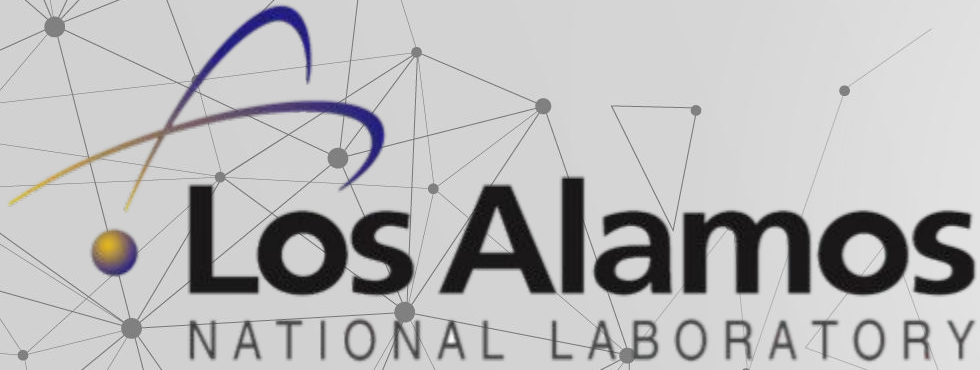
Can it scale? : Metadata Performance Testing of Lustre Dynamic Namespaces



Megan Booher
Colorado State University

Seema Kulkarni
University of Texas Austin

LA-UR-21-28065

CONTENTS

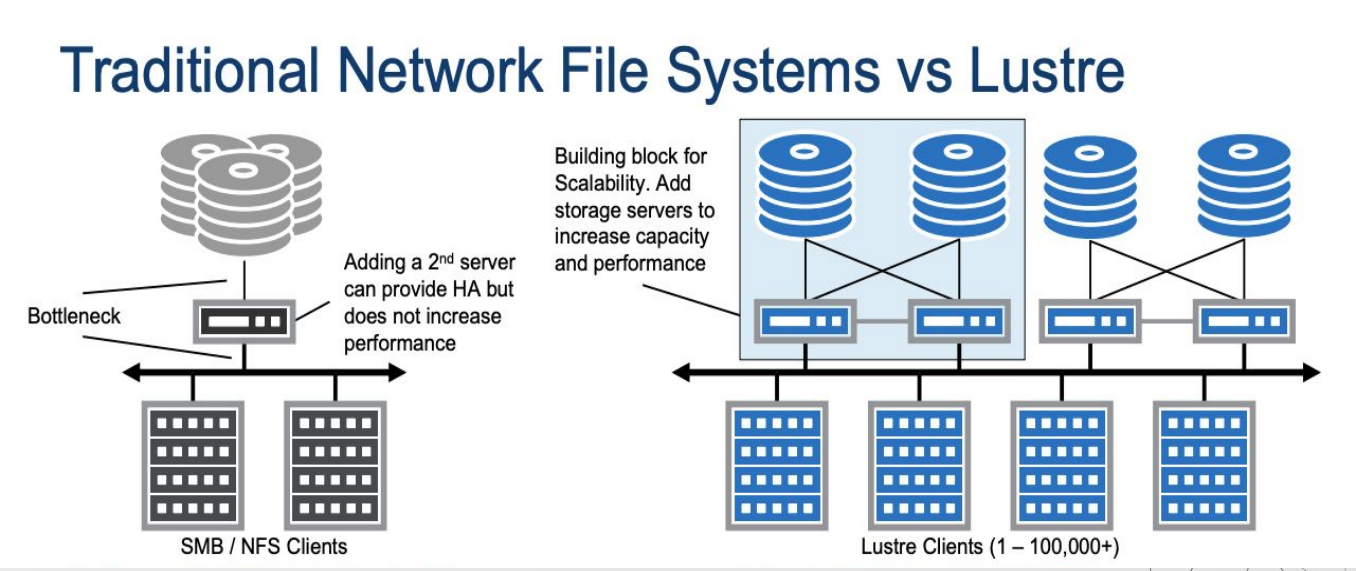
01

BACKGROUND

Background: Lustre

- Lustre HPC filesystem
  - Used to support the most demanding data-intensive applications

- Lustre provides massive scalable storage some of the largest supercomputers.



Background: Lustre

POINTS OF INTEREST	LUSTRE	NFS
EASE OF STORAGE SERVER ADDITION	✓ Seamless server addition	✗ Storage server addition creates separate file system
NAMESPACE	✓ Single coherent namespace across all servers	✗ Split namespace
CLIENT & SERVER RELATIONSHIP	✓ One to many	✗ One to one

Lustre Building Blocks

Management Server (MGS)

- Stores configuration information, and file system registries

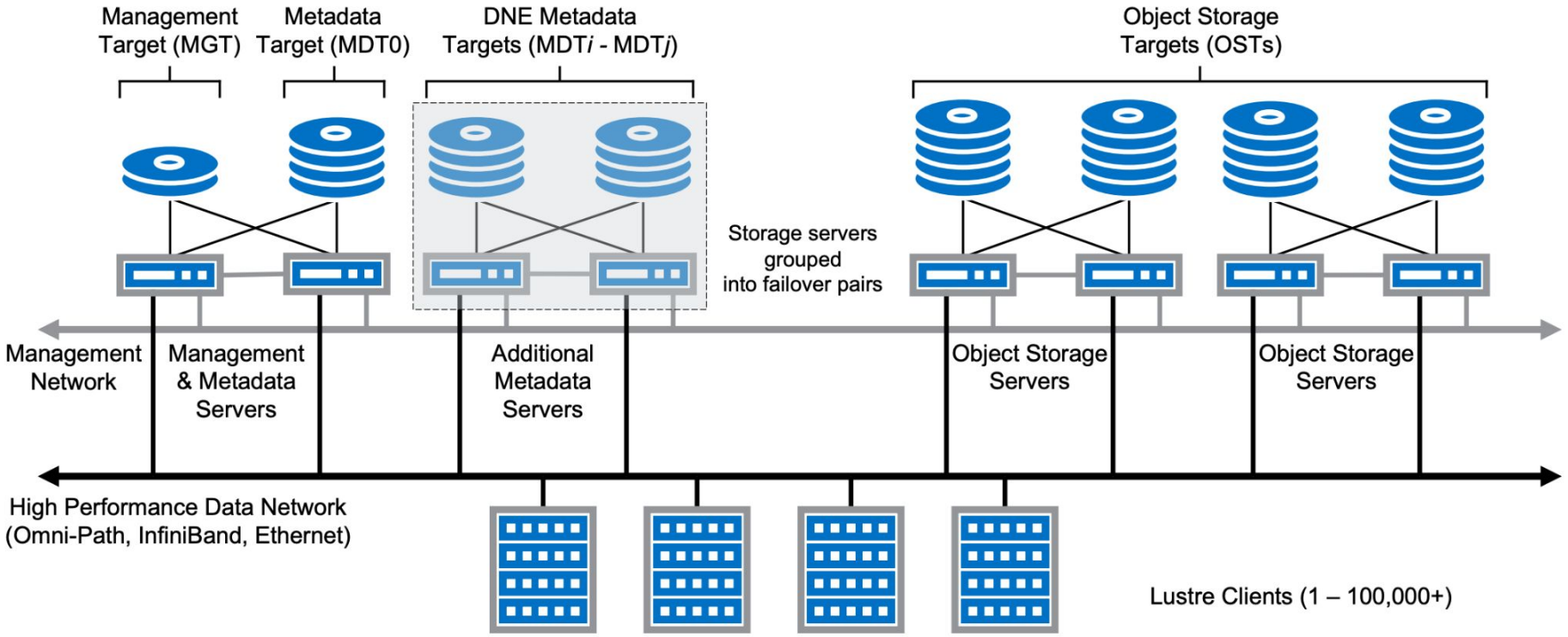
Object Storage Server (OSS)

- Record file content by striping across object storage targets (OST) for scalable performance

Metadata Server (MDS)

- Records Dynamic Namespace (DNE) and file system index of the filesystem

Lustre Scalable Storage



DNE Versions

DNE Lustre v1:

- Manual metadata striping across MDTs
- Difficult for inexperienced users

DNE Lustre v2:

- Automated metadata striping across MDTs
- Previous versions of v2 have not shown scaling when the number of MDTs increases

MDTest on Five MDTs

DNE v1:

```
# create v1 MDTs
for i in {0..4}
    lfs mkdir -c 1 -i $i mdt0$i

# run MDTest
mdtest -I <operations> -i <iterations> -u -d mdt00@mdt01@mdt02@mdt03@mdt04
```

DNE v2:

```
# create v2 MDT
lfs mkdir -c 5 mdt_v2
lfs mkdir -c 5 -D mdt_v2

# run MDTest
mdtest -I <operations> -i <iterations> -u -d mdt_v2
```

02

PROBLEM

Los Alamos
NATIONAL LABORATORY

Problem

DNE v2 has been updated in the recent past to improve performance.

We are interested in benchmarking metadata performance differences in Lustre DNE v1 and the latest version of DNE v2.

Ideally, v2's performance will scale linearly with increasing MDTs, similar to v1

Motivation

At LANL, Lustre is used in metadata heavy computations such as Artificial Intelligence, Climate Modeling, ect.

Ease of use

- Manual creation and specification of targets with DNE v1 vs automated target creation and specification with DNE v2

Reduce Bottlenecks

- Alleviates pressure from inefficient code, excess consumption of resources on an MDT on DNE v1 by distributing file creation on various MDTs in DNE v2

Speed up runtime

- Reduce runtime by removing walled metadata server behavior on DNE v1 and aggregating resources with DNE v2

03

METHODOLOGY

Los Alamos
NATIONAL LABORATORY

Methodology

STEP 1

Baseline Metadata Target Testing

Ensure that all MDTs are working properly

STEP 2

v1 Hero Testing

Collect data about v1 efficiency when increasing the number of MDTs used

STEP 3

v2 Hero Testing

Collect data about v2 efficiency when increasing the number of MDTs used

STEP 4

Compare & Analyze Results

How does v2 compare to v1? Is using v2 something that we recommend?

Tools Used

Metadata Target (MDT)

- High performance storage target shared by multiple Metadata Servers
- Five were used for v1 and v2

lfs mkdir

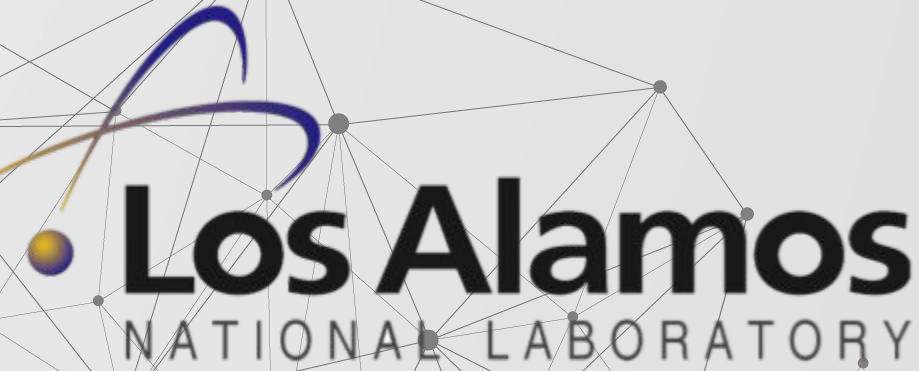
- Creates a striped directory on a specific MDT

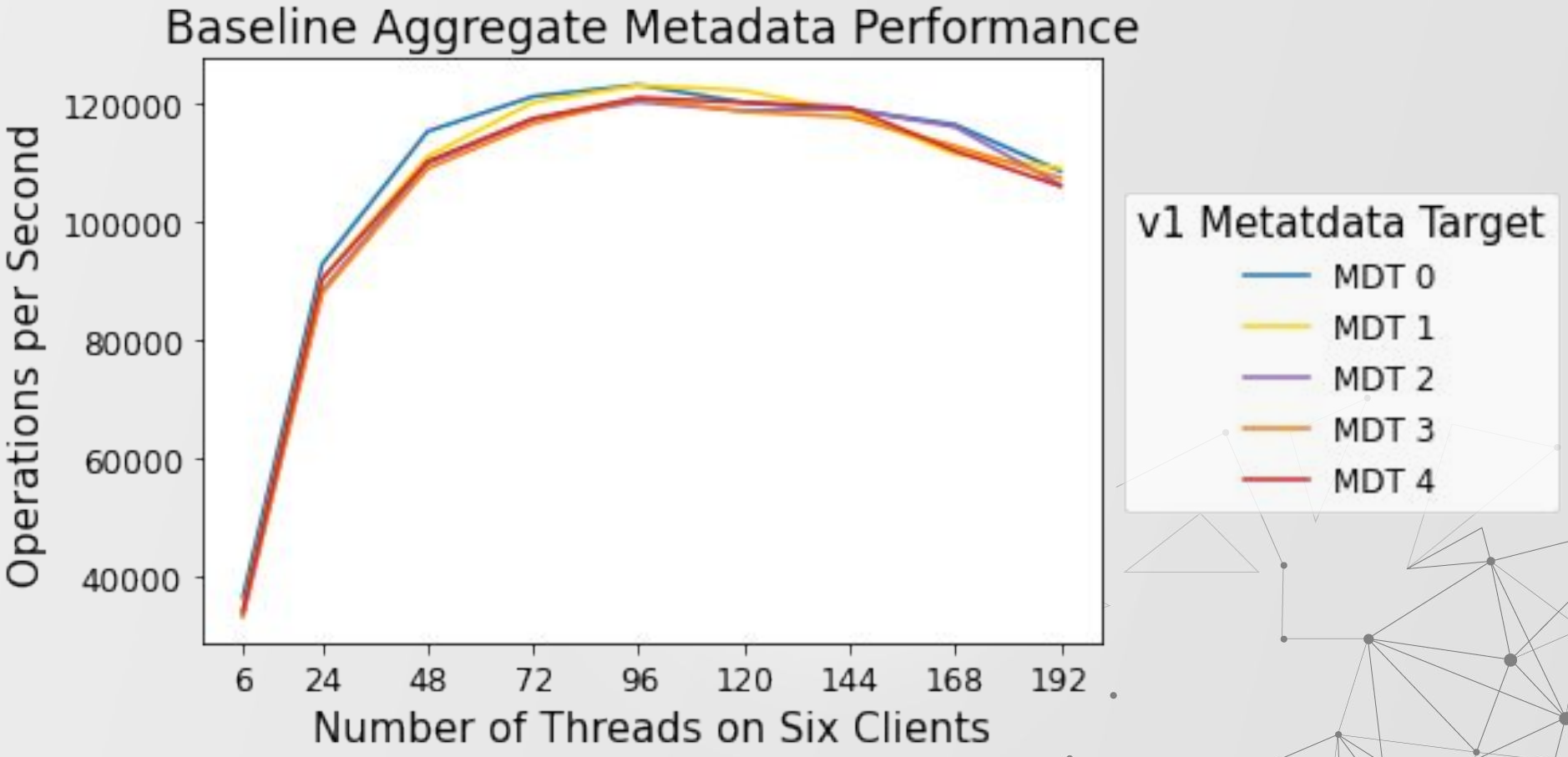
MDTest

- Evaluates the metadata performance of a parallel file system
- This tool was used to collect metadata performance speeds of v1 and v2

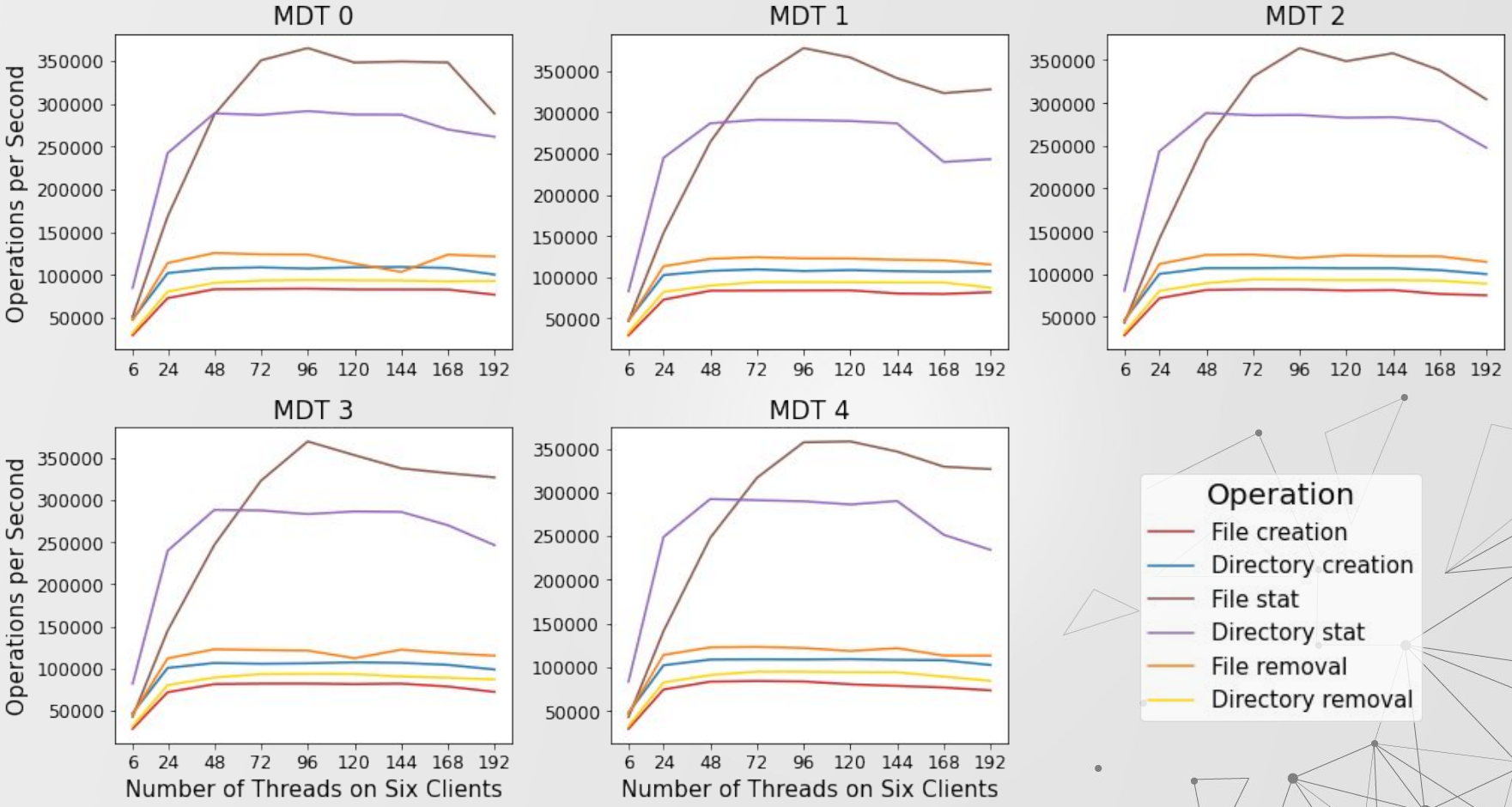
04

BASELINE TESTING




Baseline Aggregate Metadata Performance
Operations per Second
120000
100000
80000
60000
40000
6
24
48
72
96
120
144
168
192
Number of Threads on Six Clients
v1 Metatdata Target
MDT 0
MDT 1
MDT 2
MDT 3
MDT 4

Baseline Operation Metadata Performance



Baseline Conclusion

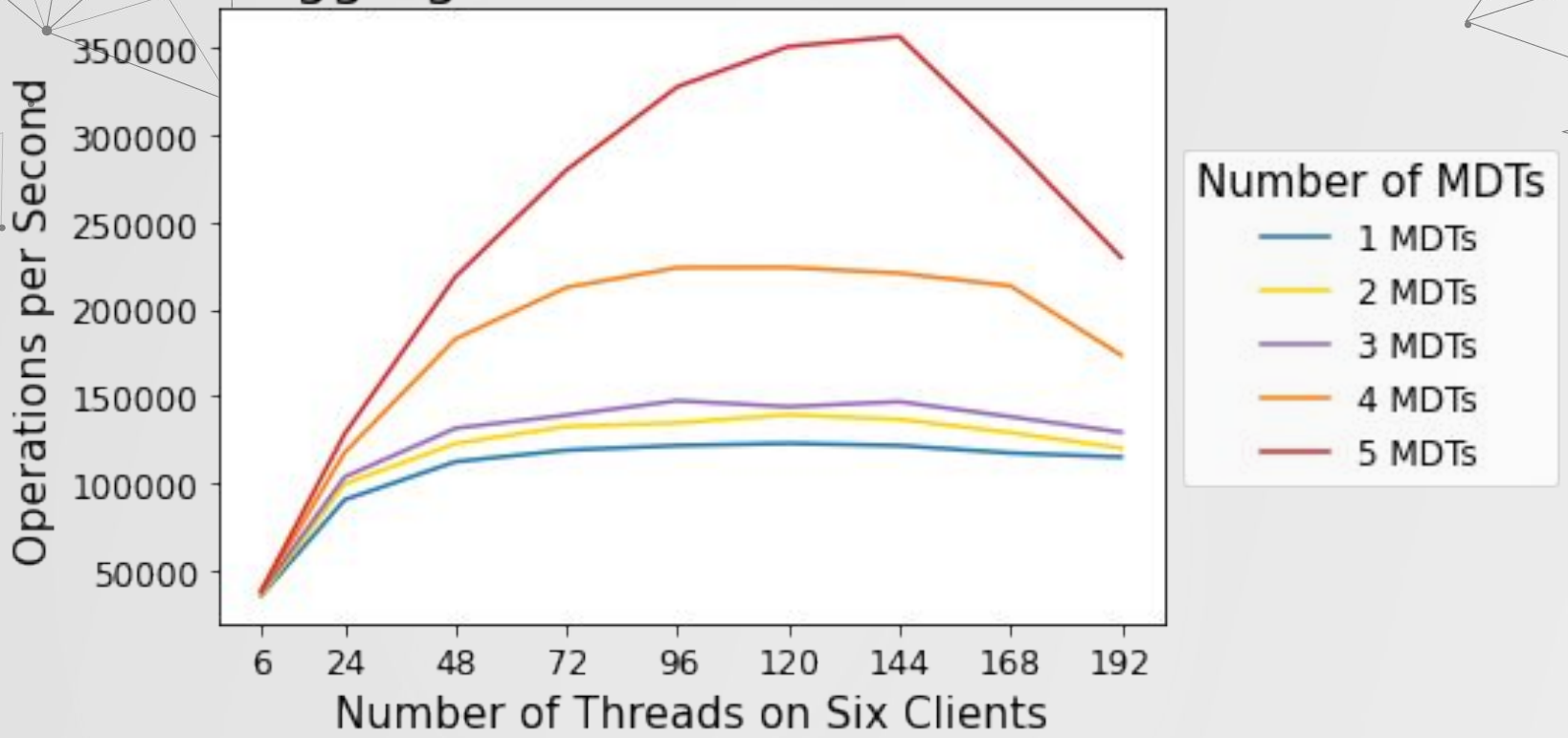
- All MDTs are operating at the same speeds, none are broken!

- This was necessary to ensure that the data we receive from later tests is as accurate as possible.

04

V1 TESTING

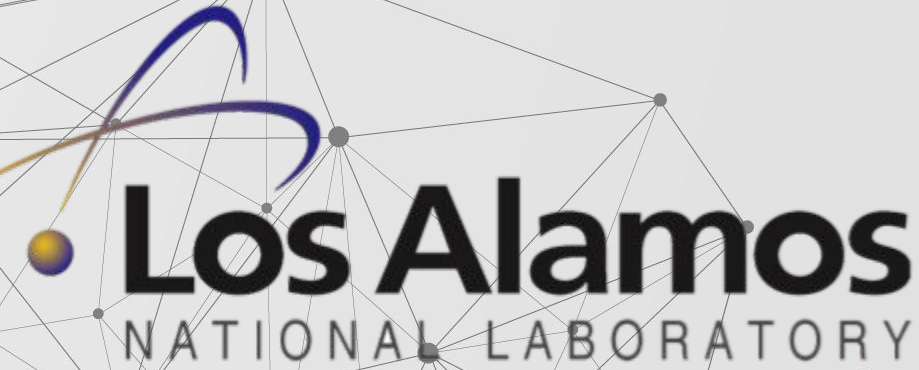
v1 Aggregate Metadata Performance

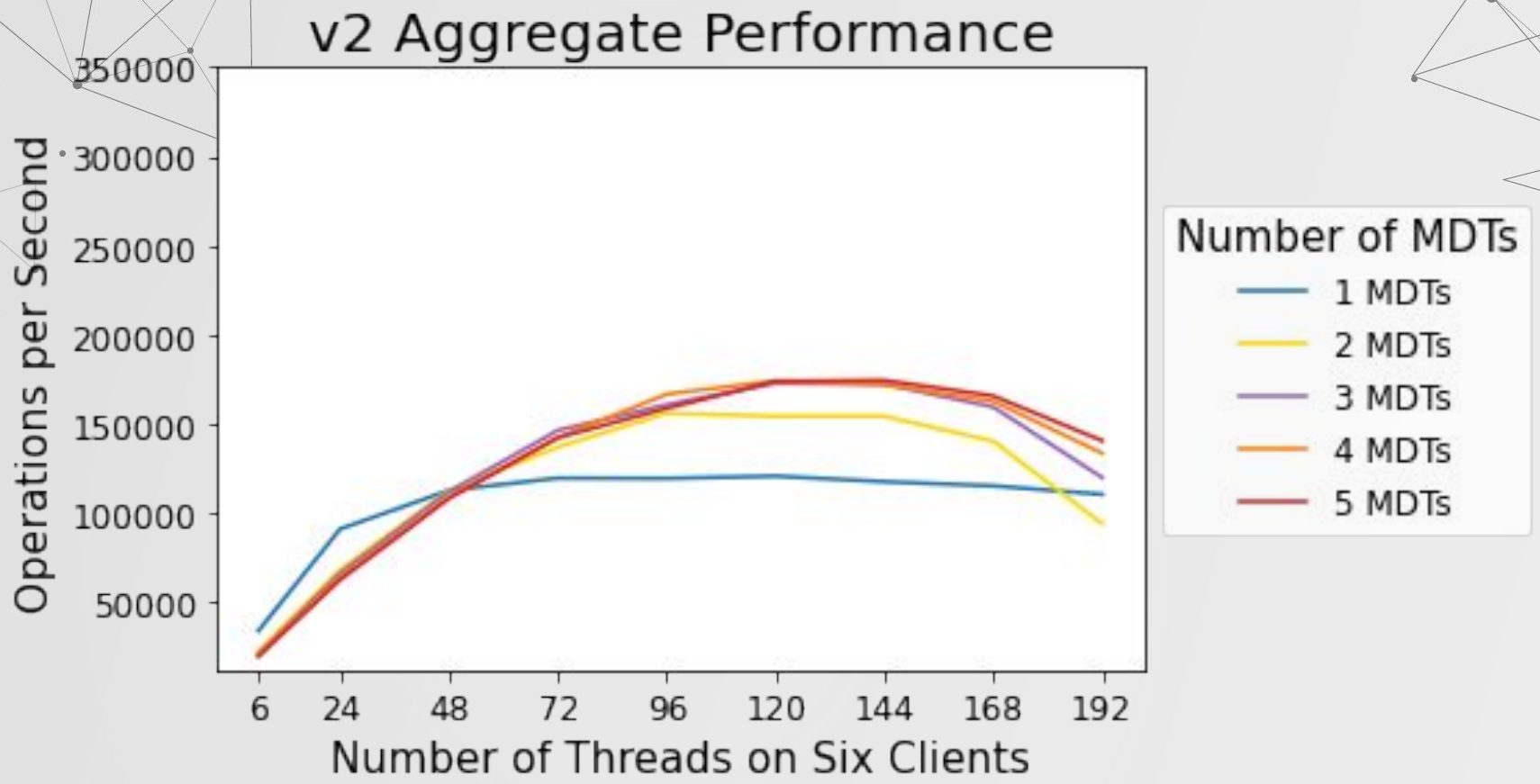


Based on our v1 metadata tests, it can be can see that Lustre's DNE v1 Metadata performance scales positively with the number of MDTs

04

V2 TESTING

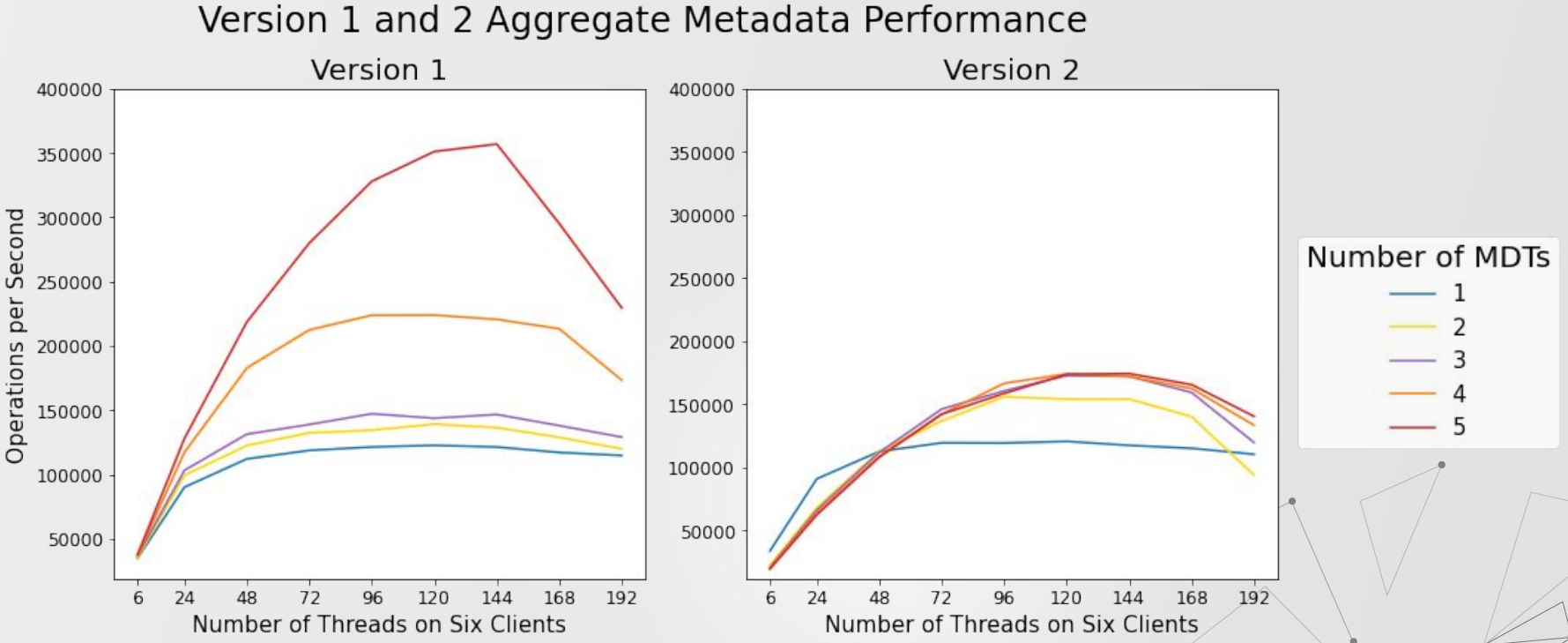




DNE v2 does not see scaling when the number of MDT's being used increases.

05

CONCLUSION



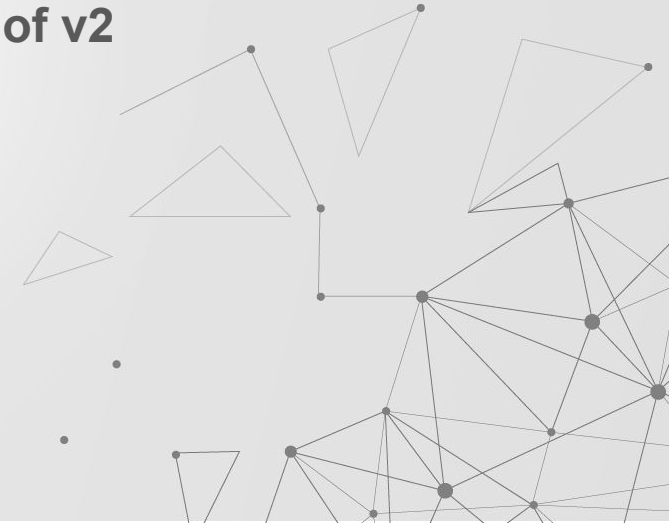
Unlike DNE v1, DNE v2 does not see linear scaling in metadata performance when the number of MDTs being used increases.

Because of this, the use of Lustre's DNE v2 is not suggested.

Future Considerations

- Understand why v2 does not scale.
  - Where are bottlenecks in v2?

- Continue testing newly released versions of v2



06
ACKNOWLEDGEMENTS

Acknowledgements

Alex Parga

Jarrett Crews

Dominic Manno

Julie Wiens

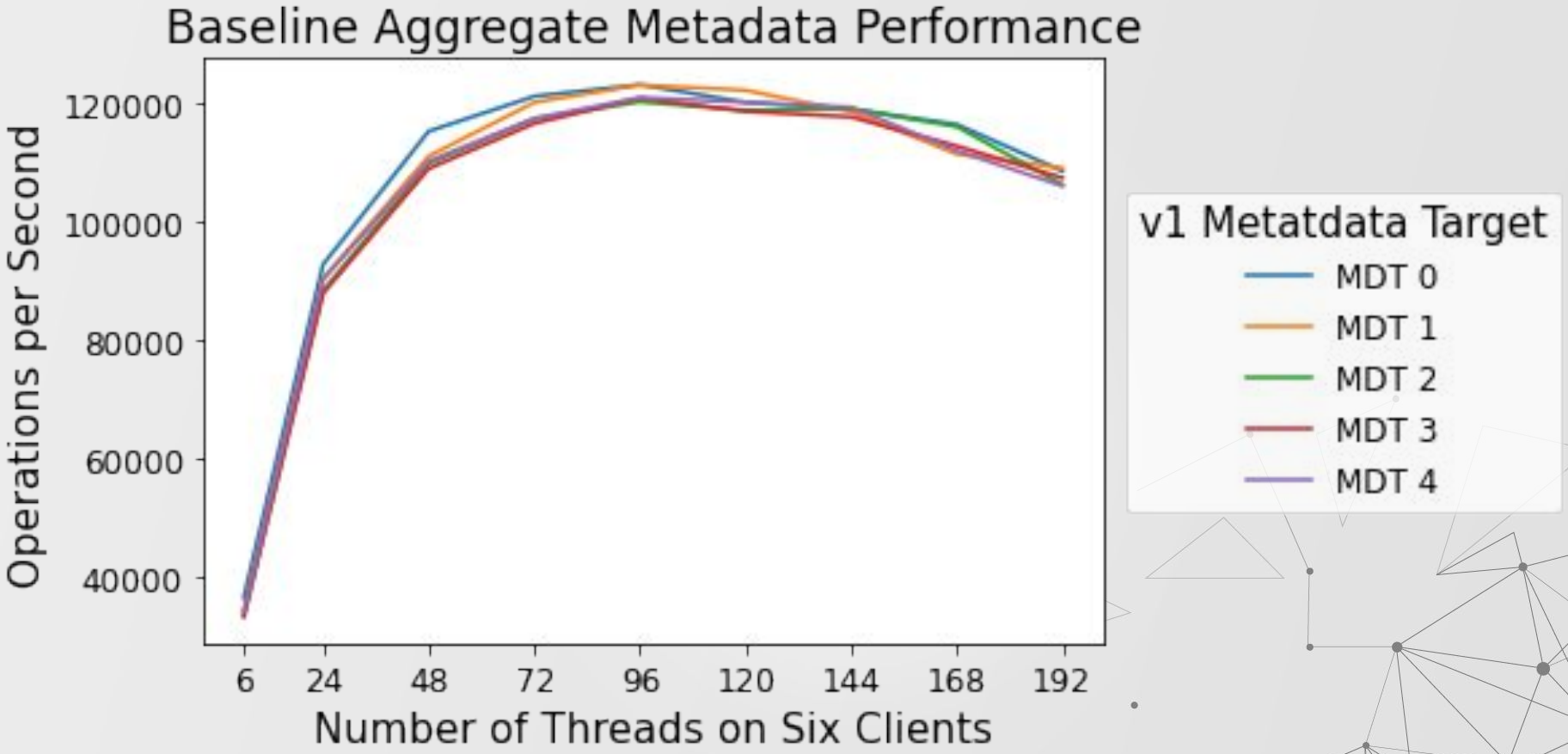
Thank you!

Megan Booher

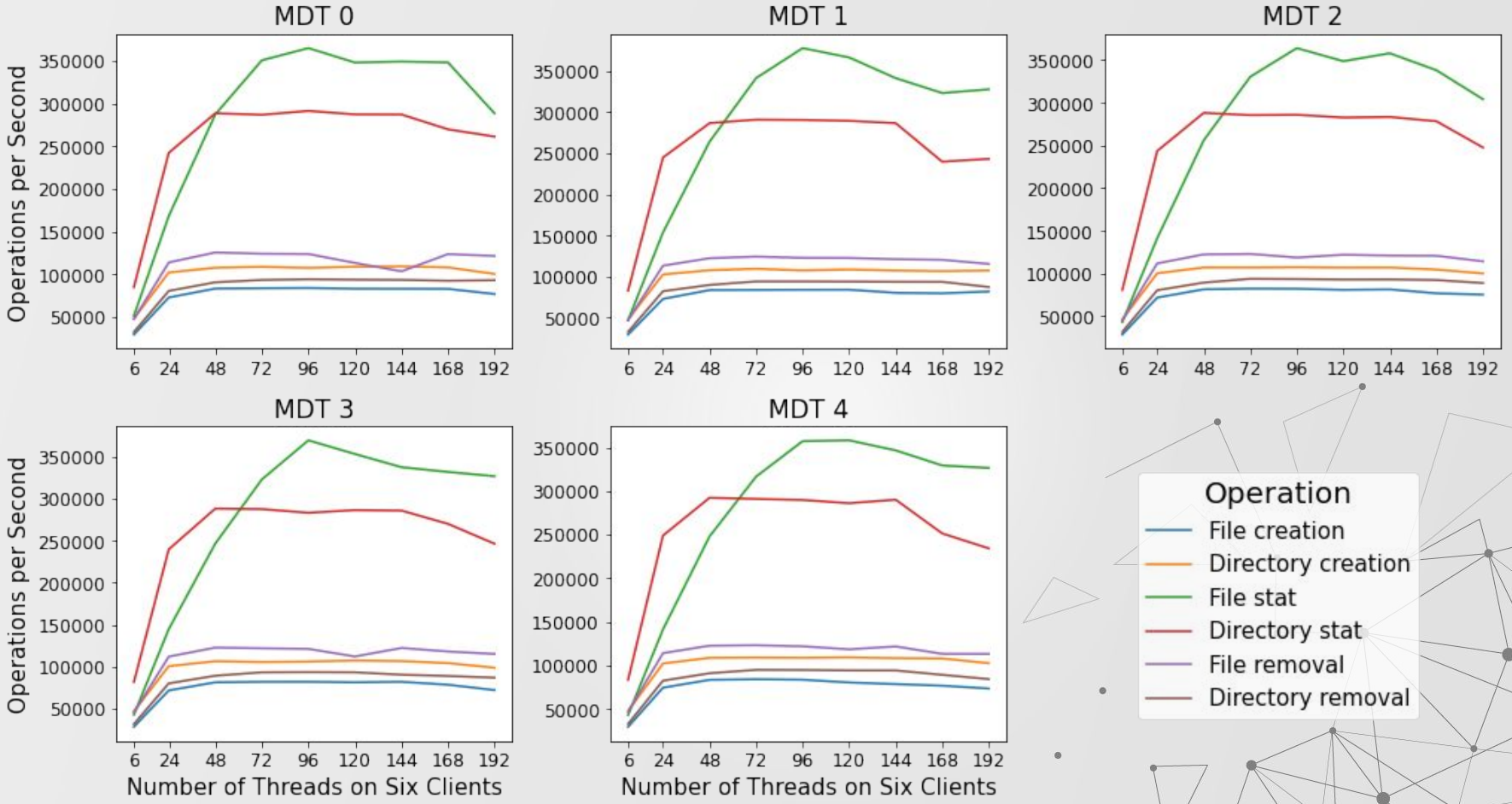
meganboo@rams.colostate.edu

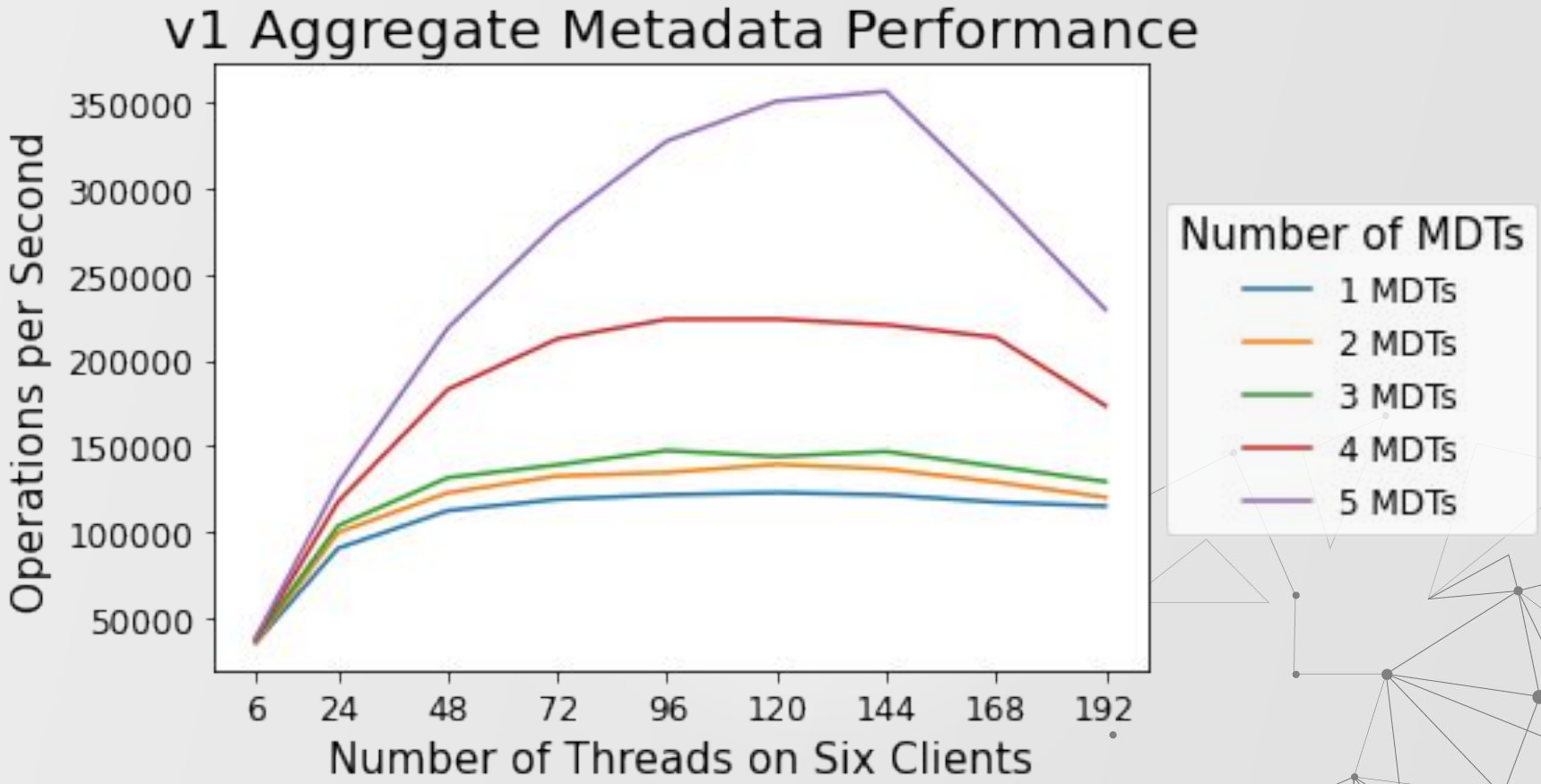
Seema Kulkarni

seemakulkarni@utexas.edu

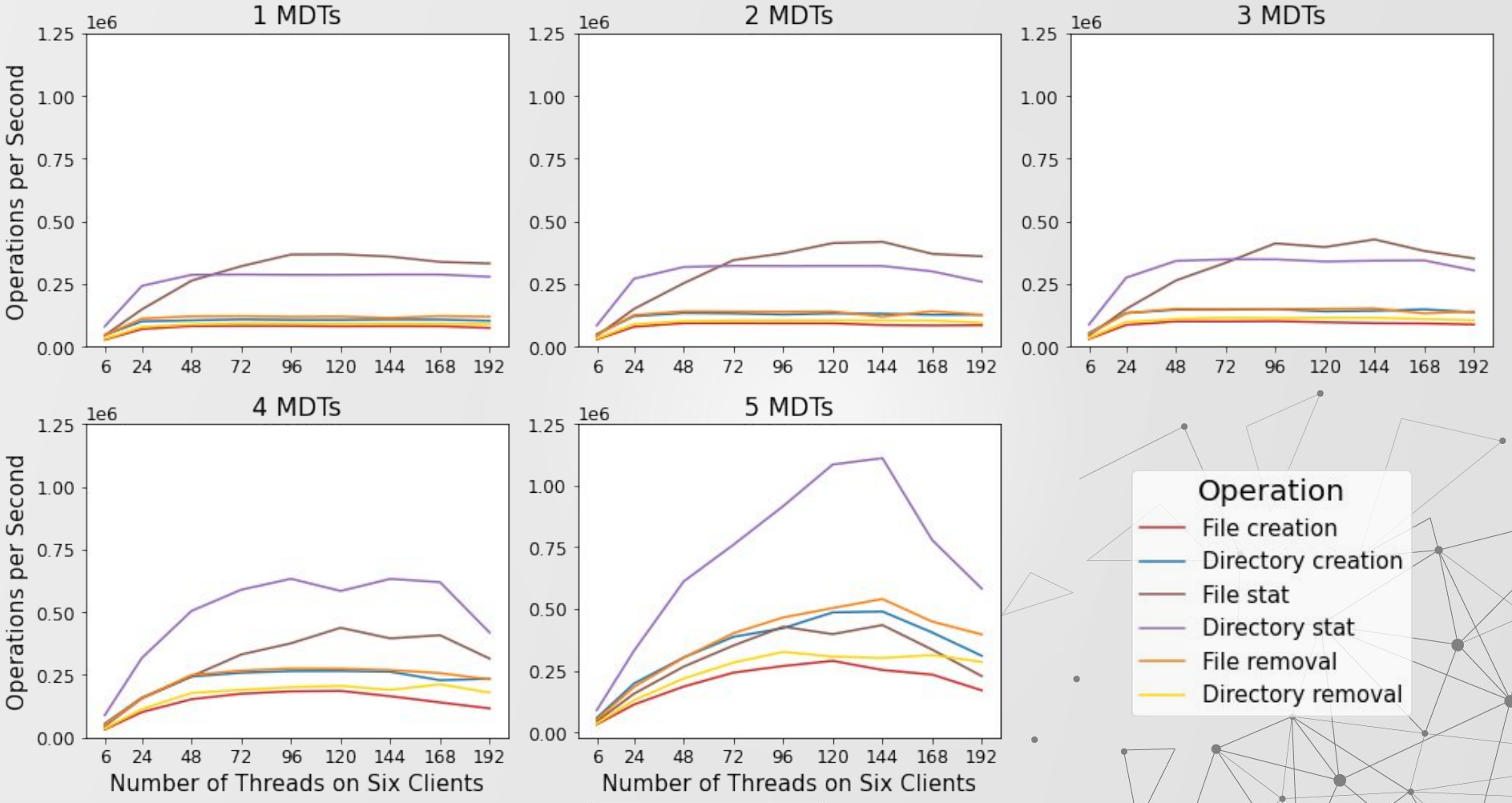

Baseline Aggregate Metadata Performance
Operations per Second
120000
100000
80000
60000
40000
6
24
48
72
96
120
144
168
192
Number of Threads on Six Clients
v1 Metatdata Target
MDT 0
MDT 1
MDT 2
MDT 3
MDT 4

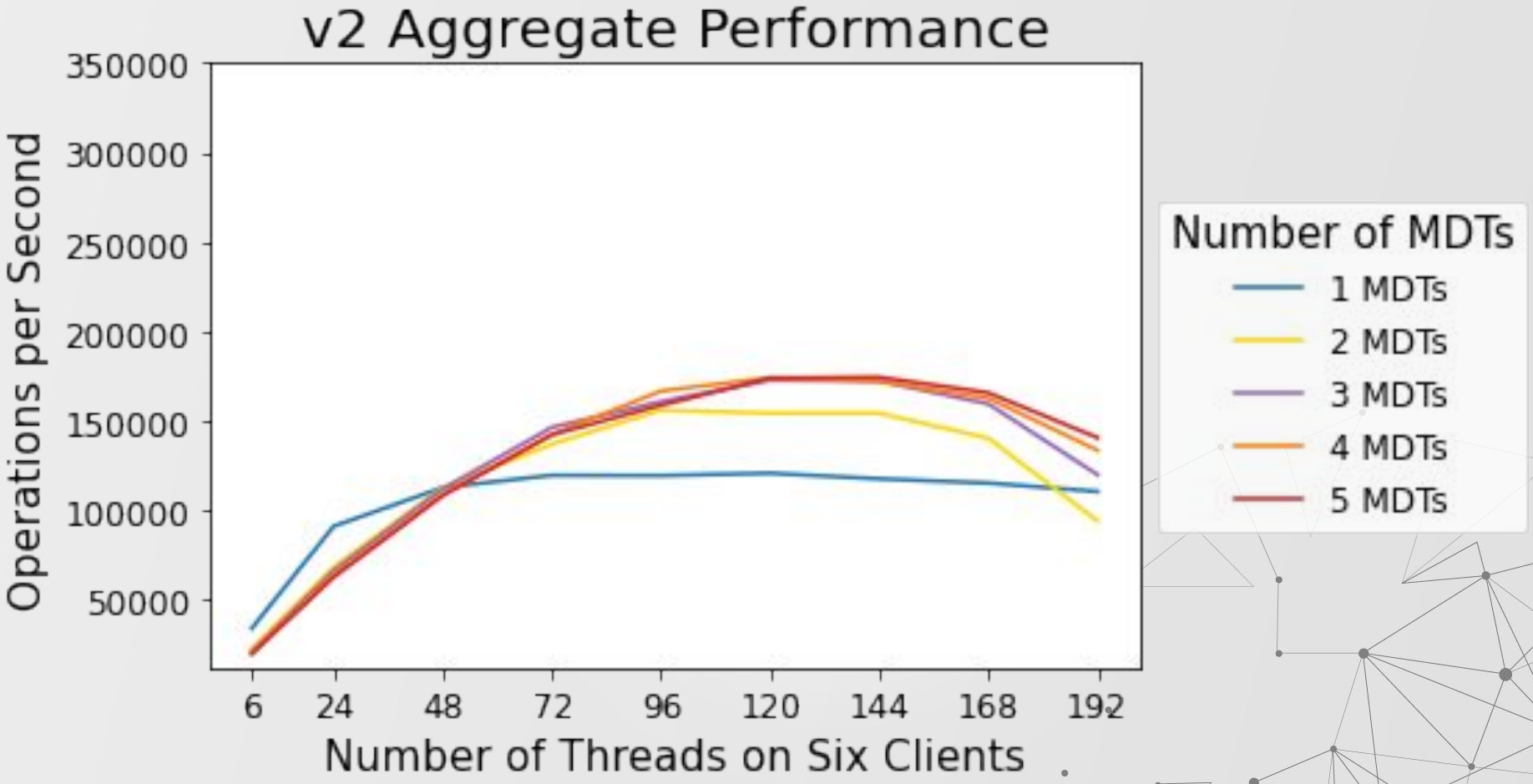
Baseline Operation Metadata Performance



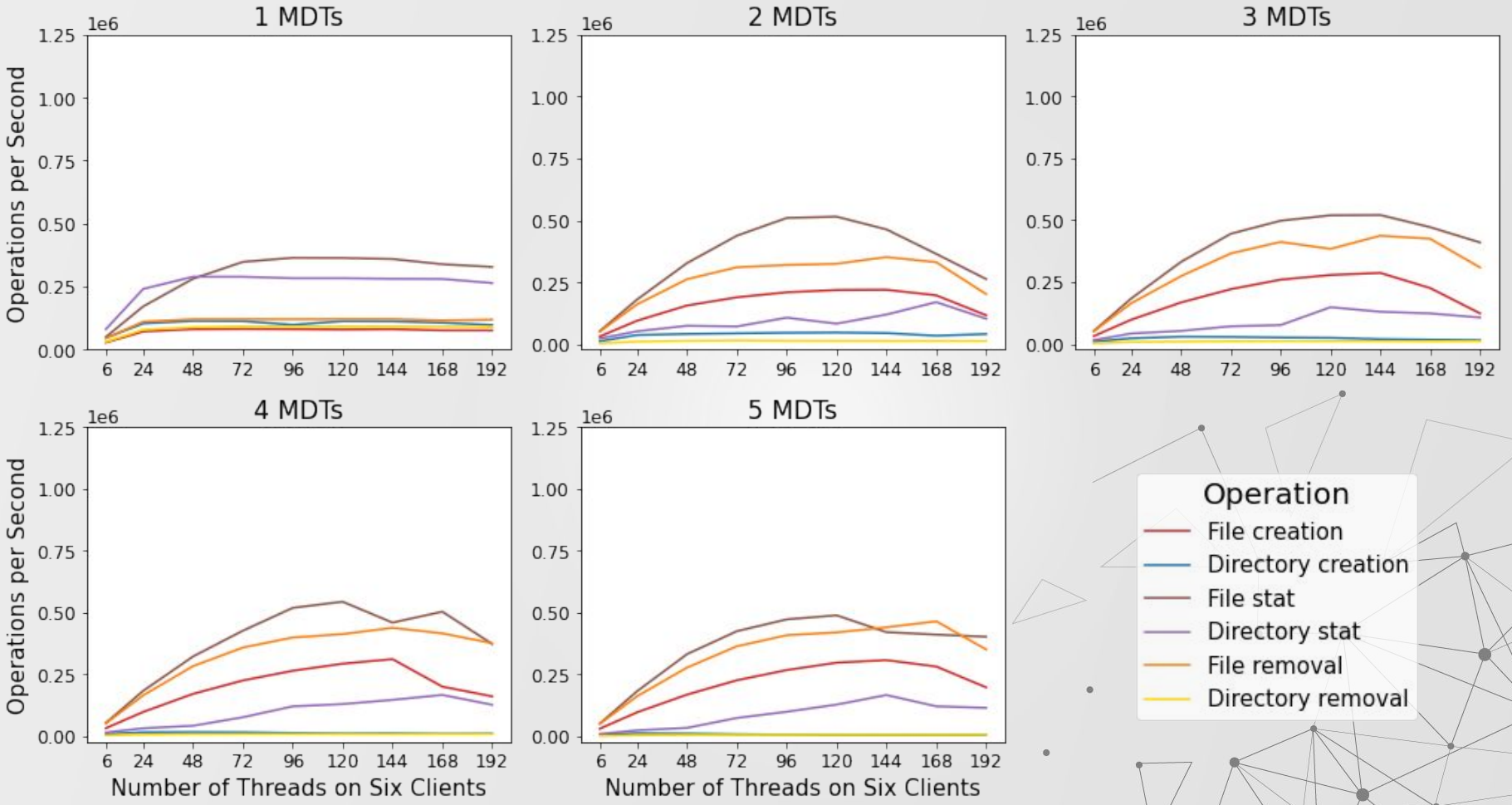

v1 Aggregate Metadata Performance
Operations per Second
350000
300000
250000
200000
150000
100000
50000
6
24
48
72
96
120
144
168
192
Number of Threads on Six Clients
Number of MDTs
1 MDTs
2 MDTs
3 MDTs
4 MDTs
5 MDTs

v1 Operation Metadata Performance

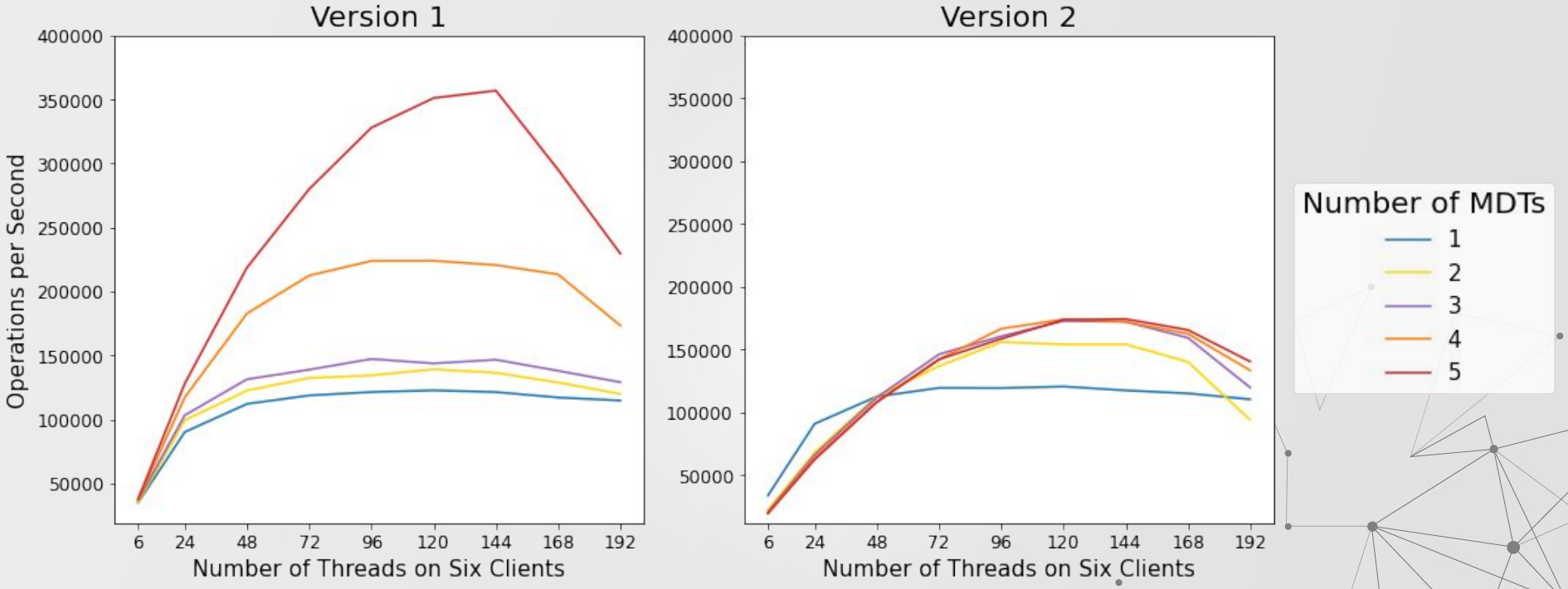


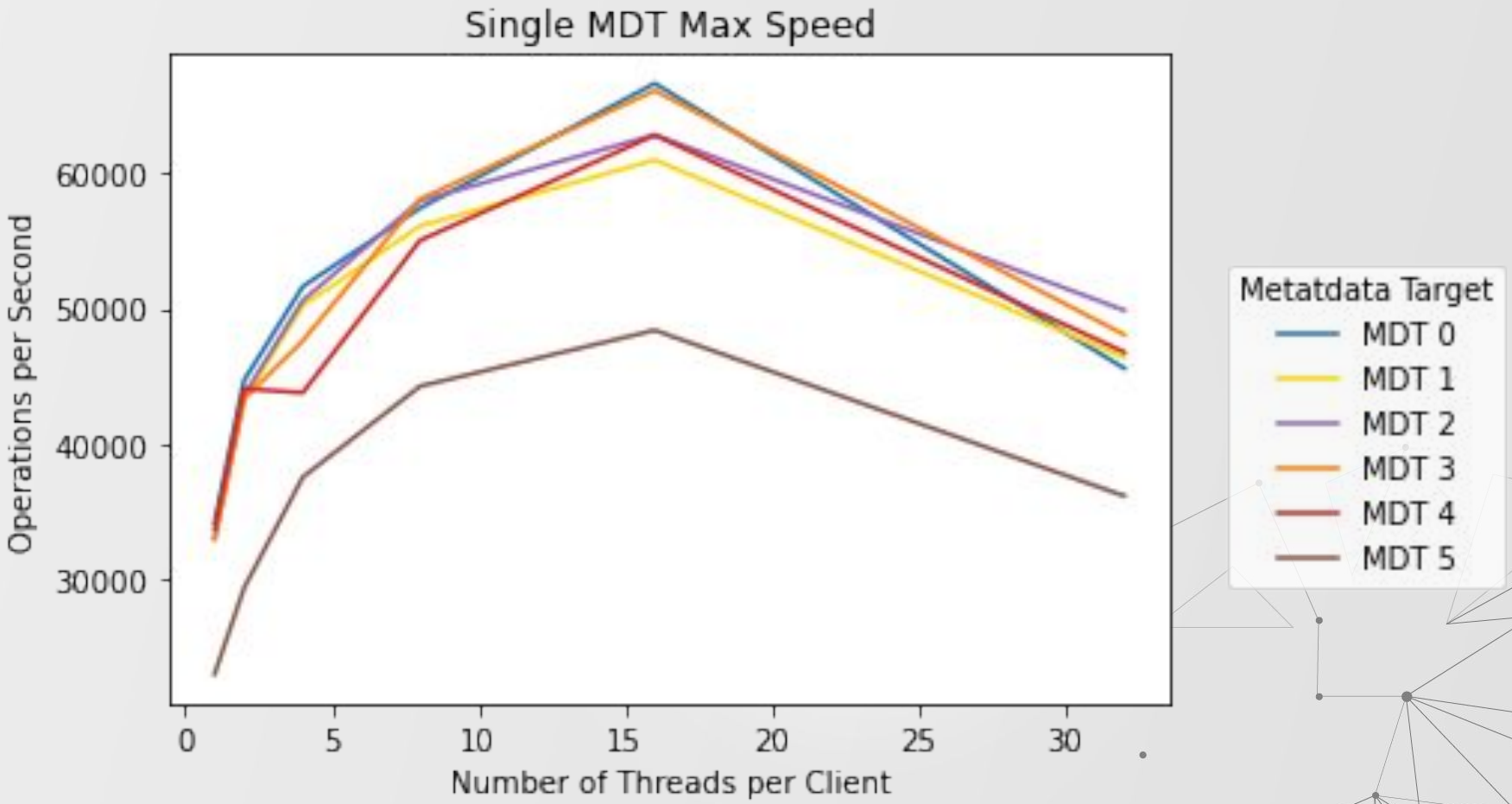

v2 Aggregate Performance
Operations per Second
350000
300000
250000
200000
150000
100000
50000
6
24
48
72
96
120
144
168
192
Number of Threads on Six Clients
Number of MDTs
1 MDTs
2 MDTs
3 MDTs
4 MDTs
5 MDTs

v2 Operation Metadata Performance



Version 1 and 2 Aggregate Metadata Performance




Single MDT Max Speed
Operations per Second
60000
50000
40000
30000
0
5
10
15
20
25
30
Number of Threads per Client
Metatdata Target
MDT 0
MDT 1
MDT 2
MDT 3
MDT 4
MDT 5