ATLAS Operation and Upgrade Plans

LHC, Detector, Computing

Tbilisi, October 2011 - Hans von der Schmitt

Computing

- In September we reported usage of computing resources to the LHCC
- The important perspective is:
  - the success of our physics program during the first months of 2011-12 is in no small part due to the success of ATLAS Computing
  - flexibility was demonstrated and effort was invested in optimizing many key factors that influence our physics and our use of resources
  - improvements in reconstruction time, particularly in the face of ever-increasing pileup, likewise event sizes, were achieved and are ongoing
  - the data distribution model includes on one hand guiding the collaboration away from ESDs to efficiently produced group derived data, and on the other hand dynamic data placement
  - trigger menu optimized and evolving
  - simulation improved

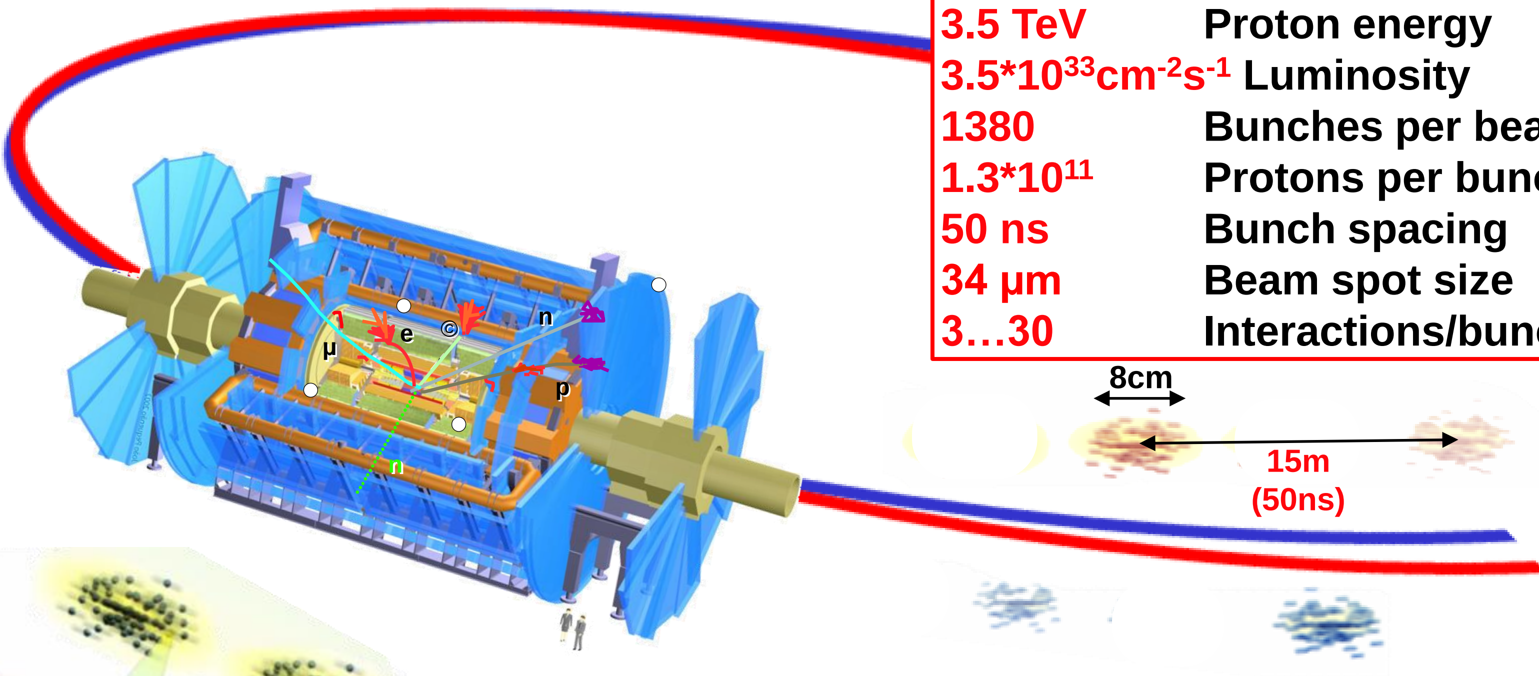
Computing resources expectation

- We are only part-way into a two-year run
  - we optimize our physics capabilities within the constraints of machine performance and computing resources
  - the instantaneous luminosity is still increasing, and already exceeds expectations. The integrated luminosity also exceeds expectations
  - we can expect LHC lifetime to improve
  - the size of our datasets are growing, and the usage patterns of analysis is still evolving, and will continue to evolve for some time
  - access to new physics studies, balancing MC and data-driven background estimates
  - flexibility of our computing system, as well as ongoing CPU and event size optimization, will continue to be important
- The 2013-14 shutdown provides *the* chance for bigger steps in software and computing

Proton-Proton collisions at LHC – parameters 2011

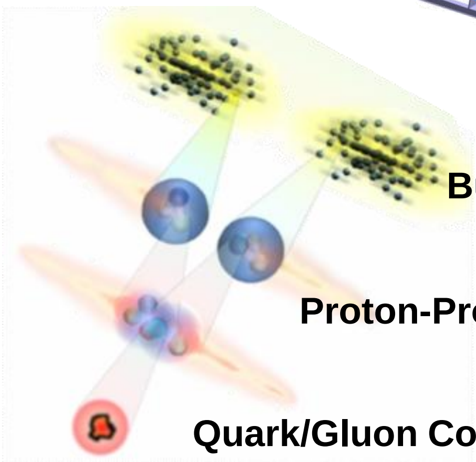
LHC parameters 2011:

3.5 TeV	Proton energy
$3.5*10^{33}cm^{-2}s^{-1}$	Luminosity
1380	Bunches per beam
$1.3*10^{11}$	Protons per bunch
50 ns	Bunch spacing
34 μm	Beam spot size
3…30	Interactions/bunch crossing



0.3 A per beam

120 MJoule per beam

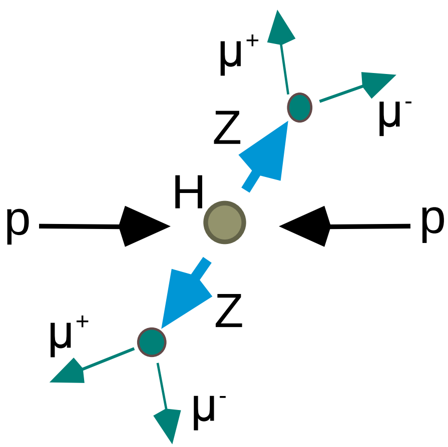


Bunch Crossings $2x10^7$ Hz

Proton-Proton Collisions $0.2*10^9$ Hz

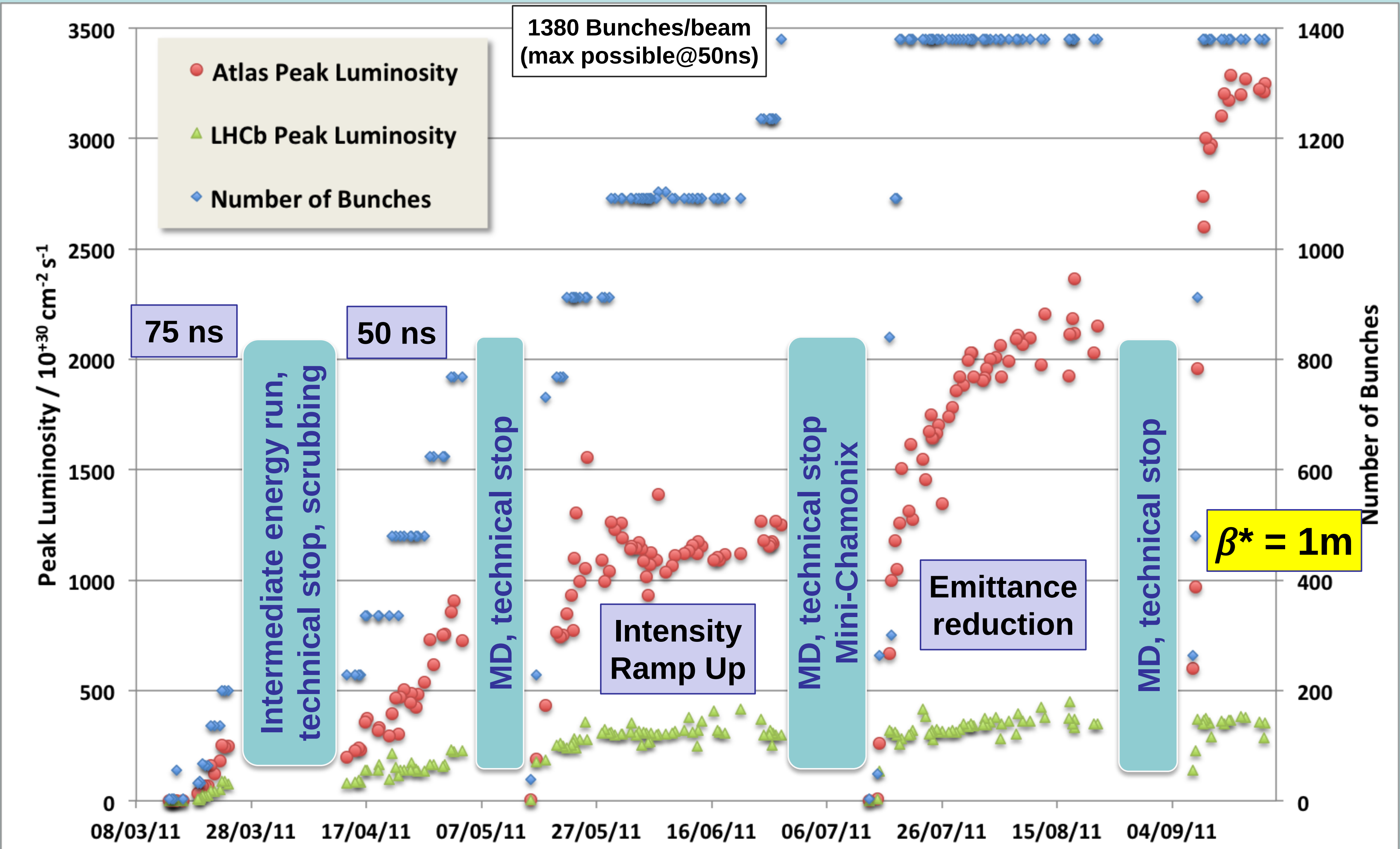
Quark/Gluon Collisions

Production of heavy particles $10^{+3…-7}$ Hz
(W, Z, t, Higgs, SUSY,…)



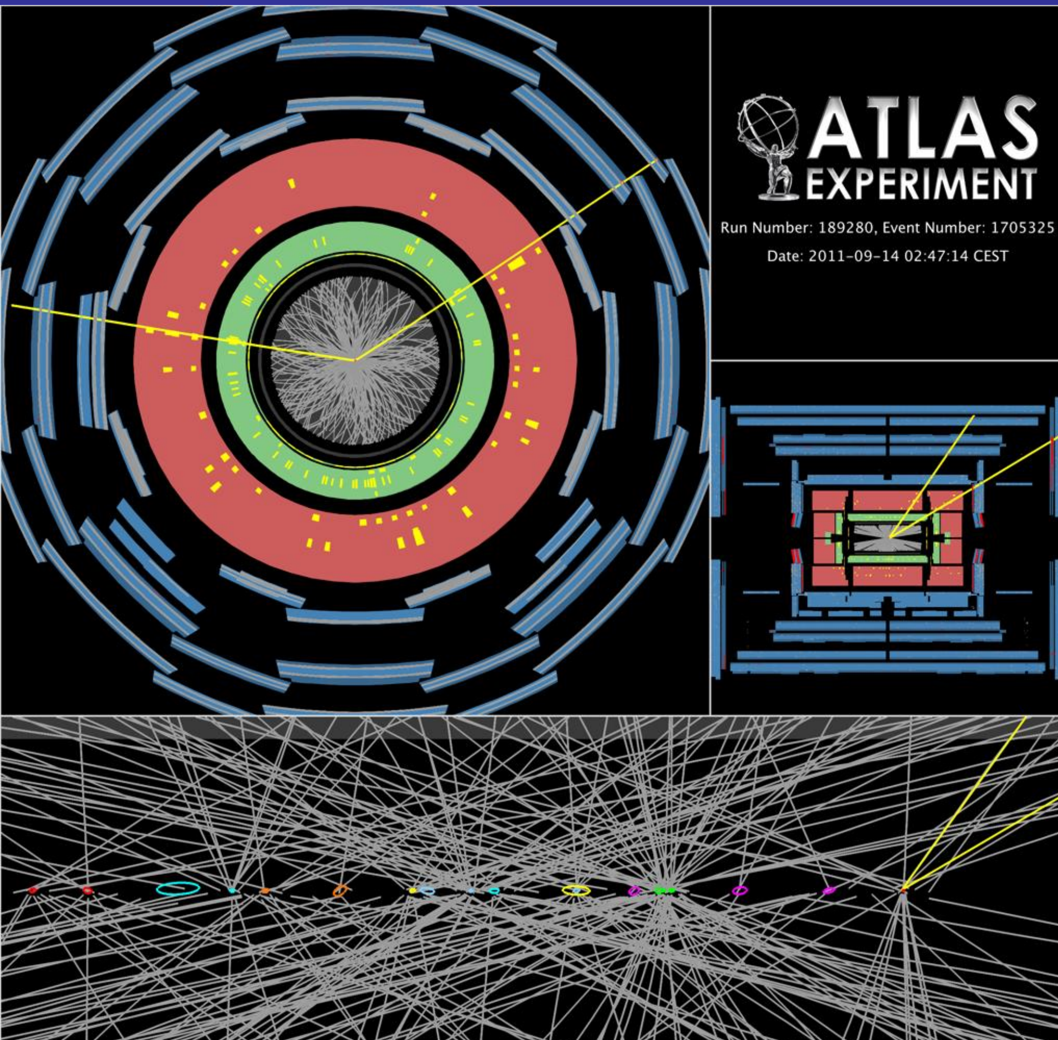
LHC status 2011 – so far

max instantaneous lumi 3.6e33, max per day 135/nb, total 4.9/fb …

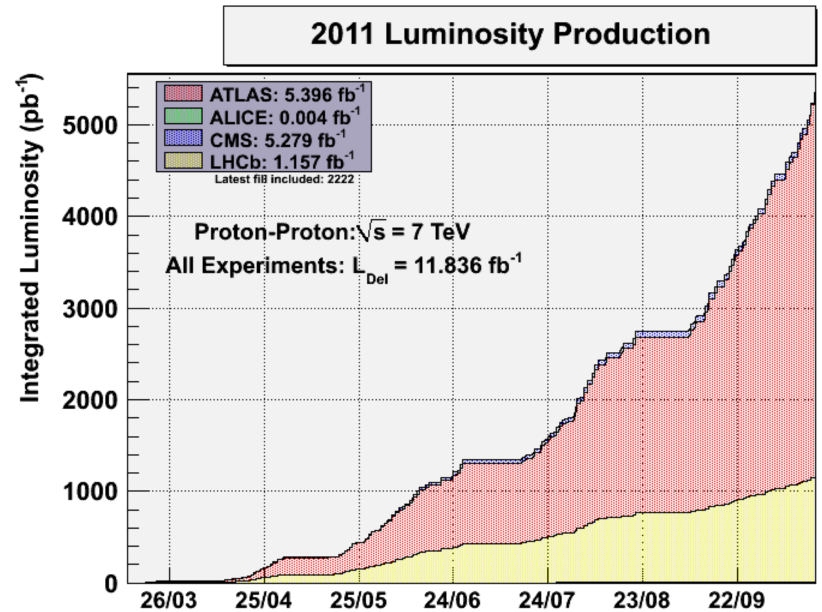


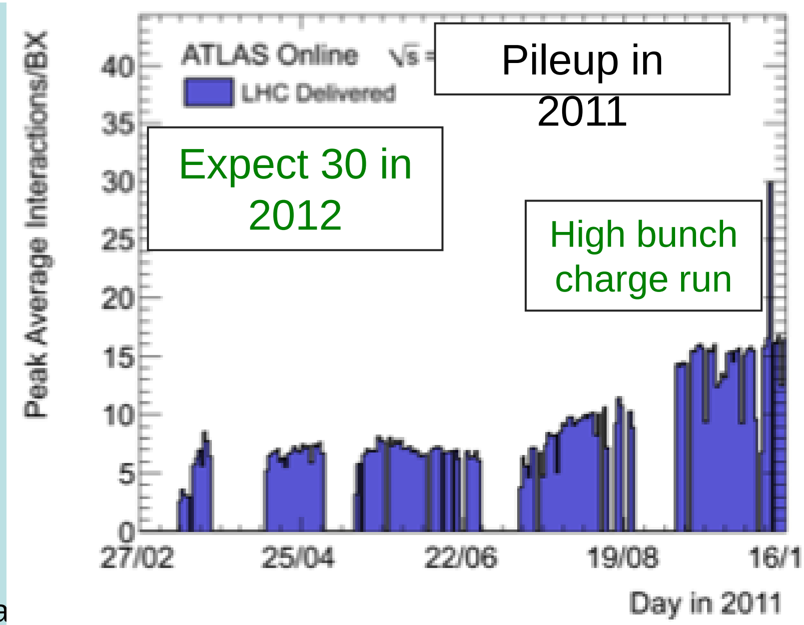
…with high pileup: <μ> ~ 16 at beginning of fill

averaged over Poisson and bunches, at start of fill



Example of Z → μμ decay with 20 primary vertices
Total scale along z is ~ ± 15 cm, p_T threshold for track reco is 0.4 GeV (ellipses have size of 20 σ for visibility)





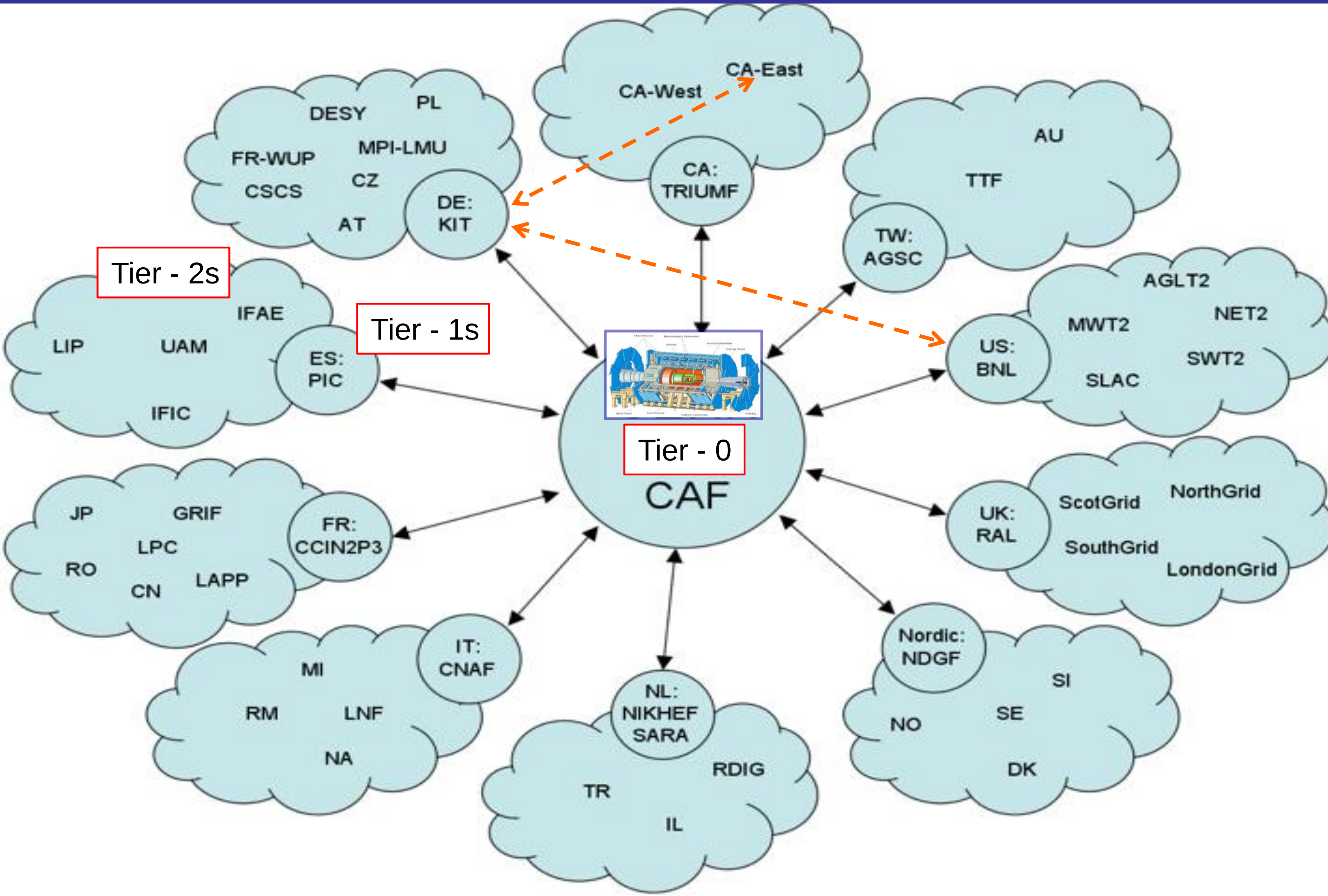
Detector status

Subdetector	Number of Channels	Approximate Operational Fraction
Pixels	80 M	97.2%
SCT Silicon Strips	6.3 M	99.2%
TRT Transition Radiation Tracker	350 k	97.5%
LAr EM Calorimeter	170 k	99.9%
Tile calorimeter	9800	98.8%
Hadronic endcap LAr calorimeter	5600	99.8%
Forward LAr calorimeter	3500	99.9%
LVL1 Calo trigger	7160	99.9%
LVL1 Muon RPC trigger	370 k	99.5%
LVL1 Muon TGC trigger	320 k	100%
MDT Muon Drift Tubes	350 k	99.8%
CSC Cathode Strip Chambers	31 k	98.5%
RPC Barrel Muon Chambers	370 k	97.0%
TGC Endcap Muon Chambers	320 k	99.1%

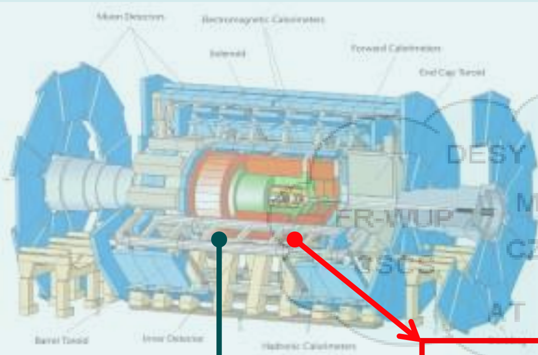
Most sub-detectors have operational fractions > 99%.

ATLAS and the worldwide Grid computing infrastructure

~80k CPU cores, ~25 PB disk worldwide



Information flow from detector to publication



Simulated data (files)

- *Generated/reconstructed on Tiers, to Tiers*

Event data (files, TAG also on DB)

- *RAW from detector reconstructed to ESD, AOD, TAG, NTUP at Tier0, to Tiers; re-reconstructed at Tier1s, to Tiers*

Metadata (DB, files) include:

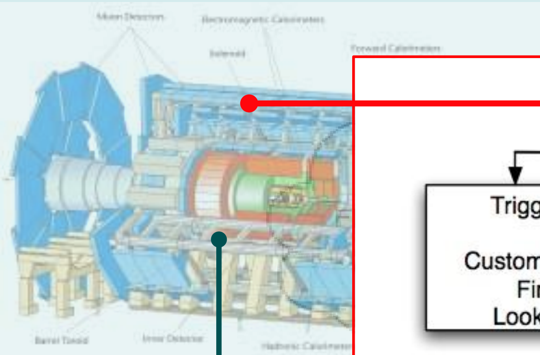
- *trigger setup (Trigger DB), dataset data*
- *running conditions / data quality (COOL DB)*
- *alignment / calibration, luminosity, backgrounds, …*

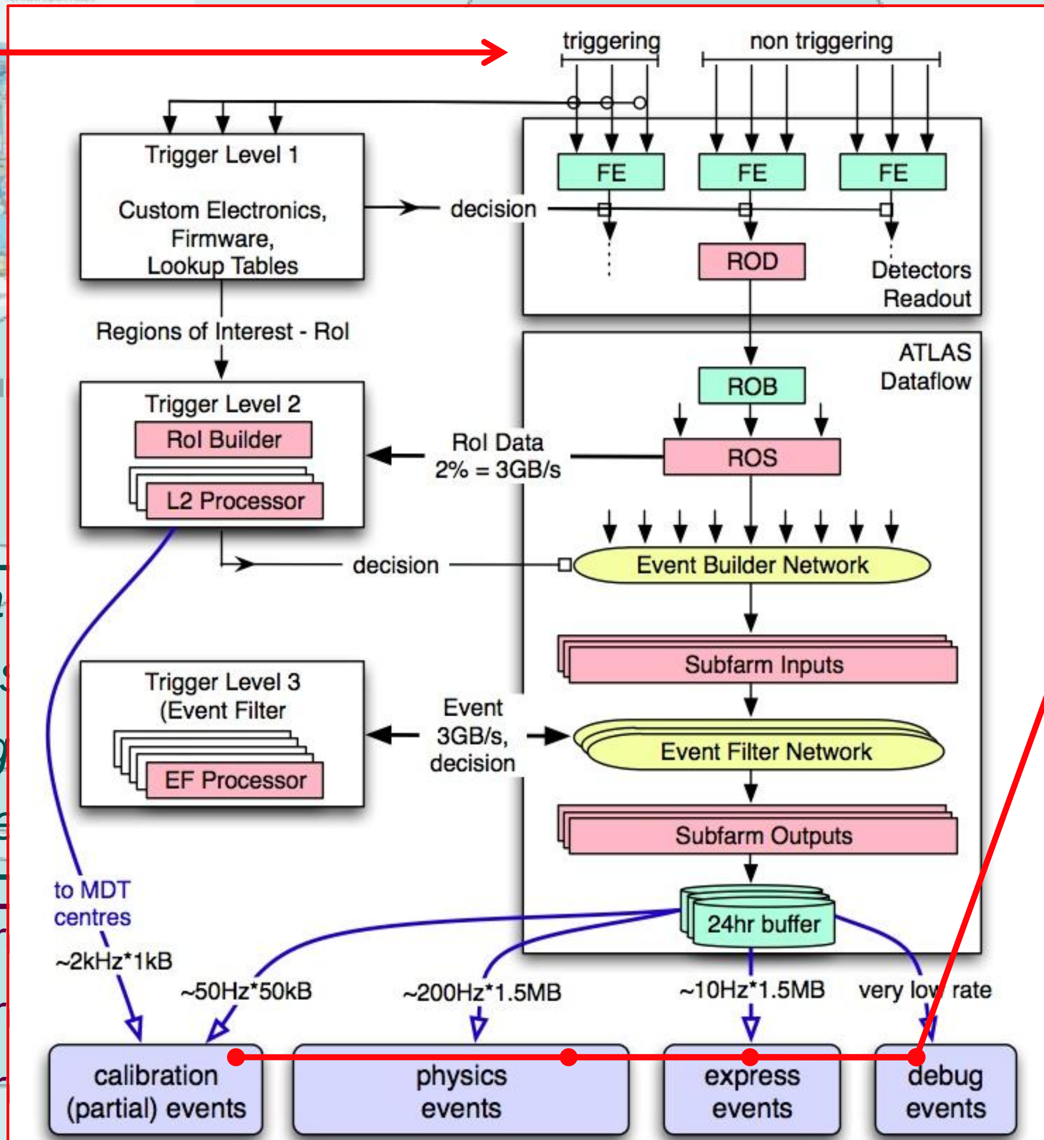
- *software (SVN)*
- *software tags used in processing*
- *less-structured input (logbooks, doc, knowledge)*

Physics analyses

Publications

Information flow – starting at Point1





...ed on Tiers, to Tiers

..., AOD, TAG, ... at Tier1s, to

Metadata

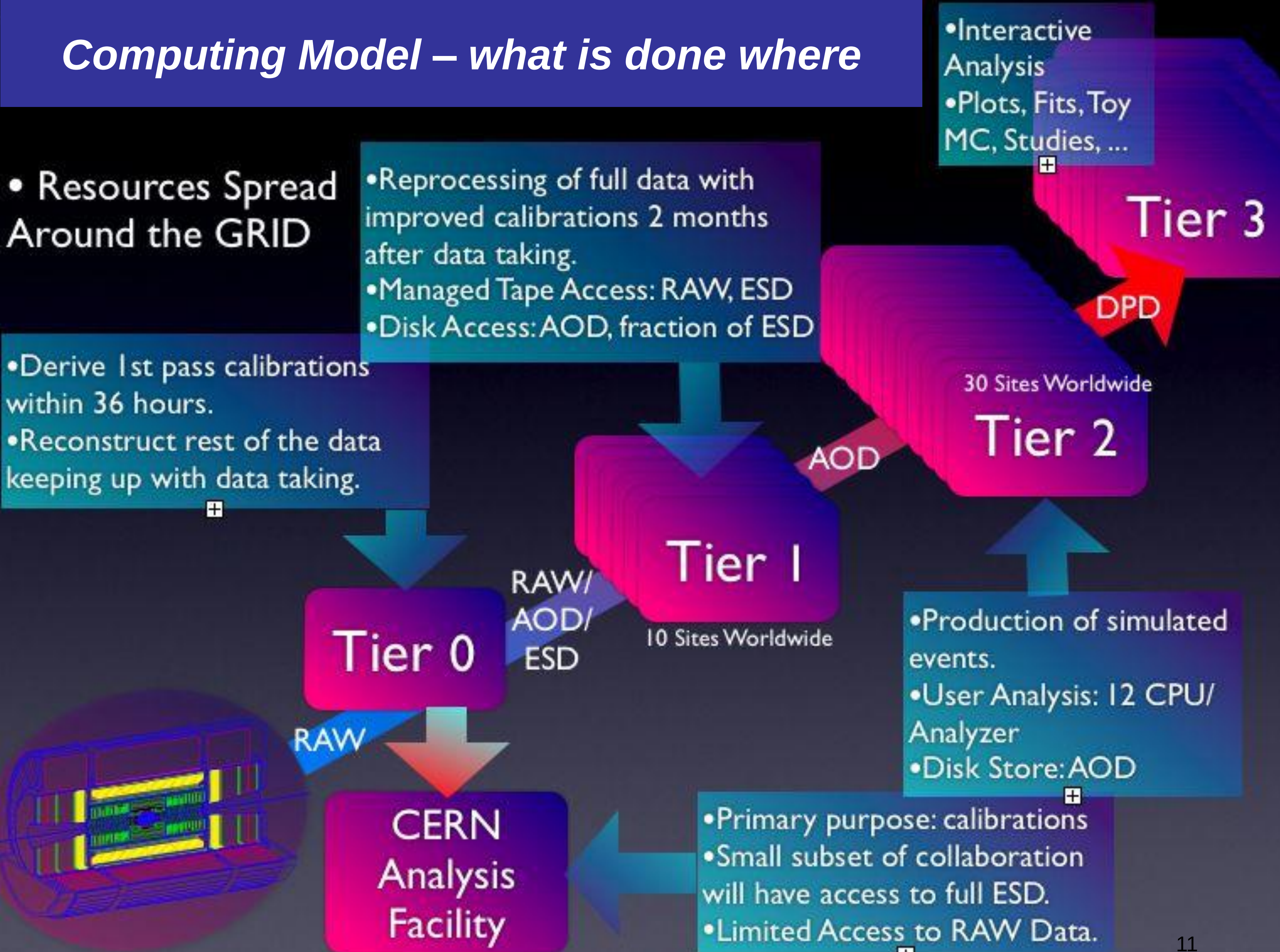
- *trigger s...*
- *running*
- *alignme...*

- *softwar...*
- *softwar...*
- *less-str...*

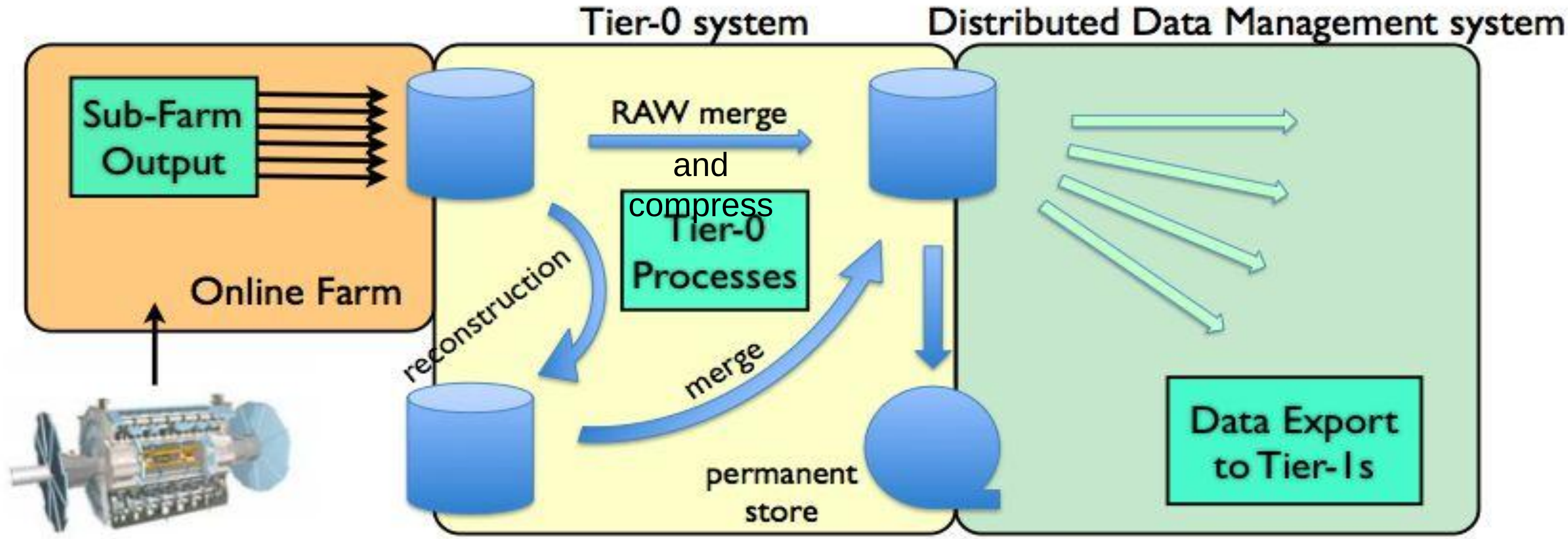
Physics analyses

Publications

Computing Model – what is done where



Data flow through the Tier-0 at CERN



Accepting data from the online system and ensuring it is archived to tape

- Merging small files to adequate size for tape archiving

Processing RAW data (event reconstruction) and archiving the products to tape

- Express stream for prompt calibration and alignment
- First-pass processing of all streams after 36h with calibration and alignment

Registering data to the ATLAS Distributed Data Management system

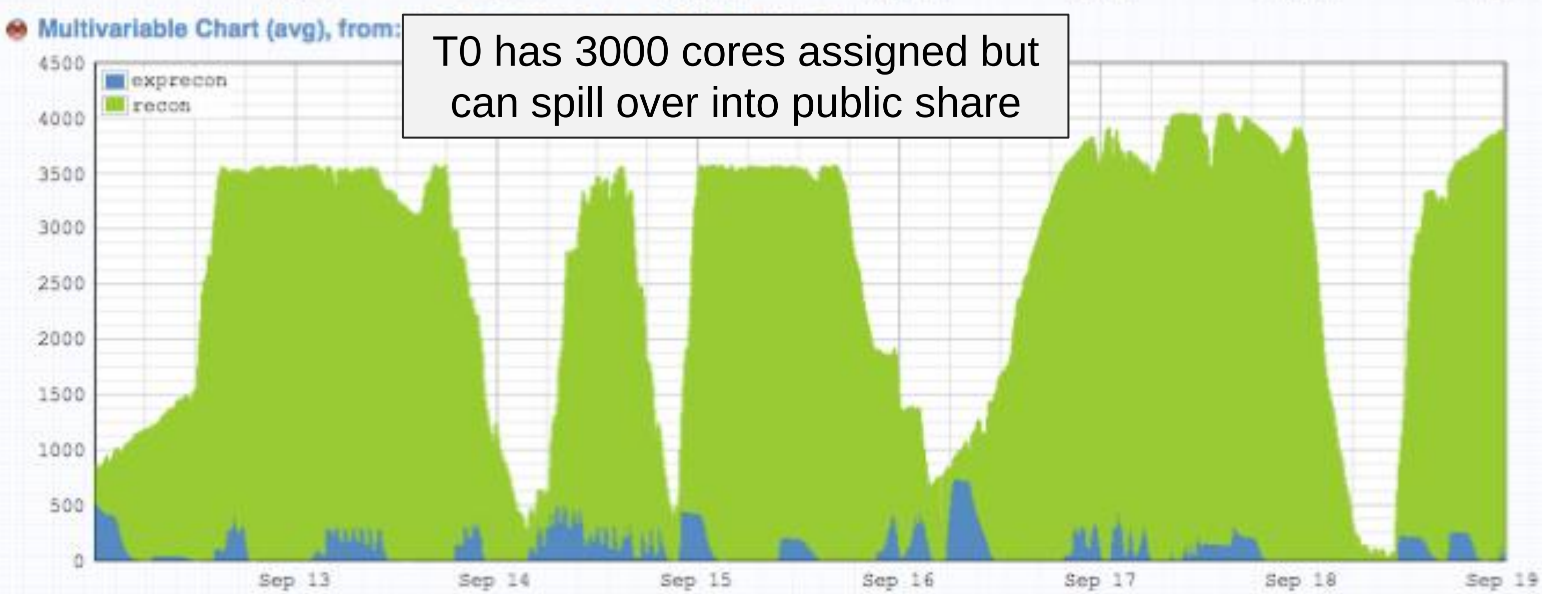
- Export data to Tier-1 and calibration Tier-2s, as well as CAF

Maximum overall I/O: 6GB/s -- including internal accesses within Tier-0

Tier-0 busy, but can keep up…

pending / running jobs

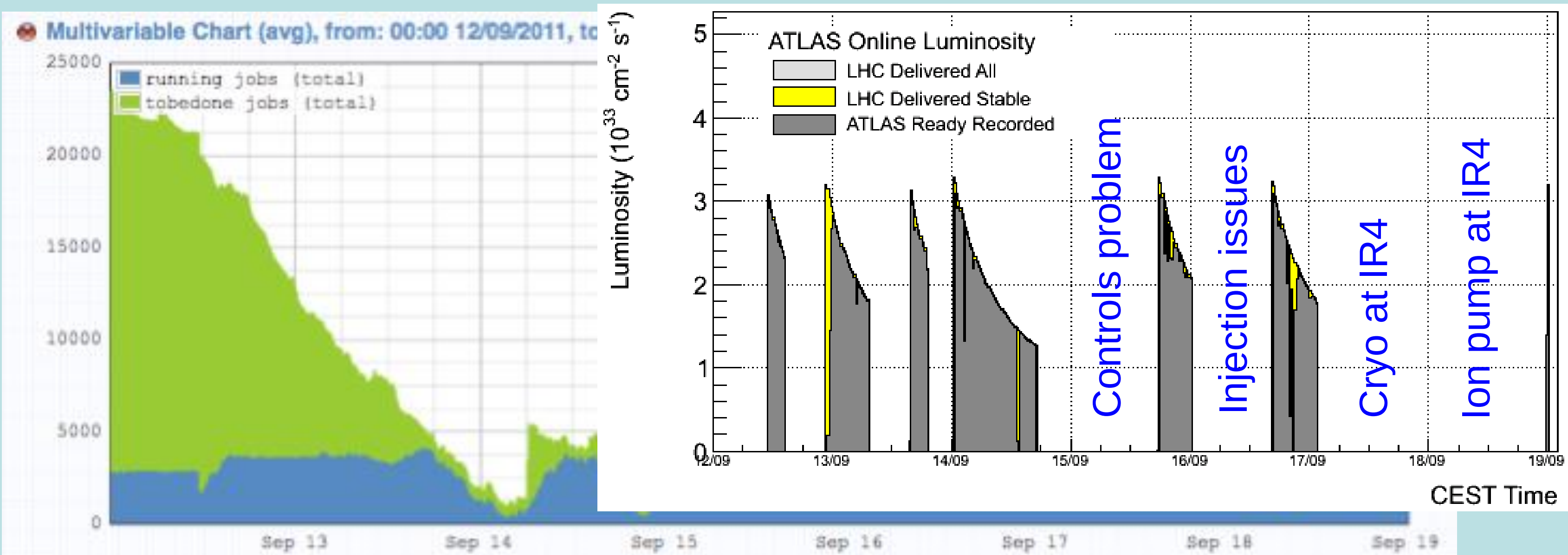
# of running jobs



T0 has 3000 cores assigned but can spill over into public share

…because not always have stable beams (~25% this year)

pending / running jobs

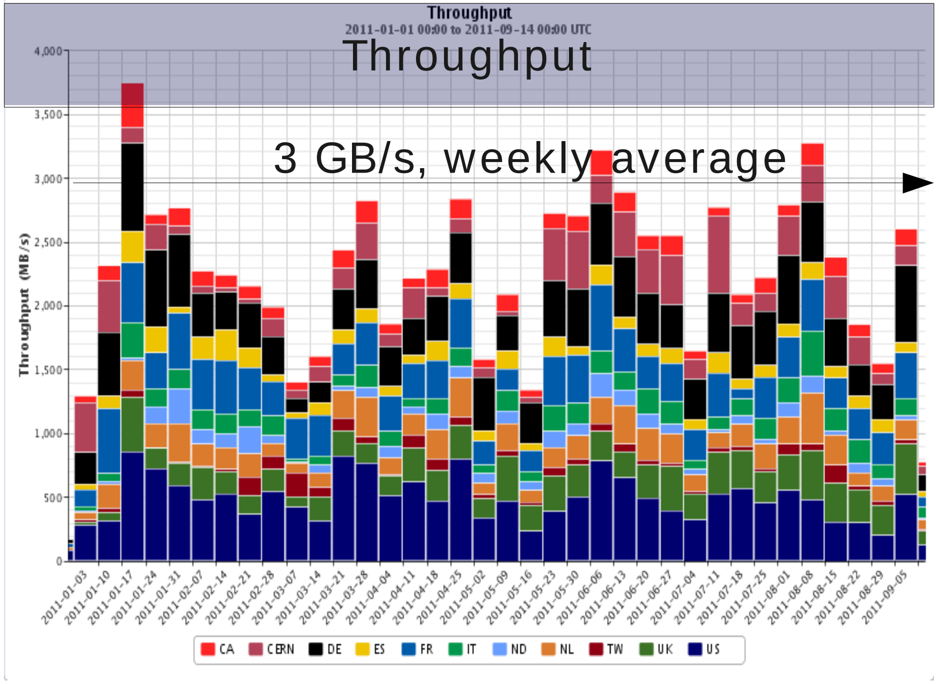


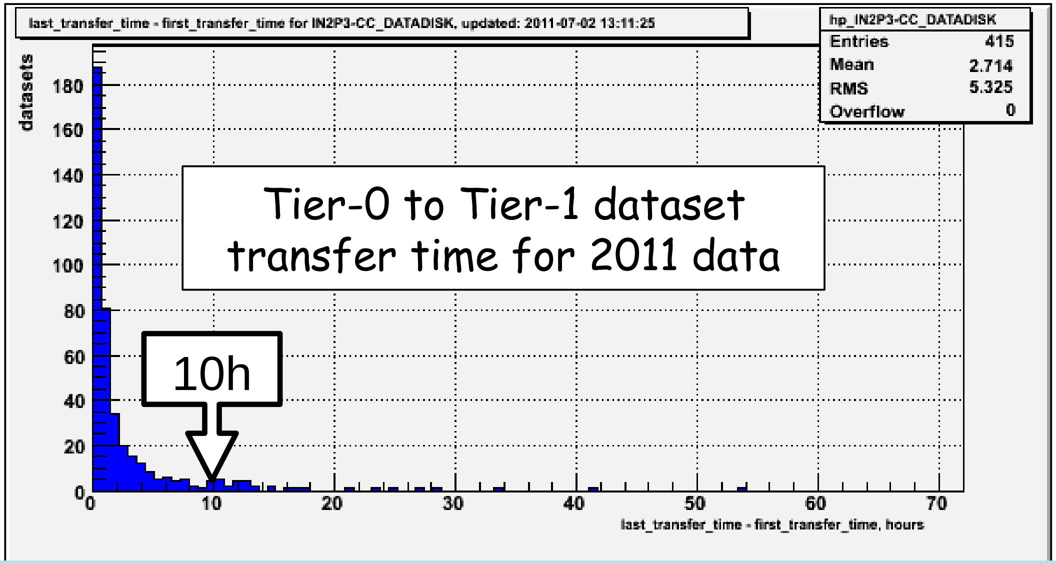
# of running jobs

Data volume handled by Tier-0 in 2011 so far:
~2.5 PB RAW recorded, ~5 PB data distributed (all data types)

ADC - Distributed Computing on the Grid: data transfers

- Data distribution
  - pre-placement
  - dynamic placement
  - user requests
- Peak throughput 10 GB/s
- Success rate 93% in 2011

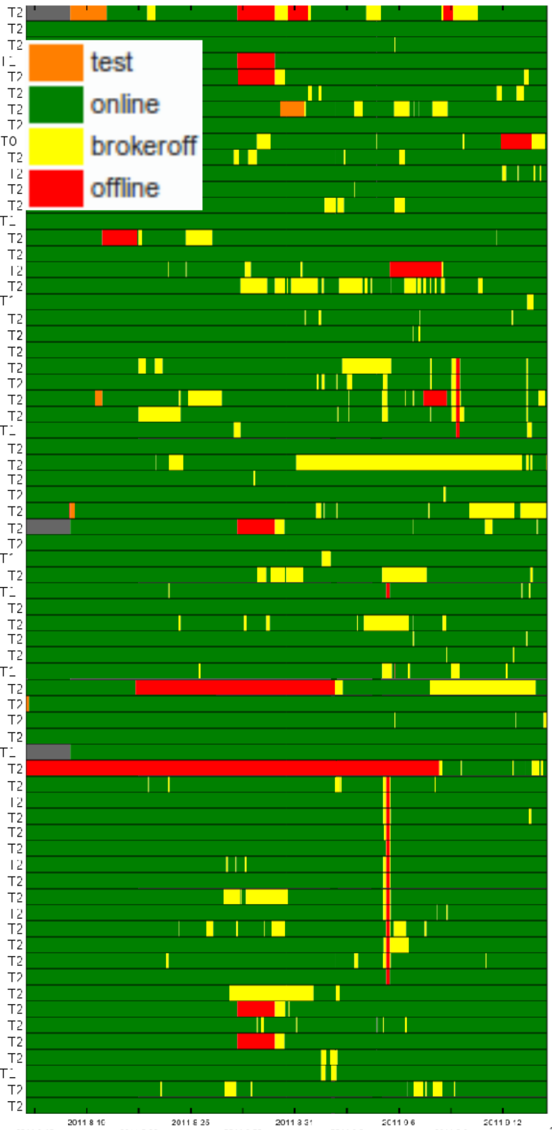


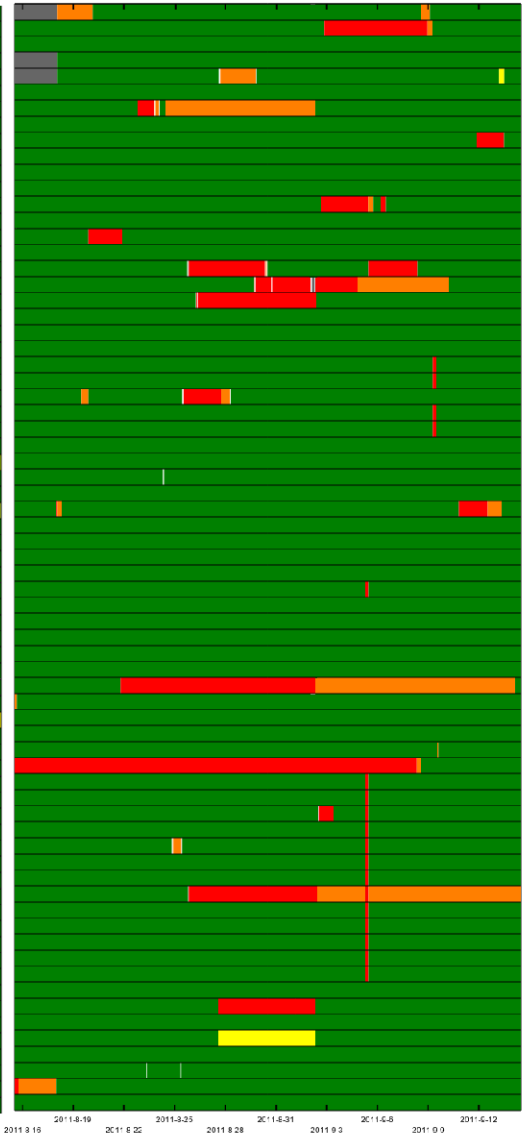


- Data are available for analysis in "almost-real" time. Example:
  - data11_7TeV AOD distribution (to one specific Tier-1 but they are all similar):
  - on average 2.7 hours to complete the dataset

ADC: data processing

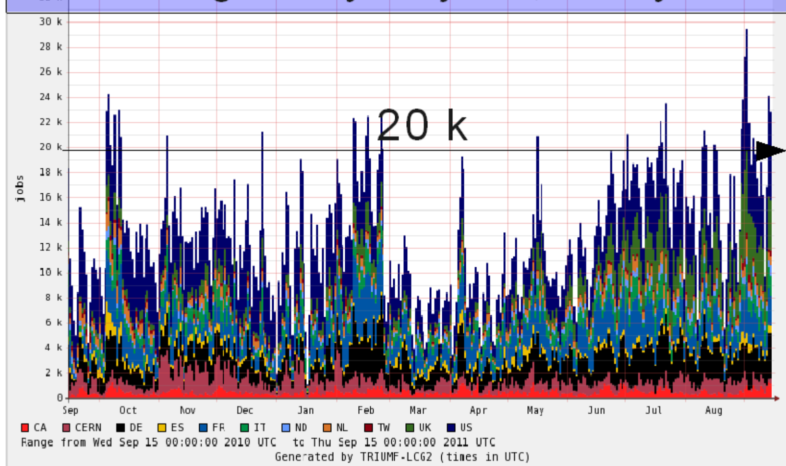
Site Status Analysis activities	Site Status Production activities





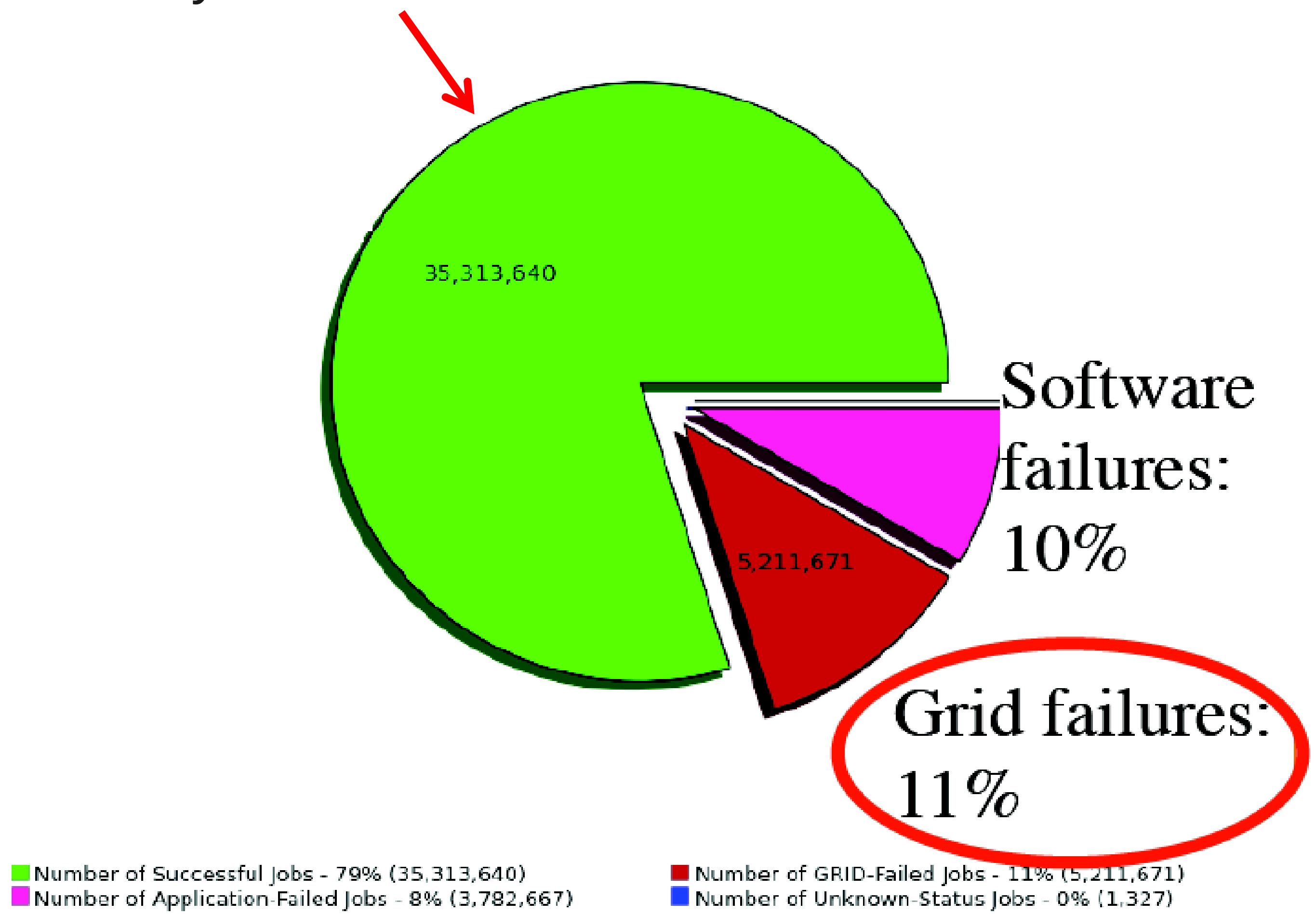
- Ca 80k jobs running simultaneously
- 12 % of CPU time spent on analysis
- Automatic job resubmission

Running analysis jobs, last year



ADC: job success rate

- Production: 90-95 % ✔
- Analysis: 79 %



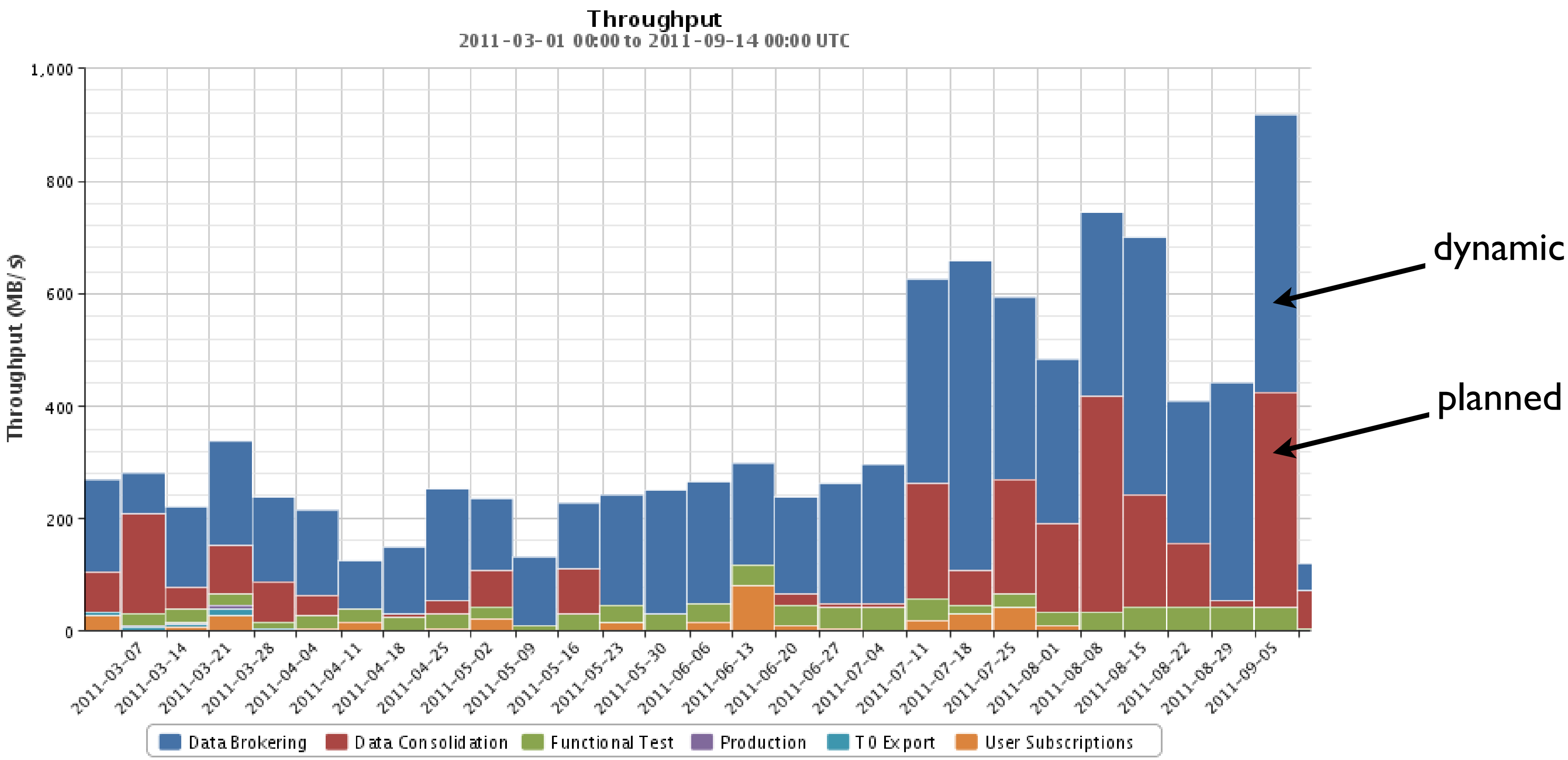
Improving success rate

- Monitor ATLAS Grid Resources 24x7
- Report to sites and cloud squads
  - 4800 GGUS tickets to the sites since 1st Jan 2010
  - Interaction with the cloud squads via e-groups, savannah tickets, ADC meetings

Cloud squad
Regional team of ADC experts; ATLAS Service Task

- Site exclusion from activities when a site is failing or when site is on downtime
- Run functional tests in activities
- Train & support ATLAS users

Tier-2 data distribution revised



Planned distribution of NTUPLE and DAOD with pre-defined share (since Jul)

Planned distribution of AOD with pre-defined share (since Sep)

Dynamic data placement with job brokerage algorithm (since last year, algorithm revised in Apr)

Dynamic data placement with pre-defined share (since Jul)

ADC: R&D and Task Forces

ATLAS Distributed Computing held a “Technical Interchange Meeting” at Dubna

- See https://indico.cern.ch/conferenceDisplay.py?confId=132486
- Cloud model relaxation
- Move towards using CVMFS (web-based file system) for software release and conditions data files distribution
- Integrate all 11 Local File Catalogues into a single catalogue at CERN
- DDM team is collecting requirements for a new generation of data management system
- Production System evolution according to new challenges
- R&D on noSQL databases
- R&D on Cloud computing – seeking simplifications usable on the Grid

Software

- Release 17 - Summer reprocessing and use at Tier-0
- Improved handling of the increased pileup
  - complete rework of the pixel clustering to better separate nearby tracks
  - further improvements, both technical or to improve physics output, are still in the pipeline
- Analysis tools: harmonize the tools used in the different physics group, support production of derived ntuples
- Use improved calibration derived from the data itself
- Introduction of extra cut levels in Inner Detector tracking reduce combinatorics
- Collaboration with Intel in tools, compilers, fine-grain parallelism (vectorise), involves offline and TDAQ, usage of GPUs for track finding and fitting
- More: redo the I/O framework, …

Data Quality – example of improvements

1st Tier-0 processing

Inner Tracking Detectors			Calorimeters				Muon Detectors				Magnets	
Pixel	SCT	TRT	LAr EM	LAr HAD	LAr FWD	Tile	MDT	RPC	CSC	TGC	Solenoid	Toroid
99.9	99.8	100	89.0	92.4	94.2	99.7	99.8	99.7	99.8	99.7	99.3	99.0

Luminosity weighted relative detector uptime and good quality data delivery during 2011 stable beams in pp collisions at √s=7 TeV between March 13th and June 29th (in %). The inefficiencies in the LAr calorimeter will partially be recovered in the future. The magnets were not operational for a 3-day period at the start of the data taking.

- Data quality close to 100% for all sub-detectors apart from LAr calorimeter in Tier0 processing
- Origin of lower data LAr quality is mostly noise bursts (and HV trips)

Reprocess

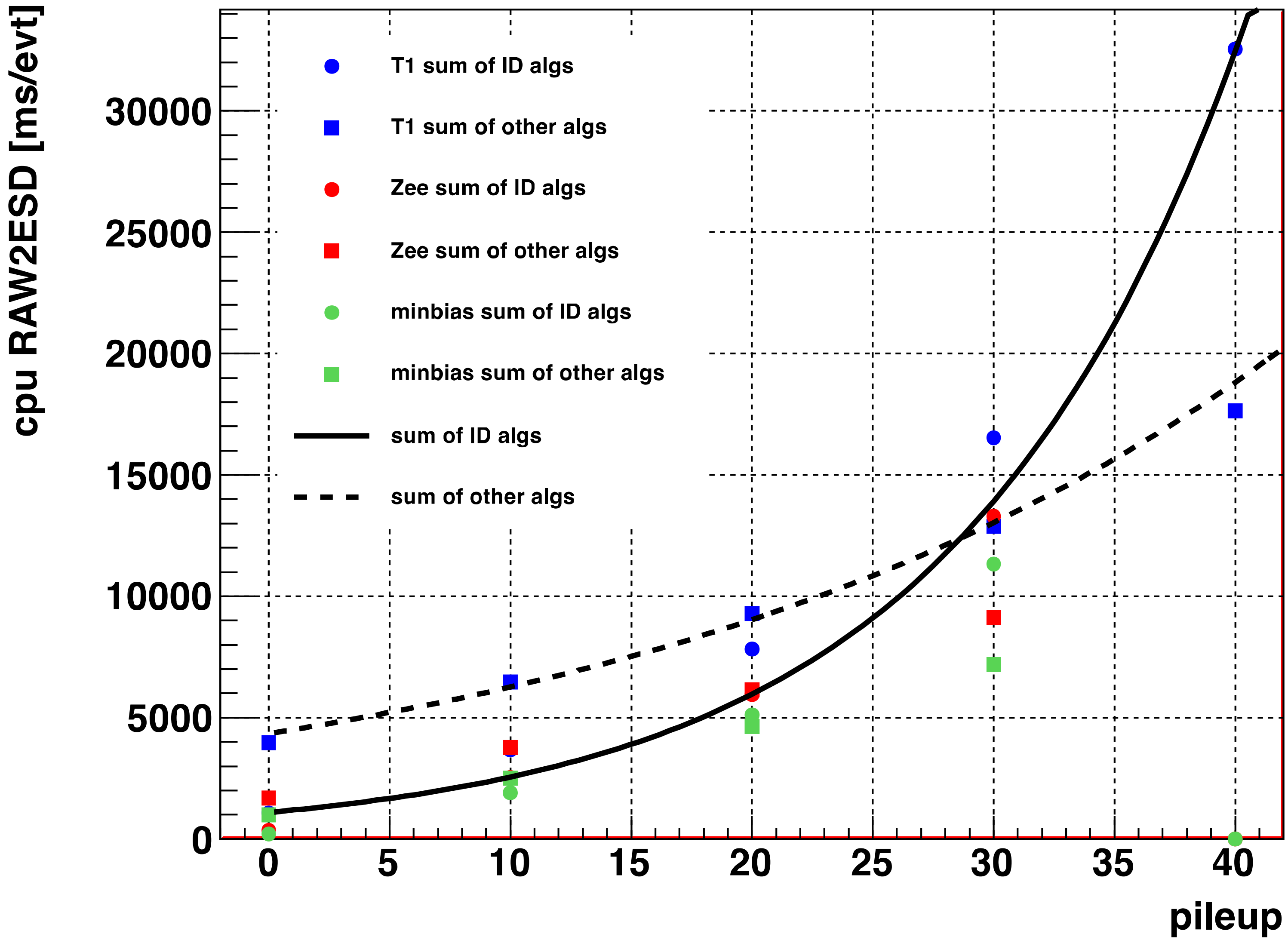
Inner Tracking Detectors			Calorimeters				Muon Detectors				Magnets	
Pixel	SCT	TRT	LAr EM	LAr HAD	LAr FWD	Tile	MDT	RPC	CSC	TGC	Solenoid	Toroid
99.9	99.8	100	96.3	98.6	98.9	99.7	99.8	99.8	99.8	99.7	99.3	99.0

Luminosity weighted relative detector uptime and good quality data delivery during 2011 stable beams in pp collisions at √s=7 TeV between March 13th and June 29th (in %).

- In reprocessing, event by event flagging of noise bursts was used
- Gain back about 7% of the data for physics analyses (now also at Tier-0)

Reconstruction time vs. <μ>

Inner Detector related algorithms will take more than the rest when <μ> > 28



For 2012: expect increase by factor 1.5-2 in max lumi and pileup ➔ <μ> = 30

How can we use our 80k CPU cores efficiently in parallel?

- Trivial in principle. Event data are independent so they can be processed easily in parallel
  - only almost true: several/many events share common files, common metadata like running conditions
- CPU boxes are coming with more and more cores, esp. true for the graphics processing units (GPU) or CPU-GPU integrated architectures
  - from now 8 cores per box soon to >100 cores per box
  - the answer is to use more fine-grain parallelism in addition to event parallelism – we are actively working on this
  - the linear algebra in the inner loops of track reconstruction are especially suited for more parallelism
  - …also the neural-network algorithms which disentangle multiple hits in the pixel detector
  - useful tools arriving – thread and array building blocks, CILK; all C++

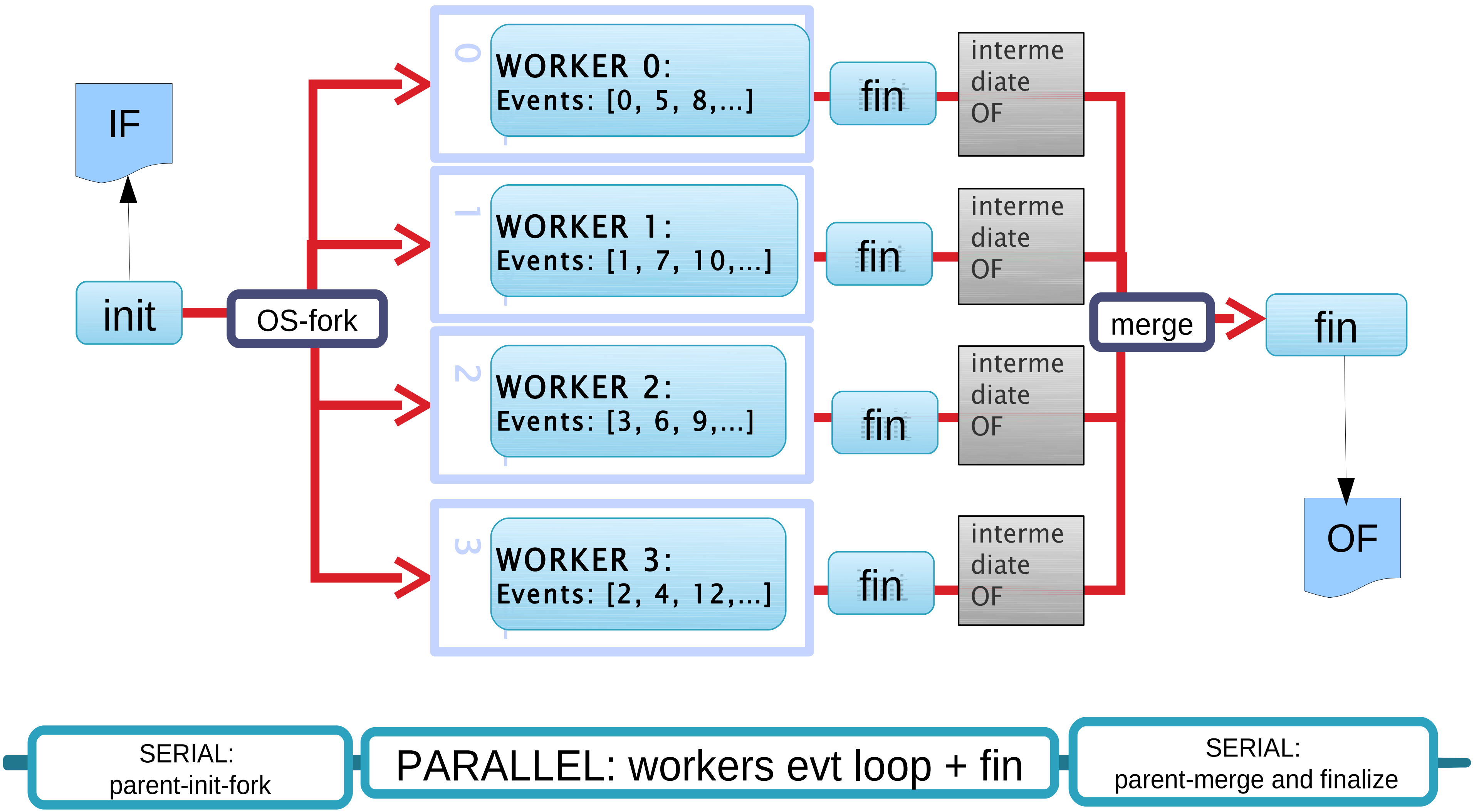
Gaining speed by saving memory…

full use of multi (8-16) to many (48++) core processors

- Multithreading would be the natural choice…
  - Turned out problematic, even in trigger, because much code was not thread-safe (new releases of external libraries, …)
- Multi-process approach: athenaMP
  - One job submitted to one empty multi-core processor
  - Forks one process per (virtual) core, sharing common memory via the copy-on-write mechanism
  - Output merged at the end
- Needs less memory per core
  - Important when using hyperthreading (within reach) and 64 bit (still difficult) so can have some speedup per physical core as well (~30%)
- Status
  - First real-life tests at Tier0 during last Technical Stop: successful
  - Hope to be production ready for the Heavy Ions run, else 2012

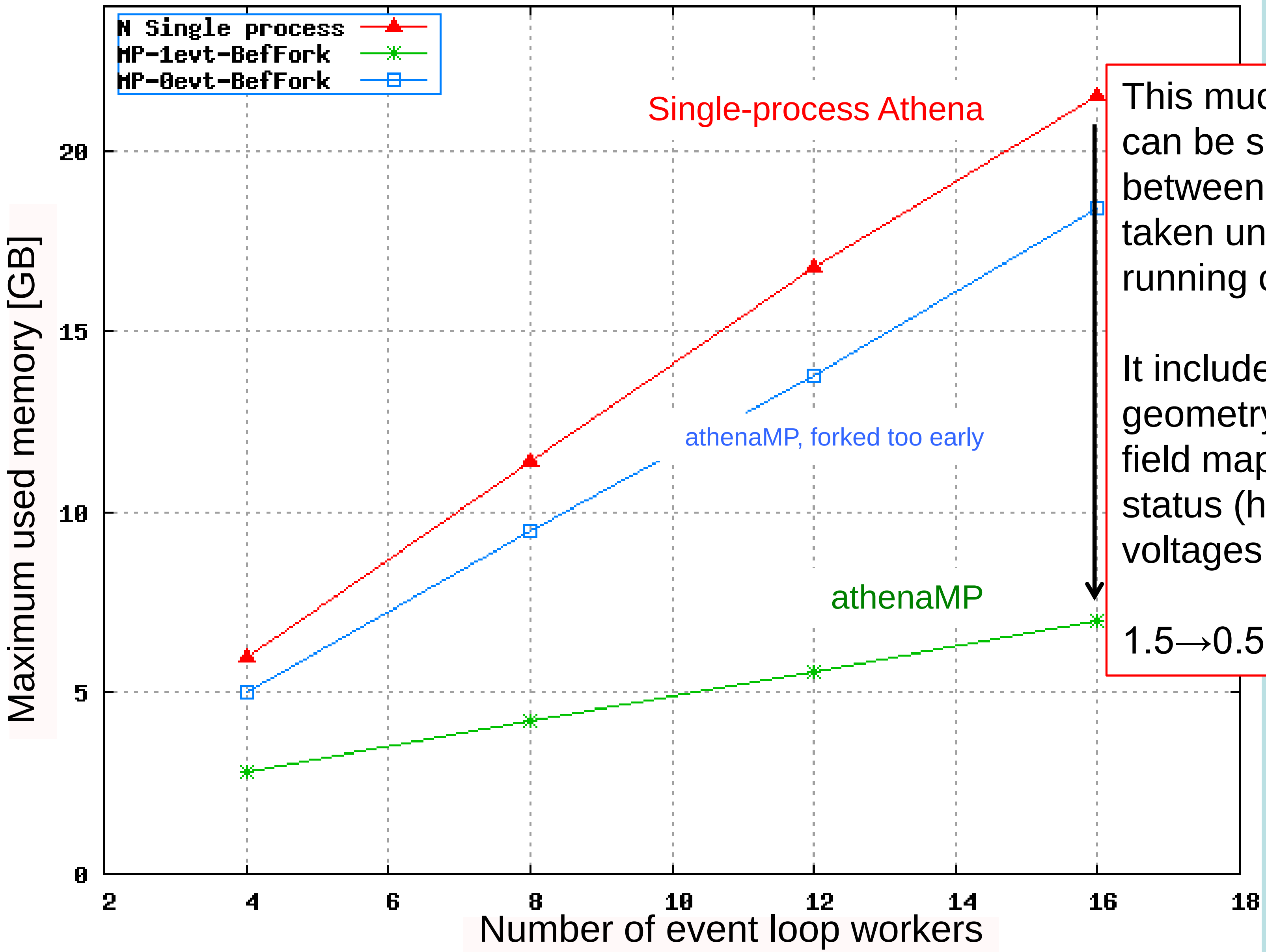
Event-level parallelism in athenaMP

https://twiki.cern.ch/twiki/bin/viewauth/Atlas/AthenaMP



Development of fast “hybrid” merging usable also outside athenaMP.
Development ongoing of shared memory usage for event passing.

Memory used (8-core machine with hyperthreading, 24GB)
forking after 1st event

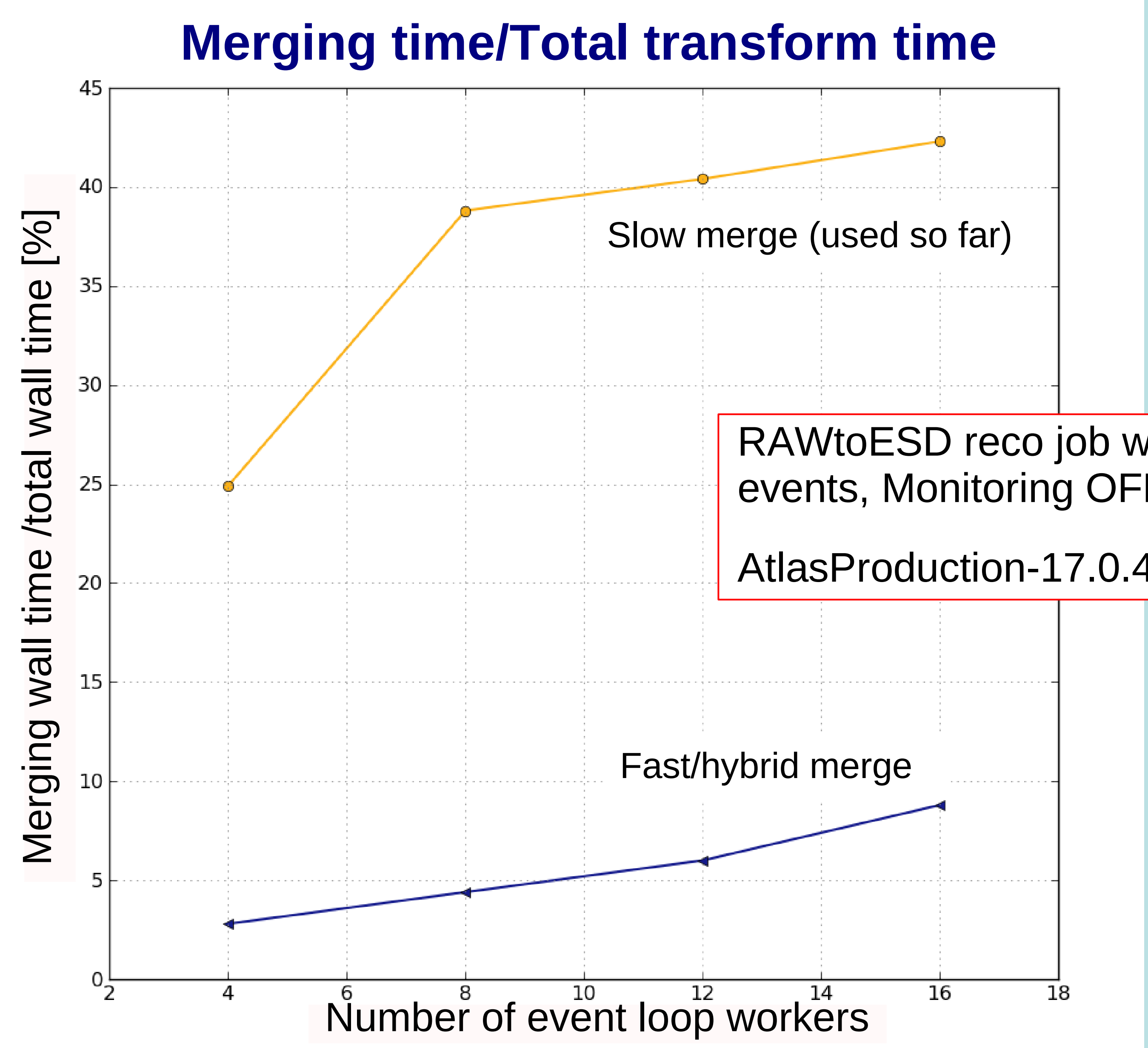


This much memory can be shared between events taken under similar running conditions.

It includes the geometry, magnetic field maps, detector status (high voltages, etc.)

1.5→0.5 GB/process

Fast (hybrid) merging for POOL files
gives 10 speedup*

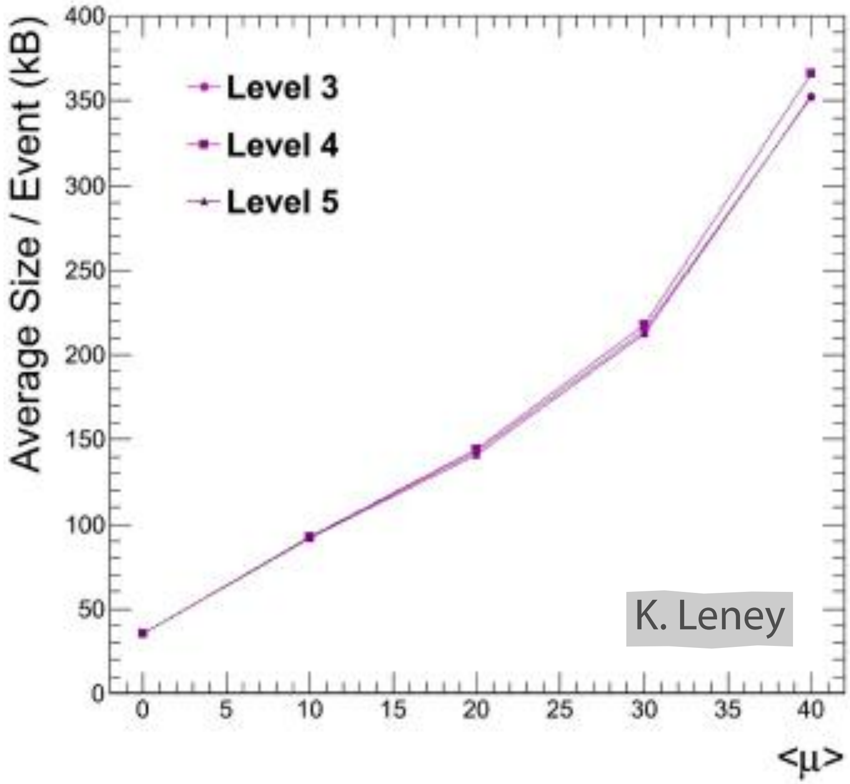


RAWtoESD reco job with 1.5 K events, Monitoring OFF

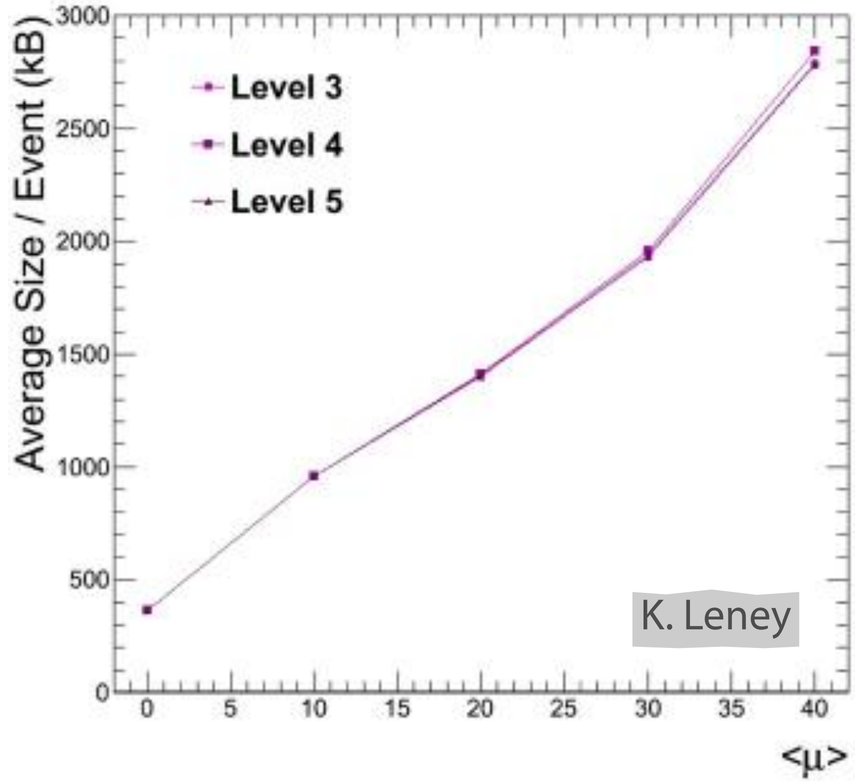
AtlasProduction-17.0.4.2

Event size vs. pileup

Average AOD event size (kB)

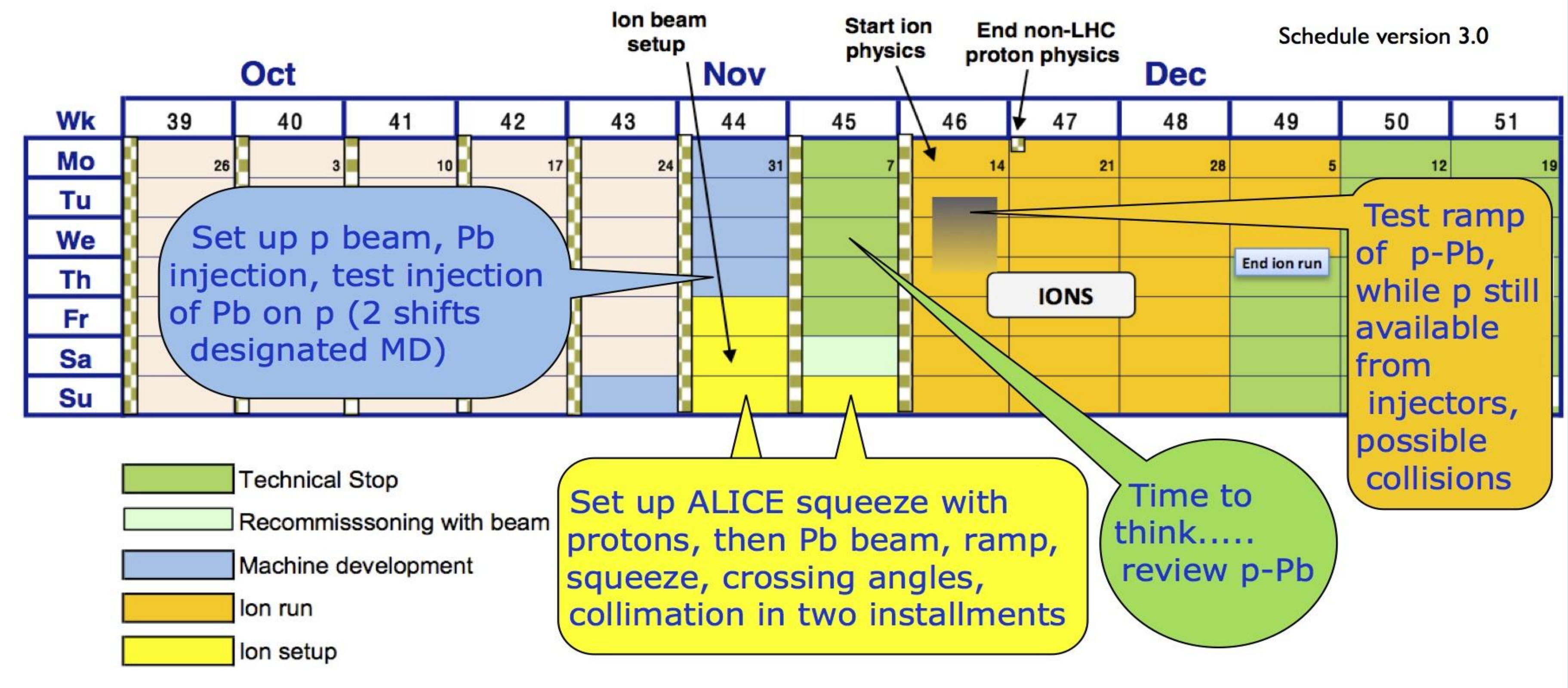


Average ESD event size (kB)



RAW event size ~700kB/event (compressed), does not depend much on <μ>
we write RAW to disk (keep for 1 year), but little ESD (6 weeks buffer)

Heavy Ions *running in 2011*



- Expecting 5-10 times more Pb-Pb luminosity than in 2010
- Need more selective triggering, including HLT
- 100 Hz of MinBias (ZDC) plus 100 Hz of High p_T
- Expected RAW size (compressed) 2.5 MB/event
- HI resource usage should be ≈ 10% of pp resources

Resource request *updated for 2012/13*

- The only change vs. the March 2011 estimate is in CERN CPU

CPU [kHS06]	2011	2012	2013
CERN	74	*73→111*	*111*
Tier-1	202	259	280
Tier-2	275	295	321
Disk [PB]			
CERN	7	9	10
Tier-1	22	27	30
Tier-2	35	49	56
Tape [PB]			
CERN	14	18	18
Tier-1	28	36	40

  - Event sizes, reconstruction times, Pileup, trigger rate, etc. have all changed and we expect that to continue
  - We adjust our computing model to fit within the resource constraints we have for 2011 and 2012
  - We expect to be able to maintain the current trigger rate with an increase of resources only at the Tier-0

For 2012: expect increase by factor 1.5-2 in max lumi and pileup ➔ <μ> = 30

(Possible) LHC time-line

		$\int L\,dt$
2009	**Start of LHC - 2009: $\sqrt{s}$ = 900 GeV**	
2010, 2011, 2012	**Run 1: $\sqrt{s}$ = 7 TeV** (2012: 8-9 TeV ?), **L = 2-3 x 10^{33} $cm^{-2}s^{-1}$** **Bunch spacing: 75/50/25 ns** (25 ns tests 2011; 2012 ?)	~10 fb^{-1}
2013, **2014**	**LHC shutdown to prepare for design energy and *nominal* luminosity**	
2015, 2016, **2017**	**Run 2: $\sqrt{s}$ = 14 TeV, L = 1 x 10^{34} $cm^{-2}s^{-1}$** **Bunch spacing: 25 ns ?** (what to do with electron clouds ?)	~50 fb^{-1}
2018	**Injector and LHC Phase-I upgrade to go to *ultimate* luminosity**	
2019, 2020, 2021	**Run 3: $\sqrt{s}$ = 14 TeV, L = 2 x 10^{34} $cm^{-2}s^{-1}$** **Bunch spacing: 25 ns**	~300 fb^{-1}
2022	**High-luminosity LHC (*HL-LHC*), crab cavities, lumi levelling, ...**	
2023, ..., **2030**	**Run 4: $\sqrt{s}$ = 14 TeV, L = 5 x 10^{34} $cm^{-2}s^{-1}$** **Bunch spacing: 25 ns**	~3000 fb^{-1}

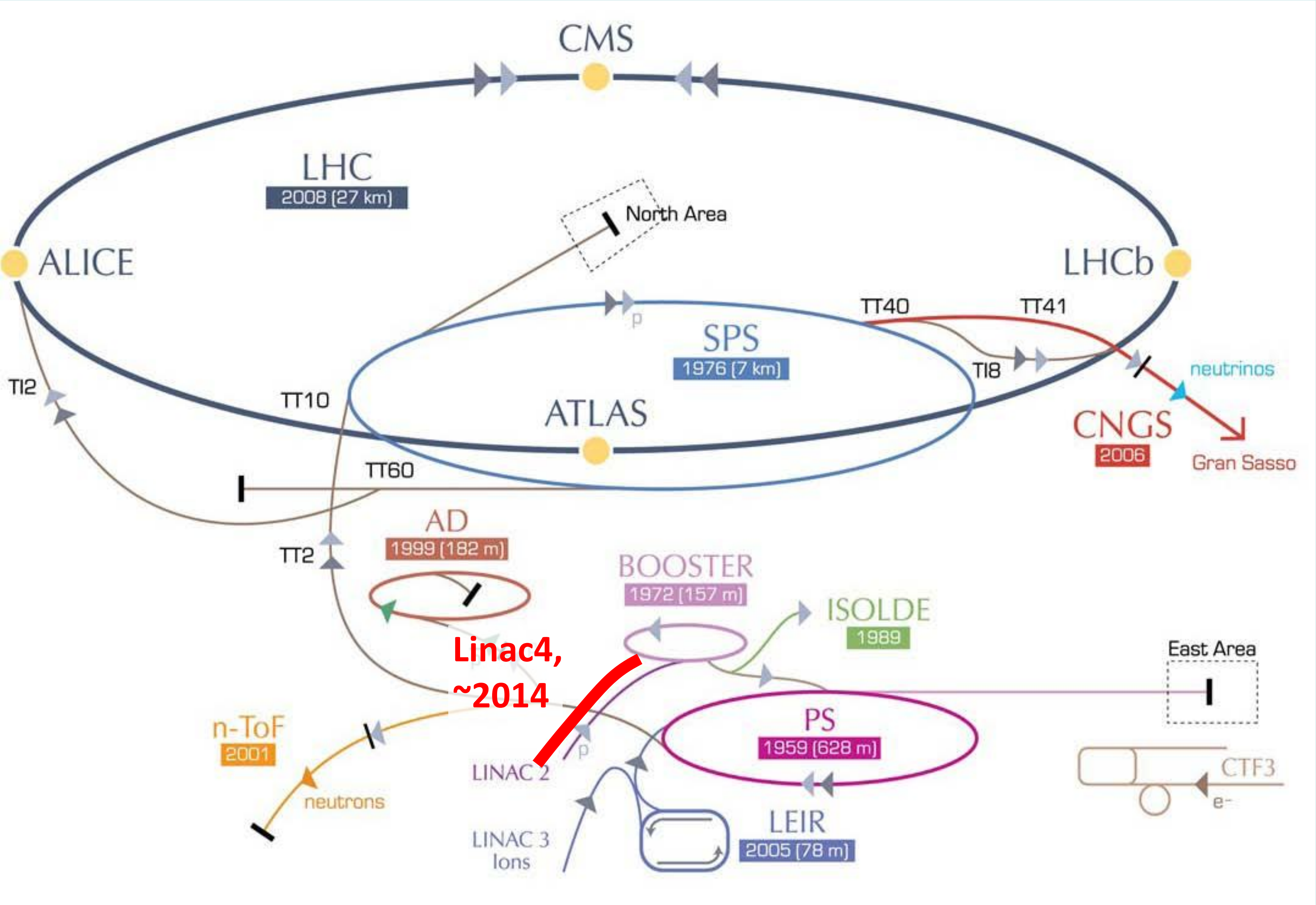
Possible ATLAS Upgrade time-line

2009	**Present ATLAS detector**
2010	Many outstanding physics and performance results, improve detector understanding and modeling in simulation
2011	
2012	
2013	**Phase 0**
2014	New innermost pixel layer (IBL) removal of Minimum Bias Scintillators, detector consolidation of Muon System, new neutron shielding
2015	
2016	
2017	**Phase I**
2018	Under consideration: new pixel detector based on IBL experience, warm miniature forward calorimeter, update of the small muon wheel, trigger adjustments (topological trigger)
2019	
2020	
2021	
2022	**Phase II**
2023	All new innermost tracking detector, forward calorimeter upgrade, additional trigger and precision chambers in the muon system, extra neutron shielding
...	
2030	

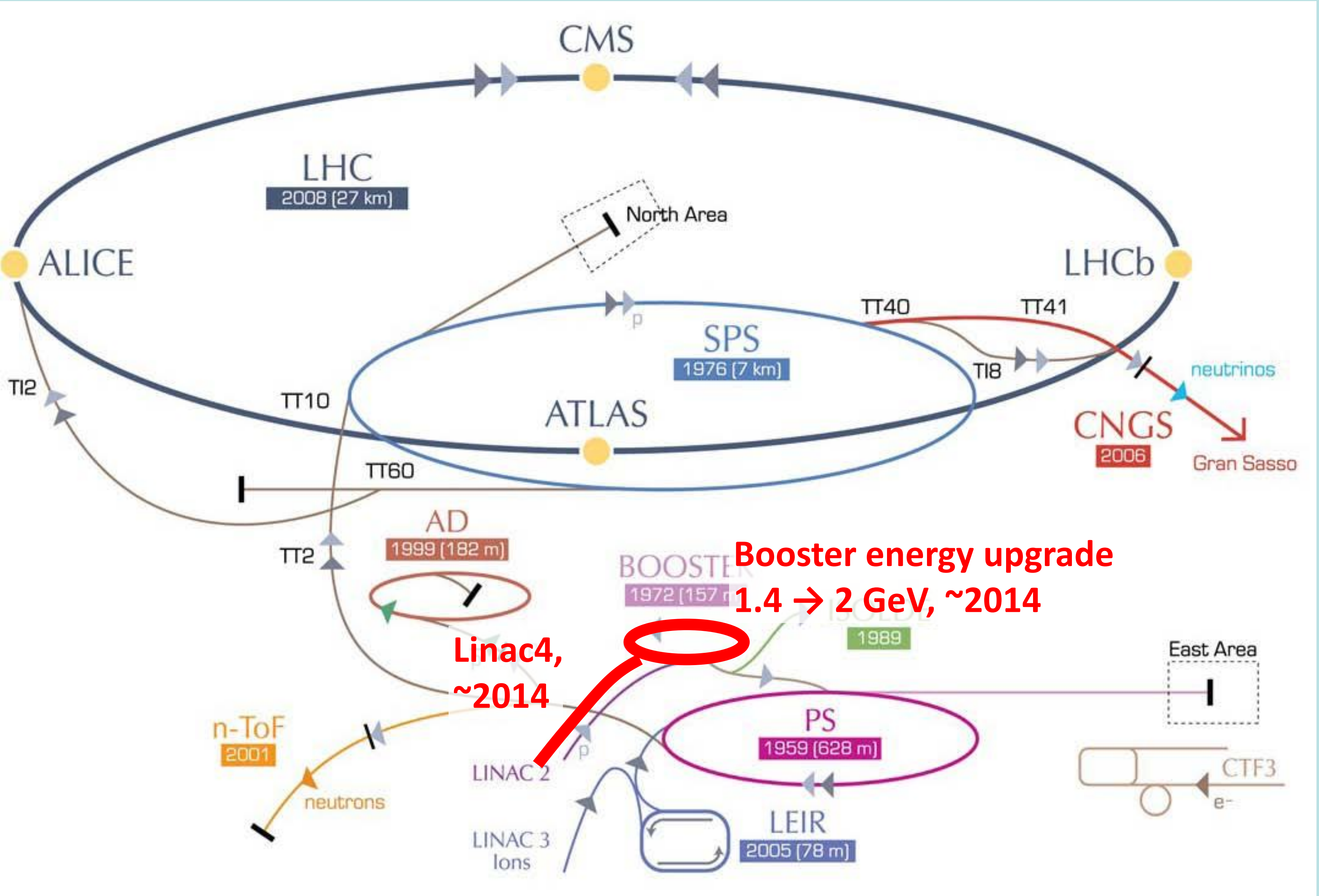
LHC has a lot in the pipeline for longer term !
achievements from the Machine Development periods

- Injection of 25 ns bunch trains with up to 216 bunches
- Luminosity leveling (i.e. constant during fill) – already in place for LHCb
- Collision of (individual) bunches with twice nominal intensity and half emittance, demonstrating 8 times nominal bunch luminosity (“fat bunch” only was used in run 190728 on 10 October)
- Injection and storage of even higher bunch intensities with nominal emittance
- Collision of 50 ns bunch trains with 4-5 sigma separation, demonstrating margin in long-range beam-beam effects
- First squeeze below 1.5 m, demonstrating β^* = 0.3 m with pilot beam, flat machine, no collisions and ATS optics (achromatic telescopic squeeze)
- Many detailed studies that were needed to achieve the above results and will make it usable (RF, injection, collimation, quench margins, optics, …)

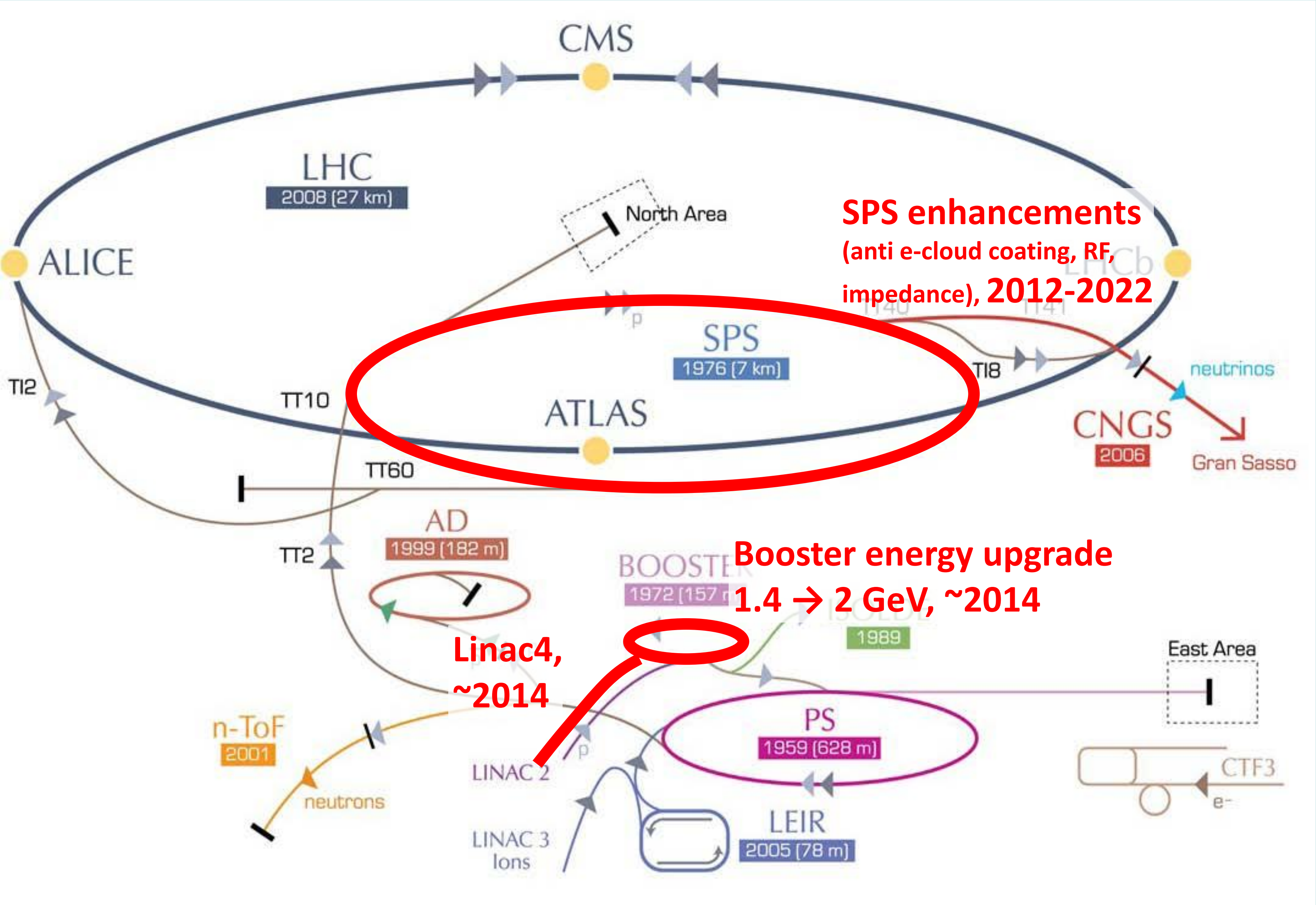
HL-LHC – accelerator modifications



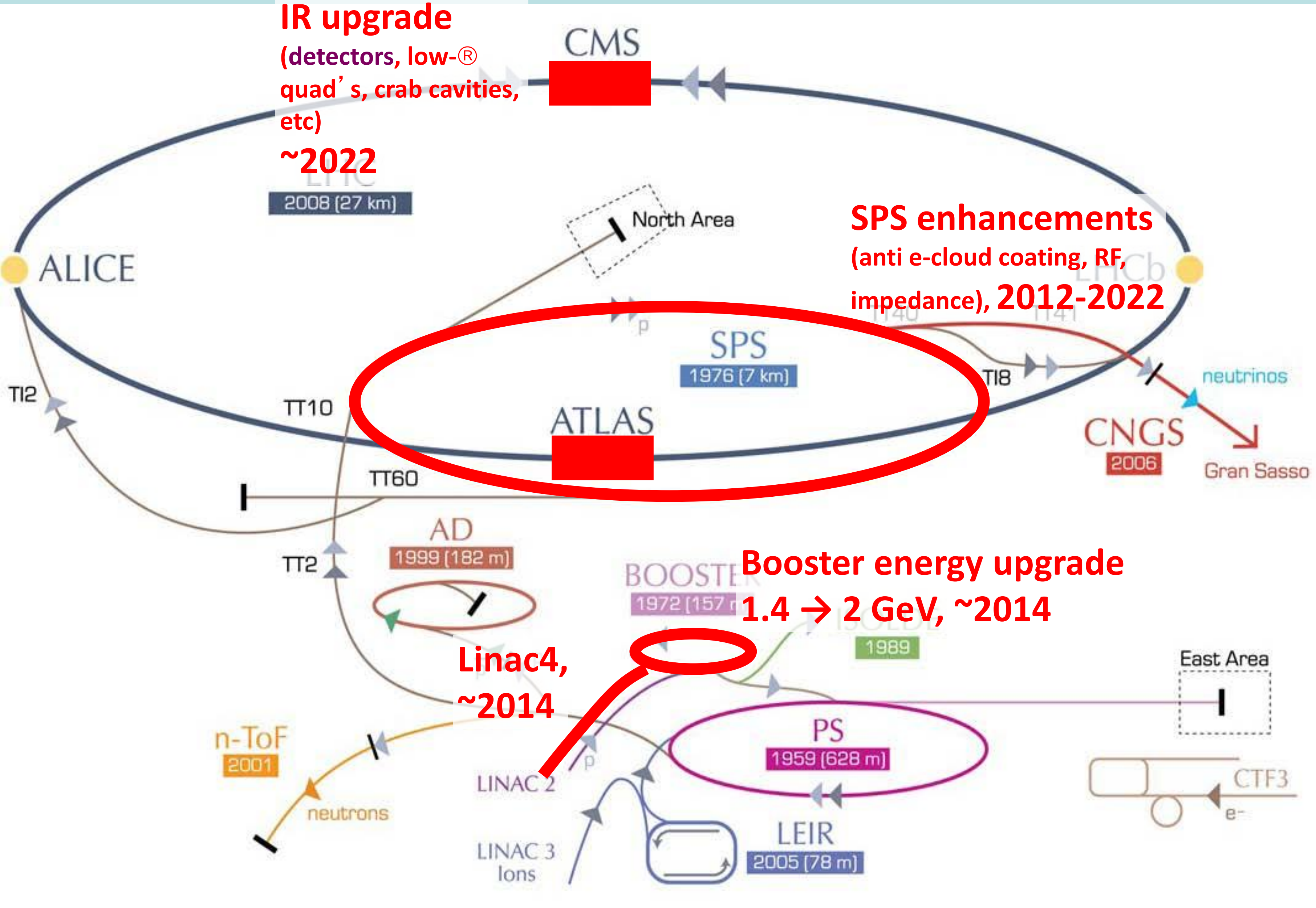
HL-LHC – accelerator modifications



HL-LHC – accelerator modifications



HL-LHC – accelerator modifications



Thanks to many people for the material !

including

Paul Collier, Vakho Tsulaia, Rolf Seuster, Christian Schmitt, Dario Barberis, Jim Shank, Janka Schovankova, Ikuo Ueda, Will Brooks, Petra Haefner, Daniel Froidevaux, Martin Wessels, Andy Salzburger

And thank you for listening!

გმადლობთ!

Luminosity of a hadron collider

$$L = \frac{N^2 k_c f}{4\pi\sigma_x\sigma_y} F = \frac{N^2 k_c f_0 \gamma}{4\pi\,\varepsilon_n \beta^*} F(\theta_c)$$

Hour glass factor: $F = 1 / \sqrt{1 + \left(\frac{\theta_c \sigma_z}{2\sigma^*}\right)^2}$

- Parameters in luminosity
  - No. of particles per bunch — N
  - No. of bunches per beam — k_b
  - No. of bunches collliding at IP ($k_c < k_b$) — k_c
  - Relativistic factor — γ
  - Normalised emittance — ε_n
  - Beta function at the IP — β^*
  - Crossing angle factor — F
    - Full crossing angle — θ_c
    - Bunch length — σ_z
    - Transverse beam size at the IP — σ^*

Equal amplitude functions:

$\beta_x^* = \beta_y^* = \beta^*$,

Geometric and normalised emittance:

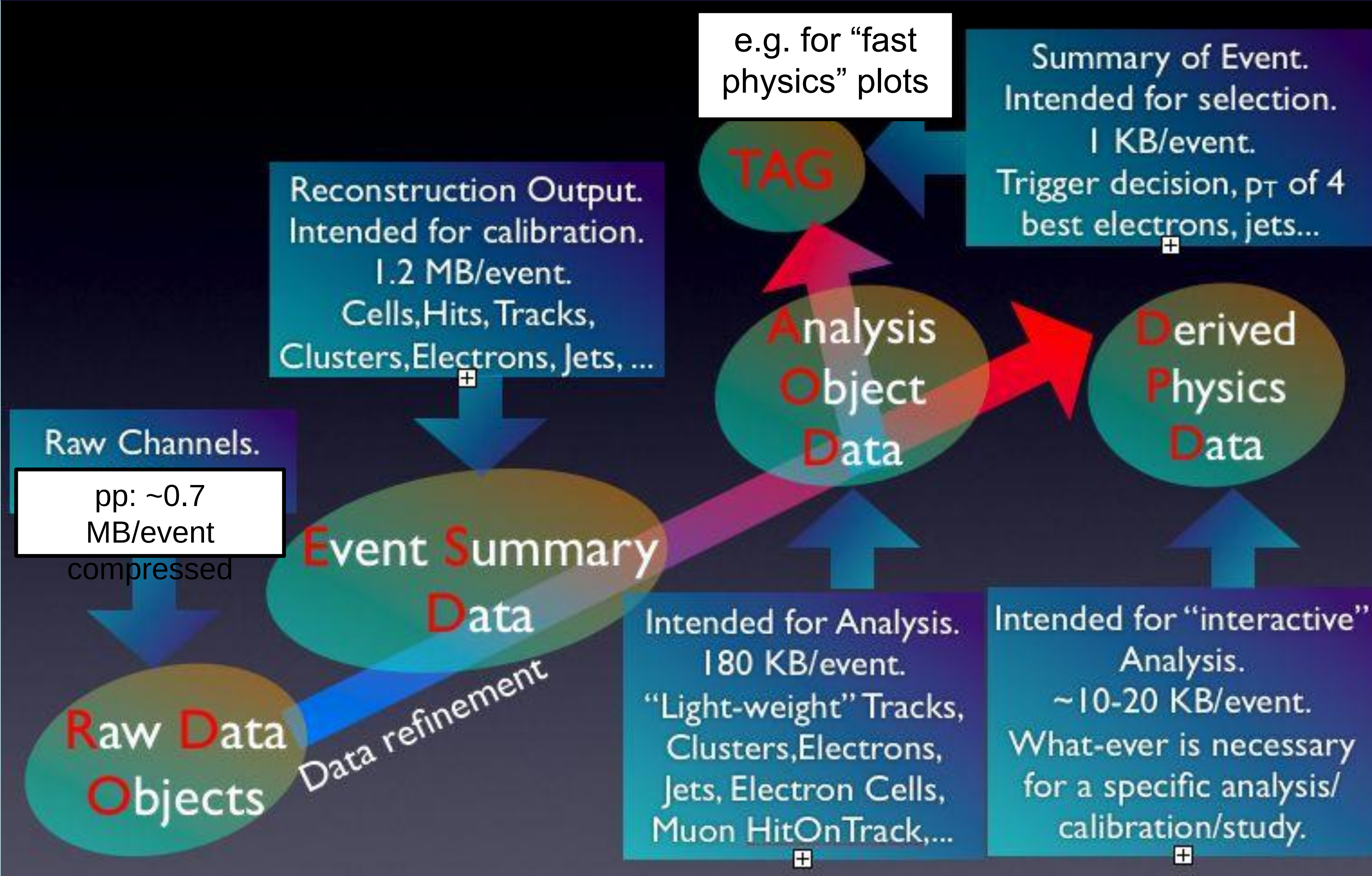
$$\varepsilon_x^* = \varepsilon_y^* = \varepsilon^* = \frac{\varepsilon_n}{\sqrt{\gamma^2 - 1}}$$

$\Rightarrow$ Round beams at IP:

$$\sigma_x^* = \sigma_y^* = \sigma^* \quad \sqrt{\frac{\beta^* \varepsilon_n}{\gamma}}$$

(N.B. LHC uses RMS emittances.)

Event Data Model – RAW, ESD, AOD, DPD, TAG data



Tier-0 Monitoring

Run Number	Project Name	Stream Name	Tag	Total # of Events	Avg. Evt. Rec. T. [s]	Input RAW Size [GB]	ESD Size [GB]	AOD Size [GB]	DESD Size [GB]	NTUP Size [GB]
186493	data11_7TeV	express_express	f393	207 314	16,2	133,4	392,6	47,3	-	141,3
186493	data11_7TeV	physics_Background	f393	98 718	8,5	49,0	57,8	5,3	-	0,7
186493	data11_7TeV	physics_CosmicCalo	f393	114 330	8,3	74,6	46,7	2,9	-	0,7
186493	data11_7TeV	physics_Egamma	f393	2 380 667	14,2	1 545,0	3 244,1	474,1	403,2	76,6
186493	data11_7TeV	physics_JetTauEtmiss	f393	2 974 572	16,0	1 978,6	4 320,0	693,3	381,0	66,3
186493	data11_7TeV	physics_MinBias	f393	216 227	11,1	130,1	225,1	26,7	41,7	1,7
186493	data11_7TeV	physics_Muons	f393	2 305 270	14,2	1 529,6	3 282,2	504,9	359,0	381,8
186493	data11_7TeV	physics_ZeroBias	f393	22 767	17,8	41,1	25,3	3,3	-	-

T2Ds

Tier-2 sites that can directly transfer any size of files with the other T1s than the one they are associated to

- First list defined in March
  - T1SC Mar 17 (http://indico.cern.ch/contributionDisplay.py?contribId=1&confId=131670)
- Revised in August
  - T1SC Sep 01 (http://indico.cern.ch/contributionDisplay.py?contribId=10&confId=153062)

T1s have been requested to configure the FTS channels accordingly

- https://twiki.cern.ch/twiki/bin/view/Atlas/DDMOperationsFTS#T2Ds_channels

T2Ds are candidates for

- multi-cloud production sites
- primary replica repository sites

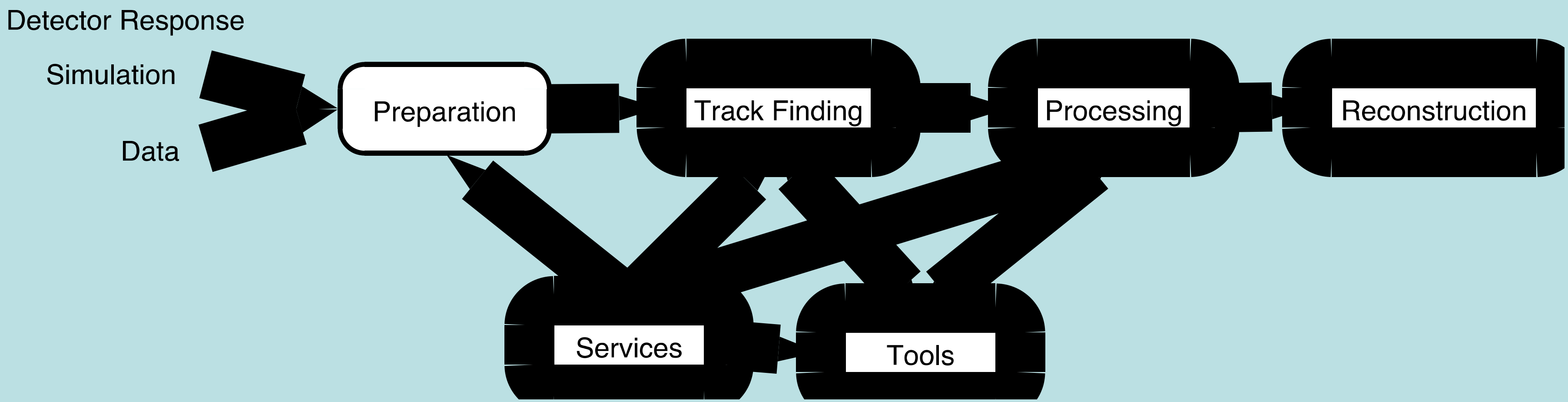
LHCONE

- Sites joining LHCONE may change their network connectivity
- Sites are requested to pre-notify ATLAS before joining LHCONE
  - so that we can do test transfers to compare the situation before and after

Specific software example

Taken from the paper *ATLAS Tracking Event Data Model*
(e.g. http://www.osti.gov/bridge/purl.cover.jsp?purl=/946305-MSIGCq)

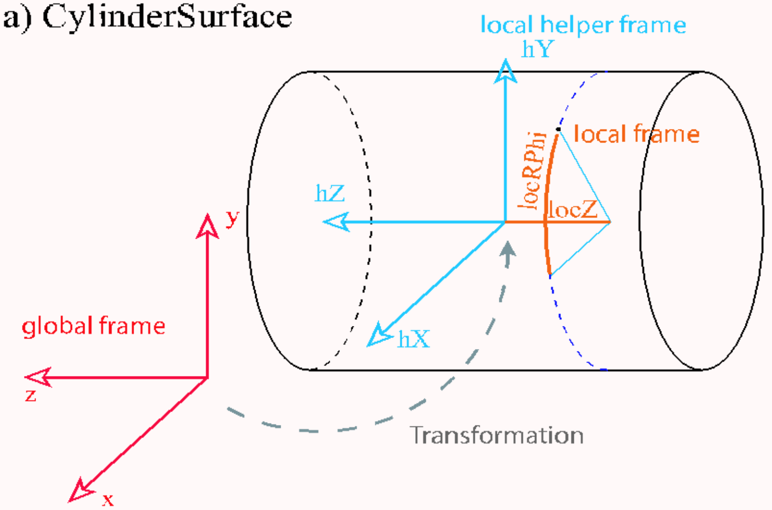
Stages involved in track reconstruction (simplified):



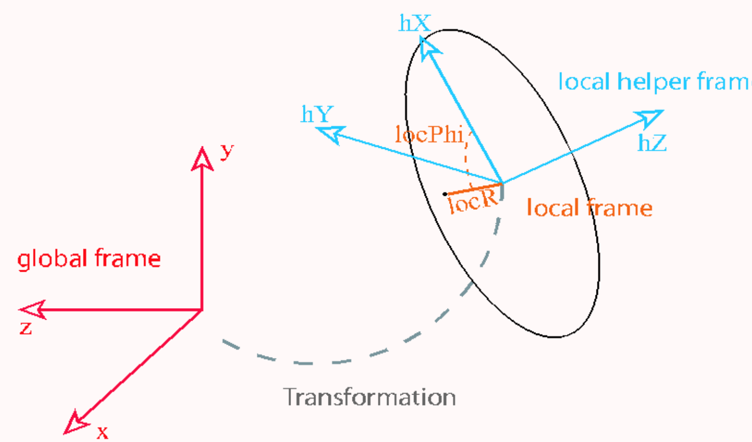
The detector information coming from either simulation or real data is prepared for the reconstruction using common classes. The track finding and the subsequent handling of the tracks can then use common services and tools due to the mutual interfaces.

Specific software example (2)

a) CylinderSurface

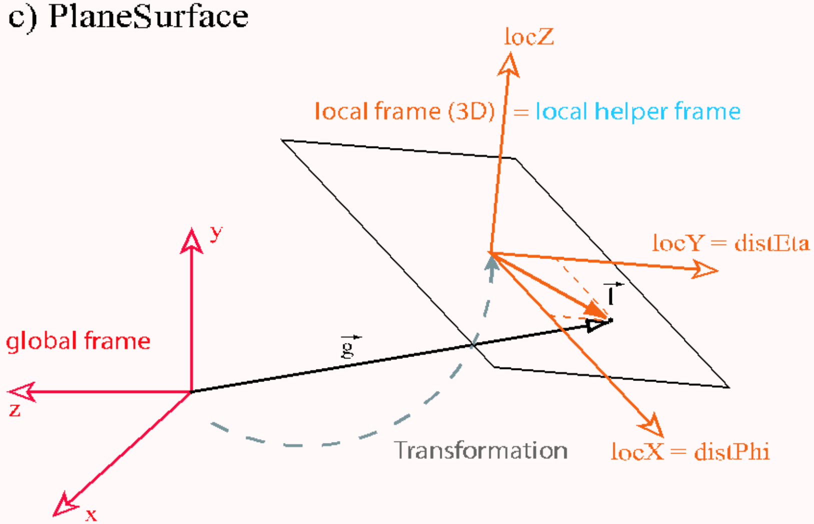


b) DiscSurface

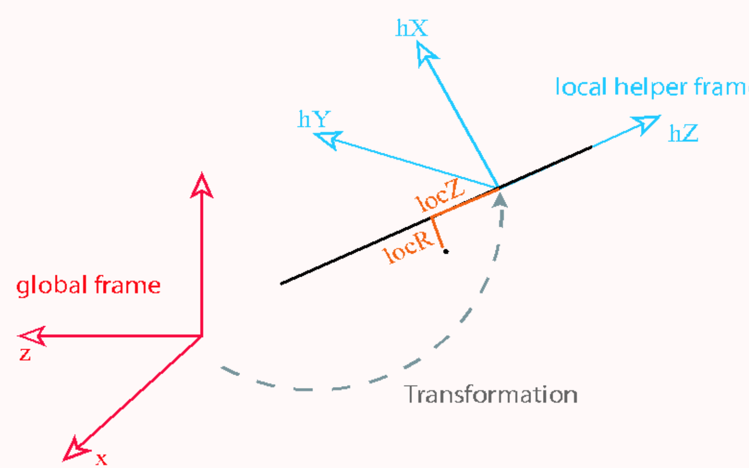


Surface types and their global to local transformations, as used in tracking

c) PlaneSurface



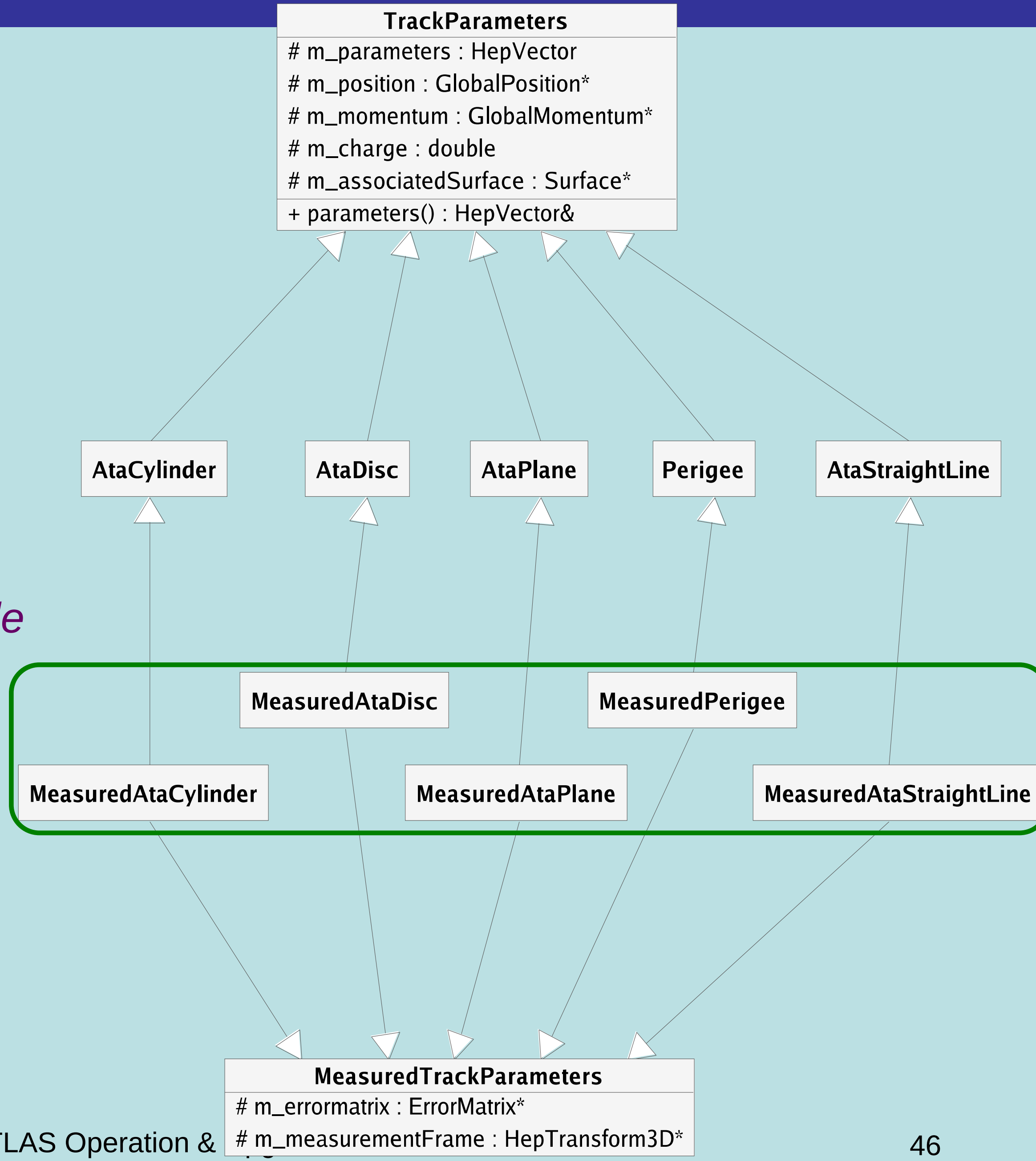
d) StraightLineSurface



Specific software example (3)

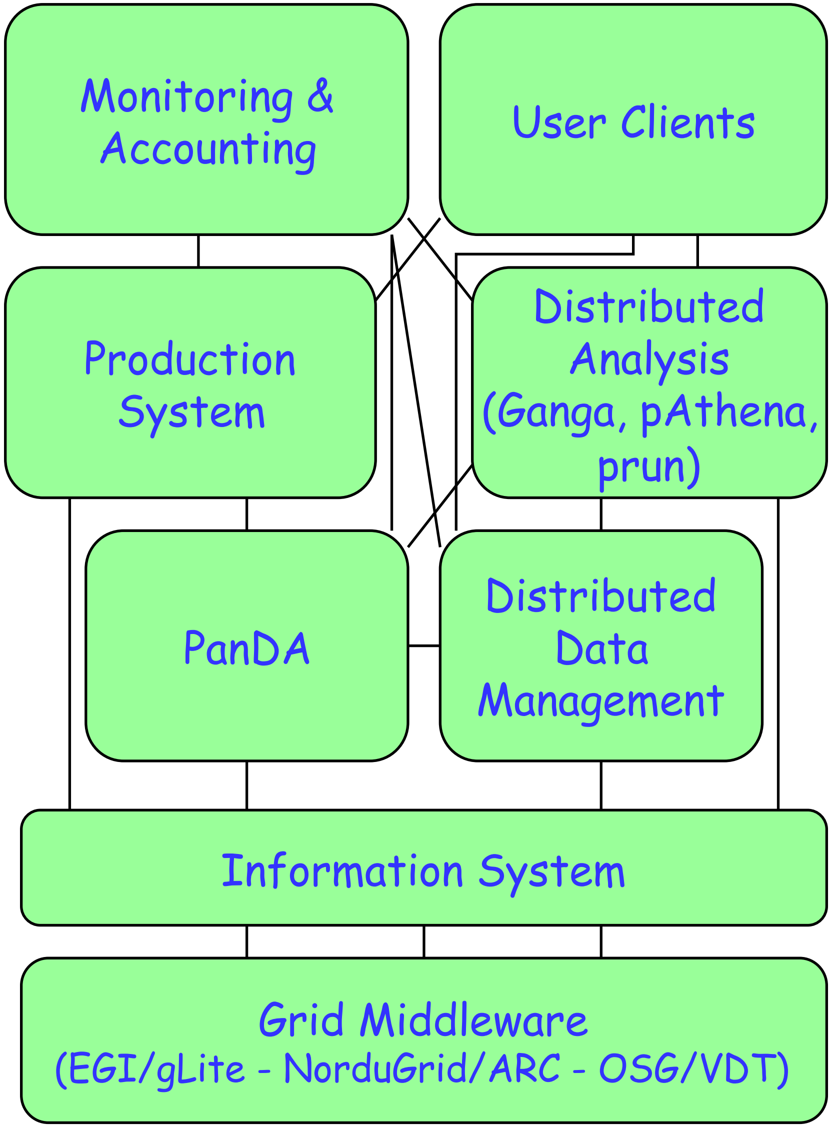
Class structure of track parameters

Track parameters exist in an unmeasured and a measured flavour. The measured track parameter classes follow a double inheritance structure, inheriting from the unmeasured class they represent and a common base class for measured track parameters, holding the error matrix description and the measurement frame definition.

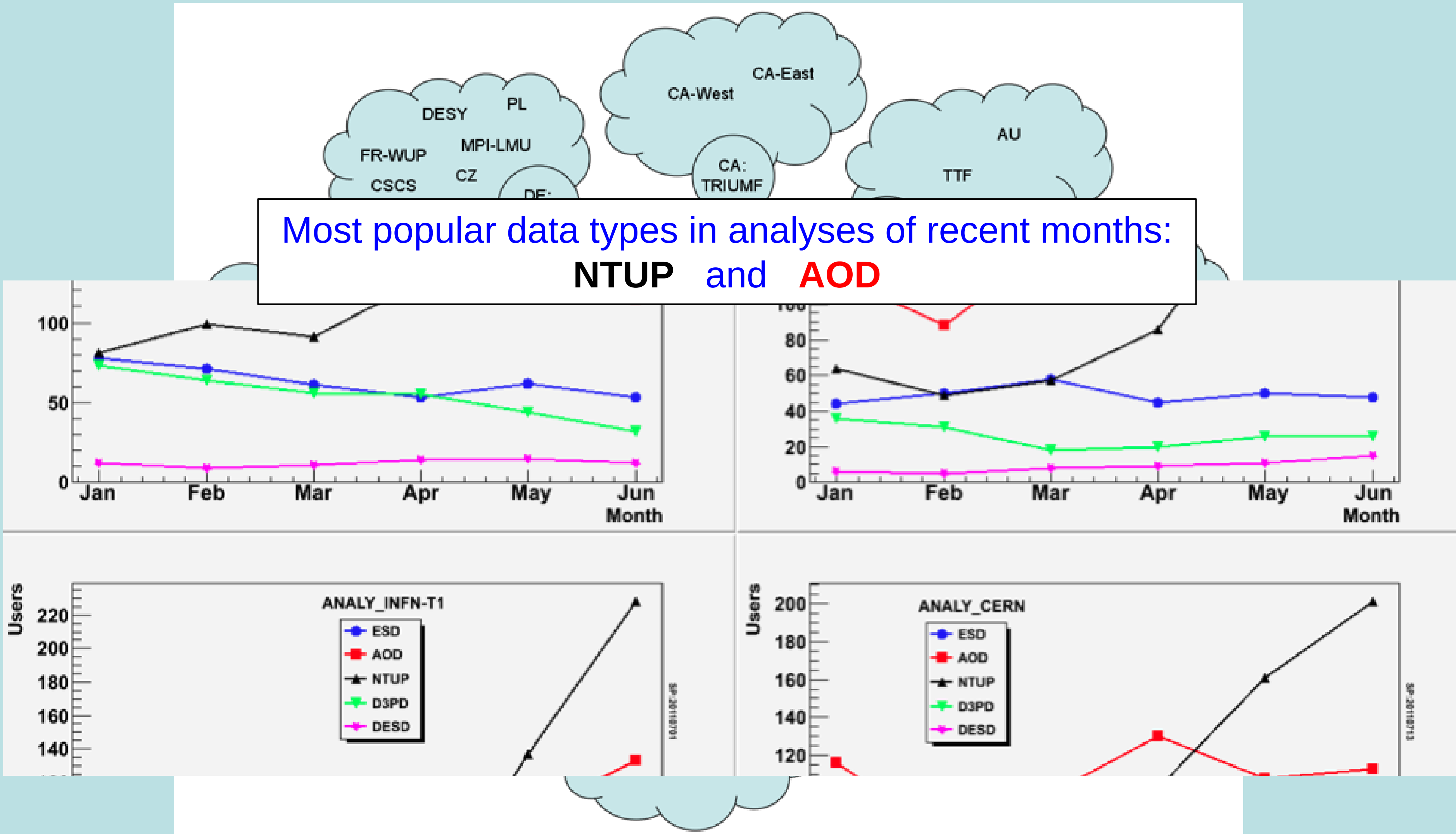


ATLAS Grid architecture

- ATLAS runs on 3 middleware suites:
  - gLite in most of Europe and several other countries (including all A-P countries)
  - ARC in Scandinavia and a few other small European countries
  - VDT in the USA
- ATLAS Grid tools interface with the middleware and shield the users from it
  - They also add a lot of functionality that is ATLAS specific
- The ATLAS Grid architecture is based on few main components:
  - Information system
  - Distributed data management (DDM)
  - Distributed production and analysis job management system (PanDA)
  - Distributed production (ProdSys) and analysis (Ganga/pAthena/prun) interfaces
  - Monitoring and Accounting tools
- DDM is the central link between all components
  - As data access is needed for any processing and analysis step!



Mitigating the old Cloud structure - more flexibility



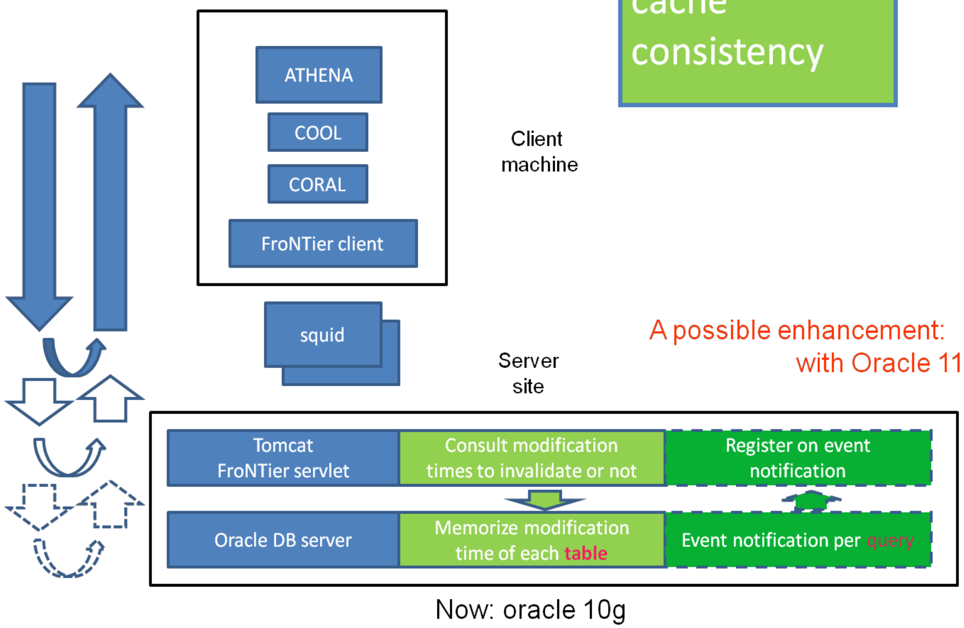
Most popular data types in analyses of recent months:
NTUP and **AOD**

Originally, we used a hierarchical model with strict “cloud” boundaries
Now, bandwidth allows inter-cloud T1 – T2 and T2 – T2 traffic of data and jobs
Data are now placed and deleted dynamically based on “popularity” - no fixed #copies

Conditions databases on the Grid

- Frontier deployed in 2009 to enable distributed access to the conditions DB
- Flow of database data:
  - Oracle: CERN online -> CERN offline -> 3D (BNL, TRIUMF, RAL, KIT, IN2P3-CC)
  - Frontier server at each of the above sites connects to local Oracle database
  - Local Squid contacts nearest Frontier server
    - ➢ With failover to next-to-nearest

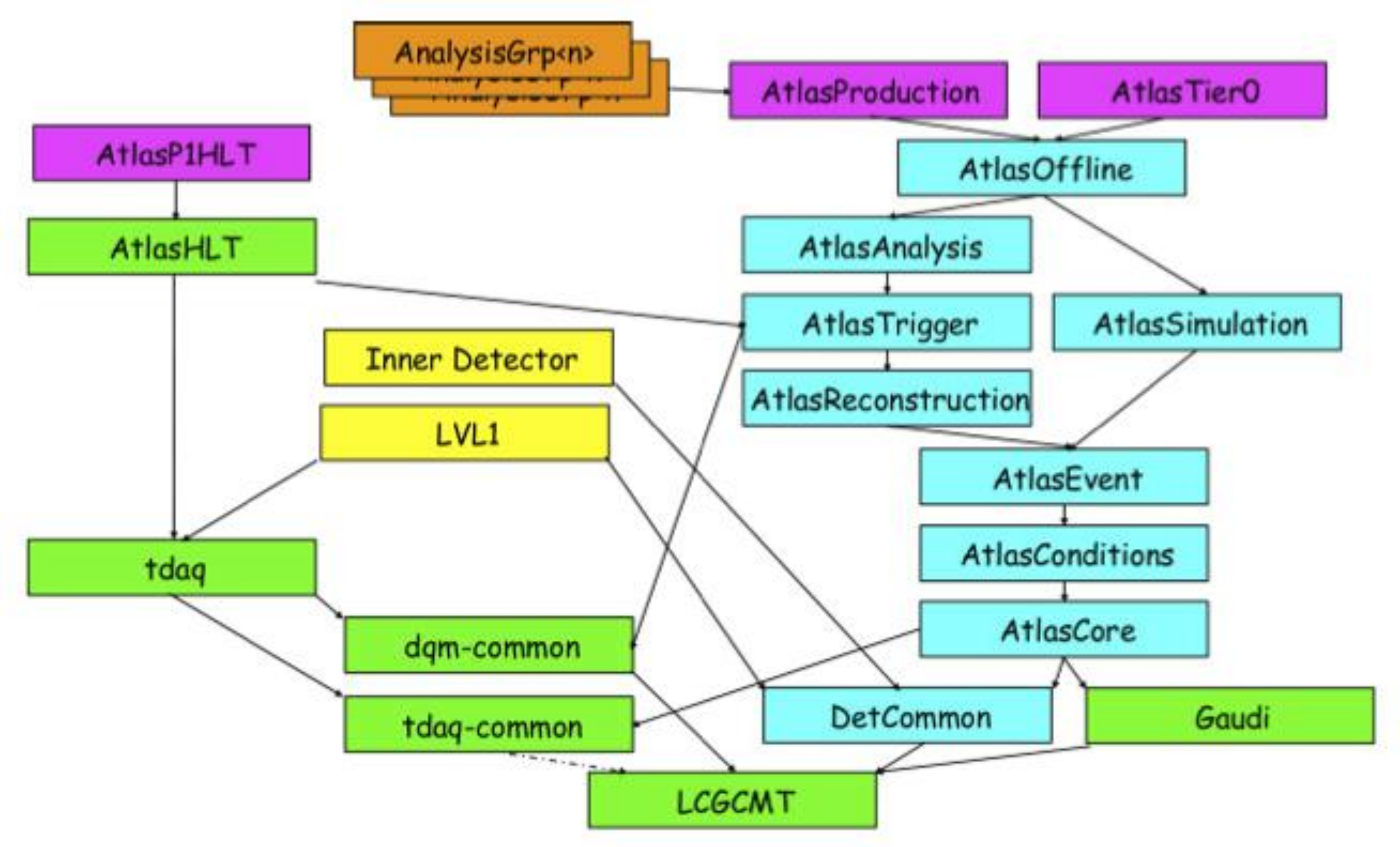
Handling cache consistency



Map of installed Squids

- Frontier reduces considerably the access time to DB data from remote sites
- It is particularly important for sites with low bandwidth and high latency towards Oracle servers

ATLAS software “projects”



ATLAS software in numbers

- ATLAS offline software is called “Athena”
  - Algorithms are used also in High-Level Trigger, under a different framework
- 2000 packages
  - sorted in several “projects” for unidirectional dependency
- 4 Million lines C++, 1.4M Python, 100k Fortran, 100k Java, ...
- 1000 developers have committed software to the offline repository in the last 3 years
- 300 developers have requested 4000 package changes in first half 2011 (25 per day)
  - It never stops: data taking, reprocessing, conferences
- 3000 users have a Grid certificate in atlas VO (able to submit job, retrieve data)