RESEARCH DATA DEPOT: INFRASTRUCTURE FOR DATA

PURDUE
UNIVERSITY

Preston Smith

Manager of Research Computing Support

9/16/2014

COMMUNITY CLUSTER FACULTY MEETING

TODAY'S AGENDA

- Welcome (Donna Cumberland, Executive Director, Research Computing)
- Introduction (Dr. Gerry McCartney, System CIO)
- Research Data Depot (Preston Smith, Manager of Research Computing Support)
- 2014-2015 Computation Plans (Michael Shuey, HPC Technical Architect)

COMPUTATION

RESEARCH COMPUTING STRENGTH

Since Steele in 2008, Research Computing has deployed many world-class offerings in computation

SIX COMMUNITY CLUSTERS

STEELE

7,216 cores

Installed May 2008

Retired Nov. 2013

COATES

8,032 cores

Installed July 2009

24 departments

61 faculty

Retired Sep. 2014

ROSSMANN

11,088 cores

Installed Sept. 2010

17 departments

37 faculty

HANSEN

9,120 cores

Installed Sept. 2011

13 departments

26 faculty

CARTER

10,368 cores

Installed April 2012

26 departments

60 faculty

#175 on June 2013 Top 500

CONTE

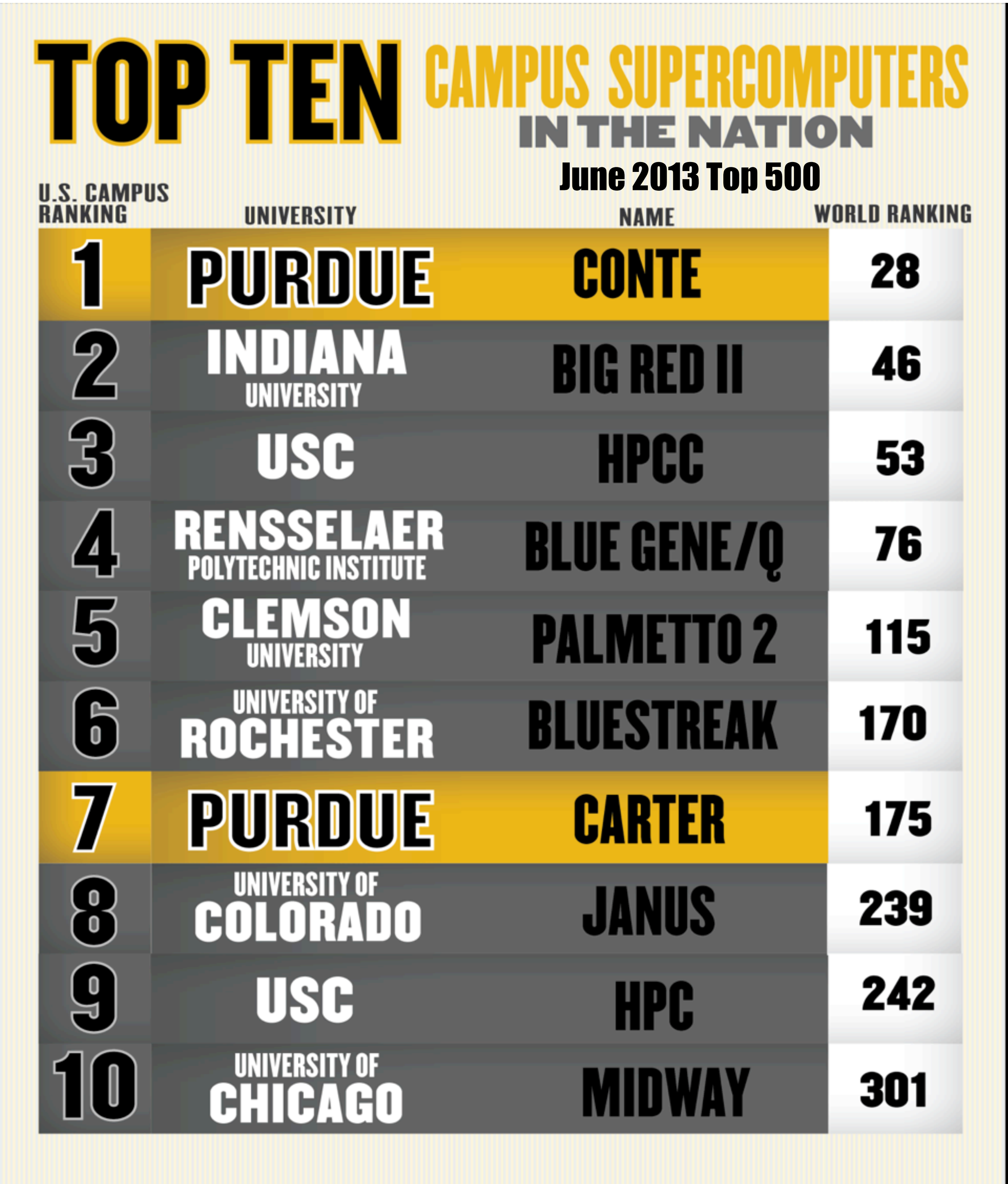
9,280 Xeon cores
(69,600 Xeon Phi cores)

Installed August 2013

20 departments

51 faculty (as of Aug. 2014)

#39 on June 2014 Top 500



TOP TEN CAMPUS SUPERCOMPUTERS IN THE NATION

June 2013 Top 500

U.S. CAMPUS RANKING	UNIVERSITY	NAME	WORLD RANKING
1	PURDUE	CONTE	28
2	INDIANA UNIVERSITY	BIG RED II	46
3	USC	HPCC	53
4	RENSSELAER POLYTECHNIC INSTITUTE	BLUE GENE/Q	76
5	CLEMSON UNIVERSITY	PALMETTO 2	115
6	UNIVERSITY OF ROCHESTER	BLUESTREAK	170
7	PURDUE	CARTER	175
8	UNIVERSITY OF COLORADO	JANUS	239
9	USC	HPC	242
10	UNIVERSITY OF CHICAGO	MIDWAY	301

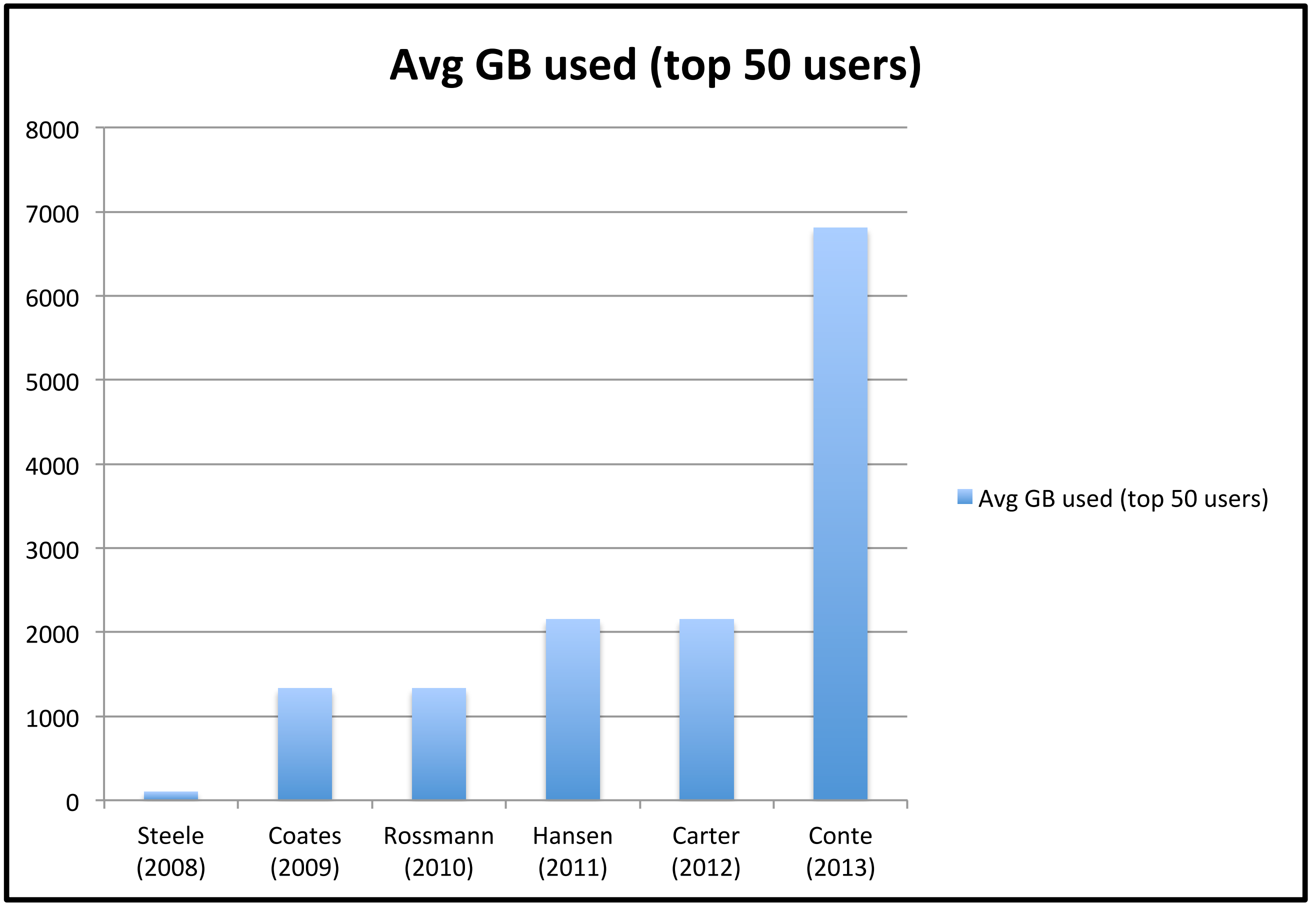
DATA STORAGE

WHAT IS AVAILABLE TODAY FOR HPC

- Research computing has historically provided some storage for research data for HPC users:
  - Archive (Fortress)
  - Actively running jobs (Cluster Scratch - Lustre)
  - Home directories

… And Purdue researchers have PURR to package, publish, and describe research data.

Scratch needs are climbing

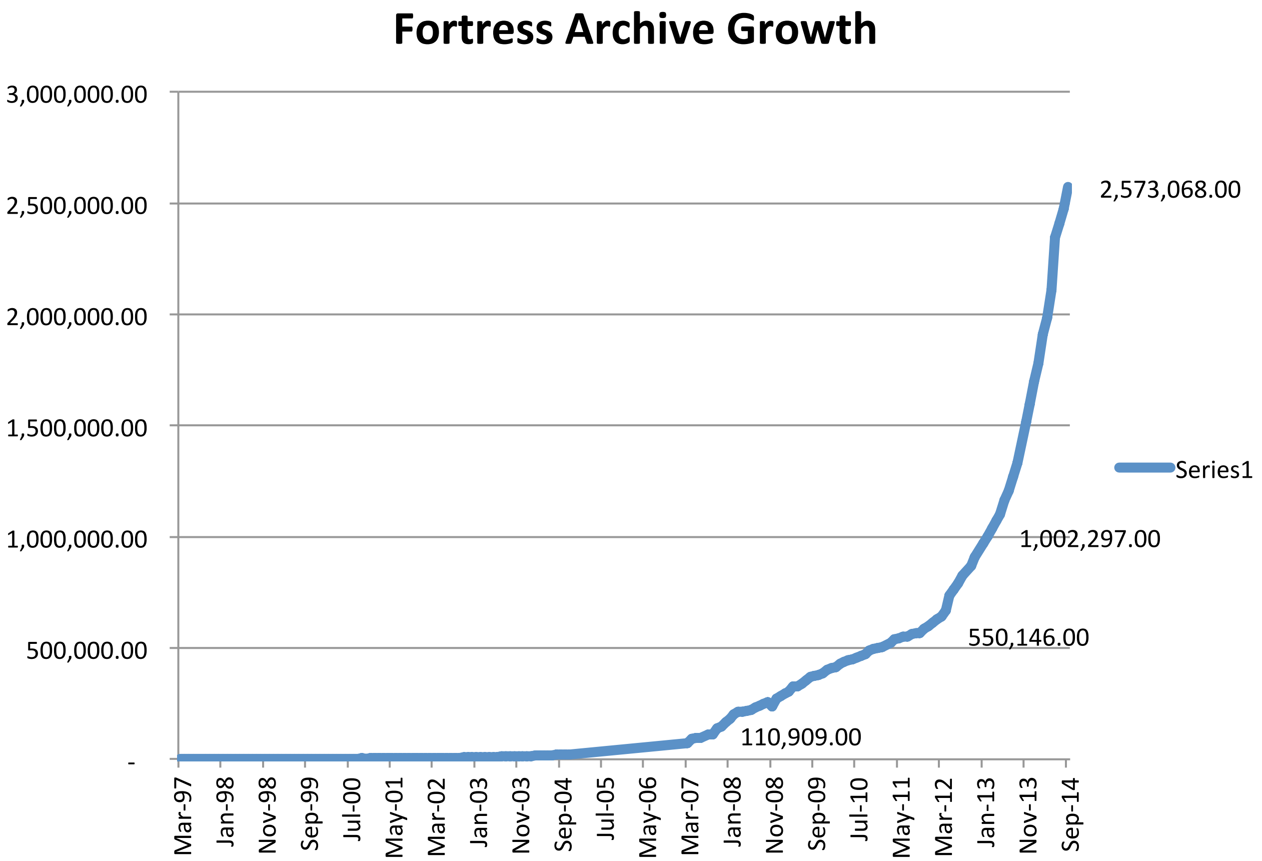


PURDUE UNIVERSITY

HPC DATA

FORTRESS

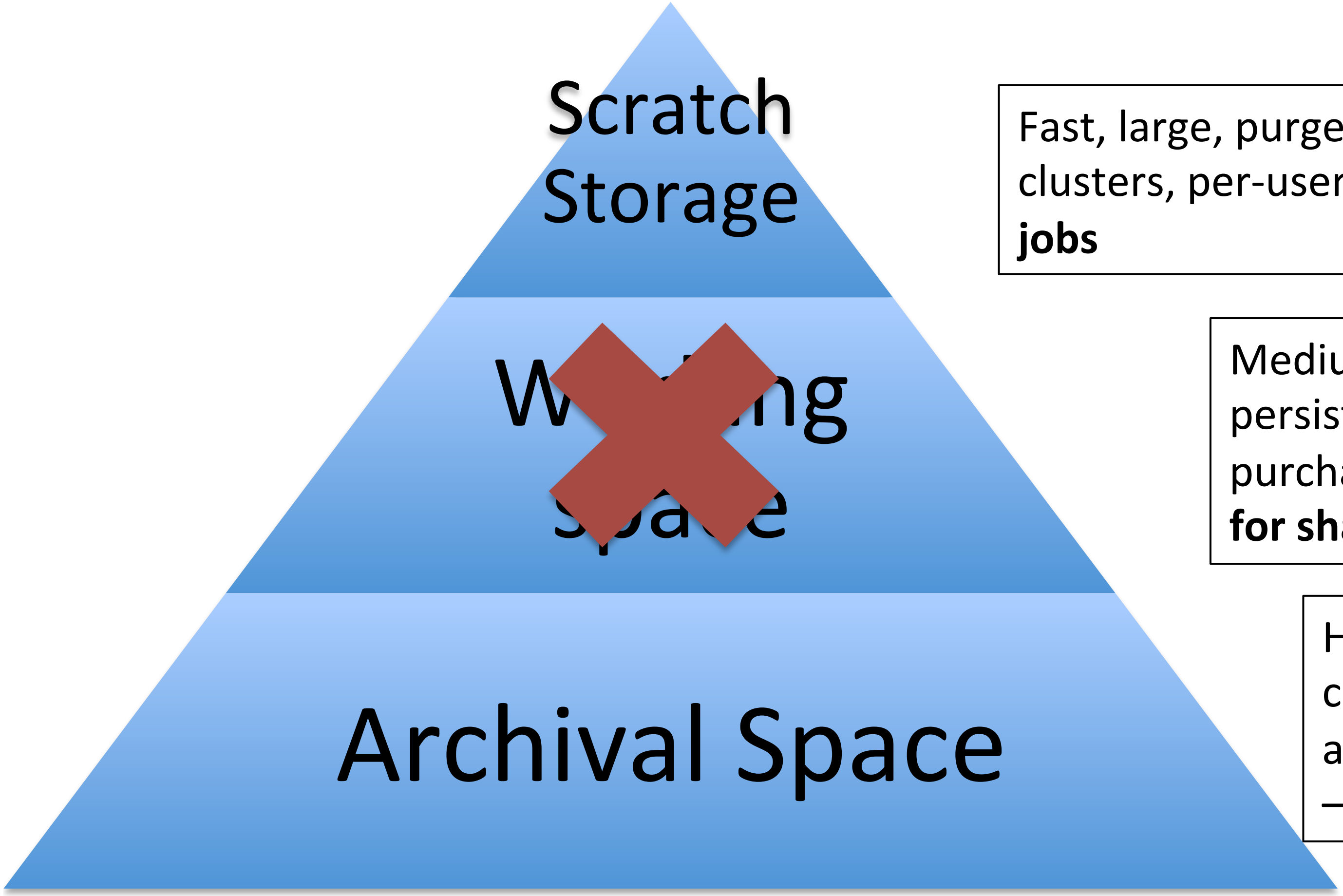
Fortress usage is skyrocketing



PURDUE
UNIVERSITY

HPC STORAGE

TIERS OF STORAGE



Fast, large, purged, coupled with clusters, per-user – **for running jobs**

Medium speed, large, persistent, data protected, purchased, per research lab – **for shared data and apps**

High-speed, infinite capacity, highly-protected, available to all researchers – **for permanent storage**

HPC STORAGE

WHAT IS AVAILABLE TODAY FOR HPC

	$HOME	/group/…	$RCAC_SCRATCH	/tmp	Fortress (HPSS)
Capacity	10-100 GB	500 GB and up	Varies by cluster… 500 GB and up	150-400 GB	unlimited
Resilience to hardware failures	yes	yes	yes	no	yes
Resilience to human errors	yes (snapshots)	yes (snapshots)	no	no	no
Subject to purging	no	no	yes	yes	no
Performance	medium	medium	high	medium to slow (Hansen)	slow to very slow
Designed for HPC (running jobs off it)	no	no	yes	no	-No (as main I/O) -Yes (for staging and archiving)
Common access within cluster	yes	yes	yes	no	yes (hsi/htar)
Common access across clusters	yes	yes	no (except front-ends)	no	yes (hsi/htar)
Advanced ACLs (beyond ugo/rwx)	no	yes	no	no	no

RESEARCH DATA ACROSS CAMPUS

CURRENT STATE

Working with other researchers across campus, we encounter many different data solutions..

From something at the department/workgroup level:

PURDUE
UNIVERSITY

RESEARCH DATA ACROSS CAMPUS

CURRENT STATE

To This

RESEARCH DATA ACROSS CAMPUS

CURRENT STATE

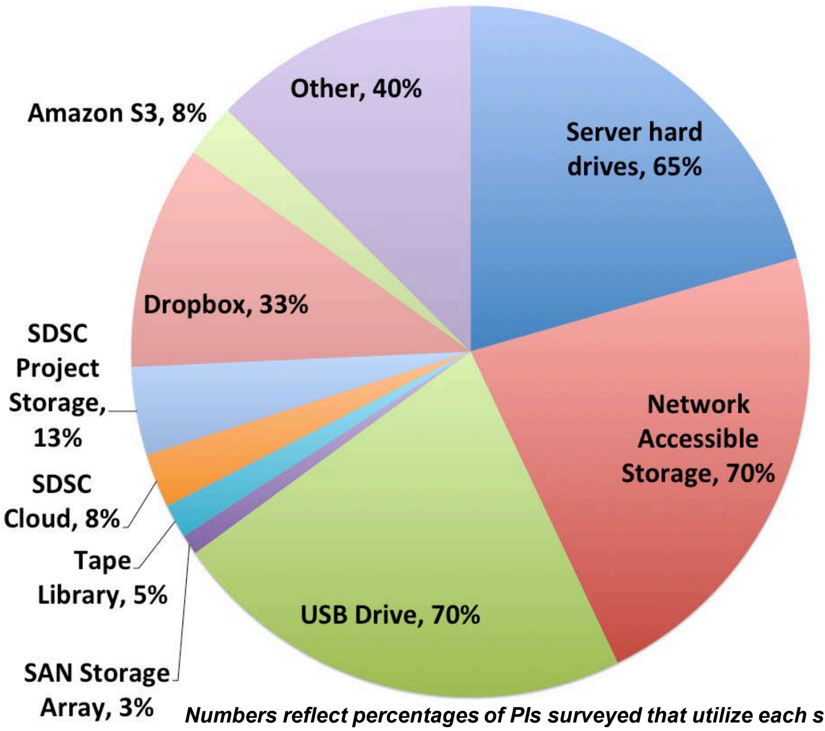
And This

RESEARCH DATA AT OTHER CAMPUSES

NOT JUST PURDUE - SURVEY FROM UCSD

PI Interview Responses: How do You Handle Data Storage/Backup?

Common Data Storage Devices and Services Utilized



Numbers reflect percentages of PIs surveyed that utilize each solution ; Individual PIs use multiple solutions, so %'s add up to >100%.

- **Storage Devices**
  - Network accessible storage (NAS), USB and server local drives dominate
  - Use of Dropbox for sharing
  - Others use Google Drive, Hadoop, XSEDE, SDSC co-location
- **Backup modes**
  - Replicated copies in two NAS
  - A copy in the NAS,
  - A copy in local hard drive (laptop/workstation),
  - And a copy in a USB drive
  - Maybe a copy in email/Dropbox
- **Problems:**
  - Out of sync
  - Lost track of its location
  - Lost version control
  - High cost of recovery

RESEARCH STORAGE GAPS

THE PROBLEM

Many central storage options have not met all the needs that researchers care about

- Departmental or lab resources are islands and not accessible from HPC clusters.
- Most are individually-oriented, rather than built around the notion of a research lab.
    - Boiler Backpack, Fortress, research homes, and scratch
- Scratch filesystems are ***also*** limited in scope to a single HPC system

RESEARCH STORAGE GAPS

COMMON QUESTIONS

Before 2013, we've heard lots of common requests:

- I need more space than I can get in scratch
- Where can I install applications for my entire research lab?
- I'm ***actively working*** on that data/software in scratch:
  - I have to go to great lengths to keep it from being purged.
  - I shouldn't have to pull I from Fortress over and over
- Can I get a UNIX group created for my students and I?
- Is there storage that I can get to on ***all*** the clusters I use?
- I have funding to spend on storage – what do you have to sell?
- I need storage for my instrument to write data into
- My student has the only copy of my research data in his home directory, and he graduated/went off the grid!

RESEARCH STORAGE GAPS

SOME SOLUTIONS

We've addressed some of these with improving scratch:

- Everybody automatically gets access to Fortress for permanent data storage
- Very large per-user quotas, beginning on Conte
- More friendly purge policy – based on the ***use*** of data, rather than creation time.

PERSISTENT GROUP STORAGE

A STORAGE SERVICE FOR HPC USERS

PURDUE
UNIVERSITY

A SOLUTION

FIRST CUT

Since early 2013, HPC researchers could at last purchase storage!

Quickly, a group storage service for research began to materialize to address many common requests:

- 500G available at no additional charge to community cluster groups
- Mounted on all clusters and exported via CIFS to labs
- *Not scratch*: Backed up via snapshots
- Data in /group is owned by faculty member!
- Sharing ability – Globus, CIFS, and WWW
- Version Control repositories
- Maintain group-wide copies of application software or shared data

A TOOL-FRIENDLY STRUCTURE

AS A STARTING POINT

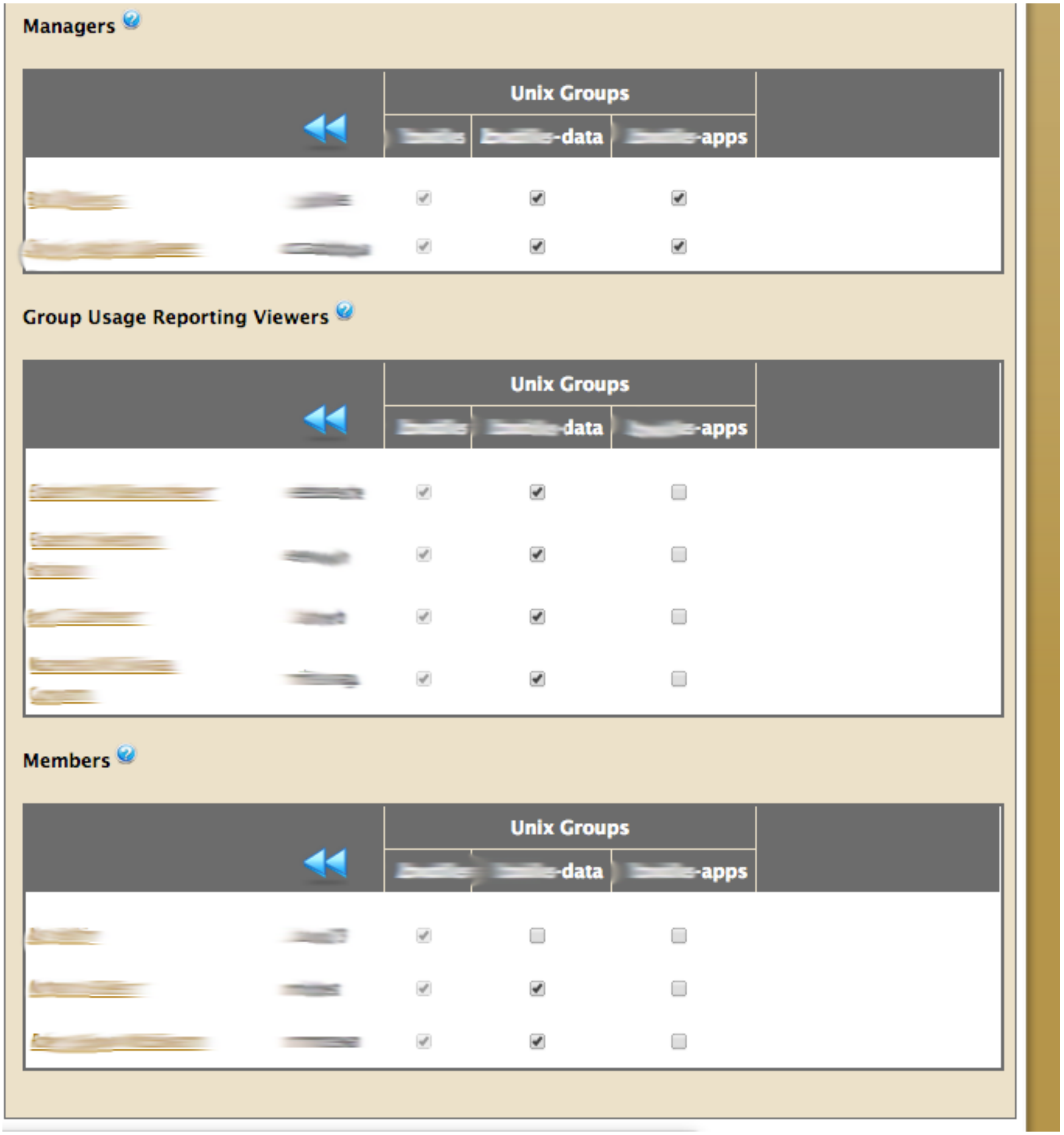
```
/group/mylab/

             +--/apps/
             |
             +--/data/
             |
             +--/etc/
             |         +--bashrc
             |         +--cshrc
             +--/repo/
             |
             +--/www/
             |
             +--/...
```

…with POSIX ACLs to overcome group permission and umask challenges!

SELF-SERVICE

MANAGE YOUR OWN ACCESS



Managers

	◀◀	Unix Groups		
		[redacted]	[redacted]-data	[redacted]-apps
[redacted]	[redacted]	☑	☑	☑
[redacted]	[redacted]	☑	☑	☑

Group Usage Reporting Viewers

	◀◀	Unix Groups		
		[redacted]	[redacted]-data	[redacted]-apps
[redacted]	[redacted]	☑	☑	□
[redacted]	[redacted]	☑	☑	□
[redacted]	[redacted]	☑	☑	□
[redacted]	[redacted]	☑	☑	□

Members

	◀◀	Unix Groups		
		[redacted]	[redacted]-data	[redacted]-apps
[redacted]	[redacted]	☑	□	□
[redacted]	[redacted]	☑	☑	□
[redacted]	[redacted]	☑	☑	□

A SOLUTION

ADOPTION

Fairly well received!

- In just over one year, over 65 research groups are participating.
  - ***Several are not HPC users!***
- Over .5 PB provisioned to date
- A research group purchasing space has purchased, on average, 8.6TB.

It’s not perfect

- ***Scalability and I/O latency are issues on clusters***
- Research computing staff have to monitor “appropriate” work
  - An I/O intensive job in the wrong directory hurts the entire campus!
- Creation and management is not automated
- Protected from “oops” – but **not** disaster-protected

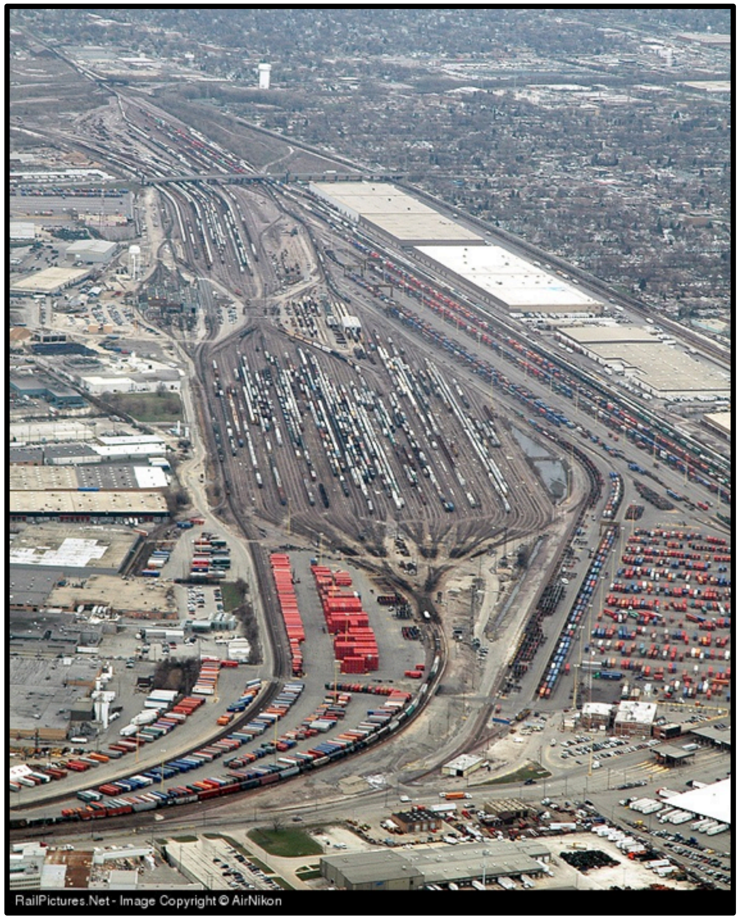
THE RESEARCH DATA DEPOT

INFRASTRUCTURE FOR RESEARCH DATA

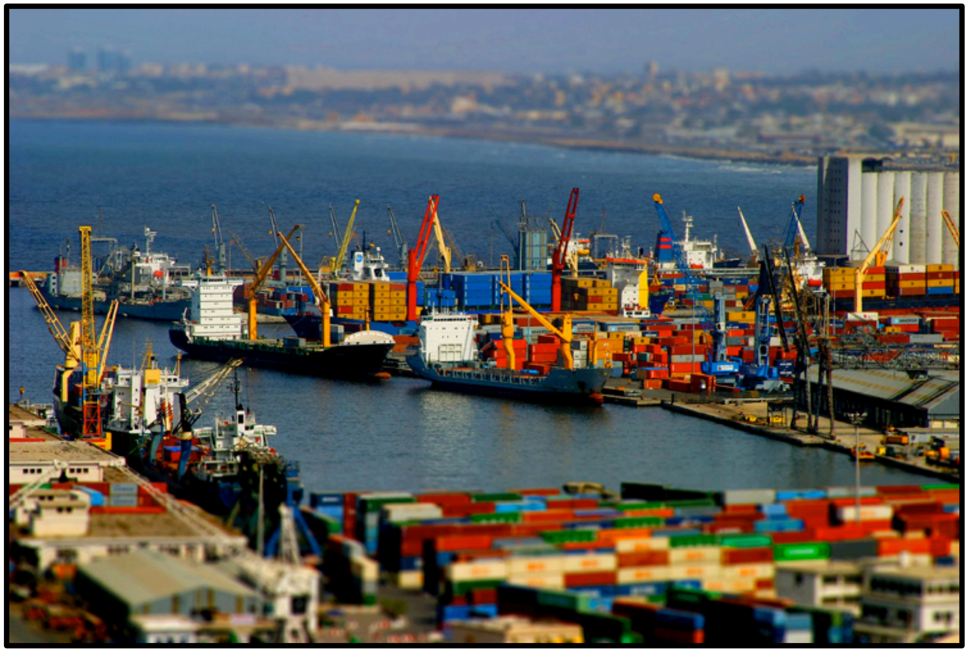
PURDUE
UNIVERSITY

DEPOT

WHY A DEPOT?


RailPictures.Net - Image Copyright © AirNikon

As a transport hub: a place where large amounts of cargo are stored, loaded, unloaded, and moved from place to place.



PURDUE UNIVERSITY

VS FORTRESS

PERMANENCE

A heavily protected and impenetrable building *(Oxford)*

DESIGN GOALS

WHAT DO WE NEED TO DO?

HPC Faculty were surveyed, we learned that they require from their storage:

- Protection from accidents (snapshots)
- Protection from disaster (replicas)
- Affordable prices

We also know that we need to provide:

- A high-performance resource for research instruments
- A business model to sell storage, if faculty wish to purchase it
- Easy ways to share data with collaborators
- Easier ways to access Fortress
- A resource that works for both HPC ***and*** non-HPC users

RFP REQUIREMENTS

WHAT DO WE NEED TO DO?

Proposals Requested from Vendors that can deliver:

Depot Requirements	What we have today
At least 1 PB usable capacity	>1 PB
40 GB/sec throughput	**5 GB/sec**
< 3ms average latency, < 20 ms maximum latency	**Variable**
100k IOPS sustained	**55k**
300 MB/sec min client speed	**200 MB/sec max**
Support 3000 simultaneous clients	Yes
Filesystem snapshots	Yes
Multi-site replication	**No**
Expandable to 10 PB	Yes
Fully POSIX compliant, including parallel I/O	**No**

SOLUTION

WHAT DID WE GET?

Approximately 2.25 PB of IBM GPFS

Hardware provided by a pair of Data Direct Networks SFA12k arrays, one in each of MATH and FREH datacenters

160 Gb/sec to each datacenter

5x Dell R620 servers in each datacenter

PURDUE
UNIVERSITY

SERVICE DETAILS

WHAT WILL IT DO?

At $150/TB per year:

- Same base capabilities as “group storage”, **PLUS**
  - Snapshots
  - Multi-site copies of your data
  - A scalable, expandable storage resource optimized for HPC
- Access to Globus data transfer service, and endpoint sharing

RELATED CAPABILITIES

EVERYTHING IS ABOUT DATA!

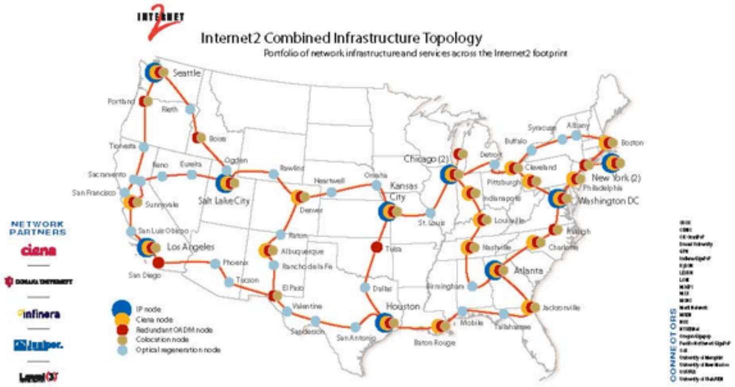
PURDUE
UNIVERSITY

IMPROVED NETWORKING

DEVELOPMENTS IN CAMPUS NETWORK

2014 network improvements

- 100 Gb/sec WAN connections
- Research Core
  - 160 Gb/sec core to each resource (up from 40)
  - 20 Gb/sec research core to most of campus
- Campus Core Upgrade



https://www.rcac.purdue.edu/news/681

PURDUE
UNIVERSITY

FILE TRANSFER AND SHARING

HOW CAN I MOVE AND SHARE LARGE SETS OF DATA?

Globus

- Move data between Purdue, other institutions, and national HPC resources
- Easily share data with collaborators around the world
  - With dropbox-like simplicity!

https://transfer.rcac.purdue.edu

BIG DATA

HADOOP AND FRIENDS

- Use the new “hathi” Hadoop cluster for prototyping big data applications
- Spark, Hbase, Hive, Pig
- On-demand *MyHadoop* clusters on your group’s Conte or Carter nodes

https://www.rcac.purdue.edu/compute/hathi/

more awesome pictures at THEMETAPICTURE.COM

CLOUD SERVICES NEAR YOUR DATA

TOO COSTLY TO MOVE YOUR DATA?

Firebox virtual servers

- Host LAMP servers, cluster login nodes, submission portals, nontraditional HPC, or interactive desktops, all within the research infrastructure
- Move your computing environment close to your research data!

https://www.rcac.purdue.edu/services/firebox/

PURDUE UNIVERSITY

GROUP STORAGE IN FORTRESS

NEED GROUP-ORIENTED ARCHIVE SPACE?

Group storage in Fortress

- Your same self-managed group for the Depot can manage access to a group fortress space
- Contact RCAC for assistance with setting this up

PURDUE
UNIVERSITY

PURDUE
UNIVERSITY

THE RESEARCH DATA DEPOT

WHAT IS NEXT?

WHAT IS IT NOT

IT DOES A LOT, BUT IS NOT EVERYTHING

The Research Data Depot is NOT:

- A 1:1 replacement for Dropbox
- Blessed for HIPAA, FISMA, ITAR, or other regulated data
- Global, paid-for, or non-purged scratch
  - Each resource will continue to have its own local, dedicated, high-performance scratch
  - Its *is* designed to not ***actively hurt*** an HPC user or others should it be used for running jobs
- A data publishing service

FUTURE DIRECTIONS

WHAT ELSE WILL WE BE ABLE TO DO WITH IT?

Potential future questions

- How can we more tightly integrate the Depot and Fortress?
- How can we bridge the space between the Depot at PURR?
- Can your Depot-using labs and instruments benefit from the campus' improved networking infrastructure?

DATA TRAINING

WHAT IS AVAILABLE?

Recent or in-development information sessions on working with data

- Sept 2: “Big Data” training session
- Globus
- Effective use of HPC storage
- HPC I/O formats: HDF5
- In collaboration with Purdue Libraries – the data life cycle and managing your research data on RCAC resources

LIBRARY DATA SERVICES

HOW CAN I MANAGE ALL MY DATA?

- Collaborations on multi-disciplinary grant proposals, both internal and external
- Developing customized Data Management Plans
- Organizing your data
- Describing your data
- Sharing your data
- Publishing your datasets
- Preserving your data
- Education on data management best practices

PURDUE
UNIVERSITY

LIBRARY DATA SERVICES

PURR

PURR is a *free* online research data collaboration platform and service solution for Purdue faculty, graduates students, and staff.

Research data - spreadsheets, images, output from sensors and instruments, transcripts, surveys, software source code and tools, video, and observation logs

PURR provides:

✓ Data management plan (DMP) resources and consultation

✓ Collaborative research data project space

✓ Dataset publication with Digital Object Identifier (DOI) *

✓ Long-term preservation and management

PURDUE
UNIVERSITY

ACCESS

HOW CAN I GET ACCESS?

The service is in early access mode.

Contact us at rcac-help@purdue.edu if you're willing to be an early access tester!

Full production on October 31

/GROUP USERS

I USE /GROUP TODAY, HOW DO I TRANSITION?

Persistent group storage owners will be transitioned into the Depot at an agreed-upon date, on a group-by-group basis.

We’ll be in touch to arrange a time!

If you’re paid up with the persistent group storage, this move will come at no additional charge until your renewal is due.

ACCESS

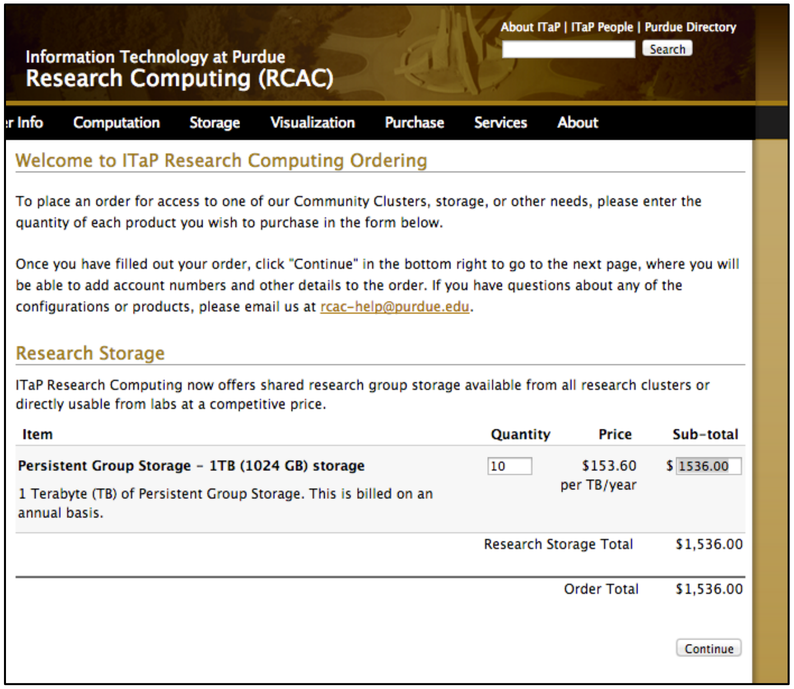
HOW CAN I GET ACCESS?

We can set your lab up with a small evaluation space upon request

Need 1TB or more?

purchase access online

- http://www.rcac.purdue.edu/order


About ITaP | ITaP People | Purdue Directory
Search
Information Technology at Purdue
Research Computing (RCAC)
r Info
Computation
Storage
Visualization
Purchase
Services
About
Continue

Welcome to ITaP Research Computing Ordering

To place an order for access to one of our Community Clusters, storage, or other needs, please enter the quantity of each product you wish to purchase in the form below.

Once you have filled out your order, click "Continue" in the bottom right to go to the next page, where you will be able to add account numbers and other details to the order. If you have questions about any of the configurations or products, please email us at rcac-help@purdue.edu.

Research Storage

ITaP Research Computing now offers shared research group storage available from all research clusters or directly usable from labs at a competitive price.

Item	Quantity	Price	Sub-total
Persistent Group Storage - 1TB (1024 GB) storage 1 Terabyte (TB) of Persistent Group Storage. This is billed on an annual basis.	10	$153.60 per TB/year	$ 1536.00
		Research Storage Total	$1,536.00
		Order Total	$1,536.00

CLUSTER COMPUTING

WHAT IS NEXT?

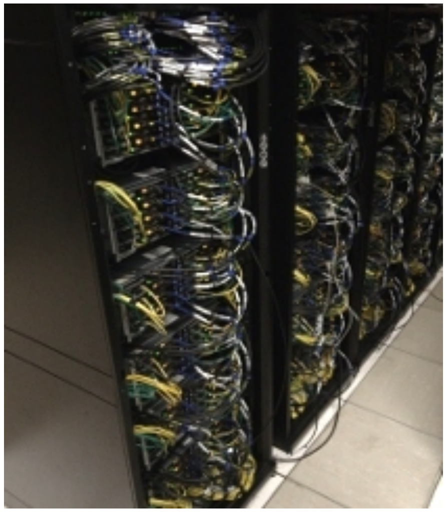
PURDUE
UNIVERSITY

TODAY

CARTER & CONTE

Carter & Conte still very current

- Ivy Bridge chips not much improvement over Carter
- Conte’s Xeon Phi co-processors extremely capable for certain applications (dense FP simulation)
- Both Carter and Conte have proven very popular!



TOMORROW

INTEL'S HASWELL PROCESSOR

Released last week:

- More cores/chip
- Slight clock speed reduction
- Additional floating-point instructions
  - Can greatly assist matrix multiplication operations
- DDR4 memory
  - Higher memory bandwidth helps just about everyone

Longer-term implications:

- Forget 2/4/8 GB/core – think about base memory/node
- Best programming model may be MPI+OpenMPI
  - Fewer MPI ranks/node, pick up extra cores/threads with OpenMP
  - Looks vaguely like Xeon Phi…

NEXT CLUSTER

ROUGH TIMELINE

Planning for next community cluster starting now

- Bringing Haswell-based loaner to Purdue
- Preparing for open bid later this fall
- Cluster assembly, go-live in mid-Spring, 2015

Open questions:

- Higher clock speed or more cores/node?
- Memory sizes?
    - Probably 64 GB base with an option for large (256G or more) nodes

THE END

QUESTIONS?

Questions?