Kempner Institute Compute Governance Guidelines

AY25

Overview

The Kempner Institute will build and maintain one of the largest academic AI clusters in the world supported by a team of research software engineers, research computing and other engineering staff with expertise in machine learning and high-performance computing. The cluster will be state of the art and structured to support a diverse portfolio of research in intelligence, with a specific focus on providing the capability to run large-scale, high-impact projects across hundreds of GPUs.

This document outlines policies governing use of and access to this equipment, access to the engineering team, contributions of hardware to the cluster, costs associated with use of the cluster, and fair share principles.

This is a working set of guidelines which will be revised as appropriate to better serve the needs of the Kempner Institute community and to help advance its scientific mission.

General principles, Open Science & Reporting

Access to the Kempner Institute cluster is based on an ongoing and direct relationship with the Kempner Institute and is exclusively intended to support mission-aligned research.

Access to the cluster includes both the ability to run jobs on the cluster, and the opportunity to work with engineers and research computing support staff employed by the Kempner Institute. Access to and use of the cluster and the engineering support will be solely to facilitate research that is directly related to the Kempner Institute's mission and strategic initiatives.

Using the cluster for research objectives that are not related to the Kempner Institute's mission may result in the loss of access to the cluster.

The Kempner Institute is committed to open, reproducible science. Any work completed on the cluster must comply with the Kempner's Open Science Policy. The policy covers sharing of data, intellectual property, software, non-software inventions, and publications. In short, work supported by the Kempner should be made freely and openly available for non-commercial, academic use, *on or before* the date of publication in a peer reviewed journal, preprint server, or other publicly available medium such as a blog or web site. This includes any code, datasets, algorithms, technology or other resultant materials. On occasion there may be constraints that prevent full public release of data, but the use of the cluster for such projects must be pre-approved.

Faculty who have access to the cluster will be required to complete a brief annual report describing the projects that were run on the cluster, and providing a list of and links to any presentations or publications which were supported by the cluster, and links to the corresponding code and data related to the project.

Access to the cluster

Access to the cluster is determined solely by the Kempner Institute. The cluster may only be used for research that is in alignment with the Kempner Institute's mission.

Access is granted to:

- Faculty co-Directors and Institute Investigators and specific members of their labs/research groups
- Associate Faculty and specific members of their labs/research groups working on Kempner-related projects
- Visiting Scientists on a per project basis
- Kempner Research Fellows
- Kempner Graduate Students for projects related to the Kempner mission*
- Kempner Undergraduate Research Programs (KURE and KRANIUM) for projects related to the Kempner mission*

Kempner Institute Investigators and Associate Faculty may request access for their lab staff each academic year by completing a brief application found here and by completing an annual report on use of the cluster as described above. Access to the cluster for external collaborators who are sponsored by any of the above faculty will be reviewed and granted on a case-by-case basis only by the Kempner Cluster Committee.

In response to a Kempner RFP, the Kempner Institute may grant other Harvard faculty access to the cluster and engineering team to advance the specific work of the approved project. This access will be limited in time, resources and access as defined in the approved proposal. In rare circumstances, the Kempner Cluster Committee may decide to give project-limited access to affiliates who engage in mission-related research which advances collaborative research projects with Kempner Institute Investigators, Associate Faculty and Fellows.

* *Student Access:* Kempner students have access to the cluster to support their own research and only for the duration of their appointment within the Kempner. Access to the cluster must be approved by their PI. If a Kempner student runs jobs on the cluster for others in order to circumvent the access limits described above, uses the cluster for purposes other than their own research, or uses the cluster to pursue research which is not aligned with the mission of the Kempner Institute, the student will lose Kempner cluster access and will only be able to access the cluster at the same level as unaffiliated students.

Fairshare

Use of the Kempner cluster for approved users is governed by fairshare. The algorithm prioritizes a balanced allocation of resources, aimed at facilitating the timely completion of tasks from various user groups (see FASRC Fairshare Accounting for a more detailed overview). This means that jobs, particularly those that are resource intensive or are being run in labs with high recent usage, may not run immediately or on demand. Fairshare guarantees access to resources averaged over a period of several months. At any given time, a user or lab may be using a greater or lesser amount of their assigned share.

Fairshare is allocated based on the user's affiliation category as described below. Each category receives a share of the available computational resources and then subdivides those resources to members within the category. Certain categories of users (e.g. undergraduates) may be limited in terms of the total resources that can be used at any given time, and/or may be

allowed to submit jobs with different run-time lengths. Furthermore, resources might occasionally be reduced due to maintenance or to facilitate approved Special Project reservations (described below).

Current allocations:

- 70% Institute Co-directors, Institute Investigators and those Associate Faculty who have contributed to cluster
- 6% Associate Faculty who have not contributed to the cluster
- 11% Kempner Research Fellows
- 12% Kempner Graduate Student Fellows
- 1% KURE and KRANIUM Students working in a lab without cluster access

Based on the number of laboratories that share these resources, the total fairshare allocation for the laboratories of Institute Investigators and Associate Faculty who have contributed funds to expand the cluster is roughly 3x the level of fairshare for Associate Faculty who have not contributed to the cluster.

Institute Investigators and Associate Faculty who contribute to the cluster will also receive an additional increase in the above fairshare relative to the size of their financial contribution as described in detail in the “Faculty Contribution” section below.

As the computing needs of the Kempner community and the computing power within the cluster change, fairshare allocations will continue to be reassessed by the Kempner Cluster Committee and adjusted. Changes will be announced, and adjustments made prospectively.

However, if users believe that they have insufficient access or that the wait time to run jobs is longer than it should be, we encourage them to contact Kempner and FASRC to help identify potential bottlenecks and troubleshoot solutions.

Special Project Reservations

A critical goal of the Kempner Institute is to provide computational resources, at scale, to facilitate groundbreaking research related to the Kempner’s strategic initiatives. To support this goal there may be times in which the Kempner Institute issues a targeted or open call for Special Projects which allow users to request access to the cluster, reserve a significant number of GPUs, engage with the engineering team, and exceed typical run-time limits.

Special Project Reservations are for new, time-limited research projects that are in close alignment with the Kempner mission or in direct response to a Kempner issued RFP.

Projects must be resource intensive, meaning that they require a scale of resources that are not freely available using any other computing facilities at Harvard University. This includes the HMS GPU cluster, FASRC and SEAS GPU or requeue GPU partitions, or NERC resources provided by University Research Computing.

If approved, projects must adhere to the Kempner Institute's Open Science Policy. Furthermore, once results are ready for dissemination, weights, code, and data must be distributed via the Kempner Institute’s channels (GitHub for code, Hugging Face for models and data). Finally, once the project is complete, a summary of the project suitable for dissemination should be prepared and provided to the Kempner Institute for sharing via a

blogpost or another relevant communication channel.

The GPUs reserved for a project should be used only to advance the project goals as described in the research proposal. Use of these project resources for any other purpose may result in the project reservation being canceled and use of the Kempner cluster restricted or removed.

Project reservations may be paused or scheduled outside a two-week window leading up to major conference submission deadlines to avoid creating resource constraints during times of significant need for the larger community.

Requests for reservations should be made using the request form available here and on the Kempner Institute website. Reservation requests are forwarded to and evaluated by the Kempner Cluster Committee.

A decision made by the committee is final, although the committee may, at its sole discretion, provide feedback on the request. If a request is granted, the committee may also make modifications to the requested resources (e.g. date, run length, GPUs, etc.).

Overhead, operating and data storage expenses

As is the current practice with users of other computational resources hosted by FASRC, operating expenses for cluster use will be charged as per existing FASRC processes. As dictated by MOUs at the school level, operating expenses are typically charged based on the primary academic appointment of the PI/user and the resources used. More information on this can be found on the FASRC website under "Offerings (Tiers of Service)".

At no time will the Kempner institute pay for data storage expenses on behalf of faculty, fellows or students using the cluster. Data storage expenses may be charged as a direct cost on Kempner Institute research funding for those that have such funding.

For those very few people who are solely affiliated with Harvard University via the Kempner Institute (i.e., Kempner Research Fellows, Kempner Visiting Scientists, and Kempner Institute computational staff) and who have no other academic home, the Kempner will pay for cluster operating expenses and storage as required.

Engineering & Research Computing Support

The primary function of the Kempner Research Engineering Team is to ensure that the Kempner Institute has the right resources, tools and technology to support innovative research. They may spend time developing open-source AI/ML software modules (typically in Python), refactoring the existing codebase, scaling distributed ML training on an HPC cluster, ensuring use of best software engineering practices for scientific software packages, optimizing system design for complex projects either on-premises or in the cloud, efficiently implementing ML algorithms, architecting big data solutions for AI/ML workflows, facilitating AI/ML workflows on the Kempner cluster, and optimizing AI/ML algorithms for specific GPU architectures. They will participate in teaching activities within the institute to disseminate best practices and will be available during office hours to provide expert level advice to Kempner Institute faculty, fellows and students on issues related to their areas of expertise.

The Research Engineering Team is available to support scientists who are working on Kempner Special Projects as determined by the institute leadership and approved as part of the Special Project request process. In this capacity they may: participate in the development, refactoring, troubleshooting and review of code; provide support to ensure code is reproducible and packaged in such a way as to encourage use and open-source collaboration; provide support for data engineering and management including the development of pipelines or data architectures; provide technical project management support including establishing and monitoring project milestones and roadmap; and advise on scalability, codebase management, user documentations and best practices for open science.

Affiliate Faculty will have only consultative access to the Kempner Engineering Team unless they have contributed hardware to the Kempner Institute cluster. For faculty who have contributed to the cluster, they may apply for project level support if they are able to provide salary support for Kempner Institute Engineers via grants or other sources of research funding.

Those who are unaffiliated with the Kempner Institute will not have access to the Kempner Institute Engineering Team.

Research computing staff funded by the Kempner will have three primary responsibilities: they will serve as a partner for the Kempner community to ensure the optimal use of the cluster; will help with system engineering of the Kempner cluster; and can support Kempner Special Projects. This role will, in some capacity, build HPC software tools and utilize frameworks or platforms to facilitate system engineering of the Kempner cluster, including GPU monitoring, network monitoring for distributed training, improving cluster reliability, and job compliance monitoring for fair usage. Additionally, this role will facilitate widely used workflows, especially in AI/ML on the cluster, by building optimized containerized solutions (Singularity), designing SLURM job submission scripts, and contributing to the Kempner Institute HPC Handbook.

Faculty Contribution to the Kempner Institute Cluster

Faculty investment in the Kempner computational infrastructure not only accelerates each faculty member's research but also fosters more ambitious research initiatives for the institute as a whole, helping to ensure that the Kempner remains at the forefront of academic computational capabilities.

When considering the benefit of contributing to the cluster versus pursuing an independent purchase, Faculty contributors benefit in the following ways:

1. The Kempner is buying at scale; thus, it gets significantly better pricing per GPU relative to more modest purchases typically made by individual faculty.
2. The Kempner Institute will provide for the installation, ensure adherence to the warranty, and ensure maintenance support for equipment.
3. Faculty who contribute receive access to the rest of the cluster at institute assigned fairshare levels, allowing them to occasionally exceed the computing capacity of their contributions and if any of the equipment fails or is taken down for maintenance, this still allows access to the rest of the cluster for research.
4. Faculty contributors are able to consult with Kempner employed engineers.

However, in accepting contributions to the cluster we must balance the desire to add to the total resources available to the Kempner community with the need for effective planning, efficient

installation, and compatibility with existing resources. Finally, in order to ensure that faculty contributions are able to expand the scale of purchases while maintaining or reducing per GPU costs, contributions to the cluster must be large enough to have a more than nominal impact to the overall purchase.

Thus, starting in AY25, any contribution to the cluster must follow the guidelines below. These terms must be acknowledged as part of the Contribution Request form (template below for reference).

It is possible that in any given year the Kempner may decide to limit or forgo faculty contribution to the cluster or limit or forgo compute purchases entirely. The Kempner cannot promise a specific date whereby equipment will be ordered, received, installed, and open for use.

However, reasonable effort will be made to ensure open communication about timelines so that faculty contributors can make appropriate plans for research.

Terms for Faculty Contribution to the Kempner Cluster, effective July 2025.

1. Only Kempner Faculty with cluster access (e.g. Institute Investigators, Associate Faculty) may request to contribute funds toward the annual purchase of equipment for the Kempner Institute.
2. Individual contribution of physical equipment will not be accepted. Only funds will be accepted.
3. Contributions less than $300,000 will not be accepted, however payment may be spread over 3 fiscal years.
4. Faculty must demonstrate that research to be undertaken on the cluster is in alignment with the Kempner Institute's research mission.
5. Faculty will be credited an increase in their fairshare on the Kempner cluster proportional to the funds they contributed against the total cost of the Kempner's annual purchase of equipment.
6. This increase in fairshare on the cluster is normalized by computational power. For example, a 5% financial contribution will result in a credit equal to 5% of the computational power of the resources purchased in that year.
7. As described in the Kempner Institute Compute Governance guidelines, all use of the Kempner Institute cluster is governed by fairshare. Priority access to resources, by definition, prevents other users from accessing GPUs. Thus, contribution of funds will not result in any priority access to GPUs. If dedicated access is needed faculty may make special project reservation requests as per Kempner Compute Governance guidelines.
8. The increase in credited fairshare will remain in place for as long as the equipment purchased is operational. Generally, equipment will be warrantied for 5 years and thus is expected to last 5 years. As a result, the incremental increase in credited fairshare is expected to last for that time period.
9. If the equipment is operational beyond the 5 year window, the incremental increase in fairshare will remain in force. However, if any portion of the hardware purchase fails or is otherwise decommissioned the fairshare credit will be reduced by a commensurate amount. For example, if faculty contributed 5% towards a purchase of 144 GPUs and 8 failed after warranty, faculty would have access to 5% of the remaining 136 GPUs as part of their fairshare allocation.
10. Kempner Institute has sole discretion to decide if and when any equipment will be

decommissioned and removed from the cluster at any time. Faculty contributors will be notified as soon as reasonably possible if and when this determination is made.
11. Once installed in the cluster, the Kempner Institute will ensure maintenance of the equipment.
12. If the faculty member leaves Harvard prior to the end of the useful life of the equipment, the equipment will be retained by the Kempner Institute for general cluster use and no funds will be provided in recompense.
13. The Kempner Institute will accept costs related to rack (power, cooling, non-blocking network) and row installation as is customary for Kempner Institute purchased equipment. Operating expenses related to use of the cluster will be charged to schools based on use, as is customary with all computing equipment managed by FASRC.
14. The Kempner strongly prefers that contributions be made using funds **provided by** the Kempner Institute (e.g. Kempner start-up funds, Kempner research support, Kempner associate faculty research funding).
15. Faculty who wish to contribute funds from other sources, including via grant support, should discuss this with Kempner Institute leadership. These contributions will be managed on a case-by-case basis. However, contributions from fund **external** to the Kempner Institute will only be considered if:
    - Funding does not limit the use or users of the equipment in any way.
    - Funding from a restricted source (grant, foundation, or donor funding) has the explicit approval from a department representative (e.g. Office of Sponsored Programs officer, or Finance Manager or Director, etc.) who can confirm that the use of the funds is an allowable expense per the terms of the funding entity.

Faculty Contribution Request Form (AY25) Template

Anticipated Total Purchase: $XX.XXM, XXX GPUs (XX nodes), XXX Total Fairshare Units, MM/YY anticipated installation date, YYYY anticipated end of warranty period

Name:
Department / School:
Website:

Amount to be contributed:
Source of Funding:
33-Digit Account String:

Describe research projects to be undertaken on the cluster (250 words):

Signature:
_____ *My signature confirms that I have read the terms outlined above and acknowledge that they are binding.*

RESTRICTED FUNDING ONLY:

Name of approver:

Signature:
_____ *My signature confirms that I have read the terms outlined above and that this is an allowable use of funds as per the funding entity guidelines or requirements*

AY24 drafted: 5/19/2023, Approved: 6/9/2023
AY25 drafted: 5/17/2024, Approved: 6/14/2024