WNoDeS
Grid / Cloud Integration Framework

INFN
Istituto Nazionale di Fisica Nucleare

WORKER NODES ON DEMANDS SERVICE

REQUIREMENTS FOR VIRTUALIZED SERVICES

Document version:	**1.0.0**
Authors:	**D. Salomoni, E. Ronchieri (ed.)**
Date:	**September 14, 2011**

This work is funded by INFN.

CONTENTS

1 INTRODUCTION

The Worker Nodes on Demand Service, a.k.a. WNoDeS, is a software framework created by INFN to integrate Grid and Cloud provisioning. All resources (whether they are offered to users via Grid interfaces, Cloud interfaces, or else) are taken from a common pool.

The WNoDeS focus is on providing scalability and flexibility to both users and resource providers in a production environment. It is a solution to virtualize computing resources and to make them available through different interfaces, supporting user interactions with computers via traditional batch systems, Grid middleware, or open cloud standards, making it possible to allocate compute, storage and network resources on a pay-as-you-go basis.

User authentication is supported via several authentication methods, whilst authorization policies are handled via gLite Argus.

The WNoDeS framework has been deployed in production at several Italian centers, including the INFN Tier-1 (CNAF, Bologna) since November 2009, where it currently manages about 2000 on-demand VMs, corresponding to 20% of the available computing power, and serving tens of different user communities (such as astro-particle and high-energy physics). WNoDeS is fully integrated into the 7500 cores Tier-1 farm. There are at the moment about 20 supported VM images that can be chosen for the instantiation of virtual machines. Some of these images are used to provide support to legacy operating systems, whilst other have been created upon specific requests of user communities expressing the need of tailored execution environments.

This document describes the requirements that were collected and that are related to the needs of both users and resource providers with regard to virtualized systems and services. These requirements have been used to architect and develop the WNoDeS framework.

For each requirement, we first provide a description of the requirement itself; we then give an explanation of how WNoDeS addresses or plans to address the requirement.

Since requirements are by their nature dynamic, and since WNoDeS is a constantly evolving product, this document must be considered a work in progress.

2 DESCRIPTION OF THE REQUIREMENTS

2.1 SUBMISSION OF JOBS TO CUSTOM COMPUTE NODES BY USERS OF A TRADITIONAL COMPUTING CENTER

2.1.1 REQUIREMENT

Users of a traditional computing center normally have a local account, used to submit computational jobs through a batch system. The requirement is to be able to submit this type of jobs to machines tailored with a user-defined execution environment. This may refer to a specific operating system or operating system version, specific libraries, or else.

2.1.2 WNODES ANSWER

WNoDeS can associate traditional batch jobs belonging to a user or to a set of users to virtual machines specifically created for them. Each virtual machine is customized to match the job requirements of the user. This is transparent for users, who do not need to change anything from their point of view. In other

words, users submit jobs using the same commands (or graphical interfaces) with or without WNoDeS: in the case of WNoDeS (depending on user requirements and on the agreements with the user or its community), jobs will just be executed on dynamically-created virtual machines.

2.2 SUBMISSION OF JOBS TO CUSTOM COMPUTE NODES BY USERS OF A GRID-BASED DISTRIBUTED INFRASTRUCTURE

2.2.1 REQUIREMENT

The Grid middleware allows users to submit jobs to distributed resources belonging to a Grid infrastructure. The requirement is to use the same middleware and the same interfaces to submit either traditional Grid jobs or Grid jobs requesting a specific execution environment.

In addition, if Grid match-making is used, these jobs should automatically be sent only to sites where the execution environment requests can be satisfied.

2.2.2 WNODES ANSWER

For WNoDeS, this is a simple extension of the case above. A Grid job will be eventually handled at a site by a Grid Computing Element (CE). The CE then passes the job and its requirements to an underlying batch system. At this point, the job can be handled by WNoDeS as a local job, simply taking into account special Grid requirements, if present. No special handling or interface change is required by either Grid users or Grid middleware.

For example, all jobs belonging to certain Virtual Organizations (VOs) could be automatically and transparently directed to pre-packaged virtual machines, agreed between a site and the VOs.

For what regards brokering of individual Grid jobs requiring execution on custom environments, a WNoDeS site will just need to publish an appropriate capability into the Grid information system pertaining, for example, to the support of a certain virtual machine. Grid users requesting automatic brokering via a Grid WMS service may then insert a special request (for example, for a specific virtual machine, or for a certain number of cores or quantity of memory to be present on a virtual machine) into their Job Description Language (JDL) script, selecting the desired capability.

2.3 SELF-ALLOCATION OF COMPUTE RESOURCES BY USERS OF A TRADITIONAL COMPUTING CENTER

2.3.1 REQUIREMENT

Most sites offering distributed remote access also have local users that need interactive access, direct and local access to the batch system, or user interfaces able to access the distributed infrastructure.

The requirement is to allow these local users to self-allocate one or more systems so that they can log onto them using their local credentials (i.e., without requiring root access). A user should be able to specify both generic (such as operating system) or site-specific (such as local filesystems to be mounted) characteristics. The provided system can then be employed by users for instance to create pools of machines for interactive analysis, personal user interfaces, or to instantiate ad-hoc services.

2.3.2 WNoDeS Answer

WNoDeS provides a mode called Virtual Interactive Pools (VIP). VIP is a scripted interface to WNoDeS, allowing local users to self-instantiate virtual machines, specifying characteristics such as machine type, number of cores, RAM, local filesystems to be mounted. This is a sort of Cloud computing applied to a traditional computing center, and like Cloud computing can be expanded (depending on the services that a site wants to offer) to cover higher-level services rather than simple hardware.

For users, this means additional flexibility. For resource centers, this speeds up provisioning and avoids to statically allocate machines for the purposes mentioned in the requirement, thus increasing overall utilization of the resource pool.

2.4 Self-allocation of compute resources by distributed users (Cloud Computing)

2.4.1 Requirement

The requirement is for distributed users to be able to autonomously self-allocate compute resources. These resources may be defined by several parameters like, for example, computing power, network connectivity, network addresses (public or private), transient or permanent storage. It should be possible to allocate the resources via either a Web-based interface, or programmatically via an Application Programming Interface (API). The resources should be billed using a pay-as-you-go model, typical of cloud computing.

2.4.2 WNoDeS Answer

WNoDeS can deliver Cloud services in two ways: firstly, through the Open Cloud Computing Interface (OCCI) API, developed within the Open Grid Forum. Secondly, through a Cloud Web portal, allowing users to self-instantiate resources (or possibly services) using a graphical front-end.

2.5 Integrated monitoring and accounting

2.5.1 Requirement

Both users and resource centers would like to have a uniform method to monitor resources and account for their usage.

At their respective levels, this translates into the requirement for a clear and single experience for what regards operational and billing workflows, regardless of the resource interface (local, Grid, Cloud) requested by the users.

2.5.2 WNoDeS Answer

WNoDeS resource requests, regardless of the customer interface used (local, Grid, Cloud), are always transparently (for the user) mediated through a batch system.

Accounting workflows with WNoDeS, therefore, do not need to change once a computing center expands its services from local to Grid or to Cloud services, because all resource requests are always tracked by

the batch system itself. For resource providers, this often means being able to re-use existing accounting workflows and to provide consistent interfaces to users.

Monitoring of resources provided by WNoDeS can also be done normally re-using existing procedures, possibly complemented by standard (libvirt-based) monitoring to check for virtual machines-specific behavior.

2.6 RESOURCE ACCESS BY DISTRIBUTED USERS THROUGH MULTIPLE AUTHENTICATION METHODS

2.6.1 REQUIREMENT

While Grids widely adopt X.509 certificates to manage authentication, this is by no means the only authentication mechanism in use.

From a user's point of view, the requirement is twofold: on the one hand, for users without Grid-based X.509 certificates, it means to be able to access services (be they Grid- or Cloud-based) using other methods, like Shibboleth federations, or local, credit-based access. On the other hand, for users already using Grid middleware, it means exploiting Cloud computing-based interfaces for resource instantiations without the need to handle multiple credentials.

From a resource provider's point of view, the requirement translates into being able to support and integrate multiple authentication methods, possibly connecting to existing distributed authentication infrastructures.

2.6.2 WNODES ANSWER

Besides providing its own Web portal for Cloud computing instantiations, WNoDeS is architected so that it can be integrated into a generic submission portal (see also the corresponding requirement elsewhere in this document).

The authentication layer is not part of the WNoDeS core; WNoDeS can therefore use any authentication method, as long as it can be eventually mapped into a known and unique resource ownership seen by a batch system. This ownership can be easily shared across different resource interfaces: for example, a Grid user can authenticate to the WNoDeS Web portal using his Grid certificate. This is the authentication method that will be used then for both Grid and Cloud resource requests for that user. Conversely, a non-Grid user, through the generic submission portal, will be able to authenticate by using for example a Kerberos or Shibboleth token. This token will eventually be translated into a local resource id, used to instantiate and account resource requests.

2.7 FLEXIBLE AND DISTRIBUTED AUTHORIZATION POLICIES

2.7.1 REQUIREMENT

It should be possible to define authorization policies in a consistent and flexible way, according to the contracted terms of service. These can refer to distributed resources, so it should be able to enforce the authorization policies in a distributed environment as well.

Within the agreed terms of service, it should be possible for local (or distributed) users or user communities to inspect and possibly modify these policies.

Resource providers, on the other hand, should be able to ultimately override users' policies in case they violate contractual terms of service or in case of security threats.

2.7.2 WNoDeS Answer

Rather than inventing a new solution for the problem of authorizing usage to resources, WNoDeS leverages EMI Argus, that is a flexible service designed to render authorization decisions considering policies in defined local and distributed environments.

Argus allows its users to define and modify authorization policies through a flexible policy definition language, and allows secure propagation of these policies across multiple centers. Since WNoDeS does not fundamentally treat different types of instantiation requests (local, Grid, Cloud) in different ways, the integration between WNoDeS and Argus provides a uniform and consistent way of handling authorization.

2.8 Support for multiple customer communities and for flexible reconfiguration of resources

2.8.1 Requirement

Large resource centers often face the need to provision computing environments, possibly accessible in a distributed way, tailored to the needs of multiple customers or customer communities. For example, the Grid world often standardizes on a common set of operating system, middleware, libraries and software to be installed on the compute nodes. But since not all users of a computing center use the Grid, this often results in statically dedicating part of the resources to Grid computing and part to other uses (for example: Cloud computing, or local access), thus decreasing overall utilization of the resources and increasing administration overhead.

The requirement is therefore to support the needs of multiple customers or customer communities, while at the same time having the flexibility to dynamically reconfigure resources so that the same resource can be seen for example at one time as a Grid resource, and at another time as a Cloud resource. This should be done without statically partitioning the available resource pool.

2.8.2 WNoDeS Answer

Since WNoDeS does not fundamentally treats different types of instantiation requests (local, Grid, Cloud) in different ways, it allows dynamic instantiation of resources, so that they can be presented to users as Grid resources, Cloud resources or local resources, as requested. In the WNoDeS framework there is no need to statically define or keep alive any number of virtual machines. They will be created on-demand, with (optional) caching mechanisms to optimize the machine provisioning time.

2.9 Virtualization techniques must be efficient, scalable and cost effective

2.9.1 Requirement

If virtualization techniques are used to offer services, they should satisfy several constraints.

Firstly, the performance penalties incurred by virtualization must be known and as low as possible for what regards key metrics like CPU, I/O and network utilization.

Secondly, any framework used to provide these services must be engineered so that it is provenly resilient and scalable; for a medium to large size computing center, this normally translates into the need to be able to concurrently support several thousands of dynamically created virtual machines, accessing several petabytes of shared storage space.

Thirdly, the framework must be cost effective. This on the one hand means that maintenance and know-how to operate it should be low. On the other hand, it is often desirable that the framework should not necessarily require the purchase of commercial licenses.

2.9.2 WNoDeS Answer

For the first point (performance), a key point is that WNoDeS does not aim to be a generic virtualization engine. It focuses on one virtualization hypervisor technology, namely Linux KVM. For this hypervisor, performance numbers are published and available. Since WNoDeS assumes that the virtualization layer is handled by KVM, it can apply known and documented ways to improve performance when virtual machines are involved, in particular in areas where virtualization penalties are higher; this is more difficult to obtain with frameworks supporting many hypervisors, and that often need to find a compromise between performance customizations and support of multiple virtualization engines.

For the second point (resiliency and scalability), WNoDeS has been used in production mode since November 2009, with more than 2000 concurrent virtual machines, at the largest INFN computing center (CNAF, Bologna), and is deployed at other sizable Italian computing centers. WNoDeS has proven to be capable of handling several tens of thousands of production allocation requests per day, and to interact with storage systems providing several petabytes of disk or storage space. Since its adoption in production mode at INFN-CNAF only, WNoDeS has handled several millions of virtual machines.

For the third point (cost effectiveness), WNoDeS is in reality a thin layer of software leveraging several other proven software technologies. This has several benefits: maintenance and developing efforts can be low because several functions are managed by other, specialized services. Know-how can be reused because many times the technologies leveraged by WNoDeS (for example, batch system configurations, accounting procedures, virtualization through Linux KVM) are already present at a medium to large size computing center. Finally, WNoDeS does not necessarily require the use of commercial software. For example, the batch systems currently supported by WNoDeS are the commercially available Platform LSF, or the open source PBS/Torque.

2.10 Flexibility in handling network isolation

2.10.1 Requirement

It should be possible for both users and resource providers to flexibly allocate resources in public or private address space, as defined in the contracted terms of service. Any framework designed to provision virtual machines should also allow the possibility to logically isolate resources across customers or customer communities, and allow traceability of virtual machines traffic or behavior.

The requirement above should also be integrated into complex networking set-ups typical of large computing centers. In particular, it should be possible to dynamically define and enforce VLAN isolation without necessarily reconfiguring the physical network switches or routers present in the computing center.

2.10.2 WNoDeS Answer

WNoDeS has been supporting VLAN isolation since its inception. The definition of virtual machines address space is not directly handled by WNoDeS, so any local policies can be applied. For example, it is possible to assign private addresses to a customer and a group of customers, and have WNoDeS honor this policy so that virtual machines instantiated for these customers can get private addresses.

When thousands of concurrent virtual machines are involved, with tens or hundreds or more of customers, it may often become practically impossible to dynamically define VLANs by reconfiguring (either manually or via switch protocols) the underlying networking layer. One of the WNoDeS development areas regards the possibility to handle, in a scalable way, a dynamically created private VLAN structure overlaying the physical networking infrastructure. This has the aim to make it possible to efficiently and dynamically define VLANs in complex networking set-ups.

2.11 Flexibility in interfacing with site-specific configurations

2.11.1 Requirement

From a resource provider point of view, it should be possible to tailor any solution to keep into account some of the key existing component of a computing center, like batch system, or storage system.

2.11.2 WNoDeS Answer

Most computing centers currently handle requests for provisioning compute resources through a batch system. WNoDeS recognizes and leverages this fact, making a batch system the center of its match-making and instantiation mechanism. Currently, WNoDeS supports two amongst the most used batch systems, namely, Platform LSF and PBS/Torque.

Storage systems may be accessed via a plethora of methods. WNoDeS currently uses Posix I/O to access files, thus supporting all distributed file systems providing a Posix interface. It is also possible to easily customize WNoDeS so that gateway systems to mediate between WNoDeS and a file system can be used; for example, a hypervisor-based NFS-to-GPFS gateway was developed, to avoid mounting a GPFS filesystem on all virtual compute nodes.

2.12 Virtual images should be site-independent, certified and possibly distributed

2.12.1 Requirement

There should be mechanisms in place for virtual images used by or within a resource center to be approved and certified. Depending on the center, it should be possible to allow or disallow users to bring in and use their own images. Images should in principle be site-independent, with mechanisms for individual resource centers to tailor them. It should be possible to re-use approved and certified images across a distributed infrastructure.

2.12.2 WNoDeS Answer

WNoDeS actively follows the work of the HEPiX working group on virtualization and of the EGI Security Policy Group; one of the main goals of these groups is to work on the definition and endorsement of

virtual machine images. WNoDeS is therefore committed to support methods that allow the use of site-independent, certified virtual machines images. Depending on resource center policies, these images may or may not allowed at a center or imported from distributed repositories.

2.13 USERS OF CLOUD SERVICES SHOULD BE ABLE TO EXPLOIT EXISTING GRID SERVICES

2.13.1 REQUIREMENT

In the past ten years, European and worldwide resource and infrastructure providers have agreed to provision substantial computing and storage capacity through Grid interfaces. According to the contracted terms, it should be possible to exploit these existing Grid-based services and infrastructures and make them available to users accessing them for example through Cloud computing interfaces.

2.13.2 WNODES ANSWER

Through the adoption of a general access portal (see relevant requirement elsewhere in this document), WNoDeS plans to transparently assign short-lived online certificates to users authenticated via several methods. Following contracted terms of service, these certificates may be inserted into Virtual Organizations (VOs) - or to subsets of VOs (using for example roles); the final result is that users accessing resources without Grid credentials may be able to access Grid services.

2.14 TRANSPARENT ACCESS TO MULTIPLE CLOUDS

2.14.1 REQUIREMENT

In the Cloud computing framework, it should be possible to access and be charged for resources belonging to multiple Clouds. This should happen taking into account security (authentication and authorization), possibly with user-transparent brokering mechanisms. In general, open standards to access these multiple Clouds are preferred by both customers and resources centers, as a way to simplify user experience and to avoid vendor lock-in.

2.14.2 WNODES ANSWER

Connecting multiple Clouds is a difficult task. Rather than re-architecting procedures for authentication, authorization or else for this purpose, WNoDeS wants to leverage on the experience gathered at several large, production Grid centers in ten years of development and operation of Grid middleware. This experience successfully lead to handling millions of Grid jobs worldwide, and to the set-up of trans-national infrastructures like EGI.

WNoDeS therefore plans to address the interconnection of multiple clouds re-using mechanisms for distributed resource discovery, inventory and brokering taken from the Grid world to the best possible extent.

2.15 TRANSPARENT SUPPORT FOR "FLASH" REQUESTS

2.15.1 REQUIREMENT

The Cloud computing model often promises "unlimited capacity", i.e., the ability to dynamically provide additional computing or storage capacity when needed.

For users, the requirement is the possibility to access additional resources at a premium in case of urgent needs (for example, in the scientific world, in close proximity to an important conference, when often unexpected computations need to carried out). This should happen in a transparent way, i.e. without necessarily resorting to other resource providers.

For resource providers, this often means the ability to act as transparent proxies on behalf of customers toward other public or private providers.

2.15.2 WNoDeS Answer

The WNoDeS architecture explicitly foresees the possibility to access external resources as if they where local through the use of a transparent proxy. One of the WNoDeS development areas focuses on developing this part, with two main goals: exploring, on the one hand, the possibility to pool WNoDeS sites; and, on the other hand, the possibility to instantiate resources on behalf of WNoDeS customers through external providers.

2.16 Long-term data access

2.16.1 Requirement

Many scientific experiments or collaborations do not have a simple way of accessing, reprocessing and analyzing data that was collected in the past, although this data may still be up-to-date and relevant. This is for example related to the fact that computers used at the time to perform scientific tasks are now outdated, or considered unsecure.

For users, the requirement is to be able to permanently access and work on this data.

For resource providers, on the other hand, this means making sure that providing systems for this purpose and allowing users to access to them does not violate privacy or security policies.

2.16.2 WNoDeS Answer

WNoDeS supports on-demand instantiation of virtual machines; it also integrates mechanisms to logically separate virtual machines (see relevant requirement elsewhere in this document). It can therefore be used to design and support long-term data access projects.

2.17 User experience should be integrated into a general access portal

2.17.1 Requirement

Computing and storage resources may be accessed under several different forms. For example, one could submit jobs to a local resource center, to a Grid-based distributed infrastructure, or one could self-instantiate services or resources (Cloud computing). The requirement is to have a single access portal where these tasks can be carried out, regardless of the specific interface (local, Grid, Cloud) used to access resources, and with consistent and integrated accounting, authentication, and authorization policies.

2.17.2 WNoDeS Answer

WNoDeS is a key part of a general access portal that is being designed by the Italian Grid Initiative (IGI) to satisfy the requirement defined above. Through this portal and WNoDeS, it will be possible to integrate Grid, Cloud and generic service requests.

2.18 Efficient handling and coexistance of real vs. virtual machines

2.18.1 Requirement

While virtualization technologies are often very useful and efficient, there are applications that are still not suited well to virtualization. These may include, for example, applications making massive use of I/O procedures, or applications requiring devices that are difficult to virtualize (like, for example, GPGPU).

For users, the requirement is to be able to dynamically choose whether to provision services based on real or on virtual machines, depending on the users' needs.

For resource providers, on the other hand, the requirement is to maximize resource utilization, avoiding to dedicate subsets of resource pools to physical machines and subsets to virtual machines.

2.18.2 WNoDeS Answer

WNoDeS recognizes that virtual machines are not currently suited for all applications. One of the WNoDeS development areas, therefore, works on supporting a WNoDeS-based method of handling resource instantiation that allows users to specify whether their requests should be satisfied with real or with virtual machines. A key point the WNoDeS team is working on, is the possibility to efficiently mix and match real and virtual machines on the same hardware, with the goal to exploit the best of both worlds, and to give resource providers the flexibility to offer multiple services to their customers and, at the same time, minimize static partitioning of their resource pools.