Table of Contents

Archival Site Survey Results

Results reformatted automatically from indivudual survey responses. As a consequence, formatting is a bit rough.

What is the site name?

ASGC	
BNL	BNL
CCIN2P3	CC IN2P3
CERN	CERN
FNAL	FNAL
GSDC-KISTI	KR-KISTI-GSDC-01 (WLCG Entry), KISTI_GSDC (ALICE Entry)
INFN-CNAF	INFN-T1
JINR	T1-JINR
KIT-GridKa	FZK_LCG2
NDGF	NDGF-T1
NIKHEF-SARA	
NRC-KI	
PIC	PIC
STFC-RAL	RAL-LCG2
Triumf	TRIUMF

Which endpoint URLs do your archival systems expose?

ASGC	
BNL	srm://dcsrm.usatlas.bnl.gov
CCIN2P3	For Atlas / CMS / LHC (dCache) srm://ccsrm.in2p3.fr For Alice (XRootd) root://ccxrdralice.in2p3.fr:1096/
CERN	srm://srm-Experiment.cern.ch and root://castorExperiment.cern.ch, where Experiment is one of alice, atlas, cms, lhcb, public. For ALICE, only the root endpoint is available.
FNAL	srm://cmssrm.fnal.gov srm://cmsdca2.fnal.gov
GSDC-KISTI	root://xht1201.sdfarm.kr:1094 (XRootD)
INFN-CNAF	srm://storm-fe.cr.cnaf.infn.it for atlas; srm://storm-fe-cms.cr.cnaf.infn.it for cms; srm://storm-fe-lhcb.cr.cnaf.infn.it for lhcb; root://alice-xrootd-tsm.cr.cnaf.infn.it for alice
JINR	se-hd02-mss.jinr-t1.ru
KIT-GridKa	srm:{atlassrm-fzk,cmssrm-kit,lhcbsrm-kit}.gridka.de
NDGF	srm://srm.ndgf.org https://dav.ndgf.org and root://ftp1.ndgf.org
NIKHEF-SARA	srm.grid.sara.nl
NRC-KI	
PIC	srm://srm.pic.es, and xrootd doors, which are typically accessed via xrootd redirectors. (experiment = atlas, cms, or lhcb)
STFC-RAL	srm-${experiment}.gridpp.rl.ac.uk ; root://${various}.gridpp.rl.ac.uk/
Triumf	srm://triumf.ca/atlas/tape/

How is tape storage selected for a write (choice of endpoint, specification of a spacetoken, namespace prefix).

ASGC	
BNL	We only serve ATLAS.
CCIN2P3	in dCache, we have different spacetokens used to select tapes pools. The XRootd endpoint for alice is tape only
CERN	Choice of endpoint and specification of a spacetoken. In some configurations (e.g. for ATLAS), the namespace prefix implies a choice of spacetoken.
FNAL	metadata tags in the namespace directory tree
GSDC-KISTI	We have different endpoint between disk and tape storage: root://xht1201.sdfarm.kr (for tape); root://alice-t1-xrdr01.sdfarm.kr (for disk)
INFN-CNAF	By endpoint and path. GPFS policies define the mapping between paths and tape pools.
JINR	The whole storage is dedicated to single client CMS
KIT-GridKa	dCache and xrootd both archive into the same tape storage backend.
NDGF	Path or spacetoken
NIKHEF-SARA	spacetoken
NRC-KI	
PIC	Depends on the VO. ATLAS and LHCb, selected by space token. CMS depends on namespace areas.
STFC-RAL	It's done by path. The path maps to a file class which maps to a tape pool.
Triumf	endpoint same as above, ATLAS only

What limits should clients respect?

ASGC	
BNL	Please send bulk requests, we prefer to do pre-staging
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	For the moment, any limits are not enforced to client side (experiment).
INFN-CNAF	
JINR	Only physical tape limits (all tapes)
KIT-GridKa	
NDGF	
NIKHEF-SARA	dCache related
NRC-KI	
PIC	For read access, the requests should come in big bulks, if possible
STFC-RAL	(RAL's response here is pretty much like CERN's because we also run CASTOR)
Triumf	We do accept any kind of recalls, but prefer bulk requests. We purposely delay requests to get processed in order to get bulk requests

---> Max number of outstanding requests in number of files or data volume

ASGC	
BNL	In theory, unlimited. We observed max record of 245k requests, and processed smoothly. Took about 5 days to complete. For my own reference: STAR 2016-09-28.

CCIN2P3	> 100 K
CERN	infinite
FNAL	queue depth is ~15k, if full, clients retry
GSDC-KISTI	
INFN-CNAF	We experienced a queue of 100000 files to recall via SRM (StoRM) that was correctly handled. We do not know a limit with xrootd.
JINR	
KIT-GridKa	dCache assigns flush and stage tasks to pools, which all have an upper limit for concurrent active tasks, usually 2k. Requests beyond that are queued. For xrootd the limit is a total of 3200 concurrent flushing and staging tasks.
NDGF	In theory unlimited, but not tested above a few million
NIKHEF-SARA	dCache related. The tape system has no limit, but we recommend <= 1000 reqs.
NRC-KI	
PIC	No limit. But if the requests are coming through SRM, there is a limit of 15k requests per VO.
STFC-RAL	infinite
Triumf	No exactly number, there was one peak number more than hundreds of k during ATLAS test, no problem for us

---> Max submission rate for recalls or queries

ASGC	
BNL	
CCIN2P3	
CERN	Up to about 10 Hz.
FNAL	no limit
GSDC-KISTI	
INFN-CNAF	Up to 15 Hz
JINR	
KIT-GridKa	
NDGF	No limit on rate.
NIKHEF-SARA	dCache related. The tape system has no upper limit
NRC-KI	
PIC	
STFC-RAL	~10Hz
Triumf	

---> Min/Max bulk request size (srmBringOnline or equivalent) in files or data volume

ASGC	
BNL	Min: prefer no less than 1000. Max is unlimited, in theory. Try sending us as many as possible.
CCIN2P3	Files: Max : 100 K Min : 1 K Volume > 100 TB
CERN	1 to 1000. The upper limit is not hard but being SRM based on XML, larger counts make requests handling heavier.
FNAL	no limit
GSDC-KISTI	
INFN-CNAF	

	We can support up to 100000 files to recall in bulk. In terms of data size, a single bulk can fill up the size of disk buffer in front of tapes; this size is different for the 4 LHC VOs.
JINR	
KIT-GridKa	
NDGF	As much as fits in an SRM request, 1k - 10k I think it is.
NIKHEF-SARA	dCache related. The tape system has no limit but we recommend <= 20 TB
NRC-KI	
PIC	We allow a minimum of 1 request to unlimited, but we recommend to group the requests >= 1k.
STFC-RAL	1-1000. Maybe we should check if anyone has submitted 1000
Triumf	5k-30k (per session) is good, even 1k is ok, but few at a time is not welcomed few TB - 200TB

Should clients back off under certain circumstances?

ASGC	
BNL	
CCIN2P3	
CERN	YES
FNAL	
GSDC-KISTI	In case of maintenance, we may request the clients to pause their actions
INFN-CNAF	Yes
JINR	
KIT-GridKa	SRM feature with dCache: A limit can be set for every request type, including srm-bring-online (10k by default). Once more requests are accumulated, SRM will block and return "overloaded" error. For xrootd, there is no such feature that would reckognise an overload situation. If a file cannot be staged from tape, xrootd will fail on each subsequent request immediately.
NDGF	Yeah
NIKHEF-SARA	yes during maintanance of the dCache or tape system
NRC-KI	
PIC	Yes. The system is dimension to work fine taking into account the PIC Tier-1 size and the experiments expectations from the site. If the load is very high, then problems might appear.
STFC-RAL	YES
Triumf	Ideally no, our HSM is able to handle hundreds of k requests without load problem, however there is a hard limit from disk buffer size, we don't use any extra disk buffer for tape operations, the disk buffer that tape use is also used for ATLAS((dcache hsm pools), space is limited to that, so realistically speaking, few TB-200TB a day is good enough, though can be reached to 500TB

---> How is this signalled to client?

ASGC	
BNL	
CCIN2P3	SRM_INTERNAL_ERROR at request level and SRM_FILE_BUSY at file level returned by SRM. Stalling client by xrootd.
CERN	SRM_INTERNAL_ERROR at request level and SRM_FILE_BUSY at file level returned by SRM. Stalling client by xrootd.
FNAL	
GSDC-KISTI	We inform via directly e-mail to experiment management
INFN-CNAF	

	In case of massive repeated errors on almost all the requests, the tape administrators may ask users to stop their activity.
JINR	
KIT-GridKa	srm-bring-online and accessing a file will fail.
NDGF	According to SRM standard signalling
NIKHEF-SARA	dCache/SRM specific
NRC-KI	
PIC	If the requests are coming through SRM, refuses occur when the requests reach 15k. This is a SRM limit, to protect the service. Reaching the limit is an exceptional situation, that rarely happens.
STFC-RAL	SRM_INTERNAL_ERROR at request level and SRM_FILE_BUSY at file level returned by SRM. Stalling client by xrootd. Or through admin processes: sysadmins communicating with experiments
Triumf	through SRM

---> For which operations?

ASGC	
BNL	
CCIN2P3	For SRM, any synchronous operation. For xrootd, any operation can be stalled by the server.
CERN	For SRM, any synchronous operation. For xrootd, any operation can be stalled by the server.
FNAL	
GSDC-KISTI	Maintenance e.g. urgent security update or required upgrade of systems: xrootd clusters or backend filesystems...
INFN-CNAF	For all operations.
JINR	
KIT-GridKa	srm-bring-online / open
NDGF	The ones giving error
NIKHEF-SARA	recalls, stores and metadata ops.
NRC-KI	
PIC	If 15k is reached through SRM, read/writes are affected.
STFC-RAL	Potentially all
Triumf	Depends on ATLAS how launch and check requests

Is it advantageous to group requests by a particular criterion (e.g. tape family, date)?

ASGC	
BNL	Yes, we constantly seeing repeat mounts in ATLAS tapes. A tape might be re-mounted again within less than 15 minutes, over 20 remounts a day, which really should be avoided. We try not to delay any request, but we may have to implement a way to delay processing such frequent mounted tapes.
CCIN2P3	
CERN	YES absolutely. This helps to avoid requesting same tape over-and-over again in a short period of time.
FNAL	
GSDC-KISTI	
INFN-CNAF	

	Yes. Grouping requests by tape familiy would reduce the mounts of the same volume in a short period.
JINR	
KIT-GridKa	In theory, yes, that would be advantageous. But we cannot guarantee that it will stay in that order or grouping. There is only a loose chronological order.
NDGF	Not really
NIKHEF-SARA	No, the system will optimize recalls
NRC-KI	
PIC	For writing, the disk servers are configured to send bunch of files per tape family to also reduce the tape re-mounts. For reads this helps to reduce the number of tape re-mounts, since datasets are stored in tapes according to predefined tape families.
STFC-RAL	YES
Triumf	tape family grouped by datatype, dataset, and date

---> What criterion?

ASGC	
BNL	Please, do the pre-staging, send us all requests once, and send them fast. This is the best practice to handle sequential access media.
CCIN2P3	Group requests by creation time in dcache: Data written in the same time are grouped on the same tapes. Reading data according creation time wil help to reduce mount/dimount of the sames tapes.
CERN	Simply grouping as many requests as possible should be enough
FNAL	
GSDC-KISTI	
INFN-CNAF	Grouping as many requests as possible
JINR	
KIT-GridKa	Timestamps would be most useful, since tape families don't necessarily match what the VOs use as classification.
NDGF	Roughly grouped by time might help a bit, but not much
NIKHEF-SARA	n.a.
NRC-KI	
PIC	By tape family.
STFC-RAL	Files recalled together should be on the same tape. For WLCG and GridPP VOs, users are expected to do this themselves; for facilities (e.g. climate), another service "above" CASTOR will aggregate files into reasonably sized chunks that can (and will) be recalled together.
Triumf	data will wait for at least 45 hours within a dataset if the dataset size not exceed a tape capacity, or will be processed when the dataset size > a single tape capacity, different datasets will be packed together by project, data type, for example: data17_900GeV, mc15_5TeV, further grouped by datatype, datatape, mctape etc..

Can you handle priority requests?

ASGC	
BNL	Yes we can
CCIN2P3	No, tape archive is shared between all VO and we not handle priority. But all recall request coming from dCache and Xrootd take benefit of our tape queuing system (TREQS : Tape Request Scheduler)

CERN	YES
FNAL	
GSDC-KISTI	No. So far we have not been asked for any priority related matters. It is because we only support one experiment (ALICE) for now.
INFN-CNAF	Not at users/groups of users level. We can handle priority for VOs.
JINR	No.
KIT-GridKa	No.
NDGF	No
NIKHEF-SARA	No
NRC-KI	
PIC	Yes, Enstore allow to modify the priority of a specific request
STFC-RAL	YES
Triumf	quite often tape is quiet, no need yet, also no priority flat in ATLAS operations, can be implemented if particular circumstance is identified

---> How is this requested?

ASGC	
BNL	Any tape that has at least 1 high priority flagged request, will be placed in front of the queue. Prioritized tape will wait and get the next available drive. All priority tapes will be processed in the same selected logic: by demand, FIFO, or LIFO.
CCIN2P3	
CERN	Selected groups of users might have higher priorities than others. However, this is balanced between experiments. Contact Castor.Support@cernNOSPAMPLEASE.ch.
FNAL	
GSDC-KISTI	
INFN-CNAF	The administrators can manually assign more or less tape drives to specific VOs. We are working on a solution of an orchestrator, integrated in our tape system (GEMSS, IBM Spectrum Protect), that would dinamically assign drives to VOs on the basis of their requests and previous usage. This would optimize the usage of shared tape drives (all our production tape drives are shared among the experiments).
JINR	
KIT-GridKa	
NDGF	If this is a strong request from one of our VOs, we would look at implementing it
NIKHEF-SARA	n.a.
NRC-KI	
PIC	This is only available for admin purposes. VOs are typically using the tape system with the same priority level.
STFC-RAL	In practice administratively. Typically, to prioritise recalls for a given user/VO, we will allocate more drives. If a lot of data needs to be recalled (petabytes), CASTOR admins can help reschedule recalls to be more efficient
Triumf	

Are there any unsupported or partially supported operations (e.g. pinning) ?

ASGC	
BNL	
CCIN2P3	
CERN	Pinning is not supported.

FNAL	
GSDC-KISTI	
INFN-CNAF	We support pinning.
JINR	All supported by dCache.
KIT-GridKa	All features that dCache and xrootd support natively should work for GridKa, too.
NDGF	Full support for pinning etc
NIKHEF-SARA	
NRC-KI	
PIC	Pinning is supported.
STFC-RAL	No pinning
Triumf	we can always pin or restore a file through dcache HSM interface, bulk or individual

What timeouts do you recommend?

ASGC	
BNL	do not set timeouts! All staging are synchronized calls, every file will be processed sooner or later. No need to re-submit. Multiple repeated requests may be transferred multiple times, as we do not drop any requests.
CCIN2P3	Timeout should be increase to 24h in order to benefit of large bulk recall
CERN	At least 1 day to overcome interventions.
FNAL	
GSDC-KISTI	at least 100 seconds? including the read-out time from cartridge and stage-in to tape buffer.
INFN-CNAF	It is recommended to put no timeouts.
JINR	No timeouts. Tasks beyond reasonable time are handled manually.
KIT-GridKa	
NDGF	At least a day
NIKHEF-SARA	this is dependand to the length of the tape request queue. The tape system at SURFsara is shared with all VOs and local users.
NRC-KI	
PIC	We recommend using high timeouts (more than 48h) or don't use timeouts. The requests will be processed sooner or later. Duplicated requests generate to process the requests multiple times causing unnecessary overload.
STFC-RAL	Our experience is that the client times out before we do. 24 hours is suggested for recalls.
Triumf	All tape requests will be served . 24 hours or even longer if large amounts data requested, few hours if requests are small number

Do you have hardcoded or default timeouts?

ASGC	
BNL	Our tape storage system do not have any timeout. Our system tracks every steps for every file, all requests will be processed eventually.
CCIN2P3	Default dcache timeout per request on tape pool is 14400 s (4h)
CERN	NO
FNAL	
GSDC-KISTI	No, we don't.
INFN-CNAF	Default timeout of backend system (GEMSS) is 4 days, but it can be changed by administrators.
JINR	No timeouts
KIT-GridKa	Yes, we have default timeouts with dCache for flushing and staging of at least 24 hours

	(may be larger on request). No timeouts are enforced with xrootd.
NDGF	Not user-visible, internal timeouts will be handled by internal retries
NIKHEF-SARA	soft timeout: 4 h, hard timeout: 24 h
NRC-KI	
PIC	Timeout for the HSM script is 864000 seconds (10 days). SRM timeouts and FTS timeouts are typically that high.
STFC-RAL	No.
Triumf	No

Can you provide total sum of data stored by VO in the archive to 100TB accuracy?

ASGC	
BNL	yes. We can provide total sum in byte accuracy. So we can convert it to any format.
CCIN2P3	Yes, accounting value is computed in byte
CERN	YES
FNAL	yes
GSDC-KISTI	2,983 TB / 3,200 TB (93%)
INFN-CNAF	Yes
JINR	6300 TB
KIT-GridKa	Yes
NDGF	Yes
NIKHEF-SARA	yes
NRC-KI	
PIC	Yes.
STFC-RAL	YES; we can do much more accurate than that: in our earlier/current information provider, we can do to byte level (but it's expensive, so we do it only once every 24 hrs)
Triumf	Yes

Can you provide space occupied on tapes by VO (includes deleted data, but not yet reclaimed space) to 100TB accuracy?

ASGC	
BNL	yes
CCIN2P3	Yes, it is the same value as above.
CERN	YES
FNAL	yes
GSDC-KISTI	2,983 TB for ALICE VO
INFN-CNAF	Yes
JINR	7190 TB
KIT-GridKa	Yes
NDGF	Yes
NIKHEF-SARA	yes, through dCache
NRC-KI	
PIC	Yes.
STFC-RAL	YES
Triumf	Yes,all deleted data on tape is logcal delete

How do you allocate free tape space to VOs?

ASGC	
BNL	In HPSS, we assign free tapes to a storage class.
CCIN2P3	We monitor the storage class usages of all VOs, and we do the allocation by bunch of 50-100 tapes when a storage class goes short of tapes
CERN	By a scriot running periodically. Defined number of new tapes (usually 1) is allocated into a tape pool per VO as needed. The check is every 15 minutes.
FNAL	quatos on a common pool
GSDC-KISTI	Currently we support only one experiment (ALICE), all free tape space is allocated to ALICE VO.
INFN-CNAF	Tape manager software (IBM Spectrum Protect) allocates a new volume from a shared scratch pool.
JINR	The whole space to CMS VO.
KIT-GridKa	We do not allocate space on tape for any VO.
NDGF	We keep track of space in hsminstances in our internal wiki, then set particular hsminstances read-only as tape space runs out
NIKHEF-SARA	we don't
NRC-KI	
PIC	After a tape purchase, we allocate the new free space to the VO according the pledge for that year. We monitor if a VO is close to exhaust the number of assigned tapes. We also have a tape pool with free tapes, used for tape migrations or if some experiment need some extra space. We also add +10% of pledges for LHC experiments, to ease tape operations (repacks).
STFC-RAL	There's an "infinite" tape pool of free tapes. Free tapes are essentially allocated as needed but we then track usage administratively, like keep an eye on when we (or the VO) need to buy more tapes, whether a VO is using too many tapes, etc.
Triumf	Through info publish

What is the frequency with which you run repack operations to reclaim space on tapes after data deletion?

ASGC	
BNL	Due to the limited drive resources, we only do massive repack as needed.
CCIN2P3	We run repack manually when tape filling is bellow 70-80 %. We also run repack when we suspect tape to generate errors on recall
CERN	Once / week on selected tape pools. On experiment tape pools, only major repack campaigns are done automatically. Exception is if experiment perform deletion campaign in which case, we can recover the space sooner.
FNAL	frequently with CMS
GSDC-KISTI	
INFN-CNAF	We do space reclamation after scheduled deletion campaigns by experiments. Otherwise, we reclaim space when we notice a certain number of volumes full and with a percentage of occupancy less than 80%.
JINR	After massive deletion.
KIT-GridKa	We have defined a threshold for "tape occupancy", which will trigger reclamation per tape.
NDGF	Continuous based on percentage used on a particular tape
NIKHEF-SARA	daily
NRC-KI	

PIC	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).
STFC-RAL	Depends on user requirements and deletion rates. Weekly.
Triumf	we do repack when space is needed or media upgrade, just done LTO5-> lto7 migration this early year, 2600 LTO5 tapes, 2.5PB(on tape since 2012) migrated to LTO7 tapes at new site, no data lost

Recommendation 1

ASGC	
BNL	do pre-stage. To prevent the data lost from un-necessary excessive accessing. We need to use the tool in the right way, the way how it was designed for.
CCIN2P3	Run prestaging (ie SRM BRINGONLINE) on large dataset with an huge timeout value.
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	It would be useful to know the expected data flow during the year, in terms of writing on and reading from tape. This would help to plan the purchase of the needed number of tapes to fulfil pledges and to optimize the usage of resources shared with other experiments. Important writing or reading activities should be announced, as sometimes happens.
JINR	
KIT-GridKa	
NDGF	Request reads in bulk, trickle-feeding requests (a handful every 10 minutes) means we have to implement longer waiting period before we have a reasonable batch of requests to send to retrieval.
NIKHEF-SARA	
NRC-KI	
PIC	Send read requests in bulks, to help for data pre-stage.
STFC-RAL	
Triumf	Bulk requests by dataset, you will get your data quicker, our tape system pick the tapes has most requests first

---> Information required by users to follow advice

ASGC	
BNL	Tape is designed for archiving, not for random access. Use it cautiously, or it may eventually be damaged by all means. Our intention is to protect the data.
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	

STFC-RAL	For WLCG/GridPP-approved experiemnts/VOs, we have a weekly meeting (using Vidyo) which it is highly recommended they join. We have mailing lists and, within the T1, lists of contacts for every VO.
Triumf	

Recommendation 2

ASGC	
BNL	
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	In general the correct usage of tape resources is to write mainly custodial data, limiting as much as possible to write data that will be removed, since intense repack is a resource-consuming activity that could limit the performance of production. Anyway, it is recommended to write non-custodial data on dedicated storage pools, in order to limit the amount of data to repack after deletions.
JINR	
KIT-GridKa	
NDGF	Use SRM, that's the only featureful tape protocol. The others will technically work, but will not have any intelligence.
NIKHEF-SARA	
NRC-KI	
PIC	It could be desirable a better description on how to dimension the disk buffers for the LHC experiments. We suspect that the disk buffers in PIC are over-dimensioned (disk buffer = disk in front of tape for reads/writes), since we are a bit conservative and want to get rid of operational troubles.
STFC-RAL	
Triumf	Let us know how you use tape data, then we can tweak our tape data packing policy, you will get fast readback and data safe protection, tape has mounting limits

Should a client stop submitting recalls if the available buffer space reaches a threshold?

ASGC	
BNL	No, because client doesn't know anything about the buffer space and its status in our dCache. They should not back off. Note - Here, I assume the buffer refers to the buffer area in dCache (dCache tape read pools), not the disk buffer of HPSS itself. That is, the data is already staged from tape to the frontend dCache.
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	Yes. Generally, when an high threashold is reached, GEMSS triggers the GPFS garbage collector that removes from buffer files starting fron the older ones. It can happen that the file system is full (till the garbage collector high threshold) of files that are written on buffer and not yet migrated on tape, e.g. in case the writing rate on disk is higher than the migration rate on tape. The same happens if the buffer is full (till the garbage collector high threshold) of recalled files all pinned until a date in the future. In both of these cases, or in a combination of them, the garbage collector can not remove any file.

JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	
STFC-RAL	To the first approximation, no. We automatically garbage collect the least recently used data on the cache. Our policy is to make the cache big enough that this isn't a problem (ATLAS and CMS have 640 TB each). If a user needs to recall 100s of TBs, then they would usually talk to the operators anyway. The cache is shared between ingest and recall - we have the ability to separate it if needed.
Triumf	

---> How can a client determine the buffer used and free space?

ASGC	
BNL	
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	SRM publishes these metrics for each storage area.
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	
STFC-RAL	They can't. We hold the buffer at 70% full - if it goes higher than that it means we either have a garbage collection problem or an unexpected tape robot outage.
Triumf	

---> What is the threshold (high water mark)?

ASGC	
BNL	
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	In this moment there is no threshold set for clients at CNAF, but it is desirable. The high threshold should be higher (e.g. 1% higher) than that used by garbage collector, This depends by the file system: alice 95%, atlas 89%, cms 97%, lhcb 97%.
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	

STFC-RAL	70% full
Triumf	

---> When should the client restart submission (low water mark)?

ASGC	
BNL	
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	The client should restart submission when occupation has reached a parcentage lower (e.g. 1%-2% lower) than the high threshold for garbage collector.
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	
STFC-RAL	We can't stop the client and don't have a meaningful low water mark because garbage collection acts as required to keep the cache at 70% full. As mentioned above, huge recalls should be done in collaboration with RAL admins. If we could stop the client, restart would depend on the size of the client's recall relative to the cache size and the amount of other recall/migration activity.
Triumf	

If the client does not have to back off on a full buffer, and you support pinning, how is the buffer managed?

ASGC	
BNL	We don't support pinning. The way it works here is: FTS sends bringonline command to dCache, which then pass to HPSS. Once the data is staged from HPSS to the dCache buffer space, bringonline command succeeds. Then FTS sends another "transfer" command, to transfer the file to the final destination, be it either within the same site but on a different disk area, or a remote site. Our buffer (dcache tape read pools) is always full, whenever new files come in, dcache purges files out of the buffer in a FIFO manner. So if the second FTS "transfer" command didn't come fast enough, and there are hugh amount of data staged from HPSS in a short time, the files can be purged out before transferred to the final destination.
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	When gargabe collector runs, it removes files no more pinned, starting from the older ones. If the buffer is full (till the garbage collector high threshold) of recalled files all pinned until a date in the future, the garbage collector can not remove any file.
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	

NRC-KI	
PIC	
STFC-RAL	We don't support pinning.
Triumf	

---> Is data moved from buffer to another local disk, either by the HSM or by an external agent?

ASGC	
BNL	By an external agent. As said above, it's triggered by a FTS transfer request.
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	Not automatically.
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	
STFC-RAL	Not by us. The users can manually copy data to other local disk resources, such as CASTOR d1t0 or ECHO.
Triumf	

Should any other questions appear in subsequent iterations of this survey?

ASGC	
BNL	
CCIN2P3	
CERN	
FNAL	
GSDC-KISTI	
INFN-CNAF	Just a clarification: the first question of the "buffer Management" session should be related to both recalls and migrations (now it refers to recalls only).
JINR	
KIT-GridKa	
NDGF	
NIKHEF-SARA	
NRC-KI	
PIC	
STFC-RAL	

| Triumf | |

This topic: HEPTape > Survey_Results
Topic revision: r5 - 2018-05-07 - OliverKeeble