First Gefion competition: Call for proposals that explore large-scale GPU computing

For the first time, the Danish Centre for AI Innovation (DCAI) is opening access to its Gefion AI supercomputer – Denmark's most powerful AI resource, powered by NVIDIA technology. During this pilot phase, we are inviting both academic and business users to submit their most ambitious AI projects for a chance to gain early access to this state-of-the-art AI supercomputer.

As part of Gefion's pilot phase, we are seeking groundbreaking AI use cases that can fully harness the potential of our new AI supercomputer. Users are invited to propose and execute ambitious large-scale GPU computational workloads on Gefion that currently cannot be launched anywhere else in Denmark. Selected participants will receive free access for up to 3 months to run their workloads, with the goal of testing and pushing the limits of Gefion's capabilities across a range of innovative applications. The chosen participants will get early access in the initial pilot phase of Gefion (tentatively planned for mid Q4-2024 to Q1-2025). We expect to select about 5 use cases as part of this competition.

Hardware, software, and restrictions in the pilot phase

The hardware architecture of Gefion is based on NVIDIA's DGX SuperPOD AI data centre and contains 191 DGX H100 nodes. Each DGX node has 8 NVIDIA H100 tensor core GPUs with Hopper-based architecture. Each GPU has 80 GB memory, implying 640 GB total GPU memory per DGX node. Each DGX node acts as a general-purpose HPC server, with 2 Intel Xeon 8480 Sapphire Rapids Platinum CPUs with 2 TB DDR5 memory, totalling 112 cores at 2.0 GHz. The DGX nodes are connected with NVIDIA Quantum-2 Infiniband architecture, and 4th generation NVLink takes care of intra-node GPU communication. Jobs to Gefion will be handled by the SLURM scheduling system. Gefion uses NVIDIA's base command manager and will have NVIDIA's Enterprise software stack installed.

During this initial pilot, it will only be possible to execute jobs containing non-sensitive data. Gefion will be able to handle use cases with sensitive data at a later date. Gefion is intended to be used for research and innovation purposes, including for training AI models, executing scientific calculations, and developing new AI applications. Proposals to use Gefion for running commercial production AI services will not be approved.

Requirements for submitting a proposal

For this pilot DCAI is especially interested in identifying use cases that highlight Danish-based research and/or innovation activities. International partners are welcome to join a consortium submitting a proposal as long as the main activity is anchored in Denmark.

Submitted use cases can stem from a wide range of topics, as long as they benefit from being accelerated on a GPU-based supercomputer like Gefion. Examples include, but are not limited to, training of AI models, digital twins, AI and computational models in areas such as life sciences, physics, chemistry, and climate, AI handling of very large data sets from large scientific experiments, etc.

The team submitting the proposal must be ready to execute the code in a short time and with minimal support. This is due to the limited resources of DCAI in this initial phase. This implies having a well-tested code that can be executed on NVIDIA's DGX H100 architecture, that uses the NVIDIA Enterprise software installed on Gefion, and that has proven to scale well on larger systems with multiple,

connected GPUs. Access to any additional resources or expertise cannot be guaranteed as part of this competition.

Process for submitting and reviewing use cases

Submissions of proposals for this competition must use a standard template, which can be requested by contacting Morten Bache (mba@novo.dk), Scientific Director in the Novo Nordisk Foundation and part of the team supporting the Gefion pilot. To qualify for this first competition, please **complete the proposal template and return it before Friday 20 September 14.00 CEST**.

The proposals will be reviewed by a committee chaired by CEO of DCAI, Nadia Carlsten. It will be the committee's objective to identify a broad range of pilot use cases, with emphasis on supporting a wide range of users and organizations. Winning proposals are expected to be selected and announced in October 2024. Workloads based on selected proposals are expected to run from mid Q4-2024 to Q1-2025.

Interested users that do not wish to submit proposals to take part in this competition but are interested in receiving additional news and information about Gefion and DCAI should send an email to gefion-interest@novo.dk

What information should a proposal contain?

Summary: Explain the use case in a few sharp sentences.

Technical impact: Explain the impact the use case will have for you (without revealing critical IP) and/or the AI community.

Benefit to the Gefion pilot: Explain why your use case should be prioritized in this pilot phase, including what Gefion staff could learn from running this use case that will help mature operations and benefit future users.

Impact on the Danish ecosystem: Explain the team's ties to Denmark and how this use case benefits the ecosystem of innovation in Denmark.

Team: Explain who is on the team and what their skills are relative to executing AI and scientific computational jobs on a large-scale GPU architecture like Gefion.

Scalability of the project: Explain why the problem is scalable to larger numbers of GPUs and thereby is meaningful to execute on a large-GPU architecture like Gefion.

Code and data readiness: Demonstrate to what degree the code (and potentially the data needed for deploying the workload) has been optimised and prepared to be ready for execution on Gefion with minimal delay. If there are any specific NVIDIA software libraries or other dependencies that are needed to execute the code, please elaborate.

Predicted wall-time and workload: Provide an estimate of how many resources (GPUs, CPUs, memory, disk space, iterations) the use case might need. Please also provide information on whether the Intel DGX server CPUs are going to be part of the simulation workload, and if so, whether this has been tested.

Support from DCAI: Let us know whether you request some kind of support from the local team at DCAI or code-optimisation/sparring on NVIDIA software before submitting the job to the queue. Note that this kind of support will only be minimal and not guaranteed.