10^{500} s

There is RJGG behind MadFS

王润基

2021.07.17 @TUNIGHT

又见面了

王润基

@wangrunji0408

@玩电脑的猫

- tuna 粉丝团成员
- 贵系五字班老年人
- 高性能所硕士二年级，导师陈康
- Rust 忠实教徒，用它写过 OS
- 热爱第八第九艺术

前情提要

IO500 SC20 List Customize

# ↑	INFORMATION				IO500				
	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS	255	18,360	7,043.99	1,475.75	33,622.19
2	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57
3	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625	938.95	174.74	5,045.33

前情提要

清華大学 Tsinghua University | 新闻 NEWS

首页 头条新闻 综合新闻 专题新闻 要闻聚焦 媒体清华 图说清华 视频空间 清华

首页 – 要闻聚焦 – 学术科研 – 内容

计算机系存储系统研发团队夺得世界IO500排行榜榜首

清华新闻网11月23日电 11月19日，由清华大学计算机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7043.99分和1129.75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。MadFS核心研发人员包括清华大学计算机系的陈康副研究员、武永卫教授、陈文光教授、郑纬民院士等教师和王润基、毛晗扬、刘艺源、陈晟祺等研究生。

半年后

首页 > 新闻动态 > 正文

计算机系存储系统研发团队再次刷新超算存储500强（IO500）排行榜世界纪录

2021.07.07 11:52

7月1日，超算存储500强（IO500）组织发布了新一期排行榜，由计算机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，连续保持了总榜和10节点榜单两个冠军位置，较去年创造的7043.99记录下，今年提高了五倍至36850.37，并与亚军的性能差距拉大近二十倍。同时在10节点榜单上获得2595.89分的好成绩，保持了第一位置。为国内系统中首次在两个榜单上蝉联冠军。

IO500 ISC21 List

ze Download

# ↑	INFORMATION							IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	512	36,864	36,850.37	3,421.62	396,872.82
2	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57

提纲

1. 去年 IO500 回顾及后续进展
2. MadFS 针对 IO500 的新优化
3. 本次打榜的故事
4. MadFS 的未来发展计划

Part I

去年 IO500 回顾及后续进展

IO500 是什么？

IO500与鹏城云脑.pptx　　使用“Keynote 讲演”打开

IO500：世界IO性能排行榜

Register　Log In

Virtual Institute for I/O

Search

Recent Changes　Sitemap

You are here: Virtual Institute for I/O » IO500 » Lists » **2019-11**

Ranked List | 10 Node Challenge | Full List | Historic data | Student Cluster Co…

2019-11

This is the official list from Supercomputing 2019. The list shows the best result for a given combination of system/institution/filesystem.

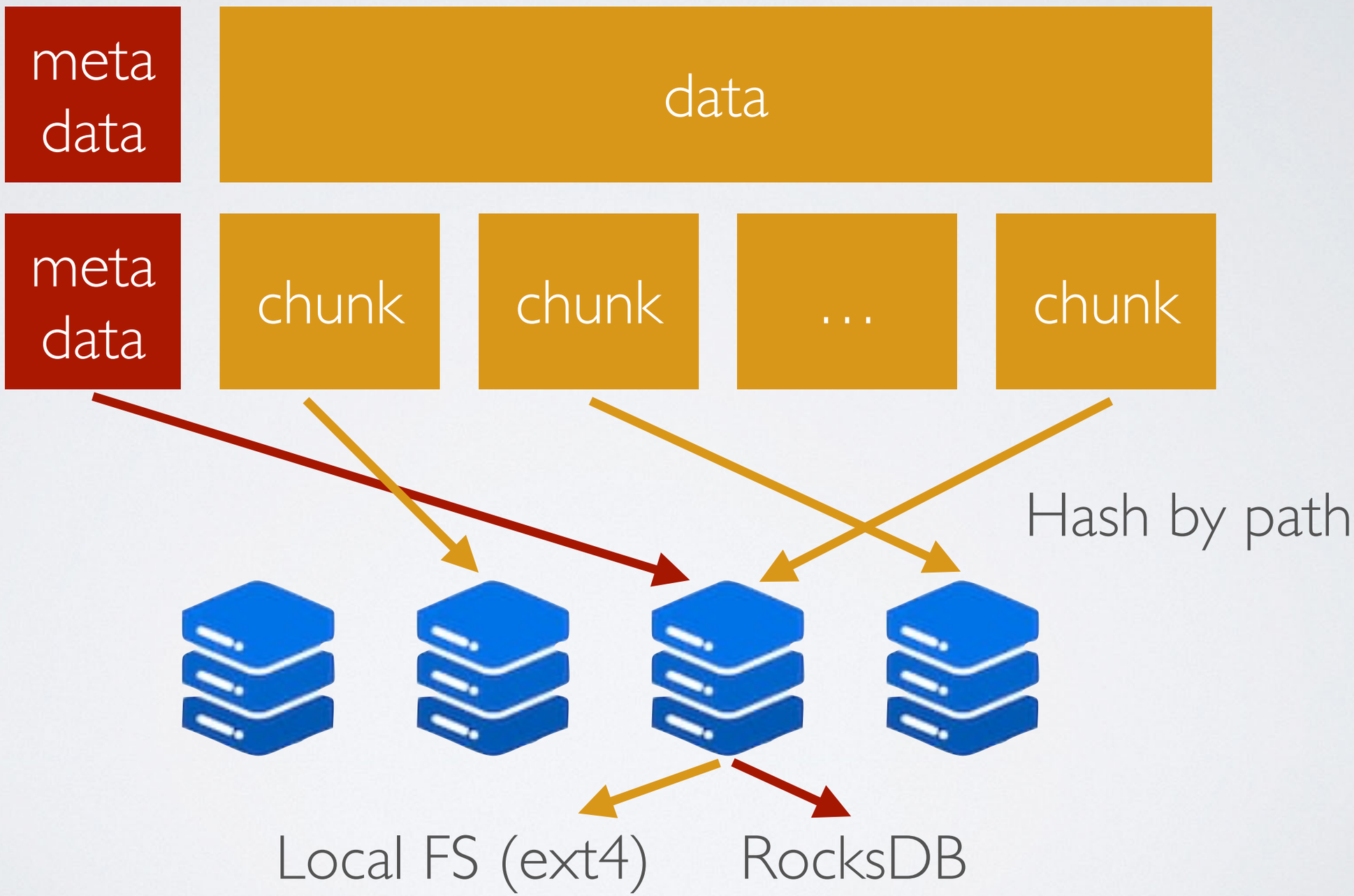
Please see also the 10 node challenge ranked list.

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	kIOP/s
1	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
2	sc19	Intel	Wolf	Intel	DAOS	26	728	zip	933.64	183.36	4753.79
3	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	453.68	209.43	982.78
4	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
5	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	229.45	131.25	401.13

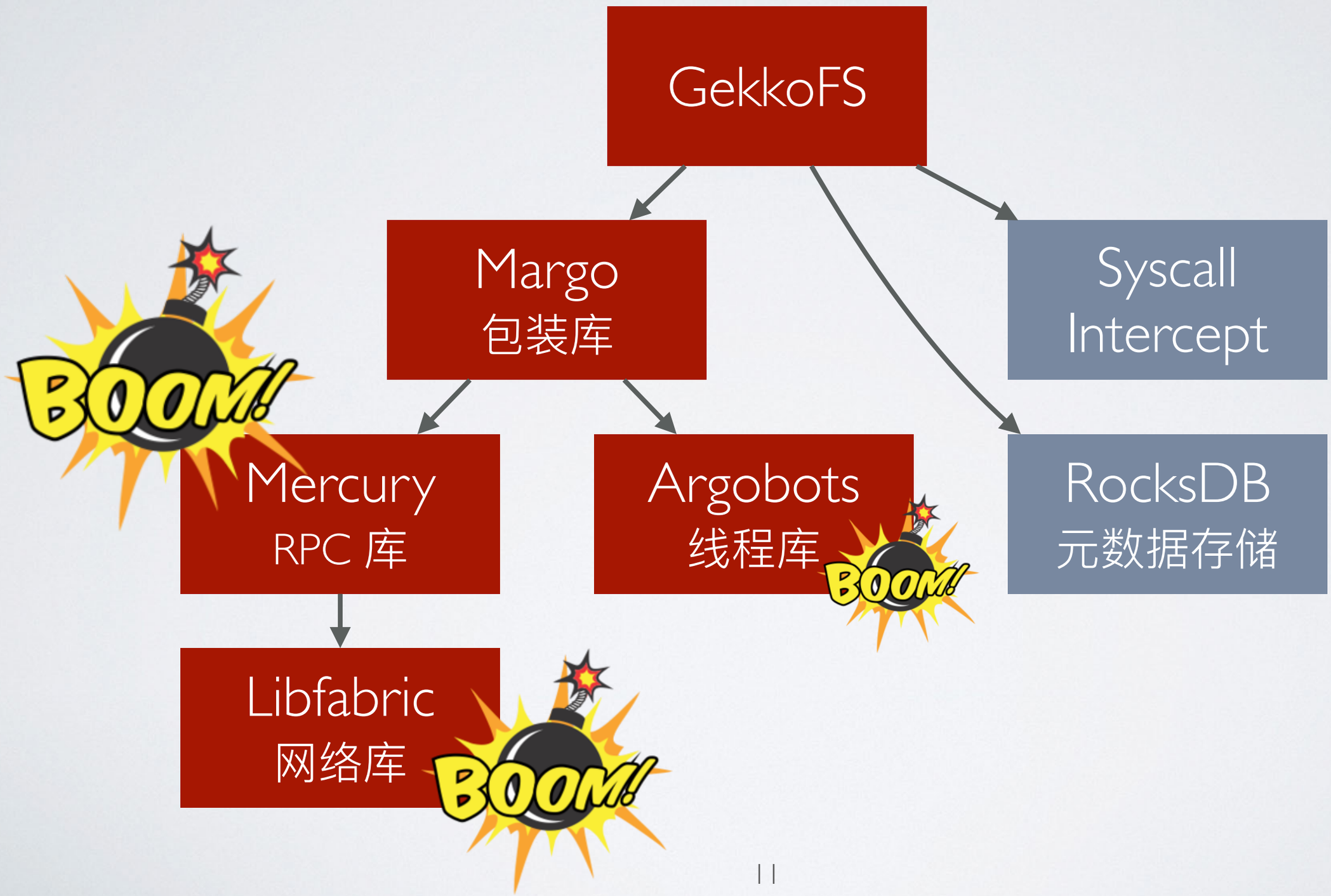
鹏城云脑 II

- 512 Nodes
- CPU: Kunpeng 920 ARMv8
  4 sockets x 48 cores, 2.6GHz
- RAM: 2T, DDR4-2666
- NIC: HiSilicon 100GE x2
  with RoCE RDMA
- SSD: NVMe SSD 3.2T x6

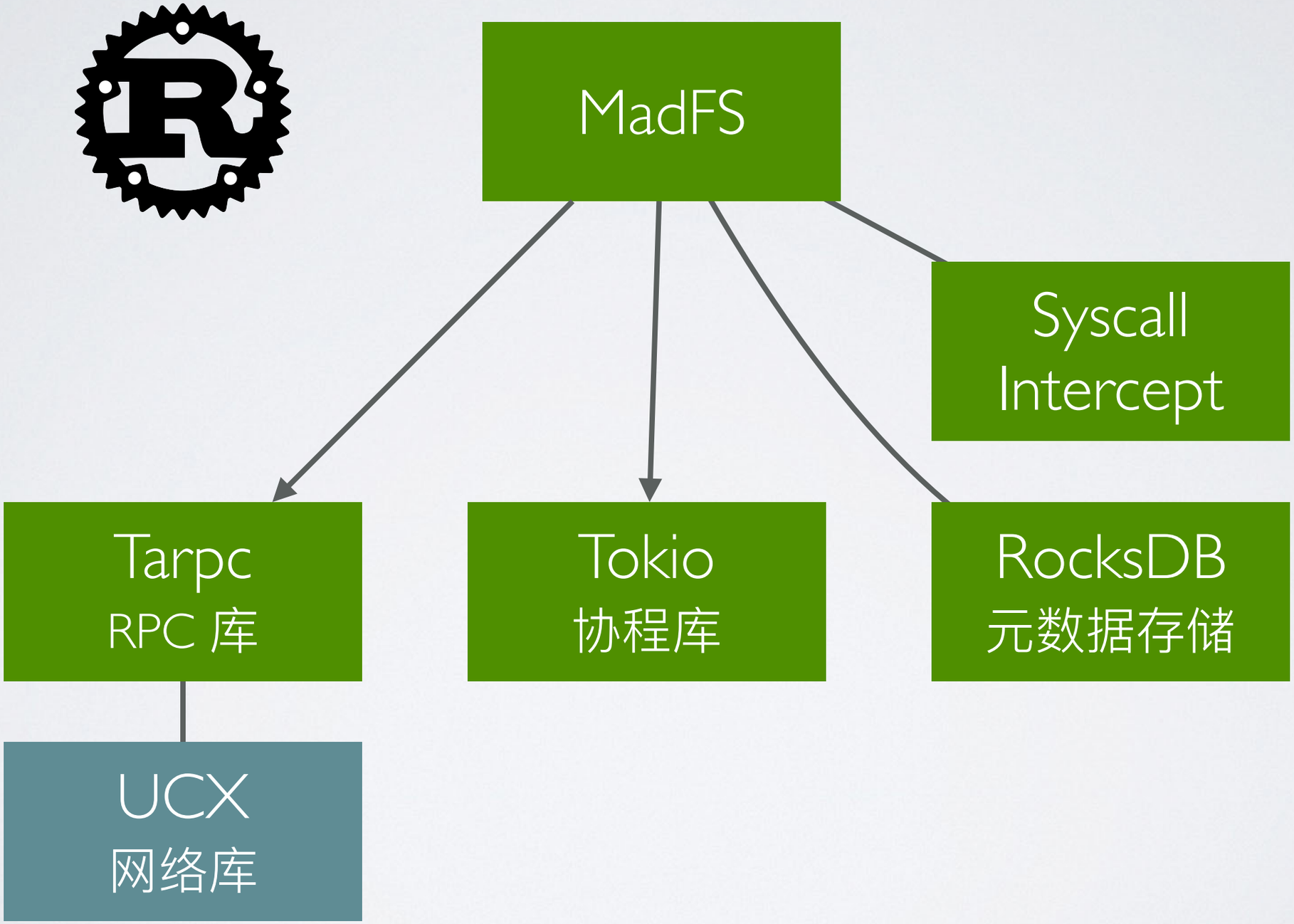
GEKKOFS 的原理


meta data
data
meta data
chunk
chunk
…
chunk
Hash by path
Local FS (ext4)
RocksDB

GekkoFS 无法运行

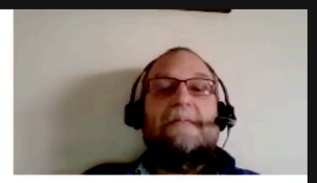


自研 MadFS


MadFS
Syscall Intercept
Tarpc
RPC 库
Tokio
协程库
RocksDB
元数据存储
UCX
网络库

性能调优

8000
6000
4000
2000
0
213
253
253
730
778
1,161
2,202
2,985
4,698
6,261
7,043
Rank1: 1793
21
22
23
24
25
26
27
28
29
30
31



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory

#1 in the IO-500 MD Score

IO500

Nov. 2020

IO-500 Steering Board

http://io500.org/list/20-11/

2:35 / 3:04

SC20

SC20 BoF, The IO500 list winners and awards

4次观看 • 2021年2月17日

0 0 分享 保存

IO500 Committee

订阅

Presented by Jay Lofstead

512 节点全系统测试

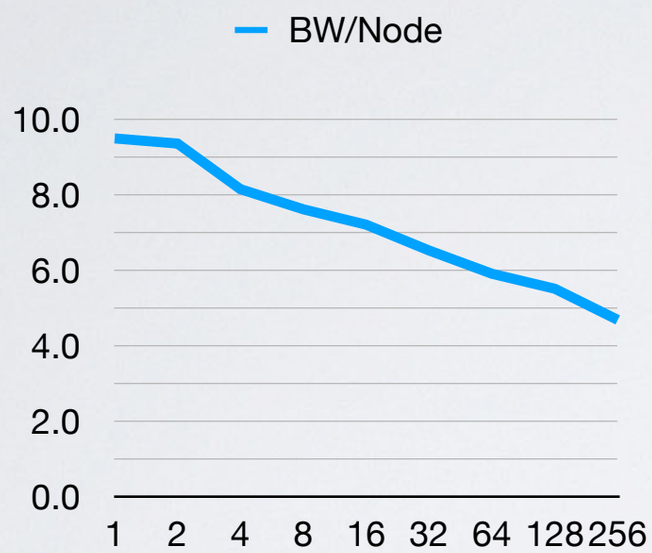
	255n12s72c	503n12s48c
ior-easy-write	1424.809	2944.813
mdtest-easy-write	60169.633	122778.106
ior-hard-write	1245.348	1374.518
mdtest-hard-write	9453.619	18844.000
find	19359.924	14694.684
ior-easy-read	2157.711	3715.881
mdtest-easy-stat	84376.480	241403.951
ior-hard-read	1238.813	1171.215
mdtest-hard-stat	106257.881	187451.307
mdtest-easy-delete	61848.108	126540.996
mdtest-hard-read	26439.065	36242.401
mdtest-hard-delete	10115.189	11733.647
Bandwidth	1475.746	2048.692
IOPS	33622.192	54921.962
Score	7043.991	10607.459

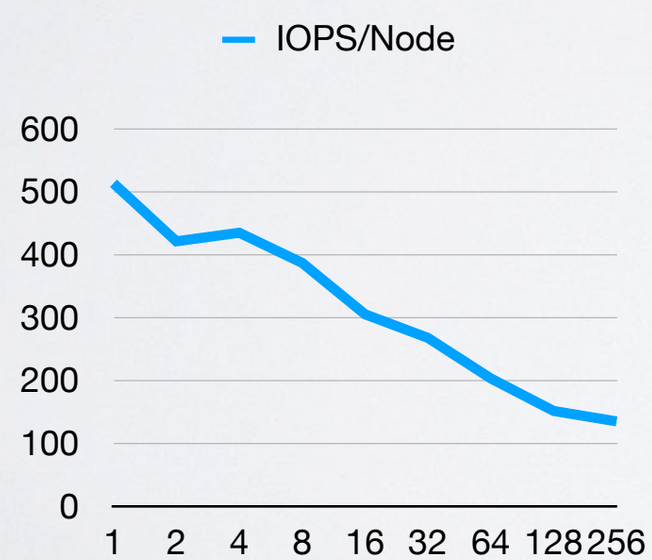
扩展性测试

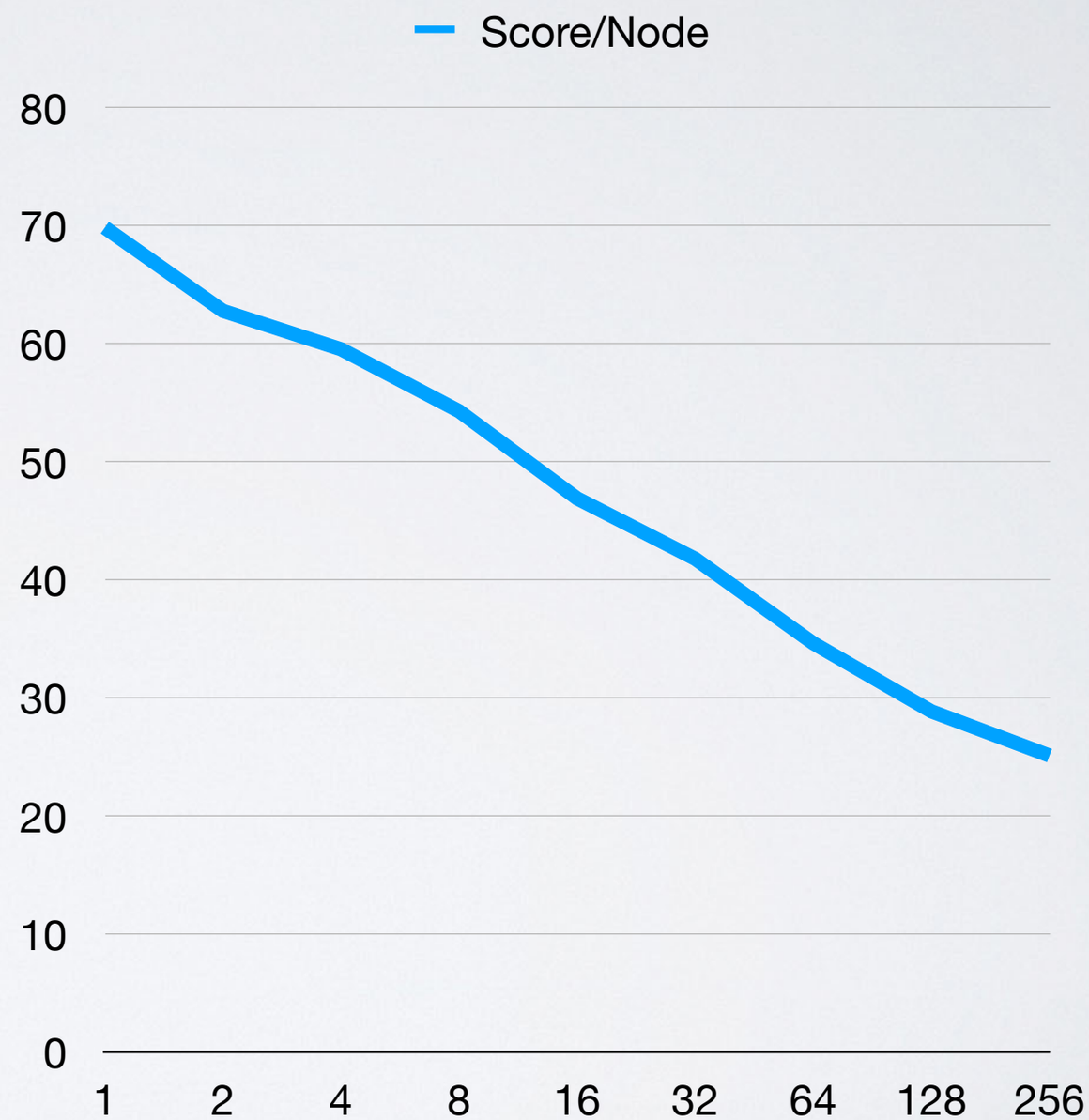
扩展性测试 (12s, 48c, 300s, 共用节点)

	1	2	4	8	16	32	64	128	256	365	365-scale
ior-easy-write	15.668	31.181	42.831	76.934	146.516	286.362	569.452	991.591	1380.423	1558.283	0.272
mdtest-easy-write	602.952	1122.227	2128.583	4090.817	7389.795	12159.529	21122.843	37935.115	72837.299	101065.176	0.459
ior-hard-write	9.099	19.043	36.956	69.530	125.979	184.306	269.090	500.628	836.611	1138.561	0.343
mdtest-hard-write	67.183	89.005	154.860	303.560	639.286	1010.322	2461.546	4523.532	9018.084	12581.030	0.513
find	12626.065	13110.963	30408.841	35995.018	29679.795	33974.614	22140.179	20598.944	18280.579	18258.818	0.004
ior-easy-read	13.789	25.424	44.281	81.662	159.876	293.570	614.272	1151.975	2111.876	2833.019	0.563
mdtest-easy-stat	1054.946	1807.320	4025.733	7907.933	13240.023	30935.608	27811.350	55242.541	139445.760	131176.322	0.341
ior-hard-read	4.152	8.166	16.194	31.724	60.311	123.385	218.701	436.030	856.297	1297.969	0.856
mdtest-hard-stat	1061.865	2123.316	3991.191	7714.356	14549.111	31251.892	44148.175	74206.616	138764.471	193362.271	0.499
mdtest-easy-delete	608.089	1198.210	2350.512	4475.549	8164.437	11990.964	18798.406	36072.989	81409.572	109067.245	0.491
mdtest-hard-read	146.614	288.706	564.659	1097.938	2132.117	4265.503	8056.708	15241.396	26953.162	32424.032	0.606
mdtest-hard-delete	93.543	143.778	386.490	619.615	673.710	1372.431	3706.461	2405.593	3794.328	11526.672	0.338
Bandwidth	9.505	18.738	32.640	61.013	115.502	209.101	378.784	706.646	1202.146	1598.193	0.461
IOPS	512.753	840.877	1736.333	3089.589	4870.690	8541.274	12955.056	19289.175	34341.575	47047.371	0.251
Score	69.812	125.523	238.064	434.173	750.050	1336.410	2215.213	3691.967	6425.230	8671.261	0.340
BW/Node	9.505	9.369	8.160	7.627	7.219	6.534	5.919	5.521	4.696	4.379	⚠️
IOPS/Node	512.753	420.439	434.083	386.199	304.418	266.915	202.423	150.697	134.147	128.897	⚠️
Score/Node	69.812	62.762	59.516	54.272	46.878	41.763	34.613	28.843	25.099	23.757	⚠️

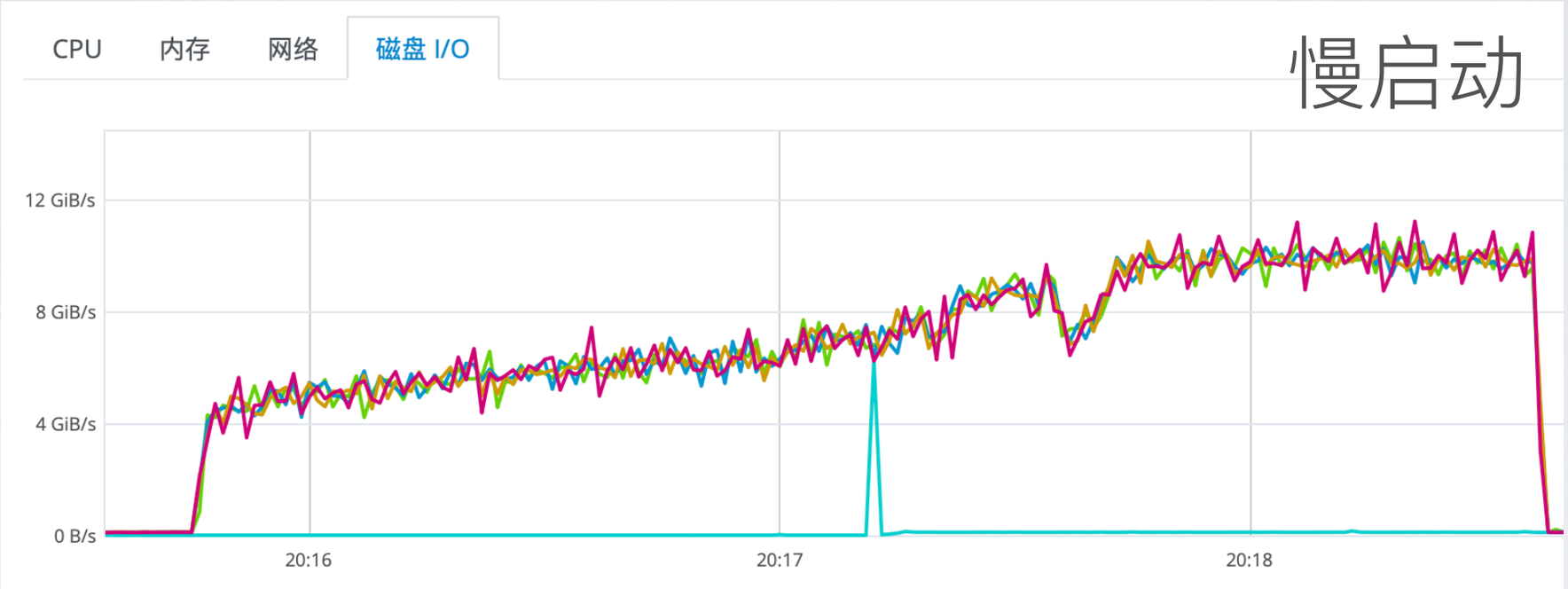
扩展性测试


BW/Node
10.0
8.0
6.0
4.0
2.0
0.0
1 2 4 8 16 32 64 128 256


IOPS/Node
600
500
400
300
200
100
0
1 2 4 8 16 32 64 128 256

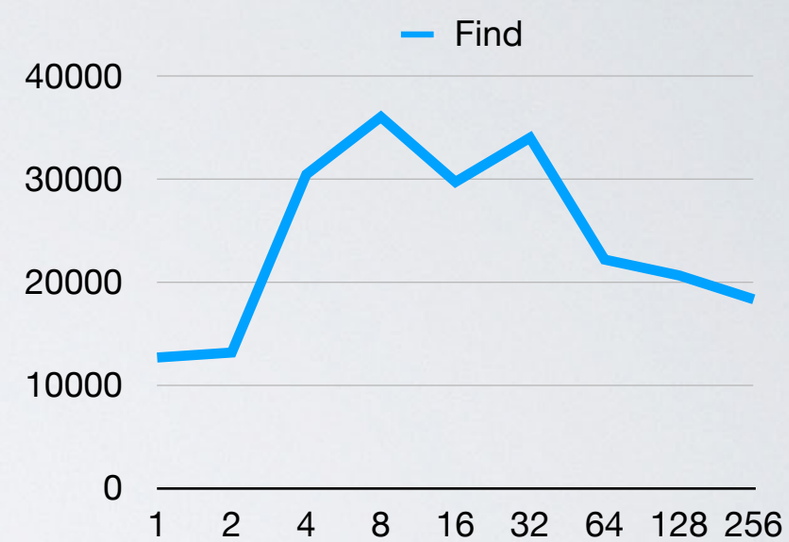

Score/Node
80
70
60
50
40
30
20
10
0
1 2 4 8 16 32 64 128 256

IOR-WRITE 带宽问题



Find 扩展性问题

```
IO500 version io500-sc20
[RESULT]       ior-easy-write     1424.808689 GiB/s : time 564.166 seconds
[RESULT]    mdtest-easy-write    60169.632729 kIOPS : time 377.775 seconds
[RESULT]       ior-hard-write     1245.347896 GiB/s : time 467.116 seconds
[RESULT]    mdtest-hard-write     9453.618969 kIOPS : time 372.813 seconds
[RESULT]                find    19359.924471 kIOPS : time 1350.331 seconds
[RESULT]        ior-easy-read     2157.710521 GiB/s : time 372.556 seconds
[RESULT]     mdtest-easy-stat    84376.480140 kIOPS : time 268.277 seconds
[RESULT]        ior-hard-read     1238.813319 GiB/s : time 466.851 seconds
[RESULT]     mdtest-hard-stat   106257.881498 kIOPS : time 33.077 seconds
[RESULT]   mdtest-easy-delete    61848.107522 kIOPS : time 365.974 seconds
[RESULT]     mdtest-hard-read    26439.064550 kIOPS : time 132.757 seconds
[RESULT]   mdtest-hard-delete    10115.188911 kIOPS : time 346.939 seconds
[SCORE] Bandwidth 1475.745858 GiB/s : IOPS 33622.191788 kiops : TOTAL 7043.991074
```

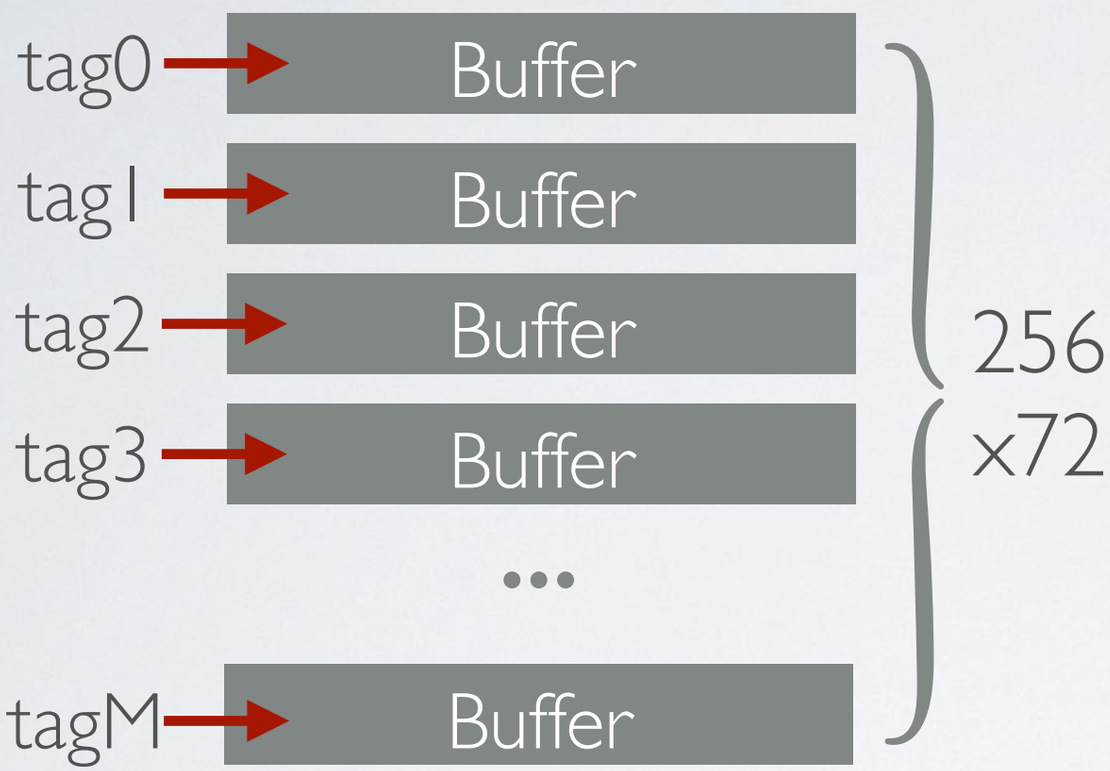


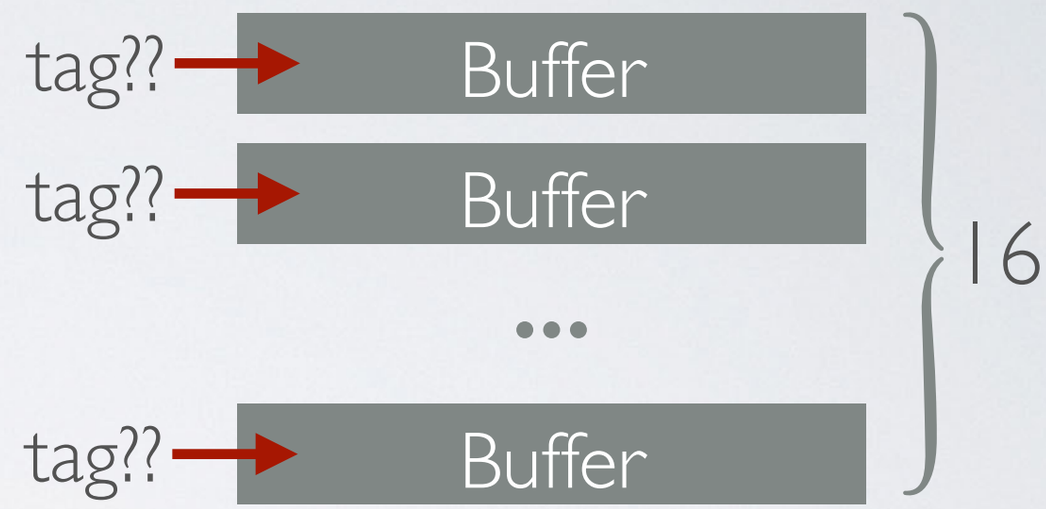
- mdtest-hard 在一个文件夹下创建了 **~3,500,000,000** 个文件
- pfind 只能用一个进程扫描它们

Part2

MadFS 针对 IO500 的新优化

内存优化





为每个链接创建一个协程+Buffer

256节点 × 72客户端 × 12服务端 × 5MB ≈ 1.1TB

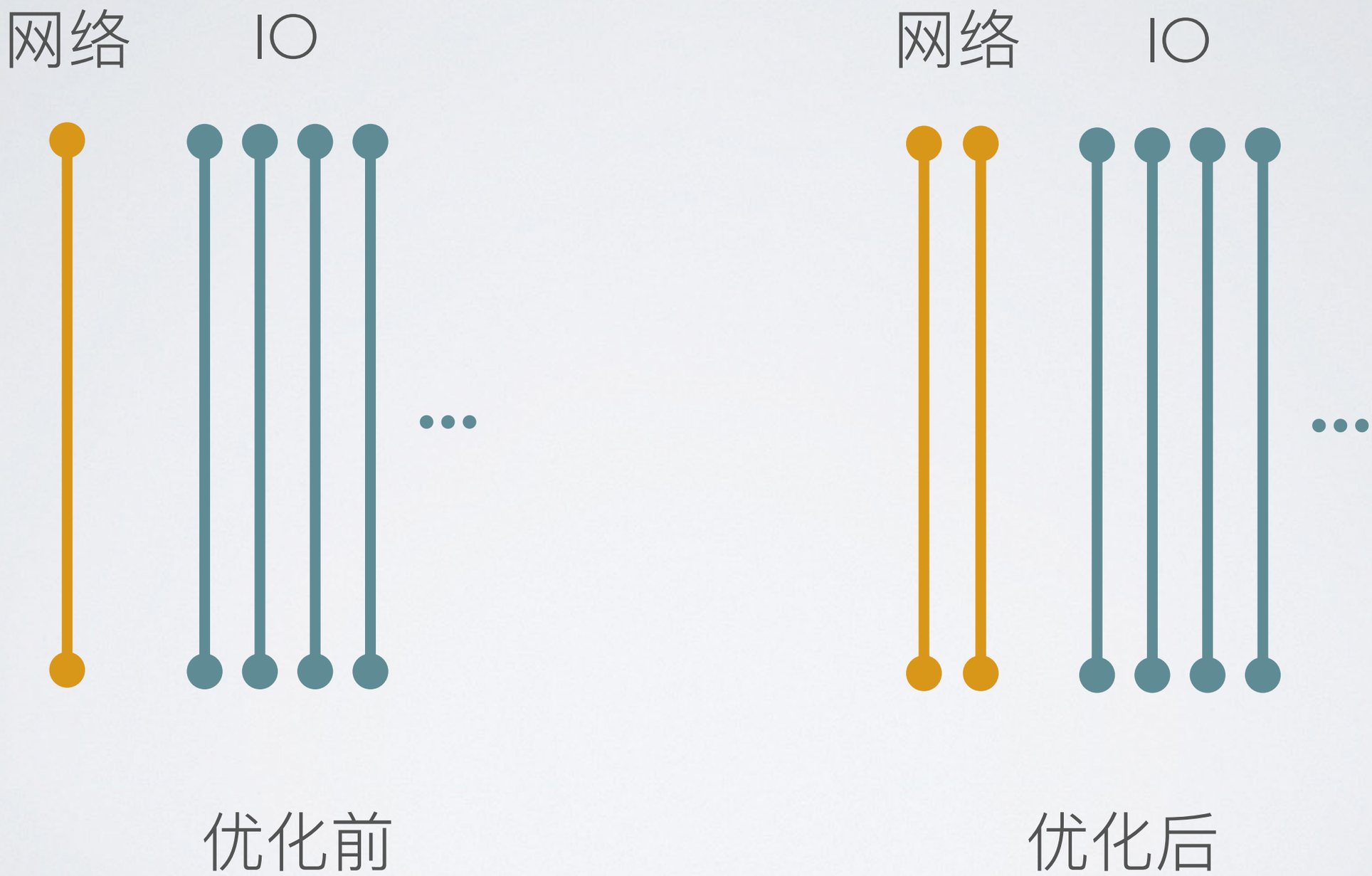
优化前

所有链接复用16个协程+Buffer

16协程 × 12服务端 × 5MB ≈ 1GB

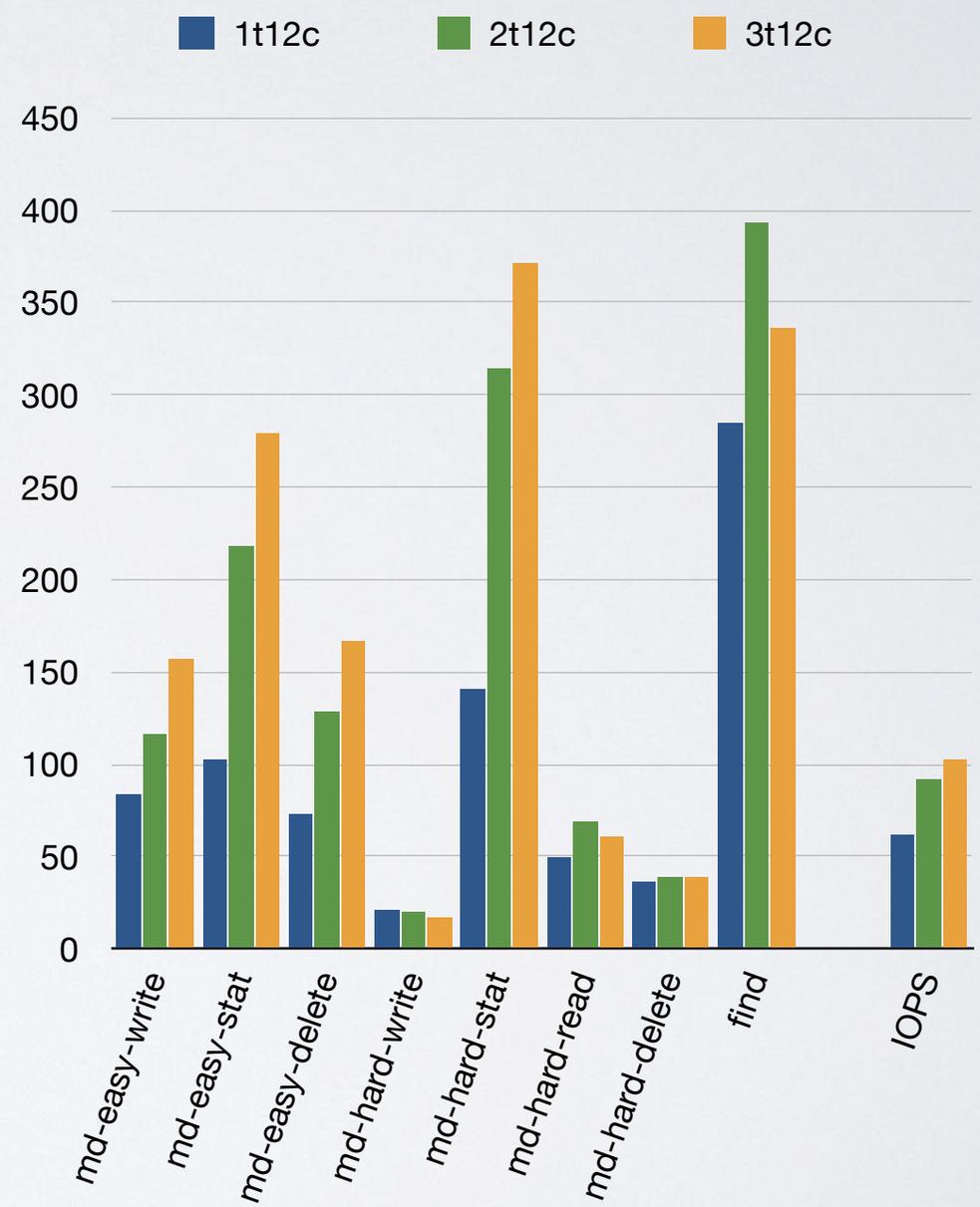
优化后

网络多线程优化



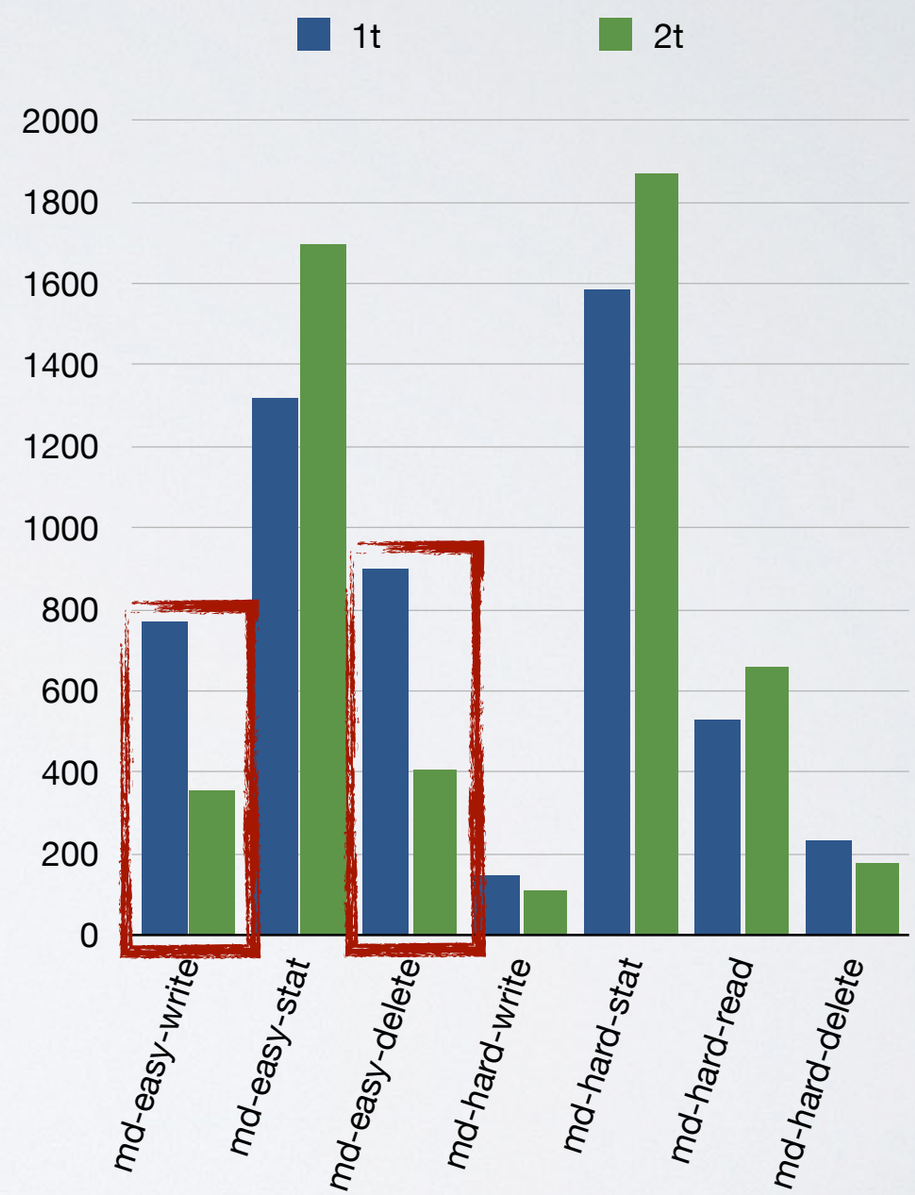
x86 效果不错

	1t12c	2t12c	3t12c
ior-easy-write	1.445	1.688	1.620
mdtest-easy-write	83.570	117.029	157.625
ior-hard-write	1.218	1.284	1.059
mdtest-hard-write	21.067	19.521	17.931
find	284.347	393.564	336.641
ior-easy-read	2.695	2.377	2.278
mdtest-easy-stat	103.330	218.089	278.688
ior-hard-read	2.120	1.847	1.867
mdtest-hard-stat	140.942	314.246	371.039
mdtest-easy-delete	72.694	128.357	166.760
mdtest-hard-read	49.364	68.354	60.904
mdtest-hard-delete	36.328	39.273	39.390
Bandwidth	1.781	1.756	1.643
IOPS	74.514	109.869	118.681
Score	11.519	13.891	13.965



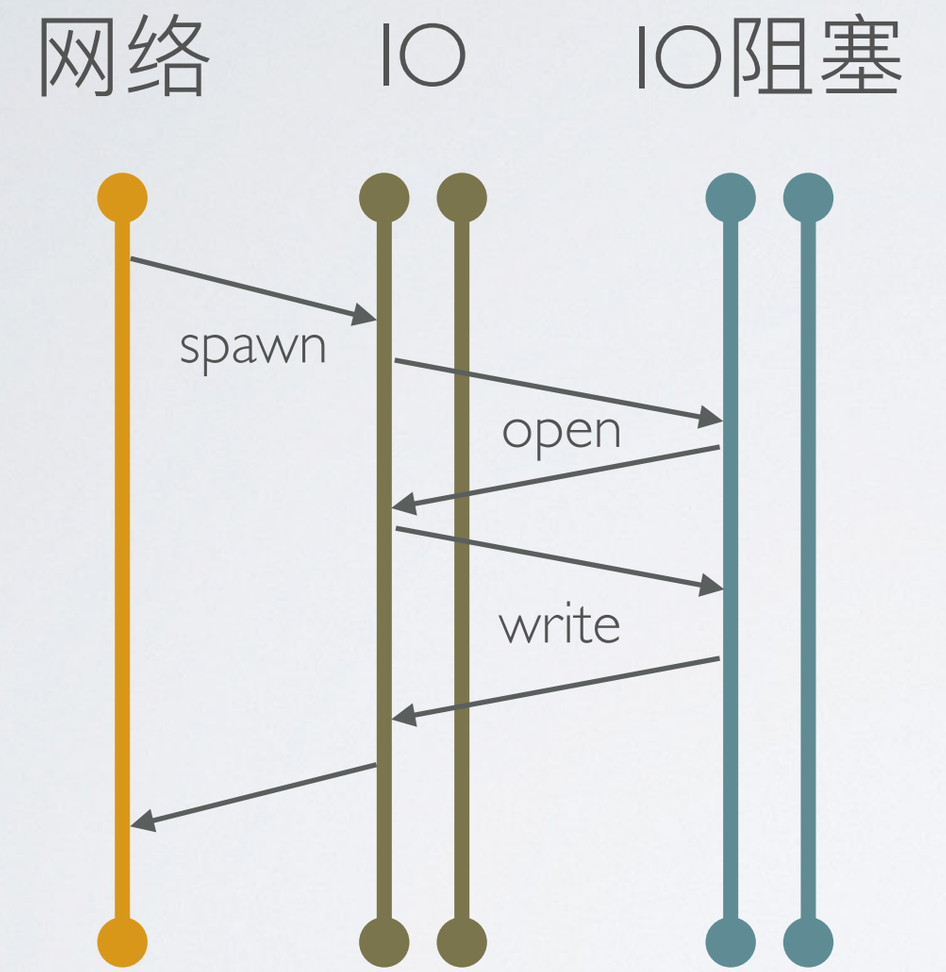
ARM 性能腰斩

	1n12s48c	1n12s2t48c
ior-easy-write	15.231	15.112
mdtest-easy-write	766.790	352.520
ior-hard-write	9.002	11.236
mdtest-hard-write	144.611	110.057
find	13597.861	8953.709
ior-easy-read	21.746	21.950
mdtest-easy-stat	1321.550	1695.604
ior-hard-read	17.539	21.508
mdtest-hard-stat	1584.197	1867.437
mdtest-easy-delete	899.968	407.251
mdtest-hard-read	528.737	655.647
mdtest-hard-delete	232.076	174.295
Bandwidth	15.122	16.826
IOPS	876.582	689.682
Score	115.135	107.726



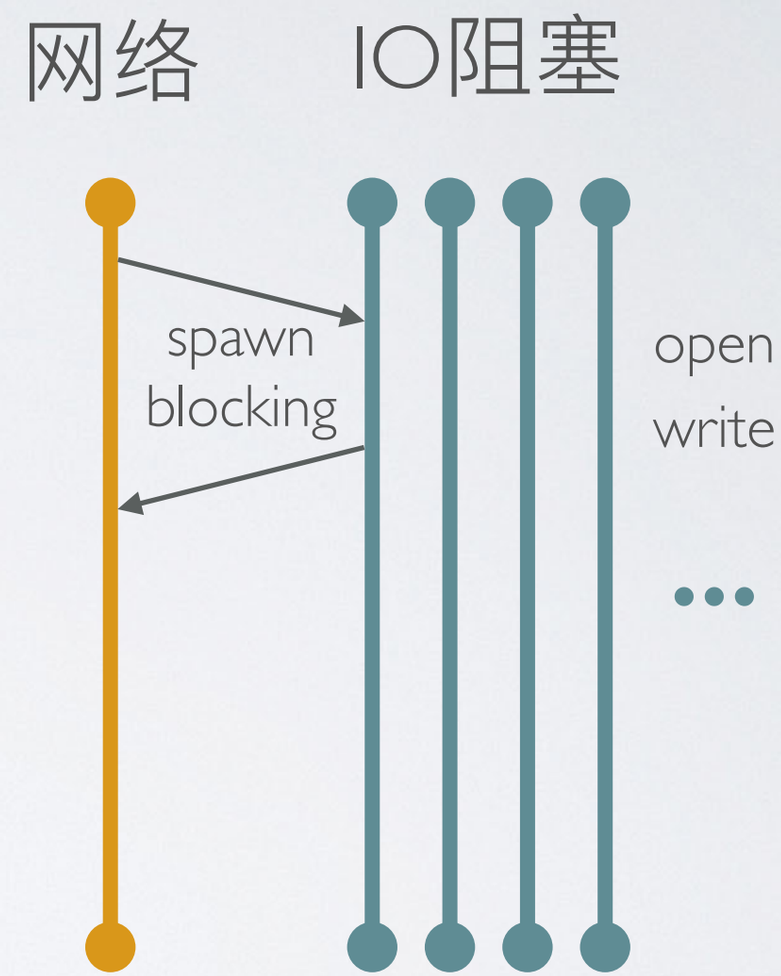
经查原因是 RocksDB 并发写冲突！

本地文件读写优化



`tokio::File::write().await`

优化前

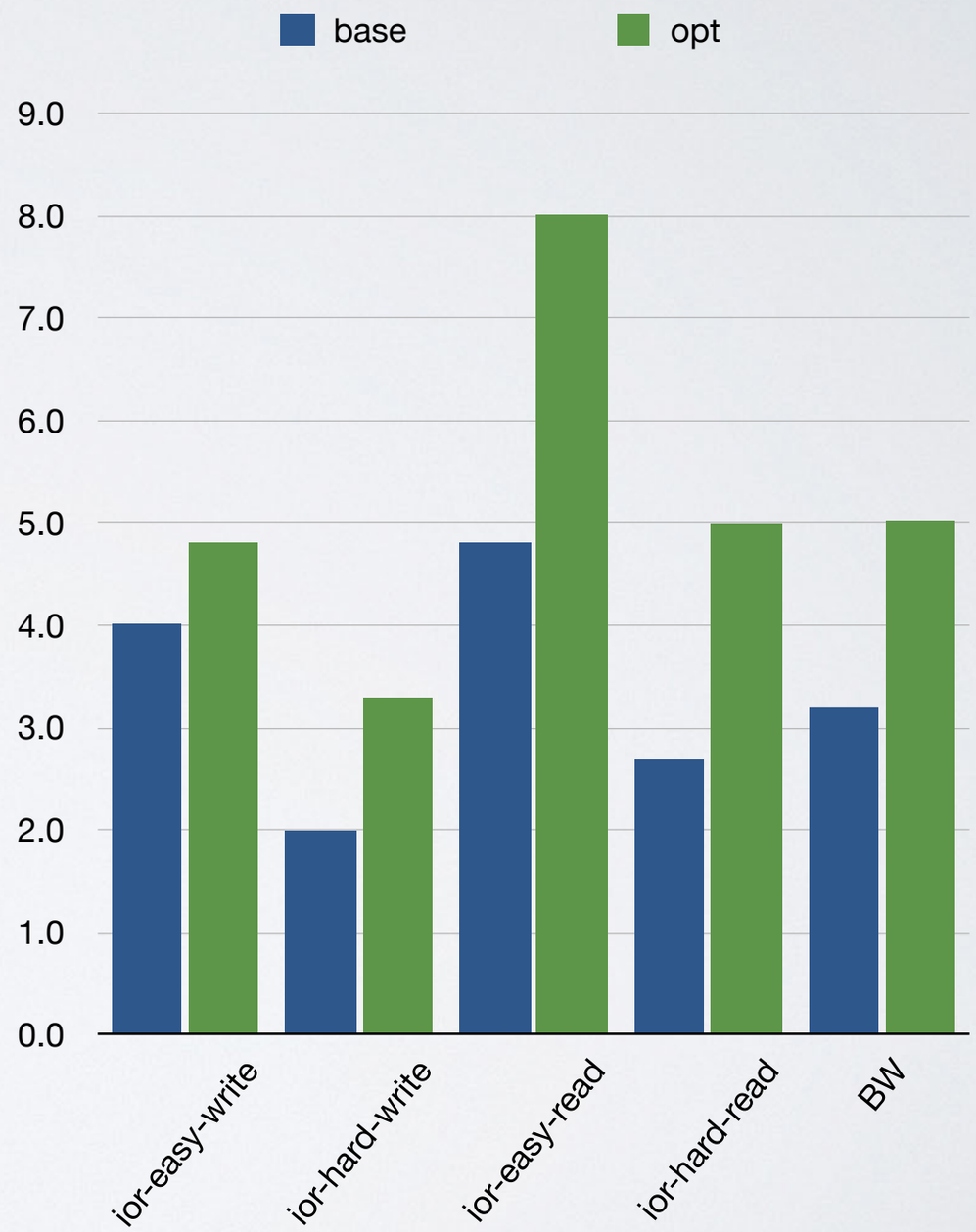


`std::File::write()`

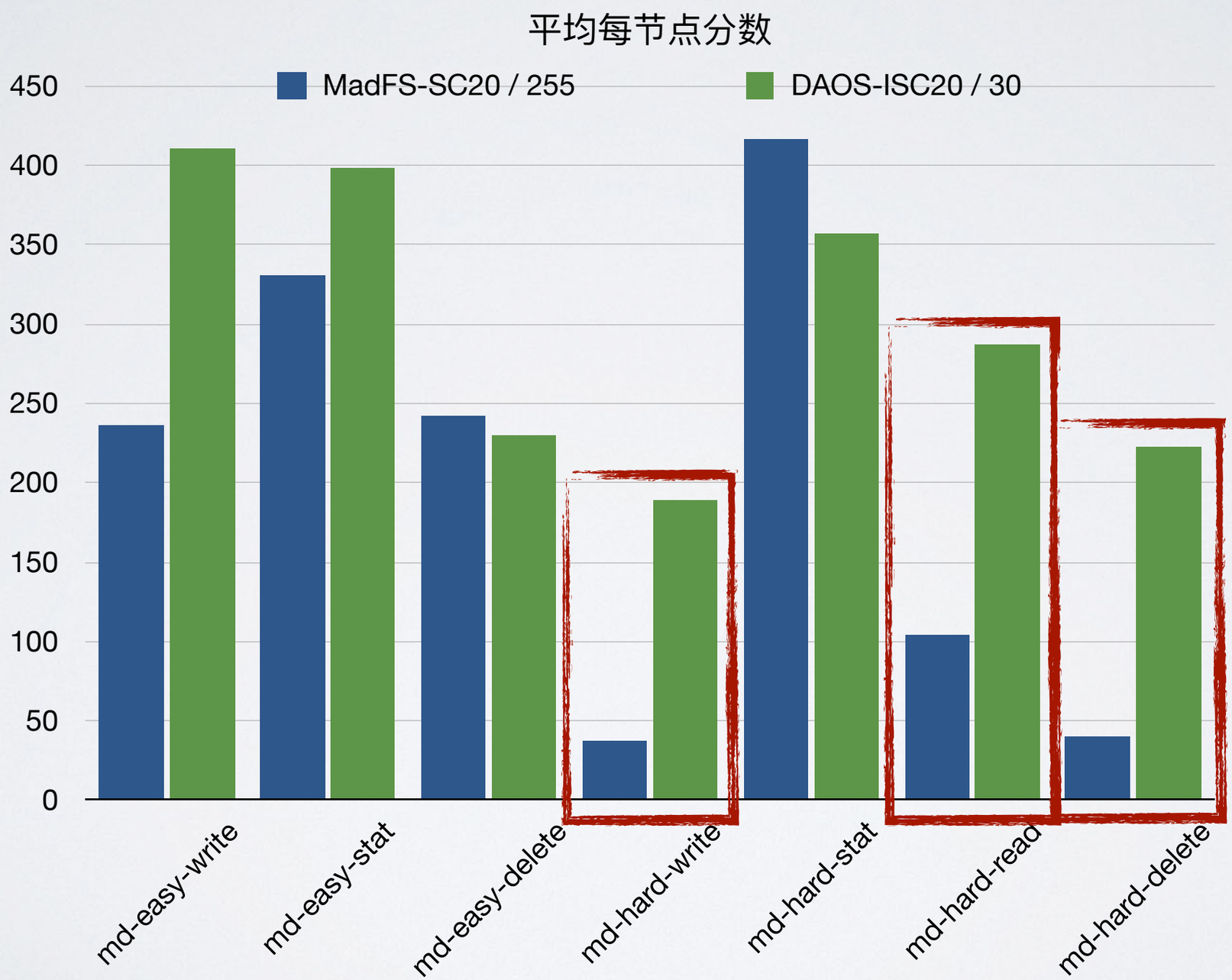
优化后

本地文件读写优化

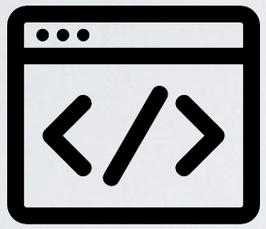
	base	opt
ior-easy-write	4.000	4.800
mdtest-easy-write	327.000	245.000
ior-hard-write	2.000	3.300
mdtest-hard-write	18.000	19.000
find	401.000	549.000
ior-easy-read	4.800	8.000
mdtest-easy-stat	497.000	405.000
ior-hard-read	2.700	5.000
mdtest-hard-stat	668.000	447.000
mdtest-easy-delete	316.000	244.000
mdtest-hard-read	104.000	164.000
mdtest-hard-delete	84.000	84.000
Bandwidth	3.191	5.017
IOPS	195.834	187.919
Score	24.998	30.705



小文件优化

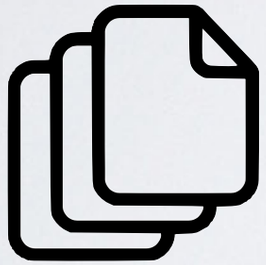

平均每节点分数
MadFS-SC20 / 255
DAOS-ISC20 / 30
450
400
350
300
250
200
150
100
50
0
md-easy-write
md-easy-stat
md-easy-delete
md-hard-write
md-hard-stat
md-hard-read
md-hard-delete

小文件优化



元数据 => RocksDB

{ 全路径 => 元数据 }



数据块 => 本地文件系统

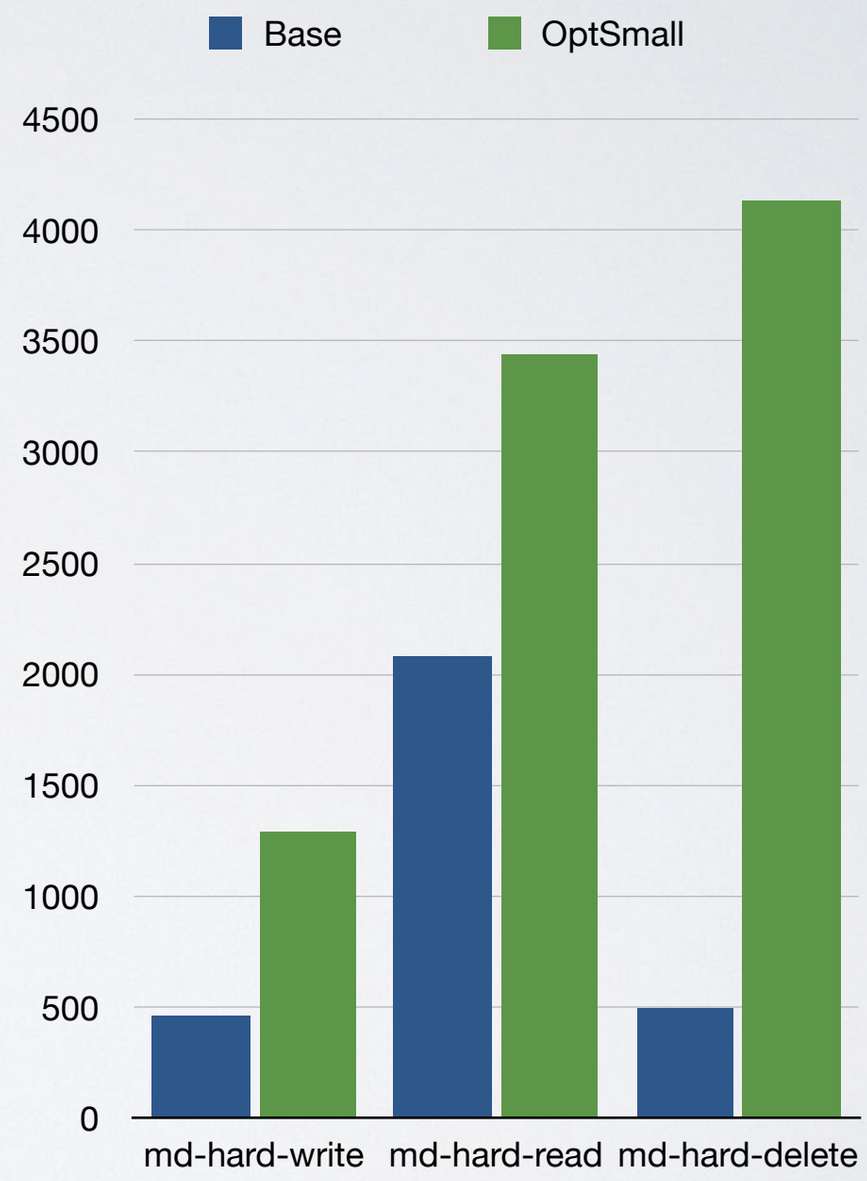
./全路径/块ID => 数据

第0块 => RocksDB

优化小文件性能

小文件优化

	Base	OptSmall	
ior-easy-write	94.122	92.497	
mdtest-easy-write	5378.439	5422.553	
ior-hard-write	80.854	78.944	
mdtest-hard-write	463.743	1285.563	+177%
find	44604.009	36128.417	
ior-easy-read	91.108	90.295	
mdtest-easy-stat	10056.221	10951.796	
ior-hard-read	60.748	59.330	
mdtest-hard-stat	10638.537	9502.525	
mdtest-easy-delete	6216.329	6212.316	
mdtest-hard-read	2080.537	3446.449	+66%
mdtest-hard-delete	490.780	4127.369	+741%
Bandwidth	80.560	79.085	
IOPS	4071.717	6245.916	+53%
Score	572.729	702.823	+23%



8 nodes x (18 server procs + 72 client procs), IO500 300s

FIND

扩展性测试 (12s, 48c, 300s, 共用节点)

	1	2	4	8	16	32	64	128	256	365	365-scale
ior-easy-write	15.668	31.181	42.831	76.934	146.516	286.362	569.452	991.591	1380.423	1558.283	0.272
mdtest-easy-write	602.952	1122.227	2128.583	4090.817	7389.795	12159.529	21122.843	37935.115	72837.299	101065.176	0.459
ior-hard-write	9.099	19.043	36.956	69.530	125.979	184.306	269.090	500.628	836.611	1138.561	0.343
mdtest-hard-write	67.183	89.005	154.860	303.560	639.286	1010.322	2461.546	4523.532	9018.084	12581.030	0.513
find	12626.065	13110.963	30408.841	35995.018	29679.795	33974.614	22140.179	20598.944	18280.579	18258.818	0.004
ior-easy-read	13.789	25.424	44.281	81.662	159.876	293.570	614.272	1151.975	2111.876	2833.019	0.563
mdtest-easy-stat	1054.946	1807.320	4025.733	7907.933	13240.023	30935.608	27811.350	55242.541	139445.760	131176.322	0.341
ior-hard-read	4.152	8.166	16.194	31.724	60.311	123.385	218.701	436.030	856.297	1297.969	0.856
mdtest-hard-stat	1061.865	2123.316	3991.191	7714.356	14549.111	31251.892	44148.175	74206.616	138764.471	193362.271	0.499
mdtest-easy-delete	608.089	1198.210	2350.512	4475.549	8164.437	11990.964	18798.406	36072.989	81409.572	109067.245	0.491
mdtest-hard-read	146.614	288.706	564.659	1097.938	2132.117	4265.503	8056.708	15241.396	26953.162	32424.032	0.606
mdtest-hard-delete	93.543	143.778	386.490	619.615	673.710	1372.431	3706.461	2405.593	3794.328	11526.672	0.338

IO500 的最后壁垒：唯一一项没有扩展性的

如果 FIND 可扩展……

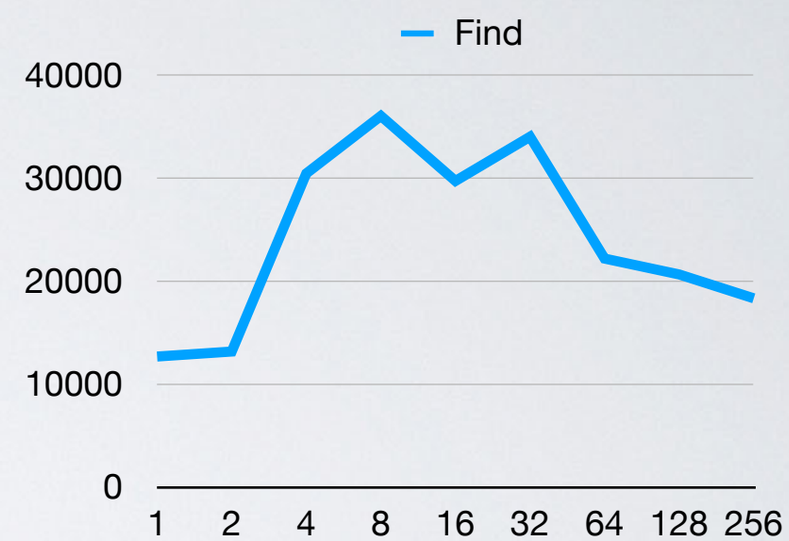
	503n12s48c	503n12s48c
ior-easy-write	2944.813	2944.813
mdtest-easy-write	122778.106	122778.106
ior-hard-write	1374.518	1374.518
mdtest-hard-write	18844.000	18844.000
find	14694.684	1469400.684
ior-easy-read	3715.881	3715.881
mdtest-easy-stat	241403.951	241403.951
ior-hard-read	1171.215	1171.215
mdtest-hard-stat	187451.307	187451.307
mdtest-easy-delete	126540.996	126540.996
mdtest-hard-read	36242.401	36242.401
mdtest-hard-delete	11733.647	11733.647
Bandwidth	2048.692	2048.692
IOPS	54921.962	97666.032
Score	10607.459	14145.233

单项可提升至少 100x

总分可提升至少 30%

Find 扩展性问题

```
IO500 version io500-sc20
[RESULT]       ior-easy-write     1424.808689 GiB/s : time 564.166 seconds
[RESULT]    mdtest-easy-write    60169.632729 kIOPS : time 377.775 seconds
[RESULT]       ior-hard-write     1245.347896 GiB/s : time 467.116 seconds
[RESULT]    mdtest-hard-write     9453.618969 kIOPS : time 372.813 seconds
[RESULT]                find    19359.924471 kIOPS : time 1350.331 seconds
[RESULT]        ior-easy-read     2157.710521 GiB/s : time 372.556 seconds
[RESULT]     mdtest-easy-stat    84376.480140 kIOPS : time 268.277 seconds
[RESULT]        ior-hard-read     1238.813319 GiB/s : time 466.851 seconds
[RESULT]     mdtest-hard-stat   106257.881498 kIOPS : time 33.077 seconds
[RESULT]   mdtest-easy-delete    61848.107522 kIOPS : time 365.974 seconds
[RESULT]     mdtest-hard-read    26439.064550 kIOPS : time 132.757 seconds
[RESULT]   mdtest-hard-delete    10115.188911 kIOPS : time 346.939 seconds
[SCORE] Bandwidth 1475.745858 GiB/s : IOPS 33622.191788 kiops : TOTAL 7043.991074
```



- mdtest-hard 在一个文件夹下创建了 **~3,500,000,000** 个文件
- pfind 只能用一个进程扫描它们

并行扫描单文件夹

README.md

Parallel find using MPI

This command is a drop-in replacement for find, implemented for using parallel access and MPI.

Usage:

```
# Scanning the homedir with a stonewall timer of 10 seconds, checking for files newer than test/0
$ mpiexec -n 2 ./pfind /home -s 10 -newer test/0.txt -name 01 -D rates -C
```

Compatibility for Single Directory Parallel Access

The tool provides parallel access to a single directory. However, this feature depends on the distribution of the "cookie" returned by telldir(). Depending on the system, it may work.

Compatibility

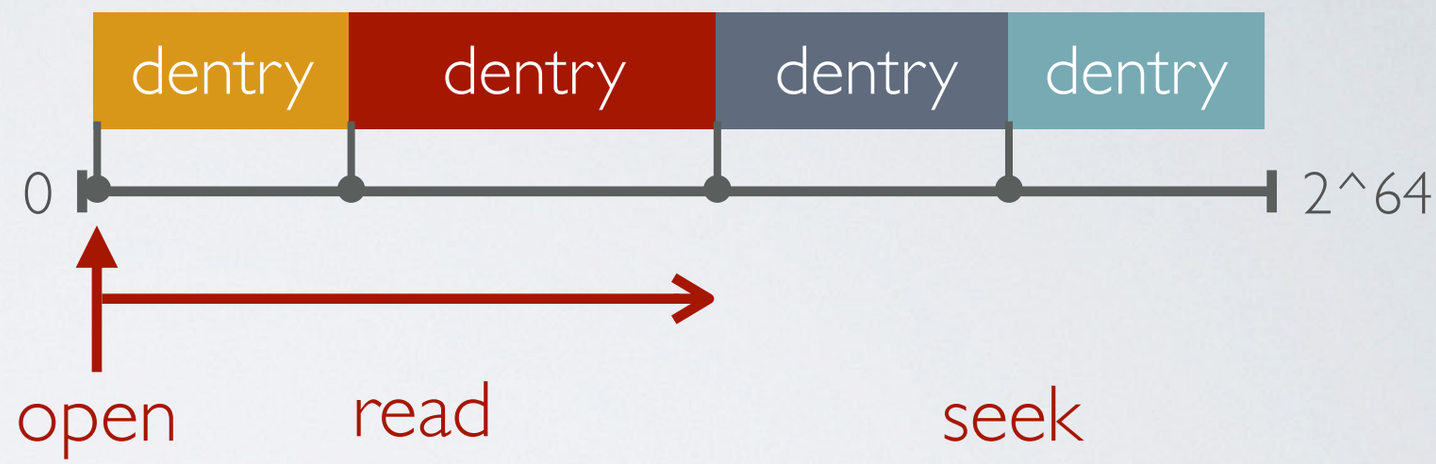
The option -H 1 works with:

- Lustre 2.5 (and later)
- NFSv3

Cookie?

POSIX 目录接口：

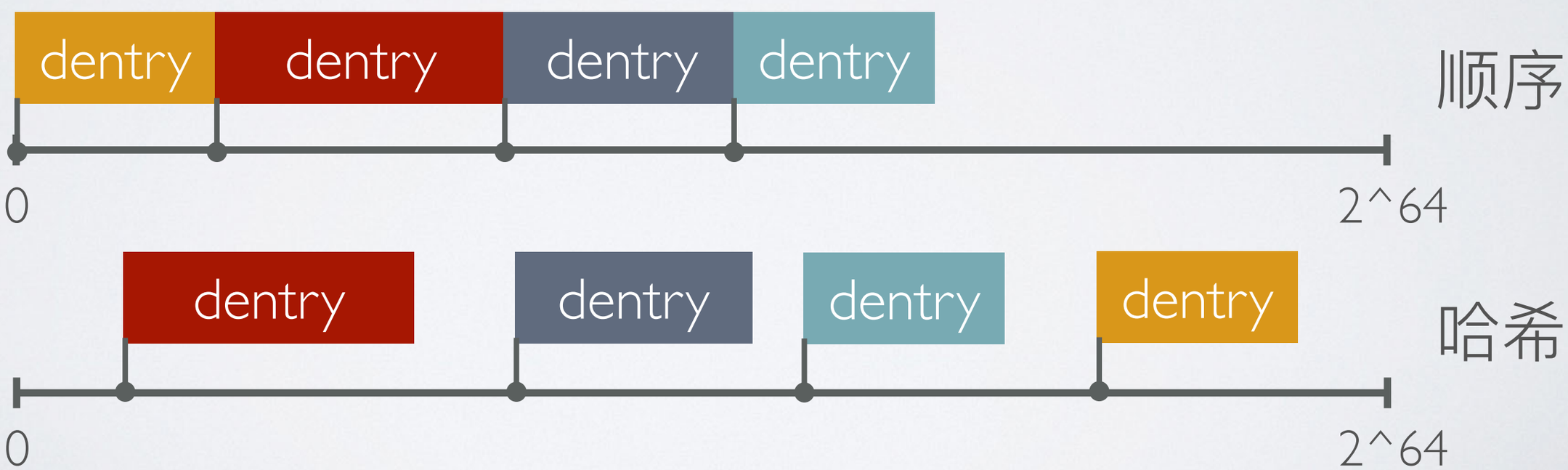
- opendir (open)
- readdir (getdents)
- closedir (close)
- **seekdir / telldir (lseek)**



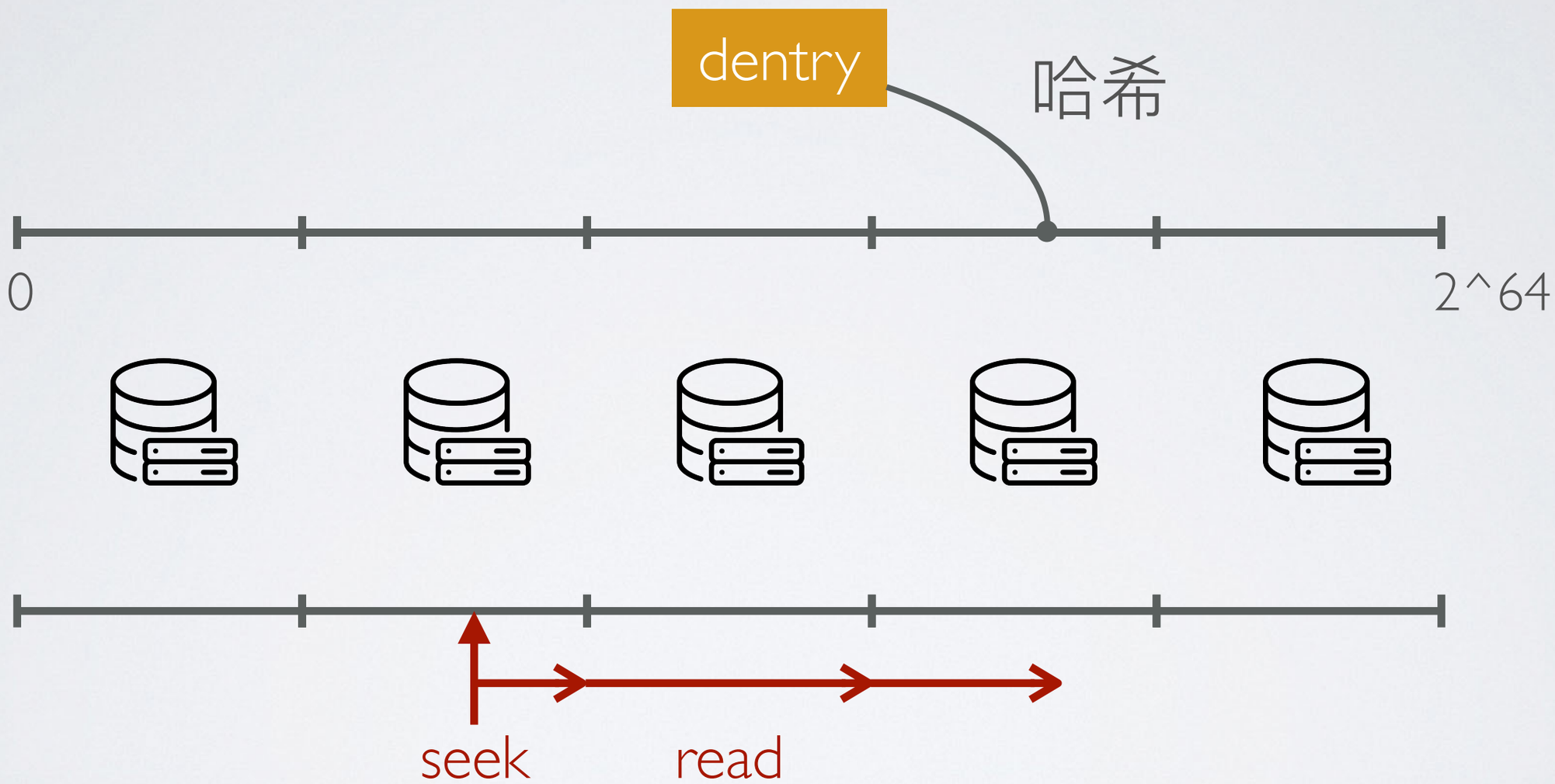

```
The getdents64() system call is like getdents(), except that its
second argument is a pointer to a buffer containing structures of
the following type:

    struct linux_dirent64 {
        ino64_t        d_ino;    /* 64-bit inode number */
        off64_t        d_off;    /* 64-bit offset to next structure */
        unsigned short d_reclen; /* Size of this dirent */
        unsigned char  d_type;   /* File type */
        char           d_name[]; /* Filename (null-terminated) */
    };
```

Cookie 的分配

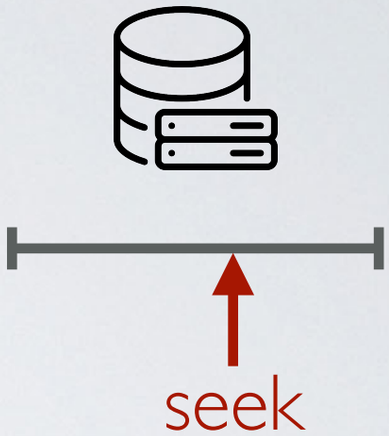
In early filesystems, the value returned by **telldir**() was a simple file offset within a directory. Modern filesystems use tree or hash structures, rather than flat tables, to represent directories. On such filesystems, the value returned by **telldir**() (and used internally by readdir(3)) is a "cookie" that is used by the implementation to derive a position within a directory. Application programs should treat this strictly as an opaque value, making *no* assumptions about its contents.

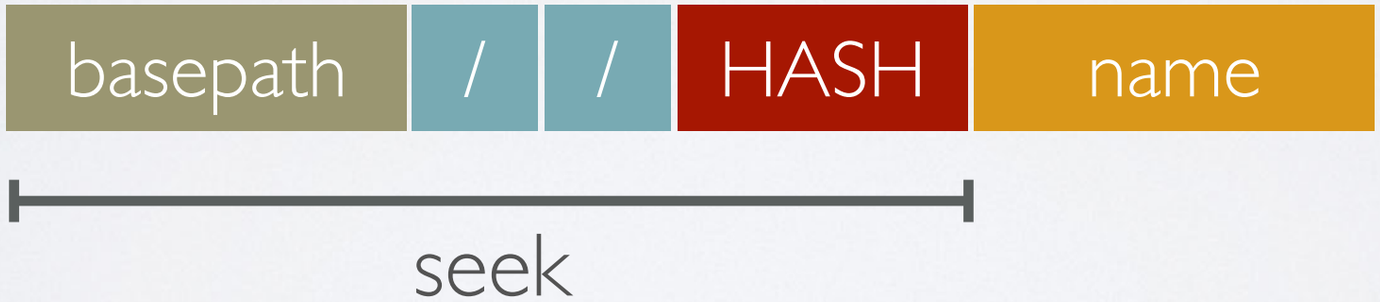


分布式系统 Cookie



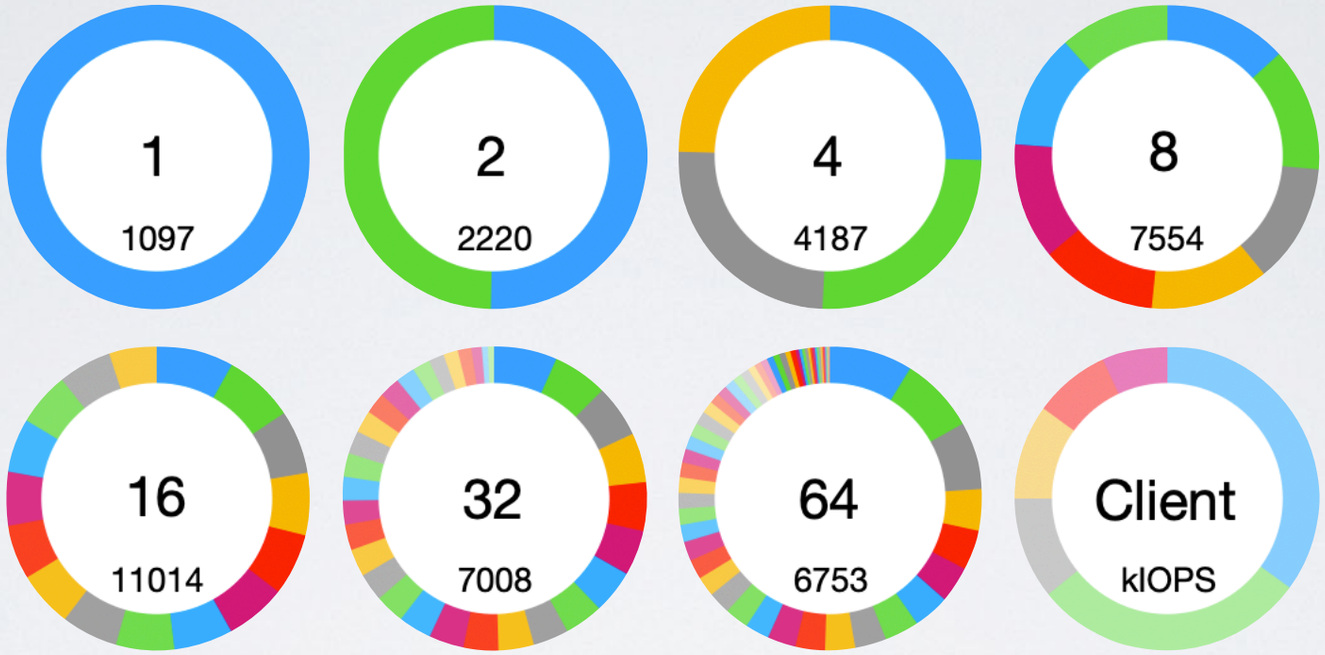
RocksDB Key

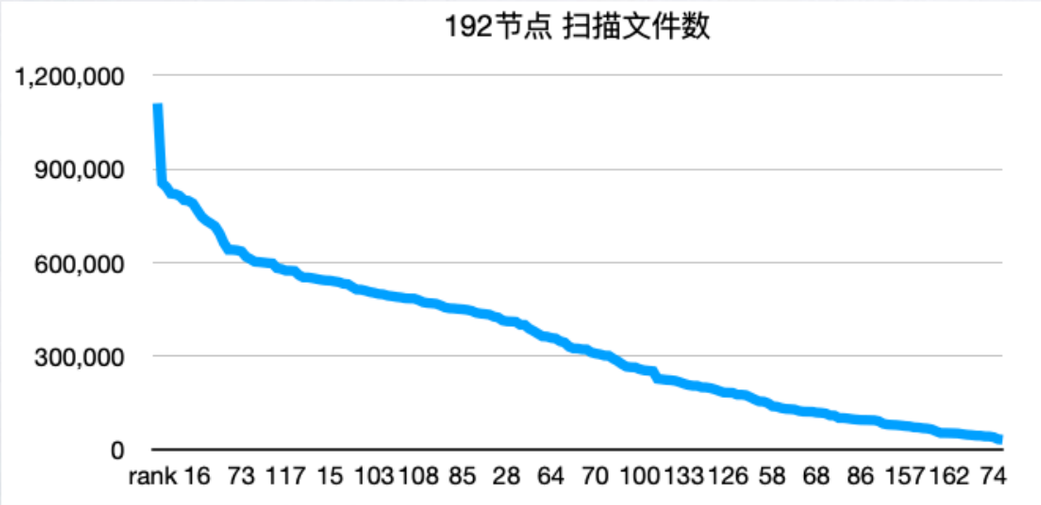


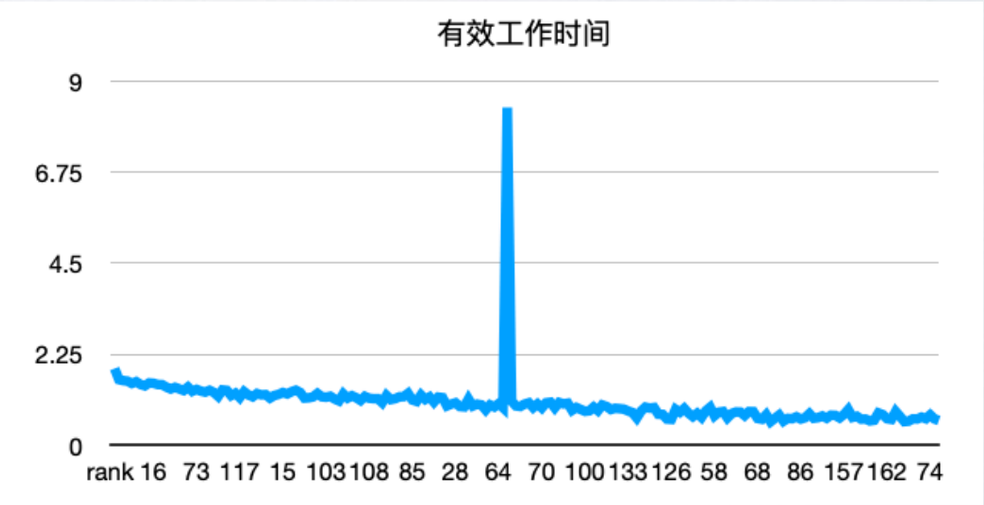


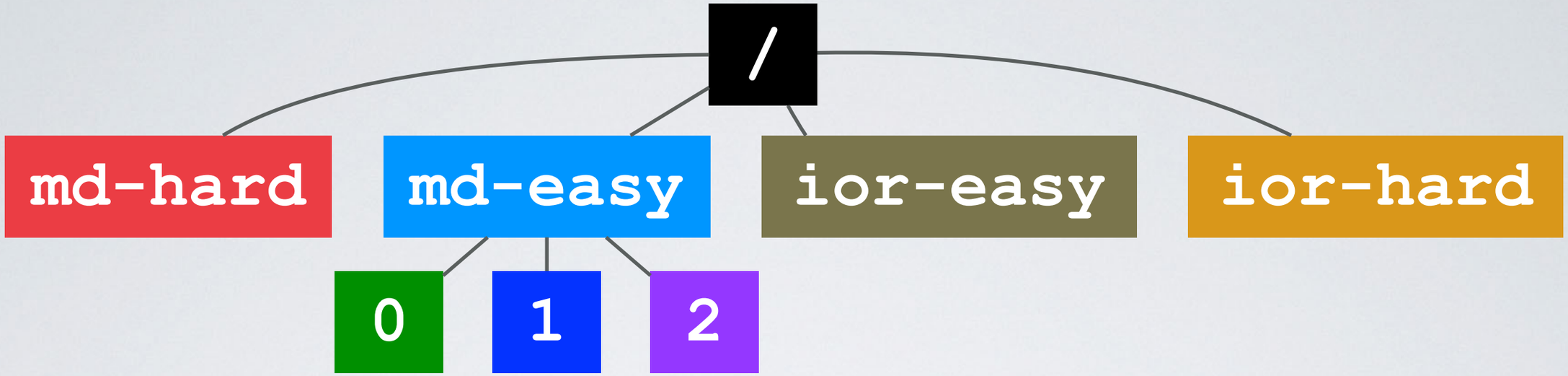
PFIND 动态任务切分

负载不均衡


1
1097
2
2220
4
4187
8
7554
16
11014
32
7008
64
6753
Client
kIOPS


192节点 扫描文件数
1,200,000
900,000
600,000
300,000
0
rank 16 73 117 15 103 108 85 28 64 70 100 133 126 58 68 86 157 162 74

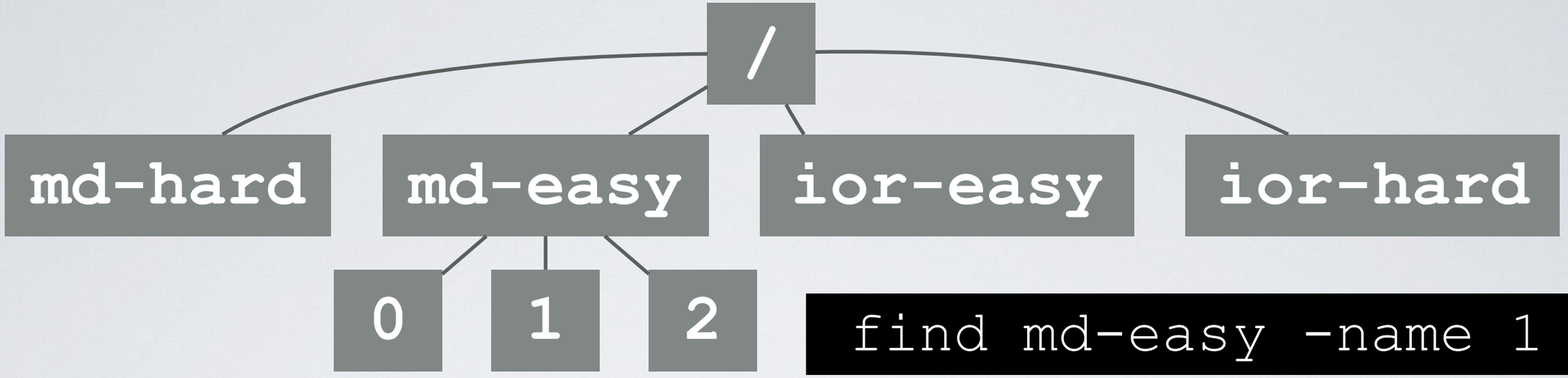

有效工作时间
9
6.75
4.5
2.25
0
rank 16 73 117 15 103 108 85 28 64 70 100 133 126 58 68 86 157 162 74



ior-easy/2
md-easy
md-easy/0/1
md-easy/0/2
md-easy/1/0
md-easy/2
md-hard/1
md-hard/3
md-hard/4

ior-easy
ior-hard/0
md-easy/0/0
md-easy/1
md-easy/1/2
md-easy/2/1
md-easy/2/2
md-hard
md-hard/2

ior-easy/0
ior-easy/1
ior-hard
md-easy/0
md-easy/1/1
md-easy/2/0
md-hard/0
md-hard/5




```
find md-easy -name 1
```

```
find md-easy -name 1
```

```
ior-easy/2
md-easy
md-easy/0/1
md-easy/0/2
md-easy/1/0
md-easy/2
md-hard/1
md-hard/3
md-hard/4
```

```
ior-easy
ior-hard/0
md-easy/0/0
md-easy/1
md-easy/1/2
md-easy/2/1
md-easy/2/2
md-hard
md-hard/2
```

```
ior-easy/0
ior-easy/1
ior-hard
md-easy/0
md-easy/1/1
md-easy/2/0
md-hard/0
md-hard/5
```

```
md-easy/0/1
md-easy/1
md-easy/2/1
md-easy/1/1
```

Madfs-find

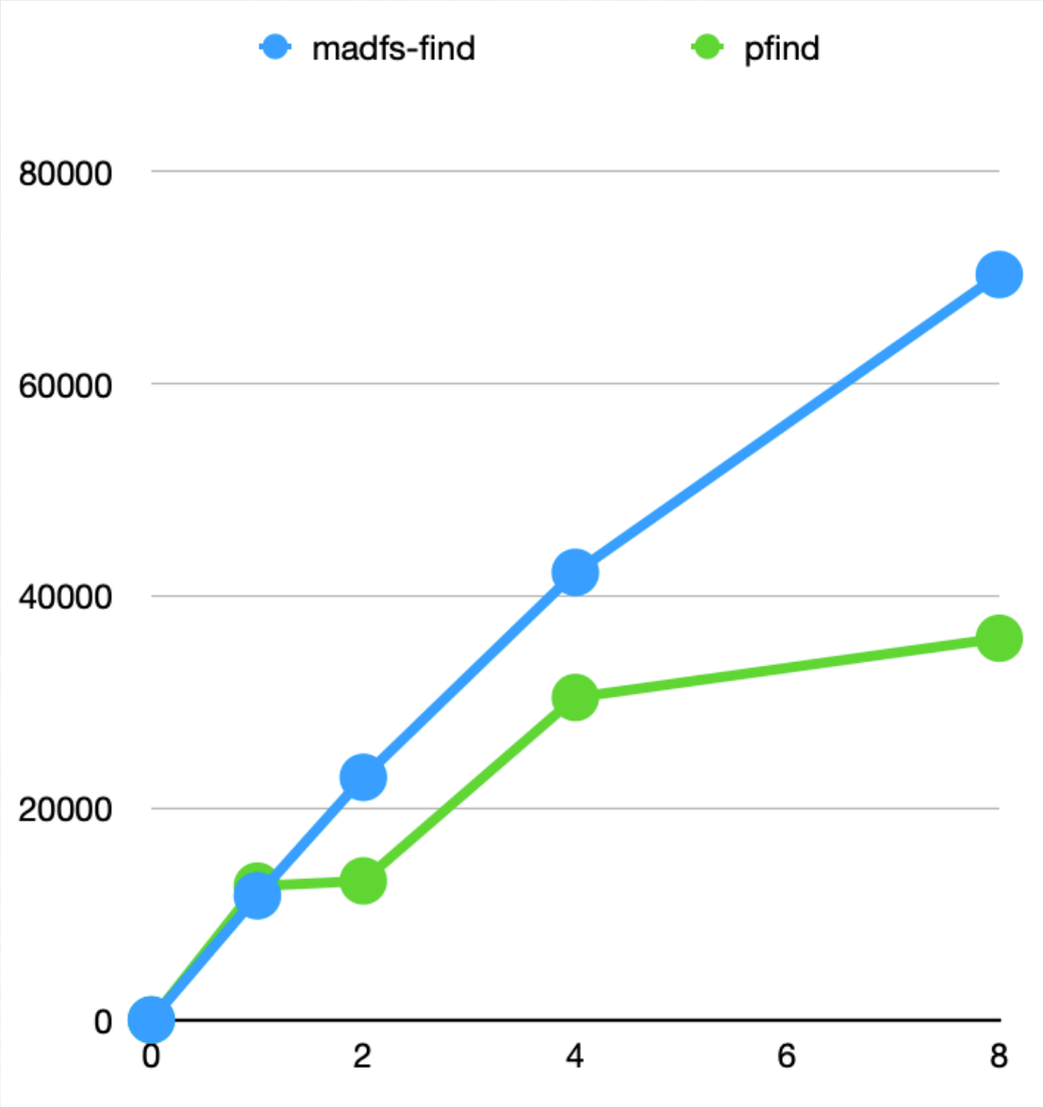
```
dir = opendir()
ioctl(dir.fd, MADFS_FIND_LIST, filter_json)
loop:
    dent = readdir(dir)
    print(dent.d_name)
closedir(dir)
```

```rust
#[derive(Debug, Serialize, Deserialize)]
struct FindFilter {
    regex: Option<String>,
    size: Option<u64>,
    ctime: Option<u64>,
}
```

```
ioctl(dir.fd, MADFS_FIND_COUNT,
      filter_json, &result)
```

```rust
/// Return info from MadFS find.
#[repr(C)]
#[derive(Debug)]
struct FindResult {
    found: u64,
    total: u64,
}
```

Madfs-find



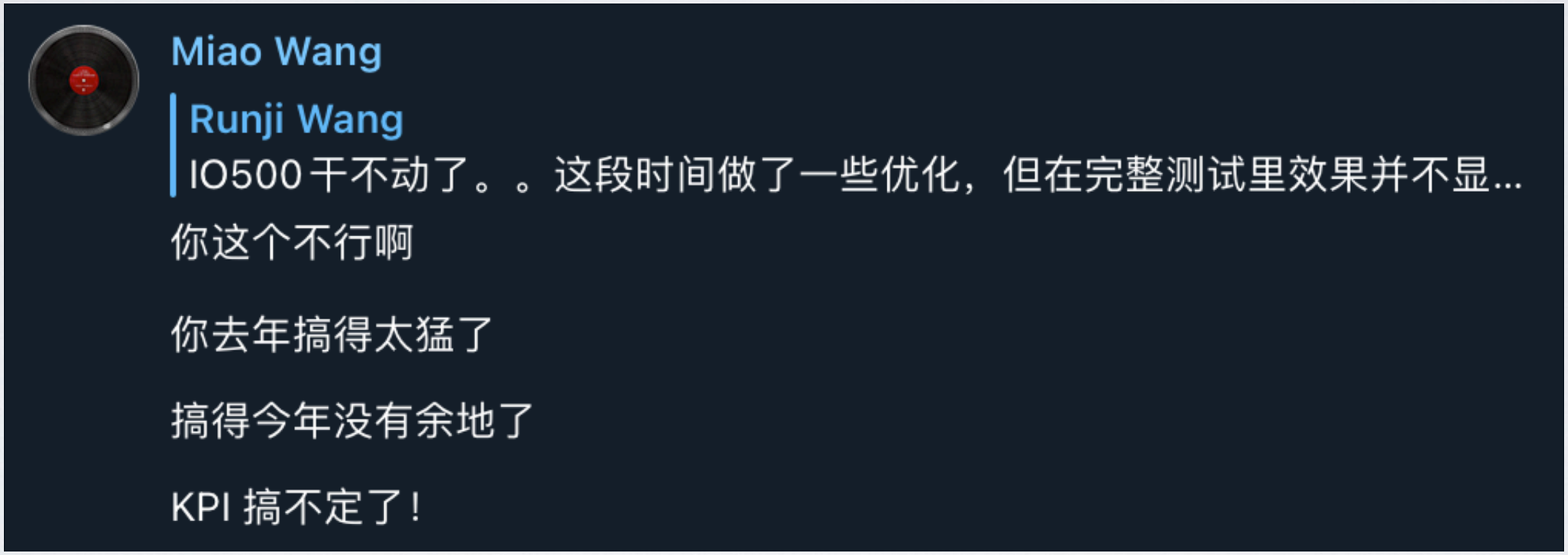
IO500 新测试

```
→  io500 git:(d2ca938) ✗ ./io500.sh config.ini
ERROR INVALID (src/phase_dbg.c:13) stonewall-time != 300s
IO500 version io500-isc21_v1 (extended)
ERROR INVALID (src/main.c:399) Runtime is smaller than expected minimum runtime
[RESULT]       ior-easy-write        8.414477 GiB/s : time 5.671 seconds [INVALID]
ERROR INVALID (src/main.c:399) Runtime is smaller than expected minimum runtime
[RESULT]        ior-rnd-write        0.838231 GiB/s : time 5.642 seconds [INVALID]
ERROR INVALID (src/main.c:399) Runtime is smaller than expected minimum runtime
[RESULT]    mdtest-easy-write      272.983070 kIOPS : time 5.492 seconds [INVALID]
[      ]   mdworkbench-create       69.112846 kIOPS : time 19.997 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.000 seconds
[RESULT]            find-easy    12534.029659 kIOPS : time 0.120 seconds
ERROR INVALID (src/main.c:399) Runtime is smaller than expected minimum runtime
[RESULT]       ior-hard-write        3.408759 GiB/s : time 6.101 seconds [INVALID]
ERROR INVALID (src/main.c:399) Runtime is smaller than expected minimum runtime
[RESULT]    mdtest-hard-write       42.897002 kIOPS : time 6.752 seconds [INVALID]
[RESULT]                 find     7684.994063 kIOPS : time 0.411 seconds
[RESULT]         ior-rnd-read        1.099620 GiB/s : time 4.301 seconds
[RESULT]            find-hard     2055.223596 kIOPS : time 0.141 seconds
[RESULT]    mdworkbench-bench      106.061921 kIOPS : time 102.489 seconds
[RESULT]        ior-easy-read       10.346266 GiB/s : time 4.612 seconds
[RESULT]     mdtest-easy-stat      475.898302 kIOPS : time 3.151 seconds
[RESULT]        ior-hard-read        6.834613 GiB/s : time 3.043 seconds
[RESULT]     mdtest-hard-stat      489.670866 kIOPS : time 0.592 seconds
[      ]   mdworkbench-delete       90.077682 kIOPS : time 15.380 seconds
[RESULT]   mdtest-easy-delete      346.467738 kIOPS : time 4.327 seconds
[RESULT]     mdtest-hard-read      175.410697 kIOPS : time 1.652 seconds
[RESULT]   mdtest-hard-delete      109.573658 kIOPS : time 2.644 seconds
[SCORE ] Bandwidth 6.710892 GiB/s : IOPS 329.709307 kiops : TOTAL 47.038745 [INVALID]
[SCOREX] Bandwidth 3.509844 GiB/s : IOPS 488.914177 kiops : TOTAL 41.424783 [INVALID]
```

- ior-rand
  4K随机读写
- find-easy/hard
  查找过滤条件
- mdworkbench
  更真实的mdtest
  侧重时延

Part3
打榜 II

坏耶


Miao Wang
Runji Wang
IO500干不动了。。这段时间做了一些优化，但在完整测试里效果并不显...
你这个不行啊
你去年搞得太猛了
搞得今年没有余地了
KPI 搞不定了！

11000
10,607
8250
7,043
5500
2750
0
10-31
12-01
05-30
31
2
3
4
7
8
9

内存泄露

+4MB/s

死循环

？？？

```
pub async fn next(&self) -> ConnectionRequest {
    poll_fn(|cx| {
        let mut incomings = self.incomings.borrow_mut();
        if let Some(connection) = incomings.items.pop_front() {
            Poll::Ready(connection)
        } else {
            incomings.wakers.push(cx.waker().clone());
            Poll::Pending
        }
    })
    .await
pub async fn next(&mut self) -> ConnectionRequest {
    self.recver.next().await.unwrap()
}
```

奇怪的bug产生了

262.0万播放 · 1128弹幕　2021-06-27 18:55:50

6 人正在看(全站20+人在看), 977 条弹幕　弹　发个友善的弹幕见证当下　弹幕礼仪 >　发送

13.0万　1.9万　5.9万　3.9万　稿件投诉　笔记

IOR 打溢出了

```
IO500 version io500-isc21_v1 (standard)
[RESULT]       ior-easy-write     1439.649245 GiB/s : time 420.304 seconds
[RESULT]    mdtest-easy-write    49604.997374 kIOPS : time 329.536 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.000 seconds
[RESULT]       ior-hard-write     1031.377546 GiB/s : time 347.253 seconds
[RESULT]    mdtest-hard-write    18188.945420 kIOPS : time 346.323 seconds
[RESULT]                 find  1392022.058925 kIOPS : time 16.267 seconds
[RESULT]        ior-easy-read     1307.214084 GiB/s : time 462.863 seconds
[RESULT]     mdtest-easy-stat   103217.687526 kIOPS : time 158.385 seconds
ERROR: read(3505597, 0x2d440000, 47008) returned EOF prematurely, (aiori-POSIX.c:639)
--------------------------------------------------------------------------
MPI_ABORT was invoked on rank 7095 in communicator MPI_COMM_WORLD
with errorcode -1.

NOTE: invoking MPI_ABORT causes Open MPI to kill all MPI processes.
You may or may not see output from other processes, depending on
exactly when Open MPI kills them.
--------------------------------------------------------------------------
```

```
if (test->filePerProc) {
  offset += i * test->blockSize;
} else {
  offset += (i * test->numTasks * test->blockSize)
}
```

64 32 32 64

```
        int hitStonewall;
-       int i, j;
+       IOR_offset_t i, j;
        IOR_point_t *point =
```

11000
8250
5500
2750
0
10,607
9,683
7,043
修复内存
泄露
10-31
12-01
05-30
31
2
3
4
7
8
9

准备出差

陈康

润基哥在机场，正在犹豫是不是要去深圳。

陈康

范高俊

先不用过来了

范高俊

现在没有疫苗不允许进实验室（领导、院士一视同仁）

陈康

润基哥已经完成两针疫苗。。。。。

陈康

@王润基 远程可以吗？远程可以的话先跑路回来？

2021 年 5 月 31 日 14:39

远程应该可以，不过还是过去吧，已经在路上了😂

陈康

让飞机飞回来！

陈康

让飞机飞回来！

陈康

4点钟的飞机的话，润基还有一段时间到深圳。

2021年5月31日 19:17

短信
今天 19:15

【长沙市文旅广电局】欢迎您来到山水洲城-快乐长沙！夜间到白天多云转小雨,21.9°C-32.5°C。

飞机真的飞回来了。。。

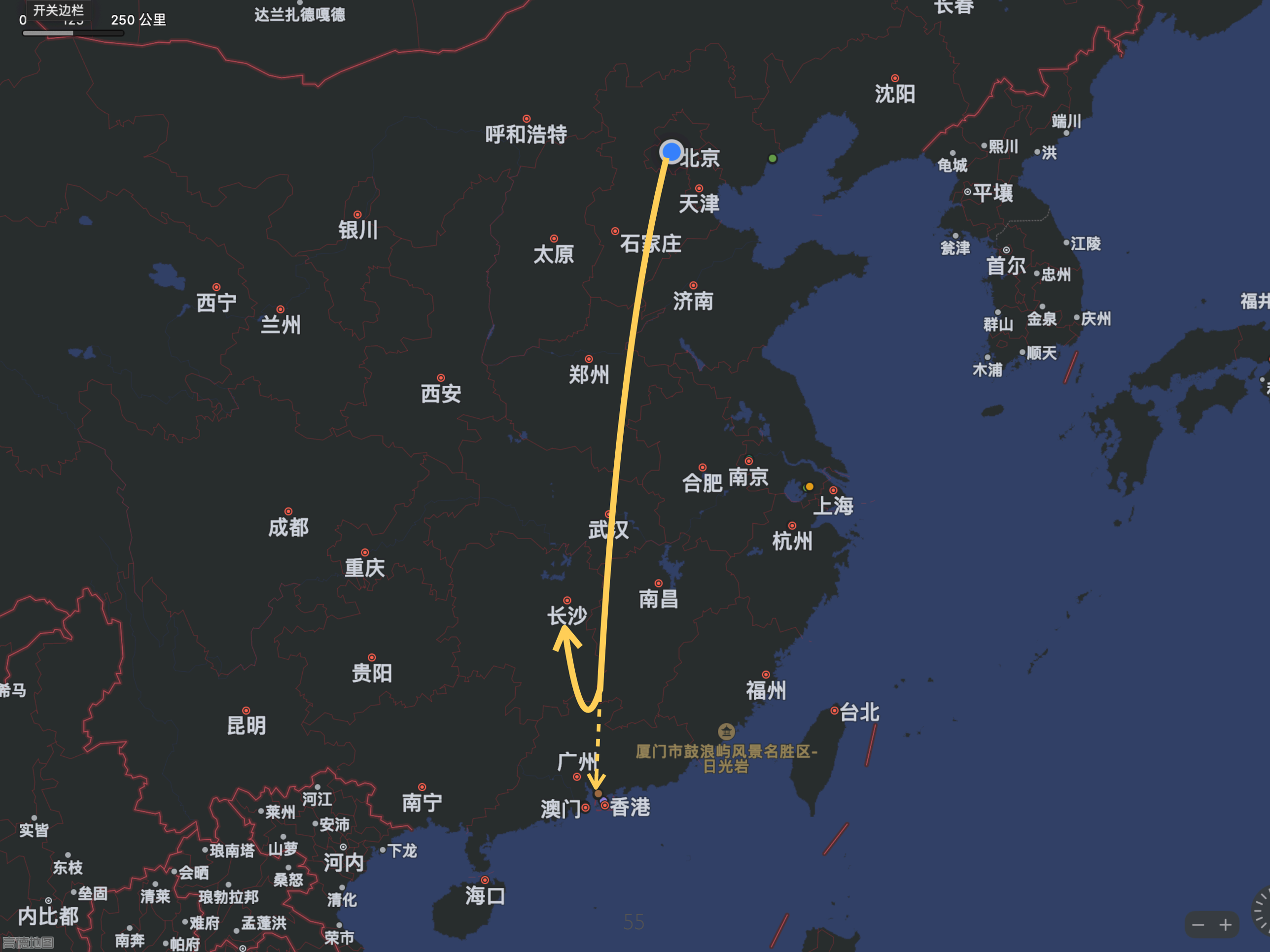
@陈康

陈康

啊?

刘永恒

深圳这会大暴雨，我比润基早走两小时就已经到了


开关边栏
0
250 公里
达兰扎德嘎德
长春
沈阳
端川
熙川
洪
龟城
平壤
呼和浩特
北京
天津
银川
太原
石家庄
瓮津
江陵
首尔
忠州
西宁
兰州
济南
群山
金泉
庆州
顺天
木浦
郑州
西安
合肥
南京
上海
成都
武汉
杭州
重庆
南昌
长沙
贵阳
福州
台北
昆明
厦门市鼓浪屿风景名胜区-日光岩
广州
南宁
澳门
香港
河江
莱州
安沛
实皆
琅南塔
山萝
下龙
河内
东枝
会晒
桑怒
垒固
清莱
琅勃拉邦
清化
海口
内比都
难府
孟蓬洪
南奔
帕府
荣市
高德地图

Boarding closes 15 minutes before departure

CZ3130Notice of delay

The flight has been delayed undefined,the departure time of Changsha is to be determined

【长沙市文旅广电局】欢迎您来到山水洲城-快乐长沙！夜间到白天多云转小雨,21.9°C-32.5°C。

湖南大学

茶颜悦色
Modern China Tea Shop
茶颜悦色
小主快来!
茶颜充值升级啦!
充100送10;充200送30;
加入会员储值消费还可以积点哦!
vivo
HONOR

嶽麓書院
於斯為盛
惟楚有材

圣地巡礼

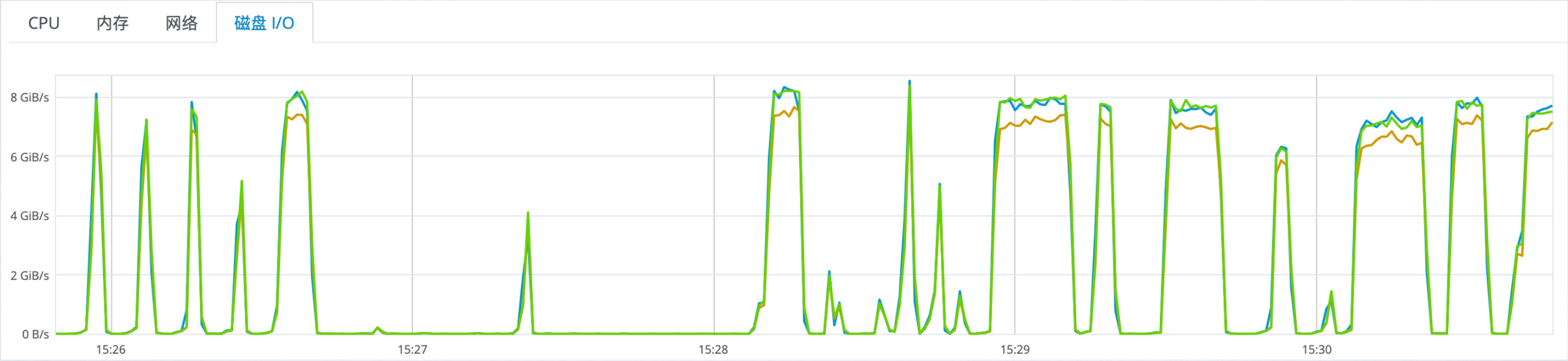
SEE YOU AGAIN

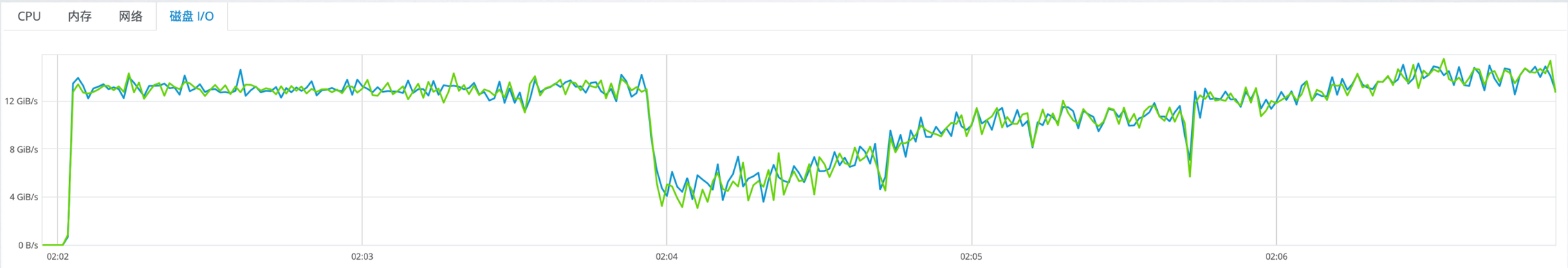
20000
15000
10000
5000
0
17,435
14,018
10,607
9,683
7,043
增大规模
10-31
12-01
05-30
31
2
3
4
7
8
9

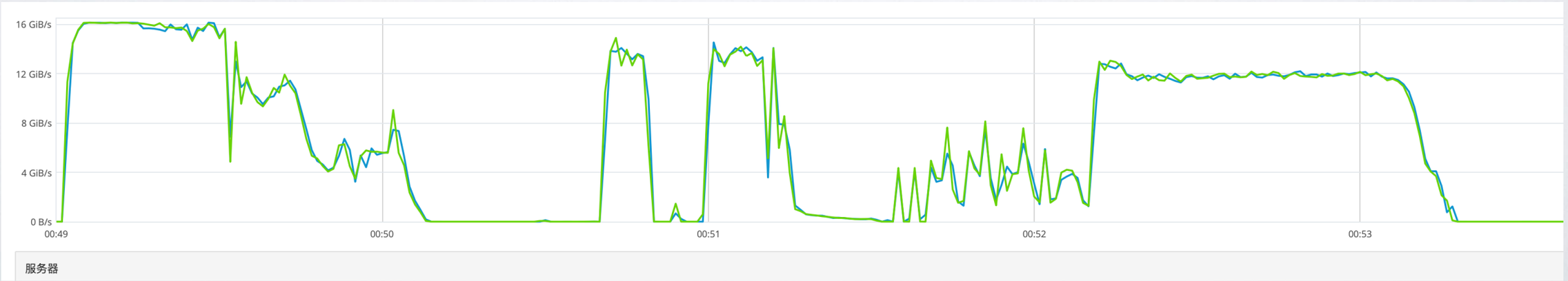
长尾问题

ior-easy-write
磁盘IO
理论带宽
实际计分带宽
8 GiB/s
6 GiB/s
4 GiB/s
2 GiB/s
0 B/s
15:39
15:40
15:41
15:42
Begin
Stonewall @ 30s
End

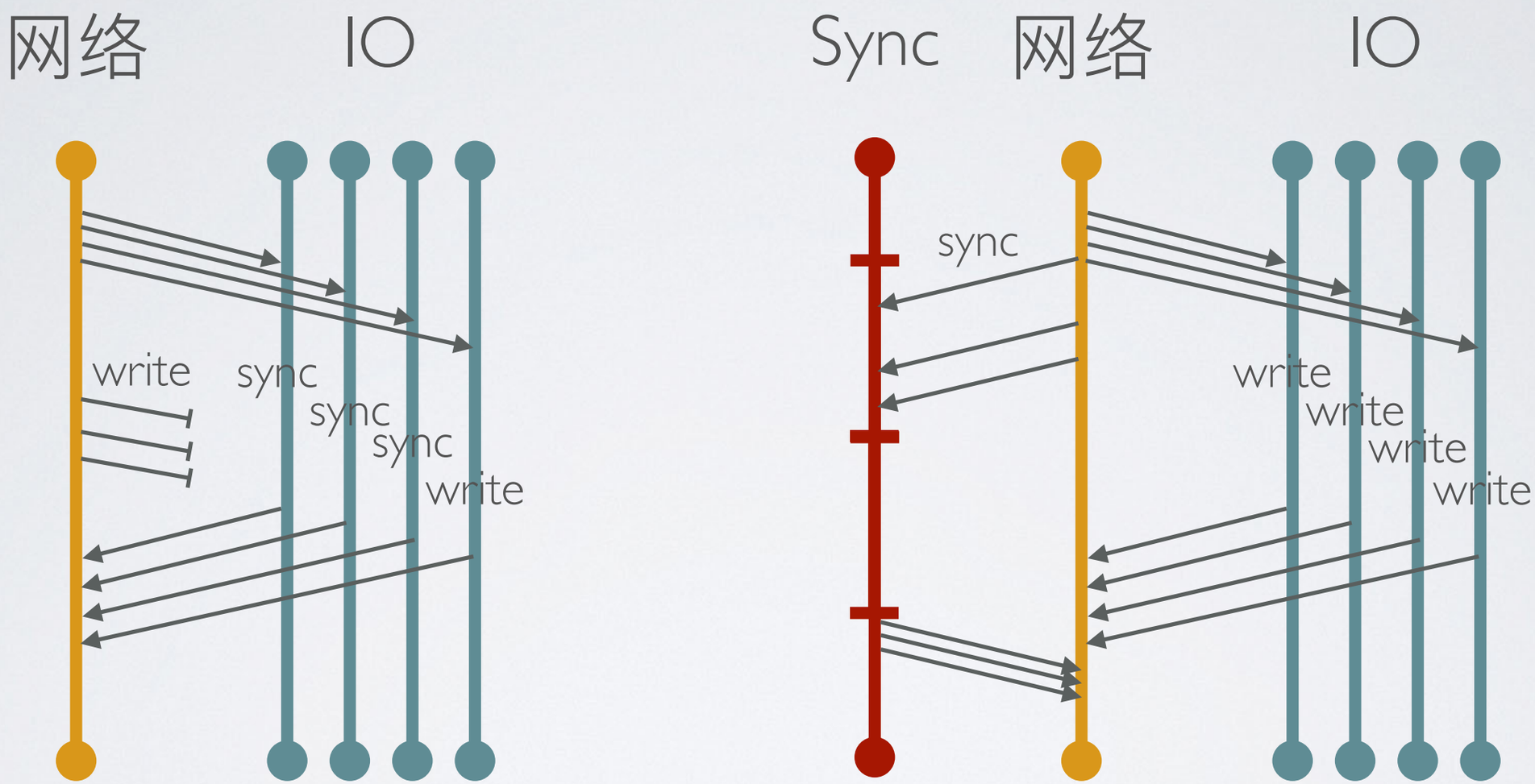
群魔乱舞


CPU
内存
网络
磁盘 I/O
8 GiB/s
6 GiB/s
4 GiB/s
2 GiB/s
0 B/s
15:26
15:27
15:28
15:29
15:30


CPU
内存
网络
磁盘 I/O
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
02:02
02:03
02:04
02:05
02:06


16 GiB/s
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
00:49
00:50
00:51
00:52
00:53
服务器

特殊处理 sync



优化前　　优化后

特殊处理 sync

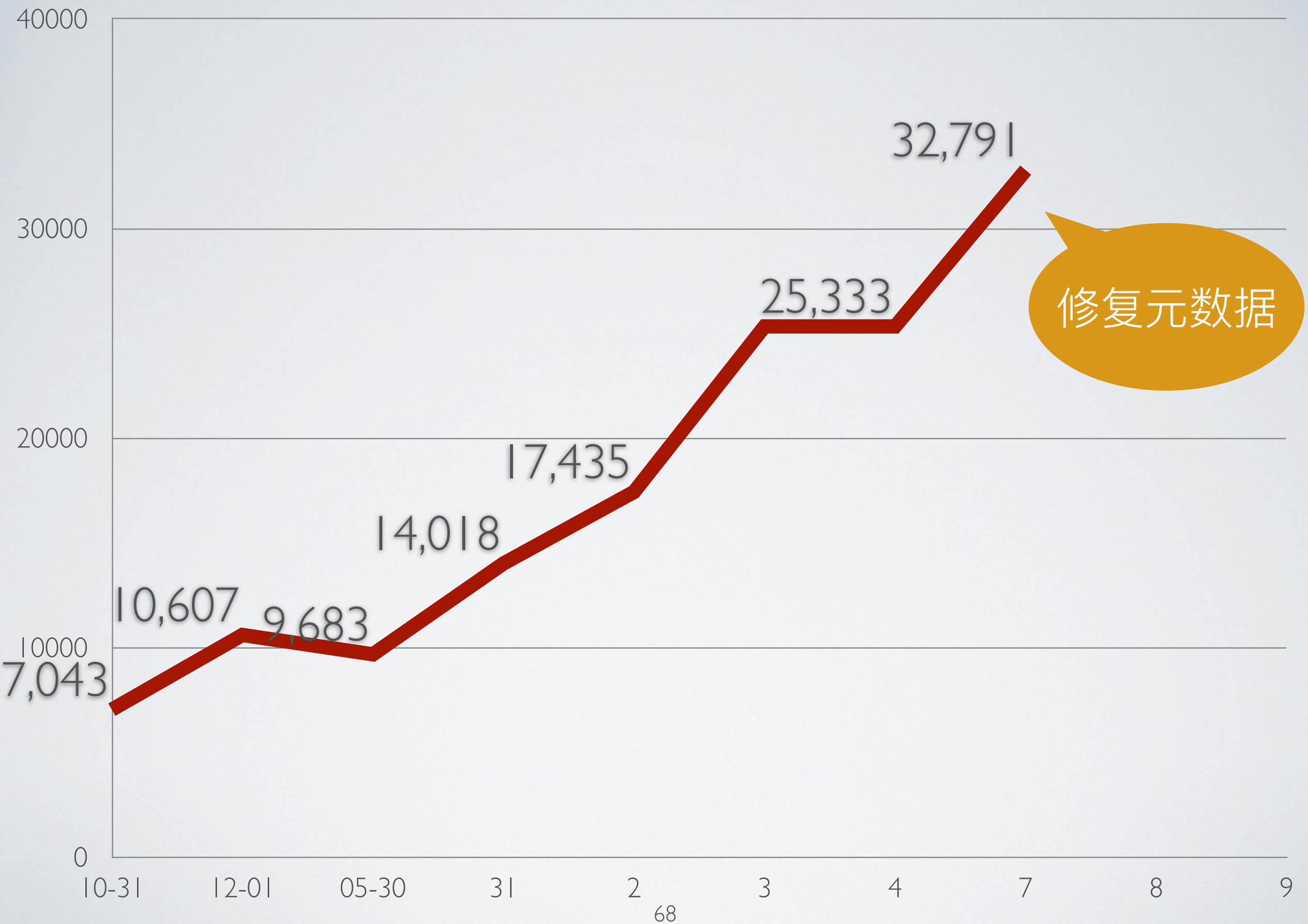
磁盘IO
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
12:08
12:09
12:10
12:11
12:12
Stonewall @ 300s
End

26000
19500
13000
6500
0
25,333
17,435
14,018
10,607
9,683
7,043
修复fsync
10-31
12-01
05-30
31
2
3
4
7
8
9

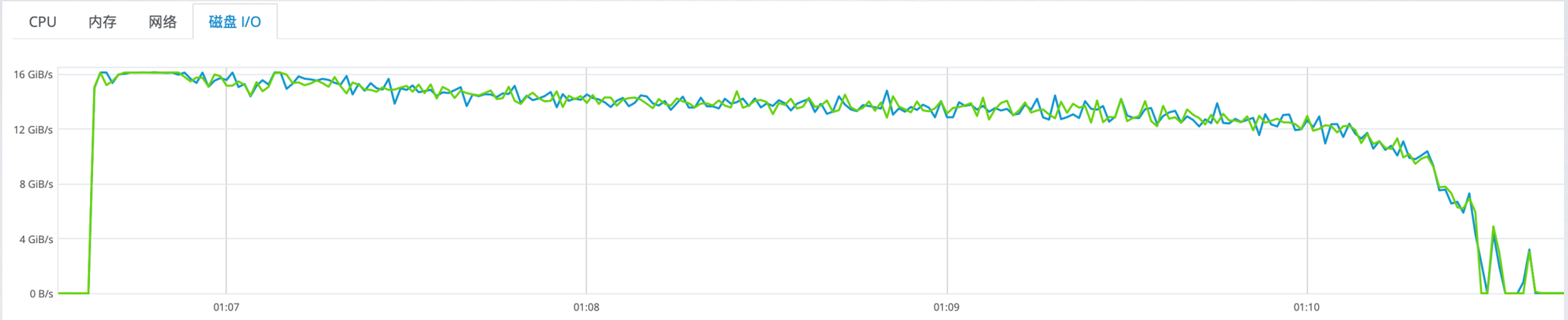
元数据性能回归

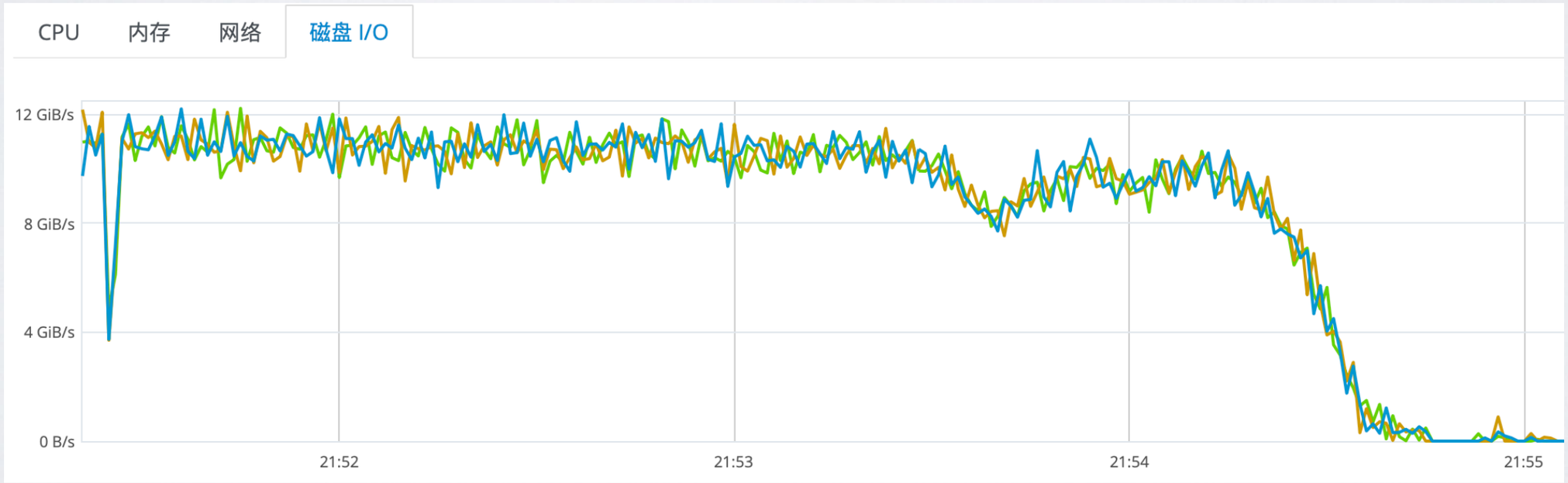
- 开了过多协程同时执行 tag_recv
- 导致 tag match 开销大增

```
+   39.25%     madfs-daemon
-   29.30%     libucp.so.0.0.0
  +   27.20%       [.] ucp_tag_exp_search_all
  +    1.14%       [.] ucp_worker_progress
       0.18%       [.] ucp_tag_send_nbx
       0.13%       [.] ucp_dt_iov_gather
       0.12%       [.] ucp_dt_pack
       0.10%       [.] ucp_tag_recv_nbx
       0.09%       [.] ucp_eager_only_handler
       0.08%       [.] ucp_tag_eager_bcopy_single
       0.08%       [.] ucp_tag_send_nb
       0.06%       [.] ucp_tag_recv_request_test
       0.05%       [.] ucp_dt_iov_gather@plt
       0.05%       [.] memcpy@plt
       0.04%       [.] ucp_tag_send_nbx@plt
+    8.89%     libc-2.28.so
+    8.37%     libhns-rdmav17.so
+    5.35%     libuct_ib.so.0.0.0
+    4.45%     libpthread-2.28.so
+    3.22%     [kernel]
+    0.63%     libucs.so.0.0.0
+    0.54%     libstdc++.so.6.0.24
```

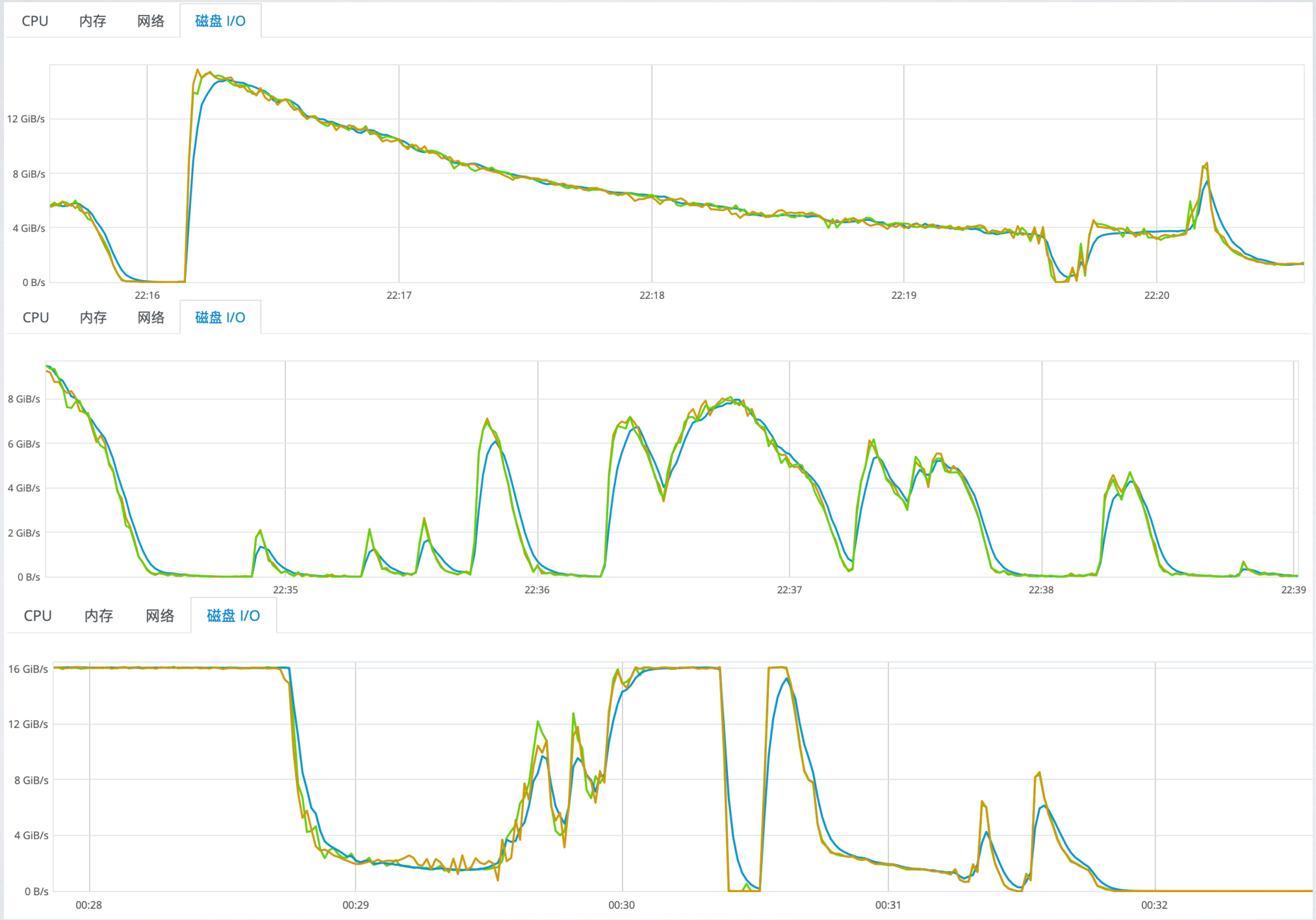

40000
30000
20000
10000
0
7,043
10,607
9,683
14,018
17,435
25,333
32,791
修复元数据
10-31
12-01
05-30
31
2
3
4
7
8
9

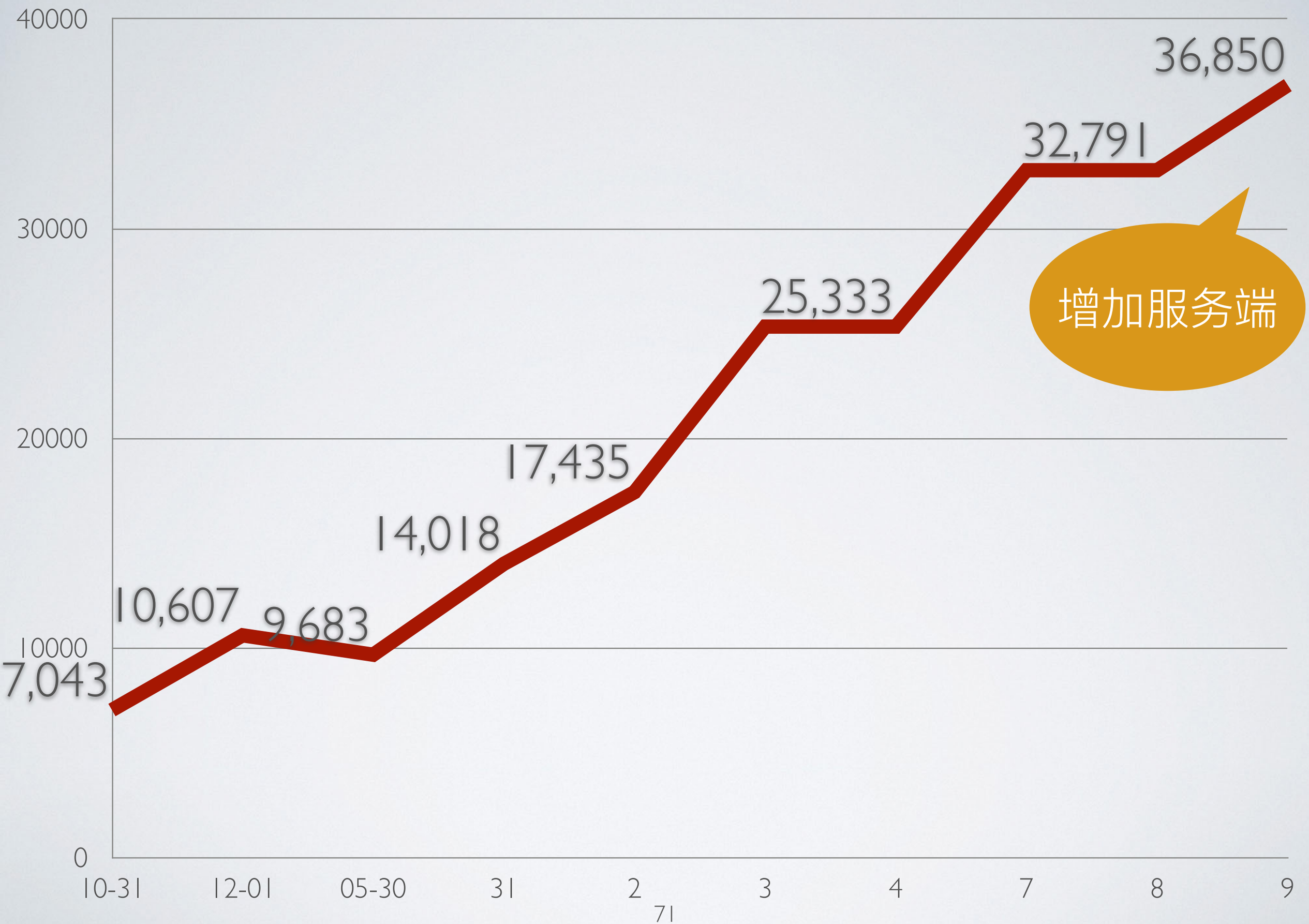
科学的带宽


CPU
内存
网络
磁盘 I/O
16 GiB/s
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
01:07
01:08
01:09
01:10


CPU
内存
网络
磁盘 I/O
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
21:52
21:53
21:54
21:55

不科学的带宽


CPU
内存
网络
磁盘 I/O
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
22:16
22:17
22:18
22:19
22:20
CPU
内存
网络
磁盘 I/O
8 GiB/s
6 GiB/s
4 GiB/s
2 GiB/s
0 B/s
22:35
22:36
22:37
22:38
22:39
CPU
内存
网络
磁盘 I/O
16 GiB/s
12 GiB/s
8 GiB/s
4 GiB/s
0 B/s
00:28
00:29
00:30
00:31
00:32


40000
30000
20000
10000
0
7,043
10,607
9,683
14,018
17,435
25,333
32,791
36,850
增加服务端
10-31
12-01
05-30
31
2
3
4
7
8
9

最终提交结果（全系统）

```
IO500 version io500-isc21_v2 (standard)
[RESULT]       ior-easy-write     3360.298674 GiB/s : time 673.312 seconds
[RESULT]    mdtest-easy-write   278101.309437 kIOPS : time 318.025 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.000 seconds
[RESULT]       ior-hard-write     3716.573255 GiB/s : time 448.352 seconds
[RESULT]    mdtest-hard-write   102419.383006 kIOPS : time 349.802 seconds
[RESULT]                 find  8991786.331160 kIOPS : time 13.770 seconds
[RESULT]        ior-easy-read     2803.716632 GiB/s : time 806.952 seconds
[RESULT]     mdtest-easy-stat   529803.484430 kIOPS : time 167.480 seconds
[RESULT]        ior-hard-read     3914.484508 GiB/s : time 415.836 seconds
[RESULT]     mdtest-hard-stat   410069.944303 kIOPS : time 88.234 seconds
[RESULT]   mdtest-easy-delete   287724.351211 kIOPS : time 307.527 seconds
[RESULT]     mdtest-hard-read   209906.918749 kIOPS : time 171.290 seconds
[RESULT]   mdtest-hard-delete   183148.570395 kIOPS : time 196.161 seconds
[SCORE ] Bandwidth 3421.624277 GiB/s : IOPS 396872.816141 kiops : TOTAL 36850.368552
```

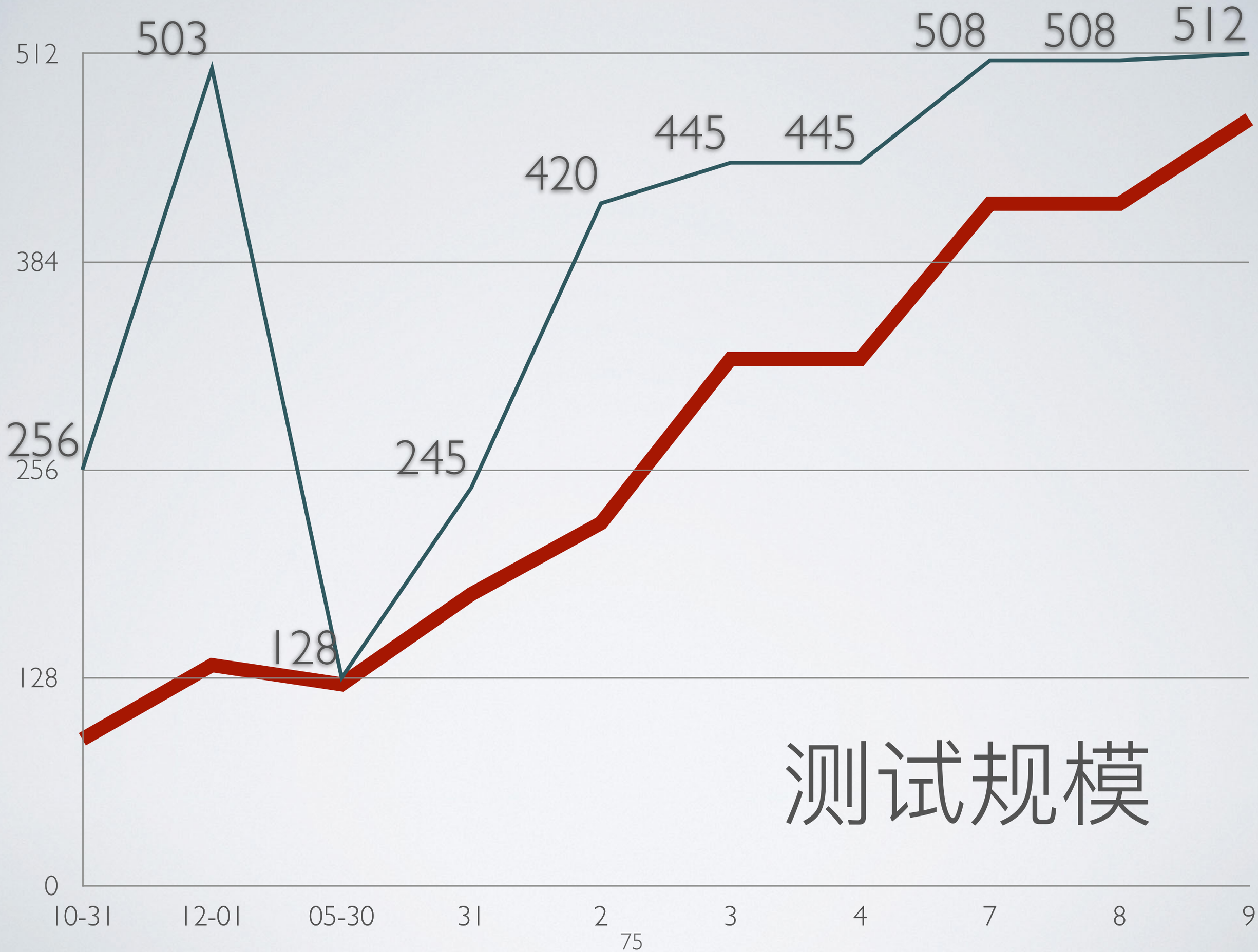
- 客户端：512 节点 × 72 进程 = 36864 进程
- 服务端：512 节点 × 24 进程 = 12288 进程

最终提交结果（10节点）

```
IO500 version io500-isc21_v2 (standard)
[RESULT]       ior-easy-write      222.987510 GiB/s : time 305.831 seconds
[RESULT]    mdtest-easy-write    19176.710145 kIOPS : time 333.678 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.000 seconds
[RESULT]       ior-hard-write      197.272000 GiB/s : time 318.301 seconds
[RESULT]    mdtest-hard-write     8494.136829 kIOPS : time 327.979 seconds
[RESULT]                 find  1584644.889026 kIOPS : time 5.778 seconds
[RESULT]        ior-easy-read      149.939929 GiB/s : time 455.657 seconds
[RESULT]     mdtest-easy-stat    36333.268045 kIOPS : time 176.582 seconds
[RESULT]        ior-hard-read      213.722239 GiB/s : time 293.621 seconds
[RESULT]     mdtest-hard-stat    35438.079693 kIOPS : time 79.381 seconds
[RESULT]   mdtest-easy-delete    26920.996932 kIOPS : time 237.966 seconds
[RESULT]     mdtest-hard-read    15692.363037 kIOPS : time 177.994 seconds
[RESULT]   mdtest-hard-delete    15239.971832 kIOPS : time 183.249 seconds
[SCORE ] Bandwidth 193.766302 GiB/s : IOPS 34777.267068 kiops : TOTAL 2595.893380
```

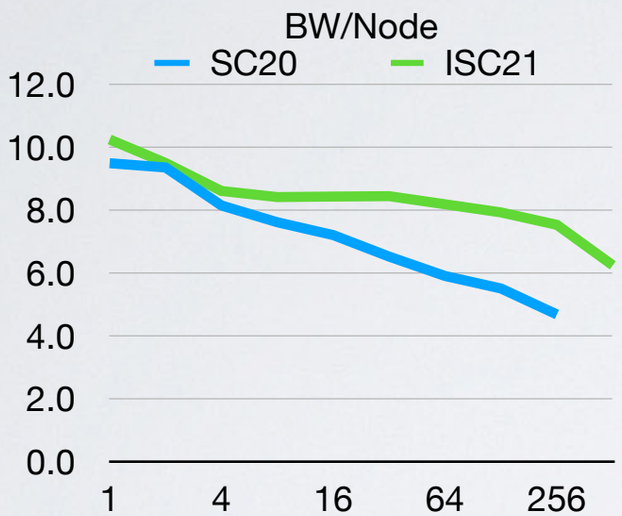
- 客户端：10 节点 × 180 进程 = 1800 进程
- 服务端：50 节点 × 24 进程 = 1200 进程

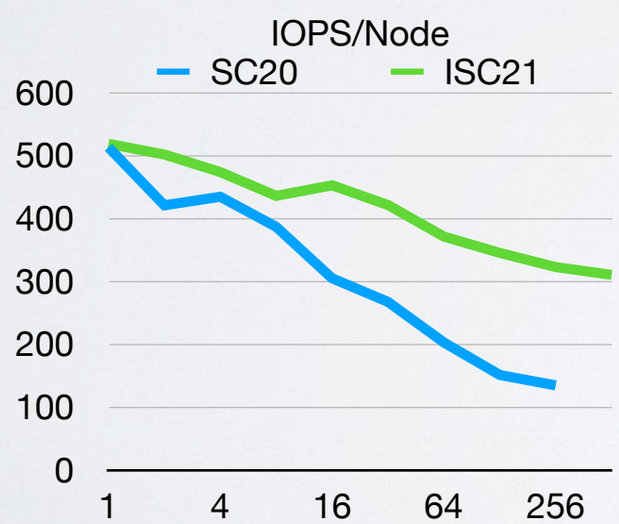
40000
30000
20000
10000
0
7,043
10,607
9,683
14,018
17,435
25,333
32,791
36,850
增大规模
修复内存
泄露
修复fsync
修复元数据
增加服务端
10-31
12-01
05-30
31
2
3
4
7
8
9

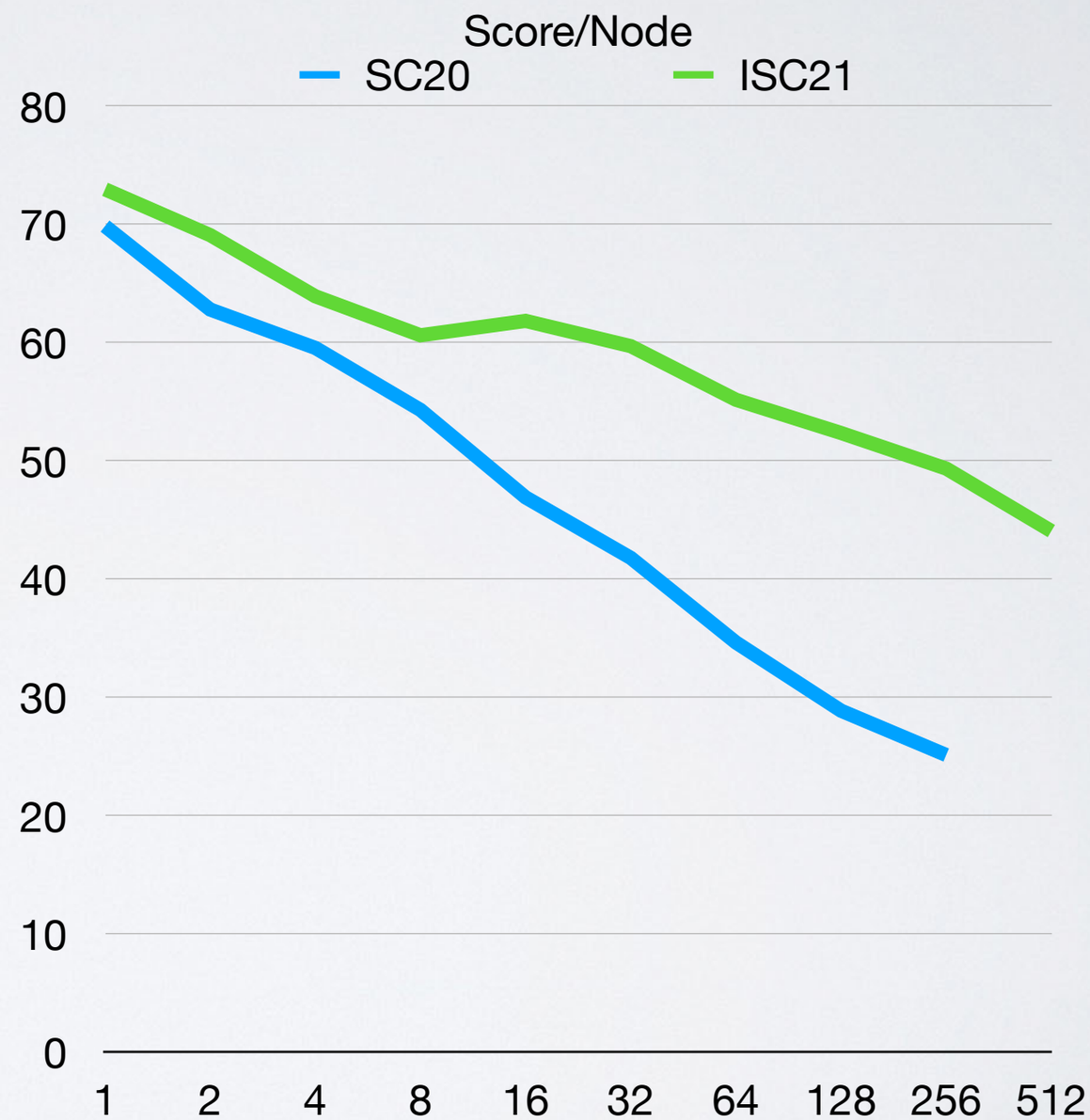

512
384
256
128
0
503
256
245
128
420
445
445
508
508
512
测试规模
10-31
12-01
05-30
31
2
3
4
7
8
9

40000
30000
20000
10000
0
213
253
778
2,202
4,698
7,043
9,683
14,018
17,435
25,333
32,791
36,850
21
23
25
27
29
31
05-30
2
4
8

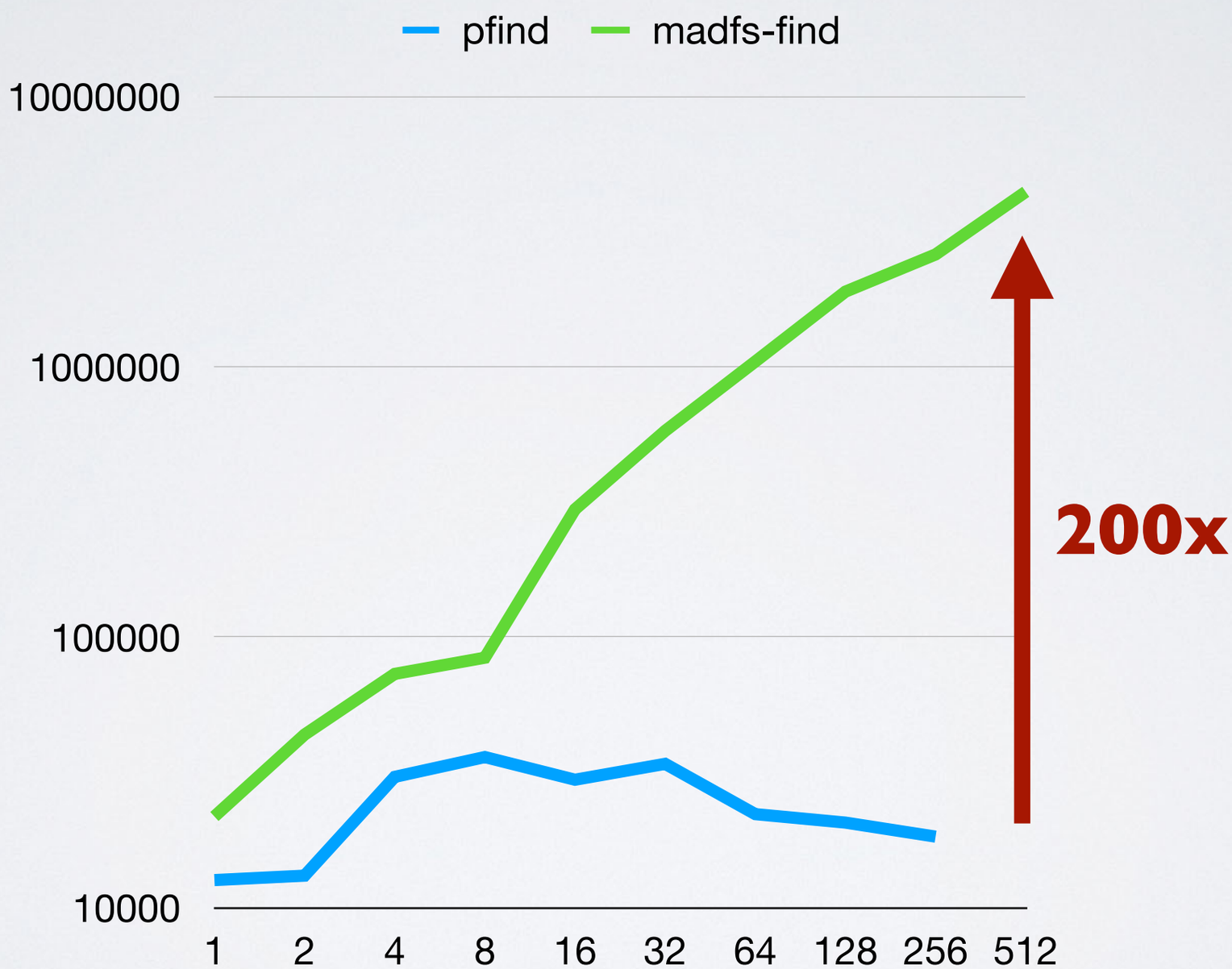
扩展性提升


BW/Node
SC20
ISC21
12.0
10.0
8.0
6.0
4.0
2.0
0.0
1
4
16
64
256


IOPS/Node
SC20
ISC21
600
500
400
300
200
100
0
1
4
16
64
256


Score/Node
SC20
ISC21
80
70
60
50
40
30
20
10
0
1
2
4
8
16
32
64
128
256
512

Madfs-Find


pfind
madfs-find
10000000
1000000
100000
10000
1
2
4
8
16
32
64
128
256
512
200x

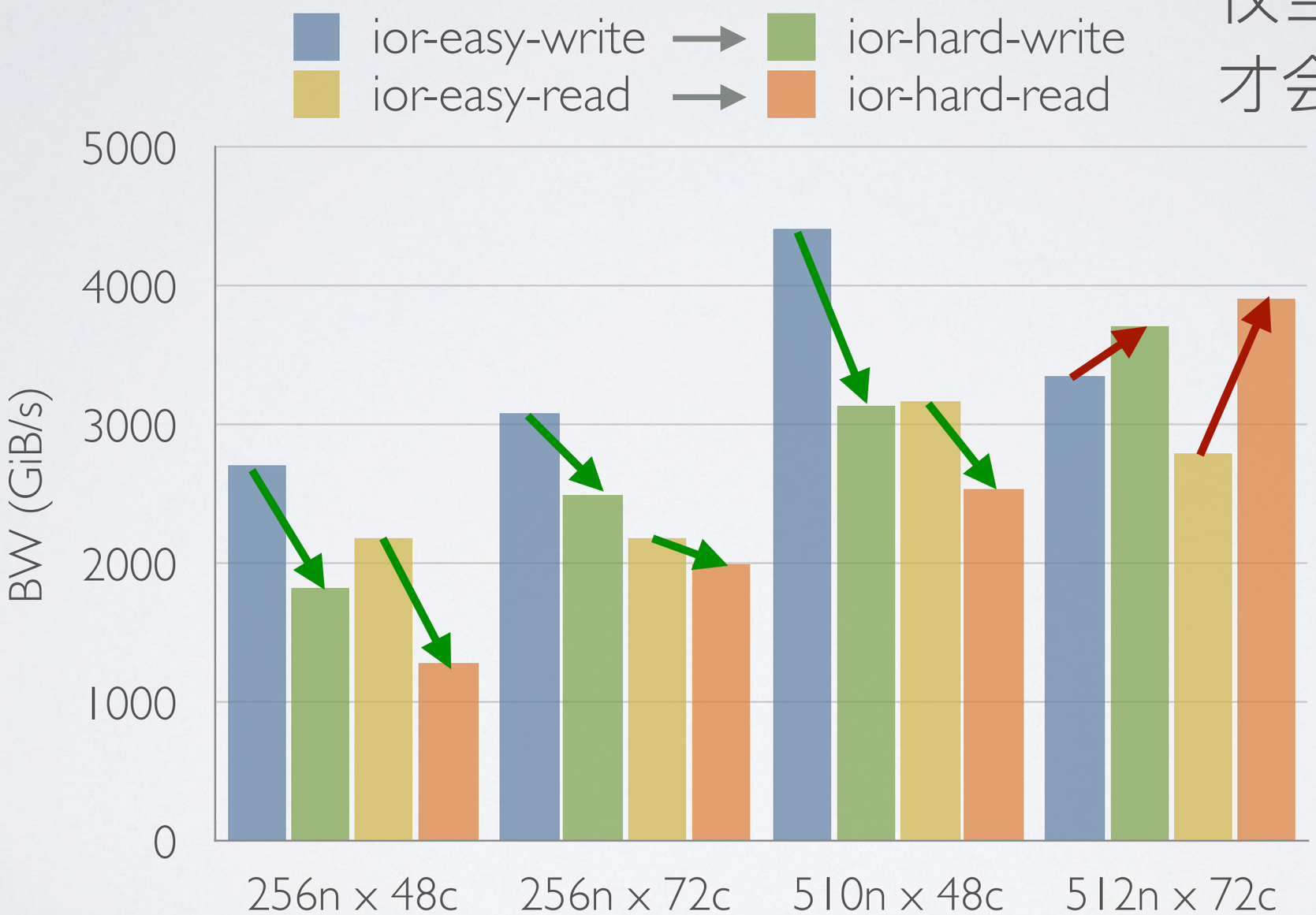
Q&A

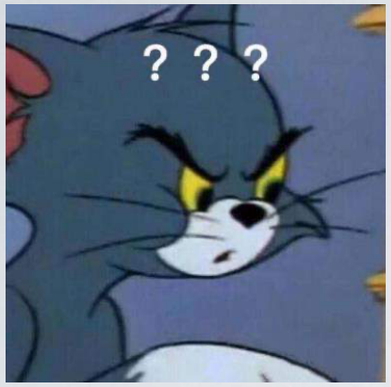
- 为什么 IOR easy 比 hard 性能低?
  (transfer_size: easy=4700800, hard=47008)

```
IO500 version io500-isc21 v2 (standard)
[RESULT]       ior-easy-write     3360.298674 GiB/s : time 673.312 seconds
[RESULT]    mdtest-easy-write   278101.309437 kIOPS : time 318.025 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.000 seconds
[RESULT]       ior-hard-write     3716.573255 GiB/s : time 448.352 seconds
[RESULT]    mdtest-hard-write   102419.383006 kIOPS : time 349.802 seconds
[RESULT]                 find  8991786.331160 kIOPS : time 13.770 seconds
[RESULT]        ior-easy-read     2803.716632 GiB/s : time 806.952 seconds
[RESULT]     mdtest-easy-stat   529803.484430 kIOPS : time 167.480 seconds
[RESULT]        ior-hard-read     3914.484508 GiB/s : time 415.836 seconds
[RESULT]     mdtest-hard-stat   410069.944303 kIOPS : time 88.234 seconds
[RESULT]   mdtest-easy-delete   287724.351211 kIOPS : time 307.527 seconds
[RESULT]     mdtest-hard-read   209906.918749 kIOPS : time 171.290 seconds
[RESULT]   mdtest-hard-delete   183148.570395 kIOPS : time 196.161 seconds
[SCORE ] Bandwidth 3421.624277 GiB/s : IOPS 396872.816141 kiops : TOTAL 36850.368552
```

Easy < hard?

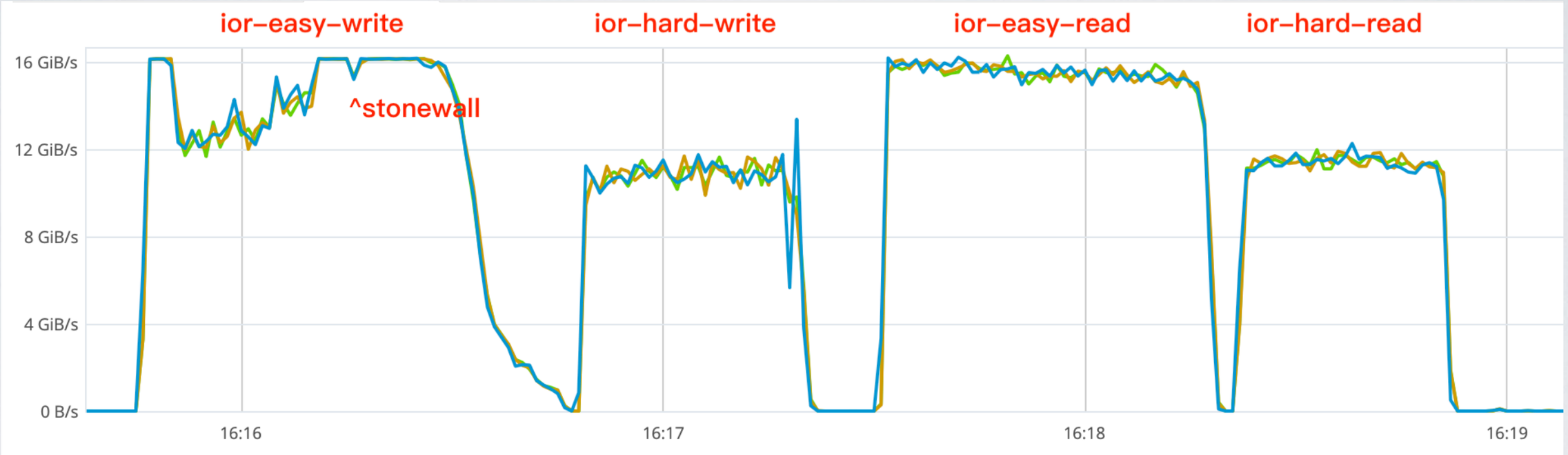
仅当规模很大时
才会出现的现象



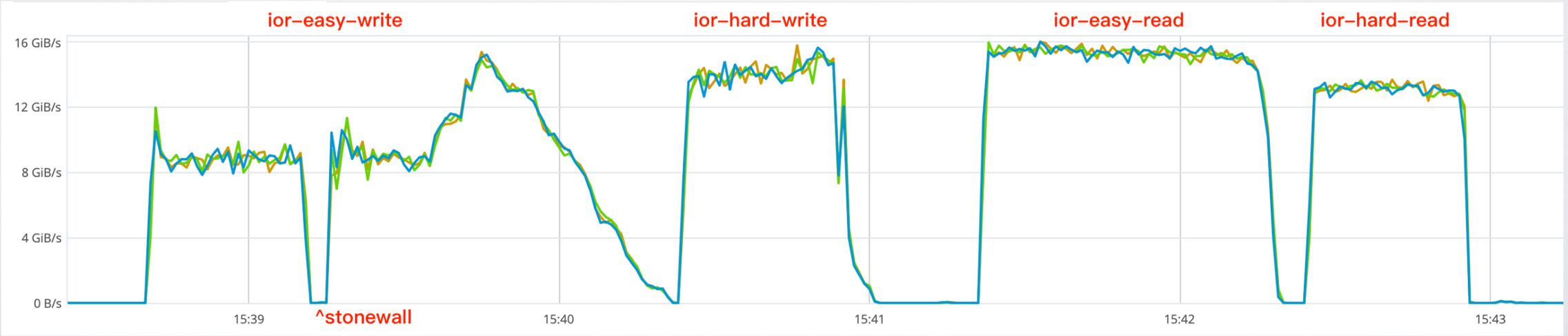


更多实验

300 nodes x 72 clients

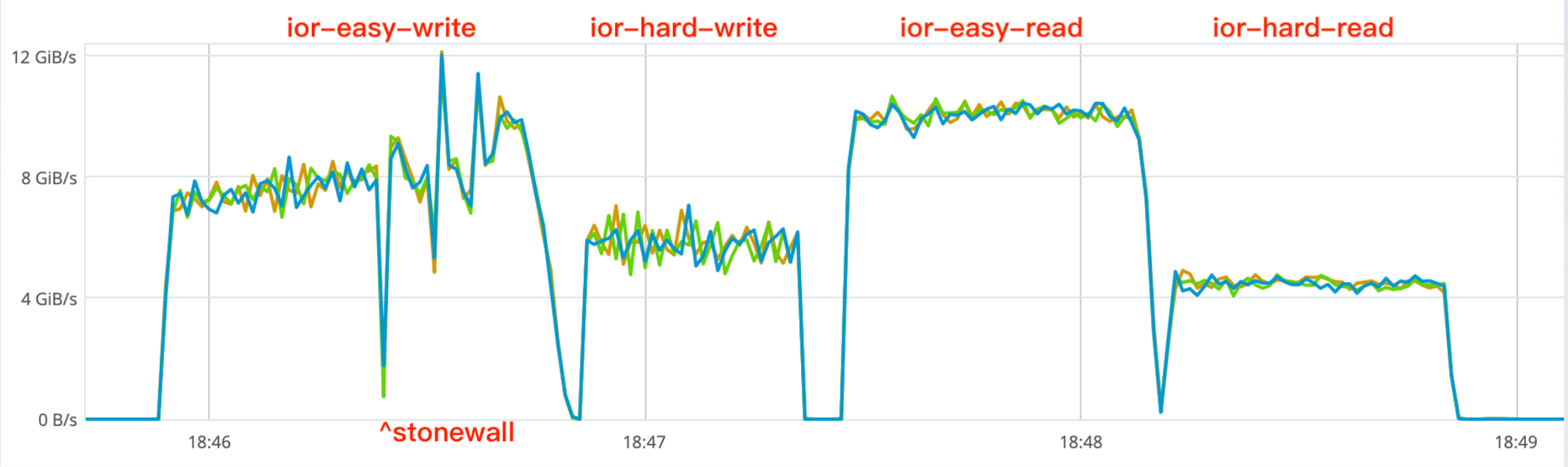


349 nodes x 72 clients

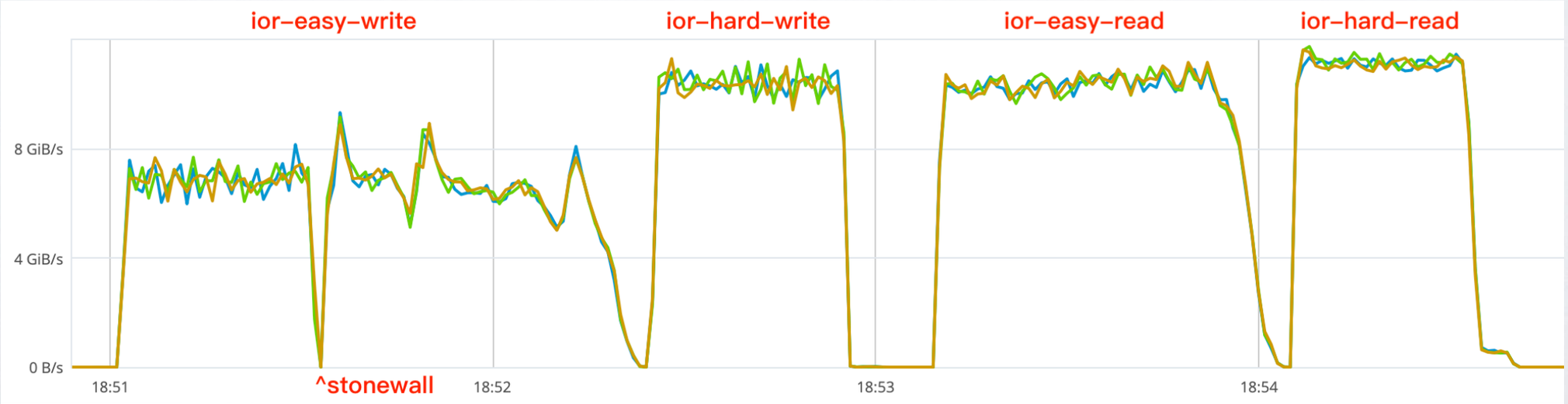


更多实验

503 nodes x **24** clients



503 nodes x **48** clients



時間就是金錢
效率就是生命
禁止攀爬
No Climbing

东
留新一路
西
Liuxin 1st Rd
NAYUKI
雪茶
奈の

北大

清華

CZ-2F

The 8th IO500 and the Virtual Institute of I/O
Andreas Dilger, Dean Hildebrand, Julian M. Kunkel, Jay Lofstead, George Markomanolis
IO500
vi4io
The IO-500 and the Virtual Institute of I/O
Chat
Questions
Polls
Kira Duwe
yes :-)
Jan-Patrick Lehr
Hi everybody
Georgios Markomanolis
Hello
Frank Godbon
Hi :)
Julian Kunkel
hi
Rui Feng
Question
When will the new IO500 list be published on the website?
Julian Kunkel
today
Adrian Jackson
We can hear you Jay
Write a message...
PA
+ 48
You

Chrome 文件 编辑 视图 历史记录 书签 个人资料 标签页 窗口 帮助

(162) IPFS: Interplanetary file s | 清华大学信息门户 | 清华大学电子邮件系统 | Attend networking events a | IO500 - Birds-of-a-Feather - IS

app.swapcard.com/event/isc-high-performance-2021-digital/planning/UGxhbm5pbmdfNDQ0Njcz

应用 Bookmarks 快科技(原驱动之家... Slashdot (16) 清华大学校工会 表格检索_SooPAT... 推酷 - IT人专属的... Coq'Art Home page Hadoop installatio... 一网通用户登录 Index of /pub/glus... TLA

00:11:21

Award Ceremony

IO500

TI BM LL KP FC MM AJ HC + 57

KC You

PA

Six Awards

- Full List
  - Bandwidth
  - Metadata
  - Overall
- 10-Node Challenge List
  - Bandwidth
  - Metadata
  - Overall

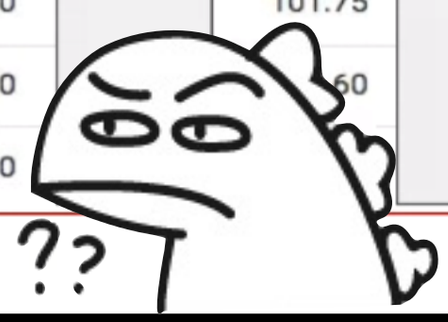
10 node challenge - Bandwidth Winner

10 Node ISC21 List

IO500 | 10 Node

Sorted by BW

#	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
			INFORMATION						IO500	
1	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	
2	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	10	1,800		193.77	
3	SC20	Forschungszentrum Juelich (FZJ)	JUWELS	HPEDDN	IME	10	400		178.11	
4	SC20	TACC	Frontera	DDN	IME	10	280		176.23	
5	ISC20	Intel	Wolf	Intel	DAOS	10	420		164.77	
6	ISC21	Supermicro		Supermicro	DAOS	10	1,120		112.17	
7	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960		105.28	
8	SC19	National Supercomputing Centre, Singapore	Aspire 1	DDN	IME	10	160		101.75	
9	SC20	DDN	DIME	DDN	IME	10	110		60	
10	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380			



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Intel (Endeavour)
#1 in the 10 Node Challenge BW Score

IO500

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/ten

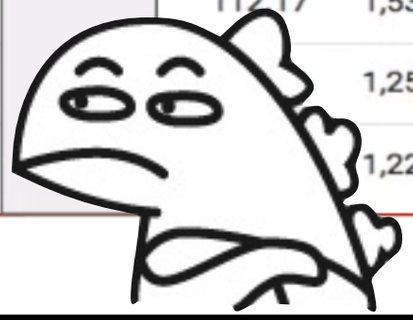
10-Node Challenge - Metadata Winner

10 Node ISC21 List

IO500 | 10 Node

Sorted by MD

#	INFORMATION								IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW (GIB/S)	MD ↑ (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	10	1,800		193.77	34,777.27
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	8,671.65
3	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960		105.28	3,567.85
4	ISC20	Intel	Wolf	Intel	DAOS	10	420		164.77	3,493.56
5	ISC20	TACC	Frontera	Intel	DAOS	10	420		79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	10	480		91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380		95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120		112.17	1,535.63
9	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800			1,251.32
10	SC20	Johannes Gutenberg University Mainz	MOGON II	JGU (ADA-FS)& BSC (NEXTGenIO)	GekkoFS	10	240			1,223.59



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the 10 Node Challenge MD Score

IO500

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/ten

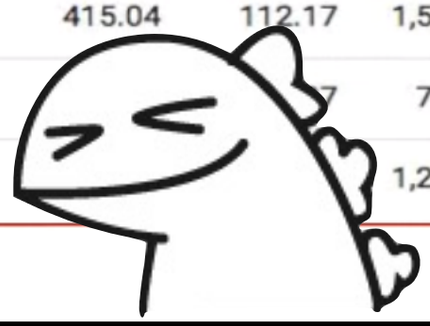
10-Node Challenge - Winner

10 Node ISC21 List

IO500 | 10 Node

Sorted by score

# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
			INFORMATION					IO500		
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	10	1,800	2,595.89	193.77	34,777.27
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
5	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	10	480	474.10	91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120	415.04	112.17	1,535.63
9	SC19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	[illegible]	[illegible]	715.76
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800	[illegible]	[illegible]	1,251.32



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the 10 Node Challenge

IO500

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/ten

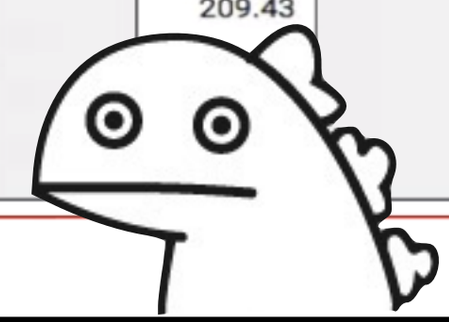
Full list - Bandwidth Winner

IO500 ISC21 List

IO500 | 10 Node

Sorted by BW

#	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW ↑ (GiB/s)	MD (KIOP/S)
		INFORMATION							IO500	
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864		3,421.62	
2	SC20	JCAHPC	Oakforest-PACS	DDN	IME	2,048	4,096		697.20	
3	ISC20	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	IME	2,048	2,048		515.59	
4	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	
5	ISC20	Intel	Wolf	Intel	DAOS	52	1,664		371.67	
6	ISC20	Oracle Cloud Infrastructure	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS	270	3,240		293.05	
7	ISC21	Google Cloud	Google	DDN	Lustre	1,000	5,000		282.78	
8	SC19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5,280		209.43	
9	ISC21	University of Florida	HiPerGator	DDN	Lustre	140	4,480			
10	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456			



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500 BW Score

IO500

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/io500

Full list - Metadata Winner

IO500 ISC21 List

IO500 | 10 Node

Sorted by MD

#	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	IO500 BW (GIB/S)	IO500 MD ↑ (KIOP/S)
			INFORMATION							
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864		3,421.6[illegible]	396,872.82
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.7[illegible]	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	52	1,664		371.67	8,649.57
4	ISC20	TACC	Frontera	Intel	DAOS	60	1,440		78.31	7,449.56
5	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456		176.37	5,545.61
6	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625		174.74	5,045.33
7	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	18	720		104.52	3,195.53
8	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544		108.19	2,668.57
9	ISC21	Supermicro		Supermicro	DAOS	10	1,120		[illegible]17	1,535.63
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800		[illegible]	1,251.32

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500 MD Score

IO500

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/io500

Full list - Winner

IO500 ISC21 List

IO500 | 10 Node

Sorted by Score

# ↑	INFORMATION							IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864	36,850.37	3,421.62	396,872.82
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456	988.99	176.37	5,545.61
5	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625	938.95	174.74	5,045.33
6	ISC20	TACC	Frontera	Intel	DAOS	60	1,440	763.80	78.31	7,449.56
7	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	18	720	577.93	104.52	3,195.53
8	ISC21	Google Cloud	Google	DDN	Lustre	1,000	5,000	569.99	282.78	1,148.90
9	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	[illegible]	[illegible]8.19	2,668.57
10	SC19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5,280	[illegible]	[illegible]	982.78

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500

IO500

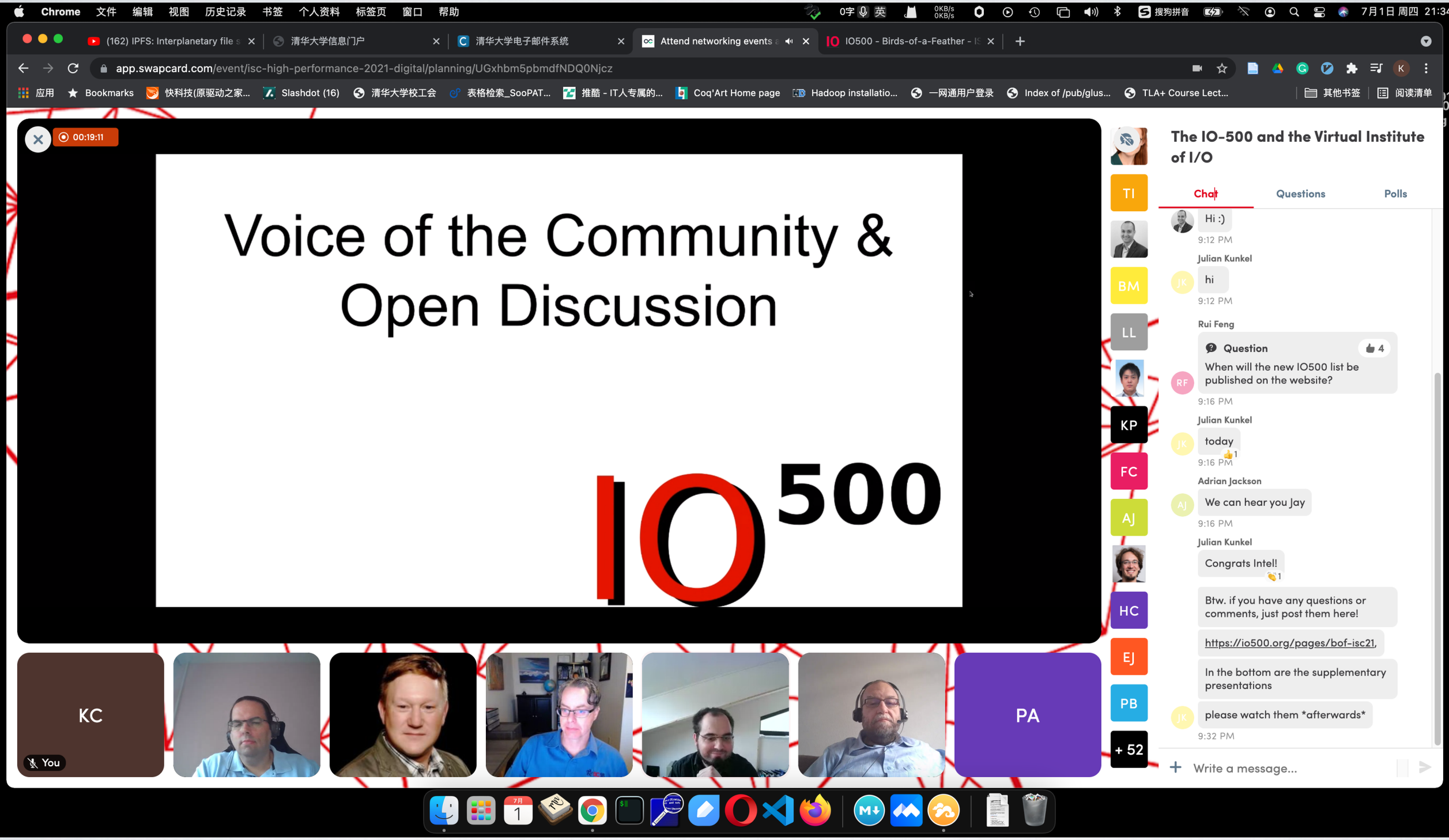
July 2021

IO-500 Steering Board

https://io500.org/list/isc21/io500

List of Awarded Systems in the Ranked Lists

10-Node	Bandwidth	Intel Endeavour	DAOS	398.77	GiB/s
	Metadata	Pengcheng Laboratory	MadFS	34777.27	kIOPS
	Overall	Pengcheng Laboratory	MadFS	**2595.89**	**score**
IO500	Bandwidth	Pengcheng Laboratory	MadFS	3421.62	GiB/s
	Metadata	Pengcheng Laboratory	MadFS	396872.82	kIOPS
	Overall	Pengcheng Laboratory	MadFS	**36850.37**	**score**


app.swapcard.com/event/isc-high-performance-2021-digital/planning/UGxhbm5pbmdfNDQ0Njcz
00:19:11
Voice of the Community &
Open Discussion
IO 500
The IO-500 and the Virtual Institute of I/O
Chat
Questions
Polls
Hi :)
9:12 PM
Julian Kunkel
hi
9:12 PM
Rui Feng
Question
When will the new IO500 list be published on the website?
9:16 PM
Julian Kunkel
today
9:16 PM
Adrian Jackson
We can hear you Jay
9:16 PM
Julian Kunkel
Congrats Intel!
Btw. if you have any questions or comments, just post them here!
https://io500.org/pages/bof-isc21,
In the bottom are the supplementary presentations
please watch them *afterwards*
9:32 PM
Write a message...
KC
You
PA
+ 52

Issues Regarding Fair Comparisons

- Production vs. benchmark-only system
- Production vs. research file system software
- Non-redundant vs. RAID/Erasure Coded storage
- Vendor submission vs. end-user submission
- Cloud vs. on-prem
- Ephemeral vs. persistent file system
- Storage media technologies

IO500 ISC21 38

The IO-500 and the Virtual Institute of I/O

Chat Questions Polls

Shall we split the list into "production" and "non-production" lists?
10 minutes ago
- Yes
- No

19 votes · 12 hours left

Shall we include new benchmarking phases (reasons to be discussed)
15 minutes ago
- Random 4K
- Find easy and hard
- Concurrent metadata load simulating real user behavior
- Others (please paste)

23 votes · 12 hours left

The IO-500 and the Virtual Institute of I/O

Chat | Questions | Polls

please watch them *afterwards*

hi ken ^-^

I just submitted a Poll

please fill it ^-^

9:39 PM

Glenn Lockwood

comment on the 4k random: if that's not a real workload, ior-hard's 47k read/write is *definitely* not a real workload for 99.99% of the world :)

1

9:40 PM

Julian Kunkel

haha, very true ^-^

George, can you move back 1 slide

9:43 PM

Kelsey Prantis

While not necessarily a comment on what should be permitted, it'd be great where there are rules if there was some mechanism in the io500 app to enforce them. For example - there are specific persistency requirements, but afaict no way looking at the results or mechanism in the tool that ensures that is the case. This leaves the benchmark potentially open to abuse. If we're adding more rules, it'd be great to also consider ability to enforce alongsid

SCROLL TO BOTTOM

The IO-500 and the Virtual Institute of I/O

Chat | Questions | Polls

splitting the list

yes Kelsey sounds great but it is very tricky to enforce persistency

9:45 PM

Kelsey Prantis

For sure tricky.

9:45 PM

Julian Kunkel

Ideas are welcome, we will surely think about what we can do!

On the defense, once we added verification of written data, it got rid of some accidential data corruption...

9:45 PM

Glenn Lockwood

re: splitting the list, maybe make it a metadata arg so stunt systems can be filtered out

9:46 PM

Julian Kunkel

can you elaborate on this

Since some of you voted "no" for splitting the list into production and non-production reasons, what arguments can you bring?

9:47 PM

Glenn Lockwood

@Julian

SCROLL TO BOTTOM

The IO-500 and the Virtual Institute of I/O

Chat | Questions | Polls

Question 1

It's a good idea this split list to avoid the controversy of HPL list with the cloud system and their clones... But what is the definition of production/ vendor system? Some vendors usually gives temporarily access to some of their system...

9:52 PM

Julian Kunkel

on-demand if it can be used by users is probably production!

if it is just ever being used for IO500 then not

1

I think of this personally a bit like for TOP500, if you build a systme that ever just runs TOP500 then is decomissioned, then it is not production.

9:54 PM

Kelsey Prantis

What if you have a system that is owned by a vendor, but is permanent, and does allow external users to login and use it to work on porting their code, benchmarking their applications, etc?

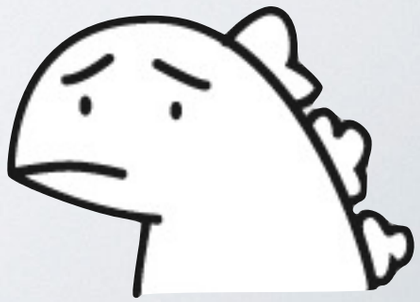
Is that benchmark or production?

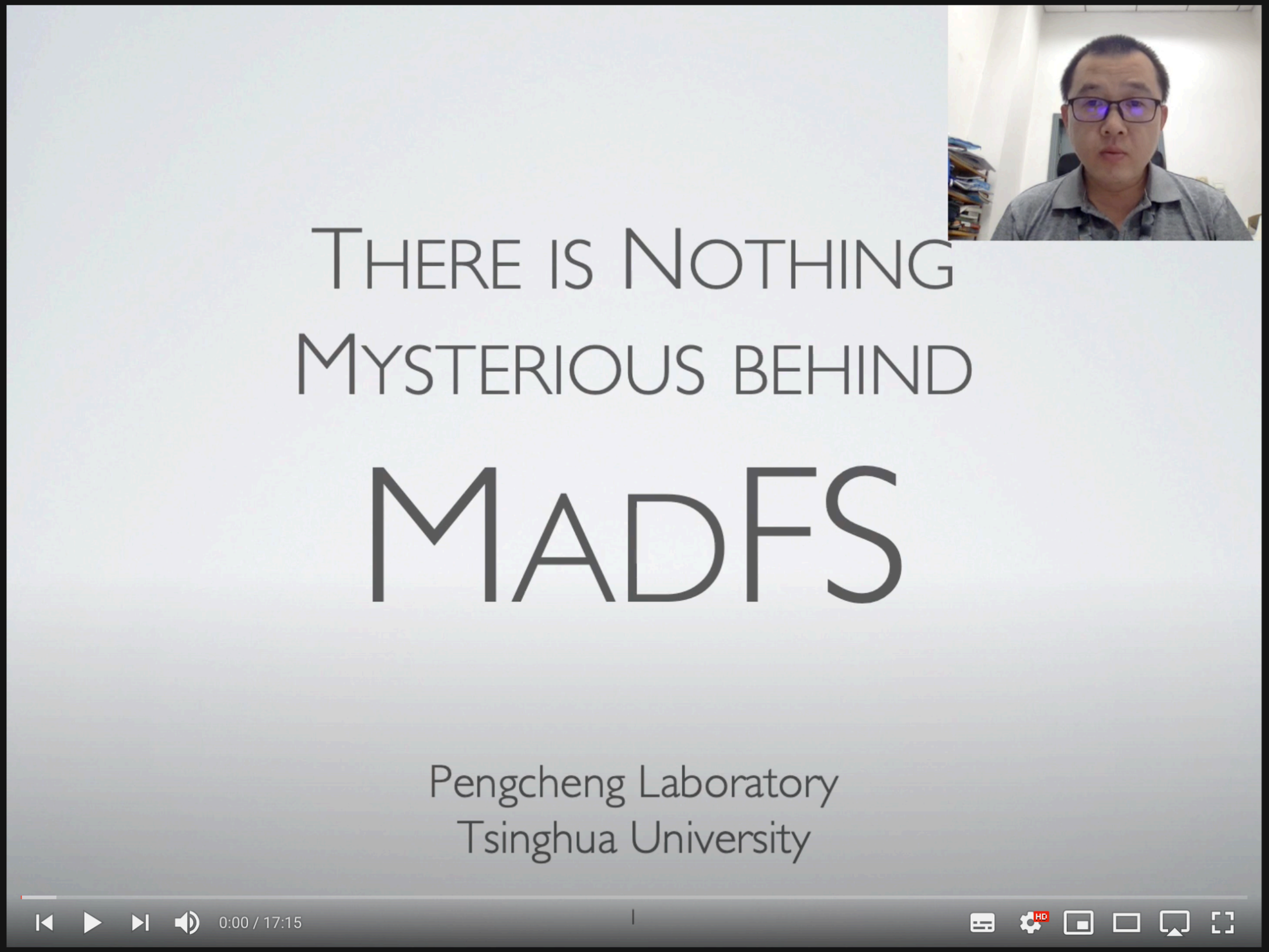
9:55 PM

Alexander Moskovsky

Systems SCROLL TO BOTTOM I and

- 如何保证数据持久化?
- 如何区分 production / benchmark 系统?





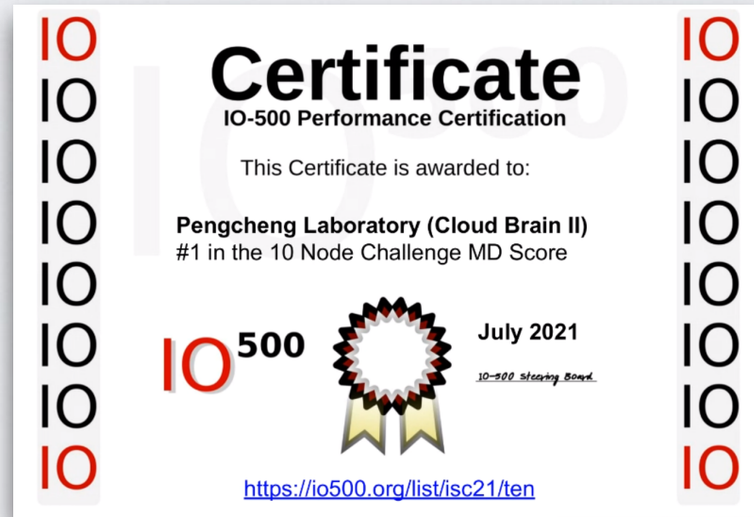
There is nothing mysterious behind MadFS

280次观看 • 2021年6月27日

11 0 分享 保存

完结撒花 🎉

Certificate
IO-500 Performance Certification
This Certificate is awarded to:
Intel (Endeavour)
#1 in the 10 Node Challenge BW Score
IO500
July 2021
IO-500 Steering Board
https://io500.org/list/isc21/ten


Certificate
IO-500 Performance Certification
This Certificate is awarded to:
Pengcheng Laboratory (Cloud Brain II)
#1 in the 10 Node Challenge MD Score
IO500
July 2021
IO-500 Steering Board
https://io500.org/list/isc21/ten

Certificate
IO-500 Performance Certification
This Certificate is awarded to:
Pengcheng Laboratory (Cloud Brain II)
#1 in the 10 Node Challenge
IO500
July 2021
IO-500 Steering Board
https://io500.org/list/isc21/ten

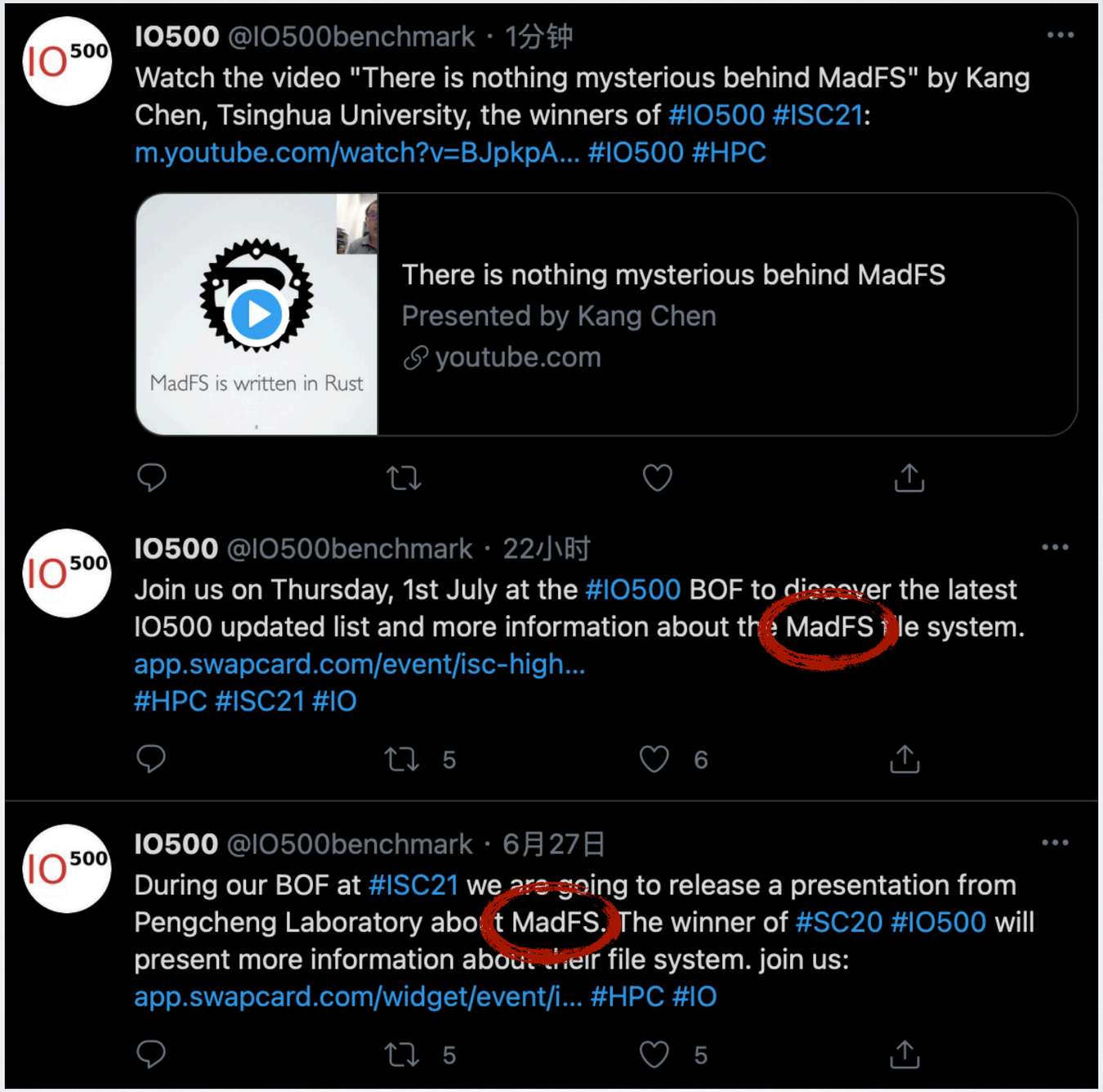
Certificate
IO-500 Performance Certification
This Certificate is awarded to:
Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500 BW Score
IO500
July 2021
IO-500 Steering Board
https://io500.org/list/isc21/io500

Certificate
IO-500 Performance Certification
This Certificate is awarded to:
Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500 MD Score
IO500
July 2021
IO-500 Steering Board
https://io500.org/list/isc21/io500

Certificate
IO-500 Performance Certification
This Certificate is awarded to:
Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500
IO500
July 2021
IO-500 Steering Board
https://io500.org/list/isc21/io500

国外舆情大赏

IO500 官方带货


IO500 @IO500benchmark · 1分钟
Watch the video "There is nothing mysterious behind MadFS" by Kang Chen, Tsinghua University, the winners of #IO500 #ISC21: m.youtube.com/watch?v=BJpkpA... #IO500 #HPC
MadFS is written in Rust
There is nothing mysterious behind MadFS
Presented by Kang Chen
youtube.com
IO500 @IO500benchmark · 22小时
Join us on Thursday, 1st July at the #IO500 BOF to discover the latest IO500 updated list and more information about the MadFS file system.
app.swapcard.com/event/isc-high...
#HPC #ISC21 #IO
5
6
IO500 @IO500benchmark · 6月27日
During our BOF at #ISC21 we are going to release a presentation from Pengcheng Laboratory about MadFS. The winner of #SC20 #IO500 will present more information about their file system. join us:
app.swapcard.com/widget/event/i... #HPC #IO
5
5

UCX 很开心

openucx @openucx · 52m

#ISC21 surprise! MadFS (the fs that holds the first place in **#IO500**) uses custom developed RPC network on top of @rustlang + @openucx + RoCE network !!! Link to the talk 👇👇👇

> **IO500** @IO500benchmark · 12h
>
> Watch the video "There is nothing mysterious behind MadFS" by Kang Chen, Tsinghua University, the winners of #IO500 #ISC21: m.youtube.com/watch?v=BJpkpA... #IO500 #HPC

1 2

openucx @openucx · 53m

Replying to @openucx

MadFS team developed asynchronous Rust bindings for UCX

madsys-dev/**async-ucx**

Async Rust UCX bindings.

3 Contributors 0 Issues 4 Stars 1 Fork

madsys-dev/async-ucx

Async Rust UCX bindings. Contribute to madsys-dev/async-ucx development by creating an account on GitHub.

github.com

1

飞得太远了

okamo(aged 6 yrs)
@hei_nyan

回复 @hei_nyan

MadFSがぶっ飛びすぎてる

由 Google 翻译自 日语

MadFS 飞得太远了

Full list - Winner

IO500 ISC21 List

IO500 | 10 Node | Sorted by Score

# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	IO500 BW (GIB/S)	IO500 MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864	36,850.37	3,421.62	396,872.82
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456	988.99	176.37	5,545.61
5	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625	938.95	174.74	5,045.33
6	ISC20	TACC	Frontera	Intel	DAOS	60	1,440	763.80	78.31	7,449.56
7	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	18	720	577.93	104.52	3,195.53
8	ISC21	Google Cloud	Google	DDN	Lustre	1,000	5,000	569.99	282.78	1,148.90
9	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	537.31	108.19	2,668.57
10	SC19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5,280	453.68	209.43	982.78

IO500 ISC21 29

下午9:31 · 2021年7月1日 · Twitter Web App

完全的黑马？？

okamo(aged 6 yrs)
@hei_nyan

MadFSってなんだ、、、完全なダークホースでは？？
由 Google 翻译自 日语
什么是MadFS.....完全的黑马？ ??

上午8:30 · 2020年11月19日 · Twitter for iPad

okamo(aged 6 yrs) @hei_nyan · May 24
Replying to @hei_nyan @glennklockwood and @suhaibkhan
MadFS seems to be based on GeckoFS.
It says "The principle is based on **GekkoFS**" (translated by DeepL).

> **okamo(aged 6 yrs)** @hei_nyan · May 24
> MadFSはGeckoFS拡張ぽいことがわかった
>
> is.gd/B1FH3j
> Show this thread
>
> MadFS
> 小巧精悍的分布式文件系统
>
> 清华大学计算机系 王润基

1 2 3

okamo(aged 6 yrs) @hei_nyan · 2020年11月19日
回复 @hei_nyan
Not foundでアクセスできないけど、引っかかった中だとこれがそうなのかな、、、
この人のレポジトリにGekkoFSもあるので、そうっぽい
github.com/wangrunji0408?...

"MadFS" yet another burst buffer file system

All Images Videos Shopping News More Settings Tools

About 24 results (0.45 seconds)

github.com › wangrunji0408 › MadFS
wangrunji0408/MadFS: Yet another burst buffer file system - GitHub
Yet another burst buffer file system. Contribute to wangrunji0408/**MadFS** development by creating an account on GitHub.

Steal MadFS

Jeff Johnson @echo5juliet · 2月10日
It'll be much easier for China to steal tech and science again when Chinese intelligence is allowed to reopen their on-campus offices.
Kevin McCarthy @GOPLeader · 2月10日
It has been a year since the Chinese Communist Party let a pandemic spread around the world.
Instead of holding them accountable for hiding the truth, the Biden Admin is rewarding China by allowing their propaganda to infiltrate our college campuses. bit.ly/3aUVuNY
1
1
Raj Ayyampalayam @raj_iampal · 2月10日
US should steal the MadFS tech from China.

黑转粉?

Glenn K. Lockwood
@glennklockwood

Man this is a really good talk. I thought MadFS would be hocus-pocus stunt software, but its creators did their homework. Very similar design to GekkoFS, DAOS, etc but had to DIY due to lack of hw support for Argobots, libfabric, etc

> **IO500** @IO500benchmark · 9小时
> Watch the video "There is nothing mysterious behind MadFS" by Kang Chen, Tsinghua University, the winners of #IO500 #ISC21: m.youtube.com/watch?v=BJpkpA... #IO500 #HPC

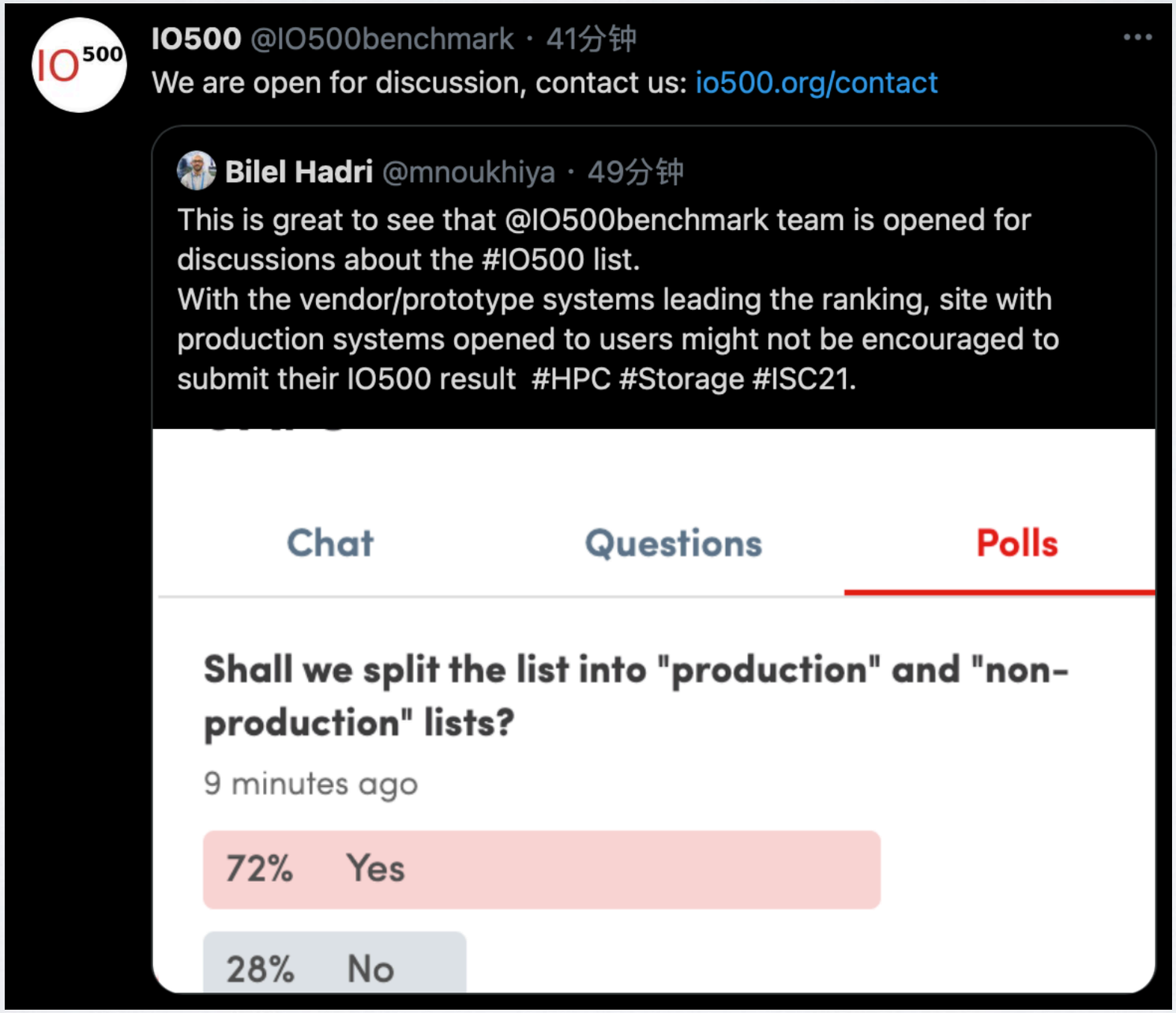
下午11:15 · 2021年7月1日 · Twitter for Mac

1 转推 **15** 喜欢次数

Glenn K. Lockwood @glennklockwood · 8小时
回复 @glennklockwood
Also makes me want to look closer at Rust. Their answer to Argobots (async I/O with blocking-like API) was to use a language that provided native support for it. Their choice to use POSIX intercept as the client isn't great (security/clone/etc break down)

8

把榜打裂了


IO500 @IO500benchmark · 41分钟
We are open for discussion, contact us: io500.org/contact
Bilel Hadri @mnoukhiya · 49分钟
This is great to see that @IO500benchmark team is opened for discussions about the #IO500 list.
With the vendor/prototype systems leading the ranking, site with production systems opened to users might not be encouraged to submit their IO500 result #HPC #Storage #ISC21.
Chat
Questions
Polls
Shall we split the list into "production" and "non-production" lists?
9 minutes ago
72% Yes
28% No

MadFS

还开源吗?

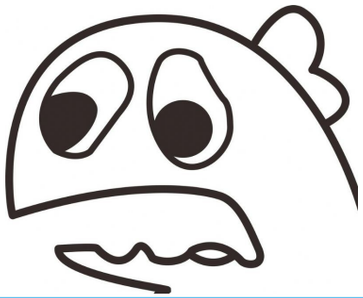
https://github.com/madsys-dev/MadFS

还打不打了?

DAOS: Primary Storage on Aurora

Aurora DAOS configuration

- Capacity: 230PB
- Bandwidth: >25TB/s



"Combined in Aurora, the Intel compute system, Cray Slingshot network, and the Intel DAOS storage open new possibilities for accelerating the scientific research needed to solve critical human challenges such as cancer and disease. DAOS enables the creation of new storage data models tailored specifically to applications like the Cancer Distributed Learning Environment (CANDLE) which provide a powerful platform to advance a wide array of scientific challenges using deep learning."

– Rick Stevens, Associate Laboratory Director for Computing, Environment and Life Sciences

"The Argonne Leadership Computing Facility is excited to be the first major production deployment of the DAOS storage system as part of Aurora, an US exascale system coming in 2021. As designed, it will provide us unprecedented levels of metadata operation rates and extremely high bandwidth for I/O intensive workloads."

– Susan Coghlan, ALCF-X Project Director/Exascale Computing Systems Deputy Director

Part4

MadFS 未来发展计划

MadFS
为了性能牺牲了什么

牺牲功能

路径	服务端存储全路径	目录逐层引用
Lookup	一次查询	多次路径解析通信
Find	服务端，完全并行	客户端，负载不均衡
Rename	需要移动数据	O(1)

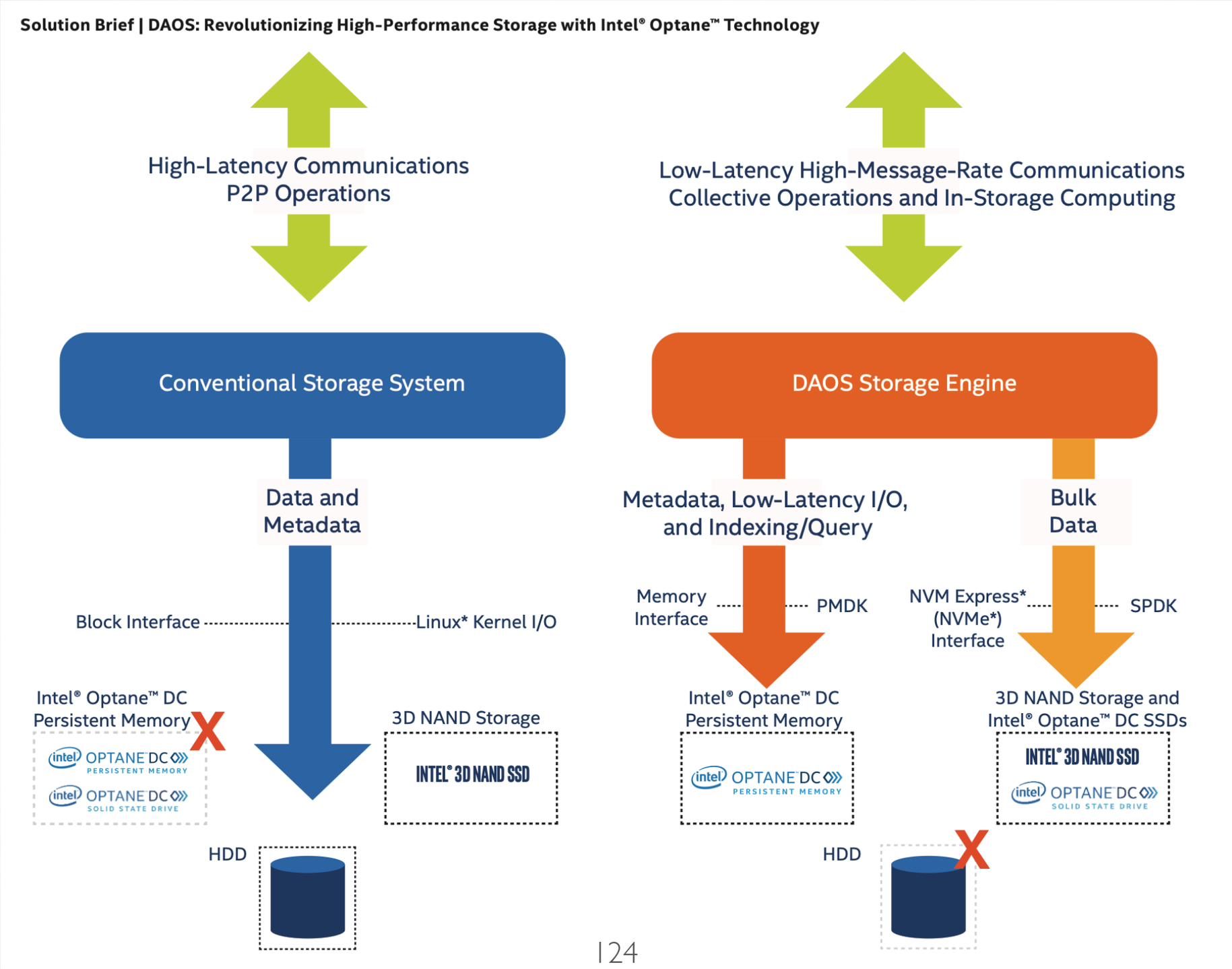
牺牲一致性

文件大小	每次写入更新	关闭文件时更新	不在元数据中存
并发随机写	元数据写冲突	无开销	无开销
并发追加写	正确	写覆盖	需广播查询
查询大小	强一致	同时打开文件时不一致	需广播查询

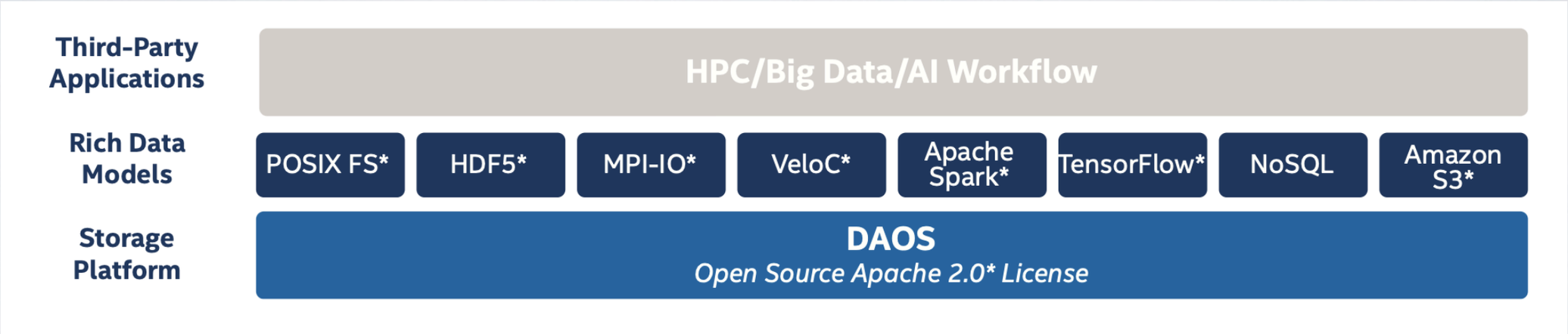
DAOS

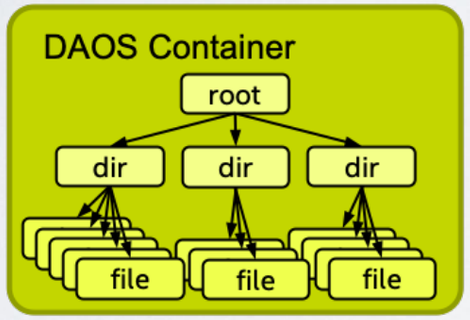
Distributed Asynchronous Object Storage

核心卖点

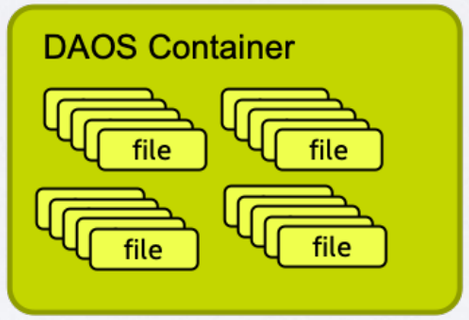

Solution Brief | DAOS: Revolutionizing High-Performance Storage with Intel® Optane™ Technology
High-Latency Communications
P2P Operations
Low-Latency High-Message-Rate Communications
Collective Operations and In-Storage Computing
Conventional Storage System
DAOS Storage Engine
Data and Metadata
Metadata, Low-Latency I/O, and Indexing/Query
Bulk Data
Block Interface
Linux* Kernel I/O
Memory Interface
PMDK
NVM Express* (NVMe*) Interface
SPDK
Intel® Optane™ DC Persistent Memory
3D NAND Storage
INTEL® 3D NAND SSD
Intel® Optane™ DC Persistent Memory
3D NAND Storage and Intel® Optane™ DC SSDs
INTEL® 3D NAND SSD
HDD
HDD

应用层

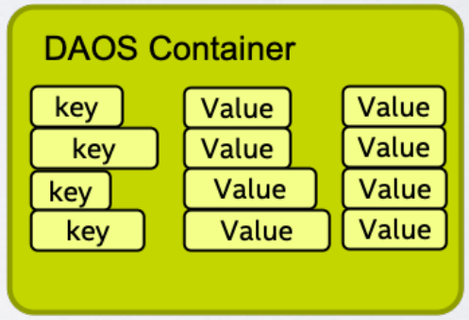




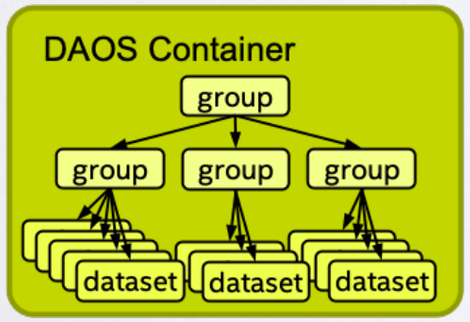
Encapsulated POSIX Namespace



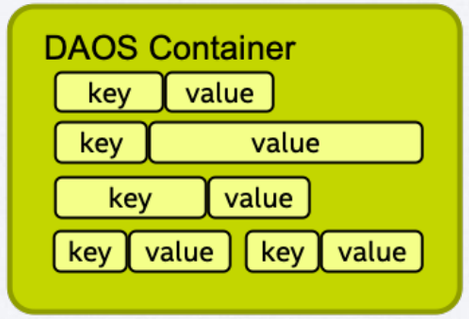
File-per-process



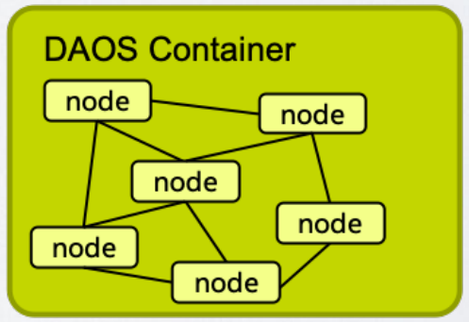
Columnar Database



HDF5 « File »



Key-value store

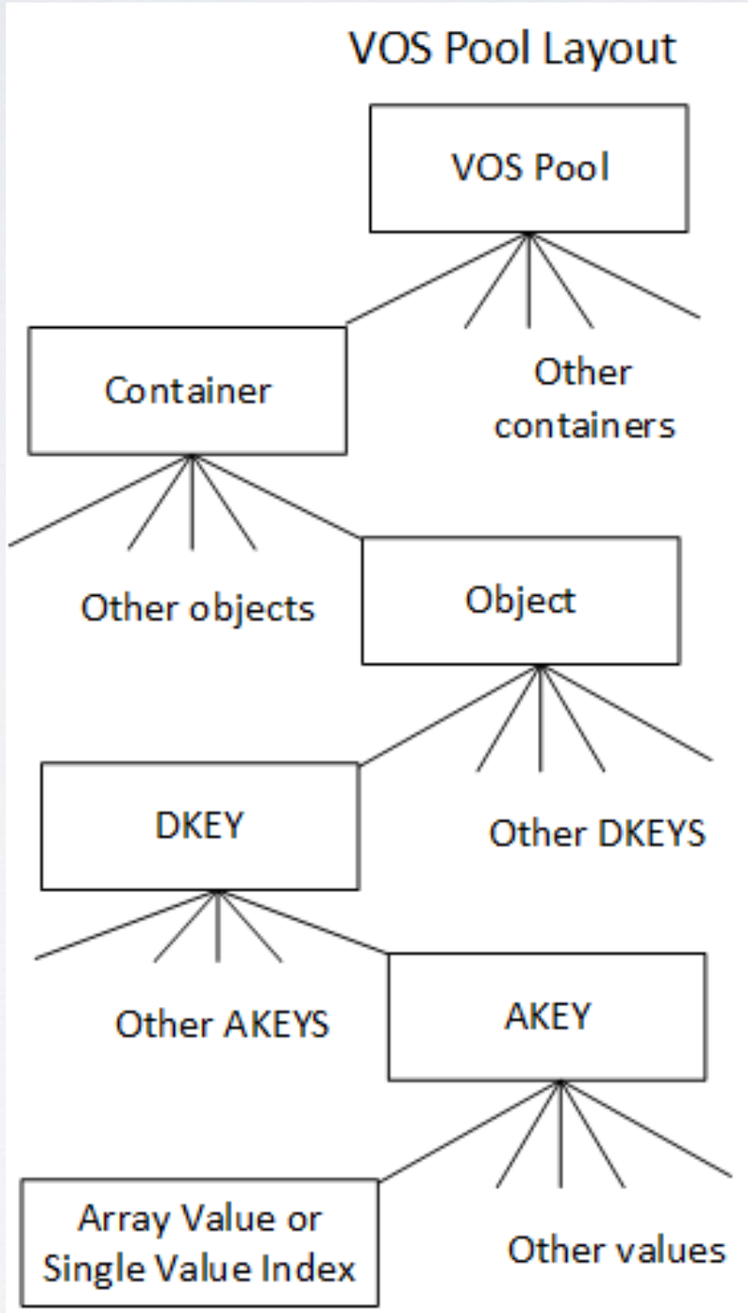


Arbitrary Connected Graph

对象存储模型

数据模型

数据分布

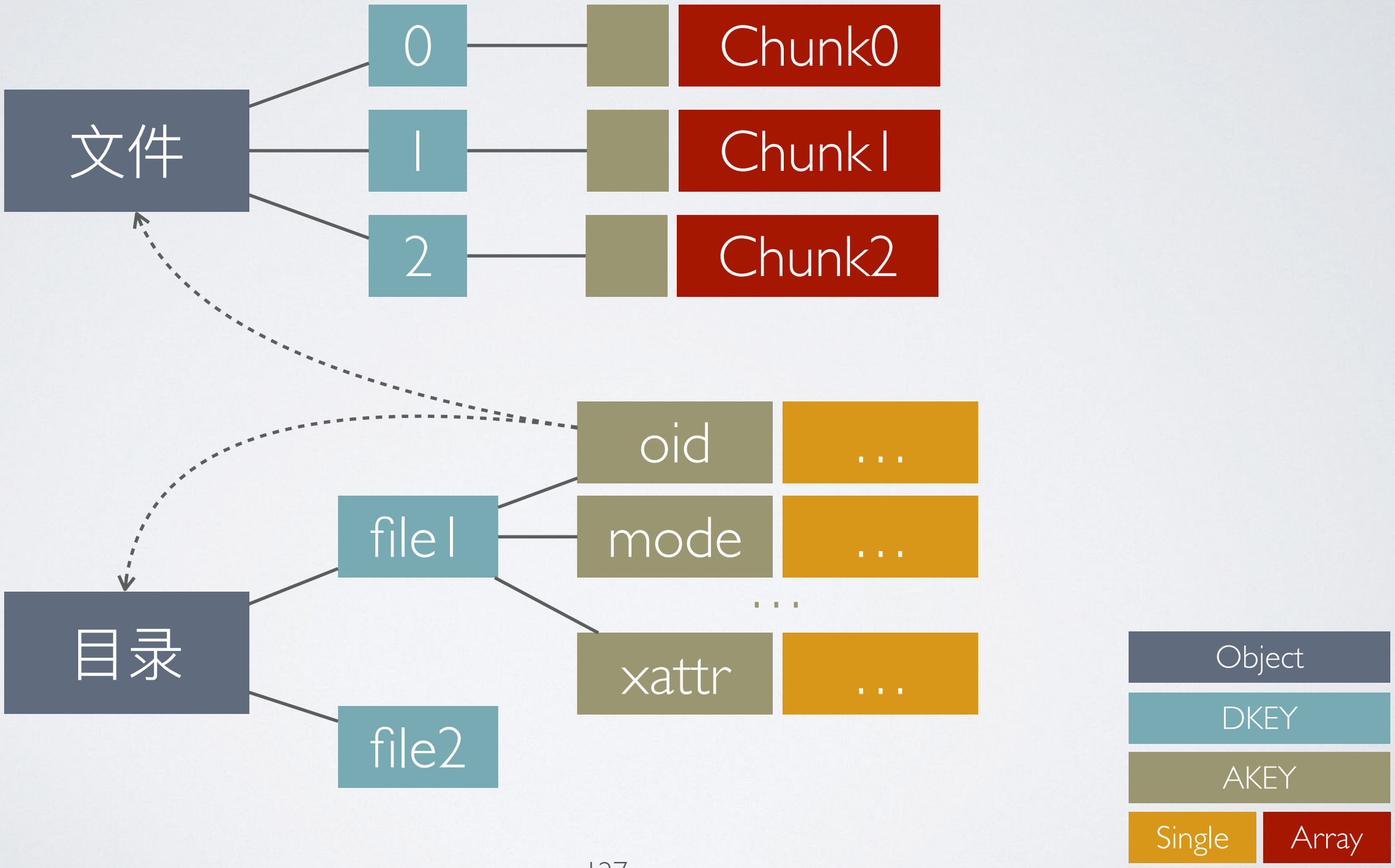


用户管理

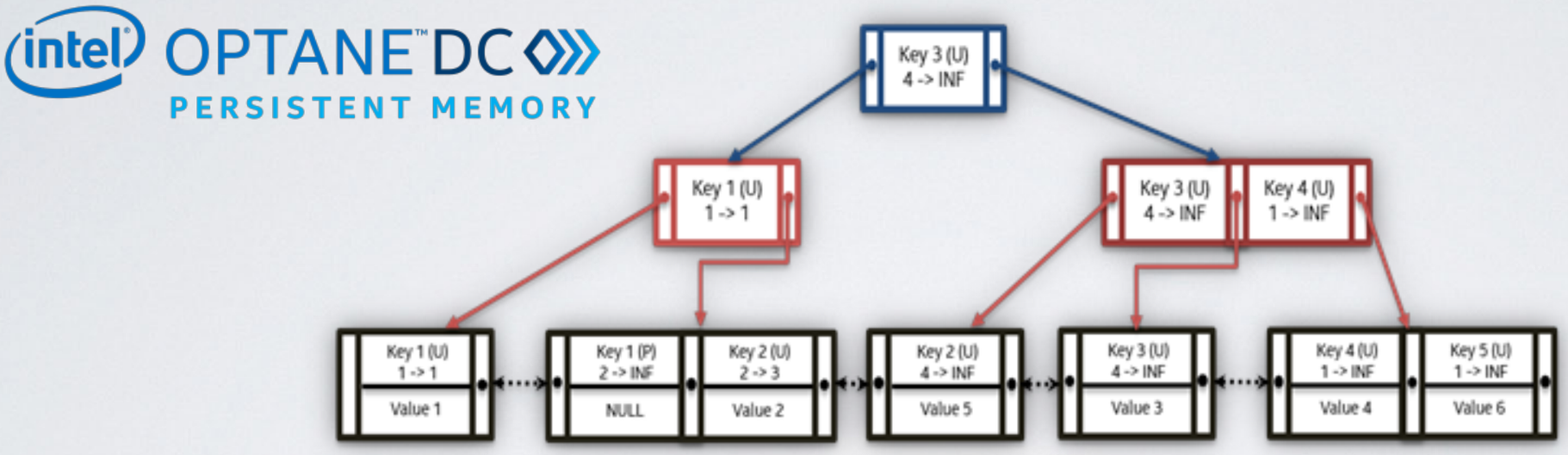
对象存储

键值存储

POSIX 文件系统

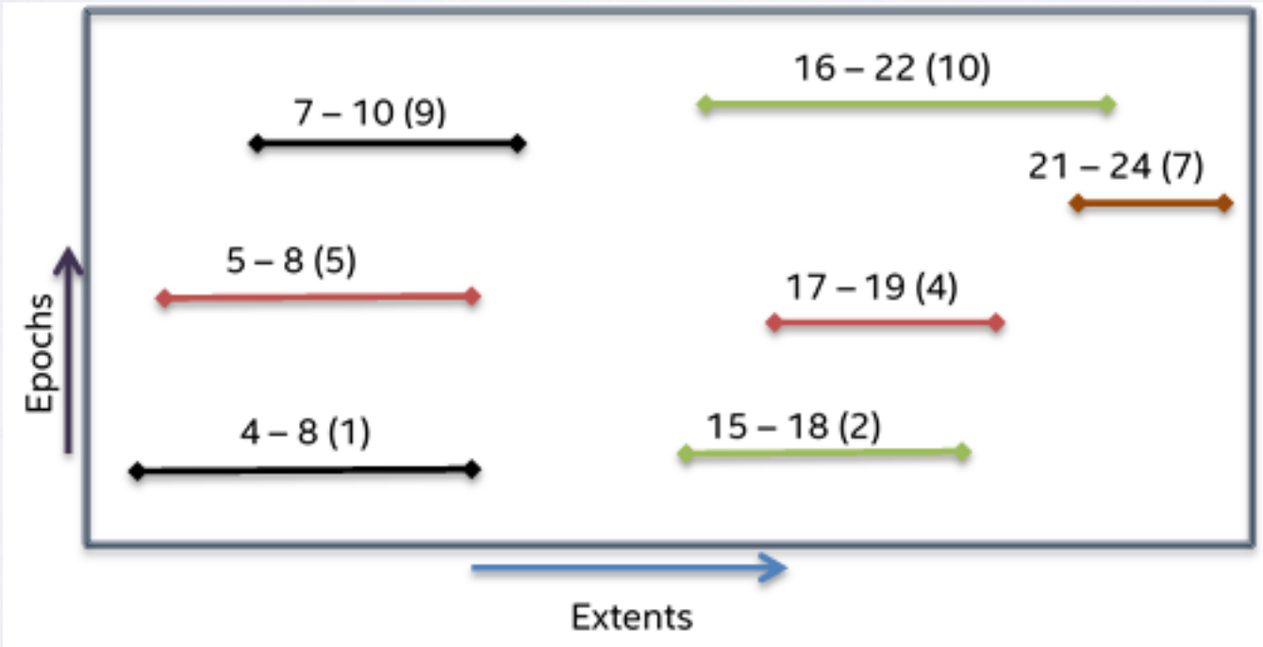

文件
0
1
2
Chunk0
Chunk1
Chunk2
oid
...
mode
...
...
xattr
...
目录
file1
file2
Object
DKEY
AKEY
Single
Array

对象的物理存储结构



Key-Value

B+树



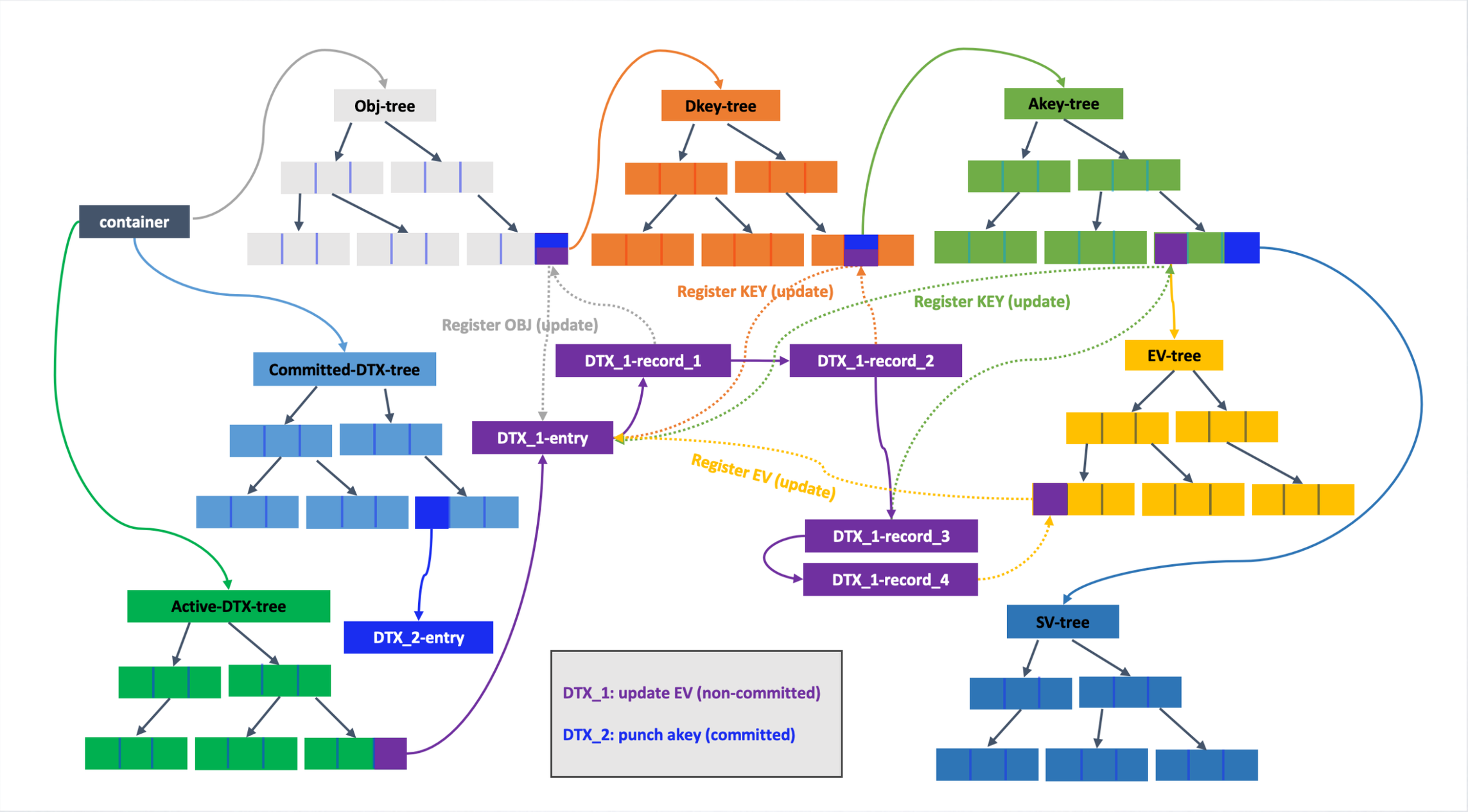
Key-Array

EV树

(R树变体)

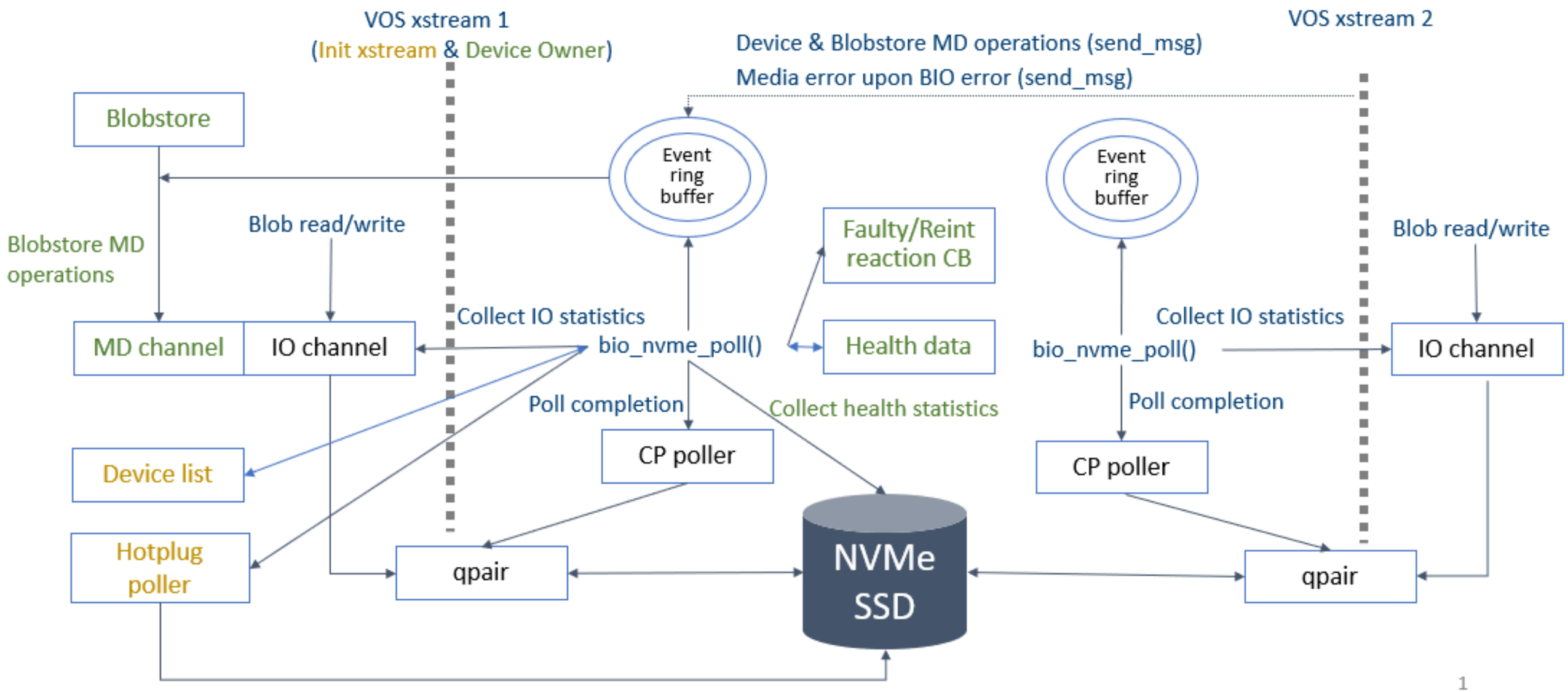
看起来挺简单的?

再加上副本



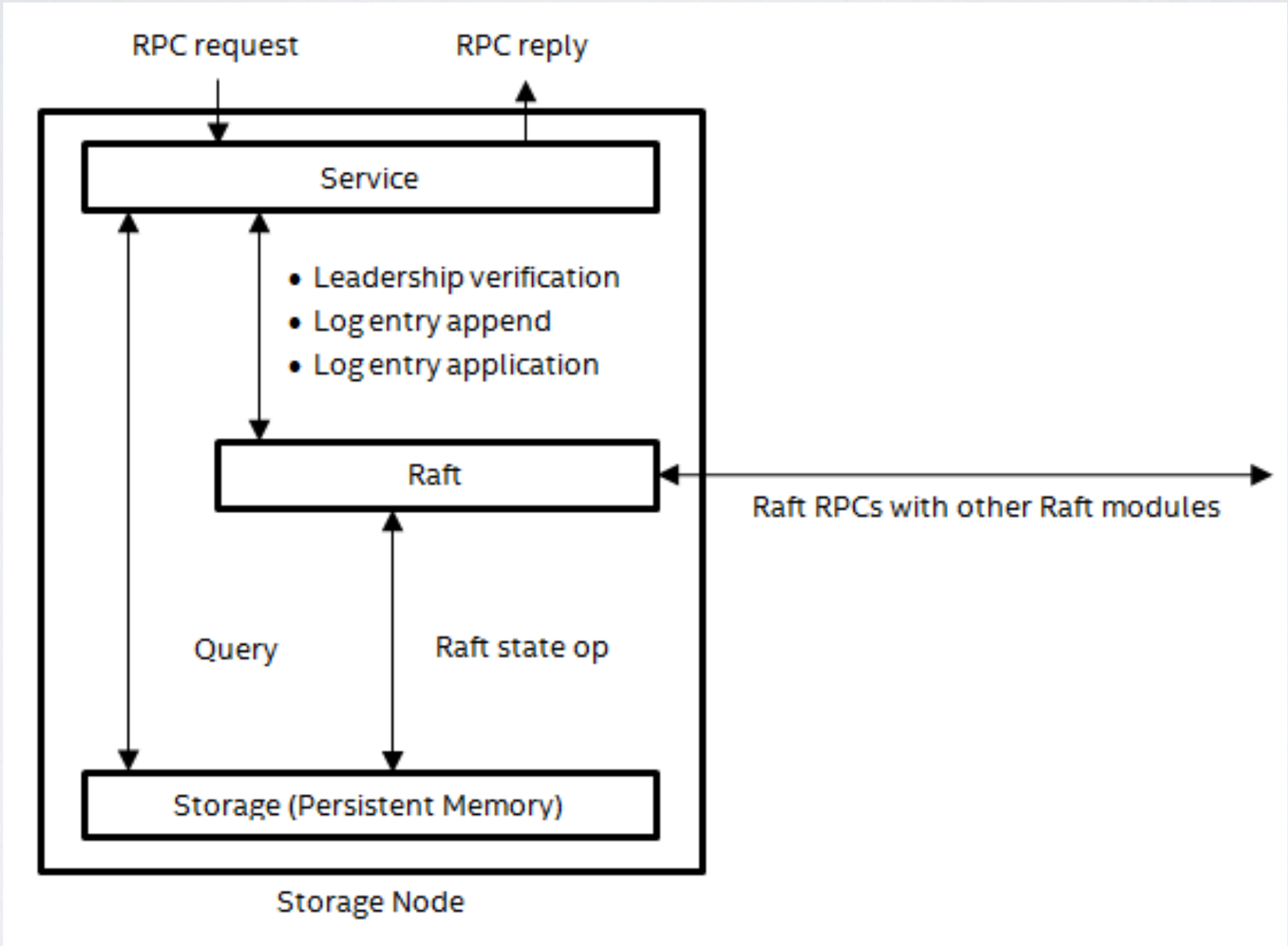
DAOS 使用基于 2PC 的三副本容错

再加上 SPDK



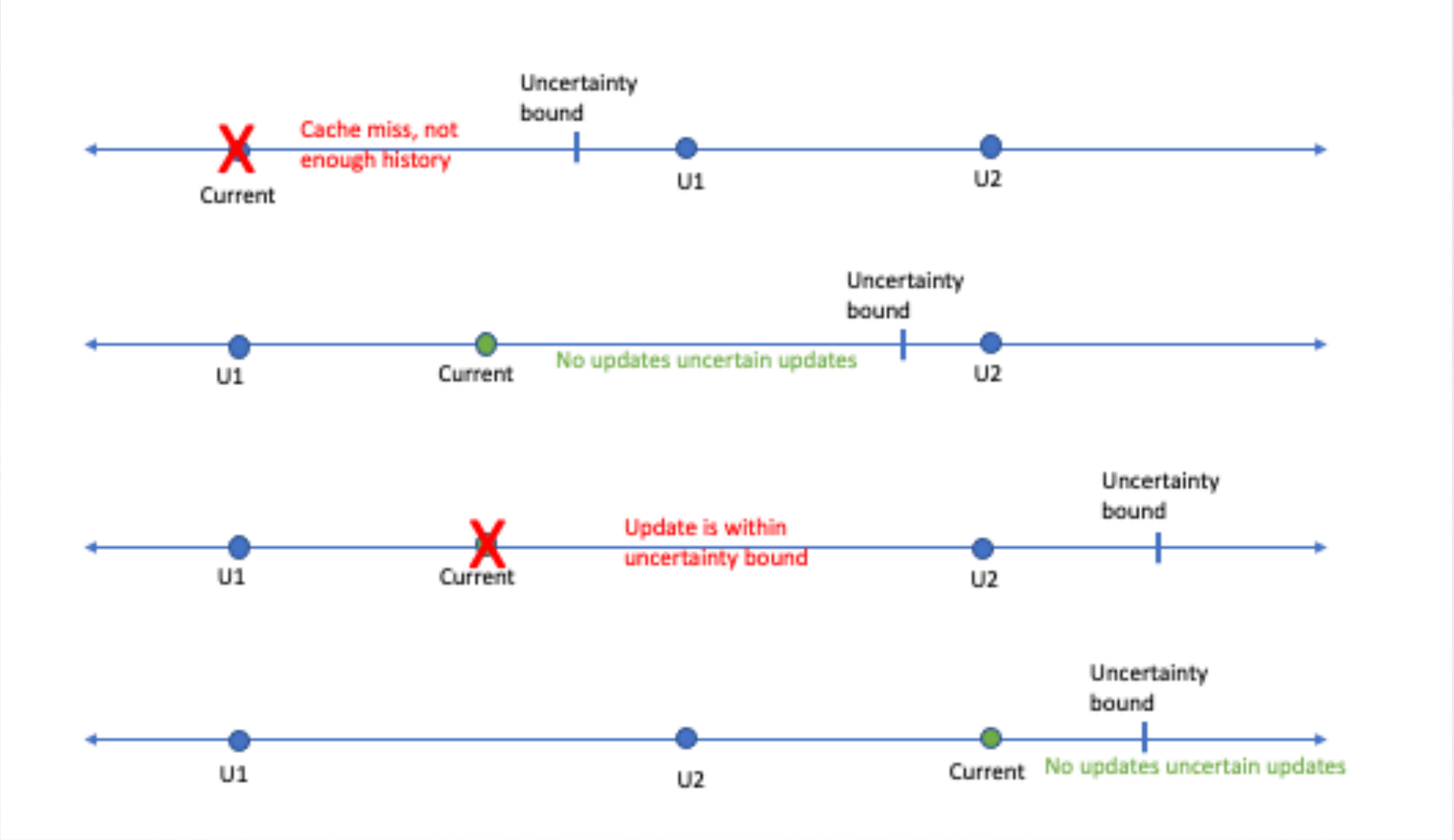
DAOS 使用 SPDK 存储大块数据

再加上分布式共识



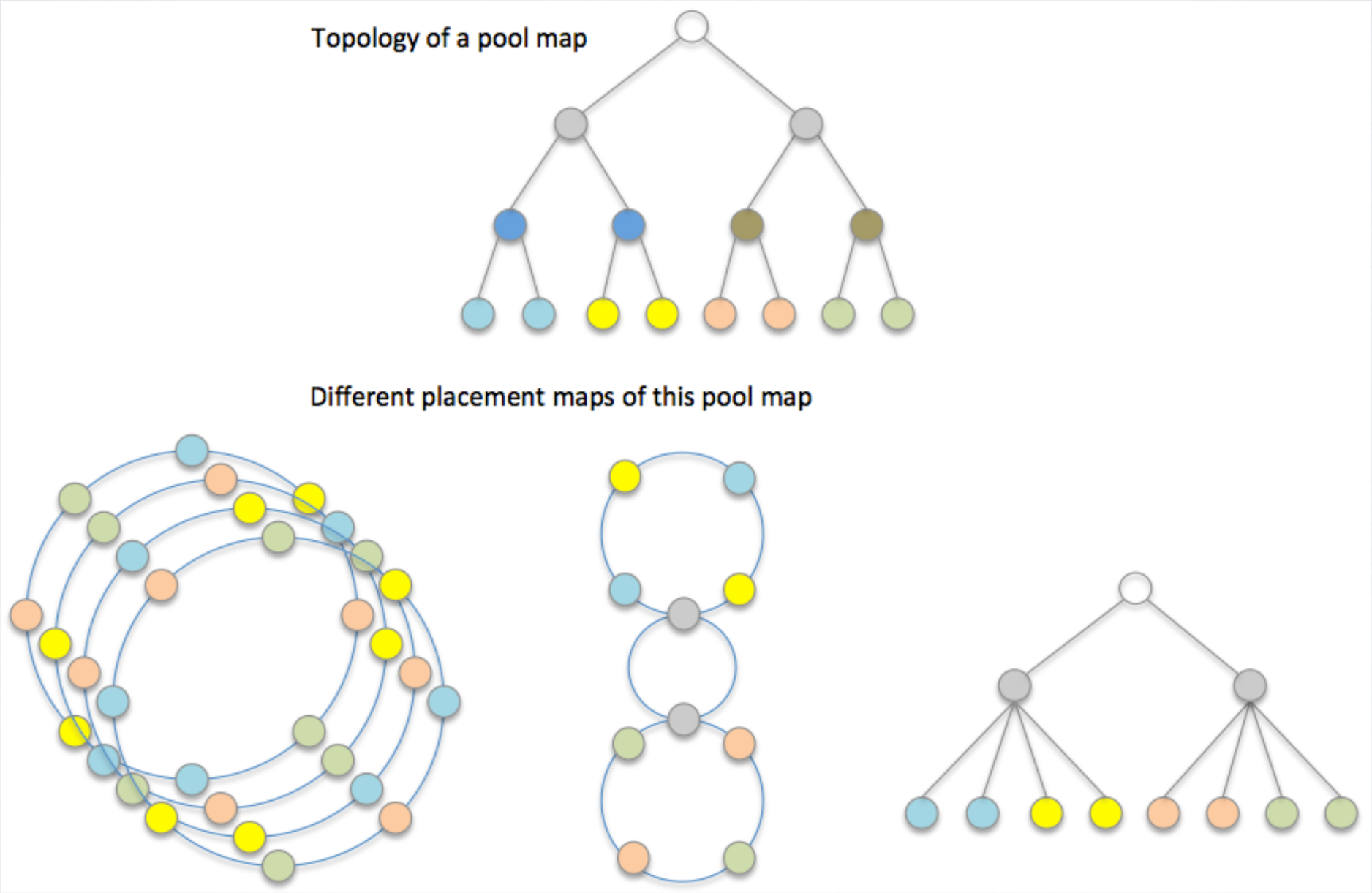
DAOS 使用 Raft 维护集群配置信息

再加上分布式事务



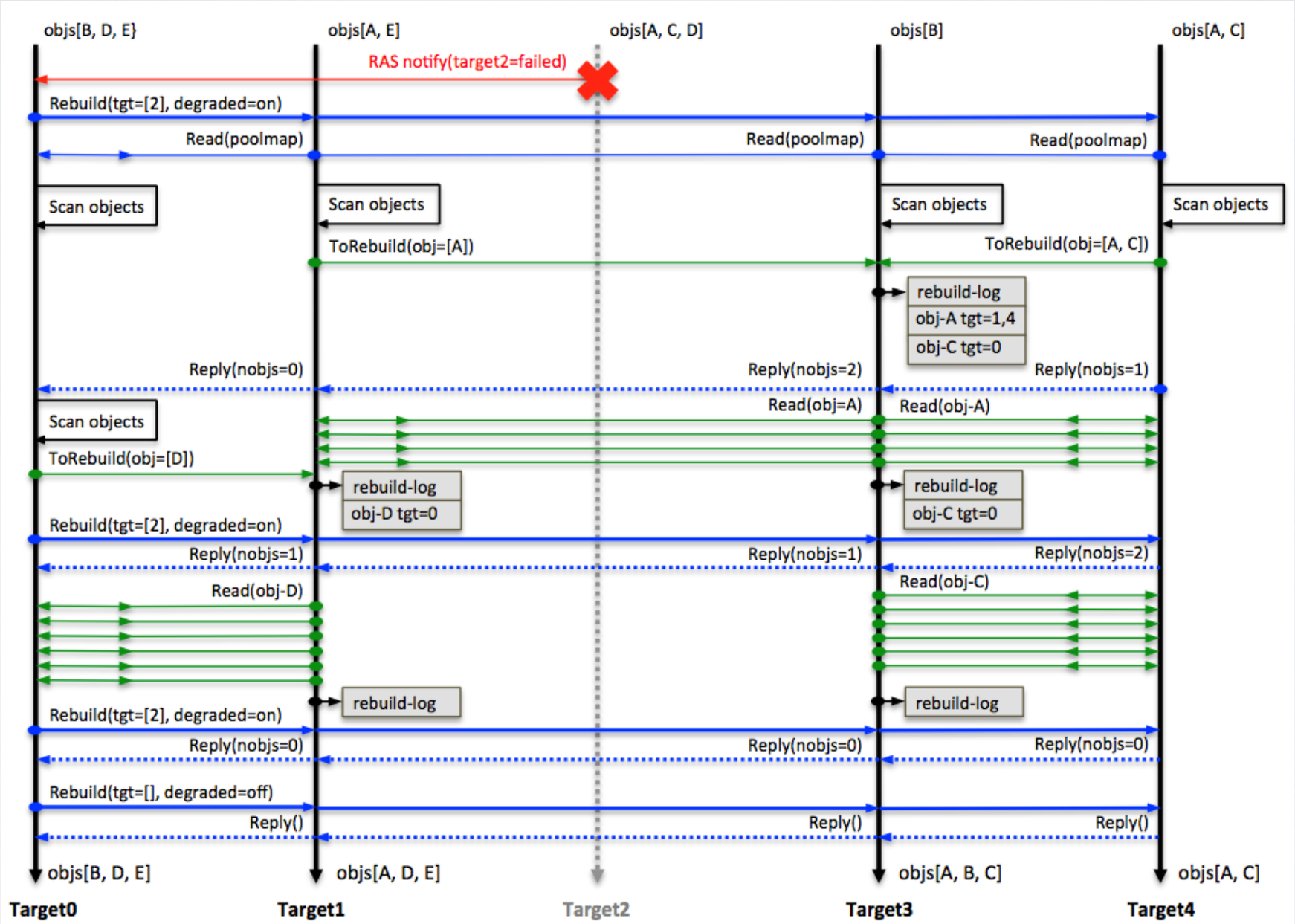
DAOS 使用基于 MVCC 的分布式事务

再加上放置算法



DAOS 使用一致性哈希计算副本位置

再加上错误恢复



DAOS 使用 @#%¥*%...# 方法恢复数据

课程公告提醒

你所学习的课程有新发布的课程公告。

课程名称：分布式系统导论（0）
公告标题：Dont forget to drop the course

DAOS Community Roadmap

1Q20	2Q20	3Q20	4Q20	1Q21	2Q21	3Q21	4Q21	1Q22	2Q22	3Q22
1.0		1.2			2.0		2.2		2.4	

Released on June 18, 2020

DAOS:
- CentOS7 support
- NVMe & DCPMM support
- Per-pool ACL
- UNS in DAOS via dfuse
- Replication & self-healing (Preview)

Application Interface:
- MPI-IO Driver
- HDF5 DAOS Connector
- Basic POSIX I/O support

Released on April 30, 2021

DAOS:
- + Leap 15 support
- End-to-end data integrity
- Per-container ACL
- Replication & self-healing
- Quiescent server reintegration
- Erasure Code (Preview)

Application Interface:
- HDF5 DAOS VOL Connector
- POSIX I/O improvements
- POSIX data mover
- Apache Spark support

DAOS:
- + CentOS8 support
- Erasure code
- Telemetry
- Improved usability

DAOS:
- Catastrophic recovery tools
- Multi OFI provider support
- Progressive layout
- Placement optimizations

DAOS:
- Checksum scrubbing
- Distributed transactions
- Lustre/UNS integration
- Online server addition/reintegration

Application Interface:
- POSIX I/O with distributed transaction support
- HDF5 data mover
- Container parking/serialization

NOTE: All information provided in this roadmap is subject to change without notice.

MadFS v2

咋做啊?

MadFS ~~v2~~

还是接着打榜吧

感谢

陈康

刘艺源

毛晗扬

HUAWEI

华为公司

深圳网络空间科学与技术省实验室
鹏城实验室
PCL
Peng Cheng Laboratory

鹏城实验室

感谢

tuna

和

清华大学学生
网络与开源软件协会

您