HARVARD
UNIVERSITY

Handling BigData: filesystems and transfers?

Scott Yockel, PhD
Harvard - Research Computing

Processing & Analyzing Data
DataFest'17

1

HARVARD
UNIVERSITY

FASRC

Research Computing Resources

- Manage 71,000 cores
- Over 2 Petaflops SP GPGPU
- 35.0PB of storage
- 600+ virtual machines (KVM)
- 2MW of research computing equipment in 3 data centers
- 20 FTE in 4 groups
  - ARCS: Advanced Research Computing Support Group
  - HPC: High Performance Computing
  - Software as Infrastructure (OpenNebula/VMs, Containers)
  - Data Center & Operations
- Supporting 600 research groups and 3000+ users across FAS, SEAS, HSPH, HBS, GSE.

2

HARVARD UNIVERSITY — FASRC

Outline

- Part 1: Where does it come from?
  - Traditional HPC : Numerically Intensive
  - External Repositories : Data Intensive
  - Modern Instruments : Data + Numerically Intensive
- Part 2: How do we deal with it?
  - File systems
  - File transfers

HARVARD UNIVERSITY — FASRC

Traditional / Historical Computing

- Dating back to the Manhattan Project solving Physics, Chemistry & Engineering problems have predominately created data on-the-fly / at runtime.
- 1966, during Robert Mulliken's Noble Prize acceptance speech: "I would like to emphasize strongly my belief that the era of computing chemists, when hundreds if not thousands of chemists will go to the computing machine instead of the laboratory for increasingly many facets of information is already at hand."

HARVARD UNIVERSITY FASRC

Traditional / Historical Computing

- Solving complex mathematical equations takes lots of memory and storage
  - Schrödinger's Eqn: Eigenvalue problem of Quantum Mechanics

$$\boxed{H\Psi = E\Psi \longrightarrow \hat{H} \sum_{i=1}^{n} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{n}\sum_{A=1}^{M} -\frac{Z_A}{r_{iA}} + \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} + \sum_{A=1}^{M} -\frac{1}{2}\nabla_A^2}$$

  - Navier-Stokes Eqn: Differential Equation used in a great variety of physical systems, including: Geophysics, Oceanography, Atmospheric Sciences, Aerodynamics, Plasma Physics, Astrophysics

$$\boxed{\frac{\partial u}{\partial t} + u \cdot \nabla u = -\frac{1}{\rho}\nabla \bar{p} + \nu \nabla^2 u + \frac{1}{3}\nu \nabla(\nabla u \cdot u) + g}$$

  - Maxwell's Eqn: Differential Equation used to describe electricity and magnetism

$$\boxed{\begin{array}{ll} \nabla \cdot D = \rho & \nabla \times E = -\frac{\partial B}{\partial t} \\ \nabla \cdot B = 0 & \nabla \times H = J + \frac{\partial D}{\partial t} \end{array}}$$

5

HARVARD UNIVERSITY FASRC

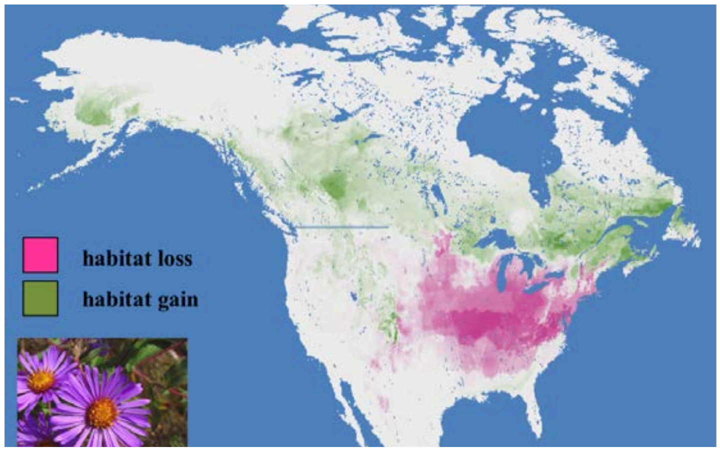
External Repositories

- 1990s: Large scale collections & high-performance networks (Internet2) begin to make it feasible for universities to access datasets that were once isolated to only a few institutions.
- At large access to data created an explosion of interest in disciplines that are data intensive
  - Biroinformatics
  - Climate modeling
  - Seismic modeling
  - Data Science (mining, analysis, map reduce, …)
  - Astrophysical Image Analysis

6

HARVARD UNIVERSITY FASRC

Species Migration (Davis - OEB/Habaria)



- 12.5k species @ ~1GB/species = 12.5 TB
- 6 different models = 75 TB

7

HARVARD UNIVERSITY FASRC

Modern Instrument Data

- Processing raw data and statistical analysis used to be done by high-end workstation when:
  - Data sets were smaller: 1-2 GB
  - Statistical algorithms were less rigorous

8

HARVARD UNIVERSITY

Multi-SEM (Lichtman - MCB)

3 TB

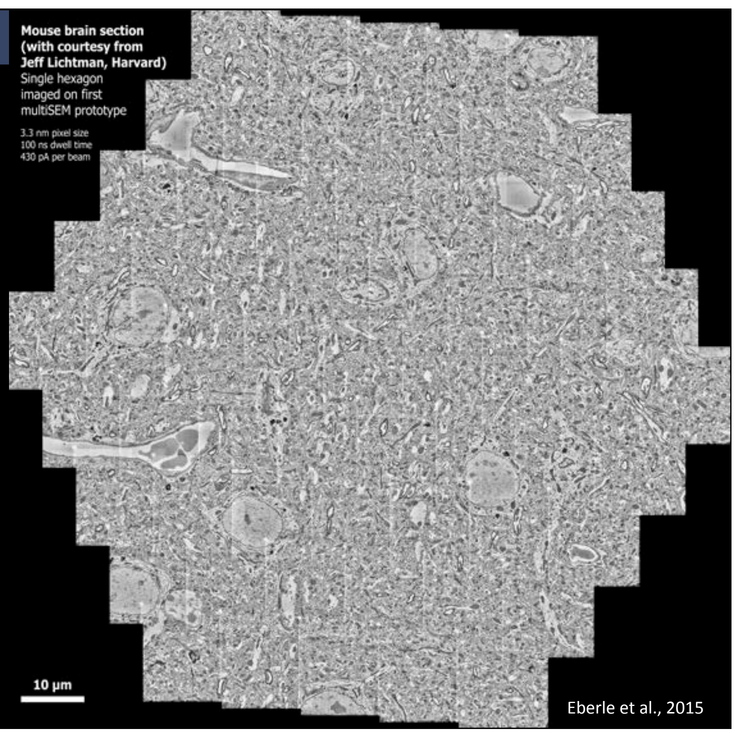
Every

Single

HOUR!

HARVARD UNIVERSITY

61 Beam Hexagonal Image 3.3 nm pixel resolution



What took 5 hours now takes 5 minutes!

Eberle et al., 2015

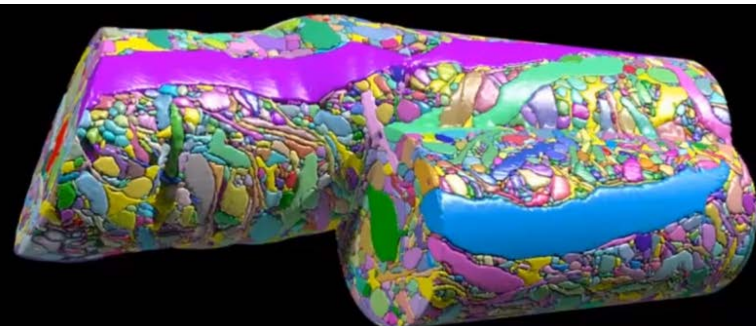
HARVARD UNIVERSITY

FASRC

Intersection Where Biology, Engineering & Computer Science Meet

11

HARVARD UNIVERSITY



Speedup with fully automated segmentation 25 hours of “tracing” by CPU cluster using a convolutional neural net algorithm vs 1.25 centuries of human tracing!

HARVARD UNIVERSITY FASRC

Outline

- Part 1: Where does it come from?
  - Traditional HPC : Numerically Intensive
  - External Repositories : Data Intensive
  - Modern Instruments : Data + Numerically Intensive
- Part 2: How do we deal with it?
  - File systems.
  - File transfers.

13

HARVARD UNIVERSITY FASRC

Filesystem Concepts Comparisons

local	remote
single server	clustered / distributed
meta-data - block maps, time stamps, file attributes, extended attributes	data blocks - the contents of the file
cache: fast access of recent data - server memory, controller cards, …	disk writes: - final resting place of data
IOPS: In/Out Operations per second	read/writes: bandwidth of filesystem
backups - disaster recovery	snapshots - deletion protection

14

HARVARD UNIVERSITY FASRC

Filesystem Concepts Comparisons

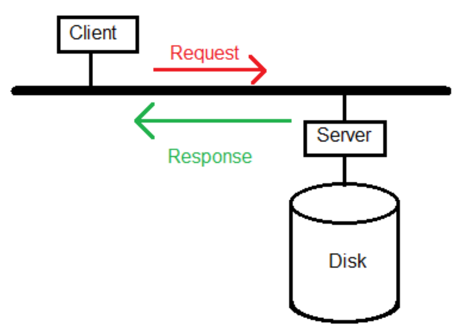
local	remote
single server	clustered / distributed
meta-data - block maps, time stamps, file attributes, extended attributes	data blocks - the contents of the file
cache: fast access of recent data - server memory, controller cards, …	disk writes: - final resting place of data
IOPS: In/Out Operations per second	read/writes: bandwidth of filesystem
backups - disaster recovery	snapshots - deletion protection

15

HARVARD UNIVERSITY FASRC

Single Server Filesystem

- NFS: Network File System
- SMB/Samba, CIFS
- RPC: Remote Procedure Call



16

HARVARD
UNIVERSITY

FASRC

Clustered Filesystem

- Isilon: OneFS
- Red Hat: Gluster, Ceph
- IBM: GPFS
- Apache: HDFS (Hadoop)
- Panasas: Panfs
- Open Source: Lustre, BeeGFS, OrangeFS

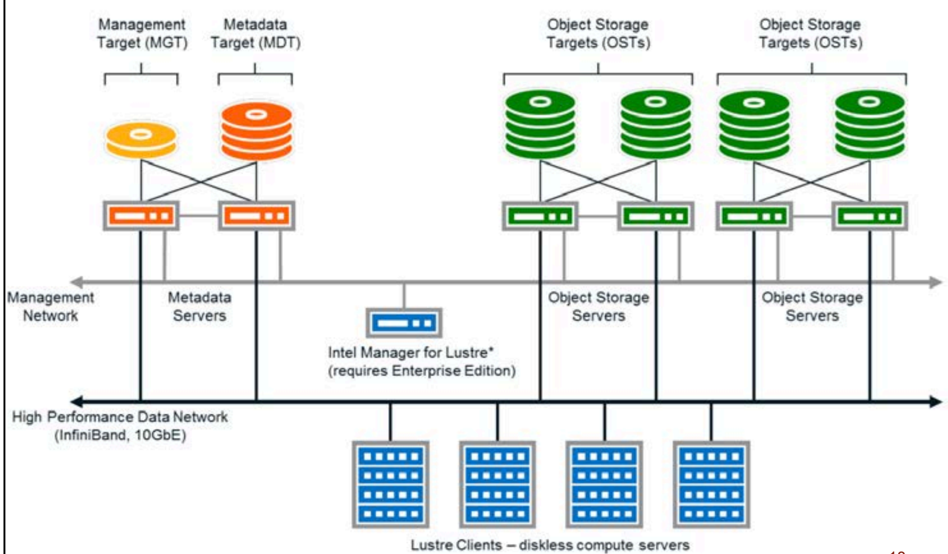
- GOALS: Failure tolerant, scalability, migration, replication, concurrency.

17

HARVARD
UNIVERSITY

FASRC

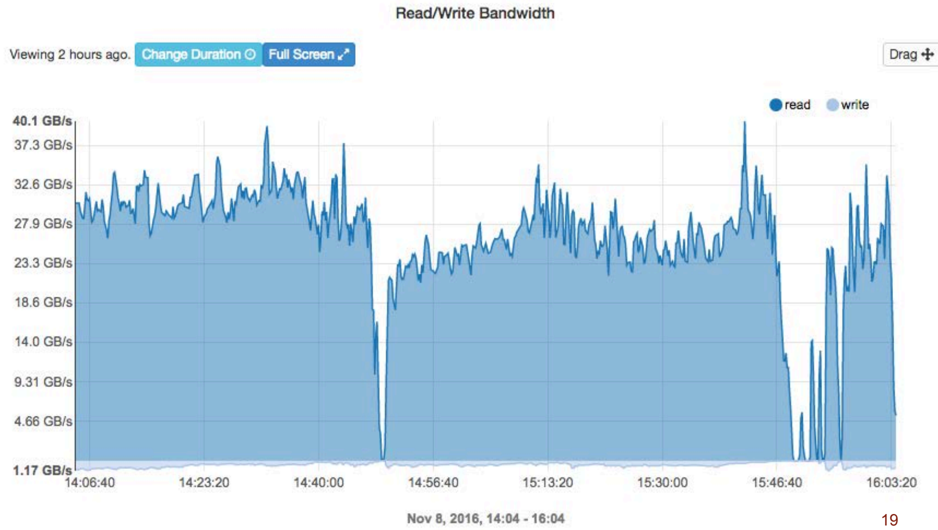
Lustre Filesystem



18

HARVARD UNIVERSITY FASRC

Because sometimes your scratch filesystem takes a beating!



19

HARVARD UNIVERSITY FASRC

Single Thread Filesystem Tests

Size	Local	Remote	Lustre
10 MB	0.025 s, 423 MB/s	0.341 s, 30.7 MB/s	0.0614 s, 171 MB/s
1 GB	17.1 s, 61.2 MB/s	13.6 s, 76.9 MB/s	5.35 s, 196 MB/s
10 GB	190.3 s, 55.1 MB/s	118. s, 89.0 MB/s	76.8 s, 137 MB/s

20

HARVARD UNIVERSITY FASRC

Single Thread Filesystem Tests

Size	Local	Remote	Lustre
10 MB	0.025 s, 423 MB/s	0.341 s, 30.7 MB/s	0.0614 s, 171 MB/s
1 GB	17.1 s, 61.2 MB/s	13.6 s, 76.9 MB/s	5.35 s, 196 MB/s
10 GB	190.3 s, 55.1 MB/s	118. s, 89.0 MB/s	76.8 s, 137 MB/s
40 kB	0.00103s, 39.7 MB/s	0.0302 s, 1.4 MB/s	0.0122 s, 3.4 MB/s
4 kB	0.000325 s, 12.6 MB/s	0.00374 s, 1.1 MB/s	0.00236 s, 1.7 MB/s

21

HARVARD UNIVERSITY FASRC

Be kind to your filesystem

- zillions of tiny files
  - consider creating records inside of files
  - create archive files for data not in use
- > 2k files in a directory
  - avoid listing file metadata
    - Example: 75k dirs: 13s to list vs 31m to list details (metadata)
- excessive nested directories
  - 32k sub-directories is the ext3 limit
- avoid 100s of simultaneous access of same file
- avoid special characters and spaces: $ * , < > : ^ ! | &
  - can cause issues when moving between OSs.
- Filesystems are typically built with specific needs in mind
  - scratch vs repository vs archive
  - IOPs vs read/writes

22

HARVARD UNIVERSITY FASRC

Outline

- Part 1: Where does it come from?
  - Traditional HPC : Numerically Intensive
  - External Repositories : Data Intensive
  - Modern Instruments : Data + Numerically Intensive
- Part 2: How do we deal with it?
  - File systems.
  - File transfers.

23

HARVARD UNIVERSITY FASRC

File Transfers

- FTP: File Transfer Protocol circa 1980
  - Still one of the most popular ways to transfer files
  - Secured with SSH (SFTP)
  - Every hop in the network slows performance
  - Single thread server and client
- HTTP, wget, cURL, …
  - similar performance to FTP
  - Single threaded both server and client
- SCP: Secure Copy
  - Secured with SSH (SFTP)
  - Single threaded client
- TCP: Transmission Control Protocol
  - This is the network backbone of the Internet
  - Every packet sent must recieve an acknolegment packet back
  - The performance decreases with distance (number of hops)

24

File Transfers

- Multi-threaded: Just send more TCP data streams
- rsync: two server synchronization and file transfer program for Linux
  - add SSH for security, add zlib for compression
  - PhD project of Andrew Tridgell, who also wrote SMB
  - Ex: single 10G file, 105 MB/s on 40Gb network
- bbcp: developed at Stanford
  - similar usage to scp
  - encrypts auth, but not the data
  - Ex: single 10G file, 200 MB/s on 40Gb network
- robocopy: robust file copy command-line tool in Windows
  - similar performance to multi-channel SMB
  - 4 threads need to saturate 1Gbps network

HARVARD UNIVERSITY FASRC

File Transfers

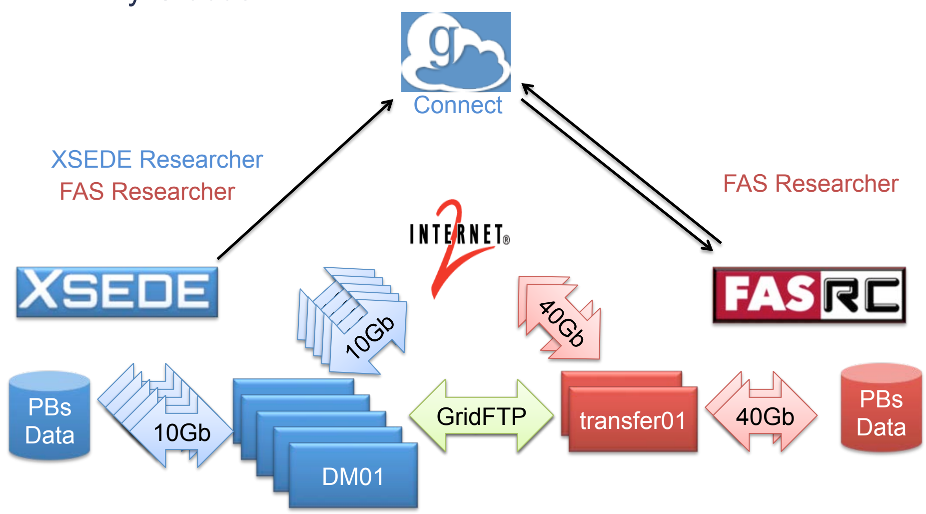
- Drop TCP all together and go with UDP: User Datagram Protocol
  - Simple connectionless transmission model, minimum protocol
  - No acknowledgements, use checksums for data integrity
- GridFTP & Globus Connect ($$)
  - Multiple servers, multiple threads
  - Federated authentication
  - encryption optional, checksums, and automated retransmit
  - endpoints and sharing
- Aspera: Acquired by IBM ($$)
  - ascp: secured by RSA, data transferred by UDP
  - NCBI Ex: 40 Mb/s FTP vs 4 x 1-2Gb/s ascp on single 10GbE

HARVARD UNIVERSITY

FASRC

Getting Data In/Out

- Why Globus?



27

HARVARD UNIVERSITY

FASRC

Data Transfers Considerations

- Resumable
- Multi-threaded/channel
- Data Integrity: network (tcp ack) or software (checksums)
- Data encryption (or not)
- Access controls
- Data locality

- More on transfering large amounts of data via network:
  - http://moo.nac.uci.edu/~hjm/HOWTO_move_data.html
  - Harry Mangalam, UC Irvine

28

HARVARD
UNIVERSITY

Questions ???

Scott Yockel, PhD
Harvard - Research Computing

SIGHPC: BigData
Supercomputing'16

29