A Scalable Artificial Intelligence system for Machine and Deep Learning Research and Training at Sapienza Università di Roma

Proponente: S. Giagu
Co-proponenti: R. Faccini, S. Leonardi

SAPIENZA
UNIVERSITÀ DI ROMA

PARTICIPANTS

- **PIs:** S. Giagu, R. Faccini, S. Leonardi
- **Participants:** P. Bagnaia, R. Belibani, C. Bini, A. Capone, F. Cesi, G. D'Agostini, S. De Cecco, D. Del Re, A. Di Domenico, F. Fattaposta, P. Gauzzi, I.R. Giardina, F. Lacava, E. Longo, V. Loreto. C. Luci, V. Marinari, F. Meddi, G. Organtini, A. Policicchio, M. Raggi, S. Rahatlou, F. Ricci Tersenghi, E. Rogora, F. Santanastasio, A. Vitaletti
- **Departments:** Fisica, DIAG, Architettura e Progetto, Neurologia e Psichiatria, Matematica

EQUIPMENT DESCRIPTION

- Advanced Machine Learning algorithms and in particular Deep Neural Networks made incredible advances in the past few years transforming society and shaping the way researchers analyze data

- As the panorama of Artificial Intelligence applications expands the algorithms are getting increasingly complex and demand unprecedented levels of computer power

- The NVIDIA DGX-2 system allows Sapienza's researchers to:
  - contribute and compete worldwide to the design and development of the next generation of Machine Learning algorithms and applications
  - to train the future experts in the field of AI

THE NVIDIA DGX-2 SYSTEM

- state of the art super computer projected to develop AI/DL algorithms and for HPC
- put Sapienza University in the “top500 world-wide supercomputer system” list (position #325 at November 2018), together in Italy with ENI, CINECA e ENEA

- computing power: 2 petaFLOPS
- 16 GPUs interconnected with an high performance bus
- DL performances: x10 the best AI available systems
- software stack designed for ease of use & scalability

SPECIFICHE DEL SISTEMA

GPU	**16 NVIDIA® Tesla V100**
Memoria della GPU	**512 GB totali**
Prestazioni	**2 petaFLOPS**
Core NVIDIA CUDA®	**81920**
NVIDIA Tensor Core	**10240**
NVSwitch	**12**
Consumo energetico massimo	**10 kW**
CPU	**Dual Intel Xeon Platinum 8168, 2,7 GHz, 24 core**
Memoria di sistema	**1,5 TB**
Rete	**8 Infiniband 100 Gb/sec/ Ethernet 100 GigE Dual 10/25 Gb/sec**
Spazio di archiviazione	**SO: 2 SSD NVME 960 GB Memoria interna: SSD NVME 30 TB (8 x 3,84 TB)**
Peso del sistema	**154,2 kg**
Dimensioni del sistema	**Altezza: 440 mm Larghezza: 482,3 Lunghezza: 795,4 mm**

NVIDIA DGX-2

16 TESLA V100 32 GB COMPLETAMENTE CONNESSE

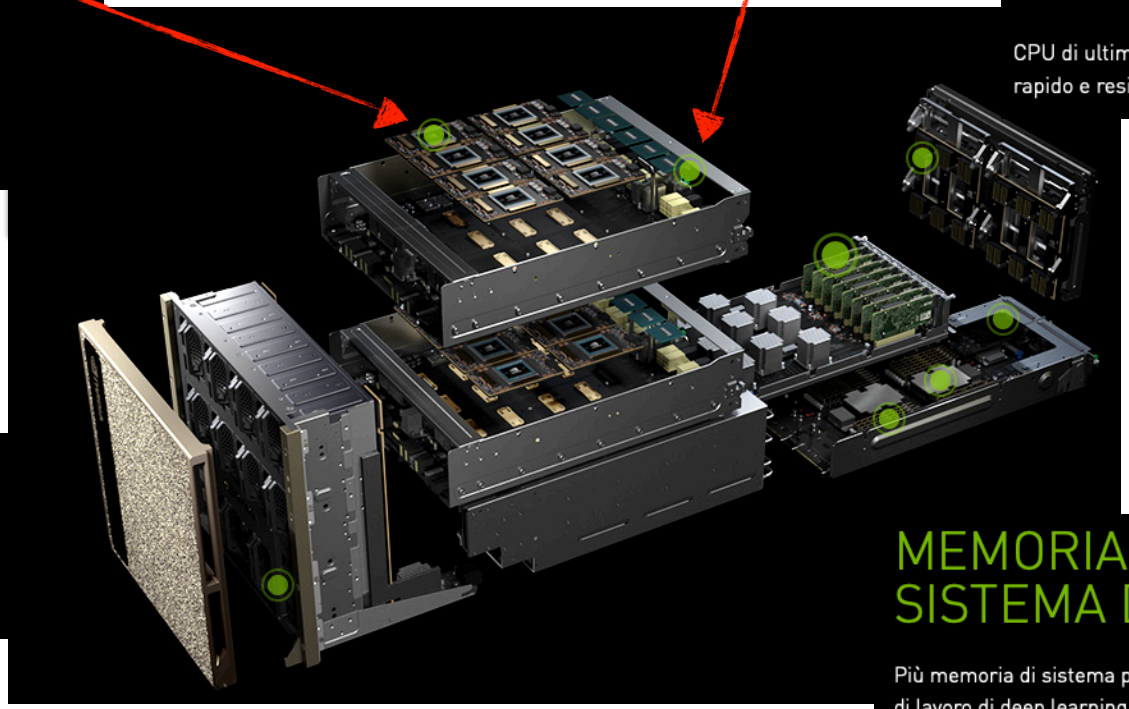
Memoria totale a elevata larghezza di banda da 0,5 TB per modelli di deep learning più complessi

12 NVSWITCH

Larghezza di banda bi-sezione da 2,4TB/s

2 XEON PLATINUM

CPU di ultima generazione per boot più rapido e resiliente e gestione storage



SSD NVME 30 TB

Integrazione rapida di gradi set di dati in cache

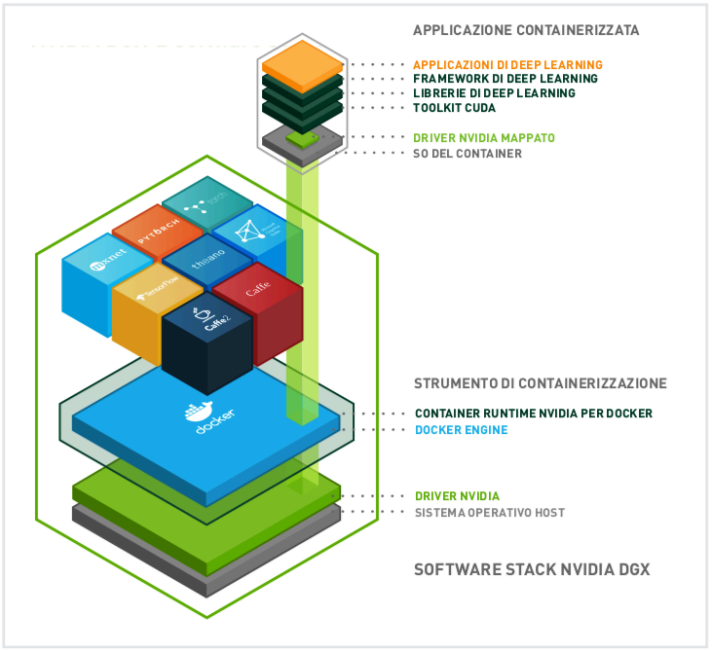
MEMORIA DI SISTEMA DA 1,5 TB

Più memoria di sistema per gestire carichi di lavoro di deep learning di grandi dimensioni

POSSIBILE APPLICATIONS AND USES

- algorithms and applications for AI/ML/Deep-NN in different fields: basic science, applied science, engineering, robotics, medicine, neuroscience, architecture, economics and finance, social sciences, politics, applications in civil and commercial law, …
- High Performance Computing, simulations of complex systems and fluid dynamics, …
- teaching activities related to data-science, machine learning, physics, mathematics, …

- the DGX software stack supports all the popular AI/DL and HPC applications and supports virtualization for easy personalization of each specific research project



LOGISTIC AND EQUIPMENT AVAILABILITY

- when available the system will be hosted in the computing center of the “Servizio Impianti Calcolo e Reti INFN” in the Dipartimento di Fisica G. Marconi (CU013)

- hosts the computing/network infrastructure of the physics department
- rack space for the DGX-2 system already prepared
- SICR/INFN provides all needed IT support for the system

- plan to initiate the tender procedure as soon as possible
  - currently only two companies (NVIDIA and λ-Lab) have in catalogue systems with equivalent specs as the DGX-2
  - goal: system ready to be used by end of 2019 / start of 2020

ACCESS AND USAGE RULES

- to manage and regulate the use of the system in the most efficient way, a regulations governing use document has been prepared
- setup a dedicated Resource Management committee that:
  - manages the subdivision of the calculation resources of the system (machine time for dedicated runs, percentage of resources usable in terms of GPU, memory and disk storage space) on the basis of requests for use received, on a half-yearly basis
  - regulates the methods of access and use of the system, the responsibilities and obligations of users in terms of information security and protection of health and safety at work
- detailed instructions for users including description and documentation of the available hardware and software resources, usage examples, etc. will be available on a dedicated web page hosted on the site of the physics department (under preparation)

INITIAL RESEARCH LINES

- an initial set of research projects has been submitted with the proposal, each of great interest in the field of basic research and/or in applied AI and/or in teaching and training aspects
- the computing power and available resources provided by the DGX-2 system allow for seamless support of many additional research lines/projects

1. Sviluppo e applicazione di reti neurali innovative per inferenza ultra-veloce su FPGA per applicazioni real-time: use-case “sistemi di filtro in tempo reale per esperimenti delle alte energie al Large Hadron Collider del CERN”

2. Sviluppo di algoritmi basati su Deep Generative Adversarial Networks (GAN) e su Variational Auto Encoders (VAE) per la simulazione veloce di eventi: use-case: “simulazione di processi fisici in esperimenti di fisica delle particelle, data-augmentation di immagini mediche (RMN, CT, …) per addestrare algoritmi di segmentazione immagini”

3. Sviluppo di applicazioni di Machine Learning in diagnostica medica e monitoraggio di pazienti in terapia domiciliare tramite algoritmi di sentiment analysis

4. Uso di Deep Learning per l'integrazione e l'embedding di dati biomedici nella network medicine e la predizione delle funzioni dei geni nella medicina di precisione

5. Simulazione di modelli di Deep Neural Network in connessione con modelli di fisica statistica per cercare di spiegarne e capirne il funzionamento

6. Studio di metodi algoritmici efficienti di riduzione della dimensione spaziale delle rappresentazioni di informazioni testuali per Deep Learning

7. Applicazione di metodi di Deep Learning per la decifrazione di lingue antiche a partire dall'analisi delle immagini delle scritture

8. Sviluppo di progetti di ricerca e partecipazione a competizioni di machine learning da parte di studenti dei corsi di laurea in data science

9. Studio di nuovi metodi per il calcolo su sistemi HPC eterogenei (CPU+GPU)

CONTACTS

- PIs:
  - Stefano Giagu:
    - stefano.giagu@uniroma1.it
    - Dipartimento Fisica – G.Marconi – ufficio 318 – telefono: +39 0649914407
  - Riccardo Faccini:
    - riccardo.faccini@uniroma1.it
    - Dipartimento Fisica – G.Marconi – ufficio 254 – telefono: +39 0649914798
  - Stefano Leonardi
    - leonardi@diag.uniroma1.it
    - DIAG – ufficio B205 – telefono: +39 0677274022

- indirizzo fisico del laboratorio e contatto tecnico:
  - Dipartimento di Fisica G. Marconi, sala calcolo SICR – piano terra - stanza 029
  - contatto locale: Dr. E. Pasqualucci responsabile centro SICR. tel +39 064991-4873/4411