S&C Overview

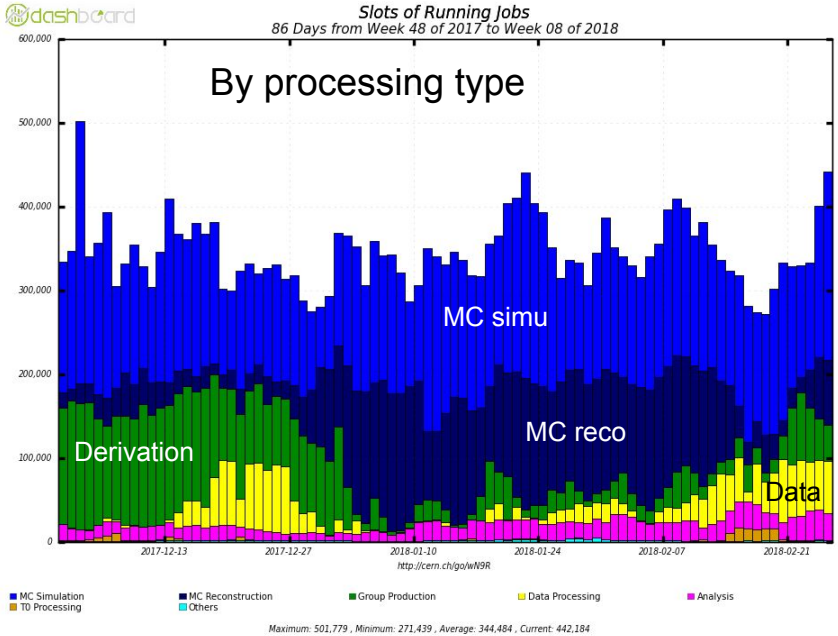
Torre Wenaus (BNL), Davide Costanzo (Sheffield)

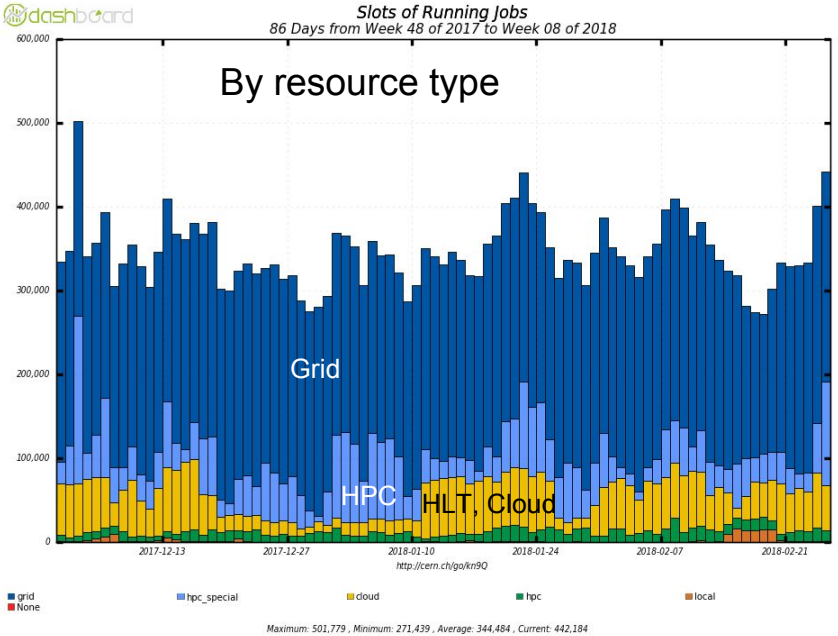
ATLAS S&C Technical Meeting Week
March 5 2018

BROOKHAVEN

The University Of Sheffield.

Processing in last few months

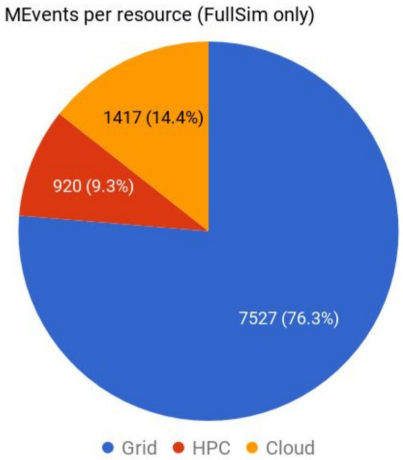




- Sustained production with smooth operations, ~300-350k cores
- HLT and HPC are important contributors as well as grid
- CNAF and its data are back, tape and disk working

Throughput metrics

- Working on updating our throughput metrics
- Concurrent core count becomes less useful with every new HPC generation
- Because HPC cores get slower and slower!
- Normalizing event processing throughput across all resource types and workflows is a complex problem



From ATLAS 2018 proposal to DOE for HPC time

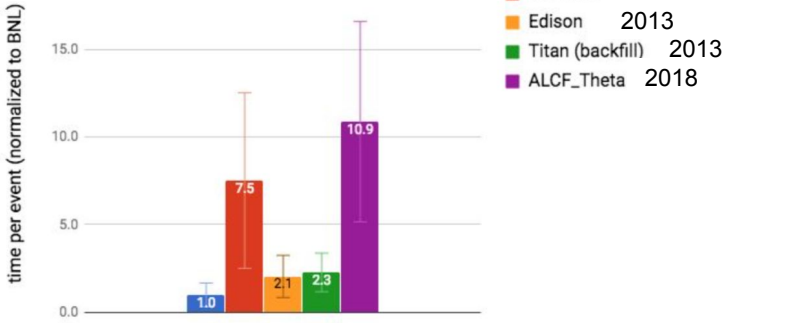
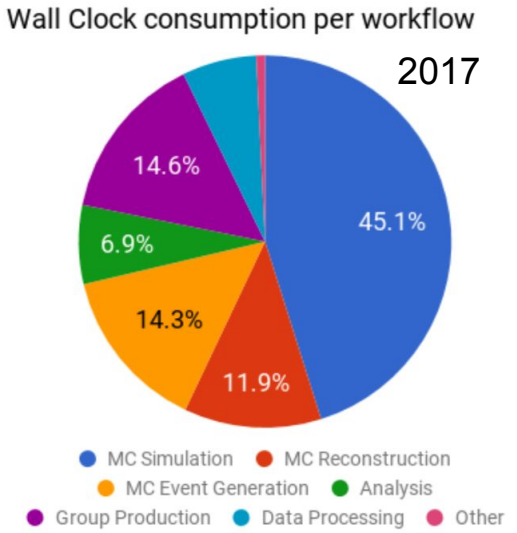


Figure 4: Average processing time per event (per CPU-core) using 10,000 ttbar events from the same physics sample run at HPC sites to gauge performance on different architectures. The Brookhaven Tier-1 was used as a baseline for performance.

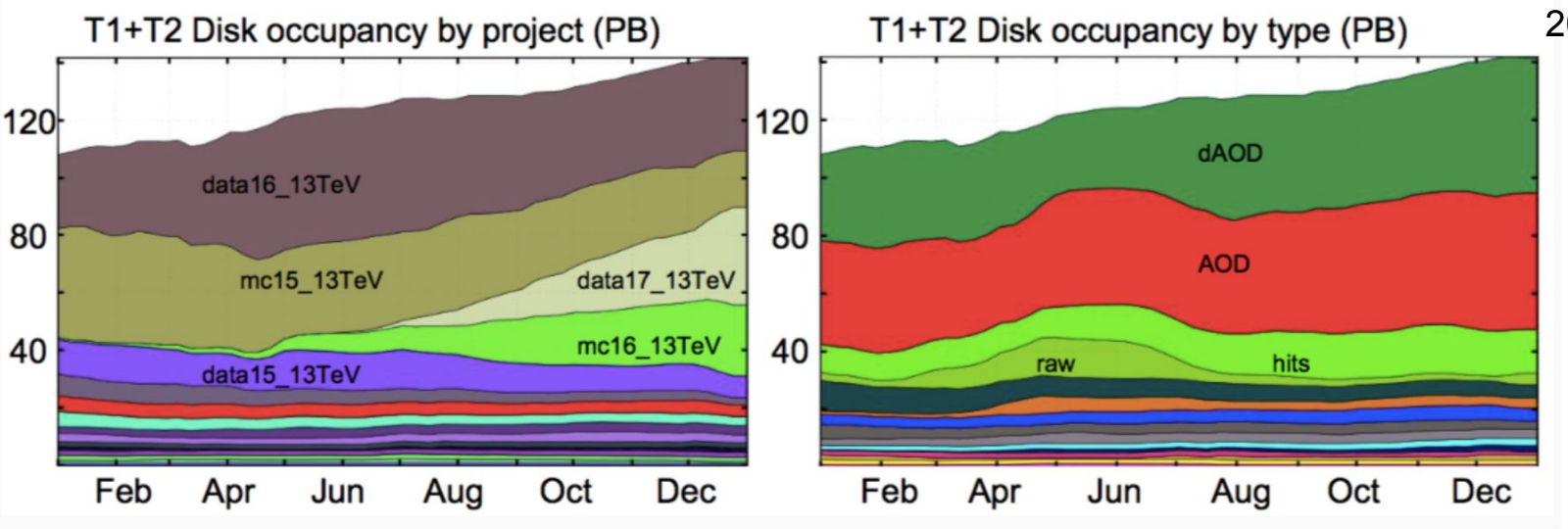
Processing campaigns

- **Derivation reprocessing** second half of campaign underway
- **2017 data reprocessing** ditto, second half (8b4e) underway now
- **MC16d campaign** reconstructing MC with final mu profile (very different from spring estimation) largely completed, MC16e starting soon
- Steady flow of **new/extended evgen/MC** samples
- Remainder of 2018:
  - Simulate/reconstruct ≈10 B events
  - Derivation train production as needed
  - Reprocess simulation with latest mu profile at year end
  - Possible data reprocessing before Christmas



Data management

2017



Moving >1 PB, >20 GB/s, 1.5-2M files per day

Disk will be tight this year, not a crisis but ADC operations is managing the space closely and developing mitigation measures

C-RSG scrutiny and LHCC review

- Our Spring 2018 C-RSG report submitted on time a few weeks ago
  - Draft; can revise in light of feedback
  - We made no changes to 2019 requests between October and now
- C-RSG had preliminary questions this week
  - HI run plans, coping with +20% lumi, pilot and multicore efficiency, tape usage, operational impact of Tier-0 spillover to grid, tape based workflows (‘data carousel’)
  - No questions or challenges to our resource requests (so far)
  - Further discussion this week
- Meetings with LHCC computing referees also took place last week
  - Went well, supportive of mitigation measures for tight resources, analysis preservation, improvement programs like SPOT (ATLAS) and data format optimization (CMS nanoAOD)
  - Outline of HL-LHC strategy document was presented by IT

Efficiency & performance

- New memory reduction and efficiency measures coming into use
  - Make **more memory shared**, reducing total memory needs and expanding the pool of resources available for memory intensive processing such as high-mu events *(fork after first event)*
  - **Eliminate costly end-of-job merge**; instead, merging is done within the event processing loop *(SharedWriter)*
    - As of yesterday: the SharedWriter is commissioned for derivations and we aim to use it in production
- Introducing a **new compression scheme** (LZMA) which reduces file sizes by up to 10%, at an affordable CPU cost, in production in coming months
- **Meltdown/Spectre** impact: ~no slowdown in MC simulation, ~3% slowdown in reconstruction, slightly more in file merging (with relatively tiny wallclock)
- **MC speedups** expected this year: static linking, compiler optimization, improvements in geometry primitives and model optimization, and (possibly) pileup overlay
- **Software Performance Optimization Team** (SPOT) has continued to ramp up and is very busy, recently expanding its active scope to I/O

Software developments

- First major step in the MT migration -- using MT-compliant data access across all systems (DataHandles) -- due at end 2017 and mostly finished, see Ed & Walter's talk
- Very tight effort levels an ongoing problem, c.f. Ed Moyse's talk last Friday in ATLAS Week (and probably the next talk)
- Reviewing the plan for integrating ACTS into Athena - March
- Planning an I/O and persistency review, towards simplification and better performance - June
  - Flesh out plans for it this week
- Fast sim advancing, but slowly. Management will help try to inject more effort.
  - Hackathons like last week help.
- RTT to ART migration of R21 releases advanced but not complete. Aim for May 1?
- Need follow-through on the doc workshop!

Software Milestones

- 2017 Q2: First AthenaMT developers workshop inviting subsystem developers
- **2017 Q4: Finish migration to MT compliant event data access (DataHandles)**
- **2018 Q1: Start ACTS integration in Athena**
- 2018 Q2: MT compliant conditions data retrieval
- 2018 Q4: Public algorithm tools thread-safe
- 2018 Q4: Make Services thread-safe
- 2018 Q4: Tracking code migrated to next-gen MT infrastructure ACTS
- 2018 Q4: TDAQ milestone: First integration with online, concurrent data access demonstrated
- 2019 Q2: Start physics validation of MT vs. ST vs. R21
- 2019 Q3: MT compliant data quality monitoring
- 2019 Q4 TDAQ milestone: Algorithms migrated & tested, multiple threads working and in use
- 2020 Q1-Q4: Bug fixes, optimization & full validation
- 2021 Q1: Release 22 in production for Run 3

Open Athena software

- The plan for making the Athena software open source, following the software policy, was presented and discussed during ATLAS Week, endorsed by the CB
  - Strongly supported by the management
- Distributed software already open source
- In the CB meeting the statement was made that in the UK, students hold the copyright to code they write
- Looking into the actual policy at a few UK institutes, they in fact include provisos for situations like ours, in which the code is written as part of a collaborative project that has a copyright policy of its own, in which case the copyright of the project supersedes
  - e.g. Sussex: "It is recognised that where a student is externally sponsored or part of a research group that is subject to obligations to a funder, the terms of that sponsorship or funding may override this position, and require the student to assign to the sponsoring organisation or funder."
  - https://www.sussex.ac.uk/webteam/gateway/file.php?name=exploitation-policy-of-ip.pdf&site=377
- As the plan says, we will sort out copyright corner cases over the next months.
- Writing a FAQ to address questions

Plan outline, most activities concurrent:

- Repo containing (only) code to go public - done
- Get code ready to open: repo cleanup, copyright check, acknowledgements, … (Apr)
- License-aware dependency analysis: Apache 2 where possible, GPL where required (May)
- Get public-facing documentation ready (May)
- Address copyright corner cases (ongoing as needed)
- Review readiness for going public (June)
- Open the software (June)
- Tell people! ATLAS outreach, HSF, etc

Rucio workshop

https://indico.cern.ch/event/676472/
Live notes

- Workshop last week at CERN presenting Rucio as a data management system to a wide community
- Info exchange between developers and interested communities
- Seed collaborations, new Rucio users, collect feedback & requirements
- Rucio already operating as an open source project with contributions from well beyond ATLAS
  - Facilitated by the github workflow that allows core developers to review code submissions
  - https://github.com/rucio/rucio
  - 28 github contributors at 10 institutes
  - 42 people on rucio.slack.com
- 70+ registrations
- Looking forward to hearing how it went!

Using in production:

- ATLAS — confirmed!
- AMS++ (ASGC) — confirmed
- Xenon1T/XenonNT — confirmed

Actively evaluating:

- CMS — confirmed
- OSG — confirmed
- Belle II — pending
- LIGO — confirmed

Interested:

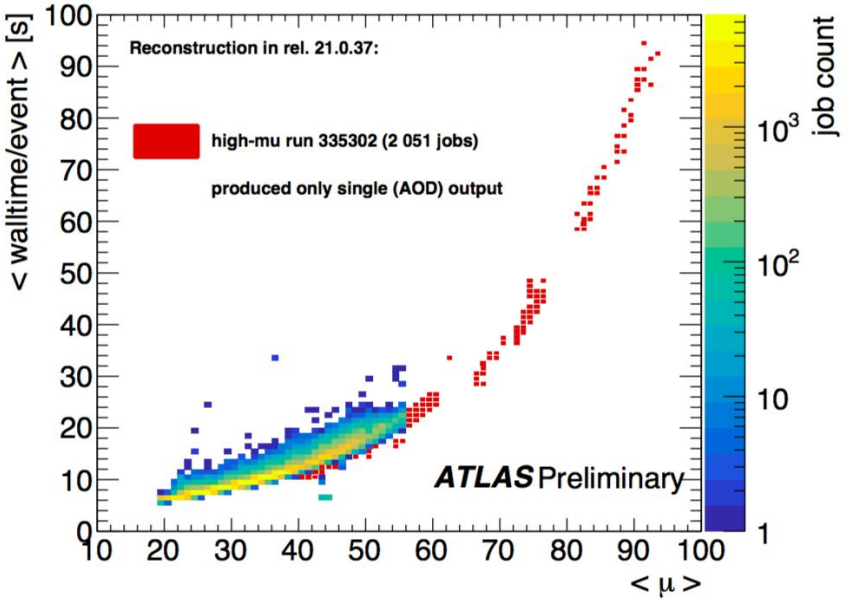
- IceCube (Vidyo) — confirmed
- Eiscat3D/NeiC — confirmed
- NA62 — confirmed
- Compass — confirmed
- DUNE/FIFE — confirmed
- CTA — confirmed
- AENEAS (SKA) — confirmed
- LSST — pending

Conditions DB for Run-3 and beyond

- Had a very useful and productive review of our plans in December
- Decided the cost/benefit for sustaining COOL through Run 3 is strongly favorable
- Will allow a slower adiabatic migration of subsystems to the new system, to take over during LS3
- Architecture of the new system, ‘Crest’, viewed favorably by the reviewers
- Recommended an ‘evolutionary’ approach
  - Adapt Crest to first serve as a REST service for COOL
  - Focuses on the biggest immediate issue with the present system, the scaling problems in the Frontier based infrastructure
- Current effort is too low to execute the plan, to be addressed
  - With less urgency given the stretched timeline, but we feel the effects now and need to address it. Effort needed!

2018 running plans

- ATLAS baseline 2018 running scenario established, maximizing the physics within the various constraints
  - Based on LHC plan, 25ns BCMS, 2500b, 1.3e11 ppb, L=2.2e34
  - Will level lumi to 2.0e34 (mu = 56)
  - Trigger menu based on 2017 1.7e34 menu
  - Implications for Tier-0, Castor tape storage, data export, downstream resources, etc were part of the considerations
- Our Tier-0 processing model shows 20% capacity shortfall for processing all of physicsMain
- Grid spillover will be fully commissioned, validated and used in steady state during 2018 running for e.g. for B physics stream
- Spillover will require more operational effort, for both computing and data prep
  - Up to 0.5 FTE additional for ADC ops
  - Second reprocessing coordinator needed

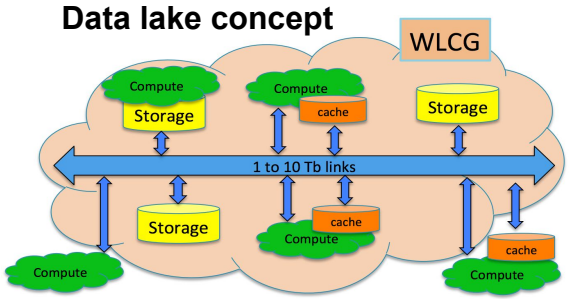


Strong wish to keep physicsMain processing at Tier-0 to ensure the timely integrity of the physics_Main data quality assessment cycle

CERN IT has responded favorably to our +20% Tier-0 CPU request

Looking forward

- R&D planning and activity informed by the white papers has begun
- Plans will develop at the joint WLCG-HSF workshop in March
- WLCG's HL-LHC strategy document coming end March
- One initiated project is in '**data lakes**'
  - Integrated consolidation of distributed storage (and compute) facilities, leveraging high-bandwidth networks
- R&D collaboration with CERN IT, led by Simone Campana
- ATLAS recently initiated a 'Data Ocean' R&D project with Google (lakes are too small for Google ;-)
  - Prototype and testbed for HL-LHC directed solutions
  - Provide near term value in e.g. using the Google cloud to store grid-produced analysis outputs: 100% availability the morning after a big run rather than 95%
- Dipping into the lake for a few slides...

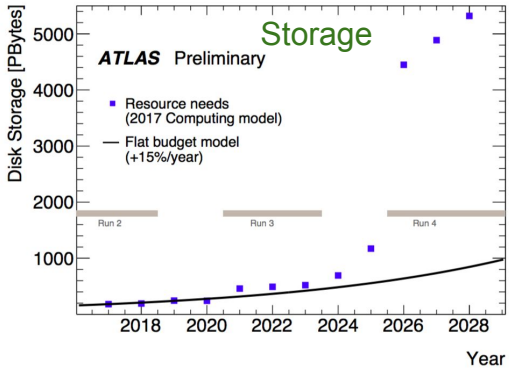


ATLAS - Google Data Ocean R&D project description

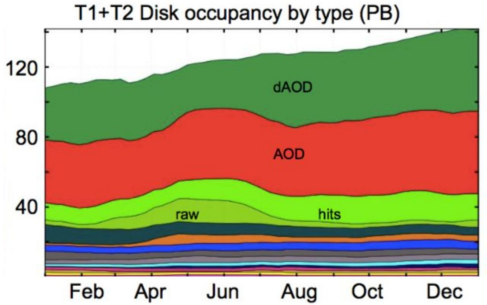
Scaling to HL-LHC: Storage

- **Storage is the biggest challenge for HL-LHC computing**
- O(10)x shortfall based on extrapolating current approaches has held steady
- We need new approaches!
- ATLAS disk usage is currently ⅓ reco output (AOD), ⅓ derived analysis objects (dAOD), and ⅓ everything else
- Within the '⅓ everything else' are samples that reside mostly on tape, rotating onto disk cache when needed for processing (e.g. simu hits)
- To reduce dramatically our storage footprint: extend this 'tape carousel' approach to AOD and ultimately dAOD
  - Or, make AODs 10x smaller a la CMS; not a cultural fit for ATLAS
- Tightly limiting replica counts won't get us all the way there
  - Having 1.0 replicas of the most current data on disk will be *too much*
- This is very difficult because tape introduces delay, and complicates workflow orchestration, and (d)AOD workflows are time critical and highly complex already
- Also, tape is a deep but geographically limited resource (Tier-1s), while our processing resources are much more widely distributed

Disk storage ~6x short at HL-LHC



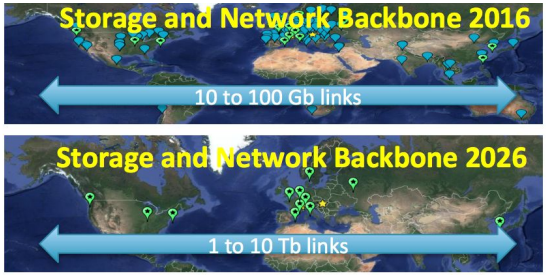
ATLAS disk usage 2017



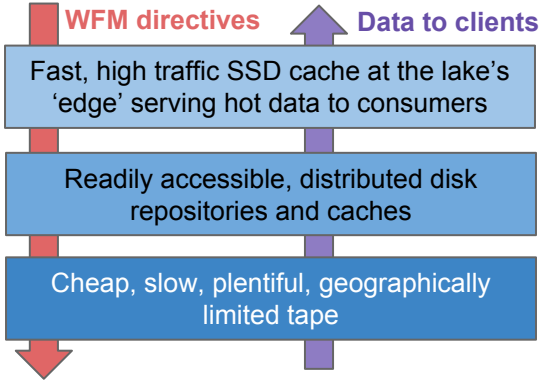
Solving the storage problem

Key elements to solving the storage problem:

- Our sites linked with (ever higher) high-bandwidth networking
- A '**data lake**': integrated consolidation of *distributed* storage (and compute) facilities, leveraging a high-bandwidth network
- The data lake encompasses facilities with several levels of storage, and can place data optimally according to (dynamic) need
  - **Tape**, at a relatively limited number of sites
  - Standard disk, at large storage repositories and smaller caches
  - Fast SSD cache for the hottest data
- PanDA/Prodsys has foreknowledge of the data to be processed
  - Use that knowledge to drive preparing needed data in the lake, transparently to the processing, e.g. tape staging, or placing hot data on fast SSD cache
  - Cache hot data 'close' to available processing
- We also know what data is hot/popular and factor that into automated placement decisions (ATLAS has led this approach for years)
- *Instead of >=1 replicas on disk today, aim for dynamic, managed availability of actively used data via the lake, replica count <<1*

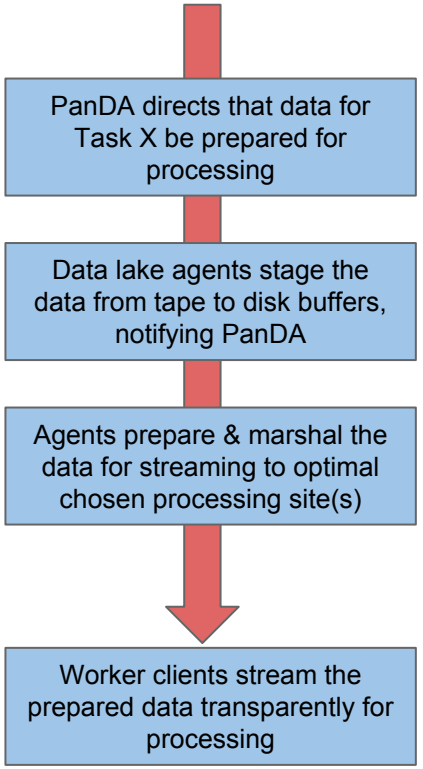


Data lake hierarchy



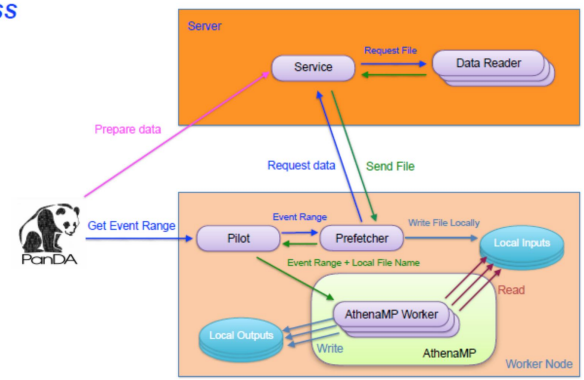
Serving data from the lake

- The data lake model depends on highly dynamic, no-waste data flows
  - Move only the data you need, when you need it, to a client ready to consume it
  - Hide the latencies involved from the processing
- Applies to processing going on both *in the take* and *outside the lake*
- Many resources are *in the lake*, sharing a fat pipe to lake-resident data
- Many, especially opportunistic and smaller tier resources, will be *outside the lake*
- Both inside/outside consumers can be most efficiently served by *streaming data flows* that do not require large files to be moved from A to B before processing can begin, and do not require that a complete large file be fully processed at B
- Instead, data *streams* from the source to the client. The streams
  - can use knowledge of the task to marshal and send only the needed data
  - begin immediately, and terminate when the processing resource goes away
  - are (re)directed to workers at different or multiple processing resources to complete tasks ASAP, without long slow tails
- Data streaming goes hand in hand with *fine-grained data processing* that partitions work into small pieces rather than the large-file level of traditional processing



Streaming from the lake

- The *fine-grained streaming* approaches to event processing that ATLAS is developing are a very good fit for the data lake model
  - Event Service (ES) is in early production for ATLAS simulation
  - Event Streaming Service (ESS) is in early R&D
  - Both clients for the emerging Event Whiteboard (EWB) now in early R&D and prototyping
- Support agile, dynamic and automatic processing and data flows so that work goes to the optimal locations of the moment, taking account of the resources available and the work to do
- Insulate processing from the latencies of the WAN by fetching data asynchronously and as-needed in near real time
- Enable the full and efficient utilization of opportunistic resources and HPCs; “the sand to fill the processing cracks”
- Opens the door to the ultimate storage saver (at a CPU cost), “virtual data”
  - Don’t save it, just (re)generate it when you need it
  - Can be feasible if MC simulation migrates to (mostly) “fast chain”, as is the ultimate plan



This week (a subset)

- EventIndex workshop on now
- Annual Sites Jamboree, a rich agenda
- Hackathon: progress on DataHandles milestone
- Conditions DB: Run 2, Frontier stress tests, DCS, Crest
- SIT day: half a comprehensive SIT session, half for CI experts
- Core: GeoModel ⇔ json, I/O, G4MT profiling, ...
- ADC+SW: I/O, SW on HPCs, workflow analysis, ART
- DCC: Event whiteboard, analysis preservation, …
- Event service: session on where to from here
- ICB meeting
- Fondue on Wednesday! Please doodle
- Friday plenary will be summaries & conclusions

Supplemental

Open software FAQ snapshot

- What are the pros of opening the software? The cons?
  - See our Pros and Cons document. Comments in the document are welcome. If there's an important Pro or Con we missed, let us know.
- Exactly what software (that isn't already open) will be opened?
  - The ATLAS Athena repository.
- By what process will the software be opened?
  - See the ATLAS Open Software Plan.
- I am a student at a UK university. Don't I have control over the copyright of the ATLAS code I write?
  - You may, and you may not. In many cases UK university policy includes exceptions to the general rule of students controlling the copyright. The spirit of of these exceptions as we read them (not being lawyers we cannot comment on the legality) cover cases like collaboration with ATLAS: when the student contributes code in the context of a project or collaboration that has its own copyright policy, that policy supersedes student ownership. Examples:
    - Glasgow: "The University's policy is that PGR students who are not employed by the University own their IP unless this is governed by a third party agreement (e.g. funding or sponsorship) or other factors which confer an interest in the IP."
      - https://www.gla.ac.uk/research/ourresearchenvironment/prs/intellectualproperty/.
    - Sheffield: "[Exceptions to student ownership include] where the Intellectual Property is generated as a result of collaborative work, for example with other students or with members of staff (or where the work being undertaken derives from the Intellectual Property of staff)
      - https://www.sheffield.ac.uk/lets/pp/policy/ip
    - Sussex: "It is recognised that where a student is externally sponsored or part of a research group that is subject to obligations to a funder, the terms of that sponsorship or funding may override this position, and require the student to assign to the sponsoring organisation or funder."
      - https://www.sussex.ac.uk/webteam/gateway/file.php?name=exploitation-policy-of-ip.pdf&site=377
  - If in your case you do control the copyright, see the answer below regarding Nordic universities, it applies to you.
- I work for a Nordic university. Don't I have control over the copyright of the ATLAS code I write?
  - Yes you do. Your rights include the right to transfer the copyright to that specified by the ATLAS Software Policy, and ATLAS requests that you respect the Collaboration's policy and transfer the copyright. If you decline to do so, nothing drastic will happen but ATLAS will review on a case by case basis whether the code should be listed for eventual removal/replacement (on a timeline determined case by case).
- How do I transfer copyright to that of the ATLAS Software Policy?
  - Use our copyright in your code and/or submit a merge request to change the copyright of already committed code to ours.