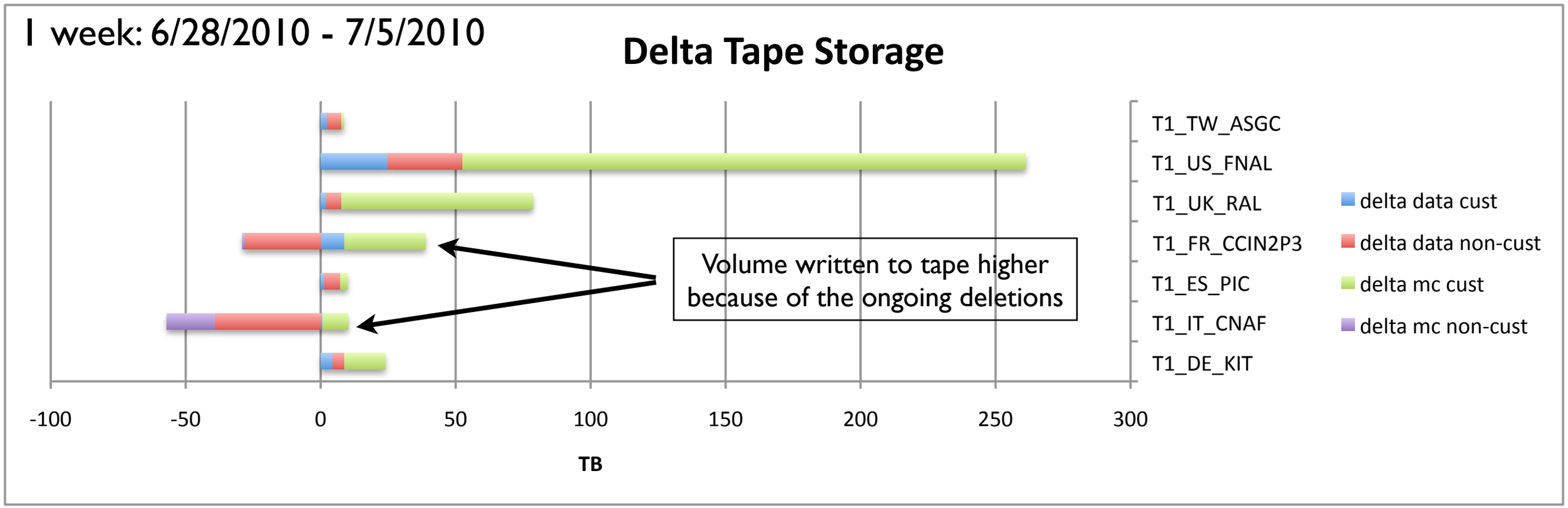
Tier-1: Workflow Efficiency and IO

WLCG Collaboration Workshop 7-9 July 2010
CMS Experiment Jamboree

Oliver Gutsche
for
CMS Data Operations

Introduction



- In preparation for ICHEP, CMS produced large volumes of data and MC and wrote them to tape at T1 sites
- We got reports from sites that at the current scale we are running into tape migration backlogs
  - These backlogs affect how fast and efficiently data from the detector can be archived and disturb the whole site operations
- In the following, we want to analyze how much data or different workflows are writing
- With this information and the site tape writing capabilities, we can determine new working points in terms of parallel jobs for sustained operation

Workflows

- Primary Tier-1 workflows
  - Data re-reconstruction
  - Data skimming
  - MC re-digitization / re-reconstruction

Data Re-reconstruction

Input:
RAW input data dataset

Outputs (separate datasets):
RECO
ALCARECO (multiple)
DQM

MC re-digitization / re-reconstruction

Input:
RECO input data dataset

Skimmed Outputs (separate datasets, depends on configuration):
RECO
RAW-RECO
USER

MC re-digitization / re-reconstruction

Input:
GEN-SIM-RAW input MC dataset

Outputs (separate datasets):
GEN-SIM-RAW (new)
GEN-SIM-RECO
AODSIM
DQM

Procedure

- Investigate how much data the different workflows produce per executable/core and per sec.

- Use success tarballs from the ProdAgent to extract properties from FrameworkJobReport per successful processing job
  - Sum of size of all output files
  - Job length (AppEndTime-AppStartTime)

- Important:
  - Script gives average of processing jobs writing into unmerged area
  - Everything has to be merged afterwards if not written directly to merged
    - This effect is neglected

Results

Average output per executable [MB/s]

Label	Workflow	Input	Output [MB/s]
Rereco	Jun14thReReco	Commissioning10 MinimumBias RAW	**0.183**
	Jun14thSkim	Commissioning10 MinimumBias Jun14thReReco RECO	**0.207**
SD/CS Skim	SD/CS Jun14thSkim	Commissioning10 MinimumBias Jun14thReReco RECO	**0.412**
	GOODCOLL Jun9thSkim	Commissioning10 MinimumBias Jun9thReReco RECO	**0.322**
Redigi/ rereco	S09 Redigi/rereco in 362	Summer09 QCD_EMEnriched_Pt20to30 RAW	**2.109**

- Re-digitization / re-reconstruction writes out the most data volume per sec. per core
  - Need to discuss with Physics/Offline if writing out RAW is actually needed
    - Could reduce the output volume significantly
- But also other workflows are writing out data at a higher rate than naively expected

Projections

Output projections

Slots	Output in 1 hour [TB]			Output in 24 hour [TB]			Rate [MB/s]		
	ReReco	SD/CS Skim	Redigi/rereco	ReReco	SD/CS Skim	Redigi/rereco	ReReco	SD/CS Skim	Redigi/rereco
500	0.31	0.71	3.62	7.54	16.97	86.89	91.50	206.00	1,054.50
1000	0.63	1.41	7.24	15.08	33.95	173.78	183.00	412.00	2,109.00
2000	1.26	2.83	14.48	30.16	67.90	347.55	366.00	824.00	4,218.00
6000	3.77	8.49	43.44	90.47	203.69	1,042.66	1,098.00	2,472.00	12,654.00

- Numbers speak for themselves

- Todo for DataOps:
  - Use writing performance of tape systems at Tier-1 sites to determine working points
    - Running a single workflow alone
    - Combining different workflows
    - Taking into account necessary parallel tape activity
      - Archiving data from Tier-0, MC from T2 level
      - Reading back data/MC from tape