


Computing commissioning plans for DC14 and ADC perspectives for Run2

Simone Campana CERN-IT/SDC
on behalf of ADC



Perspectives for Run-2

The Challenges of Run2

- LHC operation
 - Trigger rate 1 kHz (~400)
 - Pile-up up above 30 (~20)
 - 25 ns bunch spacing (~50)
 - Centre-of-mass energy $\times \sim 2$
- Different detector
- Constraints of 'flat budget'
 - Both for hardware and for operation and development
 - Data from Run1

Where to optimize?

- Simulation

- CPU

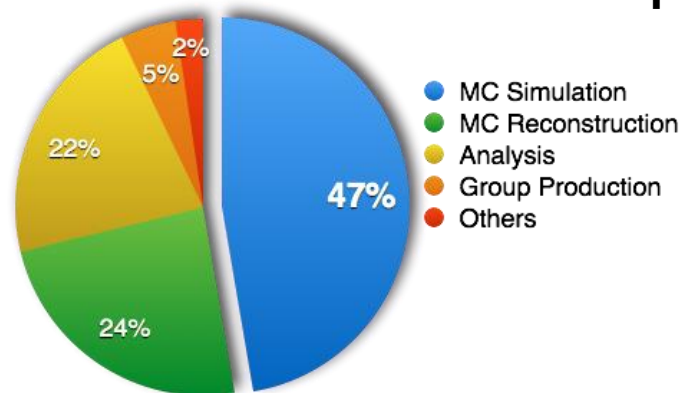
- Reconstruction

- CPU, Memory

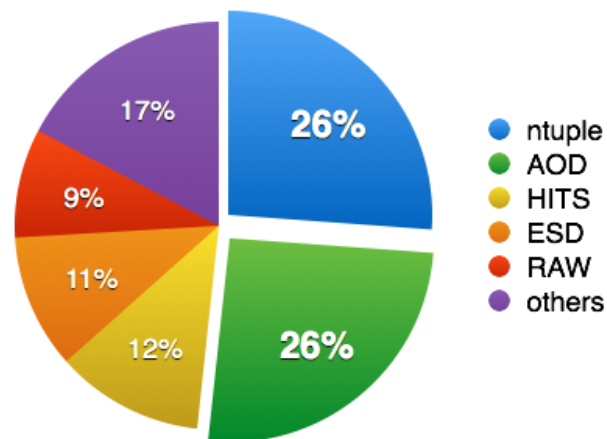
- Analysis

- CPU, Disk Space

CPU Consumption

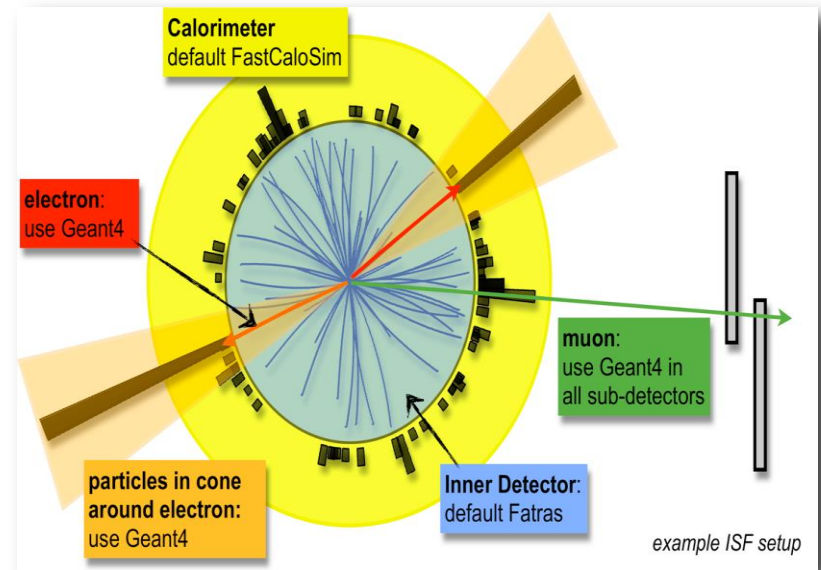


Disk usage at T1s & T2s



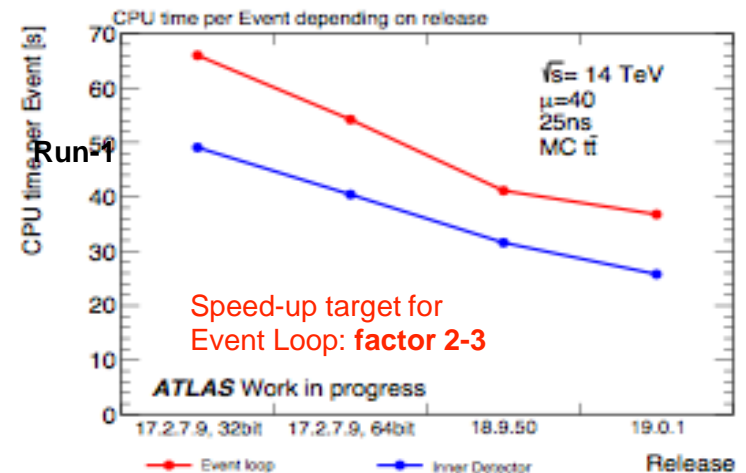
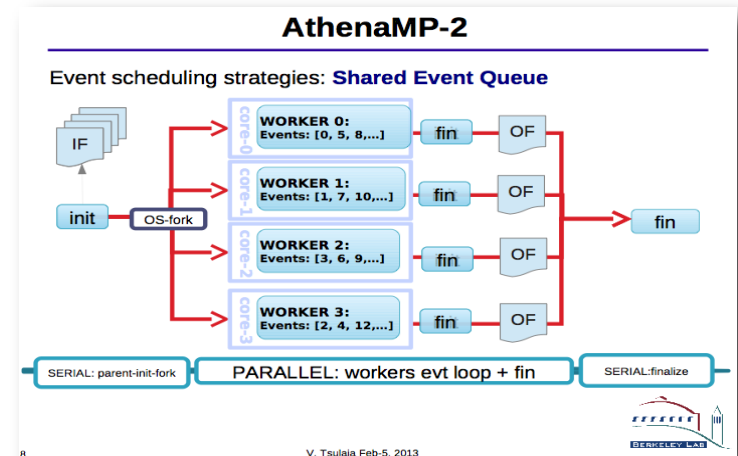
Simulation

- Simulation is CPU intensive
 - Integrated Simulation Framework (ISF)
 - Mixing of full GEANT & fast simulation within an event
 - Baseline for MC production
-
- More events per 12h job, larger output files, less transfers/merging, less I/O
 - Or shorter, more granular jobs for opportunistic resources



Reconstruction

- Reconstruction is memory eager
 - And requires non negligible CPU
- AthenaMP default from 2014
 - Keeps memory under control
 - Allows to run of native 64 bits (gain in CPU speed)
 - New hardware comes with less memory/core
- Optimization in code and algorithms

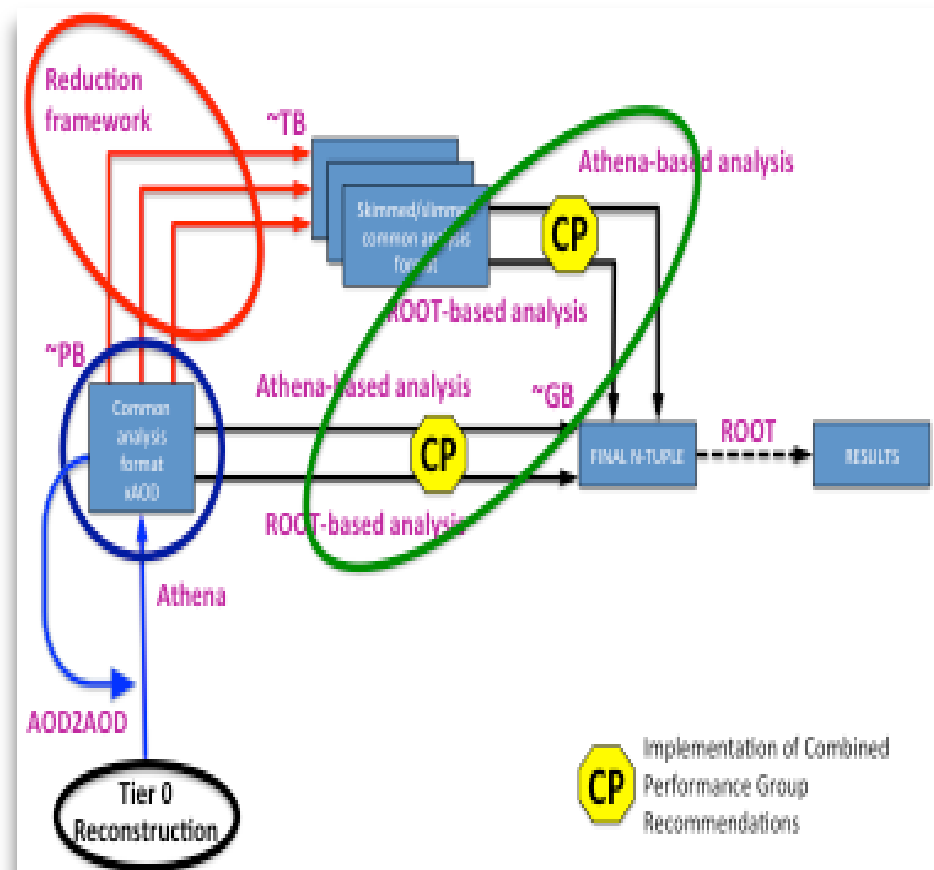


Running AthenaMP on the grid

- MultiCore resources scheduling is not an easy tasks for sites
 - Statically allocating resources for multicore jobs is not what sites want
 - To limit inefficiencies, dynamic allocation needs
 - a steady flow of long multicore jobs
 - a steady flow of short single core jobs
- Target would be
 - (Almost) all production run on multicore
 - Analysis on single core
- Need to work in this direction and progressively involve more sites
 - ADC should not expect sites to deploy multicores on demand in 1 week
 - Sites as well ...

Analysis Model

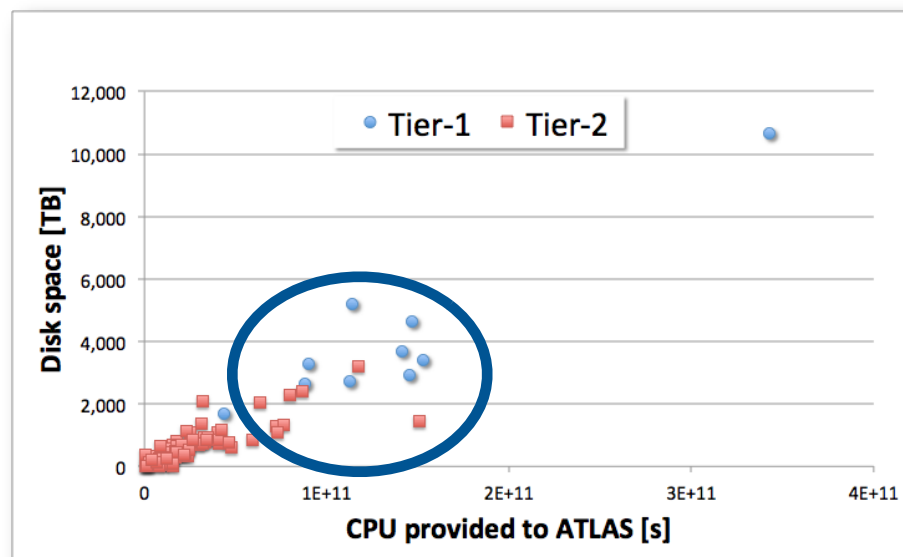
- Common analysis data format: xAOD
 - replacement of AOD & group ntuple of any kind
 - Readable both by Athena & ROOT
- Data reduction framework
 - AthenaMP to produce group data sample
 - Centrally via Prodsys
 - Based on train model
 - one input, N outputs
 - from PB to TB



Computing Model: data processing

- Optional extension of first pass processing from T0 to T1s in case of resource shortage at T0
- T1s and some T2s used for the most demanding workflows : high memory and I/O intensive tasks
- Data reprocessing & MC reconstruction also performed at some T2s
- Still one full reprocessing from RAW / year, but multiple AOD2AOD reprocessings/year
- Derivation Framework (train model) for analysis datasets

Need More Flexibility



Some **T2s** are equivalent to **T1s** in term of disk storage & CPU power

Computing Model: data placement

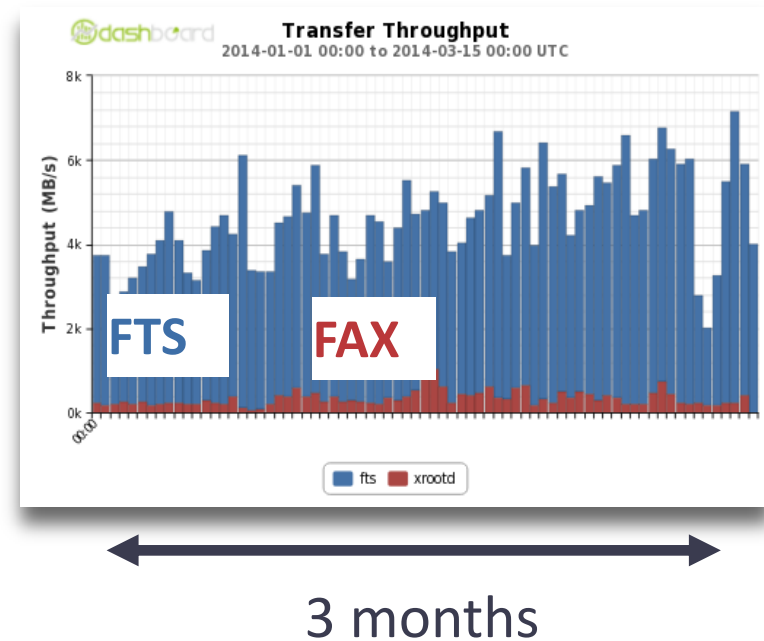
- End of Run-1: 2 copies of data (one at T1s, one at T2s)
- Run-2 : Non popular data will be archived to tape at T1s
- Higher demand on T1 tape systems
- Only popular dataset will be kept on disk
- User access to data on tape granted through centralized tools only

Distributed Data Management: Rucio

- Current DDM system delivered expected service (and beyond).
- However we will face new challenges after LS1:
 - New requirements to optimize space and network utilization are hardly compatible with current design
 - Multiple file/dataset ownership, rule based engine
 - Integrating “new” technologies hard as well
 - Protocols e.g. xroot/http; technologies e.g. NoSQL
 - Operational experience indicated some weak aspects of the system, difficult to cure
 - Proliferation of space tokens and fragmentation of space, overlapping datasets, ..
- Rather than insisting in patching current system, a more fundamental redesign was agreed
- Rucio commissioning will be carried over in the scope of DC14

Storage Federations/WAN data access

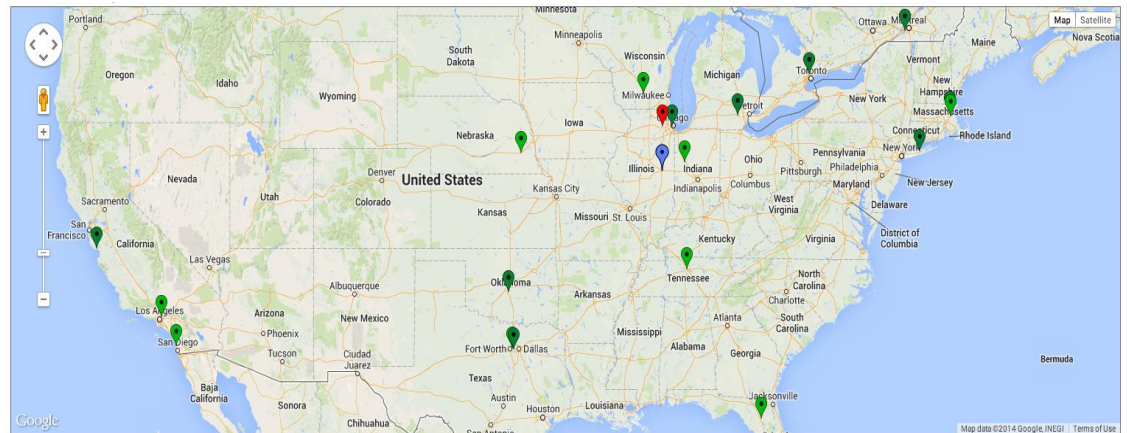
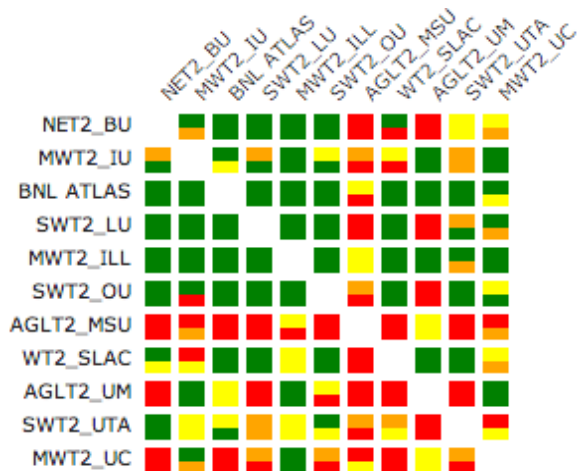
- Jobs access data on remote storage resources via WAN
 - More flexibility in using free CPU resources
 - Bandwidth and network stability required
- FAX (Federating ATLAS data stores using Xrootd)
 - Job fail-over in case of access failure activated
 - Future: generalized WAN access, throttled so that other activities/sites are not impacted
 - 10% of WAN accesses is the target/limit
 - All sites should join FAX.
- HTTP/WebDAV will be initially used for dq2-get use case
 - All sites should deploy WebDAV



Network Monitoring

- All sites are asked to install perfSONAR and configure it properly
- Situation in USATLAS is rather good

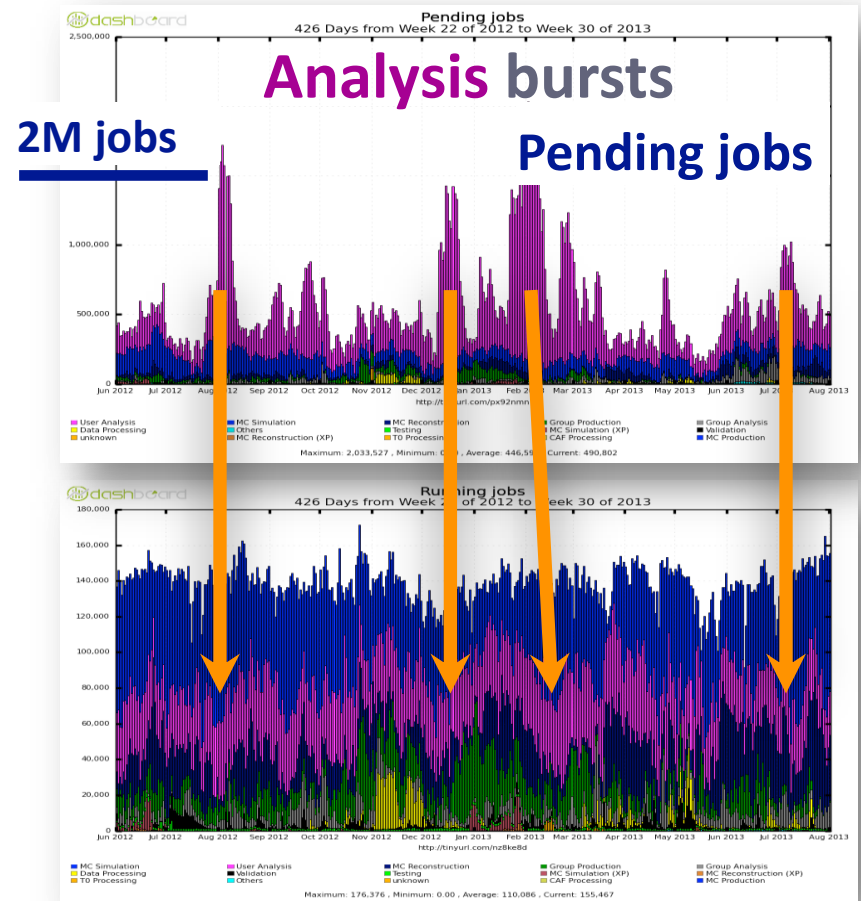
■ Throughput ≥ 900 Mbps
 ■ Throughput < 900 Mbps
 ■ Throughput ≤ 500 Mbps
 ■ Unable to retrieve data
 ■ Check has not yet run



■ Loss rate is ≤ 0
 ■ Loss rate is ≥ 0
 ■ Loss rate is ≥ 0.01
 ■ Unable to retrieve data
 ■ Check has not yet run

New production system: Prodsys2

- Same engine for analysis and production
 - Currently analysis vs production shares managed by sites not by ATLAS
 - Better reactivity to analysis load
- Minimized data traffic
 - Merging at T2s
- Optimized job to resource matching
 - for better use of computing resources

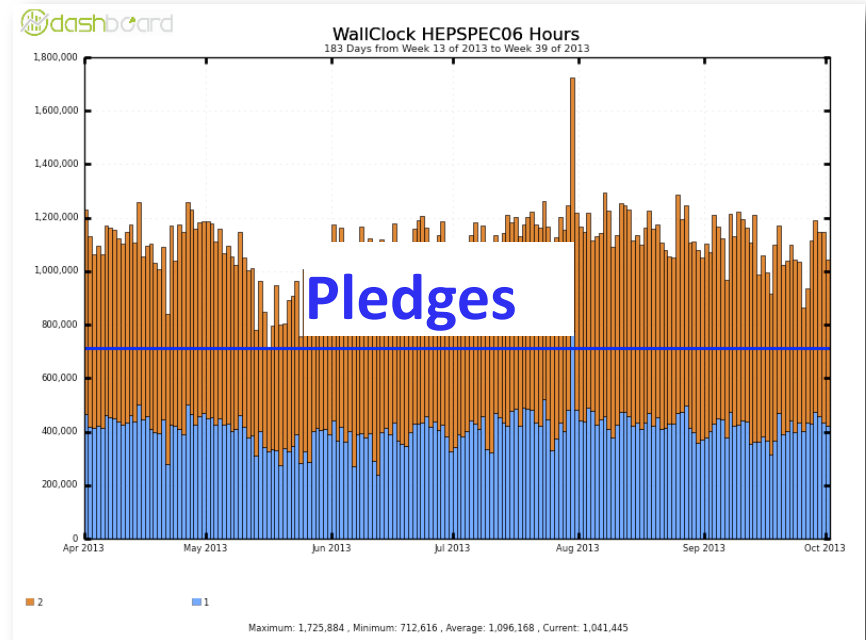


No increase of running jobs

Opportunistic resources

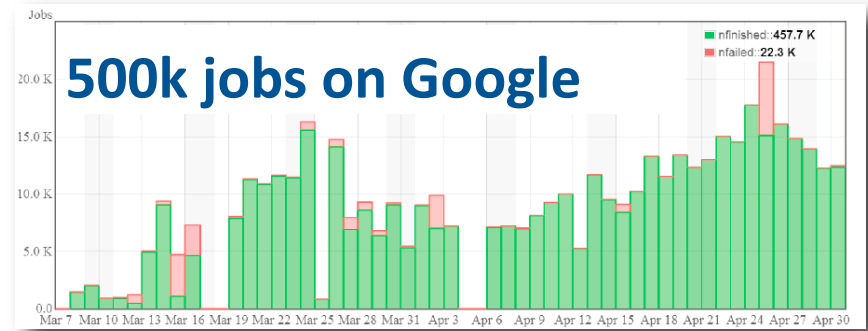
- In Run-1 lots of benefits from unpledged resources
 - Provided through Grid interfaces
- For Run-2, new interfaces and facilities will be available
 - HLT farm at P1
 - Cloud computing
 - HPC (High Performance Computing) centers
 - Volunteer computing: ATLAS@home, also useful for T3 sites

CPU consumption above pledges both at T1s and T2s



Cloud computing

- Cloud Computing R&D moved now to production
 - We have a cloud operations team
 - Resource provisioning relies heavily on Condor
- Plan for use of 'academic' clouds and opportunistic use of 'cheap' commercial is possible
 - Some cloud computing providers start to propose cost-competitive offers (with some limitations)
- Long jobs and large I/O are not well suitable in many cases



Opportunistic Resources: HPCs

- HPC offers important and necessary opportunities for HEP
 - Possibility to parasitically utilize empty cycles
- Bad news: very wide spectrum of site policies
 - No External connectivity
 - Small Disk size
 - No pre-installed Grid clients
 - One solution unlikely to fit all
- Good news: from code perspective, anything seriously tried so far did work
 - Geant4, ROOT, generators
- Short jobs preferable for backfilling



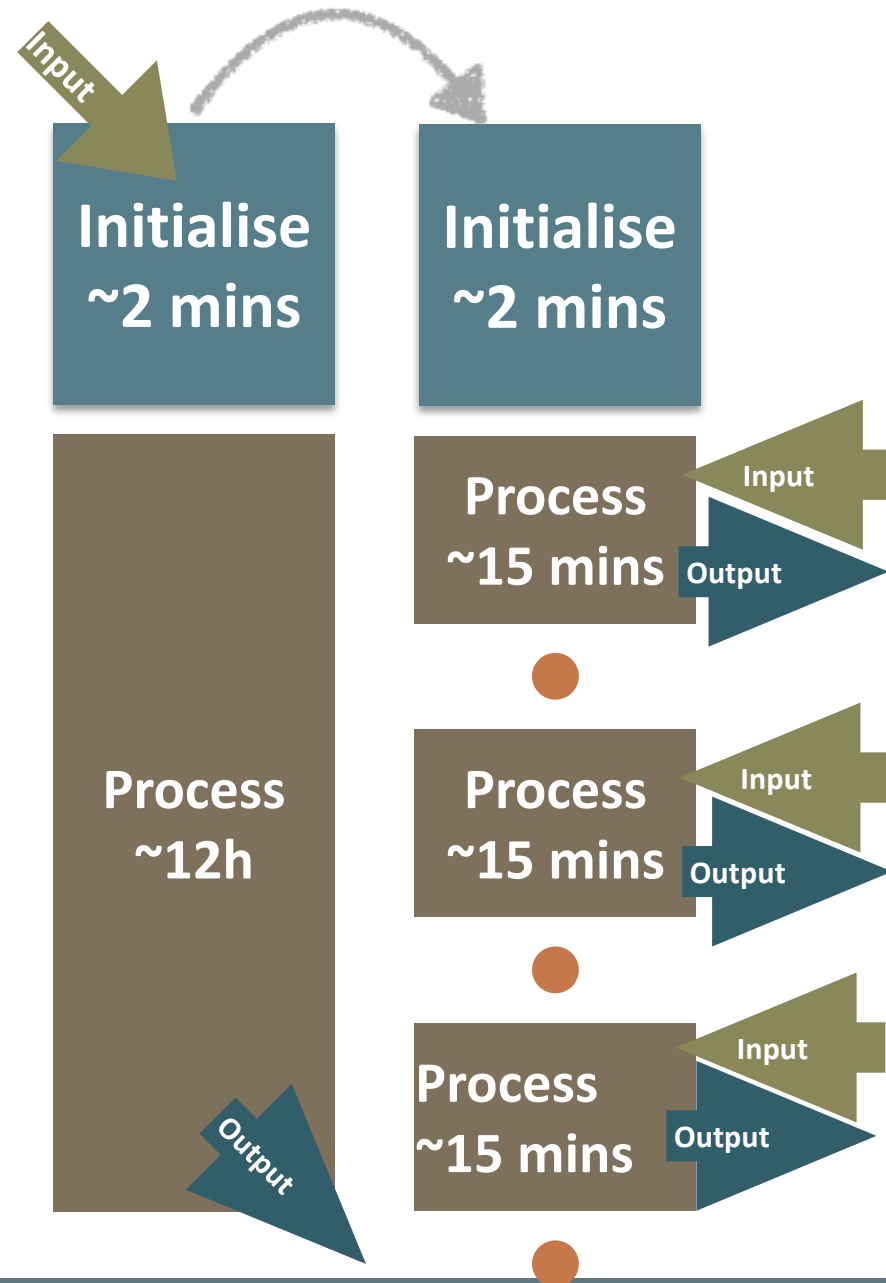
Oak Ridge Titan System

Architecture:	Cray XK7
Cabinets:	200
Total cores:	299,008 Opteron Cores
Memory/core:	2GB
Speed:	20+ PF
Square Footage	4,352 sq feet

HPC exploitation is now a coordinated ATLAS activity

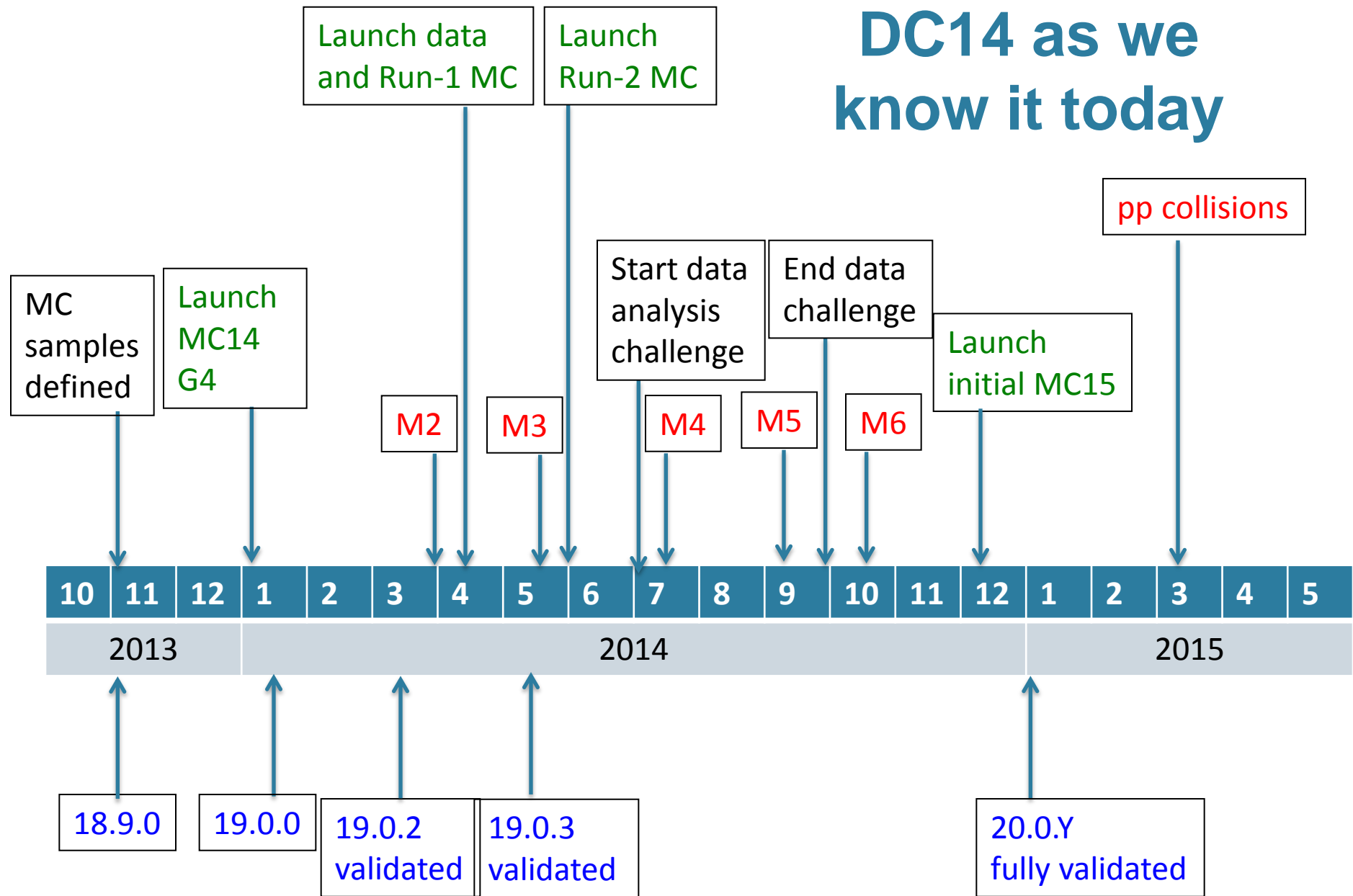
Event Service

- In development : software and distributed computing effort
- Feed VMs with short $O(15\text{min})$ work allocations, e.g. single simulation events
- Usages :
 - Backfilling of HPC centers
 - Opportunistic use of commercial clouds
 - Volunteer computing (ATLAS@home)

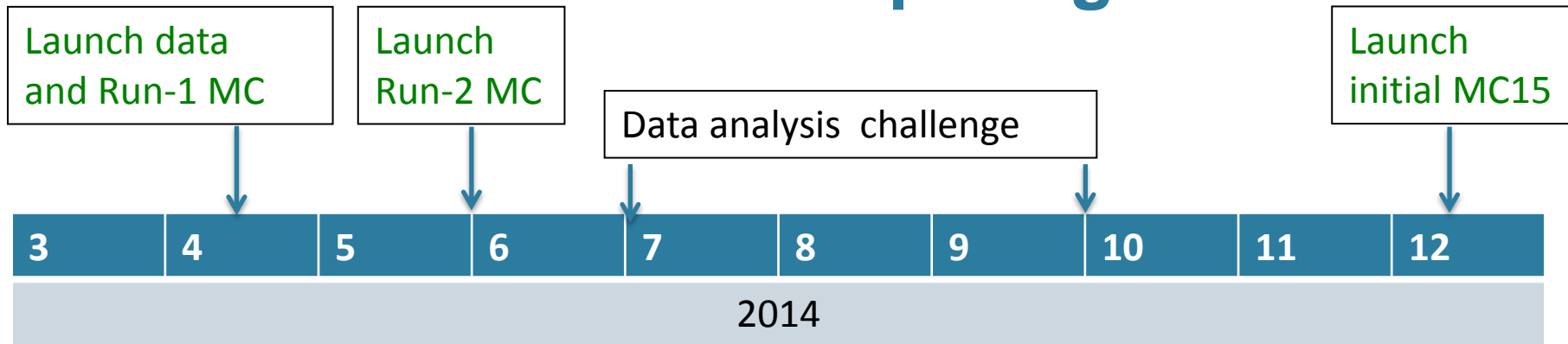


DC14 schedule and ATLAS Distributed Computing plans

DC14 as we know it today



DC14 for Distributed Computing

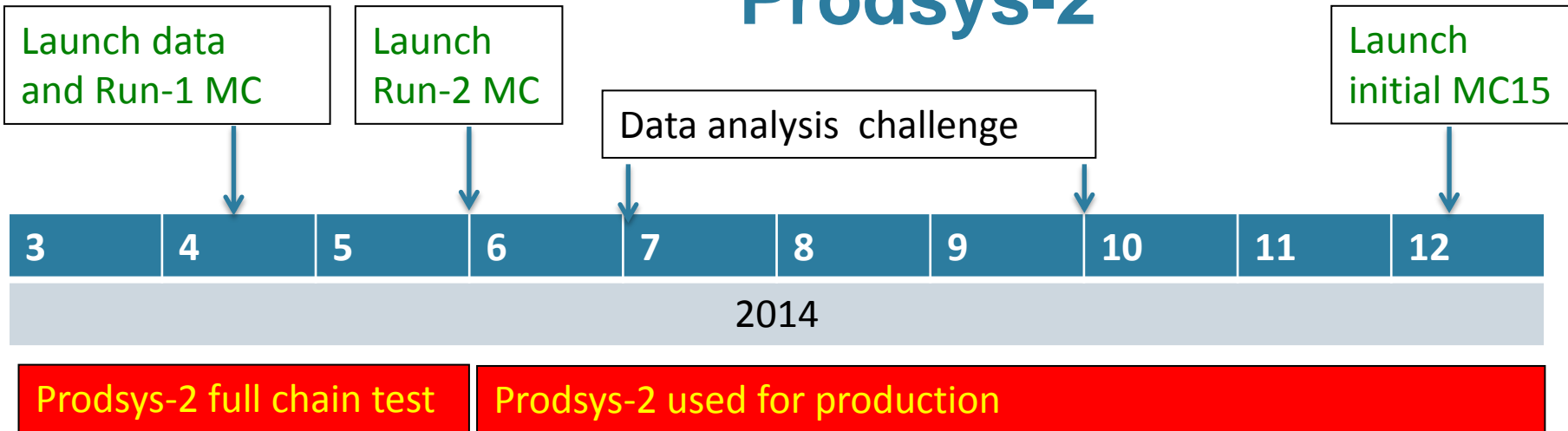


- ATLAS Distributed Computing commissioning decoupled from Software timeline
 - Schedule is tight for both, need to avoid delays
- We will try to perform as much as possible of the DC14 exercises with the new ADC components
- Main interested components:
 - Tier-0
 - Prodsys-2: new generation of ATLAS production system
 - Rucio: new generation of ATLAS Distributed Data Management system (replacing current DQ2)
 - Databases

Tier-0

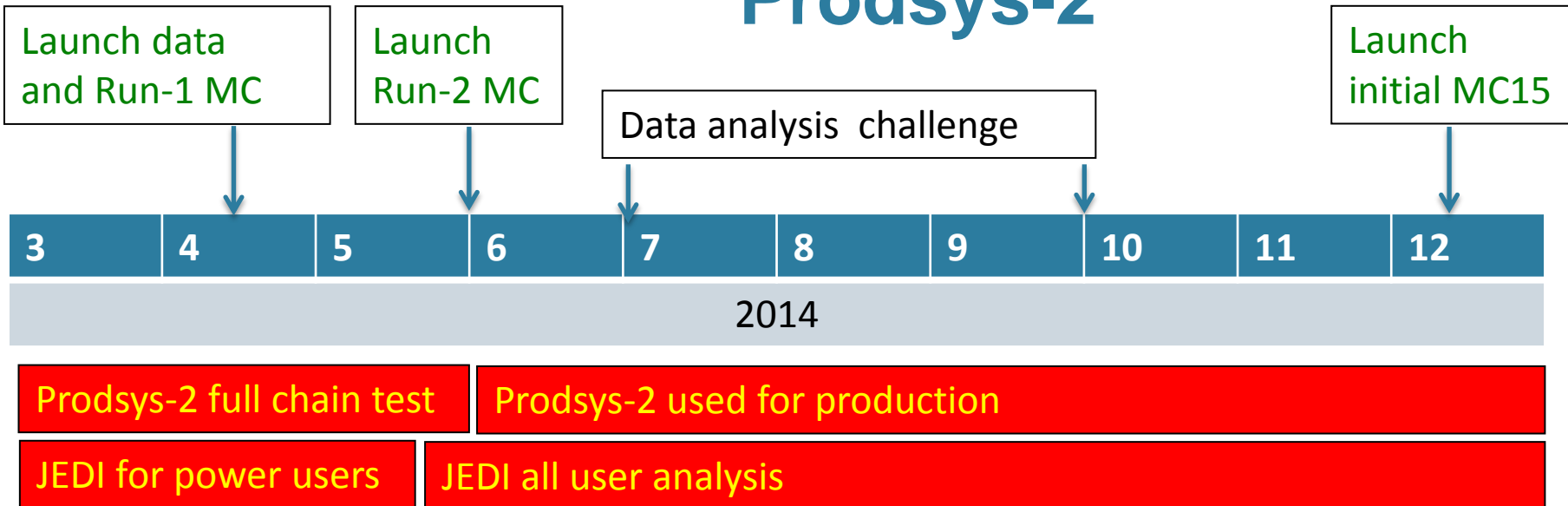
- Testing of shared vs dedicated LSF master
 - Ongoing, done by mid April – “shared” already done
- Testing of the new storage model
 - EOS “hot storage”, CASTOR archive-only, Distributed Data Management system (DDM) as transfer engine
 - Involvement of the ATLAS Online community
- Testing of new streams and workflows
 - E.g. the “fat stream”, xAOD production
- Spill-over to T1s needs some thinking

Prodsys-2



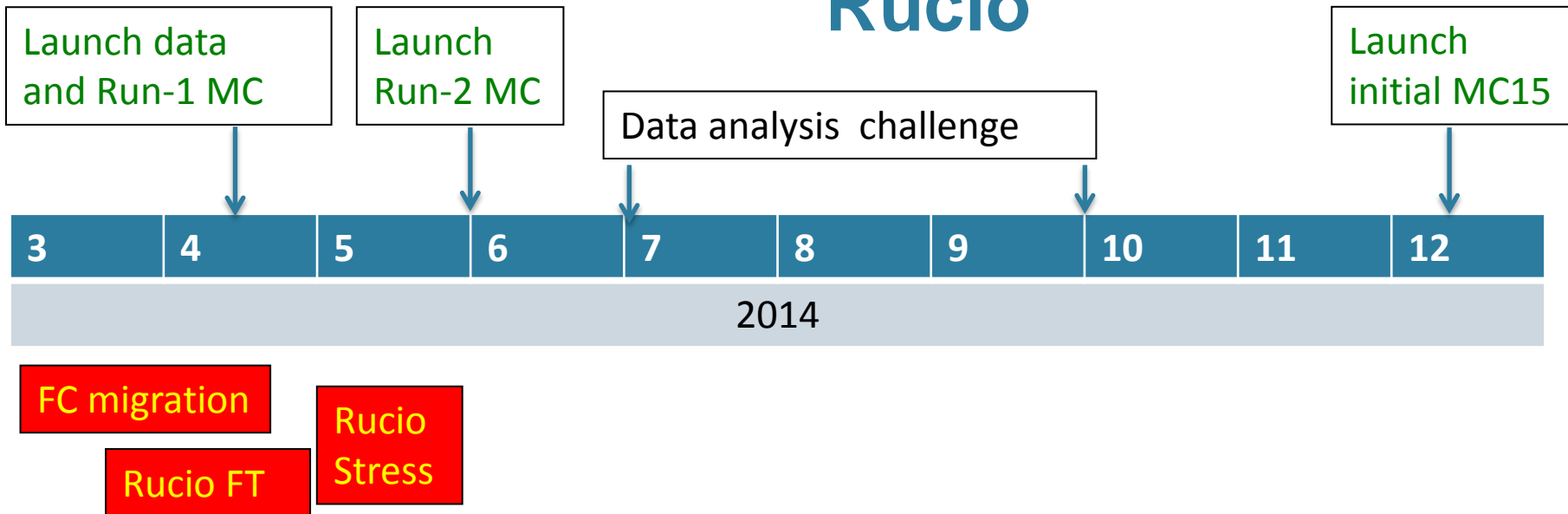
- Many new components
 - Request I/F: user interface for requesting a production workflow
 - DEfT: translates request in one or more chains of tasks
 - JEDI: generates job definition from task definition and injects them in PanDA
- Beta version of Request I/F+DEfT+JEDI (a.k.a. Prodsys-2) tested for full chain
 - Could be used for data and DC14 Run1 MC but we need 2 months to consolidate monitoring (June 1st).
 - Surely to be used for DC14 MC Run-2

Prodsys-2



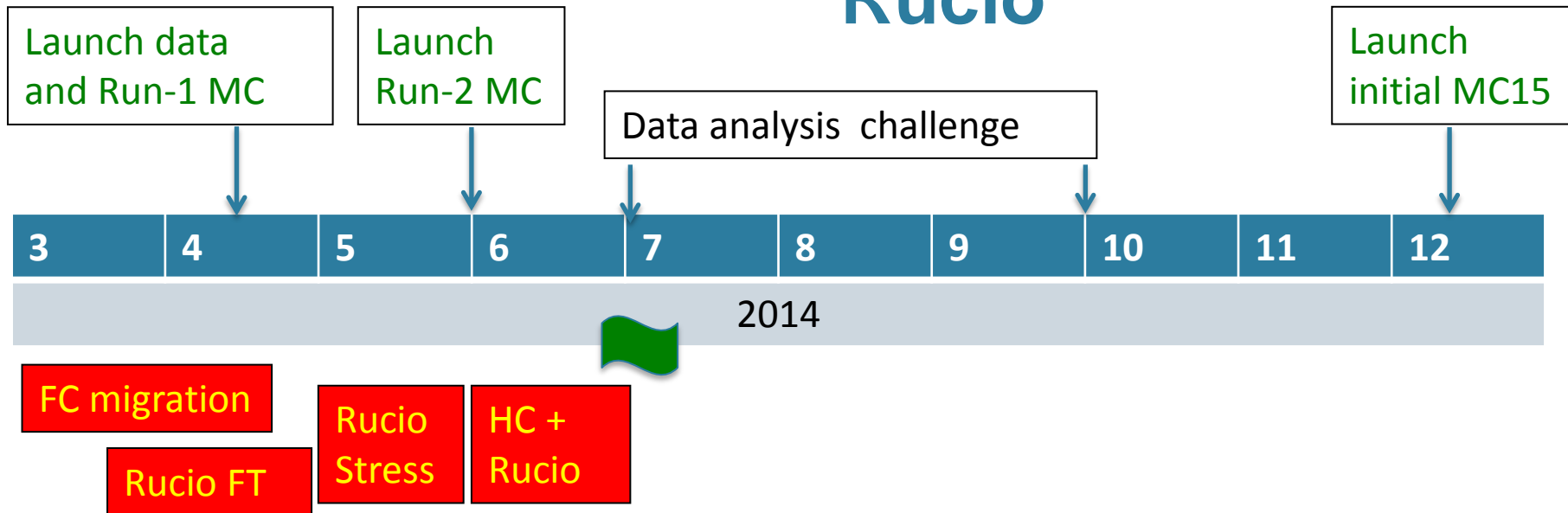
- JEDI will be used also for analysis
 - Will implement the concept of “analysis task”
- JEDI is ready for analysis
 - Used in Functional Tests, exposed to beta users (more users to come)
 - Surely to be used for DC14 Analysis Challenge

Rucio



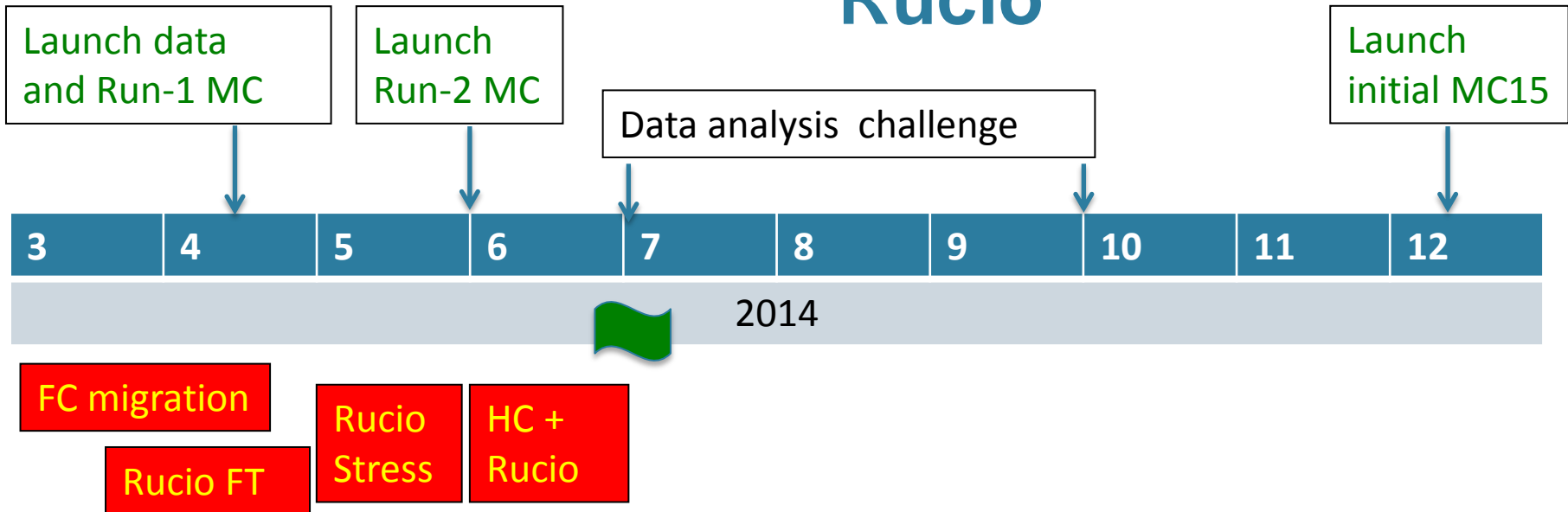
- FC migration: moving from current DQ2 file catalog (LFC) to Rucio file catalog
 - Site-by-site, ongoing. Done by Mid April
- Rucio Functional and Stress tests in April-May
 - Full chain of data export/distribution at nominal rate + deletion
 - Storage resource utilization needs to be planned

Rucio



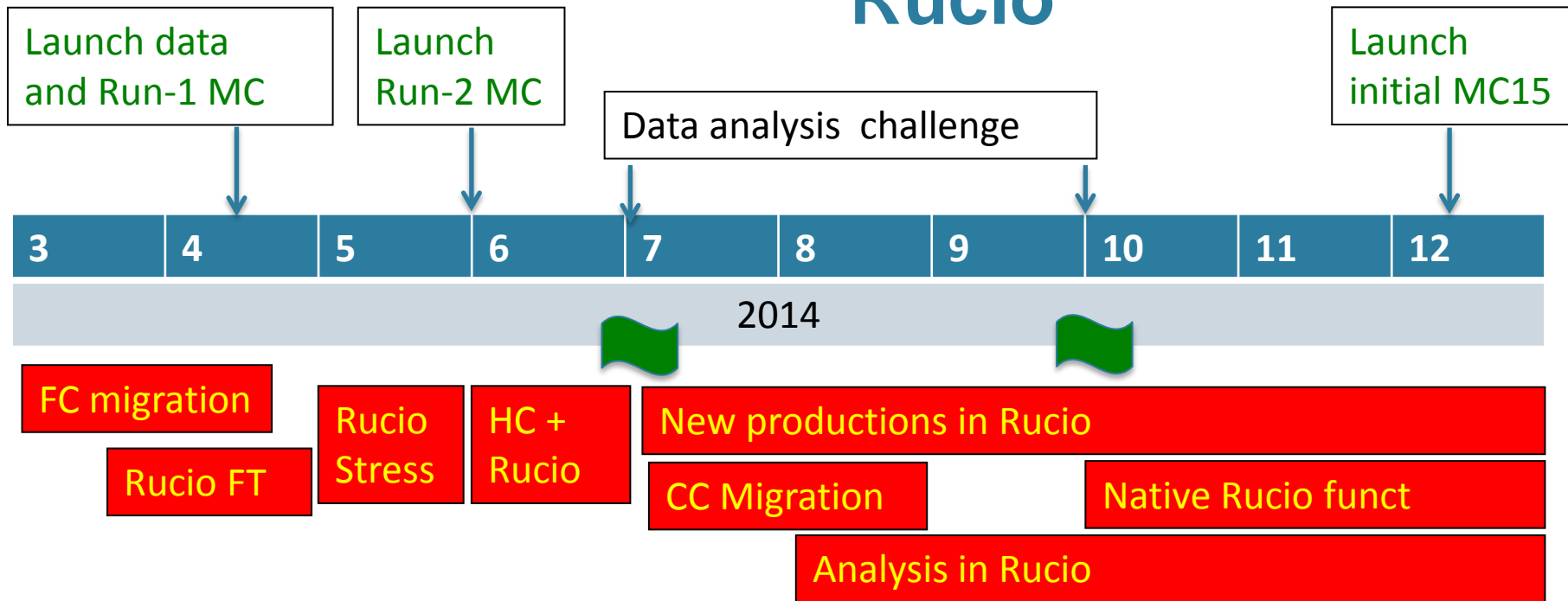
- **Test Production/Analysis against Rucio**
 - Move production and analysis HammerCloud tests to use Rucio
 - HammerCloud (HC) is our framework for analysis and production jobs functional and stress tests
- **If all tests are successful, Rucio can be considered commissioned**

Rucio



- Migrate production data from DQ2 Central Catalog to Rucio
 - Implies using the full Rucio machinery for subscriptions/transfers/deletion/location
 - Can not be done site by site, but dataset by dataset
 - DDM and Rucio Catalog can coexist with transparent client fallback
 - Rollback is possible

Rucio



- Progressively define new Monte Carlo production workflows in Rucio
- Migrate Central Catalog for existing data
 - Transparent client fallback
- Move analysis to Rucio
- Utilize native Rucio functionalities
 - This is the point of no return, decommissioning of DQ2

Databases

- Use of the new COOL instance CONDBR2 instead of COMP200
 - clean start for Run2
- No more DB releases for production
 - all DB access with Frontier
- Commissioning of the EventIndex infrastructure in the second half of 2014
 - a complete catalogue of all ATLAS events in any format
 - Lookup, skimming, completeness and consistency checks
- Access to Combined Performance calibration/alignment/efficiency data files from the new common repository

Conclusions

- We defined a commissioning work plan for ADC components in DC14
- Some DC14 exercises will be run with the current system, some others with the new system
- A shift in DC14 schedule would not be a problem for ADC. A compression of the DC14 schedule needs discussion.
- The ADC schedule is tight, a bit aggressive and non-compressible
- Manpower is a concern: lack of newcomers and maintaining the current expertise