Network Implications of Cloud Computing

Joe Weinman
Hewlett-Packard
Basking Ridge, NJ
joeweinman@{hp, hotmail, gmail}.com

***Abstract*—Cloud Computing is often described as "resources accessed via a browser over the Internet." However, this definition has become increasingly insufficient to characterize the breadth of applications and use cases for the cloud, and the networks that must support them.**

A broadening range of endpoints are accessing the cloud: browser-free device apps, multimedia endpoints such as video and game consoles, sensor networks, servers, and storage.

The wireline and wireless network requirements—e.g., jitter, latency, packet loss, protocol support—for these uses vary, and imply that a variety of network capabilities are sometimes necessary: e.g., MPLS for quality of service via class of service to support interactive high definition video in the cloud; optical transport for native protocols such as Fibre Channel for data integration in hybrid cloud scenarios; route control for country compliance issues. Also, distributed topologies and optimized routing are required due to application latency constraints.

Moreover, wireless sensor networks and hybrid cloud scenarios such as cloudbursting that require a variety of complex distributed data approaches are driving new transport requirements: guaranteed bandwidth, dynamic bandwidth on demand, and usage-sensitive pricing for fine-grained quantities and duration of bandwidth.

Cloud Computing, either as an integrated service or in support of pure-play customers must drive service providers' international telecommunications infrastructure evolution as well as BSS/OSS.

Keywords-Networks; Cloud Computing; QoS; Bandwidth; Performance; NP-Complete; OpenFlow; Cloudonomics

I. INTRODUCTION

Cloud computing represents the latest step in a half century of computing technology evolution, beginning with the mainframe and then client-server, PC, Internet, and mobility. The term "cloud computing" was coined by Ram Chellappa in 1997, who observed that "computing has evolved from a mainframe-based structure to a network-based architecture" [1], and, even though a recent survey of definitions proposed a new definition that surprisingly omitted any explicit mention of networking [2], generally, the "cloud" in cloud computing is understood as deriving somehow from the traditional use of an abstract cloud to represent a network—but what network?

A number of innovators in this space have been the so-called "over-the-top" players, who offer services and resources over the Internet, leading to a widespread view by academics [3] and analysts [4,5] that "cloud computing" may be defined as "services delivered over the Internet," often with the added proviso that they be accessed "from a Web browser" [6]. The Internet certainly has many wonderful attributes, but is not the only approach to networking. Similarly, while the Internet is certainly important for cloud computing, it is not the only network relevant to the cloud: a more general definition is offered via Axiomatic Cloud Theory [7], or by the National Institute of Standards and Technology: "Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources…that can be rapidly provisioned and released with minimal management effort or service provider interaction" [8].

More than a matter of semantics, emerging cloud-compatible endpoints, application profiles, and usage scenarios have important implications for network services, architecture, and interfaces.

Focusing first on endpoints, it isn't just browsers that are accessing cloud services, but also apps, sensors, voice and video terminals, and, at the extreme, distributed high-performance servers and storage in hybrid, intercloud, and large-scale globally dispersed architectures. These may drive network requirements not traditionally seen in a cloud context, for example, support for protocols such as Fibre Channel. Distributed networks of millions, billions, or even trillions of sensors, actuators, and other intelligent devices such as, say, connected photoframes and pill bottles will require new types of signaling, charging, and billing for wireless networks.

Application profiles are shifting in virtually every dimension. Rather than just text and graphics delivered via HTML pages, they include voice, video, and other real-time multimedia applications. These often place stringent requirements on end-to-end latency, packet loss, and jitter. Online gaming and new scenarios such as augmented reality and interactive entertainment where high resolution video is composited with rich, dynamically generated 3D graphics will require increases in many performance attributes.

And, use cases for emerging distributed architectures will also drive new network requirements. As an example, the "cloudburst" scenario addresses spikes in demand too large to be handled by owned or dedicated resources in a data center by "spilling over" workloads into a public cloud provider with "on-demand" resources. While rapidly spinning up new virtual servers is known in the art, and so is adding them to a distributed cluster, there is a substantial challenge in data management. Today, small amounts of data may be sent over the Internet to public cloud providers within reasonable timeframes, but it is suggested to express-mail hard drives when data quantities are non-trivial [9]. Focusing on

accelerating virtual server provisioning by a few seconds when the data that the virtual server needs may not arrive for a day or so is ironic, to say the least. In a perfect world, an infinite amount of bandwidth between the enterprise data center and the cloud would be available for an infinitesimal period and affordably priced. Short of that, protocol and WAN optimization together with more flexible networks with bandwidth on demand would be useful.

In short, for the range of applications and use cases for the cloud to fulfill their promise will require a range of enhancements to the "best efforts" Internet approach in use by many over-the-top providers today.

II. Network Requirements for The Cloud

The major benefit of the Internet is ubiquitous access via wireline and mobile devices to cloud resources such as webmail, search, social network tools, and software and platform services.

However, there are a number of issues that limit the applicability of the Internet. These include limited or no:

- class of service markings, end-to-end quality of service
- control over routing, and therefore
- control over latency, and also
- fine-grained, application-aware control of cross-border transport due to privacy laws and cyber-attack threats
- guarantees, e.g., bandwidth, packet loss, throughput
- control over dynamic bandwidth allocation required for data-intensive cloud scenarios

Consequently, more robust network mechanisms such as MPLS and VPLS as well as existing and emerging network transport protocols are required to support cloud scenarios such as data center bridging, intracloud connectivity, and inter-cloud federation.

Moreover, the industry will need to move to pay-per-use, dynamic networks where possible to improve the economic benefits of cloud scenarios.

III. The Evolution of Software as a Service

The current model of "software as a service" may be described as either loading web pages comprising largely text and graphics, or perhaps a mobile application that acquires updates asynchronously, e.g., a news reader or email reader on a tablet acquiring new headlines or emails as a background process. However, this represents the past, not the future.

A. Increased Interactivity

Today's applications are increasingly interactive and thus increasingly require low latency. For example, AJAX (Asynchronous Javascript and XML—eXtensible Markup Language or JSON—JavaScript Object Notation)) applications process user events such as mousemoves and keydowns, trigger Javascript, request partial data from the server, and update a web page dynamically, *without* a full page load. The quantity of data retrieved interactively is increasing. For example, web search originally required a full page load. Then, search *terms* were returned with each key press. Then, search *results* were returned. Now, web pages including images are being loaded progressively as a URL is typed.

Moreover, other applications, such as remote virtual desktops and telemedicine also require low latency due to their highly interactive nature. Finally, one can view real-time applications, ranging from high-frequency proprietary and algorithmic trading to remote control via closed-loop sensor-actuator feedback circuits as representing the ultimate in interactivity.

Human performance studies show that 200 to 250 milliseconds is acceptable for multimedia conferencing and collaboration applications. However, interactive tasks such as keystrokes and mousedowns must be responded to within about 150 milliseconds [10], and emerging online games require even lower latencies.

There are also empirical results showing the importance of low latency not just in terms of user experience, but in terms of revenue. Lower latency directly correlates with increased revenue [11].

B. Increased Bandwidth Intensity

Moreover, many industry studies show that bandwidth demands are increasing. For example, the Cisco Visual Network Index Forecast (2010-2015) [12], released June 1, 2011, projects that the total amount of global Internet traffic will quadruple in four years.

There is no end in sight to this traffic increase. Much of the world is moving to 1080p HD video, but waiting in the wings are 4K and 8K Digital Cinema standards, as well as Quad HD and Ultra-HD (a 4x4 matrix of 1080p HD screens) with a nominal resolution of 32 Megapixels. Moreover, increased frame rates for smooth motion and 3D will further increase the bandwidth beyond today's 5-7 Megabits per second for a 30 frame-per-second 1080p stream for "talking heads."

C. More Endpoints

Not only is there more interactive, higher bandwidth required per endpoint, there are also an increasing number of endpoints. In the microcosm, each person has more devices, and each household has an increasing number of connected endpoints. These endpoints will increasingly comprise multiple stream components, e.g., multistream cards for personal video recorders, and multiple monitors for immersive video conferencing or ambient video, including permanent connected "video tunnels" and immersive videoconferencing, also known as telepresence or visual collaboration.

Sensor networks, empowered by the IPv6 address space, will multiply the number of endpoints. While some are low bandwidth, e.g., temperature and humidity sensors, others will be high bandwidth, for example, video surveillance for security, merchandising, digital signage audience metrics, or traffic flow optimization. One city alone—Chongqing, China—plans to deploy over 500,000 video surveillance cameras as part of an initiative called "Peaceful Chongqing." [13]. Add in Wi-Fi light-bulbs, RFID tags, and the like, and

there will be numerous endpoint devices, with or without "users," connected to the cloud.

All these together imply a worsening mix of more demanding response times and lower latencies for more endpoints in more places with more bandwidth. This not only implies greater demands on the network, but their interaction may cause emergent effects such as congestion and spikier network utilization.

IV. The Evolution of Infrastructure as a Service

Beyond user endpoints, machine-to-machine networking in a world of "big data" will drive substantial bandwidth. Rather than just sensors, machine-to-machine scenarios include distributed databases, sharding, remote data access, and a variety of dynamic data migration scenarios.

Hybrid clouds consist of an enterprise data center "private cloud" coupled to an on-demand "public" or "virtual private" cloud with on-demand, pay-per-use capacity. They have been shown to be economically optimum under most expected application profiles, even when the unit cost of cloud resources per unit time is higher than the unit cost of "private," i.e., dedicated or owned, resources [14].

There are numerous scenarios involving hybrid clouds, such as a "front-end / back-end" architecture where a distributed set of cloud resources is used to run highly interactive web servers, app servers, or content servers, and a back-end is used for more throughput intensive tasks such as database servers or legacy systems. In this case, reliable interconnection between the front-end and back-end typically requires at least an MPLS VPN.

Even more demanding however, is the emerging "cloudbursting" scenario, where an environment that may exist only within an enterprise data center runs out of resources, for example, due to a demand spike, and dynamically enlists additional capacity in the cloud. However, the cloud resources must cooperate in some way with the data center resources, driving innovative data networking requirements.

There are a variety of scenarios for cloudbursting, in varying degrees of use and realization, with differing network requirements [15].

A. Decoupled Parallelizable Tasks

In the simplest scenario, the private data center resources do not need to communicate much with the cloud ones. For example, the application may not use or generate much data, and that which is generated may not need to be synchronized across locations. Or, the data paths may be primarily between the users and the resources. For example, a user may upload one or more photos to an online editing program, and after conducting a set of operations just receive the edited photo or a video montage based on the photos.

B. Remote Data Access

Often however, there may be a large data set resident at the enterprise data center, for example a customer data warehouse, seismic analysis data, equity trading history, or the like. For remote access at the application layer, secure low latency connections are required. Otherwise, lower layer protocols may be necessary, for example, native protocol access from the cloud to the enterprise data center storage.

C. Dynamic Data Partitioning

On cloudbursting, a portion of a large data set may be migrated over to the cloud from the enterprise data center. For example, customers whose names begin with A-Q might remain, whereas those with names R-Z might migrate over. As load builds and ebbs, more "letters" might move over or back. This might be done over fixed bandwidth connections, but would be best served by dynamic bandwidth.

D. Full Data Migration

The ultimate requirement would be to migrate or replicate an entire data warehouse to the cloud. Given that cloud virtual servers can be spun up in a matter of seconds, this drives the, to put it mildly, rather challenging scenario of near infinite bandwidth for a nearly infinitesimal amount of time. Failing this, the most bandwidth that can be provided on an end-to-end pay-per-use basis would be desirable.

E. Bidirectional Migration with Synchronous Mirroring

Falling short of this, a hybrid scenario would leverage investment in a business continuity copy, or mirror, of the data to pre-position such a large data set. The challenge is then that updates provided to either copy must be synchronously mirrored to the other location on a bidirectional basis, with the appropriate level of either fine-grained locking or eventual consistency required by the application.

V. The Need for Flexible Control

Flexibility, agility, and pay-per-use, coupled with intelligent automation—for example, autoscaling and VM migration—are key attributes of the cloud. However, the traditional focus of this flexibility has been within the enterprise private cloud or public cloud service provider data center. It is time, however, for the same degree of controlled flexibility to come to the network.

Traditionally data networks have operated at either the extreme of inflexibility, e.g., private lines with fixed point-to-point routing and bandwidth, or what might be characterized as wide-open chaos, with packets "seeking" any way to get to their destination, via protocols such as OSPF.

Emerging approaches can enable a happy medium. For example, OpenFlow [16] enables policy-based routing by decoupling the hardware used for line-speed packet data flows—which may be traditional switches or Network Field-Programmable Gate Array cards—from a controller module that can edit entries in the flow tables in accordance with policies. In a base operations mode, flows are handled as may be expected: either forwarded to a port or dropped. However, at the core of the OpenFlow approach is the notion of forwarding some flows, either new flows or selected flows, to an OpenFlow controller. In the case of a new flow, the controller may use intelligent policies to allow or deny the flow, and adjust flow-table entries according to a policy. For

flows which continue to be forwarded to the controller, additional actions may be performed, such as deep packet inspection, statistical analysis, or even application layer functionality.

The OpenFlow approach may be viewed as a generalization of AT&T's Intelligent Route Service Control Point (IRSCP), [17] which uses a control plane acting as a route reflector to provide fine-grained control within IP networks. OpenFlow can support IP, but is not restricted to it, as it is implemented at Layer 2 and it has been proposed for use with other protocols, including experimental ones.

Rather than viewing an approach such as OpenFlow as merely interesting from a research perspective, it is best viewed as an enabler of a fine-grained, high-performance, intelligent network fabric that can be coupled with flexible cloud resources. In fact, such a fabric is emerging. The ability for OpenFlow to scale to wide-area networks of continental or even global scale coupled with flexible compute nodes under integrated control has been demonstrated recently using the GENI (Global Environment for Network Innovations) [18] test bed sponsored by the U.S. National Science Foundation and the PlanetLab "planetary-scale" computing service platform [19]. A team comprising Stanford and BBN demonstrated [20] an Integrated Control Framework that manages both wide-area network and compute resources.

In a similar spirit, the PHOSPHORUS-LUCIFER (Lambda User Controlled Infrastructure for European Research) [21] has detailed interfaces for Grid Network Services, such as Grid-GMPLS, helping to tie together grid (distributed compute and storage) with network resources in a single control plane and framework.

Optimally managing such a distributed infrastructure, unfortunately, is NP-Complete [22], as the present author has shown [23]. Specifically, the simple problem of allocating varied indivisible processing (or storage) demand from multiple customers to multiple resources over a network, which may be termed the "Cloud Computing Demand Satisfiability" problem, is NP-Complete. This was demonstrated via a transformation from BOOLEAN 3-SATISFIABILITY, where constraints in allocating cloud resources are used to model truth-setting components, satisfaction-testing components, and communications links between these components.

If we assume that each workload is of equal size or is divisible, then the problem is essentially the *maximum cardinality matching problem* for bipartite graphs, also known as the *marriage problem*, which is known to be solvable in polynomial time, e.g., using the $O(|E| \times \sqrt{|V|})$ Hopcroft-Karp algorithm [24]. However, dividing up workloads can incur issues as well, e.g., latency, performance, and data transport costs associated with inter-processor communication and data migration. In fact, placing distributed service nodes to meet a latency (and thus distance) constraint is also computationally intractable [25]. The implication is that assuming P≠NP, the issues of an integrated compute, storage, and network cloud fabric entail more than creating inherent infrastructure flexibility, an API, and the integrated management frameworks to intelligently control that flexibility, but also require the development of algorithms and heuristics for "good-enough," if suboptimal, solutions to allocate workloads in real time.

VI. Consolidation and Dispersion Trade-Offs

Given that the cloud is fundamentally a distributed processing architecture, should processing nodes be consolidated or dispersed?

There are a number of potential benefits to consolidation: economies of scale, ability to handle the largest monolithic workloads, and the ability to achieve higher utilization through greater workload smoothing via enhanced statistical multiplexing [26].

There are also benefits to dispersion: less economic value at risk in the event of a "smoking hole" disaster, and proximity to customers or end-users for latency-sensitive interactive applications using chatty protocols. Perhaps the two chattiest and/or most latency-sensitive applications are financial market data, where sub-millisecond response times are necessary, and synchronous mirroring protocols for storage, where I/O operations per second are limited by the time for one or two round trips per written block.

For highly interactive applications potentially leveraging parallel processing as well as dispersion, the total response time T may be approximated as [27]:

$$T = F + \frac{N}{\sqrt{n}} + \frac{P}{p}$$

where F is a constant for the endpoint and serial portion of the workload, N is the round-trip latency with one service node, n is the number of well-distributed service nodes, P is the processing time for the parallel portion of the workload, and p is the number of processors. The minimum of this function, i.e., the optimum latency, is reached when the number of nodes is $n = \sqrt[3]{\left(\frac{QN}{2P}\right)^2}$.

From a network-centric perspective, this implies that dispersion initially leads to response time gains, but these gains become more costly relative to gains due to application parallelization.

From a comprehensive architecture perspective, this means that there are tradeoffs to be made based on the nature of aggregate workloads, and the interactivity requirements of those workloads. In terms of implementation, it also suggests that network service providers have inherent advantages: the ability to cost-effectively leverage network nodes for storage and processing, as well as the ability to integrate distance-constrained protocols, such as Fibre Channel for storage access, over existing local facilities such as metropolitan Dense Wave Division Multiplexing service channels. Interestingly, such dispersion also may help offload backbone networks—this is the equivalent of the "locavore" movement for consumption of cloud services. For closely-coupled processes where distance matters, emerging capabilities such as the Data Location Service and Resource Telemetry Service implemented in the

Open Cirrus [28] federated cloud architecture and its foundation components such as Zoni and Tashi can play a role.

VII. Requirements and Research Agenda

A number of issues have been identified that revolve around hyperscale distributed systems generally [29] as well as specifics such as problems with use of static-IP addressing in a dynamic cloud environment [30]. In light of the scenarios and issues described above, there are a number of key requirements, and research opportunities accompanying them.

- **End-to-End Quality of Service** – For linking highly interactive, bandwidth-intensive, packet-loss-, jitter-, and latency-sensitive fixed, nomadic, and mobile endpoints to the cloud. And, for mixing various types of real-time traffic, for example, video and voice over IP. A good example would be cloud-based bridging for HD interactive video conferencing, also known as telepresence.
- **Multi-Protocol Support** – For enabling some storage- or data-intensive hybrid cloud scenarios, low-level support for protocols such as Fibre Channel may be key. Encapsulation and tunneling may not meet stringent performance objectives, so native support, e.g., Fibre Channel over DWDM, may be essential.
- **Bandwidth on Demand** – For migrating replicas of large objects, such as files, movies, databases or even data warehouses to the cloud for business continuity and disaster recovery purposes, or for supporting and enabling compute-oriented cloudbursting through dynamic data replication or remote data access. Bandwidth on demand has arguably been available in a number of scenarios, such as Virtual Concatenation (VCAT) and the Link Capacity Adjustment System (LCAS) used in conjunction with the Generic Framing Procedure (GFP) in SONET/SDH networks. It is increasingly possible: the Optical Transport Network [31] (OTN standards, e.g., ITU-T G.709 and G.872) decouples service bit rate from switching bit rate, enhancing scalability and flexibility. In addition, multi-degree Reconfigurable Optical Add Drop Multiplexers can switch links and thus reroute at the photonic mesh layer.
- **Pay-Per-Use Pricing** – Pay-per-use pricing, i.e., pricing and charging based on actual use, is just as important as flexibility. The economics of cloud computing—or Cloudonomics [32]–can be largely dependent on dedicated resources complemented by on-demand resources with pay-per-use pricing, to cost-optimally support variable workload demand levels.
- **OSS/BSS Integration and APIs** – For such usage-sensitive pricing, it is also important that the variable charges are accurately rated and reflected in bills. In addition, data regarding such variable use needs to be aggregated for capacity planning and engineering purposes.
- **Dynamic Pricing, Yield Management and Real-Time Capacity Management** – To maximize resource utilization, economic supply-demand approaches can be useful [33]. In the same way that airline tickets are dynamically priced to maximize utilization and revenue, it has been proposed to do the same with networks [34]. Such dynamic pricing for yield management requires a real-time interaction between usage monitoring, yield management, dynamic pricing, and a portal or API. Other cloud resources have moved away from static prices, with constructs such as spot auctions [35]; networks need to follow.
- **Route Control** – Routing is important in two senses: one, the ability of the network to determine the nearest location and the best path between a user or device and a cloud resource, such as a service or an object replica. In addition, *avoiding* certain locations is a scenario that may become important as privacy and compliance laws continue to shift.
- **New Signaling and Pricing Strategies** – It might be said that mobile networks and pricing plans were designed for a different era: voice and voice minutes, or even data and tiered data plans. With billions of connected devices such as sensors, traditional mobile signaling approaches may not scale; and, $60/month plans may not be cost effective, say, for each e-book reader, light bulb, or temperature sensor in one's house. Moreover, many of these devices will be able to scavenge bandwidth during off-peak periods or interstitially [36]. Consequently, the mobile cloud will have a greater range of performance requirements, and require a broader portfolio of pricing, than has traditionally been available.
- **Security** – Security—usually the number one concern of cloud prospects—can be enhanced by the network. MPLS labels cannot be spoofed, unlike IP addresses. The inherently shared nature of cloud resources implies that end-to-end security in hybrid and intercloud architectures must be a focus, and can be enhanced at a variety of levels. At the network level, VLANs and MPLS VPNs are important. At the large-scale system level, the ability of the network to enable network-based firewalls and implement at-scale anti-DDoS can be critical to protecting cloud-based compute and storage resources.
- **Autonomic, Integrated Development, Operations, and Management Frameworks** – Rather than developing for endpoints, such as mobile devices or web-TVs, network functions, and cloud services independently, it will become increasingly important to have a holistic view. This view must begin with development, but continue through operations and ongoing management, for example, for performance and capacity. The possibility of a Network Services Interface abstraction, where network services with specific parameters are requested, set up, and torn down, is actively under development [37].
- **Distributed System Performance and Cost Optimization** – We have already seen that even the simplest supply and demand matching is computationally intractable. Now consider the challenges of a global, distributed environment from multiple vendors—the Intercloud—with widely dispersed users and devices accessing an ever-changing portfolio of applications which

may be more or less compute-, storage-, memory-, and network-intensive, with varying resource availability at any given location, and dynamic pricing for many or all resources. Such a federated cloud environment may exist due to specialization, where one cloud provider offers certain apps that are mashed up with another's, or for resource sharing, e.g., for mutual overflow protection. Making this even more complex, policies may conflict, e.g., load-balancing policies may distribute workloads across physical servers for performance, power-management ones may consolidate them for cost [38].

VIII. CONCLUSION

Achieving the full promise of the cloud will require greater attention to network scenarios. While many early cloud implementations have focused on so called "over-the-top" solutions leveraging the Internet, robust, mission-critical cloud solutions will require a variety of diverse multi-layer networking technologies, self-service and APIs for network controls, and integration not only with billing and operations support systems, but also emerging adjunct systems, such as yield management and dynamic pricing. This new world has the potential to increase performance while optimizing efficient resource allocation and utilization, but also opens up new challenges in frameworks and tools for global cost and performance optimization.

BIBLIOGRAPHY

[1] R. Chellappa, "Intermediaries in Cloud Computing: A New Computing Paradigm," http://meetings2.informs.org/Dallas97/TALKS/MD19.html, INFORMS Dallas, October 26-29, 1997.

[2] L. Vaquero, L. Rodero-Merino, J. Caceres, M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," ACM SIGCOMM Computer Communication Review, Vol. 39, No. 1, January, 2009.

[3] M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html, Feb. 10, 2009.

[4] F. Gens, "Defining Cloud Services and Cloud Computing," http://blogs.idc.com/ie/?p=190, September 23, 2008.

[5] "Gartner Highlights Five Attributes of Cloud Computing," http://www.gartner.com/it/page.jsp?id=1035013, Jun 23, 2009.

[6] "Cloud Computing," http://en.wikipedia.org/wiki/Cloud_computing, accessed July 11, 2011.

[7] J. Weinman, "Axiomatic Cloud Theory," July, 29, 2011, at http://www.joeweinman.com/Resources/Joe_Weinman_Axiomatic_Cloud_Theory.pdf.

[8] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," v. 15, http://www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf, Oct. 7, 2009.

[9] Amazon, "AWS Import/Export," http://aws.amazon.com/importexport/.

[10] J. Dabrowski , E. Munson, "Is 100 Milliseconds Too Fast?," CHI '01 extended abstracts on Human factors in computing systems, 2001.

[11] J. Hamilton, "Perspectives: The Cost of Latency," October 31, 2009, at http://perspectives.mvdirona.com/2009/10/31/TheCostOfLatency.aspx

[12] Cisco, "Global Internet Traffic Predicted to Quadruple by 2015," http://newsroom.cisco.com/press-release-content?type=webcontent&articleId=324003

[13] L. Chao and D. Clark, "Cisco Poised to Help China Keep an Eye on Its Citizens," *The Wall Street Journal,* July 5, 2011, p. A1.

[14] J. Weinman, "Mathematical Proof of the Inevitability of Cloud Computing," at http://joeweinman.com/Resources/Joe_Weinman_Inevitability_Of_Cloud.pdf.

[15] J. Weinman, "4 ½ Ways To Deal with Data During Cloudbursts," at http://gigaom.com/2009/07/19/4-12-ways-to-deal-with-data-during-cloudbursts/.

[16] N. McKeown, et al., "OpenFlow: Enabling Innovation in Campus Networks", March 14, 2008, ACM SIGCOMM Computer Communication Review, Volume 38, Number 2, April 2008, pp. 69-74.

[17] J. Van der Merwe, et al., "Dynamic Connectivity Management with an Intelligent Route Service Control Point," ACM SIGCOMM '06 Workshops, September 11-15, 2006.

[18] http://www.geni.net/

[19] http://www.planet-lab.org/about

[20] http://www.openflow.org/wp/2010/07/gec8/

[21] "Deliverable Reference Number D.2.7: Grid-GMPLS Network Interfaces Specification," PHOSPHORUS-LUCIFER, at http://www.ist-phosphorus.eu/files/deliverables/Phosphorus-deliverable-D2.7_M17.pdf

[22] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Co., San Francisco, 1979.

[23] J. Weinman, "Cloud Computing is NP-complete," at http://joeweinman.com/Resources/Joe_Weinman_Cloud_Computing_Is_NP-Complete.pdf

[24] J. Hopcroft and R. Karp, "An $n^{5/2}$ Algorithm for Maximum Matching in Bipartite Graphs," SIAM J. Computing, 1975, pp. 225-231, referenced in Shimon Even, *Graph Algorithms*, Computer Science Press, 1979.

[25] N. Megiddo and K. Supowit, "On the Complexity of Some Common Geometric Location Problems," SIAM J. Computing, Vol. 13, No. 1, February 1984.

[26] J. Weinman, "Smooth Operator: The Value of Demand Aggregation," at http://www.joeweinman.com/Resources/Joe_Weinman_Smooth_Operator_Demand_Aggregation.pdf

[27] J. Weinman, "As Time Goes By: The Law of Cloud Response Time," at http://joeweinman.com/Resources/Joe_Weinman_As_Time_Goes_By.pdf.

[28] A. Avetisyan, et al., "Open Cirrus: A Global Cloud Computing Testbed," IEEE Computer Magazine, April, 2010, 35-43.

[29] K. Birman, G. Chockler, and R. van Renesse, "Towards a Cloud Computing Research Agenda," Toward a cloud computing research agenda, ACM SIGACT News, v.40 n.2, June 2009

[30] I. Sriram and A. Khajeh-Hosseini, "Research Agenda in Cloud Technologies," submitted to the 1st ACM Symposium on Cloud Computing, arXiv:1001.3259v1

[31] T. Walker, "Optical Transport Network Tutorial," http://www.itu.int/ITU-T/studygroups/com15/otn/OTNtutorial.pdf

[32] http://www.cloudonomics.com/

[33] R. Wolski, J. Brevik, J. Plank, and T. Bryan, "Grid resource allocation and control using computational economies," pp. 747-771, in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G. Fox, A. Hey, Eds., Wiley, 2001.

[34] J. MacKie-Mason and H. Varian, "Pricing Congestible Network Resources," IEEE Journal on Selected Areas in Communications, (13)7:1141-1149, Sept. 1995.

[35] "Amazon EC2 Spot Instances," at http://aws.amazon.com/ec2/spot-instances/ .

[36] A. Plummer Jr., M. Taghizadeh, and S. Biswas, "Statistical Bandwidth Scavenging for Prioritized Device Coexistence," 29th IEEE International Performance Computing and Communications Conference.

[37] Open Grid Forum, "Network Service Interface WG," at http://www.gridforum.org/gf/group_info/view.php?group=nsi-wg

[38] B. Rochwerger, et al., "Reservoir—When One Cloud Is Not Enough," IEEE Computer, March 2011, pp. 44-51.