Working with data on the Farm

Ruth Marinshaw // April 25, 2018 // GlobusWorld 2018

In my short time with you, we'll cover

- Brief overview of our current landscape (Ruth)
- Challenges (Ruth)
- Approaches and suggestions (**YOU** and Ruth)

View of our Current Landscape

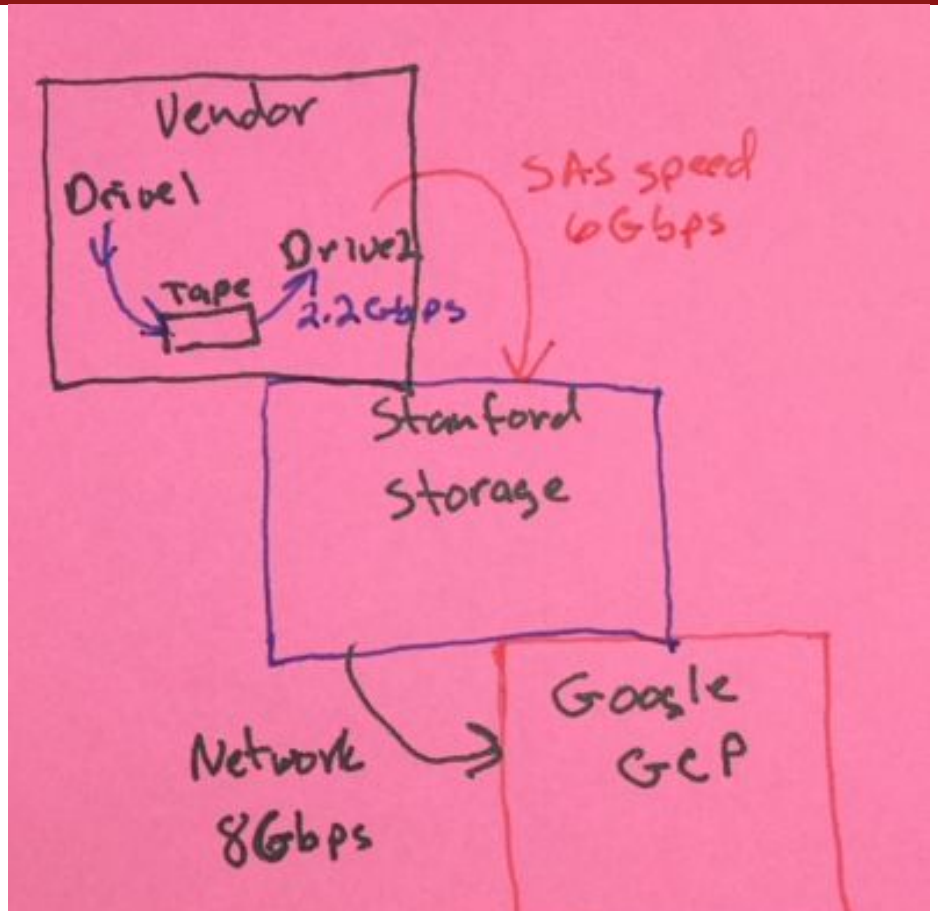
- We run a lot of stuff (clusters, storage systems, switches, etc)
- And support a lot of stuff (user & admin software tools, applications, activities)
- For >10K users with a BOATLOAD of stuff (datasets, applications, questions, ideas, requirements, demands, deadlines)
- Expectations exceed capabilities in across almost every dimension … especially time

Example: Genomics project data transfer needs

- Move 200GB per compressed file, 9000 files
- From point A to point C, via point B (Yes, you can do the math in your head)
- Where the path from A to B is LTO 8 tape
- And point C is a commercial cloud

Rejected ideas included cloud vendor transfer appliance; direct data transfer from A to C over network; network transfer A to B to C ….

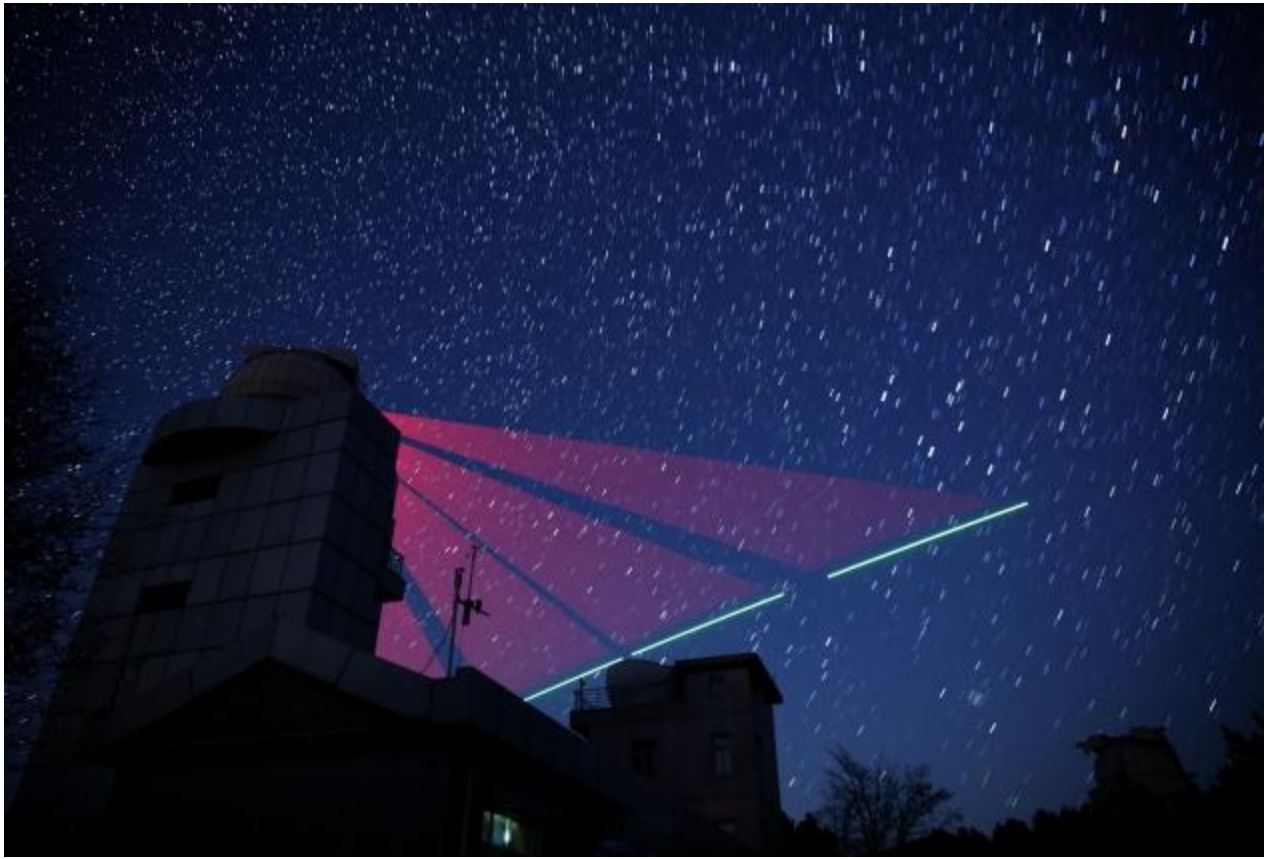
Meeting Notes



Challenges

- Frequency of need: rare, occasional, common
- Technical challenges
- Policy challenges (organizational, regulatory, assumed)
- In the presence of … incomplete information on all sides
- Goals assumed to be a mix of minimizing time and cost.
- With time being critical … “WHAT? 76 days??”

Data teleportation: not yet. We're good - but we aren't wizards.



Example: “It’s slow”

- Performance bottleneck moving data from campus device to AWS S3 endpoint. Daily workflow requires data movement.
- Differences in which Stanford network was used and which path was used to reach which AWS region (US-WEST-1 vs US-EAST-2)
- Guess what? If not behind firewall, it’s faster (~5.5 Gbs)
- And CLI on local endpoint wasn’t tuned for multi-part uploads and larger chunk sizes
- But the user thought all had been done correctly

Approaches/suggestions to educate & set expectations?

- Offer a “known” path: Stanford Research Network
- Have tools to help minimize the pain when possible
  - *“Thanks, Zhiyong. Globus works like a dream.”*
  - *SHUB/Sing. Registry and Globus integrations (V. Sochat’s work)*
- *Note to self: don’t say that we are making the process of moving data “seamless”*
- Document … but realize that 1:1 consultations will be required. Trust, but verify? Never trust?
- We (Stanford) need more examples (paths, speeds/feeds, etc)

How do YOU set user expectations???

Some examples of online docs around moving data:

- http://moo.nac.uci.edu/~hjm/HOWTO_move_data.html
- https://www.sherlock.stanford.edu/docs/user-guide/storage/data-transfer
- http://fasterdata.es.net/home/requirements-and-expectations/
- https://cloud.google.com/transfer-appliance/ (43 days/PB)

~~Questions?~~ Answers? Send to

ruthm@stanford.edu

Thank you!


LOSTLETTERMEN.CO