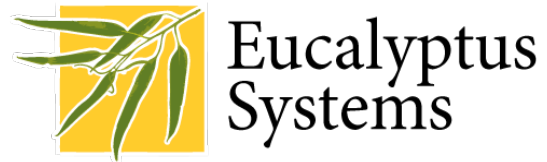


Five Steps to Enterprise Cloud Computing



Eucalyptus Systems, Inc.
© 2010

Overview

Significant technological advances are often made during periods of crisis and change. Thus it is unsurprising that today's CIOs and IT professionals, confronted with extraordinary challenges—spiking energy bills, underutilized data centers, accelerated data growth—during a time of restricted capital and economic uncertainty are gravitating towards innovative efficiency-enhancing technological models. Cloud computing is one such model. Originally proposed as a public utility computing model, “on-premise” or “private” cloud computing is emerging as a new technology for the IT-managed data center. It deploys as a complete platform for supporting scalable applications in a way that improves the efficiency of both IT management and operations.

This paper discusses the use of cloud computing in the enterprise data center, the potential friction points associated with the adoption of cloud computing and steps to take to initiate the development of an enterprise “private cloud” as an IT optimizing technology. In particular, it focuses on the operational and IT processes within the enterprise and how private and hybrid clouds can bring about

greater efficiencies within the enterprise data center.

What is a Private Cloud?

A private cloud is a software infrastructure that enables end-users to acquire, configure, and ultimately release data center resources on-demand, using automated self-service tools and software services within an enterprise's data center. One of the easiest ways to understand how a private cloud functions is by analogy with web-based e-commerce.

Today, customers expect to be able to shop for and purchase goods and services via the Internet. Successful e-commerce companies (e.g., Amazon.com, Google, eBay, etc.) implement highly scalable web services that are designed to allow as many customers as possible to make separate purchasing or rental transactions simultaneously. Furthermore, to keep sales overhead as low as possible, these e-commerce venues are fully automated and self-service. That is, the goal is to have the web services and the infrastructure (and not a sales person or support person) handle the vendor side of each transaction automatically.

By analogy, a private cloud is a service venue that allows end-users (customers) to search for (shop for) compute infrastructure that is customized to their specific needs (products), to acquire that infrastructure, and when it is no

longer needed, to release it back to the IT organization. In the same way that an e-commerce site must support the transactions put forth by many simultaneous customers, a private cloud must be able to scale to handle many simultaneous end-user requests and commands automatically, without human intervention. Similarly, to keep management overhead to a minimum, private clouds support self-service interfaces and tools so that the cloud services (and not system administrators) implement each user's request directly and automatically. The "product" in this rental analogy, is typically a "virtual machine" that has either been pre-configured with a specific set of software applications, or can be customized by the end-user directly once the acquisition transaction is complete.

Gating Concerns: Operational Changes, Governance, and Costs

For the IT organization, a self-service approach to infrastructure management can offer organization efficiency gains, but not without consideration for potential changes to operational processes, governance policies, and cost structure.

Operational changes

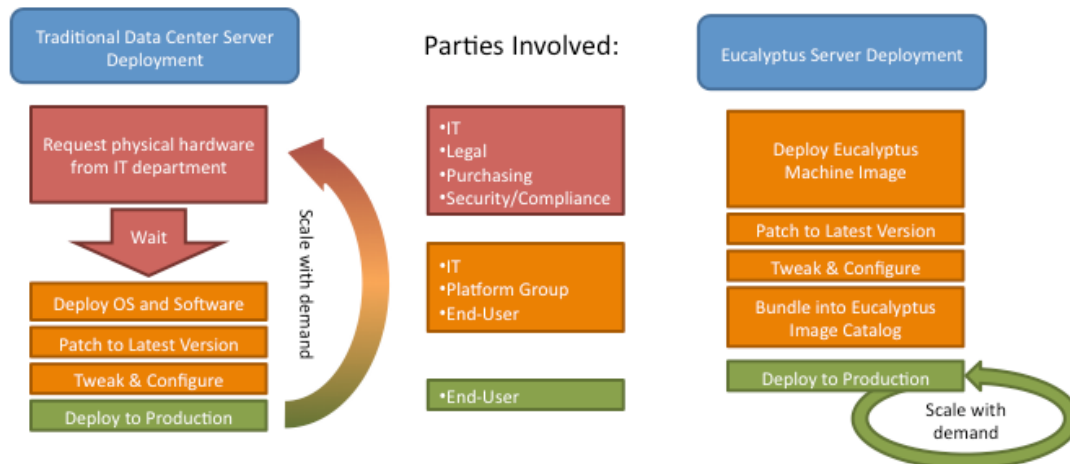
Achieving the full efficiency benefits of a private cloud typically requires a change in the activities and

responsibilities of both users and IT staff. Users must develop the skills and knowledge necessary to operate a self-service resource-provisioning interface. Note that without a cloud, resource provisioning is usually a personnel-intensive activity in which written requests for resources must navigate the organizational structure from end-user to IT professional, and back again. Often employees from different organizational units become involved during different phases of a request (requirements gathering, budgeting, security, recharge, etc.) The advantage of this approach is that specialized personnel can each focus his or her ability on a particular subtask necessary ultimately to allocate a new resource to a user. The potential disadvantage is expense (in the form of the personnel overhead) and delay (which manifests as lost productivity) between the time of the request and when it is satisfied. A private cloud automates the mechanism for provisioning new resources to individual users. The resources are virtualized (i.e. a user is allocated use of a collection of virtual machines (VMs) rather than a set of physical servers) so that the cloud infrastructure can "site" them (allocate them to physical servers) at the behest of the user. Furthermore, the user does not control where the VMs run but instead specifies a quality of service expectation with a Service Level Agreement (SLA) that the cloud infrastructure must respect for the user's VMs when making decisions about where they are to be sited.

The advantage of this self-service approach is that it is fully automated, making it possible to provision full machine, network, and storage collections in minutes, transactionally and simultaneously, for a large user community. The disadvantage is that

additional complexity is added to the duties of the end-user who becomes the only human in the provisioning loop. All of the provisioning functions need to be handled by the cloud automatically or managed by the end-user directly (**Figure 1**).

Figure 1. Operational Changes. The diagram below compares the server deployment process in a traditional IT organization to that of an organization with a fully implemented Eucalyptus private cloud. In a traditional IT setting (left-side of diagram) each request is subject to the actions of personnel across multiple departments (e.g. IT, legal, purchasing, support, security). The time required in manually performing these tasks combined with differing departmental priorities, requirements, and constraints can significantly delay resource deployment.



With a Eucalyptus private cloud (right-side of diagram) repeated iterations of this personnel-intensive deployment process are eliminated. Once appropriate organizational approvals and the cloud architecture are in place, users simply deploy and scale their own virtual resources (i.e. machines, application stacks, network, storage) on demand via an automated self-serve Web API. IT staff activities shift to optimizing cloud performance and enhancing the end-user experience with image creation, archival maintenance, interface customization, virtual network oversight, and capacity planning.

With a cloud, then, the separation of concerns associated with resource provisioning changes. The cloud infrastructure itself implements the mechanisms necessary to automate the process such as security credential management, accounting and recharge billing, network security policy enforcement, data integrity and provenance, etc. IT professionals concentrate on “baking into” the cloud (by constructing VMs for users and/or operating cloud policy interfaces) the policies that the cloud needs to enforce automatically. Finally, users must be able to choose the right set of VMs, network, and storage resources necessary to support a particular application that the cloud must provision on their behalf.

Governance

With new roles and responsibilities associated with a private cloud comes the need for new governance policies. In particular, methods of oversight, including management standards, operational guidelines, and best practices, which ensure the proper functioning and use of automated self-service activities are critical. Once provisioned, however, cloud-hosted applications behave much like non-cloud-hosted ones. The chief difference is the dynamic nature with which cloud-applications and users can change their provisioning profile.

In a cloud, for example, an end-user can change the rules governing “firewalls” isolating his or her VMs

from extra-cloud network traffic and from each other. Clearly the cloud cannot allow users to violate site-wide security policies so any changes must be automatically vetted against firewall rules for the cloud itself before they are permitted. Still, within the blanket security policies for the cloud, best practices (particularly for fault isolation and/or intrusion quarantine) dictate that inter-VM communication should be restricted. The cloud must audit and report the degree to which these practices are observed, but it is up to the organization to define policies and remediation strategies governing their use.

Is it a violation of cloud policy, for example, if a user drops an internal firewall rule between VMs for a short period (e.g., minutes) while debugging a network configuration problem? For how long should this debugging period be allowed? What is the response if the period is exceeded? How is that response implemented? Because users have control of (virtualized) infrastructure, and because the cloud can make changes to this infrastructure at machine speeds at the behest of individual users, a new set of governance policies and practices may be necessary.

Costs

Because private clouds depend, fundamentally, on virtualization technologies for isolation, they can implement server consolidation in the

same way that data center virtualization technologies do—by stacking several different VMs on each physical server. In cloud parlance, this capability is termed “multi-tenancy” and the cloud infrastructure must take care to isolate VMs owned by different users from each other when they share a common resource. The result, however, is the same as in a virtualized data center in that more computing, networking and storage can securely use fewer resources. The chief difference is that in a cloud, the cloud infrastructure and not the system administrator must automatically implement and manage multi-tenancy and do so while a large collection of users is permitted simultaneous access.

Another important cost-saving capability implemented by private clouds is the ability to temporarily exceed or “burst” over resource quotas. For example, if a particular marketing campaign suddenly generates larger than expected in-bound web traffic, it is possible for the cloud to “double up” less critical VMs temporarily until a more permanent decision about maximum resource footprint for the marketing VMs can be made.

However, even with multi-tenancy and temporary internal bursting, when the resource capacity of the cloud is exceeded user requests must be denied until sufficient resources become available. In this case, it is possible to consider bursting into one

or more public clouds, thereby forming a hybrid. Clearly, security policies must be in place to define the exact conditions and attributes governing what and when private cloud load can be burst into an external public cloud. Budgeting and cost controls, however, must also be in place.

Specifically, public clouds often charge for tenancy (e.g., rental by the hour), storage capacity and access frequency, and bandwidth in and out of the public cloud venue. It is critical to understand these resource usage characteristics on a per application basis to be able to predict the dollar cost that will be incurred when an application is burst from private to public cloud, and also when it is retracted and its data is transferred back to the private cloud.

Understanding this performance profile can be challenging. Moreover, if cost containment for hybrid operation is deemed to be critical, application development so as to minimize the expense of public cloud deployment may be necessary. That is, it is possible (but perhaps more complex) to develop applications in a way that deliberately minimizes their hosting expense in a public cloud so that if and when a public cloud deployment is triggered, the resulting cost is minimized.

Private-public cloud interoperability is also critical to a hybrid model. If the private cloud is to trigger a public cloud deployment automatically, the

application code and data must be portable between the private and public clouds under software control. Thus the private and public clouds must be able to interoperate.

Five Steps to Building a Private Cloud

If the efficiency gains through automation and self-service that private clouds offer are to be realized, IT professionals today are often interested in what steps they should take to build and deploy a private cloud. Because cloud computing is still nascent, the steps described below should be considered more of a guideline than a prescription but if followed, they will ultimately result in a functional private cloud.

Step 1: Adopt a Machine Virtualization Technology

Clouds, today, use machine virtualization as the basic technology for isolating resource usage between users. A “virtual machine” is a full operating systems stack that executes as if it is running on the hardware directly. In fact, each stack is running in a container that is exported by a software layer running underneath the operating system called a hypervisor. Systems services and tools can then be used by system administrators to manipulate virtual machines externally (e.g., move them start them, stop them, etc.) as if they are separate software processes while

the applications inside each “think” they are each running on a dedicated machine.

The first step in deploying a private cloud, then, is to choose a particular OS virtualization technology to use to implement cloud-hosted VMs. There are several choices, each offering a different price-point, feature set, and level of stability and reliability. Once a virtualization platform is chosen, the IT staff ultimately responsible for administering the private cloud can become familiar with the use of virtual machines, their failure modes, networking interactions, security interfaces, etc. as a platform for user applications.

Step 2: Profile Application Compute, Memory, and Storage Usage and Performance Requirements

One of the key impediments to deploying cloud applications surrounds the semantics associated with a more scalable and dynamic resource usage model, particularly for storage. Often compute and networking resource will change little in a cloud version of an application, but the cloud storage abstractions can be a source of non-trivial porting effort. Clouds must be able to scale both with resource count and concurrent user transaction rate. To do so, they implement storage abstractions that are different than the “standard” file system abstractions used by applications not running in a cloud. Porting applications to the

cloud requires a fundamental understanding of how these abstractions work. Further, to ensure that applications achieve the desired performance and robustness levels in the presence of dynamically changing cloud-provisioning activity, a clear understanding of their resource usage (particularly for storage) is needed.

Step 3: *Design a VM Development Consultancy*

Users and application development groups will need help in identifying, developing, and debugging the virtual machines they will ultimately use to host their applications. Often, private cloud administrators provide a base set of pre-configured VMs from which users may choose, particularly when the cloud is first deployed. These initial VMs need to be developed and cataloged in a way that allows users to understand their usage. As the cloud matures, users will want to create their own VMs either from scratch, or by modifying the “images” that have been pre-installed. To help users with these two new requirements, an organizational unit with expertise in operating system and machine configuration is needed. The cloud provides a self-service interface for provisioning and running virtual machines. Building and customizing virtual machines still requires infrastructure expertise, although because they are software abstractions, this expertise can be offered as a consultancy rather than as

a service provided by data center operations.

Step 4: *Develop Accounting and Recharge Policies Adapted to Self-service*

Automatic self-service carries with it a different set of incentives for resource usage than in a traditional data center setting. If users can simply acquire the machines they want to use, they may not always release them when no longer needed, or worse (if resource shortfalls occur) they may choose to hoard those they have been allocated. In a public cloud, rental is charged by the allocated hour so users who fail to return their resources are simply charged until they do. In a private cloud, where resource efficiencies are paramount and users bank accounts are not charged directly, an accounting and resource policy must be developed to incentivize responsible resource usage. For example, quotas on occupancy (e.g., leases) can be implemented. However, application termination or suspension due to a quota violation may not be the best response by the system. A policy that informs the errant user of a quota violation and discourages the quota-abusing user for repeated misuse is necessary for the system to be efficient.

Step 5: *Architect a Deployment and Deploy a Private Cloud Infrastructure*

Private clouds, like other data center hosted software services, can be

architected to leverage the compute, storage, and networking resources on which they run. Key architectural design elements include the mix of direct-attached and network attached storage, the topology of cloud service components with respect to network connectivity, the interaction between hosted VMs and local network security policies, and the management and routing of inter-VM network traffic. Ideally, private clouds are highly configurable so that they can take advantage of existing infrastructure if it is present, or use an infrastructure specifically designed to act as a cloud in the most efficient way possible. All private cloud platforms support a “universal” baseline configuration that can be used to get an initial deployment “up” and functioning. Like all data center infrastructure, a design and deployment plan will be needed to achieve maximum effectiveness in a production setting.

The Eucalyptus Open Source Private Cloud

Eucalyptus is a Linux-based open source software architecture that implements private and hybrid clouds within an enterprise’s existing IT infrastructure.

A Eucalyptus private cloud is deployed across an enterprise’s “on-premise” data center infrastructure and is accessed by users over enterprise intranet. Initially

developed to support the high performance computing (HPC) research of Professor Rich Wolski’s research group at the University of California, Santa Barbara, Eucalyptus is engineered according to design principles that ensure compatibility with existing Linux-based data center installations. Thus Eucalyptus can be deployed without modification on all major Linux OS distributions, including Ubuntu, RHEL, CentOS, and Debian. Further, Ubuntu distributions now include the Eucalyptus software core as the key component of the Ubuntu Enterprise Cloud.

The benefits of the Eucalyptus cloud

The Eucalyptus open source private cloud gives IT organizations the features so essential to improving the efficiency of an IT infrastructure, including the following:

- **Data center optimization.** Eucalyptus optimizes existing data center resources with consolidation through virtualization of all data center elements, including machines, storage and network. Eucalyptus is compatible with most widely used virtualization technologies, including Xen and KVM hypervisors.
- **Automated self-service.** Eucalyptus automates computer resource provisioning by allowing users

- to access their own flexible configurations of machines, storage, and networking devices as needed through standardized web service protocols.
- **Web services based.** Eucalyptus uses universally accepted Web service protocols internally, making its installation, operation, and maintenance similar to that of a high-quality e-commerce site.
 - **Scalable data center infrastructure.** Eucalyptus clouds are highly scalable, which enables an organization to efficiently scale-up or scale-down data center resources according to the needs of the enterprise.
 - **Elastic resource provisioning.** The elasticity of a Eucalyptus cloud allows users to flexibly reconfigure computing resources as requirements change. This helps the enterprise workforce remain adaptable to sudden changes in business needs.
 - **Open source innovation.** Highly transparent and extensible, Eucalyptus' open source core architecture supports value-adding customizations and innovations provided by the open source development community. The Eucalyptus open source software core is available for free download at www.eucalyptus.com.
 - **Hybrid cloud capability.** Engineered to emulate Amazon Web Services (AWS), Eucalyptus interacts seamlessly with Amazon public cloud services, including EC2 and S3, with no software modification required. This allows IT organizations to quickly “cloudburst” into the public cloud space without purchasing additional data center hardware during very large spikes in enterprise resource demand.

Eucalyptus Systems

Eucalyptus systems, Inc. offers enterprise-grade technology solutions that build upon the Eucalyptus open source software core with efficiency-enhancing additions, including customized user interfaces, enhanced automated provisioning with automated legacy support, image management, auto-scaling, auditing, metrics and accounting tools, and support for SLAs.

Now available, Eucalyptus Enterprise Edition, Eucalyptus EE 1.6, includes support for proprietary virtualization technologies, including VMware's vSphere, ESX and ESXi.



Eucalyptus consulting, training, and support services are available online at www.eucalyptus.com, via phone at 1 (866) 456-3822 (EUCA), via email at support@eucalyptus.com.

Or, visit our Eucalyptus open source community site at <http://open.eucalyptus.com>.

Eucalyptus Systems, Inc.
130 Castilian Drive, Goleta, CA 93117 USA
1 (866) 456-3822 (EUCA)
www.eucalyptus.com

Copyright © 2010
Eucalyptus Systems, Inc. All rights reserved.
Eucalyptus is a registered trademark of Eucalyptus Systems, Inc.