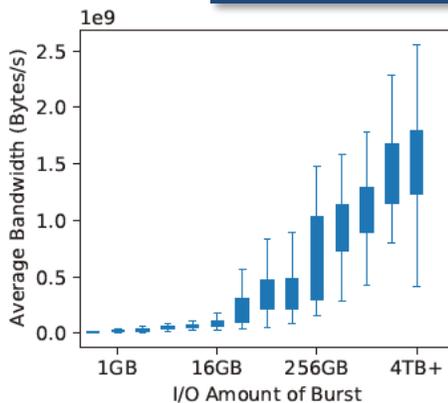
- CS-19-F上突发I/O的间隔较长, Tianhe-1A上突发I/O的间隔较短
- CS-19-F上突发I/O的间隔主要集中在2~32个时间片
- Tianhe-1A上突发I/O主要为间隔1个时间片的密集I/O和间隔128个时间片以上的稀疏I/O
- Tianhe-1A运行超14年, I/O负载较轻

2. CN级突发I/O

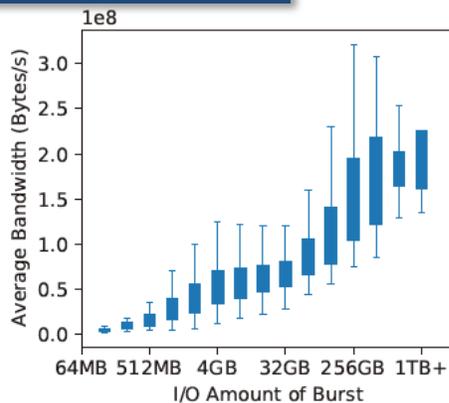


■ 突发I/O的带宽与I/O量的关系

I/O量与平均I/O带宽



(a) CS-19-F



(b) Tianhe-1A

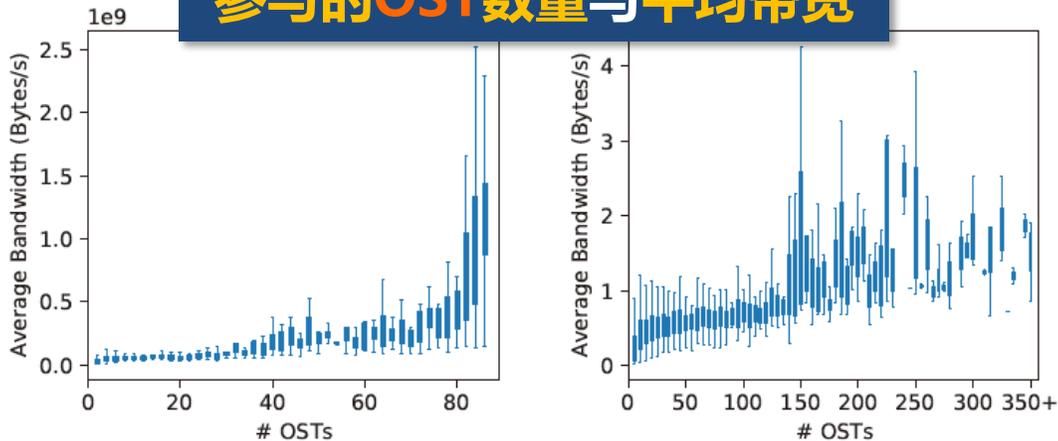
- 图中为所有突发I/O的统计箱线图
- 两个集群的趋势类似
- 突发I/O访问的数据量越大，其能获得I/O带宽也越高

一般而言，突发I/O的数据量越大，其I/O粒度也会越大。证明Lustre文件系统更擅长处理大I/O。



■ 突发I/O的带宽与OST数量和CN数量的关系

参与的OST数量与平均带宽

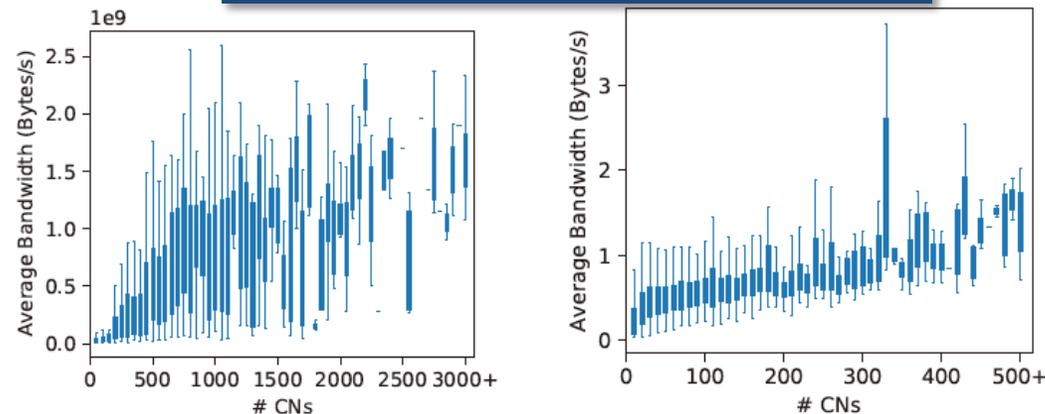


(a) CS-19-F

(b) Tianhe-1A

- 突发I/O中，参与的OST数量越多，其I/O带宽越高

参与的CN数量与平均带宽



(a) CS-19-F

(b) Tianhe-1A

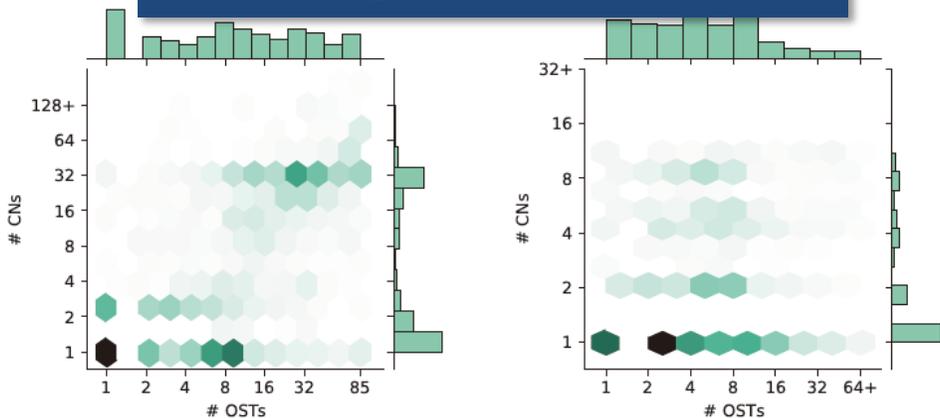
- 突发I/O中，参与的CN数量越多，其I/O带宽越高
- 当CN数量过多时，竞争导致带宽起伏不定

为了充分发挥并行文件系统的性能，需要大量的客户端并发访问大量的OST。但过多的客户端将会导致剧烈的I/O竞争，造成性能降级。



■ 作业使用的CN数量和OST数量

CN数量与OST数量频数分布



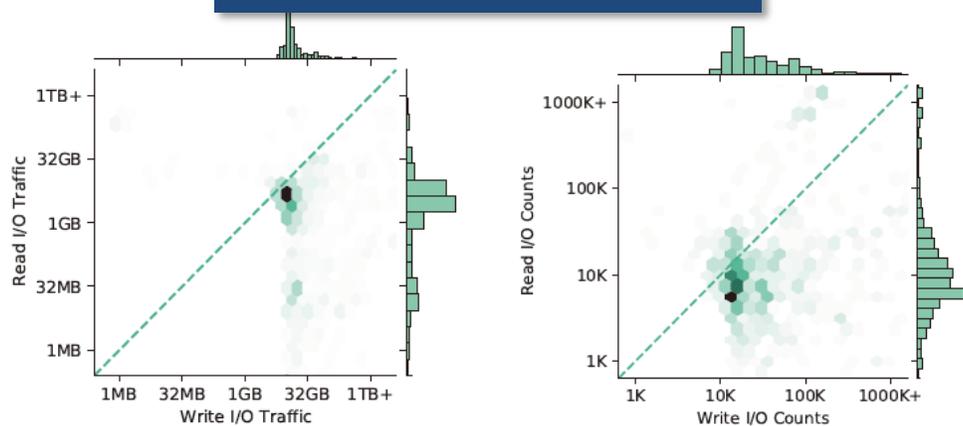
- 图中颜色越深代表作业数量越多
- 最常见的作业I/O模式是一个CN访问一个OST
- CS-19-F中绝大部分作业使用的CN数量少于32个
- Tianhe-1A中绝大部分作业使用的CN数量少于8个
- 两集群中大多数作业访问的OST数量少于16个

虽然只有当大量CN访问大量OST时，才能发挥文件系统的最大潜能，但在实际生产系统中，作业I/O活动的并发度仍然受限。



■ 作业的读写分布

CS-19-F读写分布

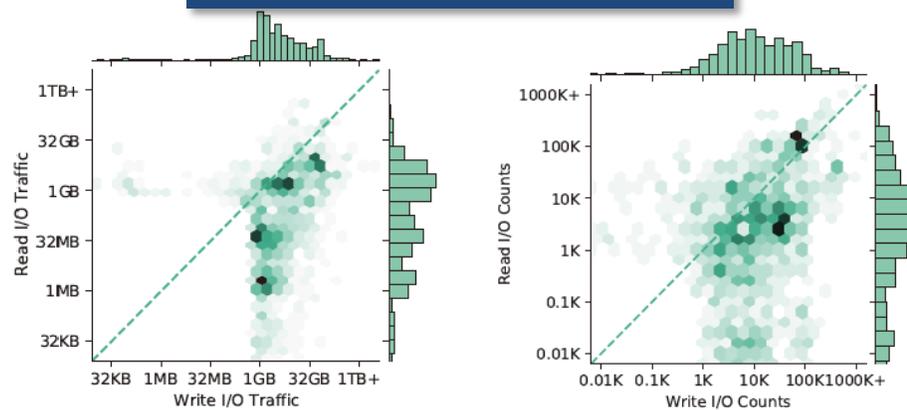


(a) I/O Amount

(b) I/O Count

- 偏左上为重读，偏右下为重写

Tianhe-1A读写分布



(a) I/O Amount

(b) I/O Count

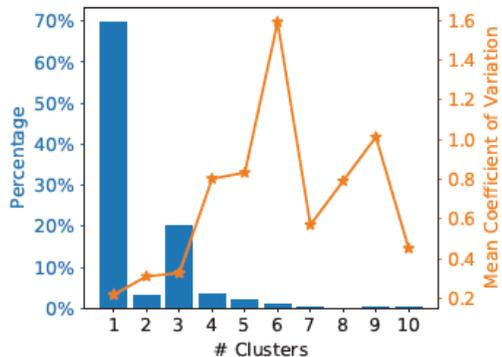
- 颜色越深代表作业数量越多

尽管两个集群的应用领域不同，它们运行的作业都是写密集的。其中，规模更大的集群（CS-19-F）的写数据量更大。

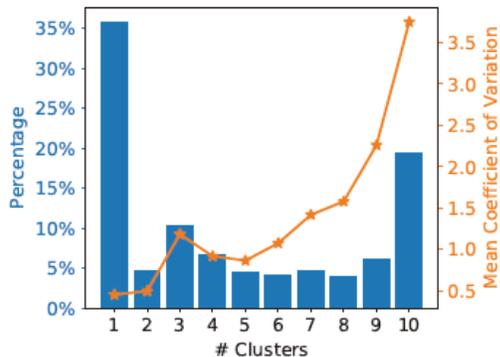


■ 作业写I/O的相似性与周期性

写I/O聚类结果



(a) CS-19-F



(b) Tianhe-1A

- 根据突发I/O的I/O量、I/O次数、使用的CN数、访问的OST数聚类
- 读I/O聚类效果较差，写I/O聚类效果很好
- 柱状图表示聚类的类数量，大部分作业的写I/O都可聚成一类，证明它们的写I/O的形态很相似
- 折线图表示作业写I/O间隔的波动性，值越低表明间隔的波动越小，证明它们的写I/O具有很强的周期性

作业的写突发I/O具有较强的相似性与周期性，主要由作业写检查点的行为造成。作业的写突发I/O是优化作业性能的关键。

目录

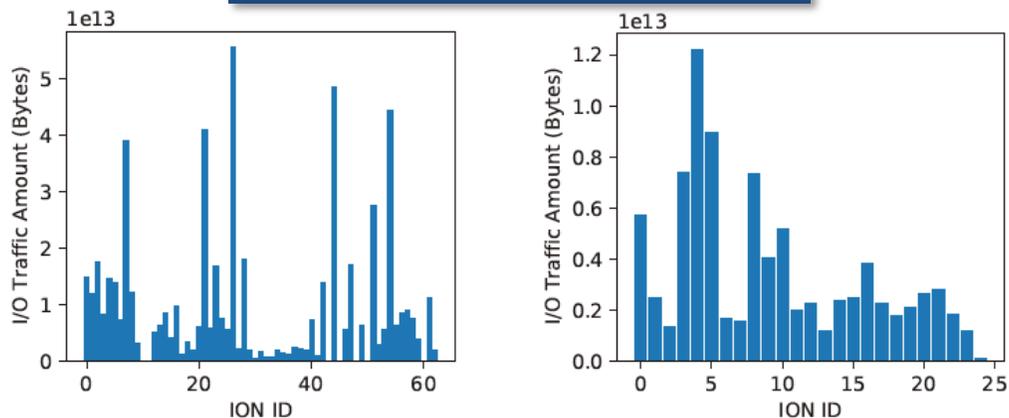
Overview

- 01 背景介绍
- 02 CN级突发I/O
- 03 ION级突发I/O**
- 04 OST级突发I/O



■ 突发I/O在I/O转发层的**拥塞**

ION上的I/O总量分布

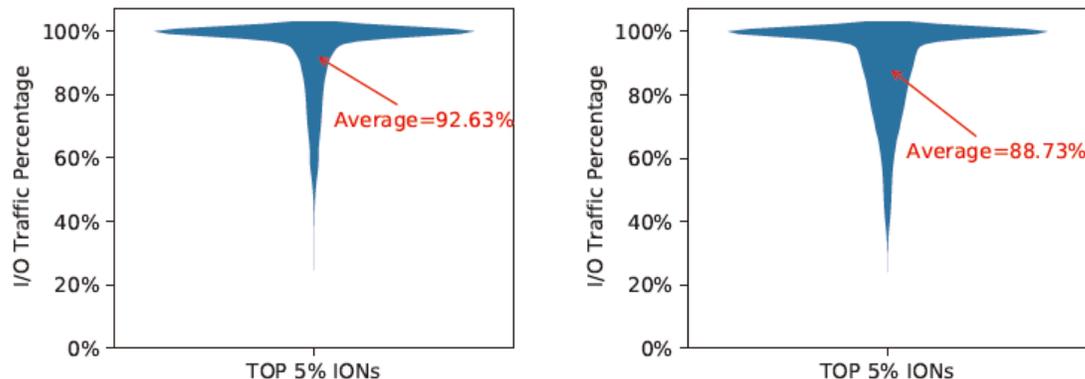


(a) CS-19-F

(b) Tianhe-1A

- 当前主流I/O转发架构为，一个ION服务一组相邻的CN
- ION间的I/O**负载严重不均**

最忙的5%ION的流量比例



(a) CS-19-F

(b) Tianhe-1A

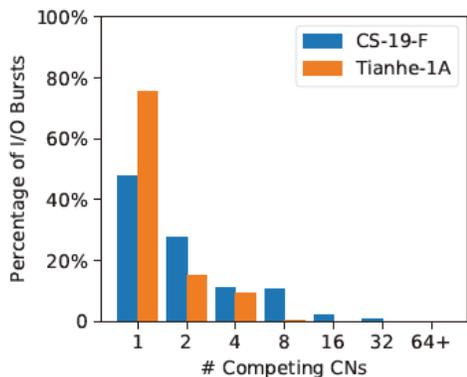
- 每个时间片上选出最忙的5%的ION，分析它们承担的I/O量比例
- 任意时刻，**5%的ION约承担了90%的I/O量**

I/O转发层的引入，使一个ION替一组数量有限的CN访问OST，有效缓解了海量CN同时并发访问OST的问题。但是这种**ION被独占使用**的方式，也造成了ION上负载不均。

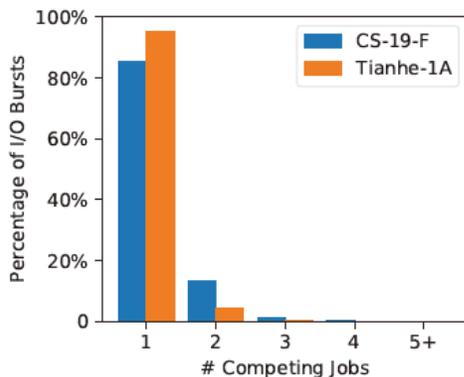


■ 写I/O拥塞的原因

突发I/O中相互竞争的CN数与作业数



(a) Competing CNs



(b) Competing Jobs

- 突发I/O中，相互竞争的CN数量分布在1~32个
- 相互竞争的作业数量分布在1~4个，其中绝大部分只有1个作业
- 即：ION上的突发I/O中，大部分时候只有同属一个作业的多个CN在互相竞争

原因分析：

1. 作业的规模相对较小，绝大部分作业的规模小于64 CN
2. Slurm为了减少通信距离，优先为作业分配相邻的CN
3. 当前I/O转发架构下，相邻的CN一般都使用同一个ION
4. 导致作业只能利用少量的ION进行I/O
5. 作业I/O具有突发性，同一时刻只有少量作业在I/O，造成ION负载不均

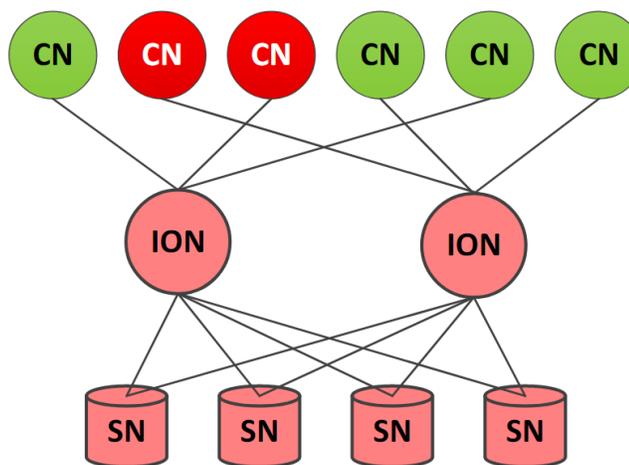
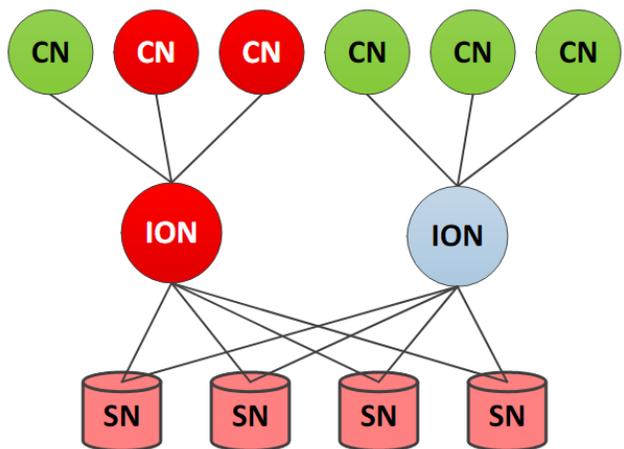
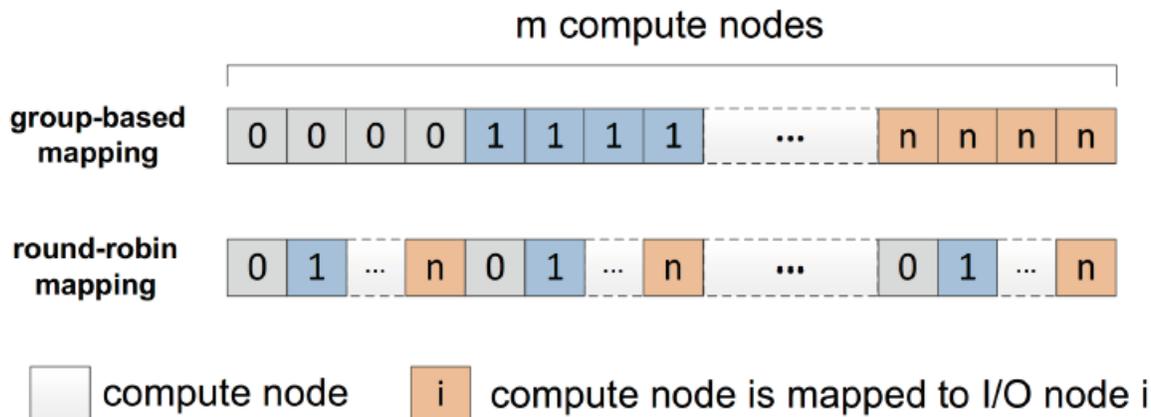
3. ION级突发I/O



■ 分散ION上的突发I/O

轮询I/O转发架构:

- 将相邻的CN映射到不同的ION
- 作业的CN能够使用数量更多、分布范围更广的ION
- 将拥塞的I/O流量分摊到更多的ION上

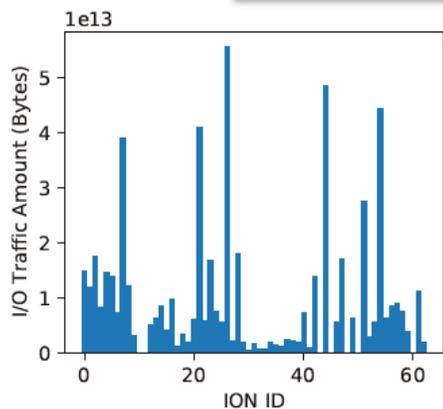


3. ION级突发I/O

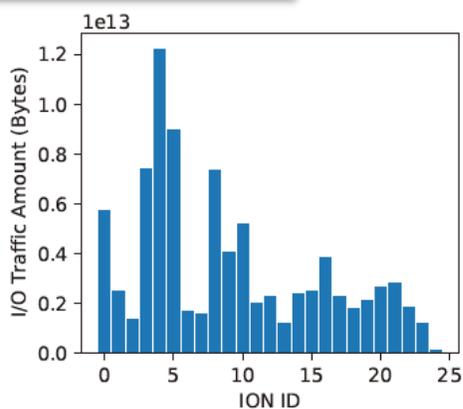


■ 轮询I/O转发架构**优化效果**

优化前：ION负载不均

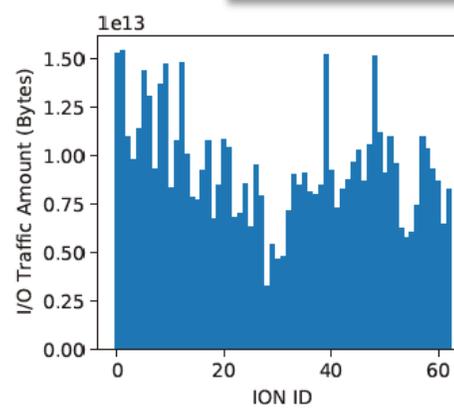


(a) CS-19-F

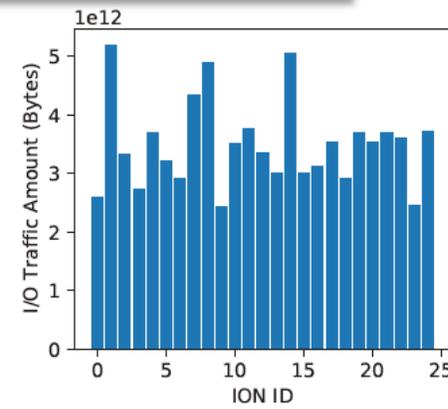


(b) Tianhe-1A

优化后：ION负载较均衡



(a) CS-19-F



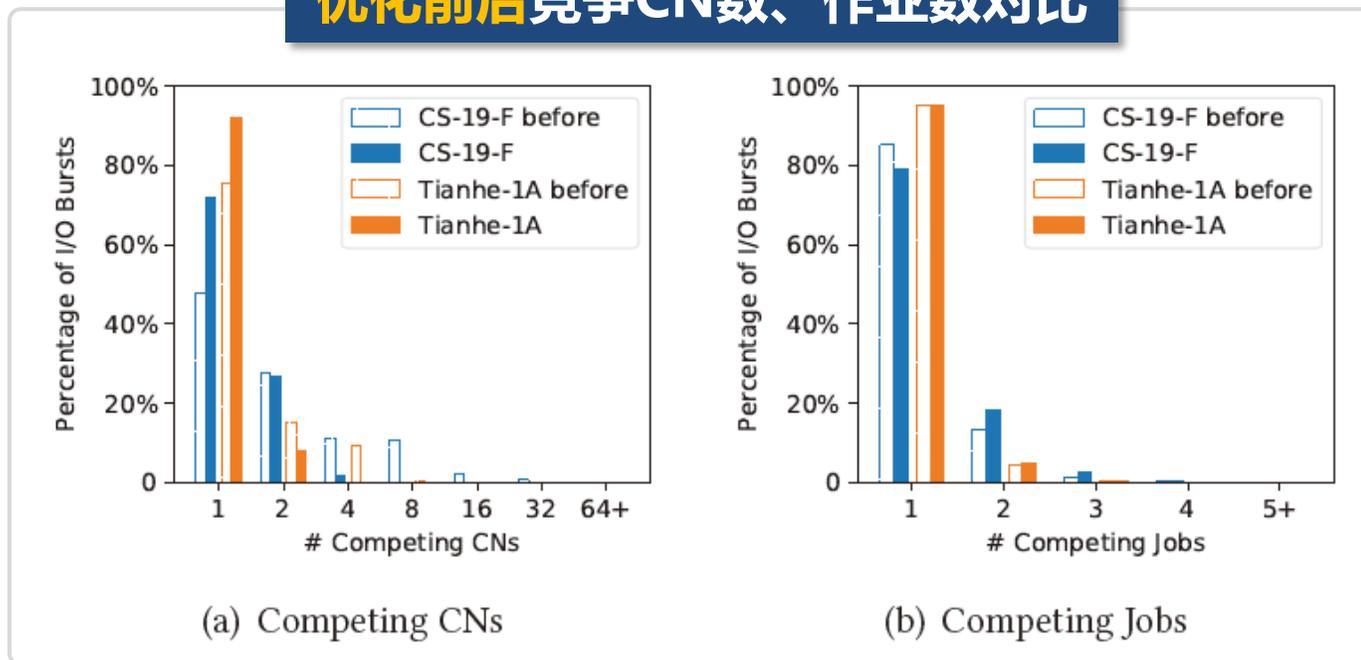
(b) Tianhe-1A

已经在本单位、天津超算、长沙超算、核动力研究院的集群中部署实用，有效改善了ION上的负载均衡状况，提升了I/O转发性能。



■ 轮询I/O转发架构**优化效果**

优化前后竞争CN数、作业数对比

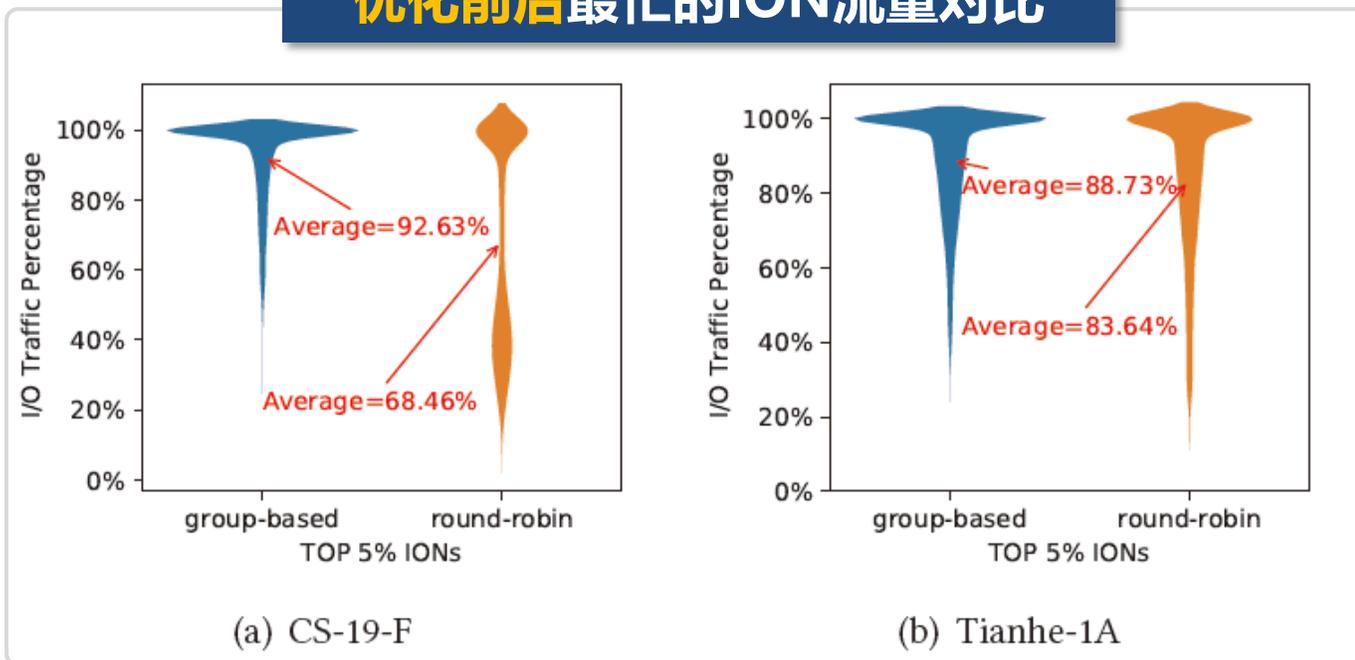


ION上的突发I/O中，竞争作业数**不变**，竞争CN数**变少**，说明本方法成功地缓解了相同作业在ION上的竞争。



■ 轮询I/O转发架构优化效果

优化前后最忙的ION流量对比

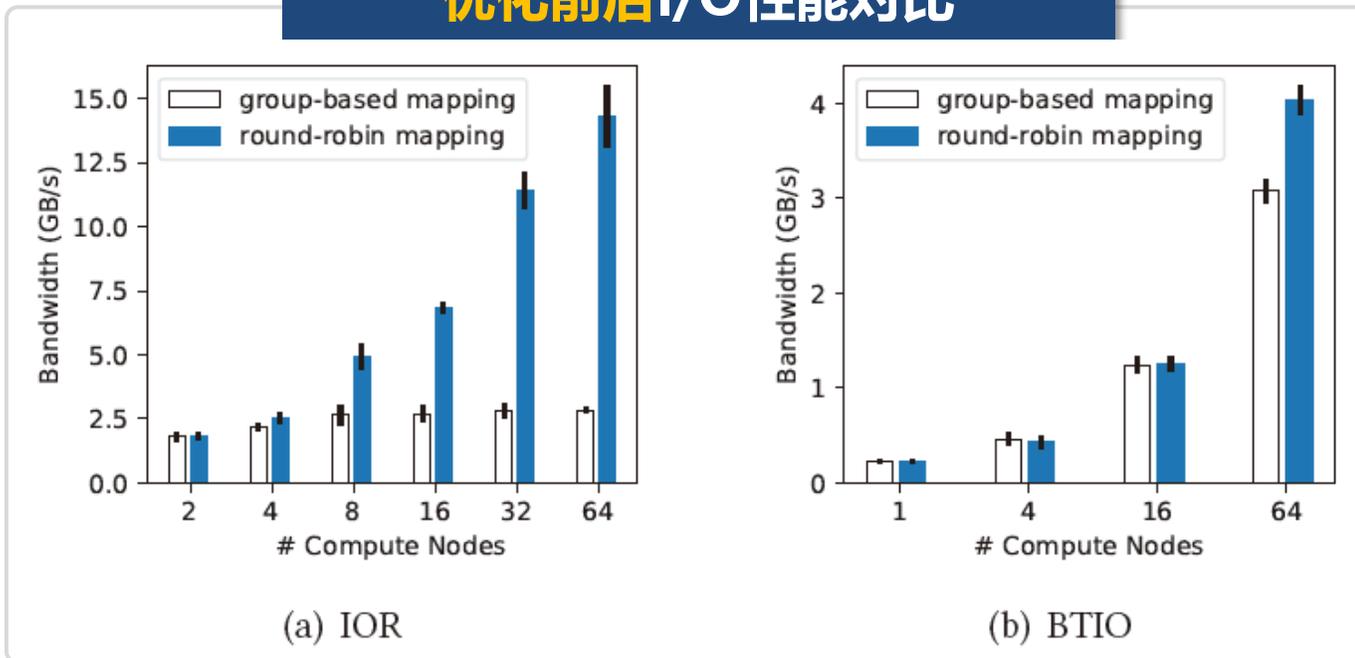


最忙的5%的ION所承担的I/O量大幅降低，避免成为性能热点。



■ 轮询I/O转发架构优化效果

优化前后I/O性能对比



虽然作业的CN数相同，但优化后作业使用的ION数**大幅增加**，作业从Lustre中访问数据的客户端数量增加，因此性能提升。

目录

Overview

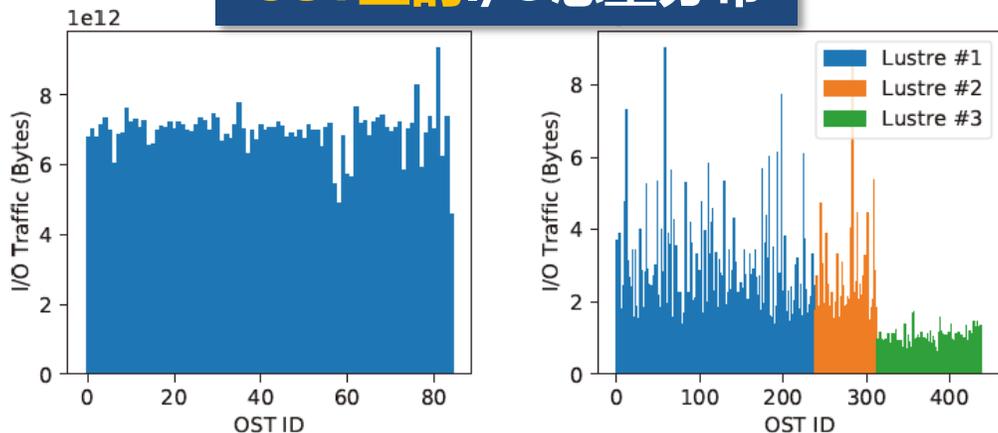
- 01 背景介绍
- 02 CN级突发I/O
- 03 ION级突发I/O
- 04 **OST级突发I/O**

4. OST级突发I/O



■ 突发I/O在存储系统层的**拥塞**

OST上的I/O总量分布

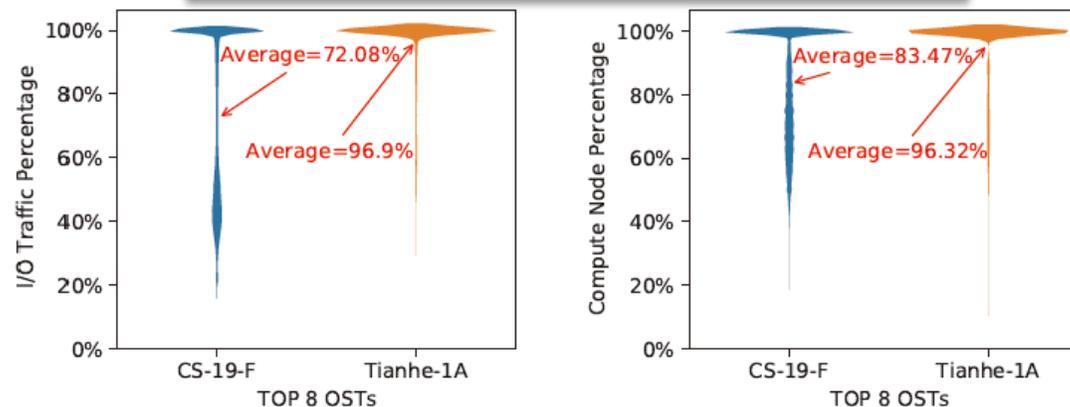


(a) CS-19-F

(b) Tianhe-1A

- 各OST上的I/O问题分布较为均衡
- Lustre随机选择OST存储文件以确保各OST用量均衡

最忙的8个OST的瞬时负载情况



(a) Proportion of processed I/O traffic.

(b) Proportion of served CNs.

- 每个时间片上选出最忙的8个OST，分析它们负载
- 任意时刻，最忙的OST约**承担了70%以上的I/O量**，**服务了80%以上的CN**

虽然Lustre具有负载平衡机制，但OST的瞬时负载仍然**极不平衡**，严重制约了Lustre的并行性能。



■ 文件条带 (File Striping) 策略

- 文件切分成数据条带，分散存储在多个OST上
- 充分发挥多个OST的并行I/O性能
- 例如：1GB文件，切分成1MB条带，以轮询方式存储在多个OST上

■ 生产系统中，大部分并不开启文件条带

- Cori, CS-19-F, Livermore Computing Center, ORLC, 天河大系统



文件不条带化存储



文件条带化存储



■ 文件条带策略的考量因素

稳定性方面

- 若开启文件条带，CN访问数据时需要与更多的OST交互，导致OST同时服务的CN数量增加，导致OST**运行不稳定**；
- 如果一个OST故障，其存储的数据条带无法访问，会影响更多文件的可用性，增加了数据的**安全风险**。

性能方面

- 当作业访问**单个大文件**时，文件条带可提升I/O效率；
- 但当作业同时访问**多个小文件**时，即使不用文件条带，作业仍然能够调动多个OST的并发访问能力；
- 此时开启文件条带，反而会**增加I/O竞争**。

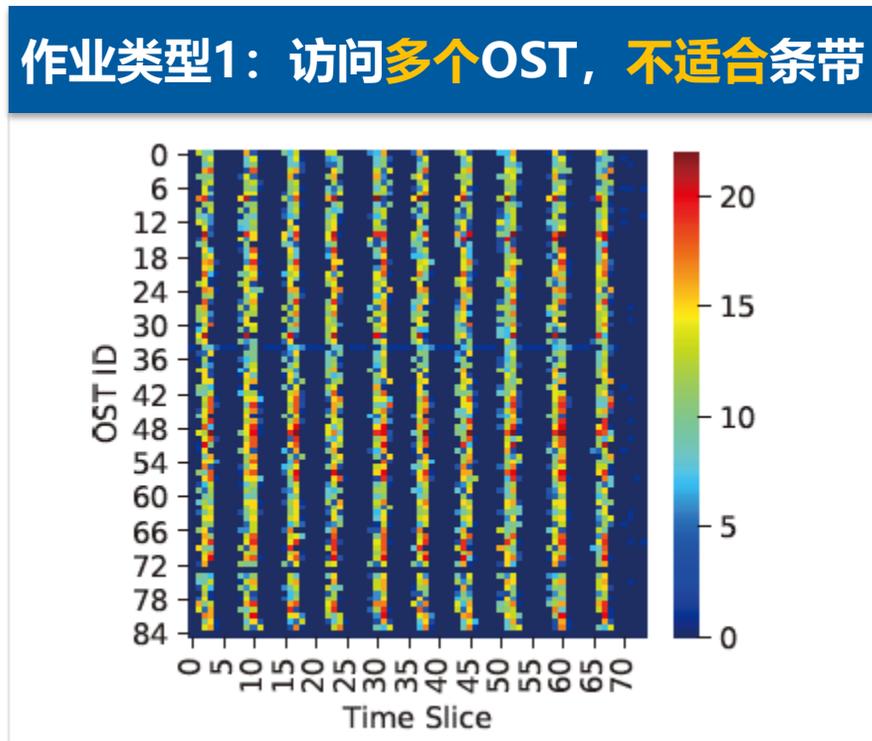
文件条带策略需要考量多个因素，它可以提升访问单个文件的并行I/O性能，也可能降低访问多个文件的性能，并影响存储系统的**运行稳定性**。

4. OST级突发I/O

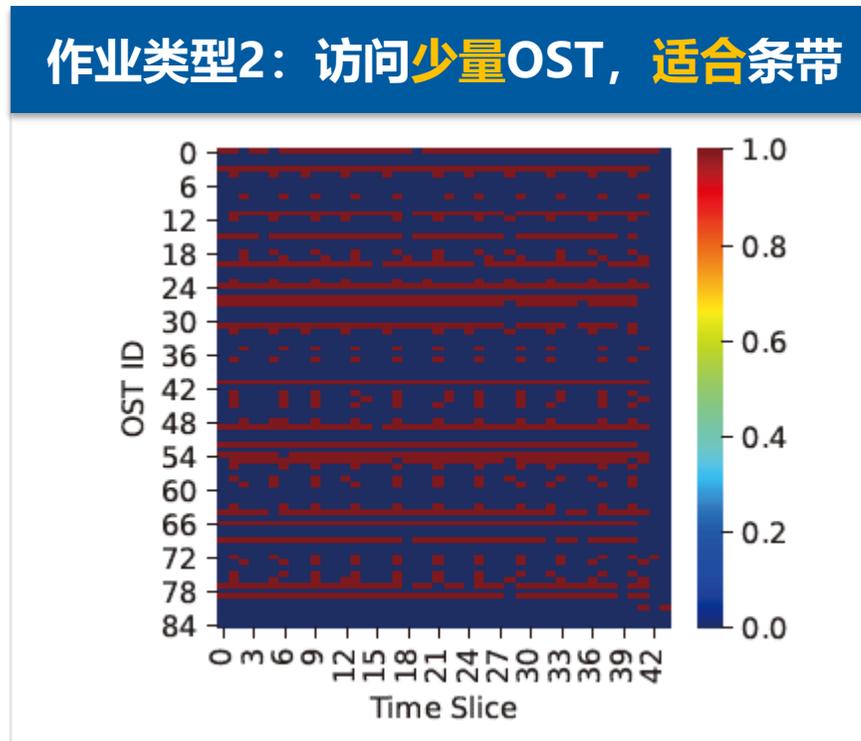


■ 作业感知的文件条带策略

作业类型1：访问多个OST，不适合条带



作业类型2：访问少量OST，适合条带



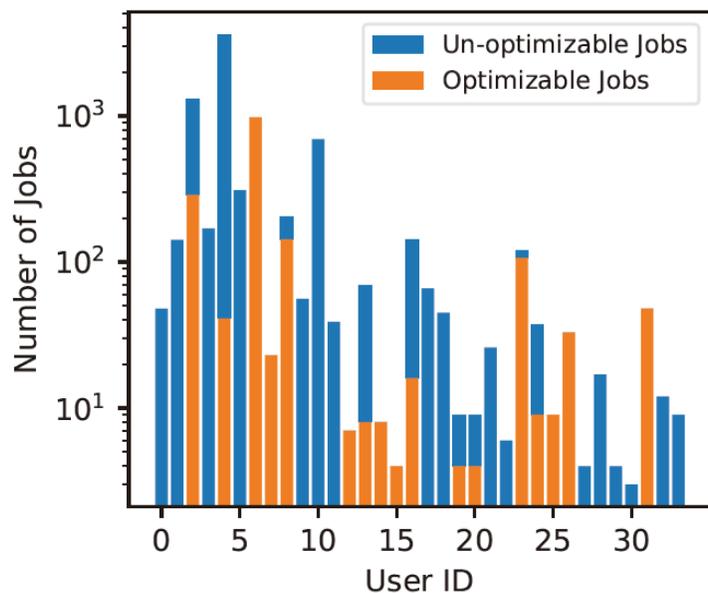
通过识别作业I/O类型，针对性地为作业选择文件条带策略。



■ 作业感知的文件条带策略 —— 静态条带化

- 选取部分适合条带化的作业，在不影响数据安全性的前提下，优化其I/O性能
- 若作业90%的I/O量集中在少于10%的OST上，则认为**适合条带化**

CS-19-F上**适合条带化**的作业比例



静态条带化:

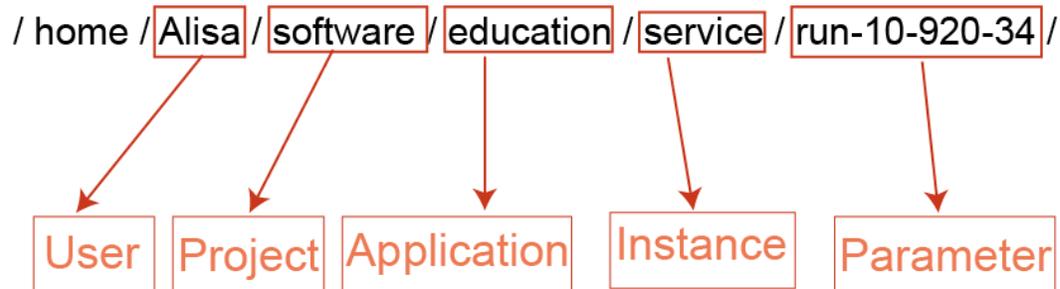
- 23.5%的用户的所有作业均适合条带化
- 占总作业数量的8.4%
- 对于这部分用户，只需**修改其目录的条带设置**，即可优化其I/O性能。



■ 作业感知的文件条带策略 —— 动态条带化

- 对于部分作业适于条带化的用户，我们利用作业的属性信息预测其I/O模式
- 属性信息包括：作业名、用户名、提交时间、作业规模、**作业提交路径**等
- 在作业提交时收集相关信息进行预测，预测**准确率达87.3%**
- 对于预测适合条带化的作业，调整其相关目录的条带设置，并在作业结束后恢复

序号	作业路径
1	/home/alisa/bob1/645pv/RESU/20210112-0814-6ms,fly,1step,change_mix_to_0.063/postprocessing/temperature-3
2	/home/carol/david2/dual-mode/xyzma1-2d/new-src+wing/64g-tow-ke/warm-xiy200-a8e1.51-wall-450-sct0.7
3	/home/eva/c34-YJ1.3-F06.09-RI2.30-K43-sedr-Ma0.8-3.3-2.6mtn/u12-YJ1.3-F06.09-RI5.25-K45-sedr-Ma1-3.3-2.8mtn
4	/home/franklin/gomez/deltawing_sharpedge/ggy_des_thirdcomputation1/topviz/022650
5	/home/hellen/cfdtest/0725CoarseGrid_gbs_testLimiter_1/cfd3



作业路径中可蕴含丰富的作业信息，如子用户名、工程名、数据集名、部分运行参数等，可**更丰富地表征**作业特征。

--IPDPS



敬请指正

中国空气动力研究与发展中心

■ 报告人： 喻 杰