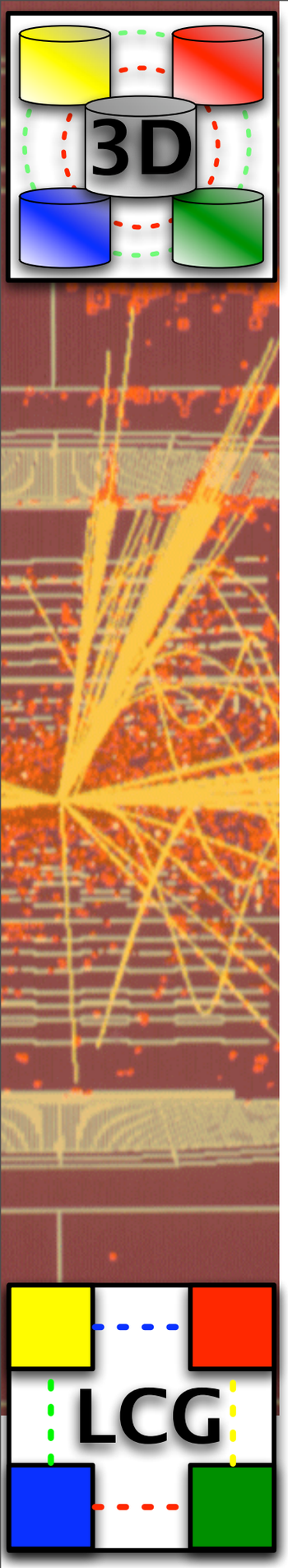


Production Experience with Distributed Deployment of Databases for the LHC

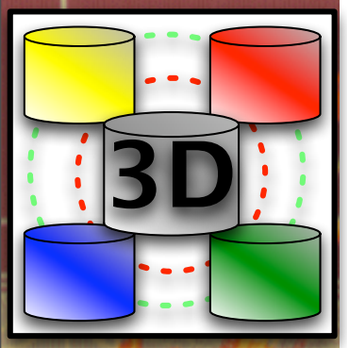
Dirk Duellmann, CERN IT
on behalf of the LCG 3D project
http://lcg3d.cern.ch

CHEP '07, September 5th
Victoria, Canada

PSS

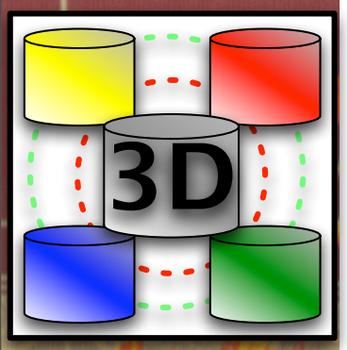
Related CHEP'07 Contributions

CERN IT Department



- 171 - Production Experience with Distributed Deployment of Databases for the LHC Computing Grid
- 172 - CERN Database Services for the LHC Computing Grid
- 122 - Relational databases for conditions data and event selection in ATLAS
- 161 - Building a Scalable Event-Level Metadata System for ATLAS
- 186 - Development, Deployment and Operations of ATLAS Databases
- 319 - Alignment data streams for the ATLAS Inner Detector
- 333 - Large Scale Access Tests and Online Interfaces to ATLAS Conditions Databases
- 90 - ATLAS Conditions Database Experience with the LCG COOL Conditions Database Project
- 236 - Control and monitoring of alignment data for the ATLAS endcap Muon Spectrometer at the LHC
- 294 - An Inconvenient Truth: file-level metadata and in-file metadata caching in the (file-agnostic) ATLAS distributed event store
- 462 - Database architecture for the calibration of ATLAS Monitored Drift Tube Chambers
- 358 - Distributed Interactive Access to Large Amount of Relational Data
- 430 - The ATLAS METADATA INTERFACE
- 110 - Experience and Lessons learnt from running high availability databases on Network Attached Storage
- 109 - Oracle RAC (Real Application Cluster) application scalability, experience with PVSS and methodology
- 265 - LHCb Distributed Conditions Database
- 213 - LHCb experience with LFC database replication
- 89 - LHCb Online Interface to a Conditions Database
- 322 - CMS Conditions Data Access using FroNTier
- 325 - The CMS Dataset Bookkeeping Service
- 182 - Distributed Database Access in the LHC Computing Grid with CORAL
- 181 - Development Status and Plans for the LCG Common Database Access Layer (CORAL)
- 204 - COOL Software Development and Service Deployment Status
- 205 - COOL Performance Tests and Optimization
- 350 - Nightly builds and software distribution in the LCG / AA / SPI project
- 447 - Implementing a Modular Framework in a Conditions Database Explorer for ATLAS
- 292 - Explicit state representation and the ATLAS event data model: theory and practice

3D Service Architecture



Oracle Streams
http cache (Squid)
Cross DB copy and MySQL/SQLight files

T0
– autonomous reliable service

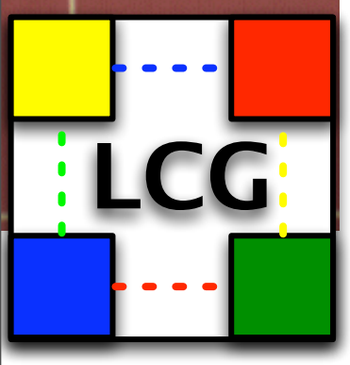
T1 – DB backbone
– all data replicated
– reliable service

T2 – local DB cache
– subset data
– only local service

Online DB
– autonomous reliable service

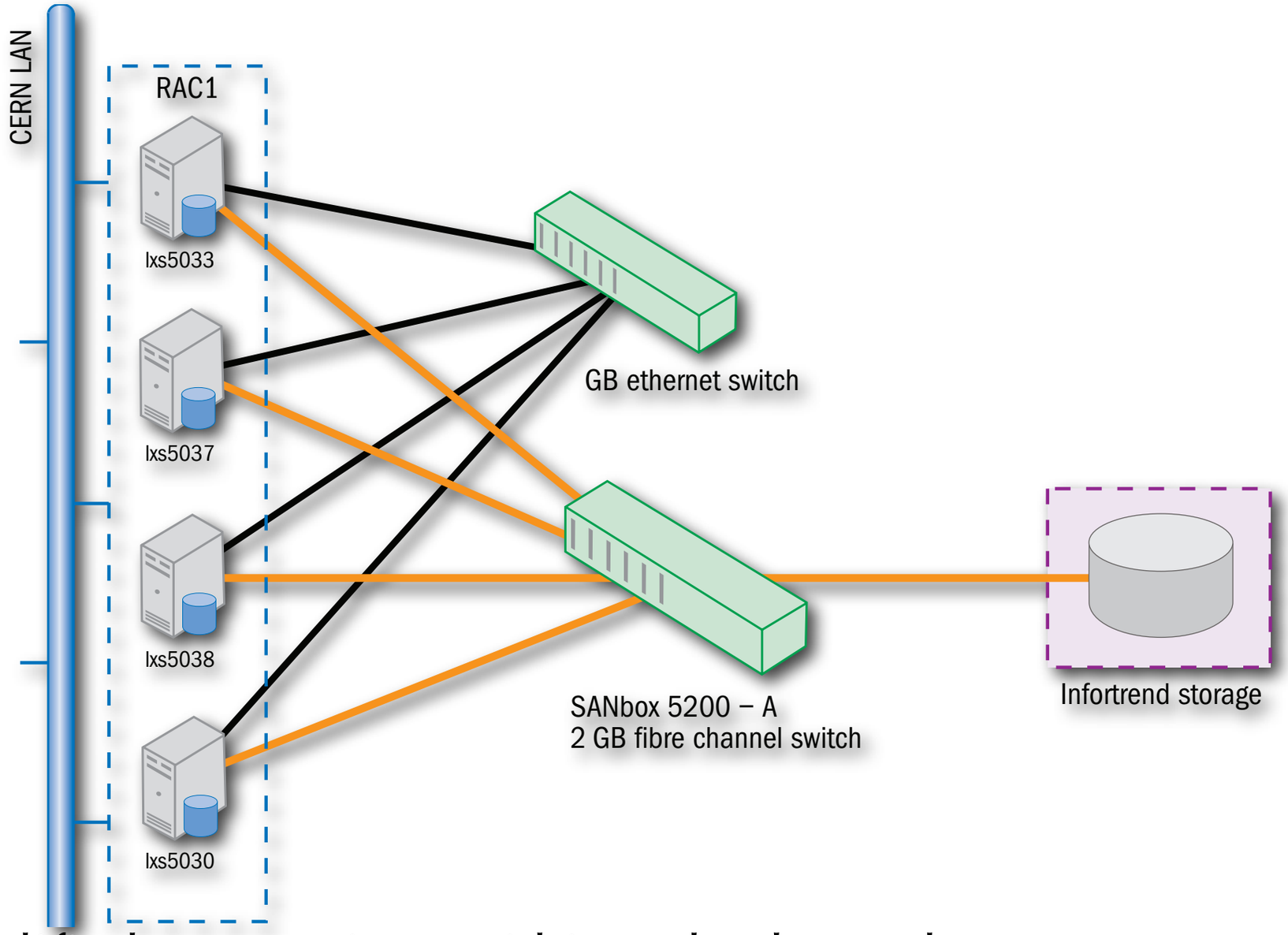
Read-only access at Tier-1/2 (at least initially)

O database (DB) cluster
M MySQL/SQLight DB file
S Squid
F FroNTier



Dirk.Duellmann@cern.ch

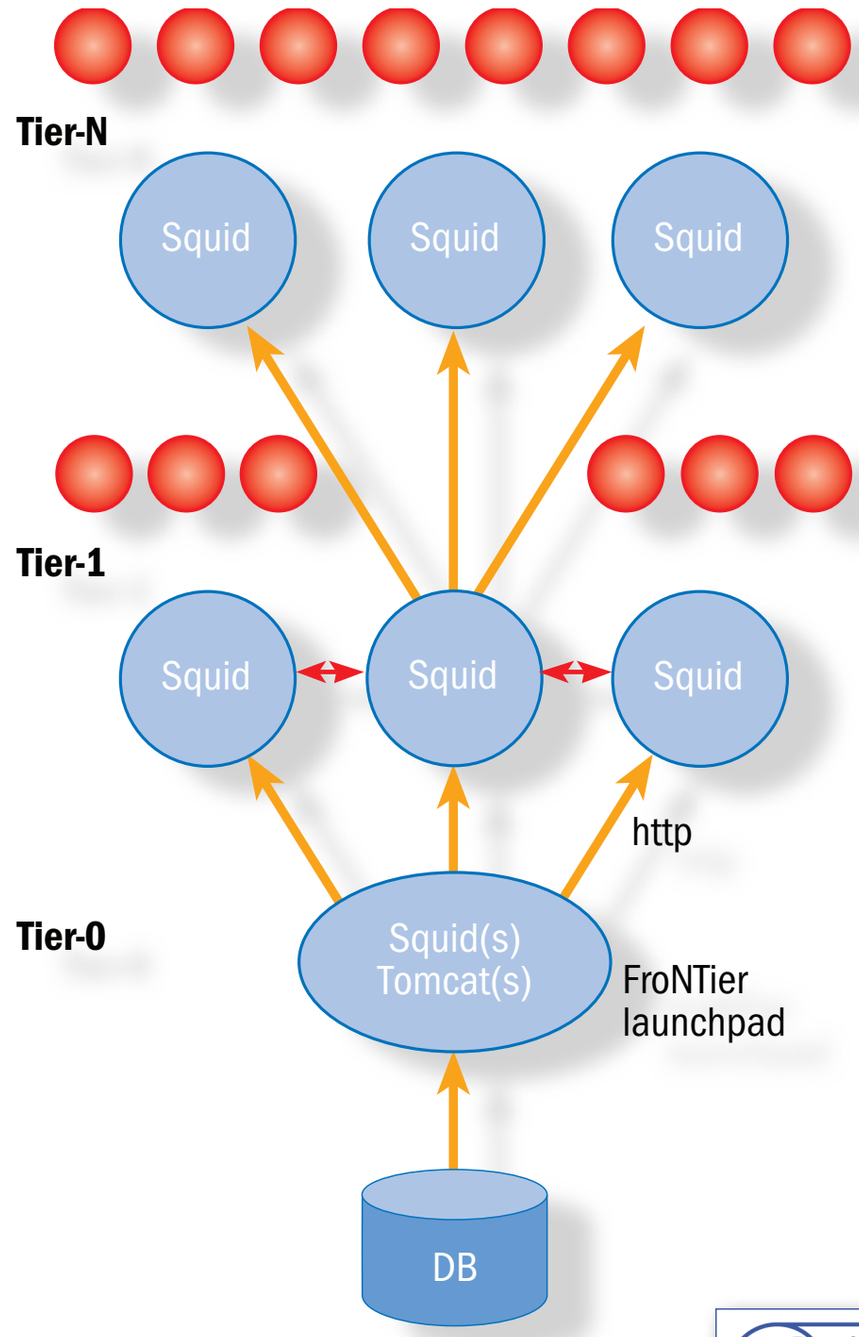
Building Block - Database Clusters



- Simplified connections within a database cluster
- Tier 0: all networking and storage redundant
- ☆ Storage and CPU scale independently
- ☆ Maintenance operations w/o down time

Frontier/Squid

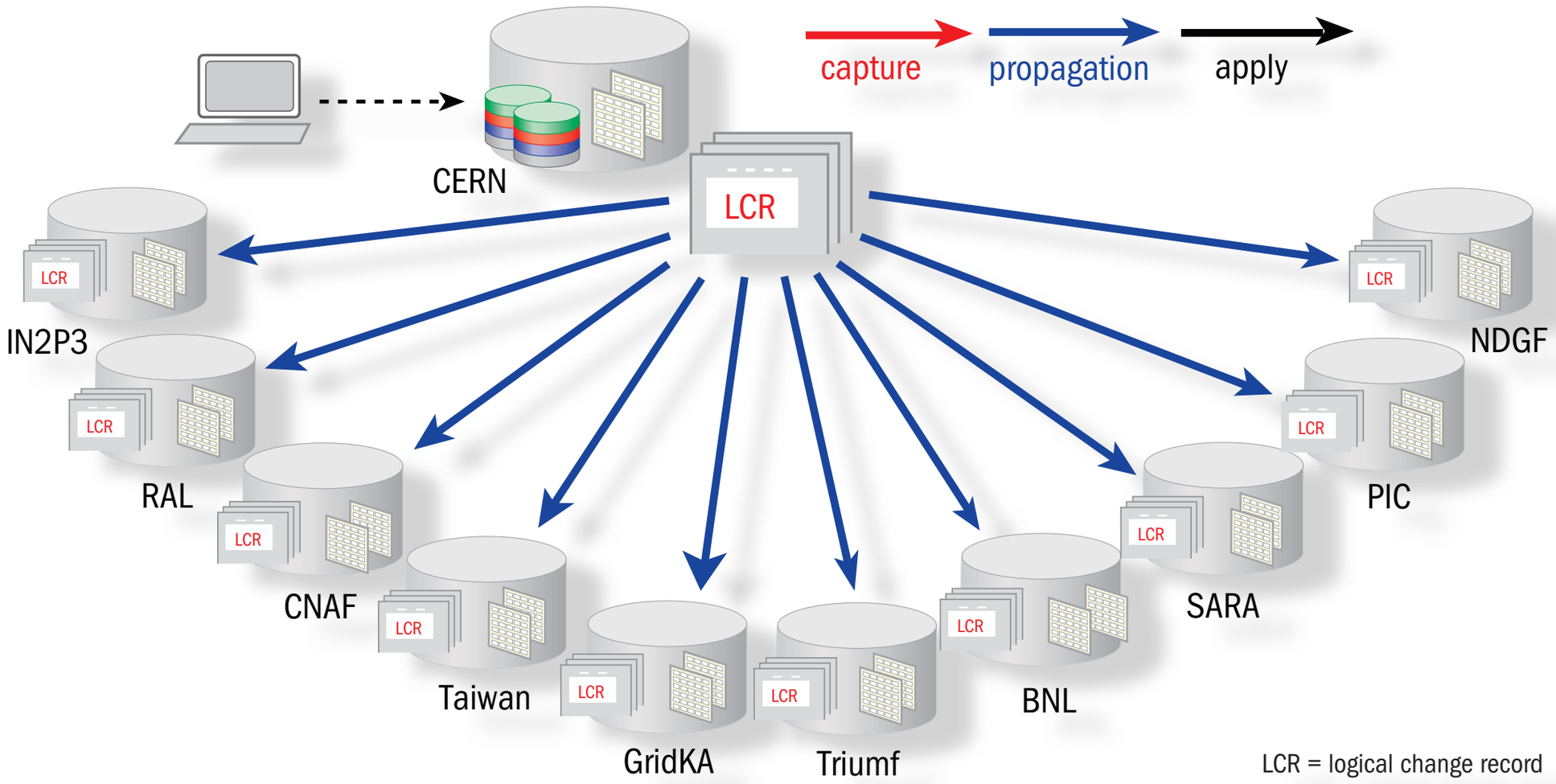
- Successfully used in CMS CSA'06
- Since then CMS reported numerous significant performance improvements
  - experiment data model
  - frontier client/server software
  - CORAL integration
- CMS is confident that possible cache coherency issues can be avoided by
  - Cache expiration windows for data and meta-data
  - Policy implemented by the client applications



Dirk.Duellmann@cern.ch

PSS Oracle Streams

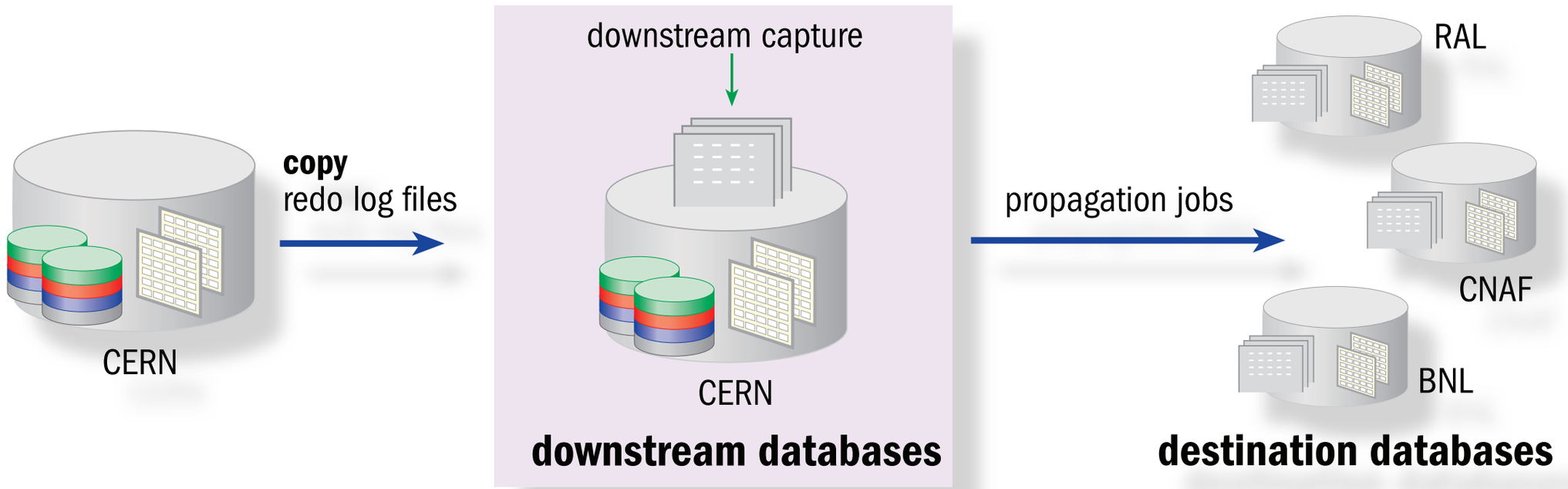
CERN IT Department



- Database changes captured from the redo-log and propagated asynchronously as Logical Change Records (LCRs)
- All changes are queued until successful application at all destinations
  - need to control change rate at the source in order to minimise the replication latency
  - 2GB/day user data to Tier 1 can be sustained with the current DB setups
- significant overheads between user data and redo-log volume apply

Downstream Capture & Network Optimisations

- Downstream Database Capture to de-couple Tier 0 production databases from Tier 1 or network problems in place for ATLAS and LHCb
- Optimising redo log retention on downstream database to allow for sufficient re-synchronisation window without recall from tape (eg 5 days)
- TCP and Oracle protocol optimisations yielded significant throughput improvements (almost factor 10)



Dirk.Duellmann@cern.ch

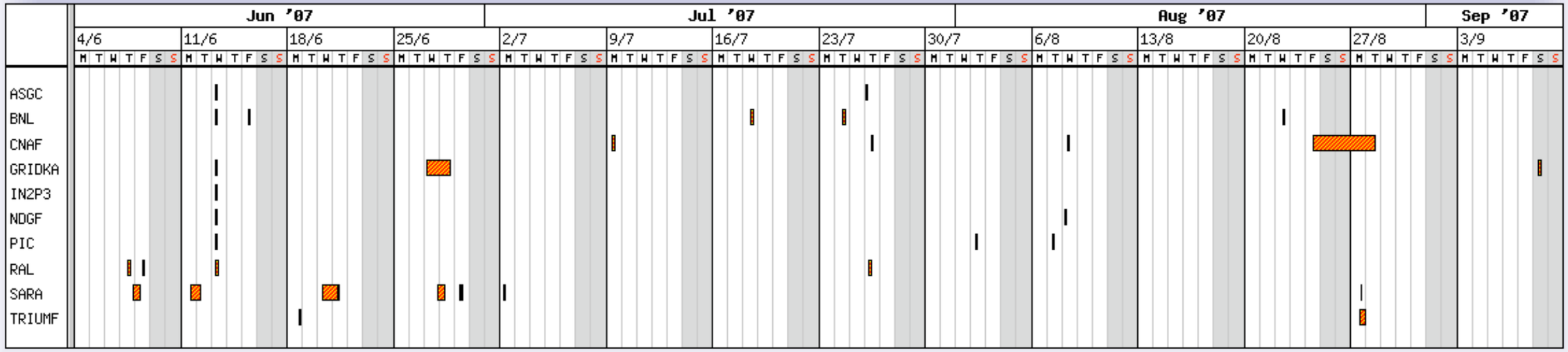
Database & Streams Monitoring

- Weekly/monthly database and replication performance summary has been added
  - extensive data and plots about replication activity, server usage, server availability available form the 3D wiki site (html or pdf).
  - summary plot with LCR rates during last week is show on the 3D home page and could be referenced/included into other dashboard pages
- Complemented by weekly Tier 0 database usage reports which are in use already since more than one year

PSS Intervention & Streams Reports

CERN IT Department

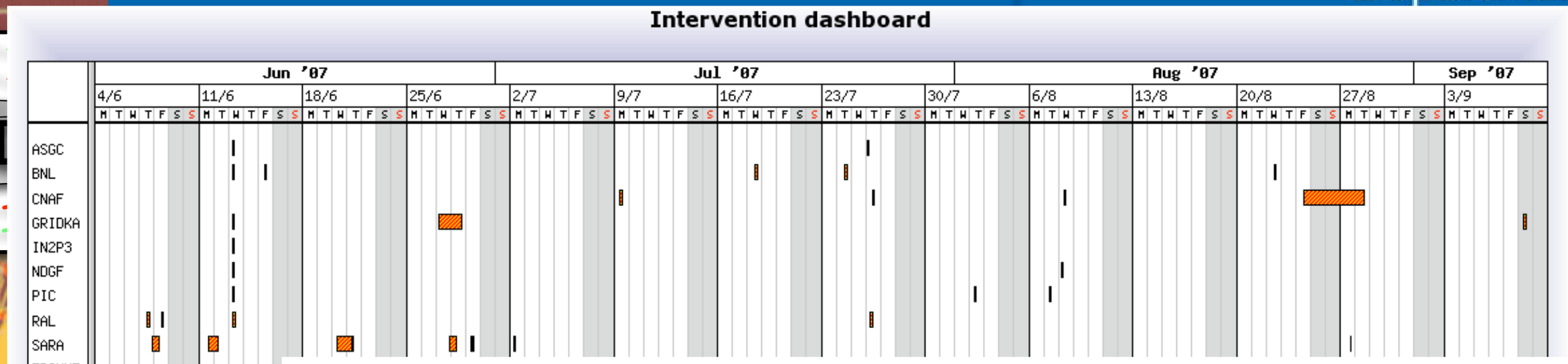
Intervention dashboard

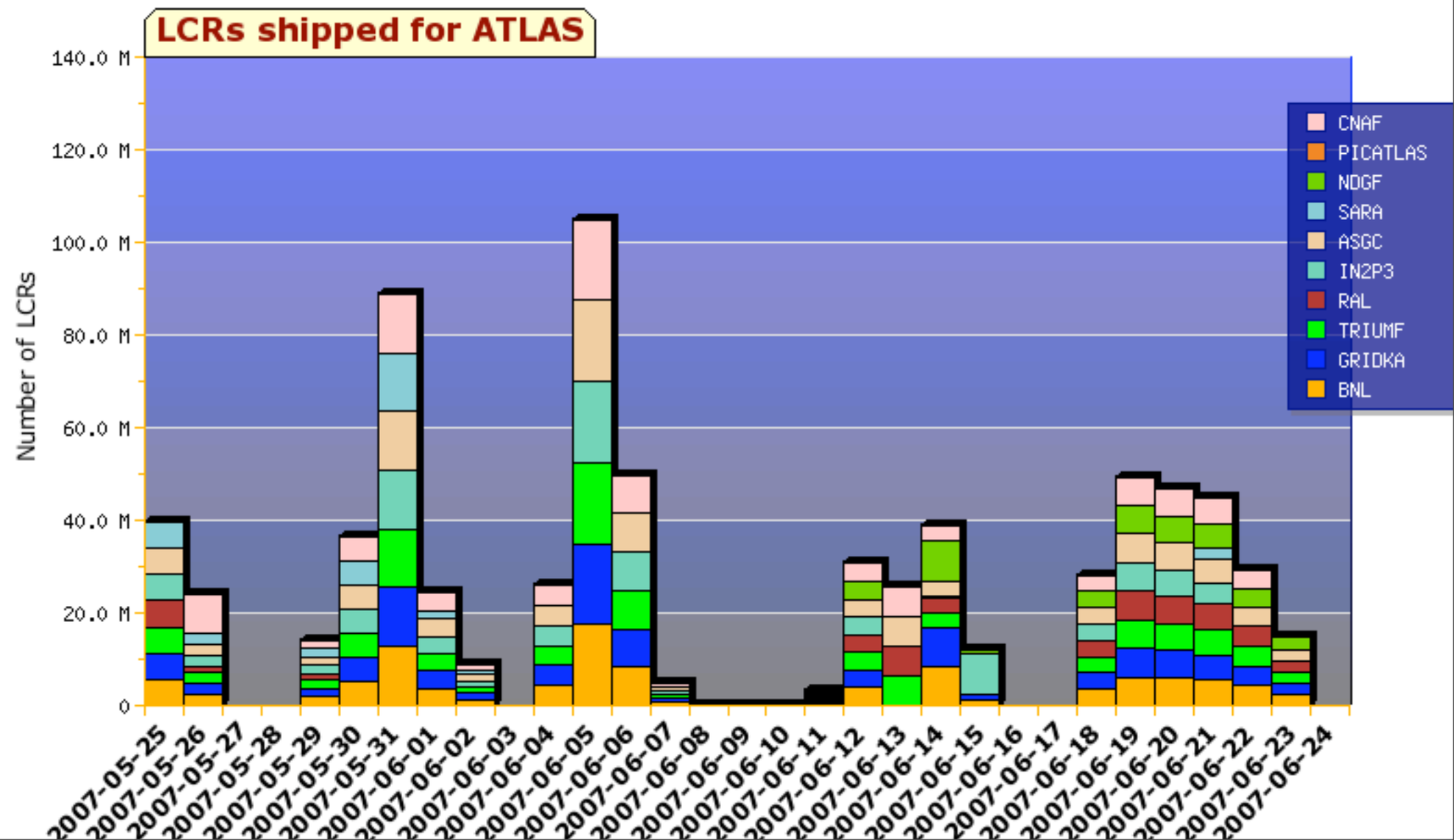


Dirk.Duellmann@cern.ch

PSS Intervention & Streams Reports

CERN IT Department

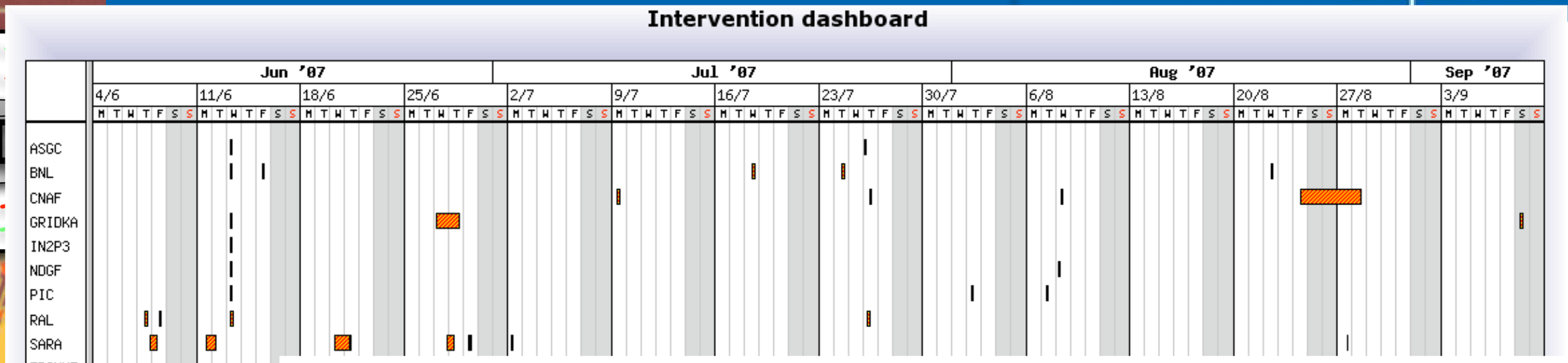


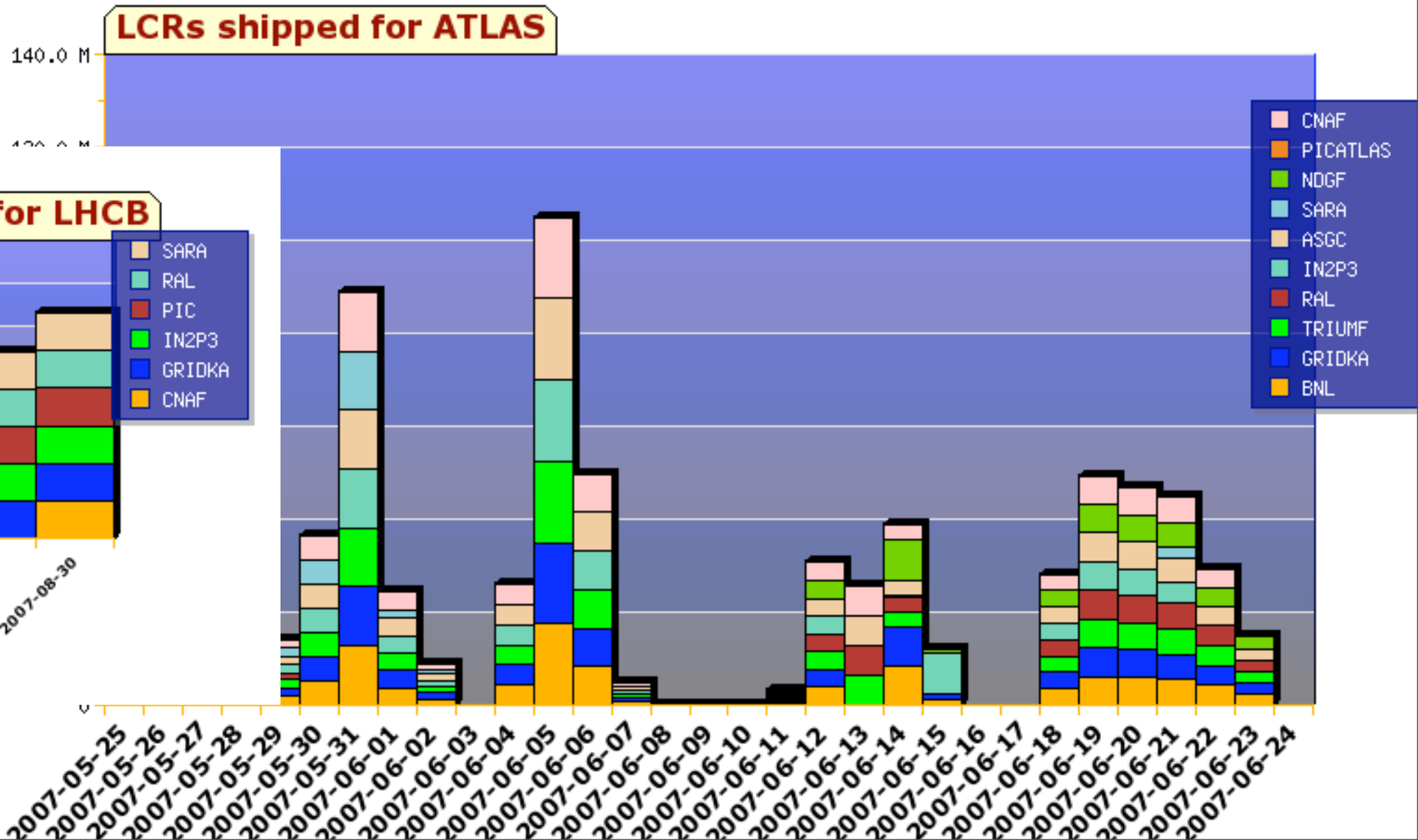


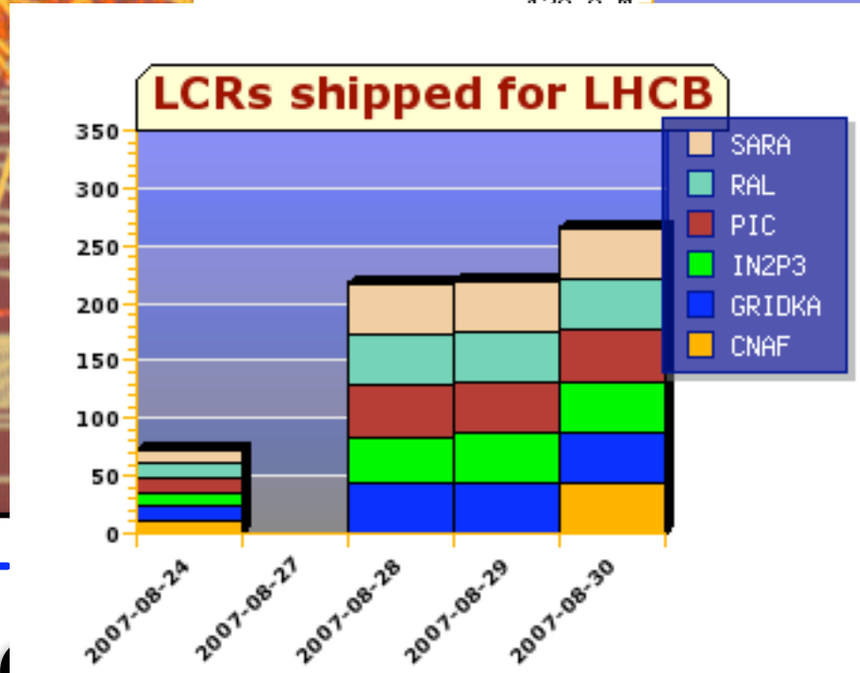
Dirk.Duellmann@cern.ch

PSS Intervention & Streams Reports

CERN IT Department







Dirk.Duellmann@cern.ch

Integration with WLCG Procedures and Tools

- 3D monitoring and alerting has been integrated with WLCG procedures and tools
  - dedicated workshop at SARA focussing on this
- Alerts since April being sent as GGUS tickets
  - Latency in some cases high for quick resolution
  - For now using direct email notification as well
- Interventions announced according to the established WLCG procedures
  - eg EGEE broadcasts, GOCDB entries
- To help reporting to the various coordination meetings we collect all 3D intervention plans also on the 3D wiki
  - Web based registration will be replaced as soon as a common intervention registry is in production

Tier 1 DB Scalability Tests

- Since spring year the experiments have started to evaluate/ confirm also the (so far) estimated size of the server resources at Tier 1
  - number of DB nodes CPUs
  - memory, network and storage configuration
- Need realistic work-load which now becomes available as experiment s/w frameworks approach complete coverage of their detectors
- ATLAS conducted two larger tests with ATHENA jobs against IN2P3 (shared solaris server) and CNAF (Linux)
  - Total throughput of several thousand jobs/h achieved with some 50 concurrent test jobs per Tier 1
- LHCb s/w framework integration done and scalability tests starting as well
  - Lower throughput requirements assumed than for ATLAS
  - Tests with several hundred concurrent jobs in progress

Dirk.Duellmann@cern.ch

Database Resource Requests

- Experiment resource requests for T1 unchanged since more than one year
  - 2 (3) node DB cluster for LHCb (ATLAS)
    - fibre channel based shared storage
  - 2 squid nodes for CMS
    - standard worker node with local storage
- Setup shown to sustain replication throughput required for conditions data (1.7 GB/d ATLAS)
  - ATLAS has moved production DB to the 3D setup
- LHCb is hosting their LFC replica databases on their Tier 1 databases
  - Successful replication tests with LFC and conditions since several month
- Updated requests for LHC startup have been collected during the WLCG w/s last week
  - No major h/w extension required

Dirk.Duellmann@cern.ch

Online-Offline Replication

- Oracle Streams used by ATLAS, CMS and LHCb
- Joint effort with ATLAS on replication of PVSS data between online and offline
  - required to allow detector groups to analyse detailed PVSS logs without adverse impact on online database
- Significantly higher rates required than for COOL based conditions
  - some 6 GB of user data per day
- Test over two months so far showed only minor problems (fixed by now)
  - e.g. switching between PVSS tablespaces
- Oracle Streams seems an appropriate technology also for this area

Backup & Recovery Exercise

- Organised several dedicated database recovery exercises
  - last exercise during the CNAF workshop
- Main aims
  - show that database implementation and procedures at each site are working
  - show that coordination and re-synchronisation after a site recovery works
  - show that replication procedures continue unaffected while some other sites are under recovery
- Exercise ahead of time has been well appreciated by all participants
  - several set-up problems haven been resolved during this hands on activity with all sites present
  - six sites have now successfully completed a full local recovery and resynchronisation
  - remaining sites will be scheduled after CHEP

Dirk.Duellmann@cern.ch

LFC and FTS DB back-ends

- LFC replica databases for LHCb in place
- Several sites plan to move their LFC and/or FTS database back-ends to database clusters similar to the set-up for experiment DBs suggested by 3D
  - service using the existing db clusters is well understood at the sites
  - increase availability and inclusion in the existing 3D monitoring/alerting
- Collected remaining questions about expected resource requirements
  - Service teams invited to dedicated 3D meeting about FTS
  - Wiki page addressing database configuration for FTS 2.0 has been produced and discussed with the T0 and T1 database administrators

Service Level Agreements and Policies

- DB Service level according to WLCG MoU
  - need more production experience to confirm manpower coverage at all T1 sites
  - piquet service being set-up at Tier 0 to replace existing 24x7 (best effort) service
    - streams interventions for now 8x5
- Proposals from CERN Tier 0 have been accepted also by the collaborating Tier 1 sites
  - Backup and Recovery
    - main items: RMAN based backups mandatory, data retention period 1 month, and volume
  - Security
    - security patch frequency and application procedure
    - database software upgrade procedure
    - patch validation window

Dirk.Duellmann@cern.ch

Summary

- The LCG 3D project has setup a wold-wide distributed database infrastructure for LHC
  - Close collaboration between LHC experiments and LCG sites
  - with some 100 DB nodes at CERN + several tens of nodes at Tier 1 sites this is one of the largest distributed database deployments world-wide
- Large scale experiment test have validated the experiment resource requests implemented by the sites
- Backup & recovery tests have been performed to validate the operational procedures for error recovery
- Regular monitoring of the database and streams performance is available to the experiments and sites
  - Integrated with EGEE problem reporting
- Tier 0+1 ready for experiment ramp-up to LHC production
  - Next steps: Replication of ATLAS Muon conditions back to CERN

Dirk.Duellmann@cern.ch

More detail

- LCG 3D wiki
  - interventions, performance summaries
    - http://lcg3d.cern.ch
- Recent LCG 3D workshops
  - Monitoring and Service Procedures w/s @ SARA
    - http://indico.cern.ch/conferenceDisplay.py?confId=11365
  - Backup and Recovery w/s @ CNAF
    - http://indico.cern.ch/conferenceDisplay.py?confId=15803

Other Areas of Progress

- License Agreement for Oracle licenses for Tier 1 LCG services
  - Instant client distribution clarified
- Grid certificate protected DB access implemented (in CORAL)
  - VOMS credentials to authenticate and check DB access permission using the LCG Catalog (LFC)
- Experiment estimates for condition data volumes and DB server CPU improving
- All experiments have online-to-offline connection in place and exercise
  - Online setups being replaced by full scale DB cluster
  - Promising streams test with PVSS (6GB/d)