Snowmass2021 - Letter of Interest

IceCube and IceCube-Gen2 Event Management Service

Thematic Areas: (check all that apply □/■)
□ (CompF1) Experimental Algorithm Parallelization
□ (CompF2) Theoretical Calculations and Simulation
□ (CompF3) Machine Learning
■ (CompF4) Storage and processing resource access (Facility and Infrastructure R&D)
■ (CompF5) End user analysis
□ (CompF6) Quantum computing
□ (CompF7) Reinterpretation and long-term preservation of data and code

Contact Information:
Benedikt Riedel - briedel@icecube.wisc.edu

Authors (alphabetical):
Claudio Kopper - koppercl@msu.edu
Benedikt Riedel - briedel@icecube.wisc.edu
David Schultz - dschultz@icecube.wisc.edu
on behalf of the IceCube[1] and IceCube-Gen2[2] Collaboration [analysis@icecube.wisc.edu]

Abstract:
The IceCube Neutrino Observatory is a cubic kilometer neutrino detector deployed at the South Pole, focused on detecting GeV-EeV energy neutrinos. IceCube measures neutrinos by detecting the optical Cherenkov photons produced in neutrino-nucleon interactions. With the increasing data volume for processing and analysis selection, we propose a new system for storing, searching, and retrieving event data. We attempt to answer the questions of metadata-data mapping, managing access to storage systems, and mapping the workflow to resource pools as efficiently as possible.

[1] Full author list available at https://icecube.wisc.edu/collaboration/authors/snowmass21_icecube
[2] Full author list available at https://icecube.wisc.edu/collaboration/authors/snowmass21_icecube-gen2

The IceCube Neutrino Observatory [1] faces an ever-increasing data volume, both due to general accumulation of observations and from the IceCube Upgrade [2] and Gen2 [3]. We propose a new system to help access and process events, both for production data and simulation processing and as an initial data source for user analysis.

Issues Addressed

Metadata-Data Mapping: Campus or national-scale data storage resources, e.g. XSEDE storage allocations, generally require users to build their own data management system or rely on the data management capabilities of the file system employed by the respective service provider. Popular distributed file systems, e.g. Lustre or GPFS, do not support detailed metadata-based data searches, rather the user typically relies on directory structures, file names, or a combination thereof. This significantly reduces metadata fidelity and usefulness in exploratory data analysis and sharing data. Mapping metadata to data will only be exasperated as object storage systems replace traditional file systems. Object storage systems do not support direct file access and nested directory structures, disconnecting researchers from their data even further.

Managing Storage Allocations: A single storage system or allocation is or will soon no longer be sufficient or scalable to host all of a researcher's or research group's data. Managing heterogeneous and\or distributed data storage systems is a daunting task for even large-scale international scientific collaborations.

Mapping Workflow to Resource Pool: This challenge is two-fold: optimizing resource utilization by mapping task resource requirements onto the resource pool and moving data between different stages of a workflow. Optimizing the resource utilization addresses a long-standing research question within computer science. Efficient scheduling of tasks across distributed and heterogeneous compute resources is also a must for researchers running large-scale workflows. For example, running a workload designed for a fast GPU on a single CPU core may take weeks. Another option is enabling partial completion of work unit sets or tasks, allowing for a much better exploitation of resources, esp. those with limited resources, preemptable resources, or remaining compute time.

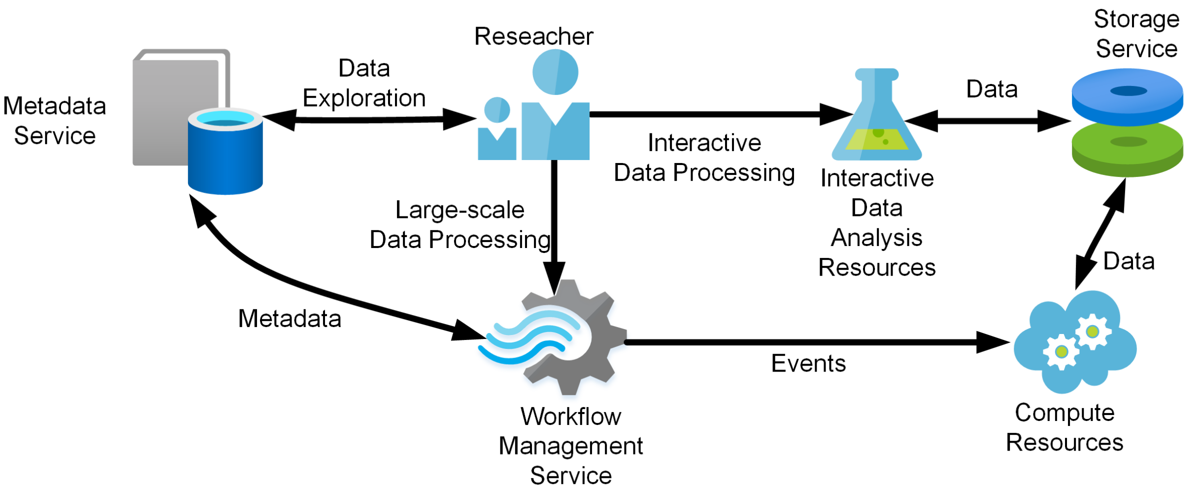
Moving data between different stages of a workflow programmatically can be difficult. It requires network connectivity between different compute sites, orchestrating data movement and clean-up, ensuring data integrity and provenance, and scaling the system easily and quickly. This is a tall task for even the most experienced user. While we cannot necessarily change network connectivity or available bandwidth, the rest can be tackled with a well-designed system.

Event Management Service

The Event Management Service (EMS) will address these challenges in three key ways:

- exploring approaches to make work units as small as possible
- distributing and managing these work units using commercially-supported and standardized tools (such as message brokers and NoSQL databases)
- presenting an interface to facilitate the separation of researchers from the details of the storage system(s)

EMS architecture and function is based on a ”micro-service” architecture and an event-based data processing paradigm, in which the interaction between distributed services is facilitated through the interchange of “messages” or retrieval of information through key-value lookup. We want to use and compare these technologies to allow significantly more fine-grained data and workflow management, where a task fetches the input data or simulation settings according to the available resources rather than crudely approximating the required resources for a given task.


Metadata Service
Data Exploration
Reseacher
Interactive Data Processing
Data
Storage Service
Interactive Data Analysis Resources
Large-scale Data Processing
Metadata
Data
Events
Workflow Management Service
Compute Resources

Event lookup and retrieval is specifically designed to abstract away the storage system(s). Metadata can be richer and more contextual, as well as easier and faster to search. This should hopefully lead to requesting only necessary data, instead of scanning large data samples for a small set of events. The physical storage abstraction allows more freedom to system administrators to architect storage systems, as POSIX compliance is unnecessary and object store systems popularized by cloud providers handle large-scale data volumes better. Additionally, users of the system need not be aware of exactly which storage system the data comes from, allowing more data mobility and easing migrations.

Technical Challenges

One challenge for IceCube is the size and number of events. Each day of detector data averages 40 million events, so a catalog of 10 years of data is O(100 billion) events. Just storing event id and location information in metadata would be 10+ TB, so a rich metadata for performing analysis selections is easily 100s of TBs. The scale is daunting, but several industry databases handle this scale. There is also the option to exclude events unlikely to be useful, as the majority of IceCube data is background. This is routinely done by research working groups as a first step in processing.

Another problem is the actual storage of events. The ideal case is storing individual events per file or storage object, but the small size of events makes this challenging. The average event size is O(1 KB), and most filesystems are slower at storing and retrieving files that small. It will likely be necessary to have a cache layer of flash storage and\or in-memory storage for the more popular events. Fortunately, the abstraction between users and storage access should allow a more seamless tiering experience.

Conclusions

The Event Management Service is an integrated data organization and distribution system that will allow researchers to separate metadata from the data itself, reduce the burden of data movement, and better match computing resources and data. The EMS is an effort being initiated by the IceCube Collaboration, but feedback and collaboration with other large scientific groups and industry partners is being developed.

References

[1] M. G. Aartsen et al. The IceCube Neutrino Observatory: Instrumentation and Online Systems. *JINST*, 12(03):P03012, 2017.

[2] Aya Ishihara. The IceCube Upgrade – Design and Science Goals. *PoS*, ICRC2019:1031, 2020.

[3] M.G. Aartsen et al. IceCube-Gen2: The Window to the Extreme Universe. 8 2020.