LA-UR-22-30125

Title: Accelerating Scientific Data Analytics with SK hynix KV-CSD

Author(s): Zheng, Qing
Manno, Dominic Anthony

Intended for: Report

Issued: 2022-09-29 (Draft)

Accelerating Scientific Data Analytics with SK hynix KV-CSD

High-Performance Computing (HPC) storage is rapidly evolving. Unlike what it used to be in the early 90s, today's HPC shares many of the concerns, characteristics, and requirements of modern big data and cloud computing platforms allowing technologies and innovation to flow in both directions benefiting both sides. While massively-parallel computing and bulk data transfers between compute and pooled storage nodes will continue to be important to HPC, being able to efficiently create, store, and manage massive amounts of small objects and handle complex queries with high data selectivity are becoming as important as emerging applications such as AI, machine learning, and experimental science pipelines become more widespread and traditional simulations are run with rising resolutions thanks to rapidly increasing compute power and memory capacity overall.

For many years, Los Alamos National Laboratory (LANL) has been communicating and leading this transition towards managing both big and small data objects, producing award-winning technologies such as PLFS, IndexFS, GUFI, and DeltaFS with sustained impact on the storage industry. KV-CSD, a recent collaboration between memory giant SK hynix and LANL, is a latest example of this endeavor. The goal is to leverage computational storage capabilities found in tomorrow's all-flash memory enclosures to greatly reduce query latency when high volumes of small scientific data records previously written by a massively-parallel scientific simulation are subsequently read for interactive data analytics.

Background: Big Data is Getting Bigger

Modern HPC data centers are no strangers to big data and this data is only getting bigger. Today's largest HPC sites can routinely generate petabytes of data on a daily basis. To reduce total storage cost, many data centers such as the Trinity supercomputer at LANL — the lab's current biggest machine and the world's fastest computer back in 2015 — have adopted a tiered storage design with warmer tiers made up of flashes and cooler tiers provisioned with high-density rotational drives. With over 900,000 CPU cores and 2PB of main memory, a single Trinity simulation application can easily produce multi-petabyte of data by periodically saving its in-memory state to flash at 3.2 TB/s, finishing each full-memory dump in just about 10 minutes. This rate will soon be dwarfed by Trinity's successor, an exascale supercomputer expected to be at least 25 times larger than Trinity, let alone the very next-next supercomputer that the lab is currently planning for the year 2027.

As simulations run, their in-memory state is periodically saved to storage for two reasons. One is to enable quick restart from failures. The other is to enable post-hoc data analytics — ones that take place *after* a simulation. By periodically recording a simulation's state as it runs, a scientist is able to later analyze its development by tracing its state over time and potentially configure new simulations to start from previously recorded states for scientific exploration.

Emerging Trends: Data Analytics Increasingly Selective

Scientists study their data by running queries against it. These queries have traditionally required seeing all of a simulation's data output over time and are typically processed by reading back an entire output dataset from storage for parallel in memory operations (such as movie making) on compute nodes. However, as simulations become larger and increasingly fine-grained, their analysis has consisted more frequently of queries whose predicates tend only to match a tiny subset of data on storage, which makes reading back an entire dataset no longer necessary and instead calls for efficient mechanisms that allows scientists to read back just the data of interest from a large dataset without having to perform a full data scan which is becoming increasingly time consuming as data size grows.

To avoid performing a full data scan for data analytics, a reader will need to know the locations of the tiny subset of data on storage that matches a query. A reader will be able to do that when the dataset it searches has been preordered on the query key and one way to make this happen is to have the simulation write the data in that order.

Unfortunately, modern scientific applications write their data without necessarily having the ability to consider the performance of the queries that follow the writes. The top priority of the writer is almost always to write the data as quickly as possible so that a simulation can spend most of its job time doing useful computation rather than waiting on storage I/O. To achieve this goal, many scientific applications are programmed to write their data in the most convenient order that allows for the best write performance, which does not necessarily optimize any followup queries. As a result, when the data is indeed queried on a key that does not happen to match its writer (a case we expect to be increasingly true for exascale applications with post-hoc data analytics requirements), a reader will continue to have to perform queries out of order, continue to have to read back an entire dataset from storage even when query selectivity is high, and continue to have to incur large data transfers resulting in long query processing times.

Case Study: Tracing the State of Few High-Energy Particles

One real-world example of the long delays experienced by domain scientists performing data analytics featuring highly-selective queries is the Vector Particle-In-Cell (VPIC) application widely used at LANL.

VPIC is a parallel particle simulation framework that divides the simulated space into cells and distributes ownership and management of each cell among the processes in the simulation. Within each cell a process manages a set of moving particles based on underlying principles from physics. Individual particles often move between cells as the simulation progresses. This is done by transferring the particle state between two processes managing neighboring cells.

Large-scale VPIC simulations powered by the world's largest high-performance computing platforms manage the state of trillions of particles across hundreds of thousands of CPU cores.

VPIC simulations run in timesteps. Every few timesteps VPIC stops and writes the state of all particles to storage. Typically, the analysis of a VPIC simulation run occurs after the simulation concludes. The problem the team of VPIC scientists are looking at involves the trajectories of a tiny subset of particles that end a simulation with an unusually high energy. The trajectory of a particle includes its travel path through the simulated space over time and its state (such as the particle's kinetic energy) for each step of the path. High energy particles of interest are identified at the conclusion of a simulation.

Finding the trajectories of a few high energy particles in a large simulation is a challenge for several reasons. First, the identity of the high energy particles of interest is not known in advance, so they cannot be marked or traced when the simulation starts. Second, as particles migrate between simulation processes during the course of a simulation, a given particle's state is scattered across the nodes in the cluster running the simulation. Third, once the simulation completes and high energy particles are identified, the entire simulation output needs to be read back and scanned in order to extract the needed trajectories.

Reading back an entire simulation output and filtering out relevant information in memory can be prohibitively expensive as VPIC simulation size grows. For example, previous results show that using full data scans it would take as many as 40 minutes to lookup a single high-energy particle trajectory from a 2-trillion-particle dataset using the Trinity supercomputer at LANL. A total of 131,072 Trinity CPU cores would be required to concurrently read back the dataset in order to finish the query in 40 minutes. Scientists would experience even longer delays if fewer CPU cores were used for the query and yet even further longer delays if the particle dataset was stored in a cooler storage tier (such as the lab's campaign storage) rather than the warm storage tier as used in the experiment.

Towards Ordered Key-Value Based Computational Storage

To enable scientists to more easily analyze their simulation output, LANL is collaborating with SK hynix to re-imagine storage for next-generation HPC platforms. KV-CSD — SK hynix's Key-Value based Computational Storage Drive and the world's first ordered key-value computational flash storage — allows scientists to represent data as simple key-value pairs and insert data to storage in arbitrary write orders while the flash drive has its own on-storage compute power — currently in the form of an FPGA card and a set of four ARM processor cores — to reorder and index the data for fast query operations. KV-CSDs sort data in key order and can efficiently answer both point and range queries. Additional secondary indexes can be configured by scientists at data insertion time to enable fast multi-dimensional queries (queries with predicates on two or more data attributes). KV-CSDs perform data sorting and index operations near the flash chips that store the data so that data can be reorganized efficiently with minimal data movement between storage and the compute nodes and much reduced storage software overheads currently seen in parallel file systems such as Lustre and software-based key-value stores such as LevelDB and RocksDB.

With KV-CSDs performing data sorting and indexing on behalf of the simulation, scientists are able to efficiently query their data and read back just the data of interest as opposed to having to stream back an entire dataset spending excessive time and compute resources on the query phase. KV-CSDs enable massive query acceleration by significantly reducing the amount of data transferred between a reader and the storage, a capability that will only become more important as highly selective data analytics become more frequent and simulations become bigger and more complex.

Co-Demonstration at Flash Memory Summit

To demonstrate the potential of ordered key-value based computational storage for large-scale scientific data analytics, we have presented KV-CSDs at this year’s Flash Memory Summit with a demonstration of using KV-CSDs to accelerate VPIC particle trajectory queries.

We prepared a VPIC particle simulation dataset consisting of 16 timesteps and 15M particles per timestep. Without KV-CSDs, to lookup a particle a scientist will have to read back the entire dataset which takes around 55 seconds with the demonstration machine. With KV-CSDs, on the other hand, looking up a particle only takes 0.3 second, a 183 times speedup. We expect much higher speedups when data size increases to production levels.