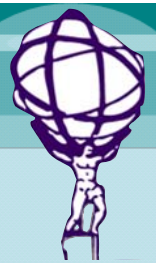
ATLAS recent production : Achievements and Lessons

SC3 Planning Workshop

CERN, June 14, 2005

P. Nevski (BNL)



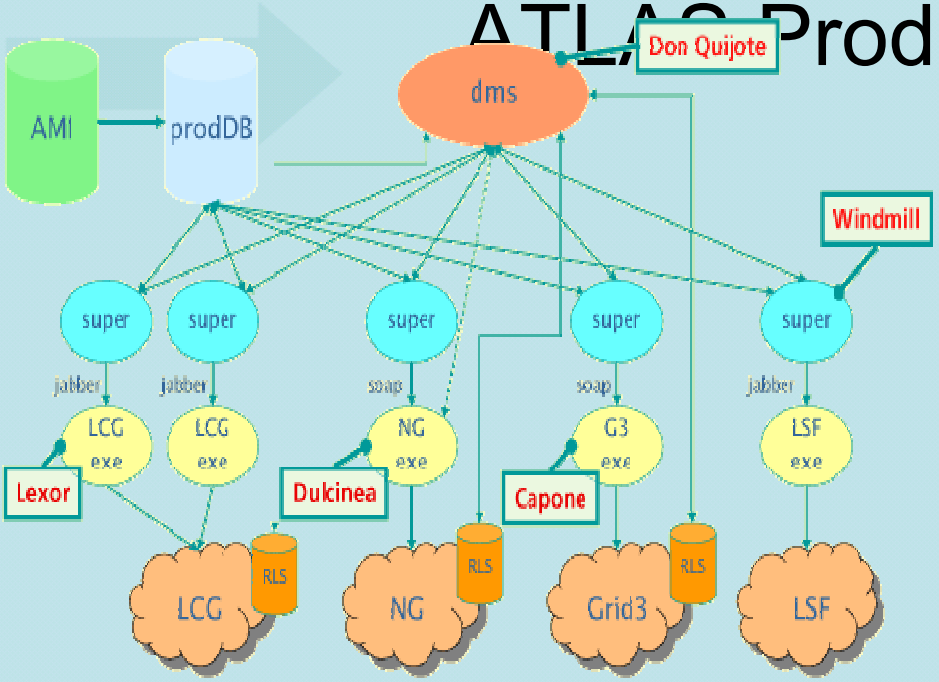
Outline

- Grid Production infrastructure
- Expectations and realities
- Databases issues
- Bottlenecks and their resolution
- Roadmap to more redundant and robust services

Physicist view on Production System

- Produce simulated data for Physics study (Rome workshop)
- Push production system to its breaking point ==========➔
- DC2: 10M evts simulated, 2M piled-up, no reconstruction on GRID
- Rome: 173 data sets containing 6.1M events simulated and reconstructed (without pile-up)
- Total simulated data: 8.5M events
- Pile-up is still ongoing (1.3M done, 50K reconstructed)

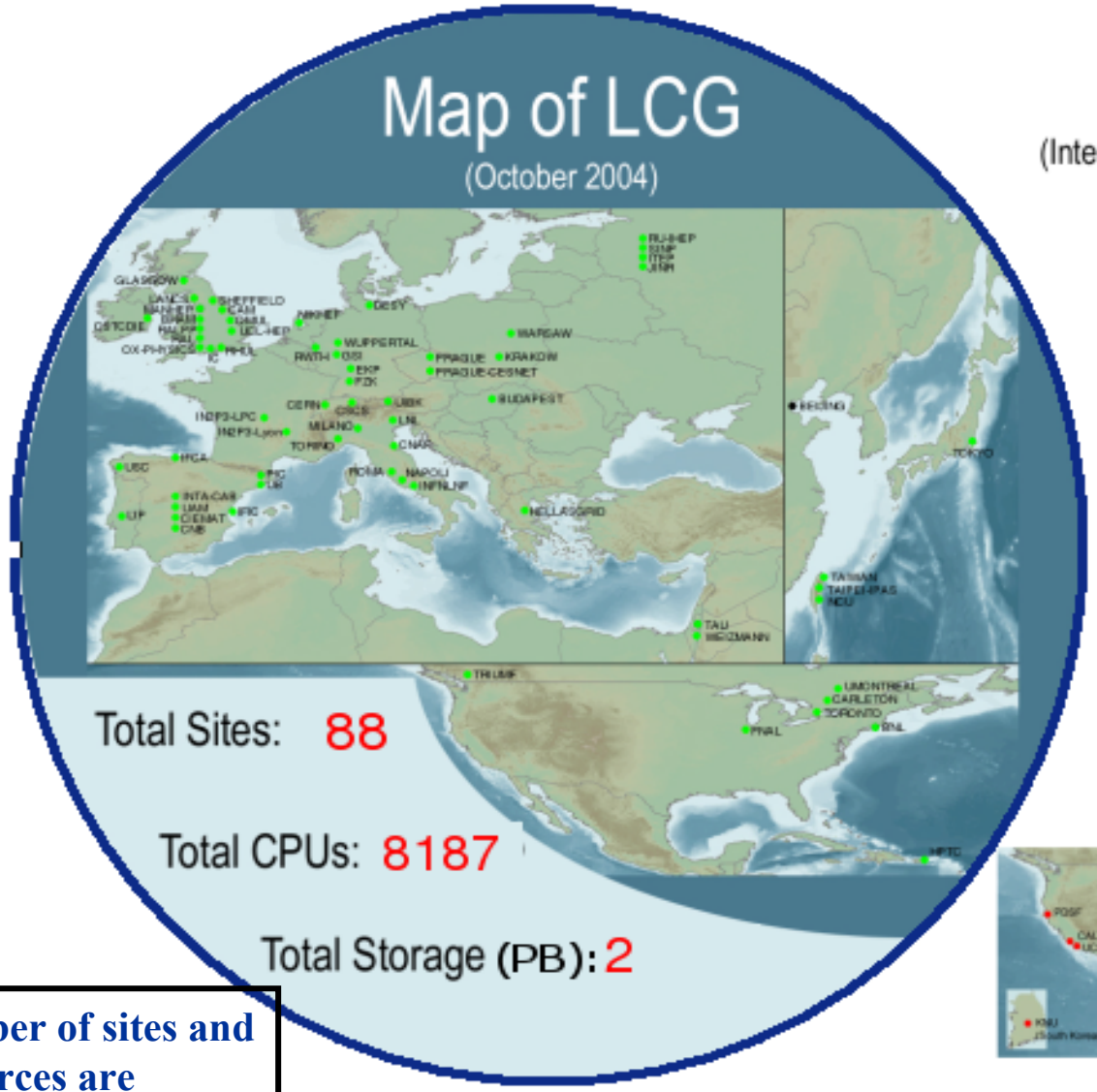
ATLAS Production System



- The production database, which contains abstract job definitions
- The Windmill supervisor that reads the production database for job definitions and present them to the different Grid executors in an easy-to-parse XML format
- The Executors, one for each Grid flavor, that receives the job-definitions in XML format and converts them to the job description language of that particular Grid
- DonQuijote, the ATLAS Data Management System, moves files from their temporary output locations to their final destination on some Storage Elements and registers the files in the Replica Location Service of that Grid

- In order to handle the task of ATLAS DC2 an automated Production system was developed.
- It consists of 4 components:

The 3 Grid flavors: LCG/CG



NorduGrid
(Interoperating with LCG)

Grid3
(Interoperating with LCG)

Number of sites and resources are evolving quickly

The 3 Grid flavors: NorduGrid

- NorduGrid is a research collaboration established mainly across Nordic Countries but includes sites from other countries.
- They contributed to a significant part of the DC1 (using the Grid in 2002).
- It supports production on non-RedHat 7.3 platforms

•11 countries, 40+ sites, ~4000 CPUs,
•~30 TB storage

s: Grid3

Sep 04
- **30 sites, multi-VO**
- **shared resources**
- **~3000 CPUs (shared)**

- The deployed infrastructure has been in operation since November 2003
- At this moment running 3 HEP and 2 Biological applications
- Over 100 users authorized to run in GRID3

ATLAS Production: countries & sites

- Australia (1)
- Austria (1)
- *Canada* (4)
- CERN (1)
- Czech Republic (2)
- Denmark (4)
- France (1)
- Germany (1+2)
- Italy (7)
- Japan (1)
- Netherlands (1)
- Norway (3)
- Poland (1)
- Slovenia (1)
- Spain (3)
- Sweden (7)
- Switzerland (1)
- Taiwan (1)
- UK (7)
- USA (19)

20 countries
69 sites

LCG

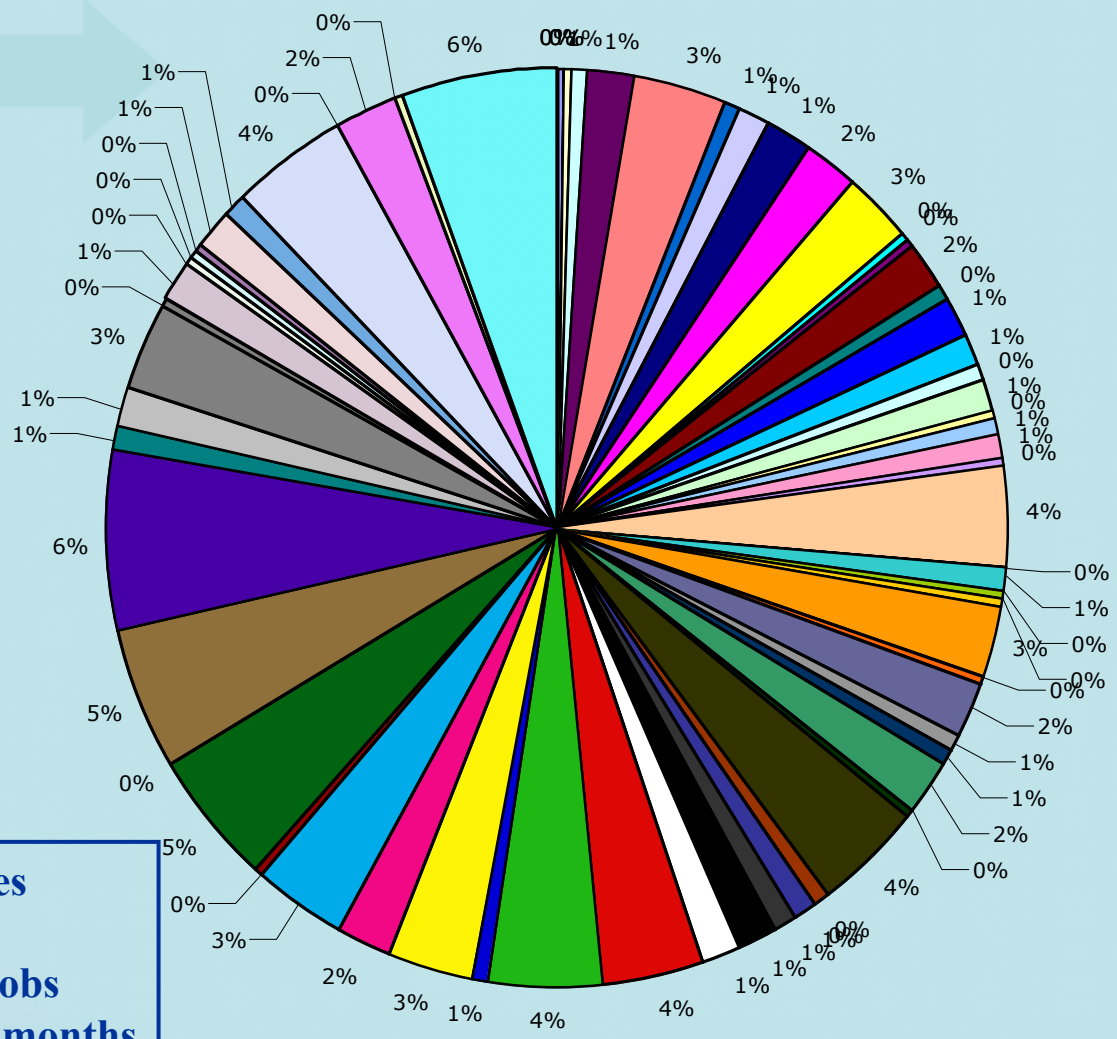
13 countries; 31 sites

NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

7 countries; 19 sites

Grid3

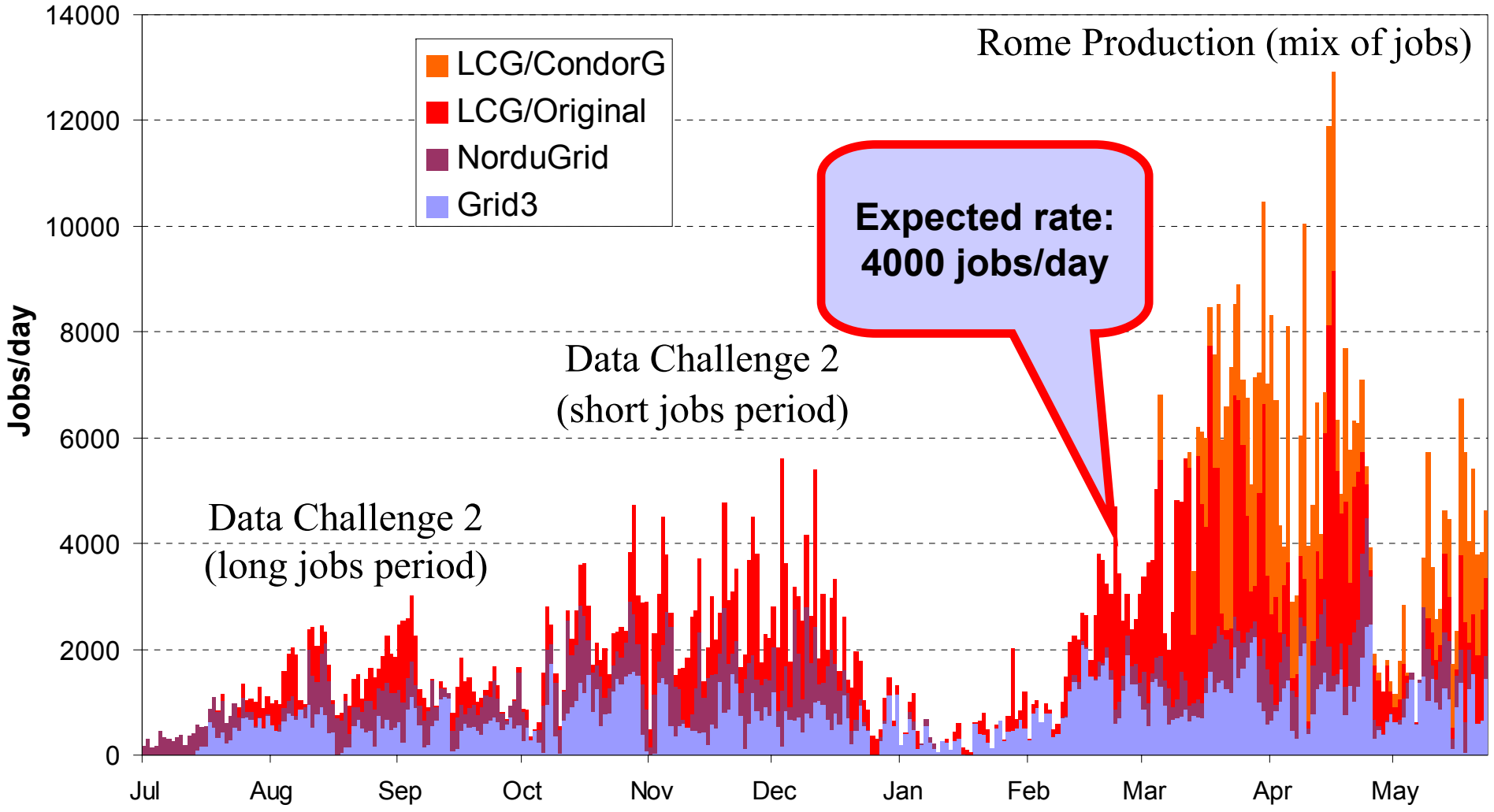
Jobs Total



- at.uibk
- ca.alberta
- ca.montreal
- ca.toronto
- ca.triumf
- ch.cern
- cz.cesnet
- cz.golias
- de.fzk
- es.ifae
- es.ific
- es.uam
- fr.in2p3
- it.cnaf
- it.lnf
- it.lnl
- it.mi
- it.na
- it.roma
- it.to
- jp.icepp
- nl.nikhef
- pl.zeus
- tw.sinica
- uk.cam
- uk.lancs
- uk.man
- uk.pp.ic
- uk.rl
- uk.shef
- uk.ucl
- au.melbourne
- ch.unibe
- de.fzk
- de.lrz-muenchen
- dk.aau
- dk.dcgc
- dk.nbi
- dk.sdu
- no.uib
- no.grid.uio
- no.hypatia.uio
- se.hoc2n.umu
- se.it.uu
- se.lu
- se.lunarc
- se.nsc
- se.pdc
- se.unicc.chalmers
- si.ijs
- ANL_HEP
- BNL_ATLAS
- BU_ATLAS_Tier2
- CalTech_PG
- FNAL_CMS
- IU_ATLAS_Tier2
- OU_OSCER
- PDSF
- PSU_Grid3
- Rice_Grid3
- SMU_Physics_Cluster
- UBuffalo_CCR
- UCSanDiego_PG
- UC_ATLAS_Tier2
- UFlorida_PG
- UM_ATLAS
- UNM_HPC
- UTA_dpcc
- UWMadison

20 countries
69 sites
~ 260000 Jobs
~ 2 MSi2k.months

Production Rate Growth



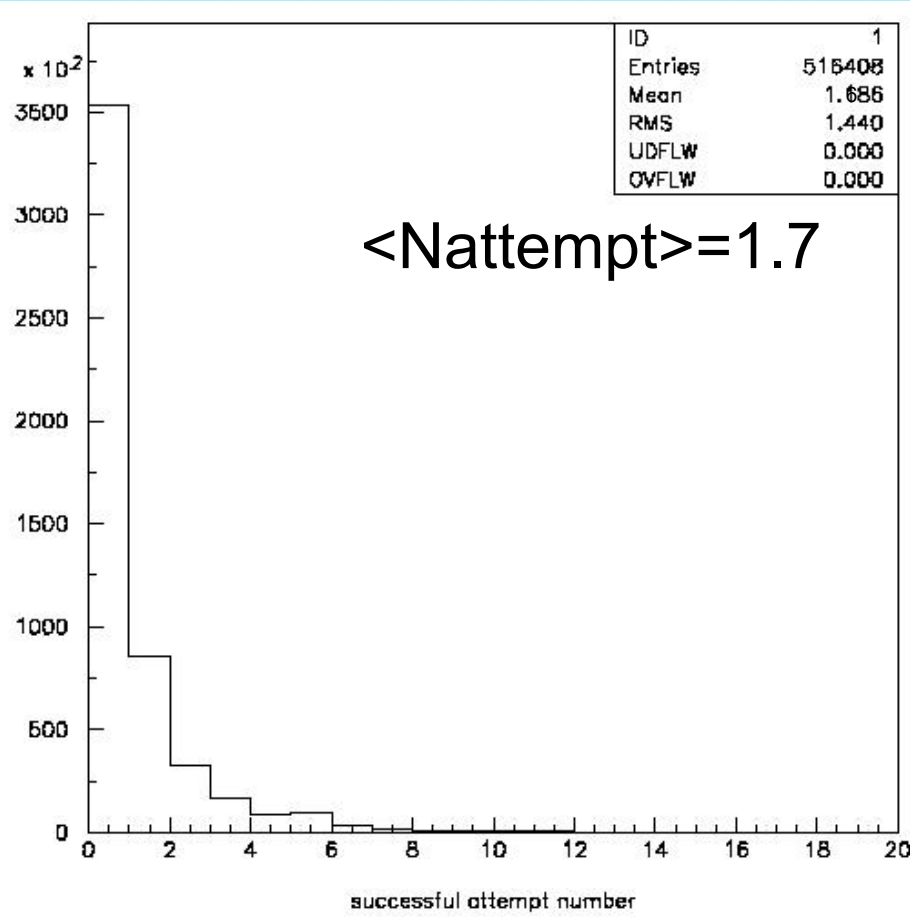
GRID Job statistics

- 516450 jobs done ===>
- 60259 jobs NOT done
- 75872 jobs have no input
- 36085 jobs aborted (bad definitions)

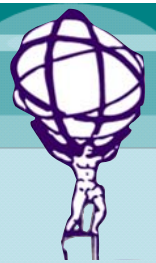
Not really bad !

- LCG:CG:GRID3:NG

~ 40: 30: 20: 10



<Nattempt>=1.7

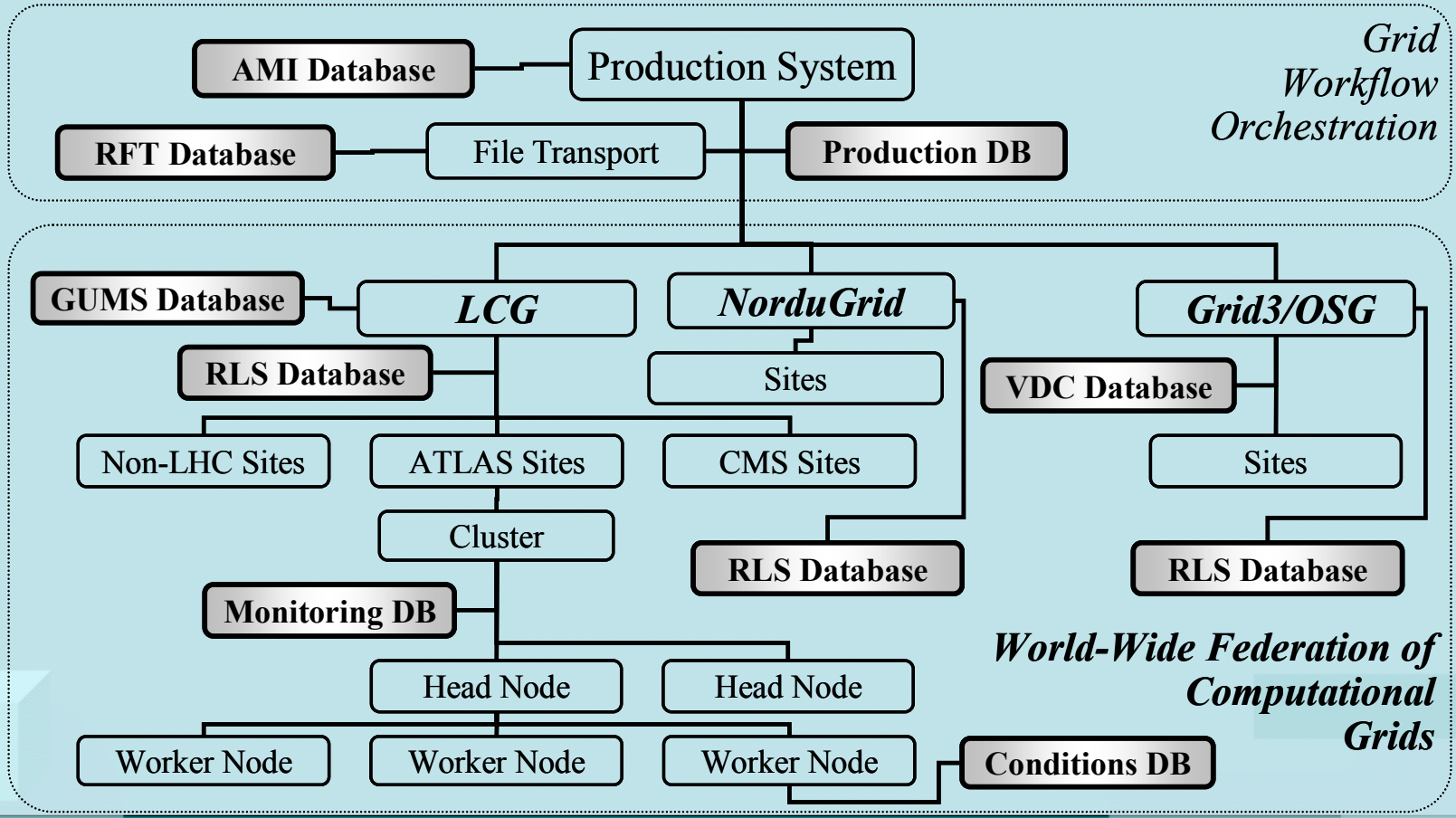


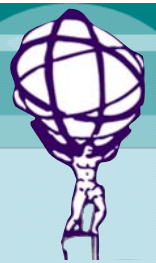
Why not a constant rate ?

Few reasons:

- Job control issue
- Software, installation
- Databases issue
- Data movement issue

Emerging Hyperinfrastructure



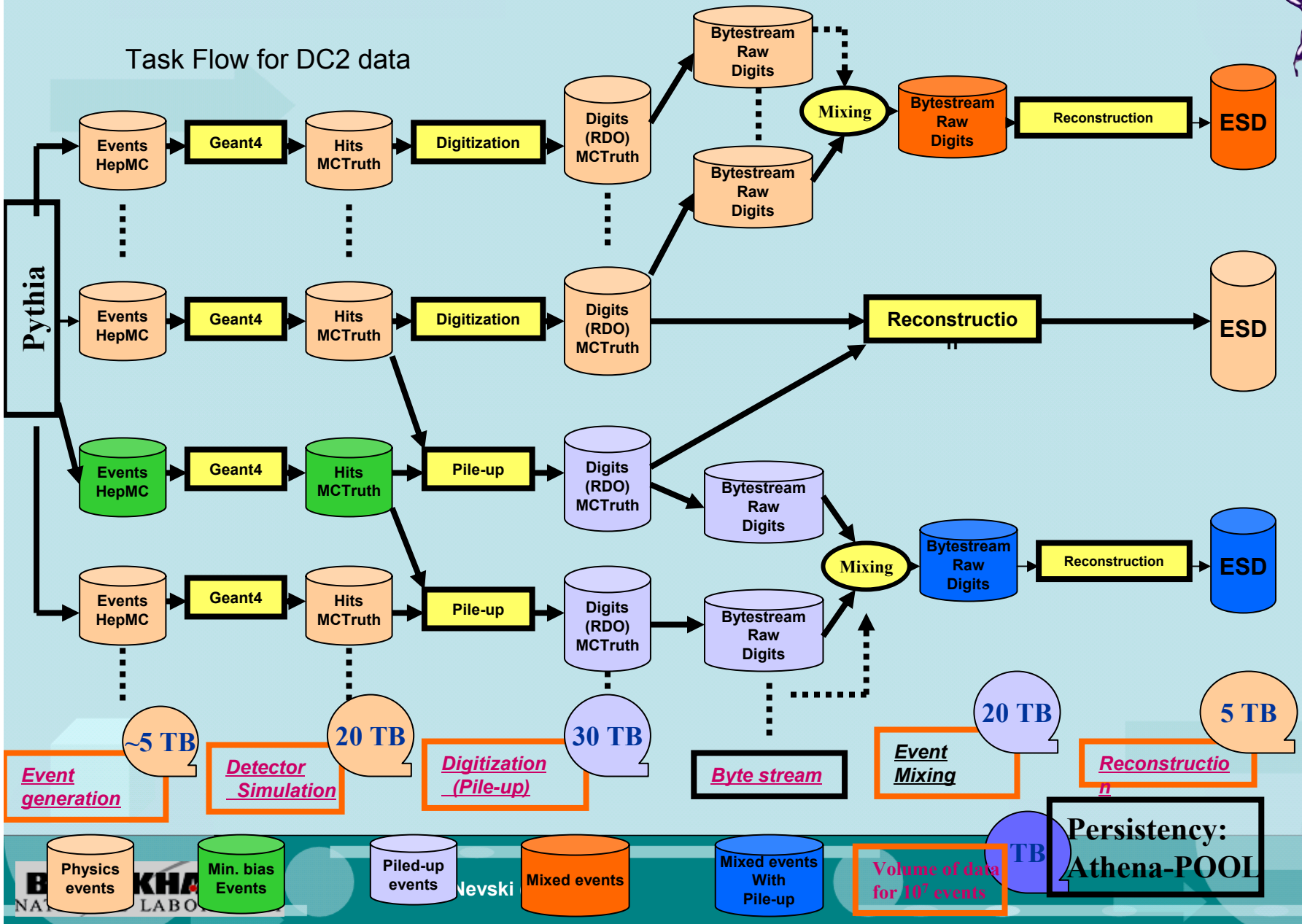


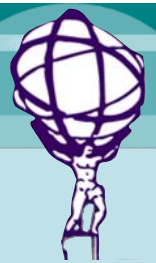
- General Production Organization

Lessons Learned (GRID3)

- Scalability
  - New problems discovered every few weeks as we increased scale
  - Needed constant interaction with software developers
  - Dependencies - solution to some problems introduced new ones
- **Testing of new releases**
  - **Need to plan for 2 week validation of new releases on grid**
  - **Required 2-8 new transformations after deployment on grid**
- **Continuous validation and physicist involvement**
  - **Some problems found after dataset completion**
  - **Many datasets aborted or restarted**
  - **Need involvement of physicists for Quality Control – thanks to Ian and Davide, this was much better for Rome**

Production phases





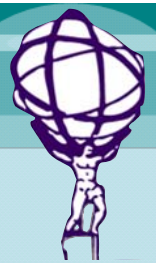
- Software and Installation Issue

List of Some Needed Improvements (GRID3)

- **Infrastructure**
  - **Automated fail-safe software installation**
  - **Not waste people on 'bad' sites**
  - **VO flexibility - ability to set different site priorities for production and other ATLAS users**
- Capone/GCE
  - Persistency
  - Resource broker
  - SE management/selection
  - Improve maximum rate

In general: everything worked

- ATLAS software problem
- The resources available via ARC are heterogeneous and not ATLAS-dedicated.
- These resources have to maintain software of different users, they need to keep the database of the installed software. This is done by using RPMs.
- **As the complexity of the ATLAS s/w increases, RPM creation becomes a full-time job, while the NorduGrid manpower is limited.**
- Result: delay in the start of the production and we lost almost half of the resources.
- **It would be a grate advantage to have the official kit packed as RPM.**



•JOB Control

- Inadequate overall control
- In case of job failure output (including Log) is not saved, no easy way to understand problems
- Jobs do not repeat their outputs
- Non repetitiveness of jobs and high failure rate may introduce *physics bias*

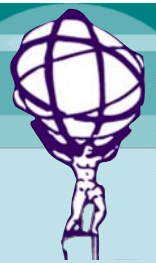
Issues: workload management

- Information schema inadequate
  - missing “per VO” information (solved by new Glue?)
- Submission rate too slow (~10 jobs/min for serial submission)
  - need to experiment with parallel (multithread) submission
- Submission rate degrades to ~1.5 jobs/min under heavy load (~4000 jobs in the system)
  - Not true if RB and LB running on different machine
- The job submission through the RB gets very slow if too many CEs are present in the BDII
  - observed once, to be investigated
- All issues under study by ATLAS and ECGI people

Experience from Rome production (NorduGrid)

- A lot of manual work is still needed:
  - Keep track of the jobs and their logs
  - interpret the error codes
  - change database entries for maxattempt or max memory etc
  - the executor had trouble with unreadable replacement characters
  - validate files on crashed storage elements
  - kill looping jobs on clusters that run Condor

- Some ATLAS related problems, e.g. with database



•Data Transfer

- Many Storage Elements had massive failure during production cycle
- DQ had a “single point of failure” (holidays), no easy way to understand its status.
- RLS catalog did not scale

Experience from Rome production (NorduGrid)

- Solved problems:

gacl problem – allowing other executor owners to read files is done by the Dulcinea.

upload failures – upload attempts are repeated with different time intervals.

Data management: badly missing, esp. when many SEs go down for e.g. maintenance, no easy way to quickly create replicas.

New people are involved in the production.

Production team: Mattias Ellert, Samir Ferrag, Farid Ould-Saada

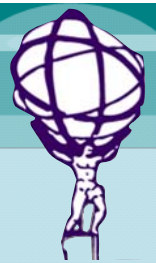
Katarina Pajchel, Alex Read, Oxana Smirnova

Issues: configuration / services

- Site related problems
  - crashing of SEs disks (Russia, Lyon, LNF, …)
  - crashing of MSS (Sinica, CNAF, …)
  - errors in NFS mounting, misconfiguration
- Network bottlenecks / server overload (too many connections on the same SE)
- RLS down (heavy load, not yet understood why)
- Jabber server down (no loss of jobs, just slowering the production)

Issues: data management

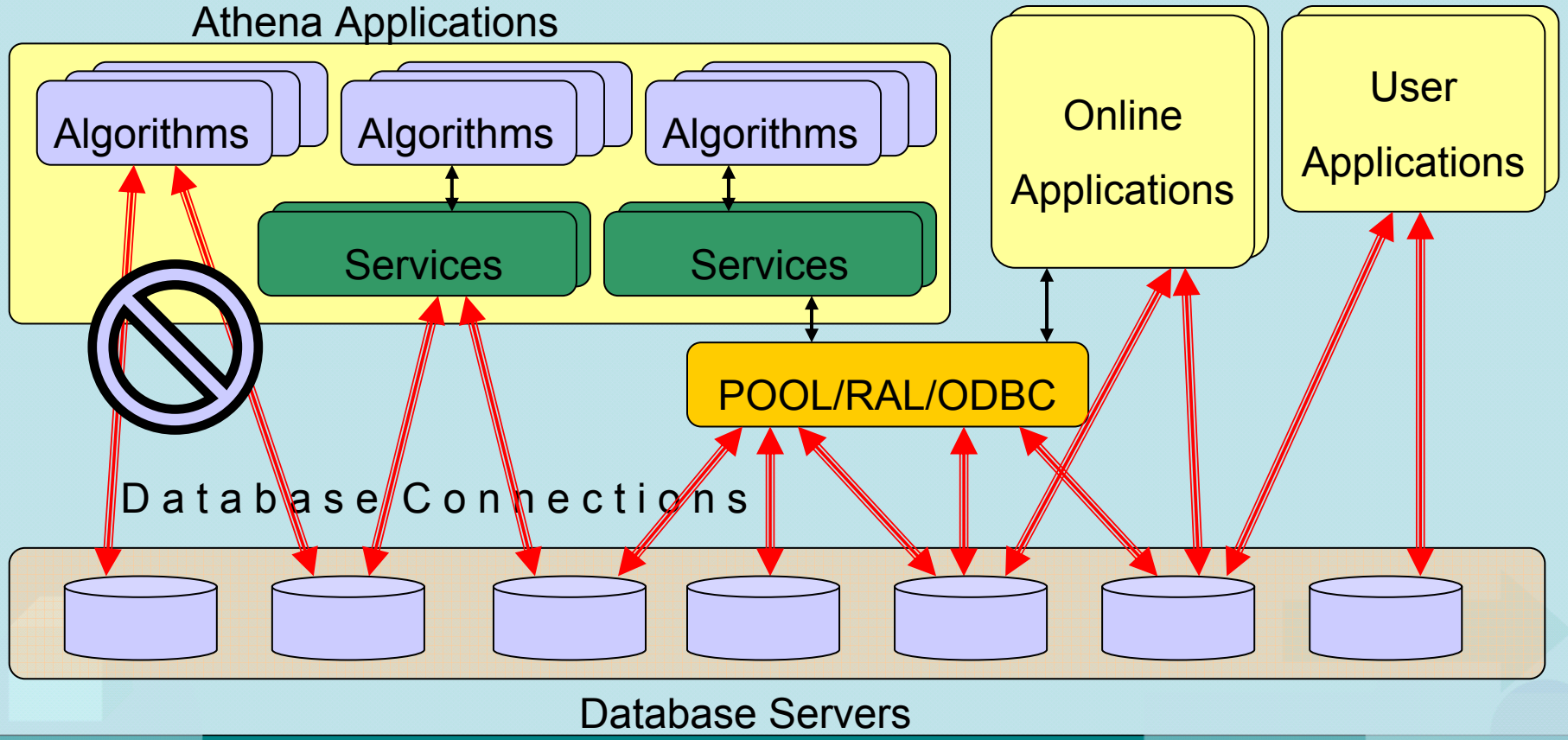
- LCG data management tools:
  - lcg-gt: calls the BDII and checks infos about a SE; if the SE is down or the BDII is down, the command crashes. Used a workaround
  - lcg-cp: if the connections times out, the command hangs. Used a workaround
  - lcr-cr: On some failures, it can leave the file catalog in an inconsistent status. Moreover it hangs if the connection times out. Used a workaround for the time-out
- Due to these inefficiencies, we should have spread the data over more sites by hands
  - Most failures due to DM problems



DATABASES

- Complex usage pattern
- Few examples of fast feedback

Complicated Access Sequence



Database Applications Involved

Every job must have the database access

The following database applications are used by release 9.0.x jobs:

Application	Athena Interface		Server Defaults		Database	Account	Transformation		
	Service	Technology	Name	Technology			simu	digi	reco
GeometryDB	RDBAccessSvc	HVS/RAL	pdb01	Oracle	ATLASDD	atlasdd_reader	x	x	x
ConditionsDB IOV	IOVDbSvc	Lisbon CondDB	atlasdev1	MySQL	LArIOVDC2	readerLArIOV	x	x	x
ConditionsDB payload	LArCondCnv	NovaBlob	atlasdev1	MySQL	LArNBDC2	reader	x	x	x

Three database applications are used by release 10.0.x jobs:

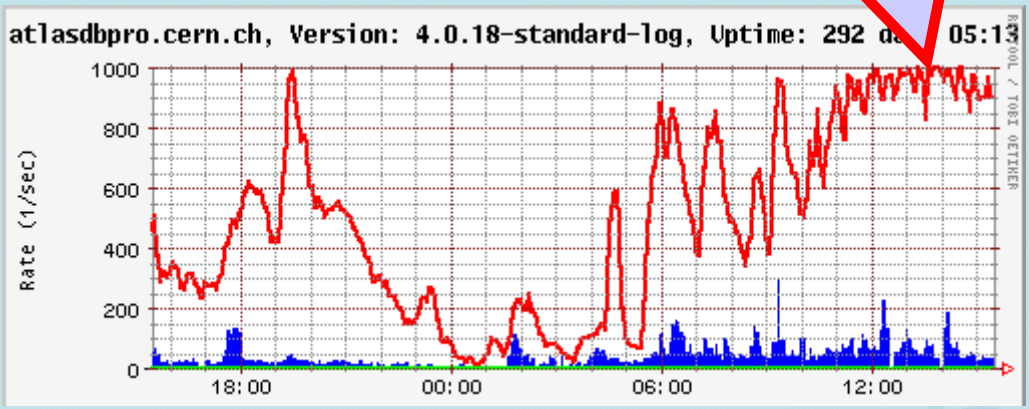
Application	Athena Interface		Server Defaults		Database	Account	Transformation		
	Service	Technology	Name	Technology			simu	digi	reco
GeometryDB	RDBAccessSvc	HVS/RAL	atlas	Oracle	ATLASDD	atlasdd_reader	x		x
ConditionsDB IOV	IOVDbSvc	Lisbon CondDB	atlasdbpro	MySQL	LArIOVDC2	readerLArIOV	x	x	x
ConditionsDB payload	LArCondCnv	NovaBlob	atlasdbpro	MySQL	LArNBDC2	reader	x	x	x

Conditions DB Bottleneck

- Fraction of the shorter digi/reco jobs increased
  - more frequent database access
- Production rates exceeded expected levels

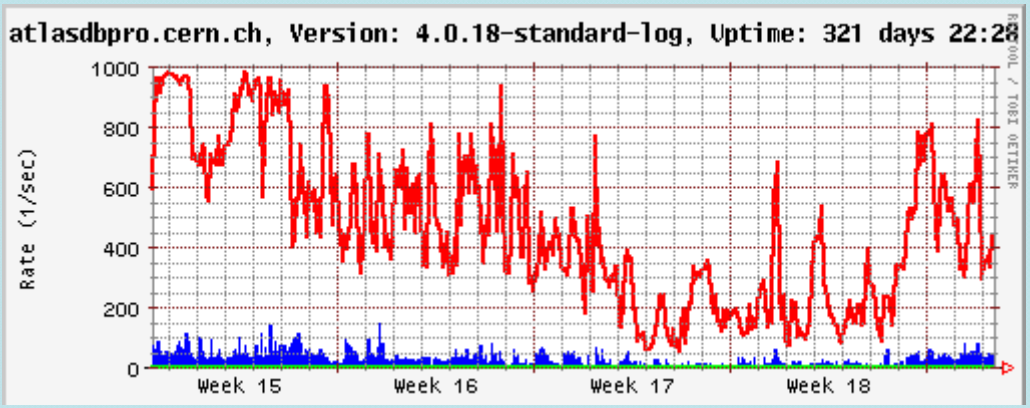
MySQL database access become a bottleneck

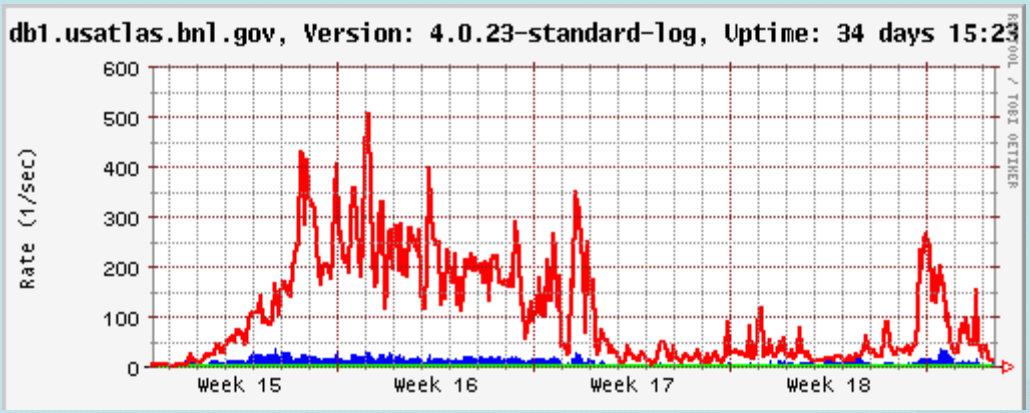
- Hard limit of a 1000 connections to a server
- No capability to switch to a replica



Bottleneck Resolved

- A significant fraction of jobs failed
- New transformation were introduced
  - *(thanks to Allessandro and Davide)*
- **atlasdbpro** offloaded to the **db1** replica
- Production bottleneck was resolved



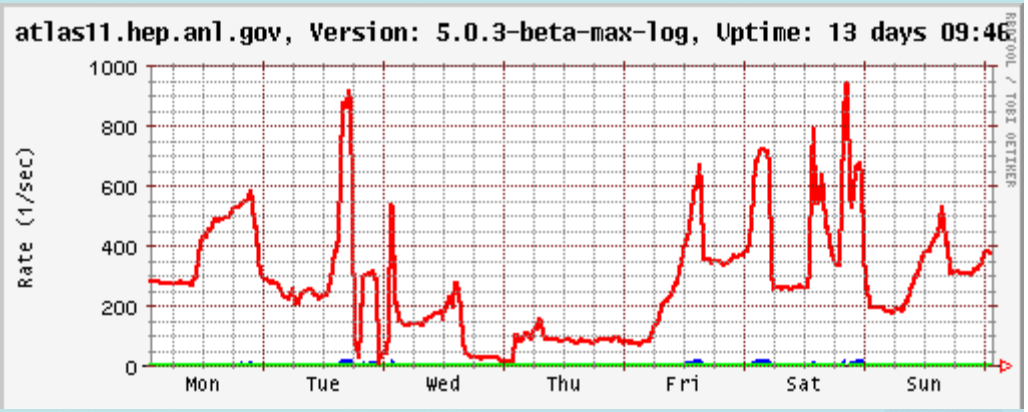


Offloading Power Users

- In addition to official Rome production an increasing number of groups and individual physicists - the "power users" - engaged in a medium scale production on their local production facilities and world-wide grids

https://uimon.cern.ch/twiki/bin/view/Atlas/PowerUsersClub

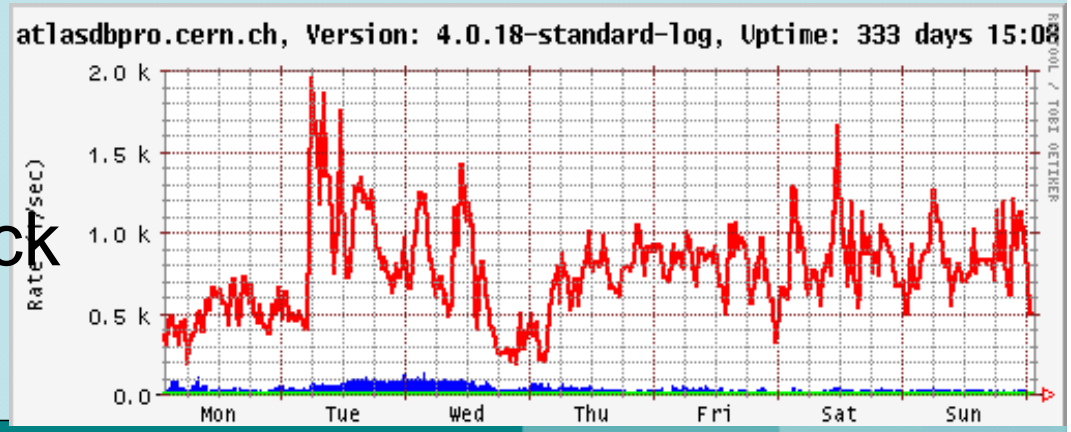
- The dedicated replica server was deployed for power users
- A significant fraction was offloaded



Overcoming Connections Count Limit

- Following discussions with MySQL performance group leader the way to overcome the 1000 connections count limit was identified, tested and deployed

- MySQL servers are no longer a production bottleneck



Oracle Server Limits

- In May increased Oracle server load from Rome production affected other experiments (CMS)
- Oracle server resource allocation for ATLAS was limited to 60% (more than a fair share)
- The technical implementation of resource limitation was introduced unexpectedly and affected a significant fraction ATLAS jobs at the beginning of May
- Improvements in CERN database services were recommended

Servers List

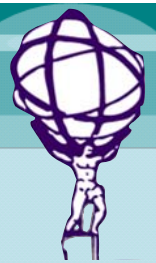
- Database servers list is now available on ATLAS Wiki page
- Please add your server to the list

Server	Location	Technology	Address	Port	Comment	Database Applications								
						Geometry		Conditions		File Catalog		Collections		DB Catalog
						HVS	NOVA	IOV	Nova	POOL	RLS	MySQL	RAL	
pdb01 atlas	CERN	Oracle9i	ask	default	2-node cluster no load balancing	x								
devdb	CERN	Oracle9i	ask	default	for development	x							x	
devdb10	CERN	Oracle10g	ask	ask	for development	x								x
cooldev	CERN	Oracle10g	ask	ask	COOL validation			x						
coolpro	CERN	Oracle10g	ask	ask	for COOL								x	
intdb10g	CERN	Oracle10g	ask	ask	for integration								x	
dbdevel1	BNL	MySQL 4.0	ask	default	ANSI-compliant	x								
atlmysql01	CERN	MySQL 4.0	ask	default	for CTB			x	x					
atlobk02	CERN	MySQL 4.0	ask	default	for CTB		x	x	x	x				
pcatm020	CERN	MySQL 3.23	ask	default	CTB muon only			x						
atlasdev1 atlasdbdev lxfs6131	CERN	MySQL 4.0	ask	default	deprecated alias actual name		x	x	x					
atlasdbpro lxfs6031	CERN	MySQL 4.0	ask	default	 actual name		x	x	x					
lxfs6021	CERN	MySQL 4.0	ask	default	behind firewall							x		
db1	BNL	MySQL 4.0	ask	default			x	x	x					
adbpro	BNL	MySQL 4.1	ask	default	2+2-node cluster					x				x
adbpro01	BNL	MySQL 4.1	ask	default	ANSI-compliant									
adbpro02	BNL	MySQL 4.1	ask	default	ANSI-compliant									
atlaspc4	U Montreal	MySQL 4.0		default	behind firewall			x	x					
in20	U Montreal	MySQL 4.0		default	behind firewall	x								
atlasdb	IJS	MySQL 4.0	ask	default	DC2 replica		x	x	x					
acdc	U Buffalo	MySQL 4.0	10.1.1.132	default	DC2/firewall		x	x	x					
mcfarm2	SMU	MySQL 4.0	ask	default	DC2 replica		x	x	x					
mcfarm	SMU	MySQL 4.0	ask	default	Rome replica			x	x					
mcfarm	SMU	MySQL 4.0	ask	ask	ANSI-compliant	x								
atlas10	ANL	MySQL 4.0	ask	default	grid-enabled									
atlas11	ANL	MySQL 5.0	ask	default	for early adopters			x	x					
atlas12	ANL	MySQL 5.0	ask	default	for Rome pileup			x	x					
lxn1190	CERN	MySQL 4.0	ask	default	DDM development					x				
lxshare070d	CERN	MySQL 4.0	ask	default	POOL project					x		x		

BROOKHAVEN NATIONAL LABORATORY

Roadmap to Redundancy

- Central Deployment:
  - Most of ATLAS current experience in production
    - Scalability problem: advanced planning for capacities required
    - Remote site firewall problem
- Replica deployment on Worker Node:
  - Extensive experience in ATLAS Data Challenge 1
    - Replica update problem
- Replica deployment on the Head Node (gatekeeper):
  - Use of grid tools to deploy database server replica
  - Proof-of-the-principle deployment performed
    - ***(thanks to Yuri Smirnov)***
  - **Details are in the talk at the Data Management session on Thursday**



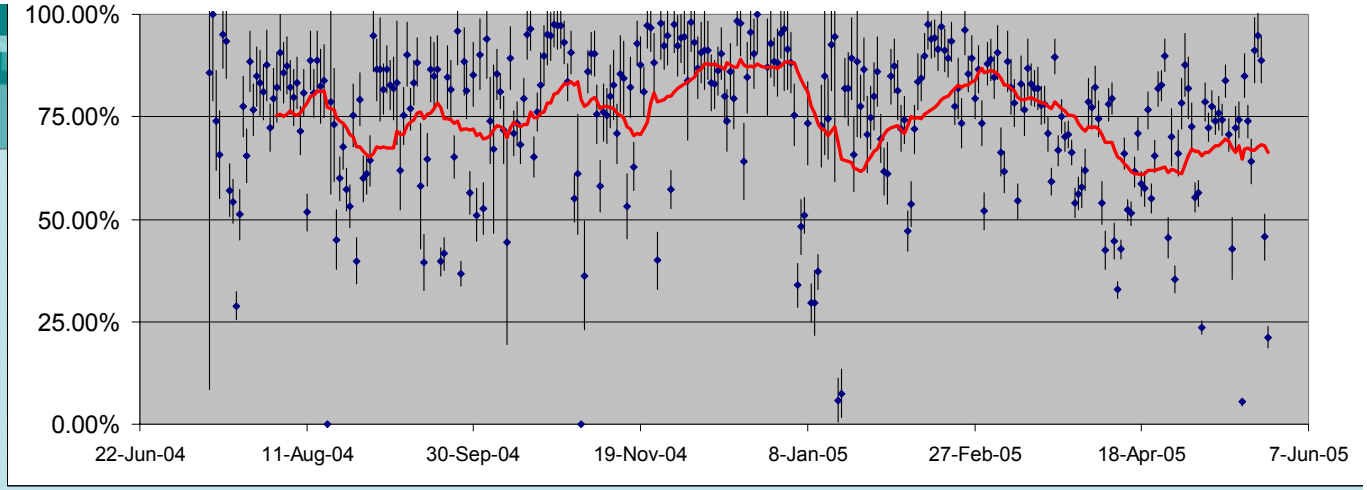
- Other Issues

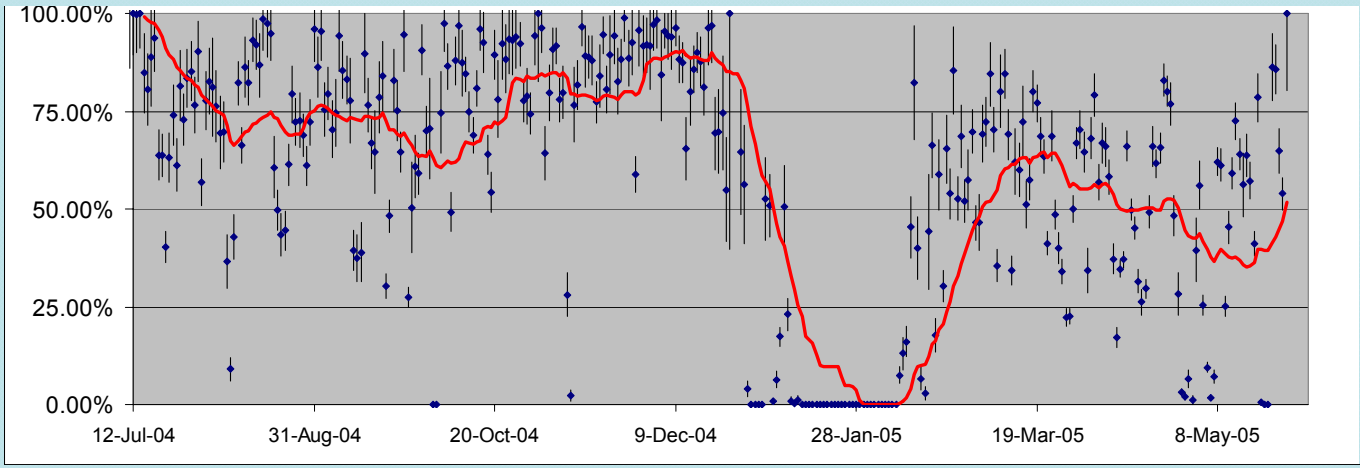
Production efficiency

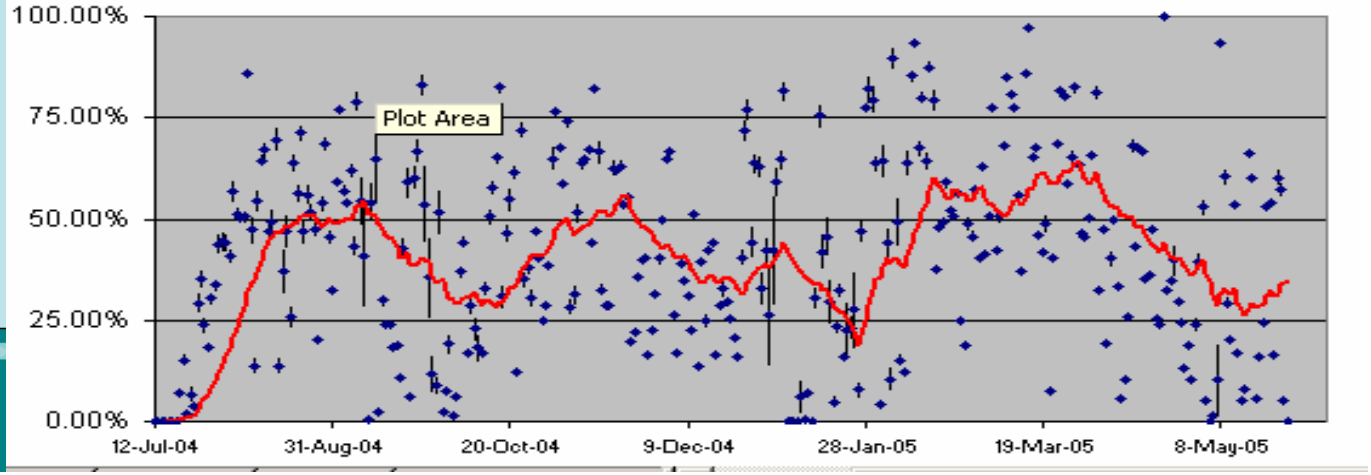
Depends on many factors….

GRID3 made most of the testing for Rome production

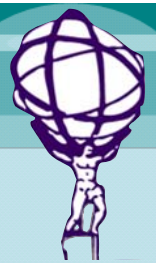
NG had personel change between DC2 and Rome …







BROOKHAVEN NATIONAL LABORATORY



TIER0 problems

- No mechanism exists to balance GRID and local load, TIER0 exercise was partially blocked by the Rome production
- RLS catalog is a real bottleneck, no easy way to get full list of files in a storage element
- Need a special treatment for analysis disc space (Castor, Dcache ?)

Conclusion (GRID3)

- Impressive success in spite of many problems – thanks to hard working Grid3 production team
- Need to provide feedback to software developers after Rome workshop – but not wait too long!
- Need improvements of ALL systems and software
- Scalability issues will only be discovered as we scale up
- Need continuous challenges from now till 2007

Some Frequent Recommendations

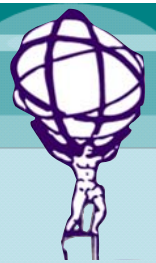
- Event generation should be done in the common framework (with general bookkeeping and more QA control)
- Generic tools (GRID flavor independent) are needed to monitor production status
- Do not abandon failed jobs,

  make all steps reproducible

  to avoid physics bias

Conclusion

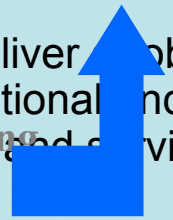
- General – better planning (software readyness, evgen production, QA)
- Job submission/control – has to be significantly improved (Eowyn, Windmill, monitoring…)
- Data management – became critical, needs more efforts
- ATLAS DBs – good progress, although a lot of work ahead
- Software installation – can be improved
- Better communications with Tier0/1….



Back-up slides

The 3 Grid flavors

- LCG (http://lcg.web.cern.ch/LCG/)
  - The job of LHC Computing Grid Project - LCG - is to prepare the computing infrastructure for the simulation, processing and analysis of LHC data for all four of the LHC collaborations. This includes both the common infrastructure of libraries, tools and frameworks required to support the physics application software, and the development and deployment of the computing services needed to store and process the data, providing batch and interactive facilities for the worldwide community of physicists involved in LHC.
- Grid3 (http://www.ivdgl.org/grid2003/)
  - The Grid3 collaboration has deployed and international Data Grid with dozens of sites and thousand of processors. The facility jointly by the US Grid project iVDGL, GriPhyN and PPDG and the US participants in the LHC experiments ATLAS and CMS.
- NorduGrid (http://www.nordugrid.org/)
  - The aim of the NorduGrid collaboration is to deliver a robust, scalable, portable and fully featured solution for a global computational and data Grid system. NorduGrid develops and deploys a set of tools and services - the so-called ARC middleware, which is a free software.



- Both Grid3 and NorduGrid have similar approaches using the same foundations (GLOBUS) as LCG with slightly different middleware

DC2 production phases

Process	No. of events	~ Event size	~ CPU time per event	~ Volume of data
		MB	kSI2k-s	TB
Event generation	5×10^7	0.06		3
Simulation	10^7	2.	520	30
Pile-up Digitization	3×10^6 10^7	3. 2.	150 15	6 20
Event mixing & Byte-stream	10^7	2.	5.4	20

- The simulation part was finished by the end of September and the pile-up and digitization parts by the end of November
- 10 million events were generated, fully simulated and digitized and ~2 Million events were “piled-up”
- Event mixing and reconstruction was done for 2.4 Million events in December.
- The Grid technology as provided the tools to perform this “massive” worldwide production

Lessons learned from DC2

- Main problems
  - The production system was in development during DC2
  - The beta status of the services of the Grid caused troubles while the system was in operation
    - For example the Globus RLS, the Resource Broker and the information system were unstable at the initial phase
  - Specially on LCG, lack of uniform monitoring system
  - The mis-configuration of sites and site stability related problems
  - But also
    - Human errors (for example “expired proxy”; bad registration of files)
    - Network problems (connection lost between two processes)
    - Data Management System problems (eg. connection with mass storage system)

Lessons learned from DC2

- Main achievements
  - To have run a large scale production on Grid ONLY, using 3 Grid flavors
  - To have an automatic production system making use of Grid infrastructure
  - Few “10 TB” of data have been moved among the different Grid flavors using DonQuijote (ATLAS Data Management) servers
  - ~260000 jobs were submitted by the production system
  - ~260000 logical files were produced and ~2500 jobs were run per day

Conclusions

- The generation, simulation and digitization of events for ATLAS DC2 have been completed using 3 flavors of Grid Technology (LCG; Grid3; NorduGrid)
  - They have been proven to be usable in a coherent way for a real production and this is a major achievement
- This exercise has taught us that all the involved elements (Grid middleware, production system, deployment and monitoring tools, …) need improvements
- From July to end November 2004, the automatic production system has submitted ~260000 jobs,

Ian: “Best laid plans of mice and men…”

3 November 2005: I suggest this:

- Start Generation November 30
- Start Simulation December 15

  Note that simulation can start before all samples are generated
- Expect simulation to be validated by end of November
- Complete background simulation 1 February 2005
- Start generation of group samples January 2005

  or earlier if private resources are available

  BUT simulation did not start until January and only in “calorimeter mode”, 9.0.4 was 1 month late

Revised plan: January 20 and February 16

- Descoped to 5M events, add AOD->Tag

Production efficiency – Human factor ?

Depends on many factors….

GRID3 made most of the testing for Rome production

NG had personel change between DC2 and Rome