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

Chen Zou, Andrew A. Chien
Department of Computer Science, University of Chicago
Chicago, USA
{*chenzou@, achien@cs.*}*uchicago.edu*

Robert Gardner, Ilija Vukotic
Enrico Fermi Institute, University of Chicago
Chicago, USA
{*rwg, ivukotic*}*@uchicago.edu*

***Abstract*—Large Hadron Collider (LHC) produces collision data at 100 PB/year which needs to be stored and analyzed for high energy physics (HEP) theories. We reconsider the design choices of HEP data centers and evaluate different upgrade options to improve their analysis capacity.**

Results show that computational storage to be the cost-effective and power-efficient upgrade option. Computational disks in the storage cluster deliver a 9.3-fold speedup for Higgs Boson analysis. This exceeds the speedup from all other upgrades considered (faster network: 100 to 1000 Gbps, upgrade from HDDs to SSDs).

***Keywords*-High energy physics, Computational storage**

I. Introduction

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 analyze and test different theories of high-energy physics (HEP).

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 the continuously growing data. By developing a parametric and detailed model of a tier-2 data center, which is the workhorse for analysis workloads for each university or research institute, we carefully evaluate several different upgrade options to match the increasing analysis capacity demand, especially for the luminosity increase [2] (and hence the collision rate and HEP data collection) after the long shutdown of LHC.

II. Methodology

Model. We build the performance model after the UChicago tier-2 data center(named as UChicagoT2), which is shown in Figure 1 with different configurations. The data center has one compute cluster 'UCT2' and one storage cluster 'DCache' which manages the collision event data with DCache [3]. Hard disks are modeled to have 200 MB/s read bandwidth and 100 MB/s write bandwidth.

This work was funded in part by the National Science Foundation Cooperative Agreement OAC-1836650.

Workload. Higgs boson analysis with 47TB CMS dataset is used as the workload. It independently maps two steps of computation to each collision event. First, apply filters to muons, global muons, electrons tracks. Second, calculate statistics on filtered muons, global muons, electrons tracks in the event for potential Higgs bosons. This parallel mapping are amenable to acceleration [4], [5]. The processing of each file (∼3GB, ∼10000 events) CMS dataset is treated as independent jobs. Each job is of the following stages:

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

Stage 1 and Stage 3 are optional and only applicable when computational storage [6] upgrade is considered. The performance number to model stage 4 is extrapolated from running Higgs Boson analysis on one actual UCT2 machine.

Approach. Above configurations and performance numbers model the current UChicago T2 data center, which is the baseline. We further consider different upgrades:

- Backbone networks from 100 Gbps to 1000 Gbps.
- Replacing HDDs with SSDs in a cluster.
- Adding computational storage [6] in DCache cluster.

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.

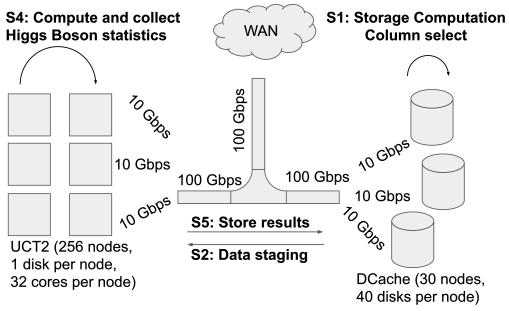

Figure 1. UChicagoT2 data center modeling

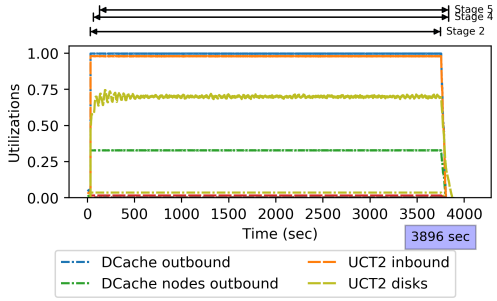

Figure 2. Performance and resource utilization in baseline

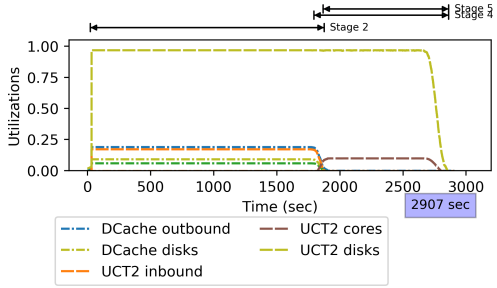

Figure 3. Upgrade backbone network: 100 Gbps to 1000 Gbps

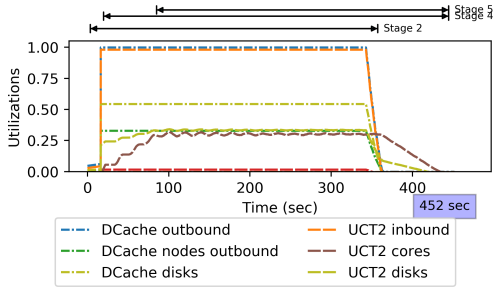

Figure 4. 1000 Gbps backbone network + UCT2 SSDs

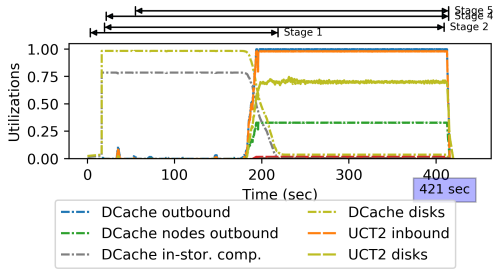

Figure 5. Computational storage in DCache cluster

III. Experiments

All the figures plot utilization of different resources in UChicagoT2 and follow the same organization. We include a legend for the resources with notable utilization under each plot. On top of each plot, there are marks showing the period when tasks of each stage are being processed.

Baseline. Figure 2 details its Higgs boson analysis performance. Each line plots the utilization of a specific resource over time. '3896 sec' is the overall analysis latency. DCache outbound 100 Gbps link and UCT2 inbound 100 Gbps link were almost always at nearly 100% utilization, performing Stage-2 tasks of transferring data from DCache to UCT2.

Improve backbone network from 100GbE to 1000GbE. To address the performance bottlenecks in baseline, backbone network is upgraded. As shown in Figure 3, the upgrade increases analysis performance by 1.3x. Backbone network (DCache outbound and UCT2 inbound) utilization drop significantly, while UCT2 disks become the system bottleneck. This led to the stall of Stage 4 tasks that read events from UCT2 disks while Stage 2 tasks are intensively writing to UCT2 disks during the DCache to UCT2 event file transfer (see stage duration at the top of the figure).

Further replace HDDs with SSDs in UCT2. On top of the network upgrade, we consider replacing HDDs with SSDs for UCT2 cluster targeting its disks bottleneck. SSDs are modeled as 4 GB/s for read and 2 GB/s for write, as opposed to 200 MB/s and 100 MB/s for HDDs as discussed in Section II. This led to a further 6.4x latency reduction, and 8.6x from baseline, as shown in Figure 4. The performance bottleneck circles back to backbone networks.

Employ computational storage. We consider a non-conventional upgrade directly from baseline: adding in-storage computation capability to hard disks in DCache cluster to perform column-selecting (slimming) such that only selected fields in each tuple needed for statistics calculation are transferred to UCT2 cluster. Each computational storage disk embeds an in-order single-scalar core modeled as an Ibex core [7] to compute on top of the data streams from storage media. Spike [8] is used to simulate column-select tasks and emit execution traces for performance analysis. Figure 5 shows a 9.3x speedup in this case. Because data is selectively transferred out from DCache storage, the bandwidth requirement on the later resources (backbone networks, UCT2 disks, UCT2 compute) is largely reduced.

References

[1] "The grid: A system of tiers," https://home.cern/science/computing/grid-system-tiers.

[2] "High-luminosity lhc," https://home.cern/science/accelerators/high-luminosity-lhc.

[3] "dcache," https://www.dcache.org/.

[4] Y. Fang, C. Zou, A. J. Elmore, and A. A. Chien, "Udp: a programmable accelerator for extract-transform-load workloads and more," in *MICRO' 17*. IEEE, 2017, pp. 55–68.

[5] Y. Fang, C. Zou, and A. A. Chien, "Accelerating raw data analysis with the accorda software and hardware architecture," *VLDB' 19*, vol. 12, no. 11, pp. 1568–1582, 2019.

[6] C. Zou and A. A. Chien, "Empowering architects and designers: A classification of what functions to accelerate in storage," *UChicago CS TR-2020-02*.

[7] P. D. Schiavone and F. Conti, "Slow and steady wins the race? a comparison of ultra-low-power risc-v cores for internet-of-things applications," in *PATMOS'17*. IEEE, 2017, pp. 1–8.

[8] "Spike RISC-V ISA Simulator," https://github.com/riscv/riscv-isa-sim.