Massachusetts Green High Performance Computing Center
100 Bigelow Street, Holyoke, MA

MGHPCC

MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER

Request for Proposal

AI Compute Resource Infrastructure System

March 17, 2025

BOSTON UNIVERSITY · HARVARD UNIVERSITY · Massachusetts Institute of Technology · Northeastern University · UMASS · Yale

Contents

1 Introduction

This AI Compute Resource (AICR) Infrastructure System RFP covers the first of 3 tranches of funding. The tranches target a sequence of systems (ST1, ST2, and ST3) aimed to be in early operation by August 2025, January 2027, August 2028, respectively. The response to this RFP should address ST1. Responders are encouraged to comment on all 3 tranches and address how newer generations of technology might be introduced and integrated with previous generations.

Formally, we expect to execute full agreements only for the procurement of system ST1 in this RFP. A projected investment of $10M to $20M has been planned for ST1. Similar or larger amounts may be available for ST2 and ST3. The total for all three tranches has not yet been fully determined, but will be part of the overall Massachusetts AI Hub program which envisions an investment that exceeds $100M.

The final decision on the nature of the procurement processes for ST2 and ST3 will be informed by the ST1 tranche activity and on external factors, including for example actual technologies that may emerge in the intervening periods. However, future RFPs will ask respondents to explicitly consider how systems ST2 and ST3 will integrate into ST1 infrastructure.

The RFP allows for responses that span exclusively on-premise system solutions to exclusively virtual cloud solutions. Hybrid solutions that span cloud and on-premise will also be considered for any of ST1, ST2 or ST3. There is some expectation by the RFP review team that price competitive ST1 proposals will involve an on-premise element, but the selection process will consider and evaluate all proposals fully and seriously. More than 80% of the funds supporting this RFP are from sources that require their use for capitalizable expenditures under accepted accounting principles. To be competitive in this RFP, respondents should propose a project and any payment terms that are consistent with largely capitalized funding.

The RFP is contains four principal sections: (1) **Background** that provides context for responders; (2) **Instructions** regarding timeline and submission format; (3) **Technical Specifications** that cover quantitative and qualitative areas that responders should address; and **Responses Guidance** that provides additional detail on response format, content anytime line, and evaluation criteria.

Quantitative capability targets are defined for the ST1 system. Responses will be evaluated on their ability to meet some or all of these targets. Responses may also include ideas addressing ST2 and ST3 tranches. This could include outlining how future enhancements might be integrated into these systems to build on quantitative capabilities.

In addition, the RFP has two appendices. The first appendix contains a set of use cases. The use cases are by no means exhaustive, but they are intended to guide responders around the sorts of workloads the ST1 (and beyond) system is anticipated to support. The second appendix contains a hypothetical “reference architecture concept” for the ST1 system. The architecture

appendix has been used to develop quantitative metrics that are in the RFP system requirements sections. There is no requirement that responders propose a system that matches the architecture appendix one to one in detail. The appendix is intended to be reference material, to help responders form thoughtful proposals that address our project requirements well.

2 Background

A. Context

The ST1, ST2 and ST3 systems are the foundational infrastructure for the Massachusetts AI Hub initiative that was announced in December 2024. The announcement can be found at: https://www.mass.gov/news/governor-healey-announces-massachusetts-ai-hub-to-make-state-global-leader-in-applied-ai-innovation.

This initiative is a public-private partnership with sizable investments from the Commonwealth of Massachusetts, as well as Boston University, Harvard University, the Massachusetts Institute of Technology, Northeastern University, the University of Massachusetts system, and Yale University. As such, the application base for this system includes a wide range of applied AI/ML computational research and innovation. This range spans the academic community; economic development in the regional early stage startup community; and strategic areas of the established regional economy – including robotics, national security, financial innovation, education, health/biotech. Emerging areas, including quantum, fusion, next-generation materials and nanotechnologies, and beyond, are also expected to be potential areas of impact.

B. Scope of Work

- Responses to this RFP should address a system and platform architecture that can support a broad range of AI/ML workloads and use cases of the sort found across the constituencies outlined in the preceding **Context** paragraph.

- Appendix 1 of this RFP describes some representative, but not exhaustive, use cases that we expect to target to help guide proposers. Responses to this RFP should define a ST1 system that tries to address this range of scenarios in a cost effective manner. RFP responses should focus on a basic system with core capabilities (outlined in subsequent sections) to meet these needs, including standard systems and user support software for efficient technical and AI/ML workload. The focus of the evaluation criteria for this RFP is basic computational data, and system management capabilities that provide a foundation for higher-level services. Optionally, responses may include ST2 and ST3 projections for capabilities and strategies for integration of ST1, ST2, and ST3 phases. The evaluation criteria will focus on ST1 proposals.

- Appendix 2 outlines a high-level set of reference architecture blocks. This diagram provides an abstract representation of the key subsystems a response should

encompass. Responses may or may not have distinct subsystems for each of these logical blocks. Responses may choose to propose solutions in which several subsystem logical capabilities are addressed by a single actual system. The description of the proposed system should describe which elements of that system are addressing each of the subsystems shown in Appendix 2. Capability targets, that responses must address, are given with respect to the different subsystems defined in Appendix 2.

- A separate AICR Operations and Engineering Request for Information solicits information about capabilities related to operations and maintenance of a basic software stack. Responders to this AICR Infrastructure System RFP may also respond to the Operations and Engineering RFI, but are not required to do so.

- A separate Infrastructure FAQ provides information about site conditions for ST1.

3 Instructions

A. RFP Timeline

The Infrastructure System RFP, the Operations and Maintenance RFI, and associated FAQ pages were posted on the MGHPCC web site on March 17, 2025 at https://www.mghpcc.org/ai-compute-resource-system

Questions concerning the RFP and RFI may be submitted to mghpcc-ioc-inquiries@mghpcc.org. For questions that MGHPCC determines to be of general interest, answers will be posted to an FAQ page.

Prospective responders may also request a one hour meeting with members of the RFP evaluation team to ask questions or present ideas, with focus on ensuring the best possible response to the RFP.

The planned response timeline is as follows:

April 17, 2025 - RFP and (optionally) RFI responses due

May 15, 2025 - RFP responses evaluated and vendor selected*

August 1, 2025 - System available for early user production workloads

*The RFP evaluation team may extend the selection date at its discretion.

B. Proposal Format

Proposal responses should include a technical document describing a system that addresses the computing, memory, storage, and networking requirements outlined in the **Technical Specifications** section. Where relevant, this section should include:

- A detailed bill of materials for all the system building blocks and a high-level summary bill of materials for the ST1 system

- Detailed system diagrams showing logical and physical layouts, as appropriate, for the ST1 system

- Details of power and cooling connectivity and all requirements and assumptions regarding physical facilities infrastructure for the ST1 system

- The expected potential scaling of key capabilities in any ST2 and ST3 system plan, if a responder chooses to include this. We anticipate the ST2 and ST3 systems will include some capacity to handle regulated data that is not present in ST1.

- A separate document should detail total system costs for the ST1 system and a cost breakdown by key subsystems. Appendix 2 of this document defines a set of logical subsystems. The Appendix 2 subsystems may or may not map to specific subsystems that an RFP response includes. However, the expectation is that any RFP response will consist of several identifiable subsystems and the costs of these should be detailed.

- A document describing installation, commissioning and training/handover services included in the overall cost, as well as the cost of those services. These services should include bringing the entire system into a working state and demonstrating its functioning through successful execution of a series of tests. These tests may be the same as the acceptance tests the AICR team will use for authorizing payment. Other tests may also be proposed.

- A document detailing service level agreements and terms that will apply, which includes correcting system problems for a minimum of 5 years from acceptance. This includes warranties/maintenance agreements, and clear information on any third party warranties not supported by the primary vendor (for example CDUs and manifolds).

- A one page summary document showing total price, number and type of Tier 1 and Tier 2 accelerators, high-speed and commodity storage capacity, delivery date, and installation complete estimated date.

- A tabular summary showing all **attributes** in the RFP with the capability the response is proposing.

C. Submission Details

Proposal Responses:

- Proposal responses should be submitted by email as an attached zip file or a zip file download link to mghpcc-ioc-aicr-rfp@mghpcc.org

RFP Questions:

- Questions concerning the RFP may be submitted to mghpcc-ioc-inquiries@mghpcc.org
    - Responses to questions will be posted to a public FAQ page on the https://mghpcc.org web site.

4 Technical Specifications

A. Capability and Capacity Requirements

Many of the technical specifications expressed below are given as aggregate system wide capabilities. This is intended to give responders flexibility to design solutions that achieve as many of the desired aggregate capability metrics as possible. Response evaluation will include how well these metrics have been met, as well as the performance of individual components.

Responses should address how their proposed system supports a hypothetical, logical architecture shown and described in Appendix 2. The logical blocks in Appendix 2 need not be distinct actual subsystems, but responses should map these blocks to their proposed solutions.

B. Overall System Architecture

The proposal must provide a detailed architectural overview of the proposed system, including text, diagrams, and tables as needed. The description should cover at least the following, as applicable:

- **Compute Processors and Nodes**: Types, quantities, connectivity, memory, NICs, and local storage.

- **Component Architecture:** Processors, memory, storage, and network interconnects.

- **Node, Board, & Blade Design (as applicable):** Integration architecture and details. Management
- **Rack & Cabinet Architecture:** Organization, interconnects, and scalability.

- **Interconnect Topology:** High-speed network connectivity across all components.

- **Management Nodes:** Hardware for system operations, ensuring accessibility even if compute nodes are down.

- **Interactive and persistent services nodes:**
    - Ability to handle a large number of interactive users and typical pre- and post-processing activities (including compilation, plotting and analysis). The response should articulate a range of the number of interactive users that the system can support (e.g., 1000-5000).

    - Ability to support persistent services such as science gateways and other self-service or centrally operated custom portals.

    - Ability to support infrastructure management software such as metric collection/archiving and alerting, as well as administrative management systems such as account management, reporting, and billing.

Responses that exceed performance targets given below are expected to rank higher, responses that do not meet all the targets below are expected to rank lower. Consistent with industry practices, most targets are given in terms of hardware peak characteristics. Unless otherwise indicated, performance target numbers are numbers to meet or exceed.
A separate set of *deliverable performance benchmarks* are also required (see subsection **Performance Benchmarks**). Contractual system acceptance is expected to require acceptable performance on satisfactory *deliverable performance benchmarks* results measured on site.

Compute Requirements:

Targets for for the sum of the capabilities of Tier-1 and Tier-2 logical blocks described in Appendix 2, for IEEE-754 FP64 arithmetic, IEEE-754 FP32 arithmetic, TFP32 arithmetic, TFP16 arithmetic, FP8 and FP4 arithmetic:

- **FP64:** 8.2 Pflop/s aggregate
- **FP32:** 90 Pflop/s aggregate
- **TFP32:** 390 Pflop/s aggregate
- **TFP16 (IEEE and BF):** 746 Pflop/s aggregate
- **FP8:** 1.5 Eflop/s aggregate
- **CPU:** 16 cores per AI/ML accelerator
- **FP4:** 2.1 Eflop/s aggregate

Compute targets for service node (persistent services, gateways, operations) logical blocks described in Appendix 2:

- **SNCPU:** 4000 physical cores aggregate
- **SNNUM**: 30 or more separate service nodes

Memory:

Memory targets for the sum of the capabilities of Tier-1 and Tier-2 logical blocks from Appendix 2:

- **T1MEM:** 34TB HBM3e (or at least equivalent) aggregate
- **T2MEM:** 38TB GDDR6 (or at least equivalent) aggregate
- **HOSTMEM:** 256 GB DDR per AI/ML accelerator

Memory targets for the service node (persistent services, gateways, operations) logical blocks from Appendix 2:

- **SNMEM:** 32TB aggregate DDR
- **SNHOSTM:** 8GB/core

Storage:

Storage targets for Tier-1 and Tier-2 AI/ML blocks

- **T1T2LS**: 3.84TiB NVMe per AI/ML accelerator

Storage targets for service node (persistent services, gateway, system operations services) logical blocks from Appendix 2:

- **SNLS:** 3PiB aggregate
- **SNHOSTLS:** 100TiB/node

Storage targets for the high-speed storage logical block in Appendix 2:

- **HSSCAP:** 4PiB aggregate
- **HSSRBW**: 100GB/s aggregate
- **HSSWBW**: 70GB/s aggregate
- **HSSIOPS**: 3.10^7 IOPS
- **HSSWE**: 1DWPD

Storage targets for the commodity storage logical block in Appendix 2:

- **CSCAP:** 10PiB aggregate
- **CSRBW**: 5GB/s aggregate
- **CSWBW**: 5GB/s aggregate
- **CSIOPS**: 2.10^5 IOPS

Responses may include expected performance for other standard file system operations in order to demonstrate the capability of their solutions. These operations include total number of files supported, additional measures of meta-data capability (e.g. time to delete 10 million small files), number of directory entries (where relevant), filesystem specifics around corruption

detection and around snapshot capabilities, indexing performance over deep and shallow hierarchies (filesystem stat performance over a tree of billion files in various directory structure levels).

In particular, the response should list the estimated time taken to:

- Write half of the total memory from Tier 1 and Tier 2 systems to the high-speed shared storage system block (ideally we expect half of the memory can be stored in the storage system within 10 mins to enable use cases that rely on fast checkpointing to shared storage).

- Cold start 1000 independent PyTorch sessions across Tier 1 and Tier 2 systems concurrently.

The response should provide performance measurement results for the time it takes to create one million files (each 64KB) from each compute node. A file creation operation should consist of the following metadata operations: open, write, read, close. The response should quantify file system performance (aggregate bandwidth and latency from compute nodes) at different utilization levels (file system 50% full, 75% full, and 90% full).

Internal Networking:

- **INT1LAT:** <20us latency single AI/ML accelerator to single AI/ML accelerator in Tier-1 pool for 8-byte message

- **INT2LAT:** <40us latency single AI/ML accelerator to single AI/ML accelerator in Tier-2 pool for 8-byte message

- **INT1BIBW:** 6TB/s bi-section bandwidth across the Tier-1 AI/ML accelerator pool

- **INT2BIBW:** 6TB/s bi-section bandwidth across the Tier-2 AI/ML accelerator pool

Services Nodes:

Proposals should describe solution elements that address the services nodes in the logical diagram in Appendix 2. Two flavors of service blocks are shown. One hosts services to which non-privileged, regular research use accounts attach (the Persistent Services and Gateway block). The other hosts system management services that are used by system operations to manage the operations of the proposed solution.

SNGPCAP: Support for at least 1000 concurrent sessions (across web portal/science gateway, terminal and data transfer) for at least 10 distinct organizations.

SNSCAP: Infrastructure to support hosting a full suite of system services and operational automations. As relevant to a particular RFP response, these include:

- User home directories serving, core software serving
- Identity services (e.g. LDAP/AD etc…)
- Name resolution services (e.g., DNS/BIND etc..)
- Provisioning services (e.g., TFTP, DHCP, PXE etc…)
- Imaging services (e.g. xcat, Warewulf4 etc…)
- License services
- Workflow management/orchestration services (e.g. SLURM, NOMAD/Portainer etc…)
- Monitoring and alerting services (e.g. CHMK, nagios, ufm etc…)
- Logging services (e.g. LOGSTASH/ELKSTACK etc…)
- Data management and life cycle services (e.g. starfish/mediaflux etc…)
- Network gateways (e.g. NAT, firewall etc…)
- Account registration (e.g. Coldfront etc…)
- Billing and allocation/quota services (e.g. Coldfront etc…)
- Resources to host overall management systems (e.g. Base Command Manager etc…)

System Software:

Proposals should address the following system software requirements:

AI/Ops System Provisioning and Management

The proposal should present a **comprehensive software solution** for system provisioning, management, and monitoring. This solution must address all the needs of a modern AI/Ops team, including:

- **Monitoring and Alerting**: Covering chiller and cooling distribution units, as well as network, storage, and compute hardware.
    - **Proactive Capabilities**: Clearly outlining any proactive alerting and automated response mechanisms.

- **Workload Management and Hardware Support:**
    - The system must **support multiple users with heterogeneous workloads**. Proposers should detail their approach to workload scheduling and management.

    - The proposal should describe the **maturity of AI/ML accelerator components and associated hardware**, ensuring full compatibility with **PyTorch, JAX, and TensorFlow workloads**.

    - If the proposal includes **non-x86 and/or non-CUDA-based solutions**, it must specify:
        - The level of coverage guaranteed for current and emerging AI/ML tools.

        - Any gaps in compatibility, which should be explicitly identified.

Open-Source and Software Architecture Considerations:

Proposals emphasizing a **maximally open-source approach**, where practical, will be ranked higher than those incorporating proprietary solutions without clear justification. The proposal must include:

- A **high-level software architecture diagram**, highlighting major components, dependencies, and classifications (open, shared, or closed source).
- A commitment to **timely sharing** of software before general availability, where applicable.
- Any plans for **integration, coordination, testing, and release schedules** for different software types.
- If bespoke software is part of the solution, an explanation of how it will be developed.
- List of international open-source and proprietary software that are expected to be a part of the default offering.

Licensing and Source Code Accessibility:

The proposal must outline:

- The **software licenses** used and the rationale for their selection.
- Open-source software that will be **coordinated upstream**, along with any reliance on external development.
- The **extent and nature of access** to source code, build environments, and updates (including firmware, compilers, and third-party software).
- A clear specification of **included and excluded licensing terms**.

Power and Cooling:

- **PWRSYS:** <650KW maximum power draw for system at full real-world workload. This power draw should include any internal cooling devices such as in-rack CDU units.

The requirement above and the site conditions detailed in the Infrastructure FAQ can potentially be adjusted depending on legitimate system needs. Any response that does not fit the above requirement or the conditions detailed in the Infrastructure FAQ should clearly highlight the deviation and explain the advantages of the deviation.

Reliability and Resilience:

- **RRPWR:** Responses should take into account the site conditions detailed in the Infrastructure FAQ, and should clearly indicate:
    - Which components have N+1 (or greater) power feed redundancy. This redundancy should allow continued operation when one leg of an A/B power feed architecture is disabled for maintenance.
    - Any components that have partial or no power feed redundancy

Physical Form Factor:

Responses should include detailed floor plan and rack diagrams for any physical hardware based proposed system or system component physical layout.

- **RRFRM:** <16 full racks or equivalent

- **RRWT:** <2000kg per rack

Items that are to be run on generator backed UPS circuits (critical storage, service nodes and gateways) should be clearly marked. For any physical hardware, these may be located in a different section of the install site from compute resources that are not UPS-backed.

External Systems Integrations:

Responses should include a description of external system integration options. This should cover both hardware that can act as external gateways and any software capabilities that can help with integration. Particular areas of interest include:

- Integrations that allow secure mapping of data from other resources within a trusted domain onto the proposed resource's namespace(s) and administrative domain.

- Resources and software for transparent/semi-transparent interfacing of Slurm workflows with external systems.

External systems integrations should address external interface elements for two families of external systems:

- Elements for interfacing for up to 10 systems all within the Massachusetts Green High Performance Computing Center (MGHPCC) data center and operated by trusted peer organizations that are members or close collaborators of the AICR consortium.

- Elements for interfacing globally with collaborating activities all over the world. These connections include academic enterprises, traditional hyperscale clouds, AI hyperscalers and regional synergistic data center facilities.

For both families, the proposed solution's ability to support shell access, web portal access, and bulk data transfer access should be described.

- **ESGW:** 100Gb/s Ethernet ingress/egress capabilities to both trusted peer MGHPCC systems and globally.

- **ESDTN:** Dedicated data transfer services (Globus, rclone, rsync servers)

- **ESISO:** Ability of external interfaces to add/remove and accept/reject layer 3 VLAN tags that may be used for logical traffic isolation.

C. Performance Benchmarks

Responses should include a detailed quantitative performance evaluation of the following key performance benchmarks. The figure of merit is end-to-end performance or application-specific performance metrics (as applicable). The performance evaluation should include single-node and multi-node performance for Tier-1 and Tier-2 systems.

High-priority critical benchmarks:

- **ML training and inference**

  https://mlcommons.org/benchmarks/training/

  https://mlcommons.org/benchmarks/inference-datacenter/

- **HPL: High Performance Computing Linpack Benchmark**

  https://www.netlib.org/benchmark/hpl/ 64-bit and 32-bit equivalent.

- **LAMMPS: Molecular Dynamics application**

  https://www.sandia.gov/ccr/focus-area/molecular-dynamics/

  https://docs.lammps.org/stable/Speed_bench.html

  The reported results should include:

  - The number of atoms, timesteps/s, and atom-step/s.
  - The configuration file listing boundary conditions, size of the box, number of atoms/molecules, and the performance log at the end of the execution.

- **IO500 benchmark for benchmarking IO system:**

  https://io500.org/about

- **Communication benchmarks**

  https://mvapich.cse.ohio-state.edu/benchmarks/ and any equivalents (e.g. NCCL https://github.com/NVIDIA/nccl-tests ; RCCL https://github.com/ROCm/rccl-tests etc.)

- **Estimated Tier-1 LLama7b training time and AI/ML accelerator resource needs** following the recipe at: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/dgxc-benchmarking/resources/llama2-dgxc-benchmarking

Guidelines for reporting benchmark results:

As much as possible, the reported results should correspond to the original source code of these benchmarks without significant code modifications. The proposal should explicitly document all hardware-specific optimizations (including compiler flags).

D. Scalability and Future Growth

ST2 System Potential Growth

- Respondents can choose to describe any capabilities to enable an ST2 system to be integrated with the ST1 system. Additionally, any information on the possible capabilities of an ST2 system may be included. Any plans and alignment with the technology roadmaps may be articulated and described – including any previous experience and demonstrations of such capabilities.

ST3 System Outline

- Respondents can choose to describe any capabilities to enable an ST3 system to be integrated with the ST1 and ST2 systems. Additionally, any information on possible capabilities of an ST3 system may be included. Any plans and alignment with the technology roadmaps may be articulated and described – including any previous experience and demonstrations of such capabilities.

We expect the combination of ST1, ST2 and ST3 to have a total power demand of under 2MW. Any additional information on likely power profiles may be included.

E. Energy Efficiency

Responses should describe:

- Any energy management features that are part of the solution, such as the ability to automatically rate limit processing to optimize energy use.

- Any included features related to real-time energy tracking and correlation with workload manager allocations should be described.

- Any ability to monitor and report energy usage per job for different compute and memory systems. Ability to monitor and report energy usage when multiple jobs are co-located.

- Any available estimates of carbon footprint of compute, memory, and storage systems (per component and aggregate) – both embodied and operational carbon footprint should be included and any innovative techniques and architectural support toward the broader sustainable HPC goal.

F. Security

The ST1 system should be able to handle basic isolation of the sort required to separate data access through Linux groups. There may not be an explicit expectation that the ST1 system will handle regulated data initially. It is expected that pilot work on a regulated data element of the system will begin within one year of initial deployment and that the ST2 system is expected to introduce a sizable element that can be allocated to regulated data use cases.

- Responses should describe any persistent storage system features for hardware based encryption at rest. If the proposed solution does not include this option or if the option is separately priced, responses should make this clear.

- Respondents may choose to provide documentation of their relevant experience with systems for regulated data and compliance processes, particularly any experience centered on NIST800-53 and NIST800-171.

G. Support and Services

Responses should include the following:

- A detailed statement describing the work the respondents will undertake ahead of time, and (where relevant) on-site, to install, configure the system and run acceptance and burn-in tests.

- A plan for providing necessary training to operations and maintenance staff. The plan should include the number of hours of training that will be provided and to how many people.

- Warranty coverage. All hardware components should come with five year manufacturer warranty coverage.
    - Describe the terms of that warranty coverage and clarify the terms of all the service-level-agreements governing the warranty.
    - Outline any technical support and maintenance that is included.
    - Outline all details of the warranty services that are provided by third-party organizations.

- Comprehensive details about any additional dedicated value-added support services, such as a Technical Account Manager (TAM), offered either to AICR or MGHPCC as a whole.

- Outline strategies for addressing persistent systemic issues that affect system capability and capacity, including remediation methods, preventive measures, and long-term solutions.

H. Documentation

Respondents should describe their plans to provide as-built documentation for handover. If possible, include references to projects that have received similar documentation, and attach templates and examples illustrating prior experience and work.

5. Response Guidelines

A. Budget and Pricing

Cost Breakdown:

- Every response should contain a section that clearly an concisely summarizes overall pricing and the pricing of each major subsystem. This section should be in addition to a detailed breakdown of individual component building block prices and a bill of materials description.

- The summary section should be laid out so that the response evaluation team can easily see the overall proposed project price and how much each major sub-component or activity contributes to that price. As far as possible, this summary should be designed to allow the evaluation to understand the impacts of fine-tuning the proposal response to alter the balance or mix of capacity in each subsystem.

Pricing Model and Payment:

- Pricing quoted in responses is expected to be best and final and will be treated as such. As far as possible the evaluation team does not expect to further negotiate pricing when reviewing responses.

- It is expected that the RFP will result in a fixed price contract with the selected respondent.

- Responses should include any restrictions around payment terms and schedules that may apply. Final terms will be subject to negotiation as part of vendor selection. However, responses should make clear any terms that would typically be expected by the respondent.

- Responses should make clear which proposed benchmarks and any other criteria that are anticipated to be used in evaluating any contract payments for the system. This set and the measures may be subject to negotiation in reaching a final agreement.

Contingencies and assumptions:

- Respondents should include their current or expected practices regarding unexpected external events such as tariffs, currency exchange rate swings, supply chain disruptions, manufacturing capacity, and unanticipated hardware or manufacturing quality issues that could impact the eventual cost of the project and that are not included in the pricing model. This information will be considered as part of any selection process and any contract that is entered into will include terms addressing external uncertainties. Any factors that may impact ongoing operational costs should be highlighted.

B. Timeline

- Respondents should include a project timeline that describes a proposed week by week schedule covering the time period from contract agreement to system acceptance. The timeline should clearly indicate any activities that will entail customer support staff and resource commitment.

- Respondents should summarize contingency options, if any, that may be exercised in the event of unexpected delays.

C. Evaluation Criteria

Proposals will be ranked and categorized according to multiple criteria. These criteria and overall programmatic considerations will form the basis of a selection.

Technical and Programmatic Fit:

Each proposal will be evaluated on how well the overall solution of compute/storage/network capabilities, software, and support services meets the project needs. These capabilities will be assessed by how well responses align with the bolded all caps metric criteria listed in the Capability and Capacity section of the RFP. Reviewers will also rate proposed systems according to how well the approach can meet the needs of a mixed workload with significant variability in computing, system memory, and networking demands while also maximizing utilization. The user experience and a range of use cases will be considered. The system needs to be flexible enough to serve a heterogeneous user community with heterogeneous workloads. An ST1 design that scales to our emerging understanding of our users over time will be valuable. Please see Appendix 1 below for more information.

Performance:

Forecasted performance on system benchmark results and scalability tests will be used to rank proposals in this category. The performance ranking will take into account which benchmarks are proposed to be included in the acceptance criteria for payment.

Cost:

Value for money and budget alignment with target capacity and capability metrics in preceding sections will be ranked across all responses. Responses that meet or exceed more of these metrics within the overall budget will be ranked more highly.

Additionally payment terms and price certainty guarantees will be taken into account for evaluating proposals. Respondents should make clear any risks from internal or external factors that could alter agreed pricing between a project contract agreement and final payment.

Support and Service:

Evaluation will consider material demonstrating the vendor's prior experience and reputation in executing projects similar to the work solicited in this RFP. Respondents are encouraged to provide references of installations and projects to help with this evaluation. The terms of service, support service level agreement contracts, and the completeness of coverage will also be taken into account when ranking responses.

Energy Efficiency:

Evaluation will also rank responses according to justifiable estimates of energy use of the system under full load. These estimates should be for system energy use when all the proposed AI/ML accelerators are in use.

Software Stack:

The overall scope, manageability and long-term sustainability of any proposed software stack will be considered. This will include consideration of appropriate use of open source tools where possible and consideration of any justification for use of proprietary software. The suitability of the stack to provide a foundation of production services with adequate basic security for a mixed use system as outlined in Appendix 1 will be considered.

D. Solution Diagrams, Bill of Materials Documentation and Template Statements of Work

Responses should include diagrams showing physical layouts (where relevant) and logical layouts of the system. For physical layouts, the configuration of the complete system in racks with all cabling should be shown.

An additional complete bill of materials should be included for every response. This should include all physical devices where applicable. A breakdown of all logical device types and their counts should also be included.

E. Legal and Contractual

The system proposed will be procured by the Massachusetts Green High Performance Computing (MGHPCC) organization. Contractual terms will follow MGHPCC standard practices. We expect to engage with competitive respondents to finalize detailed contractual and acceptance terms governing payment and legal commitments as part of the evaluation and vendor selection process.

More than 80% of the funds supporting this RFP are obligated from sources that require their use for capitalizable expenditures under accepted accounting principles. To be competitive in this RFP, respondents should propose a solution and payment terms that are consistent with this requirement.

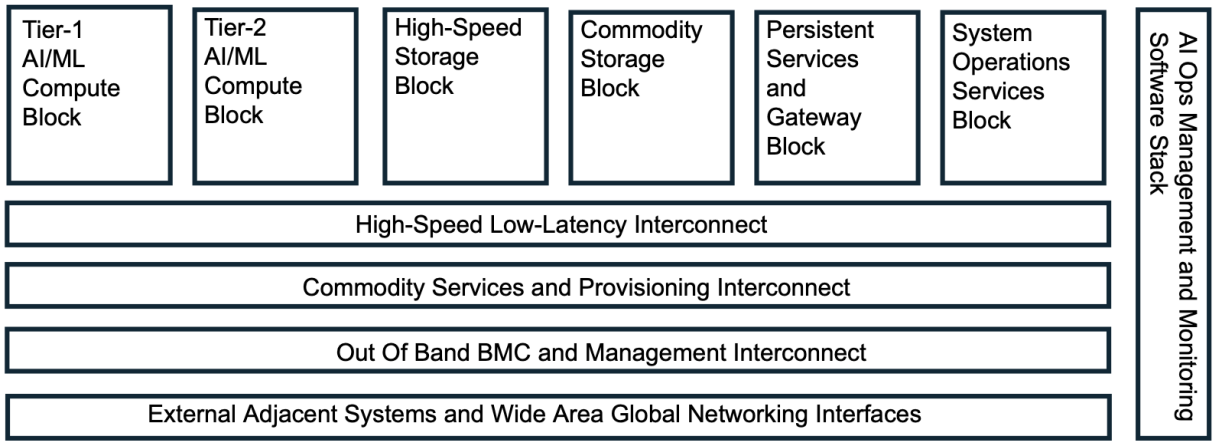
Proposals received after the deadline will be returned to the proposer unopened. MGHPCC reserves the right to reject any or all proposals submitted, with no obligation to explain the reasoning, and to make the award where it appears it will be to the best interest of MGHPCC. The successful vendor will be expected to abide by MGHPCC policies on procurement, security, and confidentiality. All awards are subject to additional documentation and definitive agreement.

Appendix 1 - Use Cases

AICR will serve a variety of entities, including large academic and research communities, small and medium businesses, start-ups, and entrepreneurs. As a result, vendors should anticipate a highly diverse community of users with very different needs and a wide range of experience with AI/ML development, application, and deployment. Vendors should consider how to serve this heterogeneous workload as they consider hardware, software, networking, support, security, and storage.

- Graduate students working across a range of disciplines with limited or moderate technical knowledge mostly running single GPU or single node jobs;
- Innovators and researchers who are expert users, building complex, large scale models requiring high-data transfer speeds;
- Entrepreneurs or researchers partnering with industry who are working to design novel generative models to facilitate drug-discovery using confidential and proprietary data;
- Academic researchers interested in replicating recently released open source models, iteratively evaluating performance and testing innovations during training and inference;
- Lightly capitalized startups and pre-seed enterprises building out initial capabilities using novel AI models for proof-of-concept prior to industry-level scaling;
- Public service entities working with open source models and open data to fine-tune for new applications;
- Individual students or research professionals from non AI/ML fields wanting to leverage emerging AI/ML tools;
- Multi-participant teams collaborating on common goals and working across skills and AI/ML experience;
- Undergraduate students learning how to design and deploy modern AI algorithms using interactive notebooks on multi-instance GPUs.

Appendix 2 -
Hypothetical system conceptual block diagram



The block schematic in Appendix 2 shows key logical blocks that the ST1 system is envisioned to contain. RFP respondents should detail how their solution maps to these blocks. A solution does not need to have distinct physical or logical subsystems for each block. For example, a single storage solution could address all the storage needs. However, respondents should show what part(s) of their solution supports each block. The blocks are briefly described below.

Tier-1 Compute:

The tier-1 compute block is a set of AI/ML hardware optimized for multi-node and multi-device training and inference. It is envisioned to be able to support training across all the elements of the block by a single optimization and inference across all the elements for highly linked query/reasoning/agentic activities. This block can also support numerous independent activities, but it is distinct because it has strong support for large integrated AI/ML problems. Tier-1 compute elements may include ephemeral local storage to hold training and model data that can be accessed at maximum speeds.

Tier-2 Compute:

The tier-2 compute block targets single or few device AI/ML activities. This block is intended to support large numbers of different projects carrying out different, independent AI/ML activities. It is envisioned that the network bandwidth and latency needs per AI/ML device of tier-2 are lower than tier-1. Tier-2 compute elements may also include ephemeral local storage to hold training and model data that can be accessed at maximum speeds.

High-Speed Storage:

The high-speed storage block is a subsystem that provides shared storage visible to all parts of the system and that supports high IOPS rates and high IO bandwidth. It is intended for holding models, training data and other resources that need fast and potentially random access as part of AI/ML workflows. This storage block is not intended for long-term storage.

Commodity Storage:

The commodity storage block is intended to be a lower cost per unit capacity, lower performance but higher capacity shared storage subsystem. It is not intended to hold active training data, or to serve models that are being used intensively.

Persistent and Gateway Services:

Persistent and gateway services host user accessed services that are longer duration than typical AI/ML experiments. These services include home directories, login services and also web portals and gateways serving AI/ML models and tools.

System Services:

System services host administrator managed services that are core to supporting workloads and day-to-day operations. As noted in the technical section, these service include; home directories serving, core software serving, identity services (e.g. LDAP/AD etc…), name resolution services (e,g, DNS/BIND etc..), provisioning services (e,g, TFTP, DHCP, PXE etc…), imaging services (e.g. xcat, Warewulf4 etc…), license services, workflow management/orchestration services (e.g. SLURM, NOMAD/Portainer etc…), monitoring and alerting services (e.g. CHMK, nagios, ufm etc…), logging services (e.g. LOGSTASH/ELKSTACK etc…), data management and life cycle services (e.g. starfish/mediaflux etc…), network gateways (e.g. NAT etc…), account registration (e.g. Coldfront etc…), billing and allocation/quota services (e.g. Coldfront etc…). This also includes resources to host overall management systems (e.g. Base Command Manager etc…)

AI Ops Management and Monitoring Stack:

The AI Ops Management and Monitoring Stack block is a collection of software needed to effectively administer the system operations and carry out monitoring and alerting to ensure any fault notifications are promptly reported for response.

High-Speed Low Latency Interconnect:

The High-Speed Low Latency Interconnect block is a physical/logical internal interconnect that links AI/ML accelerators to support efficient, scalable distributed training across a significant fraction of the Tier-1 and Tier-2 blocks. The interconnect is also expected to be the fabric that connects AI/ML accelerators to storage for loading models and for sampling training data that is larger than local node storage.

Commodity Services and Provisioning Interconnect:

The Commodity Services and Provisioning Interconnect block is a physical/logical internal interconnect that supports generic services such as VLAN segmentation and traffic isolation, DHCP, PXE boot, TFTP, that are used for provisioning and potentially higher level isolation services.

Out of Band BMC Interconnect:

The Out of Band BMC Interconnect block is a physical/logical internal interconnect that supports out of band access to systems for power control, access to control BIOS as appropriate, and system hardware level monitoring and reporting.

External Adjacent Systems and Wide Area Global Network Interfaces:

The External Adjacent Systems and Wide Area Global network Interfaces block is a physical/logical block that supports ingress/egress traffic to external systems and connects to the internal interconnects. It is envisioned to consist of a series of gateway nodes with interfaces that route layer 3 traffic to external paths that connect to systems within the MGHPCC environment as well as to systems hosted at MGHPCC member institutions, AI Hub, and partner institutions.