Journal of large-scale research facilities

Journal of large-scale research facilities, 7, A182 (2021) http://dx.doi.org/10.17815/jlsrf-7-182

Published: 29.10.2021

JURECA: Data Centric and Booster Modules implementing the Modular Supercomputing Architecture at Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre [*]

Instrument Scientists:
- Supercomputing Support, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 282, sc@fz-juelich.de

Abstract: JURECA is a Pre-Exascale Modular Supercomputer operated by Jülich Supercomputing Centre at Forschungszentrum Jülich. The system combines a flexible Data Centric (DC) module, based on the Atos BullSequana XH2000 with a selection of best-of-its-kind components, and a scalability-focused Booster module, delivered by Intel and Dell Technologies based on the Xeon Phi many-core processor. With its novel architecture, it supports a wide variety of high-performance computing and data analytics workloads.

1 Introduction

Since July 2015, the Jülich Supercomputing Centre (JSC) at the Forschungszentrum Jülich (Forschungszentrum Jülich, 2021b) operates the JURECA (Jülich Research on Exascale Cluster Architectures) system. From 2015 to 2017, the JURECA Cluster served as a general-purpose supercomputing resource and, in accordance with Forschungszentrum Jülich's dual architecture strategy, augmented the leadership-class highly scalable IBM Blue Gene/Q system JUQUEEN (Forschungszentrum Jülich, 2015). In 2017, the JURECA was itself augmented with a many-core processor based Booster module to enable highly scalable applications to leverage the system more efficiently. Funding for both JURECA modules was granted by the Helmholtz Association (Helmholtz Association, 2021) through the program "Supercomputing & Big Data". At the end of 2020 the JURECA Cluster module was replaced by the Data Centric module JURECA-DC (see Figure 1) which was funded through the PPI4HPC project (Public Procurement of Innovative Solutions for High Performance Computing, 2021).

[*] **Cite article as:** Jülich Supercomputing Centre. (2021). JURECA: Data Centric and Booster Modules implementing the Modular Supercomputing Architecture at Jülich Supercomputing Centre. *Journal of large-scale research facilities, 7*, A182. http://dx.doi.org/10.17815/jlsrf-7-182

The DC and Booster module are tightly integrated and operated as a single system following the modular supercomputing paradigm, pioneered by JSC in the context of the DEEP series of EU-funded projects (Eicker et al., 2016). The modular supercomputing concept enables users to distribute their workloads flexibly across different, architecturally diverse, modules in order to place different phases or subroutines of their workload on the hardware best suited for the execution. The software features required to leverage this architecture is made available since 2018 for the wider JURECA user community after being piloted on the DEEP prototype systems.

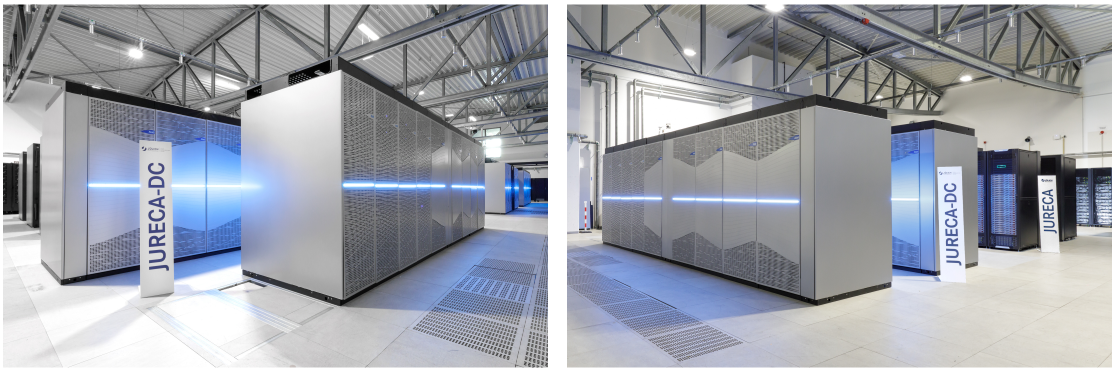

Figure 1: Jülich Research on Exascale Cluster Architectures (JURECA) at Jülich Supercomputing Centre. The left picture shows the JURECA-DC module. The right picture shows the JURECA-DC module in the front and the JURECA-Booster module in the back. Copyright: Forschungszentrum Jülich / Ralf-Uwe Limbach.

The JURECA-DC module as a result of the European PPI4HPC procurement was designed by JSC and Atos (Atos, 2021a). The JURECA-Booster module was designed by JSC and Intel (Intel Corperation, 2021) in 2016 as a highly-scalable compute architecture leveraging the latest available Intel networking and processor technology. The system was delivered by Intel together with its partner Dell Technologies (DELL Technologies, 2021) in 2017.

2 JURECA system details

The JURECA modular supercomputer consists of two separate, but tightly integrated, compute modules. The architecture of the DC module is combining the most advanced commodity hardware and software technologies available in the industry. JURECA-DC is itself a heterogeneous system offering nodes with different memory sizes (512 GiB as well as 1 TiB), nodes with graphics processing unit (GPU) accelerators as well as GPU-equipped login nodes for visualization and other post-processing workloads. The architecture of the Booster module is designed to best serve highly-scalable simulation workloads that are able to leverage the high core counts and wide vector units of the Intel Xeon Phi many-core processors.

2.1 DC module

The JURECA-DC is a BullSequana XH2000 (Atos, 2021b) supercomputer. The BullSequana XH2000 series by Atos provides a high-density node integration with warm-water direct-liquid cooling capabilities. It follows a scalable hierarchical cell-based design.

The JURECA-DC consists of 576 compute nodes of the type Atos BullSequana X2410 Blade as well as 192 GPU-accelerated BullSequana X2415 blades hosted in BullSequana XH2000 Compute Cabinets (see

Figure 2). Moreover, 12 Atos X440-A5 login and visualization nodes are available.

All systems feature two AMD EPYC Rome 7742 processors with 64 cores (CPUs) each. In the JURECA-DC module, 3200 MT/s DDR4 memory technology is used.

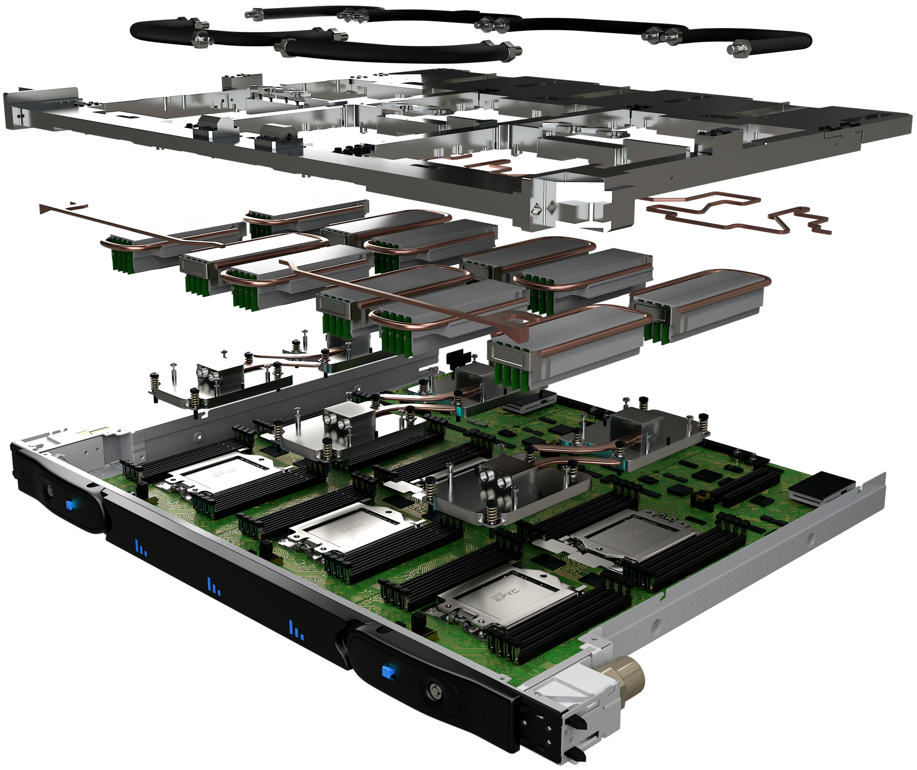

Figure 2: Atos BullSequana X2410 Blade as used in JURECA-DC module. Copyright: Atos

Applications that support the use of GPU accelerators can take advantage of the additional four NVIDIA A100 graphics processing units connected via NVLink and available in 192 compute nodes. The GPUs are connected with PCI Express Generation 4. Each A100 GPU is equipped with 40 GB HBM memory. The 12 login and visualization nodes are equipped with two NVIDIA Quadro RTX8000 GPUs intended for remote visualization usage.

The JURECA-DC compute nodes are connected with NVIDIA Mellanox HDR InfiniBand providing 100 Gb/s (12.5 GB/s) link bandwidth and MPI latencies around one microsecond. The computes nodes equipped with GPUs are connected with two times NVIDIA Mellanox HDR InfiniBand providing 200 Gb/s (25 GB/s) link bandwidth. The host channel adapters (HCAs) are connected via PCI Express Generation 4. The Cluster components are interconnected in a DragonFly+ topology which combines the advantages of DragonFly and Fat Tree topologies for appropriate communication patterns.

A particular emphasis during the design of JURECA-DC has been put on the storage connection in order to meet the increasing data requirements of simulation sciences as well as the needs of emerging data-intense sciences. All offered global (parallel) file systems on JURECA are mounted from the central Jülich Storage Cluster (JUST) (Forschungszentrum Jülich, 2021e) using IBM Spectrum Scale. Users with access to several systems in the supercomputing facility at JSC work with the same file systems on all systems so that data movement is minimized and workflows are simplified. The storage network connection of the Cluster is realized using InfiniBand-to-Ethernet gateways bridging the internal InfiniBand network with the facility's Terabit Ethernet backbone. This connection type was selected as it allows for >350 GB/s aggregate file system bandwidth as well as a high per-node file system performance.

The JURECA-DC interconnect also integrates the High Performance Storage Tier (HPST) I/O subsystem. The HPST provides a flash-based buffering layer between the disk-based storage on JUST, and the computes nodes on JURECA-DC. This layer is optimized for random access and very large bandwidth. A full description will be available in a separate article describing the storage facility at JSC. The HPST storage servers are directly attached to the InfiniBand DragonFly+ network at JURECA-DC, with HDR100 links between distributed L1-switches and storage nodes (44 nodes, each one with 2 HDR100 links).

2.2 Booster module

The JURECA-Booster consists of 1,640 compute nodes of type Dell PowerEdge C6320P (see Figure 3). All systems feature one Intel Xeon Phi 7250-F CPU (Intel Corporation, 2018) with 68 cores, a base frequency of 1.4 GHz, and 4 hardware threads per physical core. The processor package includes 16 GiB of high-bandwidth, multi-channel DRAM (MCDRAM) with a bandwidth of up to 500 GB/s. The peak performance of a Booster compute node is 3 TFlop/s. Each node is equipped with additional 96 GiB DDR4 memory clocked at 2400 MHz.

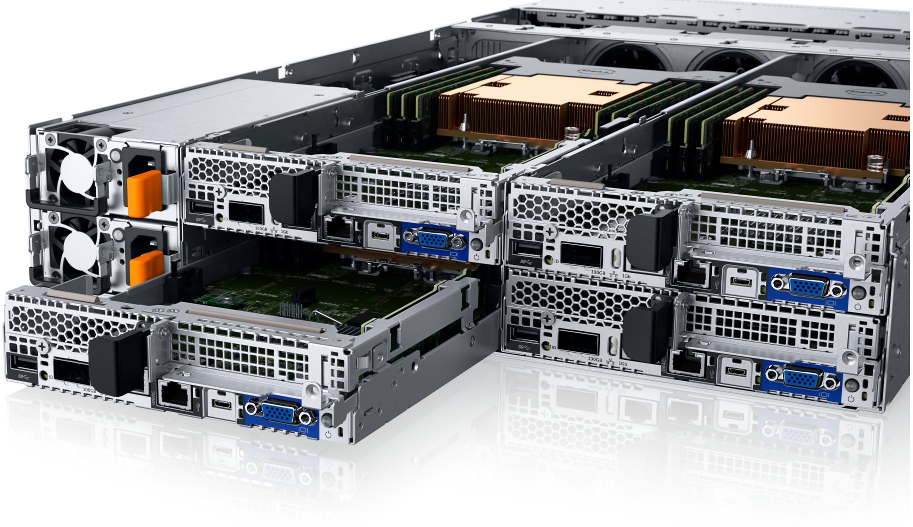

Figure 3: Example of a Dell C6320P server system. The model used in the JURECA-Booster slightly deviates from the shown version due to the utilized processor type. Copyright: Dell Technologies.

The Booster compute nodes are connected with 100 Gb/s Intel Omni-Path Architecture (OPA). The host fabric interfaces (HFI) are integrated in the CPU on the package but internally connected with PCI Express Generation 3.0 (16 lanes). The Booster nodes are interconnected in a three-level full fat tree topology. The DC and Booster modules are linked through 198 MPI router nodes equipped with one InfiniBand HCA and one OmniPath HFI, enabling Cluster-Booster communication with up to 19.8 Tb/s (2.5 TB/s) bandwidth.

The Booster connects to the JUST cluster to access the same file systems as are available on the DC module. The storage connection is realized with 26 router nodes equipped with two HFI ports and two 40 Gigabit Ethernet connections to the facility Ethernet fabric. The nominal network speed of the storage connection is 260 GB/s.

2.3 Software

The JURECA software stack, as well as the software stack in all the systems operated by JSC, is largely based on open-source software. On the login and compute nodes a CentOS 8 Linux operating system is used. Since the DC module compute nodes are disk-less, only a stripped-down operating system is available on them.

All JSC-operated software is deployed and configured via Ansible (Red Hat Inc, 2021), a software provisioning and configuration management tool. The use of Ansible enables to reuse configuration across nodes and systems easily, and facilitates homogeneity, traceability and reproducibility.

JURECA uses the open-source Slurm workload manager (SchedMD LLC, 2021) in combination with the ParaStation resource management which has a proven track record in scalability, reliability and performance on several clusters operated by JSC. The ParTec ParaStation ClusterSuite (ParTec AG, 2021) is used for node imaging and health monitoring.

The management of the scientific software stack in JURECA relies on EasyBuild (Hoste et al., 2012). Both modules, DC and Booster, have GCC, Intel and NVHPC compilers available. Support for AMD Optimizing C/C++ Compiler (AOCC) on the DC module is in development at the moment. The Message Passing Interface (MPI) implementations supported are mainly ParaStationMPI and OpenMPI, both being CUDA-aware for efficient internode GPU communication. On the Booster side, IntelMPI is also offered. Different compilers, optimized mathematical libraries and pre-compiled community codes are available. We refer to the JURECA webpage (Forschungszentrum Jülich, 2021d) for more information. Monitoring of batch jobs is possible using the latest version of the LLview (Forschungszentrum Jülich, 2021f) graphical monitoring tool.

Scientists can also use UNICORE (UNICORE, 2021) to create, submit and monitor jobs on JURECA. In addition, the system can be accessed via the Jupyter@JSC service (Forschungszentrum Jülich, 2021c).

The software functionality required for high-speed communication between Cluster and Booster via MPI is implemented in ParaStation. At the time of the Booster deployment in 2017, the software was available at a proof-of-concept level. It is matured in the course of the year 2018 and was made available, along with the necessary enhancement of the workload manager also for the JURECA-DC and JURECA-Booster integration.

2.4 Hardware components

As of this writing, JURECA consists of the following hardware components. An up-to-date description of the hardware (and software) configuration of the system is maintained on the JURECA webpage (Forschungszentrum Jülich, 2021d).

2.4.1 DC module

- 480 standard compute nodes
  - 2× AMD EPYC 7742, 2× 64 cores, 2.25 GHz
  - 512 (16× 32) GB DDR4, 3200 MHz
  - InfiniBand HDR100 (NVIDIA Mellanox Connect-X6)
  - diskless
- 96 large-memory compute nodes
  - 2× AMD EPYC 7742, 2× 64 cores, 2.25 GHz
  - 1024 (16× 64) GB DDR4, 3200 MHz
  - InfiniBand HDR100 (NVIDIA Mellanox Connect-X6)
  - diskless
- 192 accelerated compute nodes
  - 2× AMD EPYC 7742, 2× 64 cores, 2.25 GHz
  - 512 (16× 32) GB DDR4, 3200 MHz
  - 4× NVIDIA A100 GPU, 4× 40 GB HBM2e
  - 2× InfiniBand HDR (NVIDIA Mellanox Connect-X6)
  - diskless
- NVIDIA Mellanox InfiniBand HDR (HDR100/HDR) DragonFly+ network
  - NVIDIA Mellanox ConnectX-5 single and dual port host channel adapters in nodes
  - 102× 40-port Mellanox HDR switches
  - 7× Fat tree non-blocking interconnection network inside each compute cell (each cell consists of two racks)
    - 8× HDR leaf switches (L1)
    - 6× HDR spine switches (L2)
  - Service/Storage Island (attached via Up/Down chain routing)
    - 8× Mellanox Skyway Storage (GPFS) Gateways
    - 4× 40-port Mellanox HDR switches

2.4.2 Booster module

- 33 racks organized in three rows
  - 1,640 compute nodes
    - Intel Xeon Phi "Knights Landing" 7250-F CPU
      - 68 cores, 1.4 GHz base frequency
      - AVX-512 instruction set architecture extension
    - 16 GiB multi-channel DRAM (MCDRAM)
    - 96 GiB DDR4 memory clocked at 2400 MHz (6 channels)
  - 26 storage router nodes
    - Dual-port Intel Omni-Path host fabric interface cards
    - 2×40 Gigabit Ethernet connection to facility fabric
- Intel Omni-Path Architecture network organized in a three-level full-fat tree topology
  - On-package Omni-Path host fabric interface
  - 48-port Intel Omni-Path Edge Switch 100 switches
  - 3× Intel Omni-Path Director Class Switch 100 core switches

2.4.3 Joint infrastructure

- 12 login nodes
  - 2× AMD EPYC 7742, 2× 64 cores, 2.25 GHz
  - 1024 (16× 64) GB DDR4, 3200 MHz
  - 2× NVIDIA Quadro RTX8000
  - InfiniBand HDR100 (NVIDIA Mellanox Connect-X6)
  - 100 Gigabit Ethernet external connection
- 24 service nodes for system management
- 198 Cluster-Booster bridge nodes
  - 1× Mellanox ConnectX-5 single port host channel adapter connected to edge switches
  - 1× Intel Omni-Path host fabric interface card connected to edge switches

2.5 Software components

- CentOS 8 enterprise-grade Linux operating system
- ParTec ParaStation Modulo ClusterSuite
- Slurm batch system with ParaStation resource management
- Intel and ParTec ParaStation Message Passing Interface implementations
- Support for OpenMP, NVIDIA CUDA, OpenCL and OpenACC programming models

2.6 Benchmark results

In 2015, on the old JURECA Cluster a Linpack performance of 1.42 PFlop/s was measured, using 1,764 compute nodes without accelerators, placing the system on spot 50 in the November 2015 Top500 list (Top500, 2015).

In 2017, following the installation of the Booster module, a combined Linpack performance of 3.78 PFlop/s was measured with 1,760 Cluster and 1,600 Booster compute nodes. The upgrade placed the system on spot 29 in the November 2017 Top500 list (Top500, 2017). With an average 2.81 GFlop/s/W, the system ranked on spot 55 in the Green500 list in November 2017 (Green500, 2017).

In 2021, following the installation of JURECA-DC, a Linpack performance of 9.33 PFlop/s was measured with 186 DC GPU equipped compute nodes. This result placed the JURECA-DC module on spot 43 in the June 2021 Top500 list (Top500, 2021). With an average 24.291 GFlop/s/W, the system ranked on spot 8 in the Green500 list in June 2021 (Green500, 2021). On the High Performance Conjugate Gradients (HPCG) benchmark, JURECA-DC achieved 273.784 TFlop/s in 2021 corresponding to place 29 in the June 2021 HPCG list (HPCG, 2021).

3 Access to JURECA

Scientists and engineers interested in using the capacity and capability of the JURECA-DC module and the JURECA-Booster module for their research are invited to apply for computing time resources by submitting an adequate proposal in answer to the corresponding computing time calls published twice a year in January/February and July/August. These calls are conducted jointly by peers in computational science and engineering at Forschungszentrum Jülich and RWTH Aachen University, accepting proposals from these two institutions only (so-called JARA and VSR Call) (Jülich-Aachen Research Alliance, 2018).

The JURECA-DC module was procured as part of the PPI4HPC project, in which computing centres from four European countries have united to purchase new, innovative supercomputer systems through a joint procedure – for the first time at the European level. The EU is supporting the process of Public

Procurement of Innovative Solutions (PPI) by assuming 35 % of the costs that are incurred. Therefore, JURECA-DC is also available for European scientists.

All applications undergo a comparative peer-review process. The scientific quality and significance of the applications is being evaluated by national and international scientists who are experts in their respective scientific fields. Additionally, the technical feasibility of the applications is ensured by a technical assessment to enable an efficient use of the JURECA supercomputer. Details on how to get access to JURECA and how to apply for computing time, including the given requirements, are given on the webpage of the JSC (Forschungszentrum Jülich, 2021a).

References

Atos. (2021a). *Atos.* Retrieved from https://www.atos.net

Atos. (2021b). *Atos Bullsequana XH2000 product webpage.* Retrieved from https://atos.net/en/solutions/high-performance-computing-hpc/bullsequana-x-supercomputers#bullsequana-xh2000

DELL Technologies. (2021). *DELL Technologies.* Retrieved from https://www.delltechnologies.com

Eicker, N., Lippert, T., Moschny, T., & Suarez, E. (2016). The DEEP Project An alternative approach to heterogeneous cluster-computing in the many-core era. *Concurrency and computation*, *28*(8), 2394–2411. http://dx.doi.org/10.1002/cpe.3562

Forschungszentrum Jülich. (2015). JUQUEEN: IBM Blue Gene/Q Supercomputer System at the Jülich Supercomputing Centre. *Journal of large-scale research facilities*, *1*, A1. http://dx.doi.org/10.17815/jlsrf-1-18

Forschungszentrum Jülich. (2021a). *Computingtime@JSC webpage.* Retrieved from https://www.fz-juelich.de/ias/jsc/computingtime

Forschungszentrum Jülich. (2021b). *Forschungszentrum Jülich.* Retrieved from https://www.fz-juelich.de

Forschungszentrum Jülich. (2021c). *Jupyter@JSC webpage.* Retrieved from https://jupyter-jsc.fz-juelich.de

Forschungszentrum Jülich. (2021d). *JURECA webpage.* Retrieved from https://www.fz-juelich.de/ias/jsc/jureca

Forschungszentrum Jülich. (2021e). *JUST webpage.* Retrieved from https://www.fz-juelich.de/ias/jsc/just

Forschungszentrum Jülich. (2021f). *LLview webpage.* Retrieved from https://www.fz-juelich.de/jsc/llview

Green500. (2017). *Green500 November 2017 list.* Retrieved from https://top500.org/lists/green500/2017/11/

Green500. (2021). *Green500 June 2021 list.* Retrieved from https://top500.org/lists/green500/2021/06/

Helmholtz Association. (2021). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF).* Retrieved from https://www.helmholtz.de

Hoste, K., Timmerman, J., Georges, A., & De Weirdt, S. (2012). EasyBuild: Building Software with Ease. In *2012 SC Companion: High Performance Computing, Networking Storage and Analysis* (p. 572-582). http://dx.doi.org/10.1109/SC.Companion.2012.81

HPCG. (2021). *HPCG June 2021 list.* Retrieved from https://top500.org/lists/hpcg/2021/06/

Intel Corperation. (2021). *Intel Corperation.* Retrieved from https://www.intel.com

Intel Corporation. (2018). *Intel Xeon Phi Processor 7250-F.* Retrieved from https://ark.intel.com/products/94035/Intel-Xeon-Phi-Processor-7250-16GB-1_40-GHz-68-core

ParTec AG. (2021). *ParTec webpage.* Retrieved from https://www.par-tec.com

Public Procurement of Innovative Solutions for High Performance Computing. (2021). *PPI4HPC project information at Horizon 2020.* Retrieved from https://www.ppi4hpc.eu

Red Hat Inc. (2021). *Ansible Configuration Manager webpage.* Retrieved from https://www.ansible.com

SchedMD LLC. (2021). *Slurm Workload Manager webpage.* Retrieved from https://slurm.schedmd.com

Top500. (2015). *Top500 November 2015 list.* Retrieved from https://www.top500.org/lists/2015/11

Top500. (2017). *Top500 November 2017 list.* Retrieved from https://www.top500.org/lists/2017/11

Top500. (2021). *Top500 June 2021 list.* Retrieved from https://www.top500.org/lists/2021/06

UNICORE. (2021). *Uniform Interface to Computing Resources (UNICORE) webpage.* Retrieved from https://www.unicore.eu