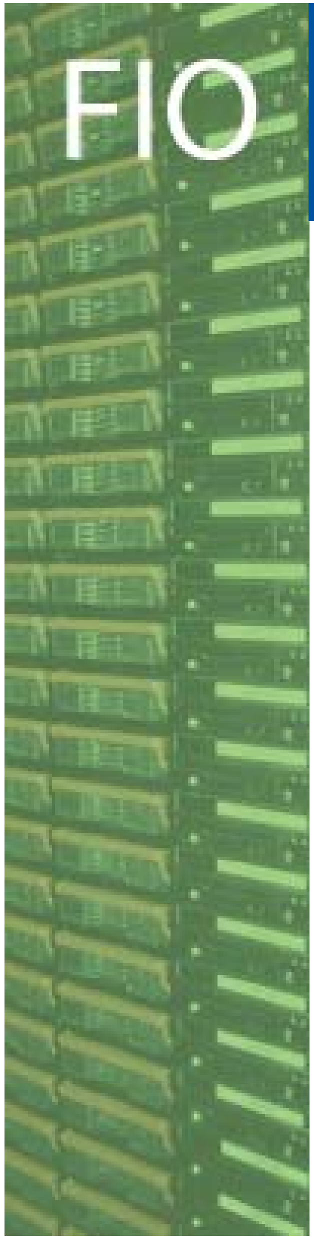


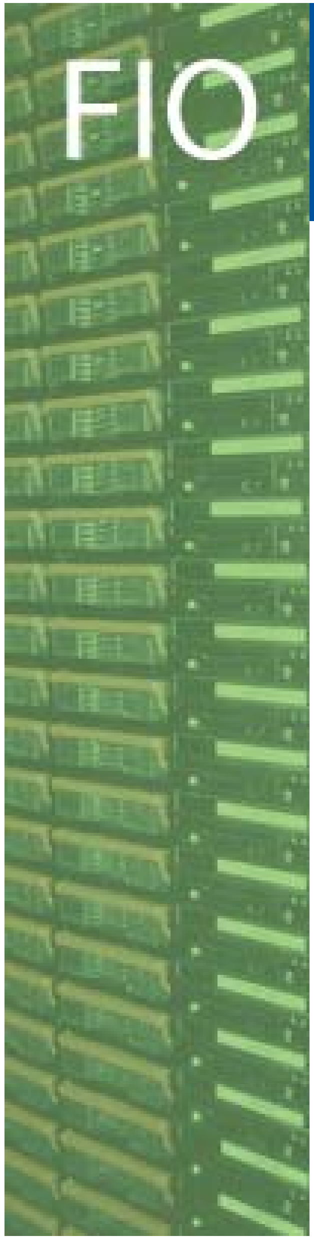
Fabric Infrastructure
and Operations

CERN IT
Department

Tape Efficiency

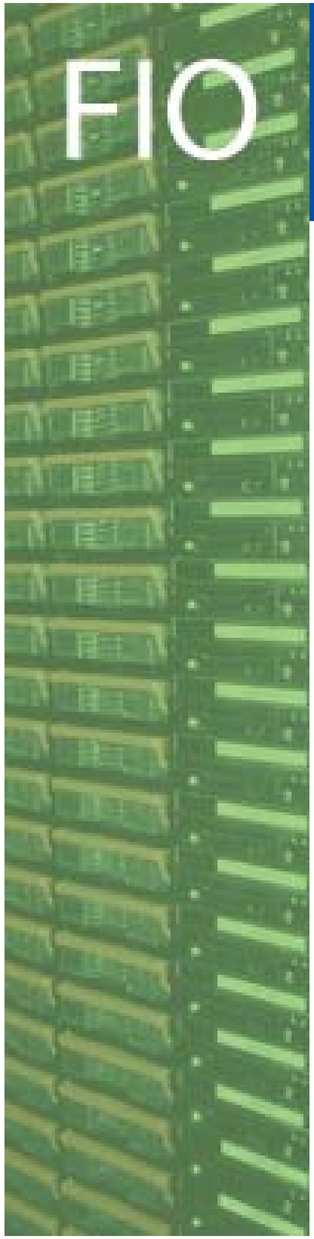
Tim Bell
January 2008

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it



We have a real problem

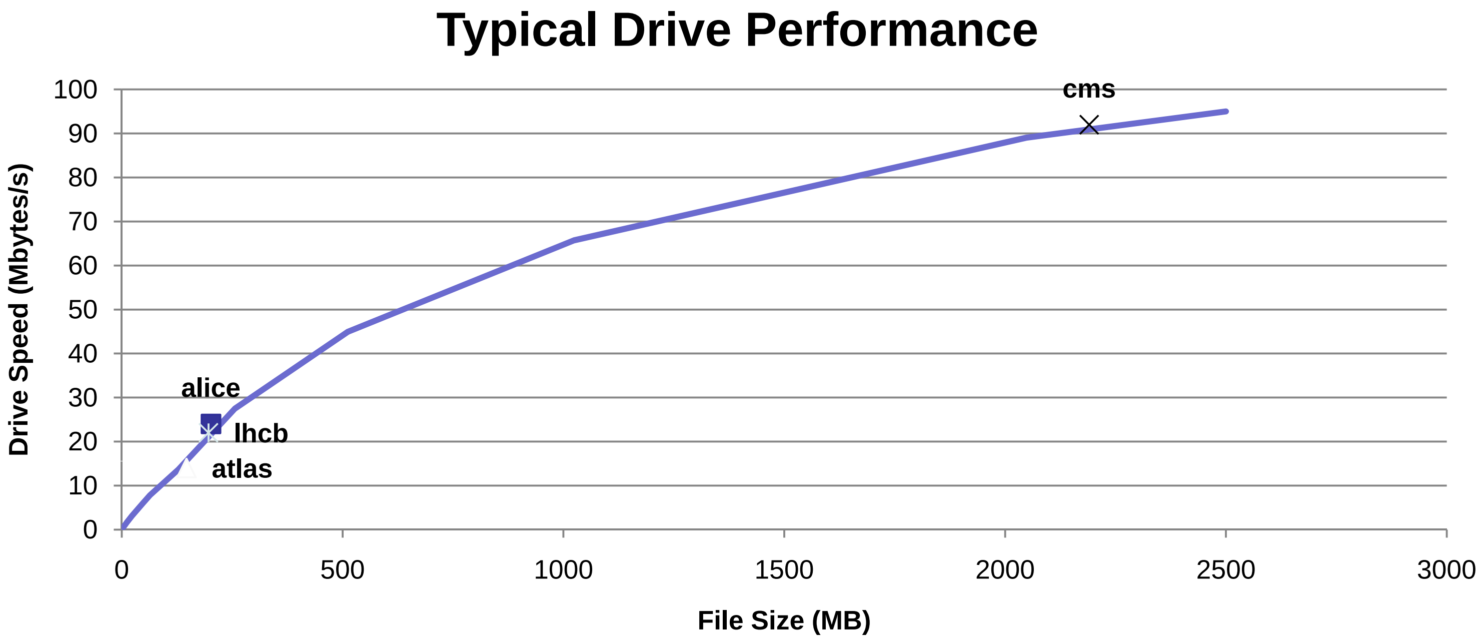
- User complaints
  - Long stage-in time during challenges
  - Data on tape unavailable
- Low batch efficiency
  - Long queues waiting for tape data staging
  - CPU jobs waiting for tape data to be read
- High failure rate of robotics
  - Drives and robot arms require maintenance
  - Tapes are often disabled needing repair



Analysis

- Data collected during Nov/Dec 2007
  - Distribution of file sizes on tape
  - Tape mounts and performance
  - Production tapes only (no user tapes)
- Root causes identified
  - Small file sizes
  - Repeated mounting

File size and performance



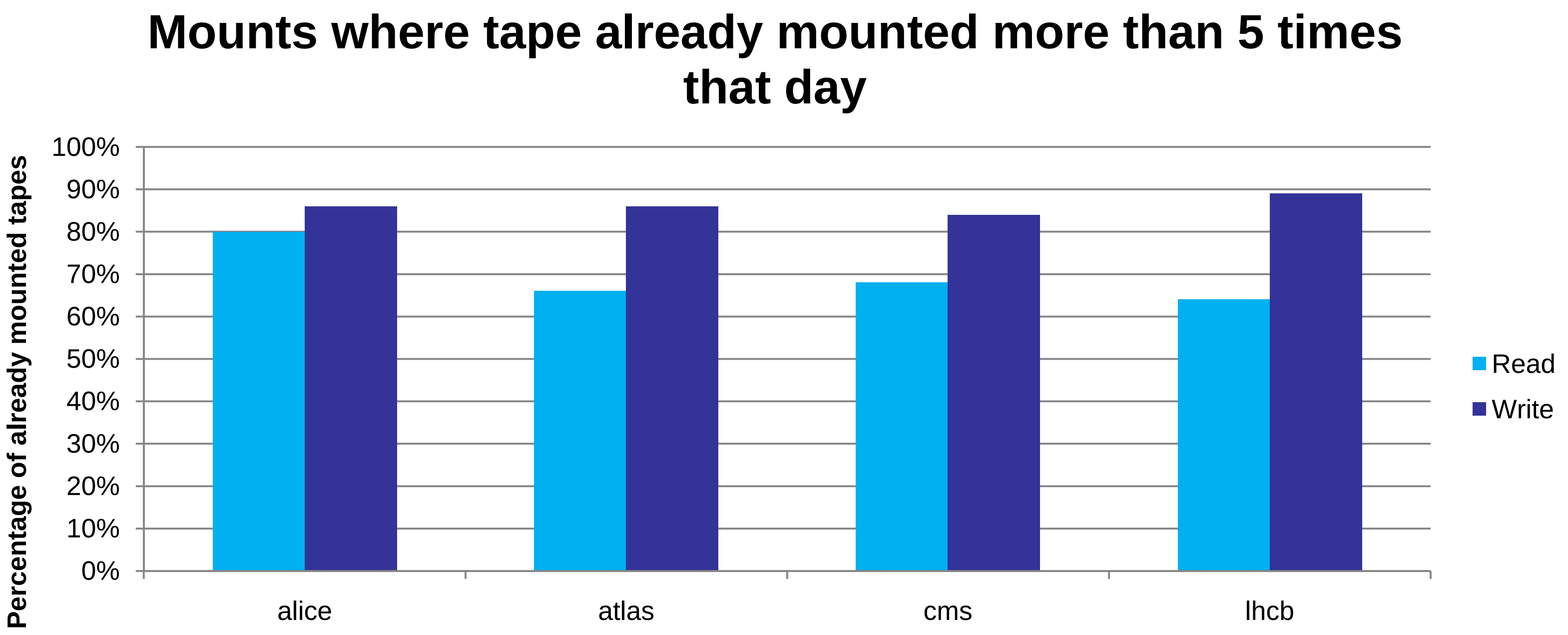
Alice	Atlas	CMS	LHCb
200 MB	150 MB	2200 MB	200 MB

•Tape drives need to stream at high speeds to achieve reasonable performance.

•Per-file overheads from tape marks lead to low data rates for small files

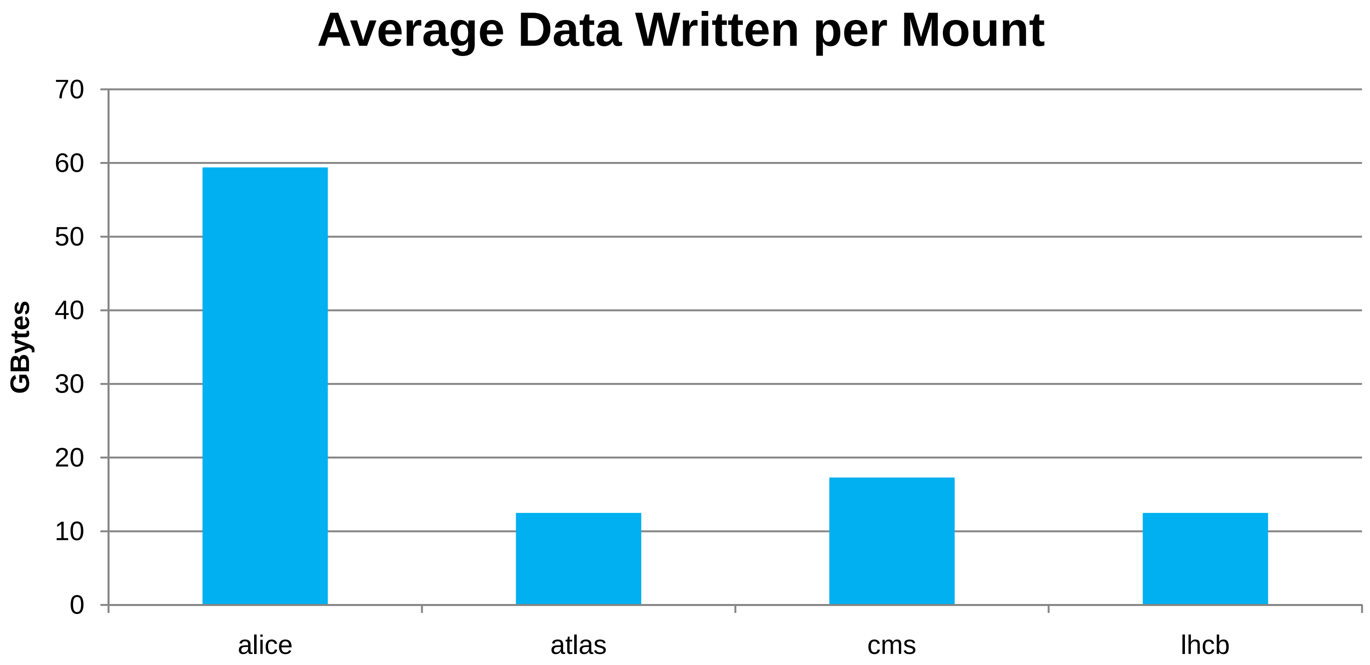
•LHC tape infrastructure sizing was based on 1-2GB files.

Repeat Mounting



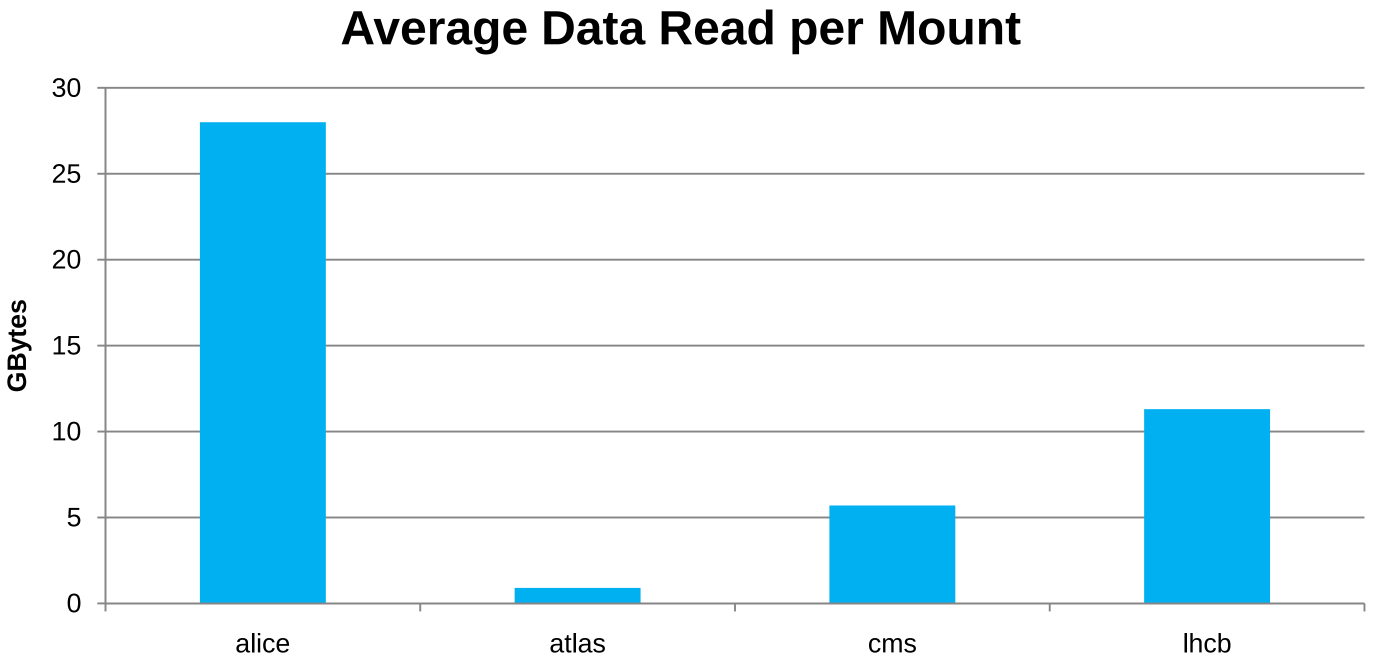
- Tapes are being repeatedly mounted/unmounted.
- Takes around 4 minutes to mount a tape compared to 100 minutes to write a complete tape
- Increases wear/tear on robots and drives along with risk of tape media issues.

Repeat Mounting - Write



- Write migration to tape is currently triggered by Castor based on the modification date of the file (typical setting is 30 minutes)
- Current policy was chosen to write files to tape quickly but this leads to inefficient short mounts
- Need to move to a migration based on volume of data (one 700GB tape) to write along with a maximum delay. (8 hours)
  - For CDR, at 100MB/s, the expected would be 2 hours to start migration and 2 hours to complete writing to tape

Repeat Mounting - Read

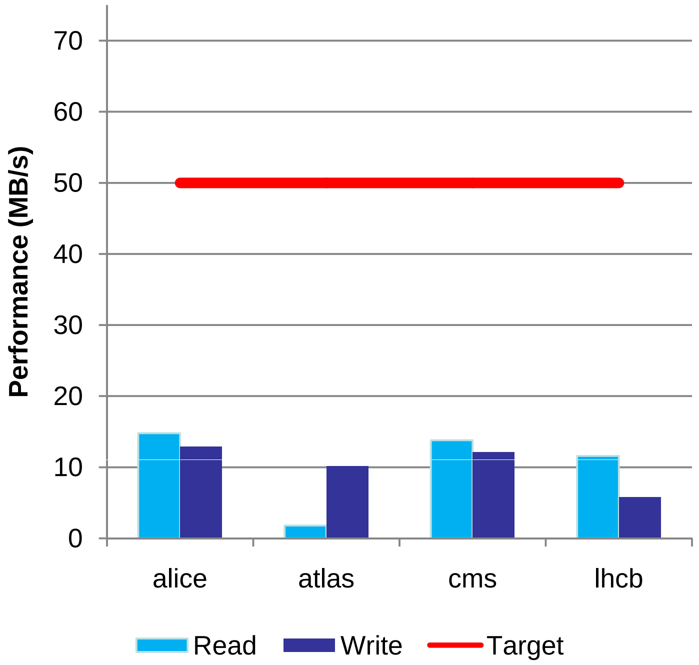


- Very limited pre-staging of data means that tapes are being re-mounted for each file. Small files makes situation worse.
- Queuing overhead to get to a drive further increases the batch job inefficiency and job performance.

Total performance to tape

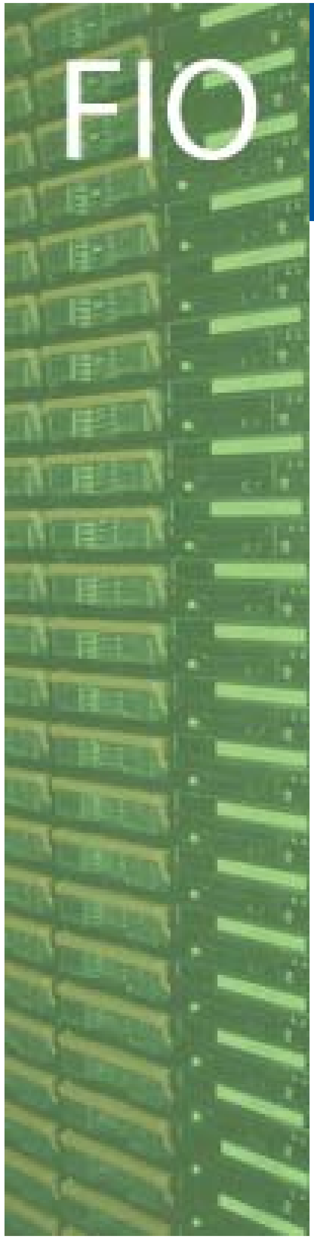
CERN IT Department

- Planning was based on total performance of 50MB/s.



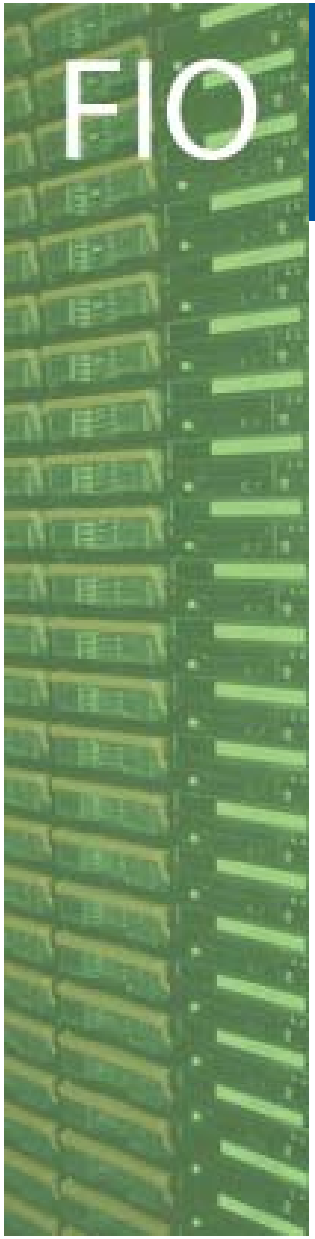
VO	File Size	Mounting Overhead
Alice	✗✗	✓
Atlas	✗✗	✗✗
CMS	✓✓	✗✗
LHCb	✗✗	✗

- Total performance is based on the sum of data transferred against the total time spent on drives (including mount unmount time).



Proposal

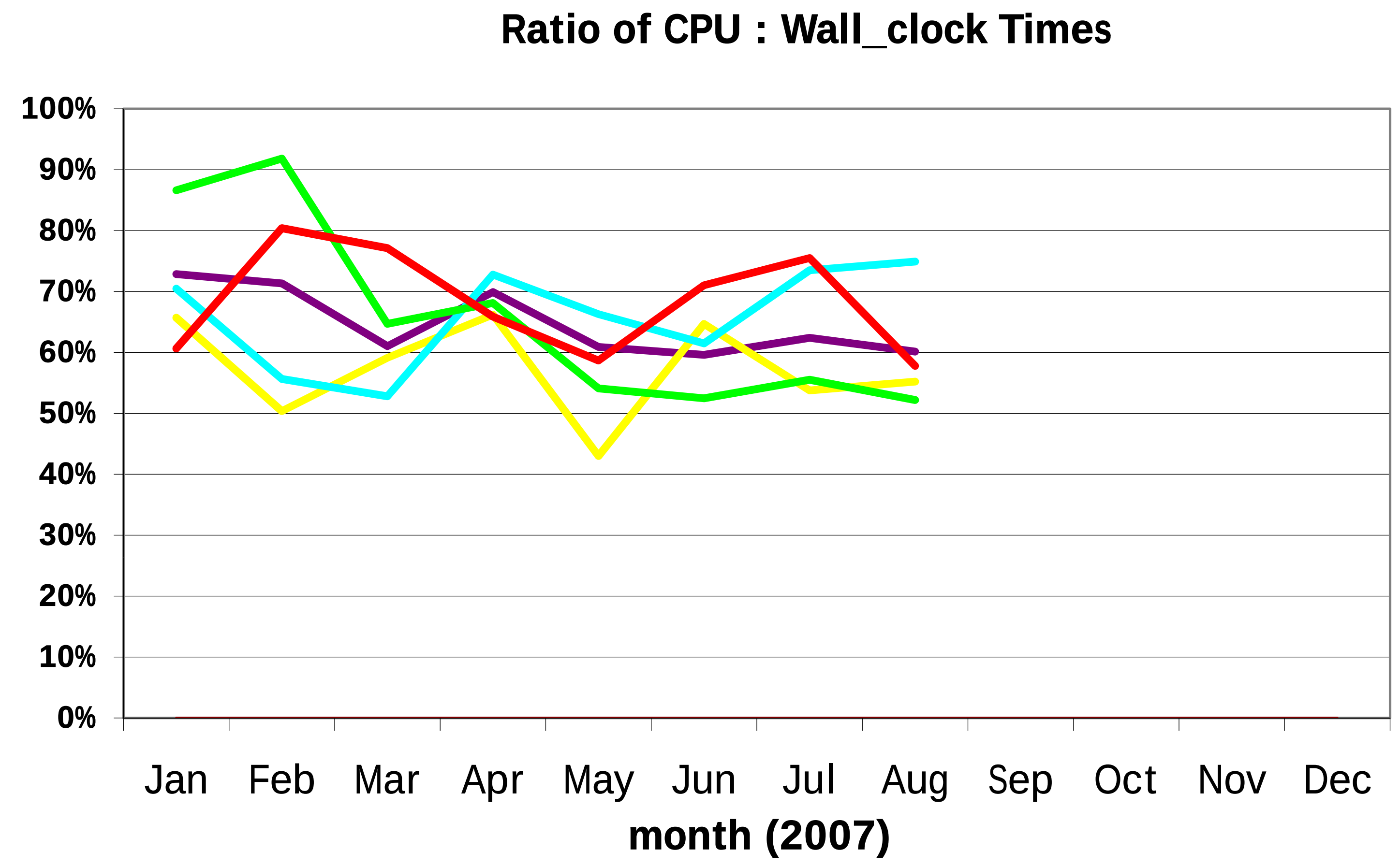
- Experiments should
  - Move to 2GB files for tape transfers
  - Ensure that pre-staging is standard for all applications
- Castor Operations will change policies for CCRC
  - Write policy of at least one tape of data with 8 hours maximum delay
  - Limit mounting for reads unless at least 10GB or 10 files requested for each read mount or if a request is 8 hours old
- Monitor February CCRC performance and cover shortfall with
  - Major drive purchases and dedication for experiments
    - Fixed budget! Implies reduction in CPU/disk capacity



Backup and Background

Batch Efficiency

CERN IT Department


Ratio of CPU : Wall_clock Times
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month (2007)