MEIM | MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

IN COLLABORATION WITH MIT MANAGEMENT SLOAN SCHOOL

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

MASTER MEIM 2022-2023

# DIGITAL TECH

# High Performance Computing

## Lesson 2

Prof. Giulio Giunta

Prof. of Scientific Computing, Università degli Studi di Napoli Parthenope

# Do we really need very powerful computers ?

**Supercomputing** has the potential to be the underlying layer to support solutions for many of the world's most pressing contemporary challenges:
global privacy and identity issues, stalemates in medical research, sustainable supply-chain logistics, genomics, weather forecast, energy-smart power grid, carbon capture, astrophysics, data analysis and market economics, blockchain, high-frequency trading, and cryptography,…

https://www.energy.gov/science/doe-explainsexascale-computing

# Do we really need very powerful computers ?

with the proliferation of data, we need to **extract insights from big data** and escalating customer demands for near-real-time service

the need for systems that can handle **AI technologies** such as deep learning is making high-performance computing a must-have for the enterprise market

since the number of connected devices (IoT) is growing at a rapid pace, we expect that by 2025 there will be more than 10 billion devices and sensors interconnected via the internet

all these devices will generate around 4.4 trillion GB of data, i.e. 4.4 Zettabyte

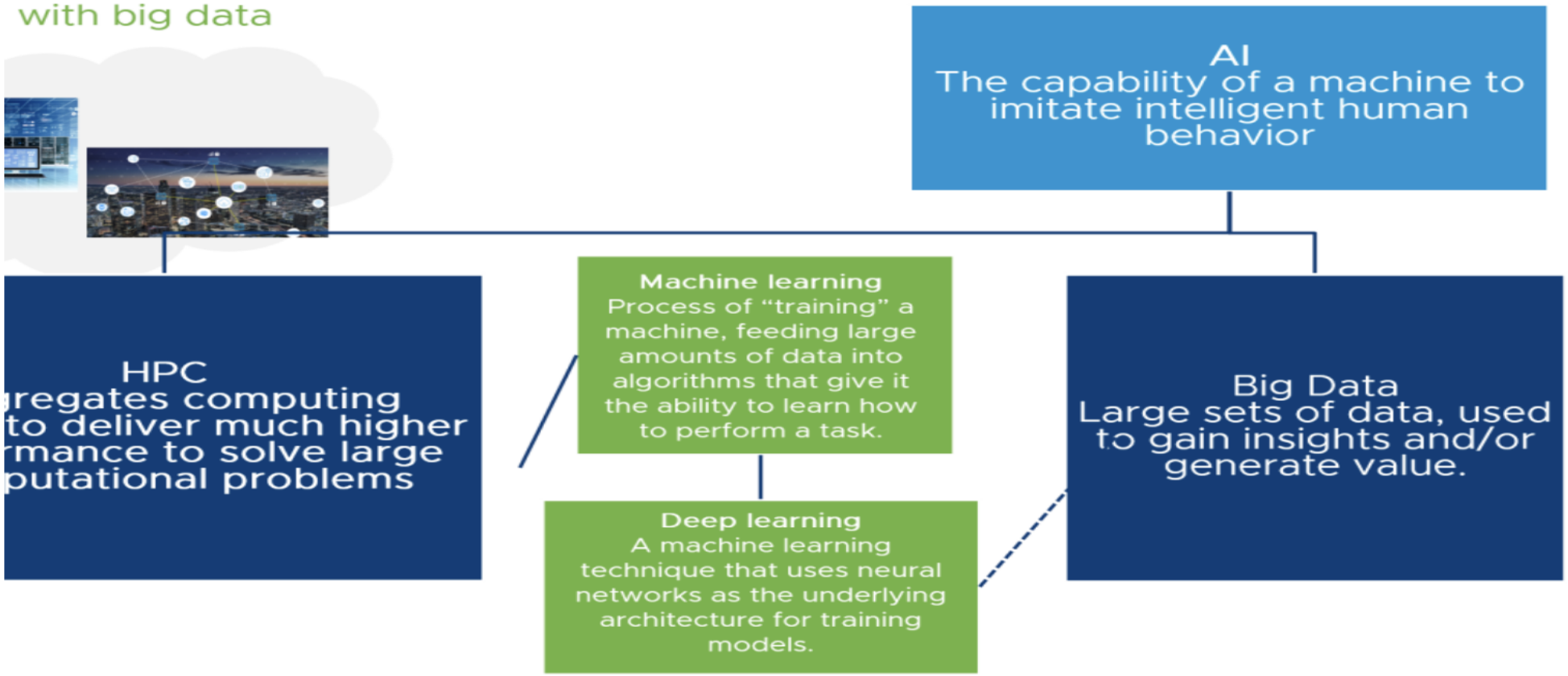
# Do we really need very powerful computers ?

MARKETSANDMARKETS™

High Performance Computing (HPC) Market by Component (Solutions (Servers, Storage, Networking Devices, and Software) and Services), Deployment Type, Organization Size, Server Prices Band, Application Area, and Region - Global Forecast to 2025

https://www.marketsandmarkets.com/Market-Reports/Quantum-High-Performance-Computing-Market-631.html

# Do we really need very powerful computers ?




https://www.scientificamerican.com/article/new-exascale-supercomputer-can-do-a-quintillion-calculations-a-second/

# Do we really need very powerful computers ?

To have an idea of the range of applications of HPC, let's give a look to the scientific journal [The International Journal of High Performance Computing Applications]

Articles in the current (May 2022) issue:

- Data-driven global weather predictions at high resolutions
- Development of a hardware-accelerated simulation kernel for ultra-high vacuum with Nvidia RTX GPUs
- A GPU-accelerated adaptive FSAI preconditioner for massively parallel simulations
- A massively parallel time-domain coupled electrodynamics–micromagnetics solver
- Large-scale ab initio simulation of light–matter interaction at the atomic scale in Fugaku
- Large-scale ab initio simulation of light–matter interaction at the atomic scale in Fugaku
- SAM++: Porting the E3SM-MMF cloud resolving model using a C++ portability library
- Parallel encryption of input and output data for HPC applications
- Resiliency in numerical algorithm design for extreme scale simulations

# Do we really need very powerful computers ?

Artificial intelligence (Machine Learning, Natural Language Processing, Robotics) has a wide range of potential applications in entrepreneurship and innovation management

1. Predictive Analytics: AI can be used for predictive analytics, where historical data is analyzed to make predictions about future outcomes. Identify trends, forecast demand, optimize pricing, and make data-driven decisions

2. Chatbots and Virtual Assistants: AI-powered chatbots and virtual assistants can automate customer service, provide personalized recommendations, answer common inquiries, improve customer engagement, reduce response times

3. Autonomous Systems: perform tasks without human intervention. For example, autonomous robots can be used in manufacturing processes, autonomous vehicles can revolutionize transportation, and autonomous drones can be used for various applications like delivery or surveillance

4. Natural Language Processing: natural language processing capabilities can be used to analyze and understand unstructured data, such as customer feedback, social media posts, and market trends

# Do we really need very powerful computers ?

Artificial intelligence (Machine Learning, Natural Language Processing, Robotics) has a wide range of potential applications in entrepreneurship and innovation management

5. Recommender Systems: recommender systems that provide personalized recommendations to customers based on their preferences, behavior, and historical data. This can enhance the customer experience, drive sales, and increase customer loyalty

6. Fraud Detection: detect patterns and anomalies in financial transactions, helping entrepreneurs and innovators prevent fraud and mitigate risks. This is particularly relevant for e-commerce platforms and financial services

7. Innovation and Idea Generation: AI techniques such as natural language processing and machine learning can assist in idea generation and innovation processes. They can analyze vast amounts of data, identify patterns, and suggest new ideas or improvements based on market insights and customer preferences

# Do we really need very powerful computers ?

popular Apps on smartphones centered on the GPT or framework to provide users with interactive and engaging conversational experiences

**1. ChatSonic**

ChatSonic is a chat platform designed to simplify communication and boosts productivity. It is integrated with Google Search to create content with the latest information. It is used for text and image creation, an AI-powered platform from Writesonic. ChatSonic is trained on large datasets from customer feedback which is refined for the most accurate and up-to-date results

**2. Alissu: Chat with AI**

Alissu is an Android application that effortlessly responds to every query. Additional features are the app's help in teaching languages, the basics of programming, and various new skills. Alissu utilizes OpenAI's API

**3. Rapid ChatGPT**

It can generate content as well as answer queries. It is a cutting-edge conversational AI tool that leverages the power of OpenAI's advanced language model. It enables users to engage in fast, dynamic, and intelligent conversations, providing a seamless and efficient experience for a wide range of applications and interactions

# Do we really need very powerful computers ?

popular Apps on smartphones centered on the GPT or framework to provide users with interactive and engaging conversational experiences

**4.ChatOn- AI Chat with GPT-3**

This chatbot is powered by advanced AI and is a conversational agent to provide intelligent responses and assist with various tasks. It can comprehend user preferences and make appropriate adjustments

**5. Frank: AI Chat Assistant**

It is an AI-powered search and chat AI bot and also a content creation engine that provides users with an ads-free search, brainstorming, text, and image creation experience. It is built on groundbreaking ChatGPT and voice technology, it uses NLP and ML to provide users with accurate and up-to-date information. It also can generate images and summaries of text for users to save time

**6. Roboco**

A chatbot assistant that provides a more user-friendly experience with AI algorithms. It is a ChatGPT alternative that quickly responds to consumer inquiries. A unique quality of Roboco is its capacity to learn from human interactions

# Do we really need very powerful computers ?

popular Apps on smartphones centered on the GPT or framework to provide users with interactive and engaging conversational experiences

**7. ChatGod**

It is an advanced online Chatbot that uses ChatGPT AI technology to answer questions instantly. It is also a content generator capable of creating engaging text based on input. It  provides accurate and high-quality language processing services like speech-to-text, text-to-speech, image, and video generation

**8. Open Chat: AI Chatbot App**

It acts as a virtual assistant to answer questions, provide information, and also carry out tasks based on commands. This Chatbot understands natural language, meaning, users can ask questions or issue commands in a conversational style rather than on specific commands or syntax

**9. Wisdom AI – AI Assistant for iOS**

It is a research management software for researchers, academics, and students. It uses GPT-3 and Dall-E technologies. The app can generate an image, write a blog post, can track down a bug in code, find answers to questions, etc.

**10. Jasper Chat**

It is an AI writer and AI art generator. This app has learned from billions of articles and other content in 29 languages

# How powerful is a powerful computer ?

Cray 1, 1977

133 Mops, no graphics, 60Kw, 56kb/s

AMD 8 core, GPU 10.3 Teraflops, 5 watt, network connection, cost: 399 Euro

PS5 Sony, 2022

# How powerful is a powerful computer ?

supercomputers ?

2023

supercomputer

1977

Univac 1106,
0.15 MegaFLOPS
2Mbyte

# How powerful is a powerful computer ?

- Computers have been increasing steadily in performance since the 1940s
- The **Colossus** vacuum tube computer was the first electronic computer in the world
- Built in Britain during the Second World War, **Colossus** ran at 500,000 FLOPS
- **CDC 6600** in 1964 was the first supercomputer with 3 MegaFLOPS
- **Cray-2** in 1985 was the first supercomputer to reach over 1 GigaFLOPS
- **ASCI Red** in 1996 was the first massively parallel computer reaching over a TeraFLOPS
- **Roadrunner** in 2008 first supercomputer to reach 1 PetaFLOPS

# How powerful is a powerful computer ?

today's treshold

PetaFLOPS scale ( $10^{15}$ ) → ExaFLOPS scale ( $10^{18}$ )

pre-Exascale era

# How powerful is a powerful computer ?

in June 2022, he supercomputer **FRONTIER** has broken the ExaFLOPS treshold ( $10^{18}$ ) in solving a standard basic computational problem

we are entering the Exascale era

# How powerful is a powerful computer ?

in June 2022, he supercomputer **FRONTIER** has broken the ExaFLOPS treshold ( $10^{18}$ ) in solving a standard basic computational problem

just how massive is that?
you'd have to perform one calculation every second for 31,688,765,000 years to match what a 1 ExaFLOPS computer system can do in just one second

# Key development lines

**Hybrid computing architecture**: Developments in this direction include the use of CPUs and GPUs in tandem, as well as the integration of specialized accelerators like FPGAs and ASICs (i.e. Google's Tensor Processing Unit -TPU), for specific AI applications. These hybrid systems can significantly improve computing speed and efficiency

**High-speed interconnections**: High-speed networking technologies, like InfiniBand and 100 Gbps Ethernet, are becoming increasingly important for connecting computing nodes in an HPC system. These high-speed interconnections are crucial for supporting AI and LLM applications that require the rapid sharing of large amounts of data

**High-speed data storage**: High-speed storage technologies, like NVMe SSDs, are crucial for managing the large volumes of data used in AI and LLM applications

# Key development lines

**Optimization software and algorithms**: Resource management software and optimization algorithms play a crucial role in maximizing the performance of HPC systems. These tools help to balance the workload and manage computing resources, making the entire system more efficient

**HPC as a Service (HPCaaS)**: With the growing interest in cloud computing, HPC as a service is becoming increasingly popular. This model allows companies to access high-performance computing resources on demand, without having to manage and maintain their own HPC infrastructure

**Green HPC**: As HPC continues to grow, so does its environmental impact. As a result, there is a growing interest in "green HPC" technologies and practices, which aim to reduce the energy consumption and carbon footprint of supercomputers

# legenda

**GPU (Graphics Processing Unit)**: A GPU is a type of processing chip specialized for processing graphics, but also used for other applications that require a large number of parallel calculations, such as data processing for artificial intelligence and machine learning

**FPGA (Field Programmable Gate Array)**: An FPGA is a programmable integrated device that can be configured to perform a wide range of functions. Unlike CPUs and GPUs, which have fixed architectures, an FPGA can be "reprogrammed" to perform specific functions or algorithms, making it extremely flexible

**ASIC (Application Specific Integrated Circuit)**: An ASIC is a custom integrated circuit (IC) for a specific application, rather than for general purposes. Because an ASIC is designed for a specific task, it can be much more energy and performance efficient than a generalist CPU or GPU

**SSD NVMe (Non-Volatile Memory Express Solid State Drive)**: An NVMe SSD is a type of storage device that uses solid state technology (SSD) and the NVMe interface. SSD technology has no moving parts and can read and write data much faster than traditional hard drives. NVMe is an interface that was specifically designed to make the most of SSD speed, offering significantly faster data access times than other interfaces like SATA

# What is the computer more powerful in the world today ?

http://www.top500.org

TOP500® SUPERCOMPUTER SITES

the 500 fastest computers at solving a predetermined set of standard problems (benchmark)

**LINPACK** Benchmark

High-Performance Conjugate Gradient (**HPCG**) Benchmark

# What is the computer more powerful in the world today ?

November 2022

**first EXAFLOPS computer**


**FRONTIER, HPE CRAY**
Power: **1,102** ExaFLOPS

**DOE/SC/Oak Ridge National Laboratory**
**USA**

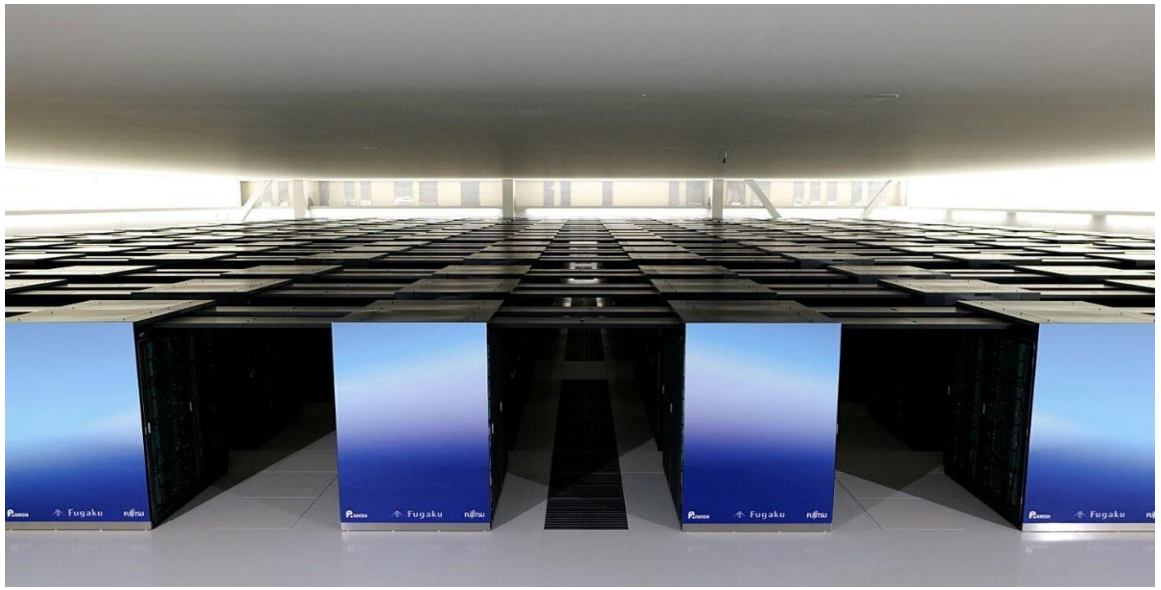
https://www.olcf.ornl.gov/frontier/

8.730.112 core  HPC and AI Optimized 3rd Gen AMD EPYC CPU 4 Purpose Built AMD Instinct 250X GPUs, 10 PB Ram

# What is the computer more powerful in the world today ?

November 2022

second in the ranking



**Fugaku, FUJITSU**
power: **442** PetaFLOPS
**Riken Center for Computational Science, Kobe, Japan**

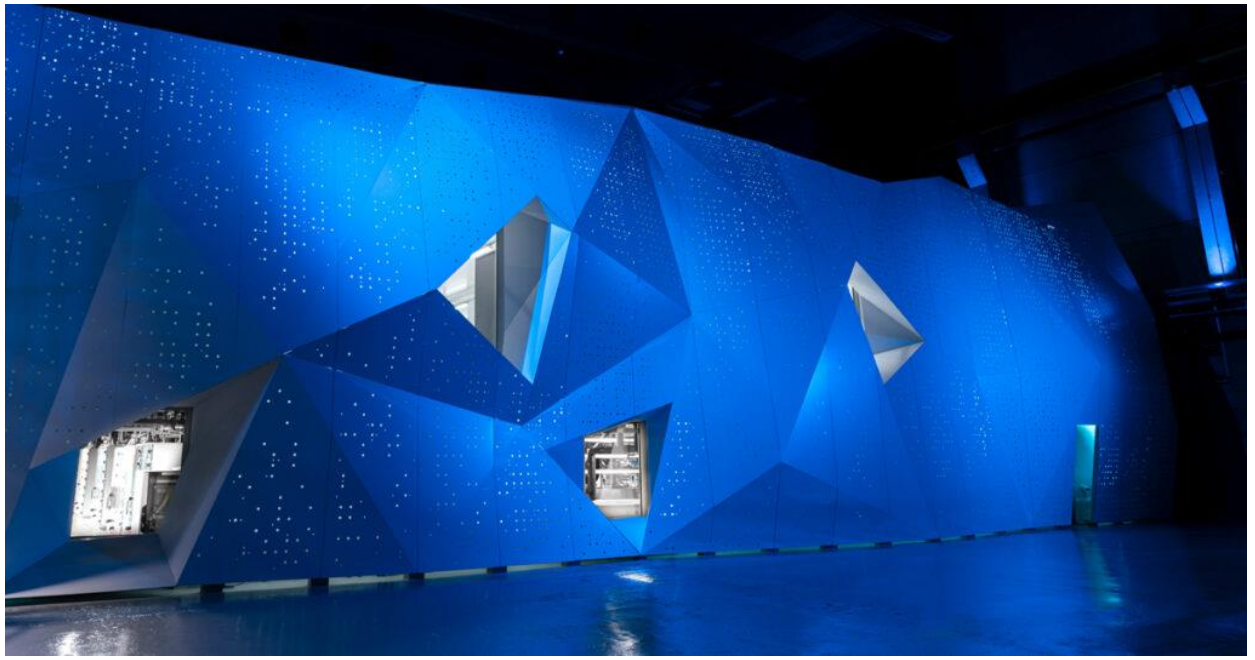
https://www.r-ccs.riken.jp/en/ fugaku/project

7.630.848 core ARM, 160.000 nodes A64FX 48C, 160 PB Ram

# What is the computer more powerful in the world today ?

November 2022

third in the ranking



**LUMI, HPE Cray**
power: **151** PetaFLOPS
**EuroHPC/CSC**
**Kajaani, Finland**

https://www.lumi-supercomputer.eu/

1.110.144 core AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, 8 PB Ram

# What is the computer more powerful in the world today ?

November 2022

fourth in the ranking


**Leonardo, Atos**
power: **174** PetaFLOPS
**CINECA**,
**Italy**

https://www.hpc.cineca.it/hardware/leonardo

1.463.616 core Xeon Platinum, Nvidia A100 SXM4, 10 PB Ram

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

TOP 500

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# What is the computer more powerful in the world today ?

November 2022

fifth in the ranking


**Summit, IBM**
power: **148** PetaFLOPS
**Oak Ridge National Laboratory,**
**USA**

https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

2.414.592 core Power9, 4600 nodes Nvidia Volta, 10 PB Ram

# What is the computer more powerful in the world today ?

November 2022

sixth in the ranking

**SIERRA** , IBM
power: **94** PetaFLOPS
Lawrence Livermore National Laboratory, USA

https://hpc.llnl.gov/hardware/platforms/sierra

1.572.000 core, 710 TB Ram

# What is the computer more powerful in the world today ?

November 2022

seventh in the ranking


**Sunway TaihuLight, NRCPC**

power: **93** PetaFLOPS

**National Supercomputing Center**, Wuxi, Cina

http://www.nsccwx.cn/

10.649.600 core, 1.3 PB Ram

# What is the computer more powerful in the world today ?

November 2022

13° place

**HPC5 , IBM**
power: **35** PetaFLOPS
ENI SpA, Pavia, Italia

https://www.eni.com/it_IT/attivita/green-data-center-hpc5

669.760 core, nodes Power Edge Xeon + Nvidia Tesla V100, Infiniband

# What is the computer more powerful in the world today ?

November 2022

24° place

**Marconi-100 , IBM**
power: **21** PetaFLOPS
CINECA, Bologna, Italia

https://www.hpc.cineca.it/hardware/marconi100

347.766 core, nodes IBM Power9

# TOP500

November 2022

| Rank | System | Cores | Rmax (PFlop/s) | Rpeak (PFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 8,730,112 | 1,102.00 | 1,685.65 | 21,100 |
| 2 | **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu<br>RIKEN Center for Computational Science<br>Japan | 7,630,848 | 442.01 | 537.21 | 29,899 |
| 3 | **LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>EuroHPC/CSC<br>Finland | 2,220,288 | 309.10 | 428.70 | 6,016 |
| 4 | **Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos<br>EuroHPC/CINECA<br>Italy | 1,463,616 | 174.70 | 255.75 | 5,610 |
| 5 | **Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 2,414,592 | 148.60 | 200.79 | 10,096 |

MEIM | MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

TOP500

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# TOP500

November 2022

| | | | | | |
|---|---|---|---|---|---|
| 6 | **Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox<br>DOE/NNSA/LLNL<br>United States | 1,572,480 | 94.64 | 125.71 | 7,438 |
| 7 | **Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC<br>National Supercomputing Center in Wuxi<br>China | 10,649,600 | 93.01 | 125.44 | 15,371 |
| 8 | **Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE<br>DOE/SC/LBNL/NERSC<br>United States | 761,856 | 70.87 | 93.75 | 2,589 |
| 9 | **Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia<br>NVIDIA Corporation<br>United States | 555,520 | 63.46 | 79.22 | 2,646 |
| 10 | **Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT<br>National Super Computer Center in Guangzhou<br>China | 4,981,760 | 61.44 | 100.68 | 18,482 |
| 11 | **Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES)<br>France | 319,072 | 46.10 | 61.61 | 921 |

MEIM | MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

TOP500

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# TOP500

November 2022

HPCG
Benchmark

| Rank | TOP500 Rank | System | Cores | Rmax (PFlop/s) | HPCG (TFlop/s) |
|---|---|---|---|---|---|
| 1 | 2 | **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu<br>RIKEN Center for Computational Science<br>Japan | 7,630,848 | 442.01 | 16004.50 |
| 2 | 1 | **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 8,730,112 | 1,102.00 | 14054.00 |
| 3 | 3 | **LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>EuroHPC/CSC<br>Finland | 2,220,288 | 309.10 | 3408.47 |
| 4 | 5 | **Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 2,414,592 | 148.60 | 2925.75 |
| 5 | 4 | **Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos<br>EuroHPC/CINECA<br>Italy | 1,463,616 | 174.70 | 2566.75 |

# TOP500

November 2022

GREEN 500

| Rank | TOP500 Rank | System | Cores | Rmax (PFlop/s) | Power (kW) | Efficiency (GFlops/wat |
|---|---|---|---|---|---|---|
| 1 | 405 | **Henri** - Lenovo ThinkSystem SR670 V2, Intel Xeon Platinum 8362 2800Mhz (32C), NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo<br>Flatiron Institute<br>United States | 5,920 | 2.04 | 31 | 65.091 |
| 2 | 32 | **Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 120,832 | 19.20 | 309 | 62.684 |
| 3 | 11 | **Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES)<br>France | 319,072 | 46.10 | 921 | 58.021 |
| 4 | 15 | **Setonix – GPU** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>Pawsey Supercomputing Centre, Kensington, Western Australia<br>Australia | 181,248 | 27.16 | 477 | 56.983 |

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

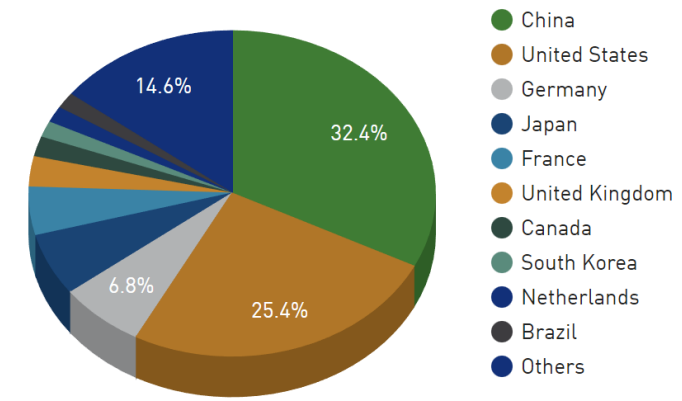
TOP 500

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# TOP500

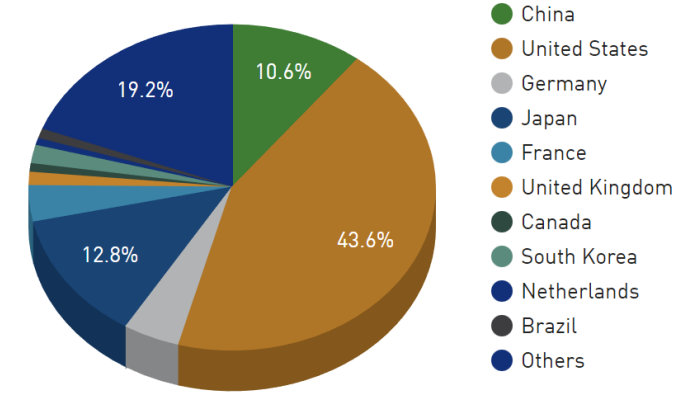
November 2022

**Countries System Share**



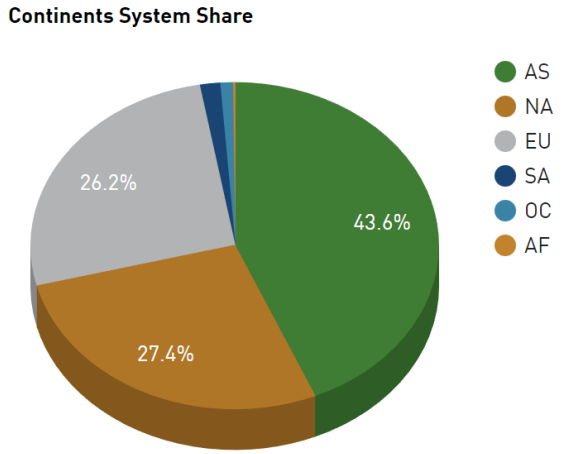

distribution by countries of supercomputing power

**Countries Performance Share**

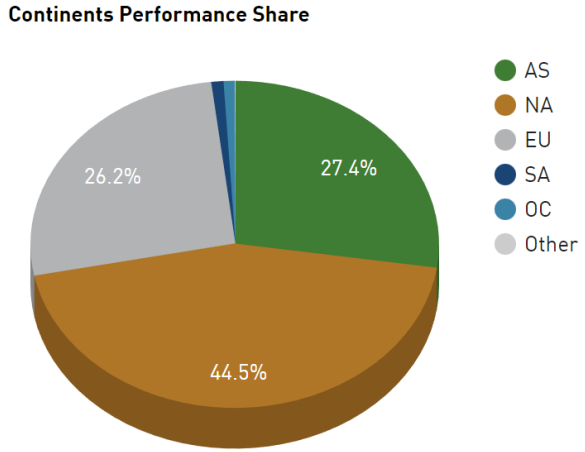


# TOP500

November 2022



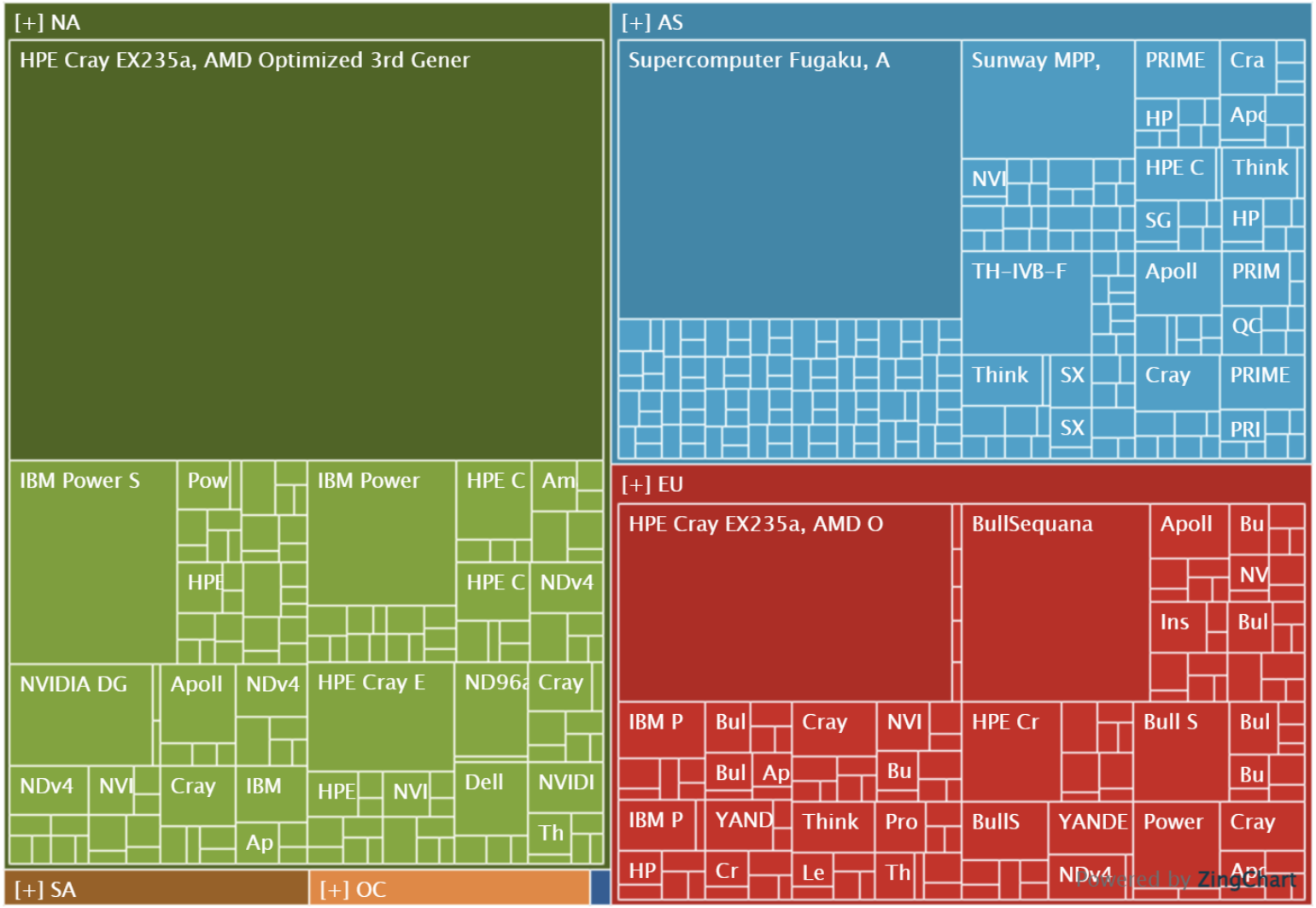

distribution by continents of supercomputing power



# TOP500

November 2022

Treemap of Continents




https://www.top500.org/statistics/treemaps/

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

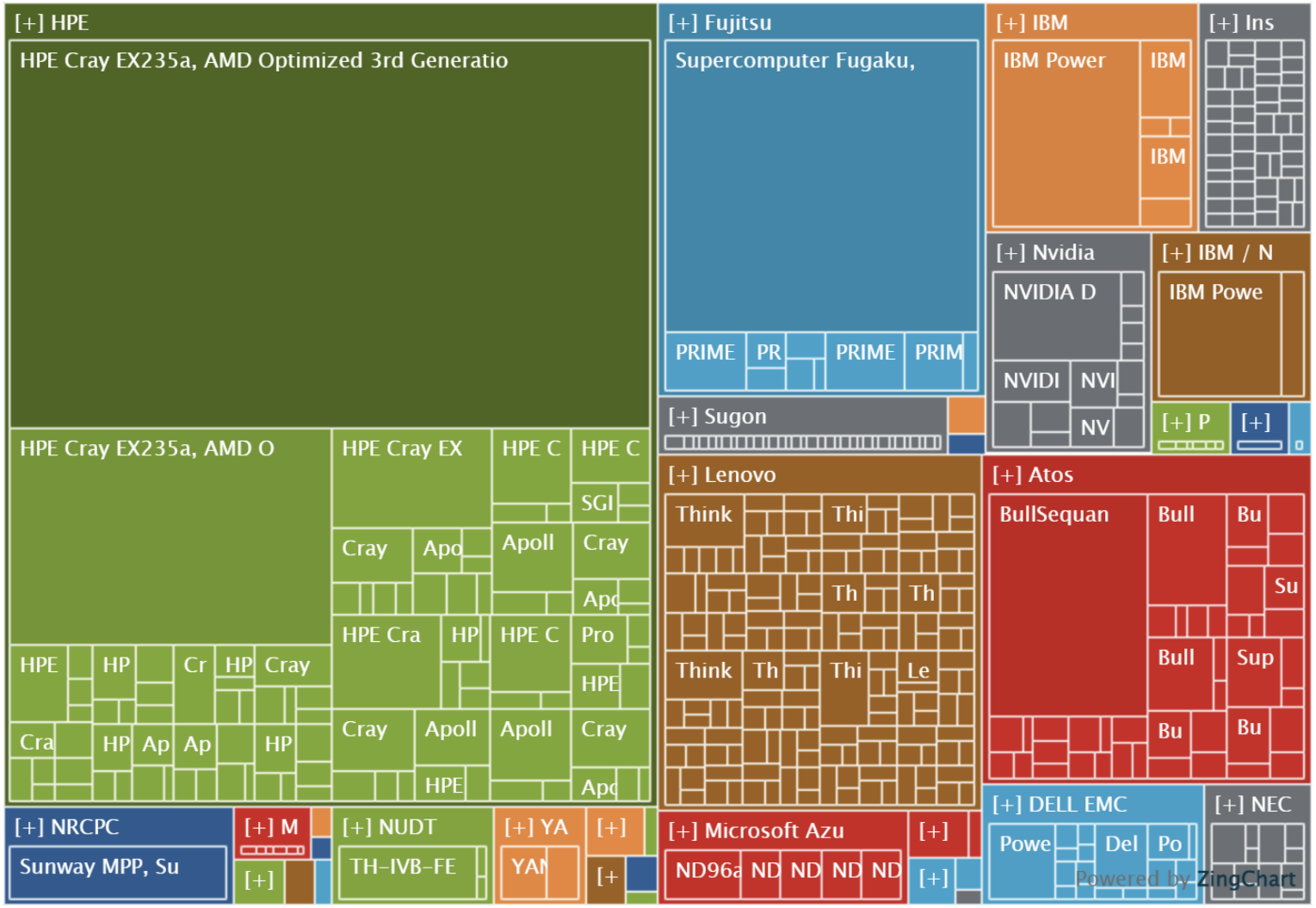
TOP500

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# TOP500

November 2022

Treemap of Vendors




**https://www.top500.org/statistics/treemaps/**

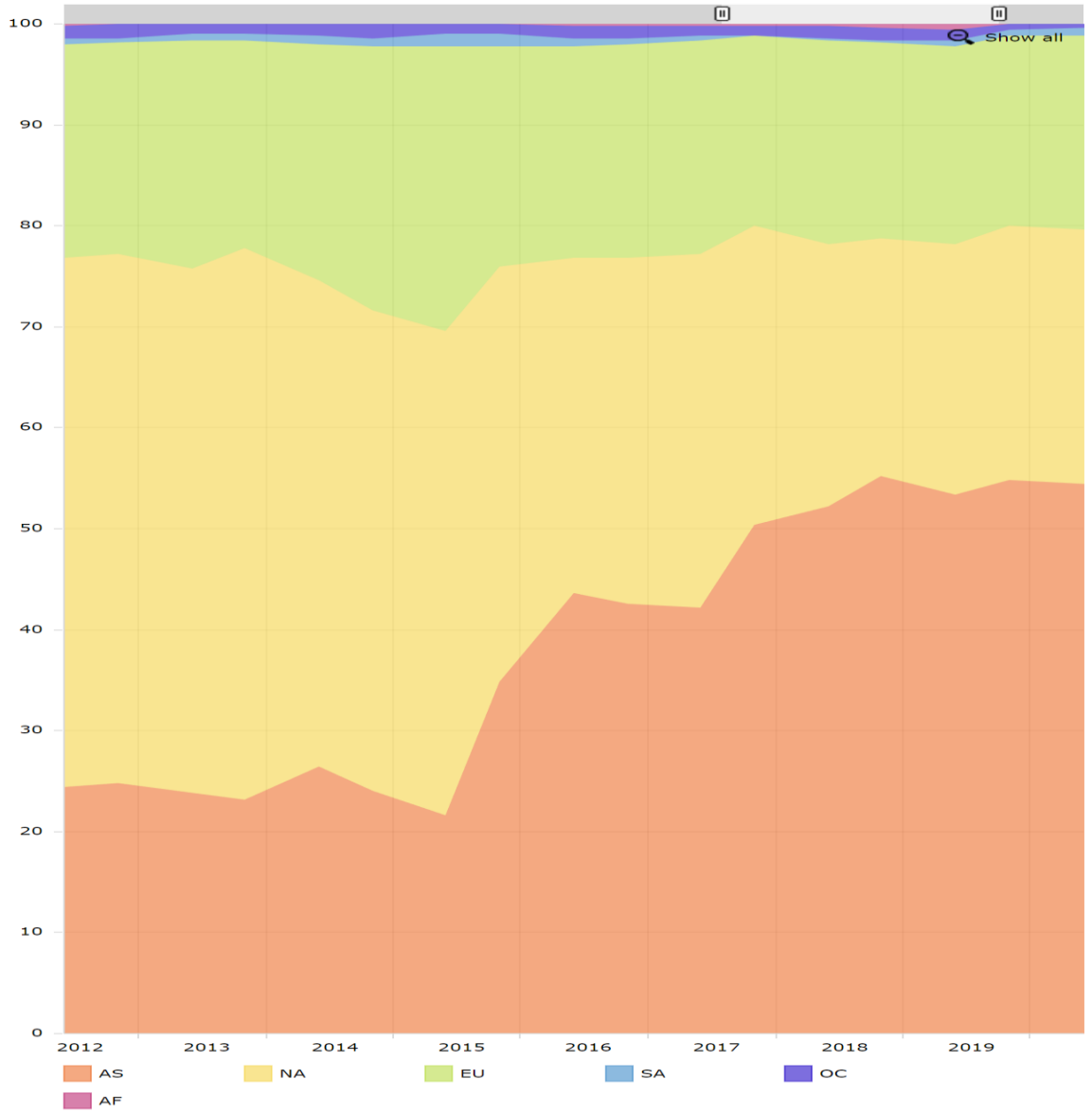
MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# TOP500

November 2022

Overtime
Continents
number of
systems share




https://www.top500.org/statistics/overtime/

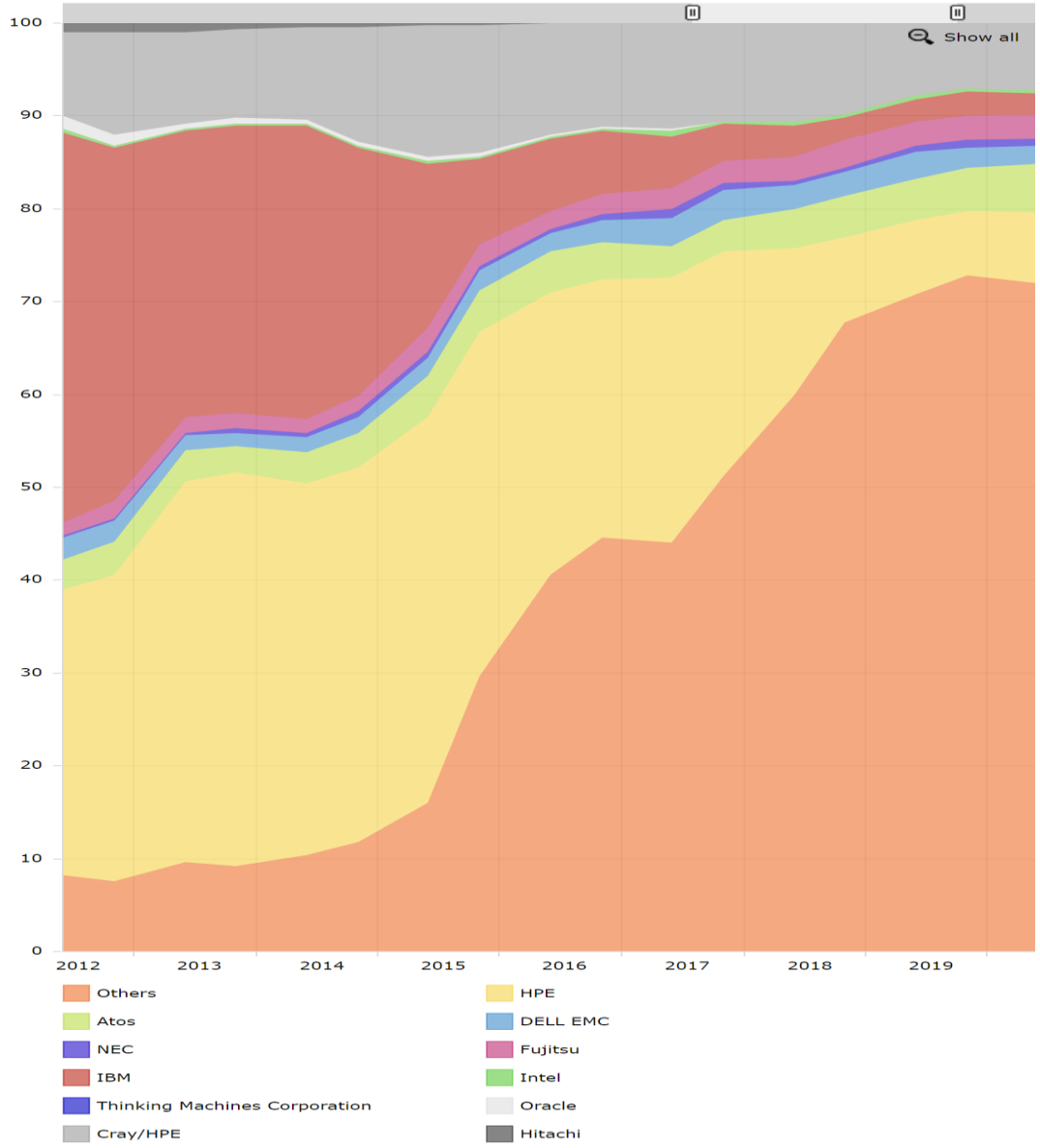
MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# TOP500

November 2022

Overtime
Vendors
systems share




https://www.top500.org/statistics/overtime/

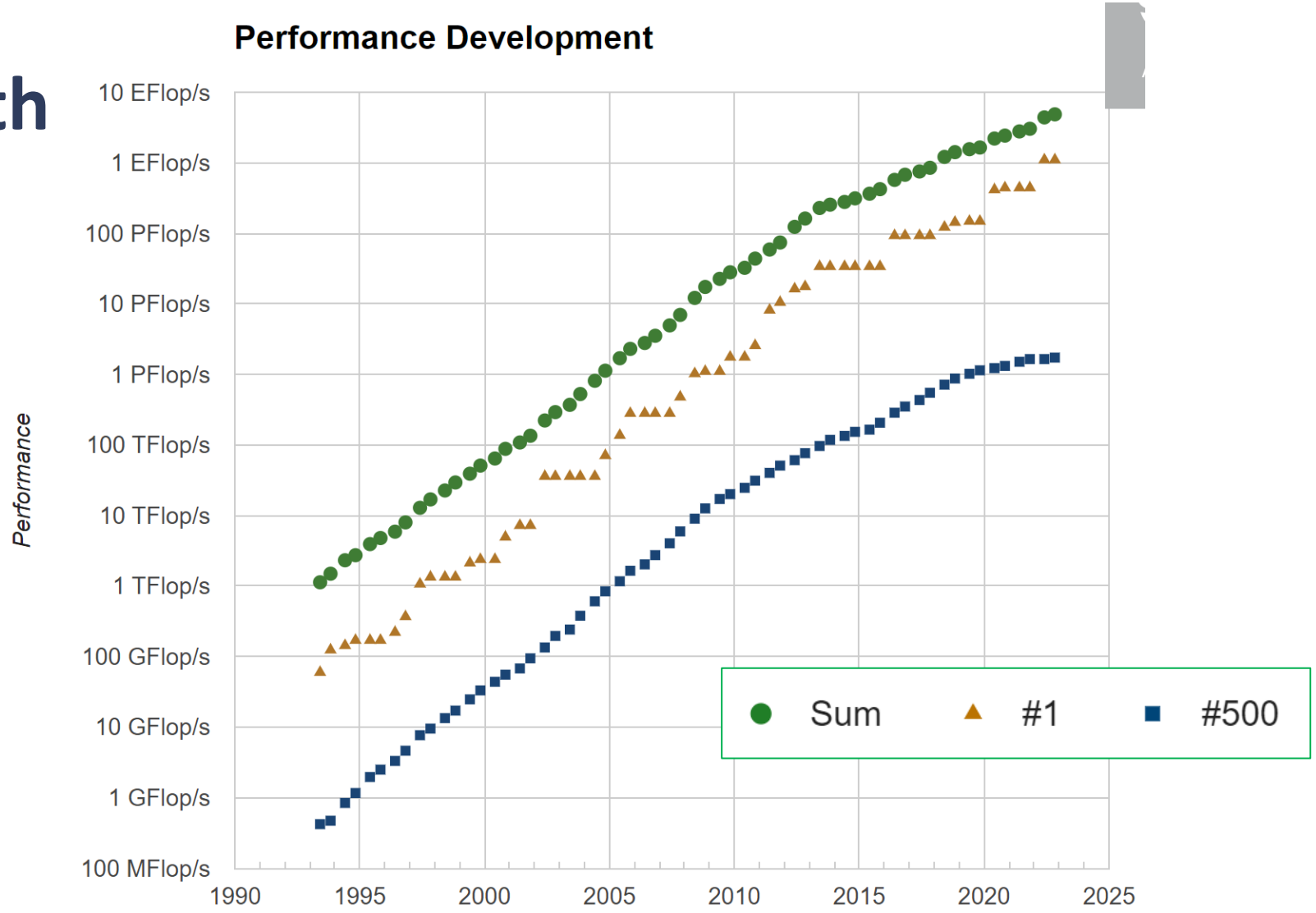
# Rate of growth

what computing powers will be reached in the next 3 years ?

- 10 **ExaFLOPSs** ($10^{19}$ FLOPS)
- 50 - 150 **Exabyte** HD ($10^{19}$ - $10^{20}$ byte)

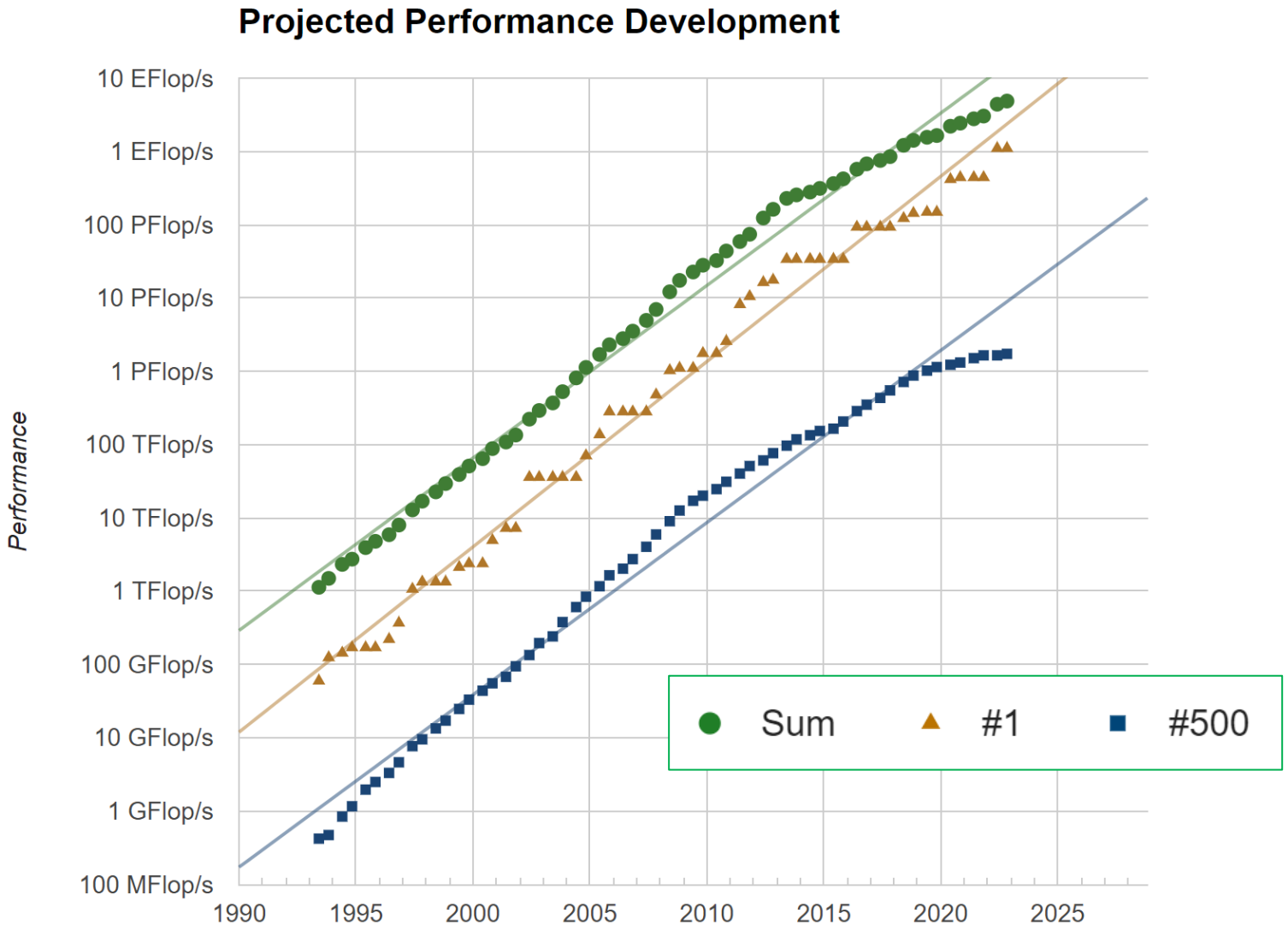
# Rate of growth

November 2022



# Rate of growth

November 2022



# Exascale computing

**Exascale computing** refers to computing systems capable of calculating at least $10^{18}$ floating point operations per second (1 ExaFLOPS)

**Fugaku** is the first supercomputer in history to hit this milestone (HPL-AI benchmark)

in April 2020, the distributed computing network **Folding@home** attained one ExaFLOPS on a real-world problem (genomic analysis)

# Exascale computing

the supercomputer **Summit** was the first supercomputer to reach one ExaFLOPS speed during a genomic analysis in 2021, achieving 1.88 ExaFLOPS in 2021
Summit is expected to reach 3.3 ExaFLOPS using mixed-precision calculations

the supercomputer **Frontier** was the first supercomputer to achieve 1.1 ExaFLOPS on a standard LINPACK benchmark (solution of a linear system of equations) in June 2022, with 64 bit arithmetic

# Exascale computing

**Exascale** computing is a significant achievement in IT

primarily it will allow improved scientific applications and better prediction such as in weather forecasting, climate modeling and personalized medicine

**Exascale** also reaches the estimated processing power of the human brain at the neural level, a target of the Human Brain Project

Wikipedia

**Human Brain Project (Wikipedia)**

**https://www.humanbrainproject.eu/en/**

# Exascale computing → Zettascale?

if advancements continue at a similar pace, Zettascale computing (performing one zettaFLOP, or a thousand times the capacity of an exascale system) could potentially be within reach by 2028

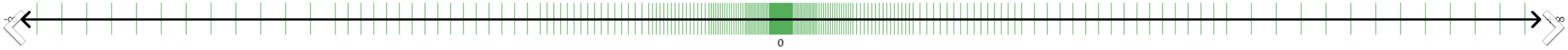
in terms of data storage, advancements in technologies such as solid-state drives (SSDs), including NVMe, and storage-class memory (SCM), are increasing both the speed and capacity of data storage. The introduction of DNA and holographic storage could potentially revolutionize data storage, offering massive capacities in incredibly small physical spaces. However, these are still largely in the research and development phase and may not be commercially viable by 2028

it's also worth noting that the rise of distributed computing and cloud storage has somewhat shifted the focus from individual machine capabilities to the capacities of interconnected systems, which can harness the combined power and storage of many machines

# A quick look at floating-point numbers

a **floating-point number** (or float) is a 64-bit representation of a real number, in which the 1-bit sign of the number, the first 52 nonzero bits of its binary representation and its 11-bit base-2 exponent are stored

any number in the interval $[-10^{308}, 10^{308}]$ can be represented with a relative error less or equal than $10^{-16}$




there are near $2^{64} \approx 10^{19}$ floating-point numbers in the interval $[-10^{308}, 10^{308}]$

# A quick look at floating-point numbers

**mixed-precision arithmetic** is for operating on inaccurate numbers with a small width and expanding them to a larger, more accurate representation
For example, two half-precision such as **fp16** (16-bit 1-5-10) or **bfloat16** (16-bit 1-8-7) numbers may be multiplied together to result in a more accurate single-precision (32-bit) number or double-precision (64-bit) number

Some platforms, including Nvidia and AMD GPUs, provide mixed-precision for this purpose, using a coarse float when possible, but expanding them to higher precision when necessary.
Iterative algorithms (that compute a sequence of improved approximations of the solution of a problem) are perfect for mixed-precision, since a coarse guess can be made and refined over many iterations until the smallest possible error in that precision is reached. Then the precision can be increased, which allows for smaller increments to be used for the approximation

Some supercomputers, e.g. Summit, utilize mixed-precision arithmetic to be more efficient with regards to memory and processing time, as well as power consumption

# Distributed supercomputers

a **distributed computer system** consists of multiple computers located in different places that are networked together and work as a single system. These computers, or nodes, share tasks and resources to achieve common goals

**World Wide Web**: probably the most recognized distributed system
**Cloud Computing Platforms**: services like Amazon Web Services (AWS), Google Cloud Platform, and Microsoft Azure
**Distributed Databases**: systems like Google's Spanner, Apache Cassandra, and Amazon's DynamoDB
**Distributed File Systems**: systems like Google File System (GFS) and Hadoop Distributed File System (HDFS)
**Blockchain Networks**: as used in cryptocurrency systems like Bitcoin and Ethereum
**Scientific Computing Projects**: projects like Folding@Home or SETI@Home utilize the unused processing power of personal computers around the world to analyze complex scientific data
**Content Delivery Networks**: systems of distributed servers that deliver content to users based on their geographic location, particularly by media-rich sites (like video streaming services), eCommerce platforms, and any site with a large, global user base

# Distributed supercomputers


**Folding@home** (**FAH** or **F@h**) is a **distributed computing** project aimed to help scientists develop new therapeutics for a variety of diseases by the means of simulating protein dynamics

this includes the process of protein folding and the movements of proteins and is reliant on simulations run on volunteers' personal computers

Folding@home is currently based at Washington University in St. Louis (USA)

https://foldingathome.org/

# Distributed supercomputers


**Folding@home** is one of the world's fastest computing systems

with heightened interest in the project as a result of the COVID-19 pandemic, the system achieved a speed of approximately 1.22 ExaFLOPS by late March 2020, and reached 2.43 ExaFLOPS by April 12, 2020 making it the world's first ExaFLOPS computing system

https://foldingathome.org/

# Distributed supercomputers

a great example of **citizen science**


https://foldingathome.org/

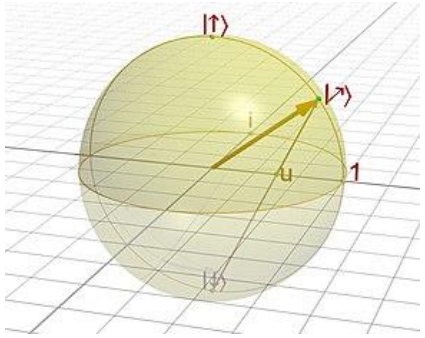
# …and beyond

## and then ZETTAscale ($10^{21}$ FLOPS)

List of hypothetical technologies - Wikipedia

- **Zettascale** computers will be able to accurately forecast the global weather for the period of approximately 2 weeks. Climate change models will become more accurate than ever before, further reducing uncertainty about temperature increases and future impacts
- **Zettascale** computing will also be able to reduce the time required for astrophysical simulations of such phenomena as black holes, neutron star mergers, and supernovae. For example, a computation which takes 1 million hours on PetaFLOPS computers and 1000 hours on ExaFLOPS machines, can be done in just one hour on ZettaFLOPS systems
- **Zettascale** ($10^{21}$ Flops) or **Yottascale** ($10^{24}$ Flops) systems might be able to accurately model the whole human brain

# Quantum computers

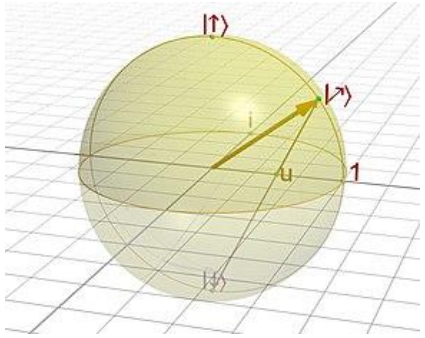



IBM Quantum System One, 20 Q-bits
the first commercial Quantum Computer

**Quantum computing** is a field of computer science that explores the use of quantum mechanics principles to process information

https://en.wikipedia.org/wiki/Quantum_computing/
https://www.dw.com/en/ibm-unveils-first-quantum-computer-in-germany/a-57909494

# Quantum computers




representation of a **qubit** as a Bloch's sphere

IBM Quantum System One, 20 Q-bits the first commercial Quantum Computer

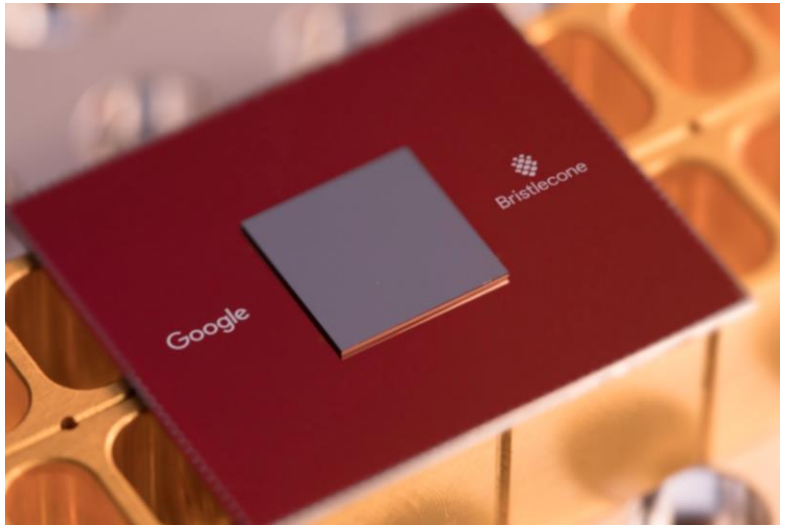
traditional computers use bits to process information, where each bit is either a 0 or a 1, while quantum computers use quantum bits or **qubits**

qubits can exist in a state of superposition, where they can be both 0 and 1 at the same time: this unique property allows quantum computers to process a vast number of computations simultaneously

qubits have the property of entanglement, where the state of one qubit becomes instantly correlated with the state of another, no matter how far apart they are

# Quantum computers

https://www.technologyreview.com/s/613507/the-new-benchmark-quantum-computers-must-beat-to-achieve-quantum-supremacy




Google Bristlecone, 72 Q-bits

https://ai.googleblog.com/2018/03/a-preview-of-bristlecone-googles-new.html

https://www.technologyreview.com/s/610274/google-thinks-its-close-to-quantum-supremacy-heres-what-that-really-means/

# Quantum computers

from McKinsey:


https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/a-game-plan-for-quantum-computing

https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/quantum-computing-use-cases-are-getting-real-what-you-need-to-know

https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/tech-forward/quantum-computing-is-coming-how-can-your-company-prepare

# Supercomputers and strategic policies

2015, US President B. Obama:
executive order authorizing the creation of new supercomputing research initiative called the **National Strategic Computing Initiative**, or **NSCI**

https://obamawhitehouse.archives.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing

https://www.wired.com/2015/07/obama-supercomputing/

# Strategic projects

## THE NATIONAL STRATEGIC COMPUTING INITIATIVE (NSCI) - USA

https://www.nitrd.gov/nsci/

https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf

Report - NATIONAL STRATEGIC COMPUTING INITIATIVE UPDATE: PIONEERING THE FUTURE OF COMPUTING (nitrd.gov)

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# Strategic projects

**The NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT OROGRAM (NITRD)**

NITRD is the US primary source of federally funded research and development in advanced Information Technologies in computing, networking, and software

https://www.nitrd.gov

MEIM | MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# Strategic projects

- AI – Artificial Intelligence
- CHuman – Computing-Enabled Human Interaction, Communication, and Augmentation
- CNPS – Computing-Enabled Networked Physical Systems
- CSP – Cyber Security and Privacy
- EdW – Education and Workforce
- ENIT – Electronics for Networking and Information Technology
- EHCS – Enabling-R&D for High-Capability Computing Systems
- HCIA – High-Capability Computing Infrastructure and Applications
- IRAS – Intelligent Robotics and Autonomous Systems
- LSDMA – Large-Scale Data Management and Analysis
- LSN – Large-Scale Networking
- SPSQ – Software Productivity, Sustainability, and Quality

**https://www.nitrd.gov/coordination-areas/**

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# Strategic projects

THE CONVERGENCE OF HIGH PERFORMANCE COMPUTING, BIG DATA, AND MACHINE LEARNING.

https://www.nitrd.gov/pubs/Convergence-HPC-BD-ML-JointWSreport-2019.pdf

# Strategic projects

## THE EXASCALE COMPUTING PROJECT- USA


**The Compelling Case for Exascale Computing (**U.S. Department of Energy (**DOE**): Office of Science (SC), National Nuclear Security Administration (NNSA)**)**

Scientific applications and software for Quality of Life, Strong Economy, National Security

https://www.exascaleproject.org/

# Strategic projects


## INTERNATIONAL EXASCALE SOFTWARE PROJECT

“The International Exascale Software Roadmap," Dongarra, J., Beckman, P. et al., Volume 25, Number 1, 2011, International Journal of High Performance Computer Applications

An international research community that organized a series of bi-annual meetings from 2009 to 2012 to develop ideas for hardware/software design to move to exascale science and to produce an overview of the state of different national and international projects to develop exascale computing

https://www.exascale.org/iesp/Main_Page

# Strategic projects

EuroHPC Joint Undertaking

## THE EUROPEAN HIGH PERFORMANCE COMPUTING JOINT UNDERTAKING (EUROHPC JU) - EU

joint initiative between the EU, European countries and private partners to develop a World Class Supercomputing Ecosystem in Europe

https://eurohpc-ju.europa.eu/

https://en.wikipedia.org/wiki/Supercomputing_in_Europe

# Strategic projects

PRACE PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

## PRACE – PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

The mission of PRACE is to enable 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 whose representative organisations create a pan-European supercomputing infrastructure, providing access to computing and data management resources and services to researchers from academia and industry for large-scale scientific and engineering applications.

https://prace-ri.eu/about/introduction/

https://prace-ri.eu/wp-content/uploads/PRACE_Annual_Report_2020.pdf

# An interesting report

## Exascale Computing Market Analysis, Innovation Trends and Current Business Trends by 2026

The research report assesses ongoing market trends, as well as the factors that are poised to enhance the market growth during the analysis timeframe. It encompasses major market restraints which may hamper the market growth.
The report also comprises of the key manufacturers which formulate the competitive terrain of the Exascale Computing market and also highlights the major market segmentations

https://www.business-newsupdate.com/exascale-computing-market-328354

# Supercomputers for AI

**the race for AI supremacy between nations and top global corporations**

Artificial Intelligence (AI) relies on and benefits from the global race for faster supercomputers

Tesla, Facebook, Microsoft, and NVIDIA are working on huge supercomputers to power their artificial intelligence R&D and product potential

MEIM | MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# a TESLA project

## Tesla Dojo is a supercomputer for computer vision video processing and recognition

1.1 ExaFLOP (ExaPOD)
1.3 terabytes of SRAM
13TB of high-bandwidth DRAM

https://www.youtube.com/watch?v=1mFva7qa-5Q&ab_channel=TheTeslaSpace

# a META project

Meta AI

## AI RSC - Artificial Intelligence Research Super Cluster

"Meta (Facebook) is announcing that they have designed and built the AI Research SuperCluster (RSC) — which will be among the fastest AI supercomputers running today and will be the fastest AI supercomputer in the world when it's fully built out in mid-2022. Meta's researchers have already started using RSC to train large models in natural language processing (NLP) and computer vision for research, with the aim of one day training models with trillions of parameters"

https://www.youtube.com/watch?v=fZnykn1tDSE

https://ai.facebook.com/blog/ai-rsc/

MEIM | MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

# a META project

Meta AI

## AI RSC - Artificial Intelligence Research Super Cluster

at full strength, it achieves almost 5 ExaFLOPS of computing power

this level of performance is achieved through the use of 2,000 NVIDIA DGX A100 systems as RSC's compute nodes, for a total of 16,000 NVIDIA A100 Tensor Core GPUs, connected via an NVIDIA Quantum InfiniBand 16 Tb/s fabric network

80 PB of cache, over a half exabyte of bulk storage, offering up to 16TB/s

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

https://ai.facebook.com/blog/supercomputer-meta-research-supercluster-2023/

# a META project

Meta AI

## AI RSC - Artificial Intelligence Research Super Cluster

- ✓ LLaMA
- ✓ No Language Left Behind
- ✓ Universal speech translator
- ✓ Theorem proving

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

https://ai.facebook.com/blog/supercomputer-meta-research-supercluster-2023/

# Microsoft AI supercomputer

Microsoft

## Microsoft's AI at Scale initiative

the supercomputer developed for OpenAI is a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server

it would rank in the top five of the top500 list

it is hosted in Azure

https://www.youtube.com/watch?v=Rk3nTUfRZmo&ab_channel=MicrosoftMechanics/

# NVIDIA

**DGX GH200**, massive memory supercomputing for emerging AI

1 ExaFLOPS of FP8 AI

256 NVIDIA Grace Hopper Superchips, all connected with NVIDIA NVLin

144 terabytes (TB) shared-GPU-memory

GPU-to-GPU bandwidth 900 GB/s

https://www.nvidia.com/en-us/data-center/dgx-gh200/

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

OpenAI

## Sam Altman: CEO of OpenAI calls for US to regulate artificial intelligence

Sam Altman, the CEO of OpenAI, the company behind ChatGPT, testified before a US Senate committee on May 17-th 2023 about the possibilities - and pitfalls - of the new technology

In a matter of months, several AI models have entered the market

Altman said a new agency should be formed to license AI

However, the technology is moving so fast that legislators also wondered whether such an agency would be capable of keeping up

He also admitted the impact that AI could have on the economy, including the likelihood that AI technology could replace some jobs, leading to layoffs in certain fields

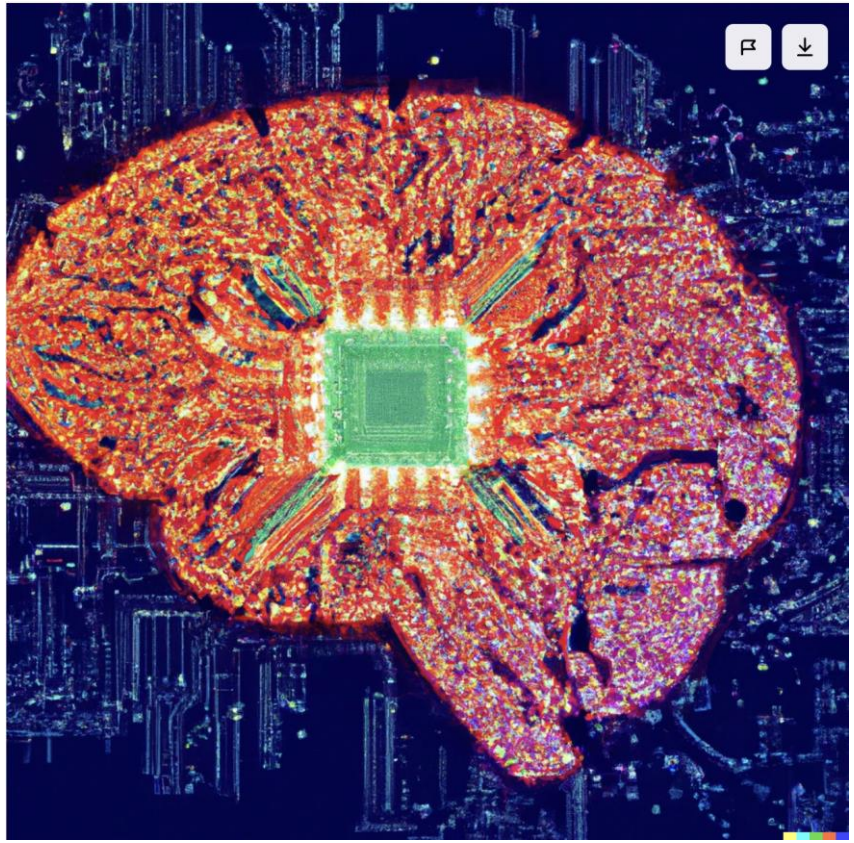
Altman added, however, that he is "very optimistic about how great the jobs of the future will be"

https://www.bbc.com/news/world-us-canada-65616866

MASTER MEIM 2022-2023

# Thanks for your attention

# …… Digital Tech HPC continues



DALL_E prompt: a supercomputer depicted as a brain for Artificial intelligence applications