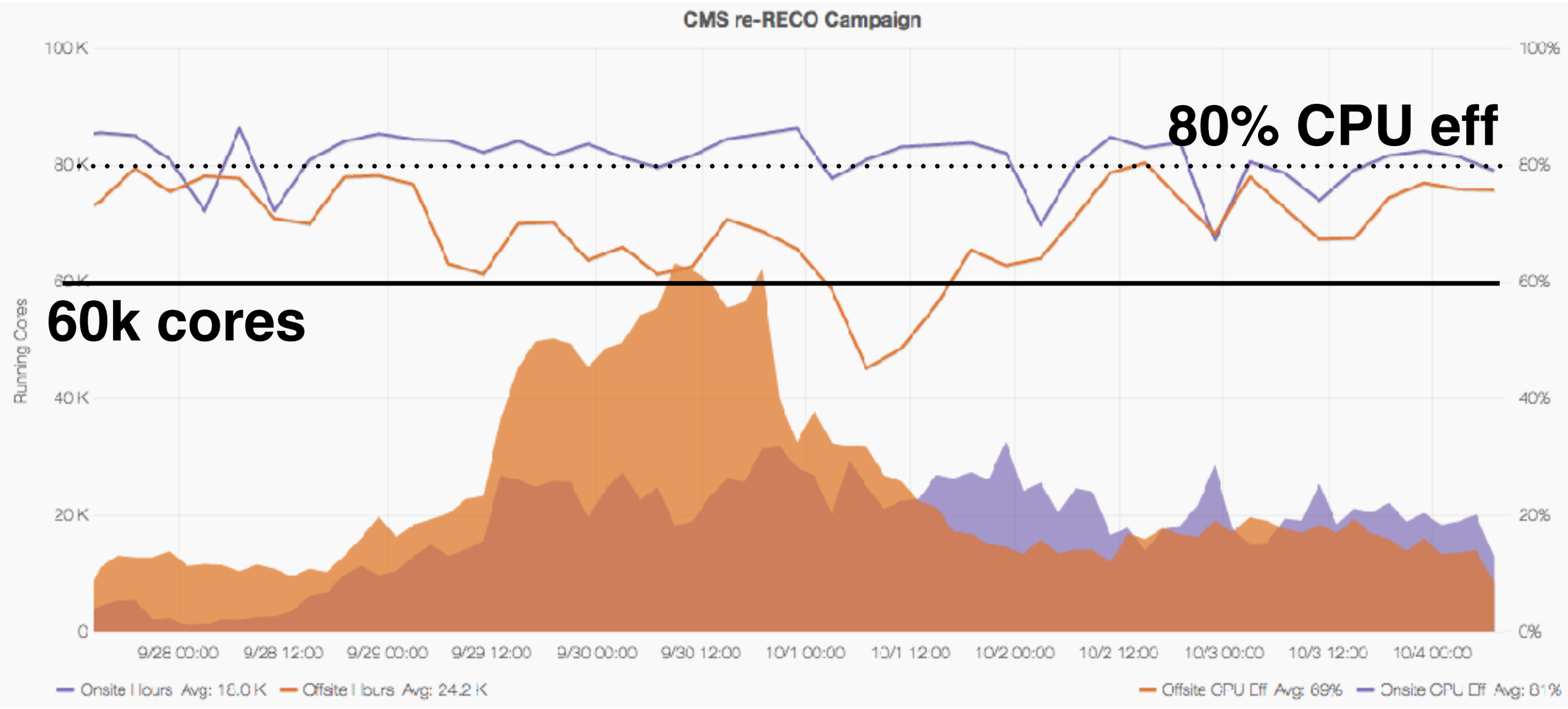
CMS Input

Brian Bockelman

How can we reduce data & storage costs?

- CMS has aggressively invested in software and computing infrastructure to be able to **stream data from remote centers**.
    - In Run2, more data-intensive workflows are “WAN-ready”.
- The large bulk of our organized processing workflows *no longer require local data*.
    - Local data provides a modest CPU efficiency boost.
    - Current reprocessing campaign: **Average transfer rates are ~.5Mbps/core**; instantaneous rates are ~35Mbps/core.
- Production can effectively utilize **large CPU resources without storage** for production campaigns.
- **Stageout not as robust** and efficient as streaming inputs. Note that Amazon EC2 has higher costs for stageout than streaming inputs.



Current CMS reprocessing campaign:

- By streaming remotely, more sites can participate.
- More core-hours spent on “offsite” jobs than onsite.
- Offsite CPU efficiency hit is noticeable but manageable.
- CPU efficiency somewhat independent of data rates.

Data Analysis

- Analysis jobs are far more difficult to characterize and outlier workflows can span multiple order-magnitude of scale.
    - A recent user submitted a workflow requiring 2.6B file-opens…
    - Many well-formed user analysis jobs can effectively stream over the WAN (there are clusters of sites that are "nearby" in RTT). **Many user analysis jobs cannot**.
- Question: can we instead **cache** the analysis datasets locally? That would potentially provide much higher bandwidth and provide protection against "clever users".

Caching Pilot Goals

- Evaluate XRootd Caching Software at large scale:
    - Performance evaluation, i.e., simultaneous read/write vs number of clients
    - Operational cost evaluation: operate it in production for 3-6 months before data taking restarts in 2017.
    - Write interim report after 2-3 months: decide on whether to continue.
    - Write final report after 6 months: continue ops if deemed worthwhile.
- Use findings as input to future planning

Scale of Pilot

- 300+ TB in SoCal (Caltech & UCSD)
- 30-50 Gbps IO performance for simultaneous read/write with up to 20,000 clients reading.
  - 100+ disks, diverse hardware & filesystems
  - estimate to need ~20 disks to fill 10Gbps pipe
- Operations Phase:
  - Host part of CMS namespace, most likely all MINIAOD from currently most used release
  - default access for all jobs in SoCal to hit cache for this namespace

Interested in collaborations with others!!

Where does this data live?

- Caching and remote streaming are lovely: **the data still must be on disk somewhere!**
  - Unlikely that the “storage element” model will go away anytime soon for large sites.
- A few operational improvements for traditional CMS storage elements:
  - Tier-2s no longer need anything besides a high-performance GridFTP endpoint*.
  - USCMS has long run a centralized PhEDEx site agents for Tier-3s; Tier-2 sites could do the same.

*Has been true for awhile (first site was T2_UK_Bristol); now you can pass all the SAM tests…

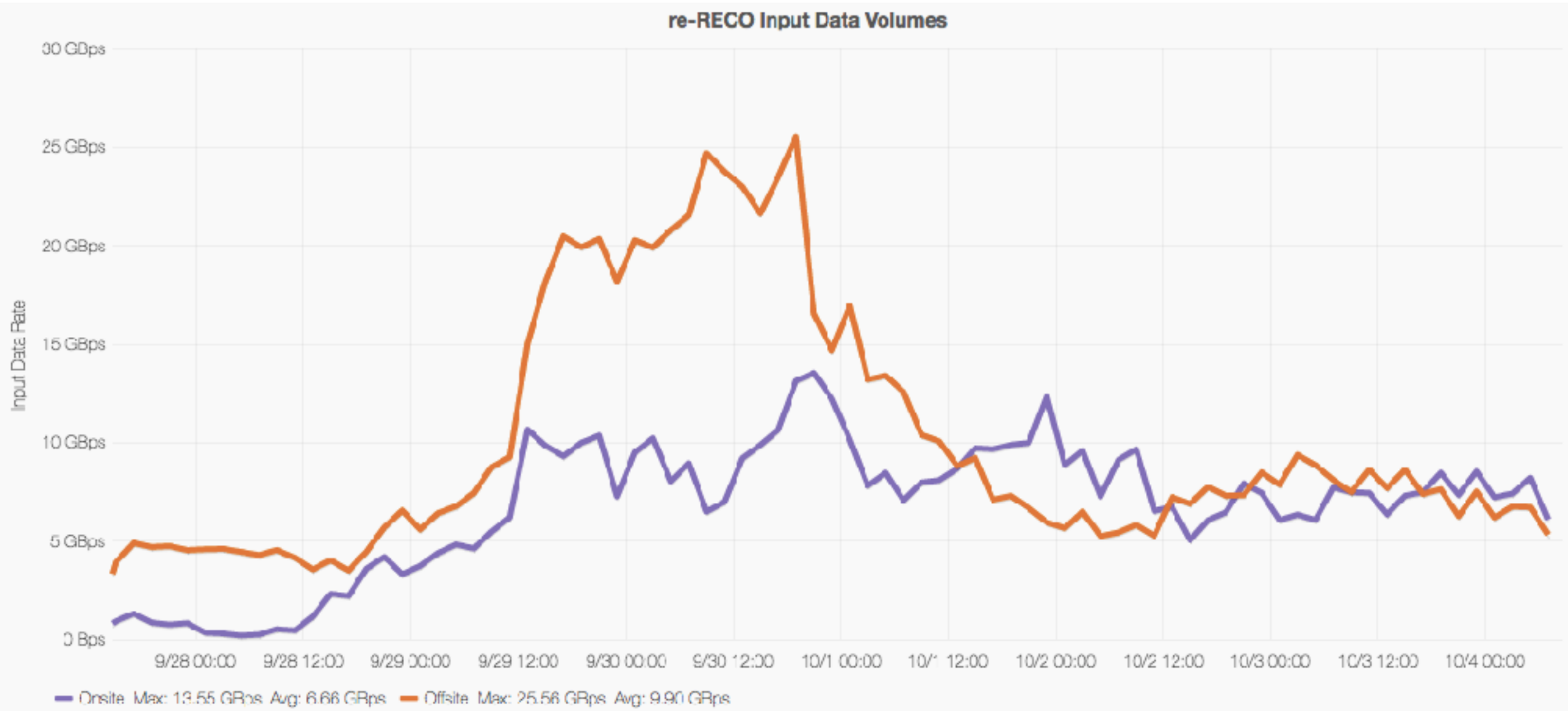
Take-home message:

CMS can more effectively utilize CPU-only sites

CMS is working to simplify storage management, particularly at smaller T2 sites.

CMS is interested in continued improvements in the effectiveness of its use of storage

Backup



Current CMS reprocessing campaign data rates:

- CPU time appears to increase quadratically with pileup.
- Data size appears to increase approximately linearly.
- Current WAN transfer rates (~25Gbps) are manageable.
- CPU efficiency should go up & WAN rates decrease with pileup increases.