ATLAS Computing: projections

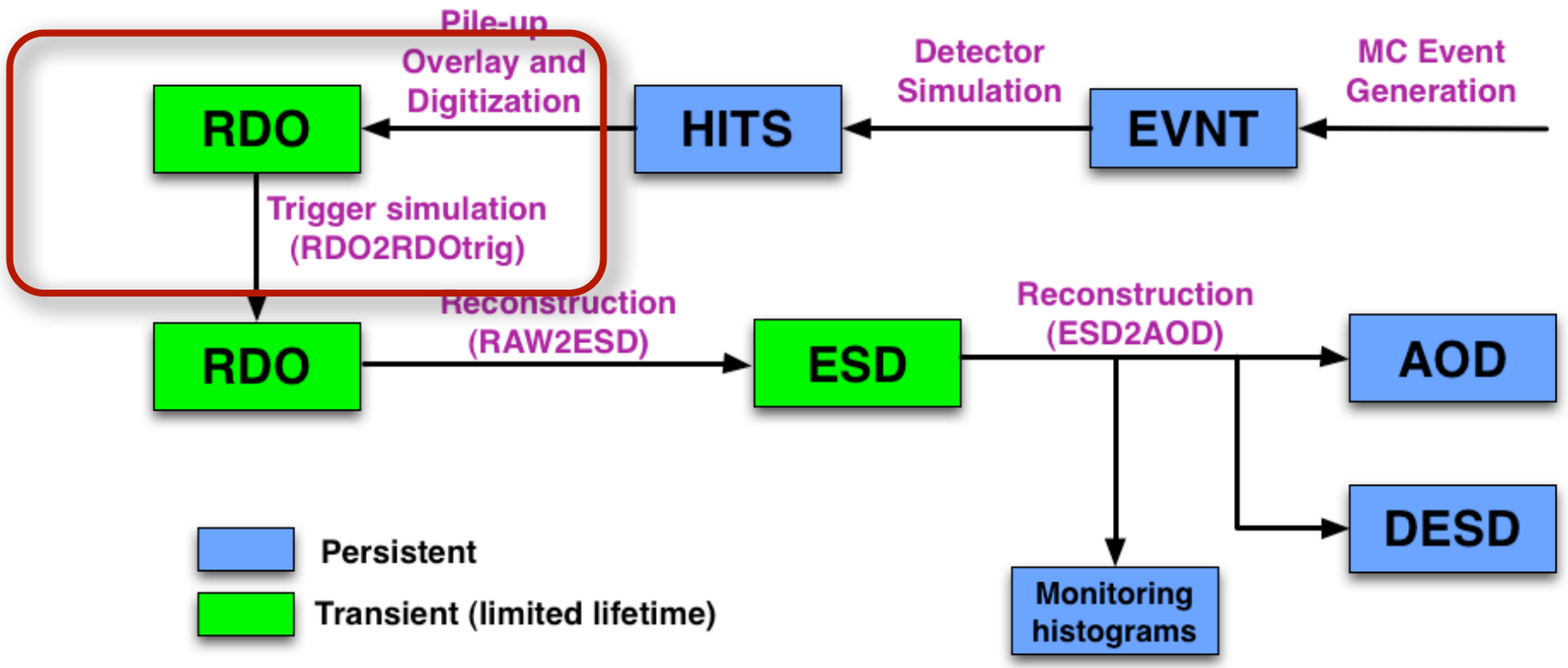
Roger Jones

Computing Model Update for Run 2

- **LHCC asked the four experiments and WLCG to provide a common document with updated computing models, demonstrating that we are doing our best to optimally use the resources foreseen for Run 2 (2015-2017) in the difficult financial climate (i.e. having 'flat budgets' for Computing resources).**
- Focus on the updates to the model and related resource implications.
  - The resource projections might still be refined, with additional information and synchronization with other experiments (e.g. on running conditions).
- Focus is on introducing more **dynamic** (as opposed to static) resource handling by **using operational experience from Run 1** combined with further software development, as an example:
  - **Analyses take a while to complete, i.e. can be a year between data taken and the analysis finishing - and the data and MC need to stay on disks.**

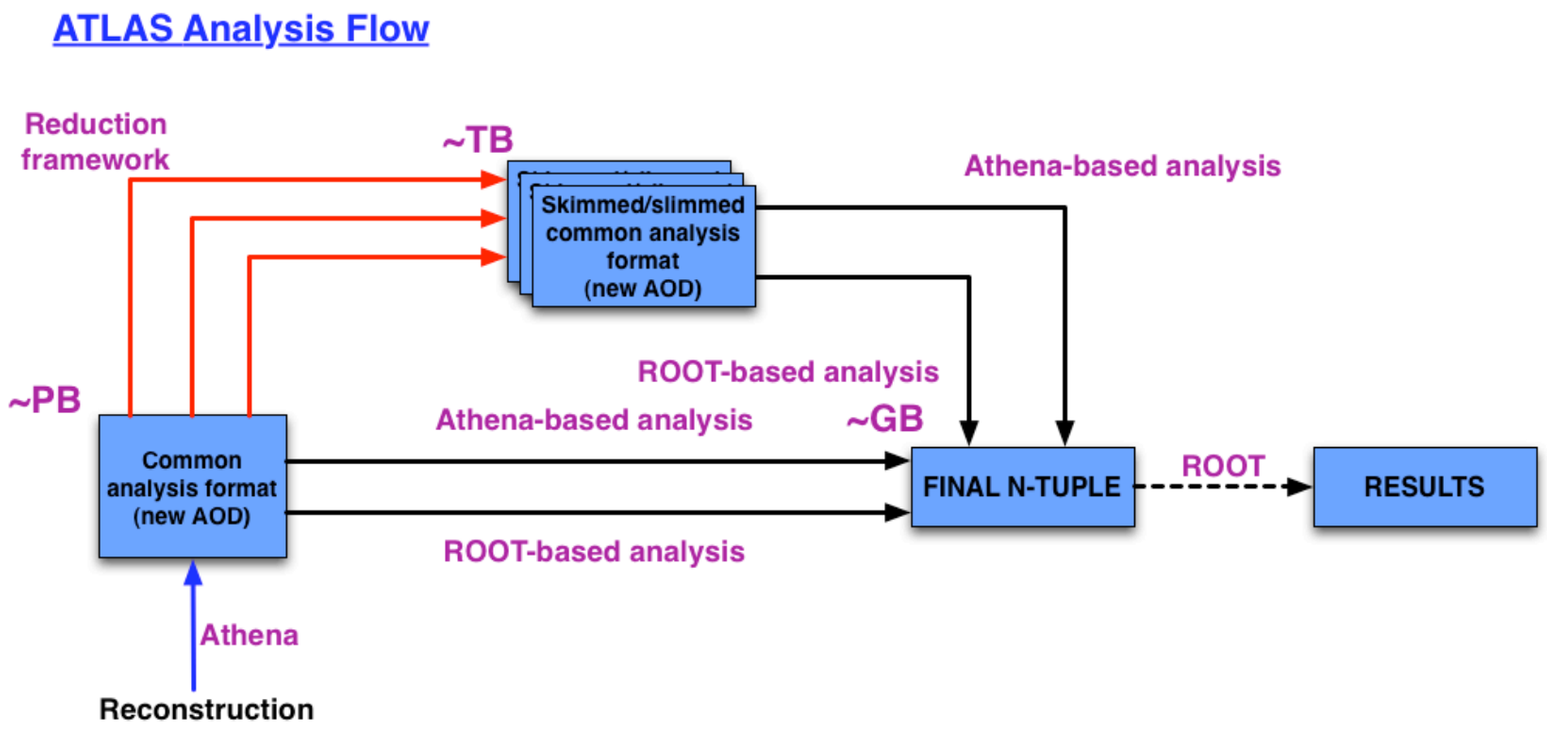
Software Improvements

- **We rely heavily on software improvements to keep within a reasonable resource budget.**
  - The processing times (CPU/event) will in reconstruction need to improve by at least a factor of 2 compared to direct use of current Run 1 software in Run 2 conditions (**1 kHz data recording rate, 14 TeV, 25 ns bunch spacing, average pile-up of 25 to 40**).
  - The event sizes will need to be optimized to be kept (approximately) at Run-1 values.
  - Work is needed on memory consumption of the reconstruction chain.
  - The development of fast simulation flavors within Integrated Simulation Framework as well as fast digitization and fast reconstruction is of high importance.
    - We will have to **rely more on fast simulation** than in Run 1 to achieve reasonable Monte Carlo statistics for analysis.

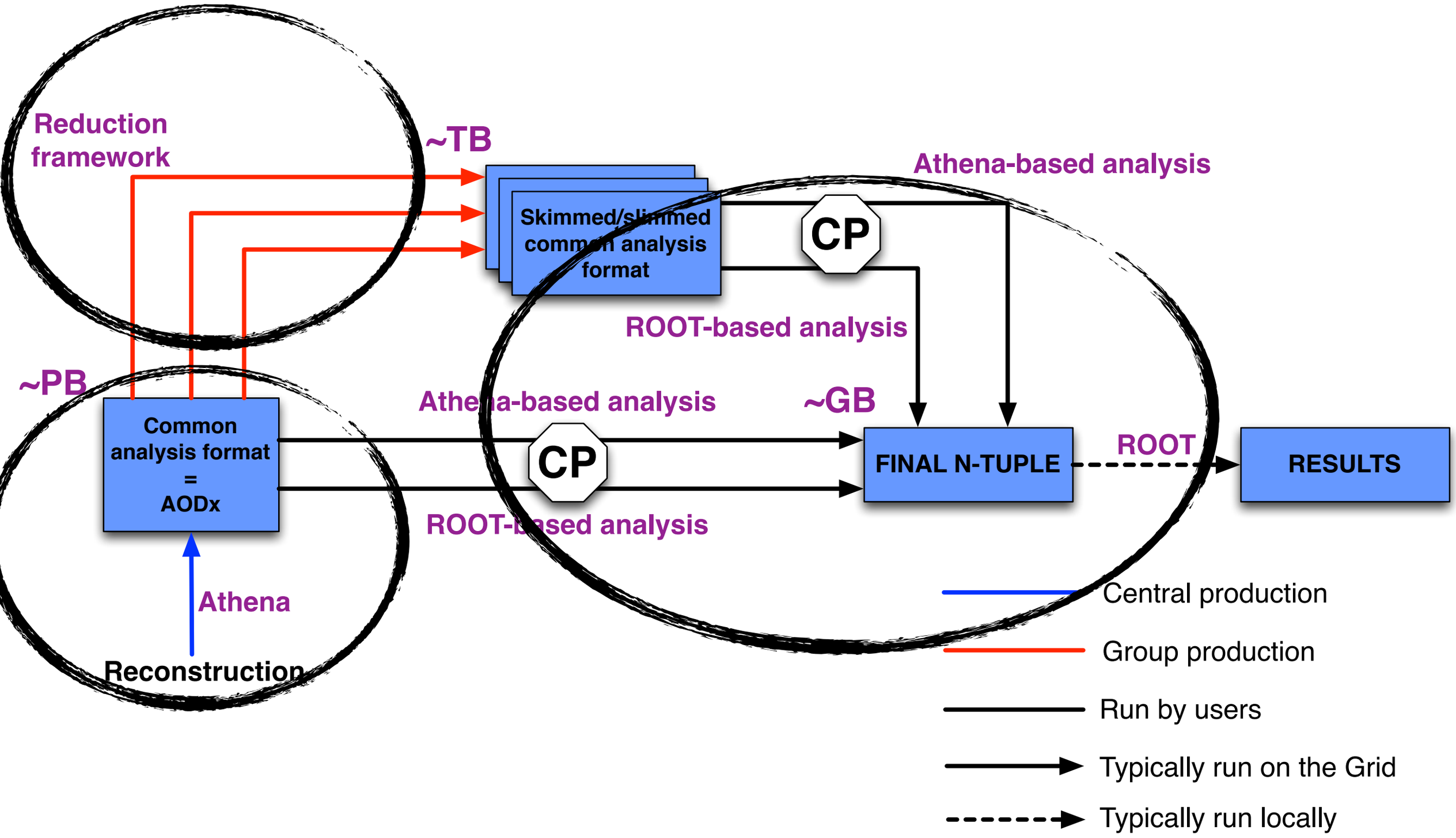


Workflow Changes

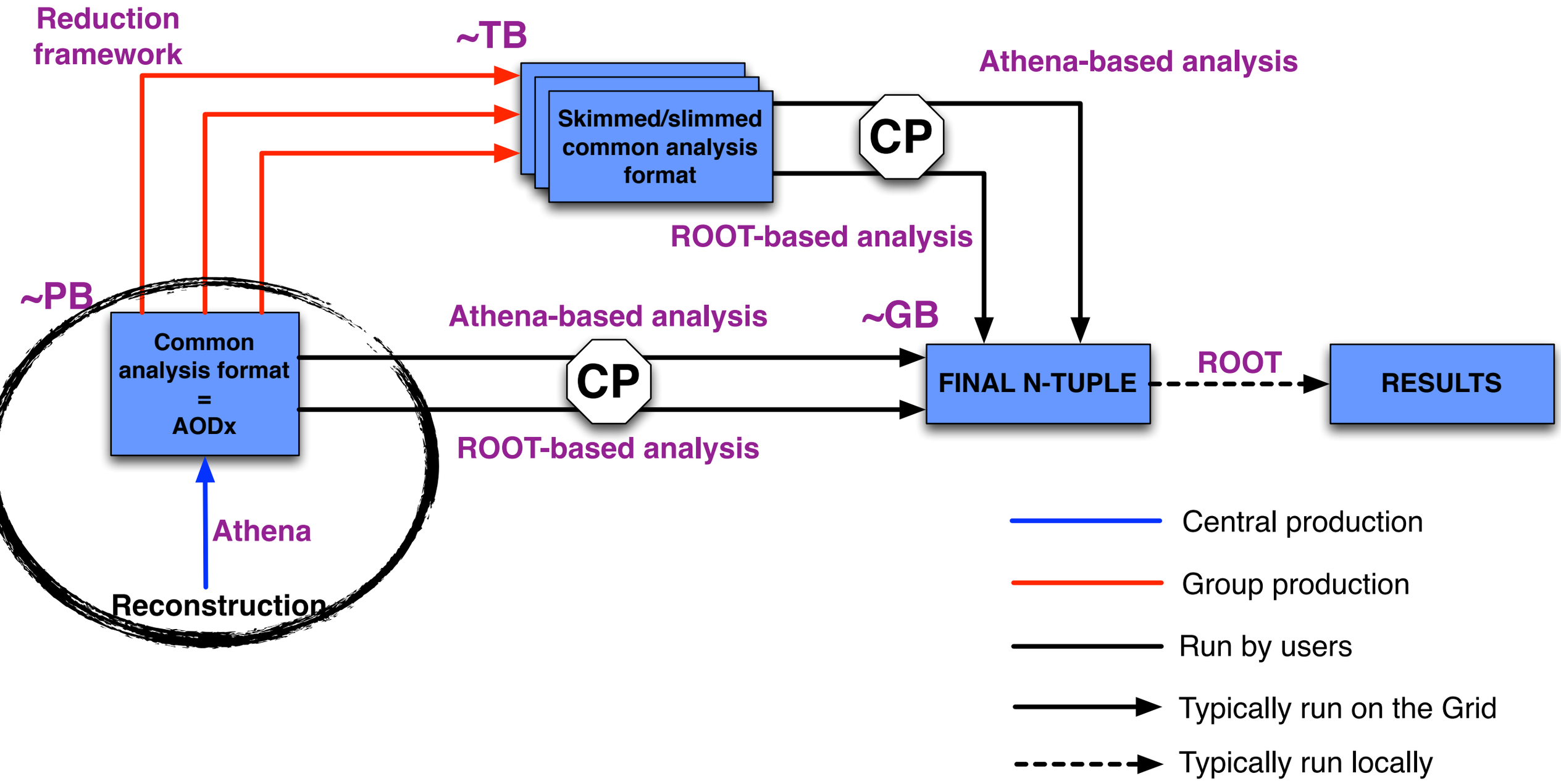
- The changes in workflows are motivated by the Analysis Model Study Group report which are being implemented for Run-2; in brief:
  - Updating the AOD format for the analysis objects to become directly/simply ROOT-readable.
  - Introduction of AOD2AOD reprocessing(s), to optimize the resource usage .
  - Introduction of an ADC-integrated 'Derivation framework', to produce the group specific formats by skimming/slimming/thinning the AODs in a train model (replacing the current DPD production).
  - Introduction of a common analysis framework, providing uniform interfaces to analysis tools and simpler/better Grid integration.



What is the New Analysis Model?



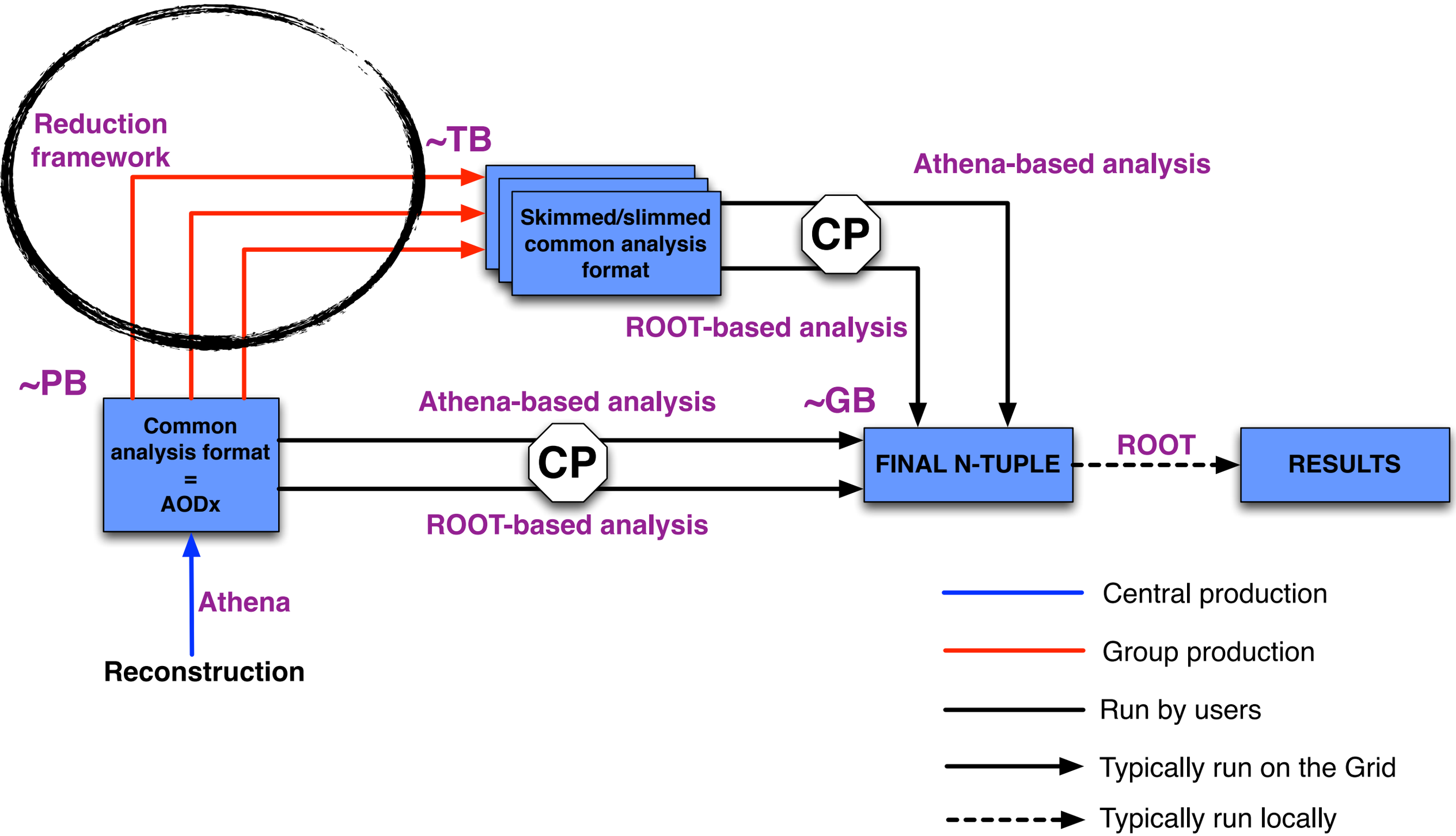
What is the new analysis format?



What is the new analysis format?

- **Combine all of the analysis properties from the existing AOD and D3PDs into a single format**
- **Why? Because we cannot sustain all these D3PDs (too little storage, too many versions/replicas)**
  - **Simplifies the CP & analysis tools (currently poorly documented)**
  - **Easier to learn and use**
  - **Fast - D3PD speeds**
  - **Flexible - easy and fast to drop objects and variables**
- **Must be readable with ROOT as well as Athena - the aim is for vanilla root**
  - **I suspect we may have a recompiled version**
- **Compatible with PROOF**
  - **Progress towards this about half-way; next step a stand-alone demonstrator**
  - **XAOD ready for 2015 data**

What is the Derivation Framework?



What is the Derivation Framework?

- **Derivation framework: Athena-based software for processing reconstructed data/MC in bulk to produce smaller formats suitable for specific physics analyses, by**
  - **skimming (removing whole events)**
  - **slimming (removing data from events)**
  - **thinning (removing objects from events)**
  - **augmenting with group-defined data**

UK involved

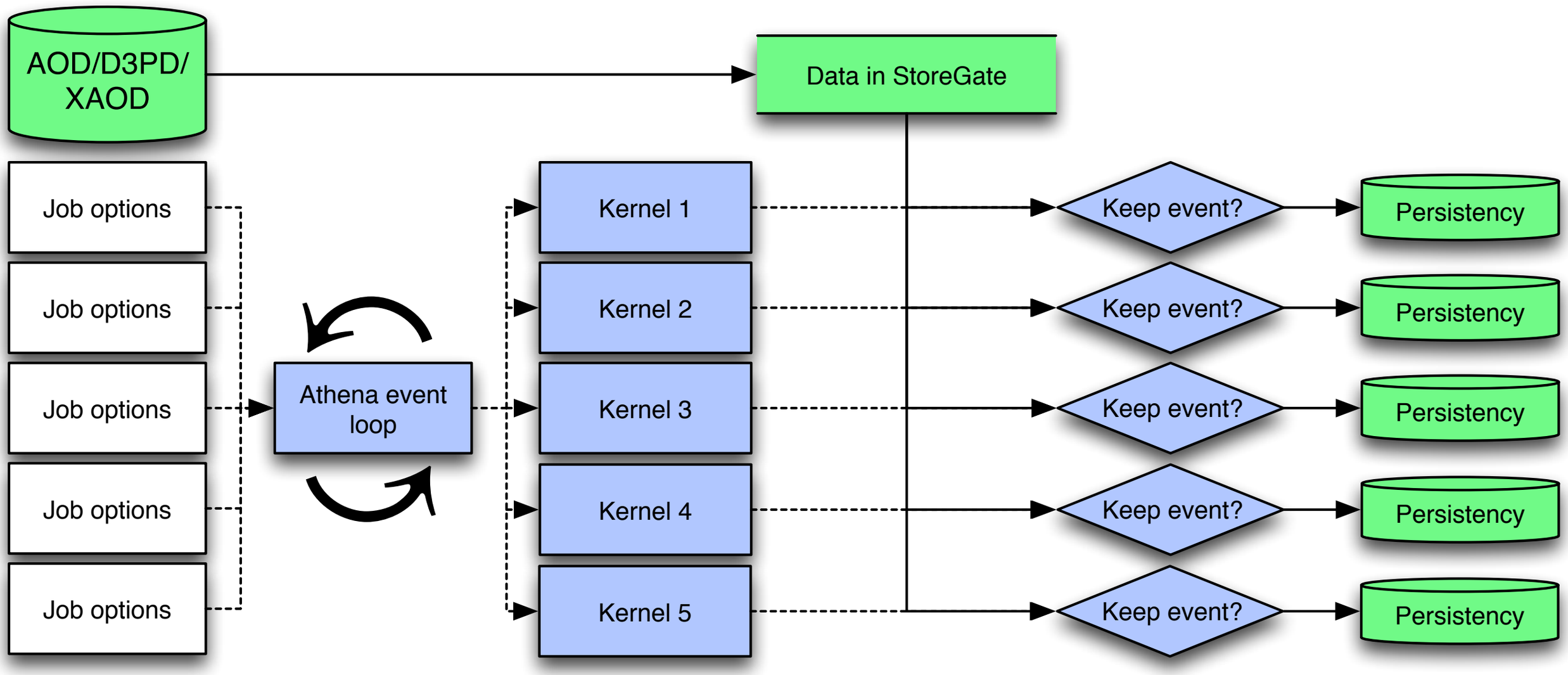
- **Will be primarily run on the production system in a centrally organized way, although it can also be run locally as with all ATLAS software**
- **More than 100 different derivations, tailored to specific analyses, are foreseen for 2015, each of which should have a total size of ~1-10TB**
- **Multiple formats being produced in the same ProdSys task (train model)**
- **Intention is to use it in 2013-2014 for deriving from either AOD or NTUP_COMMON, moving to XAOD in time for the data challenges**
  - **So before XAOD it will be a two-headed beast....**

Train production

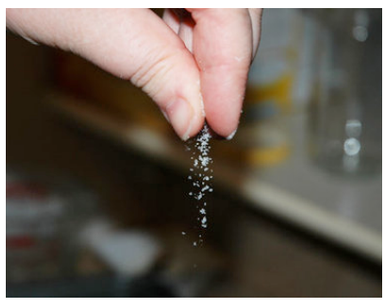
- **A train departs on a regular basis and has several carriages**
  - **Train production = several/many derivations (carriages) in each job; tasks are launched on a regular basis**
  - **In this way the requirements of several groups can be met at once, rather than having to set up separate tasks for each individual derivation**
  - **Can have one-off trains, open ended trains that depart as soon as a run finishes, fast trains with very light CPU requirements, slow trains with CPU-heavy tasks, etc**
  - **Scheduling and configuration of trains will be the responsibility**

Depending on where you are from, your notion of train production may vary....

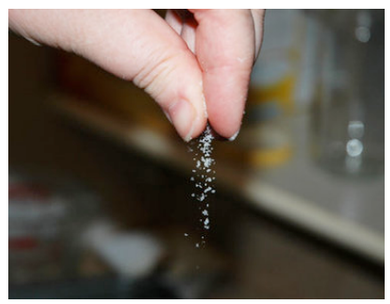
Train production



How many carriages can we have?



1000 muon stream AOD events on interactive LXPLUS
Each carriage is *identical* in this test:
selecting all events with ≥3 muons, pT≥1GeV
(26 events accepted in each carriage)



Number of carriages	Wall clock [min] whole job	Average CPU [ms] per event step	Fraction (CPU*nEvents)/ (wall clock)	VMEM [MB] at finalize
2	2.8	130	77%	1410
50	6.4	296	77%	3102

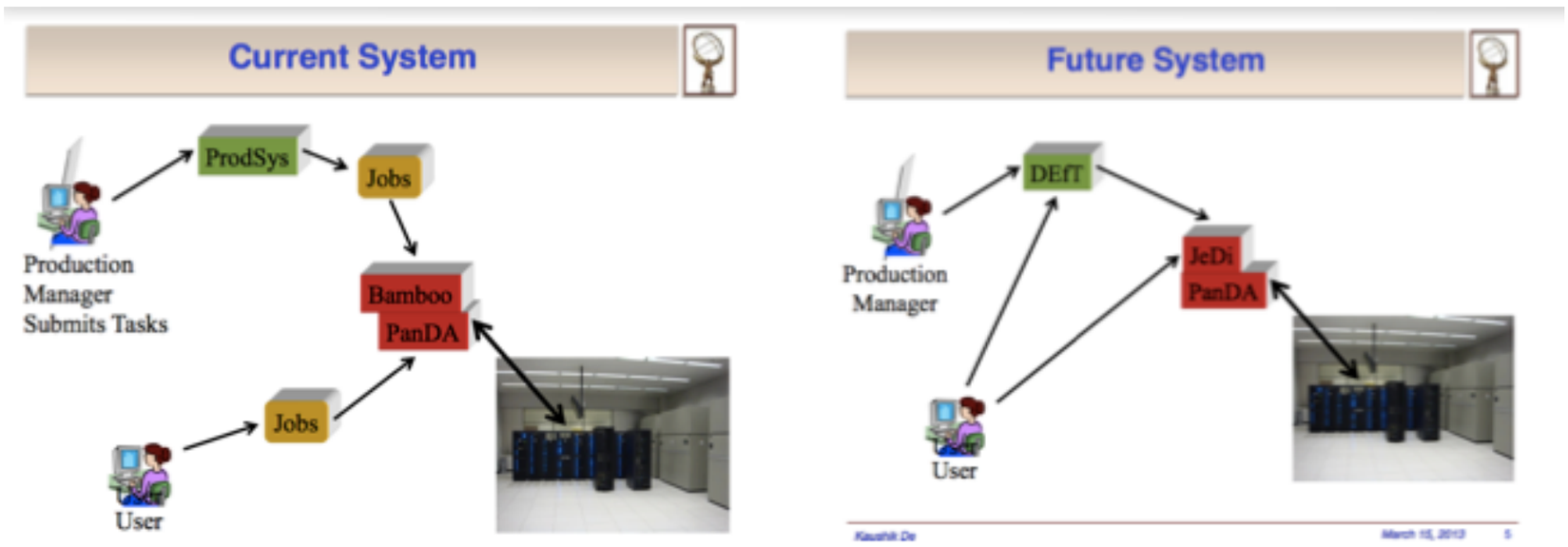
The hope is for ~100 streams
One train every 2 weeks
Fast and slow trains will be required

Data Placement Policies

- Replication and data placement policies remain largely the same as in Run 1, with reductions/corrections/improvements:
  - **The number of pre-placed (primary) replicas of AODs will go down to 1 replica at Tier-1 and 1 replica at Tier-2.**
    - **We will rely even more on popularity-based dynamic data placement (PD2P) and automatic cleanup of secondary copies (Victor).**
    - **Consequently, more emphasis on the network use is expected.**
  - **We will introduce an automated algorithm to clean (set to secondary) the pre-placed replicas, for example:**
    - If a data set was not accessed for six months set the Tier-1 replica as secondary.
    - If a data set was not accessed for one year set the Tier-2 replica as secondary.
    - The parameters will need operational fine-tuning.
    - **As a consequence example, we will use tape retrieval from Tier-1s more :**
      - **e.g. the unpopular AODs will remain only on tape, retrieved for AOD2AOD reprocessing or derivation of group-specific data..**

Rucio and Prodsys-2

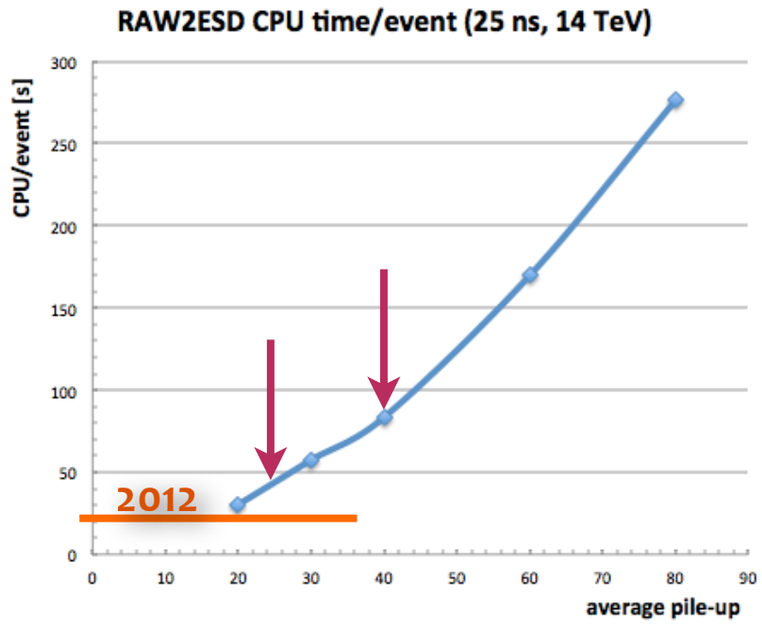
- A few points to stress:
  - **The new data management system (Rucio)** will introduce more granular handling of ATLAS data (file replicas rather than dataset replicas) together with a more intelligent implementation of data ownership for users and groups. This will allow optimization of the space usage while offering more functionalities in terms of local and global quotas, thus e.g. **relaxing the need for dedicated group storage SRM space tokens at sites. Dedicated sites will continue.**
  - **The new PanDA+Jedi+DEfT** will simplify handling of different workflows (e.g. production vs analysis jobs and pilots).
  - The combination of Rucio and Prodsys-2, will also provide the functionality for the **use of federated resources and remote access**, i.e. xRootD, HTTP federations…
  - "Event Server", combining functionalities of Panda/Prodsys2, the xROOTd federation and AthenaMP, is under development for Run 2.

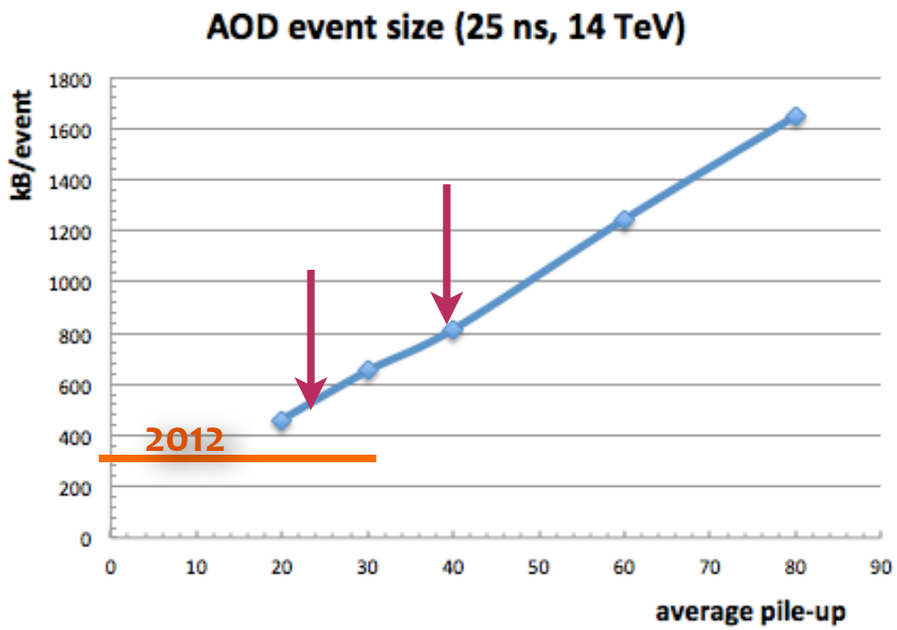


Model Parameters

- The LHC scenario agreed between experiments
  - **A lot** of work needed to reach the target CPU times/event, event sizes etc..
    - **needed** to achieve a reasonable match to available resources!

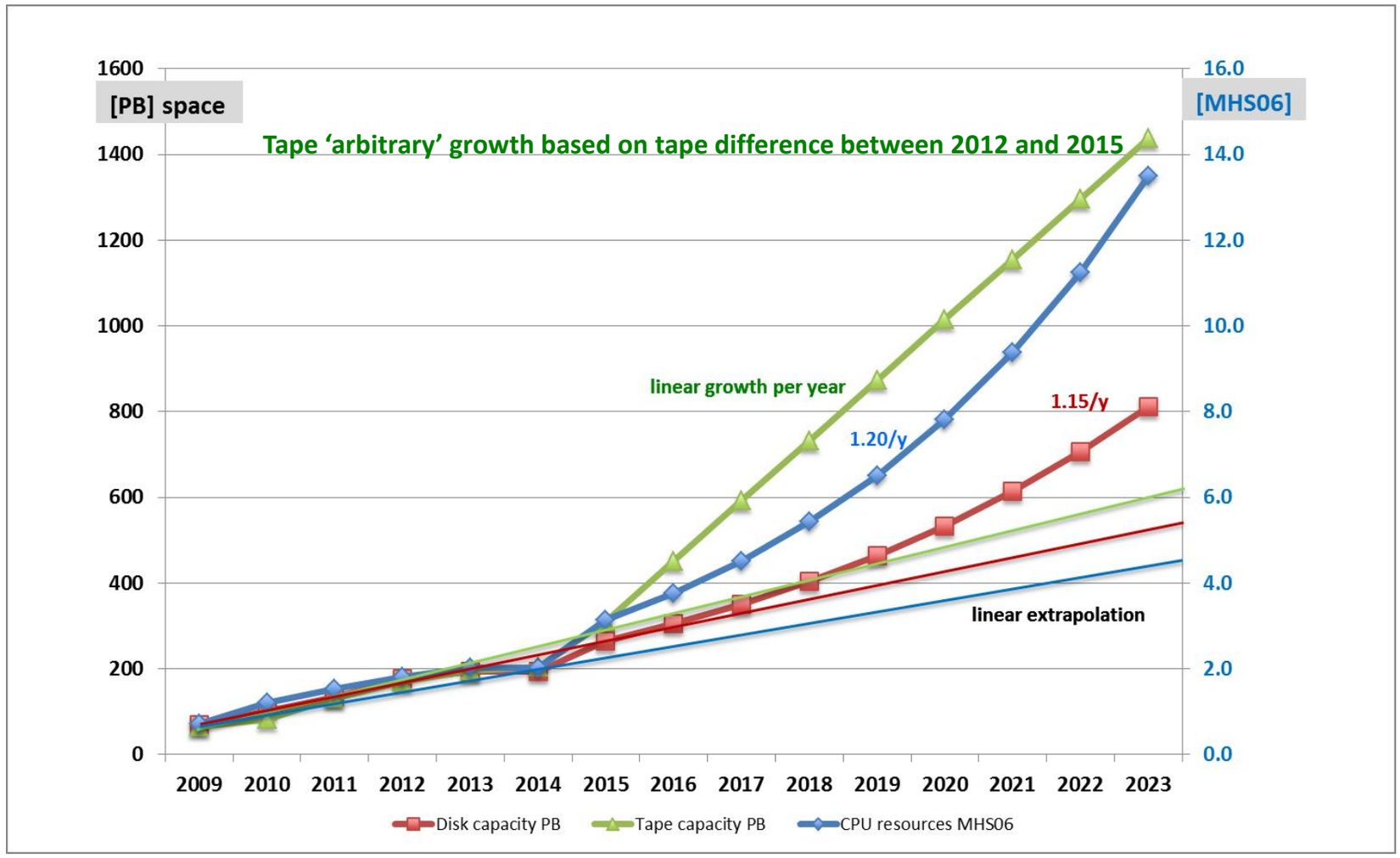
LHC and data taking parameters		2012 pp actual	2015 pp mu=25 @ 25 ns	2016 pp mu=40 @ 25 ns	2017 pp mu=40 @ 25 ns
Rate [Hz]	Hz	400 + 150 (delayed)	1000	1000	1000
Time [sec]	MSeconds	6.6	3.0	5.0	7.0
Real data	B Events	3.0 + 0.9 (delayed)	3.0	5.0	7.0
Full Simulation	B Events	2.6 (8 TeV) + 0.8 (7 TeV)	2	2	2
Fast Simulation	B Events	1.9 (8TeV) + 1 (7 TeV)	5	5	5
Simulated Data					
Event sizes					
Real RAW	MB	0.8	0.8	1	1
Real ESD	MB	2.4	2.5	2.7	2.7
Real AOD	MB	0.24	0.25	0.35	0.35
Sim HITS	MB	0.9	1	1	1
Sim ESD	MB	3.3	3.5	3.7	3.7
Sim AOD	MB	0.4	0.4	0.55	0.55
Sim RDO	MB	3.3	3.5	3.7	3.7
CPU times per event					
Full sim	HS06 sec	3100	3500	3500	3500
Fast sim	HS06 sec	260	300	300	300
Real recon	HS06 sec	190	190	250	250
Sim recon	HS06 sec	770	500	600	600
AOD2AOD data	HS06 sec	0	19	25	25
AOD2AOD sim	HS06 sec	0	50	60	60
Group analysis	HS06 sec	40	2	3	3
User analysis	HS06 sec	0.4	0.4	0.4	0.4





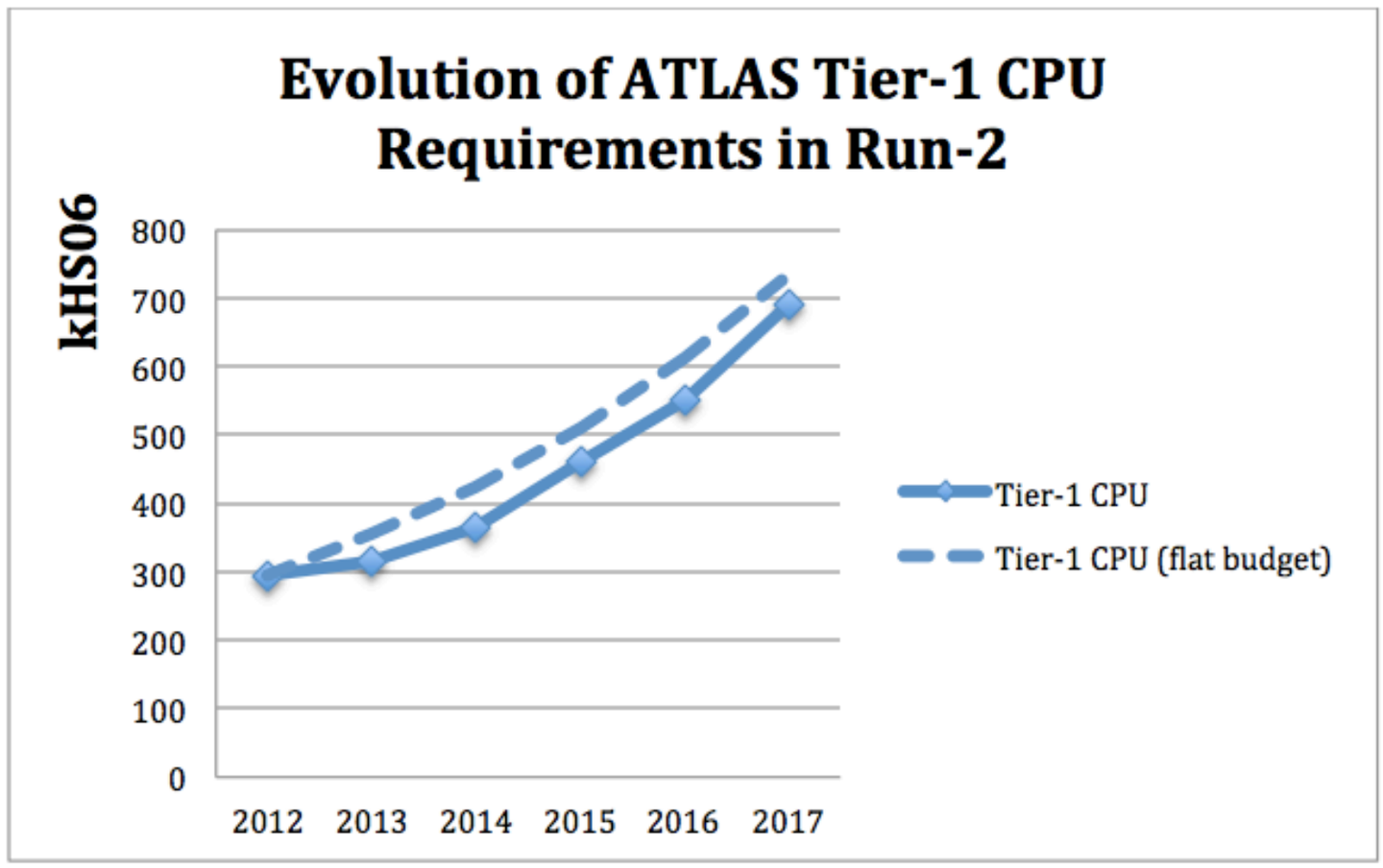
‘Flat Budget Interpretation’

- The ‘flat budget’ resource increase projections made by Bernd Panzer
  - evaluated at factors of **1.2/year for CPU and 1.15/year for disk.**
  - this appears in the technology chapter of the LHCC document, the plot is taken from that draft.



Resource Evolution (preliminary)

Tier-1 CPU (kHS06)	*2015*	*2016*	*2017*
Re-processing	38	30	43
Simulation production	154	89	102
Simulation reconstruction	194	245	280
Group (+user) activities	76	187	267
Total	***462 [478]***	***552***	***691***

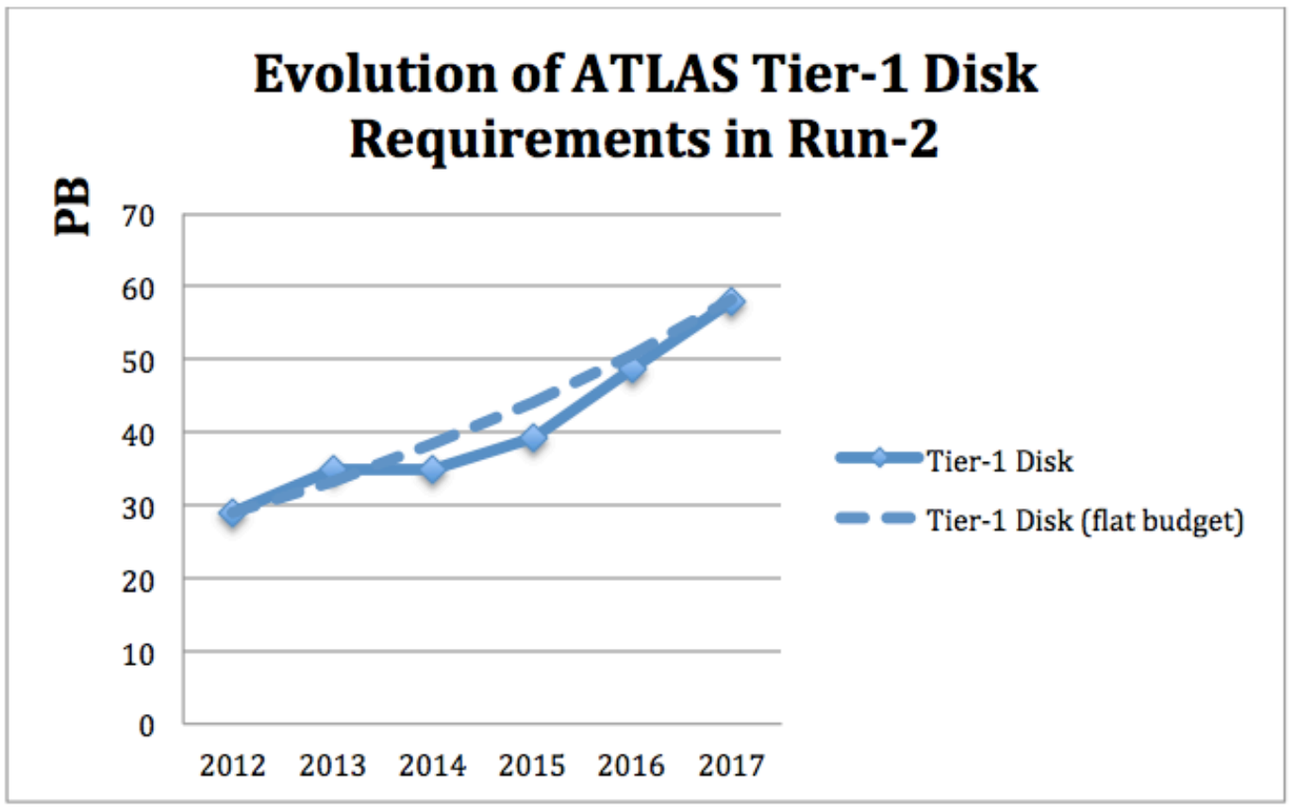


Flat CPU budget: factor 1.2/year

Resource Evolution cont'd

Tier-1 Disk (PB)	*2015*	*2016*	*2017*
Current RAW data	2.4	5.0	7.0
Real ESD+AOD+DPD data	5.6	7.9	11.1
Simulated RAW+ESD+AOD+DPD data	9.2	11.4	11.4
Calibration and alignment outputs	0.3	0.3	0.3
Group data	7.5	8.0	10.4
User data (scratch)	2.0	2.0	2.0
Cosmics	0.2	0.2	0.2
Processing and I/O buffers	3.0	3.0	3.0
Dynamic data buffers (30%)	9.0	10.9	12.6
Total	***39 [47]***	***49***	***58***

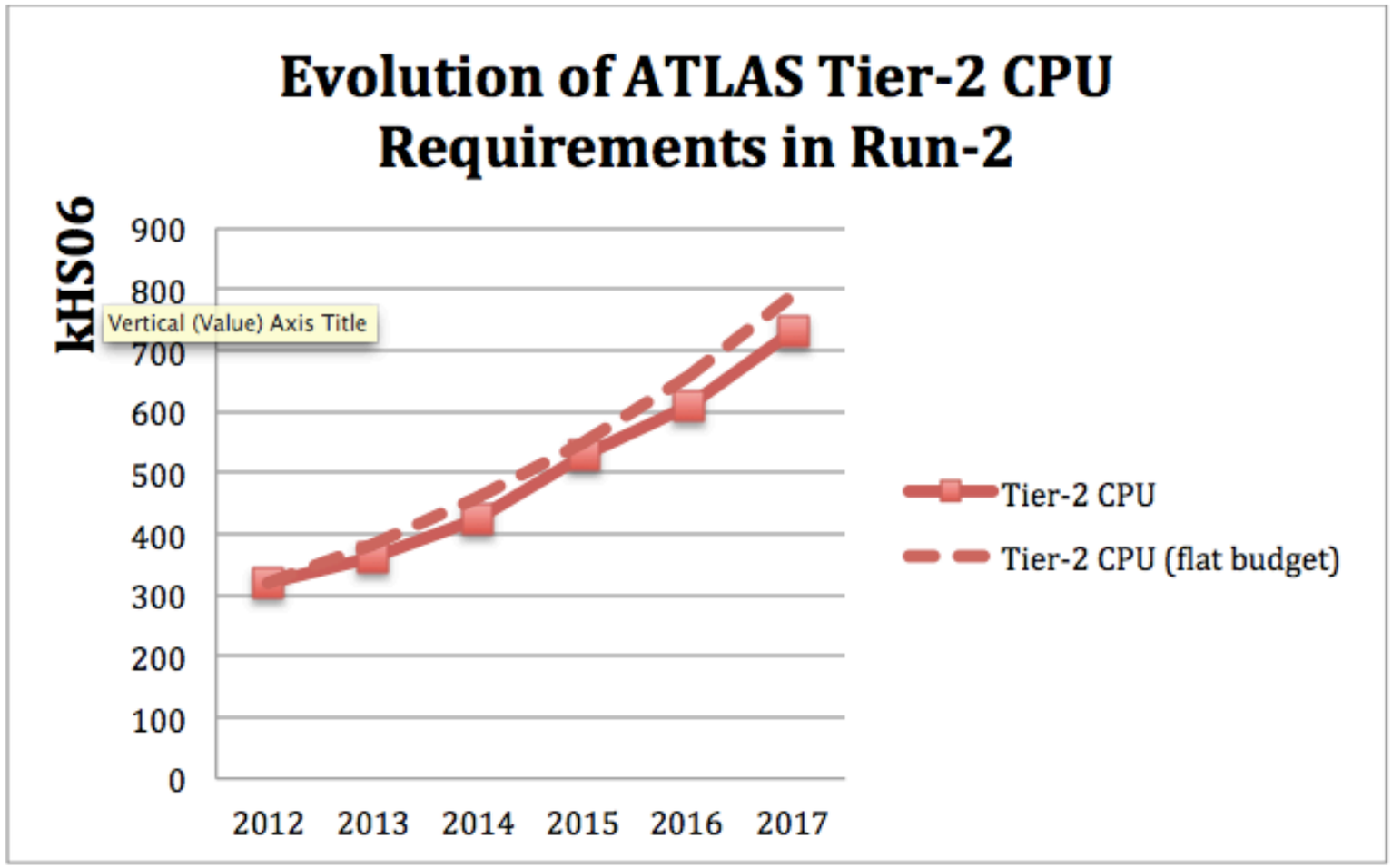
Tier-1 Tape (PB) Cumulative	*2015*	*2016*	*2017*
Real RAW+AOD+DPD data	18	28	42
Cosmics and other data	4	4	4
Group + User	7	8	9
Simulated HITS+AOD data	39	53	66
Total	***68 [74]***	***92***	***121***



Flat disk budget: factor 1.15/year

Resource Evolution cont'd

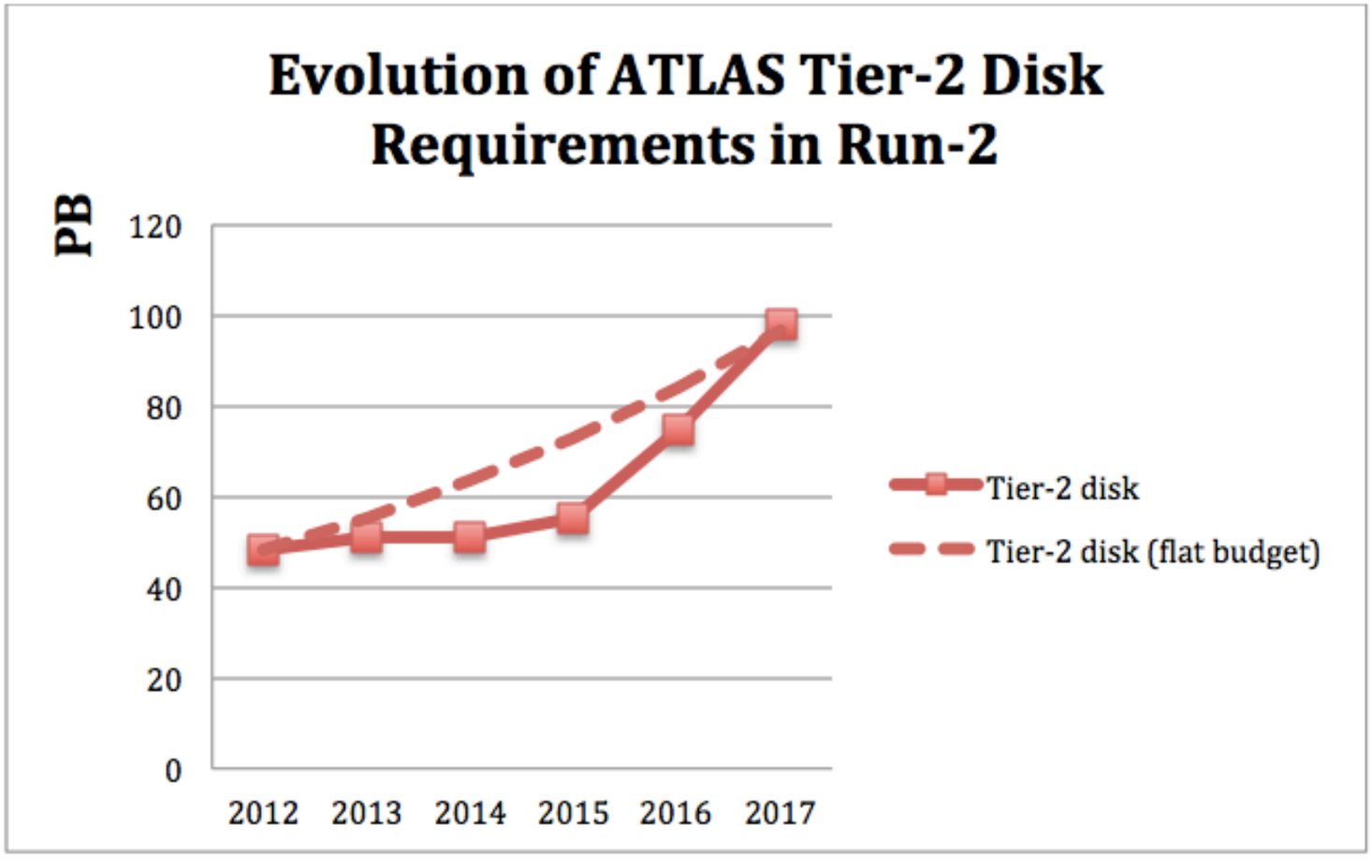
Tier-2 CPU (kHS06)	*2015*	*2016*	*2017*
Re-processing	20	33	47
Simulation production	338	347	396
Simulation reconstruction	77	61	70
Group + User activities	96	166	219
Total	***530 [522]***	***608***	***732***



Flat CPU budget: factor 1.2/year

Resource Evolution cont'd

Tier-2 Disk (PB)	*2015*	*2016*	*2017*
Real AOD+DPD data	4.1	6.3	10.6
Simulated HITS+RDO+ESD+AOD	10.6	16.6	21.6
Calibration and alignment outputs	0.2	0.2	0.2
Group data	20.4	29.3	41.6
User data (scratch)	4.0	4.0	4.0
Processing and I/O buffers	3.0	3.0	3.0
Dynamic data buffers (30%)	12.7	15.3	16.8
Total	***55 [65]***	***75***	***98***



Flat disk budget: factor 1.15/year

Roles of Tiers

- **In Run 1 the border between Tier-1s and Tier-2s in terms of performing different job types became 'soft':**
  - This will continue in Run 2, i.e. workflows will be optimally shared between Tier-1s and Tier-2s to optimize disk and CPU availability, network traffic…
- **With the technological progress of WAN ATLAS has already in Run 1 gradually moved away from the hierarchical association of Tier-2s to one Tier-1 (MONARC model) and is now able to associate workflows between a well-connected Tier-2 and several Tier-1s, based on monitored connectivity metrics.**
  - This functionality will be further pursued in Run 2 and is being incorporated into the production system and distributed data management developments.
- **Tier-1s continue to keep the essential role as archival facilities and are being relied upon for the most difficult workflows (I/O,CPU,memory..)**
  - A new feature in Run 2 is the envisaged scenario of **'overspilling' the prompt reconstruction from Tier-0 in case of CPU congestions to Tier-1s.**
    - In order to ensure the readiness of Tier-1s for the prompt data processing a small stream of data will be constantly processed at Tier-1s and validated by comparing the outputs with the Tier-0 processing outputs.

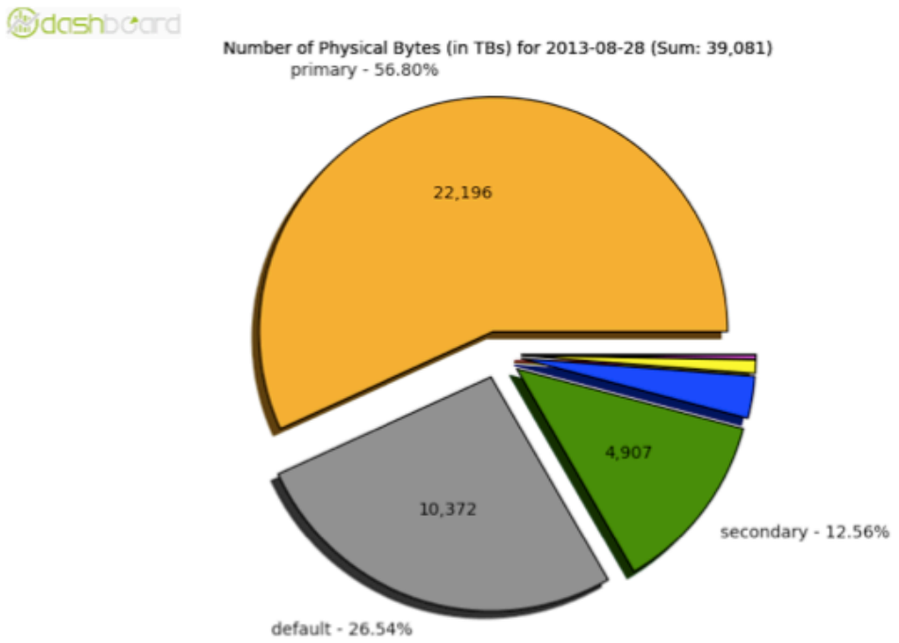
Tier-2 Status

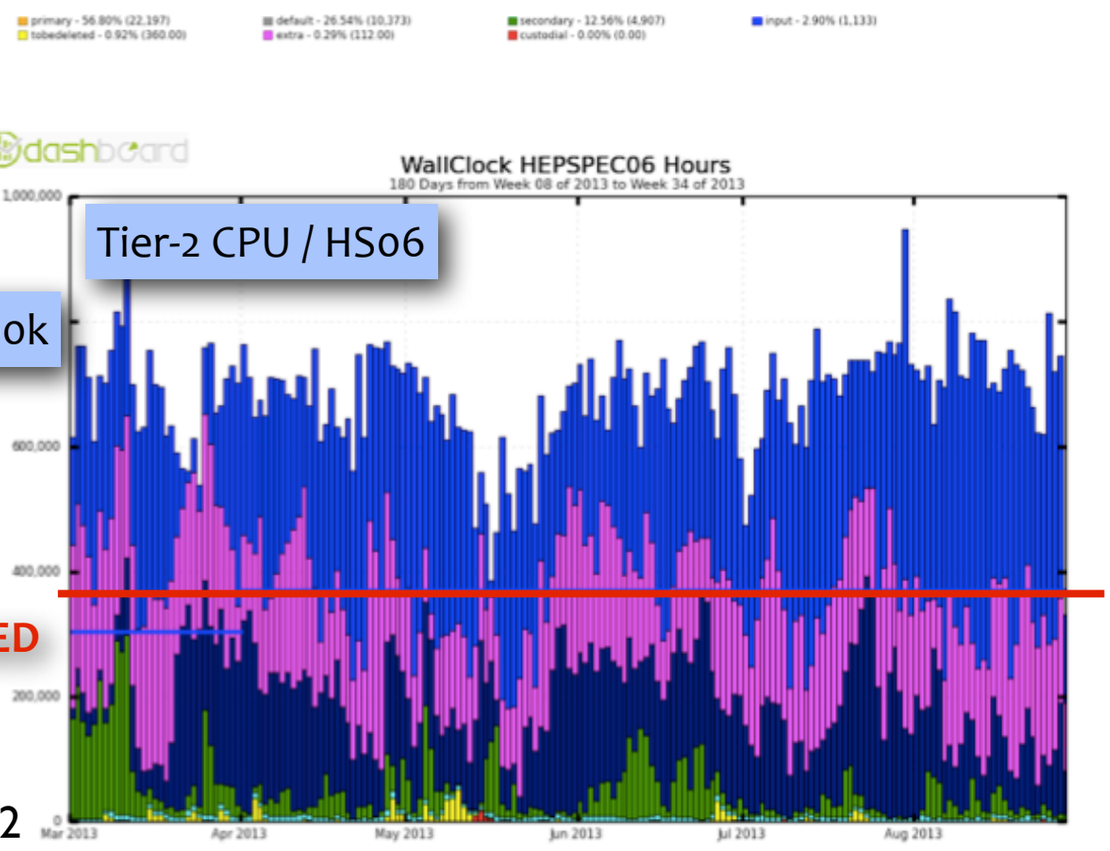
WE HAVE MORE DISK AND CPU THAN PLEDGED, MANY THANKS TO TIER-2s!

Tier-2 Disk [PB]	*Requested*	*Used*
Real AOD+DPD	7	8
Simulated data	21	19
Calibration and alignment output	0.2	0.2
Group+User data	20	13
Processing, scratch and I/O buffers	3.4	7.2
Total	***51 [52]***	***48***

Tier-2 CPU [kHS06]	*Requested*	*Used*
Simulation Production	158	308
Simulation Reconstruction	42	207
Group + User activities	160	198
Total	***360***	***713***

DISKS ARE ALSO FULL - WE ARE TAKING ACTIONS



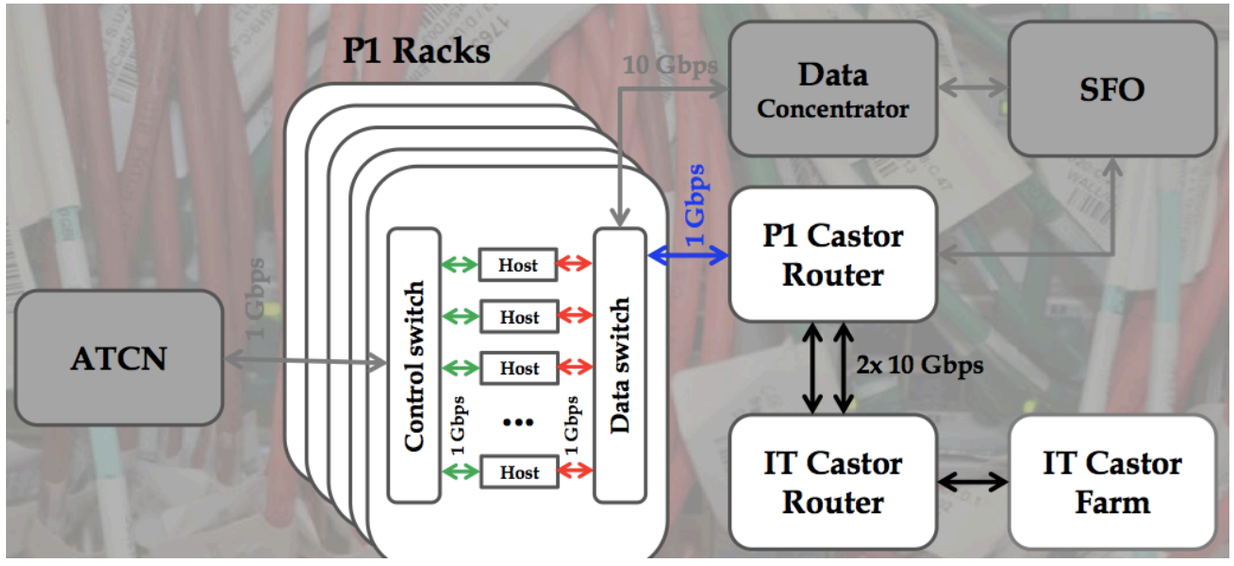


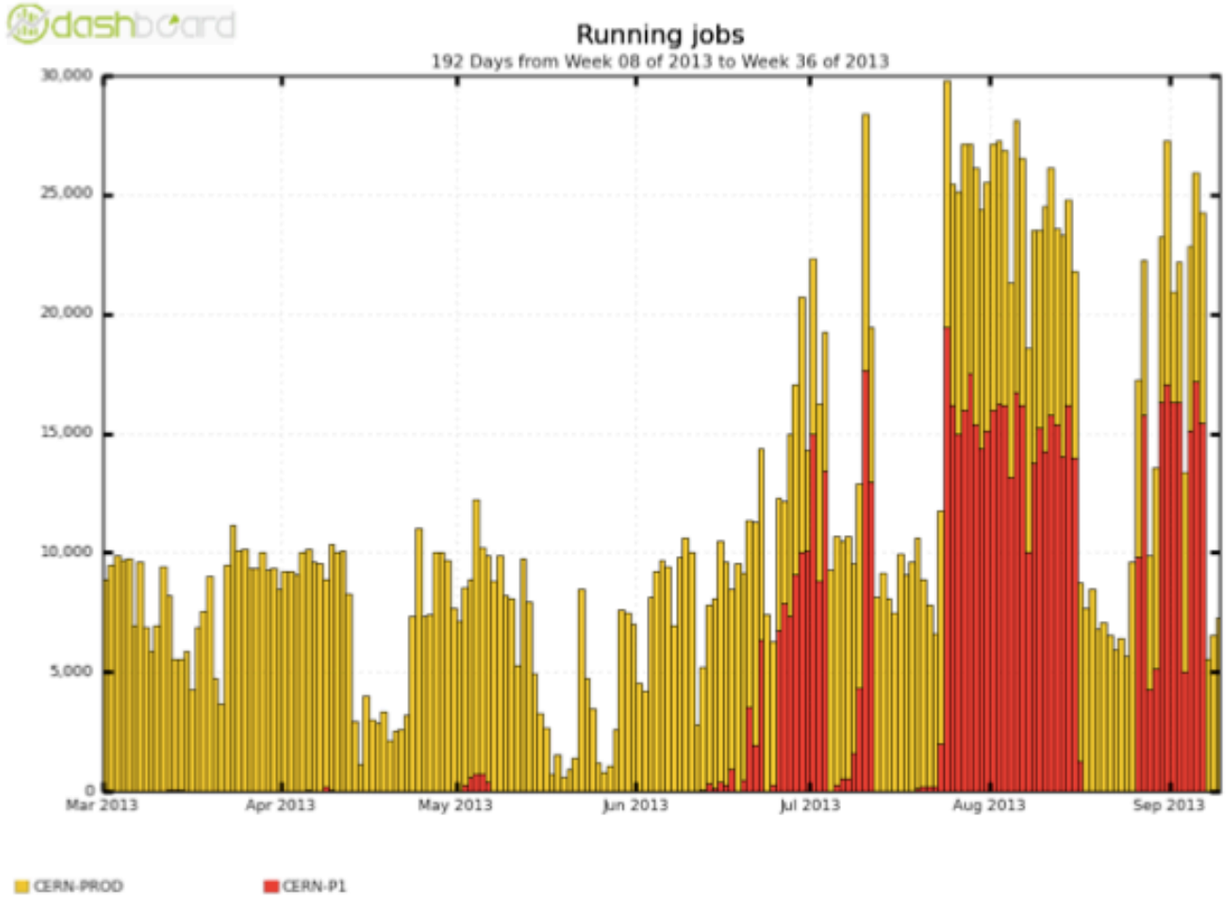
Opportunistic and Un-pledged Resources

- The resource planning is based upon the physics programme that can be accomplished within **achievable pledged resources, corresponding to a 'flat' spending budget**; the hope is that our centres and funding agencies will continue to provide ATLAS with the **invaluable resources beyond those pledged that will allow us to accomplish an optimal research programme and physics productivity.**
  - **With 'flat' resources we reach the MC/data ratio of 1.4, in 2012 data analyses the ratio is by now equal to 2.**
- In Run 2, more work on **opportunistic 'Cloud'/IAAS resources or High Performance Computing (HPC) centres** *for appropriate workloads* and the creation of a transparent and straightforward interface between these and the ATLAS distributed computing environment.
  - to complement our existing resources to **off-load CPU intensive and low I/O workloads which under-use our I/O optimized grid resources.** A typical example would be 'Monte Carlo generation of hard-process events' (Alpgen, Sherpa Monte Carlo event generators), which took ~15% of CPU resources on the ATLAS Grid sites in 2012. The Grid resources can then be better used for I/O intensive work: (simulation, reconstruction, group production, analysis), important especially during peak demand.
  - In addition, **if ATLAS Grid sites want to move to use cloud middleware** (e.g. OpenStack) or **provide dual-use resources** ('batch'+HPC) the ATLAS versatility gained by the ongoing R&D provides this option.
    - **We use this approach at our HLT farm during LS1.**

Highlight: Using the HLT Farm during LS1

- The computing experts from TDAQ, IT/SDC and BNL/T1 have set up the ATLAS HLT Farm to be used as a Grid 'site' during the LS1 as an **'opportunistic resource'**:
  - ~ 15,000 jobs = a big Tier-2!
  - A requirement from the C-RSG (RRB).
  - over the period of July-August: **140kHS06 28.3 billion seconds of wall clock time, job efficiency of 94.7%, completing 815k jobs.**
  - **A big success! Many thanks to all involved parties!**
  - **The idea of using Cloud middleware (OpenStack) as the overlay infrastructure:**
    - CERN IT (Agile), CMS (HLT Farm) and BNL all on OpenStack:
      - Similar use cases:
        - support if needed,
        - sharing experiences,
        - BNL has already part of its resources 'cloudified' and ATLAS is successfully using them!
    - An excellent use-case to gain experience with Cloud technologies!





Multi-Core Jobs and Queue Configurations

- **The increased processing times and memory footprints of the ATLAS software due to higher energy and pile-up in Run 2 processing suggest the use of parallel/concurrent processing sharing (a fraction of) the required memory.**
  - AthenaMP provides a solution to efficiently run multi-processing jobs, using less memory, which will be exploited further by the future concurrent framework.
  - For reconstruction at high pile-up, the RSS memory/core will, according to the current projections, with AthenaMP still be **around 3 GB/core.**
    - Other workflows will use much less RSS memory.
- **The current commissioning scenario of multi-core jobs assumes each site to allocate a pool of resources to exclusively run multi-core jobs.**
  - The optimal number of cores to be used **is at present evaluated to be eight** but further studies will be performed to understand the optimal scenario.
- **This will be extended by a dynamic setup, where the batch system will allocate multi-core resources on demand to be able to balance between single and multi-core jobs automatically. Most of the batch systems used by WLCG sites already support dynamic resource allocations and some sites already use it in production in 2013.**
  - The full deployment of multi-core setup on all sites will follow the schedule of AthenaMP deployment with the commissioning of the ATLAS Run-2 software environment.

What are the longer term challenges?

- **Limited cash for resources**
- **Datasets are growing**
  - **Incremental growth with time**
  - **Higher trigger rates**
  - **High pile up/larger events**
- **More complex events**
- **Hardware and architectures evolving**
  - **More cores per processor**
  - **Wider registers**
  - **Limited memory bandwidth & cache sizes**
  - **New architectures (ARM chips, co-processors)**
  - **More bandwidth**

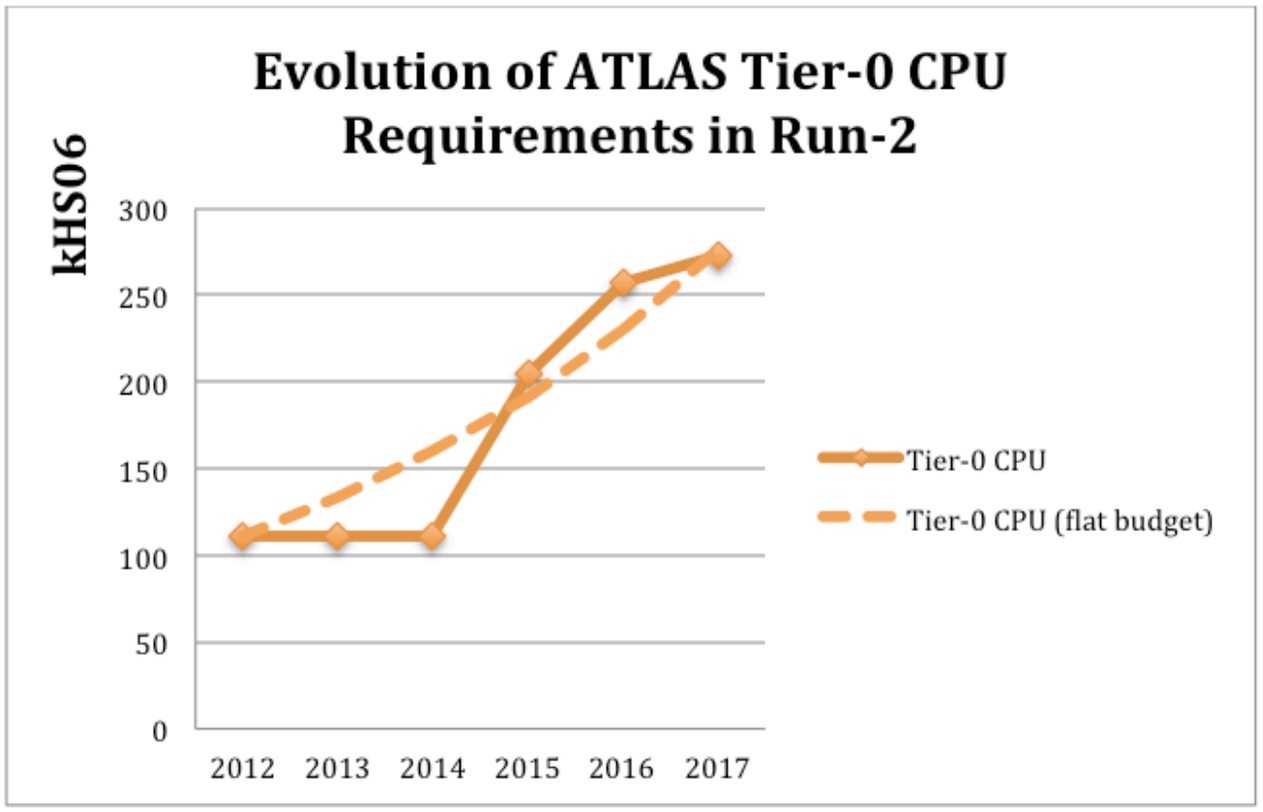
Summary

- The Computing model update is optimizing the ATLAS resource use w.r.t. the limiting resource constraints expected in the future.
- The update is fine-tuning the existing Run 1 model, profiting from the operational experience and software development (both in ADC and offline software).
- The resource requirement trends are aiming to demonstrate that we can efficiently perform our Physics program in Run 2 within these resource constraints, as mandated by the LHCC (and Computing RRB)..
- Loosened hierarchy - T2s do more
- The rise of the T3, federated data.

Backup

Resource Evolution cont'd

CERN CPU (kHS06)	***2015***	***2016***	***2016***
CERN CPU Total	***205 [240]***	***257***	***273***
Tier-0 subtotal	***156***	***199***	***199***
T0: Full reconstruction	133	175	175
T0: Partial processing and validation	12	12	12
T0:Merging and monitoring	4	5	5
T0: Automatic calibration	5	5	5
T0: Servers	2	2	2
CAF subtotal	***49***	***58***	***73***
CAF: Partial reconstruction, debugging and monitoring	13	18	18
CAF: Non-automatic calibrations	4	4	4
CAF: Group activities	15	19	27
CAF: User activities	5	6	13
CAF: Servers	12	12	12



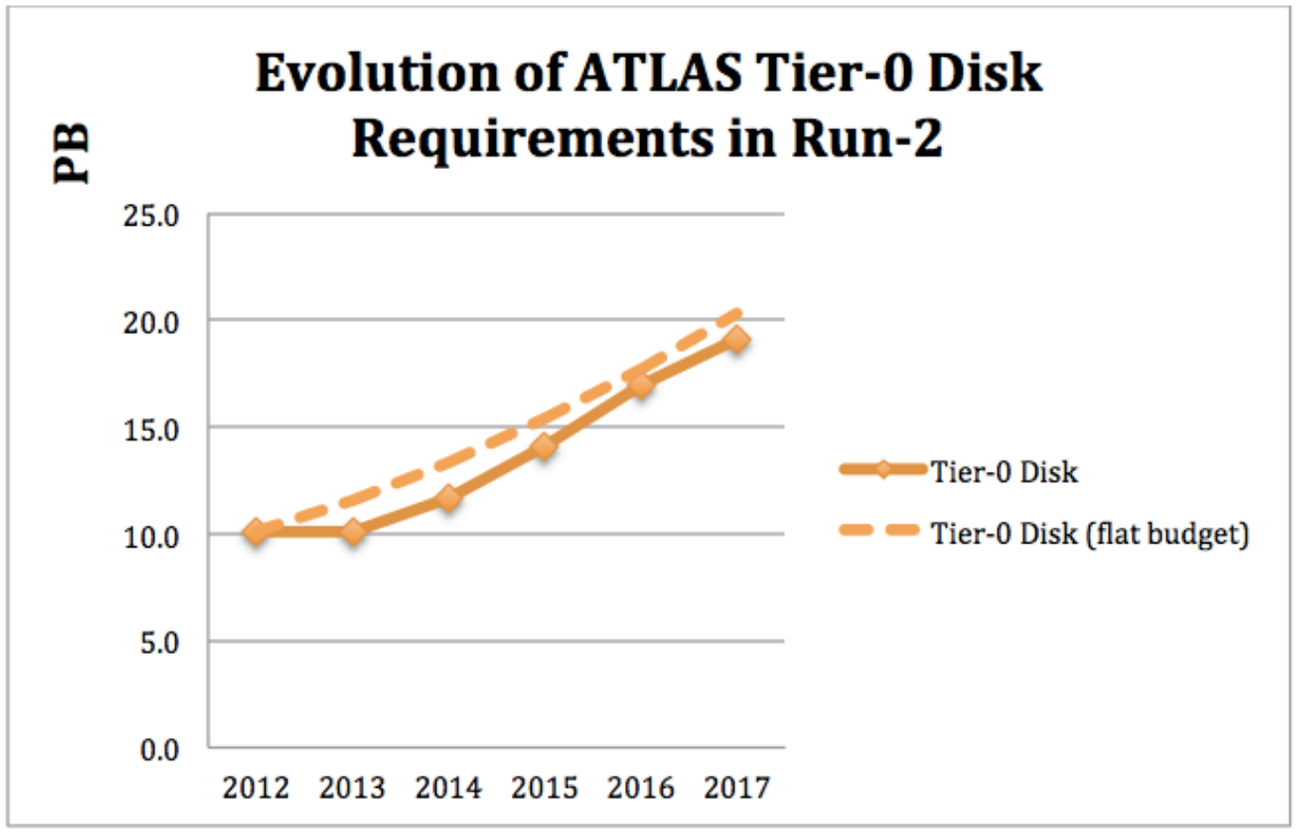
Flat CPU budget: factor 1.2/year

Resource Evolution cont'd

CERN Disk (PB)	*2015*	*2016*	*2017*
CERN Disk Total	***14.1 [15.3]***	***17.0***	***19.1***
Tier-0 Disk Subtotal	***3.40***	***3.40***	***3.40***
Buffer for RAW and processed data	3.00	3.00	3.00
Buffers for merging	0.30	0.30	0.30
Tape buffer	0.10	0.10	0.10
CAF Total	***10.7***	***13.6***	***15.7***
CAF: Calibration and alignment	0.5	0.5	0.5
CAF: Derived detector data	2.0	2.8	3.9
CAF: Derived simulated data	6.7	8.8	8.8
CAF: Group data	1.0	1.0	2.0
CAF: User data	0.5	0.5	0.5

CERN Tape (PB) Cumulative	*2015*	*2016*	*2017*
Total	***30 [35]***	***41***	***55***

Flat disk budget: factor 1.15/year



The implications

- **Many of the changes will be prior to the end-user analysis**
  - **Group production sees the biggest changes**
- **But the changes will feed-down**
  - **Limited event information**
  - **Limited formats**
  - **Fewer replicas**
- **Planning so far is up to LS2**
  - **But most of the new ideas should remain thereafter**

Software structure: kernels and job options

- ***Kernel*: Athena algorithm which steers the derivation**
  - **Takes a list of tools from the user for doing skimming, thinning and augmenting (slimming done separately because it is the same for every event)**
  - **Individual setting up a derivation never needs to modify the kernel - it is simply a shell for scheduling the tools in Athena (and becomes an AcceptAlg)**
  - **The tools can do whatever the user wishes, provided they inherit from the interface class**
    - **In particular at the moment they can be written either for D3PD or AOD types**
- **Each derivation is encoded by a set of job options, which sets up the kernel and the required tools and settings. These are executed by Athena**

Examples of possibilities for derivations

- **Slimming only**
  - **for analyses which do not want to remove events (e.g. tau, jet/ETMiss, top) just remove branches/containers that are not used**
  - **augment with variables that can lead to further reduction (example: number of tracks around a tau candidate)**
  - **Questions: interested in *thinning*? Trigger skimming?**
- **Skimming+slimming**
  - **Do event selection on basis of trigger decision and/or offline objects**
  - **Skimming via decay candidates (e.g. W, Z, J/ψ etc)**
    - **Augment data with candidate information**
  - **Most analyses fall into this category (?)**

Division of labour between the derivation framework and user analysis

- **The derivation framework is primarily for reducing PB of data down to ~TB (that might mean anything from 1TB to upwards of 10TB) on the production system**

- **It is necessary to balance the following:**

  - **The size of each derivation should be small enough to process within a day or less with normal user grid jobs, and/or to fit on a local Tier-3 if you're lucky enough to have access to one**

  - **On the other hand you should keep enough information to avoid having to re-do production-system level derivations too often**

  - **This might mean**

    - **keeping enough information to re-do CP group requirements / calibrations if they change late, rather than trying to do them in the derivation step**

    - **making your cuts loose enough that calibration changes can be absorbed without re-skimming**

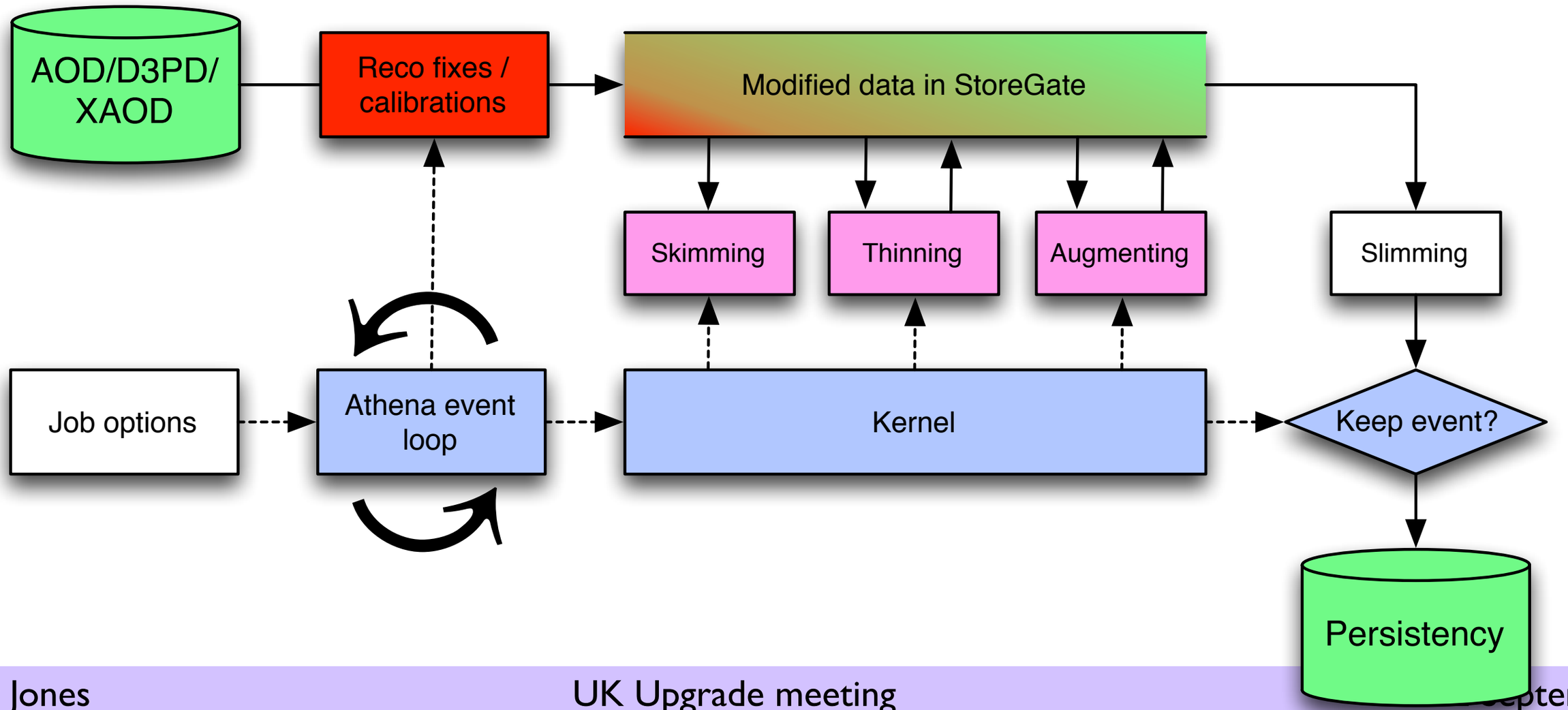
Division of labour between the derivation framework and user analysis: recommendation

- **The suggestion is to avoid applying**
  - **Calibrations**
  - **CP group recommendations/smearings/scalings**

  in the derivation framework, but to do these tasks at user analysis level using the Analysis Framework
- **This is the model originally envisaged by the AMSG**
- **However, this may not be appropriate to all derivations (especially those based on jets/MET), so calibrations can be scheduled before the derivation framework starts up**
  - **In a train, it would be necessary to create a new StoreGate instance specific to the carriage needing the calibrated collection, to avoid "contaminating" other carriages (unless they also wanted the same calibration)**
- **Application of systematic shifts must continue to be done in user analysis**
- **Major fixes done in an AOD to AOD processing**

Augmentation

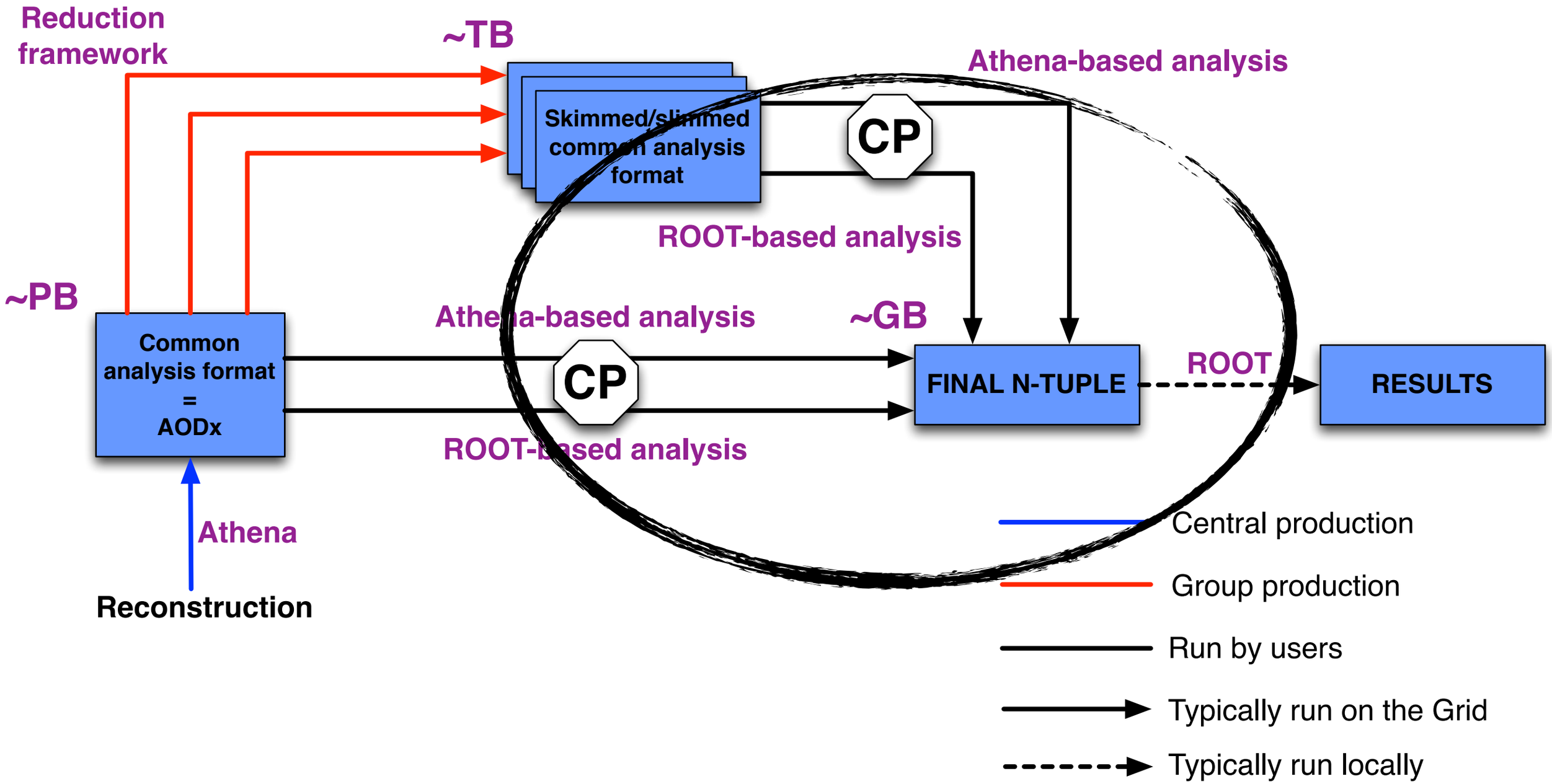
- **Three kinds of augmentation in derivation framework are envisaged**
  - **New quantities calculated from the existing data and inserted into the derived format ("user data") - *done via the kernel and tools***
  - **New object collections of an already-defined kind (e.g. new jet collections etc) - *scheduled in Athena before the derivation framework starts***
  - **Application of calibrations to the data - *scheduled in Athena before the derivation***



Other open points (i)

- **Truth slimming**
  - **For MC a major part of derivations will involve slimming/ thinning the truth information**
  - **This is a quite complex operation and so there is likely to be a range of pre-sets for different physics studies - to be discussed with groups/MC truth task force**
- **Pre-set slimming**
  - **Idea is to avoid people have to come up with detailed lists of containers/branches to be kept, but rather make use of pre-sets suited to what they need to do**
    - **Verbosity levels?**
    - **Lists for a given tool/operation - maybe autogenerated by the tools**

What are the new analysis tools?



The new analysis tools

- **Note I am reluctant to say 'framework' (bad history!)**
- **First steps**

UK: Get involved!

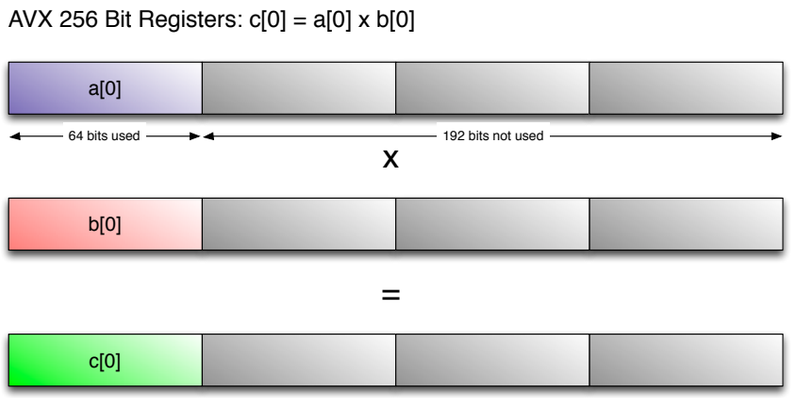
  - **Convert existing CP tools for the new XAOD**
  - **Create full-blown examples in ATHENA and ROOT**
- **Analysis releases**
  - **The full releases are huge and mainly unused in analysis**
  - **Create a small O(100MB) release for use on desktop/laptop**
    - **All CPU-intensive steps to be in the derivation framework**
    - **Must be supported on all POSIX platforms, perhaps Windows**
    - **Main tools as binary, everything else can be recompiled**
    - **Must be forward-compatible so analysis is reproducible into**

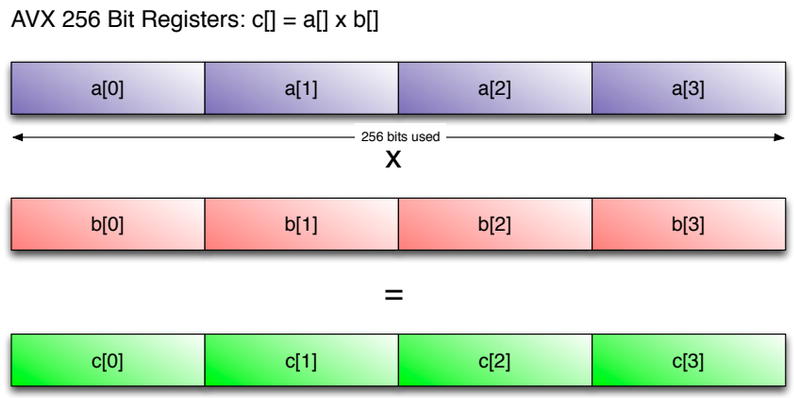
New Architectures and Hardware: Many-cores

- **The first challenge with many- and multi-core processors is the memory bloat**
  - **We crudely load one event per core, and reinstantiate everything in memory**
  - **Already have 64 bit registers already pose a challenge for a single core**
- **First step: AthenaMP (UK upgrade worked on this)**
  - **Mother process loads most things needed in common**
  - **Forked event workers only read memory for new items or when a page changes (Linux Copy-on-Write)**
  - **Total memory goes like 3GB+1.5GB/worker - about half**
- **Good - but not good enough**

New Architectures and Hardware

- **Threading is hard**
- **Use libraries**
  - **Use libraries e.g. Intel has Threading Building Blocks**
  - **Use frameworks; thread at the event and algorithm level**
    - **Gaudi-Hive is being investigated (UK upgrade will contribute)**
- **Vectorize (UK upgrade work here)**
  - **Complements threading; libraries like Eigen, VDT, MKL**
  - **Do more each clock cycle - Single Instruction Multiple Data**





The Local Environment

- **Your datasets will grow**
- **But so will bandwidth**
  - **You will be pulling data across the the network (already starting with XROOT)**
- **You will use more fast simulation modes**
  - **Ed boards will have to lump it!**
- **Your ntuples will need to read faster**
  - **Faster storage media? Flash, newer solid state storage???**
- **Will you be using Clouds**
  - **Commercially: Well, maybe - but only if they redesign their offering and drop prices considerably**
  - **Privately: Almost certainly, the private cloud is a fuller implementation of the original Grid vision, and probably easier to manage**