Outline

- Network related issues and thinking for FAX
- Cost among sites, who has problems
- Analytics of FAX meta data, what are the problems

➢ The main object is to bring up discussion on WAN related issue when using FAX infrastructure

Also see URL for much broader SW options when running over WAN

https://indico.in2p3.fr/getFile.py/access?contribId=3&resId=0&materialId=slides&confId=6941

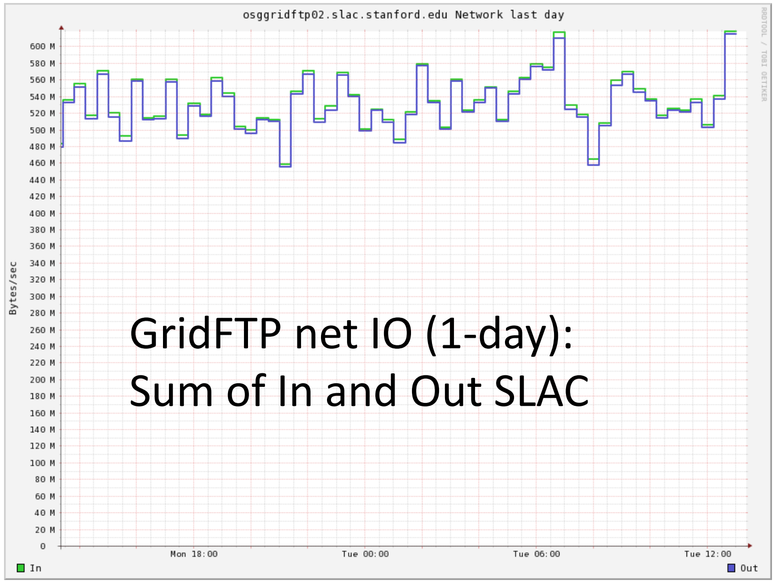
FAX and WAN
Requirement and Configuration

Wei Yang

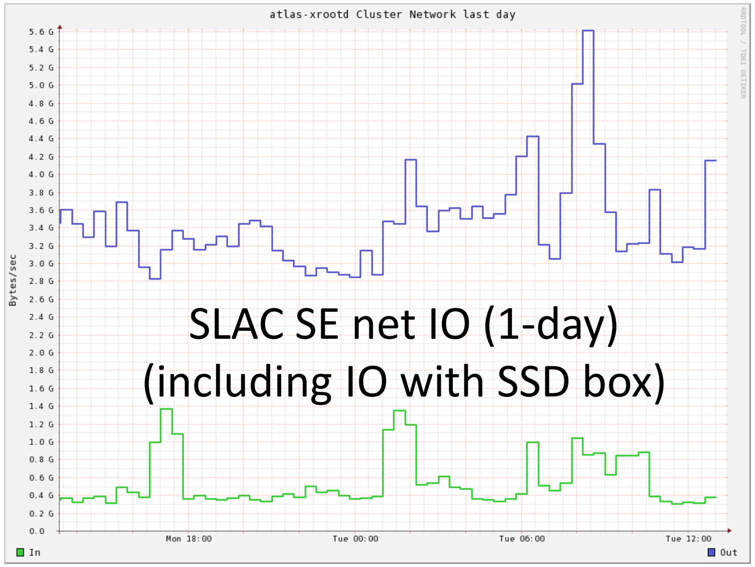
Diff: FAX and GridFTP

WAN Usage Difference:

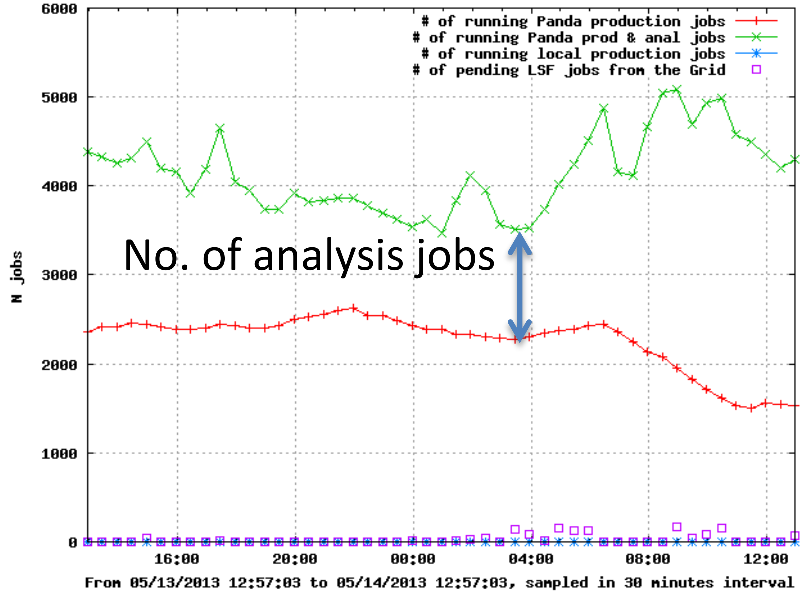
- Job driven: no 0-access files
- On demand, un-managed, no flow control
  - No help from FTS, demand go up and down quickly
- ➢ Q: will file level caching at server site and client site help?
  - Working dataset is very large. Need large cache, or very smart caching algorithm
- NOT to replace data pre-placement
  - ➢ Q: can we stop GridFTP data distribution and use FAX to pull
  - haven’t given it a thought



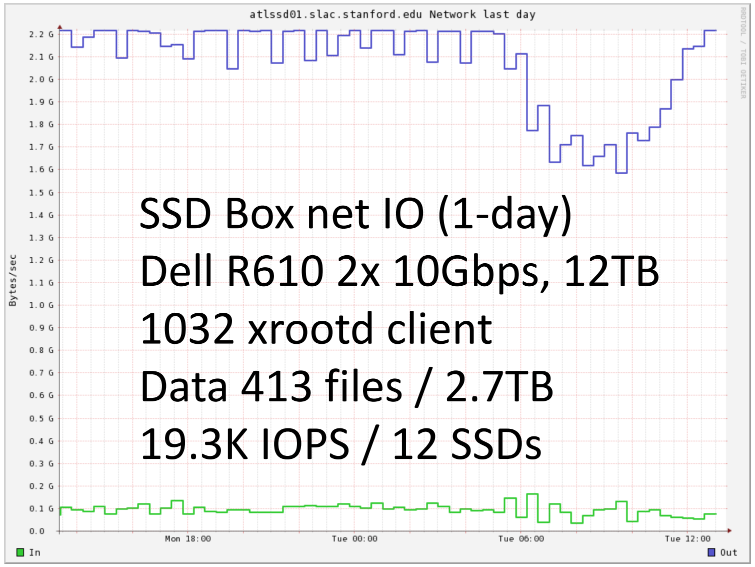
GridFTP net IO (1-day):
Sum of In and Out SLAC



SLAC SE net IO (1-day)
(including IO with SSD box)



No. of analysis jobs



SSD Box net IO (1-day)
Dell R610 2x 10Gbps, 12TB
1032 xrootd client
Data 413 files / 2.7TB
19.3K IOPS / 12 SSDs

Tuning Network Parameters for FAX

- GridFTPs are dedicated for WAN transfer
  - Buffer size, etc. well tuned for high latency WAN
  - Large IO and transfer block size
  - Multiple TCP streams
- With FAX, batch nodes won't be dedicated for WAN transfer
  - Auto tuning on batch nodes are required
    - ➢ Q: Any LAN related tuning?
  - Small IO block size, single TCP stream for remote direct IO
  - Good read ahead algorithm is important, otherwise:

    Assuming average 8KB read size:
    - 0.2ms (LAN) : < 8 * 1/0.2ms = 40MB/s
    - 20ms (WAN) : < 8 * 1/20ms = 0.4MB/s (extreme case)
- About RTT

  In general, RTT is proportional to geographic distance, but not always:
  - SLAC -> AGLT2, BU, MWT2 ~ 80ms
  - SLAC -> BNL ~70ms

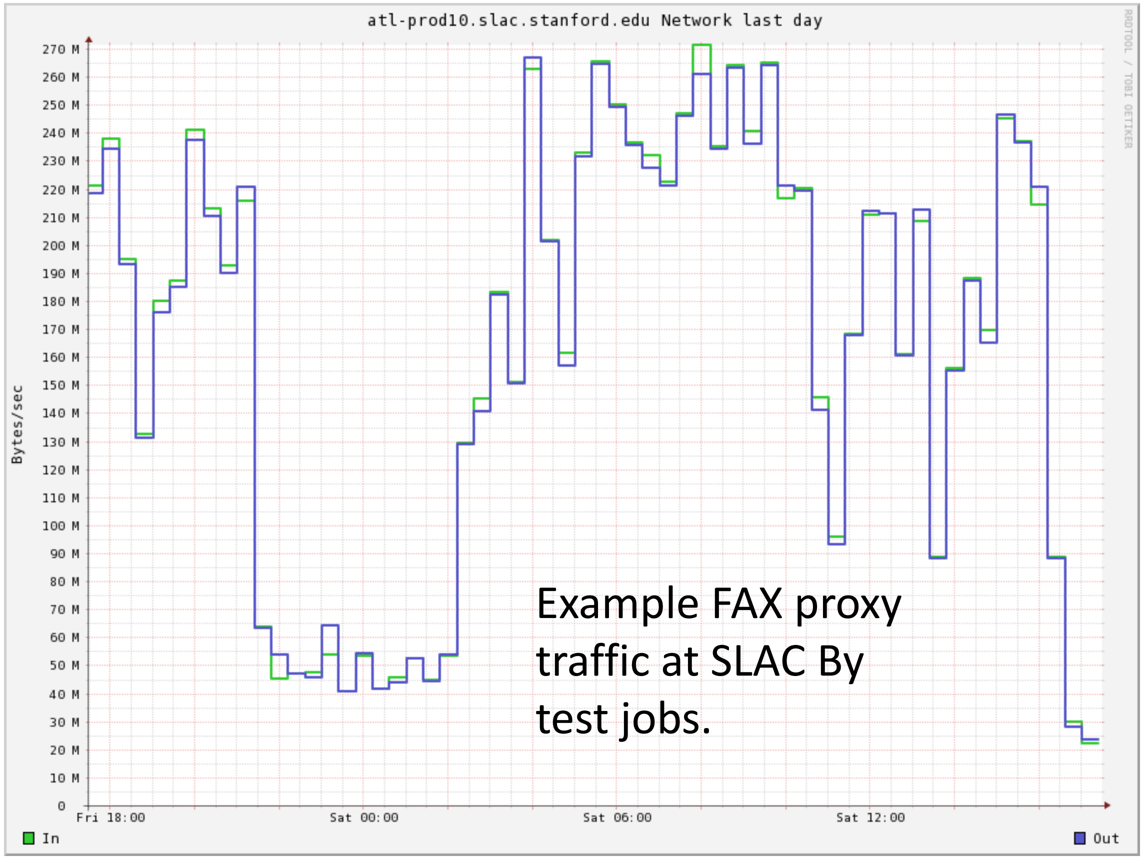
Site Models

Expose SE to remote client

- Data fly between client and storage data server, better performance
- requires N2N and GSI handled by storage element
  - GSI needs to be implemented to prevent circumvention
- Example: DPM xrootd door; dCache xrootd door at AGLT2 and MWT2
- Likely need helper (xrootd/cmsd pairs) to join FAX

Proxy between SE and remote client

- Proxy is easy to setup
  - N2N and GSI handled by Proxy (more N2N optimization work to do)
  - It may limit the performance, but also be a natural throttling mechanism
  - Expandable with Proxy cluster (BNL, SLAC).
- Proxy is useful if data servers are behind firewall
- Reverse proxy is useful if batch nodes have no outbound TCP connection
  - What is a reverse proxy
  - Large RTT between reverse proxy and FAX sites, So readv() passage is important (Xrootd release 4.x)



Example FAX proxy traffic at SLAC By test jobs.

All FAX sites need to prepare SE for some direct (remote) IO

- Small IO block, random/sparse
- Large # of open files hanging around, etc.
- Adequate WAN bandwidth
- Operational experience

Above the Site Level

Improve Redirection Algorithm

- Cost awareness redirection? There are Pros and Cons
  - Cost matrix measure the past, average but not current cost
    - Cost may includes RTT, SE/net performance, bandwidth/congestions, etc.
  - LAMBDA project propose research on real time identification of most capable sites, avoid bad choice
- Efficiency of the Redirection Network:
  - Topology, Parameters, Algorithm, real-time info

Optimal data pre-placement

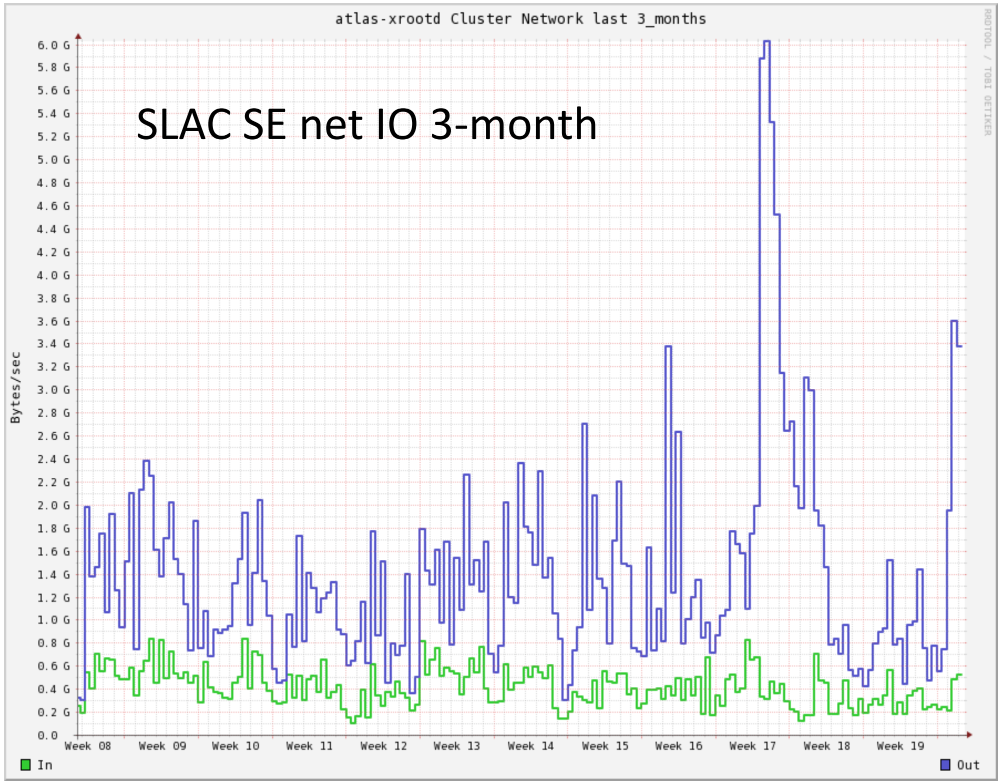
- Need to consider site storage and network capacity/performance
- Nearby sties’ CPU and network capacity
- Cost to nearby sites

- Spread files in a (hot) dataset to multiple sites? DDM nightmare

WAN Load by FAX

Rough estimation on how much load FAX will put on WAN

- Start with SLAC's internal network load: RTT ~ 0.2ms
- Think of this as the demand from jobs
- Q: what if RTT = 2ms or 20ms? -- for all are remote jobs, 50% remote jobs, etc.
  - E.g. 40% 0.2ms + 30% 10ms + 20% 20ms + 5% 40ms + 5% 80ms



- Each bin is a day (daily avg.)
- On average days, 1-1.5GB/s
- Much large than GridFTP IOs
- But there are many >2GB/s days

What can FAX offer to compensate the WAN cost?

- More slots to reduce waiting/pending time
- In theory FAX reduce job failure rate caused by missing files
- Caching mechanism (TTree, etc) may help

What about those >2GB/s bins (days)

- Are sites ready for this?
- Is data locality aware scheduling still necessary?
- and those 6GB/s bins?