Distributed analysis with CRAB: the client-server architecture evolution and commissioning

Giuseppe Codispoti
Dipartimento di Fisica, Università di Bologna & INFN

On behalf of CMS Collaboration

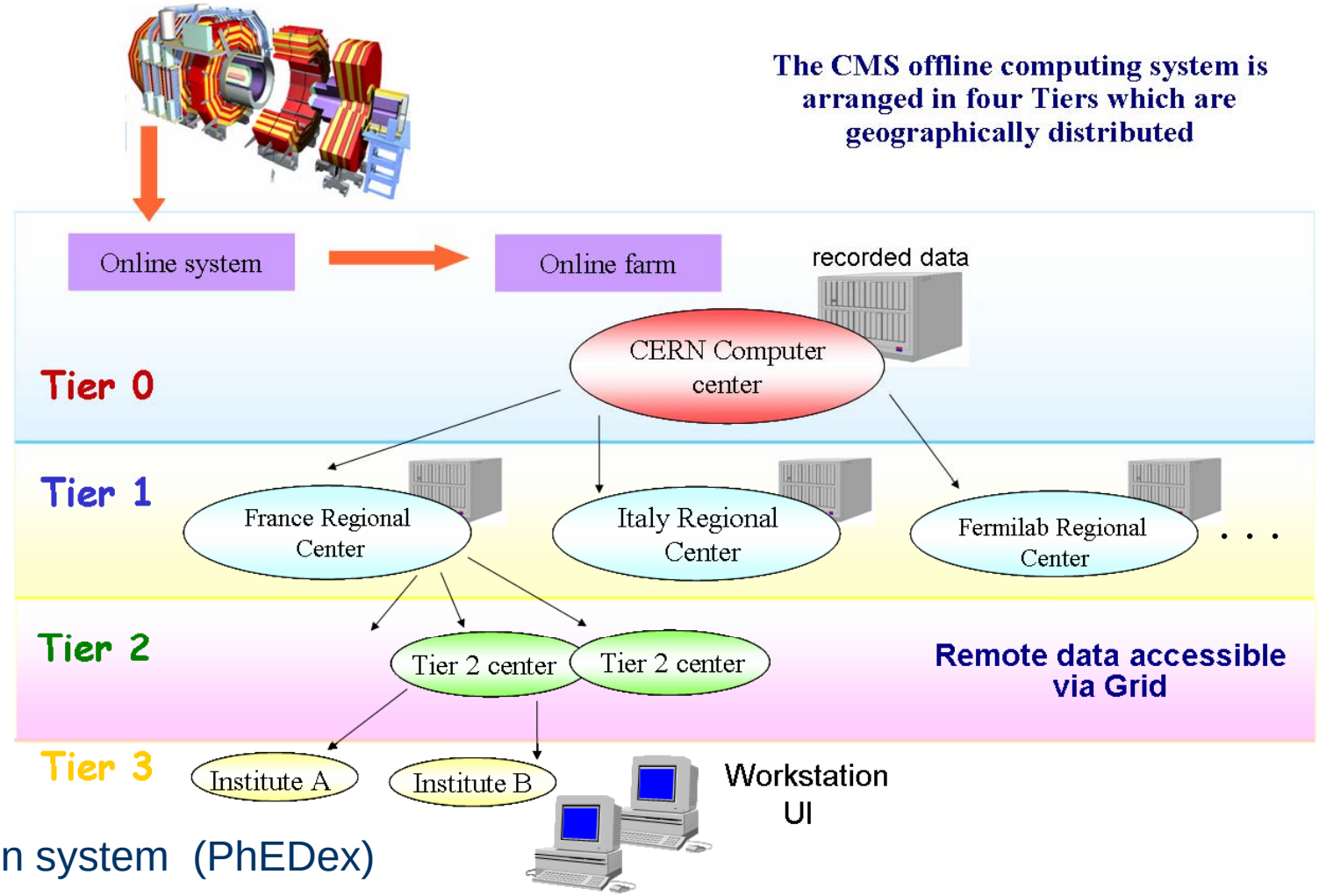
Acat 2008

Ettore Majorana Foundation and Centre for Scientific Culture, Erice (Sicily), Italy
November 3-7, 2008

Overview

- CMS and the Distributed Computing
- The Analysis Model
- CRAB
- Motivation for the Analysis Server
- The Server Architecture
- The Server User Interfaces
- Server Scale Tests
- CRAB Usage
- What about the Future?

The CMS Distributed Environment



-CMSSW

-Data distribution system (PhEDex)

-Data Bookkeeping system (DBS)

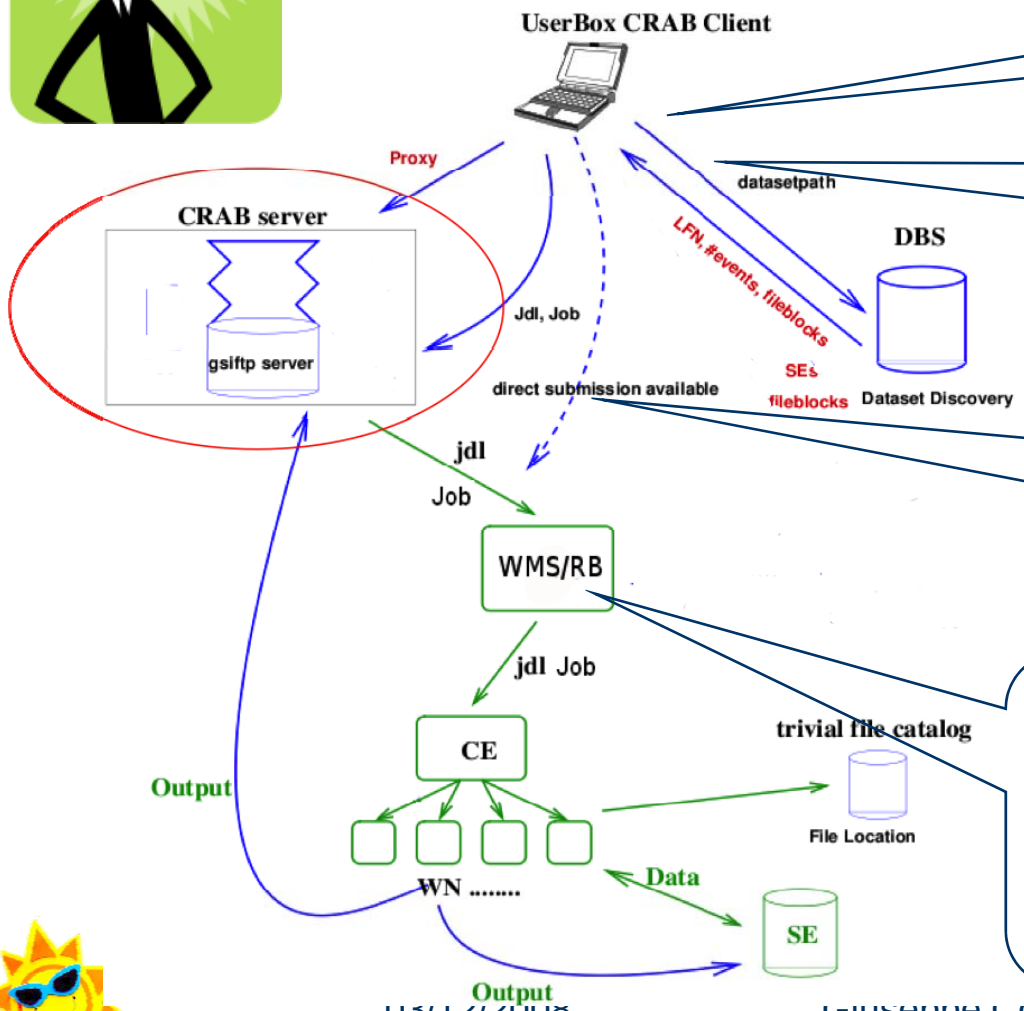
-Grid flavours and local batch system (WLCG, OSG, lsf..)

CRAB and the User Analysis Model

- CMS Analysis model:
  - User runs interactively on small samples in the local environment in order to develop his analysis code and test it
  - User selects a large (whole) sample to submit the very same code to analyze zillions of events (no remote user code compilation)
  - User's analysis code is transported to the sites where sample of interest is located, following the CMS data location driven computing model
  - Results are made available to the user to be analyzed interactively to produce the final plot
- Cms Remote Analysis Builder (CRAB):
  - A friendly interface for the physics distributed analysis
    - interacts with the user code and in general with the Analysis software (CMSSW)
    - interacts with the Data distribution system (PhEDex)
    - interacts the Data Bookkeeping System (DBS)
    - interacts with the Grid flavours and local batch system (WLCG, OSG, lsf..) for job execution.

CRAB Workflow

Write analysis code;
Select the dataset to be analyzed;
Decide job parameters (e.g. event number per job) and other CRAB configuration



1) Finds data information (physical files location and their content)

2) Splits the user task in many jobs, according to the user configuration and the data location
Packs the user code

3) Submit jobs to the Grid: the data location is used by Grid Workload Management tools to match remote resources

4) through Grid tools: Track jobs status, Retrieve output or move it to a user defined Storage Element, Allow more jobs action such as cancelling, resubmitting, retrieve verbose log for job failures

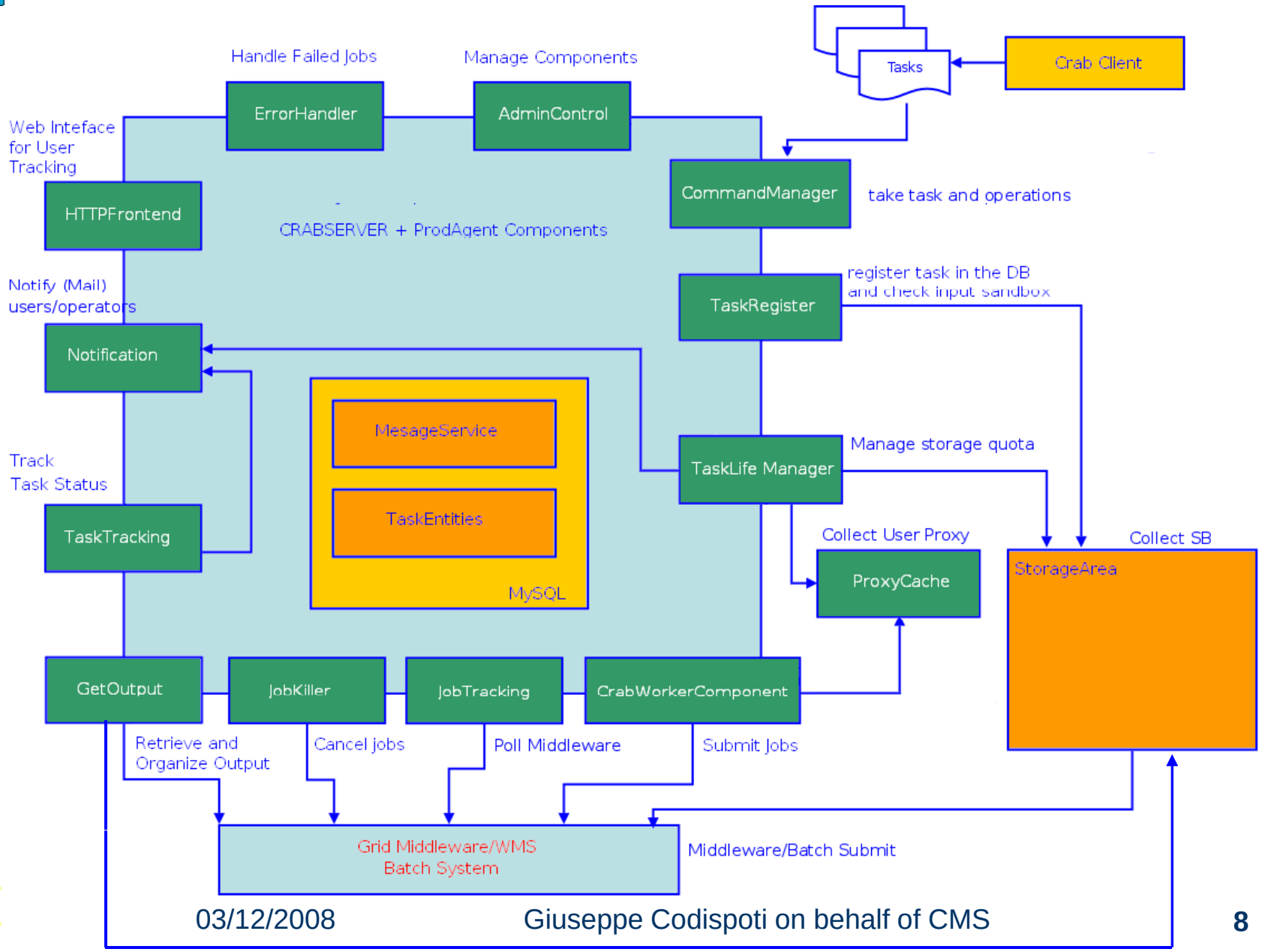
Analysis Server

- Provide logging information about user jobs as well as the system usage
- Automate as much as possible the whole analysis workflow
  - Reduce the unnecessary human load, moving all possible actions to server side, reducing to a minimum those on client side
- Automate as much as possible the interaction with the Grid
  - perform submission, resubmission, error handling, output retrieval, post-mortem operations, etc. . .
  - Allow better job distribution and management
- More in general:
  - Improve scalability of the whole system providing to CMS specific functionalities which cannot to be not included in the Grid middleware but are heavily used by the generic CMS analysis job

Server Architecture (1)

- The server adopts a modular software approach:
  - independent components implemented as agents
  - where needed, multithreading approach in the agent implementation
  - communication through an asynchronous and persistent message service (publish & subscribe model)
- The system core is a MySQL DB (message service, job logging & bookkeeping …)
  - other external components used for data storage and transfer:
    - default GridFTP but also: rfio, dcache…
- The architecture is as similar as possible to the other CMS Workload Management Tools: reusing code, sharing efforts

Server Architecture (2)



Server from the User Point of View

- CRAB command line
  - batch-like system
- Mail notification
- HTTPfrontend

```
[lxplus239] ~/scratch0/WorkOK > crab -create -submit
crab. crab (version 2.4.2) running on Tue Oct 28 16:37:16 2008

crab. Working options:
  scheduler           glite
  job type            CMSSW
  working directory   /afs/cern.ch/user/s/spiga/scratch0/WorkOK/crab_0_081028_163716/

crab. Contacting Data Discovery Services ...
crab. Requested dataset: /Zmumu/CSA08_CSA08_S156_v1/GEN-SIM-RECO has 16862 events in 1 blocks.

crab. May not create the exact number_of_jobs requested.
crab. 10 job(s) can run on 100 events.

crab. List of jobs and available destination sites:

Block     1: jobs                1-10: sites: srm.ciemat.es,storm.ifca.es,t2-srm-02.lnl.infn.it,srm.cern.ch,cmssrm.hep.wisc.edu,srm.grid.sinica.edu.tw

crab. Creating 10 jobs, please wait...

crab. Total of 10 jobs created.

crab. Registering a valid proxy to the serv
crab. Proxy successfully delegated to the s

crab. Starting sending the project to the s
crab. Task crab_0_081028_163716 successfull

crab. Total of 10 jobs submitted
```

Event: **CRAB_Cmd_Mgr:NewTask**

date: 1224186408.2

txt: Arrived task: fanfani_StressTest-Round4_16_83b838ac-1f8d-4d56-8a41-66bd77628428

Event: **TaskRegisterComponent:NewTaskRegistered**

date: 1224186427.28

txt: Task in submission queue: fanfani_StressTest-Round4_16_83b838ac-1f8d-4d56-8a41-66bd77628428

Event: **CrabServerWorkerComponent:FatWorkerResult**

code: 0

reason: Full Success for fanfani_StressTest-Round4_16_83b838ac-1f8d-4d56-8a41-66bd77628428

time: 61.7795169353

date: 1224186492.37

txt: Submission completed: fanfani_StressTest-Round4_16_83b838ac-1f8d-4d56-8a41-66bd77628428

Event: **Reached 1**

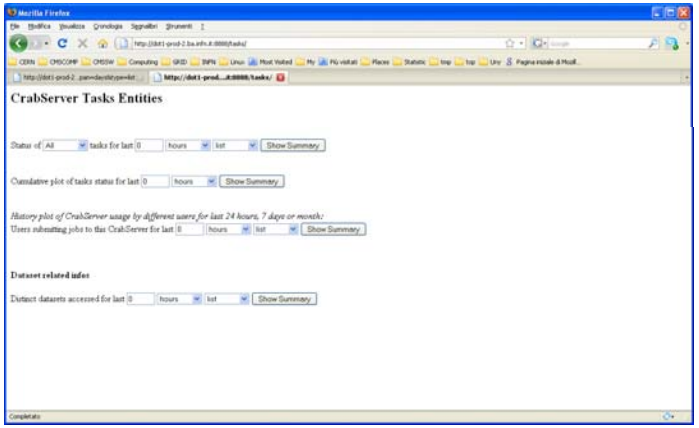
date: 1224186571.54

txt: publishing task success (sending e-mail to)

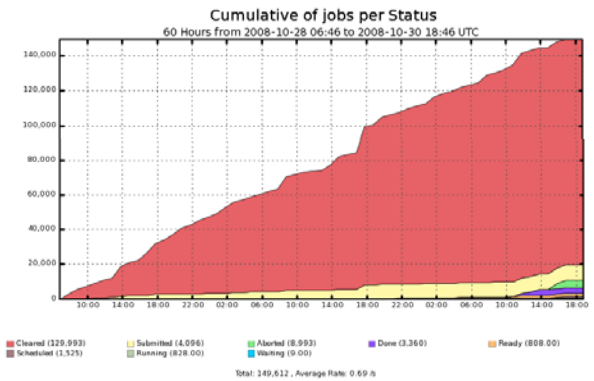
```
From:     crab@lnl.infn.it
Subject:     "CrabServer@crabas-cms-1.cr.cnaf.infn.it Notification: The task [crab_0_081028_125019] is completed at 100%

Date:     October 28, 2008 1:12:19 PM GMT+01:00

To:     Undisclosed recipients:;

    The task 'crab_0_081028_125019' owned by Giuseppe Codispoti and composed by 40 job(s)
    is completed at: 100%

    Status Report:
    40 Job(s) in status [Done]
    1 Job(s) in status [Killed]
```

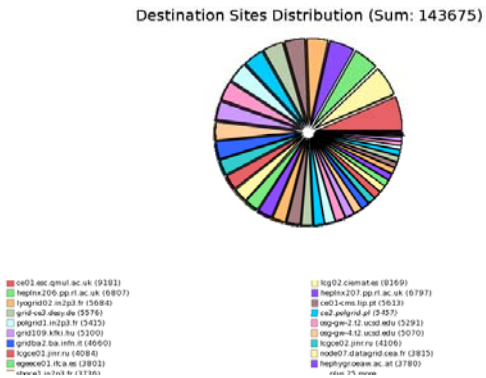
Server from the Admin Point of View

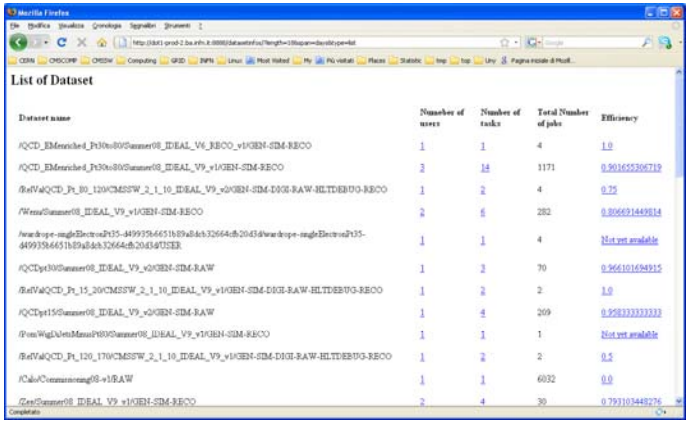
- HTTPfrontend:
  - Shows Tasks & Jobs Information:
    - Jobs cumulative status
    - Datasets accessed
    - Jobs destinations
    - User monitoring



  - Monitor Components and Services status:
    - Display the status of components and active services
    - Allow to access components logs through web







Scale and Performance Tests

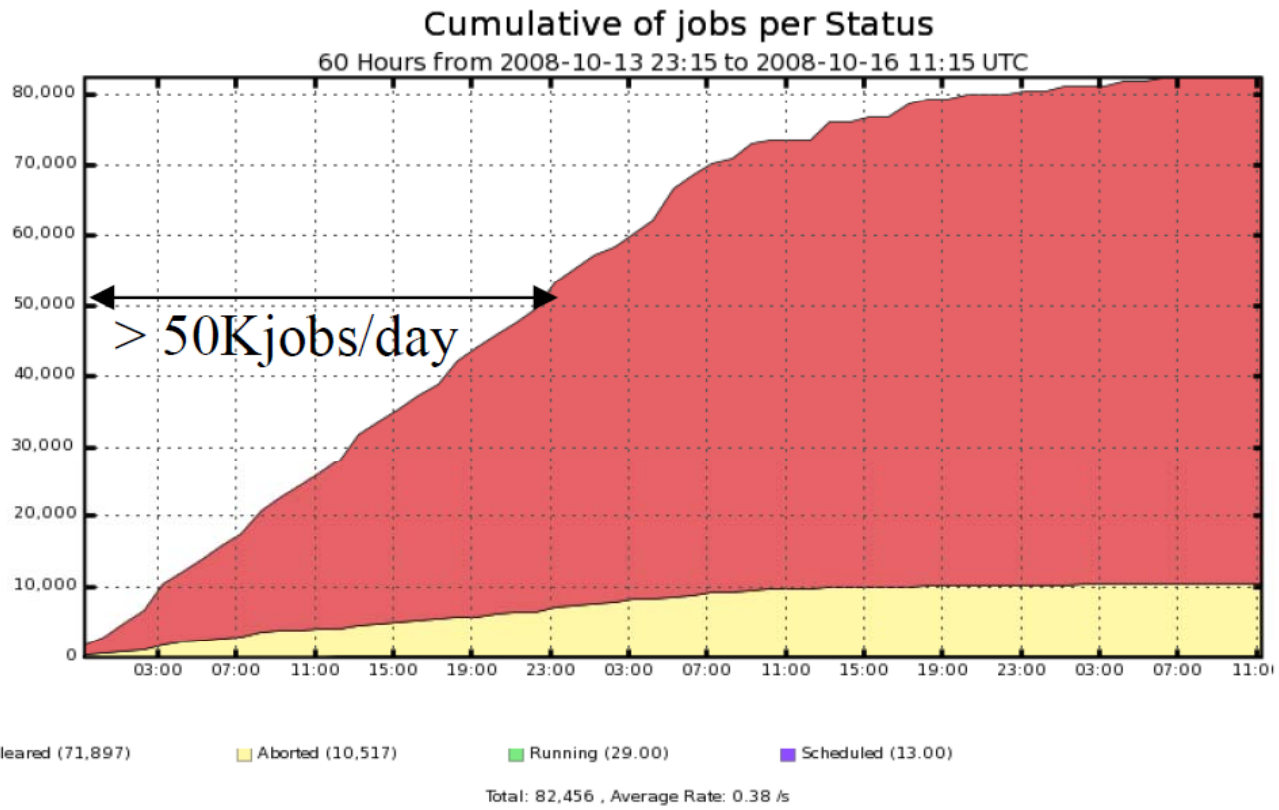
- Setup of a dedicated test environment:
  - Controlled job submission for few days
  - starting with a single GLite WMS, adding more up to an eventual breaking point
- Collecting monitoring info from various sources (WMS, GridFTP server, CRABserver)
- Emulate multi-user environment, after an initial single user phase

Single User Scale Tests Setup

- Very short jobs not reading an input dataset
- Zipped Sandbox of about 8MB
- Submitted to all sites excluding T1s
  - Constant rate : 500-600 jobs every 20min
  - Plus Peaks of 1000-2000 jobs every 5 hours
- From 40Kjobs/day to 50Kjobs/day submitted

Single User Scale Test Results

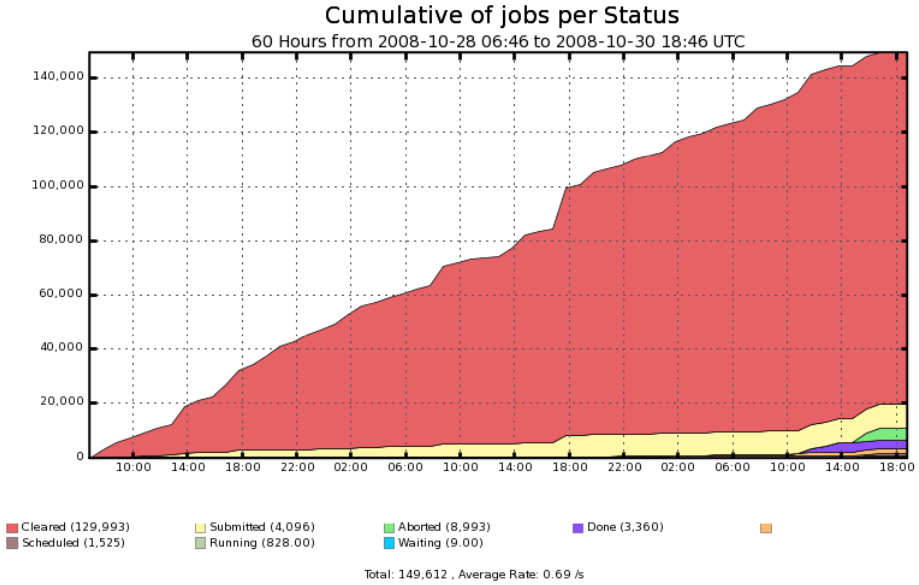
- ~ 200Kjobs handled
- More than 50 CEs
- CRAB Server scales with 2 WMS!
- There is no indication of reaching a breaking point: It can likely handle more!!!



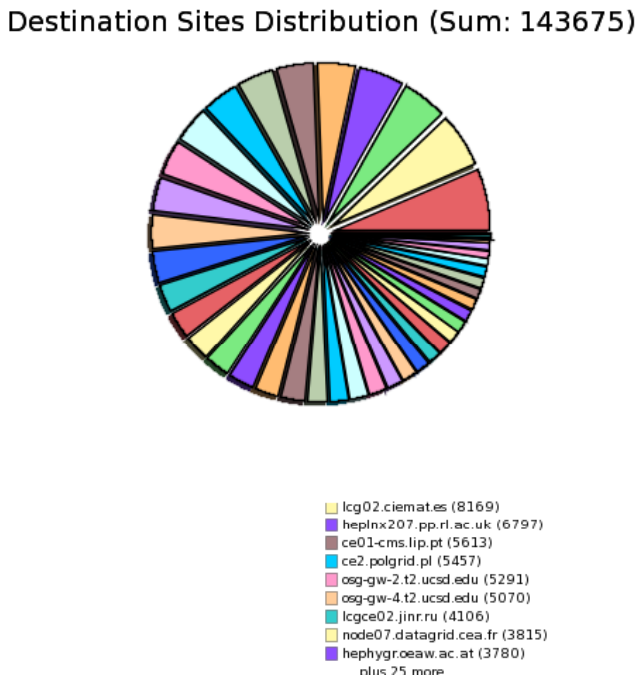
MultiUser Test Setup

- Job submission from different user certificates using more WMSs
- 6 users with a variety of submission patterns:
  - multi users scheduled submission:
    - 1 user submitting a task of 100jobs every 15min
    - 1 user submitting a task of 500 jobs every 20min
    - 1 user submitting a task of 600 jobs every 3hours
    - 1 user submitting a task of 2000 jobs every 6hours
  - random submissions: 2 volunteer users submitting at their will
- Increase in the DB content, starting from a not empty DB (about 200Kjobs from previous testing)

MultiUser Test Results



- 120Kjobs in 48 hours (with 2 WMS) with peaks of 2-3 Hz
- No evident problems with DB and gridFTP server

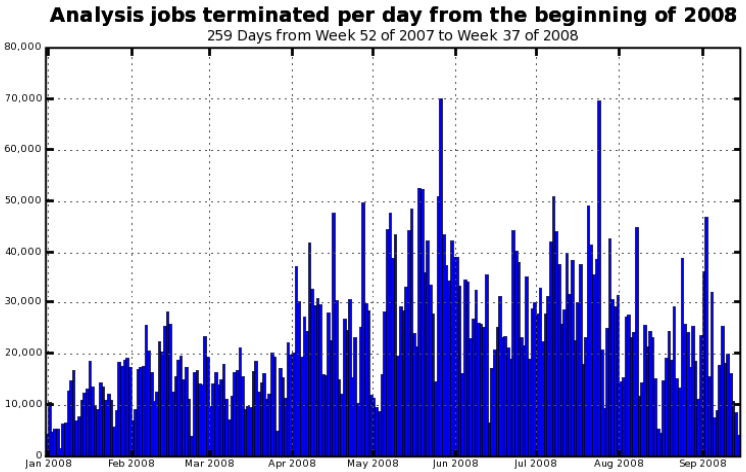


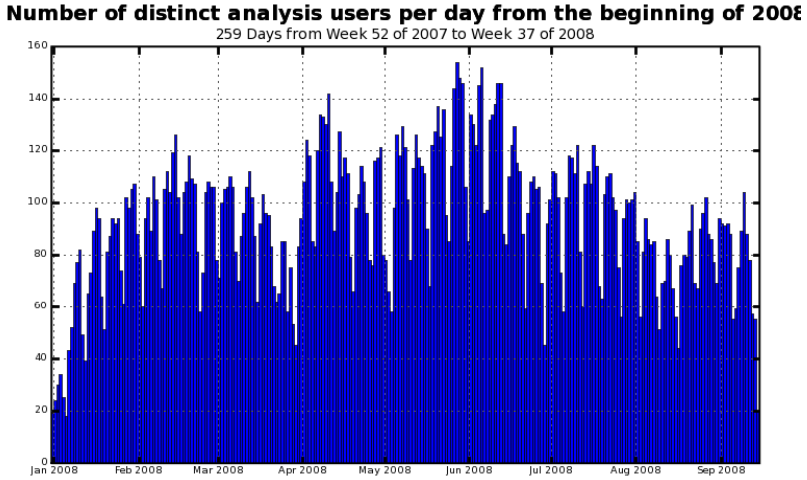
- But also indication for further improvements:
  - We are confident we can do better and we know how

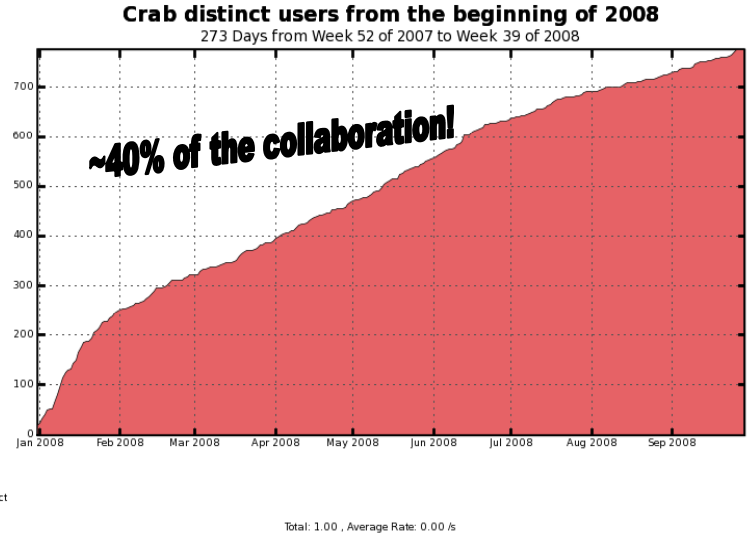
Outlook

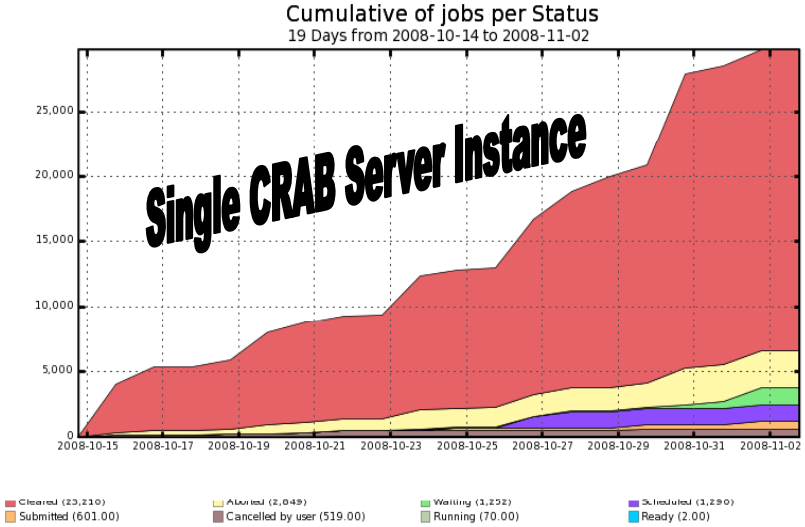
- Realized a friendly interface for the physics distributed analysis (see next slide for actual usage)
- Realized an analysis server automating the interaction with the Grid, reducing the unnecessary human load, scalable and reliable
- Stressed the system up to high rates, trying to reach the expected rate for CMS operations

CRAB Usage


Analysis jobs terminated per day from the beginning of 2008
259 Days from Week 52 of 2007 to Week 37 of 2008
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Jan 2008
Feb 2008
Mar 2008
Apr 2008
May 2008
Jun 2008
Jul 2008
Aug 2008
Sep 2008


Number of distinct analysis users per day from the beginning of 2008
259 Days from Week 52 of 2007 to Week 37 of 2008
160
140
120
100
80
60
40
20
0
Jan 2008
Feb 2008
Mar 2008
Apr 2008
May 2008
Jun 2008
Jul 2008
Aug 2008
Sep 2008


Crab distinct users from the beginning of 2008
273 Days from Week 52 of 2007 to Week 39 of 2008
~40% of the collaboration!
700
600
500
400
300
200
100
0
Jan 2008
Feb 2008
Mar 2008
Apr 2008
May 2008
Jun 2008
Jul 2008
Aug 2008
Sep 2008
Direct
Total: 1.00 , Average Rate: 0.00 /s


Cumulative of jobs per Status
19 Days from 2008-10-14 to 2008-11-02
Single CRAB Server Instance
25,000
20,000
15,000
10,000
5,000
0
2008-10-15
2008-10-17
2008-10-19
2008-10-21
2008-10-23
2008-10-25
2008-10-27
2008-10-29
2008-10-31
2008-11-02
Submitted (601.00)
Cancelled by user (519.00)
Running (70.00)
Ready (2.00)

What about the future?

- Post-mortem operations
  - output merge, data movement etc…
- Extending user interface (web & graphics)
- Improving monitoring tools also for the administrator
- Workflow automate
- Anything that can be automated

Thanks
To
Everybody

Giuseppe