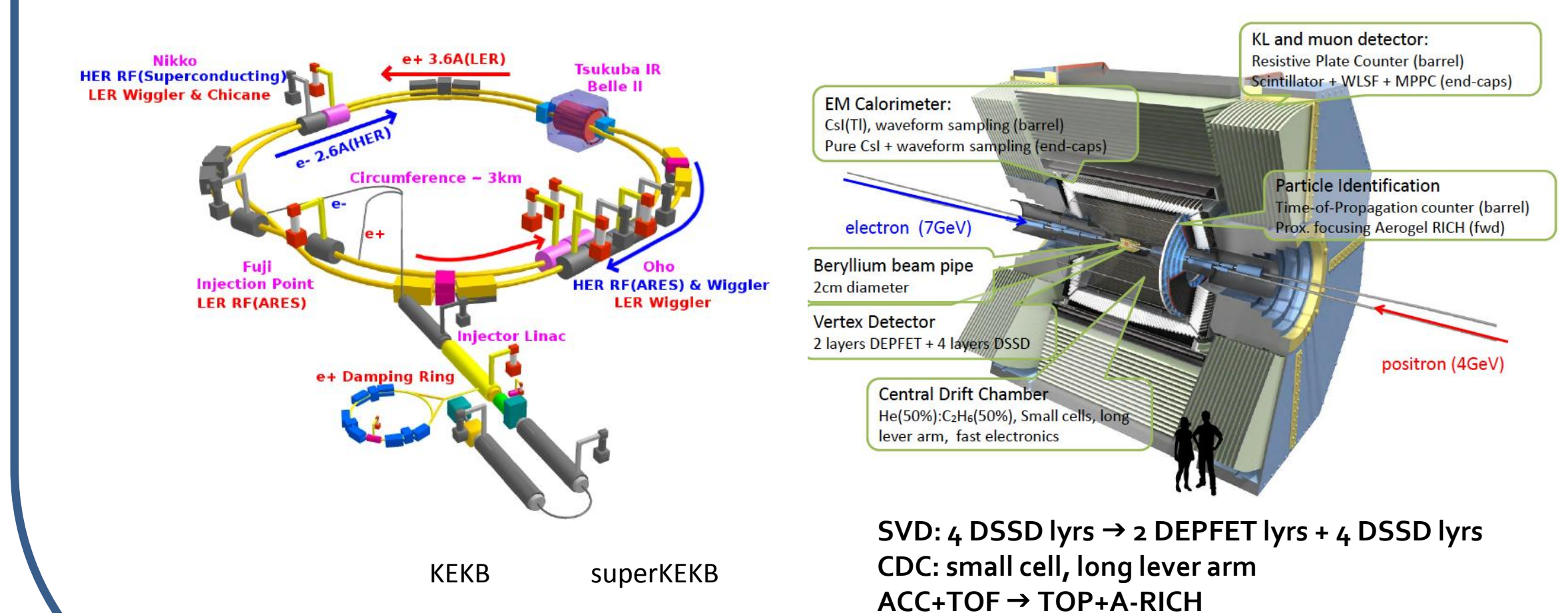
Job monitoring on DIRAC for Belle II distributed computing

Y. Kato (KMI,Nagoya U.), K. Hayasaka (KMI, Nagoya U.), T. Hara (KEK),
H. Miyake (KEK), I. Ueda (U. Tokyo/KEK) for Belle II computing group

Kobayashi-Maskawa Institute for the Origin of Particles and the Universe

Introduction

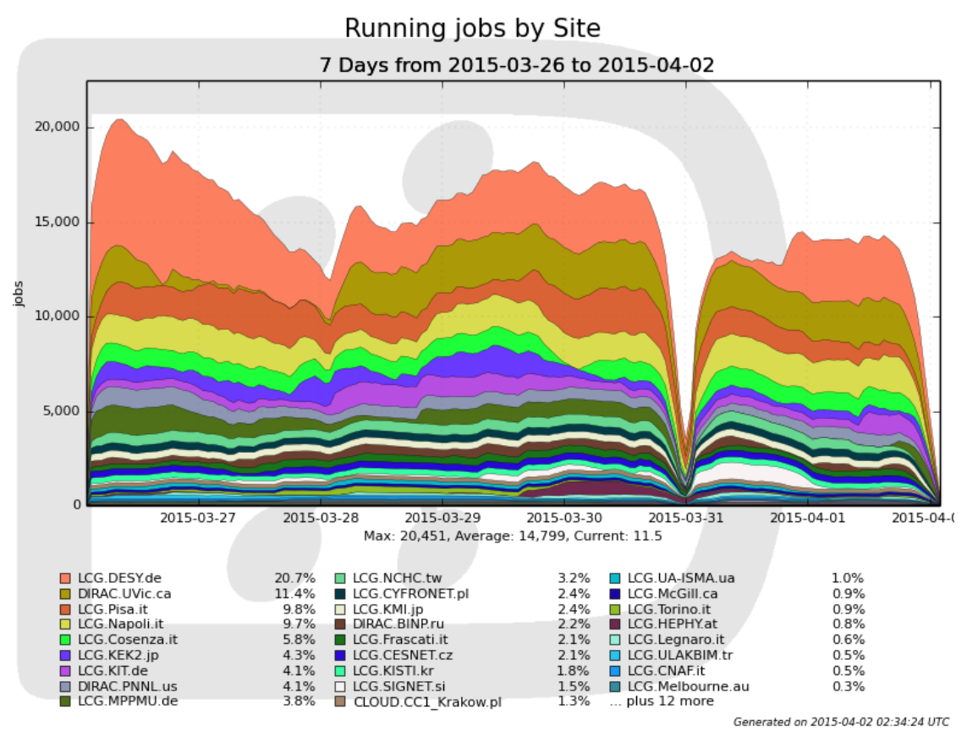
- Belle II experiment is a next-generation B-factory at KEK in Japan, which will start for physics run without vertex detector in 2017, where 50 ab^{-1} data sample will be collected for 10 years, which corresponds to about 5x10^{10} B$\bar{B}$-pair events.
- We roughly need to handle 1MHS06 cpu resources ,100PB storage for one set of raw data and 100 PB one for MC/analysis data, finally.
- In order to utilize these huge resources, we adopt distributed computing technique.

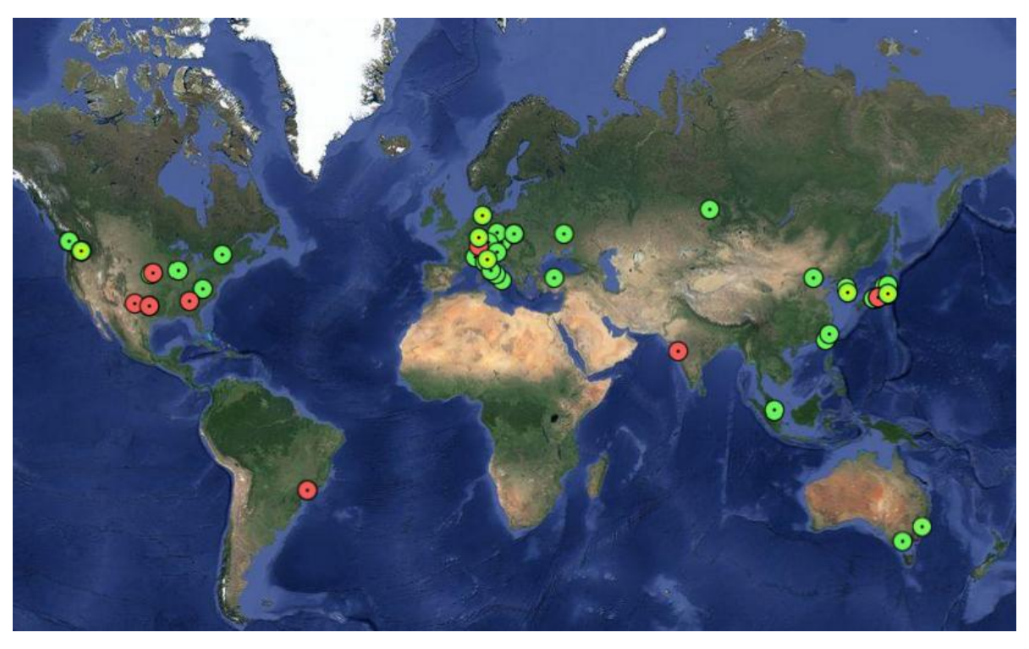


SVD: 4 DSSD lyrs → 2 DEPFET lyrs + 4 DSSD lyrs
CDC: small cell, long lever arm
ACC+TOF → TOP+A-RICH
ECL: waveform sampling, pure CsI for end-caps
KLM: RPC → Scintillator +SiPM (end-caps)

Belle II computing

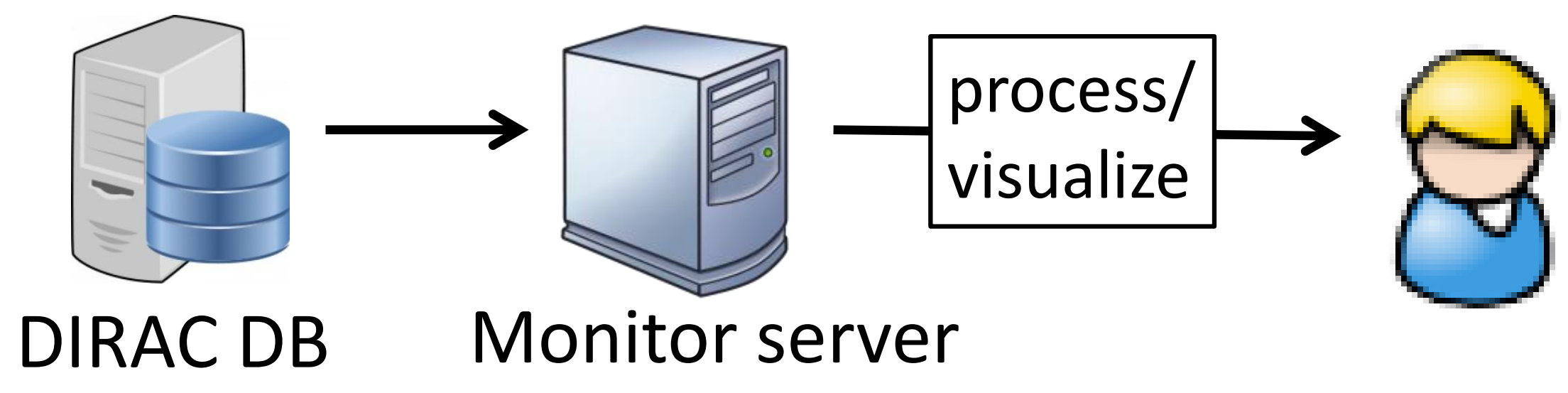
- Belle II has adopted DIRAC as the distributing computing software framework, which can handle grid, cloud and local cluster resources. (http://diracgrid.org/)
- CVMFS is used to provide Belle II software and libraries.
- At the present, around 40 sites participates (LCG, OSG, HPC, cloud and traditional cluster) and more than 25K concurrent jobs are handled at peak.





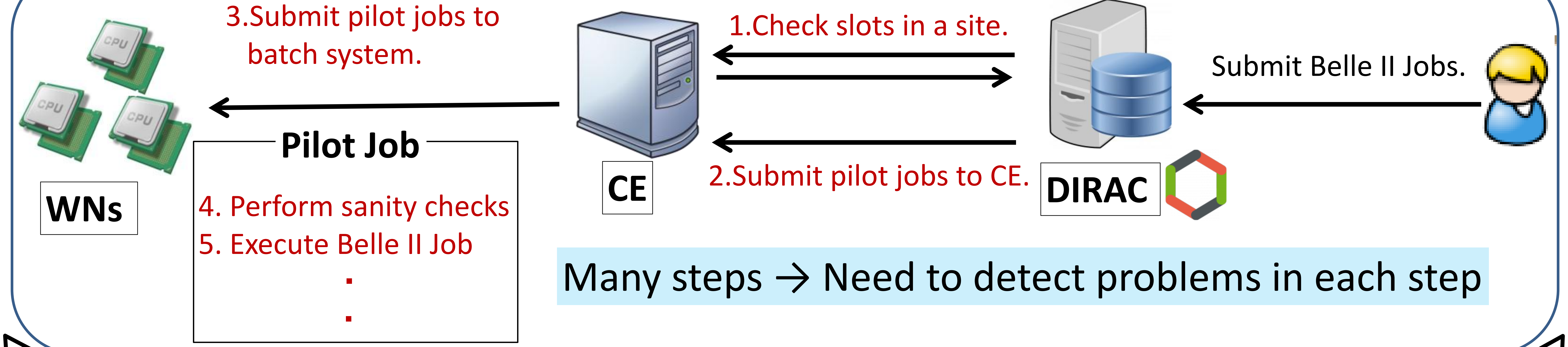
Monitoring

- For the effective use of huge resources, a monitor system for detecting problems quickly and identifying the source is necessary.
- In this poster, we introduce passive monitors, where data existing in DIRAC DB are retrieved and then processed and visualized to detect problems.
- In some cases, necessary information are not stored in DIRAC DB. In such cases, DIRAC agents which collect information are developed.



- For active way, please visit poster by K. Hayasaka (sessionB, poster 314, booth 18).

Workload management flow in the DIRAC

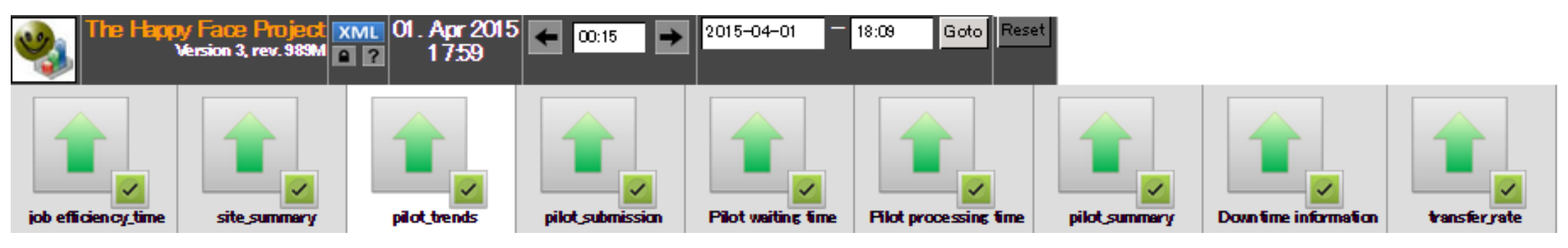

3.Submit pilot jobs to batch system.
1.Check slots in a site.
Submit Belle II Jobs.
CPU
WNs
CE
DIRAC
2.Submit pilot jobs to CE.

Pilot Job

4. Perform sanity checks
5. Execute Belle II Job
·
·

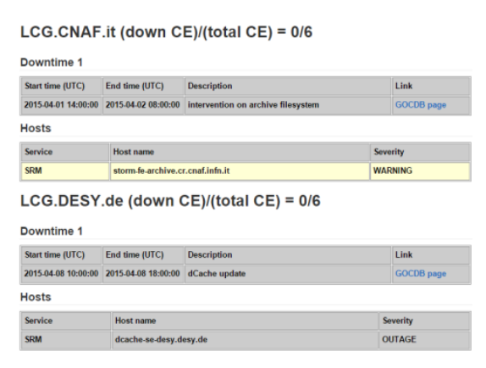
Many steps → Need to detect problems in each step

HappyFace as a platform



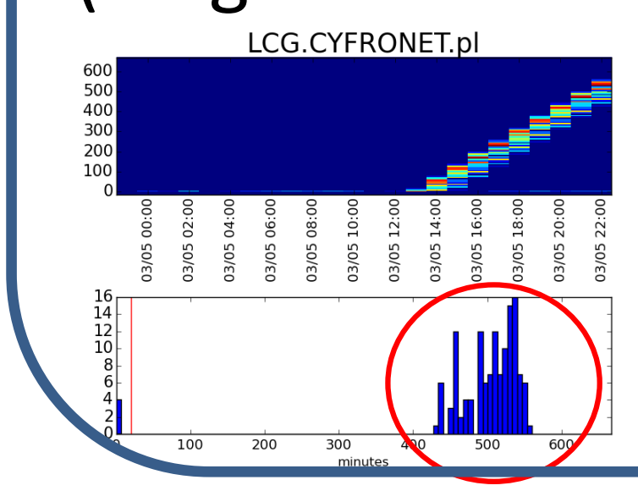
https://ekptrac.physik.uni-karlsruhe.de/trac/HappyFace/

- Developed at Karlsruhe Institute of Technology
- Modular structure.
- In the Belle II HappyFace instance, not only for workload management issue but also downtime etc are shown.



1.Check slots in a site.

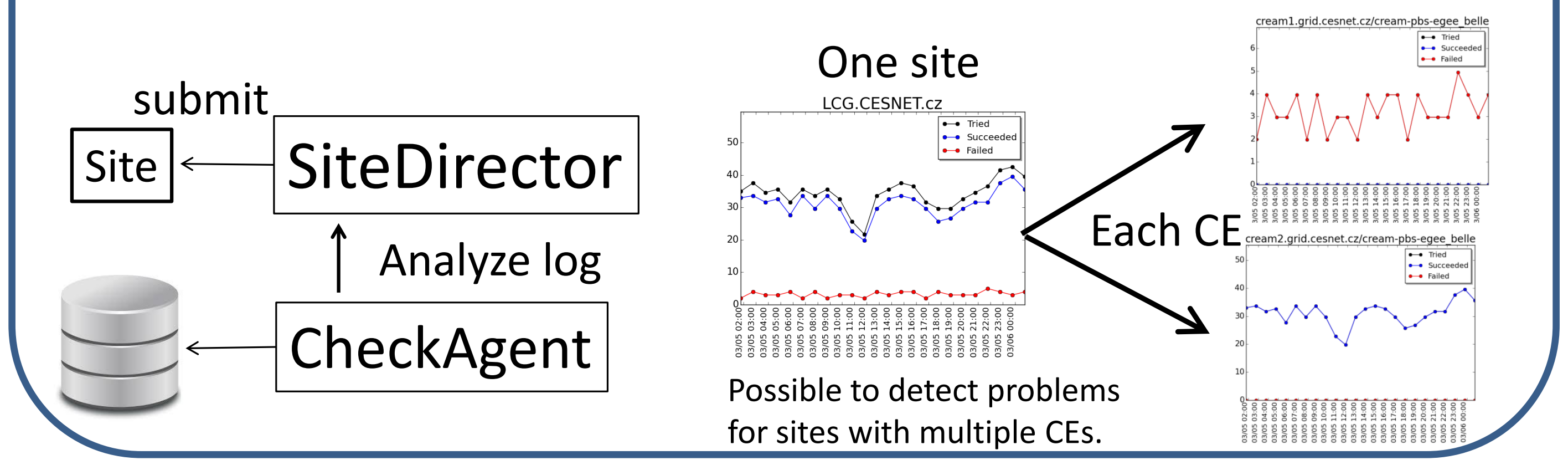
Sometimes, **CE reports incorrect # of running pilot jobs** due to the problem of CREAM etc. In such cases, DIRAC misunderstands site is full and stops to send jobs. This problem can be characterized by long-keeping-silent pilot jobs (long time since last communication with DIRAC).



Pilot silent time distribution (in minutes). Red line shows the possible maximum silent time for normal pilot jobs. In this case, CREAM-CE recognizes finished job as running.

2.Submit pilot jobs to CE

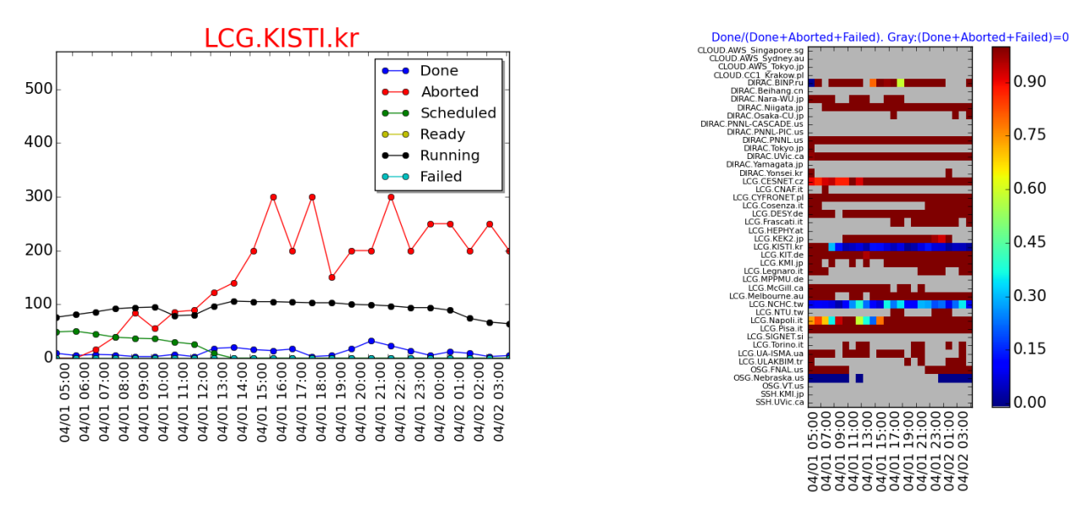
Submission of pilot jobs to CE often fails because of **CE down or problem on VOMS proxy** etc. Pilot jobs are sent by "SiteDirector" agent but activity is not stored in DB. DIRAC agent to monitor the activity of SiteDirector is developed and visualized.



Possible to detect problems for sites with multiple CEs.

3.Submit pilot jobs to batch system

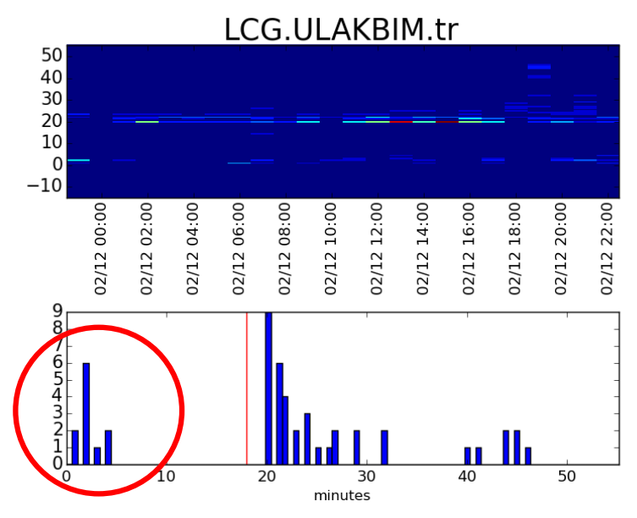
Submission to batch server often fails because of **problem on the batch system.** If it is failed, status of pilot job becomes "Aborted".



Example of error message:
[BLAH error: submission command failed (exit code = 1) (stdout:) (stderr:qsub: Queue is not enabled MSG=queue is disabled.]

4.Perform sanity checks

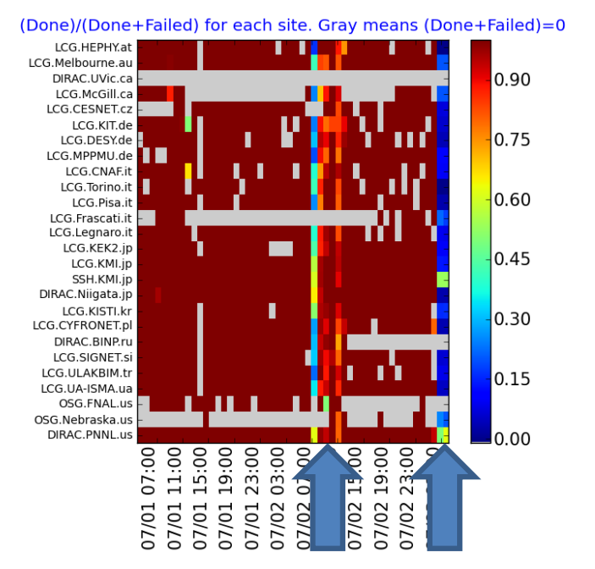
At the beginning of the pilot job, DIRAC client is installed to communicate with DIRAC server. Then, sanity checks of the computing node are performed. If a problem is found, the pilot job stops immediately.
Ex. **CVMFS not properly mounted, disk full, failed to download DIRAC client etc..**



Pilot life time distribution (in minutes). Redline is possible minimum life time for normal pilot jobs. In this case, one of WNs does not have enough disk space.

5.Execute Belle II Jobs

Payload jobs may fail with many reasons. **For example, failed to contact meta data server (AMGA), failed to handle input/output files, and problem on program itself.**



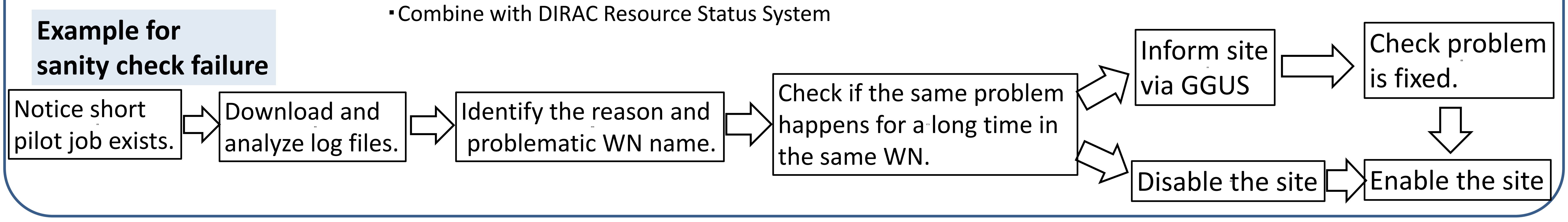
"Job efficiency" for each site. Simultaneous failure for all the site means problem on central server. In this case, AMGA was down.

Now, we can detect problems in each step!

Automate the process (work in progress)

- Next step is to identify reason (as much as possible) and inform/disable each site.
- These process should be automated.
- Combine with DIRAC Resource Status System

Example for sanity check failure



Notice short pilot job exists. ⇒ Download and analyze log files. ⇒ Identify the reason and problematic WN name. ⇒ Check if the same problem happens for a long time in the same WN. ⇒ Inform site via GGUS ⇒ Check problem is fixed. ⇒ Enable the site

Check if the same problem happens for a long time in the same WN. ⇒ Disable the site ⇒ Enable the site

We aim to resolve the problem quickly and maximize the availability of Belle II computing system!