



Running the CMS data analysis on the Grid

The workflow and the experience

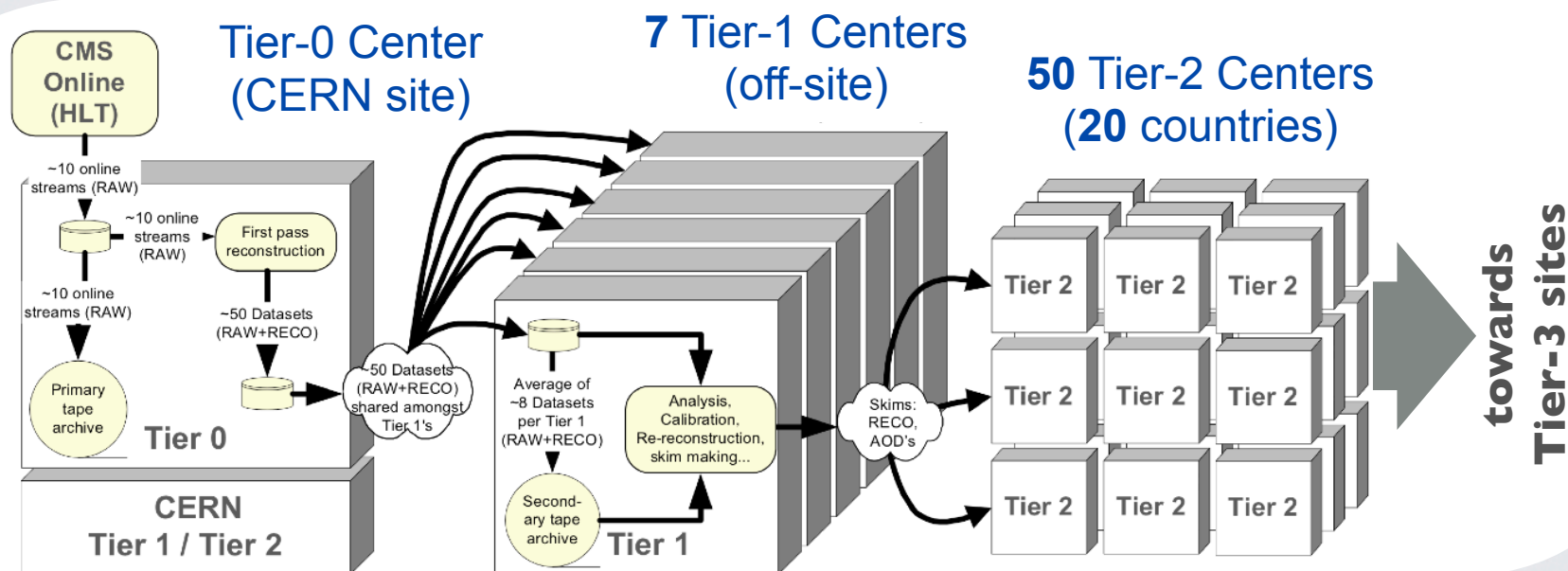
Predrag Milenovic,
ETH Zurich

The LHC Grid and the CMS computing model

- The LHC experiments will produce approximately **15 PB of data each year** (physicists around the world want to access and analyze this data)
- Computing and storage resources for the LHC Grid are distributed at **11 Tier-1 sites, ~160 Tier-2 sites** (32 countries)

The CMS computing model

Worldwide LHC Computing Grid Infrastructure



The analysis workflow

- A typical analysis workflow



- Several **questions** and **tools** to answer them:

- **Where** is my data located? Where can I run my jobs? —————→ **DBS**
- How do I **transfer** data to where I want to run? —————→ **PHEDEX**
- Is the right software available at that site? —————→ **CRAB**
- How time consuming is my job?
- Are my jobs running? **Monitoring?** —————→ **Dashboard**

Finding the dataset

- The tool named **Dataset Bookkeeping Service (DBS)** provides means to **describe, discover and use** the CMS event data

The screenshot shows the DBS web interface. At the top, there are navigation tabs: Dashboard, DBS Discovery (selected), DataTransfer, SiteDB, CondDB, T0Mon, and Support. Below the tabs is a breadcrumb trail: Home - aSearch - Navigator - RSS - Status - Runs - Admin - Tools - Help - Contact - TinyURL. The main content area is divided into two sections. The first section is titled "ADVANCED KEYWORD SEARCH" and contains a dropdown menu for "DBS instances" set to "cms_dbs_prod_global", a "HELP" button, and a search input field with the text "find dataset where dataset like *Run2010A-PromptReco-v4/RECO* and datas". There are "Search" and "Reset" buttons. The second section is titled "MENU-DRIVEN INTERFACE" and contains two dropdown menus: "Physics groups" set to "Any" and "Data tier" set to "Any".

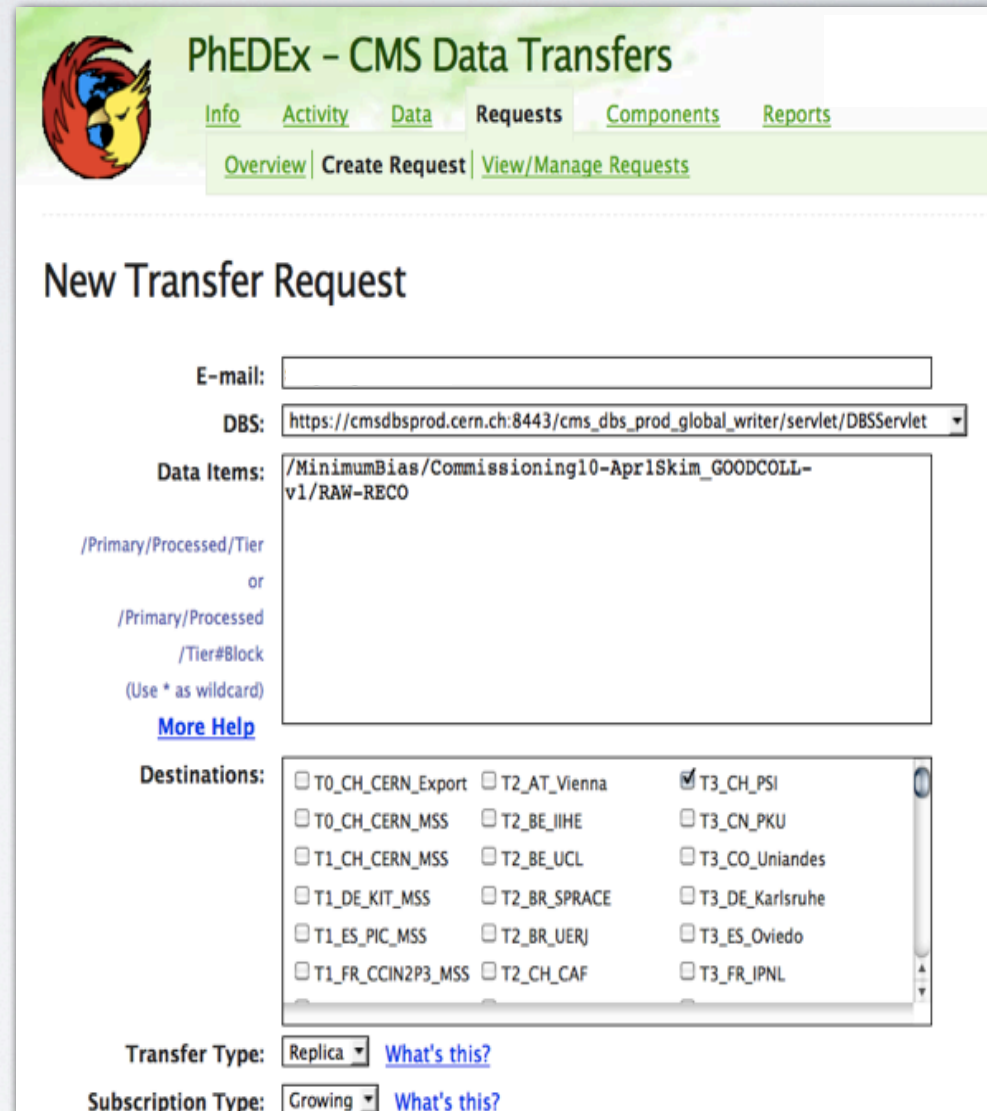
- Contains all the necessary information on:
 - locations
 - num. of events / num. of files / size
 - date of production
 - configuration used for production of dataset
 - software versions...
- Accessible via **web-interface & command line**

/Mu/Run2010A-PromptReco-v4/RECO
Created 10 Jun 2010 15:01:28 GMT, contains 23977520 events, 1608 files, 516 block(s),
[Release info](#), [Block info](#), [Run info](#), [Conf. files](#), [Parents](#), [Children](#), [Description](#), [PhEDEx](#)

Location	Events	Files	size	LFNs
T2_CH_CSCS : storage01.lcg.cscs.ch	23977520	1608	7.2TB	cff plain
T2_ES_IFCA : storm.ifca.es	23977520	1608	7.2TB	cff plain
T0_CH_CERN : srm-cms.cern.ch	23977520	1608	7.2TB	cff plain
T2_IT_Bari : storm-se-01.ba.infn.it	22780230	1541	6.8TB	cff plain
T2_CN_Beijing : srm.ihep.ac.cn	21538024	1468	6.4TB	cff plain
T2_US_MIT : se01.cmsaf.mit.edu	23977520	1608	7.2TB	cff plain
T2_UK_London_IC : gfe02.grid.hep.ph.ic.ac.uk	22831760	1544	6.8TB	cff plain
T2_US_UCSD : bsrm-1.t2.ucsd.edu	23977520	1608	7.2TB	cff plain
T1_ES_PIC : srmcms.pic.es	23977520	1608	7.2TB	cff plain
T2_CH_CAF : caf.cern.ch	23977520	1608	7.2TB	cff plain
T3_FR_IPNL : lyogrid06.in2p3.fr	23977520	1608	7.2TB	cff plain
T2_US_Purdue : srm-dcache.rcac.purdue.edu	22831760	1544	6.8TB	cff plain
T2_IT_Legnano : t2-srm-02.lnl.infn.it	23977520	1608	7.2TB	cff plain
T1_IT_CNAF : storm-fe-cms.cr.cnaf.infn.it	23977520	1608	7.2TB	cff plain
T2_DE_RWTH : grid-srm.physik.rwth-aachen.de	23977520	1608	7.2TB	cff plain
T2_RU_JINR : lcgse01.jinr.ru	23977520	1608	7.2TB	cff plain
T2_DE_DESY : dcache-se-cms.desy.de	23977520	1608	7.2TB	cff plain
T2_ES_CIEMAT : srm.ciemat.es	23977520	1608	7.2TB	cff plain
T1_US_FNAL : cmssrm.fnal.gov	23977520	1608	7.2TB	cff plain
T2_US_Florida : srm.ihepa.ufl.edu	23977520	1608	7.2TB	cff plain
T3_CH_PSI : t3se01.psi.ch	23977520	1608	7.2TB	cff plain
T2_FR_IPHC : sbgse1.in2p3.fr	23977520	1608	7.2TB	cff plain
T2_US_Wisconsin : cmssrm.hep.wisc.edu	23977520	1608	7.2TB	cff plain
T2_US_Nebraska : srm.unl.edu	23977520	1608	7.2TB	cff plain

Data transfer

- The tool for data transfer is called **PhEDEx** (Physics Experiment Data Export)
- **Create subscription** requests on the **web interface**. Choose:
 - dataset, destination
 - transfer and subscription type...
- Request handled by site data managers (storage quotas)
- If approved - transfer starts (handled by the PhEDEx system)
- User can **monitor the transfer** performance using **web interface**



The screenshot shows the PhEDEx web interface for creating a new transfer request. The page title is "PhEDEx - CMS Data Transfers" and it features a navigation menu with tabs for "Info", "Activity", "Data", "Requests", "Components", and "Reports". Below the menu are links for "Overview", "Create Request", and "View/Manage Requests".

The main form is titled "New Transfer Request" and contains the following fields:

- E-mail:** An empty text input field.
- DBS:** A dropdown menu with the value "https://cmsdbsprod.cern.ch:8443/cms_dbs_prod_global_writer/servlet/DBSServlet".
- Data Items:** A text area containing the path "/MinimumBias/Commissioning10-Apr1Skim_GOODCOLL-v1/RAW-RECO". Below this are instructions: "/Primary/Processed/Tier" or "/Primary/Processed /Tier#Block" and "(Use * as wildcard)". A "More Help" link is provided.
- Destinations:** A list of checkboxes for various sites. The checked destination is "T3_CH_PSI". Other destinations include T0_CH_CERN_Export, T2_AT_Vienna, T3_CN_PKU, T0_CH_CERN_MSS, T2_BE_IJHE, T3_CO_Uniandes, T1_CH_CERN_MSS, T2_BE_UCL, T3_DE_Karlsruhe, T1_DE_KIT_MSS, T2_BR_SPRACE, T3_ES_Oviedo, T1_ES_PIC_MSS, T2_BR_UERJ, T3_FR_IPNL, T1_FR_CCIN2P3_MSS, and T2_CH_CAF.
- Transfer Type:** A dropdown menu set to "Replica" with a "What's this?" link.
- Subscription Type:** A dropdown menu set to "Growing" with a "What's this?" link.

CRAB jobs

- CMS uses a python front-end tool called the **CMS Remote Analysis Builder (CRAB)** for job creation and submission in a **user-friendly way**:
- User only needs to specify some **minimal information**:
 - configuration of the analysis
 - dataset name
 - number of events/jobs
 - output handling
- The rest is handled **by CRAB**
 - splitting of jobs
 - finding appropriate sites
 - submission
 - basic monitoring
 - output retrieval
- The **CRAB** is operated as a **command line tool**

CRAB configuration file

```
[CMSSW]
### The data to access (defined in multicrab cfg)
#datasetpath=

### The parameter set to use and additional options
pset                = ../../ntupleproducer_cfg.py
pycfg_params        = runon=data recoType=RECO

### Splitting parameters (defined in multicrab cfg)
total_number_of_events = -1
events_per_job         = 20000
number_of_jobs         = 300
total_number_of_lumis = -1
lumis_per_job          = 30
split_by_run           = 1

### run number restrictions (defined in multicrab cfg)
runselection=132440-135735
lumi_mask=Cert_132440-135735_7TeV_StreamExpress_Collisions10_JSON.txt

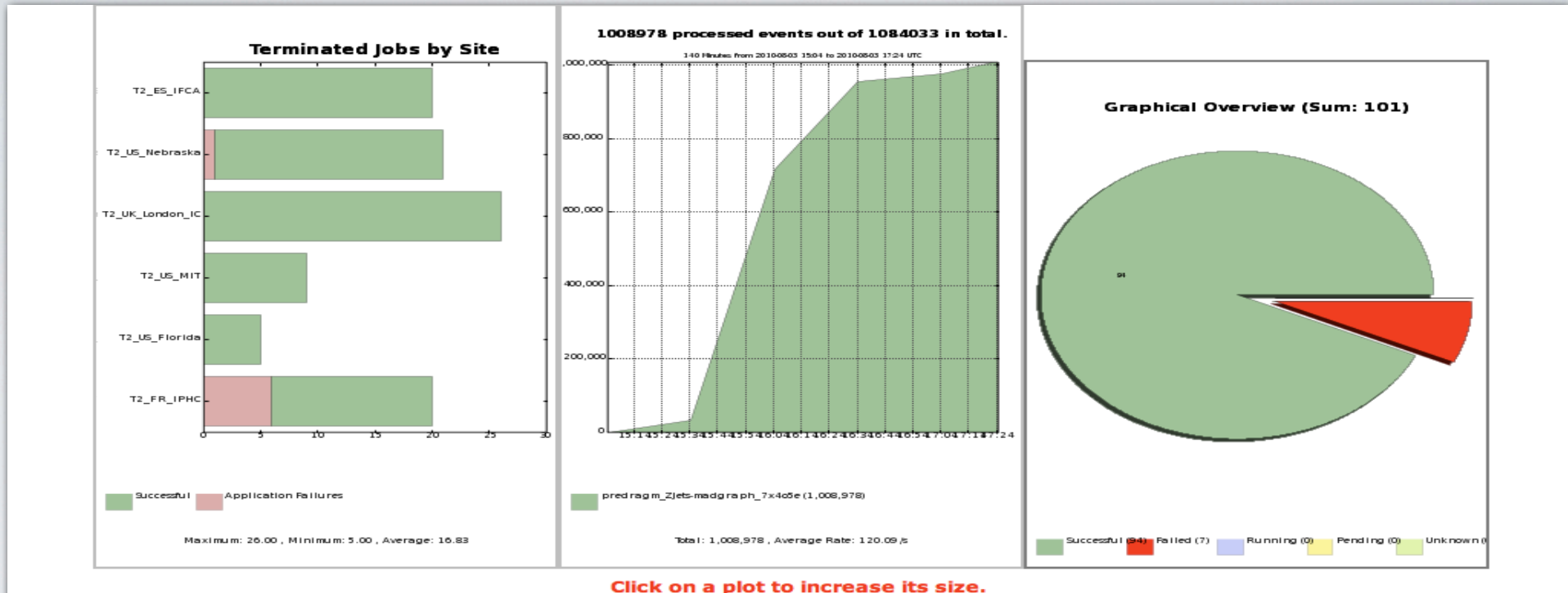
### The output files (defined in ntupleproducer_cfg)
#output_file =

[USER]
### output back into UI
return_data        = 0
#ui_working_dir    = data

### output files into a SE
copy_data          = 1
storage_element    = t3se01.psi.ch
storage_path       = /srm/managerv2?SFN=/pnfs/psi.ch/cms/trivcat/
user_remote_dir    = ntuples/data/
```

Job monitoring - Dashboard

- General overview of all running tasks and a detailed overview of jobs provided also via the CMS Dashboard

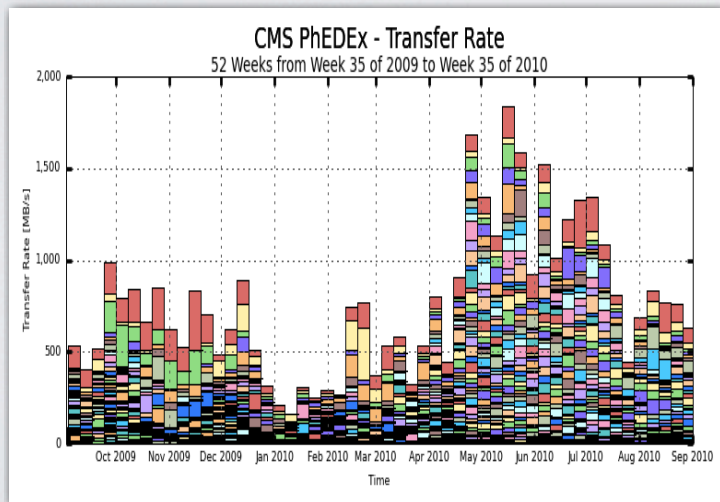
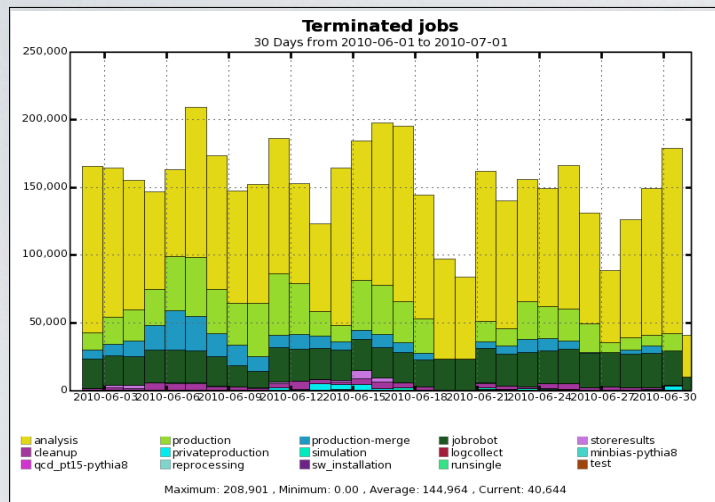


Click on a plot to increase its size.

SchedulerJobId	Id in Task	Appl Status	Appl Exit Code	Grid End Status	Retries	Site	Submitted	Started	Finished
https://prod-lb-01.ct.infn.it:9000/ckr7LPZ7Me1TjTxNrGa5Eg	1	Appl Succeeded	0	Unknown	1	T2_US_Nebraska	2010-08-03 16:20:47	2010-08-03 16:20:47	2010-08-03 16:53:59
https://prod-lb-01.ct.infn.it:9000/ad5pAJStBe5FvEi5VKWsKg	2	Appl Succeeded	0	Unknown	1	T2_US_MIT	2010-08-03 16:20:21	2010-08-03 16:20:21	2010-08-03 16:41:07
https://prod-lb-01.ct.infn.it:9000/HULJiswijgBijhJUI-lbaw	3	Appl Succeeded	0	Unknown	1	T2_UK_London_IC	2010-08-03 16:58:46	2010-08-03 16:58:46	2010-08-03 17:18:42
https://prod-lb-01.ct.infn.it:9000/_V1Z_P60aEC8wVIxgx7V_g	4	Appl Succeeded	0	Unknown	1	T2_ES_IFCA	2010-08-03 16:18:24	2010-08-03 16:18:24	2010-08-03 16:46:48

Analysis activities

Routinely delivering 100k jobs per day



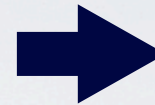
Hundreds of active users running analysis



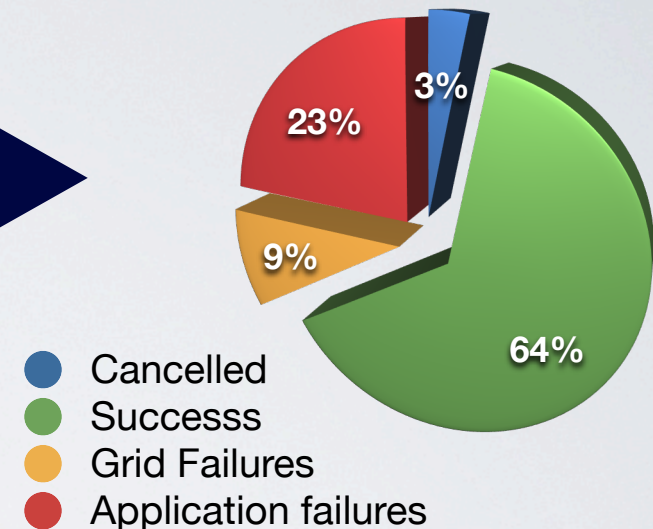
General user experience

- The sources of analysis failures:

- Grid-related problems
(not easy to control or have influence on)
- CMS user-related errors and problems:
 - in **configuration files**
 - in **analysis code**
 - issues with **stage-out of job outputs**(good support from the CRAB team)



Analysis Jobs (2008-2009)



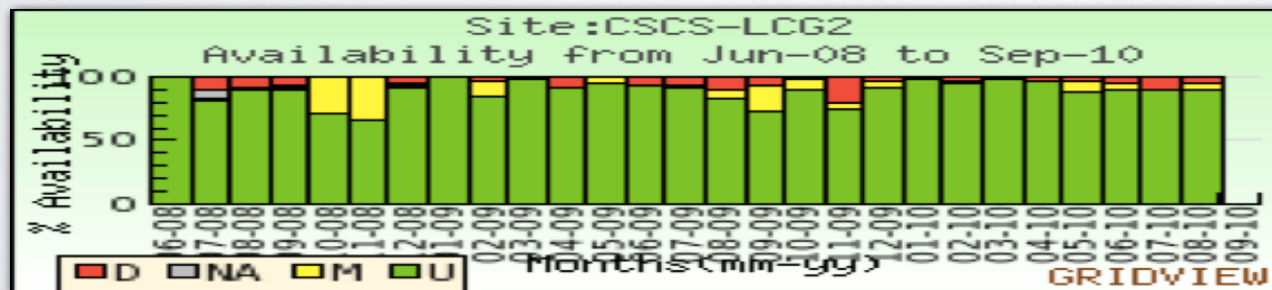
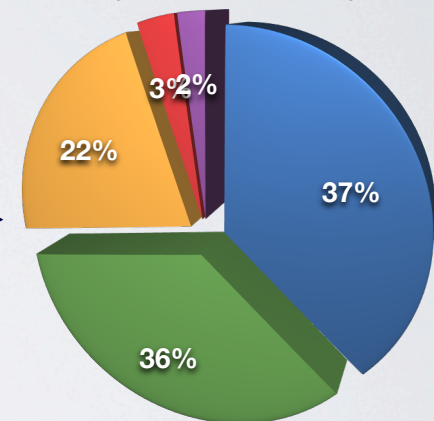
- Users would like to have improved:

- the total time needed for 100% of jobs to be completed
 - ➔ e.g. we can have **90% of jobs finished within ~2-3 hours**
and **the other 10% of jobs** could be on the Grid for **up to 20-30 hours!**
(various reasons could lead to this issue - not easy to tackle this problem)
- better synchronization of the monitoring information provided by our two tools: the CRAB and the Dashboard

Swiss CMS Tier-2

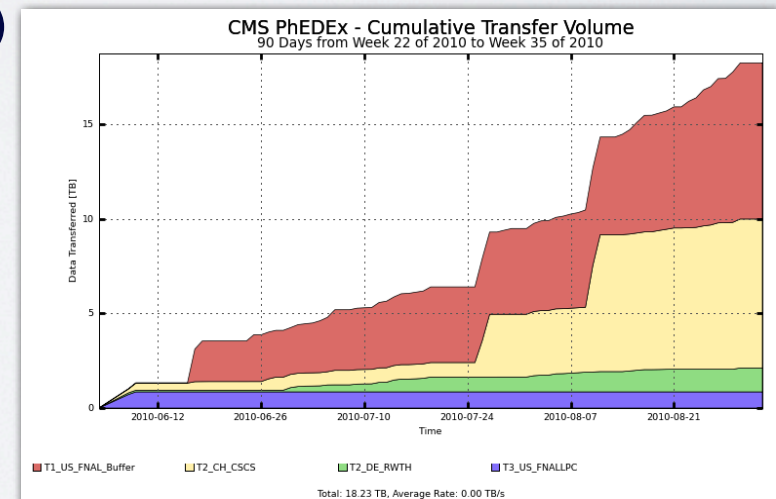
- **CSCS at Manno** hosts a **CMS Tier-2** which offers a total of **768 cores** and about **800 TB** of storage space (reached this upgrade phase in Q1/2010)
- Network traffic via **two redundant lines of 10Gbps** (to CERN and Europe)
- Tier-2 is up & has been **in stable operation for years!**
- Shares of resources per VO are **pretty well balanced**. (spare cycles given to other VOs: H1, theory...)
- Typical reliability and availability in last two years are **higher than 95%**

CPU share per VO
(2007-2010)



Swiss CMS Tier-3

- **The PSI hosts a CMS Tier-3** which offers ~250 TB of storage space and 28 nodes with 224 cores. It is available to all CMS users from Swiss institutes (ETH Zurich, University of Zurich and PSI)
- Datasets can be transferred via PhEDEx from some Tier centers directly to the storage element (SE) and from the others via the Swiss Tier-2.
- Analysis jobs are typically submitted and **run as CRAB jobs**.
 - if the output is small, it is retrieved directly on the PSI SE (or at the UI)
 - otherwise, it is preferred to use CSCS SE for stage-out (and transfer data to PSI SE with tools provided by our T3 team)
- The analysis/simulation jobs can be submitted directly to the local batch farm on worker nodes of the PSI Tier-3
- It has become **The Tool** for members of swiss institutes (CMS)



Conclusions

- The CMS has entered the data analysis era in a full swing!
- All the challenges we had in CMS in recent years have led to many improvements and resulted in a **reliable system** for analysis
 - exercised in many computing aspects...
(DC04, CS06, CSA07, CSA08, STEP09, OctoberExercise ...)
 - ...and in most of the analysis aspects (dataset skimming, analysis at T2s etc.)
- The complete infrastructure has sustained an enormous load and **successfully supported** the analysis “marathon” prior to the ICHEP 2010
- The Swiss Tier-2 and Tier-3 centers are **extensively used** and play an essential role in our everyday-analysis life
- The main building blocks for data analysis are fully in place and working in a production environment (and **being debugged and regularly updated to improve the user experience**)