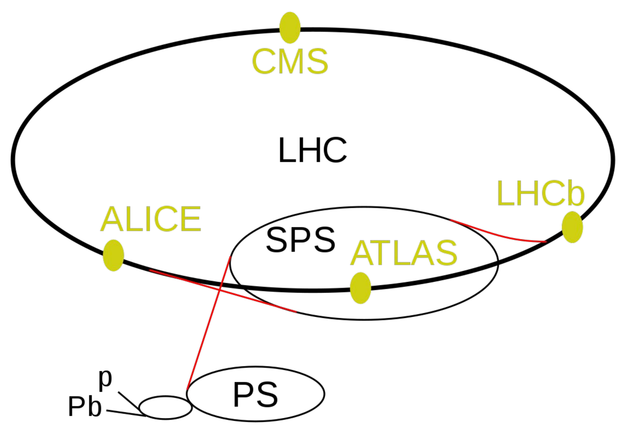
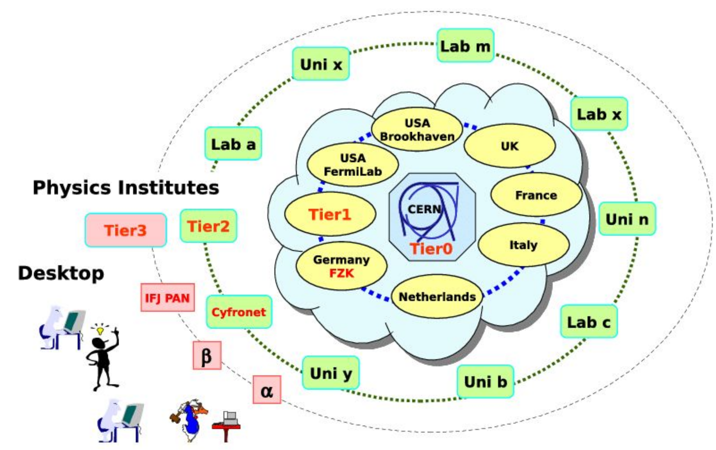
Computational Storage to Increase the Analysis Capability of Tier-2 HEP Data Sites

Chen Zou, Andrew A. Chien Robert Gardner, Ilija Vukotic

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

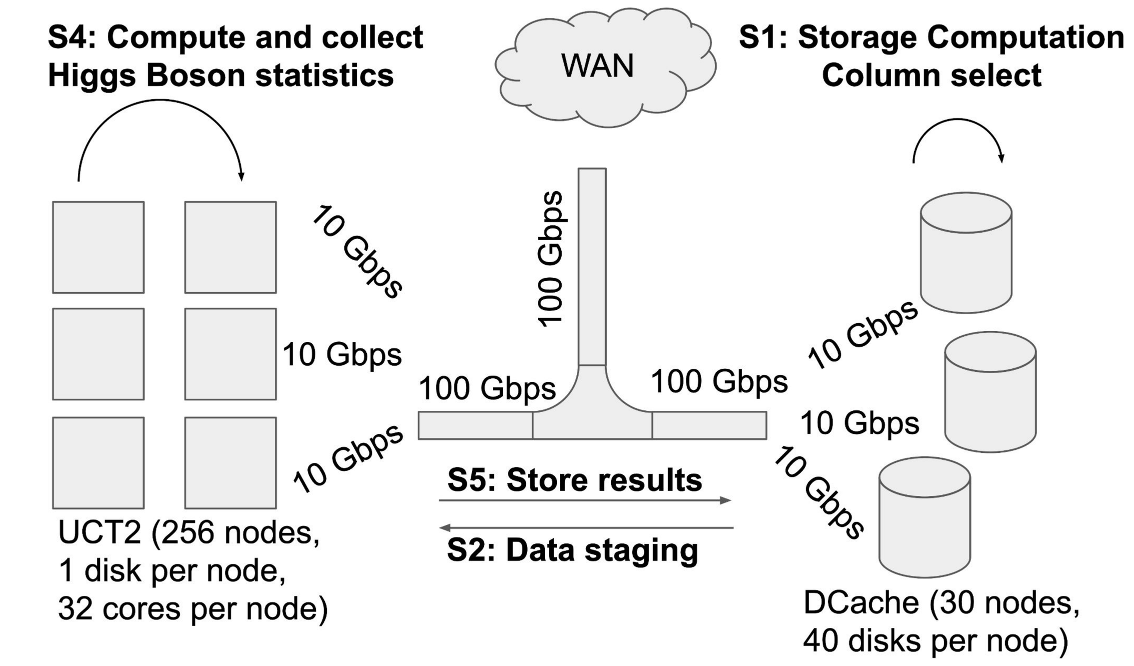
CERN and HEP community built a tiered grid [1] around the globe to meet the storage and analysis demands.

- Limited by cost considerations, the conventional wisdom on developing the HEP data centers is to buy disks to host 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.

- Developed a parametric model of a tier-2 data center, workhorse for analysis.
- Evaluated upgrade options to increase analysis capacity including computational storage

Methodology



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

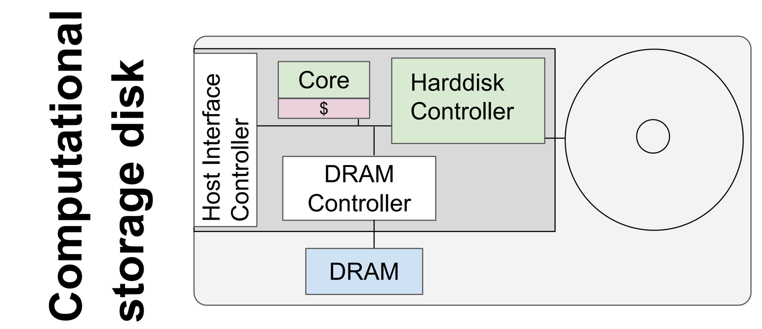
1. In-storage computation at DCache cluster
2. Data staging: DCache→UCT2
3. In-storage computation at UCT2 cluster
4. Higgs boson stats compute at UCT2 cluster
5. Store back results: UCT2→DCache

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

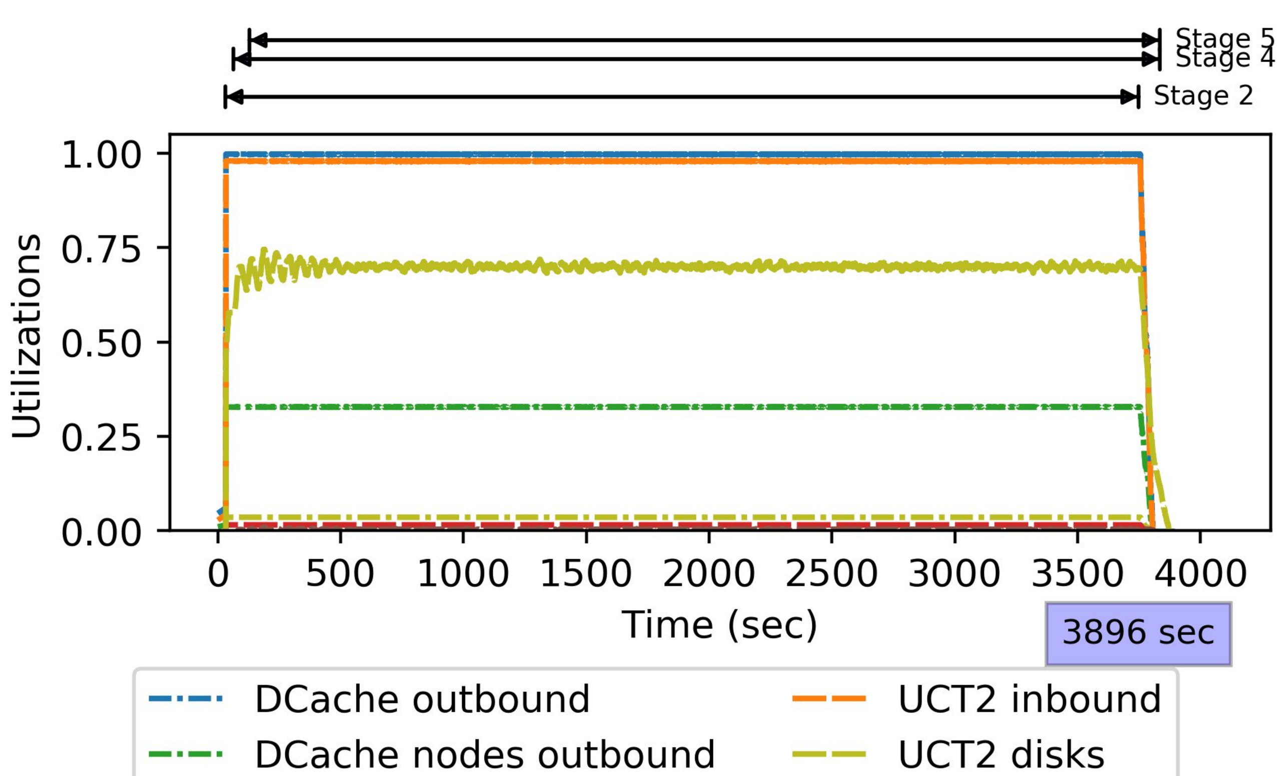
Workload.Higgs boson analysis with 47TB CMS dataset is used as the workload. It independently maps two steps of computation to each collision event.

- Apply filters to muons, global muons, electrons tracks.
- Calculate statistics on filtered muons, global muons, electrons tracks in the event for potential Higgs bosons.

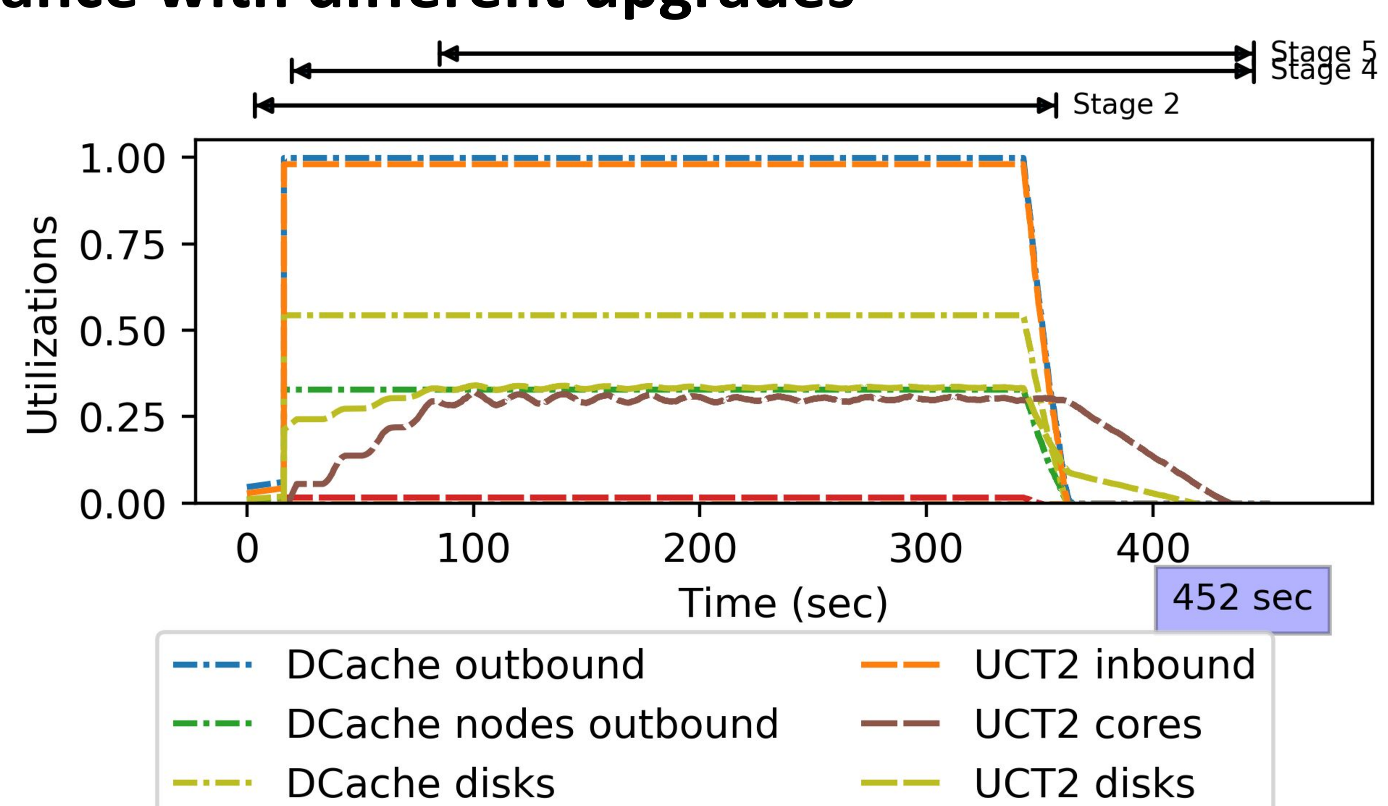
Approach. A C++ simulator implements the performance model with task-level granularity, where each stage for a job is instantiated as a task. It simulates task progress under the modeled customizable resource properties and produce the application performance measured in latency.



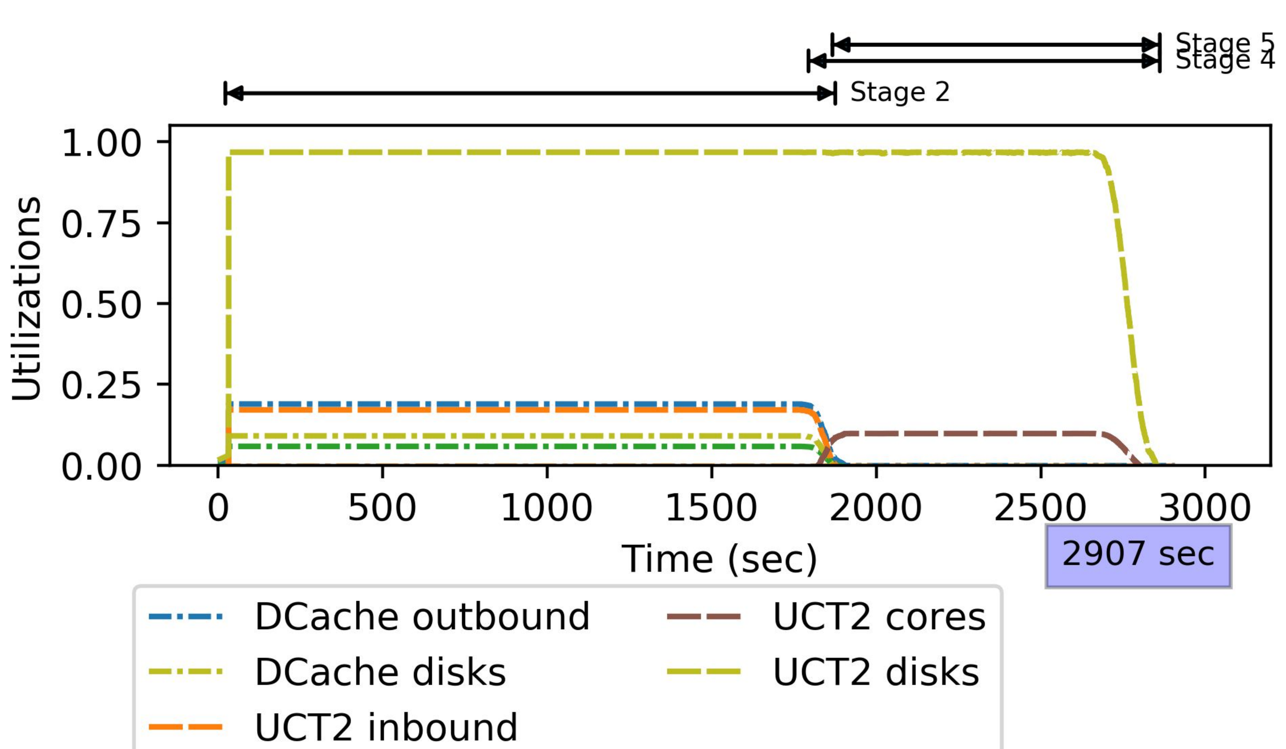
Higgs boson analysis performance with different upgrades



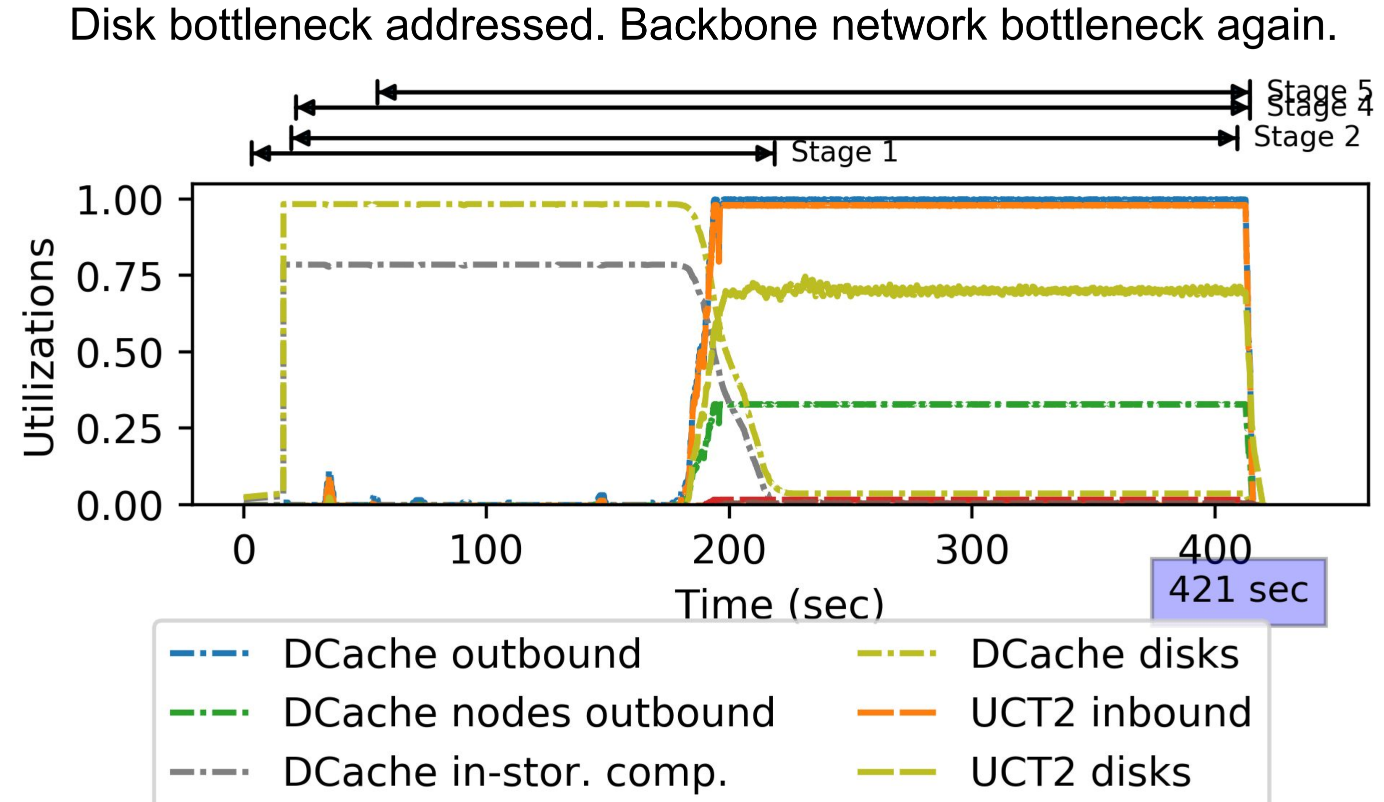
①**Baseline. 1x**
Backbone network bottleneck.



③**Backbone network 1000 Gpbs & UCT2 SSDs. 8.62x**
Disk bottleneck addressed. Backbone network bottleneck again.



②**Backbone network 100 Gpbs -> 1000 Gbps. 1.34x**
Network bottleneck addressed. UCT2 disk bottleneck



④**Computational storage in DCache disks to column select. 9.25x**
Performance improves nearly 10 times.

Analytical model

Because of the highly parallel nature of the workload, we could derive the following analytical model for analysis performance:

$$bn \doteq BackboneNetworkBandwidth$$
$$dd \doteq DCacheAggregateDiskReadBandwidth$$
$$ud \doteq UCT2AggregateDiskReadWriteBandwidth$$
$$uc \doteq UCT2AggregateComputeThroughput$$
$$ApplicationLatency = \frac{EventDataSize}{min(bn, dd, ud, uc)}$$

Comparing analytical model and simulation

	Baseline	1000 Gbps Backbone	1000 Gbps Backbone UCT2 HDD -> SSD
Data Size	47.2 TB		
Bottleneck	Backbone network	UCT2 disks	Backbone network
Bandwidth	100 Gbps	136.7 Gbps	1000 Gbps
Analytical Latency (s)	3776	2762	378
Simulated Latency (s)	3896	2907	452

Our analytical model for analysis performance agrees well with the simulation results
For ③ 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 hat breaks 'Liebig's barrel' trade-off as a local optimum.

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.

	Area(mm^2)	Power(mw)
Computational storage disks	1.58	9.64
Extra CPU in DCache	63.41	23216.00
Ratio	40.2	2408.5

Compute elements in computational storage disks is modeled after ibex cores [4]. Numbers for CPU are extrapolated from Skylake-SP specs and die-shots [5].

Acknowledgement

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

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] C. Zou and A. A. Chien, "Empowering architects and designers: A classification of what functions to accelerate in storage," UChicago CS TR-2020-02.
[4] 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
[5] "Skylake SP Die Shot," https://i.imgur.com/Na64wWe.jpg