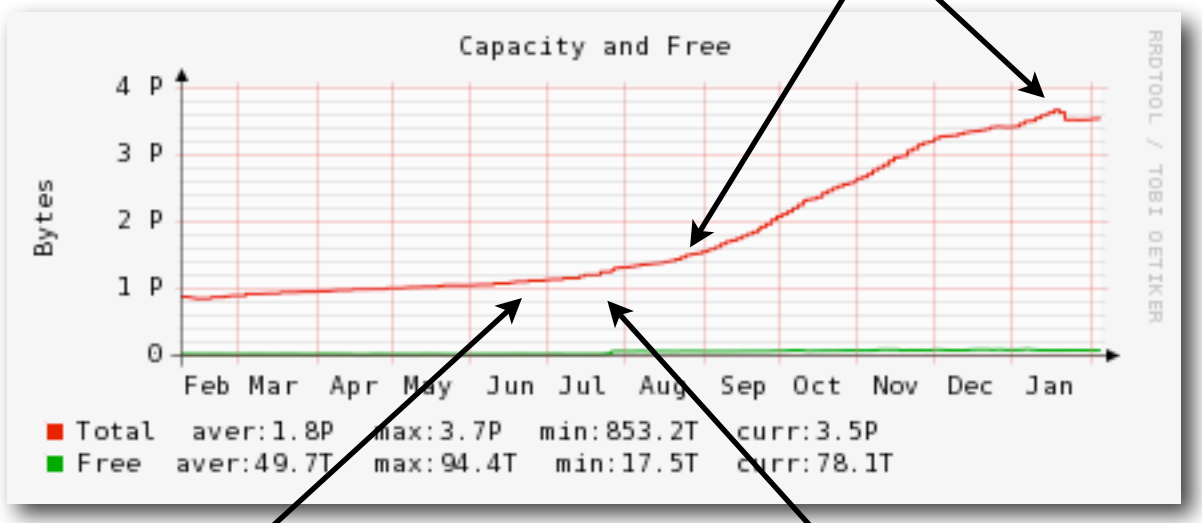
Tape effic. at Cern: CMS production system

Mike Miller (MIT)
for the CMS Tier0 team

CMS: Heavy user of cern tape system



Outline

- Brief overview of cms tape use cases at cern
  - "Beam on"
  - "Beam off reprocessing"
  - current pre-beam activities
- Some "feedback" on production experiences with tape
- Plans to optimize tape writes, reads, deletes

“Beam on” paradigm

- CMS computing model calls for “write-only” mode during data taking
  - Tier 0 exclusively used for calibrations and prompt reco
  - => 500 MB/sec written to tape from t0export pool
  - “prompt” export of raw/reco data to Tier1
    - retrieve from tape to t1transfer pool only as fallback to recover from transfer problems
- Only production and transfer systems will use tape
  - no user analysis, no grid analysis
  - (except CAF, which is currently being defined)

"Beam off" reprocessing mode

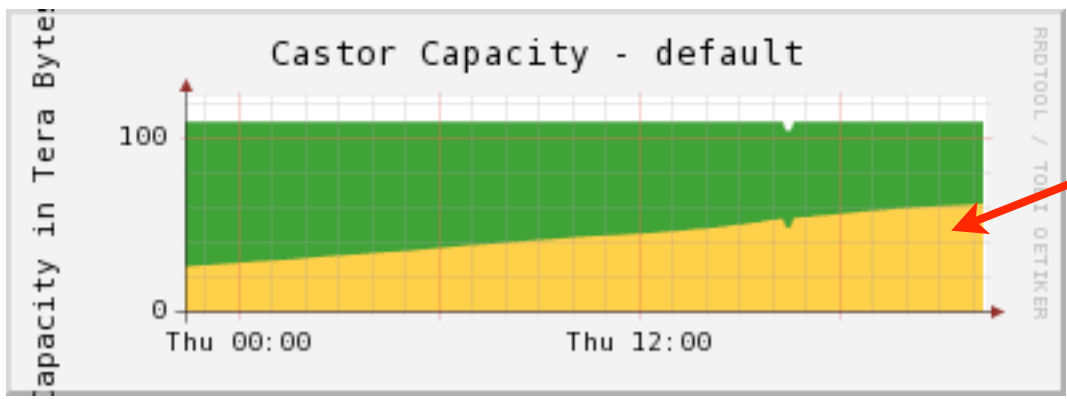
- CMS computing model does not explicitly plan for re-reconstruction at CERN
  - but we will most certainly do it in the early running
- Pre-stage driven processing:
  - prestage a dataset chunk (~50-100 TB)
  - process, wipe, iterate
  - We currently use Tier 0 in this mode
  - current "manual" procedures being automated
- Never, never launch job w/o prestaging data!

Current pre-beam setup

- Tier 0 production system in reprocessing mode
- Transfer system:
  - trickle MC imports from offsite
  - trickle exports, many from tape. Working on optimizing pre-staging
- Users:
  - supporting high rate of local and grid job submission
  - No pre-staging in analysis jobs. Jobs no nothing of tape system
  - We need to start weaning users away from cern
- Users/transfer probably dominate tape mounts
  - at least for reads

e.g., last Thursday on cms default pool

- default pool flushed last Wednesday
- pool used for CRAB analysis jobs and most local analysis, not used by production/transfer system



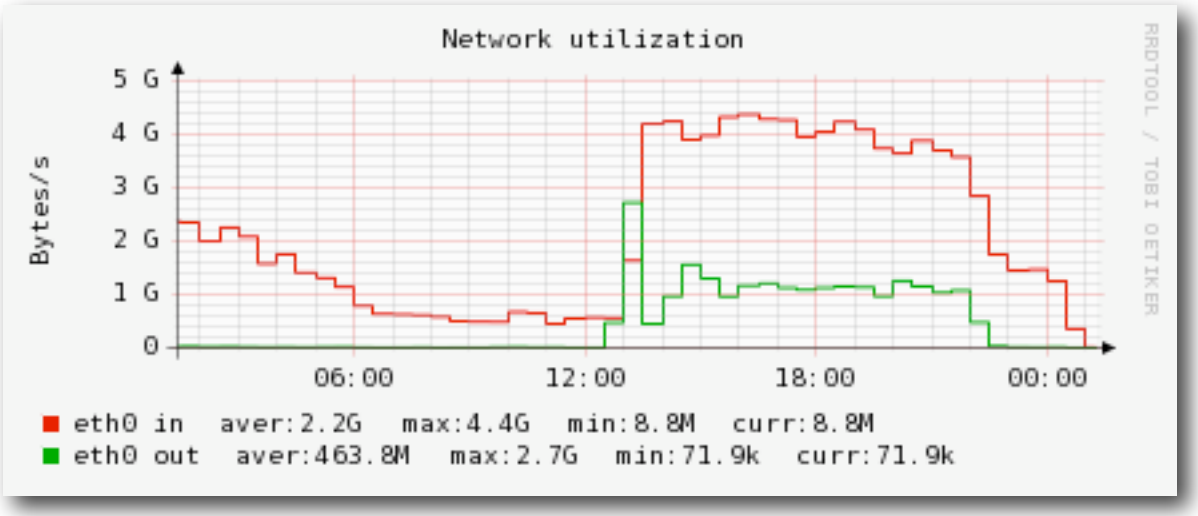
40 TB / day

- Predominantly user recalls from tape (judging by several angry emails I received ;)
- I predict cms tape mounts would plummet if we turned off users, leaving only production/transfer system

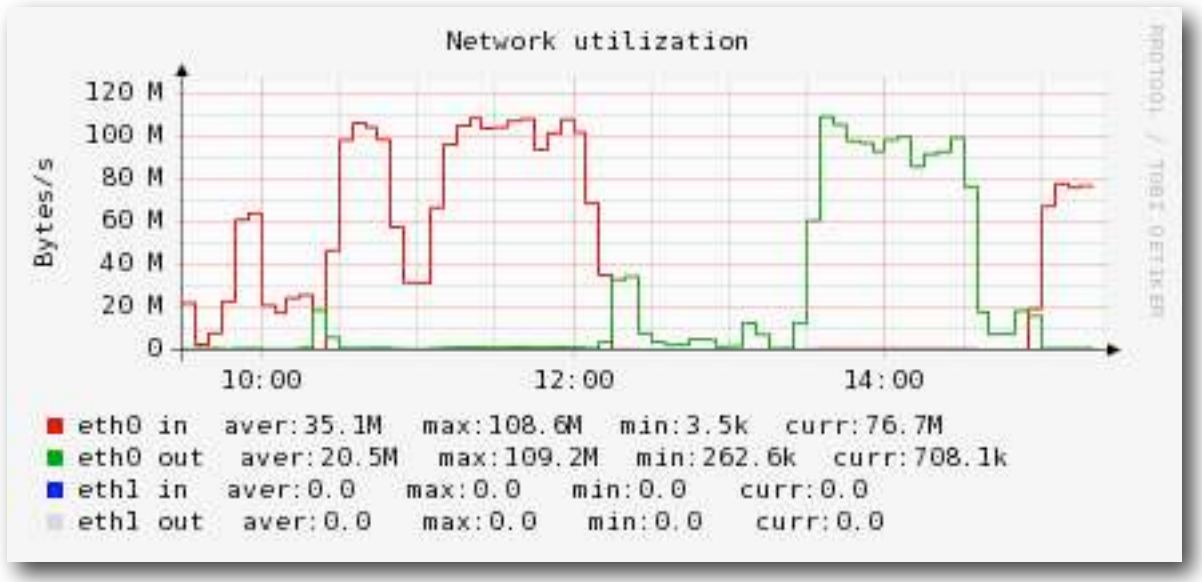
Feedback on production experiences

- Tape readback and tape write have been highly efficient, not limiting factor in production
  - Typical cms merged file is 2 GB,
  - typical CSA07 dataset is 10-80 TB
  - 10k-40k files / dataset
- Prestage process:
  - Sort pfn's by tape, issue stager_get for each pfn
  - Limiting factor in stager get is callback per pfn
  - Have to sort requests by tape to get all requests for a tape to the stager before it begins tape mounts

Feedback on production experiences

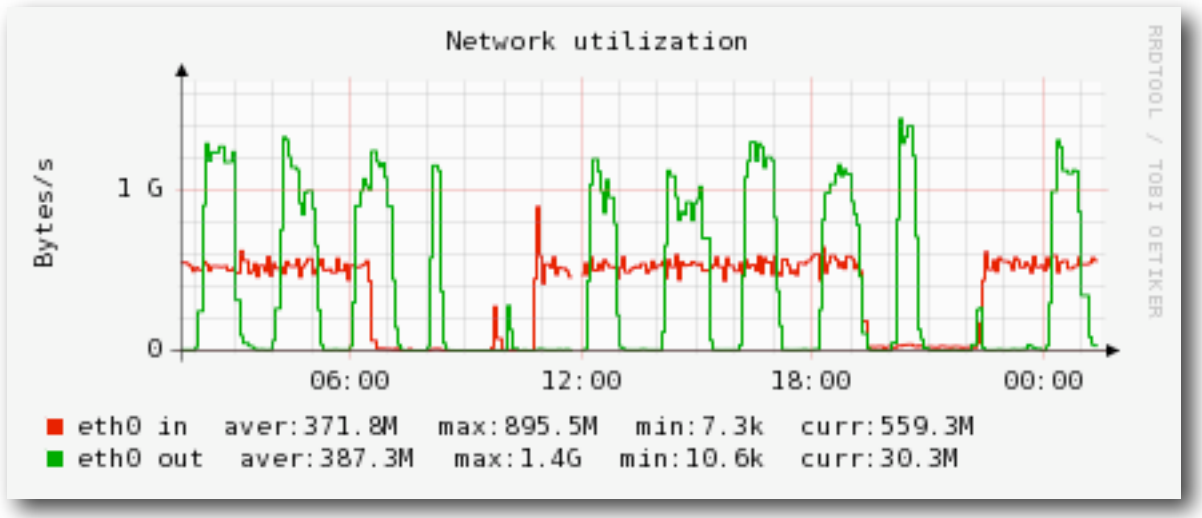


Stagein of 81 TB from 41k files in just over 6 hours



"It is interesting to see how fast the drives can go with good size files, once they are running at full speed. Here is your recent transfer which had some good size files in it." (Tim Bell)

Feedback on production experiences



- CCRC08 test setup pushes 8*3.4 GB files / 37 seconds to t0export to hit nominal 500 MB/sec rate (will scale test to 1+ GB/sec)
- migHunter runs every 2 hours, clearing 3.6 TB/hour to tape (1 GB/sec)
  - will slowly raise rate throughout this week to find upper limit

Plans for file organization optimization

- CSA07 Lesson:
  - We have to organize files at tape write to optimize
    - tape read -- we access files from same "dataset" together
    - tape delete -- we delete files from same "dataset" together
- Towards atomic "chunks" of homogenous files
  - informal discussions with castor team lead to
  - formal discussion in recent Lyon data management workshop
  - Goal: work with castor dev team on migration strategies
  - "dataset" concept defined at file creation time
  - Blend of LFN, tape class, file class, and migration logic
  - Rectify cms fileblock with tape

Summary

- Moving towards exclusively organized production at cern
- Statistical analysis of tape usage would benefit from correlation with user and/or disk pool
- Tape system handles 500 MB/sec nominal write rate
- Tier 0 reprocessing not limited by tape stagein time
- But work to be done if we ever hope to delete files from tape