EuroCC workshop

Big Data analysis with Hadoop and RHadoop

03-03-2022

This course is an EuroCC event jointly organised by EuroCC Slovenia and EuroCC Austria.

EuroCC project

- National competence cetres HPC in 33 countries
- Tasks:
  - 1: (National Competence Centre) - Management
  - 2: Training and Skills Development
  - 3: Technology Transfer/Business Development
  - 4: Collaboration with Industry
  - 5: Mapping of HPC/Big Data/AI Technical Competences
  - 6: Facilitation of access to scientific and technical expertise and knowledge pools
  - 7: Awareness Creation and Collaboration
- to fill the gaps in HPC knowledge in different areas:
  - Use in industry,
  - School education,
  - Higher education institutions and
  - Enthusiastic individuals.

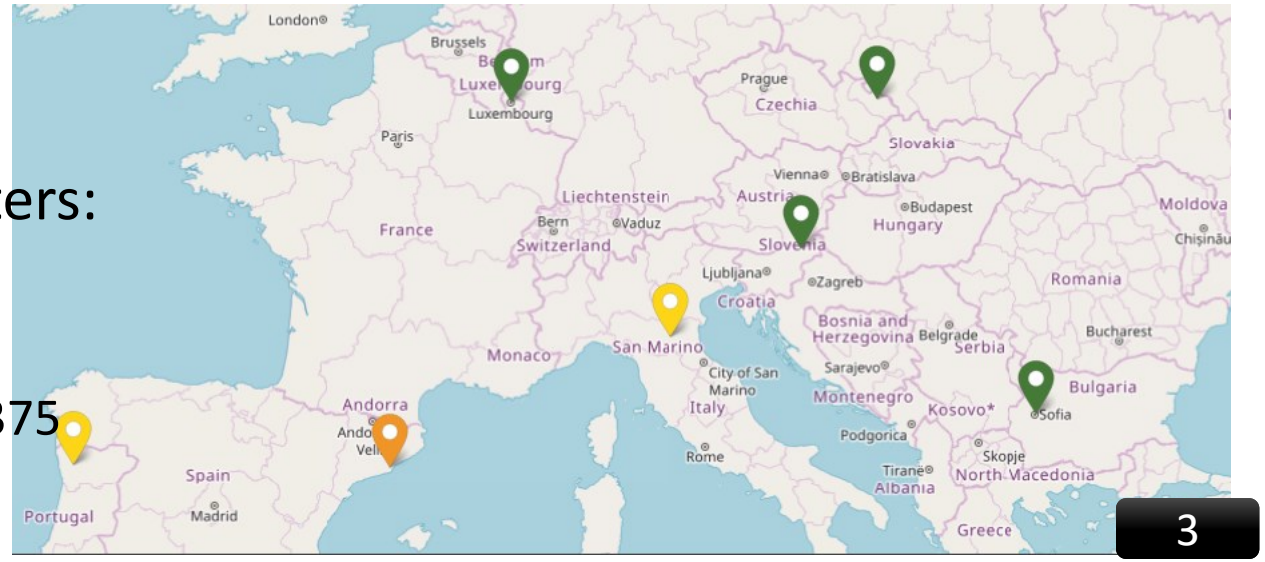
EuroHPC Joint Undertaking

HPC landscape

#EuroHPC (high performance computing) **Joint Undertaking**

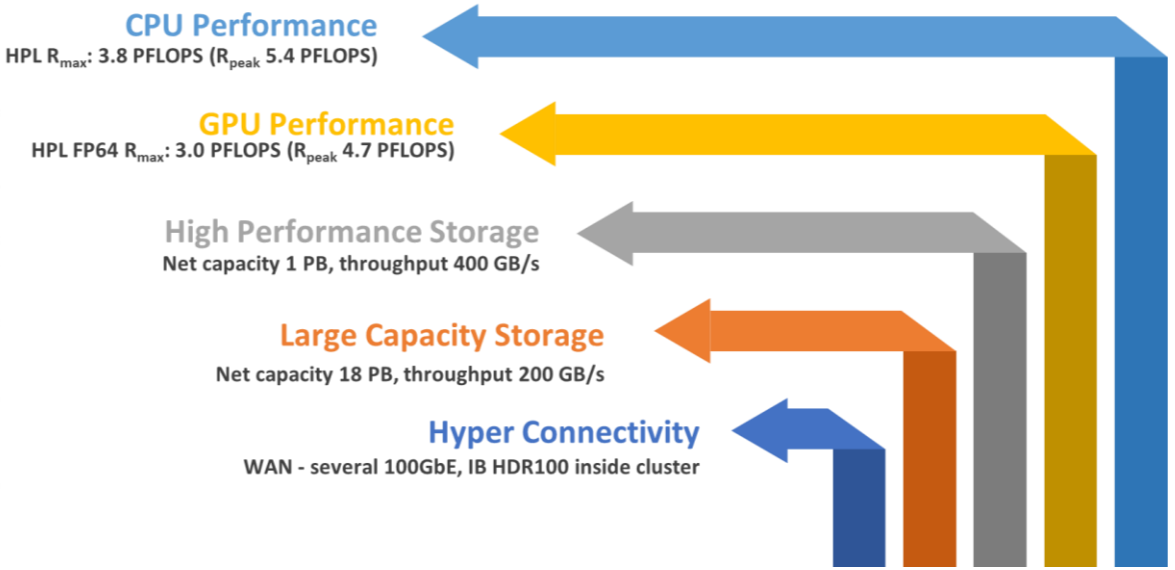
- Five petascale supercomputers:
  - Deucalion, Portugal (under construction)
  - Karolina, Ostrava, Czech Republic (2021) – 15,7 peteflops
  - MeluXina, Luxembourg (2021) – 10 petaflops
  - Discoverer in Sofia Tech Park, Bulgaria (2021) – 4,5 petaflops
  - Vega, Maribor, Slovenia (2021) - 10,1 petaflops

- Three precursors-to-exascale supercomputers:
  - MareNostrum5, Barcelona, Spain
  - Leonardo, hosted in Bologna, Italy (2022)
  - LUMI, Kajaani, Finland (under construction) – 375 petaflops



HPC landscape

EuroHPC system VEGA



Operational since early 2021

computing capacity 10 PetaFLOPS .
Dual-processor compute nodes with more than 100,000 processor cores.
Dual-processor compute nodes with additional graphics processing units with more than 600,000 cores.
Disk array with at least 4 PB of fast and at least 30 PB of permanent storage space.

https://www.hpc-rivr.si/system/

EuroHPC system Discoverer

Operational since late 2021

Dicoverer is a Petascale supercomputer capable of executing more than 4,2 Petaflops Rmax and over 6 petaFLOPS Rpeak

Architecture and Parameters:

- The size of the system is combining 12 computing Direct Liquid Cooling BullSequana racks
- The platform is built on AMD EPYC processors with hot water cooling and its representing 376 computing nodes
- The number of cores is 144,384 where size of the RAM reaches 300TB
- 2 PB of fast disk storage DDN space is guaranteeing optimal operability for store procedures
- The entire system is backed up against a power failure by an uninterruptible power supply with an output of 1 MW
- The whole complex weighs over 30 tons.

https://sofiatech.bg/en/petascale-supercomputer/

HPC landscape

EuroHPC system MeluXina

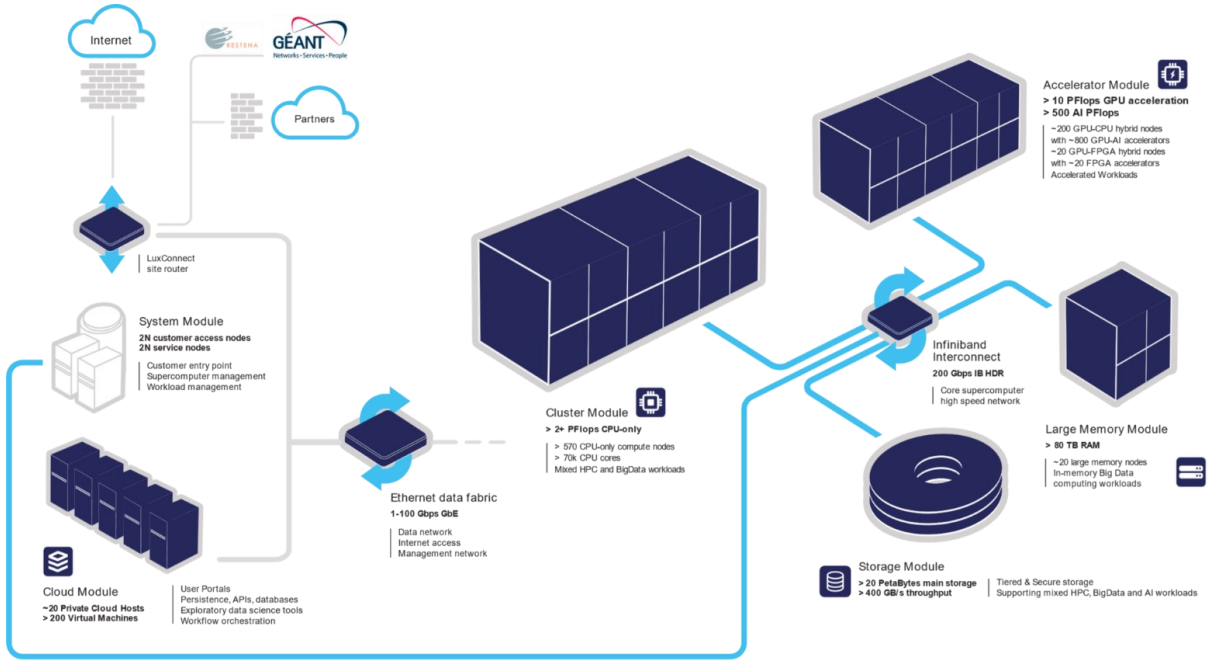
Operational since summer 2021

Cluster – CPU
573 nodes, each equipped with 2 AMD Rome CPUs (2x 64 cores @ 2.6 GHz) and 512 GB of RAM, for a total of 73344 cores and 293 TB RAM

Accelerator - GPU
200 GPU-accelerated nodes, each with 2 AMD Rome CPUs (2x 32 cores @ 2.35 GHz) and 4 NVIDIA A100 GPUs with 40 GB HBM. Each node also includes 512 GB of RAM, a local SSD of 1.92 TB and 2 links to the InfiniBand HDR fabric.

Accelerator - FPGA
20 FPGA nodes, each composed of 2 AMD Rome CPUs (2x 32 cores @ 2.35 GHz) and 2 Intel Stratix 10MX 16 GB HBM FPGA cards. Each node has 512 GB of RAM, a local SSD of 1.92 TB and dual links to the InfiniBand fabric.

Large Memory
Each large memory node is composed of 2 AMD Rome CPUs (2x 64 core @ 2.6 GHz), has 4 TB of memory (4096 GB), 1.92 TB of local storage, and is dual-linked into the InfiniBand fabric



https://luxprovide.lu/

EuroHPC system Karolina

THEORETICAL PEAK PERFORMANCE	15,690 Tflop/s
OPERATING SYSTEM	Centos 64 bit 7.x
COMPUTE NODES	831
CPU	720x 2x AMD 7H12, 64 cores, 2,6 GHz, 92,160 cores in total 72x 2x AMD 7763, 64 cores, 2,45 GHz, 9,216 cores in total 72x 8x NVIDIA A100 GPU, 576 GPU in total 32x Intel Xeon-SC 8628, 24 cores, 2,9 GHz, 768 cores in total 36x 2x AMD 7H12, 64 cores, 2,6 GHz, 4,608 cores in total 2x 2x AMD 7452, 32 cores, 2,35 GHz, 128 cores in total
RAM PER COMPUTE NODE	256 GB / 1 TB (GPU) / 24 TB fat node 320 GB HBM2 (8 x 40 GB) GPU
ACCELERATORS	576x NVIDIA A100
STORAGE	30.6 TB / home (1.93 GB/s sequential write performance, 3.10 GB/s sequential read performance), 1,361 TB / scratch (NVMe, 730.9 GB/s sequential write performance, 1, 198.3 GB/s sequential read performance)
INTERCONNECT	Infiniband HDR 200 Gb/s

Installed in 2021

https://www.it4i.cz/en/infrastructure/karolina

EuroHPC system Leonardo

Leonardo will be built from Atos' BullSequana XH2000 supercomputer nodes, each with four NVIDIA Tensor Core GPUs and a single Intel CPU. It will also use NVIDIA Mellanox HDR 200Gb/s InfiniBand connectivity, with smart in-network computing acceleration engines that enable extremely low latency and high data throughput to provide the highest AI and HPC application performance and scalability.

Leonardo will feature nearly 14 000 NVIDIA Ampere architecture-based GPUs. It will deliver 10 exaflops of FP16 AI performance.

- More than 136 BullSequana XH2000 Direct Liquid cooling racks
- 250 PFLOPs HPL Linpack Performance (Rmax)
- 10 ExaFLOPS of FP16 AI performance
- 3456 servers equipped with Intel Xeon Ice Lake and NVIDIA Ampere architecture GPUs
- 1536 servers with Intel Xeon Sapphire processors
- 5PB of High Performance storage
- 100PB of Large Capacity Storage

Installed in 2022 (comming soon)

https://www.cineca.it/en/hot-topics/Leonardo-announce

Workshop description

The training event will consist of two 4-hour trainings in two consecutive days.

The first day will focus to big data management and data analysis with Hadoop. The participant will learn how to

(i) move big data efficiently to a cluster and to Hadoop distributed file system, and

(ii) how to perform simple big data analysis by Python scripts using MapReduce and Hadoop.

The second day will focus to big data management and analysis using Rhadoop.

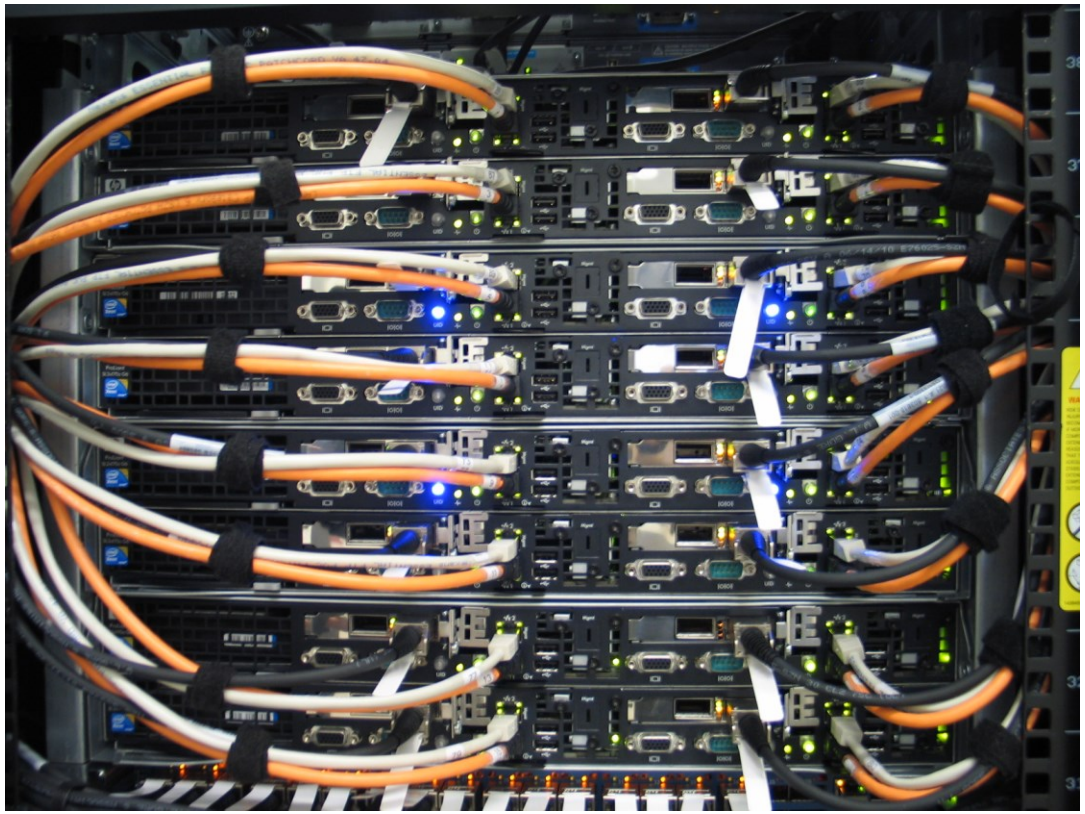
We will stick to work within RStudio and will write all scripts within R using several state-of-the-art libraries for parallel computations, like parallel, doParallel and foreach and libraries to work with Hadoop, like rmr, rhdfs and rhbase.

About the workshop

- Skills to be gained:
  - connect to a supercomputer using NoMachine tool;
  - Move big data to a supercomputer and store it to a distributed file system;
  - Writing Python scripts to perform basic data management and data analysis tasks by Hadoop;
  - Writing R scripts to perform basic data management and data analysis tasks by Rhadoop libraries like rmr, rhdfs and rhbase;

HPC-FS

- Hardware:
  - nadzorni strežniki (master and login nodes)
  - delovna vozlišča (worker nodes)
  - datotečni strežnik (ZFS, NFS, LUSTRE
  - omrežni del (Infiniband, Ethernet)
- Software:
  - Ansys Multiphysics
  - Ansys CFX, Fluent, Maxwell, HFSS
  - OpenFOAM CFD + extend
  - VisIt in ParaView postprocesor
  - Intel F90, CC
  - TotalView, Allinea DDT
  - Modelirnik Siemens NX in ProEngineer
  - Octave, R, Mathematica
  - OpenMP, OpenMPI, HPMPI, IntelMPI
  - ATLAS, BLAS, BLACS, FFTW, GOTO, MUMPS, NetCDF, HDF5, Sparsekit, Scalapack, ...



Trainers

- Prof. Janez Povh; University of Ljubljana, Slovenia
- Dr. Giovanna Roda ; EuroCC Austria, BOKU, and TU Wien, Austria
- Liana Akobian ; TU Wien, Austria

Day 1, 03 March 2022

Beginning	End	Description
13:00	13:15	Introduction
13:15	14:00	Introduction to HADOOP • Introduction to Big Data • The Hadoop Distributed Computing Architecture
14:00	14:15	Break
14:15	15:00	• HDFS: the Hadoop Distributed File System • MapReduce
15:00	15:15	Break
14:15	15:00	Hands-on demos
16:00	16:15	Break
16:15	16:00	Hands-on demos

Day 2, 04 March 2022

Beginning	End	Description
13:00	13:15	Introduction to Day 2
13:15	14:00	Introduction to R • Connecting to RStudio web server at HPC@UL • Creating and running own R scripts; Creating, retrieving, saving data files; • Standard data management operations on data frames; Data management with dplyr, magritt
14:00	14:15	Break
14:15	15:00	Advanced and Big data management with R • Dana manipulations with apply functions apply, lapply, sapply, vapply, tapply, and mapply • Big Data management with function for efficient parallel loops parLapply, parSapply, mcLapply and for each-dopar
15:00	15:15	Break
14:15	15:00	Big data management with RHadoop • Preparing and storing big data to HDFS using rhdfs library • Retriving from and managing big data in HDFS by plyrmr and rhdfs library
16:00	16:15	Break
16:15	16:00	Big data analysis with RHadoop Preparing map-reduce scripts to make basic data analysis tasks (extreme values, counts, mean values, dispersions, visualisations) using rhdfs library
16:00	16:05	Wrap-up

EURO AUSTRIA

EURO SLING

Thank you for attention!

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 951732. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, United Kingdom, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Switzerland, Turkey, Republic of North Macedonia, Iceland, Montenegro

EuroHPC Joint Undertaking