

# CMS Computing Report

16 February 2010  
LHCC Review  
Ian Fisk





# CMS Computing Model

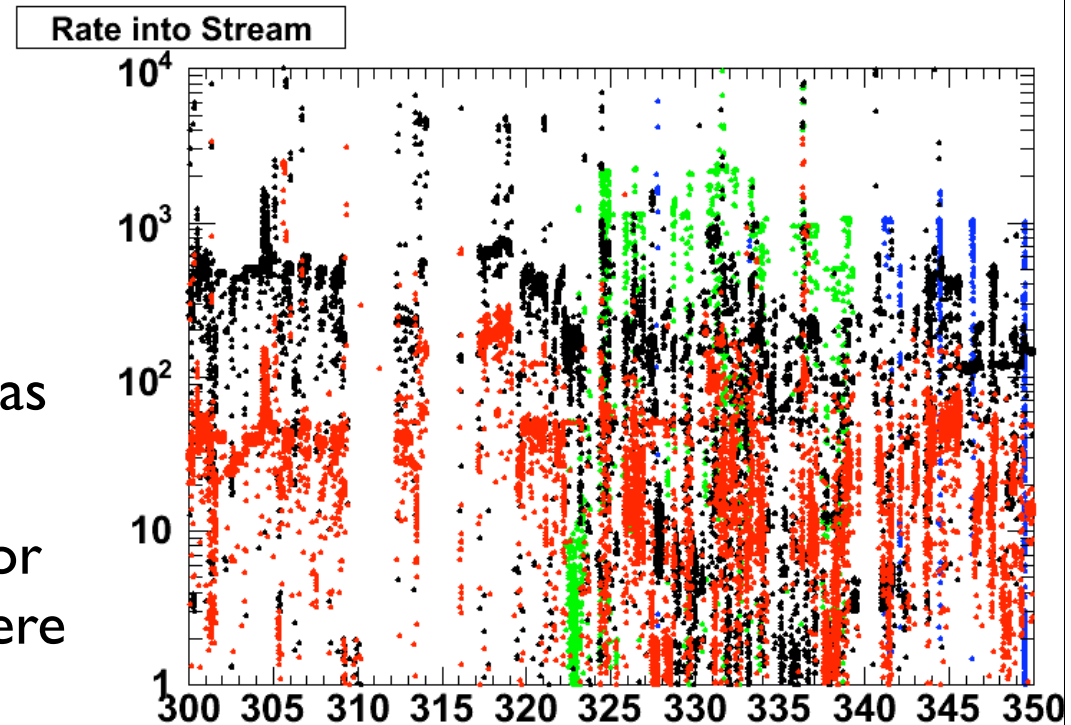
- ▶ During the fall data period CMS performed workflows and activities that were predicted in the Computing Model
- ▶ Computing Tiers performed the specified workflows
  - ▶ Tier-0 promptly reconstructed, the Tier-1s reprocessed and served data, and the Tier-2 centers were used for simulated event production and analysis
  - ▶ An additional online stream was added
- ▶ The predicted workflows were executed much more frequently.
  - ▶ Reprocessing and analysis were exercised frequently
  - ▶ Data was subscribed to many T1s and more T2s
    - ▶ Replication and processing went well
  - ▶ Event complexity and fraction of “interesting” physics events was much lower than expected in the planning



# Planning/Observation

- ▶ The Computing Model planning defined 2 periods
  - ▶ Period 1 is Oct-November 2009 to April of 2010
  - ▶ Period 2 was the remainder of 2010
- ▶ In Period 1
  - ▶ 100 days at 20% livetime (20 days)
  - ▶ 1.5MB/event RAW and 0.5RECO
  - ▶ Total Number of Events 726M
  - ▶ Total Volume of Data
    - ▶ ~1PB RAW
    - ▶ 359TB Reco
    - ▶ A few 10pb-1
  - ▶ Rate of Data from P5 450MB/s
- ▶ 2009 to Present for the Minimum Bias Sample
  - ▶ There are nearly 16k lumi sections on the RAW Minbias PDS
    - ▶ 17 days (90M events)
    - ▶ 2400 Files
    - ▶ 7.8TB
    - ▶ 10 inverse micro-barn collected
  - ▶ Stable beams with all detectors on and timed-in is smaller
    - ▶ 22 hours
    - ▶ 6.8M events (Around a 1TB)

- ▶ Data in CMS is sent from the Online as streams
- ▶ **Express** is expected to be about 40Hz. Generally stayed within 40-60Hz with occasion spikes to 3kHz
- ▶ **Stream A** is the source of the Primary Datasets (PDS). In the planning was expected at 300Hz, was 200Hz with spikes to 1kHz.
- ▶ Expectation in the planning was for 10 PDS. In the first run there were only 2 populated.
- ▶ **Stream B** was proposed before the run as insurance. It's a very high rate stream of ZeroBias Data. Averages 1kHz. **Stream B** was also buffered





# Data volume: Streams

Stream	#Events	Size [GB]
HLTMON	19,454,692	3,935.81
Express	80,478,349	12,335.44
B_Buffered	130,167,201	<b>25,478.95</b>
A	731,269,373	<b>98,467.30</b>
B	278,019,843	<b>20,111.04</b>
RPCMON	145,150,042	540.61
FEDErrors	457	0.17
Calibration	209,228,981	24,186.10
ALCAP0	40,154,649	401.21
ALCAPHISYM	253,569,603	2,488.50
OnlineErrors	89,297	24.87
	<b>1,887,582,487</b>	<b>187,970.01</b>

- ▶ Planning for period I called for about 725M events
- ▶ 770M Simulated
- ▶ The corresponding data volume per month was IPB over 6 months
- ▶ Event size and complexity of processing much lower than planned
- ▶ The fraction of interesting to taken events much much lower



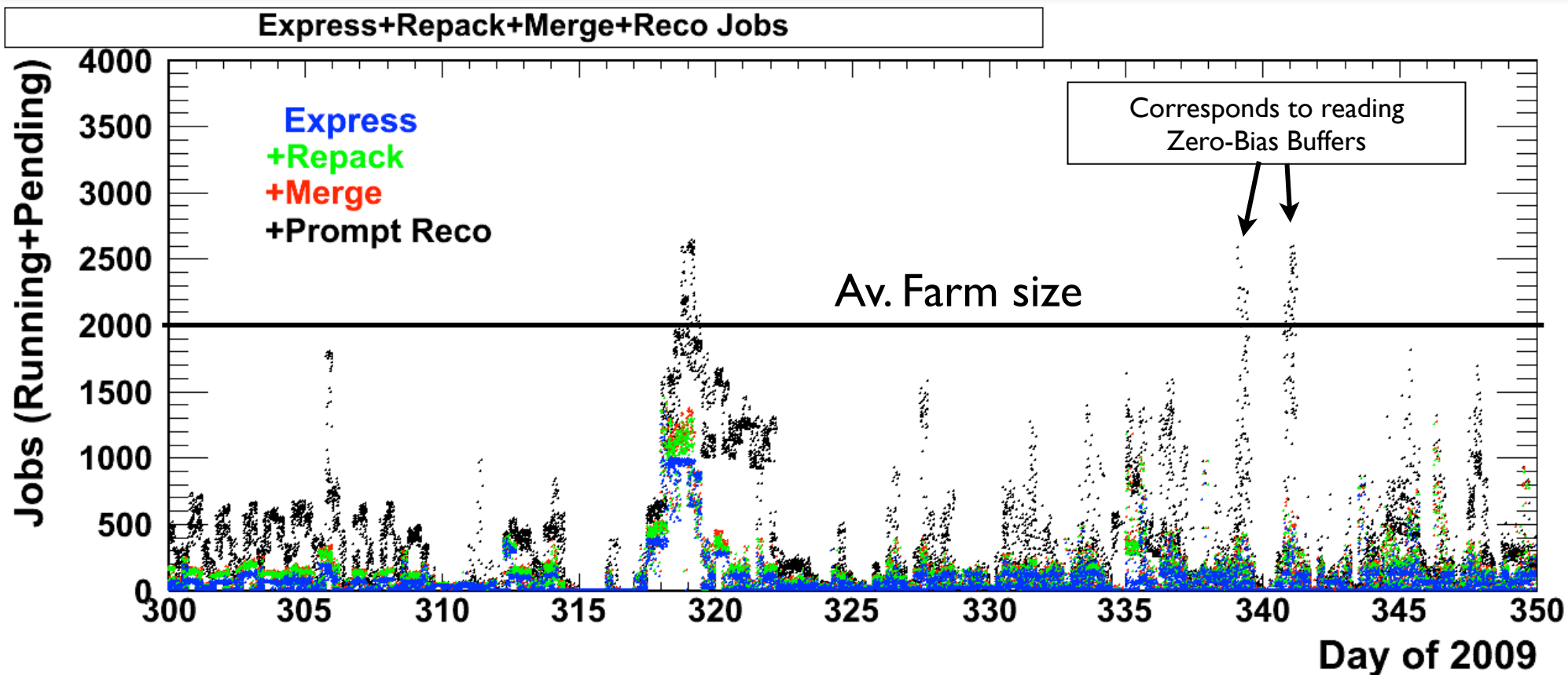
# Data volume: PDs

PD	RAW		Prompt Reco	
	#Events	Size [GB]	#Events	Size [GB]
MinimumBias	90,052,125	7,822.32	89,791,258	6,964.36
RandomTriggersO	46,969	0.04		
ρ				
RandomTriggers	47,444,572	5,441.41		
ZeroBias	78,065,537	6,445.00	78,038,521	4,671.24
ZeroBiasB	404,057,754	37,999.02	20,124,018	1,389.00
LogMonitor	86,462	19.61		
TestEnables	209,228,219	19,352.02		
PhysicsMuonBkg	91,890,670	9,411.89	67,388,127	4,536.38
BeamHalo	123,852	17.40	80,971	31.21
Test	763,109	147.33		
AlCaPhiSymEcal	253,569,603	1,896.82		
MinimumBiasNoCa	0	0.02		
l				
AlCaP0	40,154,649	328.85		
HcalHPDNoise	1,674,393	257.93		
RPCMonitor	145,150,042	219.71		
FEDMonitor	3,293	0.39		
Calo	117,967,688	14,634.55	81,452,296	4,428.22
Cosmics	407,437,569	42,558.37	363,689,277	31,803.55
HcalNZS	6,916,109	1,203.36	61,619	8.77
ZeroBiasBnotT0	4,129,290	473.47		
	<b>1,898,761,905</b>	<b>148,229.52</b>	<b>700,626,087.00</b>	<b>53,832.73</b>

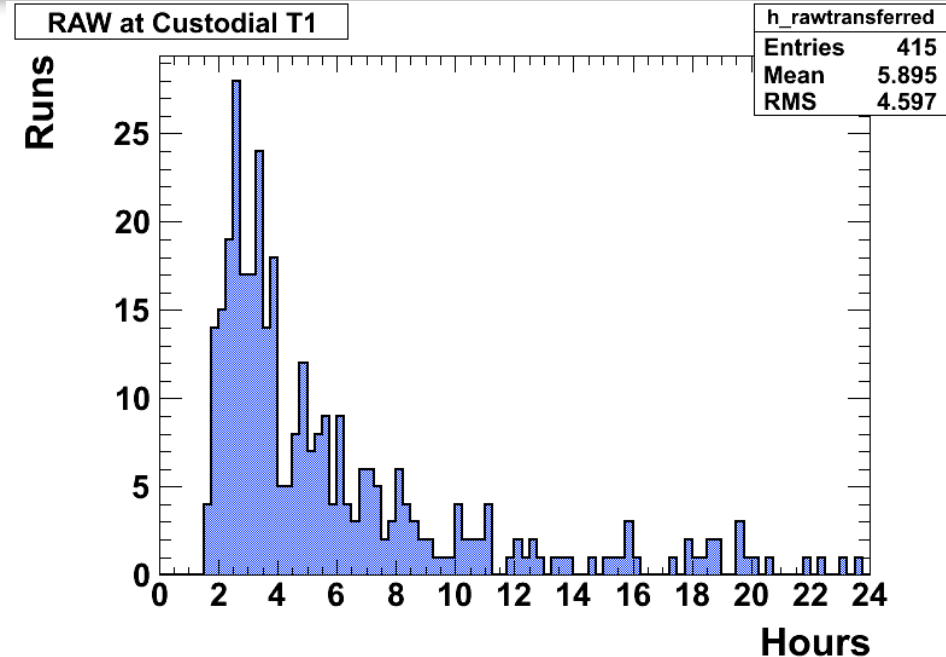
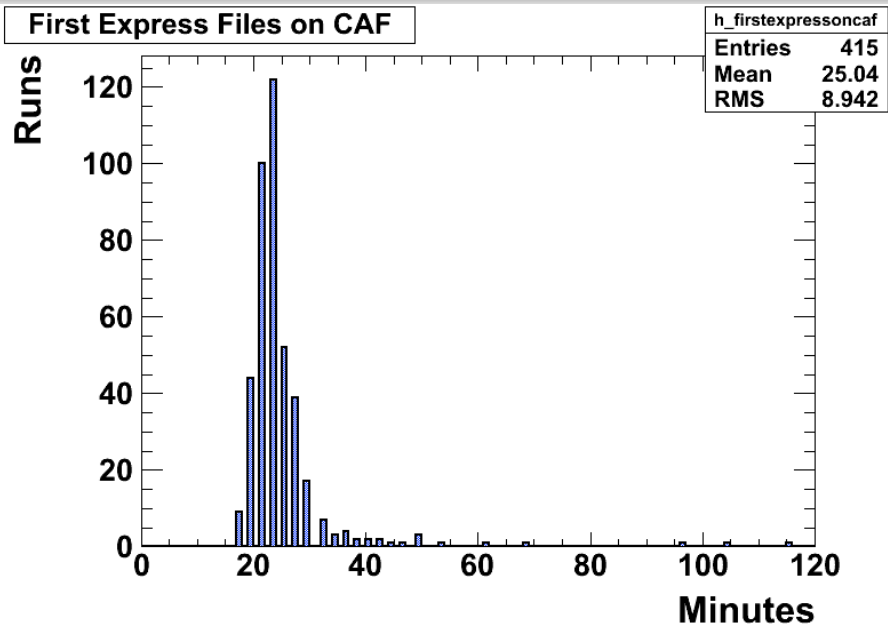
Average Event Size in Minimum Bias and Zero Bias Stream: 100k

► Agreement between real and Simulated event size and processing time for 900GeV is excellent

# T0 queue utilization



- ▶ Each of the points are cumulative: black corresponds to all jobs
- ▶ Includes pending jobs (only jobs released from queue recorded)



- ▶ Time to first express files on the CAF relative to receiving first streamers of run at T0:

- ▶ The design spec for this time is one hour.
- ▶ Average more 25 min. With very small tails

- ▶ Latency for RAW data successfully transferred to the Tier-I for custodial storage:

- ▶ Average is within the model expectations
- ▶ The tails are understood and most are caused by specific technical issues





# Data Distribution

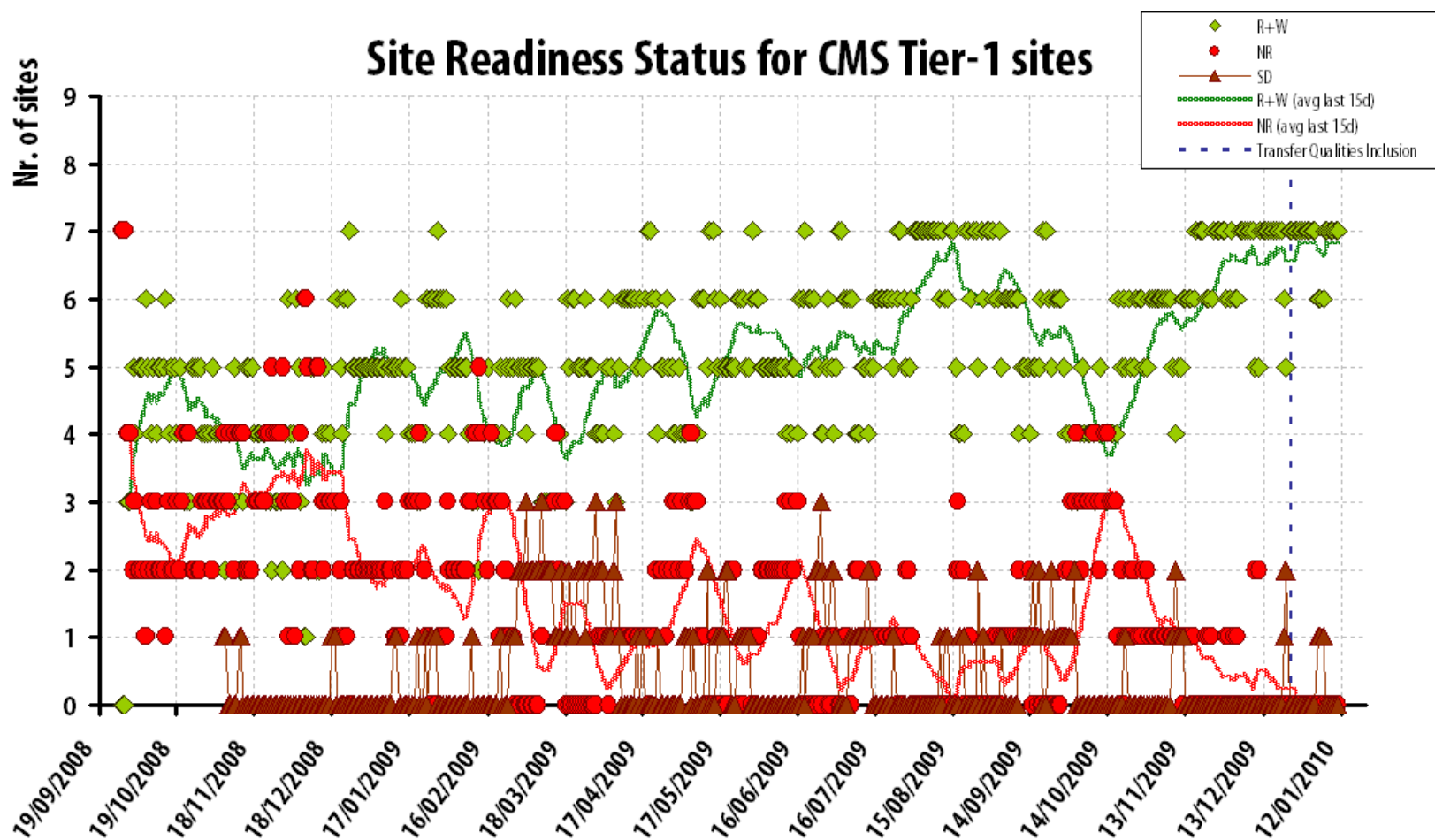
## BeamCommissioning09

Primary Dataset	Custodial Site	Other Sites	RecoSequence	Comment
AlCaP0	RAL	FNAL	dedicated ALCARECO	-
AlCaPhiSymEcal	RAL	FNAL	dedicated ALCARECO	-
Cosmics	FNAL	RAL	cosmics	-
FEDMonitor	PIC	FNAL	None	-
HcalNZS	FNAL	-	collision+hcalnzs	-
HcalHPDNoise	FNAL	-	None	-
LogMonitor	FNAL	-	None	-
MinimumBias	IN2P3	FNAL, PIC, CNAF	collision	-
PhysicsMuonBkg	FNAL	IN2P3	collision	new
RandomTriggers	PIC	FNAL	None	-
RPCMonitor	PIC	FNAL,CNAF	None	-
TestEnables	FNAL	-	None	-
ZeroBias	CNAF	FNAL	collision	new - 100 Hz in A stream
ZeroBiasB	KIT	FNAL	collision or None depending on resources	new - 1 kHz in B stream
ZeroBiasBnotT0	RAL	-	None	new

► MinimumBias was replicated to 4 TI sites in total

# Tier-1 Readiness

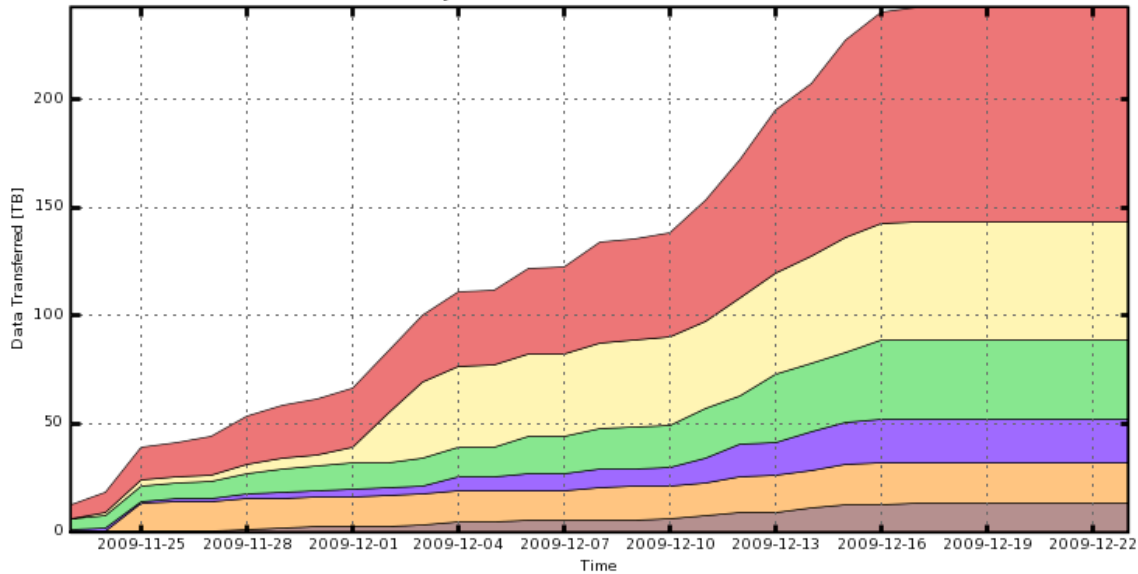
- ▶ Site readiness of the Tier-1s has improved, but CMS ran with only 6 sites receiving custodial data during the 2009 collision data
- ▶ Goal is to send data to 2 sites plus an archival copy at CERN for as long as the resources permit in 2010





# Transfers: T0→T1

CMS PhEDEx - Cumulative Transfer Volume  
30 Days from 2009-11-23 to 2009-12-23



■ T1\_US\_FNAL\_Buffer   
 ■ T1\_FR\_CCIN2P3\_Buffer   
 ■ T1\_DE\_KIT\_Buffer   
 ■ T1\_IT\_CNAF\_Buffer   
 ■ T1\_UK\_RAL\_Buffer  
■ T1\_ES\_PIC\_Buffer

Total: 242.43 TB, Average Rate: 0.00 TB/s

Site	Total Transfer Volume [TB]	Percentage	Expected
T1_DE_KIT	36.68	15	12
T1_ES_PIC	13.03	5	7
T1_FR_CCIN2P3	54.33	22	12
T1_IT_CNAF	20.23	8	8
T1_UK_RAL	19.08	8	10
T1_US_FNAL	99.07	41	40
	<b>242.42</b>	<b>99</b>	<b>89</b>

Doesn't include ASGC

- ▶ This only shows data originating at T0
- ▶ 6 T1 sites received data



# TI processing

- ▶ 5 rereco passes so far (not counting the first small one)

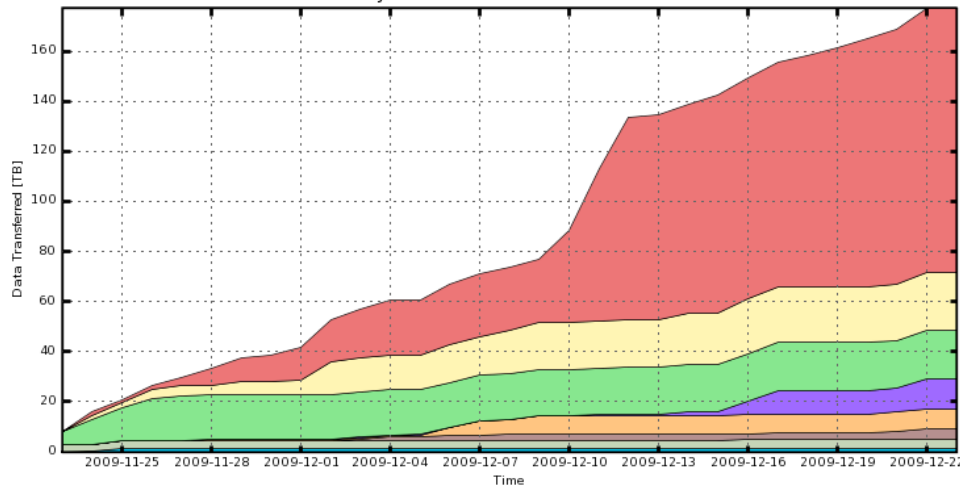
Rereco Pass	#Runs	#Events	#Input-Files = #Jobs	Rough Expectations
First rereco (GR09_P_V7)	11	998,545	76	
Dec09	14	3,499,729	144	
Dec14	37	15,885,562	701	
Dec19	52	19,681,382	925	250M Events

- ▶ total number of produced events: 150M
- ▶ total output size: 34TB (includes all data tiers)
- ▶ Latency: ~1-2 days (Planning expectations 1-2 weeks)
- ▶ Main time consumption:
  - ▶ Long running jobs (many events in input file while splitting by file to keep lumi sections intact)
  - ▶ Debugging and bookkeeping
- ▶ CPU efficiency: ~80-90% for reprocessing jobs
- ▶ Still some errors in monitoring and memory applications



# Transfers: T1->T1

CMS PhEDEX - Cumulative Transfer Volume  
30 Days from 2009-11-23 to 2009-12-23



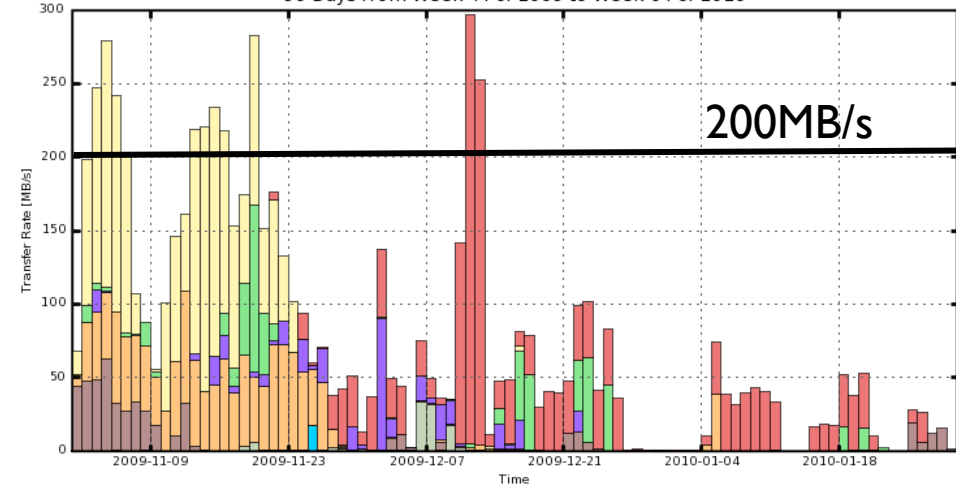
■ T1\_FR\_CCIN2P3\_Buffer   
 ■ T1\_CH\_CERN\_Buffer   
 ■ T1\_UK\_RAL\_Buffer   
 ■ T1\_US\_FNAL\_Buffer   
 ■ T1\_TW\_ASGC\_Buffer  
■ T1\_IT\_CNAF\_Buffer   
 ■ T1\_DE\_KIT\_Buffer   
 ■ T1\_ES\_PIC\_Buffer

Total: 176.92 TB, Average Rate: 0.00 TB/s

(\*) includes 2.93 TB transfers from T1\_DE\_FZK to T1\_DE\_KIT, 7.51 TB to repair samples at ASGC, 23.31 TB going to T1\_CH\_CERN

Destination Site	Total Transfer Volume [TB]
T1_DE_KIT	0.39
T1_ES_PIC	1.51
T1_FR_CCIN2P3	105.15
T1_IT_CNAF	4.55
T1_UK_RAL	19.27
T1_US_FNAL	12.29
<b>Total</b>	<b>143.16</b>

CMS PhEDEX - Transfer Rate  
90 Days from Week 44 of 2009 to Week 04 of 2010



■ T1\_FR\_CCIN2P3\_Buffer   
 ■ T1\_DE\_KIT\_Buffer   
 ■ T1\_US\_FNAL\_Buffer   
 ■ T1\_CH\_CERN\_Buffer   
 ■ T1\_UK\_RAL\_Buffer  
■ T1\_IT\_CNAF\_Buffer   
■ T1\_TW\_ASGC\_Buffer   
■ T1\_ES\_PIC\_Buffer

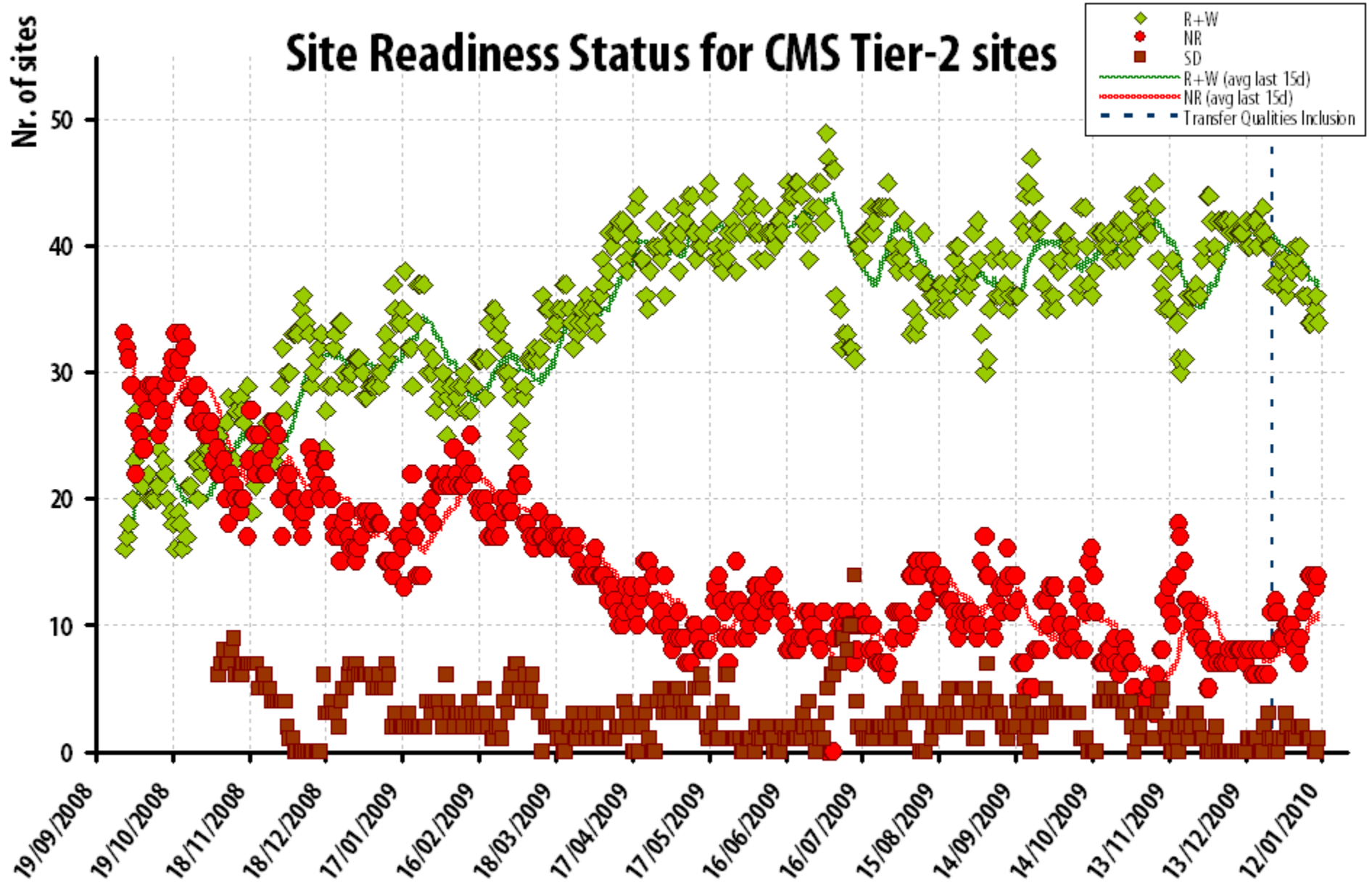
Maximum: 297.19 MB/s, Minimum: 0.00 MB/s, Average: 76.30 MB/s, Current: 1.03 MB/s

► Data transfer between T1 sites dominated by repopulation of IN2P3



# Tier-2 Readiness

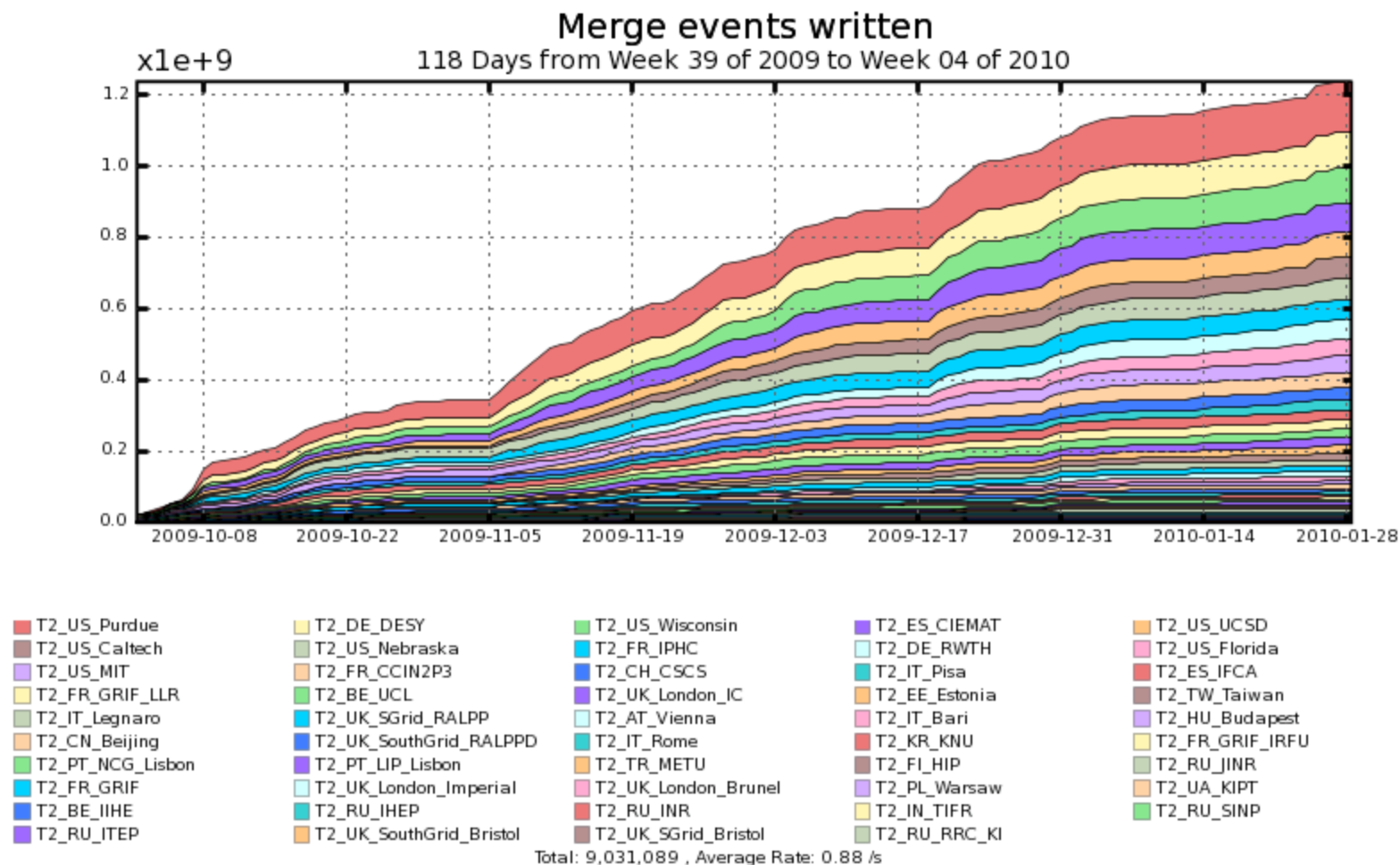
► Tier-2 Readiness has plateaued





# MC production

- ▶ Planning period I started in Oct
- ▶ 1.2B Events = ~ 400M individual simulation events roughly scales where we expected to be
- ▶ 3 months through 6 month period we have half of 780M

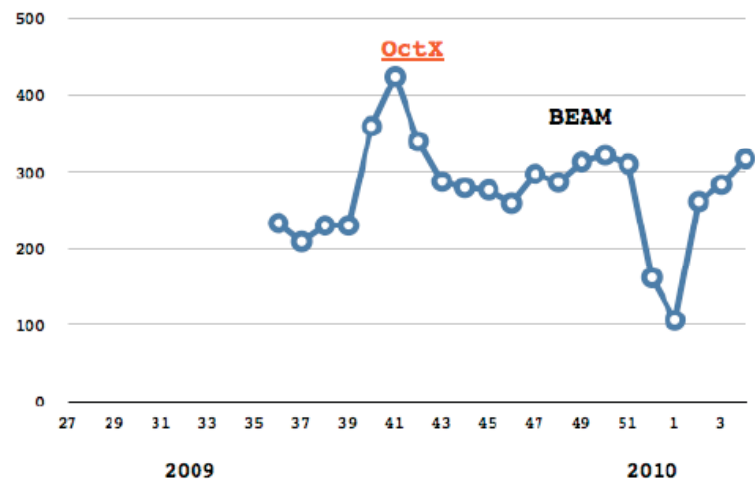




# Analysis

- ▶ Since last review CMS has formed an Analysis Operations Team
  - ▶ Provide technical support for analysis infrastructure
  - ▶ Subscribe samples to centrally controlled space at Tier-2s
  - ▶ Analysis Ops has access to 50TB of space at ~50 sites
    - ▶ Currently ~1PB of space is utilized

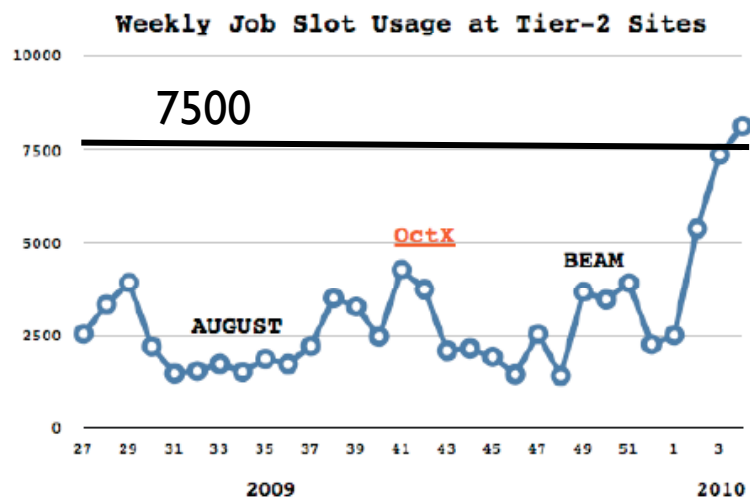
Number of Analysis Users at Tier-2 Sites Each We



- ▶ Total number of people submitting distributed analysis jobs in a given week ~300
- ▶ Bump after the October analysis exercise

Type of Data	Total Size (TB)	All Replicas		AnalysisOps Requested	
		Size (TB)	Sites	Size (TB)	Sites
Collision Data	446.9	688.8	29	218.0	12
900 GeV MC	44.8	159.8	32	70.7	14
2.36 TeV MC	31.1	72.6	24	48.7	10
<b>All Data Sets</b>	<b>9.3 PB</b>		<b>50</b>	<b>1.04 PB</b>	<b>38</b>

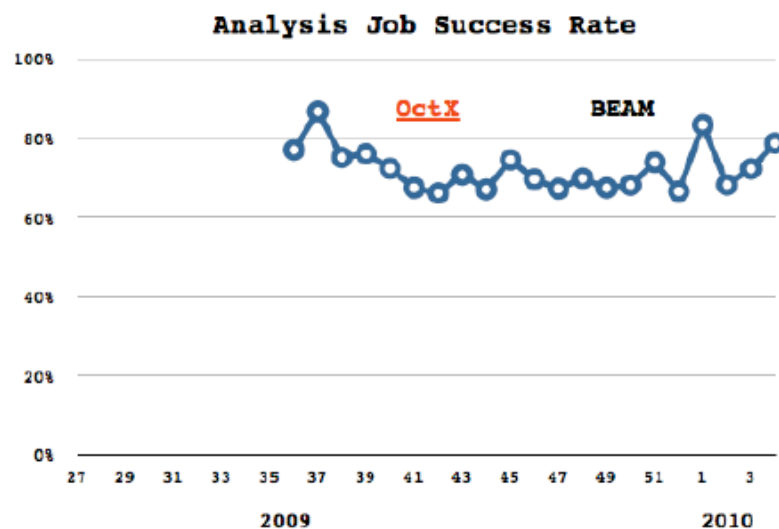




▶ Roughly 11k jobs slots are available for analysis

▶ Reaching 75% utilization toward the beginning of the year

▶ In any given week 47+/-2 Tier-2 sites have analysis jobs



▶ Success rate remains a persistent issue

▶ Improvement over last year where we had ~65%

▶ Half of errors are related to remote stage-out of produced files

- ▶ Activities like MC and Analysis that are driven by external factors are making reasonably high use of the available resources
  - ▶ Analysis is currently running at 75% level
  - ▶ MC roughly on planning
- ▶ Activities like Re-reconstruction and skimming that are driven by available data are not fully utilized
  - ▶ Data Volume is much lower than planned for
  - ▶ Transfers lower on average. Good peaks. Partially compensated by over subscription
  - ▶ Tier-I utilization for activities like Cosmic reprocessing was high. On average lower than planned.



# Looking Forward

- ▶ New preliminary 2010 guidance is more seconds of livetime than 2010 from last year's guidance
- ▶ More data, but the available resources are unchanged
- ▶ Items that scale with instantaneous luminosity appear to be well matched with the available resources
- ▶ Tier-0 and CAF look reasonable in 2010
- ▶ Items that scale as total number of seconds of data and total events will be squeezed
- ▶ Particularly reprocessing at Tier-1s and Simulated Event production at Tier-2s
- ▶ Adjusting the activities to match the resources
- ▶ **All 2011 Predictions are preliminary and in the process of being checked**

Some Relevant Parameters		
Year	2010	2011
Trigger(Hz)	300Hz	300Hz
PDS Overlap	1.4	1.2
RAW(MB)	1.5	1.5
RECO(MB)	0.5	0.5
AOD(MB)	0.2	0.2
RECO(HS)	100	140
SIM(HS)	828	970
Days for Re-RECO	60	90



# Tier-0 2010-2011

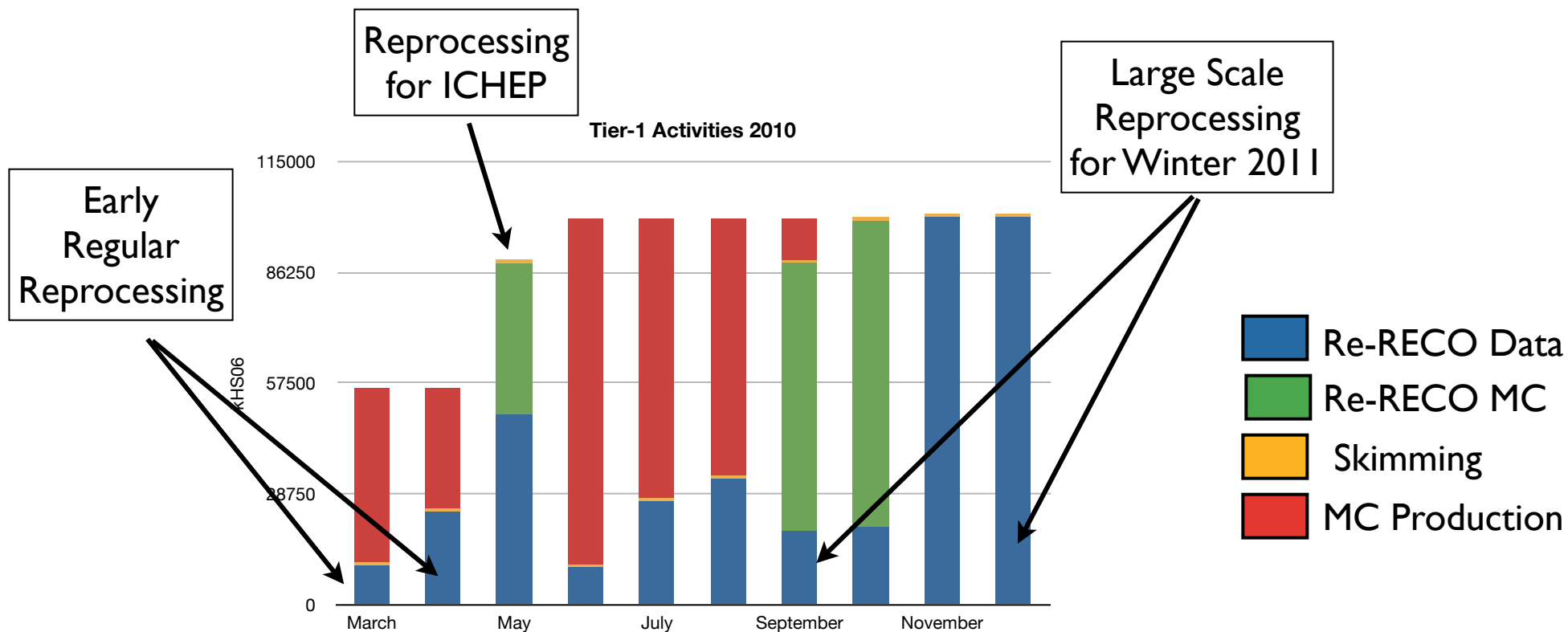
## ▶ Tier-0

- ▶ Matches well to guidance because the elements that drive the calculation are unchanged
- ▶ Prompt Reconstruction measured similarly to last year and the Computing Model
- ▶ Trigger rate is the same

- ▶ Looking forward to 2011, we are trying to get an accurate estimate for the prompt reconstruction time as a function of luminosity
- ▶ CMS is clearly sensitive to this and the reco time will be longer

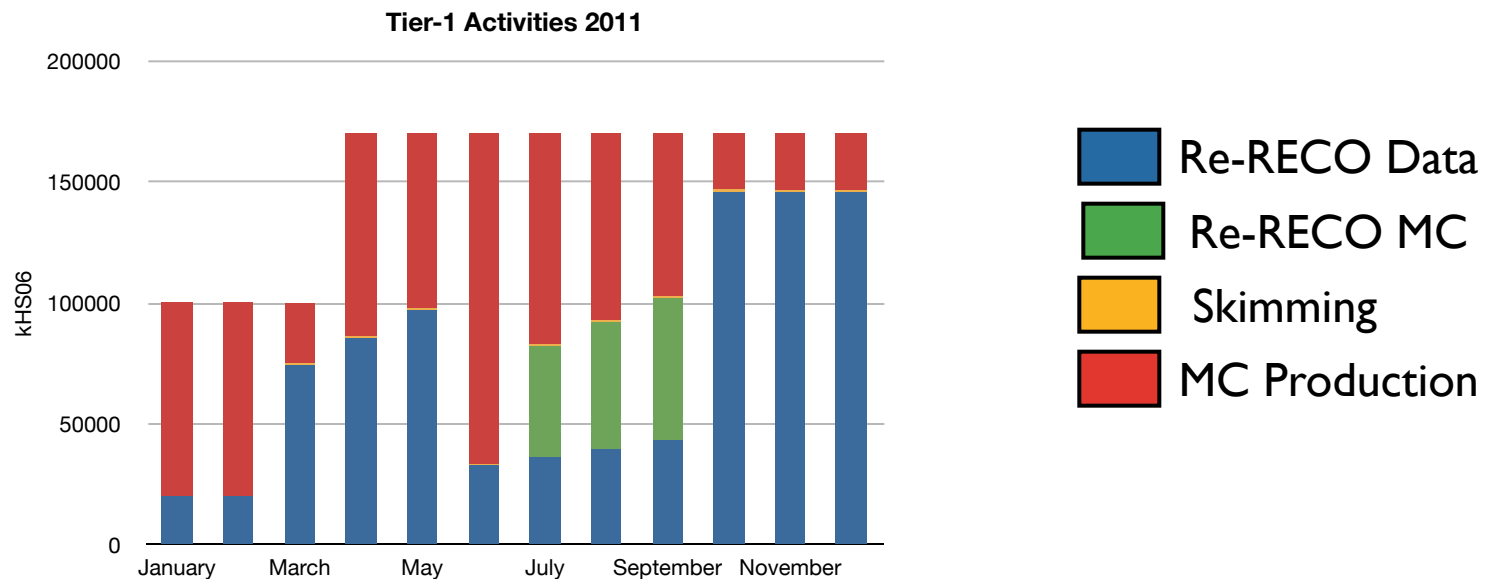
Tier-0 CPU	2010	2011
Express	5333	7467
Prompt RECO	42000	50400
Repack	2000	2000
Alca Workflows	630	540
ReVal	6000	6000
VOBoxs	6000	7000
Total	63973	73407
Request	61900	

- ▶ The Tier-1s have 3 main periods of activities



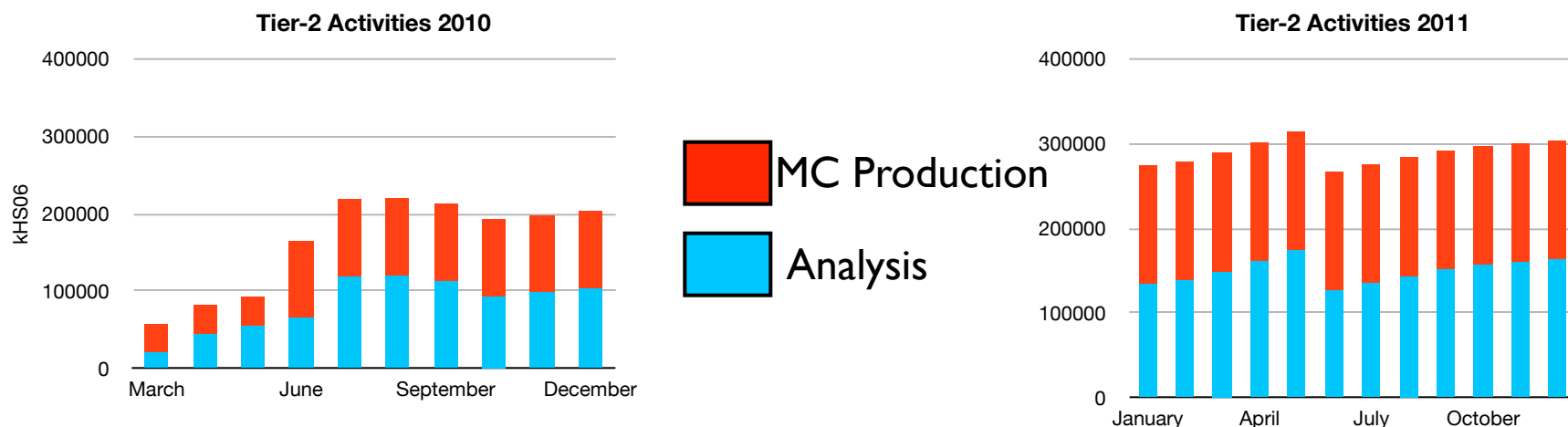
- ▶ Between these use Tier-1 resources for MC Production to make up for shortfall in capacity at Tier-2s
- ▶ To accommodate the increased simulation need with increased number of events in 2010 some simulation will be moved to Tier-1s
- ▶ Averaged over the year it's less than 40% of the Tier-1 capacity.

- ▶ In 2011, the total number of events increases, but the time for a reprocessing pass also increases
- ▶ Two passes through the data before major conference seasons



- ▶ Time between reprocessing is used for simulated event capacity
- ▶ A 60-80% increase is needed at Tier-1s

- ▶ CMS is striving to maintain a balance between analysis and simulation at the Tier-2s
- ▶ Using the model used previously for analysis CPU we can almost maintain 50% analysis and 50% simulation at the Tier-2s (if Tier-1s contribute)



Relies on a migration to summarized data as outlined in the CMS ECoM report

- ▶ Roughly a 50% increase in the Tier-2 CPU is required in 2011 to accommodate the data analysis and MC production needs



# Heavy Ion

- ▶ CMS now has a detailed Heavy Ion Computing Model
  - ▶ Data is reconstructed initially at the Tier-0
    - ▶ Estimates for 2010 and 2011 are 4 days and 18 days at the Tier-0
      - ▶ Both should fit in the existing Tier-0 request
  - ▶ Data is transferred to FNAL for archiving
    - ▶ ~200TB of tape storage
  - ▶ Data Reprocessing will be performed at a new Tier-2 facility in Vanderbilt
    - ▶ 3200 HS06 2010 and 5800 HS06 2011
  - ▶ Simulation and Analysis will be performed at a combination of 5 Tier-2s
    - ▶ 9000 HS06 Total in 2011 growing to 10500 HS06 in 2011





# Outlook

- ▶ CMS is looking at approximately 1M times more integrated luminosity by the start of the summer
  - ▶ 10 inverse micro-barn to 10 pb<sup>-1</sup>
- ▶ While many elements of the computing model accurately reflect the activities and the experience,
  - ▶ We have little experience with a resource constrained system
  - ▶ We don't have experience with large quantities of very interesting physics events
  - ▶ We are hoping for a huge increase in data volume.