

HPC support and use of GPUs for scientific applications

Gioacchino Vino (INFN, Bari)

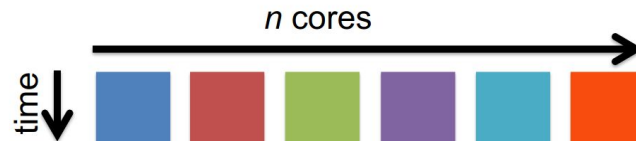
2° Congresso della Sezione INFN e del Dipartimento di Fisica di Bari, 03-04 Feb 2022

Computing: HTC vs HPC

There are two different approaches to scientific computing:

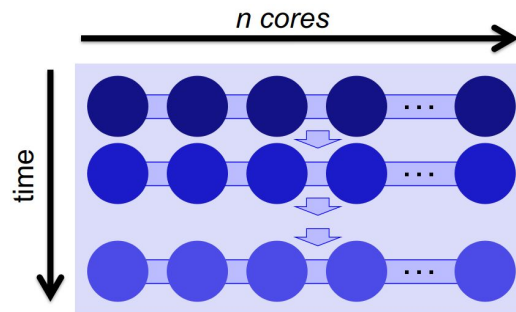
- **High-Throughput Computing (HTC)**

- Large workflows of numerous, small and independent tasks
- Executes on physically distributed resources using grid-enabled technologies



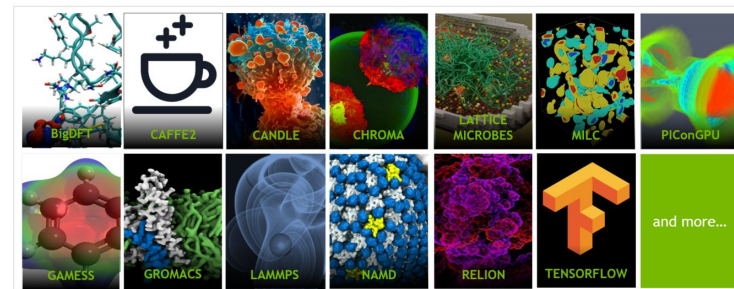
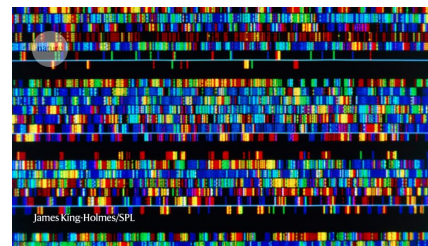
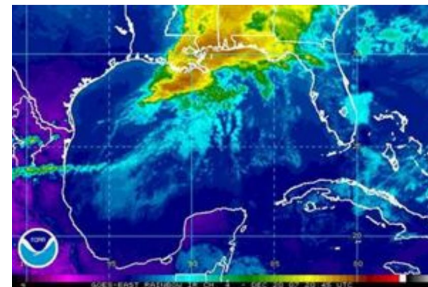
- **High-Performance Computing (HPC)**

- Large workflows of highly-interdependent sub-tasks
- Frequent and rapid exchanges of intermediate results is required to perform the computations
- Parallel processing over many processors



HPC Scientific Applications

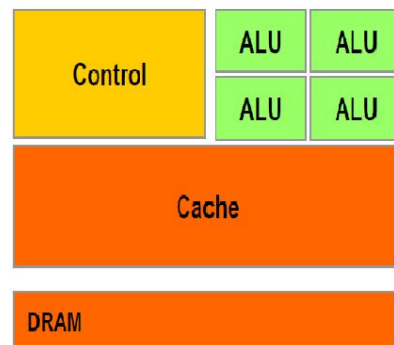
- Bioinformatics
- Artificial Intelligence algorithms
- Simulations of physical phenomena such as:
 - Weather forecasting
 - Earthquake forecasting
 - Galaxy formation
 - Molecular dynamics
- Almost all problems that involve many floating point operations.



Why GPU in HPC?

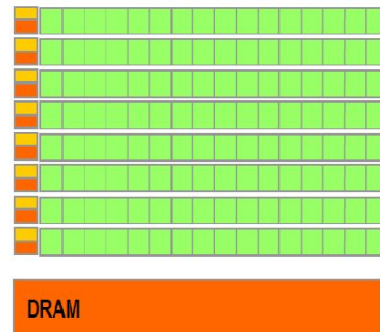
Control Processing Unit (CPU):

- Designed to handle complex tasks
- Low-level parallelism (<100 cores)



Graphical Processing Unit (GPU):

- Massively parallel hardware architecture (> 5000 cores)
- High performance of floating point arithmetic



Make them suited for many scientific workloads on HPC clusters

ReCaS HPC/GPU Cluster

Hardware Facility:

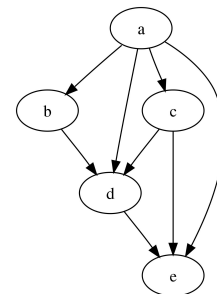
- Nodes: 10
- GPUs: 18 (V100 and A100 Nvidia GPU)
- Cores: 1755
- RAM: 13.7 TB
- Local Storage: 55 TB (SSD/HDD)
- Parallel File System: ReCaS storage based on IBM GPFS (3800TB)
- Bandwidth between nodes: 10 Gbps



ReCaS HPC/GPU Cluster: Services

Services ready-to-use:

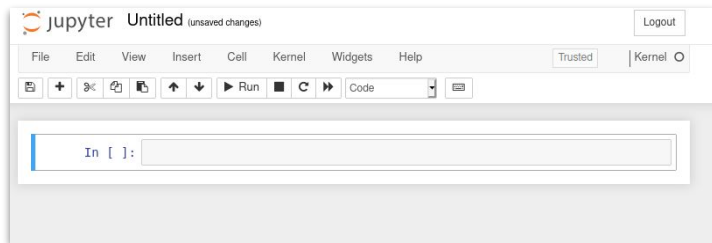
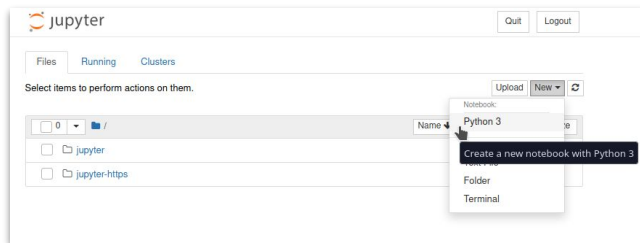
- Interactive remote GPU-based IDE services:
 - Jupyter Notebook
 - “web service for interactive computing across all programming languages”
 - Rstudio
 - “An integrated development environment for R”
- Job Scheduler:
 - Support to GPU-based workflows represented as Directed Acyclic Graphs (DAG)



ReCaS HPC/GPU Cluster: Services

Jupyter Notebook remote IDE

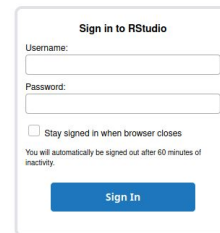
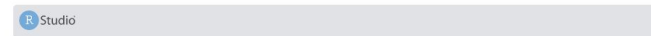
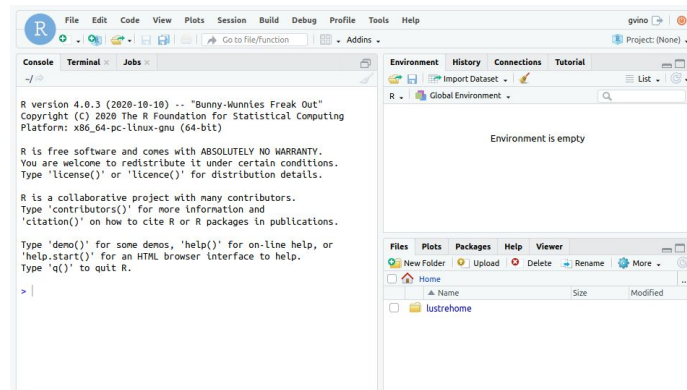
- After authentication, users have access to their home directory in the ReCaS distributed storage (GPFS)
- Users can immediately create a new Python3 script
- The Jupyter IDE (Integrated Development Environment) will be available and users can already write code and execute it
- Python modules can be installed directly within the code



ReCaS HPC/GPU Cluster: Services

RStudio remote IDE

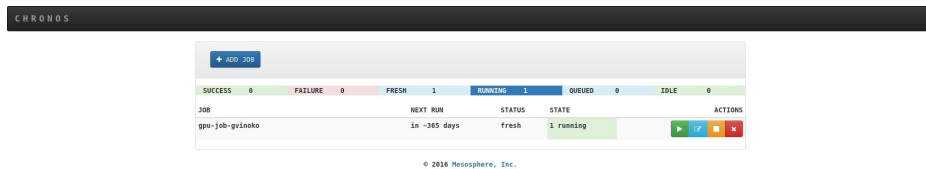
- After authentication, users have access to their home directory in the ReCaS distributed storage (GPFS)
- The Rstudio IDE (Integrated Development Environment) will be available and users can already write code and execute it
- R modules can be installed directly within the code

A 'Sign in to RStudio' dialog box. It contains a 'Username:' label followed by a text input field, and a 'Password:' label followed by a password input field. Below these is a checkbox labeled 'Stay signed in when browser closes'. A note states: 'You will automatically be signed out after 60 minutes of inactivity.' At the bottom is a blue 'Sign In' button.

ReCaS HPC/GPU Cluster: Services

Job Scheduler (Chronos)

- Provides an intuitive and simple User Interface (UI) where to check job status
- New jobs can be submitted using UI or via command line using a JSON file describing the job
- Manages heterogeneous requests:
 - 2 GPU / 4 CPU / 20 GB RAM
 - 100 CPU / 8GB RAM

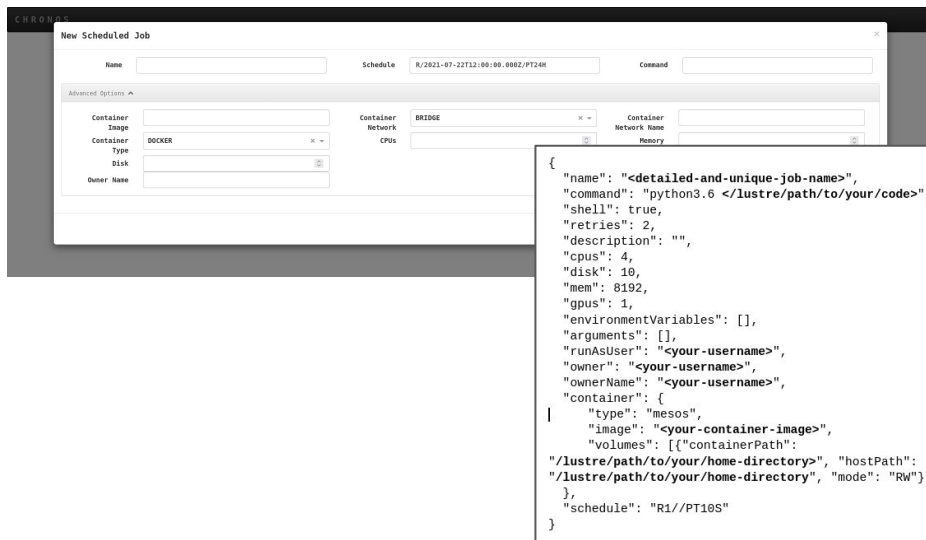


CHRONOS

+ ADD JOB

SUCCESS	FAILURE	FRESH	RUNNING	QUEUED	IDLE	
0	0	1	1	0	0	
JOB	NEXT RUN	STATUS	STATE	ACTIONS		
gpu-job-gvino	in ~365 days	fresh	1 running			

© 2016 Mesosphere, Inc.



CHRONOS

New Scheduled Job

Name Schedule R/2021-07-22T12:00:00.000Z/PT24H Command

Advanced Options

Container Image Container Type DOCKER Container Network Name Container Network CPU BRIDGE

```
{
  "name": "<detailed-and-unique-job-name>",
  "command": "python3.6 </lustre/path/to/your/code>",
  "shell": true,
  "retries": 2,
  "description": "",
  "cpus": 4,
  "disk": 10,
  "mem": 8192,
  "gpus": 1,
  "environmentVariables": [],
  "arguments": [],
  "runAsUser": "<your-username>",
  "owner": "<your-username>",
  "ownerName": "<your-username>",
  "container": {
    "type": "mesos",
    "image": "<your-container-image>",
    "volumes": [{"containerPath":
"/lustre/path/to/your/home-directory", "hostPath":
"/lustre/path/to/your/home-directory", "mode": "RW"}]
},
  "schedule": "R1/PT10S"
}
```

ReCaS HPC/GPU Cluster: Under the hood

Apache Mesos:

- Unifies all cluster resources in a single virtual entity
- Multi-users
- High Availability
- Manages a lot number of nodes



Marathon:

- Runs long running services on top of Apache Mesos
- High Availability
- Load balancing



Chronos:

- Job scheduler for Apache Mesos
- Supports depending and periodic jobs



ReCaS HPC/GPU Cluster: Under the hood

Docker container:

- Contains software, code, libraries and dependencies
- Isolates applications from the machine where it is executed
- Images are light, standalone and contain all necessary to be run
- Official images are available (Nvidia, TensorFlow, ...)



ReCaS HPC/GPU Cluster policy on Docker containers:

- Mandatory for security purpose
- Jupyter Notebook and Rstudio containers have been developed in-house because the majority of the supported use cases needs them
- Not all users' containers can be developed in-house
- An [INFN course](#) and a [ReCaS tutorial](#) are available to speed-up the user learning process

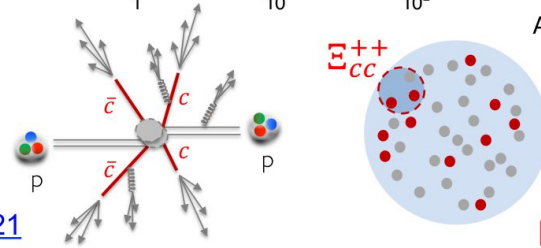
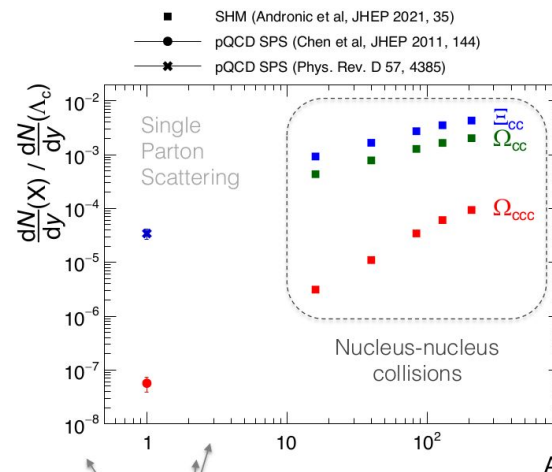
Motivation and challenges

Working Team:

- Domenico Elia
- Annalisa Mastroserio
- Domenico Colella
- Gioacchino Vino
- David Chinellato (CERN)

Multi-charm baryons: from low to high density QCD

- Charm production in general: almost exclusive to hard scatterings due to large mass ($\sim 1275 \text{ MeV}/c^2$)
- Formation of Ξ_{cc}^{++} , Ω_{cc}^{+} , Ω_{ccc}^{++} : extremely unlikely in single parton scattering (unlike e.g. J/ψ)
- Multi-parton interactions and multi-charm: **multiple charm quarks combine into hadrons**
- In nuclear collisions:
 - High density of charm quarks leads to much larger multi-charm population
 - Described by SHM (g_c) and coalescence
 - Enormous dynamic effect!



[D. D. Chinellato, ALICE 3 workshop 18-19.10.2021](#)

Multi-charm baryons in ALICE 3



2

Use case: Multi-charm reconstruction with ML in ALICE 3

ML method and analysis chain

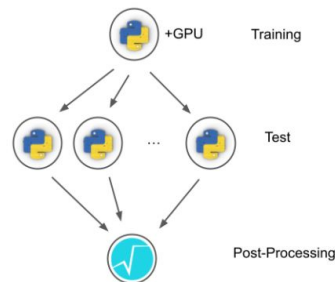
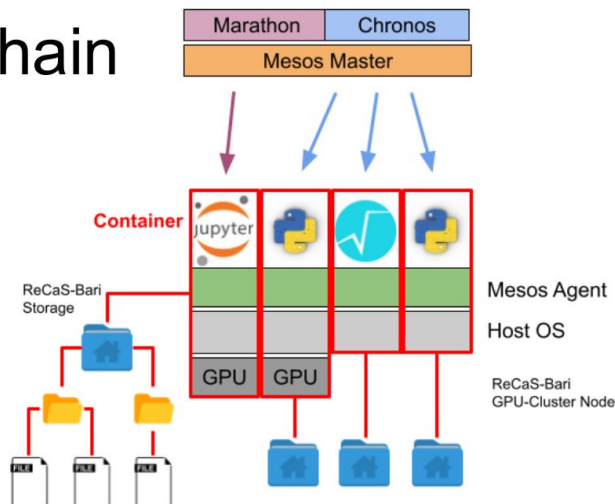
ML analysis chain (Gioacchino):

- fully developed from scratch
- **ReCaS-Bari GPU-Cluster used:**
 - ✓ nodes equipped with NVIDIA A100 or V100
 - ✓ cluster managed with Apache Mesos
 - ✓ services deployed with Mesos and Docker Containers

Each container has access to ReCaS-Bari Storage (3.8 PB)

Remote IDE (Jupyter Notebook and Rstudio) with access to GPU are available with Marathon (Development phase)

Analysis can be submitted as a set of dependent tasks (Directed Acyclic Graph) with Chronos (Production phase)



Domenico Elia

ALICE 3 HF WG meeting / 19.1.2022

5

Use case: Multi-charm reconstruction with ML in ALICE 3

Preliminary results:

- better than standard selection at low p_T
- up to a factor of 4-5x improvement for $p_T < 2$ GeV/c

Impact on future measurements:

- ML-based selection has potential to allow measurement of multi-charm down to 0 p_T
- included in the ALICE 3 Letter of Intent currently under preparation

Work ongoing:

- still room to improve ML, in particular at low p_T , eg with p_T -dedicated training

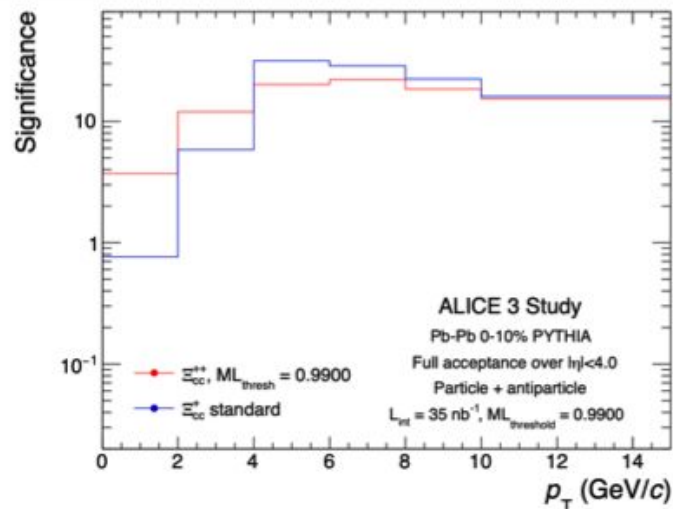


Figure 33: Ξ_{cc}^{++} significance in 0-10% central Pb-Pb collisions at $\sqrt{s_{NN}} = 5.52$ TeV as a function of p_T with a 2.0 T magnetic field using standard selections and using machine learning.

Other use cases Cluster

Current running applications and their speed-up vs CPU execution:

- CNR/IREA: x100
- Whole-genome Sequencing: x10
- Machine Learning Algorithms: x40

Strong interest of other groups to use the ReCaS-Bari HPC/GPU Cluster:

- GPU applications (Pompili, ...)
- AI applications

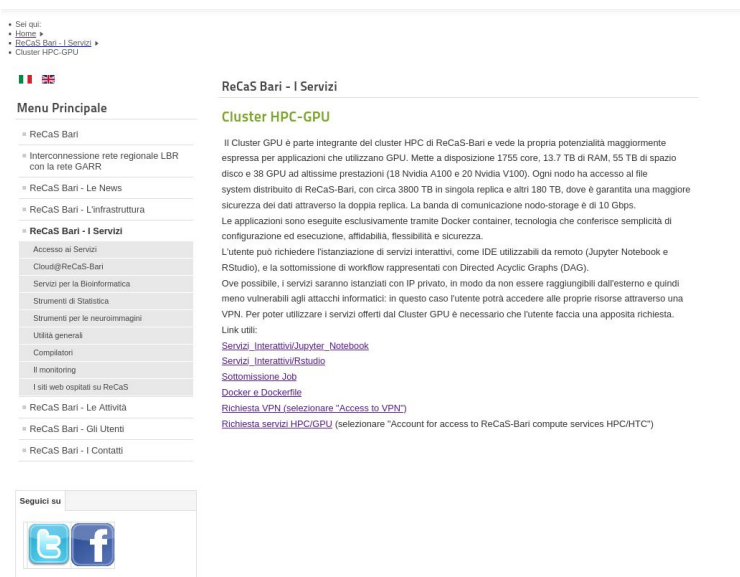
Future Developments

- **Kubernetes** will replace Apache Mesos since it overcomes some known limitations
- Chronos will be replaced with a more complex workflow scheduler, like **Apache Airflow**
- Adding **Apache Spark** to the service portfolio



Additional information

At this [URL](#) additional information about the ReCaS HPC/GPU Cluster can be found, together with some guides and tutorials about Docker



The screenshot displays the ReCaS Bari website interface. At the top, there is a navigation bar with links: "Sei qui", "Home", "ReCaS Bari - I Servizi", and "Cluster HPC-GPU". Below this, a "Menu Principale" section lists various services and resources, including "ReCaS Bari", "Interconnessione rete regionale LBR", "ReCaS Bari - Le News", "ReCaS Bari - L'Infrastruttura", and "ReCaS Bari - I Servizi". The "ReCaS Bari - I Servizi" menu is expanded, showing a list of services such as "Accesso ai Servizi", "Cloud@ReCaS-Bari", "Servizi per la Bioinformatica", "Strumenti di Statistica", "Strumenti per le neuroimmagini", "Utilità generali", "Compilatori", "Il monitoring", "I siti web ospitati su ReCaS", "ReCaS Bari - Le Attività", "ReCaS Bari - Gli Utenti", and "ReCaS Bari - I Contatti". Below the menu, there is a "Seguici su" section with social media icons for Twitter and Facebook.

ReCaS Bari - I Servizi

Cluster HPC-GPU

Il Cluster GPU è parte integrante del cluster HPC di ReCaS-Bari e vede la propria potenzialità maggiormente espressa per applicazioni che utilizzano GPU. Mette a disposizione 1755 core, 13.7 TB di RAM, 55 TB di spazio disco e 38 GPU ad altissime prestazioni (18 Nvidia A100 e 20 Nvidia V100). Ogni nodo ha accesso al file system distribuito di ReCaS-Bari, con circa 3800 TB in singola replica e altri 180 TB, dove è garantita una maggiore sicurezza dei dati attraverso la doppia replica. La banda di comunicazione nodo-storage è di 10 Gbps. Le applicazioni sono eseguite esclusivamente tramite Docker container, tecnologia che conferisce semplicità di configurazione ed esecuzione, affidabilità, flessibilità e sicurezza. L'utente può richiedere l'istanziamento di servizi interattivi, come IDE utilizzabili da remoto (Jupyter Notebook e RStudio), e la sottomissione di workflow rappresentati con Directed Acyclic Graphs (DAG). Ove possibile, i servizi saranno istanziati con IP privato, in modo da non essere raggiungibili dall'esterno e quindi meno vulnerabili agli attacchi informatici: in questo caso l'utente potrà accedere alle proprie risorse attraverso una VPN. Per poter utilizzare i servizi offerti dal Cluster GPU è necessario che l'utente faccia una apposita richiesta. Link utili: [Servizi Interattivi/Jupyter Notebook](#), [Servizi Interattivi/RStudio](#), [Sottomissione Job](#), [Docker e Dockerfile](#), [Richiesta VPN \(selezionare "Access to VPN"\)](#), [Richiesta servizi HPC/GPU \(selezionare "Account for access to ReCaS-Bari compute services HPC/HTC"\)](#)

**THANKS
FOR YOUR
ATTENTION**