Cloud Network Infrastructure Design Approach

Vassil Gourov*, Elissaveta Gourova**, Borislav Lazarov*, Georgi Kostadinov*
* E-fellows Ltd., Sofia, Bulgaria
** Sofia University 'St. Kl. Ohridski', Sofia, Bulgaria
vgourov@efellows.bg
elis@fmi.uni-sofia.bg
blazarov@efellows.bg
gkostadinov@efellows.bg

***Abstract*—**During the past decade the cloud service market is one of the fastest growing segments around the world. The amount of companies that turned to the cloud has been steadily growing, since paying for a "shared" Cloud service over a given period of time reduces the capital expenditures and turned out to be better than using a dedicated hardware. This paper is focused on the architecture and the design of a shared public cloud service provider. The primary goal of the paper is to present a complete integrated solution for a single communication platform providing Cloud services to end-users. The paper, first, provides a literature review of some Cloud computing aspects, including the requirements for Cloud computing services and the key performance indicators to be evaluated. Second, it is described a real network infrastructure design approach, which allows smooth implementation of additional services and functionalities. All of the applicable functionalities are built to be managed and maintained separately for various independent customers in isolated mode of operations (Multi-Tenancy). The design allows optimal performance assuring high-availability of the service. Finally, the modular approach used in the design allows future optimisation and capacity upgrade plan of all key infrastructure components.

I. INTRODUCTION

Since the invention of the telephone more then a century ago rapid developments in all fields of science and technology have been witnessed. The trends in Information and Communication Technologies (ICT), especially, created enormous opportunities for fast access to data and information and their processing. While the development of telecommunications technologies facilitated the data transfer at high speed and regardless of geographical location, the Information Technologies (IT) created many new opportunities for increasing work efficiency of individuals, groups and organisations, for establishing new models of doing business, working, learning or entertaining.

The fast hardware developments following the Moore's law, on the one hand, and the increasing demands for computer power of the emerging applications, on the other, have forced organisations and individuals to regularly change their IT equipment. The heavy investments in technology, often underutilised [2], and the expected return of investments, as well as the increasing requirements for skills and knowledge for their deployment are among the driving forces in the uptake of Cloud computing [1]. This is closely related to a phenomenon that computers have become more powerful and less expensive, however, the pervasiveness of ICTs and the increasing management complexity is linked to growing expenses for organisations [2]. Subsequently, many organisations prefer nowadays to focus on their core capabilities and to outsource other functions introducing high overhead, such as ICTs. This is especially important for Small and Medium Enterprises (SME) which have difficulties to find highly-skilled professionals to maintain ICTs in-house [2], [4]. Thus, the opportunities to hire infrastructure or applications according to the real demands are decreasing the entry barriers for technology adoption in SMEs, and the time to market. Cloud computing offers also benefits like fast access to hardware resources without new capital investments, enhanced opportunities for scaling of services [2], as well as uninterrupted services and easier management [3]. At the same time, there are many evidences that the hardware emissions due to extensive IT use could be decreased by using Cloud services, contributing to more broad challenges of present-day economy like environmental sustainability [1], [3].

While the early providers of Cloud computing services are mainly in the United States [5], [7], there are evidences of their uptake also in Europe. Recently, a new Cloud infrastructure was developed in Bulgaria by eFellows Ltd. - a company established 10 years ago with a main focus on development and implementation of solutions in the field of information security, network and communication infrastructure, storage systems and data consolidation. Today the company has customers throughout the world in various fields.

The primary goal of the paper is to present the design approach followed by eFellows Ltd. for developing a Cloud-based network infrastructure. The first part of the paper provides an overview of Cloud computing and especially the concepts for Infrastructure as a Service (IaaS), and the challenges and requirements for its design. Second, a specific IaaS solution focused on meeting customer's needs is presented with special emphasis on physical and logical structure, and the orchestration options.

II. TRENDS IN INFRASTRUCTURE AS A SERVICE

A. Understanding of Cloud Computing

The Cloud computing phenomenon emerged in the 2000s, however, its core concepts could be traced back to the remotely connected terminals to mainframe computers, and later to the grid computing in academic institutions [1], [6]. Some authors [3] link this phenomenon to the advances in telecommunications, and in particular, the provision of Virtual Private Network

(VPN) services with comparable quality of service at a much lower cost. In fact, the emergence of Cloud computing was enabled by the uptake of three core technologies – Virtualisation, Multi-Tenancy and Web services [2]. By virtualisation the physical characteristics of the computing infrastructure or platform are abstracted and encapsulated, thus, hidden from end-users, and can be configured on demand, maintained and replicated very easily [2], [14]. In some cases these resources are partitioned (1:N) in multiple virtual elements, and in others – aggregated in a single virtual resource (N:1) [6]. In fact, through the virtualisation technique Cloud computing services are characterised by a front-end, seen by end-users, and a back-end, where the physical resources are configured, monitored and administrated by the providers [3]. This facilitates the Multi-Tenancy approach, whereas different users are served by the same resources without having any interference between them, and using independently the virtual resources allocated to them [2], [6]. Both, virtualisation and Multi-Tenancy, allow better utilisation of system resources available, thus leading to lower upfront and operational costs [2], [10], [14]. The Web-Services technology, on its side, facilitates the interoperability and interaction of different resources providing them standard interfaces over the network [2].

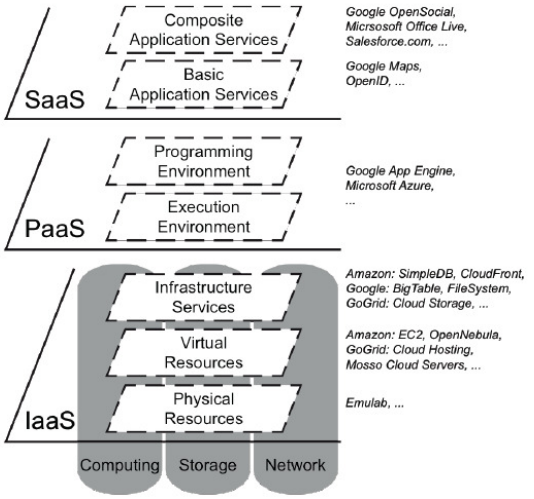

Figure 1. Layered Cloud computing infrastructure [6]

One of the widely quoted definitions of Cloud computing given by the National Institute of Standards and Technology [8] considers it as *"a model for enabling convenient, on-demand network access to a shared pool of configuration computing resources that can be rapidly provisioned and released with minimal management effort or service provider (SP) interaction."* It is widely accepted [3], [5], [6], [9], [11] that Cloud computing has three service layers (Fig. 1):

- *Software as a Service (SaaS)* ensures users access through a client network interface to commercially available software applications, which users do not need to install and maintain on their own computers.
- *Platform as a Service (PaaS)* ensures a computing environment where end-users can deploy their own software applications using the Cloud infrastructure of the provider, thus, avoiding additional costs for obtaining and supporting own computer platforms.
- *Infrastructure as a Service (IaaS)* provides end-users ICT infrastructure that is dynamically scalable on-demand, thus, allowing users to manage network and fundamental computing resources (e.g. servers and data centers) on which to run their operating systems and applications. IaaS providers manage the physical servers and network resources, and normally offer virtualised infrastructure as a service.

In comparison to computer grids, which are more research oriented, Cloud computing is more commercially oriented. Its basic characteristics comprise [2], [6], [14]: user friendliness and on-demand services, resource pooling, virtualised physical resources, architecture abstraction, dynamic scalability of resources, elastic and automated self-provisioning of resources, ubiquity, operational expense model, e.g. pay-per-use model, and Service Level Agreement (SLA). Specific for Cloud computing is the Multi-Tenancy, e.g. the provision of services to different types of users [9]: individual users, business users (large enterprises and governments, educational and research organisations and SMEs), developers and independent software vendors. The Cloud computing providers operate in two business models: variable (pay-for-your-usage) plans and fixed plans [5]. Variable plans allow customers to pay only for the resources actually consumed (e.g., instance hours, data transfer), and are consider as one of the major benefits of Cloud computing [9].

B. Requirements for Cloud services

The research literature [12] suggests that when designing an Internet-based system along with its functional characteristics, specific emphasis should be made on balancing its non-functional requirements: *availability* (depending on the reliability of all system components to work without any failure, and its robustness to cope with a failure), *performance* (measured with response time to users requests), *scalability* (ability to ensure performance when the number of users grows), *security* (ensuring normal system operation by controlling the access to its functionality while providing a reasonable degree of privacy), *manageability* (ability to monitor and alter the system runtime behavior), *maintainability* (how easy is to fix system problems or upgrade its components during runtime), *flexibility* (ability to produce new system versions or re-configure it) *and portability* (easy migration to a new environment). In the case of Cloud computing, most of these requirements are stressed by researchers as well [1], [6], [9], [13]. For example, Venters and Whitley [1] provide an overview of the specific requirements to Cloud computing systems (Table I). It is interesting to note their emphasis on providing equivalent opportunities to Cloud users in terms of security, response time and performance compared to similar services, and at the same time, offering benefits linked to service variety and scalability. In addition to these requirements, Develder et al. [6] consider the specific needs of different applications (targeted at scientific, business or individual users) in terms of resource volume and granularity (e.g. storage volumes, CPU performance and network bandwidth), elasticity and multiple tasks opportunities. As a specific issue is pointed out also the ability of Cloud services to be integrated with those available in-house [6].

TABLE I.
REQUIREMENTS TO CLOUD COMPUTING, ADAPTED [1]

technological dimensions	
Security Equivalence	at least equivalent in security to that experienced when using a locally running server
Availability Equivalence	at least equivalent in availability to that experienced when using a local server
Latency Equivalence	at least equivalent in latency to that experienced when using a locally running server
Variety	provides variety corresponding with the use for which the service will be put
Abstraction	abstract away unnecessary complexity for the service they provide
Scalability	service which is scalable to meet demand
service dimensions	
Efficiency	helps users be more efficient economically
Creativity	aids innovation and creativity
Simplicity	simple to understand and use

Garg et al. [13] point out that the Virtual Machine (VM) performance often varies from the promised values in the SLA, which reflects on the Quality of Services (QaS) offered to clients. Therefore, the authors propose a Service Measurement Index (SMI) comprising a set of indicators which could be used for taking a decision which Cloud SP better meets the specific user's requirements. In particular, for assessment of IaaS providers the authors suggest the following Key Performance Indicators (KPI):

- *Service response time* - measured in terms of the response time for making the service available for usage. In the case of IaaS, this includes provisioning the VM, booting the VM, assigning an IP address and starting application deployment;
- *Sustainability* - defined in terms of the environmental impact, e.g. can be measured as the average carbon footprint or energy efficiency of the Cloud service;
- *Suitability* - the degree to which users requirements are met, including both functional and non-functional requirements;
- *Accuracy* - measures the degree of proximity to the user's actual values when using a service compared to the expected values given in the SLA, e.g. the frequency of failure in fulfilling the promised SLA in terms of Compute units, network, and storage;
- *Transparency* - indicates the extent to which usability is affected by any changes in service, and can be measured as a time for which the performance of the user's application is affected during a change in the service or the frequency of such effects;
- *Interoperability* - ability of a service to interact with other services offered 0by the same provider or other);
- *Availability* - percentage of time a customer can access the service;
- *Reliability* - reflects how a service operates without failure during a given time and conditions, based on the mean time to failure promised by the provider and previous failures experienced by the users;
- *Cost* - depends on service acquisition and usage, e.g. could be measured according to the cost of one unit of CPU, storage, RAM, and network bandwidth;
- *Adaptability* - ability of the Cloud provider to adjust changes in services based on users requests;
- *Elasticity* - defined in terms of how much a Cloud service can be scaled during peak times;
- *Usability* - the ease of use could be measured in terms of the average time experienced by the previous users of the Cloud service to operate, learn, install and understand it;
- *Throughput* - evaluate the performance of infrastructure services, and depends on several factors that can affect execution of a task, e.g. number of tasks of the user application and the number of machines on which it runs;
- *System efficiency* - indicates the effective utilisation of leased services, e.g. its higher value is linked to a smaller overhead;
- *Scalability* - determines whether a system can handle a large number of application requests simultaneously, e.g. the ability to scale resources horizontally (e.g. 'scale out' Cloud resources of the same types) or vertically ('scale up' different Cloud resources assigned to a particular Cloud service).

In order to meet users' requirements, and more specifically, ensure guaranteed QoS and meet the provisions of the SLA, it is essential to undertake special monitoring and management efforts. As stressed by Mohamaddiah et al. [11], resource management in Cloud computing should focus on three interrelated processes: monitoring (infrastructure management and control using KPIs), allocation (assigning available recourses) and discovery (determining the most appropriate resources to meet SLA). Generally, the Services providers assign specific resources to the incoming job requests from end-users. This requires real-time information on the load and availability of physical resources. Therefore, the Infrastructure Provider monitors and controls the infrastructure performance and takes care of its optimal utilisation. In case of alack of physical resources, these could be hired from other IP or in case of higher availability – offered to them [11]. The management of Cloud resources is very important not only for the services provided to end-users, but also for the interrelations among Cloud providers in federated Cloud structures mediated by brokers [10].

III. DEVELOPMENT OF CLOUD INFRASTRUCTURE

A. Background

The main goal is to design a complete integrated solution for a single communication platform providing Cloud services to end users and internal users (company employees). The network infrastructure design should allow smooth implementation of additional services and functionalities. All of the applicable functionalities should ensure separate management and maintenance for different independent customers in isolated mode of operations (Multi-Tenancy), meeting all the Cloud services providers' special requirements. Some functionalities should be automatically activated and configured by the end users through a Self-Service Portal. All self-service features should be realised through a custom built "Orchestration" application, developed to automatically execute predefined template scripts (workflows) for configuration of specific functionalities and settings. Orchestration application should be tightly

integrated with the self-service portal to be also custom built in-house.

Other requirements for the infrastructure solution comprise:

- the logical and the physical solution designs should allow optimal performance, graceful degradation (fault tolerance) and increased availability and capabilities of the devices and interconnections between them;
- the solution should be built with redundancy at minimum 2:1, so that no Single Points of Failure components to be allowed;
- the solution should ensure scalability, e.g. to ensure future optimisation and capacity upgrade opportunities of all key infrastructure modules due to the need of growth, increasing number of customers/users or other system requirements.

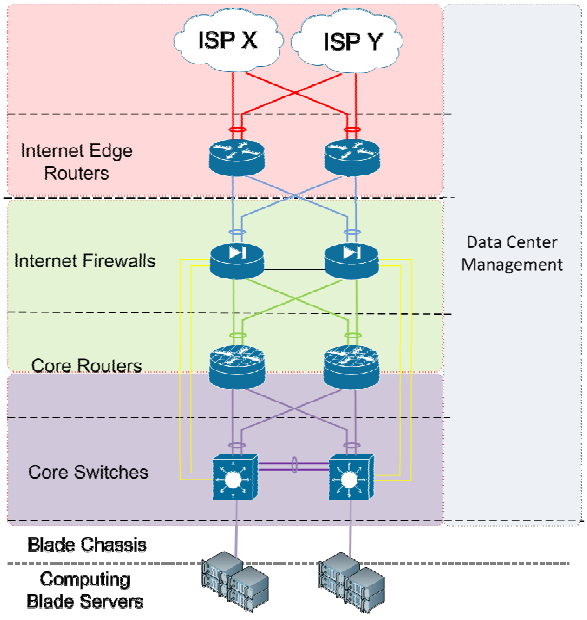

Figure 2. Simplified Physical Network Diagram

B. *Concept for IaaS*

In order to ensure these requirements, during the design phase were taken into consideration the best world and industry practices and manufacturers' recommendations. Special emphasis was put on the physical and logical network design. In order to meet the requirements for high performance, flexibility, scalability and high availability of the entire network infrastructure, it was separated into layers, including (Fig. 2):

- *Blade Chassis Networking:* consisting of some blade servers and converged network modules, whereas the latter provide additional layer of abstraction by presenting to the blade servers virtual Network Interface Controllers (NIC) over the physical 10 gbps NICs for both data and storage. They also provide external network connectivity over 1G/10G Ethernet and FibreChannel.
- *Data Center Switches:* The Core Switches have a function to provide switching between all physical and virtual hosts within the Cloud computing environment. A Data Center Access and management Switch: providing network access for single servers (outside the blade chassis) as well as Out-of-Band (OOB) management interfaces of any Data Center equipment (servers, storage, network).
- *Data Center Core Routers:* used to provide routing for all the subnets within the datacenter (both for clients/tenants and for internal/system use), as well as to provide VPN termination for Site-to-Site and Remote Access VPNs.
- *Internet Firewalls:* including devices that perform Network Address Translation (NAT) between private Virtual Machines (VM) IP addresses and public IP addresses, ensure Internet access control and security policies.
- *Wide-Area Network (WAN) Switches:* providing any-to-any full mesh connectivity between all WAN links (such as Internet Service Providers (ISPs), Metropolitan Area Network (MAN) transport, etc.) and the corresponding termination devices (Internet Edge Routers, Data Center Core Routers, other). In addition, these switches perform also the function of Edge Interconnect Switches in order to achieve resource optimisation.
- *Internet Edge Routers:* connecting to Upstream ISPs and exchange Border Gateway Protocol (BGP) routing information, and guaranteeing, thus, outbound and inbound network reachability for the Provider Independent IP network and company Internet Autonomous System.

The Logical network design (Fig. 3) determines how the entire network is divided into individual segments and how packets are routed between each of these segments. Generally, the network is divided into zones containing one or more individual networks with similar functions, connectivity and security needs. For the customers the internet firewalls split the cloud network into four security zones. The outside zone is the connection with the Internet. There are two demilitarised zones (DMZ) namely public and private. In the public zone all publicly available common and shared resources that need to be access from external sources over the Internet are placed here – e.g. Public DNS servers. The private DMZ is where all common and shared cloud resources, that need to be accessed by all or most of the internal client VMs are be placed. There are one or more subnets with specific access policies assigned to this zone and they are used for services like DHCP, DNS and etc. Furthermore, there is the inside zone, where the client networks reside. . Each of those networks represent a separate 802.1Q VLAN and are also reside in a different L3 routing table. This is how complete tenant isolation is archived. Additionally, there is a dedicated transport network between the core routers and the firewalls. All Client VM traffic destined to the Internet or to the Public/Private DMZ resources will be routed via this network. iBGP dynamic routing protocol is used to provide network exchange information and to ensure high availability and load balancing between the devices. Finally, there is a specialised zone where all remote access VPN (Similar to Dial-UP) users will be terminated and will have their VPN IP Address allocated. In the design five VPN client use cases are considered – office Local Area Networks (LAN), private Cloud, home LAN, mobile worker and home worker.

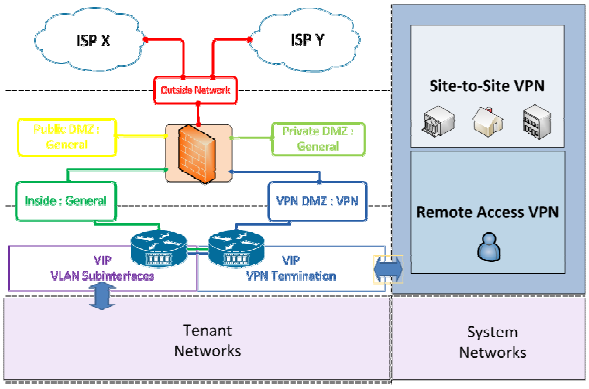

Figure 3. Logical Network Diagram

This design supports two VPN types: Site-to-site VPN and Remote Access VPN. The former provides permanent secure connectivity between the client's resources within the Cloud and a remote location based on infrastructure device(s) such as router or firewall. The Remote Access VPN type provides on-demand secure access to client resources within the Cloud via the Internet similar to dial-up, whereas users authenticate with username and password, configured via the self-service portal.

Internet connectivity is ensured by registering own Autonomous System and public IP Address ranges. BGP peering with any ISP can be established using them to guarantee adequate performance and high availability for the users. The design and the equipment supports adding new connections without service disruption.

All client Virtual Machines are allowed for outbound Internet connectivity using NAT performed by the Firewall by default. No additional configuration is required. Special features such as Anti-Virus and Web Content Filtering are supported by the Firewall and can be configured manually on demand.

Inbound Internet connectivity for client virtual machine is possible using static NAT (Virtual IPs) performed by the firewall. It can be configured automatically by the orchestration service on user requests in the self-service portal. Users are able to activate static or dynamic public IPs bound to a particular internal IP address of a virtual machine in the Cloud and to apply security filters based on IP protocol and port.

Management and Monitoring tools ensure:

- *Log Collection:* All infrastructure devices are configured to export event logs to an external server using standard SYSLOG protocol, which is useful both in routine troubleshooting and in incident handling.
- *Configuration Management and Backup:* All infrastructure devices are configured to archive configuration changes and to automatically backup configuration.
- *Monitoring:* All vital infrastructure components are being continuously monitored and graphed. If any component fails or falls out of acceptable thresholds an automatic notification is being sent to the support team.
- *Management Access to equipment:* Infrastructure devices' management interfaces require authentication with valid credentials. Only valid users who are members of a specific domain group are entitled with management access. Furthermore, there are access lists set on every device that allow access only from certain IP address ranges – used by the support team.

All backend operations and configurations that should be performed on the infrastructure triggered by end-user actions via self-service portal. The Orchestration options in case of Physical Networking are depicted in Fig. 4. All self-service features are realised through a custom built "Orchestration" application, developed to automatically execute predefined template scripts (workflows) for configuration of specific functionalities and setting. The communication between the application and the infrastructure is archived via standard command line interfaces (CLI) such as SSH and PowerShell.

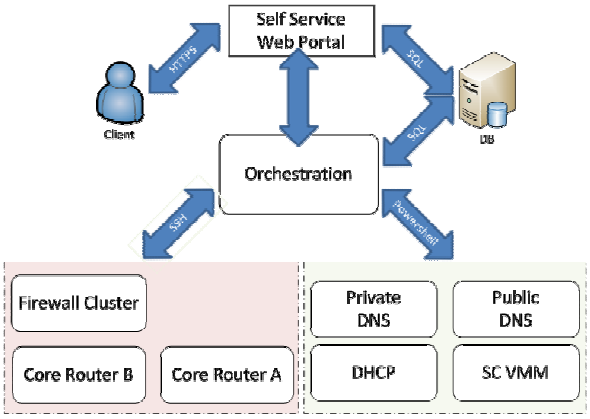

Figure 4. Physical Network Orchestration workflow

C. *Main features of the approach*

Usability: The self-service portal is easy-to-use and allows the users to build the server they want right away. From there on the clients can install and configure any software they want.

Reliability: Both, the logical and the physical solution designs allow optimal performance, graceful degradation (fault tolerance) and increased availability and capabilities of the devices and interconnections between them. Also, the solution has been built with redundancy at minimum 2:1, so that there are no Single Points of Failure components.

Scalability: The design includes future optimisation and capacity upgrade plan of all key infrastructure modules (components based or overall solution oriented) due to the need of growth, increasing number of customers/users or other system requirements.

Performance: All components of the infrastructure are in Active/Active state. This means that the traffic from the VMs is balanced trough one of the device pairs based on network segment the traffic comes from. The network throughput is 20 Gbps for inter-VM traffic, 5 Gbps for L3 interconnections and 600 Mbps for the Internet links.

IV. CONCLUSIONS

The paper provides an overview of the current cloud service requirements and evaluation. It gives an architectural description of a newly deployed cloud service provider in Bulgaria. The main features of the

suggested design are High Availability, Load Balancing and Multi-Tenant Isolation. Furthermore, the architecture uses a modular approach so the components can be easily upgraded.

The main problem of the approach is that it support a limited number of tenants. It is possible to add more devices horizontally, but this is not a very cost effective solution. This has been taken into consideration

REFERENCES

[1] W. Venters, E. A. Whitley, A critical review of cloud computing: researching desires and realities, *Journal of Information Technology*, 27, 2012, pp. 179–197.

[2] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, A. Ghalsasi, Cloud computing — The business perspective, *Decision Support Systems*, 51, 2011, pp. 176–189.

[3] Y. Jadeja, K. Modi, Cloud Computing - Concepts, Architecture and Challenges, *International Conference on Computing, Electronics and Electrical Technologies*, 2012, pp. 877-880.

[4] E. Gourova, V. Kadrev, A. Stancheva, G. K. Petrov, M. Dragomirova, Adapting educational programmes according to e-competence needs: the Bulgarian case", *Interactive Technology and Smart Education*, 11(2), 2014, pp. 123-145.

[5] Y. Han, Cloud Computing: Case Studies and Total Costs of Ownership, *Information Technology and Libraries*, *30*(4), 2011, pp. 198-206.

[6] Ch. Develder, M. De Leenheer, B. Dhoedt, M. Pickavet, D. Colle, P. Demeester, Optical Networks for Grid and Cloud Computing Applications, *Proceedings of the IEEE*, *100*(5), 2012, pp. 1149-1167.

[7] Schubert L. *The future of Cloud Computing: Opportunities for European Cloud Computing beyond 2010*. Exper Group Report: public version 1.0. European Commission, 2011.

[8] P. Mell, T. Grance, *The NIST Definition of Cloud Computing*, NIST, http://csrc.nist.gov/groups/SNS/cloud computing/ (accessed Oct. 21, 2010).

[9] S. Patidar, D. Rane, P. Jain, A Survey Paper on Cloud Computing, *Second International Conference on Advanced Computing & Communication Technologies*, 2012, pp. 394-397.

[10] D. Villegas, N. Bobroff, I. Rodero, J. Delgado, Y. Liu, A. Devarakonda, L. Fong, S. M. Sadjadi, M. Parashar, Cloud federation in a layered service model, *Journal of Computer and System Sciences*, 78, 2012, pp. 1330–1344.

[11] M. H. Mohamaddiah, A. Abdullah, M. Hussin, S. Subramaniam,A Proposed Architectural Framework for Resource Provisioning Mechanism in Cloud Computing, *1st International Conference of Recent Trends in Information and Communication Technologies*, Sept. 2014, pp. 312-327.

[12] P. Dyson, A. Longshaw, *Architecting Enterprise Solutions: Patterns for High-Capability Internet-Based Systems*, John Wiley & Sons, 2004.

[13] S. K. Garg, S. Versteeg, R. Buyya, A framework for ranking of cloud computing services, *Future Generation Computer Systems*, 29, 2013, pp. 1012–1023.

[14] U. Divakarla, G. Kumari, An Overview Of Cloud Computing In Distributed Systems, *AIP Conference Proceedings*, 1324(1), 2010, pp. 184-186.