
Chapter 1

Foreword

1

Foreword



The buzz and excitement around cloud computing has been steadily building over the last few years. There is general agreement that something big and profound is going on out there, although we may not be totally sure what it is yet. "There is a clear consensus that there is no real consensus on what cloud computing is," was one of the key conclusions at a recent conference on the subject.

So, what in the world is cloud computing? Is it the evolution of the Internet? Is it a new model of computing? Is it a way of delivering just about anything "as a service"? Does cloud computing represent the *industrialization* of IT, much as happened with electricity one hundred years ago as it became widely used across the economy and society?

Cloud computing, in my opinion, is all the above and then some. It is like the fable of the blind men and the elephant. Each one touches a different part of the elephant. They then compare notes on what they felt, and learn that they are in complete disagreement.

To begin with, cloud is the natural evolution of the Internet. As we all know, the original Internet was primarily developed as a TCP/IP network. It later added a number of communication oriented applications like e-mail and file transfer. The advent of the World Wide Web in the early 1990s transformed the Internet into a huge source of information and content, and coupled with the browser, ushered the Internet into the wider commercial world a few years later. Later in the decade, companies started to leverage the Internet for all kinds of *e-business* applications. Irrational exuberance and the dot-com bubble followed.

The bursting of the bubble barely slowed down the continuing advances of the Internet. A number of new initiatives were aimed at making it easier to access IT resources and applications over the Internet, including virtualization, grid computing, service oriented architectures and utility computing. Other initiatives focused on making the Internet much more pervasive and accessible over a wide variety of devices beyond personal computers, including smartphones, mobile devices and sensors.

Cloud computing is essentially turning the Internet into a major part computing platform, significantly extending and improving the technologies and capabilities first introduced in these earlier initiatives. The Cloud is becoming *the platform* for applications, information and services for the billions of smart devices and trillions of smart sensors connected to the Internet.

Cloud thus represents the emergence of a new model of computing in the IT industry. This is a big deal, because in the fifty to sixty years since there has been an IT industry, this would be only the third such model, with centralized and client-server computing being the two previous ones.

In its early decades, the '50s, '60s and '70s, just about all computing was centralized, typically consisting of mainframes and supercomputers located behind the glass walls of the data center. Generally, these computers were quite expensive, shared by many users, and managed by a central IT organization. Minicomputers were smaller and less expensive versions of these central computers designed to be used by departmental functions in large enterprises as well as smaller companies.

The 1980s saw the emergence of increasingly powerful and inexpensive microprocessors, personal computers and Unix-based workstations. These technologies paved the way for the new distributed client-server model of computing. The architecture of these client-server systems was quite different from the architecture of the mainframes of the central computing model. The designs were optimized for low costs and simplicity, rather than efficiency and reliability.

The underlying assumption in the client-server model was that because the individual systems were fairly inexpensive, you could add as many systems as you needed to support the various applications and users in the installation. Over time, companies ended up with very large numbers of relatively small servers, distributed over various departments in the organization, each dedicated to a single application. Because the server were not shared among multiple application or a large enough group of users, they often ran at low utilizations, between 10% and 20%. These factors eventually led to significantly increased management complexities and costs.

The web-based applications that began to appear in the mid 1990s generally followed a client-server model. The much larger number of users now able to access these web applications required servers that were significantly more scalable and reliable, as well as offering significantly better systems management. We saw the rise of very large web sites from Google, Amazon, Yahoo and others, that provided all kinds of consumer services to huge numbers of users, including search, maps, shopping and news. Later on, we saw the rise of Web 2.0 concepts like blogs and wikis, and social networking sites like MySpace and Facebook, which rapidly grew to support large numbers of users communicating and sharing information with each other.

Through the years, we kept adding features to the client-server IT infrastructures to make them more scalable and easier to manage. I believe that what finally gave us unmistakable signals that client-server was running out of gas was the explosive rise of mobile devices in the last few years, as well as the prospect of even larger number of sensors and other digital technologies, each with their own IP address, that were started to be embedded into myriads of things in the physical world, like cars, appliances, medical equipment, cameras, roadways, pipelines, and pharmaceuticals.

Client-server computing was not up to the massive scalability and low costs required to support these billions of new mobile devices and trillions of sensors. A

new model of computing, no longer optimized around individual PCs but around the Internet as a whole was needed. As is often the case, it has taken a number of years for the marketplace to reach consensus and finally give this new computing model a name around which everyone can rally. Cloud computing has emerged as the name most people have settled on for this new, Internet-based computing model.

The scale and reach of cloud computing is driving a major revolution in the way services, applications and information are delivered and consumed. Cloud is driving a much needed industrialization of IT data centers and of the IT infrastructure in general. Thirty years ago, we saw something similar happen in manufacturing. Before that time, most manufacturing plants were fairly inefficient by almost any measure, and were turning out products of varying quality. Then, driven by the huge success of Toyota and other companies around the world, the industrial sector and academia discovered the merits of applying engineering discipline as well as a holistic, systems-wide approach to manufacturing processes.

Data centers are the manufacturing plants for the 21st century services and information economy. But, with the exception of a few, relatively young, *born-to-the-cloud* companies, the data centers of most enterprises are in the same pre-industrialization phase that manufacturing was thirty years ago. They have not exercised the needed engineering discipline in their IT operations. They have allowed different departments in their organization to architect their own systems and application, which often do not interoperate with each other. Unless these data centers significantly improve the quality and efficiency of their operations, they will just not be competitive. Many will not be able to do so, and will instead rely on professional service providers for many of their IT operations, much as has happened in manufacturing.

But, perhaps the major revolution that cloud computing will usher is in the design of the services and applications themselves, to make them much easier for people to consume and interact with, often on the go with a mobile device and a relatively small screen. While many talk about Cloud as IT-as-a-service, platforms-as-a-service, or software-as-a-service, the reality is that most people would rather not have to know anything about IT, platforms and software to deal with information, applications and IT-based services in general. What they really want is well-designed-services-as-a-service to help them in their everyday work and life - be it dealing with money, health matters, communications, entertainment and so on. I fully expect that the years ahead will usher in a plethora of innovative new services, which will be highly useful and a pleasure to use, as well as ubiquitously available at very reasonable costs.

I like the way *The Economist* described cloud computing in its introduction to a recent special report on the subject:

"In the beginning computers were human. Then they took the shape of metal boxes, filling entire rooms before becoming ever smaller and more widespread.

Now they are evaporating altogether and becoming accessible from anywhere [...] Computing is taking on yet another new shape. It is becoming more centralized again as some of the activity moves into data centers. But more importantly, it is turning into what has come to be called a 'cloud', or collections of clouds."

"Computing power will become more and more disembodied and will be consumed where and when it is needed [...] it will also profoundly change the way people work and companies operate. It will allow digital technology to penetrate every nook and cranny of the economy and of society, creating some tricky political problems along the way."

Irving Wladawsky-Berger

Chairman Emeritus, IBM Academy of Technology.