OURRstore: The OU & Regional Research Store

Henry Neeman, University of Oklahoma
Director, OU Supercomputing Center for Education & Research (OSCER)
Associate Professor, Gallogly College of Engineering
Adjunct Associate Professor, School of Computer Science

CADRE 2020, Friday April 17 2020

Short Version, Long Version

- This talk will be the short version of the OURRstore talk.
- The long version's slides are in this slide deck,
  after the short version's.
- A recording of the long version of this talk can be found at:

https://www.youtube.com/watch?v=bF1MPhdFag0

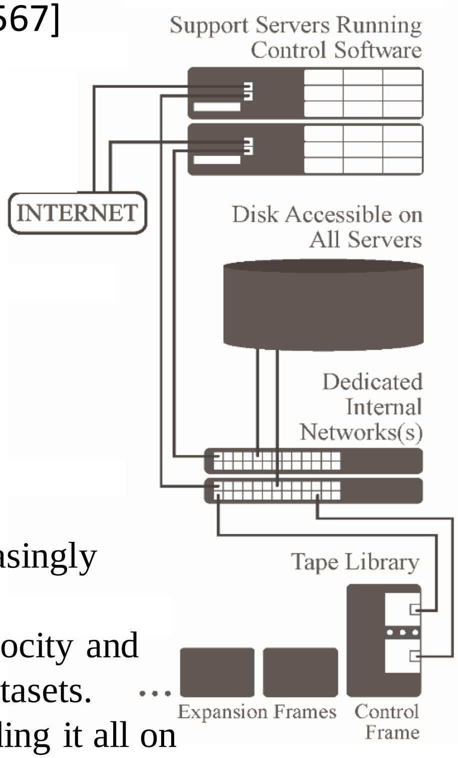
THE SHORT VERSION

MRI: Acquisition of a Regional Resource for Long-Term Archiving of Large Scale Research Data Collections

PI Henry Neeman, University of Oklahoma [Award #1828567]

OU & Regional Research Store ("OURRstore")

- Large scale (many PB), long term (8+ year), multi-copy, multi-institution
- Hundreds of students, staff and faculty can:
  - Build large and growing data collections
  - Share and publish datasets, making them discoverable and searchable
  - EPSCoR states and territories (and others) nationwide



Support Servers Running Control Software
INTERNET
Disk Accessible on All Servers
Dedicated Internal Networks(s)
Tape Library
Expansion Frames
Control Frame

Intellectual Merit:

- Science, Technology, Engineering and Mathematics (STEM) research is increasingly data-intensive, with massive growth of research data collections.
- Many universities and colleges are underprepared not only for the volume, velocity and variety of data, but especially for long term stewardship of rapidly growing datasets.
- Many STEM research projects need big storage during their experiments: holding it all on disk is too expensive.

Broader Impacts:

- A national model for affordable, large scale, long term, resilient, multi-institutional storage.
- 200+ users, advisors etc are women, African Americans, Hispanics, Native Americans, disabled, and/or US veterans.
- OU's "Supercomputing in Plain English" webinars, which in 2018 had 1200+ registrants in every US state and worldwide, includes a section on the storage hierarchy.
- Research librarians facilitate data management best practices, and hold a monthly peer mentoring call.

Outline

- The Challenge of Physical Data Management
- A Business Model for Physical Management of Big Data
- OURRstore Features
- Acknowledgements

The Challenge of Physical Data Management

Large Data Volume Choices

I’ve got tens of TB of data (or hundreds of TB or PB or …).

Why can’t I just buy a bunch of USB drives
at my local big box store (or online)?

Large Data Volume Choices

You can enter a NASCAR race on a riding lawnmower, but:

- you probably won’t win;
- you probably will get killed.

http://express.howstuffworks.com/gif/exp-nascar-2.jpg

http://uslmra.org/wp-content/uploads/2009/09/HowardLawnMowerRacing.jpg

In a Nutshell ….

=

http://www.sportsbet.com.au/blog/wp-content/uploads/money1.jpg

http://gadgetsin.com/uploads/2010/02/samsung_ecogreen_f3eg_hard_disk_drive_1.jpg

What Does a PetaByte Cost?

1 PB = 1000 TB = 1M GB = a quadrillion (10^{15}) bytes

Example: 1 PB of content for 8 years, at least dual copies:

- IBM V5010E: 2 copies, rebuy @ yr 6 = ~$920,000 at current pricing: **38×**
- 45drives: 2 copies, rebuy @ yr 6 = ~$254,000 at current pricing: **11×**
- Dell MD: 2 copies, rebuy @ yr 6 = ~$234,000 at current pricing: **10×**
- Google: 2+ copies = ~$102,000 at current pricing: **4.3×**
- Amazon: 2+ copies = ~$84,000 at current pricing: **3.6×**
- Microsoft: 2+ copies = ~$84,000 at current pricing: **3.6×**
- USB disks: 2 copies, rebuy @ yr 6 = ~$60,000 at current pricing: **2.5×**
- OURRstore: 2 copies = ~$23,600 at current pricing: **base**

  (assumes dual copies for all non-cloud solutions; some include IDC; equipment gets bought again at the start of year 6)

  - Explanation shortly

What About Tape’s Fixed Costs?

- Tape has **huge fixed costs**: In OURRstore, ~80% of total equipment costs are tape/disk/switch/software/media components that are needed regardless of the number of tape cartridge slots (this takes into account 8 years of warranty/support/maintenance):
  - tape library base frame;
  - tape library drive expansion frame;
  - tape drives;
  - servers;
  - disk;
  - network switches;
  - software.
- Only ~20% of the OURRstore budget is for tape cartridge slots (~11,000 of them).

A Business Model for Physical Management of Big Data

Business Model

First PetaStore (2010-20), then OURRstore (2020-28)

- **Grant**: hardware, software, initial warranties on everything
- **Institution (CIO)**: space, power, cooling, labor, maintenance after the initial warranty period
- **Researchers**: media (tape cartridges)

- Compared to roll-your-own disk, for our researchers OURRstore tape is:
  - cheaper per TB @ dual copies;
  - more reliable (bit rot rate is 10% to 0.1% as high as disk);
  - less labor (treat it as a filesystem);
  - requires less expertise to set up and use (~1 hour training);
  - slower (moderate bandwidth, very high latency: 1 minute, not 10 ms).

OURRstore Technology Strategy

- Distribute the costs among a research funding agency, the institution, and the research teams.
- Archive, not live storage: “Write once, read seldom if ever.”
- Independent, standalone system: not part of a cluster.
- Spend grant funds on many slots but few tape cartridges.
- Media slots are available on a first come first serve basis.*
- Software cost should be a modest fraction of total cost.
- Maximize media longevity.
- Globus for file transfers, file sharing, ~~file publishing~~, discoverability etc. (Institutional repositories for file publishing.)
- LTFS (tiny file catalog on each tape cartridge):
  Ship secondary copies to the data owner – if anything goes wrong, it’s under $3K to buy an LTO tape drive, and the software is free.

OU’s NSF MRI Archive Grants

- **PetaStore**: “MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research,” OCI-1039829, $792,925, 10/1/2010-9/30/2013 (+ 1 year no cost extension)
  - That grant was 3 years, took a 1 year no cost extension;
its archive was meant to be 6 years
but is currently in the first half of year 9 of full production.
  - 12 research teams on the proposal, 55 now.
- **OURRstore**: “MRI: Acquisition of a Regional Resource for Long-term Archiving of Large Scale Research Data Collections,” OAC-1828567, $967,755, 9/1/2018 - 8/31/2021
  - The grant is 3 years; its archive will be for 8+ years.
  - 87 research teams on the proposal; a few have dropped, many will join.

Who’s Eligible? Who’s In?

- Institutions in Great Plains Network (GPN) states. (AR,KS,MO,NE,OK,SD)
- Institutions in the 28 EPSCoR jurisdictions.
  (AK, AL, AR, DE, GU, HI, IA, ID, KS, KY, LA, ME, MS, MT, ND, NH, NE, NM, NV, OK, PR, RI, SC, SD, VI, VT, WV, WY)
- Institutions (and consortia) in non-EPSCoR jurisdictions, if they buy their own cartridge expansion cabinet (limited space!).
- In the proposal, 87 research teams at 27 institutions in 17 states, including 29 research teams at OU. (A few have dropped, many will be added.)
- <u>16 teams will each need at least 1 PB</u>:
  8 at OU, 1 in another GPN state, 7 in non-GPN EPSCoR states.
  By contrast, the original PetaStore proposal included only 12 teams *total*, regardless of capacity need.

How Much Need?

Per the proposal:

- Capacity needed: 134 PB (~17,500 LTO-7 Type M cartridges)
  - $31M in on-premise spinning disk RAID, OR
  - $11M in a commercial cloud, OR
  - $8M in USB disk drives (Good luck managing that!), OR
  - $3M in tape cartridges
    - If we bought the full 134 PB today.
- Active funding of these projects: $162M
- Pending/planned funding: $140M
- Faculty: 250+
- Staff: 150+
- Postdocs: 100+
- Graduate students: 500+
- Undergraduate students: 500+

Yeah, But Tape Sucks!

- Well, yes, tape does suck:
  - Retrieval has very high latency (typically 1 minute per file, vs ~10 ms per file for spinning disk).
  - Tape medium inside a tape cartridge can break -- tinsel!
- How to resolve?
  - Only store large files (OURRstore minimum is 1 GB).
    - So, you have to create Zip files or compressed tar files.
  - Offline storage: Download file to disk before using.
  - Think hierarchically:
    - small amount of very fast disk;
    - medium amount of slower disk;
    - large amount of tape.

OURRstore Features

OURRstore New Features

- Disk Caching
- Globus
- LTFS
- Research Data Management
- Archive Longevity Approach

Disk Caching

- Files that have been requested recently will stay on disk as long as possible, until disk capacity needs to be cleared out to make room for other files.
- We can do this because we’re being very careful stewards of both our NSF grant funds and our CIO’s cost share funds, and we found a way to get:
    - much much lower disk cost than in the winning bid response;
    - much more disk capacity than requested in the RFP (~600 TB vs 100-200 TB);
    - faster than requested in the RFP (~2.5 GB/sec vs 1-2 GB/sec).

Globus

- Globus is a software product that has a long history.
  - Funded in the late 1990s/early 2000s by the NSF.
  - Became a web-based large file transfer service in the early 2010s.
  - Now an independent, standalone software system (still uses web).
  - The software features below would cost most of our institutions ~$20K per year – but on OURRstore, only the grant pays (and then OU’s CIO, starting in year 5, but at a much lower rate).
- File sharing: A file owner can decide, for any file or directory, who can access it, not only among OURRstore users, but also outsiders who want to download the data.
  - The incremental download cost is zero.
- ~~File Publishing~~: No longer a supported Globus feature.

LTFS

- Starting with LTO-5, every LTO tape cartridge has a little chunk of capacity that's set aside for a file catalog that describes every file on that tape cartridge.
- For PetaStore, we didn't use LTFS, because the software we could afford didn't understand LTFS; for OURRstore, we will.
- So, for second copies, we'll require a prepaid shipping label, and we'll export the secondary copy and ship it to the file owner.
    - The reviewers insisted we require at least dual copies for all files, on the fear that, if we offered single copy, that's all anyone would use, and OURRstore would hold the sole copy of many datasets.
- Because of shipping you the secondary copy of your files, if something goes wrong at OU, you can still access your files, for less than $3000 for a tape drive, with free software.

Research Data Management

- Monthly research data management conference calls will start once OURRstore is in production, led by Dr. Mark Laufersweiler, OU Libraries’ Research Data Specialist.
- We need to identify and recruit research data librarians at many OURRstore institutions:
  - Clemson U (SC)
  - Idaho State U
  - New Mexico State U
  - North Dakota State U
  - U Louisiana Lafayette
  - U Montana
  - U New Hampshire
  - U New Mexico
  - U North Dakota

Acknowledgements: People

- NSF MRI OURRstore Leadership
  - Co-PIs Laura Bartley, Kendra Dresback, Amy McGovern, Horst Severini
  - Senior Personnel Mark Laufersweiler
- NSF MRI OURRstore Advisory Committees
- NSF MRI OURRstore Research Data Librarians
- NSF MRI OURRstore users
- OSCER Team: David Akin, Patrick Calhoun, Jim Ferguson, Kali McLennan, Horst Severini, Jason Speckman
  - Recent Former: Debi Gentis, George Louthan, Ashish Pai, John Shelton, Matt Younkins
- OU CIO David Horton
- OU Interim Dean of University Libraries Carl Grant

Acknowledgements: Grants

Portions of this material are based upon work supported by the National Science Foundation under the following grants:

- Grant no. OAC-1828567, “MRI: Acquisition of a Regional Resource for Long-term Archiving of Large Scale Research Data Collections,” PI H. Neeman, $968K, 2018-21
- Grant no. ACI-1440783, “CC*IIE Engineer: A Model for Advanced Cyberinfrastructure Research and Education Facilitators,” PI H. Neeman, $400K, 2014-17
- Grant no. ACI-1341028 , “CC-NIE Integration: OneOklahoma Friction Free Network,” PI H. Neeman, $500K, 2013-17
- Grant no. OCI-1039829, “MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research,” PI H. Neeman, $792K, 2010-14
- Grant no. EPS-1006919, “Oklahoma Optical Initiative,” PI H. Neeman, $1.17M, 2010-13
- Grant no. OAC-1531128, “MRI: Acquisition of Shared High Performance Compute Cluster for Multidisciplinary Computational and Data-Intensive Research,” PI D. Brunson, OSU, $952K, 2015-18
- Grant no. ACI-1126330, “MRI: Acquisition of a High Performance Compute Cluster for Multidisciplinary Research,” PI D. Brunson, OSU, $909K, 2011-15
- Grant no. ACI-1229107, MRI: Acquisition of a High Performance Computing Cluster for Research and Education,” PI J. Snow, Langston U, $250K, 2013-16
- Grant no. ACI-1440774, “CC*IIE Region: Leveraging Partnerships Across the Great Plains to Build Advanced Networking and CI Expertise,” PI G. Monaco, Co-PI J. Deaton, $130K, 2014-17
- Grant no. ACI-1429702, “MRI: Acquisition of a High Performance Computing Cluster for Research at a Predominantly Undergraduate Institution,” PI. E. Lemley, UCO, $305K, 2015-17
- Grant no. ACI-531270, “MRI: Development of Heterogeneous Cluster for Cyber-Physical System Hybrid Analytics,” PI J. Hale, TU, $181K, 2015-17
- Grant no. 1659235, “CC* Network Design: Multiple Organization Regional One Oklahoma Friction Free Network (MORe OFFN), PI R. Reif, NSU, $334K

Bibliography

1. S. P. Calhoun, D. Akin, B. Zimmerman and H. Neeman, 2016: “Large Scale Research Data Archiving: Training for an Inconvenient Technology.” *Journal of Computational Science*. DOI: 10.1016/j.jocs.2018.07.005.
2. H. Neeman, K. Adams, J. Alexander, D. Brunson, S. P. Calhoun, J. Deaton, F. Fondjo Fotou, K. Frinkle, Z. Gray, E. Lemley, G. Louthan, G. Monaco, M. Morris, J. Snow and B. Zimmerman, 2015: “On Fostering a Culture of Research Cyberinfrastructure Grant Proposals within a Community of Service Providers in an EPSCoR State.” *Proc. XSEDE’15*, article 19. DOI: 10.1145/2792745.2792764.
3. H. Neeman, D. Akin, J. Alexander, D. Brunson, S. P. Calhoun, J. Deaton, F. Fondjo Fotou, B. George, D. Gentis, Z. Gray, E. Huebsch, G. Louthan, M. Runion, J. Snow and B. Zimmerman, 2014: “The OneOklahoma Friction Free Network: Towards a Multi-Institutional Science DMZ in an EPSCoR State.” *Proc. XSEDE’14*, article 49. DOI: 10.1145/2616498.2616542.
4. S. P. Calhoun, D. Akin, J. Alexander, B. Zimmerman, F. Keller, B. George and H. Neeman, 2014: “The Oklahoma PetaStore: A Business Model for Big Data on a Small Budget.” *Proc. XSEDE’14*, article 48. DOI: 10.1145/2616498.2616548.
5. H. Neeman, D. Brunson, J. Deaton, Z. Gray, E. Huebsch, D. Gentis and D. Horton, 2013: “The Oklahoma Cyberinfrastructure Initiative.” *Proc. XSEDE’13*, article 70. DOI: 10.1145/2484762.2484793.

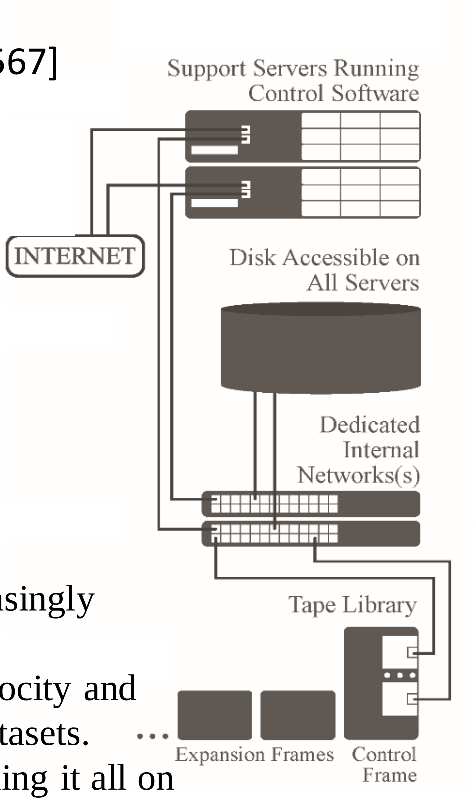
Thanks for your attention!

QUESTIONS?

THE LONG VERSION

MRI: Acquisition of a Regional Resource for Long-Term Archiving of Large Scale Research Data Collections

PI Henry Neeman, University of Oklahoma [Award #1828567]



Support Servers Running Control Software
INTERNET
Disk Accessible on All Servers
Dedicated Internal Networks(s)
Tape Library
Expansion Frames
Control Frame

OU & Regional Research Store (“OURRstore”)

- Large scale (many PB), long term (8+ year), multi-copy, multi-institution
- Hundreds of students, staff and faculty can:
  - Build large and growing data collections
  - Share and publish datasets, making them discoverable and searchable
  - EPSCoR states and territories (and others) nationwide

Intellectual Merit:

- Science, Technology, Engineering and Mathematics (STEM) research is increasingly data-intensive, with massive growth of research data collections.
- Many universities and colleges are underprepared not only for the volume, velocity and variety of data, but especially for long term stewardship of rapidly growing datasets.
- Many STEM research projects need big storage during their experiments: holding it all on disk is too expensive.

Broader Impacts:

- A national model for affordable, large scale, long term, resilient, multi-institutional storage.
- 200+ users, advisors etc are women, African Americans, Hispanics, Native Americans, disabled, and/or US veterans.
- OU's "Supercomputing in Plain English" webinars, which in 2018 had 1200+ registrants in every US state and worldwide, includes a section on the storage hierarchy.
- Research librarians facilitate data management best practices, and hold a monthly peer mentoring call.

Outline

- A Little History
- The Challenge of Physical Data Management
- Researchers in the Wild
- A Business Model for Physical Management of Big Data
- OURRstore Architecture
- OURRstore's New Features
- LTO Weirdness
- Acknowledgements

Please feel free to ask questions at any time. I like interacting.

A Little History

A Little History

- 2015: We submitted a $5M NSF Data Infrastructure Building Blocks (DIBBs) proposal, for a pair of OURRstores, which got just about the worst reviews I've ever gotten in my life! (Poor, Poor, Fair, Fair – it's like getting 2 Fs and 2 Ds = F+ avg!)
- 2016: We submitted an $800K NSF Major Research Instrumentation proposal, which got interesting reviews but wasn't funded. (Excellent/Good, Very Good, Fair, Fair – 2 Bs and 2 Ds = C average!)
- 2017: Another try at the $800K NSF MRI, not funded. (Excellent, Very Good, Good – A, B, C = B average!)
- 2018: $968K NSF MRI, excellent reviews, funded! (3 As, 1 B).

The Challenge of Physical Data Management

Large Data Volume Choices

I’ve got tens of TB of data (or hundreds of TB or PB or …).

Why can’t I just buy a bunch of USB drives
at my local big box store (or online)?

Large Data Volume Choices

You can enter a NASCAR race on a riding lawnmower, but:

- you probably won’t win;
- you probably will get killed.

http://express.howstuffworks.com/gif/exp-nascar-2.jpg

http://uslmra.org/wp-content/uploads/2009/09/HowardLawnMowerRacing.jpg

Why Not Roll-Your-Own?

- For small data collections, roll-your-own is perfectly reasonable:
  - USB disk drives are cheap: https://www.walmart.com/ip/Seagate-8TB-EXPANSION-DESK-STEB8000100/52752803
    - 8 TB USB 3.0 = $141.88 walmart.com 4/7/2020 => ~$20 per usable TB
  - Buy two and copy everything to both drives (getting user compliance on buying and making the secondary copy isn’t always trivial).
    - Remember that you have to pay IDC on equipment under $5000.
- Slightly bigger than that (10s of TB): You can do a small, cheap RAID enclosure for mirroring or RAID6, **BUT**:
  - Price per TB starts going way up:
    - 45drives.com: 60×16TB = ~$42K 4/7/2020 => ~$51 per usable TB
    - Dell MD1400: 12×14TB = ~$10K 4/7/2020 => ~$86 per usable TB
      - 45drives.com = 2.5×USB, Dell = 4.3×USB, for price per usable TB per copy
  - Need much more expertise to configure and manage.
  - Risk is higher because a failed system loses lots of data – or buy two, doubling your costs, and then buy again 5 years later.

Enterprise-Class Disk

Enterprise-class disk

- IBM Storwize V5010E (entry/midrange enterprise product)
  - 196 x 14 TB = ~$355K MSRP = ~$146 per usable TB per copy
- This doesn’t include the servers or switches to attach it to.
- Of course, you’ll need dual copies, so double this.
- And, after 5 years, you’ll need to buy it again (with far fewer drives, because the capacity per drive will be higher).
- Of course, you’ll need a proper data center for space, power, cooling, fire suppression etc.
- And then there’s labor (and expertise) ….

In a Nutshell ….

=

http://www.sportsbet.com.au/blog/wp-content/uploads/money1.jpg

http://gadgetsin.com/uploads/2010/02/samsung_ecogreen_f3eg_hard_disk_drive_1.jpg

Why Not Commercial Cloud?

- You definitely can do metered storage in a commercial cloud.
- But, it can be pricey over the long term (including IDC @ ~50%):
  - Microsoft Azure Blob Archive Storage
    - $0.001 per GB per month = ~$57 per TB used over 8 years (really $84) assuming Moore's Law doubling period is 2 years (**NO LONGER TRUE**)
    - https://azure.microsoft.com/en-us/pricing/details/storage/blobs/
  - Google Coldline Storage
    - $0.0012 per GB per month = ~$69 per TB used over 8 years (really $102) assuming Moore's Law doubling period is 2 years (**NO LONGER TRUE**)
    - Costs 5 cents per GB to retrieve.
    - https://cloud.google.com/storage/archival/
  - Amazon Glacier
    - $0.004 per GB per month = ~$230 per TB used over 8 years (really $339) assuming Moore's Law doubling period is 2 years (**NO LONGER TRUE**)
    - https://aws.amazon.com/glacier/pricing/
  - NOTE: All of these keep multiple copies.
    – But, how will you pay after the grant ends?

What About Longer Term?

What happens when the grant that generated the data runs out?

- Can you get a new grant to pay for archiving your old data?

Example: 1 PB of content for 8 years, at least dual copies:

IBM V5010E:	2 copies, rebuy @ yr 6	= ~$920,000 at current pricing:	**38×**
Amazon:	2+ copies	= ~$339,000 at current pricing:	**14×**
45drives:	2 copies, rebuy @ yr 6	= ~$254,000 at current pricing:	**11×**
Dell MD:	2 copies, rebuy @ yr 6	= ~$234,000 at current pricing:	**10×**
Google:	2+ copies	= ~$102,000 at current pricing:	**4.3×**
Microsoft:	2+ copies	= ~$84,000 at current pricing:	**3.6×**
USB disks:	2 copies, rebuy @ yr 6	= ~$60,000 at current pricing:	**2.5×**
OURRstore:	2 copies	= ~$23,600 at current pricing:	**base**

(assumes dual copies for all non-cloud solutions; some include IDC; equipment gets bought again at the start of year 6)

  - Explanation shortly

What About Tape's Fixed Costs?

- Tape has **huge fixed costs**: In OURRstore, 72% of total equipment costs are tape/disk/switch/software/media components that are needed regardless of the number of tape cartridge slots (this takes into account 8 years of warranty/support/maintenance):
  - tape library base frame;
  - tape library drive expansion frame;
  - tape drives;
  - servers;
  - disk;
  - network switches;
  - software.
- Only 28% of the OURRstore budget is for tape cartridge slots (~11,000 of them).

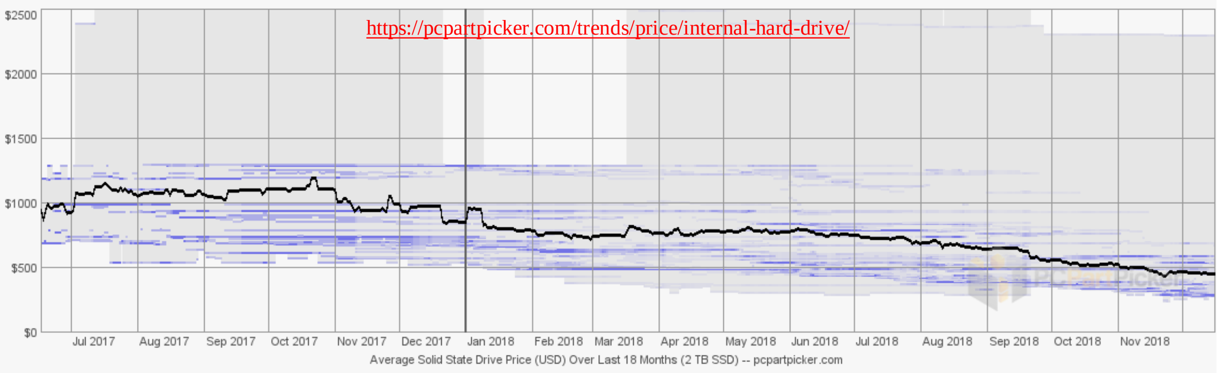
What About Longer Term?

Will Moore’s Law save you?

- No, because your datasets will grow much faster:
  - Computing speed and capacity doubles every 24 months.
  - Next Generation Sequencing improves 10x every 16 months.
- No, because spinning hard disk drive price improvements are slowing down: http://www.mkomo.com/cost-per-gigabyte
  - 1980-2009: $/TB halved every 14 months (1 $^{1}/_{6}$ years)=>116×/8yrs
  - 2009-2013: $/TB halved in 44 months (3 $^{2}/_{3}$ years)=> 4.5×/8yrs
  - 2013-2020: $/TB halved in 58 months (4 $^{5}/_{6}$ years)=> 3.1×/8yrs
  - Tape (LTO):$/TB halved every 30 months (2 ½ years)=> 6.3×/8yrs
    - Tape used to improve slower than disk, but now tape improves faster than disk.
    - Tape’s areal density is about 1% of spinning disk’s, so it has much more room to grow in capacity than spinning disk (6-7 doublings). https://spectrum.ieee.org/computing/hardware/why-the-future-of-data-storage-is-still-magnetic-tape

What About SSD?

- SSD prices were roughly cut in half in calendar 2018.



https://pcpartpicker.com/trends/price/internal-hard-drive/

Average Solid State Drive Price (USD) Over Last 18 Months (2 TB SSD) -- pcpartpicker.com

- However, this was caused by a lack of demand for components: manufacturers made too much capacity, so
  the price of components dropped.
- We can’t count on that happening on an ongoing basis.
- SSD is still very expensive: ~$1M per PB for 8 years
  (5 × XL60 w/ 60 × 7.68 TB SSD mix use × 2 copies + rebuy) =
  ~4 × spinning hard disk price per TB

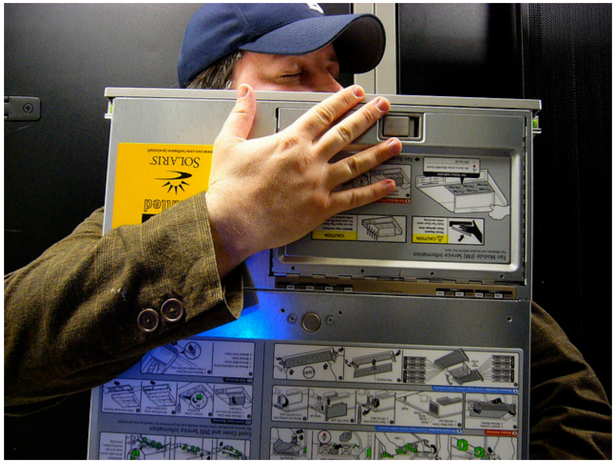
Researchers in the Wild

How Do Researchers Behave in the Wild?

- Territoriality
- Affordability
- No data management strategy
- Why?

Territoriality

- Some researchers like to hug their toys – because they don’t trust others (a) to provide shared resources to a large community, while simultaneously (b) serving each user’s specific needs well (and at high priority).



http://gigaom2.files.wordpress.com/2012/10/jason-server-hug.jpeg

http://enterprise.media.seagate.com/files/2009/09/computerhug460x276-300x180.jpg

Affordability

- Some researchers perceive roll-your-own as cheaper[?] than a central resource – even when it's actually **more expensive** because of non-obvious (non-hardware) costs.
    - Space, power, cooling: rack-in-a-closet isn't plausible any more.
    - Labor: requires expertise far beyond a typical STEM researcher.
    - Maintenance: not cheap, especially after 3 to 4 years.
    - But they have to stretch their research funds as far as possible.

ClipartOf.com/1057687

Data Mgmt Strategy?

- Some research teams store their research data on a single hard drive in the PC under a grad student’s desk:
  - The faculty member doesn’t know what format the data is in, where to find it, nor how to read it – so when the grad student graduates, the data essentially becomes unusable.
  - May be rarely if ever backed up.
- Some research teams have a box full of USB disk drives, in which case:
  - There’s no guarantee that the drives still work.
  - The files aren’t searchable, unless someone has bothered to keep an up-to-date inventory – which they probably haven’t.
  - May be rarely if ever backed up.

Why?

- **Perception**: Some researchers perceive their administrations (especially but not only central IT) as barriers to their progress, instead of partners in their progress.
  - In some cases, this is based on direct negative experience and/or advice/anecdotes from colleagues.
- **Mindset**: For some users, the bulk of their hands-on computing experience is with personal computing (PCs, laptops, tablets, phones), which typically are relatively straightforward to manage with tiny capital, labor and expertise cost (e.g., increase phone storage by inserting MicroSD card; install software with a few taps, for a few dollars or free).
- **Cost**: Grad student labor is (relatively) cheap.
- **Incentives**: At most academic institutions, faculty incentives are based on graduating students, publishing papers, and getting external funding – **NOT** on having well-managed IT resources.

How to Be, and Seem, Cheaper?

- **Distribute the costs among multiple entities.**
  - That way, no one has to bear the whole burden.
  - Therefore, the cost for each becomes affordable.
- Find ways to **leverage the funding** to get other funding.

A Business Model for Physical Management of Big Data

Business Model

First PetaStore (2010-20), then OURRstore (2020-28)

- **Grant**: hardware, software, initial warranties on everything
- **Institution (CIO)**: space, power, cooling, labor, maintenance after the initial warranty period
- **Researchers**: media (tape cartridges)

- Compared to roll-your-own disk, for our researchers OURRstore tape is:
  - cheaper per TB @ dual copies;
  - more reliable (bit rot rate is 10% to 0.1% as high as disk);
  - less labor (treat it as a filesystem);
  - requires less expertise to set up and use (~1 hour training);
  - slower (moderate bandwidth, very high latency: 1 minute, not 10 ms).

OURRstore Technology Strategy

- Distribute the costs among a research funding agency, the institution, and the research teams.
- Archive, not live storage: “Write once, read seldom if ever.”
- Independent, standalone system: not part of a cluster.
- Spend grant funds on many slots but few tape cartridges.
- Media slots are available on a first come first serve basis.*
- Software cost should be a modest fraction of total cost.
- Maximize media longevity.
- Globus for file transfers, file sharing, ~~file publishing~~, discoverability etc. (Institutional repositories for file publishing.)
- LTFS (tiny file catalog on each tape cartridge):
  Ship secondary copies to the data owner – if anything goes wrong, it’s under $3K to buy an LTO tape drive, and the software is free.

OU’s NSF MRI Archive Grants

- **PetaStore**: “MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research,” OCI-1039829, $792,925, 10/1/2010-9/30/2013 (+ 1 year no cost extension)
  - That grant was 3 years, took a 1 year no cost extension;
    its archive was meant to be 6 years
    but is currently in the first half of year 9 of full production.
  - 12 research teams on the proposal, 55 now.
- **OURRstore**: “MRI: Acquisition of a Regional Resource for Long-term Archiving of Large Scale Research Data Collections,” OAC-1828567, $967,755, 9/1/2018 - 8/31/2021
  - The grant is 3 years; its archive will be for 8+ years.
  - 87 research teams on the proposal; a few have dropped, many will join.

Who’s Eligible? Who’s In?

- Institutions in Great Plains Network (GPN) states. (AR,KS,MO,NE,OK,SD)
- Institutions in the 28 EPSCoR jurisdictions.
  (AK, AL, AR, DE, GU, HI, IA, ID, KS, KY, LA, ME, MS, MT, ND, NH, NE, NM, NV, OK, PR, RI, SC, SD, VI, VT, WV, WY)
- Institutions (and consortia) in non-EPSCoR jurisdictions, if they buy their own cartridge expansion cabinet (limited space!).
- In the proposal, 87 research teams at 27 institutions in 17 states, including 29 research teams at OU. (A few have dropped, many will be added.)
- 16 teams will each need at least 1 PB:
  8 at OU, 1 in another GPN state, 7 in non-GPN EPSCoR states.
  By contrast, the original PetaStore proposal included only 12 teams *total*, regardless of capacity need.

How Much Need?

Per the proposal:

- Capacity needed: 134 PB (~17,500 LTO-7 Type M cartridges)
  - $31M in on-premise spinning disk RAID, OR
  - $11M in a commercial cloud, OR
  - $8M in USB disk drives (Good luck managing that!), OR
  - $3M in tape cartridges
    - If we bought the full 134 PB today.
- Active funding of these projects: $162M
- Pending/planned funding: $140M
- Faculty: 250+
- Staff: 150+
- Postdocs: 100+
- Graduate students: 500+
- Undergraduate students: 500+

Yeah, But Tape Sucks!

- Well, yes, tape does suck:
  - Retrieval has very high latency (typically 1 minute per file, vs ~10 ms per file for spinning disk).
  - Tape medium inside a tape cartridge can break -- tinsel!
- How to resolve?
  - Only store large files (OURRstore minimum is 1 GB).
    - So, you have to create Zip files or compressed tar files.
  - Offline storage: Download file to disk before using.
  - Think hierarchically:
    - small amount of very fast disk;
    - medium amount of slower disk;
    - large amount of tape.

Investment Protection

- PetaStore (current archive) will reach end-of-life when OURRStore gets to full production.
  - And what happens when OURRstore gets to end-of-life?
    - We’re actually already thinking about this:
      If we fill OURRstore, will we need to run 2 archives side by side?
- Faculty may not have funds for purchasing new media in the next archive for their old data (because the old data may not be relevant to their new grants).
- Need to provide for buying up front instead of recurring charges.
- How to handle the tape?

Original Longevity Strategy

- OURRstore would have to be backward-compatible with the PetaStore, in the sense of allowing LTO, including LTO-5 and LTO-6.
    - Tape cartridges are good for the earliest of:
        - 15+ years
        - 5000 load/unload cycles
        - 200 complete tape read/writes
    - So far, only 6 PetaStore tape cartridges (<< 1%) are in danger of wearing out in less than 15 years.
- OURRstore had to include some LTO-6 drives, which could read and write both LTO-6 and LTO-5, but new tape cartridges would be LTO-7 Type M (9 TB).
- Unlike disk drives, tape cartridges can migrate from archive to archive.

New Longevity Strategy

- Which is cheaper?
  - Keep the old tape cartridges (LTO-5, LTO-6)?
  - Replace old with new tape cartridges (LTO-7 Type M)?
- Replacing the old (LTO-5 and LTO-6) tape cartridges with new (LTO-7 Type M) tape cartridges is cheap.
- It would cost more the keep the old tape cartridges,
  even though the old tape cartridges have already been paid for!
  - They consume a lot more tape cartridge slots, which aren’t cheap!
  - They require old tape drives, which can be brought over from the old tape archive, but we have to pay annual maintenance on them, which isn’t cheap!
  - Buying new tape cartridges and dumping old saves mid-5-figures!

Longevity Mechanism

Once OURRstore is in full production:

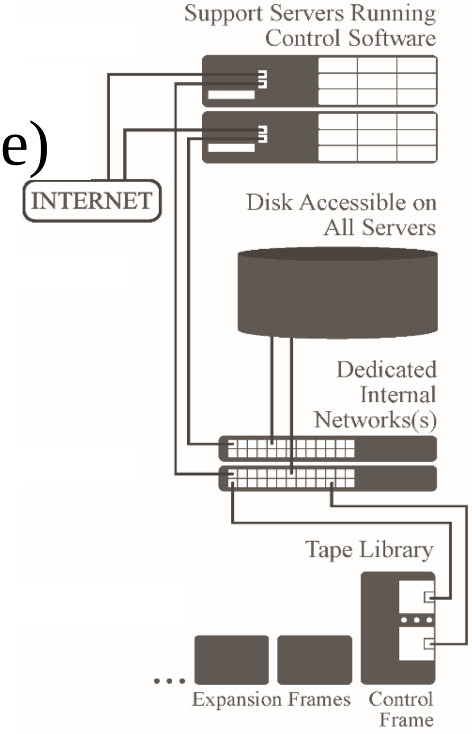
- Set the PetaStore to read-only.
- On the PetaStore, for each tape cartridge, identify all its files.
- Copy all those files to OURRstore's new tape cartridges.
- When all files are copied (months, maybe a year), decommission the PetaStore.
    - Best case is 3 weeks; expected is several months.

We hope to use this same procedure at OURRstore's end of life (c. 2028).

OURRstore Architecture

OURRstore Architecture

- Tape library: base frame, 28 tape drive slots, expansion frames
- Tape drives: 6 × LTO-8 (1.8 GB/sec aggregate)
- Servers: 3 for disk SW, 2 for tape SW
- Disk: ~600 TB, ~2 GB/sec
  - Slow on purpose: it's an archive.
- Network switches: dual FC32, dual 10GE
- Software
  - Tape control SW: IBM Spectrum Archive
  - Disk control SW: IBM Spectrum Scale (GPFS)
  - Globus


Support Servers Running Control Software
INTERNET
Disk Accessible on All Servers
Dedicated Internal Networks(s)
Tape Library
...
Expansion Frames
Control Frame

OURRstore’s New Features

OURRstore New Features

- Disk Caching
- Globus
- LTFS
- Research Data Management
- Archive Longevity Approach

Disk Caching

- Files that have been requested recently will stay on disk as long as possible, until disk capacity needs to be cleared out to make room for other files.
- We can do this because we're being very careful stewards of both our NSF grant funds and our CIO's cost share funds, and we found a way to get:
  - much much lower disk cost than in the winning bid response;
  - much more disk capacity than requested in the RFP (~600 TB vs 100-200 TB);
  - faster than requested in the RFP (~2.5 GB/sec vs 1-2 GB/sec).

Globus

- Globus is a software product that has a long history.
  - Funded in the late 1990s/early 2000s by the NSF.
  - Became a web-based large file transfer service in the early 2010s.
  - Now an independent, standalone software system (still uses web).
  - The software features below would cost most of our institutions ~$20K per year – but on OURRstore, only the grant pays (and then OU’s CIO, starting in year 5, but at a much lower rate).
- File sharing: A file owner can decide, for any file or directory, who can access it, not only among OURRstore users, but also outsiders who want to download the data.
  - The incremental download cost is zero.
- ~~File Publishing~~: No longer a supported Globus feature.

LTFS

- Starting with LTO-5, every LTO tape cartridge has a little chunk of capacity that's set aside for a file catalog that describes every file on that tape cartridge.
- For PetaStore, we didn't use LTFS, because the software we could afford didn't understand LTFS; for OURRstore, we will.
- So, for second copies, we'll require a prepaid shipping label, and we'll export the secondary copy and ship it to the file owner.
  - The reviewers insisted we require at least dual copies for all files, on the fear that, if we offered single copy, that's all anyone would use, and OURRstore would hold the sole copy of many datasets.
- Because of shipping you the secondary copy of your files, if something goes wrong at OU, you can still access your files, for less than $3000 for a tape drive, with free software.

Research Data Management #1

- The first OURRstore proposal (2016) didn’t discuss research data management; the reviewers complained about that.
- For the second OURRstore proposal (2017), we determined that:
  - we’d better address research data management in the proposal, BUT
  - we couldn’t solve research data management for the dozens of disciplines that would use OURRstore.
- Solution: Help each OURRstore user identify a research data librarian at their own institution who could help.
  - This will allow people who have the right expertise to provide the right guidance – instead of us trying to have to learn an entirely new discipline in zero time.

Research Data Management #2

- Monthly research data management conference calls will start once OURRstore is in production, led by Dr. Mark Laufersweiler, OU Libraries' Research Data Specialist.
- We need to identify and recruit research data librarians at many OURRstore institutions:
  - Clemson U (SC)
  - Idaho State U
  - New Mexico State U
  - North Dakota State U
  - U Louisiana Lafayette
  - U Montana
  - U New Hampshire
  - U New Mexico
  - U North Dakota

OURRstore Longevity Approach #1

- A tape library can last ~15 years.
  - Vendor support will be available most or all of that time.
  - But, the product line may become obsolete during that period, in which case new components (e.g., tape drives for newer versions of LTO) wouldn’t be available from the manufacturer.
    - TS3500 (PetaStore): available for sale 2008-17.

https://www.manualslib.com/manual/1262732/Ibm-Ts3500.html?page=12#manual

https://www.parkplacetechnologies.com/end-of-service-life/ibm/

- Servers, disk and network switches don’t last nearly as long.
- Strategy (we have budget for this):
  - Buy the tape library, servers, disk, network switches and software at the beginning of the grant period.
  - Again buy servers, disk and maybe network switches at the end of the grant period.
    - But retain the tape library and software, just paying maintenance.

OURRstore Longevity Approach #2

- New versions of LTO come out every ~2½ years on average.
- LTO-8 drives came out in late 2017, and we have 6 drives, so we already can read and write LTO-8, LTO-7 Type M and LTO-7.
- LTO-9 drives are expected in mid-2020 (during grant period).
  - We've budgeted for 4 × LTO-9, but there's no advantage to buying: LTO-8 tape cartridges will become breakeven per TB after LTO-9 drives are released (probably a year later or so).
- LTO-10 drives: expected late 2022/early 2023 (will require up to 2 years of No Cost Extensions to our grant)
  - We probably can afford 4 × LTO-10 drives (can read/write LTO-9).
- LTO-11 drives: mid-2025 (during the planned life of OURRstore)
- LTO-12 drives: early 2028 (near OURRstore likely end-of-life)
  - So LTO-11 cartridges will probably become breakeven in 2028-29.

https://en.wikipedia.org/wiki/Linear_Tape-Open

LTO Weirdness

LTO Versions

NOTE: LTO-N tape drives can (usually) write N-1 and read N-2.

- LTO-5: 1.5 TB raw, 140 MB/sec, 2/2010 release
  http://www.backupworks.com/Data-Storage-backup-news.aspx
- LTO-6: 2.5 TB raw, 160 MB/sec, 12/2012 (22 months)
- LTO-7: 6.0 TB raw, 300 MB/sec, 12/2015 (36 months)
  - LTO-7 “Type M:” 9 TB raw, 300 MB/sec – **NO upcharge!**
    - LTO-7 “Type M” in LTO-8 tape drives only
- LTO-8: 12 TB raw, 360 MB/sec, 12/2017 (36 months)
  - LTO-8 tape cartridges became available after a lawsuit.
    https://insight.rpxcorp.com/litigation_documents/12445782
- LTO-9: 24 TB raw, 708? MB/sec, 7/2020?
- LTO-10: 48 TB raw, 1100? MB/sec, 12/2022?
- LTO-11: 96 TB raw, ? MB/sec, 7/2025?
- LTO-12: 192 TB raw, ? MB/sec, 12/2027?

Weirdness: LTO-7 Type M

If you have an LTO-8 tape drive, an LTO-7 tape cartridge, and an LTO-7 "Type M" barcode for that LTO-7 tape cartridge, then:

- in the LTO-8 tape drive, you can format the LTO-7 tape cartridge to be LTO-7 Type M;
- that tape cartridge's raw capacity will be 9 TB instead of 6 TB;
- that tape cartridge will only be readable and writable by an LTO-8 drive (non-LTO-8 won't be able to read nor write it);
- that tape cartridge's speed will be LTO-7 speed (300 MB/sec), even though it'll be in an LTO-8 tape drive (360 MB/sec).

So, the price per TB is cut by a third, at no dollar cost and almost no effort, in exchange for a 17% reduction in bandwidth. (And the cost advantage is even better when you count the cost of a tape cartridge slot in your tape library.)

Will There Be Another Type M?

- Apparently, LTO-7 Type M was able to happen because LTO-7 and beyond use a different material for their magnetic particles than LTO-6 and older (BaFE for LTO-7 and beyond, vs metal particulate for older).
- No one knows whether there'll be another Type M – this may be one time only, in which case we might never, or only briefly, recommend LTO-8 cartridges for OURRstore users.
  - This is because it'll probably take a lot longer than usual for LTO-8 to become breakeven with LTO-7 in price per TB.
    - Usually LTO-(N-1) becomes the cheapest per TB many months (perhaps more than a year) after LTO-N is released.
    - Right now, LTO-8 is more than 50% more expensive per TB than LTO-7 Type M.
      - Which means that LTO-8 is almost breakeven with regular LTO-7.

Acknowledgements

Acknowledgements: People

- NSF MRI OURRstore Leadership
    - Co-PIs Laura Bartley, Kendra Dresback, Amy McGovern, Horst Severini
    - Senior Personnel Mark Laufersweiler
- NSF MRI OURRstore Advisory Committees
- NSF MRI OURRstore Research Data Librarians
- NSF MRI OURRstore users
- OSCER Team: David Akin, Patrick Calhoun, Jim Ferguson, Kali McLennan, Horst Severini, Jason Speckman
    - Recent Former: Debi Gentis, George Louthan, Ashish Pai, John Shelton, Matt Younkins
- OU CIO David Horton
- OU Interim Dean of University Libraries Carl Grant

Acknowledgements: Grants

Portions of this material are based upon work supported by the National Science Foundation under the following grants:

- Grant no. OAC-1828567, “MRI: Acquisition of a Regional Resource for Long-term Archiving of Large Scale Research Data Collections,” PI H. Neeman, $968K, 2018-21
- Grant no. ACI-1440783, “CC*IIE Engineer: A Model for Advanced Cyberinfrastructure Research and Education Facilitators,” PI H. Neeman, $400K, 2014-17
- Grant no. ACI-1341028 , “CC-NIE Integration: OneOklahoma Friction Free Network,” PI H. Neeman, $500K, 2013-17
- Grant no. OCI-1039829, “MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research,” PI H. Neeman, $792K, 2010-14
- Grant no. EPS-1006919, “Oklahoma Optical Initiative,” PI H. Neeman, $1.17M, 2010-13
- Grant no. OAC-1531128, “MRI: Acquisition of Shared High Performance Compute Cluster for Multidisciplinary Computational and Data-Intensive Research,” PI D. Brunson, OSU, $952K, 2015-18
- Grant no. ACI-1126330, “MRI: Acquisition of a High Performance Compute Cluster for Multidisciplinary Research,” PI D. Brunson, OSU, $909K, 2011-15
- Grant no. ACI-1229107, MRI: Acquisition of a High Performance Computing Cluster for Research and Education,” PI J. Snow, Langston U, $250K, 2013-16
- Grant no. ACI-1440774, “CC*IIE Region: Leveraging Partnerships Across the Great Plains to Build Advanced Networking and CI Expertise,” PI G. Monaco, Co-PI J. Deaton, $130K, 2014-17
- Grant no. ACI-1429702, “MRI: Acquisition of a High Performance Computing Cluster for Research at a Predominantly Undergraduate Institution,” PI. E. Lemley, UCO, $305K, 2015-17
- Grant no. ACI-531270, “MRI: Development of Heterogeneous Cluster for Cyber-Physical System Hybrid Analytics,” PI J. Hale, TU, $181K, 2015-17
- Grant no. 1659235, “CC* Network Design: Multiple Organization Regional One Oklahoma Friction Free Network (MORe OFFN), PI R. Reif, NSU, $334K

Bibliography

1. S. P. Calhoun, D. Akin, B. Zimmerman and H. Neeman, 2016: "Large Scale Research Data Archiving: Training for an Inconvenient Technology." *Journal of Computational Science*. DOI: 10.1016/j.jocs.2018.07.005.
2. H. Neeman, K. Adams, J. Alexander, D. Brunson, S. P. Calhoun, J. Deaton, F. Fondjo Fotou, K. Frinkle, Z. Gray, E. Lemley, G. Louthan, G. Monaco, M. Morris, J. Snow and B. Zimmerman, 2015: "On Fostering a Culture of Research Cyberinfrastructure Grant Proposals within a Community of Service Providers in an EPSCoR State." *Proc. XSEDE'15*, article 19. DOI: 10.1145/2792745.2792764.
3. H. Neeman, D. Akin, J. Alexander, D. Brunson, S. P. Calhoun, J. Deaton, F. Fondjo Fotou, B. George, D. Gentis, Z. Gray, E. Huebsch, G. Louthan, M. Runion, J. Snow and B. Zimmerman, 2014: "The OneOklahoma Friction Free Network: Towards a Multi-Institutional Science DMZ in an EPSCoR State." *Proc. XSEDE'14*, article 49. DOI: 10.1145/2616498.2616542.
4. S. P. Calhoun, D. Akin, J. Alexander, B. Zimmerman, F. Keller, B. George and H. Neeman, 2014: "The Oklahoma PetaStore: A Business Model for Big Data on a Small Budget." *Proc. XSEDE'14*, article 48. DOI: 10.1145/2616498.2616548.
5. H. Neeman, D. Brunson, J. Deaton, Z. Gray, E. Huebsch, D. Gentis and D. Horton, 2013: "The Oklahoma Cyberinfrastructure Initiative." *Proc. XSEDE'13*, article 70. DOI: 10.1145/2484762.2484793.

Thanks for your attention!

QUESTIONS?