CMS reprocessing tests at STEP09

Claudio Grandi (INFN Bologna)

CMS Tier-1 Coordinator

Introduction

- Goals of the tests:
  - Determine the effect of accessing data on disk vs tape on reprocessing efficiency
  - Verify that the sites could stage data from tape at a rate compatible with the reprocessing needs
  - Gain experience with organized pre-staging of data
  - Verify that the sites are able to provide CMS with the pledged resources at any time (fair-share algorithms)

MSS description

Site	Total number of tape drives	Average fraction for CMS (percent)	Nominal write speed of a tape drive (MB/s)	Nominal read speed of a tape drive (MB/s)	Is read/write performance per VO monitored?	Size of CMS disk buffer (TB)
T1_DE_FZK dCache/ TSM	Currently 24 (+12 to come)	CMS share: 7-8 drives (4 to come)	LTO4 drives (~50 MB/sec), write speed varies widely on use case, absolute number not that meaningful	LTO4 drives (~50 MB/sec), read speed varies widely on use case, absolute number not that meaningful	Monitored on dCache level	~600 TB tape read pools, 50 TB tape write pools, 100 TB disk-only pools
T1_ES_PIC dCache/ Enstore	Currently 13 (+4 to come) - 2 9940B STK - 7 LTO3 IBM - 2 LTO4 IBM - 2 LTO4 STK - (+4 LT04 STK to come, not before STEP09)	'storage_group_limits': {'vo-cms' : 2, 'vo-atlas' : 2, 'vo-lhcb' : 1}	LTO3 80 MB/s, LTO4 120 MB/s, 9940B 30MB/s	LTO3 80 MB/s, LTO4 120 MB/s, 9940B 30MB/s	Yes	47 TBs reserved for tape-recall ; 340 TBs in front of tape (almost all disk is buffer)
T1_FR_IN2P3 dCache/ HPSS	Currently 36 drives T10k +30 drives T10kb (Jun 8th, ONLY for migration, no staging)	Driver are not dedicated to VOs	50 MB/s	100 MB/s	No at HPSS level, but at dcache one	413 TB
T1_IT_CNAF Castor	T10000B: 20 9940b: 10 (to be dismissed)	~25%	T10000B: 100 MB/s 9940b: 20 MB/s	T10000B: 100 MB/s 9940b: 30 MB/s	Aggregated network throughput in Lemon	156 TB
T1_TW_ASGC Castor	6 LTO4 (increase before 2010 yet to be quantified)	50%	~75 MB/s	~ 85 MB/s overall	yes (thanks tape logging facility from castor)	120TB for farm read
T1_UK_RAL Castor	5 for CMS	5 for CMS	50 MB/s	50 MB/s	We monitor data rates in / out of each disk pool, and can monitor the rate on drives that are dedicated to the VO.	~450TB (Farm), ~200TB (Import), ~100TB (Export)
T1_US_FNAL dCache/ Enstore	25 LTO4	100%	~60 MB/s overall - see https://twiki.cern.ch/twiki/bin/view/LCG/MssEfficiencyUS-FNAL-CMS	~65 MB/s overall - see https://twiki.cern.ch/twiki/bin/view/LCG/MssEfficiencyUS-FNAL-CMS	Yes	3.2 PB

Tests description

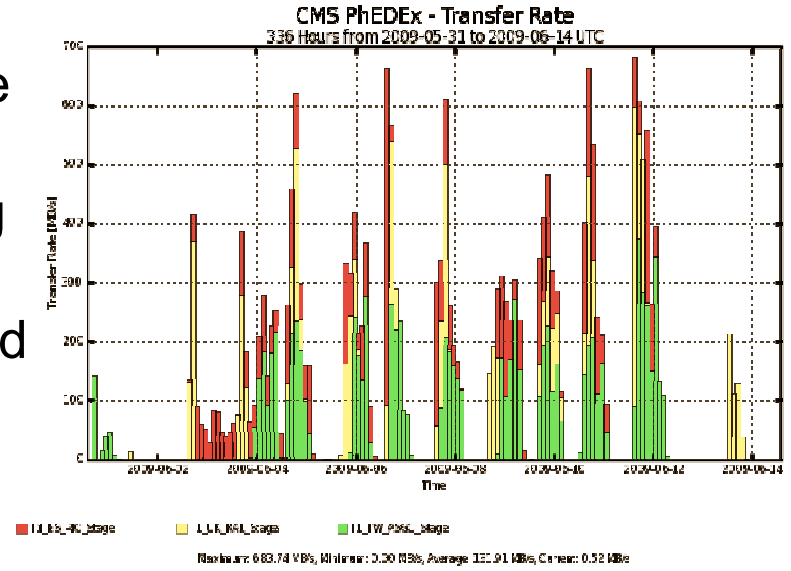
- Pre-stage data using different methods:
  - Central script based on SRM commands (CNAF)
  - PhEDEx pre-stage agent (ASGC, PIC, RAL)
  - Site-operated (FNAL, FZK, IN2P3)
- Every day at 16:00:
  - wipe from disk the data of day N+1
  - pre-stage the data of day N
  - process the data of day N-1

	Pre-stage	Process	Purge from disk
June 3	day00		
June 4	day01	day00	
June 5	day02	day01	day00
June 6	day03	day02	day01

- Measure the staging time and find the rates
- Monitor the number of slots provided to CMS at each site
- Measure the processing efficiency (CPT/WCT) at each site
- On the last day: process data not pre-staged and compare the processing efficiency

Pre-stage methods

- Central script based on SRM commands
  - Issues a bulk *BringOnline* request containing all the SURLs
  - At regular (configurable) time intervals, it queries
    - the status of the files in the request via *StatusOfBringOnline*
    - the locality of all SURLs via *Ls*
  - StatusOfBringOnline does not work well on Castor
    - unless Castor-SRM 2.8 and Castor 2.1.8 are used
- PhEDEx
  - Based on a special configuration of the standard transfer system
  - The buffer used by the jobs for reading data is treated as a PhEDEx node
  - Exploit the existing pre-stage agent and the control and monitoring infrastructure
    - New pre-stage agent for d-Cache

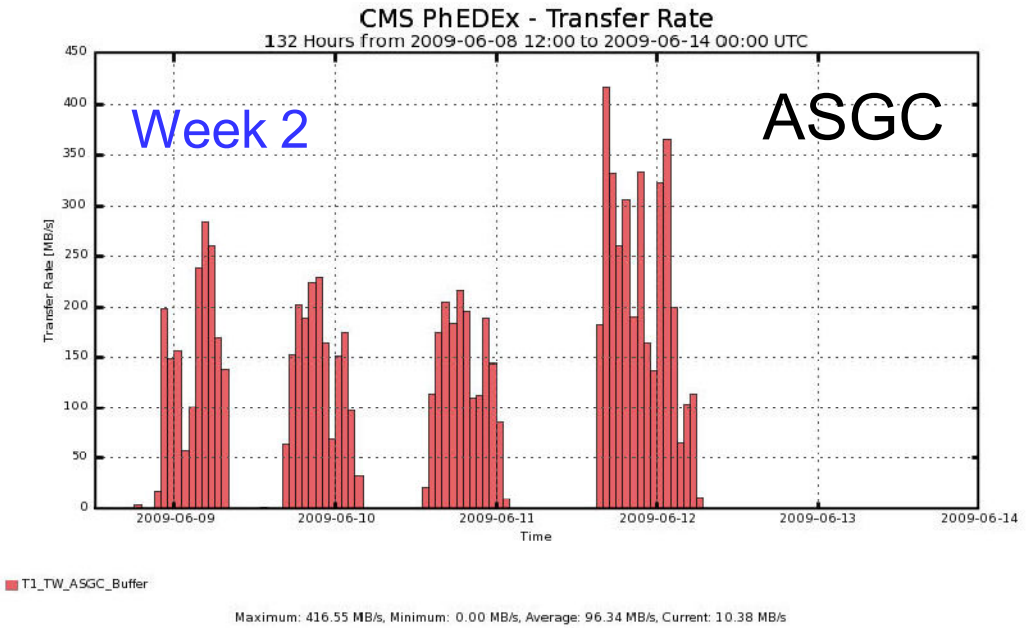


- Site-operated pre-stage
  - The list of files to be pre-staged is communicated to the site manager that starts the procedure locally

Data samples

- The samples used for the pre-stage and reprocessing tests were real data collected during the 2008 cosmic global runs
- The size of the samples to be used daily at sites depends on the expected MSS read rate
  - that in turn depends on the requested processing rate
- Average file size was 2.5 GB

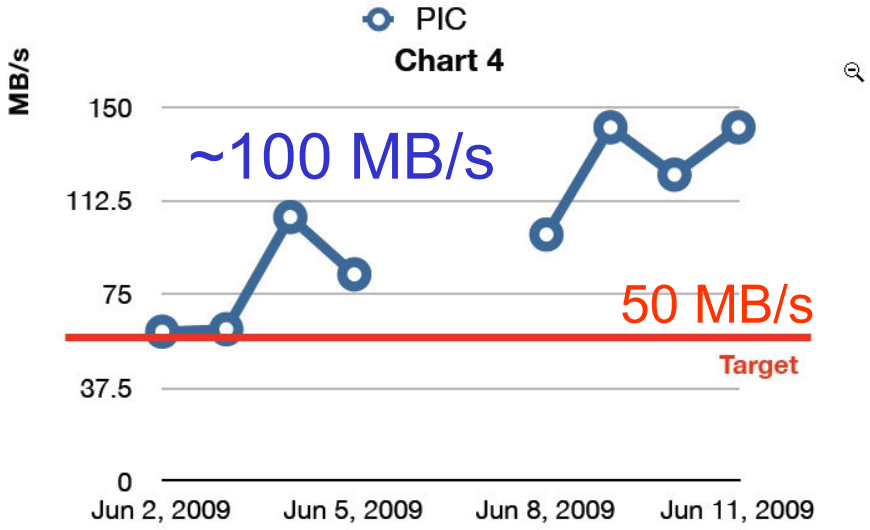
	Size (GB)	Expected rate (MB/s)
FZK	7179	85
PIC	4225	50
IN2P3	4364	52
CNAF	4727	56
ASGC	6131	73
RAL	3346	40
FNAL	20365	242

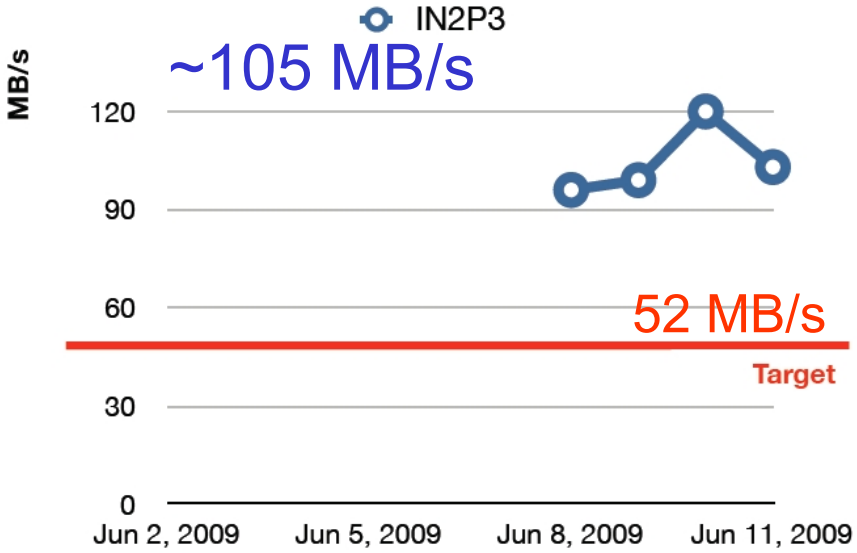


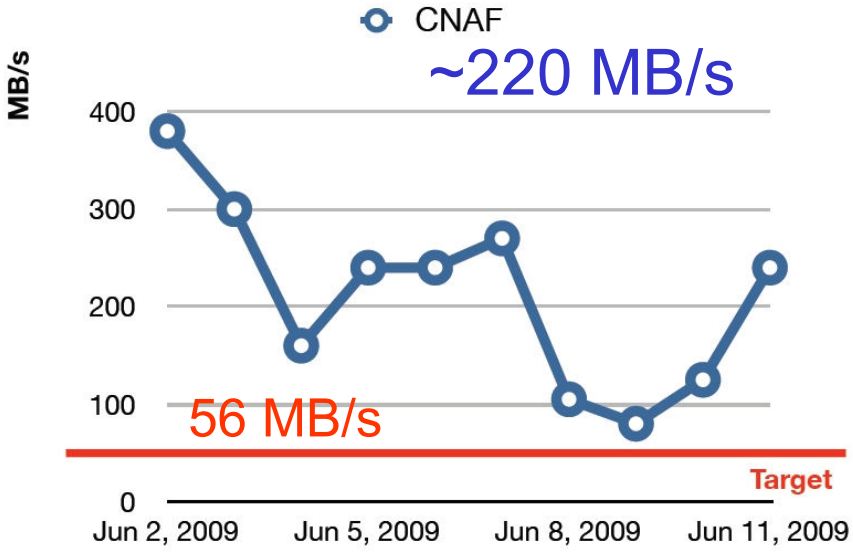
- The time needed to stage in the whole sample is measured

Pre-stage rates 1/2

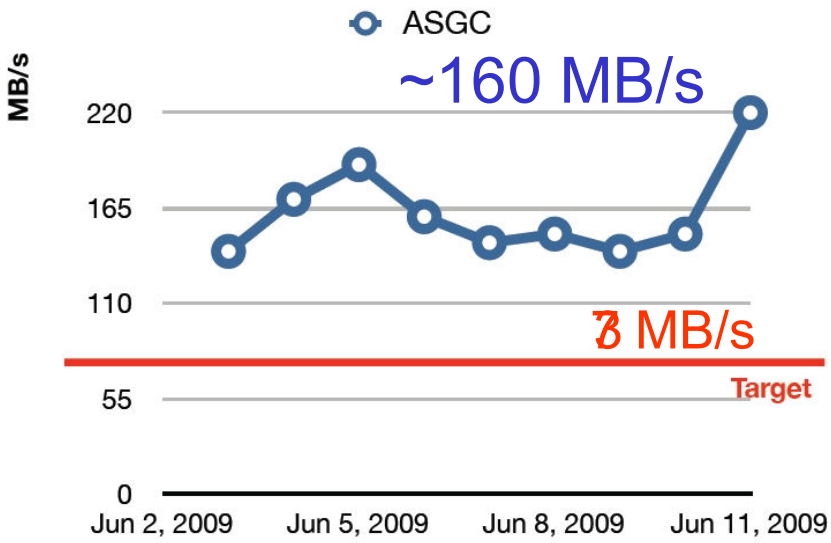
No data for FZK
(overall CMS rate
from tape ~120 MB/s)

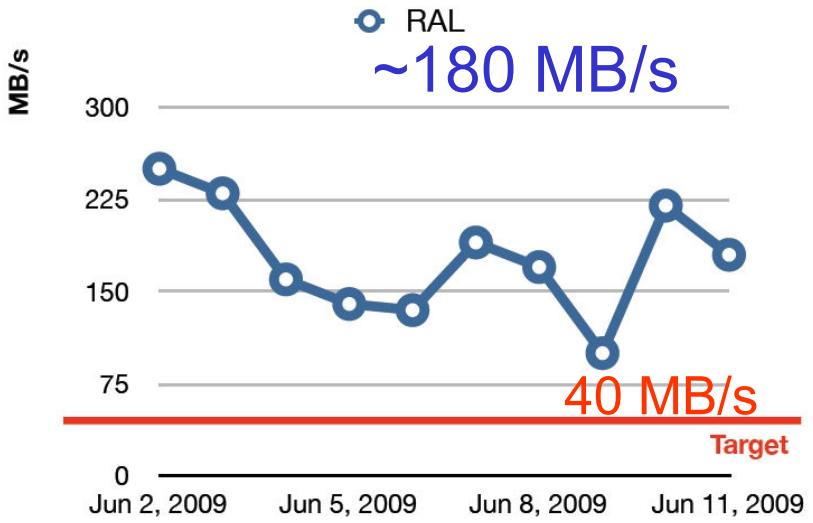


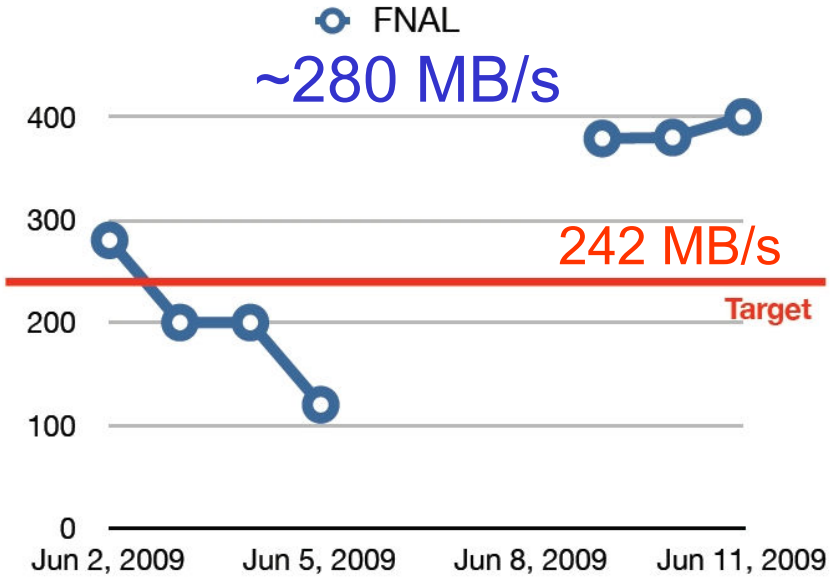




Pre-stage rates 1/2





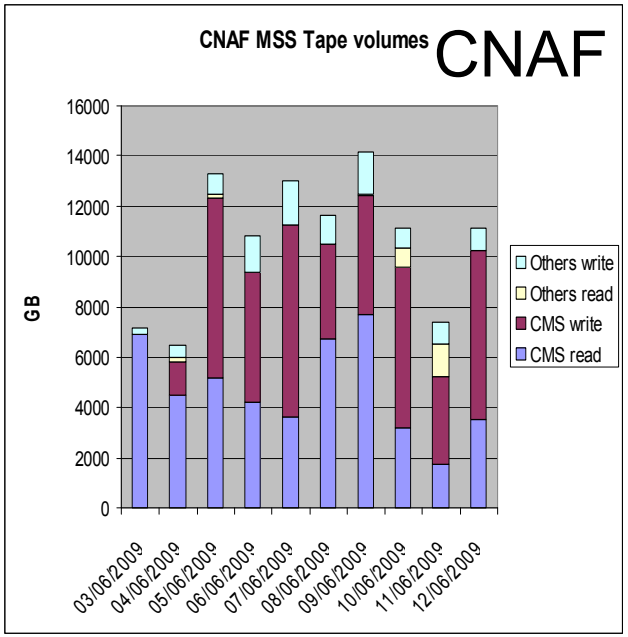


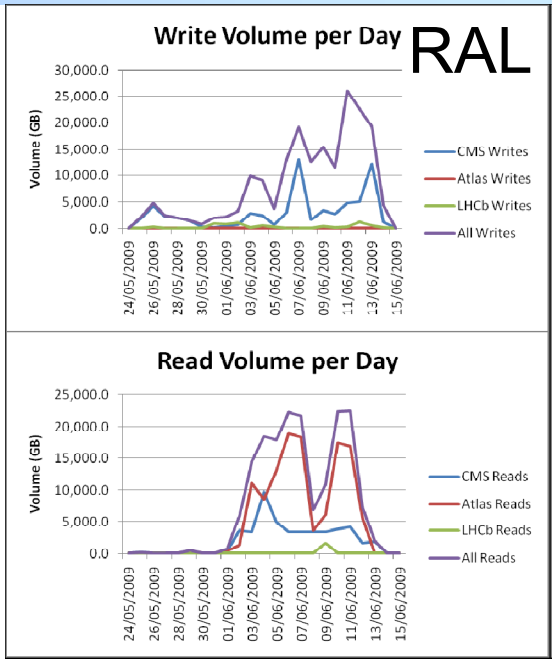
All sites exceeded the expected rate

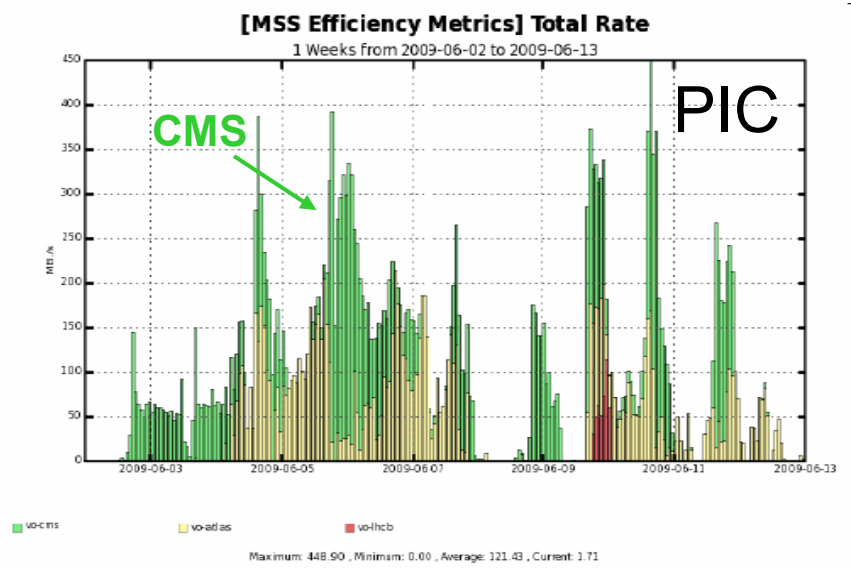
Need to redo the tests at FZK

Tape usage

- Tape Read:
  - pre-stage
  - data transfers
    - T1-T1, T1-T2
  - non-STEP09 activities
- Tape Write:
  - re-processing output
  - data transfers
    - T0-T1, T1,T1
  - non-STEP09 activities
    - including T2-T1 transfers
- At several sites CMS was the main user of the tape systems
  - Use by ATLAS was heavy on the disk pools but not on the tape system

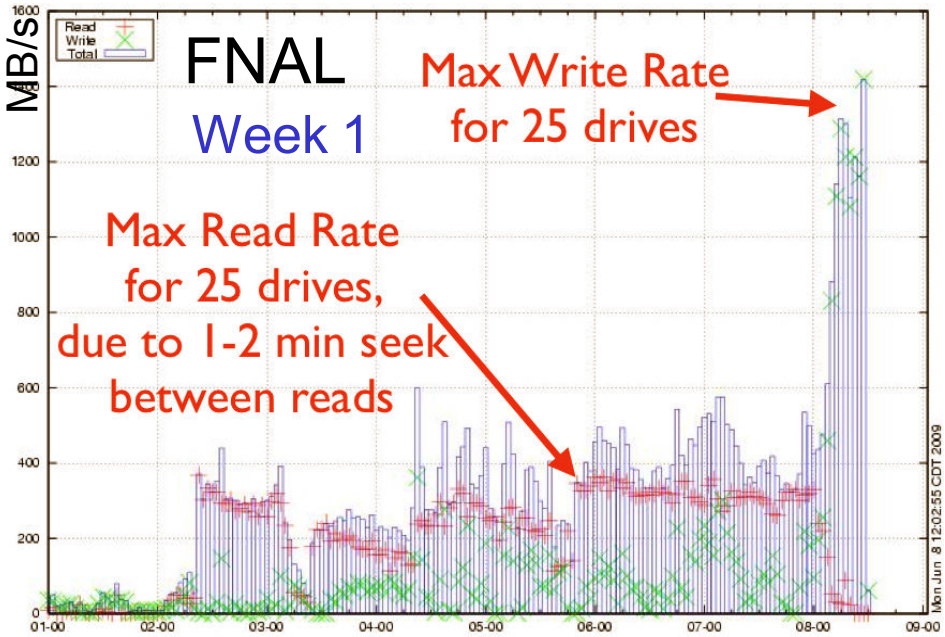


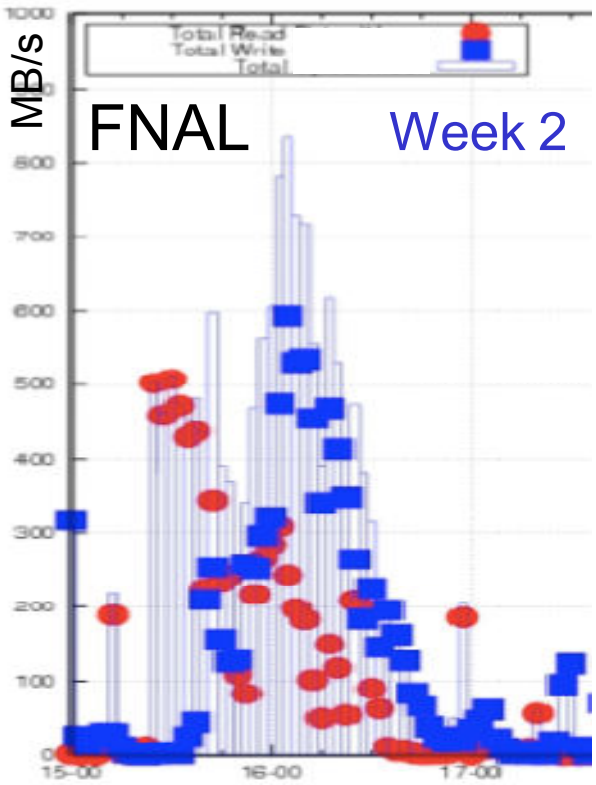




Effect of MSS optimization: FNAL

- Week 1: problems in staging from tape due to the competition between read and write (with higher priority for writes)
  - High rate of "seeks" in combination with a 1 minute delay in reads
    - 'normal' for non adjacent files

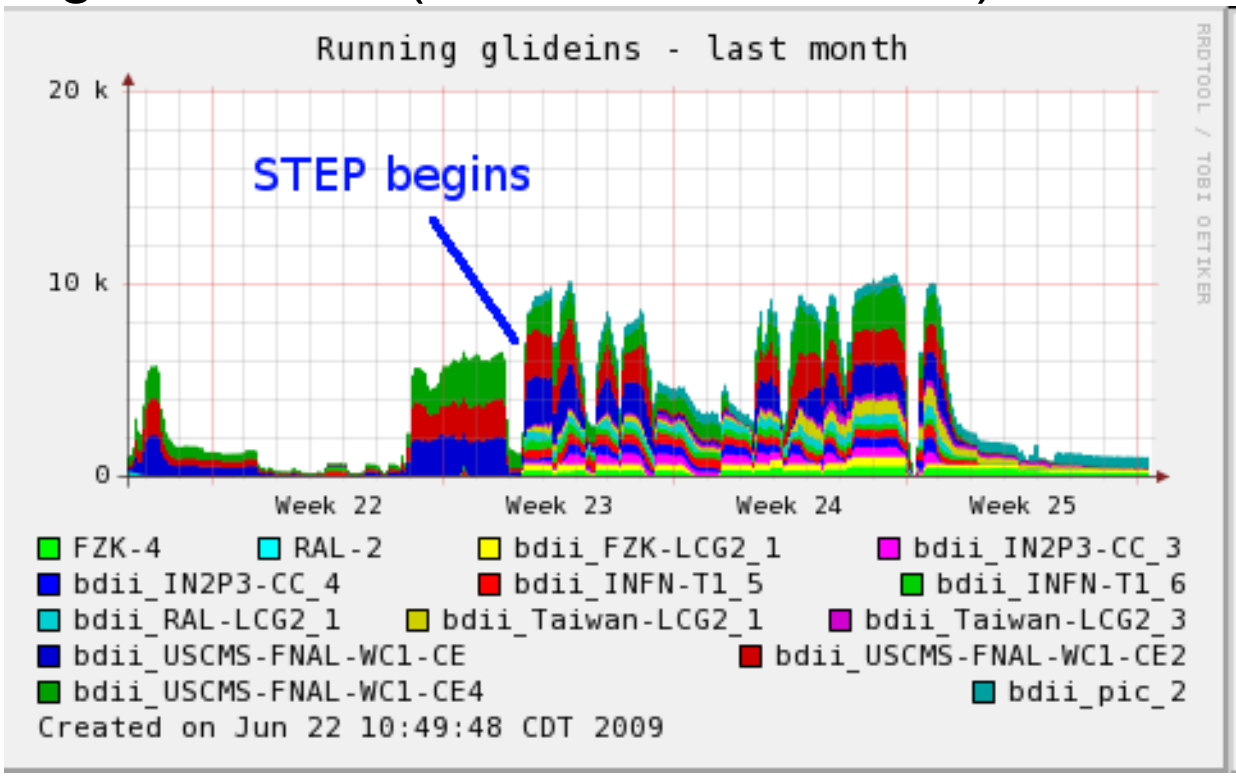




- Week 2: Optimization at several levels
  - more tape ordering for staging;
  - acting on # active transfers per node;
  - kernel changes for buffer allocation and % of dirty pages in mem;
  - tests on encp buffer size to achieve encp disk rates of ~90 MB/s on average per drive overall in reads
- Read rate increased of 25%; stage-in rate 400 MB/s

Reprocessing

- Reprocessing controlled by a single operator submitting jobs via glide-in
- Number of jobs kept close to the level of the pledge at sites
- Jobs analyzing 1 or 2 files (6 to 15 hours)
- During the first 9 days of the test the unfinished jobs were killed before submitting new ones (before 16:00 CEST)
- On day 10 (with prestaging) and 13 (without prestaging) the jobs have been left running for 2 days and used for the comparison with and without pre-staging

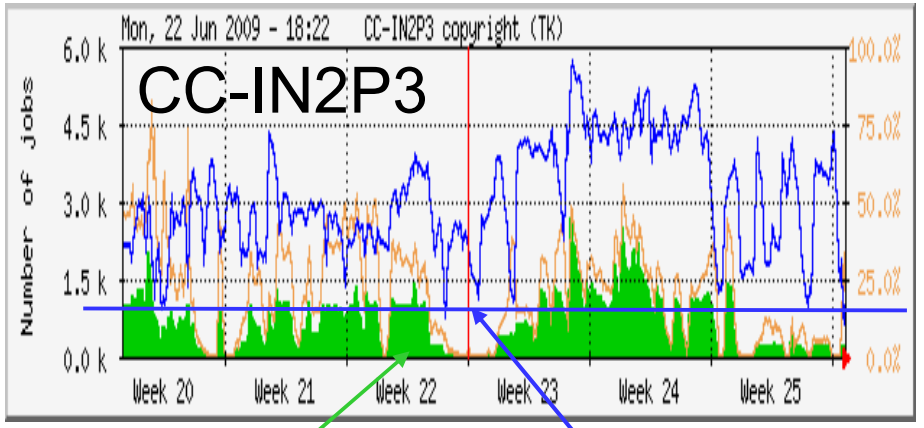


Fair shares at CMS Tier-1s

- FZK: Two-step mechanism, a combination of a short (1 day) fair-share adjustment with a long-term (180 days) accounting of the consumed resources. Limit to [0.5, 2] times the nominal VO shares
- PIC: the historical data is broken in 14 slots, 12h each, with a decay rate of 20% between them
- CC-IN2P3: the fairshare is on a7days weighted average
- CNAF: the fairshare is on a 2 days weighted average
- ASGC: the fairshare is on a7days weighted average
- RAL: the fairshare is on a 9 days weighted average with the previous days usage counting most and the 9th day counting only at the 5% level
- FNAL: no fairshare, CMS site only

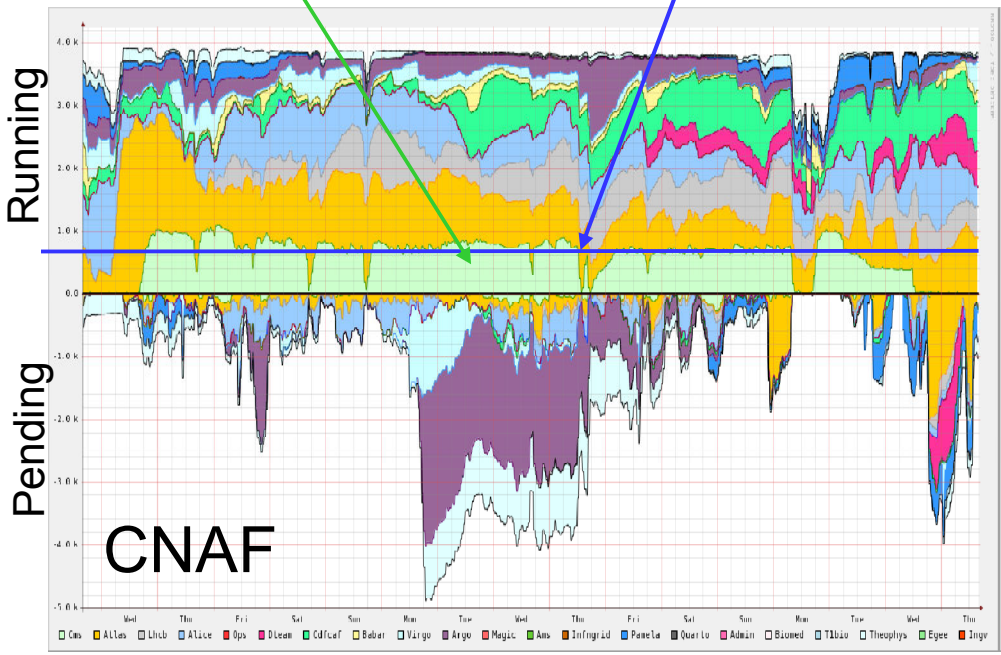
Fair shares

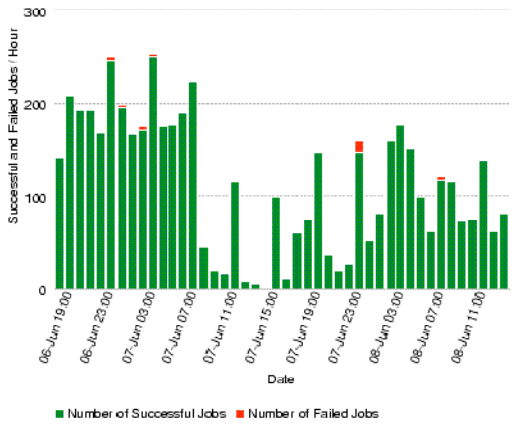
- CMS could get its share of resources basically at all sites notwithstanding the competition with other VOs
  - Only some difficulties at ASGC on the first days
  - Recommendation for short (~2 days) “cool-off” period
- Good job efficiency at all sites

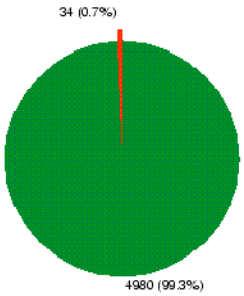


CMS

Pledge



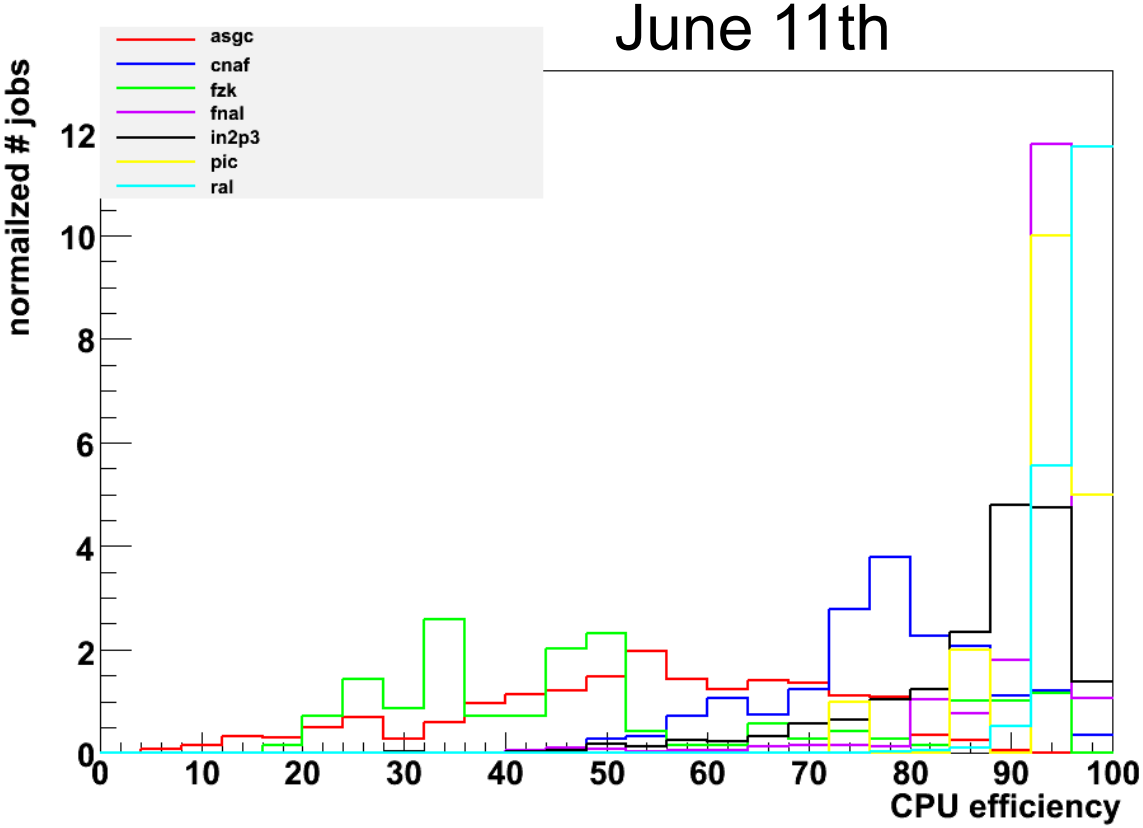




FZK

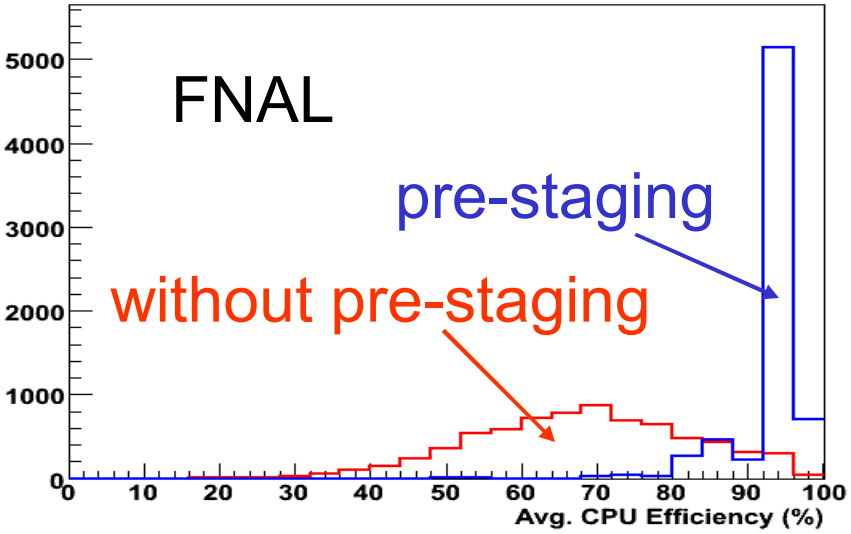
Reprocessing efficiency

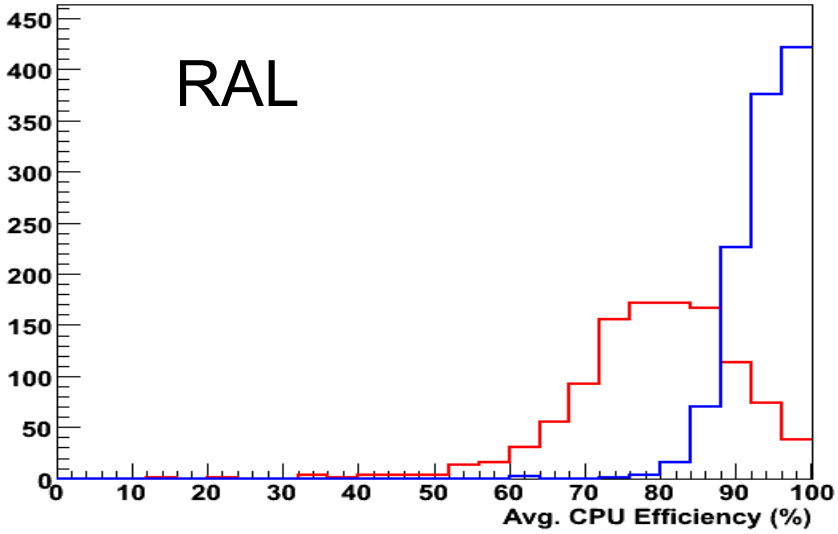
- Measure CPT/WCT for all jobs
- Very different performances at different sites

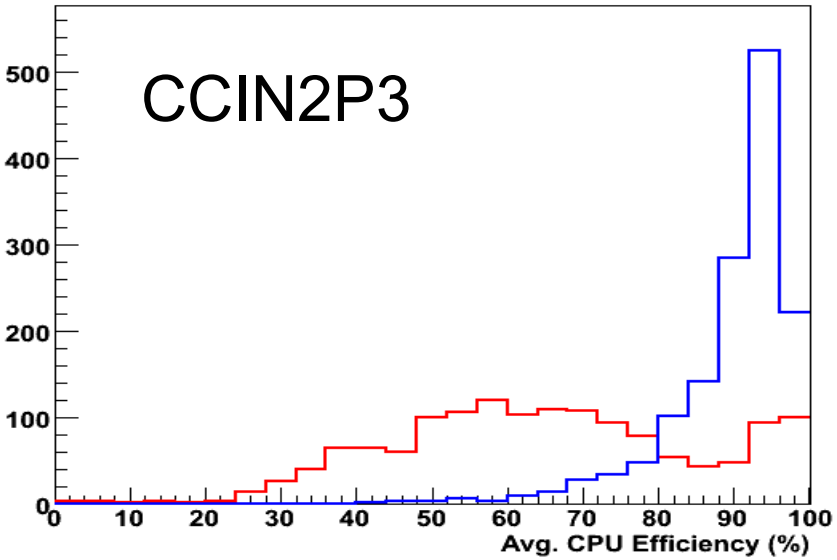


- Good performance for FZK, PIC, IN2P3, RAL, FNAL
  - This particular day was bad for FZK but in general did well

Efficiency with and w/o pre-staging

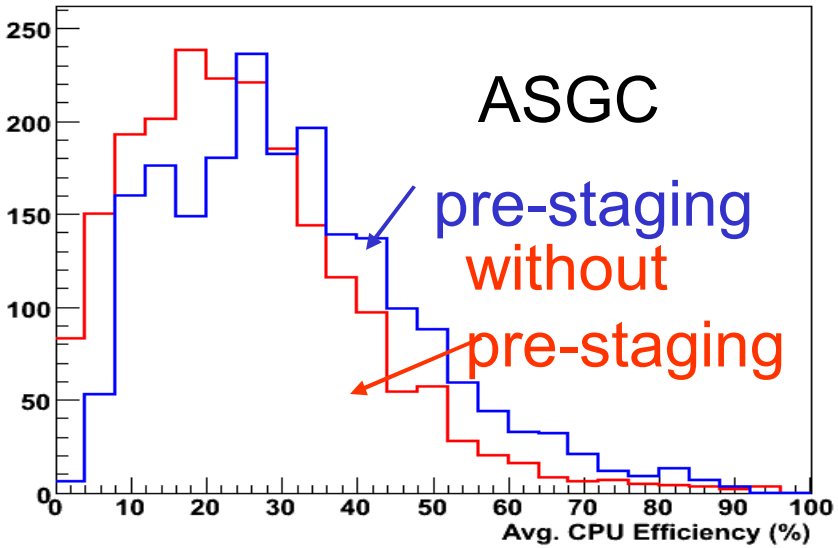


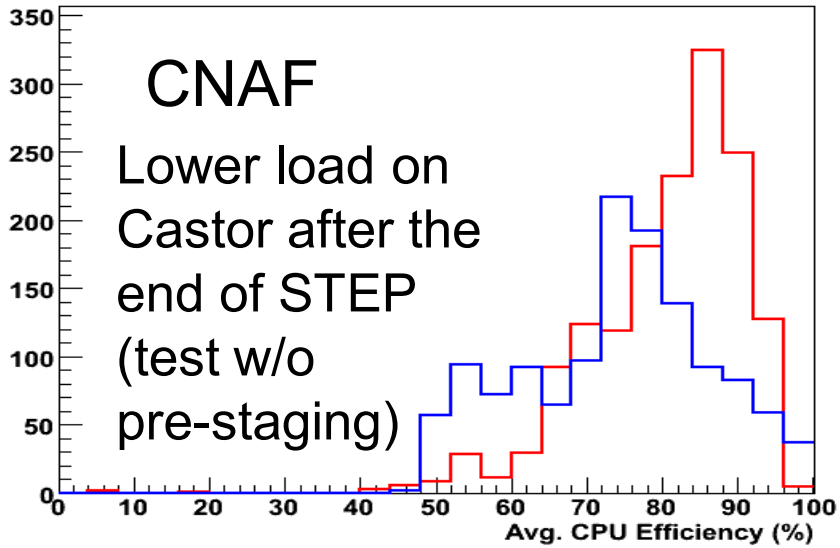


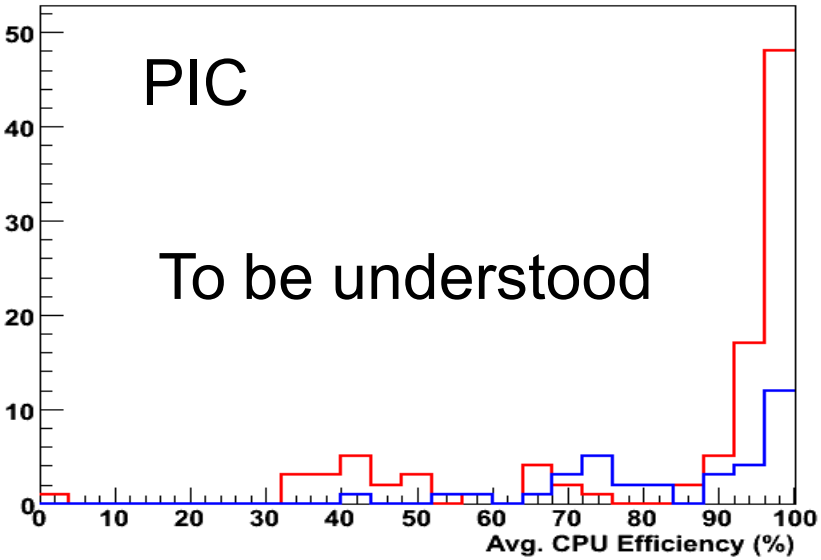


- Clear indications from the sites that had good efficiency

Efficiency with and w/o pre-staging







FZK

No meaningful data without pre-staging

To be redone

Conclusions

- All sites (but FZK) demonstrated to be able to stage data well above the required rate
  - But the competition with ATLAS was low as they were stressing in particular the disk pools, not the MSS
- Reprocessing ran smoothly and could get the expected number of slots at most sites
  - At ASGC it was difficult to get the required number of batch slots especially at the beginning
  - At CNAF and ASGC the job efficiency was low
- Sites with good efficiency when processing from disk clearly indicate that without pre-staging the performance degrades significantly
  - PIC and CNAF behavior to be understood
  - FZK test not significant