**大规模云虚拟机配置:挑战和方法**

**摘要：**全球数据中心市场规模近年来一直呈爆炸性。随着市场的发展，对虚拟资源的快速配置的需求，以支持弹性，易于管理，经济的云计算变得很高。快速配置大型的虚拟机（VM），尤其是保证服务质量（QoS）的关键。在本文中，我们系统地回顾现有的虚拟机配置方案，并将它们分为三个主要类别。我们讨论每一类的特点和研究现状，介绍了最近的两个解决方案，虚拟机和虚拟机 +，两者都能每秒提供给数百个虚拟机。

**关键词：**云计算IaaS，规模大，虚拟机配置

**1引言**

对基础设施即服务（IaaS）的快速膨胀的云，如亚马逊EC2和Windows Azure在美国，和阿里云的云、清云、UCloud、美团云在中国，培育大市场各类企业和个人的虚拟计算。不想投资在自己的计算基础设施的用户可以租用虚拟机（VM）云。此外，全球数据中心市场是巨大的：2013的数据中心的全球销售额为1430亿美元，预计将在2014达到1490亿美元；IaaS的市场份额在2013和2014分别是200亿美元和240亿美元。

与租用物理机传统的服务相比，虚拟机的规模可根据用户的需求进行调整。这种管理方式为用户提供了一种计算需求的弹性、易管理和经济的方法。

然而，IaaS云提供商面临着支持高效的虚拟机配置的挑战，特别是在一个非常大的规模。在虚拟资源配置效率对服务质量（QoS）直接影响IaaS云[ 1 ]。当用户请求和VMS已付款，这是他们能够使用VMS立即满意的运行性能非常好。因此，任何IaaS云提供商需要提供大规模集群系统尽可能快的保证运行时的性能。

然而，在实践中，虚拟机引导的时间可能比用户期望的时间长得多，根据虚拟机请求的规模和实际系统负载[ 2-4 ]。一个重要的因素，虚拟机启动时间是VM图像传输（VMI）从存储节点，通过数据中心网络，到选定的计算节点。例如，一个典型的VMI包含几个GB的数据，而虚拟机集群可能有几十或几百个VM，和总的数据可能涉及到几个TB的数据。因此，引导一个虚拟机集群将触发峰值数据访问。此外，该峰可以根据国标以太网处理时间。因此，如果数据峰值不能满足数据中心的基础设施，那么虚拟机配置的性能将会受到。缓解这个问题的关键是减少数据中心网络的数据传输总量。

在虚拟机配置问题中有三个挑战：（1）计算虚拟化技术与存储虚拟化技术之间的性能差距，（2）虚拟资源配置和运行时性能之间的权衡，以及（3）“成熟”的工业平台和新技术。

有虚拟资源配置的几种方法，可分为三大类：（1）图像引导配置，（2）记忆分叉配置和（3）轻量级虚拟化技术。此外，数据传输的优化可以适用于上述三种方法。最广泛使用的方法是图像引导的虚拟机配置，它侧重于图像管理，以减轻问题，在传输高容量的数据量。记忆的分叉方法菌种直接运行VM主机。轻量级虚拟化技术，如集装箱、码头工人，提出了一个不同于前两个范畴，但可能不成熟的实现快速配置解决方案。

本文的贡献是双重的。首先，我们提出了一个系统分类的状态的艺术虚拟机配置问题。其次，我们介绍我们的解决方案。当一个图像引导的方法和其他方法记忆分叉，都共享相同的存储基础设施。

论文的其余部分组织如下。第2节介绍的问题和大型虚拟机配置的挑战。3节介绍了配置方法的基本分类：图像引导的解决方案，在1节中，记忆分叉解第1和3款，新的虚拟化技术。还介绍了数据传输优化。4节介绍我们的解决方案，称为虚拟机和虚拟机加（虚拟机 +）。此外，该虚拟机溶液（图像引导）已经发表在以前的文章中虚拟机：快速配置大规模虚拟机集群。（TPDS，2014.11）[ 4 ]，而虚拟机 +（记忆分叉）是第一个发布的解决方案。最后，我们讨论了未来的研究趋势，在第5节和6节结束的文件。

**2关键问题和挑战**

**2.1关键问题**

IaaS用户需要一套完整的服务水平协议。用户需求的快速响应时间，所以配置延迟必须是短的。此功能有助于云服务的弹性，因为服务可以从一个租户切换到另一个租户。

虚拟机配置的延迟通常指的是用户在登录到虚拟机时提交请求的时间段。快速的虚拟机配置是具有挑战性的图像或规模的需求是非常大的。

**2.2研究声明**

虚拟机配置的问题是不相关的虚拟机调度问题[ 5-7 ]。虚拟机的配置通常是在调度之后。对于每个主机在局域网内共享相同的网络拓扑结构，并具有大致相同的传输延迟到服务器，不同的调度方案的影响的配置非常小。相比之下，虚拟机配置可以跨机架内的集群（带宽可能会有所不同，当传输交换机从一个机架内跨机架）。而调度的趋势把外来的虚拟机机架内，大规模的虚拟机配置可能需要用几架。跨多个机架的配置不是本文的一个关键性障碍。

**2.3挑战**

近年来，许多研究已经对弹性和IaaS云计算效率。IaaS平台与虚拟化技术构建。由于一台物理机器运行的虚拟系统大多在本地内存，虚拟机不能开始在一台物理机上没有必要的数据本地。对数据传输的需求是密集的当VMS的数量变得非常大，导致在存储服务器负担重。

第一个挑战是批量交付的虚拟图像。传统的解决方案使用一个文件或块设备来模拟物理磁盘，因此我们称之为虚拟映像，它比普通文件更大。一个虚拟的图像通常需要几个字节，有时甚至百万兆字节。交通量很大时，虚拟机配置生成的虚拟资源，大型主机。不幸的是，网络带宽可能会有所不同，网络拥塞是很难预测的，导致不可预测的延迟，在虚拟图像传输。此外，用户的操作系统版本的请求可能会不同的和自定义的虚拟图像与自己的需求。

其次的挑战是虚拟资源配置和运行时性能之间的权衡。快一些的配置方法可以使用不同的网络存储，将数据从远程设备和取决于网络拥塞，和伤害的表现。我们认为这种配置加速是基于性能的牺牲。然而，供应加速和性能之间的权衡，不能被量化的简单。

最后一个挑战是现有的工业实践和新技术之间的矛盾。在长期的历史范围，新的技术可能需要大量的变化，在现有的做法，从而可能在现实世界中的采用面对强大的阻力。例如，容器技术下降了传统的大图像（它与主机共享系统映像）。该方法从根本上减轻该问题涉及到大数据量的传输，但它可能需要很长的一段路要得到广泛认可。

**3分类的虚拟机配置**

有三种主要的方法来从一个小库中生成成百上千的虚拟机器：重量级和轻量级的方法。重量级的方法包括启动虚拟机模板的图像（图像引导的方法）和产卵的分叉的记忆正在运行的虚拟机（内存叉法）。轻量级的方法是引入轻量级虚拟机管理程序（轻量级虚拟化技术）。

引导虚拟机需要种子。不幸的是，种子（模板图像）通常是大。我们将在第3.1款中显示这种方法。产卵运行虚拟机需要一个活运行的虚拟机被克隆。不幸的是，内存的成本可能会很大，克隆和VMS的存储也需要稳定的、有效的，和个人的解决方案。我们将在第3.2节中介绍这种方法。新的轻量级虚拟化技术，如LXC，Linux的容器，已被引入到虚拟的世界，它有更少的数据前要准备配置。这一类的新技术将在第3.3款中引入。

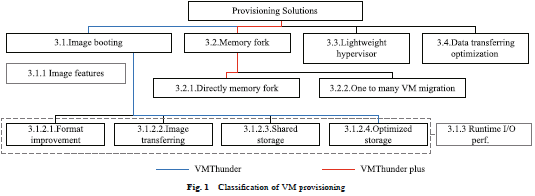
此外，将大量的数据从一个节点传送到许多可能需要其他的网络支持，例如，在以太网中广播或多播。这些技术提供了一个数据复制选项，复制许多副本具有更短的延迟，我们将显示这个（可选）优化在第3.4。值得注意的是，广播或组播的好处是不免费的，因为它们可能会影响其他主机或应用程序的性能。

图1显示所有分类解决方案，我们的解决方案了。我们将在下一小节详细的分类。

**3.1图像引导解决方案**

图像引导解决方案是最广泛使用的虚拟机配置方法。他们是透明的，目前的虚拟化工具，我们将介绍最基本的知识，图像格式，在第1。基于图像的格式，我们将显示在第2款中的图像管理的功能。

图像引导解决方案的核心问题是如何和在哪里存储的图像，以及如何提高效率的图像数据传送到每个虚拟机。在介绍图像格式和图像管理，以下四个小节将目前的图像引导的解决方案四大类，分别包括1款格式的改进，在2节传送图像，在3款共享存储，并优化存储在第4。格式改进的目的是提供一种新的格式，以节省数据量。图像传输方法的重点是有效的移动或复制的虚拟图像，每个主机和每个客人。共享存储方法存储在存储系统中的图像，并将图像提供给远程的每个客户。



最后，我们将给出一个定性的运行时间I/O在第3款中的这些解决方案之间的性能比较。

**3.1.1图像特征**

**3.1.1.1图像格式**

图像格式有三种主要类型：原始图像、增量稀疏图像和基于内容的图像。

原始图像是天真的，它将原来的位图上的物理磁盘文件或虚拟块设备。这类图像不需要I/O地址转换。这种格式有2个典型的实现，一个基于文件和其他基于块设备。文件为基础的原始格式是最幼稚的解决方案，但它可能面临的问题，嵌套的文件系统[ 8 ]。基于块设备的虚拟机使I/O直接从虚拟机到物理设备，但基于块设备的格式是不灵活的文件为基础的。

典型的增量稀疏图像QCOW2）和qcow23），都由QEMU [ 9 ]。用一个格式是消除磁盘空闲空间。当我们使用用磁盘，虚拟文件只占实际使用的虚拟磁盘空间。QCOW2提供副本上写的用磁盘性能，使虚拟磁盘文件快照显示表保持多个访问和支持备份/回滚操作版本。这两种类型的图像使磁盘地址压缩比原来的线性磁盘格式，并采取较少的物理磁盘空间。

相比与RAW格式图像的稀疏，增量释放不成文的块（零块）。然而，使用空间可能仍然包含字节的数据。此外，稀疏的虚拟格式引入的数据地址映射问题。数据请求需要一个翻译层来获得原始磁盘上的物理地址。这种额外的要求可能会降低性能。

基于内容的图像可以提高虚拟机配置。它是基于图像中的应用和文件，并减少了图像数据，我们将显示详细信息，在第1款的优点和缺点。

最后，一个相关的图像技术，快照技术，应该被提及。使数据跟踪、可靠、灵活，是一种有效的方法。它被广泛用于存储，数据库和虚拟化[ 10，11 ]。快照被定义为在特定时间内的系统状态。它有不同的实现，在本文中，我们提出的重定向写快照，这是一个最经常使用的副本写（牛）方法[ 11 ]。简单地说，当客户机虚拟机写入时，它记录了快照上的差异（与原始数据相比）；当它读取时，它读取快照，如果有的话，以及来自本地缓存或其他主机的原始数据来重新构造数据。

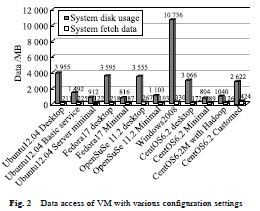
**3.1.1.2图像管理的特点**

图像的使用包括三个主要功能。首先，图像被使用：一个图像可以被分为不同的区域的访问模式的基础上。二，图像之间的相似性意味着不同的图像可能会有大量相同的数据。第三，在一个图像被部署和使用，它是单独使用从其他图像。

•形象的需求

当一个虚拟机启动时，虚拟机访问的数据块遵循一定的模式。此外，虚拟机相同的类型（即，相同的虚拟操作系统）检索相似的数据块的启动过程中，即使这些虚拟机可能有不同的配置设置。此功能减少了虚拟机引导的数据传输量，并显着提高了响应时间的云服务。

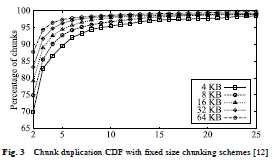
我们测试了流行的操作系统，包括窗口和各种系统的引导程序。如图2所示[ 4 ]，操作系统配置完整的视觉像桌面GNOME或KDE需要大约3 GB的磁盘空间，但只有约200 MB的数据访问过程中引导。操作系统与简单的配置，如文本的接口和基本服务需要较少的磁盘空间（约800兆），但仍访问约100的数据在启动过程中。在所有的测试中，我们观察到，并不是所有的数据要求的应用程序是获取的，只有一个相对较小的部分的数据块，而不是整个虚拟机的图像被转移过程中的虚拟机引导。



•图像是相似的

一些作品（例如，[ 12-14 ]）透露，图像共享大量的共同部分。重复数据删除技术（例如，[ 15–17 ] ）也利用这一现象减少冗余的云。不同的图像之间的相似性是由于流行的操作系统和应用程序。此外，用户可以创建新的图像的基础上存在的[ 12 ]。

要清楚地说明这个功能，我们使用的温度模型（热，温暖或冷）来描述图像的使用量在块级（逻辑图像分割在特定的规则）。该模型已被一些作品，例如，VDN [ 13 ] [ 14 ]和松鼠。图像块的温度测量其使用的虚拟机，或所有的用户在云中。图3显示了块复制的分布。可以看出，大多数的块有一个以上的重复。



直观地说，复制每一块的数量比例的温度可以缓解网络拥塞，并压缩冷块可以减少存储的使用。

•单独使用图像

虽然在一个云的图像共享共同的块，图像独立使用。这是因为一个图像上的多个用户的多址接入会导致不一致的问题。

此外，个别图像具有强烈的隔离和安全需求。虽然图像单独使用，安全问题需要特别照顾。例如，虚拟图像文件可以被未经授权的用户从较低层访问，因为较低的层可以给予高级特权。

对单个图像的版本控制是另一个重要的问题，这可能会增加图像的管理和处理复杂度[ 10 ]。

**3.1.2解**

**3.1.2.1图像格式的改进**

基于图像格式[ 20 ] 18–内容旨在improvevm配置性能。它根据应用程序和单个文件组织虚拟磁盘，以形成一组元数据。类似的想法也出现在图像，非内容例如，重复数据删除[ 17 ]将图像分成若干块，识别常见的块。相比于非内容的图像，基于内容的格式更准确地识别普通块，可以实现较高的压缩比。然而，这种图像格式需要显著改变现有的虚拟化技术和需要改变的成熟的管理程序。

只有大型公司像IBM [ 18–20]实现了基于内容的图像格式。幻影[ 18 ]和[ 20 ]改变FVD格式内容和改变图像引导的元数据。他们需要修改相应的管理程序，因此不容易扩展到新的平台。此外，他们介绍了一个新的数据地址转换层的基础上，其格式的策略。这可能会导致性能下降。图像格式越复杂，它对运行时性能的影响越大。

**3.1.2.2图像传输**

在云时代初期，亚马逊EC2采用图像传输方法来部署在EC2工具箱岩[ 21 ]虚拟机。一些早期的IaaS系统[ 6，22 ]也使用这种方法。图像传输是最幼稚的方法，是不相关的图像格式。

由于虚拟机需求的规模增加，配置延迟可能成为不可接受的。在亚马逊EC2的早期解决方案，超过20个虚拟机是对用户的要求[ 3 ]。同样的情况也出现在桉树[ 22 ]，和桉树的早期版本的表现表明，它没有花任何精力解决这个问题。

对等点对图像的传播是一种简单而有效的方法。它已被用于在雪球树[ 24 ]，wartel等人。[ 25 ]，对等网络[ 26 ]，比特洪流[ 27 ]，和虚拟机洪流[ 28 ]。此外，图像可以被压缩之前，它的传输。

图像传输的最突出特点是，虽然它可以分割成块的图像，它不能区分数据块，有助于虚拟机配置，从而需要将整个图像传输到本地磁盘。这个问题在虚拟机配置，然而，导致更好的运行时性能，因为整个图像抓取到本地磁盘的虚拟机启动前。启动后，不需要对图像进行进一步的远程访问。

**3.1.2.3共享存储**

共享存储支持虚拟机配置VIE的远程访问。这是类似于传统的远程启动，例如，物理机没有本地磁盘可以启动远程存储，称为无盘远程启动Linux（DRBL）[ 29 ]），通过一些网络协议连接到远程存储。最具代表性的一个使用etherboot或PXE从TFTP和NFS获得内存。

相比传统的远程引导，通过共享存储引导虚拟机更灵活。远程，共享存储看起来像一个黑盒子或一个资源池的虚拟机。共享存储有多种类型。在下面，我们将共享存储分为不同的类别。基于抽象层，有文件系统级和块设备级。基于计算节点和存储节点的耦合，有集中式存储和耦合存储。

•基于抽象层

1）文件系统级的存储解决方案的最新的方法是基于现有的分布式文件系统，如NFS，Ceph的[ 30 ]，PVFS，ZFS，光泽，全球FS，FS，FS Gluster，驼鹿。例如，投诉警察课[ 31 ] [ 32 ]和LVD是基于NFS；dvce [ 33 ]利用驼鹿FS；尼古拉[ 10 ]以PVFS [ 34 ]；松鼠用ZFS [ 14 ]。图像存储在文件系统中作为一个正常的文件，当虚拟机配置要求，虚拟机文件系统中的文件从启动。

作为共同的特点，大多数这种类型的解决方案使用当前流行的文件系统。然而，没有一个文件系统最初设计的虚拟机配置的目的，和文件系统可能是一个繁琐的基础设施的虚拟机。首先，分布式文件系统通常有一个固定数量的文件副本（大多为容错）。当请求被请求时，图像文件或图像块文件也应该被复制，这也会导致大量的数据传输。其次，分布式文件系统包括机制来解决数据的一致性问题。正如我们所提到的，虚拟机映像是独立使用的，因此，在分布式文件系统中的输入/输出操作可能是不必要的沉重的虚拟机配置。第三、文件系统通常包括日志机制来提高系统的可靠性。然而，这可以在虚拟机内完成，主机文件系统对这8个方面的影响较小。

总之，原来的文件系统的设计是不是最佳的虚拟机配置，并有更多的空间，进一步改善。例如，快照和缓存可以用于增强文件系统的解决方案，我们将此改进为优化存储的优化存储解决方案。

2）块设备级存储解决方案有些研究，而不是使用块设备。管弦乐队[ 35 ]和小提琴[ 36 ]采用了类似的堆叠系统的块设备。视差[ 37 ]在处理数据不一致的问题上花了很大的精力。

块设备级解决方案遵循虚拟机的隔离使用，并消除不必要的抽象级别（文件系统）。它们不会导致嵌套的文件系统的问题[ 8 ]，使虚拟机管理和使用更容易，更灵活。

在块设备的研究和实现，现在是很好的发展。许多技术已被用于和Linux内核所接受，如逻辑卷管理（LVM，设备映射器（DM），iSCSI。LVM和DM实现几个重要设备块的技术，如软RAID（线性/镜子/团队（0–10）），软加密、剥离、多径、快照。这些技术给了使用的能力，组装成不同的功能组件的原始块设备。这种类型的存储是可靠的，可扩展的，灵活的，易于管理。此外，块设备已被广泛使用，以提供用户的额外存储超出了系统的图像。

块设备逐渐变得流行。除了文件系统解决方案，Ceph [ 30 ]还提供块设备级的解决方案，称为RBD。和最受欢迎的块设备服务目前是Ceph的RBD [ 30 ]，sheepdog4），亚马逊EBS服务），迅速openstack6）。

•基于计算和存储节点的耦合

1）集中存储解决方案这一类中最流行的解决方案是存储区域网络（SAN）或网络附加存储（NAS）。SAN和NAS包含了大量的磁盘阵列内，提供可靠的连接和高带宽接口（S）的用户。

虽然企业级集中存储解决方案提供了非常好的可靠性，吞吐量和可扩展性，这可以提供高网络带宽的配置。然而，当规模足够大，立交桥的出口带宽，存储也是一个瓶颈。此外，企业解决方案通常是昂贵的硬件成本和专业化的服务，这是一个限制的普遍需求。

2）耦合的存储解决方案，如今，一个数据中心可能成为一个大型的SAN或NAS向远程用户提供存储在数据中心内的应用提供存储。在大多数情况下，存储主机也被用来作为一个计算节点，即，无论是存储和计算功能耦合在一起。这是由于在数据中心的商品硬件的广泛使用，这是更经济比集中存储。

在配置的情况下，耦合存储解决方案有更多的优势比集中式的。由于数据流是在节点之间传输的，良好的供应结构可以减轻单主机数据流的压力。例如，一个文件或一个数据块可以复制到多个副本中，或被拆分成小部分。此外，数据复制和拆分可以灵活地配置和需求。因此，耦合存储解决方案变得越来越流行，比集中式的。

然而，随着商品的硬件，主机与耦合存储和计算功能可能会面临其他问题。由于商品主机通常是不可靠的，一个失败的主机应该被另一台主机所取代，服务应该是透明的。因此，它对存储结构的设计和开发具有非常高的要求。

**3.1.2.4优化存储**

正如我们上面提到的，一般的存储系统是不是只能用于vmprovisioning并有进一步改进以满足虚拟机配置的特殊需要的空间。此外，VM图像上使用需求和遵循类似的数据使用模式。为了这个原因，一些作品已经设计出更好的解决方案，为图像存储，以提高虚拟机配置性能。VDN [ 13 ]，尼古拉[ 10 ]，和[ 31 ]，松鼠[ 14 ] [ 38 ] [ 39 ]，液体，内容块缓存[ 40 ]和LiveDFS[ 41 ]都属于这分类。

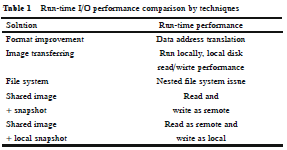
在优化的存储解决方案的最常用的方法包括剥离、缓存、快照。支持按需存取图像数据，一个是使用VMM（虚拟机监视器/虚拟机管理程序）来获取数据的透明，和其他图像分割成几块，转移，复制块子集。因为图像可以被分割，流行，高温块将有更多的复制或缓存，和冷兵器将重复数据删除以节省存储空间。有时，图像文件/内容可以被压缩，以减少网络流量。

优化的存储解决方案支持指定的虚拟图像存储。这种存储，但是，有它自己的挑战。例如，替换的副本和高速缓存的图像访问模式的基础上的实证分析。在某些情况下，如紧急事件处理，冷块可能会被要求，导致网络拥塞。

**3.1.3运行时I/O性能不同的解决方案**

由于许多解决方案可以采用几种不同的技术组装一个完整的系统，我们只是比较每个技术的关键因素，可能会影响运行时的我/输出性能最。定性比较列于表1。

本地/输出性能比远程输入/输出性能更稳定、有效，因为网络会导致长时间的延迟，而主机会在同一个集群中影响对方。基于这个原因，我们更关注是否运行时，我/输出操作可能是本地或远程。



**3.2内存叉解决方案**

3.2内存叉解决方案

克隆运行的虚拟机是虚拟机配置的另一种方式。我们把这种类型的解决方案分为2类：内存叉方法和虚拟机迁移。

**3.2.1直接记忆叉方法**

SnowFlock [ 42 ]和[ 43 ]是闪烁的具有代表性的两种记忆叉方法。

记忆的分叉的解决方案不需要时间准备申请，因为他们的克隆虚拟机作为原来的完全一样，包括内存的运行状态。我们想强调的是，VM叉方法在根本上不同于其他通用的虚拟机配置工具。虚拟机叉克隆一个正在运行的虚拟机复制到多个副本具有完全相同的运行状态。这种特殊的语义导致其应用中的许多限制。例如，SnowFlock必须特别小心因为使用时出现的多个进程在同一个虚拟机同时invokevm分叉可能冲突；孩子VMS是短暂的实体的内存和虚拟磁盘被丢弃，一旦他们存在，这样的孩子不能委托VMS复制和缓存的数据由于其短暂性[ 42 ]。

**3.2.2一对多虚拟机迁移**

虚拟机迁移是一种流行的技术，将虚拟机从一台主机移到另一台主机，无论是活的还是不活的。虚拟机迁移的本质是将虚拟机状态复制到远程主机，这样，状态也可以复制到多个主机上。代表一个是虚拟机分散[ 44 ]。它本来是用来设置虚拟机从一台主机到多台主机，因为它去重复的内存页在VMS并传输给多个主机通过组播。。

**3.2.3小结**

内存分叉的解决方案通常实现更快的配置。然而，这种类型的方法是受以下问题。

首先，要传送的虚拟机的数量数据是大于的图像引导方法。这是因为在虚拟机迁移过程中，运行虚拟机以及缓存可能会继续产生数据。此外，它是很难完全支持数据访问内存叉解决方案的需求。在内核中读取可能会引发更多的页面故障，如果页面错误无法及时处理，则会引发更多的页面故障。

第二，记忆分叉配置技术的实现是不容易的。现有的解决方案都需要管理程序的支持。和实施都需要内核级编程，处理工艺水平分叉或页面错误。

第三、记忆开始分叉解VMS与现有的虚拟机的状态，其中可能包含故障或潜在不稳定的模块。这是不是一个新启动的虚拟机安全。

最后，大多数现有的解决方案忽略了存储访问所造成的问题。当虚拟机了，只有内存数据催生了多个主机而不移动任何数据到相应的本地磁盘。这可能会导致更大的网络流量在网络上，因为以后的应用程序需要访问远程存储的数据。

**3.3轻量级虚拟化技术**

除了传统的管理技术，如KVM、Xen、VMware、一些新的、轻量级的虚拟机监控技术已成为越来越流行。最具代表性的是码头工人，基于Linux容器（LXC）技术[ 45 ]。

容器类似于一个虚拟机，其应用程序在开发人员不影响整个系统的基础上。然而，不同的是，容器没有配备一个操作系统，像一个虚拟机。它伴随着一个已经运行的主机，并运行像主人一样，像子系统。换句话说，它与主机共享资源，不仅是处理器/内存/网络/存储，而且还软件。该院有一个好的主机用户隔离、内核级实现，如C组（对照组）。另外，它使用AUFS主机存储和使用相同的文件或模块与主机操作系统。它有2个显著的好处。首先，资源利用效率更高。其次，容器容易创建和销毁。

新的轻量级虚拟机技术是一种完全不同的解决方案。包含整个操作系统的孤立映像可能永远不会被遇到，并且传统图像管理的部署瓶颈和挑战已经不再是一个问题了。

这种类型的解决方案也有一些问题。除了在主机上运行的一个容器外，容器不能支持其他操作系统。因此，灵活性成为一个值得关注的问题。其次，作为改进图像格式的方法相同，新的虚拟机管理程序技术不与已经被许多用户已经通过传统的虚拟机管理程序技术兼容。

**3.4数据传输优化**

如何减少数据传输量是虚拟机配置中的一个关键问题。广播和多播能帮助。随着广播，源节点可以发送数据到同一子[ 46 ]的所有节点，组播，源节点可以指定一组选定的接收器内部子网。

因此，一些虚拟机配置解决方案倾向于使用广播/组播生成虚拟机，像SnowFlock [ 42 ]，闪烁[ 43 ]，和[ 31 ]，Haizea [ 5 ]，等，由于数据集是类似的图像引导和记忆叉解。

请注意，广播和多播可能导致稳定性的问题，因为它们可能会影响其他主机时，大数据量的广播/组播在集群和数据中心

**4我们的解决方案：虚拟机和虚拟机加（虚拟机 +）**

根据分类，对虚拟机配置大尺度问题进行了研究，在学术界和工业社会实践。我们发现，最成熟的解决方案的配置延迟不能满足现实需求，因为最快的仍然需要分钟甚至数小时，延迟可能会增加显着的规模增加。此外，一些可能会达到短的延迟，但是，与难以忍受的运行时性能或不可接受的软件实现，这不能有助于实际的需求。然而，这是不可能的所有的解决方案的定量比较。我们将选择一些有代表性的解决方案来比较评价部分。

在本文中，我们将提出我们的解决方案的大规模虚拟机配置两。一是优化的图像引导的解决方案，为虚拟机，和另一种是memoryforking方案，命名为虚拟机加（虚拟机 +）。虚拟机是一种传统的图像引导的解决方案，使用专用的图像传输架构，而虚拟机 +使用内存叉VM具有相同的图像存储。

**4.1虚拟机**

虚拟机是基于以下的实验观察：当一个客户机虚拟机启动，数据块访问的虚拟机遵循一定的模式。和虚拟雷是一个通用的解决方案，可以适用于任何类型的存储解决方案。由于虚拟机是基于块设备的图像流结构，在该图像数据可以被访问的请求的要求，缓存在中继节点，并形成了一个对等加速结构。

**4.1.1基于块的图像流**

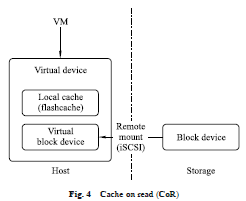
虚拟机使用传统的图像格式（例如，RAW格式），这是独立的存储类型。使用传统的图像格式，使我们的解决方案一般适用。该解决方案不需要任何图像拆分或剥离。

**4.1.1.1转移需求**

一个虚拟机访问虚拟磁盘的数据，这可能是本地或远程。数据的位置是透明的虚拟机中的应用程序，和数据传输的需求（即，只有当它是需要的）。

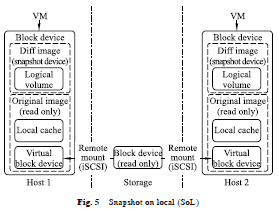
**4.1.1.2缓存读**

每个主机包含已获取数据的本地缓存，以避免不必要的网络传输。如图4所示，缓存存储已读取的数据。当相同的数据被请求时，数据将被复制到缓存中，以避免不必要的网络流量。为了更好的可靠性和性能，我们引入了两层高速缓存，其中可以发现在[ 4 ]的细节。



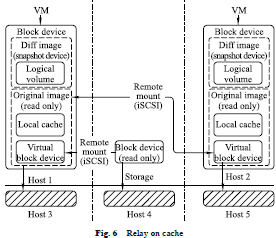
**4.1.1.3快照的地方**

为了让多个用户使用一个图像，并有效地管理图像，我们在每个主机上为每个虚拟客人创建快照的块设备堆栈。如图5所示，他的快照可以被写入，但原始图像被保留为只读。在本地的快照，我们的方法消除了存储服务器的复杂性，推送数据的一致性问题，主机。它大大减轻负担从存储服务器，使其扩展到同时支持数百个虚拟机。



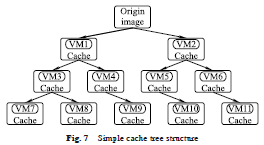
**4.1.1.4中继缓存**

如图2所示，每台主机上的缓存都包含原始图像的（部分）副本，并且只读取。因此，在逻辑上，该缓存与原始图像具有相同的状态，尽管它没有完全的数据。因此，我们可以使这个高速缓存服务的其他主机与原始图像相同。请注意，主机是另一个主机的上游主机，如果前者比后者沿覆盖树的根更接近。高速缓存可以被其他主机安装在图6所示。这种机制被称为高速缓存（中华民国）。如果本地或下游的读取请求未命中缓存，该读请求会触发另一个读请求到上游设备。这种级联的方法来处理读请求与基于树的对等流媒体方案完美地对齐。



**4.1.1.5缓存上的中继树覆盖**

在对等网络中，有各种方法可以组织对等网络。为简单起见，我们采用树结构虚拟机，包括明星（单层树），链，二叉树（2叉树（树），4 -树），和10棵树。图7显示了一个简单的缓存中继树结构。我们将测试和评估的性能和可扩展性的每一个树结构在4。



请注意，我们的应用程序上下文不同于传统的对等网络，对等点可以加入和任意离开。在我们的情况下，主机和它们的（网络）位置的数目是已知的，并且作为这样的我们不需要一个复杂的机制，动态拓扑管理。相反，我们可以预先设计任何树拓扑结构，并配置使用管理层壳脚本的拓扑结构。

**4.1.1.6 I/O性能优化**

中继树结构为每个主机提供一个缓存和若干快照。从逻辑上讲，每一个缓存和快照是相同的原始图像。所以，我们可以启动虚拟机的逻辑图像（即双缓存和快照）。物理上，每一对缓存和快照都需要从原始图像中提取数据。基于高速缓存中继结构，本次取的处理是通过读取前面的机制加速的，并且性能不受输入/输出路径长度的延长而不受。最长路径缓存性能描述和VM迅雷[4 ]和[ 47 ]说明raflow。

**4.1.1.7树覆盖的容错性**

在节点故障的存在下，该树结构可能不健壮。如果一个主机上的树失败，所有下游主机的主机不能再工作了。为了使树健壮，我们使用多路径来处理这个问题。每个节点记录到根的多个路径，在特定的路径，它的祖父，路径，其祖父，等等。当一个主机的直接上游主机失败，搜索是发出它的外祖父主机。如果外祖父还活着，那么它就用它的祖父主人的路。否则，同样的搜索被重复的根。以这种方式，与小的开销，记录多个路径，我们可以保证树的鲁棒性时，某些节点失败。

**4.2虚拟雷+**

虚拟机是一种增强版本的虚拟机，它可以有效地减少启动后的应用程序引发的虚拟机迅雷的延迟。关键的思想是，用以从模板虚拟机到大范围的应用程序初始化的应用程序的内存状态。虚拟机+使用相同的基础设施与虚拟机。它留下了图像存储的问题，并着重于内存数据流。因此，它也是一个存储独立的解决方案。相对于SnowFlock [ 42 ] [ 43 ]和闪烁，不需要修改程序，单独存放，并保证运行时的性能。

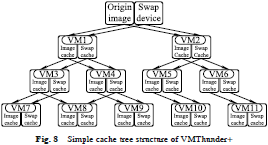
**4.2.1冬眠**

作为所有现代操作系统的本地机构，Hibernate技术的成熟和广泛应用。Hibernate是最初推出的电源管理。Hibernate技术动作中的所有数据存储到磁盘空间和停顿的运行系统。这可以释放所有的物理资源，如处理器，内存和网络，除了持久性存储。冬眠可以随时当操作系统正在运行，这个方法能够保持系统静态状态。当虚拟机重新启动时，它首先检查冬眠状态和恢复，如果有。

在持久化存储中有不同的保存内存数据的方法。例如，Linux使用的交换磁盘和Windows使用Hibernate文件（pagefile .sys）。

**4.2.2树结构**

我们可以建立一个相同的缓存继电器结构保存Hibernate磁盘/文件。实际上，这个交换分区和系统分区可以在同一个设备上。然而，控制结构简单（例如，交换分区可以通过记忆/ SSD加速），我们把他们分开。和树的结构如图8所示。在虚拟机 +结构有两个独立的保护树。例如，可以把数据从VM8 VM3和VM4分开。



基于高速缓存，每个虚拟机将建立它自己的快照。和所有的新写操作的交换分区将被重定向到快照，和所有的高速缓存设备可以被配置为只读。

**4.2.3大规模恢复**

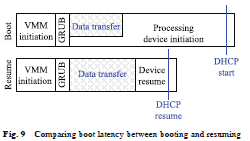
一样的地方冬眠恢复过程，当大规模的简历是预制的，在每个物理机每个虚拟机将恢复其本地快照和缓存对。在恢复，所有冬眠的内存中的数据将从从模板互换每个交换快照/缓存通过树结构对，最后恢复到内存。

重要的是要注意，每一个虚拟机的苹果地址应该被替换后，大规模的恢复，或所有的人将有相同的网络通信地址。我们使用热插拔技术卸载旧网卡和安装一个新的，在线的

恢复其他主机上似乎带来的排斥现象。然而，从冬眠的内存数据恢复基本上是相同的与主机之间迁移。因此，休眠/恢复方法只要主机之间的差异不超过迁移限制。

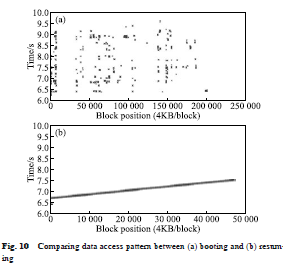
**4.2.4启动延迟**

图9显示了启动和恢复虚拟机的不同阶段。VMM启动GRUB阶段是VM的赔偿，这是不可避免的（这是一样的BIOS启动物理机启动）。GRUB的启动分区选择后，启动方法的靴子从”/“挂载点而恢复方法恢复从交换分区。引导方法从存储和加载/处理数据到内存中读取原始数据。的esuming方法只是负载休眠数据到内存。这两种方法都需要处理设备的问题，包括设备的初始化和设备的唤醒。最后，在客人vmobtains从DHCP服务器的IP地址，用户可以登录到虚拟机。



这2种方法在数据量和处理/设备延迟上也有不同。引导方法读取原始数据，然后处理，并且它传输的数据比恢复方法少。例如，CentOS 6.4minimal客读7.4 MB启动DHCP客户端获取IP的前。相反，恢复方法触发184.3 MB的数据交换设备（清洁系统无需用户应用）。尽管如此，数据不需要再进行处理，所以恢复方法可以节省相应的时间。

此外，图10说明了这两种方法的磁盘读取模式。X轴表示偏移对存储块（引导盘设置在第一个1G空间数据），和Y轴表示存取时间（从供应请求）。从图中，我们可以看到，启动方法会触发多个进程/线程存储的多个区域，恢复方法将按顺序加载数据。



**4.2.5运行时性能**

在虚拟机中，当虚拟机恢复后，所有的数据都将被转移到内存中。因为虚拟机和虚拟机都使用相同的数据流结构，它们的运行时性能几乎是相同的。对虚拟机运行时的性能是在[ 4 ]和运行性能优化的细节显示的是在[ 47 ]了。

**4.3实施**

虚拟机迅雷/虚拟机迅雷+的实施是高度模块化。VM雷实施包括各功能模块与管理脚本集成超过3200行的Python，Linux shell（bash）代码。虚拟机的数据流结构是基于虚拟机的脚本。

我们使用的设备映射模块，包括模块的快照，和多路径I/O模块。我们还集成了许多第三方模块像Facebook的flashcache7）（我们修改缓存实现本地缓存。），iSCSI启动器和目标，等原始图像被存储在一个主机在一个LVM卷。

虚拟机+需要的原始图像和交换设备配对，因为根据不同的图像交换会导致磁盘和内存之间的数据一致性问题，并引发内核恐慌。在我们的测试中，我们配置交换缓存在SSD器件

值得注意的是，Hibernate的功能（暂停）通常用于操作系统的电源管理和不适合的配置。由于休眠/挂在云的虚拟化环境的需求没有那么大，这些功能都不能很好的支持。例如，libvirt（10.2以前的版本）有错误造成的恢复虚拟机内核恐慌。

两虚拟机和虚拟机+需要的系统配置。最重要的是网络配置。克隆的虚拟机与原来的虚拟机具有相同的苹果机地址，我们使用热插拔技术，植入一个新的网络接口到新主机上的克隆。

**4.4评价**

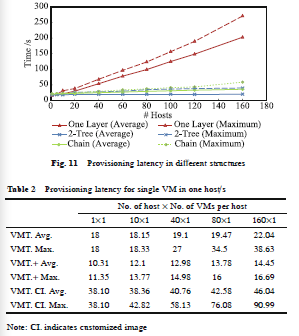
我们评估虚拟机VM迅雷和迅雷+在天津国家超级计算中心的集群NSCC天津（8）。我们已经部署了虚拟机多达160个主机在云计算集群，并部署虚拟机雷+上20的主机。

主机有双六核心处理器，48 GB的内存，以太网卡连接高速华为H3C S12 518字节 IP交换机，和2 TB的SATA硬盘。测试虚拟机 +主机配备SSD。虚拟机主机安装CentOS 6.2，和虚拟机 +主机CentOS 6.4。所有的主机都配置了LVM2，并格式化为ext4。虚拟机管理程序是KVM。客人VMS安装相同操作系统的主机（CentOS 6.2 / 6.4），并已安装QEMU I/O驱动程序。该虚拟机的配置包括4个核心虚拟处理器，8 GB的虚拟内存和100 GB的虚拟磁盘。此外，我们设置了一个比较自定义的图像（词），目前的情况时，配置不同的图像。它是基于原始图像和客人定制的应用程序安装（例如，adoop）。

我们记录启动时间自动发送一个HTTP请求从新启动的虚拟机，这是发送正确的DHCP客户端获得IP地址后。对于评估，我们测量的平均启动延迟和最大启动延迟。对于每一个测试场景，性能测试结果均获得10多个测试运行。

**4.4.1供应延迟**

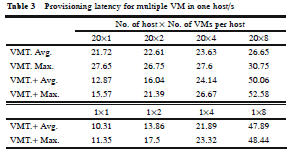
图11显示了平均和最大的虚拟机启动延迟。我们测试了在不同网络拓扑结构虚拟机 虚拟机，包括一层（星），树，和链，使用相同的图像数据172mb。在一个层的拓扑结构，所有主机直接从存储服务器获取数据。链结构代表了一个极端的情况来说明的好处，继电器缓存（中华）机制。结果表明，虚拟机在树更好的链结构比单层结构。这说明了对等网络流的好处，特别是在缓存（中华）机制的中继的利益。此外，对于定制图像具有424 MB安装数据，我们建立了一个树的拓扑结构，其结果如表2所示。它得到了几乎相同的延迟启动从本地磁盘和从虚拟机。表2也显示虚拟机 +绩效。这表明，虚拟机 +可以开机速度比图像的方法在单独的虚拟机主机上启动。



我们进行了实验测试，通过在一台物理机器上部署多个虚拟机。在这个测试中，我们选择4树结构来运行多个虚拟机（CentOS 6.2台式机）20台主机。测试结果如表3所示。结果表明，虚拟机 +失去优势，当虚拟机数量的增加。这是因为，从休眠恢复VM的多个虚拟机将引发更高的资源竞争。

**4.4.2配置掉队**

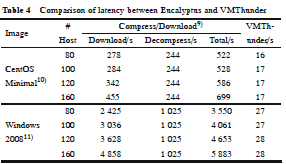
在实验中，我们观察到的95%多个虚拟机启动5秒后第一个靴子。我们把剩下的散兵游勇。我们发现掉队被均匀地分布在多个运行的主机之间，表明它们不是由一些缓慢的主机造成的。我们还标签的散兵游勇的启动顺序，并发现他们的启动顺序是独立的树的水平位置。这是很难确定的离散现象的真正原因。我们猜想，这可能是由随机因素引起的硬件或软件如一些后台进程运行在主机。然而，我们的实验表明，流浪者的数目是小的（小于5%的要求VMS）。在实践中，我们可以解决Straggler问题在工程，即延迟阈值提高后所需的VMS总数和杀死人。



4.5比较

我们比较虚拟机的性能与现有的其他虚拟机的配置方法。在[ 3 ]所示，在要求vminstances商业平台的上限。例如，EC2和Azure只允许一个用户请求多达20个虚拟机各一次，而Rackspace公司只允许一个。性能的EC2和Azure的大型虚拟机配置（即大于20 VMS）尚不清楚，因此很难比较虚拟机这些方法在大规模。

然而，我们能够比较虚拟机与其他流行的开源项目，如OpenStack、OpenNebula、Eucalyptus，，提供相同的功能像EC2。据我们所知，这些项目基本上忽视了大规模的供应问题，他们简单地处理这个问题的一些naivemethods如图像压缩与传输。这些方法中的启动延迟包括花时间下载，解压缩，而操作系统启动。下载时间VMS的数量按比例增加。即使使用一些对等的方法来加速下载阶段，在整个图像被下载之前，客人操作系统不能引导。表4显示了vmprovisioning延迟虚拟机和桉树。ROM的结果，虚拟机执行几个数量级的速度比桉树。



Ceph的[ 30 ]是一个开源的分布式存储系统。它可以提供分布式文件系统服务，分布式的块设备服务，以及分布式对象服务。在这里，我们将比较Ceph分布式块设备（CEPH由于块设备RBD），与虚拟机。Ceph的RBD是一个分布式的块设备服务系统没有中心控制节点。它实现了数据的多副本，剥离，如文件对象结构的负载平衡，安置组和对象存储设备等的块设备，可以直接使用虚拟机，同样在虚拟机块设备。

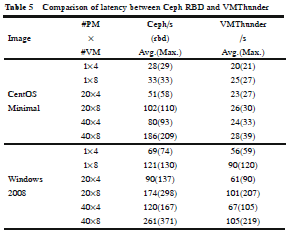
在对比实验中，我们使用一个40节点的Ceph集群，与ubuntu14.04，内核版本3.14，头孢0.79版。上述对比实验图像的相同类型，一个是CentOS最小的，另一个是Windows 2008。的虚拟机配置在Ceph的配置也有其他的选择，但是，我们选择最近的一虚拟机。所以，我们在原始图像上配置一个快照，以保持它的只读。为了保持可比性，我们设置的快照与QCOW2无缓存（QCOW2缓存=没有），表5中描述的比较，结果表明，虚拟机仍有优势。

此外，Ceph的写入性能低，通过我们的测试15MB/s。所以，在Ceph RBD服务新形象的制备具有低得多的效率比虚拟机。

**5未来趋势**

虚拟机的配置是在IaaS云服务的一个重要组成部分。在这一节中，我们指出未来趋势的虚拟机配置技术。

首先，大型虚拟机配置问题的本质是将大量的数据从一个或多个主机传输到一个更大范围的主机。一个趋势的研究将减少数据传输的量，并限制了所涉及的主机的影响。此外，随着硬件的发展，如固态硬盘和10千兆以太网，配置问题最终将变得微不足道。



第二，大型虚拟机的配置在几秒钟内是一个里程碑式的成就，用这样的小延迟，用户的满意程度将大大提高。正如我们所知，虚拟机的运行时性能，应考虑与虚拟机配置。因此，追求更高的运行时性能和运行时的稳定性将是虚拟机配置问题的下一个关键步骤。

第三，虚拟存储资源配置的问题，这可能包括弹性虚拟存储配置，数据集配置，软件更新配置等所有这些问题都是由虚拟存储的瓶颈造成的虚拟存储的问题，有着密切的关系。在我们看来，所有的扩展配置服务可以重用类似的方法的图像引导虚拟机配置。然而，现有的产业实践的具体配置问题可能是发展的障碍，解决所有配置问题的通用解决方案。

第四、为IaaS云服务日趋成熟和稳定，虚拟存储将是技术瓶颈。这个瓶颈已经和可能仍然是一个问题，在虚拟机配置，特别是在多租户环境中，我们得出的结论是，虚拟存储将是一个活跃的领域，在学术界和工业社会。

第五，新技术如集装箱和码头工人将有良好的发展前景。然而，由于兼容性问题，现有的工业实践可能会阻碍新技术的广泛采用。

第六，内存虚拟化在过去几年中发展相对缓慢，由于内存管理和系统架构的复杂性[ 48 ]。例如，内存为基础的虚拟机叉和远程热虚拟机复制已经提出了几年在学术界，虽然他们没有被用于工业。我们相信，内存虚拟化将是一个很好的方向，虚拟机配置和虚拟机本身。

总之，大型虚拟机配置的技术正在变得成熟，和类似的技术将扩展到其他虚拟资源配置。有一些未来的研究热点，包括虚拟存储的性能，可靠性，可扩展性，新的VM管理程序技术和存储虚拟化技术。

**6结论**

快速虚拟机配置是保证服务质量（QoS）关键的IaaS云服务。本文系统地回顾了现有的虚拟机配置方案。我们把它们分为三大类。我们讨论每一类的特点和研究现状，介绍了最近的两个解决方案，虚拟机和虚拟机 +，两者都能提供数百个虚拟机在秒。我们还讨论了未来趋势的虚拟机配置的问题和相关的技术。

这部分工作是由国家重点基础研究发展计划（2011cb302601），国家高技术研究发展计划（2011aa01a202）、中国国家自然科学基金（批准号：61222205），教育部新世纪优秀人才支持计划，和霍英东教育基金（141066）。

**工具书类**

1. Lu X, Wang H, Wang J, Xu J, Li D. Internet-based virtual computing environment: beyond the data center as a computer. uture Generation Computer Systems, 2013, 29(1): 309–322

2. Ahmed W, Wu Y. Estimation of cloud node acquisition. Tsinghua Science and Technology, 2014, 19(1): 1–12

3. Mao M, Humphrey M. A performance study on the VM startup time in the cloud. In: Proceedings of IEEE International Conference on Cloud Computing (CLOUD). 2012, 423–430

4. Zhang Z, Li Z,Wu K, Li D, Li H, Peng Y, Lu X. VMThunder: fast provisioning of large-scale virtual machine clusters. IEEE Transactions on Parallel and Distributed System, 2014, 25(12): 3328–3338

5. Sotomayor B, Keahey K, Foster I. Combining batch execution and leasing using virtual machines. In: Proceedings of the 17th International Symposium on High Performance Distributed Computing. 2008, 87– 96

6. Sotomayor B, Montero R S, Llorente I M, Foster I. Virtual infrastructure management in private and hybrid clouds. IEEE Internet Computing, 2009, 13(5): 14–22

7. Li J, Li D, Ye Y, Lu X. Efficient multi-tenant virtual machine allocation in cloud data centers. Tsinghua Science and Technology, 2015, 20(1): 81–89

8. Le D, Huang H, Wang H. Understanding performance implications of nested file systems in a virtualized environment. In: Proceedings of USENIX Conference on File and Storage Technologies. 2012, 8

9. Bellard F. Qemu, a fast and portable dynamic translator. In: Proceedings of USENIX Annual Technical Conference. 2005, 41–46

10. Nicolae B, Bresnahan J, Keahey K, Antoniu G. Going back and forth: efficient multideployment and multisnapshotting on clouds. In: Proceedings of ACM Symposium on High Performance Distributed Computing. 2011, 147–158

11. Xiao W, Liu Y, Yang Q, Ren J, Xie C. Implementation and performance evaluation of two snapshot methods on iSCSI target storages. In: Proceedings of IEEE Conference on Mass Storage Systems and Technologies. 2006

12. Jayaram K R, Peng C, Zhang Z, Kim M, Chen H, Lei H. An empirical analysis of similarity in virtual machine images. In: Proceedings of the Middleware. 2011, 6

13. Peng C, KimM, Zhang Z, Lei H. Vdn: Virtual machine image distribution network for cloud data centers. In: Proceedings of IEEE Infocom. 2012, 181–189

14. Razavi K, Ion A, Kielmann T. Squirrel: Scatter hoarding VM image contents on IaaS compute nodes. In: Proceedings of the 23rd International Symposium on High-performance Parallel and Distributed Computing. 2014, 265–278

15. Jin K, Miller E L. The effectiveness of deduplication on virtual machine disk images. In: Proceedings of SYSTOR. 2009, 7

16. Ng C H, Ma M, Wong T Y, Lee P P C, Lui J C S. Live deduplication storage of virtual machine images in an open-source cloud. In: Proceedings of the 12th International Middleware Conference. 2011, 80–99

17. Srinivasan K, Bisson T, Goodson G, Voruganti K. iDedup: Latencyaware, inline data deduplication for primary storage. In: Proceedings of the 10th USENIX Conference on File and Storage Technologies. 2012, 12: 1–14

18. Ammons G, Bala V, Mummert T, Reimer D, Zhang X. Virtual machine images as structured data: the mirage image library. In: Proceedings of USENIX HotCloud. 2011

19. Reimer D, Thomas A, Ammons G, Mummert T, Alpern B, Bala V. Opening black boxes: using semantic information to combat virtual machine image sprawl. In: Proceedings of International Conference on Virtual Execution Environments. 2008, 111–120

20. Tang C. Fvd: a high-performance virtual machine image format for cloud. In: Proceedings of USENIX Annual Technical Conference.2011

21. Papadopoulos P. Extending clusters to Amazon EC2 using the rocks toolkit. International Journal of High Performance Computing Applications, 2011, 25(3): 317–327

22. Nurmi D, Wolski R, Grzegorczyk C, Obertelli G, Soman S, Youseff L, Zagorodnov D. The eucalyptus open-source cloud-computing system. In: Proceedings of the 9th IEEE/ACM International Symposium on CCGrid. 2009, 124–131

23. LiD,Cao J, LuX, Chen K. Efficient range query processing in peer-topeer systems. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(1): 78–91

24. Zhang Z, Lu X, Peng Y, Li H. A reality check of multiple snowball tree file dissemination in large scale cloud cluster. In: Proceedings of the 15th IEEE International Symposium on Object/Component/Service- Oriented Real-Time Distributed Computing Workshops. 2012, 76–80

25. Wartel R, Cass T, Moreira B, Roche E, Guijarro M, Goasguen S, Schwickerath U. Image distribution mechanisms in large scale cloud providers. In: Proceedings of IEEE CloudCom. 2010, 112–117

26. Chen Z, Zhao Y, Miao X, Chen Y, Wang Q. Rapid provisioning of cloud infrastructure leveraging peer-to-peer networks. In: Proceedings of the 29th IEEE International Conference on Distributed Computing Systems Workshops. 2009, 324–329

27. O’Donnell C M. Using bittorrent to distribute virtual machine images for classes. In: Proceedings of the 36th Annual ACM SIGUCCS Fall Conference: Moving Mountains, Blazing Trails. 2008, 287–290

28. Reich J, Laadan O, Brosh E, Sherman A, Misra V, Nieh J, Rubenstein D. Vmtorrent: scalable P2P virtual machine streaming. In: Proceedings of ACM Conference on Emerging Network Experiment and Technology. 2012, 289–300

29. Morgan Jr T. Drbl: Diskless remote boot in Linux. NETWORK, 2006, 192

30. Weil S A, Brandt S A, Miller E L, Long D D, Maltzahn C. Ceph: A scalable, high-performance distributed file system. In: Proceedings of the 7th Symposium on Operating Systems Design and Implementation. 2006, 307–320

31. Shamma M, Meyer D T,Wires J, Ivanova M, Hutchinson N C,Warfield A. Capo: recapitulating storage for virtual desktops. In: Proceedings of USENIX Conference on File and Storage Technologies. 2011

32. Liao X, Xiong X, Jin H, Hu L. Lvd: A lightweight virtual desktop management architecture. Systems and Virtualization Management. Standards and New Technologies, 2008, 25–36

33. Wo T,Wang H, Hu C, Cui Y. Dvce: the virtual computing environment supported by distributed VM images. In: Proceedings of ISORC iVCE Workshop. 2012

34. Carns P H, Ligon III W B, Ross R B, Thakur R. PVFS: A parallel file system for linux clusters. In: Proceedings of the 4th Annual Linux Showcase Conference. 2000, 391–430

35. Flouris M D, Lachaize R, Bilas A. Orchestra: Extensible block-level support for resource and data sharing in networked storage systems. In: Proceedings of the 14th IEEE International Conference on Parallel and Distributed Systems. 2008, 237–244

36. Flouris M D, Bilas A. Violin: a framework for extensible block-level storage. In: Proceedings of the 22nd IEEE Goddard Conference on Mass Storage Systems and Technologies. 2005, 28–142

37. Meyer D, Aggarwal G, Cully B, Lefeb vre G, Feeley M, Hutchinson N, Warfield A. Parallax: virtual disks for virtual machines. In: Proceedings of EuroSys. 2008

38. Razavi K, Kielmann T. Scalable virtual machine deployment using vm image caches. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. 2013, 65

39. Zhao X, Zhang Y, Wu Y, Chen K, Jiang J, Li K. Liquid: A scalable deduplication file system for virtual machine images. IEEE Transactions on Parallel and Distributed System, 2014, 25(5): 1257–1266

40. Morrey III C B, Grunwald D. Content-based block caching. In: Proceedings of the 23rd IEEE Conference on Mass Storage Systems and Technologies. 2006

41. Ng C H, Ma M, Wong T Y, Lee P P C, Lui J C S. Live deduplication storage of virtual machine images in an open-source cloud. In: Proceedings of the 12th International Middleware Conference. 2011, 80–99

42. Lagar-Cavilla H A, Whitney J A, Scannell A M, Patchin P, Rumble S M, De Lara E, Brudno M, Satyanarayanan M. Snowflock: rapid virtual machine cloning for cloud computing. In: Proceedings of the 4th ACM European Conference on Computer Systems. 2009, 1–12

43. Zhu J, Jiang Z, Xiao Z. Twinkle a fast resource provisioning mechanism for internet services. In: Proceedings of IEEE Infocom. 2011, 802–810

44. Cui L, Li J, LiB, Huai J, HuC, Wo T,Al-Aqrabi H, Liu L.VMScatter: Migrate virtual machines to many hosts. In: Proceedings of International Conference on Virtual Execution Environments. 2013, 63–72

45. Merkel D. Docker: Lightweight linux containers for consistent development and deployment. Linux Journal, 2014, 2014(239): 2

46. Zhao Y,Wu J, Liu C. On peer-assisted data dissemination in data center networks: Analysis and implementation. Tsinghua Science and Technology, 2014, 19(1): 51–64

47. Zhang Z, Wu K, Li H, Feng J, Peng Y, Lu X. Raflow: Read ahead accelerated I/O flow through multiple virtual layers. In: Proceedings of the 9th IEEE International conference on networking, architecture and storage. 2014, 33–42

48. Zhang P, Chu R, Wang H. Swapcached: An effective method to promote guest paging performance on virtualization platform. In: Proceedings of the 7th IEEE International Symposium on Service Oriented System Engineering. 2013, 379–384

|  |  |
| --- | --- |
| Zhaoning Zhang received his MS in computer science at the National University of Defense Technology (NUDT), China in 2009. He was a visiting scholar in the Computer Science Department, University of Victoria, Canada and is currently a PhD student in the Computer School of NUDT. His research focuses on virtualization and block-level distributed storage. | Dongsheng Li received his PhD (with honor) at National University of Defense Technology (NUDT), China in 2005. He is currently a professor in the Computer School of NUDT. His research interests include cloud computing, computer networks and data management. |
| Kui Wu received his BS and the MS in computer science from Wuhan University, China in 1990 and 1993, respectively, and his PhD in Computing Science from the University of Alberta, Canada in 2002. He joined the Department of Computer Science at the University of Victoria, Canada in 2002 and is currently a Professor there. His research interests include mobile and wireless networks, network performance evaluation, and cloud computing. |  |