MRI: Acquisition of the Heterogeneous Accelerator Lab (HAL) at University of Wisconsin-Madison

Project Description

Instrument Location: University of Wisconsin-Madison, Computer Sciences Department, B380

Computing cluster focusing on accelerators for artificial intelligence and general-purpose computing

1. Introduction

In recent years, the University of Wisconsin-Madison (UW-Madison) campus has experienced a dramatic increase in the demand for accelerators, specialized computing hardware that provides orders of magnitude faster computation on a wide spectrum of research problems. Faculty and students from a wide spectrum of research domains and disciplines use accelerators to enhance their research and support their education mission but are limited by the availability of resources as their project cyberinfrastructure (CI) needs expand. The proposed **Heterogeneous Accelerator Laboratory** (HAL) will offer a campus-wide shared source of cost-effective accelerator capacity. It will also serve as a laboratory for innovative CI technologies for managing an evolving, multi-user, multi-discipline, shared pool of heterogeneous accelerator capacity.

Accelerators are specialized instruments: different scientific problems can have different optimal hardware deployment. Therefore, a variety of hardware configurations is required to enable exploration of artificial intelligence (AI), simulation, and data processing techniques across a wide variety of science domains. To address the challenges of maximizing return on past and future investment in such a fast-evolving market of applications and technologies, the proposed major instrument will consist of heterogeneous hardware and co-processors and build a community around common computational needs. HAL focuses on the deployment, management, and operation of a dynamic and heterogeneous pool of accelerator capacity.

AI applications are a particular focus of HAL. Today, AI activities span the entire campus research enterprise—teaching, and method application and development—where the common thread is the ability to utilize hardware accelerators to train and evaluate models. This AI-driven work requires major commitments in hardware. The proposed investment in HAL will offer UW-Madison faculty and students the capacity they need to maximize educational and research outcomes HAL's proposed configuration includes nodes with different coprocessors (GPUs and FPGAs), large memory, and high-I/O systems to support a variety of science drivers. This is complemented by diverse user interfaces backed by the HTCondor Software Suite (HTCSS) [1, 2]. Resource allocation and job scheduling across HAL will be also provided by HTCSS, which was developed by the UW-Madison Center for High Throughput Computing (CHTC). This positions HAL as a unique open-access platform in which researchers are empowered to create feedback-loops between AI, data, simulations, and insights.

The UW-Madison campus has a long history of leveraging instruments for a wider vision and federating distributed computing resources into a shared environment that democratizes access to computing capacity. HAL is a natural extension to the UW-Madison research computing environment that was seeded by two MRI acquisition awards for PI Livny in 2003 [3] and 2007 [4] for the Grid Laboratory of Wisconsin (GLOW). These awards ultimately formed the core of campus research computing and have since served over 400 campus groups; we expect HAL to have a similar campus-wide, lasting impact. More recently, co-PI Gitter has deployed the "GPU Lab" [5], a seed investment that has demonstrated the viability of the scheduling and software environment necessary for investments like HAL. HAL will serve as a campus-wide resource, managed jointly by a collaboration of research centers and campus IT: American Family Insurance Data Science Institute (AFI DSI), CHTC, Division of Information Technology (DoIT), Morgridge Institute for Research (MIR), and Wisconsin IceCube Particle Astrophysics Center (WIPAC). HAL is designed to have impact beyond the UW-Madison campus boundaries. By sharing its resources via the Open Science Pool (OSP) [6] compute federation, HAL will make regional, national, and international impact by providing resources to off-campus collaborators, such as the Compact Muon Solenoid (CMS), IceCube, and Laser Interferometer Gravitational-Wave Observatory (LIGO) collaborations. Through tight integration into a national organization with a long history of national impact, HAL is well-positioned to be a premier accelerator resource for the nation's science and engineering (S&E) community.

2. Intellectual Merit

HAL's science collaboration consists of UW-Madison researchers across NSF directorates, four of NSF's 10 Big Ideas—Growing Convergence Research, Harnessing the Data Revolution, Understanding the Rules of Life, and Windows on the Universe—, and the S&E community. The applications of HAL can be broken into three categories: training or inference of AI models for research and education; leveraging accelerators

in general-purpose computing and visualization; and using it as a laboratory for offering distributed, high-throughput computing (HTC) resources to the campus and nation. Table 1 summarizes the current CPU and GPU and expected GPU and FPGA usage with HAL. HAL will be more flexible in switching coprocessors, vendor, or software environment and cost-effective compared to public cloud [7, 8].

	2021 Usage		Expected Usage with HAL	
Scientific Domain	CPU k-Hours	GPU k-Hours	GPU k-Hours	FPGA k-Hours
Atmospheric Science	43,830	79	100	
Bioinformatics	3,506	316	876	
Chemical and Biological Engineering	140	35	45	
Computer Systems	281	24	35	
Cosmology	140	35	88	
Cryo-EM	701	175	263	
Data Science	631	158	175	
Digital Agriculture	1,753	438	657	
Large Hadron Collider	126,426	1,534	1,753	26
ML Method Development	20	10	30	
Molecular Dynamics	6,136	1,534	1,753	
Multi-Messenger Astrophysics	115,128	7,013	8,766	26
Simulated Environments	4,208	1,052	1,314	
Others	12,276		10	1
Total	315,176	12,970	15,865	53

Table 1. CPU, GPU, and FPGAs usage from past data (Calendar Year 2021; including both resources on campus and accessible via OSP off-campus) and user-reported future need.

Atmospheric Sciences: The **Space Science and Engineering Center (SSEC) at UW-Madison** is engaged in numerous research efforts to apply machine learning (ML) methods to meteorological satellite imagery and other large atmospheric datasets. These efforts include image de-noising [9], sea ice discrimination, severe storm prediction [10], cloud classification [11], aircraft turbulence detection [12] and tropical cyclone characterization [13]. All of this work is constrained by limited GPU availability and memory size. Thus the science drivers at SSEC for a next-level GPU cluster would be 1) satellite image research that employs full multi-channel imagery in an adequate batch size to train properly-sized deep learning algorithms (>20 channels at >128x128 image sizes), 2) image-to-image algorithms (inputting satellite imagery and outputting two-dimensional prediction fields) that operate smoothly at the size of contemporary satellite imagery (5500x5500 pixels for geostationary satellites and larger for some polar orbiters), and 3) successful integration of four-dimensional numerical weather modeling datasets with multichannel satellite imagery into predictive models. A larger supply of accelerators will enable these models to explore a much larger design space to find new and previously indiscernible patterns in earth imagery and will lead to more efficient inferencing.

Cooperative Institute for Meteorological Satellite Studies (CIMSS) research scientist **Leigh Orf** and his research team study tornadoes and supercell thunderstorms via tornado-resolving cloud modeling, utilizing resources such as the NSF-sponsored Frontera supercomputer [14]. Orf's team has developed a system for saving and visualizing large amounts of simulation data at extremely high spatial and temporal resolution and uses GPUs for interactive and scripted visualization of model data with tools such as Paraview, VisIt, and VAPOR. Orf's team also develops NVIDIA CUDA C++ research code that leverages GPUs to follow the motion of the air and calculate properties through already-conducted simulations.

Data scientist **Iain McConnell** (AFI DSI and SSEC) uses ML emulation of Numerical Weather Prediction (NWP) system components to enhance computational performance [15], which must be fast to be relevant to forecasters. The aim is to emulate current infrared (IR) image components via Convolutional Neural Networks (CNNs) to leverage spatial information within terabytes of IR satellite images. Additional work on public attention to weather forecasting will augment predictions with classification, sentiment analysis, and

geolocation of Twitter data via BERT and other language models [16]. Model improvements require GPU compute over training and testing data sets on the order of tens to hundreds of gigabytes of data.

Bioinformatics: Prof. Gitter uses ML to study biological images, biological sequences, and biochemistry [17]. His research group has designed CNNs to analyze microscopy images of T cells, detecting active subpopulations of cells [18], to assess effectiveness of treatments. His group is creating new AI models to predict how a protein's sequence mutations impact functional activity [19]. These models can be used to understand protein evolution and design novel proteins. In biochemistry, his group has used AI to accelerate scientific discovery [20, 21, 22]. This includes methods to predict how chemicals will impact the activity of a disease-related protein target as well as generative models to invent new candidate chemicals. In all these application areas, Gitter routinely trains O(10,000) independent ML models on GPUs. His colleagues in the Department of Biostatistics and Medical Informatics similarly require accelerated AI training for a wide variety of biomedical applications such as first-person vision [23], genetic variant interpretation [24], neuroimaging [25], biomarker prediction [26], and learning from electronic health records [27].

Chemical and Biological Engineering: Prof. Zavala uses 3D CNNs for characterizing the spatio-temporal response of liquid crystals [28, 29]. This allows his group to understand underlying phenomena that govern these responses and enables the design of new types of chemical sensors that can detect air contaminants at low concentrations. The training process of 3D CNNs is a highly computationally expensive task that involves handling of high-dimensional data objects and that requires the solution of optimization problems with hundreds of thousands to millions of parameters.

Computer Systems: PI Livny leads the CHTC and is responsible for the development, implementation, operation, and evaluation of distributed HTC (dHTC) technologies that drive the HTCSS. These technologies are used to manage the execution of O(100,000) ensembles of jobs on large collections of heterogeneous, distributed resources. The CHTC and the OSP serve as the primary platforms for the experimental work. In recent years, HTCSS has evolved to address the growing demand for accelerator capacity. These GPU capabilities were widely adopted to provide the following foundational capabilities:

1. Auto-detection of GPU devices, GPU drivers and CUDA libraries installed on a server and advertising their quantity and characteristics.
2. Schedules and provisions GPU devices. Jobs declare the number of GPU devices required, and a job is matched to a resource that currently has the requested number of GPU devices available.
3. Instantiate the job execution environment. HTCSS sets standardized environment variables to communicate to GPU software libraries (CUDA, HIP, OpenCL) which device the job should use.
4. While a GPU job runs, HTCSS continuously monitors GPU resource utilization (both GPU processor utilization and GPU memory utilization), and publishes this information, enabling policies to improve capacity planning based on job history information.
5. Administrators can tell HTCSS to take a subset of GPU devices in a server "offline".
6. Administrators can configure HTCSS to run multiple jobs of a specific designation on one GPU device, i.e., timeshare the device.

HAL, as part of the CHTC environment, will provide a laboratory with users and applications to both evaluate existing and develop new capabilities as benchmarks and synthetic workloads cannot capture the diversity of real users and workloads. HAL will provide an ideal laboratory for the several expected development:

1. Allow users to request the GPU memory required by the job to enable HTCSS to make decisions about packing multiple jobs onto a single GPU if there's sufficient GPU memory.
2. Improve the usage of NVIDIA Multi-Instance technology to partition a single GPU device into multiple smaller devices with good isolation.
3. Offload some aspects of HTCSS GPU management to new generations of container runtimes.

Workloads which include jobs that use accelerators introduce new scheduling and resource allocation profiles. The heterogeneity of the HAL resources and users' workflows will require new resource management mechanisms and new methods to evaluate allocation policies. Working closely with HAL management and users will provide a powerful laboratory for experimental studies of innovative approaches and methods.

Prof. Sinclair's group in the Computer Sciences department is working on novel parallel GPU algorithms and microarchitecture. In recent years, GPUs have evolved into more fully fledged co-processors that support a wide variety of applications. For example, recent work has shown how to use GPUs for persistent kernels [30], buddy allocation [31], and particle partitioning [32], while other work fuses GPU kernels or

uses concurrent streams. These GPU applications often utilize fine-grained synchronization across many threads. Hence, they make use of synchronization primitives, such as barriers, mutexes, and semaphores [33, 34]. Unfortunately, fine-grained synchronization is inefficient on modern GPUs. Unlike multi-core CPUs, GPUs have limited OS support for synchronization and use simple, software-driven coherence protocols that make synchronization expensive. Additionally, synchronization overheads are exacerbated by the level of parallelism on GPUs: GPUs run kernels with up to billions of threads, leading to significant contention for synchronization variables. Consequently, fine-grained synchronization can be prohibitively expensive, and is often a bottleneck for workloads that utilize it.

To address this inefficiency, Sinclair proposes to rethink how synchronization primitives such as barriers and semaphores are designed for GPUs. By co-designing synchronization primitives with the GPU's unique cache coherence protocol, memory consistency model, and memory hierarchy, most threads can use cheap, local synchronization, thereby reducing overhead and improving both scalability and performance. Compared to the vendor provided synchronization methods based on CUDA Cooperative Groups (CCG) [35], preliminary results on older GPU architectures show up to 41% improvement. Modern GPUs, such as NVIDIA's Ampere architecture, which are moving towards chiplet-based designs require further refinement and enhancement of these preliminary ideas. HAL will provide Sinclair's group with the platform to do so.

CMS: **Prof. Bose** and **Prof. Dasu** from the UW-Madison CMS group have been leaders in the design, operation, and upgrades to the CMS trigger system. As the team works on the upgrade toward the High Luminosity-Large Hadron Collider (HL-LHC) era, which will result in a 10x increase in data rates and potentially 100x increase in computing needs, the cost of the computing hardware is a major limitation for the CMS trigger system. If radical reductions in the computation cost are realized, CMS is far more likely to achieve its physics goals for HL-LHC; one leading approach is to translate algorithms to accelerators. While progress has been made with GPUs, FPGAs are more suitable for the CMS workloads as they provide both low-latency and high throughput data processing. High Level Synthesis is enabling production of firmware for FPGAs using C++-code that CMS physicists use for data processing. Studies done by their CMS collaborators including FPGAs-as-a-Service Toolkit (FaaST) [36] and High Level Synthesis for Machine Learning (HLS4ML) [37] could point a way forward. The proposed instrument will allow the group to bring the FPGA studies to fruition quicker: having a large-scale instrument instead of a single machine in the lab allows one to quickly iterate through several different algorithm designs. Further, a cluster provides the ability to test the middleware for scheduling the processing jobs as efficient use of FPGA accelerators also requires their co-scheduling alongside CPUs . UW-Madison groups are well-versed in providing computing facilities and middleware for the international collaboration of CMS for over 15 years. This team will build on their track record: HAL's integration with the OSP means the FPGAs become available to the entire CMS ecosystem, including national and international collaborators and efforts like FaaST.

Cosmology: New instruments are being designed to survey the large-scale structure of the Universe using the redshifted spectral lines of common atoms and molecules. This approach, called line intensity mapping, creates 3D tomographic maps which can be used to study fundamental questions in cosmology, such as the nature of dark matter, dark energy, and the Big Bang itself. Dedicated radio interferometers use the 21 cm line of neutral hydrogen gas to study the accelerated expansion of the Universe caused by dark energy. At millimeter wavelengths, spectral lines from CO and CII are used to study galaxy expansion and formation. At these relatively long wavelengths, the optical design of antennas and telescopes requires a full diffraction analysis, for which commercial electromagnetic simulation codes exist. For hydrogen surveys, **Prof. Timbie** and his group use CST [38] finite element analysis software to simulate antenna patterns for the existing Tianlai radio array in China, and the proposed PUMA array [39], which will include thousands of dish antennas. Prof. Timbie also uses CST for simulating the design of mm and sub-mm wave telescopes designed for CO and CII intensity mapping, such as NASA's EXCLAIM mission. A single GPU can speed-up the simulation up to a factor of 7. Nevertheless, these simulations can take days to complete.

Cryo-EM: **Prof. Wright** directs the Cryo-Electron Microscopy Research Center (CEMRC) at UW-Madison. **Dr. Larson**, a senior member of the CEMRC team, administers the computation resources including automated pre-processing of incoming data, temporary data storage, data processing with CEMRC and CHTC resources, and transfers to end-users. Cryo-electron microscopy (cryo-EM) can achieve atomic resolution structures of biomolecular complexes. The development of cryo-EM methods was awarded a Nobel Prize in Chemistry in 2017 and studies in 2020 have achieved resolutions up to 1.2 Å of specimens in the vitrified state without the requirement of crystallization [40]. The resources of the CEMRC include a

Titan Krios 300-kV cryo-transmission electron microscope (cryo-TEM), a Talos Arctica 200-kV cryo-TEM, 120-kV Talos L120C cryo-TEM instruments, and an Aquilos 2 cryo-FIB Dualbeam instrument.

Due to significant improvements in equipment and computation, cryo-EM single-particle analysis (SPA) has rapidly evolved as an alternative to X-ray crystallography and NMR. Biomolecular machines are highly dynamic and cryo-EM techniques reconstruct the multiple discrete states observed to reveal the 3D conformations of processes in motion. GPU computing is a requirement for corrections between sub-frames of movie stacks, for 2D class averaging, and for 3D structural modeling.

Cryo-Electron Tomography (cryo-ET) techniques give 3D visualizations of structures at sub-nanometer and nanometer-level resolutions while preserving their biological context within cells and tissues. CEMRC is increasing in scope to include the Midwest Center for Cryo-ET (MCCET). MCCET will coordinate the NIH National Network of Cryo-ET Centers with Stanford, the University of Colorado Boulder, and the New York Structural Biology Center. This network will collectively increase researcher access to cryo-ET instrumentation, training, and computation. The frontier technique of sub-tomogram averaging derives sub-nanometer resolutions by extracting particle images *in situ* to solve component structures at high-resolution. This is an extremely demanding computational process that requires GPU acceleration with 12-24GB memory per GPU for the large numbers of sub-tomogram volumes. Each TEM instrument produces up to 8 Terabytes of uncompressed data per day that often requires more than 30 GPU-days for processing. The MCCET facility expansion completing in Q1 of 2022 now adds a next-generation Titan Krios G4 300-kV cryo-TEM, a cryo-FIB, and light microscopy instruments for the National Network of Cryo-ET Centers. HAL will provide an innovative resource for convergent research involving cryo-EM, data analysis, and all-atom molecular dynamics simulations to understand dynamics of the resolved biological systems.

Data Science: AFI DSI [41], led by **Co-PI Cranmer**, collaborates with researchers across UW-Madison to address a variety of computing needs at scale. Individual researchers and their labs work with a data science facilitator to identify options for HTC, which will in the future include HAL. Multi-disciplinary research teams work closely with AFI DSI to plan and execute large-scale projects that require very large datasets and heterogeneous computing. HAL will enable these teams to, for instance, dramatically scale up simulation experiments designed to train and test models using a variety of vision and natural language algorithms that are currently prohibitive due to costs of cloud computing. AFI DSI will bring additional data science and subject-area experts to multiple research teams across campus to leverage HAL resources. As these resources come online, data science facilitators will develop new workshops to meet demand.

AFI DSI coordinates research relationships with American Family Insurance (AFI) in a model that will soon expand to other industry partnerships. Already funded research relies on computing for location privacy, weather forecasting, and natural language processing. More ambitious projects rely heavily on heterogeneous computing, and DSI works closely with AFI to meet those needs efficiently with a mix of on-prem and cloud resources, both at AFI and UW-Madison. HAL would move this conversation to a new level, enhancing the value of AFI's investment in UW research activities. It would also attract new industry research partnerships to take advantage of this resource. See Future Funding Opportunities below.

Co-PI Cranmer is also involved with the ATLAS experiment at the LHC. He is a leader in using AI, Deep Learning, and GPU-accelerated statistical fitting in ATLAS and the particle physics community in general.

Digital Agriculture: Prof. Dorea (Animal and Dairy Science) manages the Livestock Computer Vision Lab located at the UW-Madison research farm. He uses computer vision for real-time monitoring of animal identification, behavior, and growth in livestock systems [42]. His research group uses CNNs to analyze multimodal datasets, including images (RGB, Depth, Hyperspectral) and wearable sensors (Wireless Sensor Nodes). His research projects are focused on high-throughput phenotyping to improve farm management decisions and animal breeding programs. Training CNNs with high dimensional data requires extensive computational resources for timely and efficient processes. The real-time animal monitoring system used by Dorea generates millions of images, which have a high degree of similarity and, thus, may not be useful for training purposes. In this context, data optimization of large image datasets is performed to define the best training set and to eliminate image redundancy in the data storage.

Prof. Zhang (Biological Systems Engineering) develops ML models for high-throughput plant phenotyping. Her research group uses deep neural networks (DNNs) to process time-series drone-based hyperspectral images to estimate multiple plant traits to support plant breeding programs [43]. Recursive neural network (RNNs) models have been extensively used in her current research and they are trained on high-memory NVIDIA GPU cards. Various other neural network architectures are being evaluated such as CNNs,

Bayesian neural networks, and attention-based models to further increase the phenotype prediction performance. For multi-environment phenotyping purposes, transfer learning models are being developed to increase the model generalizability across spatial and temporal domains.

Machine Learning: Prof. Li's group in Computer Science develops statistical and algorithmic solutions for building reliable open-world ML, which aims to recognize unknowns and generalize to the new classes so that the model will become more knowledgeable over time. Her current works explore, understand, and mitigate the many challenges where failure modes can occur in deploying ML models in the open world [44]. From a computing perspective, Li's lab primarily works with deep neural networks. The standard models in development are trained on high-memory NVIDIA GPU cards, using a large collection of images. Common image datasets used for prototyping include ImageNet, MS-COCO, CIFAR-10, CIFAR-100, NIH Chest X-ray, SVHN. Various neural network architectures such as ResNet and DenseNet are evaluated, with number of FLOPs (multiplication and addition) up to 11.3 Billion per feedforward pass (e.g., with ResNet 152-layer network). Parallel training with different parameter and objective configurations is commonly used during the model development process. In some cases, in working with extremely large model capacity and dataset, distributed training across multiple GPUs is leveraged.

Prof. Papailiopoulos' (Electrical and Computer Engineering) work lies in the intersection of ML, coding theory, and distributed systems. Papailiopoulos is interested in the theory and practice of large-scale ML and the challenges that arise when building solutions that come with robustness and scalability guarantees [45]. His research seeks to develop novel algorithmic principles for scalable ML that mitigates communication bottlenecks, specifically in the context of distributed optimization. Furthermore, he seeks to understand and develop new principles for robust decentralized learning in the presence of adversarial agents that aim to corrupt the training pipelines of a ML system. Both these directions aim towards ML that scales and works out of the box.

Molecular Dynamics: Molecular dynamics (MD) simulations have a broad application across chemistry, biology, medicine, and veterinary sciences, including understanding the processes that govern the behavior of viruses like COVID-19 on a molecular level.

Scientific Domain	Faculty	Scientists	Postdoc Researcher	Graduate Students	Undergraduate Students
MPS	20	175	28	150	80
BIO	26	30	10	35	39
CISE	7	10	3	12	32
ENG	6	3	6	42	53
Total	59	218	47	239	209

Table 2. Expected number of researchers & students to benefit initially.

Prof. Huang (Chemistry) focuses on MD simulation of proteins [46, 47]. Protein function heavily relies on conformational changes, i.e., dynamic transitions between conformational states. In recent years, GPU-accelerated MD simulations have emerged to be a powerful approach to study protein dynamics. GPUs can achieve significant speedups over CPUs for MD simulations, and GPU acceleration has been implemented in all mainstream MD software. The Huang group performs large-scale MD simulations on GPUs, and further combine them in the framework of Markov State Model to study bacterial gene transcription. They aim to understand conformational changes of bacterial RNA polymerase, and further help design potential antibiotics inhibiting bacterial gene transcription.

Prof. Kawaoka (Pathobiological Sciences) uses GPUs to simulate the structure of influenza virus proteins that are critical for understanding how our immune systems recognize and defend against this constantly evolving human pathogen. Their aim is to develop vaccines based on viruses that will circulate in the future, thereby conferring better and longer-term protection than current vaccines. This work requires conducting large screens of mutant viruses which it is attractive to conduct *in silico*. Such screens require several repeats of 100s of mutants: a project structure well suited to the high throughput model.

Multi-Messenger Astrophysics: Multi-Messenger Astrophysics (MMA) concerns the observation of astrophysical events using a combination of "messengers", e.g. neutrinos, photons, and gravitational waves, from multiple observatories and telescopes [48, 49, 50]. MMA opens a window to an uncharted portion of the universe. Joint and individual observations by facilities allow us to shed light on yet unobserved phenomena in the universe, determine the relationship between different messengers, and probe fundamental theories of nature. UW-Madison is performing a cluster hire for three faculty, one each

in the Department of Physics, Astronomy, and Statistics, to strengthen the MMA program and develop new data analysis techniques in this rapidly evolving field.

Co-PI Riedel represents WIPAC that includes scientists from IceCube Neutrino Observatory [51], High-Altitude Water Cherenkov Observatory (HAWC) [52], and Southern Wide-Field Gamma-Ray Observatory (SWGO) [53]. These projects are using or exploring AI to augment their data processing and filtering pipelines and to improve their directional and energy reconstructions as well as event classification [54]. We expect the AI usage to grow over the next 5 years by at least a factor of three. IceCube has extensively used distributed and heterogeneous computing resources for the past decade. IceCube's workflow requires extensive GPU resources to propagate Cherenkov photons through the South Polar ice cap [55]. Additionally, IceCube is exploring FPGAs to accelerate the waveform deconvolution algorithm and directional reconstruction run at the South Pole. With future detectors, such as IceCube-Gen2 and SWGO, respectively, we expect a scaling according to the size of the detector. For example, with IceCube-Gen2 [56] we expect an eight-fold increase in GPU needs compared to today.

Co-PI Riedel is part of the Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery NSF HDR Institute. HAL's resources will be made available to researchers in this institute and for research and potential teaching purposes.

Simulated Environments: Prof. Negrut (Mechanical Engineering) runs computational dynamics simulations to characterize the flow and shearing of granular materials (soft matter physics) [57]. These simulations take 5-7 days on a single GPU. Negrut recently started a NASA project in which his group will simulate billion degree of freedom tests of NASA's Viper rover which will search for frozen water at the lunar south pole in 2023. The simulations will be used to develop wheel slip controllers to prevent the rover from becoming immobilized in lunar regolith. The project calls for tens of thousands of simulations, carried out on terrain models of various fidelity (high resolution, continuum, semi-empirical, rigid) to obtain statistical insights into trafficability. Negrut's group is also involved in the simulation of autonomous vehicles interacting with each other in various traffic conditions. These physics-based simulations can be used to understand the interplay of hundreds of vehicles in urban traffic and off-road conditions [58].

University of Wisconsin–La Crosse: The University of Wisconsin–La Crosse (UWL) maintains several long-term external research and instructional partnerships that provide experiential learning opportunities for our students. These partnerships include a Cooperative Education Agreement with the USGS Upper Midwest Environmental Sciences Center and a Collaborative Seed Grant Program between Mayo Clinic Health System and UWL. These community partnerships include faculty/student research projects that would greatly benefit from accelerator computing. Examples include developing an AI-based computer diagnosis system for breast ultrasound lesion assessment and training deep neural networks for computer vision-related geospatial research on the upper Mississippi River. Other academic units interested in participating in the partnership include UWL's Computer Science, Physics, Biology, and Geography/Earth Science Departments.

Intellectual Merit Metrics:

Metric	Target	Science
Performance improvement over libcu++ and CUDA Cooperative Groups for synchronization barriers	>= 30%	CISE
Performance improvement over libcu++ for read-write semaphores	>= 50%	CISE
Increase model training throughput for architecture and hyperparameter screening	10X	BIO
Increase model number and size and dataset size	10X	BIO, PHYS, AST, GEO
Elucidate antibiotic inhibition mechanisms	3X	CHEM
Number of events simulated/reconstructed with GPU-acceleration	1000	PHYS
Size of radio antennas simulated	1000x	AST/PHYS
Neutral networks trained and improved	10X	GEO
Increase in influenza molecular dynamic simulations	10x	BIO, CHEM
Total Cryo-EM users of hardware and software	20	BIO, CHEM

Table 3. Intellectual Merit Performance Metrics for HAL.

Future Funding Opportunities: Federal funding agencies are increasingly focusing on data-driven, AI-based, and convergent research methods. These methods require flexible and diverse computational resources to allow researchers to fully utilize the possibilities. HAL will be available to external collaborators of UW-Madison researchers. We plan to attract additional external funding focused specifically on data analysis and AI, such as NSF's AI Institute program, and allow UW-Madison to continue to host and attract new experimental facilities, such as IceCube-Gen2.

While the HAL design comprises a flexible and diverse accelerator deployment, the supporting storage and network infrastructure is based on well-established technologies and leverages existing CI middleware to integrate with existing facilities. This will allow HAL to serve as a "condominium" cluster, in which individual researchers can contribute hardware funded through institutional or external sources without requiring additional investment in these areas unless the extra capacity or performance is required. This will allow us to increase the size of HAL over time as well as maximize the science impact of hardware contributions. This will allow us to recruit and retain faculty and distinguished researchers. Several senior personnel are at different stages of their academic career, such as Prof. Li, Prof. Dorea, and Prof. Huang.

Results from Prior Funding: PI Livny is PI on NSF award #1148698. **Budget**: $24,378,518.00. **Period**: June 2012 through May 2022. **Title**: "THE OPEN SCIENCE GRID The Next Five Years: Distributed High Throughput Computing for the Nation's Scientists, Researchers, Educators, and Students". **Intellectual Merit**: The OSG is a national, distributed computing partnership for data-intensive research. It provides services and a framework for sharing heterogeneous computational resources. **Broader Impacts**: The OSG underlies the computing strategy of several major NSF investments. The OSG usage is approximately split between LHC, other High Energy Physics experiments, and other sciences. **Publications**: For an overview see the 2019 OSG annual report. **Research Products**: See [59] for links to publication, software, and packaging produced by this proposal.

Co-PI Cranmer was the PI for OAC-1450310 (5/1/2015-4/30/2021) **Title**: "Collaborative Research: SI2-SSI: Data-Intensive Analysis for High Energy Physics (DIANA/HEP)" **Budget**: $939,189. **Intellectual Merit**: DIANA/HEP supported the development of various ML techniques that were abstracted from the physics setting. **Broader Impacts**: DIANA/HEP Fellows provided unique training experiences in sustainable software development for undergraduates and graduate students. **Publications**: [60, 61, 62, 63, 64, 65, 66, 67, 68, 69] [70, 71] **Research Products**: DIANA/HEP led to the conceptualization of IRIS-HEP [72] and software products such as [73, 74, 75, 76], many of which are now supported through IRIS-HEP

Co-PI Riedel is a member of the IceCube collaboration and is supported by the IceCube Maintenance and Operations grant (NSF OPP-2042807) and is a PI for a CSSI Elements award (NSF OAC-2103963) **Budget**: IceCube: $38,380,300 , CSSI: $596,051 **Title**: "Management and Operations of the IceCube Neutrino Observatory 2021-2026", "CSSI Elements: EWMS - Event Workflow Management Service" **Intellectual Merit**: IceCube is a pillar of the US MMA program and will deploy additional modules in 2023-2024. EWMS will accelerate workflows using event-driven design and message passing queues. **Broader Impacts**: WIPAC trains numerous graduate and undergraduate students per year in data analysis and software development. WIPAC also has a large education and outreach component that interacts with audiences across all age groups. **Publications and Products**: See [77]

Co-PI Shechter has not received NSF funding.

Co-PI Gitter is PI of the current award NSF DBI #1553206 starting July 2016 and ending June 2022. **Budget**: $887,230 **Title**: "CAREER: Inference in temporal signaling and transcriptional data" **Intellectual Merit**: PI Gitter's lab designed new computational methods to infer biological regulatory networks from single-cell RNA-seq and time-series phosphorylation data, including tools to make more biologically plausible predictions. **Broader Impacts**: This award supported developing the Machine Learning for Biology workshop to teach ML literacy to biological researchers and Protein Pinball, a game about protein signaling for middle school field trips. **Publications**: [78, 79, 80, 81, 82, 83] **Research Products**: [84] for software products, [85, 86] for workshop materials, and [87] for proteomic data.

3. Broader Impacts

The broad range of science drivers in this proposal precipitates an impact across society, including on the public education and outreach mission of UW-Madison.

STEM Education and Workforce Development: There are several courses that would benefit from HAL either already being taught or in development at UW-Madison and outside. We highlight seven of them here, four undergraduate and three graduate education courses:

Prof. Negrut currently teaches two classes—an undergraduate course ME459 “Computing Concepts of Applications in Engineering” and the graduate course ME/ECE 759 “High Performance Computing for Applications in Engineering” taken by approximately 130-150 students each year. Both courses rely on the availability of HPC resources for educational purposes and need 1 GPU-day per student per week. *Prof. Dorea* currently teaches an undergraduate and graduate course called “Introduction to Digital Agriculture”. The undergraduate students have so far been unable to use GPU resources in this course due to cost. For the graduate-level course each of the approximately 30 students were assigned a Microsoft Azure virtual machine with 1 GPU (Tesla K80; 24 GB of memory), for their homework and projects. Dr. Dorea has taught this course internationally (Italy, Brazil, Colombia) to more than 70 students/year as part of institutional collaboration efforts. For that, cloud-based GPUs were not able to be used due to costs. *Prof. Li* is developing a new ML undergraduate and graduate curriculum, including at least two specialized courses for the CS department. A graduate course CS762 (Advanced Deep Learning) recently was approved by the university. Her classes will include giving students practical experience in ML. The success and enrollment expansion of the courses depends heavily on a centralized, high-performing GPU cluster. *Prof. Zhang* is currently teaching a course titled “Intelligence and Automation in Agriculture” for upper-level undergraduate and graduate students. This class will give students hands-on experience in using deep neural networks to solve practical agricultural problems such as weed detection, disease detection, and yield prediction. GPU resources will be needed for course projects and homework.

HAL will be a vital resource for undergraduate and graduate student education outside the classroom including research internships and project-based classes. Students will be able to rapidly prototype ML-based data analysis techniques, explore accelerator-based computing, and learn to use modern computing infrastructure. This will help students develop skills highly in demand in today’s digital economy.

Public Engagement: HAL will be a tool to demonstrate, as part of public outreach, the importance of computing to fundamental research. Examples of outreach activities are the yearly Data Science Research Bazaar hosted by AFI DSI [88] and, The Computing in Engineering Forum [89] hosted by *Prof. Negrut* and *Zavala*. Both create cross-pollination between on- and off-campus stakeholders and partners. We will give individual and summary presentations at these to show the impact of HAL to the wider data science community in Madison as well has highlighting new hardware and software techniques developed in HAL.

WIPAC hosts the operations for the IceCube Neutrino Observatory and has extensive public outreach activities. We will work with the WIPAC outreach staff to include ML-based activities in their high school student internship and Masterclass programs [90]. These programs pair groups of high school students with mentors, typically graduate students, postdoctoral fellows, and faculty, to complete a research project over the course of a semester-long internship, or a day-long Masterclass.

HAL will also be used to improve scientific visualization to inform the broader public about complex research. CIMSS will use HAL to create 5K resolution, visualizations of tornado-producing storms for education and outreach activities. The cluster will be used to interactively visualize and interrogate simulation data modeling tornado-producing thunderstorms, and to make scripted visualizations and animations to be used in videos and educational programs hosted at SSEC.

Societal Wellbeing: HAL’s science drivers in bioinformatics, cryo-EM, and MD have or the potential to make significant societal impact. *Co-PI Gitter* has been involved in drug discovery using ML techniques; *Prof. Kawaoka* has been involved in UW-Madison’s Global Health Institute and has been working on understanding and preventing pandemic and how influenza viruses transmit between species. This research could help understand the current COVID-19 pandemic. Weather research at SSEC will allow us, among other applications: forecast crop yields, and determine the impact of global warming. Digital Agriculture will allow us to improve crop yields, while improving long-term sustainability of farming.

National Security, Government Agencies, and Industry Competitiveness: HAL has a potential impact on national security and other government agencies. Researchers in Simulated Environments have been funded through the Department of Defense to perform simulations related to national security. There are also several other government agencies that fund research for the science drivers: NASA (Negrut), USDA (Dorea and Zhang), USGS (UW-La Crosse), NIH (cryo-EM and Gitter), and NOAA (SSEC). A number of these are relevant to industry, such as remote sensing to estimate crop yields or weather prediction for insurance companies.

Regional and National Impact: We will work with regional and national partners to increase HAL’s impact beyond UW-Madison. At the regional level, we will work directly with IceCube researchers at Marquette

University in Milwaukee, WI, LIGO researchers at University of Wisconsin-Milwaukee, and USGS researchers and faculty at University of Wisconsin-La Crosse to allow them to use HAL as if they were UW-Madison researchers. We will also work with University of Wisconsin System to allow other researchers in Wisconsin public universities to have access to HAL; this capability has already been demonstrated through federation technologies like the OSP. Similarly, we will be working with educators from UW system universities to allow their students to use HAL for their class projects.

HAL will be one of the largest accelerator resource providers to the OSP and will be the first FPGA resource. The increasing adoption of ML-based data analysis and accelerator-based computing projects creates an increasing demand for accelerated resources on the OSP. This trend can already be seen in several research projects. Examples include the CMS group at MIT using public cloud-based GPU and FPGA to perform ML inference for both CMS and Deep Underground Neutrino Experiment [91] and Prof. Peri at University of Colorado, Boulder using FPGAs to accelerate computing dynamic stochastic equilibrium models in macroeconomics.

Empowerment of Underrepresented and Under-Resourced Participants: HAL is committed to ensure diversity in its users across several dimensions and to use the resource to help broaden participation in computing. Access to HAL will be provided to students and researchers at Southern University (Historically Black Colleges and University) and Whittier College (Hispanic-Serving Institution) as if they were UW-Madison researchers. Students and researchers from other underrepresented groups will be able to access HAL through the OSP interface. We will continue to gorge new alliances with other underrepresented groups during the operation of HAL. One such group at UW-Madison is a new radio astronomy data science center that partners with several HBCUs [92]. We will open HAL to this data center and their collaborators.

Broader Impact Metrics

Metric	Target	Science
Number of undergraduate students	750	BIO, MPS, CISE
Number of graduate students	200	BIO, MPS, CISE
Public engagements	20	ENV, GEO
Number of students from under-represented groups	50	BIO, MPS, CISE, GEO, CHEM
Users from outside UW-Madison	50	BIO, MPS, GEO
Number of researchers outside NSF	30	BIO, MPS, CISE, GEO, CHEM

Table 4. Engagement and Broader Impact metrics for HAL.

4. Description of the Research Instrument and Needs

System Motivation and Sizing: Our design goal is to provide a large, diverse set of accelerators for R&D by scientists and students. We focus on three different types of GPUs and FPGAs in HAL to meet a diversity of existing needs and to enable experimentation with a variety of features and software environments (such as TensorFloat-32 in NVIDIA A100s and NVIDIA's CUDA vs. OpenCL, respectively) to determine which GPU types, GPU vendors, or FPGAs are best suited to particular use cases. HAL will provide a platform for researchers to adapt their algorithms and code to take advantage of the increasingly heterogeneous compute hardware available in the national CI environment. The size of the user community ultimately drives the overall instrument size. We expect at least 100 local researchers, 300-600 students per semester, and at least two large scientific collaborations to make use of the system at any time. A larger and diverse pool of GPUs, rather than a smaller number of faster or feature-rich GPUs, will maximize the availability of resources to users and maximize the system applicability.

System Architecture: HAL is a fully integrated system of instruments: AI- and single-precision GPUs, double-precision GPUs, FPGAs, and storage. HAL will be integrated into the data processing systems of existing instruments, an educational resource and a research platform for AI-research and AI-based data analysis, joining the regional and national CI.

The baseline hardware of HAL is summarized in Table 5 and in the budget documents. HAL will consist of a total of 288 GPUs and 8 FPGAs. The NVIDIA A40 [93] and NVIDIA RTX A6000 [94] make up the majority to maximize the number of GPUs available and cost-effectiveness. Two DGX A100s [95] with 8 NVIDIA A100s each. The NVIDIA A100s will provide significant AI-acceleration, double precision floating point support, and multi-GPU linking provided through the built-in NVSwitch, enabling researchers to explore cutting-edge advancements, including TensorFloat-32, multi-GPU usage, and GPU splitting, to optimize

their workflow. Three additional servers will host 8 Xilinx Alveo U55C FPGAs [96], 8 consumer-grade NVIDIA RTX 3090 [97], and 8 enterprise-class AMD MI100 GPUs [98], respectively. FPGAs have great flexibility and can accelerate AI inference significantly, such as the CMS trigger and reconstruction outlined above, and accelerating applicable compute compared to GPUs, such as IceCube waveform deconvolution.

A 5 PB shared Ceph-based [99] filesystem will support user home directories, local data cache, and common software packages and containers, such that software and data can be shared across machines and be readily available. This will enable researchers to quickly stage training AI data and other data-intensive tasks. Educators can host sample datasets, pre-train ML models, etc. needed by students. We will locally curate a set of AI test and training datasets, such as the MNIST handwritten digit dataset [100], and pre-trained neural networks, such as the ResNet collection [101, 102].

HAL's networking will consist of two separate systems. All nodes will be connected through 100 Gbps Ethernet. The DGX A100s will be directly connected to each other at 200 Gbps Infiniband to allow for sharing of GPUs across the DGX A100s. Eight GPU servers will be located with the existing CHTC HPC infrastructure to take advantage existing Infiniband infrastructure. These servers will still be connected at 100 Gbps to the campus backbone and have direct connection to the remaining HAL infrastructure.

We will collaborate with DoIT to integrate HAL with UW-Madison's ResearchDrive [103] to enable HAL researchers to easily transfer data between long-term storage resources and HAL. ResearchDrive is a university-wide, centrally funded file storage solution for UW-Madison researchers providing secure, permanent, shareable storage space on the UW–Madison campus network. HAL will be closely tied to the Open Science Data Federation [104] through existing local data caching infrastructure. Allowing external users, such as the LIGO and CMS collaborations, to access data efficiently. for data analysis.

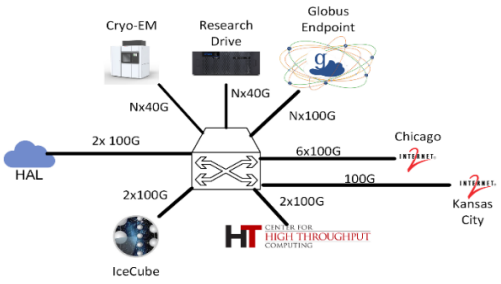

Figure 1 Network Diagram for HAL

HAL's network fits within both the local and wide area network configuration for UW-Madison, see Figure 1. HAL's internal and external network connectivity will be managed by DoIT, ensuring HAL will be part of future planning for campus networking. External data transfers will use UW's well-established connection to the ScienceDMZ [105]; this network capability was demonstrated as an integral part to recent experiments performed by IceCube and collaborators at San Diego Supercomputer Center to test workflows running in the cloud that fetched data through elastically established direct links between UW-Madison, Internet2, and three major public cloud providers [106]. This work demonstrated the team's ability to make an impact at a national scale.

HAL will have three user access interfaces: SSH with HTCondor (primarily researchers), a "notebook" interface through JupyterLab (primarily students and educators), and external interface through OSP. We expect this will evolve over time as HAL becomes a direct instrument for data exploration, processing, and filtering. The OSP interface will tie HAL into the national CI. All interfaces support container-based environments increasing accessibility.

HAL will be a steppingstone to offload time-critical computing from local to global resources delivered by commercial cloud providers and the national CI. HTCSS can expand its resource pool elastically, such as NSF's eXtreme Science and Engineering Discovery Environment (XSEDE) [107, 108] and public cloud resources [109]. CHTC facilitators and WIPAC operations team have extensive experience to help researchers migrate workloads from local to global resource pools and expanding resource pools [8, 7], respectively.

Science drives our design of HAL. The configuration of HAL draws from experiences from WIPAC, *Co-PI Gitter*, and *Senior Personnel Negrut* and was refined via surveys with science drivers and vendors (particularly NVIDIA) to determine the best price-to-performance ratio. We aim to maximize the overall utilization of each GPU recognizing that the proposed science drivers have different bottlenecks when it comes to using accelerators. ML training and inference, for example, may be constrained by limited bandwidth between storage and GPUs. Similarly, simulations may yield better throughput on multiple "slower" GPUs rather than a single "fast" GPU [8]. Thus, enabling users to split GPUs according to needs in hardware (A100) or software (A40) will yield greater scientific output per GPU. The submitted

configuration is subject to change in availability of parts, release of new hardware products from vendors, and overall cost.

There are three comparable systems within the national research CI portfolio: Chameleon [110], Delta [111], and JetStream2 [112]. Chameleon has a similar mission to HAL in providing a breadth of computing resources with a flexible software environment. The major difference is the target science drivers and mission. Chameleon is focused on CISE researchers and acts as a testbed for cloud and Internet of Things researchers. HAL is both a testbed and production resource for data- and compute-intensive researchers. We envision researchers going from using a single accelerator for development and testing to running complex workflows on HAL.

Delta, slated to be deployed first half of 2022 at the National Center for Supercomputing Applications (NCSA), will be one of the largest GPU resources in the XSEDE portfolio. Delta focuses on serving the national CI rather than a single science driver, campus, or university system. The resources will be allocated through the XSEDE Resource Allocations Committee (XRAC). Research and development of new algorithms or methods, such by Prof. Li and Prof. Dorea does not fit well within the mission of Delta. JetStream2 is focused on composable environments and education. JetStream2's available resources, software, and scheduling environment are like Chameleon's. Compute allocations and access are handled through XRAC and is the closest resource in the national CI to HAL.

From the perspective of a new user, the three resources listed above share disadvantages compared to HAL. Nearly all national CI providers are oversubscribed, even before they are available. This makes it difficult for these resources to allocate a large fraction of their resources to a single class, group, or campus. Similarly, these resources are focused on well-established research programs. Exploratory or preliminary research is hard to perform outside of a discretionary or testing allocation. HAL is meant to enable users to go from prototyping a research idea to production workload that can effectively utilize an XSEDE allocation.

Commercial cloud resources can partially fill the gap that HAL is meant to fill in both research and education. Yet, cloud resources can be prohibitively expensive compared to on-premises resources. IceCube has explored using the cloud for its Monte Carlo simulation needs and determined that **on-campus resources are 1/20th the cost of comparable cloud resources**. Dr. Dorea's Lab has an annual cloud computing cost of approximately $24,000 to keep a computer vision system, located at the UW-Madison Research Farm, performing real-time inference, and training of ML algorithms. The graduate Digital Agriculture course taught by Prof. Dorea used commercial cloud providers at a cost of $400 per student per semester. This cost precluded Prof. Dorea's undergraduate Digital Agriculture course from using the same resources. The public cloud's plethora of services and economic model can lead to surprising costs. Researchers normally do not consider the cost of networking, different tiers of storage, or whether a certain service is cost-effective for their needs. While for small scale tests, such as testing the newest GPUs, the cloud is feasible, the larger the needs or more complex the workflow, the harder public clouds are to justify.

HAL System Architecture			
Hardware Type	Nodes	Devices	Total
DGX A100	2	8x	16
NVIDIA A40	16	8x	128
NVIDIA RTX A6000	16	8x	128
NVIDIA RTX 3090 (24GB GPU RAM)	1	8x	8
AMD MI100 (32GB GPU RAM)	1	8x	8
Xilinx Alveo U55C	1	8x	8
Login	1	-	-
Service	2	-	-
Storage	16	-	-
Total	56		288 GPUs 8 FPGAs

Table 5. The proposed accelerator system architecture. DGX host will have dual AMD EPYC 7742 CPUs (128 cores total), 1 TB RAM, and 200Gbps Ethernet. All other hosts will have dual AMD EPYC 7413 CPUs (48 cores total), 1 TB RAM, and 100Gbps Ethernet. Included is 6.9 PB raw capacity Ceph filesystem and 100Gbps connectivity between all hosts and the campus backbone. The OS will be Oracle Linux 8 and significant system software will include HTCondor and will primarily utilize containers for user software.

The architecture of HAL allows it to expand to additional science domains, computing and co-processor architectures, or new user interfaces. HAL is the foundation on which UW-Madison will build new resources for researchers and educators. UW-Madison, and in particular PI Livny, have followed this model in the past with an initial external investment

(GLOW) that expanded and continues to operate well after the award period. We expect new or additional partners to contribute resources to HAL through a “condominium” style computing model. This will update and broaden the CI available through HAL, ultimately making HAL sustainable beyond the 3-year timeline of this proposal. At the same time, new researchers will be able to harness a larger shared resource pool than they could provide with their funding alone.

The exact performance of each application on the chosen hardware is hard to quantify. Each generation and version of GPU brings with its myriad optimizations in both hardware and software in addition to increases in overall performance. WIPAC and IceCube have done extensive work benchmarking and executing their code on a wide array of GPU models, including the NVIDIA A100s. From this data we expect that the chosen GPUs will perform in line with the vendor provided specifications of FLOPs for Monte Carlo simulations [7, 8]. For AI applications, there are several other factors that can impact GPU or FPGA performance, such as connection speed between storage and the GPU, AI training settings, etc.

Operations: The hosting, operations and facilitations of HAL will be split across UW-Madison DoIT, MIR, and three UW-Madison research centers: WIPAC, CHTC, and AFI DSI. DoIT will host the facility in the Computer Sciences data center and provide network services. WIPAC will operate and maintain the physical and software components of HAL. CHTC will provide consultation and facilitation services to researchers using the resource, leveraging their leadership and expertise in Research Computing Facilitation [113] to tackle the unique scheduling problems posed by the heterogeneous science drivers and hardware. Bockelman, from MIR, has committed effort to provide initial architectural guidance for the deployment of the storage and data transfer systems, operational effort to help on-board larger-scale, complex projects, and ongoing integration with the Partnership for Advanced Throughput Computing (PaTH). AFI DSI will educate researchers in ML and help coordinate curation of existing and emerging software. HAL will be available as a teaching resource across UW-Madison and the UW System, making UW-Madison one of the few campuses in the US that provides students with hands-on access broad range of AI resources and capabilities

Operational Metrics:

Metric	Target
System availability (once production status is achieved)	90%
Average Accelerator utilization	90%
Ticket responsiveness for HAL (time to first touch)	8 business hours
Active user groups (quarterly)	20
Percent of accelerators available to users in production	100% by month 15

Table 6. Metrics covering the operational aspects of the system.

5. Management Plan

Management: HAL will adopt a shared governance model to ensure user driven development and operations. Overall management and responsibility will lie with PI Livny while day-to-day operations of HAL will be coordinated by an executive committee that meets biweekly, consisting of: the HAL PI (ex-officio), the UW CTO (ex-officio), and representatives from WIPAC, CHTC, AFI DSI, and MIR.

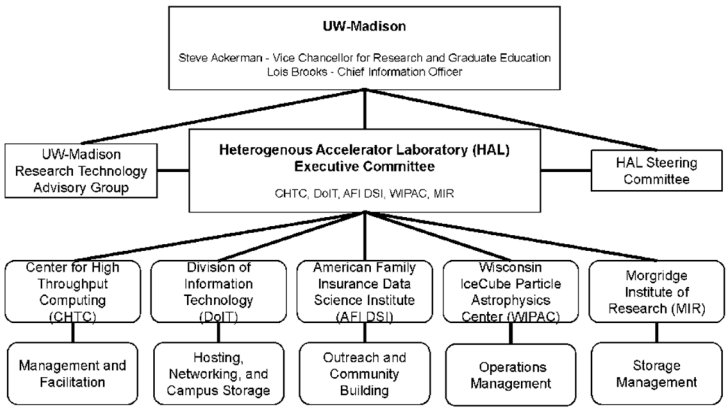

Figure 2 Overview of the management Structure for HAL.

The day-to-day responsibility is split according to expertise; DoIT hosts, WIPAC operates, CHTC does user facilitation, AFI DSI handles outreach and community building, and MIR will manage the storage and interface development. Oversight of the grant as well as the facility will be handled through the Office of the Vice Chancellor for Research and Graduate Education. See Figure 2.

Governance: The project will be governed by a steering committee which provides strategic input to the executive team. The steering committee will meet at biannually and will consist of 8 members: HAL PI (ex-officio), HAL Co-PI (ex-officio, rotating every semester), UW-Madison CIO (ex-officio), Director of AFI DSI (ex-officio), Vice Chancellor for Research and Graduate Education (ex-officio), Two HAL Research Representatives, and HAL Teaching Representatives. To balance the needs of day-to-day operations with the needs of education that provides students with hands-on access to a broad range of AI resources and capabilities, science impact, and broader impact. Research representatives will be group PIs, typically faculty, (or designated representative) chosen at random from the HAL research user pool. We randomized the selection to ensure representation across all fields over time. The HAL Teaching representative will also be chosen at random from the faculty who used HAL for teaching in current or previous semester.

The decision-making process within HAL will follow a bottom-up approach. Users will make suggestions to or request changes from facilitators or operations. The facilitators and operators are free to make changes they assess as having minor impact on HAL operations and user experience (such as adding users or changing individual user priority in case of deadline). Larger matters, such as general policy changes or scheduling policy changes, will be made in consultation with the executive committee or HAL steering committee as needed

Milestones, Metrics, and Risks: The leadership team of HAL are strong believers of using milestones, metrics, and a risk registry as programmatic tools to quantitatively track the project's progress and performance. At the beginning of the project, the management team will develop a project execution plan (PEP) detailing the project's programmatic items. The effort at MIR will be used to track the metrics, milestones, and risks and will report this information to the biannual steering committee meeting.

Project Date	Milestone
Y1Q1	First steering board meeting. Additional milestone for SB meetings every 6 months.
Y1Q2	Purchase order issued for HAL hardware
Y1Q3	First hardware begins to arrive
Y1Q4	All hardware has arrived
Y1Q4	Early science users able to access hardware via HTCSS.
Y2Q1	Full resource available to all users via HTCSS. HAL is fully deployed.
Y2Q2	JupyterHub-based access available in production.
Y2Q3	First class is taught using the JupyterHub-based interface to HAL.

Table 7. Initial set of HAL deployment milestones

For example, the initial set of metrics given in Table 3, Table 4, and Table 6 will be expanded to include information on (a) collection, (b) targets will be updated from end-of-project numbers to yearly or biannual targets, and (c) document expected management actions if the targets are missed; these expansions will turn the metrics from a reporting item to a useful tool for the management and steering committee to understand the project's progress. Similarly, a risk registry will be developed to outline the major risks to the project execution; for each risk, the project will document the time frame where the risk applies), the probability and impact, the monitoring and trigger, and the risk mitigation. For example:

Risk 1: Supply chain issues cause major delays in component availability. **Timeframe**: Year 1. **Probability**: Medium. **Impact**: Medium. **Monitoring & Trigger**: Leadership will report the percent of hardware funds ordered and percent of accelerators available to users in production at the biannual stakeholder's meeting.

This risk will trigger if less than 75% of the equipment is available by month 12. **Mitigation**: If triggered, the leadership will split the deployment to users into two phases: phase one will be a single model of GPU-only and the second, occurring in Year 2, will have the full set of accelerator resources.

Risk 2: Lack of adoption of HAL in classroom setting. **Timeframe**: Ongoing. **Probability**: Low. **Impact**: Medium. **Monitoring & Trigger**: Leadership will track the number of classes whose students have access to HAL; trigger is <3 in Year 2 and <4 in Year 3. **Mitigation**: If triggered, the leadership will reassign the facilitator effort to engage more directly with educators to integrate HAL with courses. If necessary, additional priority will be given to the JupyterHub interface and accelerator resources reserved for interactive use instead of the expected fairshare use.

Finally, the PEP will track a detailed list of milestones. An initial set of milestones is given in Table 7.

Resource Allocation and Management: The HAL science drivers will establish an initial general usage policy for HAL. The usage policy will follow established guidelines for UW-Madison's existing campus-wide resources. The major difference for HAL compared to a traditional HPC facility will be the scheduling and share allocation policy. Rather than allocating a fixed number of compute hours or a fraction of HAL to a specific user group, a user's share is set according to their user group, recent usage, and their application. This system has proven effective in sharing CHTC's resources among UW-Madison researchers, such that individual researchers and smaller research groups can readily compete for resources with large international collaborations. Additionally, this allows researchers to easily share their resources when they are not being used, increasing the overall utilization of resources.

The HAL resource allocation policy is based on five different user categories, shown in Table 8. The primary difference between the user categories is how quickly their shares decrease as a function of usage. Initially there will be an even split in shares among all users. Over time a user's shares will change according to their recent usage; the more a given user group uses the system they will lose shares compared to all other users. The rate of decrease in the shares is set by the user category, e.g., a Category I user will lose their shares slower than a Category III user. If a user uses a lot of resources in a short period of time this may drop them into a lower user category in terms of priority.

User Category	User Type	Max Job Length	Can be evicted?	Can evict
I	HAL Science Drivers	Default: 1 day Long: 7 days	No	Category VI-V
II	Educators and Students	Default: 1 day Long: 7 days	No	Category VI-V
III	UW-Madison User	Default: 1 day Long: 7 days	No	Category VI-V
IV	Evictable UW-Madison User	1 day	Yes	Category V
V	OSP	1 day	Yes	-

Table 8. User categories and the respective policy for each.

For reference on the HTCondor priority system see [114] and [115]. The overall shares and rate of decrease in shares is also affected by the type of job a user is running. If the job requires a long runtime (>1 day), it will have lower priority compared to a job that runs for short periods of time. Additionally, jobs that can be evicted/pre-empted, i.e., terminated by another job that will run in its place, will be allowed to run on unused resources and will lower a user's shares at a lower rate than jobs that cannot be evicted. This priority scheme has been well-established with CHTC-hosted resources.

Sustainability: To ensure sustainability we will recruit additional science drivers by allowing them to buy into HAL, host their hardware as part of HAL, or contribute their resources to HAL. This builds on a demonstrated approach for non-accelerator-based resources that was established in 2003 with GLOW. Adding a researcher's resources will be done in steps. An initial consultation between the researcher and the facilitators is followed by a discussion with the steering committee and the operations team. Once the researcher has decided to locate their hardware with HAL, the operations team will propose a configuration and work with researcher and vendors to ensure the researcher's needs are met.

Operations and Hosting: UW-Madison has several facilities suitable for hosting HAL. Depending on the researcher's needs, their hardware may be co-located with most of the HAL hardware or located in different space. If the additional hardware is not co-located with HAL, the project team will ensure that there will be a 100 Gbps connection through the campus networking. We have learned from our experiences with GLOW and IceCube that hardware does not have to be co-located to reach the desired level of performance so long as there is sufficient network capacity to support data movement. We anticipate that there will be workflows for which network latency is an issue; this will be considered when resources are added to the instrument.