I D C T E C H N O L O G Y
S P O T L I G H T

From Silicon to Cloud: Building Up to Cloud Computing

April 2009

Sponsored by AMD

Introduction: Cloud Formation for Enterprises

The topic of cloud computing has received tremendous attention in the past year. Cloud computing, in a nutshell, is the delivery of some aspect of IT (infrastructure, platforms, or applications) as a service that is accessed via Internet standards and infrastructure. Several factors have driven attention to this emerging computing model. From a technology point of view the maturation of virtualization software has enabled highly efficient and dense computing, key to the success of a cloud. Consumers have driven much of the initial wave of cloud server deployments, as social media and other Web 2.0 usage has skyrocketed. The next wave of cloud is expected to come from the enterprise, as the pressures of the current economy drive many companies to consider new IT models when they are forced to continue to do more with less.

Definitions

To clarify the discussion, IDC will define the following terms used throughout the document:

<u>Cloud computing services</u> – Consumer and business products, services, and solutions delivered and consumed in real-time over the Internet. Essential characteristics include:

- Shared services
- Under-virtualized management
- Accessible over the Internet (by people and other services) via Internet standards

IDC categorizes clouds into the following three general types, each offering services at a different level of the operating stack:

- Infrastructure clouds – Provides infrastructure services such as compute and storage. An example of this type of cloud would be Amazon Web Services EC2 and S3, which offers virtual machines and storage in the cloud.
- Platform clouds – Offers essentially an operating system in the cloud, providing an application development and runtime environment. Microsoft's Azure is an example of this type of cloud, offering the ability to create and run .NET applications.
- Application clouds – Hosts an entire application in the cloud, such as Salesforce.com's customer relationship management (CRM) application.

<u>Platform</u> – Hardware and software technological building blocks arranged according to the needs of a given usage model, such as virtualized servers.

Usage Model – The profile of how a system, such as a virtualized server, is used by its owner. As system development has evolved, usage models have come to dictate system design, manufacturing, configuration, software load, and sales channel. Accordingly, all of the vendors in the value chain — including semiconductor vendors and OEMs — of a system are now involved in most such system aspects.

Virtualization – The process of abstracting the underlying physical architecture (servers, networks, storage, etc.) to allow software to be decoupled from the hardware. Most commonly refers to server virtualization which uses hypervisor software to segment a server to support multiple operating environments and the related stacks of applications, application development and deployment software, and system infrastructure software. Each of the client or server operating environments is allowed to act as if it controls the entire machine, but resources are actually allocated on the basis of rules established either at the time of configuration or dynamically by related management software. Thus, multiple operating systems, even ones from different vendors, can now share the same machine without awareness of the proximity for other operating systems on the same physical hardware.

Benefits

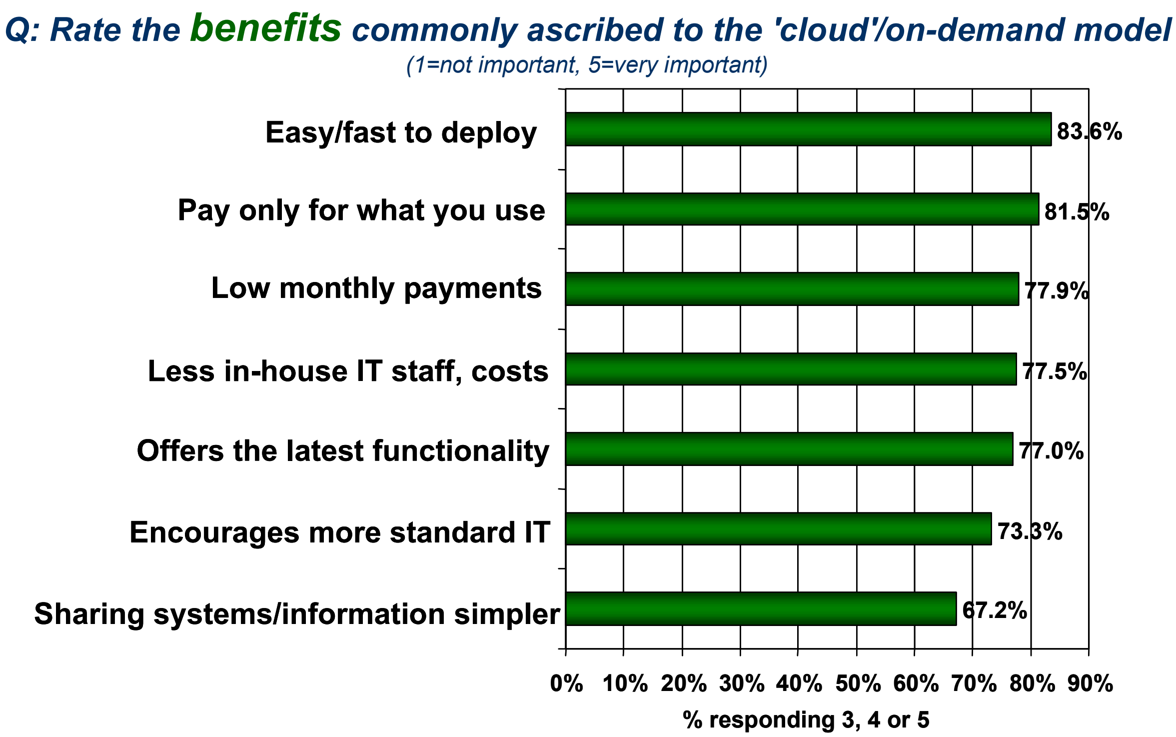
Cloud computing can offer the following important benefits to an enterprise:

- **Elasticity –** Capacity planning is a difficult exercise for IT as they must try to predict usage and acquire just enough resources to avoid excessive under- or over-provisioning. IT may also have to deal with usage spikes, such as end-of-quarter activities or seasonal fluctuations, that require short-term access to additional compute resources that might otherwise remain idle. Having to purchase compute resources to service the peak load is often highly inefficient and costly. Clouds offer computing resources on demand that can scale up or down with demands of the business, minimizing the need for capacity planning and the need to plan for peak usage.
- **Consumption-based billing –** Cloud usage is typically billed similarly to a utility, based on only what you use. This enables businesses to better align IT usage with costs, as most businesses have fluctuating loads. Typically a business would have to acquire enough IT resources to deal with the maximum load, even though that scenario may only occur infrequently. Buying IT as a service also shifts IT budgets from a capital expenditure (CAPEX) to an operational expenditure (OPEX).
- **Automation –** Clouds heavily automate operations to reduce management complexity and keep software updated. This can offer businesses the following advantages:
  - Reduce IT staff and costs needed to maintain base level of IT infrastructure.
  - Refocus IT, allowing IT to be more strategic than tactical to better align IT with business needs. More focus on business requirements and service-level management.
  - Offer the latest functionality – Updating software is often a slow and painful process, often requiring a rip and replace. Clouds offer enterprises the ability to acquire the latest and greatest software and to stay up to date automatically.
  - Speed of operations – Clouds can accelerate the speed of IT, often leveraging virtualization. As the demand for IT grows, so does the appetite for instant delivery. Clouds can offer quicker and easier deployments, instant provisioning, and on-demand scaling.

Figure 1 lists the benefits rated highest by respondents to an IDC demand-side survey.

Figure 1

Benefits of the Cloud Computing Model



Source: IDC, 2009

Trends

Two major trends in server platforms are enabling cloud computing. First, virtualization-optimized hardware and software are reducing the overhead of virtualization and increasing density. Second, hardware and software optimized for a balanced approach to power consumption and raw performance are allowing clouds to scale upwards even as datacenter power and cooling limits are reached. Some examples of such optimized hardware and software include the following:

- **Processor –** Features like clock gating, frequency adjusting, core idling, dynamically turning off part of the microprocessor, turning down the frequency to reduce voltage, or shutting down one or more cores to allow the processor subsystem to match performance to power consumption and thermal output parameters.

- **Core logic chipset –** As the intermediary between the processor and the rest of the system, the chipset determines many of the critical functions of the system, including the level of graphics support and I/O capabilities. Many of the idling and scaling features of the processor are supported by the core logic chipset. Additionally, new chipsets enable direct device assignment between a virtual machine and an I/O device and allow access to the system BIOS in order to disable clocks of idled devices that would otherwise consumer power and dissipate heat.

- **Hypervisors and system software –** Hypervisors are being continually optimized for performance and power management. Next-generation hypervisors are expected to optimized to increase virtual machine (VM) density, reduce overhead, increase performance (especially I/O), and take advantage of the latest hardware acceleration features. Also, they will likely support better power management, such as core parking; the ability to consolidate VMs onto fewer cores and turn off the unneeded cores to save power. There are multiple software layers of critical technology that may further support cloud computing. These layers include the operating system, drivers, middleware, and applications, all of which are being optimized to be virtualized, scale and balance system performance, power, and cost.

The server virtualization marketplace has been evolving rapidly over the past few years, and IDC has seen customer attitudes and stances toward virtualization mature rapidly as well. The latest IDC virtualization data predicts and shows the following:

- Virtual machines will begin to outnumber physical hosts sometime this year. By 2012, VMs will outnumber physical hosts 2:1
- VM density will increase 3x, compared to 2005 densities
- 50% of VM workload is production

As customers gain familiarity with the technology and as the technology matures, organizations are leveraging virtualization to solve far more than their server consolidation challenges. Increasingly, customers are using virtualization for disaster recovery, high availability, remote clients, and, ultimately, managing the delivery of business applications to end users. The next generation of virtualization is expected to offer new levels of automation and orchestration in the datacenter and extend beyond servers to the network and storage. It will likely be highly adaptive, service-oriented, and policy-based. Internal clouds are anticipated to be a major delivery model for this next-generation, fully virtualized datacenter.

While many compute services have been taking place "in the cloud" for many years, virtualization of all types has enabled new cloud models and changed the economics of the game.

Considering the AMD Opteron™ Processor

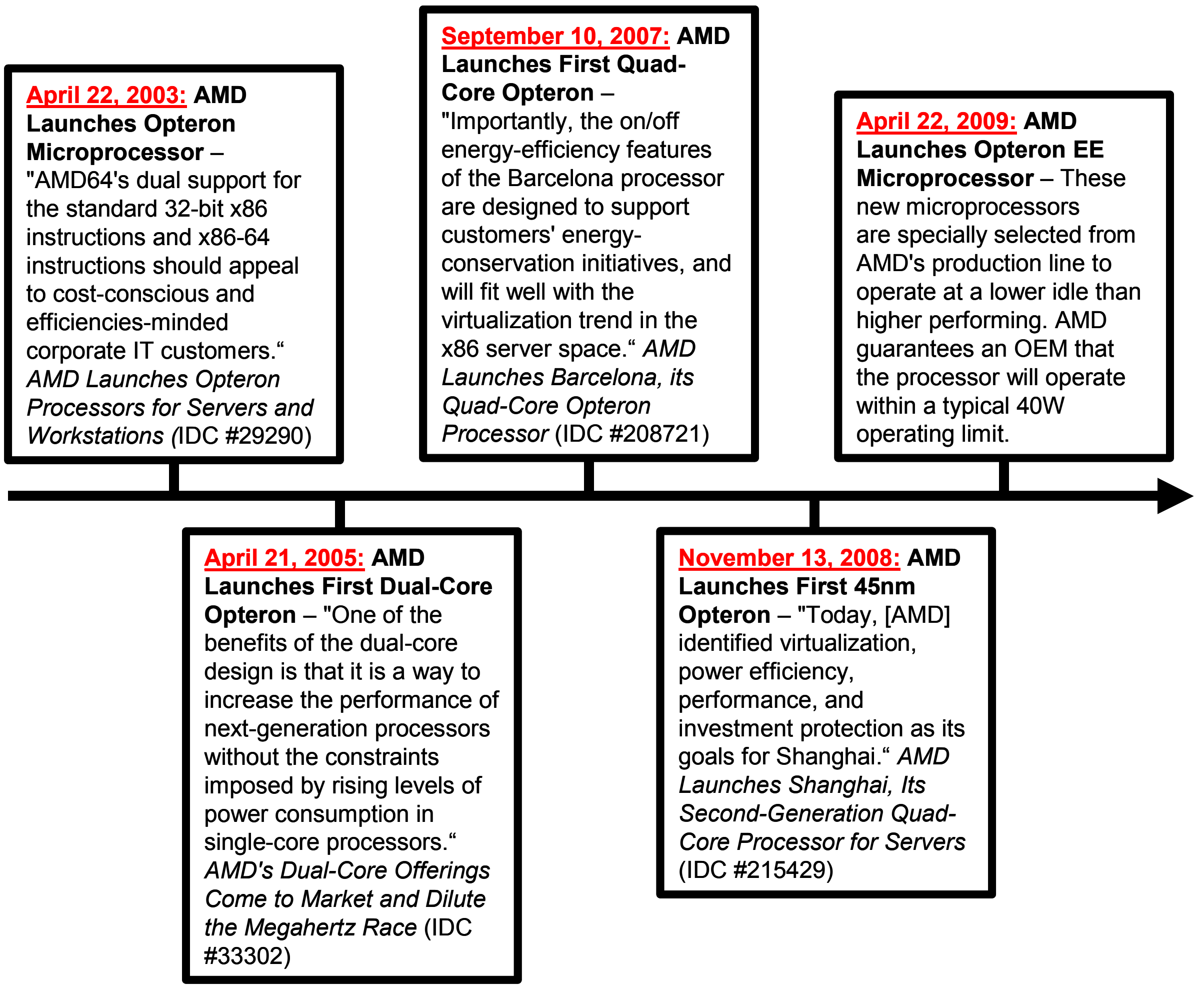
For a server to be well-suited to cloud computing, it must balance performance, power, and price. The history of server platforms based on the AMD Opteron processor serves as a case in point for the building blocks that have been put in place to assure that balance is maintained for cloud computing today.

As the first 64-bit-compatible x86 microprocessor, the AMD Opteron processor emphasized performance and foreshadowed the transition to 64-bit processing and wider memory addressing, two major elements of good performance that enable virtualization. The transition to dual-core, quad-core, and the coming transition to six-core processors reflect the ongoing effort to keep in place dynamically scalable performance resources, even as features such as clock gating and frequency scaling are introduced to turn features on or off, up or down, which balance performance with power consumption and thermal output.

Figure 2 quotes from IDC's coverage of major milestones in the history of the AMD Opteron processor to illustrate how the processor reflects the trends that lead up to cloud computing today.

Figure 2

Major Events in the History of the AMD Opteron Server Microprocessors



Source: IDC, 2009

Processor and Platform

The AMD Opteron EE (energy efficient) processor represents a new segment in AMD's current server product line up. Running at either 2.3GHz or 2.1GHz and manufactured in AMD's 45nm manufacturing process, the 40W Average CPU Power (ACP) AMD Opteron EE processors are specially selected from AMD's production line to operate at lower idle power than its higher performing models. The target environments for the EE parts — Web hosting and cloud computing — have demands that place power efficiency far ahead of raw performance, allowing AMD to sort for its most efficient processors that can deliver its best power profile while still maintaining acceptable performance for associated workloads. AMD estimates that, in comparison to the 55W Quad-Core AMD Opteron HE processor, codenamed "Shanghai," the AMD Opteron EE processor will offer up to a 13%, or 28W, savings in platform-level power consumption when measured in the exact same system, and up to a 14% reduction in processor power at idle.

Bob Weisickle, CTO of server OEM ZT Systems, says, "We've used the AMD Opteron EE processor to build 1U servers for a Web 2.0 cloud provider who needed very low power systems. In such environments, saving just a few watts per server is a big deal. When you use a low-power CPU like the AMD Opteron processor EE, the power savings ripples throughout the rest of the system. Since it produces less heat, you can reduce fan capacity, saving additional power there. Also, lower system power consumption means you can use a smaller and more efficient power supply, again saving more power."

Other semiconductor components critical to enabling cloud computing are the core logic chipset and memory. In fact, AMD plans to shortly deliver its own chipset, the AMD SR5690 chipset, so that it can better integrate chipset-resident functions into its AMD Opteron processor-based platforms. Since virtualization and cloud computing are very memory- and I/O-intensive, these systems also need robust memory support.

In upcoming platforms, AMD has announced plans to support DDR3 memory. In those new platforms, AMD plans to sustain socket longevity for multiple generations of the processor so that OEMs do not have to change their motherboards to accommodate the new processors. Further, AMD plans to keep the total power envelope of its platforms constant so that OEMs can sustain a consistent power consumption and thermal output profile for their datacenter customers.

Executive Viewpoint

Patrick Patla is the General Manager and Vice President for AMD's Server Business Unit. His team's announcement of the Opteron EE microprocessor coincides with IDC's Server Virtualization Forum. In this interview, Pat describes the benefits of the processor and its relevance to the emerging trend of cloud computing.

Q: How does the AMD Opteron EE processor serve the needs of a datacenter implementing virtualization?

A: The AMD Opteron EE processor hits the target for today's datacenters with the balance of performance and power efficiency, while not giving up any features such as our AMD-V feature set. This is important so organizations can better build to their power-consumption targets without giving up any advanced features.

Q: What are the challenges OEMs face in building systems for cloud computing?

A: The challenge is, unlike real clouds in the sky, each datacenter manager has different goals and each implementation of cloud computing can be unique — except for the common goal of lowest possible power consumption without giving up performance.

Q: What does cloud computing mean for people involved in the datacenter?

A: Cloud computing defines the flexibility each and every datacenter manager wants to be able to solve their unique IT challenges, whether it's failover capabilities, server consolidation, or the best way to maximize resources.

Source: IDC

Considerations and Challenges

In the old days of mainframes and minicomputers, computing systems were designed according to the capabilities of the technologies, not the needs of the end users. Today, the platform approach trend dictates that components are based on standardized technologies and arranged according to the needs of the end users. The ultimate expression of cloud computing is that servers and client systems support end-user data and environments in all their individualized and customized glory anytime, anywhere.

Private clouds, corporate clouds, department-level clouds, and so on each have unique requirements and will require the underlying platforms to scale and match the workloads of those end users. Such requirements put special stress for all members of the server value chain to coordinate with each other and align on what's necessary to build a successful platform.

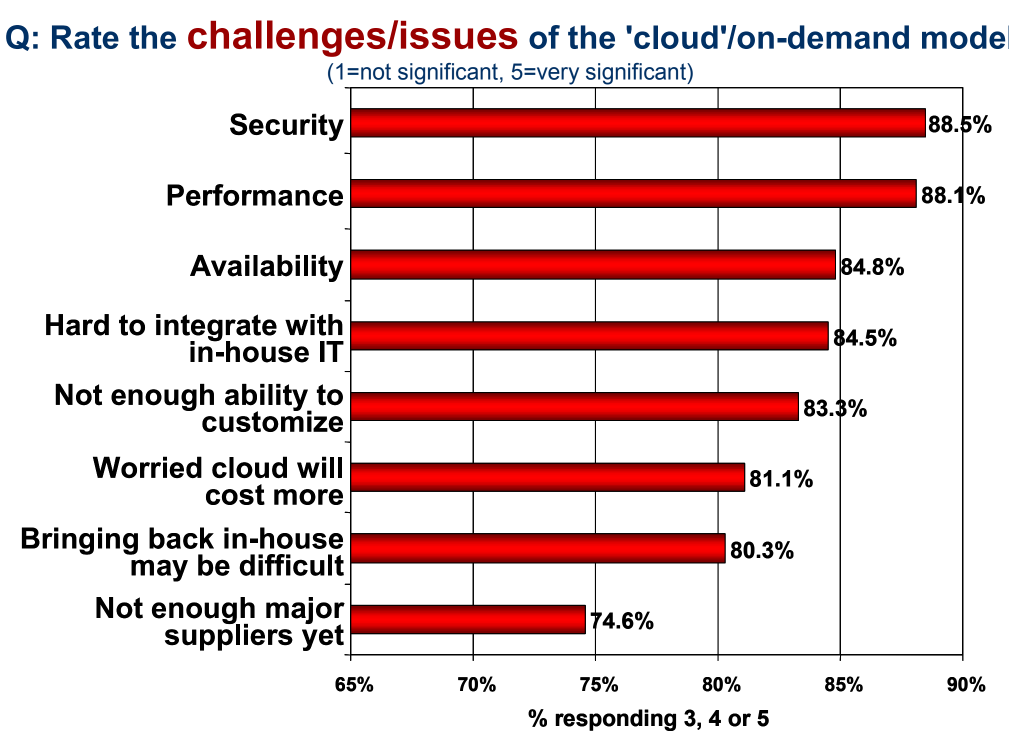
"Customers need a compelling balance of performance, power, and price when it comes to server technology," says Greg Huff, Hewlett-Packard's Chief Technology Officer for Industry Standard Servers. "Together with AMD, HP works to understand the specific application workloads our customers require. As a result, we are able to design and align our processor and server architectures so that customers benefit from a world-class solution."

Other considerations and challenges of cloud computing include the issues of security, performance, availability and integration (see Figure 3). The following key questions still to be answered are:

- Is the need for maintaining data and environment security more compelling than the cost savings of outsourcing the infrastructure to another party?
- Can clouds offer enterprise-class performance optimization and availability that matches today's enterprise datacenter?
- How will clouds integrate and interface with on-premises resources?

Figure 3

Challenges of Cloud Computing Model



Source: IDC, 2009

Conclusion

Like any system, a server is a collection of many interdependent hardware and server components. Like any product, the server must be designed for how it will be used. The need to have components work well together is why the computing industry has evolved the platform approach, which dictates that standardized technologies are arranged and applied to scale across many vectors (e.g., power, performance, size, weight, etc.) that represent the usage model.

Cloud computing is a computing model with the potential to change the way human beings interact with information technology. Moving user data and environments to the Internet, so that they can be accessed via any Internet connection, dramatically changes the nature of the underlying server and client platforms. Server platforms must balance performance with power management in view of cost and other requirements such as security and reliability.

A B O U T   T H I S   P U B L I C A T I O N

This publication was produced by IDC Go-to-Market Services. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Go-to-Market Services makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com