Data Storage and Data Analysis Workflows for Research

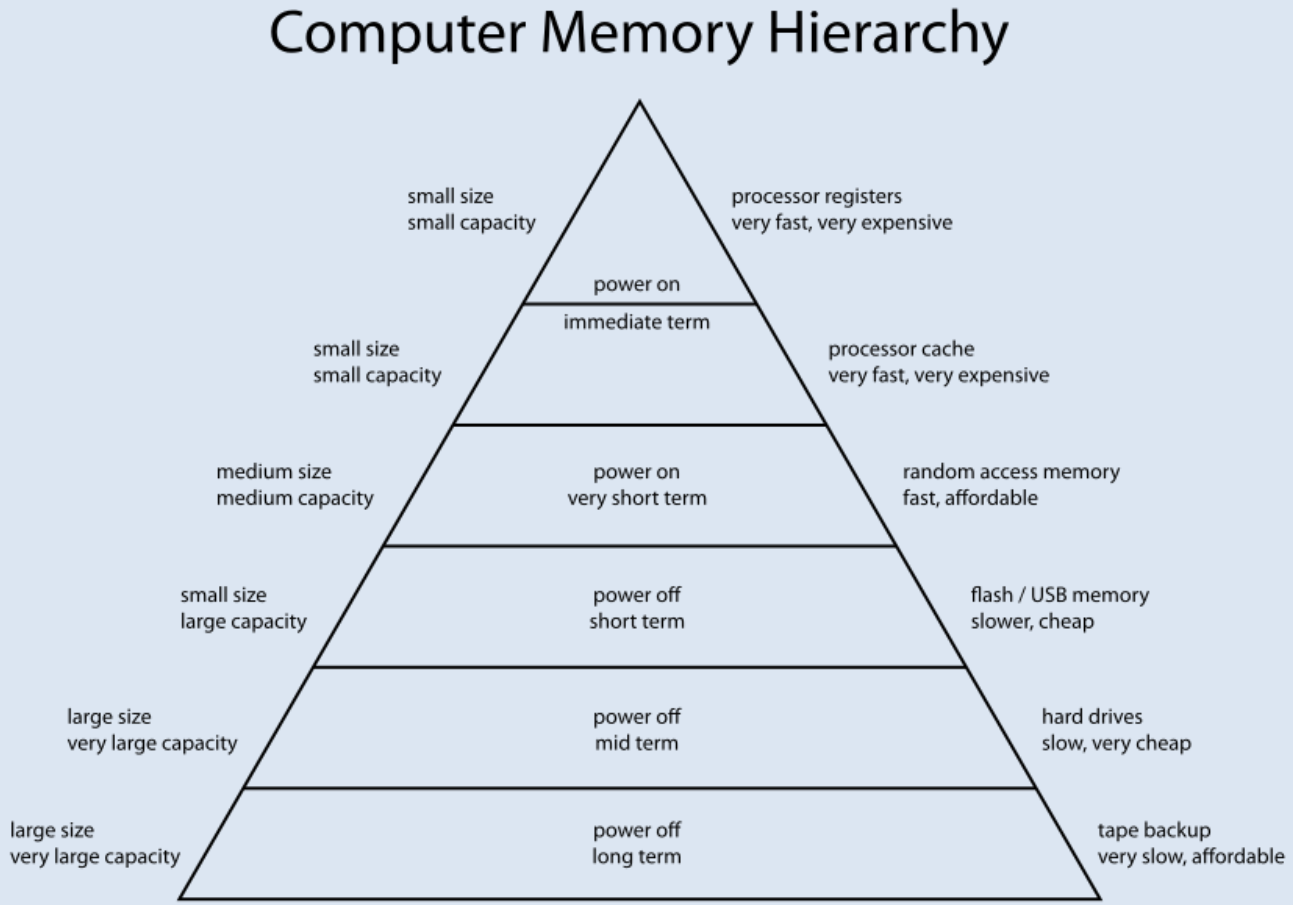
September 29, 2015

The Minnesota Supercomputing Institute for Advanced Computational Research

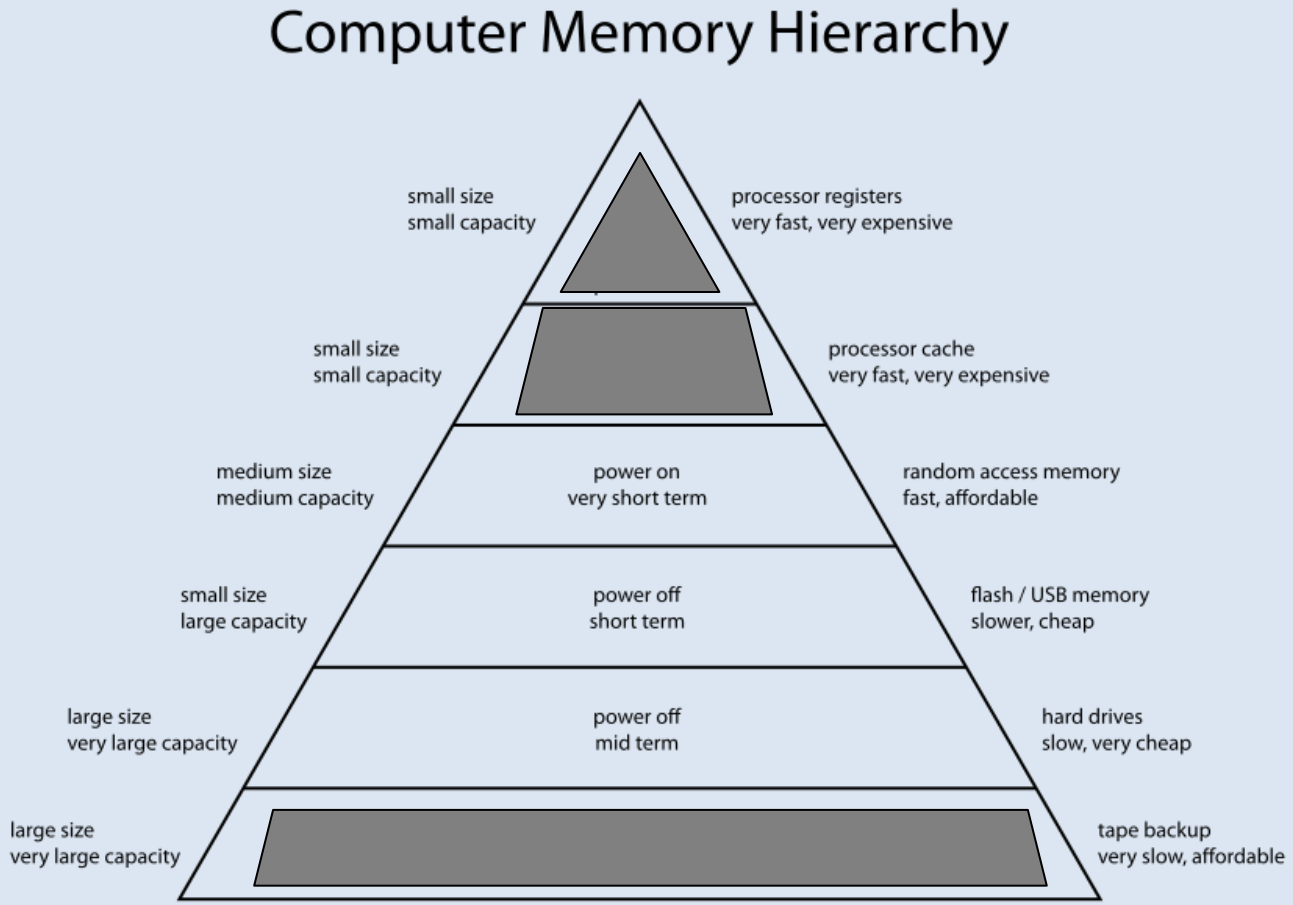
Tutorial Outline

- Hardware overview
- Systems overview
- Options at UMN
- Options at MSI
- Interfaces to MSI storage
  - Moving data on and off storage systems
- Performance issues
- Use Cases
- Hands on

Storage Technologies
Hardware



Storage Technologies
Hardware



Ask Questions First

Not all storage is created equal

- What do I want to do with the data?
  - How large are the files I'm storing?
  - How many files will I store?
  - How frequently will I access the data?
  - From what locations will I access the data?
  - In what format will the data be stored?

Storage Technologies
Hardware

- Hard Disk Drives (HDDs)
  - Rotating rigid platters on a motor-driven spindle within a protective enclosure. Data is magnetically read from and written to the platter by heads that float on a film of air above the platter.
- SATA -- Serial Advanced Technology Attachment
  - Desktop
  - Low cost
  - up to 8 TB
  - ~ 6 Gb/s
  - ~1.6 million hours MTBF
- SAS -- Serial Attached SCSI
  - Enterprise use
  - Costly
  - up to 8 TB
  - ~ 12 Gb/s
  - ~1.2 million hours MTBF

Storage Technologies
Hardware

- Solid State Drives (SSDs)
  - Use microchips which retain data in non-volatile memory chips and contain no moving parts.
  - No moving parts
  - less susceptible to physical shock
  - silent
  - very low access time
  - very expensive (Compared to HDDs)
  - ~1.5 million hours

- Hybrid HDD and SSD drives (SSHD)
  - SSDs add speed to cost effective media by acting as Cache

Storage Technologies
Hardware

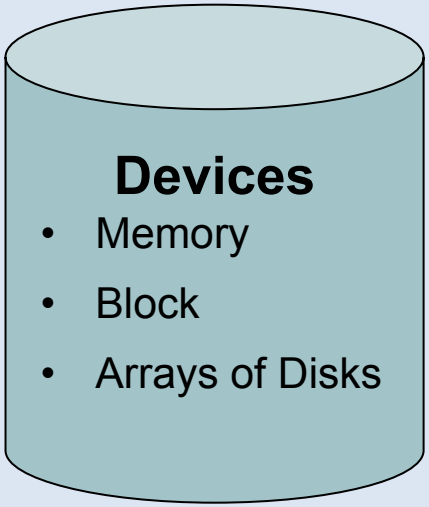
- RAM Disk
  - Block of random-access memory (primary storage or volatile memory) that a computer's software is treating as if the memory were a disk drive (secondary storage).
  - Used to accelerate processing
  - No moving parts
  - Very low access time (Compared to HDDs and SDD
  - Very expensive (Compared to HDDs and SDDs)
  - Data lost when powered off or rebooted

Storage Technologies
Future of Storage

- Better conventional HHDs
  - Helium Filled
  - Shingled Magnetic recording (SMR)
  - Heat-assisted magnetic recording (HAMR)
- Better/Cheaper Solid State solutions?
  - Phase Change Memory (PCM)
  - Could flatten complex data hierarchies?
- DNA digital data storage for archive storage
  - Very slow but extremely dense

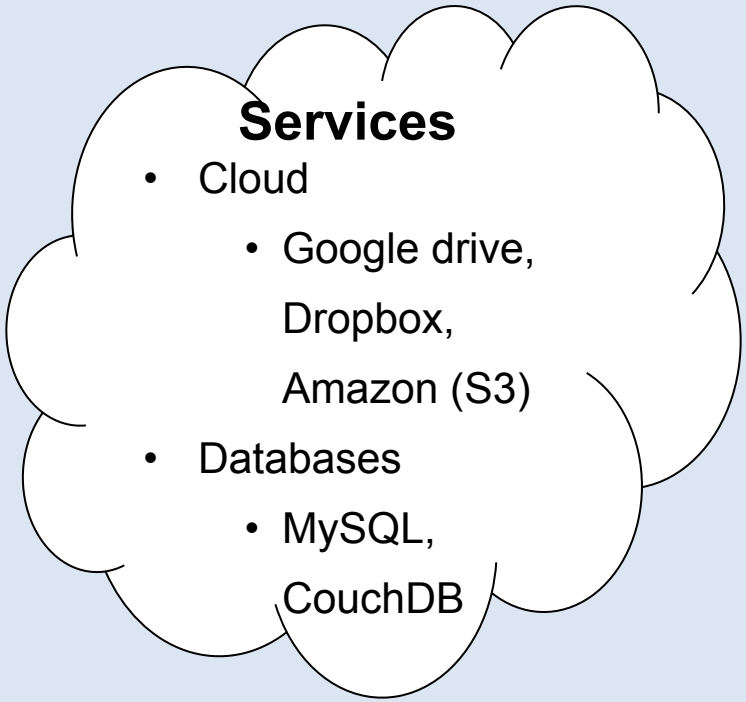
Storage Technologies

How do we use these devices?



Filesystems

- Disk File Systems
  - Ext4, ZFS
- Network File Systems
  - NFS, SMB
- Parallel File Systems
  - Panasas, Lustre, GPFS
- Special Cases
  - FUSE (Filesystem in Userspace)
  - CephFS



Storage Technologies
Order of Magnitude Guide *

Storage	Files/dir	File sizes	Band Width	IOPs
Local HDD	1,000s	GB	100 MB/s	100
Local SSD	1,000s	GB	1 GB/s	10,000+
RAM FS	10,000s	GB	10 GB/s	10,000
NFS	100s	GB	100 MB/s	100
Lustre/GPFS	100s	TB	100 GB/s	1,000
Cloud	Infinite	TB	10 GB/s	0
DB	N/A	N/A	N/A	1,000

*From SDSC 2015 Summer institute: HPC and Long Tail of Science

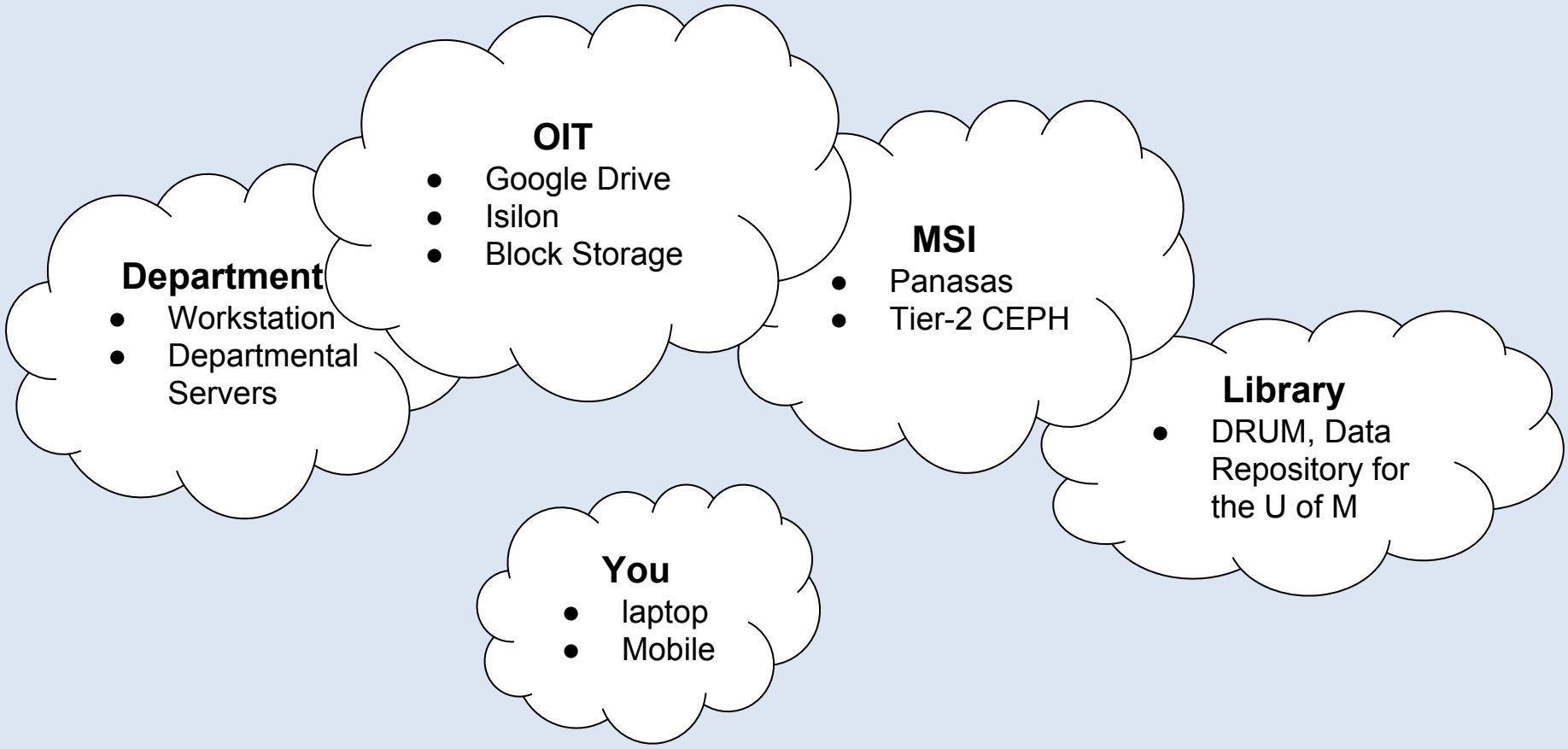
Storage Technologies
Data Redundancy

- Mirroring
  - Create identical copies of Files
- RAID (Redundant Array of Independent Disks)
  - Multiple disks pooled into a single logical unit
  - RAID with N=2 is Mirroring
  - Larger disk pools (N>2) can save storage
  - Uses a parity to recreate missing data when drive is lost
- Snapshot
  - Creates a copy of the current state of the system to disk
  - Very fast, doesn't delay subsequent writes.
- Tape backup
  - Refers the the media, portable
  - Typically less expense
  - Offline for Disaster recovery purposes.

Storage Technologies
A Cautionary Tale

http://www.youtube.com/watch?v=gDfLXAtRJfY&feature=youtu.be

Storage Options at UMN



Storage Options at UMN

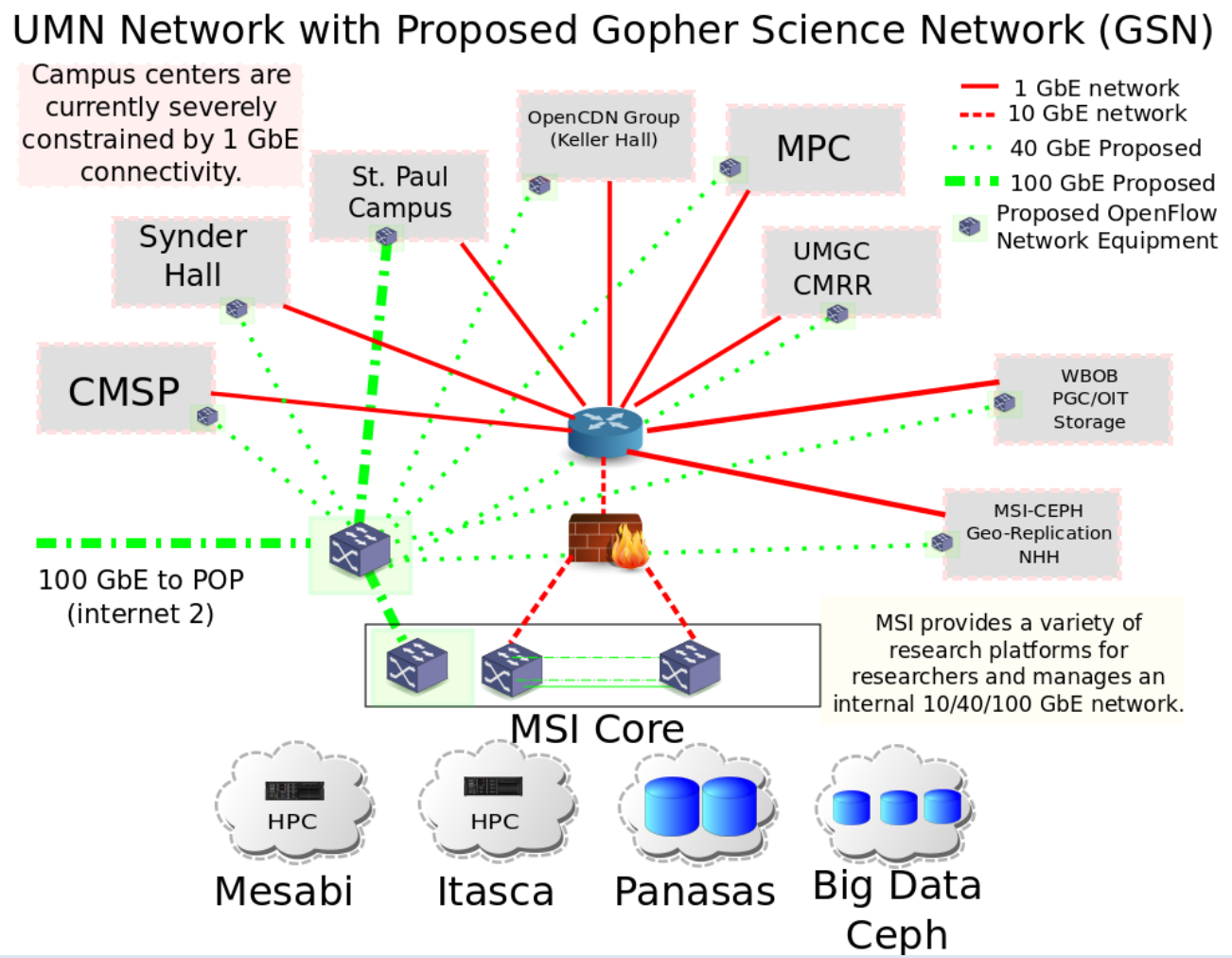
Purpose	Google Drive	OIT Isilon	OIT Block	MSI Panasas	MSI Tier-2	Dept Storage	Laptop/ Desktop
Big Data		✓?		✓	✓	?	
High Performance				✓		?	
Share access	✓				✓	?	
Archival (very long-term) storage		✓			✓?	?	
Access on Campus Laptop/Desktop		✓			✓?	?	
Access on anywhere Laptop/Desktop/Mobile	✓	✓?			✓?	?	
Access on Servers		✓	✓	✓?	✓?	?	
Legally protected data (Coming)	☒	☒	☒	☒	☒	☒	?

So Many Choices, So Much Data

Sometimes it's best to come with good questions rather than a single solution:

http://www.youtube.com/watch?v=F4OIDszDA9E!

Gopher Science Network at UMN



Storage at MSI

Store and Stage Data

What's available at MSI:

- Shared file system: PanFS
- 2nd Tier Storage: CEPH
- Databases: Web servers
- Local Disk
- RAM disk

Shared File system

What it is

PanFS: Block storage; POSIX

Visible on all MSI systems

Persistence: duration of your account at MSI

How you access it:

Directories: home, shared, public, scratch

Shell commands: cp, mv, rm, grep, …

Applications: all POSIX file IO

Shared File system

Locations & Uses

/home/<group>/<user>	Your private files
/home/<group>/shared	Share with your group
/home/<group>/public	shared with all MSI
/scratch	Temp. files for multiple hosts

Limits

/home/<group>/*	group quota (allocation)
/scratch	1 month lifetime

2nd Tier Storage

What it is

CEPH: Object storage; S3

Visible on all MSI systems and Web

Persistence: duration of allocation

How you access it:

By file only

Files organized in “buckets”

Shell: s3cmd

Web URL

https://www.msi.umn.edu/content/second-tier-storage

CEPH: S3 interface

Locations & Uses

s3://<bucket name>/<file name>

s3cmd commands: ls; get; put

Save & stage large volumes of data

Limits

CEPH write access by group allocation

CEPH read access can be granted by user

Databases & Web Services

What it is

Database services & servers managed by MSI

Visible on hosts with web access

Persistence: lifetime of project

How you access it:

Web URL

Shell: wget or database clients

Get access through a coordinated MSI project

Databases

Locations & uses

URL: www.msi.<name>

Share data with a community

Informatics applications

Limits

Capacity & bandwidth specific to project

Local Disk

What it is

Non-RAIDed Disk or SSD: POSIX

Visibili on host system only

Persistence: duration of PBS job

How you access it:

Shell commands: cp, mv, …

Applications: all POSIX file IO

Local Disk

Locations & Uses

/scratch.local

/<user>/<path>]/<file name>

Scales well to many hosts writing to their own files

Limits

Scope: local host and life of PBS job

relatively poor bandwidth, except for fragmented IO

Typical capacity: 420 GB

RAM Disk

What it is

Local system memory

Visible only on local host

Persistence: duration of PBS job

How you access it:

Shell commands: cp, mv, …

Applications: all POSIX file IO

RAM Disk

Locations & uses

/dev/shm

/path]/<file name>

Scalable to many hosts reading their own files

High bandwidth and low latency

Efficient fragmented IO

Limits

About ½ system memory (32 GB on a Mesabi node)

Scope: local to node and only during PBS job.

Data Hierarchy: Mesabi Compute Node

	Capacity	Latency	Bandwidth	Access
Cache	60 MB	~ 10 ns	~ 3 TB/s	In Process
Memory	64 GB - 1 TB	~ 100 ns	~ 30 GB/s	In Process
RAM Disk	32 GB - 512 GB	~ 0.1 ms	~ 400 MB/s * N	POSIX IO
SSD	440 GB	~ 0.26 ms	~ 400 MB/s	POSIX IO
Local Disk	420 GB	~ 24 ms	~ 100 MB/s	POSIX IO
PanFS	2.5 PB	~ 2 ms	0.4 - 25 GB/s	POSIX IO
CEPH	2.1 PB	~ 1 sec	60 - 200 MB/s	By File (S3)
WAN	→ Infinity	~ 1 sec	1 - 60 MB/s	By Web service

- Cache to register bandwidth based on HPL efficiency
- I've measured memory BW at 28 GB/s; cache: 267 GB/s
- Latencies and bandwidths are as measured in real apps.

Interfaces
(Getting Started)

Move data to and from MSI

Applications, utilities, & services

scp	can push to msi from external host
wget	Pull from within MSI only
Git	Pull or push from within MSI only
s3cmd	Push data to and pull data from CEPH
Globus	Web based control from anywhere

Access for incoming traffic

Must be within UofM domain (use UofM VPN)

Must go through MSI front end server

login msi umn.edu

Secure Copy (scp)

- *Login to MSI host*
- *Copy files & directories between a remote server and MSI*

Login to MSI

ssh <msi_user>@login.msi.umn.edu

Copy to MSI

scp <rhost_user>@<rhost>:<path>/<file> <path>

scp -r <rhost_user>@<rhost>:<path> <path>

Copy from MSI

scp <file> <rhost_user>@<rhost>:<path>

scp -r <path> <rhost_user>@<rhost>:<path>

Get Files from web (wget)

- Run client (wget) from MSI host
- Get files, source code, data posted on web

  Files must be posted on a server that support wget

  You must have the URL

On an MSI host: get a file from the web:

wget <URL>

Repositories (git)

- Sharing data & source with others: Version control
- Can run git locally or with a github
- UofM github:

On MSI host: command prompt

get add

git commit

git merge

CEPH (s3cmd)

What is it good for?

- Move large volumes of data to and from CEPH
- Stage and share data for processing
- High bandwidth: up to 250 MB/s

From MSI Linux shell (command prompt)

s3cmd mb s3://<bucket>

s3cmd put <file> s3://<bucket>

s3cmd get s3://<bucket>/<file> <directory>

s3cmd ls s3://<bucket>

https://www.msi.umn.edu/support/faq/how-do-i-use-second-tier-storage-command-line

Globus

What is it good for?

- Move data between sites across WAN
- Web GUI driven
- Move LARGE directory trees with drag and drop
- Runs in background

How to use

- Get Gobus account:
- Register your certificate ID with Globus endpoints
- Use web GUI to drag and drop between endpoints

www.globus.org

Use Cases
(HPC Workflows)

Cross OS Workflows

Use case

- Complex geometry & physics
- Computationally intensive solutions
- Use commercial software (example: ANSYS)

The issue

- ANSYS Workbench & GUIs run best on MS Windows
- ANSYS solvers scale excellently on Mesabi (Linux cluster)

The solution

- Setup model & view results w/ GUIs on Citrix VMs
- Run solvers on Linux cluster
- Use PanFS home directory as the glue

Data Intensive Workflows

Use case:

Need to process many large files

Need to access various subsets of data in many ways

The issues:

Total volume of data to large for group quota

fragmented IO slow on shared file system

MANY users on shared file system → very slow access

The Solution:

Stage full data set on CEPH in may files

Stream needed files to RAM disk in PBS jobs

Process on RAM disk and save results to PanFS or CEPH

Post processing example

Have: raw data from an MHD turbulence model.

Mesh res: 256x256x256

Full state info: (density, velocity, B-field)

300+ snapshots in time

Individual snapshot size: 470 MB

Want: Power spectra of velocity field

Post-process each time snapshot

Can be done independently

Calculation (including IO) takes ~16 s

Serial workflow

```
command          status
-------------    ----------
./do1spc 0000    FINISHED
./do1spc 0001    FINISHED
./do1spc 0002    INPROGRES
./do1spc 0003    NEW
./do1spc 0004    NEW
     ...
```

Run app. on state 0002

Raw data PanFS
Generate V-spectra
copy to output directory

e6a02-0000-000
e6a02-0001-000
e6a02-0002-000
e6a02-0003-000
e6a02-0004-000

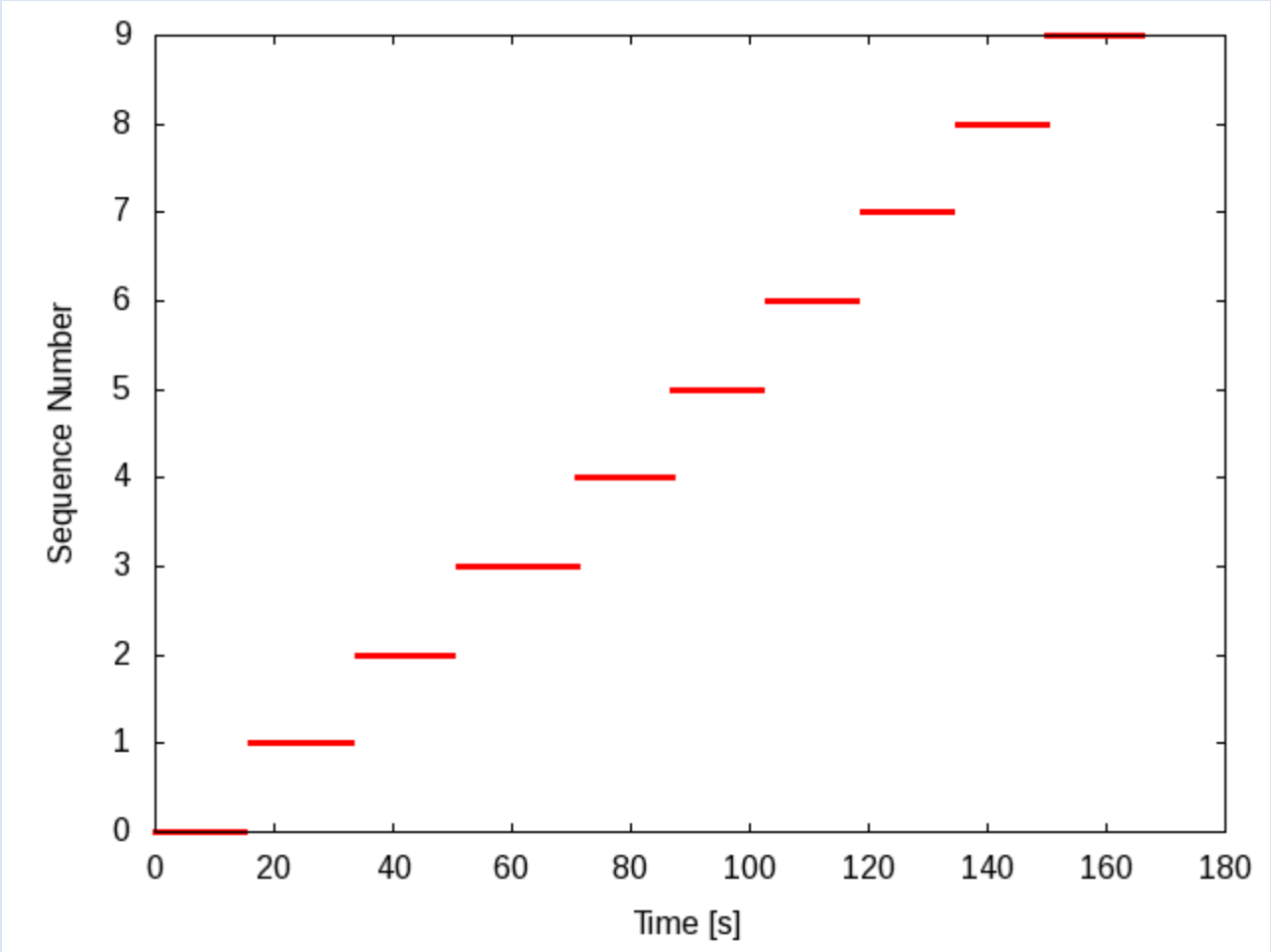
...

e6a02-0000-V3.spc3v
e6a02-0001-V3.spc3v
e6a02-0002-V3.spc3v
e6a02-0003-V3.spc3v

Serial Throughput (0-9)

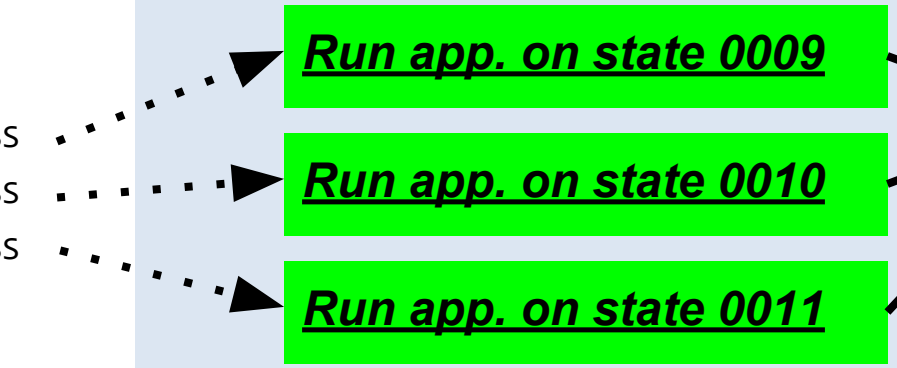
Lines show span of time each work item took

1 work item = process one time snapshot



Parallel workflow

```
    ...
./do1spc 0007  FINISHED
./do1spc 0008  FINISHED
./do1spc 0009  INPROGRESS
./do1spc 0010  INPROGRESS
./do1spc 0011  INPROGRESS
./do1spc 0012  NEW
./do1spc 0013  NEW
./do1spc 0014  NEW
    ...
```




```
    ...
e6a02-0007-V3.spc3v
e6a02-0008-V3.spc3v
e6a02-0009-V3.spc3v
e6a02-0010-V3.spc3v
e6a02-0011-V3.spc3v

    ...
```

Parallel throughput (0-40)

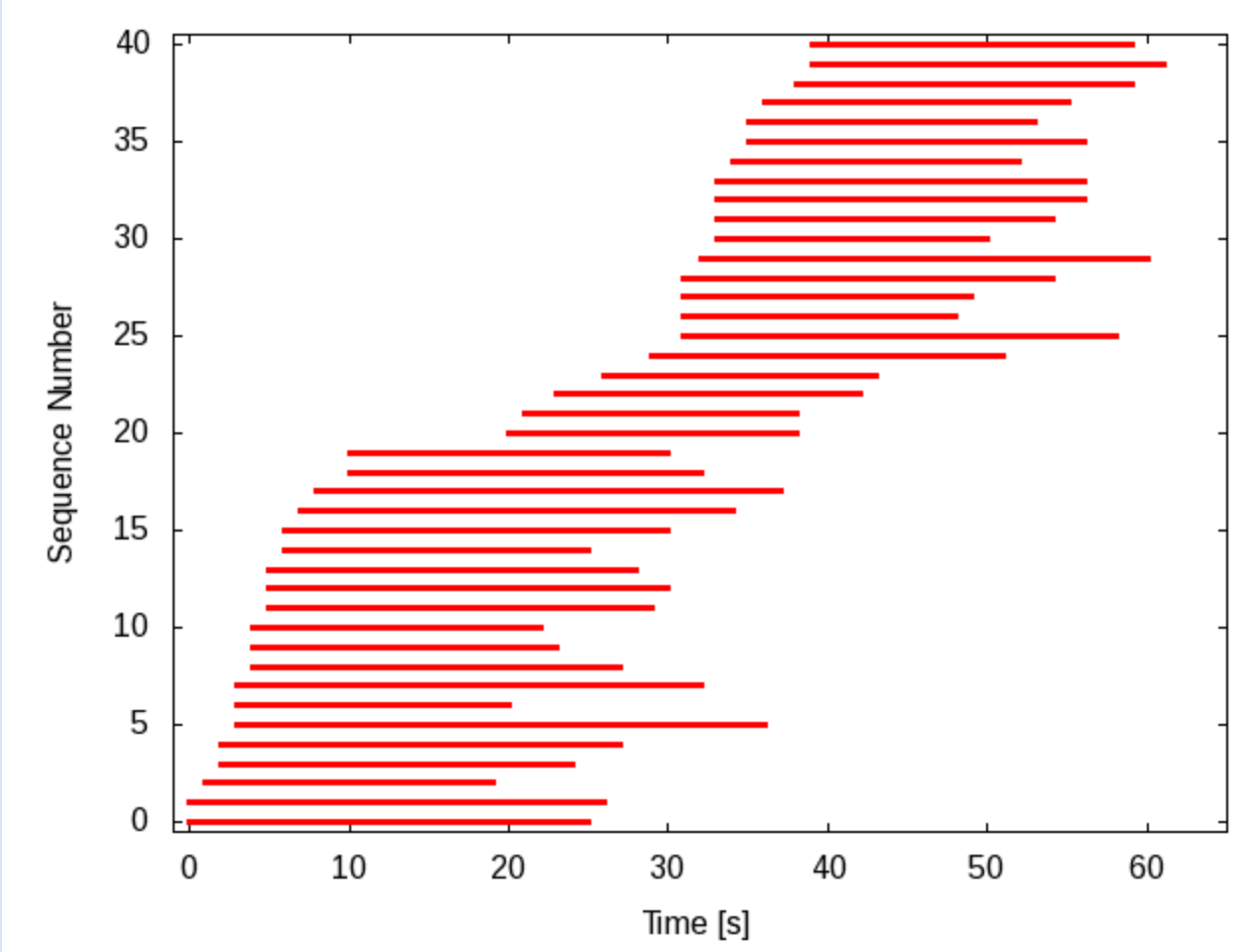
1 Mesabi node

20 Workers

Each worker grabs next work item as soon as it finishes

Variable times:

- Shared PanFS
- Variable loads



Parallel Throughput (0-299)

1 Mesabi node

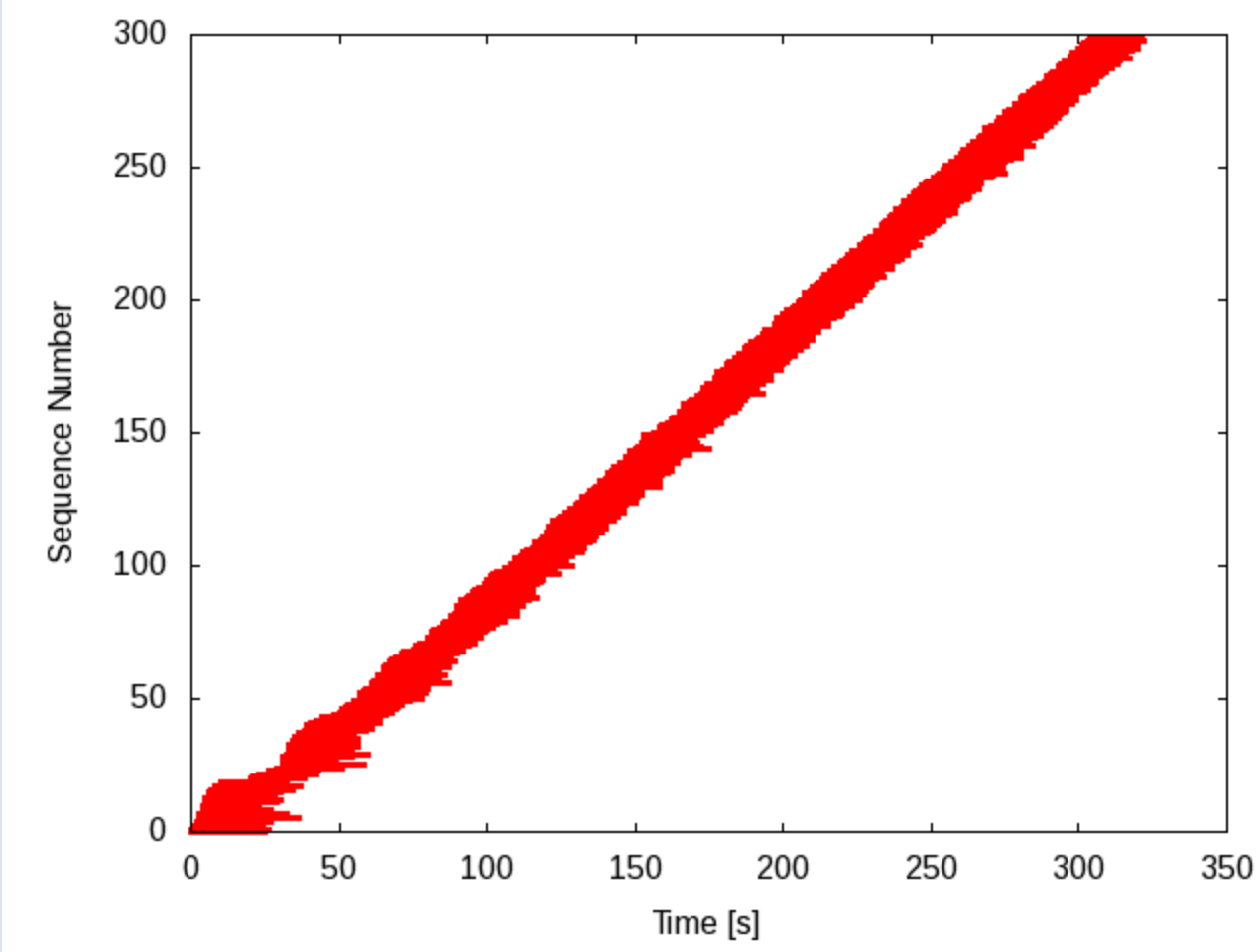
20 Workers

Processed:

- 300 files
- 330 sec.

1 worker:

- 300 files
- ~4800 sec



Process data from CEPH

Workflow with raw data on CEPH

Use s3cmd to pull raw data files

CEPH ⇒ RAM disk

Process on RAM disk then copy results to PanFS

Issue

If not staged on CEPH SSDs, getting 440MB can take ~17s

Overlap copy from CEPH with calculation

1 work item = process 5 consecutive states

work on state i while pulling state i+1

Parallel throughput from CEPH

1 Mesabi node

20 Workers

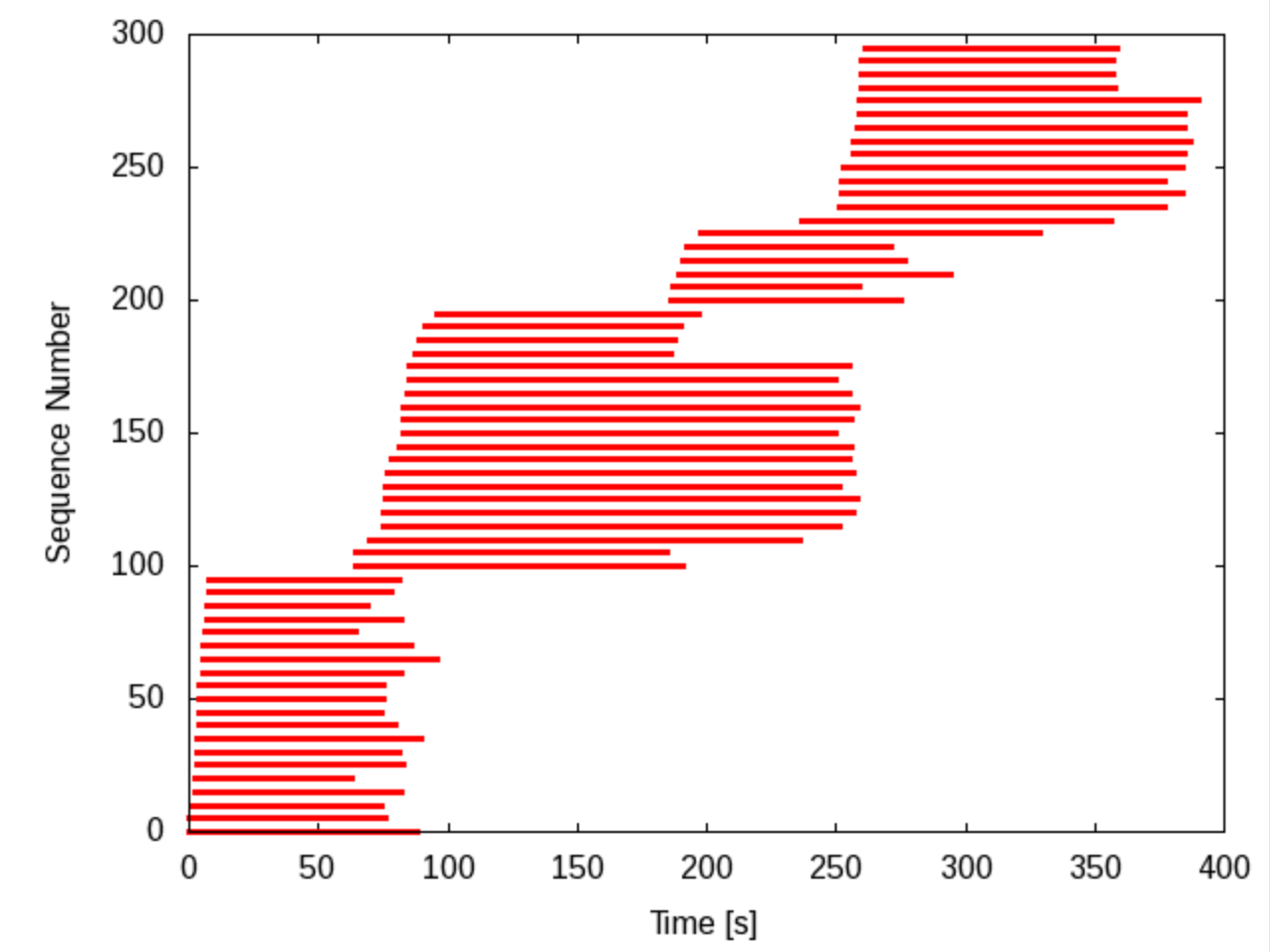
Processed:

300 files

390 sec.

Compare to same data off of PanFS:

330 sec



Thank You

Hands-On

Project lifecycle

- Get & build an application
- Run application, generate data, examine results
- Organize and save data
- Share data
- Clean up

Get Application

Get example from web & unpack

wget www.msi.umn.edu/~porter/tut/cycles.tarz

tar xvfz cycles.tarz

Go into directory and build example application

cd cycles

make

Test application

Run application to get synopsis

./cycles

Should get synopsis: usage: poly <fx> <fy>

App. takes two command line arguments.

These can be integers or floats.

Try an example

./cycles 1 2

You should get 1001 lines: 2 columns of numbers

Run a test case & plot results

Script test1:

./cycles 3 5 > cyc_3_5.dat

gnuplot -persist cyc_3_5.plt

Run it:

./test1

Try your own Parameters

Script test2

./cycles $1 $2 > cycles.dat

gnuplot -persist cycles.plt

Try several examples

./test2 2 3

./test2 13 25

./test2 2 3.02

Parameter space study

Script test3

```
#!/bin/bash
for j in $(seq 1 2 7)
do
for i in $(seq 2 2 8)
do
        ./cycles $i $j > cyc_${i}_${j}.dat
done
done
ls -l cyc*.dat
```

Run it and generate output files (cyc*.dat)

```
./test3
```

Organize & your data

Make an output directory

```
mkdir output
mv *.dat output
```

Make a zipped tar file

```
tar cvfz output.tarz output
```

Share with other members of your group

```
cp -r output ~/../share
chmod -R g=u-w ~/../share/output
```

Save data to CEPH

Make a bucket and save a file

```
module load s3cmd
s3cmd mb s3://mytest
s3cmd put output/cyc_2_1.dat  s3://mytest
```

Save all data file to bucket

```
for i in output/*
      s3cmd put $i  s3://mytest
done
```

or save tar archive

```
s3cmd put output.tarz  s3://mytest
```

Which is faster?

Use data on CEPH

Get a data file tofrom bucket

s3cmd get s3://mytest/cyc_2_3.dat .

Desktop & Web access to CEPH

https://www.msi.umn.edu/support/faq/what-are-some-user-friendly-ways-use-second-tier-storage-s3

Clean up

The situation

Immediate analysis is done.

Data is organized, saved shared and saved (on CEPH)

Assume the data is a large fraction of your group quota

Time to clean up

Fine to save source, scripts, and inputs in you home directory

Better to have them organized where you and your grouop can find it

⇒ Remove the large data files