CMS Experience with Online and Offline Databases

Dr. Andreas Pfeiffer, CERN

for the CMS experiment

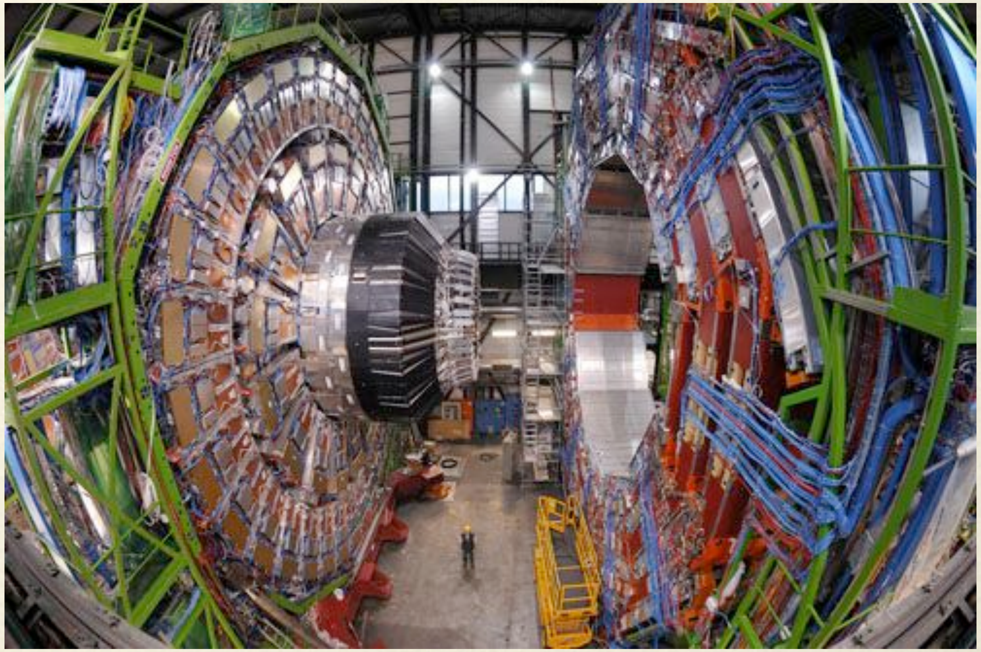
CHEP 2012, New York (NY), USA

Outline

- Overview
- The Challenge
- Conditions data: what and how
- DB Evolution and Performance
- Monitoring
- Outlook
- Summary

CMS Experiment @ CERN-LHC

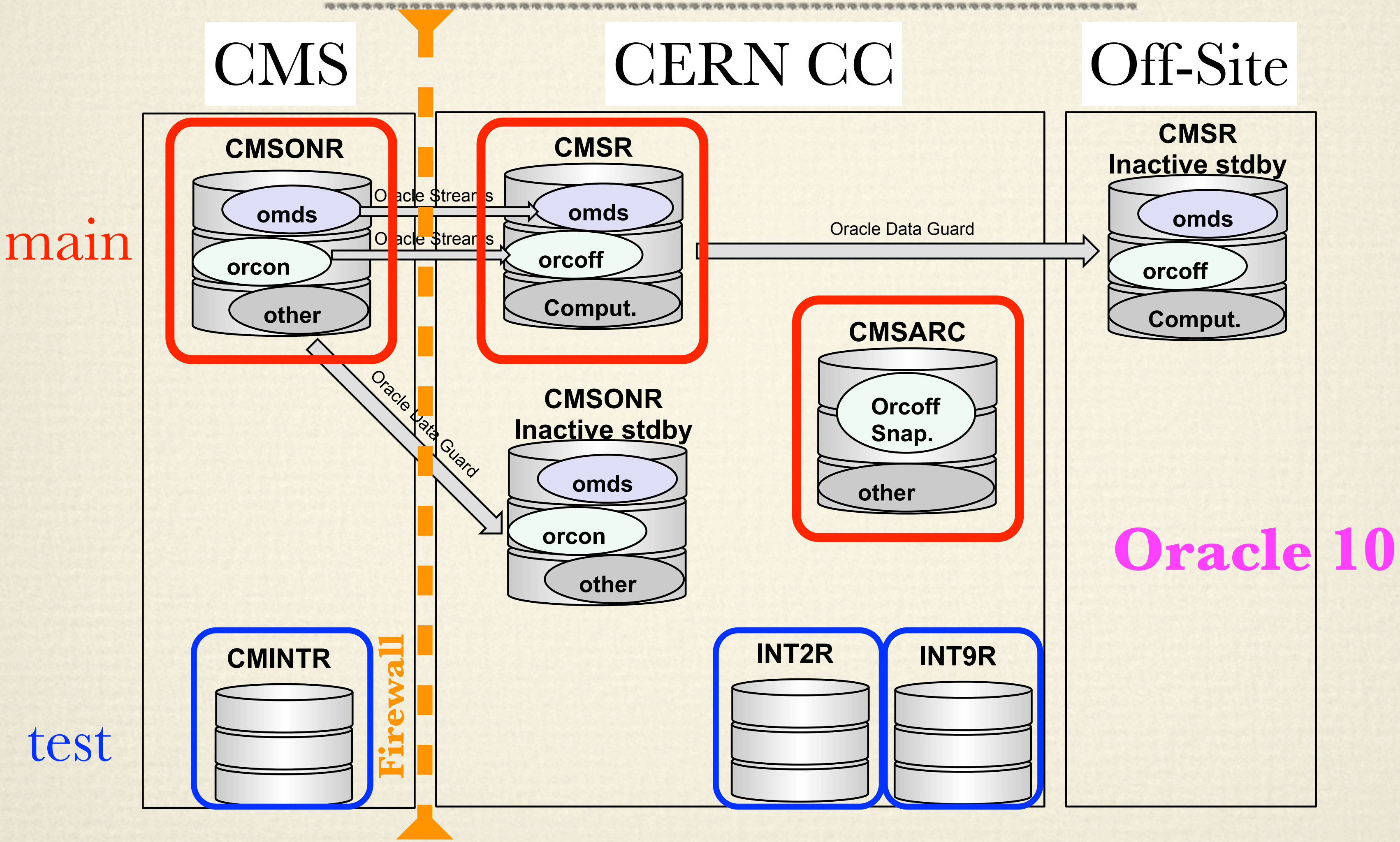
- The Compact Muon Solenoid (**CMS**) experiment at the Large Hadron Collider (**LHC**) at CERN (Geneva, Switzerland)
- 12500 t, 15 m dia., 22 m length, B 3.8T
- Around 4300 active members
  - 179 institutes
  - 41 countries



Use of DBs in CMS

- Configuration information
  - detectors, DAQ, L1 trigger, High Level Trigger (HLT)
- Run, Beam and Luminosity information
  - info on which files are written sent to Tier-0, eLog, ...
- Offline DB also hosting computing applications
  - Tier-0 workflow processing, Data distribution service (PhEDEx), Data Bookkeeping Service, ...
- **Conditions data** for offline reconstruction and analysis
  - critical data, exposed to a large community

CMS Databases until end 2011



Overview - The Challenge

- Over 75 million channels in various detectors
- Detector information for each channel
  - Conditions: Temperature, HV, LV, status, ...
  - Calibration: pedestals, charge/count, ...
  - Changes with time (temperature and radiation)
- Necessary for performance monitoring
  - by detector experts
- Subset used by offline reconstruction and physics analysis
  - Conditions data
  - need to distribute to at all Tier-N centres worldwide

Conditions Data - What

- Conditions data
  - subset of the calibration information for each of the >75 millions channels of the detector
  - plus information on calibration and alignment from offline processing
  - plus information from dedicated “express” processing
    - e.g. beam-spot fed back to online and used in HLT
- Critical for physics data reconstruction and analysis
  - data is exposed to a large community worldwide

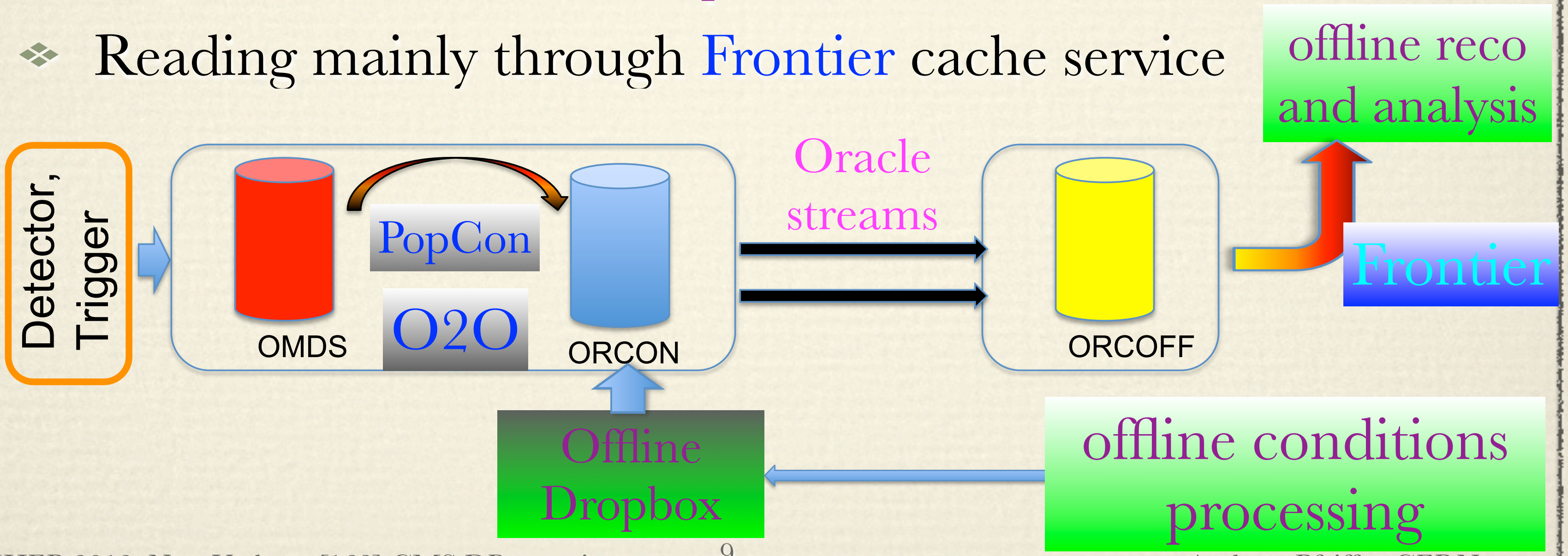
Conditions Data - How

- Conditions have Intervals Of Validity (IOV) plus a "payload" (the actual data) for each IOV
  - A specific IOV is identified/categorized by a "tag"
  - A consistent set of tags is a "Global Tag"
    - used for any kind of (re-)processing
- Consistent and transparent access to conditions via common software using object-relational mapping
  - focus on data integrity (e.g. never delete IOVs)
- Needs worldwide distribution to Tier-Ns
  - Frontier squid service

more info:
[351] G.Govi
Tue 17:50, here

Conditions handling and usage

- Online conditions are sent to offline DB via "Online-to-Offline" (O2O) jobs using the PopCon application
  - usually one job per detector, maintained by detector experts
- Offline conditions (e.g. beam-spot, alignment, ...) handled via "Offline Dropbox" *(see also: Poster [202], Talk [351])*
- Reading mainly through Frontier cache service



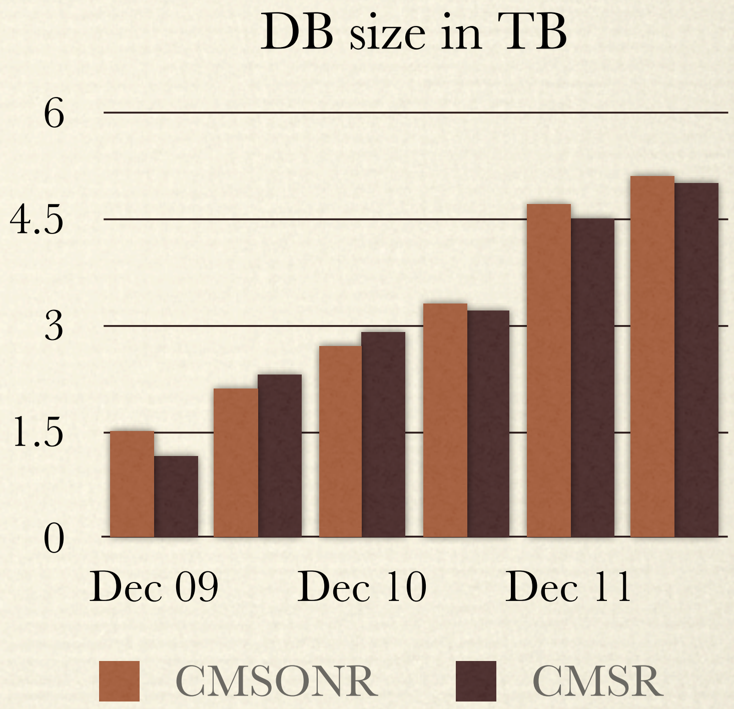
DB Clients - Frontier

- Offline reconstruction jobs on Tier-0/1 could create a large load on the Offline DB
  - tens of thousands of jobs, few hundred queries each
- Frontier squid caches minimize the direct access to Oracle servers
  - additional latency as set by the cache refresh policy
  - Frontier service for Online
    - used to distribute configuration and conditions to HLT
  - Frontier service for Offline (Tier-N)
    - reading from “Snapshot” from Offline DB
    - heavily used for reprocessing

more info:
[220] D.Dykstra
Poster, Thu

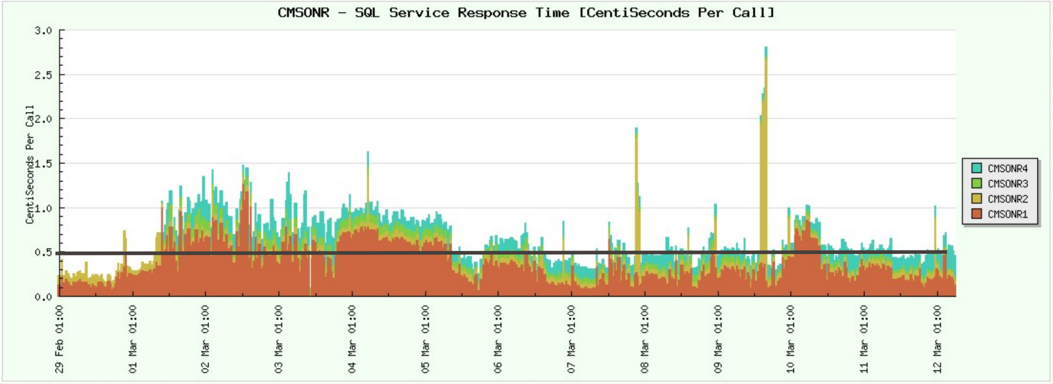
DB Space usage and Evolution

- DB growth about 1.5 TB/yr
  - both online and offline
- Condition data is only a small fraction
  - ~ 300 GB at present
  - growth: + 20 GB/yr
  - about 50 Global Tags created each month



Operations in 2011

- Very smooth running
  - CMSONR availability: **99.88 %**
    - 10.5 hours downtime overall in 2011
  - CMSR availability: **99.64 %**
    - 30.7 hours downtime overall in 2011
  - SQL query time stable (few msec)



10 ms

downtime includes all power-cuts, node reboots, hangs, (some) maintenance, …

Big Thanks to CERN DBAs !!

Essential service: Monitoring

more info:
[202] S.DiGuida
Poster, Thu

- Monitoring of services implemented for
  - Hardware and infrastructure
    - disk I/O (incl. growth), CPU, network, streams, ...
- Top level views for PopCon and Dropbox provide info for stakeholders
  - Condition DB experts: control of workflows
  - Detector experts: check status of submitted requests
- Error reporting and logs
  - active notifications of problems to experts via Nagios

Monitoring CMS DB services

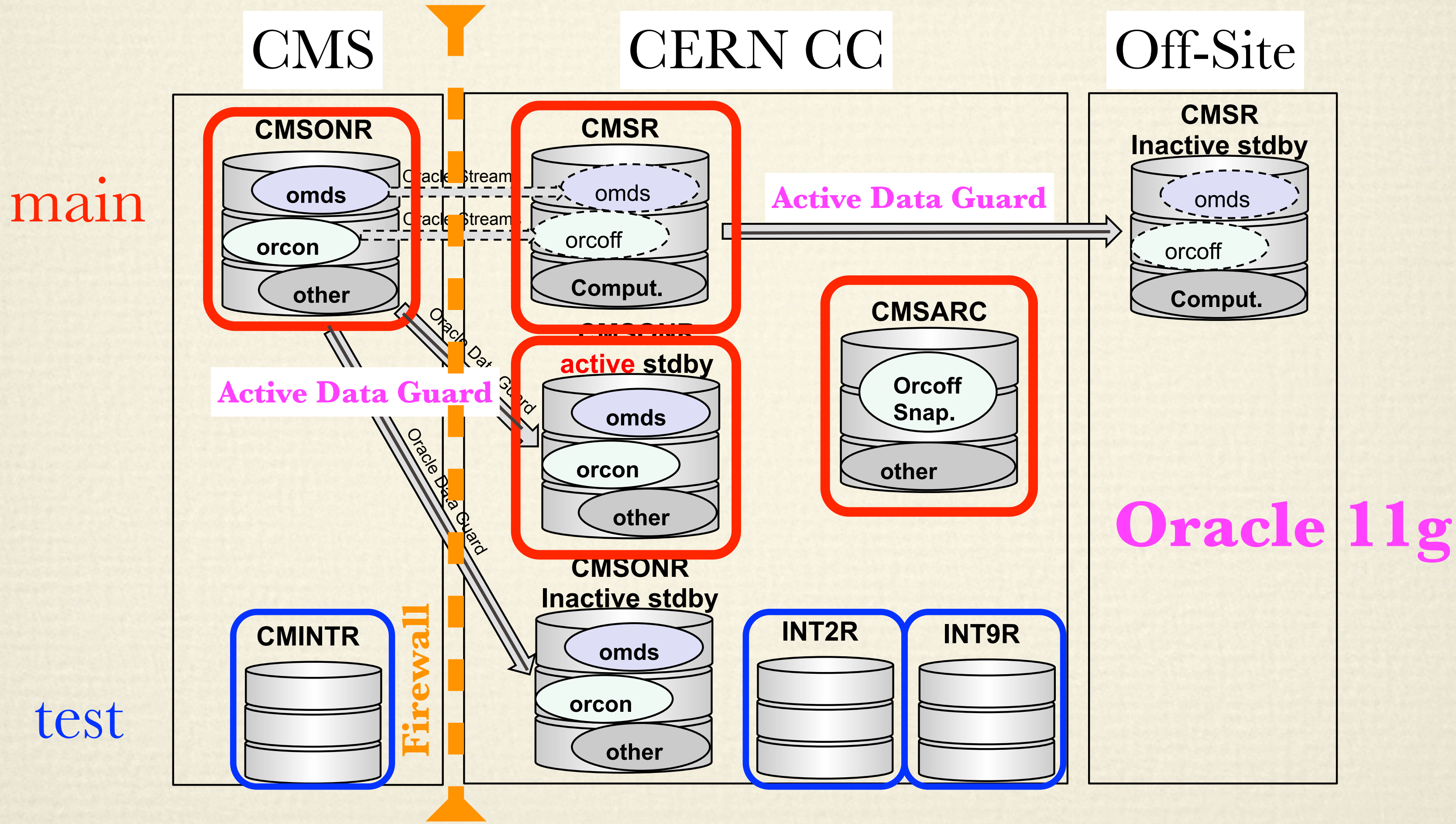
- Nagios service
  - monitoring of services and alarming of experts
- EasyMon - overview
  - http://cms-conddb.cern.ch/easymon
  - uses info from Nagios service
- Central monitoring page
  - http://cms-conddb.cern.ch/
  - Links to individual monitoring pages
    - IT page of DB status
    - Frontier monitoring (online and offline)
    - PopCon Monitoring

Outlook

- In early 2012 moved to new h/w and Oracle 11g
  - profit from new technologies (ADG)
  - improved overall redundancy, failover tests successful
- Collecting experience
  - overall positive so far (yes, there are hiccups :-))
- Clearly will continue to have an eye on performance
  - New (and updated) applications are required to be tested in INT DB before deployment in production DB
    - DBAs help to check and optimize performance
- May want to evaluate the use of "NoSQL DB"
  - "key/value" seems to map perfectly to conditions :-)

NoSQL talks/ posters: [184], [218], [352], [359]

Upgrade of DBs in early 2012



Summary

- CMS Databases are essential for operating the experiment
  - Online and Offline
- Performance overall very satisfactory
  - overall >99.5% availability in 2011
  - growth rates of ~ 50% in 2011
- New h/w and Oracle version deployed early 2012
  - positive experience so far ...
- Conditions are essential for offline reconstruction and physics analysis
  - distributed using Frontier cache service
- Good monitoring of the services is essential

Additional Info

CMS Online DB overview

- A total of 678 Schemata
  - 36 system
  - 232 for conditions (CMS_COND_...)
  - 131 for PVSS
  - 232 for “detectors”
  - 80 other

What in P5 depends on the DB ?

- detector configuration, settings ("slow control")
- trigger configurations (L1, HLT)
  - distribution for HLT via Frontier (online)
- run control, eLog, shift-list
- access control for doors
  - reads from CMS DB who is authorized to go in
  - people who are in can, of course, still go out
  - access to key to refill coffee-machine
    - access only to shift leader, shift-list read from DB
- in short: almost everything