Computing and Software

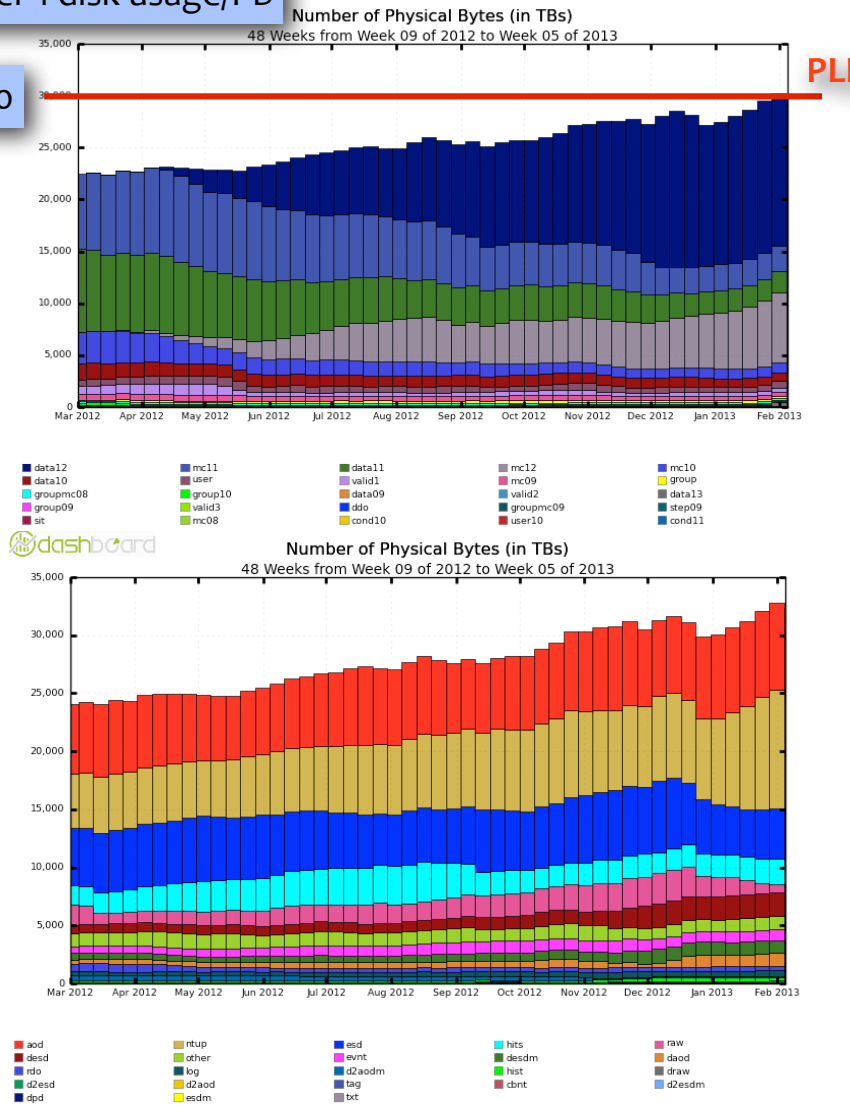
Borut Kersevan
Hans von der Schmitt

ATLAS Week,CERN, February 2013

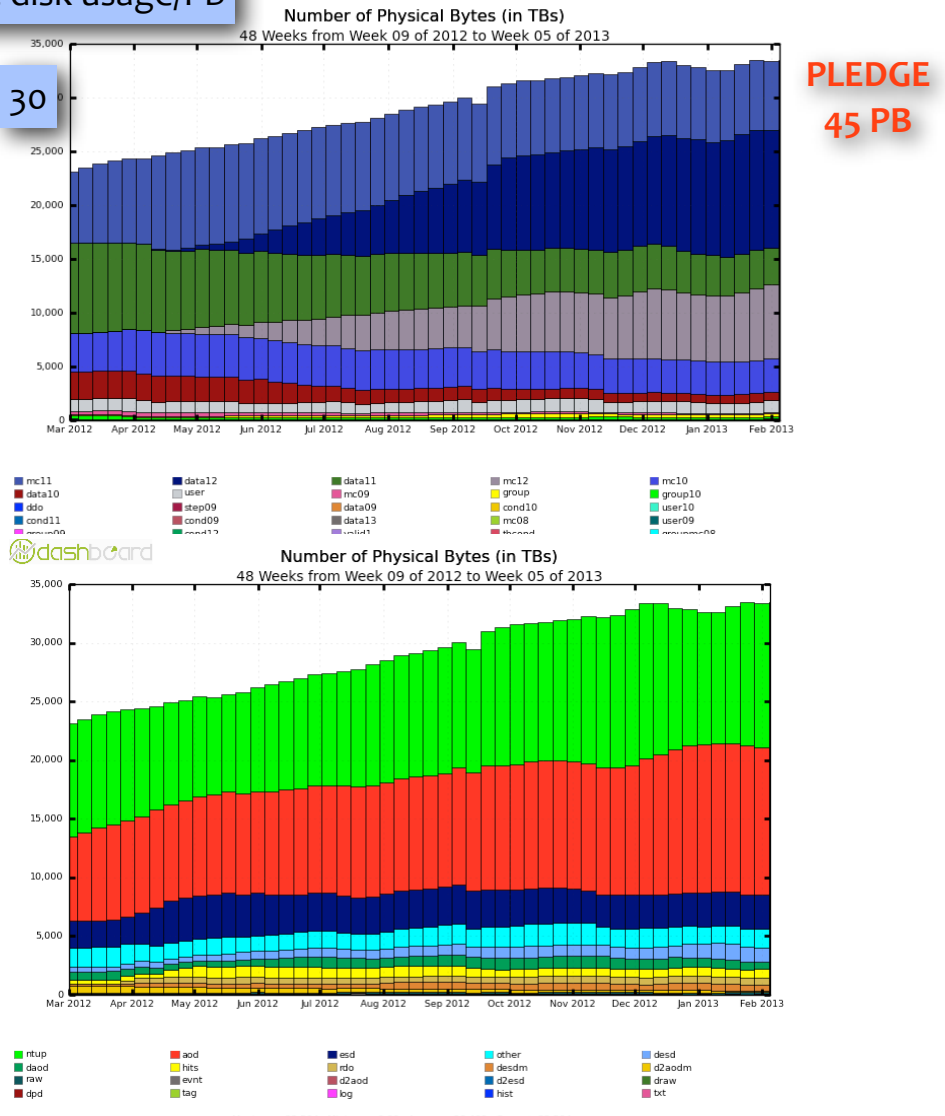
ATLAS Disk Usage in 2012

- A constant 'struggle' with full T1 disk - 'lifetime' of 2011 analyses under-estimated in our resource models.
- Complex (= work intensive) handling of workflows involving transient datasets (merging).

Tier-1 disk usage/PB

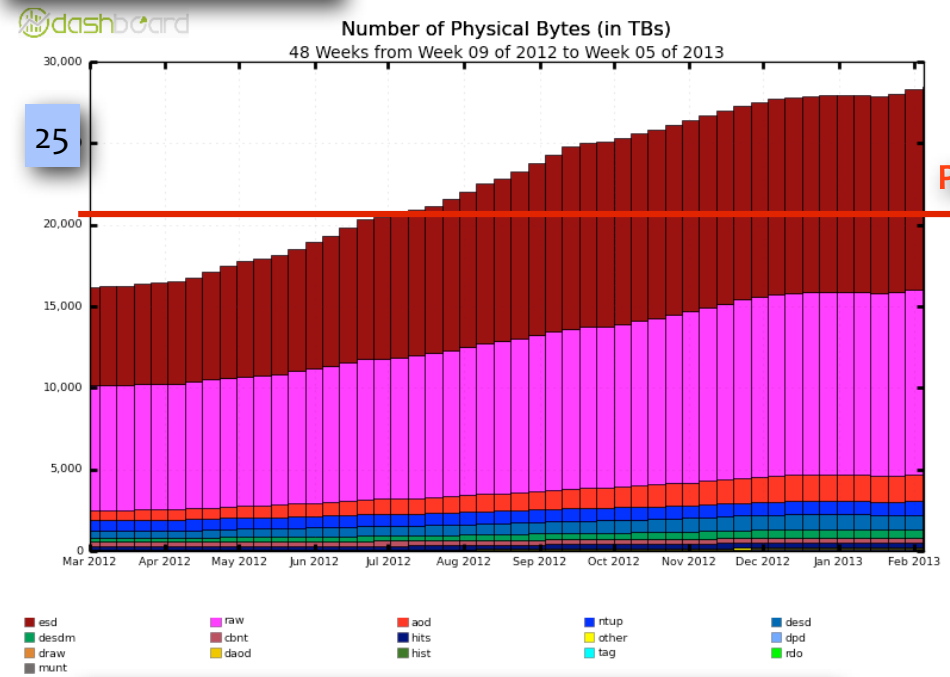


Tier-2 disk usage/PB



ATLAS Tape Usage in 2012

Tier-0 tape usage/PB

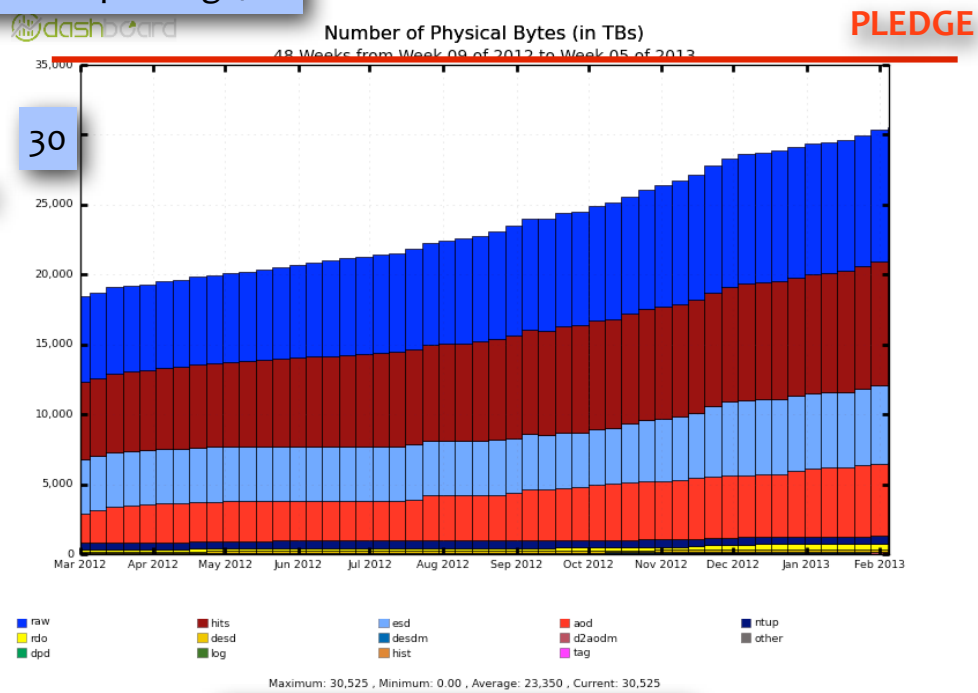


25

PLEDGE

We will clean the pp ESDs from CERN tape.
Procedure established, waiting for a good time.

Tier-1 tape usage/PB

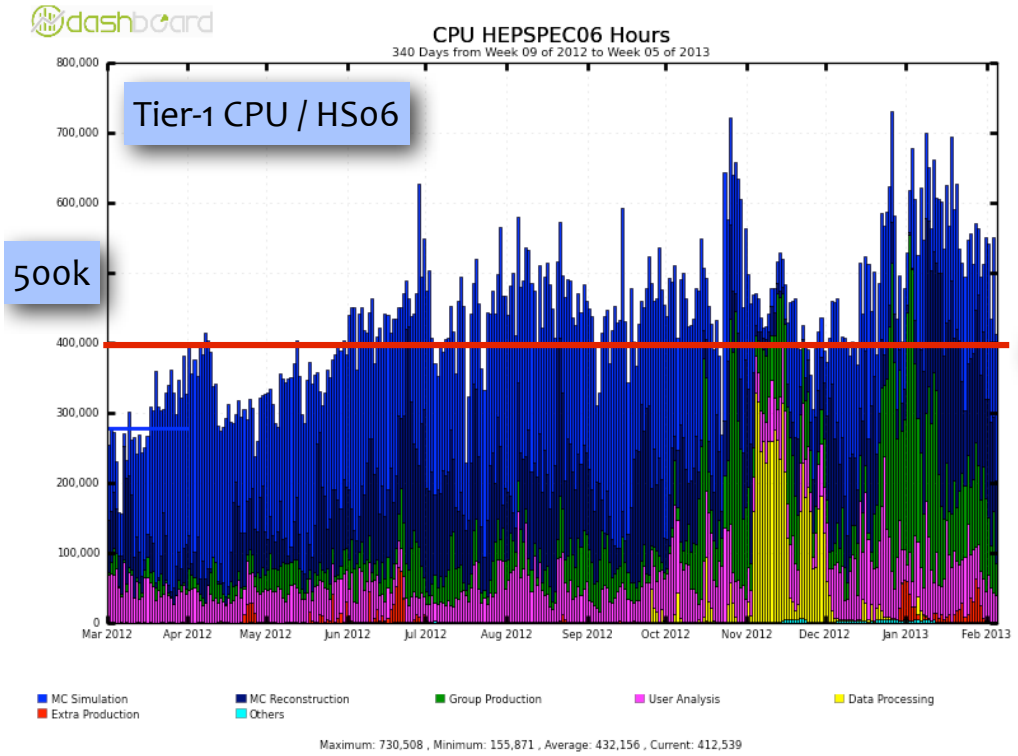


30

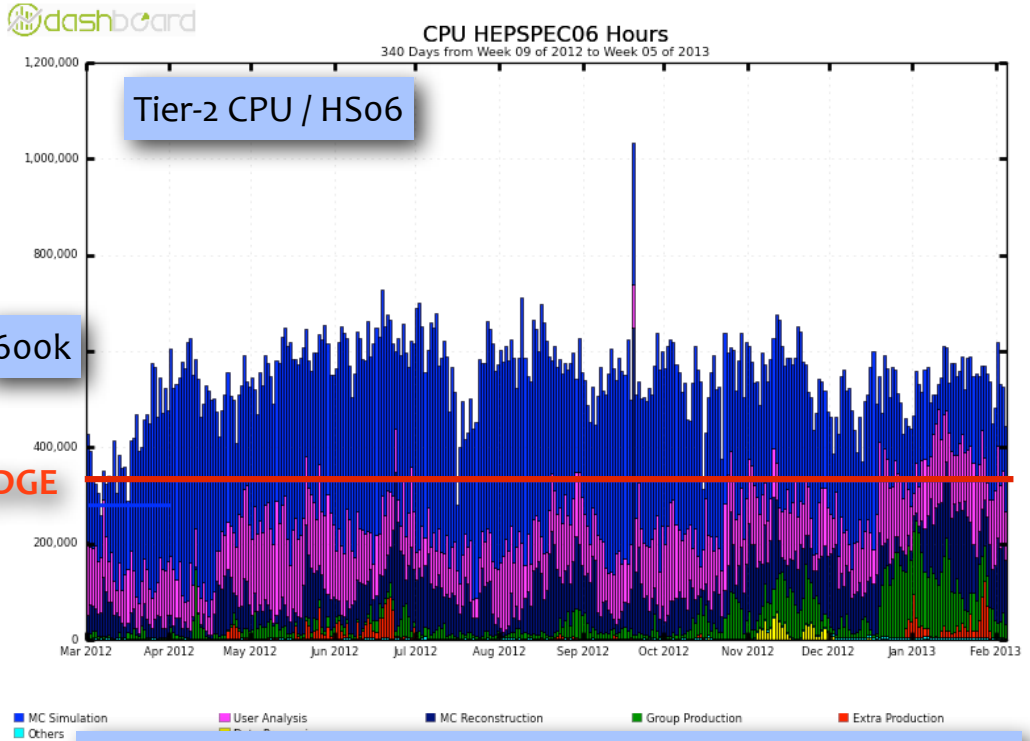
PLEDGE

Overall situation satisfactory.
Some T1 short on tape.

ATLAS CPU Utilization in 2012



- **Tier-1 and 2:** We have much more **CPU** as well as (but not that much) **disk** - ***thanks to sites for resources and excellent operating!***
  - **~1.5 times the pledge in T1.**
  - **~2 times the pledge in T2.**



ATLAS RESOURCE REQUEST TO RRB - WITH 'FLAT' BUDGET CONSTRAINTS

CPU [kHS06]	2011	2012	2013	2014	2015
CERN	74	111	111	111	228
Tier-1	202	295	319	373	502
Tier-2	275	319	355	408	540
Disk [PB]					
CERN	7	11	11	11	14
Tier-1	22	29	35	36	51
Tier-2	35	48	53	56	69
Tape [PB]					
CERN	14	21	27	31	45
Tier-1	16	35	43	53	78

up to 10% uncertainty due to WLCG parameters

Computing Resource Usage in 2012, 2013-15

- **Running well in 2012, but limitations in MC and Group production** *(=> Kevin yesterday)*
- **Extra resources for 2012 p-p run extension** were deployed, as much as available: **Many thanks to all our Grid sites!**
  - **We depend crucially on the extra resources from sites, it would be good if we can find a way to acknowledge them w.r.t. funding agencies!**
- **Brief outline of our current resource planning guidelines for 2013-2015:**
  - **In 2013:**
    - **There might be partial reprocessings of 2010-2012 data and MC for further studies.**
      - **More in Kevin's and Phil's talk.**
    - (More) new MC for analysis will be produced.
    - Very active group/user analysis.
  - **In 2014:**
    - Largish MC samples for high energy running will be produced and related physics group/user analysis.
    - The final full reprocessing of 2010-2012 data and MC, foreseen to use the evolved event formatting/data model/data distribution prepared for 2015 high-energy data taking (**data preservation!**).
    - 'Full dress rehearsal' activities, preparing for Run 2.
  - **In 2015:**
    - Processing and reprocessing of new high energy data.
    - Related production of MC samples matching the data.
    - Increased group/user activity.

Resource Projections until 2015

- **Our goal is successful processing of data taken @ 1kHz trigger rate**
- Event sizes & CPU set to be equal to 2012 values or taken from ATLAS upgrade MC samples @ 13 TeV.
  - **Not a trivial assumption, as I will show..**
- For 2015 21 weeks and 30% LHC efficiency assumed.
- relatively modest amount of simulation w.r.t. data taken assumed.

ATLAS RESOURCE PROJECTIONS WITH FIRST ESTIMATES FOR 2015

CPU [kHS06]	2011	2012	2013	2014	2015
CERN	74	111	111	111	228
Tier-1	202	295	319	373	502
Tier-2	275	319	355	408	540
Disk [PB]					
CERN	7	11	11	11	14
Tier-1	22	29	35	36	51
Tier-2	35	48	53	56	69
Tape [PB]					
CERN	14	21	27	31	45
Tier-1	16	35	43	53	78

CRSG RECOMMENDATIONS for 2013 (final) and 2014 (preliminary)

ATLAS

Resource	Site(s)	2013 ATLAS	2013 CRSG	2014 ATLAS	2014 CRSG
CPU/kHS06	T0+CAF	*111*	**111**	*111*	**111**
	T1	*319*	**319**	*373*	**355**
	T2	*355*	**350**	*408*	**350**
Disk/PB	T0+CAF	*11*	**11**	*11*	**11**
	T1	*35*	**33**	*36*	**33**
	T2	*53*	**49**	*56*	**49**
Tape/PB	T0+CAF	*27*	**23**	*31*	**23**
	T1	*43*	**40**	*53*	**44**

The CRSG/RRB is 'flattening' our requests

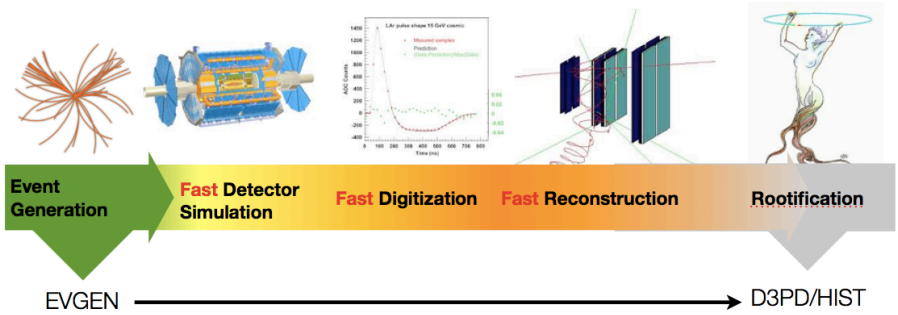
Resources in 'Critical Periods'

- All our resource planning is done using **average** CPU (and disk) consumption rates.
  - And we are **using all available resources all the time** for diverse ATLAS activities.
- The ATLAS analysis activities **peak** before big conference periods, leading to congestions and backlogs in all the Grid demands.
  - **This is very hard to present to the RRB, asking for resources to cover the peaks.**
  - We 'routinely' ask sites in such periods to provide even more CPU, which is has diminishing returns with every new crisis over the year (sites deploy what they can when first asked).
- Several venues to explore during LS1:
  - **Optimizing/changing our workflows, both in analysis and on the grid.**
    - It will necessarily involve also a change in the ways people analyze the data!
  - **Finding opportunistic resources:**
    - High Performance Computing centres have a lot of CPU available, we could use the available idle cycles for (a subset of) our activities,
      - e.g. MC event generation, possibly simulation.
    - Cloud resources: Again, for a subset of our activities, similar to HPC
      - If we are really hard pressed, even use commercial resources (?)
        - BNL T1 added 5k job slots via Amazon EC2 for the current crunch!
    - This involves quite some work from the S&C community, both venues are being explored!
      - Interesting contributions expected in Software and Computing week in March.
  - **Looking for faster solutions to speed up simulation and accommodate our MC needs.**
- **A lot of activity foreseen in Software & Computing during LS1 to tackle this.**

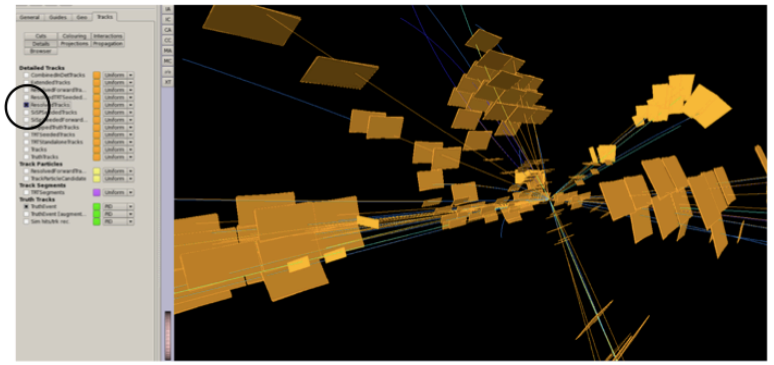
Example: ISF

- Do people really need to use the (very CPU consuming) G4 full simulation chain for all cases?
  - It is the 'safe' solution, **but it does not scale** with respect to the resources we will have available for Run 2..
- Conceptualizing and developing 'Fast simulation/digitization/reconstruction chain'..

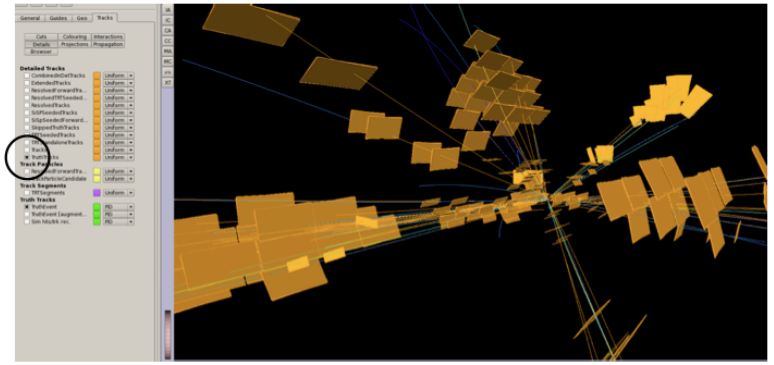
An all-in-one chain for fast MC



reconstructed tracks



truth tracks



- **More in the Simulation talk**

Example: Group Analysis Workflow

- The AOD seems not to be the 'Analysis Object Data' for a majority of analyses.
- The production of the 'Group' data formats (D3PD/NTUP) done via central production.
  - A very nice evolution of the procedures, giving us new insights..
  - Not all of the insights what we would want:

	CPU s/event	Event size (kB/ event)	Nevents data
ZPRIMEEE	0.5	8	200000000
ZPRIMEMM	1.2	105	200000000
WPRIMEE	0.5	12	200000000
WPRIMEM	0.5	6	200000000
SMQCDSLIM	0.002	0.002	xxx
EXMJ	0.002	0.002	xxx
DAOD_PHO (ZEEG, ZMMG)	0.3	300	70000000
NTUP_PHOTON (includes EGAMMA)	4.5	80	770000000
NTUP_SMQCD	0.6	80	2500000000
NTUP_TRIGBJET	0.007, 1.52 (MC)	15, 213 (MC)	870000000
NTUP_SUSYSKIM	2.6	50	2328000000
NTUP_SUSYBOOST	5	66	870000000
NTUP_SUSY	2.6	50 (data),105 (MC)	xxx
HSG2	0.28	1120 (6 outputs)	1500000000
JETMET	6.9	200	2500000000
NTUP_TOP	6.5	225	1000000000
NTUP_SMWZ	5.7	212	2328000000

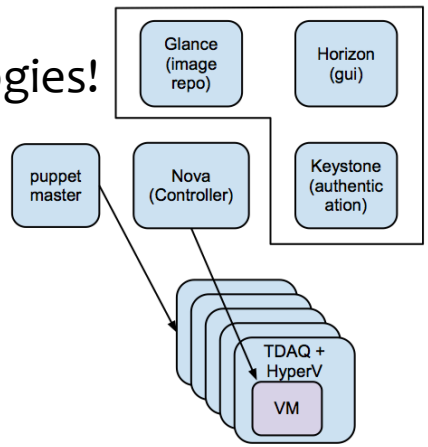
Compare to 20 sec/event and 200 kB/event for **full event reconstruction from RAW to AOD**

Current situation slows down analyses, creates grid problems filling the disks **and does not scale to 2015**

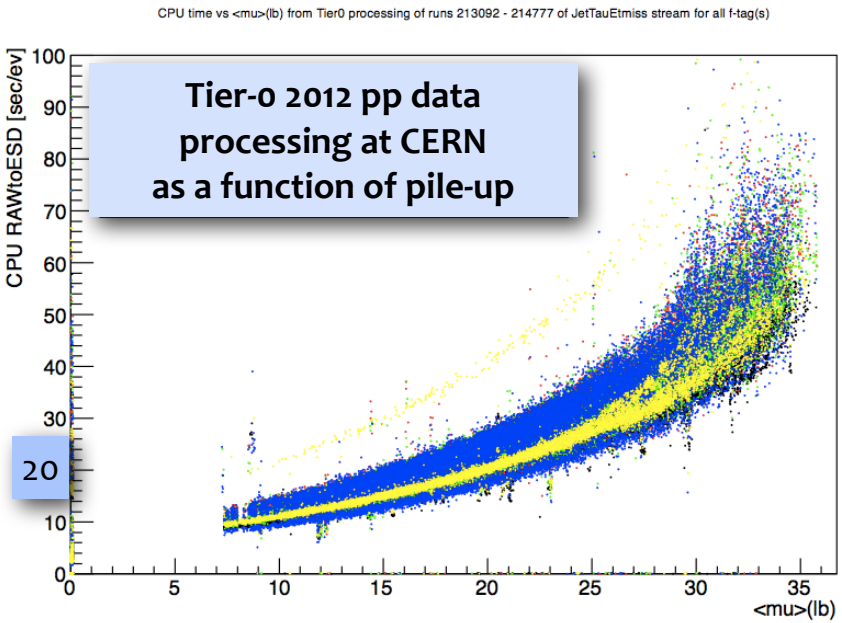
- We need to change the Analysis/Computing Model and workflows to improve on the analysis throughput. More in the AMSG talk...

Using the HLT Farm during LS1

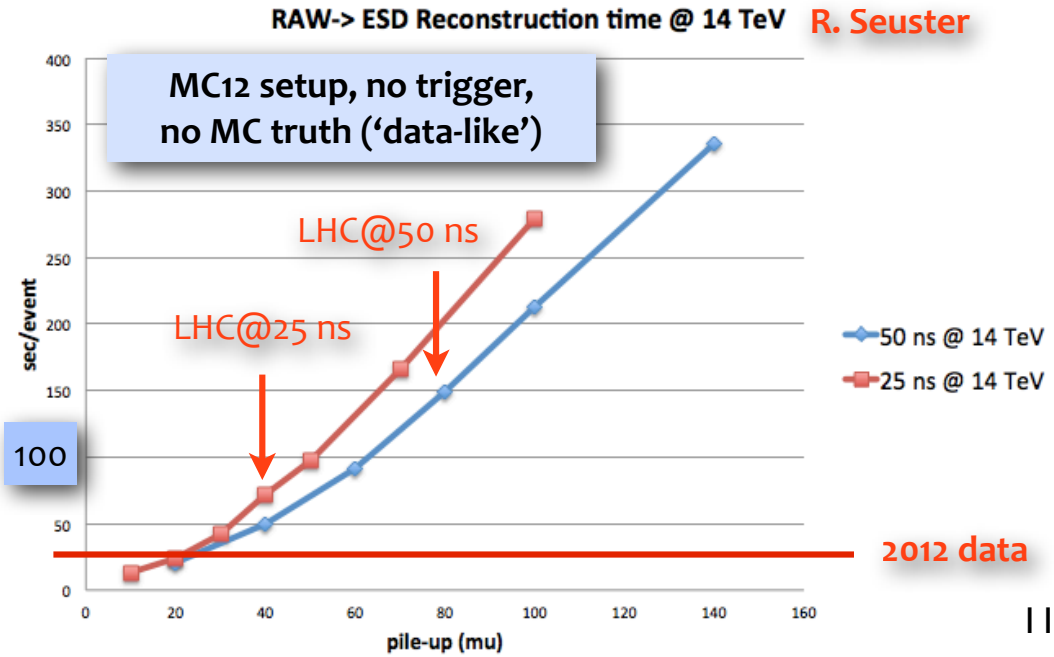
- The computing experts from TDAQ, IT/ES and BNL/T1 are setting up the ATLAS HLT Farm to be used as a Grid 'site' during the LS1 as an **'opportunistic resource'**:
  - ~ 14,000 cores (with an Grid performance factor to be estimated) = a big Tier-2!
  - A requirement from the C-RSG (RRB).
  - **The idea 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!

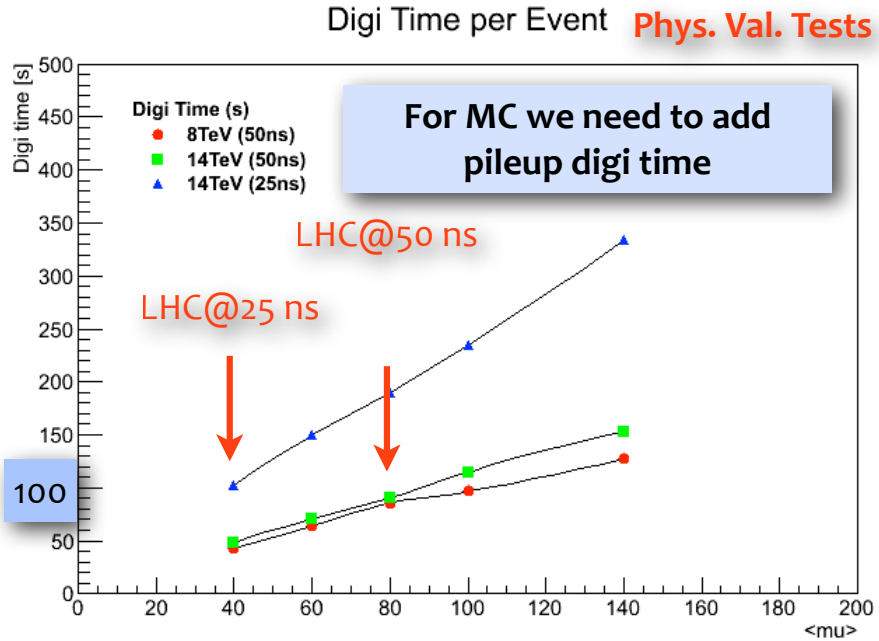


First CPU Projections towards Run 2



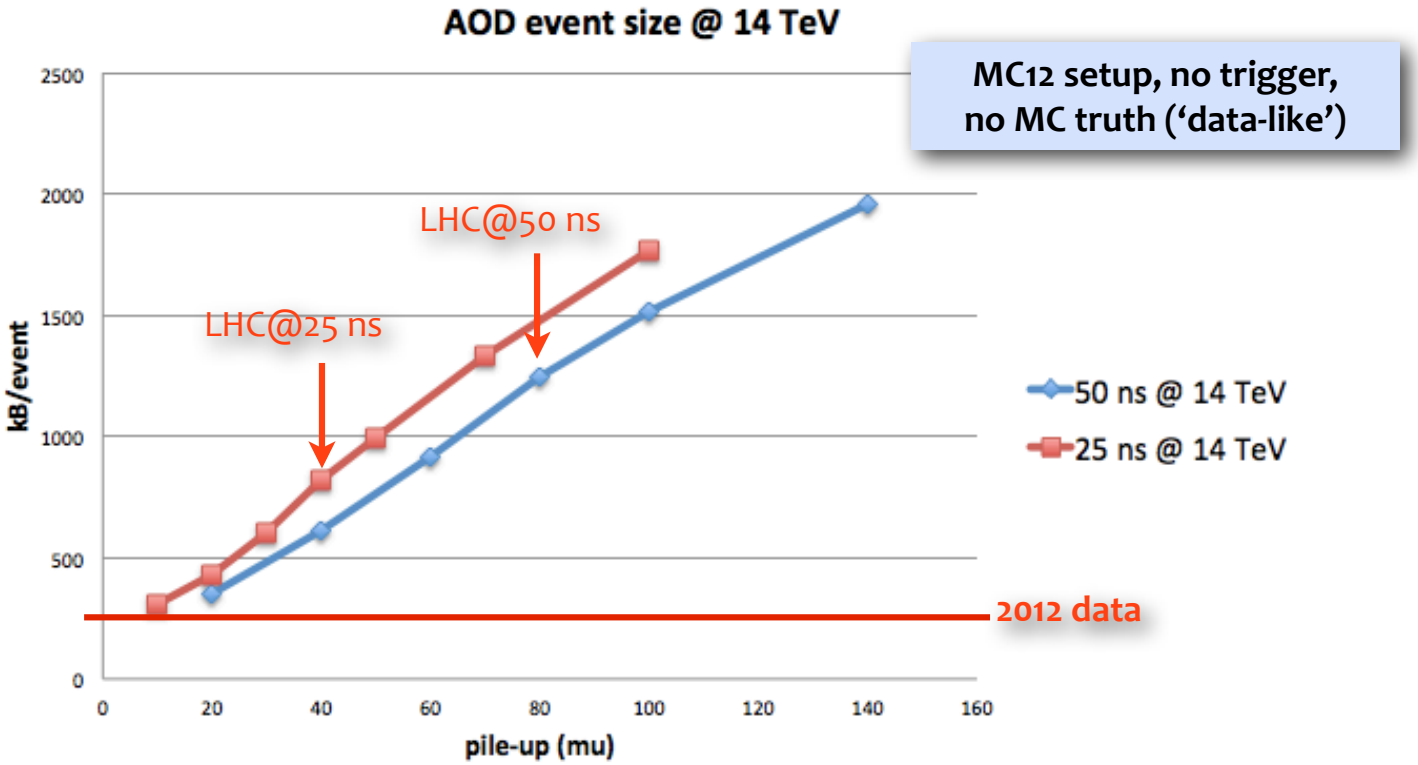
- Focusing on reconstruction: In order to match the increase in energy and pileup, **we need at least a factor 3 speedup to get to what we have now!**
  - **Code optimization a major focus of LS1:**
    - (auto) vectorization, utilizing modern CPUs better.
    - Improving algorithms (tracking, digitization....)
    - Optimizing components (BField, CLHEP)..





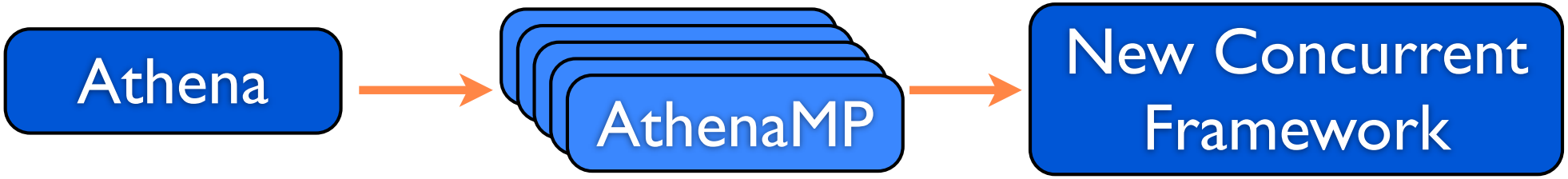
Another Concern for Run 2: event sizes

- The event sizes will also increase considerably - a major worry:
  - No big technological profits currently envisaged in compressing the AOD content further.
  - Thus two options:
    - Change the AOD content - can impact our physics results.
    - Modify the Computing Model (less copies on the Grid, better access). Strongly coupled to ATLAS analysis workflow.
- **Another major activity during LS1!**



Another Concern: Memory use

- ATLAS Grid resources by WLCG/ICB agreements limited in RAM to 2GB/core.
  - ATLAS Reconstruction SW struggles to keep this limit.
    - We still cannot run 64 bit reconstruction code except on special nodes with more memory due to this.
    - Situation certainly only deteriorates with higher energy/pileup..
- New technologies favor multi- and many-core CPUs packed into a node (soon 100 CPU cores/node); consequently, keeping memory/core requirement might be difficult.
  - Also, using opportunistic resources such as High Performance Computers (Super-..) assumes low memory footprint.
- **The LS1 plan is to:**
  - Commission **AthenaMP** (reduces memory footprint by the 'Copy-on-Write' feature)...
  - ...and start working on a **new Concurrent Framework** (Full threading, event level and algorithm level parallelism..).
    - **benefitting from collaboration with IT/OpenLab, PH-SFT and CMS in these and related fields**
  - These developments also require changes in our Distributed Computing - from queue configurations to file naming...
  - Details to be worked out by the Software and Computing Week in March.



A Note on Software Release Planning

- There will be a lot of development in all areas: simulation, digitization, reconstruction, analysis workflow…
  - We need to be **very careful** with the planning of our releases, to always have good reference 'points' for validation of all changes/developments.
    - See Phil's talk on the Simulation group challenges in this respect.
  - **A detailed plan is being evaluated and prepared.**
- We are looking into ways to speed up and simplify our software building procedures to alleviate the expected load.

Release 17.2

The current release used for 2012 data taking, and fast reprocessing of 2012 data with improved alignment, data quality and other improvements. See here for more details about the 2012 October fast reprocessing.

Release 17.4

A hypothetical release based on 17.2. with a limited and well defined set of frozen Tier0 incompatible improvements for a possible reprocessing of 2011 and 2010 data. Unclear if this will ever be needed.

Release 17.7

Switch to CLHEP 2.1.2.3. Merge in 64-bit Identifiers. This release will be used for testing current ATLAS and Upgrade Simulation using ISF. HITS simulated in this release will need to be compatible with digi-reco run in 17.3.X.Y.

Release 18

With the 2012 October fast reprocessing, there's no compelling argument for a full reprocessing of Run 1 data with a new release. Improvements w.r.t 17.2. were minor and in the near future no huge improvments are expected that could make the reprocessed data interesting for redoing analyses and publishing them.

Instead, we can think of having some software milestones archived with release 18.

- first mayor release where we use new gcc 46 or even 47, allowing us to validate of new compilers and compare physics performance with same release build with gcc43
- last release without using c++11, to ease comparison with old compiler gcc43

C++11 then could be enabled once release 18 is build and has passed some basic validation ensuring that nothing mayor broke.

Release 19

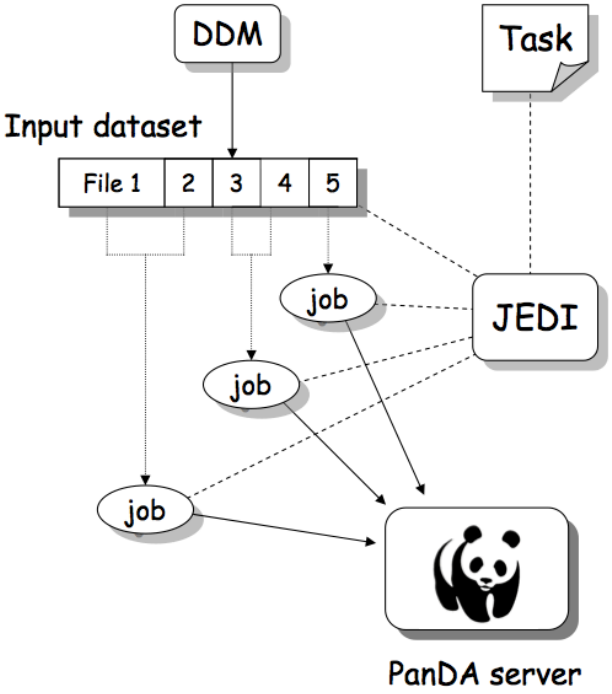
Intermediate mayor development release build during LS1, used for a preliminary evaluation of physics and software performance.

Release 20

The release used for the 2015 data taking. Previously simulated events of MC15 should be reconstructed with the same release. Further speed improvements can be implemented during data taking, but should respect frozen Tier0 policy, otherwise will have to implemented for a release 21.

Plans for LS1 in Distributed Computing

- **New Distributed Data Management system (Rucio) being implemented:**
  - Development progressing well!
  - A hands-on tutorial in last Software and Computing week.
- **WAN data access and data caching (file level, event level):**
  - Now running a 'challenge' to evaluate the XRootD federation functionality for ATLAS.
- **New MC production system (JEDI+DEFT+..):**
  - In progress: series of ongoing meetings, identifying the lessons learned in Run 1 and technical planning.
  - Technical Design Report and "final" list of requirements by May TIM in Tokyo (May 2013).
  - The first 'new' task to run in summer 2013.
- **Applications-driven usage of networks:**
  - In evaluation stages, potentially many gains.
- **Interfacing to opportunistic resources:**
  - HPC resources, Clouds,...
- **.... and many more!**
- **A challenge also to bring these all together and commissioned for Run 2!**



Summary

- **Assume we will have a flat monetary budget in future for ATLAS computing**
  - this message comes from WLCG and from the CRSG/RRB scrutiny group
  - although this may even be optimistic for some countries
  - increase in Grid resources (disk, CPU, network) by technological progress only
    - try exploit additional resources, e.g. in high-performance computing centres.
    - also seek manpower savings in Tiers operation.
- **Consequently, growing investment in software development required during LS1 (from reconstruction to distributed computing) is needed, if the computing is not to be the limiting factor in ATLAS physics results throughput:**
  - **An ambitious plan is taking shape for the S&C LS1 activities.**
    - We **need** to do it to meet the Run 2 challenges.
  - **We lack manpower in all areas (especially expert!)**
    - need to optimise the usage of the given CPU and storage very substantially, with extra effort, and we ask the collaboration for extra manpower and institutional commitments.
    - This kind of software work involved should be very tempting to a few of you …