INSIGHT

A New Approach to HPC Public Clouds: The SGI Cyclone HPC Cloud

Earl C. Joseph, Ph.D.   Steve Conway
Jie Wu

IDC OPINION

At the start of 2010, HPC clouds are being explored by many HPC centers but used by only a few. Typical cloud computing solutions will appeal to HPC sites as a way to handle workloads with minimal communications dependencies, as a way to handle overload ("surge") work without having to purchase more internal HPC resources, and as a way to develop and test applications and services before moving them into production use. SGI's new offering directly addresses the second and third bullet points below. It will be interesting to see if this changes the dynamics of clouds in HPC. In contrast, private clouds will likely see more substantial growth over the next five years in HPC as users take advantage of cloud-based technologies to provide a more seamless and easy-to-use in-house HPC environment. We are currently projecting that by 2015, HPC public clouds will represent less than 5% of all server CPU hours within HPC. There could be breakthroughs that significantly change this projection:

- If the U.S. government heavily pushes departments to use public clouds, for example, if the U.S. National Science Foundation were to provide grants to university researchers that allow them to directly purchase CPU cycles in the cloud versus buying computers at major NSF centers and then giving cycles to researchers

- If public HPC clouds are created with a lot of HPC-specific attributes, for example higher-performance interconnects, large SSI, GPGPUs, large memory nodes, full HPC software stacks, and appropriate security

- If ISV providers develop partnerships with public cloud providers and redesign their applications to fit within a cloud offering, for example scale-out HPC ISV applications with strong ease-of-use attributes

- Most important, if the pricing models close the gap with the price of doing the processing in-house

IN THIS INSIGHT

This IDC Insight first looks at how the overall public cloud market has evolved, then covers recent IDC research on the use of public clouds for HPC applications and takes a look at the new SGI HPC cloud offering, and ends with IDC's five-year forecast for public clouds in the HPC sector.

Filing Information: March 2010, IDC #222643, Volume: 1
Technical Computing: Insight

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

SITUATION OVERVIEW

While cloud computing has already made bigger inroads into the general IT sector and is growing quickly, it is only starting to gain a foothold in the HPC technical computing sector.

As with many new technology concepts, the term cloud computing is still in flux. Depending on who you ask, cloud computing can refer to the simple outsourcing of compute cycles and storage, to hosted/managed applications and services (e.g., hosted virtual machines), and to a future vision in which clients' computing environments are faithfully replicated to behave precisely like the clients′ in-house environments.

IDC Cloud Definitions

IDC has developed a set of definitions and examples of clouds for the overall IT sector. When most people talk about cloud computing, they usually refer to online delivery and consumption models for business and consumer services. These services include IT services — like software as a service (SaaS) and storage or server capacity as a service — but also many non-IT business and consumer services.

Many customers are not explicitly buying cloud computing, but the cloud services that are enabled by cloud computing environments. Cloud computing is hidden underneath the business or consumer service. And so, in our definitional framework, we distinguish between:

- **Cloud services:** Consumer and business products, services, and solutions that are delivered and consumed in real-time over the Internet

- **Cloud computing:** An emerging IT development, deployment, and delivery model, enabling real-time delivery of products, services, and solutions over the Internet (i.e., enabling cloud services)

In short, a cloud service is virtually any business or consumer service that is delivered and consumed over the Internet in real time (that's the five-second definition). Cloud computing, an important but much narrower term, is the IT environment encompassing all elements of the full "stack" of IT and network products (and supporting services) that enables the development, delivery, and consumption of cloud services.

IDC's Checklist for General Public "Cloud Services"

In the wide range of cloud offerings, the attributes that, in our view, define the new generation of commercial cloud services and provide the basis for those benefits are:

- **Offsite, provided by third-party provider.** "In the cloud" execution for most practical purposes means offsite (really, location agnostic). Specifying "third-party provider" simply highlights that the services we're focused on in our analysis are commercial cloud services.

- **Accessed via the Internet.** Standards-based, universal network access does not preclude service providers offering security or quality-of-service value-added options.

- **Minimal/no IT skills to "implement."** With online, simplified specification of services requirements, the need is eliminated for lengthy implementation services for on-premise systems that support the service (the service provider offloads this).

- **Provisioning.** Provisioning refers to self-service requesting, near real-time deployment, dynamic, and fine-grained scaling.

- **Pricing.** Pricing capability is fine grained and usage based. (As a convenience to some customers, providers may mask this pricing granularity with long-term, fixed-price agreements.)

- **User interface — browser and successors.** Browsers will evolve for a wider variety of devices and with richer capabilities, but the basic aspects of a browser **—** intuitive/easy-to-use, standards-based, application/service-independent, multiplatform **—** remain the attributes of cloud services UIs.

- **System interface.** Web services APIs provide a standards-based framework for accessing and integrating with and among cloud services (and Web services-based/enabled in-house systems). In our view, this is a critically important aspect of cloud services — that they provide well-defined, programmatic access for users, partners, and others that want to leverage the cloud service within a broader solution context.

- **Shared resources/common versions.** The shared asset approach improves supplier and customer economics; there is some ability to customize "around" the shared services, via configuration options within the service, workflow/process management among services, et al.

IDC's Definitions for General IT Public Cloud Computing

Since cloud computing is the IT foundation for cloud services, it consists of a growing list of technologies and IT offerings that enable cloud services, as defined by the attributes listed previously. A partial list includes:

- Infrastructure systems (e.g., servers, storage, and networks) that can economically scale to very high volumes, and preferably do so in a granular fashion

- Application software that provides Web-based UIs, Web services APIs, multitenant architecture, and a rich variety of configuration options

- Application development and deployment software that supports the development, integration, or runtime execution of cloud application software

- System and application management software that supports rapid self-service provisioning and configuration, load balancing, usage monitoring, et al.

- IP networks that connect end users to the cloud and the infrastructure components of the cloud to each other, leveraging network-embedded technologies for quality-of-service, security, and optimized application delivery
- For all of the above, pricing agreements for cloud services providers that scale technology costs with their cloud services volumes/revenue

In addition to supporting the unique IT requirements of cloud services, cloud computing offerings must also support the perennial "must haves" of enterprise IT environments, including manageability, reliability, availability, security, and price competitiveness. Further, because a growing number of enterprise customers will be running a portfolio of both on-premise and cloud-sourced systems, there will be increasing demand for IT offerings that span both on-premise and cloud-based systems.

HPC and Cloud Computing

IDC's Definitions for HPC Clouds

There are three major categories of HPC clouds today:

- Public clouds using industry-standard hardware and software (e.g., standard x86 nodes with Ethernet interconnects)
- Public clouds with HPC-specific hardware and software (e.g., larger memory nodes, IB interconnects between nodes, special processors, and HPC-specific middleware and applications) that include the traditional service-bureau-like offerings that are HPC focused, but are now renamed as clouds
- Private HPC clouds that are owned primarily by one organization and can also be geographically dispersed

What HPC Users Tell IDC They Need in Public Clouds

HPC Cloud Opportunities

From recent IDC studies of HPC users, a number of potential advantages are seen with public clouds. These include the ability to add HPC resources quickly, on short notice; the ability to add resources to handle peak workloads, so less needs to be invested in the user organization's internal datacenters; the ability to shift costs from capex to opex; a potential ability to lower ISV license fee costs; a capability to have broader use of HPC by scientists, engineers, and analysts by making public clouds dramatically easier to use (this is especially useful for SMBs); an ability to try new HPC technologies before buying (new systems, new processor types, new software, etc.); and access to additional expertise, talent, and skills.

HPC Cloud Concerns

HPC users are investigating the use of public clouds but have identified a number of concerns. Many HPC applications require fairly strong memory access at all levels, so clouds based on base clusters without extra capabilities tend to perform poorly on these applications. Many HPC applications are strategic applications to the organization and require a high level of security that current public clouds can't fully address. Many cloud offerings currently have pricing considerably higher than purchasing and housing HPC computers in-house. Many HPC applications have very large input and output data sets, making the data movement and storage more complex (and often slow). Some HPC users require that their application, or data, or results are fully removed and completely deleted at a point in time, and currently available public clouds are not always able to guarantee this. Finally, some current public cloud offerings are sold via the Internet focused on small purchases and aren't set up to handle larger purchases (e.g., running a 10,000CPU job that runs for five weeks). In addition, some users/sites don't have an easy way to get to a person to address technical questions.

Most Current Cloud Offerings Are Not Targeted to HPC Applications

As noted previously, at the start of 2010, HPC clouds are being explored by many HPC centers but used by only a few. Cloud computing appeals to some large HPC sites as a way to handle workloads with minimal communications dependencies and as a way to handle overload (surge) work without having to purchase more internal HPC resources. Cloud computing also appeals to some SMBs (e.g., engineering services firms) as a way to access HPC capabilities without having to purchase them or hire HPC experts to operate them.

The larger public cloud providers have put into place selling models focused on smaller organizations and smaller jobs. In many cases, they are not structured to even take orders or provide quotes for large and sometimes more complex HPC jobs. We expect that larger, more traditional HPC providers like HP, IBM, SGI and, perhaps, Sun will change this in 2010 by transforming their traditional services like service bureaus and HPC grids into HPC cloud offerings.

Developments in the private cloud area show that real demand is starting to ramp up for HPC cloud computing, and IDC believes that public cloud providers that address users' current issues with these environments will increasingly benefit from this growing demand. Some notable HPC private cloud deployments are already under way. CERN, home to the world's biggest particle accelerator (LHC), recently announced that with Platform Computing's help it is developing what may be the world's biggest scientific computing cloud to distribute data, applications, and computing resources to scientists in Europe and around the world. In February 2010, NASA announced plans to build a cloud environment with a Web portal to enable researchers to run climate models on remote systems provided by NASA. This will save NASA from having to help users build the complex models correctly on their local systems. It is important to remember that both of these advanced private cloud initiatives are aimed at handling production-oriented scientific work. And NSF is using Microsoft's cloud offering to provide HPC cycles to a broader set of researchers.

In the public cloud realm, Boeing and other companies have been remotely accessing the big supercomputer at Tata's CRL location in Pune, India, for several years. Multiple oil and gas companies are actively exploring public clouds for portions of their workloads — such as seismic processing — that don't require low-latency networking. Public clouds are also being evaluated, and in some cases already used, by organizations in the bio-life sciences, financial services, and digital content creation vertical segments.

SGI's New HPC Cloud Offering

SGI's Cyclone HPC Cloud Approach Is Called "Results on Demand"

SGI is in the enviable position of being able to provide *differentiated* cloud-computing cycles and services, especially access to a range of HPC systems that includes SMP capabilities as delivered by the company's Itanium-based Altix systems today, and that soon will include Altix UV (Ultraviolet) servers. SGI is planning to offer access to a wide spectrum of HPC system types and related resources.

SGI's Cloud Computing Is Targeted for Computational Science and Engineering

SGI's cloud is designed specifically to address the requirements of HPC applications that need an HPC server for performance. It can also include HPC applications if required. The SGI offering is designed to fulfill the part-time need for HPC resources that typically don't justify the cost of acquiring and managing a dedicated cluster. It also aims to provide better economics by having customers pay only for resources they use, when they use them — the utility model. SGI also intends to deliver results sooner than what can be achieved by ordering new dedicated hardware — SGI cloud resources can also be used as a temporary bridge until a dedicated cluster is up and running. The end goal, of course, is to allow customers to stay focused on science and research rather than on IT.

Other advantages include providing access to SGI applications experts; providing the flexibility to scale up (Altix), scale out (Altix ICE), or access hybrid, GPU-based clusters.

Two Cloud Services Model

SGI offers two models for using its cloud. First is software as a service (SaaS), where SGI sources and provides HPC applications to users over a network. The other model is infrastructure as a service (IaaS), which allows users to rent processing, storage, network capacity, and so forth, and run their own applications. The latter model is especially relevant for homegrown or already-licensed applications.

There is still more flexibility. SGI cloud services can be delivered on a pay-per-use basis, or one or more clusters may be reserved on a per-week basis. Customer data can be uploaded via a high-speed connection to an Internet backbone or via storage sent through FedEx ("sneakernet").

SGI's Cloud Services

The company offers a long list of primary HPC cloud service offerings, including:

- SGI cluster hardware provisioning
- Application tuning or benchmarking
- Virtual private network (VPN) or Secure Shell (SSH) access
- Management server node
- Networking and logical security

- 24 x 365 monitoring of critical components
- Hardware maintenance
- Customer account management
- Operating system and/or application configuration

FUTURE OUTLOOK

HPC Cloud Trends over the Next Five Years

At the start of 2010, HPC public clouds are being explored by many HPC centers but used by only a few. We expect that public cloud computing will appeal to some large HPC sites as a way to handle workloads with minimal communications dependencies and as a way to handle overload (surge) work without having to purchase more internal HPC resources. Cloud computing will also appeal to some SMBs (e.g., engineering services firms) as a way to access HPC capabilities without having to purchase them or hire HPC experts to operate them.

We currently see that there is a premium for industry-standard public clouds and an even larger premium for HPC-specialized cloud offerings. We also expect that in most cases HPC cloud offerings will be less virtual than general cloud offerings because of the need for running specific HPC middleware, specific application versions, and the prearrangement to have a larger number of cores ready to run the large job sizes. This will lead to a slower adoption of HPC clouds in general, with more rapid adoption in the SMB space and in targeted segments like financial modeling, bio-life science, oil and gas, digital content creation, and other applications that can be easily run on standards-based clusters. Clouds will also be a fit for HPC users that need to handle short-term peak projects while waiting for their next HPC system to come online.

We are currently projecting that by 2015 HPC public clouds will represent less than 5% of all server CPU hours within HPC. The amount will likely mirror HPC grids over the past five years, although the potential breakthroughs mentioned previously could significantly accelerate public cloud adoption.

In contrast, private clouds will likely see more substantial growth over the next five years in HPC, as users take advantage of cloud-based technologies to provide a more seamless and easy-to-use in-house HPC environment. Private HPC clouds allow security issues to be more directly handled at a lower cost. Similar to grid computing, we saw that only a small portion of the HPC market moved to grids, but most HPC systems incorporated grid technologies into the base system to improve system use and manageability.

SGI and other HPC vendors have distinct advantages over generic cloud services providers because HPC vendors understand the security, application profiles, and other special requirements of HPC users and, in SGI's case, because the company will provide access to a range of HPC-specific systems that can accommodate a broader spectrum of HPC applications and workloads.

LEARN MORE

Related Research

Additional research from IDC in the technical computing hardware program includes the following documents:

- *Massive HPC Systems Could Redefine Scientific Research and Shift the Balance of Power Among Nations* (IDC #219948, September 2009)
- *The Second PRACE Industry Seminar* (IDC #220029, September 2009)
- *China HPC Directions and Trends Looking at the Evolution of the China TOP100 List* (IDC #219952, September 2009)
- *The Race for the Fastest Computer Is Still On — Fujitsu's Petascale Project Plans* (IDC #lcUS21929009, July 2009)
- *I/O and Storage: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219121, June 2009)
- *HPC and Industrial Product Design: HPC User Forum, April 2009*, *Roanoke, Virginia* (IDC #219120, June 2009)
- *Petascale Computing: HPC User Forum, April 2009*, *Roanoke, Virginia* (IDC #219117, June 2009)
- *Alternative Processor Technology: HPC User Forum, April 2009*, *Roanoke, Virginia* (IDC #219118, June 2009)
- *HPC and New Energy Solutions: HPC User Forum, April 2009*, *Roanoke, Virginia* (IDC #219122, June 2009)
- *Rackable Systems Acquires Booster Rocket* (IDC #lcUS21774909, April 2009)

Copyright Notice

This IDC research document was published as part of an IDC continuous intelligence service, providing written research, analyst interactions, telebriefings, and conferences. Visit www.idc.com to learn more about IDC subscription and consulting services. To view a list of IDC offices worldwide, visit www.idc.com/offices. Please contact the IDC Hotline at 800.343.4952, ext. 7988 (or +1.508.988.7988) or sales@idc.com for information on applying the price of this document toward the purchase of an IDC service or for information on additional copies or Web rights.