

PSS



LCG 3D Project Status

Dirk Düllmann, CERN IT

(on behalf of the LCG 3D team - more detail at <http://lcg3d.cern.ch>)

LHCC Comprehensive Review
25-26 September 2006, CERN



- Full production milestone approaching in October
- Tier 1 sites are asked to provide Database and Frontier installations
 - According to experiment requirements agreed last October GDB/MB
- Seven sites from the 3D phase 1 are available
 - and included in experiment test activities
- Four additional sites were asked to join now
 - NDGF, NIKHEF/SARA, PIC, TRIUMF
- Full workshop agenda at
 - <http://agenda.cern.ch/fullAgenda.php?ida=a063213>



- Review
 - site status and any remaining issues
 - experiment / project replication tests
 - service procedures and align with existing LCG operations infrastructure
 - experiment database / frontier resource requests for production in the next 6 month
- Goal
 - experiments and sites agree on what is (can be) expected during the next 6 month



- License Requirements collected and acquired
 - All sites have Oracle s/w and support according to current experiment and project requests
- Database administrator training
 - One week course held with 14 new Oracle administrators from experiments and sites
 - OCP training being setup for November
- Streams through put tests between CERN and T1 sites
 - 10 - 100 MB/min reached (typical 30 MB/min)
 - WAN replication running at ~50% of LAN rates
 - Sufficient for planned use with conditions data
 - Need to continue to work with Oracle on rate improvement
- Experiments took over T1 setups for their replication / client access tests
 - Closing online-offline-T1 database chain



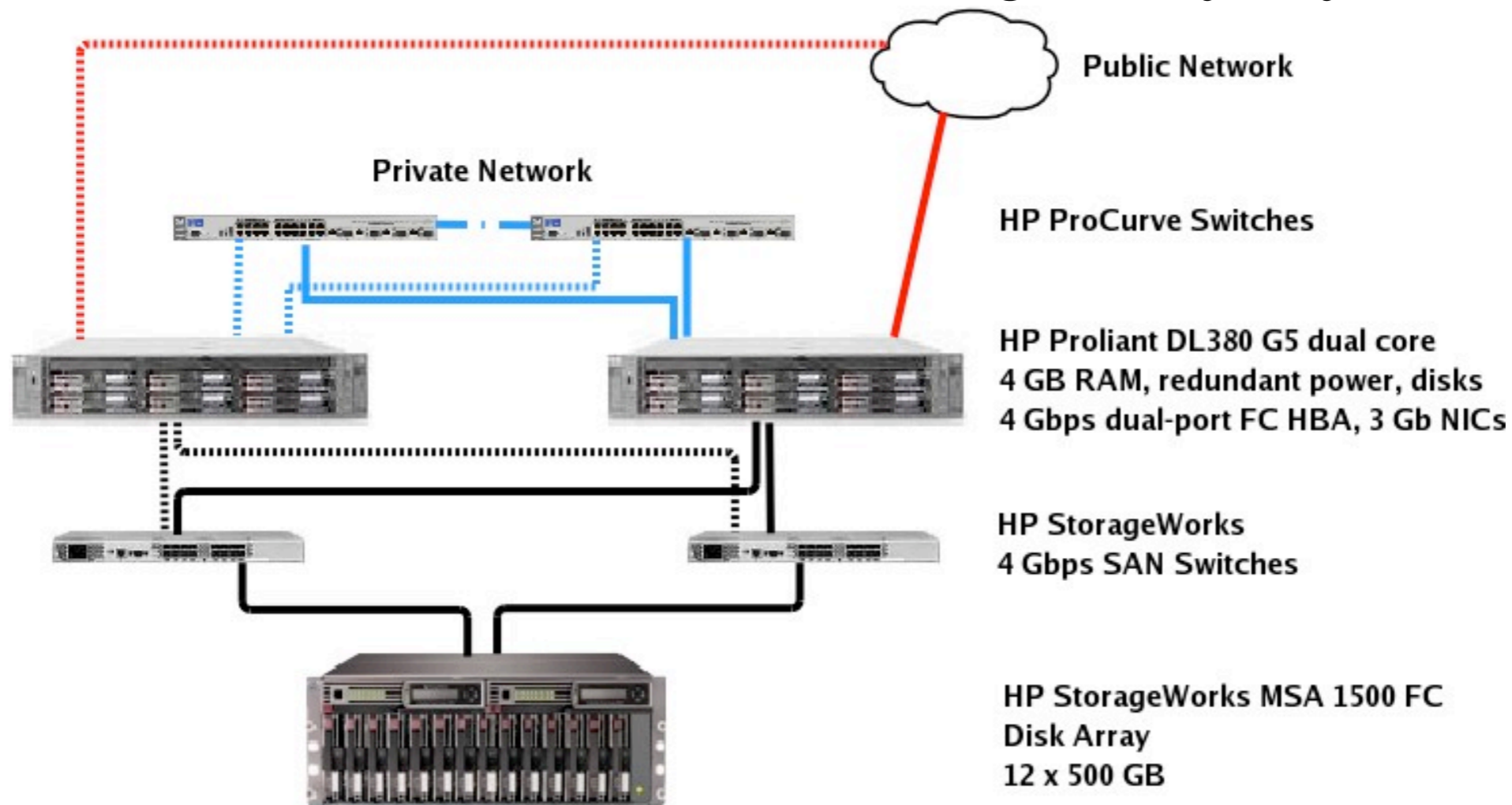
- **FroNTier and SQUID**
 - Tier 0, all Tier 1 and (almost) all Tier 2 sites are up and tested by CMS
- **Databases - T0 and all Phase 1 sites are up**
 - ASCG, BNL, CERN, CNAF, GridKa, IN2P3, RAL
 - All sites involved in tests with one or two experiments
 - Consolidation plans at some sites: RAC also for databases behind grid services
- **Remaining issues: completion of monitoring and backup set-up at some sites**



- Situation for October milestone
 - TRIUMF
 - DBA and h/w available, setting up cluster
 - actively participating in project meetings and workshops
 - NIKHEF/SARA
 - DBA and h/w available
 - waiting for SAN connection
 - NDG
 - DBA hired beginning of September
 - waiting for h/w arrival
 - PIC
 - plan to involve external company for RAC setup
 - Need to allocate DBA
- Phase 2 sites need to increase participation in 3D meetings
- Experiments need to expect delays for these sites wrt October milestone

D3D hardware

- Oracle RAC solution based on HP hardware exclusively
- Storage could be sufficient through 2008 (unlikely)
- Will be add more RAC nodes + storage arrays by end 2008-09





- 3D Oracle Enterprise Manager in place
 - collects diagnostics from all 3D sites
- Web based streams monitoring
 - Show experiment database (on-line, off-line and T1) and replication status between them
 - Database availability, throughput, latency wrt to Tier 0
 - Developed by technical student (Z. Baranowski)
- In progress
 - integration with in ATLAS dashboard and SAM availability framework
 - in contact with Oracle development for possible inclusion in Oracle Enterprise Manager

Hosts

Page Refreshed Sep 23, 2006 6:09:50 PM CEST

Search [Advanced Search](#) |

Select	Name	Status	Alerts	Policy Violations	Compliance Score (%)	CPU Util %	Mem Util %	Total IO/sec
<input checked="" type="radio"/>	ccdbcl01.in2p3.fr			5 0 0	82			
<input type="radio"/>	ccdbcl02.in2p3.fr			5 0 0	82			
<input type="radio"/>	f01-010-111-e.gridka.de		0 2	5 0 0	82	2.36	98.44	26.01
<input type="radio"/>	f01-010-112-e.gridka.de		0 0	5 0 0	82	10.29	97.35	2352.61
<input type="radio"/>	f01-010-113-e.gridka.de		0 0	5 0 0	82	2.85	98.58	24.65
<input type="radio"/>	f01-010-114-e.gridka.de		1 0	5 0 0	82	2.01	98.76	24.13
<input type="radio"/>	lcgdb01.gridpp.rl.ac.uk		1 6	5 0 0	82	12.76	82.28	49.36
<input type="radio"/>	lcgdb02.gridpp.rl.ac.uk		1 7	5 0 0	82	13.2	98.79	43.82
<input type="radio"/>	lcgdb03.gridpp.rl.ac.uk		0 3	5 0 0	82	5.21	92.74	51.03
<input type="radio"/>	lcgdb04.gridpp.rl.ac.uk		0 3	5 0 0	82	4.54	94.5	40.04
<input type="radio"/>	lxfs5591.cern.ch		1 0	5 1 0	76	1.19	97.91	11.57
<input type="radio"/>	lxfsrk421.cern.ch		0 1	5 1 0	76	4.22	98.63	32.36
<input type="radio"/>	ora-rac-02.cr.cnaf.infn.it		0 0	5 1 0	76	24.65	80.38	78.1
<input type="radio"/>	ora-rac-04.cr.cnaf.infn.it		0 0	5 1 0	76	17.9	90.14	85.86
<input type="radio"/>	oraclus01.usatlas.bnl.gov		0 0	5 3 0	70	6.51	98.42	112.8
<input type="radio"/>	oraclus02.usatlas.bnl.gov		0 1	5 3 0	70	5.02	97.55	141.86
<input type="radio"/>	rac01			5 1 0	76			
<input type="radio"/>	rac02			5 1 0	76			

 | TIP For an explanation of the icons and symbols used in this page, see the [Icon Key](#).

Related Links

[Customize Table Columns](#)[Execute Host Command](#)

- ccdb01
- ccdb02
- f01-010
- f01-010
- f01-010
- f01-010
- f01-010
- lcgdb01
- lcgdb02
- lcgdb03
- lcgdb04
- lxfs559
- lxfsrk421
- ora-rac1
- ora-rac2
- oraclus1
- oraclus2
- rac01
- rac02

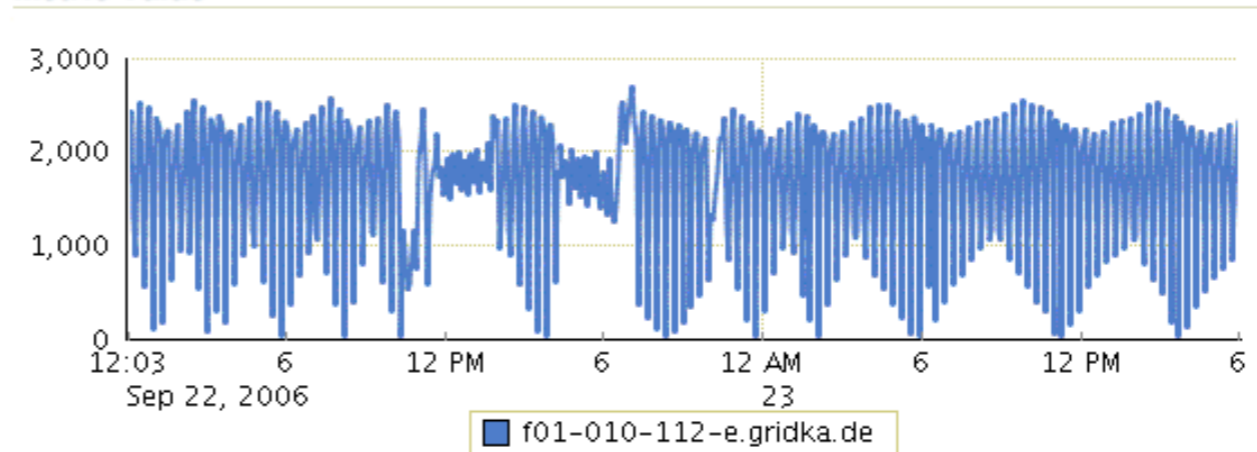
Host: f01-010-112-e.gridka.de > All Metrics >
Total Disk I/O Per Second: Last 24 hours

Last Updated **Sep 23, 2006 5:58:53 PM CEST**
 View Data Last 24 hours

Statistics

Last Known Value **2352.61**
 Average Value **1716.15**
 High Value **2705**
 Low Value **19.8**
 Warning Threshold **Not Defined**
 Critical Threshold **Not Defined**
 Occurrences Before Alert **No data**
 Corrective Action **None**

Metric Value



Alert History

Severity	Timestamp	Message	Last Comment	Details
(No alerts)				

Related Links

[Compare Targets](#) [Metric and Policy Settings](#)

TIP For an explanation of the icons and symbols used in this page, see the [Icon Key](#).

Related Links

[Customize Table Columns](#) [Execute Host Command](#)

INSTANCE (Service Name):

D3R1 @d3r1-v.cern.ch (D3R.CERN.CH)

SNAPSHOT TIME:

2006-09-21 09:44:07

< Prev Next >

Capture Processes:

ID	Name	Queue	LCRs Captured/s	LCRs Enqueued	LCRs Enqueued/s	Capture Latency	State
CP2	STRMADMIN_CAPTURE_BR	55650	1.0	2020141	0.0	6 sec	CAPTURING

Propagation Processes:

Source Queue	ID	Name	LCRs Propagated	LCRs/s	Bytes/s	State	Destination
55650	PS1	STRMADMIN_PROPAGATE_RALBR	2020117	0.0	0.0	ENABLED	STREAMS_QUEUE_AP @ OGMA.GRIDPP.PLAC.UK
	PS2	STRMADMIN_PROPAGATE_CNAFBR	2020117	0.0	0.0	ENABLED	STREAMS_QUEUE_AP @ STRMTEST.CR.CNAF.INFN.IT

Disabled:

Name	Source Queue	Error Time	Error Msg	Status	Destination
STRMADMIN_PROPAGATE_TEST1	50759	30-08-2006 14:47:18	ORA-12514: TNS:listener does not currently know of service requested in connect descriptor; ORA-06512: at "SYS.DBMS_AQADM_SYS", line 1087; ORA-06512: at "SYS.DBMS_AQADM_SYS", line 7627; ORA-06512: at "SYS.DBMS_AQADM", line 631; ORA-06512: at line 1;	DISABLED	STREAMS_QUEUE_AP @ TEST1.CERN.CH

Queues:

ID	Name	Cumulate/s	Spilled/s
Q55650	STREAMS_QUEUE_BR	0.0	0.0
Q50759	STREAMS_QUEUE_CA	0.0	0.0

ORACLE Enterprise Manager 10g Grid Control

Hosts | Databases | Application S

Host: f01-010-112-e.gridka.de > All Met

Total Disk I/O Per Second: L

Statistics

Last Known Value	2352.61
Average Value	1716.15
High Value	2705
Low Value	19.8
Warning Threshold	Not Defi
Critical Threshold	Not Defi
Occurrences Before Alert	No data
Corrective Action	None

Alert History

Severity	Timestamp
(No alerts)	

Related Links

[Compare Targets](#)

Home

Copyright © 1996, 2005, Oracle. All rights reserved. Oracle, JD Edwards, PeopleSoft, and Retek are trademarks of Oracle Corporation and/or its affiliates. Other brands and product names are trademarks of their respective owners.

[About Oracle Enterprise Manager](#)

TIP For an explanation of the icons and symbols used in this page, click here.

Related Links

[Customize Table Columns](#)

INSTANCE (Service Name):

LUGH2 (LUGH.GRIDPP.RL.AC.UK)

SNAPSHOT TIME:

2006-09-21 09:48:56

< Prev Next >

Propagation Receivers:

Destination Queue	ID	LCRs Propagated	LCRs/s	Total time	Source
53506	PR1	871004	0.0	23 hr	STREAMS_QUEUE_CA @ INT4R.CERN.CH

Apply Processes:

ID	Name	Queue	LCRs Dequeued	LCRs /s	Latency	Transaction Applied	Transaction App/s	Total Latency	State
AP2	STRMADMIN_APPLY_LHCB	53506	871004	0.0	0	133800	0.0	36 sec	IDLE

Queues:

ID	Name	Cumulate/s	Spilled/s
Q53506	STREAMS_QUEUE_LHCB_AP	0.0	0.0

Main Page TWiki

Snapshot Time: 2006-09-21 09:48:56

ING

Destination
JE_AP @ OGMA.GRIDPP.RL.AC.UK
JE_AP @ STRMTEST.CR.CNAF.INFN.IT

Status	Destination
DISABLED	STREAMS_QUEUE_AP@ TEST1.CERN.CH



- Resource Review for next 6 month
 - Scope database and FroNTier resources for Oct'06 - Mar'07
- Summary: Tier 1 resources are still adequate for ATLAS, CMS and LHCb
 - Significant increase in medium term expected eg by ATLAS (some 10 TB/y)
- Propose review via 3D regularly
 - eg every 6 month
 - Approval of new requests via LCG GDB/MB



- Each site is responsible to setup database backup and recovery infrastructure via Oracle RMAN
 - This may include on-disk backups and should include tape backups and associated media
- Backups should be performed online and with a retention period which is compatible with the time window for point-in-time recovery required by the experiments
 - Eg 1 month or 3 month?
- This is required to allow for a standard recovery procedure including streams re-synchronisation
- The T1 sites are responsible for performing
 - Standard database recovery (eg after media faults)
 - Point-in-time recovery within the agreed time window in case of logical data corruption
- Tier 0 is responsible for streams re-synchronisation once a site is locally consistent again



- Database Recovery with Streams
- Collected DB / streams recovery scenarios
 - Recovery after T1 data loss - **OK**
 - RAL recovered and re-synchronised
 - Replication CENR to CNAF continued unaffected
 - Recovery after T0 data loss - **OK**
 - Next: coordinated point-in-time recovery
 - Procedure defined, will validate asap
- Service procedure documented on 3D wiki
- Planning local and 3D wide recovery exercise as soon as all sites have backup system in place
- Need to include procedures into T0 operational procedures



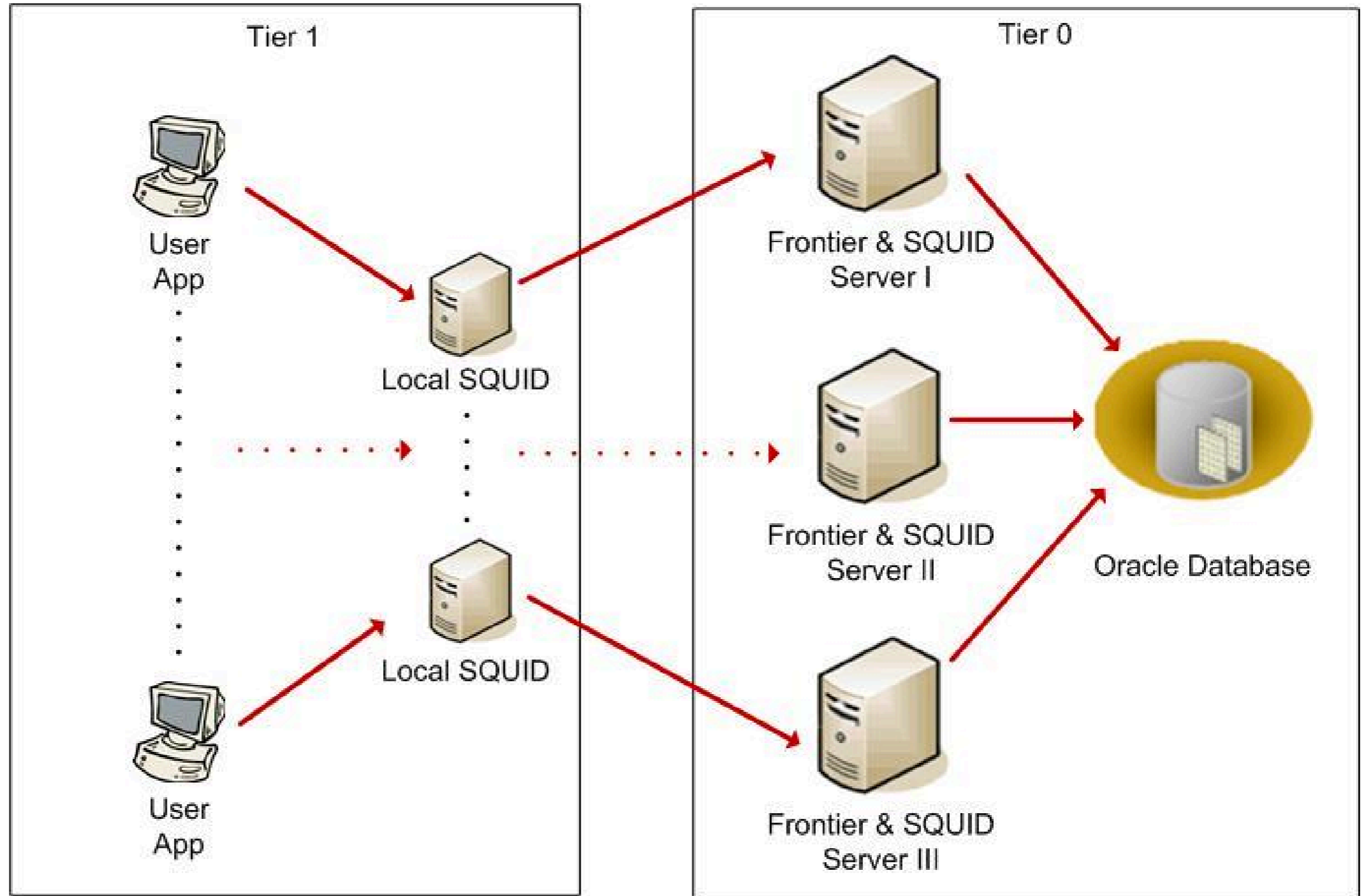
- 3D replication expertise at T0
 - trainee (FroNTier expert)
 - oracle fellow (Oracle streams expert)
 - technical student (Oracle streams monitoring)
 - plus consultancy from Physics DB team
- Moving from R&D to service phase
 - Streams management/recovery tasks will need to move soon to T0 service staff
 - Additional effort on rather loaded physics database support staff
- Need to insure that replication expertise gathered is kept



- FoNTier infrastructure ready at T0, T1 and many CMS T2 sites
 - Performance and stability looks adequate, monitoring in place
 - Access pattern and caching policy still need validation
- Database and Replication ready at T0 and T1 (phase 1) sites
 - Replication performance sufficient for conditions data and catalogs
 - More optimisation seem possible - in contact with Oracle
 - **Several phase 2 sites are late and need to participate now!**
- Service level and responsibility proposal being discussed
 - Impact on T0 for central replication support
- Experiment access patterns are still not fully known
 - But all experiments are actively involved in testing their applications in the real infrastructure now



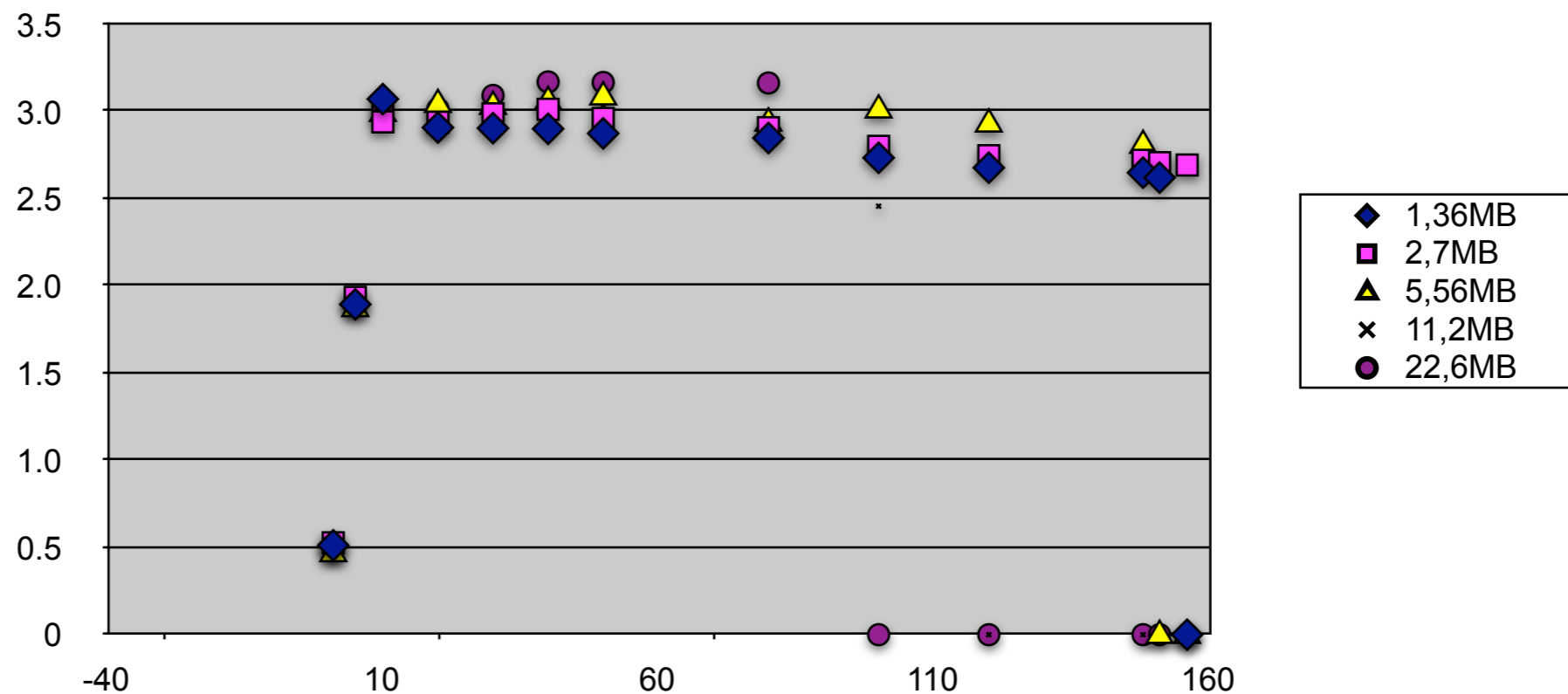
Backup slides



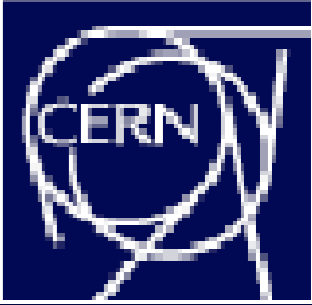
Throughput analysis Frontier Server



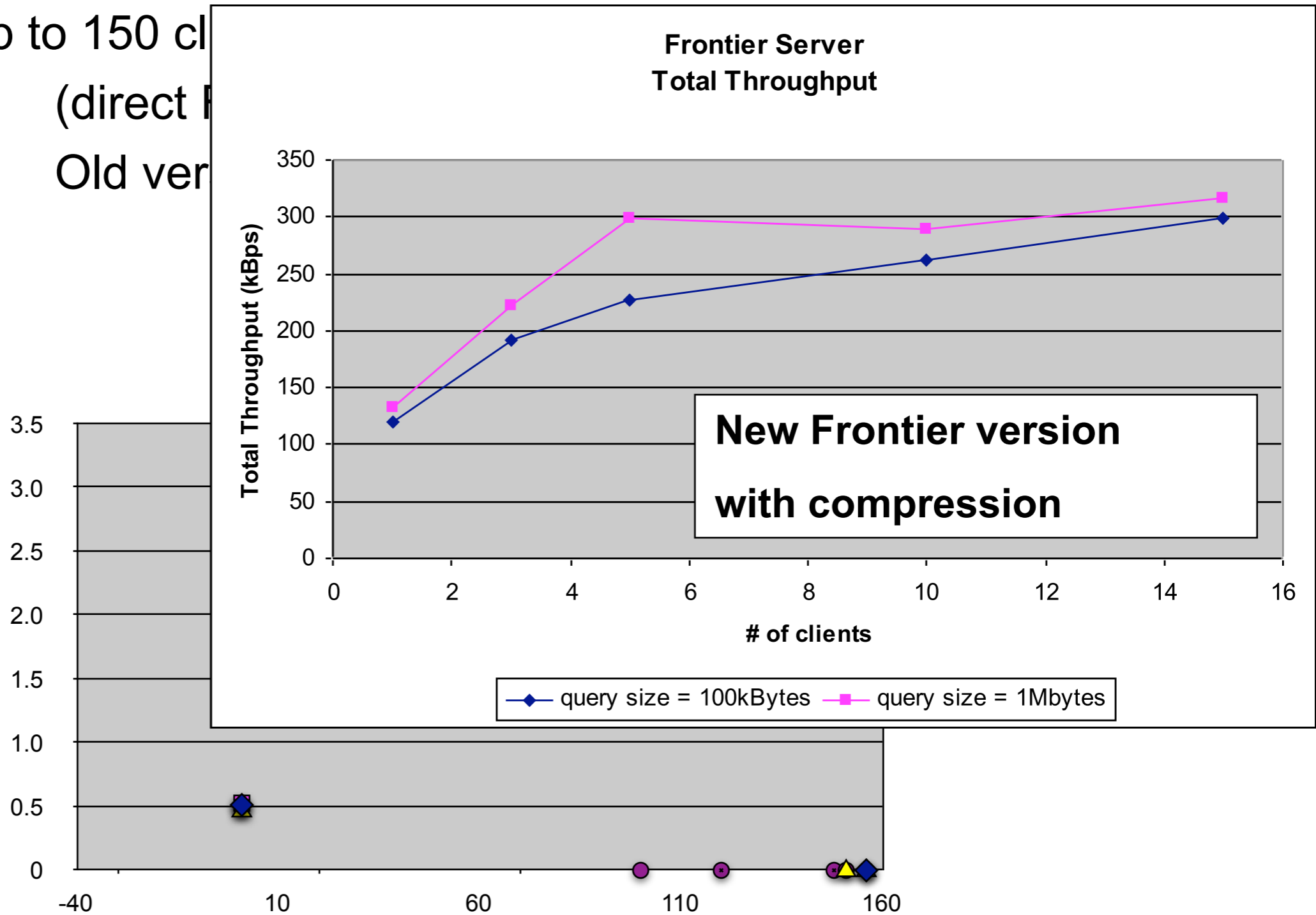
- Up to 150 clients running against a single server
(direct FroNtier server access, no Squid involved)
Old version of FroNtier -> no compression!



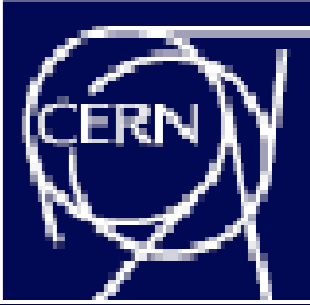
Throughput analysis Frontier Server



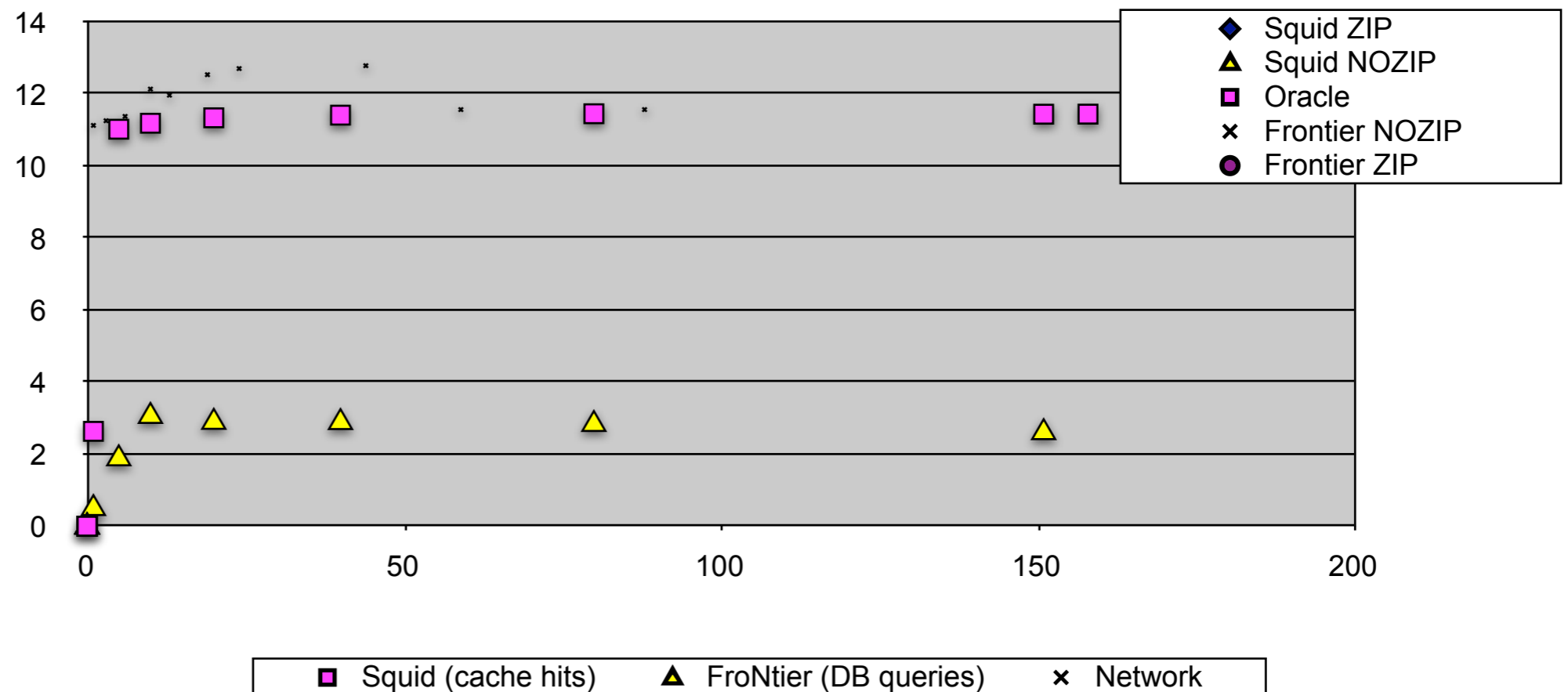
- Up to 150 clients
(direct ...)
Old version



Throughput analysis Oracle, FroNtier and Squid



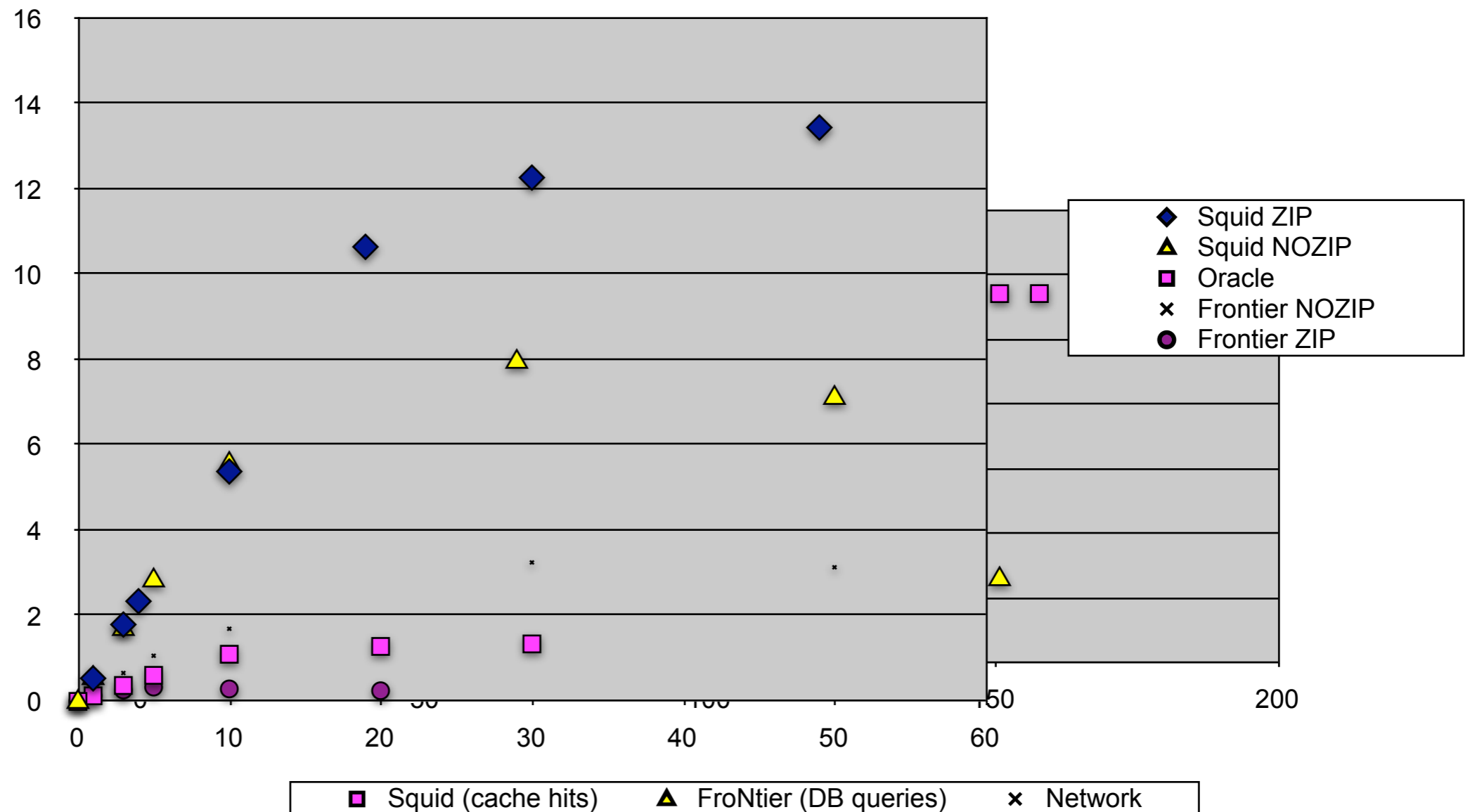
- Oracle vs Frontier Server vs Squid Cache Hits



Throughput analysis Oracle, FroNtier and Squid

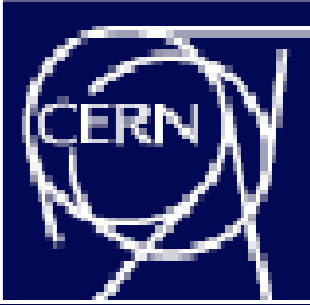


- Oracle vs Frontier Server vs Squid Cache Hits



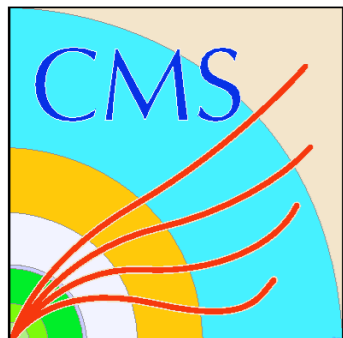
Throughput analysis

Notes on previous plots



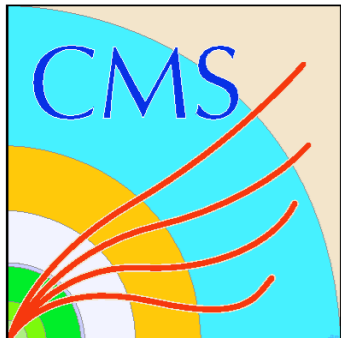
- **Direct Frontier access**
 - NOZIP version: 3MBps (bottleneck is the database)
 - ZIP version: 0,3MBps (bottleneck is the server CPU)

 - ZIP version can get **10 times slower than** NOZIP version
 - Production setup with 3 FroNtier nodes will perform better!
- **Squid access**
 - NOZIP version: 8MBps
 - ZIP version: 14MBps
 - user preceived throughput can be bigger than the network throughput (due to compression)
 - ZIP version can get **2 times faster than** NOZIP version
- **Oracle access - 1,34MBps**
 - **First guess, should be** faster then FroNtier direct access in any case!
 - **Second thought**, each client is repeatedelly creating DB connections which is quite heavy for OraclePlugin and not so much for Frontier because frontier servlet reuses connections



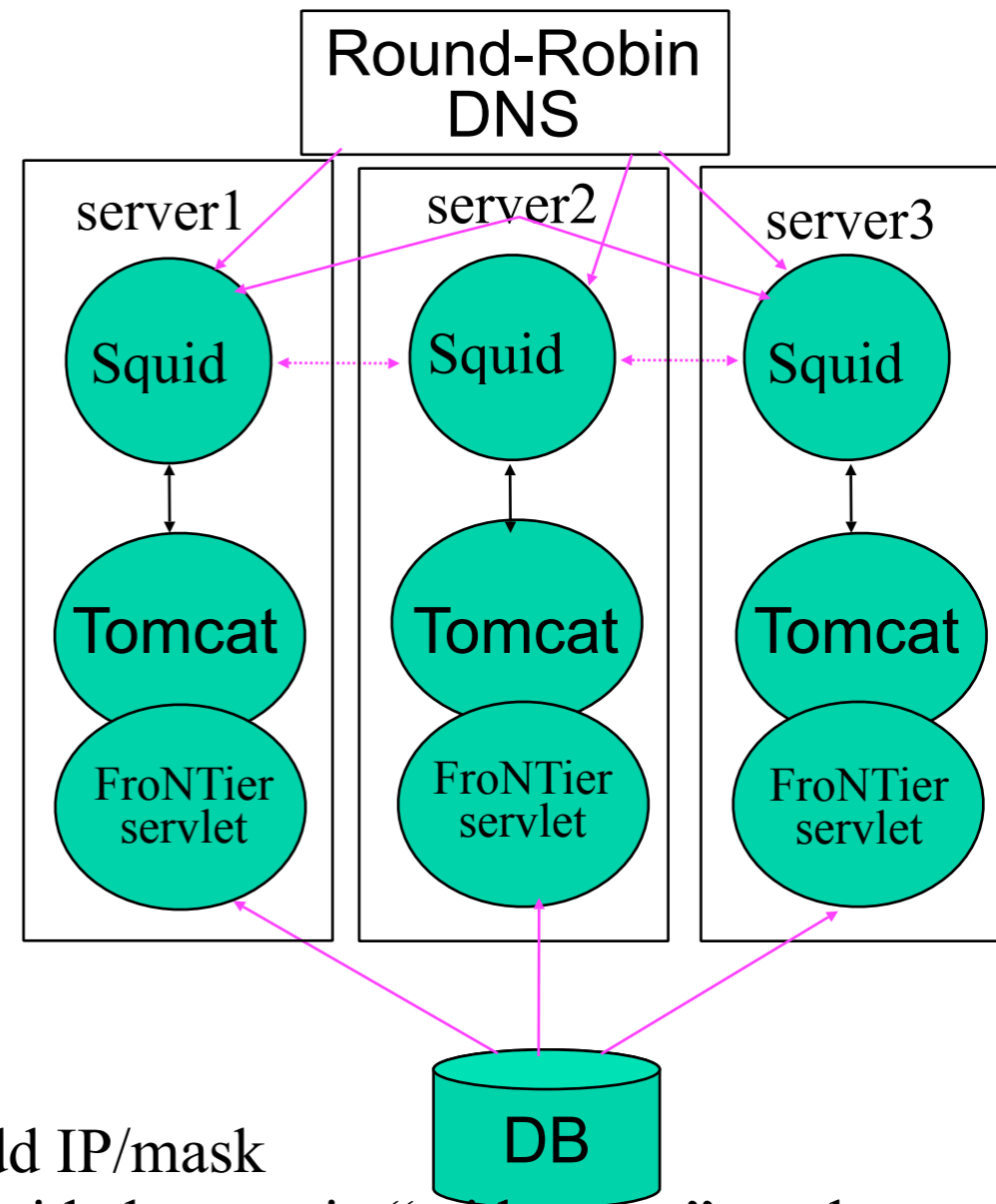
What's new in FroNTier lately

- Client can request the data be zipped by the server (compression levels 0-9)
- Keep alive signals sent to client when database is busy, avoids timeouts
- Ported to 64-bit Linux
- Parameters can come in long parenthesized connect string instead of environment vars
- Can define logical name in long string so pool file catalog can use short name

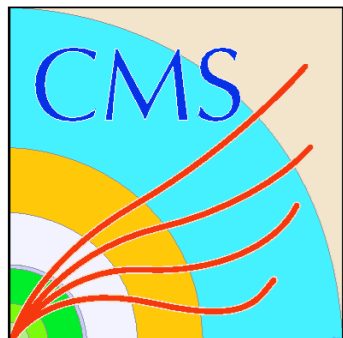


FroNTier “Launchpad” software

- Squid caching proxy
 - Load shared with Round-Robin DNS
 - Configured in “accelerator mode”
 - Peer-to-peer caching
 - “Wide open frontier”*
- Tomcat - standard
- FroNTier servlet
 - Distributed as “war” file
 - Unpack in Tomcat webapps dir
 - Change 2 files if name is different
 - One xml file describes DB connection



*In the past, we required the registration so we could add IP/mask to our Access Control List (ACL) at CERN. Recently decided to run in “wide-open” mode so installations can be tested w/o registration.



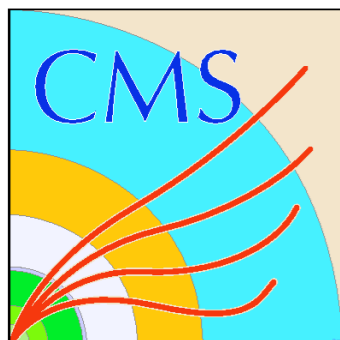
“Site” Squid details

- Hardware requirements
 - Minimum specs: 1GHz CPU, 1GByte mem, GBit network, 100 GB disk.
 - Needs to be well connected (network-wise) to worker nodes, and have access to WAN and LAN if on Private Network.
 - Having 2 machines for failover is a requirement for T-0/T-1, and a useful option for T-2. Inexpensive insurance for reliability.
- Software installation
 - Squid server and configuration
 - Site-local-config file All Details: <https://twiki.cern.ch/twiki/bin/view/CMS/CMSSquidDeployment>



Testing Sites

- ASGC
- Belgium
- CALTECH
- CERN
- CIEMAT
- CSCS
- DESY
- Estonia
- Florida
- FNAL
- GRIDKA
- Legnaro
- MIT
- Nebraska
- PIC
- Pisa
- Purdue
- RAL
- RWTH
- UCSD
- Wisconsin
- 21 sites were successfully tested
- 4 additional sites had various problems w/ squid and/or software.
- Additional 10 sites identified for install soon.



Testing Results (some examples)

Zipping turned off

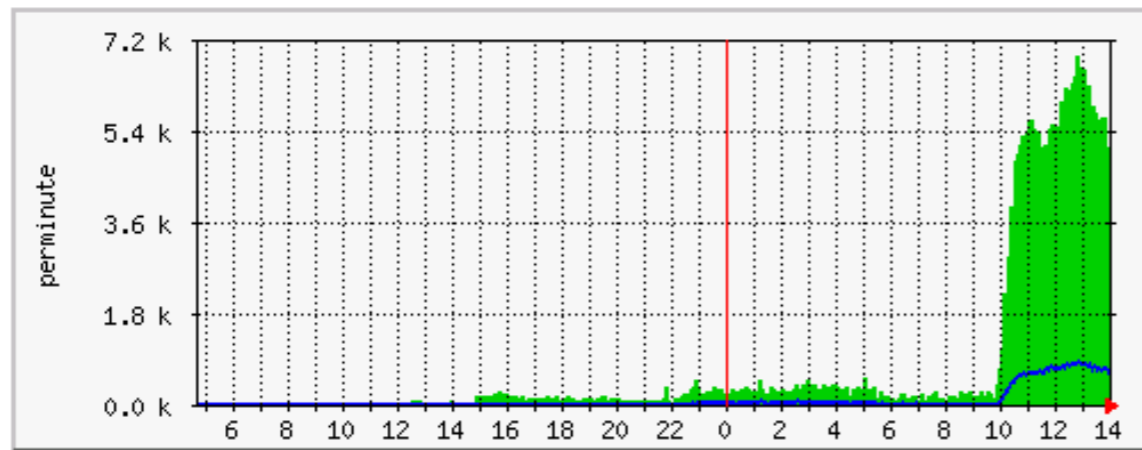
Site	Local Squid kBytes/sec	Frontier@CERN kBytes/sec
DESY	708+/-107	11.4+/-0.2
Estonia	152+/-63	6.3+/-0.8
Florida	612+/-13	40.3+/-0.4
FNAL	482+/-116	18+/-1.4
GridKa	565+/-151	23.1+/-2.3
Performance ranges	110 (Legnaro) to 826 (Belgium)	5.5 (ASGC) to 99 (Belgium)

<https://twiki.cern.ch/twiki/bin/view/CMS/CMS-Frontier-test-integration>



Monitoring Examples

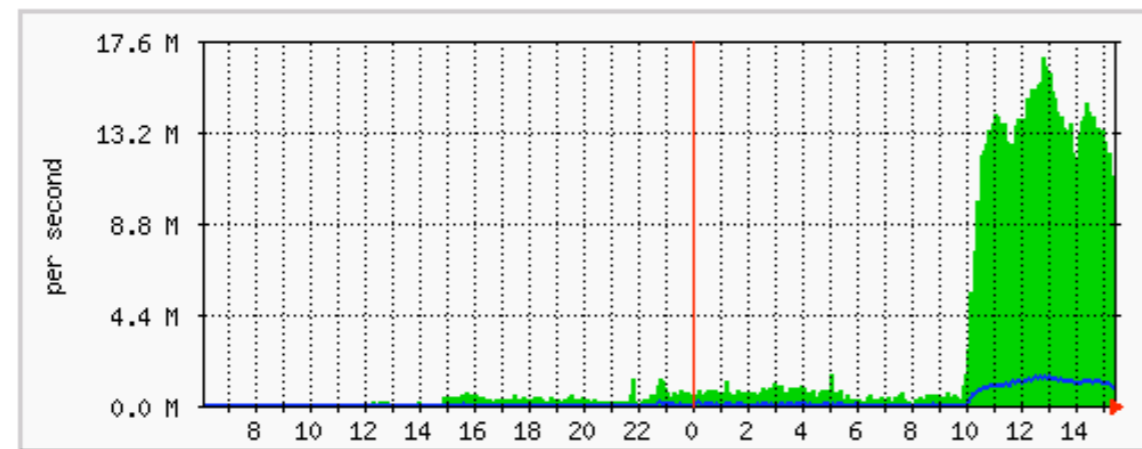
'Daily' Graph (5 Minute Average)



Max HTTP reqs 6924.0 req/min Average HTTP reqs 772.0 req/min Current HTTP reqs 4911.0 req/min
Max HTTP fetches 865.0 req/min Average HTTP fetches 105.0 req/min Current HTTP fetches 635.0 req/min

The green histogram shows the number of requests to the squid.
The blue line shows how often the request was not in the cache.

'Daily' Graph (5 Minute Average)

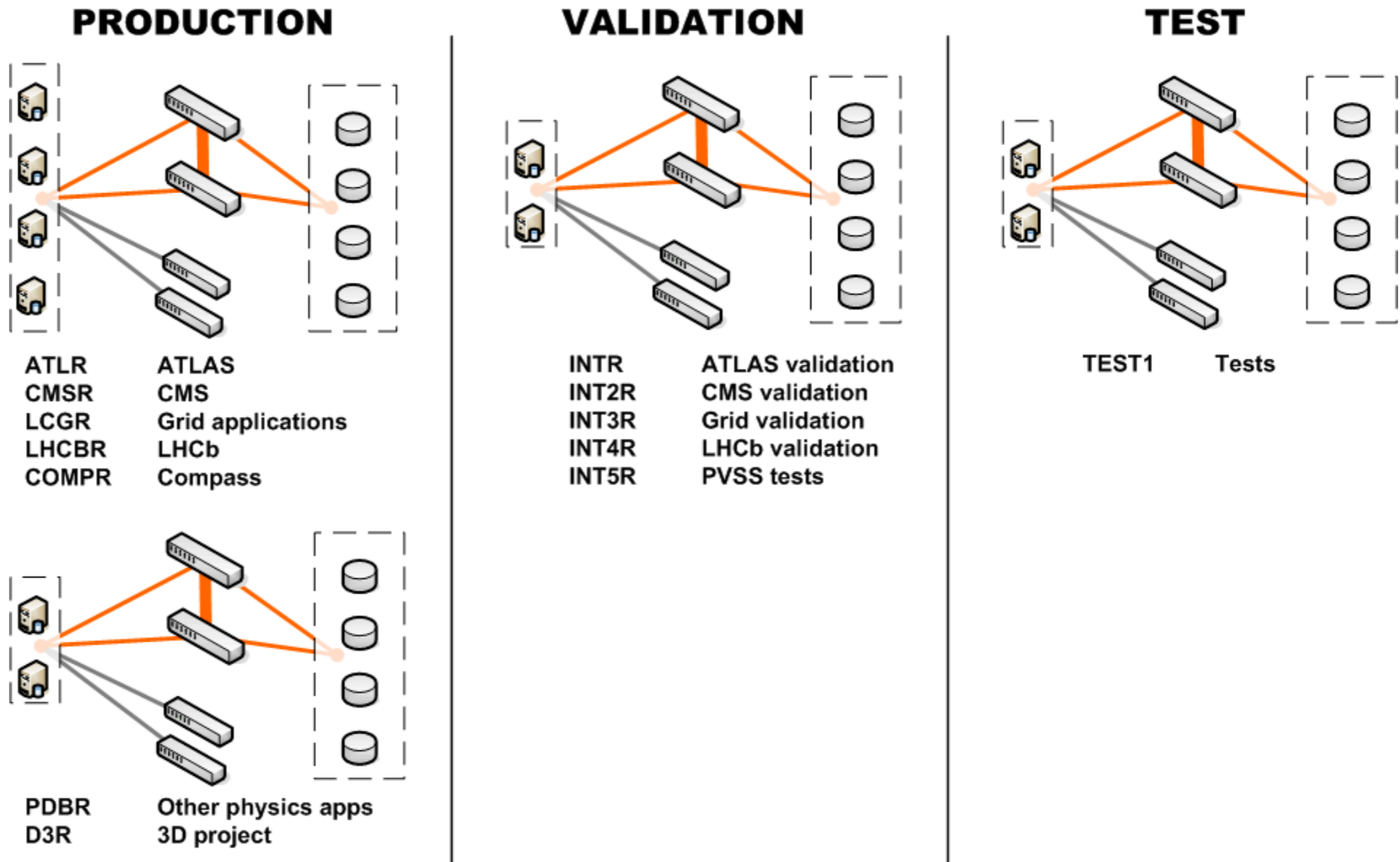


Max Total 16.9 MB/s Average Total 2500.0 kB/s Current Total 11.1 MB/s
Max Fetches 1529.0 kB/s Average Fetches 237.0 kB/s Current Fetches 948.0 kB/s

The green histogram shows how many bytes were delivered by the squid. The blue line shows how many bytes had to be retrieved from the source.

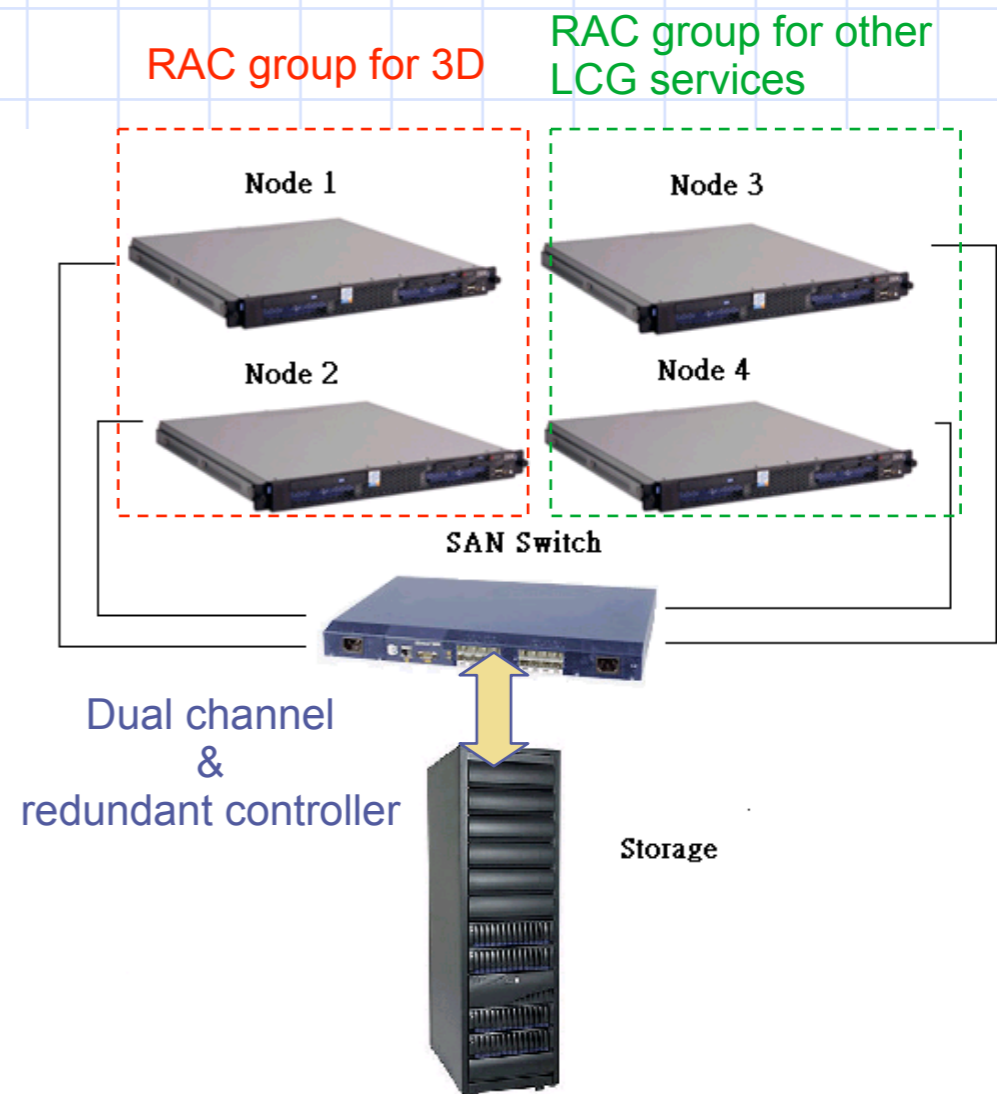


Oracle RAC on Linux project, CERN-IT-PSS 100 CPUs, 200GB RAM, 200TB disk



Hardware configuration

- **Four servers**
 - ✓ **CPU : Intel Pentium-D 830 3.0 GHz**
 - ✓ **Memory 2G (ECC)**
 - ✓ **Local Disk S-ATA2 80G 7200 rpm**
 - ✓ **Fiber Channel LSI 7102XP-LC, PCI X 1**
- **SAN Switch : Silkworm 3850 16 ports**
- **Backend Raid subsystem: StorageTek B280**
- **Each RAC group shares 1.7TB exported from SAN**



Current Status

- Streams replication
 - Replicate COOL and Tag databases (started)
- OEM Agents
 - Installed two agents connecting CERN Grid control
 - Installed two agents connecting local Grid Control
- Scheduled backup
 - Weekly incremental level 0 backup (disk)
 - Daily incremental level1 cumulative backup (disk)
 - Daily archive log backup (disk)
 - Dailly transfer to tape

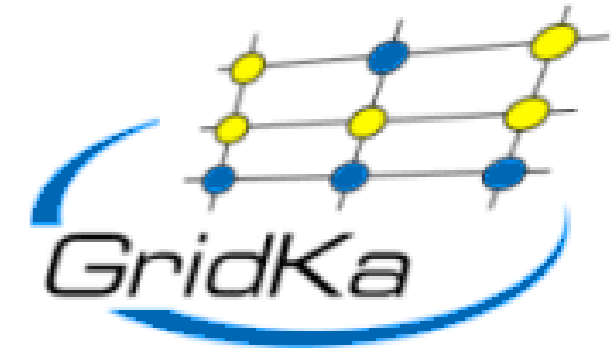
Present Resources Available

- ATLAS 2-nodes cluster
 - Dual Xeon 3,2 GHz, Qla2340 (QLogic dual port).
 - SAN storage: 900GB RAID5 Flexline partition
- LHCB 2-nodes cluster
 - Dual Xeon 3,2 GHz, Qla2340 (QLogic dual port).
 - SAN storage: 900GB RAID5 Flexline partition
- GRID 3-nodes cluster
 - DELL DL380 servers: dual Xeon 3,2 GHz QLE2462 (QLogic dual port, 4Gb).
 - SAN storage: 2 x 900GB RAID5 Flexline partitions
- All:
 - 4GB RAM two 73GB SCSI disks (RAID1), redundant power supplies, redundant FANs
 - *STK Flexline storage, RAID5 devices*
 - RedHat Enterprise 4 Update 4,ASM, Oracle 10.2.0.2

TEMPORARY



DB Hardware (in Operation)



2 RACs: 2-node-Cluster for ATLAS & 2-node-Cluster for LHCb

- 2x 2 IBM x336 machines
 - Intel Xeon dual 3.2 GHz w/ 2MB L2 cache
 - 4 GB RAM
 - 73 GB U320 hard disk
- QLogic HBA (database on SAN)
- 2x 2x548GB on StorageTEK DS280 connected by IBM INRANGE FC9000

Experiment	DB name	node1	SID1	node2	SID2	ClusterService
ATLAS	LCGDB1	f01-010-111	LCGDB11	f01-010-112	LCGDB12	LCGDB1crs
LHCb	LCGDB2	f01-010-113	LCGDB21	f01-010-114	LCGDB22	LCGDB2crs



- Each T1 site is responsible for installation and maintenance of the DB and Frontier servers according to experiment and project request
 - Requests/updates will be collected by 3D and presented to LCG GDB/MB for approval every 6 month
- The sites are responsible for
 - h/w server selection, acquisition, installation and monitoring as well as related network setup
 - s/w installation and upgrade according to the agreed evolution defined
 - regular application of security patches according to site policy
- Tier 0 is responsible for defining the streams



- In case of h/w problems of the squid setup sites are required to replace unavailable nodes, but not the cached data.
 - No cache backup is required
- In case the squid cache becomes inconsistent (eg after a power failure) sites may clear the cache (following procedures defined by CMS)
- The Frontier production setup at T0 is operated by FIO (box level) and the FNAL Frontier team (tomcat & squid)
 - Do we need to review this?



- The sites are responsible for recovery from unavailability/inconsistency caused by power or h/w problems
 - Worst case: re-import from T0 and streams resync of one site
 - Need to exercise this at each site!
- The application owners are responsible for recovering from logical corruption caused by their s/w packages
 - Worst case: point-in-time recovery and re-sync on all affected sites
 - Need to schedule a full size recovery test to estimate the unavailability caused by this.



- Each site is responsible to setup database backup and recovery infrastructure via Oracle RMAN
 - This may include on-disk backups and should include tape backups and associated media
- Backups should be performed online and with a retention period which is compatible with the time window for point-in-time recovery required by the experiments
 - Eg 1 month or 3 month?
- This is required to allow for a standard recovery procedure including streams re-synchronisation
- The T1 sites are responsible for performing



- The sites are responsible for staffing their support teams to meet the problem response time and availability numbers defined by the LCG MoU
 - Work split between DBA and other support staff should be organised by each site
- Availability numbers there are only defined indirectly
 - eg for high level application types
 - need to correlate service and database availability monitoring to understand required DB availability
- If application code does not implement DB retry/failover then the DB availability



- Security monitoring and patching is entirely site responsibility
 - Each site needs to monitor for security incidents and apply security patches according to site policy
 - This includes emergency interventions like change of compromised admin credentials
- 3D will make information about content and schedule of security patches available and collect experience with their application at the sites
 - But their selection and application is with the sites
- The application of security patches does not



- Should use as much as possible the established reporting channels
- Each site should now join the GGUS support setup
 - Expect that all service users will report problems via this channel
- Each site should pre-announce service changes/outage
 - `grid-service-databases@cern.ch`
(new list for any 3D service related discussions)
 - EGEE broadcast and other LCG lists (gmod)
- Integration with operations meetings
 - Either: All sites database team are represented via their grid teams

System evolution and experience in building Oracle RAC system



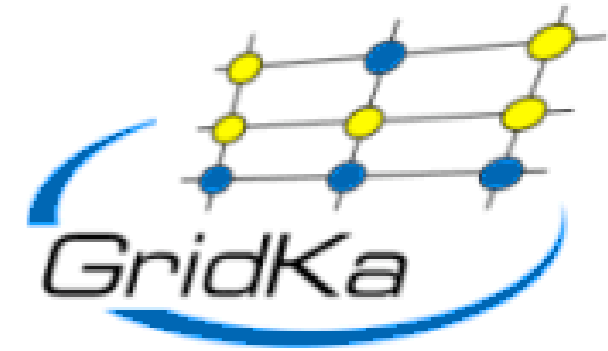
CPU	Pentium-4 3.2GHz		Itanium-II IA64 1.5GHz	Pentium-D 380 3.0GHz
OS	SLC3	SLC4	SLC4	SLC4
Nodes(#)	2	2	2	2
ASM Config	NO	YES	YES	NO
OCFS Config	YES	NO	NO	NO
OCFS2 Config	NO	YES	YES	YES

Present Resources Available

- Streamtest Cluster:
 - 2-nodes RAC (but potentially 4-nodes)
 - Xeon 2.4 GHz, 2GB RAM, 80GB disks (RAID1), Qla2312 (single port).
 - One 900GB RAID5 FastT900 partition
 - OCFS2
 - ***Will become our Test RAC***
- Oracle01: *castorstager DB*
 - Dual Xeon 3.6 GHz, 4GB RAM, 6 SCSI disks in RAID5, single instance DB.
- diskserv-san-13: dlf DB
 - Will be migrated to a new machine before October. Single instance DB.



DB status



- **Manpower**

- 1 DBA (25%): Doris.Wochele@iwr.fzk.de
- 1 DBA (100%): Andreas.Motzke@iwr.fzk.de (just started)
- Silke.Halstenberg@iwr.fzk.de (deputy)

- **Streams**

- tests with ATLAS and LHCb done

- **Problems**

- memory settings – *solved* by reducing `streams_pool_size`

`ORA-04031: unable to allocate 16 bytes of shared memory ("shared pool","select obj#,type#,ctime,mtim...","sql area","kglhin: temp")`

- ORA-Messages **unresolved**

`ORA-10388: parallel query server interrupt (failure)`

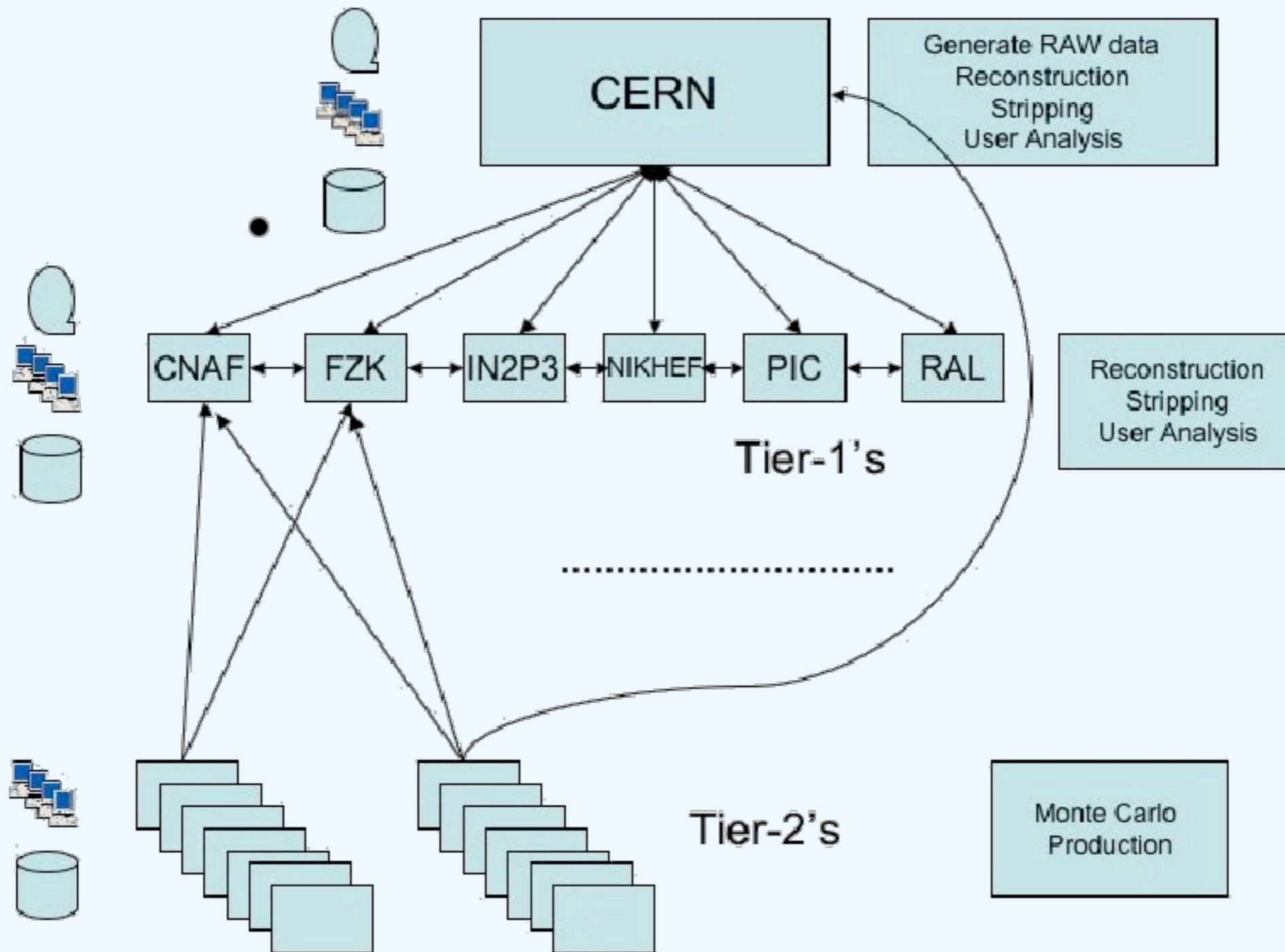
`WARNING: inbound connection timed out (ORA-3136)`

- **To do**

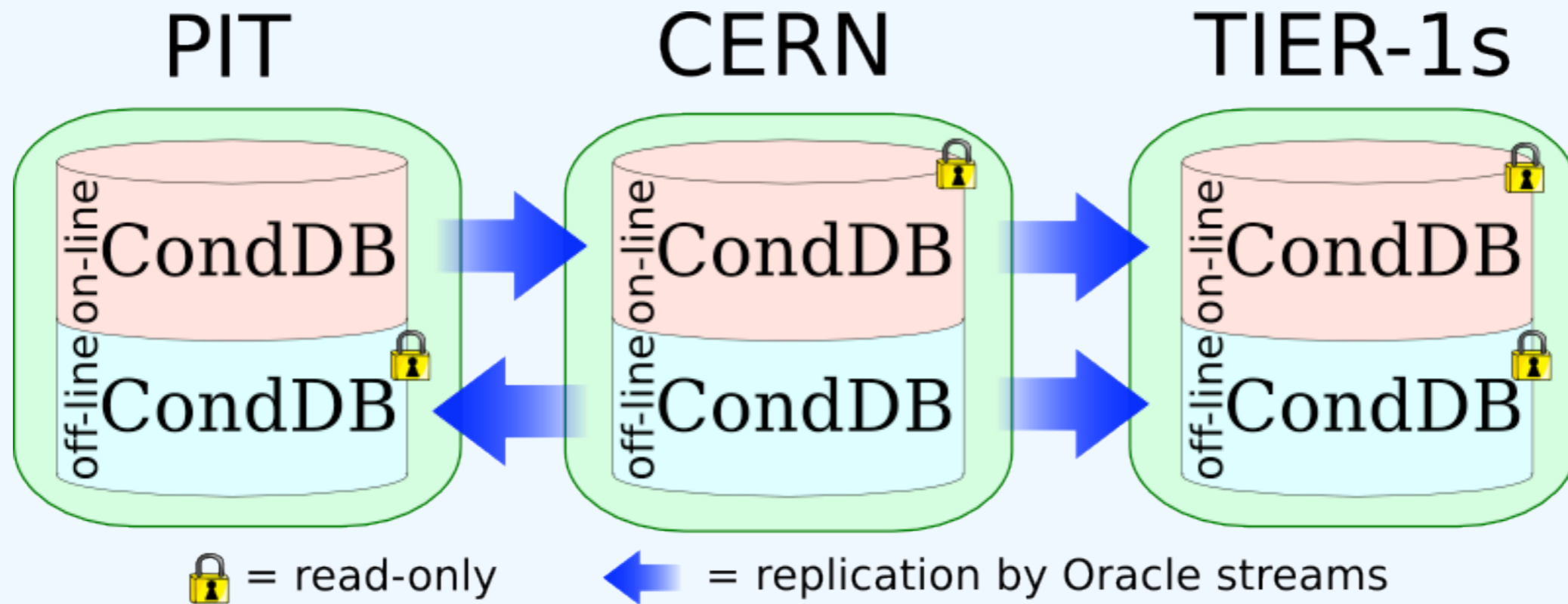
- scheduled backup
- OEM agent installation for own Grid Control
- check parameter settings (`sga_target` ?)

Andreas Motzke - GridKA

Computing Model



Deployment Strategy



- Prepare the Master CondDB
- Replicate to 3 Tier-1s
- Access the replicas from the GRID
- Set up (fake) PIT Oracle server
- Replicate from PIT to CERN and from CERN to Tier-1s
- Add the missing Tier-1s
- Full scale test on the GRID

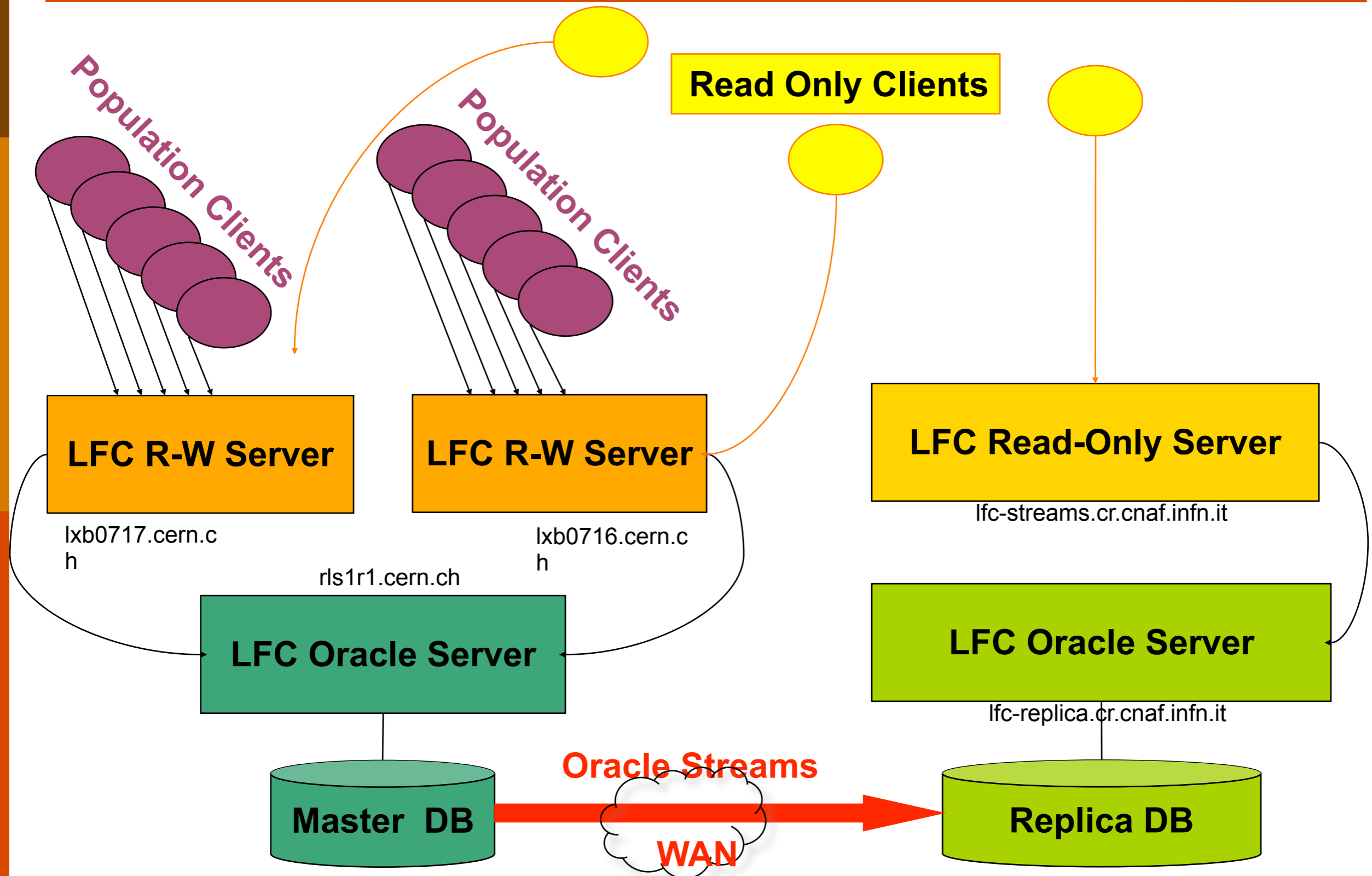
Achievements and Status (1)

- Started with RAL and GridKa
 - Creation of the schema (~10k tables)
 - 6 hour delay at RAL, memory problems at GridKa
 - Privileges replication
 - long delay few hours
- GRID access
 - Only CERN tested (no DB replica catalog)

- Added IN2P3
 - Successful export of the schema to a new slave
- Stress Test
 - 100 insertions/s distributed over 200 folders
- (fake) PIT Oracle server
 - Under preparation
 - Schema not yet imported

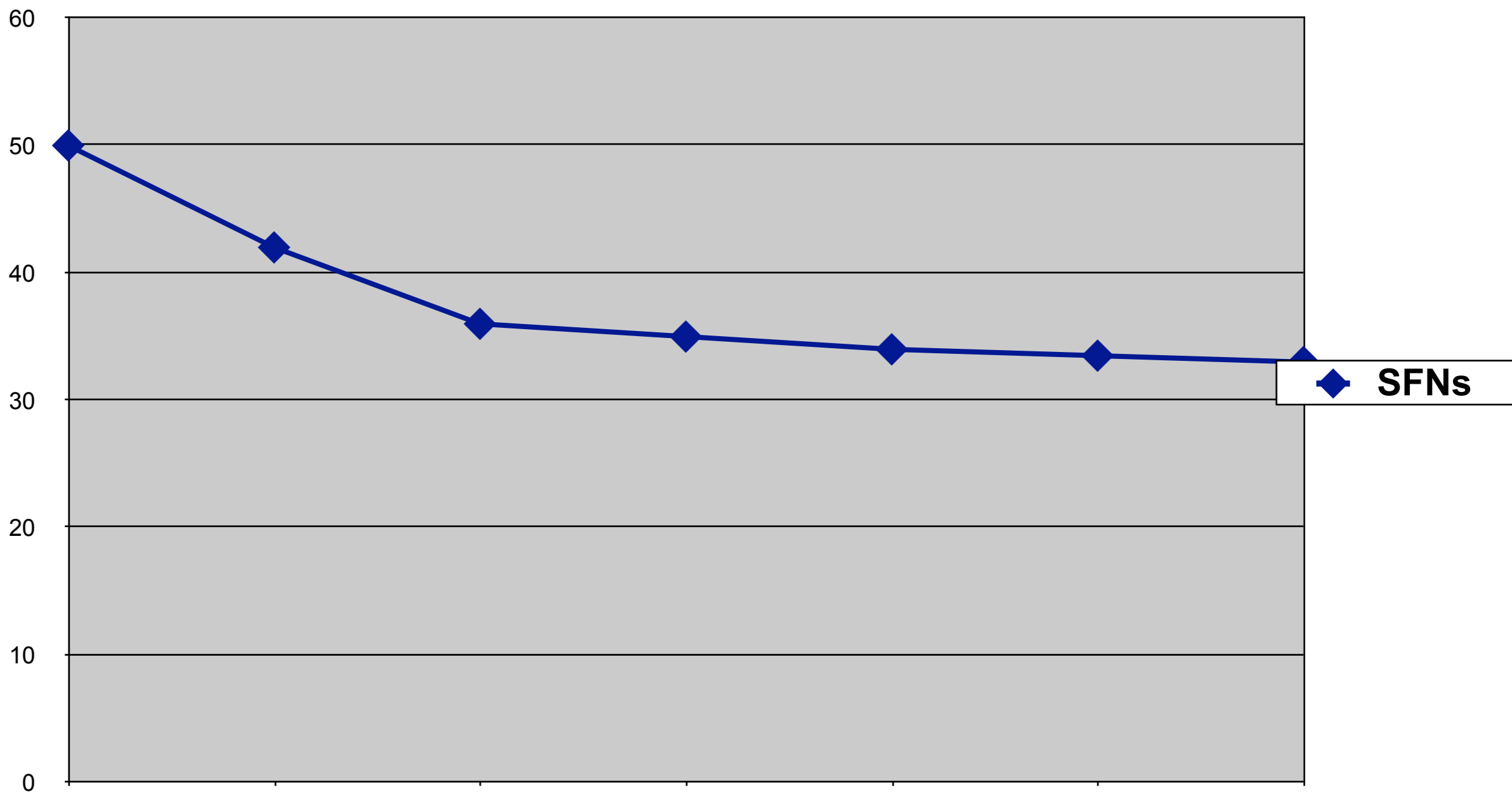
- The replication seems efficient enough for normal usage
- Slow for management tasks, but is is not an issue
- Still to do
 - Stress test on the GRID (~200 jobs/site)
 - Using fully featured CORAL library
 - 2 steps replication: PIT \Leftrightarrow CERN \Rightarrow Tier-1s
 - Replication to all sites (6 Tier-1s)

LFC Replication Testbed



Test 2: 20 Parallel Clients

- 20 parallel clients equally divided between the two LFC master servers.
- Inserted 3000 replicas per minute, 50 replicas per second.
- Apply parallelism enhanced: 4 parallel apply processes on the slave.
- After some hours the rate decreases, but reaches a stable state at 33 replicas per second.
- Achieved sustained rate of 33 replicas per second.
- No flow control on the master has been detected.



Replication Strategies at ATLAS

- Geometry database
 - Update frequency: several month
 - Replication with SQLite files
- Conditions database
 - Update frequency: seconds to hours
 - Replication with ORACLE streams
- Event data
 - Update frequency: 25 ns
 - Replication with DDM

Conditions Database Replication

- Conditions database mostly read by Tier-0/1
 - According to TDR Tier-2 will do mostly MC
- Writing to conditions database only at CERN
- Expected data volume: ~1TB/year

- TAGS database will have similar access pattern and data volume
 - Production will produce root files, inserted to database at CERN

Testing Environment for Streams

ATLAS pit

Online /
HLT
farm

ATLAS
Online RAC

Online
CondDB

ATLAS_COOL_3D

TAGS



INTR

ATLAS
Validation RAC

Oracle
streams

ATLAS_COOL_3D
ATLAS_TAGS_3D

BNL

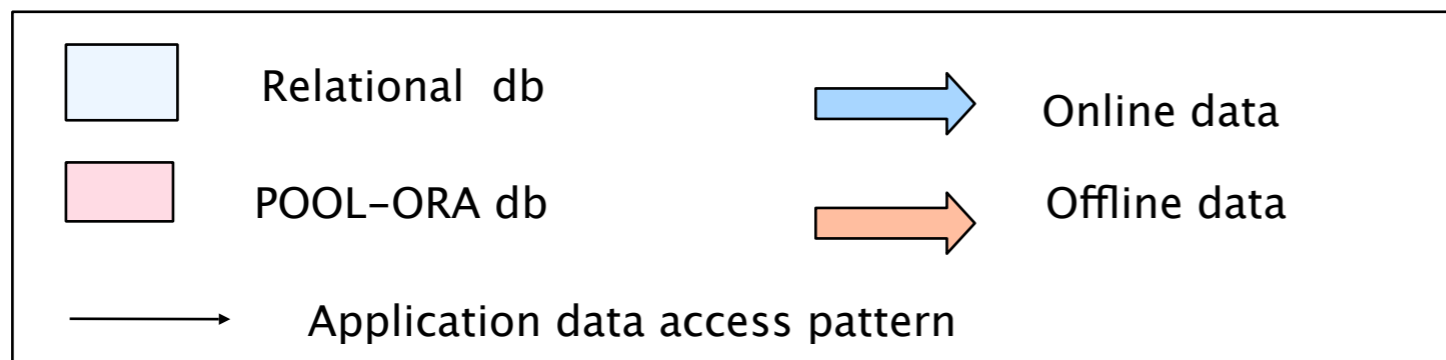
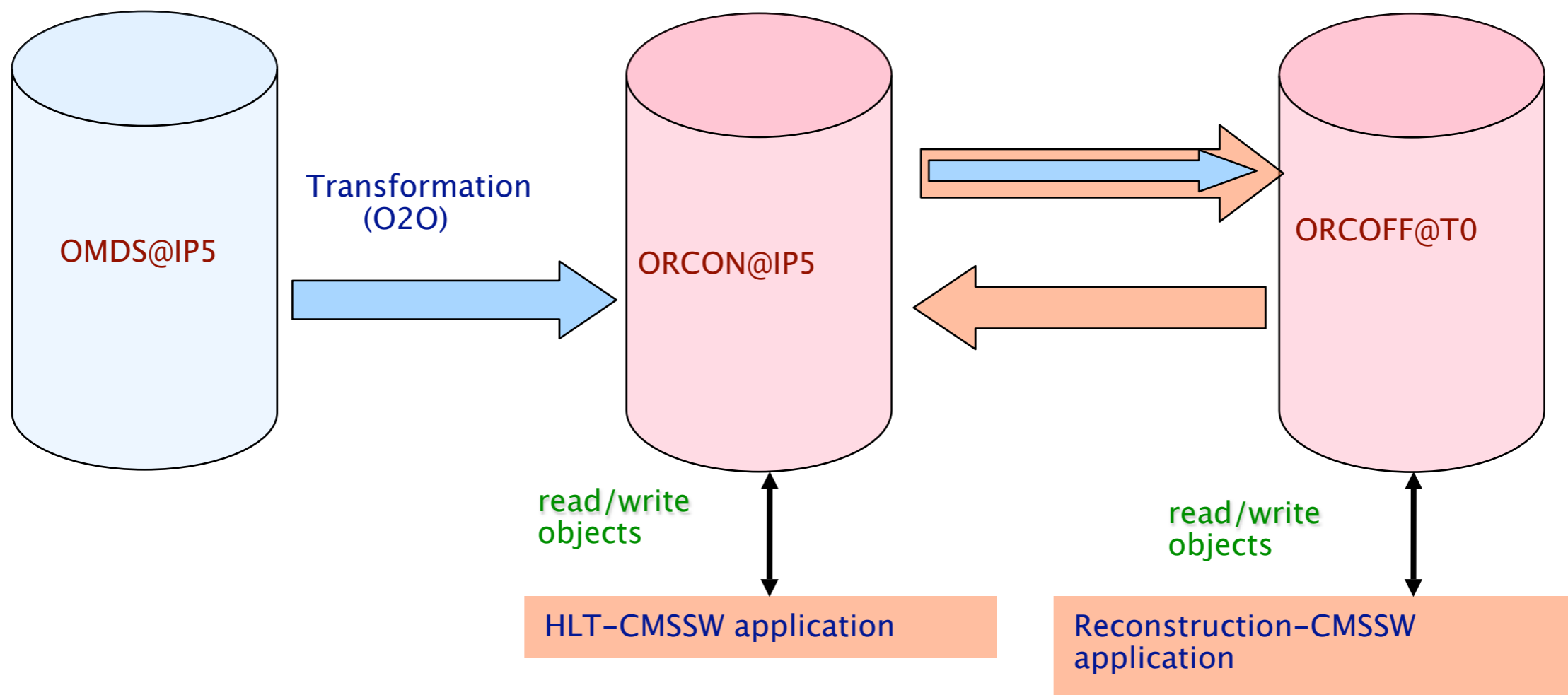
Gridk
a

ASC
G

Test load on the online RAC will be done from the ATLAS pit using David Front's Verification client and Stefan Stonjek's client



CMS O2O



slide by Zhen Xie



Conclusions



- With the help of IT our streaming performance has improved by 2.4x
- Somewhere along the line the Database Link performance fell by almost a factor of 10
 - Should be investigated
- Experience gained from these tests are valuable input for setting realistic requirements and improving our conditions data transfer/deployment model