The October Exercise

Site Perspective & Computing View

Christoph Wissing

CMS Hamburg Meeting

November 18th, 2009

Contents:

- Introduction
- Experience with jobs
- Data transfers
- Summary

Computing Model

Simplified view:

Central Processing

Tier-0 RAW → RECO

Tier-1 RECO → AOD

Primary Datasets (PD)
- based on trigger info
- archived at T1s

Secondary Datasets (SD)
Central Skims (CS)
- skim of a PD (~10%)
- ship to T2s

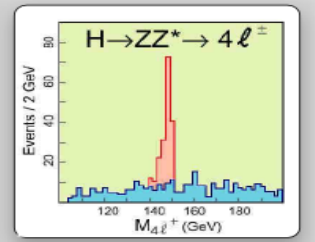
User Land

Tier-2 AOD → AOD + User

Production of Group Skims (GS)
- done at T2s

Final user analysis
- either T2 or T3

fast processing and FWLite at Tier-3



Note: - T3 not strictly defined
- "Anything beyond T2"

http://indico.cern.ch/getFile.py/access?contribId=21&sessionId=0&resId=7&materialId=slides&confId=71193

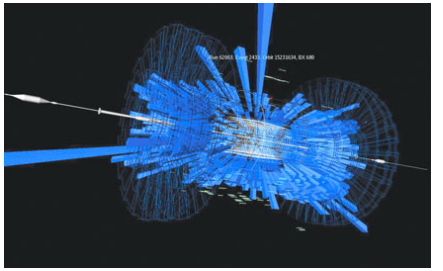
Time Line for Computing (Simplified)

STEP09

Some Holidays

???

Some Gap



Start of Beams

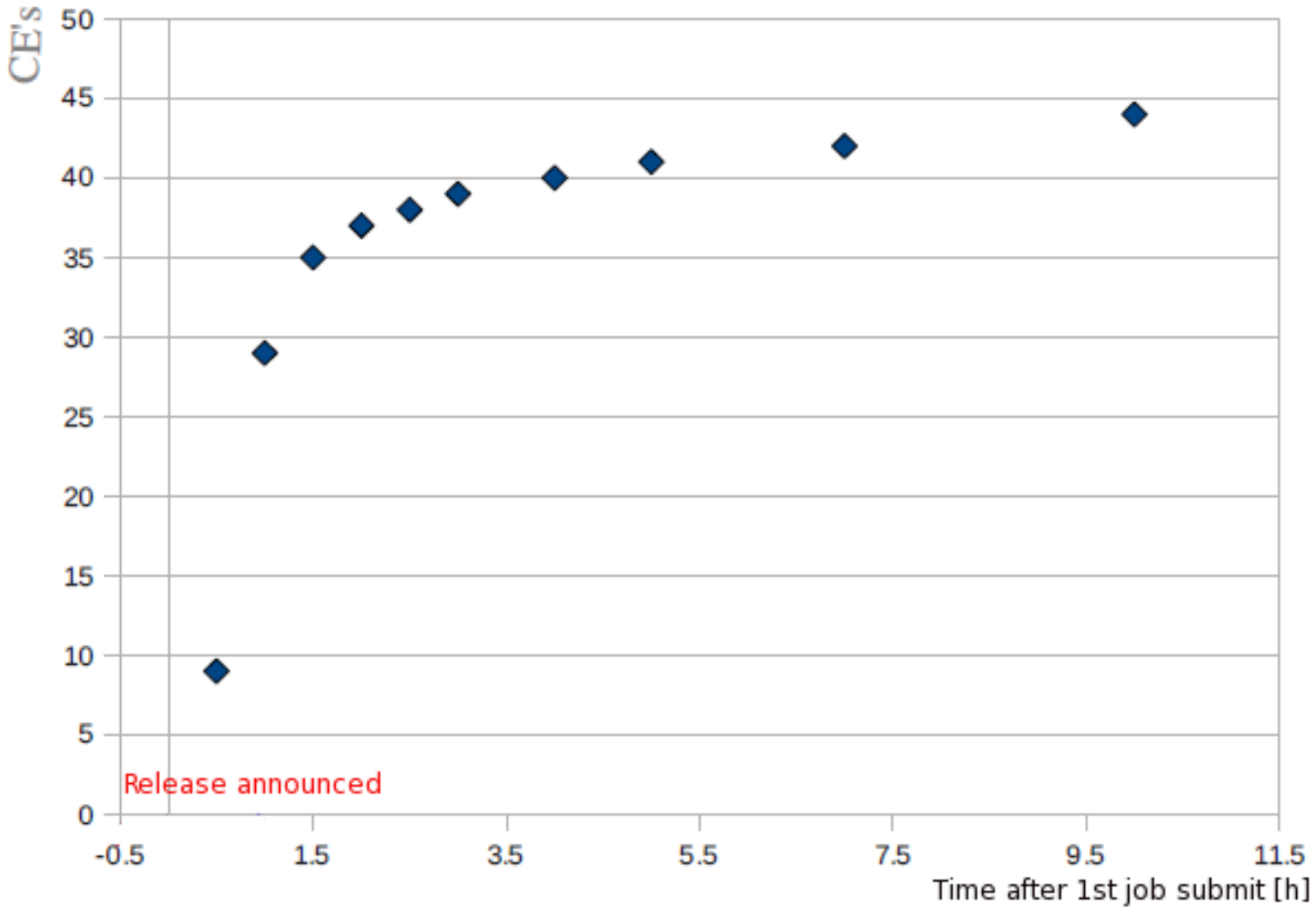
June July August September October November

- **October Exercise (OctX)**
  - **A bit more then a gap filler**
- **Full analysis exercise on the distributed infrastructure**
  - **Mainly run by physics groups**
- **Final test before (expected) beam operations**
- **Still time to fix urgent matters before it gets really serious**

October exercise: Main Goals

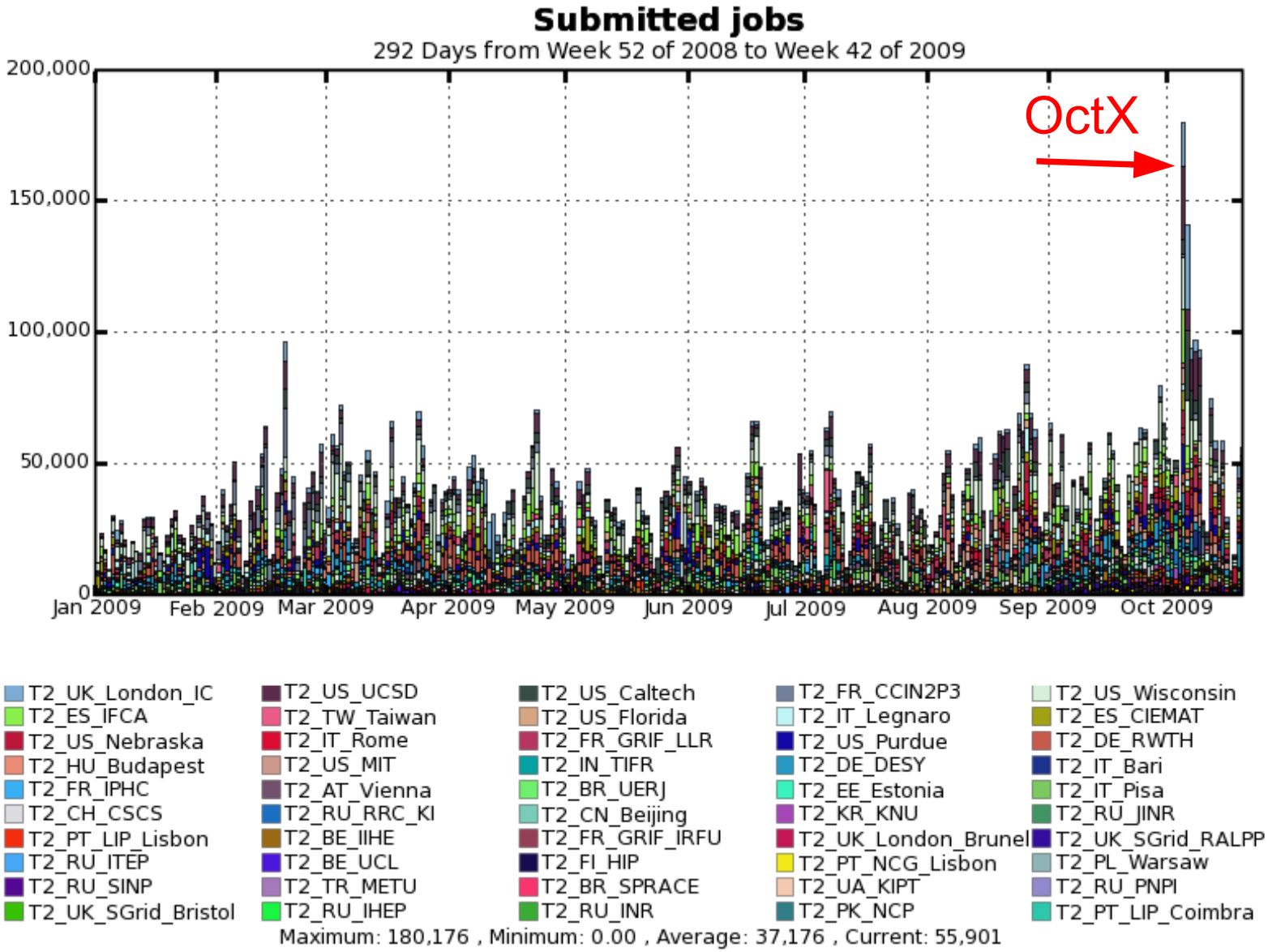
- Deploy the Tier-2 analysis model
  - Train people to do all of the tasks that are needed to enable individuals, groups, and CMS as a whole to efficiently access and analyze data.
- Scale-test (October 5th -19th)
  - Approximate the situation we will face with early data with all groups doing key analyses simultaneously
    - T1-T2 transfers of (pseudo) Secondary Datasets
    - Widespread use of CRAB server (very important to address issues)
    - Group Skims of SDs by ~2 high priority users from each group
    - T2-T2 transfers by some groups
    - T1-T2 subscriptions of MC samples by all groups
    - Analysis jobs exercising early analyses by as many people as possible
  - See if we can correct problems, overcome obstacles, eliminate bottlenecks during an intense two week period…
- Post-mortem
  - Review what has been learned and iterate ahead of data-taking
  - Continue to work on improving the system continuously

Deployment CMSSW_3_1_4



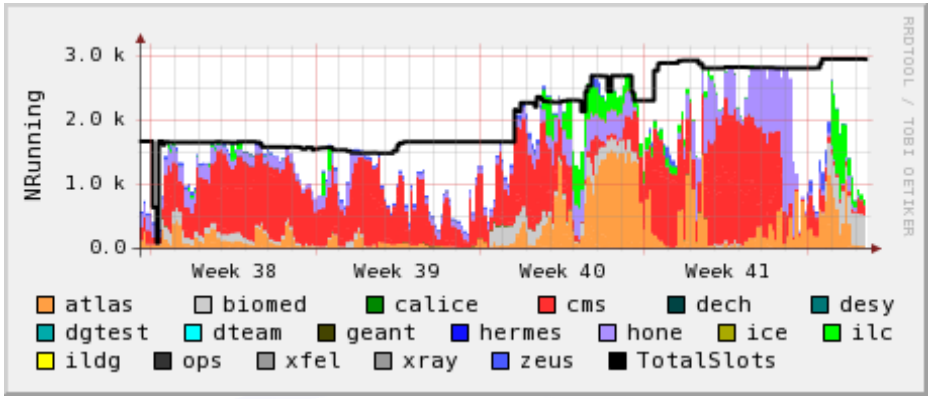
Improvement of deployment tools ongoing (Wolf Behrenhoff)

Job Activity in 2009



Jobs At DESY during OctX

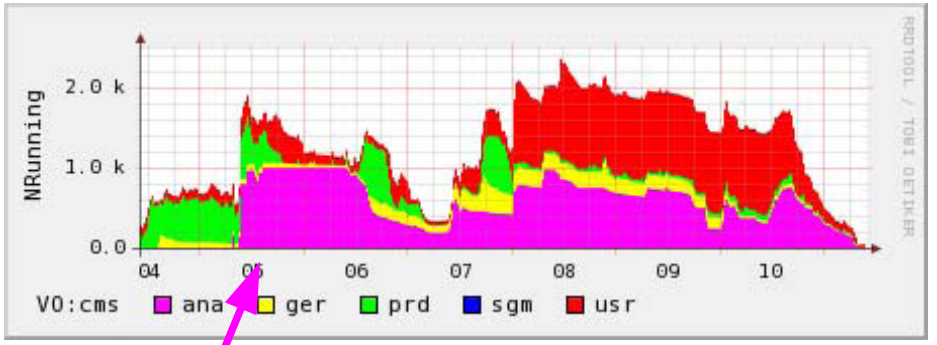
All VOs:



- Quite some CPU deployed
  - 2000 cores to 3000 cores
  - Sufficient to serve all 2010 pledges
- Many VOs active

- Various CMS job types:
  - OctX (priority + some user)
  - Production
  - Usual user analysis
  - German users
- Far beyond CMS share
  - CMS targert: ~ 1/3

CMS VO:



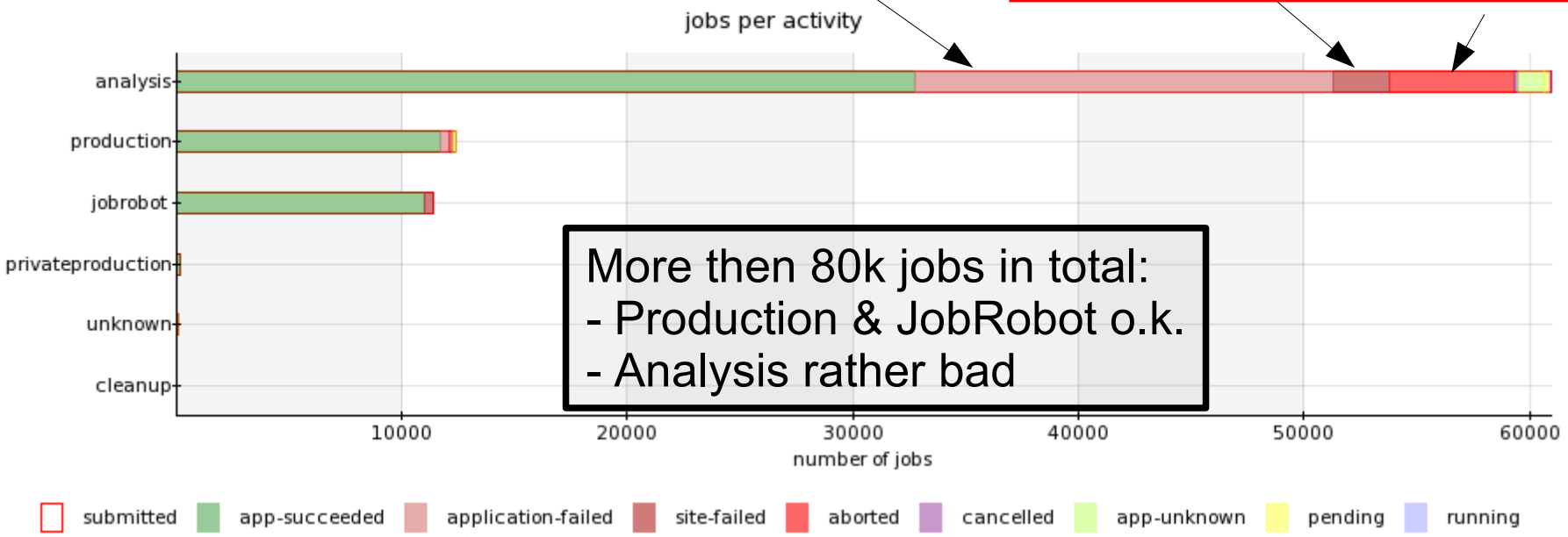
"priority user":
Production user for physics groups

Jobs At DESY during OctX

Quite some site failures
- Scalability issues of storage
- Improved during the exercise

Lot's of application failures
- Problems with remote stage-out
- “User errors”

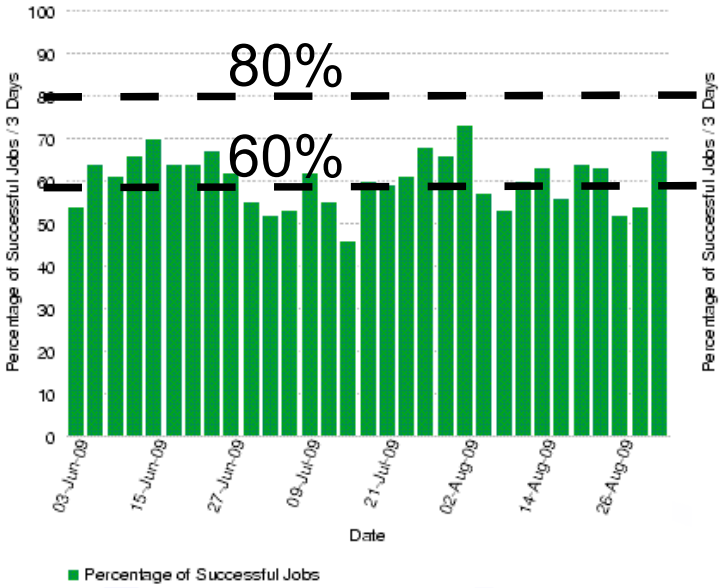
“Aborted in Grid middle-ware
- Might be a site issue
- Might be WMS related



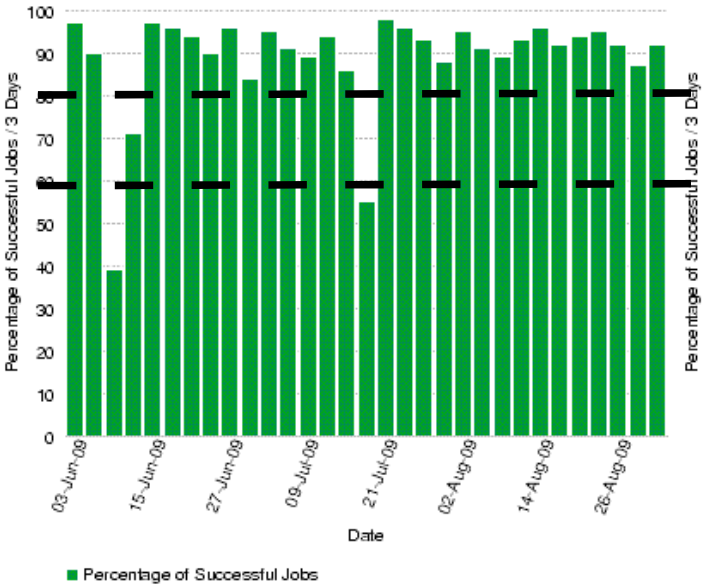
More then 80k jobs in total:
- Production & JobRobot o.k.
- Analysis rather bad

Many sites got hit badly at the start of the exercise

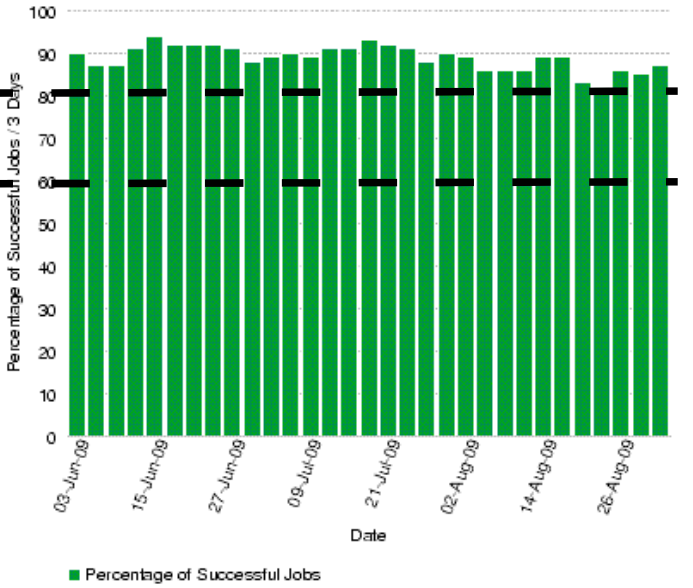
Global Job Success Rate



Analysis



MC Production

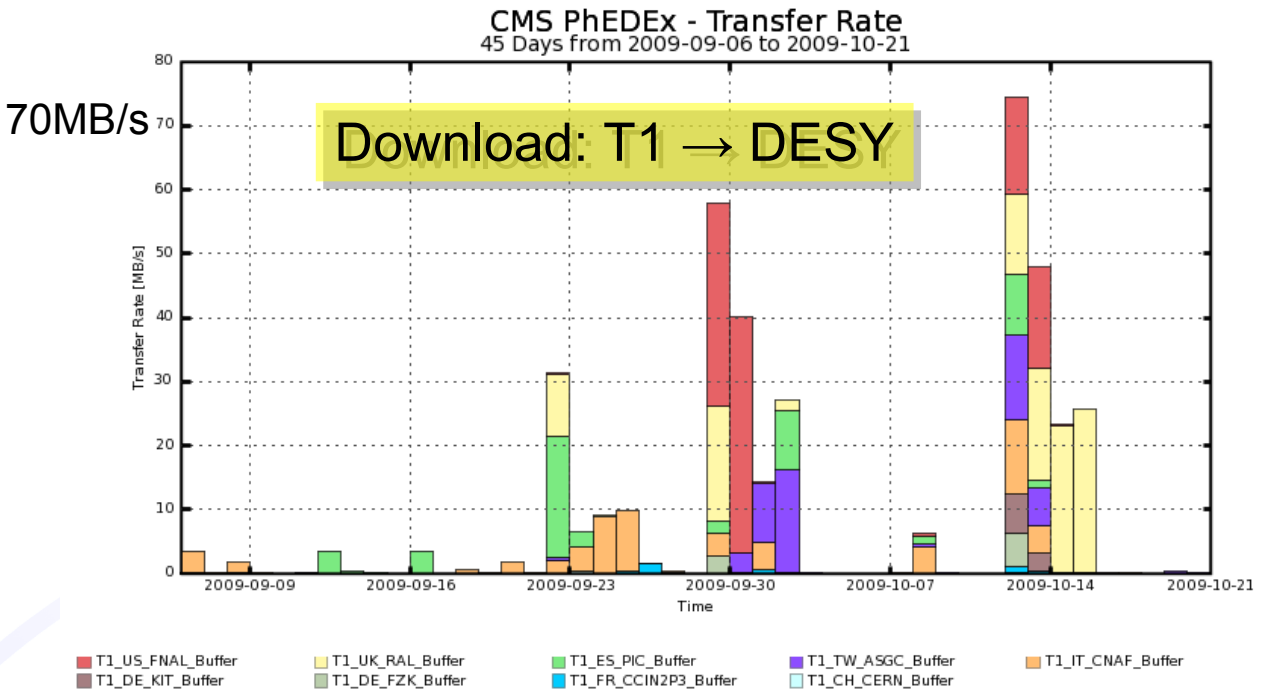


JobRobot

- **Global observation:**
  - **Acceptable success rates for production & JobRobot**
  - **Bad success rates for analysis/user jobs**
- **Analysis Operations group**
  - **Understand the reasons**
  - **Teach users to improve their efficiency**
  - **Identify weaknesses in the global approaches**

Data Transfers

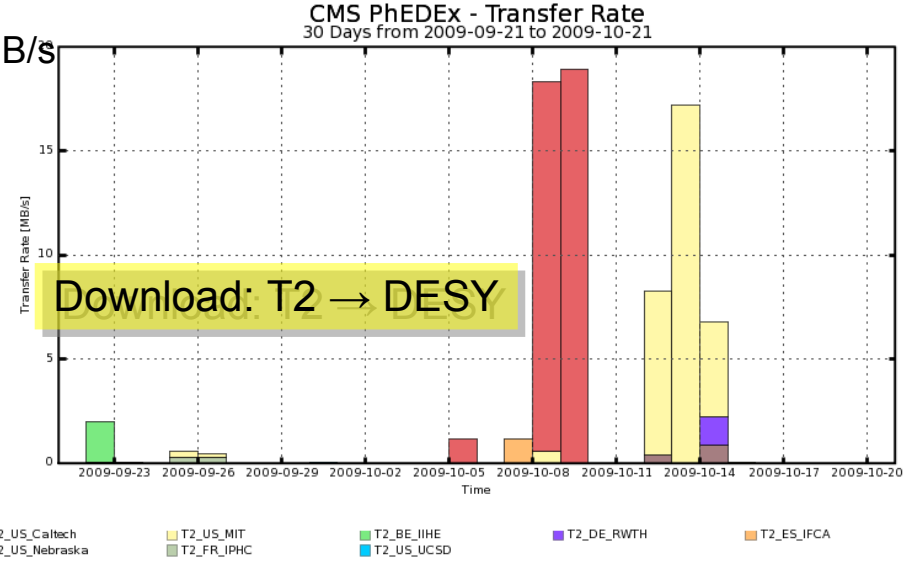
70MB/s



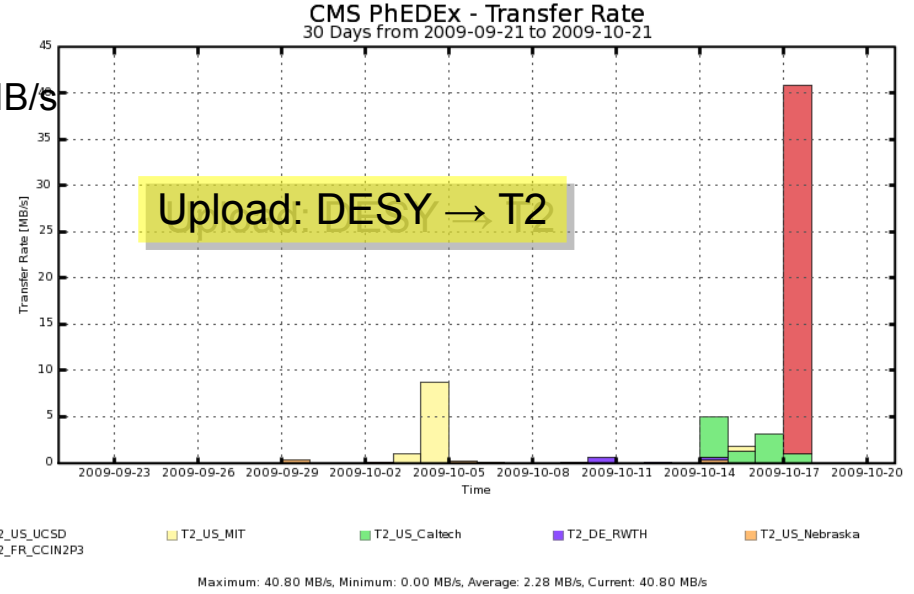
Routine Operation
- No attempt to push limits
- Good quality

T2-T2 moderate

20MB/s



40MB/s



T2-T2 Links for the "DESY" Groups

Forward

From Node ↓ To Node →	T2_BR_UERJ	T2_DE_DESY	T2_US_Wisconsin
T2_BR_UERJ			
T2_DE_DESY			
T2_US_Wisconsin			

Good T2-T2 connectivity at DESY

JetMET (not seriously started T2-T2 links)

From Node ↓ To Node →	T2_DE_DESY	T2_FI_HIP	T2_KR_KNU	T2_RU_ITEP	T2_US_Purdue
T2_DE_DESY					
T2_FI_HIP					
T2_KR_KNU					
T2_RU_ITEP					
T2_US_Purdue					

QCD

From Node ↓ To Node →	T2_DE_DESY	T2_FR_CCIN2P3	T2_US_Caltech	T2_US_MIT
T2_DE_DESY				
T2_FR_CCIN2P3				
T2_US_Caltech				
T2_US_MIT				

Top

From Node ↓ To Node →	T2_BE_IIHE	T2_DE_DESY	T2_ES_IFCA	T2_FR_IPHC	T2_US_Nebraska	T2_US_UCSD
T2_BE_IIHE						
T2_DE_DESY						
T2_ES_IFCA						
T2_FR_IPHC						
T2_US_Nebraska						
T2_US_UCSD						

Tier-2 Storage Organization

`/store/data/...`

`/store/mc/...`

Registered in **Global DBS**

Management & accounting via Phedex

Only writable by privileged production users

New transfer service
- Approval needed
- Merging of small files
- Used in OctX for the 1st time

”Home Tier2”:

`/store/user/<HyperNewsName>`

“Foreign Tier2”:

`/store/temp/user/<HyperNewsName>`

Files registered in **Local DBS**

`/store/group/<PAG>`

Registered in Local **DBS**

Writable by privileged priority user
(= 2 special users per PAG)

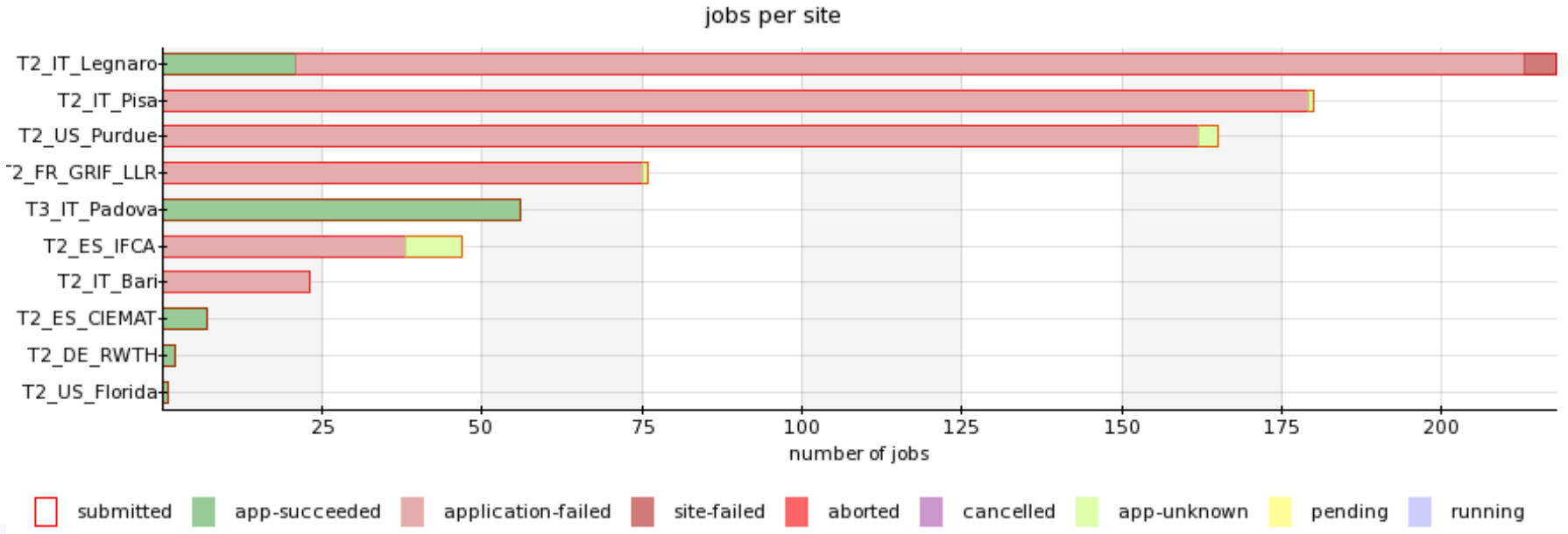
`/store/results/<PAG>`

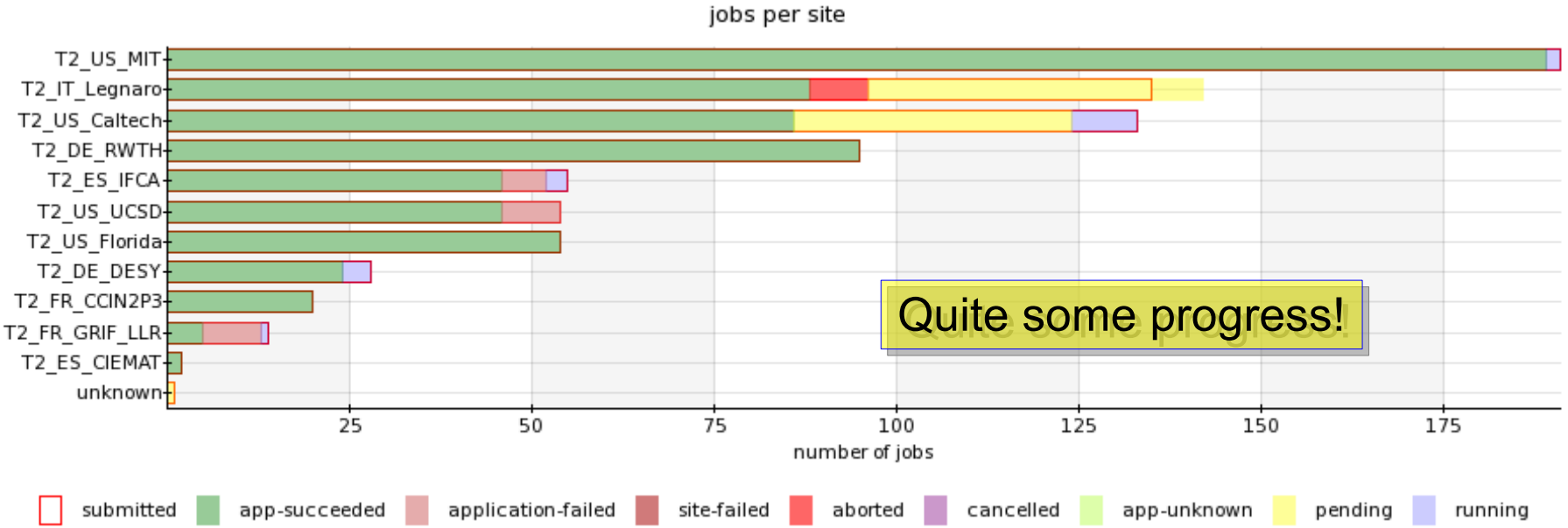
Registered in **Global DBS**

Injected in Phedex

Only writable by privileged production user

StoreResult Service 1st 2 days – last 2 days





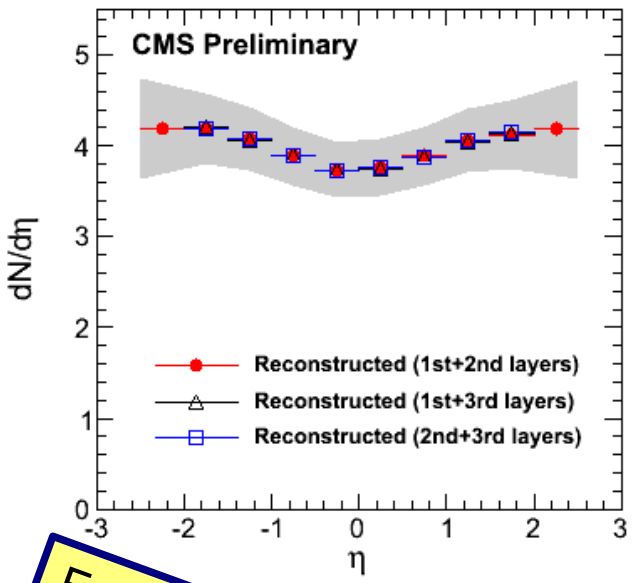
Quite some progress!

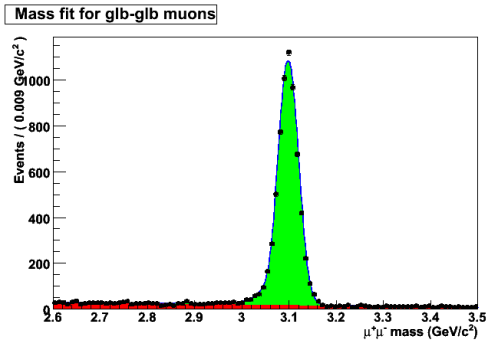
Statistics & a bit of PostMortem

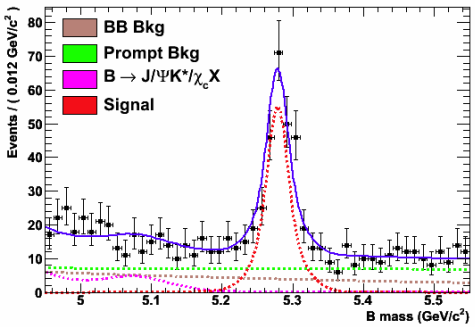
- 77 exercise in over 500 steps were planned
  - 45 finished 100% during OctX
  - Another 7 almost finished (99% done)
  - 80% of all steps done
- 230 people actively participated
- 2000 data sets subscribed to Tier-2s before OctX
  - ~900TB transferred
  - Quite some effort to prepare secondary datasets [DataOps in “hero mode” some weeks before the exercise]
- Many development items pending
  - Improve on CRAB stageout component
    - Big source of trouble in OctX
  - WMAgent should solve many things
    - Still a development project

Introduced CRAB client that checks against common user errors before job submission.

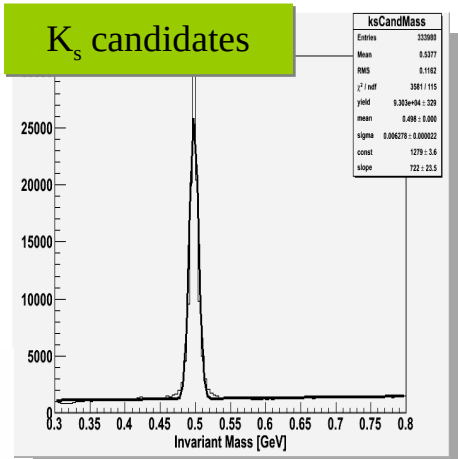
Physics



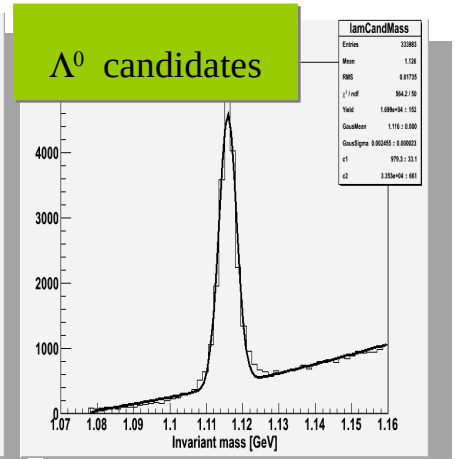


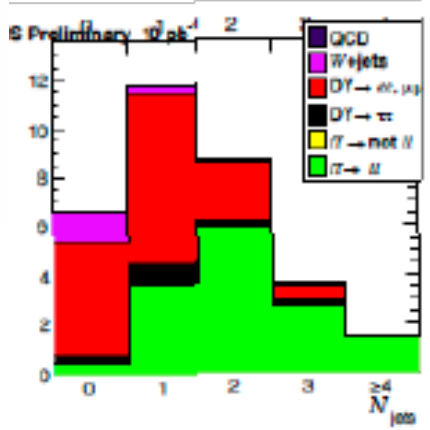


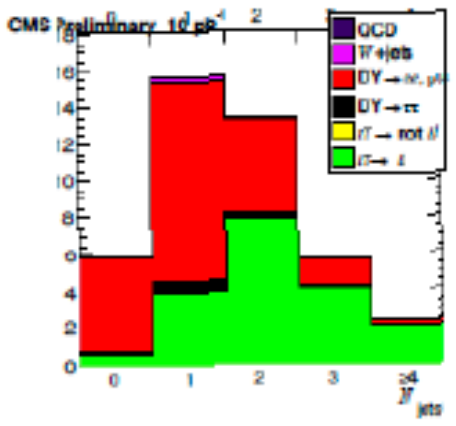
K_s candidates

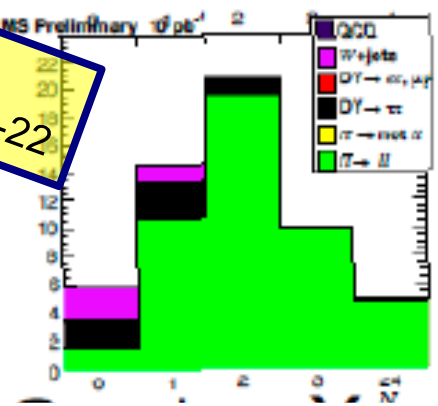


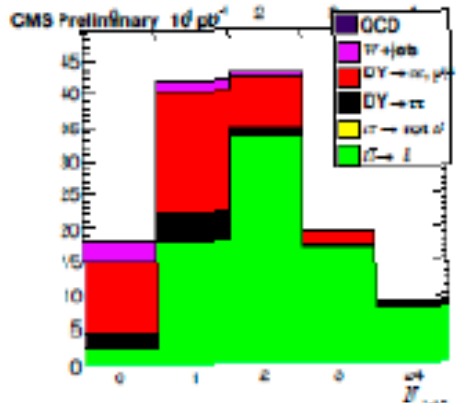
Λ^0 candidates

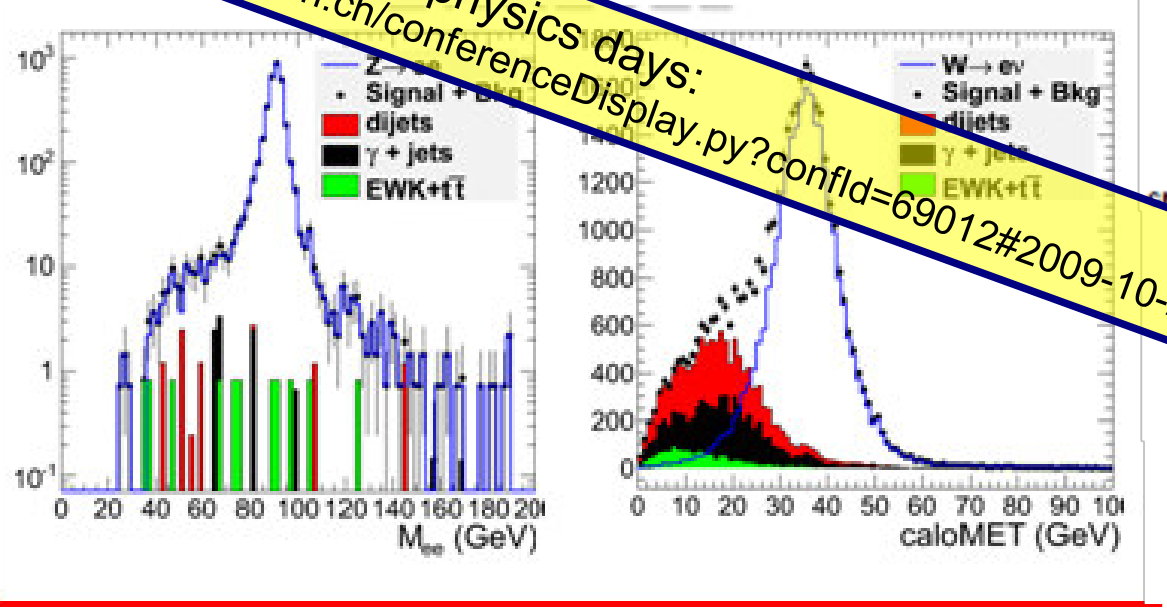












Full review during physics days:
http://indico.cern.ch/conferenceDisplay.py?confId=69012#2009-10-22

Summary (Ian's Outlook)

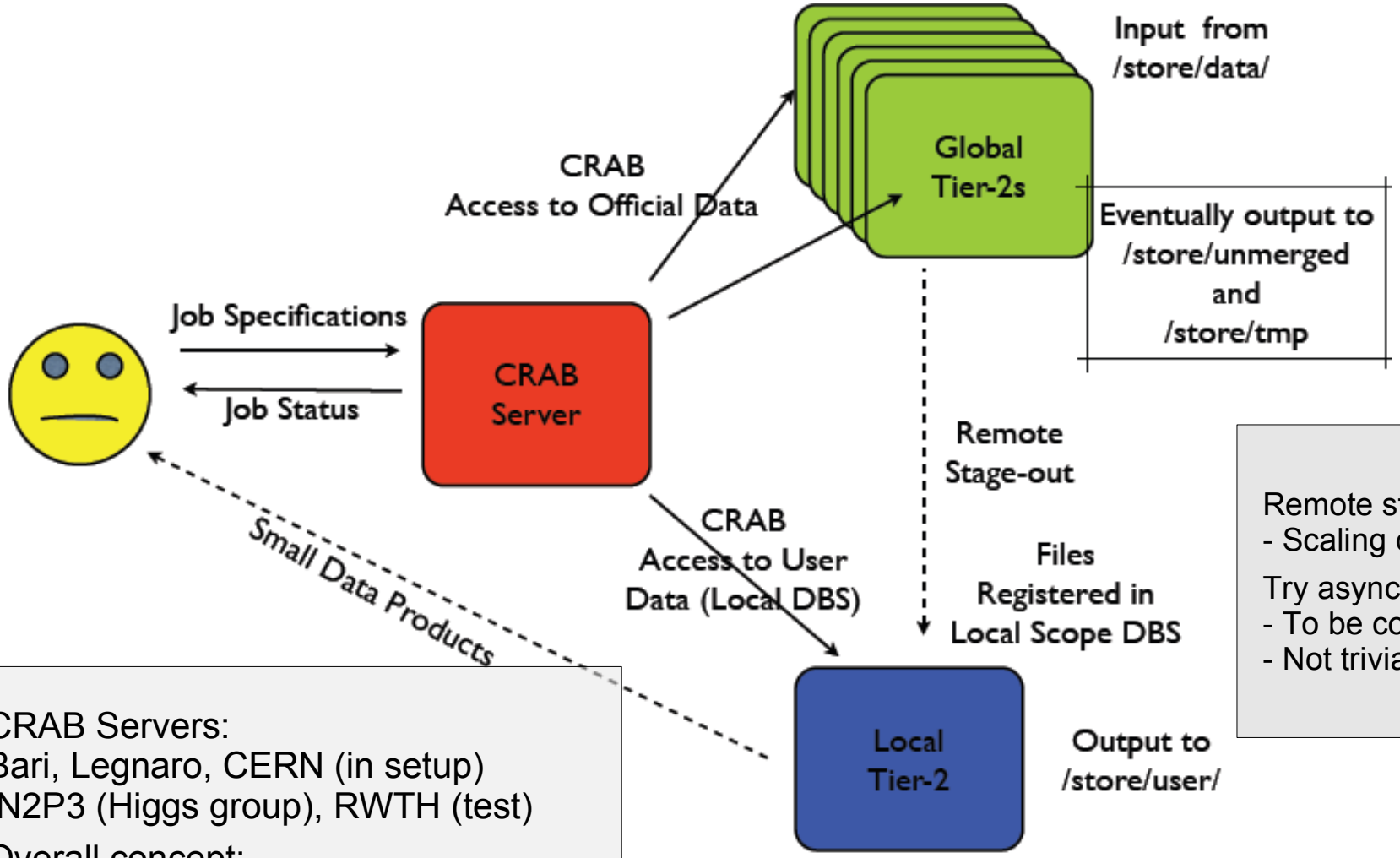
Outlook

- The Computing Project believes we are ready for the start of collision data
  - There are open issues, some of which can be addressed in the final weeks and some will need to be worked around
    - We need an agreement of the total expected data rate and an understanding of the expected rate per PDS (and an understanding of the trade-offs for some scenarios)
    - We believe we can analyze data, but important issues were identified in the Oct Ex, not all of which can be fixed in the short term
    - The Tier-1s, Tier-2s, and a growing number of Tier-3s are ready for running
      - We still see stability issues, but redundancy in the early run will alleviate some issues
    - Good team of CAT-A, CSP, CRC, and Analysis Ops Shifters though more effort will be needed for long running.

11/11/09 Computing Readiness Review 29

Backup Slides

CRAB Server



Remote stage-out critical
- Scaling of remote SE

Try asynchronous mode
- To be commissioned
- Not trivial to implement

CRAB Servers:
Bari, Legnaro, CERN (in setup)
IN2P3 (Higgs group), RWTH (test)

Overall concept:
- How many servers needed?
- Who runs/supports them?
- User distribution (region vs. group)?