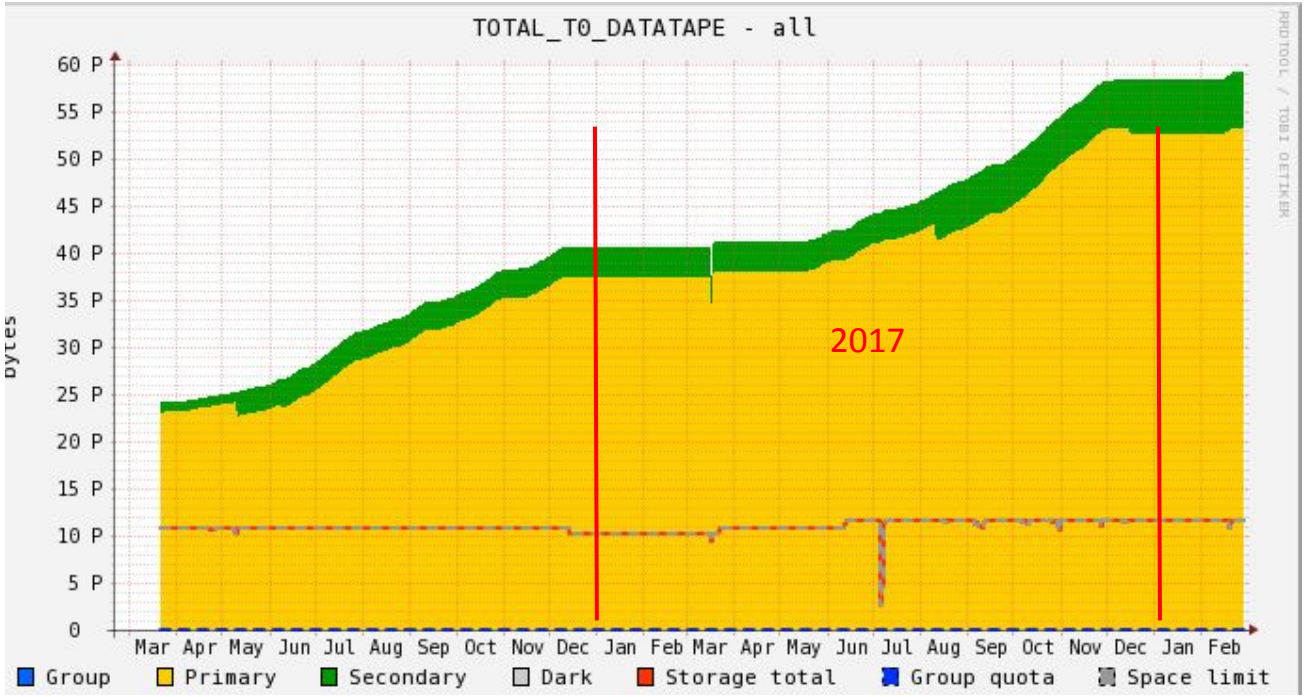
ATLAS Tape Systems 2018

DDM overview

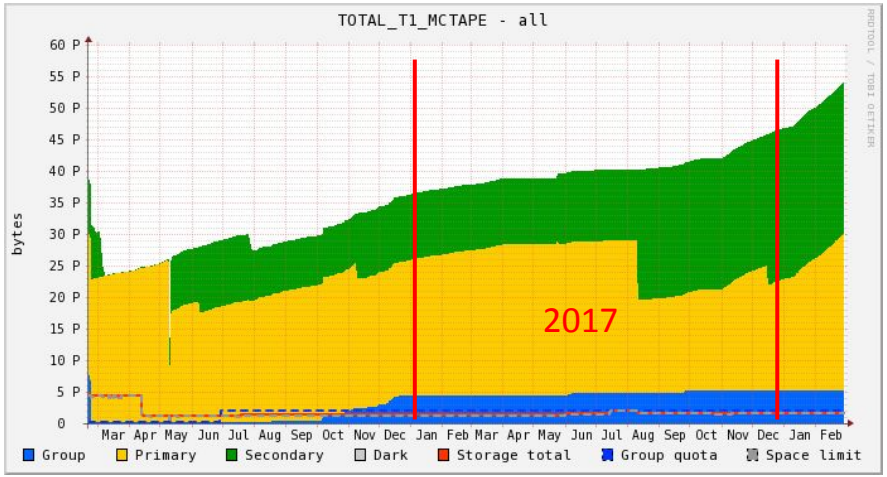
Tomas.Javurek@cern.ch
on behalf of the DDM team

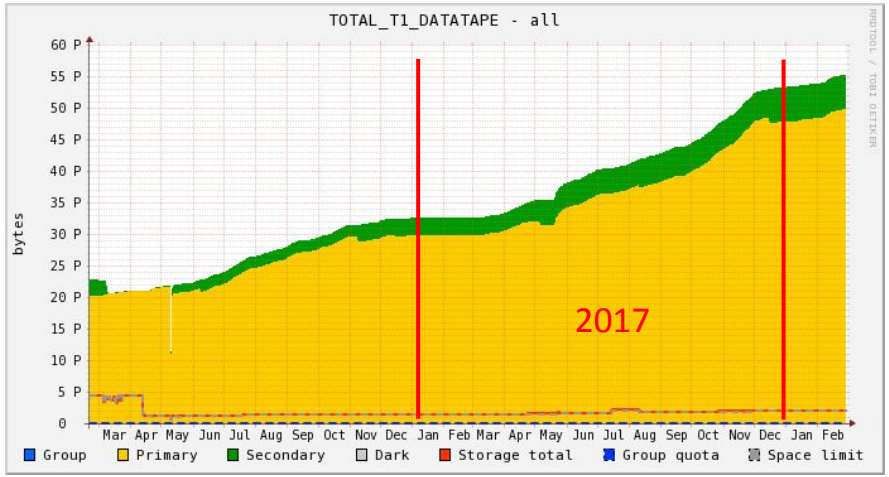
Current usage



- T0 tapes
- increase by 18 PB

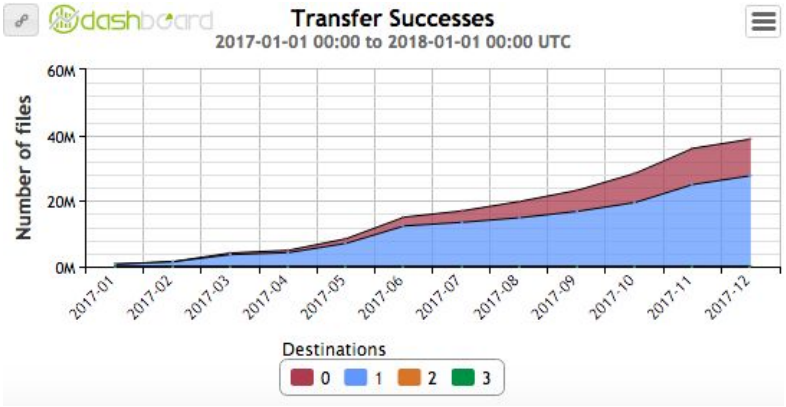
Current usage

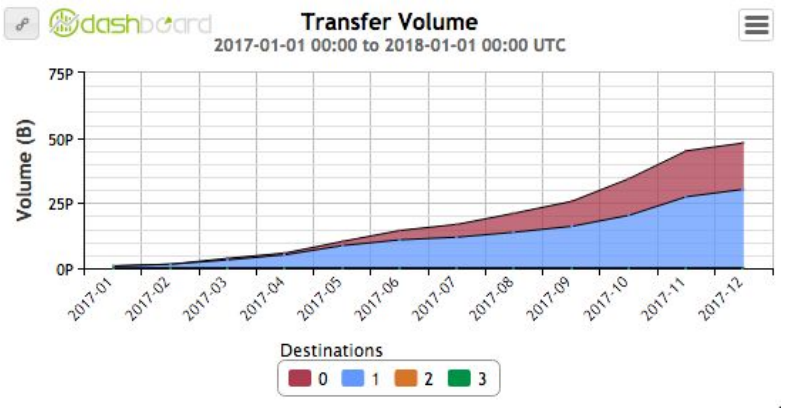




- T1 tapes
- increase by ~30 PB in 2017

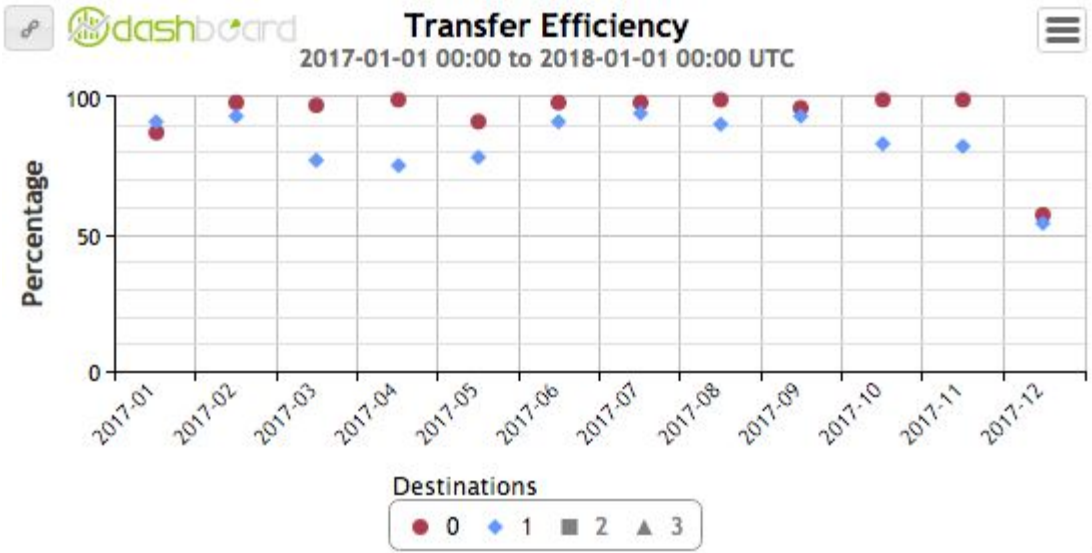
Transfers to Tapes in 2018





- agree with what is registered at Tapes

Transfers to Tapes



Writing to tape buffers is efficient.

Estimation of data written to tapes in 2018

- Longer period of data taking -> multiplication factor 1.2
- higher trigger rate (1.4kHz) -> 1.2
- same applies to reco products and mc
  - should be even more as we want to push more

=> 1.2*1.2*50 = 72 PB

Pledges

Tier	Federation	Pledge Type	ATLAS	% of Req.
Tier 0	CH-CERN	Tape (Tbytes)	105000	112%
Tier 1	CA-TRIUMF	Tape (Tbytes)	21100	11%
Tier 1	FR-CCIN2P3	Tape (Tbytes)	22000	11%
Tier 1	DE-KIT	Tape (Tbytes)	24375	13%
Tier 1	IT-INFN-CNAF	Tape (Tbytes)	17550	9%
Tier 1	NL-T1	Tape (Tbytes)	13500	7%
Tier 1	NDGF	Tape (Tbytes)	9950	5%
Tier 1	NRC-KI-T1	Tape (Tbytes)	5700	3%
Tier 1	ES-PIC	Tape (Tbytes)	8951	5%
Tier 1	TW-ASGC	Tape (Tbytes)	0	0%
Tier 1	UK-T1-RAL	Tape (Tbytes)	24375	13%
Tier 1	US-T1-BNL	Tape (Tbytes)	49000	25%

	pledge	now	2018
T0	105 PB	60 PB	90 PB
T1	197 PB	110 PB	152 PB

Discussion - site reports

- families:
  - https://docs.google.com/presentation/d/1UuE1TtfprT6x67texCfYcB7mwEqs9Duk6iFT8mmboq0/edit?usp=sharing
- bulk requests to tapse - after talk of BNL

FZK

1) will the pledge for your TAPE system be fulfilled in 2018?
Yes

2) do you plan and repacking in 2018?
Yes (more LTO -> T10K and maybe C -> D)

3) how do you organise the tape families?
For each tape-family there is a storage pool defined in TSM.
A storage pool is a collection of tape cartridges.

4) what is preferable way of accessing the tapes (big bulks, small bulks)?
Currently we are limited to 2k(??) simultaneous active recall requests, which limits the efficiency of recalls. Having another 3k requests already sent to our SRM will help us to keep the a steady 2k requests active on our tape system. Having tens of thousands of requests sent to us will not improve recall efficiency and throughput.

5) do you have any technical issues or improvements that could be share with other sites?
-

PIC

1) will the pledge for your TAPE system be fulfilled in 2018

Yes. The 2018 needs have been already tendered and we are waiting for receiving the material in time (indeed, this includes 2 new tape drives T10KD). Since Q3/2013, we even add +10% tape resources above WLCG pledges to account for operational inefficiencies (data deleted subject to repack & recycling).

2) do you plan and repacking in 2018?

Yes. We monitor this, and in particular we have a weekly digest that summarises all of the tapes subject to repack and recycle. We take actions as soon as there is a non-negligible amount of space to be recalled. Typically, several recycling/repacking campaigns are run along the year (more for CMS, though...).

3) how do you organise the tape families?

For this, we do believe in ATLAS to tell us in advance which tape families we need to create. As soon as we receive the request, we create them in our storage. This requires ATLAS to tell us in advance (well before the data is being created or transferred to PIC). We do have generic file families, so in case that the new FFs are not created in time, the data lands into these ones. The current view of tape (file) families and occupation is attached to this email.

4) what is preferable way of accessing the tapes (big bulks, small bulks)?

For read access, the requests should come in big bulks, if possible. Of course, we do have some limitations in the Enstore queue (50k requests, if I am not mistaken). We did some work to optimise, even more, our system during the ATLAS 2017 tape tests. For writes, we do buffer files in the disk pools and then we send bulk files to write to tapes. We do prefer big files to be stored into the tape system (I know this has been an issue with ATLAS in the past, not so sure how this is today at PIC. I think it improved somehow). Needless to say that we are a 5% Tier-1 in WLCG, and our tape performance in 2017 was higher than the ATLAS (and CMS) expectations. But we always try to improve even more. Esther is now working towards improving the tape re-mounts, which will also make the system more performant.

5) do you have any technical issues or improvements that could be share with other sites?

We typically report at the HEPIX workshops and specialised forums. We use those opportunities to share our tape experience with other centers, and in particular to communicate on those aspects.For the next HEPIX we will show the tape re-mounts and tape buffer optimisations. These are the topics in which we are working these days. We do meet monthly with FNAL, the developers of Enstore, to be up-to-date with the system.

Lyon

about the TAPE system at Lyon (CC-IN2P3) from our TAPE experts
- will the pledge for your TAPE system be fulfilled in 2018?
Yes we already purchase 1400 tapes in advance at the end of 2017.
- do you plan and repacking in 2018?
We plan to repack some tape for consolidating space.
We do not plan to change media type this year.
- How do you organize the tape families?
For atlas, we have a dedicate subsystems. Files are organized on 4 different class of services according the file size :

0 - 64MB : Small file, migrated on T10K-D SPORT
64MB - 512MB : Medium files, migrated on T10K-D
512MB - 2GB : Big files, migrated on T10K-D
2GB - 4TB : Huge files, migrated on T10K-D

We have 2 tapes familly (with the 4 COS) :
- atlas_dache (11,4 PB)
- atlas_dcache_archive (2,4 PB)

- What is preferable way of accessing the tapes (big bulks, small bulks)?
Big bulks are always preferable because it take full advantage of treqs queuing system.
- Do you have any technical issues or improvements that could be share with other sites?
We discovered that many tapes written in summer 2016 make errors at reading. This issue is due to on ore more faulty T10K-D drives that has made silent corruption at write (E.V.: the corruption is mechanical) .
When reading back the tapes, some files are not readable anymore. About 5 to 10 files (~20GB) can't be read.
We asked to oracle a data recovery but the tape ribbon is damaged.
So the files are irremediably lost.
For atlas, about 15 tapes presenting this issue has been identified at this time. (total lost file ~100)
But the errors only appears when reading back the full tape. We can not presume which tapes has been impacted, but we suppose that other tape still present the same issues.
We will only knows the impacted files after the full repack of all the tapes. The next repack campaign will not start before 2019 or 2020.

TRIUMF

1) will the pledge for your TAPE system be fulfilled in 2018?
Yes. currently published tape capacity is 18PB, a new library is under purchase. Total capacity will be 30PB this summer.

2) do you plan and repacking in 2018?
We've done LTO5 - LTO7 tape data migration in the end of 2017 and early 2018, 2.5PB data, 2600 LTO5 tape cartridges, not file/media damage. No repacking plan, but we do have a plan to buy a new library and will move current LTO6/7 tape cartridges into the new library, new media will be LTO8.

3) how do you organize the tape families?
We group tape data by tape family, by LFN, data type and datasets.

4) what is preferable way of accessing the tapes (big bulks, small bulks)?
Bulk staging is preferred, 5k-30k range is good. Would be even better if there is a link we can check how much data staging requests in the queue ahead of time, in case we have maintenance work to get done.

5) do you have any technical issues or improvements that could be share with other sites?
We have our own tape system with features on write and read requests reordering to get best tape performance, also have special setup to minimize hsm interface load.