Cloud Computing: An Overview

Libor Sarga

Faculty of Management and Economics

Tomas Bata University in Zlin

Czech Republic

sarga@fame.utb.cz

Abstract: *As cloud computing is gaining acclaim as a cost-effective alternative to acquiring processing resources for corporations, scientific applications and individuals, various challenges are rapidly coming to the fore. While academia struggles to procure a concise definition, corporations are more interested in competitive advantages it may generate and individuals view it as a way of speeding up data access times or a convenient backup solution. Properties of the cloud architecture largely preclude usage of existing practices while achieving end-users' and companies' compliance requires considering multiple infrastructural as well as commercial factors, such as sustainability in case of cloud-side interruptions, identity management and off-site corporate data handling policies. The article overviews recent attempts at formal definitions of cloud computing, summarizes and critically evaluates proposed delimitations, and specifies challenges associated with its further proliferation. Based on the conclusions, future directions in the field of cloud computing are also briefly hypothesized to include deeper focus on community clouds and bolstering innovative cloud-enabled platforms and devices such as tablets, smart phones, as well as entertainment applications.*

Key words: cloud computing, distributed architecture, security, economics, parallel computing

1. Introduction

Cloud computing is first and foremost a concept of distributed resource management and utilization. NIST (National Institute of Standards and Technology) defines it as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." (Mell & Grance, 2011) It aims at providing convenient endpoint access system while not requiring purchase of software, platform or physical network infrastructure, instead outsourcing them from third parties. The arrangement may beneficially influence competitive advantage and flexibility but it also brings about various challenges.

As stakeholders require ubiquitous access to and protection of data affecting their business operations, cloud became a valid option for cost-effective data redundancy, off-site backups, big data storages or IT outsourcing. However, nature of cloud computing itself may be in violation of corporate security policies regarding data retention and ability to guarantee security of physical network infrastructure. Some organizations even object to the notion of not having direct control over where the data resides. Addressing these issues is essential to understanding cloud architecture paradigm.

In recent years the number of articles providing resources for both experts and the public has been increasing steadily as have the test beds designed to facilitate research and experimental deployment of newly proposed standards. The article attempts to describe current cloud computing trends while simultaneously providing basis for further discussion. With success contingent on answering both existing and emerging issues, cloud computing may be aided by expertise and resources of the whole IT community and academia.

The rest of the article is organized as follows. In the second part we summarize definitions of cloud and grid computing models along with their differences. Due to both models being frequently thought of as identical, the distinction must be made clear. In the third part issues of currently existing parallel processing models are scrutinized in detail: uptime guarantees, security and privacy, legislature as well as economic aspects. The final part provides theoretically-based modifications to the cloud's functioning.

2. Cloud Computing

The first challenge cloud computing has to face is its definition, presenting clear and concise delineation encompassing its features and functional elements. Academic peerage in particular has been struggling to procure such universally accepted statement. Many contributions therefore surfaced focusing on just such task with results far from homogeneous in terms of quantity of features included.

Armbrust et al (2009) postulate: "Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we call a Cloud." The definition provides a comprehensive, holistic view of cloud's elements while mentioning one of several "… as a Service" models evolved when the paradigm became commercially viable for cloud operators to monetize: IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS, each of which differs in parts of hardware/software infrastructure managed by the provider and by the client. It is inherently part of a SLA (Service-Level Agreement).

Borenstein and Blake (2011) sees it as "…the use of fast, high-bandwidth Internet connections to deploy services that are centrally maintained, often by third parties, and thus minimize the cost and difficulty of IT administration and support for the organizations that consume those services." Emphasis on high-speed internet connection is of primary concern to applications requiring large data sets, possibly even Big Data – databases spanning terabytes processed by means of KDD (Knowledge Data Discovery), almost exclusively in a distributed fashion. Transferring such data poses a logistical issue requiring calculations as to whether it is viable to utilize cloud at all, or rather resolve to local data processing repositories, as well as security (data encryption) and throughput (bandwidth constraints) limitations in case the cloud is used.

Buyya et al (2008) define cloud as "a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified resources based on service-level agreements established through negotiation between the service provider and consumers." Here, cloud is inherently understood as a collection of virtualized computers allocated to each customer. Google's cloud-like database management service, BigTable, was specifically designed as being capable of dynamically allocate servers on-the-fly with no decrease in overall performance. (Chang et al., 2006) Virtualization and efficient use of limited computing resources is a subject of current research and a promising development venue for future applications with results directly testable on and applicable to the cloud platform.

Deep Kaur and Inderveer (2010) postulates it to be "…a distributed computing paradigm allowing virtualized applications, software, platforms, computation and storage to be rapidly provisioned, scaled and released instantly through the use of self manageable services that are delivered over the web in a pay-as-you-go manner." Costs, namely TCO (Total Cost of Ownership), constitute one of cloud's touted advantages over in-house IT infrastructure ownership along with data redundancy, higher processing power, on-demand access and others. It also bring about many challenges tied to its distributed nature, though (see Section 3). Application layer need to be modified in order to fully harness cloud's properties.

Foley (2008) provides this explanation: "Cloud computing is on-demand access to virtualized IT resources that are housed outside of your own data center, shared by others, simple to use, paid for via subscription, and accessed over the Web." Interfacing with the cloud is primarily done via a web browser with many vendors allowing access also via CLI (Command Line Interface) or dedicated clients executed on desktop stations, mobile phones or combination of both.

Foster et al. (2008) describes the cloud as "[a] large-scale distributed computing paradigm that is driven by economies of scale, in which pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage platforms, and services are delivered on demand to external customers over the Internet." Economies of scale is a microeconomic term denoting "[e]conomies that exist when the inputs are increased by some percentage and output increases by a greater percentage, causing units costs to fall." (Arnold, 2008) For a cloud vendor exploiting economies of scale, costs of setting up and maintaining new client's instance for the duration of and according to specifications in the SLA along with all expenditures calculated per customer is thus much lower than the expected return, primarily due to low virtualization costs.

Gartner (2008) explains it as “a style of computing where massively scalable IT-enabled capabilities are delivered ‘as a service’ to external customers using Internet technologies.” As with Armbrust et al., SaaS is explicitly mentioned in the definition. For most end-users, it is the cloud layer with which they interact most often when using Google Apps, Dropbox, iCloud, Netflix and other free and commercial services. GUI (Graphical User Interface) is usually simple and user-friendly without unnecessary technical details involved. Corporate clients may utilize IaaS, PaaS or customize existing SaaS solutions to suit their business needs.

Grandison et al (2010) understand it as a “virtualization of qualified resources”, with virtualization being “a method, process or system for providing services to multiple, independent logical entities that are abstractions of physical resources, such as storage, networking and computer cycles.” Authors uses the term virtualization, however, as it may also be understood in different ways, its definition is supplied as well, ensuring no ambiguous, confusing terms remain. Unfortunately, it is not integrated into the cloud definition itself but added as a separate statement, not conforming to the common “definition as a single sentence” approach.

IBM (2009) delimits it as “an all-inclusive solution in which all computing resources (hardware, software, networking, storage, and so on) are provided rapidly to users as demand dictates.” IBM’s definition was included due to it being one of the most successful providers of cloud services to corporate sector. IBM SmartCloud offers all three paradigms (IaaS, PaaS, SaaS), the company also cooperates with universities and is a member of several groups (The Open Group, Cloud Management Work Group, Cloud Standards Customer Council) whose efforts aim at codifying current developments in cloud computing for the purposes of industry-wide standardization.

McFedries (2008) recognizes cloud computing as an extension of personal computers and believes the time will come “in which not just our data but even our software resides within the cloud, and we access everything not only through our PCs but also cloud-friendly devices, such as smart phones, PDAs, computing appliances, gaming consoles, even cars.” As the definition was formulated in 2008, new cloud applications have already allowed the assumptions to materialize in practice. Current generation of smart phones and tablets is capable of storing data within the cloud and synchronize them wirelessly, gaming consoles utilize it as optional storage medium, and cloud-enabled vehicles have also been unveiled gathering and integrating telemetric data, weather information, and real-time map processing (itself supported by the cloud) into a package both visually and textually presented to the driver. Further commercial applications in this area are expected.

Schneider (2008) defines it as a “system of technologies and services that have commoditized IT to make it more readily consumable, scalable and cost-effective for everyone.” IT commoditization, a process of efficiently deploying IT-related services in a timely fashion and quality with minimal costs, is an approach apparent in both corporate and educational spheres. (Di Maio, 2012) As interest in specialized fields of study is dropping in countries such as the United Kingdom, Portugal, and Germany, technology is instead deployed ubiquitously in areas where it can be utilized without deep knowledge thanks to continuing consumerization.

Vaquero et al (2009) summarize proposed definitions and introduces a new one: “Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. The pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized SLAs.” Authors’ definition introduce load balancing, a system which distributes processing tasks across multiple servers in order to efficiently utilize them and potentially increase computation speed. However, a counterargument may be raised: if the task is executed on a single node, the unused ones may be turned off completely (provided no other tasks are active), thus potentially saving even more energy. Load balancing algorithms are of particular interest to green cloud computing initiatives struggling to minimize carbon footprint and greenhouse emissions from data centers’ cooling units.

Virtualization Journal provides 21 distinct definitions of cloud computing. (Geelan, 2009) However, they do not constitute definitions per se; each expert shared his ideas about cloud computing from the area s/he is involved in. None of them thus have a universal, overarching aspiration.

Warrior et al. (2008) defined cloud computing as “a layer which abstracts a service and applications by separating them from a physical resource.” Abstraction is achieved by virtualization during which the underlying hardware infrastructure provides a platform for many commoditized instances simulating run of fully-featured clients.

The differences led some authors to alternatively list its major features (Deep Kaur and Inderveer, 2010): geographically dispersed, virtualized, single or multiple administrative entities, heterogeneous, multi-tenant, loosely coupled resources, SLA driven, unlimited elasticity, service oriented, billed based on usage, high data throughput and low latency, industry-assisted, easy migration of virtual machines, commercial and content delivery applications.

Local data storages are centralized (usually on the premises) and therefore may be adequately protected. Moreover, when data are hosted locally, a fine-grain control of access privileges may be exerted, something not achievable in the cloud. When outsourced to the cloud, organizations relinquish the control in favor of decreased TCO, instead specifying the provisions in an SLA. Despite the TCO being lower, it is still necessary to ensure security of the infrastructure left in company's possession on the premises if not outsourced in its entirety.

2.1 Grid Computing

Cloud computing is based on previous attempts at distributed data processing, ubiquitous access, IT cost reduction and providing customers with infrastructure needed to carry out either day-to-day or specialized operations: cluster and grid computing.

The latter is a direct predecessor to cloud computing and its definition may therefore prove helpful in locating a vantage point for constructing definition for cloud computing. However, as grid computing followed the same pattern as cloud with its formal representation not agreed upon, such development is unlikely. As disparate efforts to unify the notion of grid computing suggest, though, lacking definition does not in any way hamper adoption by commercial, scientific and individual actors. It may therefore be more beneficial to focus on improving the existing technological background by means of standardization.

CERN (The European Organization for Nuclear Research, 2006) states: "The Grid is a service for sharing computer power and data storage capacity over the Internet." While the definition may be criticized for simplifying otherwise complex matter, it still provides communicable statement. CERN includes it in one of its press releases, not as a part of critical scientific discourse.

Foster, Kesselman, and Tuecke (2001) defined grid as a "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources…" Foster and Kesselman (2004) expanded it further: "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational abilities." The authors are also credited for introducing the term to broader academic and scientific attention in 1998. (Stockinger, 2007)

Foster and Iamnitchi (2003) propose this formalization: "Grids are sharing environments implemented via the deployment of persistent, standards-based service infrastructure that supports the creation of, and resource sharing within, distributed communities." An example of such system is BOINC (see Section 2.2) which allows communities dedicated to highly-specialized purposes to share and harness volunteer-based computing infrastructure.

IBM (2002) provides two definitions of grid computing: "[T]he ability, using a open set of standards and protocols, to gain access to applications and data, processing power, storage capacity and a vast array of other computing resources over the Internet. A Grid is a type of parallel and distributed system that enables sharing, selection, and aggregation of resources distributed across 'multiple' administrative domains based on their (resources) availability, capability, performance, cost, and users' quality of service requirements", and: "Grid Computing can be described as application processing, distributed across multiple locations, and interconnected through a shared network such as the Internet. [It] enables the sharing and coordination of disparate resources across both network and organizational boundaries."

The now-defunct NPACI (National Partnership for Advanced Computational Infrastructure, 2004) defined grid computing as "an infrastructure that enabled the integrated, collaborative use of high-end computers, networks, databases, and scientific instruments owned and managed by multiple organizations. Grid applications involve large amounts of data and/or computing and often require secure resource sharing across organizational boundaries, and are thus not easily handled by today's Internet and Web infrastructures."

Plaszczak and Wellner (2006) delimit it as "the technology that enables resource virtualization, on-demand provisioning, and service (resource) sharing between organizations." With cloud computing,

the notion of organizational-only resource sharing no longer applies as it is trivial for individuals to lease their own instance in the cloud.

Fundamental principles were devised sooner than grids came into prominence recently. In late 1960s, Leonard Kleinrock (UCLA, 1969) was quoted as saying: “As of now, computer networks are still in their infancy… [b]ut as they grow up and become more sophisticated, we will probably see the spread of ‘computer utilities’, which, like present electric and telephone utilities, will service individual homes and offices across the country.”

Widely accepted definitions of both cloud and grid computing don’t exist as of yet.

2.2 Cloud vs. Grid

Myerson (2009) states that to get the cloud to work, both grid and utility computing together with thin clients are needed where “[g]rid computing links disparate computers to form one large infrastructure, harnessing unused resources… [u]tility computing is paying for what you use on shared servers like you pay for a public utility (such as electricity, gas, and so on).” She further points out cloud computing evolved from grid computing.

This correlates with Foster et al (2008) who argue “that Cloud Computing not only overlaps with Grid Computing, it is indeed evolved out of Grid Computing and relies on Grid Computing as its backbone and infrastructure support.” Kim (2009) also believes “[g]rid computing is simply one type of underlying technologies for implementing cloud computing.”

Grid computing is by some researchers seen as an extension of continuously developing concepts (virtualization, parallelism, multithreading, load balancing, distributed computing). Cloud computing is understood as a next step in the progress of resource pooling, dynamic allocation, service-oriented architecture (SoA), utility, autonomic as well as notion of pervasive computing. Apart from technological factors the cloud has significant economic incentives: SaaS, PaaS and IaaS (see Section 2) all allow for reductions in information technology investments, benefiting from multi-tenancy, flexibility, scalability, lower infrastructure and maintenance costs.

If the cloud is indeed the grid’s next stage of development, many definitions should essentially provide similar accounts. Foster (2002) introduced a three point checklist describing prerequisites of a grid: coordination of resources that are not subject to centralized control; usage of standard, open, and general-purpose protocols and interfaces; delivery of non-trivial quality of service. Cloud computing qualifies to meet the third criterion while compliance with the first two is a matter of expert debate.

Grid computing often provides processing power for large-scale scientific initiatives. Among the notable grids BOINC (Berkeley Open Infrastructure for Network Computing), operating on a voluntary basis with users pooling their personal computers’ resources, stands out. Its average combined processing power surpasses peak values of the world’s fastest non-distributed supercomputers. (TOP500, 2011) BOINC virtual architecture houses numerous highly specialized projects (protein folding, detection of gravitational-wave sources, dynamic stellar streams models).

Cloud computing provides entry points to pooled resources for individual, commercial, and government entities. Neither the cloud or the grid require specialized hardware with general-purpose components nowadays being highly scalable, though the cloud’s capability to form a scientific platform certainly exist as the data centers are often under-utilized and over-provisioned. Both models thus seem to focus on different target groups, grid aimed at scientific application, cloud offering services to corporations, individuals and communities. However, cloud computing has been steadily gaining traction as a commercial alternative with broader deployment possibilities than specialized grid applications.

Database systems and big data processing in particular may benefit from the cloud’s scalability and storage capabilities when performing resource-intensive tasks, for instance OLAP (Online Analytical Processing) analyses or data mining as a part of KDD. (Pokorny, 2010)

3. Challenges of the Cloud

Since 2007 when cloud computing began to emerge, several challenges have arisen.

3.1 Uptime Guarantees

After client's decision to utilize a suitable model, uptime guarantees, factored into a Service level agreement (SLA) as a part of a Master Service Agreement (MSA) frequently negotiated in IT, are of primary concern.

Uptime is a timeframe during which a machine is operational and provides service up to specifications. Downtime is a period when a machine is non-operational due to corrective or preventive maintenance windows, system crashes or other factors. Business operations are not available during downtime if backup or mirroring is not employed. Two associated variables are mean time between failures (MTBF), the predicted time between two subsequent system failures; and mean downtime (MDT), the average time the system is non-operational.

In order to quantify service charges a standard was devised "[which] describes criteria to differentiate four classifications of site infrastructure topology based on increasing level of redundant capacity components and distribution paths." (Turner et al, 2009) Each of the four groups guarantees increased percentage-measured uptime availability than the previous one, forming a comprehensive framework.

Every tier is labeled according to the quality of service (Benson, 2006):

- Tier I: Basic Data Center (99.671 %),
- Tier II: Redundant Components (99.741 %),
- Tier III: Concurrently Maintainable (99.982 %),
- Tier IV: Fault Tolerant (99.995 %).

Tiers further vary in a set of 15 distinct parameters (staff, site availability, months to implement) to be taken into consideration.

Uptime guarantees are closely related to business operations continuity. As some corporations minimizes their physical exposure, instead opting in for transferring activities such as sales, Customer Relationship and Supply Chain Management into distributed processing environment, downtime minimization is essential. Availability of business functions to customers, regulators, suppliers, employees, and other parties must be assured by means of SLA or MSA with supplemental corporate contingency, disaster recovery, and risk management policies in place.

3.2 Security and privacy

In the cloud, data are available for the tenants irrespective of time, location or device via which they choose to retrieve it. Perceived on the outside as homogeneous, it provides SLA-determined Quality of Service (QoS), optionally with encrypted access. Clients don't have virtual access to instances outside their privacy zones, leaks are thus largely mitigated. However, the cloud leaser operating the infrastructure where data are physically stored may intercept, copy or modify them as a result of decentralized (multiple mirroring locations) and centralized (each location housing multi-tenant server infrastructure) storage architectures.

Itani et al (2009) propose Privacy as a Service (PasS) which introduces a cryptographic coprocessor, "[a] small hardware card that interfaces with a main computer… [in] the tamper-proof casing that encloses it and makes it resist physical attacks… [which] should reset the internal state of the coprocessor… upon detecting any suspicious physical activity on the coprocessor hardware." Distribution of such devices, each installed on a physical server running a Virtual Machine (VM), is to be carried out by an independent, both client- and leaser-trusted third party. The solution's cost is hypothesized to decrease once the technology adoption increases.

Kong (2010) compartmentalizes running guest VMs into trusted, untrusted, and protected parts with all of them relying on a hypervisor. Also proposed is memory isolation where all VMs and even modules controlled by the trusted part are strictly able to access only allocated memory space. It may hence be possible to prevent buffer overflows as memory segments form singular, dedicated units.

Angin et al (2010) introduce identity management using entity-centric approach. An entity sends encrypted personally identifiable information to the service provider, preventing collection of the plaintext and identity theft. This is called an active bundle, a container with a payload of sensitive data, metadata (provenance, integrity check, access control etc.), and a VM controlling the program code enclosed within. Other authentication forms include virtual tokens (username, full name, address),

privacy-preserving anonymous credentials together with trusted electronic identity services (OpenID). The goal is to provide fast and reliable challenge-response methods without associating the data with individual users.

Applications running within the cloud pass data through multiple layers, each constituting a distinct point of failure and a vulnerability. While user can interact with the assigned VM instance remotely, she doesn't exert any control over hardware on which it is executed. If data are thus not encrypted in transit, unauthorized third party may intercept and divert them. Encryption, however, inadvertently introduces computational overhead which may become a bottleneck for organizations requiring high-throughput transfers.

Security and privacy remains a downside for many companies, with 74.6 % out of 244 recently surveyed managers citing them as their main source of concern. (Zhou et al, 2010) For sensitive, competitive applications such as defense, aerospace, and financial brokerage industries the cloud is currently seen as unacceptable due to possible leaks of valuable data to competitors. (Chakraborty et al, 2010)

3.3 Legislature

Compliance with existing laws is obligatory for the cloud leasers operating the data centers. In the USA, the two main acts are FISMA (Federal Information Security Management Act of 2002), and SOX (Sarbanes–Oxley Act). In the European Union, The Data Protection Directive is in place.

FISMA (2002) defines information security as "protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide… integrity… confidentiality… [and] availability." The act serves as a framework for any organization which intends to provide information-processing systems to the government bodies, including cloud computing. In July 2010, Google's cloud service was the first to pass the security certification and accreditation scrutiny.

SOX (2002) was enacted as a response to lacking, fraudulent financial audits which led to high-profile cases of corporate stock devaluation and insider trading controversies. It introduced transparent, efficient system of private sector financial controls and accountable real-time corporate reporting, resulting in a boost of investors' confidence. All privately-held companies in the USA have to abide by the law.

The European Union (EU, 1995) Data Protection Directive regulates the way companies within the European Union have to handle and process personal data which comprises "any information relating to an identified or identifiable natural person ('data subject')" with further delimitation of a natural person and their discernible properties. It also established rules on cross-border transfer of such data. Major overhaul of the directive is planned in 2012.

China addresses cloud computing in several articles of law. (Murphy, 2010) Japanese government made significant investments into cloud computing research, particularly the environment-friendly variant called "green cloud computing". (Myoken, 2009)

3.4 Economics, Penetration rates, Lags

Economic concerns in case the cloud computing provider ceases to function were raised. Data retrieval, financial costs of transferring company's online presence to another site, and inaccessibility periods are all realistic estimates. To minimize economic impacts it is recommended to check the provider's financial statements, back up data located in the cloud regularly, have excess capacities ready at different locations, and prepare for application portability. (Scheier, 2009)

A recent study aimed at small-, medium-, and large-size companies researched whether purchasing local data storage capacities (hard disk drives, HDDs) is economically justifiable in comparison with storage leasing offered by cloud computing. It took into account factors such as pow

er consumption or progressive mechanical parts' degradation. (Walker et al, 2010) Apart from the results, the study concluded the latency in the cloud is not zero, therefore different storage locations are not homogeneous in terms of service quality which may incur additional costs.

Financial savings stand out among the cloud's features as a result of infrastructure outsourcing, lower maintenance and staffing expenditures, and decreased costs of associated services. Cloud service

providers need to make a transition to the new business model where products are not marketed as goods, but services. (Olsen, 2006)

Another study also showed (GfK, 2011) that cloud computing achieved higher penetration rate on the emerging markets of Brazil, China, and India while established markets in Germany, United States and United Kingdom were more reserved in its adoption. The study also stipulated cost reduction to be a reservation rather than benefit. Out of 1 800 companies, 42 % cited costs of services as their top issue along with expenditures of migrating into the cloud. One reason for such a result may be lower purchasing power parity (PPP) on emerging markets along with undifferentiated pricing of major cloud computing vendors.

As mentioned previously, since latency in the cloud is not zero, time-based economics must be considered as well. Total lag consists of a delay introduced server-side compounded with a network lag, a value considered negligible when the data storage is placed in-house. When multiple repeated requests are sent, total lag may further increase depending on connection's properties (i.e., peak and off-peak transfers) and data center's as well as leased instance's utilization. Cloud solutions are therefore not recommended to industries requiring extremely low latency access, such as HFC (High-Frequency Trading).

To address the issue, a commercial provisioning system has been introduced (Han et al., 2012) which reduces server-side lag utilizing high-speed hardware components. Authors suggest a suitable pricing model as the service is provided on an opt-in basis.

4. Hypothesized Future Changes

Companies, governments, and individuals are gradually starting to accept cloud computing as a virtualized, on-demand, economically feasible resource provisioning model. The trend, however, will make isolated changes to the underlying technology increasingly difficult as efforts call for standardization. It would be advisable for major cloud providers to consider either implementing desired alterations using remote update scheme which does not interfere with functionality and data center uptimes.

One of the cloud's features is to appear as a single space access to which is possible from any device supporting respective protocols. As the latency is not zero, though, it may be beneficial to provide users with an option to choose a data center to interact with during sessions as compared to automatic, best-match algorithmic selection. Standard users should still have the ability to let the protocol determine optimum access vector. Conversely, power users would choose it individually which could resolve problems when selecting a data center based on IP address detection while behind a proxy server in a geographically separate location. Implementation of the functionality is not uniform among different cloud vendors. With continuing penetration of high-speed computer networks, though, only those with limited network bandwidth and transfer rates may benefit from inclusion of such feature outside the expert zone.

While concise definition may benefit cloud computing, it is this author's belief that in spite of a concise delimitation, the paradigm of distributed data storage and ubiquitous access is in itself a powerful feature for both commercial and societal adoption. Several major vendors (Amazon, IBM, Microsoft) usually adopt their own suitable definitions but as the field is atomized, they will very likely not be willing to modify them. If, however, such undertaking will be attempted, cloud's properties should be agreed upon by an impartial authority or panel following discussions with experts, leading cloud providers and leasers as well as academia.

Recently proposed Cloud@Home "…envision[s]a volunteer cloud infrastructure built on resources voluntarily shared (for free or for a fee) by their owners or administrators, following a volunteer computing approach and provided to users through a cloud-service interface." (Distefano & Puliafito, 2012) Authors propose using a credit system inspired by the BOINC (see Section 2.2). In case the proposal is adopted, cloud data centers may become more widespread and economies of scale will allow vendors to offer flexible pricing options, further popularizing the use of cloud among general population.

Expanding the number of endpoints and virtualized capacities is crucial. For instance, a driver may select his destination in a dashboard computer which sends the data into the cloud data center using a Wireless Network Interface Card (WNIC). The vehicle in turn receives information containing current road blocks, detours, and local points of interest. Unlike Global Positioning System (GPS) the data, mixing coordinates with user-submitted content would be downloaded, presented, and updated in real

time. The concept of community clouds as publicly available versions of business clouds gains traction as they allow localized groups to benefit from their advantages. Universities, branch offices and other geographically dislocated entities in particular could provision such access. Nevertheless, the service should be monetized in a way that ensures competitiveness compared to "standard" cloud.

Smart and mobile phones already facilitate interaction regardless of geographic boundaries. Access scheme for mobile units (GPS, phones, notebooks, tablets) with the added benefit of saving battery power by offloading application processing away from the device into the cloud has been proposed as well, albeit not with uniform results so far. (Kumar & Lu, 2012) The most probable next point of interest will be entertainment industry. Classic model where individuals chose from predefined selections will probably transform into one in which users will select whatever content they wish on a per-per-unit basis, delivered (streamed) directly from the cloud to the device for immediate consumption. This will require industry-wide collaboration to maintain heightened bandwidth utilization as efficient as possible. Entertainment, cloud and networking are therefore intrinsically linked.

Cloud is currently based on disjointedly operating data centers. The idea of a unifying platform not unlike the Internet has already been proposed. Such model, titled Intercloud, assures interoperability in situations where data in a cloud explicitly requests data located in another one. (Bernstein et al, 2009) When processing, storing or hosting large amount of records, a single cloud's center may also become overloaded and additional data ignored. In Intercloud, the excess is transferred to another center using the pay-per-use model. Intercloud may be an option for data-intensive scientific operations (cloud–grid convergence as mentioned previously) and for organization in need of processing power surpassing a single cloud.

Nowadays, cloud computing may be perceived as homogeneous but different vendors offer varying services as demonstrated by Li et al. (2011) In this author's opinion, as cloud will permeate more aspects of commerce and science, it is inevitable some vendors will lag behind competition and will be forced to terminate their business operations altogether. This may, however, cause data replicas to become instantly unavailable across all locations (see Section 3.4). The dilemma of forgoing the predicament by keeping local copies or alternately leasing redundant cloud instances into which data will be mirrored (thus increasing infrastructure TCO, delays, bandwidth usage etc.) will probably be solved by a binding cloud data backup standard. Some concerns were raised regarding inability to forensically preserve data within the cloud.

Cloud computing and Big Data seems to be a promising development as it enables companies to store large volumes of structured as well as unstructured data for immediate access and processing. Specialized scientific applications also appeared as some vendors run cloud instances consisting solely of parallelized GPU (Graphics Processing Unit) hardware components used for high-performance computing and intensive mathematical operations.

Novel ways to exploit cloud infrastructure will no doubt appear even in areas currently outside out its scope; vehicles and transportation are just one example, however, healthcare, public administration, mobile applications and others will all likely be involved, too.

The question of what will be the next development stage after the cloud has not been satisfactorily answered yet. It will very likely be based on the same principles, namely sharing of processing power, multi-tenancy as well as further aggregation of technology into massive centers with on-demand access. On the other hand, initiatives such as Cloud@Home may shift the trend towards community clouds and localized resource pools.

5. Conclusion

Cloud architecture propagates to many diverse areas. Governments, municipal administrations and universities may benefit from it when employees are scattered in multiple locations. However, cloud providers have to assure compliance with legislation and requirements of the end-users.

It is also a field of intense scientific and commercial interest: algorithms dealing with load balancing, data replication and redundancy, parallel processing, virtualization, memory management, networking as well as security, authentication, large-scale data encryption and information retrieval are presented and benchmarked; flexible pricing models introduced; novel cloud-enabled applications integrated with existing platforms and devices (entertainment, applications for mobile access); and fields of research augmented via use of such affordable provisioning paradigm.

Future of the cloud is ever-changing as it came into prominence relatively recently. Research, testing, and academic—industry cooperation may shape it to best serve the needs of modern society and science.

6. Bibliography

Angin, P. et al, 2010: An Entity-centric Approach for Privacy and Identity Management in Cloud Computing, *29th IEEE International Symposium on Reliable Distributed Systems*, November 2, 2010, Delhi, India

Armburst, M. et al., 2009: Above the Clouds: A Berkeley View of Cloud Computing", [online], *University of California*, Berkeley, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

Armhein, D., and Quint, S., 2009: Cloud computing for the enterprise: Part 1: Capturing the cloud", [online], *IBM*, Armonk, New York, http://www.ibm.com/developerworks/websphere/techjournal/0904_amrhein/0904_amrhein.html

Arnold, R. A., 2008: *Economics,* 9th edition, Mason, South-Western Cengage Learning

Benson. T., 2006: TIA-942: Data Center Standards Overview, *ADC Telecommunications*, Minneapolis, Minnesota, http://www.adc.com/Attachment/1270711929361/102264AE.pdf

Bernestein, D. et al., 2009: Blueprint for the Intercloud - Protocols and Formats for Cloud Computing *Interoperability, 4th International Conference on Internet and Web Applications and Services (ICIW '09)*, May 24-28, 2009, Venice/Mestre, Italy

Borenstein, N., and Blake, J., 2011: Cloud Computing Standards: Where's the Beef?, *IEEE Internet Computing*, Vol. 15, No. 3, pp. 74-78

Buyya, R., Yeo, C.S., and Venugopal, S., 2008: Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities, *10th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid 2010)*, May 17-20, 2010, Melbourne, Australia

Chang, Fay, Dean, Jeffrey, Ghemawat, Sanjay, Hsieh, Wilson C., Wallach, Deborah A et al., 2006: Bigtable: A Distributed Storage System for Structured Data, *7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06)*, November 6-8, 2006, Seattle, Washington.

CERN, 2006: The Grid, [online], *CERN*, Geneva, Switzerland, http://cdsweb.cern.ch/record/976156/files/it-brochure-2006-002.pdf

Chakraborty, R., Ramireddy, S., Raghu, T.S., and Raghav Rao, H., 2010: The Information Assurance Practices of Cloud Computing Vendors, *IT Professional*, Vol. 12, No. 4, pp. 29-37

Deep Kaur, P., and Inderveer, C., 2010: Unfolding the Distributed Computing Paradigm, *International Conference on Advances in Computer Engineering (ACE 2010)*, June 21-22, 2010, Bangalore, India

Di Maio, Andrea, 2012: IT Commoditization Hits IT Education in Europe, [online] *Gartner,* http://blogs.gartner.com/andrea_dimaio/2012/03/27/it-commoditization-hits-it-education-in-europe/

Distefano, S., and Puliafito, A., 2012: Cloud@Home: Toward a Volunteer Cloud, *IT Professional*, Vol. 14, No. 1, pp. 27-31

EU, 1995: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, [online], *EU*, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:NOT.

FISMA, 2002: Federal Information Security Management Act of 2002, [online], http://csrc.nist.gov/drivers/documents/FISMA-final.pdf

Foley, J., 2008: A Definition of Cloud Computing, [online], *InformationWeek*, http://www.informationweek.com/cloud-computing/blog/archives/2008/09/a_definition_of.html.

Foster, I., 2002: What is the Grid? A Three Point Checklist, [online], *Argonne National Laboratory & University of Chicago*, http://dlib.cs.odu.edu/WhatIsTheGrid.pdf

Foster I., and Iamnitchi, A., 2003: On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing, *2nd International Workshop on Peer-to-Peer Systems (IPTPS '03)*, February 21-23, 2003, Berkeley, California

Foster, I., and Kesselman, C., 2004: *The Grid 2: Blueprint for a New Computing Infrastructure*, 2nd edition, Amsterdam, Elsevier

Foster, I., Kesselman, C., and Tuecke, S., 2001: The Anatomy of the Grid: Enabling Scalable Virtual Organizations, *International Journal of Supercomputer Applications*, Vol. 15, No. 3, Fall, pp. 200-222

Foster, I., Zhao, Y., Raicu, I., and Lu, S., 2008: Cloud Computing and Grid Computing 360-Degree Compared, *2008 Grid Computing Environment Workshop (GCE '08)*, November 12-16, 2008, Austin, Texas

Geelan, J., 2009: Twenty-One Experts Define Cloud Computing, [online], *SYS-CON Media Inc.*, http://virtualization.sys-con.com/node/612375

GfK, 2011: Cloud computing popular in emerging markets, [online], *GfK*, Nuremberg, Germany, http://www.gfk.com/group/press_information/press_releases/008354/index.en.html

Grandison, T., Maximilien, E.M., Thorpe, S., and Alba, A., 2010: Towards a Formal Definition of a Computing Cloud, *6th World Congress on Services (SERVICES-1)*, July 5-10, 2010, Miami, Florida

Han, H., Lee, Y. C., Shin, W., Jung, H., Yeom H. Y. et al., 2012: Cashing in on the Cache in the Cloud, *IEEE Transactions on Parallel and Distributed Systems* (to appear)

IBM, 2002: IBM Solutions Grid for Business Partners: Helping IBM Business Partners to Grid-enable applications for the next phase of e-business on demand, [online], *IBM*, Armonk, New York, http://joung.im.ntu.edu.tw/teaching/distributed_systems/documents/IBM_grid_wp.pdf

Itani. W., Kayssi, A., and Chehab, 2009: A. Privacy as a Service: Privacy-Aware Data Storage and Processing in the Cloud Computing Architectures, *8th International Conference on Dependable, Autonomic and Secure Computing (DASC 2009)*, December 12-14, 2009, Chengdu, China.

Kim, W., 2009: Cloud Computing: Today and Tomorrow, [online], *Sungkyunkwan University*, Suwon, South Korea, http://www.jot.fm/issues/issue_2009_01/column4/

Kong, J., 2010: A practical approach to improve the data privacy on virtual machines, *10th IEEE International Conference on Computer and Information Technology (CIT 2010)*, June 29-July 1, 2010, Bradford, United Kingdom

Kumar, K., and Lu, Y., 2012: Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?, *Computer*, Vol. 45, No. 4, (to appear)

Li, A., Yang, X., Kandula, S., and Zhang, M., 2011: Comparing Public-Cloud Providers, *IEEE Internet Computing*, Vol. 15. No. 2, pp. 50—53.

McFedries, P., 2008: The Cloud is the Computer, [online], *IEEE Spectrum*, http://spectrum.ieee.org/computing/hardware/the-cloud-is-the-computer

Mell, P., and Grance, T., 2011: The NIST Definition of Cloud Computing (Draft). Recommendations of the National Institute for Standards and Technology, [online], *NIST*, http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf

Murphy, M., 2010: Cloud Computing and Chinese Law, [online], *Gerson Lehrman Group*, New York, New York, http://www.glgroup.com/News/Cloud-Computing-and-Chinese-Law-51632.html

Myerson, J., 2009: Cloud computing versus grid computing: Service types, similarities and differences, and things to consider, [online], *IBM*, Armonk, New York, http://www.ibm.com/developerworks/web/library/wa-cloudgrid/

Myoken, Y., 2009: Cloud Computing in Japan [online], *British Embassy*, Tokyo, Japan, http://ukinjapan.fco.gov.uk/resources/en/pdf/5606907/5633632/cloud-computing-japan

NPACI, 2004: National Partnership for Advanced Computational Infrastructure: Archives, [online], *NPACI*, San Diego, California, http://npacigrid.npaci.edu/terminology.html

Plaszczak, P., and Wellner, R., 2006: *Grid Computing: The Savvy Manager's Guide*, Elsevier, Amsterdam

Olsen, E.R., 2006: Transitioning to Software as a Service: Realigning Software Engineering Practices with the New Business Model, *IEEE International Conference on Service Operations and Logistics, and Informatics, (SOLI '06)*, June 21-23, 2006, Shanghai, China

Plummer, D.C., Bittman, T.J., Cearley, D.W., and Smith, D.M., 2008: Cloud Computing: Defining and Describing an Emerging Phenomenon, *Gartner*, Stamford, Connecticut

Pokorny, J., 2010: Databases in the 3rd Millennium: Trends and Research Directions, [online], *Journal of Systems Integration*, Vol. 1, No. 1-2, pp. 3-15, http://www.si-journal.org/index.php/JSI/article/view/25

Sarbanes, P., and Oxley, M.G., 2002: Sarbanex-Oxley Act of 2002, [online], http://www.gpo.gov/fdsys/pkg/PLAW-107publ204/content-detail.html

Scheier, R.L., 2009: What to do if your cloud provider disappears, [online], *InfoWorld*, San Francisco, California, http://www.infoworld.com/d/cloud-computing/what-do-if-your-cloud-provider-disappears-508

Schneider, S., 2008: Cloud Computing Definition, [online], *Hyperic*, Palo Alto, California, http://blog.hyperic.com/cloud-computing-definition/

Stockinger, H., 2007: Defining the grid: a snapshot of the current view, *The Journal of Supercomputing*, Vol. 42, No. 1, pp. 3-17

TOP500, 2011: TOP500 List – November 2011 (1-100), [online], *TOP500 Supercomputing Sites*, http://www.top500.org/list/2011/11/100

Turner, W. P., Seader, J. H., and Renaud, W. E., 2009: Data Center Site Infrastructure Tier Standard: Topology, *Uptime Institute*, Santa Fe, New Mexico, http://professionalservices.uptimeinstitute.com/UIPS_PDF/TierStandard.pdf

UCLA, 1969: UCLA to be the First Station in Nationwide Computer Network, [online], *UCLA*, Los Angeles, California, http://www.lk.cs.ucla.edu/LK/Bib/REPORT/press.html

Vaquero, L.M., Rodero-Merino, L., Caceres, J. and Linder, M., 2009: A break in the clouds: towards a cloud definition, *ACM SIGCOMM Computer Communication Review*, Vol. 39, No. 1, January, pp. 50-55

Walker, E., Brisken, W., and Romney, J., 2010: To Lease or not to Lease from Storage Clouds, *IEEE Computer*, Vol. 43, No. 4, April, pp. 44-50

Warrior, P., O'Reilly, T., and Robinson, S., 2008: Cloud Computing: The Future Web, [online], *Web 2.0 Summit 2008*, November 5-7, 2008, San Francisco, California, http://www.web2summit.com/web2008/public/schedule/detail/6238

Zhou, M. et al., 2010: Security and Privacy in Cloud Computing: A Survey, *Sixth International Conference on Semantics, Knowledge and Grids*, November 1-3, 2010, Ningbo, China

JEL Classification M15, O30

This article should be cited as:

Sarga, L., 2012: Cloud Computing: An Overview *Journal of Systems Integration* 3 (4), pp. 3 - 14. [Online] Available at: http://www.si-journal.org. ISSN: 1804-2724