HAICORE ramp-up policy

Version 1.0

Approved by the HAICORE access board on 27 Nov 2020.

Rationale and introduction

Access to computing resources is key for the Helmholtz AI platform and the wider Helmholtz AI community to accelerate innovative AI applications. Therefore, the "Helmholtz AI computing resources" (HAICORE) was initiated at the end of 2019 as a one-off measure within the framework of the Helmholtz Incubator. HAICORE comprises HGF funding of 2.5 million euros, mainly for GPU hardware, which is implemented at the Forschungszentrum Jülich (FZJ) and Karlsruhe Institute of Technology (KIT). The resources should be open to the entire AI community within the Helmholtz Association.

This document sets out the draft for the science-driven ramp-up policy of the HAICORE computing resources and will be adapted/expanded over the next few months. Overarching themes of the policy are openness, fairness, a low threshold and a community-based focus. To ensure that these points are addressed, the HAICORE access board was established and is composed of following guests and deputies: Norbert Attig (FZJ), Daniel Mallmann (FZJ), Stefan Kesselheim (FZJ), Volker Gülzow (DESY/HIFIS), Uwe Jandt (DESY/HIFIS), Ants Finke (HZB/HIFIS), Philipp Heuser (DESY/HIP), Sara Krause-Solberg (DESY/HIP), Jennifer Buchmüller (KIT) and Markus Götz (KIT). In order to ensure that the board acts on behalf of the Helmholtz AI platform, the Helmholtz AI steering board nominated the following HAICORE access board members: Guido Juckeland (HZDR), Oliver Stegle (DKFZ), Frederik Tilmann (GFZ), and Xiaoxiang Zhu (DLR)[1].

Representatives of the operating centers at FZJ and KIT will report regularly (monthly) on usage statistics to the HAICORE access board members. The Helmholtz AI steering board will be updated on the board's decisions by the members nominated by it. An annual report of HAICORE will be generated by the operating centers and a short summary included in the annual report of the Helmholtz AI platform.

The HAICORE resources fill a much needed gap in access to compute resources for scientists between local systems and tier 1-3 compute resources. They provide easy and low-barrier access outside the typical long-term compute project allocation. As such they will allow ad-hoc usage (using Helmholtz AAI infrastructure eventually) as well as a streamlined and fast-turnaround HAICORE proposal process. In order to enable such easy access, the delivery of the computing services is achieved through web-portals and traditional ssh login. If a proposal exceeds 5,000 GPU hours or a duration of one year, they will be redirected to the already well established tier 1/2 project allocation process.

[1] F Tilmann and X. Zhu will substitute for each other in attendance of meetings.

Ramp-Up Policy

The ramp-up policy is an immediate and short term solution to enable a swift start and wide use of the resources available through HAICORE: everyone within the entire Helmholtz AI community (e.g. Helmholtz AI researchers and AI consultants, researchers within Helmholtz AI projects or voucher and researchers within any other incubator platform, funding line, and AI-related project) has access to the HAICORE resources and the job submission queue. The queue is not prioritised, i.e. HAICORE operates a first come, first serve model.

After a 6 months ramp-up period or earlier if usage statistics demonstrate a need to adapt, the usage and exploitation of the HAICORE resources will be evaluated to best determine an access policy that serves community needs. To achieve this, the access policy board will review existing practices to determine the way ahead. This is done in consultation with the HAICORE infrastructure providers.

As we might experience a high and growing demand in the near future, e.g. during the ramp-up period, the queue is likely to run full. In the ramp up phase, all users will be treated equally to better understand demands and inform a potential future usage model.

All policies and regulations placed on the HAICORE resources be evaluated by the HAICORE access board regularly (at least twice a year) even after the ramp-up phase.

Preliminary HAICORE usage regulation

The respective operating rules of FZJ and KIT apply; particularly with regard to BAFA/EC regulations preventing users of so-called rogue nations from accessing high performance/HAICORE computing resources unless there is a supervisor or superior that guarantees for them (up to five years of prison for violation).

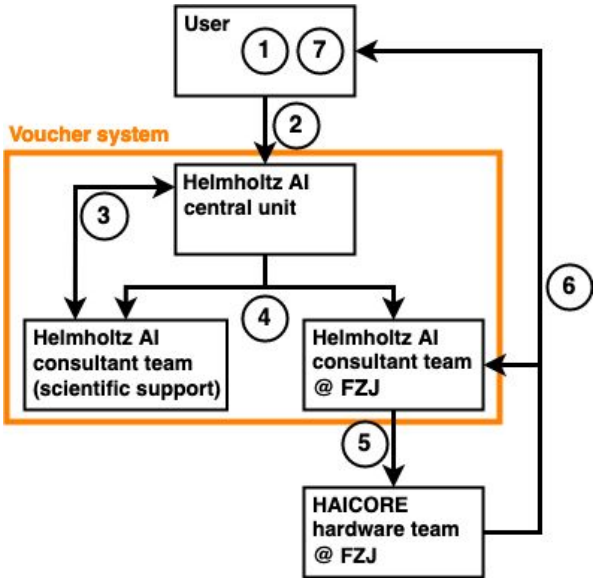
I. For ad-hoc usage (KIT is operating centre)

1. User identifies a need for computing time on HAICORE resources and navigates to a Helmholtz AI hosted website (for ramp-up, the voucher system website).
2. A link directs the user to the web-portal of the operating center (KIT).
3. Sign-up, user account creation and log-in:
    a. User submits Request for Access form to KIT
    b. User signs a Usage Agreement with KIT.
        i. This needs to be done either with a proper signature or a digital signature via a Helmholtz center issued CA certificate.
        ii. Optionally, the user needs their line manager or supervisor to declare that they vouch for BAFA-compliant use.

   c. User accounts are being created, access details sent to user and user logs in.
4. The user starts using the HAICORE resources (up to 10 GPU hours a day).

KIT will sign up to five new users per week. For ad-hoc usage technical support will be provided as a 'best effort' approach (for general access trouble by the hosting center, once the job is running by Helmholtz AI, e.g. through the voucher system using a simple helpdesk model or a mailing list). To facilitate usage and lower the need for support, tutorials (comprising e.g. a mix of videos, wiki and FAQ) will be produced across the AI consultant teams and provided for potential HAICORE users.

II. For lightweight projects up to 5,000 GPU hours (FZJ is operating centre)



1. User identifies a need for computing time on HAICORE resources, navigates to Helmholtz AI hosted website (for ramp-up, the voucher system website) and formulates a computing time proposal including the amount of resources (GPU hours) and software requirements.

2. User submits the computing time proposal as a voucher request in the Helmholtz AI voucher submission system (new voucher category to be defined).

3. Helmholtz AI central evaluates the voucher and liaises with the responsible Helmholtz AI consultant team (research field of applicant) to decide if the proposal is well-justified, i.e. i) if it is an AI-related project and ii) if the GPU hours required are within reason.

4. If the proposal is well-justified, Helmholtz AI central sends the voucher proposal to the Helmholtz AI consultant team of FZJ and the consultant of the applicants' research area.

   a. The user/s need/s to sign a Usage Agreement with FZJ, the project head/s (Principle Investigator, PI) additionally a PI Agreement.

   b. The latter one requests a proper signature or a digital signature via a Helmholtz center issued CA certificate, the Usage Agreement a digital acknowledgement of the user only.

5. Local action: Helmholtz AI consultant team at the resource provider (FZJ) informs the HAICORE hardware team onsite about the project, associated users and resource assignment.
   a. User accounts are being created, compute time projects (FZJ) are created and the home identity provider (IdP) of the user's center is informed about setting the respective entitlement in the IdP/HAAI tokens.
6. HAICORE hardware team sends information about granted access to the requested HAICORE resources to the user and local Helmholtz AI consultant team
7. The user logs into the system using the center's credentials/2FA (access via ssh or JupyterHub).

For scientific support, e.g. through Helmholtz AI consultants, the user should submit a 'realisation voucher' in the voucher system. For lightweight projects only basic technical support should be provided by the operating centers.

III. For larger projects (> 5,000 GPU hours)

1. User identifies a need for computing time of more than 5000 GPU hours
2. User follows the tier 1/2 project allocation process outside of HAICOREs limit, i.e. regular compute time proposals at one of the national HPC centers

IV. Technical considerations

- Storage is not part of HAICORE and is optionally sponsored by the hosting sites. It is limited to non-temporary 1 TB/user at KIT (up to 240 days) and 16 TB/project FZJ.
- The maximum number of users that can use the resources concurrently is limited to 72 (due to the number of GPUs) at KIT and unlimited in slots at FZJ (but compute time limited).
- Helmholtz AAI is not yet supported or integrated with the FZJ or KIT.
- Additional one-time passwords, i.e. second factor, is necessary at KIT. For this, the Google Authenticator app is used. A two-factor authentication at FZJ is planned but not yet in place.
- Computing on KIT's machines requires an entitlement in the login token. This needs to be set by the IDP provider at the user's Helmholtz center once after account creation or in the future by HAAI when supported.
- At FZJ, a local login (JuDoor) must be created preliminarily, until HAAI supported groups (virtual organizations) can be mapped to compute projects.

Proposal for acknowledgement

For any publication resulting from projects calculated on HAICORE resources, Helmholtz AI and HAICORE should be mentioned in the acknowledgements as follows:

> This work was supported by Helmholtz AI computing resources (HAICORE) of the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI.