



Use of MySQL in the ALICE data-acquisition system

sylvain.chapeland@cern.ch

PH/AID/DA

Database Futures Workshop
6-7 June 2011
CERN

Outline

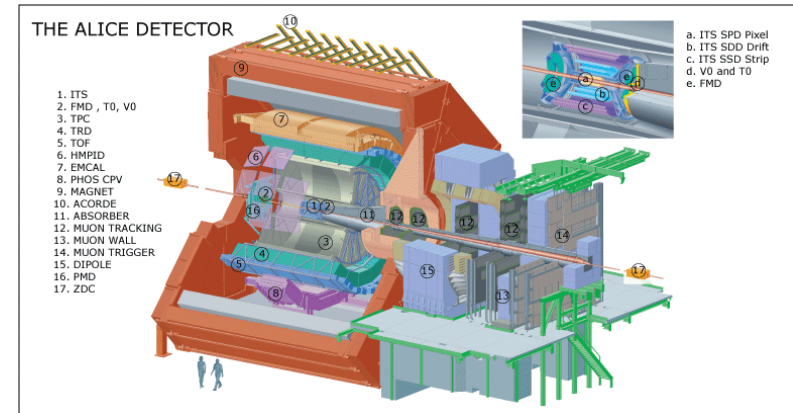
- Introduction
- The DAQ databases: description and access pattern
- Hardware and benchmarks
- Software: interfaces, features
- Conclusion

Introduction

- ALICE

- 18 sub-detectors
- 5 online systems

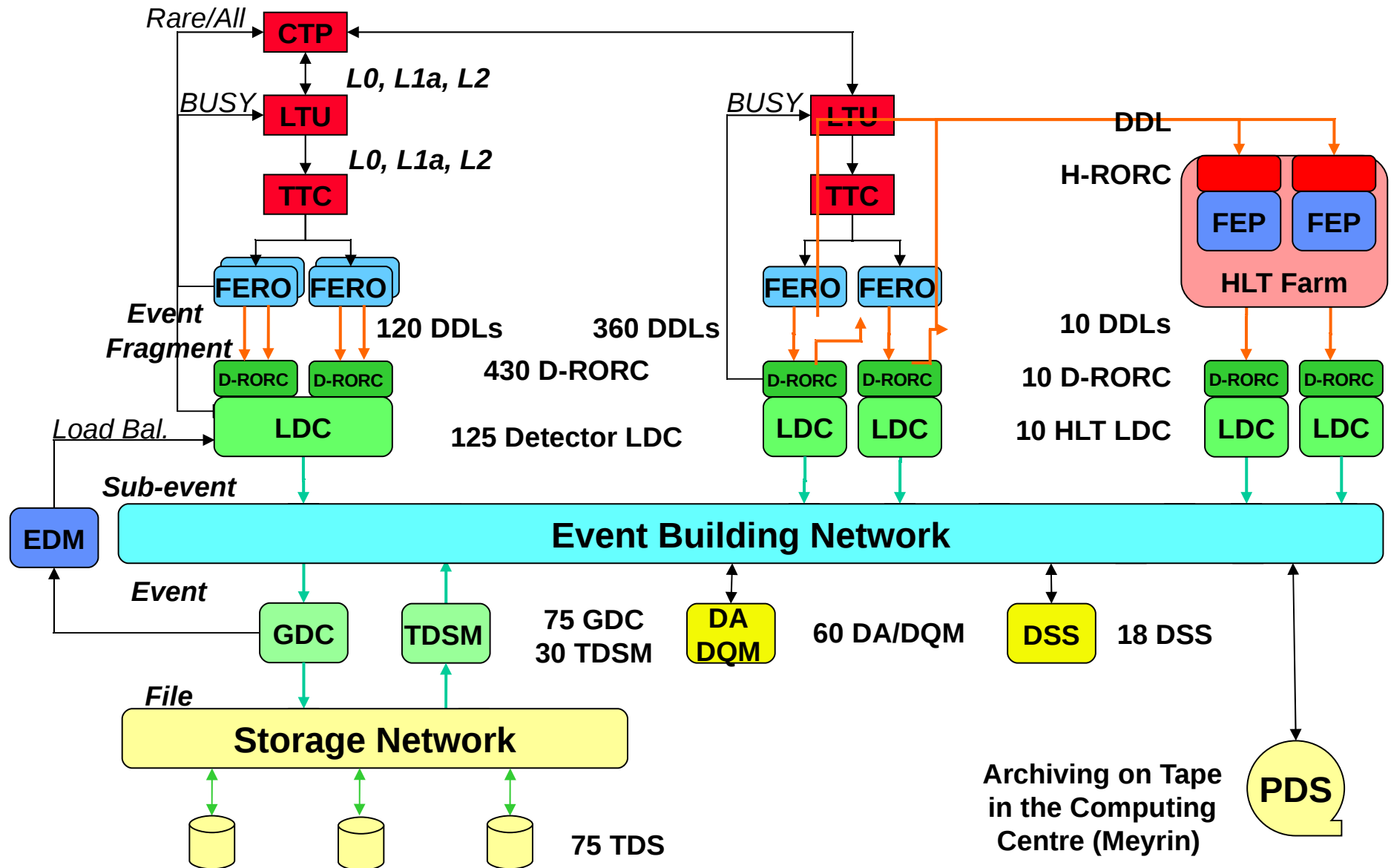
ECS, DAQ, DCS, TRI, HLT



- DAQ in numbers

Readout bandwidth: 115 GB/s, plan up to 50GB/s used with HLT filter
 Event Building bandwidth: 8 GB/s
 In 2010: 1.2 GB/s in p-p up to 2.5 GB/s in Pb-Pb
 Storage bandwidth performance:
 4.5 GB/s writing, 2.5 GB/s reading and archiving to CASTOR
 Amount of data recorded in 2010
 Physics : 1.6 PB 1.8 x 10⁹ events 1000 hours data taking
 All : 4.7 PB 23 x 10⁹ events 11000 hours data taking
 Facilities
 460 Detector Data Links (DDL) in, 360 out (to HLT)
 400 PCs
 180 TB transient storage (soon 400TB)

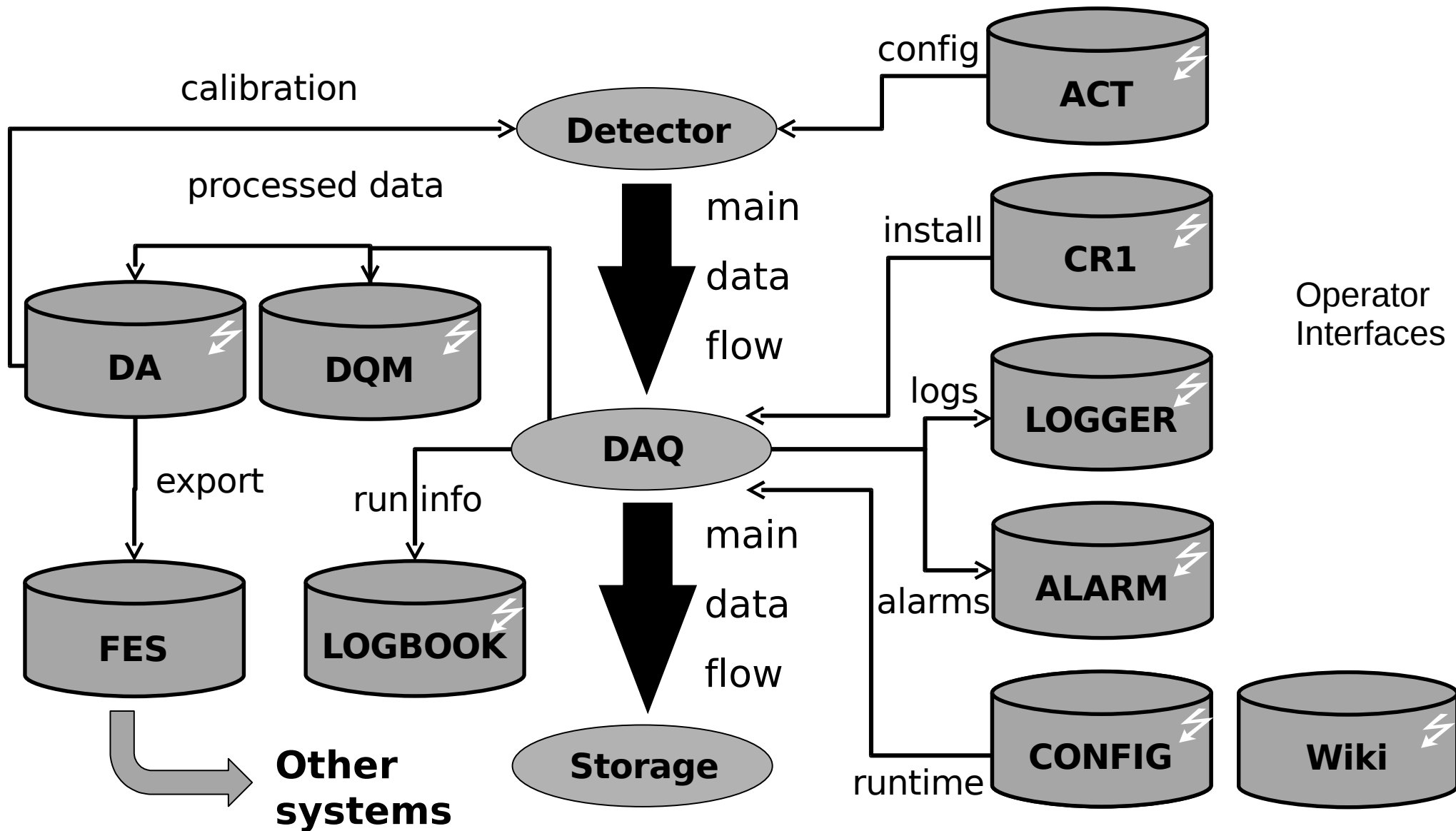
DAQ components



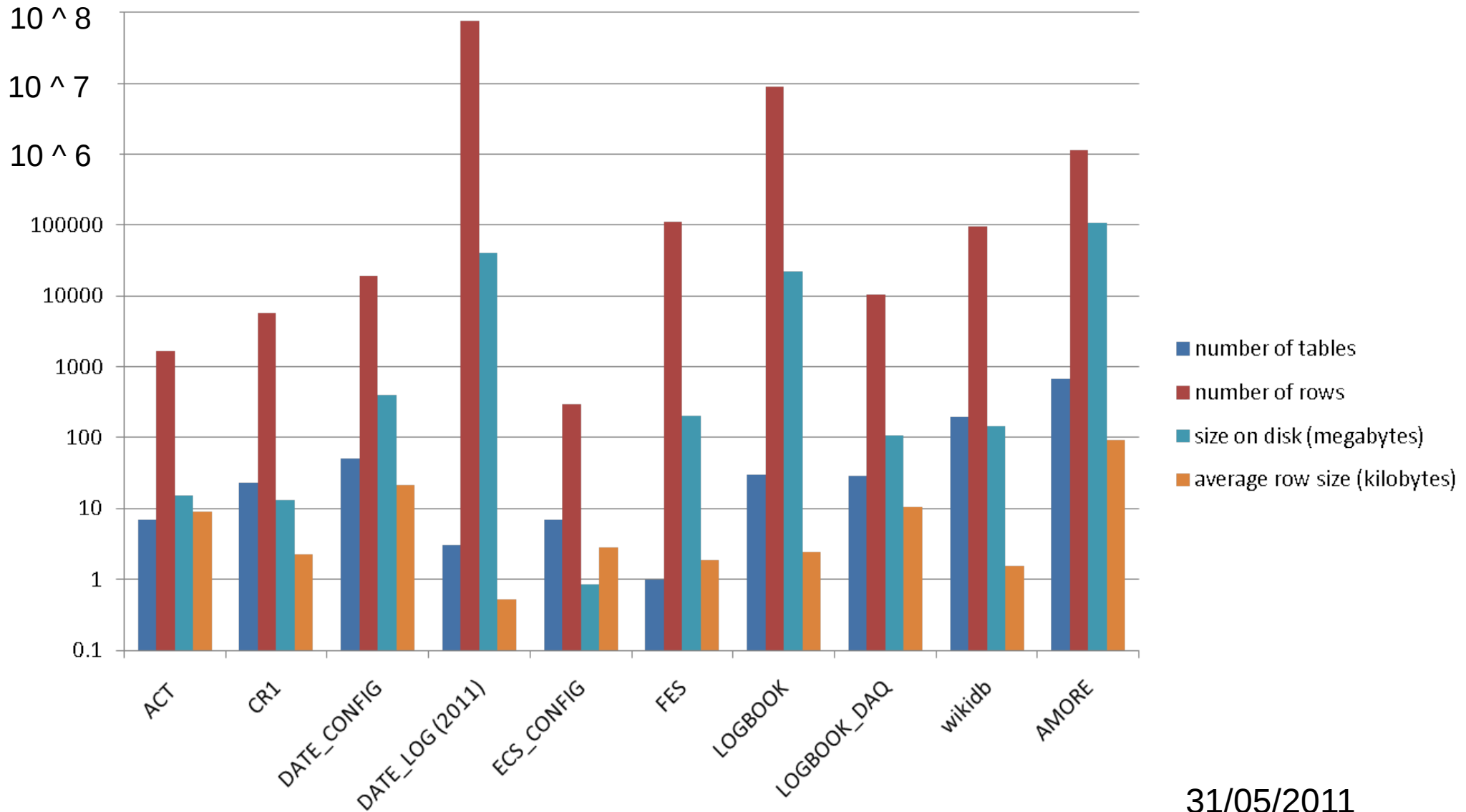
DAQ DB history

- Introduction of databases in 2004 to replace text-based configuration files
- MySQL selected based on
 - Performance
 - Lightweight installation for our multiple sites
 - Ease of use and maintenance
 - Know-how
- Extensive use in many other DAQ components
 - DB now indispensable to data taking
- Deployment
 - P2, lab, ~10 developers sites, ~20 user sites (development + production test beams)

DAQ databases overview



DB content



31/05/2011

DB usage profiles

- CONFIG, CR1, ACT
 - Low data volumes, total data size is few MB, 20000 rows (with some rare larger entries)
 - Read peaks (installation, Start Of Run), 100s of concurrent clients
- LOGGER
 - Write intensive, single insert client
 - Peaks (Start/End Of Run), ~10000 inserts in few seconds
 - Full indexing for field search by interactive clients.
 - Millions of rows per week, ~300MB/day, archiving
 - Large query results
- DQM
 - Large data volume (objects in MegaBytes), total data > 100GB
 - Concurrent read/write
- LOGBOOK
 - Complex queries
 - Distributed insert/replace
 - Increasing size (now 22GB)

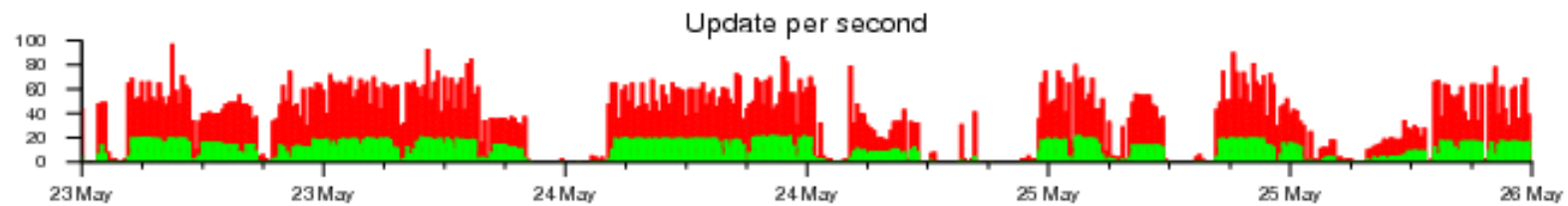
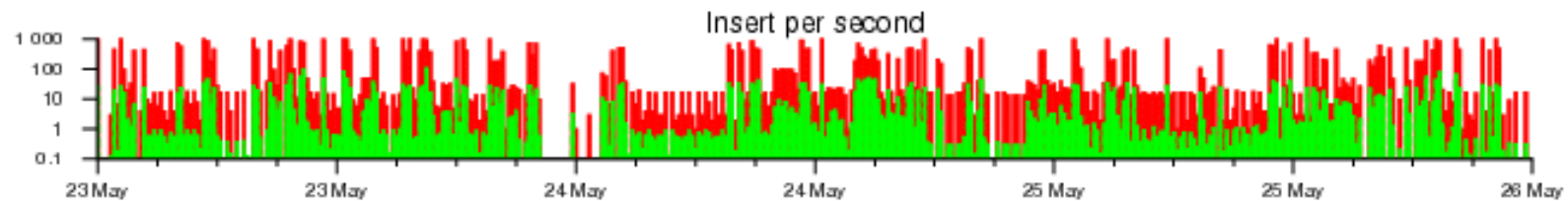
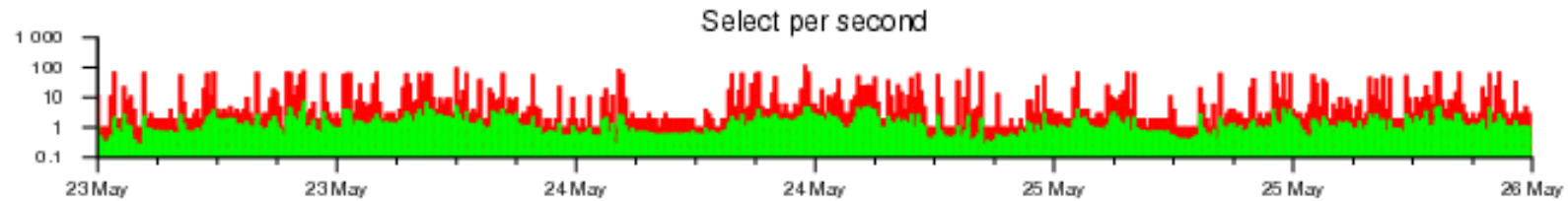
Server monitoring

- 1 server (aldaqdb) hosting all databases but DQM
- 1 server hosting only DQM
- Values presented:
 - Server variables sample time = 10s
 - Peak: max value per sample period
 - Average: average on 10 minutes

Queries

Database node 'aldaqdb'- MySQL statistics

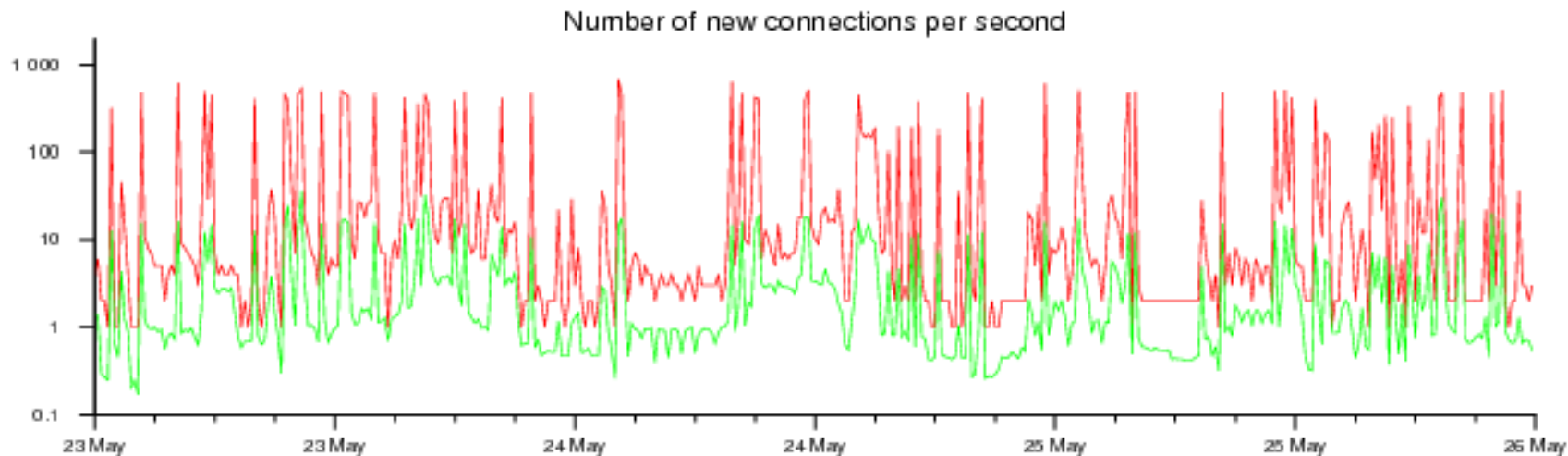
max 
average 



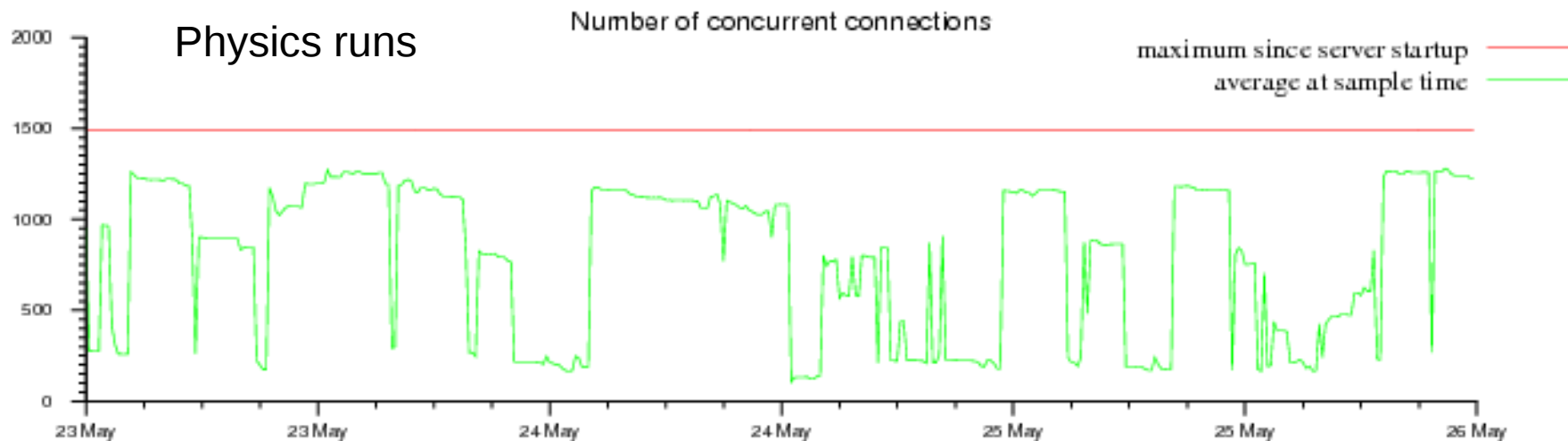
Connections to server

Database node 'aldaqdb'- MySQL statistics

max ———
average ———

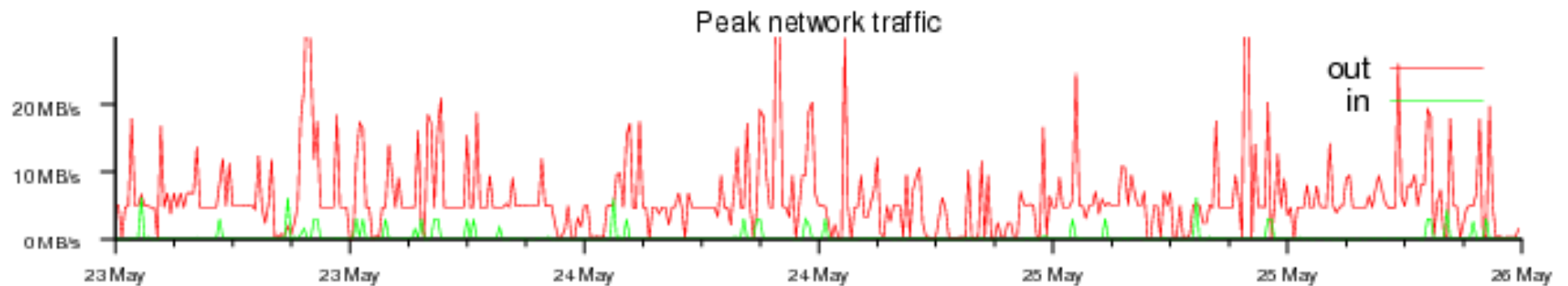


Start of run
Physics runs



Network traffic

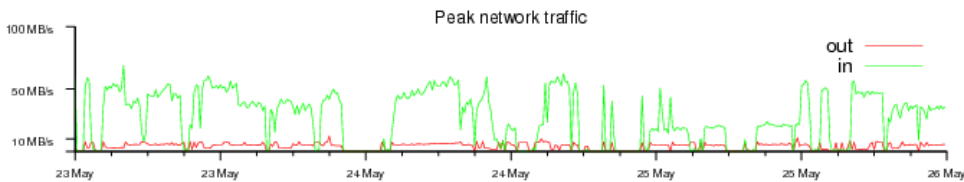
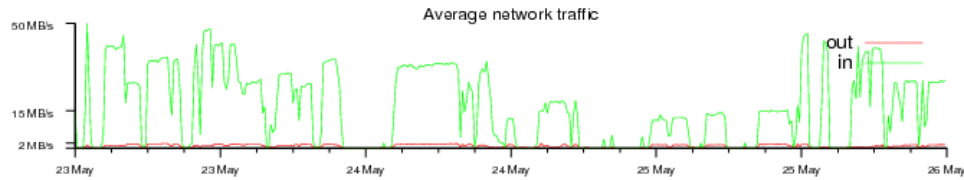
Database node 'aldaqdb' resource usage



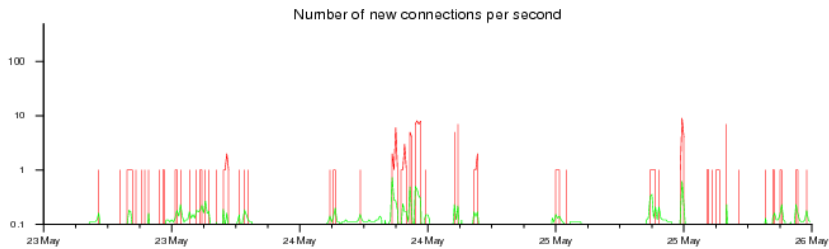
Daily backup

DQM DB

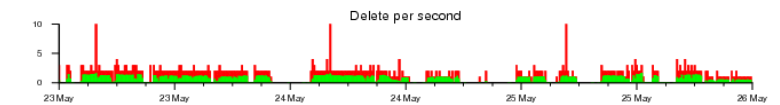
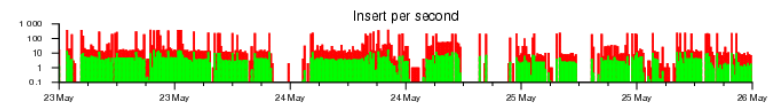
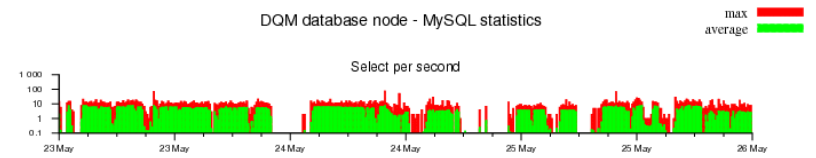
DQM Database node resource usage



DQM database node - MySQL statistics

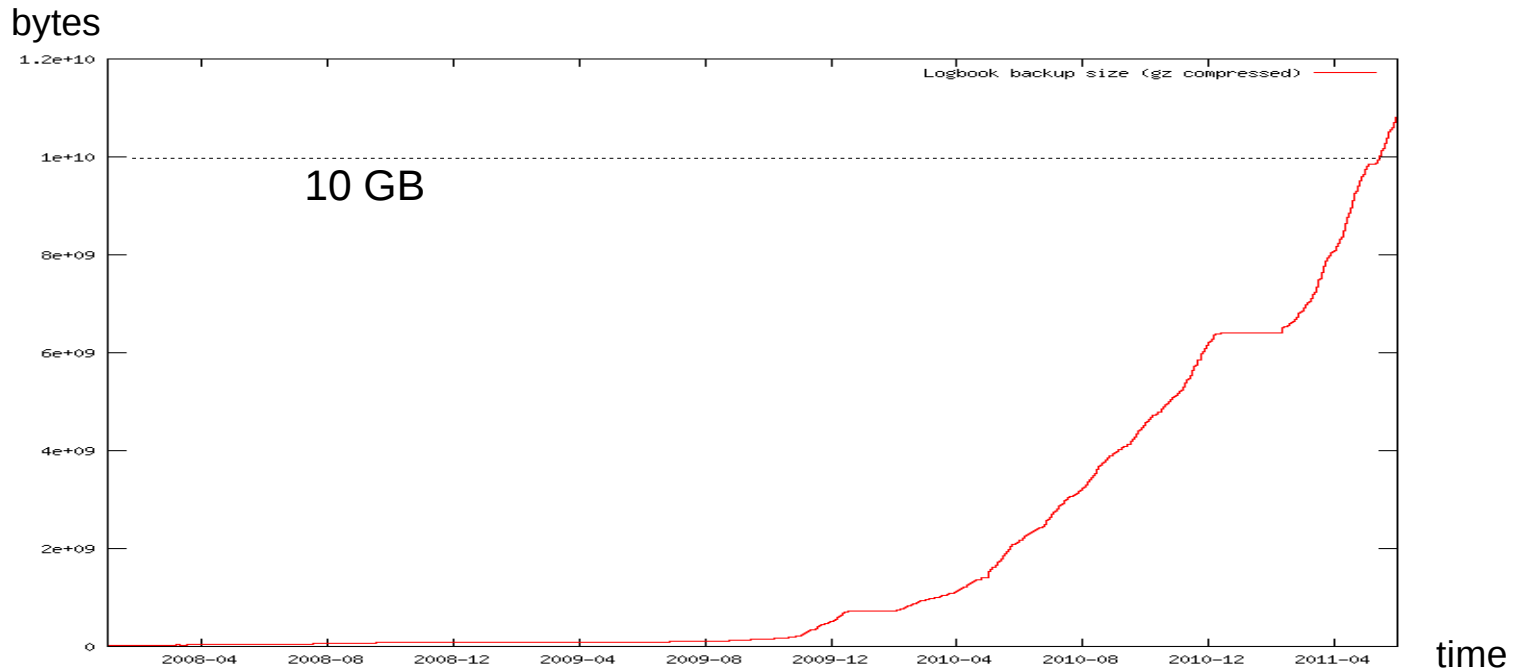


DQM database node - MySQL statistics



Different pattern:
more I/O, less clients

Data keeps growing



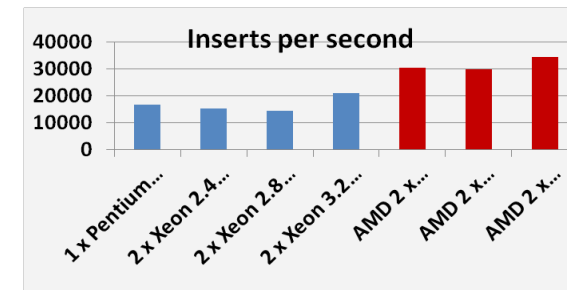
- LOGBOOK data increase with number of runs
- Size on disk = 2x compressed backup, 22GB at the moment, doubled in past 6 months (physics data taking + new features)
- Some large objects will be removed and stored as normal files
- Daily backup was heavy in the end, with previous hardware
- Replication was fine as replacement
- Various strategies to reduce size / availability, e.g. separation of online data and split history

Hardware

Extensive benchmarks performed to select production hardware

2006 (2007 in production)

- 2x dual-core AMD Opteron 275 @ 2.2 GHz
- FiberChannel RAID6 disk array 500GB
- SLC4-64



2010 (2011 in production) – HP DL380G7

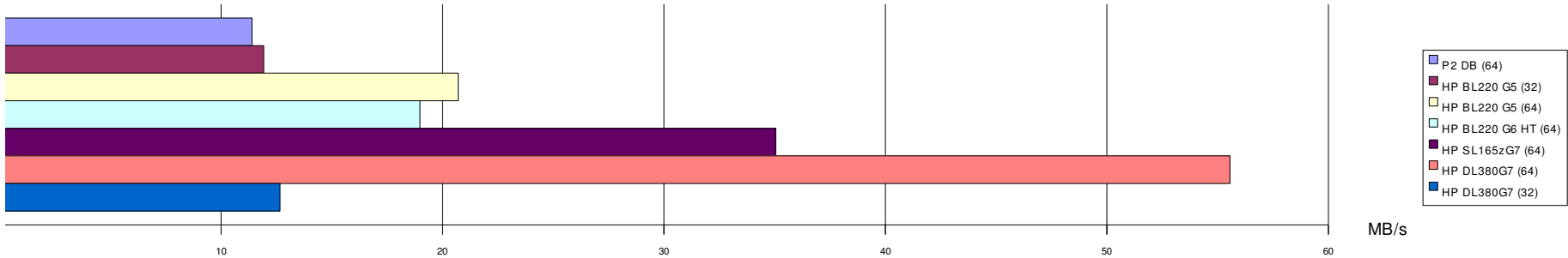
- 2x quad-core Intel Xeon X5677 @ 3.5 GHz
- internal RAID 0 + 1 SAS 15k 200GB
- Disk controller with flash write cache backup
- SLC5-64

2010 benchmarks - machines tested

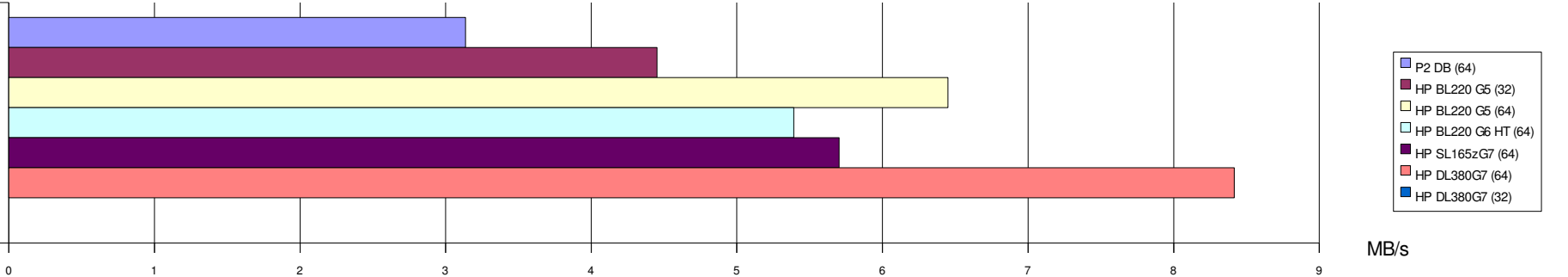
Type	Description	CPU	Cores (physical - logical)	Memory	OS
P2 DB (64)	P2 DB (2007)	2x 2 cores AMD Opteron 275 @ 1.00 Ghz	4 – 4	6G	SLC4 64
HP BL220 G5 (32)	blade G5	2x 4 cores E5450 @ 3.00GHz	8 – 8	16G	SLC4 32
HP BL220 G5 (64)	blade G5	2x 4 cores Intel E5450 @ 3.00GHz	8 – 8	16G	SLC5 64
HP BL220 G6 HT (64)	blade G5	2x 4 cores HT Intel E5530 @ 2.40GHz HyperThreading ON	8 – 16	16G	SLC5 64
HP SL165zG7 (64)	AMD G7	2x 12 cores AMD Opteron 6174 @ 2.20 Ghz	24 -24	32G	ubuntu 64
HP DL380G7 (64)	Intel g7	2x 4 cores E5640 @ 2.67GHz	8 – 8	24G	ubuntu 64
HP DL380G7 (32)	Intel g7	2x 4 cores E5640 @ 2.67GHz	8 – 8	24G	SLC4 32

Benchmark results – single threaded applications

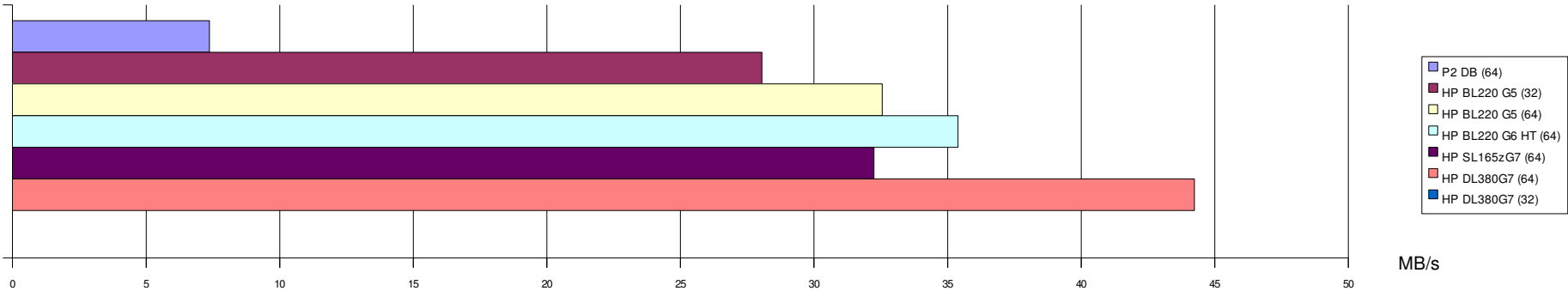
word count, file in memory



Gunzip

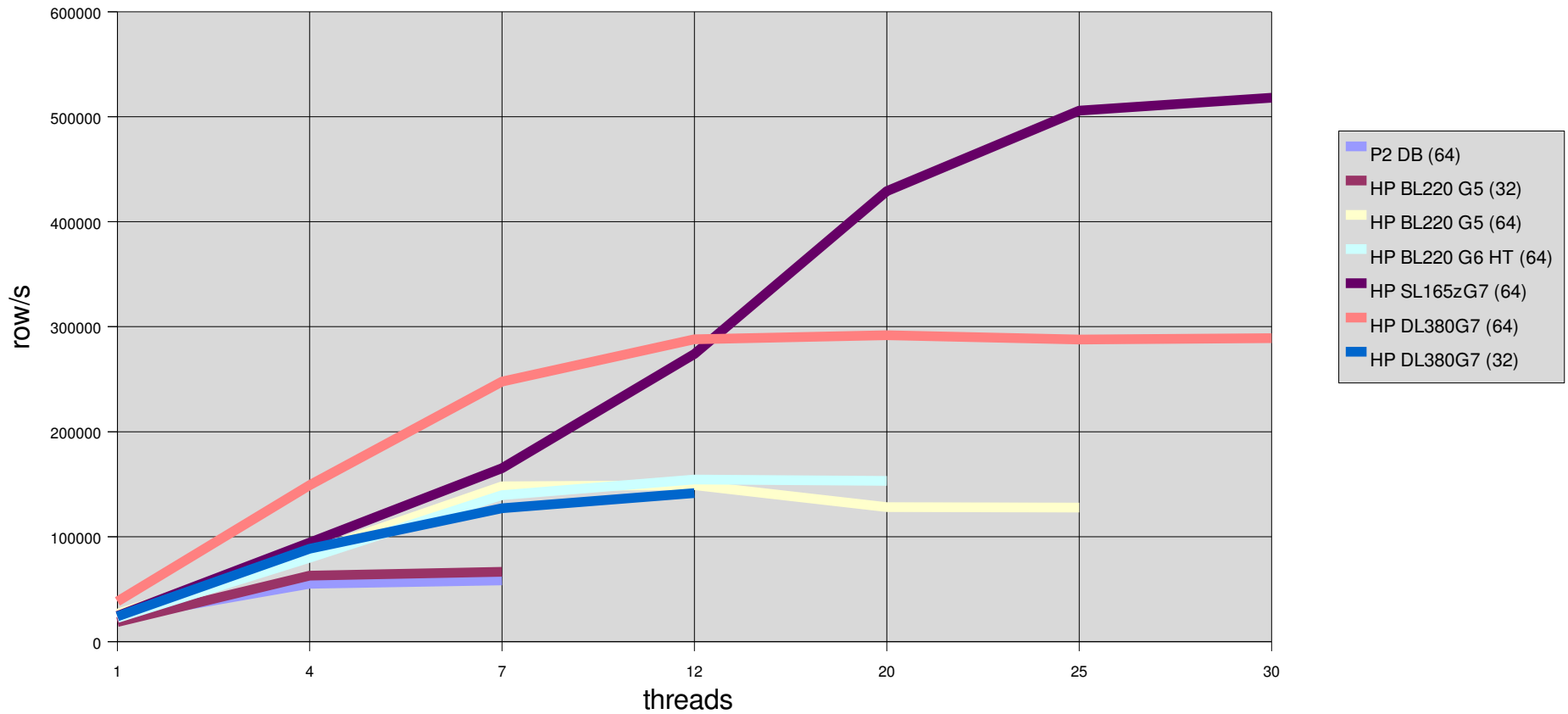


SQL file processing



Benchmark results – multi threaded applications

Number of rows inserted per second



Benchmark conclusions

(quote from report 05/2010)

*“As expected, machines must run with **64 bit OS** to get the best performance, as shown by the results of same hardware with different OS installed. Hyper-threading does not seem to improve nor reduce performance at higher thread count.*

*The **G7 Intel CPU** offers by far the best per-core performance.*

The high-density G7 AMD CPU performs decently compared to previous generation CPU of higher CPU frequency, but not as well as the Intel G7. However, it scales very well with the number of threads, and finally largely outperforms the Intel model when all cores are active.

One can expect a best-case improvement of a factor 6 to 10 with the latest CPUs compared to the hardware in production at the moment. However, these numbers can be reached only in optimal parallelism situations. A factor 2-4 looks more realistic. It will depend very much on the type of applications.

*Peak power / minimal response time would probably be achieved with the top-end **high-frequency** Intel models (not tested here) and to a lesser extent with the E5640, whereas overall performance in serving a large number of clients would fit best to the AMD 6174.”*

MySQL software

- Packaging
 - We take binary RPMs from the MySQL site
 - Use of the latest 'production' version, now 5.5
 - SLC repositories usually behind (SLC 4 mysql 3, SLC5 mysql 4, now 5.0)
- Support
 - Extensive documentation and knowledge base available
 - No problem so far
 - Reported 1 minor bug (change in packaging). Serious and rapid follow-up.
- Free version
 - 'community' version, no support / tools
- No (or little) tuning needed for performance
 - e.g. number of connections, maximum packet size, cache size, ...
- Lightweight maintenance (time spent close to zero)

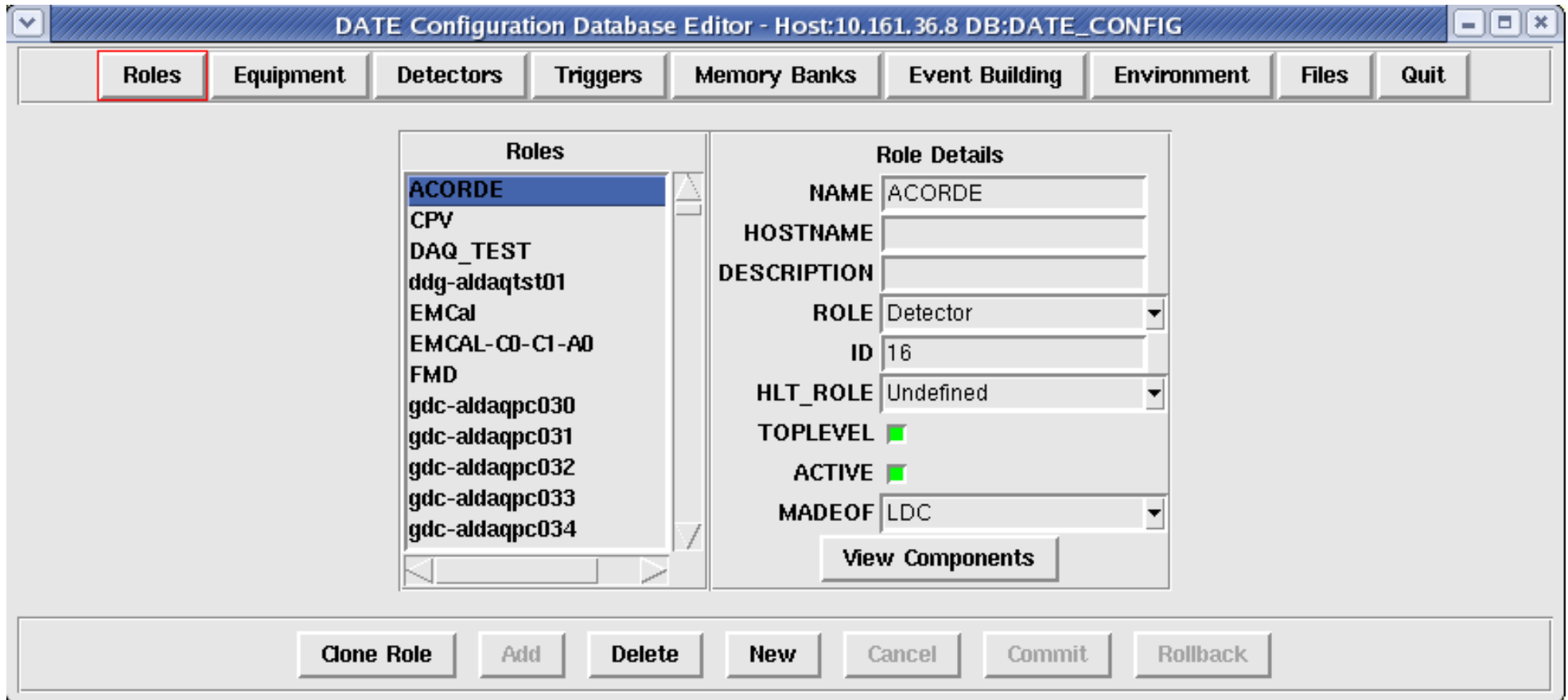
What about ...

- Security
 - Application specific credentials
 - No encryption (private network, no sensitive data)
- Data safety
 - Daily backups
 - RAID 0+1 and cold disk spare
 - Single case of severe crash last year after power cut / disk array file system corruption, where storage engine could not resume from log, fully recovered from replicated DB
- Availability
 - Several machines, services can be moved
 - Restored from backup or empty DBs

Data interfaces

- APIs
 - C for main applications
 - connect/disconnect, query, prepared statements
 - Tcl (in particular from UI / runControl)
 - Mysqлтcl (we package the SLC5 rpm)
 - Direct SQL through command line client
 - From shell scripts
 - For interactive (debug/expert) queries
- GUIs
 - Tcl / Tk
 - PHP / HTML
 - MySQL QueryBrowser / Administrator (now called 'workbench')

GUI snapshots



GUI snapshots

infoBrowser - DATE_SITE = /dateSite

Level
 Date
 Time
 decimals
 Host
 Role
 Pid
 Username
 System
 Facility
 Stream
 Run
 Message

Level	Time	Host	Facility	Run	Message
Info	16:12:56	aldaqpc110	readout	68885	Readout SOR fifoGetFree: (TOT: 2) - (OK: 2) - (BAD: 0)
Info	16:12:56	aldaqpc110	readout	68885	Readout SOR bmAllocate: (TOT: 2) - (OK: 2) - (BAD: 0)
Info	16:12:56	aldaqpc110	readout	68885	Readout EOR fifoGetFree: (TOT: 2) - (OK: 2) - (BAD: 0)
Info	16:12:56	aldaqpc110	readout	68885	Readout EOR bmAllocate: (TOT: 2) - (OK: 2) - (BAD: 0)
Info	16:12:56	aldaqpc110	ReadoutShell	68885	Readout exited, status: 0
Info	16:12:57	aldaqpc110	recorder	68885	recorder exited with status: 0
Info	16:12:59	aldaqpc143	evb	68885	Run 68885
Info	16:12:59	aldaqpc143	evb	68885	--- Event Builder summary on host:aldaqpc143 role:gdc-aldaqpc143 ---
Info	16:12:59	aldaqpc143	evb	68885	INFORMATION:
Info	16:12:59	aldaqpc143	evb	68885	Run start time:Wed May 6 16:04:19 2009, end time:Wed May 6 16:12:59 2009 Total:520 s (0 h 8 m 40 s)
Info	16:12:59	aldaqpc143	evb	68885	Event counters (IN/OUT): SOR:2/2 EOR:2/2 PHY:893636/893636 SOD:1/1 EOD:1/1 All events:893642/893642
Info	16:12:59	aldaqpc143	evb	68885	Individual breakdown by LDC:
Info	16:12:59	aldaqpc143	evb	68885	ldc-FMD-0=aldaqpc110 893642 Events 59.367 GB
Info	16:12:59	aldaqpc143	evb	68885	DAQ configuration: 1 LDC(s) 1 GDC(s) 1 recording stream(s) ldcPattern:70 gdcPattern:6 thisMachineId:6 maxErrors:10
Info	16:12:59	aldaqpc143	evb	68885	EVb runtime:dbMaxTriggerMaskId:49 dbMaxLdcId:82
Info	16:12:59	aldaqpc143	evb	68885	eventType:SOR all-events targetMask:70:(aldaqpc110) no-build applied on 2 event(s)
Info	16:12:59	aldaqpc143	evb	68885	eventType:EOR all-events targetMask:70:(aldaqpc110) no-build applied on 2 event(s)
Info	16:12:59	aldaqpc143	evb	68885	eventType:SOR_F all-events targetMask:70:(aldaqpc110) no-build
Info	16:12:59	aldaqpc143	evb	68885	eventType:EOR_F all-events targetMask:70:(aldaqpc110) no-build
Info	16:12:59	aldaqpc143	evb	68885	eventType:PHY all-events targetMask:70:(aldaqpc110) build applied on 893636 event(s)
Info	16:12:59	aldaqpc143	evb	68885	eventType:CAL all-events targetMask:70:(aldaqpc110) build
Info	16:12:59	aldaqpc143	evb	68885	eventType:SOD all-events targetMask:70:(aldaqpc110) no-build applied on 1 event(s)
Info	16:12:59	aldaqpc143	evb	68885	eventType:EOD all-events targetMask:70:(aldaqpc110) no-build applied on 1 event(s)
Info	16:12:59	aldaqpc143	evb	68885	eventType:FORMAT_ERROR all-events targetMask:70:(aldaqpc110) no-build
Info	16:12:59	aldaqpc143	evb	68885	eventType:SST all-events targetMask:70:(aldaqpc110) build
Info	16:12:59	aldaqpc143	evb	68885	eventType:DST all-events targetMask:70:(aldaqpc110) build
Info	16:12:59	aldaqpc143	evb	68885	1 input channel(s) active
Info	16:12:59	aldaqpc143	evb	68885	EDM host not present
Info	16:12:59	aldaqpc143	evb	68885	Recorded:893642 event(s) total:59.428 GB at 114.285 MB/s 1.718 KEV/s average recordingDevice:'/dev/hull' numStreams:1 maxFileSize:0
Info	16:12:59	aldaqpc143	evb	68885	893642 event(s) required 893642 write(s) for an average of 1.000000 write(s)/event
Info	16:12:59	aldaqpc143	evb	68885	Memory system NUM_OF_LDCS:1 totalSize:1.903 GB privateSize:6.000 MB pipelineDepth:3 publicSize:1.897 GB maxEventSize:4.000 MB
Info	16:12:59	aldaqpc143	evb	68885	# Fulls/Allocations Public pool:0/0 70-aldaqpc110:0/893642
Info	16:12:59	aldaqpc143	evb	68885	No events injected for monitoring
Info	16:12:59	aldaqpc143	evb	68885	*****
Info	16:12:59	aldaqpc143	evb	68885	Event builder daemon terminating
Info	16:13:00	aldaqpc113	LAUNCHER	68885	/dateSite/configurationFiles/FMDda_BASE.sh ^FMD succesfully completed
Info	16:13:01	aldaqcs01	runControl	68885	Run stopped
Info	16:13:01	aldaqcr37	runControlHI	68885	Stop processes time : 6 seconds
Info	16:13:01	aldaqcs01	DCA	68885	End of STANDALONE_RUN

Archive Filters

min. match
 max. exclude

Status : Idle
 Query : SELECT * from messages WHERE timestamp>1241615158 ORDER BY timestamp
 452 messages, 4 errors

Online
 Auto Clean

GUI snapshots

1-20 of 134 (Page 1 of 7) Local filters Partition: PHYSICS Start Time: [10/09/2008 00:00:00..15/09/2008 00:00:00]

Quick Access Export... Fields...

Statistics **Detectors** Trigger Clusters Overview

Run	Start Time	Duration	# of LDCs	# of GDCs	# of Detectors	Partition	Run Type	Total SubEvents	SubEvent Rate	Total Events	Event Rate	HLT Mode	EOR Reason	Shuttle	Data Migrated	Total Data Recording (MB)	Data Rate Recording (MB/s)
58394	12/09/2008 05:35:48	5 h	13	1	4	PHYSICS	PHYSICS	59	< 0.01	135	< 0.01	A	Operator_Request	✓	Yes	42	< 0.01
58071	10/09/2008 19:17:08	4 h	4	1	1	PHYSICS	PHYSICS	2 738	0.21	2 760	0.21	A	Operator_Request	✓	Yes	33	< 0.01
58378	12/09/2008 02:35:49	3 h	13	1	4	PHYSICS	PHYSICS	39	< 0.01	115	0.01	A	Operator_Request	✓	Yes	27	< 0.01
58020	10/09/2008 13:26:34	3 h	4	1	1	PHYSICS	PHYSICS	77	< 0.01	96	< 0.01	A	Operator_Request	✓	Yes	7	< 0.01
58559	13/09/2008 06:23:28	2 h	13	1	4	PHYSICS	PHYSICS	54	< 0.01	130	0.02	A	Operator_Request	✓	Yes	37	< 0.01
58793	14/09/2008 09:17:12	2 h	58	9	6	PHYSICS	PHYSICS	794 785	96.13	795 107	96.17	A	Operator_Request	✓	Yes	318 586	38.53
58327	11/09/2008 20:22:48	2 h	9	1	3	PHYSICS	PHYSICS	55	< 0.01	107	0.01	A	Operator_Request	✓	Yes	10	< 0.01
58343	12/09/2008 00:04:43	2 h	13	1	4	PHYSICS	PHYSICS	261	0.04	337	0.05	A	Operator_Request	✓	Yes	179	0.03
58512	12/09/2008 22:23:47	2 h	6	1	2	PHYSICS	PHYSICS	49	< 0.01	83	0.01	A	Operator_Request	✓	Yes	48	< 0.01
58780	14/09/2008 07:00:19	2 h	57	9	6	PHYSICS	PHYSICS	583 414	95.86	583 727	95.91	A	DAQ_Request	✓	Yes	180 124	29.60
58430	12/09/2008 12:03:03	2 h	13	1	4	PHYSICS	PHYSICS	477	0.09	553	0.10	A	Operator_Request	✓	Yes	327	0.06
58558	13/09/2008 04:29:08	1 h	13	1	4	PHYSICS	PHYSICS	17	< 0.01	93	0.02	A	Operator_Request	✓	Yes	12	< 0.01
58643	13/09/2008 12:29:07	1 h	6	6	3	PHYSICS	PHYSICS	12 135 040	2 903.12	12 135 075	2 903.13	A	Operator_Request	✓	No data		
57954	10/09/2008 01:14:12	1 h	53	5	8	PHYSICS	PHYSICS	155 565	39.04	155 881	39.12	A	Operator_Request	✓	No data		
58501	12/09/2008 21:18:03	1 h	6	1	2	PHYSICS	PHYSICS	473	0.13	507	0.14	A	Subsystem_failure:HLT	✓	Yes	279	0.08
58064	10/09/2008 18:14:59	1 h	4	1	1	PHYSICS	PHYSICS	131	0.04	150	0.04	A	Operator_Request	✓	Yes	4	< 0.01
58479	12/09/2008 13:42:56	60 m	13	1	4	PHYSICS	PHYSICS	27	< 0.01	103	0.03	A	Operator_Request	✓	Yes	18	< 0.01
58056	10/09/2008 16:58:13	53 m	4	1	1	PHYSICS	PHYSICS	19	< 0.01	38	0.01	A	Operator_Request	✓	Yes	1	< 0.01
58338	11/09/2008 23:03:33	53 m	13	1	4	PHYSICS	PHYSICS	249	0.08	325	0.10	A	Operator_Request	✓	Yes	171	0.05
58693	13/09/2008 16:25:40	42 m	46	7	6	PHYSICS	PHYSICS	523 283	207.41	523 512	207.50	A	Operator_Request	✓	Yes	103 572	41.05

List of detectors for partition 'PHYSICS'

FMD	SSD	473
HMPID	TO	131
SDD	TPC	27
SPD	ZDC	

Physics Trigger Clusters Info

FMD	PHYSICS	249
HMPID	PHYSICS	523 283
SDD	triggered by SPD	
TPC		
TRIGGER		
ZDC		

Some features we use

- Constraints & foreign keys
 - Configuration data integrity
- Transactions
 - Mainly from interactive clients (lock & rollback)
 - In APIs: updates of shared counters, multiple steps operations
- Partitioning
 - Automatic split of tables on a variable (e.g. log timestamp)
- Indexing
 - Needed for fast response time on queries, seen little effect on insert (but quite heavy on size)
- Triggers, Events (c.f. cron), stored procedures
 - e.g. to update global counters or lists, or for shared logic between different APIs
- Storage engine types
 - InnoDB (constraints, transactions) MySQL (raw performance) RAM (fast transient data)

Some features we use

- Replication (1 master, several slaves)
 - Easy to configure
 - Good as backup replacement or hot spare setup
 - Remove query load from main server
 - NB: Cluster feature ('shared-nothing' redundancy) of MySQL seems nice, but not tried / needed for our system
- Backup
 - Crontab: dump database to SQL file and archive (RAID6 + tape)
 - Easy to reload, may take time (indexes)
 - Careful definition of mysqldump options
- Monitoring
 - Easy access to key server metrics (inserts, connects, etc)
- Extended server log
 - Punctual enabling of full query logs and analysis allowed to spot several client implementation issues

Conclusion and perspectives

- We are happy MySQL users
 - Performance and features right “out of the box”
 - Fits our (simple?) needs for a large range of data patterns
 - Heterogeneous DBs and usages demand careful planning and testing
 - After more than 3 years in production, excellent feedback on stability, reliability, performance
- No big change expected on requirements / needs
 - More DAQ components will use databases
 - Close look on existing DB growth over time
 - What about a SQL DB with data subscribe / notification interface ?