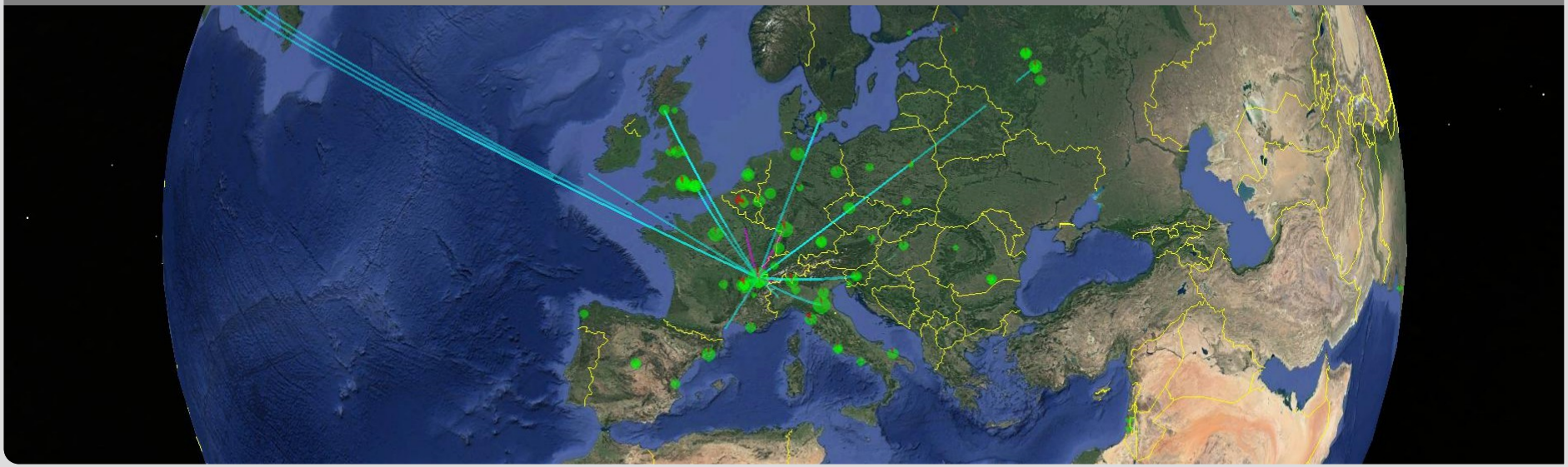
Karlsruhe Institute of Technology

Boosting data-intensive end-user analyses: TOpAS – a dedicated high throughput cluster

ErUM-Data meeting | 2020-04-17

C. Heidecker, R. Caspart, T. Feßenbecker, M. Fischer, M. Giffels, M. Horzela, G. Quast, M. J. Schnepf, R. F. von Cube

INSTITUTE OF EXPERIMENTAL PARTICLE PHYSICS (ETP) · DEPARTMENT OF PHYSICS

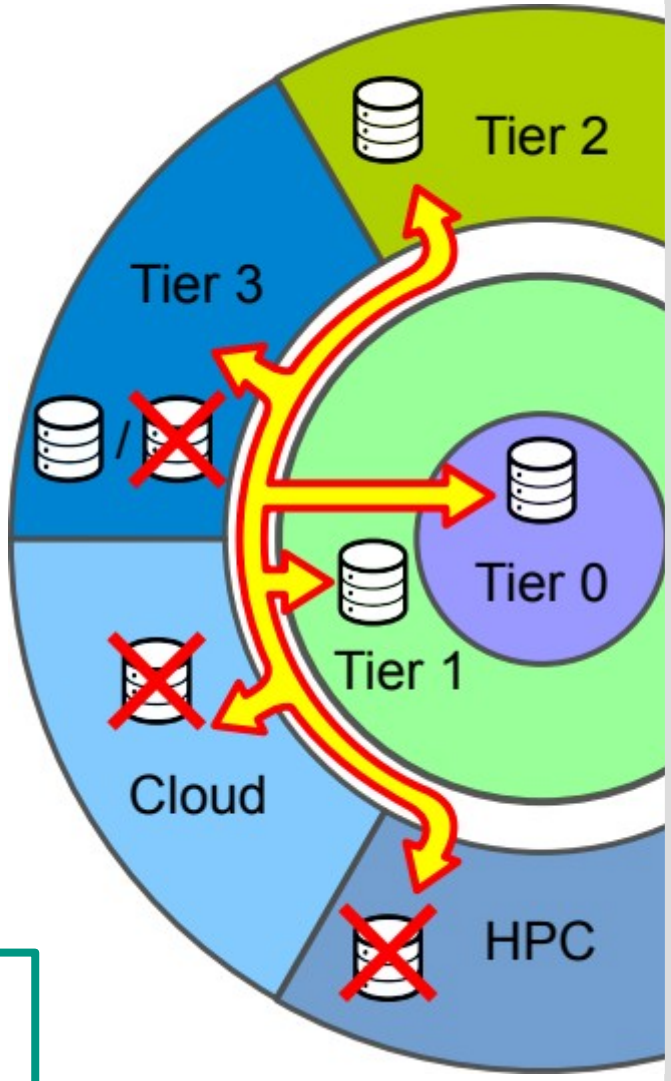


Computing challenge for data analysis

- heterogeneous infrastructure
  - specialized centers providing Grid storage
  - diverse computing resources for analysis processing providing no dedicated storage

- Challenges
  - usually no centrally managed Grid **storage** elements at Tier 3 resources

  - **network** interconnection may be **limited**
    → end-user analysis work-flows suffer from the need of remote data access



Data Throughput must be optimized for end-user analyses!

Data flow optimization

- common solution for repeatedly accessed data:
  Caching data as close as possible to the processing CPU

- concept is suitable
  - for HEP workflows that process same datasets frequently
  - for computing resources with no centrally managed storage

Data flow optimization

- common solution for repeatedly accessed data:
  Caching data as close as possible to the processing CPU

- concept is suitable
  - for HEP workflows that process same datasets frequently
  - for computing resources with no centrally managed storage

- works on dedicated resources with “naive” caches
  ⇨ **throughput optimized computing clusters**
  preferred for processing data intensive analysis workflows

Propose installation of well-balanced local caches

Throughput Optimized Analysis System

Design criteria:

- number of users and typical size of a data set
- required i/o per core (ROOT deserialization limit at ~50MB/s)

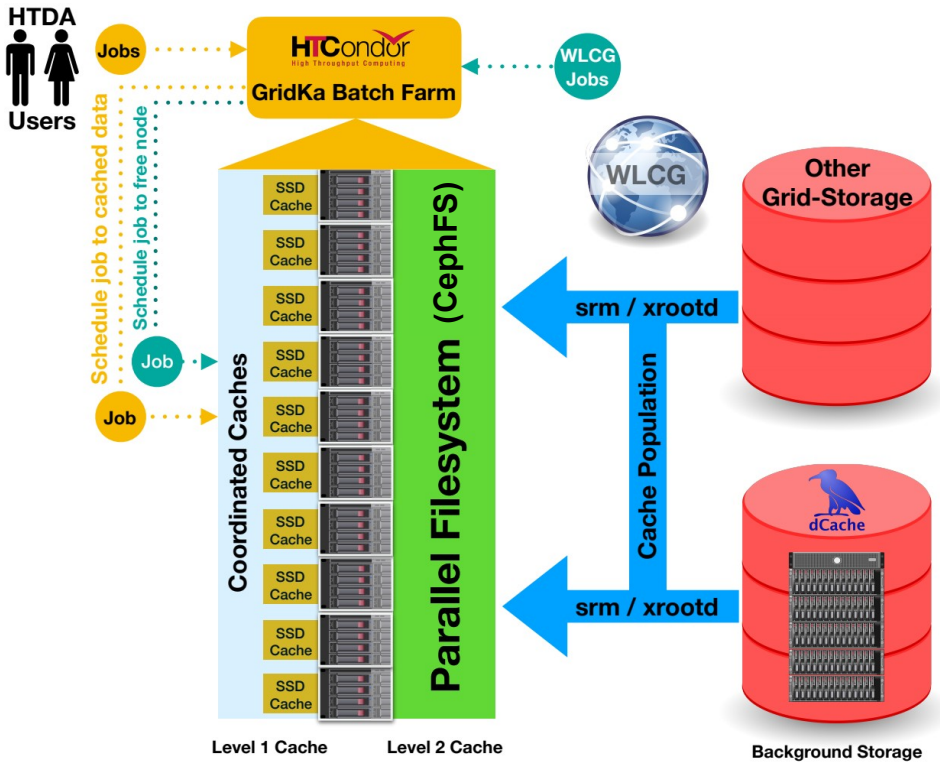
Throughput Optimized Analysis System

Design criteria:

- number of users and typical size of a data set
- required i/o per core (ROOT deserialization limit at ~50MB/s)

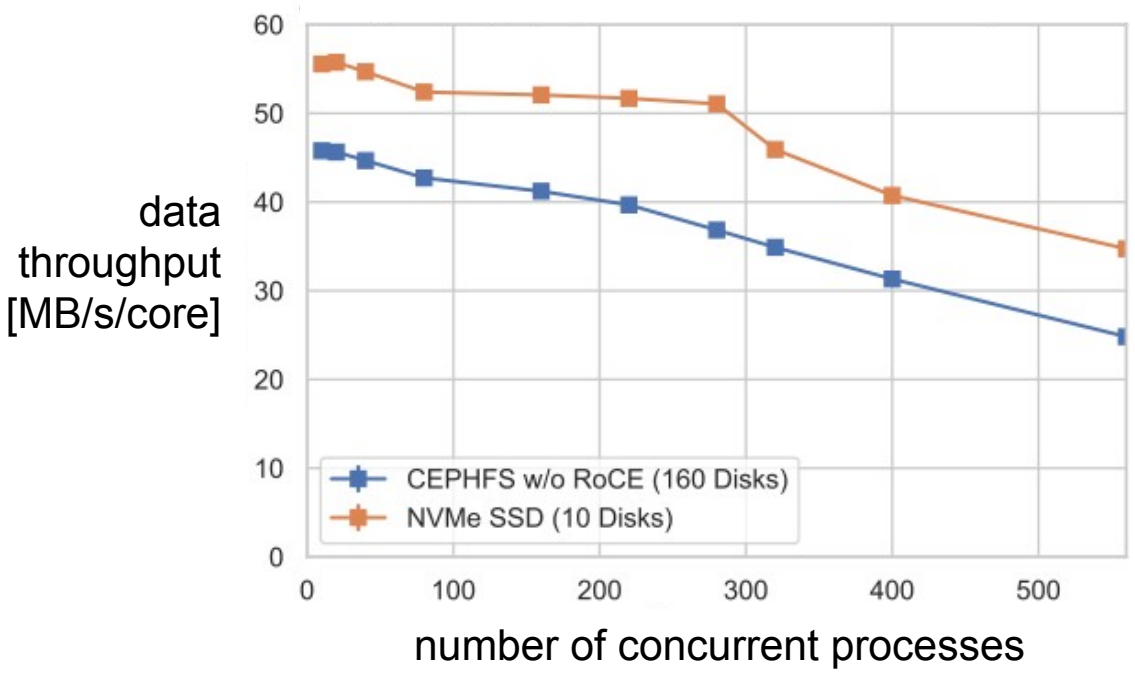
TOpAS, a production system for data-intensive end-user analysis jobs

- designed to process ~1 PB data per day
- built of 11 nodes:
  - 42 CPU cores/node (including HT, 462 cores in total)
  - 252 GB RAM per node (6GB/core)
  - 100 Gbit/s direct link to GridKa dCache
  - 1 TB NVME SSD/node
  - 16 x 5.5TB HDDs/node



Performance benchmarks

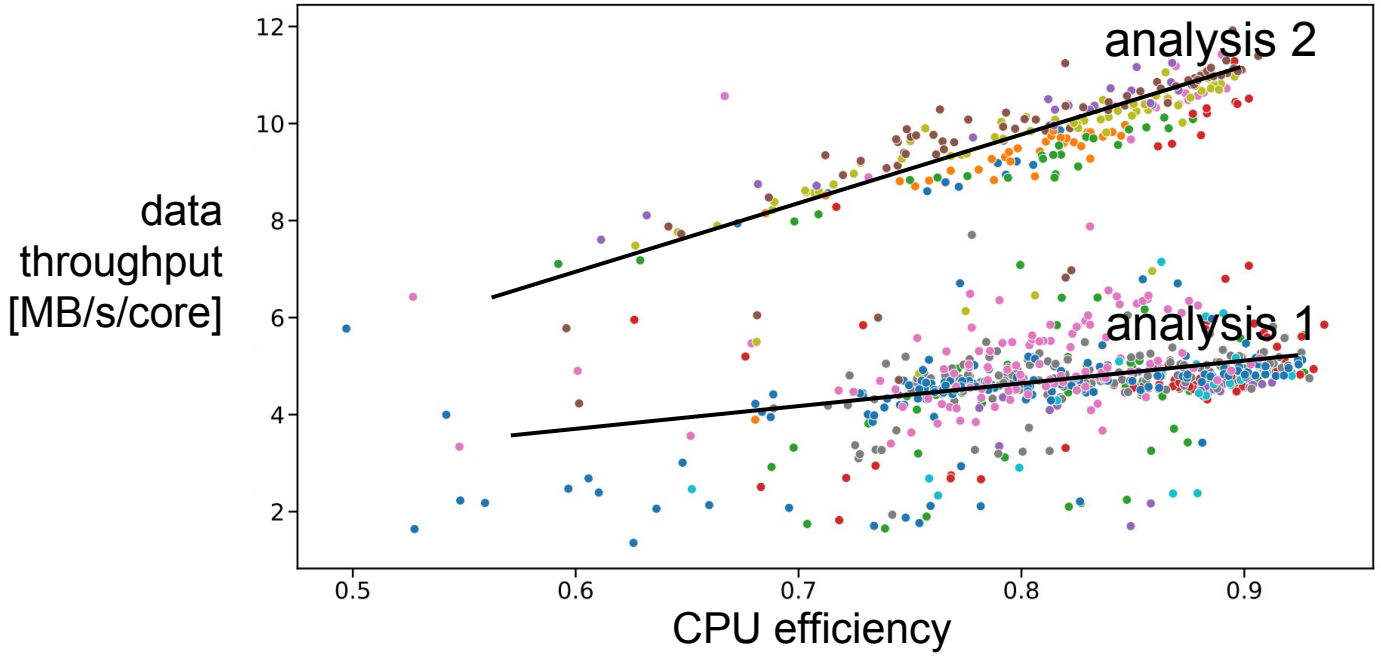
measured maximum performance by reading root files and applying minimal data processing



- both cache types, NVME SSD and CephFS, offer a stable read rate >25 MB/s when processing data with all cores of the cluster
- a combination of cache and remote access allows even higher data throughput

Optimizing throughput boosts CPU efficiency

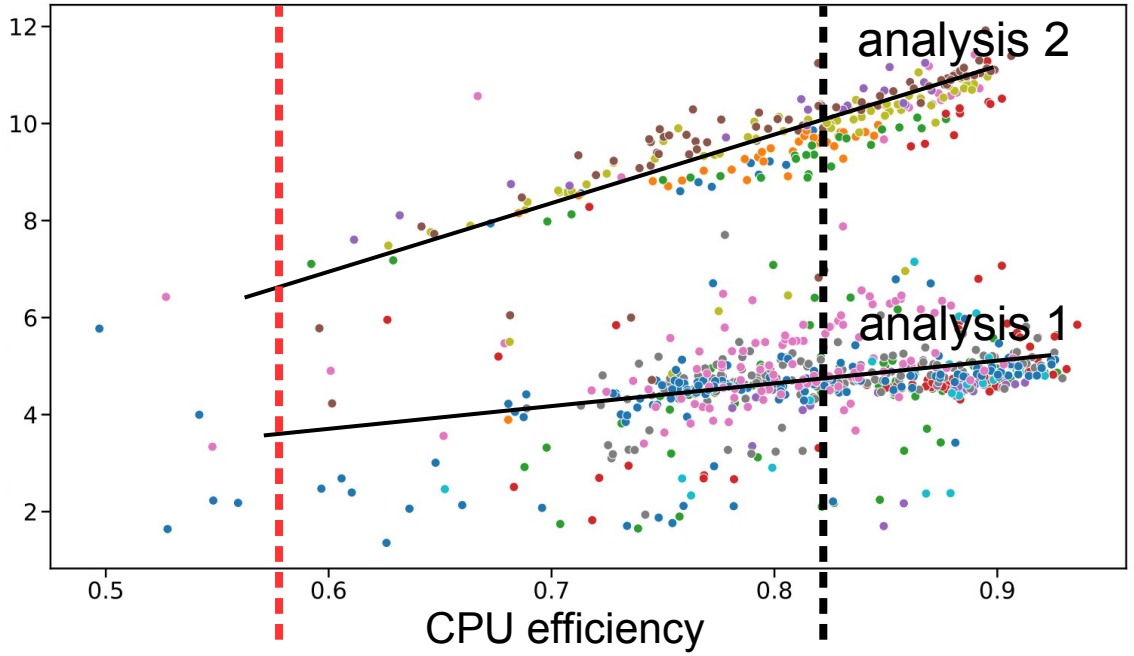
data analysis work-flow: data-intensive CMS jet energy calibration
(~17 TB of data in 2548 jobs)



overall code efficiency limits CPU efficiency to < 95%

Optimizing throughput boosts CPU efficiency

data analysis work-flow: data-intensive CMS jet energy calibration
(~17 TB of data in 2548 jobs)



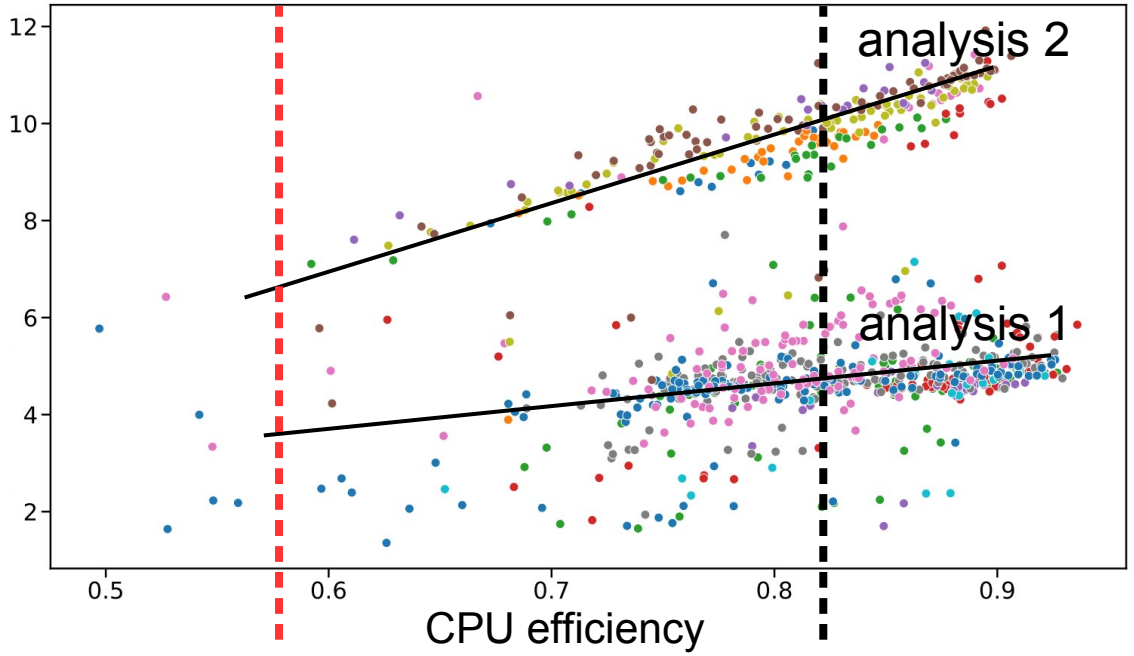
overall code efficiency limits CPU efficiency to < 95%

average for not optimized institute cluster

average for throughput optimized TOpAS

Optimizing throughput boosts CPU efficiency

data analysis work-flow: data-intensive CMS jet energy calibration
(~17 TB of data in 2548 jobs)



overall code efficiency limits CPU efficiency to < 95%

Ø runtime per job reduced by 40%

CPU efficiency and, thus, runtime of jobs get highly boosted by processing jobs on throughput optimized computing clusters!

Conclusion

I hope that I managed to convince you that our

Throughput **Op**timized **A**nalysis **S**ystem (TOpAS),

a dedicated high throughput cluster,

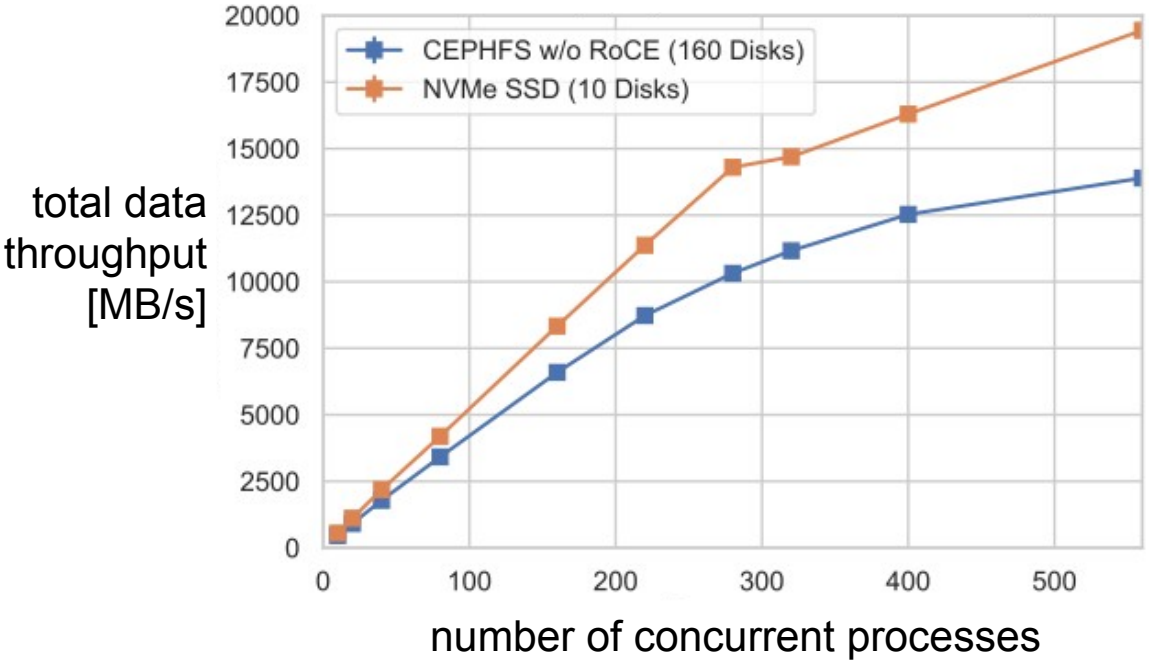
fulfills its purpose of boosting data throughput

- designed for processing ~1PB of data per day
- end-user analysis profit from massive improvement of analysis runtime

Backup Slides ...

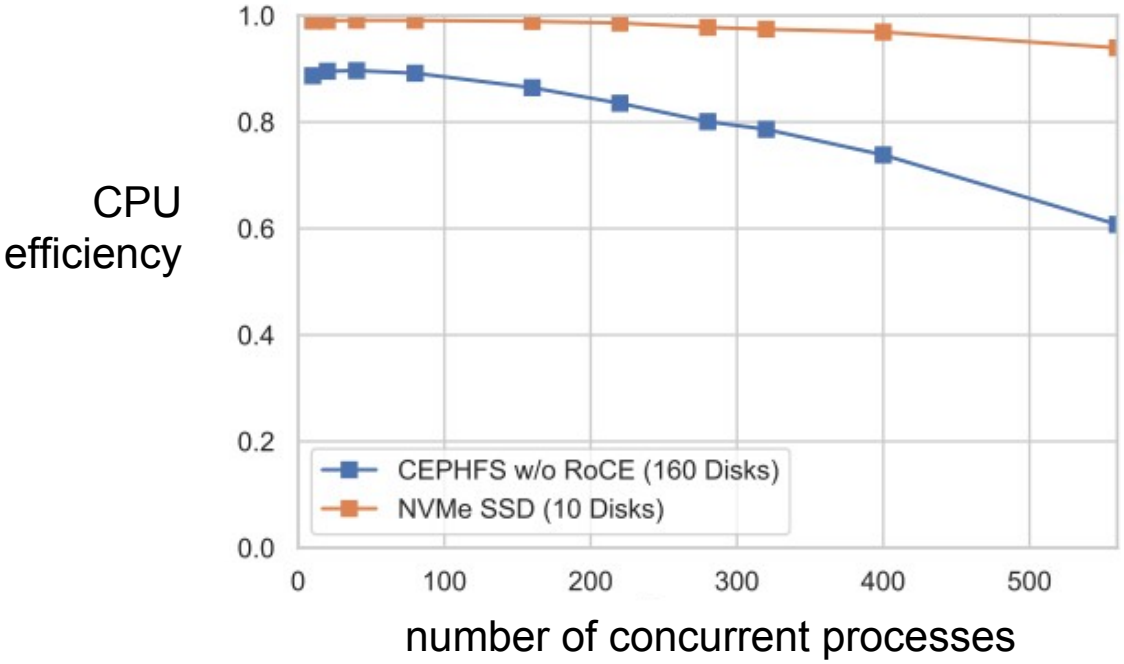
Performance benchmarks

measured maximum performance by reading root files and applying minimal data processing



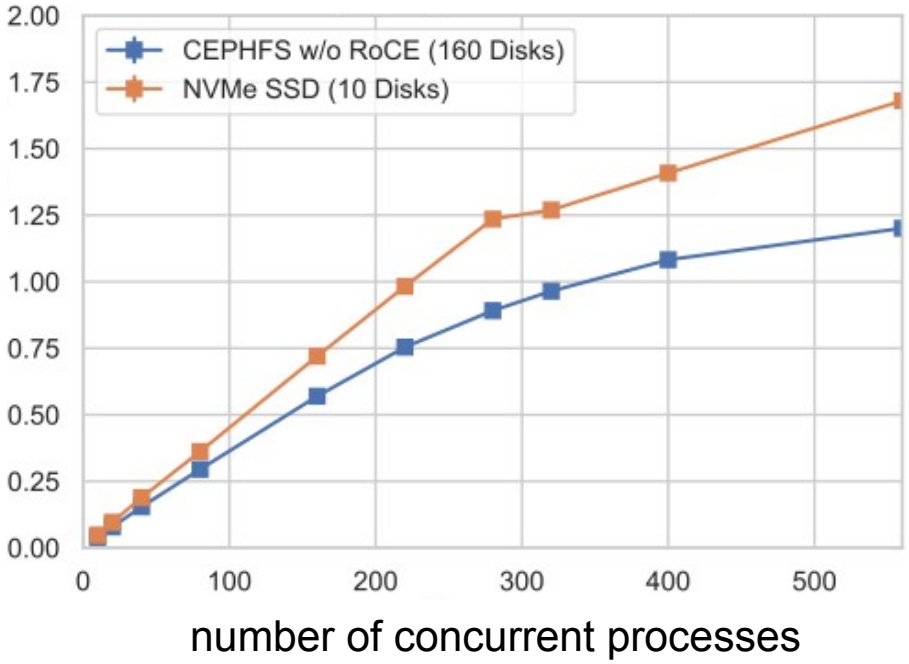
Performance benchmarks

measured maximum performance by reading root files and applying minimal data processing



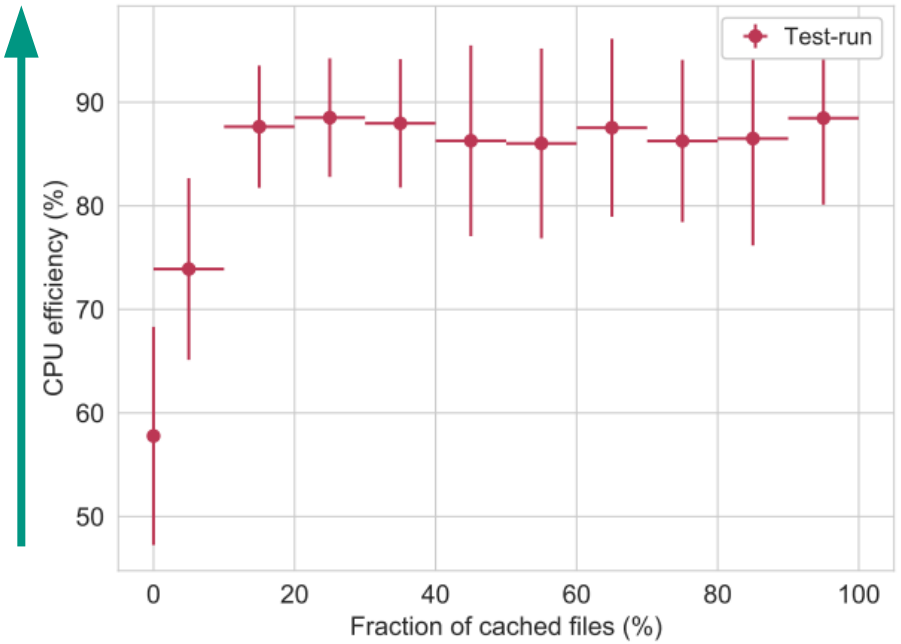
Performance benchmarks

measured maximum performance by reading root files and applying minimal data processing



ETP Institute cluster – a small version of TOpAS with 3 nodes

better



→ Bars show variance of test results

ETP Institute cluster

- 3 worker nodes, 60 job slots
- 3 x 2 TB SSD RAID0 caches, ~ 1.2 GB/s cache read rate
- ~ 750 MB/s shared internet connection

- Measured performance achieved by data-intensive HEP workflow
- Caching increases CPU efficiency and, thus, optimizes resource utilization

Significant performance boost for CMS Jet Energy calibration workflow by factor of 1.6 limited by advanced data processing