Computational Storage to Increase the Analysis Capability of Tier-2 HEP Data Sites

Chen Zou, Andrew A. Chien

Introduction to HEP analysis

Detecting collision events from 8 different experiments (e.g., ATLAS, CMS, LHCb, ALICE) around the ring, the Large Hadron Collider (LHC) produces a huge amount of data. Despite the filtering from purposely-built real-time triggering mechanisms, the collected events still mount up to 100PB per year for physicists around the world. To meet the growing analysis demands, the LHC and HEP community built a tiered grid [1] around the globe to host continuously growing data. The large disk pools offer an opportunity to improve analysis performance.

To meet the growing analysis capacity especially for the luminosity increase [2] (and hence the collision rate and HEP data collection) after the long shutdown of LHC,

- Limited by cost considerations, the conventional wisdom on developing the HEP data centers is to buy disks to host the large disk pools.
- Developed a parametric model of a tier-2 data center, workhorse for analysis.
- Evaluated upgrade options to increase analysis capacity including computational storage disks.

CERN and HEP community built a tiered grid [1] around the globe to meet the storage and analysis demands.

- Limited by cost considerations, the conventional wisdom on developing the HEP data centers is to buy disks to host continuously growing data.
- The large disk pools offer an opportunity to improve analysis performance.

Detecting collision events from 8 different experiments (e.g., ATLAS, CMS, LHCb, ALICE) around the ring, the Large Hadron Collider (LHC) produces a huge amount of data. Despite the filtering from purposely-built real-time triggering mechanisms, the collected events still mount up to 100PB per year for physicists around the world. To meet the growing analysis demands, the LHC and HEP community built a tiered grid [1] around the globe to host continuously growing data. The large disk pools offer an opportunity to improve analysis performance.

To meet the growing analysis capacity especially for the luminosity increase [2] (and hence the collision rate and HEP data collection) after the long shutdown of LHC,

- Limited by cost considerations, the conventional wisdom on developing the HEP data centers is to buy disks to host continuously growing data.
- The large disk pools offer an opportunity to improve analysis performance.

To meet the growing analysis capacity especially for the luminosity increase [2] (and hence the collision rate and HEP data collection) after the long shutdown of LHC,

- Limited by cost considerations, the conventional wisdom on developing the HEP data centers is to buy disks to host continuously growing data.
- The large disk pools offer an opportunity to improve analysis performance.

To meet the growing analysis capacity especially for the luminosity increase [2] (and hence the collision rate and HEP data collection) after the long shutdown of LHC,

- Limited by cost considerations, the conventional wisdom on developing the HEP data centers is to buy disks to host continuously growing data.
- The large disk pools offer an opportunity to improve analysis performance.

Higgs boson analysis performance with different upgrades

Backbone network 1000 Gbps & UCT2 SSDs. 8.62x

Disk bottleneck addressed. Backbone network bottleneck again.

Baseline. 1x

Backbone network bottleneck.

Backbone network 100 Gbps -> 1000 Gbps. 1.34x

Network bottleneck addressed. UCT2 disk bottleneck

Analytical model

Because of the highly parallel nature of the workload, we could derive the following analytical model for analysis performance, for DCache Aggregate Disk Read Bandwidth bandwidth of DCache Aggregate Disk Read Bandwidth, and UCT2 Aggregate Compute Throughput application latency.

\[
\text{Application Latency} = \text{Event Data Size} \times \frac{\text{Event Data Size}}{\text{Bandwidth}} + \text{Application Level Cost} + \text{Network Level Cost} + \text{Data Staging Cost} + \text{Disk Level Cost}
\]

Comparing analytical model and simulation

Our analytical model for analysis performance agrees well with the simulation results for \(\#\) case. Diff is the compute latency.

By preprocessing and emits only the needed data, computational storage reduces the data size for all resources used in the later stage. This is the out-of-box solution that has been used to improve `Liebig's barrel` trade-off as a local optimum.

Methodology

Stages. The processing of each file (~3GB - 10000 events) CMS dataset is treated as independent jobs. Each job is of the following stages:

1. In-storage computation at DCache cluster
2. Data staging: DCache → UCT2
3. In-storage computation at UCT2 cluster
4. Higgs boson stats compute at UCT2 cluster
5. Store back results: UCT2 → DCache

Stage 1 and Stage 3 are optional and only applicable when computational storage [3] upgrade is considered for column select (slimming).

Approach. A C++ simulator implements the performance model with task-level granularity, where each stage for a job is instantiated as a task. It simulates task progress under the modeled customizable resource properties and produce the application performance measured in latency.

Silicon area and power

We compare silicon area and power consumptions of employing CPUs or computational storage disks to perform the column selecting tasks before sending out from DCache cluster. This work was funded in part by the National Science Foundation Cooperative Agreement OAC-1836650.

References