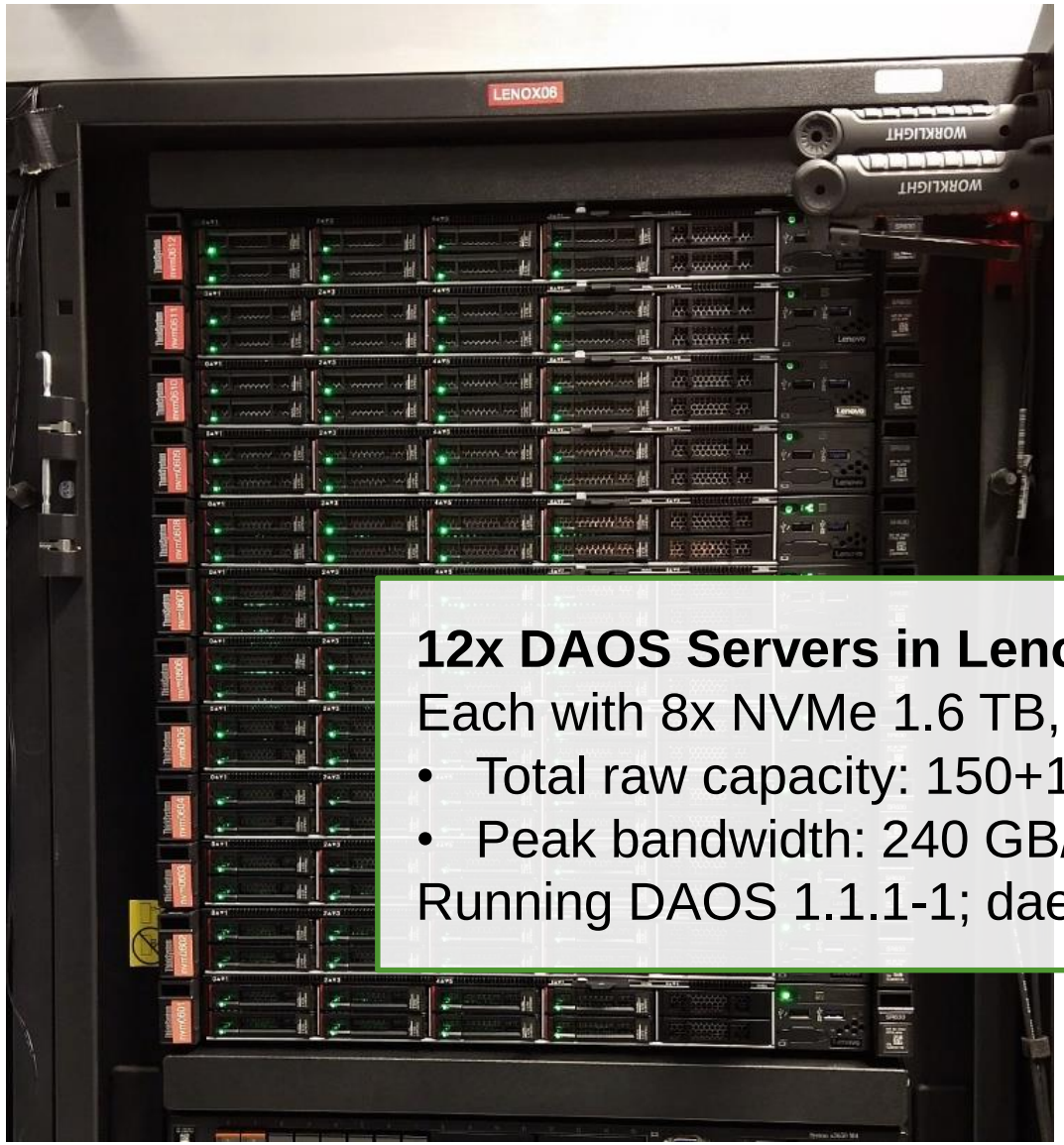
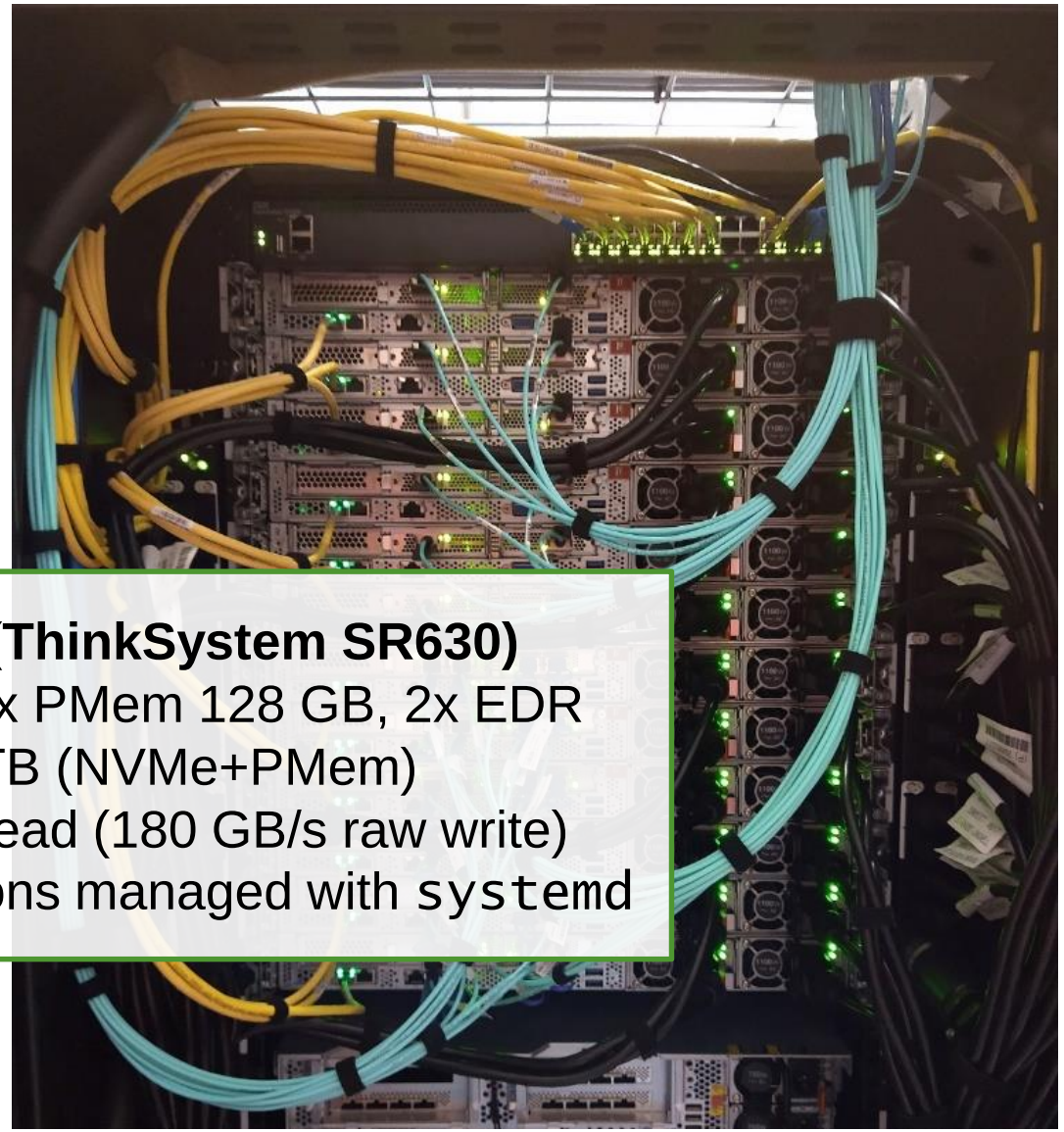
DAOS in Lenovo's HPC Innovation Center

Michael Hennecke | DUG'20, 19-Nov-2020

DAOS in Lenovo‘s HPC Innovation Center Stuttgart



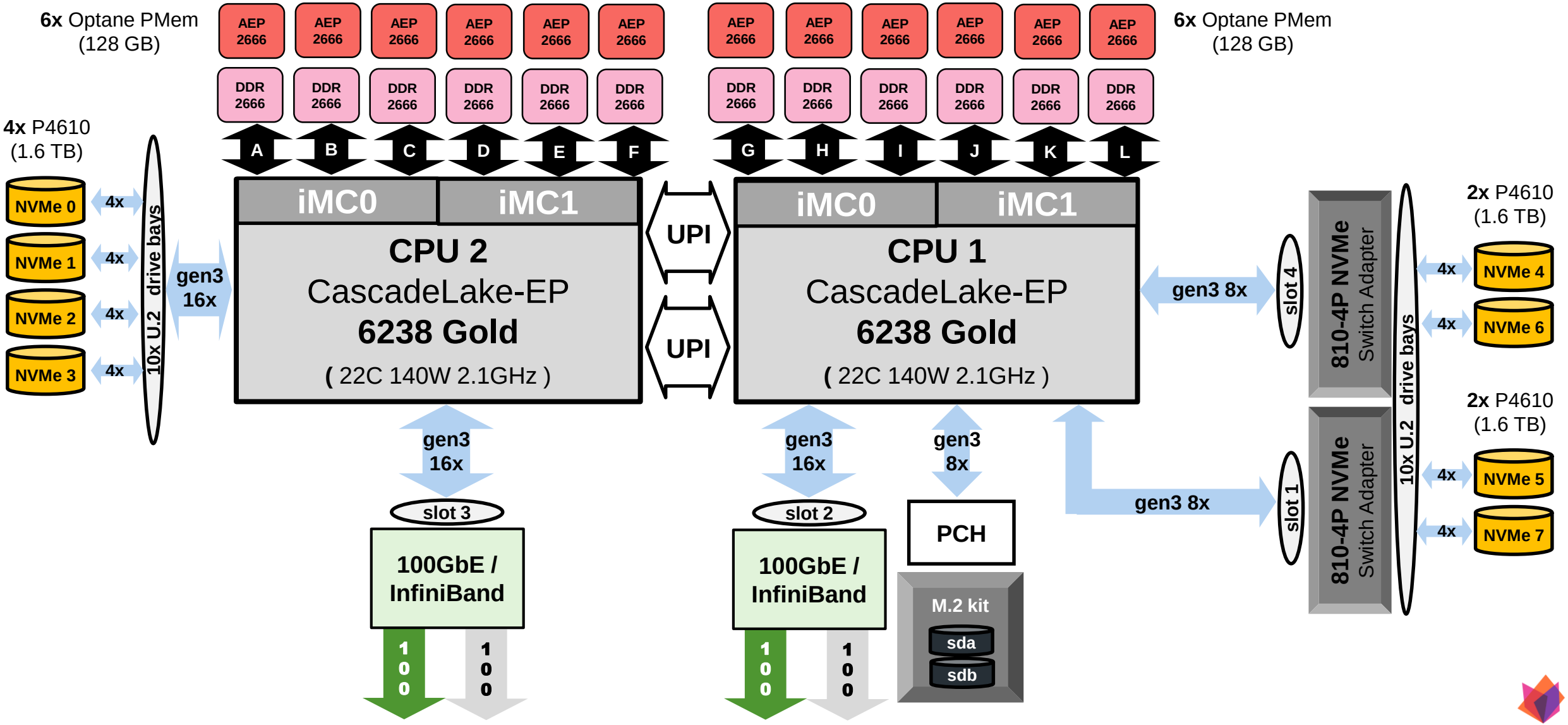


12x DAOS Servers in Lenox (ThinkSystem SR630)
Each with 8x NVMe 1.6 TB, 12x PMem 128 GB, 2x EDR

- Total raw capacity: 150+18 TB (NVMe+PMem)
- Peak bandwidth: 240 GB/s read (180 GB/s raw write)

Running DAOS 1.1.1-1; daemons managed with `systemd`

DAOS Server Architecture: Lenovo ThinkSystem SR630



SC20 Everywhere we are | more than hpc.

Lenovo‘s Slurm Integration: Ephemeral DAOS Containers

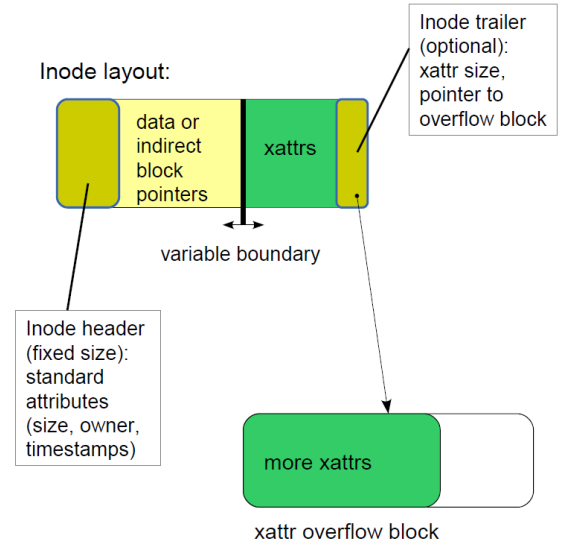
1. In `slurm.conf`, add: `Licenses=daos_mlx:1,daos_dev1_mlx:1,daos_dev2_mlx:1`
   – Define multiple DAOS clusters as Slurm „Licenses“, to manage dedicated batch job access
2. User jobs request a DAOS cluster, e.g. `#SBATCH --licenses=daos_mlx`
3. Slurm `Prolog` and `Epilog` call `daos_mounts.sh {prolog|epilog}`
4. Lenovo‘s `daos_mounts.sh prolog`:
   – Creates YML config files for the requested Slurm license, and starts `daos_agent`
   – On 1st node, provisions a DAOS pool and DAOS POSIX container for the user
   – On 1st node, saves pool and container UUIDs in per-job dotfiles (in user‘s `$HOME/.daos/`)
   – Dfuse-mounts the POSIX container for the user
5. Lenovo‘s `daos_mounts.sh epilog`:
   – Umounts the POSIX container
   – On 1st node, destroys the pool, including its container(s)
   – On 1st node, removes the per-job dotfiles (from `$HOME/.daos/`), and stops `daos_agent`

Lenovo‘s Slurm Integration: daos_mounts.sh prolog

```
# On the first node in the job (Slurm BatchHost):
dmg pool create -s 768g -n 6390g -u $D_USER -g $D_GROUP \
 > /tmp/dmg-pool-create.$D_USER.log 2>&1

LINE=`grep $UUID_MATCH /tmp/dmg-pool-create.$D_USER.log`
[[ $LINE =~ $CREATE_MATCH ]]
D_POOL=${BASH_REMATCH[1]}
D_SVC=${BASH_REMATCH[2]}

dmg pool update-acl --pool=$D_POOL -e 'A::root@:rw'  #  also -e 'A::daos@:rw‘

D_CONT=`uuidgen`
su - $D_USER \
 -c "daos cont create --pool=$D_POOL --svc=$D_SVC --cont=$D_CONT --type=POSIX“

# On all nodes in the job:
su - $D_USER \
 -c "dfuse --pool $D_POOL --svc=$D_SVC --container $D_CONT -m /daos/$D_USER"
```

DAOS Unified Namespace with Spectrum Scale (1/2)

- DAOS „Unified Namespace“ Concept:
  1. Store DAOS *pool UUID* and *container UUID* as extended attributes (XATTR‘s) of the „mount point“ directory
  2. When this „mount point“ is in a global parallel filesystem, `dfuse` can use this instead of --pool and --container
- IBM Spectrum Scale supports Extended Attributes (XATTR‘s), both for internal features and for user metadata
  - Stored in inode if small, or in „overflow“ EA block (≤64 kiB)




```
$ daos cont create --pool=$D_POOL --svc=$D_SVC --cont=$D_CONT \
    --type=POSIX --path /home/mhennecke/daos_tmp
```

```
$ mmlsattr --dump-attr /home/mhennecke/daos_tmp
file name: /home/mhennecke/daos_tmp
user.daos
$ mmlsattr --get-attr user.daos /home/mhennecke/daos_tmp
file name: /home/mhennecke/daos_tmp
user.daos: "DAOS.POSIX://c0c99a8c-5453-4950-9bbd-1d9d784b51c0/7b6ff2f2-b52d-4a25-8565-285006572c96?“
```

?

DAOS Unified Namespace with Spectrum Scale (2/2)

- **daos** can query the Spectrum Scale mountpoint directory‘s XATTR‘s on each node where the „containing“ Spectrum Scale filesystem is mounted:

```
$ daos cont query --path /home/mhennecke/daos_tmp --svc 0
Pool UUID:      c0c99a8c-5453-4950-9bbd-1d9d784b51c0
Container UUID: 7b6ff2f2-b52d-4a25-8565-285006572c96
Number of snapshots: 0
Latest Persistent Snapshot: 0
Highest Aggregated Epoch: 16057835397947207 68
DAOS Unified Namespace Attributes on path /home/mhennecke/daos_tmp:
Container Type: POSIX
Object Class:   SX
Chunk Size:     1048576
```

- The **dfuse** mount command can use the path without `--pool` and `--cont`:

```
$ dfuse -m /home/mhennecke/daos_tmp --svc 0
$ df|grep daos
dfuse    13980468750    727    13980468024    1% /gpfs/gss1/home/mhennecke/daos_tmp
```

DAOS Unified Namespace with Spectrum Scale (2/2)

- **daos** can query the Spectrum Scale mountpoint directory‘s XATTR‘s on each node where the „containing“ Spectrum Scale filesystem is mounted:

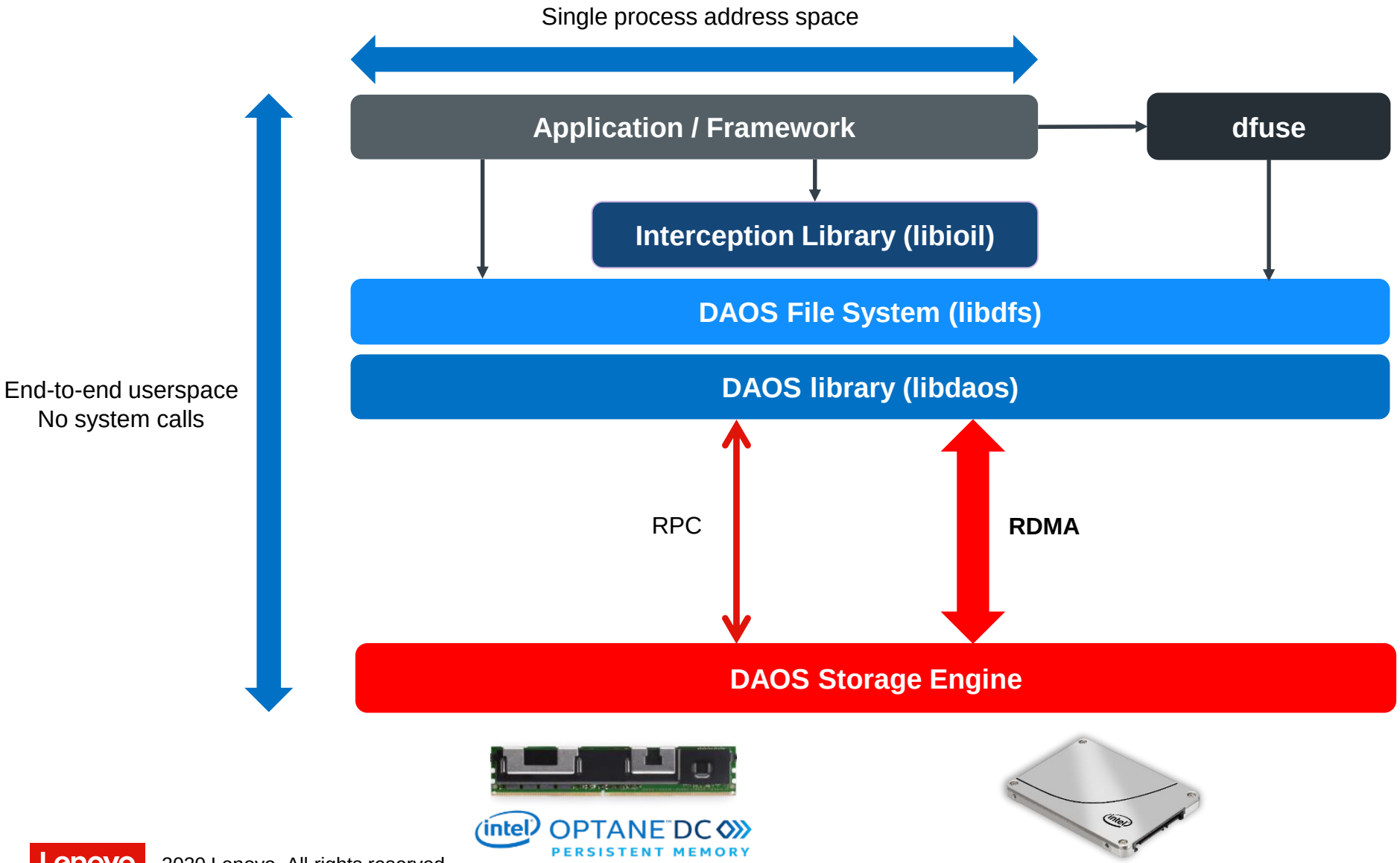
```
$ daos cont query --path /home/mhennecke/daos_tmp --svc 0
Pool UUID:      c0c99a8c-5453-4950-9bbd-1d9d784b51c0
Container UUID: 7b6ff2f2-b52d-4a25-8565-285006572c96
Number of snapshots: 0
Latest Persistent Snapshot: 0
Highest Aggregated Epoch: 16057835397947207 68
DAOS Unified Namespace Attributes on path /home/mhennecke/daos_tmp:
Container Type: POSIX
Object Class:   SX
Chunk Size:     1048576
```

The `fusermount3 -u` (unmount) fails if the Spectrum Scale file system is **mounted with root-squash**...

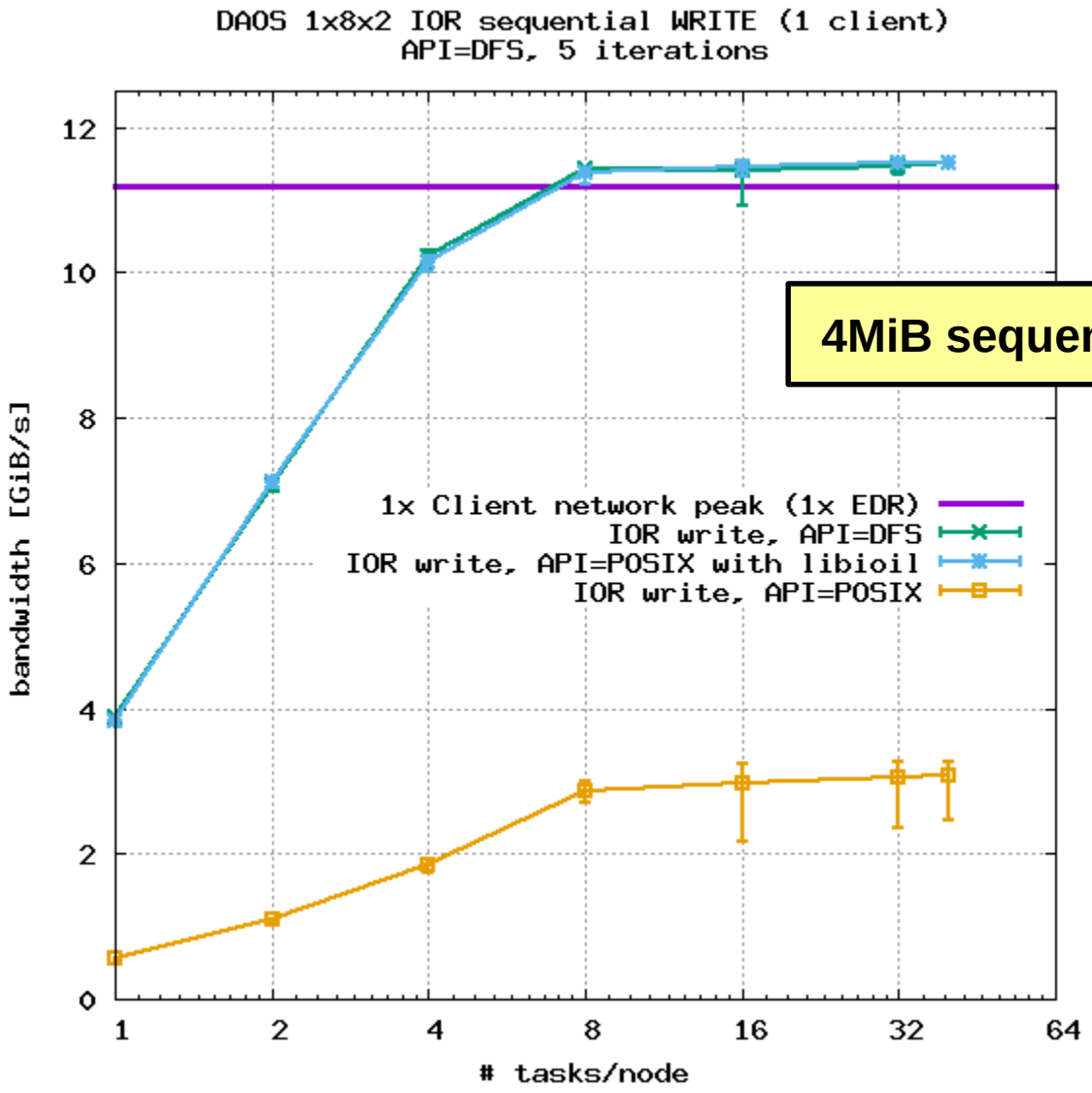
- The **dfuse** mount command can use the path without `--pool` and `--cont`:

```
$ dfuse -m /home/mhennecke/daos_tmp --svc 0
$ df|grep daos
dfuse    13980468750    727    13980468024    1% /gpfs/gss1/home/mhennecke/daos_tmp
```

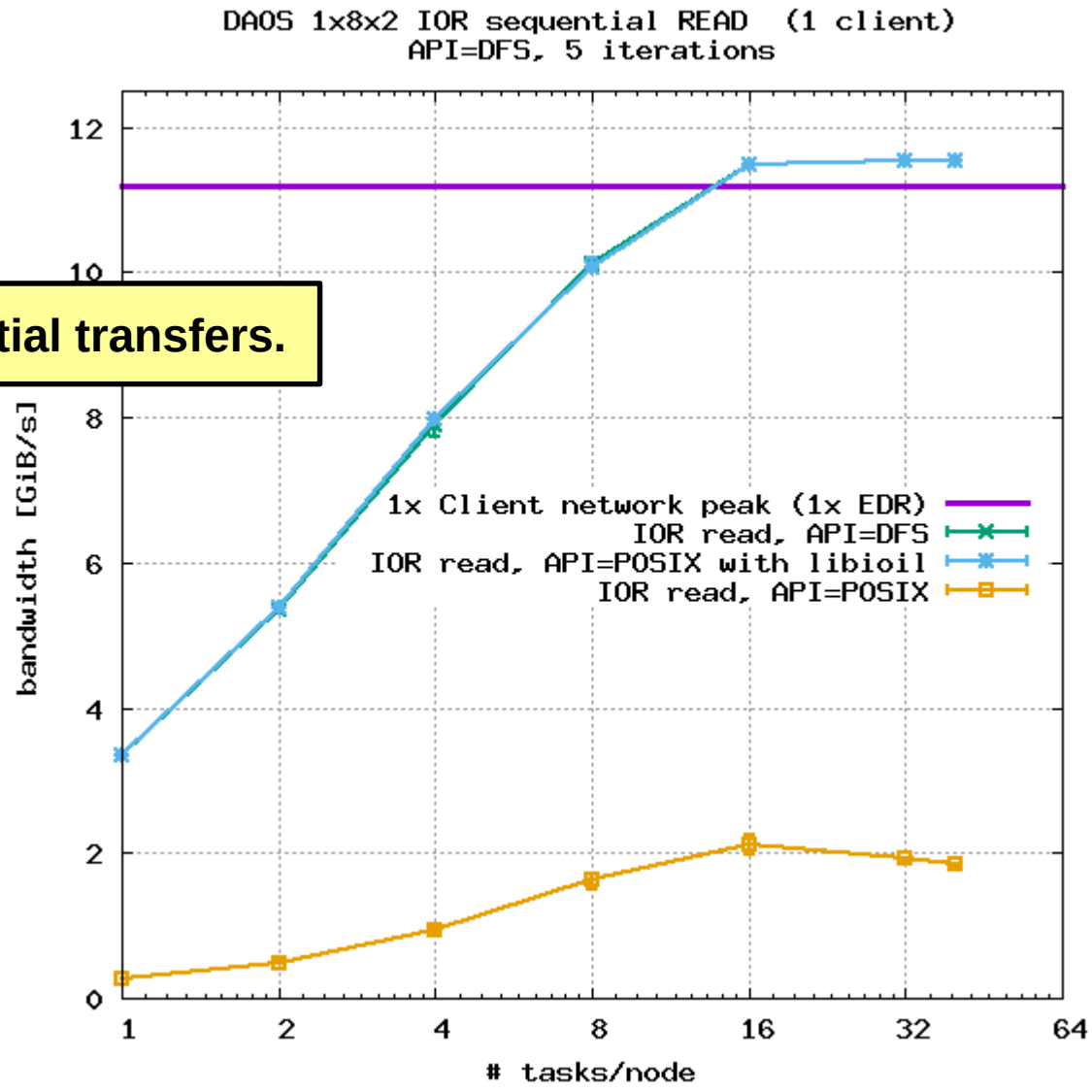
Three Ways of POSIX Filesystem Support in DAOS



DAOS ior-easy – 1 Server (P4610 3.2TB), 1 Client

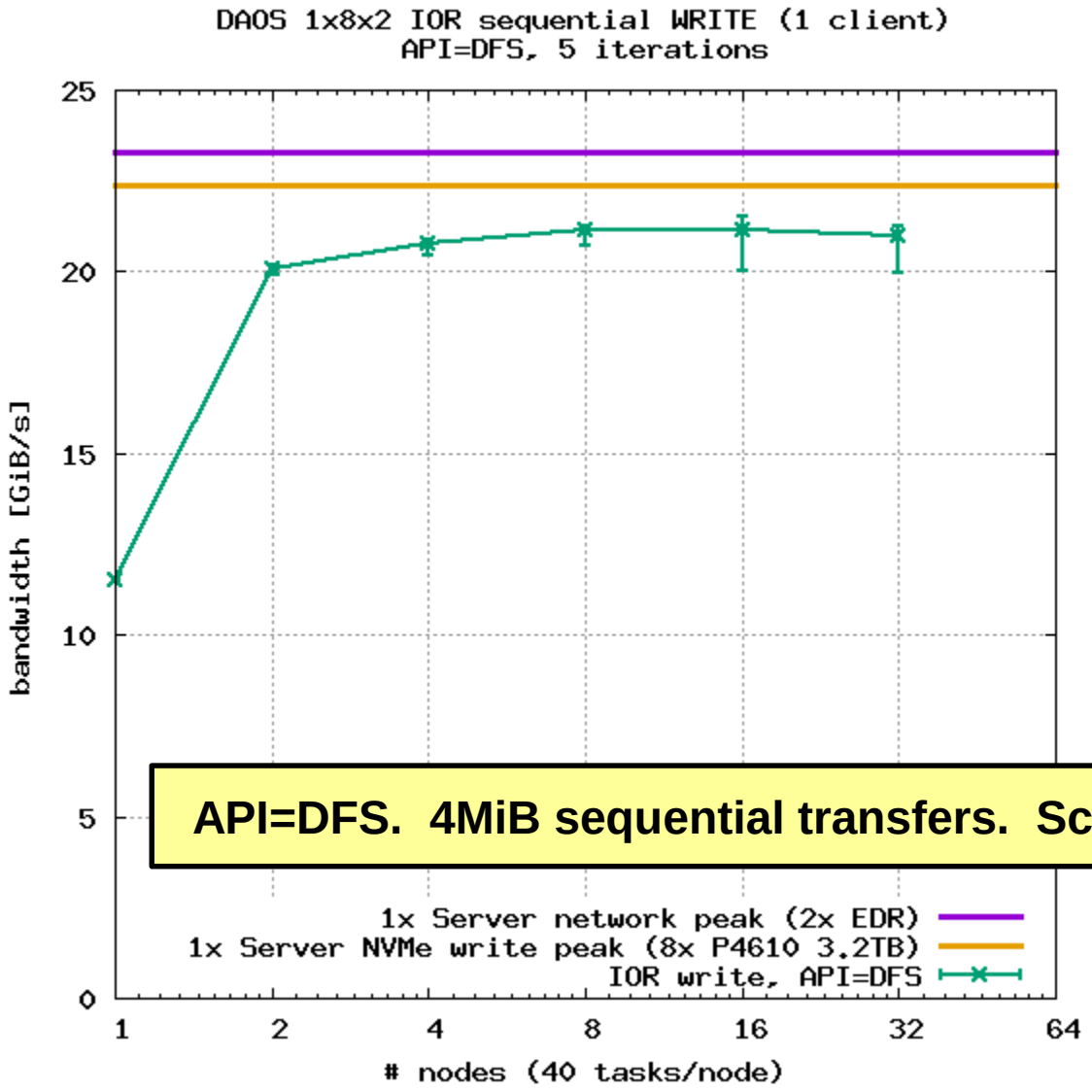


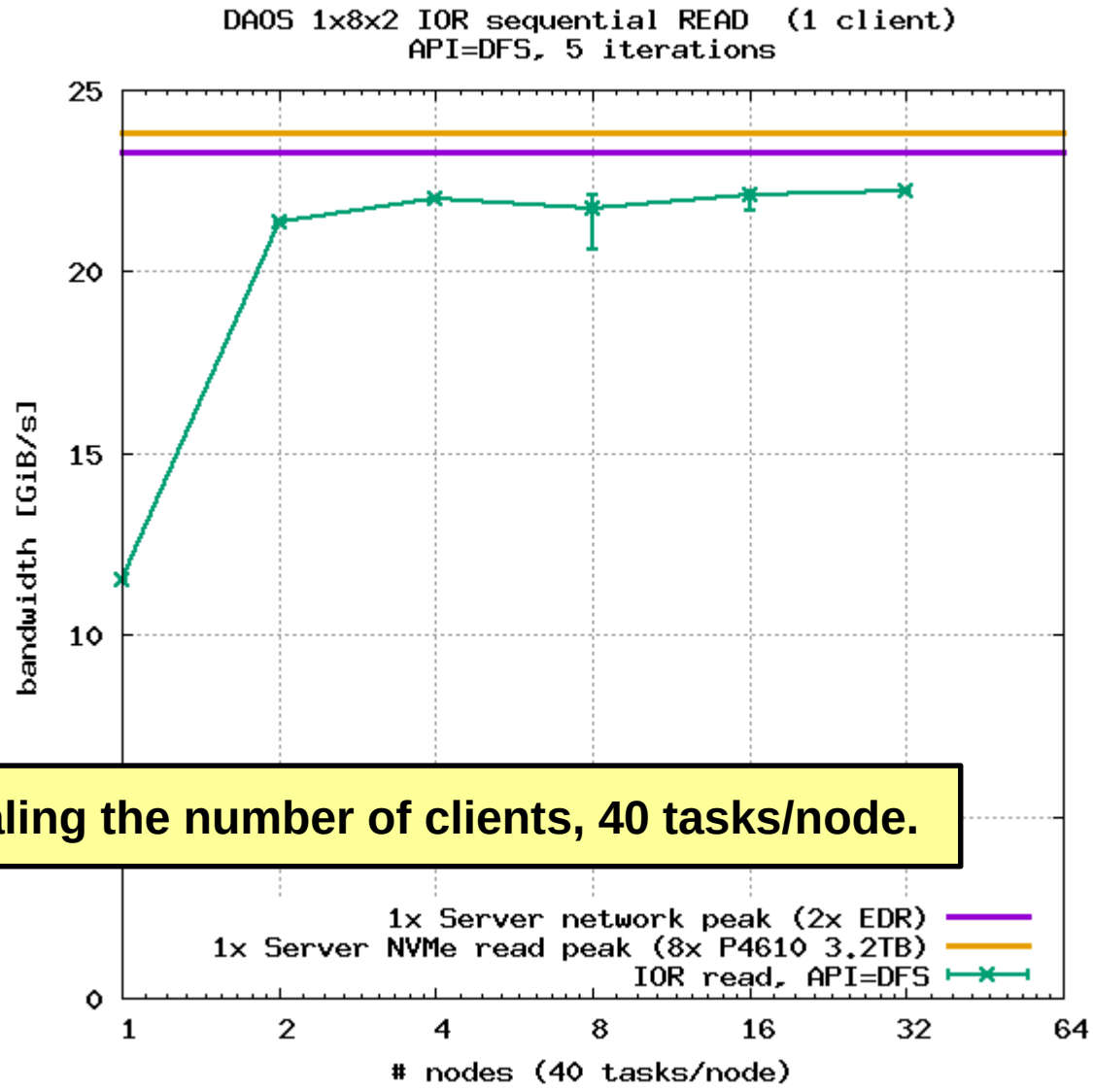
4MiB sequential transfers.



SC20 Everywhere we are | more than hpc.

DAOS ior-easy – 1 Server (P4610 3.2TB), N Clients





API=DFS. 4MiB sequential transfers. Scaling the number of clients, 40 tasks/node.

DAOS IO500 – 1 Server (P4610 3.2TB), 10 Clients

IO500 with API=DFS, and Mohamad's DFS-enabled `find` from `mpifileutils`:

```
IO500 version io500-sc20_v3
[RESULT]       ior-easy-write      20.754597 GiB/s : time 315.288 seconds
[RESULT]    mdtest-easy-write     586.050492 kIOPS : time 308.214 seconds
[RESULT]       ior-hard-write       8.015282 GiB/s : time 316.094 seconds
[RESULT]    mdtest-hard-write     120.679218 kIOPS : time 320.813 seconds
[RESULT]                 find     328.553089 kIOPS : time 657.437 seconds
[RESULT]        ior-easy-read      21.865060 GiB/s : time 298.510 seconds
[RESULT]     mdtest-easy-stat     919.974294 kIOPS : time 192.983 seconds
[RESULT]        ior-hard-read       9.767739 GiB/s : time 258.846 seconds
[RESULT]     mdtest-hard-stat     532.842517 kIOPS : time  73.066 seconds
[RESULT]   mdtest-easy-delete     389.111423 kIOPS : time 467.045 seconds
[RESULT]     mdtest-hard-read     186.604589 kIOPS : time 207.250 seconds
[RESULT]   mdtest-hard-delete     370.730738 kIOPS : time 192.598 seconds
[SCORE] Bandwidth 13.729175 GiB/s : IOPS 363.768691 kiops : TOTAL 70.669966
```


ior-hard used `--dfs.chunk_size=470080` (10x the transferSize)

SC20 Everywhere we are | more than hpc.

DAOS IO500 – 1 Server (P4610 1.6TB), 10 Clients

IO500 with API=DFS, and Mohamad's DFS-enabled `find` from `mpifileutils`:

```
IO500 version io500-sc20_v3
[RESULT]       ior-easy-write      14.385031 GiB/s : time 323.882 seconds
[RESULT]    mdtest-easy-write     583.580500 kIOPS : time 307.465 seconds
[RESULT]       ior-hard-write       7.995398 GiB/s : time 315.677 seconds
[RESULT]    mdtest-hard-write     120.932885 kIOPS : time 319.958 seconds
[RESULT]                 find     326.808035 kIOPS : time 660.728 seconds
[RESULT]        ior-easy-read      21.764214 GiB/s : time 213.669 seconds
[RESULT]     mdtest-easy-stat     866.138808 kIOPS : time 204.900 seconds
[RESULT]        ior-hard-read       7.026412 GiB/s : time 358.190 seconds
[RESULT]     mdtest-hard-stat     511.331836 kIOPS : time  76.019 seconds
[RESULT]   mdtest-easy-delete     356.626439 kIOPS : time 510.617 seconds
[RESULT]     mdtest-hard-read     187.117240 kIOPS : time 207.295 seconds
[RESULT]   mdtest-hard-delete     367.907065 kIOPS : time 193.228 seconds
[SCORE] Bandwidth 11.516139 GiB/s : IOPS 354.741268 kiops : TOTAL 63.915957
```


ior-hard used `--dfs.chunk_size=470080` (10x the transferSize)

SC20 Everywhere we are | more than hpc.

DAOS IO500 – 9 Servers (P4610 1.6TB), 10 Clients

IO500 with API=DFS, and Mohamad's DFS-enabled `find` from `mpifileutils`:

```
IO500 version io500-sc20_v3
[RESULT]       ior-easy-write      92.534946 GiB/s : time 308.122 seconds
[RESULT]    mdtest-easy-write    3135.992682 kIOPS : time 314.497 seconds
[RESULT]       ior-hard-write      50.643138 GiB/s : time 340.693 seconds
[RESULT]    mdtest-hard-write     690.243412 kIOPS : time 325.546 seconds
[RESULT]                 find    1349.765412 kIOPS : time 881.429 seconds
[RESULT]        ior-easy-read      86.753938 GiB/s : time 328.958 seconds
[RESULT]     mdtest-easy-stat    2011.338604 kIOPS : time 482.063 seconds
[RESULT]        ior-hard-read      33.106979 GiB/s : time 520.042 seconds
[RESULT]     mdtest-hard-stat    1792.537325 kIOPS : time 126.983 seconds
[RESULT]   mdtest-easy-delete    1927.080799 kIOPS : time 515.165 seconds
[RESULT]     mdtest-hard-read     892.651111 kIOPS : time 252.147 seconds
[RESULT]   mdtest-hard-delete    1816.380945 kIOPS : time 235.734 seconds
[SCORE] Bandwidth 60.570169 GiB/s : IOPS 1547.648039 kiops : TOTAL 306.172016
```


ior-hard used `--dfs.chunk_size=470080` (10x the transferSize)

SC20 Everywhere we are | more than hpc.

mhennecke @ lenovo.com

thanks.