ATLAS - Networking

2020 LHCOPN LHCONE Workshop

ATLAS computing

Ale Di Girolamo - 13 Jan 2020

Introduction

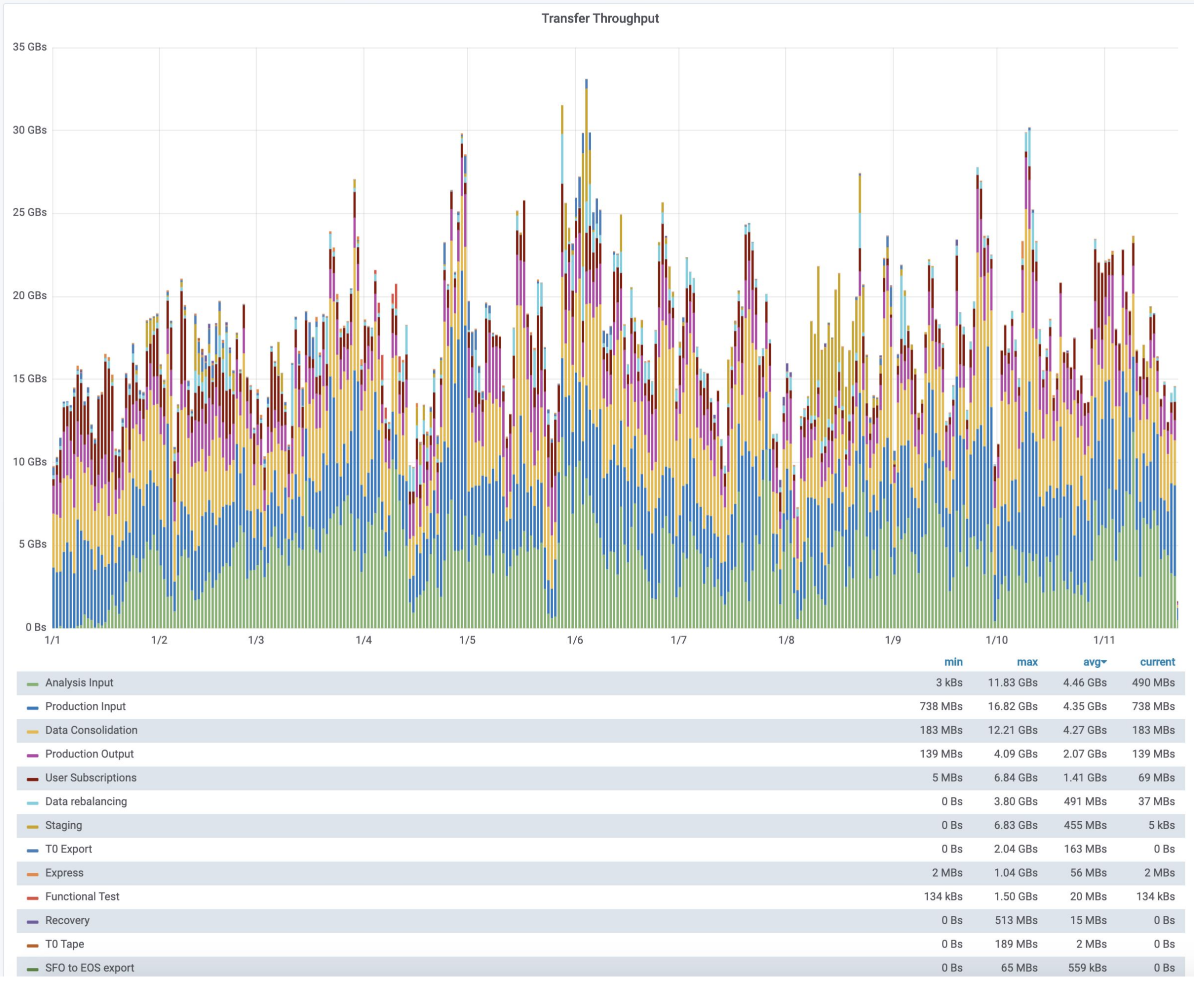
- Networking is and has been one of the rock-solid highly reliable building block of ATLAS computing successes
  - Not without tremendous amount of work from many people
  - Not without issues
- Up to now, Networking has been considered (almost) infinite
  - Saturations here and there were found
  - Most of the issues/limitations are coming from services (i.e. storages and third-party transfer manager FTS) rather than from Network fabric
- Here today to discuss with Networking community
  - Computing model
    - What we have been doing in these years
    - Requirements (n.b. ballpark!) for HL-LHC: evolution of Computing Model
  - From R&Ds to production: make sure that we are preparing ourselves for HL-LHC
    - Data Management and Workflows: better integration (e.g. Data Carousel)
    - Infrastructure (e.g. diskless sites, lakes/oceans/clouds)

ATLAS network usage: synch and asynch, LAN and WAN

- Synch and Asynch (wrt compute job slots) data transfers
  - Today Synch **almost** matching LAN. And Asynch **almost** = WAN
    - N.b. the almost for Synch - more later on diskless sites and more R&Ds
    - the almost for Asynch: needed inside some sites, e.g. Disk to/from Tape
  - Synch: jobs start when data are at the site, and read the files from local storage.
    - Read can be an open and stream directly from compute node to grid storage, or copy on local scratch of compute node
  - Asynch: third party transfers. FTS
    - FTS is the building block, driven by Rucio
    - Rucio trigger FTS transfers from site A to site B (e.g. to consolidate datasets at a site, or driven by Workflow management WFMS requests, i.e. Panda)
    - And then message Panda that job needing that file can start
  -
- All activity recorded in Rucio traces
  - Lot of work done with traces already, but still a lot of useful info could be coming out from there
  - Does anyone have spare Data Scientists?
    - We do have examples already set up!
  - ….

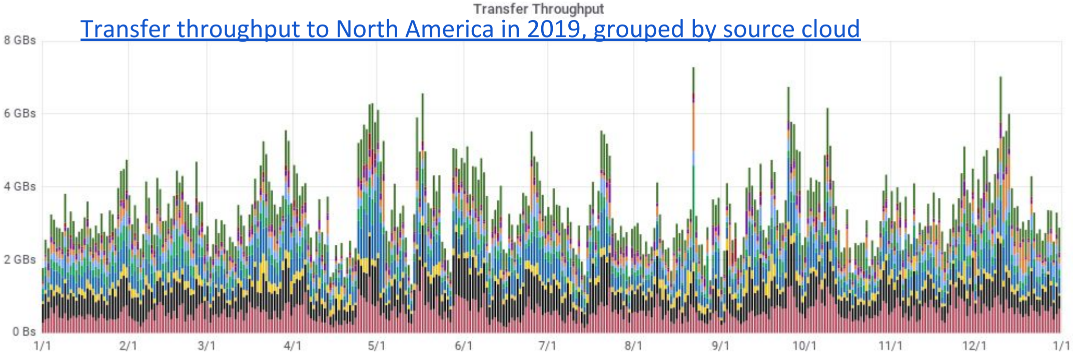
Network usage - last year - Asynch(ThirdParty)

- Data transfers between sites
  - Grouped by activity
- ~ 20 GB/s on average
  - peaks up to 35 GB/s,
  - ~ 20 Hz -> 2M files per day
- 4-5 GB/s Analysis input:
  - Scheduled analysis inputs data movements of ~10% of analysis jobs:
    - would increase proportionally if we would increase this 10%. Do not see the need of increase for now (i.e. the possible positive impact)


Transfer Throughput

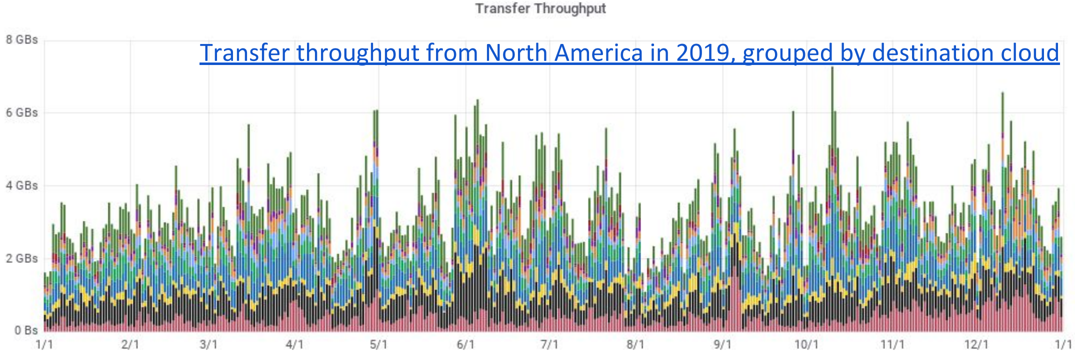
	min	max	avg	current
Analysis Input	3 kBs	11.83 GBs	4.46 GBs	490 MBs
Production Input	738 MBs	16.82 GBs	4.35 GBs	738 MBs
Data Consolidation	183 MBs	12.21 GBs	4.27 GBs	183 MBs
Production Output	139 MBs	4.09 GBs	2.07 GBs	139 MBs
User Subscriptions	5 MBs	6.84 GBs	1.41 GBs	69 MBs
Data rebalancing	0 Bs	3.80 GBs	491 MBs	37 MBs
Staging	0 Bs	6.83 GBs	455 MBs	5 kBs
T0 Export	0 Bs	2.04 GBs	163 MBs	0 Bs
Express	2 MBs	1.04 GBs	56 MBs	2 MBs
Functional Test	134 kBs	1.50 GBs	20 MBs	134 kBs
Recovery	0 Bs	513 MBs	15 MBs	0 Bs
T0 Tape	0 Bs	189 MBs	2 MBs	0 Bs
SFO to EOS export	0 Bs	65 MBs	559 kBs	0 Bs

Network usage - last y - Asynch - Transoceanic Transfers



Transfer throughput to North America in 2019, grouped by source cloud

	min	max	avg	current
CERN	135 MBs	1.814 GBs	573 MBs	462 MBs
DE	127 MBs	1.738 GBs	674 MBs	325 MBs
ES	41			
FR	135			
IT	37			
ND	38			



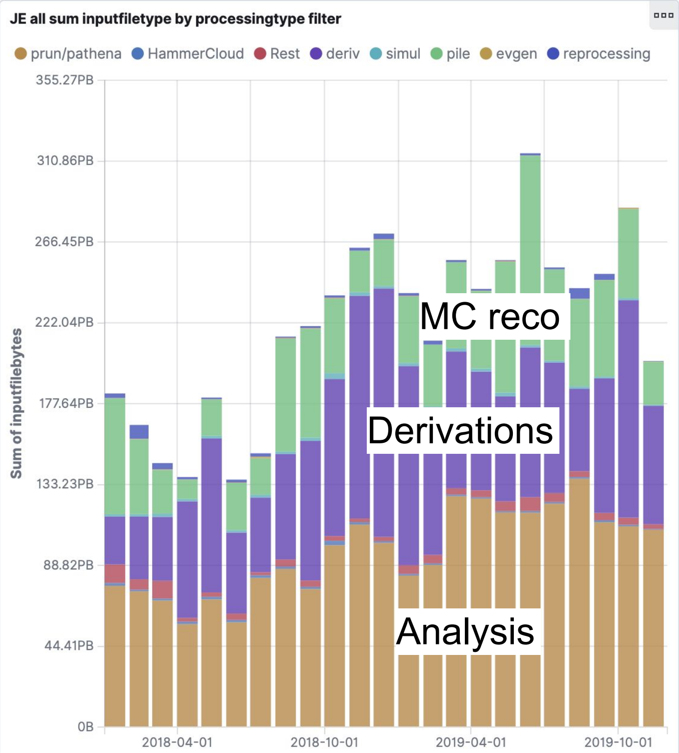
Transfer throughput from North America in 2019, grouped by destination cloud

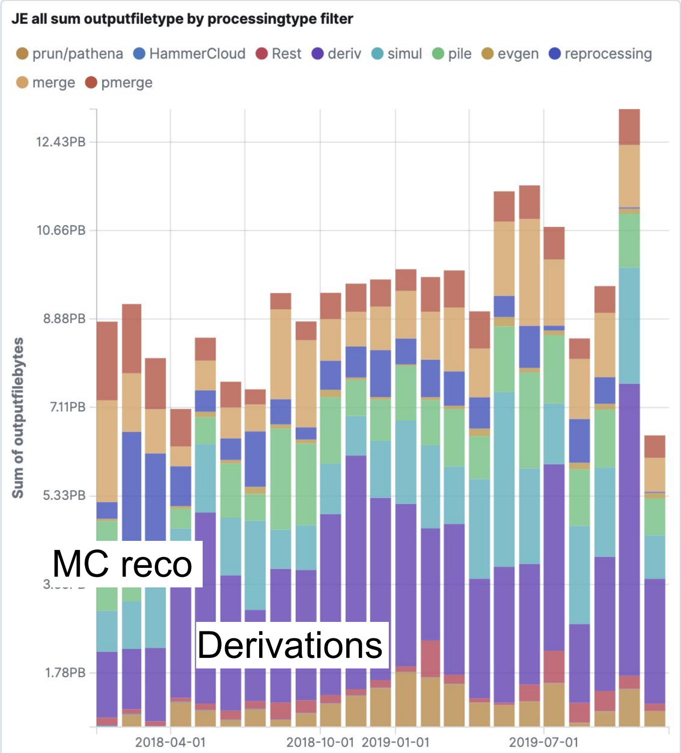
	min	max	avg	current
CERN	71 MBs	1.796 GBs	380 MBs	528 MBs
DE	190 MBs	2.022 GBs	706 MBs	523 MBs
ES	37 MBs	561 MBs	200 MBs	67 MBs
FR	120 MBs	1.829 GBs	579 MBs	446 MBs
IT	64 MBs	1.030 GBs	344 MBs	272 MBs
ND	15 MBs	958 MBs	172 MBs	188 MBs

Ale Di

Jobs data access

- Jobs reading/writing
    - Overall Grid
- (basically) LAN
- ~ 250PB/month
    - Half of it Analysis (~10% of compute power)
    - ~130 PB/month -> 50 GB/s on 25k job slots
    - -> 2 GB/s for 10kHS06 or 1k job slots



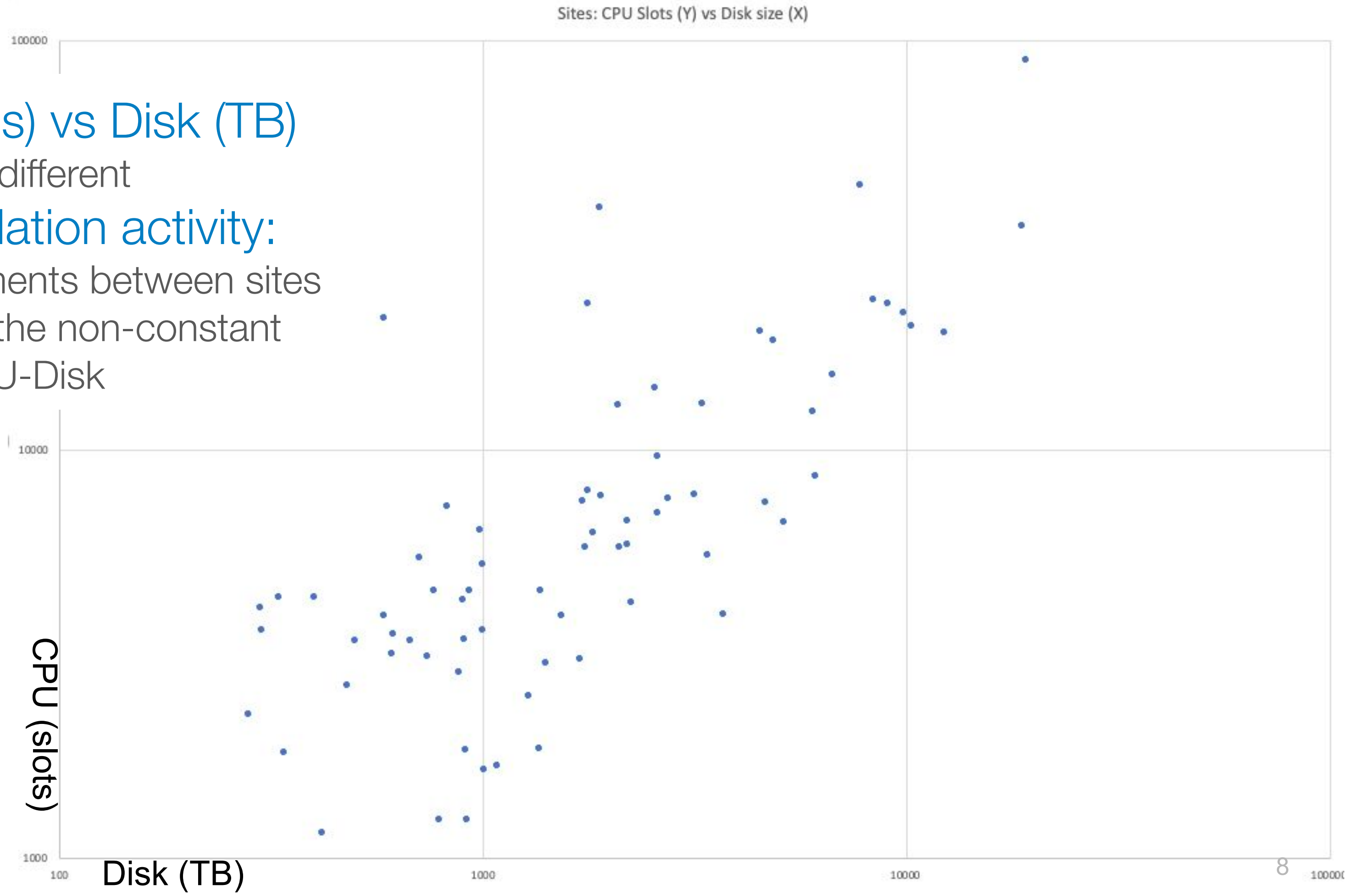


Bursting resources - I/O needs

- Already now it happens we have some periods with e.g. 100-300k slots for few hours/days
  - Obvious case, P1 (the ATLAS TriggerDAQ farm - 45k cores). This
  - we have sites now delivering their CPUs through cloud resources (e.g. Tokyo)
  - HPCs in US and soon in Europe (Prace)
- Bursting (opportunistic or why-not? pledged) resources - ballpark calculation based on today's number.
  - N.b. opportunistic CPUs, but never storage: need to transfer inputs and consolidate back output
    - Simulation only: 1GB/core/day, e.g. 10Gb/s for 100k cores
      - That's the reality today for big resources
    - Everything, including high I/O jobs: 5Gb/s for 2000 cores, or 250Gb/s with 100k cores
      - 250Gb/s is in case everything is remote and if small cache (few tens of TBs) is used.
    - PB or more storage would save network:
      - To have an idea we can scale e.g. BNL WAN usage. Or given that we have 300k I/O intensive cores and traffic of 30GB/s (300Gb/s), need at least 100Gb/s for 100k cores on e.g. HPCs with pledged storage

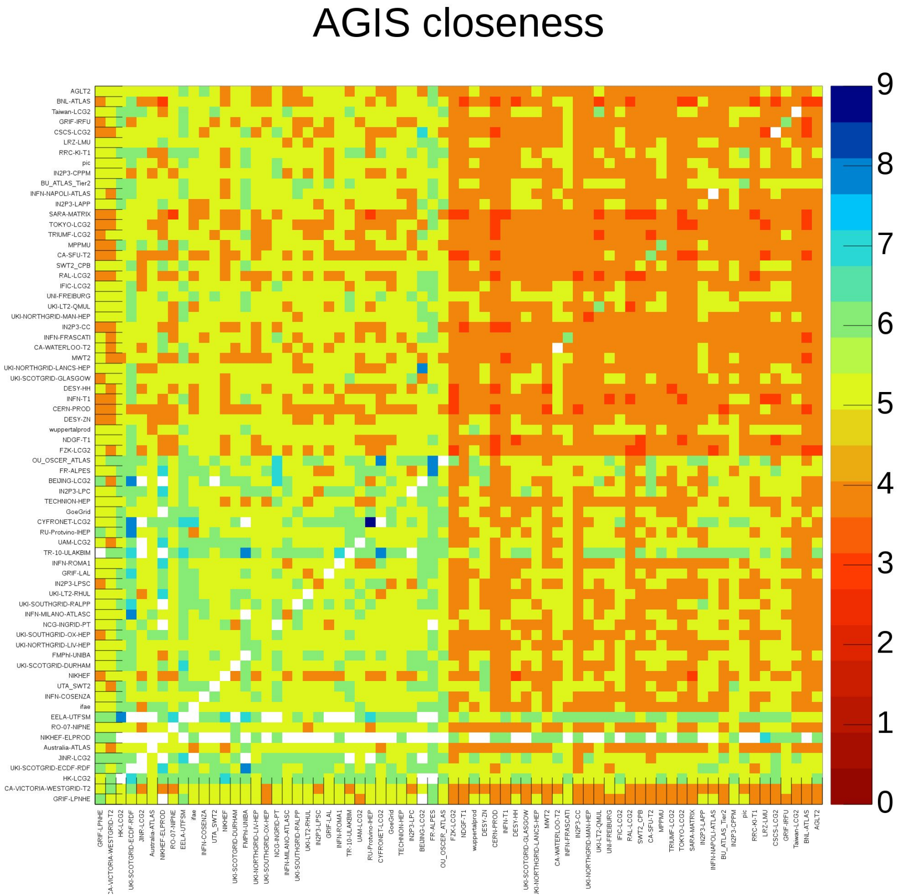
ATLAS Sites: Storage vs CPU

- Plot: CPU(slots) vs Disk (TB)
  - Each site is different
- Data Consolidation activity:
  - data movements between sites also due to the non-constant balance CPU-Disk



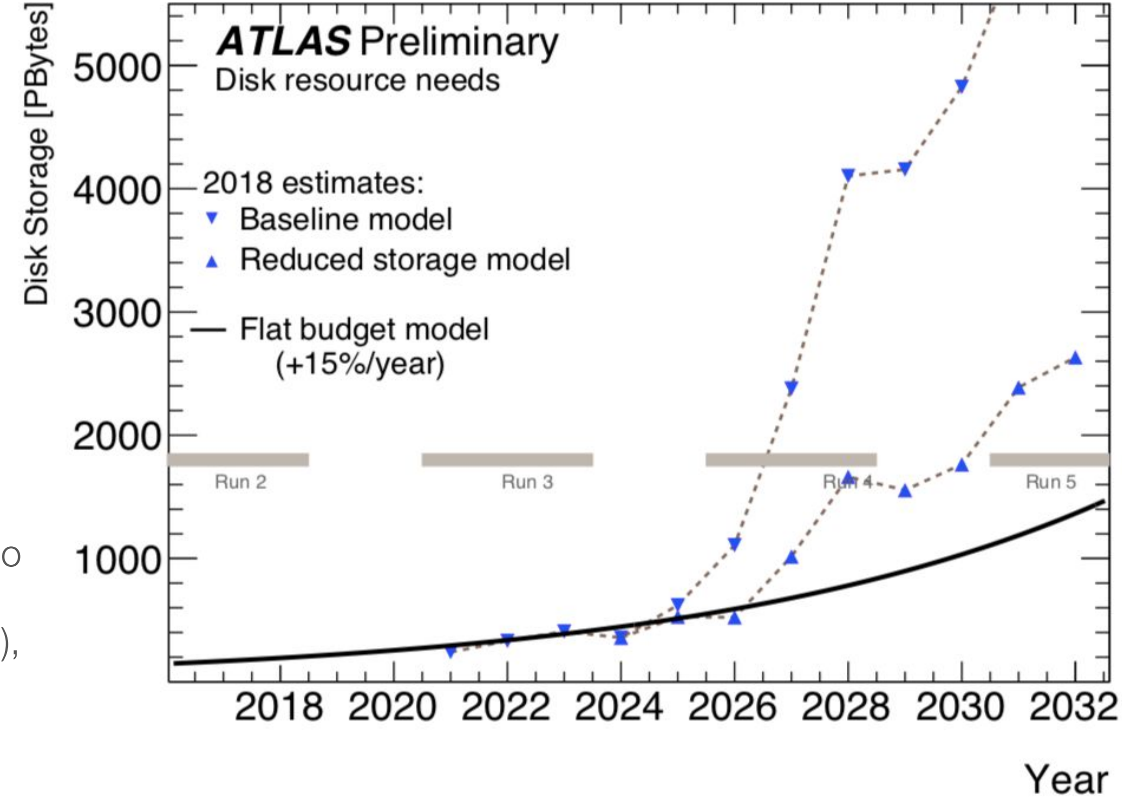
Site Closeness

- ATLAS Model: Nuclei and Satellites
  - Some ~25 Nuclei, each of them with (dynamic) Satellites
  - Nuclei:
    - sites with stable storage and good storage support
  - Satellite:
    - sites close (from Network point of view) to the Nucleus, used to speed up the completion of assigned tasks
    - (Nuclei can be also satellites of other Nuclei)
- AGIS closeness:
  - Maximum measured throughput Src-Dst over past month
  - The red(der) the better
    - 0 → >1 TB/s (there are none - yet!)
    - 3 → >1 GB/s (few links)
    - 6 → >1 MB/s



HL-LHC data volumes extrapolation

- Plot will be updated in coming months
- Good enough for this forum, where we want to have an idea of the ballpark
- Crystal balls, seers and oracles heavily used in the preparation of this plot
- Assumptions
  - Data events: 40B (2026), 50B (2027), 60B (2028), 75B (2031), 75B (2032)
  - MC: double the data, with 80B also in 2025
  - <μ>: 100 (2026), 140 (2027, 2028), 200 (2031, 2032)



HL-LHC Data Challenge - more details - ballpark!

ATLAS EXPERIMENT

- Just logical, one year (i.e. no replication factors in below numbers)
- Data Taking
  - 300PB of RAW per year
    - 50B events
  - ~30 PB of AOD
- + MC (~100B events)
  - 90 PB of AOD (AOD MC bigger than data)
  - Plus HITS (~100PB)
- New format for physics analysis
  - DAOD_PHYSLITE, ~10 KB/ev
  - Contains pre-calibrated physics objects
  - Aimed at around 80% of physics analyses in Run 4
  - ~20PB (Data + MC)

New Analysis model

- All to be discovered
  - Essential to gain experience during Run 3
  - Not yet clear how frequently the analysis format DAOD_PHYSLITE will need to be made (for data)
  - Not yet clear how (for example) do instrumental systematic uncertainty evaluation with such a small format
- Analysis techniques, facilities
  - Likely that columnar analysis techniques will be much more common by Run 4; the PHYSLITE format is already well suited to this
  - The technologies needed are still being developed in a variety of contexts (e.g. IRIS-HEP in the US)
  - Possibility of dedicated analysis facilities to deliver fast access to column-wise data for analysis
  - Likely that deep machine learning will be more heavily used during HL-LHC, also imposes some demands and uncertainties on the resources needed for analysis

HL-LHC Data Challenge - Site view

- Huge Data challenges ahead:
  - CERN: internal traffic and throughput to Tier1s is at the Tb/s (ATLAS only!)
    - Throughput to tape will be massive (20-40GB/s, as compared to today 2-3GB/s)
  - Tier1s and big Tier2s:
    - throughput to Tier1s 20-40GB/s, plus jobs read/write needs will bring the requirements to several 100Gb/s WAN
- N.b. these numbers are likely to change in the coming years
  - need to be ready to adapt

R&Ds towards HL-LHC

- is Network "free"?
  - It has never be for free! now starting to take it into account in e.g. brokering algorithm
  - For sure we are not alone in this universe: maybe we considered ourselves unique in the past, but soon we will have to coexist, to share with other sciences. Need better detailed understanding of usage, tools to orchestrate the needs.
  - WLCG, IRIS-HEP and Experiments have launched several R&D projects to address HL-LHC data challenges ~all of them relying on stable and performant networking
- Data Lake
  - Consolidate geographically distributed data storage systems connected by fast network with low latency. Evolution to reduce storage and operational cost
- intelligent Data Delivery Service (iDDS).
  - iDDS will deliver events instead of "just" 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 w/o SRMless

R&Ds towards HL-LHC - more

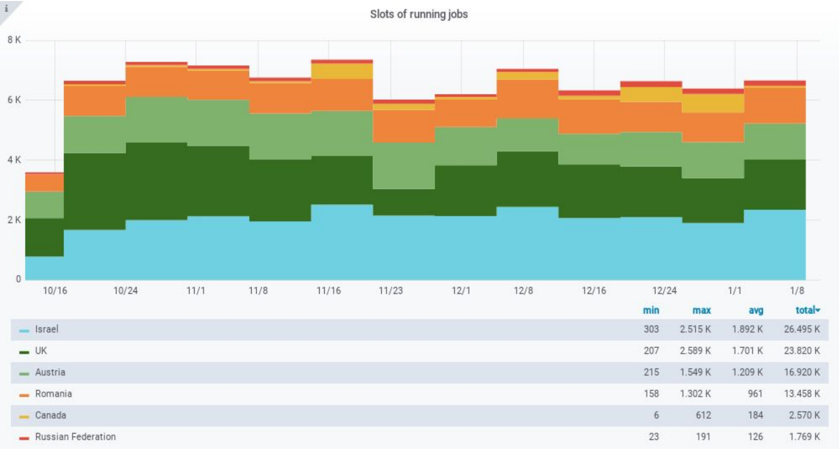
- Different data formats (columnar) and remote streaming leading to different network traffic profile?
  - Need capability of measuring the improvements on changes.
- Rucio awareness of network topology, becoming an active rather than passive user
  - Network shaping and orchestration in conjunction with FTS
  - Dynamic link reservation triggered by workflow changes (e.g. urgent reprocessing needing staging from tape)
- Google-HEP:
  - data placement and migration between “Hot-Cold” storage using data popularity/data usage information.
- Data Carousel:
  - orchestration between workflow/workload management, data management and data archiving services whereby a bulk production campaign with its inputs resident on tape (read tape= less expensive than today's disk storage).

Data Carousel

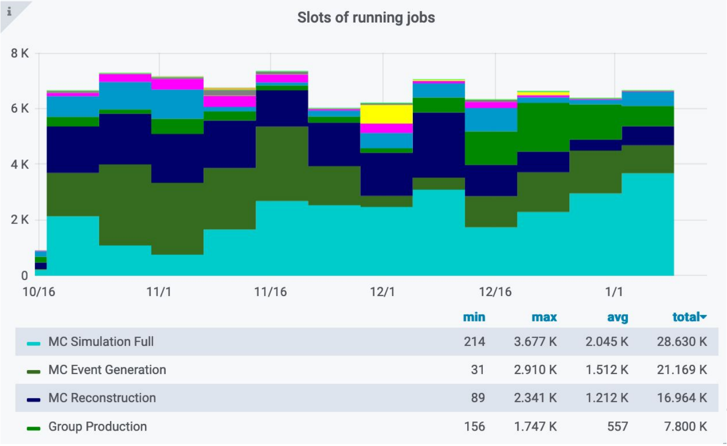
- Orchestration between workflow management, data management and data archiving services:
  - bulk production campaign with inputs resident on tape
  - executed by staging and promptly processing a sliding window of X% (5%?, 10%?) of inputs onto buffer disk
  - only ~ X% of inputs are pinned on disk at any one time.
- Ultimate goal : use tape more efficient and active
- "Side effects":
  - trade storage (pinned disk) with network (data movements)
- On the positive side:
  - Network can be "informed", i.e. WFMS and DDM has the info of the total volume of the campaign ongoing (but not yet all the details of where/which sites are best suited to run)

Diskless sites: More and more

- Possibility for sites to contribute ATLAS Grid computing without managing critical Grid storage
  - Usually triggered by reduction of efficient local manpower
  - ATLAS recommends (do not impose!) diskless sites for storages < 0.6PB (2020)
- Traffic to/from remote storage (usually within same NREN) :
  - Still FTS for transfers with other storages -> increased international traffic
  - WAN access from remote sites: caching mechanism (xcache, arc-cache) helps to minimise



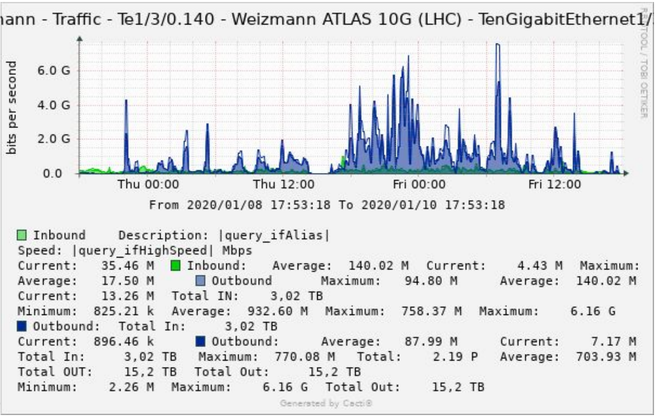
	min	max	avg	total▾
Israel	303	2.515 K	1.892 K	26.495 K
UK	207	2.589 K	1.701 K	23.820 K
Austria	215	1.549 K	1.209 K	16.920 K
Romania	158	1.302 K	961	13.458 K
Canada	6	612	184	2.570 K
Russian Federation	23	191	126	1.769 K



	min	max	avg	total▾
MC Simulation Full	214	3.677 K	2.045 K	28.630 K
MC Event Generation	31	2.910 K	1.512 K	21.169 K
MC Reconstruction	89	2.341 K	1.212 K	16.964 K
Group Production	156	1.747 K	557	7.800 K

Diskless sites: Federated Israel T2 example

- Challenging example:
  - Before summer 2019: 3 similar sites for storage (~500 TB) and CPUs (1.5k slots) running independently
    - Decided to evaluate the possibility to aggregate storage in 1 site accessed by WNs from 3 sites
  - Objective: run (even high I/O) jobs on remote sites without loss of power
  - Central storage hardware config reinforced
    - AGIS offers possibility to control I/O traffic to ramp up smoothly activity on remote sites
- Impact on network :
  - Burst of national traffic with visible impact on site efficiency (asynchronous before)
    - Saturation to be avoided (less visible with asynchronous)
  - All international traffic aggregated on 1 computing facility
- General point:
  - As of now, not weighted the WAN access of the diskless sites to the attached storage, i.e. WNs of the 3 sites are considered all equally connected.



```
From 2020/01/08 17:53:18 To 2020/01/10 17:53:18

🟩 Inbound     Description: |query_ifAlias|
Speed: |query_ifHighSpeed| Mbps
Current:   35.46 M  🟩 Inbound:    Average:  140.02 M  Current:    4.43 M  Maximum:
Average:   17.50 M     ▨ Outbound       Maximum:   94.80 M     Average:  140.02 M
Current:   13.26 M  Total IN:    3,02 TB
Minimum:  825.21 k  Average:  932.60 M  Maximum:  758.37 M  Maximum:    6.16 G
■ Outbound:   Total In:      3,02 TB
Current:  896.46 k     ■ Outbound:      Average:   87.99 M     Current:    7.17 M
Total In:     3,02 TB   Maximum:  770.08 M   Total:     2.19 P   Average:  703.93 M
Total OUT:    15,2 TB  Total Out:     15,2 TB
Minimum:    2.26 M  Maximum:    6.16 G  Total Out:    15,2 TB
Generated by Cacti®
```

R&Ds - NOTED

- Build a repository of shared network knowledge that network operators can look up, and then act locally to improve situation
  - Central service, local BGP actions
- Demonstrated that it works (e.g., CERN-NLT1 with LHCOPN)
  - Now working on automating it!
- Plus: Exploiting backup links in particular pre-planned situations
  - Site decommissioning, disaster recovery
  - Large-scale data rebalancing
- But the network is usually not the issue: filling the network is difficult!
  - Enough throughput from/to source/destination storage systems?
- Most importantly, how can we make sure that we are not "blocking something else" by doing this?

Ops, Monitoring and Analytics

- As of today, whenever there is a problem it is very much often seen too high level, too many layers.
  - Not good enough proactive monitoring and alerting
  - Not because of lack of work! It's just a very complex highly dynamic problem
  - We are not alone in this universe, when things go wrong in/with network, things avalanche and understanding the root cause is very complicated
- Multi-VO: tagging of traffic
  - Need to have as much as possible traffic VO (and application) tagged
- Orchestrating:
  - need a central place where we orchestrate, prioritize, schedule the traffic. (FTS next slide)
- Monitoring and Analytics:
  - Key to understand what we are doing and decide where to invest more.
    - We should have plenty of Data Scientists, each physicist (in theory) is a Data Scientist...
    - Back on my point of Rucio traces, understanding the details of our traffic and data usage and reusage.
    - Important to find dedicated people on this topic.

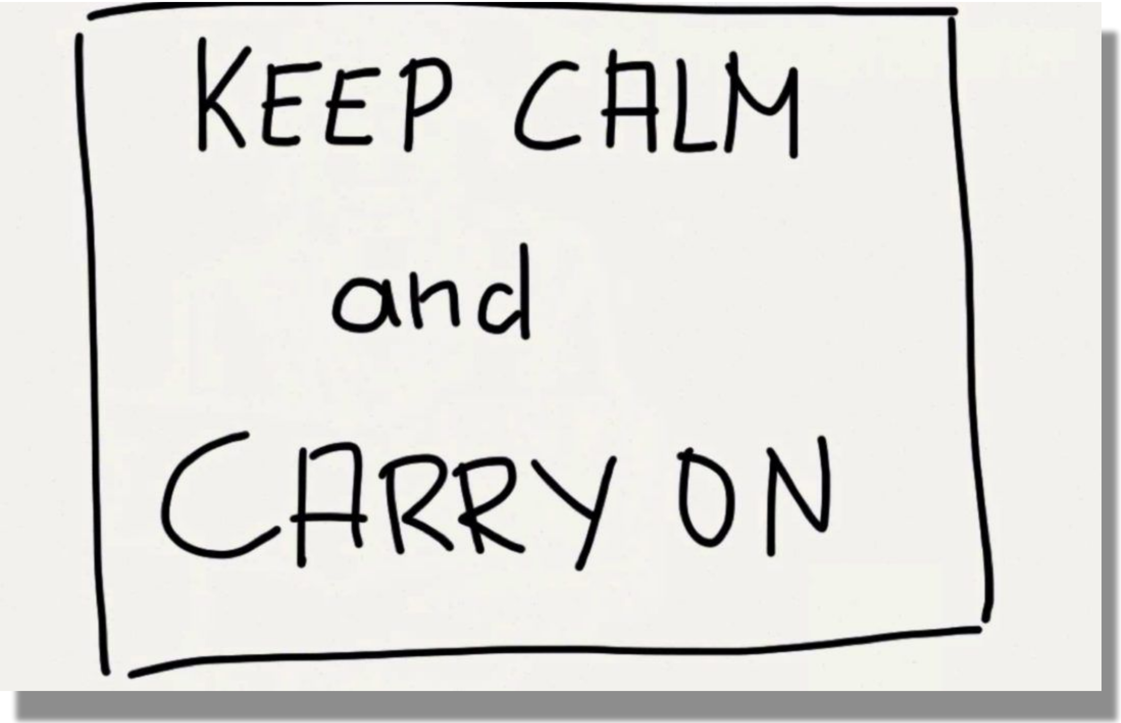
FTS - the elephant in the room

- FTS is THE service to manage third-party transfers
  - Used by ATLAS, Belle2, CMS, LHCb, and many others
  - Very complex role: FTS acts between the storages and the DDM frameworks
  - Demonstrated to work very well at the scale presented
    - Not without issues
- FTS: we can have major improvements for the infrastructure
  - "central" instance
    - Not one single instance, but capability of knowing what the other instances are doing
    - Across experiments - i.e. do not just separate experiments.
  - Better handling of the variegate storages we have
    - Especially in case of performance degradation
  - Also manage Tapes (need careful orchestration)
    - But with limitations, work ongoing
  - Networking understanding?
    - At least basic topology (e.g. sites), and "group" of sites
  - Need deep analytics studies to understand our complex infrastructure
    - To be able to be effective with the changes

Conclusions

- Nothing is for free
  - for sure not the Network!
  - Start including it as a resource in our frameworks algorithms
- Huge Data challenges ahead for HL-LHC
  - Might reach order of Tb/s between big sites
- FTS seems to be the natural central service which interacts "directly" with the network
  - Scheduling and orchestrating
- Monitoring and Analytics keys to understand the impact of the R&Ds

- Need coherent across-experiments (multi-science) approach



KEEP CALM

and

CARRY ON

Jan 2017 LHCONE/OPN meeting

- pre-GDB on Networking 10 Jan 2017
  - https://indico.cern.ch/event/571501/
- ATLAS Computing outlook 10 July 2019
  - https://indico.cern.ch/event/739880/contributions/3470870/attachments/1877809/3092930/ATLAS_Computing_Outlook_-_GDB_-_201907103.pdf

Jobs I/O in ATLAS

- Very rough, ATLAS specific (internal) - draft

- I/O within jobs (production only!) summarized in: https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/IO_in_ATLAS
  - low IO site (a priori diskless= only WAN) : 0.5 Gb/s for 10 kHS06 (1k core, e.g. plot 9)
  - high IO site : 5 Gb/s for 10 kHS06 (1k core)
  - Assuming ratio 1 WAN to 5 LAN (difference between Nucleus or not ?) :
  - WAN : 1 Gb/s for 10k HS06
  - LAN : 5 Gb/s for 10k HS06
- I/O within analysis jobs: rouge estimation 130 PB/month -> 50 GB/s on 25k job slots -> 2 GB/s for 10k HS06 or 1k job slots (N.B. 2 MB/s per job - should be much faster for sparse reading)

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

Data Carousel R&D Project. Cont’d

- DDM system : Rucio → more intelligent tape I/O
  - Bulk data staging requests handling
  - Use FTS features on more intelligent way
- File Transfer Service → optimize scheduling of transfers between tape and other storage endpoints
- DDM / WFM/ Facilities integration. Optimize data placement to tape
  - Define tape families for files known to be re-read from tape (data grouping)
  - Optimize file size (Larger file size, 10GB+ preferred)
  - Use novel (or request new) features of storage systems (dCache, EOS, CTA,…)

Data Carousel. Lessons Learned

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

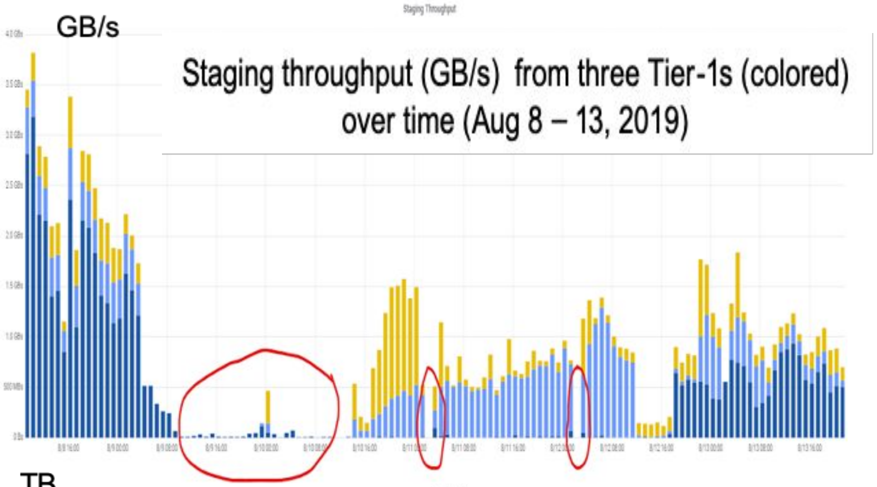
FTS limitations:

- 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

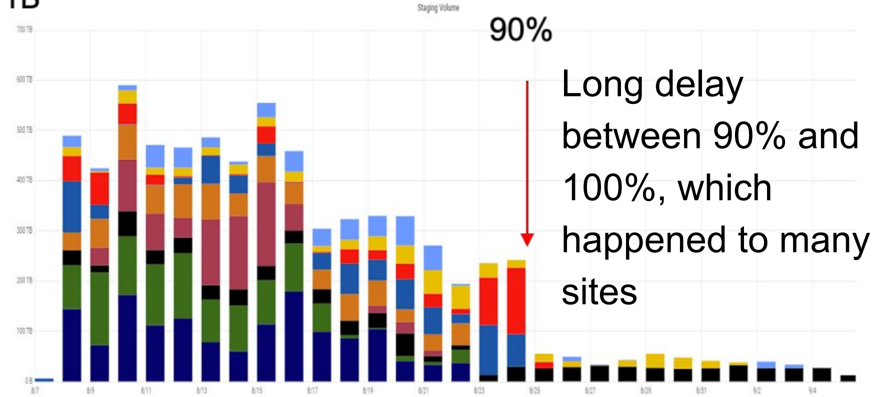
Tape frontend (dCache) limitations

- 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) had one day technical discussion in December with dCache, CTA and FTS experts. The next Data Carousel phase is starting ~today.



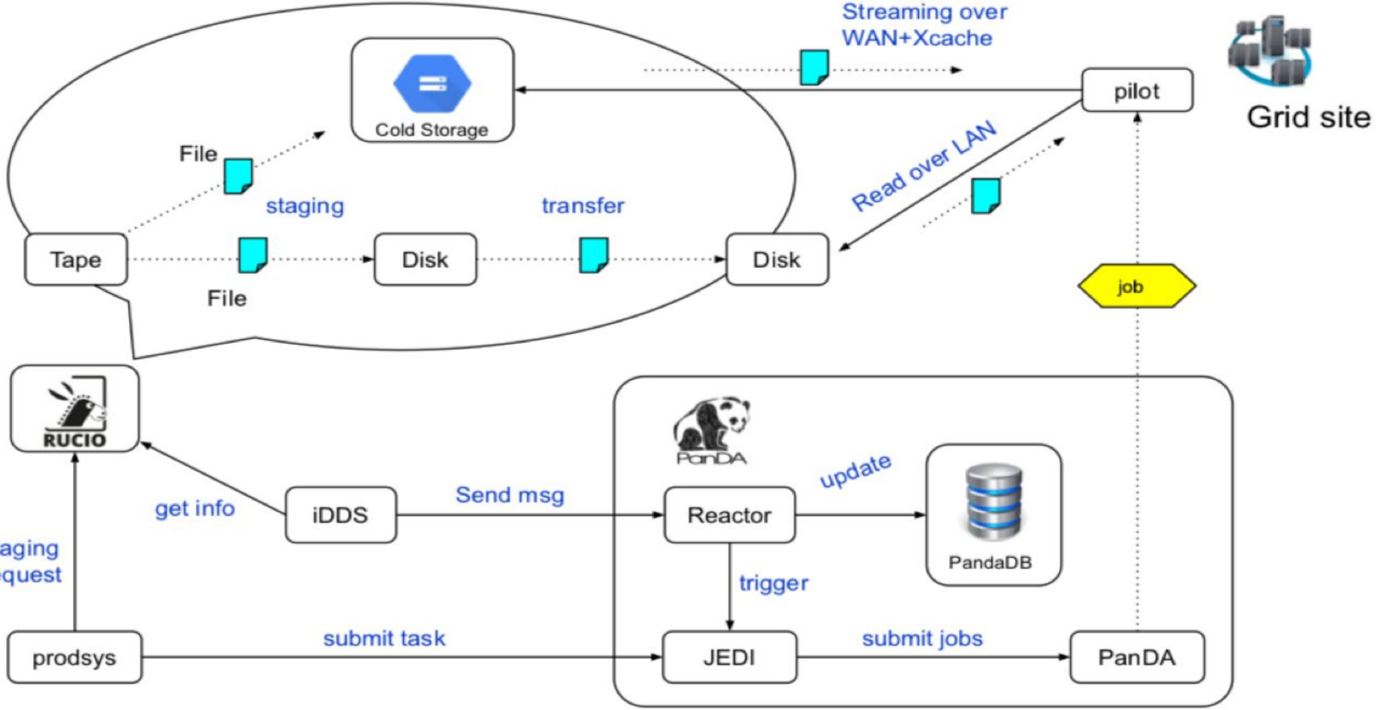
Staging throughput (GB/s) from three Tier-1s (colored) over time (Aug 8 – 13, 2019)



Long delay between 90% and 100%, which happened to many sites

Only topics relevant to networking, but there are more

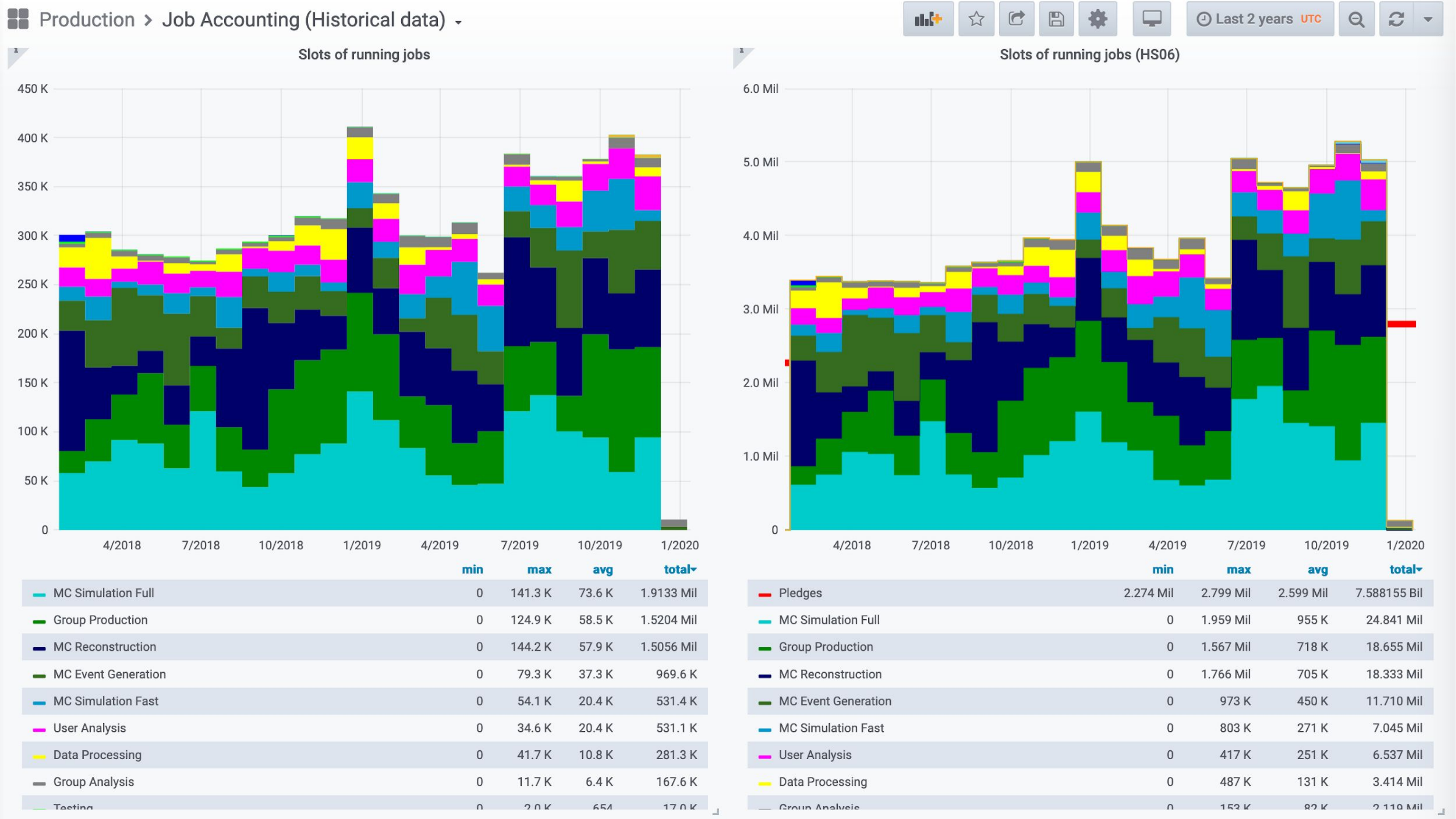
Data Carousel and iDDS R&Ds and Networking



Improve data streaming granularity : from (sub)dataset to file

A new PanDA component – Reactor. To receive messages from iDDS via ActiveMQ and take actions, e.g. update file and task information, dictate file grouping, and trigger job generation

Running jobs - last 2 years

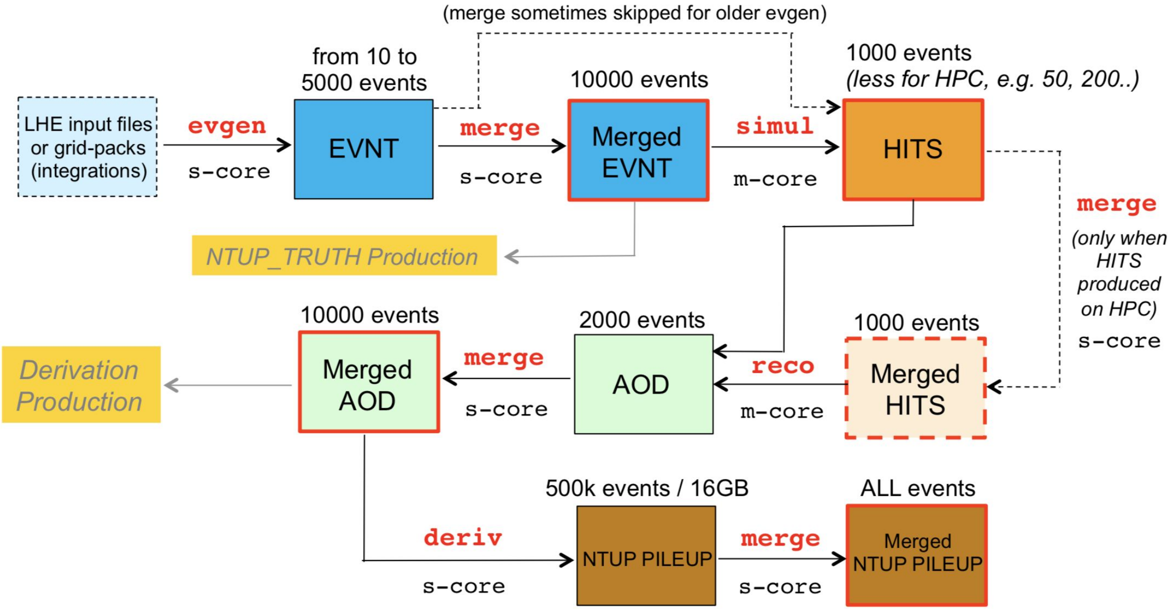


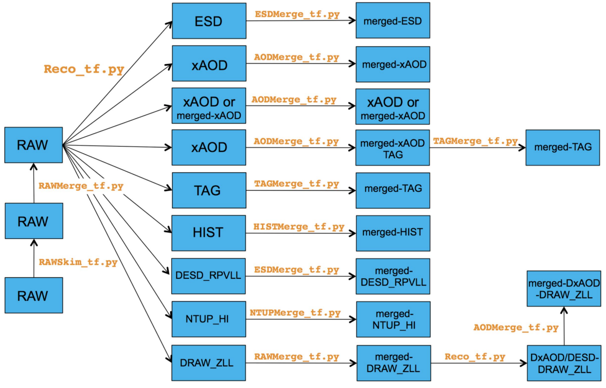
	min	max	avg	total
MC Simulation Full	0	141.3 K	73.6 K	1.9133 Mil
Group Production	0	124.9 K	58.5 K	1.5204 Mil
MC Reconstruction	0	144.2 K	57.9 K	1.5056 Mil
MC Event Generation	0	79.3 K	37.3 K	969.6 K
MC Simulation Fast	0	54.1 K	20.4 K	531.4 K
User Analysis	0	34.6 K	20.4 K	531.1 K
Data Processing	0	41.7 K	10.8 K	281.3 K
Group Analysis	0	11.7 K	6.4 K	167.6 K
Testing	0	2.0 K	654	17.0 K

	min	max	avg	total
Pledges	2.274 Mil	2.799 Mil	2.599 Mil	7.588155 Bil
MC Simulation Full	0	1.959 Mil	955 K	24.841 Mil
Group Production	0	1.567 Mil	718 K	18.655 Mil
MC Reconstruction	0	1.766 Mil	705 K	18.333 Mil
MC Event Generation	0	973 K	450 K	11.710 Mil
MC Simulation Fast	0	803 K	271 K	7.045 Mil
User Analysis	0	417 K	251 K	6.537 Mil
Data Processing	0	487 K	131 K	3.414 Mil
Group Analysis	0	153 K	82 K	2.119 Mil

Overview of workflows

- highly chained workflow, multi-steps, passing inputs and outputs around the network
- Future workflows, where we may or may not keep intermediate data formats, maybe concentrated on single sites, potentially reducing network traffic.





...

- ...
  - ...
  -

- ..
  - ..

- ,..
  - ...