ATLAS DC2 and Rome Production Experience

LCG Database Deployment and Persistency Workshop

October 19, 2005
CERN, Geneva, Switzerland

Alexandre Vaniachine, David Malon (ANL)

ARGONNE NATIONAL LABORATORY
U.S Department of Energy
Office of Science Department of Energy
THE UNIVERSITY OF CHICAGO

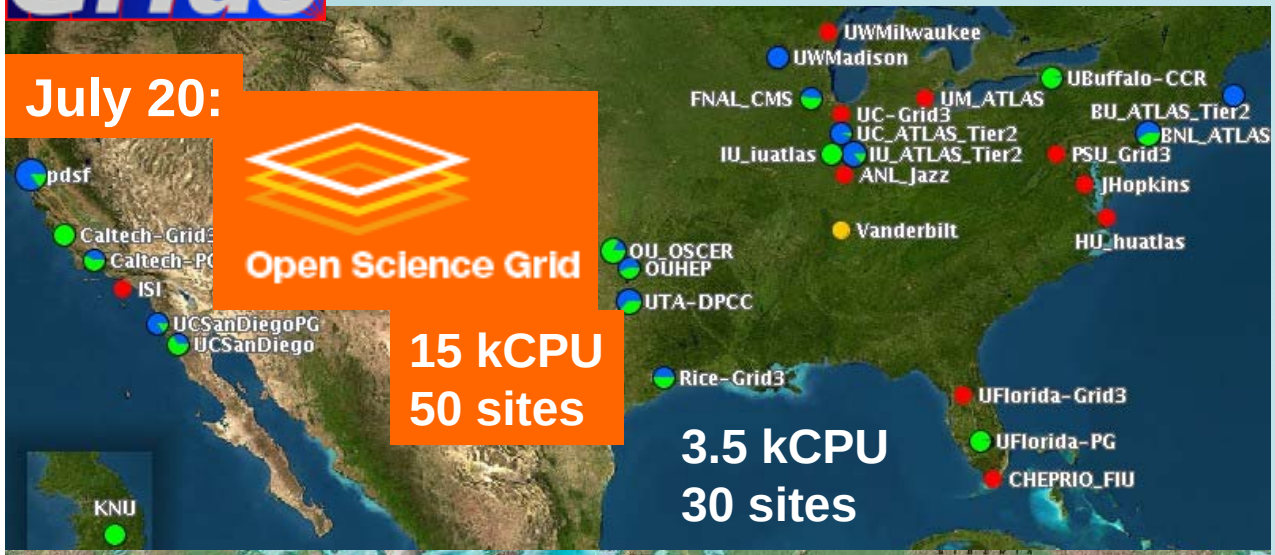
Outline and Introduction

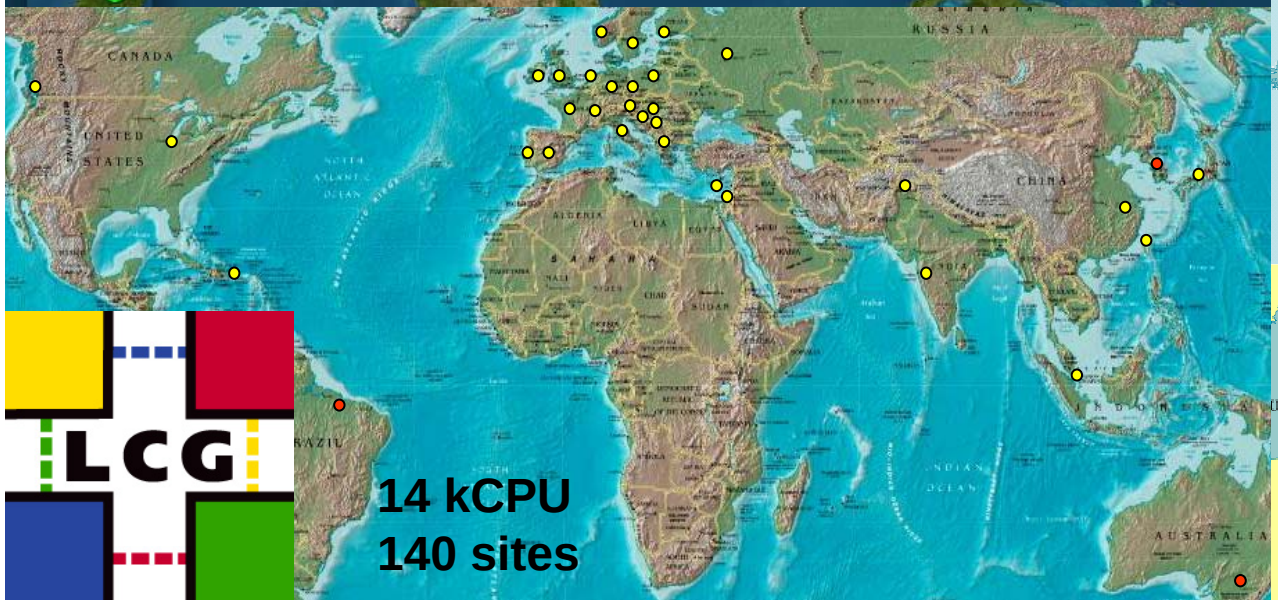
- ATLAS Experience:
  - Increased fluctuations in server load on the grid
    - Requires on-demand DB services deployment
- ATLAS Solutions:
  - Client-side: database client library
  - Server-side: on-demand grid-enabled DB services

Introduction

- ATLAS productions run on a world-wide federation of computational grids harnessing the power of more than twenty thousand processors

ATLAS Federation of Grids

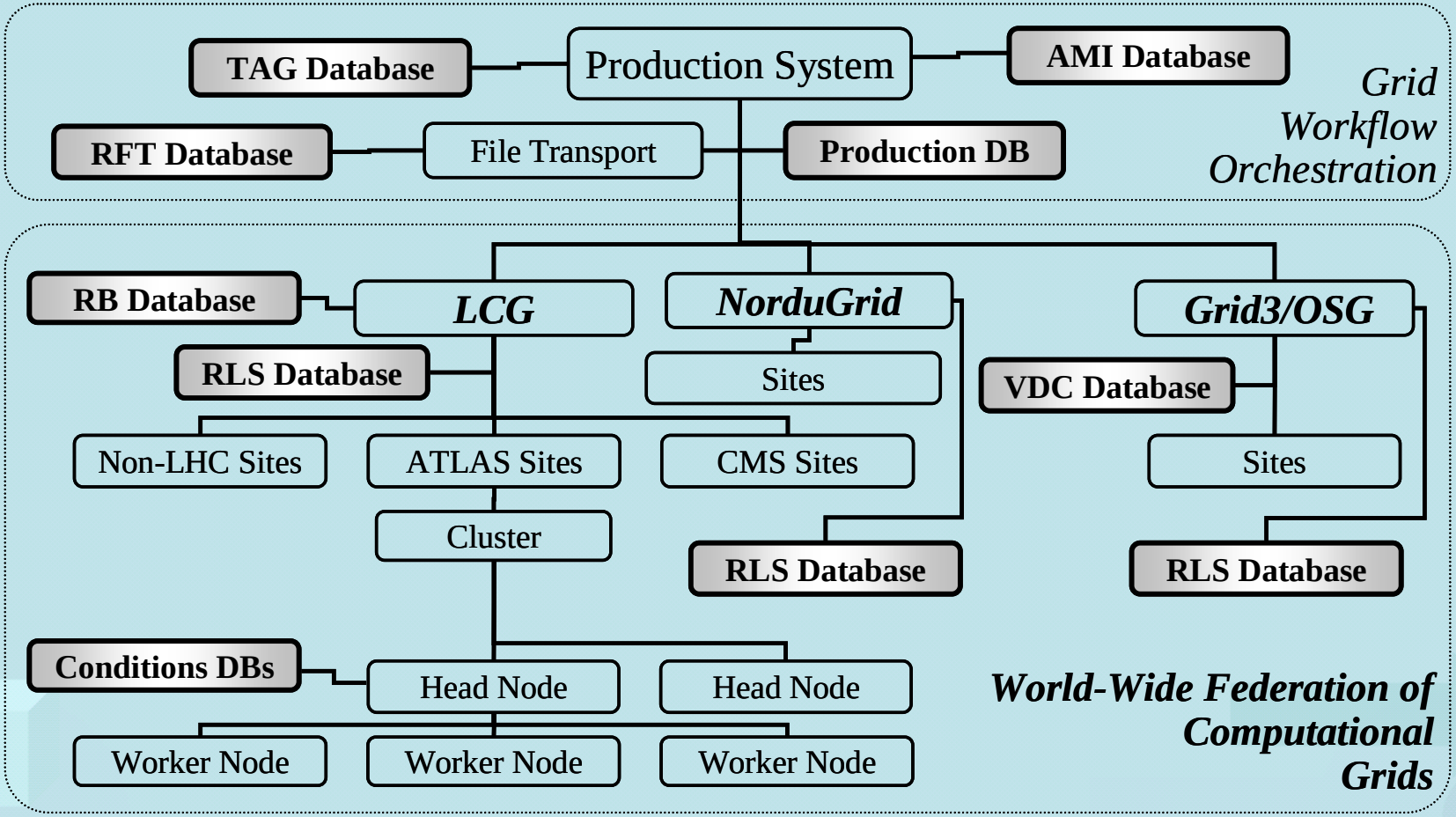




Data Workflow on the Grids

- An emerging hyperinfrastructure of databases on the grid plays the dual role: both as a built-in part of the middleware (monitoring, catalogs, etc.) and as a distributed production system infrastructure orchestrating scatter-gather workflow of applications and data on the grid
- To further detail the database-resident data flow on the grids ATLAS Data Challenges exercise the Computing Model to identify potential bottlenecks

Databases and the Grids



Expectations and Realities

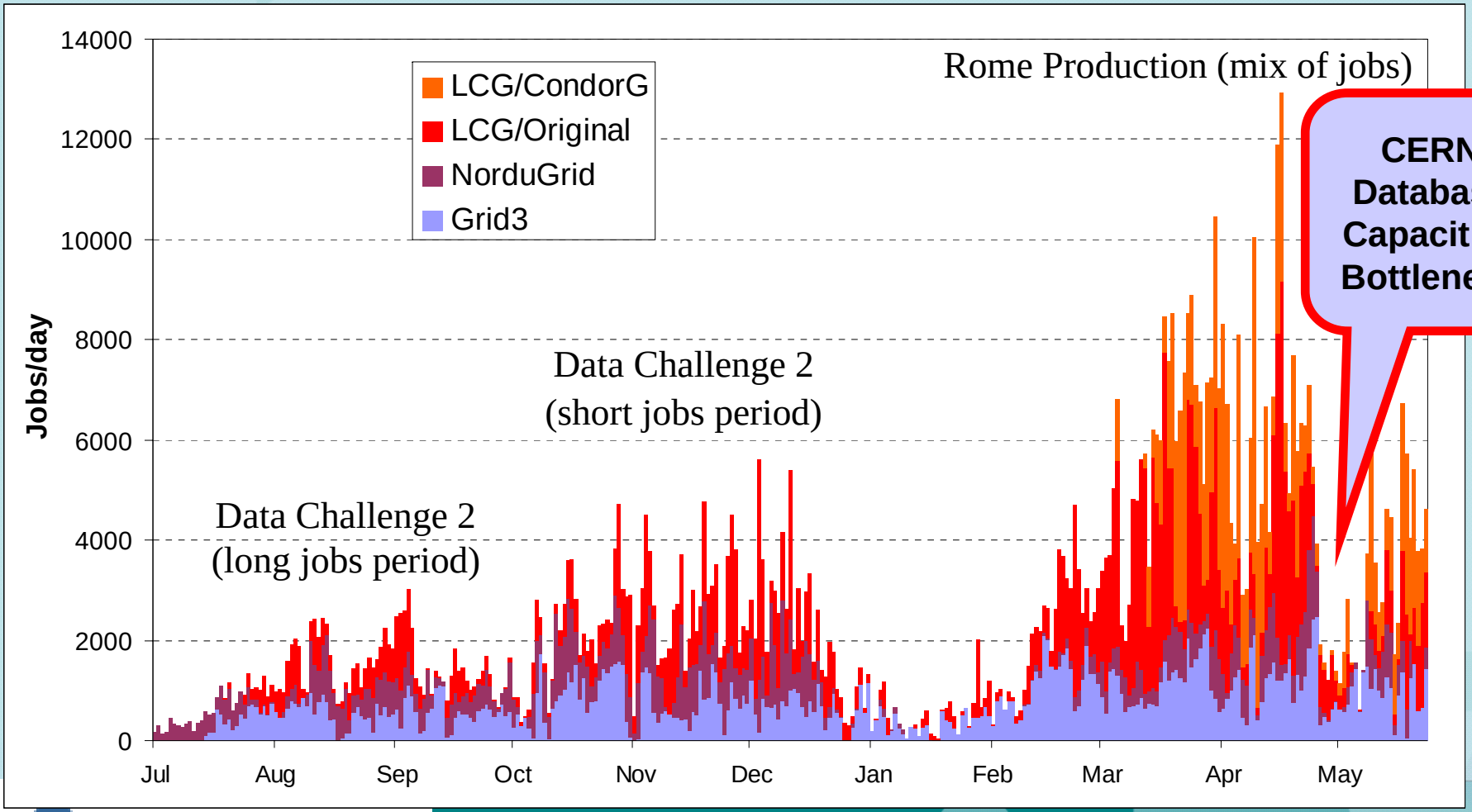
Expectations

- Scalability achieved through replica servers deployment
- Database replicas will be deployed down to each Tier1 center
- To ease an administration burden the database server replica installation was reduced to a one line command
- The needed database services level is known in advance

Realities

- Only the centers that has problems with access to the central databases (because of firewalls or geographical remoteness resulting in low data throughput) deployed the replica servers
- Difficulties in switching to the replica servers deployed
- Difficult to know the needed database services level in advance

Production Rate Growth

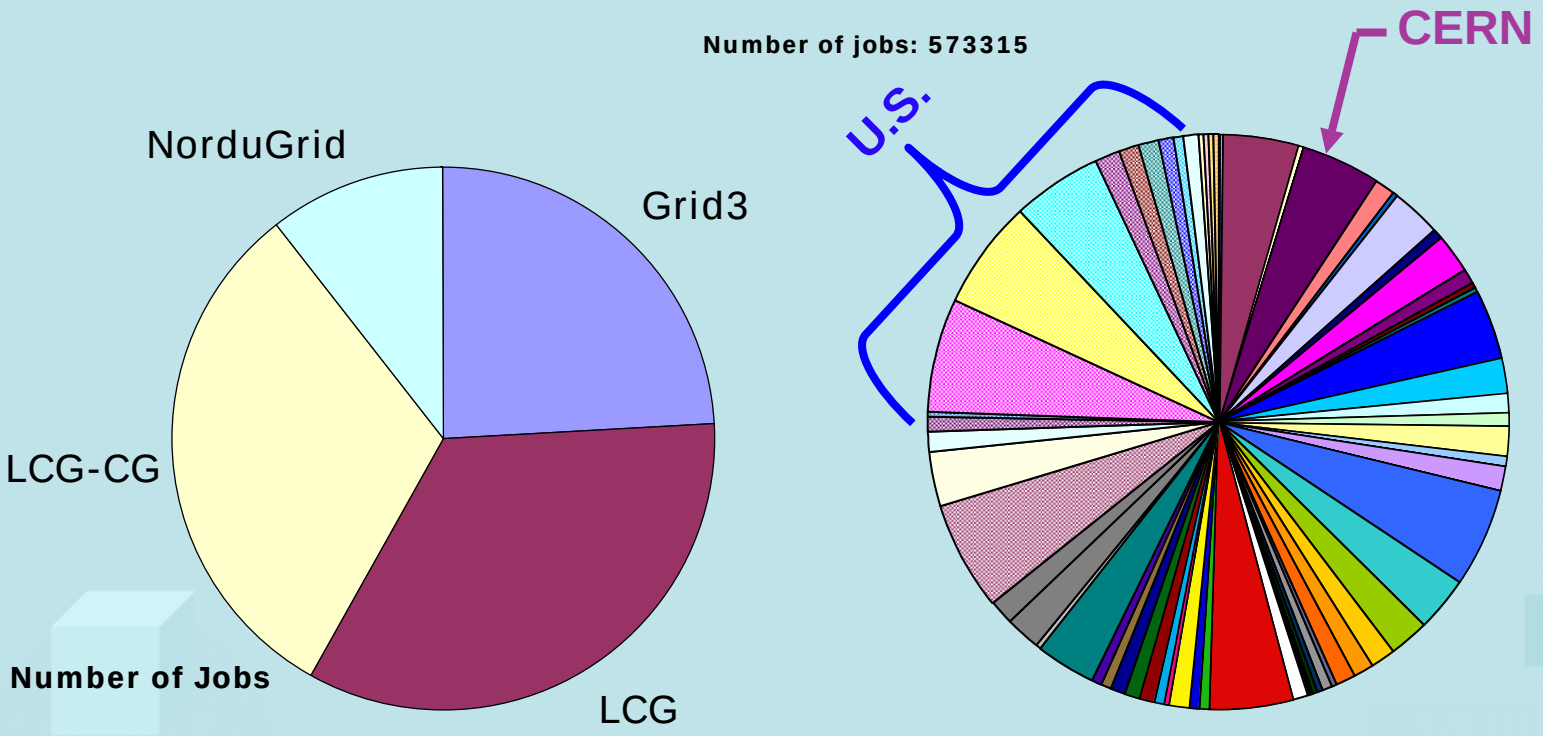


First Lessons Learned

- Most of the jobs request the same data
  - server-side in-memory query caching eliminates hardware load
- Database workload depends on the mix of jobs
- With short jobs (reconstruction) the load increased
  - remote clients failed to connect because of their local TCP/IP socket timeouts settings
    - Similar to the Denial-of-Service attack condition
  - the world-wide spread of production (see next slide)
    - Makes impractical to the TCP/IP socket timeouts adjustments
  - The only solution is to deploy even more servers

Spread of Grid Production

ATLAS Rome Production: 84 sites in 22 countries



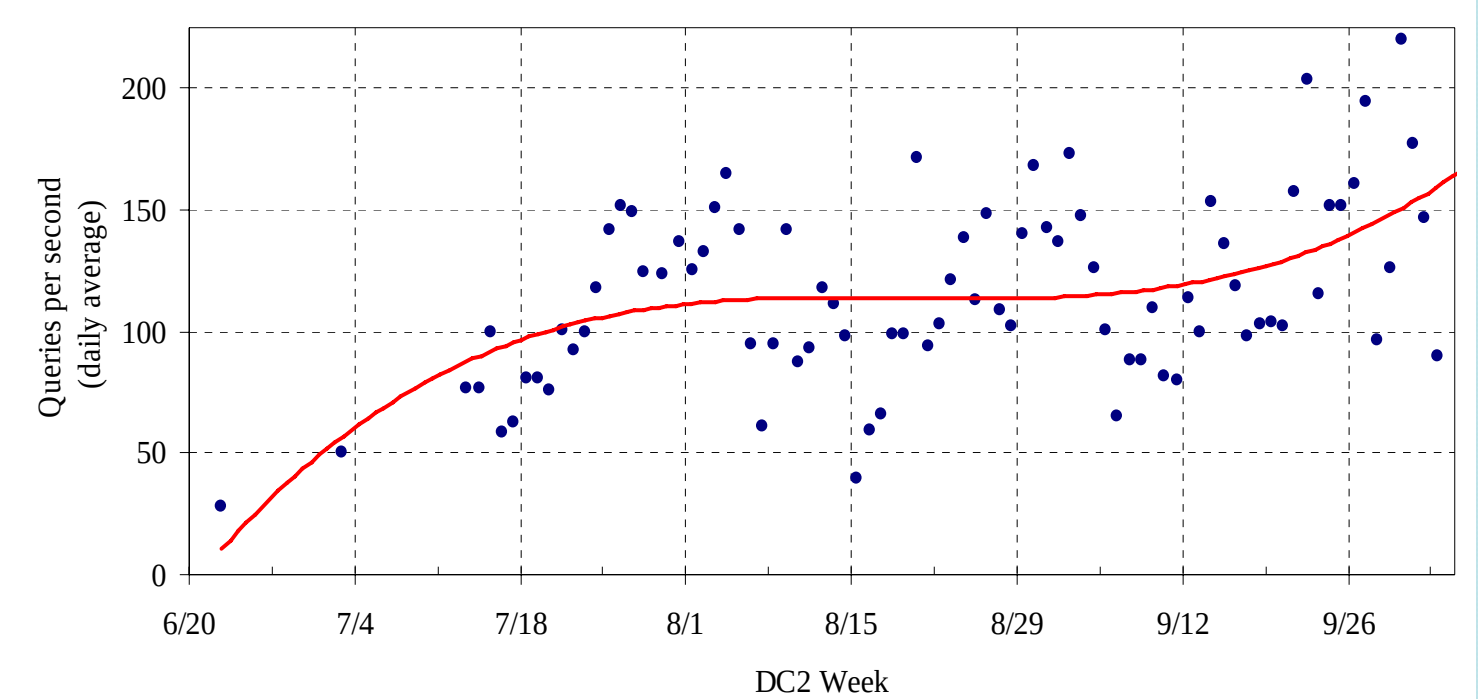
Data Mining the Experience

- The data-mining of the collected operations data reveals a striking feature – a very high degree of correlations between the failures:
  - if the job submitted to some cluster failed, there is a high probability that a next job submitted to the cluster would fail too
  - if the submit host failed, all the jobs scattered over different clusters will fail too
- Taking these correlations into account is not yet automated by the grid middleware
- That is why production databases and grid monitoring data that are providing immediate feedback on the Data Challenge operations to the production operators is very important for efficient utilization of the grid capacities

Increased Fluctuations

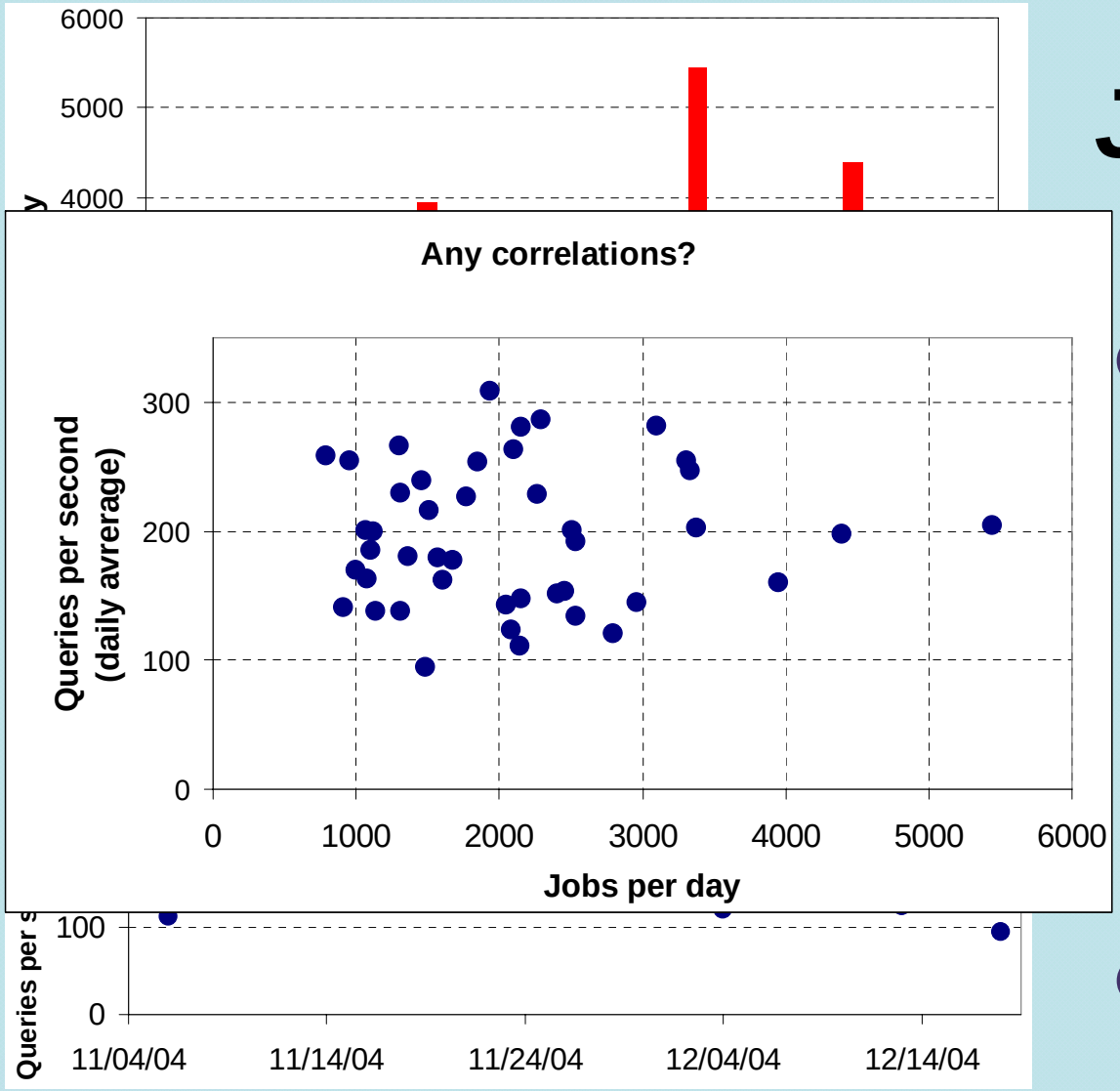
- Among the database experience issues encountered is the increase in fluctuations in database servers' workloads due to the chaotic nature of grid computations
- The observed fluctuations in database access patterns are of general nature and must be addressed by grid-enabling database services

Scalability Challenge



The chaotic nature of Grid computing increases fluctuations in demand for database services: 14×statistical

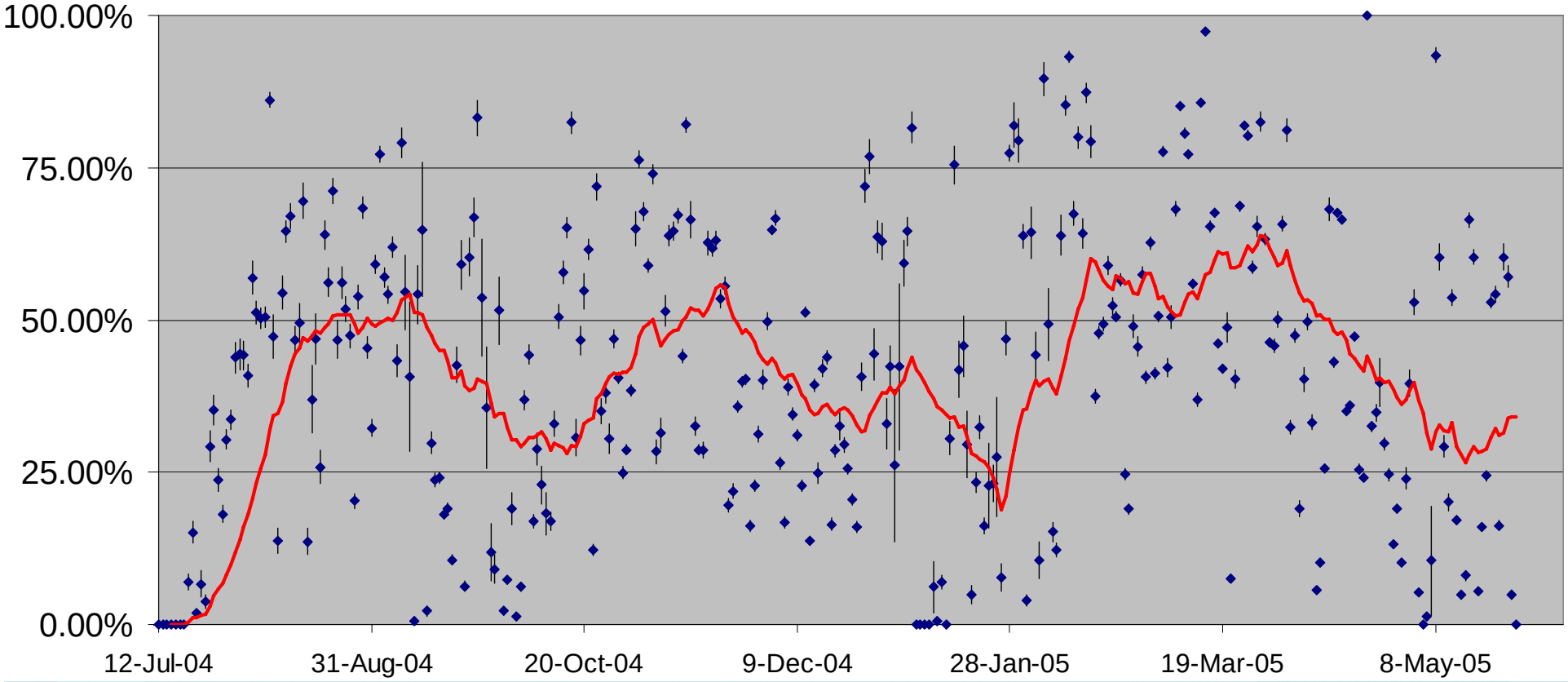
- Database services capacities should be adequate for peak demand



Jobs Load

- No apparent correlations between jobs and database server load fluctuations
- Job failures?

ATLAS Efficiency on LCG



- Efficiency fluctuates because clusters are “large error amplifiers”?

Power Users

- The organized Rome Production was for the ATLAS biannual Physics Workshop in Rome
  - Preparations for the Rome Workshop expedited large growth in ATLAS software use by physicists
  - An increasing number of groups and individual physicists - the Power Users - engage in a medium scale production on their local production facilities and world-wide grids
- Empowered by grids, occasionally, ATLAS Power Users created bottlenecks on ATLAS and/or CERN/IT shared database resources
- To coordinate database services workload ATLAS created the Power Users Club:
  - http://uimon.cern.ch/twiki/bin/view/Atlas/PowerUsersClub
- Is it only in ATLAS we got the Power Users – how do other LHC experiments deal with chaotic workload from physicists' activities?

Server Indirection

- One of lessons learnt in ATLAS Data Challenges is that the database server address should NOT be hardwired in data processing transformations
- The logical-physical indirection for database servers is now introduced in ATLAS
  - Similar to the logical-physical file Replica Location Service indirection of the grid file catalogs
- Logical-Physical resolution can be done against the local file updated from Catalogue on configurable time period or directly against the Catalogue residing on Oracle DB

Client Library

- To improve robustness of database access in a data grid environment ATLAS developed the application-side solution: database client library
  - indirection, etc
- *ConnectionManagement* module (***by Yulia Shapiro***) is adopted by POOL/CORAL

ATLAS Database Project

DISTRIBUTED DATABASE SERVICES CLIENT

Authors:	Yulia Shapiro, Alexandre Vaniachine (editor), Torre Wenaus
Date:	July 8, 2004
Project:	ATLAS Database
Activity:	11. Distributed Database Services
Document link:	http://atlas.web.cern.ch/Atlas/GROUPS/DATABASE/project/services/client.pdf

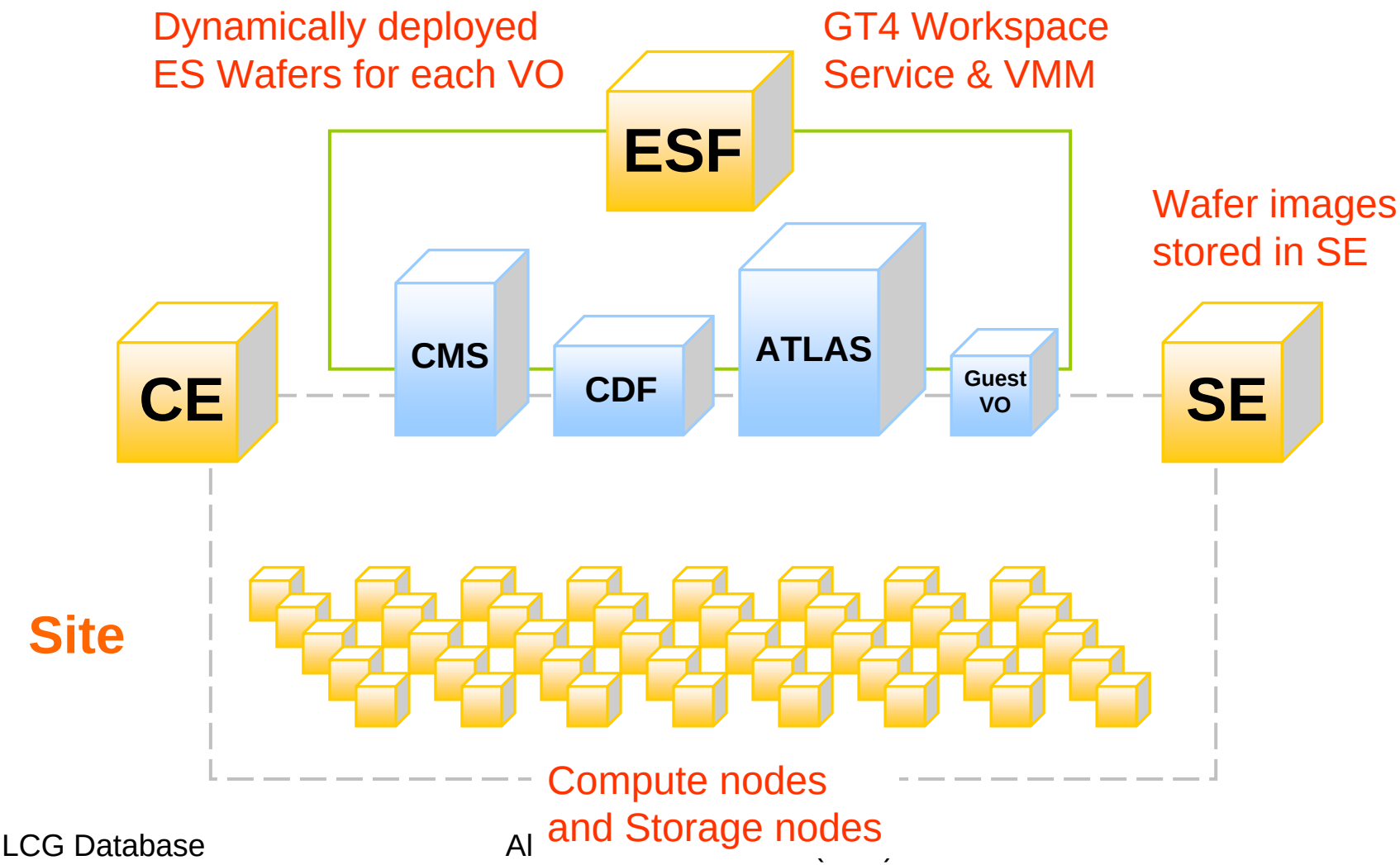
Abstract: This document defines the database client library software layer for distributed database services access in ATLAS Database Project. The Project plan prioritizes rationalization and cleanup of how server specification is done in applications which access database servers. The client library implements a consistent strategy for database server access. The Distributed Database Services client library serves as a unique layer for enforcing policies, following rules, establish best practices and encode logic to deliver efficient, secure and reliable database connectivity to applications in a heterogeneous distributed database services environment. This document collects requirements, outlines architecture and the workplan. The implementation responsibilities are also discussed.

Deployment Experience

- Central Deployment:
  - Most of ATLAS current experience in production
    - 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:
  - Use of grid commands to deploy database server replica
    - Requires dedicated Head Node allocation
- ➲ The new most promising option:
  - on-demand Edge Services Framework from OSG (see next slide)

Snapshot of ES Wafers implemented as Virtual Workspaces

ESF - Phase 1



LCG Database Deployment and

AI

Edge Services Framework

- ATLAS is involved in ESF OSG Activity:
  http://osg.ivdgl.org/twiki/bin/view/EdgeServices
  - The objective for the next month is to deploy a prototype of ESF, dynamically deploying a workspace configured with a real ESF service **and** have it run in a scenario in which the Edge Service will normally be used
- The first ATLAS grid-enabled application scheduled for deployment:
  - mysql-gsi-0.1.0-5.0.10-beta-linux-i686 database release built by the DASH project:
    http://www.piocon.com/ANL_SBIR.php

Replication Experience

- Since the 3D Kick-off Workshop ATLAS is using JDBC Octopus replication tools with increasing confidence
  http://uimon.cern.ch/twiki/bin/view/Atlas/DatabaseReplication
- With ATLAS patches Octopus Replicator now reached the production level
  - Used routinely for Geometry DB replication from Oracle to MySQL
    - Replication to SQLite was done recently with ease
- Octopus allows easy configuration of the datatype mapping between different technologies which is important for the POOL/RAL applications
- Last month Octopus was used for the large-scale Event TAGS replication of the Rome Physics Workshop data within CERN and Europe
  - Replication from CERN Tier0 to US ATLAS Tier1 and BNL was also accomplished successfully but was slower
- We have found that replication of the data on a large-scale and over large distances will benefit from the increased throughput achieved by improvements on the transport-level of MySQL JDBC drivers

POOL Collections: Setup

Data

- 38 datasets, 2.3M events from ATLAS Rome Workshop
- ROOT collections (1000 events each) generated from AOD
- ROOT collections merged into 38 collections, 1 per dataset
- Merged collections created as tables in Oracle and MySQL
- 38 collections merged into 2 master collections (1 indexed)
- Total size of 40 collections in Oracle ~ 3*(7.2 GB) ~ 22 GB

Databases

- Oracle collections used RAL collection schema and dedicated COOLPROD integration server (configured by CERN IT/PDS).
- MySQL collections used non-RAL schema and dedicated lxfs6021 server configured by ATLAS.

Slide by Kristo Karr

POOL Collections: Replication

Tools

- All collection merging and replication performed via the POOL collections utility: CollAppend

Rates

- Merge of ROOT Collections from Castor (CERN)
  ROOT ->Oracle: 49 Hz ROOT->MySQL: 85 Hz.
- Replication of Collection to Same Database (CERN)
  Oracle->Oracle: 79 Hz MySQL->MySQL: 127 Hz.
- Replication of Collection to Different Database (CERN)
  Oracle->MySQL: 67 Hz MySQL->Oracle: 62 Hz.

Also measured replication to MySQL server at US ATLAS Tier1 at BNL

- Replication of Collection to Different Database (CERN to BNL)
  MySQL->MySQL: 8 Hz
- Replication of Collection to Different Database (CERN from BNL)
  MySQL->MySQL: 89 Hz

Notes:

- Timings are without bulk insertion, which improves rates by a factor of 3-4 according to preliminary tests (no cache size optimization).
- Internal MySQL→MySQL rates are 713 Hz, internal Oracle→Oracle 800 Hz.
- Octopus and CollAppend rates agree.

Slide by Kristo Karr

Feedback on the Current Services at CERN Tier 0

- Almost a year ago the IT DB Services for Physics section was combined with Persistency Framework POOL
  - A very positive experience through the whole period starting with oracle.support@cern.ch mechanisms and now using the physics-database.support@cern.ch
  - Built strong liaisons through Dirk Geppert
  - Various Oracle issues were resolved successfully
  - All LHC experiments should be serviced on equal footing
  - A long-in-advance planning for services is unrealistic
    - Oracle RAC should deliver the on-demand services
  - A once-per-year workshop is not enough for efficient information sharing between the experiments:
    - An LCG technical board of liaisons from IT/PDS and experiments?

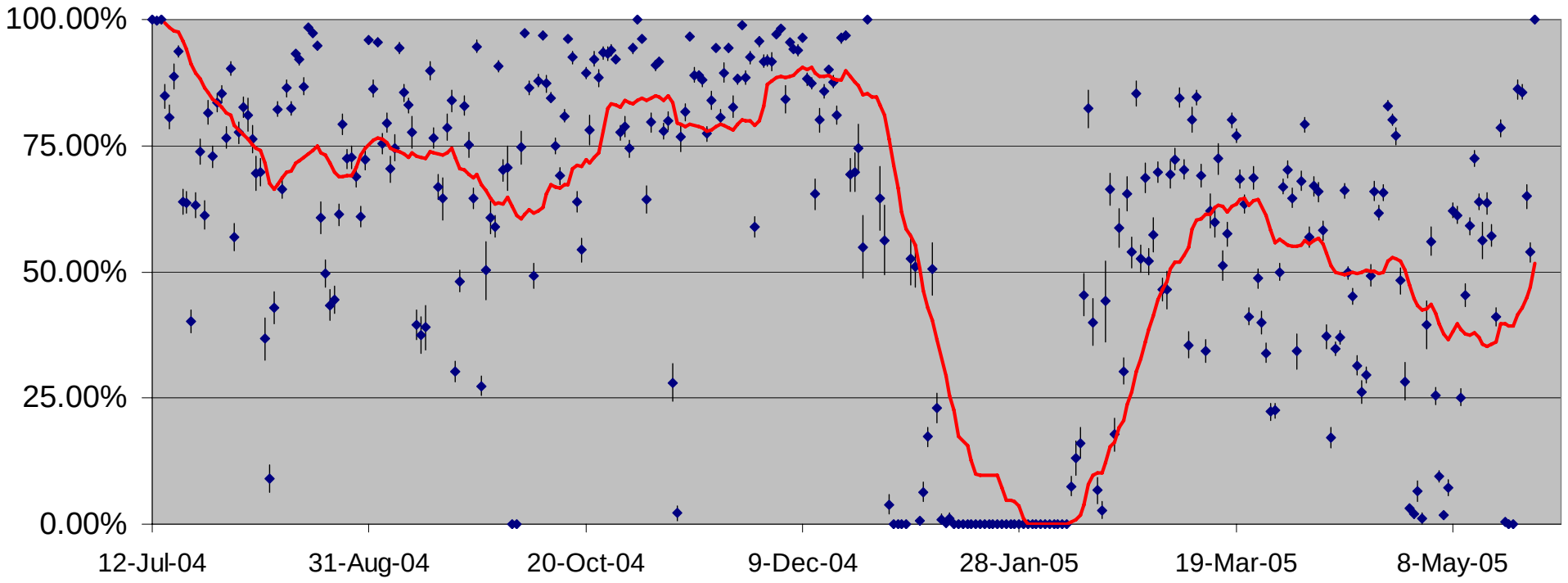
Summary

- The chaotic nature Grid computing increases fluctuations in database services demand
- Grid Database Access requires changes on the
  - Client-side:
    - ATLAS Client Library
  - Server-side:
    - On-demand deployment: ESF
    - Scalable technology: DASH (and/or FroNTier)

Backup Slides

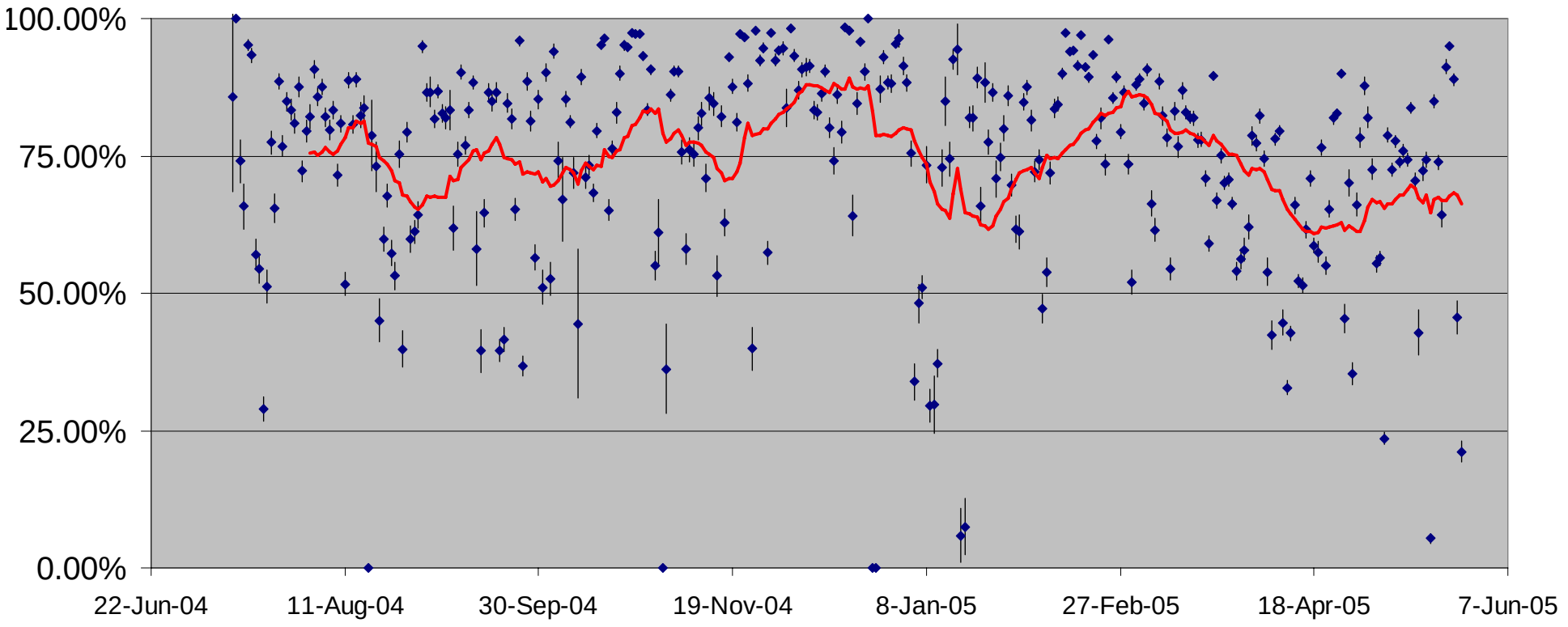
NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

ATLAS Efficiency on NorduGrid



- Fluctuations are limited when efficiency is close to the 100% boundary

ATLAS Efficiency on Grid3



- ATLAS achieved highest overall production efficiency on Grid3 (now OSG)

CERN Tier 0 Services Feedback September-December 2004

A spectrum of Oracle issues was addressed in the period:

- Continued Oracle prodDB monitoring, performance, troubleshooting
  - *Yulia Shapiro* is now helping *Luc Goossens*
- Best practices for pioneering ATLAS database application - Geometry DB:
  - Completed review of publishing procedures with the help of CERN IT/DB experts
- Another ATLAS Oracle database application DB application in development:
  - established CollectionsT0 DB on **devdb** (for T0 exercise + LCG3D testbed)
- Oracle client is now in the software distribution and production:
  - Thousands of distributed reconstruction jobs used new Oracle GeometryDB
- A very positive experience with oracle.support@cern.ch mechanisms, e.g.
  - emergency restoration of GeometryDB on development server
  - very helpful in resolving Oracle “features” encountered by GeometryDB users: vSize (*V. Tsulaia*), valgrind (*David Rousseau, Ed Moyse*), CLOB (*S. Baranov*)
- IT DB Services for Physics section is combined with Persistency Framework (POOL)
- *Dirk Geppert* is a newly established ATLAS Liaison

CERN Tier 0 Services Feedback January – May 2005

- Many thanks to CERN/IT FIO/DS for their excellent MySQL servers support
  - **atlasdbdev** server was serviced by vendor following disk alert
  - **atlasdbpro** server is continuously up and running for 336 days
- Thanks to CERN/IT ADC/PDS many Oracle performance issues with ProductionDB/ATLAS_PRODSYS on **atlassg** and (GeometryDB/ATLASDD) on **pdb01/atlas** were investigated and improved
  - indexing, caching, internal OCI error, deadlocks, etc
  - service interruption rescheduled to minimize effect on production

CERN Tier 0 Services Feedback January – May 2005

- Suggested improvements in the policies for the service interruptions announcements
- Suggested improved information sharing on Oracle database usage experience between the LHC experiments
- Requested notification when the tnsnames.ora file at CERN is changed
  - setup a monitoring system at BNL ***(Alex Undrus)***
- Requested support for the tnsnames.ora independence in POOL/RAL

CERN Tier 0 Services Feedback June - September 2005

- Many thanks to CERN/IT FIO/DS group for their excellent support of ATLAS MySQL servers
  - In July **atlasdbpro** server was serviced by vendor after **379 days of uninterrupted operations** that contributed to the success of DC2 and Rome production
  - Later in the reporting period both operating systems and the server versions were upgraded
- Thanks to CERN/IT ADC/PDS group various ATLAS Oracle database applications issues and security patches were quickly addressed and resolved: ***Dirk Geppert et al.***