EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

ComFac4LT – Computing Facilities for LT

Authors	Gaurish Thakkar, Marko Tadić, German Rigau
Organisation	University of Zagreb, Faculty of Humanities and Social Sciences

Dissemination level	Public
Date	31-03-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	ComFac4LT – Computing Facilities for LT
Authors	Gaurish Thakkar, Marko Tadić, German Rigau
Organisation	University of Zagreb, Faculty of Humanities and Social Sciences
Type	Report
Number of pages	39
Status and version	Final
Dissemination level	Public
Date of delivery	31-03-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2023 ELE2 Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AI4EU	AI4EU (EU project, 2019-2021)
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CRACKER	Cracking the Language Barrier (EU project, 2015–2017)
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DLE	Digital Language Equality
EC	European Commission
ECSPM	European Civil Society Platform for Multilingualism
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality

ELE2	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELEXIS	European Lexicographic Infrastructure
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
ELT	European Language Technology
EP	European Parliament
ERIC	European Research Infrastructure Consortium
ESCO	European Skills, Competences, Qualifications and Occupations classification
EuroHPC JU	The European High Performance Computing Joint Undertaking
GDPR	General Data Protection Regulation
GPU	Graphic Processing Unit
HPC	High Performance Computing
KPI	Key Performance Indicator
LT	Language Technology/Technologies
LUMI	Large Unified Modern Infrastructure
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NCP	National Contact Point
NLP	Natural Language Processing
PRACE	Partnership For Advanced Computing in Europe
SME	Small and Medium Enterprise
SRIA	Strategic Research and Innovation Agenda
STOA	Science and Technology Options Assessment

Abstract

High-performance computing is an essential service required in multiple research areas. This report performs desk-research and an online survey to map the HPC landscape available to LT researchers. The report enumerates the various infrastructures available as well as presents the views of the LT researchers obtained via an online survey. In Section 2, background knowledge related to HPC is presented. This information is segmented into six sections and describes HPC along with associated terminologies, HPC classification schemes, HPC initiatives, and aspects related to HPC access and users. In Section 3, methodology related to desk research and an online survey is described. In Section 4, the HPC landscape is described using the information obtained from desk research. In Section 7, responses from the online survey are analysed. The report concludes with recommendations, conclusion, and limitations.

1. Introduction

This deliverable summarizes the overall landscape of high-performance computing available to users/researchers in academia as well SME/industry sector. The results will serve as input for a strategic research, innovation, and deployment agenda (SRIA) and roadmap, in order to tackle the striking imbalance between European languages in terms of the support they receive through LTs by 2030.

Language Technology (LT) is a highly researched field, having high socio-economic impacts. The analysis processes applied to the text enable knowledge gain and strategic decision-making. The knowledge extracted from the text have been largely attributed to the advances in the field of natural language processing. Over the years, the field of language processing has been researched to solve the task of classification, language generation and many more. 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 architectures, which has led to progress in the field of language technology. One recent aspect associated with 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 boom of AI-based data driven techniques for NLP is the data crunching ability using an efficient hardware in the form of Graphic Processing Units (GPUs). In neural language model training, the cost component realised in the form of hardware and its operation. This directly results in organisations with access to these hardware resources having access to research and development of LT technologies (Ahmed and Wahed, 2020).

BLOOM (Scao et al., 2022) is a big 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 Language Technology (LT) researchers, conducted by the EU project European Language Equality 2 (ELE2). In addition to the different HPC infrastructures that are available, the report 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. The overall objective of obtaining digital language equality can be attained with the aid of high-performance infrastructure. For research and innovation in the field of LT, a comprehensive understanding of the available infrastructure alternatives is essential.

The report has been developed in the frame of the European Language Equality (ELE) 2 project. The ELE 2 project develops a strategic research, innovation, and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

2. Background

2.1. High Performance Computing – HPC

HPC, also called a supercomputer, provides the opportunity to solve complex problems in different applications[2]. 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 computations that require floating-point calculations[3]. 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 gigaflops (GFLOPS) computer system is capable of performing one billion (109) floating-point operations per second.
- A 1 teraFLOPS (TFLOPS) computer system is capable of performing one trillion (1012) floating-point operations per second.
- A 1 petaflops (PFLOPS) computer system is capable of performing one quadrillion (1015) floating-point operations per second.
- A 1 exaflops (EFLOPS) computer system is capable of performing one quintillion (1018) floating-point operations per second.

An HPC solution is made up of the following components[4]:

- 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 above-mentioned components, there are accelerated nodes, i.e., computer nodes with GPUs or any other accelerator like a Xeon Phi[5]. At the time of writing this report, **Frontier**[6]

[2] https://www.ff4eurohpc.eu/en/about/what-is-hpc/
[3] https://en.wikipedia.org/wiki/FLOPS
[4] https://www.netapp.com/data-storage/high-performance-computing/what-is-hpc/
[5] https://en.wikipedia.org/wiki/Xeon_Phi
[6] https://en.wikipedia.org/wiki/Frontier_(supercomputer)

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).

2.2. HPC initiatives in Europe

The European High Performance Computing Joint Undertaking (**EuroHPC-JU**) 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 easing access to potential HPC users like SMEs and HPC experts across Europe. To date, five supercomputers[7] 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 and **Karolina** in the Czech Republic. Two supercomputers are underway: **Deucalion** in Portugal, **Supek** in Croatia, and **MareNostrum5** in Spain. The list of EuroHPC-JU public members can be found in Appendix A. Figure 1 depicts the location of EuroHPC-JU public members.

Figure 1: EuroHPC-JU member states and associated countries.

PRACE[8] (Partnership for Advanced Computing in Europe) is a not-for-profit international association that aims to facilitate the 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 [9]

[7] https://digital-strategy.ec.europa.eu/en/policies/high-performance-computing-joint-undertaking
[8] https://prace-ri.eu/
[9] https://prace-ri.eu/about/members/

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 5 PRACE members (BSC representing Spain, CINECA representing Italy, ETH Zurich/CSCS representing Switzerland, GCS representing Germany, and GENCI representing France). Figure 2 shows PRACE member countries.

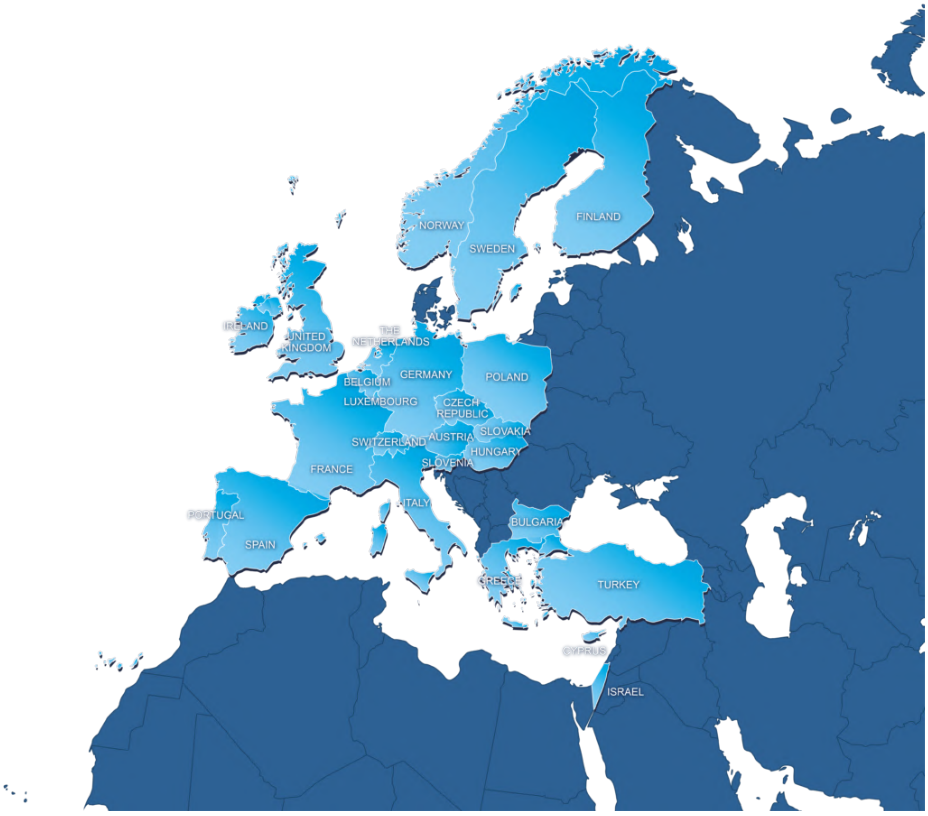


Figure 2: PRACE members

LUMI consortium [10] 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 3 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.

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

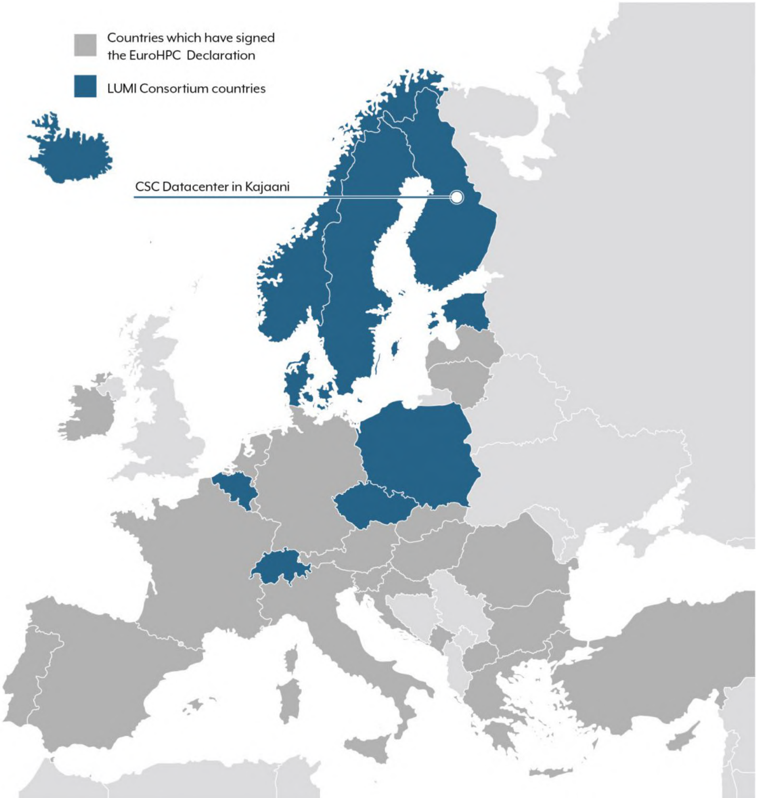


Figure 3: LUMI consortium members

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[11] call. NCCs represent a focal point for HPC in the participating country, liaising with national initiatives in the area of HPC, facilitating access of national stakeholders to European HPC competences and opportunities in different industrial sectors and domains.

2.3. HPC classifications

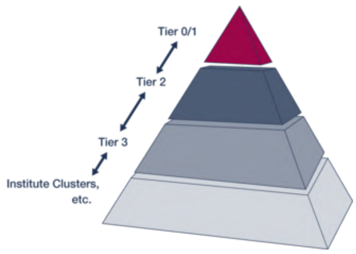


Figure 4: HPC hierarchy

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 allocations of CPU. Another HPC category called Tier-3 exists to denote a university cluster. For example, Paderborn University's Noctua 1 [12] provides access to the members of Paderborn University.

2.4. HPC calls

The 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 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 eligibility of 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. The access to these systems are provided via open calls, where computing and data management resources are awarded through a peer review process[13]. The access to systems under EuroHPC JU are provided via open calls on the PRACE website.

[11] https://www.eurocc-access.eu/
[12] https://pc2.uni-paderborn.de/hpc-services/available-systems/noctua1/
[13] https://prace-ri.eu/hpc-access/project-access/project-access-the-peer-review-process/

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, 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[14] 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[15]: suitable for SMEs with the potential of using HPC. This access mode aims to help SMEs benefit from the expertise and knowledge developed within the 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. The 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[16] program is open to all European researchers and research organizations needing resource allocations regardless of funding sources.

Figure 5 shows the PRACE call for proposal 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.

[14] https://prace-ri.eu/hpc-access/preparatory-access/preparatory-access-information-for-applicants/
[15] https://prace-ri.eu/hpc-access/shape-access/shape-access-information-for-applicants/
[16] https://prace-ri.eu/hpc-access/collaborative-calls/

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 5: PRACE - Call for Proposals for Project Access

- Fast Track Access for Academia
- Fast Track Access for Industry Access

The calls are announced on the PRACE website[17]. Applicants interested in applying to any of the EuroHPC-JU calls need to apply via the PRACE Peer-Review Platform[18]. 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 PRACE website[19]-[20]-[21]. Figure 6 summarises the various EuroHPC-JU access modes. Figure 7 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 are eligible to apply. Access to commercial companies and public organisations is provided solely for open R&D purposes.

3. Methodology

The methodology to study the various aspects related to the computing facilities for language technology was performed using two distinct studies. The first part deals with desk research to study existing HPC facilities. In the second part, a survey is conducted to study the aspects related to HPC in practice. 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.

3.1. Desk research

A list of HPC 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 segment 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

17 https://prace-ri.eu/hpc-access/eurohpc-access/
18 https://pracecalls.eu/
19 https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-ju-regular-access-mode/regular-access-applicant-information/
20 https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-extreme-scale-access/extreme-scale-applicant-information/
21 https://prace-ri.eu/hpc-access/eurohpc-access/eurohpc-ju-benchmark-development-access-calls/benchmark-development-applicant-information/

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 6: EuroHPC-JU access modes

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 7: EuroHPC-JU - Call for Proposals for Regular Access Mode

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/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 so as 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.

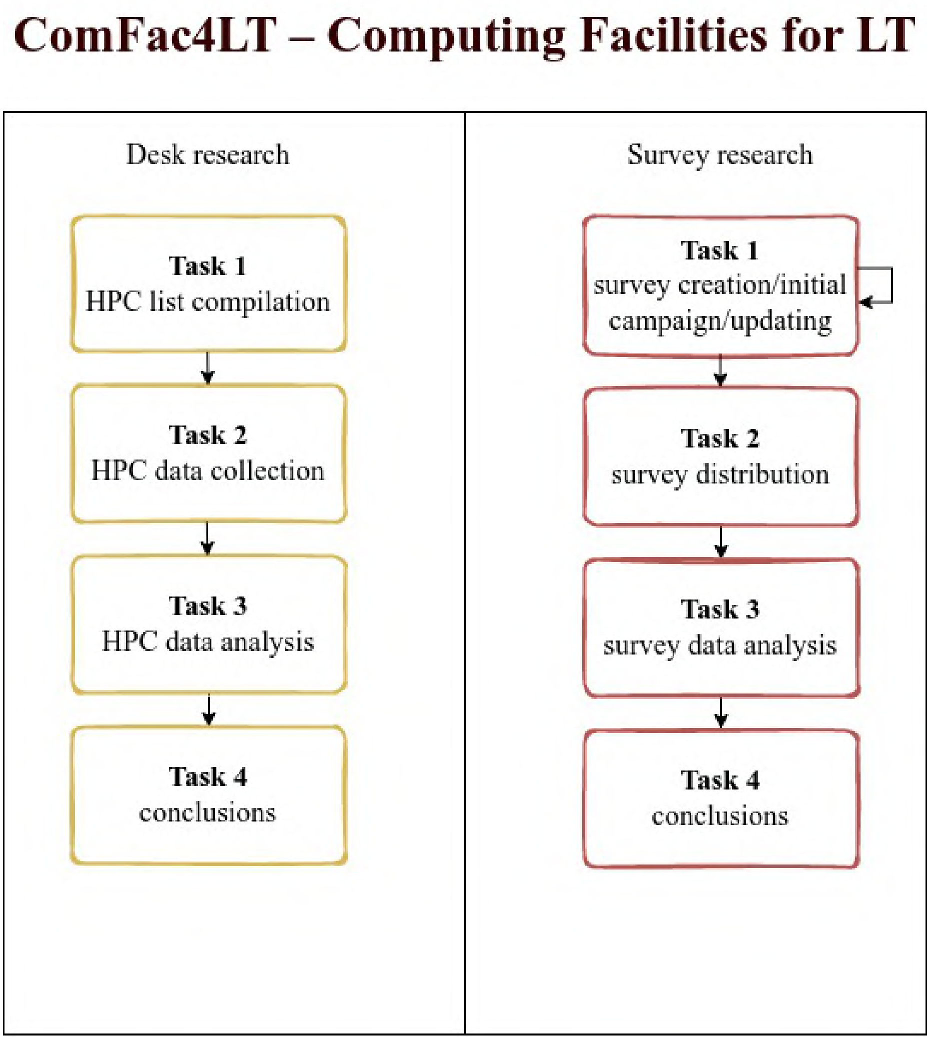


Figure 8: Methodology

Question types	Total
Closed	4
Open-ended	7
Total	11

Table 1: Types of survey questions

None of the questions in the survey were mandatory.

- Part A. Respondent profiling: the first part of the survey included question 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/problems in relation to computational facilities

The survey was designed using Google Forms and underwent three iterations to capture verbose details. The full survey, as published online, is presented in Appendix A. The survey was distributed by European Language Technology via a monthly newsletter in addition to mailing lists like META-NET-all, Corpora-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 language technology, are analysed in report.

4. The HPC landscape (desk research)

4.1. Analysis of HPCs

Given the requirement for performing language processing experiments necessitates the presence of GPUs, i.e., accelerated nodes, HPCs not having 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 9. 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.

The relationship between countries and their tiers is shown in Figure 10. 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 11. According to the HPC list, Germany had the most systems, followed by Poland, Belgium, Sweden, and other countries.

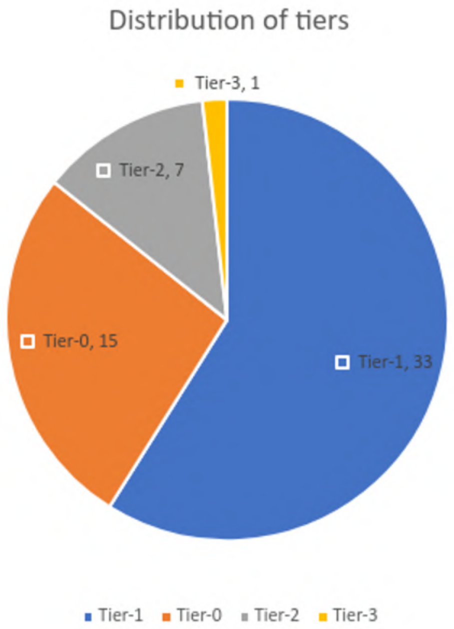


Figure 9: Distribution of tiers

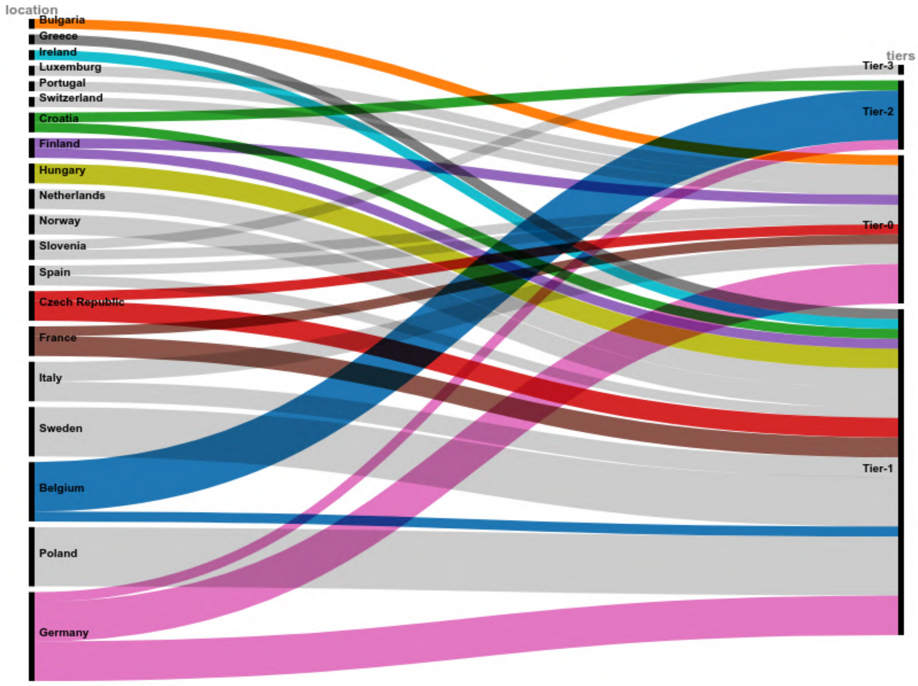


Figure 10: Country vs tiers

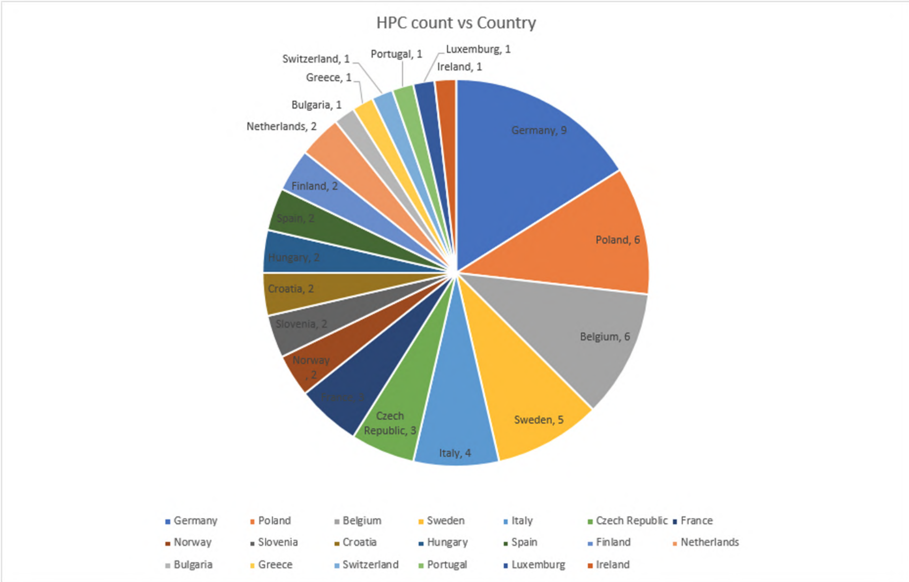


Figure 11: Countries and the number of HPC present in the study.

4.1.2. Hardware performance

The countries are listed along with their overall HPC performance in Figure 12. 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.

4.1.3. GPU performance

This section analyses the GPUs used in HPC. In figure 14, 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 15. Finland, Italy, France, and Germany have the highest cumulative VRAM values. A comparison of number of nodes vs GPU cumulative VRAM is depicted in Figure 16.

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[22] are free-

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

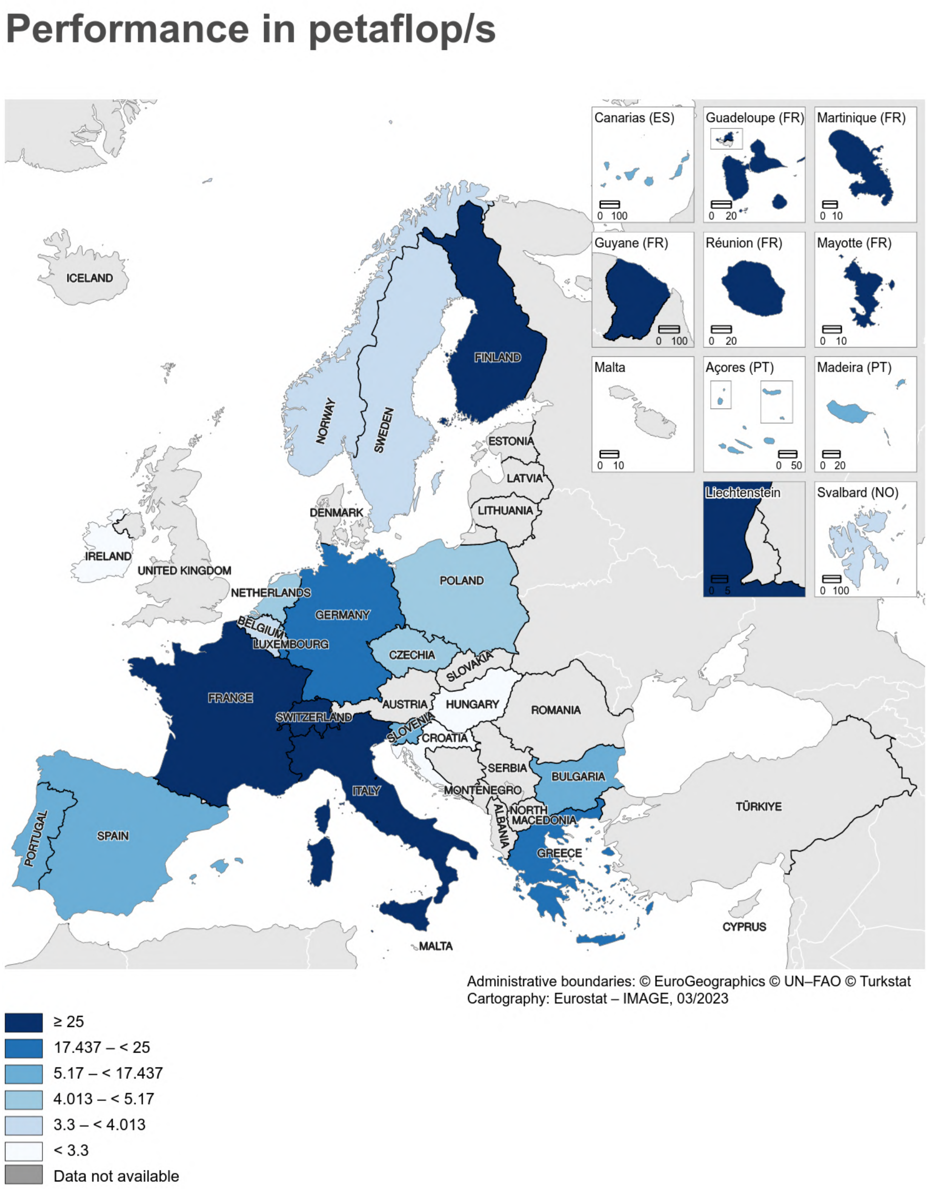


Figure 12: Countries and their cumulative performance

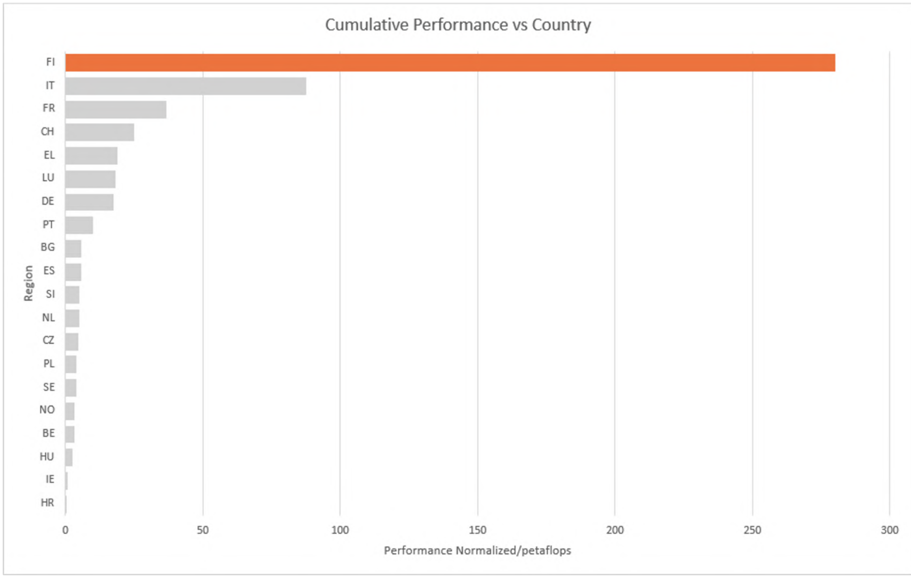


Figure 13: Countries and their cumulative performance

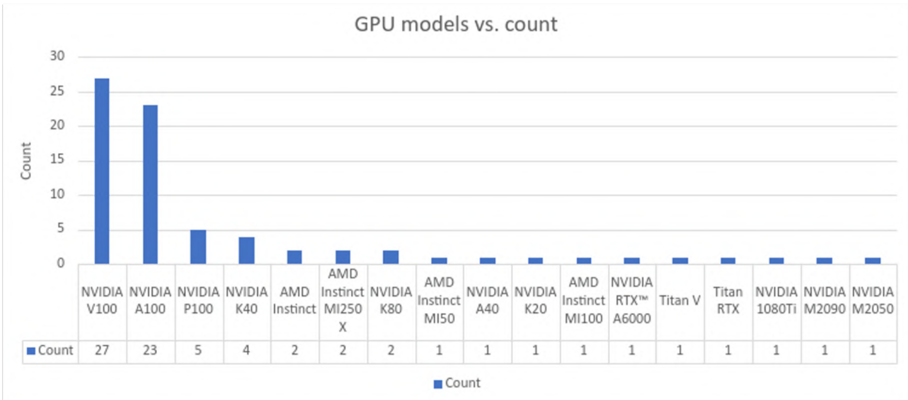


Figure 14: Distribution of GPU models

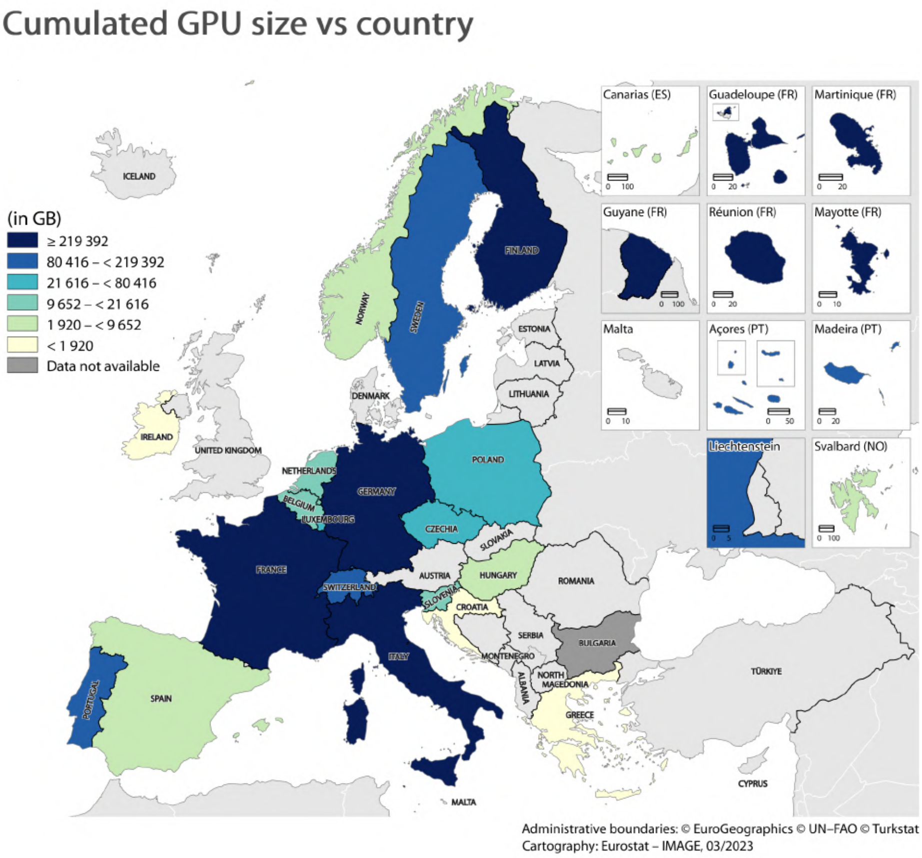


Figure 15: Countries and their cumulative GPU VRAM

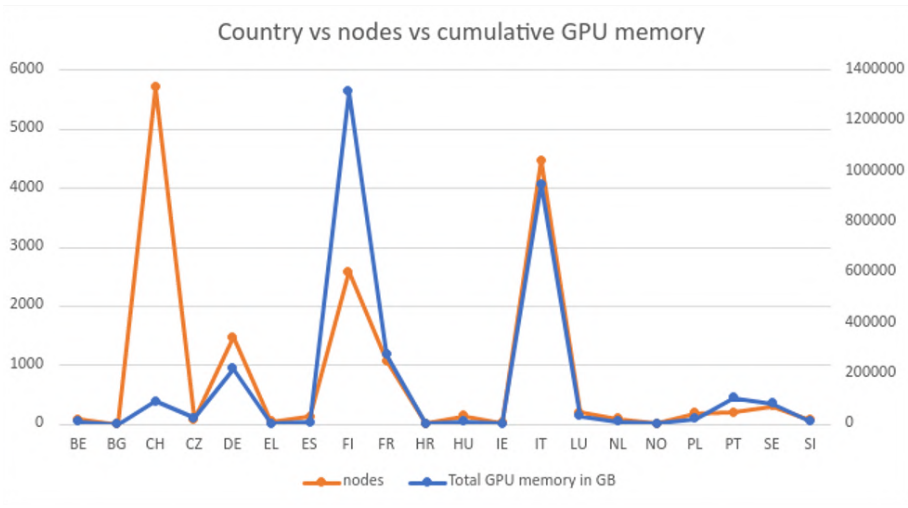


Figure 16: Countries with their total node count and GPU VRAM

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

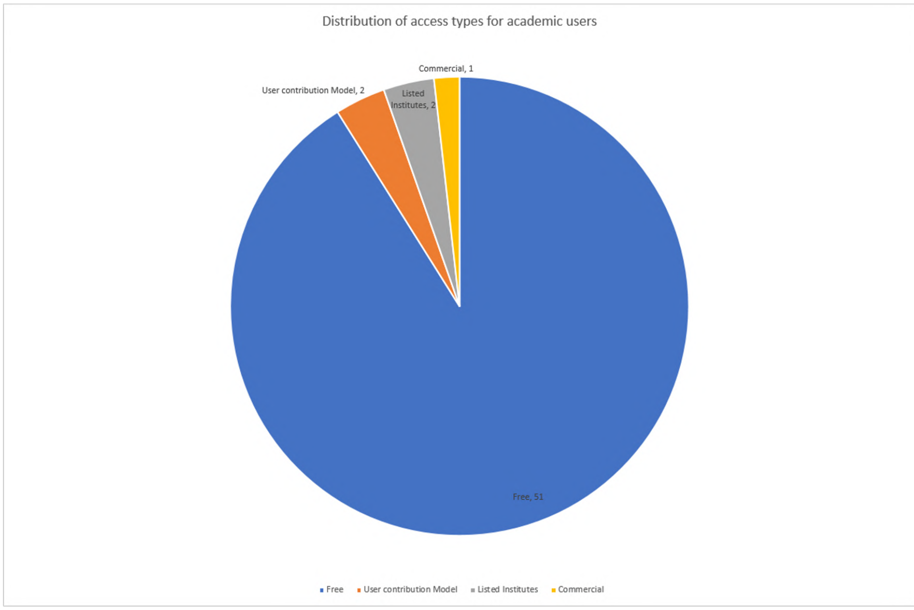


Figure 17: Distribution of different types of academic access

Access to the industry can be classified into three categories. First, free access 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 to industrial users. The breakdown of access for industry is shown in Figure 18.

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 19 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[23]. 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.

[23] https://ni4os.eu/

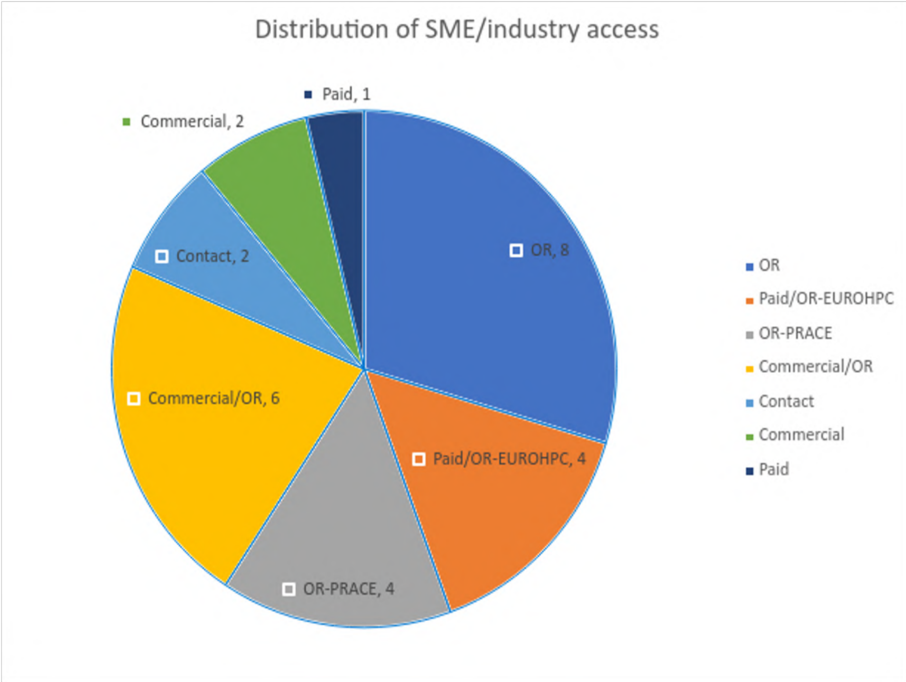


Figure 18: Distribution of different types of access for SME/industry users

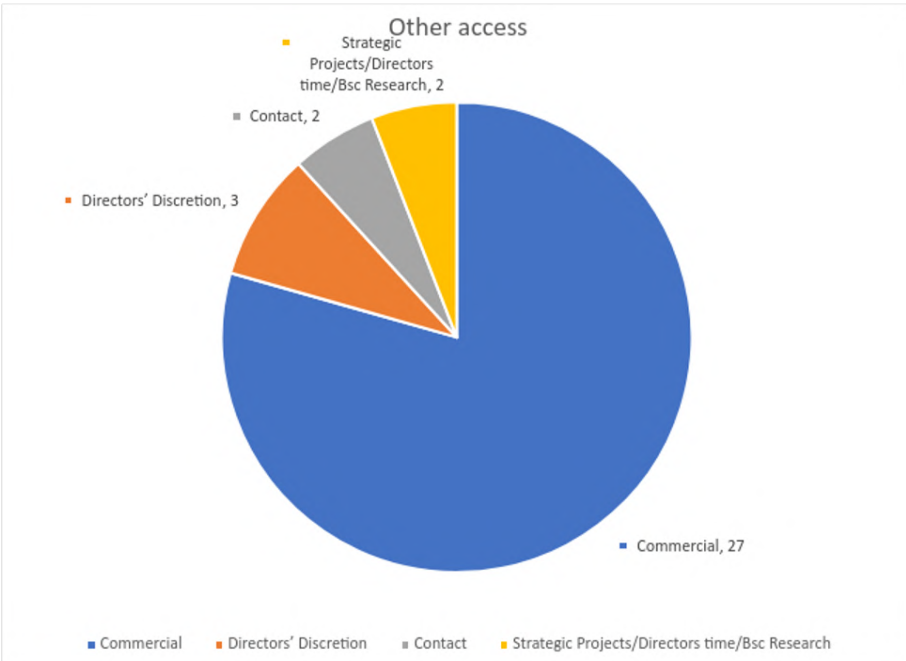


Figure 19: Distribution of other types of access

Analysing all the HPCs in the list, the calls for access can be broadly divided into types:

- Experimental/benchmarking/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 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 in continously. 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 lasts 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.

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

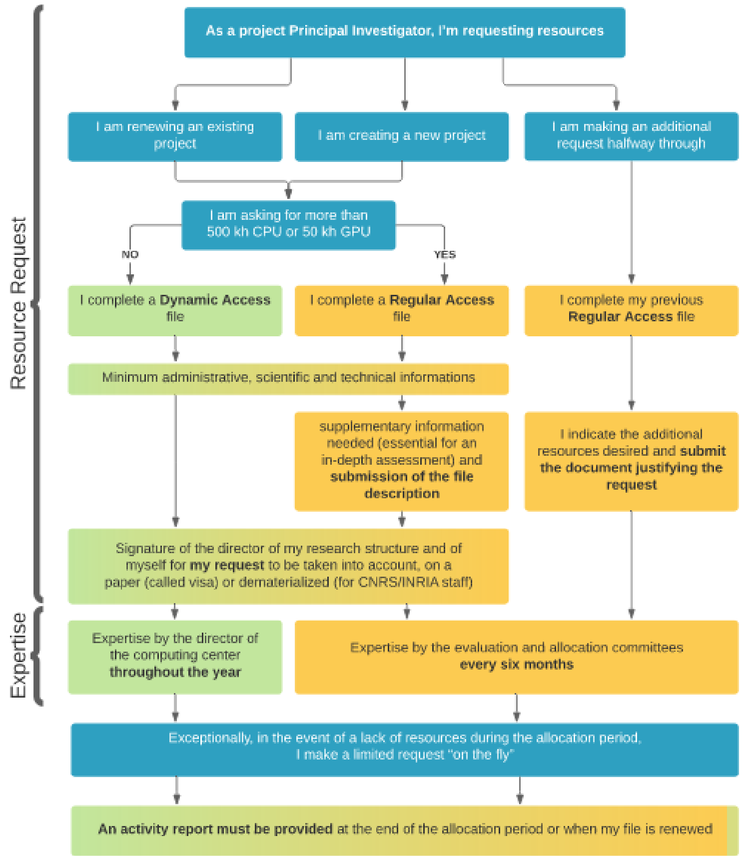


Figure 20: Reference card for access via edari.fr [24]

4.2. Comparison study

Let's examine two countries in greater detail to clarify the points presented in the previous section. Consider the two countries listed below:

- Croatia
- France

At the time of writing this report, 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 French HPC exceeded 110 petaflops, while the cumulative performance of Croatian HPC was close to 0.42 petaflops. The total GPU VRAM accessible in France remained at 273152 GB, whereas its Croatian counterpart recorded 480 GB. In addition, France's LT research benefits from having dynamic access to a vast quantity of resources. It is vital 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

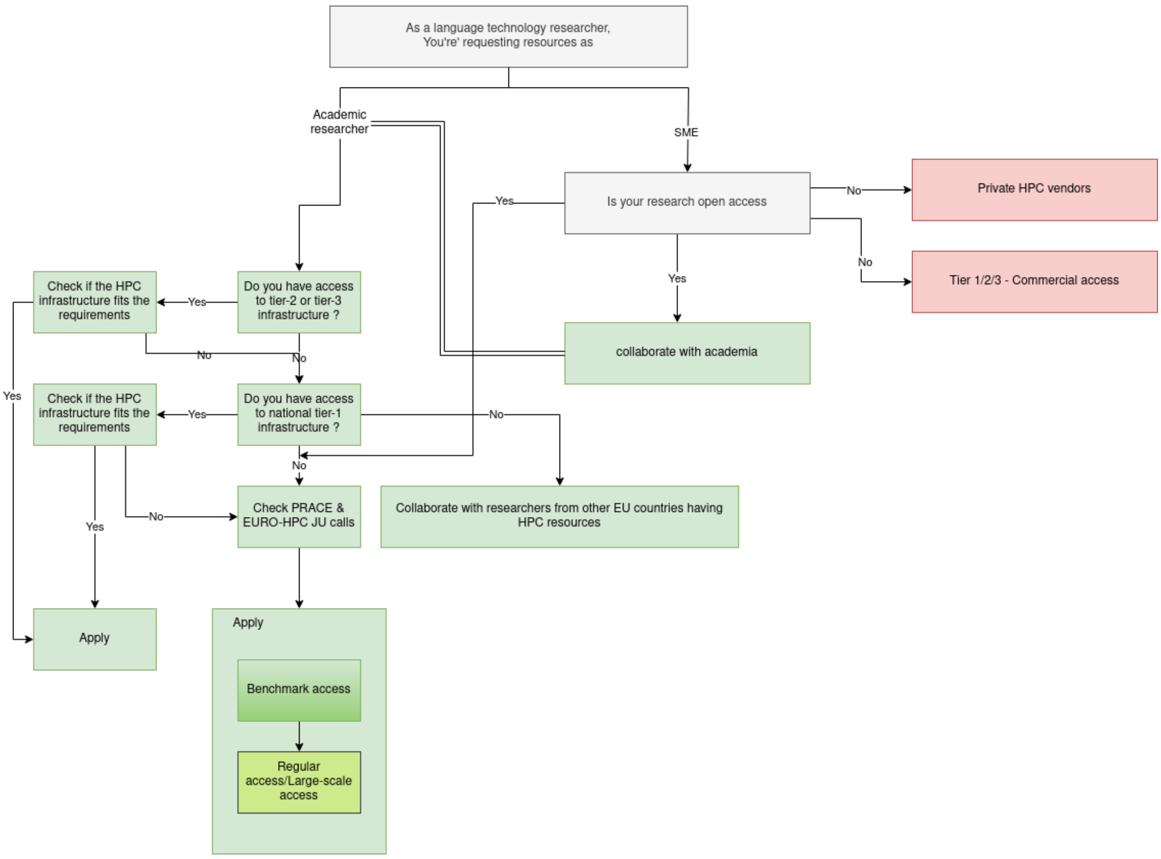


Figure 21: Overall summary

Figure 21 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 detailed application, as all the researchers are provided with fair-share quota access. If you require more resources than your centre can provide, 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.

Figure 22 shows the breakdown of answers.

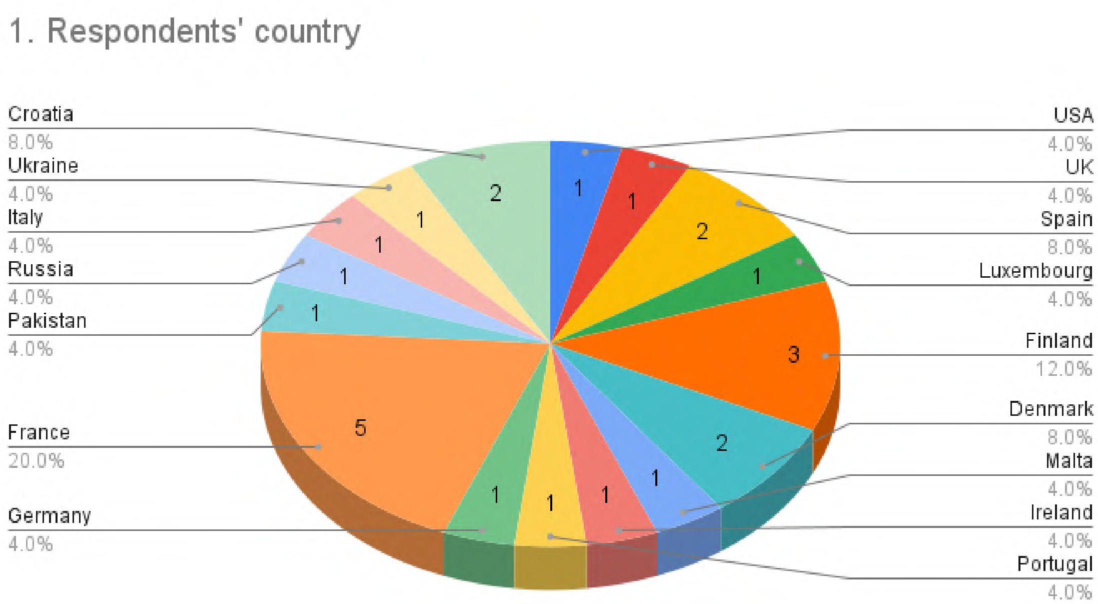


Figure 22: Location

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. Figure 23 shows the respondents associated with the area of NLP/LT.

Figure 23: Active area of research

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 Figure 24.

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 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 GPU, 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 GPU. Another question was posed regarding memory requirements, and a variety of responses were provided.

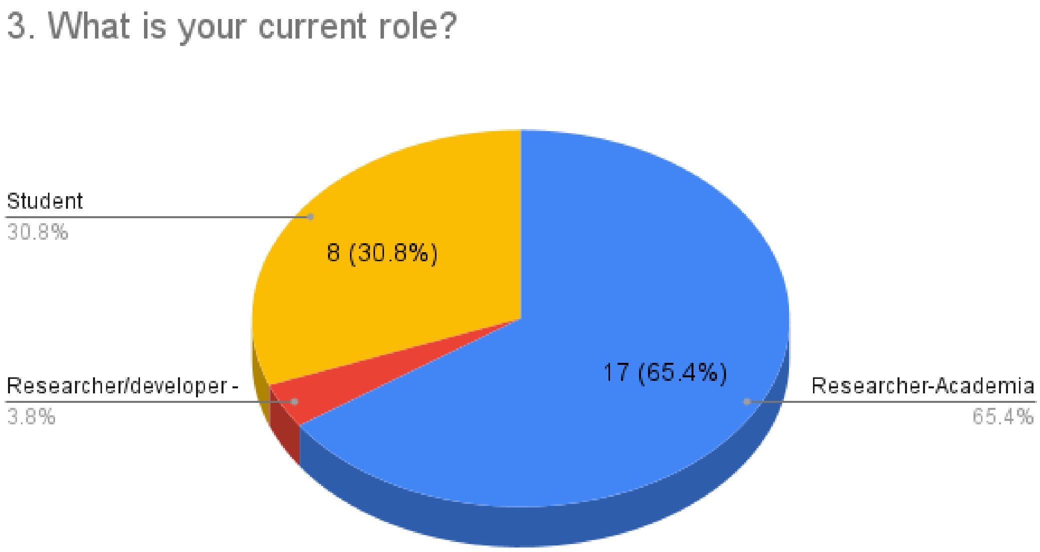


Figure 24: Respondents' role

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 usage of the service.

5.1.4. Respondents' comments/suggestions/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 number of GPUs and more users.
- opacity with respect to job scheduling

5.2. Summary

The data from the survey supports the data from the desk research. 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 responders had GPU requirements of fewer than 100 hours per month on average. Concerning, however, is the lack of

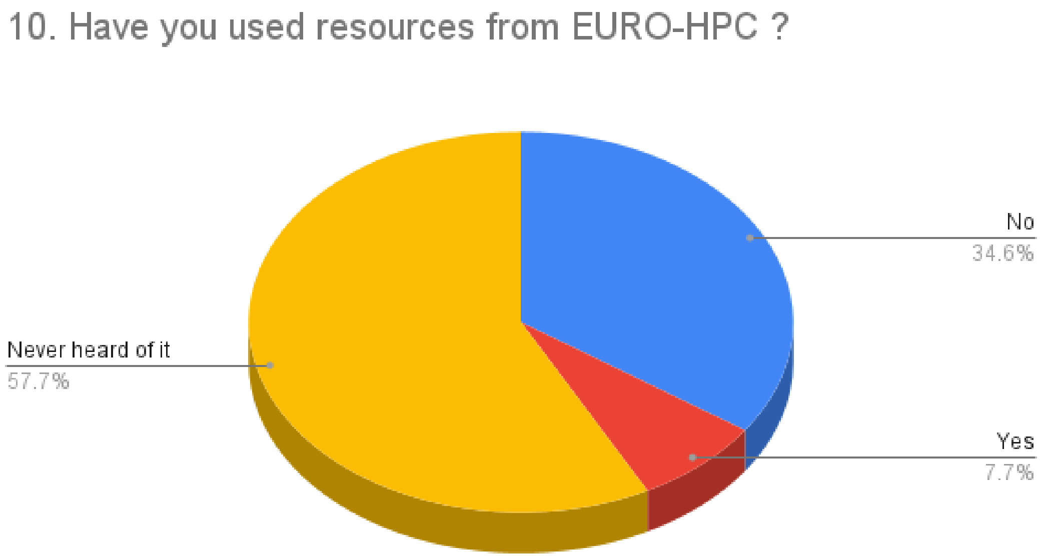


Figure 25: EuroHPC-JU usage

knowledge of the European-level HPC services available to users. While 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

High-performance computing 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 high-performance computing (HPC) nowadays, like EuroHPC-JU, PRACE, LUMI, national consortia, etc. Yet, for the subject of LT, a unified research of the many facets of HPCs was required.

In this report, 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.
- 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/lis 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 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 otherwise. 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[25] on LLMs such as BLOOM. As the model needs 352 GB in bf16 (bfloat16) weights (176*2), the most efficient set-up is 8x80 GB A100 GPUs. Also, 2x8x40 GB A100s or 2x8x48 GB A6000 can be used. The inability to employ these LLM models without access to a large number of GPUs does provide a challenge. Mosaic ML[26] 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.

[25] https://huggingface.co/blog/bloom-inference-pytorch-scripts
[26] https://www.mosaicml.com/platform

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 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 report compilation, a number of new HPC systems[27] became operational and were not included in the analysis.
- 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.

References

Nuri Mahmoud Ahmed and Muntasir Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *ArXiv*, abs/2010.15581, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

OpenAI. Gpt-4 technical report, 2023.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/arXiv.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

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

A. Appendix A

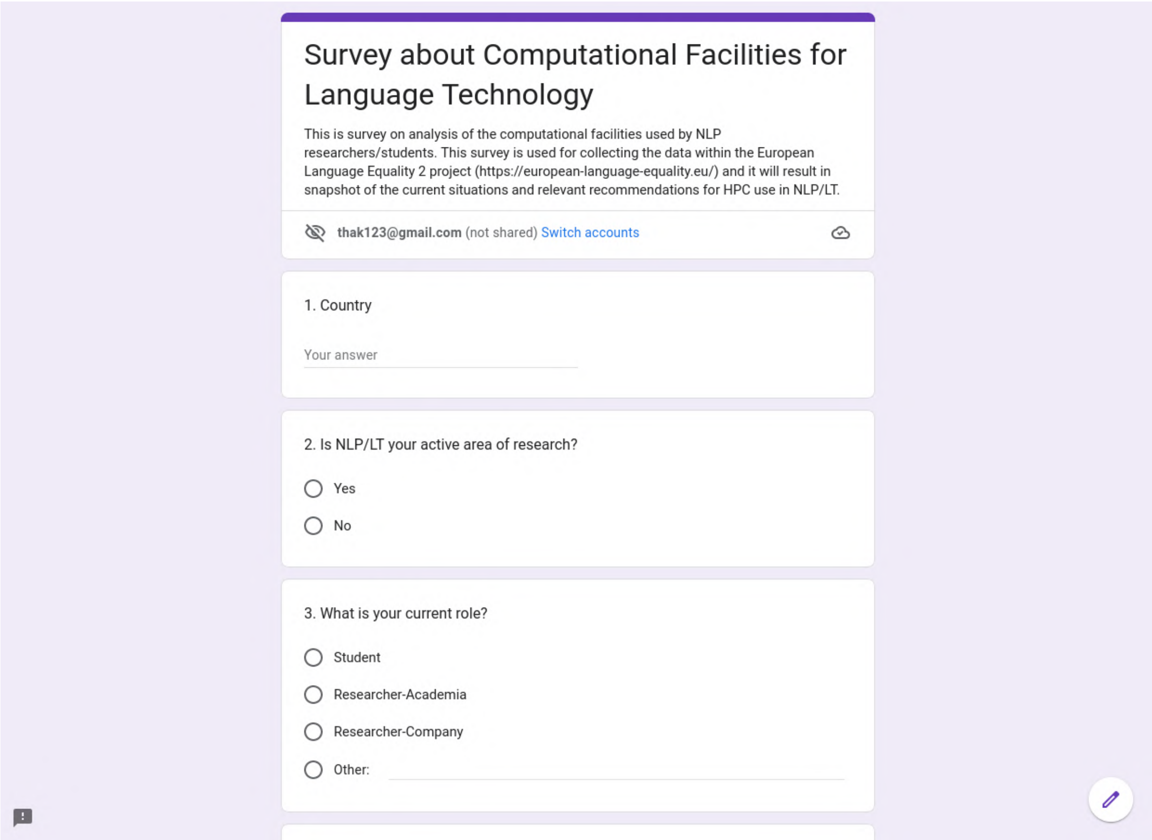
A.1. EuroHPC-JU members

The EuroHPC Joint Undertaking is composed of public and private members:

- Public members: the European Union (represented by the Commission), Member States and Associated Countries that have chosen to become members of the Joint Undertaking: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, the Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, and Türkiye.
- Private members: representatives from the three participating private partners, the European Technology Platform for High Performance Computing (ETP4HPC), the Big Data Value association (BDVA) and the European Quantum Industry Consortium (QuIC). The JU also relies on collaboration with key European actors, such as PRACE (Partnership for Advanced Computing in Europe) and GEANT (the pan-European high-speed network for research and education).

B. Appendix B

B.1. Survey about computational facilities for language technology


Survey about Computational Facilities for Language Technology

This is survey on analysis of the computational facilities used by NLP researchers/students. This survey is used for collecting the data within the European Language Equality 2 project (https://european-language-equality.eu/) and it will result in snapshot of the current situations and relevant recommendations for HPC use in NLP/LT.

thak123@gmail.com (not shared) Switch accounts

1. Country

Your answer

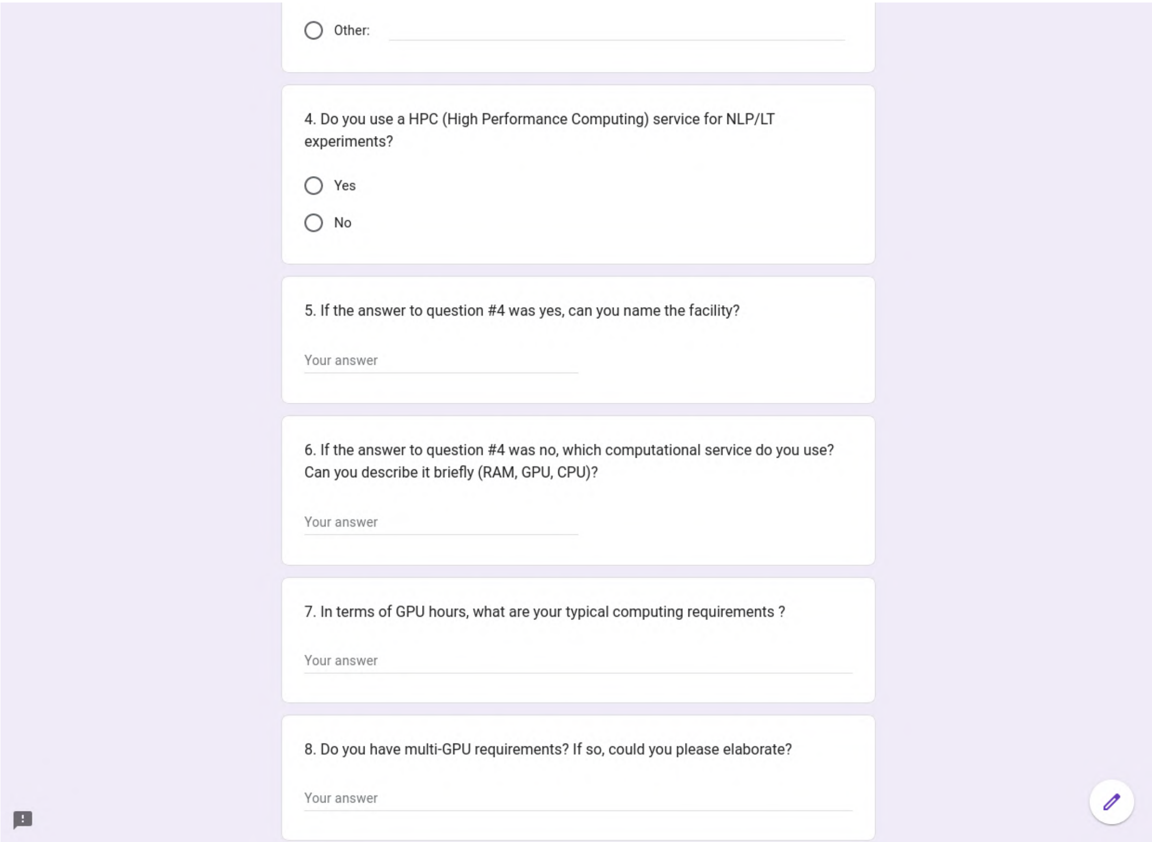
2. Is NLP/LT your active area of research?

- Yes
- No

3. What is your current role?

- Student
- Researcher-Academia
- Researcher-Company
- Other:

Figure 26: Full survey as published (page 1/3)



Other:

4. Do you use a HPC (High Performance Computing) service for NLP/LT experiments?

Yes

No

5. If the answer to question #4 was yes, can you name the facility?

Your answer

6. If the answer to question #4 was no, which computational service do you use? Can you describe it briefly (RAM, GPU, CPU)?

Your answer

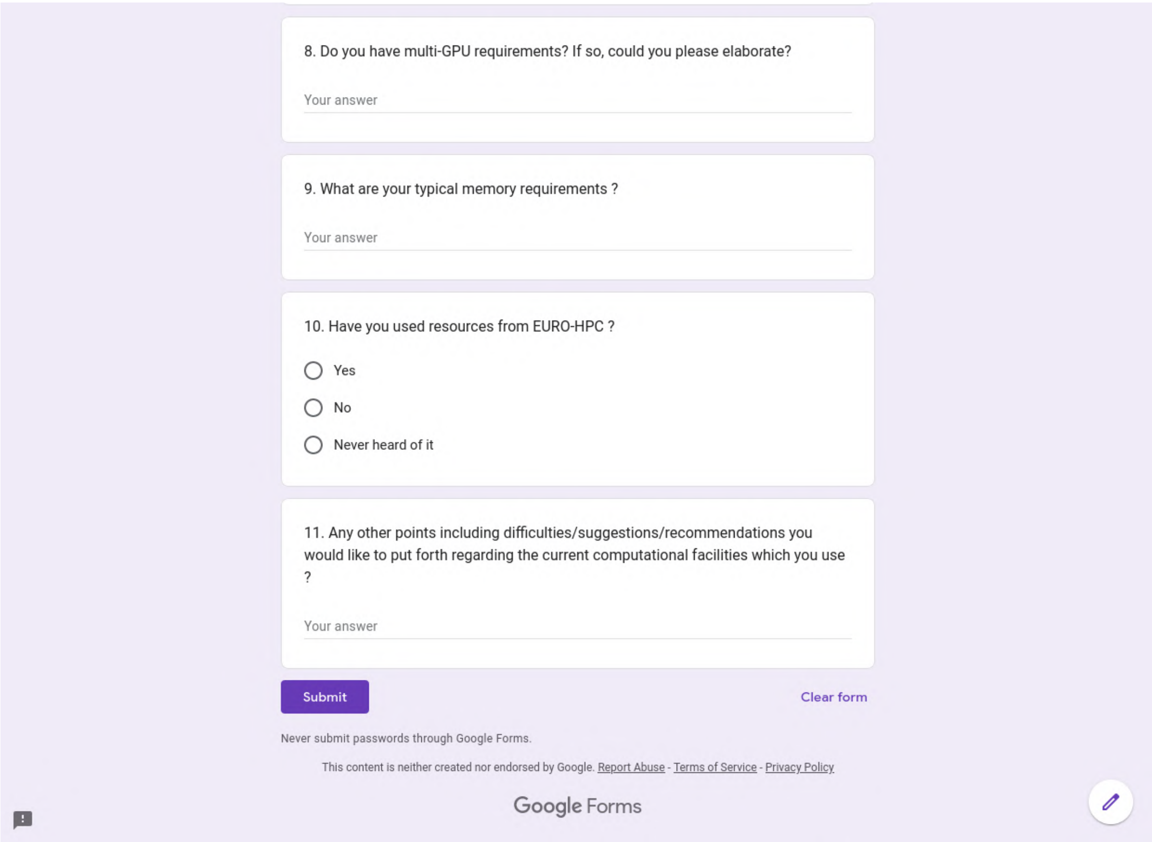
7. In terms of GPU hours, what are your typical computing requirements ?

Your answer

8. Do you have multi-GPU requirements? If so, could you please elaborate?

Your answer

Figure 27: Full survey as published (page 2/3)



8. Do you have multi-GPU requirements? If so, could you please elaborate?

Your answer

9. What are your typical memory requirements ?

Your answer

10. Have you used resources from EURO-HPC ?

- Yes
- No
- Never heard of it

11. Any other points including difficulties/suggestions/recommendations you would like to put forth regarding the current computational facilities which you use ?

Your answer

Submit

Clear form

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms

Figure 28: Full survey as published (page 3/3)

C. Appendix C

C.1. List of websites tracking HPCs

- Tracking[28] mostly Tier-2 and few Tier 0 and Tier-1 HPC centres from 44 countries.
- List of HPCs in Germany [29]

[28] https://atlas-cric-dev.cern.ch/core/rcsite/list/
[29] https://gauss-allianz.de/en/hpc-ecosystem