Some (biased) observations

Miguel Branco

Disclaimer

- The views in this set of slides are my own, and not necessarily endorsed by my employer, employees, former CERN colleagues or friends.

Disclaimer

- The views in this set of slides are my own, and not necessarily endorsed by my employer, employees, former CERN colleagues or friends.

- Information is provided “AS IS” without warranty of any kind. The information could include technical inaccuracies and/or typographical errors.

Disclaimer

- The views in this set of slides are my own, and not necessarily endorsed by my employer, employees, former CERN colleagues or friends.

- Information is provided “AS IS” without warranty of any kind. The information could include technical inaccuracies and/or typographical errors.

- Changes are periodically made to the information herein; these changes may (or may not) be incorporated in future editions of this publication.

My background

- I spent 6 years at CERN doing my best, but in practice, really making physicists' life miserable.

Things are moving fast

- … lots of testing!
- … lots of work!
- … lots of ideas!

- Some convergence of the underlying stacks:
  - OpenStack & friends nearly ubiquitous
  - A lot of ownCloud

Open Source

- Nothing but open source.

- Nothing new here, but that’s an important message for vendors.
  - Core software open source
  - Higher-end features open source
  - Everything open source!

A comment on “tech focus”

- 1) SysAdmins seem to want a “Dropbox” that “I” can manage
- 2) Storage people want to provide “Ease-to-use Storage”
- I’ve seem more of (1) in talks (a more accessible problem right now?)
- But (1) and (2) are NOT the same thing.
  - Suggest you consider explicitly where you are focusing.
  - Focus on what the science workflow needs.
  - **Where are scientists spending 90% of their time?** Is it sync’ing Word?

A comment on “tech focus”

- 1
- 2
- I’ ow?)

Google Drive is 2y old. How will doc handling look like in 2y? Will Word look like it does today? At EPFL, some 18y CS students do not know what “Office” is. They all know Google Docs. (And GitHub.)

- But (1) and (2) are NOT the same thing.
  - Suggest you consider explicitly where you are focusing.
  - Focus on what the science workflow needs. Where are scientists spending 90% of their time? Is it sync’ing Word?

A comment on “tech focus”

- 1
- 2
- I’ ow?)

Google Drive is 2y old. How will doc handling look like in 2y? Will Word look like it does today? At EPFL, some 18y CS students do not know what “Office” is. They all know Google Docs. (And GitHub.)

Vendor tech is !=. Large vs small-ish files? Updates: delta sync? Peer-to-peer: locality? Low latency? High throughput? Security? Scope: metadata? Provenance? Future features: notifications: push VS pull?

Verification

- Sync is complicated
  - How good is good enough?
  - From Word files to mission-critical data?
  - A Higgs analysis from data stored on “xxxBox”
  - Mission-critical data in the scientific context is likely read-only (easier problem)
  - … but datasets are bunches of files & if some missing, results biased (hard problem).
  - So, well, this is not obvious.

- “Academic” view: bugs/weirdness OK (they are obligatory) as long as understood
- My gut feeling: if your efforts focus on “scientific mission-critical data”, then verification is the distinguishing factor to “other” solutions
  - “For our requirements – see here – our system works exactly as expected”

Security: Authentication, Authorization, Accounting

- Academic research groups really are hierarchical.
- So are companies, so whatever companies use, should work for you as well.
- Except, well, companies are directed acyclic graphs and science collaborations are not ☺

- These issues were mentioned in passing in nearly all talks, but inconclusive.

- A non-trivial problem to address.

Federation

- Another dimension of complexity.
- Banks are also federated. So whatever works for them…
- Ah no. Science often needs to crosscut admin/bureaucratic barriers.
  - A “no no” in banking.

- How do you handle this?

- Intuitively, it is huge: “scientists” ‘freely’ sharing data, cross institution

Encryption

- Client-side encryption???
- Needed? Not needed?
- Depends on what is the data being stored.

CERNbox

- Representative of the various efforts around.
  - I appreciate the focus on “look-and-feel”!
- Convergence would be beneficial
  - This stuff is hard – particularly the last 20% - , so not point in diverging.
- Like every other system presented, I have my reservations:
  - Trying to support multiple, seemingly widely different backends / use cases
  - Assumingly a 90% solution. Fair enough, but which 90% exactly?

Let’s sync & share what we learn!

- Workshop also indirectly really useful to:
  - Document systems’ behavior (in presentations)
  - Learn experiences; deployment models under consideration (future “best practices”)
  - [Do more of this; particularly the templates Massimo/Jakub sent around]

- Future collaboration themes:
  - Dev side: Verification, Testing infrastructures (Deterministic even?)
  - Policy side: AAA (Authentication, Authorization, Accounting), Federation
  - SysAdmin side: Deployment experiences, Interoperability w/ legacy

Analysis on top of file sharing

- Baby steps still – e.g. Ganga
- Potentially game changer to how physicists work
- (I've seen the alternative… it is NOT pretty)

- Exciting to me to have this as the "underlying infrastructure", because it enables the *#reallycoolstuff* on top.
  - Interactive analysis
  - Share code, data, plots, histograms, results, reports together
  - Storage that gives you meaningful data
- Suggest you spend time looking into this explicitly: it is a game changer.

And what about you?

- What are your views?