Journal of Physics: Conference Series **664** (2015) 022032 doi:10.1088/1742-6596/664/2/022032

A prototype Infrastructure for Cloud-based distributed services in High Availability over WAN

C. Bulfon[1], G. Carlino[2], A. De Salvo[1], A.Doria[2], C. Graziosi[1],S.Pardi[2], A. Sanchez[2], M.Carboni[3,4], P.Bolletta[3], L. Puccio[3], V.Capone[5], L.Merola[6].

[1]INFN-Roma Unit - Italy

[2]INFN-Napoli Unit - Italy

[3]GARR the Italian NREN - Italy

[4]INFN-LNF Unit- Italy

[5]GÉANT – Cambridge UK

[6]University of Naples Federico II

E-mail: spardi@na.infn.it

Abstract. In this work we present the architectural and performance studies concerning a prototype of a distributed Tier2 infrastructure for HEP, instantiated between the two Italian sites of INFN-Roma1 and INFN-Napoli. The network infrastructure is based on a Layer-2 geographical link, provided by the Italian NREN (GARR), directly connecting the two remote LANs of the named sites. By exploiting the possibilities offered by the new distributed file systems, a shared storage area with synchronous copy has been set up. The computing infrastructure, based on an OpenStack facility, is using a set of distributed Hypervisors installed in both sites. The main parameter to be taken into account when managing two remote sites with a single framework is the effect of the latency, due to the distance and the end-to-end service overhead. In order to understand the capabilities and limits of our setup, the impact of latency has been investigated by means of a set of stress tests, including data I/O throughput, metadata access performance evaluation and network occupancy, during the life cycle of a Virtual Machine. A set of resilience tests has also been performed, in order to verify the stability of the system on the event of hardware or software faults.
The results of this work show that the reliability and robustness of the chosen architecture are effective enough to build a production system and to provide common services. This prototype can also be extended to multiple sites with small changes of the network topology, thus creating a National Network of Cloud-based distributed services, in HA over WAN.

1. Introduction

The enhancement of classes of service provided by the National Networks for Research and Education, together with the new distributed computing paradigms, allows to re-designing connections among the e-infrastructure supporting high energy physics experiments and other scientific applications. Today, the possibility to access to high bandwidth links among the sites is enhanced with additional services like traffic encapsulation, jitter and latency control, setup of overlay networks as the case of LHCONE[1], or by implementing long-distance layer-two connections. These

functionalities allow to better control the communication layer among the sites, to improve the performances, and to study the traffic flow behaviours. Moreover, the pervasive use of Cloud Computing technologies in data centres offers an additional degree of freedom and a more flexible way to distribute computing and storage services over distributed physical devices, stressing the paradigm of resource virtualization.

The mix of the Cloud Computing paradigm and the flexibility currently achievable in the geographic network connections opens a set of new opportunities and possible scenarios. In this work, we present the investigation activity carried on within the Italian National Project STOA, which involves two Tier2 of the ATLAS Experiment, located in the Roma and Napoli. The experiment takes advantage of the new national network GARR-X, provided by the Italian NREN GARR. More specifically, we focus on implementing a virtual Tier2, using Cloud Computing technologies, deployed over hardware resources in Napoli and Roma, directly connected through a geographic layer 2 link.

The distributed resources are highly coupled, thanks to a common storage area, replicated synchronously between the two sites, using the features of GlusterFS[2]. The goal is to create a completely redundant infrastructure able to host high availability services, in which we can migrate virtual machines transparently from a site to another. One of the main factor to take into account is the impact of the latency due to the distance between the involved resources. In this work we will show that the designed model can guarantee a high level of resilience and availability for the tested setup. Moreover, an additional study will measure the relative impact on the global performances, in term of throughput capabilities in performing basic operation, with respect to the increase of the distance among the sites.

The rest of the paper is organized as follows: in section 2 we describe the network infrastructure used to setup our testbed. In section 3 we introduce the design and we individuate the possible technologies to be used. Section 4 shows the current implementation of the designed model and the other possible setup. Then in Section 5 we present the tests made over our infrastructure, demonstrating the effectiveness of our model and of our implementation. Finally, in section 6 we summarize the results and we discuss the possible extension of our experience.

2. The Network Infrastructure

In the last years, the Italian Network provider GARR has greatly improved the national connectivity and has incremented the set of services that can be offered at research centres and Universities. In this section we will describe the technical parameters and the specific network technologies as reference for the successive analysis and future comparisons.

The new network infrastructure, called GARR-X, is based on three main elements:

- Dense Wavelength Division Multiplexing (DWDM).
- Optical Transport Network (OTN) ITU-T G.709, which provides protection mechanism, monitoring and signalling.
- Reconfigurable Optical Add/Drop Multiplexer (ROADM).

The OTN hierarchy make possible to multiplex the TDM (Time Division Multiplexing) services with different bit rate, introducing the additional overhead of control, monitoring and notification. On the GARR-X network is then very easy to create dedicated optical channel and setup high quality services between adjacent ROADM nodes.

Regarding the present work, we take advantage from an end-to-end service of 1Gbps that carries Ethernet frames between the client interfaces of INFN-Napoli and Roma La Sapienza. In figure 1 we show the elements of the GARR-X infrastructure that implement the link. More specifically, the transit node RM2 regenerates the ODU0 frame, then on the RM2-NA1 path the DWDM channel are regenerated by intermediate amplification stations.

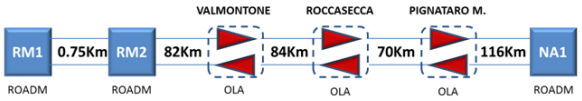

Fig. 1. The optical elements that realize end-to-end service between Napoli and Roma Sites.

At the setup time, we measured the link performances using common benchmarks, like IPERF and ping, the results showing a very stable link with a minimum latency of 5ms that guaranteed performance similar to a campus network. The main parameters are resumed as follow:

- 1Gbps link
- RTT: 5 ms
- JITTER (measured with IPERF in UDP) : 0.08 ms
- Throughput measured with IPERF between two servers: 938Mbit/s
- Packet Loss 0% on $1x10^6$ trials (measured with flooding ping)

At a later time, we reorganized the optical path up to obtain a 37ms link, in order to study the incidence of the higher latency on our model. In that case we used a very long and uncommon path, going through 22 optical equipment for a total of more than 2000 km of optical path.

The link characteristics and any variation on it are absolutely transparent for the sites, all the traffic is simply encapsulated in a separate VLAN (i.e. ID 155) and propagated as a local VLAN where we use common private RFC1918 IP address space, without conflicts with the already existing local network configuration. The figure 3 shows a logical representation of the network infrastructure.

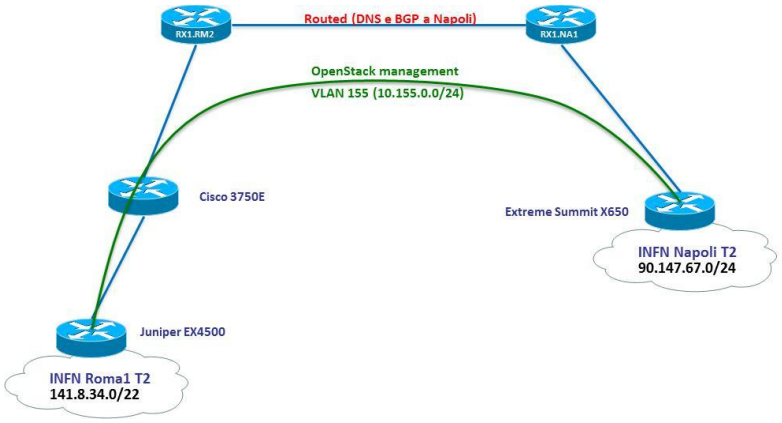

Fig. 2. The logical connection between INFN-Napoli and INFN-Roma through VLAN-155

3. A model for a distributed Tier2

The Tier 2 facilities of Napoli and Roma use the Cloud Computing paradigm, compliant with the NIST definition[9], to deploy a large part of the core services. Some examples are: the Grid frontend, user interfaces, information systems, accounting, and all the components that can take advantage from the flexibility that the virtualization provides, without suffering of the related, even if minimal, performance degradation.

The main idea is to share a common Cloud Computing infrastructure deployed over hardware distributed in the two remote sites. In that way, assuming that the network infrastructure works properly, we can guarantee a complete redundancy of the core services with the possibility to move them between two failure domains totally independent in term of power, cooling system, hardware etc. The network is obviously the key issue. More specifically, in order to design such a model, a common

IP address space is needed. Moreover, an advanced management of the storage sharing is needed to reach the goal of transparent service migration from a site to another.

The designed model can be described as a four-layer stack, represented in figure 3. The low level is the geographic network connection, which requires a layer 2 link deployable with several technologies. On top of this we have a Network File System with synchronous replication. The second layer enables the deployment of the Cloud Infrastructure with live migration capability that represent the layer 3 of our architecture. The final level is composed by the Tier 2 services deployable over the virtualized infrastructure.

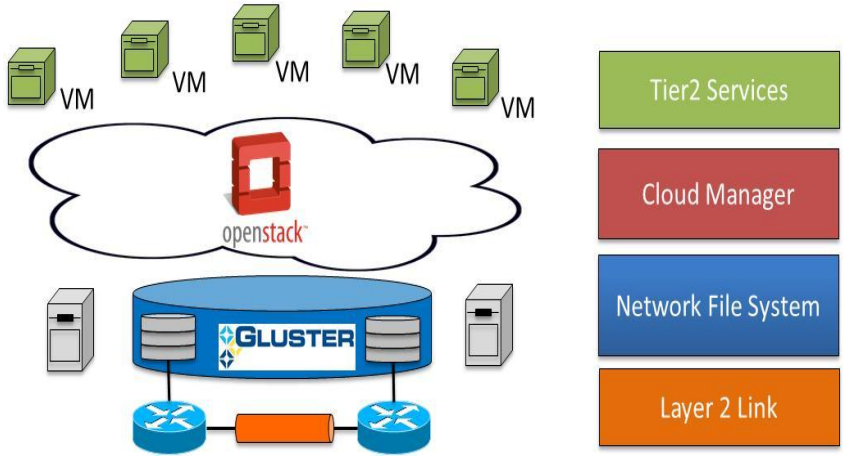

Fig. 3 The architecture of a distributed Tier 2 over a layer 2 geographic link

3.1. Technologies

The described model can be implemented using off-the-shelf technologies. As regarding the Layer 2 link among two or more distributed sites, several options are available, including MPLS pseudo-wire circuit, overlay network, as well as the usage of static techniques, as used in our implementation.

For the network file system there are several options that include open-source and commercial software that offer the synchronous replication data capability in a LAN environment; some well known examples are: GlusterFS, CEPH[3], Sheepdog[4], BeeGFS[5], as well as GPFS and several other solutions.

As regarding the Cloud Manager System, we have still several components that can help to complete the infrastructure layer of the system: Openstack, OpenNebula, Cloudstack and VMware are the most popular products that can be interchangeable due the independence of the stack levels individuated in our architecture.

Finally, we have a plethora of services that we can build over this distributed Tier2 class infrastructure and that can take advantage from the flexibility of the underlying infrastructure. In the next section we will introduce our specific implementation as an example of a real working system.

4. The Tier2 implementation between Roma and Napoli

We have created an example of a Tier 2 distribution facility (see figure 4), by selecting a set of software and hardware components with the goal to be: stable, reliable, functional and cost effective.

4.1. The hardware

We mainly used a set of servers already present in the two sites of Roma and Napoli. More specifically:

- The storage is implemented over a couple of SuperMicro servers, one per site, configured as general-purpose multi-homed nodes with 6TB of internal disk space and 8-cores Intel Xeon E5405. In figure 4 they are identified with the labels st-na-01 and st-rm-01. Each server is connected to the site's LAN with public addresses and to the shared LAN over the dedicated VLAN 155. We have used the private network 10.155.0.0/16 to address all the hosts in the common LAN, while the public IP addressing and the connection to the WAN is managed by the Napoli sites through the local router.
- Hypervisors are a set of non-homogeneous general-purpose multi-core nodes that host the computing services of the cloud system. We used KVM virtualization solution over CentOS 6.X
- Network: Both sites use as border router the Extreme X650 with 24 10Gbps. The geographic layer 2 connection, has been setup using one of the four additional gigabit ports provided by the equipment.

The local DNS server in the Napoli site is used to manage name lookup services for all the distributed system, including the FQDN for the virtual machines running on the private network.

4.2. The software

The software components for our testbed are: GlusterFS (version 3.4.3) as the distributed file system, and OpenStack (Icehouse version) as the Cloud Manager System. All the machines are configured with CentOS 6.X. We have selected these products among the open-source solutions, looking for the more stable and more widely supported solutions. They can be easily integrated and offer all the features needed to setup our model. However, thanks to the complete modularity of the designed architecture, each component can be removed and replaced with a more performing one, if needed.

Finally, at the top level we use the standard WLCG middleware already deployed over the standard Tier 2 facilities in Napoli and Roma.

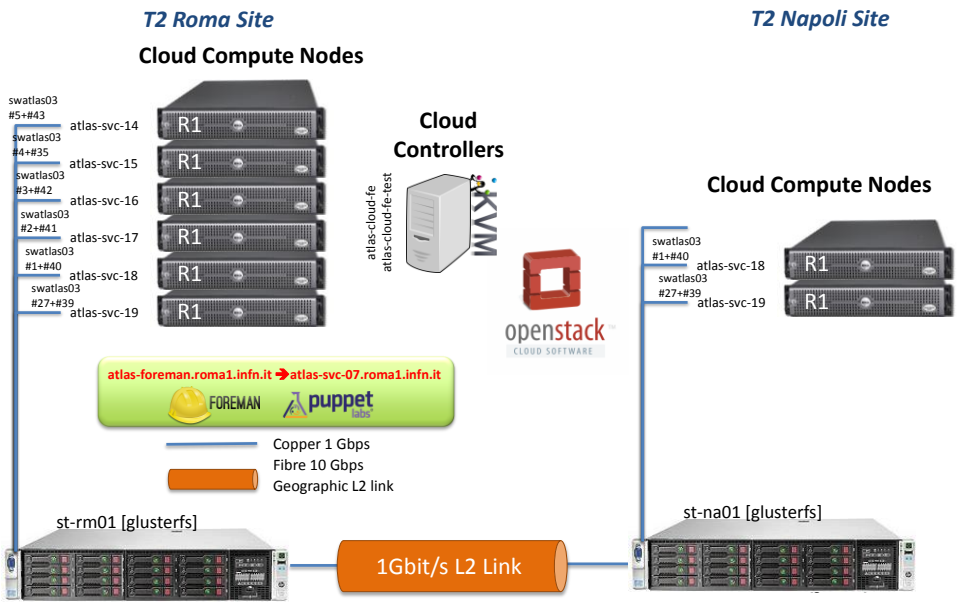

Fig. 4 – A representation of the Tier2 facility implemented between Napoli and Roma.

5. Testing the Cloud Infrastructure

In order to verify the stability and the effectiveness of our infrastructure, we performed a large set of tests, measuring different aspects of the whole system. As follows, we describe the goal of each test and the achieved results.

5.1. Stress test

In order to measure the resilience, and verify the correct behaviours of the storage system, we performed a set of stress tests using IOZone and Bonnie++, along with additional monitoring tools. All the benchmarks have demonstrated the resilience of the distributed storage with geographic synchronization by completing all the cycles without failures.

In figure 5 we show the traffic as measured on the geographic link during an intensive I/O test session made with IOzone Benchmark. The goal is to understand the network usage during the normal operation of read and write over the GlusterFS. We remember that the GlusterFS has been configured with a replica factor two over the two storage, so that when we perform normal operations of read and write, we are agnostic about the filesystem usage from the user point of view, i.e. we don't know if files are read from the local copy or from the remote one. In the same way, in case of write operation, we are not aware if the second copy of the file, needed to guarantee the requested replica factor two, is performed synchronously or not . A full network analysis during the IOZone benchmark, can show the exact behavior as describe as follow.

The traffic has been collected from the Ethernet card of the storage server in Naples. We noticed that, during the reading test, the node read from its local storage avoiding wasteful network usage, while during the writer activity he used 80% of the bandwidth with a single writing process. This behaviour is an effect of the storage system saturation, mainly due by the overhead introduced by the fuse module, used to mount the GlusterFS at both sites. The usage of the native API of Gluster in the next deploy will improve that behaviour.

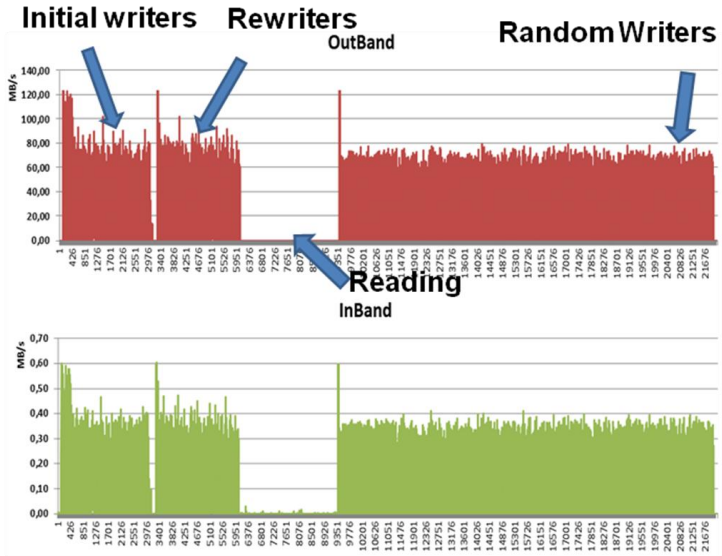

Fig. 5 Network Traffic OutBand and InBand measure on the storage serve of Napoli during and IOZone run. The benchmark has been configured to perform an a first session of write stress test (Initial writres), then a Rewriting session (Rewriters) followed by a reading phase and finish with a random writers test thate measures the performance of writing a file with accesses being made to random locations within the file. [10]

5.2. Live Migration

One of the most interesting features of our testbed is the possibility to live-migrate any Virtual Machine running on the infrastructure from a site to another, thanks to the distributed synchronously replicated storage between the two remote data centres.

We tested this capacity through the Live Migration functionality offered by KVM and libvirt for both link configurations: 5ms and 37ms of latency. Tests don't show any evident impact in term of

functional or performance due the increased latency. The monitoring system shows bandwidth saturation of the layer2 geographic link during a single machine migration, which guarantees the maximum performance during each single operation.

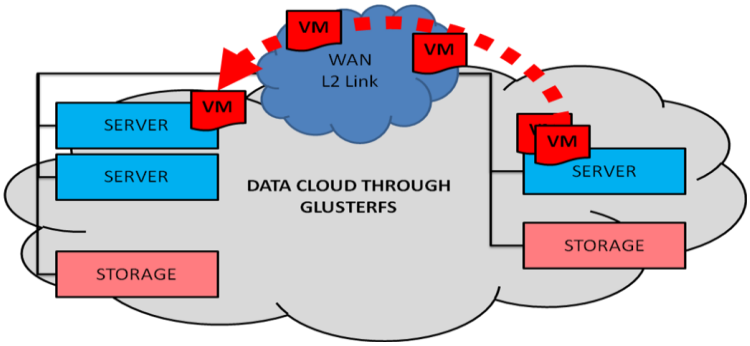

Fig. 6 Live migration scheme

We measured 10 seconds on average to move a basic and active Virtual Machine to another site, which could be considered as the lower time limit expected for each transaction. Notice that, as we show in figure 7, during the live migration process the maximum bandwidth is reached, guaranteeing that the process is completed at the maximum achievable speed.

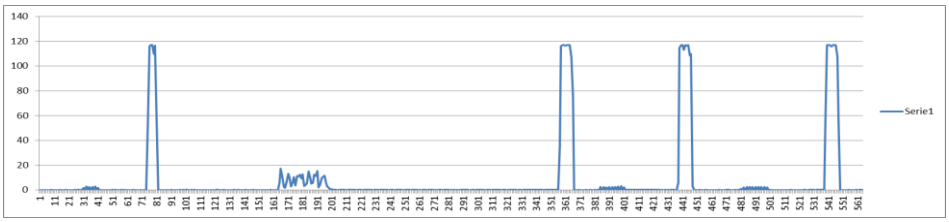

Fig. 7 Network monitoring during the live migration of a Virtual Machine from the Napoli to the Roma site.

5.3. The impact of latency

In order to study the bearing of network latency on the file system performances, we changed the end-to-end circuit topology over the GARR-X infrastructure to obtain a link with 37ms of latency instead of the 5ms of the optimal setup. We tested the new configuration through the smallfile[6] benchmark, a metadata stressing test. It works by creating a large set of small files (10 kb each one) and measures the performance of the following file system operations: Create, Read, Rename and Append.

We compared the results in three different configurations:

- Local File System
- Gluster file system 5ms latency
- Gluster file system 37ms latency

The graph in figure 8 shows that the performance decreases dramatically when the latency grows up to 37ms, however we noticed that the larger delay does not impact on the stability.

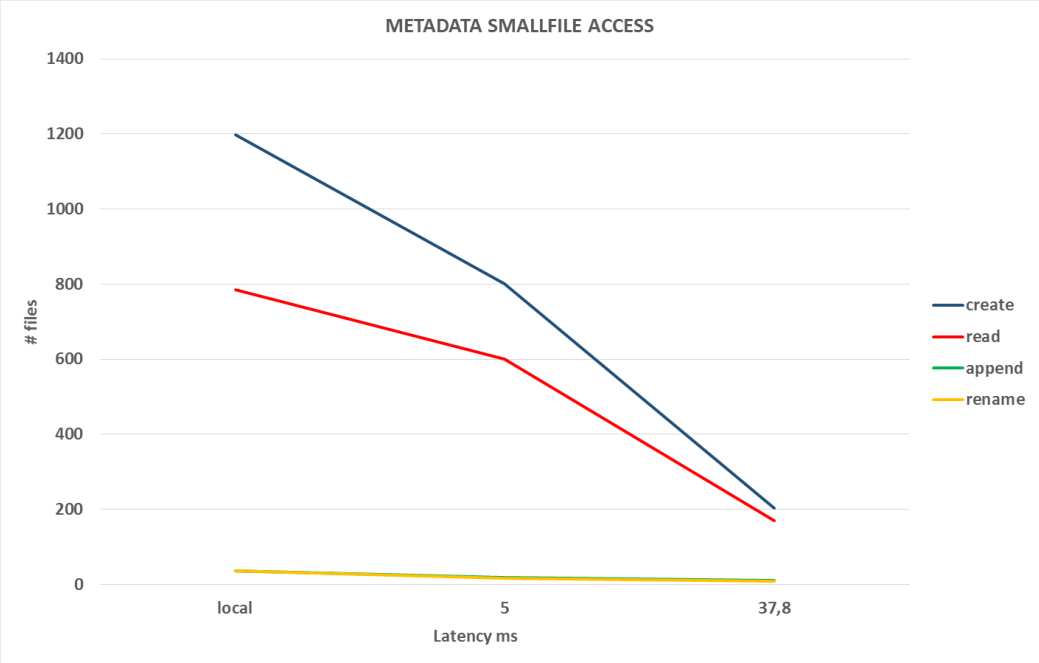

Fig. 8 In the graph we show the average number of operations for second in function of the latency.

6. Conclusion

The model presented in this work opens new scenarios to deploy Tier2-class computing infrastructures, taking advantage of the new features provided by the modern network infrastructures and by the pervasive usage of Cloud Computing technologies. Our implementation of the proposed model offers a concrete case study and an environment for testing and designing new classes of services.

The stress tests performed in both sites show that reliability and robustness of the chosen architecture are effective enough to build a real system and to provide common services. This prototype can also be extended to multiple sites, by changing the network topology and switching to the MPLS technology, following previous studies[6,7], thus creating a National Network of Cloud-based distributed services in HA over WAN.

Acknowledgment

This work was developed in the framework of the PRIN Project "/STOA-LHC 20108T4XTM/", /CUP: I11J12000080001,/ and partly supported by it.

References

[1]The official web page of the LHC Open Network Environment http://lhcone.net/

[2] The GlusterFS Home Page - http://www.gluster.org/

[3] The official web page of CEPH http://ceph.com/

[4] The Sheepdog Project http://sheepdog.github.io/sheepdog/

[5] The official web page of BeeGFS Bhttp://www.beegfs.com/content/

[6] The smallfile benchmark https://github.com/bengland2/smallfile

[7] F. Palmieri, S. Pardi: "Towards a federated Metropolitan Area Grid environment: The SCoPE network-aware infrastructure." Future Generation Computer Systems 26.8 (2010): 1241-1256.

[8] D. DelPrete, S. Pardi and G. Russo - "A Grid monitoring model over Network-Aware IaaS Cloud Infrastructure"- Journal: Int. J. of High Performance Computing and Networking, 2013 Vol.7, No.3, pp.195 – 204

[9] Peter Mell and Timothy Grance "The NIST Definition of Cloud Computing" - SP 800-145 - Sep 2011 http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[10] IOZone File System Benchmark http://www.iozone.org/docs/IOzone_msword_98.pdf