Evaluating Cloud Infrastructure as a Service

There has been a proliferation of sourcing options and service providers for infrastructure as a service (IaaS). These service offerings come in a wide array of technical architectures, with a broad variety of technical features, and are delivered via a diverse set of business models. IT managers and architects need to understand the possible variations on these services in order to know what questions to ask in an RFP, and to identify the right solution for their needs.

Key Findings

• Traditional data center outsourcers, hosting providers and “pure play” cloud providers all offer forms of IaaS. Many IT organizations are also building IaaS offerings internally.

• There are no longer clean distinctions between “data center outsourcing,” “hosting” and “cloud IaaS,” and the boundaries between customer needs, delivery models, providers and markets will continue to blur over time. Although these are distinct markets, their services exist on a spectrum.

• IaaS providers are increasingly offering options for dedicated equipment and other forms of isolation, blurring the distinction between “public” and “private” cloud.

• The technical implementation of IaaS affects availability, performance, security and cost, and it is impossible to generalize about implementation approach and quality based on the type of provider.

Recommendations

• When evaluating an IaaS offering, look at both its technical characteristics and its business model to determine its appropriateness for your needs.

• IaaS offerings are evolving very quickly. Re-evaluate service providers and their offerings every six months.

• Expect to choose more than one provider if you have multiple use cases for IaaS. No single provider is superior across all use cases.

• Expect that the IaaS provider you select now as the best fit for a particular use case might not be your best long-term vendor for IaaS. The market is shifting unpredictably, and vendors are evolving rapidly.
ANALYSIS

Cloud computing is fundamentally transforming the hosting and outsourcing markets and, in turn, the hosting and data center outsourcing markets are influencing the pure-play cloud providers. Before the advent of cloud concepts, by 2005, hosting was already moving toward an IaaS model with the introduction of on-demand utility hosting services, and a parallel transformation toward the infrastructure utility (IU) was happening in the outsourcing market. The transformation of these markets has been dramatically accelerated by the interest in cloud computing; the end result is a gradual convergence of cloud IaaS, hosting and IU. This has created an increasingly confusing sourcing landscape for cloud IaaS.

Common Business Models for Infrastructure as a Service

IaaS can be insourced or outsourced, externally hosted with a third party or internally hosted within the customer’s own data center, and be public, private or community (shared with a restricted group of participants). See “Cloud Infrastructure as a Service: An Essential Overview” for more detailed definitions.

All IaaS offerings share the common trait of being on demand – additional capacity can quickly be provisioned or deprovisioned, and pricing is based on some usage metric. There are four common business models for outsourced IaaS, but not all these services are necessarily cloud IaaS. They are:

- **Self-managed IaaS.** In this model, the provider is responsible for the network, hardware, virtualization layer and related tools, such as monitoring, the orchestration engine and customer portal. The customer is responsible for everything above that, including the guest OS and all applications running on his or her virtual machine.

- **Managed hosting.** This model is identical to traditional managed hosting, but is on demand; not all offerings of this sort are cloud IaaS. Often, this uses virtual machines instead of (or along with) dedicated physical servers. In this model, the provider is responsible for everything that is in self-managed IaaS, but also provides additional services, usually on a customized basis. Typically, the provider will manage the guest OS, including activities such as patch management. The provider may also manage the middleware, such as application servers and database servers, including providing database administration services. The customer is responsible for his or her application.

- **IU for a particular application.** In this model, the provider is responsible for all the infrastructure components underlying a particular application, which will include everything other than the customer’s application code itself. Service providers with this kind of model often offer application management as well. This model often abstracts away much of the underlying infrastructure on behalf of the customer, and it may be priced in a manner similar to software as a service, based on units such as number of users, number of transactions or some other abstraction of application usage. However, many IU offerings are not cloud IaaS.

- **Foundation for data center outsourcing (DCO).** In this model, the provider implements IaaS as if it were going to be self-managed. However, the provider then offers DCO services on top of that infrastructure. In effect, this is traditional DCO, but with a dynamic, on-demand infrastructure, which may use virtual machines instead of, or in addition to, physical servers. Many offerings of this sort are not cloud IaaS.

These models are neither pure, nor mutually exclusive, nor are they the only possible models. Furthermore, they imply little about the technical characteristics of the service, and only a little about the business model. Also, providers can offer services in more than one of these types of models, and often do.

The Myth and Reality of IaaS Offerings

**Myth:** “Cloud” offerings are massively scalable and elastic, but are implemented in a “good enough” fashion, and are inexpensive, commoditized, bought with a credit card with nonnegotiable terms of service, without meaningful SLAs and without managed services. “Hosted” or “outsourced” or “IU” offerings are modestly scalable and offer some on-demand flexibility, and have higher-quality and more secure implementations, but are more expensive, are bought on longer-term customized contracts, and offer lower levels of service.

**Reality:** A provider’s origin – pure-play cloud provider, host or outsourcer – influences only its initial approach to the market and its service differentiators. It is impossible to make general assertions about what IaaS model a pure-play cloud provider offers versus what a host offers versus what an outsourcer offers, or generalize about their implementations. For instance, cloud offerings can commonly be bought with all the formality of traditional outsourcing, while outsourcers are increasingly moving toward offering global-class architectures via lightweight on-demand sales. Evaluate each service on its own merits; make no assumptions based on who is offering it.

© 2011 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. or its affiliates. This publication may not be reproduced or distributed in any form without Gartner’s prior written permission. The information contained in this publication has been obtained from sources believed to be reliable. Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This publication consists of the opinions of Gartner’s research organization and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice. Although Gartner research may include a discussion of related legal issues, Gartner does not provide legal advice or services and its research should not be construed or used as such. Gartner is a public company, and its shareholders may include firms and funds that have financial interests in entities covered in Gartner research. Gartner’s Board of Directors may include senior managers of these firms or funds. Gartner research is produced independently by its research organization without input or influence from these firms, funds or their managers. For further information on the independence and integrity of Gartner research, see “Guiding Principles on Independence and Objectivity” on its website, http://www.gartner.com/technology/about/ombudsman/omb_guide2.jsp
Every provider architects differently, offers different features, offers a different assortment of managed services, and has a different business model. A given provider may have several different types of offerings. The marketing terms attached to a service do not necessarily imply much about either its technical characteristics or its business model.

Furthermore, technical characteristics and business model are not necessarily coupled. A “global class” implementation does not imply that the service will be commoditized, inexpensive, on-demand, and come with no contract, paid for with a credit card. Nor does an “enterprise class” implementation imply that the service will be expensive, offer limited scale and elasticity, and have negotiable long-term contracts. A provider might create an IaaS implementation that does not have significant scale and elasticity from a technical point of view, yet be willing to allow users to buy that service on demand at the same low prices that a provider with a global-class implementation offers. Conversely, a provider might create a global-class system but provide a complex, highly differentiated offering, and require that customers buy it on long-term contracts.

Public, Private or Hybrid?

Fully shared cloud IaaS offerings are the most commonplace, but IaaS providers are increasingly offering outsourced private cloud services. These solutions utilize the same architecture and tools that the providers use for their public cloud offerings, and are generally “cookie cutter” rather than customized, but are deployed on hardware that is physically dedicated to a single customer. Regardless of whether the offering is public or private, similar architectural considerations apply.

For customers who do not need highly elastic capacity, dedicated resources may not be significantly more expensive than shared resources. For the provider, the cost to provide the base service is similar, as the hardware resources used are predictable and constant. However, some features, such as automated failover between resource pools in different data centers, might not be cost-effective when the resources are dedicated, and therefore would need to be sacrificed in this model.

The cost differences between dedicated and shared resources are likely to increase over time. For instance, today, service providers generally do not actively manage their capacity pools and power utilization by consolidating running compute and shifting it to the most efficient physical location; such tactics will help to exacerbate the cost-efficiencies of shared infrastructure.

We believe that the differences between private and public clouds will blur over time, as customers become increasingly comfortable sharing resources in a multitenant fashion, and as service providers implement the technologies necessary to provide temporary isolation of resources for a single customer. We expect customers to be much more comfortable sharing network and storage infrastructure than they are compute infrastructure. Therefore, we expect service providers to implement business rules in provisioning that allow customers the option of “temporarily private” compute resources. In this type of arrangement, a customer with this service would be guaranteed that when he or she provisions a virtual machine, it will not share a physical server with another customer’s virtual machines. This would potentially waste some of the resources of a single physical server but, as an overall percentage of a typical customer’s usage, it would be trivial, and could readily be made up for by charging somewhat more for this option.

Hybridization – mixing shared and dedicated resources, and interoperation between multiple infrastructure environments – will be the most common form of infrastructure in the future. Many providers already allow interconnection between their IaaS offering and a customer’s colocation or dedicated hosting environment, and can allow a customer to use private or virtual private network (VPN) connectivity to extend its internal data center’s network to the IaaS environment. This allows a customer to choose the level of resource isolation appropriate to each particular need and to interoperate seamlessly between those different environments.

In the future, “single pane of glass” management – a single unified management tool or set of management tools that can manage infrastructure across multiple environments – will be a key customer requirement, but such tools are only in the embryonic stage at present. Strategically, VMware intends its vCloud initiative to provide unified management tools across service provider and internal IaaS environments.

Geographic Considerations

When purchasing cloud compute IaaS, you will normally know where your infrastructure is located; you don’t need to worry about your infrastructure being migrated to arbitrary locations. It’s located in specific data centers, and normally, unless otherwise explicitly specified, your workloads will always run in the single, specific data center that you have chosen.

Cloud IaaS is evolving differently in different regions of the world. The options that you have available for services will depend on the country in which you want your infrastructure to be located, and what other restrictions you have governing the country of origin for the service provider.

U.S.-based providers have the most sophisticated services and the greatest scale, and often have global IaaS footprints. However, some customers do not want to use a U.S.-based provider, even in a data center outside the U.S., because they do not want to be subject to the provisions of the USA Patriot Act (the U.S. anti-terrorism law). Some U.S.-based companies may also prefer to use Canadian-based providers and data centers to avoid the Patriot Act and other U.S. laws.

Both global and regional service providers often have a highly limited number of locations in a given region. While they are gradually increasing the breadth of their data center deployments in Europe and Asia, these deployments are typically concentrated in just a handful of places – notably London, Dublin, Amsterdam, Hong Kong, Singapore and Tokyo. There are many countries in which there are no cloud IaaS data centers. If you have a regulatory requirement to keep your infrastructure in-country, your service provider choices may be highly limited, or there may be no cloud IaaS options for you at all.
**Purchasing a Cloud IaaS Offering**

Cloud IaaS is typically purchased in one of the following ways:

- **Online.** You sign up with a credit card online, click through a standard agreement, and have instant access to the cloud. You make no commitment whatsoever.

- **Zero-commit contract.** You negotiate a contract with a service provider, which usually specifies a pricing and discounting structure, but do not commit to any specific amount of purchased services. You may need to pay some form of fee to be a customer of the service; typically, this is something trivial, such as $250 per month. The typical contract length is one year.

- **Standard contract.** You negotiate a contract with a service provider, and commit to a certain amount of revenue. However, you can use any combination of cloud services to reach that revenue commitment. You also usually agree to a discounting structure for higher volumes of services. The typical contract length is one year, but some providers may try to encourage you to sign two- or even three-year contracts.

Cloud IaaS contracts are usually quite brief – a few pages at most. Contract terms are, however, always negotiable. Furthermore, even if a provider has an established and publicly published discounting structure, you may be able to negotiate better discounts as part of a contract.

**Evaluating a Cloud IaaS Offering**

When evaluating a cloud IaaS offering for computing, look at both its technical characteristics as well as its business model to determine its appropriateness for your needs. The following reports examine the spectrum of common possibilities and the details of how these services are implemented:

- **Compute** ("Cloud IaaS: How Compute Resources Are Delivered").
- **Storage** ("Cloud IaaS: Adding Storage to Compute").
- **Networking** ("Cloud IaaS: Networking Options").
- **Security and compliance** ("Cloud IaaS: Security Considerations").
- **Availability, performance and SLAs** ("Cloud IaaS: Service-Level Agreements").
- **Service, support and portal capabilities** ("Cloud IaaS: Service and Support Models").

When considering these services, keep in mind that although IaaS tries to remove many concerns from the user of the service, the underlying implementation still strongly affects the technical traits of the service. There is considerable variance in service provider design goals, the quality of the technical implementations, and the cost-effectiveness and the value for money of those implementations.

**Evidence**

Data for this series of notes was obtained through primary research in the form of vendor surveys and interviews, as well as secondary research in the form of public information and Gartner client conversations.