

EURO **Greece**

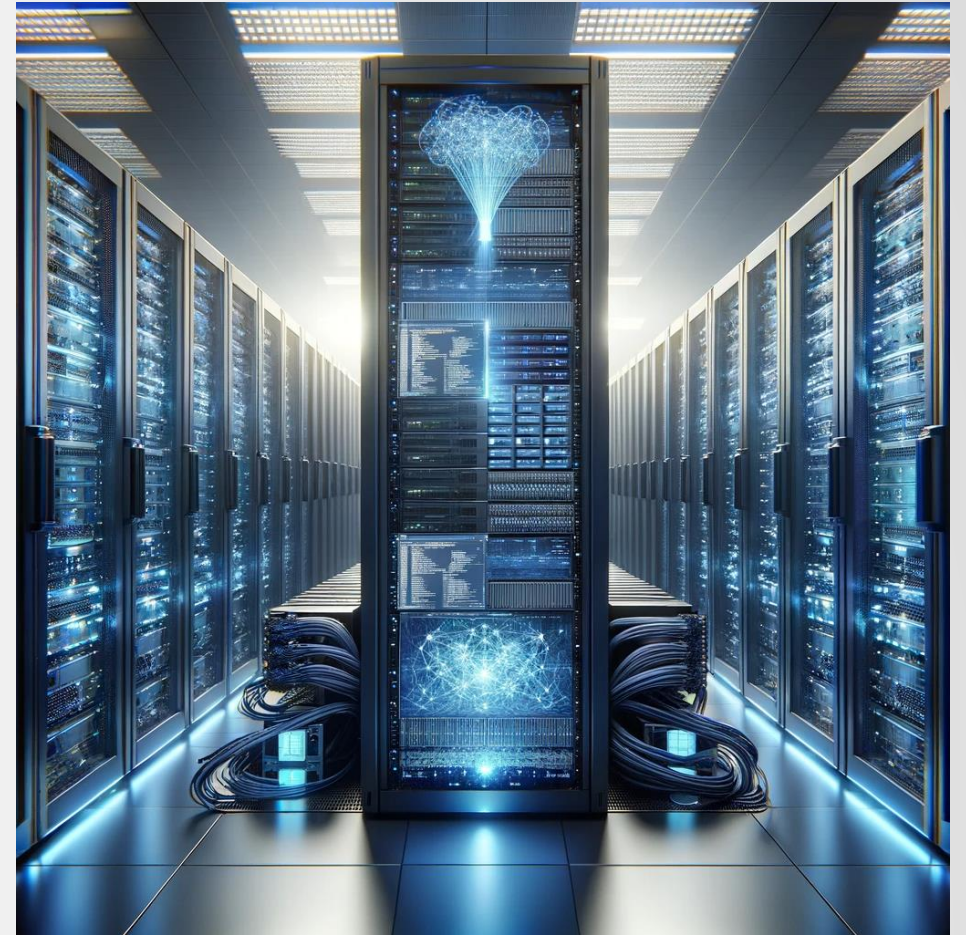
**Access to HPC resources for
AI-driven / compute-intensive applications
Dr Nikos Bakas**

Access to HPC infrastructures

Dr Nikos Bakas, GRNET

Contents

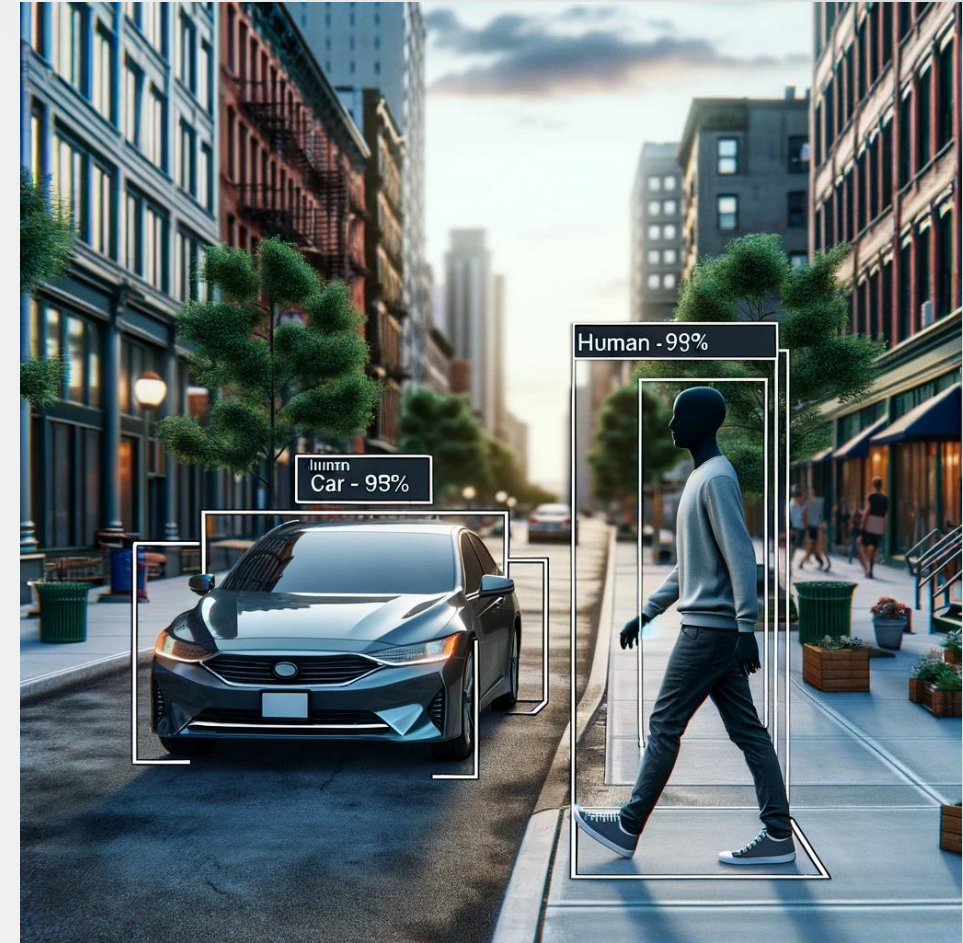
1. Artificial Intelligence (AI)
2. High-Performance Data Analytics (HPDA)
3. High-performance simulations
4. EuroCC success stories
5. Computing in Parallel at Scale
6. European Supercomputers
7. EuroHPC JU Access Types
8. How to Apply



Artificial Intelligence (AI)

Dr Nikos Bakas, GRNET

- **Access to Cutting-edge Technology:** SMEs gain access to advanced AI computing resources, enabling them to **develop and refine AI models** with efficiency and unprecedented speed.
- **Innovation and Competitive Edge:** Leveraging AI through supercomputers helps SMEs innovate, creating **new products and services**, thus staying competitive in a rapidly evolving digital landscape.
- **Large Language Models:** Enable **fine-tuning for custom applications**, enhancing precision and relevance in industry-specific contexts, driving targeted outcomes and efficiencies.
- **AI Skill Development:** Offers **training and skill development** in AI technologies for SMEs, helping them to build in-house expertise and apply AI solutions effectively.



generated with <https://chat.openai.com/>

High-Performance Data Analytics (HPDA)

Dr Nikos Bakas, GRNET

- **Data Processing at Scale:** HPDA equips SMEs with the capability to process and **analyze large datasets**, overcoming the limitations often faced due to smaller infrastructures
- **Large Image Analysis for Climate Studies:** HPDA enables the analysis of large and complex image datasets (such as satellite imagery) for SMEs involved in **environmental research, climate studies, or related fields.**
- **Healthcare Applications:** accelerate the analysis of **medical images or genetic data**, leading to faster **diagnosis and personalized treatments.**
- **Retail Markets:** process customer data to **tailor marketing strategies** or improve **personalized customer experience.**

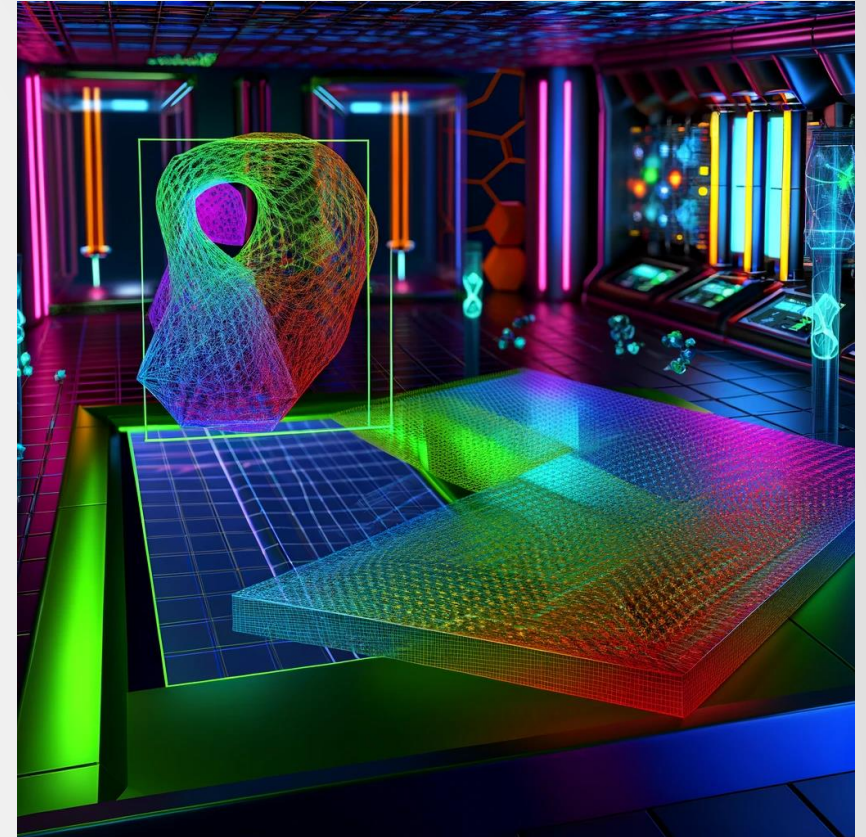


generated with <https://chat.openai.com/>

High-performance Simulations

Dr Nikos Bakas, GRNET

- **Engineering Simulations:** Utilizing methods like **finite elements** to analyze and design complex structures, mechanisms, or systems in various engineering fields.
- **Molecular Simulations:** Modeling and understanding the behavior of **molecular systems**, including the discovery of **new materials**. By simulating the interactions at the atomic or molecular level, researchers can predict the properties of materials before they are synthesized, leading to advances in materials science.
- **Drug Discovery:** By simulating the interaction between **drug molecules and biological targets**, researchers can identify promising drug candidates, significantly speeding up the **drug discovery** process and reducing the need for early-stage laboratory experiments.



generated with <https://chat.openai.com/>

EuroCC success stories

Dr Nikos Bakas, GRNET

The **EuroCC ACCESS success stories** showcase a variety of successful experiments conducted within the EuroCC projects, highlighting the **business benefits** derived from these collaborations. These experiments involve **partners from industry, society, and academia**, and cover a wide range of applications such as:

- Transfer and optimization of **CFD calculations** workflow in HPC environment
- Anomaly Detection in **Time Series Data**: Gambling prevention using Deep Learning
- Multimodal Prediction of **Alexithymia** from Physiological and **Audio Signals**
- Enabling HPC Usage for Expensive ML Tasks on **Manufacturing** Environments
- Rebar cutting **optimisation** using a cloud computing environment
- The Estonian **COVID-19 Data Portal** and KoroGeno-EST
- ... and many more...

Success Story: Transfer and optimization of CFD calculations workflow in HPC environment

- **Shark Aero company** designs and manufactures ultralight sport aircrafts with two-seat tandem cockpit.
- The computational and memory requirements rise with the **3rd power** of the number of **mesh vertices**.
- Workflow transfer into High-Performance Computing (HPC) environment was thus undertaken, with a special focus on the investigation of computational tasks **parallelization efficiency** for a given model type.

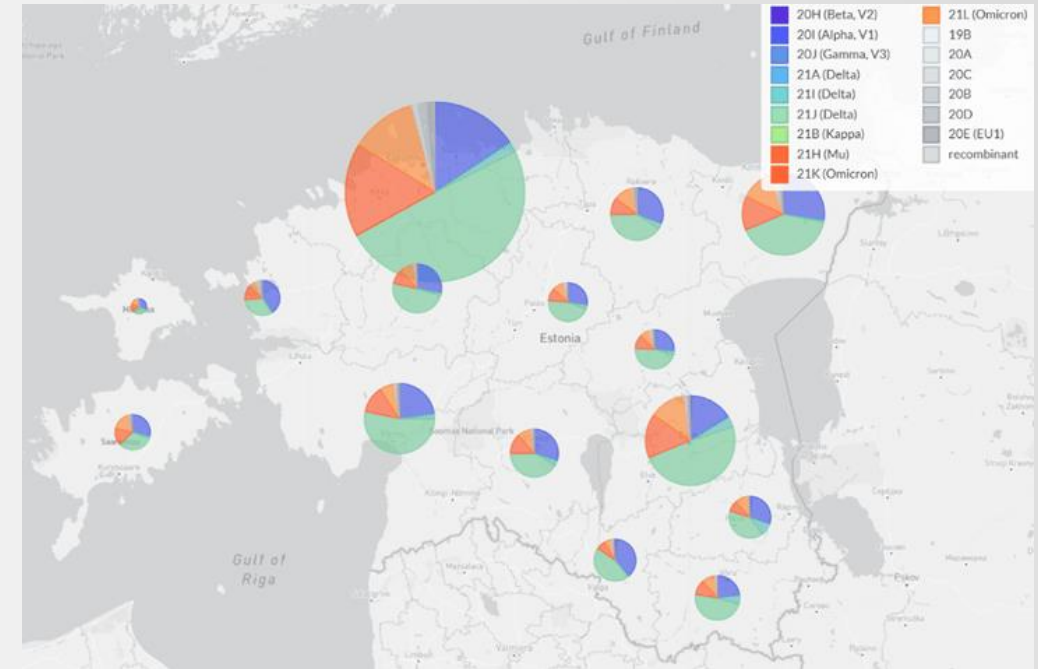


<https://www.eurocc-access.eu/success-stories/use-case-transfer-and-optimization-of-cfd-calculations-workflow-in-hpc-environment/>

Success Story: The Estonian COVID-19 Data Portal and KoroGeno-EST

- Analyze the **complete genomes of SARS-CoV-2** that have caused and are causing infections in Estonia
- Perform a **molecular epidemiological analysis** on them.
- Working with this problem involves a **very large-scale data analysis**, that can be conducted only by applying high-performance computing resources and relevant methods.

<https://www.eurocc-access.eu/success-stories/success-story-the-estonian-covid-19-data-portal-and-korogeno-est/>



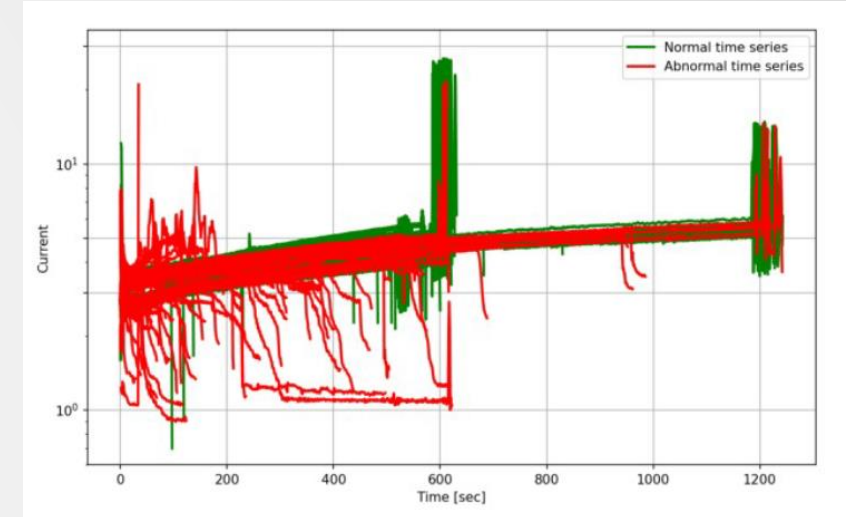
Industrial
Organisations
Involved



Success Story: Anomaly Detection of Noisy Time Series

- **Automatic detection of breaking fibres** during the **fibreglass winding** process.
- By analyzing the winding motor's current with **advanced time series analysis and machine learning**, they developed a **high-precision anomaly detection system**.
- The results helps the company management to **take informed decisions** on the development of their **break-detection strategy**.
- The company is better prepared to **assess the cost-benefit-relationship** of further educating their staff in data analytics, machine learning and modeling.

<https://www.eurocc-access.eu/success-stories/success-story-anomaly-detection-of-noisy-time-series/>



Industrial Organisations
Involved:



Computing in Parallel – Why it Matters

Example: Brightness increase of 1_000_000 Images

```
using Base.Threads, Images
```

```
image_paths = ["image1.jpg", "image2.jpg", ..., "image1_000_000.jpg"]
```

```
brightness_increase = 50 # Define your brightness increase
```

```
@threads for path in image_paths
```

```
    image = load_image(path) # Load the image
```

```
    adjusted_image = clamp.(image .+ brightness_increase, 0, 255) # Adjust brightness
```

```
    save_image(adjusted_image, "adjusted_$path") # Save the adjusted image
```

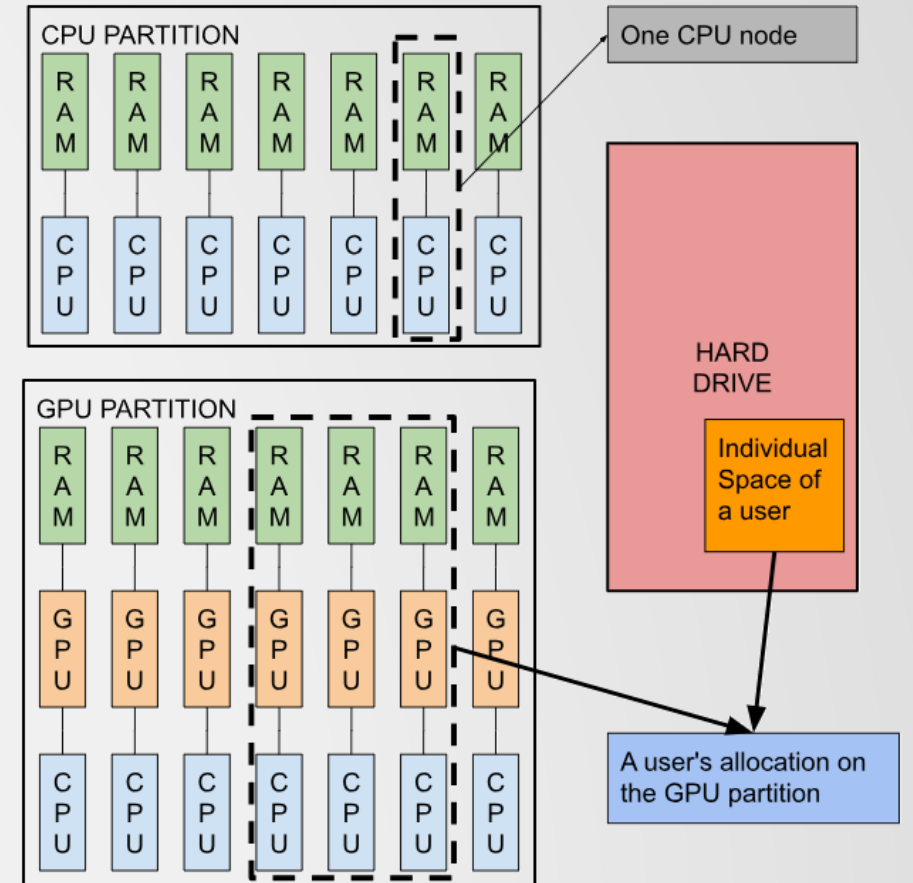
```
end
```

- ✓ **Parallelism:** Distributes image processing across threads, making 1,000,000 images as fast to process as one, assuming sufficient resources.
- ✓ **Vectorized Computations:** Modern processors use vectorized instructions (SIMD) to operate on multiple data points simultaneously within a single instruction.
- ✓ **Computational Complexity:** Turning an $O(n)$ operation into an effective $O(1)$ operation. ←



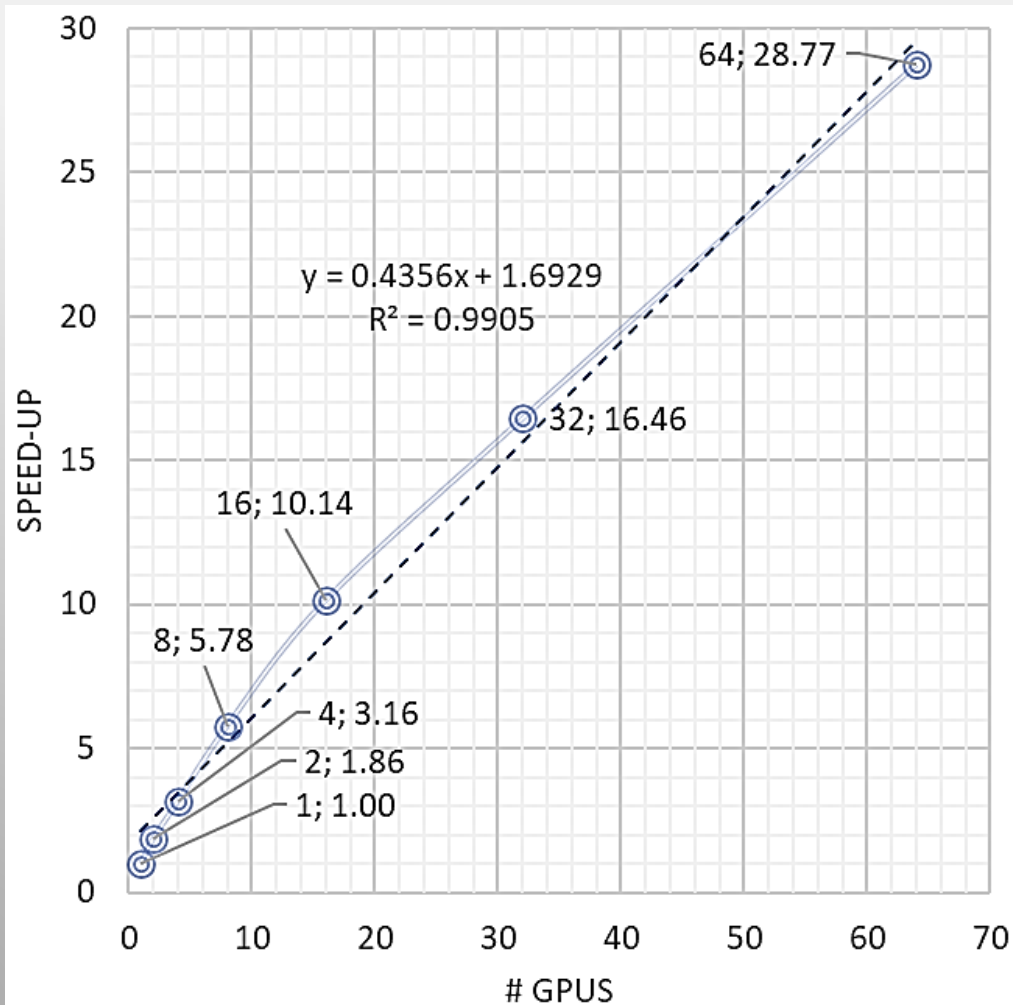
Computing in Parallel at Scale

- A cluster divided into **CPU and GPU partitions**, each with dedicated RAM and CPUs/GPUs.
- Individual **user space** is allocated on the **hard drive** for storage.
- A CPU node highlighted - **multiple CPU nodes** exist within the CPU partition.
- **User's allocation in the GPU partition** is indicated, showing a combination of RAM and GPU resources assigned to a user.



```
sbatch --gres=gpu:4 -N 16 --ntasks-per-node=4 my_gpu_job.sh
```


1-64 GPUs



# GPUs	train minutes (10 epochs)	Speed-Up	% valid. Accuracy	Accuracy Loss
1	44.82	1.00	89.00	0.00%
2	24.07	1.86	89.14	0.16%
4	14.18	3.16	89.11	0.12%
8	7.76	5.78	88.90	-0.11%
16	4.42	10.14	88.46	-0.61%
32	2.72	16.46	88.69	-0.35%
64	1.56	28.77	86.40	-2.92%

- image size: 512x512
- train batch size: 52 (per GPU)
 - learning rate: 1e-4
 - weight decay: 1e-6

European Supercomputers

Dr Nikos Bakas, GRNET



LUMI
FINLAND



LEONARDO
ITALY



MELUXINA
LUXEMBOURG



KAROLINA
CZECH REPUBLIC



DISCOVERER
BULGARIA



VEGA
SLOVENIA



DEUCALIO
PORTUGAL



MARENOSTRUM 5
SPAIN

The EuroCC project, Europe aims to **democratize access** to supercomputing resources, enabling **SMEs** across the continent to leverage **AI, high-performance data analytics, and simulations**, thereby fostering innovation, competitiveness, and growth in the European digital economy.

European Supercomputers

Dr Nikos Bakas, GRNET

1. LUMI (CSC, Finland)

- LUMI-C: 1536 nodes, 128 cores/node, 256-1024 GB RAM/node
- GPU: 2560 nodes, 64 cores/node, 4 GPUs, 128 GB GPU-RAM
- Visualization: 64 nodes, 1 GPU, 48 GB GPU-RAM
- Peak Performance: 550 petaflops
- URL: <https://www.lumi-supercomputer.eu/lumis-full-system-architecture-revealed/>

2. Leonardo (Cineca, Italy)

- Booster Module: 3456 nodes, 32 cores/node, 512 GB RAM/node, 4 GPUs, 64 GB GPU-RAM
- Data Centric Module: 1536 nodes, 112 cores/node, 512 GB RAM/node
- Peak Performance: 323.4 petaflops
- URL: <https://leonardo-supercomputer.cineca.eu/hpc-system/>

3. MareNostrum 5 (Barcelona Supercomputing Center, Spain)

- General Purpose Partition: 6408 nodes, 112 cores/node, 256 GB RAM/node
- Accelerated Partition: 1120 nodes, 64 cores/node, 512 GB RAM/node, 4 GPUs, 64 GB GPU-RAM
- Peak Performance: 314 petaflops
- URL: <https://www.bsc.es/innovation-and-services/marenostrum/marenostrum-5>

4. MeluXina (LuxProvide, Luxembourg)

- Cluster: 573 nodes, 128 cores/node, 512 GB RAM/node
- Accelerator-GPU: 200 nodes, 64 cores/node, 512 GB RAM/node, 4 GPUs, 40 GB GPU-RAM
- Large memory: 20 nodes, 128 cores/node, 4096 GB RAM/node
- Peak Performance: 18.29 petaflops
- URL: <https://docs.lxp.lu/system/overview/>

European Supercomputers

Dr Nikos Bakas, GRNET

5. Karolina (IT4I, Czech Republic)

- CPU: 828 nodes, 128 cores/node, 256-24000 GB RAM/node
- GPU: 72 nodes, 8 GPUs, 40 GB GPU-RAM
- Peak Performance: 15.69 petaflops
- URL: <https://www.it4i.cz/en/infrastructure/karolina>

6. Vega (IZUM, Slovenia)

- GPU partition: 60 nodes, 128 cores/node, 512 GB RAM/node, 4 GPUs, 40 GB GPU-RAM
- CPU node Standard: 768 nodes, 128 cores/node, 256 GB RAM/node
- CPU node Large Memory: 192 nodes, 128 cores/node, 1000 GB RAM/node
- Peak Performance: 10.05 petaflops
- URL: <https://doc.vega.izum.si/architecture/>

7. Deucalion (Guimarães, Portugal)

- ARM cluster: 1632 nodes, 48 cores/node
- X86 cluster: 500 nodes, 48+ cores/node
- Accelerated partition: 33 nodes
- Peak Performance: 10 petaflops
- URL: <https://macc.fccn.pt/resources#deucalion>

8. Discoverer (Sofia Tech Park, Bulgaria)

- CPU: 1128 nodes, 128 cores/node, 256 GB RAM/node
- CPU-Fat: 18 nodes, 128 cores/node, 1000 GB RAM/node
- Peak Performance: 5.94 petaflops
- URL: https://docs.discoverer.bg/resource_overview.html

EuroHPC JU Benchmark Access

Dr Nikos Bakas, GRNET



EuroHPC JU Benchmark Access

The purpose of the EuroHPC JU Benchmark Access calls is to support researchers and HPC application developers by giving them the opportunity **to test or benchmark their applications** on the upcoming/available EuroHPC Pre-exascale and/or Petascale system prior to applying for an Extreme Scale and/or Regular Access. The EuroHPC Benchmark call is designed **for code scalability tests or for test of AI applications** and the outcome of which is to be included in the proposal in a future EuroHPC Extreme Scale and Regular Access call. Users receive a **limited number of node hours**; the maximum allocation period is **three months**.

EuroHPC JU Development Access

Dr Nikos Bakas, GRNET



EuroHPC JU Development Access

The purpose of the EuroHPC JU Development Access calls is to support researchers and HPC application developers by giving them the opportunity **to develop, test and optimise their applications** on the upcoming/available EuroHPC Pre-exascale and/or Petascale system prior to applying for an Extreme Scale and/or Regular Access. The EuroHPC Development call is designed for projects **focusing on code and algorithm development and optimisation**, as well as **development of AI application methods**. This can be in the context of research projects from academia or industry, or as part of large public or private funded initiatives as for instance Centres of Excellence or Competence Centres. Users will typically be allocated a small number of node hours; the allocation period is **one year** and is renewable up to two times.

EuroHPC JU Regular Access

Dr Nikos Bakas, GRNET

EuroHPC JU Regular Access

The Regular Access mode is designed to serve **research domains, industry open R&D and public sector applications that require large-scale resources** or that require more frequent access to substantial computing and storage resources. This access mode distributes resources, mostly from the EuroHPC JU **petascale** systems.

This Regular Access Call offers three distinctive application tracks:

- **Scientific Access** – Intended for applications from the academia and public research institutes.
- **Industry Access** – Intended for applications with Principal Investigator (PIs) coming from industry.
- **Public Administration Access** – Intended for applications with PIs coming from the public sector.

EuroHPC JU Extreme Access

Dr Nikos Bakas, GRNET

EuroHPC JU Extreme Access

The Extreme Scale Access Mode call is targeting HPC **applications with high-impact and high-gain innovative research**. This access mode distributes resources, from the EuroHPC pre-exascale systems.

This call offers three distinctive application tracks:

- **Scientific Access** – Intended for applications from the academia and public research institutes.
- **Industry Access** – Intended for applications with Principal Investigator (PIs) coming from industry.
- **Public Administration Access** – Intended for applications with PIs coming from the public sector.

EuroHPC JU Access Call for AI and Data-Intensive Applications

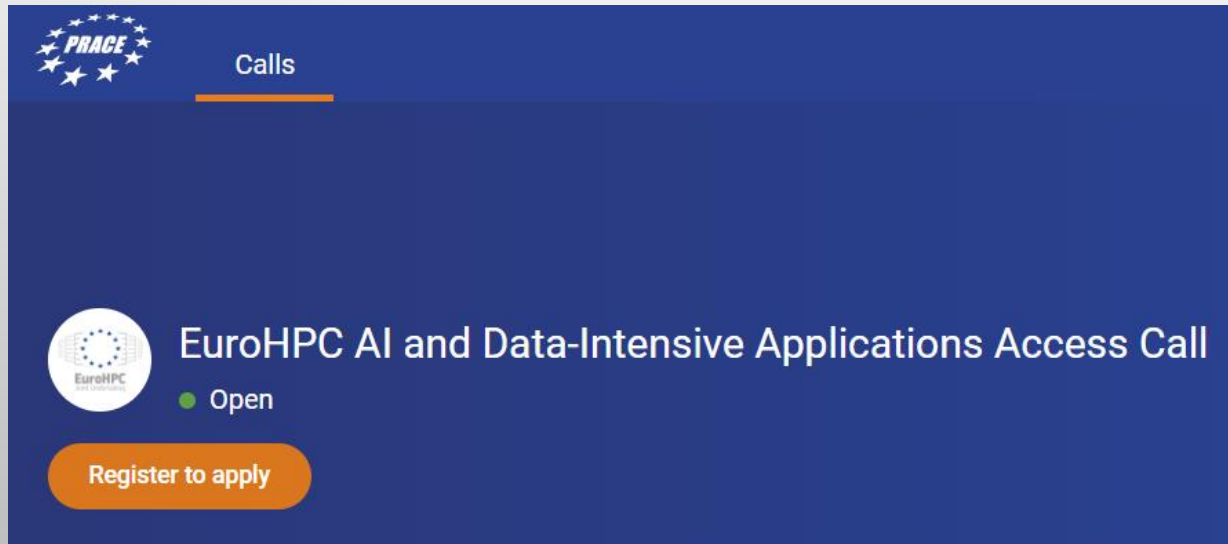
EuroHPC JU Access Call for AI and Data-Intensive Applications

https://eurohpc-ju.europa.eu/eurohpc-ju-access-call-ai-and-data-intensive-applications_en

SYSTEM*	SITE (COUNTRY)	ARCHITECTURE	PARTITION	TOTAL RESOURCES**	FIXED ALLOCATION
 MN5 MARENOSTRUM	BSC (ES)	Atos BullSequana XH3000	MN5 ACC	129 377	32 000
 LEONARDO CINECA	CINECA (IT)	Atos BullSequana XH2000	Leonardo Booster	545 865	50 000
 LUMI	CSC (FI)	HPE Cray EX	LUMI-G	351 455	35 000
 MELUXINA HIGH PERFORMANCE COMPUTING IN LUXEMBOURG	LuxProvide (LU)	Atos BullSequana XH2000	MeluXina GPU	25 000	25 000
KAROLINA	IT4I VSB-TUO (CZ)	HPE Apollo 2000 Gen10 Plus and HPE Apollo 6500	Karolina GPU	7 500	7 500
 VEGA HPC	IZUM Maribor (SI)	Atos BullSequana XH2000	Vega GPU	7 100	7 100

EuroHPC JU AI and Data-Intensive Applications

Dr Nikos Bakas, GRNET



The screenshot shows a dark blue interface for PRACE Calls. At the top left is the PRACE logo (stars and the word PRACE). To its right is the word 'Calls' with an orange underline. Below this, there is a circular EuroHPC logo. To the right of the logo is the text 'EuroHPC AI and Data-Intensive Applications Access Call' and a green dot with the word 'Open'. At the bottom left of the card is an orange button with the text 'Register to apply'.

<https://pracecalls.eu/calls/32>

Call Details

The EuroHPC JU AI and Data-Intensive Applications Access call aims to support ethical **artificial intelligence**, **machine learning**, and in general, data-intensive applications, with a particular focus on foundation models and **generative AI** (e.g. large language models).

The call is intended to serve **industry** organizations, small to medium enterprises (**SMEs**), **startups**, as well as **public sector** entities, requiring access to supercomputing resources to perform artificial intelligence and data-intensive activities.

EuroHPC JU AI and Data-Intensive Applications

Dr Nikos Bakas, GRNET



Partition information

Partition name*

MeluXina GPU

Requested amount of resources (node hours)*

25 000

This value is pre-defined

Previous Benchmark or Development Access allocations

Project ID

Please provide an ID in case you have previously obtained data relevant for this proposal via the Benchmark and/or Development Access call

+ Previous Benchmark or Development Access allocations

Frequently Asked Questions (FAQ)

Dr Nikos Bakas, GRNET

— Which organisations are eligible for access to EuroHPC machines?

Any European organisation is eligible for access to perform Open Science research (the results of the work are made available for open access). This includes public and private academic and research institutions, public sector organisations, industrial enterprises and SMEs.

+ What documents are required for access?

— What is the cost?

Currently access is free of charge.

— What are the participation conditions?

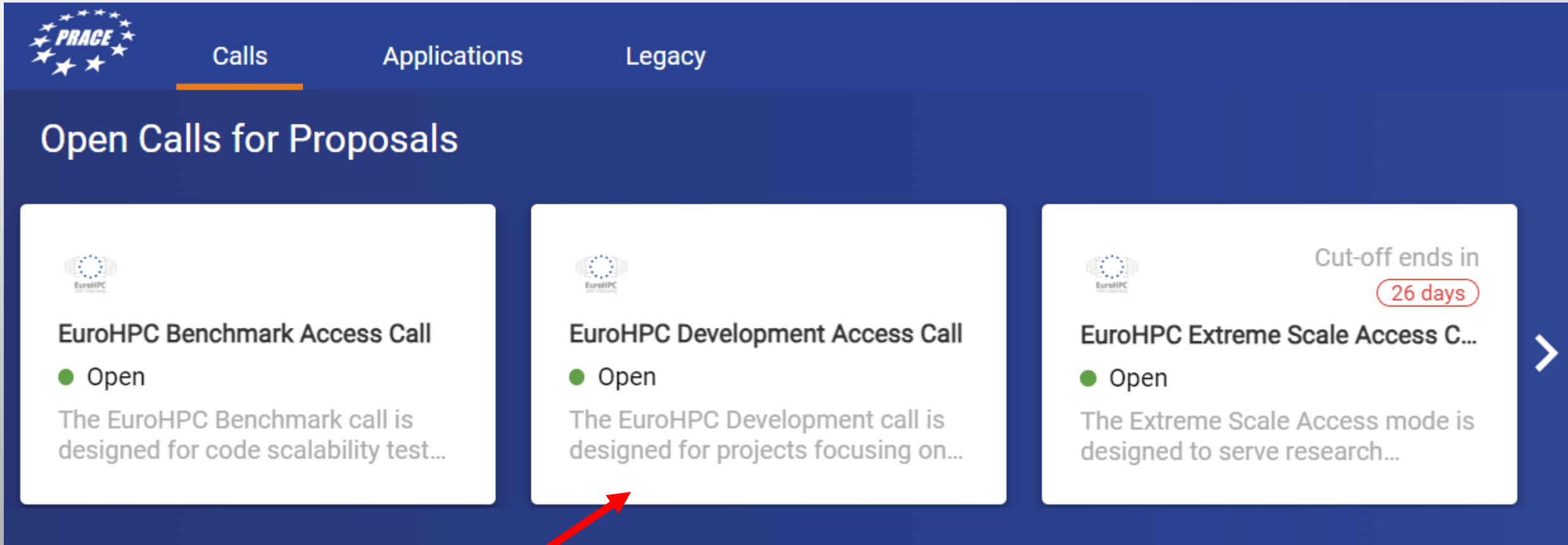
Participation conditions depend on the specific access call that a research group has applied. In general users of EuroHPC systems commit to:

- acknowledge the use of the resources in their related publications,
- contribute to dissemination events,
- produce and submit a report after completion of a resource allocation.

More information on participation conditions can be found in the call's Documents section.

How to Apply

Dr Nikos Bakas, GRNET




The screenshot shows the PRACE website's 'Calls' section. The navigation bar includes 'PRACE', 'Calls', 'Applications', and 'Legacy'. The 'Calls' section is titled 'Open Calls for Proposals' and features three call cards:

- EuroHPC Benchmark Access Call**: Status: Open. Description: The EuroHPC Benchmark call is designed for code scalability test...
- EuroHPC Development Access Call**: Status: Open. Description: The EuroHPC Development call is designed for projects focusing on... (indicated by a red arrow).
- EuroHPC Extreme Scale Access C...**: Status: Open. Description: The Extreme Scale Access mode is designed to serve research... (includes a 'Cut-off ends in 26 days' warning and a right-pointing arrow).

Access to HPC infrastructures

Dr Nikos Bakas, GRNET



PARTNERSHIP FOR ADVANCED
COMPUTING IN EUROPE

Welcome

Please, login or register for free

E-mail*


Password*

Log in



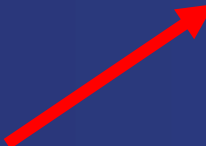
Register

[Forgot password?](#)

[Terms of Use & Privacy Policy](#)



Calls Applications Legacy



EuroHPC Development Access Call

● Open

Apply to Call

<https://pracecalls.eu/auth/login>

Project Application

- The Project**
- Principal Investigator
- Contact Person and Team Members Information
- Partitions
- Code Details and Development
- Data Consent

The Project

Project details

Project title*

HPC Support on Machine Learning and Generative AI Algorithms

Project summary (abstract)*

This project aims to develop, test, and provide open-access Machine Learning (ML) and Generative AI algorithms tailored for HPC environments. We aim to boost the efficiency, accessibility, and innovation capacity of HPC users in Greece for a wide range of scientific and engineering applications. By offering open access to cutting-edge algorithms, we aim to empower researchers and developers to optimize HPC workloads and enhance computational efficiency. The initiative will focus on creating and testing scalable and easily integrated ML models and generative AI frameworks that can be readily adopted by the HPC community in Greece. This approach ensures that the benefits of advanced computational techniques are not confined to institutions with significant resources but are available to all researchers and developers.

Explain the scientific case of the project for which you intend to use the code(s)*

Through this project, we aim to assist our efforts for open and collaborative HPC computing, fostering innovation and breakthroughs across various disciplines. Access to supercomputing facilities for testing and refining these algorithms is critical. It will allow us to ensure their performance at scale and their applicability across different HPC systems and research domains.

Deadline

01/04/2024 12:00:00 PM

Documents

Delete Application



Save Changes


Next

Apply for HPC Access

Dr Nikos Bakas, GRNET

Proposal for civilian purposes* 

Is any part of the project confidential?*


Yes No 

Research fields

Research field title*

PE6 Computer Science and Informatics 

Research field sub-title*

PE6_7 Artificial intelligence, intelligent systems, natural language processing 

Research field share (%)*

50

The sum of all research fields should not exceed the total of 100%

Apply for HPC Access

Dr Nikos Bakas, GRNET

Research fields #2

Research field title*

PE6 Computer Science and Informatics

Research field sub-title*

PE6_11 Machine learning, statistical data processing and applications using signal

Research field share (%)*

50

The sum of all research fields should not exceed the total of 100%

Remove

+ Research fields

AI set of technologies selection

Machine Learning

Natural Language Processing

Deep Learning

If applicable, please select used AI technologies. This is a multi-select field so you are able to choose more than one option.

Apply for HPC Access

Dr Nikos Bakas, GRNET

Please specify how does your project ensure ethical principles and addresses potential societal impacts associated with the development and deployment of AI technologies*

We prioritize transparency by openly sharing our AI methodologies and limitations. By providing open access to our algorithms, we democratize access to advanced computational resources. Our project serves as a catalyst for innovation across scientific and engineering disciplines. By enhancing computational efficiency, we enable researchers to tackle complex problems, from climate change to healthcare, contributing to societal well-being. We commit to responsible Generative AI deployment, adhering to ethical guidelines.

Submission details

Project duration*

6 months 12 months

Preferable starting date*

01-04-2024



Industry involvement*

As Team Member



Public sector involvement*

As Team Member



Apply for HPC Access

Dr Nikos Bakas, GRNET

Partitions

Partition name*

MeluXina CPU

Code(s) used*

XGBoost MPI Horovod Pytorch

This field is a multi-text field, for adding another code separate it with a comma

Requested amount of resources (node hours)*

4 000

Average number of processes/threads*

128

Average job memory (total usage over all nodes in GB)*

400

Maximum amount of memory per process/thread (MB)*

10 000

Total amount of data to transfer to/from (GB)*

100

Apply for HPC Access

Dr Nikos Bakas, GRNET

I/O libraries, MPI I/O, netcdf, HDF5 or other approaches*

Custom Python scripts with multiprocessing for parallel data loading and process.

Frequency and size of data output and input*

Data input/output occurs hourly, with inputs averaging 10 GB per batch and outputs around 5 GB, optimized for real-time ML model training and analysis.

Number of files and size of each file in a typical production run*


Around 500 files, each averaging 20 MB, facilitating efficient batch processing for ML training on HPC.

Total storage required (GB)

100

Apply for HPC Access

Dr Nikos Bakas, GRNET

Partitions #2 

Partition name*

MeluXina GPU 

Code(s) used*

Llama Falcon Mistral

This field is a multi-text field, for adding another code separate it with a comma

Requested amount of resources (node hours)*

800 

Average number of processes/threads*

64

Average job memory (total usage over all nodes in GB)*

800

Maximum amount of memory per process/thread (MB)*

12 500

Apply for HPC Access

Dr Nikos Bakas, GRNET

Code Details and Development

This tab should overall include the following: description of main algorithms, how they have been implemented and parallelized, and their main performance bottlenecks and the solutions to the performance issues you have considered. For each code that needs to be optimized, please provide the details below. Codes can be added by clicking on the Add code button.

Development of the code(s) description*

For machine learning we will be using custom codes such as <https://github.com/nbakas/hpmlt/> for hyperparameter tuning in parallel. For Gen-AI we will be using open access LLMs, such as Mistral, Llama, and Falcon

Code details

Name and version of the code

High Performance Machine Learning Algorithms for Tabular Datasets
<https://github.com/nbakas/hpmlt/>
Version 1.0

Apply for HPC Access

Dr Nikos Bakas, GRNET

Scalability and performance ←

Describe the scalability of the application and performance of the application*

Large language models (LLMs) to be used, like Mistral, Llama, and Falcon are designed with scalability in mind, allowing them to handle extensive computational loads efficiently. These models can scale across multiple GPUs and nodes in a high-performance computing environment, enabling parallel processing of large datasets and complex computations. E.g. see:
<https://huggingface.co/blog/falcon-180b>
<https://www.truefoundry.com/blog/benchmarking-llama2-falcon-and-mistral>

What is the target for scalability and performance?*

Our target for scalability and performance involves optimizing our application to efficiently utilize HPC resources, aiming to support concurrent processing of large datasets and ensuring rapid response times for machine learning tasks involving open-access LLMs like Mistral, LLaMA, and Falcon.

i.e. what performance is needed to reach the envisaged scientific goals

Apply for HPC Access

Dr Nikos Bakas, GRNET

Optimization of the work proposed

Explain how the optimization work proposed will contribute to future Tier-0 projects*

The proposed optimization work enhances computational efficiency, reduces runtime, and scales effectively across HPC resources, directly supporting the ambitious goals of future Tier-0 projects by enabling more complex and larger-scale computations.

Describe the impact of the optimization work proposed - is the code widely used; can it be used for other research projects and in other research fields with minor modifications; would these modifications be easy to add to the main release of the software?*

The optimization work aims to make the code highly versatile and adaptable, allowing for broad use across various research fields with minimal adjustments. These modifications will be made available to the users we support.

Describe the main algorithms and how they have been implemented and parallelized*

LLMs implement a variety of sophisticated algorithms, primarily based on the Transformer architecture, which is renowned for its self-attention mechanism. This architecture allows the models to weigh the importance of different words within the input text, enabling a deeper understanding of the context. They are implemented using frameworks that support parallel processing, such as PyTorch and TensorFlow. These frameworks provide built-in support for data, model, and pipeline parallelism, facilitating the development and training of large models on distributed computing resources.

Apply for HPC Access

Dr Nikos Bakas, GRNET

Performance

Main performance bottlenecks*

The main performance bottlenecks in training and deploying LLMs typically include memory constraints, data loading and I/O throughput, inter-node communication overheads, and the computational cost of forward and backward passes.

Describe possible solutions you have considered to improve the performance of the project*

Pretrained LLMs already incorporate advanced parallel processing, model compression techniques like quantization and pruning, and efficient data loading and caching mechanisms to optimize performance and address computational and memory constraints.

Describe the application enabling/optimization work that needs to be performed to achieve the target performance*

To achieve target performance, optimization work includes implementing dynamic batching for efficient GPU utilization, fine-tuning parallel processing strategies, optimizing memory allocation, and incorporating adaptive learning rate schedulers.

Which computational performance limitations do you wish to solve with this project?*

This project aims to solve computational performance limitations related to high memory demands, slow model training and inference speeds, and inefficient data handling. Our goal is to enable users to efficiently run and benefit from the latest LLMs like Mistral, LLaMA, and Falcon, making these advanced AI technologies more accessible and practical for diverse research and development purposes.

Apply for HPC Access

Dr Nikos Bakas, GRNET

https://eurohpc-ju.europa.eu/eurohpc-ju-call-proposals-development-access_en

Data Consent

The EuroHPC JU will process personal data in accordance with Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC.

In case the proposal is awarded, EuroHPC JU would like to publish the Principal Investigator's and Team Members' names and organizations. This may involve sharing this information on our website, social media channels, or in other promotional materials related to the project. Please provide your consent below.*

I consent I do not consent

In order to submit the proposal, you must accept the terms and conditions stated in the Access Policy, hence confirming that you have read and understood the call procedures. The documentation can be found at https://eurohpc-ju.europa.eu/eurohpc-ju-call-proposals-development-access_en*

I accept the terms and conditions stated in the Access Policy

Back

Save Changes

Submit

Deadline

01/04/2024 12:00:00 PM

  Documents

Delete Application

How to Apply



<https://www.youtube.com/watch?v=g5jOio006-E>



<https://www.youtube.com/watch?v=N1QqMh7H0mQ>

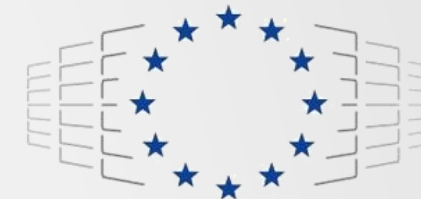
<https://eurocc-greece.gr/how-to-apply-for-access-to-eurohpc-ju-supercomputers/>

Thank you!

Access to HPC resources for AI-driven / compute-intensive applications Dr Nikos Bakas - GRNET



Co-funded by
the European Union



EuroHPC
Joint Undertaking

Funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia under grant agreement No 101101903.