Are Repositories Impeding Big Data Reuse?

Zhiwu Xie[1], Andrej Galad[2], Yinlin Chen[1, 2], and Edward Fox[2]
[1]University Libraries and [2]Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, USA

June 14, 2016, Open Repositories 2016, Trinity College, Dublin, Ireland

VirginiaTech
Invent the Future®

Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse (LCI)

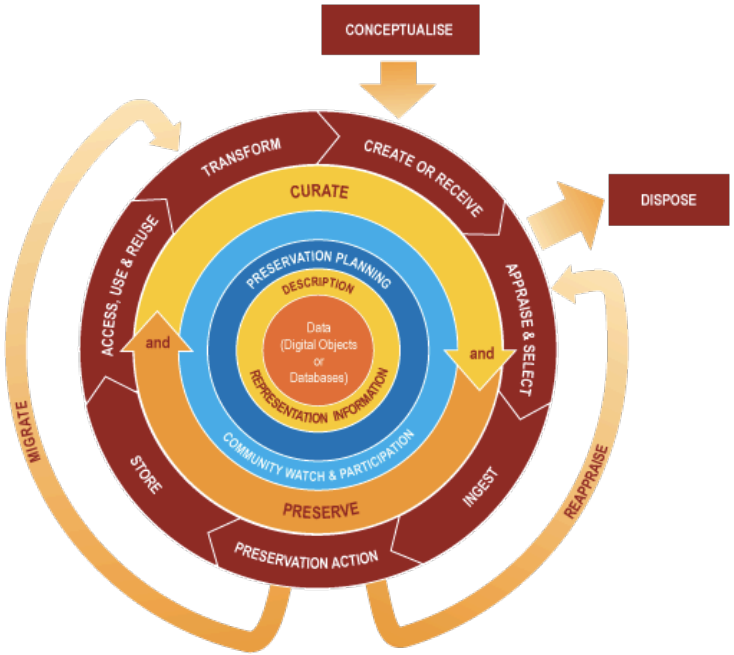
- IMLS LG-71-16-0037, $308,175 from June 2016 to May 2018
- A National Leadership Grant research project in the National Digital Platform program, Lead by VT and UNT
- Evaluate 3 library big data reuse patterns against 5 typical IT infrastructure
- This presentation focuses on operating the “bridge” reuse pattern in commercial clouds
- Lead to a decision tree to guide library IT strategy
- http://lci.lib.vt.edu

Big Data Management

- Libraries participate the management of big data
- Varieties -> volume and velocity
- More of a question for IT infrastructure

Reuse Driven Big Data Management

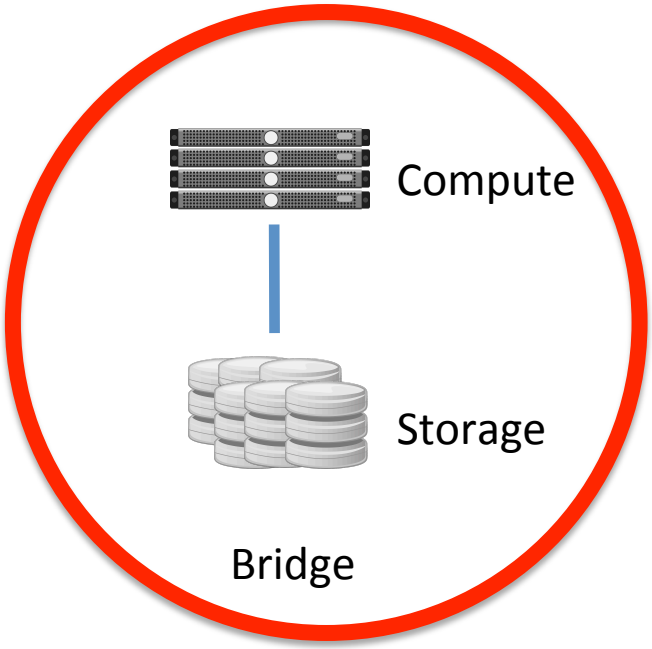
- Sharing patterns change
- Much more than just long-term storage
- Storage must be paired with sufficient processing capacity
- Warehouse -> Workshop
- Xie, Zhiwu, et al. "Towards use and reuse driven big data management." Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries. ACM, 2015.

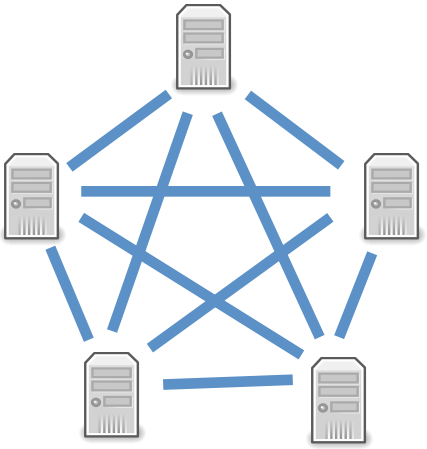


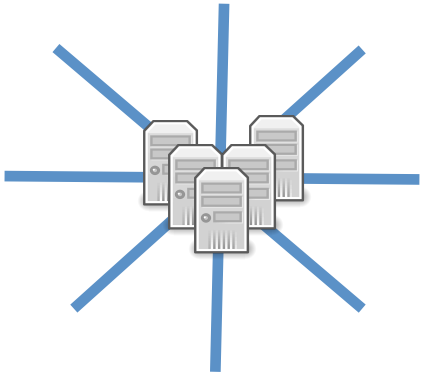
http://www.dcc.ac.uk/resources/curation-lifecycle-model

Chris 73 / Wikimedia Commons

Library Big Data Reuse Patterns





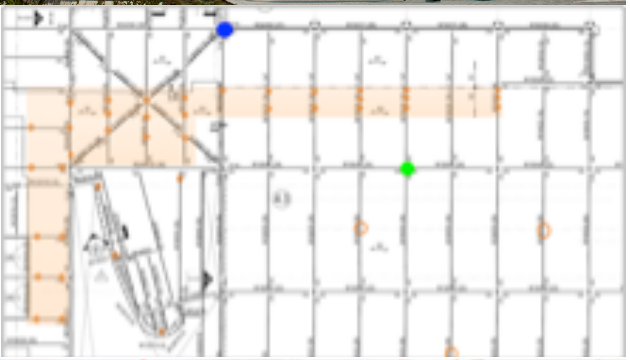


The Bridge Pattern

- Data are explicitly stored in a different facility from where they are processed
- Network links must be provided between the two
- Upon reuse, data must be moved from storage to processing via the links
- Assuming the data processing is vastly parallelizable, the link usually becomes the performance bottleneck

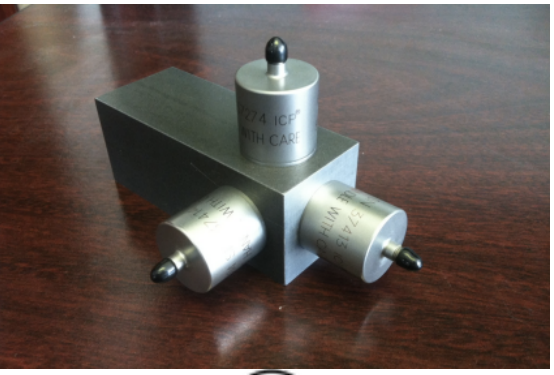
5 Shared IT Infrastructure Options

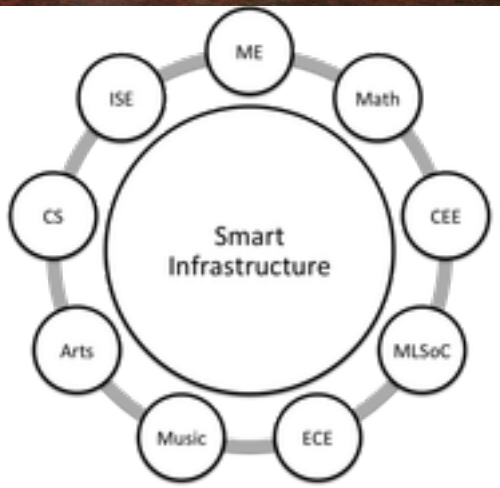
- Local clusters, e.g., DLRL Hadoop Cluster
- Institutional HPC resources
- National HPC resources, e.g., XSEDE TACC
- National research computing clouds, e.g., XSEDE Chameleon
- Commercial clouds, e.g., AWS, Azure



Goodwin Hall Living Lab

- Goodwin Hall Living Lab
- A 160,000-sf new building wired with >240 different sensors
- Sensor mounts were directly wielded to the structural steel during the building construction
- Sensors are strategically positioned and sufficiently sensitive to detect human movements
- Will be the most instrumented building for vibration

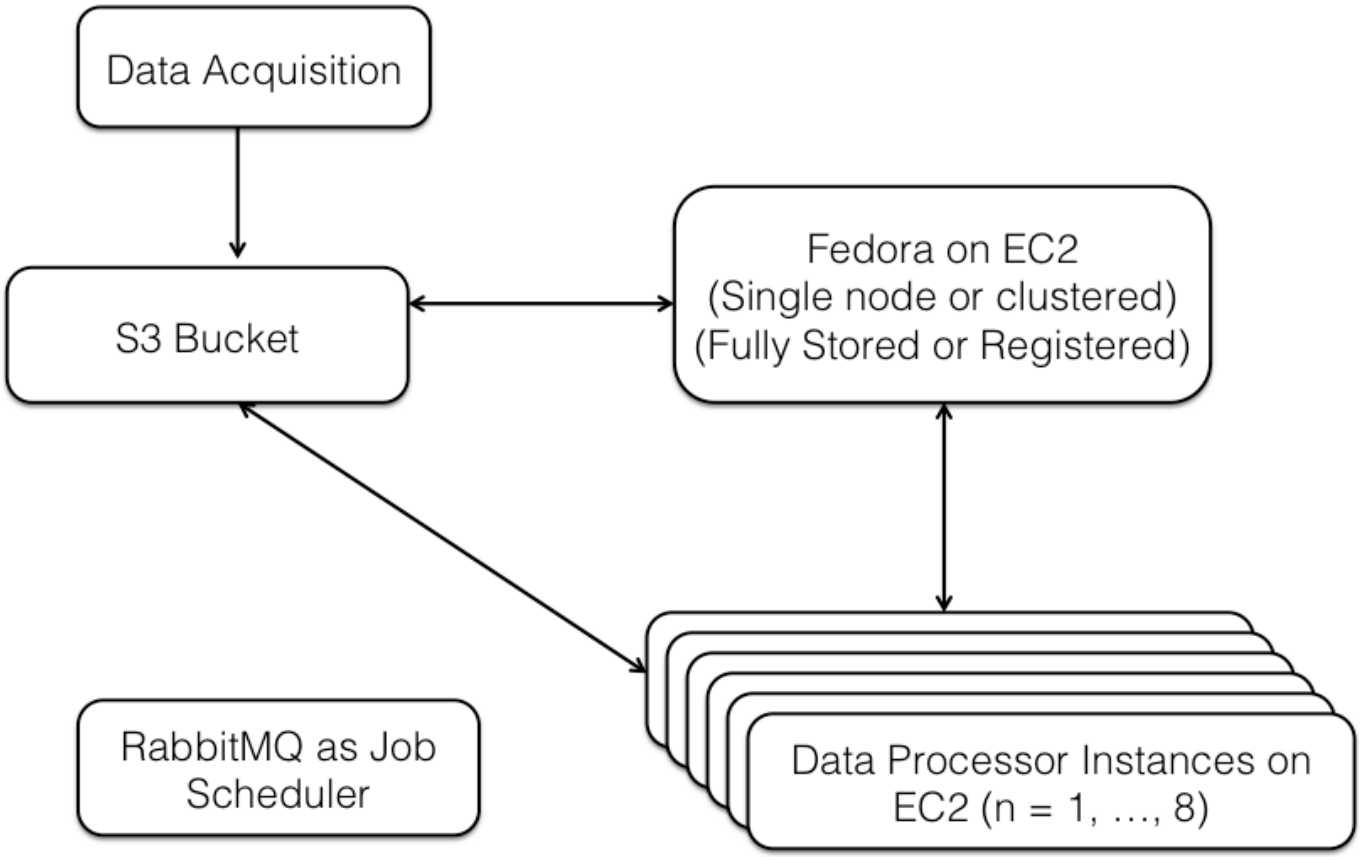




Challenges

- How fast will the repository allow us to move data to computing facilities?

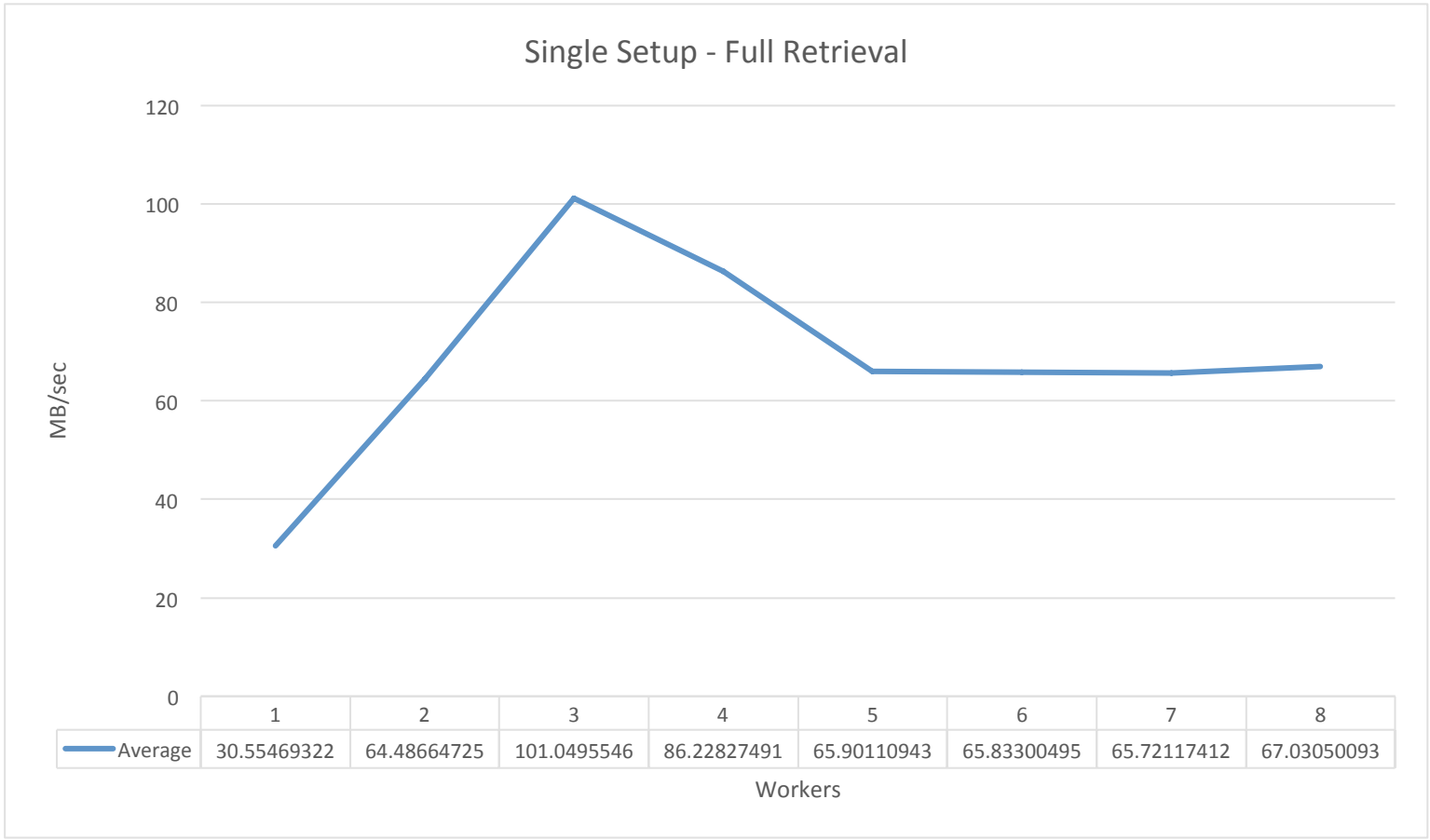
Benchmarking System Architecture



Benchmarking System Configuration

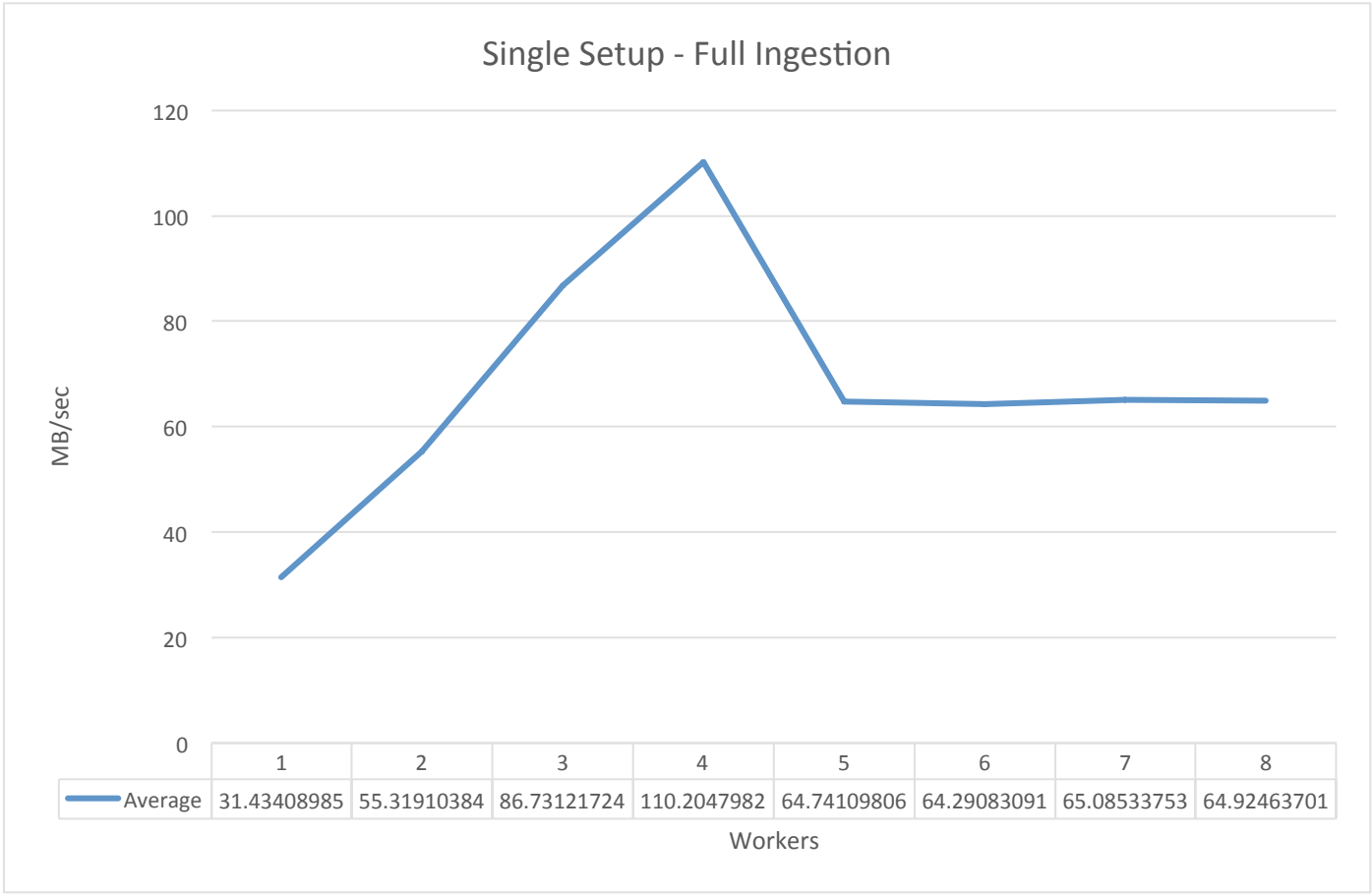
- 1 full day's data collected at moderate sampling frequency, compressed, stored in multiple hdf5 files, total ~240G
- Fedora - m4.2xlarge (8 vCPUs, 32GB RAM, 400GB drive, based on doc: max throughput: 1000Mbps, max bandwidth: 125MB/s)
- RabbitMQ - m4.xlarge (4 vCPUs, 16GB RAM, 8GB drive, based on doc: max throughput: 750Mbps, max bandwidth: 93.75MB/s)
- Workers - t2.medium (2 vCPUs, 4GB RAM, 8GB drive)

Single Node Fedora Retrieval Speed



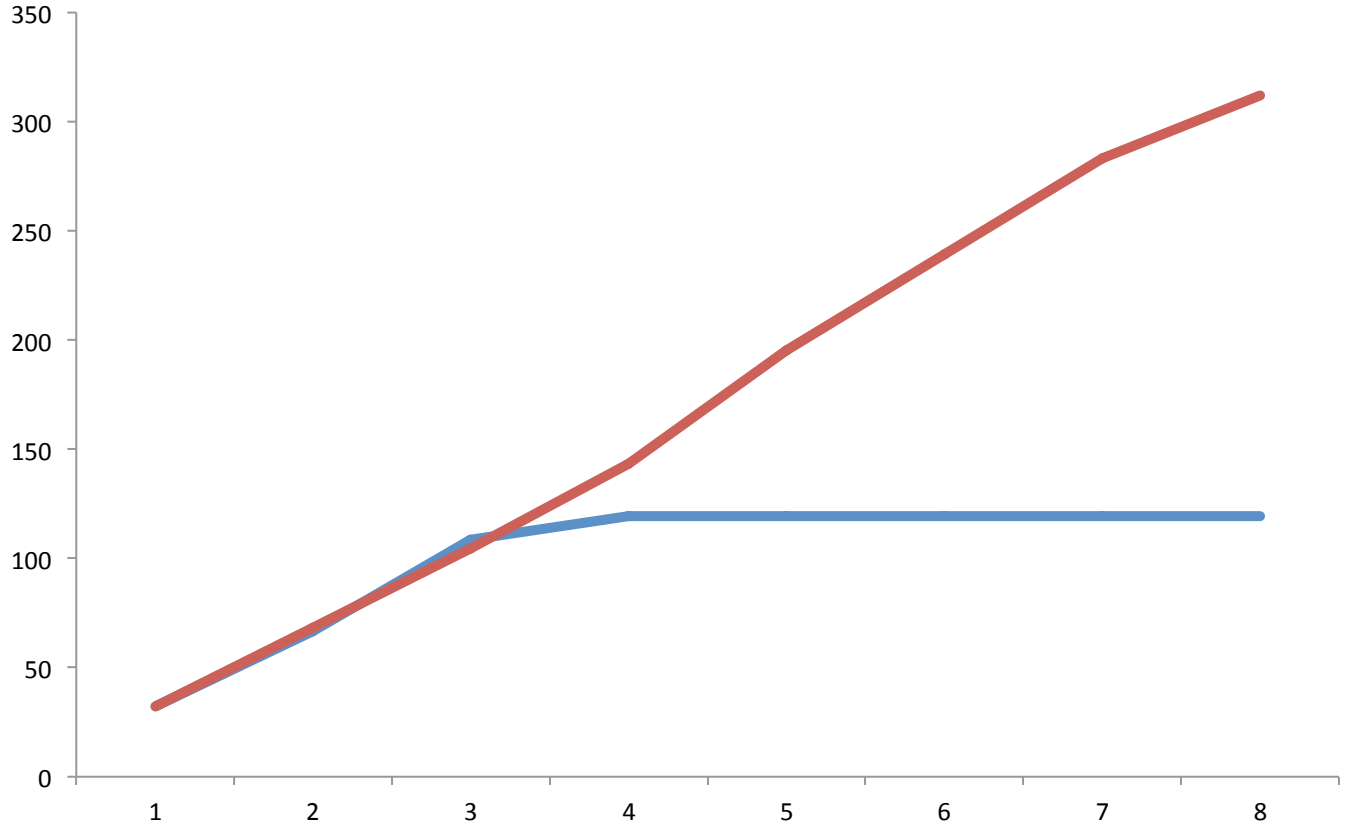
	1	2	3	4	5	6	7	8
Average	30.55469322	64.48664725	101.0495546	86.22827491	65.90110943	65.83300495	65.72117412	67.03050093

Single Node Fedora Ingestion Speed

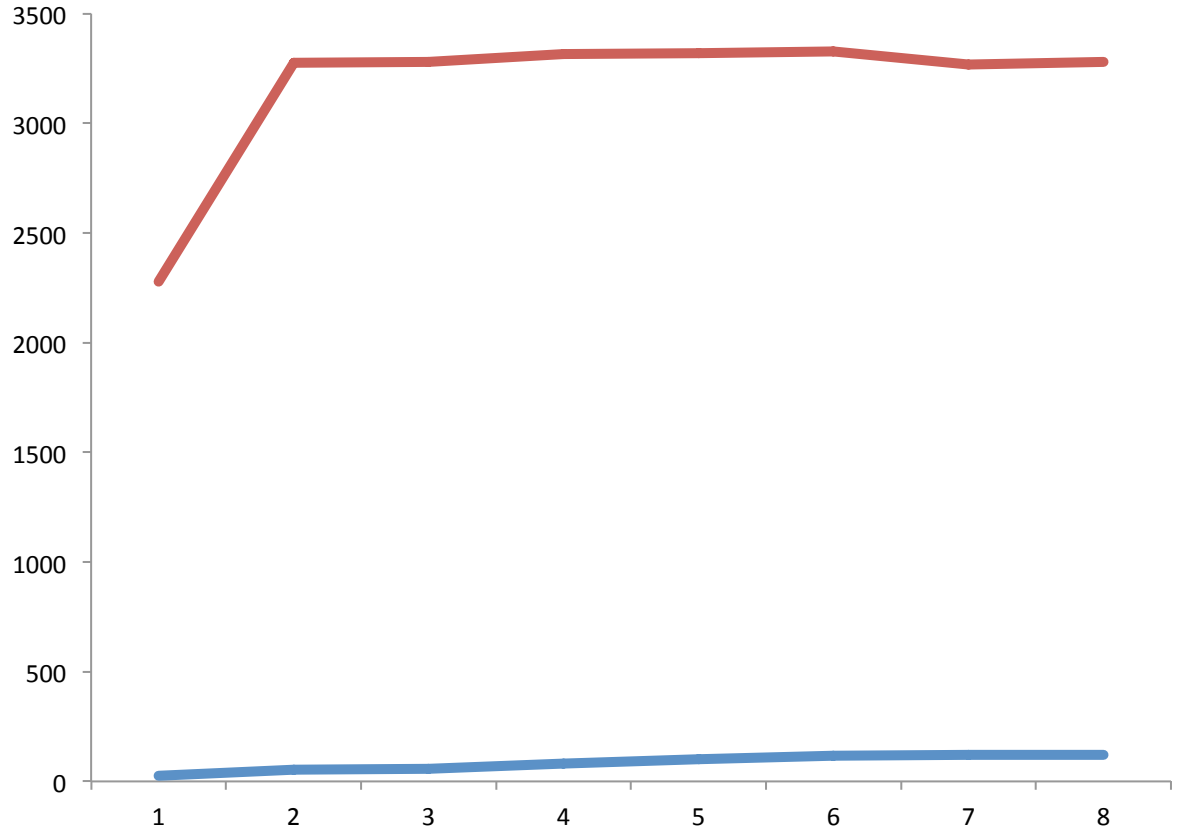


	1	2	3	4	5	6	7	8
Average	31.43408985	55.31910384	86.73121724	110.2047982	64.74109806	64.29083091	65.08533753	64.92463701

2-Node Cluster Retrieval Speed: Store in Fedora vs. Register with Fedora



2-Node Cluster Ingestion Speed: Store in Fedora vs. Register with Fedora



Summary

- Single node repository bottlenecks on the network bandwidth
- Multi-node repository cluster still bottlenecks on the network bandwidth, but scales better than the single node system
- It is much faster to store data in highly scalable storage systems such as Amazon S3 then register with Fedora than directly store large data sets in Fedora

Questions

zhiwuxie@vt.edu @zxie