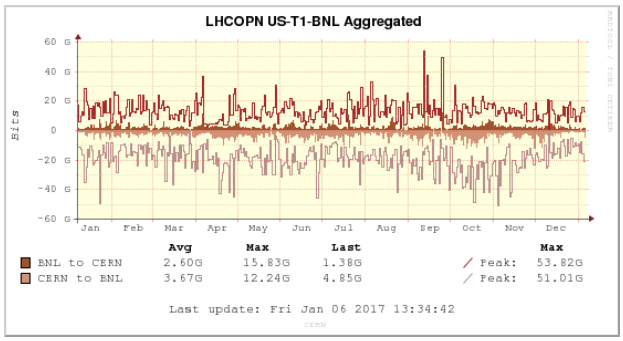
ATLAS Networking pre-GDB input

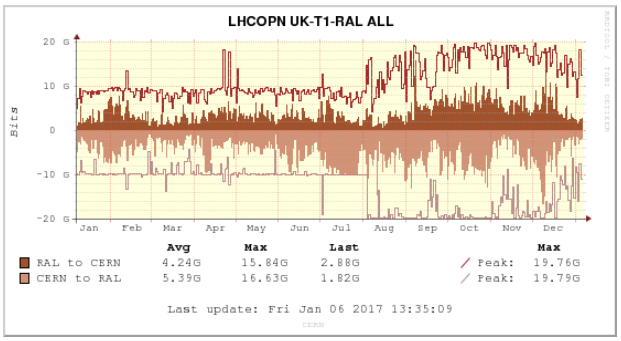
https://indico.cern.ch/event/571501/

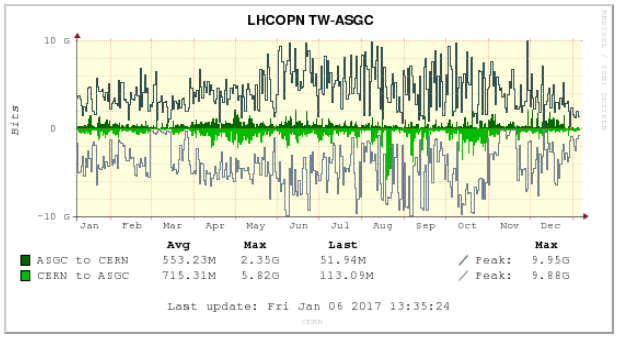
1. Network capacities used and anticipated by the experiments in the LHCOPN [Tier-0, Tier-1 Network] in the short-term (rest of Run2)

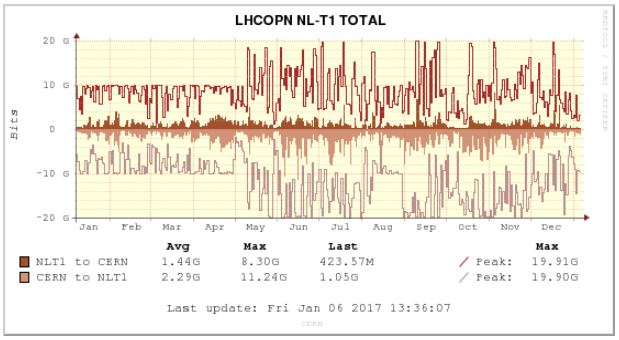
Overall Data Distribution mesh (LHCOPN, LHCONE, and all the Site-to-site connections): during 2016 ATLAS data distribution was stable around 20GB/s, peaking at 50GB/s. Tier-0 data taking does not really put a sensible contribution per-se, but we have seen effects in particular related to automatic processing of the data taking (e.g. derivations).
Overall, for the rest of Run2, we do not foresee a huge increase in the network traffic, but we foresee an increase in the usage of remote I/O (e.g. for storage-less or cache sites).

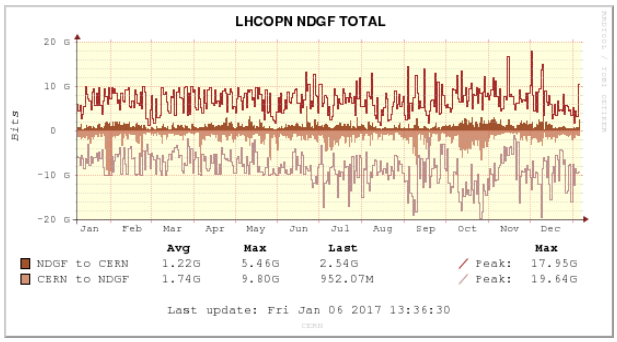
Tier-0 <-> Tier1s. LHCOPN plots below :

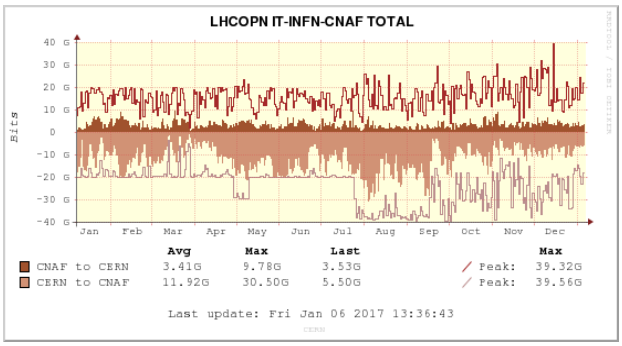


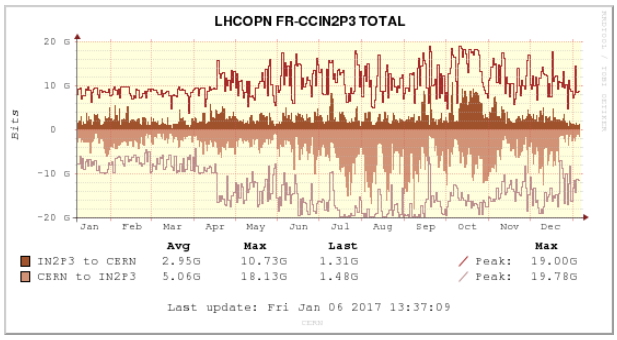


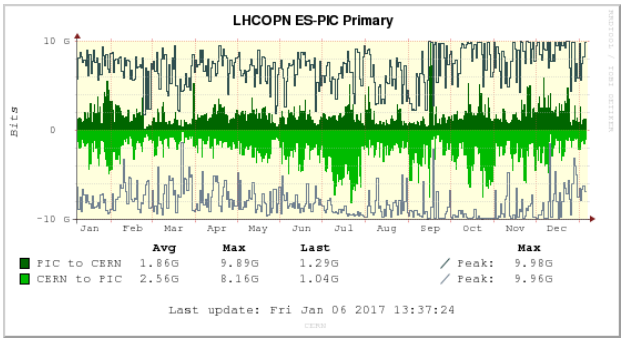


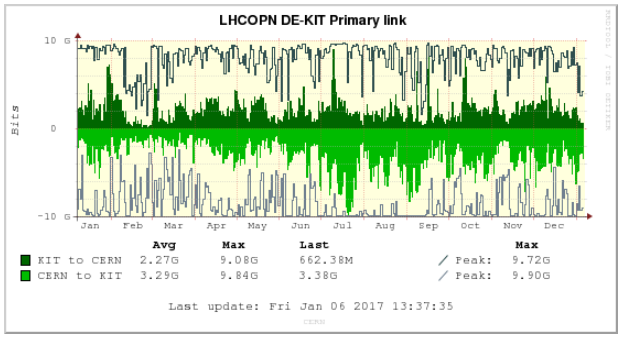


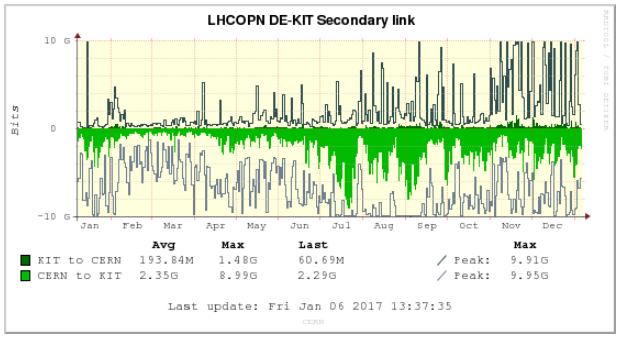


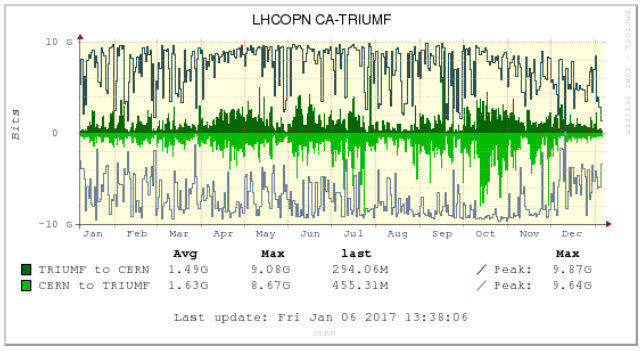












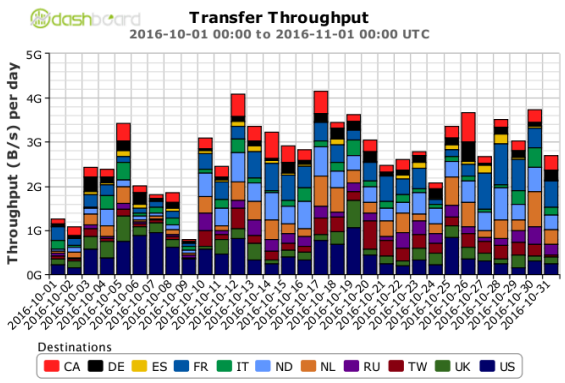
Saturation observed on CERN to T1 for some T1s (e.g. RAL). Only 3 T1s (PIC, Taiwan, Triumf) with only 10 Gb links. Probably enough for the rest of run2

Side remark : The CNAF plot is strange. They have a 40 Gb link and use a large part of the bandwidth !?!?

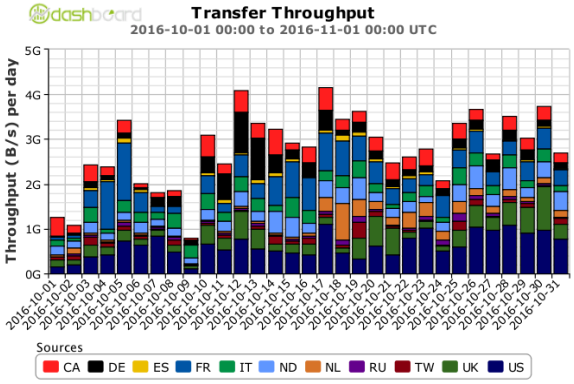
Tier1s <-> Tier1s

Plot from October :

Destination :



Source :



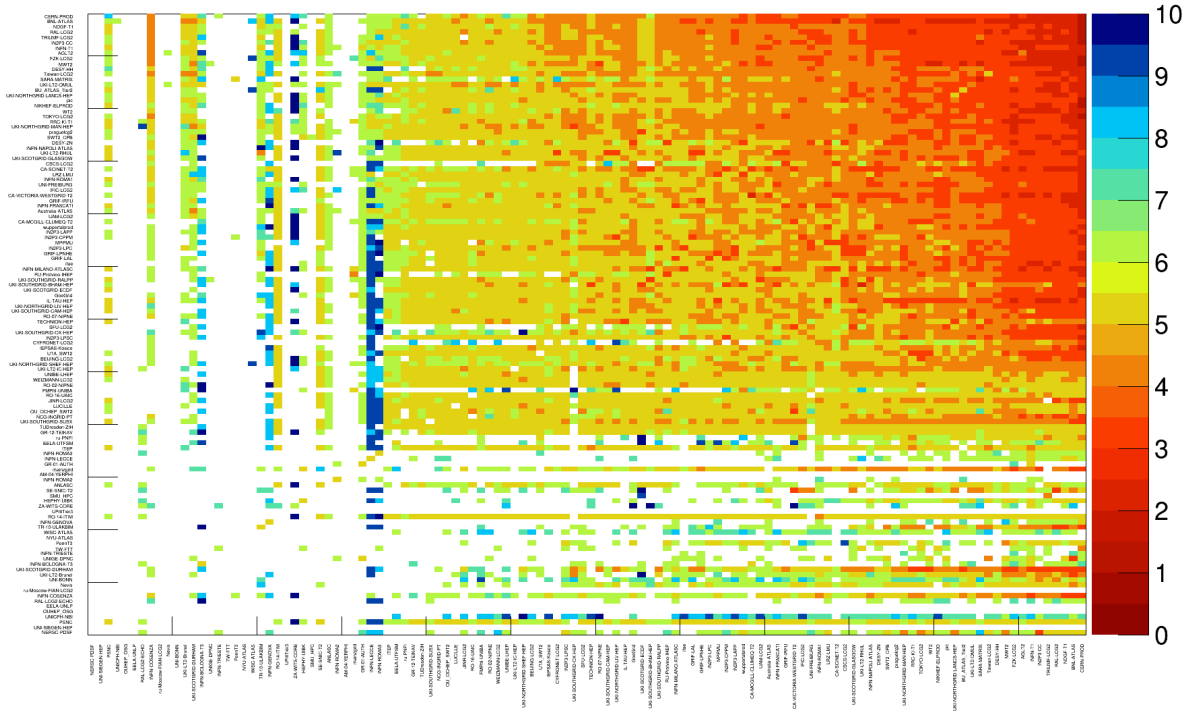
For some T1s (e.g. BNL), regularly around 1 GB/s

2. Details on the volumes transferred and anticipated with decomposition into specific geographical areas (EU, US, Asia).

ATLAS do not plan to have hardcoded geographical restrictions:
ATLAS have already dynamic data distribution and job brokerage based on Networking measurements (perfSonar - rtt/latency, last Nhours/days throughput , etc).

- We take this opportunity to say that e.g. the data of Perfsonar are not integrated enough with ATLAS, e.g. why packet loss is not in the DDM dashboard? We are aware they are in the new infrastructure, but still we don't have the same functionalities from there.
- ATLAS uses Network matrix (closeness) to optimize the job brokering

AGIS closeness

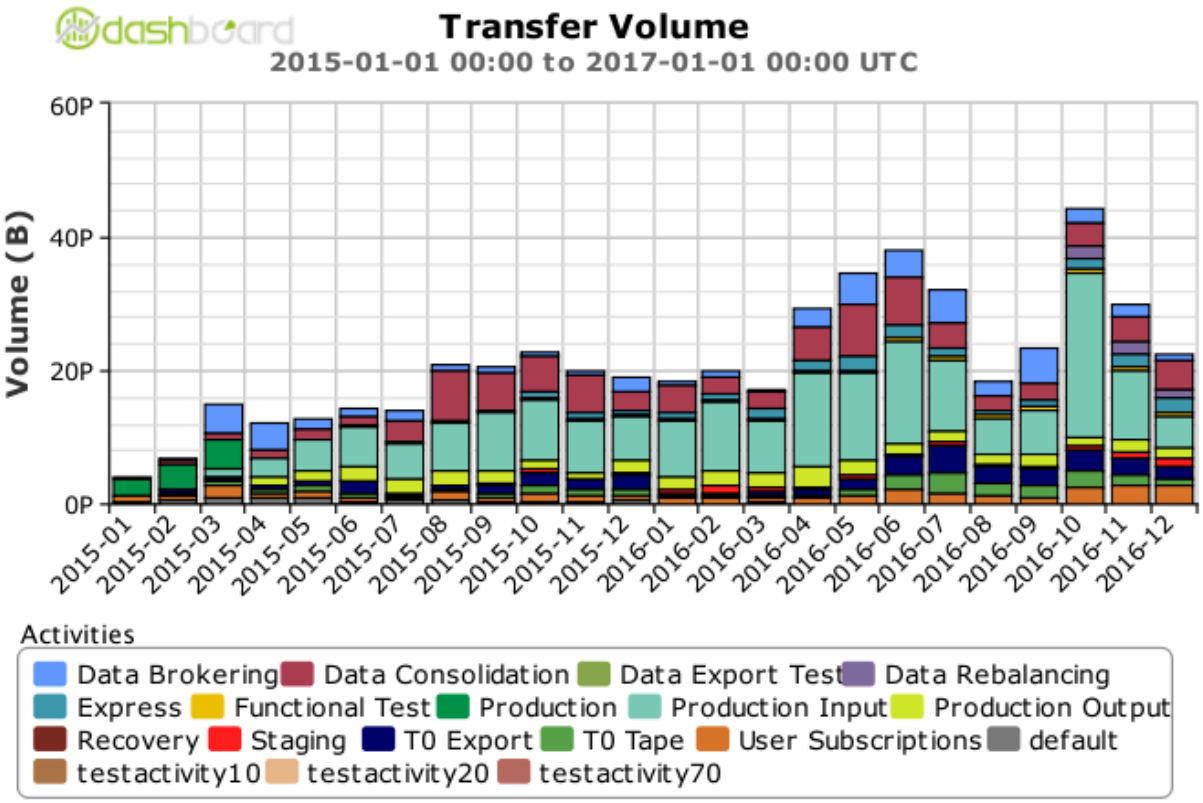


Closeness matrix for November 2016

Closeness matrix :

- Take the maximum throughput over 1 hour on every links on a period of one month.
- Put the number in a matrix that ranges from 0 (1 TB/seconds) to 10 (0.1 kB/seconds) (log scale)

And here is a split by activity Tier-0/ProdInput/DataConsolid/ etc.
Basic split is between pre-placed data (from DDM) and data moved to optimize workflows.



Performance of transoceanic links (n.b. We cannot be precise - this is what ATLAS sees):
Observed performance from DDM point of view:

1) During 2016, ATLAS transferred:
    a) 52 PB FROM US/CA clouds (1.5GB/s average)
    b) 406 PB TO US/CA (12.8GB/s average)
2) Looking at peaks in traffic for last week (with 10 mins bin granularity):
    a) 21 GB/s TO US/CA
    b) 3.3 GB/s FROM US/CA
3) Comparing the maximal throughput from FTS with measured values in Rucio, the link seems to be saturated in direction to US/CA most of the time.

3. Medium term view on the network capacities required in LHCOPN and LHCONE [Tier-1, Tier-2, Tier-3 Network] (Run 3)

During Run3 we foresee to collect 5 times more luminosity than Run2, trigger rate should stay at 1KHz.
Implication for network.

- Luminosity x5 - richer samples: WAN proportional to physics samples to be processed (~factor of 5, mostly for data).

- Cpu requirements x3 (flat budget) - expecting required inputs to be x3 larger (mostly for MC)
- Data growth faster than disk (under flat budget): this implies smaller replication factor, which implies more network usage.
- We are already exploiting use of federated storage and remote access: sites need to dimension internally themselves to cope with also remote data access. How much? Not clear how to give a definite number, because it depends on how well does it work. Order of 10-20% for now.
- Few points:
    - Pledge: up to now it was not needed to pledge network resources because sites have demonstrated to be able to properly dimension disk/cpu/network to be optimally exploited.
    - To give examples of ATLAS sites: HW: 6PB storage 12k slots, network usage observed at the local storage is about 4GB/s .
- Our understanding is that currently for most countries, typically, we have 10Gb/s to 100Gb/s WAN due to NREN - no direct implication for flat budget: from ATLAS we would benefit of big storage sites with 100Gb/s which will be able to run all the ATLAS workflows.

4. Anticipated growth in the transferred volumes and planned changes in the data management workflows (changes in the ways data will be transferred, data management patterns, data access patterns).

a. Network capacities planning process and review; guidance for NRENs and sites.
b. Network capabilities that are missing today that the experiments would like to have?

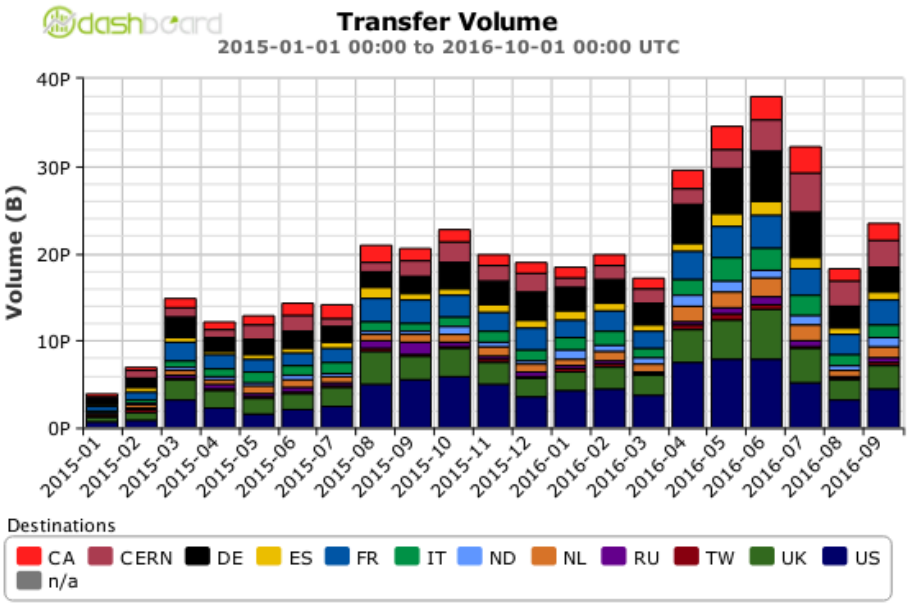
Some of the changes under discussion for the next years:

- Overall, an approximation of the data growth for Run3 can be between of a factor 5 and 10.
- Event Service moving into production, streaming events in/out WNs from/to few(3?) ObjectStores
- WAN exploiting, after having “understood” it (exploit more, but only after we are able to compute ETT for set of files with e.g. 95% accuracy)
- Exploit SDN: e.g. situation where we know in advance we need to move X files (events/part of events) from A to B, or we might be able to minimize the processing time of specific tasks if needed.
    - Related to the 3 points above we are working on the Event Streaming Service concept: access data in-place when possible through WAN moving data in small chunks (only the data/branches needed, and in parcels representing a few minutes of processing) instead of moving the whole file, and with the transfers isolated from payload processing such that processing is insensitive to WAN latency. (n.b. For this point we are also working on the ATLAS SW to make it less sensitive to WAN latencies and more robust to data retrieval hiccups)

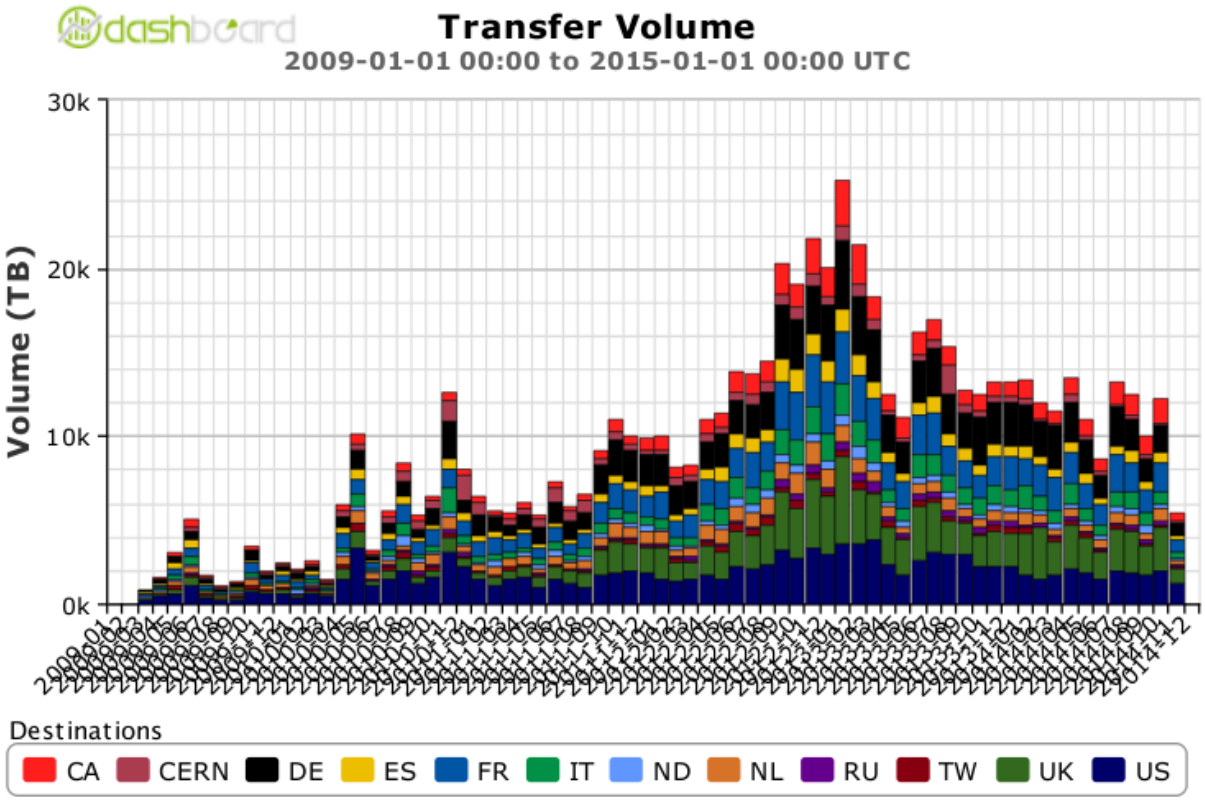
- Use more TAPEs to store and re-read data more often. This would mean having more disk caches, and the need of more network bandwidth (for these data -- the one which will be only on tape, recalled when needed-- we foresee a factor 2 increase in network usage)

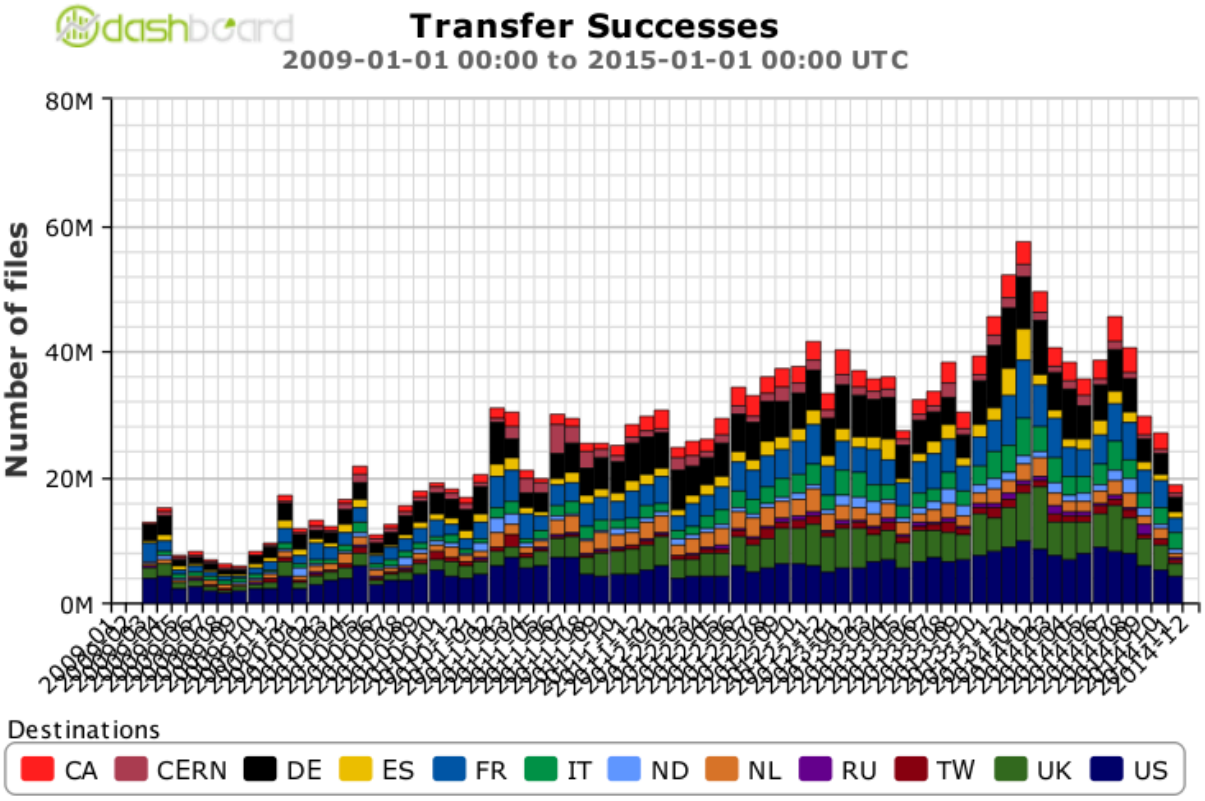
BACKUP

Network Use in ATLAS



Historic plots hereafter (1st Jan 2009 to 1st Jan 2015) :




dashboard
Transfer Successes
2009-01-01 00:00 to 2015-01-01 00:00 UTC
Number of files
80M
60M
40M
20M
0M
Destinations
CA
CERN
DE
ES
FR
IT
ND
NL
RU
TW
UK
US