Leveraging the Computational Fabric for Science-Driven Data Refactoring

Lipeng Wan, Scott Klasky, Matthew Wolf, Norbert Podhorszki, Jieyang Chen, Ben Whitney, Jong Choi, Ruonan Wang

As we compare growth in velocity and volume of data generated by simulations and instruments over time to computing resources, we see that data management will become a severe bottleneck that hinders the scientific discoveries. To address this critical challenge, our research has shown that we must provide services that refactor the data (reorder and reduce data according to its information) and prioritize delivery of scientific insights based on scientists' intentions. Moreover, such services need to guarantee certain accuracy and prevent the refactored data from leading to faulty conclusions. Our vision is that the data refactoring service will be a necessary technique to facilitate future large-scale science.

Since data refactoring services are often computationally expensive and might degrade the overall performance of scientific workflows if they were to share the computational resources with the simulation and analysis codes, they should be handled in any computational fabric that is near the data. Experiences with emerging technologies such as "in-network computing" and "in-storage computing" give us some insight into the opportunities for future in-fabric data refactoring service. By leveraging such technologies, the data generated by scientific experiments and simulations can be automatically refactored during its movement from federated instruments or HPC systems on the way to storage or to where scientists run their analytics routines.

Specifically, we see in-fabric data refactoring service as providing some of the following key capabilities: 1) decomposing the data into different pieces and identifying what are the most valuable data pieces for scientists using AI/ML algorithms; 2) moving and storing all these data pieces to different locations based on the requirements of each scientific campaign; 3) building efficient metadata or index structures to tag and track all these data pieces so that scientists can easily find the data piece they need no matter where it is located; 4) capturing the provenance of all scientific campaigns and leverage AI/ML techniques to learn and understand what scientists would like to do with different types of data in the future. Most importantly, since the service is supported in fabric, these capabilities will be available for not only immediate data access but also longer data lifecycle.

We provide two current examples that demonstrate the opportunities for this vision. DCA++ is an HPC simulation for solving quantum many-body problems with cutting edge quantum cluster algorithms. Now scientists have had to restrict themselves to looking at a statically selected 2D slice of the tensor, but as the complexity of their science scenarios have grown, they are interested in new techniques that allow them to access more of the tensor data (normally a six-order of magnitude increase in size). If data refactoring service is supported in fabric, it will enable scientists to dynamically and efficiently access the tensor data at different resolutions based on their performance requirements, thus reducing the potential loss of scientific insights caused by the restricted data selection.

Similar trends exist as well for observational and experimental data. Based on early results from working with the Square Kilometer Array (SKA) radio astronomy project (https://www.icrar.org/summit/), in-fabric data refactoring will be critical for their success in the future. In the next 10-20 years, SKA will generate over 1 PB/s 24x7. Their current workflow already involves customized in-fabric computing (e.g., the correlator done in a FPGA) to remove noise and redundant data. Although this can reduce the original data by a factor of 100-1000, the remaining are still too huge to process. Moreover, radio astronomy data is particularly challenging to manage compared to other types of scientific data since it has much longer lifecycle. This suggests that we need to fully take advantages of in-fabric data refactoring to further concentrate the information content in SKA data, and extract, move and store only the most relevant pieces.

In summary, having the capability to generate and store a huge amount of data does not necessarily lead to great scientific discoveries. Instead, having the capability to efficiently and intelligently refactor the scientific data so that scientists can easily obtain the most valuable scientific insight is more critical. Therefore, we envision science-driven, in-fabric data refactoring will play an important role in changing the way how people do science in the future.