evolve

Leading the Big Data Revolution

# D2.2: Description and access to converged hardware platform

This project has received funding from the European Union's Horizon H2020 research and innovation programme under grant agreement No 825061

DDN Storage, BULL, IBM, FORTH, OnApp, Institute of Communications and Computer Systems, MemoScale, webLyzard technology, LOBA, Thales Alenia Space, Space Hellas, CybeleTech, Neurocom Luxembourg, MemEX, Tiemme SPA, Virtual Vehicle, AVL List GmBH, BMW AG, KOOLA

# Deliverable D2.2

## Description and access to converged hardware platform

## *Contributors*

| Name |
| --- |
| Huy Nam Nguyen |
| |
| |
| |

## *Peer reviews*

| Name |
| --- |
| |
| |
| |
| |

## *Revision history*

| version | date | reviewer |
| --- | --- | --- |
| V1 | 25/11/2019 | Huy Nam Nguyen |

# Table of Contents

# 1 Executive Summary

This document represents the deliverable D2.2 *Description and access to converged hardware platform*. It presents the main features of the Nova-S5 HPC platform developed by Bull, enhanced with storage access acceleration technology provided by DDN and made available to the partners of Evolve project.

# 2 Platform Hardware

The NOVA S5 platform (Cf. Figure 1) has been built from the following major components:

- Standard HPC Bullx compute nodes w/wo Hw accelerators;
- IO nodes;
- Storage system enhanced with DDN burst buffers;
- Interconnect Fabric.

The platform breakdown into processing nodes as follows:

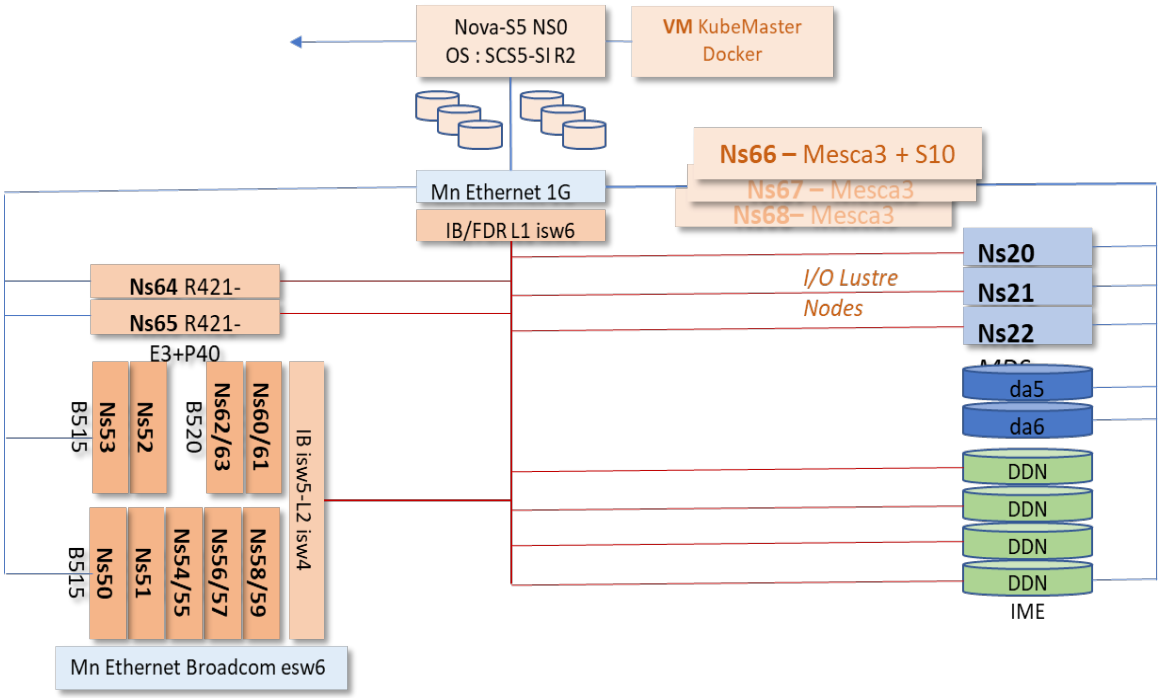


*Figure 1 : NOVA-S5 Hardware Architecture*

## 2.1 Processing nodes

### 2.1.1 Management Node

Bullx R423-E3 (ns0) Cluster management node and login node is an octa-core, bi-socket server, equipped with:

- 2 Intel® Xeon® octa-core E5-2665 (SandyBridge) processors (2.4 GHz, 20MB L3 cache, 1600 MHz)

- 64 GB DDR3-1600 ECC SDRAM
- 2 500 GB SATA3 Hard Disk Drive (7.2 Krpm)
- 8 X 8TB SATA3 SHD Hard Disk
- ConnectX-2 IB Mellanox card single port 4x FDR PCIe gen2-x8
- LightPulse PCI-express Fibre Channel Host Bus Adapter 8Gb/s dual channel

### 2.1.2 Compute Nodes

2 R423-E3 (ns20, ns21) octa-core, - I/O, Lustre OST node- , equipped with:
- Intel® Xeon® octa-core E5-2665 Processors (Sandy Bridge) (2.4 GHz) with 12 MB
- L2 and operating at 1333 MHz
- 64 GB DDR3-1600 ECC SDRAM
- 500 GB SATA3 Hard Disk Drive (7.2 Krpm)
- ConnectX-2 IB Mellanox card single port 4x FDR PCIe gen2-x8
- LightPulse PCI-express Fibre Channel Host Bus Adapter 8Gb/s dual channel

One Bullx R423-E4i (ns22), I/O, Lustre MDS node equipped with:
- 2 Intel® Xeon® octa-core E5-2667v3 8c (3.2GHz-9.6GT-20M-135W).
- 8GB DDR4-2133 ECC SDRAM (1x8GB)DR 1,2V
- 2disks 64GB SATADOM
- 960GB 2.5" SATA3 SSD Server (8 disks)
- ConnectX-3 card dual ports FDR 56GB PCIe gen3-x8

One blade system including (cmc0 holding the compute nodes): with 2 NVIDIA KEPLER & 2 INTEL XEON PHI blades including:
- 2 bullx 515 Equipped with Nvidia Kepler K20X GPUs (ns50-51)
- 2 Intel Xeon E5-2470 8c (SandyBridge) operating at 2.3GHz.
- 12 X 16GB DDR3-1600bullx ECC SDRAM
- 2 X 256GB 2.5” SATA3 flash SSD

2 Bullx 515 Equipped with Intel Xeon Phi (ns52-53):
- 2 Intel Xeon E5-2470 8c (SandyBridge)operating at 2.3GHz.
- 12 X 16GB DDR3-1600 ECC SDRAM
- 2 X 256GB 2.5” SATA3 flash SSD

Bullx B520 double compute blades (ns54-63), each equipped with:
- 2 Intel Xeon E5-2690 v3 12c (Haswell) at 2.60GHz
- 16 X 16GB DDR4-RDIM 2133 DR
- 2 X 256GB 2.5” SATA3 flash SSD

Two R421-E3servers (ns64, ns65) including
- Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz 8c
- 2 disques 935GB in Raid1
- 1 Mellanox Technologies card MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR / 10GigE]

- 2 GPUs NVIDIA with the following characteristics:
  The GPU integration starts with the current Nvidia/P40 with the following characteristics
  - CUDA Cores : 3,840
  - Memory Size : 24 GB GDDR5
  - Virtual Deskstops : 24
  - Interface : PCIe 3.0

  which will be upgraded soon to the following Nvidia/V100:
  - Nvidia Tensor Cores : 640
  - Nvidia CUDA Cores : 5120
  - Memory : 16GB HBM2
  - Interface : PCIe Gen3

One (to be extended soon to 3) Mesca3 (NS66)

- Intel 8-cores ® SKL CPU operating at 3.5 GHz
- Bittware 520N FPGA board with the following features
  - Intel Stratix 10 GX 2800
  - 4x QSFP28s for 400Gbps
  - 32 GB DDR4
  - BittWare-optimized OpenCL BSP

As the introduction of acceleration technologies represent an innovative feature of the project, we illustrate it more in detail in the following Figure 2.

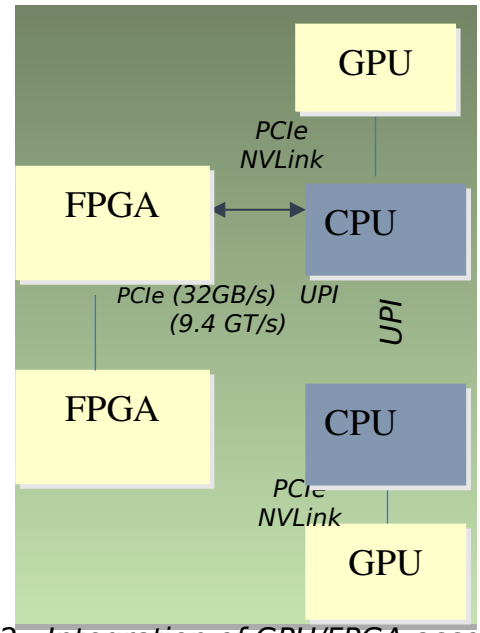


*Figure 2 : Integration of GPU/FPGA accelerators*

## 2.2 Interconnection Network

The network connects the nodes so they can communicate to share data, the state of the solution to the problem, and possibly the instructions that need to be executed. High performance systems, typically include a network dedicated to the computing aspect, along with other networks dedicated to management services, control operations, data storage, and I/O. Factors that come into play in the overall performance include the interconnection topology and communication protocole. The main interconnect network is built upon the following devices:

- InfiniBand FDR (56Gb/s) Interconnect (ISR 9024D and ISR 9024D-M)
- Cisco catalyst 3560G 44 ports

## 2.3 Storage

While the local storage in each node can be as simple as an SDD device to hold the OS, the application and the data, clusters of storage devices should be made available to the whole system for the purpose of checkpoint, archiving, etc. In order to separate the performance and capacity aspects, the insertion of flash storage, e.g. DDN/IME burst buffer, that streamline the application I/O and perform data cache, represents a plus. The storage devices include:

- 2 NetApp 2700
- 60 disks (1.8Gb each): 98T 24 disks (1.8Gb each): 47T

In comparison against the preliminary platform described in the previous report D2.1.V3, it is worth to note an imporvmenet of the burst buffer with the add-on of 4 IME240 servers with 24 NVMe software IME1.4. This evolution provides a faster tiering increased capacity whil providing support for container technology.

# 3 Platform Software

Compared to previous versions, the Nova-S5 software is built upon the Bull SuperComputer suite Version 5 (Cf. Figure 3). The Bull SCS 5 is a scalable, open, and robust software suite that meets the requirements of even the most challenging high performance computing (HPC) environments, which also require enhanced security. SC-S5 provides a complete software stack for the HPC, from small systems up to the largest ones and it targets the actual generation (Petascale) and the next one (Exascale).

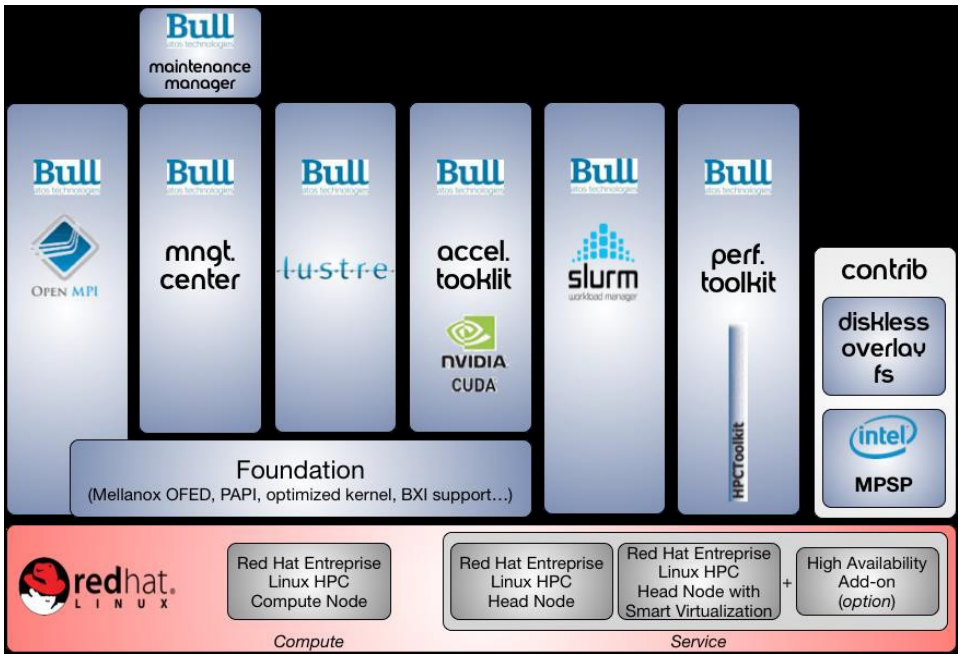


*Figure 3 : Bull SCS5 Software Architecture*

## 3.1 Operating System

Bull supercomputer suite 5 runs on Red Hat Enterprise Linux 7 which has proved its efficiency in HPC environments for years with its previous versions.
Atos and Red Hat technical experts have been working closely together for years, to make RHEL the ideal software environment for Bull hardware platforms.
Atos – Red Hat customers have access to professional-class worldwide support services provided by high level specialists who have a long experience of deploying large scale supercomputers.

## 3.2 Bull Foundation

This module includes MOFED, PAPI, advanced IPMI tools and specific modules related to Bull advanced products:
MOFED is the InfiniBand fabric management (OFED) stack from Mellanox.
PAPI (core) is enhanced by Atos to support the latest CPUs technologies in the period between CPU introduction and general support by the operating system.
IPMI tools are delivered with enhanced management functionalities, power management and inventory.

## 3.3 Bull Management Center

Bull Management Center is the administrative component of SCS 5 and integrates all tools needed to install, configure and manage a supercomputer. Depending on system size, the management will be done by:

- a management unit for supercomputers composed of up to 1,000 elements,
- a master management unit coupled with distributed management units for groups of equipment, for supercomputers with more than 1,000 elements.

The management infrastructure is designed to be scalable with a distributed and hierarchical environment.
Diskless operating system is available to ease deployment and enhance configuration.
The high availability functionality is introduced for management nodes, thanks to HA support in RHEL add-on.
*Security is greatly improved thanks to SELinux that is activated for supercomputer management and under specific conditions for compute nodes.*

## 3.4 Bull OpenMPI

The Bull Open MPI is based on open source MPI stack Open MPI 2.x, which is a standard-compliant library for message passing and hybrid programming.
Bull Open MPI provides key functionalities such as:

- run time scalability improvement with PMIx support (PMI Exascale),
- integration and support of Mellanox MXM and FCA accelerators,
- MPI 3.1 standard conformance.
- support of THREAD_MULTIPLE and Fortran 64 bit integer,

- integration of Portals 4 BTL and MTL for Bull eXascale Interconnect.

## 3.5 Bull Slurm

This batch manager is based on Slurm, the open source resource manager. Major enhancements in version 15.08, to which Bull is a major contributor, include:

- a hierarchical implementation based on hardware topology using the interconnect network for all communications to improve security and availability;
- support of Kerberos authentication through *AUKS* module;
- power adaptive scheduling for applications to enhance power capping by managing unused nodes and reducing CPU frequency:
- energetic fairshare scheduling based on energy consumption accounting;
- hyperthreading support to extend actual placement (socket and core) to hyperthread level.

Heterogeneous resources management and MPMD (Multi Process Multi Data) will be the next step of development for future versions.

## 3.6 Bull Lustre

This parallel file system is based on the Intel® Enterprise Edition for Lustre (IEEL) core, providing high performance and large storage solutions for big data workloads.
Extra functionalities were added by Atos for Lustre 2.7:

- Integration of Lustre client and router with MOFED® stack,
- Shine centralized administration tool,
- Monitoring with Shinken and Graphite,
- High Availability integration based on pacemaker.

## 3.7 Bull Performance Toolkit

The Bull Performance toolkit include Bull products such as HPC Toolkit (with Bull extensions), PAPI, and third-party products among which Intel® Parallel Studio XE software:

- HPCToolkit features Bull extensions that make it possible to detect processes with various behaviors and to compare successive runs.
- PAPI provides an open source API that gives access to the hardware performance counters available in modern processors, including latest generation thanks to Bull Foundation.
- A complementary offer for the development environment can be purchased separately: the Intel® Parallel Studio XE development environment software suite.

## 3.8 Computing Acceleration

This module includes specific packaged drivers (e.g. Nvidia/CUDA, Altera/Quartus, etc) and additional software (e.g. OpenCL) respectively for accelerators such as Nvidia/GPUs or Nvidia/FPGAs.

## 3.9 Bull Maintenance Manager

The Maintenance Manager provides the specific Bull tool Argos for the maintenance of your system, to keep it up-to-date and alive with fine tracking of maintenance operations.

# 4 Opening to BigData and Cloud

The Evolve platform allows end user not only to run traditional HPC codes as detailed in previous sections but also to use containers and interact with the cloud. Indeed, moving HPC solutions to the cloud while performing convergence with Big Data requires above all the introduction of new software layers including VM and Cloud services containers (Dockers and Kubernetes) together with middleware management to support Big Data applications. More details of such add-on features can be found in other deliverables (e.g. D2.3, D7.1, D7.3, etc.).

## 4.1 Working around root access

Container is a technology coming from the Cloud where root access is not as strictly controlled than in HPC. Therefore most if not all Docker command implies a root access. In order to work around this constraints and still allow end-user to run Docker command without compromising root access, the platform support a Docker groups. All Evolve partners are added to the docker group. Only the command prefix by # need to be run as root.

*= Create docker groups*
*# sudo groupadd* ***docker***
*= Adding user to the Docker group*
*#sudo gpasswd -a my_user* ***docker***
*Adding user my_user to group docker:*
*= Refresh groups belonging to take modification into account*
*% groups*
*my_user adm cdrom sudo dip plugdev lpadmin sambashare*
*%newgrp docker*
*%groups*
***docker*** *adm cdrom sudo dip plugdev lpadmin sambashare my_user*

## 4.2 Fetching Docker image

In the container world the code is not built locally on the platform, it is considered as portable hence built on a remote platform and fetch for the execution phase. In such a case, a remote access from the node is mandatory since the container will be refreshed just prior to its execution. In order to implement this workflow application partners need to set-up a docker repository. The repository is accessed dynamically to fetch the latest version of the image. With respect to classic cluster configuration, it means that the executing node need to have access to the internet. The example bellow is taken from the CybeleTech application:

*% docker login -u teddy.debroutelle registry.cybeletech.fr*
*Password:*

*WARNING! Your password will be stored unencrypted in /home/user/.docker/config.json.*
*Configure a credential helper to remove this warning. See*
*https://docs.docker.com/engine/reference/commandline/login/#credentials-store*
*Login Succeeded*
*% docker pull registry.cybeletech.fr/cybeletech/image/image:dev*
*dev: Pulling from cybeletech/image/image*
*Digest:*
*sha256:51d8ce7862eaf74f9aab4a3209de45a401b5553046a986521254117f8dbc85fb*
*Status: Image is up to date for registry.cybeletech.fr/cybeletech/image/image:dev*

Once the code has been executed, it is considered as good practice to remove the configuration file created during the login stage.

*% docker logout registry.cybeletech.fr*
*Removing login credentials for registry.cybeletech.fr*
*% rm /home/user/.docker/config.json*

# 5 Access to Nova-S5

Access to the Nova-S5 cluster is dedicated to users having an authorization to connect to the cluster via Internet (the list of Evolve users is provided in the annex). A user group is created for each partner, and each user is associated with a user group.
On Nova-S5, jobs and resources are managed by Slurm that enables to request resources, to submit jobs, to query their status, and more generally to make an effective use of the cluster resources.

## • Access to Nova-S5

Direct SSH connections need to be known by the Atos/Bull and AGARIK firewall. Thus, Bova-S5 users must declare their IP address to ATOS/Bull (by filling the "Request to get access to Nova" form and send it to the Project Manager and the cluster Administrators) before being able to connect.
Conditions of exploitation:

- **No backup of user data**

Due to the large amounts of data sometimes required for user applications, the Platform Administration Service does not provide any backup of data, user applications and additional software that may be installed in user spaces and shared spaces.

- **Storing personal data allowing the identification of persons is forbidden**

It is strictly forbidden to store or deposit on the platform, even temporarily, data allowing the identification of persons. The platform is not designed with a sufficient level of security to host and protect this type of data.

## • How to connect to Nova-S5

The connections are made via the Secure SHell (SSH) protocol. The cluster address enables to connect to the Login node, ns0 (the entry point for the cluster).
The commands to connect to ns0 are:
*ssh login_name@92.43.249.197*
*or*

*ssh 92.43.249.197  -l login_name*

Access to compute nodes is done via submission of  Slurm requests from the login node ns0. Other nodes of the cluster can be accessed through SSH without giving the password:
*ssh ns[x]  (x represents the chosen node number)*

When connecting from a Linux client, file transfers can be made via the scp command, for example to make a transfer from the local machine to the entry point for the cluster, the command to be used is:

*scp example.tgz login-name@92.43.249.197:~*

When connecting from a MS-Windows client, file transfers can be made using WINscp or filezilla.

## • Resource Request and Job Submission

Resource reservation is handled by slurm  (on the Login node : ns0)
For specific needs (e.g. long jobs requiring a large set of compute nodes) a request must be sent to the administrators. Should a conflict arise, the Administrators will negotiate with conflicting requesters and make the decision.

## • Job session scheme

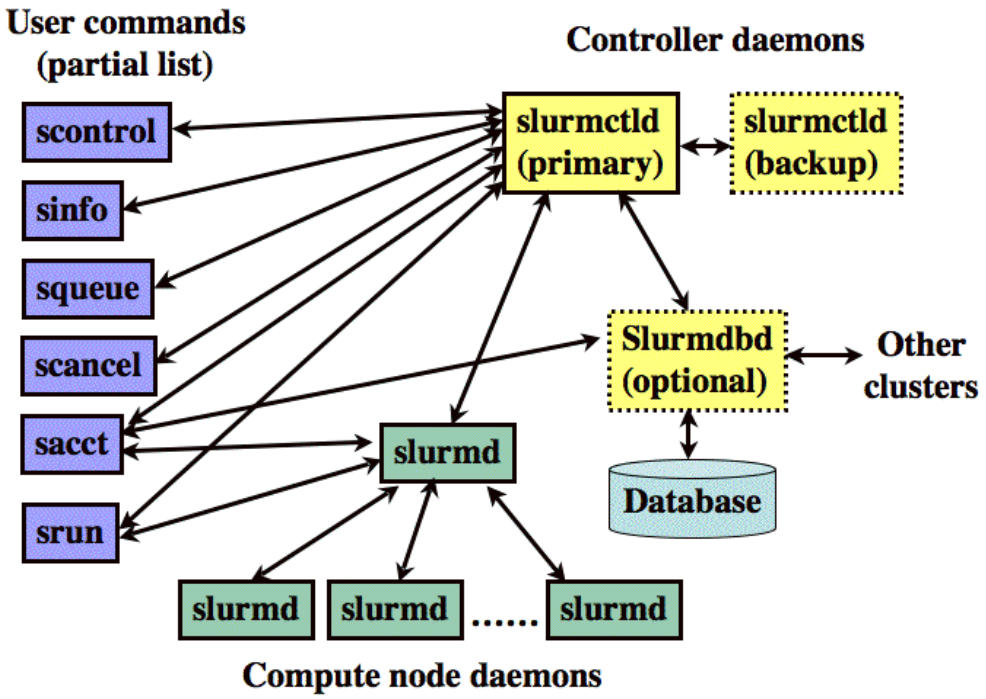


*Figure 1: Job session scheme*

- ## Job submission

srun [OPTIONS...]  executable [args...]
Slurm returns a job identifier (a job number)

There are two types of Slurm Jobs:

- Batch Job: A script that contains commands or tasks to execute site specific applications
- Interactive Job: Considered like a batch job but, when eligible to run, the user's terminal input and output are connected to the execution similar to a login session. Useful for users needing to work interactively, for instance to debug their job script.

Jobs typically pass through several states (PENDING, RUNNING, SUSPENDED, COMPLETING, and COMPLETED) in the course of  their  execution.

- ## How to request nodes and CPUs (cores)

On the Nova-S5 cluster, it is not possible to connect directly by ssh on a computing node and then start a job : It is necessary to make requests from the Slurm management node ns0 for the nodes and CPUs using the **salloc** or **srun** commands.

Examples of exploitation:

To request 2 nodes with 8 CPUs each in an **interactive** way
*[londaitl@ns0 ~]$* ***salloc -N2 -n2 -c8***
*salloc: Granted job allocation 115*
*[londaitl@ns0 ~]$*
*or :*
*[londaitl@ns0 ~]$ salloc --nodes=2 --ntasks=2 --cpus-per-task=8*
*salloc: Granted job allocation 116*
*[londaitl@ns0 ~]$*
*To view job status of slurm allocation*
*[londaitl@ns0 ~]$ sinfo*
*PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST*
*test*  up   infinite  2  alloc ns[60-61]*

# 6 Conclusions

The Evolve platform described in this document represents an HPC solution dedicated to support the execution of the project use cases. its particularity lies in the extension of its features to the bigdata as well as the cloud together with the integration of accelerated computing technologies such as GPU and FPGA. The implementation of these features represents an important amount of works of WP2 in cooperation with other WPs of the project. The platform is now up and starting to host the many applications of the project.

# 7 Annex : Users of the Evolve Platform

| Access Address | Institution | User Identification | Mail Address |
|---|---|---|---|
| 92.43.249.197 | ATOS | NGUYEN Huy-Nam | huy-nam.nguyen@atos.net |
| 92.43.249.197 | ATOS | fekhr-eddine.keddous | fekhr-eddine.keddous@atos.net |
| 92.43.249.197 | ONAPP | Michail Flouris | michail.flouris@onapp.com |
| 92.43.249.197 | ONAPP | Stelios Louloudakis | stelios.louloudakis@onapp.com |
| 92.43.249.197 | MICROLAB (ICCS) | Ioannis Oroutzoglou | ioroutzoglou@gmail.com |
| 92.43.249.197 | MICROLAB (ICCS) | Dimosthenis Masouros | demo.masouros@microlab.ntua.gr |
| 92.43.249.197 | MICROLAB (ICCS) | Konstantina Koliogeorgi | konstantina@microlab.ntua.gr |
| 92.43.249.197 | MICROLAB (ICCS) | Sotirios Xydis | sxydis@microlab.ntua.gr |
| 92.43.249.197 | CYBELETECH | Teddy Debroutelle | teddy.debroutelle@cybeletech.com |
| 92.43.249.197 | CYBELETECH | Glennie Vignarajah | glennie.vignarajah@cybeletech.com |
| 92.43.249.197 | FORTH | Antonis Papaionnou | papaioan@ics.forth.gr |
| 92.43.249.197 | FORTH | Christos Kozanitis | kozanitis@ics.forth.gr |
| 92.43.249.197 | FORTH | Manos Pavlidakis | manospavl@ics.forth.gr |
| 92.43.249.197 | BMW | Stefan Feit | Stefan.feit@bmw.de |
| 92.43.249.197 | IBM | Christian Pinto | christian.pinto@ibm.com |
| 92.43.249.197 | DDN | Jean-Thomas Acquaviva | jtacquaviva@ddn.com |
| 92.43.249.197 | DDN | Konstantinos Chasapis | kchasapis@ddn.com |
| 92.43.249.197 | MEMOSCALE | Kjetil Babington | Kjetil.babington@memoscale.com |

| | | | |
|---|---|---|---|
| 92.43.249.197 | PIXYL | Alan Tucholka | alan@pixyl.ai |
| 92.43.249.197 | ICHEC | Paddy Ó Conbhuí | padraig.oconbhui@ichec.ie |
| 92.43.249.197 | NEUROCOM | Clément Rey | c.rey@neurocom.lu |
| 92.43.249.197 | SPACE | George Vamvakas | gvamvakas@space.gr |
| 92.43.249.197 | THALES Alenia Space | Michelle Aubrun | Michelle.aubrun@thalesaleniaspace.com |
| 92.43.249.197 | THALES Alenia Space | Andres Troya Galvis | Andres.troya-galvis@thalesaleniaspace.com |
| 92.43.249.197 | MemEX Italie | Claudio Disperati | claudio.disperati@memexitaly.it |
| 92.43.249.197 | A2C2 | Alexander Stocker | alexander.stocker@v2c2.at |
| 92.43.249.197 | A2C2 | Andreas Festl | andreas.festl@v2c2.at |
| 92.43.249.197 | A2C2 | Christian Kaiser | Christian.Kaiser@v2c2.at |