Object Storage Basics and Performance Testing

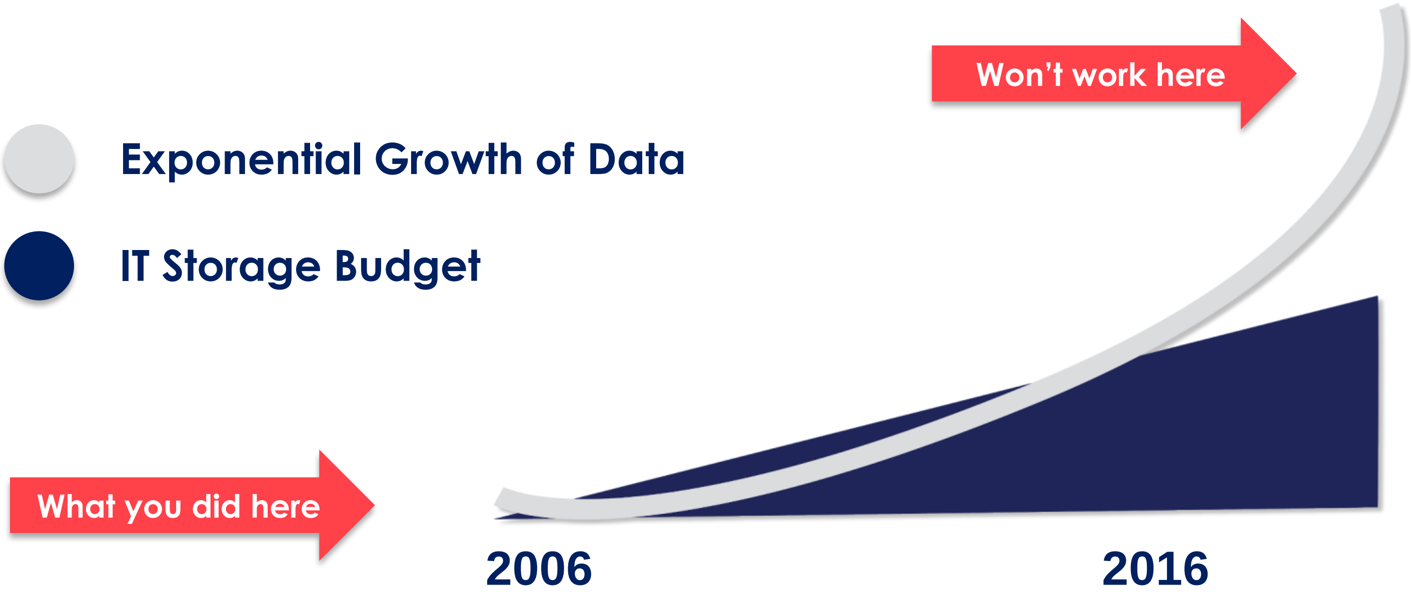
Colette Downey

Caringo Performance Engineer

linkedin.com/in/colette-downey

caringo

Storage Paradigm Shift

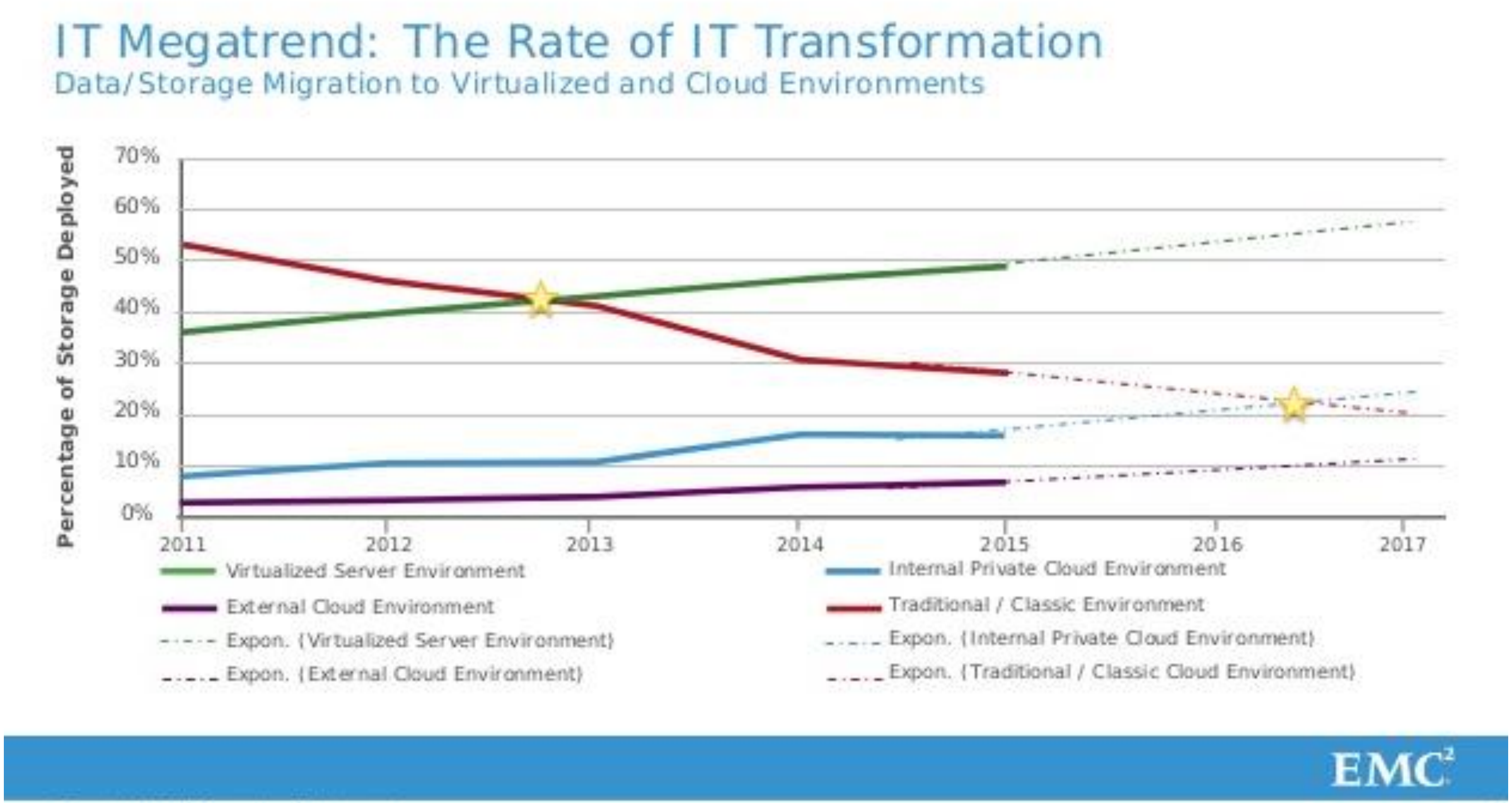


○ Exponential Growth of Data

● IT Storage Budget

Won’t work here

What you did here



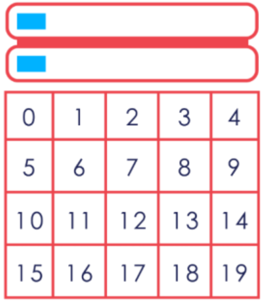
IT Megatrend: The Rate of IT Transformation

Data/Storage Migration to Virtualized and Cloud Environments

Source: EMC trend report: Managing Information Storage – Trends, Challenges and Options, 2013 – 2014

Basics of Object Storage

Block



Client Via OS

Fixed Sys Attributes

Transactional Data

Performance

File

0 1 2 3 4
5 6 7 8 9
10 11 12 13 14
15 16 17 18 19

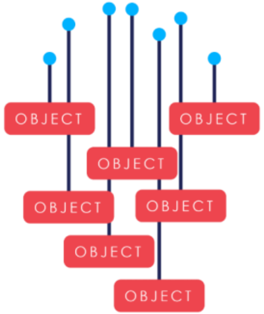
Client Via OS

Fixed FS Attributes

Shared Changing File

Access, Single Site

Object



Client is App

Custom Metadata

Shared Semi-static File

Scalable, Multi-Site

Comparing Current Storage Technology

	SAN	NAS	Object Storage	Tape
Architecture	Block	File	Object	Block
Economics (TCO and size)	$$$$, < PB structured data	$$$, 1–20 PBs structured or unstructured data	$$, EBs unstructured data	$, EBs unstructured data
Performance	High performance block in milliseconds	High performance file in milliseconds	Moderate performance in milliseconds to seconds More efficient for complex data type representations	Low performance in seconds to minutes
Accessibility	Via OS Centralized access control	Via OS Centralized access	Via API Call De-centralized access control	Via OS
Search	Block attributes No higher level metadata	File Attributes Limited higher level metadata	Object attributes and unrestricted customizable metadata	Segment attributes
Cloud Enabled	Need other components	Need other components	Yes	Need other components
Durability	RAID/Replication Long RAID rebuild times	RAID/Replication Long RAID rebuild times	EC/Replication/Content level protection Efficient regeneration	Mobile/Relocation

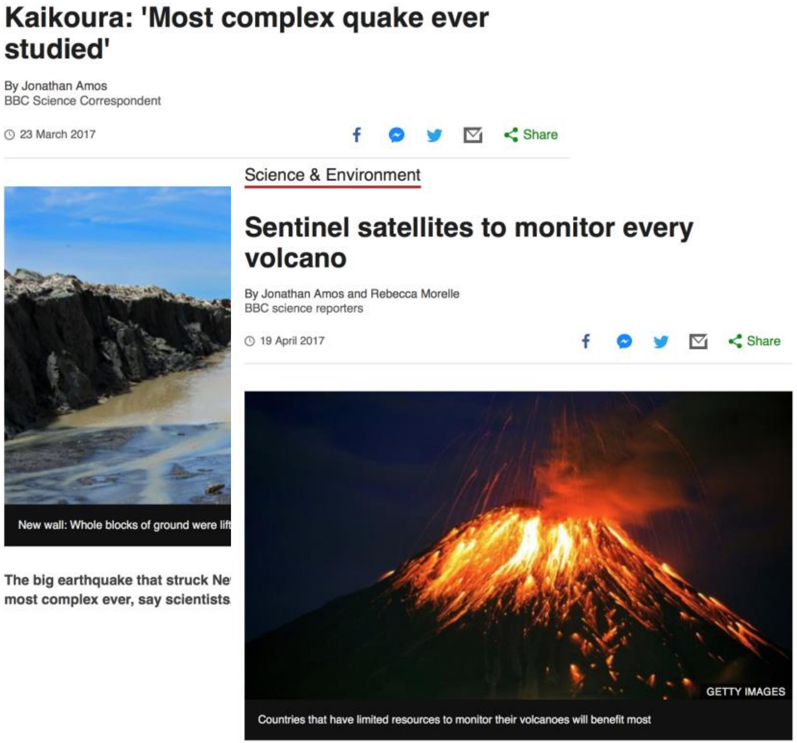
Use Case Mapping

	SAN	NAS	Object	Tape
High Transaction	●	●	○	○
Editing & Collaboration	●	●	○	○
Distribution & Delivery	○	◒	●	○
Active Archive	○	◒	●	○
Cold Archive	○	○	◒	●

● **High** ◒ **Medium** ○ **Low**

STFC JASMIN Super Data Cluster

- Goal: Store it all!
  - Geospatial data: ~10TB/day,12 satellites
  - Datasets of 150TB
- Provides:
  - More bandwidth to disk than normal clusters
  - Storage (disk and tape) and
  - Computing (batch, hosted and cloud)
- Supports:
  - 1,700 Researchers, 900 PhDs
  - 50+ universities in UK, Europe, Japan and US
  - 3,500+ users, 2,000 experiments
  - 900+ publications



Kaikoura: 'Most complex quake ever studied'

By Jonathan Amos
BBC Science Correspondent

23 March 2017

Share

New wall: Whole blocks of ground were lift

The big earthquake that struck Ne
most complex ever, say scientists

Science & Environment

Sentinel satellites to monitor every volcano

By Jonathan Amos and Rebecca Morelle
BBC science reporters

19 April 2017

Share

GETTY IMAGES

Countries that have limited resources to monitor their volcanoes will benefit most

Increasing capacity needs + growing research base =
<u>necessary</u> shift in access from traditional POSIX protocols to RESTful object interfaces.

Testing Environment

Elimination of network and storage bottlenecks => high performance

Clients (20)

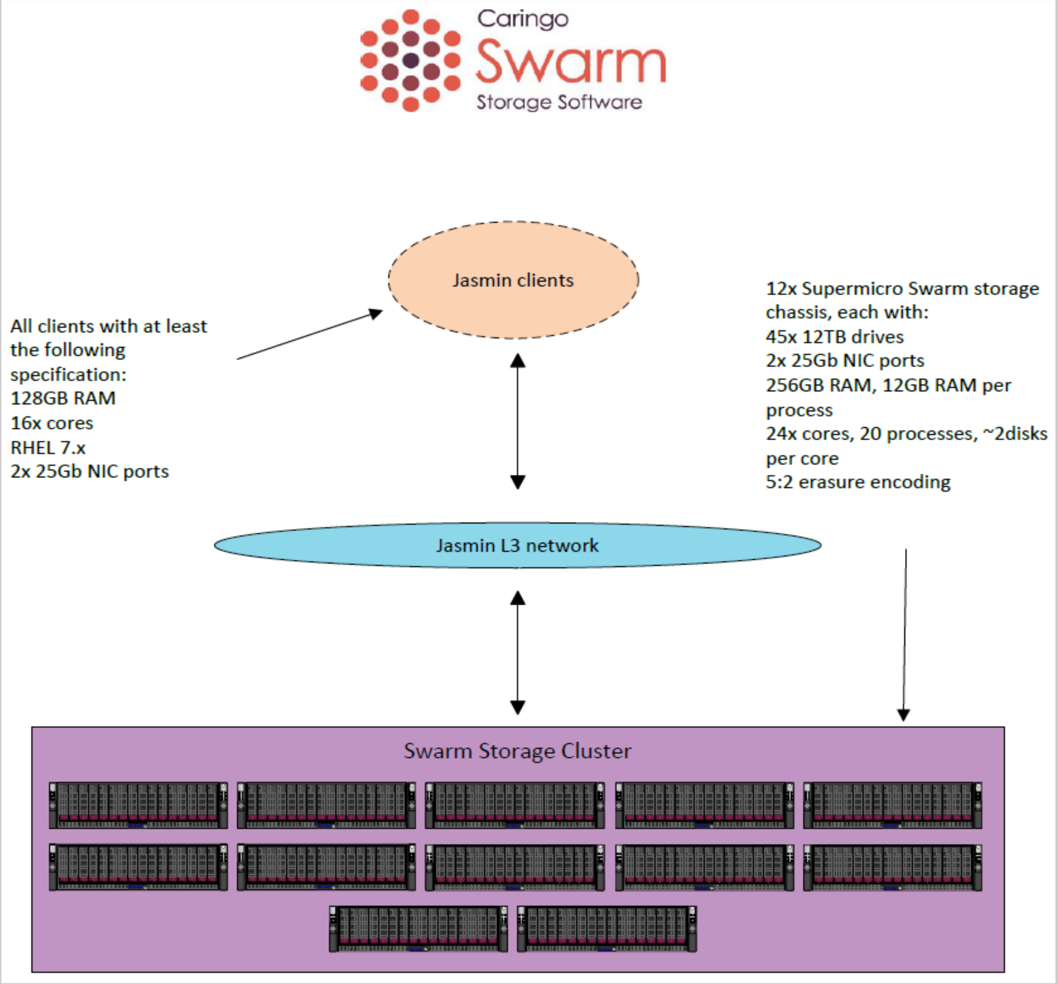
- S3 API testing, 2GB objects
- NFSv4 testing, 10GB objects

Super-low latency, non-blocking network

- Leaf-spine network; all switches 100Gb
- 3 network switch hops between any network endpoint

Swarm cluster (12)

- Highly parallelized
- Distributed shared index in RAM
- Automatic internal load balancing



Caringo Swarm Storage Software

Jasmin clients

All clients with at least the following specification:
128GB RAM
16x cores
RHEL 7.x
2x 25Gb NIC ports

12x Supermicro Swarm storage chassis, each with:
45x 12TB drives
2x 25Gb NIC ports
256GB RAM, 12GB RAM per process
24x cores, 20 processes, ~2disks per core
5:2 erasure encoding

Jasmin L3 network

Swarm Storage Cluster

Testing Methods: Object Performance

COSBENCH

- Open source, distributed load testing framework
- Web-based, real-time performance monitoring
- Flexible workload definition
- S3 API

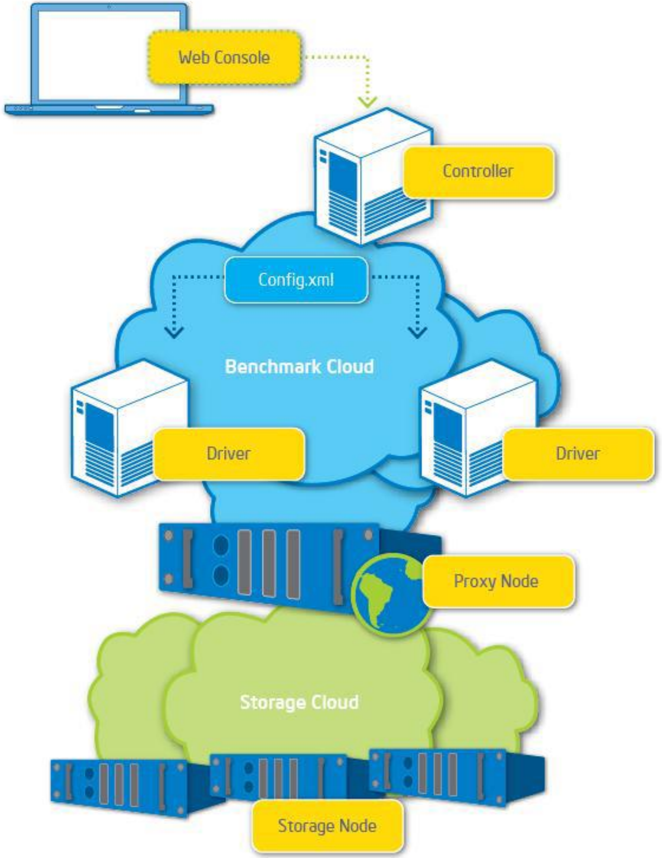
But….

- Assumes all endpoints are behind a load balancer
- Single http error will stop an entire workload
- Throughput reported includes ramp up /down
- Missing newer S3 API features:

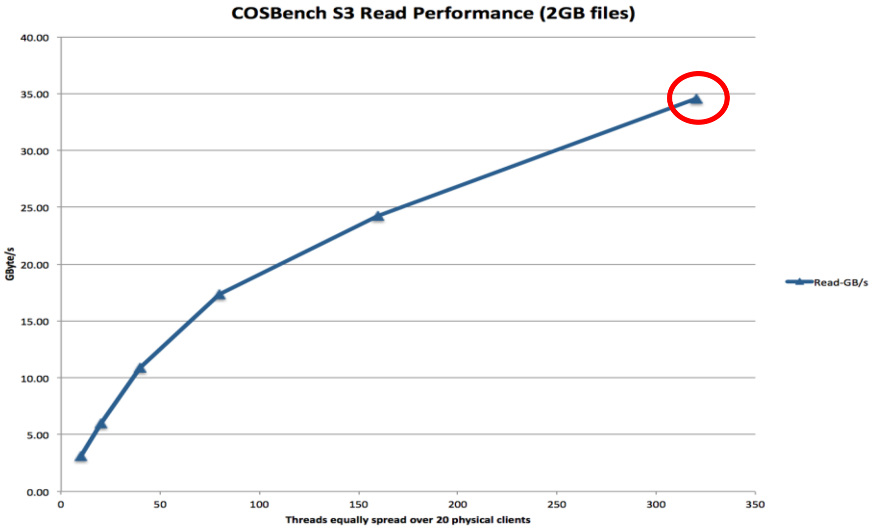
    latest authentication methods

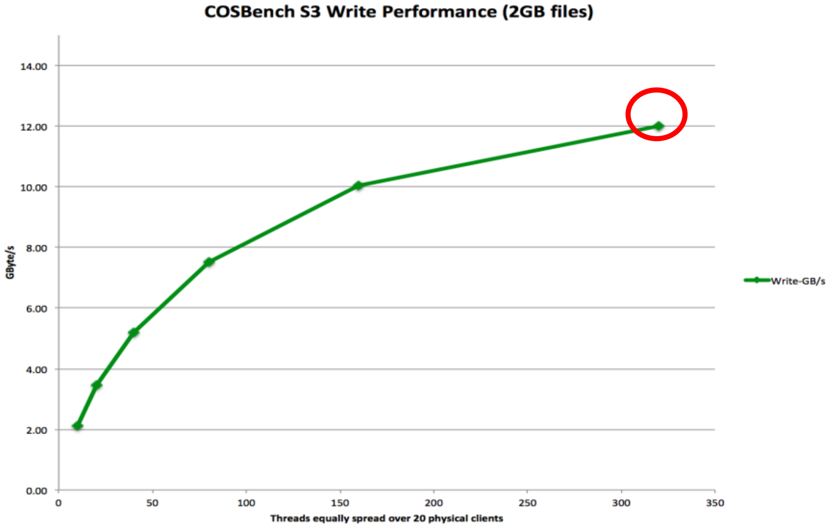
    parallel multipart write

    parallel range reads



Performance Scaling of Object Reads and Writes





	Performance Requirements	Caringo Results	% Above Goal
Object Read	**21.5 GB/s**	35 GB/s	63%
Object Write	**6.5 GBs**	12.5 GB/s	92%

Testing Methods: NFS Performance

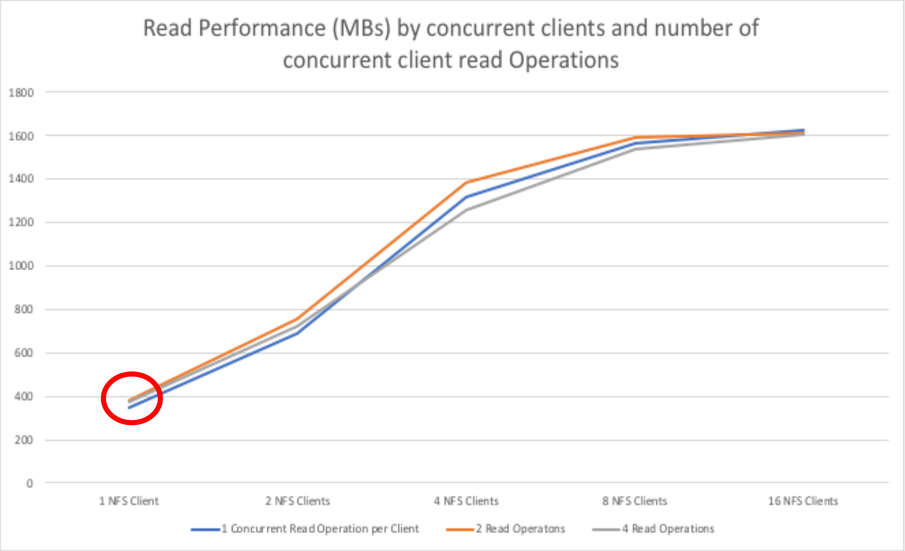
NFSv4 Single Instance Throughput

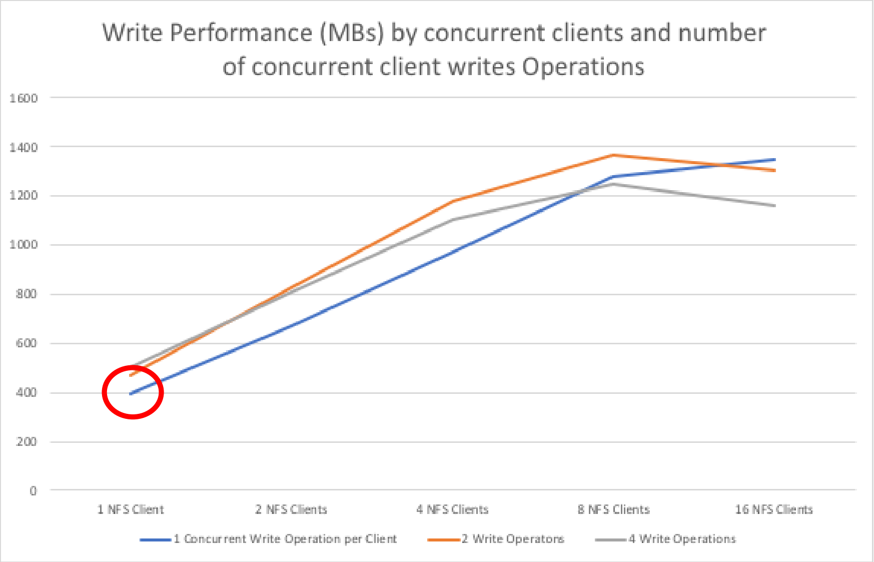
- One SwarmNFS server
- Up to 16 physical clients
- Very large files (10GB)

Custom scripts

- File cp for writes, 10GB files in client's ramdisk
- File dd for reads
- SwarmNFS restarted and kernel caches cleared between tests

NFSv4 Single Server Performance Scaling





	Performance Requirements	Caringo Results	% Above Goal
NFS Read	**150 MB/s**	349 MB/s	132%
NFS Write	**110 MBs**	392 MB/s	256%

STFC Benchmarks

Object AggregateThroughput

	Performance Requirements	Caringo Results	% Above Goal
Object Read	**21.5 GB/s**	35 GB/s	63%
Object Write	**6.5 GBs**	12.5 GB/s	92%

NFSv4 Single Instance Throughput

	Performance Requirements	Caringo Results	% Above Goal
NFS Read	**150 MB/s**	349 MB/s	132%
NFS Write	**110 MBs**	392 MB/s	256%

For More Information

STFC Scientific Computing Department Deploys Swarm

https://www.caringo.com/resources/stfc-benchmarking

Join us February 26, 1pm CT

TechTuesday Webinar: Using Metadata with Object Storage

www.caringo.com/webinars/

Colette Downey

linkedin.com/in/colette-downey

www.caringo.com