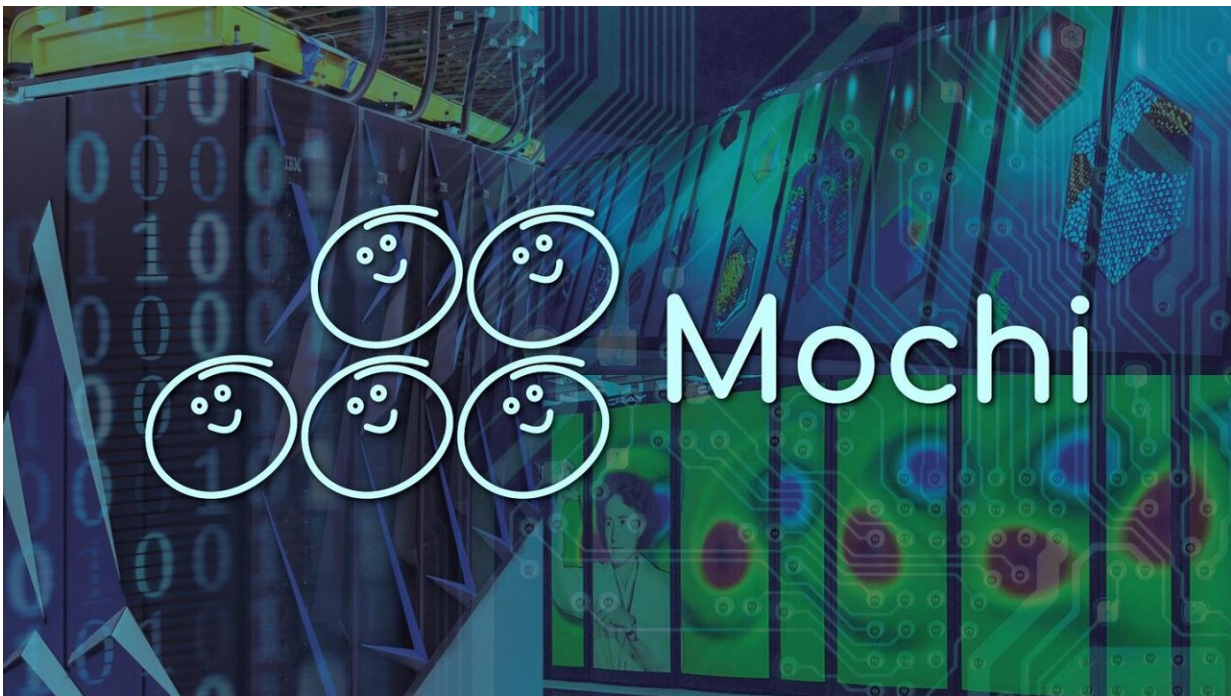# Argonne's new menu of data storage software helps scientists realize findings earlier

June 2 2020, by Jo Napolitano



An Argonne-led collaboration developed Mochi, a novel data navigation system that will allow scientists to rapidly customize a suite of data services to suit the needs of a specific domain and problem, shaving weeks or months off the time needed to produce actionable information from collected data. Credit: Argonne National Laboratory

Most scientists, no matter their discipline, rely on data storage systems to help them draw conclusions from their work.

But their needs are vastly different. A scientist studying weather, who collects data from instruments spread across the world, might want to sort the findings by date or region, while another, studying the molecules that make up a virus, might generate a single large data set to evaluate the virus's response to potential treatments.

It's nearly impossible to build a single data storage system that would satisfy both—a tweak that might help one scientist could make the system less efficient for another.

"Anyone can imagine a custom storage system to solve a particular science problem, but it would take years to get it fully complete and ready for production," said Phil Carns, principal software development specialist in the Mathematics and Computer Science (MCS) division at the U.S. Department of Energy's (DOE) Argonne National Laboratory.

Carns is technical lead of a team set to solve this problem by identifying a collection of building blocks scientists can pull together to craft a data storage system designed to address their own specific needs. Rob Ross, senior computer scientist in MCS, is principal investigator for the new technology, which he and Carns call Mochi. The Mochi team includes researchers at Argonne, DOE's Los Alamos National Laboratory, Carnegie Mellon University and The HDF Group, an Illinois-based nonprofit dedicated to advancing state-of-the-art open source data management technologies.

"We're doing this so that when someone wants to build something new, they are not starting from scratch," Carns said. "They are selecting from a menu of things they need to suit their data."

For example, the scientist studying weather data may choose a component that can index information along multiple dimensions and combine it with another component that can aggregate data from many

sources, while the scientist studying molecular data may choose a component that caches frequently used information on local devices to speed up machine learning algorithms.

Each scientist benefits from using a specialized storage service without having to create one from scratch.

Regardless of which components are used, they all share the same underlying communication framework, known as Mercury, to efficiently move large volumes of data between storage and compute resources.

The technology is in high demand as scientists around the world prepare for DOE's first exascale supercomputers, Aurora at Argonne and Frontier at DOE's Oak Ridge National Laboratory. Each will be able to complete a billion billion (i.e., a quintillion) calculations per second, making them a million times faster than a high-end desktop computer.

Mochi, which already has proof of concept, is currently in the testing phase. Its source code, examples and documentation are available on the project website for scientists who need to access large volumes of data to do their work.

Carns, who has been working on the project since it kicked off in 2015, said many scientists struggle with managing the data their experiments generate.

"A common problem across the sciences is that researchers are capable of creating data faster than it can be analyzed," he said. "Identifying those few bits of data that are particularly interesting and relevant to the problem they're trying to solve can significantly slow the process of making a discovery. For some scientists, improving their ability to process data could shave weeks or months off of the time needed to produce actionable information from their research."

Already, the technology is being evaluated to analyze data from particle accelerators, which has applications in fields such as medicine and materials science; study particle simulation data, with the goal of finding new sources of energy, such as nuclear fusion; and store machine learning data that can be used to identify cancer treatments.

 **More information:** www.mcs.anl.gov/research/projects/mochi

Robert B. Ross et al. Mochi: Composing Data Services for High-Performance Computing Environments, *Journal of Computer Science and Technology* (2020). DOI: 10.1007/s11390-020-9802-0

Jerome Soumagne et a. Advancing RPC for Data Services at Exascale. sites.computer.org/debull/A20mar/p23.pdf

Provided by Argonne National Laboratory