

SDO data storage and transfer.

Planning longterm solutions.

1

- This is a discussion, so I don't want to talk alone. Share you ideas and impressions.
- We sent a mail about it a few weeks ago, maybe it was too vague.
- Everyone just finished installing their system. We don't want to change everything tomorrow but start planning for the future.
- So here I will try to give some ideas of mine to trigger the discussion.

Discussion Points

- 5 months of experience with the real data.
- What is the state now (netDRMS, JMD, VSO)
- Problems and difficulties current and encountered.
- What can we do about it?
- Day to day fixes vs. Longterm solutions.

2

- Acknowledge the work of Stanford, Igor and Joe. We use their soft graciously.
- The NetDRMS is full of good ideas (keywords in DB)
- European institutes have not contributed much yet. Are we ready to jump in and work on it?
- We are a chain, what one does has an impact on the others, so we need to plan and work together.
- 3 catogeries of problem that I have noticed :

Problem and difficulties (I)

- Retrieving data is hard!
- Scientists must be "formed" just to retrieve a single image
- `export_as_fits` → Not meant for users
- Lookdata → Complex query language and interface
- Fuse filesystem → Complex setup
- VSO → maybe the easiest to use

3

- Example of trying to get a set of 3 hours of consecutive AIA data. Did not work :
- Lookdata too slow and timeout
- `export_as_fits` is bugged, fits files were bad
- How other institutes are doing working with data ?
- SAO export every file as fits

Problem and difficulties (II)

- NetDRMS is still unstable.
- Hard to set-up.
- sum_svc has many bugs, developed to run @ Stanford, lacks features, documentation?
- In case of crash of the sums database, all data is lost
- When the netDRMS is down, the data is useless.

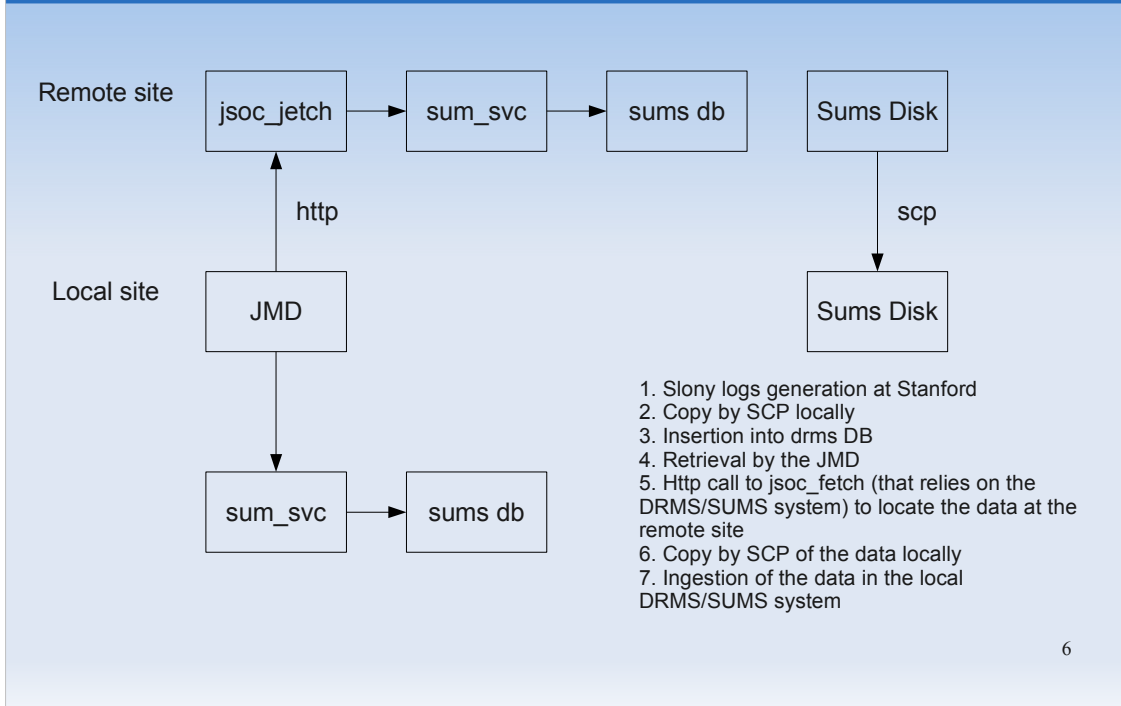
4

- sum_svc going crazy using GB of RAM and 100% of cpu
- sum_svc has sum_put, but no sum_delete
- Update of the sum_partn_avail by sum_rm
- Servers names hard encoded, filenames to home directory of Phil
- sums cannot run on different server than drms
- When crash at SAO, hard to recover. Records falsely reported online.
- export_as_fits bugs

Problem and difficulties (III)

- NetDRMS interfacing to other programs
- It 's a complex task → leads to time/effort/bugs tradeoffs (VSO, JMD, Soteria, ...)

Example : Data transfer



- In 5 month every module has failed at 1 point or the other
- EIT and Stereo data transfer use rsync and ftp, very simple
- Because huge amount of data, we cannot accept long down time, we need more robustness
- When ROB is failing → UCLAN downloads from SAO, when ROB is back up, UCLAN and ROB download from SAO
- Tried to download hmi_test.ic_45s but all sunum went into pending
- How much time every day you spent monitoring if everything is ok
-

What can we do ? (I)

- Help Stanford improving the NetDRMS!
- Make the NetDRMS more open source :
 - Bug tracking system
 - Implication in the design
 - More flexibility

7

- It is already somewhat open source : sources availables, versionning system, webserver, mailing list
- Make it a more explicit open source project model
- Allow people outside of Stanford contributing to the code
- The NetDRMS is a fork of the JSOC software meant to run at remote sites → could/should be more simple
- Most people don't need all it's possibilities (Tape storage)
- Too rigid (2 DB, filesystems under /)

What can we do ? (II)

- Start the design from a blank canvas.
- What do we really need?
- Using our experience and the developpments already made.
- With better use of the existing capabilities of filesystems, DB, ...
- Is it worth it to start all over?

8

- Sums made for easy storage to tape → but we don't have it
- Filesystem is a db of data localisation, especially built for that purpose
- Separation of data and meta-data
- Keywords change that often ? Maybe easier to update keywords...

What I would like

- Data integrity
 - Direct access to the data (keywords in the fits headers)
 - Comprehensible file names/path
aia.lev1/2010/07/01/H0100/aia.lev1.date.time.wavelength.fits
- Robustness of the data transfer mechanism
- Simplification of the whole system
- Less time on the system administration = more time on the research :)

9

- Simple : ftp vs. Ssh, rsync, ...
- 1 DB vs 2 DB

Conclusions

- What features would you want ?
- What are the priorities ?
- Make a plan for the future.
- Who can take time for this ?
- What are your skills ?