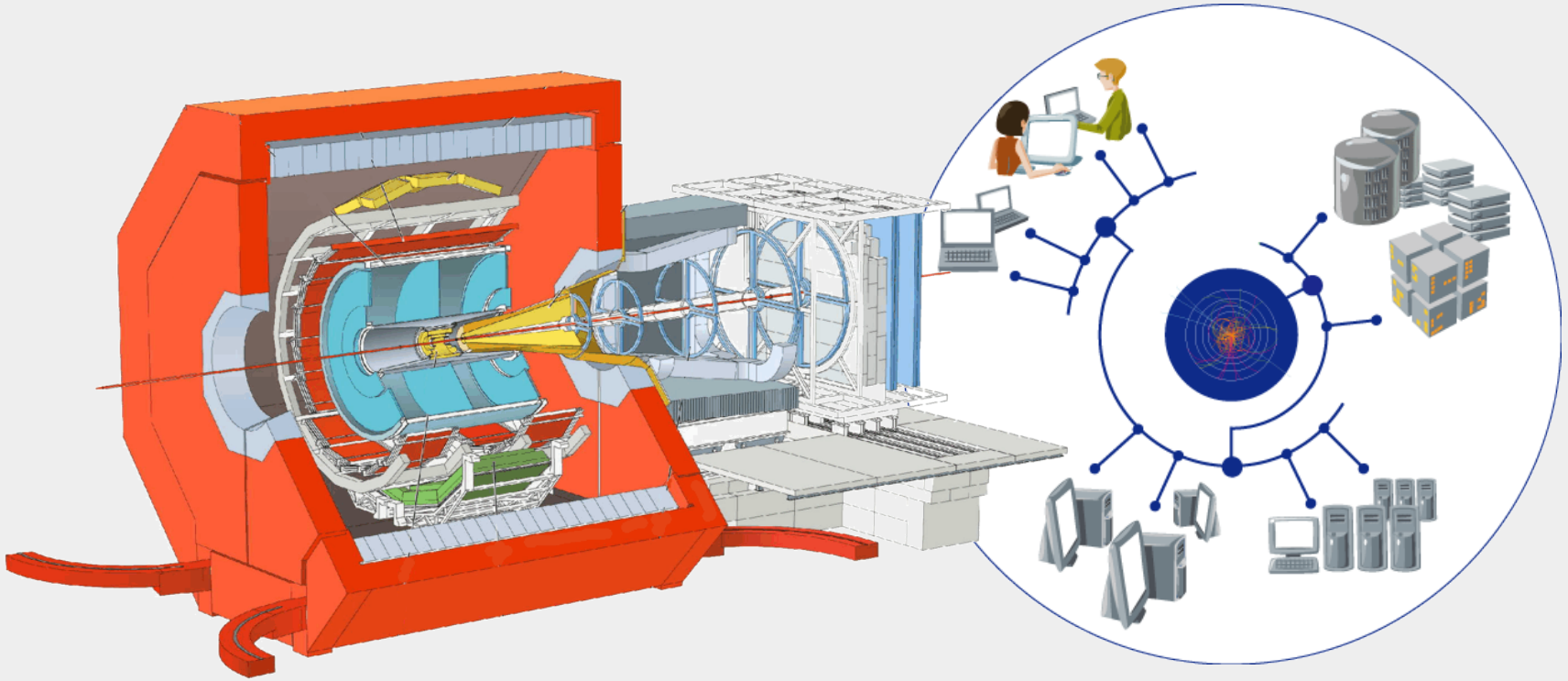


ALICE EXPERIENCE WITH FIRST DATA

S. Bagnasco, INFN Torino

Workshop “Stato e prospettive del calcolo scientifico”
L.N.L. Feb 16-18, 2011

THE ORIGINAL ALICE COMPUTING MODEL

- For **pp** similar to the other experiments
  - Quasi-online data distribution, calibration and first reconstruction at Tier-0
  - Further reconstructions at Tier-1's
- For **HI** different model
  - Online calibration, alignment, pilot reconstructions and partial data export during data taking
  - Data distribution and Pass1 reconstruction at Tier-0 in the four months after HI run (during shutdown)
  - Further reconstruction passes (one) at Tier-1's

THE ORIGINAL ALICE COMPUTING MODEL

- Three kinds of data analysis
  - **Fast pilot analysis** to tune the first reconstruction at CERN Analysis Facility (CAF)
  - **Scheduled batch analysis** on the Grid (Analysis Trains: ESDs and AODs)
  - **End-user interactive or batch analysis** on AAFs and GRID (AODs and ESDs)

- **T0 (CERN)**
  - Does: first pass reconstruction; calibration and alignment
  - Stores: one copy of RAW, calibration data and first-pass ESDs
- **T1s**
  - Does: reconstructions and scheduled batch analysis
  - Stores: second collective copy of RAW, one copy of all data to be kept, disk replicas of ESDs and AODs
- **T2s**
  - Does: simulation and end-user analysis
  - Stores: disk replicas of AODs and ESDs

THREE JOB CLASSES

- MC simulation & reco production
  - Low I/O, high CPU efficiency
  - Data export after job completion
  - Managed, scheduled
- Analysis Trains
  - Optimized I/O (read once, do many tasks)
  - Streamlined code (as much as possible…)
  - Managed, scheduled
- User jobs
  - Lowest CPU efficiency
  - Variable job duration, lots of failures, far-from-perfect code
  - Unmanaged, chaotic

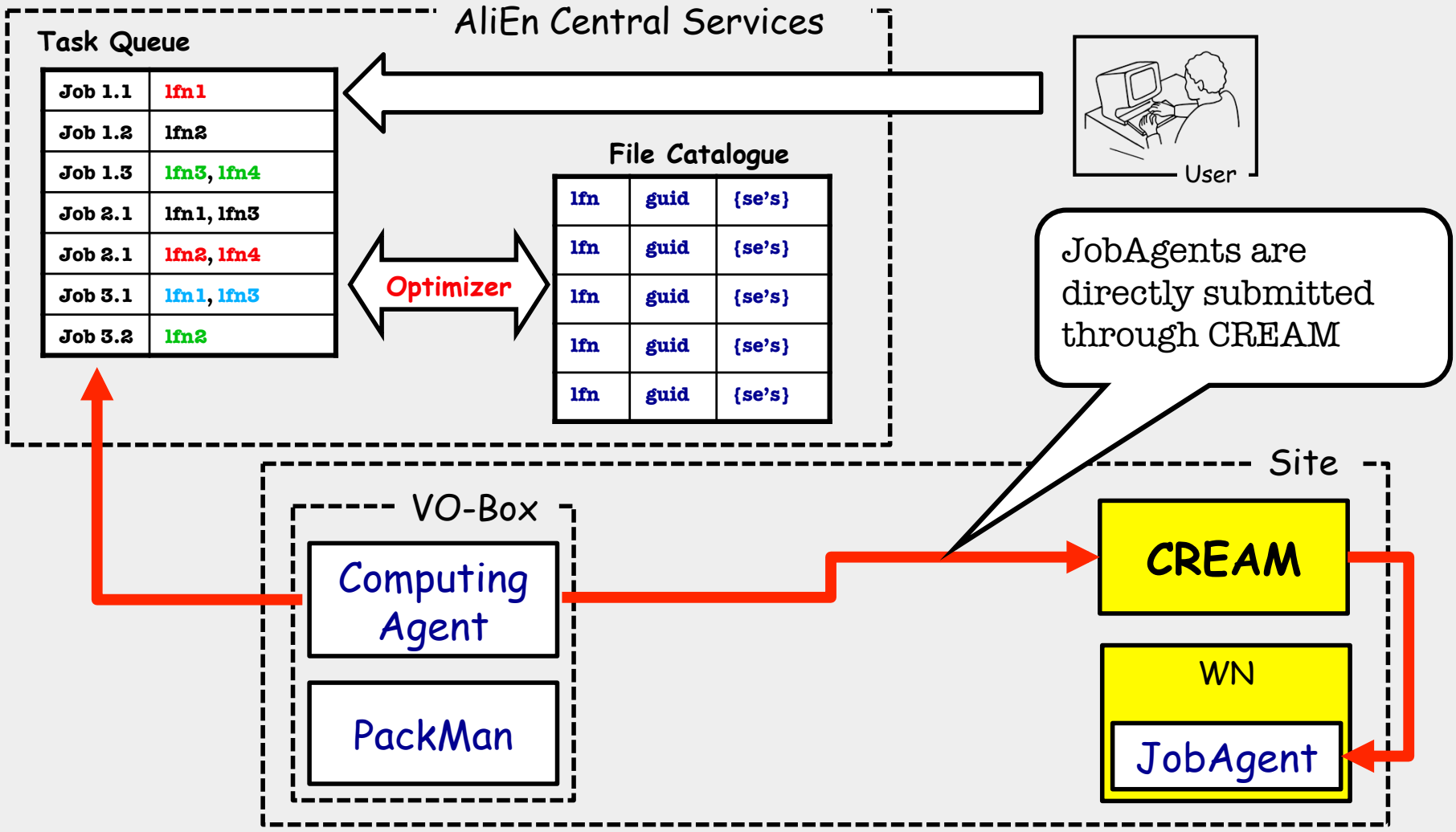
COMPUTING STRATEGY

- AliEn as a common front-end for all distributed resources
  - Using transparent interfaces to different grids where needed
  - Xrootd as a common file access protocol
- Jobs are assigned where data is located
  - All policies (data & CPU) enforced on central servers
  - WMS efficiency not a big issue thanks to JAs strategy
- Resources are shared
  - No “localization” of groups
  - Fairshare Group/Site Contribution and Consumption *will* be regulated by accounting system
  - Prioritisation of jobs in the central ALICE queue
- Data access only through the GRID
  - No backdoor access to data
  - No “private” processing on shared resources
  - No “private” resources outside of the grid

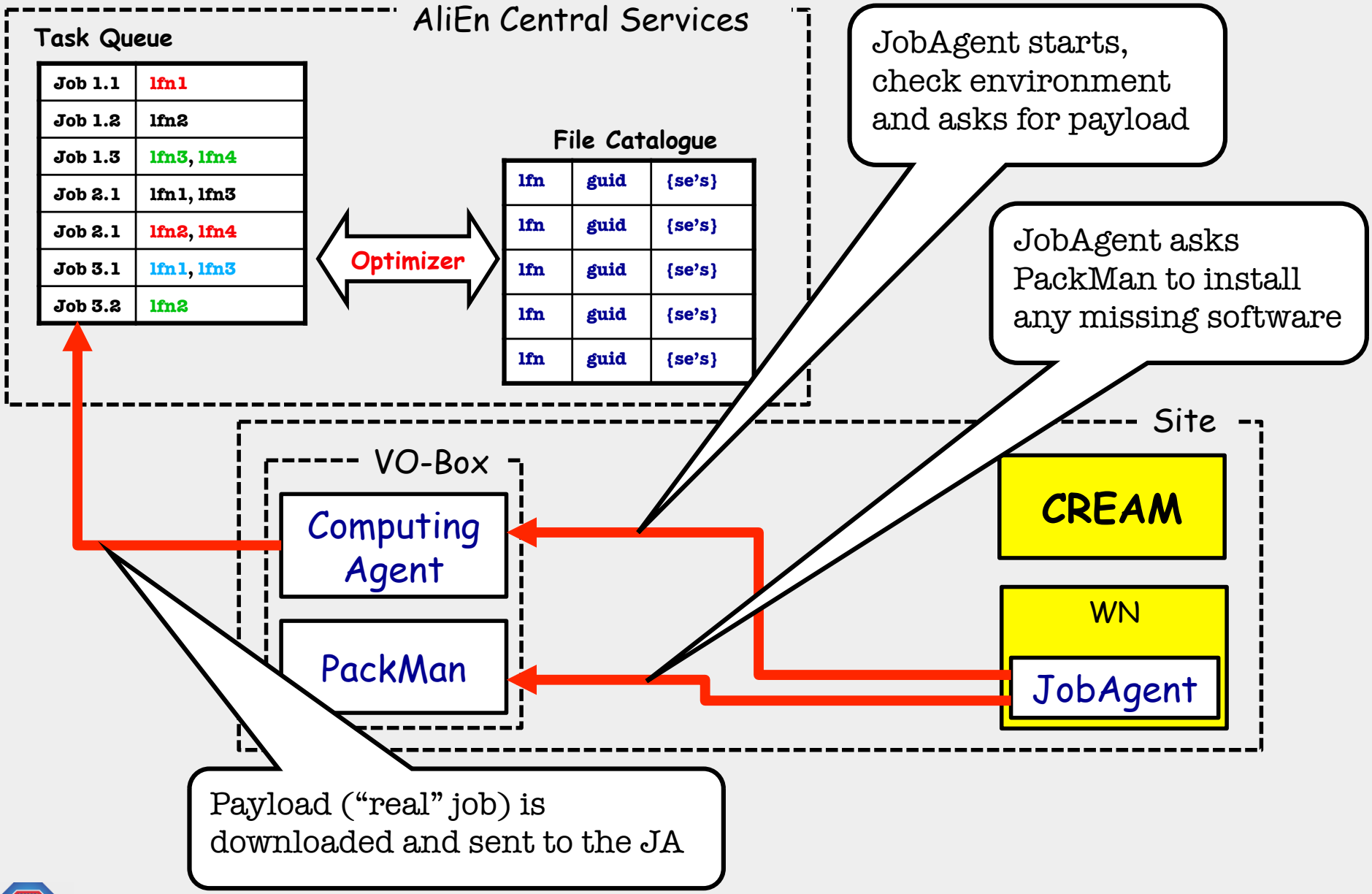
JOB MANAGEMENT KEY CONCEPTS

- Task Queue and Optimizers
  - Central DB of jobs to be executed
  - Optimizers split and arrange jobs according to input data, priority policies, job quotas and/or user defined criteria
- Site VO-Box
  - Thin interfaces to underlying Grid site services
  - Submits JobAgents to CREAM keeping batch queues constantly populated, monitors site health
  - Takes care of proxy management
- JobAgent
  - Runs on WNs, downloads payload from the TQ and executes it
  - JAs create a “virtual grid” on top of existing Grid infrastructures (gLite-based, but also ARC or raw LRMS)

JOB MANAGEMENT I



JOB MANAGEMENT II



DATA MANAGEMENT KEY CONCEPTS

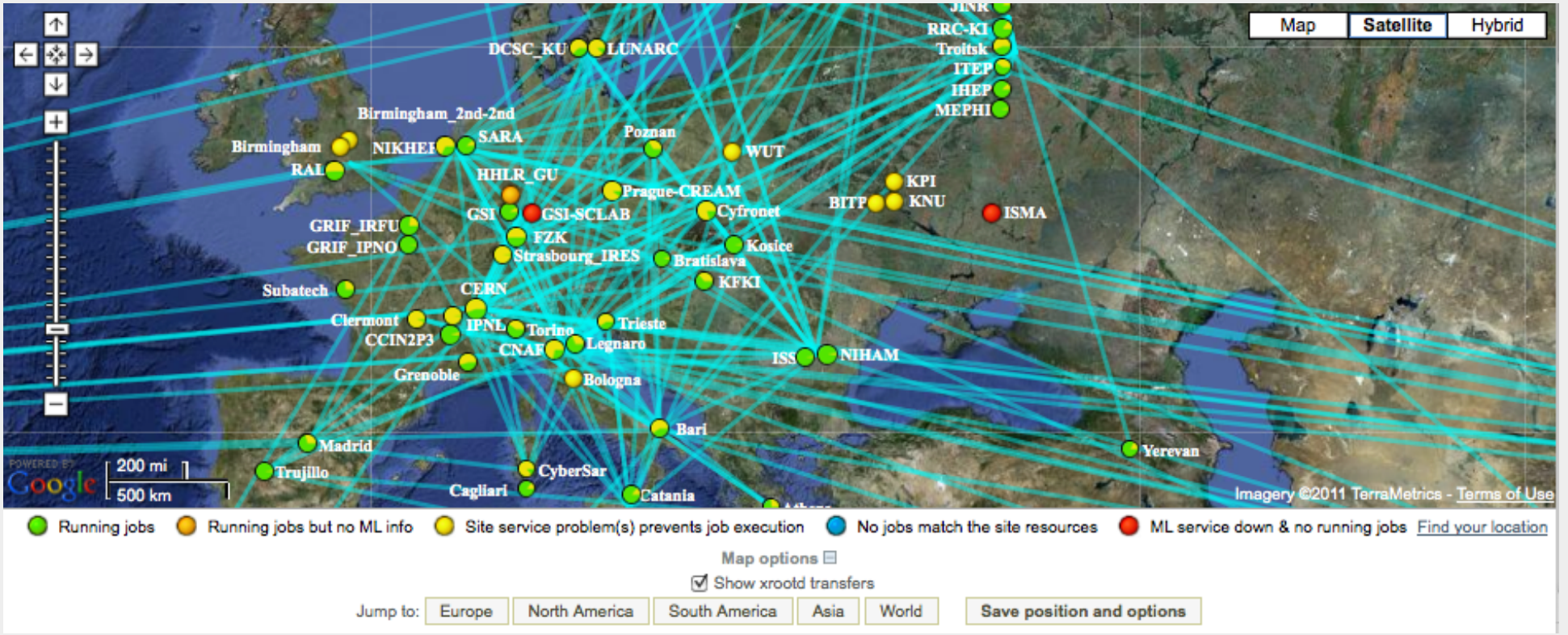
- Central File Catalogue
  - Central DB of all file produced
  - Enforcement of access rights, quotas, policies etc.
  - FS-like browsable interface for users
- Calibrations and conditions data are no different
  - Root files accessible via catalogue entries
- Xrootd as uniform access protocol
  - Across **sites**, **storage architectures** and **use cases**
  - Run the same analysis macro locally, on PROOF or on the Grid accessing data regardlessly of their physical location
- Central transfers queue (FTD)
  - Manages data transfers
  - Uses xrd3cp for transfers

DATA MANAGEMENT

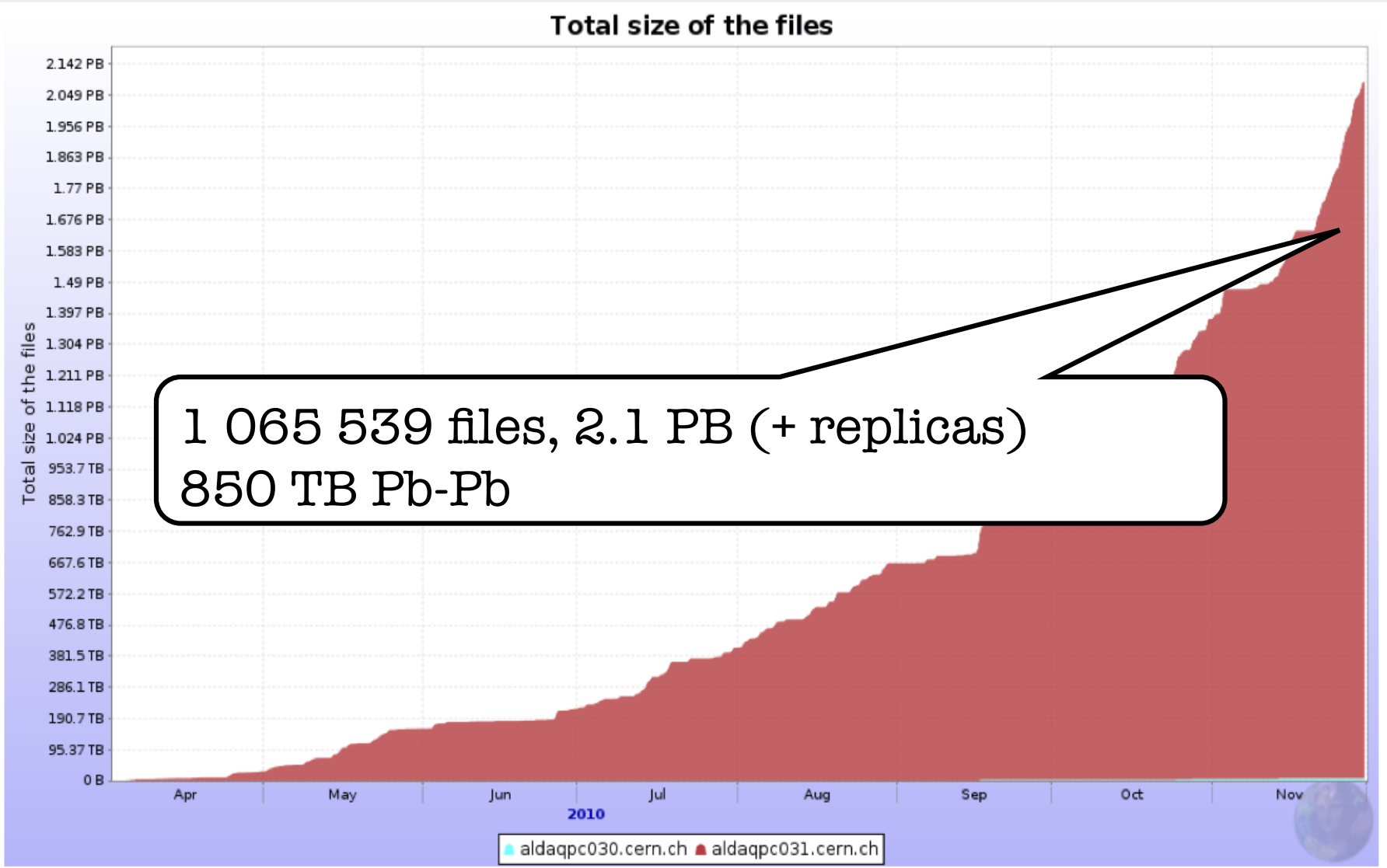
OTHER BITS AND PIECES

- Monitoring
  - MonALISA repository at CERN
  - **Everything** has a sensor and sends data
- Interactive parallel analysis facilities
  - Based on PROOF
  - More later
- Torrent-like package distribution
  - Also, automated AliEn upgrade on VO-Boxes

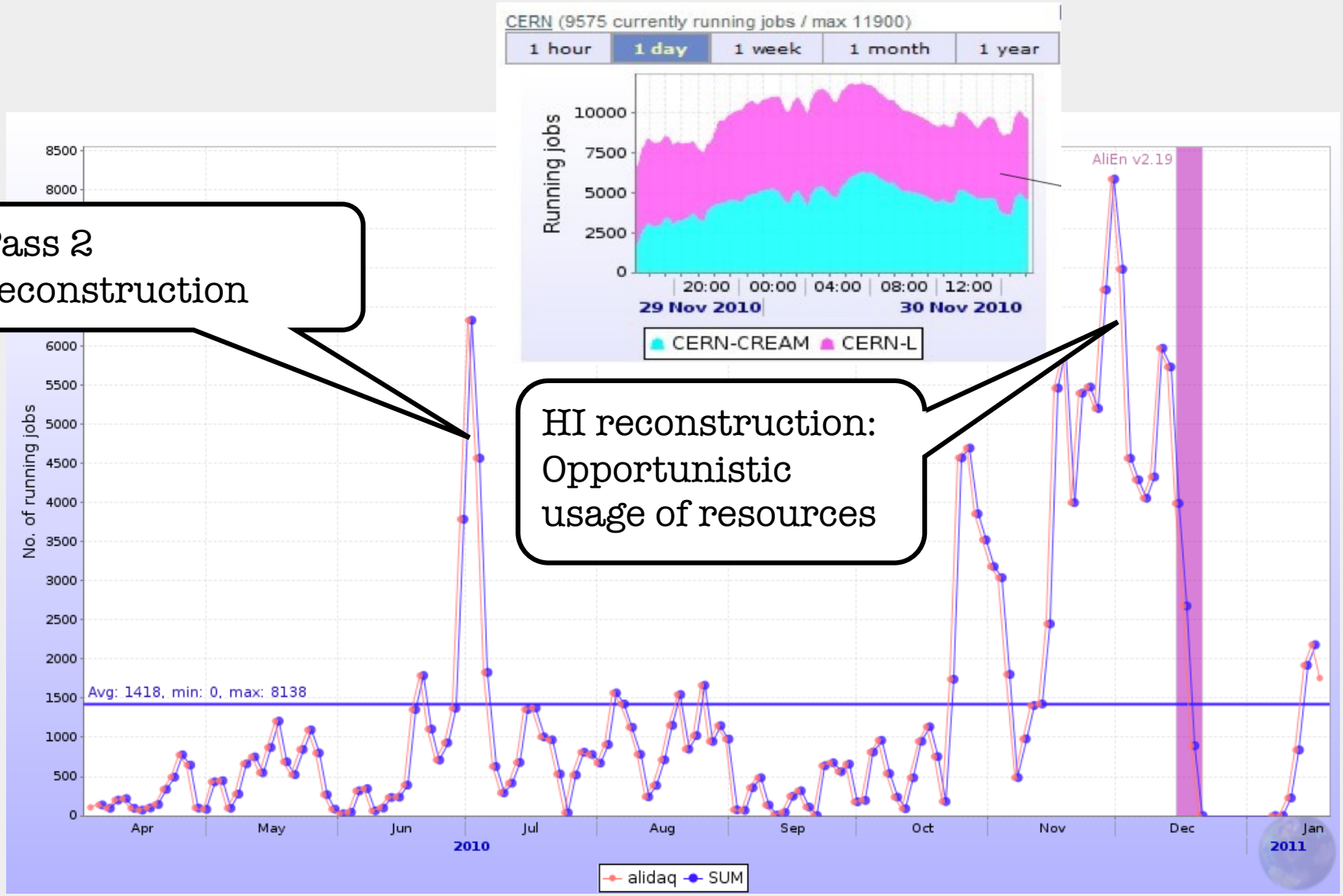
MANDATORY GOOGLE MAP



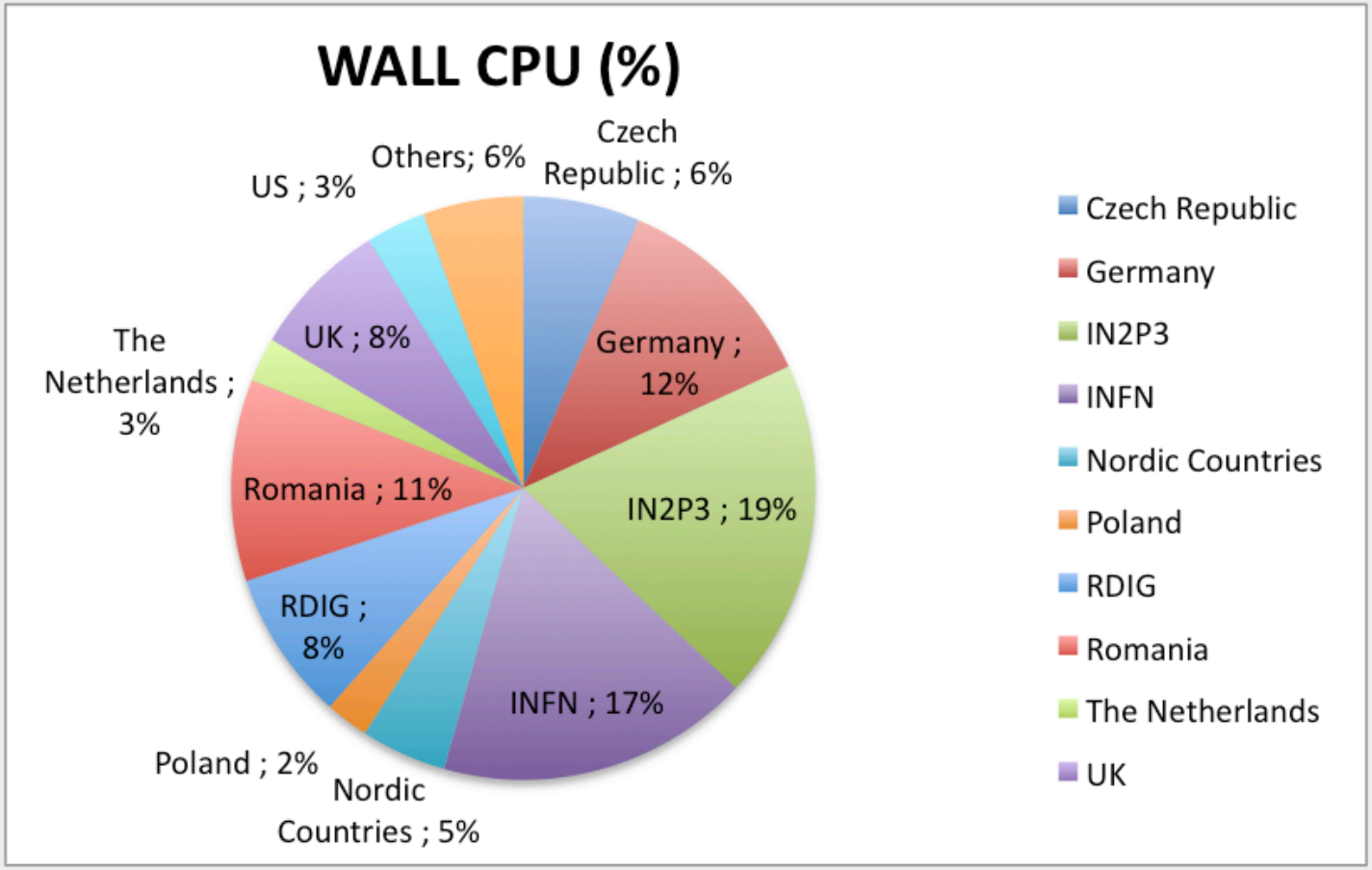
RAW DATA SAMPLE



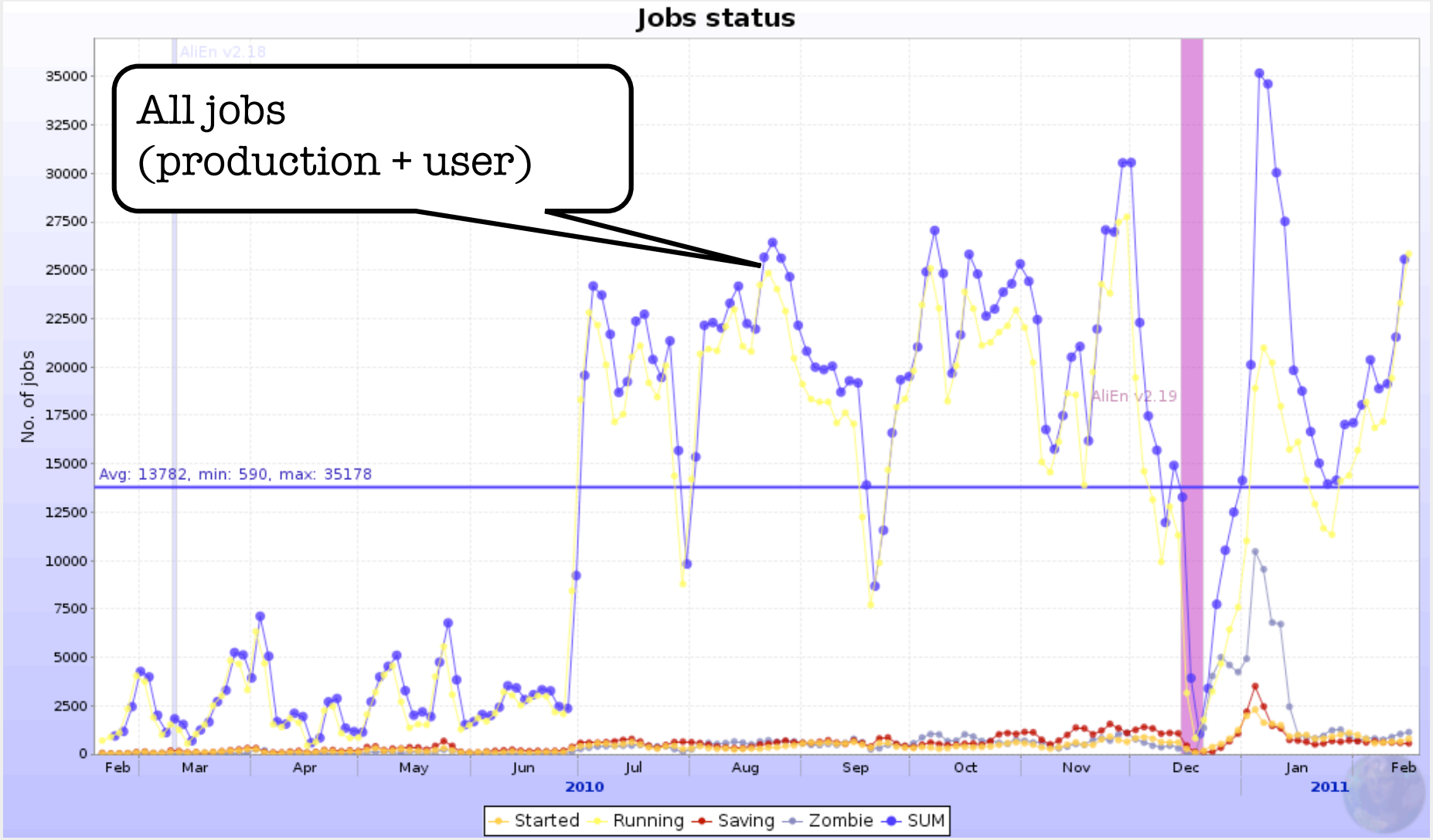
HI DATA PROCESSING



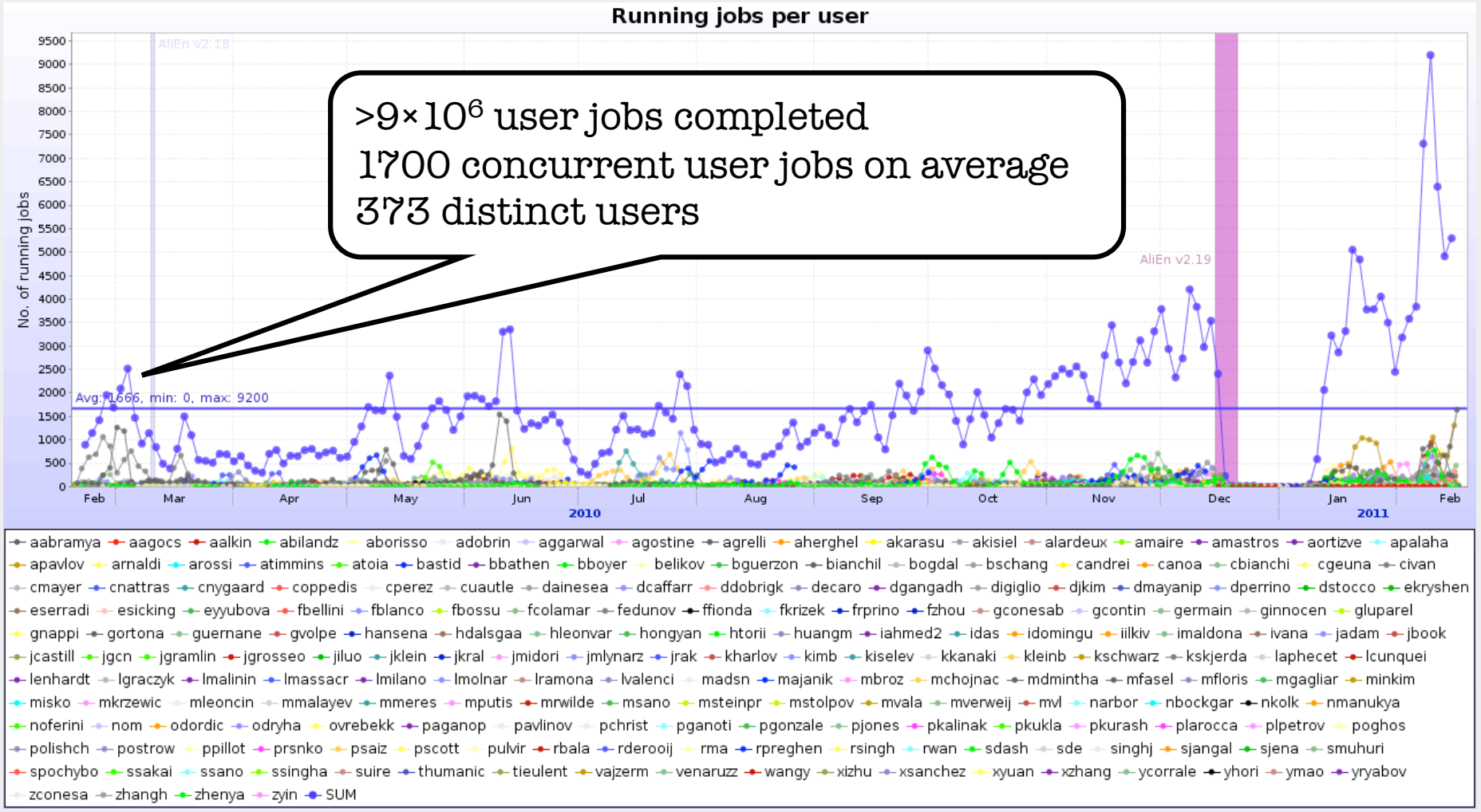
- July 17, 2010 to January 16, 2011



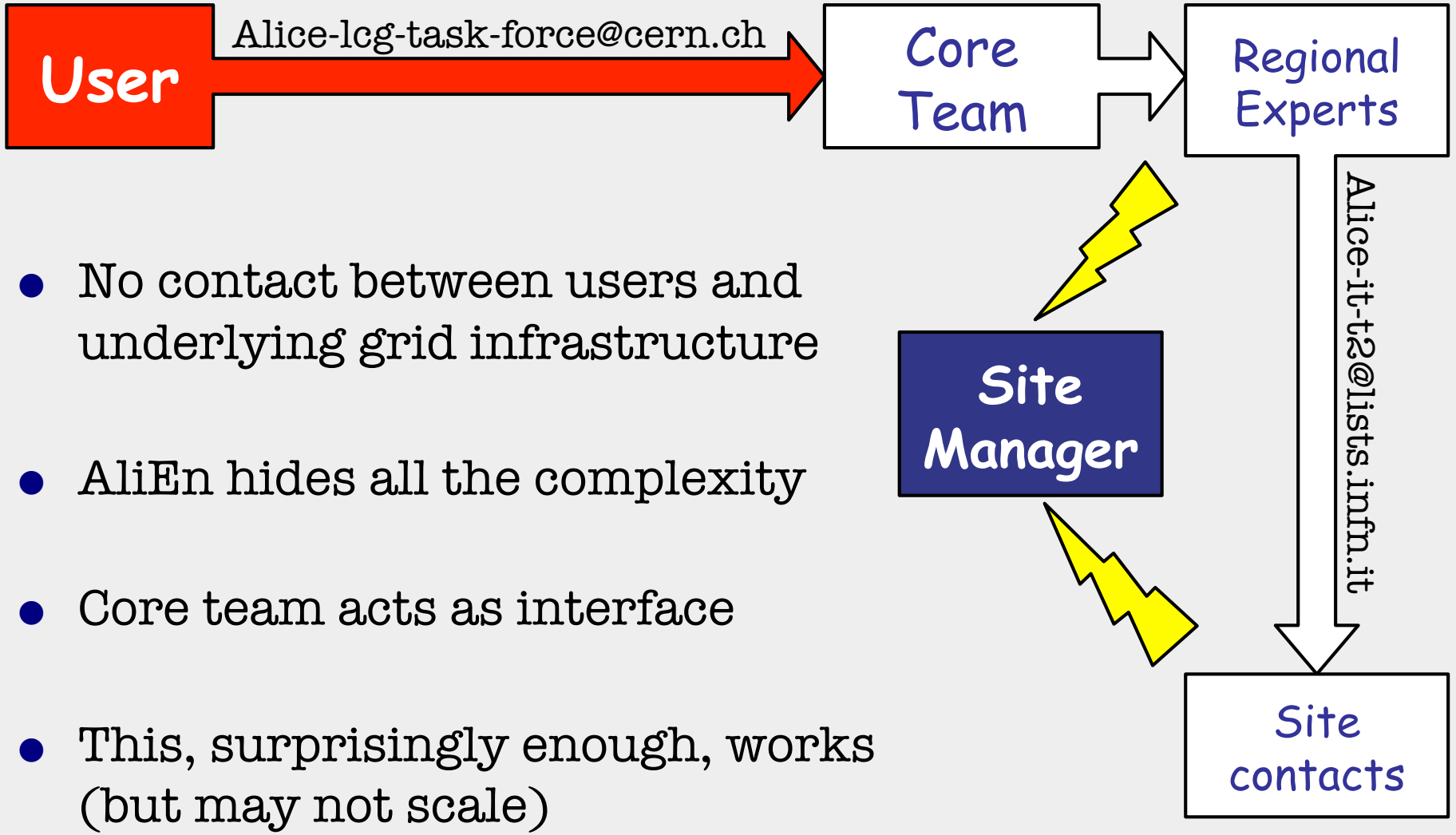
ONE YEAR RUNNING PROFILE



USER ANALYSIS ON THE GRID



USER SUPPORT EXPERIENCE



- No contact between users and underlying grid infrastructure
- AliEn hides all the complexity
- Core team acts as interface
- This, surprisingly enough, works (but may not scale)

SITE MANAGEMENT EXPERIENCE

- VO-Box management
  - Remarkably stable
  - Automatic upgrades
  - Maybe too automatic: “site contacts” have no experience to fix the occasional problem
- Xrootd management
  - See also G. Donvito’s talk
  - Stable, but lacks documentation
  - Unclear responsibility: site admins or ALICE site contact?

USER JOBS ARE MESSY

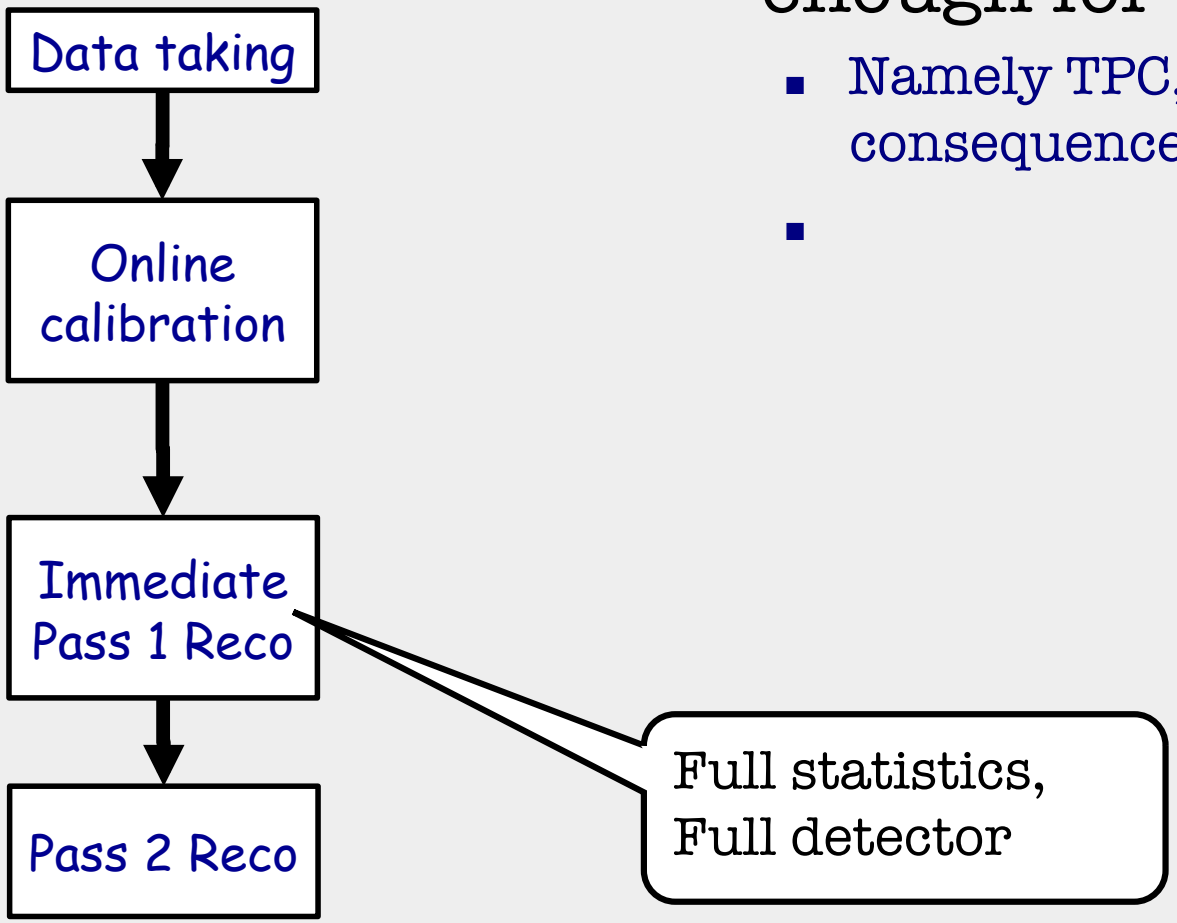
- Diverging memory allocation
  - Killed jobs or even stuck WNs
  - A safety is in place with new AliEn release, see below
- Coding and JDL errors
  - Private code is never tested enough
  - Thousands of jobs failing **very** quickly
  - User problem or site problem?
- A user will do anything with an open query
  - E.g. queries with o(10^6) files
  - Whether it makes sense or not
  - Protections need to be in place everywhere

HEAVY IONS DATA PROCESSING

- Pass 1 reconstruction worked well since the beginning
  - Both for p-p and Pb-Pb
- Some extra tasks added to original CM
  - Offline calibration
  - QA Analysis trains
  - See next slides
- Everything worked but computing model is evolving to take this into account

COMPUTING MODEL EVOLUTION

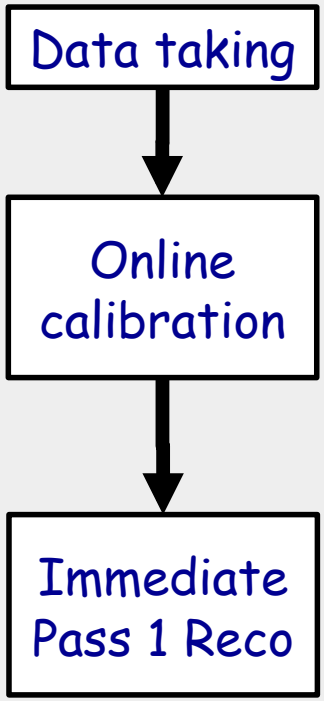
Original computing model



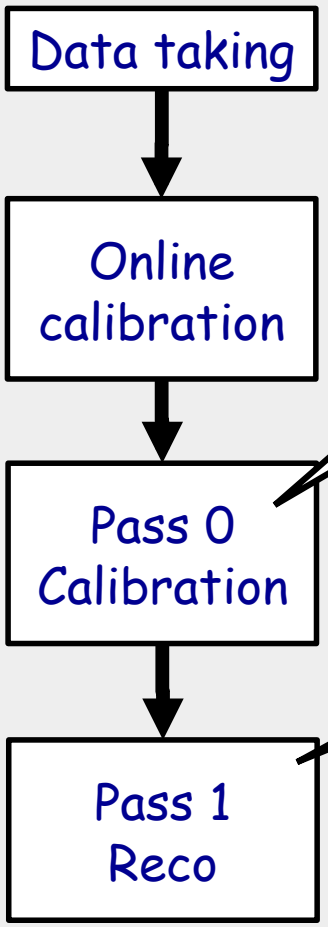
- Online calibration not enough for some detectors
  - Namely TPC, others as a consequence
  -

COMPUTING MODEL EVOLUTION

Original computing model



Current implementation

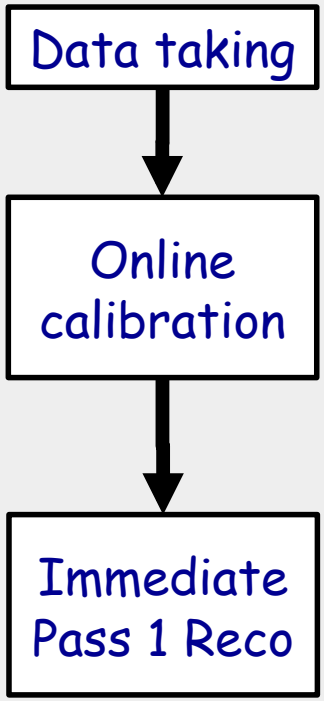


Partial sample o(10%), Only relevant subdetectors

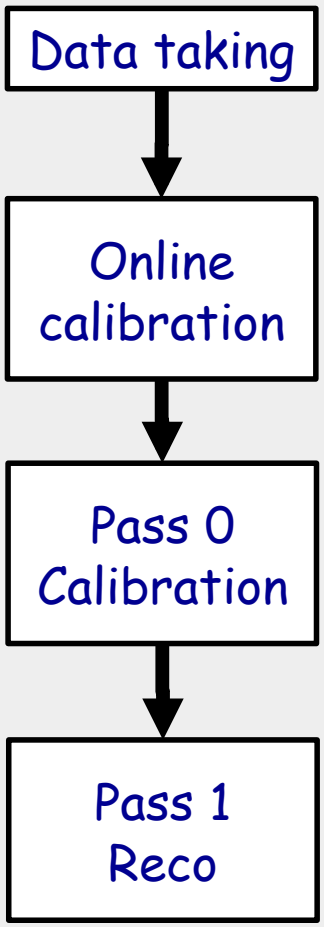
Full statistics, Full detector

COMPUTING MODEL EVOLUTION

Original computing model



Current implementation



Under discussion

UPDATED CM PARAMETERS

BASE PARAMETERS

System	Events	RAW (TB)	RAW/ev (MB)	Reco (KHEP06×s)	MC (KHEP06×s)
p–p	1.5×10^9	1100	0.8	0.07	1.30
Pb–Pb	9×10^7	880	11.7	6.50	150.00

- ADDITIONS TO TRD AND EMCAL INCREASE RAW DATA SIZE BY 11% FOR PP AND BY 20% FOR PBPB
- 25% MORE CPU TIME DUE TO PASSO
- EFFECT OF ADDITIONS TRD AND EMCAL ON CPU TIME TAKEN INTO ACCOUNT
- SELECTIVE TRIGGERS (INSTEAD OF MB) EFFECT ON PBPB SIZE TAKEN INTO ACCOUNT

F. Carminati

COMPUTING MODEL EVOLUTION

- More files than ever anticipated
  - Original model:
    1 RAW → 1 ESD → 1 AOD (×3 passes)
  - Current cascade:
    1RAW → 5× ESD-related (×3 passes) →
    → 6× AOD-related (per train) (×*N* passes)
  - MC is more difficult to describe, but also a substantial generator of files
  - Users are prolific generators of files (*.root)
  - For one year we have accumulated **25×10⁶** files in the catalogue (RAW is 1.1 ×10⁶)
  - The physical replication of the above is about **4.2**

COMPUTING MODEL EVOLUTION

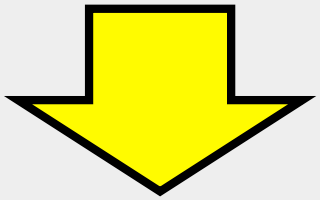
- More complex job structure
  - Added few more reconstruction passes and analysis trains to the original processing model
  - MC is increasing in complexity and is more fragmented (PWG requests,...)
  - User access strongly depends on the file fragmentation from the productions
  - In general, the **jobs are becoming more complex and demanding** on the entire Grid infrastructure

COMPUTING MODEL EVOLUTION

- More access to calibration
  - OCDB is 5× bigger (in number of objects) than originally anticipated
  - Access to OCDB is ~30 more frequent than original projections
  - Will increase substantially with more Pass0, Trains and Tenders – but how much?
  - All of the above has increased the load on the AliEn catalogue and access services dramatically
  - In addition to the massive file access within and outside of the Grid framework

ALIEN V.2-19

- Computing Model evolution
  - See previous slides
- GRID services and operations need to be less expert-dependent
  - Manpower always an issue
- System security needed improvement
  - Sites need to “trust” ALICE



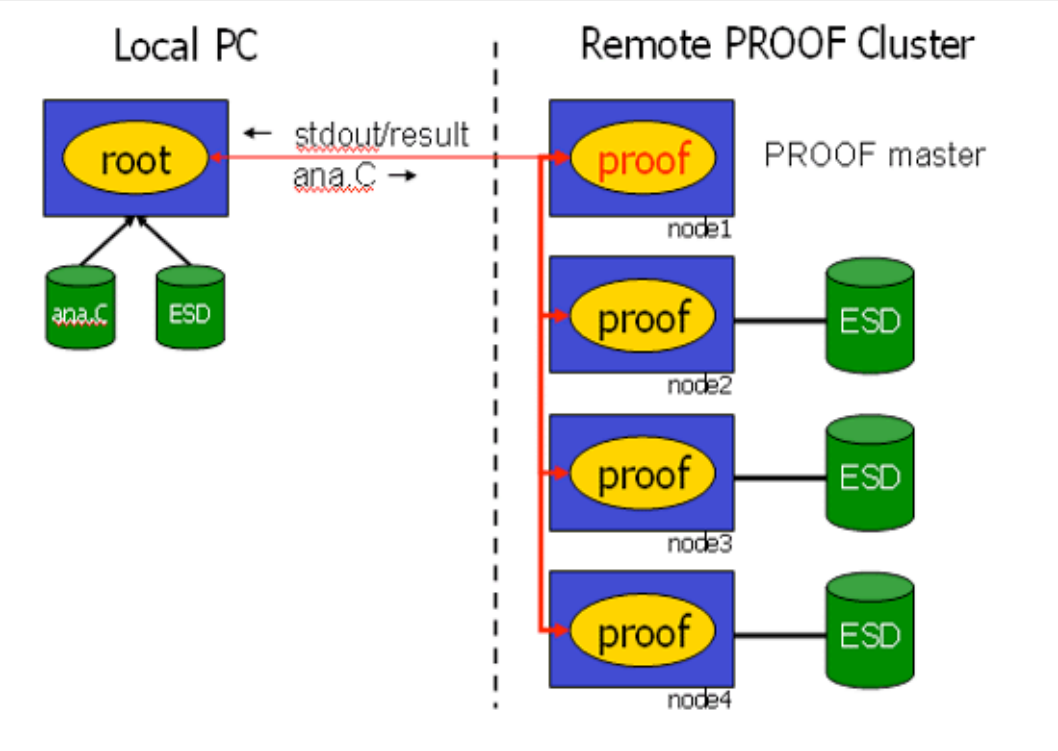
AliEn v.2-19

ALIEN V.2-19

- Security
  - All client to central services communications are now over https
  - The file security envelopes are entirely refurbished
- Services
  - Consolidation and fewer central services
  - Job splitting substantially (×4) faster
- File Catalogue
  - DB schema better adapted to large directories with more files
  - 'booking tables' for files to be removed physically from storage
- Access services for user clients
  - Fully backward compatible (not an easy task)
- (and more...)

PROOF-BASED ANALYSIS FACILITIES

- Interactive parallel processing of events
  - Local storage: data need to be staged from SEs
  - No resource management.
    User priorities managed “by hand”
  - Statically allocated resources



PROOF-BASED ANALYSIS FACILITIES

- CAF at CERN, SKAF in Košice
  - But also under development in Russia, Korea, France
  - And, very recently, in Torino

ALICE PROOF Clusters

What is this about?

Cluster list

	Cluster					ROOT	Aggregated disk space			AF xrootd		xrootd
Name	Online	Status	Proof master	Workers	Users	Version	Total	Free	Used	Running	Latest	Version
1. CAF		Stable	alice-caf.cern.ch	160	3	v5-27-06c	115.7 TB	40.77 TB	74.89 TB	1.0.38	1.0.38	20100510-1509_dbg
2. JRAF		Maintenance sin...	jraf.jinr.ru	8	0	v5-27-06c	2.014 TB	1.858 TB	160 GB	1.0.38	1.0.38	20100510-1509_dbg
3. KIAF		Maintenance sin...	kiaf.sdfarm.kr	48	0	v5-27-06c	798.2 GB	199.3 GB	598.9 GB	1.0.38	1.0.38	20100510-1509_dbg
4. SAF		Maintenance sin...	nansafmaster.in2p3.fr	48	0	v5-27-06c	12.07 TB	2.63 TB	9.442 TB	1.0.38	1.0.38	20100510-1509_dbg
5. SKAF		Stable	skaf.saske.sk	60	0	v5-27-06c	53.72 TB	30.41 TB	23.31 TB	1.0.39	1.0.39	20100510-1509_dbg
6. SKAF_TEST		Testing	skaf-test.saske.sk	2	0	v5-27-06c	815.9 GB	676.7 GB	139.2 GB	1.0.38	1.0.38	20100510-1509_dbg
7. TAF		Open to local u...	pmaster.to.infn.it	16	1	v5-27-06c	3.914 TB	96.83 GB	3.819 TB	1.0.38	1.0.38	20100510-1509_dbg
Total				**342**	**4**		**189 TB**	**76.62 TB**	**112.3 TB**			

EVOLUTION

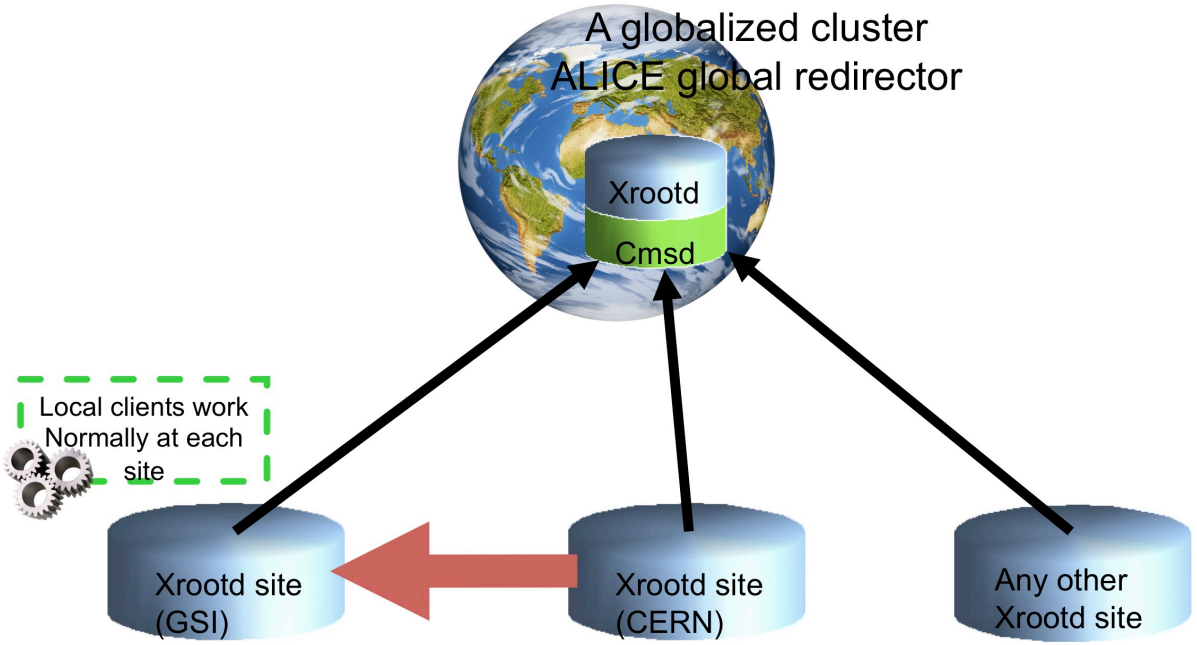
- Statically allocated resources may be a waste
  - A few "AF-on-demand" projects ongoing
  - Virtual machines? Special grid jobs? Cloud-like systems?
- PROOF is mature and stable
  - But not all desirable features are there
  - E.g. failsafe mechanisms
  - Some are already coming (sub-mergers, hierarchical masters,...)
- Also framework needs some improvement
  - For better integration between PROOF, AliEn, GRID and local environments
  - And more functionalities
- Is direct access to SE (no staging) efficient?
  - Xrootd's "Virtual Mass Storage" feature

ALICE

DM

More than Globalization: The VMSS

CERN IT Department



Missing a file?
Ask to the global redirector
Get redirected to the right collaborating cluster, and fetch it.
Immediately.

F. Furano, A. Hanushevsky - Scalla/xrootd WAN globalization tools: where we are. (CHEP09)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

F. Furano

CONCLUSIONS

- The Computing Model structure holds and is working
  - But differences between the role of Tier-1s and Tier-2s are shrinking
- Exact parameters and details are being reviewed
  - Extra calibration passes
  - More files than expected
- AliEn development follows this evolution
  - Catalogue not yet a bottleneck
  - For how much longer?
- User analysis on the grid is successfully working
  - Millions of jobs, hundreds of users
  - Job quota system in place
- Interactive Analysis Facilities are being deployed for rapid-turn-around analyses

THANKS

- Questions?