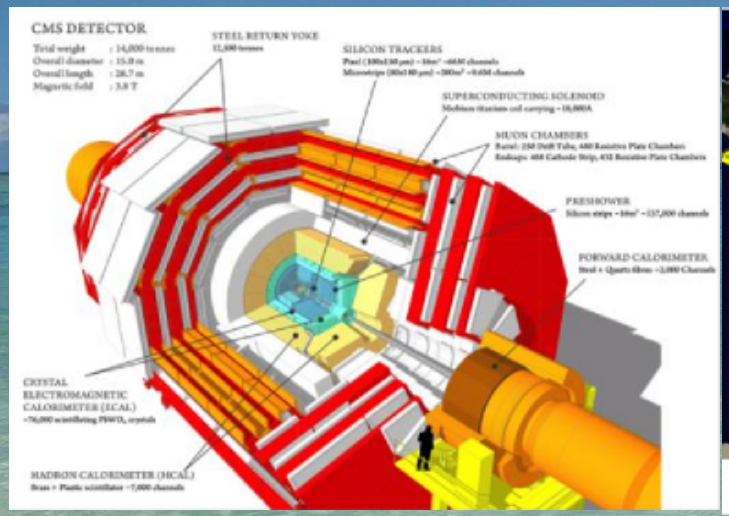
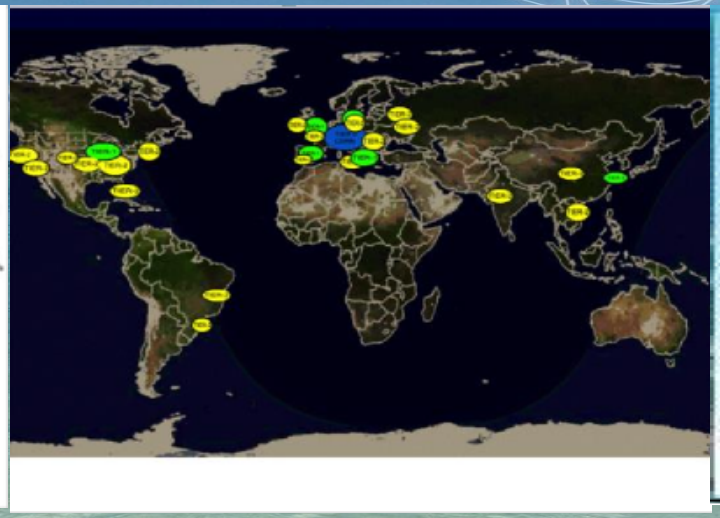
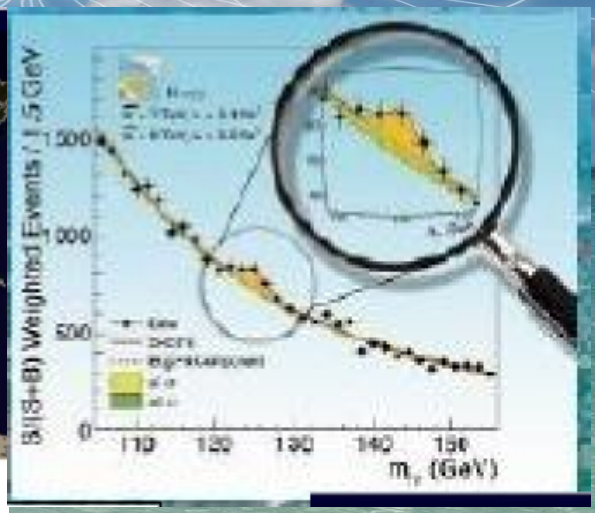
WLCG Workshop

Okinawa, Japan







CMS Report

Christoph Wissing (DESY)
for CMS Computing & Offline

April 11th, 2015

HELMHOLTZ ASSOCIATION

DESY

Some Pre-Remarks

- There will be a complementing presentation during CHEP by Ian Fisk - Improvements in the CMS Computing System for Run2
- Many subjects will only briefly be touched in this talk
  - Here are some advertisements for more detailed CHEP presentations:

CHEP talk: D.Lange -Simulation and Reconstruction Upgrades for the CMS experiment

CHEP talk: C.Jones - Using the CMS Threaded Framework In A Production Environment

CHEP talk: A.Perez-Calero Yzquierdo -Evolution of CMS workload management towards multicore job support

CHEP talk: J.Letts - Using the glideinWMS System as a Common Resource Provisioning Layer in CMS

CHEP talk: D. Hufnagel - The CMS Tier-0 goes Cloud and Grid for LHC Run 2

CHEP Poster: J.Letts -
Using HTCondor and glideinWMS to 200K+ Jobs in a Global Pool for CMSbefore LHC Run 2

CHEP talk: D.Colling - The diverse use of clouds by CMS

CHEP talk: D.Hufnagel - Enabling opportunistic resources for CMS Computing Operations

CHEP Poster: M.Zvada - MS Experience with a World-Wide Data Federation

CHEP talk: C.Paus - Dynamic Data Management for the Distributed CMS Computing System

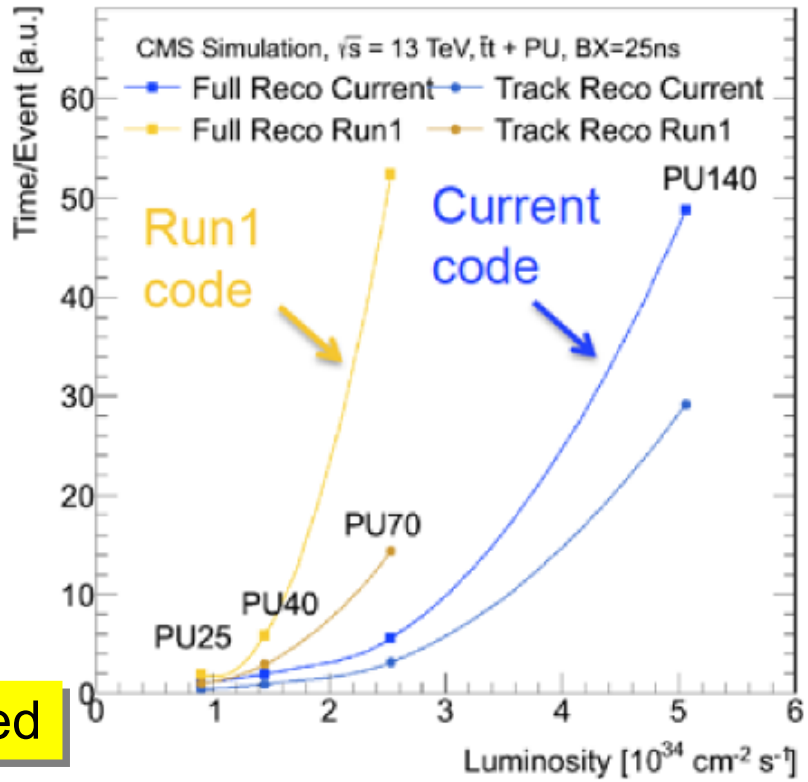
CHEP talk: Ch. Wissing - Pooling the resources of the CMS Tier-1 sites

CHEP talk: C. Vuosalo -A new analysis data format for CMS

CHEP talk: M. Mascheroni- CMS Distributed Data Analysis with CRAB3

Computing Challenges during Run2 Conditions

- New beam conditions
  - Increased center of mass energy: 8TeV → ~13TeV
  - Increased luminosity: ~0.7*1034cm-2s-1 → ~1.5*1034cm-2s-1
  - Higher pile-up rate
- Increased data logging rate from ~400Hz to 1kHz
  - Keep trigger thresholds close to Run1
- Increased event complexity
  - Higher Memory requirements
  - More CPU demanding
- Big effort to improve software performance



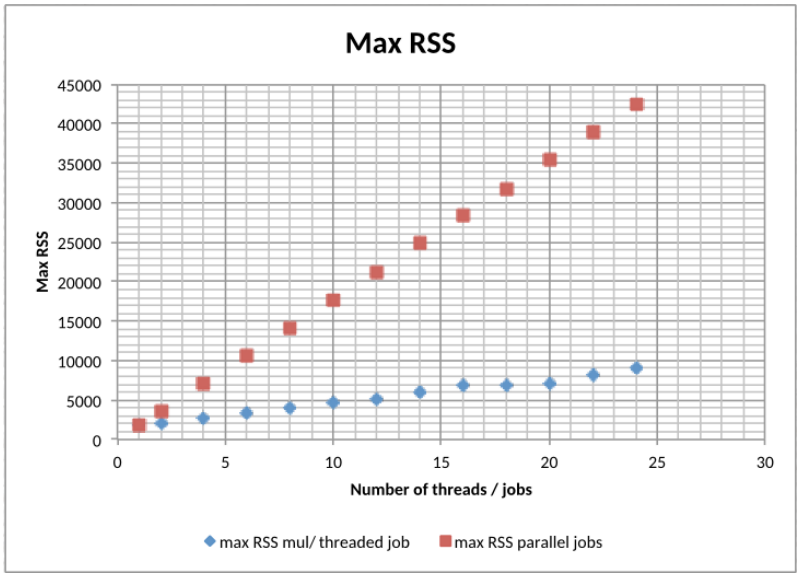
Computing during Run2 will be resource constrained

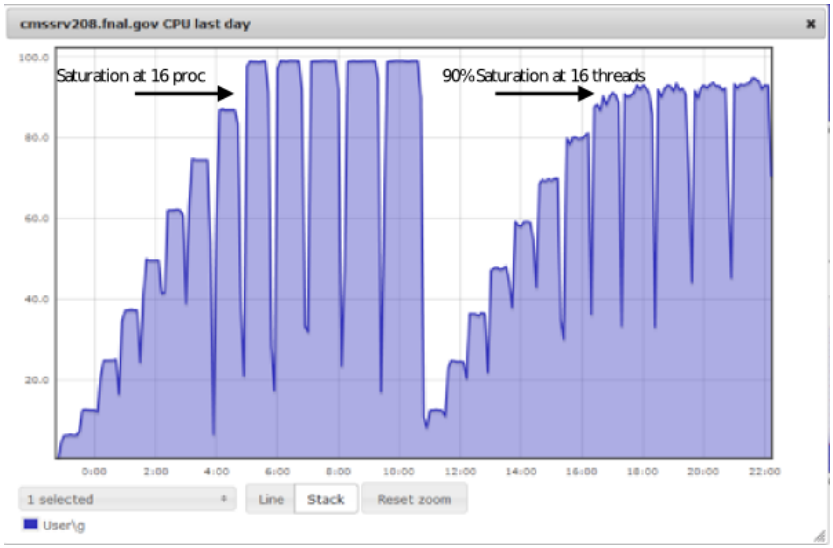
CHEP talk: D.Lange -Simulation and Reconstruction Upgrades for the CMS experiment

CHEP talk: C.Jones - Using the CMS Threaded Framework In A Production Environment

Multi-core Processing

- Need for multi-core Jobs
  - With Run2 conditions RECO for one “Lumi Section” will not fit into usual 48-hour batch queue
    - Cannot split easily beyond Lumi section boundaries
  - Large potential to save memory pro CPU core
  - Less pilots to be sent –
    should increase scaling capability of infrastructure
- Efficiency of multi-core applications
  - Big fraction of code needs to be thread-safe [Amdahl's law]
  - Achieved good CPU efficiency in recent software releases
- Mixture of multi-core & single-core jobs
  - Partitionable pilots
    - 8 core pilot can execute e.g. 1x4+1x2+2x1 core-jobs

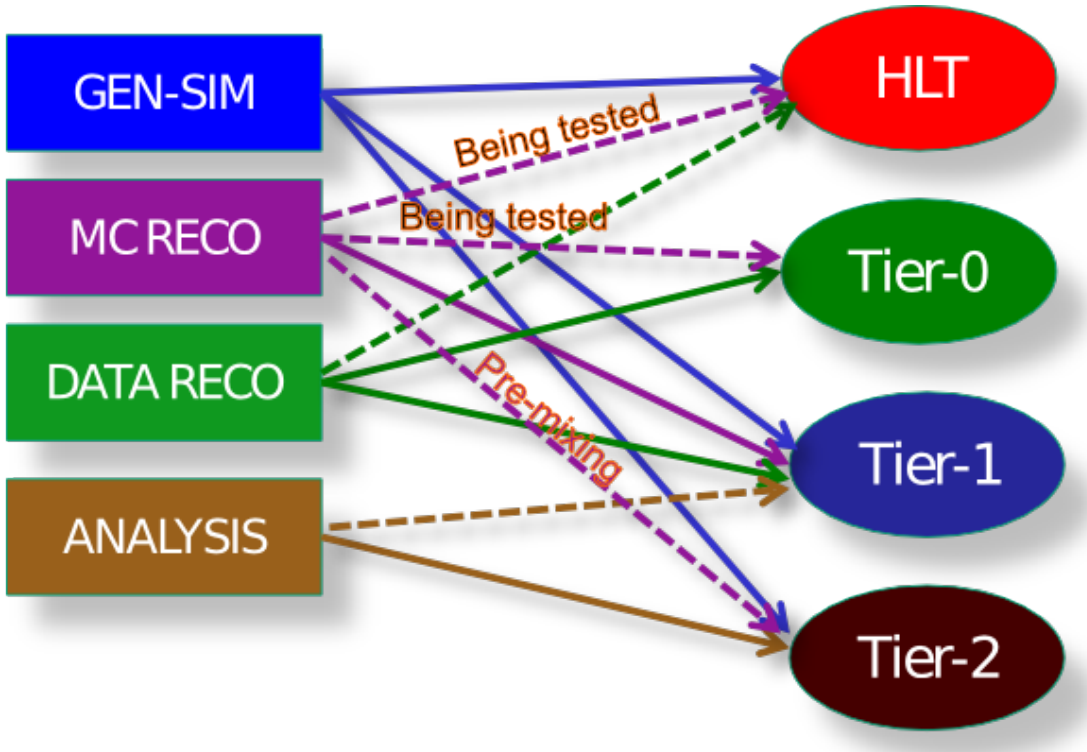




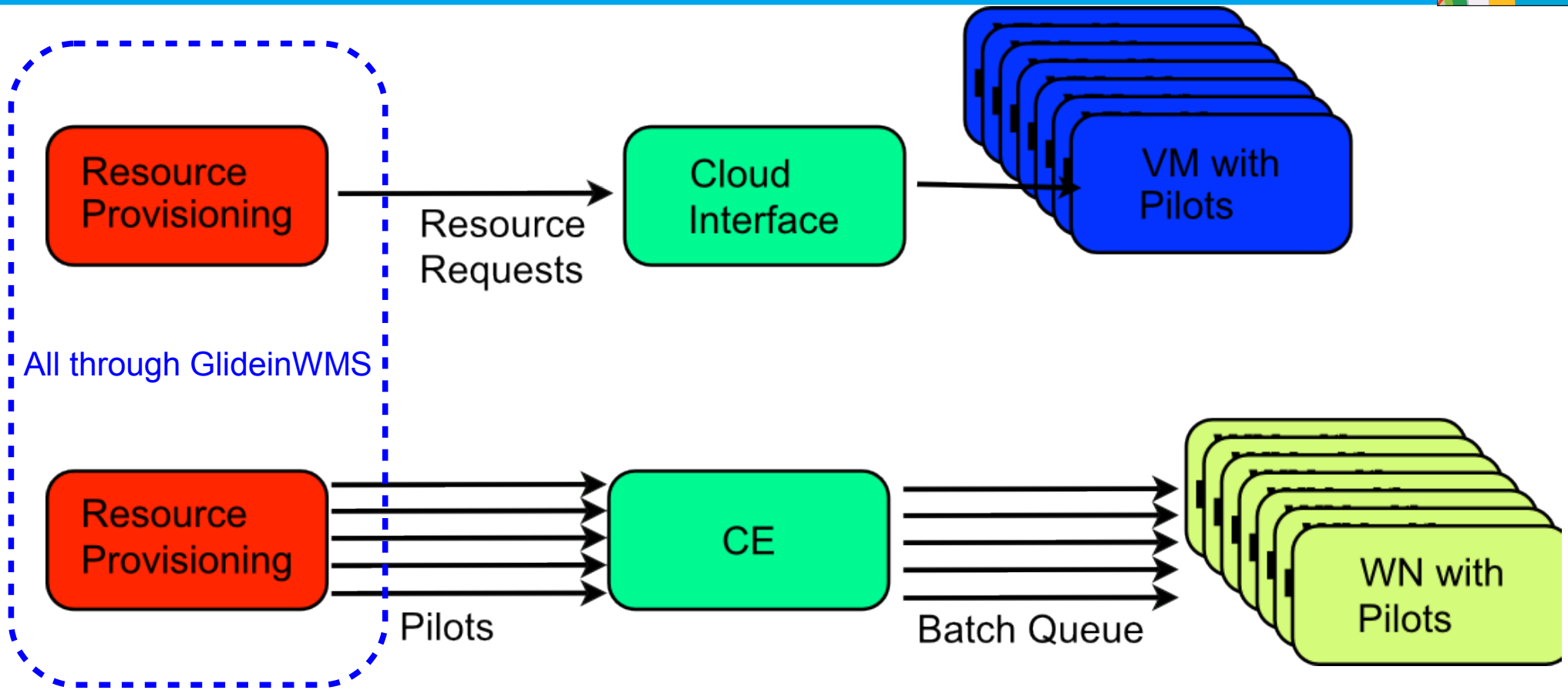
CHEP talk: A.Perez-Calero Yzquierdo -Evolution of CMS workload management towards multicore job support

Lowering Site Boundaries

- Gaining flexibility where to run which kind of workflow
- Data federation
  - Allows remote access of data
  - Paradigm “jobs go to the data” becomes less strict
- Resource allocation
  - All through GlideinWMS
  - One global HTCondor pool
- In progress of including “non-Grid” resources
  - High Level Trigger Farm (HLT)
  - Other opportunistic resources
    - HPC farm
    - (Academic) Clouds



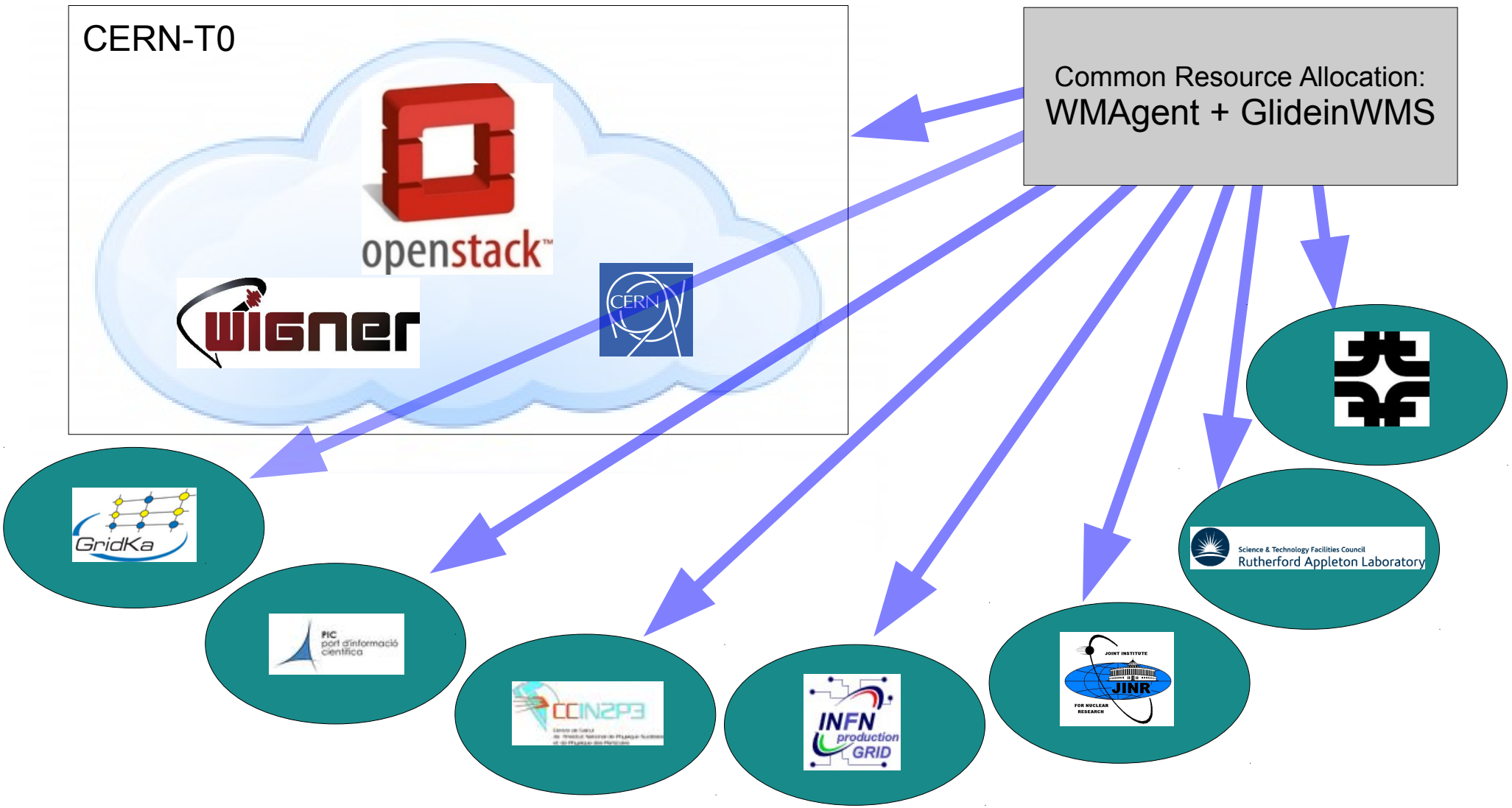
Resource Provisioning



- All resource allocation through GlideinWMS
- In production
  - Grid sites with various CE flavours
  - OpenStack Cloud interface
- Expanding to opportunistic resources

CHEP talk: J.Letts: Using the glideinWMS System as a Common Resource Provisioning Layer in CMS

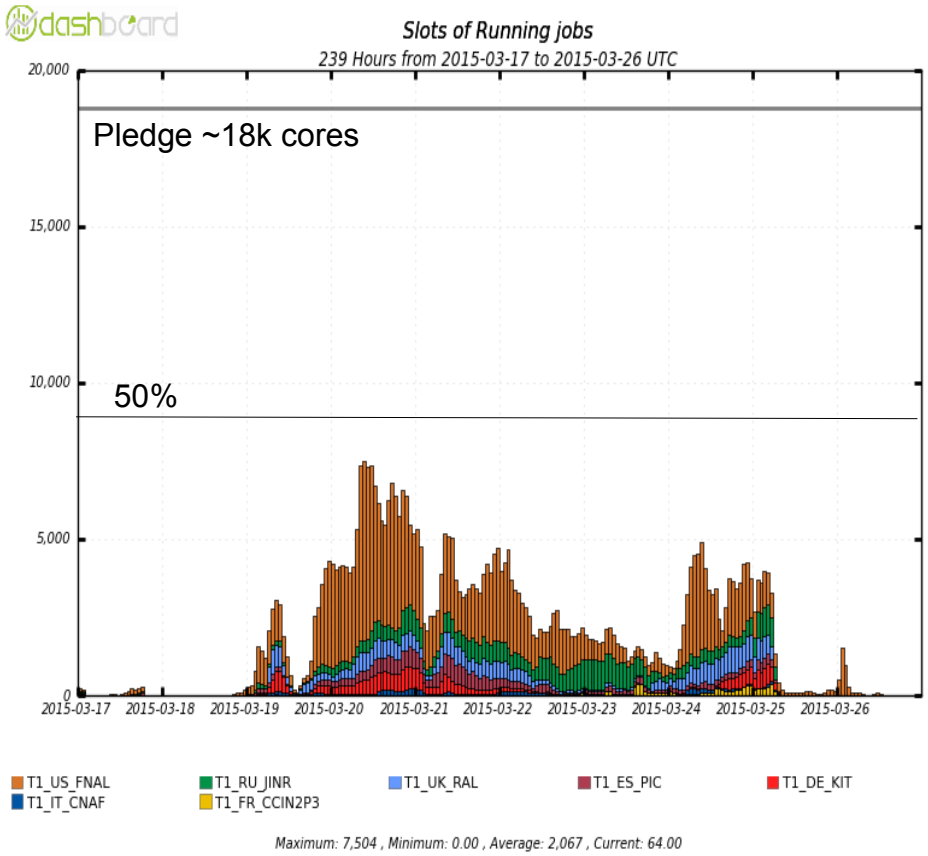
Tier-0 Application: Resource View



PromptRECO will use (up to) 50% of the Tier-1 CPU resources

Tier-0/PromptRECO Testing

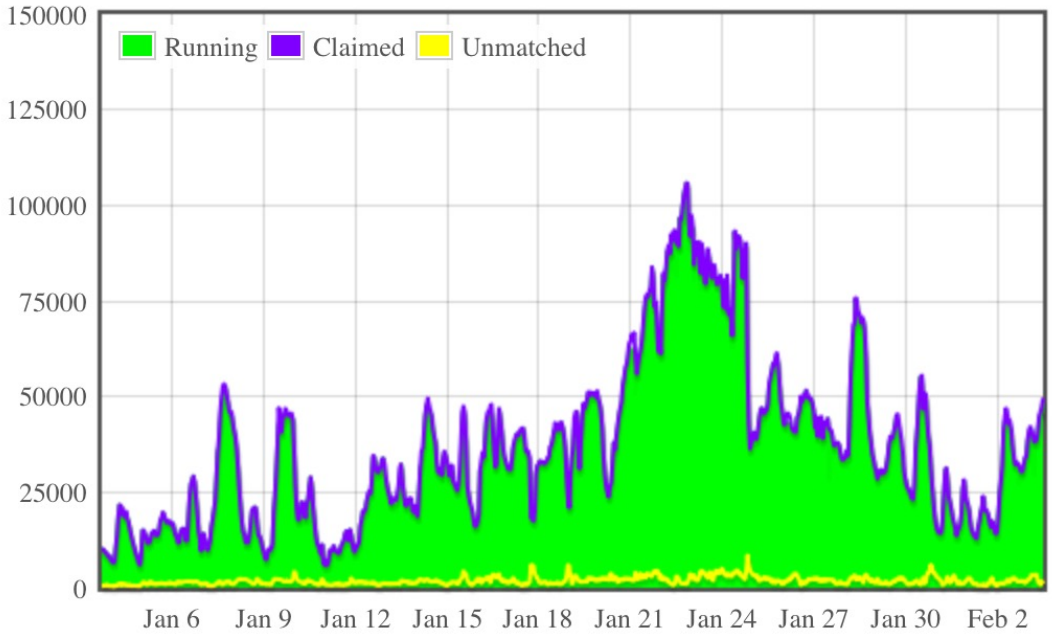
- PromptRECO will be multi-threaded
  - 8-core pilots
  - 4-thread application
- CERN Agile Infrastructure
  - Reached ~6000 cores in Nov./Dec. 2014
  - Want/need to double the scale
  - Working on remaining Meyrin vs Wigner issues
    - File merging badly affected by long latency when input is not local
- Tier-1 resources
  - Overall pledge ~18k cores
  - Reached peak of ~8k utilized cores

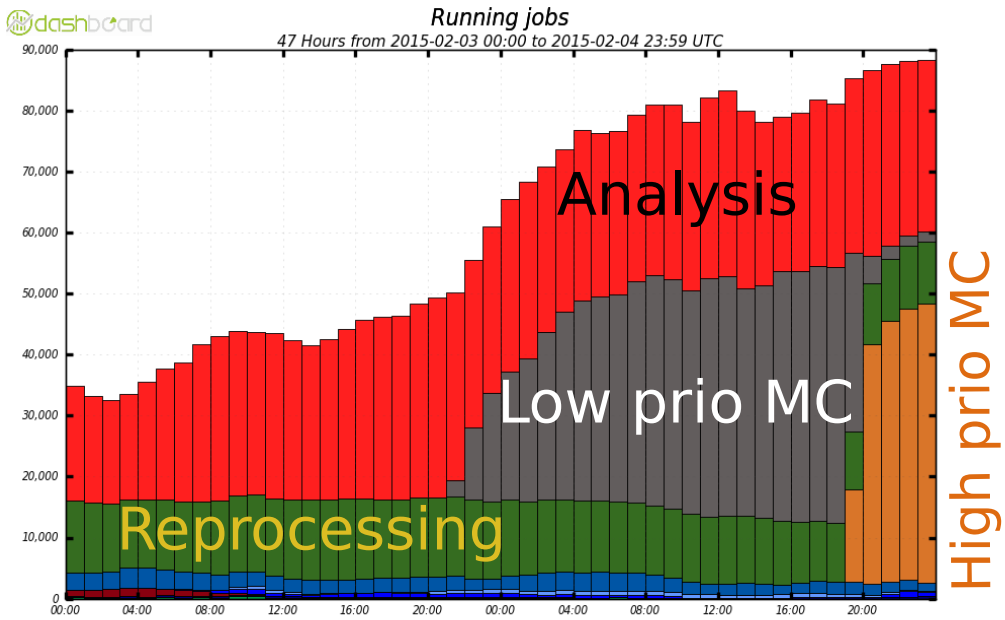


CHEP talk: D. Hufnagel - The CMS Tier-0 goes Cloud and Grid for LHC Run 2

Global Pool

- Global Pool
  - One single pool for production and analysis
  - All priorities handled inside the pool (not at the sites)
- Status
  - Migration done “on the fly”
  - Reached 100k concurrent jobs!
- Pilots at Sites
  - Tier-2 only VOMS role “pilot”
  - Tier-1 for the time being still mixture of roles “pilot” (for analysis) and “production”





CHEP Poster: J.Letts -
Pushing HTCondor and glideinWMS to 200K+ Jobs in a Global Pool for CMSbefore LHC Run 2

Reminder: Local Fair Share Configuration

Addresses only pledged CPU resources

Tier-1:

Priority oder	Share in %	VOMS role FQAN	Comment
highest priority	"small"	/cms/Role=lcgadmin	role for SAM tests, a few short jobs only, needs (almost) no fair share
default priority	95	/cms/Role=production	role for production workflows
default priority	5	/cms/Role=pilot	pilot role used for analysis jobs sent by Glidein
lower priority	0	any other role and no role	Should get resources only when the above roles are not active

Tier-2:

Resource Percentage	VOMS role FQAN	Comment
80%	/cms/Role=pilot	pilot role currently used by the Global Pool for Analyis **and** Production jobs
10%	/cms/Role=production	role for some legacy stuff still using that role
9%	any other role and no role	analyzers through gLite WMS
about 1%	/cms/Role=lcgadmin	Almost no share but **highest priority** to execute SAM tests

https://twiki.cern.ch/twiki/bin/view/CMSPublic/CompOpsPoliciesVOMSRoles

Using the HLT for Processing & Production

Sizeable resource:

- >200kHS06 / ~15k cores
- All Tier-1 WLCG pledges: ~300kHS06 (2015)

Openstack Cloud resource

- During Technical Stops
- Interfill mode still under study

No attached mass storage

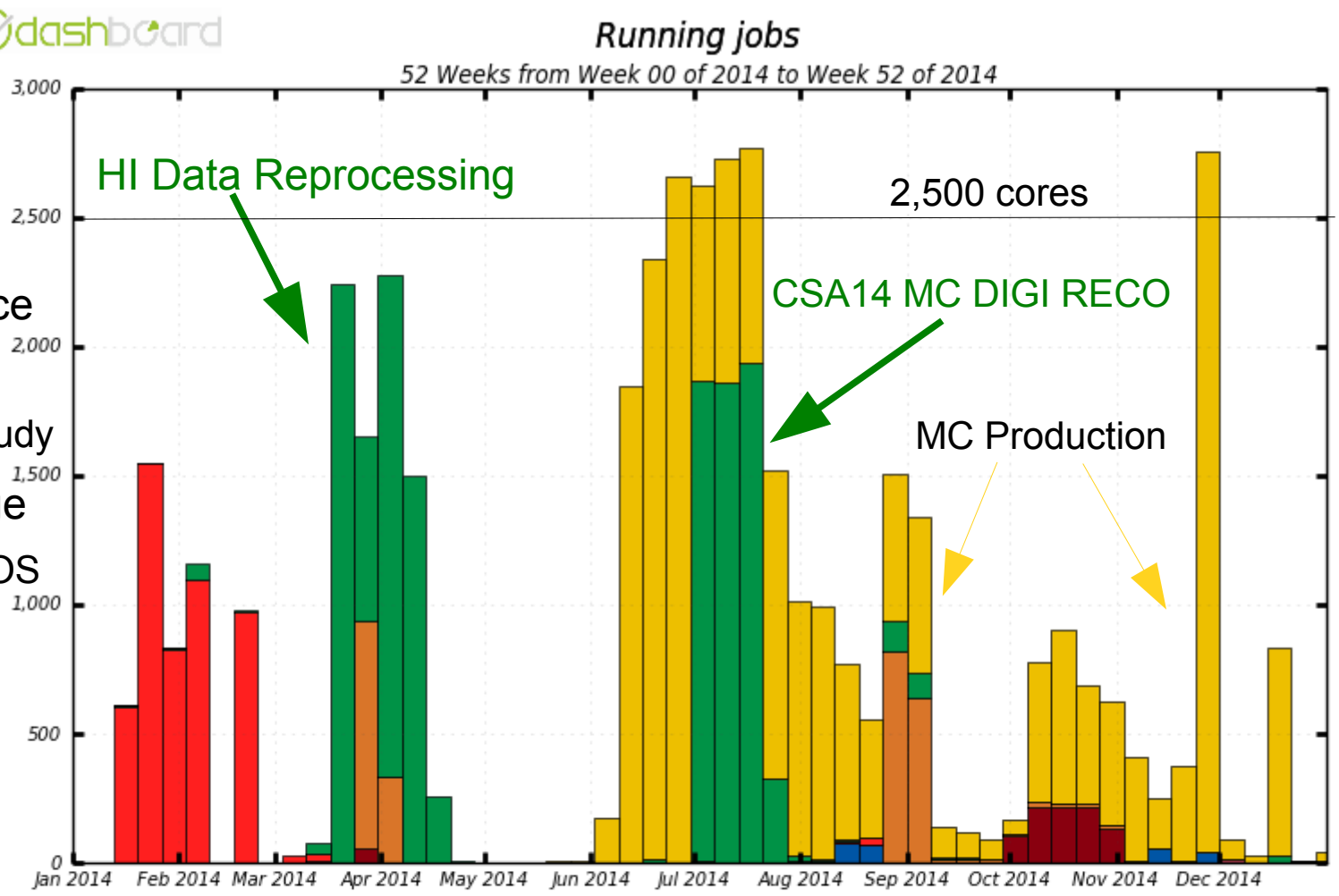
- Reading/Writing from/to EOS

Strong network link

- 60Gbit/s
- 120Gbit/s in progress

Workflows

- MC production (little I/O)
- Data RECO ~50kB/s in/out per core [Can scale to full HLT] ✓
- MC DIGI-RECO with high PU adds ~50kB/s per unit of PU ✗



CHEP talk: D.Colling - The diverse use of clouds by CMS

Opportunistic Resources

- Resources are tight for Run2
  - Have to restrict physics program to the 'necessary'
  - Resources beyond WLCG pledges (=opportunistic) offer some opportunities
- Various kinds of resources and access possibilities
  - Grid resources beyond WLCG pledges – we used to have Tier-3s also in the past
    - CMS sites and non-CMS sites
    - Access is straight forward and supported
  - Access to academic clouds
    - Base experiences from HLT cloud
    - Likely to find other places
    - GlideinWMS provides EC2 connection
  - Temporary access to High Performance Clusters
    - Usually do no have Grid interfaces like Compute Elements
    - BOSCO interface GlideinWMS can start pilots through SSH connection
    - Won't have root access to install e.g. CVMFS for software distribution
    - Parrot allows “installation” of CVMFS as non-root user at run-time
    - Can be complex, e.g. Linux different flavour different from SL
  - Commercial clouds (might be cheap when weakly utilized)

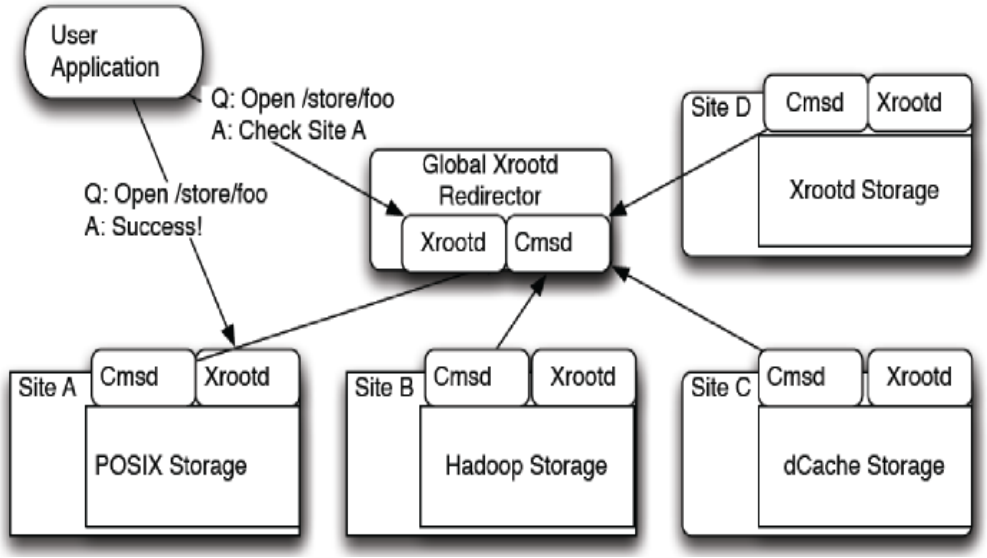
CHEP talk: D.Hufnagell - Enabling opportunistic resources for CMS Computing Operations

Data Management in Run2

- AAA – Any data, anytime, anywhere
  - Built on xrootd federation technology
  - Breaks “job always goes to the data” paradigm
  - Allows a more flexible job scheduling
- Dynamic data placement and cache release
  - Create and remove replicas of datasets based on popularity
  - Achieve more efficient usage of disks
  - Finally spend less effort in data management
- Disk Tape separation at Tier-1 sites
  - CMS computing operations controls what's on disk vs. what's on tape
  - Enhance flexibility where to process
  - Open Tier-1 sites for end user jobs – no risk of accidental tape re-calls

AAA - Status

- Fallback file open
  - Attempt to access file locally
  - Ask regional re-director, if local open fails
  - Activated at all CMS Tier-1 and (almost) all Tier-2 sites

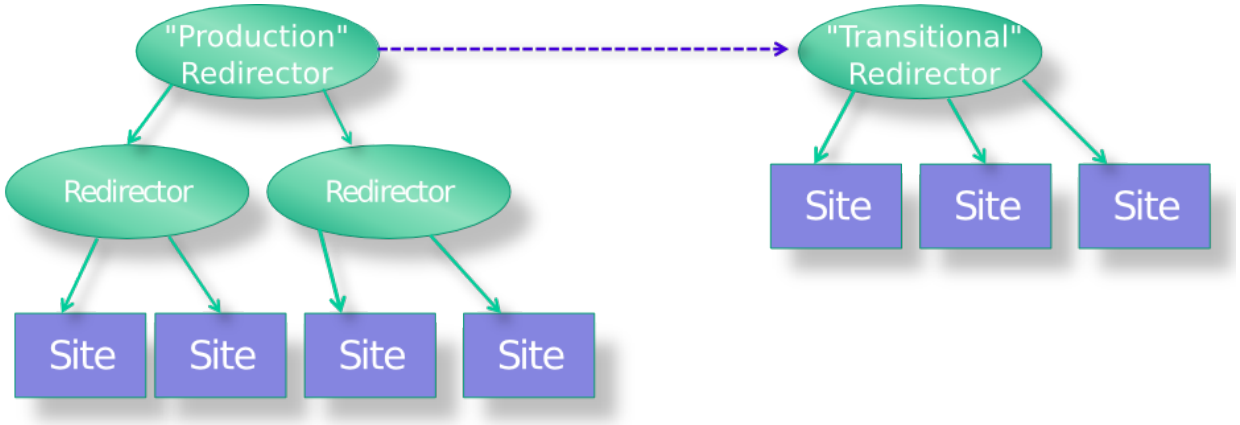


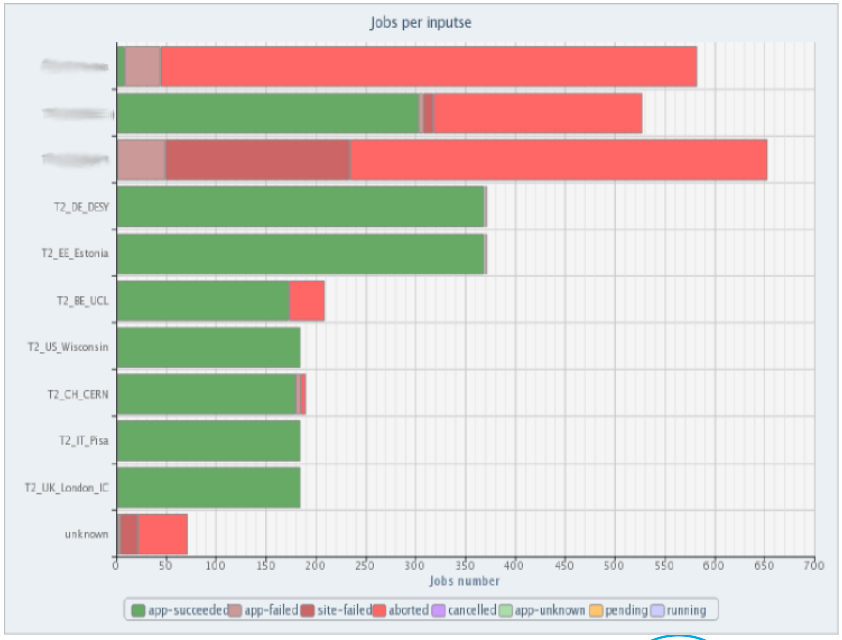
- Joining the data federation
  - Requires xrootd infrastructure with proper configuration at the sites
    - Detailed monitoring needed in addition
  - All CMS Tier-1 sites are part of the federation now
  - Most of the CMS Tier-2 sites are in the federation
    - Missing sites are in general smaller (and/or less robust)
    - ~96% of all datasets are available

CHEP Poster: M.Zvada - MS Experience with a World-Wide Data Federation

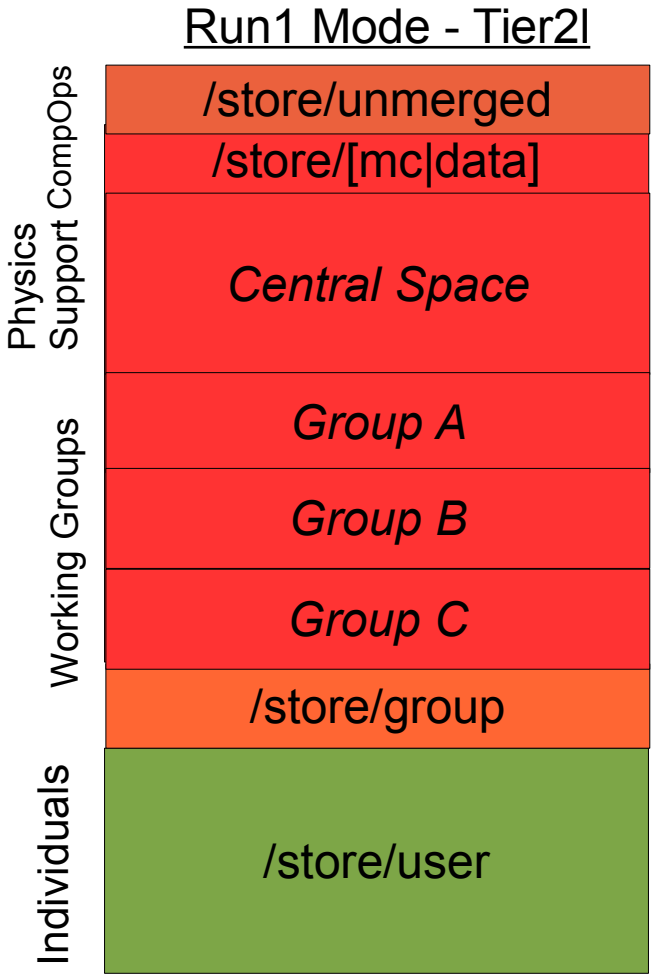
Transition Data Federation

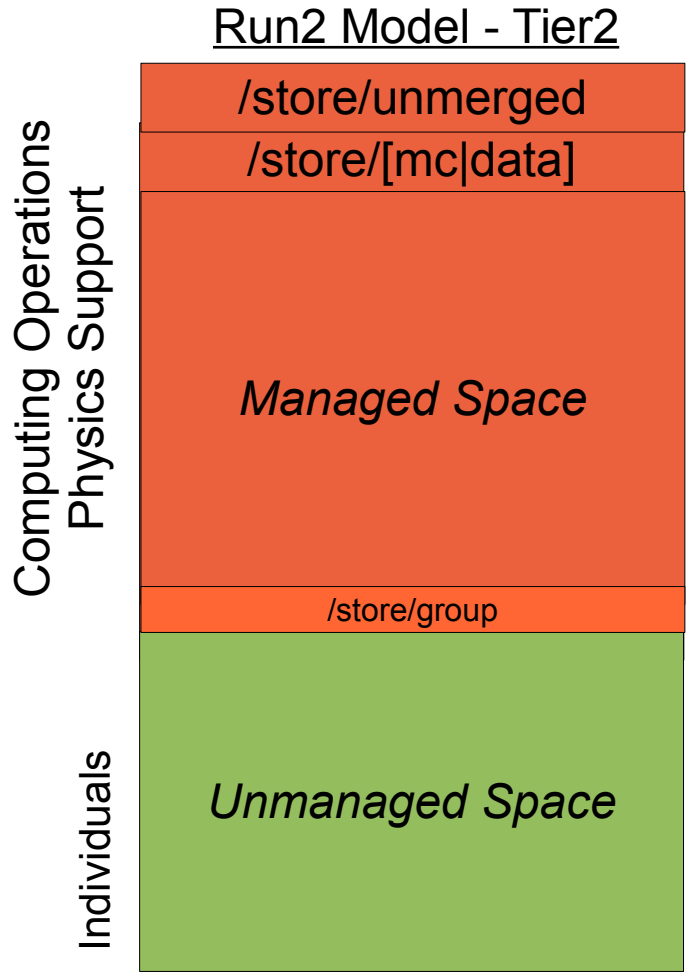
- Improve stability
  - Avoid file opening at “bad” sites
  - Still allow access to as many files as possible
  - Attempt open via “Production” redirector, fallback to “Transitional” redirector
- Employ most recent features of XROOTD4.1
- Metric for “Production” based on
  - Dedicated performances measurements
  - SAM tests
  - Special HammerCloud enforcing remote file opening





Dynamic Data Management

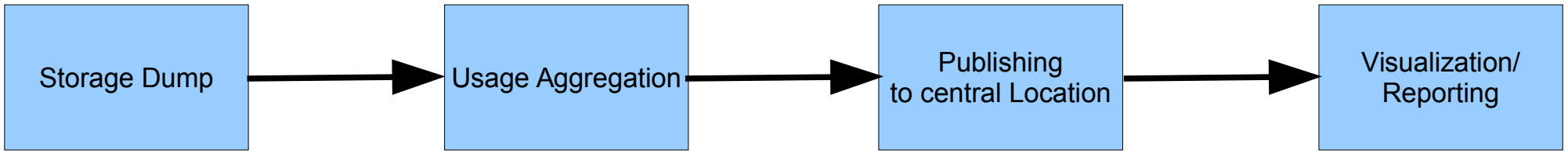




- All spaces formally managed by groups transferred into centrally managed space
- Central space is ~60% of pledged disk space
- Least popular data gets cleared when site reaches a quota value
- Create replicas for often accessed data
- Overall system is deployed including all Tier-1 sites

Space Monitoring at Sites

- CMS data transfer systems only knows about “official” datasets
  - Files registered and can be moved
- User or group produced files/datasets only partly registered
  - Registering of user/group files is optional
  - Cannot be moved around with CMS transfer system
- Dynamic Data Management depends on ~60% of pledges available
  - Presently no easy way to verify and monitor
  - Relaying on local site staff
- Space monitoring to be deployed at sites

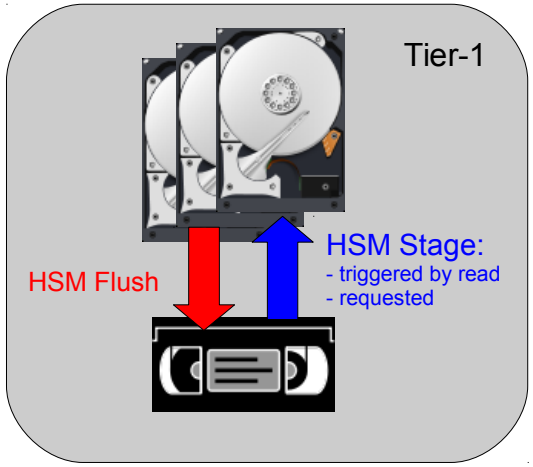


- In contact with sites to deploy it

Disk Tape Separation

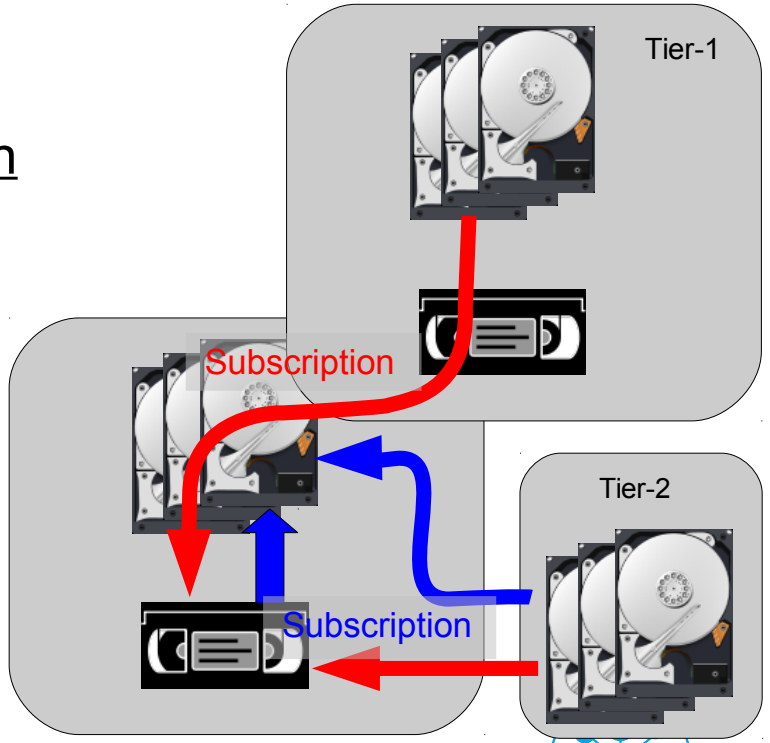
Run1: Disk-Tape coupling at Tier-1

- Uncontrolled access to files on tape forbids user analysis at Tier-1
- Produced files go to tape “immediately”
  - > Cannot use site without writing to tape → inefficient and inflexible



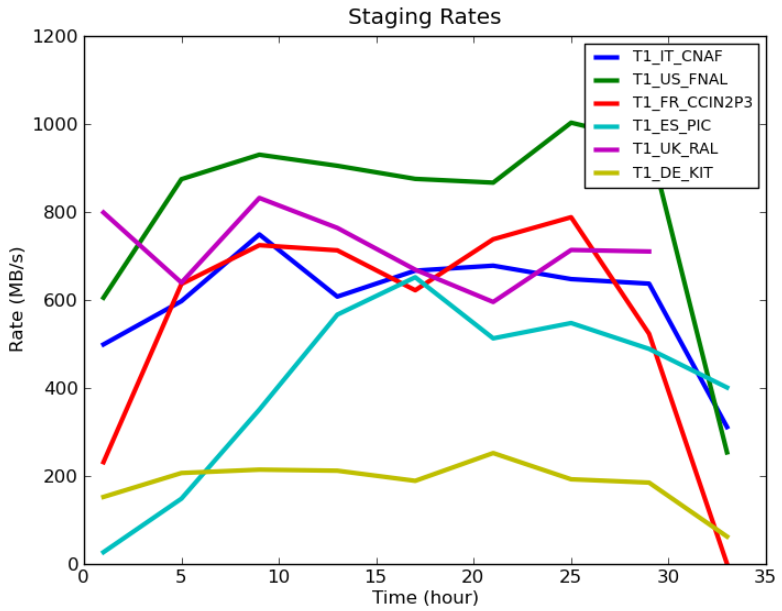
Run2: Separate disk and tape

- Large disk pool and small tape read/write pool
- Staging and flushing from/to tape is a subscription in CMS data management tool
- Files on disk can be read via WAN xrootd (never trigger tape operations)
- Tier-1s can be open for 'chaotic' analysis jobs
- Production and (tape) archiving location now independent
- Completed at all Tier-1 sites



Tape Staging Tests

- Tape archiving plans
  - RAW data will be archived with 2 copies (as before)
  - Plan to archive AOD/AODSIM, GEN-SIM
  - RECO will not be archived by default – plan to archive ~10%
- Tape staging rates assumed to be proportional to pledged capacity
- Recent exercise tested reading – assumes similar write performance

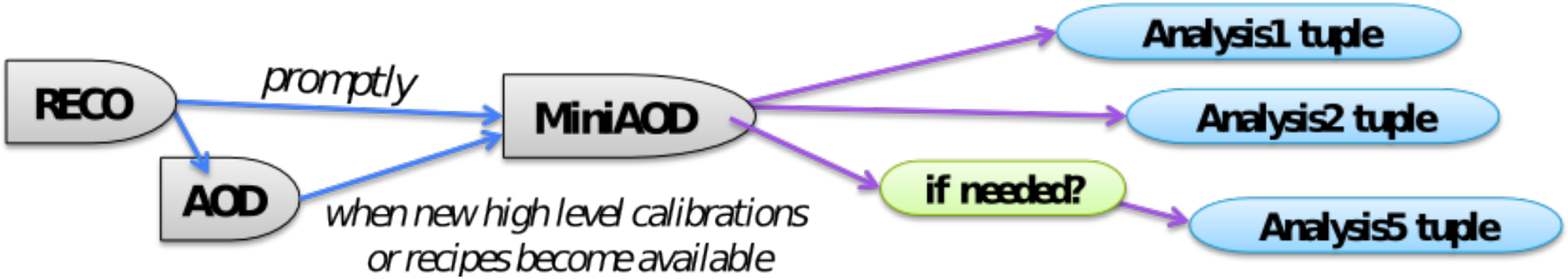


Site	Expected Rate (MB/s)	Achieved Rate (MB/s)
FNAL	650	~900
CNAF	210	~630
JINR*	150	*
KIT	150	~200
RAL	135	~700
IN2P3	135	~650
PIC	75	~500

All tape rates well above needs

Some testing at CERN soon

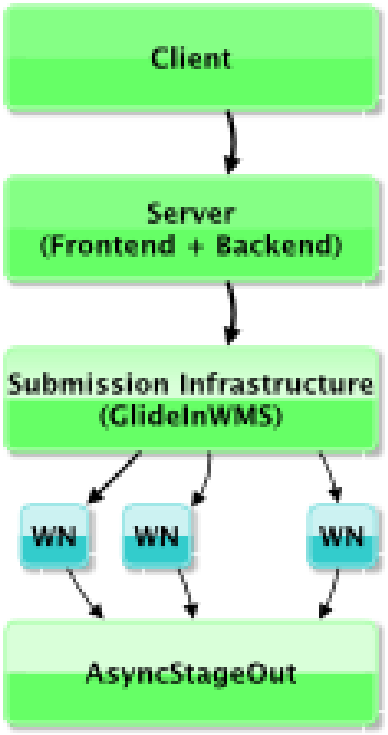
A new Data-Tier: miniAOD



- Replace the dozens of “Group ntuples/trees”
- Size ~50kB/events
  - High level physics objects (Leptons, Jets...)
  - Particle Flow candidates in packed format (to allow for re-clustering, jet-structures etc)
- Should satisfy ~80% of all analysis cases
- Large exercises during CSA14
  - Good experiences by analysis users
- MiniAOD production
  - Additional output of PromptRECO
  - Can be re-done a few times per year

CHEP talk: C. Vuosalo -A new analysis data format for CMS

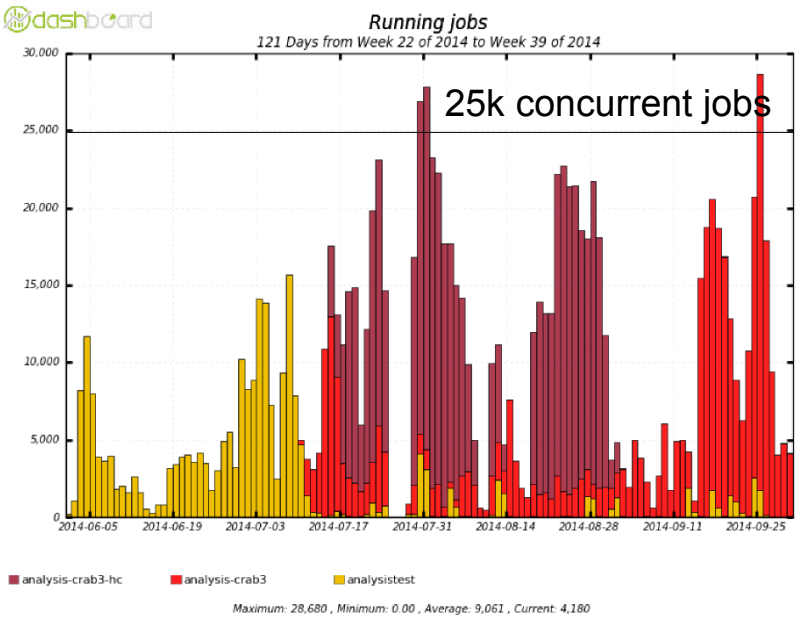
User Analysis Job Submission: CRAB3

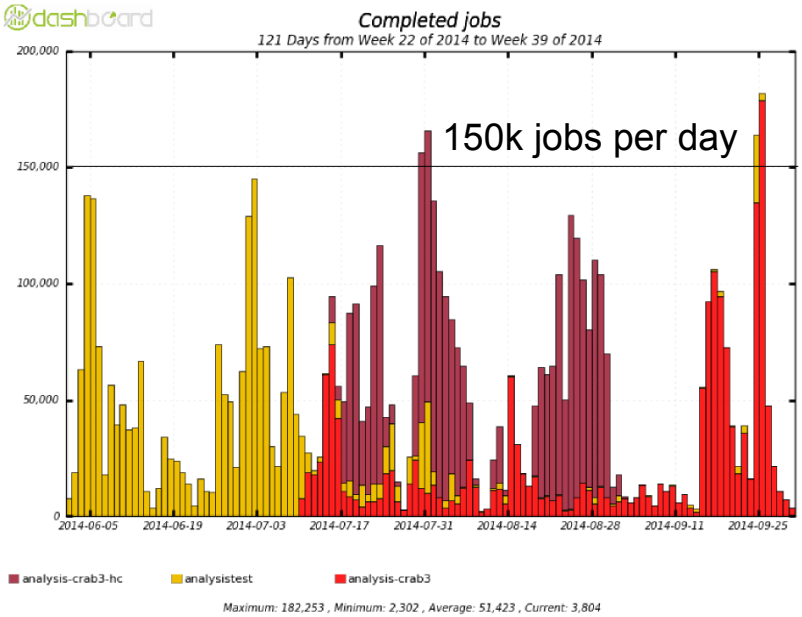


From M.Mascheroni

Submission Use Case

- Client sends a submit request to the server
- The server frontend checks auth/authz and store the request. The backend "prepares the jobs" and send them to the submission infrastructure
- The jobs are sent to the Grid. They contact the AsyncStageOut (ASO) server to start the transfer
- ASO transfers the files (using FTS) and it publishes them





- Default tool for user analysis in Run2
- ~30k parallel jobs reached
- Processed up to 150k jobs per day

CHEP talk: M. Mascheroni- CMS Distributed Data Analysis with CRAB3

Summary

- Coming Run2 will be resource constrained
  - Improve CPU performance of applications
  - Increase the flexibility of resource usage
- A number of efforts to finish for Run2
  - Completely Cloud/Openstack based Tier-0
  - Expanding PromptRECO to Tier-1 sites
  - Commissioning of the HLT for Production and Processing
  - Data federation to ease remote data access
  - Dynamic data management to improve usage of disk space
  - A new tool for user analysis
- Experiments want to discover new physics in Run2
  - Computing relies on good planning
  - Difficult to plan for the unknown

Will have an exiting to time to deal with this “conflict”

Appendix

WLCG Resources 2015 & 2016 Request

	2014	Increase from 2013	2015	Increase from 2014	2016 (C-RSG Oct 14)	Increase from 2015
Tier-0 CPU (kHS06)	121	0%	256	111%	302	18%
Tier-0 Disk (TB)	7000	0%	3000	Reallocated to CAF	3250	0%
Tier-0 Tape (TB)	26000	0%	31000	31%	38000	23%
CAF CPU (kHS06)	0	0%	15	-	15	17%
CAF Disk (TB)	0	0%	12000	-	13100	9%
CAF Tape (TB)	0	0%	4000	-	6000	50%
T1 CPU (kHS06)	175	0%	300	71%	400	33%
T1 Disk (TB)	26000	0%	26000	4%	35000 (33000)	30%
T1 Tape (TB)	55000	11%	74000	34%	100000	35%
T2 CPU (kHS06)	390	14%	500	25%	700	40%
T2 Disk (TB)	27000	4%	29000	16%	40000 (38000)	37%