# Cloud Computing

## *Distributed Internet Computing for IT and Scientific Research*

**Marios D. Dikaiakos and George Pallis**
*University of Cyprus*

**Dimitrios Katsaros**
*University of Thessaly*

**Pankaj Mehra**
*Hewlett-Packard Labs*

**Athena Vakali**
*Aristotle University of Thessaloniki*

One vision of 21st century computing is that users will access Internet services over lightweight portable devices rather than through some descendant of the traditional desktop PC. Because users won't have (or be interested in) powerful machines, who will supply the computing power? The answer to this question lies with *cloud computing*.

Cloud computing is a recent trend in IT that moves computing and data away from desktop and portable PCs into large data centers. It refers to applications delivered as services over the Internet as well as to the actual cloud infrastructure — namely, the hardware and systems software in data centers that provide these services.

The key driving forces behind cloud computing are the ubiquity of broadband and wireless networking, falling storage costs, and progressive improvements in Internet computing software. Cloud-service clients will be able to add more capacity at peak demand, reduce costs, experiment with new services, and remove unneeded capacity, whereas service providers will increase utilization via multiplexing, and allow for larger investments in software and hardware.

Currently, the main technical underpinnings of cloud computing infrastructures and services include virtualization, service-oriented software, grid computing technologies, management of large facilities, and power efficiency. Consumers purchase such services in the form of infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), or software-as-a-service (SaaS) and sell value-added services (such as utility services) to users. Within the cloud, the laws of probability give service providers great leverage through statistical multiplexing of varying workloads and easier management — a single software installation can cover many users' needs.

We can distinguish two different architectural models for clouds: the first one is designed to scale out by providing additional computing instances on demand. Clouds can use

 1089-7801/09/$26.00 © 2009 IEEE 

these instances to supply services in the form of SaaS and PaaS. The second architectural model is designed to provide data and compute-intensive applications via scaling capacity. In most cases, clouds provide on-demand computing instances or capacities with a "pay-as-you-go" economic model. The cloud infrastructure can support any computing model compatible with loosely coupled CPU clusters. Organizations can provide hardware for clouds internally (*internal clouds*), or a third party can provide it externally (*hosted clouds*). A cloud might be restricted to a single organization or group (*private clouds*), available to the general public over the Internet (*public clouds*), or shared by multiple groups or organizations (*hybrid clouds*).

A cloud comprises processing, network, and storage elements, and cloud architecture consists of three abstract layers. *Infrastructure* is the lowest layer and is a means of delivering basic storage and compute capabilities as standardized services over the network. Servers, storage systems, switches, routers, and other systems handle specific types of workloads, from batch processing to server or storage augmentation during peak loads. The middle *platform* layer provides higher abstractions and services to develop, test, deploy, host, and maintain applications in the same integrated development environment. The *application* layer is the highest layer and features a complete application offered as a service.

## Key Challenges

In 1961, John McCarthy envisioned that "computation may someday be organized as a public utility." We can view the cloud computing paradigm as a big step toward this dream. To realize it fully, however, we must address several significant problems and unexploited opportunities concerning the deployment, efficient operation, and use of cloud computing infrastructures.

### Software/Hardware Architecture

Cloud computing services's emergence suggests fundamental changes in software and hardware architecture. Computer architectures should shift the focus of Moore's law from increasing clock speed per chip to increasing the number of processor cores and threads per chip. Industry and academia must design novel systems and services that would exploit a high degree of parallelism. Software architectures for massively parallel, data-intensive computing, such as MapReduce (http://labs.google.com/papers/mapreduce.html), will grow in popularity. In terms of storage technologies, we'll likely shift from hard disk drives (HDDs) to solid-state drives (SDDs), such as flash memories, or, given that completely replacing hard disks is prohibitively expensive, hybrid hard disks — that is, hard disks augmented with flash memories, which provide reliable and high-performance data storage. The biggest barriers to adopting SSDs in data centers have been price, capacity, and, to some extent, the lack of sophisticated query-processing techniques. However, this is about to change as SSDs' I/O operations per second (IOPS) benefits become too impressive to ignore, their capacity increases at a fast pace, and we devise new algorithms and data structures tailored to them.

### Data Management

The shift of computer processing, storage, and software delivery away from desktop and local servers, across the Internet, and into next-generation data centers results in limitations as well as new opportunities regarding data management. Data is replicated across large geographic distances, where its availability and durability are paramount for cloud service providers. It's also stored at untrusted hosts, which creates enormous risks for data privacy. Computing power in clouds must be elastic to face changing conditions. For instance, providers can allocate additional computational resources on the fly to handle increased demand. They should deploy novel data management approaches, such as analytical data management tasks, multitenant databases for SaaS, or hybrid designs among database management systems (DBMSs) and MapReduce-like systems so as to address data limitations and harness cloud computing platforms' capabilities.

### Cloud Interoperability

Cloud interoperability refers to customers' ability to use the same artifacts, such as management tools, virtual server images, and so on, with a variety of cloud computing providers and platforms.

Cloud interoperability will enable cloud infrastructures to evolve into a worldwide, transparent platform in which applications aren't restricted to enterprise clouds and cloud service

providers. We must build new standards and interfaces that will enable enhanced portability and flexibility of virtualized applications. Up to now, significant discussion has occurred around open standards for cloud computing. In this context, the "Open Cloud Manifesto" (www.opencloudmanifesto.org) provides a minimal set of principles that will form a basis for initial agreements as the cloud community develops standards for this new computing paradigm.

### Security and Privacy

In cloud computing, a data center holds information that end-users would more traditionally have stored on their computers. This raises concerns regarding user privacy protection because users must outsource their data. Additionally, the move to centralized services could affect the privacy and security of users' interactions. Security threats might happen in resource provisioning and during distributed application execution. Also, new threats are likely to emerge. For instance, hackers can use the virtualized infrastructure as a launching pad for new attacks. Cloud services should preserve data integrity and user privacy. At the same time, they should enhance interoperability across multiple cloud service providers. In this context, we must investigate new data-protection mechanisms to secure data privacy, resource security, and content copyrights.

**We must build new standards and interfaces that will enable enhanced portability and flexibility of virtualized applications.**

### Service Provisioning and Cloud Economics

Providers supply cloud services by signing service-level agreements (SLAs) with consumers and end-users. Cloud service consumers, for instance, might have an SLA with a cloud service provider concerning how much bandwidth, CPU, and memory the consumer can use at any given time throughout the day. Underestimating the provision of resources would lead to broken SLAs and penalties. On the other hand, overestimating the provision of resources would lead to resource underutilization and, consequently, a decrease in revenue for the provider. Deploying an autonomous system to efficiently provision services in a cloud infrastructure is a challenging problem due to the unpredictability of consumer demand, software and hardware failures, heterogeneity of services, power management, and conflicting signed SLAs between consumers and service providers.

In terms of cloud economics, the provider should offer resource-economic services. Novel, power-efficient schemes for caching, query processing, and thermal management are mandatory due to the increasing amount of waste heat that data centers dissipate for Internet-based application services. Moreover, new pricing models based on the pay-as-you-go policy are necessary to address the highly variable demand for cloud resources.

## In this Issue

Given the continued, intense activity in the cloud arena, we invited researchers and practitioners to submit articles to this special issue of *IC* describing research efforts and experiences concerning the deployment, efficient operation, and use of cloud computing infrastructures. From among the 42 submissions, and after rigorous review, we selected the following four articles as representative of ongoing research and development activities.

The first article, "Virtual Infrastructure Management in Private and Hybrid Clouds," by Borja Sotomayor, Rubén S. Montero, Ignacio M. Llorente, and Ian Foster, presents two open source projects for private and hybrid clouds. OpenNebula is a virtual infrastructure manager that can be used to deploy virtualized services on both a local pool of resources and on external IaaS clouds. Haizea is a resource lease manager that can act as a scheduling back end for OpenNebula, providing advance reservations and resource preemption.

"Harnessing Cloud Technologies for a Virtualized Distributed Computing Infrastructure," by Alexandre di Costanzo, Marcos Dias de Assunção, and Rajkumar Buyya, presents the realization of a system — termed the InterGrid — for interconnecting distributed computing infrastructures by harnessing virtual machines. The article provides an abstract view of the proposed architecture and its implementation. Experi-

ments show the scalability of an InterGrid-managed infrastructure and how the system can benefit from using cloud infrastructure.

In "Content-Centered Collaboration Spaces in the Cloud," John S. Erickson, Susan Spence, Michael Rhodes, David Banks, James Rutherford, Edwin Simpson, Guillaume Belrose, and Russell Perry envision a cloud-based platform that inverts the traditional application-content relationship by placing content rather than applications at the center, letting users rapidly build customized solutions around their content items. The authors review the dominant trends in computing that motivate the exploration of new approaches for content-centered collaboration and offer insights into how certain core problems for users and organizations are being addressed today.

The final article, "Sky Computing," by Katarzyna Keahey, Maurício Tsugawa, Andréa Matsunaga, and José A.B. Fortes, describes the creation of environments configured on resources provisioned across multiple distributed IaaS clouds. This technology is called sky computing. The authors provide a real-world example and illustrate its benefits with a deployment in three distinct clouds of a bioinformatics application.

Cloud computing is a disruptive technology with profound implications not only for Internet services but also for the IT sector as a whole. Its emergence promises to streamline the on-demand provisioning of software, hardware, and data as a service, achieving economies of scale in IT solutions' deployment and operation. This issue's articles tackle topics including architecture and management of cloud computing infrastructures, SaaS and IaaS applications, discovery of services and data in cloud computing infrastructures, and cross-platform interoperability.

Still, several outstanding issues exist, particularly related to SLAs, security and privacy, and power efficiency. Other open issues include ownership, data transfer bottlenecks, performance unpredictability, reliability, and software licensing issues. Finally, hosted applications' business models must show a clear pathway to monetizing cloud computing. Several companies have already built Internet consumer services such as search, social networking, Web email, and online commerce that use cloud computing infrastructure. Above all, cloud computing's still unknown "killer application" will determine many of the challenges and the solutions we must develop to make this technology work in practice.

**Acknowledgments**

We would like to express our gratitude to the authors of all submitted articles and the reviewers for their contributions to this special issue. We thank Fred Douglis, *IC*'s editor in chief, and Michael Rabinovich, associate editor in chief, for their support of the special issue, and also the production staff at the IEEE Computer Society who made this issue possible.

**Marios D. Dikaiakos** is an associate professor at the University of Cyprus, Nicosia. His research interests include network-centric computing, with an emphasis on grids, vehicular computing, and Web technologies. Dikaiakos has a PhD in computer science from Princeton University. He's a senior member of the ACM, and a member of the IEEE Computer Society and the Technical Chamber of Greece. Contact him at mdd@cs.ucy.ac.cy.

**Dimitrios Katsaros** is a lecturer at the University of Thessaly, Greece. His research interests include distributed systems, such as the Web and Internet, social networks, mobile and pervasive computing, and wireless ad hoc and wireless sensor networks. Katsaros has a PhD in informatics from Aristotle University of Thessaloniki. Contact him at dkatsar@inf.uth.gr.

**Pankaj Mehra** is a principal member of the technical staff at Hewlett-Packard Labs. His research focuses on the design of systems and networks for large-scale enterprise applications. Contact him at pankaj.mehra@hp.com.

**George Pallis** is a visiting lecturer at the University of Cyprus, Nicosia. His research interests include distributed systems, such as the Web and grids, content distribution networks, information retrieval, and data clustering. Pallis has a PhD in informatics from Aristotle University of Thessaloniki. Contact him at gpallis@cs.ucy.ac.cy.

**Athena Vakali** is an associate professor at the Aristotle University of Thessaloniki. Her current research interests include Web usage mining, content delivery networks, Web and social Web data clustering, and Web data caching/outsourcing. Vakali has a PhD in informatics from the Aristotle University. Contact her at avakali@csd.auth.gr.