ComFac4LT – Computing Facilities for Language Technologies

Gaurish Thakkar, Marko Tadić
University of Zagreb
Faculty of Humanities and Social Sciences
Ul. Ivana Lučića 3, 10000, Zagreb
{gthakkar, marko.tadic}@ffzg.hr

German Rigau
University of the Basque Country
HiTZ Center - Ixa
M. Lardizabal 1, 20080, Donostia
german.rigau@ehu.eus

Abstract. *High-performance computing (HPC) is an essential service required in multiple research areas. This paper maps the existing HPC landscape and enumerates the various access types available to researchers working in Europe. Furthermore, the present study examines the perspectives of researchers in the field of language technology gathered through an online survey. Requesting resources from HPC service providers can be a time-consuming process; therefore, a dynamic access mode should be made available to reduce the request time. Collaboration with academia is a viable option for a SME user who wishes to utilise HPC services. Academic institutions in a nation without HPC resources should rely on international partnerships for HPC services. Users should have access to centralised information regarding HPCs to facilitate the location of services.*

Keywords. High-Performance Computing, Language Technologies, PRACE, EuroHPC-JU, Natural Language Processing

1 Introduction

Language Technology (LT) is a highly researched field with high socio-economic impacts. Textual analysis processes facilitate knowledge acquisition and strategic decision-making. The additional knowledge extracted from the text has been largely attributed to advances in the field of. All of these advancements have been made possible by the availability of data, improved data processing techniques (algorithms), and processing capabilities made available to researchers over time.

There has been a clear shift away from knowledge-based and human-engineered methods and towards data-driven methods, which has led to progress in the field of LT. One recent aspect associated with the paradigm shift in language processing is the use of large language models. Large-scale monolingual and/or multilingual text data is used to train language models. Pre-trained large language models, like BERT (Devlin et al., 2019), GPT (Brown et al., 2020), GPT-4 (OpenAI, 2023), and XLM-Roberta (Conneau et al., 2020), have offered a framework for using the knowledge acquired during the training process to be later applied to newer tasks.

As previously stated, one aspect associated with the boom of AI-based data-driven techniques for NLP is the ability to crunch data using efficient hardware in the form of Graphic Processing Units (GPUs). In neural language model training, the cost component is realised in the form of hardware and its operation. This directly results in organisations, with facilitated access to these hardware resources, having access to the research and development of LT (Ahmed and Wahed, 2020).

BLOOM (Scao et al., 2022) is a large language model with 176 billion parameters that can write text in 46 natural languages. The model was trained using the Jean Zay public supercomputer with 384 NVIDIA A100 80 GB GPUs (48 nodes) for 117 days. Building such models with numerous parameters that are learned during training necessitates an equally capable system with capable hardware.

This study reports on the results of an investigation of the available high-performance computing (HPC) facilities available to LT researchers. In addition to the different HPC infrastructures that are available, the paper looks at aspects like access protocols, calls, and eligibility. The compatibility of existing models is directly correlated with the available GPU. Hence, GPU hardware and associated details form another slice of data that is useful from an LT point of view. For research and innovation in the field of LT, a comprehensive understanding of the available infrastructure alternatives is essential.

The primary objective of this research is to evaluate the capacity of various high-performance computing (HPC) initiatives, including EuroHPC JU, PRACE, LUMI, and national consortia, in supporting contemporary LT. Specifically, the paper aims to:

- assess the hardware capabilities, especially the number of nodes and GPUs, of these HPC systems for both small and large-scale experiments.
- evaluate the accessibility protocols offered by these initiatives.

- examine the resource request and allocation procedures for various user roles.
- survey the various access modes provided by these HPCs, with a specific focus on accessibility for Small and Medium Enterprises (SMEs)..

2 Background

2.1 High-Performance Computing (HPC)

HPC, also called a supercomputer, provides the opportunity to solve complex problems in different applications[1]. Running applications in parallel to speed up performance is required for highly computational tasks such as pre-training a neural language model. In computing, floating point operations per second (FLOPS, flops, or flop/s) is a measure of computer performance, useful in fields of scientific computation that require floating-point calculations (Dolbeau, 2018). To put it into perspective, a laptop, or desktop with a 3 GHz processor can perform around 3 billion calculations per second. HPC solutions, on the other hand, can perform quadrillions of calculations per second, i.e., 1 Petaflops (10^{15}). The orders of magnitude in computer performance can be understood as follows:

- A 1 petaflops (PFLOPS) computer system is capable of performing one quadrillion (10^{15}) floating-point operations per second.
- A 1 exaflops (EFLOPS) computer system is capable of performing one quintillion (10^{18}) floating-point operations per second.

The following are the components[2] of an HPC solution:

- **Server**: responsible for computing
- **Network**: interconnection between the servers, responsible for high-speed transfers between servers and storage units.
- **Storage**: store for feeding data to servers, as well as persisting data received as output of the processing operation.

The collection of such servers (each server is a node) forms an HPC cluster. In addition to the abovementioned components, there are accelerated nodes, i.e., computer nodes with GPUs or any other accelerator like a Xeon Phi[3]. At the time of collecting data for the paper, **Frontier**[4] HPC in the USA is rated at 1.685 exaFLOPS (Rpeak) and is the world's fastest supercomputer in operation. **Fugaku** HPC in Japan comes in second with 537 PFLOPS (Rpeak) followed by **LUMI** in Finland with 428 PFLOPS (Rpeak). The recent trend being followed by the infrastructure providers is to shift computing to the level of exascale (10^{18} floating point operations per second). Benchmarking (Luszczek et al., 2005; Jiang et al., 2019, 2021), performance monitoring (Truong et al., 2001), resource utilisation (Asch et al., 2018), user surveys and feedback (Wolter et al., 2006), and documentation review (Lathrop et al., 2019) are some methods used to investigate HPC facilities. Previous attempt at mapping the HPC landscape in Europe can be linked to Berberich et al. (2019); Eicker et al. (2020). However, visualising the HPC landscape with LT as the primary focus is an unexplored field.

2.2 HPC Initiatives in Europe

The European High-Performance Computing Joint Undertaking (**EuroHPC-JU**) (Skordas, 2019) is a joint project that brings together the resources of the European Union. It is involved in activities such as the procurement and installation of supercomputers throughout Europe. In addition, it is involved in developing sustainable HPC technologies for efficient and cleaner computing. Other objectives of EuroHPC-JU are to design and develop applications and algorithms for HPC services, as well ease access to potential HPC users like SMEs and HPC experts across Europe. To date, five supercomputers[5] are now fully operational: **LUMI** in Finland (which ranks number 3 in the world), **LEONARDO** in Italy (which ranks number 4 in the world), **Vega** in Slovenia, **MeluXina** in Luxembourg, **Discoverer** in Bulgaria, **Karolina** in the Czech Republic, and Supek in Croatia which was opened right after we completed this research, so, unfortunately, it wasn't included in this data collection. Two supercomputers are underway: **Deucalion** in Portugal, and **MareNostrum5** in Spain. The list of EuroHPC-JU public members can be found at `https://eurohpc-ju.europa.eu/about/discover-eurohpc-ju_en`.

PRACE[6] (Partnership for Advanced Computing in Europe) (Hutton et al., 2019) is a not-for-profit international association that aims to facilitate access to a research infrastructure that enables high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. It has 25 member countries [7] whose representative organisations create a pan-European supercomputing infrastructure, providing access to computing and data management resources and services for large-scale scientific and engineering applications at the highest performance level. The computer systems and their operations accessible through PRACE are provided by five PRACE members (BSC representing Spain, CINECA representing Italy, ETH Zurich/CSCS representing Switzerland, GCS representing Germany, and GENCI representing France). Figure 1 shows PRACE member countries.

[1] https://www.ff4eurohpc.eu/en/about/what-is-hpc/

[2] https://www.netapp.com/data-storage/high-performance-computing/what-is-hpc/

[3] https://en.wikipedia.org/wiki/Xeon_Phi

[4] https://en.wikipedia.org/wiki/Frontier_(supercomputer)

[5] https://digital-strategy.ec.europa.eu/en/policies/high-performance-computing-joint-undertaking

[6] https://prace-ri.eu/

[7] https://prace-ri.eu/about/members/

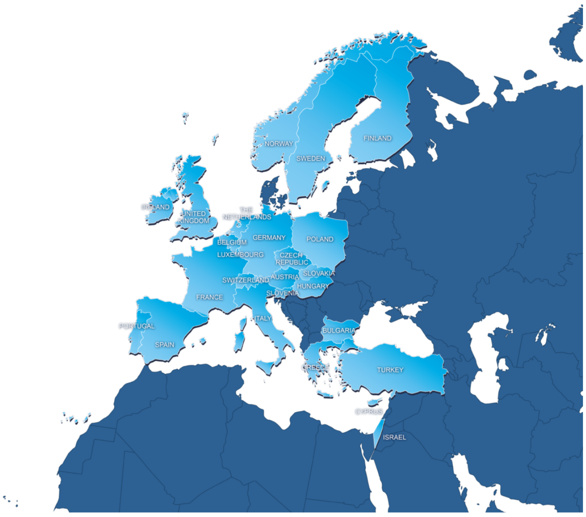


Figure 1: The list of PRACE members is displayed in the figure.

LUMI consortium [8] consists of ten European countries and provides a high-quality, cost-efficient, and environmentally sustainable HPC ecosystem based on true European collaboration. The LUMI (Large Unified Modern Infrastructure) consortium countries are Finland, Belgium, the Czech Republic, Denmark, Estonia, Iceland, Norway, Poland, Sweden, and Switzerland. Half of the LUMI resources belong to the EuroHPC Joint Undertaking, and the other half of the resources belong to the participating countries, i.e., the LUMI consortium countries. Each consortium country has a share of the resources based on its contribution to the LUMI funding. The shares for each of the countries are allocated according to local considerations and policies, so LUMI is seen and handled as an extension of national resources. The LUMI shares belonging to the EuroHPC-JU are allocated by a peer-review process (comparable to that used for PRACE Tier-0 access). Figure 2 shows the LUMI consortium members.

Apart from the previously mentioned entities, individual countries in Europe provide and support HPC services to their respective researchers via national HPC centres or infrastructure managed via open research communities like universities.

There have been active attempts by EuroHPC-JU to increase the exposure of HPC to the existing member states. For example, the creation of new national competence centres (NCC) for HPC was taken up in the EuroCC[9] call. NCCs represent a focal point for HPC in the participating country, liaising with national initiatives in the area of HPC and facilitating access for national stakeholders to European HPC competences and opportunities in different industrial sectors and domains.

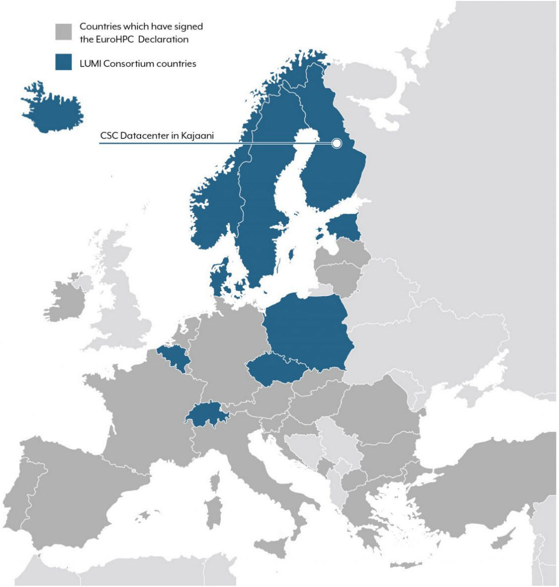

Countries which have signed the EuroHPC Declaration
LUMI Consortium countries
CSC Datacenter in Kajaani

Figure 2: The list of LUMI consortium members is displayed in the figure.

2.3 HPC Classifications

PRACE categorises European HPC facilities into three tiers: Tier-0 are European centres with petaflop machines; Tier-1 are national centres; and Tier-2 are regional centres. The resources under Tier-0 categorization are exclusively distributed via open-access calls and peer-review procedures. For Tier-1 HPC services, access is provided via DECI (Distributed European Computing Initiative), which is a programme under PRACE designed for research projects that require different resources from those currently available in the principal investigator's (PI) own country. At the same time, those projects should not require resources on the very largest (Tier-0) European supercomputers or very large CPU allocations. Another HPC category called Tier-3 exists to denote a university cluster. For example, Paderborn University's Noctua 1 [10] provides access to the members of Paderborn University.

2.4 HPC Calls

HPC resources are allocated either via open calls or by registering requests with the responsible authority via email or portals. In the case of national resources, i.e., Tier-1 HPC, which are linked to pan-European supercomputing infrastructure and ecosystems, the calls are divided into national and European calls. Researchers from universities, research institutions, and enterprises who match the eligibility conditions can utilise HPC

[8] https://www.lumi-supercomputer.eu/lumi-consortium/

[9] https://www.eurocc-access.eu/

[10] https://pc2.uni-paderborn.de/hpc-services/available-systems/noctua1/

services for free. An open-research agreement in which the results are made public serves as one of the most important prerequisites for free access to the resources. A third category of access to national resources exists in some cases where the work does not fall under the scope of the previous two types. In this case, the access is paid for, and the costs are calculated after analysis of the access request. Access to these systems is provided via open calls, where computing and data management resources are awarded through a peer review process[11]. Access to systems under the EuroHPC-JU is provided via open calls on the PRACE website.

2.5 Access Calls

2.5.1 PRACE

The following forms of access are available to PRACE systems:

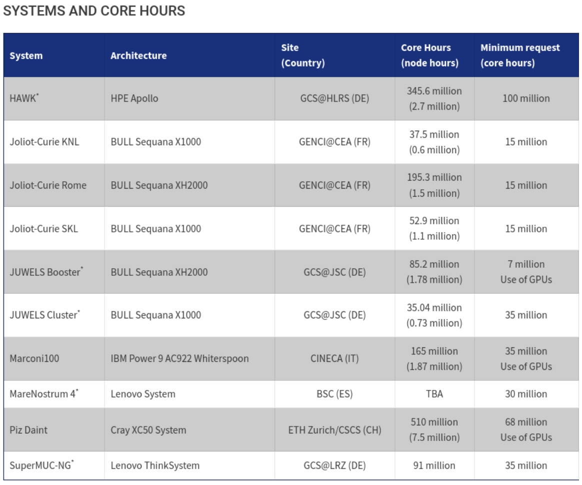
- Preparatory Access is intended for short-term access to resources, for code-enabling and porting, and is required to prepare proposals for Project Access and to demonstrate the scalability of codes. The Call for Proposals for PRACE Preparatory Access is a continuously open call, with cut-off dates every 3 months. Preparatory Access[12] allows PRACE users to optimise, scale, and test codes on PRACE Tier-0 systems before applying to PRACE calls for Project Access.
  - **Benchmark Access** is designed for code scalability tests, the outcome of which is to be included in the proposal in a future EuroHPC Extreme Scale and Regular call. Users receive a limited number of node hours; the maximum allocation period is three months.
  - **Development Access** is intended for projects centred on the development and optimisation of code and algorithms. Users will typically be allocated a few node hours; the allocation period is one year and is renewable up to two times.

- **SHAPE Access**[13]: suitable for SMEs with the potential to use HPC. This access mode aims to help SMEs benefit from the expertise and knowledge developed within PRACE RI.

- **Distributed European Computing Initiative** (DECI): Suitable for smaller-scale projects that do not require Tier-0 systems. This access mode provides Tier-1 users access to supercomputing architectures from another European country for smaller-scale projects. Proposal submissions are accepted in response to annual calls.

- **Project Access** is intended for individual researchers and research groups and is suitable for established Tier-0 users. Access can be granted for 1-year production runs, as well as for 2-year or 3-year (multi-year access) production runs. Proposal submissions are accepted in response to biannual calls.

- The PRACE ICEI[14] is open to all European researchers and research organisations needing resource allocations, regardless of funding sources.



SYSTEMS AND CORE HOURS

System	Architecture	Site (Country)	Core Hours (node hours)	Minimum request (core hours)
HAWK*	HPE Apollo	GCS@HLRS (DE)	345.6 million (2.7 million)	100 million
Joliot-Curie KNL	BULL Sequana X1000	GENCI@CEA (FR)	37.5 million (0.6 million)	15 million
Joliot-Curie Rome	BULL Sequana XH2000	GENCI@CEA (FR)	195.3 million (1.5 million)	15 million
Joliot-Curie SKL	BULL Sequana X1000	GENCI@CEA (FR)	52.9 million (1.1 million)	15 million
JUWELS Booster*	BULL Sequana XH2000	GCS@JSC (DE)	85.2 million (1.78 million)	7 million Use of GPUs
JUWELS Cluster*	BULL Sequana X1000	GCS@JSC (DE)	35.04 million (0.73 million)	35 million
Marconi100	IBM Power 9 AC922 Whiterspoon	CINECA (IT)	165 million (1.87 million)	35 million Use of GPUs
MareNostrum 4*	Lenovo System	BSC (ES)	TBA	30 million
Piz Daint	Cray XC50 System	ETH Zurich/CSCS (CH)	510 million (7.5 million)	68 million Use of GPUs
SuperMUC-NG*	Lenovo ThinkSystem	GCS@LRZ (DE)	91 million	35 million

Figure 3: PRACE - Call for Proposals for Project Access. The provided figure illustrates the specifications of the systems included in the call, encompassing information such as name, location, and the range of available computing resources, denoted by minimum and maximum limits.

Figure 3 shows the PRACE call for proposals with the minimum number of core hours to be requested.

2.5.2 EuroHPC-JU Access Modes

- Extreme Scale Access (one-year or two-year projects)
- Regular Access (single-year projects)
- Benchmark Access are designed for code scalability tests, the outcome of which is to be included in the proposal in a future EuroHPC-JU Extreme Scale and Regular call.
- Development Access is designed for projects focusing on code and algorithm development and optimisation.
- Fast Track Access for Academia
- Fast Track Access for Industry Access

The calls are announced on the PRACE website[15]. Applicants interested in applying to any of

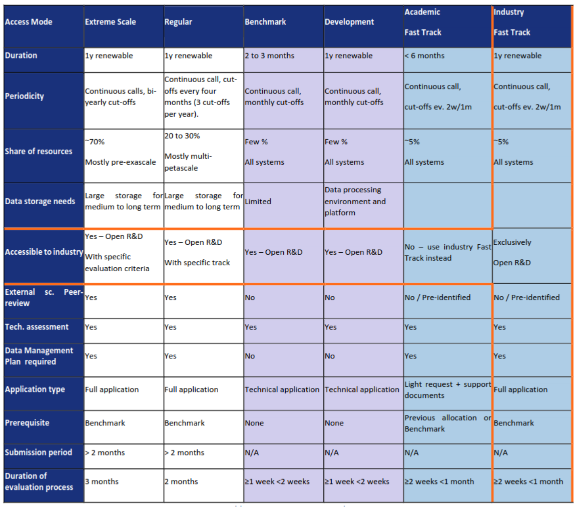
[11] https://prace-ri.eu/hpc-access/project-access/project-access-the-peer-review-process/

[12] https://prace-ri.eu/hpc-access/preparatory-access/preparatory-access-information-for-applicants/

[13] https://prace-ri.eu/hpc-access/shape-access/shape-access-information-for-applicants/

[14] https://prace-ri.eu/hpc-access/collaborative-calls/

[15] https://prace-ri.eu/hpc-access/eurohpc-access/



Access Mode	Extreme Scale	Regular	Benchmark	Development	Academic Fast Track	Industry Fast Track
Duration	1y renewable	1y renewable	2 to 3 months	1y renewable	< 6 months	1y renewable
Periodicity	Continuous calls, bi-yearly cut-offs	Continuous call, cut-offs every four months (3 cut-offs per year).	Continuous call, monthly cut-offs	Continuous call, monthly cut-offs	Continuous call, cut-offs ev. 2w/1m	Continuous call, cut-offs ev. 2w/1m
Share of resources	~70% Mostly pre-exascale	20 to 30% Mostly multi-petascale	Few % All systems	Few % All systems	~5% All systems	~5% All systems
Data storage needs	Large storage for medium to long term	Large storage for medium to long term	Limited	Data processing environment and platform		
Accessible to industry	Yes – Open R&D With specific evaluation criteria	Yes – Open R&D With specific track	Yes – Open R&D	Yes – Open R&D	No – use industry Fast Track instead	Exclusively Open R&D
External sc. Peer-review	Yes	Yes	No	No	No / Pre-identified	No / Pre-identified
Tech. assessment	Yes	Yes	Yes	Yes	Yes	Yes
Data Management Plan required	Yes	Yes	No	No	Yes	Yes
Application type	Full application	Full application	Technical application	Technical application	Light request + support documents	Full application
Prerequisite	Benchmark	Benchmark	None	None	Previous allocation or Benchmark	Benchmark
Submission period	> 2 months	> 2 months	N/A	N/A	N/A	N/A
Duration of evaluation process	3 months	2 months	≥1 week <2 weeks	≥1 week <2 weeks	≥2 weeks <1 month	≥2 weeks <1 month

Figure 4: EuroHPC-JU access modes. The diagram provides a comprehensive overview of the various elements pertaining to the access modes of EuroHPC-JU, including the duration of access, frequency of open calls, and prerequisites, among others.

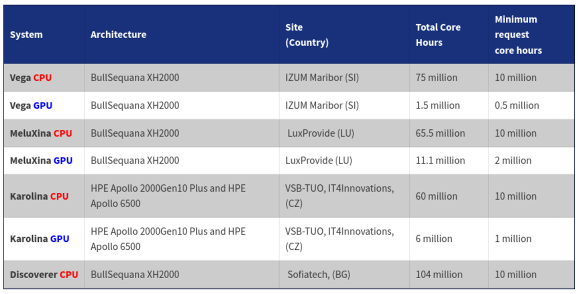
the EuroHPC-JU calls need to apply via the PRACE peer-review platform[16]. The project scope and plan are also required for regular and extreme-scale access. More information about eligibility, access tracks, peer-review process, and scoring criteria can be found in the EuroHPC-JU access section on the PRACE website[17][18][19]. Figure 4 summarises the various EuroHPC-JU access modes. Figure 5 shows the EuroHPC-JU call for proposal with the minimum number of core hours to be requested.

2.6 HPC Users

EuroCC-JU classifies all the eligible users into the following categories:

- Academic users
- Industrial users
- Public Research Institutes

Researchers from academia, research institutes, public authorities, and industry established or located in a Member State or in a country associated with Horizon 2020/Horizon Europe are eligible to apply. Access to commercial companies and public organisations is provided solely for open R&D purposes.



System	Architecture	Site (Country)	Total Core Hours	Minimum request core hours
Vega CPU	BullSequana XH2000	IZUM Maribor (SI)	75 million	10 million
Vega GPU	BullSequana XH2000	IZUM Maribor (SI)	1.5 million	0.5 million
MeluXina CPU	BullSequana XH2000	LuxProvide (LU)	65.5 million	10 million
MeluXina GPU	BullSequana XH2000	LuxProvide (LU)	11.1 million	2 million
Karolina CPU	HPE Apollo 2000Gen10 Plus and HPE Apollo 6500	VSB-TUO, IT4Innovations, (CZ)	60 million	10 million
Karolina GPU	HPE Apollo 2000Gen10 Plus and HPE Apollo 6500	VSB-TUO, IT4Innovations, (CZ)	6 million	1 million
Discoverer CPU	BullSequana XH2000	Sofiatech, (BG)	104 million	10 million

Figure 5: EuroHPC-JU - Call for Proposals for Regular Access Mode. The provided figure illustrates the specifications of the systems included in the call, encompassing information such as name, location, and the range of available computing resources, denoted by minimum and maximum limits

3 Methodology

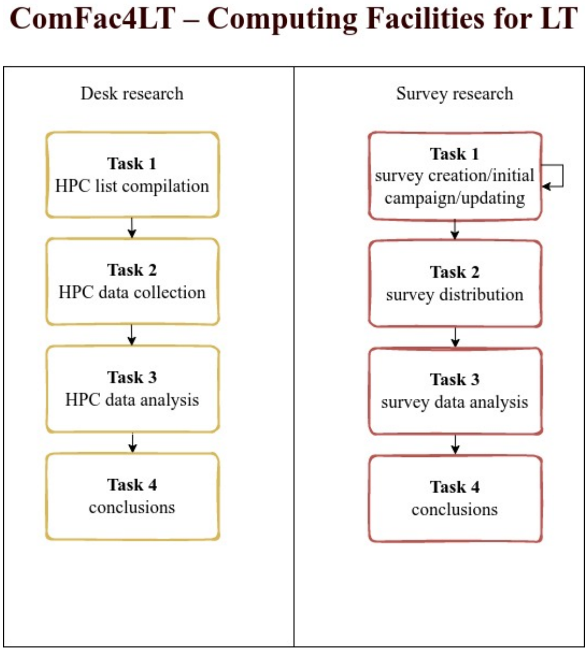


Figure 6: The methodology employed in this study encompasses two main approaches: desk-research and survey research.

The methodology used to study the various aspects related to the computing facilities for LT was performed using two distinct studies. The first part deals with desk research to study existing HPC facilities. This was accomplished by visiting each HPC's website and compiling information from the accompanying documentation, as well as contacting the appropriate authorities when necessary. This step allowed us

[16] https://pracecalls.eu/

[17] https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-ju-regular-access-mode/regular-access-applicant-information/

[18] https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-extreme-scale-access/extreme-scale-applicant-information/

[19] https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-ju-benchmark-development-access-calls/benchmark-development-applicant-information/

to concentrate on crucial aspects such as GPU specifications and access protocols. In the second part, a survey is conducted to study aspects related to HPC in practise. The survey captured the user's computational requirements as well as information about their computational facilities. The survey also gathered inquiries and comments from users regarding their existing HPC facilities. The survey method was chosen because it is critical to obtain up-to-date information on topics such as user hardware requirements and knowledge of EU-HPC initiatives.

3.1 Desk Research

A list of HPCs from the website Top500.org served as the seed list for the desk research. The Top500.org website publishes statistical lists of supercomputers twice a year. The website also includes metadata with the data releases, such as the HPC's location, ranking, and other hardware specifications. Two constraints were imposed to create the filtered list. First, only the European HPCs were retained. Second, HPCs belonging to the academic and research segments were retained. HPCs provided by vendor, private entities, others were not considered in the study as they did not relate to it directly. To this list, the supercomputers provided by EuroHPC-JU and PRACE (Partnership for Advanced Computing in Europe) were added. Given the nature of Tier-0 and Tier-1 HPCs being shared across EU member states and Horizon 2020-allied countries, we chose to focus on these supercomputers.

Desk research was conducted for each HPC in the fields listed below:

- Name
- Tier
- Performance (in petaflops)
- Location
- Hosting Institute
- HPC website link
- Types of access available to academic researchers, SMEs, and others
- Link to apply, register, or contact for the resources
- Manufacturer and specifications of GPU nodes, i.e., number of nodes, number of GPUs, and size of GPUs attached to each node
- Types of access provided as part of the institute: regular, benchmark, fast track, and others
- Additional notes or important points about the service

3.2 Online Survey

The survey, addressed to the LT researchers, sought to elicit the respondents' views to capture the real-world scenario. The survey had 11 questions in total. Two questions depended on previous answers. Table 1 shows an overview of the types of questions.

Table 1: Types of survey questions

Question types	Total
Closed	4
Open-ended	7
Total	11

None of the questions in the survey were mandatory.

- Part A. Respondent Profiling: The first part of the survey included questions for the demographic profiling of the respondents, with emphasis on
  - Country of the respondent
  - Active research field
  - Current role of the respondent
- Part B. HPC usage and requirements: assessed the current needs and usage patterns of computational facilities, i.e.,
  - HPC requirements for research and their specifics
  - Computational infrastructure
  - Hardware specifications in terms of processing time and memory
  - Awareness about Euro-HPC infrastructure
- Part C. Comments and suggestions: respondents' opinions, recommendations, and problems in relation to computational facilities

The survey was designed using Google Forms and underwent three iterations to capture verbose details. The survey was distributed by European Language Technology via a monthly newsletter in addition to mailing lists like META-NET-all, Corpora-List, MT-List, and In-Atala. The survey was open from March 7 to March 22. In total, 26 responses were collected. The responses collected as part of the survey, representing the views of the researcher in the field of LT, are analysed in the paper.

4 The HPC Landscape (desk research)

4.1 Analysis of HPCs

Given that performing language processing experiments necessitates the usage of GPUs, i.e., accelerated nodes, HPCs without GPUs were not considered. In total, the primary list contained 80 HPCs, but only 56 were analysed, as the ones that were filtered out were either offline or did not have GPUs.

4.1.1 Profile

In total, the final list accounts for 56 HPC from 20 EU member states, consisting of a mixture of various tiers. The distribution of tiers is depicted in Figure 7. The most common HPC level was Tier-1 (33) and then Tier-0 (15). This could be related to the predominance of EuroHPC-JU and PRACE supercomputers on the initial filtered list.

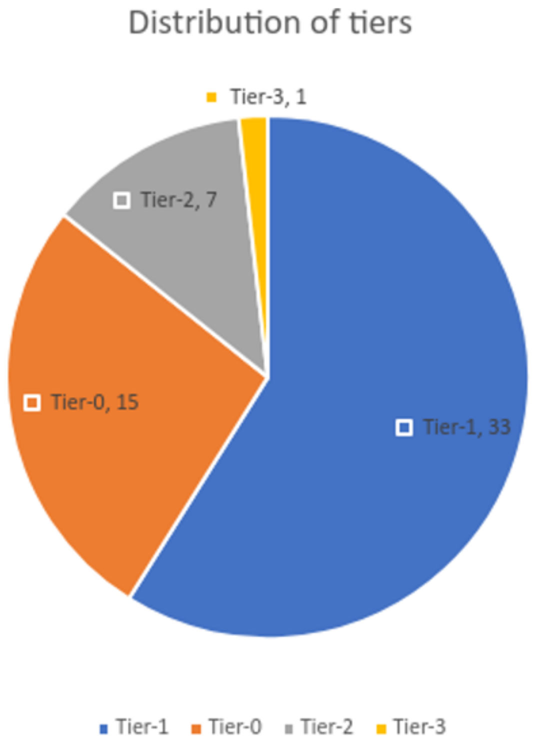


Figure 7: Distribution of tiers. This figure shows the distribution of tiers across different high-performance computing (HPC) systems.

The relationship between countries and their tiers is shown in Figure 8. Germany had HPCs in all three tiers. France, Italy, Czech Republic, and Spain had Tier-0 and Tier-1 HPCs. The number of HPCs from each country that were analysed is shown in Figure 9. According to the HPC list, Germany had the most systems, followed by Poland, Belgium, Sweden, and other countries.

4.1.2 Hardware Performance

The countries are listed along with their overall HPC performance in Figure 10. With a performance of more than 25 petaflops, Finland is host to the third-fastest HPC LUMI in the world. Italy and France are ranked second and third. It was a clear finding that members of PRACE HPC hosting nations have higher cumulative performance than non-hosting nations. It is important to know that HPCs that are part of a consortium, such as LUMI, allocate their resources based on each country's contribution share.

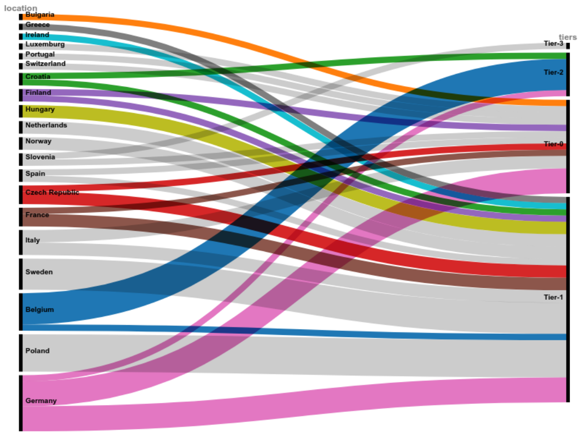


Figure 8: Country vs tiers. This diagram depicts the connections between the various tiers in various countries.

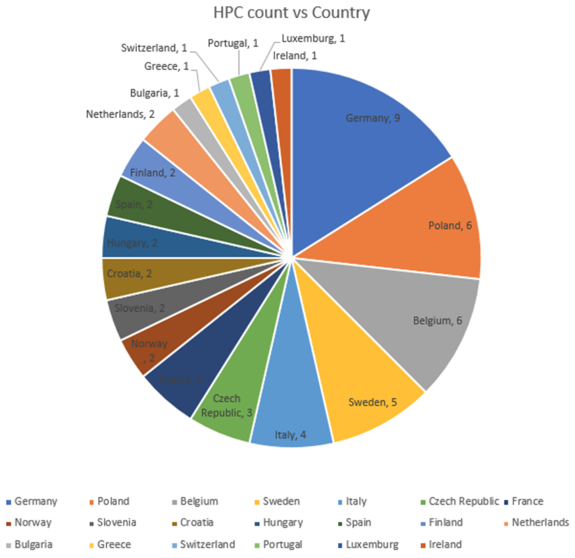


Figure 9: The figure depicts the countries, as well as the total number of HPCs that were included in the study.

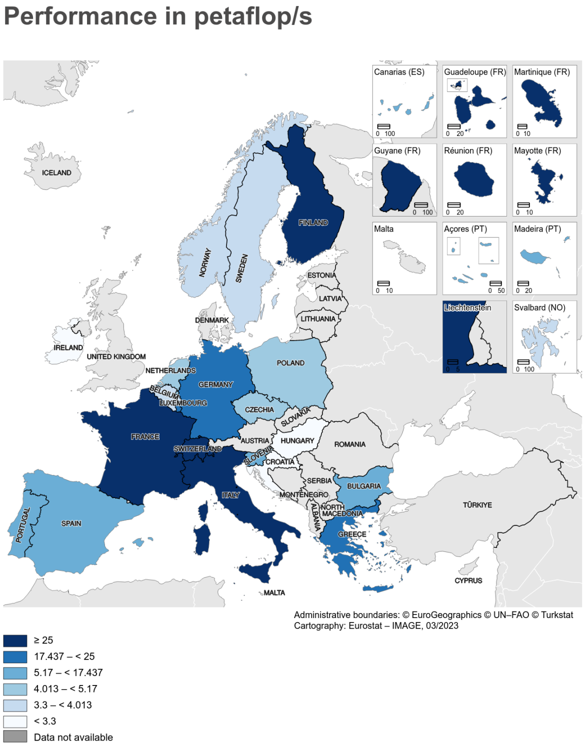


Figure 10: Countries and their cumulative performance. The figure presents focuses on the collective performance of various countries in terms of petaflops.

4.1.3 GPU Performance

This section analyses the GPUs used in HPC. In figure 12, we depict various GPU models. Nvidia V100 (28), A100 (23), and P100 were the most often installed graphics cards. There were six AMD Instinct cards in total. Nvidia CUDA with deep learning libraries such as PyTorch and TensorFlow enables easy access to GPU hardware, whereas ROCm is used to access AMD GPUs using such frameworks.

GPU cumulative VRAM is displayed per country in Figure 13. Finland, Italy, France, and Germany have the highest cumulative VRAM values. A comparison of the number of nodes vs. GPU cumulative VRAM is depicted in Figure 14.

4.1.4 HPCs Access

The HPC providers can also be grouped into the following categories:

- Open access to all researchers: An HPC provider grants open access to all the researchers linked to public research institutes like universities.
- Access to listed institutes: An HPC provider gives open access to universities or research institutes that have signed an agreement. For example, CSC's services[20] are free-of-charge for users affiliated with a Finnish higher education institution (universities, universities of applied sciences), or a state research institute. This is based on the agreement made with the Ministry of Education and Culture.
- Paid access or user contribution model: An HPC provider charges the users for the service based on the services being used.

In Figure 15, a breakdown of types of academic access is shown. The information shown pertains to non-PRACE access. In the vast majority of instances, academic users are granted free access.

Access to the industry can be classified into three categories. First, free access is granted if the research is publicly available and access is gained from the HPC service provider in the industry's respective country. Second, the HPC service provider does not offer access to industry users, but PRACE provides access. Third, access for commercial purposes is available to industrial users. The breakdown of access for industry is shown in Figure 16.

There is also strategic and discretionary access allocated for emergency-related work, such as research on pandemics such as COVID-19. Commercial access is typically an option for users that wish to utilise HPC services. Figure 17 depicts the breakdown of the different other types of access.

4.1.5 HPCs Access Calls

The calls to access computational resources for an HPC can be divided into national and international calls. International calls are handled through PRACE or a publicly funded project[21]. For the national-level calls, all the HPCs provide electronic means of submitting applications. The applications are usually accompanied by a project proposal listing requirements like a detailed plan of experiments, benchmarking scores, hardware needs, etc. The project proposal is then subjected to a technical feasibility assessment and scientific review.

The access call for HPCs can be classified as follows:

- **Experimental, benchmarking, or testing call**: The calls are usually open throughout the year and processed in a stepwise fashion. The call provides access to hardware with a few hours of computation

[20] https://research.csc.fi/free-of-charge-use-cases

[21] https://ni4os.eu/

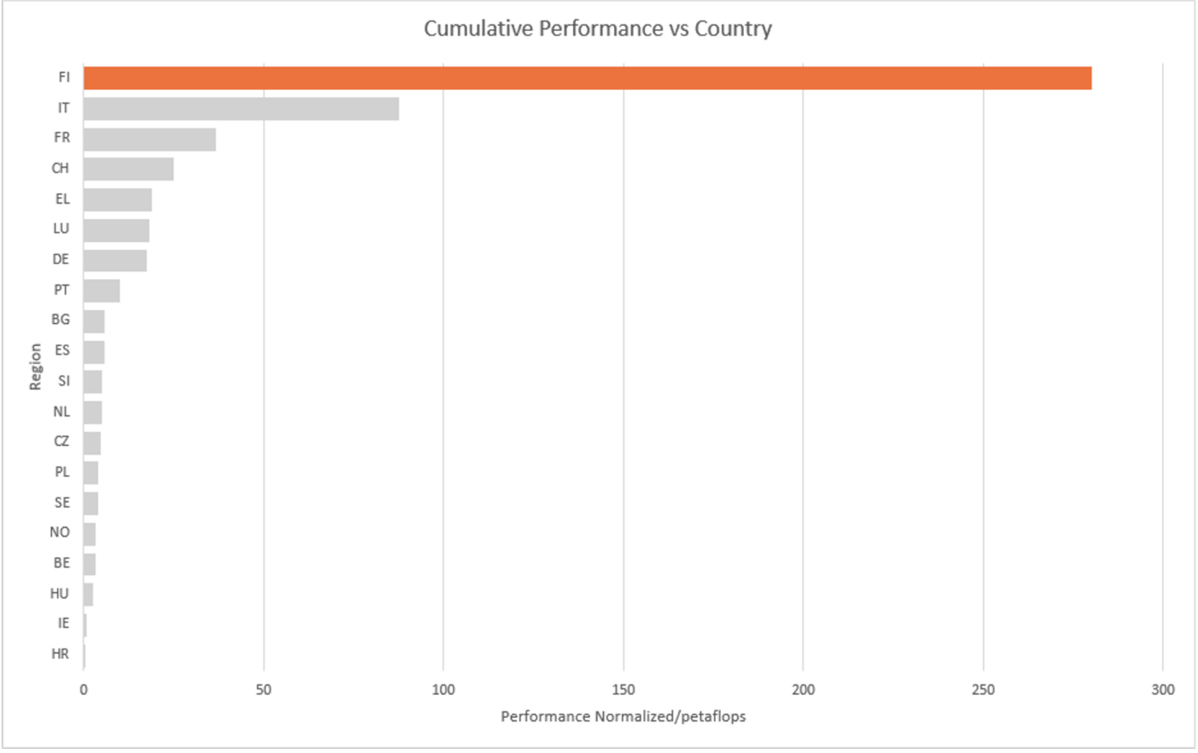


Figure 11: Countries and their cumulative performance

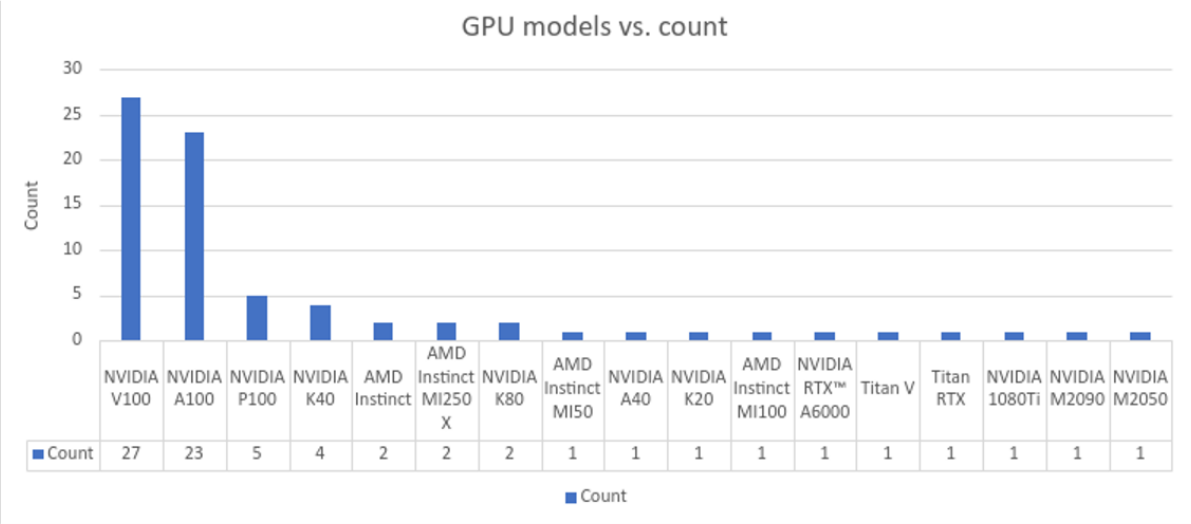
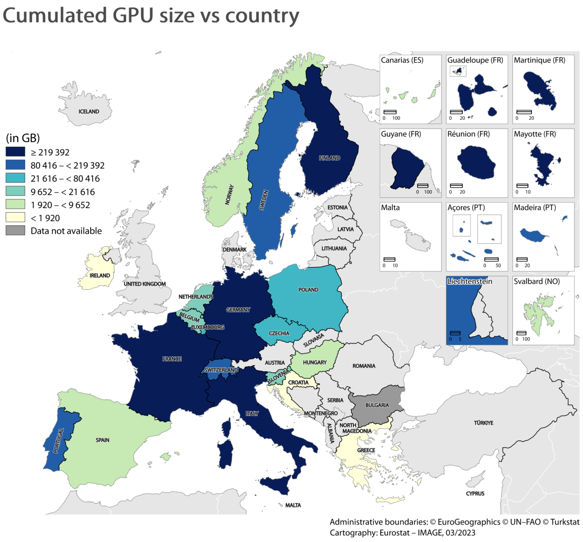


Figure 12: Model distribution of GPUs. The graph depicts the various GPU models and their counts, illustrating current GPU preference trends.



Administrative boundaries: © EuroGeographics © UN–FAO © Turkstat
Cartography: Eurostat – IMAGE, 03/2023

Figure 13: Countries and their cumulative GPU VRAM. Figure shows the availability of GPU memory in BG across various countries.

resources to test and compute the performance of the experiments. The benchmarking figures are then used in the application of regular access.

- **Fast track calls**: The calls are targeted towards users for projects that need fast access to HPC resources, which is limited in time and smaller in resources when compared with regular access call projects.
- **Regular calls**: The calls are opened for projects needing high-performance computational resources. The projects can last from 9 to 12 months, and calls are typically opened 2–4 times per year. Depending on the HPC service provider, requests can also be processed continuously. If the allocated resources are depleted, they can be extended.
- **Large-scale**: The call is similar to a regular access call but requires resources over a longer duration of time. Empirical estimates suggest that it could be more than 2% of total resources of the full HPC setup, computed over a year. The runtime of the projects ranges from 1 year to 3 years.
- **Director's Discretion/Discretionary Access**: A portion of the computational resources are reserved and made available upon project approval. An application can be submitted at any time. The computational resources are allocated irregularly based on evaluation by the management.
- **Extreme-scale**: The call is for the sectors to justify the need for and capacity to use extremely large allocations in terms of compute time, data storage, and support resources.

Each call has a processing period, which is the time it takes to look at the proposal and come to a decision. After this time, the applicants are given the resources they asked for.

4.1.6 Dynamic Access

The eDARI portal is used to request resource hours at French national computing centres. The portal allows two types of access to resources.

- Regular access
- Dynamic access

Depending on the number of requested hours, the requested access will be either Dynamic Access or Regular Access. If the number of hours requested is <= 50,000 GPU hours (and/or 500,000 CPU hours), it will be Dynamic Access (AD). If the amount is larger than these values, the request will be considered Regular Access (AR). The Dynamic Access skips the need for additional supplementary details. Requests for resources for Dynamic Access files may be made throughout the year and are renewable. Two project calls for Regular Access are launched each year.

This mode of access is discussed in a distinct section, since its general accessibility is so conducive to research. This access mode is a great choice due to the streamlined procedure and minimal documentation, particularly for CPU and GPU hours with moderate demands.

4.2 Comparison Study

In order to clarify the points made in the previous section, let us examine two countries in greater depth. Consider these two nations: Croatia, France.

At the time the study was conducted, two HPCs were publicly listed for Croatia, one for Tier 1 (SRCE) and the other for Tier 2 (BURA). There was one Tier 0 (Joliot-Curie IRENE) and two Tier 1s for France (Jean Zay, Adastra). The cumulative hardware performance of the French HPC exceeded 110 petaflops, while the cumulative performance of the Croatian HPC was close to 0.42 petaflops. The total GPU VRAM accessible in France remained at 273152 GB (number of GPU = 4616), whereas its Croatian counterpart recorded 480 GB (number of GPU = 24). In addition, France's LT research benefits from having dynamic access to a vast quantity of resources. It is critical to highlight that access to the Croatian HPC service (SRCE) is easier via the application portal. This disparity in the availability of HPC resources to researchers must be addressed if we are to realise the aim of language equality.

4.3 Summary

Figure 19 shows a summary of all the current HPC services that academic researchers and small and medium-sized enterprises (SME) can use. Users should use their local resources if they are available and appropriate. If the requirements are not very high, in most local HPC centres (Tier-2 and Tier-3), there is no need to write a

[22] https://www.edari.fr/schema/acces/ressource

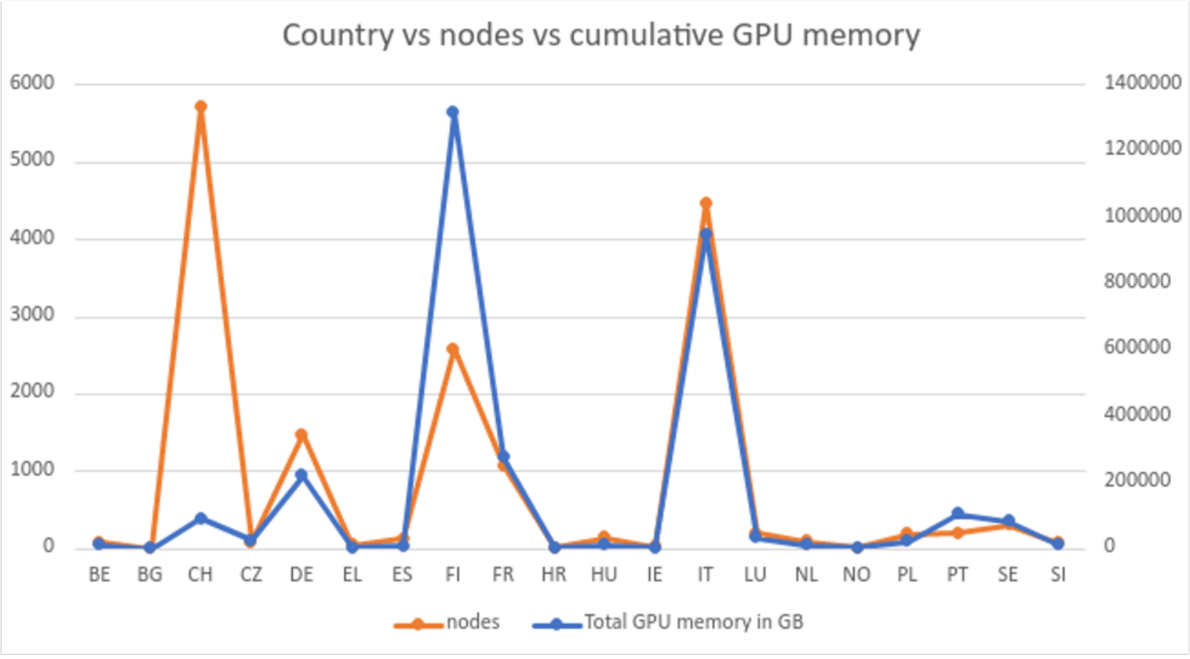


Figure 14: Countries with their total node count and GPU VRAM. The provided figure illustrates the quantity of nodes and GPU memory allocated to each country, effectively representing the sizes of their high-performance computing (HPC) nodes.

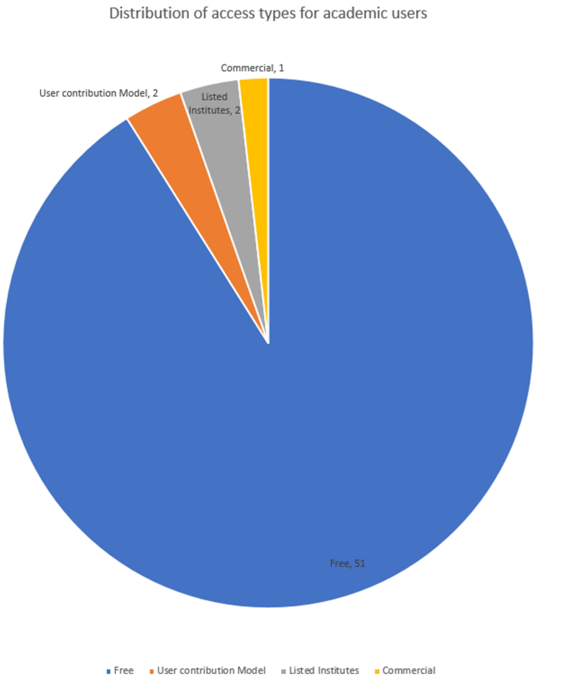


Figure 15: Distribution of different types of academic access. The majority of academic users' access is free of charge.

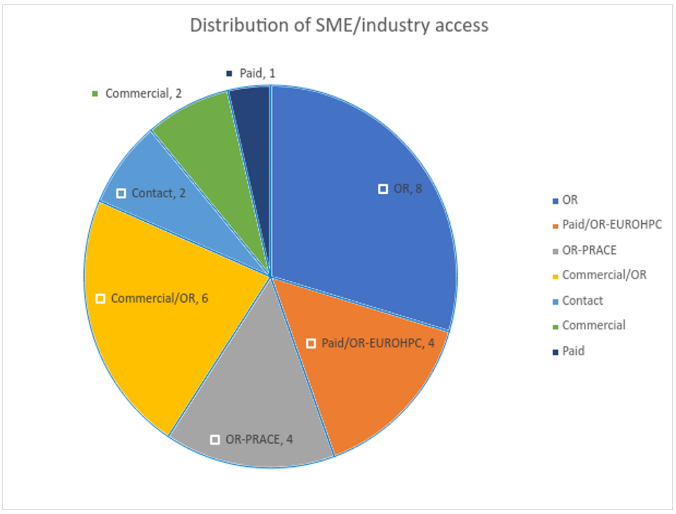


Figure 16: Distribution of different types of access for SME/industry users. The majority of SME/Industry users' access is paid. The free access is provided when SME users are involved in open research.

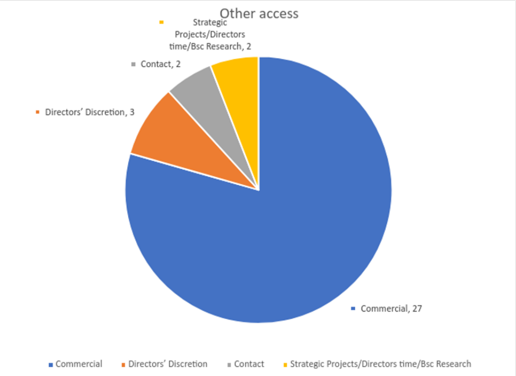


Figure 17: Distribution of other types of access. Access to the High-Performance Computing (HPC) service is exclusively granted by the governing authorities, specifically to users who have submitted a specialised application.

detailed application, as all the researchers are provided with fair-share quota access. If you require more resources than your centre can provide, and you don't have a local HPC centre, or you identify special needs (e.g., larger memory, more Cores/CPU, GPUs), you may contact another HPC centre or apply for compute time at a higher level (e.g., Tier-2/Tier-1). Only very experienced users with well-scaling codes and high demands on compute time should apply for large-scale projects on the Tier-1/Tier-0 level. In any case, you can contact your local HPC support with your queries. If the research is open, which means that the results will be available to the public after publication, a researcher from industry can work with academics or apply to the resources on their own through PRACE. In the case where research is private, the option of commercial access to resources provided by private vendors and other HPC providers is available.

5 Analysis

5.1 Survey Responses

5.1.1 Respondents' Profile

The majority of the answers came from European countries, except for a few. States covered via the survey include: Croatia, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, Malta, Portugal, Spain, the UK, Pakistan, the USA, Ukraine, Russia.

Table 2 shows the breakdown of answers.

The majority of respondents had LT as the active area of research, with 25 actively associated with NLP and 1 marking themselves as not active NLP researchers. Table 3 shows the respondents associated with the area of NLP/LT.

Table 2: The geographical distribution of the survey participants

Country	Count	Country	Count
Croatia	2	Denmark	2
Finland	3	France	5
Germany	1	Ireland	1
Italy	1	Luxembourg	1
Malta	1	Pakistan	1
Portugal	1	Russia	1
Spain	2	UK	1
Ukraine	1	USA	1

Table 3: Participant's active area of research. "Is NLP/LT your active area of research".

Answers	Total
Yes	25
No	1

Most of the individuals who responded are either academic researchers or students. One respondent identified himself as a public sector researcher. No responses from the researchers working in industry were received. The breakdown of respondent associations is shown in Table 4.

Table 4: Respondents' role. "What is your current role ?".

Roles	Total
Researcher-Academia	17
Student	8
Researcher/developer - Public sector	1

5.1.2 Respondents' LT Infrastructure and Requirements

The majority of the survey respondents reported using HPCs for their experiments. The responses can be classified into three types.

- HPC users
- Cloud service users like Google Colab
- Local hardware or personal computer users

Respondents who didn't use an HPC said they did the experiments on a single GPU or a setup with more than one CPU. When asked how many GPU hours were needed, the answers ranged from "it depends on the experiment" to a precise number that suggested a certain number of hours per day, week, or month. Regarding the multi-GPU requirement, most respondents wished to use more GPUs, especially for the task of machine translation (text and speech). 50% replied they do not have the multi-GPU requirement, while 50% reported they do wish to use more GPUs. Another question was posed regarding memory requirements, and a variety of responses were provided.

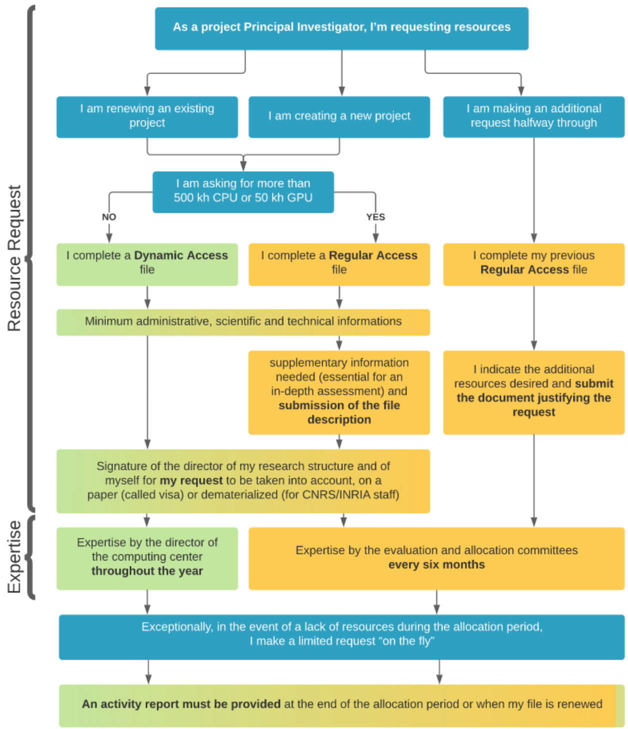

As a project Principal Investigator, I'm requesting resources
I am renewing an existing project
I am creating a new project
I am making an additional request halfway through
I am asking for more than 500 kh CPU or 50 kh GPU
NO
YES
I complete a Dynamic Access file
I complete a Regular Access file
I complete my previous Regular Access file
Minimum administrative, scientific and technical informations
supplementary information needed (essential for an in-depth assessment) and submission of the file description
I indicate the additional resources desired and submit the document justifying the request
Signature of the director of my research structure and of myself for my request to be taken into account, on a paper (called visa) or dematerialized (for CNRS/INRIA staff)
Resource Request
Expertise
Expertise by the director of the computing center throughout the year
Expertise by the evaluation and allocation committees every six months
Exceptionally, in the event of a lack of resources during the allocation period, I make a limited request "on the fly"
An activity report must be provided at the end of the allocation period or when my file is renewed

Figure 18: Reference card for access via edari.fr [22]. The provided diagram serves as a reference for individuals who are seeking to apply for High-Performance Computing (HPC) services.

5.1.3 EuroHPC-JU Usage

On the question of using the resources from EuroHPC-JU, the majority of answers suggested not using EuroHPC-JU. 57% reported never hearing about it, while 34.6% responded, negating the use of the service.

Table 5: EuroHPC-JU usage. "Have you used resources from EuroHPC-JU?".

Answer	Total
Yes	2
No	9
Never heard of it	15

5.1.4 Respondents' Comments, Suggestions, and Recommendations

The open-ended question capturing the comments and suggestions with respect to the computational facilities used by users presented multiple aspects, which can be described in the following points.

- one of the respondents said, "Euro HPC applications are extremely heavyweight and not fit for our field". Aspects like "time to solution" and terminology, like simulation, which relates to the field of physics, used during the application process introduce non-conformity.
- users' access to HPC is temporary and linked to a project.
- unavailability due to the number of GPUs and more users.
- opacity with respect to job scheduling

5.2 Summary

Although only 26 responses were received, the study can be considered a preliminary step in mapping the current HPC facilities available to the LT researchers. The country respondents, such as those from Finland, France, Italy, and Germany, use the HPC resources collected and analysed during desk research. The majority of respondents had GPU requirements of fewer than 100 hours per month on average. Concerning, however, is the lack of knowledge of the European-level HPC services available to users. The conditions set by the responders do not correspond to the minimum amount of hours required to be requested in the PRACE calls.

6 Conclusions

HPC services enable the solution of computational tasks at a rate exponentially greater than that of a desktop computer. These services have existed and have proven to be crucial in advancing the state of the art in LT. There are numerous projects that provide HPC nowadays, like EuroHPC-JU, PRACE, LUMI, national

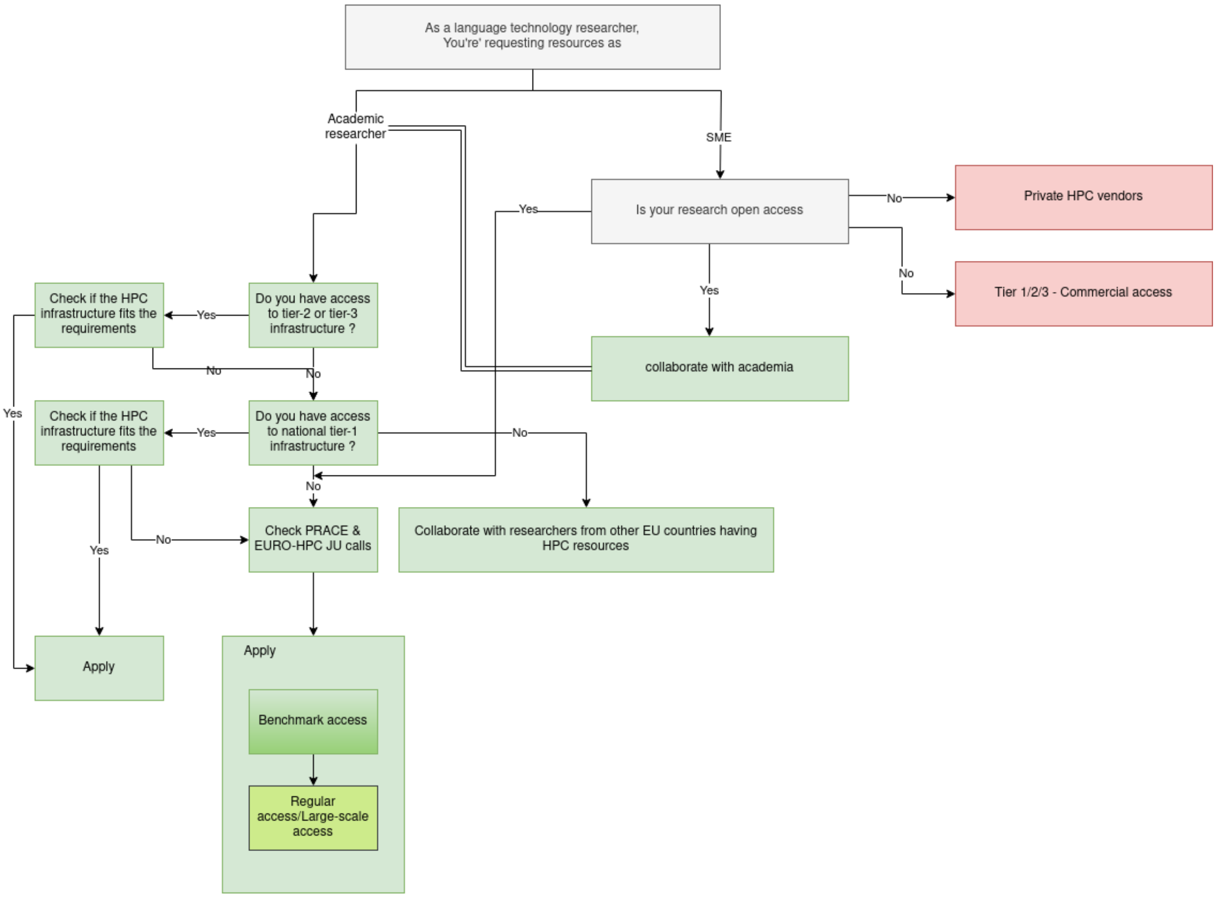

As a language technology researcher,
You're' requesting resources as
Academic researcher
SME
Is your research open access
No
Private HPC vendors
No
Tier 1/2/3 - Commercial access
Yes
Yes
collaborate with academia
Check if the HPC infrastructure fits the requirements
Yes
Do you have access to tier-2 or tier-3 infrastructure ?
No
No
Yes
Check if the HPC infrastructure fits the requirements
Yes
Do you have access to national tier-1 infrastructure ?
No
No
Yes
No
Check PRACE & EURO-HPC JU calls
Collaborate with researchers from other EU countries having HPC resources
Apply
Apply
Benchmark access
Regular access/Large-scale access

Figure 19: Overall summary. The provided diagram compiles all the aspects studied in the desk-research and serves as a reference for individuals who are seeking to apply for High-Performance Computing (HPC) services.

consortia, etc. Yet, for the subject of LT, unified research on many facets of HPCs was required.

In this paper, we give an overview of HPC services for LT research. We focused on elements such as available hardware, access types, and the requirements associated with each access type. In addition, we provide a simple reference card to be followed while seeking HPC services.

Before going into detail with all the conclusions of our analysis, we emphasise two points that, in our opinion, will be particularly critical to ensuring digital language equality in Europe:

- HPCs are important for LT research and development. Thus, competence with HPCs is a fundamental requirement for language equality[23].
- Availability of HPC for smaller and larger requirements is crucial. Users can access the European, national, and regional HPC services. Thus, efforts should be focused on facilitating easier and quicker access for these users.

Next, we present a summary of the key insights and recommendations regarding HPC services and access to them in the context of digital language equality in Europe.

- **Access to HPC resources for light-weight requirements**. As seen previously, the PRACE and EuroHPC-JU calls demand very high minimum node hours in the request. Although EuroHPC-JU offers academic fast-access, these calls are difficult to get. As previously indicated, dynamic access is an excellent solution for providing easy and quick access to requirements that are not demanding in terms of node hours. Hence, we suggest an access mode similar to dynamic access to speed up the process of resource request and allocation.
- **Collaboration within the EU community and SMEs**. An alternate way of accessing HPC resources from EU countries is through collaboration. The LT community should provide the required tools for collaboration, particularly with nations lacking HPC resources. This would be advantageous not only for academic researchers, but also for industry users.
- **Centralise access to HPC related information**. Websites like https://atlas-cric-dev.cern.ch/core/rcsite/list/ and https://gauss-allianz.de/en/hpc-ecosystem give centralised information on HPCs accessible in the country, including Tier classification and hardware specifications. Our desk study helps move in this direction, but we recommend a centralised website that would allow

[23] More about the European Language Equality project you can find at https://www.european-language-equality.eu.

users to locate and filter HPC prospects based on requirements and particular criteria.

- **Hardware absence == No LT exploration**. Last but not least, a relatively uniform image can be obtained from the survey of LT users about HPC usage. We can fairly assume that the LT researchers will stick to tasks that fit their current hardware availability rather than anything else. For instance, if a GPU is capable of fine-tuning a model, a researcher is more likely to pursue fine-tuning than machine translation or language modelling. Even in the case where the minimum hardware is available, users fiddle with hyperparameters like batch-size to finish training. This does increase the overall time required as compared to using an HPC service. Another issue related to this point is the capacity to execute inference[24] on LLMs such as BLOOM. As the model needs 352 GB in bf16 (bfloat16) weights (176*2), the most efficient set-up is 8×80 GB A100 GPUs. Also, 2×8×40 GB A100s or 2×8×48 GB A6000 can be used. The inability to employ these LLM models without access to numerous GPUs does provide a challenge. Mosaic ML[25] makes it easy to train a billion parameter models in hours instead of days, with no lock-in to a single vendor and coordination across multiple clouds. With the ability to scale across multiple providers, the OOM can be prevented. We recommend such an infrastructure be realised in the context of EuroHPC-JU and PRACE systems to cater to the dynamic needs of various NLP tasks from different strata of LT users.

Finally, HPC services are available at multiple levels (regional, national, and European). At the same time, addressing the challenges related to the availability and accessibility of HPC is of the utmost priority. Our hope is that this document provides enough overview and insights into existing HPC resources available to academia and industry. We also hope our recommendation in the final section will provide enough pointers for the stakeholders to plan and implement future steps effectively. Private providers such as Azure and Amazon are significant players in the LT market because they offer commercial access to a huge number of GPUs. This type of access is feasible with sufficient funds. This element was not addressed in this study; it's expected to be the subject of future research.

7 Limitations

Following are some of the study's immediate limitations:

- The list of HPCs is not exhaustive. For the desk research, a curated list of HPCs was compiled from the websites top500.org, PRACE, and EuroHPC-JU. Hence, the majority of the analysed systems were either Tier-0 or Tier-1, with fewer Tier-2 and even fewer Tier-3 systems. Thus, our observations and reasoning may be influenced by the HPCs analysed. Covering the EuroHPC-JU and PRACE systems does provide a European-level perspective, but country-specific observations cannot be confirmed with the same degree of certainty. During the time of data compilation, a number of new HPC systems[26] became operational and were not included in the analysis.
- The survey was intended to reach a larger audience, but due to time constraints, the sample size was insufficient, making it difficult to generalise findings to a larger group.
- In comparison to the number of LT researchers in the EU, the size of the survey's sample is significantly smaller. There were no answers from industry researchers, who may have had different computational needs or perspectives on HPC usage and access.

Acknowledgments

The work presented in this paper has received funding from the European Union under the grant agreements No. LC-01641480 – 101018166 (ELE) and No. LC-01884166 – 101075356 (ELE 2).

References

Ahmed, N. M. and Wahed, M. (2020). The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *ArXiv*, abs/2010.15581.

Asch, M., Moore, T., Badia, R., Beck, M., Beckman, P., Bidot, T., Bodin, F., Cappello, F., Choudhary, A., De Supinski, B., et al. (2018). Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *The International Journal of High Performance Computing Applications*, 32(4):435–479.

Berberich, F., Liebmann, J., Nominé, J.-P., Pineda, O., Segers, P., and Teodor, V. (2019). European hpc landscape. In *2019 15th International Conference on eScience (eScience)*, pages 471–478. IEEE.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language

[24] https://huggingface.co/blog/bloom-inference-pytorch-scripts

[25] https://www.mosaicml.com/platform

[26] https://www.tportal.hr/tehno/clanak/u-zagrebu-predstavljeno-najjace-super-racunalo-u-hrvatskoj-pokrenuta-nova-generacija-nacionalne-e-infrastrukture-hr-zoo-foto-20230328

models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dolbeau, R. (2018). Theoretical peak flops per instruction set: a tutorial. *The Journal of Supercomputing*, 74(3):1341–1377.

Eicker, N. et al. (2020). Performance monitoring and analysis of eurohpc supercomputers. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '20).*

Hutton, R. et al. (2019). Prace: A european hpc ecosystem. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '19).*

Jiang, Z., Gao, W., Tang, F., Wang, L., Xiong, X., Luo, C., Lan, C., Li, H., and Zhan, J. (2021). Hpc ai500 v2. 0: The methodology, tools, and metrics for benchmarking hpc ai systems. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 47–58. IEEE.

Jiang, Z., Gao, W., Wang, L., Xiong, X., Zhang, Y., Wen, X., Luo, C., Ye, H., Lu, X., Zhang, Y., et al. (2019). Hpc ai500: a benchmark suite for hpc ai systems. In *Benchmarking, Measuring, and Optimizing: First BenchCouncil International Symposium, Bench 2018, Seattle, WA, USA, December 10-13, 2018, Revised Selected Papers 1*, pages 10–22. Springer.

Lathrop, S., Mendes, C., Enos, J., Bode, B., Bauer, G., Sisneros, R., and Kramer, W. (2019). Best practices for management and operation of large hpc installations. *Concurrency and Computation: Practice and Experience*, 31(16):e5069.

Luszczek, P., Dongarra, J. J., Koester, D., Rabenseifner, R., Lucas, B., Kepner, J., McCalpin, J., Bailey, D., and Takahashi, D. (2005). Introduction to the hpc challenge benchmark suite. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

OpenAI (2023). GPT-4 technical report. *CoRR*, abs/2303.08774.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., and et al. (2022). BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Skordas, T. (2019). Toward a european exascale ecosystem: the eurohpc joint undertaking. *Communications of the ACM*, 62(4):70–70.

Truong, H.-L., Fahringer, T., Madsen, G., Malony, A. D., Moritsch, H., and Shende, S. (2001). On using scalea for performance analysis of distributed and parallel programs. In *Proceedings of the 2001 ACM/IEEE conference on Supercomputing*, pages 34–34.

Wolter, N., McCracken, M. O., Snavely, A., Hochstein, L., Nakamura, T., and Basili, V. (2006). What's working in hpc: Investigating hpc user behavior and productivity. *CTWatch Quarterly*, 2(4A):9–17.