
INNOVATION & TECHNOLOGY
SUMMIT
2017

Large Scale Data Archiving: Big Data on a Small Budget

Why Storage for Big Data is Hard

Henry Neeman
Director, OU Supercomputing Center for Education & Research (OSCER)
University of Oklahoma

U Texas Dallas Innovation & Technology Summit 2017, Tuesday May 16 2017

Outline

- The Challenge of Physical Data Management.
- Researchers in the Wild
- A Business Model for Physical Management of Big Data

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

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?

- If a research team's data sizes are small, roll-your-own is perfectly reasonable:
  - USB disk drives are cheap:
    - 6 TB USB 3.0 = $170 pricewatch.com 4/9/2017 => ~$33 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).
- Slightly bigger than that: can do a small, cheap RAID enclosure for mirroring or RAID6, **BUT**:
  - Price per TB starts going way up:
    - Backblaze 60x10TB = ~$49K web 4/9/2017 => ~ $91 per usable TB
    - Dell MD3060e 60x10TB = ~$76K web 4/9/2017 => ~$141 per usable TB
      - Backblaze 2.7 x USB, Dell 4.3 x 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.

Enterprise-Class Disk

Enterprise-class disk

- IBM Storwize V5010 (entry/midrange enterprise product)
  - 36 x 8 TB = ~$183K MSRP = ~$800 per usable TB per copy
- Of course, you'll need dual copies, so double this.
- Of course, you'll need a proper data center for space, power, cooling, fire suppression etc.
- And then there's labor ….

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 Metered?

- You definitely can do metered storage.
- But, it can be pricey over the long term:
  - Amazon Glacier
    - $0.004 per GB per month = ~$135 per TB used over 5 years, assuming Moore's Law doubling period of 2 years (**NO LONGER TRUE**)
  - Google Coldline Storage
    - $0.007 per GB per month = ~$236 per TB used over 5 years assuming Moore's Law doubling period of 2 years (**NO LONGER TRUE**)
    - Costs 1 cent per GB to retrieve.
  - Microsoft Azure Cool Storage
    - $0.010 per GB per month = ~$337 per TB used over 5 years assuming Moore's Law doubling period of 2 years (**NO LONGER TRUE**)
  - NOTE: All of these keep multiple copies.

What About Longer Term?

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

- Can you archive it at a national center (e.g., iPlant, NCBI)?
- Can you get a new grant to pay for archiving your old data?

Example: At OU, we currently have, in our long term archive, ~1 PB of content (ignoring secondary copies):

- IBM enterp: 5 years 1 PB = ~$1,600,000 at current pricing: 47x
- Microsoft: 5 years 1 PB = ~$337,000 at current pricing: 9.9x
- Dell: 5 years 1 PB = ~$282,000 at current pricing: 8.3x
- Google: 5 years 1 PB = ~$236,000 at current pricing: 6.9x
- Backblaze: 5 years 1 PB = ~$182,000 at current pricing: 5.4x
- Amazon: 5 years 1 PB = ~$135,000 at current pricing: 4.0x
- USB disks: 5 years 1 PB = ~$66,000 at current pricing: 1.9x
- OU internal: 5 years 1 PB = ~$34,000 at current pricing (assumes dual copies for all non-cloud solutions)
  - Explanation shortly

Even Longer Term

What about the 2nd 5 years (years 6-10)?

- IBM enterp: 2nd 5 years 1 PB = ~$283,000 if Moore's Law
- Microsoft: 2nd 5 years 1 PB = ~$60,000 if Moore's Law
- Dell: 2nd 5 years 1 PB = ~$50,000 if Moore's Law
- Google: 2nd 5 years 1 PB = ~$42,000 if Moore's Law
- Backblaze: 2nd 5 years 1 PB = ~$32,000 if Moore's Law
- Amazon: 2nd 5 years 1 PB = ~$24,000 if Moore's Law
- USB disk: 2nd 5 years 1 PB = ~$12,000 if Moore's Law
- OU internal: 2nd 5 years 1 PB = -- $0 -- if Moore's Law
  (assumes dual copies for all non-cloud solutions)

This again assumes a Moore's Law doubling period of 2 years, which, again, is **NO LONGER TRUE**.

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 per 16 months.
- No, because disk drive price improvements are slowing down:
  - From 250 GB to 3 TB: mean doubling period = ~23 months
  - Starting with 4 TB: mean doubling period = ~36 months
  - Tape (LTO format): mean doubling period 30 months
    - Tape used to improve slower than disk, but now tape improves faster than disk.

Researchers in the Wild

COMPUTERWORLD 100

PC MAGAZINE TOP WIRED COLLEGES 2006

COMPUTERWORLD 100 BEST PLACES TO WORK IN IT 2006

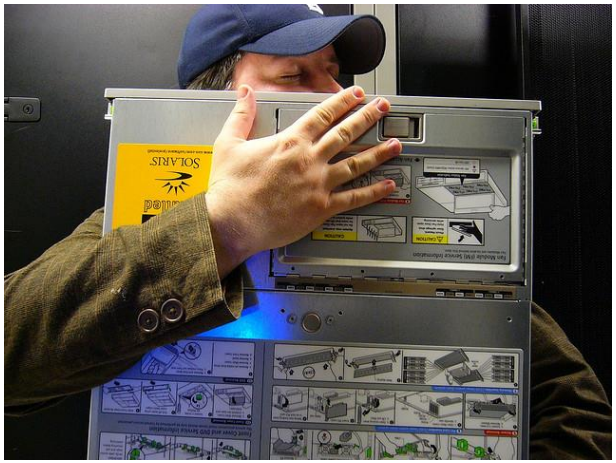
OU it INFORMATION TECHNOLOGY
THE UNIVERSITY OF OKLAHOMA

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

No Data Mgmt Strategy

- Some research teams store their research data is on a single hard drive in the PC under a grad student's desk, in which case:
  - 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 some 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

COMPUTERWORLD 100

PC MAGAZINE TOP WIRED COLLEGES 2006

COMPUTERWORLD 100 BEST PLACES TO WORK IN IT 2006

OU it INFORMATION TECHNOLOGY
THE UNIVERSITY OF OKLAHOMA

Business Model

Oklahoma PetaStore

- **Grant**: hardware, software, 3 year warranties on everything
- **Institution (CIO + VPR)**: space, power, cooling, labor, maintenance after the 3 year warranty period
- **Researchers**: media (tape cartridges)
- Compared to roll-your-own disk, for researchers PetaStore tape is:
  - cheaper
  - more reliable
  - less labor
  - requires less training (~1 hour)
  - slower (moderate bandwidth, very high latency)

OK PetaStore 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 media slots but few media (tape cartridges, disk drives).
  - Most of the media that the grant has purchased have been allocated to the research projects in the proposal.
- Media slots are available on a first come first serve basis.
- Software cost should be a small fraction of total cost.
- Under the OneOklahoma Cyberinfrastructure Initiative, this is also true for academic institutions statewide (and also many non-academic institutions).
- Maximize media longevity.

NSF MRI Grant

"Acquisition of Extensible Petascale Storage for Data Intensive Research"

National Science Foundation grant no. OCI-1039829

10/1/2010 - 9/30/2013, no cost extension to 9/30/2014

NSF MRI Grant: Summary

OU was awarded a National Science Foundation (NSF) Major Research Instrumentation (MRI) grant in 2010.

It features 15 faculty and staff from 12 projects in 10 departments.

We purchased and deployed a combined disk/tape bulk storage archive from IBM:

- the NSF budget paid for most of the hardware and software, plus warranties/maintenance for 3 years;
- OU cost share and institutional commitment pay for space, power, cooling and labor, as well as maintenance after the 3 year project period;
- individual users (e.g., faculty across Oklahoma) pay for the media (disk drives and tape cartridges).

Yeah, But Tape Sucks!

- Well, yes, tape does suck:
  - Retrieval has very high latency (typically 1 minute per file).
  - Tape medium inside a tape cartridge can break!
- How to resolve?
  - Only store large files (PetaStore 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 “slow” disk
    - Large amount of tape

The Storage Hierarchy

- Fast things are expensive, therefore you can't afford much of them.
- Slow things are cheap, therefore you can afford a lot of them.
- Why?

Longevity

- The current PetaStore system will end-of-life roughly 2017/18.
- Faculty may not have funds for purchasing new media in PetaStore II for their old data.
- How to handle the tape?

Longevity Strategy

- PetaStore II has to be backward-compatible with PetaStore I, in the sense of allowing LTO, including LTO-5 and LTO-6 (could also allow non-LTO, if desired).
  - Tape cartridges are good for the earliest of:
    - 15 years
    - 5000 load/unload cycles
    - 200 complete tape read/writes
  - So far, only 6 tape cartridges (<< 1%) are in danger of wearing out in less than 15 years.
- PetaStore II must include some LTO-6 drives, which can read and write both LTO-6 and LTO-5.

Acknowledgements

NSF MRI Participants and External Advisory Group

OSCER Operations Team: David Akin, Brett Zimmerman, Patrick Calhoun, Kali McLennan, Jason Speckman, Kyle Dudgeon; Brandon George (now at DDN), Joshua Alexander (now at MSCI)

OU CIO/VPIT Loretta Early, Asst VPIT Eddie Huebsch
OU VP for Research Kelvin Droegemeier

OU IT: Fred Keller, Gensheng Qian, cable crew, etc.

IBM: Jim Herzig (now retired), Mike Kane (now at Verizon), Frank Lee, Tu Nguyen, Ray Paden, Eric Ballard

Acknowledgements

Portions of this material are based upon work supported by the National Science Foundation under the following grant: Grant No. OCI-1039829, “MRI: Acquisition of Extensible Petascale Storage for Data Intensive Research.”

Bibliography

- H. Neeman, S. P. Calhoun, B. Zimmerman, D. Akin, 2016: “Large Scale Research Data Archiving: Training for an Inconvenient Technology.” *Journal of Computational Science*, to appear.
- 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.

Thanks for your attention!

QUESTIONS?