Artificial Intelligence Accelerated Cyberinfrastructrure as a Research Multiplier at Small and Midsized Institutions

Mike Austin[1,2], Jim Lawson[1,2], Andrew Evans[1], Andrea Elledge[1], and
Adrian Del Maestro[1,3]
[1]Vermont Advanced Computer Core, University of Vermont, Burlington, VT, 05405
[2]Enterprise Technology Services, University of Vermont, Burlington, VT, 05405
[3]Department of Physics, University of Vermont, Burlington, VT, 05405

The increasing complexity of scientific data-driven workflows has led to an evolution in the types of tools both employed and demanded by interdisciplinary researchers. A survey of queue wait times and user needs at the University of Vermont Advanced Computing Core (VACC) identified an unmet demand for GPU supercomputing. This gap was recently addressed through the NSF Major Research Infrastructure (MRI) program via the design and deployment of a special purpose device delivering 80 NVIDIA V100 GPUs with a custom-built NVMe over fabric filesystem.

Adoption of the device by advanced users in our cestommunity was rapid, and high utilization developed within months. However, we identified a substantial portion of users that were interested in employing machine learning in their research, but have minimal previous experience with high performance computing and were unable to transition their scientific workflows to GPUs. Thus, the efficient and broad utilization of NSF-supported cyberinfrastructure requires coordinated and simultaneous investment in hardware, research computing facilitation, and community tools that can be deployed for training purposes. We have identified a number of areas where such efforts can have a multiplying effect on research, training, and workforce development.

1. Training workshops that provide a low-level introduction to scientific programming and data analysis that include curricular material on machine learning frameworks with domain-agnostic examples. These should be broadly advertised, especially to those outside of traditional STEM disciplines, and highlight that previous programming experience is not necessary for attendance.

2. Deployment of web-based interfaces (such as Open OnDemand [1]) to research computing assets with environments pre-configured for artificial intelligence applications including Jupyter notebooks. We have seen broad utilization of these technologies by researchers, and high demand for deployment in the classroom by

faculty teaching courses with a computational component, requiring little to no software installation on student laptops.

3. GPU virtualization technologies (e.g. NVIDIA vCompute Server) that enable GPU sharing can lead to higher utilization of expensive physical resources, especially for widely used off-the-shelf applications in materials science and chemistry with speedups reliant on a combination of CPU and GPU, that may not be able to exercise all of the processing capability of a single V100.
4. Campus and region-wide coordination of networking, compute, and storage infrastructure to ensure successful implementation of next-generation imaging applications such as cryogenic electron microscopy and light sheet fluorescence microscopy. There is currently a disconnect between the managers and users of these facilities and the cyberinfrastructure expertise required for sustainability.

The procurement and implementation of specialized cyberinfrastructure with advanced artificial intelligence capabilities is playing an increasing and fundamental role in the university research ecosystem. Enhanced planning and support, especially around training and research software development, is required to ensure broad utilization of smart technologies.

[1] D. Hudak, D. Johnson, A. Chalker, J. Nicklas, E. Franz, T. Dockendorf, and B. L. McMichael. “Open OnDemand: A web-based client portal for HPC centers.” J. Open Src. Soft. 3, **622** (2018). https://dx.doi.org/10.21105/joss.00622