DAQExpert

An expert service to increase CMS data-taking efficiency

Maciej Gladki
maciej.gladki@cern.ch

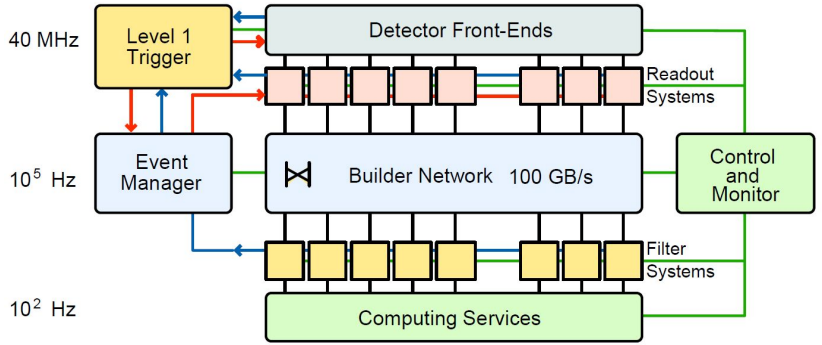
ON BEHALF OF THE CMS DAQ GROUP

CHEP 2019 Nov 4-8
24th International Conference on Computing in High Energy & Nuclear Physics

Track: 1. Online and Real-time Computing

CMS DAQ - Data Acquisition

- read out the data
- bunch crossing 40 MHz rate
- event size 1-2MB
- 2-level triggering
- hardware trigger selects 100 kHz
- full events built at 200GB/s
- 35 000 cores in HLT farm select O(1 kHz)

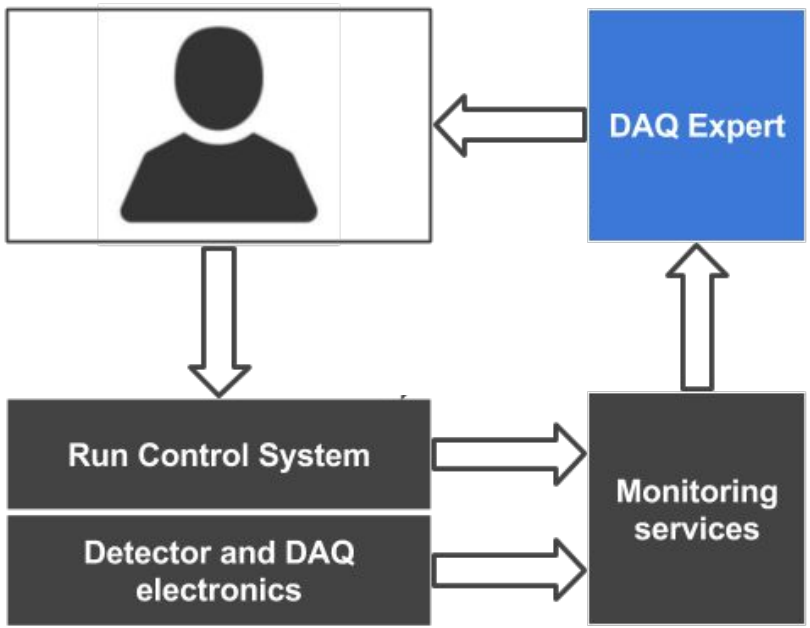


DAQ operations

- System issues are expected
  - Controlling 100s crates of electronics, detectors
  - hardware/software/network problems
  - datataking is stuck
  - recovery procedures
  - operators in control room 24/7
  - on-call experts 24/7
- Human factor
  - operators will make mistakes under time pressure
  - operators will add latency
  - on-calls don’t like to be woken up in the middle of the night
- We need a tool to automate it

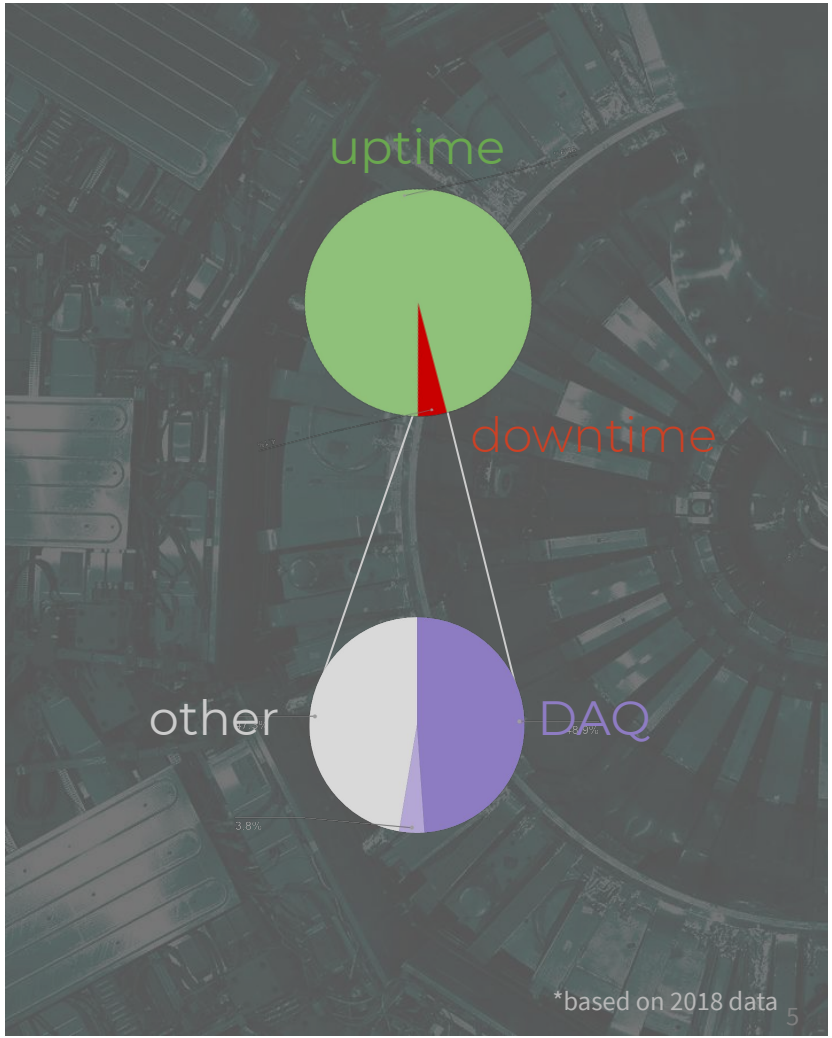
DAQExpert service

- Enables system experts to define potential dataflow problems and recovery procedures
- Identifies the problem from monitoring data
- Provides shifters the guidance
- The goal is to improve data taking efficiency



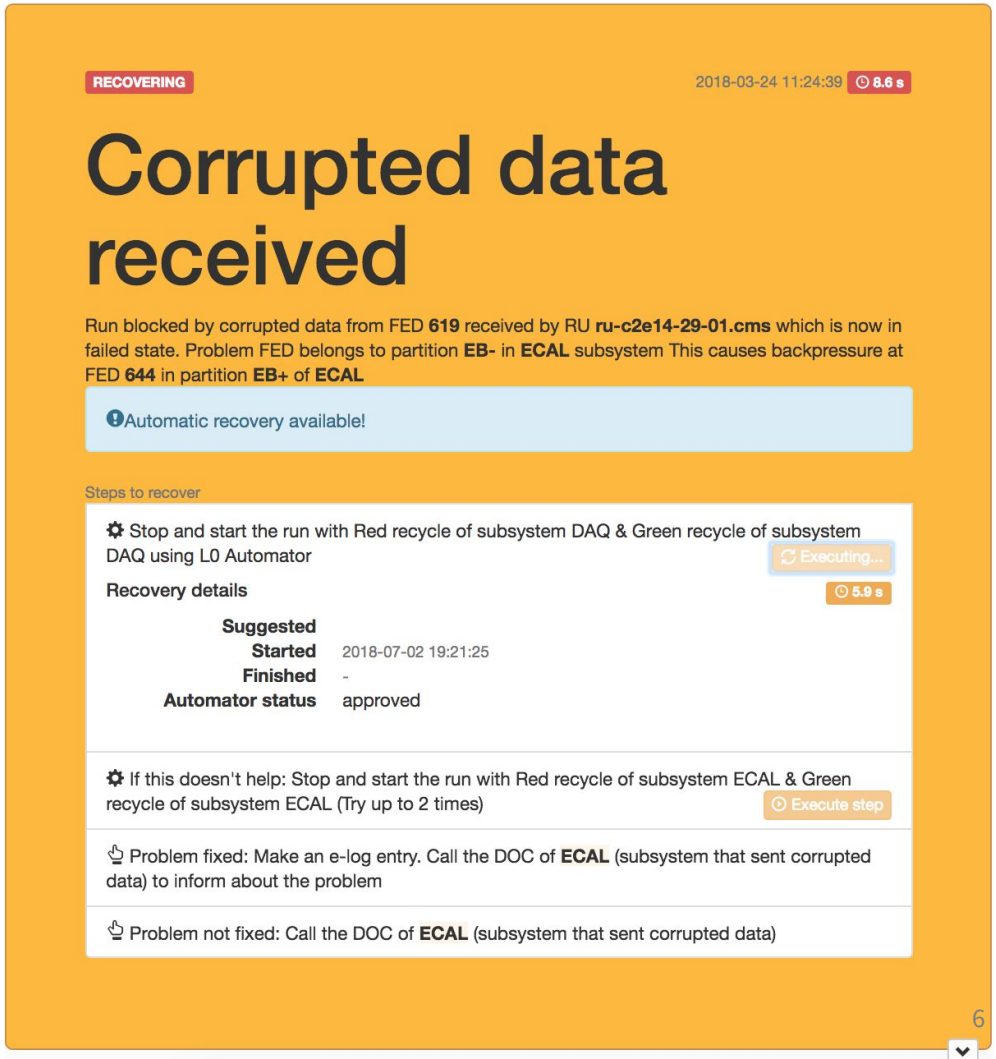
Datataking efficiency

- CMS: **95.87% uptime**, **4.13% downtime**
- Counts during Stable Beams
- **2184 hours** of Stable Beams delivered in 2018 (25% of the year)
- Total downtime was **90h**
  - power supply, infrastructure, LHC, DAQ...
- Downtime attributed to **DAQ: 46h**
  - Sub detectors 93%, central DAQ 7%
  - Scope of DAQExpert (rest is outside of its influence)



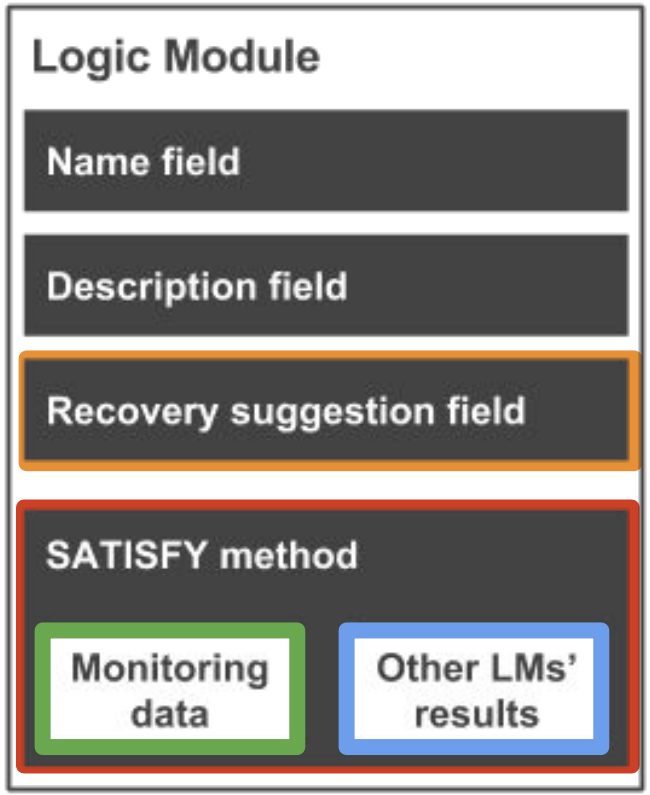
Dashboard

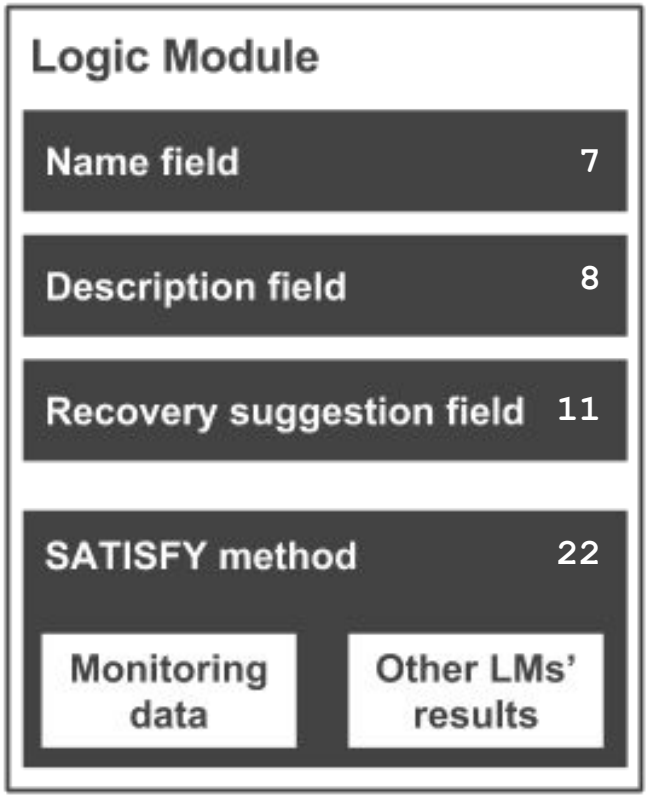
- Main DAQExpert view for control room
- Suggestions to shifters (with sound alarm)
  - Reduce reaction time
  - Avoid wrong decisions
- Suggestion format
  - Description of the problem
  - What's best action to take



Reasoning

- Expert knowledge encapsulated in logic modules (LM)
- Each LM defines
  - **name and description of the problem**
  - **recovery procedure**
  - **dataflow problem condition**
- Input data for condition
  - **monitoring data**
  - **output of other LMs**
- Modularity and imperative language



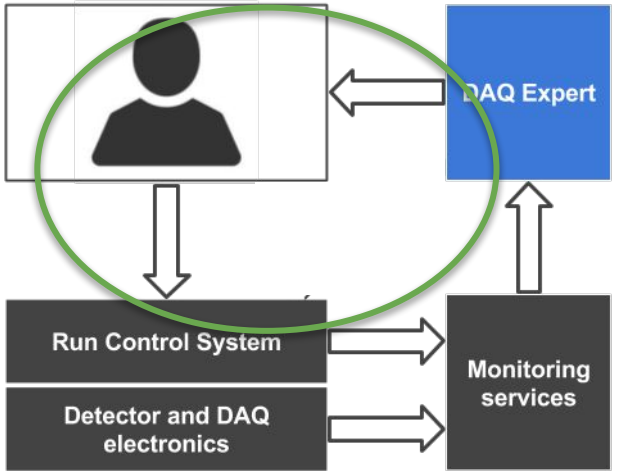



```java
/**
 * Logic module identifying corrupted data received
 */
public class CorruptedData extends KnownFailure {

 public CorruptedData() {
  this.name = "Corrupted data received";
  this.description = "Run blocked by corrupted data from " +
   "FED(s) {{PROBLEM-SUBSYSTEM}}/{{PROBLEM-PARTITION}}/{{PROBLEM-FED}}";

  this.recovery = new RecoveryProcedure(
   "<<StopAndStartTheRun>> with <<RedAndGreenRecycle::DAQ>>",

   "If this doesn't help: <<StopAndStartTheRun>> with both " +
   "<<RedAndGreenRecycle::DAQ>> and " +
   "<<RedAndGreenRecycle::{{PROBLEM-SUBSYSTEM}}>>";

   //.. other recovery steps
  }

  @Override
  public boolean satisfied(DAQ daq, Map < String, Output > results) {

   for (SubSystem subSystem: daq.getSubSystems()) {
    //... find a RU - Read out Unit that is failed
    if ("Failed".equalsIgnoreCase(ru.getStateName())) {
     //.. find a FED that has received corrupted data
     if (fed.getRuFedDataCorruption() > 0F) {
      contextHandler.register("PROBLEM-FED", fed);
      contextHandler.register("PROBLEM-SUBSYSTEM", subSystem);
      //.. register other context information
      result = true;
     }
    }
   }
   return result;
  }
}
```

*parts of the code have been removed in order to increase readability

Impact 1

- Guidance to operators → reduces intervention time
- Reaction time is significant part
- Main metric to measure impact of DAQExpert



Problem occurs
Data-flow stuck

Operator reacts
Problem identified, recovery issued

Recovery completes
Either successful or not

Reaction time

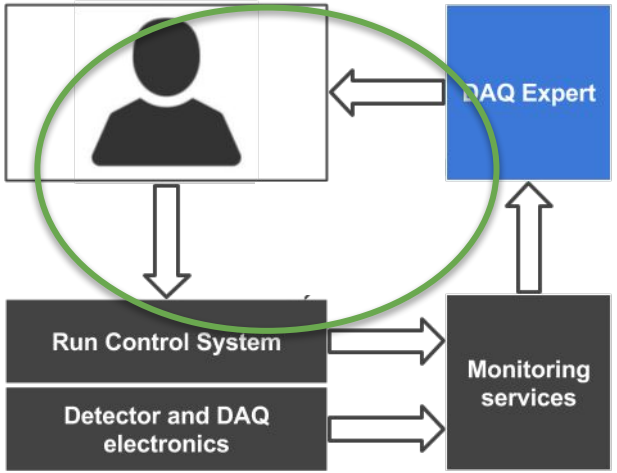
Recovery time

Intervention time

Impact 2

- Gradually introduced during Run2
  - DAQExpert introduced in 2017
  - Improvements in 2017 and 2018 (coverage and UX)
- Reduces reaction time
  - Mean reaction time: 101s (2016) → 65s (2017) → 47s (2018)

percentile	Reaction time			Reduction 2016 to 2018
	2016	**2017**	**2018**	
95th	322s	177s	132s	**59%**
75th	100s	78s	41s	**59%**
50th	85s	49s	29s	**65%**
25th	46s	23s	21s	**54%**

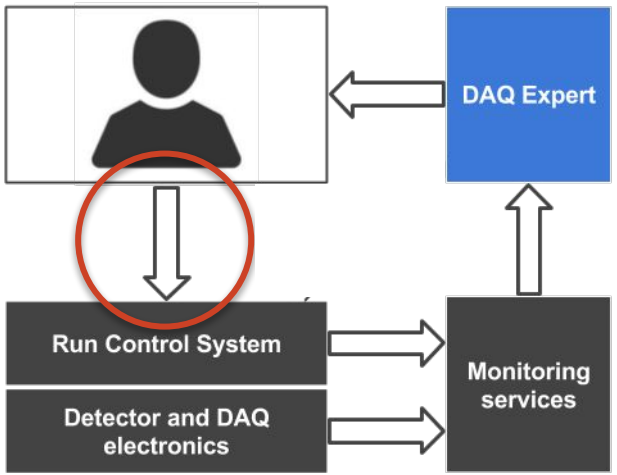


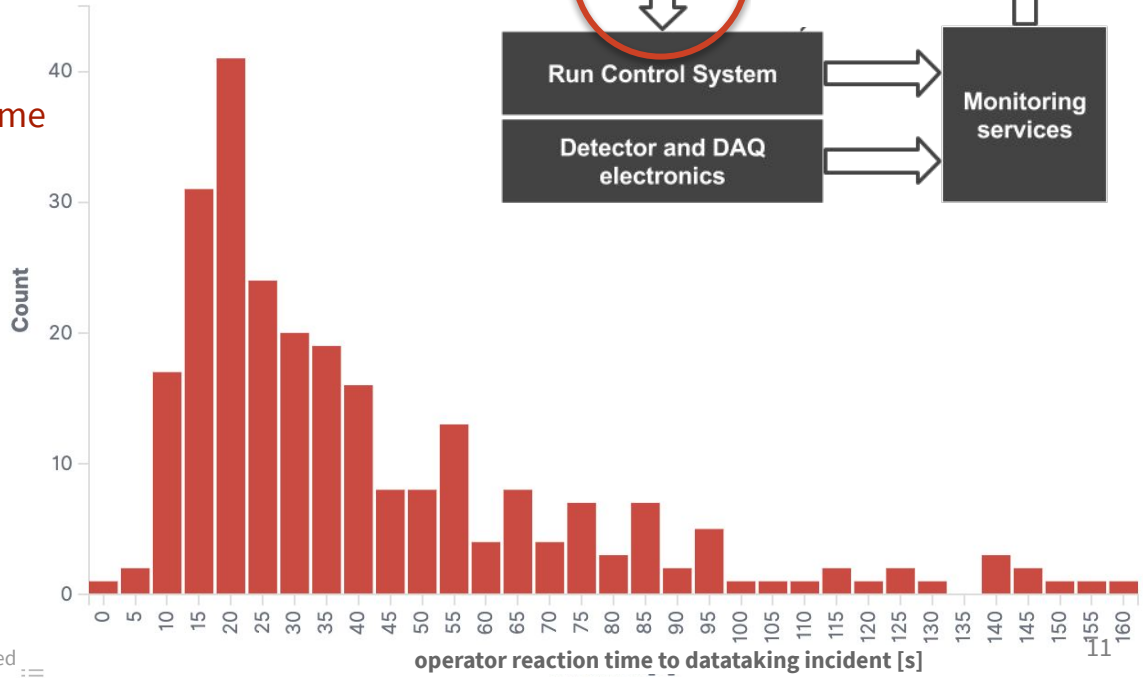
Based on:

- 297 interventions
- data from 2016-2018
- only failures in smooth datataking (at least for 5 minutes) = operators not in alarmed state

Limitations

- Operator has a final decision
- Reaction time latency
  - Reaction time of **20-25sec**
  - Significant spread
  - Accumulates to ~4-6h of downtime per year*
- Wrong decision overhead
  - ~1h of downtime per year**
- Improper usage of tools
  - ~1h of downtime per year**
- The most impactful way to improve: bypass the operator

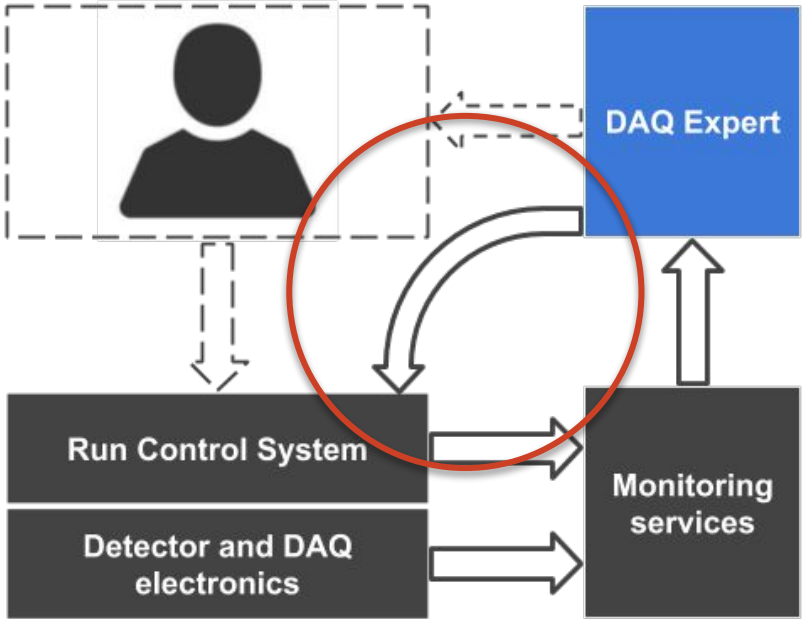




*based on 2017 and 2018 data (266 interventions, at least 5 mins of smooth data taking = operators not in alarmed state)
**based on 2018 August data - detailed case by case analysis had to be performed

Automatic recovery

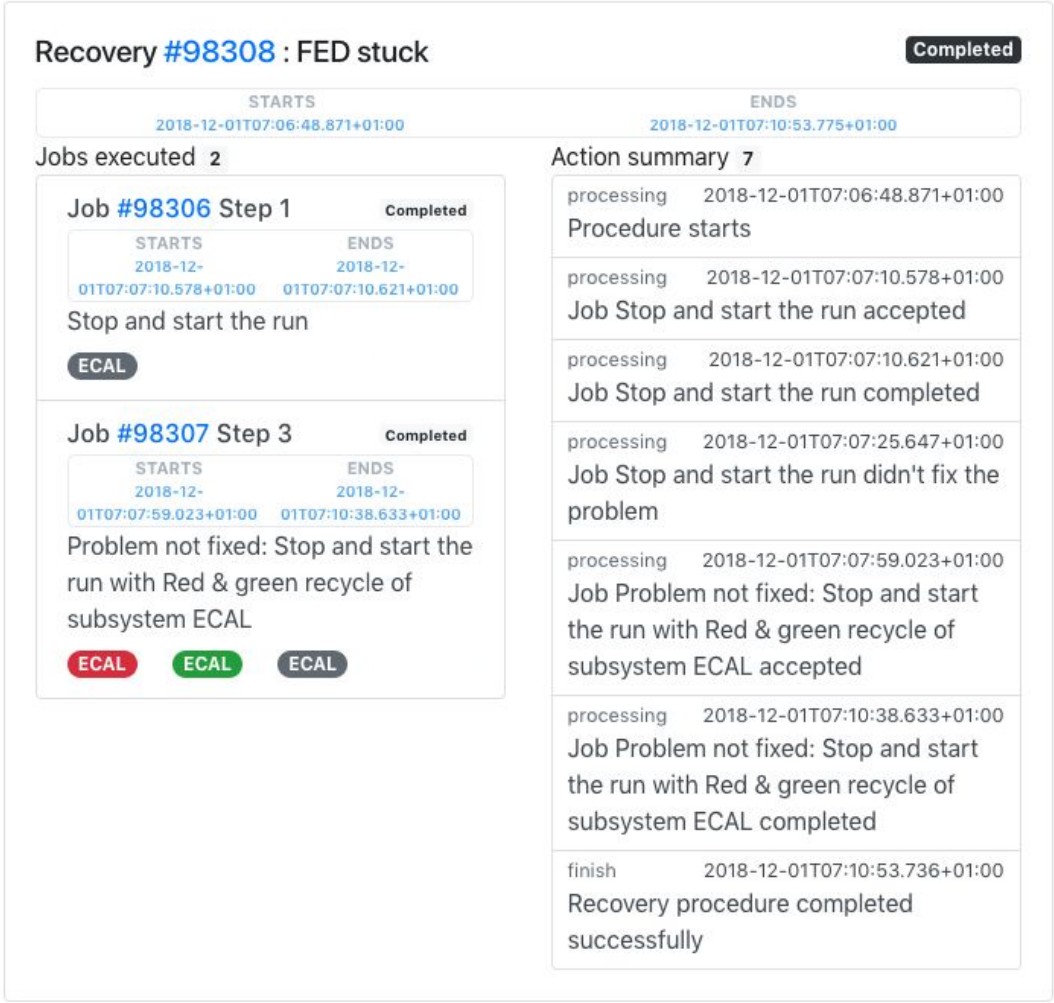
- Recovery driven by DAQExpert service
- No operator involved
- Commissioned in the end of run 2
- First successful automated jobs observed
- Estimated to reduce the downtime by **8h/year ***:
  - Reaction time
  - Wrong decision overhead
  - Improper usage of tools
  - **17%** of DAQ downtime
  - **9%** of total CMS downtime



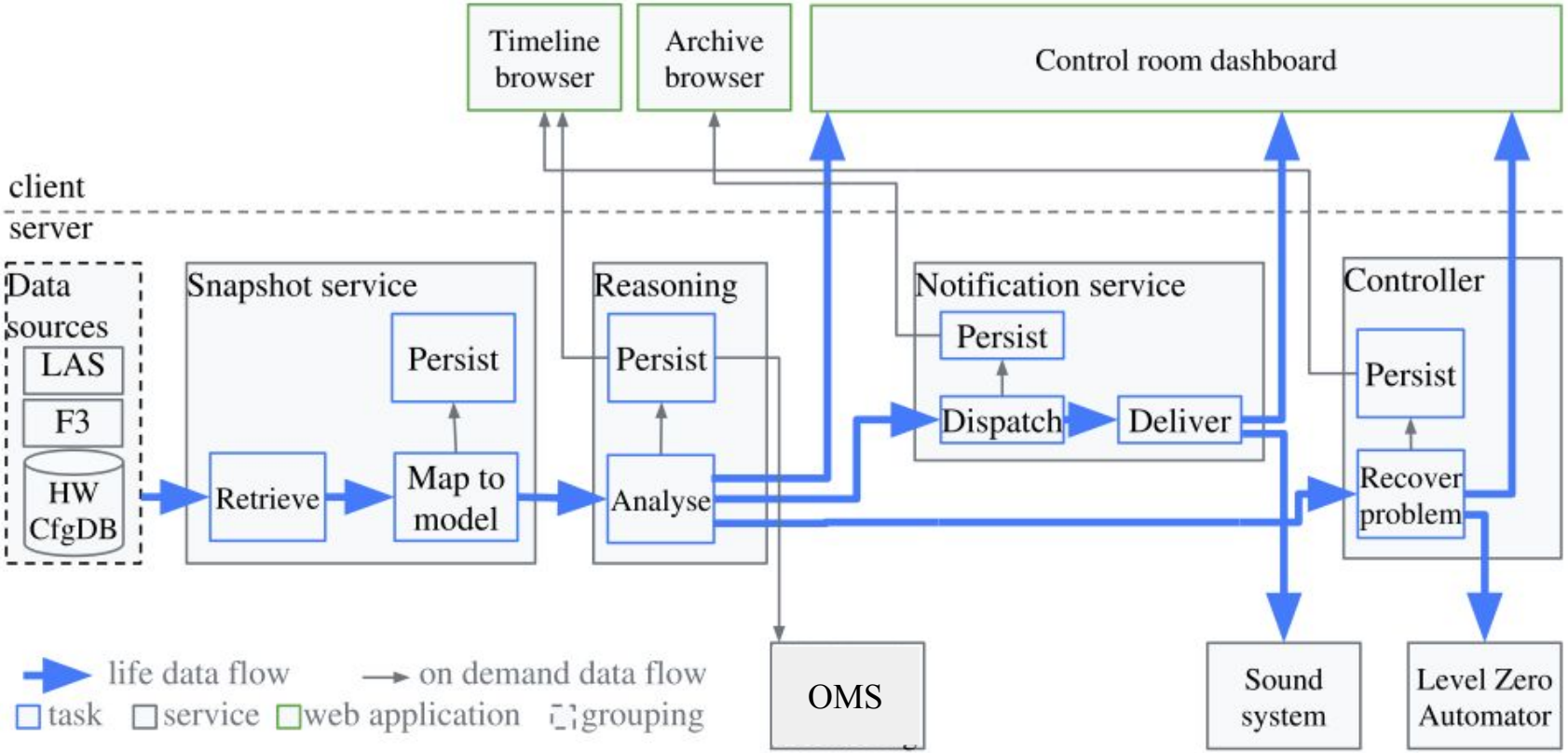
*based on 2017 and 2018 data (266 interventions, at least 5 mins of smooth data taking = operators not in alarmed state)
**based on 2018 August data - detailed case by case analysis had to be performed

Automatic recovery

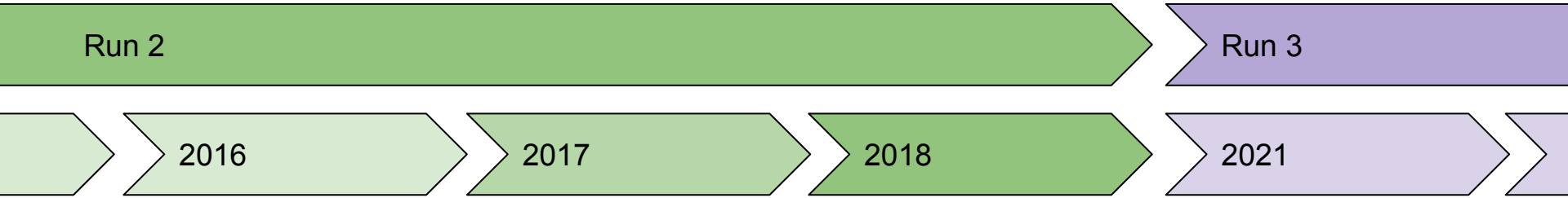
- Recovery steps picked from LM recovery procedure
- Avoid human reaction time and human error
- Only the monitoring delay



Architecture



Timeline of DAQExpert



- DAQ Expert developed
- first release Aug

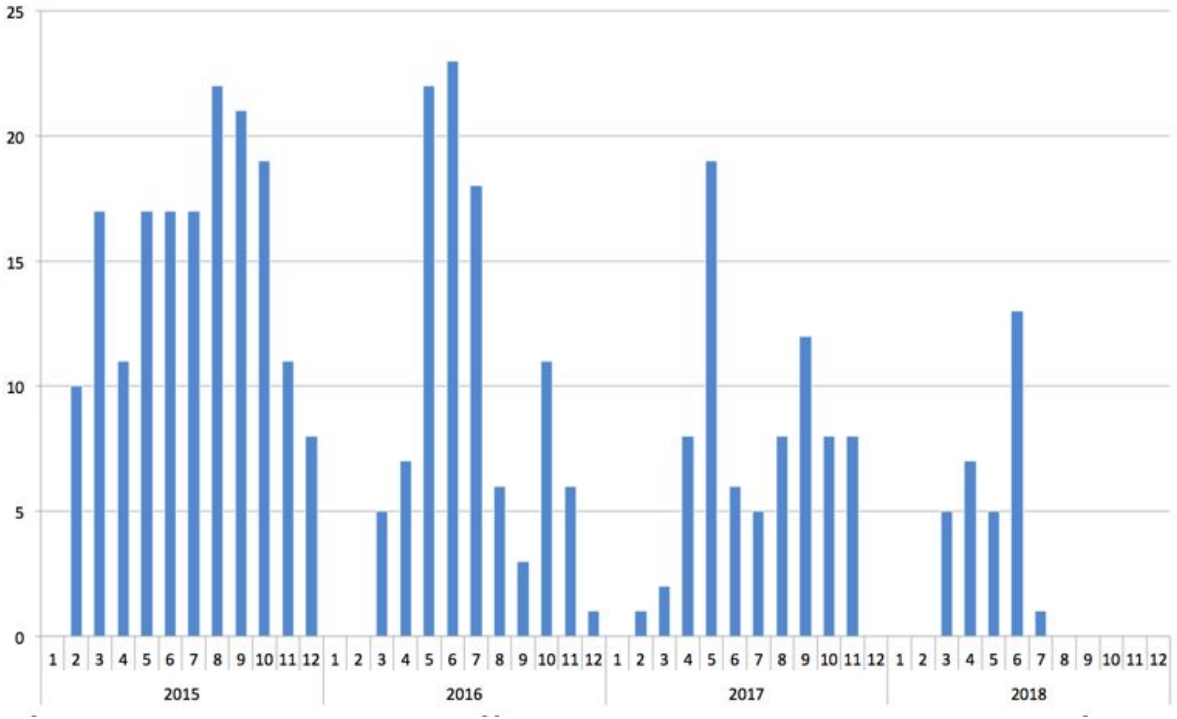
+ first guidance with a new tool
+ UX improvements
+ tools for post mortem analysis

+ coverage improved
+ 1 click recoveries
+ first fully automated recoveries (end of the Run 2)

+ fully automated recoveries wherever possible

On-call help demand

- Number of night-time calls to the on-call
- Clear trend
- Result of DAQExpert guidance and other improvements to the DAQ system



* Based on data provided by IT / CS, data from 2018 partly provided

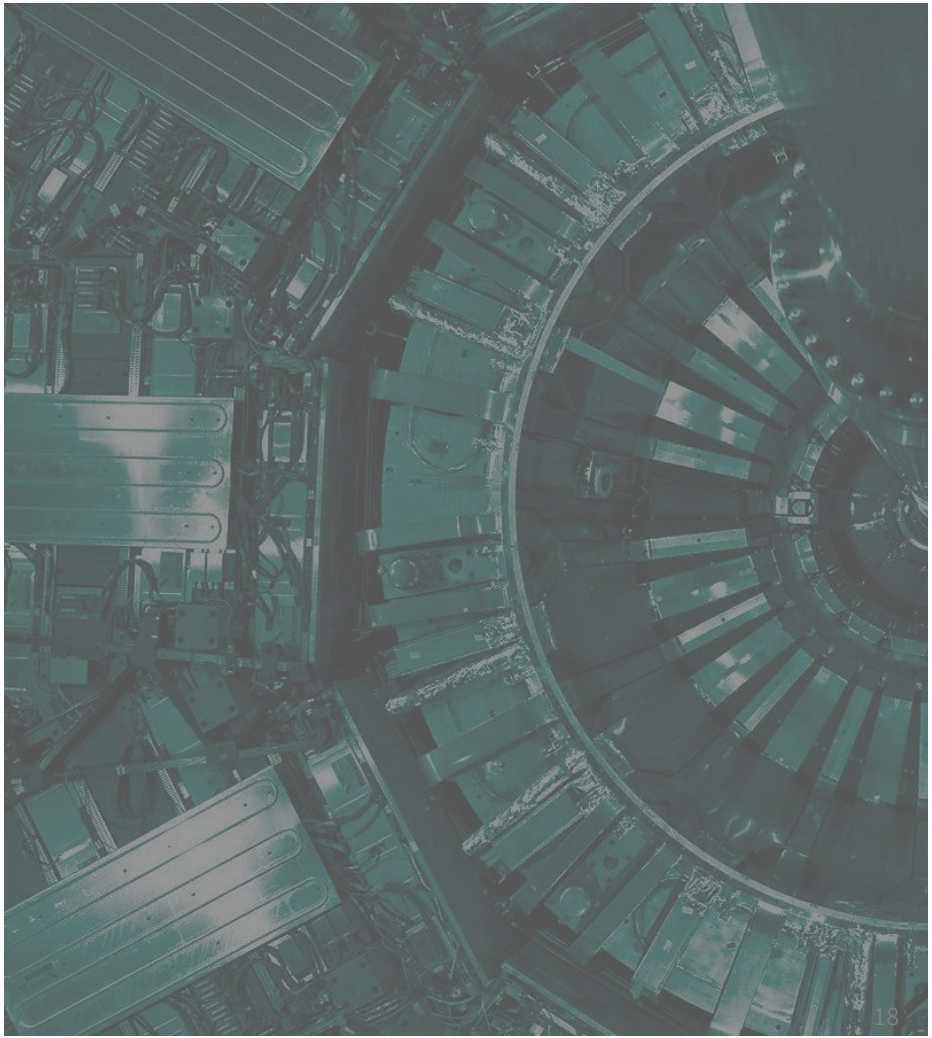
Summary

Maciej Gladki
maciej.gladki@cern.ch

1. **Expert tool is being improved**
   Automatic recoveries introduced in the end of Run2
2. **Successful at Run 2**
   Minimizes reaction time of operators
   - **101s** (2016) → **65s** (2017) → **47s** (2018)

   and reduces downtime.
   Circumstantial evidence (e.g number of night calls to DAQ on-calls reduced)
3. **Next in Run 3**
   Expecting to reduce even more downtime with automated recoveries
   - **17%** of DAQ downtime
   - **9%** of total CMS downtime

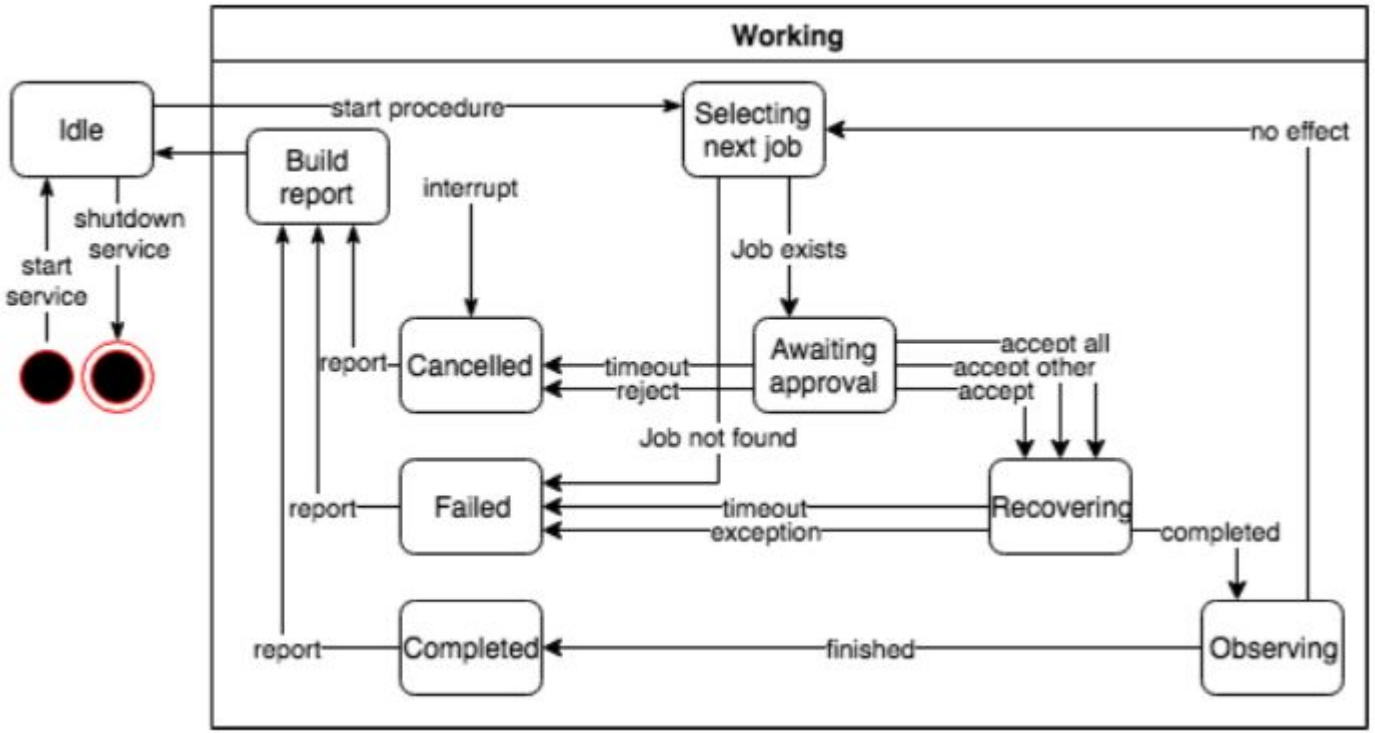
Extra slides



Abstract

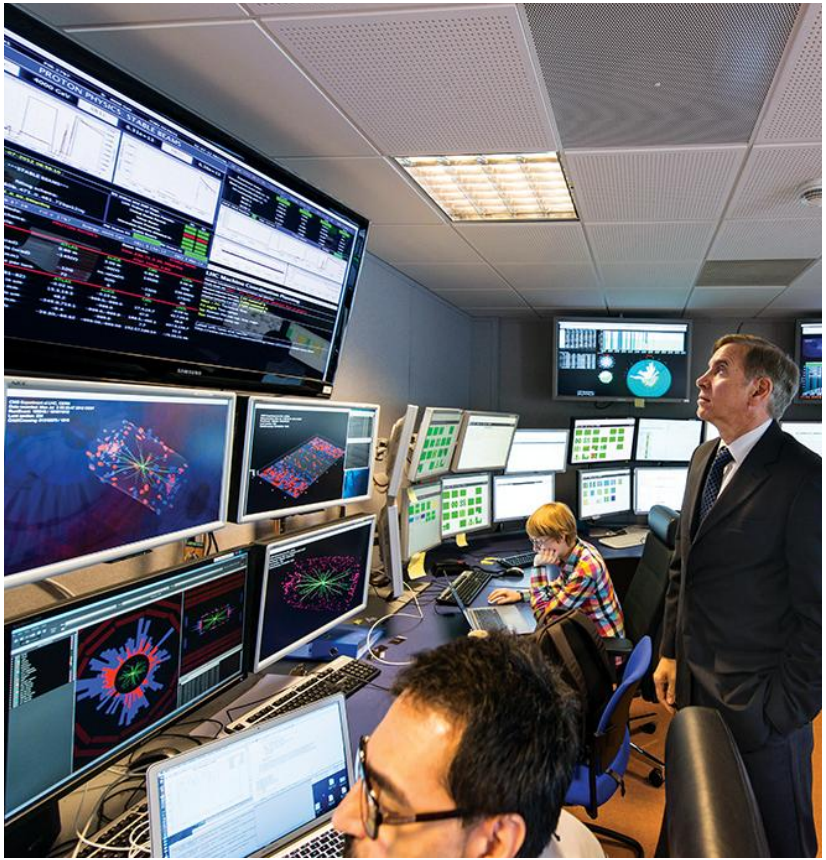
The Data Acquisition (DAQ) system of the Compact Muon Solenoid (CMS) experiment at LHC is a complex system responsible for the data readout, event building and recording of accepted events. Its proper functioning plays a critical role in the data-taking efficiency of the CMS experiment. In order to ensure high availability and recover promptly in the event of hardware or software failure of the subsystems, an expert system, the DAQ Expert, has been developed. It aims at improving the data taking efficiency, reducing the human error in the operations and minimising the on-call expert demand. Introduced in the beginning of 2017, it assists the shift crew and the system experts in recovering from operational faults, streamlining the post mortem analysis and, at the end of Run 2, triggering the fully automatic recoveries without a human intervention. DAQ Expert analyses the real-time monitoring data originating from the DAQ components and the high-level trigger updated every few seconds. It pinpoints the data flow problem and recovers it automatically or after given operator approval. We analyse the CMS downtime in the 2018 run focusing on what was improved with the introduction of automated recoveries; present challenges and design of transforming the expert knowledge to automated recovery jobs. Furthermore, we demonstrate the web-based, ReactJS interfaces that ensure an effective cooperation between the human operators in control room and the automated recovery system. We report on the operational experience with automated recoveries.

Controller FSM



Key quantities to monitor

- System availability (uptime)
- MTTR (reaction time)
- Wrong decisions overhead
- External help demand



Cumulative reaction time

- >400 DAQ related, datataking-problem interventions in 2018*
- 1-3 necessary human actions per intervention
- 20-25s reaction time
- > ~3h - 7h cumulative reaction time per year

*This does not include all interventions (no assignment by OMS of <30s downtimes -> case by case analysis needed)