Portuguese **Distributed Computing** Infrastructure

Lisbon node deployed and ready (Mi6.1)

Execution Report

State:	**FINAL**
Dissemination:	**PUBLIC**
Authors:	**Jorge Gomes**
URL:	**http://www.incd.pt**

Date	Author	Comments
20-05-2018	Jorge Gomes	Initial version
10-07-2018	Jorge Gomes	Technical deployment information
05-02-2019	Jorge Gomes	Added appendix II with new hardware

Cofinanciado por:

COMPETE 2020 · PORTUGAL 2020 · UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional · Lisb@20²⁰ · FCT Fundação para a Ciência e a Tecnologia

Executive Summary

INCD aims to support a wide range of scientific domains and projects with very different requirements. Supporting them across all stages of simulation, processing, analysis and data sharing requires a flexible infrastructure on top of which new added value services and specific virtual research environments can be implemented and delivered.

The activity 6 (deployment) includes: the deployment of the services designed in the activity 5 (development), as well as the installation of equipment to be purchased and/or committed by the INCD users.

During the first year a set of main services were identified and made available to the pilot users aiming at gathering experience and collecting feedback for further improvements. This milestone document provides an overview of the INCD main services made available for evaluation by the pilots during the first project year. The three main services made available are:

- High Throughput Computing: provided by the INCD batch farm
- High Performance Computing: provided by the INCD batch farm
- Cloud Computing: provided by the INCD OpenStack IaaS cloud whose architecture is described in the milestone Mi5.1.

The INCD batch service

The classic batch systems have been the preferred choice to manage computing farms. They excel at exploiting the aggregated capacity of multiple compute nodes. Batch systems are suited to execute applications whose execution time is well defined and relatively short, in opposition to services where execution is continuous. Applications are submitted to the batch system as jobs that are queued and whose sequence of execution is managed by the batch system according to scheduling policies defined by the system administrator. The batch systems enable different levels of quality of service to be specified and enforced. Furthermore the queuing mechanism helps ensuring a continuous flow of jobs that maximises the exploitation of the existing processing capacity by keeping compute nodes constantly occupied. This makes batch processing a highly efficient mechanism to efficiently share computing capacity. The batch system keeps track of existing resources and their availability so that the queued jobs can be scheduled to compute nodes according to the defined policies. Batch systems constitute a well-proven approach to manage and share computing capacity. Existing batch systems can reliably scale to thousands of nodes, and still constitute the hearth of modern supercomputers. Batch processing is very common in research environments where it addresses many of the scientific computing requirements. However scientific research is becoming increasingly complex and collaborative and new requirements related to communication, collaboration and sharing of information have emerged. Therefore nowadays batch systems are only one of the several required tools. They must be combined and or complemented by other solutions to provide increased flexibility and effectively address the user needs.

INCD choose to offer two batch services oriented to support High Throughput Computing and High Performance Computing applications. The requirements and technical aspects of supporting these two computing approaches are similar. High Performance Computing is characterized by tightly coupled parallel jobs that require low latency communication between processes, while High Throughput Computing is characterized by loosely coupled applications whose processes usually do not communicate across different hosts. Both may require a high bandwidth network for data access. However High Performance Computing often requires low latency network interconnects for fast communication between application instances running in different compute nodes. Besides this difference, both computing approaches can be served using the same type of batch system. Therefore INCD choose to use the same batch system scheduler to support both the High Throughput Computing and High Performance Computing services. This approach provides greater flexibility in managing the resources and makes possible the opportunistic use of idle High Performance Computing nodes by backfilling them with short High Throughput Computing jobs. Backfilling is achieved by using CPUs that would be otherwise idle while waiting for other CPUs to become available to start larger parallel jobs. The use of the same batch scheduler for both HPC and

HTC also reduces the operational costs and maximizes the availability of expertise needed to manage the system.

The batch system selected for the initial INCD setup is SoGE. This choice was based on the partner's extensive experience with this batch system.

The use of SoGE also minimizes the impact on some large user communities which were already using SoGE. This is the case of the Portuguese research groups that participate in the CERN Large Hadron Collider (LHC) experiments (ATLAS and CMS) whose national computing services including the national Tier-2 for simulation and analysis has been integrated in INCD. The Tier-2 for ATLAS and CMS delivers non-stop data processing services integrated in the Worldwide LHC Computing Grid and has complex dependencies on grid computing middleware which is typically batch oriented. The Tier-2 was already using SoGE in combination with a set of middleware services. Several other user communities inherit from the Portuguese National Grid Initiative and now piloting INCD services were also using SoGE and therefore keeping the same batch system facilitates continuity and avoids major disruptions. Meanwhile the migration to the Slurm batch system is also being strongly considered and evaluated. Slurm is a more recent development than SoGE but also has its own limitations and issues. The interest in Slurm, and the reason for an eventual change is mainly related with the level of support by the SoGE development community, better resource management control and the increased popularity of Slurm that may facilitate interoperability and support for future users. For the current development and piloting phase of INCD the current solution based on SoGE has been considered adequate.

INCD is also considering a deployment of HTCondor to aggregate computing capacity available in the several services and across datacenter locations. A global pool of HTC resources can be especially useful for the second and third years when additional capacity will be deployed in the north and center regions of the country.

Batch Service Login Nodes Service

The head nodes are interactive Linux nodes with configurations similar to the batch system compute nodes have been deployed to support the login of the end-users. Users can login via SSH in these nodes and use them to prepare the applications and interface with the batch system. INCD imposes the use of ssh keys during login for increased security. Individual home directories are provided for each user using a shared file system. In some cases dedicated head nodes are setup to address the particular needs of the users. The head nodes can also be used to transfer data between the INCD file systems and the user premises.

Batch Service High Throughput Computing (HTC) Service

The HTC service supports the execution of single threaded and multi-threaded jobs including OpenMP applications. The HTC jobs are submitted to batch queues managed by the SoGE batch system. The queue scheduling policy is mainly fair-share. The jobs can be executed in the set of compute nodes reserved for HTC applications or in the HPC machines as backfilling. Most compute nodes are physical machines but INCD is also exploiting the capacity of the INCD IaaS cloud computing service to provide virtual compute nodes that extend the base capacity of the HTC service, this will likely become the main source of compute capacity for the HTC service at a later project stage. This approach provides elasticity and flexibility both for the cloud and for the HTC service and allows capacity to be shift between both services.

Batch Service High Performance Computing (HPC) Service

The HPC service supports the execution of parallel processing applications. Multiple parallel environments have been configured in the batch system and made available to the end-users these include MVAPICH, OpenMPI, PVM and others. This service is meant to support execution of tightly coupled applications across multiple computing hosts. To achieve scalability and good performance the communication latency must be minimized. INCD addressed these requirements by scheduling these applications to physical hosts to minimize delays and overheads introduced by virtualization, and by having a fraction of the batch compute nodes interconnected via an Infiniband low latency network. A second set of physical compute nodes interconnected via a low latency Ethernet network is also available for parallel applications with less demanding latency requirements.

Furthermore the need to support multiple computing environments to facilitate the application support lead to the adoption and deployment of Linux container technologies that are now being used in the INCD batch system to support both HPC and HTC applications proving lightweight virtualization for the batch compute nodes without the overheads of full virtualization.

Batch Service Distributed Parallel Filesystem Service

The INCD Parallel FIlesystem Service is based on Lustre (lustre.org), an open source file system that provides POSIX semantics. Lustre supports large scale deployments composed of multiple data and metadata nodes and enables parallel I/O from applications. Lustre is widely used in HPC environments. The INCD deployment of Lustre is composed of several storage spaces dedicated to specific purposes and tuned for the corresponding workloads. These include storage spaces tuned for High Throughput Computing and High Performance Computing. Splitting the Lustre deployment into several storage spaces also contributes to contain the impact of maintenance interventions and technical issues. The Lustre architecture is mostly adapted to support access to large files. Data and metadata are kept separately and access to metadata is frequently limited by the access to the metadata servers. The INCD deployment includes multiple metadata servers to spread the metadata load and minimize bottlenecks.

Batch Service NFS Filesystem Service

INCD provides a Network File System (NFS) service which is used to support home directories and spaces for sharing software across groups of users. Although Lustre is more suitable for large scale storage, NFS behaves better with smaller files as it is generally the case with software.

Batch Service CVMFS Service

The CernVM File System provides a read-only, scalable, reliable and low-maintenance software distribution service that can be used locally or over Wide Area Networks. The system is especially suited to share compiled software and small files across the Internet. CFMFS provides a mountable read-only filesystem with aggressive caching. Files are transferred on-demand as they are accessed via a hierarchy of HTTP proxies enabling caching at several levels. INCD supports mounting of CVMFS areas in its infrastructure services, and also offers a stratum zero service through which files can be made available for access over the Internet.

Batch Service GPGPU Computing Service

INCD provides access to a set of compute nodes equipped with Nvidia GPUs. Although this is a small experimental service it is being used by a wide range of use cases. The growing interest in machine learning has significantly increased the demand on this service. The computing nodes equipped with GPUs can be made available as virtual machines via the IaaS cloud service or as batch system hosts both physical and virtual. The GPGPUs are mapped to virtual compute nodes via PCI pass-through. The possibility of implementing these nodes with LXD container based hypervisors in the IaaS cloud is being considered. The container technologies adopted by INCD for use in the batch system can also be used to execute GPGPU applications. Currently both CUDA and OpenCL are supported. INCD aims to expand this service during the second and third years.

Batch Service Linux Containers Service

Containers became widely used as a means to encapsulate both services and also user applications, providing isolation from the underlying execution hosts environment. INCD supports several container technologies. Containers are supported for the execution of HTC and HPC applications using udocker (https://github.com/indigo-dc/udocker). This tool enables end-users to execute Linux containers without the need of complex kernel features or privileges. INCD is actively contributing to the development of udocker that is freely available in github. Udocker encapsulates several execution methods empowering users to execute applications encapsulated in Docker containers. In addition INCD also supports Singularity when required by its use cases although this solution raises security concerns given the use of privileges and vulnerabilities identified in the past.

Batch Service Grid Computing Service

The INCD batch system in Lisbon is fully integrated into the European Grid Infrastructure (EGI) and into the Iberian Grid Infrastructure (IBERGRID). INCD supports both access of Portuguese researchers to the EGI and IBERGRID services and enables the sharing of national computing resources in the context of international collaboration projects. In technical terms INCD operates several middleware services which enable remote job submission, remote data storage and data access. To this end the INCD batch system has been integrated in the EGI and IBERGRID infrastructures by means of CREAM Computing Elements that enable submission of jobs to the INCD batch system via a RESTful interface. The data storage integration has been accomplished by using the StoRM middleware that allows the storage and transfer of files between the local Lustre filesystem and other grid Storage Elements. INCD is also providing a file catalogue service for data management and authentication and authorization services to register users in Iberian grid user communities under IBERGRID.

Batch Service Software and Application Environments

INCD supports a wide range of open source tools and application environments. Based on use case requirements the INCD support team has been compiling and maintained a growing set of libraries and tools which are made available to the users using the modules software. The software is then made available via the INCD CVMFS service.

Batch Service Physical resources

The INCD Lisbon services are housed at the sala-grid datacenter. The datacenter is operated by FCT-FCCN while the INCD services are developed, deployed and maintained by LIP. The center houses a mixed set of AMD and Intel processors with 3GB to 4GB of memory per CPU core. The Lustre storage is composed of multiple storage nodes connected to disk arrays complemented by multiple metadata nodes to spread the load related to metadata access.

The INCD Cloud system

Batch systems have limitations in what concerns the execution of long-running services that often need to be exposed to external networks (such as scientific portals, databases, web services, among many others), furthermore new types of complex applications need tailored environments often requiring heterogeneous setups that are difficult to support in homogeneous cluster environments.

INCD aims to support a wide range of scientific domains and projects with very different requirements. Supporting them across all stages of simulation, processing, analysis and data sharing, requires a flexible infrastructure on top of which new added value services and specific virtual research environments can be implemented and delivered. These services can be specific to user groups and/or communities, or generic aimed to the wider scientific community.

The INCD cloud service is based on the OpenStack cloud management framework and its architecture is defined in Mi5.1 (First implementation of base cloud services). The deployment aspects are here summarized.

OpenStack IaaS service for conventional virtual machines

This service enables the management of the life cycle of virtual machines. The INCD IaaS cloud uses the KVM virtualization hypervisor. This enables virtual machines to be instantiated from pre-defined or user uploaded images using a multitude of image formats supported by OpenStack. Virtual machines can be paused, restarted, rebooted, rebuilt, deleted, resized and shelved. Remote access to the system consoles of the virtual machines is also supported which facilitates the management and recovery.

OpenStack IaaS service for block storage

This service enables the allocation of block storage to support the reliable execution of virtual machines with live migration. In addition this service enables further block devices to be added to virtual machines thus extending their storage capacity. The block storage is provided by a Ceph storage system and is made available and managed through the OpenStack Cinder component. The Block storage can also be used to perform snapshots of existing virtual machines and/or volumes.

OpenStack IaaS service for cloud networking

This service enables mainly the creation of private overlay networks to interconnect virtual machines running in the IaaS service. It supports GRE, VXLAN and datacenter provided VLANs. The service also supports SRIOV therefore mapping virtual instances of network adapters into virtual machines for bare-metal network performance. The INCD cloud networking service also enables mapping of public IP addresses into the virtual machine IP addresses and the allocation of VLANs having public IP addresses to support applications with special connectivity requirements.

OpenStack IaaS service and Advanced Computing

The INCD cloud infrastructure is designed to support high performance computing and demanding data access requirements. The INCD cloud service offers support for: GPGPUs, bare-metal performance with containers and bare-metal performance for network intensive applications.

GPGPUs

The virtualization service includes support for the allocation of GPGPUs to virtual machines using PCI pass-through. This setup allows each GPGPU PCI device to be made directly visible inside a virtual machine. Cloud compute nodes with NVIDIA GPUs are available and virtual machines can be instantiated with GPGPUs.

Bare-metal performance

Bare metal like performance is provided by combining Linux containers technology with OpenStack. The INCD cloud architecture comprises the use of LXD as bare metal hypervisors. Compared with other container technologies LXD provides virtual machine like capabilities using Linux containers, LXD is thus suitable to run almost complete operating systems. For integration with OpenStack the nova-LXD driver is available. The benefit of using containers is reduced memory footprint and less overhead for compute and I/O operations.

Network intensive applications

The INCD cloud service supports the following high performance network enhancements:

- Support for network interface cards with SR-IOV capabilities. With SR-IOV physical network interface cards (PF) can show additional virtual instances of themselves in the PCI bus (VF) that can be mapped to virtual machines. This approach allows the virtual machines to access the network interface cards directly providing much higher performance.
- Tuning of the network packet size that has been enlarged using jumbo frames to reduce the packet processing overheads.
- Support for VLAN based virtual networks to exploit the native performance of the datacenter networks and/or access other datacenter services.

Federation

The INCD cloud is federated and integrated with the EGI (European Grid Infrastructure) AAI service which currently can accept multiple identity providers including social networks, EDUGAIN, ELIXIR AAI, DARIAH AAI, IGTF X.509 certificates and ORCID. Besides the authentication and authorization aspects, the INCD cloud was integrated with the EGI accounting service and with the EGI images catalogue. In this context the INCD cloud is ready for the EOSC (European Open Science Cloud) where it participates as an EGI federated cloud provider capable of supporting international user communities.

Resiliency and load balancing

The INCD cloud architecture includes resiliency and load balancing at several levels.

1. The SQL database supporting the OpenStack components is provided by a cluster of three machines running MariaDB
2. The messaging system used by OpenStack was setup using a redundant RabbitMQ.
3. The API and controller nodes for the several OpenStack services are split across a cluster of three machines.
4. Load balancing and high availability both for MariaDB and for the OpenStack API and controller nodes is implemented by a fault-tolerant *haproxy* setup that includes VRRP provided by *keepalived*.
5. Multiple compute nodes are available and virtual machines stored in Ceph can be live migrated between the nodes, in case of compute node failure these machines can be easily restarted in another node.
6. Neutron network gateways are setup in a fault tolerant setup enabling redundant routing.

7. Storage gateways have been deployed in a scalable redundant setup enabling access to images and object storage.
8. The Ceph storage system has each data block replicated across three storage servers.
9. The Ceph monitor nodes are setup in a Paxos cluster of three nodes.

Appendix I – endpoints

INCD Batch Service 1st implementation

Location	Service	Endpoint
Lisbon	SRM (Grid Storage)	https://srm01.ncg.ingrid.pt:8444
Lisbon	GridFTP (grid ftp transport for SRM)	gridftp://gftp01.ncg.ingrid.pt gridftp://gftp01.ncg.ingrid.pt
Lisbon	Xrootd (remote file access)	xroot.ncg.ingrid.pt:1094 xroot-data01.ncg.ingrid.pt:1094 xroot-data02.ncg.ingrid.pt:1094 xroot-data03.ncg.ingrid.pt:1094
Lisbon	WebDAV (remote file access)	
Lisbon	Computing Element (Grid Computing)	https://ce04.ncg.ingrid.pt:8443/ce-cream/services https://ce05.ncg.ingrid.pt:8443/ce-cream/services https://ce06.ncg.ingrid.pt:8443/ce-cream/services
Lisbon	Site BDII (Grid Site Information System)	ldap://sbdii01.ncg.ingrid.pt:2170/GLUE2DomainID=NCG-INGRID-PT,o=glue
Lisbon	Top BDII (Grid Replica of Top Information System)	ldap://topbdii01.ncg.ingrid.pt:2170/GLUE2GroupID=grid,o=glue
Lisbon	MyProxy (X.509 credential storage)	myproxy://px01.ncg.ingrid.pt:7512/
Lisbon	VOMS (authN/AuthZ VO mgmt)	https://voms01.ncg.ingrid.pt:8443/voms
Lisbon	Argus (authN/AuthZ policies)	https://argus.ncg.ingrid.pt:8152/authz
Lisbon	LFC (file catalogue)	lfc://lfc01.ncg.ingrid.pt:5010

INCD Cloud Service 1st implementation

Location	Service	Endpoint
Lisbon	Horizon (web dashboard)	https://nimbus.ncg.ingrid.pt
Lisbon	Heat (orchestration)	https://nimbus.ncg.ingrid.pt:8000/v1
Lisbon	Cinder (block storage)	https://nimbus-stor.ncg.ingrid.pt:8776/v2
Lisbon	Nova (compute)	https://nimbus.ncg.ingrid.pt:8774/v2.1
Lisbon	Ceilometer (metering)	https://nimbus.ncg.ingrid.pt:8777
Lisbon	Neutron (networking)	https://nimbus.ncg.ingrid.pt:9696
Lisbon	Glance (vm images)	https://nimbus-stor.ncg.ingrid.pt:9292

Lisbon	Keystone (authN/authZ)	https://nimbus.ncg.ingrid.pt:35357/v3
Lisbon	Swift (object storage)	https://nimbus-stor.ncg.ingrid.pt:8080/swift/v1

Appendix II – hardware purchase for the Lisbon node

During the first project year, INCD has identified the pilot use cases that will drive the infrastructure development and the validation of the services. This work was performed in activity 4 (Pilots). This information was combined with the development activities taking place in the activity 5 (development) to define the services to be made available and drive the purchase of new equipment within activity 6 (Deployment).

The actual process for purchase of new equipment for the Lisbon node started in the last quarter of the first project year. The process was affected by the administrative and legal complexity of the international public tender, especially because INCD is a new legal entity and this was the first public tender ever performed by the INCD association. The list of equipment delivered includes:

- More than 25 compute nodes each with:
  - Two AMD Epyc 7501 processors having a total of 64 CPU cores
  - 384 GB of RAM per compute node
  - Two SSD with 480GB each 3DWPD
  - Four 2TB SATA3 disks
  - Dual 10GbE SFP+
  - Infiniband
- Eight compute nodes with:
  - Two AMD Epyc 7501 processors having a total of 64 CPU cores
  - 384 GB of RAM per compute node
  - Two SSD with 960GB each 3DWPD
  - Four 2TB SATA3 disks
  - Dual 10GbE SFP+
  - Infiniband Mellanox FDR
- Six storage nodes for Lustre with:
  - Two Intel Xeon Silver 4110 having a total of 16 CPU cores
  - 64 GB of RAM per storage node
  - Two SSD with 480GB each 3DWPD
  - Dual 25GbE SFP+
  - 24x 8TB SATA3 disks
- Six storage nodes for Ceph with:
  - Two Intel Xeon Silver 4110 having a total of 16 CPU cores
  - 192 GB of RAM per storage node

  - Two SSD with 480GB each 3DWPD
  - Dual 25GbE SFP+
  - 24x 8TB SATA3 disks
- Three servers for Lustre metadata with:
  - Two Intel Xeon Silver 4110 having a total of 16 CPU cores
  - 192 GB of RAM
  - Two SSD with 960GB each 3DWPD
  - Dual 10GbE SFP+
- Six servers for supporting services (I/O and network gateways) with:
  - Two Intel Xeon Silver 4110 having a total of 16 CPU cores
  - 32 GB of RAM
  - Two SSD with 960GB each 3DWPD
  - Dual 10GbE SFP+
- One spare storage chassis for Lustre or Ceph with:
  - Two Intel Xeon Silver 4110 having a total of 16 CPU cores
  - 32 GB of RAM per storage node
  - Two SSD with 480GB each 3DWPD
  - Dual 25GbE SFP+
  - 12x 8TB SATA3 disks
- Three Layer 3 Ethernet managed network switches with:
  - 48x 25GbE SFP+ ports
  - 18x 100GbE QSFP+ ports
- One Infiniband managed switch with:
  - 36x 56Gb/s ports

The hardware was installed and tested between December 2018 and January 2019. The equipment is now fully operational and integrated in the HPC, HTC, cloud and data services.