

Reference ID: 11220834587_Orf

Reference ID: 11220834587_Orf

Submission Date and Time: 12/13/2019 1:29:14 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Consent Statement: "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

Consent answer: I consent to NSF's use and display of the submitted information.

Author Names & Affiliations

Submitting author: Leigh Orf - University of Wisconsin

Additional authors: None

Contact Email Address (for NSF use only): (hidden)

Research domain(s), discipline(s)/sub-discipline(s)

Atmospheric science; numerical weather prediction; mesoscale meteorology; supercell thunderstorms

Title of Response

The challenge of post-hoc data analysis and the need for lossy data compression

Abstract

In the physical sciences, supercomputer simulations can produce immense amounts of data that then need to be analyzed to have scientific value. While the approach of in-situ data analysis, where data are analyzed while the simulation is active, can be useful in some contexts, analysis from saved model data is crucial for complex simulations that will require months/years to analyze. Hence, a robust CI

infrastructure for the efficient saving, storage, and massively parallelizable post-hoc analysis techniques is crucial. A hierarchy of storage platforms such as SSD, spinning disks, and tape should interoperate seamlessly. Data should be able to be present for analysis on site, rather than requiring it to be moved offsite, as moving PB of data is not feasible. Further, scientists must go beyond their comfort zones with data and strongly consider the effective use of lossy floating point compression in their saved data. GPU technology can play a role in post-hoc analysis, and efforts to make GPUs available and easier to use are highly encouraged.

Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s). Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

In my field of atmospheric science modeling, and in other modeling fields, supercomputing technology has enabled simulations to be run at very large scale. In order to analyze these simulations, large amounts of data must be saved, in the absence of in-situ data analysis techniques (that are not always feasible). I have spent nearly a decade figuring out a way to efficiently save tornadic supercell data at high spatial and temporal resolution in order to visualize and analyze this data sufficiently, with temporal data being successfully saved as frequently as every model time step. As such I have relied on the large storage capacity of supercomputers, self-describing data formats such as HDF5, and lossy floating point data compression. Further, I rely on analysis tools that are designed to visualize and analyze data at large scale. The main challenge is not being overwhelmed with data. I have found the time to accomplish meaningful analysis is directly proportional to the amount of saved data, which in my work, can get into the petabyte range quickly. How can we analyze such an enormous amount of data and pick out the processes that will move the field of atmospheric science (and forecasting) forward? I have found that it is easier to get published when working on simpler simulations with fewer degrees of freedom - however these coarse simulations miss critical scales of flow that are most certainly physically important in reality. Further, new hardware topologies including GPUs provide a promising platform for conducting analysis, but require a whole new programming paradigm. Finding effective ways to exploit GPUs is a major challenge.

Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s). Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data

integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

It is best when large data outputs from supercomputing simulations can be analyzed and visualized without being moved to another site. Further, because it can take months to years to analyze large simulations, having a semi-permanent place to store data without fear of destruction is critical. Currently scratch space on supercomputers is scrubbed regularly, and moving data to storage archives, even if on the same network as the supercomputer, can be slow. I would be useful for there to be different hierarchies of storage, from fast SSD to large but slower spinning disk arrays, to high density but very slow tape storage. Movement between these three levels should be as seamless as possible, and not necessarily require creating TAR files that need to be unpacked. There is a movement in computer science towards the development of in-situ data analysis techniques. The advantage to in-situ data analysis is that you are doing your visualization and analysis while the simulation is running on the supercomputer. However, for my own type of research where the simulation is very complex, analysis techniques that are applied to saved model data are crucial. In-situ analysis is nearly worthless to my own work because I cannot get meaningful analysis done in an ongoing simulation; the analysis I need to do requires many different looks at the full completed simulation, something that can only be done realistically from saved data (not everyone can commandeer a supercomputer for days at a time). Hence, it is critical that analysis techniques working on saved data are robust and parallelizable. I have developed my own file system for this purpose (it is similar in function to ADIOS). I have found that the amount of resources required to analyze my storm data is a fraction of what was required to create it, hence, more modest hardware can be put to the task of analysis and visualization. The data therefore contains all of the treasure, as the simulation itself is fleeting. So we need a CI that provides both compute capability, immense semi-permanent storage capability, and post-hoc (as opposed to in-situ) data analysis capability. GPUs, I believe, can play an important role in the latter, and I am encouraged by the Frontera supercomputer topology which will include a CPU side and a GPU side. But, effectively utilizing the GPU side for analysis will be a challenge, and perhaps tools can be developed that are cross-disciplinary for analyzing multidimensional data. Another challenge is getting scientists to use lossy floating point compression in their saved data. I have found that many scientists are very averse to compressing their data. However, with lossy floating point algorithms such as ZFP (that I use) careful compression, specially tailored to each variable, can result in data savings of 20-50 times uncompressed, opening the door to new kinds of analysis while also reducing pressure on storage facilities. Further, scientists should be aware that they are able to specify how much accuracy is required for each variable and apply compression accordingly. However this is a "cultural" issue that will require perhaps carrot/stick types of incentives (i.e., if you don't have a really good reason to NOT compress your data lossily, but insist on it, you won't get a full allocation etc.). Regarding cross-disciplinary/domain agnostic approaches, to a certain extent I believe this is possible, but there are fundamental differences between, say, output from a large climate model simulation and output from an earthquake / galaxy / genetic / cell / molecule simulation. I recommend focusing on strategies that will overlap between disparate fields, but it will not always be possible.

Question 3 (maximum 300 words) – Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

In-situ data analysis is not the solution to the data problem and will only work in specific cases. Hence, we are "stuck" with having to analyze data on disk, and, due to exploding supercomputing capability, this data is huge. Further, new data saving strategies are necessary in order to save data efficiently (i.e., not taking an inordinate amount of wallclock time) and for the resulting data to be in a format that is straightforward to analyze. A pipeline for data analysis should be part of future supercomputing allocations. Simply saving large amounts of uncompressed data from a simulation shouldn't cut it anymore. Scientists should explore lossy compression approaches and also come up with an analysis pipeline that is realistic. However this cannot happen without tools and hardware that can facilitate this. Every scientist cannot be expected to come up with his/her completely new data analysis pipeline. My thought at this point is that GPUs can serve as a good platform for doing a lot of the post-processing/analysis/visualization. While GPUs are being looked at as primarily a visualization and AI platform, GPUs are basically little supercomputers if programmed efficiently. So, anything that can help assist scientists in using GPUs efficiently will be of use. There will still be a burden on the scientist to learn new things, but it doesn't have to be done in a vacuum - training programs and analysis software tuned to GPUs could be of great value.

-- End Submission --