Using Scale to Improve Stewardship of Research Data Management & Storage

2022 Spectrum Scale User Group NYC & IBM

September 20, 2022

Shailesh Shenoy
Senior Associate, Department of Cell Biology
Assistant Dean for Einstein Information Technology
shailesh.shenoy@einsteinmed.edu

Albert Einstein College of Medicine

- Research Intensive Medical School
  - Our mission is to prepare a diverse body of students to become knowledgeable, compassionate physicians and innovative scientific investigators, and to create new knowledge

- About 1,900 Faculty & 1,000 Students (MD & PhD)
- Funding Primarily from the National Institutes of Health
- Part of Montefiore Medicine Academic Health System

Technology Opportunity

- Support Innovative Education, Learning, and Research
- Attract & Retain Top Researchers
- Achieve Faster Results
- Maintain Cyber Resilient Data
- Enable Robust Collaboration
- Provide Value

Legacy Ecosystem

- 2013 DDN GRIDScaler (Started with GPFS v3.5)
- Consolidated Various Systems
- Single Site
- Grew to ~10PB, ~2.5B files
- Shared Storage: Serves HPC & Home Directories
  - Compute cluster: 4,300 Cores / 56 A100 GPUs
  - NFS
  - SMB/CIFS

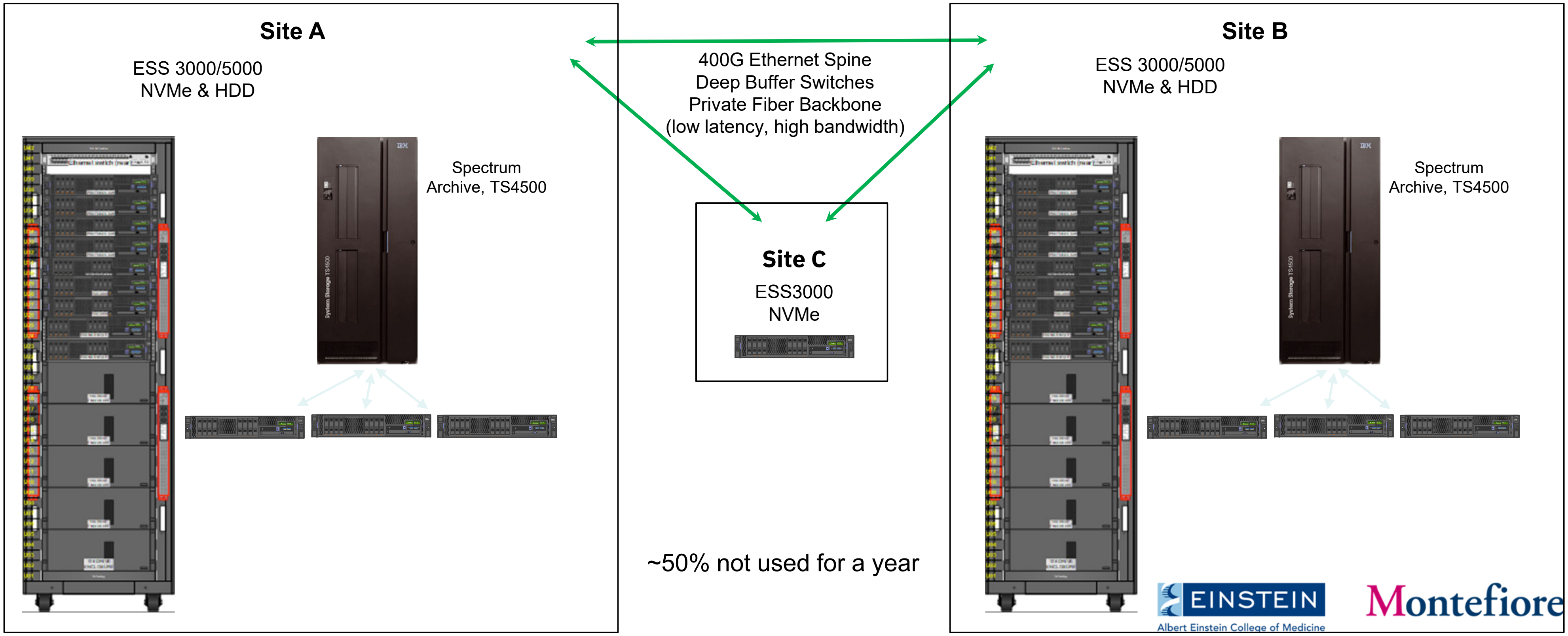
Business Requirements

- Single Namespace
- Mature, Robust, & Innovative High-Performance Filesystem
- Tiering for Cost Management
- Hardware/Software Ecosystem with Simple Scaling & Management
- Hybrid Cloud Capabilities
- Skilled Solution Partners
- Best of Class Support Organization
- High Speed Backup, Archiving and Disaster Recovery

Data Hygiene Goals

- Tiering: NVMe, HDD, Archive (Single Namespace)
- High Availability at Each Tier: No Single Point of Failure & Site Resiliency
- Air Gapped Backups Daily
- Disaster Recovery: Separate Daily Backup Copy Offsite with Ability to Restore
- Immutable File System Snapshots
- Independent File System for Protected Research Data
  - PII (Personally Identifiable Information) and PHI (Protected Health Information)

Stretch Cluster w/3-Site inode Replication



Site A

ESS 3000/5000
NVMe & HDD

Spectrum
Archive, TS4500

400G Ethernet Spine
Deep Buffer Switches
Private Fiber Backbone
(low latency, high bandwidth)

Site C

ESS3000
NVMe

Site B

ESS 3000/5000
NVMe & HDD

Spectrum
Archive, TS4500

~50% not used for a year

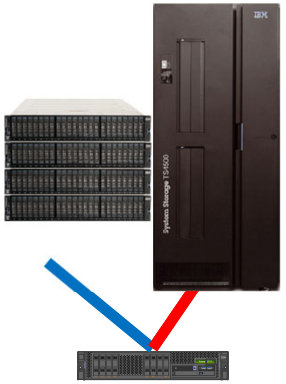
EINSTEIN
Albert Einstein College of Medicine

Montefiore

Backup w/Off-Campus Copy & Ability to Restore

Site C

Backup
Spectrum Protect, FS7200,
TS4500, ESS3000

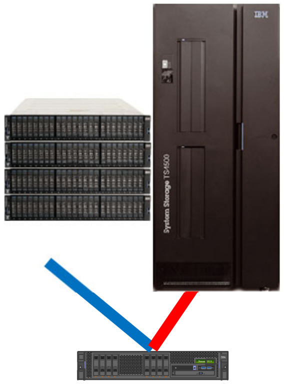


Backup with a Spectrum Protect Instance so that back-up data is in a different location from the stretch cluster locations

Private, Encrypted Fiber
(low latency, high bandwidth)

Site D

Spectrum Protect, FS7200,
TS4500, ESS3000/5000
NVMe & HDD



Off Campus:
separate power
grid

Leveraging Spectrum Protect Plus at Sites C & D
for Enterprise Backup and Disaster Recovery

EINSTEIN
Albert Einstein College of Medicine

Montefiore

Data Migration

- IBM Technology Services
- Moved from v4.2.3 -> v5.1.3
- Optimized Data Layout
  - Independent & Dependent File Sets
- Standardize ACLs
- Established Master Policy Engine to Manage File Placement
- Operational Readiness & Transition Team

Stewardship Next Steps

- Grafana: Analytics & Visualization Stakeholder Dashboards
- Collaboration: Enhance Integration of Aspera & Globus
- Qradar: Monitor Access Patterns, Trigger Safe-Guard Copy
- Cloud Scale Instance: AFM over S3
- Spectrum Discover: Data Classification
- Guardian: Document Data Lineage

Acknowledgements

- Einstein Team
  - Ian Grant – Network Services Manager
  - Donni Frid – Director of Infrastructure & Applications
  - Brian Hammond – Director of Scientific Computing Systems

- Data in Science Technologies Team
  - Andrew Gauzza III
  - Bill Pappas

- IBM Team
  - Richard Rupp
  - Dave Cooper
  - Joe Sanjour (IBM Technology Services)
  - Todd Blight
  - Mark Sternefeld

Questions

- Thank you