文章
Michael Lei · 五月 12, 2021 阅读大约需 16 分钟

InterSystems 最佳实践系列之数据平台和性能 – 第 9 篇 InterSystems IRIS VMware 最佳实践指南

本贴提供了在 VMware ESXi 5.5 及更高版本的环境中部署 Caché 2015
及更高版本时，关于配置、系统规模调整和容量规划等方面的指南。

我假定您已经了解 VMware vSphere 虚拟化平台，所以直接给出推荐。
本指南中的推荐不特定于任何具体硬件或站点特定的实现，也不应作为规划和配置 vSphere
部署的全面指南，而是一份您可以做出选择的最佳实践配置清单。 我希望您的 VMware
专家实施团队能针对具体站点对这些推荐进行评估。

这里是 InterSystems 数据平台和性能系列的其他帖子的列表。

*注：*本帖更新于 2017 年 1 月 3 日，强调必须为生产数据库实例设置虚拟机内存预留，以保证 Caché
有足够内存可用，并且不会出现内存交换或膨胀而对数据库性能产生负面影响。
更多详细信息，请参见下面的*内存*部分。

参考

本文中的信息基于经验和对公开的 VMware 知识库文章及 VMware 文档（例如 VMware vSphere 性能最佳实践
）的研究以及对 Caché 数据库部署要求的反映。

ESXi 支持 InterSystems 的产品吗？

针对处理器类型和操作系统（包括虚拟化操作系统）验证和发布 InterSystems 产品是 InterSystems 的策略和程序。
有关详情，请参见 InterSystems 支持策略和版本信息。

> 例如，在 x86 主机上的 ESXi 中，支持在 Red Hat 7.2 操作系统上运行 Caché 2016.1。

注：如果您不编写自己的应用程序，还必须检查应用程序供应商支持策略。

支持的硬件

在配合使用当前服务器和存储组件的情况下，VMware 虚拟化很适合 Caché。 使用 VMware 虚拟化的 Caché
已在客户站点成功部署，并且性能和可伸缩性已在基准测试中得到验证。
在配置正确的存储、网络和搭载较新型号的英特尔至强处理器（具体为英特尔至强 5500、5600、7500、E7 系列和
E5 系列，包括最新的 E5 v4）的服务器上使用 VMware 虚拟化不会对性能产生明显影响。

通常，Caché 和应用程序在客户机操作系统上的安装和配置方式与裸机操作系统上的安装和配置方式相同。

客户有责任查看 VMware 兼容性指南，以了解所使用的特定服务器和存储。

虚拟化架构

VMware 常用于 Caché 应用程序的两种标准配置：

- 主生产数据库操作系统实例在 “ 裸机 ” 集群上，而 VMware 只用于其他生产和非生产实例，如 Web 服务器、打印、测试、训练等。
- 所有操作系统实例（包括主生产实例）均已虚拟化。

本帖可用作任一方案的指南，不过重点是第二个方案，即包括生产实例在内的所有操作系统实例均已虚拟化。
下图显示了该配置的典型物理服务器设置。

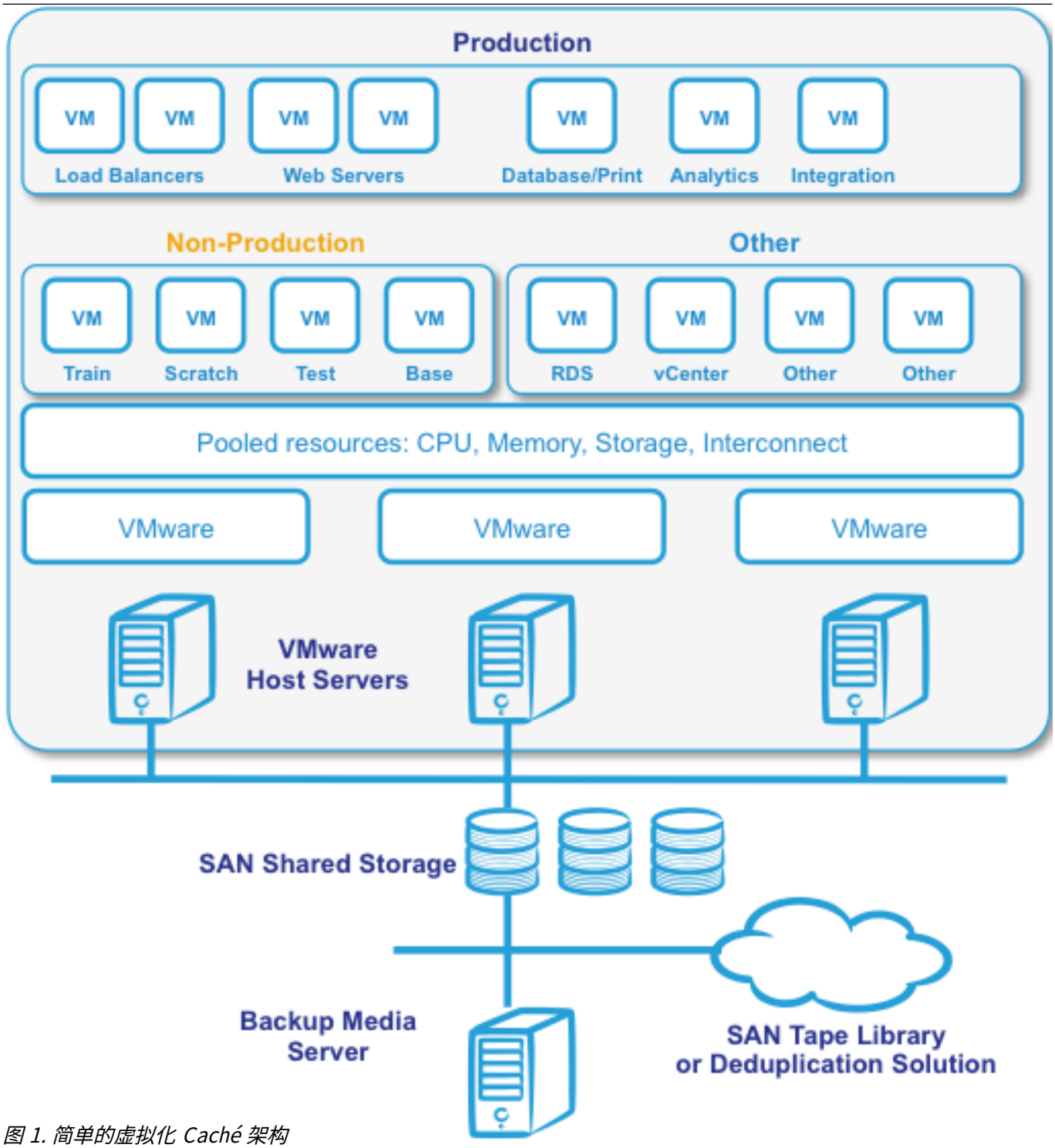

图 1. 简单的虚拟化 Caché 架构

图 1 显示了一个常见的包含至少三台物理主机服务器的部署，可通过 VMware HA 集群中的主机服务器提供 N+1 容量和可用性。 要扩展资源，可能要向集群添加额外的物理服务器。
备份/还原媒体管理和灾难恢复也可能需要额外的物理服务器。

有关特定于 *VMware vSAN*、VMware 的超融合基础架构解决方案的推荐，请参见以下帖子：第 8 部分 超融合基础架构容量和性能规划。 本贴中的大部分推荐可以应用于 vSAN，除了下面的 “ 存储 ” 部分有一些明显差异。

VMWare 版本

下表给出了针对 Caché 2015 及更高版本的主要推荐：

vSphere 是包括 vCenter Server 在内的产品套件，允许通过 vSphere 客户端对主机和虚拟机进行集中系统管理。

> 本帖假定将使用 vSphere，而不是 “ 免费的 ” ESXi Hypervisor 单独版本。

VMware 有多种授权模式；最终选择取决于哪个模式最适合您当前和将来的基础架构规划。

我通常推荐 “ 企业版 ”，因为它增加了功能，例如让硬件得到更有效利用的动态资源调度 (DRS)，以及用于存储阵列整合（快照备份）的存储 API。 VMware 网站给出了版本比较。

还有高级套件，允许捆绑 vSphere 的 vCenter Server 和 CPU 许可证。
套件有升级限制，因此通常只推荐给不期望增长的小型站点。

ESXi 主机 BIOS 设置

ESXi 主机是物理服务器。 在配置 BIOS 前，您应该：

- 向硬件供应商核实服务器是否正在运行最新的 BIOS
- 检查是否有服务器/CPU 型号特定的 BIOS 设置专门用于 VMware。

服务器 BIOS 的默认设置对于 VMware 可能不是最佳设置。
以下设置可以用来优化物理主机服务器以获得最佳性能。 并非下表中的所有设置在所有供应商的服务器上都可用。

内存

内存分配应考虑以下关键规则：

在单台物理主机上运行多个 Caché 实例或其他应用程序时，VMware 有多种技术可以实现高效的内存管理，如透明页共享 (TPS)、膨胀、交换和内存压缩。
例如，当多个操作系统实例在同一主机上运行时，TPS 可去除内存页的冗余副本，从而允许内存过载而不会降低性能，这使得虚拟机运行所占用的内存少于物理机所需内存。

> 注：必须在操作系统中安装 VMware Tools，才能利用 VMware 的这些功能和许多其他功能。

虽然这些功能的存在是为了允许内存过载，但建议务必先调整所有虚拟机的 vRAM 大小，以适应可用的物理内存。
在生产环境中，尤其重要的是要仔细考虑内存过载的影响，只有在收集数据以确定可能的过载量后，才能进行内存过载。 要确定给定 Caché 实例的内存共享有效性和可接受的过载程度，请运行工作负载并使用 Vmware 命令 resxtop 或 esxtop 来观察实际的内存节省情况。

在规划 Caché 实例内存要求时，建议回顾一下本系列中关于内存的第四个帖子。 尤其是 “ VMware 虚拟化注意事项 ” 部分，其中的重点是：

在生产系统上设置 VMware 内存预留。

您想要*必须*避免任何共享内存交换，所以将生产数据库虚拟机内存预留设置为至少是 Caché 共享内存的大小加上 Caché 进程和操作系统及内核服务的内存。如果有疑虑，则**预留整个生产数据库虚拟机内存（100% 预留）**，以保证 Caché 实例有足够内存可用，这样就不会出现内存交换或膨胀而对数据库性能产生负面影响。

注：较大的内存预留将影响 vMotion 运行，因此在设计 vMotion/管理网络时必须考虑到这一点。
只有目标主机的可用物理内存大于或等于预留内存的大小时，才能使用 Vmware HA
实时迁移虚拟机或在另一台主机上启动虚拟机。 这对于 Caché 生产虚拟机特别重要。 例如，要特别注意 HA 准入控制策略。

确保容量规划允许在 HA 发生故障转移时分配虚拟机。

对于非生产环境（测试、训练等），可以应用更积极的内存过载，但不要让 Caché
共享内存过载，而是通过减少全局缓冲区来限制 Caché 实例的共享内存。

当前的英特尔处理器架构具有 NUMA 拓扑。
处理器有自己的本地内存，并且可以访问同一主机中其他处理器上的内存。
毫无疑问，访问本地内存的延迟要低于远程内存。 有关 CPU
的讨论，请查看本系列的第三个帖子，*评论部分*中包括了关于 NUMA 的讨论。

如上面的 BIOS 部分所述，实现最佳性能的策略是，在理想情况下，将虚拟机规模调整为只达到单个处理器支持的最大核心数和内存量。 例如，如果容量规划显示最大的生产 Caché 数据库虚拟机将具有 14 个 vCPU 和 112 GB 内存，则考虑具有 2 个 E5-2680 v4（14 核处理器）和 256 GB 内存的服务器集群是否合适。

理想情况下，调整虚拟机规模以使内存在 NUMA 节点本地。 但不要太过执着于此。

如果需要一个比 NUMA 节点大的“ 怪兽虚拟机 ”，那也没问题，VMware 将管理 NUMA 以获得最佳性能。
合理调整虚拟机的规模并且分配的资源不超过所需资源也很重要（参见下文）。

CPU

虚拟 CPU 分配应考虑以下关键规则：

Caché 生产系统的规模应根据实际客户站点的基准测试和测量结果来确定。
对于生产系统，使用一开始就将系统规模调整为与裸机 CPU
核心数相同的策略，并根据最佳实践进行监测，看是否可以减少虚拟 CPU (vCPU)。

超线程和容量规划

根据物理服务器的规则调整**生产数据库**
虚拟机规模的一个良好起点是，针对启用了超线程功能的目标处理器计算物理服务器 CPU
要求，然后简单地进行转换：

一个物理 CPU（包括超线程）= 一个 vCPU（包括超线程）。

一个常见的误解是，超线程以某种方式使 vCPU 容量增加了一倍。 这对于物理服务器或逻辑 vCPU 来说并不正确。
裸机服务器上的超线程可能会比没有超线程的相同服务器提升 30%
的性能，但这也会根据应用情况的不同而有所不同。

对于初始规模调整，假定 vCPU 具有完整的内核专用资源。 例如，如果您有一台 32 核（2 个 16 核）E5-2683 V4 服务器 – 总共多达 32 个 vCPU 的容量规模，且可能还有余量。 此配置假定在主机级别启用了超线程。 VMware 将管理主机上所有应用程序和虚拟机之间的调度。 在您花时间监测应用程序、操作系统和 VMware 在峰值处理期间的性能后，您可以决定是否进行更高度的整合。

授权

在 vSphere 中，可以为虚拟机配置一定数量的插槽或核心。 例如，如果有一个双处理器虚拟机（2 个 vCPU），则可以将其配置为两个 CPU 插槽，或一个具有两个 CPU 核心的插槽。
从执行的角度看，并没有多大区别，因为虚拟机监控程序将最终决定虚拟机是在一个还是两个物理插槽上执行。
但是，指定双 CPU 虚拟机实际有两个核心而不是两个插槽，会对软件许可证产生影响。 注：Caché 许可证以核心数为准（而非线程数）。

存储

> 本部分适用于更为传统的使用共享存储阵列的存储模型。 有关 *vSAN* 的建议，另请参见以下帖子：第 8 部分 超融合基础架构容量和性能规划

对于存储，应考虑以下关键规则：

针对性能调整存储大小

存储瓶颈是影响 Caché 系统性能的最常见问题之一，对于 VMware vSphere 配置也是如此。
最常见的问题是简单地按 GB 容量来调整存储大小，而不是分配足够多的磁盘主轴来支持预期的 IOPS。 在 VMware 中，存储问题可能更加严重，因为可能有更多主机通过相同的物理连接访问同一存储。

VMware 存储概述

VMware 存储虚拟化可以分为三层，例如：

- 存储阵列是底层，由物理磁盘组成，以逻辑磁盘（存储阵列卷或 LUN）的形式呈现给上面的层。
- 下一层是 vSphere 占用的虚拟环境。 存储阵列 LUN 以数据存储的形式呈现给 ESXi 主机，并格式化为 VMFS 卷。
- 虚拟机由数据存储中的文件组成，所包含的虚拟磁盘以磁盘的形式呈现给客户机操作系统，可以对这些磁盘进行分区并在文件系统中使用。

VMware 为管理虚拟机中的磁盘访问提供了两种选择 — VMware 虚拟机文件系统 (VMFS) 和原始设备映射 (RDM)，两者的性能相似。 为了简化管理，VMware 通常推荐 VMFS，但在某些情况下可能需要 RDM。
作为一般建议 – 除非有特别的理由使用 RDM，否则请选择 VMFS，*VMware 的新开发针对 VMFS 而非 RDM*。

虚拟机文件系统 (VMFS)

VMFS 是 VMware 开发的一种文件系统，专门针对集群虚拟环境（允许从多个主机进行读/写访问）和大型文件存储进行优化。 VMFS 的结构使虚拟机文件可以存储在一个文件夹中，从而简化了虚拟机管理。 VMFS 还支持 vMotion、DRS 和 VMware HA 等 VMware 基础架构服务。

操作系统、应用程序和数据存储在虚拟磁盘文件（.vmdk 文件）中。 vmdk 文件存储在数据存储中。
一个虚拟机可以由分布在几个数据存储上的多个 vmdk 文件组成。
如下图中的生产虚拟机所示，一个虚拟机可以包括分布在多个数据存储上的存储。 对于生产系统，每个 LUN 使用一个 vmdk 文件可实现最佳性能，对于非生产系统（测试、训练等），多个虚拟机 vmdk 文件可以共享一个数据存储和一个 LUN。

虽然 vSphere 5.5 支持的最大 VMFS 卷大小为 64TB，VMDK 大小为 62TB，但在部署 Caché 时，通常使用映射到不同磁盘组上的 LUN 的多个 VMFS 卷来分离 IO 模式并提高性能。 例如，随机或顺序 IO 磁盘组，或者将生产 IO 与其他环境的 IO 分开。

下图显示了与 Caché 一起使用的 VMware VMFS 存储示例的概览：

图 2. VMFS 上的 Caché 存储示例

RDM

RDM 允许将原始 SCSI 磁盘或 LUN 作为 VMFS 文件来进行管理和访问。 RDM 是 VMFS 卷上的特殊文件，用作原始设备的代理。 对于大多数虚拟磁盘存储，推荐使用 VMFS，但在某些情况下可能需要使用原始磁盘。 RDM 仅适用于光纤通道或 iSCSI 存储。

用于阵列集成的 VMware vStorage API (VAAI)

为获得最佳存储性能，客户应考虑使用支持 VAAI 的存储硬件。 VAAI 可以在几个方面（包括虚拟机置备和精简置备的虚拟磁盘）提高性能。
对于较旧的阵列，可以向阵列供应商索取固件更新来支持 VAAI。

虚拟磁盘类型

ESXi 支持多种虚拟磁盘类型：

厚置备 – 在创建时分配空间。 进一步分为：

- 快速置零 – 将 0 写入整个驱动器。
  这会增加创建磁盘所需的时间，但可获得最佳性能，即使是第一次写入每个块。
- 延迟置零 – 首次写入每个块时写入 0。 延迟置零缩短了创建时间，但第一次写入块时会使性能降低。
  不过，后续写入的性能与快速置零的厚磁盘相同。

精简置备 – 在写入时分配空间并置零。 对未写入的文件块进行第一次写入时，会有较高的 I/O 成本（类似于延迟置零的厚磁盘），但在随后写入精简置备的磁盘时，性能与快速置零的厚磁盘相同。

在所有磁盘类型中，VAAI 都可以通过将操作卸载到存储阵列来提高性能。
某些阵列还在阵列级别支持精简置备，请勿在精简置备的阵列存储上精简置备 ESXi 磁盘，因为置备和管理可能存在冲突。

其他备注

如上文所述，为了符合最佳实践，请使用与裸机配置相同的策略；生产存储可以在阵列级别分成多个磁盘组：

- 随机访问 Caché 生产数据库
- 顺序访问备份和日志，但也可以放置其他非生产存储，如测试、训练等

请记住，数据存储是存储层的抽象表示，因此，它是存储的逻辑表示而不是物理表示。
创建专门的数据存储以隔离特定的 I/O工作负载（无论是日志还是数据库文件），并且不隔离物理存储层，并不会对性能产生预期的影响。

虽然性能是关键，但共享存储的选择更多取决于站点现有或计划的基础架构，其次才是 VMware 的影响。
与裸机实现一样，FC SAN 的性能最好，推荐使用。 对于 FC，建议至少使用 8Gbps 的适配器。
只有在适当的网络基础架构到位的情况下，才能支持 iSCSI

存储，这包括：服务器和存储之间的网络中的所有组件上都必须支持至少 10Gb 的以太网和巨型帧 (MTU 9000)，并且与其他流量分开。

对于数据库虚拟机或具有高 I/O 负载的虚拟机，使用多个 VMware 准虚拟 SCSI (PVSCSI) 控制器。 PVSCSI 可以提高整体存储吞吐量，同时降低 CPU 利用率，从而带来显著效益。 通过使用多个 PVSCSI 控制器，可以在客户机操作系统内执行多个并行 I/O 操作。 此外，建议通过单独的虚拟 SCSI 控制器将日志 I/O 流量与数据库 I/O 流量分开。 作为最佳实践，您可以将一个控制器用于操作系统和交换，另一个控制器用于日志，另外一个或多个控制器用于数据库数据文件（取决于数据库数据文件的数量和大小）。

众所周知，对齐文件系统分区是针对数据库工作负载的存储最佳实践。 物理机和 VMware VMFS 分区上均对齐分区可以防止因 I/O 跨越磁道边界而导致的 I/O 性能下降。 VMware 测试结果显示，将 VMFS 分区与 64KB 磁道边界对齐，可降低延迟并提高吞吐量。 按照存储和操作系统供应商的建议，使用 vCenter 创建的 VMFS 分区应与 64KB 边界对齐。

网络

对于网络，应考虑以下关键规则：

如上文所述，VMXNET 适配器比默认的 E1000 适配器具有更好的性能。 VMXNET3 允许 10Gb，并且使用更少的 CPU，而 E1000 只支持 1Gb。 如果各主机之间只有 1 Gb 的网络连接，那么客户端与虚拟机的通信不会有太大变化。 但是，VMXNET3 将允许同一主机上的虚拟机之间建立 10Gb 网络连接，这确实会带来改变，特别是在多层部署或者实例之间有高网络 IO 要求的情况下。
在计划关联性和反关联性 DRS 规则以将虚拟机保持在相同或不同的虚拟交换机上时，也应考虑此功能。

E1000 使用可用于 Windows 或 Linux 的通用驱动程序。 在客户机操作系统上安装 VMware Tools 之后，即可安装 VMXNET 虚拟适配器。

下图显示了一个典型的具有四个物理网卡端口的小型服务器配置，在 VMware 中配置了两个端口用于基础架构流量：用于管理和 vMotion 的 dvSwitch0，两个端口用于 VM 使用的应用程序。 使用网卡绑定和负载平衡来实现最佳吞吐量和 HA。

图 3. 典型的具有四个物理网卡端口的小型服务器配置。

客户机操作系统

推荐如下：

> 将 VMware Tools 加载到所有虚拟机操作系统中并保持该工具为最新是非常重要的。

VMware Tools 是一套实用工具，可增强虚拟机的客户机操作系统的性能，并改进对虚拟机的管理。 如果客户机操作系统中未安装 VMware Tools，则客户机性能会缺乏重要功能。

在所有 ESXi 主机上正确设置时间至关重要 - 这最终会影响客户机虚拟机。
虚拟机的默认设置是不将客户机的时间与主机同步 -
但在某些时候，客户机还是会与主机同步时间，并且如果超时，则已知会导致重大问题。 VMware 建议使用 NTP 而不是 VMware Tools 定期进行时间同步。 NTP 是行业标准，可确保客户机的计时准确。 可能需要开放防火墙 (UDP 123) 才能允许 NTP 流量。

DNS 配置

如果 DNS 服务器托管在虚拟化基础架构上并且不可用，它会阻止 vCenter 解析主机名，使虚拟环境无法管理 - 然而虚拟机本身保持运行，没有问题。

高可用性

高可用性由 VMware vMotion、VMware Distributed Resource Scheduler (DRS) 和 VMware High Availability (HA) 等功能提供。 Caché 数据库镜像也可以用于增加正常运行时间。

为 Caché 生产系统设计 n+1 个物理主机是非常重要的。 在单个主机发生故障时，必须有足够的资源（例如 CPU 和内存）让其余主机上的所有虚拟机都能运行。 在服务器发生故障时，如果 VMware 无法在其余服务器上分配足够的 CPU 和内存资源，VMware HA 将不会在其余服务器上重新启动虚拟机。

vMotion

vMotion 可以与 Caché 一起使用。 vMotion 允许以完全透明的方式将一个正常运行的虚拟机从一台 ESXi 主机服务器迁移到另一台。 虚拟机中运行的操作系统和 Caché 等应用程序没有服务中断。

使用 vMotion 进行迁移时，只有虚拟机的状态和内存（及其配置）会移动。
虚拟磁盘不需要移动；它保留在相同的共享存储位置。 虚拟机迁移后，它将在新的物理主机上运行。

vMotion 只能在共享存储架构（如共享 SAS 阵列、FC SAN 或 iSCSI）下工作。 由于 Caché 通常配置为使用大量共享内存，因此为 vMotion 提供充足的网络容量是非常重要的，1Gb 网络可能还可以，但也可能需要更高带宽，或者可以配置多网卡 vMotion。

DRS

分布式资源调度器 (DRS) 是一种通过在集群中的不同主机服务器之间共享工作负载来在生产环境中自动使用 vMotion 的方法。 DRS 还具有对虚拟机实例实施 QoS 的能力，通过阻止非生产虚拟机过度使用资源来保护生产虚拟机的资源。 DRS 会收集有关集群主机服务器使用情况的信息，并通过在集群的不同服务器之间分配虚拟机工作负载来优化资源。 这种迁移可以自动或手动进行。

Caché 数据库镜像

对于要求最高可用性的任务关键型一级 Caché 数据库应用实例，还应考虑使用 InterSystems 同步数据库镜像 。同时使用镜像的其他优势包括：

- 最新数据的单独副本。
- 故障转移在数秒内完成（比先重启虚拟机再重启操作系统再恢复 Caché 要快）。
- 在应用程序/Caché 发生故障（VMware 未检测到）时进行故障转移。

vCenter Appliance

vCenter Server Appliance 是基于 Linux 的预配置虚拟机，针对 vCenter Server 和相关服务的运行进行了优化。 我一直建议拥有小型集群的站点使用 VMware vCenter Server Appliance 作为在 Windows 虚拟机上安装 vCenter Server 的替代方案。 在 vSphere 6.5 中，建议将该设备用于所有部署。

总结

本帖是在 VMware 上部署 Caché 时应考虑的关键最佳实践的综述。 这些最佳实践大多不是 Caché 独有的，而是可以应用于 VMware 上的其他一级业务关键型部署。

如果您有任何问题，请通过下面的评论告诉我。

#InterSystems 业务解决方案和架构 #系统管理 #Caché #InterSystems IRIS #InterSystems IRIS for Health #文档

源 URL: https://cn.community.intersystems.com/post/intersystems-%E6%9C%80%E4%BD%B3%E5%AE%9E%E8%B7%B5%E7%B3%BB%E5%88%97%E4%B9%8B%E6%95%B0%E6%8D%AE%E5%B9%B3%E5%8F%B0%E5%92%8C%E6%80%A7%E8%83%BD-%E2%80%93-%E7%AC%AC-9-%E7%AF%87-intersystems-iris-vmware-%E6%9C%80%E4%BD%B3%E5%AE%9E%E8%B7%B5%E6%8C%87%E5%8D%97