Enabling Faster and More Cost-Effective Data Management Using flexFS™ with the Allotrope Data Model

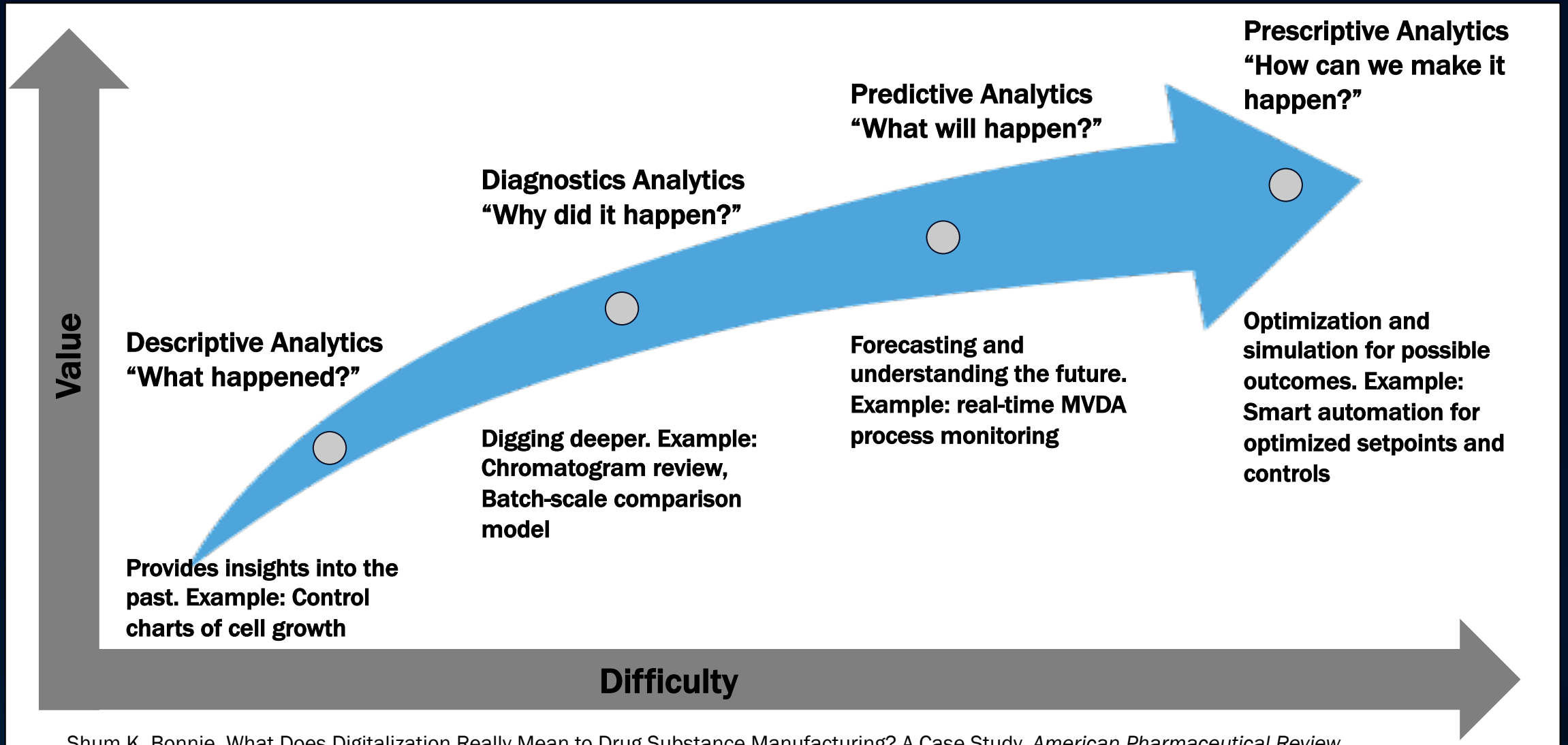
Sep 15, 2021

Agenda

- Digitization in Drug Substance Manufacturing
  - Prescriptive analytics
  - Workflows to optimize testing
- Overview of flexFS
- Use Cases with the Allotrope Data Format
  - Read/Write performance with LC-MS data
  - QC workflow at a single site
  - QC workflow at multiple sites with a Single Point of Truth (SPOT) data server
- Data storage and querying challenges
- Conclusions

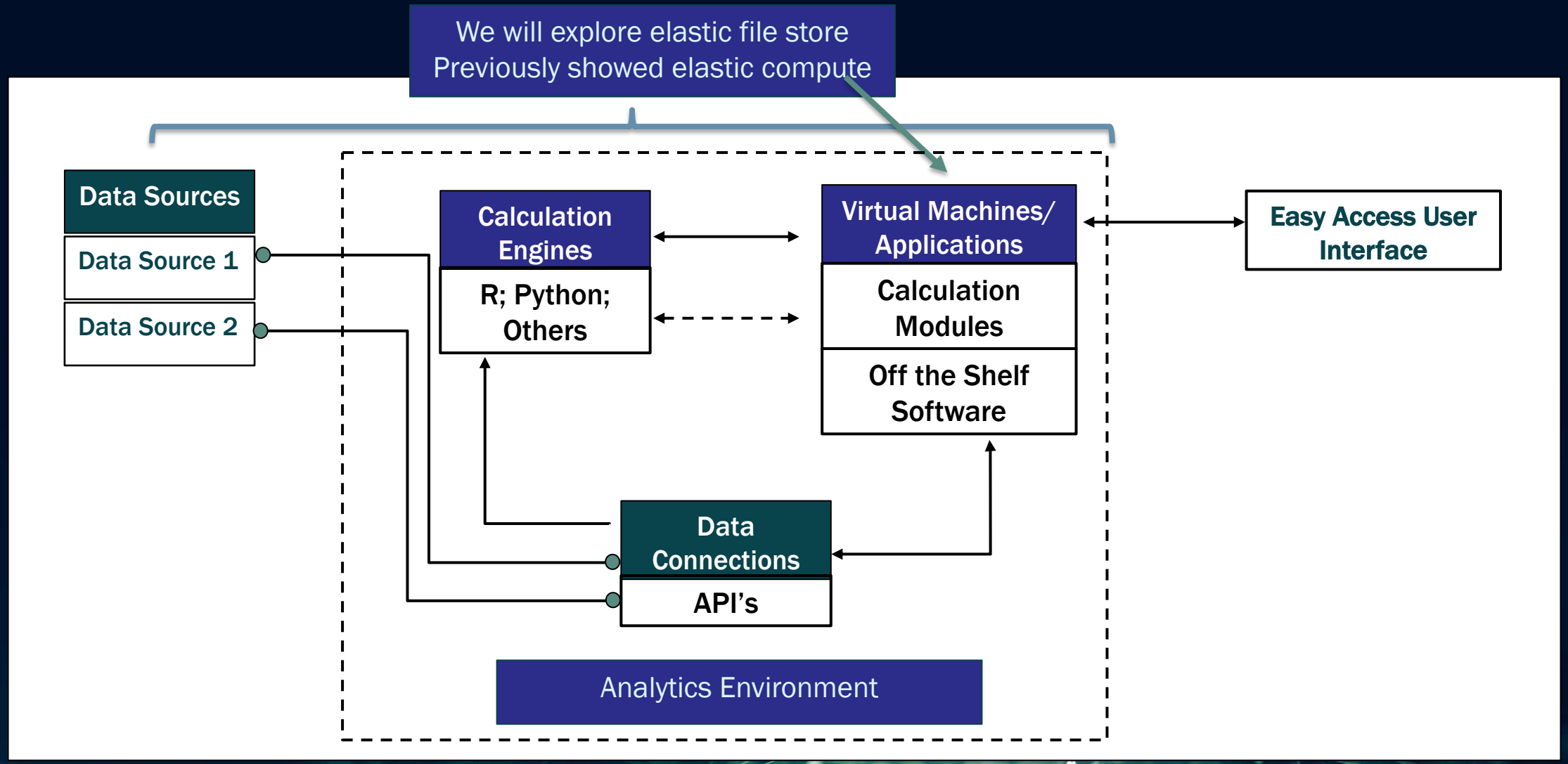
Goal is Prescriptive Analytics

From Bonnie Shum (Genentech), "What does digitalization really mean to drug substance manufacturing" Am. Ph. Rev. Aug 2020



Shum K. Bonnie. What Does Digitalization Really Mean to Drug Substance Manufacturing? A Case Study. *American Pharmaceutical Review*. August 11, 2020.

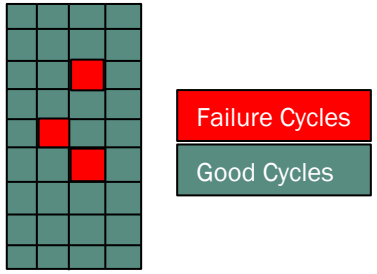
Shum outlines the digital environment



Shum suggests workflows to optimize testing

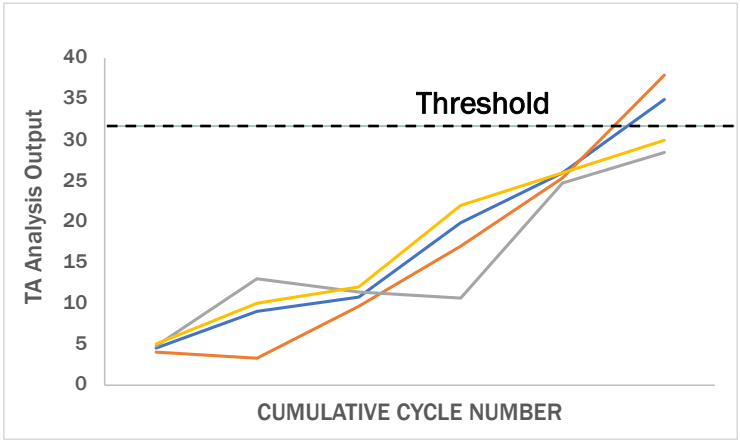
Require scalable compute and easy access to data

1 Data Collection



2 Generate Overlays

3 Determine Optimal Threshold



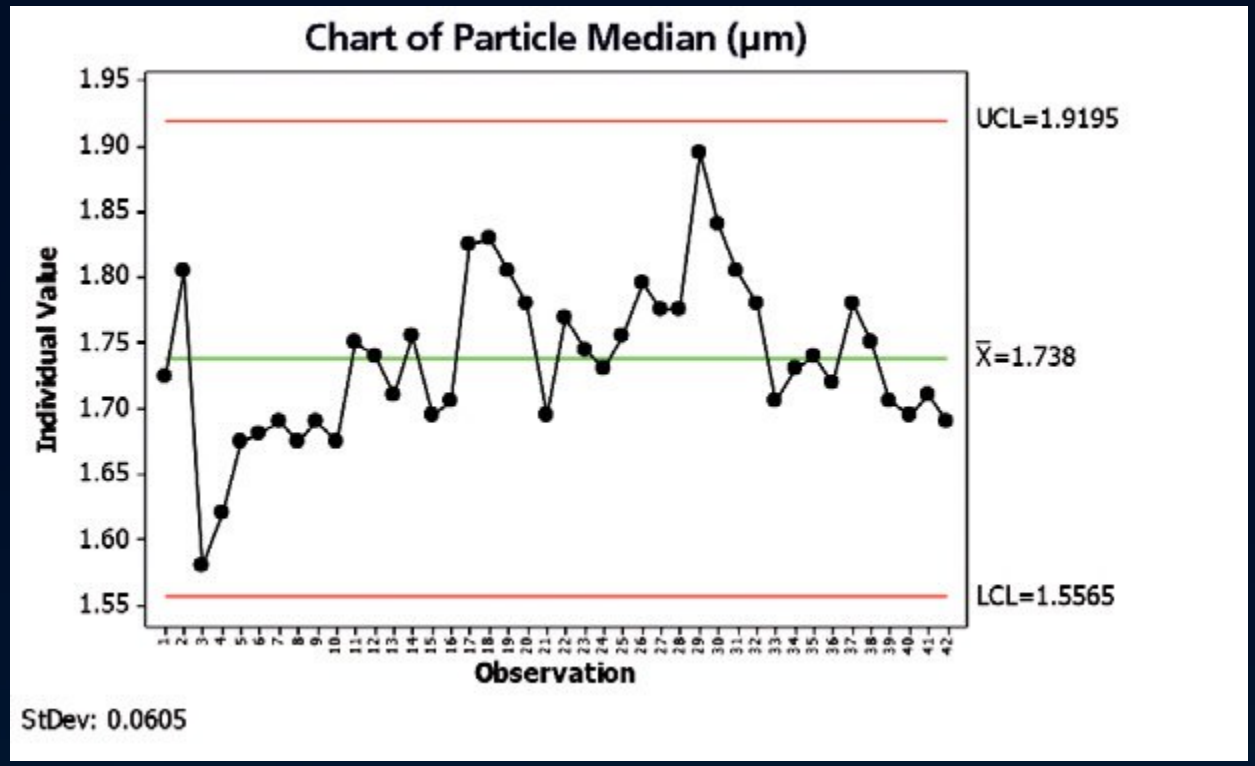
4 Evaluate Accuracy

		Predicted Class	
		N	Y
Observed Class	N	TN	FP
	Y	FN	TP

5 Compare Methods

Method	Accuracy (%)	Prediction (%)	Failures Detected (%)	Evaluation
A	95	35	90	Excellent
B	91	25	80	Good
C	94	30	60	Good
D	84	30	50	Poor

With optimal testing – acceptance testing can be established



$$C_{pk} = \min\left[\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right]$$

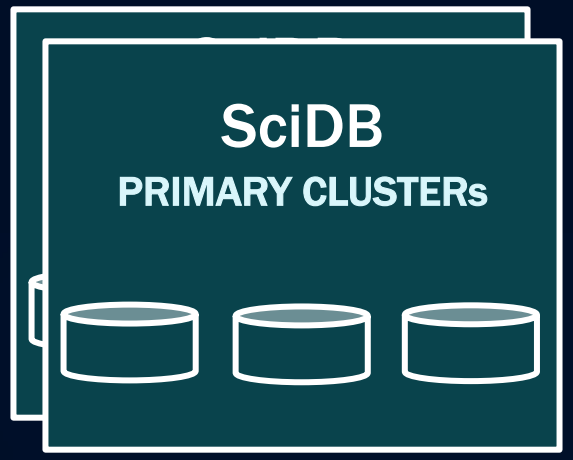
1. "min" notation means to take the lesser of the two values. This reduces over-estimation
2. Assumes a normal distribution (not true).

Slide from Brent Donovan, circa 2012.

Levey, Stanley; Jennings, E. R. (November 1, 1950). "The use of Control Charts in the Clinical Laboratory". American Journal of Clinical Pathology. 20 (11): 1059–1066. doi:10.1093/ajcp/20.11_ts.1059. PMID 14783086

Challenges to make scalable compute cost effective

Data Analysis & Querying Challenges



Interactive querying

- Metadata
- Annotations
- UV-LC
- LC-MS
- Results data
- Transactional (ACID)

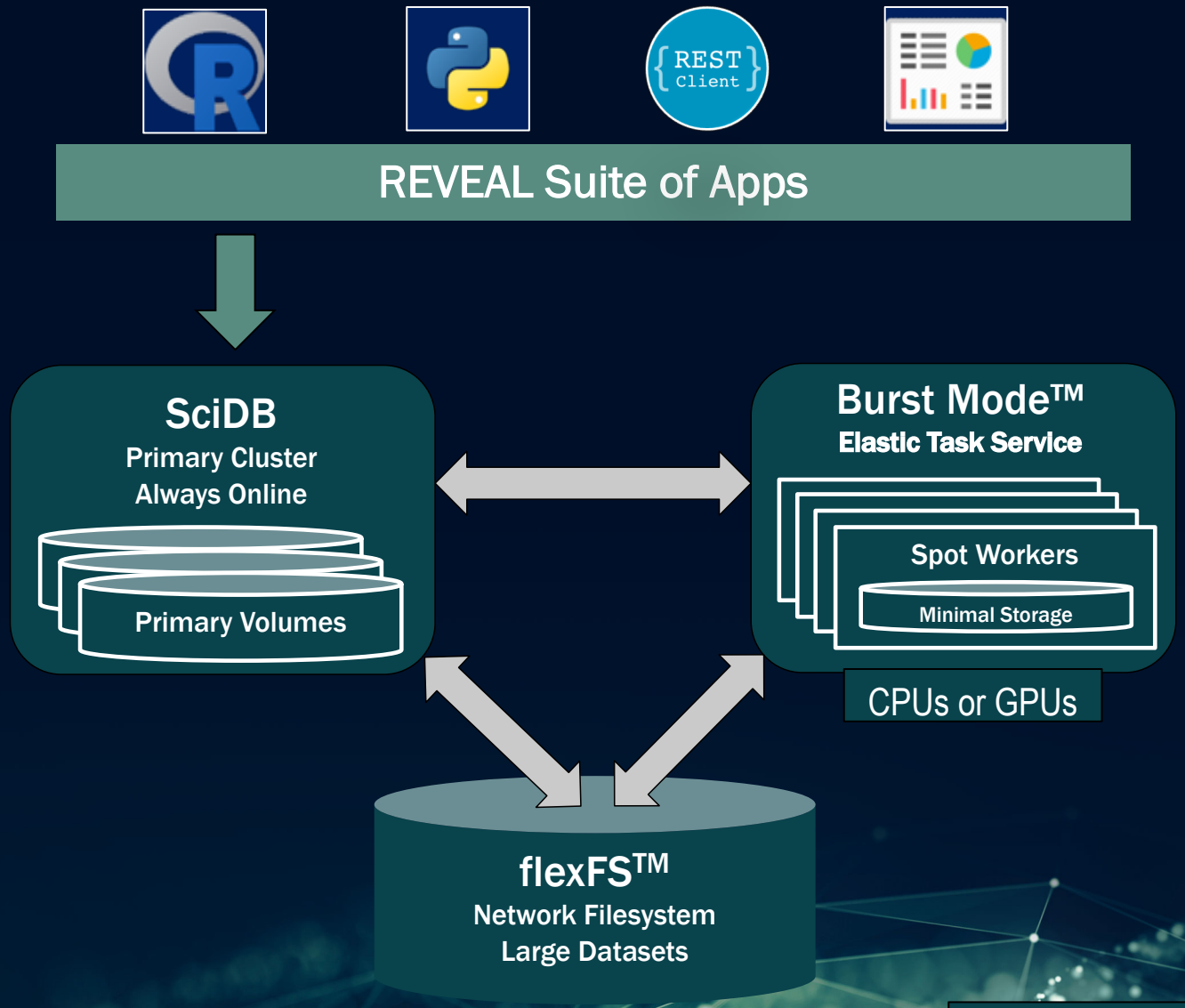
Analytics built in – R and Py APIs

Challenges to meet at minimal cost:

- How do you scale for machine learning?
  - Goal is in the range of millions to billions of calculations
  - Algorithms needed to be scalable
- Storage of different files – redundancy of data in different file formats (solved with ADF, thank you)
- Analysis results needed to be readily available and easily queryable
  - For multiple users
  - From multiple sites
- Data governance –
  - ACID, extended ACLs (beyond POSIX)
  - Encrypted

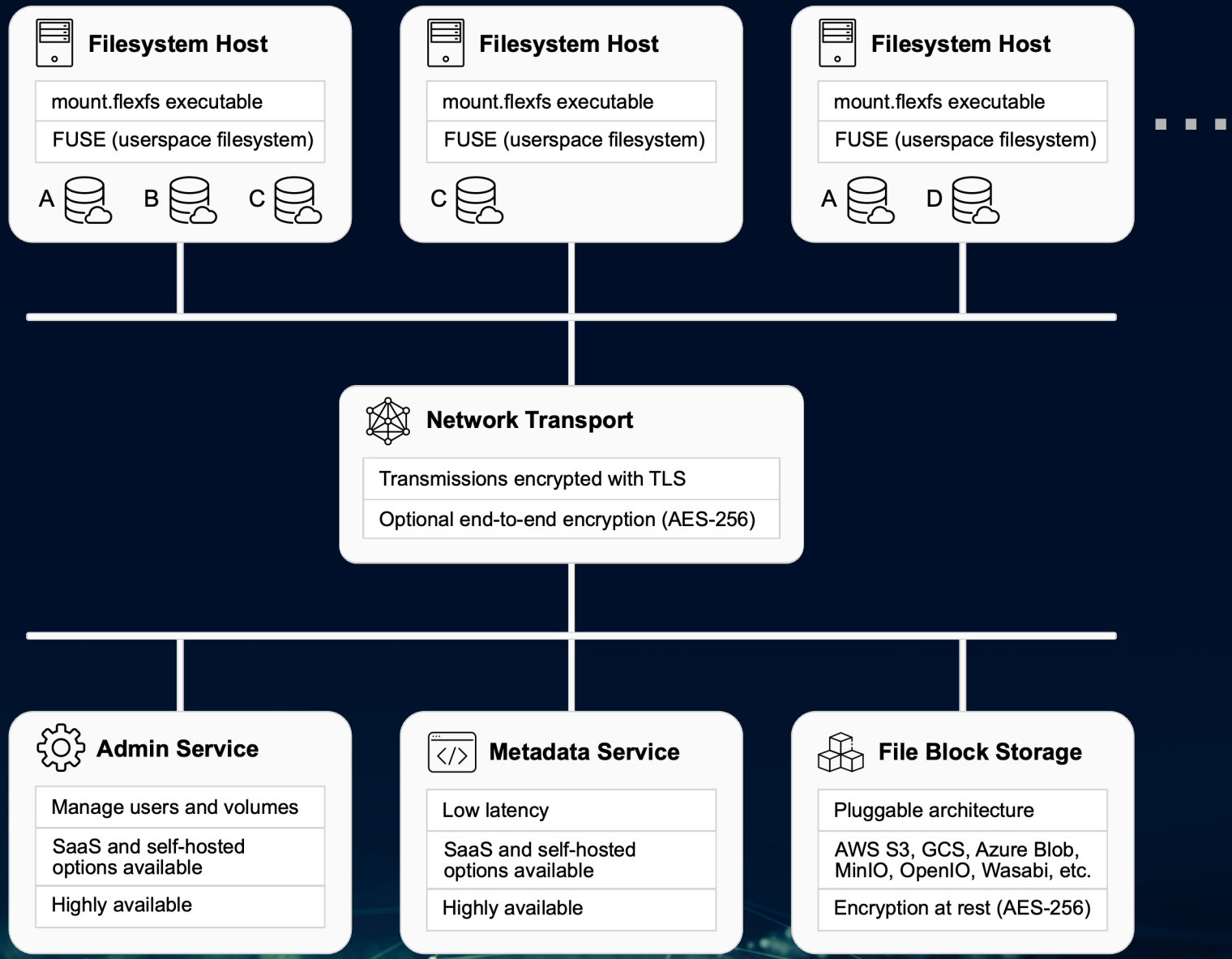
REVEAL™ Architecture

in the cloud: no code R API, combined with array native indexing, elastic compute and elastic file storage



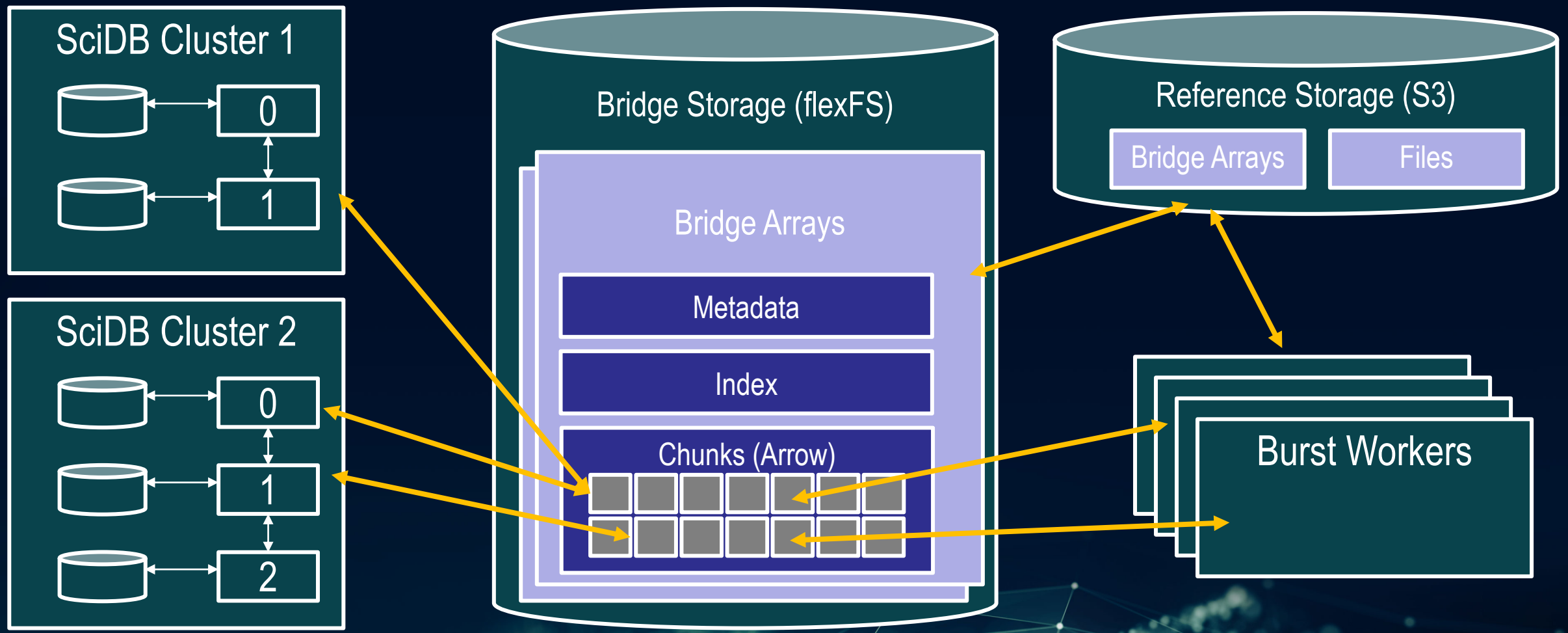
Works with cloud management services

Architecture

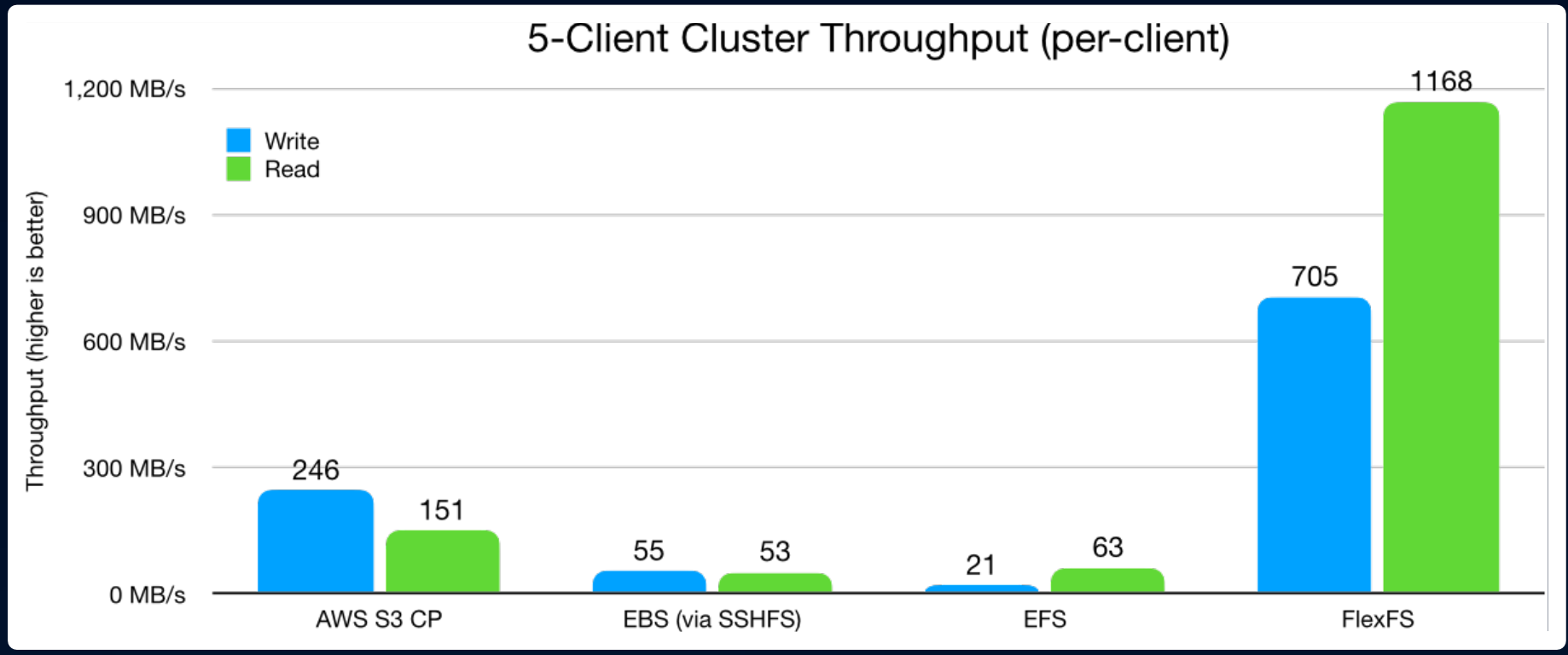


SciDB+ flexFS + SciDB Bridge

High-scalability shared arrays accessible to multiple SciDB clusters and burst workers



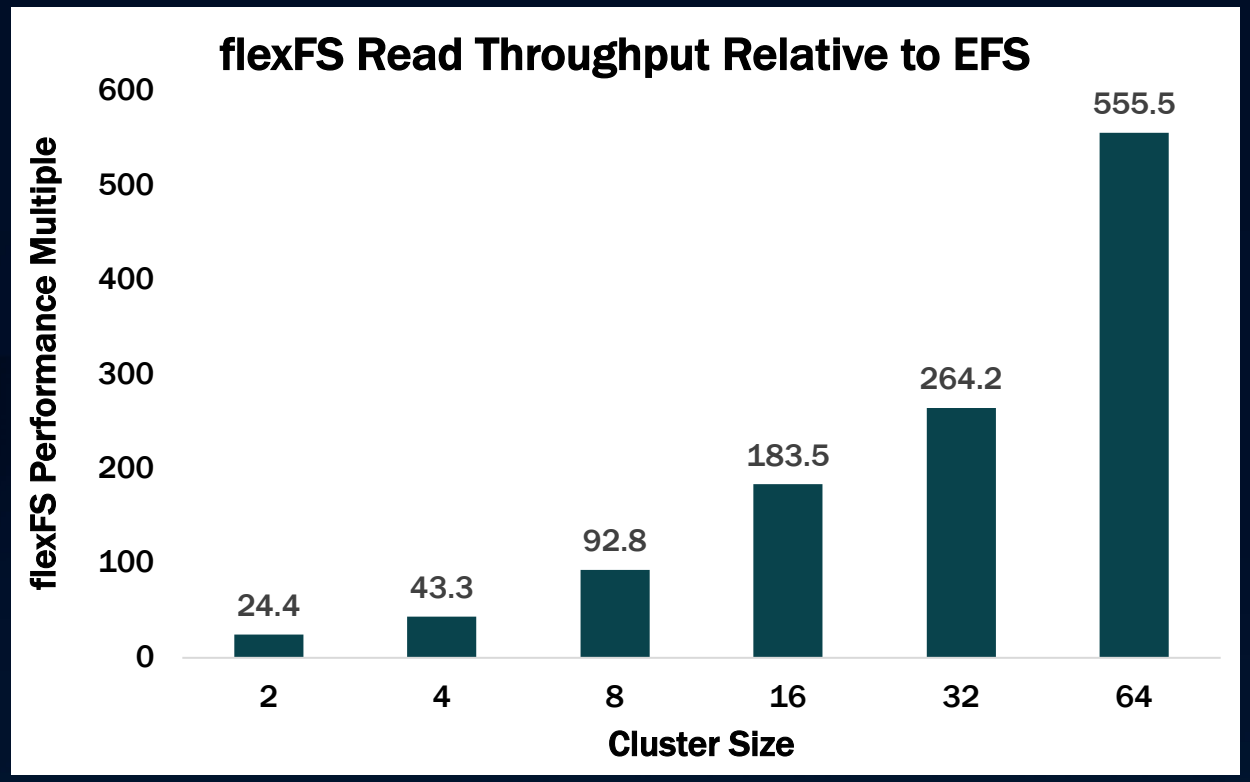
5-Client Cluster Throughput (per client)

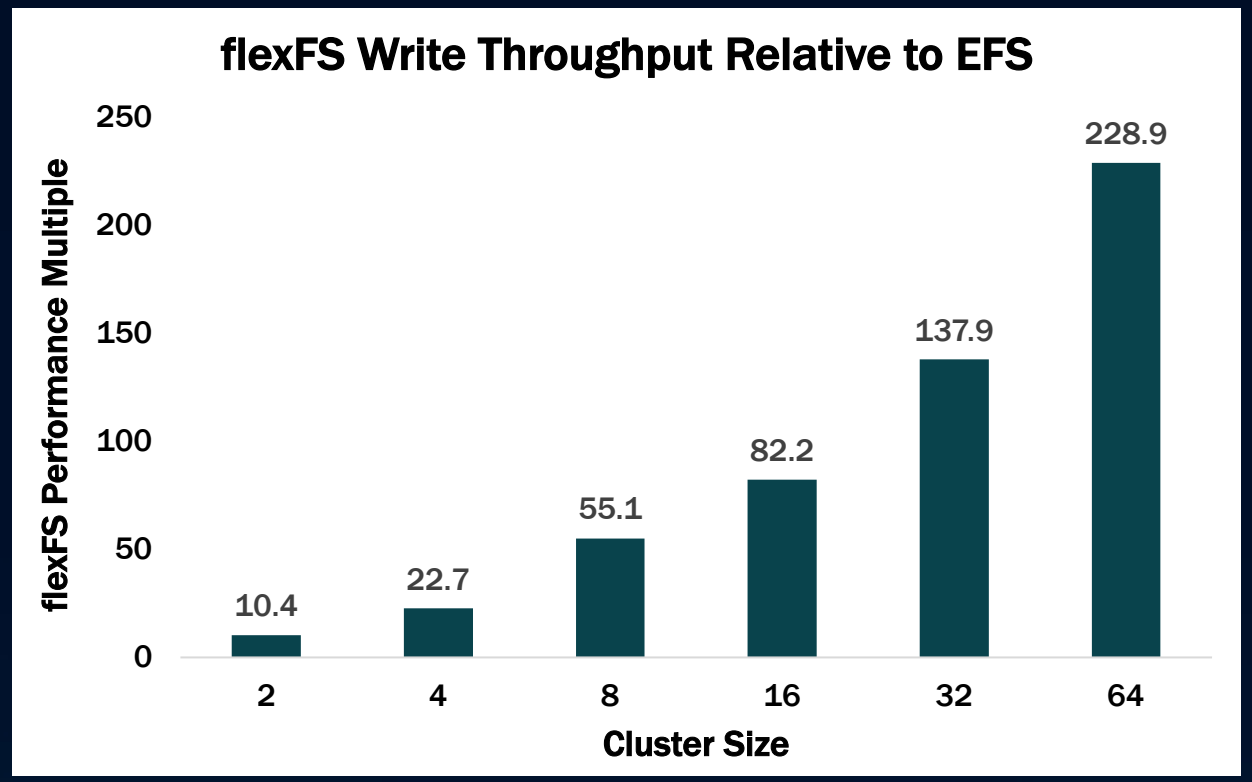


Despite there now being 5 concurrent clients, each client gets roughly the same throughput with AWS S3 CP and flexFS as it would if it were the only client. However, EBS (via SSHFS) and EFS both show roughly 5x decreases in per-client throughput. This decrease in throughput is linearly proportional to the (actively loaded) cluster size.

Use Case 1 – Read/Write Performance (LC-MS Data)

Operations performed on a ~6 GB LC-MS HDF5 file



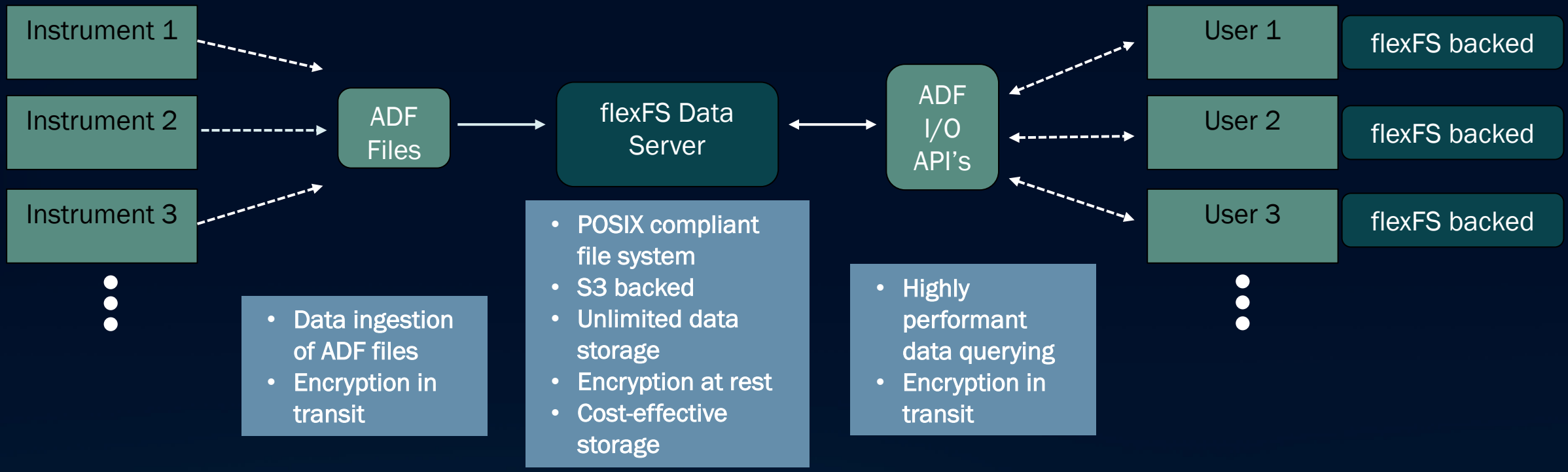


- Plot shows the multiples by which flexFS outperforms EFS

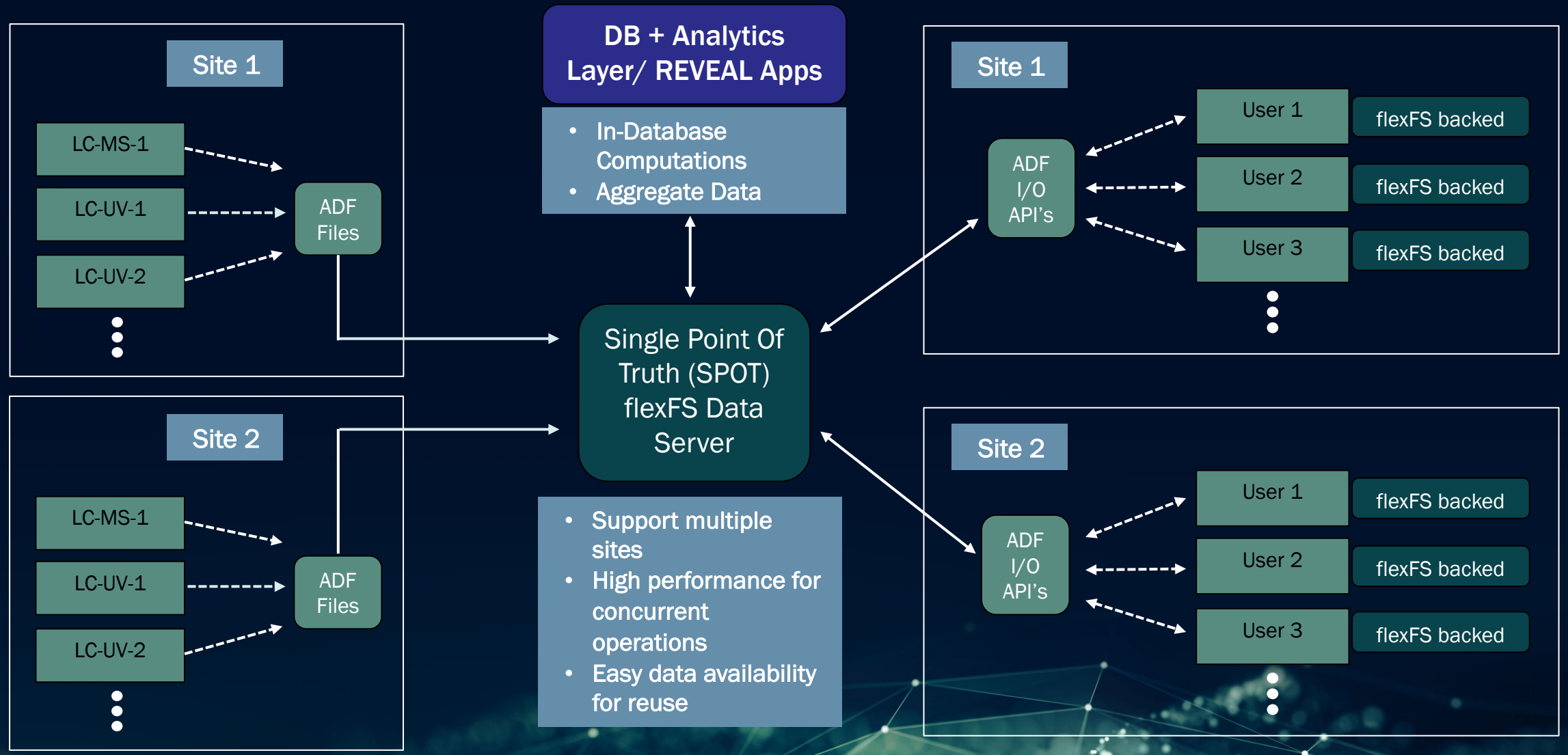
- Increase in relative performance of flexFS with cluster size
- c5n.18x large EC2 instance

Use Case 2 – Single-Site QC Workflow

a la Dr. Shum's paper



Use Case 3 – Multi-Site QC Workflow

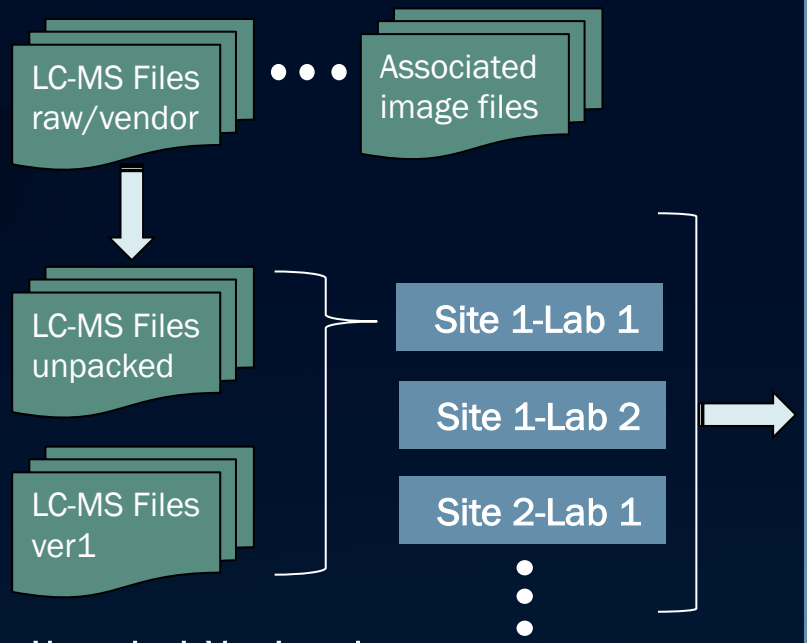


Storage Challenges with Data Growth

Data Generation, Storage, Accessibility

Raw; Vendor Format – e.g., 2 GB/ file

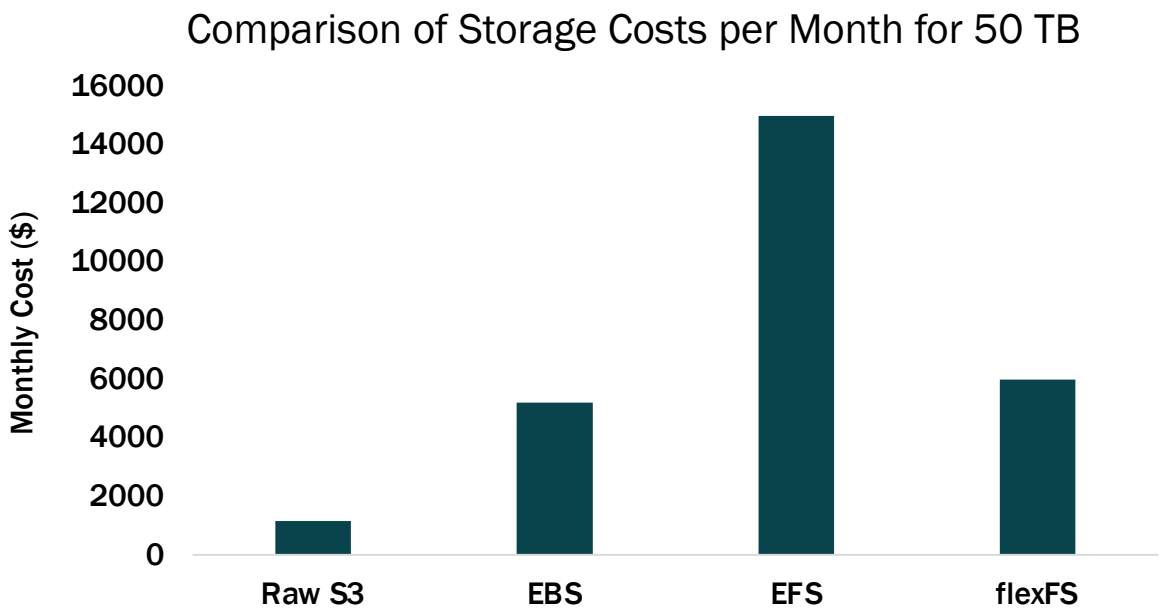
Study can have associated pathology images – e.g., 10 GB/ file



Unpacked; Versioned; Processed → ~3-5 X expansion – e.g., 10 GB/ file

- 10's of TB's of data generated/ year
- Storage required over time – raw, processed, analyzed
- High accessibility required to data
- Statistical algorithms helpful on central data server (SPOT)
- ADF format vital for federation

Current Landscape of Storage Solutions



- Estimates for storage of 50 TB of file data, actively accessed
- Comparisons made using multiple-availability-zone pricing
- Same region data storage & access (no transfer charges)
- S3 API costs, while not necessarily trivial were not considered
- On-demand pricing discounted by 30% is used for all server cost calculations
- File and metadata servers are all normalized to m5.2xlarge instances for consistency

flexFS Installation and Usage

Repeat on as many servers as desired.

```
# 1) Install FUSE userspace tools if missing.
$ sudo yum install -y fuse

# 2) Download and install the flexFS mount client
$ curl http://admin.flexfs.io/<path_to_dir>/mount.flexfs -o
mount.flexfs
$ chmod +x mount.flexfs
$ sudo mv mount.flexfs /sbin

# 3) Initialize the flexFS volume
$ sudo mount.flexfs init --admin-addr
<yourdomain>.admin.flexfs.io:443

# 4) Mount the flexFS volume
# Replace <mountpoint> with relevant folder (e.g.
/flexfs/<mountname>).

$ sudo mount.flexfs start <mountname> <mountpoint>
```

- Show mounted flexFS volume on drive
- Example operations using h5dump, h5ls, h5copy on HDF5 file
  - View root groups and structure
  - View specific group information
  - View specific dataset
  - Create new file from subset
- File transfers between flexFS backed server and client machines
- m5.8x large EC2 instances

Conclusion

- Based on work by Dr. Shum and others, it's clear that digitization will require
  - Elastic data management and
  - Elastic analysis
- There are many strategies for achieving elasticity
- We show the
  - Comparison of the scaling capability of
    - a posix compliant file store, flexFS, that works with block storage found in the cloud
    - Vs EFS found in the AWS cloud
    - The comparison in scaling showed the advantages of a solution with parallel read and write capabilities
  - The potential for savings using a file store, like flexFS for managing large file storage.


paradigm4
Thank You!