# Scalability Testing of DNE2 in Lustre 2.7

**Tom Crowe, Nathan Lavender, Stephen Simms**

Research Technologies
High Performance File Systems
hpfs-admin@iu.edu
Indiana University

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

# Abstract

"Increasing Lustre's metadata performance is something that Indiana University HPC users greatly desire. Because of the many user comments and requests, the High Performance File Systems group at IU has been looking into metadata performance using solid state storage devices. The most recent tests that we have performed involved the use of multiple metadata servers and the striped directory functionality provided by DNE2.

This presentation will feature the data we have gathered on the relationship between metadata performance and MDT count using DNE2."

# Distributed Namespace Environment (DNE)

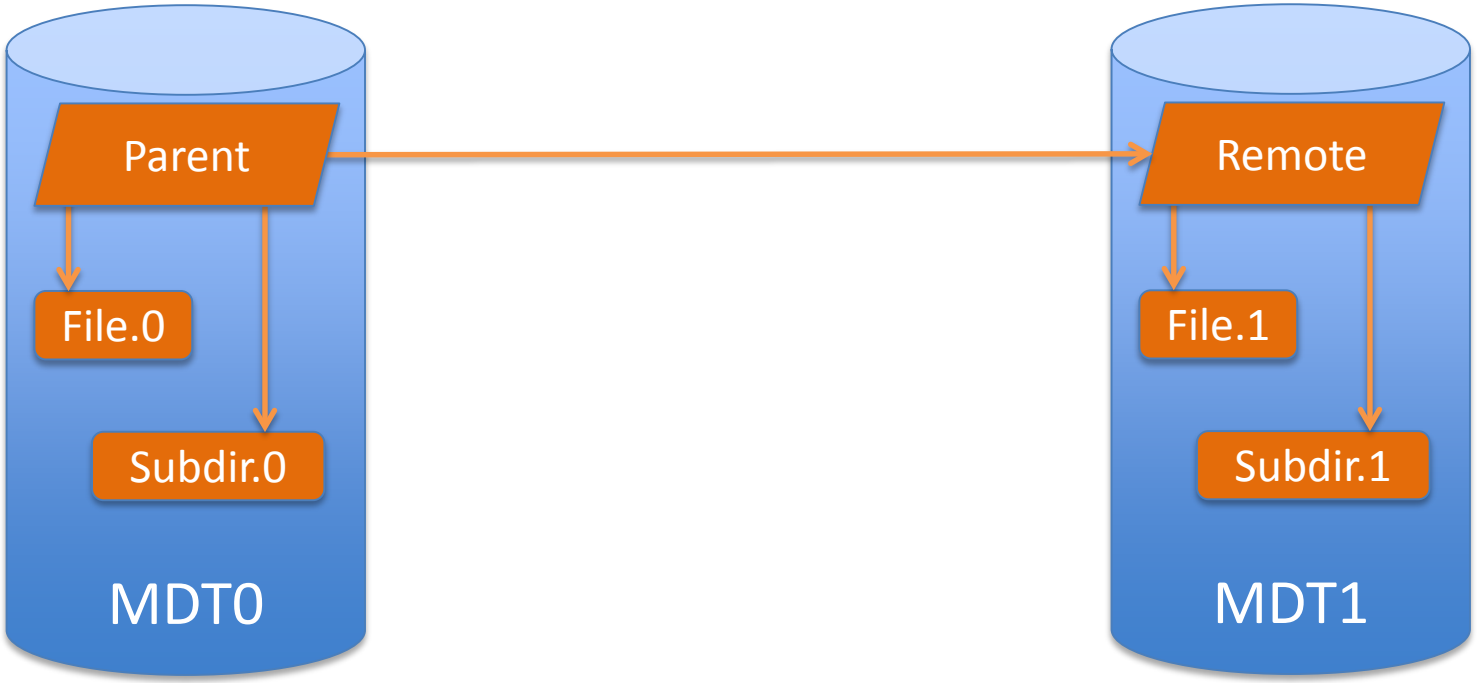
**DNE Phase 1 – Lustre 2.4**

- Enables deployment of multiple MDTs on one or more MDS nodes
- create directories on a specific remote MDT

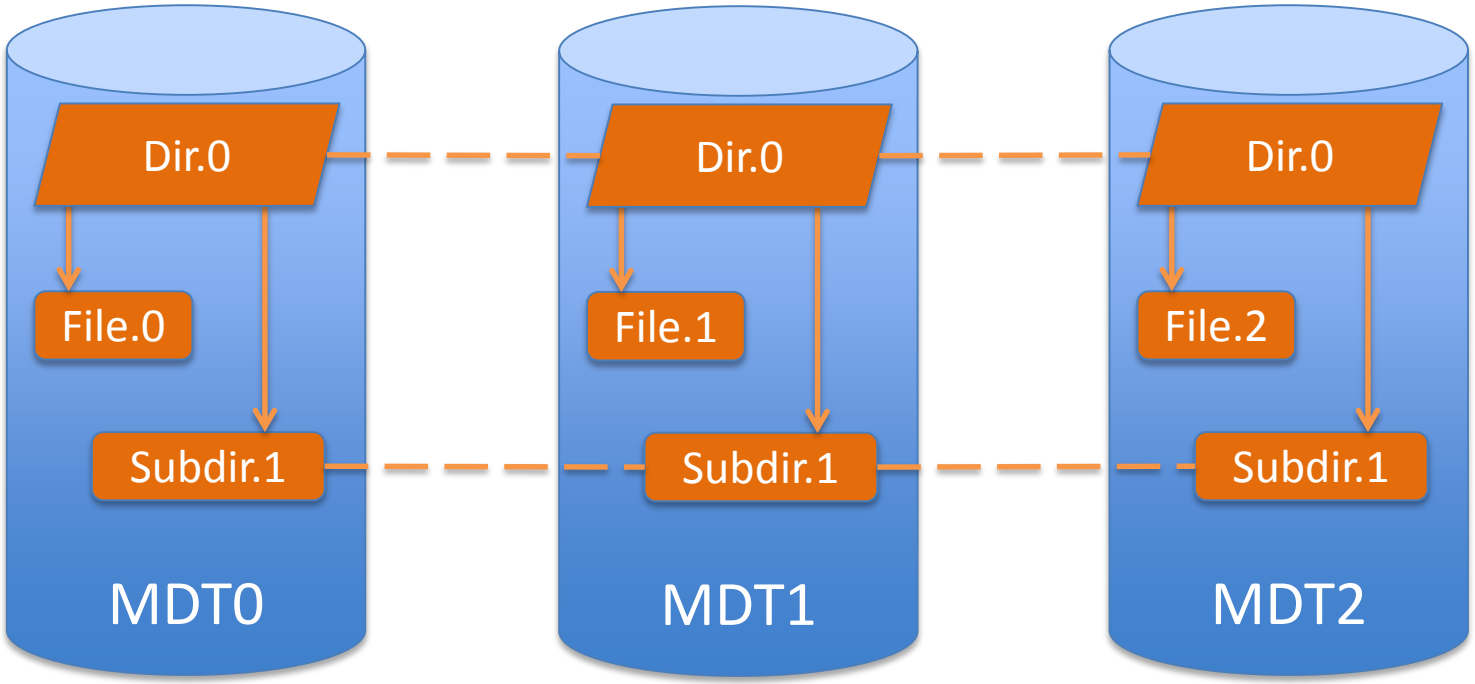
**DNE Phase 2 (preview) – Lustre 2.6/2.7**

- Enables deployment of striped directories on multiple MDS nodes
- Improved versatility

# Distributed NamespacE (DNE) – Remote Directory




# Distributed NamespacE (DNE) – Striped Directory



# Indiana University Metadata – Current Status

**Multiple Compute Clusters**

- Over 150 Disciplines served
- Mixed workloads, various I/O patters

**Current Metadata Challenge**

- Single MDS/MDT comprised from 24 SAS drives (RAID-10)
- +979,000,000 inodes
- Lustre 2.1.6 with plans to move to 2.5.X soon.

**Bottom Line - more metadata performance please**

- SSD + DNE2 = goodness?

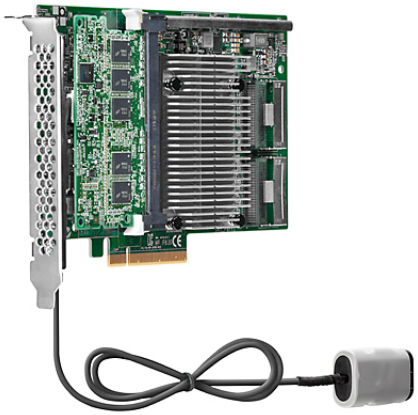
# Building Blocks

**(6) Servers, identical specs**

- HP ProLiant DL380p Generation8 (Gen8)
- Dual socket Intel(R) Xeon(R) 2x E5-2667v2  "Ivy Bridge-EP" @ 3.30GHz 8 core
- 128GB - (16) 8GB @ 1866MHz memory
- HP Smart Array P830 controller with 4GB battery backed cache
- (6) Intel SSD DC S3500 drives (800GB drives)
- (1) SAS drive (146GB, 15,000 RPM)
- Mellanox ConnectX-3

# Building Blocks (cont)

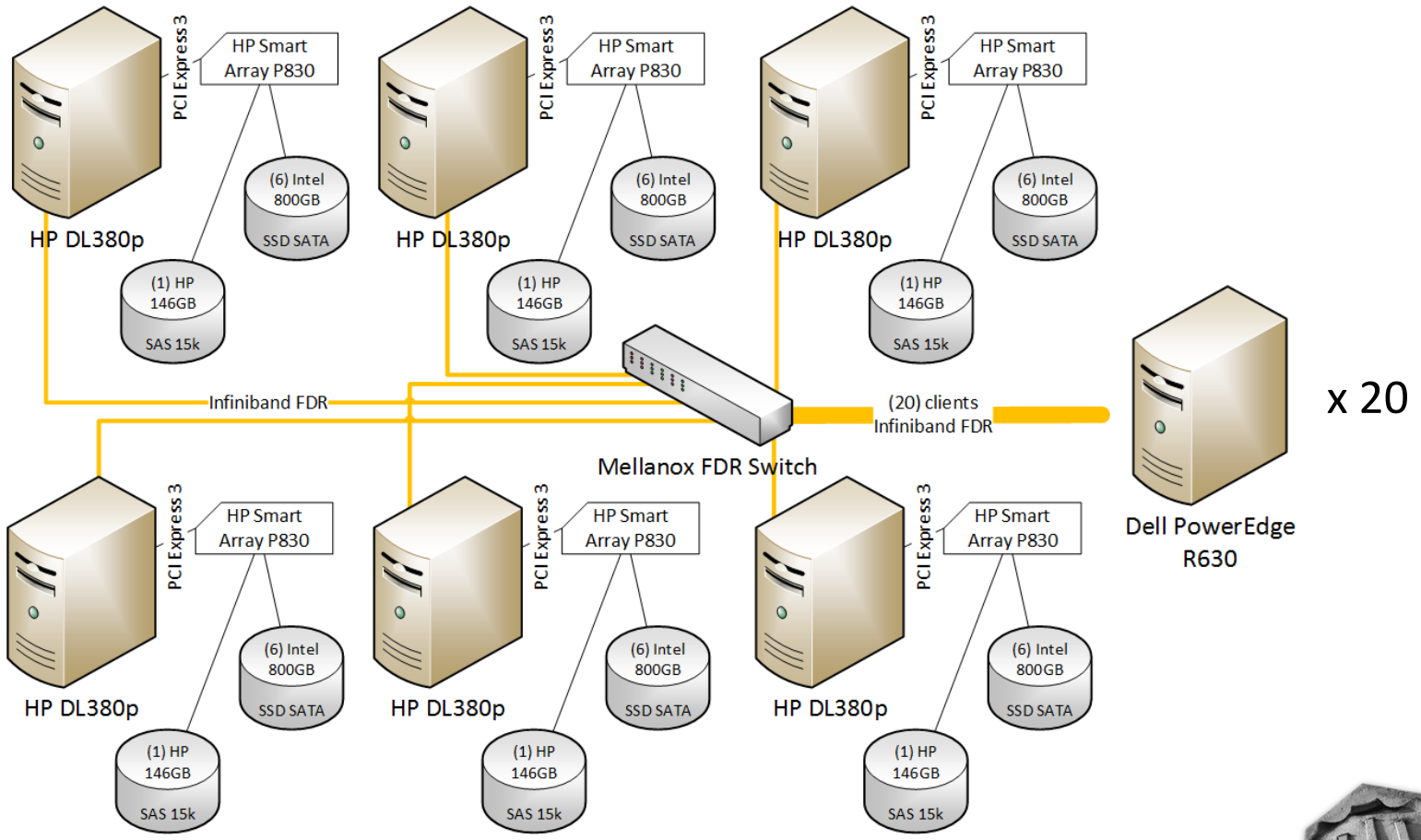
(6) HP DL380p G8 Servers



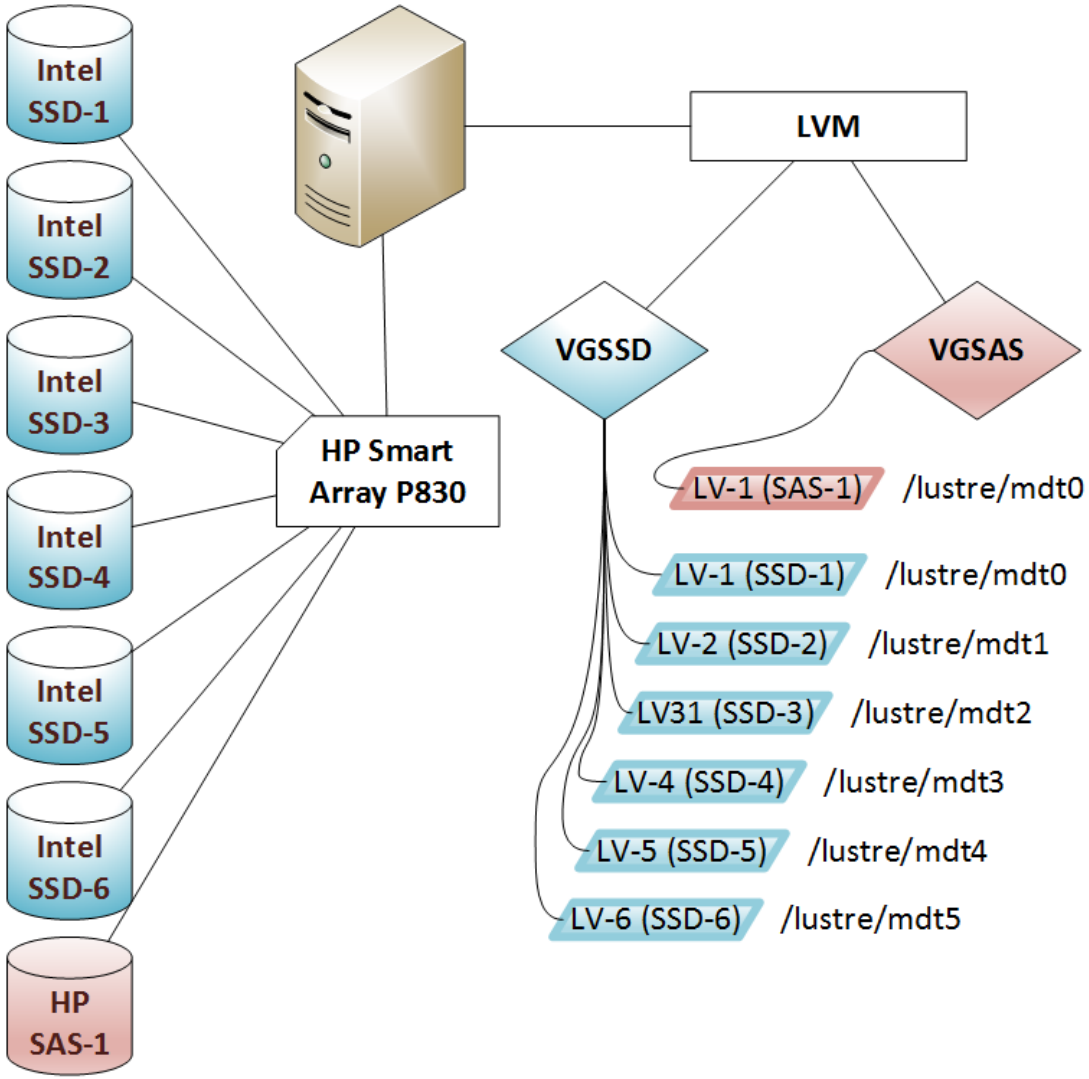
HP Smart Array
830p controller


Intel SSD DC S3500

# Building Blocks (cont)



# Logical Setup



# Logical Setup

**Block Devices**

- 50GB LUNs were provisioned from each drive, preserving 1:1 layout
  - » 50GB LUNs allowed mkfs to complete in a reasonable time

**File System Options**

- 8GB journal
- lazy_itable_init=0
  - » Enabled by default resulting in file system activity directly following mkfs/mount

# Methodology

**Software**

- mdsrate – lustre aware metadata benchmark in Lustre test suite
- operation - mknod (create with no OST object allocation)

**Wide parameter sweep**

– 20 clients, 32 mounts each, for 640 mounts simulating 640 clients

– Varied number of directories from 1 to 128 by powers of 2

– 4 threads per directory, each on a separate mount point

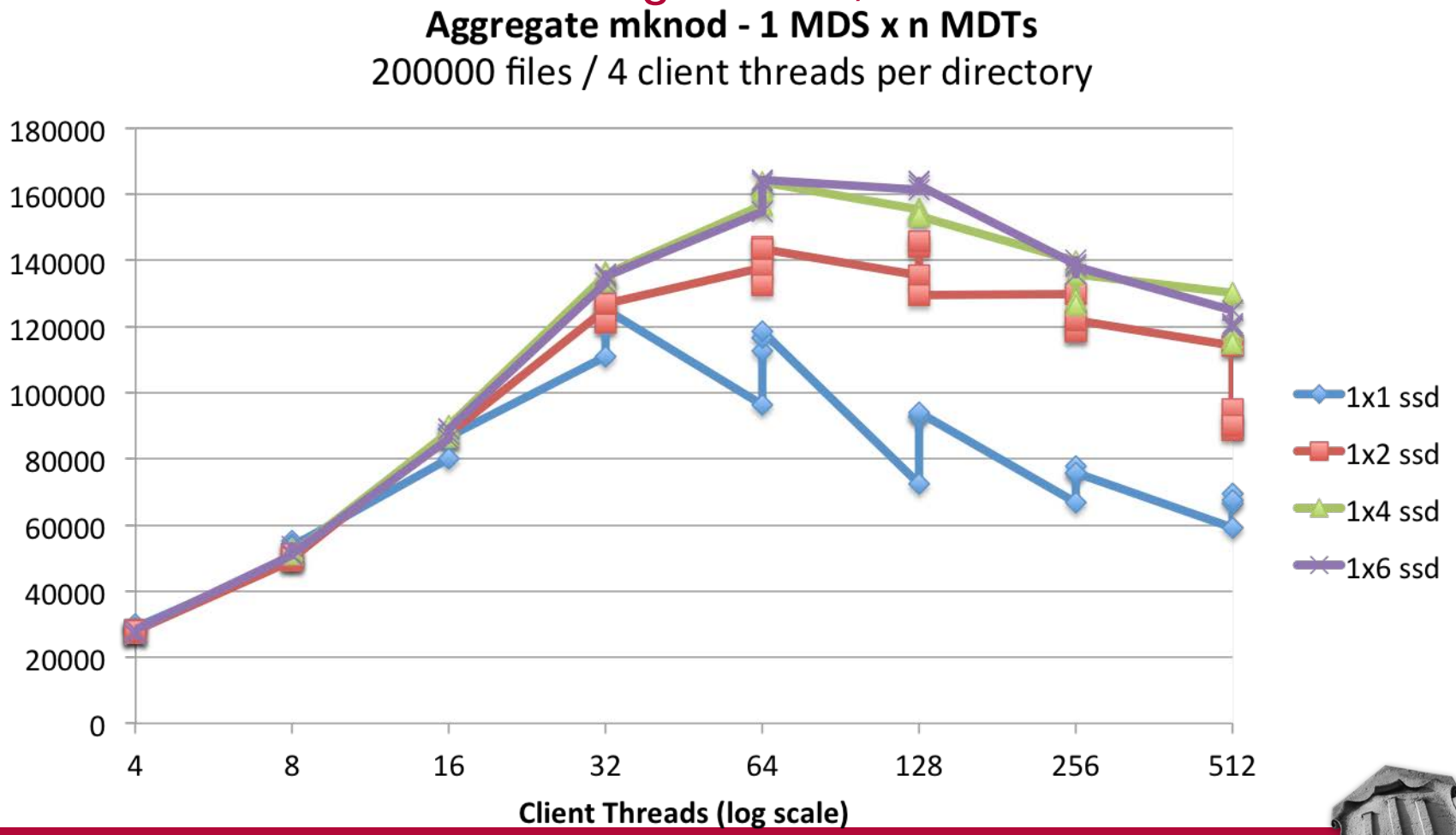
– Directory stripe count increased matching MDT count

**Hardware Configurations Tested**

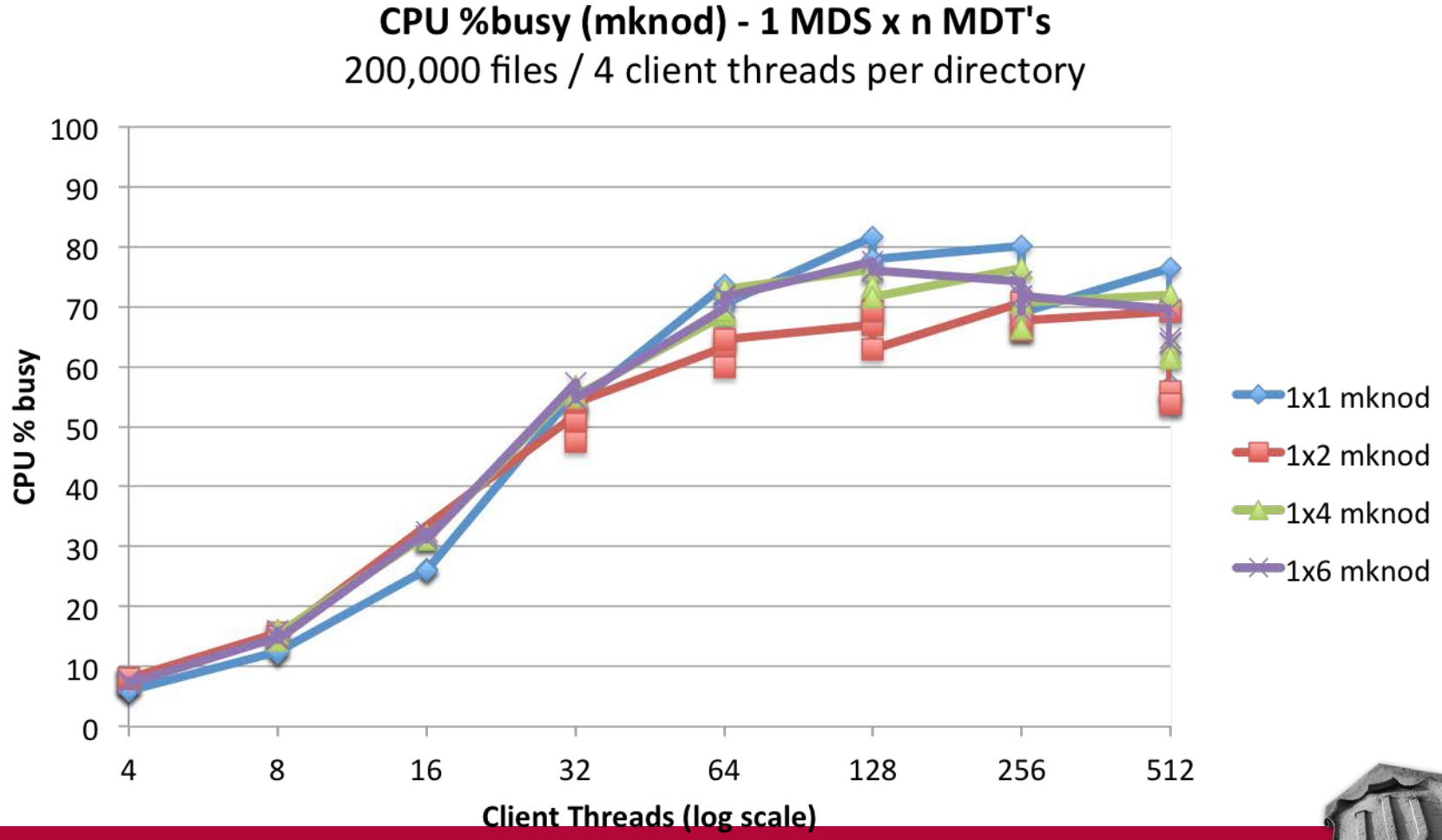
- Single MDS, multiple MDTs
- Multiple MDSs, single MDT per MDS
- Multiple MDSs, multiple MDTs per MDS

# Single MDS with Multiple MDTs

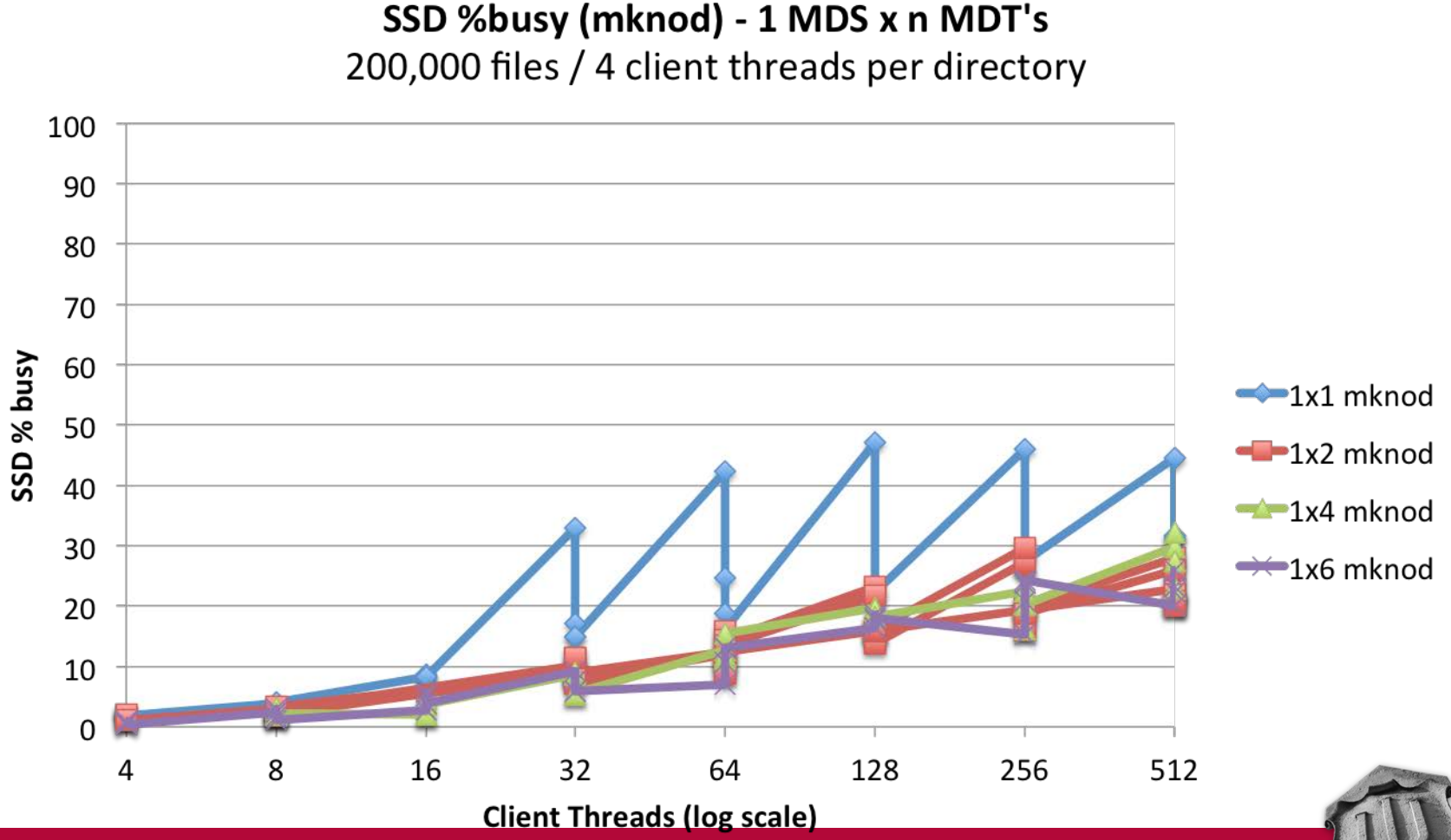
# Results - Single MDS, Multi-MDT



# Results - Single MDS, Multi-MDT



# Results - Single MDS, Multi-MDT



# Summary – Single MDS, Multi-MDT

**Lessons Learned**

- DNE2 verification
  - Metadata performance improves by increasing MDTs
  - Diminishing returns beyond 4 MDTs per MDS
- Single MDS performance peaks at 80% CPU
- SSD %busy averages don't show journal flushing IO spikes

**Why?**

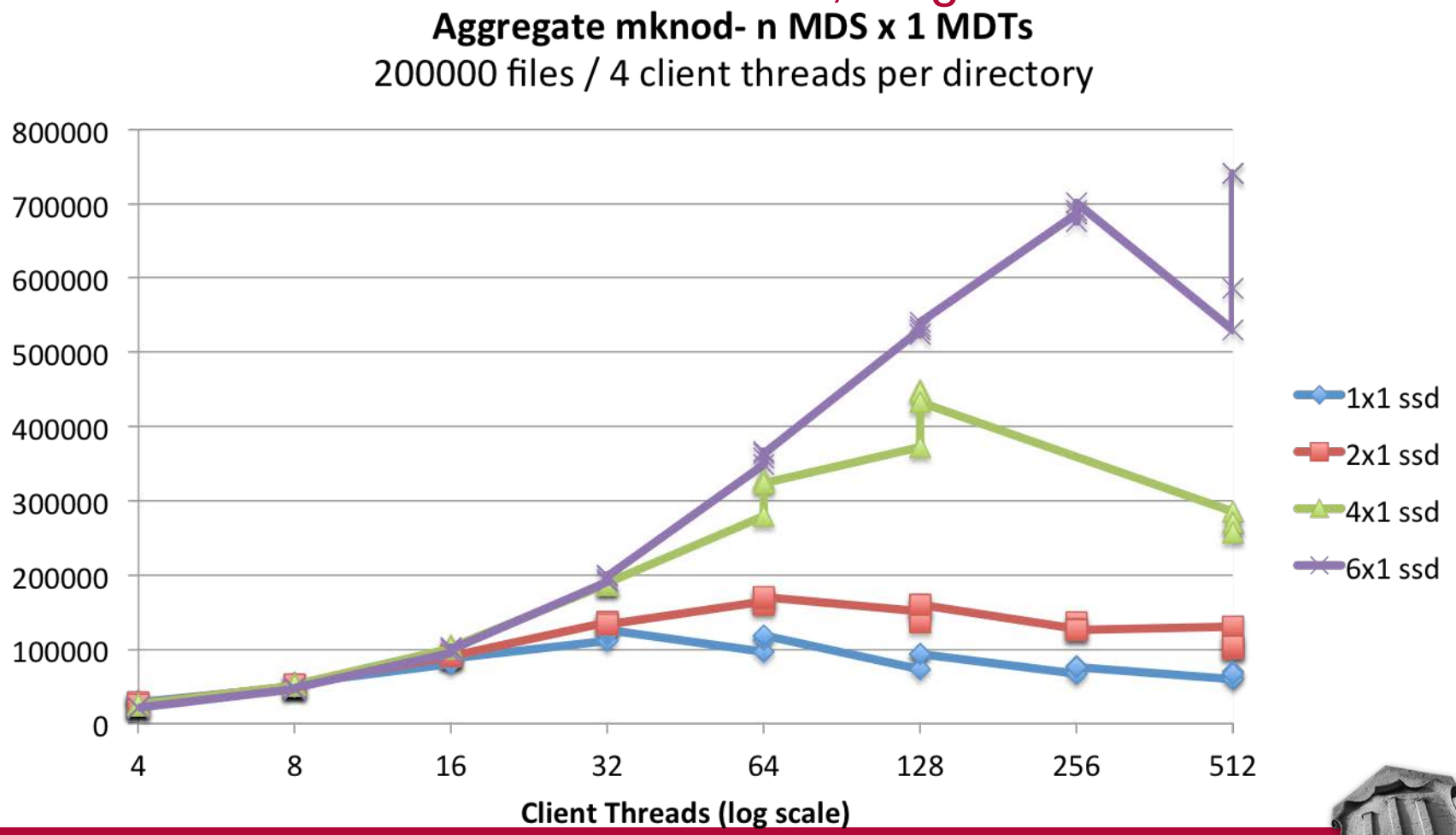
– First trial shows decreased performance

– Decrease correlates with spike in drive I/O

– File system "warm-up"?

# Multiple MDSs with Single MDT

# Results – multi-MDS, single MDT



# Summary – multi-MDS, single MDT

**Lessons Learned**

- DNE2 verification
  - Metadata performance improves increasing MDSs
- Multiple MDS scales with same client workload
- Possible client contention for 512 trials
  - 20 client x 24 cores = 480 physical, 960 hyper
- Scaling linear between 2x1 and 4x1 threads >64

**Question**

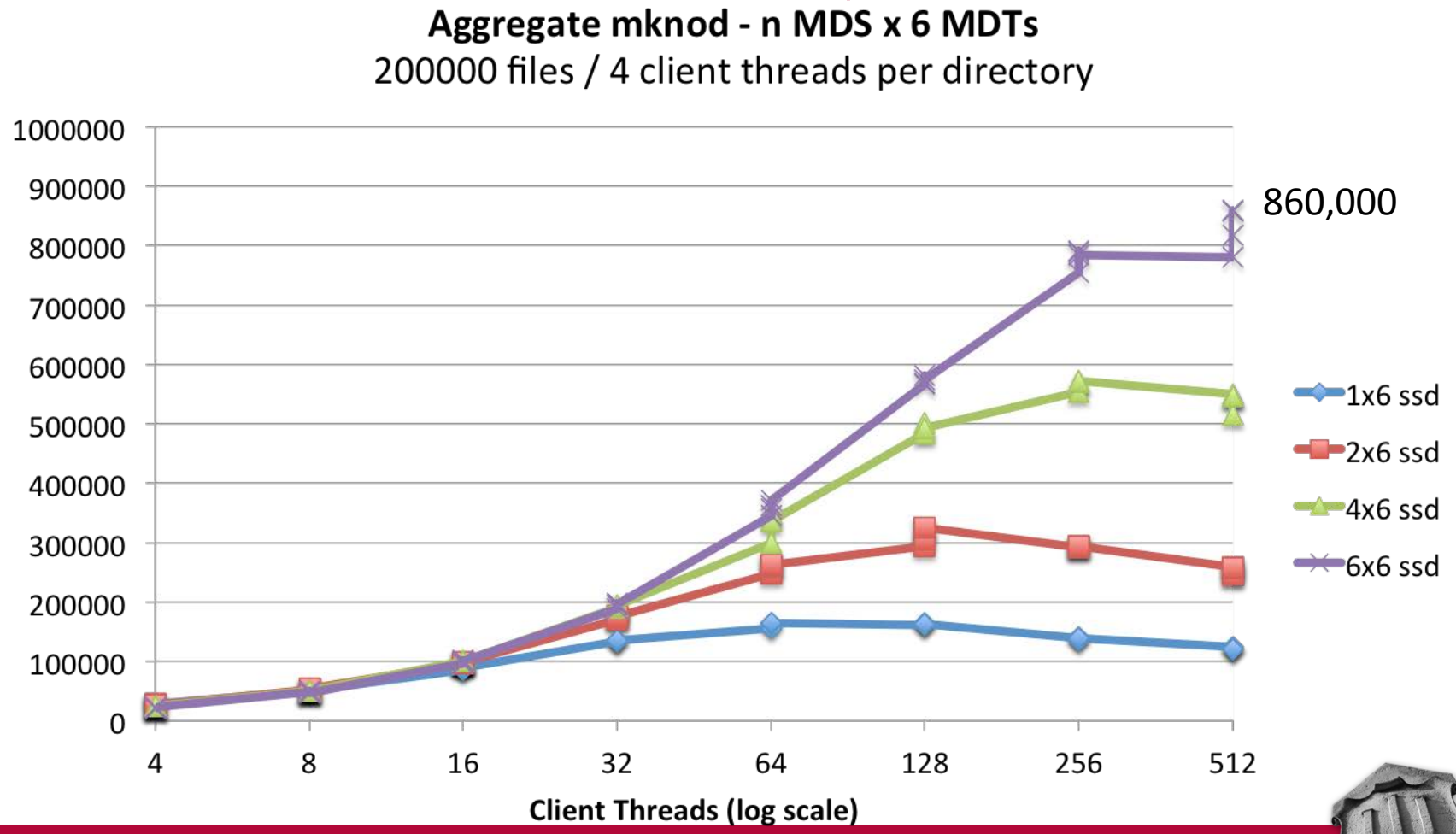
– Is it possible to increase multi-MDT performance?

» seen 80% CPU – drives aren't busy

» How do we increase efficiency?

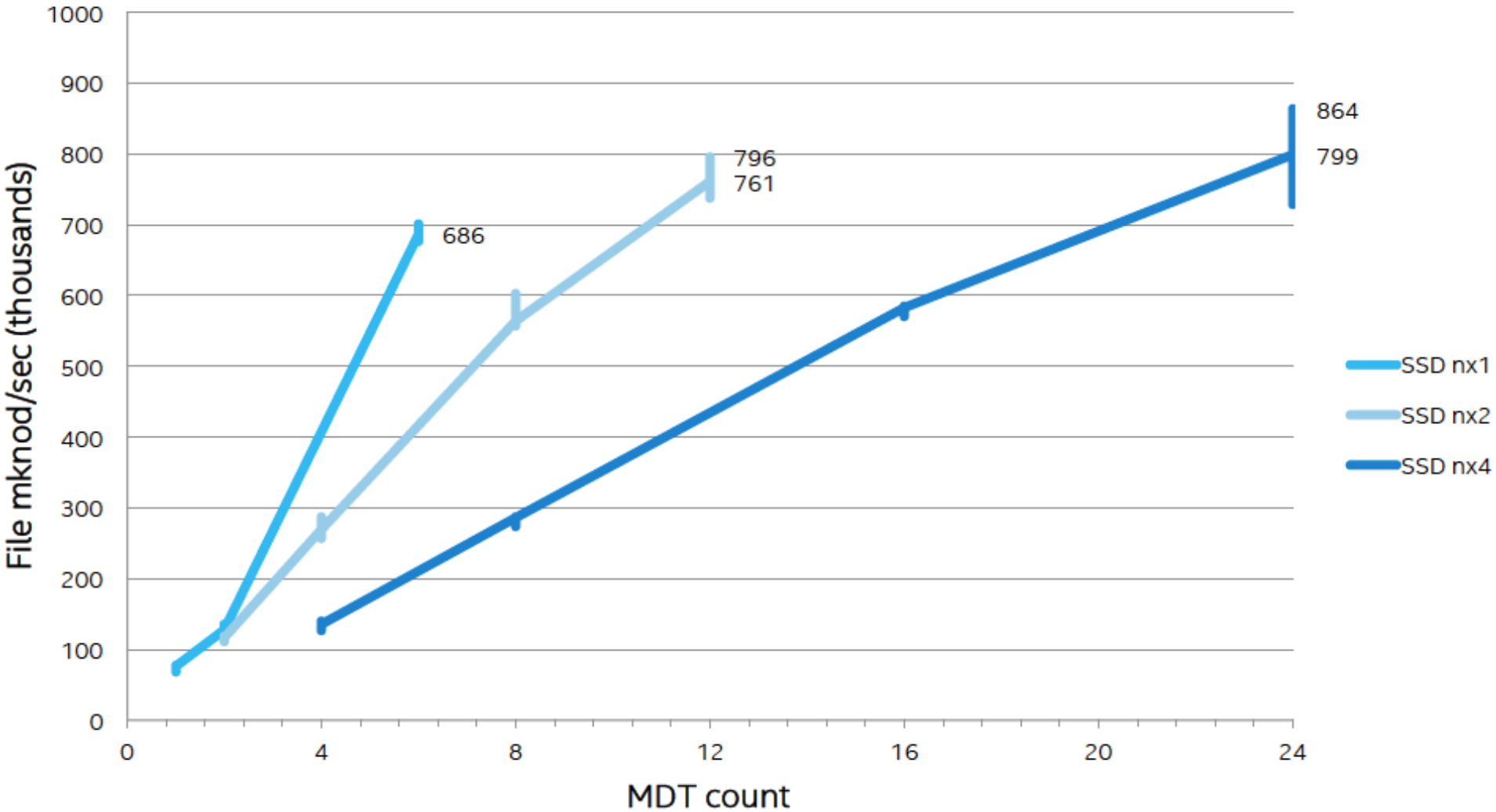
# Multiple MDSs with Multiple MDTs

# Results – multi-MDS, multi-MDT



# Results - mknod scaling with increasing MDS count

**256 threads**



# Summary – multi-MDS, multi-MDT

**Lessons Learned**

- DNE2 verification
  - Metadata performance improves
  - At 256 threads, scaling is linear with MDS count

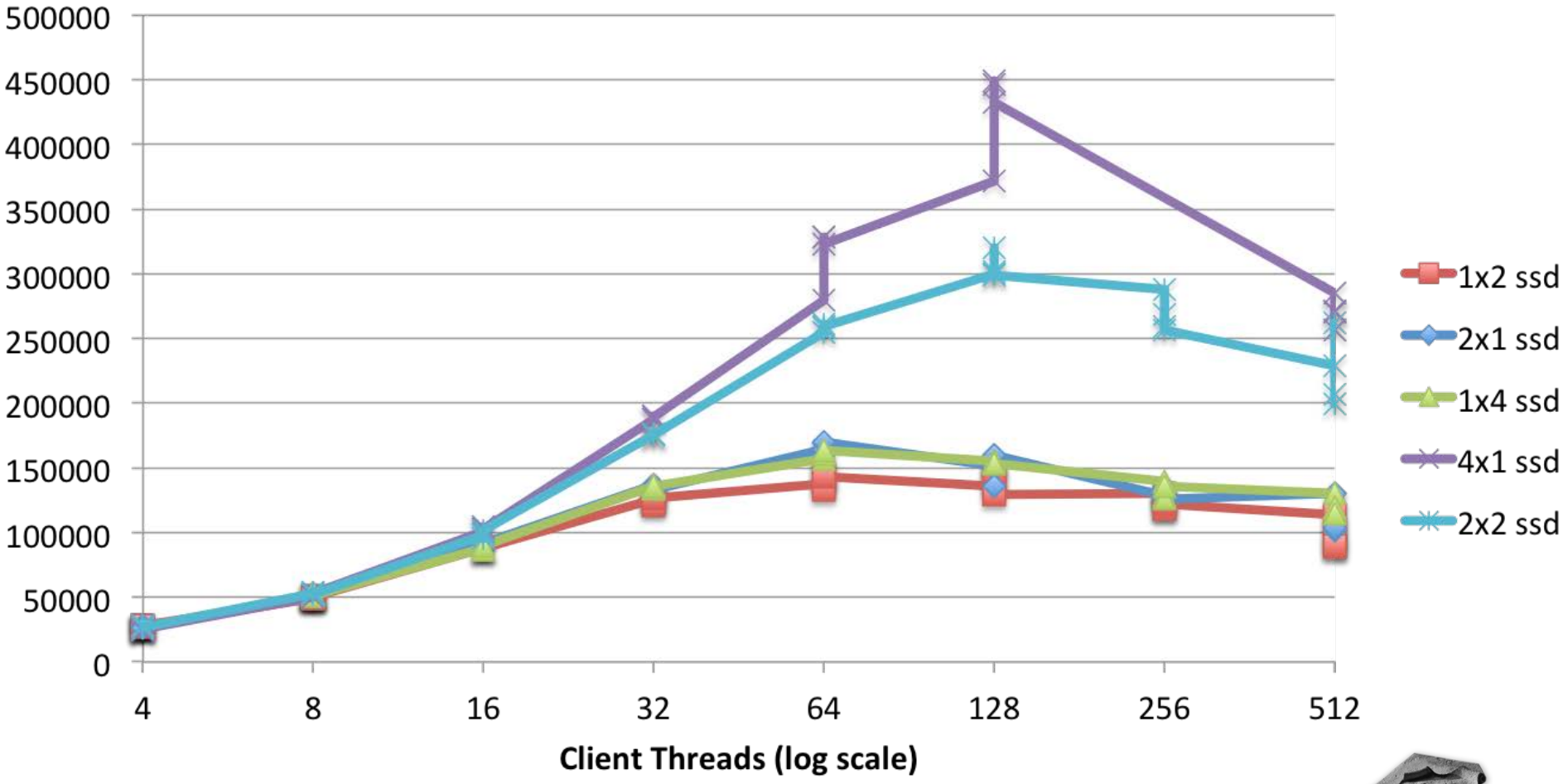
**Unanswered Questions**

– Can DNE2 scale beyond 6 MDS and 6 MDT keep linear performance?

# Comparison

# Aggregate mknod - 1 MDS x n MDTs vs n MDS x 1 MDT

200000 files / 4 client threads per directory



# Comparison Summary

- The difference between 1x2 and 2x1 isn't earth shattering
- The difference between 1x4 and 4x1 is staggering

- Extra MDTs will give more capacity but marginal performance increase
- Extra MDSs will improve performance more significantly

# Possible solution for IU's multi-discipline workload

# Methodology – DNE2 performance in a single directory

**Software**

- mdsrate - part of Lustre test suite
- operations - mknod (no OST allocation)

**Fixed parameter**

– Single striped directory

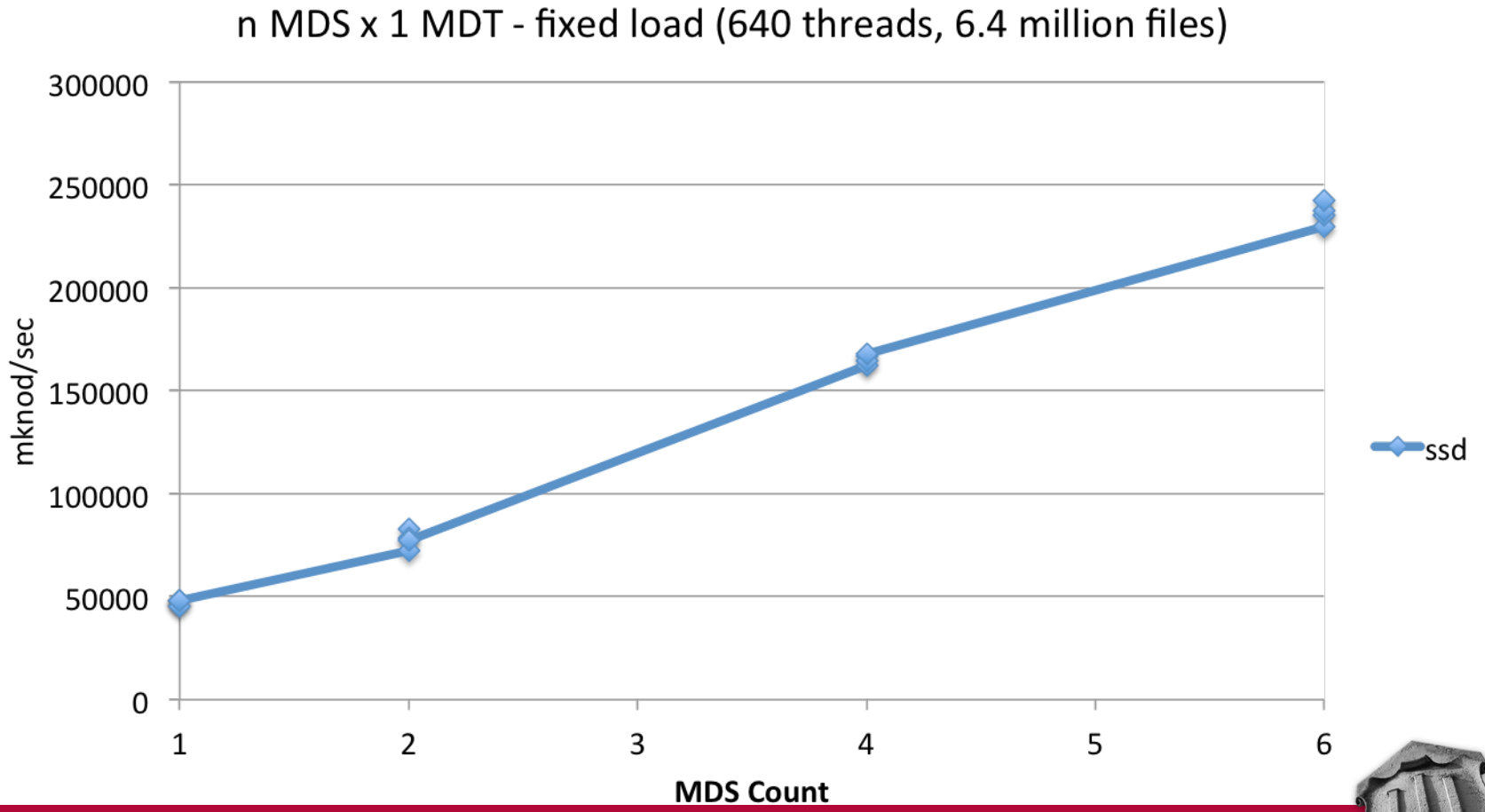
– 20 clients, 32 mounts each, total 640 mounts

– Fixed parameters on mdsrate (constant workload)

**Hardware Configurations Tested**

- Multiple MDSs, single MDT

# Results - single directory



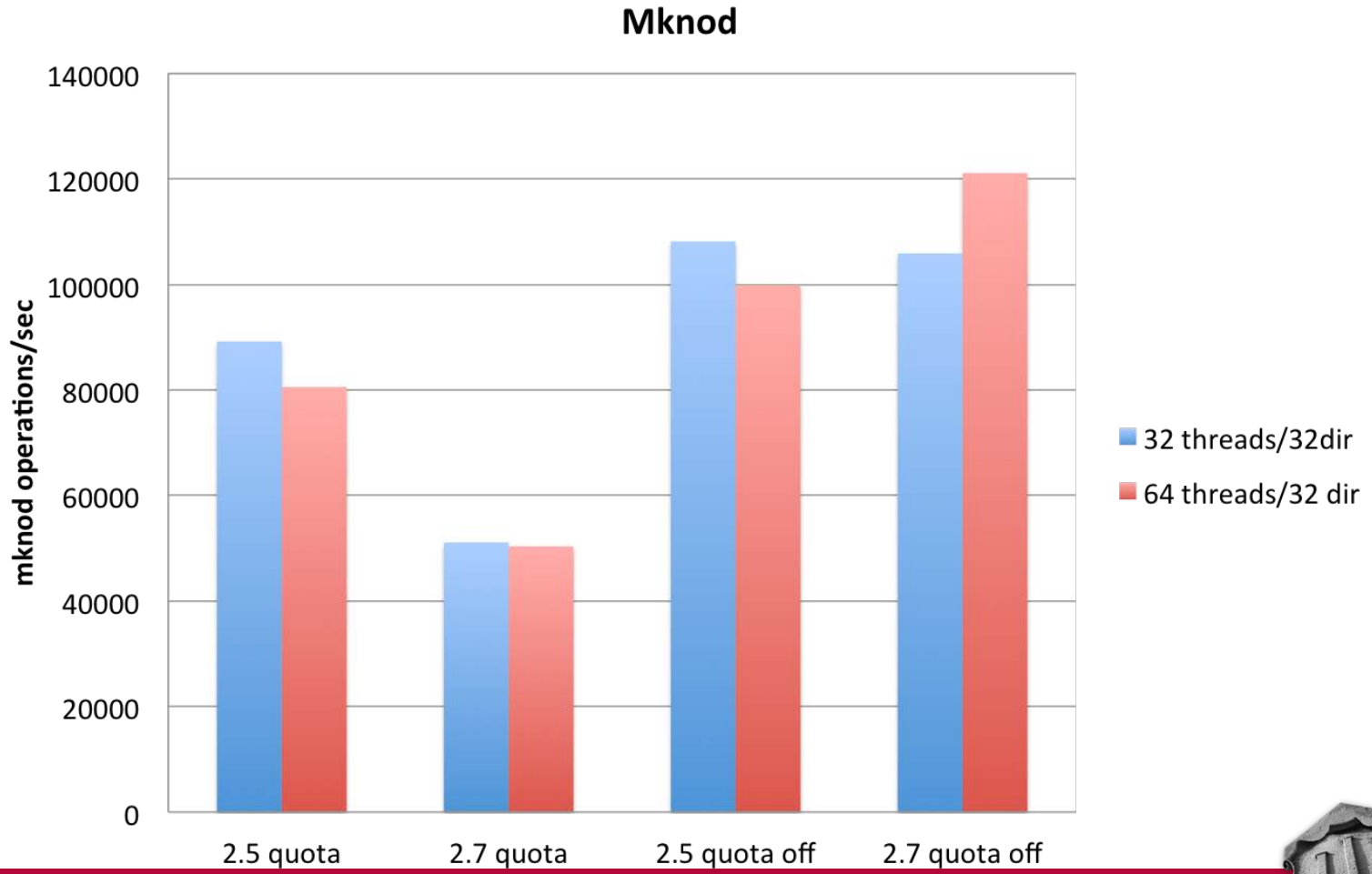
# Scaling

For years it has been possible to scale aggregate I/O performance by increasing OSTs

DNE2 makes it possible to scale aggregate metadata performance to the level you would like.

# Serendipity

# Quota Discovery – LU-6381



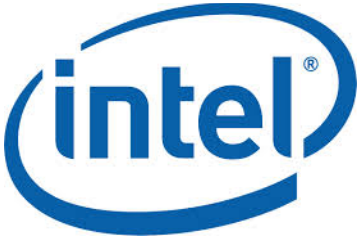
# lfs setdirstripe discovery – LU-6378

```
# lctl dl

  0 UP mgc MGC10.10.0.47@o2ib 33a61db5-fa78-ad99-9aab-bc0832e08270 5
  1 UP lov dnetwo-clilov-ffff88205ee92400 30f66e4d-cc7b-069c-2d57-74d520462cb7 4
  2 UP lmv dnetwo-clilmv-ffff88205ee92400 30f66e4d-cc7b-069c-2d57-74d520462cb7 4
...
 25 UP mdc dnetwo-MDT0016-mdc-ffff88205ee92400 30f66e4d-cc7b-069c-2d57-74d520462cb7 5
 26 UP mdc dnetwo-MDT0017-mdc-ffff88205ee92400 30f66e4d-cc7b-069c-2d57-74d520462cb7 5
 27 UP osc dnetwo-OST0000-osc-ffff88205ee92400 30f66e4d-cc7b-069c-2d57-74d520462cb7 5

# lfs setdirstripe -c 24 /lustre/dnetwo0/dir_stripe
# lfs setdirstripe -c 24 -D /lustre/dnetwo0/dir_stripe
# lfs getdirstripe /lustre/dnetwo0/dir_stripe

stripe_count: 5 lmv_stripe_offset: 0
mdtidx         FID[seq:oid:ver]
     0         [0x240000401:0x1:0x0]
     3         [0x280000400:0x1:0x0]
     8         [0x2c0000400:0x1:0x0]
     9         [0x300000400:0x1:0x0]
    10         [0x340000400:0x1:0x0]
```

# Future Work

- More data to be taken
- ZFS on MDT comparison
- Adding file creation to the mix
  - Mdsrate create in lieu of mknod
- Application testing
  - Trinity BIO code for example

# Acknowledgments



- IU's High Performance File System Team
- IU Scientific Application and Performance Tuning Team
- Matrix Integration
- Intel
- HP
- IU's Wrangler grant (NSF 13-528) partners TACC and ANL

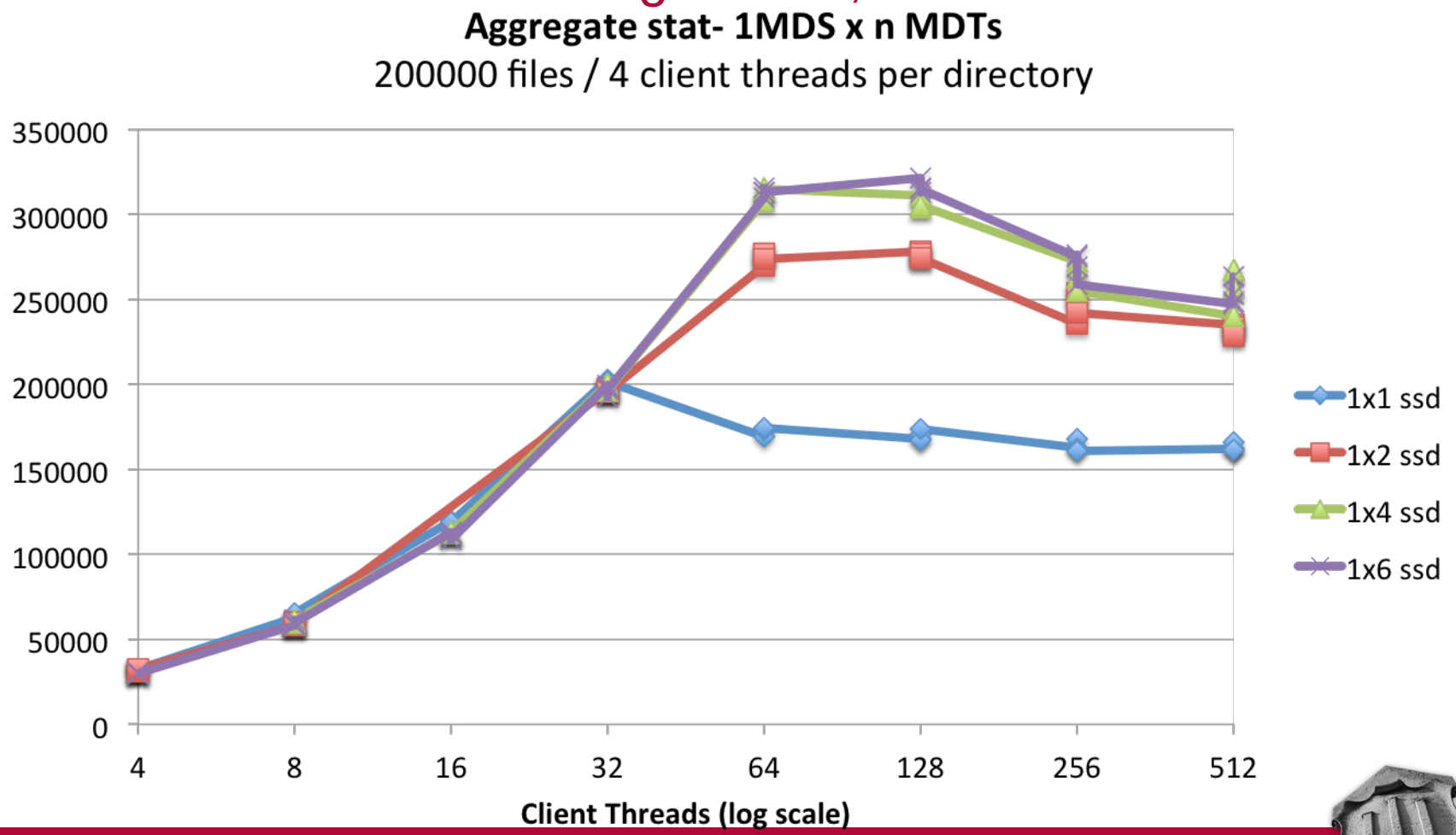
This material is based in part upon work supported by the National Science Foundation under Grant No. NSF 13-528. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

# Thank You!

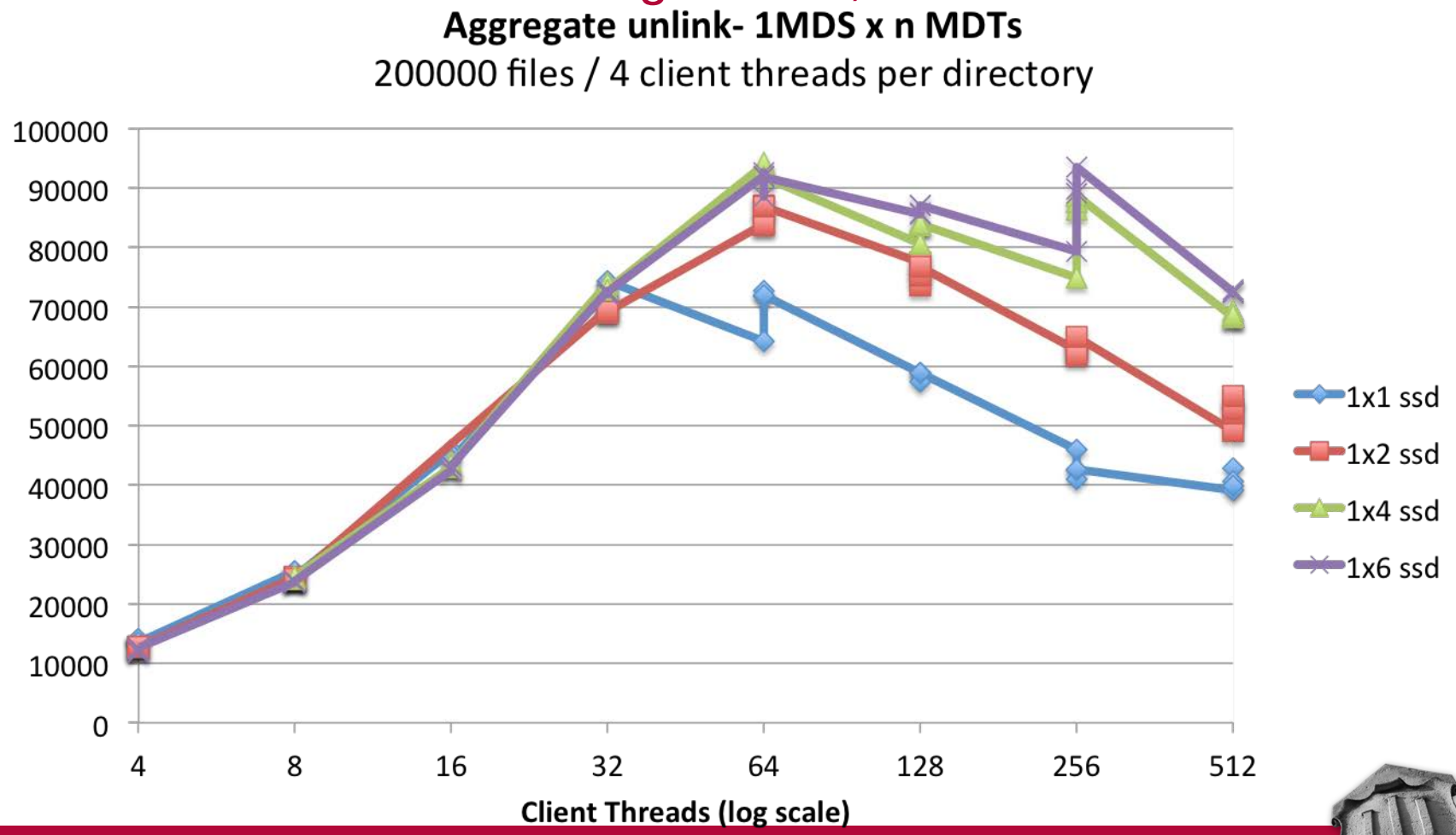
# Questions?

# Appendix

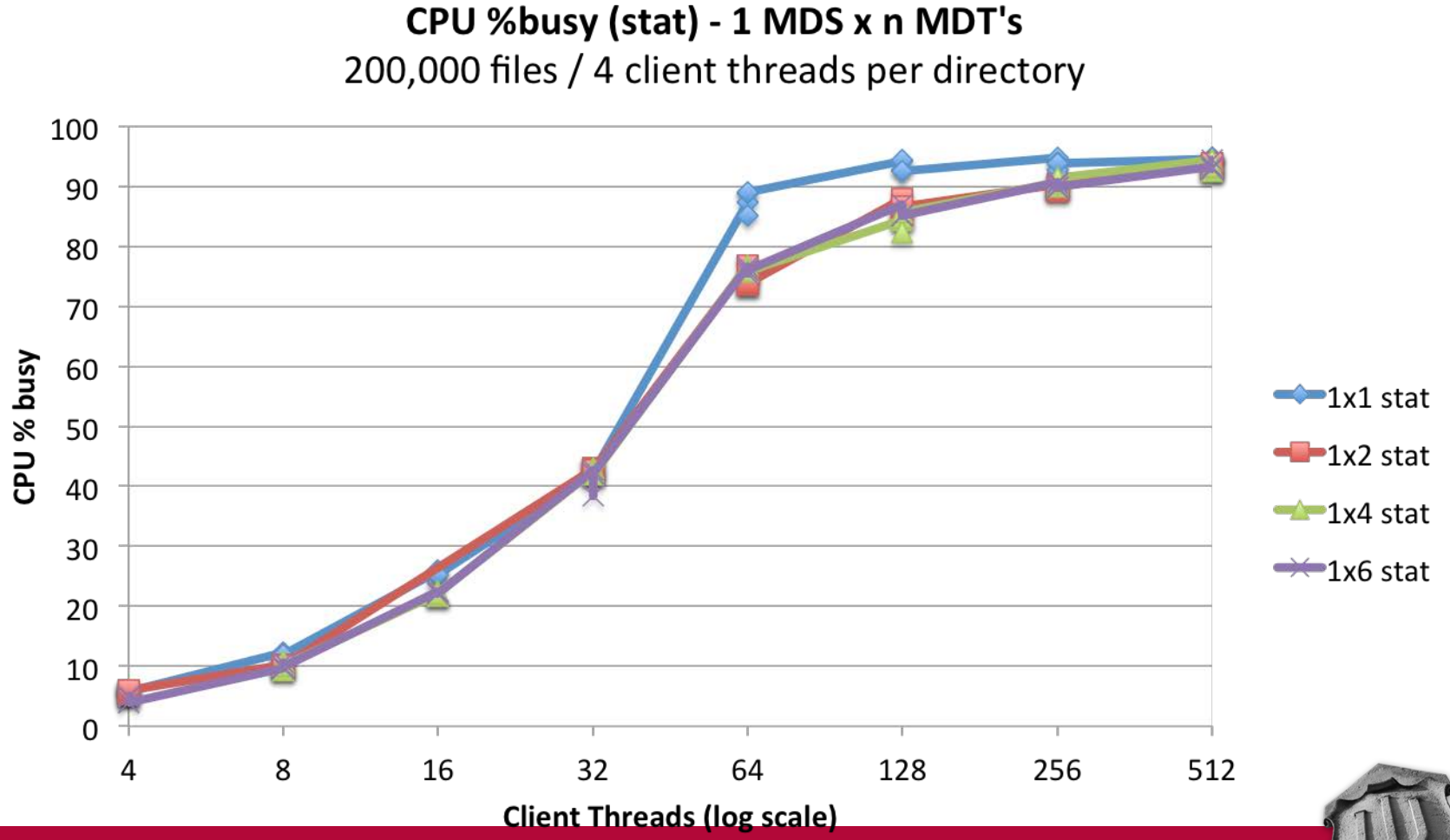
# Results - single MDS, multi-MDT



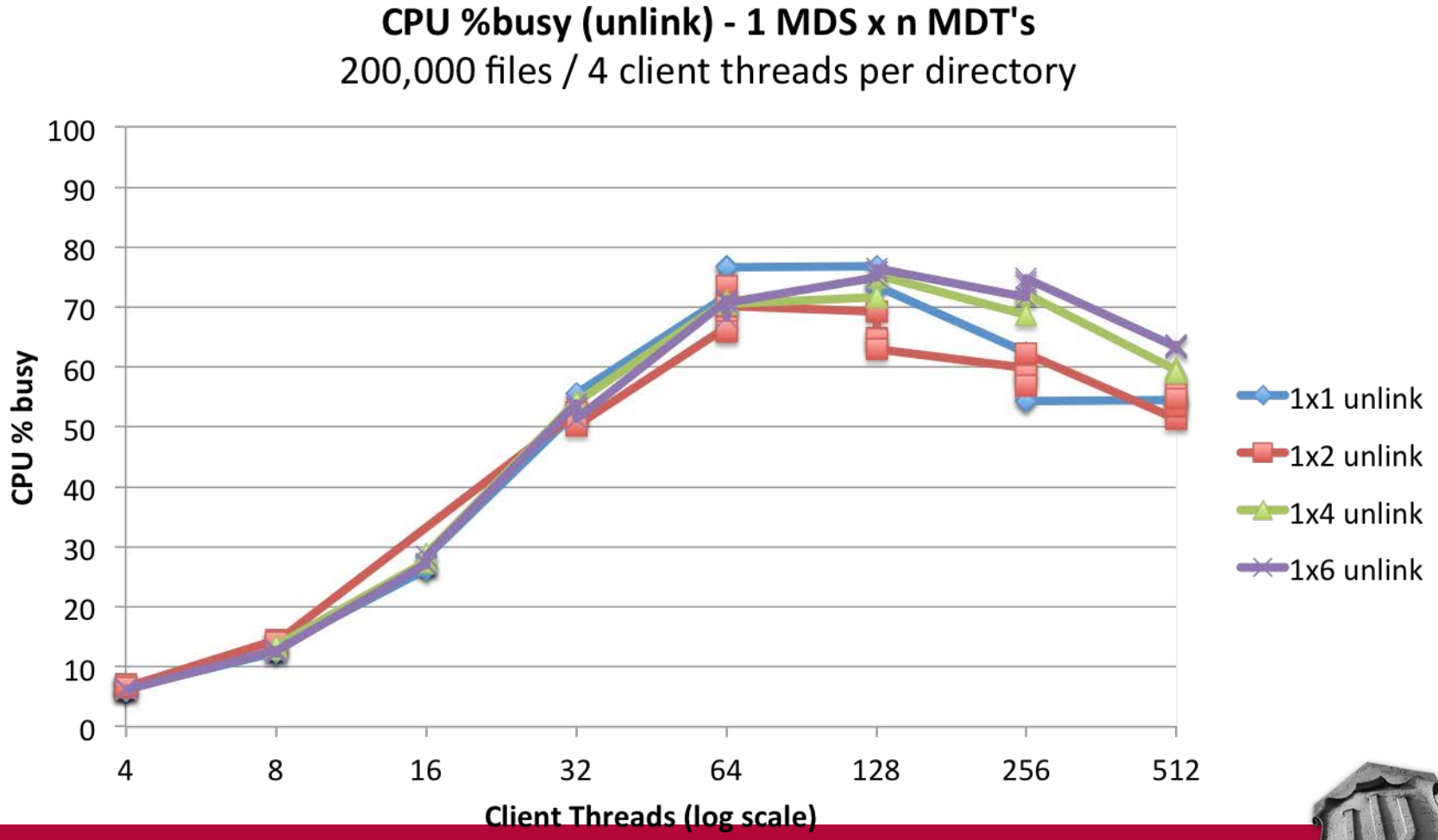
# Results - single MDS, multi-MDT



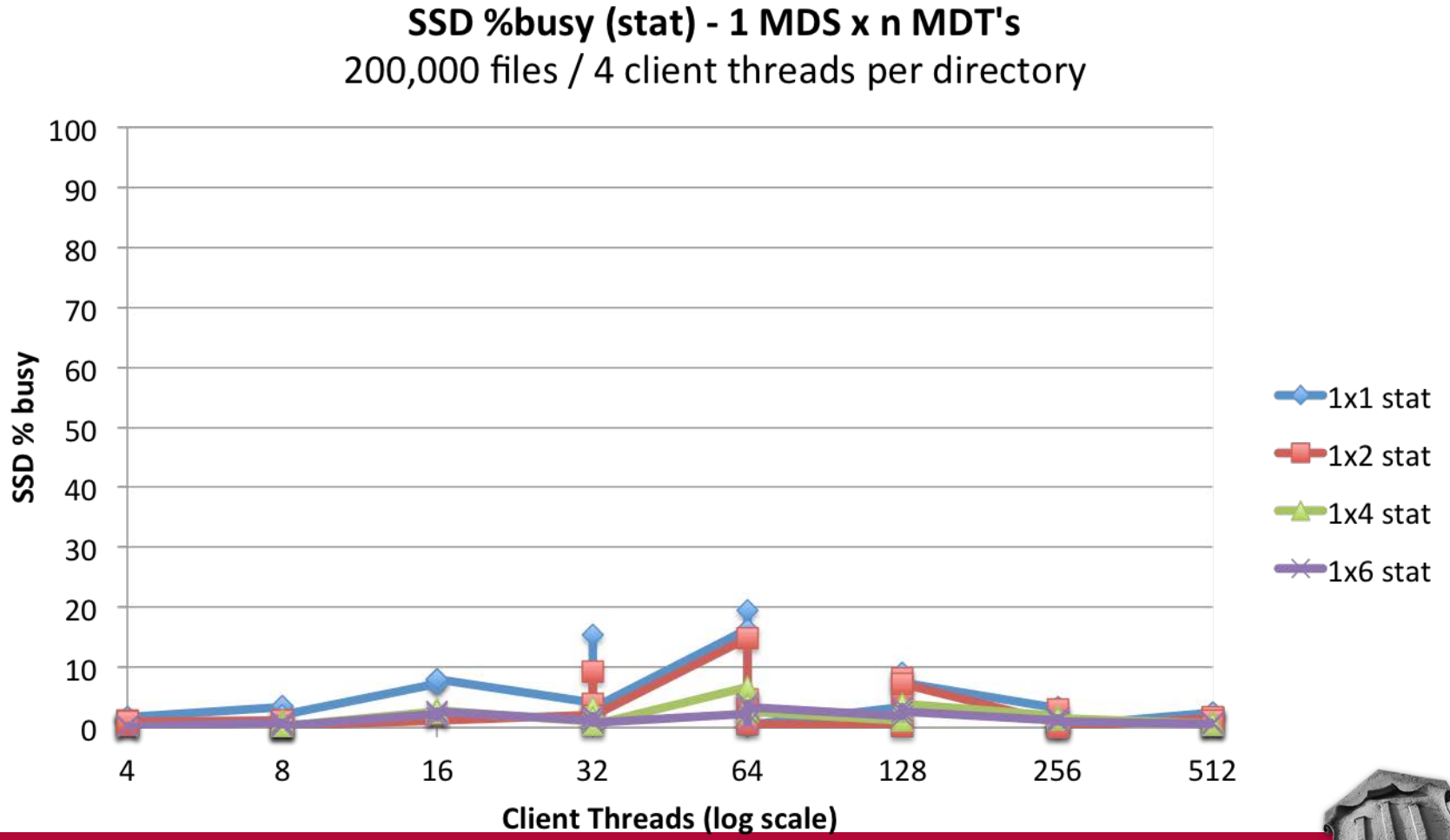
# Results - single MDS, multi-MDT



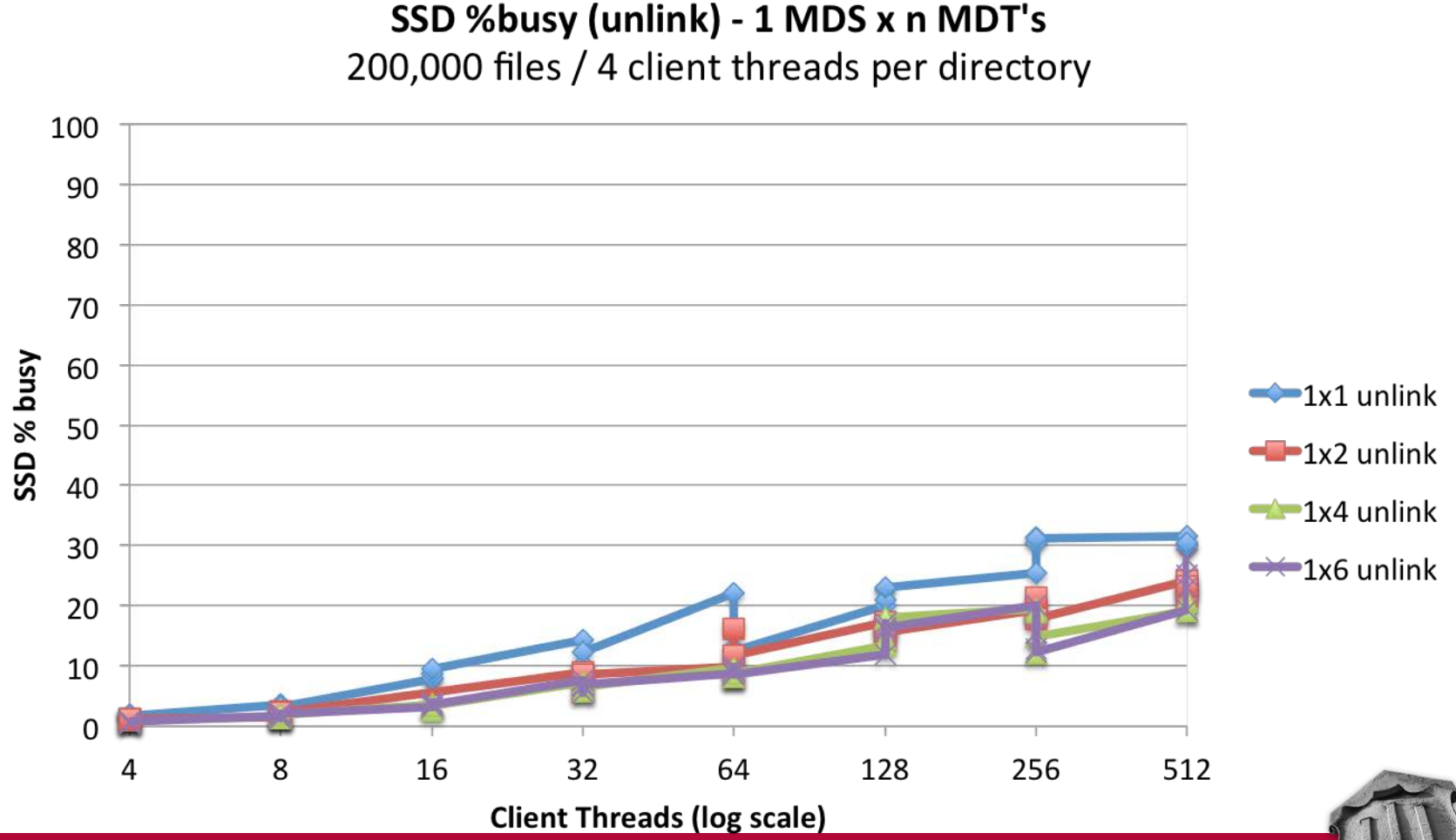
# Results - single MDS, multi-MDT



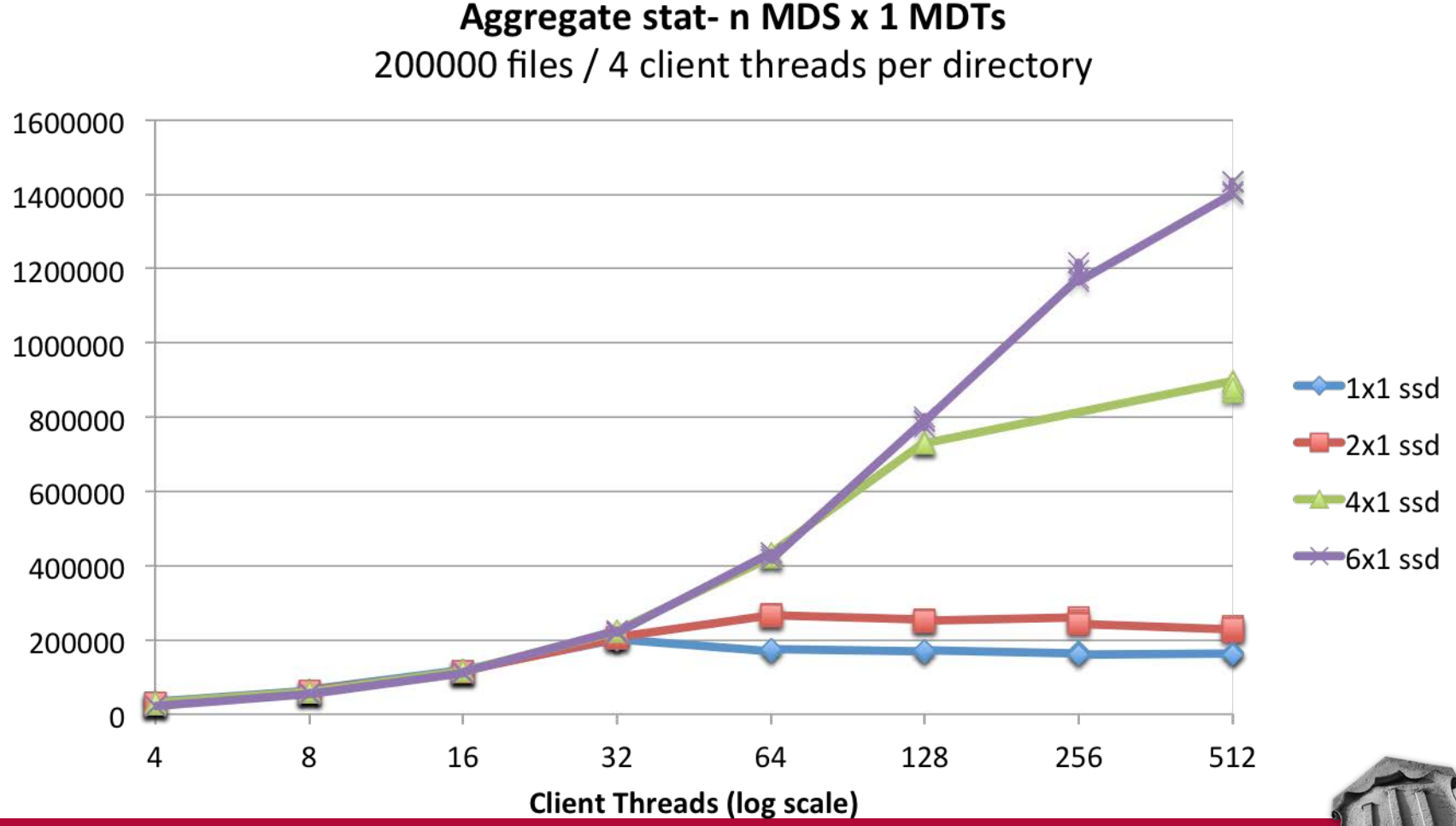
# Results - single MDS, multi-MDT



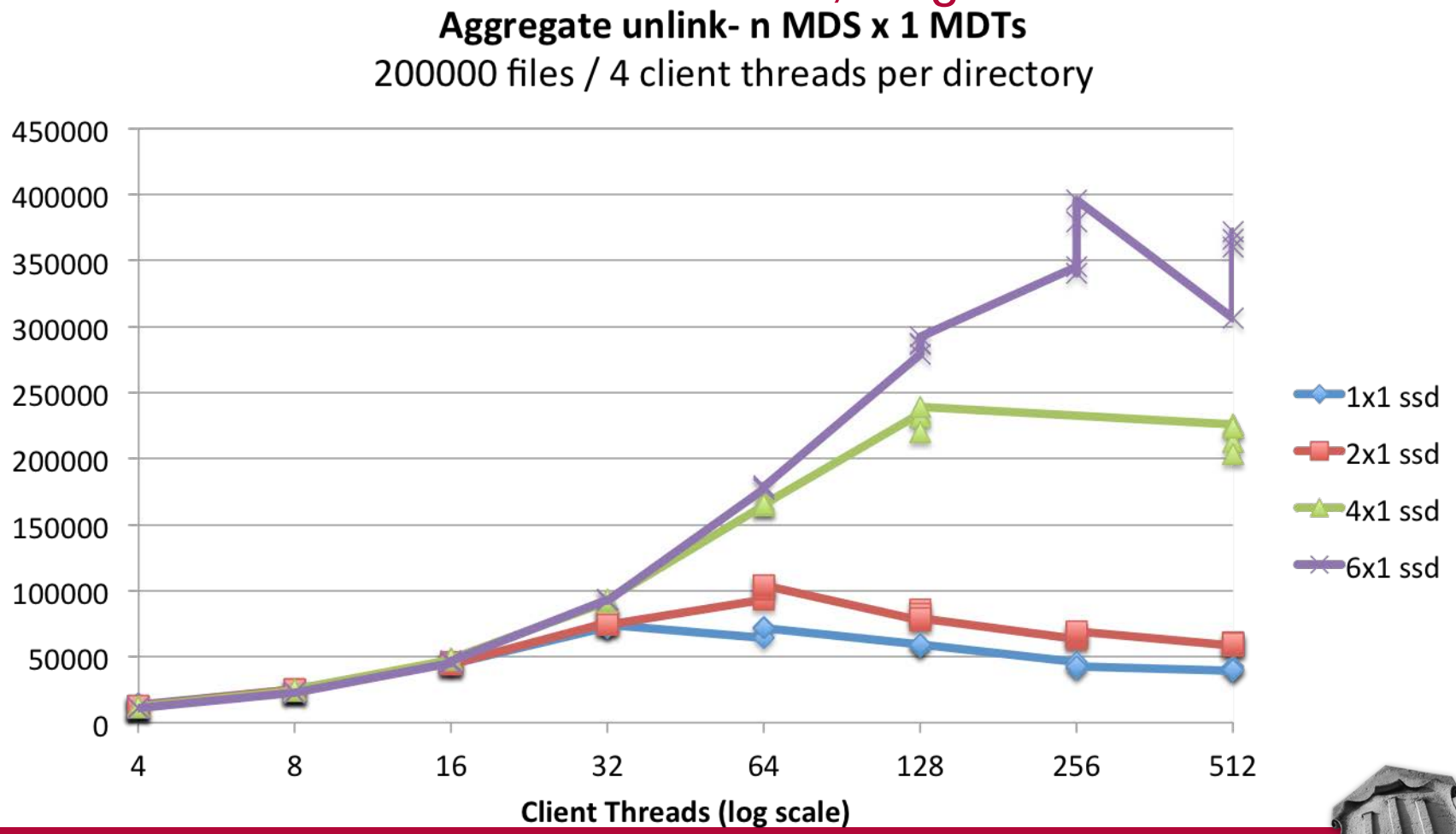
# Results - single MDS, multi-MDT



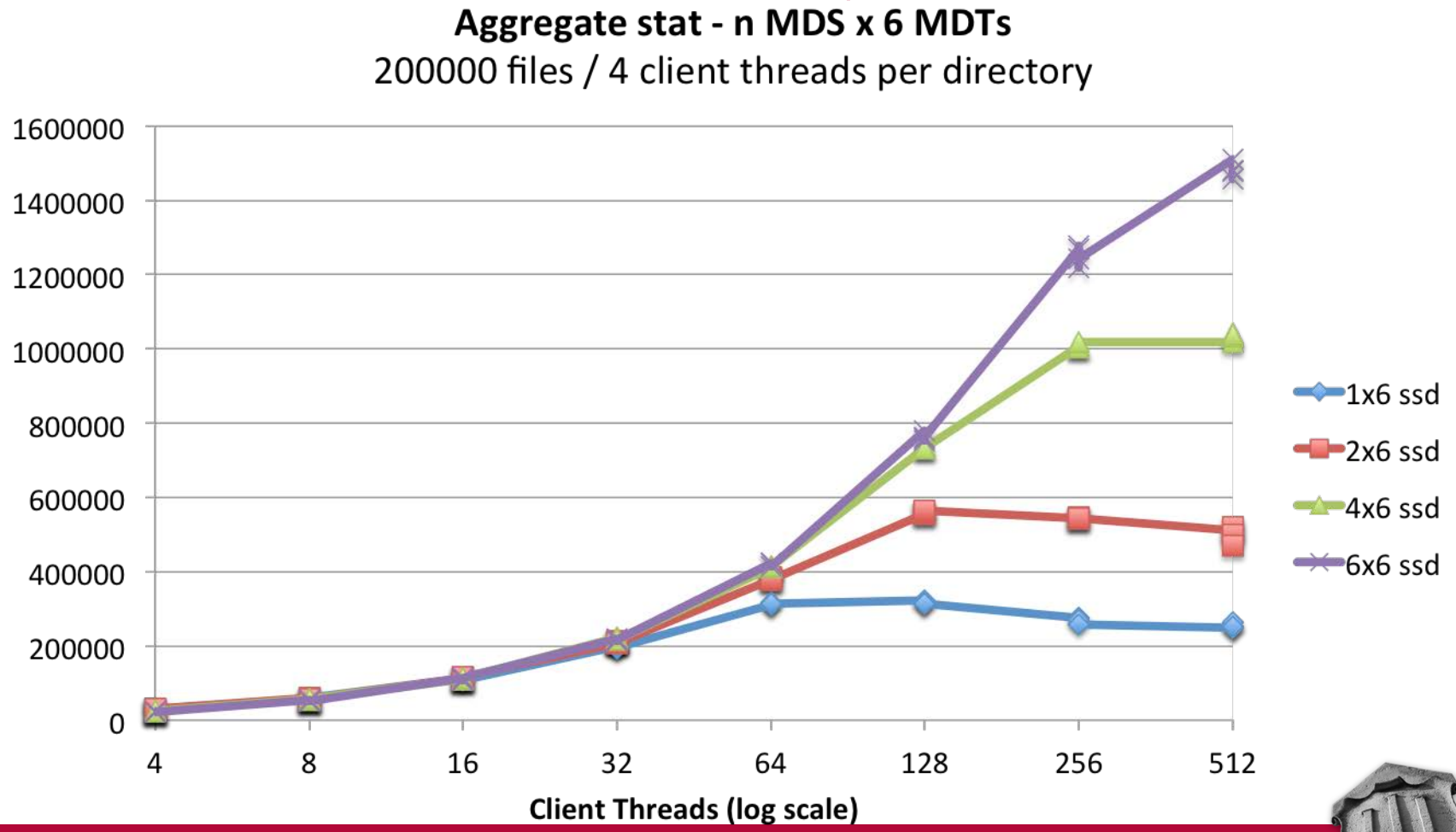
# Results – multi-MDS, single MDT



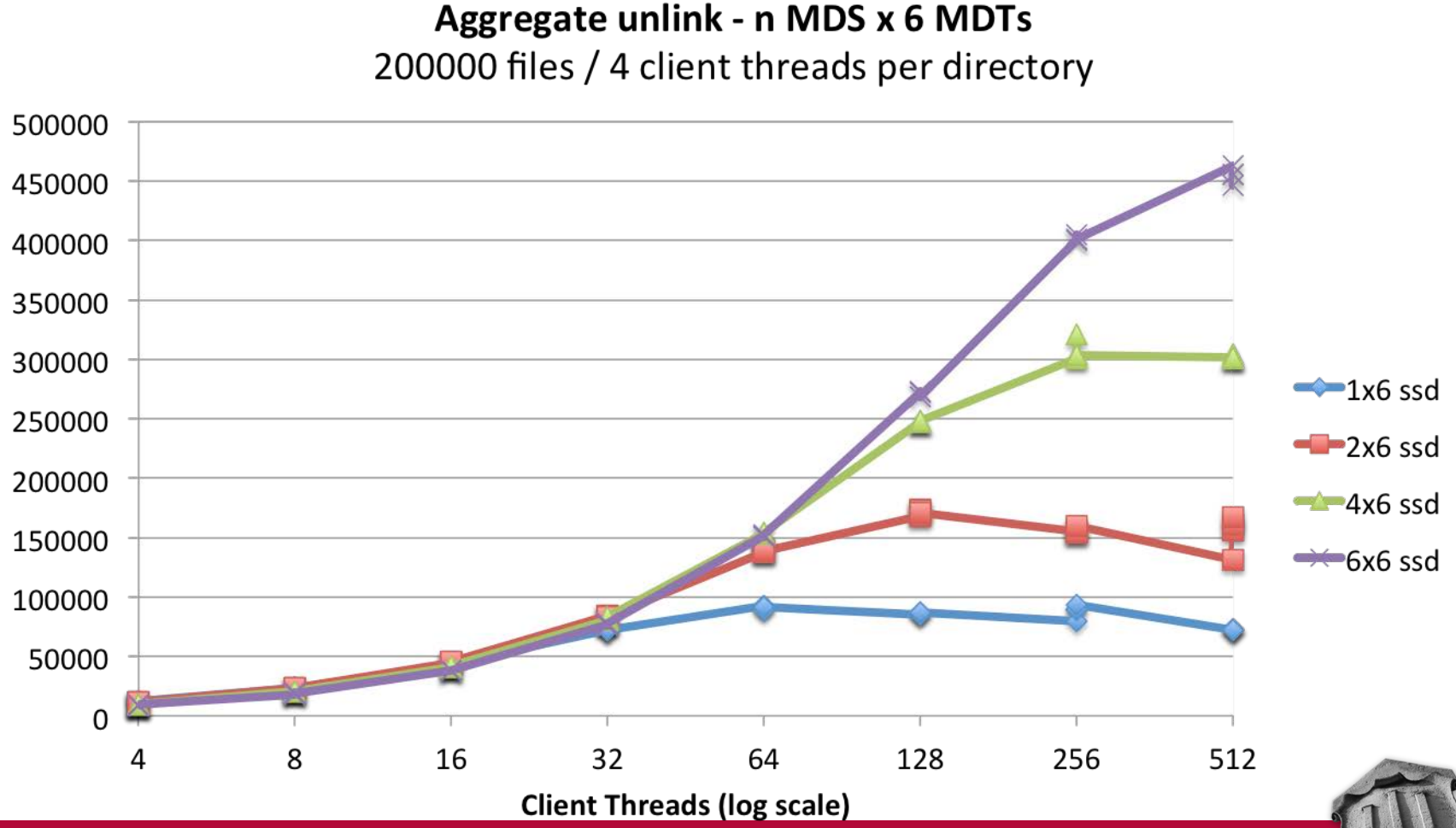
# Results – multi-MDS, single MDT



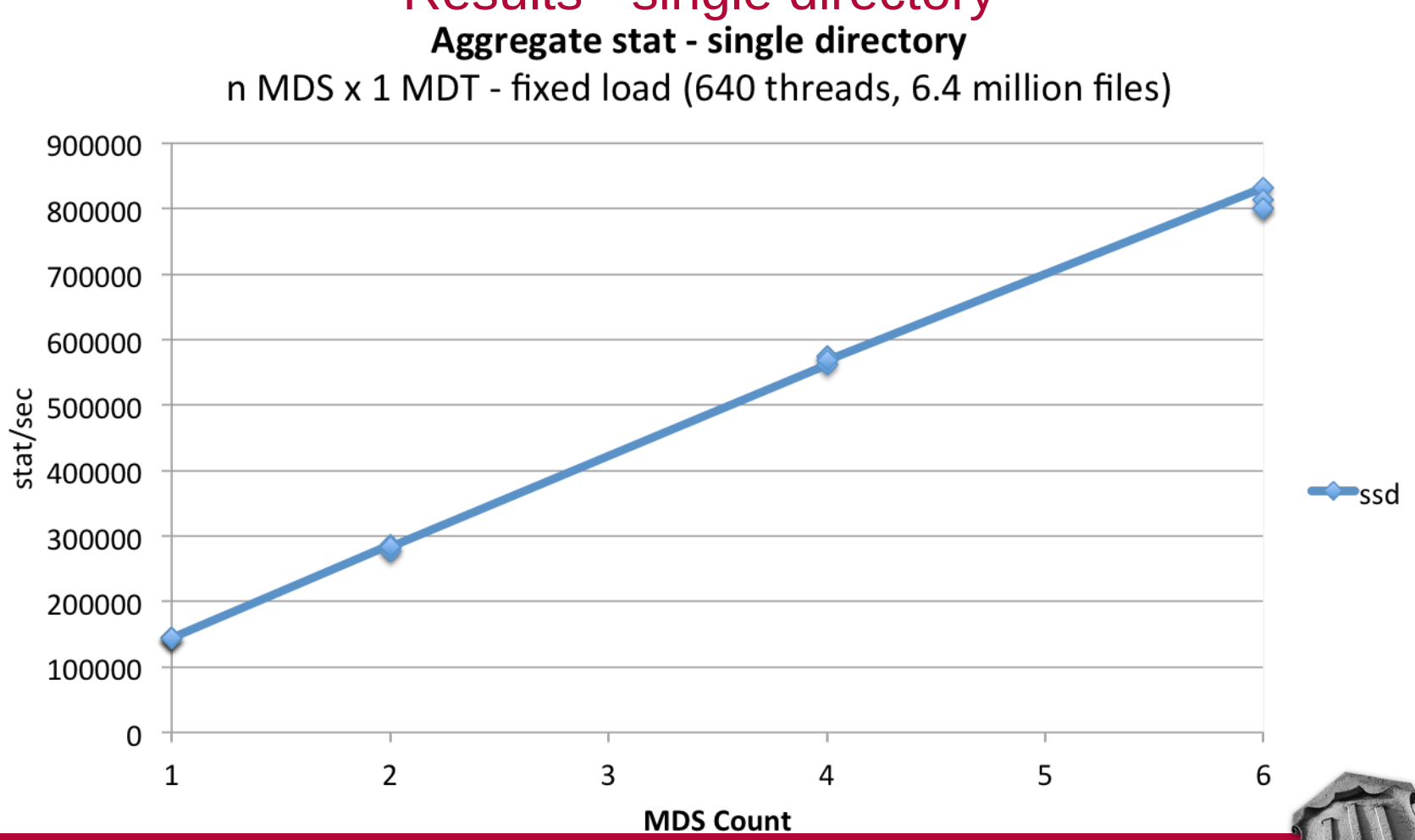
# Results – multi-MDS, multi-MDT



# Results – multi-MDS, multi-MDT



# Results - single directory



# Results - single directory