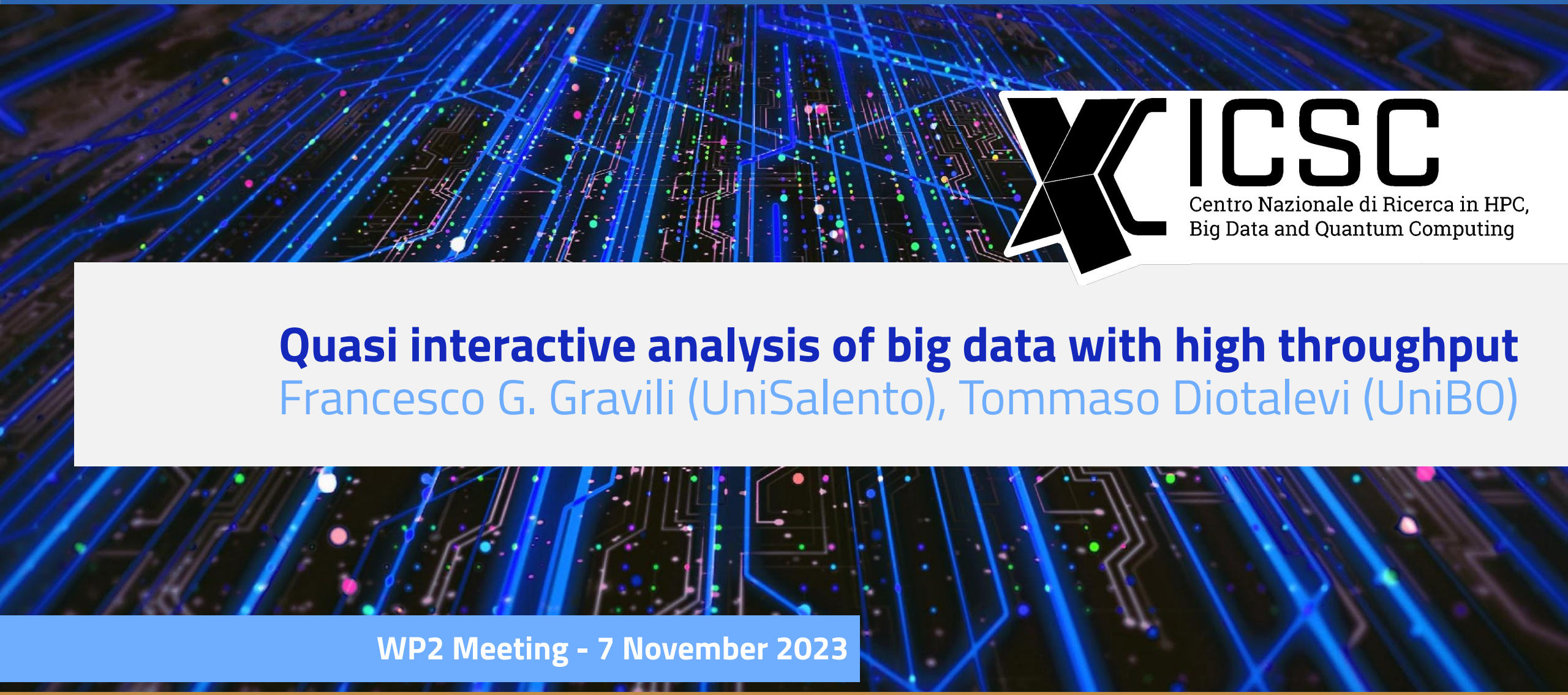
Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing



Quasi interactive analysis of big data with high throughput

Francesco G. Gravili (UniSalento), Tommaso Diotalevi (UniBO)

WP2 Meeting - 7 November 2023

Introduction to Leonardo HPC cluster: CINECA Course 27th, October 2023

Main page and resources available at this link: login as guest to access the uploaded material

Key points to start thinking, sooner or later:

- User guide for shared resources
- Technical details about the Leonardo HPC Infrastructure (mainly for WP5 interests)
- Filesystems based on Lustre: important to organize your tools, i.e. code, input/output, temporary files
- Access to clusters based on 2FA by default, using smallstep Identity Provider
- Structure splitted in login and computing nodes: the operative scheduler is SLURM on Leonardo. Conflicts with Dask?
- Accounts based on a CPU-hours budget: important to evaluate job cost, see slide 7

Current status

Activities ongoing

CMS:

- Vector Boson Scattering ssWW analysis in hadronic tau and light lepton (*INFN e Univ. Perugia*);
- Heavy Neutral Lepton search on heavy neutrinos in the D_S decays (*INFN e Univ. Bologna*);
  - Studies for performance comparison speed-up, using in-site metrics and RDF monitoring, talk WP5 (3 Nov);
- top quark+MET analysis (*INFN e Univ. Napoli*). Still in an explorative phase, talk WP5 (24 May);
- Search of rare events in $\tau \rightarrow 3\mu$ (*INFN e Univ. Bari*). Explorative phase with RDF, currently porting the old ntupliser to flat data format.

} Joint WP2 meeting (2 May)

LATEST

ATLAS:

- Search for new phenomena in events with two opposite-charge leptons, jets and missing transverse momentum (*UniSalento e INFN Lecce*). Interest towards an interactive analysis, trying to solve issue related to data format and policy. *Currently in a joint effort with INFN e Univ. Napoli.*
- *UniCal*: possible applications within several subjects: top physics, flavour tagging, trigger analysis.

Future colliders:

- Feasibility studies started on FCCee pseudo-data (*INFN e Univ. Napoli*), exploding Naples facility (*see next slide*). Approach already adopting *Dask, Jupyter and ROOT RDataFrame*. talk WP5 (13 Oct).

LATEST

LHCb:

- Survey still ongoing (planning to contact more institutions). If someone is interested to the UC activities, please contact us!

KPI - Key Performance Indicator

These parameters, which are measurable objects, will reflect the final success of the entire use case.

KPI ID	Description	Acceptance threshold	Current Status
KPI2.2.2.1	Implementation of *N* data analyses in the AF	$N \geq 2$	> 50%
KPI2.2.2.2	Reference documentation of the AF	≥ 1 dedicated web site	0%
KPI2.2.2.3	Hands-on workshops for AF users	≥ 1 workshops	0%
KPI2.2.2.4	Scaling up the testbed AF infrastructure, serving *k* tenants, for a total of *N* data analyses	≥ (200·*N*) cores	0%
KPI2.2.2.5	Talks at conferences/workshops about AF activities	≥ 1 talk	100%*

* Talk at CCR in Loano (22 May).
For other conference opportunities, here the WP2 Tracking document. Please contact us for abstract submissions, to coordinate a "flagship collaboration"!

Computing resources

What we have now (not using ICSC resources):

- CMS-AF testbed: CMS Tier-2 of Legnaro (*non-production*): 3 nodes, each with 32 logic CPUs - 128GB of RAM - 1Gb/s network. For testing purposes: CMS Tier-2 of Legnaro (*production*): 1 node, 96 logic CPUs - 128 GB of RAM - 10Gb/s network.

- AF@Naples testbed: 2 virtual machines, one controller and one worker, each one with 12 logic CPUs and 64GB of RAM. Rocky Linux 8.6 OS.

WP5 is working to handle all the expertise and resources, converging in a single cloud setup, which will be ***experiment agnostic***.

- This effort is starting with some cloud resources already available. The setup is in progress, the configuration code will be made available in the ICSC-Spoke2 organization.

- Everything will be handled according to RAC rules and guidelines.

WP5 synergies

- Deployment of an Analysis Facility (INFN e Univ. Napoli);
  - Recent progresses made by the involved group: started testing first analysis applications and transition to nationwide cloud resources is undergoing. Talk WP5 (13 Oct).

LATEST

- New scholarship in *INFN Perugia*, operational on WP2 and WP5. Among the developers of the INFN-CMS Analysis Facility, currently in touch with Naples group for the first setup. Talk WP5 (13 Oct).

LATEST

- From *UniCal*, new manpower hired on Spoke 2 recently. Their involvement is currently under internal investigation:
  - Newly hired (1 Nov) with competences on infrastructure deployment, with ReCaS (Cosenza) resources.

Thank you

cn1-spoke2-wp2-analysisfacility@lists.infn.it (click here to subscribe or get in touch in case of issues)

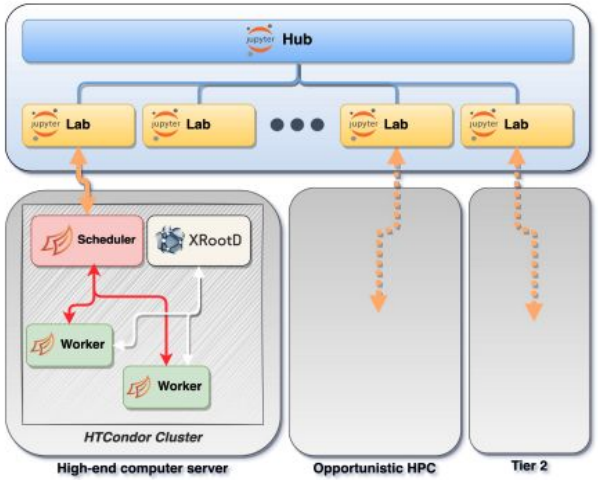
Participating Institutions

- **PIs:** Tommaso Diotalevi (Università di Bologna), Francesco G. Gravili (Università del Salento)
- **INFN** (including all sections)
- **Università del Salento**:
  - Search for new phenomena in events with two opposite-charge leptons, jets and missing transverse momentum, using LHC Run2 data. Porting of the analysis in the AF framework.
- **Università di Bologna:**
  - Search of a heavy neutral lepton (bump hunt) in the D_S meson decays with Run2 data. Porting of the analysis in the AF framework;
  - CMS Muon detector performance analysis: target quasi-interactive performance studies of phase space corners using large datasets; driven by the need to accurately assess analysis systematics (e.g. high-energy muons).
- **Politecnico di Bari:**
  - Testbed implementation of the AF in the ReCaS computing center
- **Università degli Studi di Bari:**
  - Charm and multi-Charm baryon measurements.
- **Università degli Studi di Napoli:**
  - Top quark + MET search. Full analysis implementation with Run3 data.
  - Analysis Facility deployment on local cloud infrastructure.
- **Università degli Studi di Firenze:**
  - Developing common analysis tools based on analysis facilities for the CMS experiment.
- **Università degli Studi Ferrara:**
  - Adopting tools for quasi interactive analysis for the IDEA experiment at the FCC-ee.
- **Università della Calabria.**

Please, check the correctness of the contact names in the document!

Analysis Facility testbeds

State of the art:



INFN-CMS *Analysis Facility*:

- Access to a single HUB and authentication via token (INDIGO-IAM)
- Based on standard industry technologies
- Customisable python kernel
- Workarea fully containarisable
- Overlay based on HTCondor (also available standalone)
- DASK library (python) for distributed computing
  - Scale the execution from 1 to N cores
- Possible implementation on heterogeneous resources (HTC/HPC/Cloud)
- Data access configurable with WLCG (xrootd, WebDAV, ...)

Similar approach for the INFN@Naples Analysis Facility:

- Jupyter Hub/Lab for interactive environment, supporting multiple accesses
- DASK library for scalability
- Kubernetes for container management and connection
  - Ongoing tests on this container technology to achieve a higher scalability