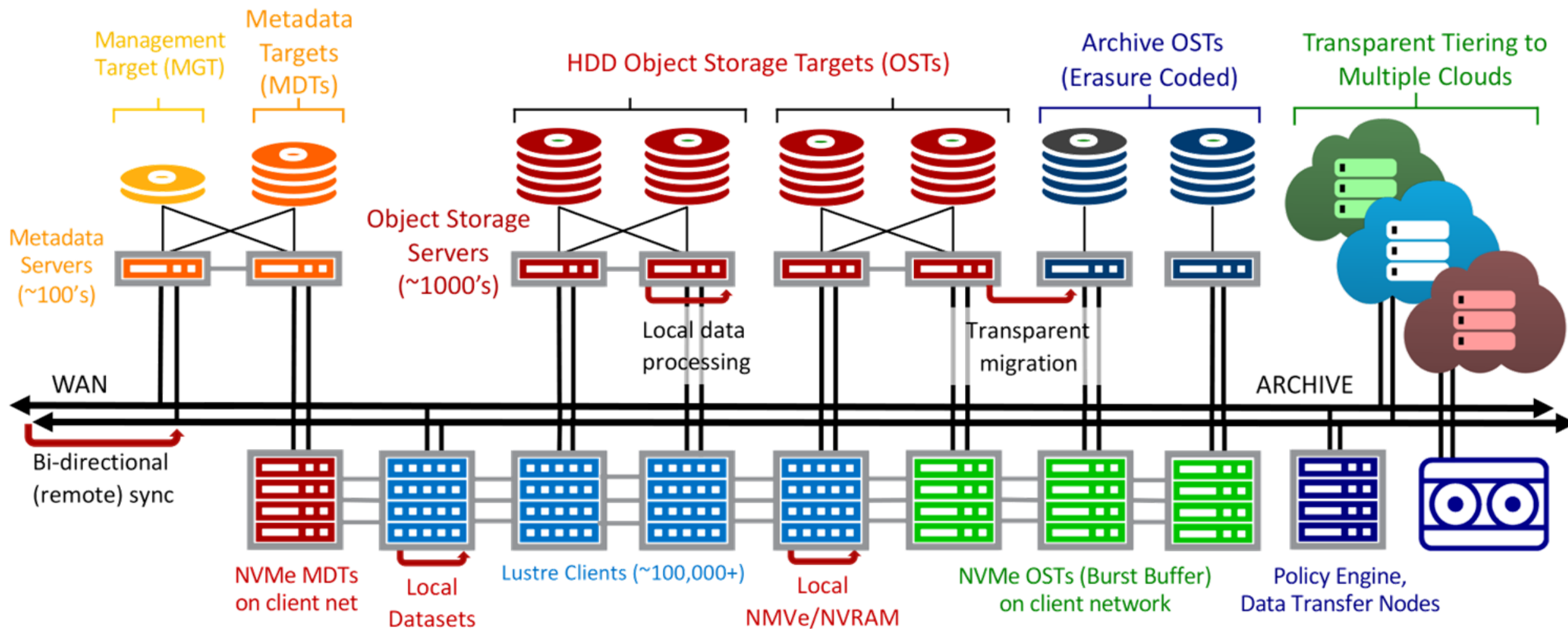


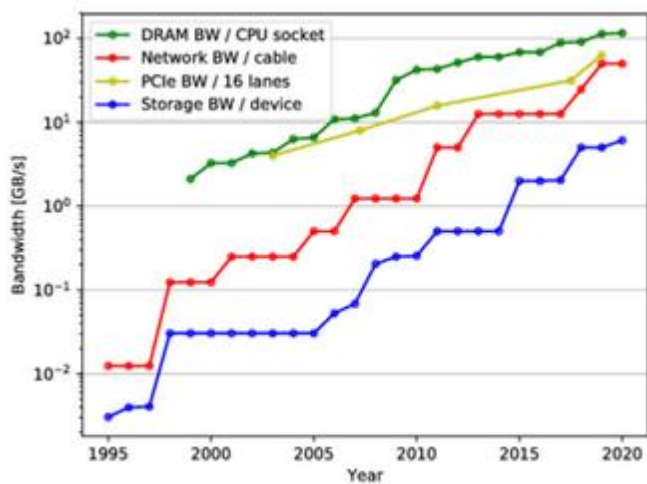
存储系统自底向上的性能测试

王 丙 强
鹏城实验室
CLUG2020

Lustre日趋复杂

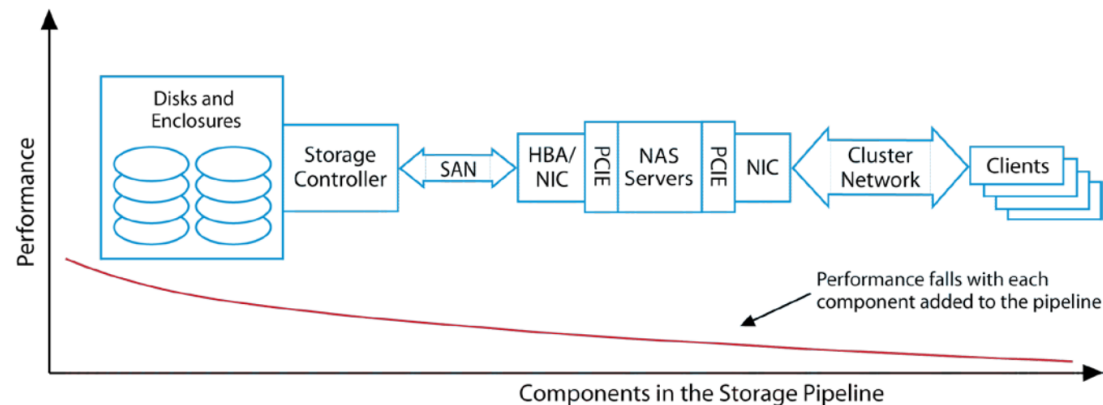


性能和可靠性挑战



• 性能挑战

- 传统科学计算的高带宽需求
- 人工智能等新领域的海量小文件吞吐
- 多用户带来的复杂使用场景
- 新的存储层次例如burst buffer的引入



• 可靠性挑战

- 架构设计上已重点考虑高可用
- 系统规模大组件众多
- 某些硬件故障往往被忽视
- 其引发的问题不易被定位

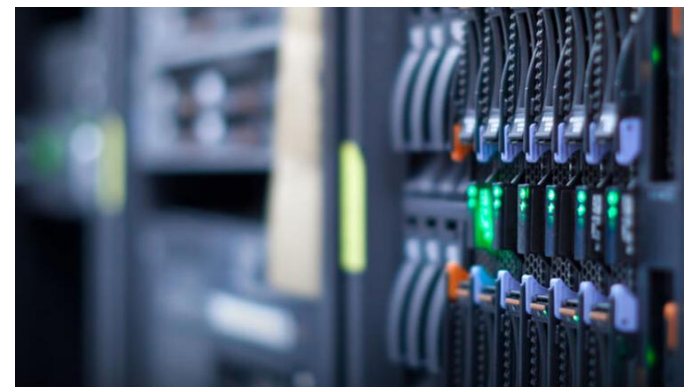
为什么要关注硬件

• 系统安装和配置时

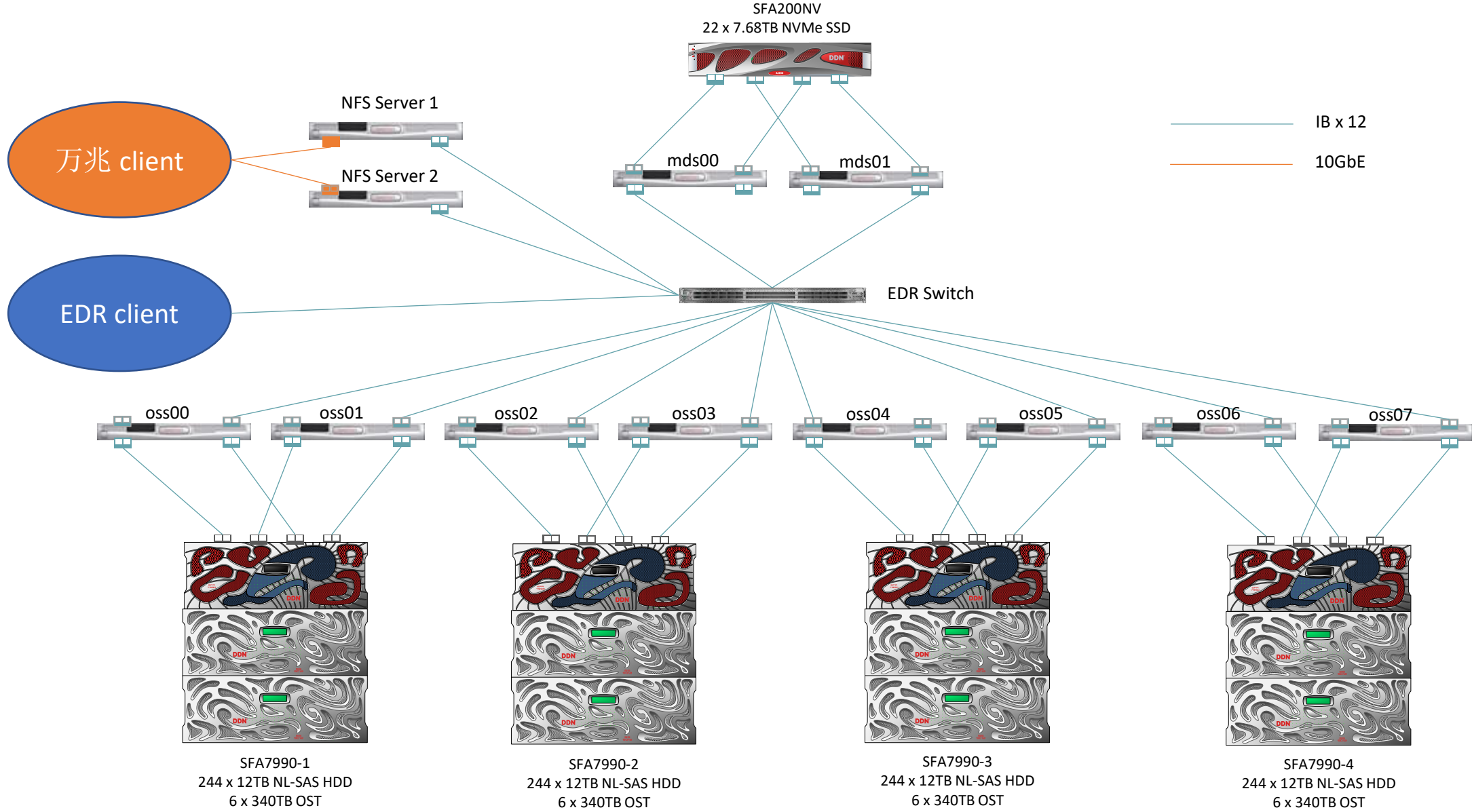
- 基本测试确认状态良好
- 压力测试确保设备稳定
- 将可能问题消灭在开始
- 提高交付和上线效率

• 系统运行时

- 实时监控设备状态
- 及时发现硬件故障
- 迅速定位解决问题
- 避免影响稳定运行



neuronfs文件系统



neuronfs硬件组成

存储	
MDT	1x SFA200NV
OST	4x SFA7990
网络	
IB EDR	100Gbps
10GbE	10Gbps
服务器	
MDS	2x
OSS	8x
NFS Gateway	2x

测什么？

- 功能
- 性能
- 稳定性

怎么测？

- 指标
- 工具
- 方法

服务器测试之前

系统BIOS设置

关闭超线程

关闭电源管理

操作系统设置

服务器测试

```
[root@oss1 ~]# ipmitool lan print 1
Set in Progress       : Set Complete
Auth Type Support     : NONE MD2 MD5 PASSWORD
Auth Type Enable      : Callback : MD2 MD5 PASSWORD
                      : User      : MD2 MD5 PASSWORD
                      : Operator  : MD2 MD5 PASSWORD
                      : Admin     : MD2 MD5 PASSWORD
                      : OEM       : MD2 MD5 PASSWORD
IP Address Source     : Static Address
IP Address            : 
Subnet Mask           : 
MAC Address           : 
SNMP Community String : public
IP Header             : TTL=0x00 Flags=0x00 Precedence=0x00 TOS=0x00
BMC ARP Control       : ARP Responses Enabled, Gratuitous ARP Disabled
Default Gateway IP    : 192.168.202.1
Default Gateway MAC   : 00:00:00:00:00:00
Backup Gateway IP     : 0.0.0.0
Backup Gateway MAC    : 00:00:00:00:00:00
802.1q VLAN ID        : Disabled
802.1q VLAN Priority  : 0
RMCP+ Cipher Suites   : 1,2,3,6,7,8,11,12
Cipher Suite Priv Max : XaaaXXaaaXXaaXX
                      : X=Cipher Suite Unused
                      : c=CALLBACK
                      : u=USER
                      : o=OPERATOR
                      : a=ADMIN
                      : O=OEM
Bad Password Threshold : 3
Invalid password disable: yes
Attempt Count Reset Int.: 300
User Lockout Interval  : 300
[root@oss1 ~]#
```

功能测试

不是每种服务器的IPMI都工作如常

```
[root@mds1 ~]# ipmitool lan print 1
Set in Progress       : Set Complete
IP Address Source     : 
IP Address            : 
Subnet Mask           : 
MAC Address           : 
SNMP Community String : TrapAdmin12#$
IP Header             : TTL=0x40 Flags=0x40 Precedence=0x00 TOS=0x10
Default Gateway IP    : 192.168.202.1
802.1q VLAN ID        : Disabled
RMCP+ Cipher Suites   : 0,1,2,3,17
Cipher Suite Priv Max : XuuaXXXXXXXXXX
                      : X=Cipher Suite Unused
                      : c=CALLBACK
                      : u=USER
                      : o=OPERATOR
                      : a=ADMIN
                      : O=OEM
Bad Password Threshold : Not Available
[root@mds1 ~]#
```

服务器测试

内存带宽测试

STREAM工具

两台MDS测试结果有较大差异

问题原因

比例大约3:2

确认有内存条没有插好

缺少了内存通道

```
-----
This system uses 8 bytes per array element.
-----
Array size = 800000000 (elements), Offset = 0 (elements)
Memory per array = 6103.5 MiB (= 6.0 GiB).
Total memory required = 18310.5 MiB (= 17.9 GiB).
Each kernel will be executed 30 times.
The *best* time for each kernel (excluding the first iteration)
will be used to compute the reported bandwidth.
-----
Number of Threads requested = 32
Number of Threads counted = 32
-----
Your clock granularity/precision appears to be 1 microseconds.
Each test below will take on the order of 113195 microseconds.
(= 113195 clock ticks)
Increase the size of the arrays if this shows that
you are not getting at least 20 clock ticks per test.
-----
WARNING -- The above is only a rough guideline.
For best results, please be sure you know the
precision of your system timer.
-----
Function      Best Rate MB/s  Avg time     Min time     Max time
Copy:          111744.7    0.141501     0.114547     0.172684
Scale:         112924.7    0.151207     0.113350     0.387974
Add:           117863.1    0.196528     0.162901     0.425280
Triad:         125445.9    0.197918     0.153054     0.409496
-----
Solution Validates: avg error less than 1.000000e-13 on all three arrays
-----
mds2
-----
STREAM version $Revision: 5.10 $
-----
This system uses 8 bytes per array element.
-----
Array size = 800000000 (elements), Offset = 0 (elements)
Memory per array = 6103.5 MiB (= 6.0 GiB).
Total memory required = 18310.5 MiB (= 17.9 GiB).
Each kernel will be executed 30 times.
The *best* time for each kernel (excluding the first iteration)
will be used to compute the reported bandwidth.
-----
Number of Threads requested = 32
Number of Threads counted = 32
-----
Your clock granularity/precision appears to be 1 microseconds.
Each test below will take on the order of 149324 microseconds.
(= 149324 clock ticks)
Increase the size of the arrays if this shows that
you are not getting at least 20 clock ticks per test.
-----
WARNING -- The above is only a rough guideline.
For best results, please be sure you know the
precision of your system timer.
-----
Function      Best Rate MB/s  Avg time     Min time     Max time
Copy:          69896.1    0.221215     0.183129     0.275000
Scale:         70419.8    0.215230     0.181767     0.261435
Add:           77114.6    0.295590     0.248980     0.511592
Triad:         77609.9    0.279005     0.247391     0.360050
-----
Solution Validates: avg error less than 1.000000e-13 on all three arrays
-----
```

服务器测试

CPU压力测试

LINPACK工具

两台MDS实测性能差距大于10%

八台OSS实测波动小于0.3%

Performance Summary (GFlops)

Size	LDA	Align.	Average	Maximal
1000	1000	4	109.2182	134.5450
2000	2000	4	324.9850	326.4865
5000	5008	4	769.1879	770.0566
10000	10000	4	1170.0965	1171.2522
15000	15000	4	1308.8422	1314.0134
18000	18008	4	1345.2404	1347.0602
20000	20016	4	1325.2502	1326.5203
22000	22008	4	1353.8134	1354.4346
25000	25000	4	1364.7605	1366.0153
26000	26000	4	1390.4541	1390.8984
27000	27000	4	1394.1431	1394.1431
30000	30000	1	1399.6139	1399.6139
35000	35000	1	1404.2392	1404.2392
40000	40000	1	1421.0432	1421.0432
45000	45000	1	1413.5212	1413.5212

Residual checks PASSED

End of tests

Done: Wed Apr 8 22:48:53 CST 2020
[root@mds1 linpack]#

Size	LDA	Align.	Average	Maximal
1000	1000	4	108.3144	131.4180
2000	2000	4	331.2943	334.2311
5000	5008	4	720.9302	731.4904
10000	10000	4	1076.9516	1077.0350
15000	15000	4	1196.3285	1197.9165
18000	18008	4	1218.0870	1219.6041
20000	20016	4	1179.8618	1186.2357
22000	22008	4	1210.6233	1212.8187
25000	25000	4	1221.0362	1221.7839
26000	26000	4	1240.2725	1241.1304
27000	27000	4	1242.5168	1242.5168
30000	30000	1	1242.9255	1242.9255
35000	35000	1	1248.1088	1248.1088
40000	40000	1	1260.9434	1260.9434
45000	45000	1	1254.0228	1254.0228

Residual checks PASSED

End of tests

Done: Wed Apr 8 22:49:53 CST 2020
[root@mds2 linpack]#

网络测试

测试工具和方法

OSU Micro Benchmarks

测量随机对之间的带宽和延迟

网络轻载或者空载情况下，一致性应该非常好

发现与解决问题

个别IB光纤工作在FDR速率(56Gbps)下

要求全部更换为Mellanox原厂IB EDR光纤

对网络进行压力测试

网络状态直接影响到存储系统稳定性

常常被忽视的影响因素

模仿真实的存储网络负载

主要测试长期运行的稳定性

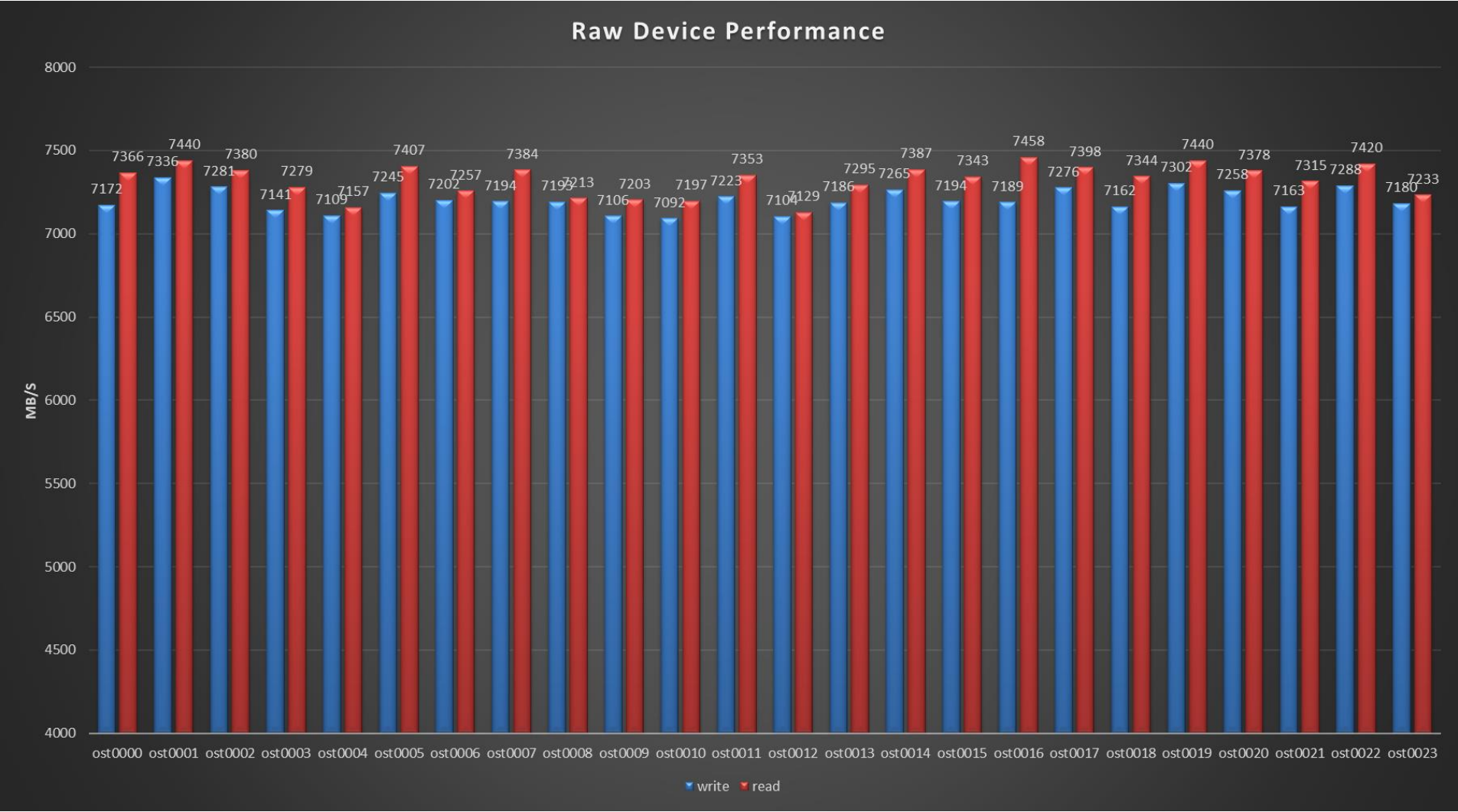
HPL对网络的压力有限

HPCG和Graph500可能是更合理的选择？

存储测试

OST性能测试结果

单位MB/s



存储测试

元数据访问性能测试结果

```
root@CPUA4-1:~# /applications/benchbin/openmpi403/bin/mpirun --prefix /applications/benchbin/openmpi403/ --mca btl_openib_allow_ib true --mca btl_openib_if_include mlx5_0:1 --allow-run-as-root -n
480 --hostfile hostfile_10node48ppn /applications/benchbin/ior_w_openmpi403/bin/mdtest -u -n 8000 -F -d
/mnt/neuronfs/dir1/@/mnt/neuronfs/dir2/@/mnt/neuronfs/dir3/@/mnt/neuronfs/dir4/@/mnt/neuronfs/dir5/@/mnt/neuronfs/dir6/@/mnt/neuronfs/dir7/@/mnt/neuronfs/dir8/
-- started at 04/20/2020 00:22:29 --

mdtest-1.9.3 was launched with 480 total task(s) on 10 node(s)
Command line used: /applications/benchbin/ior_w_openmpi403/bin/mdtest "-u" "-n" "8000" "-F" "-d"
"/mnt/neuronfs/dir1/@/mnt/neuronfs/dir2/@/mnt/neuronfs/dir3/@/mnt/neuronfs/dir4/@/mnt/neuronfs/dir5/@/mnt/neuronfs/dir6/@/mnt/neuronfs/dir7/@/mnt/neuronfs/dir8/"
Path: /mnt/neuronfs/dir1
FS: 8233.6 TiB   Used FS: 0.0%   Inodes: 32760.2 Mi   Used Inodes: 0.0%

480 tasks, 3840000 files

SUMMARY rate: (of 1 iterations)
  Operation      Max           Min           Mean           Std Dev
  -----
File creation    : 230389.246    230389.246    230389.246      0.000
File stat       : 427271.569    427271.569    427271.569      0.000
File read        : 241235.133    241235.133    241235.133      0.000
File removal     : 255399.012    255399.012    255399.012      0.000
Tree creation    : 128.284       128.284       128.284         0.000
Tree removal     : 11.933        11.933        11.933         0.000

-- finished at 04/20/2020 00:23:26 --
```

达到427K/s

存储测试

单节点访问带宽测试结果

```
root@CPUA4-1:~# time iofzone -I -i 0 -i 1 -r 16m -s 32g -t 24 -n -F /mnt/neuronfs/iz01 /mnt/neuronfs/iz02 /mnt/neuronfs/iz03 /mnt/neuronfs/iz04 /mnt/neuronfs/iz05 /mnt/neuronfs/iz06
Iofzone: Performance Test of File I/O
Version $Revision: 3.429 $
Compiled for 64 bit mode.
Build: linux-AMD64

Contributors:William Norcott, Don Capps, Isom Crawford, Kirby Collins
Al Slater, Scott Rhine, Mike Wisner, Ken Goss
Steve Landherr, Brad Smith, Mark Kelly, Dr. Alain CYR,
Randy Dunlap, Mark Montague, Dan Million, Gavin Brebner,
Jean-Marc Zucconi, Jeff Blomberg, Benny Halevy, Dave Boone,
Erik Habbinga, Kris Strecker, Walter Wong, Joshua Root,
Fabrice Bacchella, Zhenghua Xue, Qin Li, Darren Sawyer,
Vangel Bojaxhi, Ben England, Vikentsi Lapa.

Run began: Sun Apr 19 16:23:33 2020

O_DIRECT feature enabled
Record Size 16384 kB
File size set to 33554432 kB
No retest option selected
Command line used: iofzone -I -i 0 -i 1 -r 16m -s 32g -t 24 -n -F /mnt/neuronfs/iz01 /mnt/neuronfs/iz02 /mnt/neuronfs/iz03 /mnt/neuronfs/iz04 /mnt/neuronfs/iz05 /mnt/neuronfs/iz06
Output is in kBytes/sec
Time Resolution = 0.000001 seconds.
Processor cache size set to 1024 kBytes.
Processor cache line size set to 32 bytes.
File stride size set to 17 * record size.
Throughput test with 24 processes
Each process writes a 33554432 kByte file in 16384 kByte records

Children see throughput for 24 initial writers = 11125707.00 kB/sec
Parent sees throughput for 24 initial writers = 8842593.05 kB/sec
Min throughput per process = 389401.34 kB/sec
Max throughput per process = 550291.19 kB/sec
Avg throughput per process = 463571.12 kB/sec
Min xfer = 23756800.00 kB
```

达到11GB/s

测试总结

问题类型多样

事先无法预料

更为规范和系统的测试方法？

集成商和厂商以及客户的有效协作模式？

作用和意义

磨刀不误砍柴工

迅速发现并定位了硬件故障

推动了高效率高质量的交付

遵循良好的工程实践方法仍然极其重要

IO500测试

```
[Summary] Results files in /mnt/neuronfs/october/toaster/results/2020.05.16-14.56.51
[Summary] Data files in /mnt/neuronfs/october/toaster/datafiles/io500.2020.05.16-14.56.51
[RESULT] BW    phase 1          ior_easy_write          46.719 GB/s : time 301.00 seconds
[RESULT] BW    phase 2          ior_hard_write          0.073 GB/s : time 1033.48 seconds
[RESULT] BW    phase 3          ior_easy_read          41.032 GB/s : time 342.67 seconds
[RESULT] BW    phase 4          ior_hard_read          0.654 GB/s : time 115.68 seconds
[RESULT] IOPS phase 1          mdtest_easy_write      150.855 kiops : time 331.42 seconds
[RESULT] IOPS phase 2          mdtest_hard_write      71.388 kiops : time 302.58 seconds
[RESULT] IOPS phase 3          find                    1538.440 kiops : time 46.54 seconds
[RESULT] IOPS phase 4          mdtest_easy_stat       257.597 kiops : time 194.09 seconds
[RESULT] IOPS phase 5          mdtest_hard_stat       226.377 kiops : time 95.42 seconds
[RESULT] IOPS phase 6          mdtest_easy_delete     101.441 kiops : time 492.86 seconds
[RESULT] IOPS phase 7          mdtest_hard_read       89.394 kiops : time 241.63 seconds
[RESULT] IOPS phase 8          mdtest_hard_delete     38.068 kiops : time 568.85 seconds
[SCORE] Bandwidth 3.09495 GB/s : IOPS 155.021 kiops : TOTAL 21.9039
```

hosts = 18, ppn = 4

IO500测试

► Hero bandwidth

- Write and read

► Anti-hero bandwidth

- Write and read

► Hero metadata

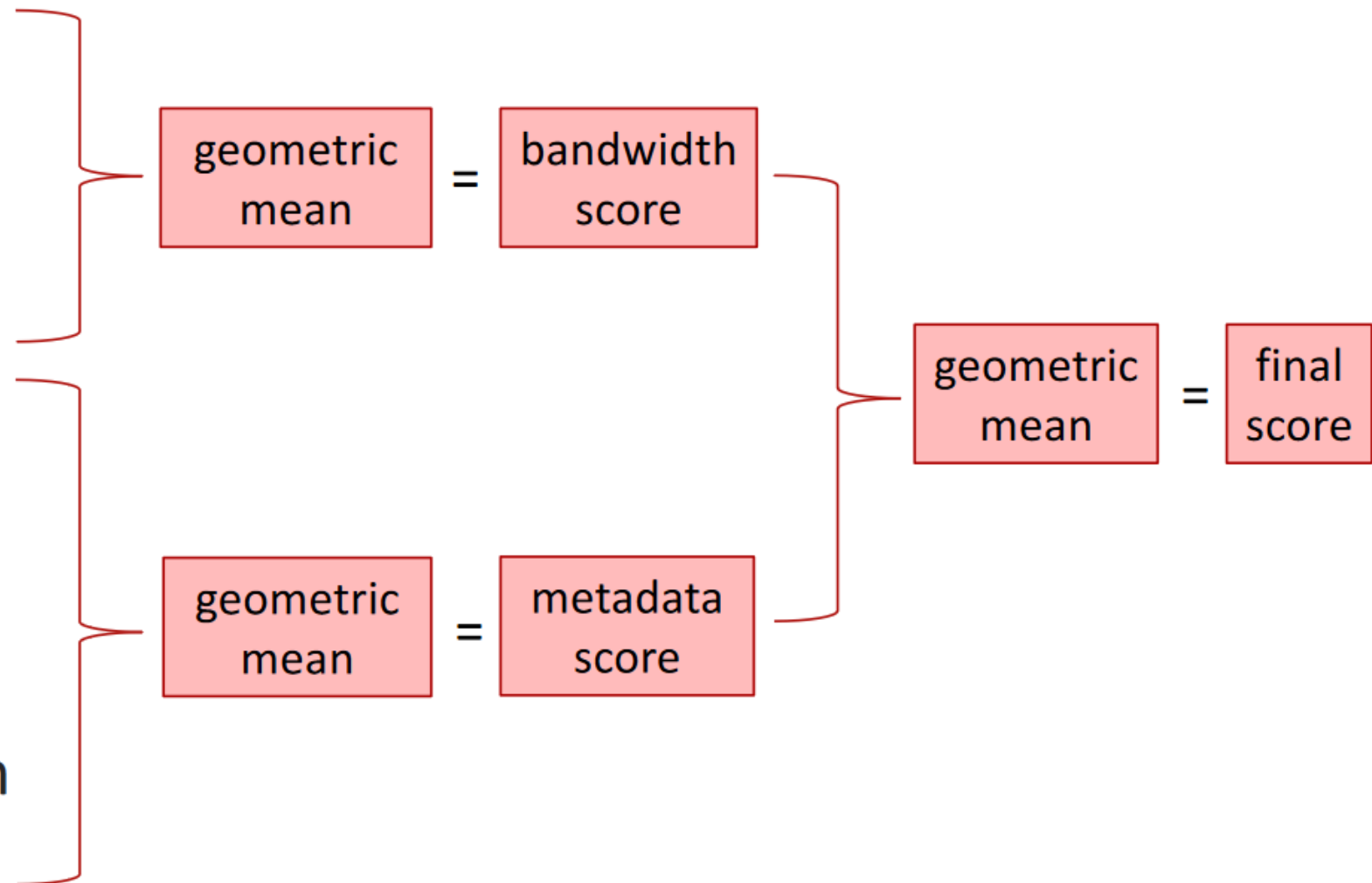
- Create, stat, delete

► Anti-hero metadata

- Create, stat, read, delete

► And a namespace search

- Search created files



IO500测试

体会

对不同情况下的带宽和吞吐充分测试

最终得分取几何平均

不能有明显短板

能否反映人工智能应用的存储压力？

作用和意义

存储系统的性能调优极为复杂

推动测试标准的统一和规范

扩展性问题

分布式锁操作

MPI-IO模式

放宽POSIX语义

规模问题

不限制系统规模

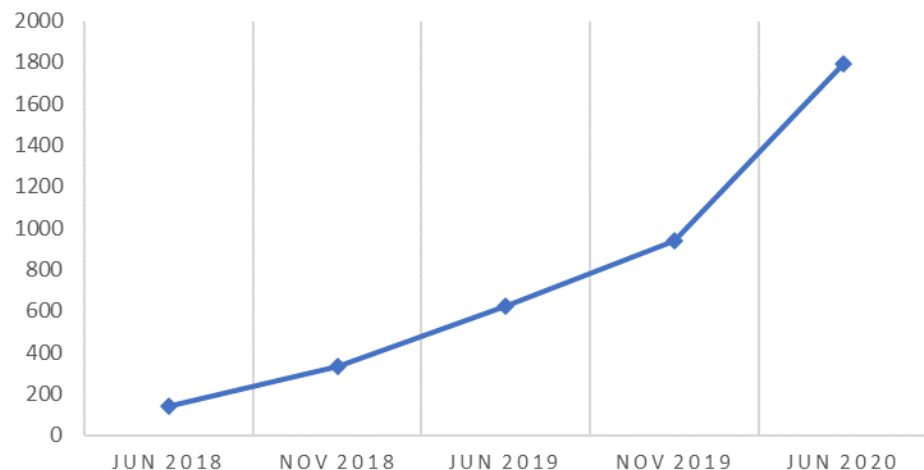
也鼓励小系统提交10-node成绩

最新IO500榜单

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	klOP/s
1	isc20	Intel	Wolf	Intel	DAOS	52	1664	zip	1792.98	371.67	8649.57
2	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
3	isc20	TACC	Frontera	Intel	DAOS	60	1440	zip	763.80	78.31	7449.56
4	isc20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	zip	537.31	108.19	2668.57
5	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	453.68	209.43	982.78
6	isc20	KISTI	NURION	DDN	IME	2048	2048	zip	282.45	515.59	154.74
7	isc20	Oracle Cloud Infrastructure	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS	270	3240	zip	267.25	293.05	243.73
8	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
9	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	229.45	131.25	401.13
10	sc19	CEA	Tera-1000	DDN	Lustre	128	4096	zip	210.26	81.01	545.74
11	isc20	JCAHPC	Oakforest-PACS	DDN		512	4096	zip	175.85	348.95	88.62
12	isc20	EPCC	NextGENIO	NextGenIO (BSC & JGU)	GekkoFS	10	280	zip	152.30	24.83	934.12
13	isc20	DDN	AI400X	DDN	Lustre	40	1280	zip	137.03	46.84	400.84
14	isc20	CSIRO	bracwell	Dell/DDN	Lustre	10	400	zip	99.64	39.90	248.81
15	isc20	Johannes Gutenberg University Mainz	MOGON II	JGU (ADA-FS)& BSC (NEXTGenIO)	GekkoFS	10	160	zip	96.46	11.61	801.60

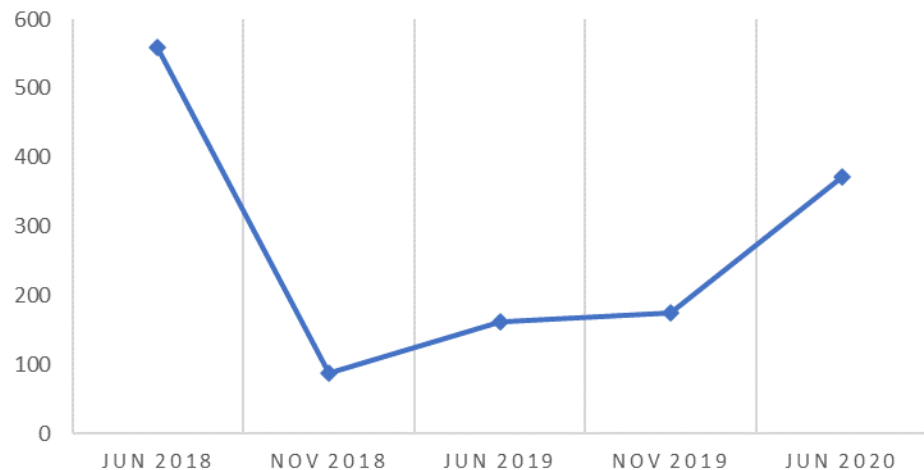
IO500发展趋势

SCORE

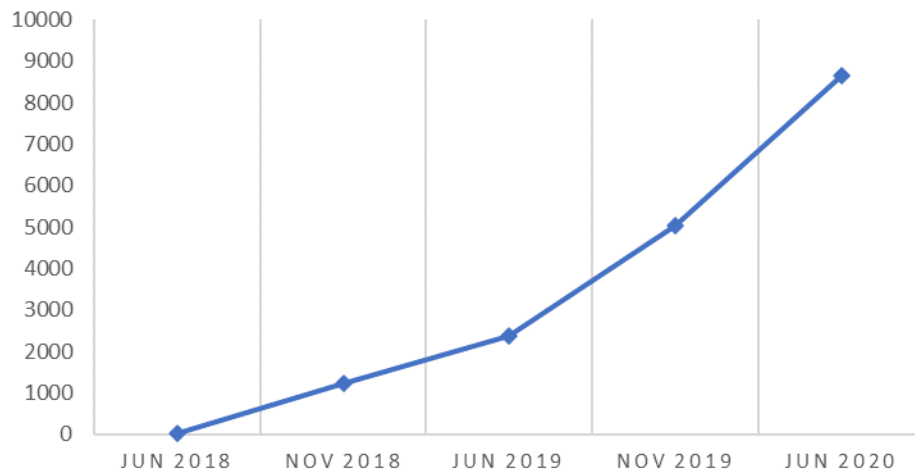


- IO500冠军跑分在稳步提升
- 带宽得分增长不多
- 元数据得分增长趋势明显
- 冠军成绩每半年翻一番

BW



MD



问题思考

- 面向存储系统的硬件测试方法
- 真实反映实际使用场景的负载模拟
- 面向具体使用场景的综合优化策略
- 多级缓存架构的部署和整合

致谢

DDN的同仁

上海交通大学、中国科学技术大学和南京大学的同行朋友



**谢谢
欢迎指正**