

Reference ID: 11227070561_Livny

Reference ID: 11227070561_Livny

Submission Date and Time: 12/16/2019 4:54:23 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: Miron Livny - University of Wisconsin-Madison

Additional authors: Brian Bockelman, Morgridge Institute for Research

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

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

Computer Science, Distributed Computing, High Throughput Computing

Title of Response

Storage Management for Data in Transit

Abstract

Throughout its life cycle, science data is moved through buffers. Close-by storage is small, fast, and often has rich POSIX semantics whereas far away storage is slow, shared, large, and often non-POSIX. We refer to these buffers as Transit Storage (TS). The life cycle of science data can be viewed as consisting of two main phases – generation and processing. In the first phase the data transits from a generation point to

Permanent Storage (PS). During the processing phase, the data travels from the PS location to the processing unit. The NSF Cyberinfrastructure (CI) ecosystem is lacking a framework and software to manage contention for TS. This void is critical for High Throughput Computing (HTC) workloads that rely on distributed processing and storage. Applications, schedulers, workload management systems, and administrators are missing means to request, wait for, and grant TS capacity. These concepts are engrained at the campus for managing computing: a framework is needed now for TS. We find an increasing number of researcher workflows that manage both compute and TS; investment in data services and infrastructure in this area will have transformative impact by integrating a broader set of campus users into the data-intensive research computing ecosystem.

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.

Considering TS in cyberinfrastructure is not a foreign concept. Large CI infrastructures like ATLAS and CMS have begun to consider the challenge: their data management software can fill site-level buffers up to a quota and track requests unfulfillable due to a lack of space. The workflow planning components can reorder processing workflows based when datasets are available from PS and otherwise plan out future workflow execution based on predicted available space. ATLAS and CMS demonstrate dealing with finite TS is inevitable - even when managing hundreds of petabytes of storage and buffers. However, these examples fall short in that their ecosystems fail to “scale down” to the level of single campuses or PI-driven labs and they do not leverage TS infrastructure - but were as bespoke systems because site storage lacks the necessary primitives. As an analogy with computing resources management, imagine how few researchers could utilize Frontera if TACC staff simply hand-assigned CPUs and relied on the PIs to build their own batch systems. The challenge is more acute at the campus level; at UW-Madison, we’ve encountered diverse examples such as: ● Botanists performing high-throughput phenotyping of corn needing to move thousands of images from scanners at greenhouse facilities to data archives and from data archives to be processed at computing facilities. ●

The Cryo-EM facility will generate up to 8TB of images per day over many experiments and place the images at onsite storage for a limited duration. Scientists must move the data to long-term storage and computing facilities. ● Dairy Science researchers are equipping research barns with cameras to record activities of individual cattle every 5 seconds. These images must be transferred from the dairy barn (over unreliable internet connections) to on-campus storage. This dataset - eventually millions of images - will be used for training machine learning algorithms to infer the health of individual cattle. In each case, the data moves from PS to TS, processed in the computing infrastructure, and generated output moved back. These workflows are not simple: the computing workflows shouldn’t be started unless there’s space at the PS for generated data; the input data must be moved piece-by-piece to support the computing workflows. These examples - whether from labs at UW or large scientific

infrastructures - share a common thread of data-driven processing, resulting in the need of storage management. Fundamentally, the challenge of managing TS is cross-disciplinary.

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?

Investment into storage management fundamentals is needed for the data-driven CI landscape. We need software and services within an intellectual framework of storage management - tracking usage, declaring policy, enforcing policy, matchmaking between providers and users, and developing the semantics of queueing and scheduling storage allocations. While the core concepts are widely translatable, we believe developing a reference implementation will help better expose them to the S&E community and illustrate their importance. New concepts and services require early buy-in from S&E stakeholders. We believe users - especially small-scale ones - provide the best feedback on system usability and whether we have captured essential needs. Working with a broad set of PI-led labs - as opposed to solely with large experiments - allows one to ensure we tackle core, fundamental problems in storage management as opposed to solving bespoke problems. We believe that in the area of data-intensive CI, starting with small projects focused on real needs is more effective in capturing the end-to-end structural needs. These services must not stand alone but rather exist in the larger data-driven CI community. Once adopted, the concept of TS management will permeate through the design of a computing infrastructure. Storage management must integrate with external software systems; for example,

- A batch system should not start a job unless there will be sufficient space to store output at the PS.
- A workflow management system must not submit jobs until it has moved input data to a TS accessible to jobs.
- Transfer management systems must be able to move data between buffers, such as the PS and TS.

It is critical that this is seen as an essential building block of an end-to-end data-intensive ecosystem and not a monolithic system in itself. The teams providing each piece must work together to adopt common data models and policy languages to communicate between the layers. In addition to the software itself, we need a production-quality distributed platform to integrate with user facilities. In our decade of experience in implementing distributed high-throughput computing services for the Open Science Grid, we have found it is extraordinarily difficult to make software services simple enough for a wide range of sites to deploy (and to support!). The advent of containers and orchestration systems such as Kubernetes has greatly simplified this process of deploying and managing software services. These platforms, supported and packaged by large-scale industry players, serve as an “adapter”: facilities must learn to support a single software service while

the service providers have a more homogenous platform to target. This reduces the investment needed by facilities to deploy new concepts such as finite storage. We believe these types of platforms, especially when co-located at larger facilities and campuses, provide a new opportunity to deploy data-centric CI and will be critical in gaining a wide footprint across the entire S&E community. As with everything in the entire CI, TS has finite capacity and thus should be managed. However, we often provide users with the illusion it is infinite. Recognizing this finite capacity drives the need for a storage management framework. Investment in a production-quality reference platform for storage management is key to demonstrating its value to the community - especially when developed in conjunction with a diverse set of use cases. Integration across the entire end-to-end (jobs, workflows, data access, data transfer) allows the full potential of the framework to be realized; using new service orchestration techniques also allows facilities to participate without becoming mired to a specific platform. Altogether, addressing the challenges in this area would make a transformative change to the NSF S&E community.

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.

Re-evaluating how the NSF S&E community approaches storage management - in particular, integrating the concept of TS - will be most transformative if done within the context of existing organizations and projects. Any explorations or investments in this direction should be campus centric as this is where the majority of S&E activity takes place. We believe the Campus Cyberinfrastructure program provides an illustrative example: it has enabled campuses, aggregating PI-led labs into a coherent plan and provided a conduit to national cyberinfrastructure providers such as the Open Science Grid (OSG). Similarly, investments in data-driven infrastructure must first and foremost solve problems that are present on campus - especially the deceptively “simple-sounding” ones such as those from the UW-Madison outlined in our response to Question 2. The approach outlined also helps workforce development in two ways: it further integrates domain scientists (especially including students working in the labs) into the cyberinfrastructure and the suggested use of industry tools such as Kubernetes at facilities will provide key skills. Engaging students in science domains in advanced cyberinfrastructure (through providing tooling and data services or direct training through venues such as the OSG summer school) will help them reach greater scales in their own area of science and will be a skill that transfers well to the workforce regardless of their future career.

-- End Submission --