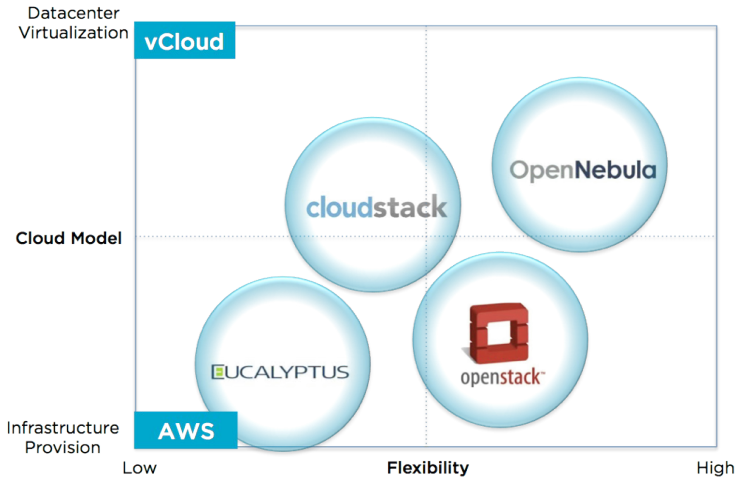
怎样用好Eucalyptus？

蒋清野（新浪微博：qyjohn_）
http://www.qyjohn.net/

来自OpenNebula项目的Ignacio M. LIorente最近发表了一篇题为**EUCALYPTUS, CLOUDSTACK, OPENSTACK AND OPENNEBULA: A TALE OF TWO CLOUD MODELS**的博客文章，从应用场景的角度分析了Eucalyptus、CloudStack、OpenStack和OpenNebula这四个云管理平台的不同点。Ignacio认为VMWare vCloud和AWS分别代表了数据中心虚拟化和按需获取计算资源的两种典型的应用场景，如上所述四个开源云管理平台基本上都是以VMWare vCloud或者AWS为参考原型但是在实现细节上又与参考原型有所差别。Ignacio将开源云平台与其参考原型之间的差异之处称为灵活性（Flexibility），并以数据中心虚拟化、按需获取计算资源、低灵活性、高灵活性为四象限将如上所述四个开源云管理平台放到不同的位置（如下图所示）。Ignacio进一步指出这个图例并不是为了说明某个开源云平台优于其他开源云平台，而是为了说明不同的开源云管理平台适用于不同的客户需求以及不同的应用场景。以目前的状况而言，私有云市场规模很大，客户需求以及应用场景之间的差别很大，并不存在一个能够通吃所有应用场景的云管理平台。未来Eucalyptus、CloudStack、OpenStack和OpenNebula这四个云管理平台之间既有竞争也会有合作，并在这种竞争与合作并存的关系中找准适合自己的市场和客户。



我基本上认同Ignacio M. LIorente的观点，就是不同的云管理平台适用于不同的客户需求以及不同的应用场景，并不存在一个能够通吃所有应用场景的云管理平台。出于同样的道理，Eucalyptus也有自己所擅长的应用场景，以及自己所不擅长的应用场景。作为Eucalyptus的员工，我自然希望各行各业的用户都使用Eucalyptus来搭建他们的私有云。但是为了能够充分发挥Eucalyptus的潜力，我建议所有潜在的客户首先了解一下Eucalyptus是什么（或者不是什么），Eucalyptus能做什么（或者不能做什么），以及应该如何规划、实施、使用基于Eucalyptus的私有云。

Eucalyptus是（不是）什么？

Eucalyptus是一个开放源代码的、与AWS高度兼容的云管理平台。以AWS为参考原型的各种云管理平台（例如OpenStack）都在某种程度上兼容AWS API，但是只有Eucalyptus将忠诚地兼容AWS API上升到企业战略与核心竞争力的层面。忠诚地兼容AWS API意味着客户能够在私有云环境中继续使用各种现有的与AWS API相兼容的工具、脚本和映像（AMI），能够在基于Eucalyptus

的私有云和AWS公有云之间迁移负载和数据，或者是将基于Eucalyptus的私有云作为开发测试环境但是将AWS公有云作为生产环境。

根据Ignacio M. LIorente的云管理平台四象限图，VMWare vCloud和AWS分别代表了数据中心虚拟化和按需获取计算资源的两种典型的应用场景。以数据中心虚拟化为应用场景的云管理平台通常采取自下而上的架构设计，旨在解决数据中心的复杂度问题；以按需获取计算资源为应用场景的云管理平台通常采取自上而下的架构设计，旨在通过简单高效的接口提供计算资源。设计理念上的差异，决定了一个云管理平台很难同时具备VMWare vCloud和AWS的种种特性。Eucalyptus与AWS的高度兼容性决定了Eucalyptus不是VMWare vCloud或者VMWare vCenter的替代品。Eucalyptus和VMWare试图解决的是不同的问题，适用于不同的应用场景，因此具有不同的功能和特性。常常有用户将Eucalyptus和VMWare vCloud或者是VMWare vCenter进行功能或者特性对比。他们没有意识到Eucalyptus和VMWare vCloud或者是VMWare vCenter完全不是同一类型的软件，是没有办法直接进行功能或者特性对比的。

值得一提的是，Eucalyptus是一个开放源代码的产品，但是按树公司并不为特定客户提供提供软件定制化服务。经常有一些潜在的客户问我们是否可以为其提供定制的版本。的确，作为一个开源项目的主要开发者，按树公司具备为特定客户提供特定版本的能力，但是按树公司通常不会这么做。为特定客户提供定制化版本意味着要对产品进行修改，也意味着使用定制版本的用户在升级到Eucalyptus后续版本时可能会遇到不可预知的风险。尽管在短期内定制化版本可能为客户解决了某些问题，但是从长期来看它所带来的问题要大于它所解决的问题。（Eucalyptus是一个开源项目，如果客户愿意并且具备相应的开发能力的话，当然也可以自己对Eucalyptus进行定制化。但是，用户自己对Eucalyptus进行定制化同样也会遇到升级的问题。）

Eucalyptus能（不能）做什么？

目前Eucalyptus的最新发行版本是3.2.1，它能够很好地用作开发测试环境，或者是用来支撑各种可扩展的Web服务。这两个应用场景的共同特点是大量地使用非持久性虚拟机实例（Ephemeral Instance），以及使用弹性块存储（EBS）来保存持久性数据。尽管Eucalyptus也支持从弹性块存储启动（Boot from EBS, BfEBS）的持久性虚拟机实例，但是由于架构设计方面的原因，在一个集群中存在大量BfEBS实例时整个集群的性能会有所下降。一个集群中BfEBS实例的数量越大，集群的性能恶化就越严重。因此，我们不建议客户在Eucalyptus上运行大量BfEBS实例。

Eucalyptus也不能很好地支持各种磁盘IO密集型应用，例如需要高速读写磁盘的数据库应用。严格地说，这不是Eucalyptus自身的问题，而是底层虚拟化技术的问题。目前各种虚拟化技术 - 例如VMWare ESX、Xen、KVM等等 - 已经较好地解决了CPU和内存的性能损失问题，但是在磁盘IO方面还是存在一定的性能损失。因此，我们不建议客户在虚拟机上运行各种磁盘IO密集型应用，包括负载较重的数据库应用。

在VMWare vCenter里面，系统管理员可以根据应用特征为应用定制网络参数。在Eucalyptus里面，如果系统管理员希望具备同样的能力，恐怕他很快就要失望了。为了以简单高效的途径提供计算资源，Eucalyptus尽可能自动化地管理整个私有云的网络配置，留给系统管理员自由发挥的空间不大。

Eucalyptus的硬件拓扑

接下来我们介绍几个典型的硬件拓扑结构，以帮助各位读者深入了解适合Eucalyptus的应用场景。在这些拓扑结构图中有一些缩写，含义如下：

CLC - 云控制器（Cloud Controller），Eucalyptus中的前端组件

Walrus - Eucalyptus中类似于Amazon S3的对象存储服务

CC - 集群控制器（Cluster Controller），管理一个Eucalyptus集群

SC - 存储控制器（Storage Controller），为一个Eucalyptus提供弹性块存储（EBS）服务

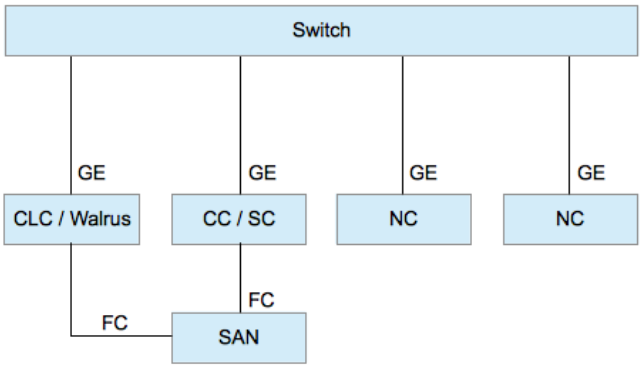
NC - 计算节点（Node Controller）

GE - 千兆网

10 GE - 万兆网

FC - 光纤通道

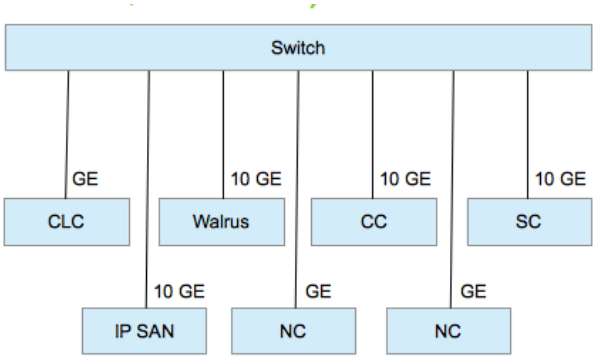
SAN - SAN存储



上面这个拓扑图展示的是一个只有一个计算集群的小型Euclayptus私有云。在这个私有云中，所有服务器都连接到一台千兆网交换机上，其中CLC和Walrus共同部署在同一台物理服务器上，CC和SC共同部署在同一台物理服务器上，SAN存储通过光纤通道连接到Walrus和SC服务器。如果为了进一步降低成本，SAN存储设备也可以替换成DAS存储设备。

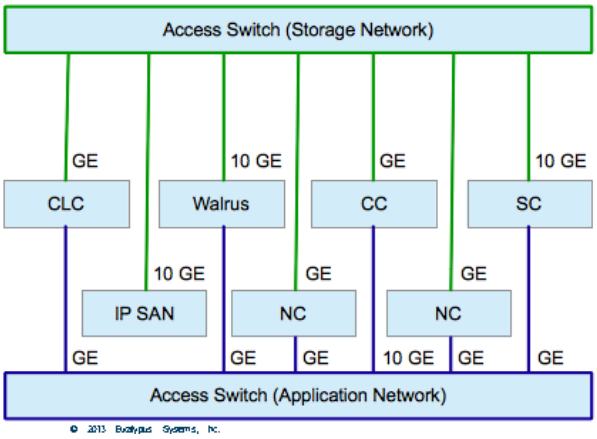
在这样一种拓扑结构下，Eucalyptus使用开源的iSCSI TGT驱动提供EBS服务。大量的实践表明，开源的iSCSI TGT驱动存在稳定性问题，在存储压力比较大的情况下会莫名其妙的崩溃掉。（这个问题不仅仅在Eucalyptus中存在，在所有使用开源的iSCSI TGT驱动的应用中都会发生。）EBS服务存在稳定性问题，意味着处于运行状态的虚拟机可能会突然访问不到挂载的弹性块存储设备，也意味着从弹性块存储启动的BfEBS实例会突然崩溃。这样的拓扑结构部署在对数据持久性要求不高的开发测试环境中的问题不是很大，但是我们不建议客户在对数据持久性要求较高的生产环境中使用。

除了EBS的稳定性问题之外，这个拓扑结构的瓶颈在于SC与整个计算集群之间的连接是一个千兆网。整个计算集群访问EBS服务的吞吐量受到千兆网带宽的限制，有效吞吐量大概在100 MB/s左右。假设每个EBS实例所造成的平均压力为2 MB/s，一个计算集群能够同时支持40到50个EBS实例。假设每个EBS实例所造成的平均压力为4 MB/s，一个计算集群只能够同时支持20到25个EBS实例。需要指出的是，4 MB/s相当于一个质量中等的U盘（USB 2.0）的吞吐能力，其性能远远不及老式笔记本电脑中常用的7200转SATA硬盘，而2 MB/s更是一个性能非常低下的极端情况了。如果客户希望在这样一种拓扑结构下大量使用EBS服务，建议将SC通过万兆网连接到交换机。（现在很多接入交换机都带2~4个万兆接口了，这样的改造成本不大。）经过这个简单改造之后，EBS服务的有效吞吐量一下子增长了10倍，基本可以消除由于带宽限制所带来的性能瓶颈。

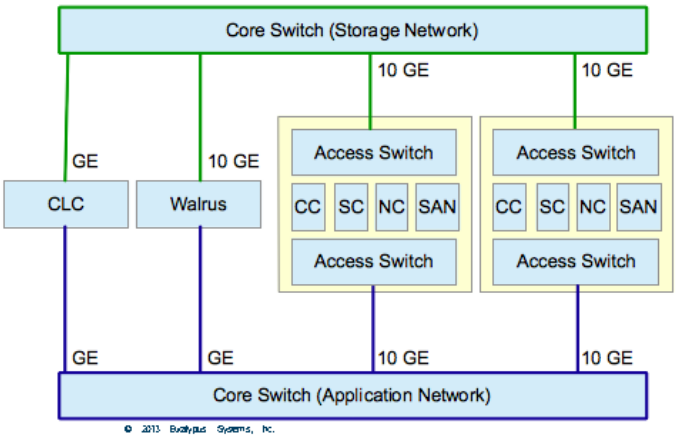


在生产环境中，我们建议客户使用基于IP SAN的存储设备来提供EBS服务。上面这个拓扑图展示的是一个可用于生产环境的Euclayptus私有云。在这个私有云中，所有Eucalyptus前端组件（CLC、Walrus、CC、SC）都部署到独立的物理服务器上，所有可能有大流量的组件（IP SAN、Walrus、CC、SC）都通过万兆网连接到私有云。目前Eucalyptus支持EMC、EqualLogic、NetApp等多个厂商的IP SAN设备，在这种拓扑结构下Eucalyptus使用官方支持的iSCSI驱动EBS服务，其稳定性和可靠性与开源的iSCSI TGT相比都有大幅度的提高。

即使如此，我们依然不建议客户在Eucalyptus上运行大量从弹性块存储启动的BfEBS实例。Eucalyptus的设计初衷是鼓励用户尽可能多地使用非持久性虚拟机实例，在这种情况下虚拟机磁盘映像被存储在计算节点上，虚拟机内部的磁盘IO不会对私有云的网络造成压力，运行在不同计算节点上的虚拟机基本上不会互相影响。由于从弹性块存储启动的虚拟机实例是持久性实例，需要频繁地与存储设备进行交互，对网络带宽的压力是很大的。因此，我们对客户的一般性建议是：（1）尽可能使用非持久性虚拟机实例；（2）在必须使用BfEBS实例的情况下，尽可能将操作系统盘做得比较小，例如10 GB；（3）将操作系统盘与数据存储盘分离，也就是首先启动一个尺寸较小的BfEBS实例，然后挂载一个尺寸较大的EBS卷用于存储持久性数据。



上面这个拓扑结构可以进一步加以改造，将存储网与业务网分离。这样存储流量就不会对业务流量造成影响，对私有云各个组件的健康监控也可以在存储网上进行。



当一个计算集群的容量达到极限的时候，可以往私有云中增加新的计算集群进行扩容，就形成了上面这个拓扑结构。

Eucalyptus的使用建议

前面我们已经说过，不同的云管理平台有不同的设计理念，适用于不同的应用场景。有些潜在的客户认为只要投资买了X硬件按照Y拓扑安装好Z软件就可以应付一切类型的应用，这种期望基本上是不切实际的。类似于AWS的云计算更多地是一种新的管理和分配计算资源的理念，而不是一种新的技术。使用类似于AWS的云计算服务要求用户了解一些基本的概念，并且通常需要对应用做一些修改才能够达到最佳的效果。我们给Eucalyptus客户提供的一般性建议包括：

（1）尽可能使用非持久性虚拟机实例；

（2）当必须使用BfEBS实例时，尽可能缩小磁盘映像的尺寸；

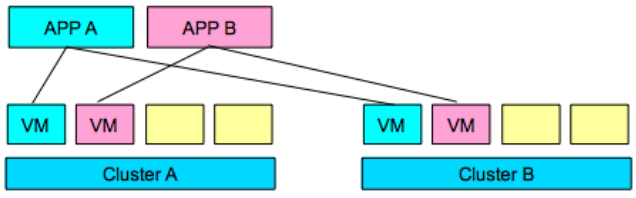
（3）使用EBS卷保存持久性数据，并且将操作系统盘和数据存储盘分离；

（4）不要在虚拟机上运行磁盘IO密集型应用；

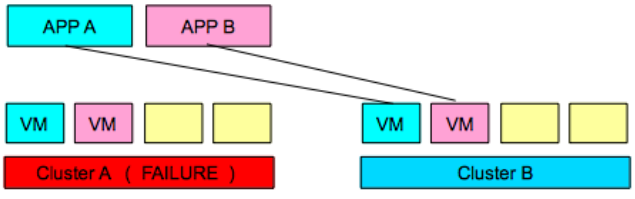
（5）不要对虚拟机进行纵向扩展，要通过横向扩展提高应用的处理能力；

（6）通过负载均衡实现应用的高可用性；

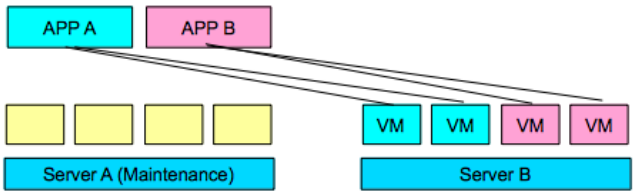
（7）避免虚拟机级别的在线迁移，当物理服务器需要进行维护时，使用Eucalyptus的维护模式。



如上图所示，我们建议用户将同一个应用部署到多个计算集群上。



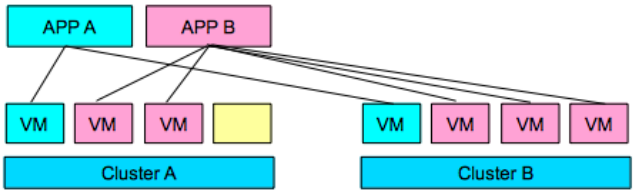
当一个计算集群发生失效的时候，应用依然是可用的，但是其处理能力降低了。



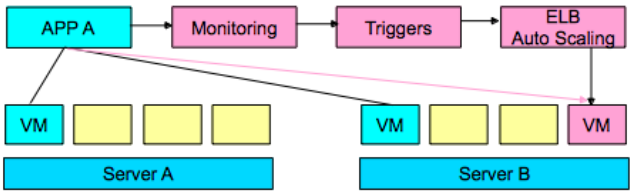
对于有计划的系统维护，可以先将应用的负载迁移到不需要进行维护的计算集群上，然后对处于空闲状态的计算集群进行维护。这样既保证了应用的可用性，又保证了应用的处理能力。

需要说明的是，上面我们所提到的"负载迁移"并不是指将一个处于运行状态的虚拟机从一个计算集群动态迁移到另外一个计算集群，而是在另外一个计算集群中基于同样的EMI创建一个新的虚拟机实例，并通过负载均衡设置将应用的负载重定向到新的虚拟机实例。到目前为止，Eucalyptus并不提供VM级别的动态迁移（类似于VMWare vMotion）的功能。对于类似于AWS的云服务来说最终用户与底层的基础设施是通过多个层次的抽象措施彻底隔离的。对于最终用户来说，他所看到的计算资源只有自己的虚拟机。至于他的虚拟机运行在什么样的物理服务器上以及该服务器上的负载情况，最终用户应该是一无所知的。因此，提供虚拟机级别的在线迁移功能，其实质是违反了通过抽象措施将最终用户与基础设施进行隔离的原则。

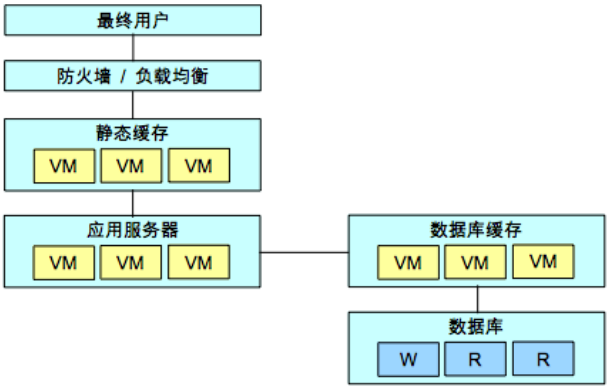
在Eucalyptus 3.3中即将提供一个称为维护模式的功能，允许系统管理员将某个计算集群中的特定物理服务器标志为维护状态。这时系统就会自动地将指定物理服务器上的所有虚拟机实例迁移到同一计算集群中的其他物理服务器上。这个功能允许系统管理员对计算集群中的特定物理服务器进行维护，同时又不必清空该计算集群中的所有负载。



当应用的负载上升时，通过横向扩展提升应用的处理能力。



在Eucalyptus 3.3中即将提供的实例监控（Monitoring）、弹性负载均衡（Elastic Load Balancing，ELB）、自动扩展（Auto Scaling，AS）功能，会使得应用的横向扩展更加容易。简单地说，实例监控功能允许用户对自己的虚拟机实例进行监控，监控对象包括虚拟机实例的CPU、内存、磁盘IO、网络IO等等；自动扩展功能允许用户设定一些简单的触发参数，并在监控参数达到触发参数要求的时候自动地创建或者是销毁虚拟机实例；弹性负载均衡功能则负责修改与应用相关的负载均衡策略，自动地添加新创建的虚拟机实例或者是移除旧的虚拟机实例。



综上所述，可以得到我们向客户所建议的一般性的应用架构设计。可以看出，这个架构设计与我们所熟悉的Web应用架构基本上是一致的。一个典型的Web应用，可能仅仅需要进行少量的改动（如果应用在设计之初就充分考虑到横向扩展的话，甚至是完全不需要改动），就可以充分利用Eucalyptus私有云的种种特性。

除此之外，在弹性块存储EBS服务的使用方面，建议各位感兴趣的用户读一读**Why EBS was a bad idea**这篇文章。尽管我并非完全同意文章中的观点，但是作者的许多观察和思考是值得云管理员和应用开发者深入思考的。

Eucalyptus的容量规划

现在各位读者已经大致了解了Eucalyptus是（不是）什么，能（不能）做什么，以及应该如何使用基于Eucalyptus的私有云。如果您觉得上面这些描述与您的应用场景相符合，并希望使用Eucalyptus来搭建您的私有云的话，我们建议您在动手之前做一些简单的容量规划。容量规划的基本方法，是将可预见的负载与物理资源进行映射，以确保私有云的容量能够承载应用和业务所带来的压力。一般来说，需要进行考察的参数包括CPU（物理核心数量）、内存、网络（吞吐量）、存储（吞吐量和IOPS）。

对于CPU和内存来说，可以通过标准化的VM产品类型（VM Types）进行简单的换算，其结果可以表达为特定的硬件设备可以支撑多少个某个类型的虚拟机实例。

对于网络来说，往往需要对即将运行在私有云上的应用的网络行为进行采样，获取其真实的流量特征，并在此基础上与私有云的网络配置进行对比。

对于存储来说，一个集群的容量往往受限于IOPS而不是吞吐量。上个月Eucalyptus公司的联合创始人之一Graziano Obertelli写了一篇题为**Will My Internet Be Faster**的博客文章，用较长的篇幅讨论了如何基于IOPS进行容量规划的问题。我将Graziano Obertelli的博客文章翻译为《我的网络会变得更快吗？》，可供参考。

服务等级协议

经常有客户问我们："用了Eucalyptus之后，是不是就可以保证我的云主机不会宕机了呢？"坦率地说，不仅Eucalyptus不能，其他的云管理平台也不能。

宕机时间，或者说不宕机时间，是数据中心领域的一个奇妙参数，通常被称为服务等级协议（Service Level Agreement，SLA）。服务等级协议是服务合同的一部分，通常用年度不宕机时间百分比（Annual Uptime Percentage）来表示。很显然，根据年度不宕机时间百分比，可以计算出一年内允许宕机的总时间。常见的SLA条款包括99%（一年可以宕机3.65天，也就是87.6小时），99.9%（一年可以宕机8.8小时），99.99%（一年可以宕机53分钟）。Linode的SLA条款为99.9%（一年可以宕机8.8小时），AWS的SLA条款为99.95%（一年可以宕机4.4小时），阿

里云的SLA条款为99.9%（一年可以宕机8.8小时），而盛大云则没有任何关于SLA的承诺。通常来说，更高的SLA条款意味这更高的硬件、软件和人力资源投入。到目前为止，尚未出现能够保障100% SLA条款的产品和技术。

其他

成功地运用云计算，需要硬件与拓扑、软件架构、应用场景、容量规划、服务等级协议等等多个方面的准备。我们在前面已经说过，类似于AWS的云计算更多地是一种新的管理和分配计算资源的理念，而不是一种新的技术。使用类似于AWS的云计算服务要求用户了解一些基本的概念，并且通常需要对应用做一些修改才能够达到最佳的效果。成功的云计算要求服务提供商和服务使用者都要为云计算做好准备。如果你只是想延用老的方式来使用云计算，基本上很难获得成功。