The ATLAS Data Carousel Project

A.Klimentov, M.Lassnig and X.Zhao
on behalf of Data Carousel and Google-HEP R&D Teams

DCC meeting
November 1, 2019

BROOKHAVEN

CERN

Team Effort --- WFM SW, Rucio, DPAs, Operations, Monitoring teams, Alessandro Di Girolamo, Johannes Elmsheuser and ADC experts, CERN T0, all T1s storage and tape experts, dCache and FTS experts

Outline

- ATLAS Distributed Computing Software and Data/WFM R&D projects
- Data Carousel
- AOD/DAOD metrics
- More challenges ahead

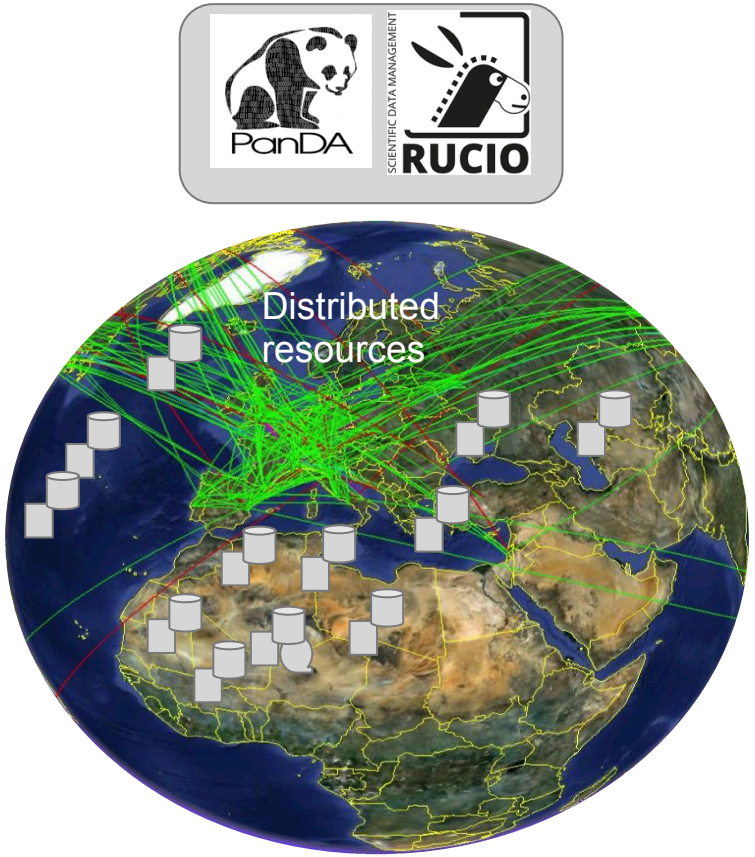
ATLAS primary distributed computing tools

Production System -

Workflow Management: "translates" physicist requests into production tasks

PanDA –

Workload Management: submission and scheduling of jobs & tasks



Rucio –

Data Management: bookkeeping and distribution of files & datasets

AGIS/CRIC-

Information System

PanDA queues and resources description

BROOKHAVEN

HL-LHC R&D Computing Projects

- HL-LHC will be a (multi) Exabyte challenge. The WLCG community needs to evaluate LHC computing model to store and manage data efficiently.
  - The technologies that will address the HL-LHC computing challenges may be applicable for other communities to manage large-scale data volume (SKA, DUNE, LSST, BELLEII, NICA, etc).
- WLCG, IRIS-HEP and experiments have launched several R&D projects to address HL-LHC data challenges :
  - ***Data Lake***. The aim is to consolidate geographically distributed data storage systems connected by fast network with low latency. The Data Lake model as an evolution of the current infrastructure bringing reduction of the storage and operational costs.
  - ***Intelligent Data Delivery Service*** (iDDS). The intelligent data delivery system will deliver events as opposed to delivering bytes. This allows an edge service to prepare data for production consumption (filtering out unnecessary events and objects), the on-disk data format to evolve independently of applications, and decrease the latency between the application and the storage.
  - ***Third Party Copy***
  - ***Google-HEP***, data placement and migration between “Hot-Cold” storage using data popularity information.
  - ***Data Carousel***

Data Carousel R&D Project

By 'data carousel', we mean an orchestration between workflow/workload management (WFMS), data management (DDM) and data archiving services whereby a bulk production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% (5%?, 10%?) of inputs onto buffer disk, such that only ~ X% of inputs are pinned on disk at any one time.

- Ultimate goal : use tape more efficient and active
  - Cycle through tape data, processing all queued jobs requiring currently staged data
    - 'Carousel engine' : job queue regulating tape staging for efficient data matching to jobs?
    - Brokerage must be globally aware of all jobs hitting tape to aggregate those using staged data
  - No pre-set target on tape throughput, instead, we focus on ***efficiently*** using the ***available*** tape capacities
    - *Introduce no or little performance penalty to tape throughput, after integrating tapes into our workflow*
    - *Improve efficiency and throughput of tape systems, by orchestrating the various components in the whole system stack, starting from better organization of writing to tapes*
    - *Solutions should scale proportionally with future growth of capacities of tape resources*
- 'Data Carousel' R&D was started in the second half of 2018 → to study the feasibility to use tape as the input to various I/O intensive workflows, such as derivation production and RAW data re-processing
  …and "tape" could be any "cold" storage

Data Carousel Project Phases

- Phase I : Tape Sites Evaluation
  - Conduct tape staging tests, understand tape system performance at sites and define primary metrics

Done

- Phase II : ProdSys2/Rucio/Facilities integration
  - Address issues found in Phase I
  - Deeper integration between workflow, workload and data management systems (ProdSys2/PanDA/Rucio), plus facilities

In Progress

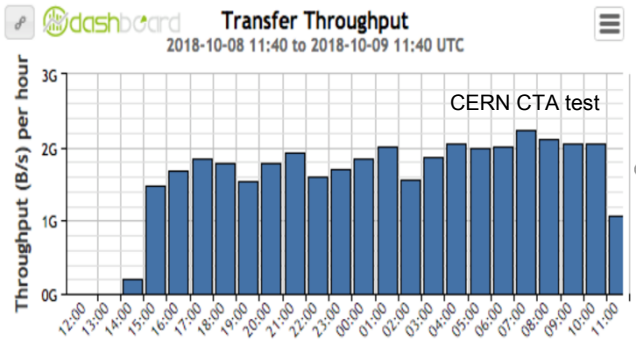
- Phase III : Run production, at scale, for selected workflows
  - Address it in cold/hot storage context

We intended to conduct an iterative data carousel exercises, and to combine them with real production campaigns, to test new ideas and reveal possible bottlenecks

Goal : to have data carousel in production for LHC Run3

Data Carousel Phase I. Jun-Nov 2018

- Established baseline measurement of current tape capacities
- All ATLAS T1s (but NRC-KI and ASGC) and CERN participated
- Overall throughput from all T1s (as of Nov, 2018) reached ~600TB/day
- CERN conducted its own Tape Archive (CTA) test, reached ~2GB/s throughput



Average Tape Throughput: throughput directly from local site tape monitoring

Stable Rucio Throughput: from rucio dashboard, over a “stable” run time

Test Average Throughput: total volume staged / total walltime of the test

Site	Tape Drives used	Average Tape (re)mounts #	Average Tape throughput	Stable Rucio throughput	Test Average throughput
BNL	31 LTO6/7	2.6	1~2.5GB/s	866MB/s	545MB/s (47TB/ day)
FZK	8 T10KC/D	>20	~400MB/s	300MB/s	286MB/s (25TB/ day)
INFN	2 T10KD	Majority tapes mounted once	277MB/s	300MB/s	255MB/s (22TB/ day)
PIC	5~6 T10KD	Some outliers (>40 times)	500MB/s	380MB/s	400MB/s (35TB/ day)
TRIUMF	11 LTO7	Very low (near 0) remounts	1.1GB/s	1GB/s	700MB/s (60TB/ day)
CCIN2P3	36 T10KD	~5.33	2.2GB/s	3GB/s	2.1GB/s (180TB/ day)
SARA-NIKHEF	10 T10KD	2.6~4.8	500~700MB/s	640MB/s	630MB/s (54TB/ day)
RAL	10 T10KD	n/a	1.6GB/s	2GB/s	1.6GB/s (138TB/ day)
NDGF	10 IBM Jaguar/ LTO-5/6 (@4 sites)	~3	200~800MB/s	500MB/s	300MB/s (26TB/ day)

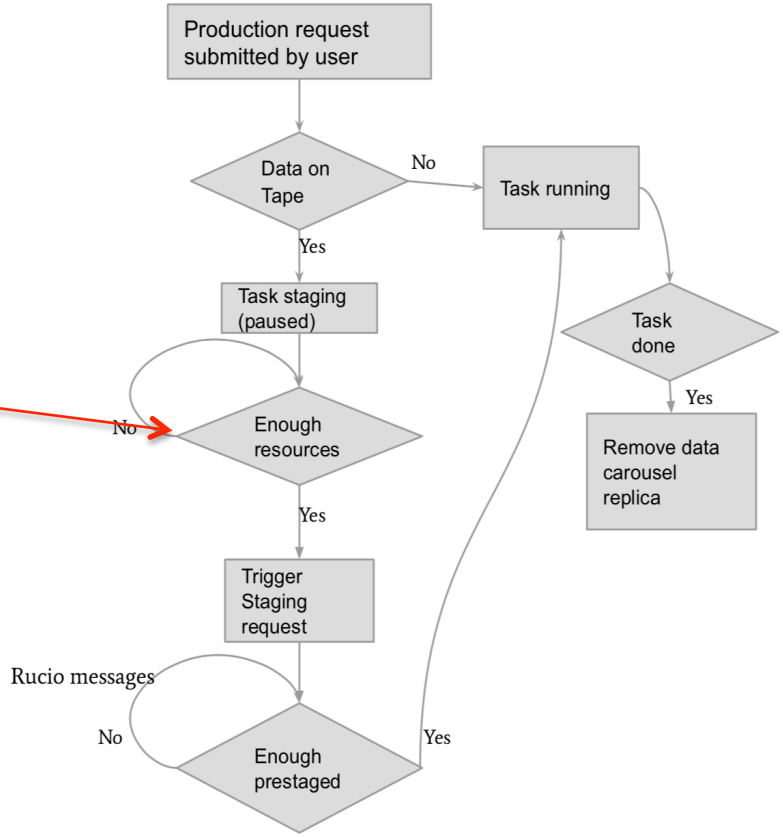
BROOKHAVEN

Data Carousel Phase I. Metrics

- Tape frontend --- a potential bottleneck for an effective tape usage
  - Identify throughput characteristics per site (tape system)
- Data organization (file placement on tape) is vital
  - Good throughput seen from sites who organize writing to tape (especially in case grouping data by datasets)
  - Usually the reason for performance difference between two sites that have similar hardware and software setup
- Define site specific I/O numbers

Data Carousel Phase II. 2019

- Use tape intensively and integrate archiving storage into ATLAS workflow
  - No more manual pre-staging campaign
  - Algorithms development for intelligent prestaging
    - Respect priorities, shares, availability of computing and storage resources…
    - Define and tune the “sliding window”
  - Define and establish Rucio/ Prodsys2 communication protocol



Data Carousel Phase II. Round2

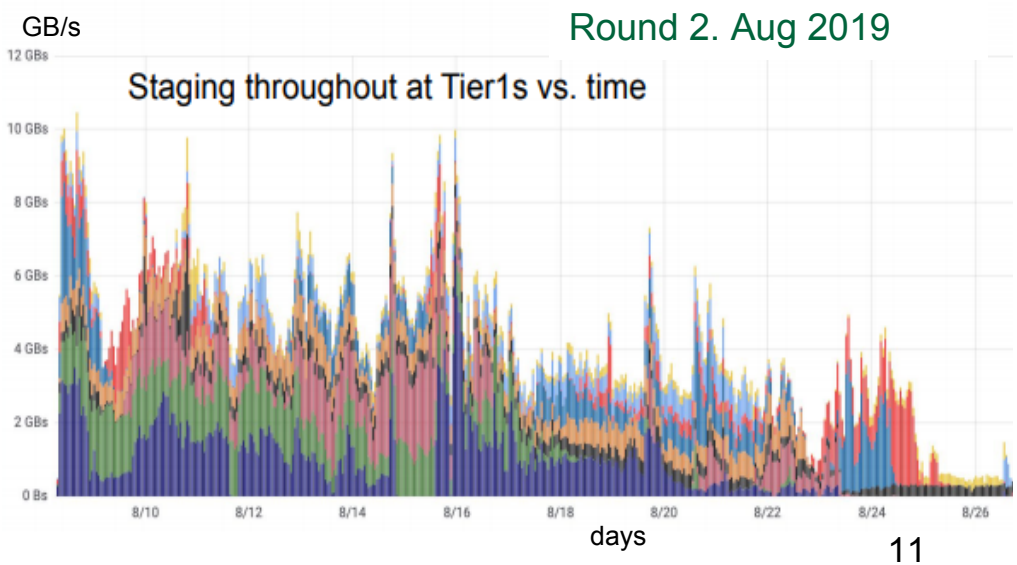
- Deeper integration of workflow/workload management (ProdSys2/JEDI/PanDA), data management (Rucio) systems and facilities
- Two rounds of data carousel exercises have been conducted :
  - the second round was combined with data reprocessing campaign
  - It took 5 days to have 70-90% data staged

We managed to finish campaign in time
(enough CPU slots)

2018 RAW RPVLL data reprocessing

- Data carousel model used, T1s (except in downtime) and T0 tape systems participated
- 238 datasets staged from tape. 6.9PB, 3.1M files, 6.4B events
- Average file size ~2GB/s

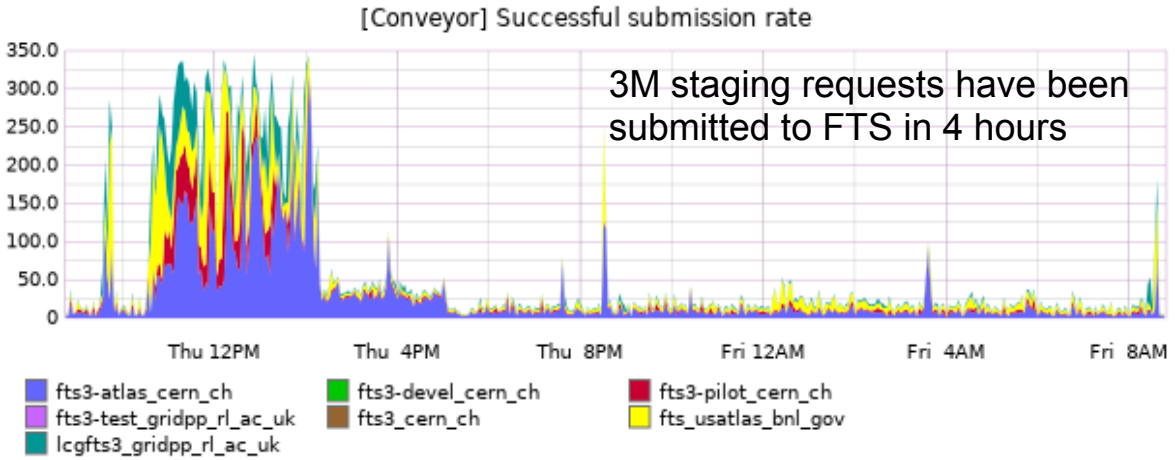
Data Carousel Phase II
Round 2. Aug 2019



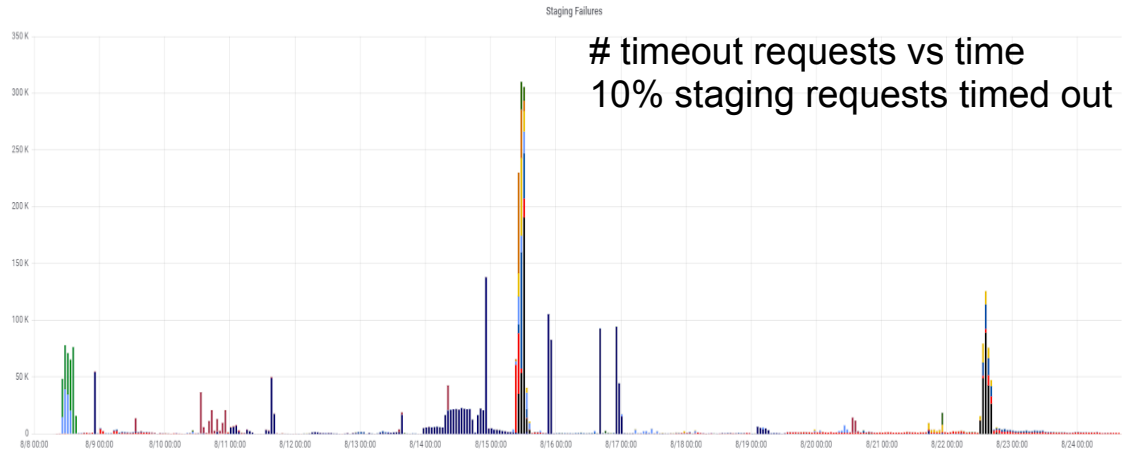
BROOKHAVEN

Data Carousel Phase II Round2. Data Staging

- Staging requests were submitted in bulk mode, but max limit per site was respected
- Evaluating (together with sites) a more intelligent scenario : staging profile per site



3M staging requests have been submitted to FTS in 4 hours



timeout requests vs time
10% staging requests timed out

BROOKHAVEN

Data Carousel Phase II Round2. Data Transfer

Staged files are purged from disk buffer (DATATAPE), before they can be transferred to the final destination

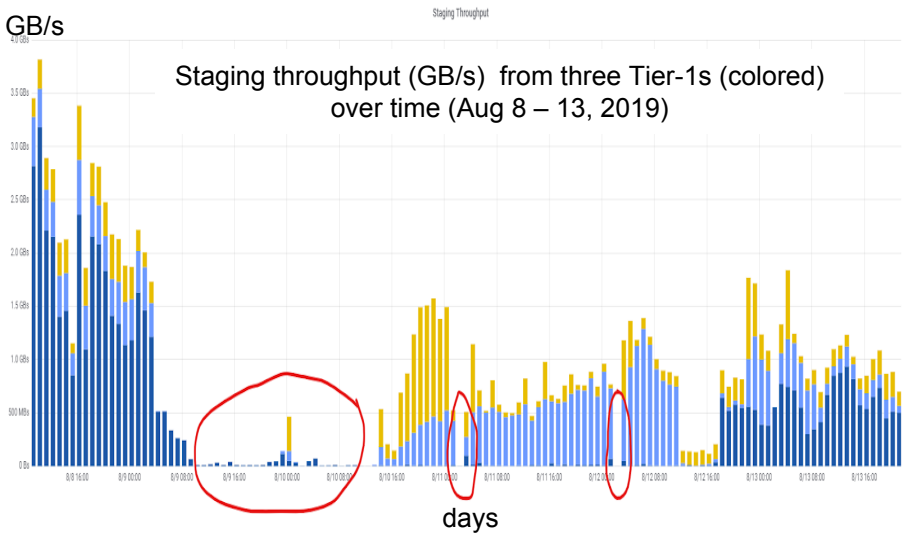
- Staging rate by site: 300MB/s ~ 2GB/s, way below any limits of disk-disk transfer

FTS limitations:

- Bulk submission of staging requests (1.5M+ in 4 hours) to single FTS instance, caused FTS scheduler degradation. Overloaded FTS DB slows down submission of transfer commands
- Purged files increased transfer failure, which in turn triggered FTS optimizer to throttle down the number of parallel transfer limits on the FTS links to minimum

Tape frontend (dCache) limitations

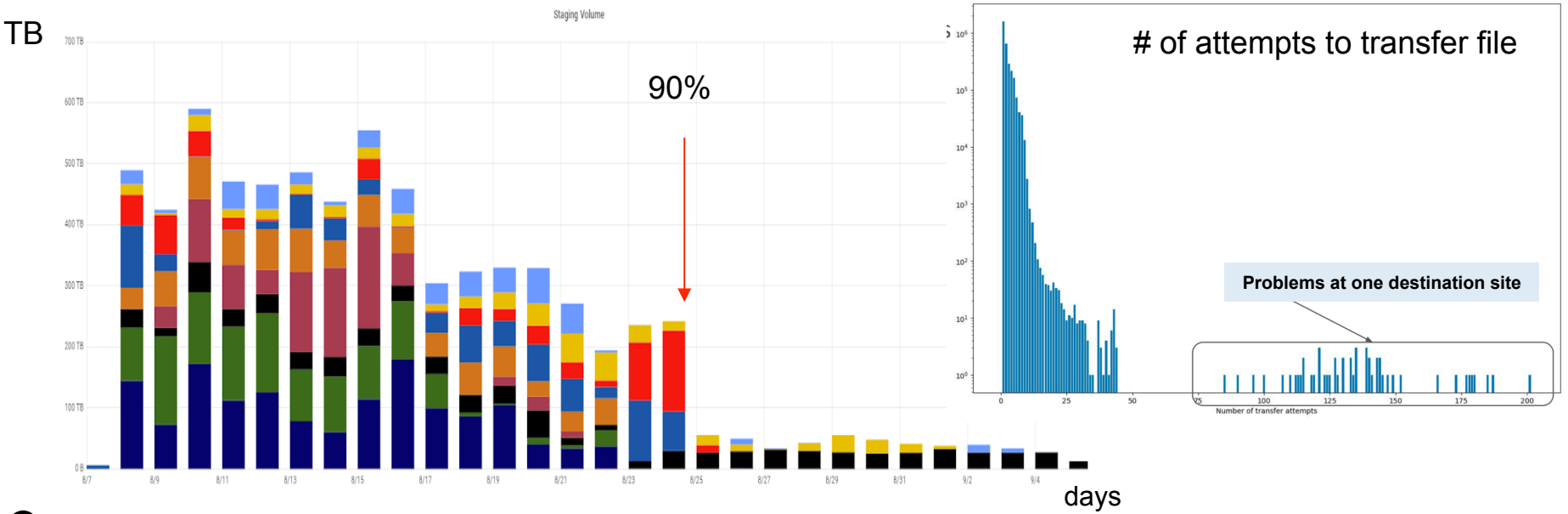
- Can't handle the bulk size, pools crashed, slow I/O nodes caused higher failure rate, which triggered FTS optimizer to reduce link limit … (not new, seen in Phase I)



We (as ATLAS) are planning 1 day technical discussion in November/December with dCache, CTA and FTS experts

BROOKHAVEN

Data Carousel Phase II Round2. Tail effect



- Long delay between 90% and 100%, which happened to many sites
  - Staging scenario
  - FTS issue as mentioned above
  - Problem at the destination (took up to 200 attempts to transfer a file)
  - Rucio and ProdSys2 parameter tuning

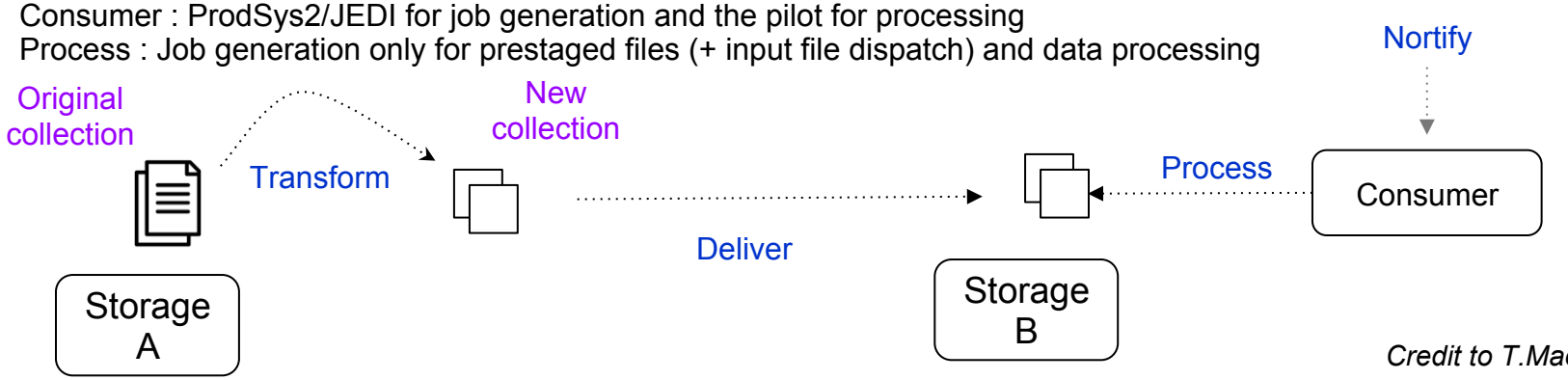
BROOKHAVEN

Data Carousel and iDDS

- A dataset is a unit of ATLAS data processing and replication
- Data carousel works with datasets and ProdSys2 sends staging request per dataset although files are used in downstream systems
  - ✓ Files in each dataset are prestaged by the tape system rather randomly

Prodsys is an upstream component which is far from actual resources → Some changes are required in downstream components for better performance and more optimal resources usage :
One of possible solutions will be orchestration by iDDS with inter-service messaging
Mapping Data Carousel to iDDS workflow :

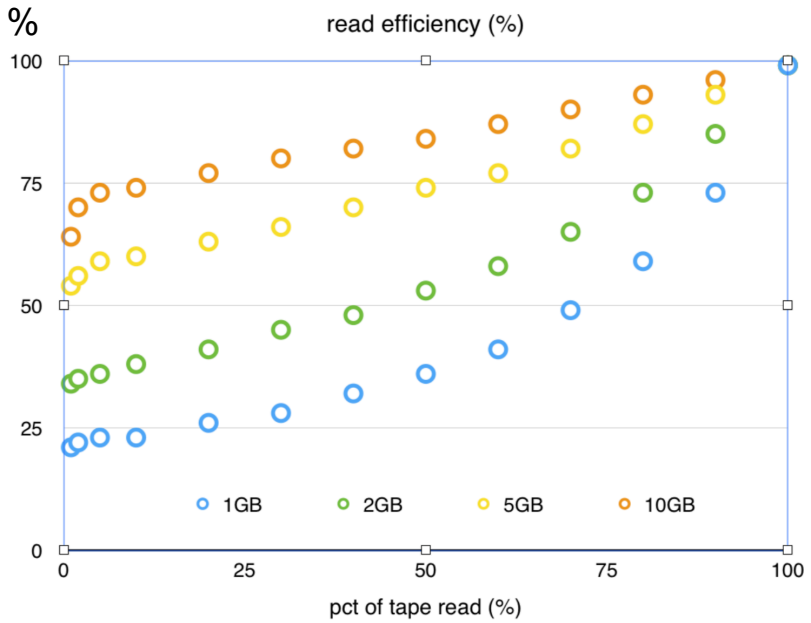
- Storage A : TAPE, Storage B : DATADISK
- Transform : program, Original collection : Files on TAPE, New collection : File replicas on DISK
- Delivery service : Rucio/FTS (near term) WAN/Xcache (for streaming mode)
- Notification : The list of prestaged files
- Consumer : ProdSys2/JEDI for job generation and the pilot for processing
- Process : Job generation only for prestaged files (+ input file dispatch) and data processing



Credit to T.Maeno

Data Carousel. Smart writing

- Efficient data carousel is not possible without smart writing
- It is a team effort between storage SW developers, sites and experiments (TRIUMF has a very interesting experience)
- Possible options
  - Tape families --- too high of a layer than datasets, won't help much
  - Bigger files
    - Zip small output files before writing to tape.
    - Target 10GB
  - Co-locating files from the same dataset on tape
    - Since they will be recalled together, equivalent to "bigger fat file"
    - We have a site that put all files of a dataset on one tape (or 1+ for bigger dataset). Reach almost stream reading speed of a tape drive per tape mount



(Plot is courtesy of Luc Goossens (CERN))

BROOKHAVEN

ATLAS Data. Rucio Statistics

Format	TeraBytes	# Files
AOD	66980	51408767
DAOD	106050	174579883
NTUP	3876	12306115

Caveat : All statistics and numbers from Sep 2019

ProdSys2 Statistics For Selected projects and formats

Project	Format	TBytes	#Datasets (deleted)	#Files
data15_13TeV	AOD	1966	2332 (1012)	1276038
data16_13TeV	AOD	3047	2346 (766)	1670613
data17_13TeV	AOD	4161	3196 (749)	1681786
data18_13TeV	AOD	935	1164 (81)	380999
Total AOD		10109	9038	3.86M
data15_13TeV*	DAOD	570	21321 (145882)	1261594
data16_13TeV*	DAOD	3634	33036 (124876)	5756126
data17_13TeV*	DAOD	5935	42544 (67509)	5995275
data18_13TeV*	DAOD	5092	36031 (26750)	5319080
Total DAOD		15231	132932	18.33M
mc16_13TeV	AOD	18790	49218 (15266)	4.02M
mc16_13TeV*	DAOD	24200	191014(118467)	10.23M

*) *merge+deriv*

AOD metrics 1/2

data AOD datasets (merge only) produced in **the last 365 days**

- Total datasets/files : 3942/379K/1.88PB (#files : <96>, <5.2GB>) used as input 6303 times, <1.6>
- AOD datasets used as input

Not used	1	2	3	4+
302	2447	649	339	205

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
820	164	238	1113	3968

- **Max used = #86, <hours> = 87, min = 11h, max = 7798h [~324 days]**

MC16 AOD datasets (merge only) produced in **the last 365 days**

- Total : 38097/2695K (#files: <65>,<5.1GB>) used as input 167384 times, <4.4>
- AOD datasets used as input

Not used	1	2	3	4+
167	1631	7742	14929	13628

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
3551	3085	7433	24730	128585

Max used = #89, <hours> = 1448, min = 0h, max = 8702h [~362 days]

Delta = [dataset used as input] – [dataset creation time]

ALL Numbers for Production Tasks

AOD metrics 2/2

Users Analysis

data AOD datasets (merge only) produced in the last 365 days

- max used : #23
- #Tasks used AOD datasets as input

Not used	1	2	3	4+
3001	21	239	459	373

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
215	385	786	1696	1554

MC16 AOD datasets (merge only) produced in the last 365 days

- max used : #181
- #Tasks used AOD datasets as input

Not used	1	2	3	4+
33384	2385	373	277	990

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
865	431	850	3414	7819

Delta = [dataset used as input] – [dataset creation time]

ALL Numbers for prun/pathena (credit to T.Maeno)

DAOD metrics 1/2

Central Production

data DAOD datasets (derivation only) produced in the last 365 days

- Total : 96770 (-4845)/12331K (<files> : 128, <0.9GB>) used as input 6145 times
- DAOD datasets used as input

Not used	1	2	3	4+
95502	6102	18	1	1

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
0	74	73	1773	4225

- **Max used : 4; <hours> = 125, min = 0h, max = 3846h [~160 days]**

MC16 DAOD datasets (derivation only) produced in the last 365 days

- Total : 204937 (-54789)/13711K (1022K) (<files> : 52, <2.4GB>) used as input 11635 times
- DAOD datasets used as input

Not used	1	2	3	4+
248143	11535	46	1	1

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
15	157	258	2873	8332

- **23PB, Max used : 5 , <hours> = 424, min = 0h, max = 7910h [~329 days]**
- Delta = [dataset used as input] – [dataset creation time]

DAOD metrics 2/2

Users Analysis

data DAOD datasets (derivation only)

- max used : #181
- #Tasks used DAOD datasets as input

Not used	1	2	3	4+
25122	11952	6573	5635	52998

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
11318	22400	65969	376831	751417

MC16 DAOD datasets (derivation only)

- max used : #1170
- #Tasks used DAOD datasets as input

Not used	1	2	3	4+
73721	25602	16762	14019	111403

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
400329	68483	151691	581119	957728

delta = [dataset used as input] – [dataset creation time]

ALL Numbers for prun/pathena (credit to T.Maeno)

(D)AOD History

Statistics for the last 365 days (Sep 2018-Sep2019); DAOD and AOD datasets used as input by users

- Users tasks : 1923609 by 1552 users
- Total datasets accessed : 406,661,395
  - Users datasets : 36,824,715
  - AOD and DAOD distinct datasets : 321125, Files : 31.6M ; 56 PB
    - +67789 deleted;
    - # AOD datasets accessed MC / data : 33625/ 55226 : <4.2> / <9.9>
    - # DAOD datasets accessed MC / data : 1501168/ 1569729: <8.6> / <12>
    - MAX number of accesses : 1046

delta = [dataset used as input] – [dataset creation time]

#dataset vs format

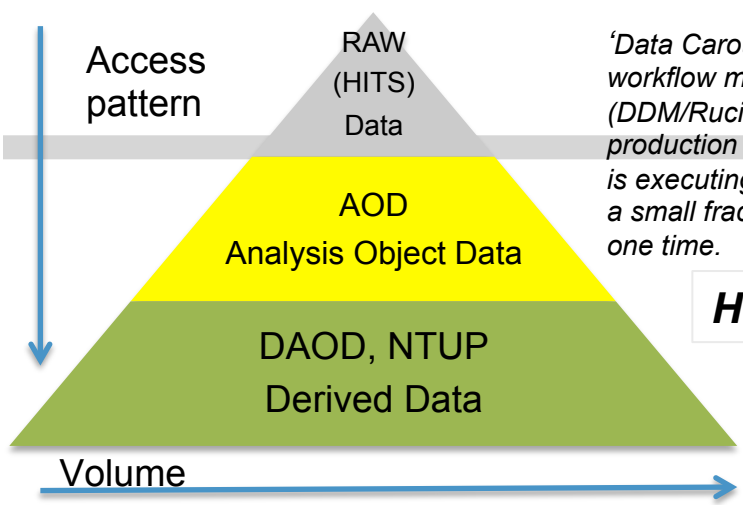
Project	AOD	DAOD	NTUP
MC	8154	173749	108
data	5588	130990	2536

#access per dataset

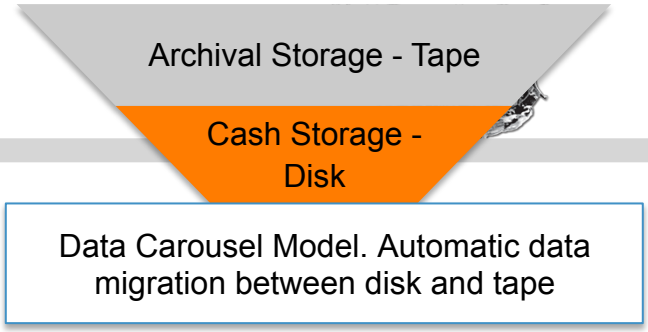
Project	1	2	3	4+
MC	42438	27283	19916	91574
data	24911	12573	13437	88013

delta

< 7 days	1-2 weeks	2 -4 weeks	1-3 months	3+ months
84041	26383	56012	81763	76926

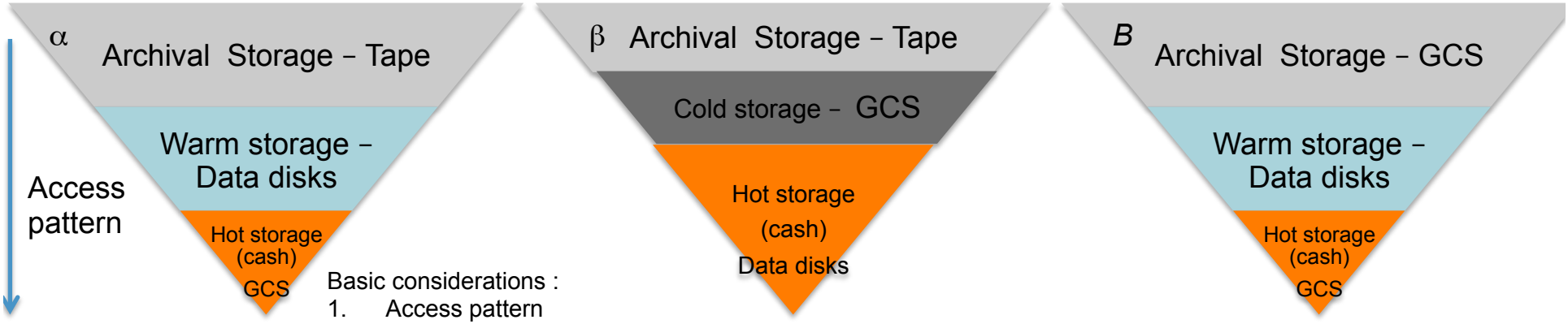


'Data Carousel' we mean an orchestration between workflow management (WFMS), data management (DDM/Rucio) and tape services whereby a bulk production campaign with its inputs resident on tape, is executing and promptly processing of inputs. Only a small fraction of inputs are pinned on disk at any one time.



Hot/Cold Storage Model

Hot/Cold storage model gives us more flexibility with data handling .
We can archive ALL data on tape and keep on disk and cash the most popular data
Plan A (α, β): Data will migrate between hot/warm/cold storage automatically
Plan B : will address the case when tape drives market will be in danger



Basic considerations :

1. Access pattern
2. Cost, performance and capability
    1. Capability = functionality.. How well requirements are managed
    2. Performance = data availability, retrieval speed and data access speed

More Challenges Ahead

- We successfully and quickly passed “a pilot project phase” between ATLAS and T0, T1 centers
  - Many unknown unknowns problem retired/solved. Known unknowns (smart writing,…) still remain
- Continue iterative data carousel exercises
  - Technical exercises with two or three sites
  - Derivation with AOD from tape for a new ATLAS Analysis model
  - New reprocessing campaigns
  - Collaborative exercise with other R&D (e.g. iDDS)
- Continue R&D with Google on hot/cold storage
- Continue R&D with Google on I/O performance optimization