Look Inside.™

Deploying a Lustre* Cluster for HPC Applications in the Cloud

Gabriele Paciucci, Robert Read, Andrew Uselton

Presentation Outline

- Motivations
- Deploying Lustre* on AWS
- Benchmarks for different cluster topologies
- Running a real application
- Conclusion and Q&A

Presentation Outline

- Motivations
- Deploying Lustre* on AWS
- Benchmarks for different cluster topologies
- Running a real application
- Conclusion and Q&A

Motivation for AWS Lustre*

Amazon is growing its HPC capabilities, and we believe there are some HPC workloads moving to the cloud.

Amazon has several storage related services, such as EBS and S3, but there is no shared file system service.

Since many existing HPC applications have been built to assume a shared file system is available, it seems there is a need for a parallel file system like Lustre.

Virtual Hardware Available

Amazon EC2 instances:

- Spot
- EBS optimized
- High network capabilities (but always 1Gbps limited)

Amazon EBS storage:

- Networked storage
- Max size 1TB per EBS volume
- Not magic
- Standard, not Provisioned in our provisioning system

VMs size	vCPU	vRAM (GB)	EBS	Network MB/sec **
M1.medium	1	3.7	N/A	94+
M1.large	2	7.5	Yes	95+
M1.xlarge	4	15	Yes	110+
M3.2xlarge	8	30	Yes	110+
CC2.8xlarge	32	60	N/A	10 GbE

NEW

EBS Storage	IOPS	Size	Performance (WRITE) **
Standard	N/A	100	24+ MB/sec
Provisioned	2000	200	35+ MB/sec
Provisioned	4000	400	50+ MB/sec

** not intended to be authoritative numbers

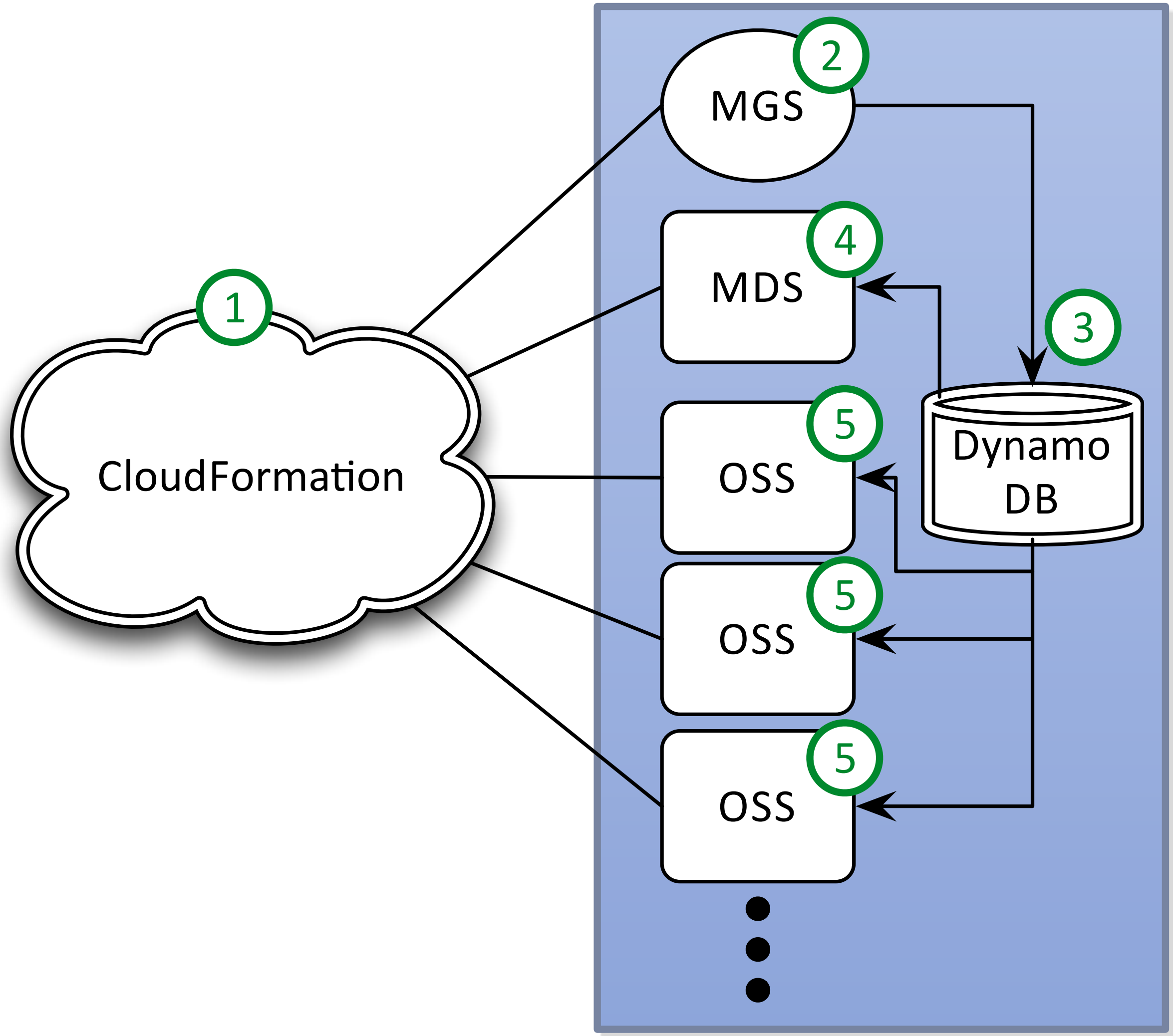
Presentation Outline

- Motivations
- Deploying Lustre* on AWS
- Benchmarks for different cluster topologies
- Running a real application
- Conclusion and Q&A

Deploying Lustre* on Amazon

- Custom Lustre Server AMI
  - Centos 6.4
  - Lustre 2.4
- Deploy cluster with Cloud Formation
  - number of nodes to create: OSS MDS Clients
  - the instance type to use: m1.xlarge / m3.2xlarge / cc2.8xlarge
  - disk size: OSS MDS
- Minimal coordination through Dynamo DB
- New file system is assembled as nodes boot
- Rich number of monitor tools available and configured
  - ltop, ganglia, lmt

Deploying workflow



1 CloudFormation creates a stack of AWS resources from a template

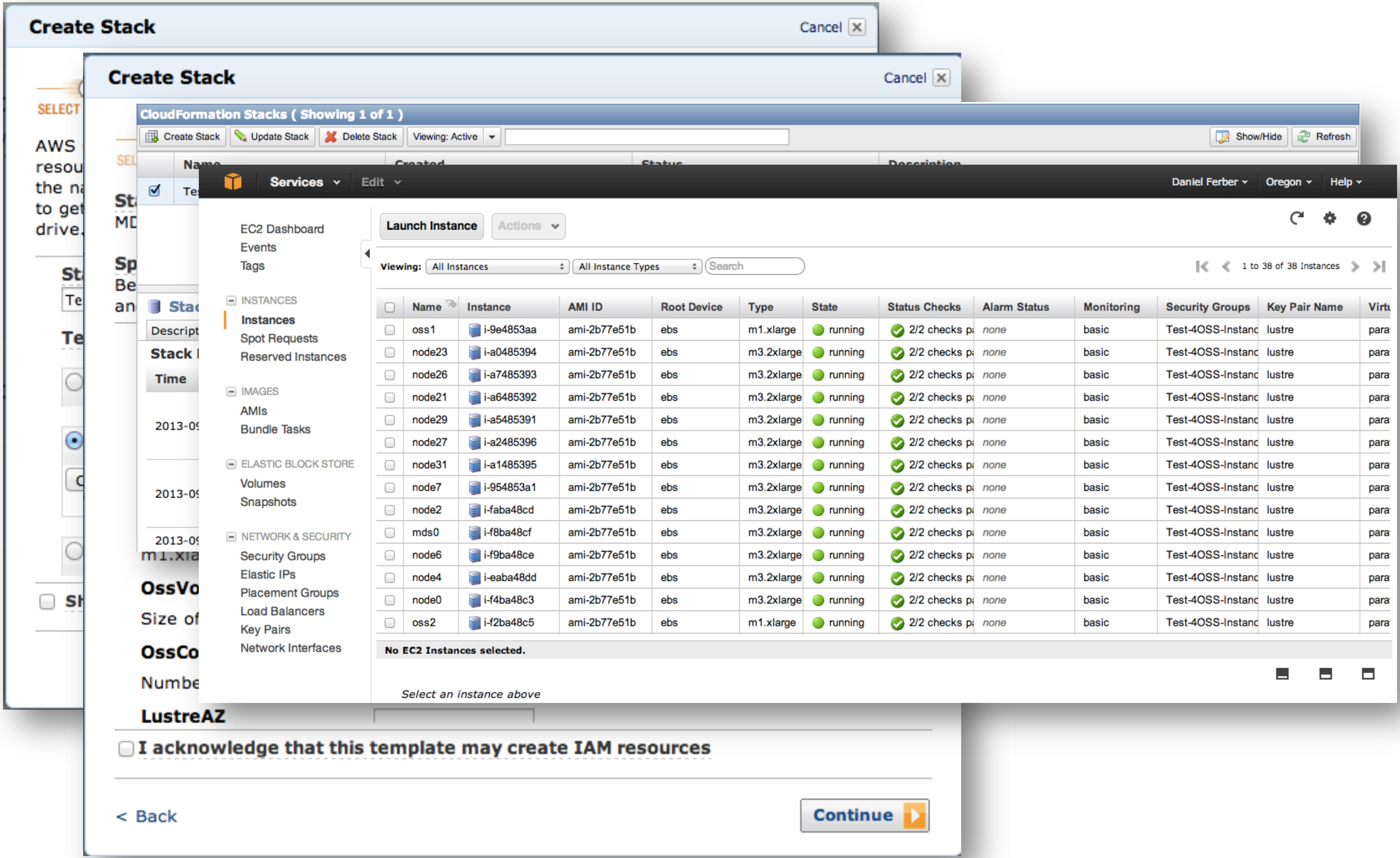
2 MGS Initializes itself

3 MGS updates DB with NID

4 MDS formats MDT, registers with MGS, updates DB.

5 OSSs format local targets, updates DB

Deploying, from a user perspective



Create Stack Cancel ☒

SELECT

AWS resou the na to get drive.

St Te

Te

○ ● C ○

□ Sh

Create Stack Cancel ☒

SEL

St MD

Sp Be an

m1.xla

OssVo

Size of

OssCo

Numbe

LustreAZ

□ I acknowledge that this template may create IAM resources

< Back

Continue

CloudFormation Stacks (Showing 1 of 1)

Create Stack | Update Stack | Delete Stack | Viewing: Active | Show/Hide | Refresh

	Name	Created	Status	Description
☑	Te			

Stac

Descript

Stack

Time

2013-09

2013-09

2013-09

EC2 Dashboard
Events
Tags

INSTANCES
Instances
Spot Requests
Reserved Instances

IMAGES
AMIs
Bundle Tasks

ELASTIC BLOCK STORE
Volumes
Snapshots

NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Load Balancers
Key Pairs
Network Interfaces

Launch Instance | Actions

Viewing: All Instances | All Instance Types | Search

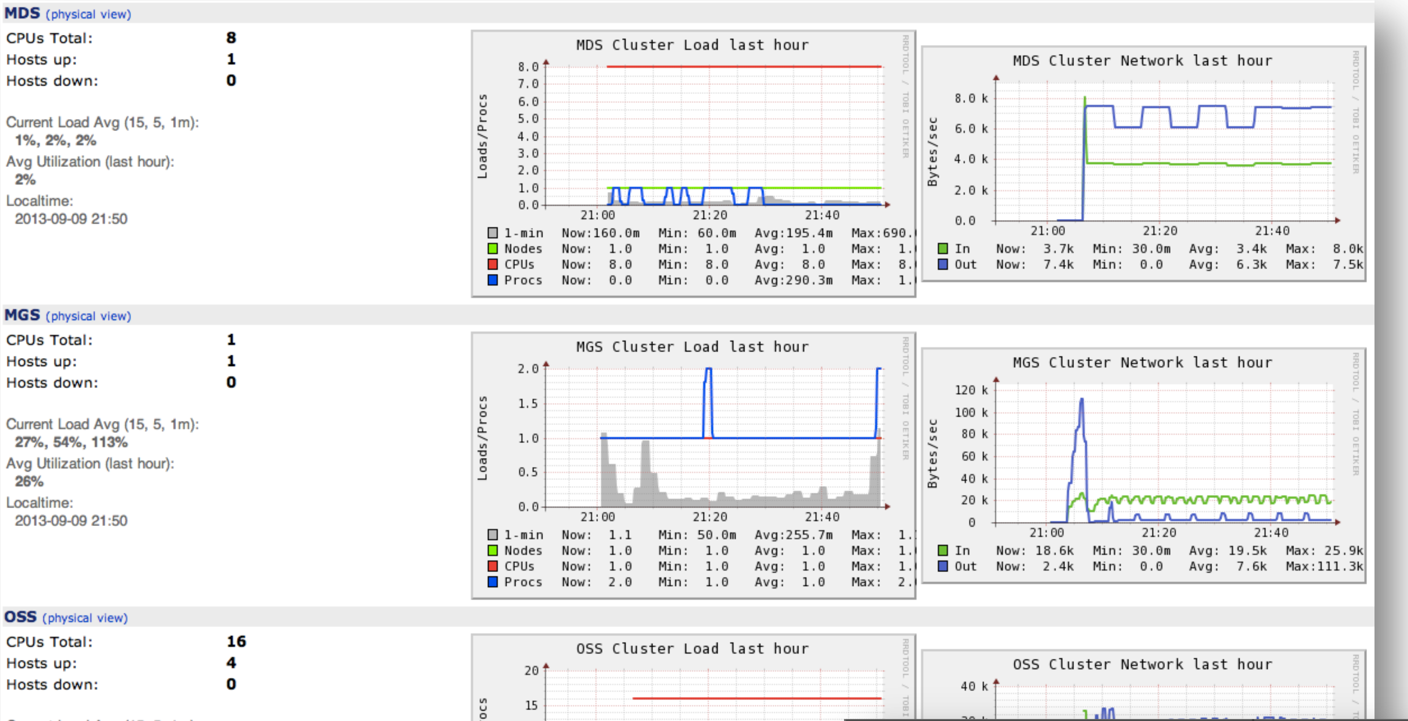
1 to 38 of 38 Instances

□	Name	Instance	AMI ID	Root Device	Type	State	Status Checks	Alarm Status	Monitoring	Security Groups	Key Pair Name	Virtu
□	oss1	i-9e4853aa	ami-2b77e51b	ebs	m1.xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node23	i-a0485394	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node26	i-a7485393	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node21	i-a6485392	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node29	i-a5485391	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node27	i-a2485396	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node31	i-a1485395	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node7	i-954853a1	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node2	i-faba48cd	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	mds0	i-f8ba48cf	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node6	i-f9ba48ce	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node4	i-eaba48dd	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	node0	i-f4ba48c3	ami-2b77e51b	ebs	m3.2xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para
□	oss2	i-f2ba48c5	ami-2b77e51b	ebs	m1.xlarge	running	2/2 checks pa	none	basic	Test-4OSS-Instanc	lustre	para

No EC2 Instances selected.

Select an instance above

Monitors tools are available



MDS (physical view)

CPUs Total:	**8**
Hosts up:	**1**
Hosts down:	**0**

Current Load Avg (15, 5, 1m):
1%, 2%, 2%
Avg Utilization (last hour):
2%
Localtime:
2013-09-09 21:50

```
□ 1-min   Now:160.0m   Min: 60.0m   Avg:195.4m   Max:690.
🟩 Nodes   Now:  1.0    Min:  1.0    Avg:  1.0    Max:  1.
🟥 CPUs    Now:  8.0    Min:  8.0    Avg:  8.0    Max:  8.
🟦 Procs   Now:  0.0    Min:  0.0    Avg:290.3m   Max:  1.
```

```
🟩 In    Now:  3.7k   Min: 30.0m   Avg:  3.4k   Max:  8.0k
🟦 Out   Now:  7.4k   Min:  0.0    Avg:  6.3k   Max:  7.5k
```

MGS (physical view)

CPUs Total:	**1**
Hosts up:	**1**
Hosts down:	**0**

Current Load Avg (15, 5, 1m):
27%, 54%, 113%
Avg Utilization (last hour):
26%
Localtime:
2013-09-09 21:50

```
□ 1-min   Now:  1.1    Min: 50.0m   Avg:255.7m   Max:  1.
🟩 Nodes   Now:  1.0    Min:  1.0    Avg:  1.0    Max:  1.
🟥 CPUs    Now:  1.0    Min:  1.0    Avg:  1.0    Max:  1.
🟦 Procs   Now:  2.0    Min:  1.0    Avg:  1.0    Max:  2.
```

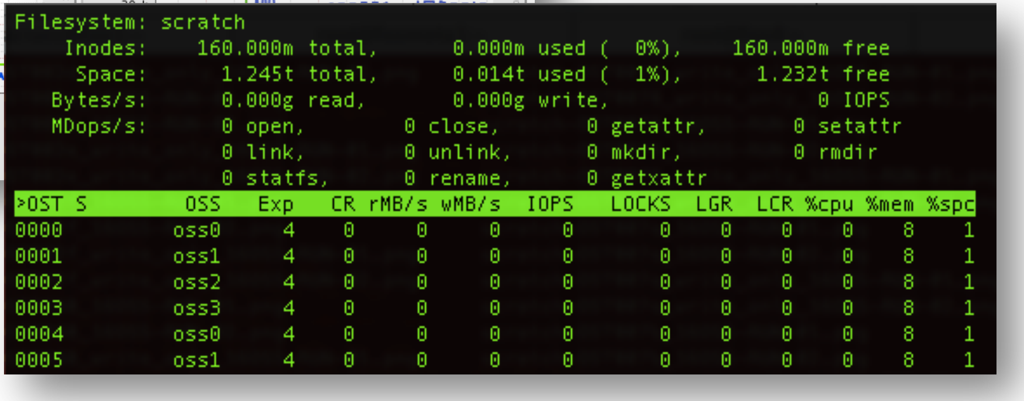
```
🟩 In    Now: 18.6k   Min: 30.0m   Avg: 19.5k   Max: 25.9k
🟦 Out   Now:  2.4k   Min:  0.0    Avg:  7.6k   Max:111.3k
```

OSS (physical view)

CPUs Total:	**16**
Hosts up:	**4**
Hosts down:	**0**

Current Load Avg (15, 5, 1m):
0%, 0%, 0%
Avg Utilization (last hour):
0%
Localtime:
2013-09-09 21:50

```
□ 1-min   Now:  0.0    Min:  0.0    Avg: 83.0m
🟩 Nodes   Now:  4.0    Min:  4.0    Avg:  4.0
🟥 CPUs    Now: 16.0    Min: 15.9    Avg: 16.0
🟦 Procs   Now:  2.0    Min:  0.0    Avg:  1.7
```



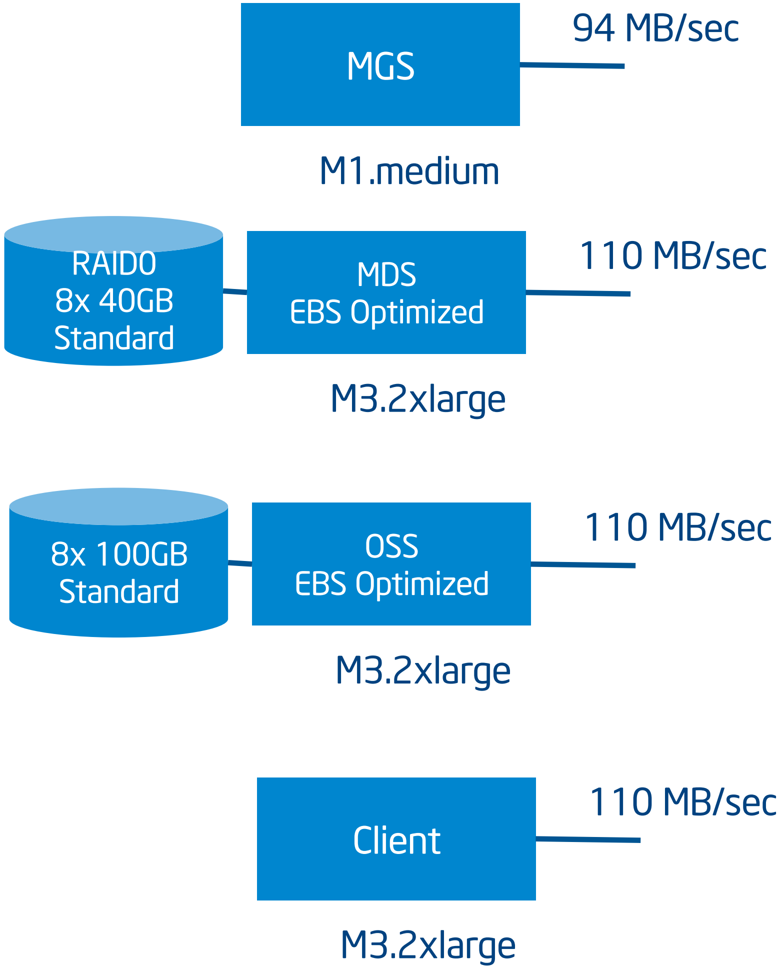
```
Filesystem: scratch
    Inodes:   160.000m total,      0.000m used (  0%),    160.000m free
     Space:     1.245t total,      0.014t used (  1%),      1.232t free
   Bytes/s:     0.000g read,       0.000g write,               0 IOPS
   MDops/s:     0 open,         0 close,        0 getattr,       0 setattr
                0 link,         0 unlink,       0 mkdir,         0 rmdir
                0 statfs,       0 rename,       0 getxattr
>OST S           OSS   Exp   CR rMB/s wMB/s  IOPS   LOCKS  LGR  LCR %cpu %mem %spc
0000            oss0     4    0     0     0     0       0    0    0    0    8    1
0001            oss1     4    0     0     0     0       0    0    0    0    8    1
0002            oss2     4    0     0     0     0       0    0    0    0    8    1
0003            oss3     4    0     0     0     0       0    0    0    0    8    1
0004            oss0     4    0     0     0     0       0    0    0    0    8    1
0005            oss1     4    0     0     0     0       0    0    0    0    8    1
```

Presentation Outline

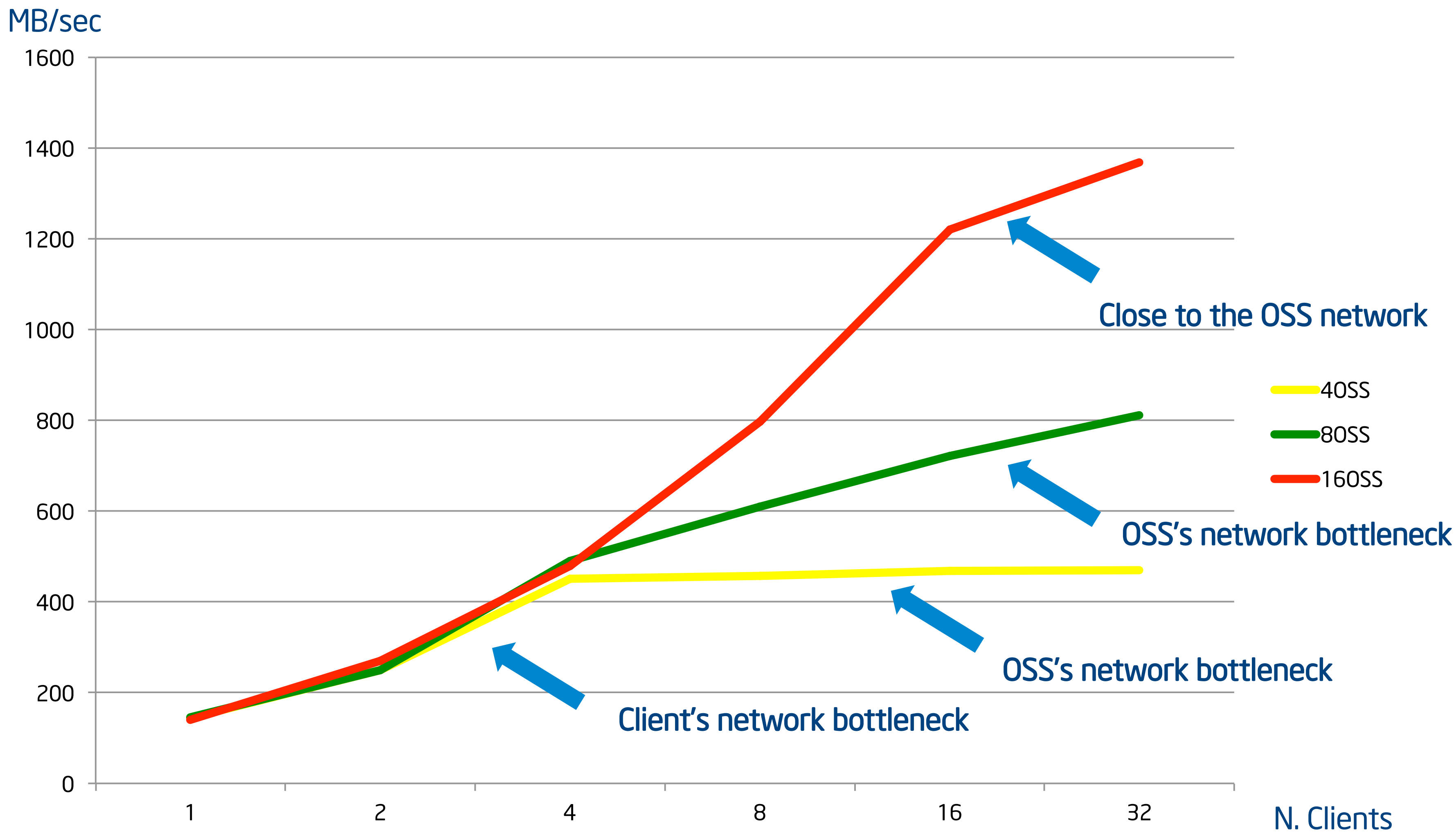
- Motivations
- Deploying Lustre* on AWS
- Benchmarks for different cluster topologies
- Running a real application
- Conclusion and Q&A

Lustre* Benchmark

- Comparing 3 Lustre cluster configuration.
- Increase the number of OSSs
  - 4 OSS
  - 8 OSS
  - 16 OSS
- Configurations of MGS and MDS are the same.
- We use 32 clients.



IOR Sequential Read FPP



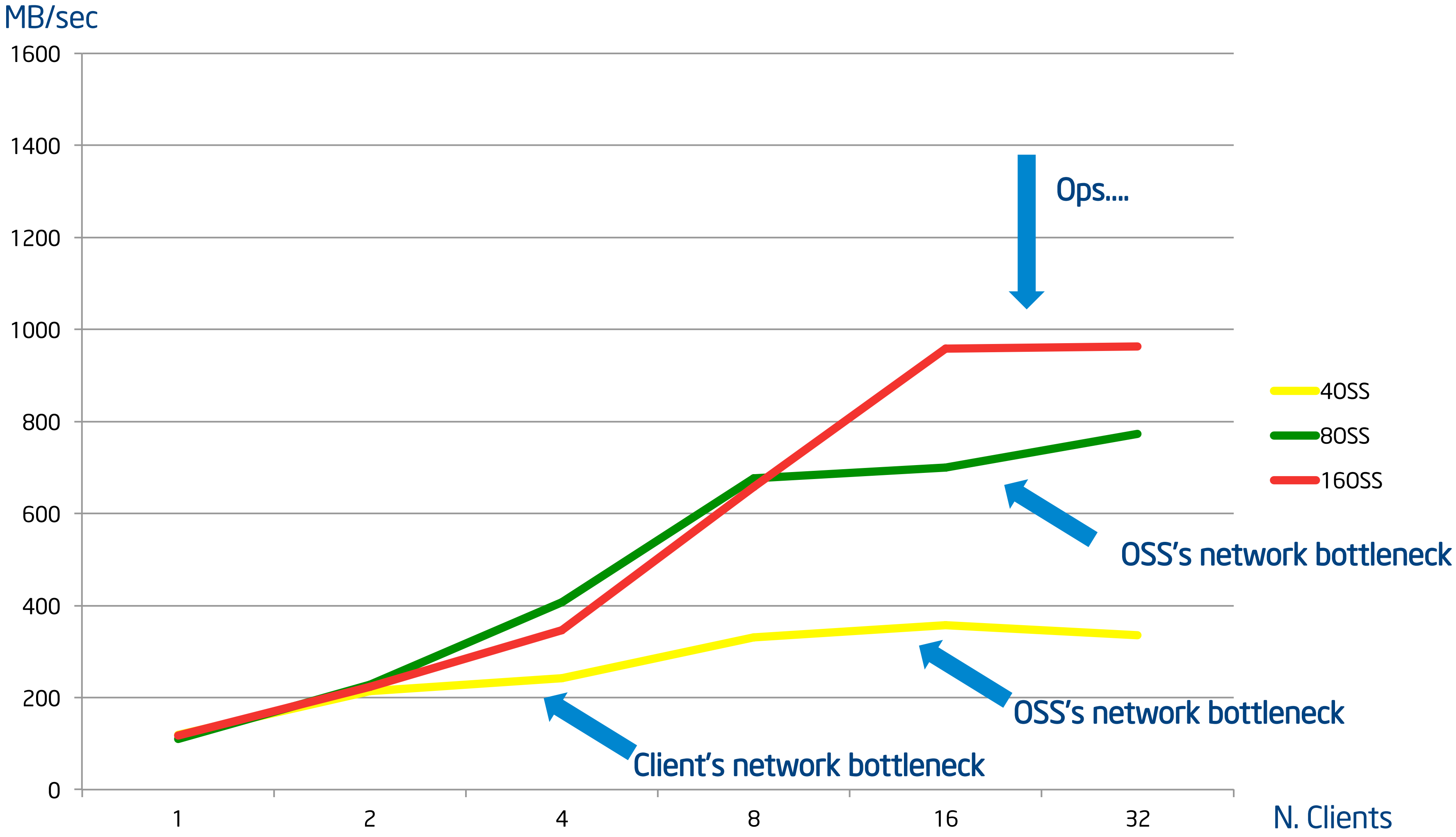
Close to the OSS network

OSS's network bottleneck

OSS's network bottleneck

Client's network bottleneck

IOR Sequential Write FPP



Ops….

OSS's network bottleneck

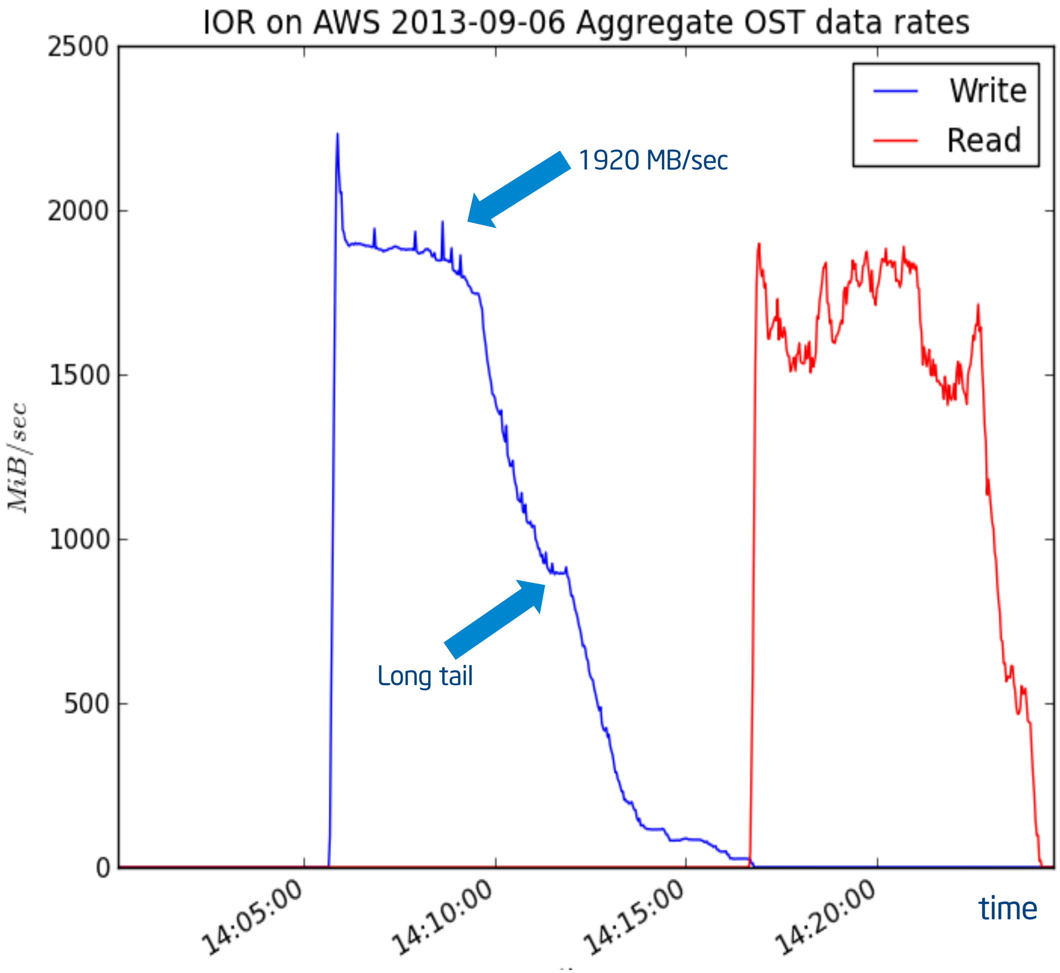
OSS's network bottleneck

Client's network bottleneck

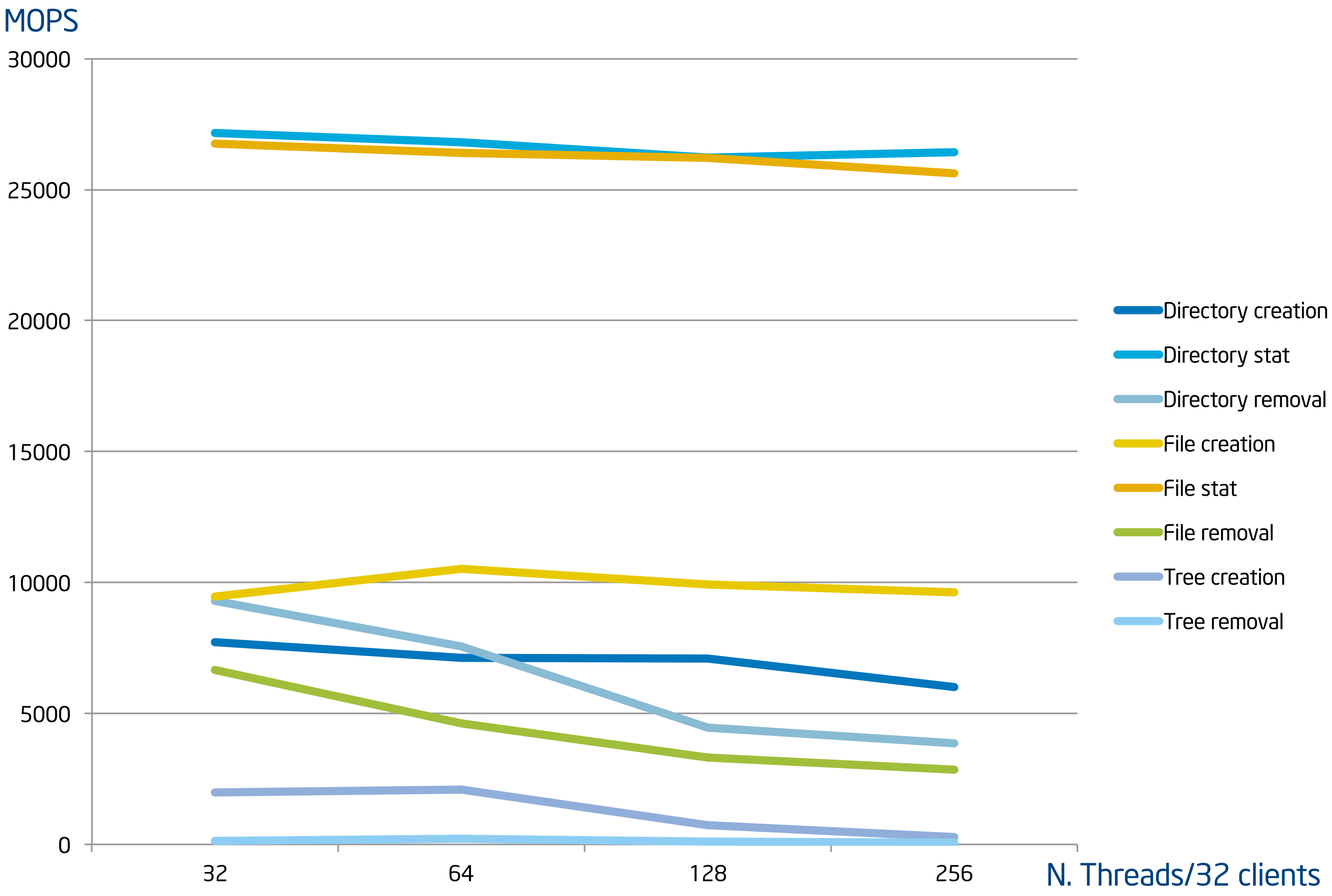
Aggregate Performance During Run

LTOP is available and we use it to record the OSTs activities during the IOR run.

With a simple python script we create this graph: “aggregate performance vs time” to analyze the problem.



MDTEST on 16 OSS Cluster Configuration

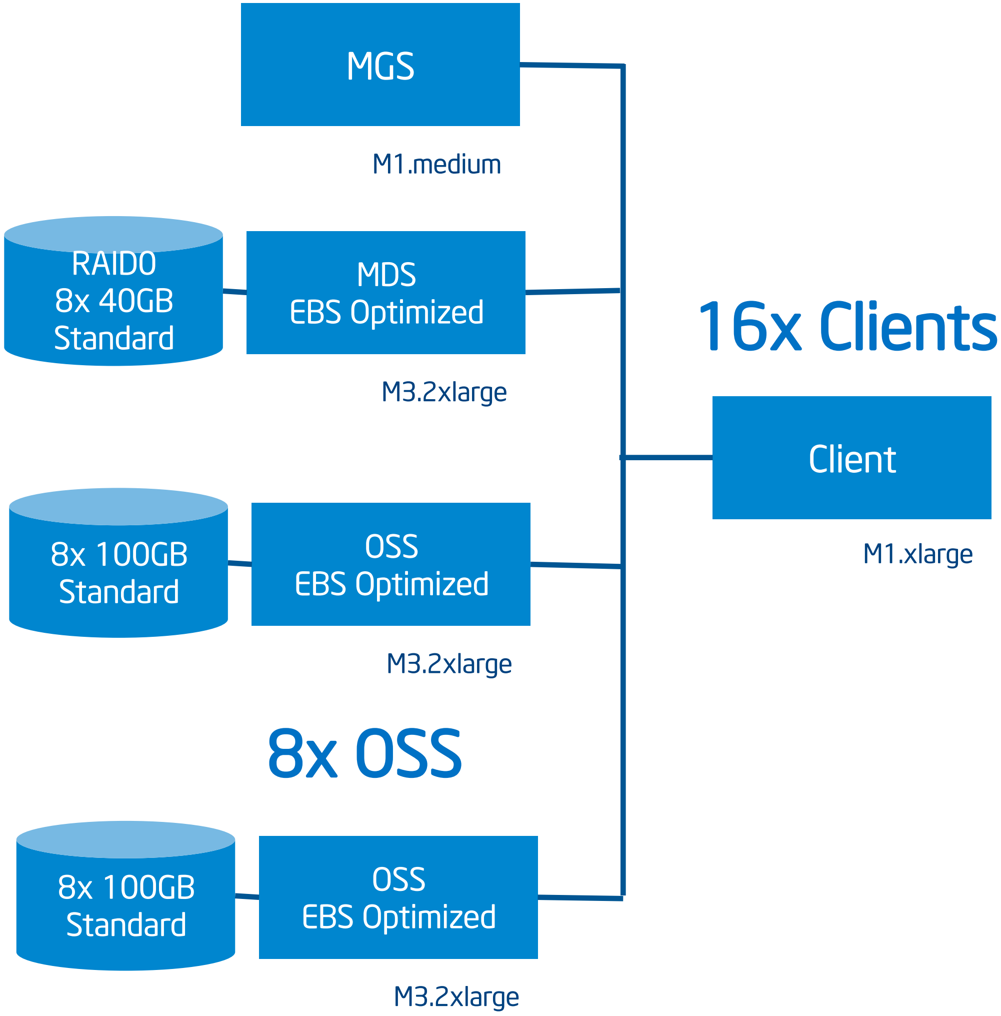


Presentation Outline

- Motivations
- Deploying Lustre* on AWS
- Benchmarks for different cluster topologies
- Running a real application
- Conclusion and Q&A

Lustre* Cluster for MADBench2

In the MADbench2 application the problem is to generate simulations of the cosmic microwave background radiation sky map. Each of those simulations involves a very large matrix inversion that is solved with an out-of-core algorithm (thus the I/O)



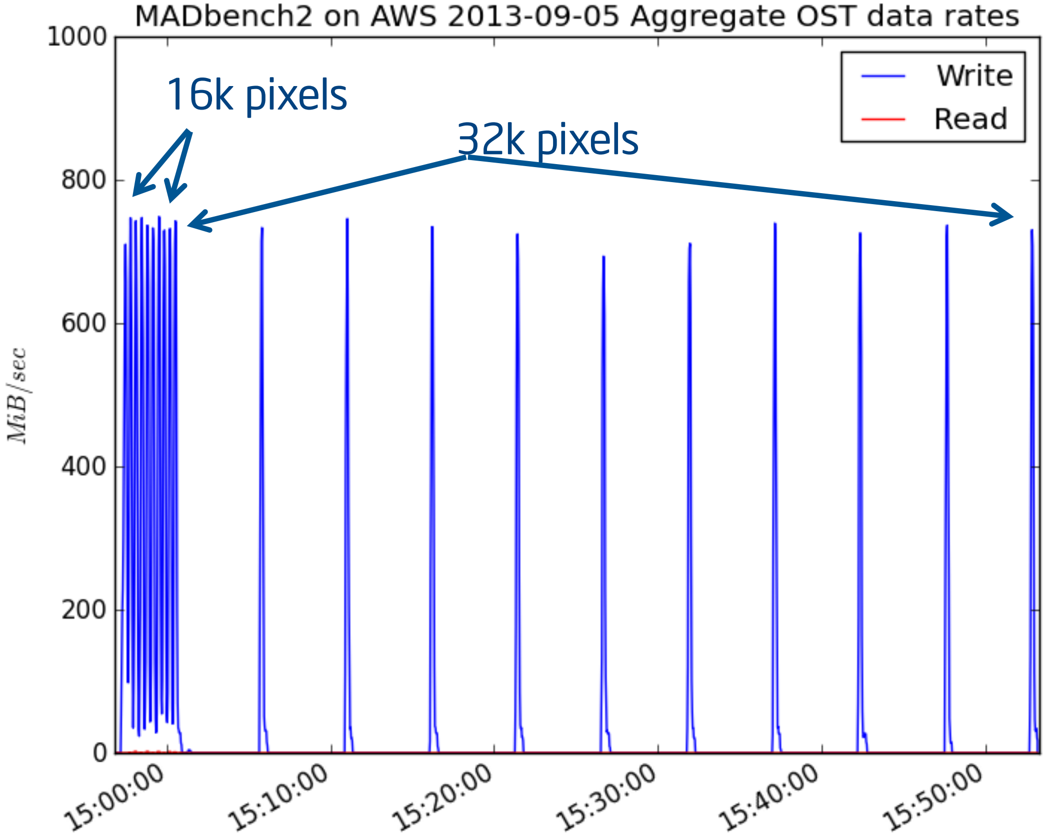
http://crd-legacy.lbl.gov/~borrill/MADbench2/

Running MADbench2

MADbench2 was run at scales of 1k, 2k, 4k, 8k, 16k, and 32k.

The 32k instance ran for about an hour.

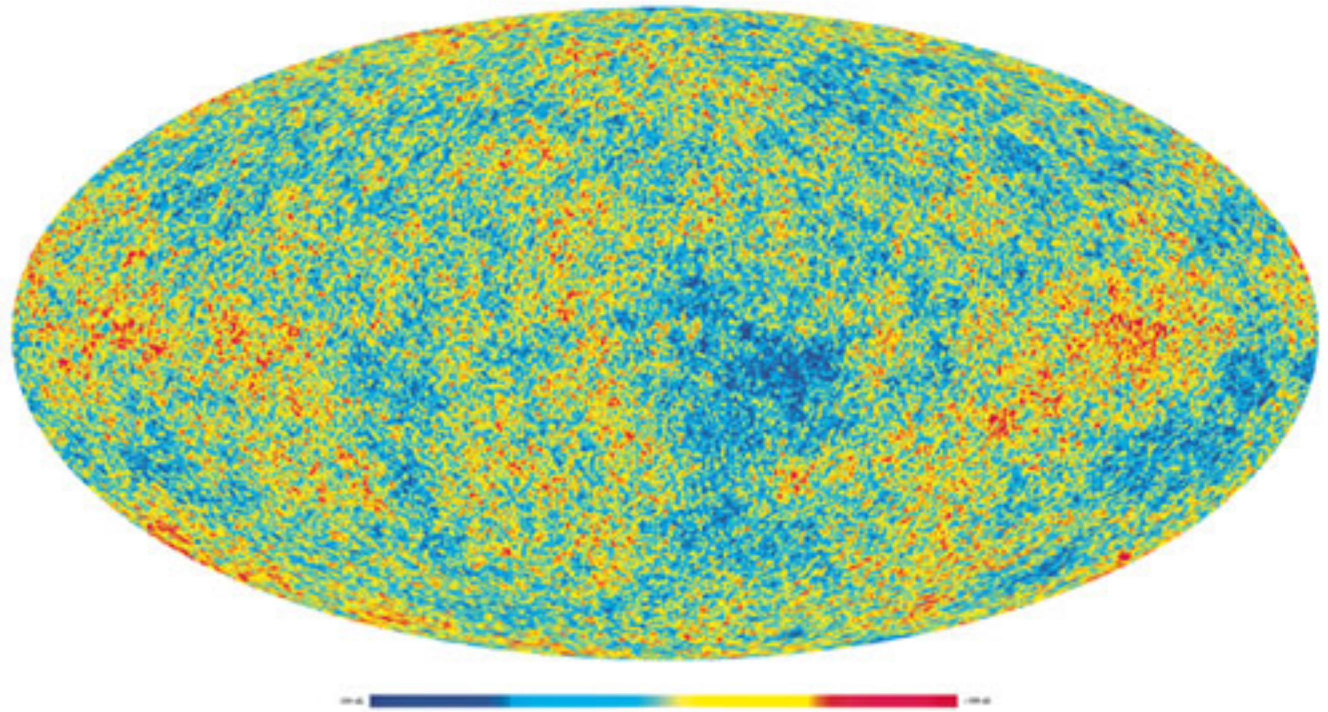
A 64k instance would probably run for about 7 hours, and a 128k instance for about 50 hours.



16k pixels

32k pixels

Cosmic Microwave Background



Presentation Outline

- Motivations
- Deploying Lustre* on AWS
- Benchmarks for different cluster topologies
- Running a real application
- Conclusion and Q&A

Actual Status

- While there is more testing that needs to be done of the different high capacity/bandwidth nodes, the results are that Lustre* runs fairly well in the cloud.
- The provisioning system is good and permits creation of a complete Lustre cluster in less then 10 minutes.
- The created Lustre file system is then complete, up and running, MPI libraries are configured, and monitoring tools like LMT/LTOP/GANGLIA are usable.
- We are alpha testing with some initial users, and a adding some features.
- We plan to make our AMIs available on AWS Marketplace soon.

Conclusion

$ 532.35

more than 1GB/sec, peak 25K MOPS, 12.TB available space, 32 clients for 3 weeks intel®