MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER

Frequently Asked Questions- RFP/RFI
AI Compute Resource
Infrastructure System

Last Updated April 17, 2025

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

Contents

*Last updated 4/17/25

The most up to date version can be found at https://www.mghpcc.org/ai-compute-resource-system/

Benchmarking

**All questions below were published on 3/28/25*

1. How is a figure of merit (FOM) calculated? Are all the benchmark results given equal value in the FOM?

At this stage proposers should choose a figure of merit that makes sense to them and explicitly document them. The relative weight given to benchmarks in any selection will favor overall capability across all benchmarks. Our goal is to deploy a system that provides value and capability to a broad community. Exemplar performance on a single benchmark at the expense of poor performance on other benchmarks will not be as highly rated as strong performance across all benchmarks.

2. Re SLURM integration: What is the scale for the job so that we can select/prepare the benchmark(s)?

Slurm should be able to comfortably schedule jobs for thousands of users running a mixed workload with median job duration in the range of an hour to a few hours. It should be able to cope reasonably (response times in seconds maximum) with bursts in which hundreds of short jobs are launched concurrently. Launching of a single job on an idle system that leverages all AI/ML resources should be well under a minute.

3. Do you have any prototype-sation for benchmark (lab use cases) (AI exp environments vs HPC cases)?

No.

4. If we propose a GPU that is not currently available (eg, NVIDIA B300), can we submit “predicted” benchmark results based on GPUs that are available?

Yes - but we are interested in systems that can have some early user reasonable fraction of production level activity in place in the August 2025 timeframe.

5. Usually, FP32 and FP64 are in a 2:1 ratio. The current aggregate targets have an 11:1 FP32:FP64 ratio. Is this correct?

Yes. The AI/ML community has substantially more needs for transistors that provide lower precision and specialist tensor and ray-tracing capabilities than traditional 64-bit arithmetic. The Tier-1 and Tier-2 blocks may be a useful abstraction for thinking about this if there are cost optimizations that enable it.

6. Re: HSSIOPS: 3.107 IOPS - Can you provide details on the workload profile?
Typical examples could be 4000 threads across the cluster each driving small I/O at a rate of 10,000 IOPs per thread during shuffled access to some shared data repository for training. Another example could be 5000 IOPS across a similar number of threads all driving small I/O against a shared library.

**All questions below were published on 4/1/25*

7. Beyond the aggregate 6TB/s bi-section bandwidth target, are there specific per-link throughput requirements for GPU-to-GPU communication?
The RFP includes showing benchmark results for accelerator to accelerator communication using tools such as the NCCL, RCCL or OSU benchmark suite as appropriate. The proposed systems performance for those tests described in responses will be used to evaluate responses. It is likely that these benchmarks will also feature in on-site acceptance tests.

8. Are these listed computational targets (FP64, FP32, TFP32, TFP16, FP8, FP4) expected to be achievable simultaneously and continuously, or are these performance goals flexible, reflecting different workload scenarios?
The aggregates reflect different workload scenarios or phases of workloads. There are certainly some codes that mix precisions for different stages of a machine learning pipeline.

9. Do you anticipate sustained workloads simultaneously requiring maximum FP64, FP32, FP8, and FP4 performance, or typically only one numeric precision at maximum load at a time? For example, your FP32 numbers, 90 PFLOP/s (90000 TF), would require around 1500 GPUs, and then FP8 1.5 EFLOP/s (1500 PF) would require 750 GPUs.
We envision the same accelerators being used to achieve the aggregate performance. Solutions that can reach multiple targets simultaneously will likely be rated higher, but for most current use cases, all the target performance numbers will not be realized simultaneously. It is up to proposers to put forward detailed designs that do their best to be able to deliver on each target. These designs could use a mix of accelerator types or could be constructed of a single type of accelerator, potentially with different software configurations.

10. Should the infrastructure be sized based on the largest single precision requirement (e.g., FP32), or can we assume precision formats are optimized individually at different times?
The performance targets do not have to be met simultaneously.

11. Can you provide more context regarding the FP32 requirement of 90PF?
The workloads are likely to include research projects exploring ML network designs and hyperparameter settings in many domains. Many training projects in this category make use of 32-bit floating point. In many cases a heterogeneous mix of accelerators can provide an optimal balance of floating point, tensor unit and ray-tracing core capabilities.

**All questions below were published on 4/8/25*

12. Are there any restrictions on permitted code versions, or chosen accelerator backend (kokkos vs. default GPU)? Specifically in using the public upstream "patch" versions of LAMMPS, with additional minor code modifications to support modern GPU architectures.
There are no restrictions on choosing publicly-available code versions or accelerator backend -provided the choices are well-specified and changes to the code base, if any, are described in detail.

13. Do you want performance results for all 5 benchmarks in speed bench?
That is our preference unless there is a strong rationale presented for a different choice.

14. Will benchmark results for larger, modern benchmarks be useful? We typically benchmark systems with 1m+ atoms to demonstrate performance.
Yes, larger and longer-running benchmarking would be helpful, esp. multi-node performance benchmarking for LAMMPS. We want to make sure the proposed system has some capacity for simulations around synthetic training data based approaches in the context of AI/ML for research acceleration. We are not expecting to leverage the system for stand-alone traditional HPC work.

15. Are we allowed to include accelerators in total peak calculations?
Yes. The proposal should disclose how peak calculations are being derived.

16. Can we report 32 bit precision results using HPL-MXP?
That is acceptable. Whatever options are used please make that clear in your response.

**All questions below were published on 4/11/25*

17. Regarding performance and acceptance testing, how important are benchmarks like MLPerf, and is there room for partial or delayed results?
MLPerf and similar benchmarks are strongly encouraged to illustrate system performance. Vendors should provide best-available projections in their proposals. It is expected that acceptance tests will be finalized during contract negotiations.

18. Will there be separate acceptance tests for HPC and AI use cases, or is MGHPCC focusing primarily on AI/ML performance?
MGHPCC's priority is the AI/ML focus, but they anticipate HPC-like usage among certain constituencies. Acceptance testing will include relevant HPC metrics if needed; the goal is to validate a balanced solution that can serve both HPC and AI efficiently.

19. Can we use different benchmarks in our proposal?
Yes, if they are in addition to and not instead of the ones we have specified.

20. Do we need to provide performance evaluation results for all benchmarks listed in the RFP?
We prefer that the performance evaluation results are as complete and as detailed as possible. However, we also recognize the challenges of conducting rigorous performance benchmarks, therefore, we encourage the responders to (a) include the benchmarking results to the extent possible, and (b) provide reasoning/rationale for remaining benchmarks where results could not be reported.

Compute

**All questions below were published on 3/28/25*

1. Does T1T2LS need RAID and backup?
This is intended to be ephemeral storage so will not be backed up. On systems with multiple AI/ML accelerators per node we envision creating a single volume that spans NVMe devices if the design employs multiple devices. We do not envision using hardware RAID. Proposers are free to suggest that, but our expectation is some form of software based solution that preserves full capability of the devices will be preferred.

2. Does SNLS storage (3PiB aggregate) need some form of redundancy for handling component failure?
This storage will be used to serve persistent systems tools and support systems operations that need to be resilient to failure; some form of RAID or RAID equivalent would make sense.. The aggregate storage amount could be a single volume or could be proposed as a set of smaller volumes. That is up to the proposer.

3. Is the evaluation committee interested in low-power accelerators for inferencing?
Inferencing research and innovation is of interest. New technologies that are proven and have viable software stacks with a reasonable level of maturity will be considered.

4. Is the project aiming to support robotics AI/ML research?
Robotics is called out as an area of interest. Solutions that include features geared toward a robotics train, sim-eval cycle without compromising in other areas would be of interest.

**All questions below were published on 4/1/25*

5. Is the project aiming to support robotics AI/ML research?
Robotics is called out as an area of interest. Solutions that include features geared toward a robotics train, sim-eval cycle without compromising in other areas would be of interest.

6. Is there a preference of processor, Intel/AMD?
Vendors are encouraged to propose whatever CPU and GPU make and model they feel is the best and most cost-effective match to the requirements stated in the RFP.

7. Do you prefer liquid or air cooled systems?
We are open to both but generally would prefer cost-effective liquid cooling solutions

8. How did you arrive at the requirement for 30 service nodes? Can we deviate from this number?
This is an approximation based on existing systems that each of our institutions currently has deployed and responses can deviate from this number.

9. Will the Service Nodes require GPUs for compilation or other activities?
This could be a possibility. Respondents are free to design service nodes in a way that they think is advantageous to their overall solution value.

10. Can persistent services be self contained and separate from the core ST1 cluster?
We are open to exploring persistent workload solutions that leverage the core ST1 cluster or are more standalone in nature.

11. Commodity Storage (CSIOPS: 2.105 IOPS): Can you provide the details on the workload profile?
Commodity storage is envisioned to be storage that is used for nearby cache of models and datasets. It is intended to provide good price per capacity but lower performance and to be used to stage things to and from storage that will be used for active computation.

12. Is SNCPU (4000 cores) specifying physical or logical cores?
This was intended to refer to physical cores, to the extent that the distinction is clear.

**All questions below were published on 4/4/25*

13. For persistent gateway services serving AI/ML models, should these have dedicated GPUs or use compute nodes allocated through the scheduler?
Both approaches are acceptable and are compatible envisioned modes of operation of the system.

14. How does MGHPCC plan to handle rapidly changing GPU technology (e.g., new releases every 18 months)?
For on-prem solutions, MGHPCC intends to purchase and deploy hardware in tranches, allowing the consortium to adopt updated GPUs as they become available. It is expected that cloud and hybrid cloud/on-prem responses will detail how new technologies will be made available to AICR as they become available in the vendor's environment

**All questions below were published on 4/8/25*

15. When sizing service nodes can we consider the number of logical cores available to applications vs physical cores?
Yes. The proposal should clearly state the types of cores being used.

16. How strongly does IOC prioritize FP64 performance for HPC relative to FP16 or BF16 for AI?
Applications may still need FP64 or FP32 for simulations centered on synthetic training data generation and fused synthetic data and training/validation coupled paradigms. AI training and inference tasks will rely on lower-precision. Vendors should showcase how both needs are addressed.

**All questions below were published on 4/11/25*

17. Model scale is an issue, size growing faster than hardware clusters have been growing in size and density - how do you address this?
We are addressing technology advances and capability increases by engaging in periodic purchases of equipment (ST1-ST3) to make the best use of our set budget.

18. Are you open to cloud "bursting" to support FP32 needs?
Yes.

19. In looking at aggregate performance, gpu, memory, and budget, are you looking to specific chip designs for Tier1 and Tier2 compute?
No.

20. What drove the requirements for the service node networking and storage?
We expect an array of different use cases for general compute such as data staging, hosting software updates, customizing the logon/access experience and other yet-to-be-identified services. We do not anticipate or want people to be running aggressive, compute-intensive workloads there.

21. What are MGHPCC's expectations for service (management) nodes, and how might these nodes be used?
Service nodes may handle logins, Jupyter sessions, data transfer utilities, and certain small-scale compute tasks in addition to the management of the cluster as a whole. Generally we want additional resources to provide limited capabilities to our users to work and develop their codes without tying up the large GPU nodes in a system.

22. Does MGHPCC want separate networks for Tier 1 (HBM-based) and Tier 2 (standard memory) GPU compute, or can one unified network fabric serve both tiers?
A single, unified fabric is acceptable if the vendor can demonstrate it meets (or exceeds) performance goals for both tiers. Alternatively, MGHPCC will also consider designs where Tier 1 and Tier 2 use separate networks.

23. Can we propose a solution with an expected real-world power consumption <=650kW but nominal max >650kW?
Yes but please provide details on your throttling strategy/capabilities to stay within the power envelope.

**All questions below were published on 4/15/25*

24. What drove the performance requirements for Tier1 and Tier2 memory?
We expect a mix of applications will benefit from reasonably large memory GPUs and assumed reasonable HBM requirements per GPU.

25. Are the aggregate performance numbers for precisions at fp32 or lower intended to reflect peak performance as measured by vendor published sparse performance numbers, or are we supposed to use dense math performance numbers instead?

The expectation is that the response includes peak performance numbers obtained via benchmarking the real chip/system. In case the chip/system is not available, projected numbers with strong justification are allowed, but but they should be clearly marked as such (that projection/estimation/raw vendor product sheet numbers were used)

26. Are the fp32 workloads running 24x7 or periodic?
Periodic. We do not expect any 24x7 workloads.

Facilities

**All questions below were published on 3/28/25*

1. Will the power to the racks and CDUs be powered by UPS?
No. The cluster will run from utility power. Sufficient UPS will be provided so that critical storage can be protected during a utility outage. It is expected that this protection will require keeping certain storage servers, management servers and networks operational until power is restored. ST1 has allocated 80kW for this purpose. Any additional UPS power needed by the solution will require facility engineering.

2. ST1 electrical distribution states limited power is available to support critical cluster equipment. What racks/custers are critical?
Critical services like base management capabilities, login and some fractions of storage should be considered critical. For any physical hardware proposal the ST1 system will have access to up to 80KW and 3 racks of floor space for critical equipment that is running against UPS generator backed power. Respondents should propose solutions that make most effective use of that capability.

3. If the CDU is providing cooling to a critical cluster, is there capacity to feed the CDU with UPS power?
The total power budget for critical equipment, including CDUs, is 80kW

4. If power to either feed to the CDU is lost, flow to the cluster could be lost for a short period, impacting cooling?
The cluster should not sustain damage if power is lost for short periods of time.

5. The fluid temperature is shown between 60-65°F. What type of fluid is being used? 100% water or a glycol/water mix?
100% water.

6. Is there a required return temp for the fluid loop?
The chilled water plant for the facility is designed for a 20°F ΔT. Accordingly, return temperatures should be between 80°F and 85°F.

7. The relative humidity is shown at 20% - 80%. Are there any units within the space designed to maintain humidity/dew point? If the dew point increases, secondary fluid network temps would need to increase to prevent condensation. Rear door heat exchangers could also experience reduced capacity or condensation.
Makeup air units humidify computer room supply air to maintain the dew point between 42°F (winter mode) and 50°F (summer mode)

**All questions below were published on 4/1/25*

8. What is the kW per rack you are looking to provide?
Please review the MGHPCC Data Center Infrastructure Frequently Asked Questions document. If you have further questions please email them to us for review.

**All questions below were published on 4/11/25*

9. Section 2 of the Facility section in the FAQ says: "ST1 system will have access to up to 80KW and 3 racks of floor space for critical equipment that is running against UPS generator backed power." Are these racks intended to be part of, or in addition to the 16 rack maximum footprint stated in the Physical Form Factor section of the RFP?
These racks are in addition to the <16 noted in the Physical Form Factor section of the RFP.

10. Are the 16-rack and 650 kW limits firm, or is there room for negotiation if a design exceeds those metrics slightly?
MGHPCC treats 16 racks and 650 kW as important guidelines for Phase 1 but can allow some flexibility if going beyond them offers significant benefits (e.g., better performance, future-proofing). Major deviations, however, would be challenging due to the project's current timeline and power constraints.

11. Are the 16 racks all located in one row on the data center floor?
Yes.

12. What is the maximum power per rack?
Please refer to the Facility FAQ.

13. “Can we propose partial racks or multiple smaller racks if that helps achieve budget/power targets?”
Yes, partial or additional racks are acceptable if they satisfy RFP requirements and provide a well-documented approach to real-world usage within 650 kW.

General

**All questions below were published on 4/1/25*

1. What is your expected timeline for future expansion (if any), and how should the architecture accommodate additional GPUs and nodes without a complete redesign?
As stated in the RFP phases ST2 and ST3 are expected to occur 18 and 36 months after the initial ST1 phase. The RFP invites proposers to suggest possible upgrade paths, if they have them.

2. Will ST2 and ST3 be scaled-down versions of ST1? Is the big difference that ST2/3 will have the ability to host regulated information?
As stated in the RFP "A projected investment of $10M to $20M has been planned for ST1. Similar or larger amounts may be available for ST2 and ST3.”

3. Will the sponsoring organizations like Boston University, Harvard, Yale, etc., bring their infrastructure into this ST1, ST2, and ST3, or will they consume the resources?
Initial plans center on participating organizations acting as customers for the resource. The evolution over time will be determined as the project proceeds.

4. We are assuming 5 years HW and SW support for all components, can your team confirm?
Yes. We expect 5 year warranty and support coverage to be included for all parts of the system.

5. Is August a hard date for ST1 operations?
Our stakeholders would like AICR to progress quickly and provide resources in August. If there are compelling reasons to receive portions of ST1 after August please state them in your response.

6. Can we ship portions of our ST1 solution at different times to mitigate supply chain and new technology release dates?
Potentially, yes. If there are compelling reasons to receive portions of ST1 at different times, please state them in your proposal.

7. Is the 'up to $20M' ST1 expected spend a hard number?
Yes, but we are open to small deviations above this amount.

8. Would you be open to leasing a solution?
Due to our funding mechanisms we need to capitalize the majority of this purchase (80%). We are open to reviewing other methods of purchasing that will meet the requirements of our funding that are not overly complex.

9. Would you be open to a direct purchase vs working through a reseller?
We favor simplicity and are open to both direct and reseller purchasing.

10. Is the complete system required by Aug 1? Is it possible to use cloud for the initial onboarding of users by Aug 1? Is it possible to use H200 for the initial onboarding of users by Aug 1 and then phase in B200?
Ideas for employing some cloud resources to align product cycles with deployment goals is a reasonable approach. The RFP and RFI are intended to be neutral to whether physical or virtual resources are employed. Respondents should pay attention to the capital funds discussion. Cloud costs that can be capitalized will be easier to incorporate than those that would be accounted for as operational costs under general accounting principles.

**All questions below were published on 4/4/25*

11. Would MGHPCC allow an off-site burn-in for acceptance tests at vendor's facility?
Final payment terms will be negotiated with RFP respondents. Our expectation is that, for any physical system, final acceptance will require testing on location. Preparatory work that can streamline final testing will be helpful.

12. How can vendors comply with capital-spend requirements if its services are usually charged as OPEX?
We encourage cloud vendors to propose creative solutions that might meet capital-expenditure rules, such as certain upfront or prepaid cloud constructs. MGHPCC is open to approaches, provided they satisfy state requirements for CAPEX.

13. Is MGHPCC open to splitting the award across different vendors (e.g., one for hardware and another for cloud)?
Yes. We are open to multi-award scenarios if that arrangement delivers the best outcomes for the consortium and the research community. Proposals that provide clear lines of responsibility and simple mechanisms for accountability for delivery of respondent commitments will score higher than proposals in which responsibility is shared and/or ambiguous.

14. Will ST1, ST2, and ST3 be hosted in the same location?
It is expected that an on-prem solution would be hosted at the MGHPCC datacenter in Holyoke, MA, though various system components and later tranches may be installed in different locations within that datacenter. It is anticipated that cloud or hybrid cloud/on-prem proposals may span locations outside of the MGHPCC datacenter

**All questions below were published on 4/8/25*

15. What are you looking for (tech, capability, services) that's different than what you currently have at your institutions?
More GPUs, specifically for AI rather than pure HPC. We want a readily accessible resource for the state—AI-specific hardware that facilitates collaboration and easy scaling for machine learning research.

16. If usage grows faster than expected, might you expedite expansions, or do you plan to hold strictly to the 18-month cycle?
The plan is an 18-month upgrade cycle. If usage or funding changes significantly, we may accelerate. However, we avoid chasing every GPU release before it’s proven stable in production settings.

17. Can we propose a partial or hybrid cloud solution? Is CAPEX strictly on on-prem hardware?
You are welcome to include cloud as an option / partial-option if you feel that it would make your proposal stronger. Please refer to CapX constraints in the FAQ when considering cloud-based options as part of your solution.

18. In addition to tracking, is cloud resource capping useful?
Yes. Automated alerting (for example, at 80% usage) and possible hard caps on consumption are both considered beneficial.

19. Should we think of this system as a sandbox for short-term use only?
We will not be supporting operational systems; when/if projects reach that state we expect them to move off the platform. Note, however, that a researcher project might run over multiple years and AICR might support a researcher over that period.

20. Would we create something separate for commercial-focused use cases (e.g. startup) and research ones?
No.

21. With cloud or hybrid there will be a WAN/direct-connect aspect, should we include that in the proposal?
Yes.

22. Is there any desire for the cloud to act as an extension of the on-prem data center?
In hybrid or fully cloud-based solutions, the ability to scale or supplement on-prem resources with cloud compute or storage should be described.

23. Is there a plan for multiple data catalogs or a single, global catalog?
Data cataloging is not the immediate focus. If a proposal includes unique data-management or cataloging features—especially for a hybrid or cloud architecture—MGHPCC welcomes such capabilities, but they are secondary in S1.

24. Do you anticipate needing to restrict access or have certain resources reserved for specific users?
Yes. Institutions and user groups will require allocation controls, quota management, and potentially different priority levels for resource usage.

25. Division of this resource between stakeholders: how are the institutions going to work together on this?
We rely on MGHPCC's governance model, which has proven effective over many years. Our goal is to maintain a simple, transparent, fair, and predictable resource-sharing structure.

26. Do you see only one path for ST2 and ST3 or could you move the timeline forward to accommodate quicker refresh?
Approximately 18 months is our planned cycle, but if the industry or user demand suggests otherwise, we may adjust. We prioritize stable, production-ready hardware rather than adopting brand-new GPUs at day one.

**All questions below were published on 4/11/25*

27. How should the solution integrate with external resources, such as commercial cloud platforms or existing HPC clusters?
Initially, the system should function independently. However, MGHPCC/IOC anticipates that users at different institutions may wish to leverage existing infrastructure or potentially connect with commercial clouds. The solution should be designed to accommodate resource sharing or federation over time, without requiring it from the outset.

28. Is the mention of specific technologies like "Star", or others in the RFP, a mandatory requirement?
Unless explicitly stated otherwise, references to particular technologies are illustrative and not prescriptive. MGHPCC/IOC is open to alternative solutions that fulfill the same functional requirements (e.g., storage, orchestration, data transfer), as long as they are proven, stable, and beneficial to end users.

29. Would MGHPCC be interested in a vendor-provided training component as part of the proposal?
Yes, MGHPCC values structured training opportunities—both for novice users and more experienced researchers seeking advanced optimization. This may include modular offerings like NVIDIA certification programs, vendor-led webinars, or targeted workshops. A clear plan for training is considered a strong asset.

30. Does MGHPCC envision the platform being heavily integrated with a broader AI or data-sharing ecosystem in Massachusetts?
Yes. The HPC/AI system is one part of the Commonwealth's broader efforts to foster AI growth, which might include a future "Data Commons" or interlinked HPC resources. MGHPCC wants a foundation that can be adapted or extended to collaborate with these additional components over time.

31. Are there any concerns about balancing advanced features with system reliability?
Absolutely. MGHPCC/IOC stresses "production readiness" above all. Any advanced or emerging feature (e.g., Kubernetes-based HPC) should be thoroughly tested, stable, and non-disruptive to core SLURM workflows. Ensuring a positive user experience from day one is crucial.

32. What are some of the key metrics that you are looking to judge the program on tech, impact, what's your measurement for success?

Success will be measured in meaningful research outputs and industry engagements by our institutions and the MA AI Hub. Secondary to this is delivering a very stable and capable AICR offering to the community allowing them to focus on their research and industry aims.

34. If you are looking to optimize time-to-result or cost?
We are working on a principle of "constrain, optimize, and accept." We have constraints on funding, which impact how we most effectively optimize for performance and we accept that we will not have the scale of capability of commercial AI providers. This allows our community to effectively leverage AI resources in a cost constrained environment.

35. The RFP implied primary access via OOD and ssh to login nodes to use SLURM, can you confirm?
Yes. The majority of our users are working in this context today and we expect that AICR will need to support this capability. This does not preclude proposals from highlighting additional workflow solutions (containers, cloud, front ends, etc) as we expect our community to increasingly adopt these alternatives in the future.

36. What is your thinking on how to scale your user base while mitigating misuse of resources?
Collectively we have significant experience managing use of resources on our existing HPC clusters and expect we will need to do the same with AICR. We are open to proposals that highlight tooling and solutions that aid in this resource management.

37. Would you be open to us proposing network and volume isolation as part of our solution?
Yes, but please articulate how this may impact performance, cost and other operational factors.

38. Do you expect growth in the number of institutions utilizing this resource in the future?
Yes we expect additional partner institutions to utilize AICR in the future.

39. Are there data center limitations that may affect future ST upgrades?
Money is the main limiter here. There is space and power available to the facility to grow AICR in the future.

40. Does MGHPCC have any preferences on HPC software stacks (e.g., bare metal vs. container-based deployments, or specific orchestrators)?
MGHPCC does not mandate a single orchestration solution, so long as standard HPC batch scheduling is possible. Vendors should propose the most appropriate software environment that supports HPC, AI/ML, and the potential mix of academic and commercial users.

41. Looking forward, do you anticipate one cluster that expands at each stage or multiple new clusters at each stage?
Expand. Note that ST2 and later anticipate supporting regulated data and system support for that could be satisfied by a physically or logically "fenced" implementation.

42. The RFP seems to imply a "white glove installation", can you confirm?
Yes but there needs to be sufficient hand off for operations in terms of documentation and/or training.

43. Do you want us to manage third-party contracts? How about spare parts?
Yes, we expect coordinated management of service contracts when possible and are open to proposals that include spare parts depots.

44. How should vendors approach expansions in future phases (ST 2, ST 3), especially concerning new GPU or network technologies?
We expect next-generation GPUs and faster network fabrics will become relevant in future ST purchases. These tranches may be different enough to require separate networking fabrics. Proposals should outline a clear path to incorporate these upgrades while preserving common storage and user workflows.

**All questions below were published on 4/15/25*

45. Will you accept proposal responses that only address parts of the RFP? For example, addressing only the storage requirements of the RFP.
We expect strong responses to cover all aspects of the systems (including storage, compute, networking, etc.). We recognize that collaborative multi-partner teams may be needed to put together a response that covers all aspects of the systems.

46. Going forward will schools still build their own clusters or is it assumed they will only leverage AICR?
Assume we will continue as we are now. Institutions may augment the capabilities of AICR in the future as the project matures and if it proves successful.

47. What is the time preference for SLAs next business day, 4hr, 2hr?
Show us the costs. Currently, our institutions employ a mix of SLAs (large amounts of NBD with critical core equipment leveraging 4hr).

48. Can we use links within the proposal?
Yes, but please be to the point there will be a lot to read. Please use the format we have provided for submissions.

Instructions

**All questions below were published on 3/28/25*

1. How do we schedule a one hour meeting?
Email mghpcc-ioc-inquiries@mghpcc.org and request.

2. What is the anticipated response time for the committee's answers to vendor questions?
We do not have a firm commitment, but are aiming to post updated responses to questions twice weekly (usually Tuesdays and Fridays) until the RFP closing date. Updated responses can be found at https://www.mghpcc.org/ai-compute-resource-system/

3. The Data center FAQ states that vendors must conduct a site visit before finalizing any system design. When should the site visit occur? Is this something to be scheduled following the RFP evaluation?
Email mghpcc-ioc-inquiries@mghpcc.org and request. It is not expected this will be necessary prior to submitting an RFP response.

4. Could a vendor (virtual or physical) respond to the RFP and RFI with a single response?
Our expectation is that vendors who chose to respond to both the RFP and RFI will submit separate responses, one for each request. If the RFI response depends on receiving an award for the RFP, it should be clearly stated in your response.

**All questions below were published on 4/4/25*

5. How should a respondent present any virtual components in a bill of materials format?
A bill of materials is envisioned to provide a clear definition of what is being purchased, what components make up the system and how those components map to the aggregate performance goals. It should contain enough information to allow reviewers

to understand how performance claims are being met. It does not have to consist of itemized physical piece part lists. For virtual components an architectural schematic specifying instance types and flavors, storage services and capacities, network options and more is expected. All the distinct items that a proposed solution contains and that go to make up the total system costs should be listed out in a clear format.

6. Please clarify where the RFP makes clear that cloud providers are eligible to submit responses.
The RFP preamble states "The RFP allows for responses that span exclusively on-premise system solutions to exclusively virtual cloud solutions."

7. Will responses that do not meet all the target performance metrics be considered?
As noted in the RFP (1) "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". (2) Respondents are free to propose acceptance tests that align with whatever target performance their proposal will exhibit. (3) 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. None of these statements requires that a proposal meets or exceeds all the performance targets. However, a proposal that closely satisfies most of them will likely be preferred to one that has major components missing.

8. How rigid is the RFP's requirement for five years of hardware/software support coverage in a cloud context?
It is expected that whichever solution is chosen will be fully operational and supported over a five-year period. For cloud-based services, the emphasis is on ensuring consistent availability and usability throughout the term.

9. How can vendors ensure continuous availability of on-demand cloud services for researchers over five years?
We expect any proposal (hardware or cloud or hybrid) to detail how the service remains functional, supported, and meets user needs over the agreed term.

10. Is there a formal evaluation or scoring rubric?
The evaluation criteria are described in the RFP. We recommend that responders read the section titled "Evaluation Criteria". Those criteria will form the initial basis of any scoring. Final decisions will be subject to contract negotiations with chosen responders. The RFP lays out very clearly sets of criteria that will be examined and used in ranking to help the decision process. Ranking based on evaluation criteria will be one

component of selection. We also expect to confer with relevant members of the customer community and consider their input along with considering factors around overall perception of the quality of the response and the proposing team.

11. Do you have a strict target budget for Tranche 1 or the entire five-year plan?
We aim to stay within an overall $10–$20 million range and also ensure future tranches can be funded without using all resources upfront. If a compelling proposal exceeds typical allocations but demonstrates value, we may still consider it.

**All questions below were published on 4/8/25*

12. Will there be an extension to the April 17th deadline?
The core project team has discussed this and decided that we will keep the deadline for responses of April 17. We recognize that this is an ambitious goal and appreciate responders' efforts in helping us reach that goal.
**Please note we have amended this response to add the following-
To help responders produce higher-quality submissions, we will allow the responders to amend their response one-time by May 1st. Please note that you must still submit a reasonable proposal by April 17th for you to leverage the opportunity to amend it by May 1st. The core project team will consider the amendment during evaluation of different responses, but cannot guarantee full consideration of the amendment. For a favorable evaluation, the responders are highly encouraged to submit a strong proposal by 11:59pm EST April 17th.

**All questions below were published on 4/11/25*

13. UPDATE- Will there be an extension to the April 17th deadline?
The core project team has discussed this and decided that we will keep the deadline for responses of April 17. We recognize that this is an ambitious goal and appreciate responders' efforts. To help responders produce higher-quality submissions, we will allow the responders to amend their response one-time by May 1st. Please note that you must still submit a reasonable proposal by April 17th for you to leverage the opportunity to amend it by May 1st. The core project team will consider the amendment during evaluation of different responses, but cannot guarantee full consideration of the amendment. For a favorable evaluation, the responders are highly encouraged to submit a strong proposal by 11:59pm EST on April 17th.

All questions below were published on 4/15/25

14. Will there be a network infrastructure component released, or should we respond within this one?
We are not planning a separate internal networking RFP. It is envisioned that the internal system network(s) will be part of the current AI Compute Resource Infrastructure System RFP responses. The system will interface with external systems and we will be reaching out separately to vendors in relation to purchases for that component.

15. Will you look at a network infrastructure only response?
Not for the internal system network(s). We expect the internal networks, their integration and their hardware and software support costs to be included in systems RFP responses.

All questions below were published on 4/16/25

16. What time are RFP submissions due on April 17th 2025?
RFP submissions are due by 11:59pm EST on Thursday April 17th 2025.
**Please note the RFP/RFI submissions have the same deadline.

All questions below were published on 4/17/25

17. What time are RFP/RFI submissions due on April 17th 2025?
RFP/RFI submissions are due by 11:59pm EST on Thursday April 17th 2025.

Legal

All questions below were published on 3/28/25

1. Is MGHPCC the buy entity?
Yes.

2. What form of agreement will be used, and will there be an opportunity to negotiate terms?
MGHPCC will apply its standard purchasing terms, with modifications that reflect the nature of the purchase. Contract terms will be made available after proposals have been considered and either a short list or a final awardee has been selected. Reasonable requests for modification of terms will be considered for negotiation.

MGHPCC is a private entity. A state government contract vehicle or a vendor's standard terms of sale will not be accepted.

3. What are the MGHPCC procurement, security and confidentiality policies?
All bid discussions are private correspondence and neither party is expected to share material openly. Procurement decisions will be evaluated on factors set by the MGHPCC team. These include the factors described in the RFP. The MGHPCC reserves the right to reject or accept any bid and/or accept no bids.

**All questions below were published on 4/1/25*

4. Do responding vendors need to be on the state contract IPT72 or any other state contract?
No. As noted previously, MGHPCC is a private entity and state contracts are not relevant.

Networking

**All questions below were published on 4/1/25*

1. What are the networking requirements to the storage? Infiniband NDR or Ethernet 100G? Or a Mix?
It is up to proposers to design an internal high-speed network and present performance and cost-effectiveness characteristics. We do not have a specific requirement for a particular networking technology.

**All questions below were published on 4/4/25*

2. Does MGHPCC require integration with existing university infrastructures (e.g., identity/access management)?
We want a solution that simplifies researcher access and can integrate with or complement institutional authentication systems.

**All questions below were published on 4/8/25*

3. Is Infiniband required, or would you be open to commodity Ethernet?
Infiniband is not a requirement; we are open to Ethernet solutions.

**All questions below were published on 4/15/25*

4. Is there a preference for having a core chassis vs leaf/spine (Top of Rack) networks?
Show us the costs. Top of Rack is probably preferable for reliability reasons but understand may be more expensive.

5. Would there be interest in vendor-supplied networking tools for the project?
Yes. Please feel free to include them in your proposal and detail their costs/capabilities.

Operations

**All questions below were published on 3/28/25*

1. Is MGHPCC requesting a full deployment and a fully managed service?
That is one option. Proposers are also free to propose staff augmentation or simply propose to provide hardware and initial install/testing.

2. Should a vendor expect to provide their own physical tools?
We expect a vendor to be able to provide any physical tools they need. MGHPCC may be able to help, but that should not be assumed. Physical hardware vendor proposals should include physical installation of the proposed solution. This should include all personnel and equipment necessary to affect the installation.

3. Is the "Service Delivery" set of items a complete list.
No. It is an example of the expected service delivery tooling. Other options that meet the base suggested in the "Service Delivery" will be considered.

4. Does MGHPCC have an existing IT Service Management stack?
No.

5. Can you clarify what "Model APIs" are referring to under "Services & Applications" on Page 4 of the RFI?
Driving models as services. e.g. operating custom OLLama (for example) and other such services.

6. Can you clarify what "domain-specific, scientific facilitation" is referring to under "Service Delivery" on Page 5 of the RFI? Please cite examples of "facilitations" that are both desired and undesired in this response.
Desired facilitation could be simple help desk services to assist with on-boarding and basic getting started questions/support. Facilitation we are not expecting to support through the RFI would be more advanced expertise in specific areas like machine

learning applied to materials discovery, or advanced development of new agentic AI algorithms.

**All questions below were published on 4/1/25*

7. The RFI is very open ended. Can you provide additional guidance on expectations?
The RFI is meant to convey what you as a vendor can bring to the table in terms of operations. You can expect we will have a management framework in place that will make policy decisions for AICR. In addition, there will be some number of core team members. The RFI does indicate that respondents should address their current capabilities in Systems Software and System Operations at a minimum and comment on experience in Service Delivery and Applications if applicable.

8. Are the workflow management/orchestration services mentioned in the RFP prescriptive?
For batch based workflows we expect SLURM as a core requirement. We are, however, open to additional orchestration solutions for more persistent workflows outside of Nomad/Portainer. We are interested to hear ideas of additional technologies that can help optimize efficient usage.

9. Can we expect that there will be existing external services (eg DNS) that can be leveraged by a proposed solution?
Yes core external services will be available that the proposed solution can leverage.

10. What are your requirements in terms of document management?
We need to develop a document management systems and are open to RFI response suggestions.

**All questions below were published on 4/4/25*

12. Can you confirm that there are no existing services that we would be required to federate with?
Federation with existing compute clusters is not required.

**All questions below were published on 4/8/25*

13. Should software licenses or purchases be included in the RFI response. For example a slurm maintenance contract, bright cluster management license or service delivery hosted service zendesk or confluence?

This may depend on the RFP solution that is selected. Including options in an RFI response would be an acceptable way to address this. Once we have responses we expect to engage with respondents with strong responses to align/refine details as part of any final selection.

14. In Section 3 (Responses), Subsection D (Response Evaluation) of the RFI, can you please clarify "respondent's demonstrated credibility in terms of quality of answers to response elements labeled A through I as described above"? We are assuming you would want the responses laid out in corresponding A-Z fashion, but we only see A-D (section2) or A-E (section 3); so it is not clear what the response elements should be.

We apologize for the confusion, this is an error. The Response Evaluation section (3.D) should read:

“The evaluation process is expected to involve review of the response for

- cost competitiveness
- respondent's demonstrated credibility in terms of the quality of answers to elements numbered 1-4 in section C, "Response Elements" above.”

15. Is the plan for the AICR project to be under the mghpcc.org email domain? Can email originating from within the project use MGHPCC outgoing email resources?

Respondents can assume that the mghpcc.org email domain will be used and that MGHPCC will make outgoing email resources available for sending AICR-originated email such as maintenance announcements.

16. Is there a desire to track usage to ensure fair allocation for users?

We anticipate needing to account for usage of the platform. Usage tracking is critical for budgetary oversight, grant reporting, and equitable resource distribution. Respondents with differentiating capabilities in this are are encouraged to include those capabilities in their response.

17. From a user-experience standpoint, how do you want users to interact with the resource (SSO, SSH, etc.)?

Minimal impact on existing workflows is key. SSH and command-line interfaces must be available for advanced users, while graphical and notebook-based environments (e.g., Jupyter, Open OnDemand) are necessary for broader accessibility.

**All questions below were published on 4/11/25*

18. Is there a desire to explore a Kubernetes deployment, and what are your thoughts on SLURM within Kubernetes or Kubernetes within SLURM?

MGHPCC/IOC welcomes modern container orchestration but insists that SLURM remain the core scheduler. SLURM is heavily used across institutions, and the user base expects it for AI/ML workloads. Any Kubernetes component must be optional and must not jeopardize the stability or simplicity of the platform.

19. What is the ratio of users to L1/L2 support personnel? Is this based on supporting GPU clusters?
Modern tooling in the Research Computing space has led to fewer L1 support requests. This has enabled us to focus on L2 support that focuses on enabling and optimizing specific workloads for the community. This has been our experience supporting both CPU as well as GPU clusters.

20. What's the structure of the AICR staffing?
We have 4 positions identified on the AICR side today, including an Executive Director, Full Stack Developer, Service Delivery Manager, and System Engineer. The remainder of the staffing is TBD and will be influenced by vendor proposals.

21. For the control plane - would you want monitoring and/or ability to limit resource utilization?
Yes.

Security

**All questions below were published on 3/28/25*

1. What is the scope of NIST800-NNN requirements?
In ST1 there is no specific NIST800-NNN requirement. Later phases will include some portion that is validated against NIST800-NNN (-171, -53) controls.

**All questions below were published on 4/4/25*

2. Which compliance standards do you intend to support in the near term?
We will initially focus on unregulated data. This is clearly stated in the RFP and we encourage respondents to read the RFP carefully. We expect to start to address HIPAA and NIST 800-171 to handle sensitive research data after the initial tranche is in reasonable production operations. We are considering the possibility of expanding to more rigorous standards such as NIST 800-53 as the project matures. We will also pay attention to NIST 800-223 and other standards depending on regulatory guidance.

**All questions below were published on 4/11/25*

3. Is MGHPCC concerned about security or data segmentation for multiple institutions in a shared environment?
We expect standard HPC user isolation mechanisms—such as job scheduling controls, and best practices / documentation. We are open to vendor recommendations on more advanced multi-tenant architectures and have an expectation for more isolated environments to support constrained workloads in the future.

Storage

**All questions below were published on 3/28/25*

1. MPI is mentioned as an example of a supported library in Appendix A, page 10 of the RFI document. To what extent does the High-Speed storage tier need to support MPI workloads?
Any high-speed storage tier should ideally be able to deal with an MPI job doing lots of I/O across many processes or nodes and/or be able to work with MPI I/O views on a filesystem where multiple processes access separate parts of the same file concurrently for write or read. I/O sequential consistency between processes/ranks of the sort typical of a standard HPC sub-system wil be adequate. We need the ability to run regular workloads that can be found on any scientific cluster to generate training data for AI, produce validation or other comparison work or potentially support in-line training etc.

2. Does the commodity storage tier require accessibility from outside the AICR system?
That is not required of proposers, but we are interested in optional approaches that can make it easier for customers to work in this environment some of the time and in university or other environments at other times. We are also interested to know of any integrated solution options that do support efficient ingress and egress of data. We expect to overlay some current standard solutions, for example Globus, on the platform if the equivalent functions are not part of any integrated approach.

**All questions below were published on 4/1/25*

3. We believe the IO requirement is 3 million IOPS for high-speed storage. Is this correct?
The RFP calls out 30 million IOPS. The goal is a system that maintains performance when several thousand parallel threads are performing distinct meta-data operations or small I/O operations against the storage. It is not expected that every proposal will be able to achieve

every target and reviewers will evaluate proposals in part of their capability across all the targets.

4. Is storage required to be on the high-speed network used for GPU-to-GPU communication?
Respondents can choose a design that makes sense to them. The performance of the storage in serving I/O needs of model training and inference cycles and the cost-effectiveness of solutions will factor in reviews. The specific network technologies are less of a review factor.

**All questions below were published on 4/8/25*

5. For the 30M IOPS requirement for storage, what is the workload (small vs large) you are expecting for this?
We are considering various workloads and ask for responses to provide details on the workload you benchmarked and the rationale for meeting or not meeting target benchmarks.

6. Will multiple external research organizations frequently import and export data sets?
Yes.

7. Are you expecting specific protocols for commodity storage?
We have no specific expectations but welcome vendors to propose multi-protocol designs that best serve HPC & AI data from internal and external research sources. We expect proposals to detail which protocols you are supporting.

8. How is storage expected to be structured—combined or split between a high-speed flash tier and a lower-cost, high-capacity tier?
MGHPCC welcomes different storage strategies. Solutions could present a unified system with intelligent tiering or physically separate high-speed and commodity storage as long as budget and performance targets are met.

**All questions below were published on 4/11/25*

9. Are you ok with solutions that utilize data reduction for your storage?
Yes, but explain how that solution meets requirements. In particular the RFP envisions a target for usable storage across a diverse base. Any assumptions about reduction through deduplication would need to be well justified as we have limited a priori insight into the data that will be hosted.

10. Can we bid a single-tier all flash storage solution?
Yes if you can do it within budget.

11. For storage, many MGHPCC institutions are using similar vendor-provided solutions. Is there value in our response indicating how they might integrate with our proposed solution?
If a vendor has interesting integration capabilities then we would be interested in hearing about them in the response.

12. Is it ok or preferred to have storage all in one rack?
You are free to distribute storage as it makes sense for the overall system design.

**All questions below were published on 4/15/25*

13. Do you have specific storage requirements for the Service Nodes?
Service nodes are intended to support persistent services, gateways and system operations; performance should be commensurate with supporting these types of operations. The RFP specifies 100TiB/node with an aggregate of 3PiB.

Use Cases

**All questions below were published on 3/28/25*

1. Can you help us further understand class of tiers or anticipated user demand for HPC workloads vs AI workloads?
The research community using this facility will be a mix across the full span of research at the university communities and the Commonwealth AI Hub community of startups and economic development participants. There is not a specific tier that this represents. We expect a healthy set of customers doing modest scale experimental work to optimize ideas on training and inference. We also expect customers requiring a sizable fraction of the resource for modest scaling up of experiments in both training and inference. We do not expect to use the system much for large parallel HPC runs in isolation, but there may be scenarios involving embedded training or online inference where such experiments are executed.

**All questions below were published on 4/1/25*

2. Is this system strictly for AI workloads? Will there be a strictly HPC type / traditional workloads?

This is a predominantly AI/ML system. However, activities like reinforcement learning that draw on physical solutions and/or embedded training may involve 64-bit precision algorithms.

3. What will your user volume / workloads look like on Day 1? What does the initial pilot period look like in terms of users and workloads?
We expect to ramp up the number of customer accounts over multiple months. In the initial phases we envision tens of early accounts that span the community, but that can provide feedback and iterate on final operational practices. Over the first 3-6 months we anticipate deliberately growing this number to thousands of eligible accounts.

4. Can you confirm that Tier 1 Compute will focus on a small number of users requiring large modeling and Tier 2 Compute will focus on a larger number of users with more diverse workloads and smaller bandwidth requirements?
The Tier 1 capability should be designed to allow effective use by a small number of users with large model activities. It should also allow larger numbers of users to undertake intermediate model training and inference experimentation. These workloads are envisioned to be managed by a workload manager to allow flexibility depending on community needs and priorities.

5. The RFP refers to supporting up to 1K users at one time. Is the expectation that all 1K users will be doing substantial work on the cluster?
No, the 1K user figure represents the number of users that could be logged into the system at one time but it is not expected that all 1K would be necessarily be running intensive workloads at the same time.

6. In addition to HPC stack requirements (SLURM), is there any requirement on AI stack for both training or inference. For example, a particular model deployment API (OpenAI API, Olamma Stack, etc?)
Not particularly. We expect use of many the AI tool interfaces. This might be through Hugging Face tools, through Python API keys etc... or a variety of other ways, ollama, vLLM and more. We also expect researchers to experiment with open source weight models like the DeepSeek family models, including reasoning configurations. Any system thinking ought to be general enough to support this flexibly. We have groups that support things simply through basic reservations and containers, others that are exploring more elaborate frameworks.

**All questions below were published on 4/8/25*

7. What kind of workloads will this system serve—pure batch HPC jobs or interactive AI sessions?
A mix. We envision standard HPC schedulers (Slurm) plus the ability for interactive (fraction-of-a-GPU) sessions, such as Jupyter notebooks. Large multi-node AI training or inference jobs remain a high priority, too.

8. Do you anticipate a small number of big projects or large number of small projects?
The RFP explicitly identifies two logical tiers of systems to accommodate large-scale, high-consumption workloads and smaller, more common GPU tasks.

9. Would you consider multiple smaller clusters, each specialized, instead of one unified system?
Potentially yes—heterogeneous clusters that look like a single resource to users is attractive. We can unify them via shared scheduling, storage, and identity. Some partitions might focus on data exploration, others on large training runs.

10. Would a separate infrastructure be needed for commercial vs. research use cases?
That is not anticipated. There is a lot of cross-collaboration between research entities and between research and industry startups. Having a similar platform to service all stakeholders would be beneficial to support this cross-collaboration.

**All questions below were published on 4/11/25*

11. What is the experience level of AI and ML users anticipated to use this system?
There is a broad range of experience levels. Some users are experts who already run large AI/ML workloads, while others are novices. MGHPCC/IOC expects the vendor to include training resources or outlines for end-user education, acknowledging that universities also have local facilitators who can assist.

12. Will this resource support research for business or research for publications?
We expect a mix of both.