



SUSE Linux Enterprise Server 15 SP7

虚拟化指南

虚拟化指南

SUSE Linux Enterprise Server 15 SP7

本指南概括介绍了虚拟化技术。其中介绍了 libvirt（统一虚拟化接口），并提供了有关特定超级管理程序的详细信息。

出版日期：2025 年 9 月 23 日

<https://documentation.suse.com> 

版权所有 © 2006–2025 SUSE LLC 和撰稿人。保留所有权利。

根据 GNU 自由文档许可证 (GNU Free Documentation License) 版本 1.2 或（根据您的选择）版本 1.3 中的条款，在此授予您复制、分发和/或修改本文档的权限；本版权声明和许可证附带不可变部分。许可版本 1.2 的副本包含在“GNU Free Documentation License”部分。

有关 SUSE 商标，请参见 <https://www.suse.com/company/legal/> 。所有第三方商标均是其各自所有者的财产。商标符号（®、™ 等）代表 SUSE 及其关联公司的商标。星号 (*) 代表第三方商标。

本指南力求涵盖所有细节，但这不能确保本指南准确无误。SUSE LLC 及其关联公司、作者和译者对于可能出现的错误或由此造成的后果皆不承担责任。

目录

前言 xix

1 可用文档 xix

2 改进文档 xix

3 文档约定 xx

4 支持 xxii

SUSE Linux Enterprise Server 支持声明 xxiii • 技术预览 xxiii

I 简介 1

1 虚拟化技术 2

1.1 概述 2

1.2 虚拟化的优点 2

1.3 虚拟化模式 3

1.4 I/O 虚拟化 4

2 虚拟化场景 6

2.1 服务器合并 6

2.2 隔离 7

2.3 灾难恢复 7

2.4 动态负载平衡 8

3 Xen 虚拟化简介 9

3.1 基本组件 9

3.2 Xen 虚拟化体系结构 10

4 KVM 虚拟化简介 12

4.1 基本组件 12

4.2 KVM 虚拟化体系结构 12

5 虚拟化工具 14

5.1 虚拟化控制台工具 14

5.2 虚拟化 GUI 工具 15

6 安装虚拟化组件 18

6.1 简介 18

6.2 安装虚拟化组件 18

指定系统角色 18 • 运行 **YaST 虚拟化** 模块 19 • 安装特定的安装软件集 20

6.3 在 KVM 中启用嵌套虚拟化 21

VMware ESX 用作 Guest 超级管理程序 23

7 虚拟化限制和支持 24

7.1 体系结构支持 24

KVM 硬件要求 24 • Xen 硬件要求 25

7.2 超级管理程序限制 25

KVM 限制 26 • Xen 限制 26

7.3 支持的主机环境（超级管理程序） 27

7.4 支持的 Guest 操作系统 28

半虚拟化驱动程序的提供 29

7.5 支持的 VM 迁移场景 30

脱机迁移场景 30 • 实时迁移场景 31

7.6 功能支持 33

Xen 主机 (Dom0) 34 • Guest 功能支持 35

II 使用 libvirt 管理虚拟机 37

8 libvirt 守护程序 38

8.1 启动和停止模块化守护程序 38

8.2 启动和停止一体化守护程序 40

8.3 切换到一体化守护程序 42

9 准备 VM 主机服务器 44

9.1 配置网络 44

网桥 44 • 虚拟网络 49

9.2 配置存储池 59

使用 **virsh** 管理存储 62 • 使用虚拟机管理器管理存储设备 67

10 Guest 安装 73

10.1 基于 GUI 的 Guest 安装 73

为虚拟机配置 PXE 引导 75

10.2 使用 **virt-install** 从命令行安装 76

10.3 高级 Guest 安装方案 79

高级 UEFI 配置 80 • 对 Windows Guest 使用内存气球 82 • 在安装中包含附加产品 83

11 基本 VM Guest 管理 84

11.1 列出 VM Guest 84

使用虚拟机管理器列出 VM Guest 84 • 使用 **virsh** 列出 VM Guest 84

11.2 通过控制台访问 VM Guest 85

打开图形控制台 85 • 打开串行控制台 87

- 11.3 更改 VM Guest 的状态：启动、停止、暂停 88
 - 使用虚拟机管理器更改 VM Guest 的状态 88 • 使用 **virsh** 更改 VM Guest 的状态 89
- 11.4 保存和恢复 VM Guest 的状态 90
 - 使用虚拟机管理器保存/恢复 92 • 使用 **virsh** 保存和恢复 92
- 11.5 创建和管理快照 93
 - 术语 94 • 使用虚拟机管理器创建和管理快照 94 • 使用 **virsh** 创建和管理快照 96
- 11.6 删除 VM Guest 99
 - 使用虚拟机管理器删除 VM Guest 99 • 使用 **virsh** 删除 VM Guest 99
- 11.7 监控 99
 - 使用虚拟机管理器进行监控 99 • 使用 **virt-top** 进行监控 100 • 使用 **kvm_stat** 进行监控 101
- 12 连接和授权 104**
- 12.1 身份验证 104
 - libvirtd authentication 105 • VNC 身份验证 109
- 12.2 连接到 VM 主机服务器 113
 - 非特权用户的“system”访问权限 115 • 使用虚拟机管理器管理连接 116
- 12.3 配置远程连接 116
 - 基于 SSH 的远程隧道 (qemu+ssh 或 xen+ssh) 117 • 使用 x509 证书进行远程 TLS/SSL 连接 (qemu+tls 或 xen+tls) 117
- 13 高级存储主题 125**
- 13.1 使用 **virtlockd** 锁定磁盘文件和块设备 125
 - 启用锁定 125 • 配置锁定 126
- 13.2 联机调整 Guest 块设备的大小 127

- 13.3 在主机与 Guest 之间共享目录（文件系统直通） 128
- 13.4 通过 libvirt 使用 RADOS 块设备 129
- 14 使用虚拟机管理器配置虚拟机 130**
 - 14.1 计算机设置 131
 - 概述 131 · 性能 132 · 处理器 133 · 内存 134 · 引导选项 135
 - 14.2 存储 136
 - 14.3 控制器 138
 - 14.4 网络 139
 - 14.5 输入设备 140
 - 14.6 视频 142
 - 14.7 USB 重定向器 143
 - 14.8 杂项 143
 - 14.9 使用虚拟机管理器添加 CD/DVD-ROM 设备 144
 - 14.10 使用虚拟机管理器添加软盘设备 145
 - 14.11 使用虚拟机管理器弹出和更换软盘或 CD/DVD-ROM 媒体 146
 - 14.12 将主机 PCI 设备分配到 VM Guest 147
 - 使用虚拟机管理器添加 PCI 设备 147
 - 14.13 将主机 USB 设备分配到 VM Guest 148
 - 使用虚拟机管理器添加 USB 设备 148
- 15 使用 virsh 配置虚拟机 150**
 - 15.1 编辑 VM 配置 150
 - 15.2 更改计算机类型 151

- 15.3 配置超级管理程序功能 152
- 15.4 配置 CPU 153
 - 配置 CPU 数量 153 • 配置 CPU 型号 155
- 15.5 更改引导选项 156
 - 更改引导顺序 157 • 使用直接内核引导 157
- 15.6 配置内存分配 158
- 15.7 添加 PCI 设备 160
 - IBM Z 的 PCI 直通 163
- 15.8 添加 USB 设备 164
- 15.9 添加 SR-IOV 设备 165
 - 要求 165 • 加载和配置 SR-IOV 主机驱动程序 166 • 将 VF 网络设备添加到 VM Guest 169 • 动态分配池中的 VF 172
- 15.10 列出挂接的设备 174
- 15.11 配置存储设备 175
- 15.12 配置控制器设备 176
- 15.13 配置视频设备 178
 - 更改分配的 VRAM 量 178 • 更改 2D/3D 加速状态 178
- 15.14 配置网络设备 179
 - 使用多队列 virtio-net 提升网络性能 179
- 15.15 使用 macvtap 共享 VM 主机服务器网络接口 179
- 15.16 禁用内存气球设备 181
- 15.17 配置多个监控器（双头） 181
- 15.18 将 IBM Z 上的加密适配器直通到 KVM Guest 183
 - 简介 183 • 本章内容 183 • 要求 183 • 将加密适配器专用于 KVM 主机 183 • 更多资料 186

16 使用 AMD SEV-SNP 增强虚拟机安全性 187

- 16.1 支持的硬件 187
- 16.2 启用机密计算模块 187
- 16.3 安装软件包并配置基础系统 188
- 16.4 验证安装 189
- 16.5 启动 AMD SEV-SNP 虚拟机 189
- 16.6 验证 AMD SEV-SNP 虚拟机 192

17 迁移 VM Guest 193

- 17.1 迁移类型 193
- 17.2 迁移要求 194
- 17.3 使用虚拟机管理器进行实时迁移 195
- 17.4 使用 **virsh** 进行迁移 196
- 17.5 分步操作示例 198
 - 导出存储区 198
 - 在目标主机上定义池 199
 - 创建卷 200
 - 创建 VM Guest 201
 - 迁移 VM Guest 201

18 Xen 到 KVM 的迁移指南 202

- 18.1 使用 **virt-v2v** 迁移到 KVM 202
 - virt-v2v** 简介 202
 - 安装 **virt-v2v** 203
 - 将虚拟机转换为在 libvirt 管理的 KVM 下运行 203
 - 运行转换的虚拟机 208
- 18.2 Xen 到 KVM 的手动迁移 209
 - 一般概述 209
 - 备份 Xen VM Guest 209
 - 特定于半虚拟化 Guest 的更改 210
 - 更新 Xen VM Guest 配置 213
 - 迁移 VM Guest 217
- 18.3 更多信息 218

III 独立于超级管理程序的功能 219

19 磁盘缓存模式 220

- 19.1 什么是磁盘缓存? 220
- 19.2 磁盘缓存的工作原理 220
- 19.3 磁盘缓存的优势 220
- 19.4 虚拟磁盘缓存模式 221
- 19.5 缓存模式和数据完整性 221
- 19.6 缓存模式和实时迁移 222

20 VM Guest 时钟设置 223

- 20.1 KVM: 使用 `kvm_clock` 223
 - 其他计时方法 224
- 20.2 Xen 虚拟机时钟设置 224

21 libguestfs 225

- 21.1 VM Guest 操作概述 225
 - VM Guest 操作风险 225 · libguestfs 的设计用途 226
- 21.2 软件包安装 226
- 21.3 Guestfs 工具 227
 - 修改虚拟机 227 · 支持的 filesystem 和磁盘映像 227 · **virt-rescue** 228 · **virt-resize** 229 · 其他 `virt-*` 工具 230 · **guestfish** 233 · 将物理机转换为 KVM Guest 234
- 21.4 查错 236
 - Btrfs 相关的问题 236 · 环境 236 · **libguestfs-test-tool** 237
- 21.5 更多信息 237

22 QEMU Guest 代理 238

- 22.1 运行 QEMU GA 命令 238
- 22.2 需要 QEMU GA 的 **virsh** 命令 239
- 22.3 增强 libvirt 命令 239
- 22.4 更多信息 240

23 软件 TPM 模拟器 241

- 23.1 简介 241
- 23.2 先决条件 241
- 23.3 安装 241
- 23.4 将 **swtpm** 与 QEMU 配合使用 241
- 23.5 将 swtpm 与 libvirt 配合使用 243
- 23.6 使用 OVMF 固件进行 TPM 测量 243
- 23.7 资源 243

24 创建 VM Guest 的崩溃转储 244

- 24.1 简介 244
- 24.2 为全虚拟化计算机创建崩溃转储 244
- 24.3 为半虚拟化计算机创建崩溃转储 244
- 24.4 附加信息 244

IV 使用 XEN 管理虚拟机 246

25 设置虚拟机主机 247

- 25.1 最佳实践和建议 247

- 25.2 管理 Dom0 内存 248
 - 设置 Dom0 内存分配 249
- 25.3 全虚拟化 Guest 中的网卡 250
- 25.4 启动虚拟机主机 250
- 25.5 PCI 直通 252
 - 配置超级管理程序以使用 PCI 直通 253 · 将 PCI 设备分配给 VM Guest 系统 254 · VGA 直通 255 · 查错 255 · 更多信息 256
- 25.6 USB 直通 256
 - 标识 USB 设备 256 · 模拟的 USB 设备 257 · 半虚拟化 PVUSB 257

26 虚拟网络 260

- 26.1 Guest 系统的网络设备 260
- 26.2 Xen 中基于主机的路由 262
- 26.3 创建伪装网络设置 264
- 26.4 特殊配置 266
 - 虚拟网络中的带宽限制 267 · 监控网络流量 267

27 管理虚拟化环境 268

- 27.1 XL — Xen 管理工具 268
 - Guest 域配置文件 269
- 27.2 自动启动 Guest 域 270
- 27.3 事件操作 270
- 27.4 时戳计数器 271
- 27.5 保存虚拟机 272
- 27.6 恢复虚拟机 273
- 27.7 虚拟机状态 273

28 Xen 中的块设备 274

- 28.1 将物理存储设备映射到虚拟磁盘 274
- 28.2 将网络存储设备映射到虚拟磁盘 275
- 28.3 基于文件的虚拟磁盘和回写设备 275
- 28.4 调整块设备的大小 276
- 28.5 用于管理高级存储方案的脚本 277

29 虚拟化：配置选项和设置 278

- 29.1 虚拟 CD 读取器 278
 - 半虚拟计算机上的虚拟 CD 读取器 278
 - 全虚拟计算机上的虚拟 CD 读取器 278
 - 添加虚拟 CD 读取器 279
 - 去除虚拟 CD 读取器 280
- 29.2 远程访问方法 280
- 29.3 VNC 查看器 280
 - 向虚拟机分配 VNC 查看器端口号 281
 - 使用 SDL 而不是 VNC 查看器 282
- 29.4 虚拟键盘 282
- 29.5 分配专用 CPU 资源 283
 - Dom0 283
 - VM Guest 284
- 29.6 HVM 功能 285
 - 指定引导时使用的引导设备 285
 - 更改 Guest 的 CPUID 285
 - 增加 PCI-IRQ 的数量 286
- 29.7 虚拟 CPU 调度 287

30 管理任务 288

- 30.1 引导加载程序 288
- 30.2 稀疏映像文件和磁盘空间 289

30.3 迁移 Xen VM Guest 系统 290
检测 CPU 功能 291 • 准备要迁移的块设备 292 • 迁移 VM Guest 系统 293

30.4 监控 Xen 293
使用 **xentop** 监控 Xen 293 • 其他工具 294

30.5 提供 VM Guest 系统的主机信息 295

31 XenStore：在域之间共享的配置数据库 297

31.1 简介 297

31.2 文件系统接口 297
XenStore 命令 298 • /vm 298 • /local/domain/<domid> 301

32 使用 Xen 作为高可用性虚拟化主机 303

32.1 使用远程存储设备实现 Xen HA 303

32.2 使用本地存储设备实现 Xen HA 304

32.3 Xen HA 和专用网桥 304

33 Xen：将半虚拟 (PV) Guest 转换为全虚拟 (FV/HVM) Guest 306

V 使用 QEMU 管理虚拟机 310

34 QEMU 概述 311

35 设置 KVM VM 主机服务器 312

35.1 CPU 的虚拟化支持 312

35.2 所需的软件 312

- 35.3 特定于 KVM 主机的功能 314
 - 使用具有 virtio-scsi 的主机存储设备 314 • 使用 vhost-net 实现加速网络 315 • 使用多队列 virtio-net 提升网络性能 316 • VFIO：对设备进行安全的直接访问 317 • VirtFS：在主机与 Guest 之间共享目录 319 • KSM：在 Guest 之间共享内存页 320

36 Guest 安装 322

- 36.1 使用 **qemu-system-ARCH** 进行基本安装 322
- 36.2 使用 **qemu-img** 管理磁盘映像 323
 - 有关 qemu-img 调用的一般信息 324 • 创建、转换和检查磁盘映像 325 • 使用 qemu-img 管理虚拟机的快照 330 • 有效操作磁盘映像 332

37 使用 **qemu-system-ARCH** 运行虚拟机 337

- 37.1 基本 **qemu-system-ARCH** 调用 337
- 37.2 一般 **qemu-system-ARCH** 选项 338
 - 基本虚拟硬件 339 • 存储和读取虚拟设备的配置 341 • Guest 实时时钟 342
- 37.3 在 QEMU 中使用设备 342
 - 块设备 343 • 图形设备和显示选项 349 • USB 设备 351 • 字符设备 352
- 37.4 QEMU 中的网络 354
 - 定义网络接口卡 355 • 用户模式网络 356 • 桥接网络 358
- 37.5 使用 VNC 查看 VM Guest 361
 - 保护 VNC 连接 363

38 使用 **QEMU** 监控器管理虚拟机 366

- 38.1 访问监控器控制台 366
- 38.2 获取有关 Guest 系统的信息 367

- 38.3 更改 VNC 口令 370
- 38.4 管理设备 370
- 38.5 控制键盘和鼠标 371
- 38.6 更改可用内存 372
- 38.7 转储虚拟机内存 372
- 38.8 管理虚拟机快照 373
- 38.9 挂起和恢复虚拟机执行 374
- 38.10 动态迁移 375
- 38.11 QMP - QEMU 计算机协议 376
 - 通过标准输入/输出访问 QMP 376 · 通过 telnet 访问 QMP 377 · 通过 Unix 套接字访问 QMP 378 · 通过 libvirt 的 **virsh** 命令访问 QMP 379

VI 查错 380

39 集成式帮助和软件包文档 381

40 收集系统信息和日志 382

- 40.1 libvirt 日志控制 382

词汇表 384

A 虚拟机驱动程序 393

B 为 NVIDIA 卡配置 GPU 直通 394

- B1 简介 394

- B2 先决条件 394

B3 配置主机 394

校验主机环境 394 • 启用 IOMMU 395 • 将 Nouveau 驱动程序加入黑名单 396 • 配置 VFIO 并隔离用于直通的 GPU 396 • 加载 VFIO 驱动程序 396 • 为 Microsoft Windows Guest 禁用 MSR 397 • 安装 UEFI 固件 398 • 重引导主机计算机 398

B4 配置 Guest 398

Guest 配置要求 398 • 安装显卡驱动程序 399

C XM、XL 工具栈和 libvirt 框架 402

C1 Xen 工具栈 402

从 xend/xm to xl/libxl 升级 403 • XL 设计 403 • 升级前的核对清单 403

C2 将 Xen 域配置导入 libvirt 404

C3 xm 与 xl 应用程序之间的差异 406

表示法约定 406 • 新的全局选项 407 • 未更改的选项 407 • 已去除的选项 411 • 已更改的选项 414 • 新选项 428

C4 外部链接 429

C5 以与 xm 兼容的格式保存 Xen Guest 配置 430

D GNU licenses 431

前言

1 可用文档

联机文档

可在 <https://documentation.suse.com> 上查看我们的联机文档。您可浏览或下载各种格式的文档。



注意：最新更新

最新的更新通常会在本文档的英文版中提供。

SUSE 知识库

如果您遇到问题，请参考 <https://www.suse.com/support/kb/> 上提供的联机技术信息文档 (TID)。在 SUSE 知识库中搜索根据客户需求提供的已知解决方案。

发行说明

有关发行说明，请参见 <https://www.suse.com/releasesnotes/>。

在您的系统上

如需脱机使用，您也可在系统的 `/usr/share/doc/release-notes` 下找到该发行说明。各软件包的相应文档可在 `/usr/share/doc/packages` 中找到。

许多命令的**手册页**中也对相应命令进行了说明。要查看手册页，请运行 `man` 后跟特定的命令名。如果系统上未安装 `man` 命令，请使用 `sudo zypper install man` 加以安装。

2 改进文档

欢迎您提供针对本文档的反馈及改进建议。您可以通过以下渠道提供反馈：

服务请求和支持

有关产品可用的服务和支持选项，请参见 <https://www.suse.com/support/>。

要创建服务请求，需在 SUSE Customer Center 中注册订阅的 SUSE 产品。请前往 <https://scc.suse.com/support/requests> 并登录，然后单击新建。

Bug 报告

在 <https://bugzilla.suse.com/> 中报告文档问题。

要简化此过程，请点击本文档 HTML 版本中标题旁边的报告问题图标。这样会在 Bugzilla 中预先选择正确的产品和类别，并添加当前章节的链接。然后，您便可以立即开始键入 Bug 报告。

需要一个 Bugzilla 帐户。

贡献

要帮助改进本文档，请点击本文档 HTML 版本中标题旁边的 Edit Source document（编辑源文档）图标。然后您会转到 GitHub 上的源代码，可以在其中提出拉取请求。

需要一个 GitHub 帐户。



注意： Edit source document（编辑源文档）仅适用于英语版本

Edit source document（编辑源文档）图标仅适用于每个文档的英语版本。对于所有其他语言，请改用报告问题图标。

有关用于本文档的文档环境的详细信息，请参见储存库的 README。

邮件

您也可以将有关本文档的错误以及反馈发送至 doc-team@suse.com。请在其中包含文档标题、产品版本和文档发布日期。此外，请包含相关的章节号和标题（或者提供 URL），并提供问题的简要说明。

3 文档约定

本文档中使用了以下通知和排版约定：

- /etc/passwd：目录名称和文件名
- PLACEHOLDER：请将 PLACEHOLDER 替换为实际值
- PATH：环境变量
- ls、--help：命令、选项和参数

- user: 用户或组的名称
- package_name: 软件包的名称
- **Alt**、**Alt + F1**: 按键或组合键。按键以大写字母显示，与键盘上的一样。
- 文件、文件 > 另存为: 菜单项、按钮
- **AMD/Intel** 本段内容仅与 AMD64/Intel 64 体系结构相关。箭头标记文本块的开始位置和结束位置。 ◁
- **IBM Z, POWER** 本段内容仅与 IBM Z 和 POWER 体系结构相关。箭头标记文本块的开始位置和结束位置。 ◁
- Chapter 1, “Example chapter”: 对本指南中其他章节的交叉引用。
- 必须使用 root 特权运行的命令。您还可以在这些命令前加上 sudo 命令，以非特权用户身份来运行它们：

```
# command
> sudo command
```

- 非特权用户也可以运行的命令：

```
> command
```

- 可以通过一行末尾处的反斜线字符 (\) 拆分成两行或多行的命令。反斜线告知外壳命令调用将会在该行末尾后面继续：

```
> echo a b \
c d
```

- 显示命令（前面有一个提示符）和外壳返回的相应输出的代码块：

```
> command
output
```

- 注意事项



警告：警报通知

在继续操作之前，您必须了解的关键性信息。向您指出有关安全问题、潜在数据丢失、硬件损害或物理危害的警告。



重要：重要通知

在继续操作之前，您必须了解的重要信息。



注意：注意通知

额外信息，例如有关软件版本差异的信息。



提示：提示通知

有用信息，例如指导方针或实用性建议。

- 精简通知



额外信息，例如有关软件版本差异的信息。



有用信息，例如指导方针或实用性建议。

4 支持

下面提供了 SUSE Linux Enterprise Server 的支持声明和有关技术预览的一般信息。有关产品生命周期的细节，请参见 <https://www.suse.com/lifecycle>。有关虚拟化支持状态，请参见第 7 章“虚拟化限制和支持”。

如果您有权获享支持，可在 <https://documentation.suse.com/sles-15/html/SLES-all/cha-adm-support.html> 中查找有关如何收集支持票据所需信息的细节。

4.1 SUSE Linux Enterprise Server 支持声明

要获得支持，您需要订阅适当的 SUSE 产品。要查看为您提供的具体支持服务，请前往 <https://www.suse.com/support/> 并选择您的产品。

支持级别的定义如下：

L1

问题判定，该技术支持级别旨在提供兼容性信息、使用支持、持续维护、信息收集，以及使用可用文档进行基本查错。

L2

问题隔离，该技术支持级别旨在分析数据、重现客户问题、隔离问题区域，并针对级别 1 不能解决的问题提供解决方法，或完成准备工作以提交级别 3 处理。

L3

问题解决，该技术支持级别旨在借助工程方法解决级别 2 支持所确定的产品缺陷。

对于签约客户与合作伙伴，SUSE Linux Enterprise Server 包含除以下项目外的其他所有软件包的 L3 支持：

- 技术预览。
- 声音、图形、字体和作品。
- 需要额外客户合同的软件包。
- **Workstation Extension** 模块随附的某些软件包仅享受 L2 支持。
- 名称以 `-devel` 结尾的软件包（包含头文件和类似的开发人员资源）只能与其主软件包一起获得支持。

SUSE 仅支持使用原始软件包，即，未发生更改且未重新编译的软件包。

4.2 技术预览

技术预览是 SUSE 提供的旨在让用户大致体验未来创新的各种软件包、堆栈或功能。随附这些技术预览只是为了提供方便，让您有机会在自己的环境中测试新的技术。非常希望您能提供反馈。如果您测试了技术预览，请联系 SUSE 代表，将您的体验和用例告知他们。您的反馈对于我们的未来开发非常有帮助。

技术预览存在以下限制：

- 技术预览仍处于开发阶段。因此，它们可能在功能上不完整、不稳定，或者**不适合**生产用途。
- 技术预览**不受支持**。
- 技术预览可能仅适用于特定的硬件体系结构。
- 技术预览的细节和功能可能随时会发生变化。因此，可能无法升级到技术预览的后续版本，而只能进行全新安装。
- SUSE 可能会发现某个预览不符合客户或市场需求，或者未遵循企业标准。技术预览可能会随时从产品中删除。SUSE 不承诺未来将提供此类技术的受支持版本。

如需大致了解产品随附的技术预览，请参见 <https://www.suse.com/releasesnotes>  上的发行说明。

I 简介

- 1 虚拟化技术 2
- 2 虚拟化场景 6
- 3 Xen 虚拟化简介 9
- 4 KVM 虚拟化简介 12
- 5 虚拟化工具 14
- 6 安装虚拟化组件 18
- 7 虚拟化限制和支持 24

1 虚拟化技术

虚拟化技术提供了一种供计算机（主机）在主机操作系统上运行另一个操作系统（Guest 虚拟机）的方式。

1.1 概述

SUSE Linux Enterprise Server 中包含最新的开源虚拟化技术：Xen 和 KVM。借助这些超级管理程序，可以使用 SUSE Linux Enterprise Server 在单个物理系统上置备、取消置备、安装、监控和管理多个虚拟机 (VM Guest)。有关详细信息，请参见[超级管理程序](#)。SUSE Linux Enterprise Server 能够创建可运行经过修改、高度优化的半虚拟化操作系统，以及未经修改的全虚拟化操作系统的虚拟机。

操作系统中实现虚拟化的主要组件是超级管理程序（或虚拟机管理器），它是直接在服务器硬件上运行的软件层。超级管理程序控制着平台资源，并通过向每个 VM Guest 提供虚拟化的硬件接口，在多个 VM Guest 及其操作系统之间共享这些资源。

SUSE Linux Enterprise 是企业级 Linux 服务器操作系统，提供两种类型的超级管理程序：Xen 和 KVM。

采用 Xen 或 KVM 的 SUSE Linux Enterprise Server 作为虚拟化主机服务器 (VHS)，支持具有各自 Guest 操作系统的 VM Guest。SUSE VM Guest 体系结构由一个超级管理程序以及多个管理组件组成，这些管理组件构成了 VHS，后者运行着许多托管应用程序的 VM Guest。

在 Xen 中，管理组件在通常称为 **Dom0** 的特权 VM Guest 中运行。在 Linux 内核充当超级管理程序的 KVM 中，管理组件直接在 VHS 上运行。

1.2 虚拟化的优点

虚拟化不仅能够提供与硬件服务器相同的服务，而且还具有许多优势。

首先，它降低了基础架构的成本。服务器主要用于向客户提供服务，而虚拟化的操作系统不仅能够提供相同的服务，同时还具备以下优势：

- 更少的硬件：您可以在一台主机上运行多个操作系统，因此可以减轻总体硬件维护工作量。
- 更低的能耗/散热成本：更少的硬件意味着当您需要更多服务时，无需在电能、备用电源和散热方面投入更多成本。
- 节省空间：由于不需要更多的硬件服务器（服务器数量比运行的服务数量更少），因此可以节省数据中心的空间。
- 更少的管理工作：使用 VM Guest 可以简化基础架构的管理。
- 灵活性和工作效率：虚拟化提供**迁移**功能、**实时迁移**和**快照**。这些功能减少了停机时间，并且可以在不造成任何服务中断的情况下，让您轻松将服务从一个位置转移到另一个位置。

1.3 虚拟化模式

Guest 操作系统以全虚拟化 (FV) 模式或半虚拟 (PV) 模式托管在虚拟机上。每种虚拟化模式都有其优缺点。

- 全虚拟化模式允许虚拟机运行未经修改的操作系统，例如 Windows* Server 2003。它可以使用二进制转换或**硬件辅助**虚拟化技术，例如 AMD* 虚拟化或 Intel* 虚拟化技术。在支持此功能的处理器上使用硬件辅助可以提高性能。

以全虚拟化模式托管的某些 Guest 操作系统可配置为使用 SUSE 虚拟机驱动程序包 (VMDP) 中的驱动程序，而不是源自操作系统的驱动程序。在 Windows Server 2003 这样的 Guest 操作系统上运行虚拟机驱动程序可以大幅提升性能。有关详细信息，请参见[附录 A “虚拟机驱动程序”](#)。

- 要使 Guest 操作系统能够在半虚拟模式下运行，通常需要根据虚拟化环境对其进行修改。不过，以半虚拟模式运行的操作系统比全虚拟化模式下运行的操作系统的性能更佳。当前已经过修改可在半虚拟模式下运行的操作系统称为**半虚拟化操作系统**，包括 SUSE Linux Enterprise Server。

1.4 I/O 虚拟化

VM Guest 不仅可以共享主机系统的 CPU 和内存资源，还能共享 I/O 子系统的此类资源。由于软件 I/O 虚拟化技术提供的性能低于裸机，因此最近开发了接近“本机”性能的硬件解决方案。SUSE Linux Enterprise Server 支持以下 I/O 虚拟化技术：

完全虚拟化

全虚拟化 (FV) 驱动程序会模拟受到广泛支持的真实设备，可以通过 VM Guest 中的现有驱动程序来使用这些设备。Guest 也称为**硬件虚拟机 (HVM)**。由于 VM 主机服务器上的物理设备可能不同于模拟的设备，超级管理程序需要先处理所有 I/O 操作，然后才能将其转交到物理设备。因此，所有 I/O 操作需要遍历两个软件层，这一过程不仅会显著影响 I/O 性能，而且还会消耗 CPU 时间。

半虚拟化

半虚拟化 (PV) 支持在超级管理程序与 VM Guest 之间直接通讯。与全虚拟化相比，它产生的开销更少，但性能却好很多。但使用半虚拟化技术时，无论是要支持半虚拟化 API 还是半虚拟化驱动程序，都必须修改 Guest 操作系统。有关支持半虚拟化的 Guest 操作系统列表，请参见第 7.4.1 节“半虚拟化驱动程序的提供”。

PVHVM

这种类型的虚拟化通过半虚拟化 (PV) 驱动程序以及 PV 中断和计时器处理增强了 HVM（请参见完全虚拟化）。

VFIO

VFIO 全称为**虚拟功能 I/O (Virtual Function I/O)**，是适用于 Linux 的新式用户级驱动程序框架。它取代了传统的 KVM PCI 直通设备分配。VFIO 驱动程序会在受安全内存 (IOMMU) 保护的环境中向用户空间公开直接的设备访问。利用 VFIO，VM Guest 可以直接访问 VM 主机服务器上的硬件设备（直通），避免性能关键型路径中的模拟操作造成性能问题。此方法不允许共享设备 — 每个设备只能分配到一个 VM Guest。VFIO 需受 VM 主机服务器 CPU、芯片组和 BIOS/EFI 的支持。

与传统的 KVM PCI 设备分配相比，VFIO 具有以下优势：

- 资源访问与 UEFI 安全引导兼容。
- 设备会被隔离，并且其内存访问受到保护。

- 提供设备所有权模型更为灵活的用户空间设备驱动程序。
- 独立于 KVM 技术，不受限于 x86 体系结构。

SUSE Linux Enterprise Server 中已弃用 USB 和 PCI 直通设备分配方法，采用 VFIO 模型来替代这些方法。

SR-IOV

作为最新的 I/O 虚拟化技术，单根 I/O 虚拟化 (SR-IOV) 结合了上述技术的优点 — 在兼顾性能的同时还能与多个 VM Guest 共享设备。SR-IOV 要求使用特殊的 I/O 设备，这些设备必须能够复制资源，使它们看似是多个独立设备。每个这样的“伪”设备都可由单个 Guest 直接使用。但对于网卡（举例而言），可使用的并发队列数有限，因此与半虚拟化驱动程序相比，SR-IOV 有可能会降低 VM Guest 的性能。在 VM 主机服务器上，SR-IOV 必须受 I/O 设备、CPU 和芯片组、BIOS/EFI 及超级管理程序的支持 — 有关设置说明，请参见第 14.12 节“将主机 PCI 设备分配到 VM Guest”。

! 重要：VFIO 和 SR-IOV 的要求

要能够使用 VFIO 和 SR-IOV 功能，VM 主机服务器需要满足以下要求：

- 需在 BIOS/EFI 中启用 IOMMU。
- 对于 Intel CPU，需要在内核命令行中提供内核参数 `intel_iommu=on`。有关详细信息，请参见 <https://github.com/torvalds/linux/blob/master/Documentation/admin-guide/kernel-parameters.txt#L1951>。
- VFIO 基础架构需可用。加载内核模块 `vfio_pci` 即可做到这一点。有关详细信息，请参见《管理指南》，第 19 章“systemd 守护程序”，第 19.6.4 节“加载内核模块”。

2 虚拟化场景

虚拟化为您的组织提供多种有用的功能，例如：

- 提高硬件使用效率
- 支持传统软件
- 操作系统隔离
- 实时迁移
- 灾难恢复
- 负载平衡

2.1 服务器合并

可用一台大型物理服务器取代许多服务器，以便整合硬件，并将 Guest 操作系统转换为虚拟机。这样还能支持在新硬件上运行旧式软件。

- 更好地利用未完全运行的资源
- 减少所需的服务器占地空间
- 更有效地利用计算机资源：将多个工作负载放到同一台服务器上
- 简化数据中心基础结构
- 简化将工作负载转移到其他主机的过程，同时可避免服务停机
- 更快、更灵敏的虚拟机置备
- 多个 Guest 操作系统可以在一台主机上运行



重要

进行服务器整合需要特别注意以下几点：

- 应该认真规划维护时段
- 存储空间非常关键：必须能够支持迁移以及不断增长的磁盘使用量
- 必须校验您的服务器是否能够支持额外的工作负载

2.2 隔离

Guest 操作系统与运行这些操作系统的主机完全隔离。因此，如果虚拟机内部出现问题，主机不会受到损害。此外，一个 VM 内部出现问题不会影响其他 VM。不会在 VM 之间共享数据。

- 可对 VM 使用 UEFI 安全引导。
- 应避免使用 KSM。有关 KSM 的更多细节，请参见 [KSM](#)。
- 可将单独的 CPU 核心指派给 VM。
- 应禁用超线程 (HT)，以避免潜在的安全问题。
- VM 不应共享网络、存储空间或网络硬件。
- 使用 PCI 直通或 NUMA 等高级超级管理程序功能会对 VM 迁移功能产生不利影响。
- 使用半虚拟化和 [virtio](#) 驱动程序可提高 VM 的性能和效率。

AMD 提供了与虚拟化安全性相关的特定功能。

2.3 灾难恢复

超级管理程序可以创建 VM 的快照，使之恢复到已知正常的状态或以前的任何所需状态。与直接在裸机上运行的操作系统相比，[虚拟化](#)操作系统对硬件配置的依赖较小，因此可将这些快照恢复到另一个服务器硬件，前提是该硬件运行的是相同的超级管理程序。

2.4 动态负载平衡

利用实时迁移，可以按需将 VM 从繁忙主机转移到容量富余的主机，轻松对基础结构中的服务进行负载平衡。

3 Xen 虚拟化简介

本章介绍并阐述您在设置和管理基于 Xen 的虚拟化环境时需要了解的组件与技术。

3.1 基本组件

基于 Xen 的虚拟化环境的基本组件包括：

- Xen 超级管理程序
- Dom0
- 任意数量的其他 VM Guest
- 用于管理虚拟化的工具、命令和配置文件

运行所有这些组件的物理计算机系统称为 **VM 主机服务器**，因为这些组件共同构成了托管虚拟机的平台。

Xen 超级管理程序

Xen 超级管理程序（有时简称为虚拟机监控器）是一个开源软件程序，用于协调虚拟机与物理硬件之间的低级别交互。

Dom0

虚拟机主机环境（也称为 **Dom0** 或控制域）由多个组件构成，其中包括：

- SUSE Linux Enterprise Server，提供图形环境和命令行环境，用于管理虚拟机主机组件及其虚拟机。



注意

术语“Dom0”指的是提供管理环境的特殊域。Dom0 能以图形模式或命令行模式运行。

- 基于 xenlight 库 (libxl) 的 xl 工具堆栈，用于管理 Xen Guest 域。
- 开源软件 QEMU，可模拟完整的计算机系统，包括处理器和多种外设。它提供以全虚拟化或半虚拟化模式托管操作系统的功能。

基于 Xen 的虚拟机

基于 Xen 的虚拟机（也称为 VM Guest 或 DomU）由以下组件构成：

- 至少一个包含可引导操作系统的虚拟磁盘。虚拟磁盘可以基于文件、分区、卷或其他类型的块设备。
- 每个 Guest 域的配置文件。这是遵循手册页 man 5 xl.conf 中所述语法的文本文件。
- 与控制域所提供的虚拟网络连接的多个网络设备。

管理工具、命令和配置文件

您可以结合使用 GUI 工具、命令与配置文件来管理和自定义虚拟化环境。

3.2 Xen 虚拟化体系结构

下图描绘了包含四个虚拟机的虚拟机主机。所示的 Xen 超级管理程序直接在物理硬件平台上运行。控制域也是虚拟机，不过相比所有其他虚拟机，它还承担了多项其他管理任务。

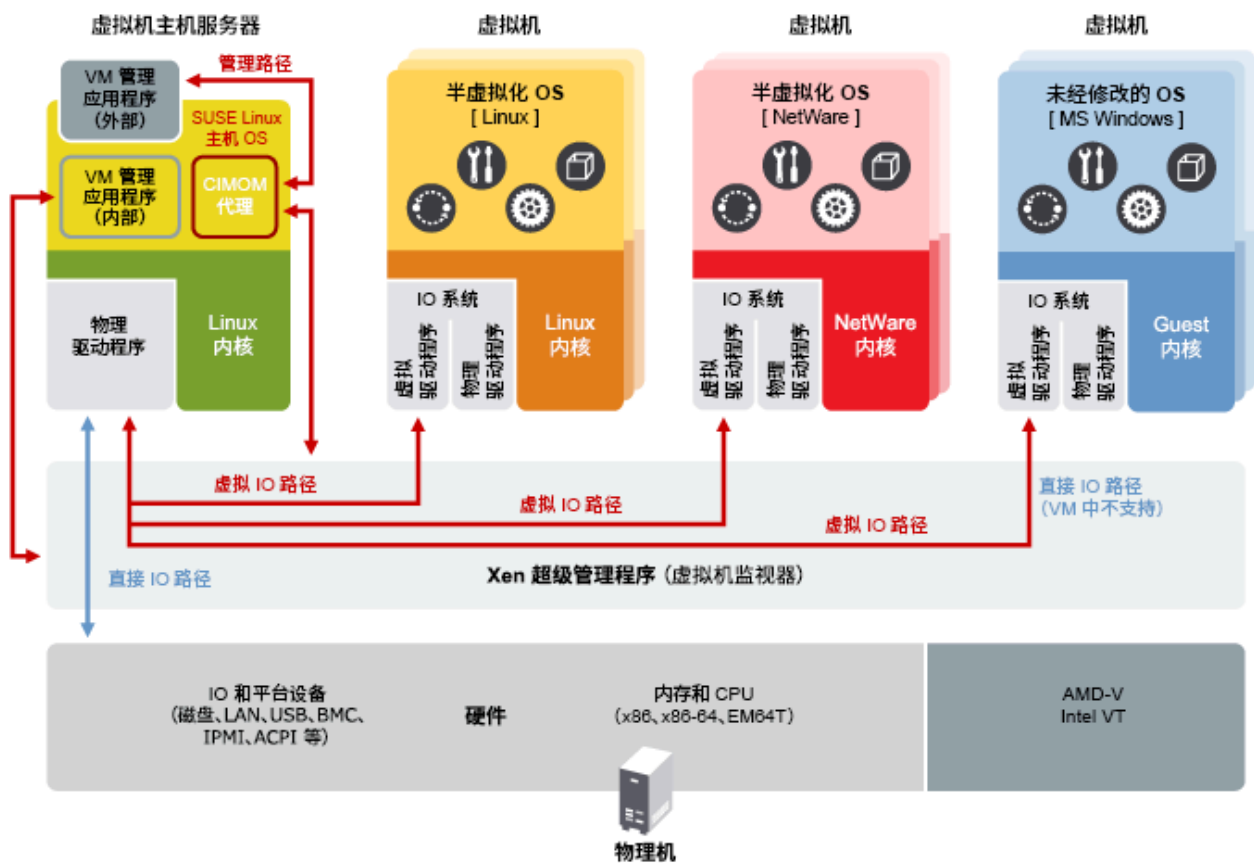


图 3.1：XEN 虚拟化体系结构

左侧所示虚拟机主机的 Dom0 运行的是 SUSE Linux Enterprise Server 操作系统。中间所示的两个虚拟机运行的是半虚拟化操作系统。右侧所示的虚拟机是全虚拟计算机，运行的是未经修改的操作系统，例如最新版本的 Microsoft Windows/Server。

4 KVM 虚拟化简介

4.1 基本组件

KVM 是支持硬件虚拟化的硬件体系结构全虚拟化解决方案（有关支持的体系结构的更多细节，请参见第 7.1 节“体系结构支持”）。

可以直接使用 QEMU 工具或使用基于 libvirt 的堆栈来管理 VM Guest（虚拟机）、虚拟存储和虚拟网络。QEMU 工具包括 qemu-system-ARCH、QEMU 监控器、qemu-img 和 qemu-ndb。基于 libvirt 的堆栈包括 libvirt 本身，以及 virsh、virt-manager、virt-install 和 virt-viewer 等基于 libvirt 的应用程序。

4.2 KVM 虚拟化体系结构

这款全虚拟化解决方案包括两个主要组件：

- 一组内核模块（kvm.ko、kvm-intel.ko 和 kvm-amd.ko），提供核心虚拟化基础架构和特定于处理器的驱动程序。
- 一个用户空间程序 (qemu-system-ARCH)，提供虚拟设备模拟以及用于管理 VM Guest（虚拟机）的控制机制。

术语 KVM 更适合表示内核级虚拟化功能，但在实践中，更多的是使用它来表示用户空间组件。

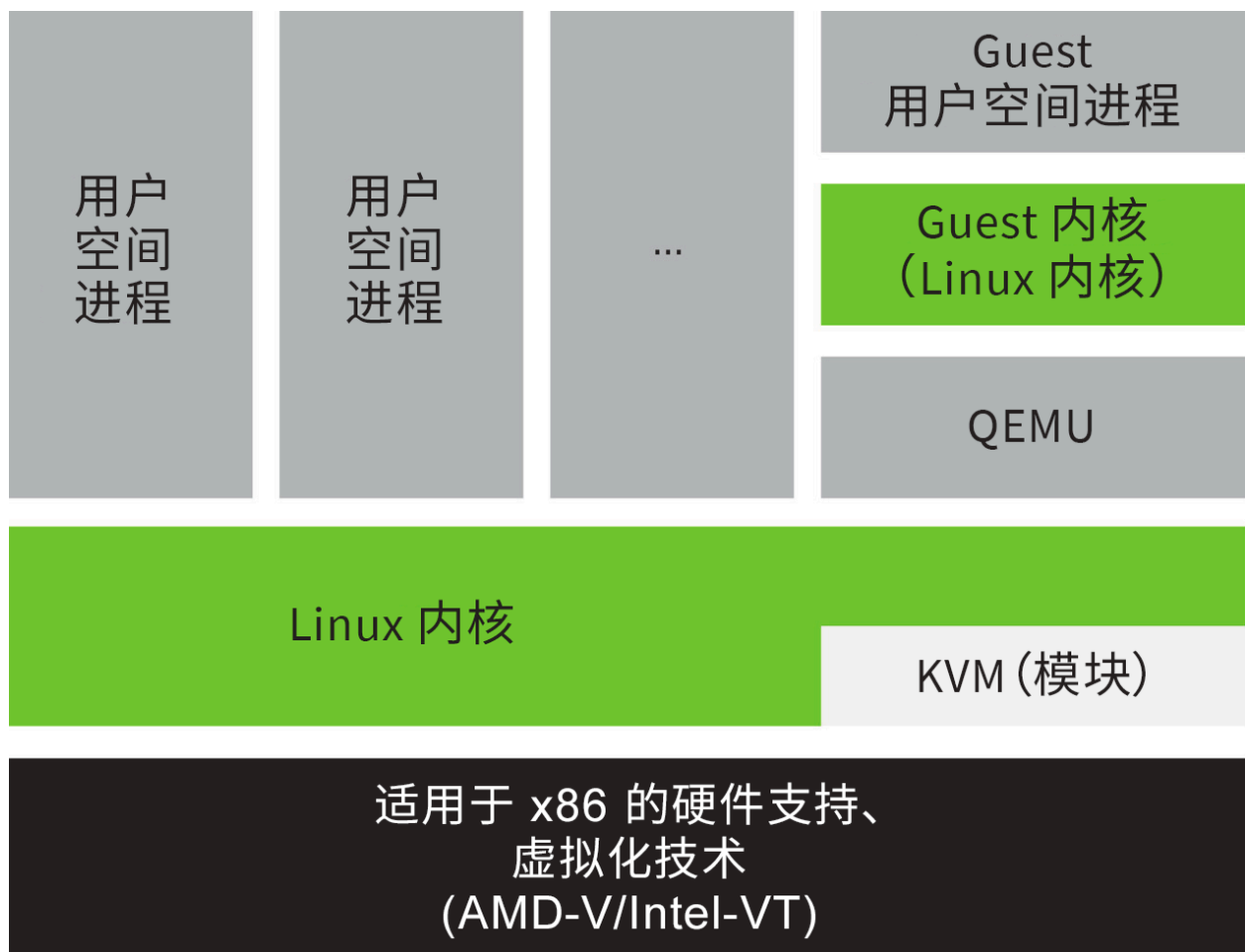


图 4.1：KVM 虚拟化体系结构

5 虚拟化工具

`libvirt` 是一个库，提供用于管理 KVM、Xen 等流行虚拟化解决方案的通用 API。该库为这些虚拟化解决方案提供规范化管理 API，以便为更高层级的管理工具提供一个跨超级管理程序的稳定接口。该库还提供用于管理 VM 主机服务器上的虚拟网络和存储空间的 API。每个 VM Guest 的配置都存储在 XML 文件中。

您还可以使用 `libvirt` 来远程管理 VM Guest。它支持 TLS 加密、x509 证书和 SASL 身份验证。这样，您便可以通过单个工作站集中管理 VM 主机服务器，无需再单独访问每台 VM 主机服务器。

建议您使用基于 `libvirt` 的工具来管理 VM Guest。`libvirt` 与基于 `libvirt` 的应用程序之间的互操作性已经过测试，SUSE 的支持原则将其视为不可或缺的一部分。

5.1 虚拟化控制台工具

`libvirt` 包含多个用于管理虚拟机的命令行实用程序。最重要的选项如下：

virsh (软件包: `libvirt-client`)

用于管理 VM Guest 的命令行工具，其功能与虚拟机管理器类似。**virsh** 可让您更改 VM Guest 的状态、设置新的 Guest 和设备，或编辑现有配置。**virsh** 还可用于编写 VM Guest 管理操作的脚本。

virsh 将第一个参数作为命令，将后续参数作为此命令的选项：

```
virsh [-c URI] COMMAND DOMAIN-ID [OPTIONS]
```

与 **zypper** 一样，您也可以调用不带命令的 **virsh**。在此情况下，`virsh` 会启动一个外壳并等待您发出命令。此模式非常适合必须运行后续命令的情形：

```
~> virsh -c qemu+ssh://wilber@mercury.example.com/system
Enter passphrase for key '/home/wilber/.ssh/id_rsa':
Welcome to virsh, the virtualization interactive terminal.
```

```
Type: 'help' for help with commands
      'quit' to quit

virsh # hostname
mercury.example.com
```

virt-install (软件包: virt-install)

用于通过 libvirt 库创建新 VM Guest 的命令行工具。它支持通过 VNC 或 **SPICE** 协议进行图形安装。如果指定了适当的命令行参数, **virt-install** 能够以完全无人照管的方式运行。这样便可以轻松地自动完成 Guest 安装。**virt-install** 是虚拟机管理器使用的默认安装工具。

remote-viewer (软件包: virt-viewer)

一个简单的远程桌面查看器。它支持 SPICE 和 VNC 协议。

virt-clone (软件包: virt-install)

一个用于通过 libvirt 超级管理程序管理库克隆现有虚拟机映像的工具。

virt-host-validate (软件包: libvirt-client)

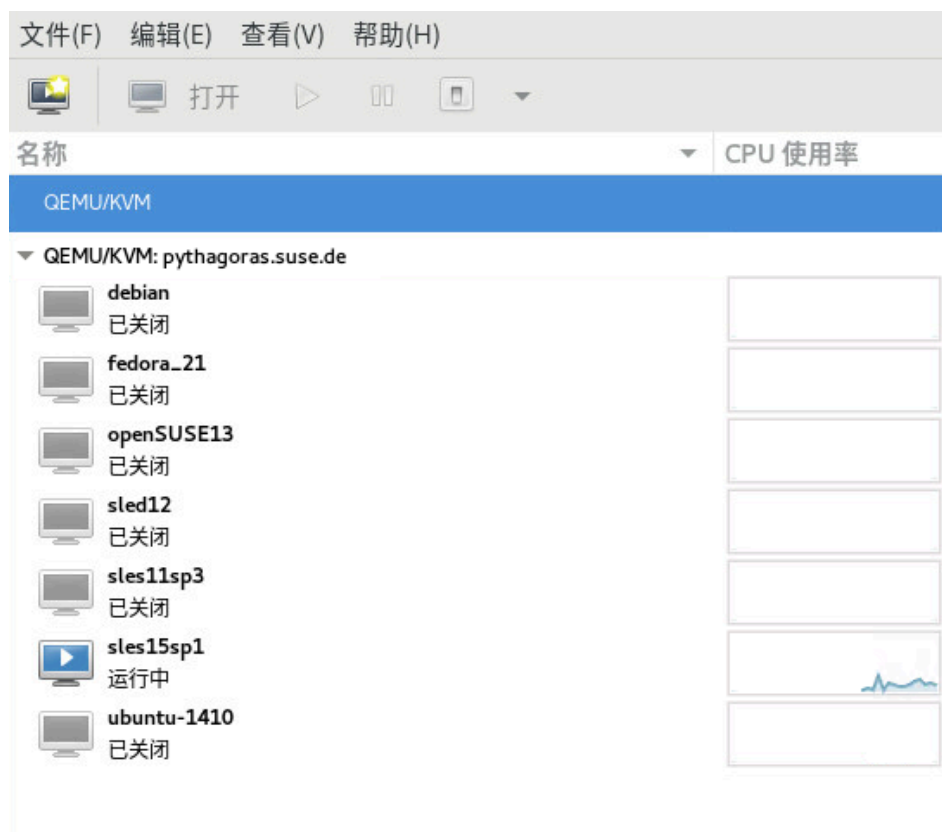
该工具用于验证主机是否经过适当的配置, 可以运行 libvirt 超级管理程序驱动程序。

5.2 虚拟化 GUI 工具

SUSE Linux Enterprise Server 提供了以下基于 libvirt 的图形工具。所有工具由带有相应工具名称的软件包提供。

虚拟机管理器 (软件包: virt-manager)

虚拟机管理器是用于管理 VM Guest 的桌面工具。此工具提供控制现有计算机生命周期 (启动/关机、暂停/继续、保存/恢复) 以及创建新 VM Guest 的功能。它可用于管理多种类型的存储设备和虚拟网络。使用它可以通过内置 VNC 查看器访问 VM Guest 的图形控制台, 以及查看性能统计数据。**virt-manager** 支持连接到本地 libvirtd 来管理本地 VM 主机服务器, 或连接到远程 libvirtd 来管理远程 VM 主机服务器。



要启动虚拟机管理器，请在命令提示符处输入 **virt-manager**。



注意

要禁用通过 spice 对 VM Guest 自动进行 USB 设备重定向的功能，请结合 `--spice-disable-auto-usbredir` 参数启动 **virt-manager**，或运行以下命令来永久更改默认行为：

```
> dconf write /org/virt-manager/virt-manager/console/auto-redirect false
```

virt-viewer (软件包: virt-viewer)

VM Guest 图形控制器的查看器。它使用 SPICE（默认已在 VM Guest 上配置）或 VNC 协议，并支持 TLS 和 x509 证书。可按名称、ID 或 UUID 访问 VM Guest。如果 Guest 尚未运行，可以告知该查看器先等待 Guest 启动，然后再尝试连接到控制台。**virt-viewer** 默认未安装，可在安装软件包 virt-viewer 后使用。



注意

要禁用通过 spice 对 VM Guest 自动进行 USB 设备重定向的功能，请使用 `--spice-usbredir-auto-redirect-filter=''` 参数添加一个空过滤器。

yast2 vm (软件包: yast2-vm)

一个 YaST 模块，可简化虚拟化工具的安装并可设置网桥：

选择要安装的虚拟机管理程序

服务器：运行虚拟机管理程序的最小系统

工具：配置、管理和监视虚拟机

禁止选择的复选框意味着 Hypervisor 项目已被安装

Xen 管理程序

☐ Xen 服务器 ☐ Xen 工具

KVM 管理程序

☒ KVM 服务器 ☒ KVM 工具

取消(C) 接受(A)

6 安装虚拟化组件

6.1 简介

要运行可托管一个或多个 Guest 系统 (VM Guest) 的虚拟化服务器 (VM 主机服务器)，需要在该服务器上安装所需的虚拟化组件。这些组件根据您要使用的虚拟化技术而异。

6.2 安装虚拟化组件

可通过以下方式之一安装运行 VM 主机服务器所需的虚拟化工具：

- 在 VM 主机服务器上安装 SUSE Linux Enterprise Server 期间选择特定的系统角色
- 在已安装并正在运行的 SUSE Linux Enterprise Server 上运行 **YaST 虚拟化** 模块。
- 在已安装并正在运行的 SUSE Linux Enterprise Server 上安装特定的安装软件集。

6.2.1 指定系统角色

您可以在 VM 主机服务器上安装 SUSE Linux Enterprise Server 期间安装虚拟化所需的所有工具。在安装期间，您会看到系统角色屏幕。



图 6.1：“系统角色”屏幕

在此屏幕中，可以选择 KVM 虚拟化主机或 Xen 虚拟化主机角色。在安装 SUSE Linux Enterprise Server 的过程中，系统会自动执行相应软件的选择和设置。



提示

这两个虚拟化系统角色都会创建专用的 `/var/lib/libvirt` 分区，并启用 `firewalld` 和 `Kdump` 服务。

6.2.2 运行 YaST 虚拟化模块

您的系统上可能未安装任何虚拟化工具，具体取决于 VM 主机服务器上的 SUSE Linux Enterprise Server 安装范围。使用 YaST 虚拟化模块配置超级管理程序时，系统会自动安装虚拟化工具。



提示

YaST 虚拟化模块包含在 `yast2-vm` 软件包中。在安装虚拟化组件之前，请校验是否在 VM 主机服务器上安装了该模块。

过程 6.1：安装 KVM 环境

要安装 KVM 虚拟化环境和相关工具，请执行以下操作：

1. 启动 YaST 并选择虚拟化 > 安装管理程序和工具。
2. 选择 KVM 服务器以安装极简的 QEMU 和 KVM 环境。同时请选择 KVM 工具以使用基于 libvirt 的管理堆栈。单击接受确认。
3. YaST 会建议自动在 VM 主机服务器上配置网桥。这可以确保 VM Guest 获得正常的网络功能。如果您同意执行此操作，请选择是，否则请选择否。
4. 安装完成后，您可以开始创建并配置 VM Guest。不需要重引导 VM 主机服务器。

过程 6.2：安装 XEN 环境

要安装 Xen 虚拟化环境，请执行以下操作：

1. 启动 YaST 并选择虚拟化 > 安装管理程序和工具。
2. 选择 Xen 服务器以安装极简的 Xen 环境。同时请选择 Xen 工具以使用基于 libvirt 的管理堆栈。单击接受确认。
3. YaST 会建议自动在 VM 主机服务器上配置网桥。这可以确保 VM Guest 获得正常的网络功能。如果您同意执行此操作，请选择是，否则请选择否。
4. 安装完成后，您需使用 **Xen 内核** 重引导计算机。



提示：默认引导内核

如果一切按预期进行，请使用 YaST 更改默认引导内核，并将支持 Xen 的内核设置为默认内核。有关更改默认内核的详细信息，请参见《管理指南》，第 18 章“引导加载程序 GRUB 2”，第 18.3 节“使用 YaST 配置引导加载程序”。

6.2.3 安装特定的安装软件集

SUSE Linux Enterprise Server 软件储存库中的相关软件包已组织成**安装软件集**。您可以使用这些软件集在已运行的 SUSE Linux Enterprise Server 上安装特定的虚拟化组件。使用 zypper 安装这些组件：

```
zypper install -t pattern PATTERN_NAME
```

要安装 KVM 环境，请考虑使用以下软件集：

kvm_server

安装带有 KVM 和 QEMU 环境的基本 VM 主机服务器。

kvm_tools

安装用于在 KVM 环境中管理和监控 VM Guest 的 libvirt 工具。

要安装 Xen 环境，请考虑使用以下软件集：

xen_server

安装基本的 Xen VM 主机服务器。

xen_tools

安装用于在 Xen 环境中管理和监控 VM Guest 的 libvirt 工具。

6.3 在 KVM 中启用嵌套虚拟化



重要：技术预览

KVM 的嵌套式虚拟化仍为技术预览版。此版本是出于测试目的提供的，我们不提供相关支持。

嵌套式 Guest 是 KVM Guest 中运行的 KVM Guest。在描述嵌套式 Guest 时，我们会用到以下虚拟化层：

L0

运行 KVM 的裸机主机。

L1

在 L0 上运行的虚拟机。由于它可以在另一个 KVM 上运行，因此称为 **Guest 超级管理程序**。

L2

在 L1 上运行的虚拟机。称为**嵌套式 Guest**。

嵌套式虚拟化具有诸多优势。它可在以下场景中为您提供便利：

- 在云环境中直接使用所选的超级管理程序管理您自己的虚拟机。
- 将超级管理程序及其 Guest 虚拟机作为单个实体进行实时迁移。



注意

不支持实时迁移嵌套的 VM Guest。

- 使用嵌套式虚拟化进行软件开发和测试。

要暂时启用嵌套功能，请去除该模块，然后使用 nested KVM 模块参数重新加载该模块：

- 对于 Intel CPU，请运行：

```
> sudo modprobe -r kvm_intel && modprobe kvm_intel nested=1
```

- 对于 AMD CPU，请运行：

```
> sudo modprobe -r kvm_amd && modprobe kvm_amd nested=1
```

要永久启用嵌套功能，请根据您的 CPU，在 /etc/modprobe.d/kvm_*.conf 文件中启用 nested KVM 模块参数：

- 对于 Intel CPU，请编辑 /etc/modprobe.d/kvm_intel.conf，在其中添加下面一行：

```
options kvm_intel nested=1
```

- 对于 AMD CPU，请编辑 /etc/modprobe.d/kvm_amd.conf，在其中添加下面一行：

```
options kvm_amd nested=1
```

如果 L0 主机能够嵌套，则您可以通过以下方式之一启动 L1 Guest：

- 使用 -cpu host QEMU 命令行选项。
- 将 vmx（对于 Intel CPU）或 svm（对于 AMD CPU）CPU 功能添加到 -cpu QEMU 命令行选项，用于启用虚拟 CPU 的虚拟化。

6.3.1 VMware ESX 用作 Guest 超级管理程序

如果您在 KVM 裸机超级管理程序之上使用 VMware ESX 作为 Guest 超级管理程序，网络通讯可能会变得不稳定。嵌套的 KVM Guest 与 KVM 裸机超级管理程序或外部网络之间特别容易发生这种问题。嵌套 KVM Guest 的以下默认 CPU 配置会导致该问题：

```
<cpu mode='host-model' check='partial' />
```

要解决问题，请如下所示修改 CPU 配置：

```
[...]
<cpu mode='host-passthrough' check='none'>
  <cache mode='passthrough' />
</cpu>
[...]
```

7 虚拟化限制和支持

! 重要

仅当与 KVM 或 Xen 超级管理程序一起用于虚拟化时，QEMU 才受支持。TCG 加速器不受支持，即使 SUSE 产品中分发了该工具。用户切勿依赖 QEMU TCG 提供 Guest 隔离或任何安全保障。另请参见 <https://qemu-project.gitlab.io/qemu/system/security.html>。

7.1 体系结构支持

7.1.1 KVM 硬件要求

SUSE 支持在 AMD64/Intel 64、AArch64、IBM Z 和 IBM LinuxONE 主机上实施 KVM 全虚拟化。

- 在 AMD64/Intel 64 体系结构上，KVM 是围绕 AMD* (AMD-V) 和 Intel* (VT-x) CPU 中包含的硬件虚拟化功能设计的。它支持芯片组和 PCI 设备的虚拟化功能，例如 I/O 内存映射单元 (IOMMU) 和单根 I/O 虚拟化 (SR-IOV)。您可以使用以下命令测试您的 CPU 是否支持硬件虚拟化：

```
> egrep '(vmx|svm)' /proc/cpuinfo
```

如果此命令未返回任何输出，则表示您的处理器不支持硬件虚拟化，或者已在 BIOS 或固件中禁用此功能。

以下网站指出了支持硬件虚拟化的 AMD64/Intel 64 处理器：<https://ark.intel.com/Products/VirtualizationTechnology>（针对 Intel CPU），以及 <https://products.amd.com/>（针对 AMD CPU）。

- 在 Arm 体系结构上，Armv8-A 处理器支持虚拟化。
- 在 Arm 体系结构上，仅支持通过 CPU 型号 host（在虚拟机管理器或 libvirt 中名为 host-passthrough）运行 QEMU/KVM。



注意：不加载 KVM 内核模块

仅当 CPU 硬件虚拟化功能可用时，才会加载 KVM 内核模块。

VM 主机服务器的一般性最低硬件要求与《部署指南》，第 2 章“在 AMD64 和 Intel 64 上安装”，第 2.1 节“硬件要求”中概述的相同。不过，对于每个虚拟化的 Guest 都需要提供额外的 RAM。此额外 RAM 量应至少与物理安装所需的 RAM 量相同。另外，强烈建议为每个运行中的 Guest 至少配备一个处理器核心或超线程。



注意：AArch64

AArch64 是个持续发展的平台。它不遵循传统的标准与合规性认证计划来实现与操作系统和超级管理程序的互操作性。请让您的供应商提供针对 SUSE Linux Enterprise Server 的支持声明。



注意：POWER

不支持在 POWER 平台上运行 KVM 或 Xen 超级管理程序。

7.1.2 Xen 硬件要求

SUSE 支持 AMD64/Intel 64 上的 Xen。

7.2 超级管理程序限制

每个服务包 (SP) 的发行说明 (<https://www.suse.com/releasesnotes/>) 中都概述了 Xen 和 KVM 的新功能和虚拟化限制。

仅支持 SUSE Linux Enterprise Server 官方储存库中包含的软件包。相反，[packagehub](https://packagehub.suse.com/) (<https://packagehub.suse.com/>) 中提供的所有可选子软件包和插件（对于 QEMU 为 `libvirt`）均不受支持。

有关每个主机的最大虚拟 CPU 总数，请参见《虚拟化最佳实践》文章，第 4.5.1 节“分配 CPU”。虚拟 CPU 总数应与可用物理 CPU 数成正比。



注意：32 位超级管理程序

我们已从 32 位版本的 SUSE Linux Enterprise Server 11 SP2 中去除虚拟化主机设施。32 位 Guest 不受影响，可以使用提供的 64 位超级管理程序为其提供全面支持。

7.2.1 KVM 限制

在 AMD64/Intel 64 上运行 Linux Guest 的 SUSE Linux Enterprise Server 15 SP7 主机所支持（且经过测试）的虚拟化限制。对于其他操作系统，请咨询具体的供应商。

表 7.1：KVM VM 限制

每个 VM 的最大虚拟 CPU 数	768
每个 VM 的最大内存	4 TiB



注意

KVM 主机限制与 SUSE Linux Enterprise Server 相同（请参见发行说明的相应章节），但以下限制除外：

- **每个 VM 的最大虚拟 CPU 数：**请参见《虚拟化最佳实践》文章, 第 4.5.1 节“分配 CPU”上的《Virtualization Best Practices Guide》中有关过量分配物理 CPU 的建议。虚拟 CPU 总数应与可用物理 CPU 数成正比。

7.2.2 Xen 限制

表 7.2：XEN VM 限制

每个 VM 的最大虚拟 CPU 数	64 (HVM Windows Guest)、128（可信 HVM）或 512 (PV)
每个 VM 的最大内存	2 TiB（64 位 Guest）、16 GiB（具有 PAE 的 32 位 Guest）

表 7.3：XEN 主机限制

最大物理 CPU 总数	1024
每个主机的最大虚拟 CPU 总数	请参见 sec-vt-best-perf-cpu-assign 上的《虚拟化最佳实践指南》中有关过量分配物理 CPU 的建议。虚拟 CPU 总数应与可用物理 CPU 数成正比。
最大物理内存	16 TiB
挂起和休眠模式	不支持。

7.3 支持的主机环境（超级管理程序）

本章介绍 SUSE Linux Enterprise Server 15 SP7 支持在不同虚拟化主机（超级管理程序）上作为 Guest 操作系统运行的情况。

表 7.4：支持以下 SUSE 主机环境

SUSE Linux Enterprise Server	超级管理程序
SUSE Linux Enterprise Server 12SP5	Xen 和 KVM（SUSE Linux Enterprise Server 15 SP6 Guest 必须使用 UEFI 引导）
SUSE Linux Enterprise Server 15 SP3 到 SP7	Xen 和 KVM

支持以下第三方主机环境

- Citrix XenServer (<https://www.citrix.com/products/citrix-hypervisor/>)
- Nutanix Acropolis Hypervisor with AOS (<https://portal.nutanix.com/page/documents/compatibility-matrix/guestos>)
- Oracle VM Server 3.4 (<https://www.oracle.com/fr/virtualization/virtualbox/>)
- Oracle Linux KVM 7, 8 (<https://www.oracle.com/linux/>)
- VMware ESXi 6.7、7.0 (<https://www.vmware.com/products/esxi-and-esx.html>)
- Windows Server 2016、2019、2022

您也可以在 [SUSE YES 认证数据库 \(https://www.suse.com/yessearch/Search.jsp\)](https://www.suse.com/yessearch/Search.jsp) 中进行搜索。

支持级别如下

- 根据相应的[产品生命周期 \(https://www.suse.com/lifecycle/\)](https://www.suse.com/lifecycle/)，对 SUSE 主机操作系统的支持级别为全面支持 L3（适用于 Guest 和主机）。
- SUSE 为第三方主机环境中的 SUSE Linux Enterprise Server Guest 提供全面 L3 支持。
- 对主机的支持以及与 SUSE Linux Enterprise Server Guest 相关的合作必须由主机系统供应商提供。

7.4 支持的 Guest 操作系统

本节列出了在 SUSE Linux Enterprise Server 15 SP7 上虚拟化且适用于 KVM 和 Xen 超级管理程序的 Guest 操作系统的支持状态。



重要

仅当 Guest 中安装了半虚拟化驱动程序时，才能通过 `libvirt/virsh` 重引导 Microsoft Windows Guest。有关下载和安装 PV 驱动程序的更多细节，请参见 <https://www.suse.com/products/vmdriverpack/>。

以下 GUEST 操作系统受到全面支持 (L3):

- SUSE Linux Enterprise Server 12SP5
- SUSE Linux Enterprise Server 15 SP2、15 SP3、15 SP4、15 SP5、15 SP6
- SUSE Linux Enterprise Micro 5.1、5.2、5.3、5.4、5.5、6.0
- Windows Server 2016、2019
- Oracle Linux 6、7、8（仅适用于 KVM 超级管理程序）

以技术预览（L2，在合理的情况下提供修复）的形式支持以下 GUEST 操作系统:

- SLED 15 SP3
- Windows 10/11

如果客户购买了 SUSE MULTI-LINUX SUPPORT，SUSE 将为 RED HAT 和 CENTOS GUEST 操作系统提供全面支持 (L3)。

- 有关可用组合和受支持版本的列表，请参见 <https://documentation.suse.com/liberty> 上的 SUSE Multi-Linux Support 文档。对于其他情况，将为这些操作系统提供有限的支持（L2，在合理的情况下提供修复）。



注意：RHEL PV 驱动程序

从 RHEL 7.2 开始，Red Hat 去除了 Xen PV 驱动程序。

所有其他 GUEST 操作系统

- 在其他组合中提供 L2 支持，但只在可行的情况下才提供修复。SUSE 会对主机操作系统（超级管理程序）提供全面支持。Guest 操作系统问题需要在相应操作系统供应商的支持下予以解决。如果修复某个问题同时涉及到主机环境和 Guest 环境，则客户需要联系 SUSE 和 Guest VM 操作系统供应商。
- 所有 Guest 操作系统既支持全虚拟化，也支持半虚拟化。但以下操作系统例外：Windows 系统，它们只支持全虚拟化（但可以使用 PV 驱动程序：<https://www.suse.com/products/vmdriverpack/>）；OES 操作系统，它们只支持半虚拟化。
- 除非另有说明，否则所有 Guest 操作系统在 32 位和 64 位环境中均受支持。

7.4.1 半虚拟化驱动程序的提供

为了提升 Guest 操作系统的性能，我们将会提供半虚拟化驱动程序（如果有）。尽管这些驱动程序不是必需的，但我们强烈建议使用。

从 SUSE Linux Enterprise Server 12 SP2 开始，我们已改用 PVops 内核。我们不再使用专用的 `kernel-xen` 软件包：

- dom0 上的 `kernel-default+kernel-xen` 已由 `kernel-default` 软件包取代。
- PV domU 上的 `kernel-xen` 软件包已由 `kernel-default` 软件包取代。
- HVM domU 上的 `kernel-default+xen-kmp` 已由 `kernel-default` 取代。

对于 SUSE Linux Enterprise Server 12 SP1 和更低版本（低至 10 SP4），半虚拟化驱动程序包含在专用 `kernel-xen` 软件包中。

半虚拟化驱动程序的提供方式如下：

SUSE Linux Enterprise Server 12/12 SP1/12 SP2

包含在内核中

SUSE Linux Enterprise Server 11/11 SP1/11 SP2/11 SP3/11 SP4

包含在内核中

SUSE Linux Enterprise Server 10SP4

包含在内核中

Red Hat

从 Red Hat Enterprise Linux 5.4 开始提供。从 Red Hat Enterprise Linux 7.2 开始，Red Hat 已去除了 PV 驱动程序。

Windows

SUSE 开发了适用于 Windows 的基于 virtio 的驱动程序，这些驱动程序包含在虚拟机驱动程序包 (VMDP) 中。有关详细信息，请参见 <https://www.suse.com/products/vmdriverpack/> 。

7.5 支持的 VM 迁移场景

SUSE Linux Enterprise Server 支持将虚拟机从一台物理主机迁移到另一台物理主机。

7.5.1 脱机迁移场景

SUSE 支持脱机迁移：关闭 Guest VM，然后将其迁移到运行不同 SLE 产品（从 SLE 12 到 SLE 15 SPX）的主机。对于以下主机操作系统组合，全面支持 (L3) 从一台主机的 Guest 迁移到另一台主机的 Guest：

表 7.5：支持的 GUEST 脱机迁移方案

目标 SLES 主机	12	12	12	15	15	15	15	15	15	15
源 SLES 主机	SP3	SP4	SP5	GA	SP1	SP2	SP3	SP4	SP5	SP6
12 SP3	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
12 SP4	✗	✓	✓	✓ ¹	✓	✗	✗	✗	✗	✗
12 SP5	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗
15 GA	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗
15 SP1	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗
15 SP2	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗
15 SP3	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
15 SP4	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
15 SP5	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
15 SP6	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

✓ 完全兼容且完全受支持

✓¹ 仅支持使用 KVM 超级管理程序

✗ 不支持

7.5.2 实时迁移场景

本节列出了在 SLES 上运行虚拟化时实时迁移方案的支持状态。另请参见受支持方案所要满足的第 17.2 节“迁移要求”。全面支持以下主机操作系统组合（根据相应的产品生命周期 (<https://www.suse.com/lifecycle>) ¹，支持级别为 L3）。

📎 注意：动态迁移

- SUSE 始终支持在运行 SLES 的主机之间实时迁移虚拟机，但这两个 SLES 版本的服务包编号必须是连续的。例如，从 SLES 15 SP4 迁移到 SLES 15 SP5。
- SUSE 致力于为以下实时迁移提供支持：在 SUSE Linux Enterprise Server 的同一主要版本中，将虚拟机从运行 LTSS 所涵盖的服务包的主机迁移到运行更新服务包的主机。例如，将虚拟机从 SLES 12 SP2 主机迁移到 SLES 12 SP5 主机。对于从 LTSS 迁移到更新的服务包的场景，SUSE 只会执行极简单的测试，建议在尝试迁移关键的虚拟机之前执行全面的现场测试。

❗ 重要：Xen 实时迁移

由于工具栈不同，不支持在 SLE 11 与 SLE 12 之间实时迁移。有关更多细节，请参见[发行说明 \(https://www.suse.com/releasenotes/x86_64/SUSE-SLES/12/#fate-317306\)](https://www.suse.com/releasenotes/x86_64/SUSE-SLES/12/#fate-317306)。

❗ 重要：机密计算

SLES 15 SP6 及更高版本中包含内核补丁，以及用于在产品中启用 AMD 和 Intel 机密计算技术的工具。由于此技术尚未完全准备好用于生产环境，因此它作为技术预览提供。

表 7.6：支持的 GUEST 实时迁移

目标 SLES 主机	12	12	15	15	15	15	15	15	15	15
源 SLES 主机	SP4	SP5	GA	SP1	SP2	SP3	SP4	SP5	SP6	SP7
12 SP3	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
12 SP4	✓	✓	✓ ¹	✗	✗	✗	✗	✗	✗	✗
12 SP5	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗
15 GA	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
15 SP1	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗
15 SP2	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗

目标 SLES 主机	12	12	15	15	15	15	15	15	15	15
源 SLES 主机	SP4	SP5	GA	SP1	SP2	SP3	SP4	SP5	SP6	SP7
15 SP3	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
15 SP4	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
15 SP5	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗
15 SP6	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓ ²

- ✓ 完全兼容且完全受支持
- ✓¹ 仅支持使用 KVM 超级管理程序
- ✓² 如果有
- ✗ 不支持

7.6 功能支持

! 重要：嵌套虚拟化：技术预览

使用嵌套虚拟化可在一个 VM 的内部运行另一个虚拟机，同时仍可利用主机的硬件加速功能。嵌套虚拟化的性能较差，并且在调试时会提高复杂性。嵌套虚拟化通常用于测试目的。在 SUSE Linux Enterprise Server 中，嵌套虚拟化是作为技术预览提供的。此版本仅用于测试，我们不提供相关支持。用户可以报告 bug，但这些 bug 的处理优先级较低。此外我们明确指出，在使用嵌套虚拟化的情况下，不支持实时迁移或者保存或恢复 VM。

！ 重要：复制后实时迁移：技术预览

复制后是一种实时迁移虚拟机的方法，旨在让 VM 尽快在目标主机上运行，并视需要随着时间的推移在后台逐步转移 VM RAM。在某些情况下，与传统的预复制方法相比，这可能是一种优化。但此方法存在一个重大的缺点：在迁移过程中发生的错误（尤其是网络故障）可能会导致整个 VM RAM 内容丢失。因此，我们建议仅在生产环境中使用预复制，如果不介意丢失 VM 状态，可将后复制用于进行测试和实验。

7.6.1 Xen 主机 (Dom0)

表 7.7：功能支持 — 主机 (Dom0)

功能	Xen
网络和块设备热插拔	✓
物理 CPU 热插拔	✗
虚拟 CPU 热插拔	✓
虚拟 CPU 固定	✓
虚拟 CPU 限制	✓
Intel* VT-x2: FlexPriority、FlexMigrate（迁移限制适用于不同的 CPU 体系结构）	✓
Intel* VT-d2（具有中断过滤和排队失效的 DMA 重新映射）	✓
AMD* IOMMU（具有 Guest 到主机物理地址转换的 I/O 页表）	✓

🔒 注意：不支持在运行时添加或去除物理 CPU

不支持在运行时添加或去除物理 CPU，但可以脱机添加或去除每个 VM Guest 的虚拟 CPU。

7.6.2 Guest 功能支持



注意：Xen PV Guest 实时迁移

对于实时迁移，源体系结构和目标体系结构均需要匹配；即处理器（AMD* 或 Intel*）必须相同。除非使用了 CPU ID 掩码（例如，使用 Intel FlexMigration），否则目标的处理器修订版应该与源相同或者比源更新。如果在不同系统之间移动 VM，那么这些规则适用于每一次移动。为了避免优化的代码在运行时或应用程序启动期间失败，源 CPU 和目标 CPU 需要公开相同的处理器扩展。Xen 透明地向 VM 公开物理 CPU 扩展。总而言之，Guest 可以是 32 位或 64 位，但 VHS 必须相同。



注意：Windows Guest

仅当使用的是 PV 驱动程序 (VMDP (<https://www.suse.com/products/vmdriverpack/>) ) 时，才支持在 Xen 和 KVM 中热插拔虚拟网络和虚拟块设备，以及收缩、恢复动态虚拟内存及调整其大小。



注意：Intel FlexMigration

对于支持 Intel FlexMigration 的计算机，CPU-ID 掩码和错误引发可让您更灵活地跨 CPU 迁移。



提示

对于 KVM，有关支持的限制、功能、建议的设置和方案的详细说明以及其他有用信息均在 `kvm-supported.txt` 中提供。此文件是 KVM 软件包的一部分，可在 `/usr/share/doc/packages/qemu-kvm` 中找到。

表 7.8：XEN 和 KVM 的 GUEST 功能支持

功能	Xen PV Guest (DomU)	Xen FV Guest	KVM FV Guest
虚拟网络和虚拟块设备热插拔	✓	✓	✓

功能	Xen PV Guest (DomU)	Xen FV Guest	KVM FV Guest
虚拟 CPU 热插拔	✓	✗	✗
虚拟 CPU 过量分配	✓	✓	✓
动态虚拟内存大小调整	✓	✓	✓
VM 保存和恢复	✓	✓	✓
VM 实时迁移	✓ [1]	✓ [1]	✓
VM 快照	✓	✓	✓
使用 GDB 进行高级调试	✓	✓	✓
对 VM 可见的 Dom0 指标	✓	✓	✓
内存气球	✓	✗	✗
PCI 直通	✓ [2]	✓	✓
AMD SEV	✗	✗	✓ [3]

✓ 完全兼容且完全受支持

✗ 不支持

[1] 请参见 第 17.2 节 “迁移要求”。

[2] 不包括 NetWare Guest。

[3] 请参见 <https://documentation.suse.com/sles/html/SLES-amd-sev/article-amd-sev.html>。

II 使用 libvirt 管理虚拟机

- 8 libvirt 守护程序 38
- 9 准备 VM 主机服务器 44
- 10 Guest 安装 73
- 11 基本 VM Guest 管理 84
- 12 连接和授权 104
- 13 高级存储主题 125
- 14 使用虚拟机管理器配置虚拟机 130
- 15 使用 **virsh** 配置虚拟机 150
- 16 使用 AMD SEV-SNP 增强虚拟机安全性 187
- 17 迁移 VM Guest 193
- 18 Xen 到 KVM 的迁移指南 202

8 libvirt 守护程序

访问 KVM 或 Xen 的 libvirt 部署需要在主机上安装并激活一个或多个守护程序。libvirt 提供两个守护程序部署选项：一体化守护程序或模块化守护程序。libvirt 始终提供单个一体化守护程序 libvirtd。它包括主要超级管理程序驱动程序，以及管理存储空间、网络、节点设备等所需的所有次要驱动程序。一体化 libvirtd 还为外部客户端提供安全的远程访问。随着时间的推移，libvirt 增加了对模块化守护程序的支持，其中每个驱动程序都在自身的守护程序中运行，使用户能够自定义其 libvirt 部署。模块化守护程序默认处于启用状态，但可以通过禁用各个守护程序并启用 libvirtd，来将部署切换到传统的一体化守护程序。

在需要最低程度的 libvirt 支持的方案中，模块化守护程序部署非常有用。例如，如果虚拟机存储空间和网络不是由 libvirt 提供的，则不需要 libvirt-daemon-driver-storage 和 libvirt-daemon-driver-network 软件包。Kubernetes 是一个极端的示例，它会处理网络、存储、cgroup 和名称空间集成等方面的所有工作。对于 Kubernetes，只需安装提供 virtqemud 的 libvirt-daemon-driver-QEMU 软件包即可。模块化守护程序允许配置仅包含用例所需组件的自定义 libvirt 部署。

8.1 启动和停止模块化守护程序

模块化守护程序按照它们运行的驱动程序命名，模式为 “virtDRIVERd”。它们是通过文件 `/etc/libvirt/virtDRIVERd.conf` 配置的。SUSE 支持 virtqemud 和 virtxend 超级管理程序守护程序，以及所有次要守护程序：

- **virtnetworkd** - 提供 libvirt 虚拟网络管理 API 的虚拟网络管理守护程序。例如，virtnetworkd 可用于在主机上创建供虚拟机使用的 NAT 虚拟网络。
- **virtnodedevd** - 主机物理设备管理守护程序，提供 libvirt 的节点设备管理 API。例如，virtnodedevd 可用于从主机分离供虚拟机使用的 PCI 设备。
- **virtnwfilterd** - 提供 libvirt 防火墙管理 API 的主机防火墙管理守护程序。例如，virtnwfilterd 可用于为虚拟机配置网络流量过滤规则。
- **virtsecret** - 提供 libvirt 机密管理 API 的主机机密管理守护程序。例如，virtsecret 可用于存储与 LUKS 卷关联的密钥。

- **virtstorage** - 提供 libvirt 存储管理 API 的主机存储管理守护程序。virtstorage 可用于创建存储池，并基于这些池创建卷。
- **virtinterfaced** - 主机 NIC 管理守护程序，提供 libvirt 的主机网络接口管理 API。例如，virtinterfaced 可用于在主机上创建绑定的网络设备。SUSE 不建议使用 libvirt 的接口管理 API，最好使用 wicked 或 NetworkManager 等默认网络工具。建议禁用 virtinterfaced。
- **virtproxyd** - 充当传统 libvirtd 套接字与模块化守护程序套接字之间的连接代理的守护程序。如果使用模块化 libvirt 部署，virtproxyd 将允许远程客户端访问类似于一体化 libvirtd 的 libvirt API。连接到一体化 libvirtd 套接字的本地客户端也可以使用 virtproxyd。
- **virtlogd** - 用于管理虚拟机控制台日志的守护程序。一体化 libvirtd 也使用 virtlogd。一体化守护程序和 virtqemud systemd 单元文件需要 virtlogd，因此无需明确启动 virtlogd。
- **virtlockd** - 用于管理虚拟机资源（例如磁盘）锁的守护程序。一体化 libvirtd 也使用 virtlockd。一体化守护程序、virtqemud 和 virtxend systemd 单元文件需要 virtlockd，因此无需明确启动 virtlockd。

一体化 libvirtd 也使用 virtlogd 和 virtlockd。出于安全考虑，这些守护程序始终是与 libvirtd 分开的。

默认情况下，模块化守护程序会侦听 /var/run/libvirt/virtDRIVERd-sock 和 /var/run/libvirt/virtDRIVERd-sock-ro Unix 域套接字上的连接。客户端库偏向于使用这些套接字而不是传统的 /var/run/libvirt/libvirtd-sock。virtproxyd 守护程序适用于需要传统 libvirtd 套接字的远程客户端或本地客户端。

virtqemud 和 virtxend 服务在 systemd 预设中处于启用状态。virtnetworkd、virtnodedevd、virtnwfilterd、virtstorage 和 virtsecret 的套接字在预设中也处于启用状态，以确保在安装相应的软件包时，这些守护程序已启用且可用。尽管为了方便起见，模块化守护程序已在预设中启用，但也可以使用其 systemd 单元文件进行管理：

- **virtDRIVERd.service** - 用于启动 virtDRIVERd 守护程序的主单元文件。如果 VM 配置为在主机引导时启动，我们建议也将服务配置为在引导时启动。
- **virtDRIVERd.socket** - 与主读写 UNIX 套接字 `/var/run/libvirt/virtDRIVERd-sock` 对应的单元文件。默认情况下，我们建议在引导时启动此套接字。
- **virtDRIVERd-ro.socket** - 与主只读 UNIX 套接字 `/var/run/libvirt/virtDRIVERd-sock-ro` 对应的单元文件。默认情况下，我们建议在引导时启动此套接字。
- **virtDRIVERd-admin.socket** - 与管理 UNIX 套接字 `/var/run/libvirt/virtDRIVERd-admin-sock` 对应的单元文件。默认情况下，我们建议在引导时启动此套接字。

使用 `systemd` 套接字激活时，将不再遵循 `virtDRIVERd.conf` 中的多个配置设置。必须通过系统单元文件控制这些设置：

- **unix_sock_group** - UNIX 套接字组拥有者，通过 `virtDRIVERd.socket` 和 `virtDRIVERd-ro.socket` 单元文件中的 `SocketGroup` 参数进行控制。
- **unix_sock_ro_perms** - 只读 UNIX 套接字权限，通过 `virtDRIVERd-ro.socket` 单元文件中的 `SocketMode` 参数进行控制。
- **unix_sock_rw_perms** - 读写 UNIX 套接字权限，通过 `virtDRIVERd.socket` 单元文件中的 `SocketMode` 参数进行控制。
- **unix_sock_admin_perms** - 管理员 UNIX 套接字权限，通过 `virtDRIVERd-admin.socket` 单元文件中的 `SocketMode` 参数进行控制。
- **unix_sock_dir** - 在其中创建所有 UNIX 套接字的目录，通过以下任意单元文件中的 `ListenStream` 参数独立控制：`virtDRIVERd.socket`、`virtDRIVERd-ro.socket` 和 `virtDRIVERd-admin.socket`。

8.2 启动和停止一体化守护程序

一体化守护程序称为 `libvirtd`，通过 `/etc/libvirt/libvirtd.conf` 配置。可使用多个 `systemd` 单元文件来管理 `libvirtd`：

- **libvirtd.service** - 用于启动 libvirtd 的主 systemd 单元文件。如果 VM 配置为在主机引导时启动，我们建议也将 libvirtd.service 配置为在引导时启动。
- **libvirtd.socket** - 与主读写 UNIX 套接字 /var/run/libvirt/libvirt-sock 对应的单元文件。我们建议在引导时启用此单元。
- **libvirtd-ro.socket** - 与主只读 UNIX 套接字 /var/run/libvirt/libvirt-sock-ro 对应的单元文件。我们建议在引导时启用此单元。
- **libvirtd-admin.socket** - 与管理 UNIX 套接字 /var/run/libvirt/libvirt-admin-sock 对应的单元文件。我们建议在引导时启用此单元。
- **libvirtd-tcp.socket** - 与用于进行非 TLS 远程访问的 TCP 16509 端口对应的单元文件。在管理员已配置适当的身份验证机制之前，不应将此单元配置为在引导时启动。
- **libvirtd-tls.socket** - 与用于进行 TLS 远程访问的 TCP 16509 端口对应的单元文件。在管理员已部署 x509 证书并选择性地配置适当的身份验证机制之前，不应将此单元配置为在引导时启动。

使用 systemd 套接字激活时，将不再遵循 libvirtd.conf 中的某些配置设置。必须通过系统单元文件控制这些设置：

- **listen_tcp** - 通过启动 libvirtd-tcp.socket 单元文件启用 TCP 套接字。
- **listen_tls** - 通过启动 libvirtd-tls.socket 单元文件启用 TLS 套接字。
- **tcp_port** - 非 TLS TCP 套接字的端口，通过 libvirtd-tcp.socket 单元文件中的 ListenStream 参数进行控制。
- **tls_port** - TLS TCP 套接字的端口，通过 libvirtd-tls.socket 单元文件中的 ListenStream 参数进行控制。
- **listen_addr** - 要侦听的 IP 地址，通过 libvirtd-tcp.socket 或 libvirtd-tls.socket 单元文件中的 ListenStream 参数独立控制。
- **unix_sock_group** - UNIX 套接字组拥有者，通过 libvirtd.socket 和 libvirtd-ro.socket 单元文件中的 SocketGroup 参数进行控制。
- **unix_sock_ro_perms** - 只读 UNIX 套接字权限，通过 libvirtd-ro.socket 单元文件中的 SocketMode 参数进行控制。

- **unix_sock_rw_perms** - 读写 UNIX 套接字权限，通过 `libvirtd.socket` 单元文件中的 `SocketMode` 参数进行控制。
- **unix_sock_admin_perms** - 管理员 UNIX 套接字权限，通过 `libvirtd-admin.socket` 单元文件中的 `SocketMode` 参数进行控制。
- **unix_sock_dir** - 在其中创建所有 UNIX 套接字的目录，通过以下任意单元文件中的 `ListenStream` 参数独立控制：`libvirtd.socket`、`libvirtd-ro.socket` 和 `libvirtd-admin.socket`。

! 重要：有冲突的服务：libvirtd 和 xendomains

如果 `libvirtd` 无法启动，请检查是否加载了 `xendomains` 服务：

```
> systemctl is-active xendomains active
```

如果该命令返回 `active`，您需要停止 `xendomains`，然后才可以启动 `libvirtd` 守护程序。如果您希望在重引导后也要启动 `libvirtd`，另外还需禁止 `xendomains` 自动启动。禁用该服务：

```
> sudo systemctl stop xendomains
> sudo systemctl disable xendomains
> sudo systemctl start libvirtd
```

`xendomains` 和 `libvirtd` 提供相同的服务，如果同时使用，可能会互相干扰。例如，`xendomains` 可能会尝试启动已由 `libvirtd` 启动的 `domU`。

8.3 切换到一体化守护程序

要从模块化守护程序切换到一体化守护程序，需要更改多个服务。在守护程序选项之间切换之前，建议停止或逐出所有正在运行的虚拟机。

1. 停止模块化守护程序及其套接字。以下示例会禁用 KVM 的 QEMU 守护程序和几个次要守护程序。

```
for drv in qemu network nodedev nwfilter secret storage
```

```
do
> sudo systemctl stop virt${drv}d.service
> sudo systemctl stop virt${drv}d{,-ro,-admin}.socket
done
```

2. 禁止将来启动模块化守护程序

```
for drv in qemu network nodedev nwfilter secret storage
do
> sudo systemctl disable virt${drv}d.service
> sudo systemctl disable virt${drv}d{,-ro,-admin}.socket
done
```

3. 启用一体化 libvirtd 服务和套接字

```
> sudo systemctl enable libvirtd.service
> sudo systemctl enable libvirtd{,-ro,-admin}.socket
```

4. 启动一体化 libvirtd 套接字

```
> sudo systemctl start libvirtd{,-ro,-admin}.socket
```

9 准备 VM 主机服务器

在可以安装 Guest 虚拟机之前，需要准备好 VM 主机服务器，以便为 Guest 提供正常运行所需的资源。具体而言，您需要配置：

- **网络**：使 Guest 能够利用主机提供的网络连接。
- 一个可从主机访问的**存储池**，便于 Guest 存储其磁盘映像。

9.1 配置网络

有两个常用网络配置可为 VM Guest 提供网络连接：

- **网桥**。这是为 Guest 提供网络连接的默认方式，也是建议的方式。
- 已启用转发功能的**虚拟网络**。

9.1.1 网桥

网桥配置为 VM Guest 提供第 2 层交换机，可以基于与端口关联的 MAC 地址在网桥上的端口之间交换第 2 层以太网包。这样 VM Guest 便可通过第 2 层连接访问 VM 主机服务器的网络。此配置类似于将 VM Guest 的虚拟以太网网线连接到与主机以及主机上运行的其他 VM Guest 共享的集线器。该配置通常称为**共享物理设备**。

当 SUSE Linux Enterprise Server 配置为 KVM 或 Xen 超级管理程序时，网桥配置便是它的默认配置。如果您只想将 VM Guest 连接到 VM 主机服务器的 LAN，则此为首选配置。

要使用哪个工具创建网桥取决于您在 VM 主机服务器上使用哪个服务来管理网络连接：

- 如果网络连接由 wicked 管理，请使用 YaST 或命令行来创建网桥。服务器主机默认使用 wicked。
- 如果网络连接由 NetworkManager 管理，请使用 NetworkManager 命令行工具 nmcli 创建网桥。台式机和笔记本电脑默认使用 NetworkManager。

9.1.1.1 使用 YaST 管理网桥

本节包含使用 YaST 添加或去除网桥的过程。

9.1.1.1.1 添加网桥

要在 VM 主机服务器上添加网桥，请执行以下步骤：

1. 启动 YaST > 系统 > 网络设置。
2. 进入概览选项卡并单击添加。
3. 从设备类型列表中选择网桥，然后在配置名称项中输入网桥设备接口名称。单击下一步按钮继续。
4. 在地址选项卡中指定网络细节，例如 DHCP/静态 IP 地址、子网掩码或主机名。
仅当您还将设备分配到与 DHCP 服务器连接的网桥时，才需要使用动态地址。
如果您打算创建不与实际网络设备连接的虚拟网桥，请使用静态分配的 IP 地址。在这种情况下，比较好的做法是使用私用 IP 地址范围（例如 192.168.0.0/16、172.16.0.0/12 或 10.0.0.0/8）内的地址。
要创建仅在不同 Guest 之间充当连接点，而不连接到主机系统的网桥，请将 IP 地址设置为 0.0.0.0，将子网掩码设置为 255.255.255.255。网络脚本会将此特殊地址作为未设置的 IP 地址来处理。
5. 进入桥接设备选项卡，然后选中您要包含在网桥中的网络设备。
6. 单击下一步返回到概览选项卡，然后单击确定进行确认。现在，新网桥应在 VM 主机服务器上处于活动状态。

9.1.1.1.2 删除网桥

要删除现有网桥，请执行以下步骤：

1. 启动 YaST > 系统 > 网络设置。
2. 从概览选项卡上的列表中选择您要删除的网桥设备。

3. 单击删除以删除该网桥，然后单击确定进行确认。

9.1.1.2 通过命令行管理网桥

本节包含使用命令行添加或去除网桥的过程。

9.1.1.2.1 添加网桥

要在 VM 主机服务器上添加新网桥设备，请执行以下步骤：

1. 在要创建新网桥的 VM 主机服务器上以 `root` 身份登录。
2. 为新网桥选择一个名称（在本示例中为 `virbr_test`），然后运行以下命令

```
# ip link add name VIRBR_TEST type bridge
```

3. 检查是否已在 VM 主机服务器上创建了网桥：

```
# bridge vlan  
[...]  
virbr_test 1 PVID Egress Untagged
```

`virbr_test` 存在，但不与任何物理网络接口相关联。

4. 启动该网桥，并在其中添加一个网络接口：

```
# ip link set virbr_test up  
# ip link set eth1 master virbr_test
```



重要：网络接口必须未被使用

只能分配尚未被其他网桥使用的网络接口。

5. （可选）启用 STP（请参见[生成树协议 \(https://en.wikipedia.org/wiki/Spanning_Tree_Protocol\)](https://en.wikipedia.org/wiki/Spanning_Tree_Protocol) )

```
# bridge link set dev virbr_test cost 4
```

9.1.1.2.2 删除网桥

要通过命令行删除 VM 主机服务器上的现有网桥设备，请执行以下步骤：

1. 在要从中删除现有网桥的 VM 主机服务器上以 root 身份登录。
2. 列出现有网桥，以识别要去除的网桥的名称：

```
# bridge vlan
[...]  
virbr_test  1 PVID Egress Untagged
```

3. 删除网桥：

```
# ip link delete dev virbr_test
```

9.1.1.3 使用 nmcli 添加网桥

本节介绍使用 NetworkManager 的命令行工具 nmcli 添加网桥的过程。

1. 列出活动网络连接：

```
> sudo nmcli connection show --active  
NAME                                UUID                                TYPE  
DEVICE  
Ethernet connection 1  84ba4c22-0cfe-46b6-87bb-909be6cb1214  ethernet  eth0
```

2. 添加名为 br0 的新网桥设备，并校验是否已创建该设备：

```
> sudo nmcli connection add type bridge ifname br0  
Connection 'bridge-br0' (36e11b95-8d5d-4a8f-9ca3-ff4180eb89f7) \  
successfully added.  
> sudo nmcli connection show --active  
NAME                                UUID                                TYPE  
DEVICE  
bridge-br0                        36e11b95-8d5d-4a8f-9ca3-ff4180eb89f7  bridge    br0  
Ethernet connection 1  84ba4c22-0cfe-46b6-87bb-909be6cb1214  ethernet  eth0
```

3. （可选）可以查看网桥设置：

```
> sudo nmcli -f bridge connection show bridge-br0
bridge.mac-address:      --
bridge.stp:              yes
bridge.priority:         32768
bridge.forward-delay:    15
bridge.hello-time:       2
bridge.max-age:          20
bridge.ageing-time:      300
bridge.group-forward-mask: 0
bridge.multicast-snooping: yes
bridge.vlan-filtering:   no
bridge.vlan-default-pvid: 1
bridge.vlans:            --
```

4. 将网桥设备链接到物理以太网设备 eth0:

```
> sudo nmcli connection add type bridge-slave ifname eth0 master br0
```

5. 禁用 eth0 接口并启用新网桥:

```
> sudo nmcli connection down "Ethernet connection 1"
> sudo nmcli connection up bridge-br0
Connection successfully activated (master waiting for slaves) \
(D-Bus active path: /org/freedesktop/NetworkManager/ActiveConnection/9)
```

9.1.1.4 使用 VLAN 接口

有时需要在两台 VM 主机服务器之间或者 VM Guest 系统之间创建私用连接。例如，将 VM Guest 迁移到其他网段中的主机时，或者创建只有 VM Guest 系统能够连接到的私用网桥（即使是在不同 VM 主机服务器系统上运行）时。构建此类连接的简单方法是设置 VLAN 网络。

VLAN 接口通常在 VM 主机服务器上设置。它们可以将不同的 VM 主机服务器系统互连，或者可以设置为其他仅限虚拟连接的网桥的物理接口。甚至可以创建一个使用 VLAN 作为物理接口（该接口在 VM 主机服务器中没有 IP 地址）的网桥。这样，Guest 系统就无法通过此网络访问主机。

运行 YaST 模块系统 > 网络设置。执行以下过程设置 VLAN 设备：

过程 9.1：使用 YAST 设置 VLAN 接口

1. 单击添加以创建新网络接口。
2. 在硬件对话框中，选择设备类型 VLAN。
3. 将配置名称的值更改为 VLAN 的 ID。请注意，VLAN ID 1 通常用于管理目的。
4. 单击下一步。
5. 选择 VLAN 设备应连接的接口虚拟局域网的真实接口。如果所需的接口未显示在列表中，请先在不指定 IP 地址的情况下设置此接口。
6. 选择将 IP 地址分配到 VLAN 设备的所需方法。
7. 单击下一步完成配置。

还可将 VLAN 接口用作网桥的物理接口。这样便可以连接多个仅限 VM 主机服务器的网络，以及实时迁移与此类网络连接的 VM Guest 系统。

YaST 并非始终允许不设置 IP 地址。但有时可能需要这种功能，尤其是应该连接仅限 VM 主机服务器的网络时。在这种情况下，请使用特殊地址 0.0.0.0 和网络掩码 255.255.255.255。系统脚本会将此地址当作未设置 IP 地址来处理。

9.1.2 虚拟网络

libvirt 管理的虚拟网络类似于桥接的网络，但通常不与 VM 主机服务器建立第 2 层连接。与 VM 主机服务器物理网络的连接通过第 3 层转发来实现，与第 2 层桥接网络相比，第 3 层转发在 VM 主机服务器上引入了额外的包处理。虚拟网络还为 VM Guest 提供 DHCP 和 DNS 服务。有关 libvirt 虚拟网络的详细信息，请参见 <https://libvirt.org/formatnetwork.html> 上的《Network XML format》文档。

SUSE Linux Enterprise Server 上的标准 libvirt 安装中已预定义了一个名为 default 的虚拟网络。该虚拟网络提供网络的 DHCP 和 DNS 服务，并可使用网络地址转换 (NAT) 转发模式连接到 VM 主机服务器的物理网络。尽管 default 虚拟网络是预定义的，但它需要由管理员明确启用。有关 libvirt 支持的转发模式的详细信息，请参见 <https://libvirt.org/formatnetwork.html#elementsConnect> 上《Network XML format》文档中的“Connectivity”一节。

libvirt 管理的虚拟网络可用于满足各种用例，但通常是在进行无线连接或动态/零星网络连接的 VM 主机服务器（例如便携式计算机）上使用。虚拟网络也适用于 VM 主机服务器网络的 IP 地址有限的情形，可用来在虚拟网络与 VM 主机服务器的网络之间转发包。不过，大多数服务器用例更适合网桥配置，其中的 VM Guest 会连接到 VM 主机服务器的 LAN。



警告：启用转发模式

如果在 libvirt 虚拟网络中启用转发模式，会将 `/proc/sys/net/ipv4/ip_forward` 和 `/proc/sys/net/ipv6/conf/all/forwarding` 设置为 1，从而在 VM 主机服务器中启用转发，而这本质上是将 VM 主机服务器转变成路由器。如果重新启动 VM 主机服务器的网络，可能会重置这些值并禁用转发。要避免这种行为，请通过编辑 `/etc/sysctl.conf` 文件并添加以下设置，在 VM 主机服务器中明确启用转发：

```
net.ipv4.ip_forward = 1
```

```
net.ipv6.conf.all.forwarding = 1
```

9.1.2.1 使用虚拟机管理器管理虚拟网络

您可以使用虚拟机管理器定义、配置和操作虚拟网络。

9.1.2.1.1 定义虚拟网络

1. 启动虚拟机管理器。在可用连接列表中，右键单击您需要为其配置虚拟网络的连接名称，然后选择细节。
2. 在连接详情窗口中，单击虚拟网络选项卡。您会看到可用于当前连接的所有虚拟网络的列表。右侧会显示选定虚拟网络的细节。



图 9.1：连接细节

3. 要添加新虚拟网络，请单击添加。

4. 为新虚拟机指定名称。



图 9.2：创建虚拟网络

5. 指定网络模式。对于 NAT 和路由类型，可以指定要将网络通讯转发到的设备。NAT（网络地址转换）会重新映射虚拟网络地址空间并允许共享单个 IP 地址，路由会将来自虚拟网络的数据包转发到 VM 主机服务器的物理网络而不进行任何转换。
6. 如果您需要 IPv4 网络，请选中启用 IPv4 并指定 IPv4 网络地址。如果您需要 DHCP 服务器，请选中启用 DHCPv4 并指定可分配的 IP 地址范围。
7. 如果您需要 IPv6 网络，请选中启用 IPv6 并指定 IPv6 网络地址。如果您需要 DHCP 服务器，请选中启用 DHCPv6 并指定可分配的 IP 地址范围。
8. 要指定与虚拟网络名称不同的域名，请在 DNS 域名下选择自定义，并在此处输入所需的域名。
9. 单击完成以创建新虚拟网络。VM 主机服务器上即会有一个新的虚拟网桥 `virbrX`，该网桥对应于新建的虚拟网络。您可以使用 **bridge link** 进行检查。`libvirt` 会自动添加 `iptables` 规则来允许传入/传出挂接到新 `virbrX` 设备的 Guest 的流量。

9.1.2.1.2 启动虚拟网络

要启动某个暂时停止的虚拟网络，请执行以下步骤：

1. 启动虚拟机管理器。在可用连接列表中，右键单击您需要为其配置虚拟网络的连接名称，然后选择细节。
2. 在连接详情窗口中，单击虚拟网络选项卡。您会看到可用于当前连接的所有虚拟网络的列表。
3. 要启动虚拟网络，请单击启动。

9.1.2.1.3 停止虚拟网络

要停止活动的虚拟网络，请执行以下步骤：

1. 启动虚拟机管理器。在可用连接列表中，右键单击您需要为其配置虚拟网络的连接名称，然后选择细节。
2. 在连接详情窗口中，单击虚拟网络选项卡。您会看到可用于当前连接的所有虚拟网络的列表。

3. 选择要停止的虚拟网络，然后单击停止。

9.1.2.1.4 删除虚拟网络

要删除 VM 主机服务器中的虚拟网络，请执行以下步骤：

1. 启动虚拟机管理器。在可用连接列表中，右键单击您需要为其配置虚拟网络的连接名称，然后选择细节。
2. 在连接详情窗口中，单击虚拟网络选项卡。您会看到可用于当前连接的所有虚拟网络的列表。
3. 选择要删除的虚拟网络，然后单击删除。

9.1.2.1.5 使用 nsswitch 获取 NAT 网络的 IP 地址（在 KVM 中）

- 在 VM 主机服务器上，安装用来为 libvirt 提供 NSS 支持的 libvirt-nss：

```
> sudo zypper in libvirt-nss
```

- 将 `libvirt` 添加到 `/etc/nsswitch.conf`：

```
...
hosts:  files libvirt mdns_minimal [NOTFOUND=return] dns
...
```

- 如果 NSCD 正在运行，请将其重新启动：

```
> sudo systemctl restart nscd
```

现在，您可以从主机按名称访问 Guest 系统了。

NSS 模块的功能有限。它会读取 `/var/lib/libvirt/dnsmasq/*.status` 文件，以在描述 `dnsmasq` 所提供的每个租约的 JSON 记录中查找主机名和对应的 IP 地址。只能使用受 `dnsmasq` 支持且由 libvirt 管理的桥接网络在这些 VM 主机服务器上执行主机名转换。

9.1.2.2 使用 **virsh** 管理虚拟网络

可以使用 **virsh** 命令行工具来管理 **libvirt** 提供的虚拟网络。要查看所有与网络相关的 **virsh** 命令，请运行以下命令

```
> sudo virsh help network
Networking (help keyword 'network'):
net-autostart                autostart a network
    net-create                create a network from an XML file
    net-define                define (but don't start) a network from
an XML file
    net-destroy               destroy (stop) a network
    net-dumpxml               network information in XML
    net-edit                  edit XML configuration for a network
    net-event                 Network Events
    net-info                  network information
    net-list                  list networks
    net-name                  convert a network UUID to network name
    net-start                 start a (previously defined) inactive
network
    net-undefine              undefine an inactive network
    net-update                update parts of an existing network's
configuration
net-uuid                     convert a network name to network UUID
```

要查看特定 **virsh** 命令的简要帮助信息，请运行 **virsh help VIRSH_COMMAND**:

```
> sudo virsh help net-create
NAME
    net-create - create a network from an XML file

SYNOPSIS
    net-create <file>

DESCRIPTION
    Create a network.

OPTIONS
    [--file] <string>  file containing an XML network description
```

9.1.2.2.1 创建网络

要创建新的**运行中**虚拟网络，请运行以下命令

```
> sudo virsh net-create VNET_DEFINITION.xml
```

VNET_DEFINITION.xml XML 文件包含 libvirt 接受的虚拟网络的定义。

要定义新虚拟网络但不激活它，请运行以下命令

```
> sudo virsh net-define VNET_DEFINITION.xml
```

以下示例说明了不同类型的虚拟网络的定义。

例 9.1：基于 NAT 的网络

下面的配置允许进行 VM Guest 传出连接（如果 VM 主机服务器上提供此功能）。当没有 VM 主机服务器网络时，此配置可让 Guest 互相直接通讯。

```
<network>
<name>vnet_nated</name> ❶
<bridge name="virbr1"/> ❷
<forward mode="nat"/> ❸
<ip address="192.168.122.1" netmask="255.255.255.0"> ❹
  <dhcp>
    <range start="192.168.122.2" end="192.168.122.254"/> ❺
    <host mac="52:54:00:c7:92:da" name="host1.testing.com" \
      ip="192.168.1.101"/> ❻
    <host mac="52:54:00:c7:92:db" name="host2.testing.com" \
      ip="192.168.1.102"/>
    <host mac="52:54:00:c7:92:dc" name="host3.testing.com" \
      ip="192.168.1.103"/>
  </dhcp>
</ip>
</network>
```

- ❶ 新虚拟网络的名称。
- ❷ 用于构造虚拟网络的网桥设备的名称。定义 `<forward>` 模式为 `"nat"` 或 `"route"` 的新网络（或者不包含 `<forward>` 元素的隔离网络）时，libvirt 将自动为网桥设备生成唯一的名称（如果未指定名称）。

- ③ 包含 `<forward>` 元素表示该虚拟网络将连接到物理 LAN。`mode` 属性指定转发方法。最常用的模式为 `"nat"`（网络地址转换，默认值）、`"route"`（直接转发到物理网络，不执行地址转换）和 `"bridge"`（在 `libvirt` 外部配置的网桥）。如果不指定 `<forward>` 元素，虚拟网络将与其他网络相隔离。有关转发模式的完整列表，请参见 <https://libvirt.org/formatnetwork.html#elementsConnect>。
- ④ 网桥的 IP 地址和网络掩码。
- ⑤ 为虚拟网络启用 DHCP 服务器，并提供 `start` 和 `end` 属性所指定的范围内的 IP 地址。
- ⑥ 可选的 `<host>` 元素指定内置 DHCP 服务器要为其分配名称和预定义 IP 地址的主机。任何 IPv4 `host` 元素都必须指定以下设置：要被分配给定名称的主机的 MAC 地址、要分配到该主机的 IP，以及 DHCP 服务器要为该主机分配的名称。IPv6 `host` 元素与 IPv4 略有不同，它没有 `mac` 属性，因为 MAC 地址在 IPv6 中没有明确的含义。IPv6 中使用 `name` 属性来标识要为其分配 IPv6 地址的主机。对于 DHCPv6，`name` 是由客户端发送到服务器的客户端主机的纯文本名称。也可以使用这种分配特定 IP 地址的方法来代替 IPv4 中的 `mac` 属性。

例 9.2：路由网络

下面的配置会在不应用任何 NAT 的情况下将流量从虚拟网络路由到 LAN。必须在 VM 主机服务器网络上的路由器的路由表中预定义 IP 地址范围。

```
<network>
  <name>vnet_routed</name>
  <bridge name="virbr1"/>
  <forward mode="route" dev="eth1"/> ❶
  <ip address="192.168.122.1" netmask="255.255.255.0">
    <dhcp>
      <range start="192.168.122.2" end="192.168.122.254"/>
    </dhcp>
  </ip>
</network>
```

- ❶ Guest 流量只能通过 VM 主机服务器上的 `eth1` 网络设备传出。

例 9.3：隔离网络

此配置提供隔离的专用网络。Guest 可以相互通讯以及与 VM 主机服务器通讯，但无法访问 LAN 上的任何其他计算机，因为 XML 说明中没有 <forward> 元素。

```
<network>
  <name>vnet_isolated</name>
  <bridge name="virbr3"/>
  <ip address="192.168.152.1" netmask="255.255.255.0">
    <dhcp>
      <range start="192.168.152.2" end="192.168.152.254"/>
    </dhcp>
  </ip>
</network>
```

例 9.4：使用 VM 主机服务器上的现有网桥

此配置说明如何使用 VM 主机服务器的现有网桥 br0。VM Guest 直接连接到物理网络。其 IP 地址全部都在物理网络的子网中，并且传入和传出连接不存在任何限制。

```
<network>
  <name>host-bridge</name>
  <forward mode="bridge"/>
  <bridge name="br0"/>
</network>
```

9.1.2.2.2 列出网络

要列出 `libvirt` 可用的所有虚拟网络，请运行以下命令：

```
> sudo virsh net-list --all
```

Name	State	Autostart	Persistent

crowbar	active	yes	yes
vnet_nated	active	yes	yes
vnet_routed	active	yes	yes
vnet_isolated	inactive	yes	yes

要列出可用域，请运行以下命令：

```
> sudo virsh list
Id      Name                                State
-----
1       nated_sles12sp3                    running
...
```

要获取运行中域的接口列表，请运行 `domifaddr DOMAIN`，或选择性地指定接口，以在输出中仅列出此接口。默认情况下，会额外输出接口的 IP 和 MAC 地址：

```
> sudo virsh domifaddr nated_sles12sp3 --interface vnet0 --source lease
Name      MAC address      Protocol  Address
-----
vnet0     52:54:00:9e:0d:2b  ipv6     fd00:dead:beef:55::140/64
-         -                 ipv4     192.168.100.168/24
```

要列显与指定域关联的所有虚拟接口的简要信息，请运行以下命令：

```
> sudo virsh domiflist nated_sles12sp3
Interface  Type      Source      Model      MAC
-----
vnet0      network  vnet_nated  virtio     52:54:00:9e:0d:2b
```

9.1.2.2.3 获取有关网络的细节

要获取有关网络的详细信息，请运行以下命令：

```
> sudo virsh net-info vnet_routed
Name:          vnet_routed
UUID:          756b48ff-d0c6-4c0a-804c-86c4c832a498
Active:        yes
Persistent:    yes
Autostart:     yes
Bridge:        virbr5
```

9.1.2.2.4 启动网络

要启动某个已定义的非活动网络，请使用以下命令查找其名称（或唯一标识符，即 UUID）：

```
> sudo virsh net-list --inactive
```

Name	State	Autostart	Persistent
vnet_isolated	inactive	yes	yes

然后运行:

```
> sudo virsh net-start vnet_isolated
Network vnet_isolated started
```

9.1.2.2.5 停止网络

要停止某个活动的网络，请使用以下命令查找其名称（或唯一标识符，即 UUID）：

```
> sudo virsh net-list --inactive
```

Name	State	Autostart	Persistent
vnet_isolated	active	yes	yes

然后运行:

```
> sudo virsh net-destroy vnet_isolated
Network vnet_isolated destroyed
```

9.1.2.2.6 去除网络

要从 VM 主机服务器中永久去除某个非活动网络的定义，请运行以下命令：

```
> sudo virsh net-undefine vnet_isolated
Network vnet_isolated has been undefined
```

9.2 配置存储池

在 VM 主机服务器本身上管理 VM Guest 时，您可以访问 VM 主机服务器的整个文件系统，以挂接或创建虚拟硬盘，或将现有映像挂接到 VM Guest。但通过远程主机管理 VM Guest 时无法做到这些。出于此原因，libvirt 支持可从远程计算机访问的所谓“存储池”。



提示：CD/DVD ISO 映像

如果希望能够从远程客户端访问 VM 主机服务器上的 CD/DVD ISO 映像，需要将这些映像也放在存储池中。

libvirt 可识别两种不同类型的存储资源：卷和池。

存储卷

存储卷是可分配到 Guest 的存储设备 — 虚拟磁盘或 CD/DVD/软盘映像。从物理上讲，它可以是块设备（例如分区或逻辑卷），也可以是 VM 主机服务器上的某个文件。

存储池

存储池是 VM 主机服务器上可用于存储卷的存储资源，类似于台式计算机的网络存储空间。从物理上而言，它可分为以下类型之一：

文件系统目录 (dir)

用于存放映像文件的目录。文件可以是支持的磁盘格式之一（raw 或 qcow2），也可以是 ISO 映像。

物理磁盘设备 (disk)

使用整个物理磁盘作为存储空间。系统会为添加到池的每个卷创建一个分区。

预格式化的块设备 (fs)

指定要使用的分区，该分区与文件系统目录池（用于存放映像文件的目录）的使用方式相同。唯一的区别在于，使用文件系统目录时，libvirt 会负责挂载设备。

iSCSI 目标 (iscsi)

在 iSCSI 目标上设置池。您需要先登录卷一次才能将卷用于 libvirt。使用 YaST iSCSI 发起端来检测和登录卷，详情请参见《存储管理指南》。不支持在 iSCSI 池中创建卷；每个现有逻辑单元号 (LUN) 都代表一个卷。每个卷/LUN 还需要一个有效的（空）分区表或磁盘标签，这样您才能使用该卷。如果没有分区表或磁盘标签，请使用 fdisk 添加：

```
> sudo fdisk -cu /dev/disk/by-path/ip-192.168.2.100:3260-iscsi-  
iqn.2010-10.com.example:[...]-lun-2  
Device contains neither a valid DOS partition table, nor Sun, SGI  
or OSF disklabel
```

```
Building a new DOS disklabel with disk identifier 0xc15cdc4e.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.

Warning: invalid flag 0x0000 of partition table 4 will be corrected by
w(rite)

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.
```

LVM 卷组 (logical)

使用 LVM 卷组作为池。您可以使用预定义的卷组，或者通过指定要使用的设备来创建组。存储卷会创建为卷上的分区。



警告：删除基于 LVM 的池

在存储管理器中删除基于 LVM 的池时，也会删除卷组。这会导致池中存储的所有数据丢失且不可恢复。

多路径设备 (mpath)

目前，多路径支持仅限于向 Guest 分配现有设备。不支持从 libvirt 内部创建卷或配置多路径。

网络导出的目录 (netfs)

指定要使用的网络目录，该网络目录与文件系统目录池（用于存放映像文件的目录）的使用方式相同。唯一的区别在于，使用文件系统目录时，libvirt 会负责挂载目录。支持的协议为 NFS。

SCSI 主机适配器 (scsi)

SCSI 主机适配器的使用方式与 iSCSI 目标基本相同。我们建议使用基于 /dev/disk/by-* 的设备名称，而不要使用 /dev/sdX。后者可能会发生变化（例如，添加或拆除硬盘时）。不支持在 iSCSI 池中创建卷。每个现有 LUN（逻辑单元号）都代表一个卷。



警告：安全考虑因素

为了避免数据丢失或损坏，请不要尝试使用也用于在 VM 主机服务器上构建储存池的资源，例如 LVM 卷组、iSCSI 目标等。无需从 VM 主机服务器连接到这些资源，也无需在 VM 主机服务器上挂载这些资源 — [libvirt](#) 将负责这些事项。

不要在 VM 主机服务器上按标签挂载分区。在某些情况下，分区可能是从 VM Guest 内部使用 VM 主机服务器上的现有名称标记的。

9.2.1 使用 **virsh** 管理存储

也可以使用 **virsh** 通过命令行管理存储。不过，SUSE 目前不支持创建存储池。因此，本节仅会介绍启动、停止和删除池以及卷管理等功能。

运行 **virsh help pool** 和 **virsh help volume** 可分别获取用于管理池和卷的所有 **virsh** 子命令的列表。

9.2.1.1 列出池和卷

执行以下命令可以列出当前处于活动状态的所有池。要同时列出非活动池，请添加选项 **--all**:

```
> virsh pool-list --details
```

可以使用 **pool-info** 子命令获取有关特定池的细节：

```
> virsh pool-info POOL
```

默认情况下，只能按池列出卷。要列出某个池中的所有卷，请输入以下命令。

```
> virsh vol-list --details POOL
```

目前，**virsh** 不提供任何可显示某个卷是否已由 Guest 使用的工具。下面的过程说明如何列出所有池中当前已被 VM Guest 使用的卷。

过程 9.2：列出 VM 主机服务器上当前使用的所有存储卷

1. 将以下内容保存到某个文件（例如 `~/libvirt/guest_storage_list.xsl`）来创建一个 XSLT 样式表：

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="text"/>
  <xsl:template match="text()"/>
  <xsl:strip-space elements="*" />
  <xsl:template match="disk">
    <xsl:text>  </xsl:text>
    <xsl:value-of select="(source/@file|source/@dev|source/@dir)[1]"/>
    <xsl:text>&#10;</xsl:text>
  </xsl:template>
</xsl:stylesheet>
```

2. 在外壳中运行以下命令：假设 Guest 的 XML 定义全部存储在默认位置 (`/etc/libvirt/qemu`)。 `xsltproc` 由软件包 `libxslt` 提供。

```
SSHEET="$HOME/libvirt/guest_storage_list.xsl"
cd /etc/libvirt/qemu
for FILE in *.xml; do
  basename $FILE .xml
  xsltproc $SSHEET $FILE
done
```

9.2.1.2 启动、停止和删除池

使用 `virsh pool` 子命令来启动、停止或删除池。在以下示例中，请将 `P00L` 替换为池的名称或其 UUID：

停止池

```
> virsh pool-destroy P00L
```



注意：池的状态不会影响挂接的卷

无论池处于什么状态（活动（已停止）或非活动（已启动）），池中挂接到 VM Guest 的卷始终可用。池的状态只会影响到能否通过远程管理将卷挂接到 VM Guest。

删除存储池

```
> virsh pool-delete P00L
```



警告：删除存储池

请参见[警告：删除存储池](#)

启动池

```
> virsh pool-start P00L
```

启用池自动启动功能

```
> virsh pool-autostart P00L
```

只有标记为 autostart 的池才会自动在 VM 主机服务器重引导时启动。

禁用池自动启动功能

```
> virsh pool-autostart P00L --disable
```

9.2.1.3 将卷添加到存储池

virsh 提供了两种将卷添加到存储池的方法：在 XML 定义中使用 `vol-create` 和 `vol-create-from` 添加，或者使用 `vol-create-as` 通过命令行参数添加。SUSE 目前不支持前一种方法，因此本节重点介绍 `vol-create-as` 子命令。

要将卷添加到现有池，请输入以下命令：

```
> virsh vol-create-as P00L ① NAME ② 12G --format ③ raw|qcow2 ④ --allocation 4G ⑤
```

- ① 卷要添加到的池的名称

- ② 卷的名称
- ③ 卷的大小，在本示例中为 12 GB。使用后缀 k、M、G、T 来分别表示千字节、兆字节、千兆字节和万亿字节。
- ④ 卷的格式。SUSE 目前支持 raw 和 qcow2。
- ⑤ 可选的参数。默认情况下，**virsh** 将创建一个按需增长的稀疏映像文件。使用此参数指定应分配的空间量（本示例中为 4 GB）。使用后缀 k、M、G、T 来分别表示千字节、兆字节、千兆字节和万亿字节。
如果不指定此参数，将生成不包含分配量的稀疏映像文件。要创建非稀疏卷，请使用此参数指定整个映像大小（在本示例中为 12G）。

9.2.1.3.1 克隆现有卷

将卷添加到池的另一种方法是克隆现有卷。请始终在原始实例所在的同一个池中创建新实例。

```
> virsh vol-clone NAME_EXISTING_VOLUME ① NAME_NEW_VOLUME ② --pool POOL ③
```

- ① 要克隆的现有卷的名称
- ② 新卷的名称
- ③ 可选参数。libvirt 会尝试自动查找现有卷。如果找不到，请指定此参数。

9.2.1.4 从存储池中删除卷

要从池中永久删除某个卷，请使用 vol-delete 子命令：

```
> virsh vol-delete NAME --pool POOL
```

--pool 可选。libvirt 会尝试自动查找卷。如果找不到，请指定此参数。



警告：删除卷时不会进行检查

无论卷当前是否已在活动或非活动的 VM Guest 中使用，都将一律被删除。无法恢复已删除的卷。

只能使用过程 9.2 “列出 VM 主机服务器上当前使用的所有存储卷”中所述的方法来检测某个卷是否已由 VM Guest 使用。

9.2.1.5 将卷挂接到 VM Guest

按照第 9.2.1.3 节 “将卷添加到存储池” 中所述创建卷后，可将其挂接到虚拟机并作为硬盘使用：

```
> virsh attach-disk DOMAIN SOURCE_IMAGE_FILE TARGET_DISK_DEVICE
```

例如：

```
> virsh attach-disk sles12sp3 /virt/images/example_disk.qcow2 sda2
```

要检查是否已挂接新磁盘，请检查 **virsh dumpxml** 命令的结果：

```
# virsh dumpxml sles12sp3
[...]
<disk type='file' device='disk'>
  <driver name='qemu' type='raw'/>
  <source file='/virt/images/example_disk.qcow2'/>
  <backingStore/>
  <target dev='sda2' bus='scsi'/>
  <alias name='scsi0-0-0'/>
  <address type='drive' controller='0' bus='0' target='0' unit='0'/>
</disk>
[...]
```

9.2.1.5.1 热插入或持久更改

可将磁盘挂接到活动和非活动的域。挂接操作由 **--live** 和 **--config** 选项控制：

--live

将磁盘热插入到活动域。挂接操作不会保存在域配置中。对非活动域使用 **--live** 会出错。

--config

持久更改域配置。下一次启动域后，挂接的磁盘将可用。

--live--config

热插入磁盘并将其添加到持久域配置中。



提示：virsh attach-device

virsh attach-device 是 **virsh attach-disk** 的更通用形式。可以使用此命令将其他类型的设备挂接到域。

9.2.1.6 从 VM Guest 分离卷

要从域中分离磁盘，请使用 **virsh detach-disk**：

```
# virsh detach-disk DOMAIN TARGET_DISK_DEVICE
```

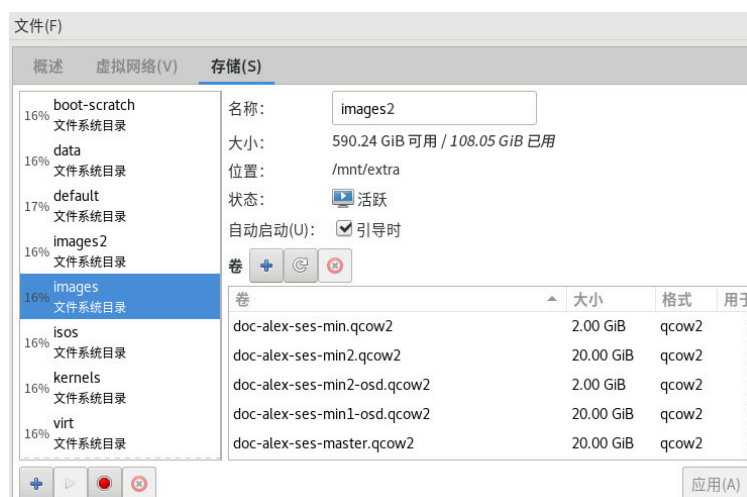
例如：

```
# virsh detach-disk sles12sp3 sda2
```

可以按照第 9.2.1.5 节 “将卷挂接到 VM Guest” 中所述，使用 `--live` 和 `--config` 选项来控制挂接操作。

9.2.2 使用虚拟机管理器管理存储设备

虚拟机管理器提供了一个图形界面，即存储管理器，用于管理存储卷和池。要访问存储管理器，请右键单击某个连接并选择细节，或者高亮显示某个连接并选择编辑 > 连接详情。选择存储选项卡。



9.2.2.1 添加存储池

要添加存储池，请执行以下操作：

1. 单击左下角的添加。添加新存储池对话框即会显示。
2. 提供池的名称（只能包含字母数字字符以及 `_`、`-` 或 `.`），然后选择类型。



3. 在下面指定所需的细节。具体的值取决于您要创建的池类型。



重要

不支持 ZFS 池。

dir 类型

- 目标路径：指定现有目录。

disk 类型

- 格式：设备分区表的格式。一般情况下，使用 `auto` 应该不会有问题。如果有问题，请通过在 VM 主机服务器上运行命令 `parted -l` 获取所需的格式。
- 源路径：设备的路径。建议使用基于 `/dev/disk/by-*` 的设备名称而不是简单的 `/dev/sdX`，因为后者可能会发生变化（例如，添加或拆除硬盘时）。您需要指定类似整个磁盘而不是磁盘上的分区（如果存在）的路径。

fs 类型

- 目标路径：VM 主机服务器文件系统上的挂载点。
- 格式：设备的文件系统格式。使用默认值 `auto` 应该不会有问题。
- 源路径：设备文件的路径。建议使用基于 `/dev/disk/by-*` 的设备名称而不是 `/dev/sdX`，因为后者可能会发生变化（例如，添加或拆除硬盘时）。

iscsi 类型

通过在 VM 主机服务器上运行以下命令来获取所需的数据：

```
> sudo iscsiadm --mode node
```

此命令将返回采用以下格式的 iSCSI 卷列表。以粗体文本显示的元素是必需的：

```
IP_ADDRESS:PORT,TPGT TARGET_NAME_(IQN)
```

- 目标路径：包含设备文件的目录。请使用 `/dev/disk/by-path`（默认值）或 `/dev/disk/by-id`。
- 主机名：iSCSI 服务器的主机名或 IP 地址。

- 源 IQN: iSCSI 目标名称 (iSCSI 限定的名称)。
- 发起端 IQN: iSCSI 发起端名称。

logical 类型

- 卷组名称: 指定现有卷组的设备路径。

mpath 类型

- 目标路径: 目前对多路径的支持仅限于使所有多路径设备可用。因此, 请在此处指定任意字符串。必须指定路径, 否则 XML 解析器将会失败。

netfs 类型

- 目标路径: VM 主机服务器文件系统上的挂载点。
- 主机名: 要导出网络文件系统的服务器的 IP 地址或主机名。
- 源路径: 正在导出的服务器上的目录。

rbd 类型

- 主机名: 导出的 RADOS 块设备所在服务器的主机名。
- 源名称: 服务器上 RADOS 块设备的名称。

scsi 类型

- 目标路径: 包含设备文件的目录。请使用 /dev/disk/by-path (默认值) 或 /dev/disk/by-id。
- 源路径: SCSI 适配器的名称。



注意: 文件浏览

从远程位置操作时, 无法通过单击浏览使用文件浏览器。

4. 单击完成以添加存储池。

9.2.2.2 管理存储池

您可以通过虚拟机管理器的存储管理器在池中创建或删除卷。您还可以暂时停用或永久删除现有的存储池。SUSE 目前不支持更改池的基本配置。

9.2.2.2.1 启动、停止和删除池

存储池的用途是提供 VM 主机服务器上的块设备，远程管理 VM Guest 时，可将这些设备添加到其中。要暂时禁止远程访问某个池，请单击存储管理器左下角的停止。停止的池会标有状态：非活动，并会在列表窗格中灰显。默认情况下，新建的池在 VM 主机服务器引导时会自动启动。要启动某个非活动的池并再次使其可从远程位置使用，请单击存储管理器左下角的启动。



注意：池的状态不会影响挂接的卷

无论池处于什么状态（活动（已停止）或非活动（已启动）），池中挂接到 VM Guest 的卷始终可用。池的状态只会影响到能否通过远程管理将卷挂接到 VM Guest。

要永久禁止访问某个池，请单击存储管理器左下角的删除。您只能删除非活动的池。删除某个池不会实际擦除其在 VM 主机服务器上的内容，而只会删除池配置。不过，在删除池时需要额外小心，尤其是删除基于 LVM 卷组的工具时：



警告：删除存储池

删除基于**本地**文件系统目录、本地分区或磁盘的存储池不会影响到这些池中当前挂接到 VM Guest 的卷的可用性。

如果删除 iSCSI、SCSI、LVM 组或网络导出的目录这些类型的池，将无法再从 VM Guest 访问这些池中的卷。尽管卷本身不会被删除，但 VM 主机服务器将不再可以访问这些资源。

创建足够大的新池或者直接从主机系统挂载/访问这些资源时，iSCSI/SCSI 目标或网络导出的目录中的卷将再次可供访问。

删除基于 LVM 组的存储池时，将擦除 LVM 组定义，并且该 LVM 组将不再存在于主机系统上。其配置将无法恢复，并且此池中的所有卷都会丢失。

9.2.2.2.2 将卷添加到存储池

借助虚拟机管理器，您可以在所有存储池（多路径、iSCSI 或 SCSI 类型的池除外）中创建卷。这些池中的卷相当于 LUN，无法从 `libvirt` 内部更改。

1. 可以使用存储管理器创建新卷，或者在将新存储设备添加到 VM Guest 时创建新卷。在以上任一情况下，请从左侧面板中选择存储池，然后单击创建新卷。
2. 指定映像的名称并选择映像格式。
SUSE 目前仅支持 `raw` 或 `qcow2` 映像。后一个选项在基于 LVM 组的池中不可用。
在最大容量的旁边，指定允许磁盘映像达到的最大大小。除非您使用 `qcow2` 映像，否则还可以设置最初应该分配的分配量。如果这两个值不同，将会创建一个按需增长的稀疏映像文件。
对于 `qcow2` 映像，可以使用构成基本映像的后备存储（也称为“后备文件”）。这样，新建的 `qcow2` 映像便只会记录对基本映像所做的更改。
3. 单击完成开始创建卷。

9.2.2.2.3 从存储池中删除卷

您只能在存储管理器中删除卷，方法是选择相应卷并单击删除卷。单击是进行确认。



警告：即使卷处于使用中状态，也能将其删除

即使活动或非活动的 VM Guest 中当前正在使用卷，也能将卷删除。无法恢复已删除的卷。

存储管理器中的使用对象列会指示某个卷是否已由 VM Guest 使用。

10 Guest 安装

VM Guest 由一个包含操作系统和数据文件的映像以及一个描述 VM Guest 虚拟硬件资源的配置文件构成。VM Guest 托管在 VM 主机服务器上并受其控制。本节提供有关安装 VM Guest 的概括说明。有关支持的 VM Guest 列表，请参见第 7 章“虚拟化限制和支持”。

与运行操作系统需要满足的要求相比，虚拟机几乎没有什么要求。如果操作系统未根据虚拟机主机环境进行优化，将只能以全虚拟化模式在硬件辅助虚拟化计算机硬件上运行，并需要加载特定的设备驱动程序。提供给 VM Guest 的硬件取决于主机的配置。

您应该了解与在多个虚拟机上运行单个已许可操作系统副本相关的任何许可问题。有关详细信息，请查阅操作系统许可协议。

10.1 基于 GUI 的 Guest 安装



提示：更改新虚拟机的默认选项

可以更改在创建新虚拟机时要应用的默认值。例如，要将 UEFI 设置为新虚拟机的默认固件类型，请从虚拟机管理器的主菜单中选择编辑 > 首选项，单击新建虚拟机，然后将 UEFI 设置为默认固件。



图 10.1：指定新 VM 的默认选项

新建虚拟机向导将帮助您完成创建虚拟机和安装其操作系统需执行的步骤。要启动该向导，请打开虚拟机管理器并选择文件 > 新建虚拟机。或者，启动 YaST 并选择虚拟化 > 创建虚拟机。

1. 从 YaST 或虚拟机管理器中启动新建虚拟机向导。
2. 选择安装源 — 本地可用的媒体或网络安装源。要从现有映像安装 VM Guest，请选择导入现有磁盘映像。

在运行 Xen 超级管理程序的 VM 主机服务器上，您可以选择是要安装半虚拟化 Guest 还是全虚拟化 Guest。您可以在体系结构选项下选择相应的选项。根据此项选择，并非所有安装选项均可用。

3. 根据在上一步中所做的选择，您需要提供以下数据：

本地安装媒体（ISO 映像或 CDROM）

在 VM 主机服务器上指定包含安装数据的 ISO 映像的路径。如果该映像是作为 libvirt 存储池中的卷提供的，您也可以使用浏览来选择。有关详细信息，请访问 [第 13 章 “高级存储主题”](#)。

或者，选择已插入到 VM 主机服务器光驱中的物理 CD-ROM 或 DVD。

网络安装（HTTP、HTTPS 或 FTP）

提供指向安装源的 URL。有效的 URL 前缀包括 [ftp://](#)、[http://](#) 和 [https://](#) 等。

在 URL 选项下，提供自动安装文件（例如 AutoYaST 或 Kickstart）的路径以及内核参数。提供 URL 后，应该就会自动正确检测到操作系统。如果情况并非如此，请取消选择基于安装媒体自动检测操作系统，并手动选择操作系统类型和版本。

导入现有磁盘映像

要从现有映像安装 VM Guest，您需要在 VM 主机服务器上指定该映像的路径。如果该映像是作为 libvirt 存储池中的卷提供的，您也可以使用浏览来选择。有关详细信息，请访问 [第 13 章 “高级存储主题”](#)。

手动安装

如果您要创建虚拟机，手动配置其组件并在稍后安装其操作系统，则适合使用这种安装方法。要将 VM 调整为特定的产品版本，请开始键入版本名称（例如 [sles](#)），然后在出现匹配项时选择所需的版本。

4. 选择新虚拟机的内存大小和 CPU 数量。
5. 如果在第一步中选择了导入现有映像，则会省略此步骤。

设置 VM Guest 的虚拟硬盘。创建新磁盘映像，或者从存储池中选择一个现有的磁盘映像（有关详细信息，请参见第 13 章“高级存储主题”）。如果您选择创建磁盘，将会创建一个 qcow2 映像，该映像默认存储在 `/var/lib/libvirt/images` 下。

设置磁盘是可选操作。例如，如果您直接从 CD 或 DVD 运行实时系统，可以通过停用为此虚拟机启用存储来省略此步骤。

6. 在向导的最后一个屏幕上指定虚拟机的名称。如果您希望能够查看和更改虚拟化硬件选择，请选中在安装之前自定义配置。在网络选择下指定网络设备。使用网桥设备时，系统会预先填充主机上的第一个网桥。要使用其他网桥，请在文本框中手动更新为该网桥名称。

单击完成。

7. （可选）如果您在上一步中保留了默认设置，则会开始安装。如果您选择了在安装之前自定义配置，会打开 VM Guest 配置对话框。有关配置 VM Guest 的详细信息，请参见第 14 章“使用虚拟机管理器配置虚拟机”。

完成配置后，单击开始安装。



提示：将组合键传递给虚拟机

安装将在一个虚拟机管理器控制台窗口中开始。某些组合键（例如 `Ctrl - Alt - F1`）会被 VM 主机服务器识别，但不会传递给虚拟机。虚拟机管理器提供“粘滞键”功能来绕过 VM 主机服务器。按 `Ctrl`、`Alt` 或 `Shift` 三次使该键成为粘滞键，然后按组合键中剩余的键便可将组合键传递给虚拟机。

例如，要将 `Ctrl - Alt - F2` 传递给 Linux 虚拟机，请按 `Ctrl` 三次，然后按 `Alt - F2`。也可以按 `Alt` 三次，然后按 `Ctrl - F2`。

在安装 VM Guest 期间以及安装之后，都可以在虚拟机管理器中使用粘滞键功能。

10.1.1 为虚拟机配置 PXE 引导

PXE 引导使虚拟机能够通过网络从安装媒体引导，而无需从物理媒体或安装磁盘映像进行引导。有关设置 PXE 引导环境的更多细节，请参见《部署指南》，第 18 章“准备网络引导环境”。

要使您的 VM 从 PXE 服务器引导，请执行以下步骤：

1. 按照第 10.1 节 “基于 GUI 的 Guest 安装” 中所述启动安装向导。
2. 选择手动安装方法。
3. 按照向导操作到最后一步，然后选中在安装之前自定义配置。单击完成确认。
4. 在自定义屏幕上，选择引导选项。
5. 检查引导设备顺序，然后选择启用引导菜单。
 - 要保留默认引导选项 VirtIO 磁盘，请单击应用进行确认。
 - 要强制虚拟机使用 PXE 作为默认引导选项，请执行以下操作：
 - a. 在引导菜单配置中选择 NIC 设备。
 - b. 使用右侧的箭头标志将其移动到顶部。
 - c. 单击应用进行确认。
6. 单击开始安装以开始安装。现在按 **Esc** 进入引导菜单，然后选择 1. iPXE。如果正确配置了 PXE 服务器，PXE 菜单屏幕将会显示。

10.2 使用 **virt-install** 从命令行安装

virt-install 是个命令行工具，可帮助您使用 libvirt 库创建新虚拟机。如果您无法使用图形用户界面，或需要自动化虚拟机创建过程，此工具十分有用。

virt-install 是个复杂的脚本，其中包含大量命令行开关。下面是必需的开关。有关详细信息，请参见 **virt-install** (1) 的手册页。

一般选项

- `--name VM_GUEST_NAME`: 指定新虚拟机的名称。该名称必须在同一连接上超级管理程序已知的所有 Guest 中保持唯一。该名称用于创建和命名 Guest 的配置文件，您可以通过 `virsh` 使用该名称来访问 Guest。该名称可以包含字母数字和 `_-.:+` 字符。
- `--memory REQUIRED_MEMORY`: 以 MB 为单位指定分配给新虚拟机的内存量。
- `--vcpus NUMBER_OF_CPUS`: 指定虚拟 CPU 数量。要获得最佳性能，虚拟处理器数量应小于或等于物理处理器数量。

虚拟化类型

- `--paravirt`: 安装半虚拟化 Guest。如果 VM 主机服务器支持半虚拟化和全虚拟化，这就是默认设置。
- `--hvm`: 安装全虚拟化 Guest。
- `--virt-type HYPERVISOR`: 指定超级管理程序。支持的值为 `kvm` 或 `xen`。

Guest 存储空间

指定 `--disk`、`--filesystem` 或 `--nodisks` 作为新虚拟机的存储类型。例如，`--disk size=10` 会在超级管理程序的默认映像位置创建 10 GB 磁盘，并将此磁盘用于 VM Guest。`--filesystem /export/path/on/vmhost` 指定 VM 主机服务器上要导出到 Guest 的目录。`--nodisks` 会安装没有本地存储空间的 VM Guest（适合使用实时 CD 的情形）。

安装方法

使用 `--location`、`--cdrom`、`--pxe`、`--import` 或 `--boot` 指定安装方法。

访问安装

使用 `--graphics VALUE` 选项指定如何访问安装。SUSE Linux Enterprise Server 支持值 `vnc` 或 `none`。

如果使用 VNC，`virt-install` 将尝试启动 `virt-viewer`。如果 `virt-viewer` 未安装或无法运行，请使用您偏好的查看器手动连接到 VM Guest。要明确阻止 `virt-install` 启动查看器，请使用 `--noautoconsole`。要定义用于访问 VNC 会话的口令，请使用以下语法：`--graphics vnc,password=PASSWORD`。

如果您使用 `--graphics none`，可以通过操作系统支持的服务（例如 SSH 或 VNC）访问 VM Guest。请参见操作系统安装手册了解如何在安装系统中设置这些服务。

传递内核和 initrd 文件

可以直接指定安装程序的内核和 Initrd，例如，指定来自网络来源的内核和 Initrd。要设置网络来源，请参见《部署指南》，第 17 章“设置网络安装源”，第 17.4 节“手动设置 HTTP 储存库”。

要传递其他引导参数，请使用 `--extra-args` 选项。此选项可用于指定网络配置。有关详细信息，请参见《部署指南》，第 8 章“引导参数”。

例 10.1：从 HTTP 服务器加载内核和 INITRD

```
# virt-install --location "http://example.tld/REPOSITORY/DVD1/" \
--extra-args="textmode=1" --name "SLES15" --memory 2048 --virt-type kvm \
--connect qemu:///system --disk size=10 --graphics vnc \
--network network=vnet_nated
```

启用控制台

默认不会对使用 `virt-install` 安装的新虚拟机启用控制台。要启用控制台，请如以下示例所示使用 `--extra-args="console=ttyS0 textmode=1"`：

```
> virt-install --virt-type kvm --name sles12 --memory 1024 \
--disk /var/lib/libvirt/images/disk1.qcow2 --os-variant sles12
--extra-args="console=ttyS0 textmode=1" --graphics none
```

安装完成后，VM 映像中的 `/etc/default/grub` 文件将会更新，在 `GRUB_CMDLINE_LINUX_DEFAULT` 行中包含 `console=ttyS0` 选项。

使用 UEFI 安全引导



注意

SUSE 仅支持在 AMD64/Intel 64 KVM Guest 上使用 UEFI 安全引导。Xen HVM Guest 支持使用 UEFI 固件引导，但不支持 UEFI 安全引导。

默认情况下，使用 **virt-install** 安装的新虚拟机会配置传统 BIOS。您可以通过 `--boot firmware=efi` 将这些虚拟机配置为使用 UEFI。系统会选择支持 UEFI 安全引导并已注册 Microsoft 密钥的固件。如果不需要安全引导，可以使用 `--boot firmware=efi,firmware.feature0.name=secure-boot,firmware.feature0.enabled=no` 选项来选择不支持安全引导的 UEFI 固件。也可以明确指定 UEFI 固件映像。有关为虚拟机使用 UEFI 的高级信息和示例，请参见第 10.3.1 节“高级 UEFI 配置”。

例 10.2：virt-install 命令行示例

以下命令行示例将创建带有 virtio 加速磁盘和网卡的新 SUSE Linux Enterprise 15 SP2 虚拟机。它将创建新的 10 GB qcow2 磁盘映像作为存储空间，源安装媒体为主机 CD-ROM 驱动器。此命令行使用 VNC 图形，并会自动启动图形客户端。

KVM

```
> virt-install --connect qemu:///system --virt-type kvm \
--name sle15sp2 --memory 1024 --disk size=10 --cdrom /dev/cdrom --
graphics vnc \
--os-variant sle15sp2
```

Xen

```
> virt-install --connect xen:// --virt-type xen --hvm \
--name sle15sp2 --memory 1024 --disk size=10 --cdrom /dev/cdrom --
graphics vnc \
--os-variant sle15sp2
```

10.3 高级 Guest 安装方案

本节提供有关超出了正常安装范围的操作（例如手动配置 UEFI 固件、使用内存气球和安装附加产品）的说明。

10.3.1 高级 UEFI 配置

虚拟机使用的 UEFI 固件由 **OVMF**（**开放虚拟机固件**）提供。`qemu-ovmf-x86_64` 软件包提供适用于 AMD64/Intel 64 VM Guest 的固件。AArch64 VM Guest 的固件由 `qemu-uefi-aarch64` 软件包提供。两个软件包都包含多个固件版本，每个固件版本都支持一组不同的特性和功能。这些软件包中还包含 JSON 固件描述符文件，用于描述各个固件版本的特性和功能。`libvirt` 支持两种选择虚拟机 UEFI 固件的方法：自动和手动。如果使用自动选择方法，`libvirt` 将根据用户指定的一组可选功能选择固件。如果未明确指定功能，`libvirt` 将选择已启用安全引导并已注册 Microsoft 密钥的固件。使用手动选择方法时，必须明确指定固件的完整路径和任何可选设置。用户可以引用 JSON 描述符文件来查找满足其要求的固件。



提示

目录 `/usr/share/qemu/firmware` 包含 `libvirt` 使用的所有 JSON 文件。此文件提供有关固件的详细信息，包括功能的功能和特性。

使用 `virt-install` 时，可以通过为 **boot** 选项指定 **firmware=efi** 参数（例如，`--boot firmware=efi`）来启用自动固件选择。可以通过请求添加或删除固件功能来影响选择过程。以下示例演示了如何在禁用 UEFI 安全引导的情况下自动选择固件。

```
> virt-install --connect qemu:///system --virt-type kvm \
--name sle15sp5 --memory 1024 --disk size=10 --cdrom /dev/cdrom --graphics vnc \
--boot firmware=efi,firmware.feature0.name=secure-
boot,firmware.feature0.enabled=no \
--os-variant sle15sp5
```



注意

为确保永久性 VM Guest 在其生命周期内始终使用相同的固件和变量存储区，`libvirt` 将在 VM Guest XML 配置中记录自动选择的固件。自动选择固件是一次性活动。选择固件后，仅当 VM Guest 管理员使用手动选择固件的方法明确更改固件时，固件才会更改。

loader 和 **nvr** 参数用于手动选择固件。**loader** 是必选参数，**nvr** 定义可选的 UEFI 变量存储。以下示例演示了如何在启用安全引导的情况下手动选择固件。


```
> virt-install --connect qemu:///system --virt-type kvm \
--name sle15sp5 --memory 1024 --disk size=10 --cdrom /dev/cdrom --graphics vnc \
--boot loader=/usr/share/qemu/ovmf-x86_64-smm-
code.bin,loader.readonly=yes,loader.type=pflash,loader.secure=yes,nvram.template=/
usr/share/qemu/ovmf-x86_64-smm-vars.bin \
--os-variant sle15sp5
```



注意

libvirt 无法修改 UEFI 固件的任何特征。例如，它无法在启用了 UEFI 安全引导的固件中禁用 UEFI 安全引导，即使指定 **loader.secure=no** 也是如此。libvirt 将确保指定的固件可以满足任何指定的功能。例如，它将拒绝使用 **loader.secure=no** 禁用安全引导的配置，而是指定启用了 UEFI 安全引导的固件。

qemu-ovmf-x86_64 软件包中包含多个 UEFI 固件映像。例如，以下子集支持 SMM、UEFI 安全引导，并已注册 Microsoft、openSUSE 或 SUSE UEFI CA 密钥：

```
# rpm -ql qemu-ovmf-x86_64
[...]
/usr/share/qemu/ovmf-x86_64-smm-ms-code.bin
/usr/share/qemu/ovmf-x86_64-smm-ms-vars.bin
/usr/share/qemu/ovmf-x86_64-smm-opensuse-code.bin
/usr/share/qemu/ovmf-x86_64-smm-opensuse-vars.bin
/usr/share/qemu/ovmf-x86_64-smm-suse-code.bin
/usr/share/qemu/ovmf-x86_64-smm-suse-vars.bin
[...]
```

对于 AArch64 体系结构，该软件包名为 qemu-uefi-aarch32：

```
# rpm -ql qemu-uefi-aarch32
[...]
/usr/share/qemu/aavmf-aarch32-code.bin
/usr/share/qemu/aavmf-aarch32-vars.bin
/usr/share/qemu/firmware
/usr/share/qemu/firmware/60-aavmf-aarch32.json
/usr/share/qemu/qemu-uefi-aarch32.bin
```

`*-code.bin` 文件是 UEFI 固件文件。`*-vars.bin` 文件是对应的变量存储映像，可用作每个 VM 的非易失性存储模板。首次创建 VM 时，`libvirt` 会将指定的 `vars` 模板复制到 `/var/lib/libvirt/qemu/nvram/` 下每个 VM 的专属路径中。名称中不包含 `code` 或 `vars` 的文件可用作单个 UEFI 映像。它们没有太大的作用，因为每次经过 VM 关开机后，UEFI 变量都不会保存。

`*-ms*.bin` 文件包含存放在实际硬件上的 UEFI CA 密钥。因此，在 `libvirt` 中它们已配置为默认设置。同样，`*-suse*.bin` 文件包含预安装的 SUSE 密钥。还有一组不包含预安装密钥的文件。

有关 OVMF 的更多细节，请参见 <http://www.linux-kvm.org/downloads/lersek/ovmf-whitepaper-c770f8c.txt>。

10.3.2 对 Windows Guest 使用内存气球

内存气球是在运行时更改 VM Guest 所用内存量的方法。KVM 和 Xen 超级管理程序都提供此方法，但需要 Guest 也支持此方法。

基于 openSUSE 和 SLE 的 Guest 支持内存气球，而 Windows Guest 需要通过[虚拟机驱动程序软件包 \(VMDP\)](https://www.suse.com/products/vmdriverpack/) (<https://www.suse.com/products/vmdriverpack/>) 来提供气球技术。要使设置的最大内存大于为 Windows Guest 配置的初始内存，请执行以下步骤：

1. 安装最大内存等于或小于初始值的 Windows Guest。
2. 在 Windows Guest 中安装虚拟机驱动程序包，以提供所需的驱动程序。
3. 关闭 Windows Guest。
4. 将 Windows Guest 的最大内存重新设置为所需值。
5. 再次启动 Windows Guest。

10.3.3 在安装中包含附加产品

某些操作系统（例如 SUSE Linux Enterprise Server）允许在安装过程中包含附加产品。如果附加产品安装源是通过 SUSE Customer Center 提供的，则无需进行特殊的 VM Guest 配置。如果安装源是通过 CD/DVD 或 ISO 映像提供的，则需要向 VM Guest 安装系统提供标准安装媒体映像和附加产品的映像。

如果您使用的是基于 GUI 的安装方法，请在向导的最后一步选择在安装之前自定义配置，并通过添加硬件 > 存储添加附加产品 ISO 映像。指定映像的路径，并将设备类型设置为 CD-ROM。

如果您是从命令行安装的，则需要使用 `--disk` 参数而不是 `--cdrom` 来设置虚拟 CD/DVD 驱动器。将使用第一个指定的设备进行引导。以下示例将 SUSE Linux Enterprise Server 15 连同 SUSE Enterprise Storage 扩展一起安装：

```
> virt-install \
  --name sles15+storage \
  --memory 2048 --disk size=10 \
  --disk /path/to/SLE-15-SP7-Full-ARCH-GM-media1.iso-x86_64-GM-
DVD1.iso,device=cdrom \
  --disk /path/to/SUSE-Enterprise-Storage-VERSION-DVD-ARCH-
Media1.iso,device=cdrom \
  --graphics vnc --os-variant sle15
```

11 基本 VM Guest 管理

使用虚拟机管理器图形应用程序或者在命令行上使用 **virsh** 可以完成大部分管理任务，例如启动或停止 VM Guest。而要通过 VNC 连接到图形控制台，就只能从图形用户界面进行。



注意：管理远程 VM 主机服务器上的 VM Guest

如果 VM 主机服务器上启动了虚拟机管理器、**virsh** 和 **virt-viewer** 这些 **libvirt** 工具，则可以使用它们来管理主机上的 VM Guest。不过，您也可以管理远程 VM 主机服务器上的 VM Guest。这需要在主机上为 **libvirt** 配置远程访问权限。有关说明，请参见第 12 章“连接和授权”。

要使用虚拟机管理器连接到此类远程主机，需要按照第 12.2.2 节“使用虚拟机管理器管理连接”中所述设置连接。如果您通过 **virsh** 或 **virt-viewer** 连接到远程主机，需要使用参数 **-c** 指定连接 URI（例如，**virsh -c qemu+tls://saturn.example.com/system** 或 **virsh -c xen+ssh://**）。连接 URI 的格式取决于连接类型和超级管理程序 — 有关细节，请参见第 12.2 节“连接到 VM 主机服务器”。

本章列出的所有示例都不包含连接 URI。

11.1 列出 VM Guest

VM Guest 列表显示 VM 主机服务器上由 **libvirt** 管理的所有 VM Guest。

11.1.1 使用虚拟机管理器列出 VM Guest

虚拟机管理器的主窗口会列出它所连接的每台 VM 主机服务器的所有 VM Guest。每个 VM Guest 项都包含计算机的名称及其状态（正在运行、已暂停或已关闭），这些信息以图标、文本和 CPU 使用率条的形式显示。

11.1.2 使用 **virsh** 列出 VM Guest

使用 **virsh list** 命令可获取 VM Guest 的列表：

列出所有正在运行的 Guest

```
> virsh list
```

列出所有正在运行的 Guest 以及非活动的 Guest

```
> virsh list --all
```

有关详细信息和其他选项，请参见 [virsh help list](#) 或 [man 1 virsh](#)。

11.2 通过控制台访问 VM Guest

可以通过 VNC 连接（图形控制台）或串行控制台（如果受 Guest 操作系统的支持）访问 VM Guest。

11.2.1 打开图形控制台

打开与 VM Guest 连接的图形控制台可与该计算机交互，就如同通过 VNC 连接与物理主机交互一样。如果访问 VNC 服务器需要身份验证，系统会提示您输入用户名（如果适用）和口令。

当您单击进入 VNC 控制台时，光标将被“捕获”，不能再在控制台外部使用。要释放光标，请按 **Alt + Ctrl**。



提示：无缝（绝对）光标移动

为防止控制台夺取光标，同时为了启用无缝光标移动，请向 VM Guest 添加绘图板输入设备。有关更多信息，请参见第 14.5 节“输入设备”。

某些组合键（例如 **Ctrl + Alt + Del**）由主机解释，不会传递给 VM Guest。要将此类组合键传递给 VM Guest，请在 VNC 窗口中打开发送键菜单，然后选择所需的组合键项。仅当使用虚拟机管理器和 **virt-viewer** 时，才能使用发送键菜单。借助虚拟机管理器，可以按照[提示：将组合键传递给虚拟机](#)中所述改用“粘滞键”功能。



注意：支持的 VNC 查看器

理论上而言，所有 VNC 查看器都可连接到 VM Guest 的控制台。但如果您是使用 SASL 身份验证和/或 TLS/SSL 连接来访问 Guest 的，那么您的选择就比较有限。**tightvnc** 或 **tigervnc** 等常见 VNC 查看器既不支持 SASL 身份验证，也不支持 TLS/SSL。唯一可替代虚拟机管理器和 **virt-viewer** 的工具是 Remmina（请参见《管理指南》，第 14 章“使用 VNC 的远程图形会话”，第 14.2 节“Remmina：远程桌面客户端”）。

11.2.1.1 使用虚拟机管理器打开图形控制台

1. 在虚拟机管理器中，右键单击某个 VM Guest 项。
2. 从弹出菜单中选择打开。

11.2.1.2 使用 **virt-viewer** 打开图形控制台

virt-viewer 是一个简单的 VNC 查看器，其中添加了用于显示 VM Guest 控制台的功能。例如，可以“wait”模式启动该查看器，在此情况下，它会先等待 VM Guest 启动，然后再建立连接。它还支持自动重新连接到重引导的 VM Guest。

virt-viewer 按名称、ID 或 UUID 对 VM Guest 进行寻址。使用 **virsh list --all** 可获取这些数据。

要连接到正在运行或已暂停的 Guest，请使用 ID、UUID 或名称。已关闭的 VM Guest 没有 ID — 您只能按 UUID 或名称与其建立连接。

连接到 ID 为 8 的 Guest

```
> virt-viewer 8
```

连接到名为 **sles12** 的非活动 Guest；Guest 启动后，连接窗口就会打开

```
> virt-viewer --wait sles12
```

如果使用 **--wait** 选项，即使 VM Guest 此刻未运行，也会保持连接。当 Guest 启动时，查看器即会启动。

有关更多信息，请参见 `virt-viewer --help` 或 `man 1 virt-viewer`。



注意：通过 SSH 建立远程连接时输入的口令

使用 `virt-viewer` 通过 SSH 来与远程主机建立连接时，需要输入 SSH 口令两次。第一次用于向 `libvirt` 进行身份验证，第二次用于向 VNC 服务器进行身份验证。第二个口令需要在启动 `virt-viewer` 的命令行上提供。

11.2.2 打开串行控制台

要访问虚拟机的图形控制台，需要在访问 VM Guest 的客户端上提供一个图形环境。或者，也可通过串行控制台和 `virsh` 在外壳中访问使用 `libvirt` 管理的虚拟机。要打开与名为 “sles12” 的 VM Guest 连接的串行控制台，请运行以下命令：

```
> virsh console sles12
```

`virsh console` 接受两个可选标志：`--safe` 确保以独占方式访问控制台，`--force` 在连接之前断开与所有现有会话的连接。这两个功能需受 Guest 操作系统的支持。

Guest 操作系统必须支持串行控制台访问，并且该操作系统也受到适当的支持，才能通过串行控制台连接到 VM Guest。有关详细信息，请参见 Guest 操作系统手册。



提示：为 SUSE Linux Enterprise Guest 和 openSUSE Guest 启用串行控制台访问

SUSE Linux Enterprise 和 openSUSE 中默认会禁用串行控制台访问。要启用它，请执行下列步骤：

SLES 12、15 和 openSUSE

启动 YaST 引导加载程序模块并切换到内核参数选项卡。将 `console=ttyS0` 添加到可选内核命令行参数字段。

SLES 11

启动 YaST 引导加载程序模块，并选择要为其激活串行控制台访问的引导项。选择编辑，将 `console=ttyS0` 添加到可选内核命令行参数字段。此外，请编辑 `/etc/inittab` 并取消注释包含以下内容的行：

```
#S0:12345:respawn:/sbin/agetty -L 9600 ttyS0 vt102
```

11.3 更改 VM Guest 的状态：启动、停止、暂停

可以使用虚拟机管理器或 **virsh** 来启动、停止或暂停 VM Guest。您还可以将 VM Guest 配置为在引导 VM 主机服务器时自动启动。

关闭 VM Guest 时，可将其正常关机或强制关机。后一种操作的效果等同于拔下物理主机上的电源插头，建议仅在没有其他办法时才这样做。强制关机可能导致 VM Guest 上的文件系统损坏或数据丢失。



提示：正常关机

要能够执行正常关机，必须将 VM Guest 配置为支持 **ACPI**。如果 Guest 是使用虚拟机管理器创建的，则可在 VM Guest 中使用 ACPI。

根据 Guest 操作系统，能够使用 ACPI 可能还不足以执行正常关机。在生产环境中使用 Guest 之前，强烈建议先对其进行关机和重引导测试。例如，openSUSE 或 SUSE Linux Enterprise Desktop 可能需要获得 Polkit 授权才能关机和重引导。确保已在所有 VM Guest 上关闭此策略。

如果在安装 Windows XP/Windows Server 2003 Guest 期间启用了 ACPI，只在 VM Guest 配置中开启 ACPI 并不足够。有关更多信息，请参见：

- <https://support.microsoft.com/en-us/kb/314088> 
- <https://support.microsoft.com/en-us/kb/309283> 

无论 VM Guest 的配置如何，始终都可以从 Guest 操作系统内部实现正常关机。

11.3.1 使用虚拟机管理器更改 VM Guest 的状态

可以通过虚拟机管理器的主窗口或 VNC 窗口更改 VM Guest 的状态。

过程 11.1：通过虚拟机管理器窗口更改状态

1. 右键单击某个 VM Guest 项。
2. 在弹出菜单中选择运行、暂停或其中一个关机选项。

过程 11.2：通过 VNC 窗口更改状态

1. 按照第 11.2.1.1 节“使用虚拟机管理器打开图形控制台”中所述打开 VNC 窗口。
2. 在工具栏或虚拟机菜单中，选择运行、暂停或其中一个关机选项。

11.3.1.1 自动启动 VM Guest

您可以在引导 VM 主机服务器时自动启动 Guest。此功能默认未启用，需要为每个 VM Guest 单独启用。无法全局激活此功能。

1. 在虚拟机管理器中双击 VM Guest 项以打开其控制台。
2. 选择视图 > 细节打开 VM Guest 配置窗口。
3. 选择引导选项，然后选中在主机引导时启动虚拟机。
4. 单击应用保存新配置。

11.3.2 使用 **virsh** 更改 VM Guest 的状态

以下示例会更改名为“sles12”的 VM Guest 的状态。

开始

```
> virsh start sles12
```

暂停

```
> virsh suspend sles12
```

恢复（已暂停的 VM Guest）

```
> virsh resume sles12
```

重引导

```
> virsh reboot sles12
```

正常关机

```
> virsh shutdown sles12
```

强制关机

```
> virsh destroy sles12
```

开启自动启动

```
> virsh autostart sles12
```

关闭自动启动

```
> virsh autostart --disable sles12
```

11.4 保存和恢复 VM Guest 的状态

保存 VM Guest 会保留其内存的确切状态。该操作类似于将计算机**休眠**。保存的 VM Guest 可以快速恢复到保存前的相同运行状况。

保存时，VM Guest 会暂停，其当前内存状态将保存到文件，然后该 Guest 停止。该操作不会复制 VM Guest 虚拟磁盘的任何一部分。保存虚拟机所需的时间取决于分配的内存量。成功保存后，VM Guest 的资源将释放回虚拟机主机服务器。

恢复操作将加载先前保存的 VM Guest 内存状态文件并启动该 Guest。该 Guest 不会引导，而是在以前保存它的位置继续运行。该操作类似于退出休眠状态。

libvirt 支持多种保存文件格式。默认格式称为 raw，由 VM Guest 内存页面的顺序流组成。raw 格式的顺序布局不太适合多读取器和写入器场景。

除 raw 保存文件格式外，libvirt 还支持多种压缩格式：zstd、lzop、gzip、bzip2 和 xz。与 raw 格式类似，压缩格式由 VM Guest 内存页的顺序流构成，但在写入保存文件或从中读取时，会通过指定的压缩算法进行压缩处理。这些格式可节省保存文件的存储空间，但会增加保存/恢复时间及主机 CPU 占用。

sparse 保存文件格式采用预计算的固定偏移量来读写 VM Guest 内存页。生成的保存文件逻辑大小约等于 VM Guest 的内存容量，实际磁盘占用空间则取决于 VM Guest 的实时内存使用情况。sparse 格式为 VM Guest 内存页预置固定偏移量，可完美支持多读写器并发操作，这对于大内存容量的 VM Guest 而言，能显著提升其保存与恢复操作的执行效率。

默认的保存文件格式可通过修改 `/etc/libvirt/qemu.conf` 中的 `save_image_format` 进行更改。此外，在使用 **virsh** 执行保存操作时也可指定格式。有关使用 **virsh** 进行保存和恢复的详细信息，请参见第 11.4.2 节“使用 **virsh** 保存和恢复”。

由于 VM Guest 的运行状态将保存至文件，请确保存储设备具备足够的可用空间。如果采用 sparse 保存文件格式，保存文件的逻辑大小约等于 VM Guest 分配的内存容量。但实际磁盘占用空间通常更小，具体取决于 VM Guest 的内存使用情况。VM Guest 中未使用的内存空间不会写入保存文件，因此我们称它为 sparse。

对于 raw 保存文件格式，其逻辑文件大小与磁盘实际占用空间相同，两者均取决于 VM Guest 的内存使用情况。无论采用 raw 还是 sparse 格式，均可通过在 VM Guest 上执行以下命令估算保存文件的磁盘占用空间（以 MB 为单位）：

```
> free -mh | awk '/^Mem:/ {print $3}'
```

采用压缩格式可减小磁盘占用空间，具体取决于指定的压缩算法的效率。



警告：始终恢复保存的 Guest

成功执行保存操作后，如果通过恢复操作以外的方式启动 VM Guest，将导致已保存的状态文件失效。保存文件可能包含未完全写入磁盘的文件系统数据。如果在 VM Guest 通过其他方式执行后再尝试恢复保存的状态，可能会导致文件系统损坏。

请务必使用相同的应用程序进行 VM Guest 的保存与恢复操作。例如，如果使用 **virsh** 保存 VM Guest，则不要使用虚拟机管理器执行恢复。在这种情况下，请务必使用 **virsh** 进行恢复。

！ 重要：恢复 VM Guest 后同步其时间

如果您在保存 VM Guest 后，经过长时间（数小时）的暂停再恢复该 Guest，其时间同步服务（例如 `chronyd`）可能会拒绝同步其时间。在这种情况下，请手动同步 VM Guest 的时间。例如，对于 KVM 主机，可以使用 QEMU Guest 代理，并使用 **`guest-set-time`** 来指示 Guest 设置时间。有关更多详细信息，请参见第 22 章 “QEMU Guest 代理”。

11.4.1 使用虚拟机管理器保存/恢复

过程 11.3：保存 VM GUEST

1. 打开 VM Guest 的 VNC 连接窗口。确保该 Guest 正在运行。
2. 选择虚拟机 > 关机 > 保存。

过程 11.4：恢复 VM GUEST

1. 打开 VM Guest 的 VNC 连接窗口。确保该 Guest 未运行。
2. 选择虚拟机 > 恢复。

如果 VM Guest 之前是使用虚拟机管理器保存的，则系统不会为您提供用于运行该 Guest 的选项。但请注意警告：始终恢复保存的 Guest 中所述的有关使用 **`virsh`** 保存的计算机的注意事项。

11.4.2 使用 **`virsh`** 保存和恢复

与虚拟机管理器相比，**`libvirt`** 为保存与恢复操作提供了更精细的控制能力。**`virsh save`** 和 **`virsh restore`** 支持多个选项，可用于调整操作行为。基本的方式是通过指定 VM Guest 的名称、ID 或 UUID 以及文件名来保存 VM Guest。例如：

```
> virsh save openSUSE-Leap /virtual/saves/openSUSE-Leap.vmsav
```

执行基本的 VM Guest 恢复操作时，只需指定保存文件名即可。例如：

```
> virsh restore /virtual/saves/openSUSE-Leap.vmsav
```

当 VM Guest 内存容量增大时，为达到理想的传输速率，保存与恢复操作可能需要使用额外选项，尤其是在保存镜像文件存储于高吞吐量存储设备时。在此类场景中，VM 主机服务器的文件系统缓存往往适得其反，应当使用 `bypass-cache` 选项来避免。例如：

```
> virsh save --bypass-cache openSUSE-Leap /virtual/saves/openSUSE-Leap.vmsav
```

```
> virsh restore --bypass-cache /virtual/saves/openSUSE-Leap.vmsav
```

通过多通道读写 VM Guest 内存页，可显著提升其在高吞吐量存储设备上的保存与恢复效率。如第 11.4 节“保存和恢复 VM Guest 的状态”中所述，使用多通道必须采用 `sparse` 映像格式。选择通道数量时需谨慎，要确保操作不会对 VM 主机服务器上的其他工作负载产生负面影响。当 VM 主机服务器资源采用静态分区时，一般建议通道数设置为 VM Guest 专属物理 CPU 的核心数。由于保存操作启动时将停止 VM Guest 的 vCPU，因此这些 CPU 资源可安全地用于保存内存页。

以下示例使用 4 个通道（同时绕过 VM 主机服务器文件系统缓存）执行保存与恢复操作：

```
> virsh save --bypass-cache --image-format sparse --parallel-channels 4  
openSUSE-Leap /virtual/saves/openSUSE-Leap.vmsav
```

```
> virsh restore --bypass-cache --parallel-channels 4 /virtual/saves/openSUSE-  
Leap.vmsav
```

映像格式已编码存储于保存映像文件中，执行恢复操作时无需指定。

有关保存/恢复操作及支持选项的更多信息，请参见 `virsh help save`、`virsh help restore` 或 `man 1 virsh`。

11.5 创建和管理快照

VM Guest 快照是整个虚拟机的快照，包括 CPU、RAM、设备的状态，以及所有可写磁盘的内容。要使用虚拟机快照，所有挂接的硬盘均需使用 `qcow2` 磁盘映像格式，并且其中至少有一个硬盘需是可写的。

快照可让您将计算机恢复到特定时间点的状态。在撤消有错误的配置或者安装大量软件包时，此功能十分有用。启动一个在 VM Guest 处于关闭状态下创建的快照后，需要引导该 Guest。在该时间点之后写入磁盘的所有更改都将在启动快照时丢失。



注意

快照仅在 KVM VM 主机服务器上受支持。

11.5.1 术语

有多个特定术语用于描述快照的类型：

内部快照

保存到原始 VM Guest 的 qcow2 文件中的快照。该文件包含保存的快照状态，以及自截取快照以来发生的更改。内部快照的主要优势是它们全都存储在一个文件中，因此方便在多个计算机之间复制或移动。

外部快照

创建外部快照时，会保存原始 qcow2 文件并将其设为只读，同时会创建一个新的 qcow2 文件用于存放更改。原始文件有时称为**后备文件**或**基础文件**，包含所有更改的新文件称为**覆盖文件**或**派生文件**。备份 VM Guest 时，外部快照很有用。但外部快照不受虚拟机管理器的支持，且无法直接通过 **virsh** 删除。有关 QEMU 中外部快照的详细信息，请参见第 36.2.4 节“有效操作磁盘映像”。

实时快照

当原始 VM Guest 正在运行时创建的快照。内部实时快照支持保存设备以及内存和磁盘状态，而使用 **virsh** 的外部实时快照则支持保存内存状态和/或磁盘状态。

脱机快照

基于已关闭的 VM Guest 创建的快照。由于 Guest 的所有进程已停止且未使用任何内存，因此此类快照可确保数据完整性。

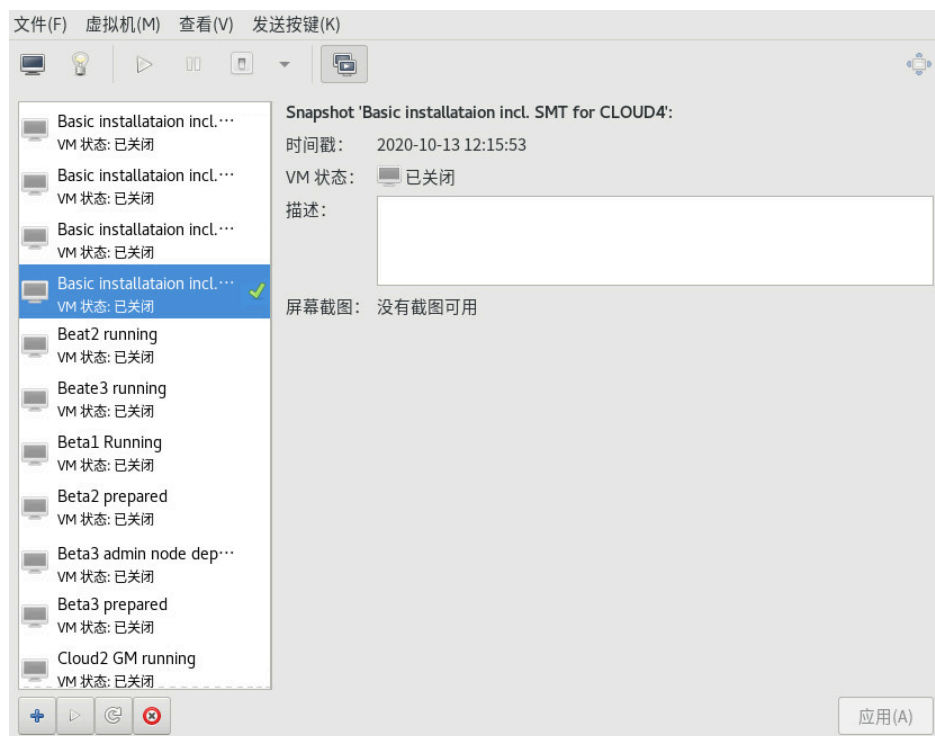
11.5.2 使用虚拟机管理器创建和管理快照



重要：仅限内部快照

虚拟机管理器仅支持实时或脱机的内部快照。

要在虚拟机管理器中打开快照管理视图，请按照第 11.2.1.1 节“使用虚拟机管理器打开图形控制台”中所述打开 VNC 窗口。现在请选择视图 > 快照，或单击工具栏中的管理虚拟机快照。



所选 VM Guest 的现有快照列表显示在窗口的左侧。上次启动的快照带有绿色对勾标记。窗口的右侧显示列表中当前标记的快照的细节。这些细节包括快照的标题和时戳、截取快照时 VM Guest 的状态以及说明。运行中 Guest 的快照还包括一个屏幕截图。可以直接在此视图中更改说明。其他快照数据不可更改。

11.5.2.1 创建快照

要截取 VM Guest 的新快照，请执行以下操作：

1. （可选）关闭 VM Guest 以创建脱机快照。
2. 单击 VNC 窗口左下角的添加。
创建快照窗口即会打开。
3. 提供名称和说明（可选）。截取快照后将无法更改该名称。为方便以后识别该快照，请使用“自述性名称”。
4. 单击完成确认。

11.5.2.2 删除快照

要删除 VM Guest 的快照，请执行以下操作：

1. 单击 VNC 窗口左下角的删除。
2. 单击是确认删除。

11.5.2.3 启动快照

要启动快照，请执行以下操作：

1. 单击 VNC 窗口左下角的运行。
2. 单击是确认启动。

11.5.3 使用 **virsh** 创建和管理快照

要列出某个域（以下示例中为 admin_server）的所有现有快照，请运行 snapshot-list 命令：

```
> virsh snapshot-list --domain sle-ha-node1
Name                               Creation Time           State
-----
sleha_12_sp2_b2_two_node_cluster 2016-06-06 15:04:31 +0200 shutoff
sleha_12_sp2_b3_two_node_cluster 2016-07-04 14:01:41 +0200 shutoff
sleha_12_sp2_b4_two_node_cluster 2016-07-14 10:44:51 +0200 shutoff
sleha_12_sp2_rc3_two_node_cluster 2016-10-10 09:40:12 +0200 shutoff
sleha_12_sp2_gmc_two_node_cluster 2016-10-24 17:00:14 +0200 shutoff
sleha_12_sp3_gm_two_node_cluster 2017-08-02 12:19:37 +0200 shutoff
sleha_12_sp3_rc1_two_node_cluster 2017-06-13 13:34:19 +0200 shutoff
sleha_12_sp3_rc2_two_node_cluster 2017-06-30 11:51:24 +0200 shutoff
sleha_15_b6_two_node_cluster      2018-02-07 15:08:09 +0100 shutoff
sleha_15_rc1_one-node             2018-03-09 16:32:38 +0100 shutoff
```

使用 snapshot-current command：命令显示上次启动的快照：

```
> virsh snapshot-current --domain admin_server
```



```
Basic installation incl. SMT for CLOUD4
```

运行 `snapshot-info` 命令可以获取有关特定快照的细节：

```
> virsh snapshot-info --domain admin_server \
    -name "Basic installation incl. SMT for CLOUD4"
Name:          Basic installation incl. SMT for CLOUD4
Domain:        admin_server
Current:       yes
State:         shutoff
Location:      internal
Parent:        Basic installation incl. SMT for CLOUD3-HA
Children:      0
Descendants:    0
Metadata:     yes
```

11.5.3.1 创建内部快照

要创建 VM Guest 的内部快照（实时或脱机快照），请如下所示使用 `snapshot-create-as` 命令：

```
> virsh snapshot-create-as --domain admin_server ❶ --name "Snapshot 1" ❷ \
    --description "First snapshot" ❸
```

- ❶ 域名。必需。
- ❷ 快照的名称。建议使用“自述性名称”，这样可以更轻松地识别该快照。必需。
- ❸ 快照的说明。可选。

11.5.3.2 创建外部快照

使用 `virsh` 可以创建 Guest 内存状态和/或磁盘状态的外部快照。

要同时创建 Guest 磁盘的实时和脱机外部快照，请指定 `--disk-only` 选项：

```
> virsh snapshot-create-as --domain admin_server --name \
    "Offline external snapshot" --disk-only
```

可以指定 `--diskspec` 选项来控制外部文件的创建方式：

```
> virsh snapshot-create-as --domain admin_server --name \
  "Offline external snapshot" \
  --disk-only --diskspec vda,snapshot=external,file=/path/to/snapshot_file
```

要创建 Guest 内存的实时外部快照，请指定 `--live` 和 `--memspec` 选项：

```
> virsh snapshot-create-as --domain admin_server --name \
  "Offline external snapshot" --live \
  --memspec snapshot=external,file=/path/to/snapshot_file
```

要创建 Guest 磁盘和内存状态的实时外部快照，请结合使用 `--live`、`--diskspec` 和 `--memspec` 选项：

```
> virsh snapshot-create-as --domain admin_server --name \
  "Offline external snapshot" --live \
  --memspec snapshot=external,file=/path/to/snapshot_file
  --diskspec vda,snapshot=external,file=/path/to/snapshot_file
```

有关更多细节，请参见 [man 1 virsh](#) 中的 `SNAPSHOT COMMANDS` 部分。

11.5.3.3 删除快照

无法使用 `virsh` 删除外部快照。要删除 VM Guest 的内部快照并恢复其占用的磁盘空间，请使用 `snapshot-delete` 命令：

```
> virsh snapshot-delete --domain admin_server --snapshotname "Snapshot 2"
```

11.5.3.4 启动快照

要启动快照，请使用 `snapshot-revert` 命令：

```
> virsh snapshot-revert --domain admin_server --snapshotname "Snapshot 1"
```

要启动当前快照（用于启动 VM Guest 的快照），使用 `--current` 便已足够，不需要指定快照名称：

```
> virsh snapshot-revert --domain admin_server --current
```

11.6 删除 VM Guest

默认情况下，使用 **virsh** 删除 VM Guest 只会去除其 XML 配置。由于默认不会删除挂接的存储设备，因此您可以在另一个 VM Guest 上重复使用该存储设备。使用虚拟机管理器还可以删除 Guest 的存储文件。

11.6.1 使用虚拟机管理器删除 VM Guest

1. 在虚拟机管理器中，右键单击某个 VM Guest 项。
2. 从上下文菜单中选择删除。
3. 一个确认窗口即会打开。单击删除会永久擦除该 VM Guest。该删除操作不可恢复。
您还可以通过激活删除关联的存储文件来永久删除 Guest 的虚拟磁盘。该删除操作也不可恢复。

11.6.2 使用 **virsh** 删除 VM Guest

要删除 VM Guest，需先将其关闭。无法删除运行中的 Guest。有关关机的信息，请参见第 11.3 节“更改 VM Guest 的状态：启动、停止、暂停”。

要使用 **virsh** 删除 VM Guest，请运行 **virsh undefine VM_NAME**。

```
> virsh undefine sles12
```

没有可自动删除挂接的存储文件的选项。如果这些文件由 libvirt 管理，请按照第 9.2.1.4 节“从存储池中删除卷”中所述将其删除。

11.7 监控

11.7.1 使用虚拟机管理器进行监控

启动虚拟机管理器并连接到 VM 主机服务器后，所有运行中 Guest 的 CPU 使用率图表将会显示。

您也可以通过此工具获取有关磁盘和网络使用情况的信息，不过必须先为首选项中激活此功能：

1. 运行 **virt-manager**。
2. 选择编辑 > 首选项。
3. 从常规选项卡切换到轮询。
4. 选中要查看的活动类型对应的复选框：轮询磁盘 I/O、轮询网络 I/O 和轮询内存统计。
5. 如果需要，还可以使用更新状态间隔：n 秒来更改更新间隔。
6. 关闭首选项对话框。
7. 在视图 > 图表下选中应显示的图表。

此后，磁盘和网络统计数据也会显示在虚拟机管理器的主窗口中。

可以从 VNC 窗口获取更精确的数据。按照第 11.2.1 节“打开图形控制台”中所述打开 VNC 窗口。在工具栏或视图菜单中选择细节。可以通过左侧树菜单中的性能项显示统计数据。

11.7.2 使用 **virt-top** 进行监控

virt-top 是一个命令行工具，与众所周知的进程监控工具 **top** 类似。**virt-top** 使用 **libvirt**，因此能够显示不同超级管理程序上运行的 VM Guest 的统计数据。建议使用 **virt-top**，而不要使用 **xentop** 等特定于超级管理程序的工具。

virt-top 默认会显示所有运行中 VM Guest 的统计数据。显示的数据包括已用内存百分比 (%MEM)、已用 CPU 百分比 (%CPU)，以及 Guest 的运行时长 (TIME)。数据会定期更新（默认为每三秒更新一次）。下面显示了某台 VM 主机服务器上的输出，该服务器包含七个 VM Guest，其中有四个处于非活动状态：

```
virt-top 13:40:19 - x86_64 8/8CPU 1283MHz 16067MB 7.6% 0.5%
7 domains, 3 active, 3 running, 0 sleeping, 0 paused, 4 inactive D:0 0:0 X:0
CPU: 6.1% Mem: 3072 MB (3072 MB by guests)
```

ID	S	RDRQ	WRRQ	RXBY	TXBY	%CPU	%MEM	TIME	NAME
7	R	123	1	18K	196	5.8	6.0	0:24.35	sled12_sp1

```

 6 R    1    0 18K    0 0.2  6.0   0:42.51 sles12_sp1
 5 R    0    0 18K    0 0.1  6.0   85:45.67 opensuse_leap
-                                     (Ubuntu_1410)
-                                     (debian_780)
-                                     (fedora_21)
-                                     (sles11sp3)

```

输出默认按 ID 排序。使用以下组合键可以更改排序字段：

Shift - P : CPU 使用率

Shift - M : Guest 分配的内存总量

Shift - T : 时间

Shift - I : ID

要使用任何其他字段进行排序，请按 **Shift - F** 并从列表选择一个字段。要切换排序顺序，请使用 **Shift - R**。

virt-top 还支持基于 VM Guest 数据生成不同的视图，按以下键可以即时更改视图：

0 : 默认视图

1 : 显示物理 CPU

2 : 显示网络接口

3 : 显示虚拟磁盘

virt-top 支持使用更多热键来更改数据视图，并支持许多可以影响程序行为的命令行开关。

有关详细信息，请访问 [man 1 virt-top](#)。

11.7.3 使用 **kvm_stat** 进行监控

kvm_stat 可用于跟踪 KVM 性能事件。它会监控 `/sys/kernel/debug/kvm`，因此需要挂载 `debugfs`。SUSE Linux Enterprise Server 上默认应该已挂载 `debugfs`。如果未挂载，请使用以下命令：

```
> sudo mount -t debugfs none /sys/kernel/debug
```

可在三种不同的模式下使用 **kvm_stat**：

```
kvm_stat          # update in 1 second intervals
```

```
kvm_stat -l                # 1 second snapshot
kvm_stat -l > kvmstats.log # update in 1 second intervals in log format
                           # can be imported to a spreadsheet
```

例 11.1：kvm_stat 的典型输出

```
kvm statistics

efer_reload          0          0
exits                11378946    218130
fpu_reload           62144       152
halt_exits           414866       100
halt_wakeup          260358        50
host_state_reload    539650       249
hypercalls           0          0
insn_emulation       6227331    173067
insn_emulation_fail  0          0
invlpg               227281        47
io_exits             113148        18
irq_exits            168474       127
irq_injections       482804       123
irq_window           51270        18
largepages           0          0
mmio_exits           6925         0
mmu_cache_miss       71820         19
mmu_flooded          35420         9
mmu_pde_zapped       64763         20
mmu_pte_updated      0          0
mmu_pte_write        213782        29
mmu_recycled          0          0
mmu_shadow_zapped    128690        17
mmu_unsync           46          -1
nmi_injections       0          0
nmi_window           0          0
pf_fixed             1553821       857
pf_guest             1018832       562
remote_tlb_flush     174007         37
request_irq           0          0
signal_exits         0          0
```

tlb_flush	394182	148
-----------	--------	-----

有关如何解释这些值的更多信息，请参见 <https://clalance.blogspot.com/2009/01/kvm-performance-tools.html> 。

12 连接和授权

如果您要管理多个 VM 主机服务器，而每个服务器又托管了多个 VM Guest，那么管理工作很快就会变得困难起来。libvirt 的一个优势是，它能够一次连接到多个 VM 主机服务器，提供单个接口用于管理所有 VM Guest 以及连接其图形控制台。

为确保只有授权用户能够建立连接，libvirt 提供了多种可与不同授权机制（套接字、Polkit、SASL 和 Kerberos）结合使用的连接类型（通过 TLS、SSH、Unix 套接字和 TCP）。

12.1 身份验证

有权管理 VM Guest 和访问其图形控制台的用户应该限定在明确定义的人员范围内。为实现此目标，可在 VM 主机服务器上使用以下身份验证方法：

- 使用权限和组所有权对 Unix 套接字进行访问控制。此方法仅适用于 libvirtd 连接。
- 使用 Polkit 对 Unix 套接字进行访问控制。此方法仅适用于本地 libvirtd 连接。
- 使用 SASL（简单身份验证和安全层）进行用户名和口令身份验证。此方法适用于 libvirtd 和 VNC 连接。使用 SASL 不需要在服务器上拥有实际的用户帐户，因为 SASL 使用自己的数据库来存储用户名和口令。通过 SASL 进行身份验证的连接会加密。
- Kerberos 身份验证。此方法仅适用于 libvirtd 连接，本手册不予介绍。有关详细信息，请参见https://libvirt.org/auth.html#ACL_server_kerberos。
- 单口令身份验证。此方法仅适用于 VNC 连接。

重要：libvirtd 和 VNC 的身份验证需要分开配置

对 VM Guest 管理功能的访问权限（通过 libvirtd）以及对其图形控制台的访问权限始终需要分开配置。如果对管理工具的访问权限施加了限制，这些限制**不会**自动应用到 VNC 连接。

通过 TLS/SSL 连接远程访问 VM Guest 时，可以通过仅允许特定的组拥有证书密钥文件的读取权限，在每个客户端上间接控制访问。有关详细信息，请参见第 12.3.2.5 节“限制访问（安全考虑因素）”。

12.1.1 libvirtd authentication

`libvirtd` 身份验证在 `/etc/libvirt/libvirtd.conf` 中配置。此处进行的配置将应用到所有 `libvirt` 工具，例如虚拟机管理器或 `virsh`。

`libvirt` 提供了两个套接字：一个只读套接字用于监控目的，一个读写套接字用于管理操作。可以单独配置对这两个套接字的访问。默认情况下，这两个套接字由 `root.root` 拥有。默认仅向用户 `root` 授予对读写套接字的访问权限 (0700)，而对于只读套接字的访问权限则完全开放 (0777)。

以下说明介绍如何配置对读写套接字的访问权限。这些说明同样也适用于只读套接字。所有配置步骤均需在 VM 主机服务器上执行。



注意：SUSE Linux Enterprise Server 上的默认身份验证设置

SUSE Linux Enterprise Server 上的默认身份验证方法是对 Unix 套接字进行访问控制。只有用户 `root` 能够进行身份验证。在 VM 主机服务器上以非 `root` 用户身份访问 `libvirt` 工具时，需要通过 Polkit 提供一次 `root` 口令。然后，系统将向您授予对当前和将来会话的访问权限。

或者，您可以配置 `libvirt`，以允许非特权用户进行“system”访问。有关详细信息，请参见第 12.2.1 节“非特权用户的“system”访问权限”。

建议的身份验证方法

本地连接

第 12.1.1.2 节“使用 Polkit 对 Unix 套接字进行访问控制”

第 12.1.1.1 节“使用权限和组所有权对 Unix 套接字进行访问控制”

基于 SSH 的远程隧道

第 12.1.1.1 节“使用权限和组所有权对 Unix 套接字进行访问控制”

远程 TLS/SSL 连接

第 12.1.1.3 节 “使用 SASL 进行用户名和口令身份验证”

无（通过限制对证书的访问在客户端控制访问权限）

12.1.1.1 使用权限和组所有权对 Unix 套接字进行访问控制

要为非 root 帐户授予访问权限，请配置特定的组（在下面的示例中为 libvirt）拥有且可访问的套接字。此身份验证方法可用于本地和远程 SSH 连接。

1. 创建应该拥有套接字的组（如果不存在）：

```
> sudo groupadd libvirt
```

重要：组需要存在

在重新启动 libvirtd 之前，该组必须存在。否则，重新启动将会失败。

2. 将所需用户添加到该组：

```
> sudo usermod --append --groups libvirt tux
```

3. 如下所示在 /etc/libvirt/libvirtd.conf 中更改配置：

```
unix_sock_group = "libvirt" ❶  
unix_sock_rw_perms = "0770" ❷  
auth_unix_rw = "none" ❸
```

- ❶ 组所有权将设置给组 libvirt。
- ❷ 设置对套接字的访问权限 (srwxrwx--)。
- ❸ 禁用其他身份验证方法（Polkit 或 SASL）。访问仅通过套接字权限来控制。

4. 重新启动 libvirtd:

```
> sudo systemctl start libvirtd
```

12.1.1.2 使用 Polkit 对 Unix 套接字进行访问控制

在 SUSE Linux Enterprise Server 上，对于非远程连接，默认的身份验证方法是使用 Polkit 对 Unix 套接字进行访问控制。因此，无需对 `libvirt` 配置进行更改。启用 Polkit 授权后，对上述两个套接字的权限将默认为 `0777`，每个尝试访问套接字的应用程序将需要通过 Polkit 进行身份验证。

重要：仅对本地连接进行 Polkit 身份验证

只能对 VM 主机服务器本身上的本地连接进行 Polkit 身份验证，因为 Polkit 不会处理远程身份验证。

有关访问 `libvirt` 套接字的策略有两个：

- **org.libvirt.unix.monitor**：访问只读套接字
- **org.libvirt.unix.manage**：访问读写套接字

默认情况下，读写套接字访问策略是使用 `root` 口令进行一次身份验证，并授予对当前和将来的会话的特权。

要向用户授予无需提供 `root` 口令即可访问套接字的权限，您需要在 `/etc/polkit-1/rules.d` 中创建一条规则。创建包含以下内容的 `/etc/polkit-1/rules.d/10-grant-libvirt` 文件，以向组 `libvirt` 的所有成员授予对读写套接字的访问权限。

```
polkit.addRule(function(action, subject) {
    if (action.id == "org.libvirt.unix.manage" && subject.isInGroup("libvirt")) {
        return polkit.Result.YES;
    }
});
```

12.1.1.3 使用 SASL 进行用户名和口令身份验证

SASL 提供用户名和口令身份验证以及数据加密（默认为 `digest-md5`）。由于 SASL 会维护自己的用户数据库，VM 主机服务器上无需存在用户。TCP 连接需要 SASL，并且在 TLS/SSL 连接上也需要 SASL。

！ 重要：普通 TCP 以及提供 digest-md5 加密的 SASL

在未通过其他方式加密的 TCP 连接上使用 digest-md5 加密并不会为生产环境提供充足的安全性。建议仅在测试环境中使用这种加密。

💡 提示：在 TLS/SSL 上进行 SASL 身份验证

通过限制对证书密钥文件的访问，可以在**客户端**间接控制通过远程 TLS/SSL 连接进行的访问。但在处理大量客户端时，这种方法可能容易很出错。对 TLS 使用 SASL 可以另外在服务器端控制访问，从而提高安全性。

要配置 SASL 身份验证，请执行以下操作：

1. 如下所示在 `/etc/libvirt/libvirtd.conf` 中更改配置：

a. 要为 TCP 连接启用 SASL，请使用以下配置：

```
auth_tcp = "sasl"
```

b. 要为 TLS/SSL 连接启用 SASL，请使用以下配置：

```
auth_tls = "sasl"
```

2. 重启 `libvirtd`：

```
> sudo systemctl restart libvirtd
```

3. libvirt SASL 配置文件位于 `/etc/sasl2/libvirtd.conf`。通常无需更改默认设置。

但是，如果在 TLS 上使用 SASL，您可以通过将设置 `mech_list` 参数的行注释掉，来关闭会话加密以避免额外的开销（TLS 连接已加密）。请仅对 TLS/SASL 执行此操作。对于 TCP 连接，此参数必须设置为 `digest-md5`。

```
#mech_list: digest-md5
```

4. 默认不会配置任何 SASL 用户，因此无法登录。使用以下命令可管理用户：

添加用户 tux

```
saslpaswd2 -a libvirt tux
```

删除用户 tux

```
saslpaswd2 -a libvirt -d tux
```

列出现有用户

```
sasldblistusers2 -f /etc/libvirt/passwd.db
```



提示：virsh 和 SASL 身份验证

使用 SASL 身份验证时，每次您发出 virsh 命令，都会收到输入用户名和口令的提示。在外壳模式下使用 virsh 可避免出现此提示。

12.1.2 VNC 身份验证

由于对 VM Guest 图形控制台的访问不是由 libvirt 控制，而是由特定的超级管理程序控制，因此始终需要另外配置 VNC 身份验证。主配置文件为 /etc/libvirt/<hypervisor>.conf。本节介绍的是 QEMU/KVM 超级管理程序，因此目标配置文件为 /etc/libvirt/qemu.conf。



注意：Xen 的 VNC 身份验证

与 KVM 相比，Xen 目前最多只能在每个 VM 上设置口令，除此之外不会提供更复杂的 VNC 身份验证。请参见下面的 <graphics type='vnc'... libvirt 配置选项。

可用的身份验证类型有两种：SASL 和单口令身份验证。如果您是使用 SASL 进行 libvirt 身份验证的，我们强烈建议也将它用于 VNC 身份验证 — 这样就可以共享同一数据库。

第三种限制对 VM Guest 的访问的方法是在 VNC 服务器上启用 TLS 加密。这要求 VNC 客户端有权访问 x509 客户端证书。通过限制对这些证书的访问，可以在客户端间接控制访问。有关细节，请参见第 12.3.2.4.2 节“基于 TLS/SSL 的 VNC：客户端配置”。

12.1.2.1 使用 SASL 进行用户名和口令身份验证

SASL 提供用户名和口令身份验证以及数据加密。由于 SASL 会维护自己的用户数据库，VM 主机服务器上无需存在用户。与对 `libvirt` 使用 SASL 身份验证一样，您可以在 TLS/SSL 连接上使用 SASL。有关配置这些连接的细节，请参见第 12.3.2.4.2 节“基于 TLS/SSL 的 VNC：客户端配置”。

要为 VNC 配置 SASL 身份验证，请执行以下操作：

1. 创建 SASL 配置文件。建议使用现有的 `libvirt` 文件。如果您已经为 `libvirt` 配置 SASL，并打算使用相同的设置（包括同一用户名和口令数据库），则使用简单的链接即可：

```
> sudo ln -s /etc/sasl2/libvirt.conf /etc/sasl2/qemu.conf
```

如果您只是为 VNC 设置 SASL，或者打算使用与 `libvirt` 不同的配置，请复制现有文件以用作模板：

```
> sudo cp /etc/sasl2/libvirt.conf /etc/sasl2/qemu.conf
```

然后根据需要编辑该文件。

2. 如下所示在 `/etc/libvirt/qemu.conf` 中更改配置：

```
vnc_listen = "0.0.0.0"
vnc_sasl = 1
sasldb_path: /etc/libvirt/qemu_passwd.db
```

第一个参数使 VNC 侦听所有公共接口（而不仅仅是本地主机），第二个参数启用 SASL 身份验证。

3. 默认不会配置任何 SASL 用户，因此无法登录。使用以下命令可管理用户：

添加用户 tux

```
> saslpasswd2 -f /etc/libvirt/qemu_passwd.db -a qemu tux
```

删除用户 tux

```
> saslpasswd2 -f /etc/libvirt/qemu_passwd.db -a qemu -d tux
```

列出现有用户

```
> sasldblistusers2 -f /etc/libvirt/qemu_passwd.db
```

4. 重启 libvirtd:

```
> sudo systemctl restart libvirtd
```

5. 重启在更改配置之前已在运行的所有 VM Guest。未重启的 VM Guest 无法对 VNC 连接使用 SASL 身份验证。



注意：支持的 VNC 查看器

目前，虚拟机管理器和 **virt-viewer** 均支持 SASL 身份验证。这两个查看器还支持 TLS/SSL 连接。

12.1.2.2 单口令身份验证

您也可以通过设置 VNC 口令来控制对 VNC 服务器的访问。可以为所有 VM Guest 设置一个全局口令，或者为每个 Guest 设置单独的口令。后一种做法需要编辑 VM Guest 的配置文件。



注意：始终设置全局口令

如果您要使用单口令身份验证，比较好的做法是设置一个全局口令，即使为每个 VM Guest 设置了口令也是如此。这样，在您忘记设置虚拟机的口令时，可以通过“回退”口令保护您的虚拟机。仅当未为计算机设置其他口令时，才会使用全局口令。

过程 12.1：设置全局 VNC 口令

1. 如下所示在 `/etc/libvirt/qemu.conf` 中更改配置：

```
vnc_listen = "0.0.0.0"
vnc_password = "PASSWORD"
```

第一个参数使 VNC 侦听所有公共接口（而不仅仅是本地主机），第二个参数设置口令。口令的最大长度为八个字符。

2. 重新启动 `libvirtd`:

```
> sudo systemctl restart libvirtd
```

3. 重新启动在更改配置之前已在运行的所有 VM Guest。未重新启动的 VM Guest 无法对 VNC 连接使用口令身份验证。

过程 12.2：设置 VM GUEST 特定的 VNC 口令

1. 按如下所示在 `/etc/libvirt/qemu.conf` 中更改配置，使 VNC 侦听所有公共接口（而不仅仅是本地主机）。

```
vnc_listen = "0.0.0.0"
```

2. 在编辑器中打开 VM Guest 的 XML 配置文件。请将以下示例中的 `VM_NAME` 替换为 VM Guest 的名称。使用的编辑器默认为 `$EDITOR`。如果未设置该变量，则使用 `vi`。

```
> virsh edit VM_NAME
```

3. 搜索包含 `type='vnc'` 属性的 `<graphics>` 元素，例如：

```
<graphics type='vnc' port='-1' autoport='yes' />
```

4. 添加 `passwd=PASSWORD` 属性，然后保存文件并退出编辑器。口令的最大长度为八个字符。

```
<graphics type='vnc' port='-1' autoport='yes' passwd='PASSWORD' />
```


5. 重新启动 `libvirtd`:

```
> sudo systemctl restart libvirtd
```

6. 重新启动在更改配置之前已在运行的所有 VM Guest。未重新启动的 VM Guest 无法对 VNC 连接使用口令身份验证。



警告：VNC 协议的安全性

VNC 被认为是不安全的协议。尽管口令是以加密方式发送的，但如果攻击者可以嗅探到已加密的口令和加密密钥，该口令可能就会被利用。因此，建议将 VNC 与 TLS/SSL 结合使用，或通过 SSH 建立隧道。**`virt-viewer`**、虚拟机管理器和 Remmina（请参见《管理指南》，第 14 章“使用 VNC 的远程图形会话”，第 14.2 节“Remmina：远程桌面客户端”）支持这两种方法。

12.2 连接到 VM 主机服务器

要使用 `libvirt` 连接到超级管理程序，需要指定统一资源标识符 (URI)。使用 **`virsh`** 和 **`virt-viewer`** 时需要此 URI（在 VM 主机服务器上以 `root` 身份操作时例外），而使用虚拟机管理器时，此 URI 为可选项。虽然可以使用连接参数（例如 **`virt-manager -c qemu:///system`**）来调用虚拟机管理器，但虚拟机管理器还是提供了一个图形界面用于创建连接 URI。有关详细信息，请参见第 12.2.2 节“使用虚拟机管理器管理连接”。

```
HYPERVERSOR ❶ +PROTOCOL ❷ ://USER@REMOTE ❸ /CONNECTION_TYPE ❹
```

- ❶ 指定超级管理程序。SUSE Linux Enterprise Server 目前支持以下超级管理程序：**`test`**（用于测试）、**`qemu`** (KVM) 和 **`xen`** (Xen)。此参数是必需的。
- ❷ 连接到远程主机时，请在此处指定协议。其值可以是：**`ssh`**（通过 SSH 隧道连接）、**`tcp`**（使用 SASL/Kerberos 身份验证进行 TCP 连接）或 **`tls`**（进行 TLS/SSL 加密的连接并通过 x509 证书完成身份验证）。
- ❸ 连接到远程主机时，请指定用户名和远程主机名。如果未指定用户名，将使用调用了该命令的用户名 (`$USER`)。有关详细信息，请参见下文。对于 TLS 连接，需要完全按照 x509 证书指定主机名。

- ④ 连接到 QEMU/KVM 超级管理程序时，接受两种连接类型：system（完全访问权限）或 session（受限访问权限）。由于 SUSE Linux Enterprise Server 不支持 session 访问权限，因此本文档将重点介绍 system 访问权限。

超级管理程序连接 URI 示例

test:///default

连接到本地测试超级管理程序。

qemu:///system 或 xen:///system

连接到本地主机上拥有完全访问权限（system 类型）的 QEMU/Xen 超级管理程序。

qemu+ssh://tux@mercury.example.com/system 或 xen+ssh://

tux@mercury.example.com/system

连接到远程主机 `mercury.example.com` 上的 QEMU/Xen 超级管理程序。连接是通过 SSH 隧道建立的。

qemu+tls://saturn.example.com/system 或 xen+tls://saturn.example.com/system

连接到远程主机 `mercury.example.com` 上的 QEMU/Xen 超级管理程序。连接是使用 TLS/SSL 建立的。

有关更多细节和示例，请参见 <https://libvirt.org/uri.html> 上的 `libvirt` 文档。



注意：URI 中的用户名

使用 Unix 套接字身份验证时，需要指定用户名（无论是使用用户/口令身份验证模式还是 Polkit）。这适用于所有 SSH 连接和本地连接。

使用 SASL 身份验证（对于 TCP 或 TLS 连接）时或者不对 TLS 连接执行额外的服务器端身份验证时，无需指定用户名。使用 SASL 时不会评估用户名 — 在任何情况下，系统都会提示您输入 SASL 用户/口令组合。

12.2.1 非特权用户的“system”访问权限

如上文所述，可以使用两种不同的协议来与 QEMU 超级管理程序建立连接：`session` 和 `system`。“`session`”连接建立时具有与客户端程序相同的特权。此类连接适用于桌面虚拟化，因为它会受到限制（例如，无 USB/PCI 设备分配、无虚拟网络设置，只能对 `libvirtd` 进行受限的远程访问）。

适用于服务器虚拟化的“`system`”连接不存在功能限制，但默认仅可供 `root` 访问。不过，在将 DAC（自主访问控制）驱动程序添加到 `libvirt` 后，现在可以向非特权用户授予“`system`”访问权限。要向用户 `tux` 授予“`system`”访问权限，请执行以下操作：

过程 12.3：向普通用户授予“SYSTEM”访问权限

1. 按照第 12.1.1.1 节“使用权限和组所有权对 Unix 套接字进行访问控制”中所述通过 Unix 套接字启用访问权限。该示例向 `libvirt` 组的所有成员授予 `libvirt` 访问权限，并使 `tux` 成为此组的成员。这样可确保 `tux` 能够使用 `virsh` 或虚拟机管理器进行连接。
2. 编辑 `/etc/libvirt/qemu.conf` 并如下所示更改配置：

```
user = "tux"
group = "libvirt"
dynamic_ownership = 1
```

这样可确保 VM Guest 由 `tux` 启动，并且 `tux` 能够访问和修改已绑定至 Guest 的资源（例如虚拟磁盘）。

3. 使 `tux` 成为 `kvm` 组的成员：

```
> sudo usermod --append --groups kvm tux
```

需要执行此步骤才能授予对 `/dev/kvm` 的访问权限，而要启动 VM Guest 就必须具有此访问权限。

4. 重启 `libvirtd`：

```
> sudo systemctl restart libvirtd
```

12.2.2 使用虚拟机管理器管理连接

虚拟机管理器对它管理的每个 VM 主机服务器都使用一个 Connection。每个连接都包含相应主机上的所有 VM Guest。默认已配置并已建立与本地主机的连接。

配置的所有连接都显示在虚拟机管理器主窗口中。活动连接带有小三角形标记，单击该标记可以收起或展开此连接的 VM Guest 列表。

非活动连接以灰色列出，带有 Not Connected 标记。可以双击或者右键单击这些连接，然后从上下文菜单中选择连接。还可以通过此菜单删除现有连接。



注意：编辑现有连接

无法编辑现有的连接。要更改连接，请使用所需参数创建新连接，然后删除“旧”连接。

要在虚拟机管理器中添加新连接，请执行以下操作：

1. 选择文件 > 添加连接
2. 选择主机的虚拟机管理程序（Xen 或 QEMU/KVM）
3. （可选）要设置远程连接，请选择连接到远程主机。有关详细信息，请访问 [第 12.3 节“配置远程连接”](#)。

如果要设置远程连接，请以 USERNAME@REMOTE _HOST 格式指定远程计算机的主机名。



重要：指定用户名

无需为 TCP 和 TLS 连接指定用户名：使用这些连接时，系统不会评估用户名。但在使用 SSH 连接时，如果您要以非 root 用户身份进行连接，则必须指定用户名。

4. 如果您不希望在启动虚拟机管理器时自动启动连接，请停用自动连接。
5. 单击连接以完成配置。

12.3 配置远程连接

libvirt 的一大优势是能够从一个中心位置管理不同远程主机上的 VM Guest。本节提供有关如何配置服务器和客户端以允许远程连接的详细说明。

12.3.1 基于 SSH 的远程隧道 (qemu+ssh 或 xen+ssh)

只需能够接受 SSH 连接，即可在 VM 主机服务器上启用基于 SSH 的远程隧道连接。确保 SSH 守护程序已启动 (`systemctl status sshd`)，并且已在防火墙中打开服务 `SSH` 的端口。

可以按照第 12.1.1.1 节 “使用权限和组所有权对 Unix 套接字进行访问控制” 中所述，使用传统的文件用户/组所有权和权限完成 SSH 连接的用户身份验证。以用户 `tux` 的身份进行连接 (`qemu+ssh://tuxsIVname;/system` 或 `xen+ssh://tuxsIVname;/system`) 是开箱即用的功能，无需在 `libvirt` 一端进行额外的配置。

通过 SSH 进行连接 (`qemu+ssh://USER@SYSTEM` 或 `xen+ssh://USER@SYSTEM`) 时，需要提供 `USER` 的口令。按照《安全和强化指南》，第 22 章 “使用 OpenSSH 保护网络操作”，第 22.6 节 “公共密钥身份验证” 中所述将公共密钥复制到 VM 主机服务器上的 `~USER/.ssh/authorized_keys` 可以避免此情况。在发起连接的计算机上使用 `gnome-keyring` 会更方便。有关详细信息，请访问《安全和强化指南》，第 22 章 “使用 OpenSSH 保护网络操作”，第 22.9 节 “使用 `gnome-keyring` 自动进行公共密钥登录”。

12.3.2 使用 x509 证书进行远程 TLS/SSL 连接 (qemu+tls 或 xen+tls)

与使用 SSH 相比，使用通过 x509 证书实现 TLS/SSL 加密和身份验证的 TCP 连接在设置上要复杂得多，不过此方法的可缩放性也高得多。如果您需要管理多个 VM 主机服务器，而这些服务器的管理员数量各异，请使用此方法。

12.3.2.1 基本概念

TLS（传输层安全）使用证书来加密两台计算机之间的通讯。发起连接的计算机一律视为“客户端”，使用的是“客户端证书”；接收方计算机一律视为“服务器”，使用的是“服务器证书”。例如，如果您通过一个中心桌面来管理 VM 主机服务器，则此方案适用。

如果连接是从两台计算机发起的，则每台计算机都需要有一个客户端证书和一个服务器证书。例如，如果您将 VM Guest 从一台主机迁移到另一台主机，就需要符合这种要求。

每个 x509 证书都有一个匹配的私用密钥文件。只有结合证书和私用密钥文件才能正确标识自身。为确保证书由宣称的拥有者颁发，该证书需由称作证书颁发机构 (CA) 的中心证书签名并颁发。客户端证书和服务器证书必须由同一个 CA 颁发。

! 重要：用户身份验证

使用远程 TLS/SSL 连接只能确保允许两台计算机进行特定方向的通讯。通过限制对证书的访问，可以在客户端间接地限制只有特定用户可进行访问。有关详细信息，请访问[第 12.3.2.5 节“限制访问（安全考虑因素）”](#)。

`libvirt` 还支持使用 SASL 在服务器上对用户进行身份验证。有关详细信息，请访问[第 12.3.2.6 节“对 TLS 套接字使用 SASL 进行集中式用户身份验证”](#)。

12.3.2.2 配置 VM 主机服务器

VM 主机服务器是接收连接的计算机，因此需要安装**服务器**证书，还需要安装 CA 证书。准备好证书后，可以为 `libvirt` 开启 TLS 支持。

1. 创建服务器证书，然后将其连同相应的 CA 证书一并导出。
2. 在 VM 主机服务器上创建以下目录：

```
> sudo mkdir -p /etc/pki/CA/ /etc/pki/libvirt/private/
```

按如下所示安装证书：

```
> sudo /etc/pki/CA/cacert.pem
> sudo /etc/pki/libvirt/servercert.pem
> sudo /etc/pki/libvirt/private/serverkey.pem
```

! 重要：限制对证书的访问

确保按照[第 12.3.2.5 节“限制访问（安全考虑因素）”](#)中所述限制对证书的访问。

3. 通过启用相关套接字并重新启动 `libvirtd` 来启用 TLS 支持：

```
> sudo systemctl stop libvirtd.service
```

```
> sudo systemctl enable --now libvirtd-tls.socket
> sudo systemctl start libvirtd.service
```

4. 默认情况下，libvirt 使用 TCP 端口 16514 来接受 TLS 安全连接。请在防火墙中打开此端口。

❗ 重要：在启用了 TLS 的情况下重新启动 libvirtd

如果您为 libvirt 启用了 TLS，则需要准备好服务器证书，否则重新启动 libvirtd 会失败。如果您更改了证书，也需要重新启动 libvirtd。

12.3.2.3 配置客户端并测试设置

客户端是发起连接的计算机，因此需要安装**客户端**证书，还需要安装 CA 证书。

1. 创建客户端证书，然后将其连同相应的 CA 证书一并导出。
2. 在客户端上创建以下目录：

```
> sudo mkdir -p /etc/pki/CA/ /etc/pki/libvirt/private/
```

按如下所示安装证书：

```
> sudo /etc/pki/CA/cacert.pem
> sudo /etc/pki/libvirt/clientcert.pem
> sudo /etc/pki/libvirt/private/clientkey.pem
```

❗ 重要：限制对证书的访问

确保按照第 12.3.2.5 节“**限制访问（安全考虑因素）**”中所述限制对证书的访问。

3. 发出以下命令测试客户端/服务器设置。请将 mercury.example.com 替换为您的 VM 主机服务器名称。指定在创建服务器证书时所用的完全限定主机名。

```
#QEMU/KVM
virsh -c qemu+tls://mercury.example.com/system list --all
```

```
#Xen
virsh -c xen+tls://mercury.example.com/system list --all
```

如果您的设置正确，则您可以看到 VM 主机服务器上已在 [libvirt](#) 中注册的所有 VM Guest。

12.3.2.4 为 TLS/SSL 连接启用 VNC

目前只能通过几个工具来支持基于 TLS 的 VNC 通讯。[tightvnc](#) 或 [tigervnc](#) 等常见 VNC 查看器不支持 TLS/SSL。唯一可替代虚拟机管理器和 [virt-viewer](#) 的工具是 [remmina](#)（请参见《管理指南》，第 14 章“使用 VNC 的远程图形会话”，第 14.2 节“Remmina：远程桌面客户端”）。

12.3.2.4.1 基于 TLS/SSL 的 VNC：VM 主机服务器配置

要通过基于 TLS/SSL 的 VNC 访问图形控制台，需按如下所述配置 VM 主机服务器：

1. 在防火墙中打开服务 [VNC](#) 的端口。
2. 创建目录 [/etc/pki/libvirt-vnc](#)，并如下所示将证书链接到此目录：

```
> sudo mkdir -p /etc/pki/libvirt-vnc && cd /etc/pki/libvirt-vnc
> sudo ln -s /etc/pki/CA/cacert.pem ca-cert.pem
> sudo ln -s /etc/pki/libvirt/servercert.pem server-cert.pem
> sudo ln -s /etc/pki/libvirt/private/serverkey.pem server-key.pem
```

3. 编辑 [/etc/libvirt/qemu.conf](#) 并设置以下参数：

```
vnc_listen = "0.0.0.0"
vnc_tls = 1
vnc_tls_x509_verify = 1
```

4. 重启 [libvirtd](#)：

```
> sudo systemctl restart libvirtd
```


! 重要：需要重新启动 VM Guest

只有在启动 VM Guest 时才会设置 VNC TLS。因此，需要重新启动更改配置前已在运行的所有计算机。

12.3.2.4.2 基于 TLS/SSL 的 VNC：客户端配置

需要在客户端执行的唯一操作是，将 x509 客户端证书放到可由所选客户端识别的位置。但是，虚拟机管理器和 **virt-viewer** 要求将证书放到不同的位置。虚拟机管理器可以从应用到所有用户的系统范围位置或者从每个用户的位置读取数据。在初始化与远程 VNC 会话的连接时，Remmina（请参见《管理指南》，第 14 章“使用 VNC 的远程图形会话”，第 14.2 节“Remmina：远程桌面客户端”）会要求提供证书位置。

虚拟机管理器 (virt-manager)

要连接到远程主机，虚拟机管理器需要使用第 12.3.2.3 节“配置客户端并测试设置”中所述的设置。要能够通过 VNC 进行连接，还需要将客户端证书放到以下位置：

系统范围的位置

```
/etc/pki/CA/cacert.pem  
/etc/pki/libvirt-vnc/clientcert.pem  
/etc/pki/libvirt-vnc/private/clientkey.pem
```

每个用户的位置

```
/etc/pki/CA/cacert.pem  
~/.pki/libvirt-vnc/clientcert.pem  
~/.pki/libvirt-vnc/private/clientkey.pem
```

virt-viewer

virt-viewer 仅接受位于系统范围位置的证书：

```
/etc/pki/CA/cacert.pem  
/etc/pki/libvirt-vnc/clientcert.pem
```

/etc/pki/libvirt-vnc/private/clientkey.pem

❗ 重要：限制对证书的访问

确保按照第 12.3.2.5 节“限制访问（安全考虑因素）”中所述限制对证书的访问。

12.3.2.5 限制访问（安全考虑因素）

每个 x509 证书都由两个部分构成：公共证书和私用密钥。只有使用了这两个部分，客户端才能通过身份验证。因此，对客户端证书及其私用密钥拥有读取访问权限的任何用户都可以访问您的 VM 主机服务器。另一方面，具有完整服务器证书的任意计算机都可以假装是 VM 主机服务器。这种情况不是您希望发生的，因此至少需要尽可能地限制对私用密钥文件的访问。要控制对密钥文件的访问，最简单的方法就是使用访问权限。

服务器证书

服务器证书需可由 QEMU 进程读取。在 SUSE Linux Enterprise Server QEMU 上，通过 libvirt 工具启动的进程由 root 拥有，因此只要 root 能够读取证书即可：

```
> chmod 700 /etc/pki/libvirt/private/  
> chmod 600 /etc/pki/libvirt/private/serverkey.pem
```

如果您在 /etc/libvirt/qemu.conf 中更改了 QEMU 进程的所有权，则也需要调整密钥文件的所有权。

系统范围的客户端证书

要控制对可在系统范围使用的密钥文件的访问，请限制只有特定的组具有读取访问权限，这样只有该组的成员可以读取密钥文件。以下示例会创建一个 libvirt 组，并将 clientkey.pem 文件及其父目录的组所有权设置为 libvirt。之后，限制只有拥有者和组具有访问权限。最后，将用户 tux 添加到 libvirt 组，如此该用户便可以访问密钥文件。

```
CERTPATH="/etc/pki/libvirt/"  
# create group libvirt  
groupadd libvirt  
# change ownership to user root and group libvirt
```

```
chown root.libvirt $CERTPATH/private $CERTPATH/clientkey.pem
# restrict permissions
chmod 750 $CERTPATH/private
chmod 640 $CERTPATH/private/clientkey.pem
# add user tux to group libvirt
usermod --append --groups libvirt tux
```

每个用户的证书

需要将用于通过 VNC 访问 VM Guest 图形控制台的用户特定客户端证书放在用户主目录下的 `~/.pki` 中。与 SSH 不同，使用这些证书的 VNC 查看器不会检查私用密钥文件的访问权限（举例而言）。因此，用户需自行负责确保密钥文件不可由其他人读取。

12.3.2.5.1 限制从服务器端的访问

默认情况下，具有相应客户端证书的每个客户端都可以连接到接受 TLS 连接的 VM 主机服务器。因此，可以按照第 12.1.1.3 节“使用 SASL 进行用户名和口令身份验证”中所述通过 SASL 来实施额外的服务器端身份验证。

还可以通过 DN（判别名）允许列表来限制访问，这样只有其证书与白名单中的某个 DN 匹配的客户端才能建立连接。

将允许的 DN 列表添加到 `/etc/libvirt/libvirtd.conf` 中的 `tls_allowed_dn_list`。此列表可以包含通配符。请不要指定空列表，因为这会导致拒绝所有连接。

```
tls_allowed_dn_list = [
    "C=US,L=Provo,O=SUSE Linux Products GmbH,OU=*,CN=venus.example.com,EMAIL=*",
    "C=DE,L=Nuremberg,O=SUSE Linux Products GmbH,OU=Documentation,CN=*" ]
```

使用以下命令获取证书的判别名：

```
> certtool -i --infile /etc/pki/libvirt/clientcert.pem | grep "Subject:"
```

更改配置后重启 `libvirtd`：

```
> sudo systemctl restart libvirtd
```

12.3.2.6 对 TLS 套接字使用 SASL 进行集中式用户身份验证

通过 TLS 无法进行直接的用户身份验证 — 只能按照第 12.3.2.5 节“限制访问（安全考虑因素）”中所述，通过证书读取权限在每个客户端上间接处理这种身份验证。但是，如果您需要采用基于服务器的集中式用户身份验证，`libvirt` 还允许在 TLS 的基础上使用 SASL（简单身份验证和安全层），来实现直接的用户身份验证。有关配置细节，请参见第 12.1.1.3 节“使用 SASL 进行用户名和口令身份验证”。

12.3.2.7 查错

12.3.2.7.1 虚拟机管理器/`virsh` 无法连接到服务器

按给定的顺序完成以下检查：

这是防火墙的问题吗（需要在服务器上打开 TCP 端口 16514）？

启动虚拟机管理器/`virsh` 的用户是否可以读取客户端证书（证书和密钥）？

是否在连接中指定了与服务器证书中相同的完全限定主机名？

是否在服务器上启用了 TLS (`listen_tls = 1`)？

是否在服务器上重启了 `libvirtd`？

12.3.2.7.2 VNC 连接失败

确保您可以使用虚拟机管理器连接到远程服务器。如果可以连接，请检查是否在启用了 TLS 支持的情况下启动了服务器上的虚拟机。以下示例中的虚拟机名称为 `sles`。

```
> ps ax | grep qemu | grep "\-name sles" | awk -F" -vnc " '{ print FS $2 }'
```

如果输出不是以类似于下面的字符串开头，则表示虚拟机启动时未启用 TLS 支持，必须将虚拟机重新启动。

```
-vnc 0.0.0.0:0,tls,x509verify=/etc/pki/libvirt
```

13 高级存储主题

本章介绍有关从 VM 主机服务器的角度操作存储设备的高级主题。

13.1 使用 virtlockd 锁定磁盘文件和块设备

锁定块设备和磁盘文件可以防止从不同的 VM Guest 并发向这些资源写入数据。它可以防范启动同一个 VM Guest 两次，或者将同一个磁盘添加到两个不同的虚拟机。这样就会减少由于配置错误导致虚拟机磁盘映像损坏的风险。

锁定操作由名为 `virtlockd` 的守护程序控制。由于此守护程序独立于 `libvirtd` 守护程序运行，在 `libvirtd` 崩溃或重新启动后，锁将会保留。甚至在更新 `virtlockd` 本身期间，锁也仍会保留，因为此守护程序能够自行重新执行。这可以确保当 `virtlockd` 更新后，**无需**重新启动 VM Guest。KVM、QEMU 和 Xen 支持 `virtlockd`。

13.1.1 启用锁定

SUSE Linux Enterprise Server 上默认未启用锁定虚拟磁盘功能。要启用锁定并在系统重引导时自动启动锁定，请执行以下步骤：

1. 编辑 `/etc/libvirt/qemu.conf` 并设置

```
lock_manager = "lockd"
```

2. 使用以下命令启动 `virtlockd` 守护程序：

```
> sudo systemctl start virtlockd
```

3. 使用以下命令重新启动 `libvirtd` 守护程序：

```
> sudo systemctl restart libvirtd
```

4. 确保引导系统时自动启动 `virtlockd`：

```
> sudo systemctl enable virtlockd
```

13.1.2 配置锁定

`virtlockd` 默认配置为自动锁定为 VM Guest 配置的所有磁盘。默认设置使用“直接”锁空间，在这种情况下，系统会根据与 VM Guest `<disk>` 设备关联的实际文件路径获取锁。例如，如果 VM Guest 包含以下 `<disk>` 设备，将直接针对 `/var/lib/libvirt/images/my-server/disk0.raw` 调用 `flock(2)`：

```
<disk type='file' device='disk'>
  <driver name='qemu' type='raw' />
  <source file='/var/lib/libvirt/images/my-server/disk0.raw' />
  <target dev='vda' bus='virtio' />
</disk>
```

可以通过编辑 `/etc/libvirt/qemu-lockd.conf` 文件来更改 `virtlockd` 配置。此文件还包含详细注释及其他信息。确保通过重新加载 `virtlockd` 来激活配置更改：

```
> sudo systemctl reload virtlockd
```

13.1.2.1 启用间接锁空间

`virtlockd` 的默认配置使用“直接”锁空间。这意味着，系统会根据与 `<disk>` 设备关联的实际文件路径来获取锁。

如果磁盘文件路径不可供所有主机访问，可将 `virtlockd` 配置为允许“间接”锁空间。这意味着，系统会使用磁盘映像路径的哈希在间接锁空间目录中创建一个文件。然后，将在这些哈希文件而不是实际的磁盘文件路径中存放锁。如果包含磁盘文件的文件系统不支持 `fcntl()` 锁，也可以使用间接锁空间。使用 `file_lockspace_dir` 设置指定间接锁空间：

```
file_lockspace_dir = "/MY_LOCKSPACE_DIRECTORY"
```

13.1.2.2 在 LVM 或 iSCSI 卷上启用锁定

如果您要锁定由多个主机共享的 LVM 或 iSCSI 卷上的虚拟磁盘，则需要按 UUID 而不是路径（默认使用路径）执行锁定。此外，需将锁空间目录放在可供共享该卷的所有主机访问的共享文件系统上。为 LVM 和/或 iSCSI 设置以下选项：

```
lvm_lockspace_dir = "/MY_LOCKSPACE_DIRECTORY"
iscsi_lockspace_dir = "/MY_LOCKSPACE_DIRECTORY"
```

13.2 联机调整 Guest 块设备的大小

有时，您需要更改（扩展或收缩）Guest 系统使用的块设备的大小。例如，当最初分配的磁盘空间不再足够时，便需要增大空间大小。如果 Guest 磁盘驻留在**逻辑卷**中，您可以在 Guest 系统运行时调整该磁盘的大小。与脱机调整磁盘大小相比（请参见 第 21.3 节 “Guestfs 工具” 软件包中的 **virt-resize** 命令），这是一项巨大的优势，因为 Guest 提供的服务在调整大小期间不会受到干扰。要调整 VM Guest 磁盘的大小，请执行以下步骤：

过程 13.1：联机调整 GUEST 磁盘的大小

1. 在 Guest 系统内部，检查磁盘（例如 `/dev/vda`）的当前大小。

```
# fdisk -l /dev/vda
Disk /dev/sda: 160.0 GB, 160041885696 bytes, 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
```

2. 在主机上，将容纳 Guest 磁盘 `/dev/vda` 的逻辑卷调整到所需大小，例如 200 GB。

```
# lvresize -L 200G /dev/mapper/vg00-home
Extending logical volume home to 200 GiB
Logical volume home successfully resized
```

3. 在主机上，调整与 Guest 磁盘 `/dev/mapper/vg00-home` 相关的块设备的大小。可以使用 **virsh list** 查找 `DOMAIN_ID`。

```
# virsh blockresize --path /dev/vg00/home --size 200G DOMAIN_ID
Block device '/dev/vg00/home' is resized
```

4. 检查 Guest 是否接受新磁盘大小。

```
# fdisk -l /dev/vda
```

```
Disk /dev/sda: 200.0 GB, 200052357120 bytes, 390727260 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
```

13.3 在主机与 Guest 之间共享目录（文件系统直通）

libvirt 允许使用 QEMU 的文件系统直通（也称为 VirtFS）功能在主机与 Guest 之间共享目录。此类目录还可由多个 VM Guest 同时访问，因此可用于在 VM Guest 之间交换文件。



注意：Windows Guest 和文件系统直通

无法通过文件系统直通在 VM 主机服务器与 Windows Guest 之间共享目录，因为 Windows 缺少挂载共享目录所需的驱动程序。

要使共享目录可在 VM Guest 上使用，请执行以下操作：

1. 在虚拟机管理器中打开 Guest 的控制台，然后从菜单中选择视图 > 细节，或者在工具栏中单击显示虚拟硬件详情。选择添加硬件 > 文件系统打开文件系统直通对话框。
2. 可以在驱动程序中选择句柄或路径模式的驱动程序。默认设置为路径。可以在模式中选择安全模型，这会影响在主机上设置文件权限的方式。有三个选项可用：

直通（默认设置）

使用客户端用户的身份凭证直接在文件系统上创建文件。这与 NFSv3 使用的设置相似。

Squash

与直通相同，但会忽略 **chown** 等特权操作的失败事件。当以 **root** 特权之外的身份运行 KVM 时，需要选择此选项。

映射

使用文件服务器的身份凭证 (**qemu.qemu**) 创建文件。用户身份凭证和客户端用户的身份凭证保存在扩展属性中。当主机和 Guest 域应该隔离时，建议使用此模型。

3. 使用源路径指定 VM 主机服务器上的目录的路径。在目标路径中输入一个字符串，作为挂载共享目录时使用的标记。此字段中的字符串仅作为标记，不是 VM Guest 上的路径。
4. 应用设置。如果 VM Guest 当前正在运行，需要将其关闭才能应用新设置（重引导 Guest 是不够的）。
5. 引导 VM Guest。要挂载共享目录，请输入以下命令：

```
> sudo mount -t 9p -o trans=virtio,version=9p2000.L,rw TAG /MOUNT_POINT
```

要使共享目录永久可用，请将下面一行添加到 `/etc/fstab` 文件中：

```
TAG /MOUNT_POINT 9p trans=virtio,version=9p2000.L,rw 0 0
```

13.4 通过 libvirt 使用 RADOS 块设备

RADOS 块设备 (RBD) 将数据存储在 Ceph 群集中。这些设备支持快照、复制和数据一致性。您可以像使用其他块设备一样，从 `libvirt` 管理的 VM Guest 使用 RBD。

有关更多细节，请参见 SUSE Enterprise Storage 《Administration Guide》中的“Using libvirt with Ceph”一章。<https://documentation.suse.com/ses/> 中提供了 SUSE Enterprise Storage 文档。

14 使用虚拟机管理器配置虚拟机

虚拟机管理器的细节视图提供有关 VM Guest 完整配置和硬件要求的详细信息。使用此视图还可以更改 Guest 配置，或者添加和修改虚拟硬件。要访问此视图，请在虚拟机管理器中打开 Guest 的控制台，然后从菜单中选择视图 > 细节，或者在工具栏中单击显示虚拟硬件详情。

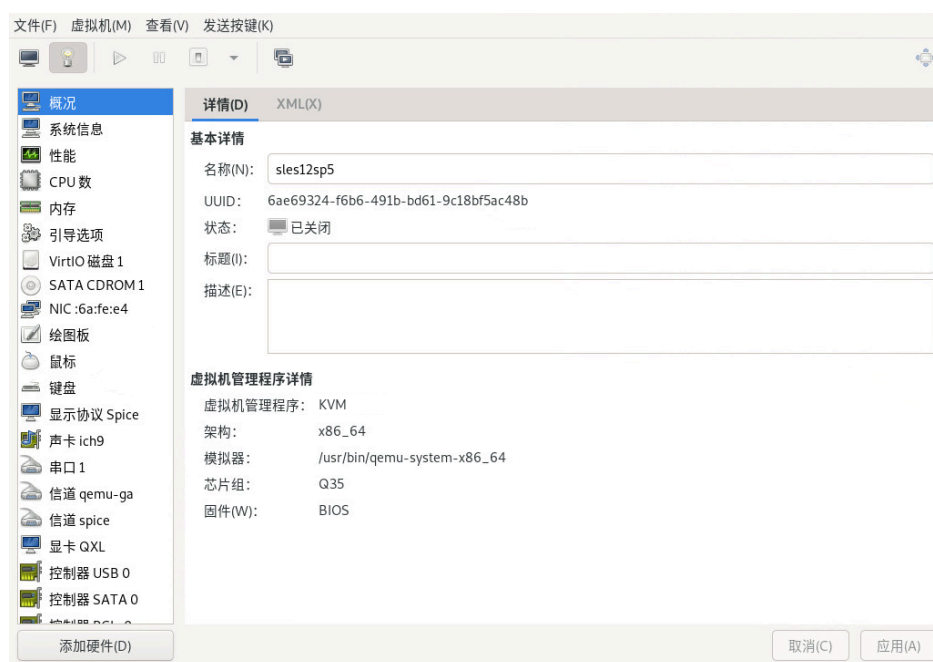


图 14.1：VM GUEST 的细节视图

窗口的左侧面板列出了 VM Guest 概述及已安装的硬件。在列表中单击某个项后，可以在细节视图中访问其详细设置。您可以根据需要更改硬件参数，然后单击应用确认更改。有些更改会立即生效，而有些更改需要重引导计算机，`virt-manager` 会预先通知您此情况。

要从 VM Guest 中去除安装的硬件，请在左侧面板中选择相应的列表项，然后单击窗口右下方的去除。

要添加新硬件，请单击左侧面板下方的添加硬件，然后在添加新虚拟硬件窗口中选择要添加的硬件类型。修改其参数并单击完成进行确认。

下列章节介绍要添加的特定硬件类型的配置选项。其中不会重点介绍如何修改现有的硬件，因为选项均相同。

14.1 计算机设置

本节介绍虚拟化处理器和内存硬件的设置。这些组件对于 VM Guest 至关重要，因此不能将其去除。本节还将介绍如何查看概览和性能信息，以及如何更改引导参数。

14.1.1 概述

概览显示有关 VM Guest 和超级管理程序的基本细节。

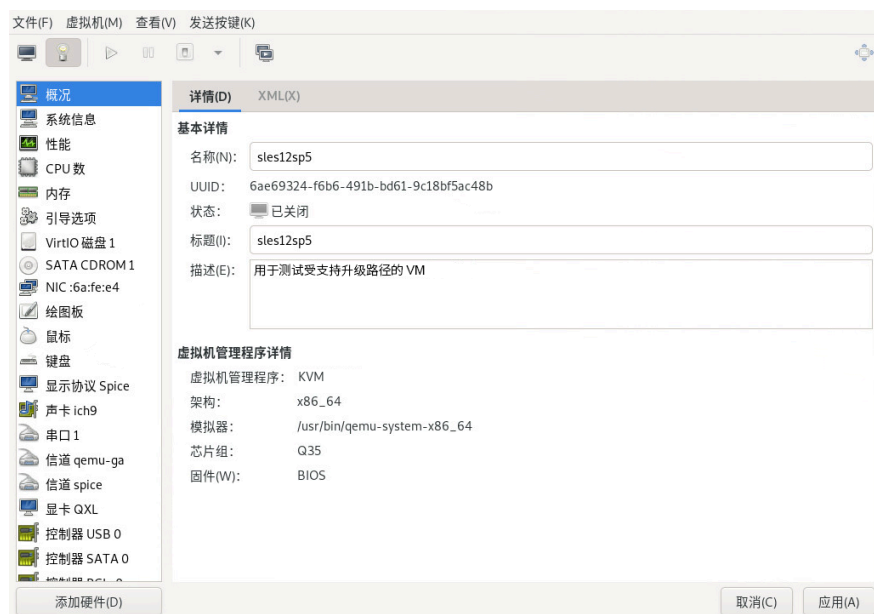


图 14.2：概览细节

名称、标题和说明均可编辑，有助于您在虚拟机管理器的计算机列表中识别 VM Guest。

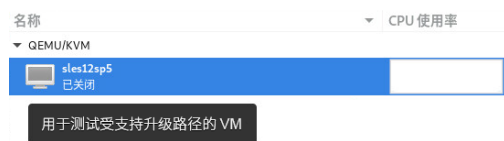


图 14.3：VM GUEST 标题和说明

UUID 显示虚拟机的全局唯一标识符，而状态显示虚拟机的当前状态 — 正在运行、已暂停或已关闭。

虚拟机管理程序详情部分显示超级管理程序类型、CPU 体系结构、使用的模拟器和芯片组类型。所有超级管理程序参数都不可更改。

14.1.2 性能

性能显示 CPU 使用率和内存用量以及磁盘和网络 I/O 的定期更新图表。

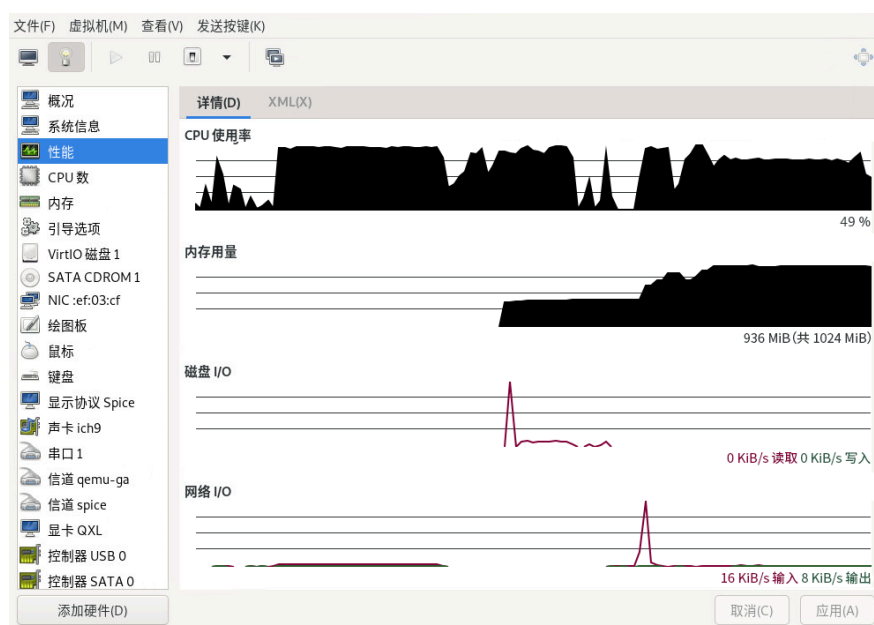


图 14.4：性能



提示：启用已禁用的图表

并非图表视图中的所有图表默认都已启用。要启用这些图表，请转到文件 > 视图管理器，然后选择编辑 > 首选项 > 轮询，检查您要查看的图表是否定期更新。



图 14.5：统计图

14.1.3 处理器

CPU 视图包含有关 VM Guest 处理器配置的详细信息。

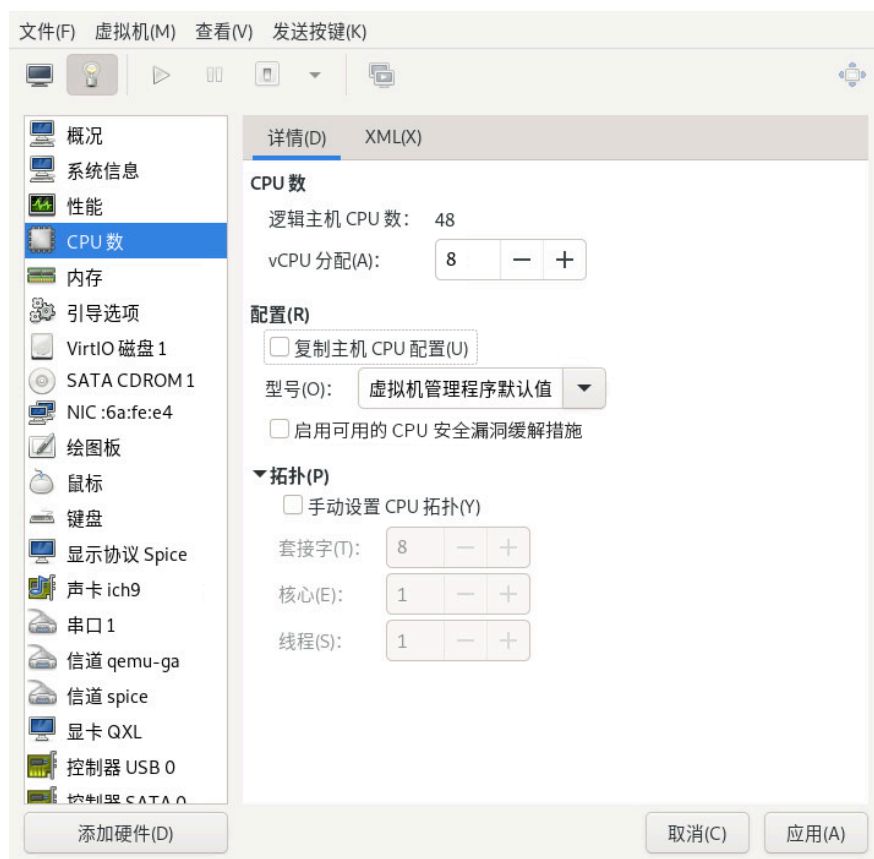


图 14.6：处理器视图

在 CPU 部分，可以配置分配给 VM Guest 的虚拟 CPU 数量。逻辑主机 CPU 显示 VM 主机服务器上的联机 CPU 和可用 CPU 数量。

配置部分可让您配置 CPU 型号和拓扑。

如果激活复制主机 CPU 配置选项，将为 VM Guest 使用主机 CPU 型号。可以在 **virsh capabilities** 命令的输出中查看主机 CPU 型号的 details。如果该 CPU 型号已停用，需要通过下拉框中的可用型号列表指定该型号。

主机 CPU 型号可在 CPU 功能和 VM Guest 迁移功能之间提供良好的平衡。**libvirt** 不会对每个 CPU 的每个方面建模，因此 VM Guest CPU 与 VM 主机服务器 CPU 并不完全匹配。但是，提供给 VM Guest 的 ABI 可重现，并且在迁移过程中，完整的 CPU 型号定义将传输到目标 VM 主机服务器，确保迁移的 VM Guest 可以在目标上看到完全相同的 CPU 型号。

`host-passthrough` 型号为 VM Guest 提供与 VM 主机服务器 CPU 完全相同的 CPU。当 `libvirt` 的简化 `host-model` CPU 不能提供 VM Guest 工作负载所需的 CPU 功能时，该型号可能很有用。`host-passthrough` 型号存在迁移能力下降的劣势。采用 `host-passthrough` 型号 CPU 的 VM Guest 只能迁移到具有相同硬件的 VM 主机服务器。

有关 `libvirt` 的 CPU 型号和拓扑选项的详细信息，请参见 <https://libvirt.org/formatdomain.html#cpu-model-and-topology> 上的《CPU model and topology》文档。激活手动设置 CPU 拓扑后，可以指定 CPU 的自定义插槽数、核心数和线程数。

14.1.4 内存

内存视图包含有关 VM Guest 可用内存的信息。

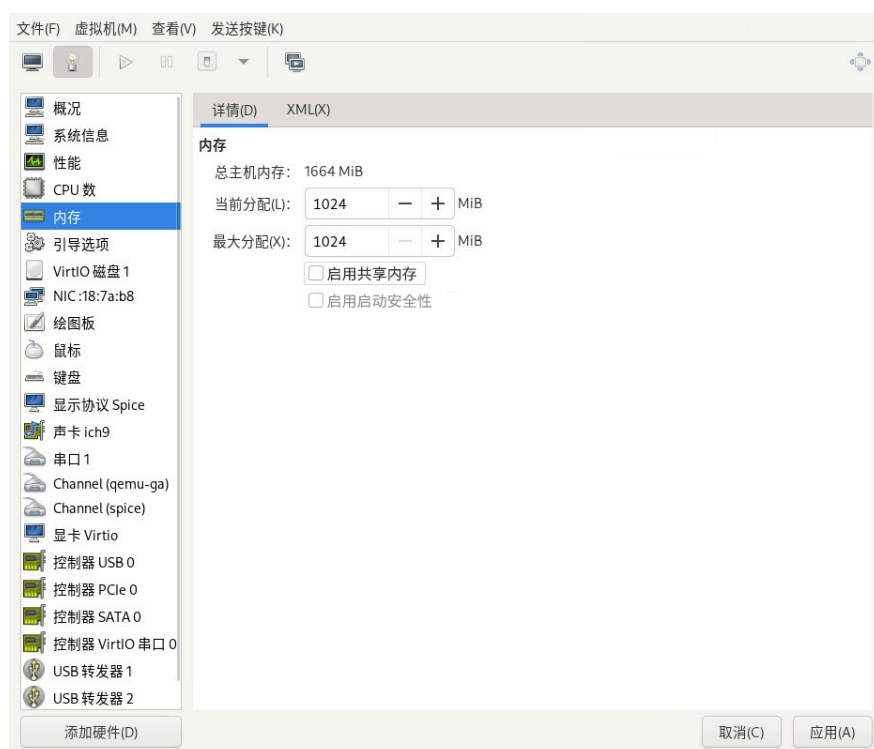


图 14.7：内存视图

主机内存总数

VM 主机服务器上安装的内存总量。

当前分配

当前可供 VM Guest 使用的内存量。可以通过增大此值（但不超过最大分配）来热插入更多内存。

启用共享内存

指定虚拟机是否可以通过基于 `memfd` 的内存使用共享内存。只有这样才能使用 `virtiofs` 文件系统。有关详细信息，请参见 <https://libvirt.org/kbase/virtiofs.html>。

最大分配

可热插入的当前可用内存的最大值。对此值所做的任何更改将在 VM Guest 下次重引导后生效。

启用启动安全性

如果 VM 主机服务器支持 AMD-SEV 技术，激活此选项可为受保护的 Guest 启用加密内存。此选项需要芯片组类型为 Q35 的虚拟机。有关详细信息，请参见《AMD 安全加密虚拟化 (AMD-SEV) 指南》文章。



重要：大内存 VM Guest

内存要求为 4 TB 或以上的 VM Guest 必须使用 `host-passthrough` CPU 模式，或者在使用 `host-model` 或 `custom` CPU 模式时明确指定虚拟 CPU 地址大小。这些模式的默认虚拟 CPU 地址大小可能不足以满足 4 TB 或以上的内存配置。只能通过编辑 VM Guest XML 配置来指定地址大小。有关指定虚拟 CPU 地址大小的详细信息，请参见第 15.6 节“配置内存分配”。

14.1.5 引导选项

引导选项引入了影响 VM Guest 引导进程的选项。

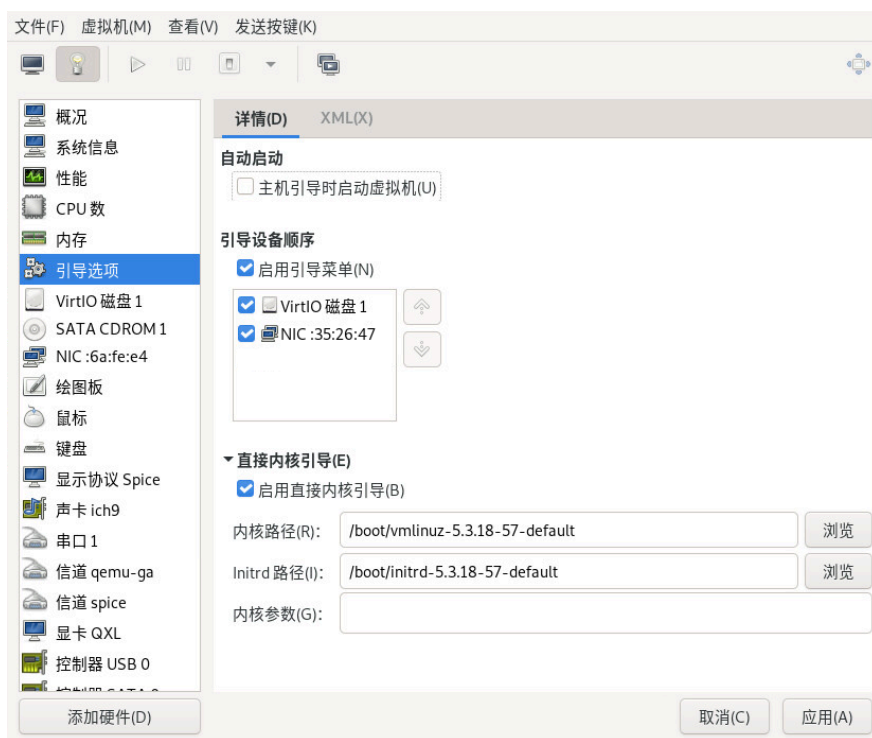


图 14.8：引导选项

在自动启动部分，可以指定是否应在 VM 主机服务器引导阶段自动启动虚拟机。

在引导设备顺序中，激活用于引导 VM Guest 的设备。可以使用列表右侧的向上和向下箭头按钮更改其顺序。要在 VM Guest 启动时从可引导设备列表中进行选择，请激活启用引导菜单。

要引导其他内核而不是引导设备上的内核，请激活启用直接内核引导，并指定替代内核以及位于 VM 主机服务器文件系统上的 initrd 的路径。您还可以指定要传递给所加载内核的内核参数。

14.2 存储

本节提供存储设备配置选项的详细说明。其中包括硬盘和可移动媒体，例如 USB 或 CD-ROM 驱动器。

过程 14.1：添加新存储设备

1. 在左侧面板下方，单击添加硬件打开添加新虚拟硬件窗口。在此窗口中选择存储。

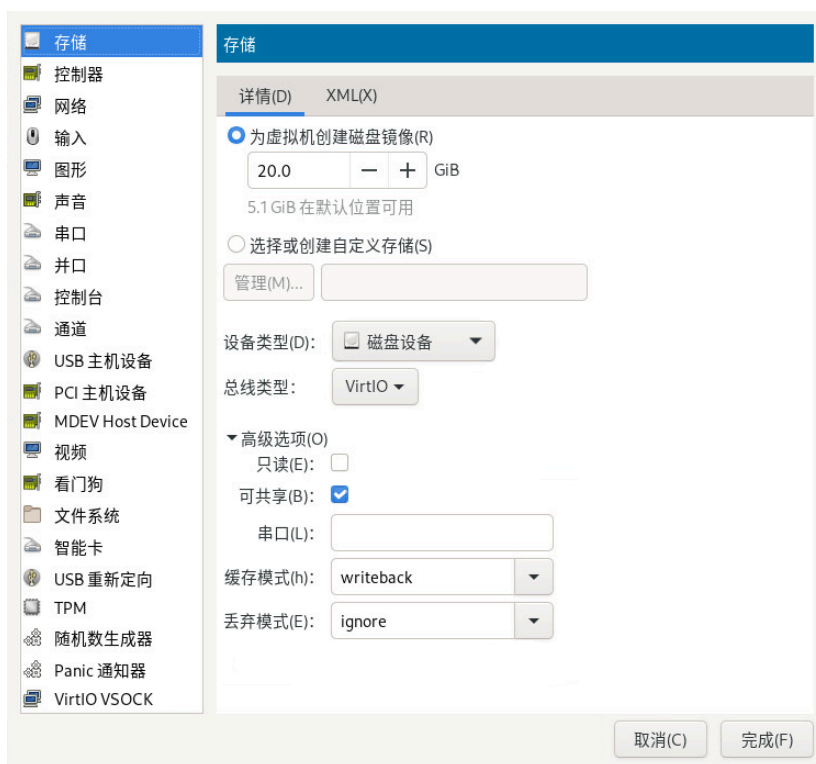


图 14.9：添加新存储设备

2. 要在默认位置创建 qcow2 磁盘映像，请激活创建虚拟机的磁盘映像，然后以 GB 为单位指定其大小。

要更细致地控制磁盘映像的创建，请激活选择或创建自定义存储，然后单击管理以管理存储池和映像。选择存储卷窗口即会打开，其中提供的功能与第 9.2.2 节“使用虚拟机管理器管理存储设备”中所述的存储选项卡基本相同。



提示：支持的存储格式

SUSE 仅支持以下存储格式：raw 和 qcow2。

3. 创建并指定磁盘映像文件后，指定设备类型。设备类型可为以下选项之一：
 - 磁盘设备
 - CDROM 设备：不允许使用创建虚拟机的磁盘映像。

- 软盘设备：不允许使用创建虚拟机的磁盘映像。
 - LUN 直连：如果想直接使用现有的 SCSI 存储而不将其添加到存储池，则需使用此选项。
4. 选择设备的总线类型。可用选项的列表取决于您在上一步中选择的设备类型。基于 VirtIO 的类型使用半虚拟化驱动程序。
 5. 在高级选项部分选择首选的缓存模式。有关缓存模式的详细信息，请参见第 19 章 “磁盘缓存模式”。
 6. 单击完成确认您的设置。新存储设备随即显示在左侧面板中。

14.3 控制器

本节重点介绍如何添加和配置新控制器。

过程 14.2：添加新控制器

1. 在左侧面板下方，单击添加硬件打开添加新虚拟硬件窗口。在此窗口中选择控制器。

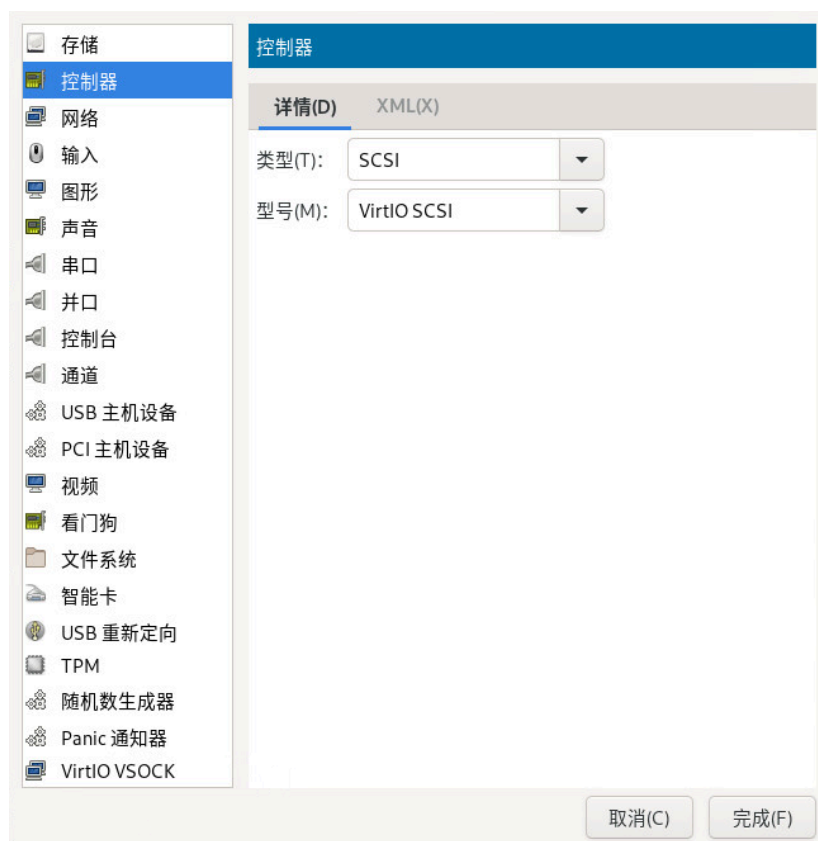


图 14.10：添加新控制器

2. 选择控制器的类型。可以选择 IDE、Floppy、SCSI、SATA、VirtIO Serial（半虚拟化）、USB 或 CCID（智能卡设备）。
3. （可选）如果选择了 USB 或 SCSI 控制器，请选择控制器型号。
4. 单击完成确认您的设置。新控制器随即显示在左侧面板中。

14.4 网络

本节介绍如何添加和配置新网络设备。

过程 14.3：添加新网络设备

1. 在左侧面板下方，单击添加硬件打开添加新虚拟硬件窗口。在此窗口中选择网络。

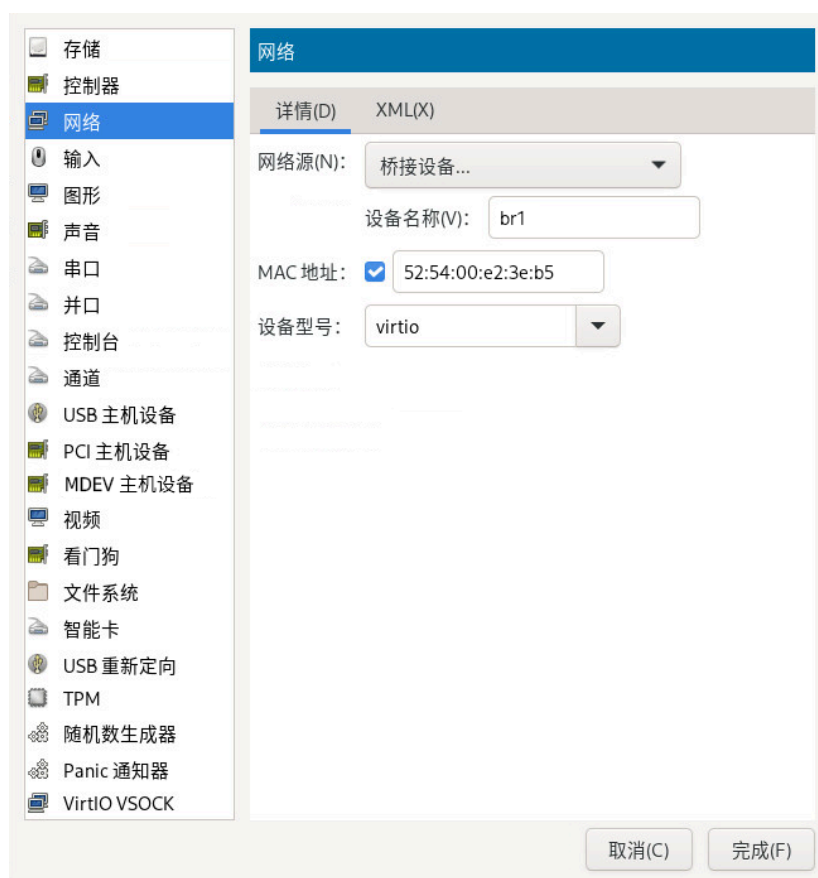


图 14.11：添加新网络接口

2. 在网络源列表中，选择网络连接的来源。该列表包含 VM 主机服务器的可用物理网络接口、网桥或网络绑定。您还可以将 VM Guest 分配到已定义的虚拟网络。有关使用虚拟机管理器设置虚拟网络的详细信息，请参见第 9.1 节“配置网络”。
3. 指定网络设备的 MAC 地址。尽管出于方便虚拟机管理器会预先填充一个随机值，但我们建议提供适合您网络环境的 MAC 地址，以免发生网络冲突。
4. 从列表中选择设备型号。可以保留虚拟机管理程序默认值，或者指定 e1000、rtl8139 或 virtio 型号中的一个。**virtio** 使用半虚拟化驱动程序。
5. 单击完成确认您的设置。新网络设备随即显示在左侧面板中。

14.5 输入设备

本节重点介绍如何添加和配置新输入设备，例如鼠标、键盘或绘图板。

过程 14.4：添加新输入设备

1. 在左侧面板下方，单击添加硬件打开添加新虚拟硬件窗口。在此窗口中选择输入。



图 14.12：添加新输入设备

2. 从列表中选择设备类型。
3. 单击完成确认您的设置。新输入设备随即显示在左侧面板中。



提示：启用无缝且同步的鼠标指针移动

在 VM Guest 的控制台中单击鼠标时，指针将由控制台窗口捕获，除非明确释放指针（按 **Alt + Ctrl**），否则无法在控制台外部使用指针。如果不想让控制台独占按键，而想在主机与 Guest 之间启用无缝指针移动，请按照[过程 14.4 “添加新输入设备”](#)中的说明将 EvTouch USB 图形数位板添加到 VM Guest。

添加绘图板的另一个好处是，在 Guest 上使用图形环境时可以同步 VM 主机服务器与 VM Guest 之间的鼠标指针移动。如果不在 Guest 上配置绘图板，您经常会看到两个有拖尾现象（一个指针拖在另一个指针后面）的指针。

14.6 视频

本节介绍如何添加和配置新视频设备。

过程 14.5：添加视频设备

1. 在左侧面板下方，单击添加硬件打开添加新虚拟硬件窗口。在此窗口中选择视频。

2.

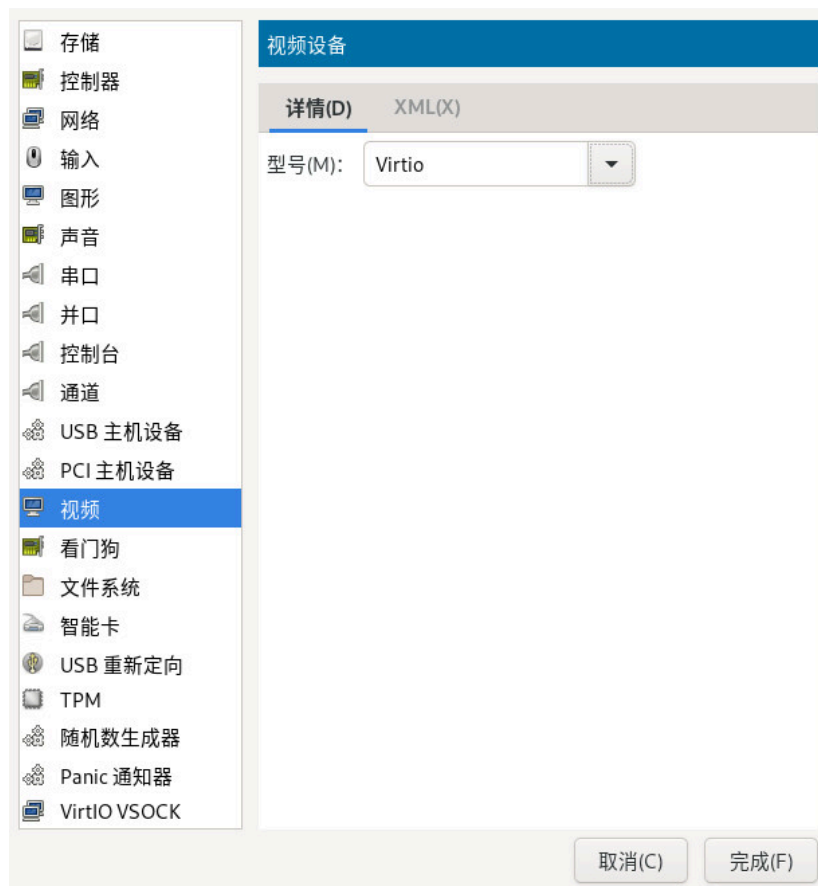


图 14.13：添加新视频设备

3. 从下拉框中选择一个型号。



注意：次要视频设备

只能将 QXL 和 Virtio 添加为次要视频设备。

4. 单击完成确认您的设置。新视频设备随即显示在左侧面板中。

14.7 USB 重定向器

可以使用 USB 重定向器将连接到客户端计算机的 USB 设备重定向到 VM Guest。

过程 14.6：添加 USB 重定向器

1. 在左侧面板下方，单击添加硬件打开添加新虚拟硬件窗口。在此窗口中选择 USB 重新定向。



图 14.14：添加新 USB 重定向器

2. 从列表中选择设备类型。根据您的配置，您可以选择 Spice 通道或 TCP 重定向器。
3. 单击完成确认您的设置。新 USB 重定向器随即显示在左侧面板中。

14.8 杂项

智能卡

可以通过 Smartcard 元素添加智能卡功能。然后，物理 USB 智能卡读卡器便可以直通到 VM Guest。

看门狗

系统还支持虚拟看门狗设备。可通过 Watchdog 元素创建这些设备。可以指定型号和设备操作。



提示：虚拟看门狗设备的要求

要使用 QA 虚拟看门狗设备，需要在 VM Guest 中安装特定的驱动程序和守护程序，否则虚拟看门狗设备将无法正常工作。

TPM

可以通过 TPM 元素添加 TPM 功能，以在 VM Guest 中使用主机 TPM 设备。



提示：虚拟 TPM

每次只能在一个 VM Guest 中使用主机 TPM。

14.9 使用虚拟机管理器添加 CD/DVD-ROM 设备

KVM 通过直接访问 VM 主机服务器上的物理驱动器或访问 ISO 映像来支持 VM Guest 中的 CD 和 DVD-ROM。要基于现有 CD 或 DVD 创建 ISO 映像，请使用 **dd**：

```
> sudo dd if=/dev/CD_DVD_DEVICE of=my_distro.iso bs=2048
```

要将 CD/DVD-ROM 设备添加到 VM Guest，请执行以下操作：

1. 双击虚拟机管理器中的某个 VM Guest 项打开其控制台，然后选择视图 > 细节切换到细节视图。
2. 单击添加硬件并在弹出窗口中选择存储。
3. 将设备类型更改为 IDE CDROM。
4. 选择选择或创建自定义存储。

- a. 要将设备分配到物理媒体，请在管理旁边输入 VM 主机服务器 CD/DVD-ROM 设备的路径（例如 `/dev/cdrom`）。或者，使用管理打开文件浏览器，然后单击本地浏览选择设备。仅当 VM 主机服务器上已启动虚拟机管理器时，才能将设备分配到物理媒体。
 - b. 要将设备分配到现有映像，请单击管理以从存储池中选择映像。如果 VM 主机服务器上已启动虚拟机管理器，您也可以单击本地浏览从文件系统上的另一个位置选择映像。选择某个映像并单击选择卷以关闭文件浏览器。
5. 单击完成以保存新虚拟化设备。
 6. 重引导 VM Guest 以使新设备可用。有关详细信息，请访问 [第 14.11 节 “使用虚拟机管理器弹出和更换软盘或 CD/DVD-ROM 媒体”](#)。

14.10 使用虚拟机管理器添加软盘设备

KVM 目前仅支持使用软盘映像 — 不支持使用物理软盘驱动器。使用 **dd** 基于现有软盘创建软盘映像：

```
> sudo dd if=/dev/fd0 of=/var/lib/libvirt/images/floppy.img
```

要创建空软盘映像，请使用以下命令之一：

Raw 映像

```
> sudo dd if=/dev/zero of=/var/lib/libvirt/images/floppy.img bs=512  
count=2880
```

FAT 格式映像

```
> sudo mkfs.msfdos -C /var/lib/libvirt/images/floppy.img 1440
```

要将软盘设备添加到 VM Guest，请执行以下操作：

1. 双击虚拟机管理器中的某个 VM Guest 项打开其控制台，然后选择视图 > 细节切换到细节视图。

2. 单击添加硬件并在弹出窗口中选择存储。
3. 将设备类型更改为软盘。
4. 选择选择或创建自定义存储，然后单击管理以从存储池中选择一个现有映像。如果 VM 主机服务器上已启动虚拟机管理器，您也可以单击本地浏览从文件系统上的另一个位置选择映像。选择某个映像并单击选择卷以关闭文件浏览器。
5. 单击完成以保存新虚拟化设备。
6. 重引导 VM Guest 以使新设备可用。有关详细信息，请访问 [第 14.11 节 “使用虚拟机管理器弹出和更换软盘或 CD/DVD-ROM 媒体”](#)。

14.11 使用虚拟机管理器弹出和更换软盘或 CD/DVD-ROM 媒体

无论您使用的是 VM 主机服务器的物理 CD/DVD-ROM 设备，还是 ISO/软盘映像，在更换 VM Guest 中现有设备的媒体或映像之前，都需要先将媒体与 Guest disconnect。

1. 双击虚拟机管理器中的某个 VM Guest 项打开其控制台，然后选择视图 > 细节切换到细节视图。
2. 选择软盘或 CD/DVD-ROM 设备，然后单击断开连接以“弹出”媒体。
3. 要“插入”新媒体，请单击连接。
 - a. 如果使用的是 VM 主机服务器的物理 CD/DVD-ROM 设备，请先更换设备中的媒体（这可能需要先在 VM 主机服务器上卸载该媒体，然后再将其弹出）。然后选择 CD-ROM 或 DVD，并从下拉框中选择设备。
 - b. 如果您使用的是 ISO 映像，请选择 ISO 映像位置，然后单击管理以选择映像。从远程主机连接时，只能选择现有存储池中的映像。
4. 单击确定以完成操作。现在便可在 VM Guest 中访问新媒体了。

14.12 将主机 PCI 设备分配到 VM Guest

可以直接将主机 PCI 设备分配到 Guest（PCI 直通）。将 PCI 设备分配到某个 VM Guest 后，除非重新分配，否则在主机上无法使用该设备，其他 VM Guest 也不能使用该设备。此功能的先决条件是 VM 主机服务器配置符合[重要：VFIO 和 SR-IOV 的要求](#)中所述的要求。

14.12.1 使用虚拟机管理器添加 PCI 设备

以下过程说明如何使用虚拟机管理器将主机计算机中的 PCI 设备分配到 VM Guest：

1. 双击虚拟机管理器中的某个 VM Guest 项打开其控制台，然后选择视图 > 细节切换到细节视图。
2. 单击添加硬件，然后在左侧面板中选择 PCI 主机设备类别。窗口右侧将显示可用 PCI 设备的列表。

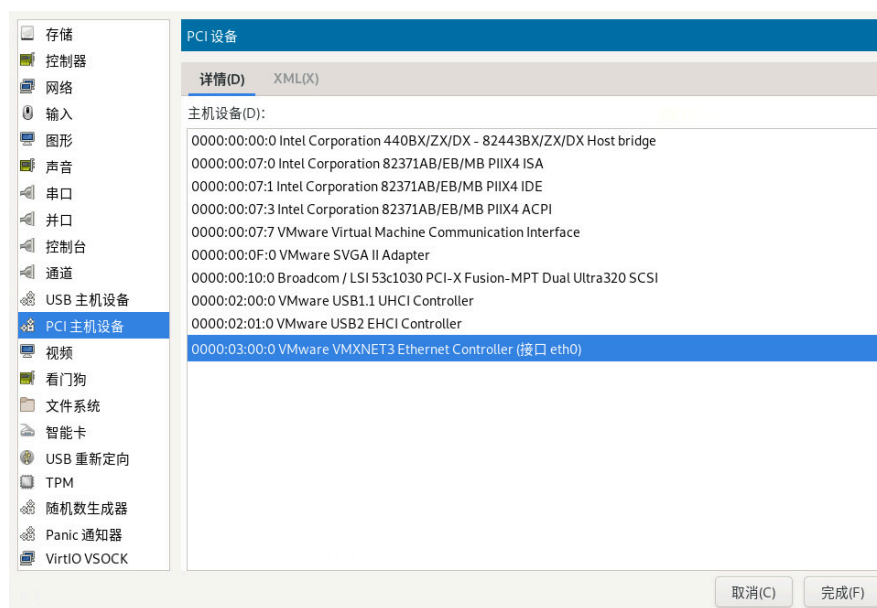


图 14.15：添加 PCI 设备

3. 在可用 PCI 设备的列表中，选择您要传递给 Guest 的设备。单击完成确认。

！ 重要：SLES 11 SP4 KVM Guest

在包含 SLES 11 SP4 KVM Guest 的新型 QEMU 计算机（pc-i440fx-2.0 或更高版本）上，默认不会在 Guest 中加载 `acpiphp` 模块。必须加载此模块才能启用磁盘和网络设备热插拔功能。要手动加载该模块，请使用 `modprobe acpiphp` 命令。也可以通过在 `/etc/modprobe.conf.local` 文件中添加 `install acpiphp /bin/true` 来自动加载该模块。

！ 重要：使用 QEMU Q35 计算机类型的 KVM Guest

使用 QEMU Q35 计算机类型的 KVM Guest 采用 PCI 拓扑，其中包含一个 `pcie-root` 控制器和七个 `pcie-root-port` 控制器。`pcie-root` 控制器不支持热插拔。每个 `pcie-root-port` 控制器支持热插拔一个 PCIe 设备。PCI 控制器无法热插拔，因此，如果要热插拔的 PCIe 设备超过七个，请做好相应规划并添加更多 `pcie-root-port` 控制器。可以添加一个 `pcie-to-pci-bridge` 控制器来支持热插拔旧式 PCI 设备。有关不同 QEMU 计算机类型的 PCI 拓扑的详细信息，请参见 <https://libvirt.org/pci-hotplug.html>。

14.13 将主机 USB 设备分配到 VM Guest

与分配主机 PCI 设备（请参见第 14.12 节“将主机 PCI 设备分配到 VM Guest”）类似，您可以直接将主机 USB 设备分配到 Guest。将 USB 设备分配到某个 VM Guest 后，除非重新分配，否则在主机上无法使用该设备，其他 VM Guest 也不能使用该设备。

14.13.1 使用虚拟机管理器添加 USB 设备

要使用虚拟机管理器将主机 USB 设备分配到 VM Guest，请执行以下步骤：

1. 双击虚拟机管理器中的某个 VM Guest 项打开其控制台，然后选择视图 > 细节切换到细节视图。
2. 单击添加硬件，然后在左侧面板中选择 USB 主机设备类别。窗口右侧将显示可用 USB 设备的列表。

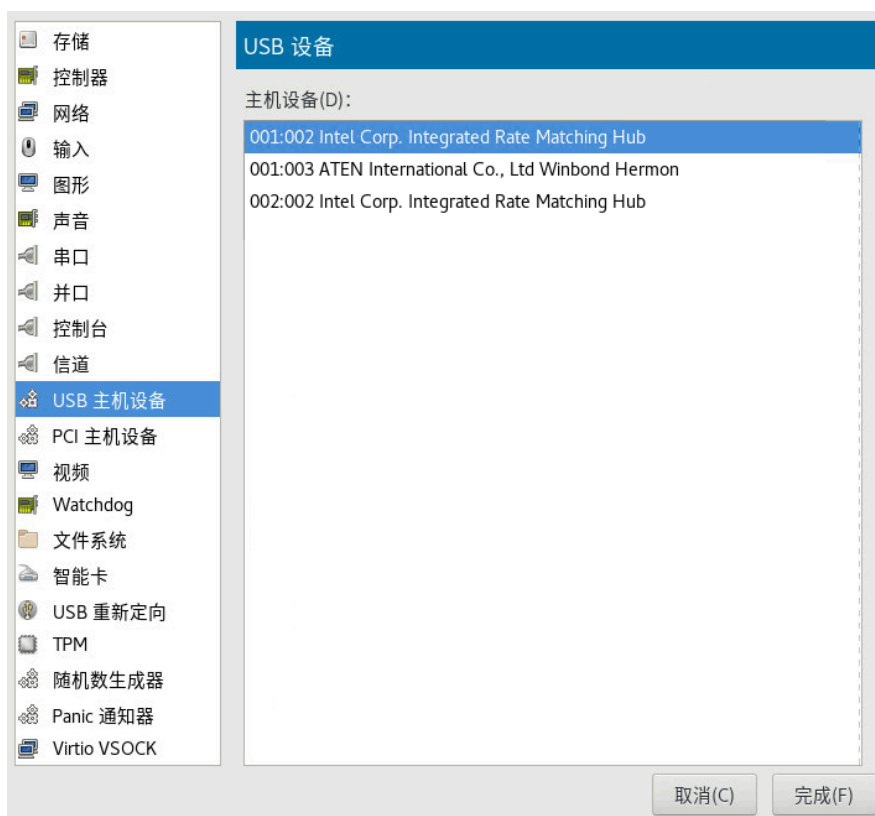


图 14.16：添加 USB 设备

3. 在可用 USB 设备的列表中，选择您要传递给 Guest 的设备。单击完成确认。新 USB 设备随即显示在细节视图的左侧窗格中。



提示：去除 USB 设备

要去除分配的主机 USB 设备，请在细节视图的左侧窗格中单击此设备，然后单击去除确认。

15 使用 **virsh** 配置虚拟机

您也可以在命令行上使用 **virsh** 来配置虚拟机 (VM)，以此替代虚拟机管理器。使用 **virsh** 可以控制 VM 的状态、编辑 VM 的配置，甚至将 VM 迁移到另一台主机。下列章节介绍如何使用 **virsh** 来管理 VM。

15.1 编辑 VM 配置

VM 配置存储在 `/etc/libvirt/qemu/` 中的某个 XML 文件内，其内容如下所示：

例 15.1：示例 XML 配置文件

```
<domain type='kvm'>
  <name>sles15</name>
  <uuid>ab953e2f-9d16-4955-bb43-1178230ee625</uuid>
  <memory unit='KiB'>2097152</memory>
  <currentMemory unit='KiB'>2097152</currentMemory>
  <vcpu placement='static'>2</vcpu>
  <os>
    <type arch='x86_64' machine='pc-q35-2.0'>hvm</type>
  </os>
  <features>...</features>
  <cpu mode='custom' match='exact' check='partial'>
    <model fallback='allow'>Skylake-Client-IBRS</model>
  </cpu>
  <clock>...</clock>
  <on_poweroff>destroy</on_poweroff>
  <on_reboot>restart</on_reboot>
  <on_crash>destroy</on_crash>
  <pm>
    <suspend-to-mem enabled='no' />
    <suspend-to-disk enabled='no' />
  </pm>
  <devices>
    <emulator>/usr/bin/qemu-system-x86_64</emulator>
```

```
<disk type='file' device='disk'>...</disk>
</devices>
...
</domain>
```

要编辑 VM Guest 的配置，请检查它是否处于脱机状态：

```
> sudo virsh list --inactive
```

如果您的 VM Guest 在此列表中，则表明您可以放心地编辑其配置：

```
> sudo virsh edit NAME_OF_VM_GUEST
```

在保存更改之前，**virsh** 会根据 RelaxNG 纲要验证您的输入。

15.2 更改计算机类型

使用 **virt-install** 工具安装时，VM Guest 的计算机类型默认为 **pc-q35**。计算机类型存储在 VM Guest 配置文件中的 **type** 元素内：

```
<type arch='x86_64' machine='pc-q35-2.3'>hvm</type>
```

以下过程示范了如何将此值更改为 **q35** 计算机类型。值 **q35** 表示一种 Intel* 芯片组，其中包括 **PCIe**，最多支持 12 个 USB 端口，并支持 **SATA** 和 **IOMMU**。

过程 15.1：更改计算机类型

1. 检查您的 VM Guest 是否处于非活动状态：

```
> sudo virsh list --inactive
Id      Name                                State
-----
-       sles15                             shut off
```

2. 编辑此 VM Guest 的配置：

```
> sudo virsh edit sles15
```

3. 请将 `machine` 属性的值替换为 `pc-q35-2.0`:

```
<type arch='x86_64' machine='pc-q35-2.0'>hvm</type>
```

4. 重新启动 VM Guest:

```
> sudo virsh start sles15
```

5. 检查计算机类型是否已更改。登录到 VM Guest 并运行以下命令:

```
> sudo dmidecode | grep Product
Product Name: Standard PC (Q35 + ICH9, 2009)
```



提示：计算机类型更新建议

每当升级主机系统上的 QEMU 版本时（例如，将 VM 主机服务器升级到新服务包时），请将 VM Guest 的计算机类型升级到最新的可用版本。要进行检查，请在 VM 主机服务器上使用 `qemu-system-x86_64 -M help` 命令。

默认计算机类型（例如 `pc-i440fx`）会定期更新。如果您的 VM Guest 仍在 `pc-i440fx-1.X` 计算机类型上运行，我们强烈建议更新到 `pc-i440fx-2.X`。这样就可以利用计算机定义中最近的更新和更正，并确保将来可以更好地兼容。

15.3 配置超级管理程序功能

`libvirt` 可自动启用一组默认的超级管理程序功能（这些功能在大多数情况下已够用），同时还允许按需启用和禁用功能。例如，Xen 不支持默认启用 PCI 直通。必须使用 `passthrough` 设置来启用此功能。可以使用 `virsh` 来配置超级管理程序功能。查看 VM Guest 配置文件中的 `<features>` 元素，并根据需要调整 VM Guest 功能。仍以 Xen 直通为例：

```
> sudo virsh edit sle15sp1
<features>
  <xen>
    <passthrough/>
  </xen>
```



```
</features>
```

保存更改并重新启动 VM Guest。

有关详细信息，请参见 <https://libvirt.org/formatdomain.html#elementsFeatures> 上 libvirt 的《Domain XML format》手册中的“Hypervisor features”一节。

15.4 配置 CPU

可以使用 **virsh** 来配置提供给 VM Guest 的虚拟 CPU 的许多属性。可以更改分配给 VM Guest 的当前和最大 CPU 数量，以及 CPU 型号及其功能集。以下小节介绍如何更改 VM Guest 的常用 CPU 设置。

15.4.1 配置 CPU 数量

分配的 CPU 数量存储在 `/etc/libvirt/qemu/` 下 VM Guest XML 配置文件中的 `vcpu` 元素内：

```
<vcpu placement='static'>1</vcpu>
```

在此示例中，只为 VM Guest 分配了一个 CPU。下面的过程说明如何更改分配给 VM Guest 的 CPU 数量：

1. 检查您的 VM Guest 是否处于非活动状态：

```
> sudo virsh list --inactive
```

Id	Name	State
-	sles15	shut off

2. 编辑现有 VM Guest 的配置：

```
> sudo virsh edit sles15
```

3. 更改分配的 CPU 数量：

```
<vcpu placement='static'>2</vcpu>
```

4. 重新启动 VM Guest:

```
> sudo virsh start sles15
```

5. 检查 VM 中的 CPU 数量是否已更改。

```
> sudo virsh vcpuinfo sled15
```

```
VCPU:          0
CPU:           N/A
State:         N/A
CPU time       N/A
CPU Affinity:  yy
```

```
VCPU:          1
CPU:           N/A
State:         N/A
CPU time       N/A
CPU Affinity:  yy
```

还可以在 VM Guest 正在运行时更改 CPU 数量。可以热插接 CPU，只要不超过 VM Guest 启动时配置的最大数量即可。同样，可以热拔除 CPU，只要不达到下限 1 即可。以下示例说明如何将活动 CPU 计数从 2 个更改为预定义的最大计数 4 个。

1. 检查当前的在线 vcpu 计数:

```
> sudo virsh vcpucount sles15 | grep live
```

```
maximum      live      4
current      live      2
```

2. 将当前或活动的 CPU 数量更改为 4 个:

```
> sudo virsh setvcpus sles15 --count 4 --live
```

3. 检查当前的在线 vcpu 计数现在是否为 4 个:

```
> sudo virsh vcpucount sles15 | grep live
```

```
maximum      live      4
```

15.4.2 配置 CPU 型号

向 VM Guest 公开的 CPU 型号往往会影晌该 VM Guest 中运行的工作负载。默认 CPU 型号派生自一种名为 `host-model` 的 CPU 模式。

```
<cpu mode='host-model' />
```

启动 CPU 模式为 `host-model` 的 VM Guest 时，`libvirt` 会将其主机 CPU 型号复制到 VM Guest 定义中。可以在 `virsh capabilities` 的输出中查看复制到 VM Guest 定义的主机 CPU 型号和功能。

另一种有趣的 CPU 模式是 `host-passthrough`。

```
<cpu mode='host-passthrough' />
```

启动 CPU 模式为 `host-passthrough` 的 VM Guest 时，将为该 VM Guest 提供与 VM 主机服务器 CPU 完全相同的 CPU。当 `libvirt` 的简化 `host-model` CPU 不能提供 VM Guest 工作负载所需的 CPU 功能时，该型号可能很有用。`host-passthrough` CPU 模式存在迁移能力下降的劣势。采用 `host-passthrough` CPU 模式的 VM Guest 只能迁移到具有相同硬件的 VM 主机服务器。

使用 `host-passthrough` CPU 模式时，仍可以禁用不需要的功能。以下配置将为 VM Guest 提供与主机 CPU 完全相同的 CPU，但会禁用 `vmx` 功能。

```
<cpu mode='host-passthrough'>
  <feature policy='disable' name='vmx' />
</cpu>
```

`custom` CPU 模式是另一种常用模式，用于定义可在群集中不同主机之间迁移的规范化 CPU。例如，在主机包含 Nehalem、IvyBridge 和 SandyBridge CPU 的群集中，可以使用包含 Nehalem CPU 型号的 `custom` CPU 模式来配置 VM Guest。

```
<cpu mode='custom' match='exact'>
  <model fallback='allow'>Nehalem</model>
  <feature policy='require' name='vme' />
  <feature policy='require' name='ds' />
</cpu>
```

```

<feature policy='require' name='acpi' />
<feature policy='require' name='ss' />
<feature policy='require' name='ht' />
<feature policy='require' name='tm' />
<feature policy='require' name='pbe' />
<feature policy='require' name='dtes64' />
<feature policy='require' name='monitor' />
<feature policy='require' name='ds_cpl' />
<feature policy='require' name='vmx' />
<feature policy='require' name='est' />
<feature policy='require' name='tm2' />
<feature policy='require' name='xtpr' />
<feature policy='require' name='pdcms' />
<feature policy='require' name='dca' />
<feature policy='require' name='rdtscp' />
<feature policy='require' name='invts' />
</cpu>

```

有关 [libvirt](https://libvirt.org/formatdomain.html#cpu-model-and-topology) 的 CPU 型号和拓扑选项的详细信息，请参见 <https://libvirt.org/formatdomain.html#cpu-model-and-topology> 上的《CPU model and topology》文档。

15.5 更改引导选项

可以在 `os` 元素中找到 VM Guest 的引导菜单，如以下示例所示：

```

<os>
  <type>hvm</type>
  <loader>readonly='yes' secure='no' type='rom'>/usr/lib/xen/boot/hvmloder</loader>
  <nvram template='/usr/share/OVMF/OVMF_VARS.fd'>/var/lib/libvirt/nvram/guest_VARS.fd</nvram>
  <boot dev='hd' />
  <boot dev='cdrom' />
  <bootmenu enable='yes' timeout='3000' />
  <smbios mode='sysinfo' />
  <bios useserial='yes' rebootTimeout='0' />
</os>

```

此示例中显示了两个设备：hd 和 cdrom。配置还反映了实际引导顺序，在示例中，hd 在 cdrom 之前引导。

15.5.1 更改引导顺序

VM Guest 的引导顺序通过 XML 配置文件中的设备顺序来表示。由于设备可以互换，因此可以更改 VM Guest 的引导顺序。

1. 打开 VM Guest 的 XML 配置。

```
> sudo virsh edit sles15
```

2. 更改可引导设备的顺序。

```
...  
<boot dev='cdrom' />  
<boot dev='hd' />  
...
```

3. 通过查看 VM Guest 的 BIOS 中的引导菜单来检查引导顺序是否已更改。

15.5.2 使用直接内核引导

使用直接内核引导可以从主机上存储的内核和 initrd 引导。您需要在 kernel 和 initrd 元素中设置这两个文件的路径：

```
<os>  
...  
<kernel>/root/f8-i386-vmlinuz</kernel>  
<initrd>/root/f8-i386-initrd</initrd>  
...  
</os>
```

要启用直接内核引导，请执行以下操作：

1. 打开 VM Guest 的 XML 配置：

```
> sudo virsh edit sles15
```

2. 在 `os` 元素内，添加一个 `kernel` 元素以及主机上内核文件的路径：

```
...  
<kernel>/root/f8-i386-vmlinuz</kernel>  
...
```

3. 添加 `initrd` 元素以及主机上 `initrd` 文件的路径：

```
...  
<initrd>/root/f8-i386-initrd</initrd>  
...
```

4. 启动 VM 以从新内核引导：

```
> sudo virsh start sles15
```

15.6 配置内存分配

您还可以使用 **virsh** 来配置分配给 VM Guest 的内存量。该配置存储在 `memory` 元素中，它定义了引导时为 VM Guest 分配的最大内存。可选的 `currentMemory` 元素定义分配给 VM Guest 的实际内存。`currentMemory` 可以小于 `memory`，这样，就可以在 VM Guest 运行时增加（或**扩大**）内存。如果省略 `currentMemory`，则其默认值与 `memory` 元素的值相同。

可以通过编辑 VM Guest 配置来调整内存设置，但请注意，更改只会在下次引导后生效。以下步骤说明如何将 VM Guest 更改为使用 4G 内存引导，但随后可以扩展到 8G：

1. 打开 VM Guest 的 XML 配置：

```
> sudo virsh edit sles15
```

2. 搜索 `memory` 元素并设置为 8G：

```
...  
<memory unit='KiB'>8388608</memory>  
...
```

3. 如果 `currentMemory` 元素不存在，请将其添加到 `memory` 元素下面，或将其值更改为 4G：

```
[...]
<memory unit='KiB'>8388608</memory>
<currentMemory unit='KiB'>4194304</currentMemory>
[...]
```

当 VM Guest 正在运行时，可以使用 `setmem` 子命令更改内存分配。以下示例显示如何将内存分配增加到 8G：

1. 检查 VM Guest 的现有内存设置：

```
> sudo virsh dominfo sles15 | grep memory
Max memory:      8388608 KiB
Used memory:     4194608 KiB
```

2. 将使用的内存更改为 8G：

```
> sudo virsh setmem sles15 8388608
```

3. 检查已更新的内存设置：

```
> sudo virsh dominfo sles15 | grep memory
Max memory:      8388608 KiB
Used memory:     8388608 KiB
```

! 重要：大内存 VM Guest

内存要求为 4 TB 或以上的 VM Guest 必须使用 `host-passthrough` CPU 模式，或者在使用 `host-model` 或 `custom` CPU 模式时明确指定虚拟 CPU 地址大小。默认的虚拟 CPU 地址大小可能不足以满足 4 TB 或以上的内存配置。以下示例说明了在使用 `host-model` CPU 模式时如何使用 VM 主机服务器的物理 CPU 地址大小。

```
[...]
<cpu mode='host-model' check='partial'>
<maxphysaddr mode='passthrough'>
</cpu>
```

[...]

有关指定虚拟 CPU 地址大小的详细信息，请参见 <https://libvirt.org/formatdomain.html#cpu-model-and-topology> 上的《CPU model and topology》文档中的 `maxphysaddr` 选项。

15.7 添加 PCI 设备

要使用 **virsh** 将 PCI 设备分配到 VM Guest，请执行以下步骤：

1. 标识要分配到 VM Guest 的主机 PCI 设备。在下面的示例中，我们要将一块 DEC 网卡分配到 Guest：

```
> sudo lspci -nn
[...]
03:07.0 Ethernet controller [0200]: Digital Equipment Corporation DECchip \
21140 [FasterNet] [1011:0009] (rev 22)
[...]
```

请记住设备 ID（在本例中为 `03:07.0`）。

2. 使用 **virsh nodedev-dumpxml ID** 收集有关设备的详细信息。要获取 ID，请将设备 ID (`03:07.0`) 中的冒号和句点替换为下划线，并在前面加上前缀 “`pci_0000_`”，例如：`pci_0000_03_07_0`。

```
> sudo virsh nodedev-dumpxml pci_0000_03_07_0
<device>
  <name>pci_0000_03_07_0</name>
  <path>/sys/devices/pci0000:00/0000:00:14.4/0000:03:07.0</path>
  <parent>pci_0000_00_14_4</parent>
  <driver>
    <name>tulip</name>
  </driver>
  <capability type='pci'>
    <domain>0</domain>
    <bus>3</bus>
```



```

<slot>7</slot>
<function>0</function>
<product id='0x0009'>DECchip 21140 [FasterNet]</product>
<vendor id='0x1011'>Digital Equipment Corporation</vendor>
<numa node='0' />
</capability>
</device>

```

记下域、总线和功能的值（请查看上面以粗体列显的 XML 代码）。

3. 从主机系统上分离设备，然后将其挂接到 VM Guest：

```

> sudo virsh nodedev-detach pci_0000_03_07_0
Device pci_0000_03_07_0 detached

```



提示：多功能 PCI 设备

使用不支持 FLR（功能级重置）或 PM（电源管理）重置的多功能 PCI 设备时，需从 VM 主机服务器分离其所有功能。出于安全原因，整个设备都必须重置。如果 VM 主机服务器或其他 VM Guest 仍在使用该设备的某个功能，[libvirt](#) 将拒绝分配该设备。

4. 将域、总线、插槽和功能值从十进制转换为十六进制。在本示例中，域 = 0，总线 = 3，插槽 = 7，功能 = 0。确保按正确顺序插入值：

```

> printf "<address domain='0x%x' bus='0x%x' slot='0x%x' function='0x%x' />
\n" 0 3 7 0

```

这会返回以下结果：

```

<address domain='0x0' bus='0x3' slot='0x7' function='0x0' />

```

5. 在您的域上运行 **virsh edit**，并使用上一步的结果在 **<devices>** 部分中添加以下设备项：

```

<hostdev mode='subsystem' type='pci' managed='yes'>
  <source>
    <address domain='0x0' bus='0x03' slot='0x07' function='0x0' />
  
```

```
</source>
</hostdev>
```



提示：managed与unmanaged模式的比较

`libvirt` 可识别 PCI 设备的两种处理模式：`managed`或`unmanaged`。在受管模式下，`libvirt` 将处理从现有驱动程序解除绑定设备（如果需要）、重置设备、在启动域之前将设备绑定到 `vfio-pci` 等事项的所有细节。当域终止或设备从域中移除时，`libvirt` 会从 `vfio-pci` 解除绑定，并在使用受管设备时重新绑定到原始驱动程序。如果设备不受管，则用户必须确保在将设备分配到域之前以及在设备不再由域使用之后，所有这些设备管理方面的操作都已完成。

在上面的示例中，`managed='yes'` 选项表示设备是受管的。要将设备切换为不受管模式，请在上面的列表中设置 `managed='no'`。如果这样做，需使用 `virsh nodedev-detach` 和 `virsh nodedev-reattach` 命令处理相关的驱动程序。在启动 VM Guest 之前，需运行 `virsh nodedev-detach pci_0000_03_07_0` 以从主机分离设备。如果 VM Guest 未运行，您可以运行 `virsh nodedev-reattach pci_0000_03_07_0`，使设备可供主机使用。

6. 关闭 VM Guest，并禁用 SELinux（如果它正在主机上运行）。

```
> sudo setsebool -P virt_use_sysfs 1
```

7. 启动 VM Guest 以使分配的 PCI 设备可用：

```
> sudo virsh start sles15
```



重要：SLES11 SP4 KVM Guest

在包含 SLES 11 SP4 KVM Guest 的新型 QEMU 计算机（`pc-i440fx-2.0` 或更高版本）上，默认不会在 Guest 中加载 `acpiphp` 模块。必须加载此模块才能启用磁盘和网络设备热插拔功能。要手动加载该模块，请使用 `modprobe acpiphp` 命令。也可以通过在 `/etc/modprobe.conf.local` 文件中添加 `install acpiphp /bin/true` 来自动加载该模块。

！ 重要：使用 QEMU Q35 计算机类型的 KVM Guest

使用 QEMU Q35 计算机类型的 KVM Guest 采用 PCI 拓扑，其中包含一个 `pcie-root` 控制器和七个 `pcie-root-port` 控制器。`pcie-root` 控制器不支持热插拔。每个 `pcie-root-port` 控制器支持热插拔一个 PCIe 设备。PCI 控制器无法热插拔，因此，如果要热插拔的 PCIe 设备超过七个，请做好相应规划并添加更多 `pcie-root-port` 控制器。可以添加一个 `pcie-to-pci-bridge` 控制器来支持热插拔旧式 PCI 设备。有关不同 QEMU 计算机类型的 PCI 拓扑的详细信息，请参见 <https://libvirt.org/pci-hotplug.html>。

15.7.1 IBM Z 的 PCI 直通

为了支持 IBM Z，QEMU 扩展了 PCI 表示形式，现在它允许用户配置额外的属性。`<zpci/>` `libvirt` 规范中额外添加了两个属性 — `uid` 和 `fid`。`uid` 表示用户定义的标识符，`fid` 表示 PCI 功能标识符。这些属性是可选的，如果不指定，系统将使用不冲突的值自动生成这些属性。要在域规范中包含 zPCI 属性，请使用以下示例定义：

```
<controller type='pci' index='0' model='pci-root'/>
<controller type='pci' index='1' model='pci-bridge'>
  <model name='pci-bridge'/>
  <target chassisNr='1'/>
  <address type='pci' domain='0x0000' bus='0x00' slot='0x01' function='0x0'>
    <zpci uid='0x0001' fid='0x00000000'/>
  </address>
</controller>
<interface type='bridge'>
  <source bridge='virbr0'/>
  <model type='virtio'/>
  <address type='pci' domain='0x0000' bus='0x01' slot='0x01' function='0x0'>
    <zpci uid='0x0007' fid='0x00000003'/>
  </address>
</interface>
```

15.8 添加 USB 设备

要使用 **virsh** 将 USB 设备分配到 VM Guest，请执行以下步骤：

1. 标识要分配到 VM Guest 的主机 USB 设备：

```
> sudo lsusb
[...]
Bus 001 Device 003: ID 0557:2221 ATEN International Co., Ltd Winbond Hermon
[...]
```

记下供应商 ID 和产品 ID。在本示例中，供应商 ID 为 0557，产品 ID 为 2221。

2. 在您的域上运行 **virsh edit**，并使用上一步的值在 `<devices>` 部分中添加以下设备项：

```
<hostdev mode='subsystem' type='usb'>
  <source startupPolicy='optional'>
    <vendor id='0557' />
    <product id='2221' />
  </source>
</hostdev>
```



提示：供应商/产品或设备地址

如果不使用 `<vendor/>` 和 `<product/>` ID 定义主机设备，您可以根据第 15.7 节“添加 PCI 设备”中所述的适用于主机 PCI 设备的操作使用 `<address/>` 元素。

3. 关闭 VM Guest，并禁用 SELinux（如果它正在主机上运行）：

```
> sudo setsebool -P virt_use_sysfs 1
```

4. 启动 VM Guest 以使分配的 PCI 设备可用：

```
> sudo virsh start sles15
```

15.9 添加 SR-IOV 设备

支持单根 I/O 虚拟化 (SR-IOV) 的 PCIe 设备能够复制其资源，因此它们看上去像是多个设备。每个“伪设备”都可以分配给 VM Guest。

SR-IOV 是外围部件互连专业组 (PCI-SIG) 联盟制定的行业规范。其中介绍了物理功能 (PF) 和虚拟功能 (VF)。PF 是用于管理和配置设备的完整 PCIe 功能。PF 还可以移动数据。VF 在配置和管理方面的作用有所欠缺 — 它们只能移动数据，提供的配置功能有限。由于 VF 不包括所有的 PCIe 功能，主机操作系统或超级管理程序必须支持 SR-IOV 才能访问和初始化 VF。理论上 VF 的最大数量为每台设备 256 个（因此，对于双端口以太网卡，最大数量为 512 个）。在实际环境中，此最大数量要少得多，因为每个 VF 都会消耗资源。

15.9.1 要求

要使用 SR-IOV，必须符合以下要求：

- 支持 SR-IOV 的网卡（从 SUSE Linux Enterprise Server 15 开始，只有网卡支持 SR-IOV）
- 支持硬件虚拟化的 AMD64/Intel 64 主机（AMD-V 或 Intel VT-x），有关详细信息，请参见第 7.1.1 节“KVM 硬件要求”
- 支持设备分配的芯片组（AMD-Vi 或 Intel VT-d）
- libvirt 0.9.10 或更高版本
- 主机系统上必须加载并配置 SR-IOV 驱动程序
- 符合重要：VFIO 和 SR-IOV 的要求中所列要求的主机配置
- 要分配到 VM Guest 的 VF 的 PCI 地址列表



提示：检查设备是否支持 SR-IOV

可以通过运行 **lspci** 从设备的 PCI 描述符中获取有关该设备是否支持 SR-IOV 的信息。支持 SR-IOV 的设备会报告类似如下的功能：

```
Capabilities: [160 v1] Single Root I/O Virtualization (SR-IOV)
```



注意：在创建 VM Guest 时添加 SR-IOV 设备

您必须已按照第 15.9.2 节“加载和配置 SR-IOV 主机驱动程序”中所述配置 VM 主机服务，才可在最初设置 VM Guest 时向其添加 SR-IOV 设备。

15.9.2 加载和配置 SR-IOV 主机驱动程序

要访问和初始化 VF，需在主机系统上加载一个支持 SR-IOV 的驱动程序。

1. 在加载驱动程序之前，请运行 **lspci** 来确保可正常检测到网卡。以下示例显示了双端口 Intel 82576NS 网卡的 **lspci** 输出：

```
> sudo /sbin/lspci | grep 82576
01:00.0 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
01:00.1 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
04:00.0 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
04:00.1 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
```

如果未检测到网卡，可能是因为未在 BIOS/EFI 中启用硬件虚拟化支持。要检查是否已启用硬件虚拟化支持，请查看主机 BIOS 中的设置。

2. 运行 **lsmod** 来检查是否已加载 SR-IOV 驱动程序。在以下示例中，用于检查是否加载了 Intel 82576NS 网卡的 igb 驱动程序的命令返回了一条结果。这表示已加载该驱动程序。如果该命令未返回任何结果，则表示未加载该驱动程序。

```
> sudo /sbin/lsmod | egrep "^igb "
igb                  185649  0
```

3. 如果已加载驱动程序，请跳过以下步骤。如果尚未加载 SR-IOV 驱动程序，需要先去除非 SR-IOV 驱动程序，然后再加载新驱动程序。使用 **rmmmod** 卸载驱动程序。下面的示例会卸载 Intel 82576NS 网卡的非 SR-IOV 驱动程序：

```
> sudo /sbin/rmmod igbvf
```

4. 随后使用 **modprobe** 命令加载 SR-IOV 驱动程序 — 必须指定 VF 参数 (max_vfs):

```
> sudo /sbin/modprobe igb max_vfs=8
```

或者, 您也可以通过 SYSFS 加载驱动程序:

1. 通过列出以太网设备确定物理 NIC 的 PCI ID:

```
> sudo lspci | grep Eth
06:00.0 Ethernet controller: Emulex Corporation OneConnect NIC (Skyhawk)
(rev 10)
06:00.1 Ethernet controller: Emulex Corporation OneConnect NIC (Skyhawk)
(rev 10)
```

2. 要启用 VF, 请向 sriov_numvfs 参数回送需要加载的 VF 数量:

```
> sudo echo 1 > /sys/bus/pci/devices/0000:06:00.1/sriov_numvfs
```

3. 校验是否已加载 VF NIC:

```
> sudo lspci | grep Eth
06:00.0 Ethernet controller: Emulex Corporation OneConnect NIC (Skyhawk)
(rev 10)
06:00.1 Ethernet controller: Emulex Corporation OneConnect NIC (Skyhawk)
(rev 10)
06:08.0 Ethernet controller: Emulex Corporation OneConnect NIC (Skyhawk)
(rev 10)
```

4. 获取可用 VF 的最大数量:

```
> sudo lspci -vvv -s 06:00.1 | grep 'Initial VFs'
Initial VFs: 32, Total VFs: 32, Number of VFs: 0,
Function Dependency Link: 01
```

5. 创建 /etc/systemd/system/before.service 文件, 用于在引导时通过 SYSFS 加载 VF:

```
[Unit]
Before=
[Service]
Type=oneshot
RemainAfterExit=true
ExecStart=/bin/bash -c "echo 1 > /sys/bus/pci/devices/0000:06:00.1/
sriov_numvfs"
# beware, executable is run directly, not through a shell, check the man
pages
# systemd.service and systemd.unit for full syntax
[Install]
# target in which to start the service
WantedBy=multi-user.target
#WantedBy=graphical.target
```

6. 在启动 VM 之前，需要创建指向 /etc/init.d/after.local 脚本（用于分离 NIC）的另一个服务文件 (after-local.service)。否则 VM 将无法启动：

```
[Unit]
Description=/etc/init.d/after.local Compatibility
After=libvirtd.service
Requires=libvirtd.service
[Service]
Type=oneshot
ExecStart=/etc/init.d/after.local
RemainAfterExit=true

[Install]
WantedBy=multi-user.target
```

7. 将此文件复制到 /etc/systemd/system。

```
#!/bin/sh
# ...
virsh nodedev-detach pci_0000_06_08_0
```

将此文件另存为 /etc/init.d/after.local。

8. 重引导计算机，然后按照本过程的第一步重新运行 **lspci** 命令，以检查是否已加载 SR-IOV 驱动程序。如果已成功加载 SR-IOV 驱动程序，您应该会看到额外的 VF 行：

```
01:00.0 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
01:00.1 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
01:10.0 Ethernet controller: Intel Corporation 82576 Virtual Function (rev
01)
01:10.1 Ethernet controller: Intel Corporation 82576 Virtual Function (rev
01)
01:10.2 Ethernet controller: Intel Corporation 82576 Virtual Function (rev
01)
[...]
04:00.0 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
04:00.1 Ethernet controller: Intel Corporation 82576NS Gigabit Network
Connection (rev 01)
04:10.0 Ethernet controller: Intel Corporation 82576 Virtual Function (rev
01)
04:10.1 Ethernet controller: Intel Corporation 82576 Virtual Function (rev
01)
04:10.2 Ethernet controller: Intel Corporation 82576 Virtual Function (rev
01)
[...]
```

15.9.3 将 VF 网络设备添加到 VM Guest

在 VM 主机服务器上正确设置 SR-IOV 硬件后，便可将 VF 添加到 VM Guest。为此，需要先收集特定的数据。

过程 15.2：将 VF 网络设备添加到现有 VM GUEST

下面的过程使用的是示例数据。请将其替换为您设置中的相应数据。

1. 使用 **virsh nodedev-list** 命令获取您要分配的 VF 的 PCI 地址及其对应的 PF。第 15.9.2 节 “加载和配置 SR-IOV 主机驱动程序” 中所示的 **lspci** 输出中的数字值（例如 01:00.0 或 04:00.1）已经过转换：添加了前缀 **pci_0000_** 并将冒号和句点替换为下划线。因此，**lspci** 列出的 PCI ID 04:00.0 会被 **virsh** 列为 **pci_0000_04_00_0**。下面的示例列出了 Intel 82576NS 网卡的第二个端口的 PCI ID：

```
> sudo virsh nodedev-list | grep 0000_04_  
pci_0000_04_00_0  
pci_0000_04_00_1  
pci_0000_04_10_0  
pci_0000_04_10_1  
pci_0000_04_10_2  
pci_0000_04_10_3  
pci_0000_04_10_4  
pci_0000_04_10_5  
pci_0000_04_10_6  
pci_0000_04_10_7  
pci_0000_04_11_0  
pci_0000_04_11_1  
pci_0000_04_11_2  
pci_0000_04_11_3  
pci_0000_04_11_4  
pci_0000_04_11_5
```

前两项表示 **PF**，其他项表示 **VF**。

2. 对您要添加的 VF 的 PCI ID 运行以下 **virsh nodedev-dumpxml** 命令：

```
> sudo virsh nodedev-dumpxml pci_0000_04_10_0  
<device>  
  <name>pci_0000_04_10_0</name>  
  <parent>pci_0000_00_02_0</parent>  
  <capability type='pci'>  
    <domain>0</domain>  
    <bus>4</bus>  
    <slot>16</slot>  
    <function>0</function>  
    <product id='0x10ca'>82576 Virtual Function</product>
```

```

<vendor id='0x8086'>Intel Corporation</vendor>
<capability type='phys_function'>
  <address domain='0x0000' bus='0x04' slot='0x00' function='0x0' />
</capability>
</device>

```

下一步需要以下数据：

- <domain>0</domain>
- <bus>4</bus>
- <slot>16</slot>
- <function>0</function>

3. 创建一个临时 XML 文件（例如 `/tmp/vf-interface.xml`），其中包含将 VF 网络设备添加到现有 VM Guest 所需的数据。该文件至少需包含如下所示的内容：

```

<interface type='hostdev'>❶
  <source>
    <address type='pci' domain='0' bus='11' slot='16' function='0'2/>❷
  </source>
</interface>

```

- ❶ VF 的 MAC 地址不固定；每次重引导主机后，MAC 地址都会改变。如果使用 `hostdev` 以“传统”方式添加网络设备，每次重引导主机后都需要重新配置 VM Guest 的网络设备，因为 MAC 地址会改变。为避免出现这种问题，`libvirt` 引入了 `hostdev` 值用于在分配设备**之前**设置网络特定的数据。
 - ❷ 请在此处指定上一步中获取的数据。
4. 如果设备已挂接到主机，则无法将它挂接到 VM Guest。要使该设备可供 Guest 使用，请先将它从主机分离：

```
> sudo virsh nodedev-detach pci_0000_04_10_0
```

5. 将 VF 接口添加到现有 VM Guest：

```
> sudo virsh attach-device GUEST /tmp/vf-interface.xml --OPTION
```

需将 GUEST 替换为 VM Guest 的域名、ID 或 UUID。--OPTION 可为下列其中一项：

--persistent

此选项始终将设备添加到域的永久性 XML 中。如果域正在运行，则会热插入设备。

--config

此选项只影响永久性 XML，即使域正在运行也是如此。设备在下次引导时才会显示在 VM Guest 中。

--live

此选项只影响正在运行的域。如果域处于非活动状态，则操作将会失败。设备不会永久保留在 XML 中，下次引导时会在 VM Guest 中变为可用状态。

--current

此选项影响域的当前状态。如果域处于非活动状态，设备将添加到永久性 XML 中，并会在下次引导时变为可用状态。如果域处于活动状态，则会热插入设备，但不会将其添加到永久性 XML 中。

6. 要分离 VF 接口，请使用 **virsh detach-device** 命令，该命令也接受上面所列的选项。

15.9.4 动态分配池中的 VF

如果您按照第 15.9.3 节 “将 VF 网络设备添加到 VM Guest” 中所述以静态方式在 VM Guest 的配置中定义了 VF 的 PCI 地址，此类 Guest 将很难迁移到另一台主机。该主机必须在 PCI 总线上的相同位置具有相同的硬件，否则每次启动之前都必须修改 VM Guest 配置。

另一种方法是使用一个包含 SR-IOV 设备所有 VF 的设备池创建 libvirt 网络。之后，VM Guest 将引用此网络，每次 VM Guest 启动时，系统都会向它动态分配单个 VF。当 VM Guest 停止时，该 VF 将返回到池中，可供其他 Guest 使用。

15.9.4.1 在 VM 主机服务器上使用 VF 池定义网络

以下网络定义示例为 SR-IOV 设备创建了一个包含所有 VF 的池，该设备的物理功能 (PF) 位于主机中的网络接口 `eth0` 上：

```
<network>
  <name>passthrough</name>
  <forward mode='hostdev' managed='yes'>
    <pf dev='eth0' />
  </forward>
</network>
```

要在主机上使用此网络，请将上述代码保存到文件（例如 `/tmp/passthrough.xml`）中，然后执行以下命令。请记得将 `eth0` 替换为 SR-IOV 设备的 PF 的实际网络接口名称：

```
> sudo virsh net-define /tmp/passthrough.xml
> sudo virsh net-autostart passthrough
> sudo virsh net-start passthrough
```

15.9.4.2 将 VM Guest 配置为使用池中的 VF

下面的 VM Guest 设备接口定义示例使用了 SR-IOV 设备的一个 VF，该 VF 来自第 15.9.4.1 节“在 VM 主机服务器上使用 VF 池定义网络”中创建的池。Guest 首次启动时，`libvirt` 会自动派生与该 PF 关联的所有 VF 的列表。

```
<interface type='network'>
  <source network='passthrough'>
</interface>
```

在第一个使用以 VF 池定义的网络的 VM Guest 启动后，校验关联的 VF 列表。为此，请在主机上运行 `virsh net-dumpxml passthrough`。

```
<network connections='1'>
  <name>passthrough</name>
  <uuid>a6a26429-d483-d4ed-3465-4436ac786437</uuid>
  <forward mode='hostdev' managed='yes'>
    <pf dev='eth0' />
```

```
<address type='pci' domain='0x0000' bus='0x02' slot='0x10' function='0x1' />
<address type='pci' domain='0x0000' bus='0x02' slot='0x10' function='0x3' />
<address type='pci' domain='0x0000' bus='0x02' slot='0x10' function='0x5' />
<address type='pci' domain='0x0000' bus='0x02' slot='0x10' function='0x7' />
<address type='pci' domain='0x0000' bus='0x02' slot='0x11' function='0x1' />
<address type='pci' domain='0x0000' bus='0x02' slot='0x11' function='0x3' />
<address type='pci' domain='0x0000' bus='0x02' slot='0x11' function='0x5' />
</forward>
</network>
```

15.10 列出挂接的设备

尽管 **virsh** 中没有任何机制可列出 VM 主机服务器中已挂接到其 VM Guest 的所有设备，但您可以通过运行以下命令列出已挂接到特定 VM Guest 的所有设备：

```
virsh dumpxml VMGUEST_NAME | xpath -e /domain/devices/hostdev
```

例如：

```
> sudo virsh dumpxml sles12 | -e xpath /domain/devices/hostdev
Found 2 nodes:
-- NODE --
<hostdev mode="subsystem" type="pci" managed="yes">
  <driver name="xen" />
  <source>
    <address domain="0x0000" bus="0x0a" slot="0x10" function="0x1" />
  </source>
  <address type="pci" domain="0x0000" bus="0x00" slot="0x0a" function="0x0" />
</hostdev>
-- NODE --
<hostdev mode="subsystem" type="pci" managed="yes">
  <driver name="xen" />
  <source>
    <address domain="0x0000" bus="0x0a" slot="0x10" function="0x2" />
  </source>
  <address type="pci" domain="0x0000" bus="0x00" slot="0x0b" function="0x0" />
</hostdev>
```



提示：列出通过 <interface type='hostdev'> 挂接的 SR-IOV 设备

对于通过 <interface type='hostdev'> 挂接到 VM 主机服务器的 SR-IOV 设备，需要使用不同的 XPath 查询：

```
virsh dumpxml VMGUEST_NAME | xpath -e /domain/devices/interface/@type
```

15.11 配置存储设备

存储设备在 `disk` 元素中定义。一般的 `disk` 元素支持多个属性。下面是两个最重要的属性：

- `type` 属性描述虚拟磁盘设备的来源。有效值为 `file`、`block`、`dir`、`network` 或 `volume`。
- `device` 属性显示如何向 VM Guest 操作系统公开磁盘。例如，可能的值可能包括 `floppy`、`disk`、`cdrom` 等。

下面是最重要的子元素：

- `driver` 包含驱动程序和总线。VM Guest 使用驱动程序和总线来操作新磁盘设备。
- `target` 元素包含新磁盘显示在 VM Guest 中时所用的设备名称。它还包含可选的总线属性，该属性定义用于操作新磁盘的总线的类型。

下面的过程说明如何将储存设备添加到 VM Guest：

1. 编辑现有 VM Guest 的配置：

```
> sudo virsh edit sles15
```

2. 在 `devices` 元素内添加 `disk` 元素，同时指定其 `type` 和 `device` 属性：

```
<disk type='file' device='disk'>
```

3. 指定 `driver` 元素并使用默认值：

```
<driver name='qemu' type='qcow2' />
```

4. 创建一个磁盘映像作为新虚拟磁盘设备的来源：

```
> sudo qemu-img create -f qcow2 /var/lib/libvirt/images/sles15.qcow2 32G
```

5. 添加磁盘来源的路径：

```
<source file='/var/lib/libvirt/images/sles15.qcow2' />
```

6. 定义 VM Guest 中的目标设备名以及磁盘所使用的总线：

```
<target dev='vda' bus='virtio' />
```

7. 重新启动您的 VM：

```
> sudo virsh start sles15
```

现在，新存储设备在 VM Guest 操作系统中应该可供使用。

15.12 配置控制器设备

libvirt 根据 VM Guest 使用的虚拟设备类型自动管理控制器。如果 VM Guest 包含 PCI 和 SCSI 设备，系统会自动创建并管理 PCI 和 SCSI 控制器。**libvirt** 还可为特定于超级管理程序的控制器（例如 KVM VM Guest 的 `virtio-serial` 控制器，或 Xen VM Guest 的 `xenbus` 控制器）建模。尽管默认控制器及其配置在一般情况下都可满足需求，但在某些使用场景中，需要手动调整控制器或其属性。例如，`virtio-serial` 控制器可能需要更多端口，或者 `xenbus` 控制器可能需要更多内存或更多虚拟中断。

`Xenbus` 控制器的独特之处在于，它充当着所有 Xen 半虚拟设备的控制器。如果 VM Guest 包含许多磁盘和/或网络设备，则控制器可能需要更多内存。Xen 的 `max_grant_frames` 属性设置要将多少授权帧或共享内存块分配给每个 VM Guest 的 `xenbus` 控制器。

默认值 32 在大多数情况下已够用，但包含多个 I/O 设备的 VM Guest 以及 I/O 密集型工作负载可能会由于授权帧耗尽而发生性能问题。**xen-diag** 可以检查 dom0 与 VM Guest 的当前和最大 `max_grant_frames` 值。VM Guest 必须正在运行：


```
> sudo virsh list
Id    Name           State
-----
0     Domain-0       running
3     sle15sp1       running

> sudo xen-diag gnttab_query_size 0
domid=0: nr_frames=1, max_nr_frames=256

> sudo xen-diag gnttab_query_size 3
domid=3: nr_frames=3, max_nr_frames=32
```

sle15sp1 Guest 仅使用了 32 个帧中的 3 个。如果您发现了性能问题并且有日志项指出帧数不足，请使用 **virsh** 提高该值。查看 Guest 配置文件中的 `<controller type='xenbus'>` 行，并添加 `maxGrantFrames` 控制元素：

```
> sudo virsh edit sle15sp1
<controller type='xenbus' index='0' maxGrantFrames='40' />
```

保存更改并重新启动 Guest。现在，`xen-diag` 命令应该会显示您的更改：

```
> sudo xen-diag gnttab_query_size 3
domid=3: nr_frames=3, max_nr_frames=40
```

与 `maxGrantFrames` 类似，`xenbus` 控制器也支持 `maxEventChannels`。事件通道类似于半虚拟中断，它们与授权帧共同构成半虚拟驱动程序的数据传输机制。它们还用于处理器间的中断。包含大量 vCPU 和/或许多半虚拟设备的 VM Guest 可能需要增大最大默认值 1023。更改 `maxEventChannels` 的方式与更改 `maxGrantFrames` 类似：

```
> sudo virsh edit sle15sp1
<controller type='xenbus' index='0' maxGrantFrames='128'
maxEventChannels='2047' />
```

有关详细信息，请参见 <https://libvirt.org/formatdomain.html#elementsControllers> 上 libvirt 的《Domain XML format》手册中的“Controllers”一节。

15.13 配置视频设备

使用虚拟机管理器时，只能定义视频设备型号。只能在 XML 配置中更改分配的 VRAM 量或 2D/3D 加速。

15.13.1 更改分配的 VRAM 量

1. 编辑现有 VM Guest 的配置：

```
> sudo virsh edit sles15
```

2. 更改分配的 VRAM 大小：

```
<video>
<model type='vga' vram='65535' heads='1'>
...
</model>
</video>
```

3. 通过查看虚拟机管理器中的数量来检查 VM 中的 VRAM 量是否已更改。

15.13.2 更改 2D/3D 加速状态

1. 编辑现有 VM Guest 的配置：

```
> sudo virsh edit sles15
```

2. 要启用/禁用 2D/3D 加速，请相应地更改 `accel3d` 和 `accel2d` 的值：

```
<video>
<model>
  <acceleration accel3d='yes' accel2d='no'>
</model>
</video>
```



提示：启用 2D/3D 加速

只有 virtio 和 vbox 视频设备支持 2D/3D 加速。无法在其他视频设备上启用此功能。

15.14 配置网络设备

本节介绍如何使用 **virsh** 配置虚拟网络设备的特定方面。

<https://libvirt.org/formatdomain.html#elementsDriverBackendOptions> 中提供了有关 libvirt 网络接口规范的更多细节。

15.14.1 使用多队列 virtio-net 提升网络性能

多队列 virtio-net 功能允许 VM Guest 的虚拟 CPU 并行传输包，因此可以提升网络性能。有关更多一般信息，请参见第 35.3.3 节 “使用多队列 virtio-net 提升网络性能”。

要为特定的 VM Guest 启用多队列 virtio-net，请按照第 15.1 节 “编辑 VM 配置” 中所述编辑其 XML 配置，并按如下所示修改其网络接口：

```
<interface type='network'>
  [...]
  <model type='virtio' />
  <driver name='vhost' queues='NUMBER_OF_QUEUES' />
</interface>
```

15.15 使用 macvtap 共享 VM 主机服务器网络接口

使用 Macvtap 可将 VM Guest 虚拟接口直接挂接到主机网络接口。基于 macvtap 的接口扩展了 VM 主机服务器网络接口，它在相同以太网段上有自己的 MAC 地址。通常，使用此功能是为了使 VM Guest 和 VM 主机服务器都直接显示在 VM 主机服务器连接的交换机上。



注意：Macvtap 不能与 Linux 网桥搭配使用

Macvtap 不能与已连接到 Linux 网桥的网络接口搭配使用。在尝试创建 macvtap 接口之前，请去除网桥中的接口。



注意：使用 macvtap 在 VM Guest 与 VM 主机服务器之间通讯

使用 macvtap 时，一个 VM Guest 可与其他多个 VM Guest 通讯，并可与网络上的其他外部主机通讯。但是，该 VM Guest 无法与用于运行它的 VM 主机服务器通讯。这是规定的 macvtap 行为，原因与 VM 主机服务器物理以太网挂接到 macvtap 网桥的方式有关。从 VM Guest 进入该网桥并转发到物理接口的流量无法回弹到 VM 主机服务器的 IP 堆栈。同样，来自 VM 主机服务器 IP 堆栈并发送到物理接口的流量无法回弹到 macvtap 网桥以转发到 VM Guest。

libvirt 通过指定接口类型 `direct` 支持基于 macvtap 的虚拟网络接口。例如：

```
<interface type='direct'>
  <mac address='aa:bb:cc:dd:ee:ff' />
  <source dev='eth0' mode='bridge' />
  <model type='virtio' />
</interface>
```

可以使用 `mode` 属性控制 macvtap 设备的操作模式。以下列表显示了该属性的可能值以及每个值的说明：

- `vepa`：将所有 VM Guest 包发送到外部网桥。如果包的目标是某个 VM Guest，而该 VM Guest 所在的 VM 主机服务器与包的来源服务器相同，那么这些包将由支持 VEPA 的网桥（现今的网桥通常都不支持 VEPA）发回到该 VM 主机服务器。
- `bridge`：将其目标与来源为同一 VM 主机服务器的包直接递送到目标 macvtap 设备。来源和目标设备需处于 `bridge` 模式才能直接递送。如果其中一个设备处于 `vepa` 模式，则需要使用支持 VEPA 的网桥。

- **private**: 将所有包发送到外部网桥；如果通过外部路由器或网关发送所有包，并且设备会将其发回到 VM 主机服务器，则将所有包递送到同一 VM 主机服务器上的目标 VM Guest。如果来源或目标设备处于 private 模式，将遵循此过程。
- **passthrough**: 可为网络接口提供更强大能力的一种特殊模式。将所有包转发到接口，并允许 virtio VM Guest 更改 MAC 地址或设置混杂模式，以桥接该接口或在该接口上创建 VLAN 接口。在 **passthrough** 模式下，网络接口不可共享。将某个接口分配到 VM Guest 会使其与 VM 主机服务器断开连接。出于此原因，在 **passthrough** 模式下经常会将 SR-IOV 虚拟功能分配到 VM Guest。

15.16 禁用内存气球设备

内存气球已成为 KVM 的默认选项。设备将明确添加到 VM Guest，因此您无需在 VM Guest 的 XML 配置中添加此元素。如果出于任何原因要在 VM Guest 中禁用内存气球，请按如下所示设置 `model='none'`：

```
<devices>
  <memballoon model='none' />
</device>
```

15.17 配置多个监控器（双头）

libvirt 支持使用双头配置在多个监控器上显示 VM Guest 的视频输出。

！ 重要：不受 Xen 支持

Xen 超级管理程序不支持双头配置。

过程 15.3：配置双头

1. 当虚拟机正在运行时，校验 `xf86-video-qxl` 软件包是否已安装在 VM Guest 中：

```
> rpm -q xf86-video-qxl
```

2. 关闭 VM Guest，并按照第 15.1 节 “编辑 VM 配置” 中所述开始编辑其 XML 配置。
3. 校验虚拟显卡的型号是否为 “qxl”：

```
<video>
  <model type='qxl' ... />
```

4. 将显卡型号规格中的 heads 参数从默认值 1 增大为 2，例如：

```
<video>
  <model type='qxl' ram='65536' vram='65536' vgamem='16384' heads='2'
    primary='yes' />
  <alias name='video0' />
  <address type='pci' domain='0x0000' bus='0x00' slot='0x01' function='0x0' />
</video>
```

5. 将虚拟机配置为使用 Spice 显示器而不是 VNC：

```
<graphics type='spice' port='5916' autoport='yes' listen='0.0.0.0'>
  <listen type='address' address='0.0.0.0' />
</graphics>
```

6. 启动虚拟机并使用 virt-viewer 连接到其显示器，例如：

```
> virt-viewer --connect qemu+ssh://USER@VM_HOST/system
```

7. 在 VM 列表中，选择您已修改了其配置的 VM，并单击连接确认。
8. 在 VM Guest 中加载图形子系统 (Xorg) 后，选择视图 > 显示器 > 显示器 2 打开一个新窗口，其中会显示第二个监控器的输出。

15.18 将 IBM Z 上的加密适配器直通到 KVM Guest

15.18.1 简介

IBM Z 计算机附带加密硬件以及一些实用的功能，例如生成随机数、生成数字签名或加密。KVM 允许将这些加密适配器作为直通设备专门用于 Guest。这意味着，超级管理程序无法监测 Guest 与设备之间的通讯。

15.18.2 本章内容

本章介绍如何将 IBM Z 主机上的加密适配器和域专用于 KVM Guest。该过程包括以下基本步骤：

- 对主机上的默认驱动程序屏蔽加密适配器和域。
- 加载 `vfio-ap` 驱动程序。
- 将加密适配器和域分配到 `vfio-ap` 驱动程序。
- 将 Guest 配置为使用加密适配器。

15.18.3 要求

- QEMU/libvirt 虚拟化环境需已正确安装且正常运行。
- 用于运行内核的 `vfio_ap` 和 `vfio_mdev` 模块需在主机操作系统上可用。

15.18.4 将加密适配器专用于 KVM 主机

1. 校验是否已在主机上加载 `vfio_ap` 和 `vfio_mdev` 内核模块：

```
> lsmod | grep vfio_
```

如有任何一个模块未列出，请手动加载，例如：

```
> sudo modprobe vfio_mdev
```

2. 在主机上创建一个新的 MDEV 设备，并校验是否已添加该设备：

```
uuid=$(uuidgen)
$ echo ${uuid} | sudo tee /sys/devices/vfio_ap/matrix/mdev_supported_types/
vfio_ap-passthrough/create
dmesg | tail
[...]
[272197.818811] iommu: Adding device 24f952b3-03d1-4df2-9967-0d5f7d63d5f2
to group 0
[272197.818815] vfio_mdev 24f952b3-03d1-4df2-9967-0d5f7d63d5f2: MDEV:
group_id = 0
```

3. 识别主机逻辑分区中您要专用于 KVM Guest 的设备：

```
> ls -l /sys/bus/ap/devices/
[...]
lrwxrwxrwx 1 root root 0 Nov 23 03:29 00.0016 -> ../../../../devices/ap/
card00/00.0016/
lrwxrwxrwx 1 root root 0 Nov 23 03:29 card00 -> ../../../../devices/ap/card00/
```

在此示例中，该设备是卡 0 队列 16。为了与硬件管理控制台 (HMC) 配置相匹配，需要将十六进制数 16 转换为十进制数 22。

4. 使用以下命令对 zcrypt 屏蔽适配器：

```
> lszcrypt
CARD.DOMAIN TYPE MODE STATUS REQUEST_CNT
-----
00 CEX5C CCA-Coproc online 5
00.0016 CEX5C CCA-Coproc online 5
```

屏蔽适配器：

```
> cat /sys/bus/ap/apmask
0xffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```



```
echo -0x0 | sudo tee /sys/bus/ap/apmask
0x7fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

屏蔽域：

```
> cat /sys/bus/ap/aqmask
0xfffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
echo -0x0 | sudo tee /sys/bus/ap/aqmask
0xfffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

5. 将适配器 0 和域 16（十进制数 22）分配到 vfio-ap：

```
> sudo echo +0x0 > /sys/devices/vfio_ap/matrix/${uuid}/assign_adapter
> echo +0x16 | sudo tee /sys/devices/vfio_ap/matrix/${uuid}/assign_domain
> echo +0x16 | sudo tee /sys/devices/vfio_ap/matrix/${uuid}/
assign_control_domain
```

6. 校验配置的矩阵：

```
> cat /sys/devices/vfio_ap/matrix/${uuid}/matrix
00.0016
```

7. 创建一个新 VM（参见第 10 章 “Guest 安装”）并等待它初始化，或使用现有的 VM。对于这两种情况，请确保 VM 已关闭。

8. 将 VM 的配置更改为使用 MDEV 设备：

```
> sudo virsh edit VM_NAME
[...]
<hostdev mode='subsystem' type='mdev' model='vfio-ap'>
  <source>
    <address uuid='24f952b3-03d1-4df2-9967-0d5f7d63d5f2' />
  </source>
</hostdev>
[...]
```

9. 重新启动 VM：

```
> sudo virsh reboot VM_NAME
```

10. 登录到 Guest 并校验该适配器是否存在：

```
> lszcrypt
CARD.DOMAIN TYPE MODE STATUS REQUEST_CNT
-----
00 CEX5C CCA-Coproc online 1
00.0016 CEX5C CCA-Coproc online 1
```

15.18.5 更多资料

- 第 6 章 “安装虚拟化组件” 中详细介绍了虚拟化组件的安装。
- <https://www.kernel.org/doc/Documentation/s390/vfio-ap.txt> 中详细介绍了 vfio_ap 体系结构。
- <https://bugs.launchpad.net/ubuntu/+source/linux/+bug/1787405> 中提供了一般概览和详细过程。
- <https://www.kernel.org/doc/html/latest/driver-api/vfio-mediated-device.html> 中详细介绍了 VFIO 调解设备 (MDEV) 的体系结构。

16 使用 AMD SEV-SNP 增强虚拟机安全性

您可以通过 AMD 安全加密虚拟化-安全嵌套分页 (SEV-SNP) 增强虚拟机的安全性。AMD SEV-SNP 功能可将虚拟机与主机系统及其他 VM 隔离，保护数据和代码。该功能会对数据进行加密，并确保能检测或跟踪 VM 中代码和数据的任何更改。由于这种做法会隔离 VM，因此其他 VM 或主机不会受到威胁影响。

本章介绍在搭载 SUSE Linux Enterprise Server 15 SP7 系统的 AMD EPYC 服务器上，启用并使用 AMD SEV-SNP 的具体步骤。

16.1 支持的硬件

运行 AMD SEV-SNP 虚拟机需要配备 AMD EPYC（第 3 代或更新版本）的系统。AMD 计算机的 BIOS 必须提供相应选项以启用平台的机密计算支持。

16.2 启用机密计算模块

AMD SEV-SNP 功能所需的软件包通过机密计算模块提供。您必须在系统安装时或之后通过 SUSEConnect 命令行工具启用该模块。

- 要检查模块是否已启用，请运行以下命令：

```
> sudo suseconnect -l
```

这将显示可用模块列表及其激活状态，以及启用非活动模块所需的命令。

非活动的机密计算模块将显示如下：

```
Confidential Computing Technical Preview Module 15 SP6 x86_64  
Activate with: suseconnect -p sle-module-confidential-computing/15.6/x86_64
```

- 要启用机密计算模块技术预览，请运行以下命令：

```
> sudo suseconnect -p sle-module-confidential-computing/15.6/x86_64  
Registering system to SUSE Customer Center
```

```
Updating system details on https://scc.suse.com ...
Activating sle-module-confidential-computing 15.6 x86_64 ...
Adding service to system ...
Installing release package ...
Successfully registered system
```

机密计算模块已启用，您现在可以安装软件包。

16.3 安装软件包并配置基础系统

机密计算模块提供了支持 AMD SEV-SNP 的替代软件包。为了确保最大程度的兼容性，这些软件包基于 SUSE Linux Enterprise Server 15 SP7 的代码流。

需要替换的三个组件包括：

- Linux 内核
- QEMU 虚拟机监视程序
- libvirt 框架

1. 要安装替代软件包，请执行以下命令：

```
> sudo zypper install --from SLE-Module-Confidential-Computing-15-SP6-Pool
--from SLE-Module-Confidential-Computing-15-SP6-Updates qemu-ovmf-x86_64
libvirt kernel-coco
```

替换软件包后，必须通过配置更改来设置系统，以使 AMD SEV-SNP 功能处于可用状态。主机端的 IOMMU 必须配置为非直通模式。这是为了防止外围设备写入属于加密 Guest 的内存，从而破坏其数据完整性。SUSE Linux Enterprise Server 15 SP7 中的默认 IOMMU 配置为 passthrough 模式。

2. 要在 SUSE Linux Enterprise Server 15 SP7 中禁用 IOMMU 配置，请打开 /etc/default/grub 文件，并将 iommu=nopt 添加到 GRUB_CMDLINE_LINUX_DEFAULT 变量。

3. 要更新引导加载程序配置，请运行命令：

```
> sudo ; update-bootloader
```

4. 系统现在可以使用机密计算内核重新启动。引导加载程序中没有选择它作为默认内核，因此请确保在启动菜单中选择它。

16.4 验证安装

您可以验证软件包的安装和配置情况。

1. 要验证系统是否已使用新内核启动，请查看命令 `uname -r` 的响应。

```
> sudo uname -r 6.4.0-150616.coco15sp6-coco
```

确保显示的内核版本包含 `coco` 标签。

2. 要在内核运行时检查 AMD 安全处理器的初始化结果，请运行以下命令查看内核日志：

```
> sudo dmesg | grep -i ccp
[ 10.103166] ccp 0000:42:00.1: enabling device (0000 -> 0002)
[ 10.114951] ccp 0000:42:00.1: no command queues available
[ 10.127137] ccp 0000:42:00.1: sev enabled
[ 10.133152] ccp 0000:42:00.1: psp enabled
[ 10.240817] ccp 0000:42:00.1: SEV firmware update successful
[ 11.128307] ccp 0000:42:00.1: SEV API:1.55 build:8
[ 11.135057] ccp 0000:42:00.1: SEV-SNP API:1.55 build:8
```

出现 SEV-SNP API 版本的相关消息表示 AMD 安全处理器已成功初始化。有时，内核日志中不显示这些消息。这种情况下，通常是由于 BIOS 设置或 IOMMU 配置的问题。

16.5 启动 AMD SEV-SNP 虚拟机

在机密计算内核启动且 AMD 安全处理器初始化完成后，您可以使用 `libvirt` 框架运行受 AMD SEV-SNP 保护的虚拟机。

`libvirt` 提供了多种设置新虚拟机的方法。本文档使用预先准备好的磁盘映像和 `virt-manager` 图形用户界面。

1. 将 `virt-manager` 连接到 AMD EPYC 主机并创建新虚拟机。

2. 在“创建新虚拟机”窗口中，选择以下详细信息：

- 选择操作系统安装方式。
- 选择 ISO 或 CD-ROM 安装介质。
- 选择内存和 CPU 设置。
- 选择所需的存储详情。

3. 在第五步中，验证详细信息并选择在安装前自定义配置。



图 16.1：创建虚拟机

4. 单击完成。

5. 在虚拟机配置窗口中选择“XML”选项卡。

在“XML”选项卡中，您可以编辑由 libvirt 后端使用的虚拟机 XML 配置。

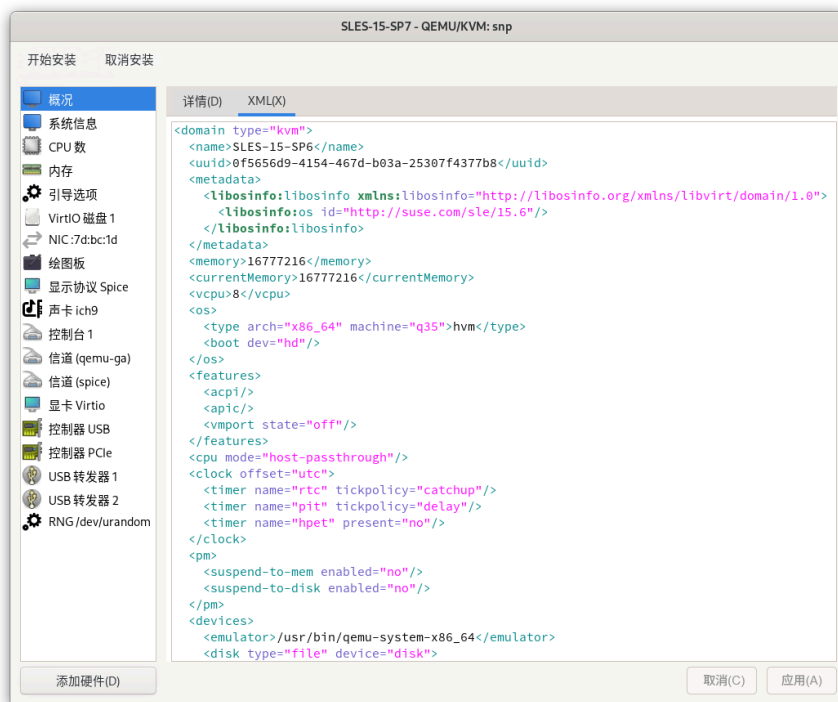


图 16.2：虚拟机配置的 XML 视图

6. 要使用 AMD SEV-SNP 保护虚拟机，请按以下所示修改 `os` 部分以设置正确的固件：

```
<os>
  <type arch="x86_64" machine="pc-q35-8.2">hvm</type>
  <loader readonly="yes" type="rom">/usr/share/qemu/ovmf-x86_64-sev.bin</loader>
  <boot dev="hd"/>
</os>
```

图 16.3：设置固件

其中 `loader` 行将固件设置为 SEV 版本的 OVMF。

7. 添加 `launchSecurity` 部分。对于 AMD SEV-SNP，该部分应如下所示：

```
<launchSecurity type="sev-snp">
  <policy>0x00030000</policy>
</launchSecurity>
```

图 16.4：LAUNCHSECURITY

8. 单击应用选项卡，然后单击细节选项卡。
9. 在左侧列表中选择 CPU，并将 CPU 型号 设置为 `host-model`：

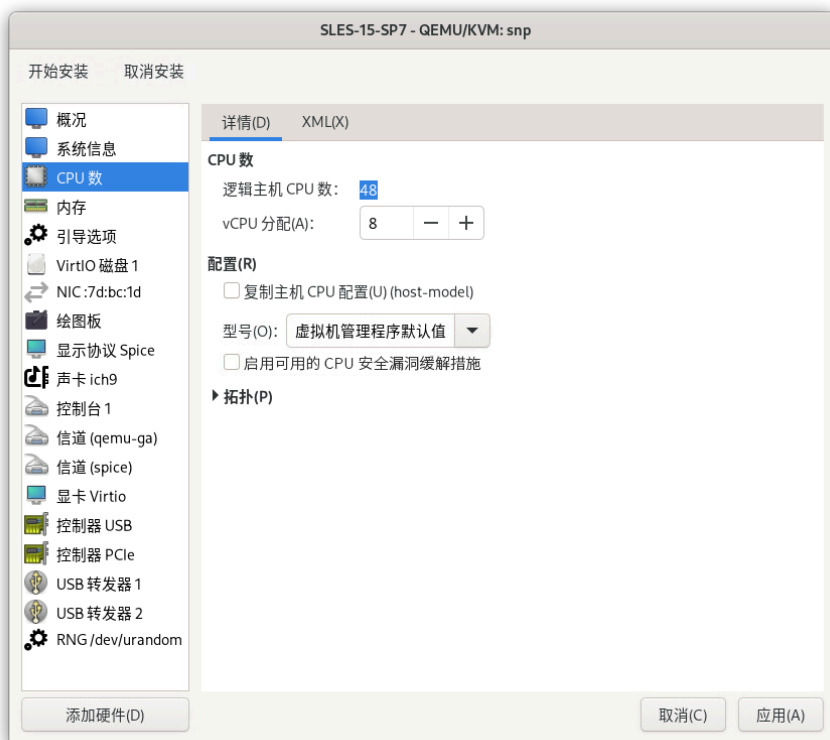


图 16.5：虚拟机配置的详情视图

10. 单击应用，然后单击开始安装。

这将启动虚拟机并根据您的设置进行安装。当上述流程完成后，虚拟机将正常启动，此时您可验证 AMD SEV-SNP 保护是否生效。

16.6 验证 AMD SEV-SNP 虚拟机

从虚拟机的外观来看，无法判断它是否在机密计算环境中运行。但可以通过以下几种方式从虚拟机内部进行验证。

要检查内核日志，请运行以下命令：

```
> sudo dmesg | grep -i sev-snp  
[ 1.986186] Memory Encryption Features active: AMD SEV SEV-ES SEV-SNP
```

内核日志中显示的 SEV-SNP 功能及其他活动内存加密特性，表明该虚拟机已启用此功能。

此外，还可以通过加密安全的方法验证 AMD SEV-SNP 环境的安全性。

17 迁移 VM Guest

虚拟化的主要优势之一是 VM Guest 可移植。当 VM 主机服务器需要维护时，或者当该主机过载时，可以轻松将 Guest 迁移到另一台 VM 主机服务器。KVM 和 Xen 甚至支持“实时”迁移，在此期间，VM Guest 仍然可用。

17.1 迁移类型

根据所需场景，您可以通过三种方式迁移虚拟机 (VM)。

实时迁移

源 VM 在将配置和内存传输至目标主机期间持续运行。传输完成后，源 VM 会暂停，目标虚拟机恢复运行。

实时迁移适用于需要保持在线不停机的 VM。



注意

I/O 负载较重或内存页写入频繁的 VM 很难进行实时迁移。此类情况建议采用非实时迁移或离线迁移。

非实时迁移

源 VM 将被暂停，其配置和内存将传输至目标主机。随后目标 VM 恢复运行。

非实时迁移相比实时迁移更为可靠，但会造成 VM 中断。如果能接受停机，对于难以实施实时迁移的 VM，可采用非实时迁移方案。

离线迁移

VM 定义会被传输至目标主机。源 VM 不会被停止，目标 VM 也不会恢复运行。

离线迁移适用于迁移处于非活动状态的 VM。



重要

采用离线迁移时，必须使用 `--persistent` 选项。

17.2 迁移要求

要成功将 VM Guest 迁移到另一台 VM 主机服务器，需符合以下要求：

- 源和目标系统必须采用相同的体系结构。
- 两台计算机必须都能访问存储设备，例如通过 NFS 或 iSCSI。有关详细信息，请访问[第 13 章 “高级存储主题”](#)。
在迁移期间连接的 CD-ROM 或软盘映像也需要符合此要求。不过，您可以按照[第 14.11 节 “使用虚拟机管理器弹出和更换软盘或 CD/DVD-ROM 媒体”](#)中所述在迁移之前断开其连接。
- 这两台 VM 主机服务器上都需要运行 `libvirtd`，并且您必须能够打开从目标主机到源主机（或反向）的远程 `libvirt` 连接。有关细节，请参见[第 12.3 节 “配置远程连接”](#)。
- 如果目标主机上在运行防火墙，则需要打开端口以允许迁移。如果您在迁移期间不指定端口，`libvirt` 将从 49152:49215 范围内选择一个端口。确保在**目标主机**上的防火墙中打开此端口范围（建议）或所选的专用端口。
- 源计算机和目标计算机应位于网络上的同一子网中，否则迁移后网络功能无法正常工作。
- 所有参与迁移的 VM 主机服务器所用 `qemu` 用户的 UID 必须相同，并且使用的 `kvm`、`qemu` 和 `libvirt` 组的 GID 必须相同。
- 目标主机上不能存在正在运行或已暂停的同名 VM Guest。如果存在已关闭的同名计算机，其配置将被覆盖。
- 迁移 VM Guest 时支持除**主机 CPU** 型号以外的所有 CPU 型号。
- [SATA](#) 类型地磁盘设备无法迁移。
- 文件系统直通功能不可迁移。
- 需在 VM 主机服务器和 VM Guest 上安装适当的计时功能。请参见[第 20 章 “VM Guest 时钟设置”](#)。
- 无法将物理设备从主机迁移到 Guest。目前，在使用具有 PCI 直通或 [SR-IOV](#) 功能的设备时，不支持实时迁移。如果需要支持实时迁移，请使用软件虚拟化（半虚拟化或全虚拟化）。

- 缓存模式设置是重要的迁移设置。请参见第 19.6 节 “缓存模式和实时迁移”。
- 不支持向后迁移（例如，从 SLES 15 SP2 迁移到 15 SP1）。
- SUSE 致力于支持将 VM Guest 从运行 LTSS 所涵盖的服务包的 VM 主机服务器，迁移到运行同一 SLES 主要版本中更新的服务包的 VM 主机服务器。例如，将 VM Guest 从 SLES 12 SP2 主机迁移到 SLES 12 SP5 主机。对于从 LTSS 迁移到更新的服务包的场景，SUSE 只会执行极简单的测试，建议在尝试迁移关键 VM Guest 之前执行全面的现场测试。
- 在两台主机上，映像目录应位于同一路径。
- 所有主机的微代码级别（尤其是 Spectre 微代码更新）应该相同。在所有主机上安装 SUSE Linux Enterprise Server 的最新更新即可实现此目的。

17.3 使用虚拟机管理器进行实时迁移

使用虚拟机管理器迁移 VM Guest 时，在哪台计算机上启动虚拟机管理器并不重要。您可以在源主机或目标主机上启动虚拟机管理器，甚至可以在这两台之外的主机上启动。对于后一种情况，您需要能够同时与目标主机和源主机建立远程连接。

1. 启动虚拟机管理器，并与目标主机或源主机建立连接。如果虚拟机管理器不是在目标主机或源主机上启动的，则需要与这两台主机都建立连接。
2. 右键单击要迁移的 VM Guest，然后选择迁移。确保该 Guest 正在运行或已暂停 — 关闭的 Guest 无法迁移。



提示：提高迁移速度

要提高迁移速度，请暂停 VM Guest。这相当于第 17.1 节 “迁移类型” 中介绍的“非实时迁移”。

3. 为 VM Guest 选择一个新主机。如果所需的目标主机未显示，请确保您已连接到该主机。要更改用于连接到远程主机的默认选项，请在连接下设置模式，以及目标主机的地址（IP 地址或主机名）和端口。如果指定端口，还必须指定地址。
在高级选项下，选择迁移是永久性的（默认设置）还是暂时性的（使用临时迁移）。

此外我们还提供了一个选项允许不安全的迁移，它允许在不禁用 VM 主机服务
器缓存的情况下进行迁移。这样可以加速迁移，但仅当当前配置能够在不使用
`cache="none"/0_DIRECT` 的情况下提供一致的 VM Guest 存储信息时，此选项才起作
用。



注意：带宽选项

在最近的虚拟机管理器版本中，去除了用于设置迁移带宽的选项。要设置特定的带
宽，请改用 `virsh`。

4. 要执行迁移，请单击迁移。

迁移完成后，迁移窗口将会关闭，该 VM Guest 随即列在新主机的虚拟机管理器窗口中。
原始 VM Guest 仍可在源主机上使用（处于关机状态）。

17.4 使用 `virsh` 进行迁移

要使用 `virsh migrate` 迁移 VM Guest，您需要能够直接访问或者通过外壳远程访问 VM 主机
服务器，因为命令需在主机上运行。迁移命令如下所示：

```
> virsh migrate [OPTIONS] VM_ID_or_NAME CONNECTION_URI [--migrateuri  
tcp://REMOTE_HOST:PORT]
```

下面列出了最重要的选项。有关完整列表，请参见 `virsh help migrate`。

`--live`

执行实时迁移。如果未指定此选项，Guest 将会在迁移期间暂停（“非实时迁移”）。

`--suspend`

在实时或非实时迁移期间，暂停目标主机上的 VM。

`--persistent`

在目标主机上保留迁移后的 VM。如果不使用此选项，当该 VM 关闭时，它不会出现在
`virsh list --all` 报告的域列表中。

--undefinesource

如果指定此选项，成功迁移后，将删除源主机上的 VM Guest 定义，但**不会**删除挂接到此 Guest 的虚拟磁盘。

--parallel --parallel-connections NUM_OF_CONNECTIONS

当单个迁移线程无法使源主机与目标主机之间的网络链接饱和时，可以使用并行迁移来提高迁移数据吞吐量。在具备 40 GB 网络接口的主机上，可能需要四个迁移线程才能使链接饱和。使用并行迁移可以缩短迁移大内存 VM 所需的时间。

以下示例使用 `mercury.example.com` 作为源系统，使用 `jupiter.example.com` 作为目标系统；VM Guest 的名称为 `opensuse131`，ID 为 `37`。

使用默认参数进行非实时迁移

```
> virsh migrate 37 qemu+ssh://tux@jupiter.example.com/system
```

使用默认参数进行瞬态实时迁移

```
> virsh migrate --live opensuse131 qemu+ssh://tux@jupiter.example.com/system
```

永久性实时迁移；删除源上的 VM 定义

```
> virsh migrate --live --persistent --undefinesource 37 \
qemu+tls://tux@jupiter.example.com/system
```

非实时迁移，使用端口 49152

```
> virsh migrate opensuse131 qemu+ssh://tux@jupiter.example.com/system \
--migrateuri tcp://@jupiter.example.com:49152
```

实时迁移，转移所有已使用的存储资源

```
> virsh migrate --live --persistent --copy-storage-all \
opensuse156 qemu+ssh://tux@jupiter.example.com/system
```

重要

使用 `--copy-storage-all` 选项迁移 VM 的存储资源时，存储资源必须位于 `libvirt` 的存储池中。目标存储池必须与源存储池类型相同、名称一致。

要获取源存储池的 XML 表示，请使用以下命令：

```
> sudo virsh pool-dumpxml EXAMPLE_VM > EXAMPLE_POOL.xml
```

要在目标主机上创建并启动该存储池，请将其 XML 文件复制到目标主机，并执行以下命令：

```
> sudo virsh pool-define EXAMPLE_POOL.xml  
> sudo virsh pool-start EXAMPLE_VM
```



注意：瞬态迁移与永久性迁移的比较

默认情况下，**virsh migrate** 会为目标主机上的 VM Guest 创建临时（瞬态）副本。已关机版本的原始 Guest 说明将保留在源主机上。瞬态副本将会在 Guest 关机后从服务器中删除。

要为目标主机上的 Guest 创建永久副本，请使用开关 `--persistent`。已关机版本的原始 Guest 说明也会保留在源主机上。将选项 `--undefinesource` 与 `--persistent` 一起使用可以实现“真正”的迁移，在此情况下，将在目标主机上创建永久副本，并删除源主机上的版本。

不建议只使用 `--undefinesource` 而不使用 `--persistent` 选项，因为这会导致两个 VM Guest 定义均会在目标主机上的 Guest 关闭后丢失。

17.5 分步操作示例

17.5.1 导出存储区

首先需要导出存储区，以便在主机之间共享 Guest 映像。可以通过一台 NFS 服务器完成此操作。在以下示例中，我们想要共享网络 10.0.1.0/24 中所有计算机的 `/volume1/VM` 目录。我们使用了一个 SUSE Linux Enterprise NFS 服务器。以 root 用户身份编辑 `/etc/exports` 文件，在其中添加以下内容：

```
/volume1/VM 10.0.1.0/24 (rw,sync,no_root_squash)
```

您需要重新启动该 NFS 服务器：

```
> sudo systemctl restart nfsserver
> sudo exportfs
/volume1/VM      10.0.1.0/24
```

17.5.2 在目标主机上定义池

在您要迁移 VM Guest 的每台主机上，必须定义池才能访问卷（其中包含 Guest 映像）。我们的 NFS 服务器 IP 地址为 10.0.1.99，它的共享是 `/volume1/VM` 目录，而我们想要将此共享挂载到 `/var/lib/libvirt/images/VM` 目录中。池名称为 **VM**。要定义此池，请创建包含以下内容的 `VM.xml` 文件：

```
<pool type='netfs'>
  <name>VM</name>
  <source>
    <host name='10.0.1.99' />
    <dir path='/volume1/VM' />
    <format type='auto' />
  </source>
  <target>
    <path>/var/lib/libvirt/images/VM</path>
    <permissions>
      <mode>0755</mode>
      <owner>-1</owner>
      <group>-1</group>
    </permissions>
  </target>
</pool>
```

然后使用 **pool-define** 命令将其加载到 `libvirt`：

```
# virsh pool-define VM.xml
```

另一种定义此池的方法是使用 **virsh** 命令：

```
# virsh pool-define-as VM --type netfs --source-host 10.0.1.99 \
  --source-path /volume1/VM --target /var/lib/libvirt/images/VM
```

```
Pool VM created
```

以下命令假设您在 **virsh** 的交互式外壳中操作，使用不带任何参数的 **virsh** 命令也可以访问该外壳。然后，可将池设置为在主机引导时自动启动（autostart 选项）：

```
virsh # pool-autostart VM
Pool VM marked as autostarted
```

要禁用自动启动，请运行以下命令：

```
virsh # pool-autostart VM --disable
Pool VM unmarked as autostarted
```

检查该池是否存在：

```
virsh # pool-list --all
Name                               State      Autostart
-----
default                            active     yes
VM                                 active     yes

virsh # pool-info VM
Name:          VM
UUID:          42efelb3-7eaa-4e24-a06a-ba7c9ee29741
State:         running
Persistent:    yes
Autostart:     yes
Capacity:      2,68 TiB
Allocation:    2,38 TiB
Available:     306,05 GiB
```



警告：池需要在所有目标主机上存在

记住：必须在您要迁移 VM Guest 的每台主机上定义此池。

17.5.3 创建卷

池已定义，现在我们需要一个包含磁盘映像的卷：


```
virsh # vol-create-as VM sled12.qcow2 8G --format qcow2
Vol sled12.qcow2 created
```

稍后将会通过 `virt-install` 使用所示的卷名称安装 Guest。

17.5.4 创建 VM Guest

让我们使用 **`virt-install`** 命令创建一个 SUSE Linux Enterprise Server VM Guest。使用 **`--disk`** 选项指定 VM 池；如果您不希望在执行迁移时使用 **`--unsafe`** 选项，建议您设置 **`cache=none`**。

```
# virt-install --connect qemu:///system --virt-type kvm --name \
  sles15 --memory 1024 --disk vol=VM/sled12.qcow2,cache=none --cdrom \
  /mnt/install/ISO/SLE-15-Server-DVD-x86_64-Build0327-Media1.iso --graphics \
  vnc --os-variant sles15
Starting install...
Creating domain...
```

17.5.5 迁移 VM Guest

一切准备就绪，现在可以执行迁移。在当前托管 VM Guest 的 VM 主机服务器上运行 **`migrate`** 命令，并选择目标。

```
virsh # migrate --live sled12 --verbose qemu+ssh://IP/Hostname/system
Password:
Migration: [ 12 %]
```

18 Xen 到 KVM 的迁移指南

随着服务器管理员越来越广泛地使用 KVM 虚拟化解决方案，他们中的许多人都需要有一个途径将其基于 Xen 的现有环境迁移到 KVM。目前还没有成熟的工具可自动将 Xen VM 转换为 KVM。不过，有一个技术解决方案有助于将 Xen 虚拟机转换为 KVM。下面的信息和过程可帮助您执行这样的迁移。

重要：不支持迁移过程

SUSE 不完全支持本文档中所述的迁移过程。我们提供的说明仅作为指导。

18.1 使用 **virt-v2v** 迁移到 KVM

本章包含可帮助您将虚拟机从外部超级管理程序（例如 Xen）导入到 libvirt 所管理的 KVM 的信息。

提示：Microsoft Windows Guest

本章重点介绍如何转换 Linux Guest。使用 **virt-v2v** 转换 Microsoft Windows Guest 与转换 Linux Guest 的过程相同，只不过对虚拟机驱动程序包 (VMDP) 的处理方式有所不同。[虚拟机驱动程序包文档 \(https://documentation.suse.com/sle-vmdp/\)](https://documentation.suse.com/sle-vmdp/)  中单独提供了有关使用 VMDP 转换 Windows Guest 的更多细节。

18.1.1 **virt-v2v** 简介

virt-v2v 是一个命令行工具，可将外部超级管理程序中的 VM Guest 转换为在 libvirt 管理的 KVM 上运行。如果可能，它会在转换的虚拟机中启用半虚拟化 virtio 驱动程序。下表列出了支持的操作系统和超级管理程序：

支持的 GUEST 操作系统

- SUSE Linux Enterprise Server

- openSUSE
- Red Hat Enterprise Linux
- Fedora
- Microsoft Windows Server 2003 和 2008

支持的源超级管理程序

- Xen

支持的目标超级管理程序

- KVM (由 `libvirt` 管理)

18.1.2 安装 `virt-v2v`

安装 `virt-v2v` 的过程很简单：

```
> sudo zypper install virt-v2v
```

请记住，`virt-v2v` 需要 `root` 特权，因此您需要以 `root` 身份或通过 `sudo` 运行该工具。

18.1.3 将虚拟机转换为在 `libvirt` 管理的 KVM 下运行

`virt-v2v` 可将 Xen 超级管理程序中的虚拟机转换为在 `libvirt` 管理的 KVM 上运行。要了解有关 `libvirt` 和 `virsh` 的详细信息，请参见第 II 部分 “使用 `libvirt` 管理虚拟机”。此外，`virt-v2v` 手册页 (`man 1 virt-v2v`) 中也对所有 `virt-v2v` 命令行选项进行了说明。

转换虚拟机之前，请务必完成以下步骤：

过程 18.1：准备转换环境

1. 创建新的本地存储池。

`virt-v2v` 会将源虚拟机的存储复制到由 `libvirt` 管理的本地存储池（原始磁盘映像保持不变）。可以使用虚拟机管理器或 `virsh` 创建该池。有关详细信息，请参见第 9.2.2 节 “使用虚拟机管理器管理存储设备” 和第 9.2.1 节 “使用 `virsh` 管理存储”。

2. 准备本地网络接口。

检查转换的虚拟机能否使用 VM 主机服务器上的本地网络接口。该接口通常是一个网桥，如果尚未定义，请选择 YaST › 系统 › 网络设置 › 添加 › 网桥创建该接口。



注意：网络设备的映射

在转换过程中，可能会将源 Xen 主机上的网络设备映射到 KVM 目标主机上的相应网络设备。例如，可能会将 Xen 网桥 `br0` 映射到默认的 KVM 网络设备。`/etc/virt-v2v.conf` 中提供了一些示例映射。要启用这些映射，请修改 XML 规则，并确保未以 `<!--` 和 `-->` 标记注释掉相应部分。例如：

```
<network type='bridge' name='br0'>
  <network type='network' name='default' />
</network>
```



提示：无网桥

如果没有可用的网桥，虚拟机管理器可以选择性创建一个。

`virt-v2v` 的基本命令语法如下：

```
virt-v2v -i INPUT_METHOD -os STORAGE_POOL SOURCE_VM
```

input_method

有两种输入方法：`libvirt` 或 `libvirtxml`。有关详细信息，请参见 `SOURCE_VM` 参数。

storage_pool

您已为目标虚拟机准备好的储存池。

source_vm

要转换的源虚拟机。其值取决于 `INPUT_METHOD` 参数：如果输入方法为 `libvirt`，需指定 `libvirt` 域的名称。如果输入方法为 `libvirtxml`，需指定包含 `libvirt` 域规范的 XML 文件的路径。



注意：转换时间

转换虚拟机时会占用大量系统资源，主要用于复制虚拟机的整个磁盘映像。转换单个虚拟机一般最长需要 10 分钟时间。对于使用大型磁盘映像的虚拟机，转换时间可能要长很多。

18.1.3.1 基于 libvirt XML 描述文件的转换

本节说明如何使用 `libvirt` XML 配置文件转换本地 Xen 虚拟机。如果主机已在运行 KVM 超级管理程序，则适合使用此方法。确保在本地主机上可以使用源虚拟机的 `libvirt` XML 文件以及其中引用的 `libvirt` 存储池。

1. 获取源虚拟机的 `libvirt` XML 描述。



提示：获取 XML 文件

要获取源虚拟机的 `libvirt` XML 文件，您必须在 Xen 内核下运行主机操作系统。如果您已将主机重引导至启用了 KVM 的环境，请将其重引导回 Xen 内核，转储 `libvirt` XML 文件，然后再重引导回 KVM 环境。

首先，识别 `virsh` 下的源虚拟机：

```
# virsh list
Id      Name                                State
-----
[...]
 2      sles12_xen                         running
[...]
```

`sles12_xen` 是要转换的源虚拟机。现在，导出其 XML 并保存到 `sles12_xen.xml`：

```
# virsh dumpxml sles12_xen > sles12_xen.xml
```

2. 从 KVM 主机的角度校验所有磁盘映像路径是否正确。在一台计算机上转换时，路径正确与否不会产生问题，但使用 XML 转储从其他主机进行转换时，可能需要手动更改路径。

```
<source file='/var/lib/libvirt/images/XenPool/SLES.qcow2' />
```



提示：复制映像

为避免将映像复制两次，请直接手动将一个或多个磁盘映像复制到 `libvirt` 存储池。更新 XML 描述文件中的源文件项。`virt-v2v` 进程将检测现有磁盘，并就地转换这些磁盘。

3. 运行 `virt-v2v` 以转换为 KVM 虚拟机：

```
# virt-v2v sles12_xen.xml ❶ \  
-i LIBVIRTXML ❷ \  
-os remote_host.example.com:/exported_dir ❸ \  
--bridge br0 ❹ \  
-on sles12_kvm ❺
```

- ❶ 基于 Xen 的源虚拟机的 XML 描述。
- ❷ `virt-v2v` 从 `libvirt` XML 文件中读取有关源虚拟机的信息。
- ❸ 用于存放目标虚拟机磁盘映像的存储池。在此示例中，映像将放置在 `remote_host.example.com` 服务器的 NFS 共享 `/exported_dir` 上。
- ❹ 基于 KVM 的目标虚拟机使用主机上的网桥 `br0`。
- ❺ 目标虚拟机将重命名为 `sles12_kvm`，以防与同名的现有虚拟机发生名称冲突。

18.1.3.2 基于 `libvirt` 域名的转换

如果您仍在 Xen 下运行 `libvirt`，打算稍后重引导到 KVM 超级管理程序，则此方法很有用。

1. 确定您要转换的虚拟机的 `libvirt` 域名。

```
# virsh list  
Id      Name                                State  
-----  
[...]  
2       sles12_xen                         running
```

[...]

sles12_xen 是要转换的源虚拟机。

2. 运行 **virt-v2v** 以转换为 KVM 虚拟机：

```
# virt-v2v sles12_xen ❶ \  
-i libvirt ❷ \  
-os storage_pool ❸ \  
--network eth0 ❹ \  
-of qcow2 ❺ \  
-oa sparse ❻ \  
-on sles12_kvm
```

- ❶ 基于 Xen 的虚拟机的域名。
- ❷ **virt-v2v** 直接通过 libvirt 活动连接读取有关源虚拟机的信息。
- ❸ 目标磁盘映像将放置在本地 libvirt 存储池中。
- ❹ 所有 Guest 网桥（或网络）将连接到本地管理的网络。
- ❺ 目标虚拟机的磁盘映像的格式。支持的选项为 raw 或 qcow2。
- ❻ 转换的 Guest 磁盘空间是采用 sparse 模式还是 preallocated 模式。

18.1.3.3 转换远程 Xen 虚拟机

如果您需要转换在远程主机上运行的 Xen 虚拟机，此方法非常有用。由于 **virt-v2v** 通过 ssh 连接到远程主机，因此请确保主机上正在运行 SSH 服务。



注意：无口令 SSH 访问

virt-v2v 要求通过无口令 SSH 连接来连至远程主机。这意味着需向 ssh-agent 添加一个使用 SSH 密钥的连接。有关更多细节，请参见 man ssh-keygen 和 man ssh-add。《安全和强化指南》，第 22 章 “使用 OpenSSH 保护网络操作” 上也提供了详细信息。

要连接到远程 libvirt 连接，请构建远程主机的相关有效连接 URI。在以下示例中，远程主机名为 remote_host.example.com，连接用户名为 root。连接 URI 的格式如下所示：

```
xen+ssh://root@remote_host.example.com/
```

有关 libvirt 连接 URI 的详细信息，请参见 <https://libvirt.org/uri.html>。

1. 确定您要转换的远程虚拟机的 libvirt 域名。

```
# virsh -c xen+ssh://root@remote_host.example.com/ list
Id      Name                                State
-----
1       sles12_xen                          running
[...]
```

sles12_xen 是要转换的源虚拟机。

2. 远程连接的 virt-v2v 命令如下所示：

```
# virt-v2v sles12_xen \
-i libvirt \
-ic xen+ssh://root@remote_host.example.com/ \
-os local_storage_pool \
--bridge br0
```

18.1.4 运行转换的虚拟机

virt-v2v 成功完成后，将以 -on 选项所指定的名称创建一个新的 libvirt 域。如果您未指定 -on，将使用与源虚拟机相同的名称。可以使用 virsh 或虚拟机管理等标准 libvirt 工具管理新 Guest。



提示：重引导计算机

如果您已按第 18.1.3.2 节“基于 libvirt 域名的转换”中所述在 Xen 下完成转换，可能需要重引导主机，并使用非 Xen 内核进行引导。

18.2 Xen 到 KVM 的手动迁移

18.2.1 一般概述

首选的虚拟机管理解决方案是基于 `libvirt` 进行管理；有关详细信息，请参见 <https://libvirt.org/>。与手动定义和运行虚拟机的方式相比，它具有多项优势 — `libvirt` 可以跨平台，支持许多超级管理程序，具有安全的远程管理功能以及虚拟网络功能，最重要的是，它可提供统一的抽象层来管理虚拟机。因此，本文着重于介绍 `libvirt` 解决方案。

一般情况下，从 Xen 迁移到 KVM 需执行以下基本步骤：

1. 创建原始 Xen VM Guest 的备份副本。
2. （可选）应用特定于半虚拟化 Guest 的更改。
3. 获取有关 Xen VM Guest 的信息，并将其更新为 KVM 对等项。
4. 关闭 Xen 主机上的 Guest，然后在 KVM 超级管理程序下运行新 Guest。



警告：无法实时迁移

当源 VM Guest 正在运行时，无法将 Xen 实时迁移到 KVM。运行新的 KVM 就绪 VM Guest 之前，建议您关闭原始 Xen VM Guest。

18.2.2 备份 Xen VM Guest

要备份 Xen VM Guest，请执行以下步骤：

1. 识别您要迁移的相关 Xen Guest 并记住其 ID/名称。

```
> sudo virsh list --all
Id Name                               State
-----
 0 Domain-0                           running
 1 SLES15SP3                           running
[...]
```

2. 关闭 Guest。可以通过关闭 Guest 操作系统或使用 **virsh** 执行此操作：

```
> sudo virsh shutdown SLES11SP3
```

3. 将其配置备份到 XML 文件。

```
> sudo virsh dumpxml SLES11SP3 > sles11sp3.xml
```

4. 备份其磁盘映像文件。使用 **cp** 或 **rsync** 命令创建备份副本。请记住，比较好的做法始终是使用 **md5sum** 命令检查副本。

5. 备份映像文件之后，您可以使用以下命令再次启动 Guest

```
> sudo virsh start SLES11SP3
```

18.2.3 特定于半虚拟化 Guest 的更改

如果您从半虚拟化 Xen Guest 进行迁移，请实施以下更改。可以使用 **guestfs-tools** 在运行中的 Guest 或已停止的 Guest 上执行此操作。



重要

应用本节中所述的更改后，与迁移的 VM Guest 有关的映像文件将无法继续在 Xen 下使用。

18.2.3.1 安装默认内核



警告：不会引导

安装默认内核后，系统无法引导 Xen Guest。

克隆 Xen Guest 磁盘映像以在 KVM 超级管理程序下使用之前，请确保该映像可在**没有** Xen 超级管理程序的情况下引导。这一点对于半虚拟化 Xen Guest 至关重要，因为它们通常包含一个特殊的 Xen 内核，并且通常未安装完整的 GRUB 2 引导加载程序。

1. 对于 SLES 11, 请更新 `/etc/sysconfig/kernel` 文件。更改 `INITRD_MODULES` 参数, 具体做法是去除所有 Xen 驱动程序, 并将它们替换为 virtio 驱动程序。替换

```
INITRD_MODULES="xenblk xennet"
```

替换为

```
INITRD_MODULES="virtio_blk virtio_pci virtio_net virtio_balloon"
```

对于 SLES 12、15 和 openSUSE, 请在 `/etc/dracut.conf.d/*.conf` 中搜索 `xenblk xennet`, 并将其替换为 `virtio_blk virtio_pci virtio_net virtio_balloon`

2. 半虚拟化 Xen Guest 运行特定的 Xen 内核。要在 KVM 下运行 Guest, 您需要安装默认内核。



注意：默认内核已安装

对于全虚拟化 Guest, 您无需安装默认内核, 因为它已安装。

在 Xen Guest 上输入 `rpm -q kernel-default`, 以确认默认内核是否已安装。如果未安装, 请使用 `zypper in kernel-default` 安装。

要用于在 KVM 下引导 Guest 的内核必须有可用的 **virtio** (半虚拟化) 驱动程序。请运行以下命令确认是否如此。不要忘记将 `6.4.0-150700.38` 替换为您的内核版本:

```
> sudo find /lib/modules/6.4.0-150700.38-default/kernel/drivers/ -name virtio*
/lib/modules/6.4.0-150700.38-default/kernel/drivers/block/virtio_blk.ko.zst
/lib/modules/6.4.0-150700.38-default/kernel/drivers/bluetooth/virtio_bt.ko.zst
/lib/modules/6.4.0-150700.38-default/kernel/drivers/char/hw_random/virtio-rng.ko.zst
/lib/modules/6.4.0-150700.38-default/kernel/drivers/crypto/virtio
/lib/modules/6.4.0-150700.38/kernel/drivers/block/virtio_blk.ko
...
```

3. 更新 `/etc/fstab`。将所有存储设备从 `xvda` 更改为 `vda`。

4. 更新引导加载程序配置。在 Xen Guest 上输入 `rpm -q grub2`，以确认 GRUB 2 是否已安装。如果未安装，请使用 `zypper in grub2` 安装。

现在，将新安装的默认内核设为默认的操作系统引导项。此外，去除/更新可能会引用特定于 Xen 的设备的内核命令行选项。您可以使用 YaST（系统 > 引导加载程序）或手动执行此操作。

- 列出所有 Linux 引导菜单项，以确定首选引导菜单项：

```
> cat /boot/grub2/grub.cfg | grep 'menuentry '
```

请记住您新安装的项目的序号（从零开始计数）。

- 将其设为默认引导菜单项：

```
> sudo grub2-set-default N
```

将 N 替换为之前查明的引导菜单项编号。

- 打开 `/etc/default/grub` 进行编辑，并查找 `GRUB_CMDLINE_LINUX_DEFAULT` 和 `GRUB_CMDLINE_LINUX_RECOVERY` 选项。去除或更新对特定于 Xen 的设备的所有引用。在下面的示例中，您可以将

```
root=/dev/xvda1 disk=/dev/xvda console=xvc
```

替换为

```
root=/dev/vda1 disk=/dev/vda
```

不要忘记去除对 `xvc` 型控制台（例如 `xvc0`）的所有引用。

5. 更新 `/boot/grub2` 或 `/boot/grub2-efi` 目录（以 VM 使用的目录为准）中的 `device.map`。将所有存储设备从 `xvda` 更改为 `vda`。
6. 要导入新的默认设置，请运行

```
grub2-mkconfig -o /boot/grub2/grub.cfg
```

18.2.3.2 更新 Guest 以在 KVM 下引导

1. 更新系统以使用默认串行控制台。列出配置的控制台，并去除 xvc? 控制台的符号链接。

```
> sudo ls -l /etc/systemd/system/getty.target.wants/  
getty@tty1.service -> /usr/lib/systemd/system/getty@.service  
getty@xvc0.service -> /usr/lib/systemd/system/getty@xvc0.service  
getty@xvc1.service -> /usr/lib/systemd/system/getty@xvc1.service  
  
# rm /etc/systemd/system/getty.target.wants/getty@xvc?.service
```

2. 更新 /etc/securetty 文件。将 xvc0 替换为 ttyS0。

18.2.4 更新 Xen VM Guest 配置

本节介绍如何导出原始 Xen VM Guest 的配置，以及为了能够将该 Guest 作为 KVM Guest 导入到 libvirt 中需对其实施的特定更改。

18.2.4.1 导出 Xen VM Guest 配置

首先导出 Guest 的配置并保存到某个文件中。例如：

```
> sudo virsh dumpxml SLES11SP3  
<domain type='xen'>  
  <name>SLES11SP3</name>  
  <uuid>fa9ea4d7-8f95-30c0-bce9-9e58ffcabeb2</uuid>  
  <memory>524288</memory>  
  <currentMemory>524288</currentMemory>  
  <vcpu>1</vcpu>  
  <bootloader>/usr/bin/pygrub</bootloader>  
  <os>  
    <type>linux</type>  
  </os>  
  <clock offset='utc' />  
  <on_poweroff>destroy</on_poweroff>  
  <on_reboot>restart</on_reboot>  
  <on_crash>restart</on_crash>
```

```

<devices>
  <emulator>/usr/lib/xen/bin/qemu-dm</emulator>
  <disk type='file' device='disk'>
    <driver name='file' />
    <source file='/var/lib/libvirt/images/
SLES_11_SP2_JeOS.x86_64-0.0.2_para.raw' />
    <target dev='xvda' bus='xen' />
  </disk>
  <interface type='bridge'>
    <mac address='00:16:3e:2d:91:c3' />
    <source bridge='br0' />
    <script path='vif-bridge' />
  </interface>
  <console type='pty'>
    <target type='xen' port='0' />
  </console>
  <input type='mouse' bus='xen' />
  <graphics type='vnc' port='-1' autoport='yes' keymap='en-us' />
</devices>
</domain>

```

<https://libvirt.org/formatdomain.html> 上提供了有关 VM Guest 描述的 libvirt XML 格式的详细信息。

18.2.4.2 对 Guest 配置的一般更改

您需要对导出的 Xen Guest XML 配置进行一些一般性更改，方可使其在 KVM 超级管理程序下运行。以下规则对全虚拟化和半虚拟化 Guest 均适用。以下 XML 元素仅为示例，不需要将其包含在您的具体配置中。



提示：使用的约定

为了表示 XML 配置文件中的节点，整个文档使用了 XPath 语法。例如，为了表示 `<domain>` 标记中的 `<name>`，

```

<domain>
  <name>sles11sp3</name>

```

```
</domain>
```

本文档使用了 XPath 对等项 `/domain/name`。

1. 将 `/domain` 元素的 `type` 属性从 `xen` 更改为 `kvm`。
2. 去除 `/domain/bootloader` 元素部分。
3. 去除 `/domain/bootloader_args` 元素部分。
4. 将 `/domain/os/type` 元素值从 `linux` 更改为 `hvm`。
5. 在 `/domain/os` 元素下添加 `<boot dev="hd"/>`。
6. 在 `/domain/os/type` 元素中添加 `arch` 属性。可接受的值为 `arch="x86_64"` 或 `arch="i686"`
7. 将 `/domain/devices/emulator` 元素从 `/usr/lib/xen/bin/qemu-dm'` 更改为 `/usr/bin/qemu-kvm`。
8. 对与半虚拟化 (PV) Guest 关联的每个磁盘进行以下更改：
 - 将 `/domain/devices/disk/driver` 元素的 `name` 属性从 `file` 更改为 `qemu`，并添加 `type` 属性以指定磁盘类型。例如，有效选项包括 `raw` 和 `qcow2`。
 - 将 `/domain/devices/disk/target` 元素的 `dev` 属性从 `xvda` 更改为 `vda`。
 - 将 `/domain/devices/disk/target` 元素的 `bus` 属性从 `xen` 更改为 `virtio`。
9. 对每个网络接口卡进行以下更改：
 - 如果 `/domain/devices/interface` 中定义了 `model`，请将其 `type` 属性值更改为 `virtio`

```
<model type="virtio">
```

- 删除所有 `/domain/devices/interface/script` 部分。
- 删除其 `dev` 属性以 `vif`、`vnet` 或 `veth` 开头的 `/domain/devices/interface/target` 元素。如果使用的是自定义网络，请将 `dev` 值更改为该目标。

10. 去除 `/domain/devices/console` 元素部分（如果存在）。
11. 去除 `/domain/devices/serial` 元素部分（如果存在）。
12. 将 `/domain/devices/input` 元素中的 `bus` 属性从 `xen` 更改为 `ps2`。
13. 在 `/domain/devices` 元素下添加以下有关内存气球功能的元素。

```
<memballoon model="virtio"/>
```



提示：设备名

`<target dev='hda' bus='ide' />` 控制在哪个设备下向 Guest 操作系统公开磁盘。`dev` 属性指示“逻辑”设备名称。我们无法保证实际指定的设备名称与 Guest 操作系统中的设备名称对应。因此，您可能需要更改引导加载程序命令行上的磁盘映射。例如，如果引导加载程序预期根磁盘为 `hda2`，但 KVM 仍将其视为 `sda2`，请将引导加载程序命令行从

```
[...] root=/dev/hda2 resume=/dev/hda1 [...]
```

更改为

```
[...] root=/dev/sda2 resume=/dev/sda1 [...]
```

对于半虚拟化 `xvda` 设备，请将其更改为：

```
[...] root=/dev/vda2 resume=/dev/vda1 [...]
```

否则，VM Guest 将拒绝在 KVM 环境中引导。

18.2.4.3 目标 KVM Guest 配置

实施上述所有修改后，您的 KVM Guest 的最终配置如下所示：

```
<domain type='kvm'>
  <name>SLES11SP3</name>
  <uuid>fa9ea4d7-8f95-30c0-bce9-9e58ffcabeb2</uuid>
```



```

<memory>524288</memory>
<currentMemory>524288</currentMemory>
<vcpu cpuset='0-3'>1</vcpu>
<os>
  <type arch="x86_64">hvm</type>
  <boot dev="hd"/>
</os>
<clock offset='utc'/>
<on_poweroff>destroy</on_poweroff>
<on_reboot>restart</on_reboot>
<on_crash>restart</on_crash>
<devices>
  <emulator>/usr/bin/qemu-kvm</emulator>
  <disk type='file' device='disk'>
    <driver name='qemu' type="raw"/>
    <source file='/var/lib/libvirt/images/
SLES_11_SP2_Je0S.x86_64-0.0.2_para.raw'/>
    <target dev='vda' bus='virtio'/>
  </disk>
  <interface type='bridge'>
    <mac address='00:16:3e:2d:91:c3'/>
    <source bridge='br0'/>
  </interface>
  <input type='mouse' bus='usb'/>
  <graphics type='vnc' port='5900' autoport='yes' keymap='en-us'/>
  <memballoon model="virtio"/>
</devices>
</domain>

```

将配置保存到主目录下的某个文件中，例如，保存为 SLES11SP3.xml。将其导入之后，它会复制到默认的 /etc/libvirt/qemu 目录中。

18.2.5 迁移 VM Guest

更新 VM Guest 配置并对 Guest 操作系统实施必要的更改之后，请关闭原始 Xen Guest 并在 KVM 超级管理程序下运行其克隆版本。

1. 在控制台中以 root 身份运行 **shutdown -h now**，以关闭 Xen 主机上的 Guest。

2. 根据需要复制与 VM Guest 关联的磁盘映像。默认配置要求将 Xen 磁盘文件从 `/var/lib/xen/images` 复制到 `/var/lib/kvm/images`。如果您之前未创建 VM Guest，则可能需要以 `root` 身份创建 `/var/lib/kvm/images` 目录。

3. 创建新域，然后将其注册到 `libvirt` 中：

```
> sudo virsh define SLES11SP3.xml
Domain SLES11SP3 defined from SLES11SP3.xml
```

4. 校验新 Guest 是否包含在 KVM 配置中。

```
> virsh list --all
```

5. 创建域之后，您可以将其启动：

```
> sudo virsh start SLES11SP3
Domain SLES11SP3 started
```

18.3 更多信息

有关 `libvirt` 的详细信息，请参见 <https://libvirt.org>。

<https://libvirt.org/formatdomain.html> 上提供了有关 `libvirt` XML 格式的更多细节。

III 独立于超级管理程序的功能

- 19 磁盘缓存模式 220
- 20 VM Guest 时钟设置 223
- 21 libguestfs 225
- 22 QEMU Guest 代理 238
- 23 软件 TPM 模拟器 241
- 24 创建 VM Guest 的崩溃转储 244

19 磁盘缓存模式

19.1 什么是磁盘缓存？

磁盘缓存是用于加速在硬盘中存储和访问数据的过程的内存。物理硬盘集成了缓存作为一项标准功能。对于虚拟磁盘，缓存使用 VM 主机服务器的内存，您可以微调其行为（例如，通过设置其类型）。

19.2 磁盘缓存的工作原理

通常，磁盘缓存会存储最近和经常使用的程序与数据。当用户或程序请求数据时，操作系统首先会检查磁盘缓存。如果磁盘缓存中有数据，则操作系统会快速将数据传递给程序，而不是从磁盘中重新读取数据。

第 1 条请求



后续请求



图 19.1：缓存机制

19.3 磁盘缓存的优势

虚拟磁盘设备的缓存会影响 Guest 计算机的整体性能。您可以通过优化缓存模式、磁盘映像格式和存储子系统的组合来提高性能。

19.4 虚拟磁盘缓存模式

如果未指定缓存模式，则默认会使用 `writeback`。每个 Guest 磁盘可以使用以下缓存模式之一：

writeback

`writeback` 使用主机页面缓存。将写入内容放在主机缓存中后，会向 Guest 报告写入已完成。缓存管理将处理对存储设备的提交。Guest 的虚拟存储适配器被告知采用 **writeback** 缓存，因此 Guest 预期会按需发送刷新命令来管理数据完整性。

writethrough

仅当已将数据提交到存储设备后，才将写入操作报告为已完成。Guest 的虚拟存储适配器被告知不存在 **writeback** 缓存，因此 Guest 无需发送刷新命令来管理数据完整性。

none

绕过主机缓存，读取和写入直接在超级管理程序与存储设备之间发生。由于实际存储设备可能在数据仅存入其写入队列时就报告写入完成，因此 Guest 的虚拟存储适配器会被告知存在 **writeback** 缓存。此模式相当于直接访问主机的磁盘。

unsafe

此模式类似于 **writeback** 模式，只不过它会忽略 Guest 发送的所有刷新命令。使用此模式意味着用户更看重性能增益，而不关心主机发生故障时丢失数据的风险。此模式在 Guest 安装期间可能很有用，但对于生产工作负载没有作用。

directsync

仅当已将数据提交到存储设备后才将写入操作报告为已完成，会绕过主机缓存。此模式类似于 **writethrough**，可用于不按需发送刷新的 Guest。

19.5 缓存模式和数据完整性

writethrough、none、directsync

当 Guest 操作系统按需使用刷新命令时，这些模式被视为最安全的模式。对于不安全或不稳定的 Guest，请使用 **writethrough** 或 **directsync**。

writeback

此模式告知 Guest 存在写缓存，并依赖于 Guest 按需发送刷新命令来保持其磁盘映像中的数据完整性。如果主机发生故障，此模式会导致 Guest 丢失数据。原因是将写入操作报告为已完成与将写入内容提交到存储设备之间存在时间差。

unsafe

此模式类似于 **writeback** 缓存，只不过它会忽略 Guest 刷新命令。这意味着，主机故障导致数据丢失的风险更高。

19.6 缓存模式和实时迁移

缓存存储数据会限制支持实时迁移的配置。目前，只能使用 raw 和 qcow2 映像格式进行实时迁移。如果使用群集文件系统，则所有缓存模式都支持实时迁移。否则，只有 none 缓存模式支持在读取/写入共享存储中进行实时迁移。

libvirt 管理层包含根据多种因素检查迁移兼容性的功能。如果 Guest 存储托管在群集文件系统上，并且是只读的或者标记为可共享，则在确定是否允许迁移时会忽略缓存模式。除非将缓存模式设置为 none，否则 libvirt 不允许迁移。不过，您可以使用 “--unsafe” 选项来覆盖对迁移 API 的此项限制，此选项也受 virsh 的支持。例如：

```
> virsh migrate --live --unsafe
```



提示

要设置 AIO 模式 native，缓存模式需设为 none。如果使用另一种缓存模式，则 AIO 模式将静默切换回默认值 threads。

20 VM Guest 时钟设置

在 VM Guest 中保持准确的时间是虚拟化的一项较为困难的工作。保持准确的时间对于网络应用程序特别重要，也是进行 VM Guest 实时迁移的先决条件。



提示：VM 主机服务器上的计时

强烈建议在 VM 主机服务器上也保持准确的时间，例如，通过使用 NTP 来实现（有关详细信息，请参见《管理指南》，第 38 章“使用 NTP 同步时间”）。

20.1 KVM：使用 `kvm_clock`

KVM 提供通过 `kvm_clock` 驱动程序支持的半虚拟化时钟。强烈建议使用 `kvm_clock`。

在运行 Linux 的 VM Guest 中使用以下命令来检查是否已加载 `kvm_clock` 驱动程序：

```
> sudo dmesg | grep kvm-clock
[ 0.000000] kvm-clock: cpu 0, msr 0:7d3a81, boot clock
[ 0.000000] kvm-clock: cpu 0, msr 0:1206a81, primary cpu clock
[ 0.012000] kvm-clock: cpu 1, msr 0:1306a81, secondary cpu clock
[ 0.160082] Switching to clocksource kvm-clock
```

要检查当前使用了哪个时钟源，请在 VM Guest 中运行以下命令。此命令应输出 `kvm-clock`：

```
> cat /sys/devices/system/clocksource/clocksource0/current_clocksource
```



重要：kvm-clock 和 NTP

使用 `kvm-clock` 时，建议同时在 VM Guest 中使用 NTP，并在 VM 主机服务器上也使用 NTP。

20.1.1 其他计时方法

半虚拟化 `kvm-clock` 目前不适用于 Windows* 操作系统。对于 Windows*，请使用 Windows Time Service Tools 进行时间同步。

20.2 Xen 虚拟机时钟设置

在 Xen 4 中，已去除用于在 Xen 主机与 Guest 之间进行时间同步的独立时钟设置 `/proc/sys/xen/independent_wallclock`。引入了新的配置选项 `tsc_mode`。此选项指定使用**时戳计数器**将 Guest 时间与 Xen 服务器同步的方法。其默认值 0 适合大多数硬件和软件环境。

有关 `tsc_mode` 的更多细节，请参见 `xen-tscmode` 手册页 (**man 7 xen-tscmode**)。

21 libguestfs

虚拟机由磁盘映像和定义文件构成。虽然您可以手动访问和操作这些 Guest 组件（在常规超级管理程序进程外部），但这么做本质上存在风险，可能会给数据完整性造成危害和风险。libguestfs 是用于安全访问和修改**虚拟机**磁盘映像的一个 C 语言库和一组相应工具 — 它能在常规超级管理程序进程之外进行操作，同时规避通常与手动编辑相关的风险。



重要

只有 AMD64/Intel 64 体系结构完全支持使用 libguestfs 工具。

21.1 VM Guest 操作概述

21.1.1 VM Guest 操作风险

磁盘映像和定义文件不过是 Linux 环境中另一种类型的文件，因此可以使用许多工具来访问、编辑这些文件以及向其中写入数据。如果正确使用，此类工具可成为 Guest 管理的重要组成部分。但是，即使是正确使用这些工具，也不一定能够杜绝风险。手动操作 Guest 磁盘映像时应考虑到如下风险：

- **数据损坏**：如果绕过虚拟化保护层，通过主机计算机或群集中的另一节点并发访问映像，可能会导致更改丢失或数据损坏。
- **安全性**：将磁盘映像挂载为循环设备需要 root 访问权限。如果映像不是循环挂载的，其他用户和进程就有可能可以访问磁盘内容。
- **管理员错误**：正确绕过虚拟化层需要对虚拟组件和工具有深入的了解。如果在做出更改后无法隔离映像或无法正确进行清理，可能会导致在恢复虚拟化控制后出现其他问题。

21.1.2 libguestfs 的设计用途

libguestfs C 库用于安全地创建、访问和修改虚拟机 (VM Guest) 磁盘映像。它还提供对 [Perl \(https://libguestfs.org/guestfs-perl.3.html\)](https://libguestfs.org/guestfs-perl.3.html)、[Python \(https://libguestfs.org/guestfs-python.3.html\)](https://libguestfs.org/guestfs-python.3.html) 和 [Ruby \(https://libguestfs.org/guestfs-ruby.3.html\)](https://libguestfs.org/guestfs-ruby.3.html) 的其他语言绑定。libguestfs 无需 root 权限即可访问 VM Guest 磁盘映像，并提供多层防御机制来防范恶意磁盘映像。

libguestfs 提供许多用于访问和修改 VM Guest 磁盘映像与内容的工具。这些工具提供如下功能：查看和编辑 Guest 内部的文件、通过脚本对 VM Guest 进行更改、监控已用/可用磁盘空间统计数据、创建 Guest、执行 V2V 或 P2V 迁移、执行备份、克隆 VM Guest、格式化磁盘和调整磁盘大小。



警告：最佳实践

切勿在实时虚拟机上使用 libguestfs 工具，这可能会导致 VM Guest 中发生磁盘损坏。libguestfs 工具会尝试阻止您这样做，但无法做到万无一失。

但是，大多数命令都具有 `--ro`（只读）选项。使用此选项可以在实时虚拟机上运行命令。结果可能比较奇怪或不一致，但可以避免磁盘损坏的风险。

21.2 软件包安装

libguestfs 通过 4 个软件包提供：

- [`libguestfs0`](#)：提供主 C 库
- [`guestfs-data`](#)：包含启动映像时使用的设备文件（存储在 `/usr/lib64/guestfs` 中）
- [`guestfs-tools`](#)：核心 guestfs 工具、手册页和 `/etc/libguestfs-tools.conf` 配置文件。
- [`guestfs-winsupport`](#)：在 guestfs 工具中提供对 Windows 文件 Guest 的支持。仅当您需要处理 Windows Guest 时才需安装此软件包，例如，在将 Windows Guest 转换为 KVM 时。

要在系统上安装 guestfs 工具，请运行：

```
> sudo zypper in guestfs-tools
```

21.3 Guestfs 工具

21.3.1 修改虚拟机

guestfs-tools 软件包中的工具集用于访问和修改虚拟机磁盘映像。此功能通过用户所熟悉的、可提供内置保护措施来确保映像完整性的外壳界面提供。Guestfs 工具外壳会公开 guestfs API 的所有功能，并使用计算机上安装的软件包以及 /usr/lib64/guestfs 中的文件即时创建设备。

21.3.2 支持的文件系统和磁盘映像

Guestfs 工具支持多种文件系统，包括：

- Ext2、Ext3、Ext4
- Xfs
- Btrfs

还支持多种磁盘映像格式：

- raw
- qcow2



警告：不支持的文件系统

Guestfs 可能还支持 Windows* 文件系统（VFAT、NTFS）、BSD* 和 Apple* 文件系统，以及其他磁盘映像格式（VMDK、VHDX...）。但这些文件系统和磁盘映像格式在 SUSE Linux Enterprise Server 上不受支持。

21.3.3 virt-rescue

virt-rescue 类似于救援 CD，但用于虚拟机，且无需提供 CD。**virt-rescue** 为用户提供救援外壳和多种简单的恢复工具，可用于检查和更正虚拟机或磁盘映像中的问题。

```
> virt-rescue -a sles.qcow2
Welcome to virt-rescue, the libguestfs rescue shell.

Note: The contents of / are the rescue appliance.
You need to mount the guest's partitions under /sysroot
before you can examine them. A helper script for that exists:
mount-rootfs-and-chroot.sh /dev/sda1

><rescue>
[ 67.194384] EXT4-fs (sda1): mounting ext3 file system
using the ext4 subsystem
[ 67.199292] EXT4-fs (sda1): mounted filesystem with ordered data
mode. Opts: (null)
mount: /dev/sda1 mounted on /sysroot.
mount: /dev bound on /sysroot/dev.
mount: /dev/pts bound on /sysroot/dev/pts.
mount: /proc bound on /sysroot/proc.
mount: /sys bound on /sysroot/sys.
Directory: /root
Thu Jun  5 13:20:51 UTC 2014
(none):~ #
```

您现在是以救援模式运行 VM Guest 的：

```
(none):~ # cat /etc/fstab
devpts /dev/pts          devpts mode=0620,gid=5 0 0
proc   /proc                proc   defaults                0 0
sysfs   /sys                  sysfs   noauto                  0 0
debugfs /sys/kernel/debug      debugfs noauto                  0 0
usbfs   /proc/bus/usb          usbfs   noauto                  0 0
tmpfs   /run                 tmpfs   noauto                  0 0
/dev/disk/by-id/ata-QEMU_HARDDISK_QM00001-part1 / ext3 defaults 1 1
```



```
Expanding /dev/sda1 using the 'resize2fs' method ...
```

```
Resize operation completed with no errors. Before deleting the old disk, carefully check that the resized disk boots and works correctly.
```

4. 确认是否已正确调整映像大小：

```
> virt-filesystems --long --parts --blkdevs -h -a outdisk.img
```

Name	Type	MBR	Size	Parent
/dev/sda1	partition	83	32G	/dev/sda
/dev/sda	device	-	32G	-

5. 使用新磁盘映像启动 VM Guest，确认其运行正常，然后删除旧映像。

21.3.5 其他 virt-* 工具

某些 guestfs 工具可以简化管理任务 — 例如查看和编辑文件，或者获取有关虚拟机的信息。

21.3.5.1 virt-filesystems

此工具用于报告有关磁盘映像或虚拟机中的文件系统、分区和逻辑卷的信息。

```
> virt-filesystems -l -a sles.qcow2
```

Name	Type	VFS	Label	Size	Parent
/dev/sda1	filesystem	ext3	-	17178820608	-

21.3.5.2 virt-ls

virt-ls 可列出虚拟机或磁盘映像中的文件名、文件大小、校验和、扩展属性等信息。可以指定多个目录名，在这种情况下，每个目录的输出将会串联起来。要列出某个 libvirt Guest 中的目录，请使用 **-d** 选项指定 Guest 名称。对于磁盘映像，请使用 **-a** 选项。

```
> virt-ls -h -lR -a sles.qcow2 /var/log/
```

d 0755	776	/var/log
--------	-----	----------

```

- 0640          0 /var/log/NetworkManager
- 0644        23K /var/log/Xorg.0.log
- 0644        23K /var/log/Xorg.0.log.old
d 0700        482 /var/log/YaST2
- 0644        512 /var/log/YaST2/_dev_vda
- 0644         59 /var/log/YaST2/arch.info
- 0644        473 /var/log/YaST2/config_diff_2017_05_03.log
- 0644        5.1K /var/log/YaST2/curl_log
- 0644        1.5K /var/log/YaST2/disk_vda.info
- 0644        1.4K /var/log/YaST2/disk_vda.info-1
[...]
```

21.3.5.3 **virt-cat**

virt-cat 命令行工具用于显示指定虚拟机（或磁盘映像）中存在的文件的内容。可以指定多个文件名，在这种情况下，这些文件名会串联到一起。指定每个文件名时，必须提供以根目录 (/) 开头的绝对路径。

```

> virt-cat -a sles.qcow2 /etc/fstab
devpts /dev/pts devpts mode=0620,gid=5 0 0
proc   /proc     proc   defaults      0 0
```

21.3.5.4 **virt-df**

virt-df 命令行工具用于显示虚拟机文件系统上的可用空间。与其他工具不同，它不仅会显示分配给虚拟机的磁盘大小，而且还会查看磁盘映像内部以显示使用的空间量。

```

> virt-df -a sles.qcow2
```

Filesystem	1K-blocks	Used	Available	Use%
sles.qcow2:/dev/sda1	16381864	520564	15022492	4%

21.3.5.5 **virt-edit**

virt-edit 命令行工具能够编辑驻留在指定虚拟机（或磁盘映像）中的文件。

21.3.5.6 **virt-tar-in/out**

virt-tar-in 可将未压缩的 TAR 归档解压缩到虚拟机磁盘映像或指定的 libvirt 域中。**virt-tar-out** 可将虚拟机磁盘映像目录打包成 TAR 归档。

```
> virt-tar-out -a sles.qcow2 /home homes.tar
```

21.3.5.7 **virt-copy-in/out**

virt-copy-in 可将本地磁盘中的文件和目录复制到虚拟机磁盘映像或指定的 libvirt 域中。**virt-copy-out** 可从虚拟机磁盘映像或指定的 libvirt 域中复制文件和目录。

```
> virt-copy-in -a sles.qcow2 data.tar /tmp/
> virt-ls -a sles.qcow2 /tmp/
.ICE-unix
.X11-unix
data.tar
```

21.3.5.8 **virt-log**

virt-log 可显示指定的 libvirt 域、虚拟机或磁盘映像的日志文件。如果安装了软件包 **guestfs-winsupport**，则 virt-log 还可显示 Windows 虚拟机磁盘映像的事件日志。

```
> virt-log -a windows8.qcow2
<?xml version="1.0" encoding="utf-8" standalone="yes" ?>
<Events>
<Event xmlns="http://schemas.microsoft.com/win/2004/08/events/
event"><System><Provider Name="EventLog"></Provider>
<EventID Qualifiers="32768">6011</EventID>
<Level>4</Level>
<Task>0</Task>
<Keywords>0x0080000000000000</Keywords>
<TimeCreated SystemTime="2014-09-12 05:47:21"></TimeCreated>
<EventRecordID>1</EventRecordID>
<Channel>System</Channel>
```



```

<Computer>windows-uj49s6b</Computer>
<Security UserID=""></Security>
</System>
<EventData><Data><string>WINDOWS-UJ49S6B</string>
<string>WIN-KG190623QG4</string>
</Data>
<Binary></Binary>
</EventData>
</Event>

...

```

21.3.6 **guestfish**

guestfish 是用于检查和修改虚拟机文件系统的外壳和命令行工具。它使用 libguestfs 并公开了 guestfs API 的所有功能。

用法示例：

```

> guestfish -a disk.img <<EOF
run
list-filesystems
EOF

```

guestfish

Welcome to guestfish, the guest filesystem shell for editing virtual machine filesystems and disk images.

Type: 'help' for help on commands
 'man' to read the manual
 'quit' to quit the shell

```

><fs> add sles.qcow2
><fs> run
><fs> list-filesystems
/dev/sda1: ext3
><fs> mount /dev/sda1 /

```

```
cat /etc/fstab
devpts /dev/pts          devpts mode=0620,gid=5 0 0
proc   /proc              proc   defaults              0 0
sysfs  /sys               sysfs  noauto                 0 0
debugfs /sys/kernel/debug debugfs noauto                 0 0
usbfs  /proc/bus/usb        usbfs  noauto                 0 0
tmpfs  /run              tmpfs  noauto                 0 0
/dev/disk/by-id/ata-QEMU_HARDDISK_QM00001-part1 / ext3 defaults 1 1
```

21.3.7 将物理机转换为 KVM Guest

Libguestfs 提供了可帮助您将 Xen 虚拟机或物理机转换为 KVM Guest 的工具。第 18 章 “Xen 到 KVM 的迁移指南” 中介绍了从 Xen 转换到 KVM 的方案。下一节将介绍一个特殊用例：将裸机转换为 KVM 计算机。

SUSE Linux Enterprise Server 尚不支持将物理机转换为 KVM 计算机。此功能仅发布为技术预览版。

转换某个物理机需要收集有关该物理机的信息，并将这些信息传输到转换服务器。要实现此目的，需在计算机上运行一个使用 **virt-p2v** 和 KIWI NG 工具准备的实时系统。

过程 21.2：使用 VIRT-P2V

1. 使用以下命令安装所需的软件包：

```
> sudo zypper in virt-p2v kiwi-desc-isoboot
```



注意

这些步骤将会阐述如何创建一个 ISO 映像用于创建可引导 DVD。或者，您也可以改为创建 PXE 引导映像；有关使用 KIWI NG 构建 PXE 映像的详细信息，请参见 **man virt-p2v-make-kiwi**。

2. 创建 KIWI NG 配置：

```
> virt-p2v-make-kiwi -o /tmp/p2v.kiwi
```

-o 定义在何处创建 KIWI NG 配置。

3. 如果需要，请编辑生成的配置中的 `config.xml` 文件。例如，在 `config.xml` 中调整实时系统的键盘布局。

4. 使用 **kiwi** 构建 ISO 映像：

```
> kiwi --build /tmp/p2v.kiwi ❶ \  
    -d /tmp/build ❷ \  
    --ignore-repos \  
    --add-repo http://URL_TO_REPOSITORIES ❸ \  
    --type iso
```

- ❶ 存放上一步中生成的 KIWI NG 配置的目录。
- ❷ KIWI NG 将用于存放生成的 ISO 映像及其他中间构建结果的目录。
- ❸ 使用 `zypper lr -d` 找到的软件包储存库的 URL。
请对每个储存库使用一个 `--add-repo` 参数。

5. 在 DVD 或 USB 记忆棒上刻录 ISO。使用此类媒体引导要转换的计算机。
6. 系统启动后，请输入 **转换服务器** 的连接详情。此服务器是装有 `virt-v2v` 软件包的计算机。
如果网络设置比 DHCP 客户端更复杂，请单击 **配置网络** 按钮打开 YaST 网络配置对话框。
单击 **测试连接** 按钮以转到向导的下一页。
7. 选择要转换的磁盘和网络接口，并定义 VM 数据，例如分配的 CPU 数量和内存容量，以及虚拟机名称。



注意

如果不定义这些数据，创建的磁盘映像默认将采用 **raw** 格式。可以在输出格式字段中输入所需格式来更改此格式。

可通过两种方法生成虚拟机：使用 **local** 或 **libvirt** 输出。第一种方法将虚拟机磁盘映像和配置放到输出存储字段中定义的路径。然后，可以通过 **virsh** 使用这些内容来定义由 libvirt 处理的新 Guest。第二种方法使用放在输出存储字段所定义的池中的磁盘映像，来创建由 libvirt 处理的新 Guest。

单击 **开始转换** 开始该过程。

21.4 查错

21.4.1 Btrfs 相关的问题

对采用 Btrfs 根分区（在 SUSE Linux Enterprise Server 中为默认设置）的映像使用 guestfs 工具时，可能会显示以下错误消息：

```
> virt-ls -a /path/to/sles12sp2.qcow2 /
virt-ls: multi-boot operating systems are not supported

If using guestfish '-i' option, remove this option and instead
use the commands 'run' followed by 'list-file systems'.
You can then mount file systems you want by hand using the
'mount' or 'mount-ro' command.

If using guestmount '-i', remove this option and choose the
filesystem(s) you want to see by manually adding '-m' option(s).
Use 'virt-file systems' to see what file systems are available.

If using other virt tools, multi-boot operating systems won't work
with these tools. Use the guestfish equivalent commands
(see the virt tool manual page).
```

发生此问题通常是因为 Guest 中存在多个快照。在此情况下，guestfs 不知道要引导哪个快照。要强制使用某个快照，请如下所示使用 `-m` 参数：

```
> virt-ls -m /dev/sda2::subvol=@/.snapshots/2/snapshot -a /path/to/
sles12sp2.qcow2 /
```

21.4.2 环境

在 libguestfs 设备中对问题进行查错时，可以使用环境变量 **LIBGUESTFS_DEBUG=1** 来启用调试消息。要以类似于 guestfish 命令的格式输出每个命令/API 调用，请使用环境变量 **LIBGUESTFS_TRACE=1**。

21.4.3 libguestfs-test-tool

libguestfs-test-tool 是一个测试程序，用于检查基本 libguestfs 功能是否正常工作。它会列显 guestfs 环境的大量诊断消息和细节，然后创建一个测试映像并尝试将其启动。如果它成功运行到完成时，则测试将近结束时应该会显示以下消息：

```
===== TEST FINISHED OK =====
```

21.5 更多信息

- libguestfs.org (<https://libguestfs.org>) ↗
- [libguestfs 常见问题](https://libguestfs.org/guestfs-faq.1.html) (<https://libguestfs.org/guestfs-faq.1.html>) ↗

22 QEMU Guest 代理

QEMU Guest 代理 (GA) 在 VM Guest 中运行，使 VM 主机服务器能够通过 [libvirt](#) 在 Guest 操作系统中运行命令。它支持许多功能 — 例如，获取有关 Guest 文件系统的细节、冻结和解冻文件系统，或者挂起或重引导 Guest。

QEMU GA 包含在 [qemu-guest-agent](#) 软件包中，默认已在 KVM 虚拟机上安装、配置并激活。

QEMU GA 安装在 Xen 虚拟机中，但默认未激活。尽管可以将 QEMU GA 与 Xen 虚拟机配合使用，但无法使用如下所述的适用于 KVM 虚拟机的 [libvirt](#) 命令来实现集成。要将 QEMU GA 与 Xen 配合使用，必须将一个通道设备添加到 VM Guest 配置。该通道设备包含 VM 主机服务器上用来与 QEMU GA 通讯的 Unix 域套接字路径。

```
<channel type='unix'>
  <source mode='bind' path='/example/path'/>
  <target type='xen' name='org.qemu.guest_agent.0'/>
</channel>
```

22.1 运行 QEMU GA 命令

QEMU GA 包含的许多内置命令没有直接对应的 [libvirt](#) 命令。请参见第 22.4 节 “更多信息” 查看完整列表。您可以使用 [libvirt](#) 的通用命令 [qemu-agent-command](#) 来运行所有 QEMU GA 命令：

```
virsh qemu-agent-command DOMAIN_NAME '{"execute":"QEMU_GA_COMMAND"}'
```

例如：

```
> sudo virsh qemu-agent-command sle15sp2 '{"execute":"guest-info"}' --pretty
{
  "return": {
    "version": "4.2.0",
    "supported_commands": [
      {
        "enabled": true,
        "name": "guest-get-osinfo",
```

```
"success-response": true
},
[...]
```

22.2 需要 QEMU GA 的 **virsh** 命令

有多个 **virsh** 命令需要 QEMU GA 才能实现其功能。例如，以下命令：

virsh guestinfo

从 Guest 的角度列显有关该 Guest 的信息。

virsh guestvcpus

从 Guest 的角度查询或更改虚拟 CPU 的状态。

virsh set-user-password

为 Guest 中的用户帐户设置口令。

virsh domfsinfo

显示正在运行的域中挂载的文件系统列表。

virsh dompm_suspend

挂起正在运行的 Guest。

22.3 增强 **libvirt** 命令

如果在 Guest 中启用了 QEMU GA，多个 **virsh** 子命令在以**代理**模式运行时，其功能会得到增强。以下列表仅包含其中某些子命令的示例。有关完整列表，请参见 **virsh** 手册页并搜索 **agent** 字符串。

virsh shutdown --mode agent和**virsh reboot --mode agent**

这种关机或重引导方法类似于 ACPI 方法，可让 Guest 为下次运行保持干净状态。

virsh domfsfreeze和**virsh domfsthaw**

指示 Guest 将其文件系统保持静止状态 — 刷新缓存中的所有 I/O 操作并将卷保持一致状态，以便在重新挂载卷时无需进行任何检查。

virsh setvcpus --guest

更改分配给 Guest 的 CPU 数量。

virsh domifaddr --source agent

在 QEMU GA 中查询 Guest 的 IP 地址。

virsh vcpucount --guest

从 Guest 的角度列显有关虚拟 CPU 计数的信息。

22.4 更多信息

- <https://www.qemu.org/docs/master/interop/qemu-ga-ref.html> 上提供了 QEMU GA 支持的命令的完整列表。
- virsh 手册页 (man 1 virsh) 包含支持 QEMU GA 接口的命令的说明。

23 软件 TPM 模拟器

23.1 简介

可信平台模块 (TPM) 是使用加密密钥保护硬件的加密处理器。对于使用 TPM 开发安全功能的开发人员而言，软件 TPM 模拟器是一种便利的解决方案。与硬件 TPM 设备相比，该模拟器对可以访问它的 Guest 数不设限制。另外，在 TPM 1.2 和 2.0 版本之间切换也很简单。QEMU 支持 swtpm 软件包中包含的软件 TPM 模拟器。

23.2 先决条件

您需要先安装 libvirt 虚拟化环境，然后才能安装并使用软件 TPM 模拟器。请参见第 6.2 节“安装虚拟化组件”并安装所提供的虚拟化解决方案之一。

23.3 安装

要使用软件 TPM 模拟器，请安装 swtpm 软件包：

```
> sudo zypper install swtpm
```

23.4 将 swtpm 与 QEMU 配合使用

swtpm 提供三种类型的接口：socket、chardev 和 cuse。以下过程重点介绍 **socket** 接口。

1. 在 VM 目录（例如 /var/lib/libvirt/qemu/sle15sp3）中创建 mytpm0 目录用于存储 TPM 状态：

```
> sudo mkdir /var/lib/libvirt/qemu/sle15sp3/mytpm0
```

2. 开始 swtmp。它将创建 QEMU 可以使用的套接字文件，例如 /var/lib/libvirt/qemu/sle15sp3：

```
> sudo swtpm socket
--tpmstate dir=/var/lib/libvirt/qemu/sle15sp3/mytpm0 \
--ctrl type=unixio,path=/var/lib/libvirt/qemu/sle15sp3/mytpm0/swtpm-sock
\
--log level=20
```



提示：TPM 版本 2.0

默认情况下，**swtpm** 会启动 TPM 版本 1.2 模拟器，并将此模拟器的状态存储在 **tpm-00.permall** 目录中。要创建 TPM 2.0 实例，请运行以下命令：

```
> sudo swtpm socket
--tpm2
--tpmstate dir=/var/lib/libvirt/qemu/sle15sp3/mytpm0 \
--ctrl type=unixio,path=/var/lib/libvirt/qemu/sle15sp3/mytpm0/
swtpm-sock \
--log level=20
```

TPM 2.0 状态存储在 **tpm2-00.permall** 目录中。

3. 将以下命令行参数添加到 **qemu-system-ARCH** 命令：

```
> qemu-system-x86_64 \
[...]
-chardev socket,id=chrtpm,path=/var/lib/libvirt/qemu/sle15sp3/mytpm0/swtpm-
sock \
-tpmdev emulator,id=tpm0,chardev=chrtpm \
-device tpm-tis,tpmdev=tpm0
```

4. 运行以下命令，以校验在 Guest 中是否可以使用 TPM 设备：

```
> tpm_version
TPM 1.2 Version Info:
Chip Version:      1.2.18.158
Spec Level:        2
Errata Revision:   3
TPM Vendor ID:     IBM
```

TPM Version:	01010000
Manufacturer Info:	49424d00

23.5 将 swtpm 与 libvirt 配合使用

要将 swtpm 与 libvirt 搭配使用，请将以下 TPM 设备添加到 Guest XML 规范中：

```
<devices>
  <tpm model='tpm-tis'>
    <backend type='emulator' version='2.0' />
  </tpm>
</devices>
```

libvirt 将自动为 Guest 启动 swtpm。您无需提前手动启动 swtpm。相应的 permall 文件是在 /var/lib/libvirt/swtpm/VM_UUID 中创建的。

23.6 使用 OVMF 固件进行 TPM 测量

如果 Guest 使用开放虚拟机固件 (OVMF)，它将使用 TPM 来测量组件。可以在 /sys/kernel/security/tpm0/binary_bios_measurements 中找到事件日志。

23.7 资源

- 维基百科对 TPM 进行了全面介绍，网址为 https://en.wikipedia.org/wiki/Trusted_Platform_Module。
- 第 6 章 “安装虚拟化组件” 中介绍了如何在 SUSE Linux Enterprise Server 上配置特定的虚拟化环境。
- swtpm 的手册页 (**man 8 swtpm**) 中提供了其用法细节。
- <https://libvirt.org/formatdomain.html#elementsTpm> 上提供了 TPM 的详细 libvirt 规范
- 第 10.3.1 节 “高级 UEFI 配置” 中介绍了如何使用 OVMF 启用 UEFI 固件。

24 创建 VM Guest 的崩溃转储

24.1 简介

每当 VM 崩溃时，有用的做法是收集 VM 内存的核心转储以进行调试和分析。对于物理机，Kexec 和 Kdump 会负责收集崩溃转储。对于虚拟机，如何收集崩溃转储取决于 Guest 是全虚拟化 (FV) 还是半虚拟化 (PV) 计算机。

24.2 为全虚拟化计算机创建崩溃转储

要查看 FV 计算机的崩溃转储，请使用适用于物理机的相同过程 — 使用 Kexec 和 Kdump。

24.3 为半虚拟化计算机创建崩溃转储

与在 FV 中不同，Kexec/Kdump 在半虚拟化计算机中不起作用。PV Guest 的崩溃转储必须由主机工具堆栈执行。如果将 **xl** 工具堆栈用于 Xen domU，**xl dump-core** 命令将生成转储。对于基于 libvirt 的 VM Guest，**virsh dump** 命令可提供相同的功能。

您可以使用 VM Guest 配置中的 on_crash 设置来配置核心转储自动收集。此设置将告知主机工具堆栈在 VM Guest 遇到崩溃时该如何处理。**xl** 和 libvirt 中的默认值均为 destroy。可自动收集核心转储的有用选项为 coredump-destroy 和 coredump-restart。

24.4 附加信息

- 第 1.3 节 “虚拟化模式” 中介绍了全虚拟化与半虚拟化虚拟机之间的差别。
- 《系统分析和微调指南》，第 20 章 “Kexec 和 Kdump” 中提供了有关 Kexec/Kdump 机制的详细信息。

- 有关 **xl** 配置语法的详细信息，请参见 `xl.cfg` 手册页 (**man 5 xl.cfg**)。
- 有关 `libvirt` XML 设置的细节，请参见 <https://libvirt.org/formatdomain.html#events-configuration> 。

IV 使用 Xen 管理虚拟机

- 25 设置虚拟机主机 247
- 26 虚拟网络 260
- 27 管理虚拟化环境 268
- 28 Xen 中的块设备 274
- 29 虚拟化：配置选项和设置 278
- 30 管理任务 288
- 31 XenStore：在域之间共享的配置数据库 297
- 32 使用 Xen 作为高可用性虚拟化主机 303
- 33 Xen：将半虚拟 (PV) Guest 转换为全虚拟 (FV/HVM) Guest 306

25 设置虚拟机主机

本节介绍如何将 SUSE Linux Enterprise Server 15 SP7 设置为虚拟机主机，并作为虚拟机主机使用。

Dom0 的硬件要求通常与 SUSE Linux Enterprise Server 操作系统的硬件要求相同。应该添加额外的 CPU、磁盘、内存和网络资源才能满足规划的所有 VM Guest 系统的需求。



提示：资源

请记住，如果 VM Guest 系统在更快的处理器上运行并且可以访问更多系统内存，其性能就会更好，这与物理机一样。

虚拟机主机要求安装多个软件包及其依赖项。要安装所需的全部软件包，请运行 YaST 软件管理，选择视图 > 软件集，然后选择 Xen 虚拟机主机服务器进行安装。也可以在 YaST 中使用模块虚拟化 > 安装管理程序和工具来执行安装。

安装 Xen 软件后，重新启动计算机，并在引导屏幕上选择新添加的具有 Xen 内核的选项。

通过更新通道可使用更新。为了确保安装最新的更新，请在完成安装后运行 YaST 在线更新。

25.1 最佳实践和建议

在主机上安装和配置 SUSE Linux Enterprise Server 操作系统时，请遵循以下最佳实践和建议：

- 如果该主机应始终作为 Xen 主机运行，请运行 YaST 系统 > 引导加载程序，并选中 Xen 引导项作为默认引导项。

- 在 YaST 中单击系统 > 引导加载程序。
 - 将默认引导更改为 Xen 标签，然后单击设置为默认值。
 - 单击完成。
- 为获得最佳性能，请仅在虚拟机主机上安装虚拟化所需的应用程序和进程。
 - 如果您打算使用挂接到 Xen 主机的看门狗设备，请每次仅使用一个设备。建议使用提供实际硬件集成的驱动程序，而不要使用通用软件驱动程序。



注意：硬件监控

Dom0 内核以虚拟化模式运行，因此 `irqbalance` 或 `lscpu` 等工具不会反映真实的硬件特征。



重要：Xen 不支持可信引导

Xen 不支持可信引导 (Tboot)。为了确保 Xen 主机可正确引导，请在 GRUB 2 配置对话框中校验启用可信引导支持选项是否已取消选择。

25.2 管理 Dom0 内存

在以前的 SUSE Linux Enterprise Server 版本中，Xen 主机的默认内存分配模式是将所有主机物理内存都分配给 Dom0，并启用自动气球式调节功能。当有其他域启动后，内存会自动从 Dom0 进行气球式调节。此行为总是容易出错，因此强烈建议将其禁用。从 SUSE Linux Enterprise Server 15 SP1 开始，默认已禁用自动气球式调节，并会为 Dom0 分配 10% 的主机物理内存加 1 GB。例如，在物理内存大小为 32 GB 的主机上，将为 Dom0 分配 4.2 GB 内存。我们仍支持并建议在 `/etc/default/grub` 中使用 `dom0_mem` Xen 命令行选项。可以通过将 `dom0_mem` 设置为主机物理内存大小，并在 `/etc/xen/xl.conf` 中启用 `autoballoon` 设置，来恢复旧行为。



警告：Dom0 内存不足

为 Dom0 预留的内存容量取决于主机上运行的 VM Guest 数量，因为 Dom0 会为每个 VM Guest 提供后端网络和磁盘 I/O 服务。计算 Dom0 内存分配时，还应考虑到 Dom0 中运行的其他工作负载。应该像确定任何其他虚拟机的内存大小一样来确定 Dom0 的内存大小。

25.2.1 设置 Dom0 内存分配

1. 确定需要为 Dom0 分配的内存。
2. 在 Dom0 中，键入 `xl info` 以查看计算机上可用的内存量。可以使用 `xl list` 命令确定当前为 Dom0 分配的内存。
3. 编辑 `/etc/default/grub` 并调整 `GRUB_CMDLINE_XEN` 选项，使其包含 `dom0_mem=MEM_AMOUNT`。将 `MEM_AMOUNT` 替换为要分配给 Dom0 的最大内存量。添加 `K`、`M` 或 `G` 以指定大小单位。例如：

```
GRUB_CMDLINE_XEN="dom0_mem=2G"
```

4. 重新启动计算机以应用更改。



提示

有关 Xen 相关引导配置选项的更多细节，请参见《管理指南》，第 18 章“引导加载程序 GRUB 2”，第 18.2.2 节“文件 `/etc/default/grub`”。



警告：Xen Dom0 内存

对 GRUB 2 中的 Xen 超级管理程序使用 XL 工具堆栈和 `dom0_mem=` 选项时，需在 `etc/xen/xl.conf` 中禁用 `xl autoballoon`。否则，启动 VM 时将会失败，并出现有关无法压缩 Dom0 内存气球的错误。因

此，如果您为 Xen 指定了 **dom0_mem=** 选项，请在 `xl.conf` 中添加 **autoballoon=0**。另请参见 [Xen dom0 内存 \(https://wiki.xen.org/wiki/Xen_Best_Practices#Xen_dom0_dedicated_memory_and_preventing_dom0_memory_balloonin](https://wiki.xen.org/wiki/Xen_Best_Practices#Xen_dom0_dedicated_memory_and_preventing_dom0_memory_balloonin)

25.3 全虚拟化 Guest 中的网卡

在全虚拟化 Guest 中，默认网卡是一个模拟的 Realtek 网卡。不过，您也可以使用分离式网络驱动程序来管理 Dom0 与 VM Guest 之间的通讯。默认情况下，这两个接口都会呈现给 VM Guest，因为某些操作系统的驱动程序要求这两个接口都存在。

使用 SUSE Linux Enterprise Server 时，默认只有半虚拟化网卡可供 VM Guest 使用。可用的网络选项如下：

模拟

要使用模拟 Realtek 网卡之类的模拟网络接口，请在域 xl 配置的 `vif` 设备部分指定 `type=ioemu`。示例配置如下所示：

```
vif = [ 'type=ioemu,mac=00:16:3e:5f:48:e4,bridge=br0' ]
```

`xl.conf` 手册页 **man 5 xl.conf** 中提供了有关 xl 配置的更多细节。

半虚拟化

如果指定 `type=vif` 但不指定型号或类型，将使用半虚拟化网络接口：

```
vif = [ 'type=vif,mac=00:16:3e:5f:48:e4,bridge=br0,backen=0' ]
```

模拟和半虚拟化

如果要为管理员提供上述两个选项，只需同时指定类型和型号即可。xl 配置如下所示：

```
vif = [ 'type=ioemu,mac=00:16:3e:5f:48:e4,model=rtl8139,bridge=br0' ]
```

在这种情况下，应在 VM Guest 上禁用其中一个网络接口。

25.4 启动虚拟机主机

如果正确安装了虚拟化软件，计算机引导时会显示 GRUB 2 引导加载程序，且其菜单中会包含 Xen 选项。选择此选项即可启动虚拟机主机。



警告

引导 Xen 系统时，dom0 的 `/var/log/messages` 日志文件或 `systemd` 日记中可能会出现如下所示的错误消息：

```
isst_if_mbox_pci: probe of 0000:ff:1e.1 failed with error -5
isst_if_pci: probe of 0000:fe:00.1 failed with error -5
```

请忽略这些错误，因为它们不会产生问题，之所以发生这些错误，是因为 ISST 驱动程序无法为虚拟机提供任何电源或频率调节功能。



注意：Xen 和 Kdump

在 Xen 中，超级管理程序会管理内存资源。如果您需要为 Dom0 中的恢复内核预留系统内存，需由超级管理程序预留此内存。因此，请务必在 `/etc/default/grub` 文件的 `GRUB_CMDLINE_XEN_DEFAULT` 变量中添加 `crashkernel=size`，然后保存文件并运行以下命令：

```
> sudo grub2-mkconfig -o /boot/grub2/grub.cfg
```

有关 `crashkernel` 参数的详细信息，请参见《系统分析和微调指南》，第 20 章 “Kexec 和 Kdump”，第 20.4 节 “计算 `crashkernel` 分配大小”。

如果 GRUB 2 菜单中不包含 Xen 选项，请检查安装步骤，并校验是否已更新 GRUB 2 引导加载程序。如果安装时未选择 Xen 软件集，安装完成后请运行 YaST 软件管理，选择软件集过滤器，然后选择 Xen 虚拟机主机服务器进行安装。

引导超级管理程序后，Dom0 虚拟机将会启动并显示其图形桌面环境。如果您未安装图形桌面，则会显示命令行环境。



提示：图形问题

有时可能会发生图形系统无法正常工作的问题。在这种情况下，请将 `vga=ask` 添加到引导参数。要激活永久设置，请使用 `vga=mode-0x???`，其中 `???` 的计算方式为 `0x100 + https://en.wikipedia.org/wiki/VESA\_BIOS\_Extensions` 中所述的 VESA 模式，例如 `vga=mode-0x361`。

在开始安装虚拟 Guest 之前，请确保系统时间正确。为此，请在控制域上配置 NTP（网络时间协议）：

1. 在 YaST 中选择网络服务 > NTP 配置。
2. 选择在引导期间自动启动 NTP 守护程序的选项。提供现有 NTP 时间服务器的 IP 地址，然后单击完成。



注意：虚拟 Guest 上的时间服务

硬件时钟并不精确。所有新式操作系统都会尝试通过一个额外的时间源来更正系统时间（对比硬件时间）。要使所有 VM Guest 系统上的时间正确，请也在每个相应 Guest 上激活网络时间服务，或确保 Guest 使用主机的系统时间。有关 SUSE Linux Enterprise Server 中的 `Independent Wallclocks` 的详细信息，请参见第 20.2 节“Xen 虚拟机时钟设置”。

有关管理虚拟机的详细信息，请参见第 27 章“管理虚拟化环境”。

25.5 PCI 直通

为了充分利用 VM Guest 系统，有时需要将特定的 PCI 设备分配给专用的域。如果使用的是全虚拟化 Guest，那么仅当系统的芯片组支持并且已在 BIOS 中激活此功能时，此功能才可用。

AMD* 和 Intel* 都提供此功能。对于 AMD 计算机，该功能称为 `IOMMU`。在 Intel 术语中称为 `VT-d`。请注意，具备 Intel-VT 技术还不足以对全虚拟化 Guest 使用此功能。为确保您的计算机支持此功能，需专门要求您的供应商提供一个支持 PCI 直通的系统。

限制

- 有些图形驱动程序使用高度优化的方法来访问 DMA。这种方法不受支持，因此使用显卡可能会存在问题。
- 访问 **PCIe** 网桥后面的 PCI 设备时，必须将所有 PCI 设备都分配到单个 Guest。这项限制不适用于 **PCIe** 设备。
- 具有专用 PCI 设备的 Guest 无法实时迁移到其他主机。

PCI 直通的配置要兼顾两个方面。首先，在引导时必须告知超级管理程序应使某个 PCI 设备可供重新分配。其次，必须将该 PCI 设备分配给 VM Guest。

25.5.1 配置超级管理程序以使用 PCI 直通

1. 选择要重分配给 VM Guest 的设备。为此，请运行 **lspci -k**，并读取设备编号以及分配给该设备的原始模块的名称：

```
06:01.0 Ethernet controller: Intel Corporation Ethernet Connection I217-LM
(rev 05)
    Subsystem: Dell Device 0617
    Kernel driver in use: e1000e
    Kernel modules: e1000e
```

在本例中，PCI 编号为 (06:01.0)，相关的内核模块为 e1000e。

2. 指定模块依赖项以确保 xen_pciback 是用于控制设备的第一个模块。添加包含以下内容的 /etc/modprobe.d/50-e1000e.conf 文件：

```
install e1000e /sbin/modprobe xen_pciback ; /sbin/modprobe \
--first-time --ignore-install e1000e
```

3. 指示 xen_pciback 模块使用 hide 选项来控制设备。编辑或创建包含以下内容的 /etc/modprobe.d/50-xen-pciback.conf 文件：

```
options xen_pciback hide=(06:01.0)
```

4. 重新启动系统。

5. 使用以下命令检查该设备是否在可分配设备列表中

```
xl pci-assignable-list
```

25.5.1.1 通过 xl 进行动态分配

为了避免重新启动主机系统，您可以利用通过 xl 进行的动态分配来使用 PCI 直通。

首先确保 Dom0 中已加载 pciback 模块：

```
> sudo modprobe pciback
```

然后使用 **xl pci-assignable-add** 使设备可供分配。例如，要使设备 **06:01.0** 可供 Guest 使用，请运行以下命令：

```
> sudo xl pci-assignable-add 06:01.0
```

25.5.2 将 PCI 设备分配给 VM Guest 系统

可通过多种方法来使 PCI 设备专用于某个 VM Guest：

安装时添加该设备：

在安装期间，在配置文件中添加 pci 行：

```
pci=['06:01.0']
```

将 PCI 设备热插入到 VM Guest 系统

可以使用 xl 命令即时添加或去除 PCI 设备。要将编号为 06:01.0 的设备添加到名为 sles12 的 Guest，请使用：

```
xl pci-attach sles12 06:01.0
```

将 PCI 设备添加到 Xend

要将设备永久添加到 Guest，请在 Guest 配置文件中添加以下代码段：

```
pci = [ '06:01.0,power_mgmt=1,permissive=1' ]
```

将 PCI 设备分配给 VM Guest 后，Guest 系统必须负责处理此设备的配置和设备驱动程序。

25.5.3 VGA 直通

Xen 4.0 和更高版本支持在全虚拟化 VM Guest 上实现 VGA 图形适配器直通。Guest 可以全面控制提供高性能全 3D 和视频加速的图形适配器。

限制

- VGA 直通功能与 PCI 直通类似，因此也需要主板芯片组和 BIOS 提供 IOMMU（或 Intel VT-d）支持。
- 只有主图形适配器（打开计算机电源时使用的图形适配器）能够与 VGA 直通搭配使用。
- 仅支持对全虚拟化 Guest 使用 VGA 直通。不支持半虚拟 (PV) Guest。
- 不能在多个使用 VGA 直通的 VM Guest 之间共享显卡 — 显卡只能供一个 Guest 专用。

要启用 VGA 直通，请在全虚拟化 Guest 配置文件中添加以下设置：

```
gfx_passthru=1
pci=['yy:zz.n']
```

其中，yy:zz.n 是使用 **lspci -v** 在 Dom0 上找到的 VGA 图形适配器的 PCI 控制器 ID。

25.5.4 查错

在某些情况下，安装 VM Guest 期间可能会出现問題。本节将说明多个已知问题及其解决方法。

系统在引导期间挂起

软件 I/O 转换缓冲区会提前在引导进程中分配一大块低速内存。如果内存请求超出了缓冲区大小，可能会导致引导进程挂起。要检查是否存在这种情况，请切换到控制台 10，并检查其输出是否包含如下所示的消息

```
kernel: PCI-DMA: Out of SW-IOMMU space for 32768 bytes at device
000:01:02.0
```

在这种情况下，您需要增加 `swiotlb` 的大小。在 Dom0 的命令行中添加 `swiotlb=VALUE`（其中 `VALUE` 指定为 slab 项数）。可以通过增大或减小该数字来找到适合计算机的最佳大小。



注意：为 PV Guest 启用 swiotlb

要在 PV Guest 上正常进行 PCI 设备的 DMA 访问，必须指定 `swiotlb=force` 内核参数。有关 IOMMU 和 `swiotlb` 选项的详细信息，请参见软件包 `kernel-source` 中的 `boot-options.txt` 文件。

25.5.5 更多信息

互联网上的一些资源提供了有关 PCI 直通的有趣信息：

- https://wiki.xenproject.org/wiki/VTd_HowTo ↗
- <https://software.intel.com/en-us/articles/intel-virtualization-technology-for-directed-io-vt-d-enhancing-intel-platforms-for-efficient-virtualization-of-io-devices/> ↗
- https://support.amd.com/TechDocs/48882_IOMMU.pdf ↗

25.6 USB 直通

可通过两种方法将单个主机 USB 设备直通到 Guest。第一种方法是使用模拟的 USB 设备控制器，第二种方法是使用 PVUSB。

25.6.1 标识 USB 设备

在将 USB 设备直通到 VM Guest 之前，需要先在 VM 主机服务器上标识该设备。使用 `lsusb` 命令列出主机系统上的 USB 设备：

```
# lsusb
Bus 001 Device 001: ID 1d6b:0002 Linux Foundation 2.0 root hub
```



```
Bus 002 Device 003: ID 0461:4d15 Primax Electronics, Ltd Dell Optical Mouse
Bus 002 Device 001: ID 1d6b:0001 Linux Foundation 1.1 root hub
```

例如，要直通 Dell 鼠标，请以 `vendor_id:device_id` (0461:4d15) 格式指定设备标记，或以 `bus.device` (2.3) 格式指定总线地址。请记得去除前导零，否则 `xl` 会将数字解释为八进制值。

25.6.2 模拟的 USB 设备

在模拟的 USB 中，设备模型 (QEMU) 会向 Guest 呈现模拟的 USB 控制器。然后，将通过 Dom0 控制 USB 设备，而 USB 命令将在 VM Guest 与主机 USB 设备之间转换。此方法仅适用于全虚拟化域 (HVM)。

使用 `usb=1` 选项启用模拟的 USB 集线器。然后使用 `host:USBID` 在配置文件中指定设备列表中的设备以及其他模拟的设备。例如：

```
usb=1
usbdevice=['tablet','host:2.3','host:0424:460']
```

25.6.3 半虚拟化 PVUSB

PVUSB 是以高性能方式实现从 Dom0 到虚拟化 Guest 的 USB 直通的新方法。借助 PVUSB，可以通过两种方式将 USB 设备添加到 Guest：

- 在创建域时通过配置文件添加
- 当 VM 正在运行时通过热插入方式添加

PVUSB 使用半虚拟化前端和后端接口。PVUSB 支持 USB 1.1 和 USB 2.0，适用于 PV 和 HVM Guest。要使用 PVUSB，Guest 操作系统中需要有 USB 前端，并且 Dom0 中需要有 USB 后端或者 QEMU 中需要有 USB 后端。在 SUSE Linux Enterprise Server 上，QEMU 随附了 USB 后端。

从 Xen 4.7 开始引入了 `xl` PVUSB 支持和热插入支持。

在配置文件中，使用 `usbctrl` 和 `usbdev` 指定 USB 控制器和 USB 主机设备。例如，对于 HVM Guest：

```
usbctrl=['type=qusb,version=2,ports=4', 'type=qusb,version=1,ports=4', ]
usbdev=['hostbus=2, hostaddr=1, controller=0,port=1', ]
```



注意

必须为 HVM Guest 的控制器指定 type=qusb。

要管理 PVUSB 设备的热插入，请使用 usbctrl-attach、usbctrl-detach、usb-list、usbdev-attach 和 usb-detach 子命令。例如：

创建版本为 USB 1.1 并包含 8 个端口的 USB 控制器：

```
# xl usbctrl-attach test_vm version=1 ports=8 type=qusb
```

在域中找到第一个可用的 controller:port，并将 busnum:devnum 为 2:3 的 USB 设备挂接到该端口；您也可以指定 controller 和 port：

```
# xl usbdev-attach test_vm hostbus=2 hostaddr=3
```

显示域中的所有 USB 控制器和 USB 设备：

```
# xl usb-list test_vm
Devid  Type   BE   state usb-ver  ports
0      qusb   0    1     1        8
  Port 1: Bus 002 Device 003
  Port 2:
  Port 3:
  Port 4:
  Port 5:
  Port 6:
  Port 7:
  Port 8:
```

分离控制器 0 端口 1 下的 USB 设备：

```
# xl usbdev-detach test_vm 0 1
```

去除具有所示 dev_id 的 USB 控制器，以及其下的所有 USB 设备：

```
# xl usbctrl-detach test_vm dev_id
```

有关详细信息，请参见 https://wiki.xenproject.org/wiki/Xen_USB_Passthrough 。

26 虚拟网络

VM Guest 系统需要通过特定的方式来与其他 VM Guest 系统或本地网络通讯。用于连接 VM Guest 系统的网络接口由一个分离式设备驱动程序构成，也就是说，任何虚拟以太网设备在 Dom0 中都有一个对应的网络接口。此接口设置为访问 Dom0 中运行的虚拟网络。SUSE Linux Enterprise Server 的系统配置中全面集成了桥接式虚拟网络，您可以通过 YaST 对该网络进行配置。

安装 Xen VM 主机服务器时，系统会在常规网络配置期间建议一种桥接式网络配置。用户可以选择在安装期间更改配置，并可根据本地需求对其进行自定义。

如果需要，可以在执行默认的物理服务器安装后，使用 YaST 中的 Install Hypervisor and Tools 模块来安装 Xen VM 主机服务器。此模块会使系统做好托管虚拟机的准备，包括调用默认网桥功能的建议。

如果使用 rpm 或 zypper 手动安装了 Xen VM 主机服务器所需的软件包，则其余的系统配置需由管理员手动完成，或者通过 YaST 完成。

SUSE Linux Enterprise Server 中默认不使用 Xen 提供的网络脚本。这些脚本仅供参考，并且已禁用。SUSE Linux Enterprise Server 中使用的网络配置通过 YaST 系统配置完成，该系统配置类似于 SUSE Linux Enterprise Server 中的网络接口配置。

有关管理网桥的更多一般信息，请参见第 9.1.1 节“网桥”。

26.1 Guest 系统的网络设备

Xen 超级管理程序可以提供不同类型的网络接口来连接 VM Guest 系统。首选网络设备应该是半虚拟化网络接口。此类网络接口的系统要求最低，却能实现最高的传输速率。最多可为每个 VM Guest 提供八个网络接口。

无法感知半虚拟化硬件的系统可能无法使用此方案。有多个模拟网络接口可用于将系统连接到只能以全虚拟化模式运行的网络。您可以自行选用以下模拟产品：

- Realtek 8139 (PCI)。这是默认的模拟网卡。
- AMD PCnet32 (PCI)
- NE2000 (PCI)

- NE2000 (ISA)
- Intel e100 (PCI)
- Intel e1000 及其衍生产品 e1000-82540em、e1000-82544gc 和 e1000-82545em (PCI)

以上网络接口全部都是软件接口。由于每个网络接口都必须具有唯一的 MAC 地址，因此已向这些接口可以使用的 XenSource 分配了一个地址范围。



提示：虚拟网络接口和 MAC 地址

虚拟化环境中的默认 MAC 地址配置会创建类似于 00:16:3E:xx:xx:xx 的随机 MAC 地址。一般情况下，可用 MAC 地址的数量应该足够大才能获得唯一地址。但是，如果您的安装规模非常大，或者要确保随机 MAC 地址分配不会造成问题，您也可以手动分配这些地址。

进行调试或系统管理时，知道 Dom0 中的哪个虚拟接口连接到正在运行的 Guest 中的哪个以太网设备可能很有帮助。可以从 Dom0 中的设备名称获得此信息。所有虚拟设备都遵循命名规则 vif<domain number>.<interface_number>。

例如，在 Dom0 中，ID 为 5 的 VM Guest 的第三个接口 (eth2) 的设备名称是 vif5.2。要获取所有可用接口的列表，请运行 **ip a** 命令。

设备名称不包含有关此接口连接到哪个网桥的任何信息，但 Dom0 中提供了此信息。要大致了解有关哪个接口连接到哪个网桥的信息，请运行 **bridge link** 命令。输出可能如下所示：

```
> sudo bridge link
2: eth0 state DOWN : <NO-CARRIER,BROADCAST,MULTICAST,SLAVE,UP> mtu 1500 master br0
3: eth1 state UP : <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 master br1
```

此示例中配置了三个网桥：br0、br1 和 br2。当前为 br0 和 br1 各添加了一个真实的以太网设备，分别是 eth0 和 eth1。

26.2 Xen 中基于主机的路由

可以将 Xen 设置为在控制 Dom0 中使用基于主机的路由，不过，YaST 目前还不能很好地支持此设置，需要手动对配置文件进行大量的编辑才能使用此设置。因此，此任务需要由高级管理员来完成。

仅当使用固定 IP 地址时，以下配置才起作用。在此过程中不能使用 DHCP，因为 VM Guest 和 VM 主机服务器系统都必须知道 IP 地址。

要创建路由式 Guest，最简单的方法就是将网络从桥接网络更改为路由网络。要执行以下过程，必须先安装一个使用桥接网络的 VM Guest。例如，VM 主机服务器名为 earth，IP 为 192.168.1.20；VM Guest 名为 alice，IP 为 192.168.1.21。

过程 26.1：配置路由式 IPV4 VM GUEST

1. 确保 alice 已关机。使用 `xl` 命令关机并检查。
2. 在 VM 主机服务器 earth 上准备网络配置：
 - a. 创建一个用于路由流量的热插拔接口。为此，请创建包含以下内容的 `/etc/sysconfig/network/ifcfg-alice.0` 文件：

```
NAME="Xen guest alice"
BOOTPROTO="static"
STARTMODE="hotplug"
```

- b. 确保已启用 IP 转发：
 - i. 在 YaST 中，转到网络设置 > 路由。
 - ii. 进入路由选项卡，选中启用 IPv4 转发和启用 IPv6 转发选项。
 - iii. 确认设置并退出 YaST。
 - c. 将以下配置应用于 `firewalld`：
 - 将 alice.0 添加到公共区域中的设备：

```
> sudo firewall-cmd --zone=public --add-interface=alice.0
```

- 告知防火墙要转发哪个地址：

```
> sudo firewall-cmd --zone=public \
--add-forward-
port=port=80:proto=tcp:toport=80:toaddr="192.168.1.21/32,0/0"
```

- 使运行时配置更改永久生效：

```
> sudo firewall-cmd --runtime-to-permanent
```

- d. 将一个静态路由添加到 alice 的接口。为此，请将以下行添加到 /etc/sysconfig/network/routes 的末尾：

```
192.168.1.21 - - alice.0
```

- e. 为确保 VM 主机服务器连接的交换机和路由器知道路由接口，请在 earth 上激活 proxy_arp。将以下行添加到 /etc/sysctl.conf 中：

```
net.ipv4.conf.default.proxy_arp = 1
net.ipv4.conf.all.proxy_arp = 1
```

- f. 使用以下命令激活所有更改：

```
> sudo systemctl restart systemd-sysctl wicked
```

3. 按照第 27.1 节 “**XL — Xen 管理工具**” 中所述，通过更改 alice 的 vif 接口配置继续完成 VM Guest 的 Xen 配置。对配置过程中生成的文本文件进行以下更改：

- a. 去除以下代码段

```
bridge=br0
```

- b. 添加以下代码段：

```
vifname=vifalice.0
```

或

```
vifname=vifalice.0=emu
```

(针对全虚拟化域)。

- c. 按如下所示更改用于设置接口的脚本：

```
script=/etc/xen/scripts/vif-route-ifup
```

- d. 激活新配置并启动 VM Guest。

- 4. 其余配置任务必须在 VM Guest 内部完成。

- a. 使用 `xl console DOMAIN` 打开与 VM Guest 连接的控制台，然后登录。
- b. 检查 Guest IP 是否设置为 192.168.1.21。
- c. 为 VM Guest 提供用于连接 VM 主机服务器的主机路由和默认网关。为此，请将以下行添加到 `/etc/sysconfig/network/routes` 中：

```
192.168.1.20 - - eth0
default 192.168.1.20 - -
```

- 5. 最后，测试从 VM Guest 到外部网络的网络连接，以及从该网络到 VM Guest 的连接。

26.3 创建伪装网络设置

创建伪装网络设置与创建路由设置的过程相似。不过，创建伪装网络设置时不需要 `proxy_arp`，并且某些防火墙规则不同。要为 IP 地址为 192.168.100.1 并且其主机在 `br0` 上有自己的外部接口的 Guest（名为 dolly）创建伪装网络，请执行以下操作。为了简化配置，此处仅将已安装的 Guest 修改为使用伪装网络：

过程 26.2：配置伪装 IPV4 VM GUEST

- 1. 使用 `xl shutdown DOMAIN` 关闭 VM Guest 系统。
- 2. 在 VM 主机服务器上准备网络配置：
 - a. 创建一个用于路由流量的热插拔接口。为此，请创建包含以下内容的 `/etc/sysconfig/network/ifcfg-dolly.0` 文件：

```
NAME="Xen guest dolly"
```



```
BOOTPROTO="static"
STARTMODE="hotplug"
```

b. 编辑文件 `/etc/sysconfig/SuSEfirewall2`，在其中添加以下配置：

- 将 dolly.0 添加到 FW_DEV_DMZ 中的设备：

```
FW_DEV_DMZ="dolly.0"
```

- 在防火墙中打开路由：

```
FW_ROUTE="yes"
```

- 在防火墙中打开伪装：

```
FW_MASQUERADE="yes"
```

- 告知防火墙要伪装哪个网络：

```
FW_MASQ_NETS="192.168.100.1/32"
```

- 从伪装例外中去除网络：

```
FW_NOMASQ_NETS=""
```

- 最后，使用以下命令重新启动防火墙：

```
> sudo systemctl restart SuSEfirewall2
```

c. 向 dolly 的接口添加一个静态路由。为此，请将以下行添加到 `/etc/sysconfig/network/routes` 的末尾：

```
192.168.100.1 - - dolly.0
```

d. 使用以下命令激活所有更改：

```
> sudo systemctl restart wicked
```

3. 继续完成 VM Guest 的 Xen 配置。

a. 按照第 27.1 节 “XL — Xen 管理工具” 中所述更改 dolly 的 vif 接口配置。

b. 去除下面一项：

```
bridge=br0
```

c. 添加以下代码段：

```
vifname=vifdolly.0
```

d. 按如下所示更改用于设置接口的脚本：

```
script=/etc/xen/scripts/vif-route-ifup
```

e. 激活新配置并启动 VM Guest。

4. 其余配置任务需在 VM Guest 内部完成。

a. 使用 **xl console DOMAIN** 打开与 VM Guest 连接的控制台，然后登录。

b. 检查 Guest IP 是否设置为 192.168.100.1。

c. 为 VM Guest 提供用于连接 VM 主机服务器的主机路由和默认网关。为此，请将以下行添加到 /etc/sysconfig/network/routes 中：

```
192.168.1.20 - - eth0
default 192.168.1.20 - -
```

5. 最后，测试从 VM Guest 到外部世界的网络连接。

26.4 特殊配置

在 Xen 中可以使用许多可行的网络配置。以下配置默认未激活：

26.4.1 虚拟网络中的带宽限制

在 Xen 中，您可以限制虚拟 Guest 在访问网桥时可使用的网络传输速率。要配置该限制，需要按照第 27.1 节“[XL — Xen 管理工具](#)”中所述修改 VM Guest 配置。

在配置文件中，先搜索连接到虚拟网桥的设备。配置如下所示：

```
vif = [ 'mac=00:16:3e:4f:94:a9,bridge=br0' ]
```

要添加最大传输速率，请按如下所示在配置中添加参数 `rate`：

```
vif = [ 'mac=00:16:3e:4f:94:a9,bridge=br0,rate=100Mb/s' ]
```

速率单位为 `Mb/s`（每秒兆位数）或 `MB/s`（每秒兆字节数）。在上面的示例中，虚拟接口的最大传输速率为 100 兆位。默认情况下，Guest 与虚拟网桥之间的带宽没有限制。

您甚至可以通过指定用于定义信用补充粒度的时段来微调该行为：

```
vif = [ 'mac=00:16:3e:4f:94:a9,bridge=br0,rate=100Mb/s@20ms' ]
```

26.4.2 监控网络流量

要监控特定接口上的流量，不妨使用一个实用的小应用程序 `iftop`，它可在终端中显示当前网络流量。

运行 Xen VM 主机服务器时，需要定义受监控接口。Dom0 用来访问物理网络的接口是网桥设备，例如 `br0`。但在您的系统上可能不是这样。要监控传输到物理接口的所有流量，请以 `root` 身份运行终端并使用以下命令：

```
iftop -i br0
```

要监控特定 VM Guest 的特殊网络接口的网络流量，请提供正确的虚拟接口。例如，要监控 ID 为 5 的域的第一个以太网设备，请使用以下命令：

```
ftop -i vif5.0
```

要退出 `iftop`，请按 `q` 键。手册页 `man 8 iftop` 中介绍了更多选项和可行做法。

27 管理虚拟化环境

除了使用建议的 `libvirt` 库（第 II 部分“使用 `libvirt` 管理虚拟机”）以外，您还可以在命令行中使用 `xl` 工具来管理 Xen Guest 域。

27.1 XL — Xen 管理工具

`xl` 程序是用于管理 Xen Guest 域的工具。它包含在 `xen-tools` 软件包中。`xl` 基于 `LibXenlight` 库，可用于执行一般的域管理工作，例如创建、列出、暂停或关闭域。只有 `root` 用户才能执行 `xl` 命令。



注意

`xl` 只能管理域配置文件指定的运行中 Guest 域。如果某个 Guest 域未运行，则您无法使用 `xl` 来管理它。



提示

为了让用户能够继续像使用已过时的 `xm` 命令那样来使用受管 Guest 域，目前我们建议使用 `libvirt` 的 `virsh` 和 `virt-manager` 工具。有关详细信息，请参见第 II 部分“使用 `libvirt` 管理虚拟机”。

`xl` 操作依赖于 `xenstored` 和 `xenconsole` 服务。请确保在引导时启动

```
> systemctl start xencommons
```

以初始化 `xl` 所需的所有守护程序。



提示：在主机域中设置 `xenbr0` 网桥

在最常用的网络配置中，需在主机域中设置一个名为 `xenbr0` 的网桥，以便为 Guest 域提供正常工作的网络。

每个 `xl` 命令的基本结构如下：

```
xl <subcommand> [options] domain_id
```

其中，<subcommand> 是要运行的 xl 命令，domain_id 是分配给域的 ID 编号或虚拟机的名称，**OPTIONS** 表示特定于子命令的选项。

如需可用 **xl** 子命令的完整列表，请运行 **xl help**。对于每个命令，都可以使用附加参数 **--help** 获取更详细的帮助。**xl** 的手册页中提供了有关相应子命令的详细信息。

例如，**xl list --help** 会显示 list 命令可用的所有选项。举例来说，**xl list** 命令会显示所有虚拟机的状态。

```
> sudo xl list
```

Name	ID	Mem	VCPUs	State	Time(s)
Domain-0	0	457	2	r-----	2712.9
sles12	7	512	1	-b-----	16.3
opensuse		512	1		12.9

State 信息指示某个计算机是否正在运行，以及处于哪种状态。最常用的标志为 **r**（正在运行）和 **b**（受阻），其中“受阻”的意思是该计算机正在等待 IO，或者由于无需执行任何操作而处于休眠状态。有关状态标志的更多细节，请参见 **man 1 xl**。

其他有用的 **xl** 命令包括：

- **xl create**，用于基于给定的配置文件创建虚拟机。
- **xl reboot**，用于重引导虚拟机。
- **xl destroy**，用于立即终止虚拟机。
- **xl block-list**，用于显示挂接到虚拟机的所有虚拟块设备。

27.1.1 Guest 域配置文件

使用 **xl** 操作域时，每个域都需有相应的域配置文件。用于存储此类配置文件的默认目录为 **/etc/xen/**。

域配置文件是一个纯文本文件。它包含多个 **KEY=VALUE** 对。某些键是必需的。通用键适用于任何 Guest，还有些键只适用于特定的 Guest 类型（半虚拟化或全虚拟化）。值可以是括在单引号或双引号中的字符串（**"string"** 形式）、数字、布尔值，或者括在方括号中的多个值的列表（**[value1, value2, ...]** 形式）。

例 27.1：SLED 12 的 GUEST 域配置文件：/etc/xen/sled12.cfg

```
name= "sled12"
builder = "hvm"
vncviewer = 1
memory = 512
disk = [ '/var/lib/xen/images/sled12.raw,,hda', '/dev/cdrom,,hdc,cdrom' ]
vif = [ 'mac=00:16:3e:5f:48:e4,model=rtl8139,bridge=br0' ]
boot = "n"
```

要启动此类域，请运行 **xl create /etc/xen/sled12.cfg**。

27.2 自动启动 Guest 域

要使 Guest 域在主机系统引导后自动启动，请执行以下步骤：

1. 创建域配置文件（如果不存在），并将其保存到 /etc/xen/ 目录中，例如 /etc/xen/domain_name.cfg。
2. 在 auto/ 子目录中创建 Guest 域配置文件的符号链接。

```
> sudo ln -s /etc/xen/domain_name.cfg /etc/xen/auto/domain_name.cfg
```

3. 系统下次引导时，domain_name.cfg 中定义的 Guest 域将会启动。

27.3 事件操作

在 Guest 域配置文件中，您可以定义在发生一组预定义的事件时要执行的操作。例如，要告知域在其关机后自行重新启动，请在其配置文件中包含下面一行：

```
on_poweroff="restart"
```

下面是 Guest 域的预定义事件列表：

事件列表

on_poweroff

指定在域自行关机后应执行什么操作。

on_reboot

当域关机并提供了请求重引导的原因代码时要执行的操作。

on_watchdog

当域由于 Xen 看门狗超时而关机时要执行的操作。

on_crash

当域崩溃时要执行的操作。

对于这些事件，可以定义以下操作之一：

相关操作列表

destroy

销毁域。

restart

销毁域，并立即采用相同的配置创建新域。

rename-restart

重命名已终止的域，然后立即采用与原始域相同的配置创建新域。

preserve

保留域。可以检查该域，以后再使用 **xl destroy** 将它销毁。

coredump-destroy

将域的核心转储写入 /var/xen/dump/NAME，然后销毁该域。

coredump-restart

将域的核心转储写入 /var/xen/dump/NAME，然后重新启动该域。

27.4 时戳计数器

您可以为 Guest 域配置文件中的每个域指定时戳计数器 (TSC)（有关详细信息，请参见第 27.1.1 节“Guest 域配置文件”）。

使用 tsc_mode 设置可以指定是要“本机”执行 rdtsc 指令（速度较快，但 TSC 敏感型应用程序有时无法正常运行），还是模拟这些指令（始终可正常运行，但性能可能受到影响）。

tsc_mode=0 (默认)

使用此设置可确保正常运行，同时提供可以实现的最佳性能 — 有关详细信息，请参见 <https://xenbits.xen.org/docs/4.3-testing/misc/tscmode.txt>。

tsc_mode=1 (始终模拟)

如果 TSC 敏感型应用程序正在运行，并且已知且可接受最坏情况下的性能下降，请使用此设置。

tsc_mode=2 (永不模拟)

如果此 VM 中运行的所有应用程序都能够灵活适应 TSC 并且需要最高性能，请使用此设置。

tsc_mode=3 (PVRDTSCP)

可以半虚拟化（修改）高 TSC 频率应用程序，以便兼顾正确性和最高性能 — 任何未经修改的应用程序都必须能够灵活适应 TSC。

有关背景信息，请参见 <https://xenbits.xen.org/docs/4.3-testing/misc/tscmode.txt>。

27.5 保存虚拟机

过程 27.1：保存虚拟机的当前状态

1. 确保要保存的虚拟机正在运行。
2. 在主机环境中输入

```
> sudo xl save ID STATE-FILE
```

其中，ID 是要保存的虚拟机 ID，STATE-FILE 是您为内存状态文件指定的名称。默认情况下，当您创建域的快照后，该域将不再运行。使用 -c 可使域保持运行状态，即使在您创建快照后也是如此。

27.6 恢复虚拟机

过程 27.2：恢复虚拟机的当前状态

1. 确保要恢复的虚拟机自您运行保存操作后始终未启动。
2. 在主机环境中输入

```
> sudo xl restore STATE-FILE
```

其中，STATE-FILE 是先前保存的内存状态文件。域在恢复后默认将处于运行状态。要在恢复后暂停域，请使用 -p。

27.7 虚拟机状态

可以通过查看 xl list 命令的结果来显示虚拟机的状态，结果中以单字符缩写形式显示状态。

- r - 正在运行 - 虚拟机当前正在运行，并在消耗分配的资源。
- b - 受阻 - 虚拟机的处理器未运行，且无法运行。虚拟机正在等待 I/O 或已停止工作。
- p - 已暂停 - 虚拟机已暂停。虚拟机不会与超级管理程序交互，但仍保有其分配的资源，例如内存。
- s - 已关闭 - Guest 操作系统正在关闭、已重引导或已挂起，并且虚拟机正在停止。
- c - 已崩溃 - 虚拟机已崩溃，未在运行。
- d - 即将死机 - 虚拟机正在关机或崩溃。

28 Xen 中的块设备

28.1 将物理存储设备映射到虚拟磁盘

域配置文件中 Xen 域的磁盘指定方式非常直接，如以下示例所示：

```
disk = [ 'format=raw,vdev=hdc,access=ro,devtype=cdrom,target=/root/image.iso' ]
```

此命令基于 `/root/image.iso` 磁盘映像文件定义一个磁盘块设备。Guest 将该磁盘视为 `hdc`，只能对它进行只读 (`ro`) 访问。设备类型为 `cdrom`，格式为 `raw`。

下面的示例定义了相同的设备，但使用的是简化的位置语法：

```
disk = [ '/root/image.iso,raw,hdc,ro,cdrom' ]
```

您可在同一行中包含更多磁盘定义，每个定义需以逗号分隔。如果未指定某个参数，系统会采用其默认值：

```
disk = [ '/root/image.iso,raw,hdc,ro,cdrom', '/dev/vg/guest-volume,,hda','...' ]
```

参数列表

target

源块设备或磁盘映像路径。

format

映像文件的格式。默认值为 `raw`。

vdev

Guest 看到的虚拟设备。支持的值为 `hd[x]`、`xvd[x]`、`sd[x]` 等。有关详细信息，请参见 `/usr/share/doc/packages/xen/misc/vbd-interface.txt`。此参数是必需的。

access

提供给 Guest 的块设备处于只读模式还是读写模式。支持的值为 `ro` 或 `r`（只读访问）以及 `rw` 或 `w`（读写访问）。对于 `devtype=cdrom`，默认值为 `ro`；对于其他设备类型，默认值为 `rw`。

devtype

限定虚拟设备类型。支持的值为 `cdrom`。

backendtype

要使用的后端实现。支持的值为 `phy`、`tap` 和 `qdisk`。一般不应指定此选项，因为系统会自动确定后端类型。

script

指定 `target` 不是常规的主机路径，而是要由可执行程序解释的信息。如果指定的脚本文件不指向绝对路径，将在 `/etc/xen/scripts` 中查找该文件。这些脚本通常名为 `block-<script_name>`。

有关指定虚拟磁盘的详细信息，请参见 `/usr/share/doc/packages/xen/misc/xl-disk-configuration.txt`。

28.2 将网络存储设备映射到虚拟磁盘

与映射本地磁盘映像（请参见第 28.1 节“将物理存储设备映射到虚拟磁盘”）类似，您也可以将网络磁盘映射为虚拟磁盘。

下面的示例展示了启用了多个 Ceph 监控器和 cephx 身份验证的 RBD（RADOS 块设备）的映射：

```
disk = [ 'vdev=hdc, backendtype=qdisk, \
target=rbd:libvirt-pool/new-libvirt-image:\
id=libvirt:key=AQDsPwtW8JoXJBAAyLPQe7MhCC\
+JPKI3QuhaAw==:auth_supported=cephx;none:\
mon_host=137.65.135.205\\:6789;137.65.135.206\\:6789;137.65.135.207\\:6789' ]
```

下面是 NBD（网络块设备）磁盘映射示例：

```
disk = [ 'vdev=hdc, backendtype=qdisk, target=nb:151.155.144.82:5555' ]
```

28.3 基于文件的虚拟磁盘和回写设备

当虚拟机正在运行时，其每个基于文件的虚拟磁盘都会占用主机上的一个回写设备。默认情况下，主机最多允许占用 64 个回写设备。

要在主机上同时运行更多基于文件的虚拟磁盘，可以通过将以下选项添加到主机的 `/etc/modprobe.conf.local` 文件，来增加可用回写设备的数量。

```
options loop max_loop=x
```

其中，`x` 是要创建的回写设备的最大数量。

重新加载模块后，更改即会生效。



提示

输入 `rmmod loop` 和 `modprobe loop` 可以卸载和重新加载模块。如果 `rmmod` 不起作用，请卸载所有现有循环设备，或重引导计算机。

28.4 调整块设备的大小

尽管您始终都可以向 VM Guest 系统添加新的块设备，但有时更合适的做法是增加现有块设备的大小。如果在部署 VM Guest 期间已经规划了此类系统修改，则应该注意多个基本事项：

- 使用可以增加大小的块设备。通常使用 LVM 设备和文件系统映像。
- 不要将 VM Guest 内部的设备分区，而是直接使用主设备来应用文件系统。例如，直接使用 `/dev/xvdb`，而不是在 `/dev/xvdb` 中添加分区。
- 确保要使用的文件系统可调整大小。有时（例如，使用 Ext3 时），必须关闭某些功能才能调整文件系统的大小。其中一个可以联机调整大小并挂载的文件系统是 XFS。增加底层块设备的大小后，使用 `xfs_growfs` 命令调整该文件系统的大小。有关 XFS 的详细信息，请参见 `man 8 xfs_growfs`。

调整分配给 VM Guest 的 LVM 设备的大小后，VM Guest 会自动获悉新的大小。您无需执行其他操作告知 VM Guest 块设备的新大小。

使用文件系统映像时，将使用一个循环设备向 Guest 挂接映像文件。有关调整该映像的大小以及为 VM Guest 刷新大小信息的详细信息，请参见第 30.2 节“稀疏映像文件和磁盘空间”。

28.5 用于管理高级存储方案的脚本

可以借助脚本来管理高级存储方案，例如 **dmmd**（“设备映射程序 — 多磁盘”）提供的磁盘环境，包括构建在软件 RAID 集基础之上的 LVM 环境，或者构建在 LVM 环境基础之上的软件 RAID 集。这些脚本包含在 `xen-tools` 软件包中。安装后，可以在 `/etc/xen/scripts` 中找到它们：

- **`block-dmmd`**
- **`block-drbd-probe`**
- **`block-npiv`**

借助这些脚本，可以在将块设备提供给 Guest 之前使用外部命令执行这些设备的特定操作或一系列操作。

以前，只能使用磁盘配置语法 `script=` 将这些脚本与 **`xl`** 或 **`libxl`** 结合使用。现在，可以通过在磁盘的 `<source>` 元素中指定块脚本的基本名称，将这些脚本与 `libvirt` 结合使用。例如：

```
<source dev='dmmd:md;/dev/md0;lvm;/dev/vg xen/lv-vm01' />
```

29 虚拟化：配置选项和设置

本节中的内容介绍可帮助技术创新者实施前沿虚拟化解决方案的高级管理任务和配置选项。这些内容仅作为参考信息，并不意味着所述的所有选项和任务都受 Novell, Inc. 的支持。

29.1 虚拟 CD 读取器

可以在创建虚拟机时设置虚拟 CD 读取器，也可以将虚拟 CD 读取器添加到现有虚拟机。虚拟 CD 读取器可以基于物理 CD/DVD 或 ISO 映像。虚拟 CD 读取器的工作方式根据其是半虚拟还是全虚拟读取器而异。

29.1.1 半虚拟计算机上的虚拟 CD 读取器

一台半虚拟计算机最多可以包含 100 个由虚拟 CD 读取器和虚拟磁盘构成的块设备。在半虚拟计算机上，虚拟 CD 读取器会将 CD 显示为允许进行只读访问的虚拟磁盘。虚拟 CD 读取器不可用于向 CD 写入数据。

访问完半虚拟计算机上的 CD 后，建议从虚拟机中去除虚拟 CD 读取器。

半虚拟化 Guest 可以使用 `devtype=cdrom` 设备类型。这可以在一定程度上模拟真实 CD 读取器的行为，并允许更换 CD。您甚至可以使用 `eject` 命令打开 CD 读取器的托盘。

29.1.2 全虚拟计算机上的虚拟 CD 读取器

一台全虚拟计算机最多可以包含 4 个由虚拟 CD 读取器和虚拟磁盘构成的块设备。全虚拟计算机上的虚拟 CD 读取器会像物理 CD 读取器一样与插入的 CD 进行交互。

将 CD 插入主机计算机上的物理 CD 读取器后，包含基于物理 CD 读取器（例如 `/dev/cdrom/`）的虚拟 CD 读取器的所有虚拟机都可以读取插入的 CD。假设操作系统具有自动挂载功能，CD 应该会自动显示在文件系统中。虚拟 CD 读取器不可用于向 CD 写入数据。它们配置为只读设备。

29.1.3 添加虚拟 CD 读取器

虚拟 CD 读取器可以基于插入 CD 读取器中的 CD，也可以基于 ISO 映像文件。

1. 请确保虚拟机正在运行，并且操作系统已完成引导。
2. 将所需 CD 插入物理 CD 读取器，或者将所需 ISO 映像复制到 Dom0 可访问的位置。
3. 在 VM Guest 中选择一个新的未使用的块设备，例如 `/dev/xvdb`。
4. 选择您要分配给 Guest 的 CD 读取器或 ISO 映像。
5. 使用真实 CD 读取器时，请使用以下命令将 CD 读取器指派给 VM Guest。在此示例中，Guest 名为 `alice`：

```
> sudo xl block-attach alice target=/dev/sr0,vdev=xvdb,access=ro
```

6. 分配映像文件时，请使用以下命令：

```
> sudo xl block-attach alice target=/path/to/file.iso,vdev=xvdb,access=ro
```

7. 新的块设备（例如 `/dev/xvdb`）即会添加到虚拟机中。
8. 如果虚拟机运行的是 Linux，请完成以下操作：

- a. 在虚拟机中打开一个终端，然后输入 `fdisk -l` 校验是否正确添加了设备。您也可以输入 `ls /sys/block` 查看虚拟机可用的所有磁盘。

虚拟机将 CD 识别为带有驱动器号的虚拟磁盘，例如：

```
/dev/xvdb
```

- b. 输入使用相应驱动器号挂载 CD 或 ISO 映像的命令。例如，

```
> sudo mount -o ro /dev/xvdb /mnt
```

会将 CD 挂载到名为 `/mnt` 的挂载点。

虚拟机应该可以通过指定的挂载点使用该 CD 或 ISO 映像文件。

9. 如果虚拟机运行的是 Windows，请重引导虚拟机。

校验虚拟 CD 读取器是否显示在 My Computer 部分。

29.1.4 去除虚拟 CD 读取器

1. 请确保虚拟机正在运行，并且操作系统已完成引导。
2. 如果已挂载虚拟 CD 读取器，请从虚拟机内部将其卸载。
3. 在 Guest 块设备的主机视图中输入 **`xl block-list alice`**。
4. 输入 **`xl block-detach alice BLOCK_DEV_ID`** 以从 Guest 中去除该虚拟设备。如果该操作失败，请尝试添加 **`-f`** 以强制去除。
5. 按下硬件弹出按钮以弹出 CD。

29.2 远程访问方法

某些配置（例如包含机架式服务器的配置）要求运行的计算机不配备视频监视器、键盘或鼠标。此类配置通常称为 headless，需要使用远程技术来管理。

典型的配置方案和技术包括：

包含 X Window 系统服务器的图形桌面

如果在虚拟机主机上安装了图形桌面（例如 GNOME），您便可以使用 VNC 查看器这样的远程查看器。在远程计算机上，使用 tigervnc 或 virt-viewer 等图形工具登录并管理远程 Guest 环境。

仅文本

您可以从远程计算机使用 ssh 命令登录到虚拟机主机并访问其基于文本的控制台。然后可以使用 xl 命令来管理虚拟机，并可使用 virt-install 命令来创建新虚拟机。

29.3 VNC 查看器

VNC 查看器用于以图形方式查看运行中 Guest 系统的环境。可以从 Dom0 使用该查看器（称为本地访问或机上访问），也可以从远程计算机使用。

可以使用 VM 主机服务器的 IP 地址和 VNC 查看器来查看此 VM Guest 的显示画面。当虚拟机运行时，主机上的 VNC 服务器将为该虚拟机分配一个端口号，用于建立 VNC 查看器连接。分配的端口号是虚拟机启动时可用的最小端口号。该端口号仅供虚拟机运行时使用。虚拟机关机后，该端口号可能会分配给其他虚拟机。

例如，如果端口 1、2、4、5 已分配给运行中的虚拟机，那么 VNC 查看器将分配最小可用端口号 3。如果虚拟机下次启动时端口号 3 仍在使用中，则 VNC 服务器将向虚拟机分配其他端口号。

要从远程计算机使用 VNC 查看器，防火墙必须允许访问 VM Guest 系统要通过其运行的端口数。即，要允许访问 5900 及更大编号的端口。例如，要运行 10 个 VM Guest 系统，则需要打开 TCP 端口 5900:5910。

要从运行 VNC 查看器客户端的本地控制台访问虚拟机，请输入以下命令之一：

- `vncviewer ::590#`

- `vncviewer :#`

`#` 是分配给虚拟机的 VNC 查看器端口号。

从 Dom0 之外的计算机访问 VM Guest 时，请使用以下语法：

```
> vncviewer 192.168.1.20::590#
```

在本例中，Dom0 的 IP 地址为 192.168.1.20。

29.3.1 向虚拟机分配 VNC 查看器端口号

尽管 VNC 查看器默认会分配第一个可用的端口号，但您应该向特定的虚拟机分配特定的 VNC 查看器端口号。

要在 VM Guest 上分配特定的端口号，请编辑虚拟机的 `xl` 设置，并将 `vnclisten` 更改为所需的值。以端口号 5902 为例，请仅指定 2，因为系统会自动加上 5900：

```
vfb = [ 'vnc=1,vnclisten="localhost:2"' ]
```

有关编辑 Guest 域的 `xl` 设置的详细信息，请参见第 27.1 节“[XL — Xen 管理工具](#)”。



提示

分配较大的端口号可避免与 VNC 查看器分配的端口号冲突，VNC 查看器使用的是最小的可用端口号。

29.3.2 使用 SDL 而不是 VNC 查看器

如果您要从虚拟机主机控制台访问虚拟机的显示画面（称为本地访问或机上访问），应使用 SDL 而不是 VNC 查看器。通过网络查看桌面时，VNC 查看器速度更快，但从同一台计算机查看桌面时，SDL 速度更快。

要设置为默认使用 SDL 而不是 VNC，请按如下所示更改虚拟机的配置信息。有关说明，请参见第 27.1 节“[XL — Xen 管理工具](#)”。

```
vfb = [ 'sdl=1' ]
```

请记住，与 VNC 查看器窗口不同，关闭 SDL 窗口会终止虚拟机。

29.4 虚拟键盘

启动虚拟机后，主机会根据虚拟机的设置创建一个与 **keymap** 项匹配的虚拟键盘。如果未指定 **keymap** 项，虚拟机的键盘将默认为“英语（美国）”。

要查看虚拟机当前的 **keymap** 项，请在 Dom0 上输入以下命令：

```
> xl list -l VM_NAME | grep keymap
```

要配置 Guest 的虚拟键盘，请使用以下代码段：

```
vfb = [ 'keymap="de"' ]
```

有关支持的键盘布局的完整列表，请参见 **xl.cfg** 手册页 **man 5 xl.cfg** 的 **Keymaps** 部分。

29.5 分配专用 CPU 资源

在 Xen 中，可以指定 Dom0 或 VM Guest 应使用多少以及哪些 CPU 核心来保持其性能。Dom0 的性能对于系统整体而言非常重要，因为磁盘和网络驱动程序都是在 Dom0 上运行。此外，I/O 密集型 Guest 的工作负载可能会消耗 Dom0 的大量 CPU 周期。不过，设置 VM Guest 的目的是为了完成某项任务，而 VM Guest 的性能对于此类任务能否完成也很重要。

29.5.1 Dom0

为 Dom0 分配专用 CPU 资源可以提高虚拟化环境的总体性能，因为这样 Dom0 便会有可用的 CPU 时间处理来自 VM Guest 的 I/O 请求。不为 Dom0 分配专用 CPU 资源可能会导致性能不佳，并可能导致 VM Guest 无法正常运行。

分配专用 CPU 资源涉及到三个基本步骤：修改 Xen 引导行，将 Dom0 的 VCPU 绑定到物理处理器，然后在 VM Guest 上配置 CPU 相关选项：

1. 首先，需要将 `dom0_max_vcpus=X` 追加到 Xen 引导行。为此，请将以下行添加到 `/etc/default/grub` 中：

```
GRUB_CMDLINE_XEN="dom0_max_vcpus=X"
```

如果 `/etc/default/grub` 已包含设置了 `GRUB_CMDLINE_XEN` 的行，请将 `dom0_max_vcpus=X` 追加到此行。

需将 `X` 替换为供 Dom0 专用的 VCPU 数量。

2. 运行以下命令更新 GRUB 2 配置文件：

```
> sudo grub2-mkconfig -o /boot/grub2/grub.cfg
```

3. 重引导以使更改生效。

4. 下一步是将 Dom0 的每个 VCPU 绑定（或“固定”）到一个物理处理器。

```
> sudo xl vcpu-pin Domain-0 0 0
xl vcpu-pin Domain-0 1 1
```

第一行将 Dom0 的 VCPU 0 绑定到物理处理器 0，第二行将 Dom0 的 VCPU 1 绑定到物理处理器 1。

5. 最后，需确保没有任何 VM Guest 使用 Dom0 的 VCPU 专用的物理处理器。假设您正在运行一个 8 CPU 系统，需将

```
cpus="2-8"
```

添加到相关 VM Guest 的配置文件。

29.5.2 VM Guest

用户常常需要分配特定的 CPU 资源供虚拟机专用。默认情况下，虚拟机会使用任何可用的 CPU 核心。为虚拟机分配合理数量的物理处理器可以提升其性能，因为分配给虚拟机后，其他 VM Guest 都不能使用这些物理处理器。假设某台计算机具有 8 个 CPU 核心，而虚拟机需要其中的 2 个，请按如下所示更改其配置文件：

```
vcpus=2  
cpus="2,3"
```

以上示例为 VM Guest 分配了 2 个专用处理器，即第三个和第四个处理器（从 0 开始算起为第 2 和第 3 个）。如果您需要分配更多的物理处理器，请使用 `cpus="2-8"` 语法。

如果您需要以热插拔方式为名为 “alice” 的 Guest 更改 CPU 分配，请在相关 Dom0 上使用以下命令：

```
> sudo xl vcpu-set alice 2  
> sudo xl vcpu-pin alice 0 2  
> sudo xl vcpu-pin alice 1 3
```

此示例为该 Guest 分配了 2 个专用物理处理器，并将其 VCPU 0 和 1 分别绑定到物理处理器 2 和 3。现在检查分配：

```
> sudo xl vcpu-list alice
```

Name	ID	VCPUs	CPU	State	Time(s)	CPU Affinity
alice	4	0	2	-b-	1.9	2-3
alice	4	1	3	-b-	2.8	2-3

29.6 HVM 功能

在 Xen 中，某些功能仅适用于全虚拟化域。这些功能极少用到，但在特定环境中仍可能具有独特的作用。

29.6.1 指定引导时使用的引导设备

与使用物理硬件时一样，有时需要从其他设备而非 VM Guest 自己的引导设备来引导该 Guest。对于全虚拟计算机，可以在域 xl 配置文件中使用 `boot` 参数来选择引导设备：

```
boot = BOOT_DEVICE
```

`BOOT_DEVICE` 可以是 `c`（表示硬盘）、`d`（表示 CD-ROM）或 `n`（表示网络/PXE）。您可以指定多个选项，系统会按指定的顺序尝试这些选项。例如，

```
boot = dc
```

从 CD-ROM 引导，如果 CD-ROM 不可引导，则回退到硬盘。

29.6.2 更改 Guest 的 CUID

要将 VM Guest 从一台 VM 主机服务器迁移到另一台 VM 主机服务器，VM Guest 系统只能使用这两个 VM 主机服务器系统上均会提供的 CPU 功能。如果两台主机上的实际 CPU 不同，可能需要在启动 VM Guest 之前隐藏某些功能。这样便可以在两台主机之间迁移 VM Guest。对于全虚拟化 Guest，可以通过配置可供 Guest 使用的 `cpuid` 来实现此目的。

要大致了解当前 CPU 的情况，请查看 `/proc/cpuinfo`。其中包含定义当前 CPU 的所有重要信息。

要重新定义 CPU，请先查看 CPU 供应商的相应 `cpuid` 定义。可从以下网站获取这些定义：

Intel

<https://www.intel.com/Assets/PDF/appnote/241618.pdf> 

```
cpuid = "host,tm=0,sse3=0"
```

语法为 `key=value` 对的逗号分隔列表，前面加上 `host` 一词。一些键接受数字值，其他所有键接受描述功能位作用的单个字符。有关 `cpuid` 键的完整列表，请参见 [man 5 xl.cfg](#)。可使用下面的值更改相应的位：

1

将对应的位强制为 1

0

将对应的位强制为 0

x

使用默认策略的值

k

使用主机定义的值

s

与 k 类似，但会在迁移后保留值



提示

请记住，位是从位 0 开始从右向左计数的。

29.6.3 增加 PCI-IRQ 的数量

如果您需要增加 Dom0 和/或 VM Guest 可用的 PCI-IRQ 的默认数量，可以修改 Xen 内核命令。使用命令 `extra_guest_irqs=DOMU_IRGS,DOM0_IRGS`。第一个可选数字 `DOMU_IRGS` 是所有 VM Guest 公用的数字，第二个可选数字 `DOM0_IRGS`（前面带有逗号）用于 Dom0。更改 VM Guest 的设置不会影响 Dom0，反之亦然。例如，要在不更改 VM Guest 的情况下更改 Dom0，请使用

```
extra_guest_irqs=,512
```

29.7 虚拟 CPU 调度

Xen 超级管理程序会将虚拟 CPU 单独调度到不同的物理 CPU 中。在每个核心均支持多个线程的新式 CPU 中，虚拟 CPU 可以在同一核心上的不同线程中运行，因此会相互影响。在一个线程中运行的虚拟 CPU 的性能可能会受到其他线程中的其他虚拟 CPU 所执行操作的显著影响。如果这些虚拟 CPU 属于不同的 Guest 系统，这些 Guest 可能会相互影响。具体影响各不相同，轻则 Guest CPU 用时统计出现偏差，重则遭遇**边信道攻击**等更坏的情况。

调度粒度可以解决此问题。您可以使用 Xen 引导参数在引导时指定粒度：

```
sched-gran=GRANULARITY
```

请将 GRANULARITY 替换为下列其中一项：

cpu

Xen 超级管理程序的常规调度。不同 Guest 的虚拟 CPU 可以共享同一个物理 CPU 核心。此为默认设置。

core

一个虚拟核心的虚拟 CPU 始终一起调度到一个物理核心中。来自不同虚拟核心的两个或更多个虚拟 CPU 永远不会调度到同一个物理核心中。因此，在某些物理核心中，可能有多个 CPU 保持空闲状态，即使存在需要运行的虚拟 CPU 也不例外。对性能的影响取决于 Guest 系统内部正在运行的实际工作负载。在我们所分析的大多数案例中，观察到的性能下降情况（尤其是在承受很高负载的情况下）比禁用超线程（使所有核心仅保留一个线程）要轻一些（请参见 <https://xenbits.xen.org/docs/unstable/misc/xen-command-line.html#smt-x86> 上的 smt 引导选项）。

socket

粒度甚至可以提高到 CPU 插槽级别。

30 管理任务

30.1 引导加载程序

引导加载程序可控制虚拟化软件的引导和运行方式。您可以使用 YaST 或者通过直接编辑引导加载程序配置文件来修改引导加载程序属性。

您可以通过 YaST > 系统 > 引导加载程序访问 YaST 引导加载程序。单击引导加载程序选项选项卡，对于默认引导项，选择包含 Xen 内核的行。



引导加载器设置

引导代码选项(D) 内核参数(K) 引导加载程序选项(L)

超时 (以秒计) (T)
8 ☐ 引导时隐藏菜单(H)

默认引导项(D)
SLES 15-SP4, with Xen hypervisor

☐ 使用密码保护引导加载器(E)
☒ 仅保护启动项不被修改(R)

GRUB2 用户 root 的口令(P) 再次键入密码(T)

帮助(H) 取消(C) 确定(O)

图 30.1：引导加载程序设置

单击确定进行确认。您下次引导主机时，引导加载程序可以提供 Xen 虚拟化环境。

您可以使用引导加载程序来指定功能，例如：

- 传递内核命令行参数。
- 指定内核映像和初始 RAM 磁盘。

- 选择特定的超级管理程序。
- 向超级管理程序传递其他参数。有关完整的参数列表，请参见 <https://xenbits.xen.org/docs/unstable/misc/xen-command-line.html>。

可以通过编辑 `/etc/default/grub` 文件来自定义虚拟化环境。将以下行添加到此文件中：`GRUB_CMDLINE_XEN="<boot_parameters>"`。编辑该文件后，请不要忘记运行 `grub2-mkconfig -o /boot/grub2/grub.cfg`。

30.2 稀疏映像文件和磁盘空间

如果主机的物理磁盘已没有可用空间，使用基于稀疏映像文件的虚拟磁盘的虚拟机将无法向其磁盘写入数据。因此，它会报告 I/O 错误。

如果发生这种情况，您应该释放物理磁盘上的可用空间，重新挂载虚拟机的文件系统，然后将文件系统重新设置为读写模式。

要检查稀疏映像文件的实际磁盘要求，请使用 `du -h <image file>` 命令。

要增加稀疏映像文件的可用空间，请先增加文件大小，然后增加文件系统的大小。



警告：在调整大小之前备份文件

改动分区或稀疏文件的大小始终要承担数据失败的风险。在未备份的情况下，请不要执行此操作。

可以在 VM Guest 运行时联机调整映像文件的大小。使用以下命令增加稀疏映像文件的大小：

```
> sudo dd if=/dev/zero of=<image file> count=0 bs=1M seek=<new size in MB>
```

例如，要将 `/var/lib/xen/images/sles/disk0` 文件的大小增至 16GB，请使用以下命令：

```
> sudo dd if=/dev/zero of=/var/lib/xen/images/sles/disk0 count=0 bs=1M  
seek=16000
```



注意：增加非稀疏映像大小

还可以增加设备的非稀疏映像文件的大小，但您必须知道上一个映像的具体结束位置。使用 `seek` 参数指向映像文件的末尾，然后使用如下所示的命令：

```
> sudo dd if=/dev/zero of=/var/lib/xen/images/sles/disk0 seek=8000 bs=1M  
count=2000
```

请务必正确使用 `seek`，否则可能发生数据丢失情况。

如果在 VM Guest 运行时执行大小调整操作，另外还需调整向 VM Guest 提供映像文件的循环设备的大小。首先使用以下命令检测正确的循环设备：

```
> sudo losetup -j /var/lib/xen/images/sles/disk0
```

然后使用以下命令调整循环设备（例如 `/dev/loop0`）的大小：

```
> sudo losetup -c /dev/loop0
```

最后使用 `fdisk -l /dev/xvdb` 命令检查 Guest 系统中块设备的大小。请将设备名称替换为已增大的磁盘的名称。

调整稀疏文件中的文件系统大小所涉及到的工具取决于实际文件系统。《存储管理指南》中对此做了详细说明。

30.3 迁移 Xen VM Guest 系统

在 Xen 中，可将 VM Guest 系统从一台 VM 主机服务器迁移到另一台 VM 主机服务器，而且几乎不会造成服务中断。例如，可以使用此功能将一个繁忙的 VM Guest 转移到一台硬件更强大或者尚无负载的 VM 主机服务器。或者，如果 VM 主机服务器的某项服务不能中断，可将此计算机上运行的所有 VM Guest 系统迁移到其他计算机，以避免该服务中断。这只是其中的两个例子 — 在您的个人使用情形中，可能还有许多其他迁移原因。

在开始之前，需要了解有关 VM 主机服务器的某些应预先注意的事项：

- 所有 VM 主机服务器系统应使用类似的 CPU。CPU 的频率并不是很重要，但它们应该使用同一 CPU 系列。要获取有关所用 CPU 的详细信息，请使用 `cat /proc/cpuinfo`。第 30.3.1 节“检测 CPU 功能”中提供了有关比较主机 CPU 功能的更多细节。
- 特定 Guest 系统使用的所有资源必须已在所有相关 VM 主机服务器系统上提供 — 例如，使用的所有块设备必须在两个 VM 主机服务器系统上均存在。
- 如果迁移过程涉及的主机在不同的子网中运行，请确保 Guest 可以使用 DHCP 中继，或者针对采用静态网络配置的 Guest 手动设置网络。
- 使用 `PCI Pass-Through` 等特殊功能可能会造成问题。为可能会在不同 VM 主机服务器系统之间迁移 VM Guest 系统的环境部署虚拟机时，请不要实施这些功能。
- 为了快速迁移，必须设置一个快速网络。如果可能，请使用 GB 以太网和快速交换机。部署 VLAN 也可能有助于避免冲突。

30.3.1 检测 CPU 功能

可以使用 `cpuid` 和 `xen_maskcalc.py` 工具，将您要从中迁移源 VM Guest 的主机上的 CPU 功能与目标主机上的 CPU 功能进行比较。这样，您就可以更好地预测 Guest 迁移是否会成功。

1. 在预期要运行或接收迁移的 VM Guest 的每个 Dom0 上运行 `cpuid -lr` 命令，然后在文本文件中捕获输出，例如：

```
tux@vm_host1 > sudo cpuid -lr > vm_host1.txt
tux@vm_host2 > sudo cpuid -lr > vm_host2.txt
tux@vm_host3 > sudo cpuid -lr > vm_host3.txt
```

2. 将所有输出文本文件复制到装有 `xen_maskcalc.py` 脚本的主机上。
3. 针对所有输出文本文件运行 `xen_maskcalc.py` 脚本：

```
> sudo xen_maskcalc.py vm_host1.txt vm_host2.txt vm_host3.txt
cpuid = [
    "0x00000001:ecx=x00xxxxxx0xxxxxxxx00xxxxxxxx",
    "0x00000007,0x00:ebx=xxxxxxxxxxxxxxxx00x0000x0x00"
```

```
] ]
```

4. 将输出的 `cpuid=[...]` 配置片段复制到迁移的 Guest `domU.cfg` 的 `xl` 配置中，或复制到其 `libvirt` 的 XML 配置中。
5. 使用**经过删减的** CPU 配置启动源 Guest。现在，Guest 只能使用每台主机上提供的 CPU 功能。



提示

`libvirt` 还支持计算用于迁移的基线 CPU。有关详细信息，请参考《虚拟化最佳实践》文章。

30.3.1.1 更多信息

<https://etallen.com/cpuid.html> 上提供了有关 `cpuid` 的更多细节。

您可以从 https://github.com/twizted/xen_maskcalc 下载最新版本的 CPU 掩码计算器。

30.3.2 准备要迁移的块设备

VM Guest 系统所需的块设备必须已在所有相关 VM 主机服务器系统上提供。要做到这一点，可以实施特定类型的共享存储，用于充当所迁移 VM Guest 系统的根文件系统的容器。常见的做法包括：

- 可以设置 `iSCSI`，以便可从不同的系统同时访问相同的块设备。有关 `iSCSI` 的详细信息，请参见《存储管理指南》，第 15 章“经由 IP 网络的大容量存储：iSCSI”。
- `NFS` 是广泛使用的根文件系统，用户可从不同的位置轻松访问该文件系统。有关详细信息，请参见《存储管理指南》，第 19 章“通过 NFS 共享文件系统”。
- 如果只涉及到两个 VM 主机服务器系统，则可以使用 `DRBD`。这会在一定程度上提高数据安全，因为此方法会通过网络镜像使用的数据。有关详细信息，请参见 <https://documentation.suse.com/sle-ha-15/> 上的 SUSE Linux Enterprise High Availability 15 SP7 文档。

- 如果提供的硬件允许对相同磁盘进行共享访问，则还可以使用 [SCSI](#)。
- [NPIV](#) 是使用光纤通道磁盘的特殊模式。但在这种情况下，所有迁移主机都必须挂接到同一个光纤通道交换机。有关 NPIV 的详细信息，请参见第 28.1 节 “[将物理存储设备映射到虚拟磁盘](#)”。一般而言，如果光纤通道环境支持 4 Gbps 或更快的连接，便可使用此方式。

30.3.3 迁移 VM Guest 系统

VM Guest 系统的实际迁移是使用以下命令完成的：

```
> sudo xl migrate <domain_name> <host>
```

迁移速度取决于将占用内存的数据保存到磁盘、发送到 VM 主机服务器并在该服务器中加载的速度。也就是说，迁移小型 VM Guest 系统可能比迁移包含大量内存的大型系统的速度更快。

30.4 监控 Xen

在对众多虚拟 Guest 进行常规运维时，检查所有不同 VM Guest 系统的健全性是必不可少的功能。除了系统工具以外，Xen 还提供了数个工具用于收集有关系统的信息。



提示：监控 VM 主机服务器

可以通过虚拟机管理器监控 VM 主机服务器的基本情况（I/O 和 CPU）。有关细节，请参见第 11.7.1 节 “[使用虚拟机管理器进行监控](#)”。

30.4.1 使用 [xentop](#) 监控 Xen

用于收集有关 Xen 虚拟环境的信息的首选终端应用程序是 [xentop](#)。请注意，此工具需要一个相当宽的终端，否则它会在显示内容中插入换行符。

[xentop](#) 提供多个命令键，可供您查看有关所监控系统的详细信息。例如：

D

更改屏幕两次刷新相隔的延迟时间。

N

同时显示网络统计数据。请注意，只会显示标准配置。如果您使用路由网络等特殊配置，将不会显示网络。

B

显示相应的块设备及其累计使用次数。

有关 **xentop** 的详细信息，请参见手册页 **man 1 xentop**。



提示：virt-top

libvirt 提供不限定超级管理程序的工具 **virt-top**，建议使用此工具来监控 VM Guest。

有关详细信息，请参见第 11.7.2 节“使用 **virt-top** 进行监控”。

30.4.2 其他工具

我们还提供了许多系统工具来帮助您监控或调试运行中的 SUSE Linux Enterprise 系统。《系统分析和微调指南》，第 2 章“系统监控实用程序”中介绍了其中一些工具。以下工具特别适合用于监控虚拟化环境：

ip

命令行实用程序 **ip** 可用于监控任意网络接口。如果您设置了路由网络或者应用了伪装网络，此实用程序特别有用。要监控名为 alice.0 的网络接口，请运行以下命令：

```
> watch ip -s link show alice.0
```

bridge

在标准设置中，所有 Xen VM Guest 系统都会挂接到虚拟网桥。使用 **bridge** 可以确定网桥与 VM Guest 系统中虚拟网络适配器之间的连接。例如，**bridge link** 的输出可能如下所示：

```
2: eth0 state DOWN : <NO-CARRIER, ...,UP> mtu 1500 master br0
8: vnet0 state UNKNOWN : <BROADCAST, ...,LOWER_UP> mtu 1500 master virbr0 \
state forwarding priority 32 cost 100
```

此输出表明系统上定义了两个虚拟网桥。一个网桥连接到物理以太网设备 `eth0`，另一个网桥连接到 VLAN 接口 `vnet0`。

iptables-save

使用伪装网络时，或者设置了多个以太网接口并与防火墙设置结合使用时，此工具特别有用，它可以帮助检查当前的防火墙规则。

iptables 命令可用于检查所有不同的防火墙设置。要列出某个链甚至整个设置的所有规则，可以使用 **iptables-save** 或 **iptables -S** 命令。

30.5 提供 VM Guest 系统的主机信息

在标准 Xen 环境中，VM Guest 系统只能获得有关运行它的 VM 主机服务器系统的有限信息。如果某个 Guest 应该了解有关运行它的 VM 主机服务器的更多信息，`vhostmd` 可为选定 Guest 提供详细信息。要设置系统以运行 `vhostmd`，请执行以下操作：

1. 在 VM 主机服务器上安装 `vhostmd` 软件包。
2. 要在配置中添加或删除 `metric` 部分，请编辑文件 `/etc/vhostmd/vhostmd.conf`。不过，默认设置也可正常工作。
3. 使用以下命令检查 `vhostmd.conf` 配置文件的有效性：

```
> cd /etc/vhostmd
> xmllint --postvalid --noout vhostmd.conf
```

4. 使用 **`sudo systemctl start vhostmd`** 命令启动 `vhostmd` 守护程序。

如果系统启动期间应自动启动 `vhostmd`，请运行以下命令：

```
> sudo systemctl enable vhostmd
```

5. 使用以下命令将映像文件 `/dev/shm/vhostmd0` 挂接到名为 `alice` 的 VM Guest 系统：

```
> xl block-attach opensuse /dev/shm/vhostmd0,,xvdb,ro
```

6. 在 VM Guest 系统上登录。
7. 安装客户端软件包 `vm-dump-metrics`。

8. 运行 `vm-dump-metrics` 命令。要将结果保存到文件中，请使用选项 `-d <filename>`。

`vm-dump-metrics` 返回的结果为 XML 输出。相应的指标项遵循 DTD `/etc/vhostmd/metric.dtd`。

有关详细信息，请参见 VM 主机服务器系统上的手册页 `man 8 vhostmd` 和 `/usr/share/doc/vhostmd/README`。在 Guest 上，请参见手册页 `man 1 vm-dump-metrics`。

31 XenStore：在域之间共享的配置数据库

本章介绍有关 XenStore 的基本信息、它在 Xen 环境中的作用、XenStore 所使用的文件的目录结构，并提供了 XenStore 命令的说明。

31.1 简介

XenStore 是在 Dom0 中所运行的 VM Guest 与管理工具之间共享的配置和状态信息数据库。VM Guest 和管理工具对 XenStore 进行读取和写入以传达配置信息、状态更新和状态更改。XenStore 数据库由 Dom0 管理，支持读取和写入密钥等简单操作。可以通过监测关注项将 XenStore 中发生的任何更改通知给 VM Guest 和管理工具。xenstored 守护程序由 xencommons 服务管理。

XenStore 位于 Dom0 上的单个数据库文件 /var/lib/xenstored/tdb（tdb 表示**树数据库**）中。

31.2 文件系统接口

XenStore 数据库内容由与 /proc 类似的虚拟文件系统表示（有关 /proc 的详细信息，请参见《系统分析和微调指南》，第 2 章“系统监控实用程序”，第 2.6 节“/proc 文件系统”）。树有三条主路径：/vm、/local/domain 和 /tool。

- /vm - 存储有关 VM Guest 配置的信息。
- /local/domain - 存储有关本地节点上的 VM Guest 的信息。
- /tool - 存储有关多个工具的一般信息。



提示

每个 VM Guest 都有两个不同的 ID 编号。即使 VM Guest 迁移到其他计算机，**全局唯一标识符** (UUID) 也保持不变。**域标识符** (DOMID) 是表示正在运行的特定实例的标识号。将 VM Guest 迁移到其他计算机后，此编号通常会改变。

31.2.1 XenStore 命令

可使用以下命令操作 XenStore 数据库的文件系统结构：

xenstore-ls

显示 XenStore 数据库的完整转储。

xenstore-readpath_to_xenstore_entry

显示指定 XenStore 项的值。

xenstore-existsxenstore_path

报告指定的 XenStore 路径是否存在。

xenstore-listxenstore_path

显示指定 XenStore 路径的所有子项。

xenstore-writepath_to_xenstore_entry

更新指定 XenStore 项的值。

xenstore-rmxenstore_path

去除指定的 XenStore 项或目录。

xenstore-chmodxenstore_pathmode

更新对指定 XenStore 路径的读取/写入权限。

xenstore-control

将一条命令（例如，用于触发完整性检查）发送到 xenstored 后端。

31.2.2 /vm

/vm 路径按每个 VM Guest 的 UUID 编制索引，用于存储虚拟 CPU 数量和分配的内存量等配置信息。每个 VM Guest 都有一个 /vm/<uuid> 目录。要列出目录内容，请使用 **xenstore-list**。

```
> sudo xenstore-list /vm
00000000-0000-0000-0000-000000000000
9b30841b-43bc-2af9-2ed3-5a649f466d79-1
```

输出的第一行属于 Dom0，第二行属于正在运行的 VM Guest。以下命令会列出与 VM Guest 相关的所有项：

```
> sudo xenstore-list /vm/9b30841b-43bc-2af9-2ed3-5a649f466d79-1
image
rtc
device
pool_name
shadow_memory
uuid
on_reboot
start_time
on_poweroff
bootloader_args
on_crash
vcpus
vcpu_avail
bootloader
name
```

要读取某个项（例如，专用于 VM Guest 的虚拟 CPU 数量）的值，请使用 **xenstore-read**：

```
> sudo xenstore-read /vm/9b30841b-43bc-2af9-2ed3-5a649f466d79-1/vcpus
1
```

所选 /vm/<uuid> 项的列表如下：

uuid

VM Guest 的 UUID，在迁移过程中不会变化。

on_reboot

指定响应重引导请求时是要销毁还是重启动 VM Guest。

on_poweroff

指定响应暂停请求时是要销毁还是重启动 VM Guest。

on_crash

指定响应崩溃事件时是要销毁还是重启动 VM Guest。

vcpus

分配给 VM Guest 的虚拟 CPU 数量。

vcpu_avail

VM Guest 的活动虚拟 CPU 的位掩码。位掩码中有多个位等于 vcpus 的值，其中有一个位是为每个联机虚拟 CPU 设置的。

name

VM Guest 的名称。

普通的 VM Guest（不是 Dom0）使用 /vm/<uuid>/image 路径：

```
> sudo xenstore-list /vm/9b30841b-43bc-2af9-2ed3-5a649f466d79-1/image
ostype
kernel
cmdline
ramdisk
dmargs
device-model
display
```

使用的项的解释如下：

ostype

VM Guest 的操作系统类型。

kernel

Dom0 上指向 VM Guest 内核的路径。

cmdline

引导时对 VM Guest 使用的内核命令行。

ramdisk

Dom0 上指向 VM Guest RAM 磁盘的路径。

dmargs

显示传递给 QEMU 进程的参数。如果您使用 **ps** 查看 QEMU 进程，看到的参数应该与 /vm/<uuid>/image/dmargs 中的相同。

31.2.3 `/local/domain/<domid>`

此路径按运行中的域 (VM Guest) ID 编制索引，包含有关运行中的 VM Guest 的信息。请记住，在迁移 VM Guest 期间，域 ID 会变化。可用的项如下：

vm

此 VM Guest 的 /vm 目录的路径。

on_reboot, on_poweroff, on_crash, name

请参见第 31.2.2 节 “/vm” 中的相同选项。

domid

VM Guest 的域标识符。

cpu

当前 VM Guest 固定到的 CPU。

cpu_weight

出于调度目的分配给 VM Guest 的权重。权重越高，使用物理 CPU 的频率就越高。

除了上面所述的各个项以外，/local/domain/<domid> 下的多个子目录也包含特定的项。要查看所有可用的项，请参见《[XenStore Reference](https://wiki.xen.org/wiki/XenStore_Reference)》（XenStore 参考）(https://wiki.xen.org/wiki/XenStore_Reference)。

/local/domain/<domid>/memory

包含内存信息。/local/domain/<domid>/memory/target 包含 VM Guest 的目标内存大小（以 KB 为单位）。

/local/domain/<domid>/console

包含有关 VM Guest 所用控制台的信息。

/local/domain/<domid>/backend

包含有关 VM Guest 所用的所有后端设备的信息。该路径包含 VM Guest 自己的子目录。

/local/domain/<domid>/device

包含有关 VM Guest 的前端设备的信息。

/local/domain/<domid>/device-misc

包含有关设备的其他信息。

/local/domain/<domid>/store

包含有关 VM Guest 的存储空间的信息。

32 使用 Xen 作为高可用性虚拟化主机

与在专用硬件上运行每台服务器的设置相比，将两台 Xen 主机设置为故障转移系统可以带来多项优势。

- 一台服务器发生故障不会导致出现重大服务中断情况。
- 购置一台大型计算机通常比购置多台小型计算机要便宜。
- 按需添加新服务器的工作简单无比。
- 服务器的利用率可得到改进，而这又会对系统的能耗产生正面影响。

第 30.3 节 “迁移 Xen VM Guest 系统” 中介绍了 Xen 主机的迁移设置。下面描述了几种典型方案。

32.1 使用远程存储设备实现 Xen HA

Xen 可以直接向相应的 Xen Guest 系统提供多个远程块设备。这些设备包括 iSCSI、NPIV 和 NBD。它们可用于执行实时迁移。存储系统准备就绪后，请先尝试使用您已在网络中使用的相同设备类型。

如果存储系统不可直接使用，但具有通过 NFS 提供所需空间的能力，您也可以在 NFS 上创建映像文件。如果所有 Xen 主机系统上都提供了 NFS，则此方法还可以实现 Xen Guest 的实时迁移。

设置新系统时，其中一个主要考虑因素为是否应该实施专用的存储区域网络。可行的方法如下：

表 32.1：XEN 远程存储

方法	复杂性	注释
以太网	低	所有块设备流量都将通过用于传送网络流量的同一以太网接口传送。这可能会使 Guest 的性能受到限制。

方法	复杂性	注释
专用于储存的以太网。	中	通过专用以太网接口运行存储流量可以消除服务器端的瓶颈。但是，为自己的网络规划自己的 IP 地址范围时，以及规划专用于存储设备的 VLAN 时，您需要考虑某些因素。
NPIV	高	NPIV 是用于虚拟化光纤通道连接的一种方法。通过使用最低支持 4 Gbit/秒的数据传输速率并允许复杂存储系统设置的适配器来提供此功能。

通常，1 Gbit/秒的以太网设备可以充分利用典型的硬盘或存储系统。使用快速存储系统时，此类以太网设备可能会限制系统的速度。

32.2 使用本地存储设备实现 Xen HA

出于空间和预算原因，有时可能需要依赖于 Xen 主机系统本地的存储设备。如果仍想实现实时迁移，需要构建镜像到两台 Xen 主机的块设备。可用于执行此操作的软件称为分布式复制块设备 (DRBD)。

如果需要设置一个使用 DRBD 在两台 Xen 主机之间镜像块设备或文件的系统，这两台主机应使用相同的硬件。如果其中一台主机的硬盘速度较慢，则两台主机的速度会受到同样的限制。

在设置期间，所需的每个块设备均应使用自己的 DRBD 设备。设置此类系统是一个复杂的任务。

32.3 Xen HA 和专用网桥

使用需要相互通讯的多个 Guest 系统时，可以通过普通的接口来进行通讯。但出于安全原因，建议您创建一个仅连接到 Guest 系统的网桥。

在还需要支持实时迁移的 HA 环境中，此类专用网桥必须连接到其他 Xen 主机。使用专用物理以太网设备和专用网络可以做到这一点。

另一种实现方法是使用 VLAN 接口。在这种情况下，所有流量都将通过普通以太网接口传送。不过，VLAN 接口不会收到普通流量，因为只有标记为要传送到正确 VLAN 的 VLAN 包会被转发。

有关 VLAN 接口设置的详细信息，请参见[第 9.1.1.4 节 “使用 VLAN 接口”](#)。

33 Xen：将半虚拟 (PV) Guest 转换为全虚拟 (FV/HVM) Guest

本章介绍如何将 Xen 半虚拟机转换为 Xen 全虚拟机。

过程 33.1：GUEST 端

要在 FV 模式下启动 Guest，需要在 Guest 中执行以下步骤。

1. 在转换 Guest 之前，安装所有待应用的补丁并重引导 Guest。
2. FV 计算机使用 `-default` 内核。如果尚未安装此内核，请安装 `kernel-default` 软件包（在 PV 模式下运行时）。
3. PV 计算机通常使用 `vda*` 这样的磁盘名称。这些名称必须更改为 FV `hd*` 语法。此更改必须在以下文件中进行：

- `/etc/fstab`
- `/boot/grub/menu.lst`（仅适用于 SLES 11）
- `/boot/grub*/device.map`
- `/etc/sysconfig/bootloader`
- `/etc/default/grub`（SLES 12、15、openSUSE）



注意：建议使用 UUID

应在 `/etc/fstab` 中使用 UUID 或逻辑卷。通过 UUID 可以方便地使用挂接的网络存储设备、多路径和虚拟化。要确定磁盘的 UUID，请使用 `blkid` 命令。

4. 为了避免在使用所需模块重新生成 `initrd` 时出现任何错误，可以使用 `ln` 创建一个从 `/dev/hda2` 到 `/dev/xvda2` 等的符号链接：

```
ln -sf /dev/xvda2 /dev/hda2
ln -sf /dev/xvda1 /dev/hda1
.....
```

5. PV 和 FV 计算机使用不同的磁盘和网络驱动程序模块。必须手动将这些 PV 模块添加到 `initrd`。需要的模块为 `xen-vbd`（用于磁盘）和 `xen-vnif`（用于网络）。这些是全虚拟化 VM Guest 仅有的 PV 驱动程序。所有其他模块（例如 `ata_piix`、`ata_generic` 和 `libata`）应该会自动添加。



提示：将模块添加到 `initrd`

- 在 SLES 11 上，可以将模块添加到 `/etc/sysconfig/kernel` 文件中的 `INITRD_MODULES` 行。例如：

```
INITRD_MODULES="xen-vbd xen-vnif"
```

运行 **dracut** 以构建包含这些模块的新 `initrd`。

- 在 SLES 12、15 和 openSUSE 上，打开或创建 `/etc/dracut.conf.d/10-virt.conf`，并按以下示例所示添加一行来使用 `force_drivers` 添加这些模块（注意前导空格）。

```
force_drivers+=" xen-vbd xen-vnif"
```

运行 **dracut -f --kver `KERNEL_VERSION-default`** 以构建包含所需模块的新 `initrd`（用于内核的默认版本）。

确定您的内核版本： 使用 `uname -r` 命令可获取系统上当前使用的版本。

6. 在关闭 Guest 之前，使用 **yast bootloader** 将默认引导参数设置为 `-default` 内核。
7. 在 SUSE Linux Enterprise Server 11 下，如果 Guest 上正在运行 X 服务器，您需要调整 `/etc/X11/xorg.conf` 文件来调整 X 驱动程序。搜索 `fbdev` 并将其更改为 `cirrus`。

```
Section "Device"
    Driver      "cirrus"
    . . . . .
EndSection
```



注意：SUSE Linux Enterprise Server 12/15 和 Xorg

在 SUSE Linux Enterprise Server 12/15 下，Xorg 将自动调整 X 服务器能够正常运行所需的驱动程序。

8. 关闭 Guest。

过程 33.2：主机端

以下步骤说明了需要在主机上执行的操作。

1. 要以 FV 模式启动 Guest，必须修改 VM 的配置以匹配 FV 配置。使用 **virsh edit [DOMAIN]** 可轻松编辑 VM 的配置。建议进行以下更改：

- 确保在 OS 部分中，将 `machine`、`type` 和 `loader` 项中的 `xenpv` 更改为 `xenfv`。更新后的 OS 部分应如下所示：

```
<os>
    <type arch='x86_64' machine='xenfv'>hvm</type>
    <loader>/usr/lib/xen/boot/hvmloder</loader>
    <boot dev='hd' />
</os>
```

- 在 OS 部分，去除所有特定于 PV Guest 的内容：

- `<bootloader>pygrub</bootloader>`

- `<kernel>/usr/lib/grub2/x86_64-xen/grub.xen</kernel>`

- `<cmdline>xen-fbfront.video=4,1024,768</cmdline>`

- 在 `devices` 部分，采用以下形式添加 `qemu` 模拟器：

```
<emulator>/usr/lib/xen/bin/qemu-system-i386</emulator>
```

- 更新磁盘配置，使目标设备和总线使用 FV 语法。这需要将 `xen` 磁盘总线替换为 `ide`，并将 `vda` 目标设备替换为 `hda`。更改应如下所示：

```
<target dev='hda' bus='ide' />
```

- 将鼠标和键盘的总线从 xen 更改为 ps2。另外添加一个新的 USB 绘图板设备：

```
<input type='mouse' bus='ps2' />
    <input type='keyboard' bus='ps2' />
<input type='tablet' bus='usb' />
```

- 将控制台目标类型从 xen 更改为 serial：

```
<console type='pty'>
    <target type='serial' port='0' />
</console>
```

- 将视频配置从 xen 更改为 cirrus，其中 VRAM 大小为 8 MB：

```
<video>
    <model type='cirrus' vram='8192' heads='1' primary='yes' />
</video>
```

- 如果需要，向 VM 的功能添加 acpi 和 apic：

```
<features>
    <acpi />
    <apic />
</features>
```

2. 启动 Guest（使用 virsh 或 virt-manager）。如果 Guest 运行的是 kernel-default（通过 uname -a 校验），计算机将以全虚拟模式运行。



注意：guestfs-tools

要编写此过程的脚本，或直接在磁盘映像上工作，可以使用 guestfs-tools 套件（有关详细信息，请参见第 21.3 节“Guestfs 工具”）。有多种工具可以帮助修改磁盘映像。

V 使用 QEMU 管理虚拟机

- 34 QEMU 概述 311
- 35 设置 KVM VM 主机服务器 312
- 36 Guest 安装 322
- 37 使用 qemu-system-ARCH 运行虚拟机 337
- 38 使用 QEMU 监控器管理虚拟机 366

34 QEMU 概述

QEMU 是一个快捷的跨平台开源计算机模拟器，可以模拟许多硬件体系结构。QEMU 可让您在现有系统（VM 主机服务器）之上运行未经修改的完整操作系统 (VM Guest)。您还可以使用 QEMU 进行调试 — 可以轻松停止正在运行的虚拟机、检查其状态、保存并在以后恢复其状态。

QEMU 主要由以下部分构成：

- 处理器模拟器。
- 模拟的设备，例如显卡、网卡、硬盘或鼠标。
- 用于将模拟的设备连接到相关主机设备的通用设备。
- 调试器。
- 用来与模拟器交互的用户界面。

QEMU 是 KVM 和 Xen 虚拟化的核心，在这些虚拟化环境中提供常规的计算机模拟。Xen 在使用 QEMU 时会对用户隐藏部分功能，而 KVM 在使用 QEMU 时会透明地公开大部分 QEMU 功能。如果 VM Guest 硬件体系结构与 VM 主机服务器的体系结构相同，QEMU 便可以利用 KVM 加速的优势（SUSE 仅支持加载了 KVM 加速的 QEMU）。

除了提供核心虚拟化基础架构以及特定于处理器的驱动程序以外，QEMU 还提供特定于体系结构的用户空间程序来管理 VM Guest。根据具体的体系结构，此程序是以下其中一项：

- `qemu-system-i386`
- `qemu-system-s390x`
- `qemu-system-x86_64`
- `qemu-system-aarch64`

在后面的章节中，此命令称为 `qemu-system-ARCH`；示例中使用的是 `qemu-system-x86_64` 命令。

35 设置 KVM VM 主机服务器

本节介绍如何设置并使用 SUSE Linux Enterprise Server 15 SP7 作为基于 QEMU-KVM 的虚拟机主机。



提示：资源

虚拟 Guest 系统所需的硬件资源与将其安装在物理机上时所需的资源相同。您打算在主机系统上运行的 Guest 越多，需要添加到 VM 主机服务器的硬件资源（CPU、磁盘、内存和网络）就越多。

35.1 CPU 的虚拟化支持

要运行 KVM，您的 CPU 必须支持虚拟化，并且需要在 BIOS 中启用虚拟化。[/proc/cpuinfo](#) 文件包含有关 CPU 功能的信息。

要确定您的系统是否支持虚拟化，请参见[第 7.1.1 节 “KVM 硬件要求”](#)。

35.2 所需的软件

需在 KVM 主机上安装多个软件包。要安装所有必要的软件包，请执行以下操作：

1. 校验是否已安装 [yast2-vm](#) 软件包。此软件包是 YaST 的配置工具，可以简化虚拟化超级管理程序的安装过程。
2. 运行 YaST › 虚拟化 › 安装管理程序和工具。

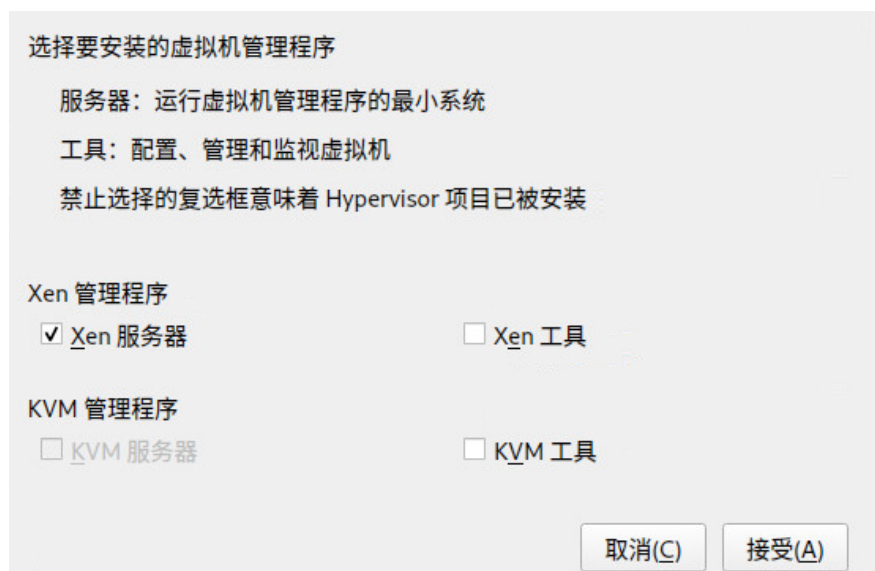


图 35.1：安装 KVM 超级管理程序和工具

3. 选择 KVM 服务器，最好也选择 KVM 工具，然后单击接受确认。
4. 在安装过程中，您可以选择让 YaST 自动为您创建网桥。如果您不打算另外为虚拟 Guest 使用一块物理网卡，那么将 Guest 计算机连接到网络的标准方式就是使用网桥。

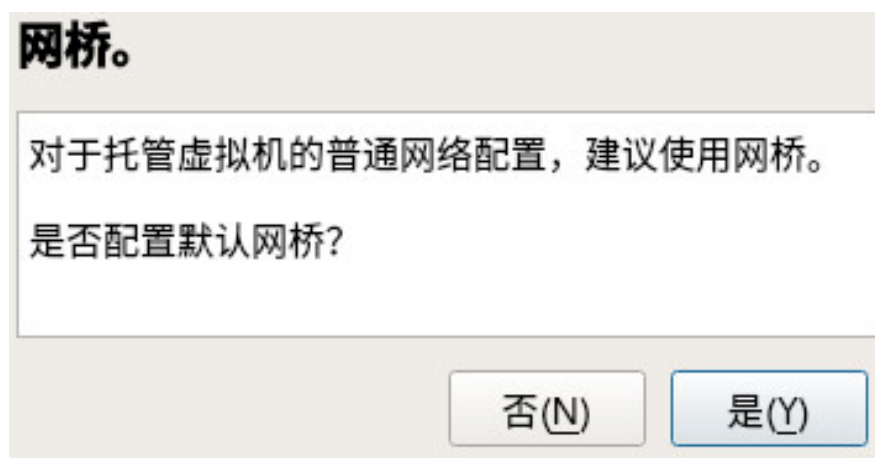


图 35.2：网桥

5. 安装所有所需的软件包（并激活新网络设置）后，尝试加载适用于 CPU 类型的 KVM 内核模块 — `kvm_intel` 或 `kvm_amd`：

```
# modprobe kvm_intel
```

检查该模块是否已加载到内存中：

```
> lsmod | grep kvm
kvm_intel          64835  6
kvm                411041  1 kvm_intel
```

现在，KVM 主机即可为 KVM VM Guest 提供服务。有关详细信息，请访问 [第 37 章 “使用 qemu-system-ARCH 运行虚拟机”](#)。

35.3 特定于 KVM 主机的功能

您可以让基于 KVM 的 VM Guest 充分使用 VM 主机服务器硬件的特定功能（**半虚拟化**），以提高这些 Guest 的性能。本节将介绍可以通过哪些方法来使 Guest 直接访问物理主机的硬件（无需模拟层），以充分利用这些硬件。



提示

本节中的示例假设读者基本了解 `qemu-system-ARCH` 命令行选项。有关详细信息，请参见 [第 37 章 “使用 qemu-system-ARCH 运行虚拟机”](#)。

35.3.1 使用具有 virtio-scsi 的主机存储设备

`virtio-scsi` 是 KVM 的高级存储堆栈。它取代了以前用于 SCSI 设备直通的 `virtio-blk` 堆栈。与 `virtio-blk` 相比，它具有多项优势：

提高了可缩放性

KKVM Guest 的 PCI 控制器数量有限，导致挂接的设备数量也受到限制。`virtio-scsi` 解决了这个限制，因为它可以将多个存储设备组合到单个控制器上。`virtio-scsi` 控制器上的每个设备以逻辑单元 (LUN) 表示。

标准命令集

`virtio-blk` 会使用 `virtio-blk` 驱动程序和虚拟机监控器均需识别的一小部分命令，因此引入新命令需要同时更新该驱动程序和监控器。

相比之下，virtio-scsi 并不定义命令，而是遵循行业标准 SCSI 规范为这些命令定义一个传输协议。此方法与光纤通道、ATAPI 和 USB 设备等其他技术共享。

设备命名

virtio-blk 设备在 Guest 中显示为 /dev/vdX，这与物理系统中的设备名称不同，可能导致迁移时出现问题。

virtio-scsi 会确保设备名称与物理系统上的名称相同，这样便可轻松重新定位虚拟机。

SCSI 设备直通

对于由主机上整个 LUN 提供支持的虚拟磁盘，最好让 Guest 直接向 LUN 发送 SCSI 命令（直通）。此功能在 virtio-blk 中受到限制，因为 Guest 需使用 virtio-blk 协议而不是 SCSI 命令直通，此外，此功能并不适用于 Windows Guest。virtio-scsi 从根本上消除了这些限制。

35.3.1.1 virtio-scsi 的用法

KVM 支持对 virtio-scsi-pci 设备使用 SCSI 直通功能：

```
# qemu-system-x86_64 [...] \  
-device virtio-scsi-pci,id=scsi
```

35.3.2 使用 vhost-net 实现加速网络

vhost-net 模块用于加速 KVM 的半虚拟化网络驱动程序。它可提供更低的延迟和更高的网络吞吐量。通过以下示例命令行启动 Guest 即可使用 vhost-net 驱动程序：

```
# qemu-system-x86_64 [...] \  
-netdev tap,id=guest0,vhost=on,script=no \  
-net nic,model=virtio,netdev=guest0,macaddr=00:16:35:AF:94:4B
```

guest0 是 vhost 驱动的设备的标识字符串。

35.3.3 使用多队列 virtio-net 提升网络性能

QEMU 提供了使用**多队列**提升网络性能的方式，来应对 VM Guest 中虚拟 CPU 数量增加的情况。多队列 virtio-net 允许 VM Guest 的虚拟 CPU 并行传输包，因此可以提升网络性能。VM 主机服务器和 VM Guest 端都需要提供多队列支持。



提示：性能优势

多队列 virtio-net 解决方案在以下情况下最有利：

- 网络流量包较大。
- VM Guest 上存在许多同时保持活动状态的连接，这些连接主要是在 Guest 系统之间、Guest 与主机之间，或者 Guest 与外部系统之间建立的。
- 活动队列数量等于 VM Guest 中虚拟 CPU 的数量。



注意

尽管多队列 virtio-net 可以增加总网络吞吐量，但由于使用了虚拟 CPU 的计算资源，因此也会增加 CPU 消耗量。

过程 35.1：如何启用多队列 VIRTIO-NET

以下过程列出了使用 **qemu-system-ARCH** 启用多队列功能的重要步骤。假设在 VM 主机服务器上设置了一个具有多队列功能（自内核版本 3.8 开始支持此功能）的 tap 网络设备。

1. 在 **qemu-system-ARCH** 中，为该 tap 设备启用多队列：

```
-netdev tap,vhost=on,queues=2*N
```

其中 **N** 表示队列对的数量。

2. 在 **qemu-system-ARCH** 中，为 virtio-net-pci 设备启用多队列并指定 MSI-X（消息信号式中断）矢量：

```
-device virtio-net-pci,mq=on,vectors=2*N+2
```

其中 MSI-X 矢量数量的计算公式源于：N 个矢量用于 TX（传输）队列，N 个矢量用于 RX（接收）队列，一个矢量用于配置目的，一个矢量用于可能的 VQ（矢量化）控制。

3. 在 VM Guest 中的相关网络接口（在本示例中为 `eth0`）上启用多队列：

```
> sudo ethtool -L eth0 combined 2*N
```

最终的 `qemu-system-ARCH` 命令行类似于以下示例：

```
qemu-system-x86_64 [...] -netdev tap,id=guest0,queues=8,vhost=on \
-device virtio-net-pci,netdev=guest0,mq=on,vectors=10
```

对于命令行中的两个选项，需指定相同的网络设备 `id (guest0)`。

在运行中的 VM Guest 内部，以 `root` 特权指定以下命令：

```
> sudo ethtool -L eth0 combined 8
```

现在，Guest 系统网络将使用 `qemu-system-ARCH` 超级管理程序的多队列支持。

35.3.4 VFIO：对设备进行安全的直接访问

将 PCI 设备直接分配到 VM Guest（PCI 直通）可以避免在性能关键型路径中进行任何模拟，从而避免性能问题。VFIO 取代了传统的 KVM PCI 直通设备分配。此功能的先决条件是 VM 主机服务器配置符合[重要：VFIO 和 SR-IOV 的要求](#)中所述的要求。

要通过 VFIO 将 PCI 设备分配到 VM Guest，需要确定该设备属于哪个 IOMMU 组。IOMMU（用于将支持直接内存访问的 I/O 总线连接到主内存的输入/输出内存管理单元）API 支持组表示法。组是可与系统中的所有其他设备相互隔离的一组设备。因此，组是 VFIO 使用的所有权单元。

过程 35.2：通过 VFIO 将 PCI 设备分配到 VM GUEST

1. 标识要分配到 Guest 的主机 PCI 设备。

```
> sudo lspci -nn
[...]
```

```
00:10.0 Ethernet controller [0200]: Intel Corporation 82576 \
Virtual Function [8086:10ca] (rev 01)
[...]
```

记下设备 ID（在本例中为 00:10.0）和供应商 ID (8086:10ca)。

2. 确定此设备的 IOMMU 组：

```
> sudo readlink /sys/bus/pci/devices/0000\:00\:10.0/iommu_group
../../../../kernel/iommu_groups/20
```

此设备的 IOMMU 组为 20。现在，您可以检查该设备是否属于同一个 IOMMU 组：

```
> sudo ls -l /sys/bus/pci/devices/0000\:01\:10.0/iommu_group/devices/
[...] 0000:00:1e.0 -> ../../../../devices/pci0000:00/0000:00:1e.0
[...] 0000:01:10.0 -> ../../../../devices/
pci0000:00/0000:00:1e.0/0000:01:10.0
[...] 0000:01:10.1 -> ../../../../devices/
pci0000:00/0000:00:1e.0/0000:01:10.1
```

3. 从设备驱动程序取消绑定设备：

```
> sudo echo "0000:01:10.0" > /sys/bus/pci/devices/0000\:01\:10.0/driver/
unbind
```

4. 使用步骤 1 中记下的供应商 ID 将设备绑定到 vfio-pci 驱动程序：

```
> sudo echo "8086 153a" > /sys/bus/pci/drivers/vfio-pci/new_id
```

随即会创建一个新设备 /dev/vfio/IOMMU_GROUP，在本例中为 /dev/vfio/20。

5. 更改新建设备的所有权：

```
> sudo chown qemu.qemu /dev/vfio/DEVICE
```

6. 现在，运行为其分配了 PCI 设备的 VM Guest。

```
> sudo qemu-system-ARCH [...] -device
vfio-pci,host=00:10.0,id=ID
```

！ 重要：不支持热插拔

截至 SUSE Linux Enterprise Server 15 SP7，尚不支持通过 VFIO 传递给 VM Guest 的 PCI 设备的热插拔功能。

`/usr/src/linux/Documentation/vfio.txt` 文件中提供了有关 **VFIO** 驱动程序的更详细信息（需要安装软件包 `kernel-source`）。

35.3.5 VirtFS：在主机与 Guest 之间共享目录

VM Guest 通常在单独的计算空间中运行 — 它们各自都有自己的内存范围、专用的 CPU 和文件系统空间。能够共享 VM 主机服务器文件系统的某些部分，就能通过简化相互数据交换来提高虚拟化环境的灵活性。网络文件系统（例如 CIFS 和 NFS）是传统的共享目录方式，但由于它们不是专为虚拟化目的而设计，因此会有重大的性能和功能问题。

KVM 引入了一种经过优化的新方法，称为 **VirtFS**（有时称为“文件系统直通”）。VirtFS 使用半虚拟文件系统驱动程序，可以避免将 Guest 应用程序文件系统操作转换为块设备操作，然后将块设备操作转换为主机文件系统操作。

VirtFS 通常可用于以下情况：

- 要从多个 Guest 访问共享目录，或者提供 Guest 到 Guest 的文件系统访问。
- 在 Guest 引导过程中，要将虚拟磁盘替换为 Guest 的 RAM 磁盘所要连接到的根文件系统。
- 要从云环境中的单个主机文件系统为不同的客户提供储存服务。

35.3.5.1 实施

在 QEMU 中，可以通过定义两种类型的服务来简化 VirtFS 的实现：

- 用于在主机与 Guest 之间传输协议消息和数据的 `virtio-9p-pci` 设备。
- 用于定义导出文件系统属性（例如文件系统类型和安全模型）的 `fsdev` 设备。

例 35.1：使用 **VIRTFS** 导出主机的文件系统

```
> sudo qemu-system-x86_64 [...] \
```

```
-fsdev local,id=expl❶,path=/tmp/❷,security_model=mapped❸ \  
-device virtio-9p-pci,fsdev=expl❹,mount_tag=v_tmp❺
```

- ❶ 要导出的文件系统的标识。
- ❷ 要导出的主机上的文件系统路径。
- ❸ 要使用的安全模型 — mapped 可使 Guest 文件系统模式和权限与主机相互隔离，而 none 会调用“直通”安全模型，其中，对 Guest 文件进行的权限更改也会反映在主机上。
- ❹ 前面使用 `-fsdev id=` 定义的已导出文件系统 ID。
- ❺ 稍后要用在 Guest 上挂载所导出文件系统的挂载标记。

可按如下所示在 Guest 上挂载此类导出的文件系统：

```
> sudo mount -t 9p -o trans=virtio v_tmp /mnt
```

其中，`v_tmp` 是前面使用 `-device mount_tag=` 定义的挂载标记，`/mnt` 是要将导出的文件系统挂载到的挂载点。

35.3.6 KSM：在 Guest 之间共享内存页

内核同页合并 (KSM) 是 Linux 内核的一项功能，可将多个运行中进程的相同内存页合并到一个内存区域中。由于 KVM Guest 在 Linux 中以进程的形式运行，KSM 为超级管理程序提供了内存过量使用功能，以提高内存的使用效率。因此，如果您需要在内存有限的主机上运行多个虚拟机，KSM 可能有所帮助。

KSM 将其状态信息存储在 `/sys/kernel/mm/ksm` 目录下的文件中：

```
> ls -l /sys/kernel/mm/ksm  
full_scans  
merge_across_nodes  
pages_shared  
pages_sharing  
pages_to_scan  
pages_unshared  
pages_volatile  
run
```



```
sleep_millisecs
```

有关 `/sys/kernel/mm/ksm/*` 文件含义的详细信息，请参见 `/usr/src/linux/Documentation/vm/ksm.txt`（软件包 `kernel-source`）。

要使用 **KSM**，请执行以下操作。

1. 尽管 SLES 在内核中包含了 **KSM** 支持，但其默认处于禁用状态。要启用该支持，请运行以下命令：

```
# echo 1 > /sys/kernel/mm/ksm/run
```

2. 现在，请在 KVM 中运行多个 VM Guest，并检查 `pages_sharing` 和 `pages_shared` 文件的内容，例如：

```
> while [ 1 ]; do cat /sys/kernel/mm/ksm/pages_shared; sleep 1; done
13522
13523
13519
13518
13520
13520
13528
```

36 Guest 安装

virt-manager 和 **virt-install** 等基于 **libvirt** 的工具提供了方便的界面来设置和管理虚拟机。这些工具充当 **qemu-system-ARCH** 命令的某种封装程序。不过，您也可以直接使用 **qemu-system-ARCH**，而无需使用基于 **libvirt** 的工具。



警告：qemu-system-ARCH 和 libvirt

使用 **qemu-system-ARCH** 创建的 虚拟机 对于基于 **libvirt** 的工具不可见。

36.1 使用 qemu-system-ARCH 进行基本安装

在下面的示例中，将为 SUSE Linux Enterprise Server 11 安装创建一个虚拟机。有关命令的详细信息，请参见相关的手册页。

如果您尚未创建要在虚拟化环境中运行的系统的映像，需要从安装媒体创建一个映像。在这种情况下，您需要准备一个硬盘映像，并获取安装媒体的映像或该媒体本身。

使用 **qemu-img** 创建硬盘。

```
> qemu-img create ❶ -f raw ❷ /images/sles/hda ❸ 8G ❹
```

- ❶ **create** 子命令告知 **qemu-img** 创建新映像。
- ❷ 使用 **-f** 参数指定磁盘的格式。
- ❸ 映像文件的完整路径。
- ❹ 映像大小，在本例中为 8 GB。该映像创建为 稀疏映像文件，会随着数据填充到磁盘中而增长。指定的大小定义映像文件可增长到的最大大小。

至少创建了一个硬盘映像后，您可以使用 **qemu-system-ARCH** 设置一个将引导到安装系统的虚拟机：

```
# qemu-system-x86_64 -name "sles" ❶ -machine accel=kvm -M pc ❷ -m 768 ❸ \  
-smp 2 ❹ -boot d ❺ \  
-drive file=/images/sles/hda,if=virtio,index=0,media=disk,format=raw ❻ \  

```

```
-drive file=/isos/SLE-15-SP7-Online-ARCH-GM-medial.iso,index=1,media=cdrom ⑦ \  
-net nic,model=virtio,macaddr=52:54:00:05:11:11 ⑧ -net user \  
-vga cirrus ⑨ -balloon virtio ⑩
```

- ① 虚拟机的名称，将在窗口标题中显示，并用于 VNC 服务器。此名称必须是唯一的。
- ② 指定计算机类型。使用 **qemu-system-ARCH -M ?** 显示有效参数的列表。**pc** 是默认的标准 PC。
- ③ 虚拟机的最大内存量。
- ④ 定义包含两个处理器的 SMP 系统。
- ⑤ 指定引导顺序。有效值为 **a**、**b**（软盘 1 和 2）、**c**（第一个硬盘）、**d**（第一个 CD-ROM）或 **n** 到 **p**（从网络适配器 1-3 进行 Ether 引导）。默认值为 **c**。
- ⑥ 定义第一个 (**index=0**) 硬盘。系统会将该硬盘作为 **raw** 格式的半虚拟化 (**if=virtio**) 驱动器来访问。
- ⑦ 第二个 (**index=1**) 映像驱动器充当 CD-ROM。
- ⑧ 定义 MAC 地址为 **52:54:00:05:11:11** 的半虚拟化 (**model=virtio**) 网络适配器。请务必指定唯一的 MAC 地址，否则会发生网络冲突。
- ⑨ 指定显卡。如果指定 **none**，将禁用显卡。
- ⑩ 定义允许动态更改内存量（最大为使用参数 **-m** 指定的最大值）的半虚拟化气球设备。

安装完 Guest 操作系统后，您无需指定 CD-ROM 设备即可启动相关的虚拟机：

```
# qemu-system-x86_64 -name "sles" -machine type=pc,accel=kvm -m 768 \  
-smp 2 -boot c \  
-drive file=/images/sles/hda,if=virtio,index=0,media=disk,format=raw \  
-net nic,model=virtio,macaddr=52:54:00:05:11:11 \  
-vga cirrus -balloon virtio
```

36.2 使用 **qemu-img** 管理磁盘映像

在上一节中（请参见第 36.1 节“使用 **qemu-system-ARCH** 进行基本安装”），我们使用 **qemu-img** 命令创建了硬盘的映像。另一方面，您可以使用 **qemu-img** 来执行一般的磁盘映像操作。本节介绍可帮助您灵活管理磁盘映像的 **qemu-img** 子命令。

36.2.1 有关 qemu-img 调用的一般信息

qemu-img（像 **zypper** 那样）使用子命令来执行特定的任务。每个子命令会识别一组不同的选项。有些选项是通用的，其中的多数子命令都可使用，而有些选项则专用于相关的子命令。有关所有受支持选项的列表，请参见 **qemu-img** 手册页 (**man 1 qemu-img**)。 **qemu-img** 使用以下一般语法：

```
> qemu-img subcommand [options]
```

支持以下子命令：

create

在文件系统中创建新磁盘映像。

check

检查现有磁盘映像是否有错误。

compare

检查两个映像的内容是否相同。

map

转储映像文件名的元数据及其后备文件链。

amend

修正映像文件名的映像格式特定选项。

convert

将现有磁盘映像转换为其他格式的新映像。

info

显示相关磁盘映像的信息。

snapshot

管理现有磁盘映像的快照。

commit

应用对现有磁盘映像进行的更改。

rebase

基于现有映像创建新的基本映像。

resize

增大或减小现有映像的大小。

36.2.2 创建、转换和检查磁盘映像

本节介绍如何创建磁盘映像、检查其状态、转换磁盘映像的格式，以及获取有关特定磁盘映像的详细信息。

36.2.2.1 qemu-img create

使用 **qemu-img create** 可为 VM Guest 操作系统创建新磁盘映像。该命令使用以下语法：

```
> qemu-img create -f fmt ❶ -o options ❷ fname ❸ size ❹
```

- ❶ 目标映像的格式。支持的格式为 raw 和 qcow2。
- ❷ 某些映像格式支持在命令上传递其他选项。可在此处使用 -o 选项指定这些附加选项。raw 映像格式仅支持 size 选项，因此可以插入 -o size=8G，而不要在命令的末尾添加大小选项。
- ❸ 要创建的目标磁盘映像的路径。
- ❹ 目标磁盘映像的大小（如果尚未使用 -o size=<image_size> 选项指定）。映像大小的可选后缀为 K (KB)、M (MB)、G (GB) 或 T (TB)。

要在 /images 目录中创建最大可增长至 4 GB 的新磁盘映像 sles.raw，请运行以下命令：

```
> qemu-img create -f raw -o size=4G /images/sles.raw
Formatting '/images/sles.raw', fmt=raw size=4294967296

> ls -l /images/sles.raw
-rw-r--r-- 1 tux users 4294967296 Nov 15 15:56 /images/sles.raw

> qemu-img info /images/sles.raw
image: /images/sles11.raw
file format: raw
virtual size: 4.0G (4294967296 bytes)
```

```
disk size: 0
```

可以看到，新创建的映像的**虚拟**大小为 4 GB，但报告的实际磁盘大小为 0，因为尚未将任何数据写入该映像。



提示：Btrfs 文件系统上的 VM Guest 映像

如果您需要在 Btrfs 文件系统上创建磁盘映像，可以使用 `nocow=on` 来减少 Btrfs 的写入时复制功能产生的性能开销。

```
> qemu-img create -o nocow=on test.img 8G
```

但是，如果您想使用写入时复制（例如，要使用此功能来创建快照或者在虚拟机之间共享快照），请在命令行中省略 `nocow` 选项。

36.2.2.2 `qemu-img convert`

使用 **`qemu-img convert`** 可将磁盘映像转换为另一种格式。要获取 QEMU 支持的映像格式的完整列表，请运行 **`qemu-img -h`** 并查看输出的最后一行。该命令使用以下语法：

```
> qemu-img convert -c ❶ -f fmt ❷ -O out_fmt ❸ -o options ❹ fname ❺ out_fname ❻
```

- ❶ 对目标磁盘映像应用压缩。只有 `qcow` 和 `qcow2` 格式支持压缩。
- ❷ 源磁盘映像的格式。系统通常会自动检测格式，因此可以省略此参数。
- ❸ 目标磁盘映像的格式。
- ❹ 指定目标映像格式相关的其他选项。使用 `-o ?` 可查看目标映像格式支持的选项列表。
- ❺ 要转换的源磁盘映像的路径。
- ❻ 转换后的目标磁盘映像的路径。

```
> qemu-img convert -O vmdk /images/sles.raw \
/images/sles.vmdk

> ls -l /images/
-rw-r--r-- 1 tux users 4294967296 16. lis 10.50 sles.raw
-rw-r--r-- 1 tux users 2574450688 16. lis 14.18 sles.vmdk
```

要查看选定目标映像格式相关的选项列表，请运行以下命令（请将 `vmdk` 替换为您的映像格式）：

```
> qemu-img convert -O vmdk /images/sles.raw \
/images/sles.vmdk -o ?
Supported options:
size                Virtual disk size
backing_file        File name of a base image
compat6            VMDK version 6 image
subformat           VMDK flat extent format, can be one of {monolithicSparse \
                    (default) | monolithicFlat | twoGbMaxExtentSparse | twoGbMaxExtentFlat}
scsi                SCSI image
```

36.2.2.3 `qemu-img check`

使用 `qemu-img check` 可检查现有磁盘映像是否有错误。并非所有磁盘映像格式都支持此功能。该命令使用以下语法：

```
> qemu-img check -f fmt ❶ fname ❷
```

- ❶ 源磁盘映像的格式。系统通常会自动检测格式，因此可以省略此参数。
- ❷ 要检查的源磁盘映像的路径。

如果未发现错误，该命令不返回任何输出，否则会显示所发现的错误的类型和数量。

```
> qemu-img check -f qcow2 /images/sles.qcow2
ERROR: invalid cluster offset=0x2af0000
[...]
ERROR: invalid cluster offset=0x34ab0000
378 errors were found on the image.
```

36.2.2.4 增大现有磁盘映像的大小

创建新映像时，必须在创建映像之前指定其最大大小（请参见第 36.2.2.1 节 “`qemu-img create`”）。安装 VM Guest 并使用一段时间后，映像的初始大小可能不再够用。在这种情况下，可向映像增加空间。

要将现有磁盘映像的大小增大 2 GB，请使用：

```
> qemu-img resize /images/sles.raw +2GB
```



注意

您可以调整 `raw` 和 `qcow2` 格式的磁盘映像的大小。要调整其他格式的映像的大小，请先使用 `qemu-img convert` 将其转换为支持的格式。

现在，该映像的最后一个分区后面包含 2 GB 可用空间。您可以调整现有分区的大小，或添加新分区。

36.2.2.5 qcow2 文件格式的高级选项

qcow2 是 QEMU 使用的主要磁盘映像格式。其大小可按需增长，仅当虚拟机需要磁盘空间时才分配磁盘空间。

qcow2 格式的文件以恒定大小的单元进行组织。这些单元称为**簇**。从 Guest 的角度而言，虚拟磁盘也可划分为相同大小的簇。QEMU 默认为 64 kB 簇，但您可以在创建新映像时指定不同的值：

```
> qemu-img create -f qcow2 -o cluster_size=128K virt_disk.qcow2 4G
```

qcow2 映像包含一组表，这些表划分为两个级别，分别称为 L1 表和 L2 表。每个磁盘映像只有一个 L1 表，而根据映像的大小，L2 表可能有很多。

要在虚拟磁盘中读取或写入数据，QEMU 需要读取其对应的 L2 表，以确定相关数据位置。由于为每个 I/O 操作读取该表会消耗系统资源，QEMU 会在内存中缓存 L2 表，以提高磁盘访问速度。

36.2.2.5.1 选择适当的缓存大小

缓存大小与分配的空间量相关。L2 缓存可以映射以下虚拟磁盘空间量：

```
disk_size = l2_cache_size * cluster_size / 8
```


使用默认 64 kB 簇大小，即

```
disk_size = l2_cache_size * 8192
```

因此，要使缓存在使用默认簇大小的情况下映射 n GB 磁盘空间，需要

```
l2_cache_size = disk_size_GB * 131072
```

QEMU 默认使用 1 MB（1048576 字节）的 L2 缓存。根据上面的公式，1 MB 的 L2 缓存涵盖了 8 GB（1048576 / 131072）的虚拟磁盘空间。这意味着，如果您的虚拟磁盘大小不超过 8 GB，则使用默认 L2 缓存大小可使性能保持正常。如果磁盘更大，则可以通过增大 L2 缓存大小来提高磁盘访问速度。

36.2.2.5.2 配置缓存大小

可以在 QEMU 命令行上使用 `-drive` 选项来指定缓存大小。或者，可以在通过 QMP 通讯时使用 `blockdev-add` 命令。有关 QMP 的详细信息，请参见第 38.11 节“QMP - QEMU 计算机协议”。

以下选项配置虚拟 Guest 的缓存大小：

l2-cache-size

L2 表缓存的最大大小。

refcount-cache-size

refcount 块缓存的最大大小。有关 **refcount** 的详细信息，请参见 <https://raw.githubusercontent.com/qemu/qemu/master/docs/qcow2-cache.txt>。

cache-size

上述两个缓存的合计最大大小。

指定上述选项的值时，请注意以下几点：

- L2 缓存和 refcount 块缓存的大小需是簇大小的倍数。
- 如果您仅设置其中一个选项，QEMU 将自动调整其他选项，使 L2 缓存比 refcount 缓存大 4 倍。

refcount 缓存的使用频率比 L2 缓存要低得多，因此您可以将 refcount 缓存设置得小一些：

```
# qemu-system-ARCH [...] \  
-drive file=disk_image.qcow2,l2-cache-size=4194304,refcount-cache-size=262144
```

36.2.2.5.3 减少内存使用量

缓存越大，消耗的内存就越多。每个 qcow2 文件都有一个单独的 L2 缓存。使用大量较大的磁盘映像时，您可能需要相当大的内存量。如果您将后备文件（第 36.2.4 节“有效操作磁盘映像”）和快照（请参见第 36.2.3 节“使用 qemu-img 管理虚拟机的快照”）添加到 Guest 的设置链，则内存的消耗甚至更严重。

正因如此，QEMU 引入了 `cache-clean-interval` 设置。此设置定义一个以秒为单位的间隔，在此间隔过后，将从内存中去除未访问过的所有缓存项。

下面的示例每 10 分钟去除一次所有未使用的缓存项：

```
# qemu-system-ARCH [...] -drive file=hd.qcow2,cache-clean-interval=600
```

如果未设置此选项，则默认值为 0，这会禁用此功能。

36.2.3 使用 qemu-img 管理虚拟机的快照

虚拟机快照是运行 VM Guest 的整个环境的快照。该快照包含处理器 (CPU)、内存 (RAM)、设备和所有可写磁盘的状态。

当您需保存特定状态的虚拟机时，快照非常有用。例如，在虚拟化服务器上配置网络服务后，您可以从上次保存的虚拟机状态快速启动虚拟机。或者，您可以在关闭虚拟机之后创建快照，以便在尝试执行某种会导致 VM Guest 不稳定的试验性操作之前创建备份状态。本节介绍后一种做法，前一种做法会在第 38 章“使用 QEMU 监控器管理虚拟机”中介绍。

要使用快照，您的 VM Guest 必须至少包含一个 qcow2 格式的可写硬盘映像。此设备通常是第一个虚拟硬盘。

虚拟机快照是在交互式 QEMU 监控器中使用 `savevm` 命令创建的。为了更方便地识别特定的快照，可为其分配一个**标记**。有关 QEMU 监控器的详细信息，请参见第 38 章“使用 QEMU 监控器管理虚拟机”。

qcow2 磁盘映像包含保存的快照后，您可以使用 `qemu-img snapshot` 命令检查这些快照。



警告：关闭 VM Guest

请不要在虚拟机正在运行时使用 `qemu-img snapshot` 命令创建或删除虚拟机快照。否则，可能会损坏包含保存的虚拟机状态的磁盘映像。

36.2.3.1 列出现有快照

使用 `qemu-img snapshot -l DISK_IMAGE` 可查看 `disk_image` 映像中保存的所有现有快照的列表。即使 VM Guest 正在运行，您也可以获取该列表。

```
> qemu-img snapshot -l /images/sles.qcow2
Snapshot list:
```

ID ❶	TAG ❷	VM SIZE ❸	DATE ❹	VM CLOCK ❺
1	booting	4.4M	2013-11-22 10:51:10	00:00:20.476
2	booted	184M	2013-11-22 10:53:03	00:02:05.394
3	logged_in	273M	2013-11-22 11:00:25	00:04:34.843
4	ff_and_term_running	372M	2013-11-22 11:12:27	00:08:44.965

- ❶ 快照的自动递增唯一标识号。
- ❷ 快照的唯一说明字符串。它以直观易懂的 ID 形式来表示。
- ❸ 快照占用的磁盘空间。运行中应用程序消耗的内存越多，快照就越大。
- ❹ 快照的创建时间和日期。
- ❺ 虚拟机时钟的当前状态。

36.2.3.2 创建已关闭虚拟机的快照

使用 `qemu-img snapshot -c SNAPSHOT_TITLE DISK_IMAGE` 可创建事先已关闭的虚拟机的当前状态快照。

```
> qemu-img snapshot -c backup_snapshot /images/sles.qcow2
```

```
> qemu-img snapshot -l /images/sles.qcow2
Snapshot list:
```

ID	TAG	VM SIZE	DATE	VM CLOCK
----	-----	---------	------	----------

1	booting	4.4M	2013-11-22	10:51:10	00:00:20.476
2	booted	184M	2013-11-22	10:53:03	00:02:05.394
3	logged_in	273M	2013-11-22	11:00:25	00:04:34.843
4	ff_and_term_running	372M	2013-11-22	11:12:27	00:08:44.965
5	backup_snapshot	0	2013-11-22	14:14:00	00:00:00.000

如果某种情况干扰了 VM Guest 的运行，而您需要恢复到所保存快照（在本示例中为 ID 5）的状态，请关闭 VM Guest 并执行以下命令：

```
> qemu-img snapshot -a 5 /images/sles.qcow2
```

下一次您使用 **qemu-system-ARCH** 运行虚拟机时，它将处于编号为 5 的快照的状态。



注意

qemu-img snapshot -c 命令与 QEMU 监视器的 **savevm** 命令（请参见第 38 章“使用 QEMU 监控器管理虚拟机”）无关。例如，对于使用 **savevm** 在 QEMU 监控器中创建的快照，无法使用 **qemu-img snapshot -a** 应用快照。

36.2.3.3 删除快照

使用 **qemu-img snapshot -d SNAPSHOT_ID DISK_IMAGE** 可删除虚拟机的旧快照或不需要的快照。这可以节省 **qcow2** 磁盘映像中的磁盘空间，因为快照数据占用的空间将会恢复：

```
> qemu-img snapshot -d 2 /images/sles.qcow2
```

36.2.4 有效操作磁盘映像

假设在实际应用中，您是一名服务器管理员，负责运行和管理多个虚拟化操作系统。其中一组系统基于一个特定的发行套件，而另一组（或多个组）基于不同版本的发行套件，甚至不同的平台（也许不是 Unix）。更复杂的是，基于同一发行套件的虚拟 Guest 系统会因部门和部署而各不相同。文件服务器使用的设置和服务通常与 Web 服务器不同，不过，两者可能仍然基于 SUSE® Linux Enterprise Server。

使用 QEMU 可以创建“基本”磁盘映像。您可以将这些映像作为模板虚拟机使用。这些基本映像将为您节省大量时间，因为您无需多次安装同一个操作系统。

36.2.4.1 基本映像和派生映像

首先，照常构建一个磁盘映像，并在其中安装目标系统。有关详细信息，请参见第 36.1 节“使用 **qemu-system-ARCH** 进行基本安装”和第 36.2.2 节“创建、转换和检查磁盘映像”。然后使用第一个映像作为基本映像来构建新映像。基本映像也称为**后备文件**。构建新的**派生**映像后，切勿再次引导基本映像，而是引导派生映像。多个派生映像可以同时依赖于一个基本映像。因此，更改基本映像可能会损坏依赖性。使用派生映像时，QEMU 会将更改写入其中，并仅使用基本映像进行读取操作。

比较好的做法是基于一个全新安装（并已根据需要注册）的操作系统创建基本映像，该操作系统中应当尚未应用任何补丁且未安装或删除其他应用程序。以后，您可以在应用最新补丁后基于原始基本映像创建另一个基本映像。

36.2.4.2 创建派生映像



注意

尽管您可以对基本映像使用 `raw` 格式，但不能对派生映像使用该格式，因为 `raw` 格式不支持 `backing_file` 选项。可对派生映像使用 `qcow2` 等格式。

例如，`/images/sles_base.raw` 是包含全新安装的系统的基本映像。

```
> qemu-img info /images/sles_base.raw
image: /images/sles_base.raw
file format: raw
virtual size: 4.0G (4294967296 bytes)
disk size: 2.4G
```

该映像的预留大小为 4 GB，实际大小为 2.4 GB，格式为 `raw`。使用以下命令创建自 `/images/sles_base.raw` 基本映像派生的映像：

```
> qemu-img create -f qcow2 /images/sles_derived.qcow2 \
-o backing_file=/images/sles_base.raw
Formatting '/images/sles_derived.qcow2', fmt=qcow2 size=4294967296 \
backing_file='/images/sles_base.raw' encryption=off cluster_size=0
```

查看派生映像的细节：

```
> qemu-img info /images/sles_derived.qcow2
image: /images/sles_derived.qcow2
file format: qcow2
virtual size: 4.0G (4294967296 bytes)
disk size: 140K
cluster_size: 65536
backing file: /images/sles_base.raw \
(actual path: /images/sles_base.raw)
```

尽管派生映像的预留大小与基本映像的大小相同 (4 GB)，但实际大小仅为 140 KB。原因是只有对派生映像内部的系统进行的更改会保存下来。运行派生的虚拟机，根据需要注册，并应用最新补丁。在系统中进行任何其他更改，例如，去除不需要的软件包或安装新软件包。然后关闭 VM Guest 并再次检查其细节：

```
> qemu-img info /images/sles_derived.qcow2
image: /images/sles_derived.qcow2
file format: qcow2
virtual size: 4.0G (4294967296 bytes)
disk size: 1.1G
cluster_size: 65536
backing file: /images/sles_base.raw \
(actual path: /images/sles_base.raw)
```

disk size 值已增长为 1.1 GB，这是文件系统（而不是基本映像）中的更改所占用的磁盘空间。

36.2.4.3 从派生映像重建基本映像

在修改派生映像（应用补丁、安装特定的应用程序、更改环境设置，等等）之后，它会达到所需的状态。此时，您可以合并原始基本映像和派生映像以创建新的基本映像。

原始基本映像 (/images/sles_base.raw) 包含全新安装的系统。它可以是经过修改的新基本映像的模板，而新的基本映像可以包含与第一个基本映像相同的系统，加上所有安全补丁和更新补丁等内容。在创建此新基本映像后，还可将它用作更专用的派生映像的模板。新基本映像便会独立于原始基本映像。基于派生映像创建基本映像的过程称为**重建基本映像**：

```
> qemu-img convert /images/sles_derived.qcow2 \
-O raw /images/sles_base2.raw
```

此命令创建了使用 `raw` 格式的新基本映像 `/images/sles_base2.raw`。

```
> qemu-img info /images/sles_base2.raw
image: /images/sles11_base2.raw
file format: raw
virtual size: 4.0G (4294967296 bytes)
disk size: 2.8G
```

新映像比原始基本映像大 0.4 GB。它不使用任何后备文件，您可以轻松基于此映像创建新的派生映像。这样，您便可以为组织中的虚拟磁盘映像创建复杂的层次结构，并节省大量的时间和工作。

36.2.4.4 在 VM 主机服务器上挂载映像

在主机系统下挂载虚拟磁盘映像的做法可能会很实用。强烈建议阅读第 21 章 “`libguestfs`”，并使用专用的工具来访问虚拟机映像。不过，如果您需要手动执行此操作，请按照本指南所述操作。

Linux 系统可以使用回写设备挂载 `raw` 磁盘映像的内部分区。第一个示例过程更复杂，但阐释得更清楚，而第二个过程则更简单直接：

过程 36.1：通过计算分区偏移来挂载磁盘映像

1. 在您要挂载其分区的磁盘映像中设置一个**循环设备**。

```
> losetup /dev/loop0 /images/sles_base.raw
```

2. 确定您要挂载的分区的**扇区大小**和**起始扇区编号**。

```
> fdisk -lu /dev/loop0
```

```
Disk /dev/loop0: 4294 MB, 4294967296 bytes
255 heads, 63 sectors/track, 522 cylinders, total 8388608 sectors
Units = sectors of 1 * 512 = 512 ① bytes
Disk identifier: 0x000ceca8
```

Device	Boot	Start	End	Blocks	Id	System
/dev/loop0p1		63	1542239	771088+	82	Linux swap
/dev/loop0p2	*	1542240 ❷	8385929	3421845	83	Linux

❶ 磁盘扇区大小。

❷ 分区的起始扇区。

3. 计算分区起始偏移：

$\text{sector_size} * \text{sector_start} = 512 * 1542240 = 789626880$

4. 删除循环，并在准备好的目录中，根据计算出的偏移挂载磁盘映像内的分区。

```
> losetup -d /dev/loop0
> mount -o loop,offset=789626880 \
/images/sles_base.raw /mnt/sles/
> ls -l /mnt/sles/
total 112
drwxr-xr-x  2 root root  4096 Nov 16 10:02 bin
drwxr-xr-x  3 root root  4096 Nov 16 10:27 boot
drwxr-xr-x  5 root root  4096 Nov 16 09:11 dev
[...]
drwxrwxrwt 14 root root  4096 Nov 24 09:50 tmp
drwxr-xr-x 12 root root  4096 Nov 16 09:16 usr
drwxr-xr-x 15 root root  4096 Nov 16 09:22 var
```

5. 将一个或多个文件复制到挂载的分区，并在完成后卸载该分区。

```
> cp /etc/X11/xorg.conf /mnt/sles/root/tmp
> ls -l /mnt/sles/root/tmp
> umount /mnt/sles/
```



警告：不要向当前正在使用的映像写入数据

切勿挂载处于 `read-write` 模式的运行中虚拟机映像的分区。这可能会损坏该分区并破坏整个 VM Guest。

37 使用 qemu-system-ARCH 运行虚拟机

准备好虚拟磁盘映像后（有关磁盘映像的详细信息，请参见第 36.2 节“使用 `qemu-img` 管理磁盘映像”），便可以启动相关的虚拟机了。第 36.1 节“使用 `qemu-system-ARCH` 进行基本安装”介绍了用于安装和运行 VM Guest 的简单命令。本章重点详细解释 `qemu-system-ARCH` 的用法，并针对更具体的任务提供解决方案。有关 `qemu-system-ARCH` 选项的完整列表，请参见其手册页 (`man 1 qemu`)。

37.1 基本 `qemu-system-ARCH` 调用

`qemu-system-ARCH` 命令使用以下语法：

```
qemu-system-ARCH OPTIONS ❶ -drive file=DISK_IMAGE ❷
```

- ❶ `qemu-system-ARCH` 接受许多选项。其中的大部分选项定义模拟硬件的参数，其他选项会影响更一般性的模拟器行为。如果您不提供任何选项，则会使用默认值，在此情况下，您需要提供所要运行的磁盘映像的路径。
- ❷ 包含要虚拟化的 Guest 系统的磁盘映像路径。`qemu-system-ARCH` 支持许多映像格式。使用 `qemu-img --help` 可列出这些格式。

❗ 重要：AArch64 体系结构

只有 64 位 ARM 体系结构 (AArch64) 上提供 KVM 支持。要在 AArch64 体系结构上运行 QEMU，您需要：

- 使用 `-machine virt-VERSION_NUMBER` 选项指定专用于 QEMU Arm® 虚拟机的计算机类型。
- 使用 `-bios` 选项指定固件映像文件。
也可以使用 `-drive` 选项指定固件映像文件，例如：

```
-drive file=/usr/share/edk2/aarch64/QEMU_EFI-pflash.raw,if=pflash,format=raw
```

```
-drive file=/var/lib/libvirt/qemu/nvram/  
opensuse_VARS.fd,if=pflash,format=raw
```

- 使用 `-cpu host` 选项指定 VM 主机服务器的 CPU（默认值为 `cortex-15`）。
- 使用 `-machine gic-version=host` 选项指定与主机相同的通用中断控制器 (GIC) 版本（默认值为 `2`）。
- 如需使用图形模式，请指定 `virtio-gpu-pci` 类型的图形设备。

例如：

```
> sudo qemu-system-aarch64 [...] \  
-bios /usr/share/qemu/qemu-uefi-aarch64.bin \  
-cpu host \  
-device virtio-gpu-pci \  
-machine virt,accel=kvm,gic-version=host
```

37.2 一般 `qemu-system-ARCH` 选项

本节介绍一般 `qemu-system-ARCH` 选项，以及与基本模拟硬件（例如虚拟机的处理器、内存、型号类型或时间处理方法）相关的选项。

`-name NAME_OF_GUEST`

指定运行中 Guest 系统的名称。该名称将显示在窗口标题中，用于 VNC 服务器。

`-boot OPTIONS`

指定定义的驱动器的引导顺序。驱动器以字母（盘符）表示，`a` 和 `b` 代表软盘驱动器 1 和 2，`c` 代表第一个硬盘，`d` 代表第一个 CD-ROM 驱动器，`n` 到 `p` 代表 Ether 引导网络适配器。

例如，`qemu-system-ARCH [...] -boot order=ndc` 首先尝试从网络引导，然后尝试从第一个 CD-ROM 驱动器引导，最后尝试从第一个硬盘引导。

-pidfile FILENAME

将 QEMU 的进程标识号 (PID) 存储在文件中。如果您从脚本运行 QEMU，此文件非常有用。

-nodefaults

默认情况下，即使您不在命令行上指定基本虚拟设备，QEMU 也会创建这些设备。此选项会关闭此功能，在此情况下，您必须手动指定每个设备，包括显卡和网卡、并行或串行端口，或虚拟控制台。默认连 QEMU 监控器都不会挂接。

-daemonize

启动 QEMU 进程后将其“守护程序化”。在 QEMU 准备好接收其任何设备上的连接后，会从标准输入和标准输出分离。

注意：SeaBIOS BIOS 实现

默认使用的 BIOS 是 SeaBIOS。您可以引导 USB 设备和任何驱动器（CD-ROM、软盘或硬盘）。SeaBIOS 支持 USB 鼠标和键盘，并支持多个 VGA 显卡。有关 SeaBIOS 的详细信息，请访问 [SeaBIOS 网站 \(https://www.seabios.org/SeaBIOS\)](https://www.seabios.org/SeaBIOS) .

37.2.1 基本虚拟硬件

37.2.1.1 计算机类型

您可以指定模拟计算机的类型。运行 `qemu-system-ARCH -M help` 可查看支持的计算机类型列表。

注意：ISA-PC

不支持计算机类型 `isapc:ISA-only-PC`。

37.2.1.2 CPU 型号

要指定处理器 (CPU) 型号的类型，请运行 `qemu-system-ARCH -cpu MODEL`。使用 `qemu-system-ARCH -cpu help` 可查看支持的 CPU 型号列表。

37.2.1.3 其他基本选项

下面是从命令行启动 **qemu** 时最常用的选项列表。要查看所有可用选项，请参见 **qemu-doc** 手册页。

`-m MEGABYTES`

指定用作虚拟 RAM 大小的 MB 数。

`-balloon virtio`

指定用于动态更改分配给 VM Guest 的虚拟 RAM 量的半虚拟化设备。上限是使用 `-m` 指定的内存量。

`-smp NUMBER_OF_CPUS`

指定要模拟的 CPU 数量。QEMU 在 PC 平台上最多支持 255 个 CPU（其中最多有 64 个 CPU 可使用 KVM 加速）。此选项还接受其他 CPU 相关的参数，例如插槽数、每个插槽的**核心数**，或每个核心的**线程数**。

下面是有效的 `qemu-system-ARCH` 命令行示例：

```
> sudo qemu-system-x86_64 \  
-name "SLES 15 SP7" \  
-M pc-i440fx-2.7 -m 512 \  
-machine accel=kvm -cpu kvm64 -smp 2 \  
-drive format=raw,file=/images/sles.raw
```

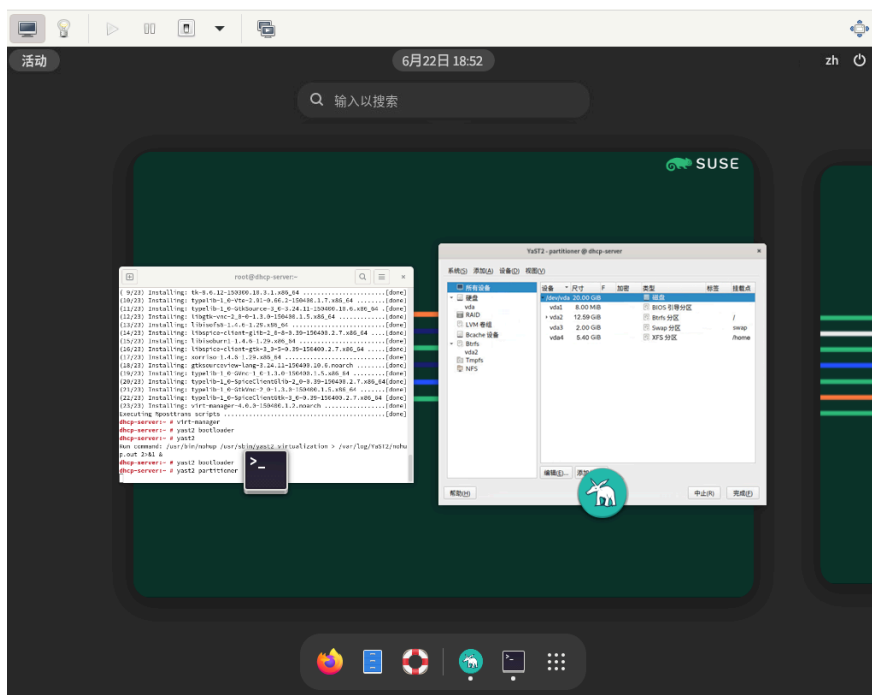


图 37.1：显示使用 SLES 作为 VM GUEST 的 QEMU 窗口

-no-acpi

禁用 [ACPI](#) 支持。

-S

QEMU 在 CPU 停止的状态下启动。要启动 CPU，请在 QEMU 监控器中输入 c。有关详细信息，请参见第 38 章 “使用 QEMU 监控器管理虚拟机”。

37.2.2 存储和读取虚拟设备的配置

-readconfig CFG_FILE

您无需每次想要运行 VM Guest 时都在命令行上输入设备配置选项，**qemu-system-ARCH** 可以从先前使用 -writeconfig 保存的或者手动编辑的文件中读取相应配置。

-writeconfig CFG_FILE

将当前虚拟机的设备配置转储到文本文件，通过 -readconfig 选项可以重复使用该文件。

```
> sudo qemu-system-x86_64 -name "SLES 15 SP7" \
```

```

-machine accel=kvm -M pc-i440fx-2.7 -m 512 -cpu kvm64 \
-smp 2 /images/sles.raw -writeconfig /images/sles.cfg
(exited)
> cat /images/sles.cfg
# qemu config file

[drive]
  index = "0"
  media = "disk"
  file = "/images/sles_base.raw"

```

这样，您便可以有条不紊地有效管理虚拟机的设备配置。

37.2.3 Guest 实时时钟

-rtc OPTIONS

指定在 VM Guest 中处理 RTC 的方式。Guest 的时钟默认自主机系统的时钟派生。因此，建议将主机系统时钟与精确的外部时钟同步（例如，通过 NTP 服务同步）。

如果您需要将 VM Guest 时钟与主机时钟隔离，请指定 `clock=vm`，而不要使用默认值 `clock=host`。

您也可以使用 `base` 选项来指定 VM Guest 时钟的初始时间：

```
> sudo qemu-system-x86_64 [...] -rtc clock=vm,base=2010-12-03T01:02:00
```

可以不指定时戳，而是指定 `utc` 或 `localtime`。前者指示 VM Guest 按当前 UTC（协调世界时，请参见 <https://en.wikipedia.org/wiki/UTC>）值启动，而后者则应用本地时间设置。

37.3 在 QEMU 中使用设备

QEMU 虚拟机会模拟运行 VM Guest 所需的所有设备。例如，QEMU 支持多种类型的网卡、块设备（硬盘和可移动驱动器）、USB 设备、字符设备（串行和并行端口）或多媒体设备（显卡和声卡）。本节介绍用于配置多种类型的受支持设备的选项。



提示

如果需要为设备（例如 `-drive`）设置特殊的驱动程序和驱动程序属性，请使用 `-device` 选项来指定，并使用 `drive=` 子选项进行标识。例如：

```
> sudo qemu-system-x86_64 [...] -drive if=none,id=drive0,format=raw \
-device virtio-blk-pci,drive=drive0,scsi=off ...
```

要获取有关可用驱动程序及其属性的帮助，请使用 `-device ?` 和 `-device DRIVER, ?`。

37.3.1 块设备

块设备对于虚拟机而言至关重要。这些设备是称作**驱动器**的固定或可移动存储媒体。通常会用连接的硬盘中的其中一块保存要虚拟化的 Guest 操作系统。

虚拟机驱动器使用 `-drive` 来定义。此选项具有许多子选项，本节将介绍其中的一些子选项。有关完整列表，请参见手册页 ([man 1 qemu](#))。

`-drive` 选项的子选项

`file=image_fname`

指定要用于此驱动器的磁盘映像的路径。如果未指定，将使用一个空（可移动）驱动器。

`if=drive_interface`

指定驱动器要连接到的接口类型。SUSE 目前仅支持 `floppy`、`scsi`、`ide` 或 `virtio`。`virtio` 定义半虚拟化磁盘驱动程序。默认值为 `ide`。

`index=index_of_connector`

指定驱动器所连接到的磁盘接口（请参见 `if` 选项）上某个连接器的索引号。如果未指定，则索引会自动递增。

`media=type`

指定媒体的类型。可以是 `disk`（表示硬盘）或 `cdrom`（表示可移动的 CD-ROM 驱动器）。

`format=img_fmt`

指定连接的磁盘映像的格式。如果未指定，系统会自动检测格式。SUSE 目前支持 `raw` 和 `qcow2` 格式。

`cache=method`

指定驱动器的缓存方法。可能的值为

`unsafe`、`writethrough`、`writeback`、`directsync` 或 `none`。要在使用 `qcow2` 映像格式时提高性能，请选择 `writeback`。`none` 会禁用主机页缓存，因此是最安全的选项。对于映像文件，默认值为 `writeback`。有关详细信息，请参见第 19 章“磁盘缓存模式”。



提示

为了简化块设备的定义，QEMU 能够识别多种简写形式，以方便您输入 `qemu-system-ARCH` 命令行。

可使用

```
> sudo qemu-system-x86_64 -cdrom /images/cdrom.iso
```

来代替

```
> sudo qemu-system-x86_64 -drive format=raw,file=/images/cdrom.iso,index=2,media=cdrom
```

和

```
> sudo qemu-system-x86_64 -hda /images/image1.raw -hdb /images/image2.raw  
-hdc \  
/images/image3.raw -hdd /images/image4.raw
```

来代替

```
> sudo qemu-system-x86_64 -drive format=raw,file=/images/image1.raw,index=0,media=disk \  
-drive format=raw,file=/images/image2.raw,index=1,media=disk \  
-drive format=raw,file=/images/image3.raw,index=2,media=disk \  
-drive format=raw,file=/images/image4.raw,index=3,media=disk
```



```
-drive format=raw,file=/images/image4.raw,index=3,media=disk
```



提示：使用主机驱动器代替映像

作为使用磁盘映像（请参见第 36.2 节“使用 **qemu-img** 管理磁盘映像”）的替代方式，您还可以使用现有的 VM 主机服务器磁盘，将其作为驱动器进行连接，然后从 VM Guest 访问它们。请直接使用主机磁盘设备，而不要使用磁盘映像文件名。

要访问主机 CD-ROM 驱动器，请使用

```
> sudo qemu-system-x86_64 [...] -drive file=/dev/cdrom,media=cdrom
```

要访问主机硬盘，请使用

```
> sudo qemu-system-x86_64 [...] -drive file=/dev/hdb,media=disk
```

VM Guest 使用的主机驱动器不可同时由 VM 主机服务器或另一个 VM Guest 访问。

37.3.1.1 释放未使用的 Guest 磁盘空间

稀疏映像文件这种磁盘映像文件的大小会随着用户在其中添加数据而增长，它所占用的磁盘空间量等于其中存储的数据量。例如，如果您在稀疏磁盘映像中复制 1 GB 数据，则此映像的大小会增长 1 GB。如果您随后删除 500 MB（举例而言）的数据，映像大小默认不会按预期减小。

正因如此，KVM 命令行上引入了 **discard=on** 选项。此选项告知超级管理程序在从稀疏 Guest 映像中删除数据后自动释放“空洞”。此选项仅对 **if=scsi** 驱动器接口有效：

```
> sudo qemu-system-x86_64 [...] -drive format=img_format,file=/path/to/  
file.img,if=scsi,discard=on
```



重要：支持状态

不支持 **if=scsi**。此接口不会映射到 **virtio-scsi**，而是映射到 **lsi SCSI 适配器**。

37.3.1.2 IOThread

IOThread 是 virtio 设备的专用事件循环线程，用于执行 I/O 请求来提高可缩放性，尤其是在包含 SMP VM Guest 并使用许多磁盘设备的 SMP VM 主机服务器上。进行 I/O 处理时，IOThread 不会使用 QEMU 的主事件循环，而是允许将 I/O 工作分散到多个 CPU 之间，因而可改善延迟情况（如果配置正确）。

可通过定义 IOThread 对象来启用 IOThread。然后，virtio 设备可将这些对象用于其 I/O 事件循环。许多 virtio 设备都可以使用单个 IOThread 对象，或者可按 1:1 映射配置 virtio 设备和 IOThread 对象。以下示例创建 ID 为 `iothread0` 的单个 IOThread，然后，该 IOThread 将用作两个 virtio-blk 设备的事件循环。

```
> sudo qemu-system-x86_64 [...] -object iothread,id=iothread0\  
-drive if=none,id=drive0,cache=none,aio=native,\  
format=raw,file=filename -device virtio-blk-pci,drive=drive0,scsi=off,\  
iothread=iothread0 -drive if=none,id=drive1,cache=none,aio=native,\  
format=raw,file=filename -device virtio-blk-pci,drive=drive1,scsi=off,\  
iothread=iothread0 [...]
```

下面的 qemu 命令行示例说明了 virtio 设备与 IOThread 之间的 1:1 映射：

```
> sudo qemu-system-x86_64 [...] -object iothread,id=iothread0\  
-object iothread,id=iothread1 -drive if=none,id=drive0,cache=none,aio=native,\  
format=raw,file=filename -device virtio-blk-pci,drive=drive0,scsi=off,\  
iothread=iothread0 -drive if=none,id=drive1,cache=none,aio=native,\  
format=raw,file=filename -device virtio-blk-pci,drive=drive1,scsi=off,\  
iothread=iothread1 [...]
```

37.3.1.3 virtio-blk 的基于 Bio 的 I/O 路径

为了优化 I/O 密集型应用程序的性能，内核 3.7 版本为 virtio-blk 接口引入了一条新的 I/O 路径。这种基于 bio 的块设备驱动程序跳过了 I/O 调度器，从而缩短了 Guest 内的 I/O 路径并降低了延迟。对于 SSD 磁盘等高速存储设备，该驱动程序特别有用。

该驱动程序默认处于禁用状态。要使用该驱动程序，请执行以下操作：

1. 在 Guest 上的内核命令行中追加 `virtio_blk.use_bio=1`。可以通过 YaST > 系统 > 引导加载程序执行此操作。

为此，您也可以编辑 `/etc/default/grub`，搜索包含 `GRUB_CMDLINE_LINUX_DEFAULT=` 的行，并在末尾添加内核参数。然后运行 `grub2-mkconfig >/boot/grub2/grub.cfg` 以更新 grub2 引导菜单。

2. 在激活新内核命令行的情况下重引导 Guest。



提示：慢速设备上基于 Bio 的驱动程序

基于 bio 的 virtio-blk 驱动程序对于机械硬盘等慢速设备没有帮助。原因在于，调度所带来的优势大于缩短 bio 路径所带来的优势。请不要在慢速设备上使用基于 bio 的驱动程序。

37.3.1.4 直接访问 iSCSI 资源

QEMU 现已与 `libiscsi` 相集成。因此，QEMU 可以直接访问 iSCSI 资源并将其用作虚拟机块设备。此功能不需要任何主机 iSCSI 发起端配置，而基于 iSCSI 目标的 libvirt 存储池设置则需要这种配置。此功能通过用户空间库 `libiscsi` 直接将 Guest 存储接口连接到 iSCSI 目标 LUN。您也可以在 libvirt XML 配置中指定基于 iSCSI 的磁盘设备。



注意：RAW 映像格式

由于 iSCSI 协议存在某些技术方面的限制，仅当使用 RAW 映像格式时，此功能才可用。

下面是用于配置 iSCSI 连接的 QEMU 命令行界面。



注意：virt-manager 限制

virt-manager 界面尚未公开基于 `libiscsi` 的存储空间置备的用法，但是可以通过直接编辑 Guest XML 对其进行配置。这种访问基于 iSCSI 的存储空间的新方式通过命令行来实现。

```
> sudo qemu-system-x86_64 -machine accel=kvm \
  -drive file=iscsi://192.168.100.1:3260/iqn.2016-08.com.example:314605ab-
  a88e-49af-b4eb-664808a3443b/0,\
  format=raw,if=none,id=mydrive,cache=none \
```

```
-device ide-hd,bus=ide.0,unit=0,drive=mydrive ...
```

下面是使用基于协议的 iSCSI 的 Guest 域 XML 的示例代码段：

```
<devices>
...
<disk type='network' device='disk'>
  <driver name='qemu' type='raw' />
  <source protocol='iscsi' name='iqn.2013-07.com.example:iscsi-nopool/2'>
    <host name='example.com' port='3260' />
  </source>
  <auth username='myuser'>
    <secret type='iscsi' usage='libvirtiscsi' />
  </auth>
  <target dev='vda' bus='virtio' />
</disk>
</devices>
```

将此代码段与使用 virt-manager 设置的基于主机的 iSCSI 发起端示例相对比：

```
<devices>
...
<disk type='block' device='disk'>
  <driver name='qemu' type='raw' cache='none' io='native' />
  <source dev='/dev/disk/by-path/scsi-0:0:0:0' />
  <target dev='hda' bus='ide' />
  <address type='drive' controller='0' bus='0' target='0' unit='0' />
</disk>
<controller type='ide' index='0'>
  <address type='pci' domain='0x0000' bus='0x00' slot='0x01'
    function='0x1' />
</controller>
</devices>
```

37.3.1.5 通过 QEMU 使用 RADOS 块设备

RADOS 块设备 (RBD) 将数据存储在 Ceph 群集中。这些设备支持快照、复制和数据一致性。您可以像使用其他块设备一样，从 KVM 管理的 VM Guest 使用 RBD。

有关更多细节，请参见《SUSE Enterprise Storage Administration Guide》中的“Ceph as a Back-end for QEMU KVM Instance”一章 (<https://documentation.suse.com/ses/html/ses-all/cha-ceph-kvm.html>) 。

37.3.2 图形设备和显示选项

本节介绍影响模拟视频卡类型的 QEMU 选项，以及 VM Guest 图形输出的显示方式。

37.3.2.1 定义视频卡

QEMU 使用 `-vga` 来定义用于显示 VM Guest 图形输出的视频卡。`-vga` 选项识别以下值：

`none`

在 VM Guest 上禁用视频卡（不模拟视频卡）。您仍可以通过串行控制台访问运行中的 VM Guest。

`std`

模拟标准的 VESA 2.0 VBE 视频卡。如果您打算在 VM Guest 上使用较高的显示分辨率，请使用此值。

`qxl`

QXL 是半虚拟显卡。它与 VGA 兼容（包括 VESA 2.0 VBE 支持）。使用 `spice` 视频协议时，建议使用 `qxl`。

`virtio`

半虚拟 VGA 显卡。

37.3.2.2 显示选项

以下选项会影响 VM Guest 图形输出的显示方式。

`-display gtk`

在 GTK 窗口中显示视频输出。此界面提供用于在运行时配置和控制 VM 的 UI 元素。

`-display sdl`

通过 SDL 在单独的图形窗口中显示视频输出。有关详细信息，请参见 SDL 文档。

-spice option[,option[,...]]

启用 spice 远程桌面协议。

-display vnc

有关更多信息，请参考第 37.5 节 “使用 VNC 查看 VM Guest”。

-nographic

禁用 QEMU 的图形输出。模拟的串行端口将重定向到控制台。

使用 -nographic 启动虚拟机后，在虚拟控制台中按 **Ctrl - A H** 可查看其他有用快捷键的列表，例如，用于在控制台与 QEMU 监控器之间切换的快捷键。

```
> sudo qemu-system-x86_64 -hda /images/sles_base.raw -nographic
```

```
C-a h    print this help
C-a x    exit emulator
C-a s    save disk data back to file (if -snapshot)
C-a t    toggle console timestamps
C-a b    send break (magic sysrq)
C-a c    switch between console and monitor
C-a C-a  sends C-a
(pressed C-a c)

QEMU 2.3.1 monitor - type 'help' for more information
(qemu)
```

-no-frame

禁用 QEMU 窗口的装饰。便于在专用桌面工作空间中操作。

-full-screen

以全屏模式启动 QEMU 图形输出。

-no-quit

禁用 QEMU 窗口的关闭按钮，防止强行关闭窗口。

-alt-grab、-ctrl-grab

默认情况下，在按 **Ctrl - Alt** 之后，QEMU 窗口会释放“捕获的”鼠标。您可以将组合键更改为 **Ctrl - Alt - Shift** (-alt-grab) 或右 **Ctrl** 键 (-ctrl-grab)。

37.3.3 USB 设备

可通过两种方式来创建可供 KVM 中的 VM Guest 使用的 USB 设备：可以在 VM Guest 中模拟新的 USB 设备，或将现有的主机 USB 设备分配给 VM Guest。要在 QEMU 中使用 USB 设备，首先需要通过 `-usb` 选项启用通用 USB 驱动程序。然后可以通过 `-usbdevice` 选项指定各个设备。

37.3.3.1 在 VM Guest 中模拟 USB 设备

SUSE 目前支持以下类型的 USB 设备：`disk`、`host`、`serial`、`braille`、`net`、`mouse` 和 `tablet`。

`-usbdevice` 选项的 USB 设备类型

`disk`

基于文件模拟大容量存储设备。可以使用可选的 `format` 选项，而不要检测格式。

```
> sudo qemu-system-x86_64 [...] -usbdevice  
    disk:format=raw:/virt/usb_disk.raw
```

`host`

直通主机设备（由 `bus.addr` 标识）。

`serial`

主机字符设备的串行转换器。

`braille`

使用 BrAPI 模拟盲文设备以显示盲文输出。

`net`

模拟支持 CDC 以太网和 RNDIS 协议的网络适配器。

`mouse`

模拟虚拟 USB 鼠标。此选项会覆盖默认的 PS/2 鼠标模拟。以下示例显示了使用 `qemu-system-ARCH [...] -usbdevice mouse` 启动的 VM Guest 上的鼠标硬件状态：

```
> sudo hwinfo --mouse  
20: USB 00.0: 10503 USB Mouse
```

```
[Created at usb.122]
UDI: /org/freedesktop/Hal/devices/usb_device_627_1_1_if0
[...]
Hardware Class: mouse
Model: "Adomax QEMU USB Mouse"
Hotplug: USB
Vendor: usb 0x0627 "Adomax Technology Co., Ltd"
Device: usb 0x0001 "QEMU USB Mouse"
[...]
```

tablet

模拟使用绝对坐标的定位设备（例如触摸屏）。此选项会覆盖默认的 PS/2 鼠标模拟。如果您要通过 VNC 协议查看 VM Guest，则绘图板设备非常有用。有关更多信息，请参见第 37.5 节“使用 VNC 查看 VM Guest”。

37.3.4 字符设备

使用 -chardev 可创建新的字符设备。该选项使用以下一般语法：

```
qemu-system-x86_64 [...] -chardev BACKEND_TYPE,id=ID_STRING
```

其中，BACKEND_TYPE 可以是

null、socket、udp、msmouse、vc、file、pipe、console、serial、pty、stdio、braille、tty 或 parport。所有字符设备都必须有一个最长为 127 个字符的唯一标识字符串。此字符串用于在其他相关指令中标识该设备。有关后端的所有子选项的完整说明，请参见手册页 (man 1 qemu)。下面是可用 back-ends 的简要说明：

null

创建一个空设备，该设备不输出数据且会丢弃收到的所有数据。

stdio

连接到 QEMU 的进程标准输入和标准输出。

socket

创建双向流套接字。如果指定了 PATH，则创建 Unix 套接字：

```
> sudo qemu-system-x86_64 [...] -chardev \
```



```
socket,id=unix_socket1,path=/tmp/unix_socket1,server
```

SERVER 子选项指定该套接字是侦听套接字。

如果指定了 PORT，则创建 TCP 套接字：

```
> sudo qemu-system-x86_64 [...] -chardev \
socket,id=tcp_socket1,host=localhost,port=7777,server,nowait
```

该命令在端口 7777 上创建一个本地侦听 (server) TCP 套接字。QEMU 不会因为等待客户端连接到侦听端口而进入阻塞状态 (nowait)。

udp

通过 UDP 协议将来自 VM Guest 的所有网络流量发送到远程主机。

```
> sudo qemu-system-x86_64 [...] \
-chardev udp,id=udp_fwd,host=mercury.example.com,port=7777
```

该命令会在远程主机 mercury.example.com 上绑定端口 7777，并从中发送 VM Guest 网络流量。

vc

创建新的 QEMU 文本控制台。您可以选择性地指定虚拟控制台的尺寸：

```
> sudo qemu-system-x86_64 [...] -chardev vc,id=vc1,width=640,height=480 \
-mon chardev=vc1
```

该命令会创建指定大小且名为 vc1 的新虚拟控制台，并将 QEMU 监控器连接到该控制台。

file

将来自 VM Guest 的所有流量都记录到 VM 主机服务器上的一个文件。必须指定 path，如果该路径不存在，系统将予以创建。

```
> sudo qemu-system-x86_64 [...] \
-chardev file,id=qemu_log1,path=/var/log/qemu/guest1.log
```

默认情况下，QEMU 将为串行与并行端口创建一组字符设备，并为 QEMU 监控器创建一个特殊控制台。不过，您可以创建自己的字符设备，并将其用于所述目的。以下选项可为您提供帮助：

`-serial CHAR_DEV`

将 VM Guest 的虚拟串行端口重定向到 VM 主机服务器上的字符设备 `CHAR_DEV`。在图形模式下，此设备默认为一个虚拟控制台 (vc)；在非图形模式下，默认为 `stdio`。`-serial` 可识别许多子选项。有关子选项的完整列表，请参见手册页 `man 1 qemu`。您最多可以模拟四个串行端口。使用 `-serial none` 可禁用所有串行端口。

`-parallel DEVICE`

将 VM Guest 的并行端口重定向到 `DEVICE`。此选项支持的设备与 `-serial` 相同。



提示

使用 SUSE Linux Enterprise Server 作为 VM 主机服务器时，您可以直接使用硬件并行端口设备 `/dev/parportN`（其中的 `N` 是端口号）。

您最多可以模拟三个并行端口。使用 `-parallel none` 可禁用所有并行端口。

`-monitor CHAR_DEV`

将 QEMU 监控器重定向到 VM 主机服务器上的字符设备 `CHAR_DEV`。此选项支持的设备与 `-serial` 相同。在图形模式下，此设备默认为一个虚拟控制台 (vc)；在非图形模式下，默认为 `stdio`。

有关可用字符设备后端的完整列表，请参见手册页 (`man 1 qemu`)。

37.4 QEMU 中的网络

将 `-netdev` 选项与 `-device` 结合使用可为 VM Guest 定义特定类型的网络和网络接口卡。`-netdev` 选项的语法为

```
-netdev type[,prop[=value][,...]]
```

SUSE 目前支持以下网络类型：`user`、`bridge` 和 `tap`。有关 `-netdev` 子选项的完整列表，请参见手册页 (`man 1 qemu`)。

bridge

使用指定的网络助手来配置 TAP 接口并将其挂接到指定的网桥。有关详细信息，请参见第 37.4.3 节“桥接网络”。

user

指定用户模式网络。有关详细信息，请参见第 37.4.2 节“用户模式网络”。

tap

指定桥接网络或路由网络。有关详细信息，请参见第 37.4.3 节“桥接网络”。

37.4.1 定义网络接口卡

将 `-netdev` 与相关的 `-device` 选项一起使用可以添加新的模拟网卡：

```
> sudo qemu-system-x86_64 [...] \  
-netdev tap ❶,id=hostnet0 \  
-device virtio-net-pci ❷,netdev=hostnet0,vlan=1 ❸,\  
macaddr=00:16:35:AF:94:4B ❹,name=ncard1
```

- ❶ 指定网络设备类型。
- ❷ 指定网卡的型号。使用 `qemu-system-ARCH -device help` 并搜索 `Network devices`：部分可获取您平台上受 QEMU 支持的所有网卡型号的列表。
SUSE 目前支持型号 `rtl8139`、`e1000` 及其衍生产品
`e1000-82540em`、`e1000-82544gc`、`e1000-82545em` 和 `virtio-net-pci`。要查看特定驱动程序的选项列表，请添加 `help` 作为驱动程序选项：

```
> sudo qemu-system-x86_64 -device e1000,help  
e1000.mac=macaddr  
e1000.vlan=vlan  
e1000.netdev=netdev  
e1000.bootindex=int32  
e1000.autonegotiation=on/off  
e1000.mitigation=on/off  
e1000.addr=pci-devfn
```

```
e1000.romfile=str  
e1000.rombar=uint32  
e1000.multifunction=on/off  
e1000.command_serr_enable=on/off
```

- ③ 将网络接口连接到 VLAN 1。您可以指定自己的编号，该编号主要用于标识目的。如果您省略此子选项，QEMU 将使用默认值 0。
- ④ 指定网卡的媒体访问控制 (MAC) 地址。它是一个唯一标识符，建议您始终指定该地址。如果未指定，QEMU 将提供自己的默认 MAC 地址，因此可能会在相关 VLAN 中造成 MAC 地址冲突。

37.4.2 用户模式网络

`-netdev user` 选项指示 QEMU 使用用户模式网络。如果未选择网络模式，则默认使用用户模式。因此，这些命令行等效于：

```
> sudo qemu-system-x86_64 -hda /images/sles_base.raw
```

```
> sudo qemu-system-x86_64 -hda /images/sles_base.raw -netdev user,id=hostnet0
```

如果您要允许 VM Guest 访问外部网络资源（例如互联网），则此模式非常有用。默认不允许任何传入流量，因此 VM Guest 对于网络中的其他计算机不可见。在此网络模式下，将不需要管理员特权。用户模式还可用于从 VM 主机服务器上的本地目录在 VM Guest 上执行网络引导。

VM Guest 会获得虚拟 DHCP 服务器分配的一个 IP 地址。VM 主机服务器（DHCP 服务器）可通过 10.0.2.2 访问，而分配的 IP 地址范围从 10.0.2.15 开始。您可以使用 **ssh** 连接到位于 10.0.2.2 的 VM 主机服务器，并使用 **scp** 在两者之间复制文件。

37.4.2.1 命令行示例

本节提供了有关如何使用 QEMU 设置用户模式网络的几个示例。

例 37.1：受限用户模式网络

```
> sudo qemu-system-x86_64 [...] \  
-netdev user①,id=hostnet0 \  
-hda /images/sles_base.raw
```

```
-device virtio-net-pci,netdev=hostnet0,vlan=1②,name=user_net1③,restrict=yes④
```

- ① 指定用户模式网络。
- ② 连接到 VLAN 1。如果省略此选项，则默认使用 0。
- ③ 指定网络堆栈的直观易懂名称。可用于在 QEMU 监控器中标识该堆栈。
- ④ 隔离 VM Guest。这样 VM Guest 将无法与 VM 主机服务器通讯，并且网络包将不会路由到外部网络。

例 37.2：使用自定义 IP 范围的用户模式网络

```
> sudo qemu-system-x86_64 [...] \  
-netdev user,id=hostnet0 \  
-device virtio-net-pci,netdev=hostnet0,net=10.2.0.0/8①,host=10.2.0.6②,\  
dhcpstart=10.2.0.20③,hostname=tux_kvm_guest④
```

- ① 指定 VM Guest 看到的网络 IP 地址，以及可选的网络掩码。默认值为 10.0.2.0/8。
- ② 指定 VM Guest 看到的 VM 主机服务器 IP 地址。默认值为 10.0.2.2。
- ③ 指定可由内置 DHCP 服务器指派给 VM Guest 的 16 个 IP 地址中的第一个。默认值为 10.0.2.15。
- ④ 指定由内置 DHCP 服务器分配给 VM Guest 的主机名。

例 37.3：使用网络引导和 TFTP 的用户模式网络

```
> sudo qemu-system-x86_64 [...] \  
-netdev user,id=hostnet0 \  
-device virtio-net-pci,netdev=hostnet0,tftp=/images/tftp_dir①,\  
bootfile=/images/boot/pxelinux.0②
```

- ① 激活内置 TFTP（提供基本 FTP 功能的文件传输协议）服务器。指定目录中的文件将以 TFTP 服务器根目录的形式显示给 VM Guest。
- ② 以 BOOTP（可提供引导映像 IP 地址和网络位置的一种网络协议，通常在无盘工作站中使用）文件的形式广播指定的文件。与 `tftp` 一起使用时，可以通过主机上的本地目录从网络引导 VM Guest。

例 37.4：使用主机端口转发的用户模式网络

```
> sudo qemu-system-x86_64 [...] \  

```

```
-netdev user,id=hostnet0 \  
-device virtio-net-pci,netdev=hostnet0,hostfwd=tcp::2222-:22
```

将主机上端口 2222 的传入 TCP 连接转发到 VM Guest 上的端口 22 (SSH)。如果 `sshd` 正在 VM Guest 上运行，请输入

```
> ssh qemu_host -p 2222
```

(其中，`qemu_host` 是主机系统的主机名或 IP 地址)，以获取 VM Guest 的 SSH 提示。

37.4.3 桥接网络

使用 `-netdev tap` 选项时，QEMU 会通过将主机 TAP 网络设备连接到 VM Guest 的指定 VLAN 来创建网桥。该网络设备的网络接口便会对网络的其余部分可见。此方法默认不会启用，需要明确指定。

首先创建一个网桥，并将一个 VM 主机服务器物理网络接口（例如 `eth0`）添加到其中：

1. 启动 YaST 控制中心并选择系统 > 网络设置。
2. 单击添加，然后从硬件对话框窗口的设备类型下拉框中选择网桥。单击下一步。
3. 选择您需要使用动态还是静态分配的 IP 地址，然后填写相关网络设置（如果适用）。
4. 在桥接设备窗格中，选择要添加到网桥的以太网设备。
单击下一步。出现有关调整已配置设备的提示时，请单击继续。
5. 单击确定以应用更改。检查是否已创建网桥：

```
> bridge link  
2: eth0 state UP : <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 master br0 \  
state forwarding priority 32 cost 100
```

37.4.3.1 手动连接到网桥

使用以下示例脚本将 VM Guest 连接到新建的网桥接口 `br0`。脚本中的数个命令通过 `sudo` 机制运行，原因是这些命令需要 `root` 特权。



提示：所需的软件

要管理网桥，需要安装 `tunctl` 软件包。

```
#!/bin/bash
bridge=br0 ❶
tap=$(sudo tunctl -u $(whoami) -b) ❷
sudo ip link set $tap up ❸
sleep 1s ❹
sudo ip link add name $bridge type bridge
sudo ip link set $bridge up
sudo ip link set $tap master $bridge ❺
qemu-system-x86_64 -machine accel=kvm -m 512 -hda /images/sles_base.raw \
-netdev tap,id=hostnet0 \
-device virtio-net-pci,netdev=hostnet0,vlan=0,macaddr=00:16:35:AF:94:4B,\
ifname=$tap ❻,script=no ❼,downscript=no
sudo ip link set $tap nomaster ❽
sudo ip link set $tap down ❾
sudo tunctl -d $tap ❿
```

- ❶ 网桥设备的名称。
- ❷ 准备新的 TAP 设备并将其分配给运行脚本的用户。TAP 设备是常用于虚拟化和模拟设置的虚拟网络设备。
- ❸ 启动新建的 TAP 网络接口。
- ❹ 暂停 1 秒，以确保新 TAP 网络接口确实启动。
- ❺ 将新 TAP 设备添加到网桥 `br0`。
- ❻ `ifname=` 子选项指定用于桥接的 TAP 网络接口的名称。
- ❼ **`qemu-system-ARCH`** 会在连接到网桥之前检查 `script` 和 `downscript` 值。如果它在 VM 主机服务器文件系统中找到了指定的脚本，将会在连接到网桥之前运行 `script`，并在退出网络环境之后运行 `downscript`。您可以使用这些脚本来设置和拆除桥接接口。默认会检查 `/etc/qemu-ifup` 和 `/etc/qemu-ifdown`。如果指定了 `script=no` 和 `downscript=no`，则会禁止执行脚本，您需要手动执行该脚本。
- ❽ 删除网桥 `br0` 中的 TAP 接口。

- 9 将 TAP 设备的状态设置为 `down`。
- 10 拆除 TAP 设备。

37.4.3.2 使用 `qemu-bridge-helper` 连接到网桥

通过网桥将 VM Guest 连接到网络的另一种方式是使用 `qemu-bridge-helper` 助手程序。该程序可为您配置 TAP 接口并将其挂接到指定的网桥。默认的助手可执行文件为 `/usr/lib/qemu-bridge-helper`。该助手可执行文件的权限要求为 `setuid root`，也就是说，只允许虚拟化组 (`kvm`) 的成员执行。因此，`qemu-system-ARCH` 命令本身并不需要以 `root` 特权运行。当您指定网桥时，会自动调用该助手：

```
qemu-system-x86_64 [...] \  
-netdev bridge,id=hostnet0,vlan=0,br=br0 \  
-device virtio-net-pci,netdev=hostnet0
```

您可以使用 `helper=/path/to/your/helper` 选项指定自己的自定义助手脚本来处理 TAP 设备配置或解除配置：

```
qemu-system-x86_64 [...] \  
-netdev bridge,id=hostnet0,vlan=0,br=br0,helper=/path/to/bridge-helper \  
-device virtio-net-pci,netdev=hostnet0
```



提示

要定义对 `qemu-bridge-helper` 的访问特权，请检查 `/etc/qemu/bridge.conf` 文件。例如，以下指令

```
allow br0
```

允许 `qemu-system-ARCH` 命令将其 VM Guest 连接到网桥 `br0`。

37.5 使用 VNC 查看 VM Guest

默认情况下，QEMU 使用 GTK（一个跨平台工具包库）窗口来显示 VM Guest 的图形输出。如果指定了 `-vnc` 选项，您可以让 QEMU 侦听指定的 VNC 显示器，并将其图形输出重定向到 VNC 会话。



提示

通过 VNC 会话操作 QEMU 的虚拟机时，使用 `-usbdevice tablet` 选项会很有用。

此外，如果您需要使用另一种键盘布局而不是默认的 `en-us`，请使用 `-k` 选项指定所需布局。

`-vnc` 的第一个子选项必须是 **display** 值。`-vnc` 选项识别以下 display 指定值：

host:display

只接受来自显示器编号 display 上的 host 的连接。随后运行 VNC 会话的 TCP 端口通常是值为 `5900 + display` 的数字。如果未指定 host，系统将接受来自任何主机的连接。

unix:path

VNC 服务器侦听 Unix 域套接字上的连接。path 选项指定相关 Unix 套接字的位置。

none

将初始化 VNC 服务器功能，但不启动该服务器本身。您稍后可以使用 QEMU 监控器启动 VNC 服务器。有关详细信息，请参见第 38 章“使用 QEMU 监控器管理虚拟机”。

可以在 display 值的后面使用一个或多个选项标志（以逗号分隔）。有效选项为：

reverse

通过**反向**连接来连接侦听方 VNC 客户端。

websocket

额外打开一个专用于 VNC Websocket 连接的 TCP 侦听端口。根据定义，Websocket 端口为 `5700+display`。

password

要求对客户端连接使用基于口令的身份验证。

tls

要求客户端在与 VNC 服务器通讯时使用 TLS。

x509=/path/to/certificate/dir

指定了 TLS 时有效。要求使用 x509 身份凭证来协商 TLS 会话。

x509verify=/path/to/certificate/dir

指定了 TLS 时有效。要求使用 x509 身份凭证来协商 TLS 会话。

sasl

要求客户端使用 SASL 向 VNC 服务器进行身份验证。

acl

打开访问控制列表，以检查 x509 客户端证书和 SASL 参与方。

lossy

启用有损压缩方法（梯度、JPEG 等）。

non-adaptive

禁用自适应编码。默认会启用自适应编码。

share=[allow-exclusive|force-shared|ignore]

设置显示共享策略。



注意

有关显示选项的更多细节，请参见 **qemu-doc** 手册页。

VNC 示例用法：

```
tux > sudo qemu-system-x86_64 [...] -vnc :5
# (on the client:)
wilber > vncviewer venus:5 &
```

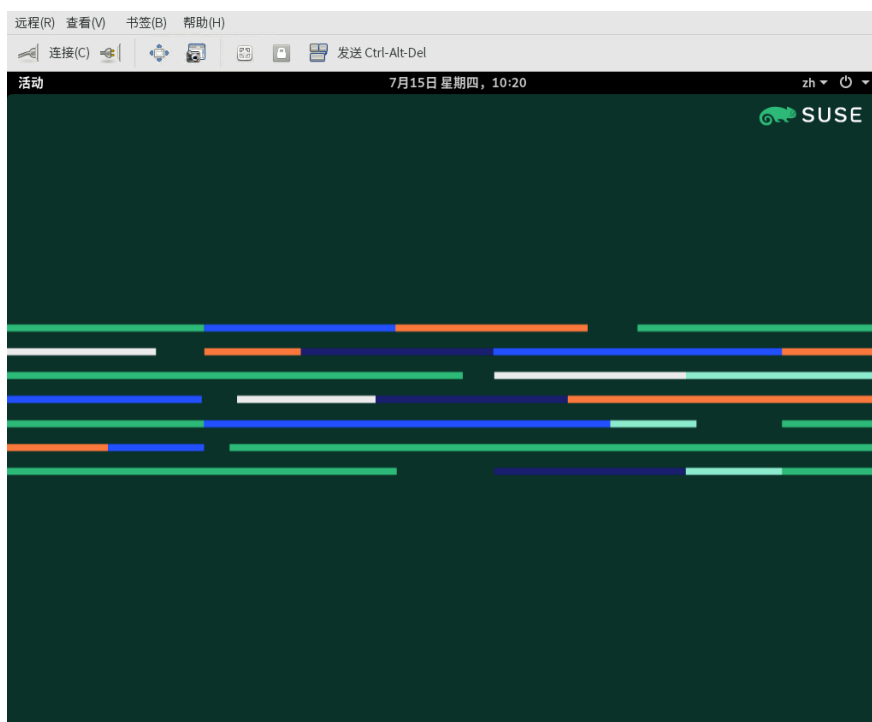


图 37.2：QEMU VNC 会话

37.5.1 保护 VNC 连接

默认的 VNC 服务器设置不使用任何形式的身份验证。在前面的示例中，任何用户都可以从网络中的任何主机连接和查看 QEMU VNC 会话。

系统提供了多个级别的安全性，可供您应用于 VNC 客户端/服务器连接。您可以使用口令、x509 证书、SASL 身份验证，甚至可在一条 QEMU 命令中结合多种身份验证方法来保护连接。

有关在 VM 主机服务器和客户端上配置 x509 证书的详细信息，请参见第 12.3.2 节“使用 x509 证书进行远程 TLS/SSL 连接 (qemu+tls 或 xen+tls)”和第 12.3.2.3 节“配置客户端并测试设置”。

Remmina VNC 查看器支持高级身份验证机制。对于此示例，我们假设服务器 x509 证书 `ca-cert.pem`、`server-cert.pem` 和 `server-key.pem` 位于主机上的 `/etc/pki/qemu` 目录中。可将客户端证书放在任何自定义目录中，Remmina 在连接启动时会要求提供这些证书的路径。

例 37.5：口令身份验证

```
qemu-system-x86_64 [...] -vnc :5,password -monitor stdio
```

在 VNC 显示器编号 5（对应于端口 5905）上启动 VM Guest 图形输出。`password` 子选项会初始化一种基于口令的简单身份验证方法。系统默认未设置口令，您需要在 QEMU 监控器中使用 `change vnc password` 命令设置一个口令：

```
QEMU 2.3.1 monitor - type 'help' for more information
(qemu) change vnc password
Password: ****
```

此处需要指定 `-monitor stdio` 选项，因为如果不重定向 QEMU 监控器的输入/输出，您将无法管理该监控器。

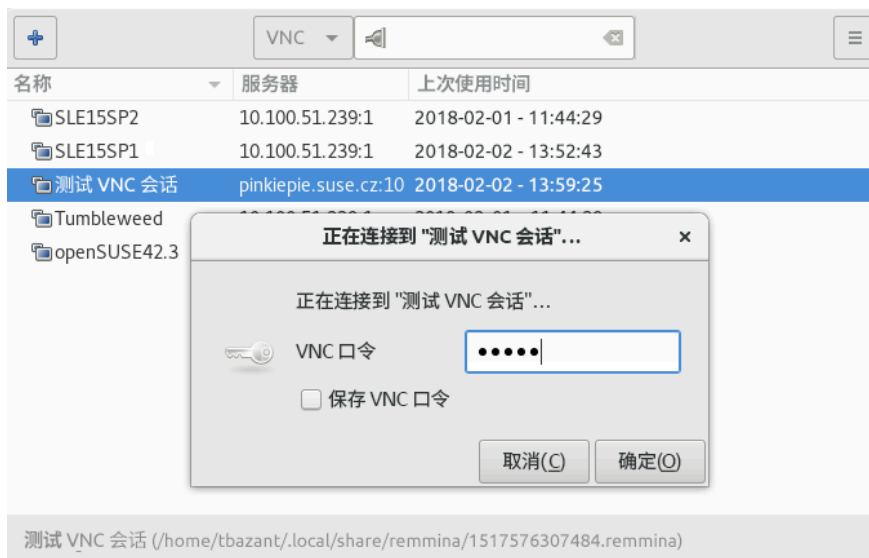


图 37.3：REMMINA 中的身份验证对话框

例 37.6：X509 证书身份验证

QEMU VNC 服务器可对会话使用 TLS 加密，并使用 x509 证书进行身份验证。服务器将要求客户端提供证书，并根据 CA 证书验证提供的证书。如果您的公司可提供内部证书颁发机构，请使用此身份验证类型。

```
qemu-system-x86_64 [...] -vnc :5,tls,x509verify=/etc/pki/qemu
```

例 37.7：X509 证书和口令身份验证

您可以将口令身份验证与 TLS 加密和 x509 证书身份验证结合使用，以便为客户端打造双层身份验证模型。运行以下命令后，请记得在 QEMU 监控器中设置口令：

```
qemu-system-x86_64 [...] -vnc :5,password,tls,x509verify=/etc/pki/qemu \
```

```
-monitor stdio
```

例 37.8：SASL 身份验证

简单身份验证和安全层 (SASL) 是互联网协议中的身份验证和数据安全性框架。它集成了多种身份验证机制，例如 PAM、Kerberos、LDAP 等等。SASL 会维护自己的用户数据库，因此 VM 主机服务器上无需存在连接用户帐户。

出于安全考虑，建议您将 SASL 身份验证与 TLS 加密和 x509 证书结合使用：

```
qemu-system-x86_64 [...] -vnc :5,tls,x509,sasl -monitor stdio
```

38 使用 QEMU 监控器管理虚拟机

通过 `qemu-system-ARCH` 命令（例如 `qemu-system-x86_64`）调用虚拟机时，会提供一个监控器控制台用来与用户交互。使用监控器控制台中提供的命令可以检查运行中的操作系统、更改可移动媒体、截取屏幕截图或音频片段，以及控制虚拟机的其他方面。



注意

下列章节列出了精选的实用 QEMU 监控器命令及其用途。要获取完整列表，请在 QEMU 监控器命令行中输入 `help`。

38.1 访问监控器控制台



提示：libvirt 没有监控器控制台

仅当您直接使用 `qemu-system-ARCH` 命令启动虚拟机并在内置 QEMU 窗口中查看其图形输出时，才可以访问监控器控制台。

如果您使用 `libvirt` 启动了虚拟机（例如，使用 `virt-manager`）并通过 VNC 或 Spice 会话查看其输出，则无法直接访问监控器控制台。不过，您可以通过 `virsh` 将监控器命令发送到虚拟机：

```
# virsh qemu-monitor-command COMMAND
```

访问监控器控制台的方式取决于您使用哪种显示设备来查看虚拟机的输出。第 37.3.2.2 节“显示选项”中提供了有关显示器的更多细节。例如，要在使用 `-display gtk` 选项的情况下查看监控器，请按 `Ctrl - Alt - 2`。同样，在使用 `-nographic` 选项时，可按以下组合键切换到监控器控制台：`Ctrl - A C`。

在使用控制台时如需帮助，请使用 `help` 或 `?`。要获取有关特定命令的帮助，请使用 `help COMMAND`。

38.2 获取有关 Guest 系统的信息

要获取有关 Guest 系统的信息，请使用 **info**。如果不结合任何选项使用该命令，将列显可能的选项的列表。选项可确定要分析系统的哪个部分：

info version

显示 QEMU 的版本。

info commands

列出可用的 QMP 命令。

info network

显示网络状态。

info chardev

显示字符设备。

info block

有关块设备（例如硬盘、软盘驱动器或 CD-ROM）的信息。

info blockstats

块设备的读取和写入统计数据。

info registers

显示 CPU 寄存器。

info cpus

显示有关可用 CPU 的信息。

info history

显示命令行历史。

info irq

显示中断统计数据。

info pic

显示 i8259 (PIC) 状态。

info pci

显示 PCI 信息。

info tlb

显示虚拟内存到物理内存的映射。

info mem

显示活动的虚拟内存映射。

info jit

显示动态编译器信息。

info kvm

显示 KVM 信息。

info numa

显示 NUMA 信息。

info usb

显示 Guest USB 设备。

info usbhost

显示主机 USB 设备。

info profile

显示分析信息。

info capture

显示捕获（音频抓取）信息。

info snapshots

显示当前保存的虚拟机快照。

info status

显示当前虚拟机的状态。

info mice

显示哪些 Guest 鼠标正在接收事件。

info vnc

显示 VNC 服务器状态。

info name

显示当前虚拟机的名称。

info uuid

显示当前虚拟机的 UUID。

info usernet

显示用户网络堆栈连接状态。

info migrate

显示迁移状态。

info balloon

显示气球设备信息。

info qtree

显示设备树。

info qdm

显示 qdev 设备型号列表。

info roms

显示 ROM。

info migrate_cache_size

显示当前迁移 xbzrle（“基于 Xor 的零运行长度编码”）缓存大小。

info migrate_capabilities

显示多个迁移功能（例如 xbzrle 压缩）的状态。

info mtree

显示 VM Guest 内存层次结构。

info trace-events

显示可用的跟踪事件及其状态。

38.3 更改 VNC 口令

要更改 VNC 口令，请使用 **`change vnc password`** 命令并输入新口令：

```
(qemu) change vnc password
Password: *****
(qemu)
```

38.4 管理设备

要在 Guest 运行时添加新磁盘（热插入），请使用 **`drive_add`** 和 **`device_add`** 命令。首先定义要作为设备添加到总线 0 的新驱动器：

```
(qemu) drive_add 0 if=none,file=/tmp/test.img,format=raw,id=disk1
OK
```

可以通过查询块子系统来确认新设备：

```
(qemu) info block
[...]
disk1: removable=1 locked=0 tray-open=0 file=/tmp/test.img ro=0 drv=raw \
encrypted=0 bps=0 bps_rd=0 bps_wr=0 iops=0 iops_rd=0 iops_wr=0
```

定义新驱动器后，需将它连接到某个设备，使 Guest 能够看到它。典型的设备是 **`virtio-blk-pci`** 或 **`scsi-disk`**。要获取可用值的完整列表，请运行：

```
(qemu) device_add ?
name "VGA", bus PCI
name "usb-storage", bus usb-bus
[...]
name "virtio-blk-pci", bus virtio-bus
```

现在添加设备

```
(qemu) device_add virtio-blk-pci,drive=disk1,id=myvirtio1
```

并使用以下命令确认

```
(qemu) info pci
```

```
[...]
Bus 0, device 4, function 0:
  SCSI controller: PCI device 1af4:1001
  IRQ 0.
  BAR0: I/O at 0xffffffffffffffff [0x003e].
  BAR1: 32 bit memory at 0xffffffffffffffff [0x0000ffe].
  id "myvirtio1"
```



提示

可以使用 **`device_del`** 从 Guest 中去除通过 **`device_add`** 命令添加的设备。如需详细信息，请在 QEMU 监控器命令行中输入 **`help device_del`**。

要释放与可移动媒体设备连接的设备或文件，请使用 **`eject DEVICE`** 命令。使用可选的 **`-f`** 可以强制弹出。

要更改可移动媒体（例如 CD-ROM），请使用 **`change DEVICE`** 命令。可以使用 **`info block`** 命令确定可移动媒体的名称：

```
(qemu) info block
ide1-cd0: type=cdrom removable=1 locked=0 file=/dev/sr0 ro=1 drv=host_device
(qemu) change ide1-cd0 /path/to/image
```

38.5 控制键盘和鼠标

如果需要，可以使用监控器控制台来模拟键盘和鼠标输入。例如，如果您的图形用户界面拦截了某些低级别的组合键（例如 X Window 系统中的 **`Ctrl - Alt - F1`**），您仍可以使用 **`sendkey KEYS`** 来输入这些组合键：

```
sendkey ctrl-alt-f1
```

要列出 **`KEYS`** 选项中使用的按键名称，请输入 **`sendkey`** 并按 **`→|`**。

要控制鼠标，可使用以下命令：

`mouse_move DX dy [DZ]`

将活动的鼠标指针移到指定的坐标 dx, dy, dz（该滚动轴可选）。

mouse_button VAL

更改鼠标按钮的状态（1=左，2=中，4=右）。

mouse_set INDEX

设置由哪个鼠标设备接收事件。可以使用 **info mice** 命令获取设备索引号。

38.6 更改可用内存

如果启动虚拟机时使用了 **-balloon virtio** 选项（如此会启用半虚拟化气球设备），您便可以动态更改可用内存。有关启用气球设备的详细信息，请参见第 36.1 节“使用 **qemu-system-ARCH** 进行基本安装”。

要在监控器控制台中获取有关气球设备的信息，并确定该设备是否已启用，请使用 **info balloon** 命令：

```
(qemu) info balloon
```

如果气球设备已启用，请使用 **balloon MEMORY_IN_MB** 命令设置请求的内存量：

```
(qemu) balloon 400
```

38.7 转储虚拟机内存

要将虚拟机内存的内容保存到磁盘或控制台输出，请使用以下命令：

memsaveADDRSIZEFILENAME

将起始地址为 ADDR、大小为 SIZE 的虚拟内存转储保存到 FILENAME 文件中

pmemsaveADDRSIZEFILENAME

将起始地址为 ADDR、大小为 SIZE 的物理内存转储保存到 FILENAME- 文件中

x /FMTADDR

创建起始地址为 ADDR 并根据 FMT 字符串设置格式的虚拟内存转储。FMT 字符串由 COUNTFORMATSIZE 这三个参数构成：

COUNT 参数是要转储的项数。

FORMAT 可以是 x (十六进制)、d (有符号十进制)、u (无符号十进制)、o (八进制)、c (字符) 或 i (汇编指令)。

SIZE 参数可以是 b (8 位)、h (16 位)、w (32 位) 或 g (64 位)。在 x86 上，可以使用 i 格式指定 h 或 w，以分别选择 16 位或 32 位代码指令大小。

xp /FMTADDR

创建起始地址为 ADDR 并根据 FMT 字符串设置格式的物理内存转储。FMT 字符串由 COUNTFORMATSIZE 这三个参数构成：

COUNT 参数是要转储的项数。

FORMAT 可以是 x (十六进制)、d (有符号十进制)、u (无符号十进制)、o (八进制)、c (字符) 或 i (汇编指令)。

SIZE 参数可以是 b (8 位)、h (16 位)、w (32 位) 或 g (64 位)。在 x86 上，可以使用 i 格式指定 h 或 w，以分别选择 16 位或 32 位代码指令大小。

38.8 管理虚拟机快照

SUSE 目前不支持在 QEMU 监控器中管理快照。本节中的信息在特定的情形下可能有帮助。

虚拟机快照是整个虚拟机的快照，包括 CPU、RAM 的状态以及所有可写磁盘的内容。要使用虚拟机快照，您必须至少有一个使用 qcow2 磁盘映像格式且不可移动的可写块设备。

当您需要保存特定状态的虚拟机时，快照非常有用。例如，在虚拟化服务器上配置网络服务后，您可以从上次保存的虚拟机状态快速启动虚拟机。您还可以在关闭虚拟机之后创建快照，以便在尝试执行某种会导致 VM Guest 不稳定的试验性操作之前创建备份状态。本节介绍前一种做法，后一种做法已在第 36.2.3 节“**使用 qemu-img 管理虚拟机的快照**”中介绍。

可在 QEMU 监控器中使用以下命令管理快照：

savevmNAME

创建一个标记为 NAME 的新虚拟机快照，或替换现有快照。

loadvmNAME

加载标记为 NAME 的虚拟机快照。

delvm

删除虚拟机快照。

info snapshots

列显有关可用快照的信息。

```
(qemu) info snapshots
Snapshot list:
ID ①      TAG ②      VM SIZE ③    DATE ④      VM CLOCK ⑤
1        booting      4.4M 2013-11-22 10:51:10  00:00:20.476
2        booted       184M 2013-11-22 10:53:03  00:02:05.394
3        logged_in    273M 2013-11-22 11:00:25  00:04:34.843
4        ff_and_term_running 372M 2013-11-22 11:12:27  00:08:44.965
```

- ① 快照的自动递增唯一标识号。
- ② 快照的唯一说明字符串。它以直观易懂的 ID 形式来表示。
- ③ 快照占用的磁盘空间。运行中应用程序消耗的内存越多，快照就越大。
- ④ 快照的创建时间和日期。
- ⑤ 虚拟机时钟的当前状态。

38.9 挂起和恢复虚拟机执行

以下命令可用于挂起和恢复虚拟机：

stop

挂起虚拟机的执行。

cont

恢复虚拟机的执行。

system_reset

重置虚拟机。效果类似于物理机上的复位按钮。这可能会使文件系统处于一种不干净状态。

system_powerdown

向计算机发送 **ACPI** 关机请求。效果类似于物理机上的电源按钮。

q 或 quit

立即终止 QEMU。

38.10 动态迁移

实时迁移过程可将任何虚拟机从一个主机系统传输到另一个主机系统，而不会对可用性造成任何干扰。您可以永久性更改主机，或者仅在维护期间更改主机。

实时迁移的要求：

- 第 17.2 节 “迁移要求” 中所述的所有要求均适用。
- 只能在具有相同 CPU 功能的 VM 主机服务器之间进行实时迁移。
- 无法使用 AHCI 接口、VirtFS 功能和 `-mem-path` 命令行选项进行迁移。
- 必须以相同的方式启动源主机和目标主机上的 Guest。
- 不应使用 `-snapshot qemu` 命令行选项进行迁移（不支持此 `qemu` 命令行选项）。



重要：支持状态

SUSE Linux Enterprise Server 中尚不支持 `postcopy` 模式。此模式仅发布为技术预览版。有关 `postcopy` 的详细信息，请参见 <https://wiki.qemu.org/Features/PostCopyLiveMigration>。

以下网站上提供了更多建议：<https://www.linux-kvm.org/page/Migration>

实时迁移过程包括以下步骤：

1. 虚拟机实例正在源主机上运行。
2. 虚拟机以冻结侦听模式在目标主机上启动。使用的参数与源主机上相同，不过还要加上 `-incoming tcp:IP:PORT` 参数，其中 `IP` 指定 IP 地址，`PORT` 指定用于侦听传入迁移的端口。如果设置了 0 作为 IP 地址，则虚拟机将侦听所有接口。
3. 在源主机上，切换到监控器控制台，并使用 `migrate -d tcp:DESTINATION_IP:PORT` 命令启动迁移。
4. 要确定迁移状态，请在源主机上的监控器控制台中使用 `info migrate` 命令。
5. 要取消迁移，请在源主机上的监控器控制台中使用 `migrate_cancel` 命令。

6. 要设置迁移时容许的最长停机时间（以秒为单位），请使用 `migrate_set_downtime` `NUMBER_OF_SECONDS` 命令。
7. 要设置最大迁移速度（每秒字节数），请使用 `migrate_set_speed` `BYTES_PER_SECOND` 命令。

38.11 QMP - QEMU 计算机协议

QMP 是基于 JSON 的协议，可使应用程序（例如 `libvirt`）能够与运行中的 QEMU 实例通讯。您可以使用 QMP 命令以多种方式访问 QEMU 监控器。

38.11.1 通过标准输入/输出访问 QMP

最灵活的 QMP 使用方式是指定 `-mon` 选项。下面的示例使用标准输入/输出创建了一个 QMP 实例。在以下示例中，`->` 标记了包含从客户端发送到运行中 QEMU 实例的命令的行，而 `<-` 标记了包含 QEMU 返回的输出的行。

```
> sudo qemu-system-x86_64 [...] \  
-chardev stdio,id=mon0 \  
-mon chardev=mon0,mode=control,pretty=on  
  
<- {  
  "QMP": {  
    "version": {  
      "qemu": {  
        "micro": 0,  
        "minor": 0,  
        "major": 2  
      },  
      "package": ""  
    },  
    "capabilities": [  
    ]  
  }  
}
```


建立新的 QMP 连接后，QMP 将发送其问候消息并进入功能协商模式。在此模式下，只有 **qmp_capabilities** 命令能够正常运行。要退出功能协商模式并进入命令模式，必须先发出 **qmp_capabilities** 命令：

```
-> { "execute": "qmp_capabilities" }
<- {
  "return": {
  }
}
```

"return": {} 是 QMP 成功时返回的响应。

QMP 的命令可以附带参数。例如，要弹出 CD-ROM 驱动器，请输入以下命令：

```
->{ "execute": "eject", "arguments": { "device": "ide1-cd0" } }
<- {
  "timestamp": {
    "seconds": 1410353381,
    "microseconds": 763480
  },
  "event": "DEVICE_TRAY_MOVED",
  "data": {
    "device": "ide1-cd0",
    "tray-open": true
  }
}
{
  "return": {
  }
}
```

38.11.2 通过 telnet 访问 QMP

如果不使用标准输入/输出，您可将 QMP 接口连接到某个网络套接字，并通过特定的端口与其通讯：

```
> sudo qemu-system-x86_64 [...] \
-chardev socket,id=mon0,host=localhost,port=4444,server,nowait \
```

```
-mon chardev=mon0,mode=control,pretty=on
```

然后运行 telnet 连接到端口 4444:

```
> telnet localhost 4444
Trying ::1...
Connected to localhost.
Escape character is '^]'.
<- {
  "QMP": {
    "version": {
      "qemu": {
        "micro": 0,
        "minor": 0,
        "major": 2
      },
      "package": ""
    },
    "capabilities": [
    ]
  }
}
```

您可以同时创建多个监控器接口。下面的示例在标准输入/输出上创建了一个 HMP 实例（可识别“常规”QEMU 监控器命令的人工监控器），并在 localhost 端口 4444 上创建了一个 QMP 实例：

```
> sudo qemu-system-x86_64 [...] \
-chardev stdio,id=mon0 -mon chardev=mon0,mode=readline \
-chardev socket,id=mon1,host=localhost,port=4444,server,nowait \
-mon chardev=mon1,mode=control,pretty=on
```

38.11.3 通过 Unix 套接字访问 QMP

使用 `-qmp` 选项调用 QEMU，并创建一个 Unix 套接字：

```
> sudo qemu-system-x86_64 [...] \
-qmp unix:/tmp/qmp-sock,server --monitor stdio
```

```
QEMU waiting for connection on: unix:./qmp-sock,server
```

要通过 `/tmp/qmp-sock` 套接字来与 QEMU 实例通讯，请在同一主机上的另一个终端中使用 `nc`（有关详细信息，请参见 `man 1 nc`）：

```
> sudo nc -U /tmp/qmp-sock
<- {"QMP": {"version": {"qemu": {"micro": 0, "minor": 0, "major": 2} [...]
```

38.11.4 通过 libvirt 的 `virsh` 命令访问 QMP

如果您在 `libvirt` 下运行虚拟机（请参见第 II 部分“使用 `libvirt` 管理虚拟机”），则可以通过运行 `virsh qemu-monitor-command` 来与它的运行中 Guest 通讯：

```
> sudo virsh qemu-monitor-command vm_guest1 \
--pretty '{"execute":"query-kvm"}'
<- {
  "return": {
    "enabled": true,
    "present": true
  },
  "id": "libvirt-8"
}
```

在以上示例中，我们运行了简单的 `query-kvm` 命令来检查主机是否能够运行 KVM，以及是否启用了 KVM。



提示：生成直观易懂的输出

要使用 QEMU 的直观易懂的标准输出格式而不是 JSON 格式，请使用 `--hmp` 选项：

```
> sudo virsh qemu-monitor-command vm_guest1 --hmp "query-kvm"
```

VI 查错

- 39 集成式帮助和软件包文档 **381**
- 40 收集系统信息和日志 **382**

39 集成式帮助和软件包文档

虚拟化软件包提供了用于多方面管理虚拟化主机的命令。记住这些命令支持的所有选项是不切实际的，也不需要这样做。Xen 或 KVM 主机的基本安装包含外壳命令的手册页和集成式帮助。文档子软件包提供了基本安装所不包含的附加内容。

外壳命令的手册页

大多数命令都随附了手册页，其中提供了有关该命令的详细信息、介绍了所有选项，有些还提供了示例命令用法。例如，要查看 **virt-install** 命令的手册，请键入：

```
> man virt-install
```

外壳命令的集成式帮助

命令还随附了集成式帮助，其中提供了更简洁的主题导向型文档。例如，要查看 **virt-install** 命令的简要说明，请键入：

```
> virt-install --help
```

集成式帮助还可用于查看特定选项的细节。例如，要查看磁盘选项类型支持的子选项，请键入：

```
> virt-install --disk help
```

文档子软件包

许多虚拟化软件包在其文档子软件包中提供了附加内容。例如，**libvirt-doc** 软件包包含 <https://libvirt.org> 上提供的所有文档，以及用于演示 libvirt C API 用法的示例代码。使用 **rpm** 命令可查看文档子软件包的内容。例如，要查看 **libvirt-doc** 的内容，请键入：

```
rpm -ql libvirt-doc
```

40 收集系统信息和日志

当虚拟化主机遇到问题时，通常需要收集详细的系统报告。可以借助 **supportconfig** 工具来实现此目的。有关 **supportconfig** 的详细信息，请参见《管理指南》，第 47 章“收集系统信息以供支持所用”。

在某些情况下，**supportconfig** 收集的信息并不足够，可能还需要收集基于自定义日志记录或调试配置生成的日志来确定问题的原因。

40.1 libvirt 日志控制

libvirt 针对库和守护程序提供了日志记录工具。可以通过调整日志级别、过滤器和输出设置来控制日志记录工具的行为。

日志级别

libvirt 日志消息分为四种优先级：DEBUG、INFO、WARNING 和 ERROR。DEBUG 级别非常详细，短时间内生成的信息就能达到 GB 级别。日志消息的数量按照 INFO、WARNING 和 ERROR 日志级别的顺序逐渐减少。ERROR 是默认的日志级别。

日志过滤器

使用日志过滤器，您可以仅记录与特定组件和日志级别匹配的消息。日志过滤器允许收集特定组件的详细 DEBUG 日志消息，但只能从系统的其余组件收集 ERROR 级别的日志消息。默认情况下未定义日志过滤器。

日志输出

日志输出允许指定要将过滤的日志消息发送到的位置。可将消息发送到文件、进程的标准错误流或 **journald**。默认情况下，过滤的日志消息将发送到 **journald**。

有关 **libvirt** 的日志控制的更多细节，请参见 <https://libvirt.org/logging.html>。

默认的 **libvirt** 安装将日志级别设置为 ERROR，未定义日志过滤器，并将日志输出设置为 **journald**。可以使用 **journalctl** 命令查看来自 **libvirt** 守护程序的日志消息：

```
# journalctl --unit libvirtd
```

默认的日志工具设置适合常规操作，可为应用程序和 `libvirt` 用户提供有用的消息，但内部问题通常需要通过 `DEBUG` 级别的消息来解决。例如，假设 `libvirt` 与 `QEMU` 监控器之间的交互存在一个潜在的 bug。在这种情况下，我们只需查看 `libvirt` 与 `QEMU` 之间的通讯的调试消息。以下示例会创建一个日志过滤器，以选择来自 `QEMU` 驱动程序的调试消息并将其发送到名为 `/tmp/libvirtd.log` 的文件

```
log_filters="1:qemu.qemu_monitor_json"
log_outputs="1:file:/tmp/libvirtd.log"
```

可在 `/etc/libvirt/libvirtd.conf` 中找到 `libvirt` 守护程序的日志控制。对该配置文件进行任何更改后，必须重新启动该守护程序。

```
# systemctl restart libvirtd.service
```

词汇表

一般

Dom0

该术语在 Xen 环境中使用，表示一个虚拟机。主机操作系统是在特权域中运行的虚拟机，可称为 Dom0。主机上的所有其他虚拟机在非特权域中运行，可称为域 U。

KVM

请参见第 4 章 “KVM 虚拟化简介”

VHS

虚拟化主机服务器

运行 SUSE 虚拟化平台软件的物理计算机。虚拟化环境由超级管理程序、主机环境、虚拟机、关联的工具、命令和配置文件构成。其他常用术语包括主机、主机计算机、主机机器 (HM)、虚拟服务器 (VS)、虚拟机主机 (VMH) 和 VM 主机服务器 (VHS)。

VirtFS

VirtFS 是全新的半虚拟化文件系统界面，旨在改进 KVM 环境中的直通方法。它以 VirtIO 框架为基础构建。

Xen

请参见第 3 章 “Xen 虚拟化简介”

xl

适用于 Xen 的一组命令，可让管理员通过主机计算机上的命令提示符管理虚拟机。它取代了已弃用的 xm 工具堆栈。

主机环境

允许与主机计算机环境交互的桌面或命令行环境。它提供命令行环境，并且还可包含 GNOME 或 IceWM 等图形桌面。主机环境作为特殊类型的虚拟机运行，拥有控制和管理其他虚拟机的特权。其他常用术语包括 Dom0、特权域和主机操作系统。

创建虚拟机向导

YaST 和虚拟机管理器中提供的一个软件程序，它提供图形界面来引导您完成创建虚拟机的步骤。该软件程序也可在文本模式下运行，在主机环境中的命令提示符处输入 **virt-install** 即可进入文本模式。

半虚拟化帧缓冲区

通过一个内存缓冲区（其中包含以半虚拟模式运行的虚拟机显示器的完整数据帧）驱动视频显示器的视频输出设备。

硬件辅助

Intel* 和 AMD* 提供虚拟化硬件辅助技术。此技术降低了 VM 输入/输出的频率（VM 陷阱更少），同时由于软件是主要开销来源，此技术也提高了效率（执行由硬件完成）。此外，此技术减少了内存占用量，可提供更好的资源控制，并可确保安全地分配特定的 I/O 设备。

虚拟化

在虚拟机上运行的 Guest 操作系统或应用程序。

虚拟机

能够托管 Guest 操作系统和关联的应用程序的虚拟化 PC 环境 (VM)，也可称为 VM Guest。

虚拟机管理器

一个软件程序，提供用于创建和管理虚拟机的图形用户界面。

超级管理程序

用于协调虚拟机与底层物理计算机硬件之间的低级交互的软件。

CPU

CPU 固定

使用 CPU 固定（也称为处理器亲和性）可将某个进程或线程绑定到一个中央处理器 (CPU) 或某个范围的 CPU，或者将其取消绑定。

CPU 热插拔

CPU 热插拔用于描述在不关闭系统的情况下更换/添加/拆除 CPU 的功能。

CPU 过量分配

利用虚拟 CPU 过量分配，可以为 VM 分配超出物理系统中实际存在的物理 CPU 数量的虚拟 CPU 数量。此过程并不会提高系统的总体性能，但在进行测试时可能有用。

CPU 限制

虚拟 CPU 限制允许您将 vCPU 容量设置为物理 CPU 容量的 1%–100%。

网络

传统网桥

主机为其提供了物理网络设备和虚拟网络设备的一种网桥。

内部网络

将虚拟机局限于其主机环境的一种网络配置。

外部网络

主机内部网络环境之外的网络。

无主机网桥

主机为其提供了物理网络设备但未提供虚拟网络设备的一种网桥。此网桥可让虚拟机在外部网络上通讯，但不能与主机通讯。这样，您便可以将虚拟机网络通讯与主机环境相隔离。

本地网桥

主机为其提供了虚拟网络设备但未提供物理网络设备的一种网桥。此网桥可让虚拟机与主机以及主机上的其他虚拟机通讯。虚拟机之间可以通过主机在外部网络上通讯。

桥接网络

这种网络连接允许在外部网络上将虚拟机标识为与其主机计算机相互独立且不相关的唯一身份。

空网桥

主机未为其提供任何物理网络设备或虚拟网络设备的一种网桥。此网桥可让虚拟机与同一主机上的其他虚拟机通讯，但不能与主机或在外部网络上通讯。

网络地址转换 (NAT)

允许虚拟机使用主机的 IP 地址和 MAC 地址的一种网络连接。

存储

AHCI

高级主机控制器接口 (AHCI) 是 Intel* 定义的一套技术标准，指定如何以不特定于实现的方式操作串行 ATA (SATA) 主机总线适配器。

xvda

分配给半虚拟计算机上的第一个虚拟磁盘的驱动器号。

原始磁盘

在单个字节级别而不是通过磁盘文件系统访问磁盘中的数据的一种方法。

块设备

以块的形式移动数据的数据存储设备，例如 CD-ROM 驱动器或磁盘驱动器。分区和卷也被视为块设备。

基于文件的虚拟磁盘

基于文件的虚拟磁盘，也称为磁盘映像文件。

稀疏映像文件

不保留其整个磁盘空间量，而是随着在其中写入数据而不断扩展的磁盘映像文件。

缩写词

ACPI

高级配置和电源接口 (ACPI) 规范为操作系统的设备配置和电源管理提供了开放的标准。

AER

高级错误报告

AER 是 PCI Express 规范提供的一项功能，用于报告 PCI 错误并在发生其中某些错误后进行恢复。

APIC

高级可编程中断控制器 (APIC) 是一个中断控制器系列。

BDF

总线:设备:功能

用于简要描述 PCI 和 PCIe 设备的表示法。

CG

控制组

用于限制、统计和隔离资源（CPU、内存、磁盘 I/O 等）使用量的功能。

EDF

最早截止期限优先

此调度程序以直观方式提供加权 CPU 共享，并使用实时算法来提供时间上的保障。

EPT

扩展页表

虚拟化环境中的性能与本机环境非常接近。不过，虚拟化确实会产生一定的开销。这些开销源自 CPU、MMU 和 I/O 设备的虚拟化。在一些新款 x86 处理器中，AMD 和 Intel 已开始提供硬件扩展来帮助弥补这种性能差距。2006 年，这两家供应商推出了采用 AMD-Virtualization (AMD-V) 和 Intel® VT-x 技术的第一代 x86 虚拟化硬件支持。最近，Intel 推出了其第二代硬件支持，其中整合了称作扩展页表 (EPT) 的 MMU 虚拟化。与使用影子分页技术进行 MMU 虚拟化相比，支持 EPT 的系统可以提升性能。如果工作负载不多，EPT 会增加内存访问延迟，这个代价可以通过在 Guest 和超级管理程序中有效使用大页来降低。

FLASK

Flux 高级安全内核

Xen 通过一个称为 FLASK 的安全体系结构，使用同名的模块来实现某种类型的强制访问控制。

HAP

高保证平台

HAP 结合了硬件和软件技术来提高工作站与网络安全性。

HVM

硬件虚拟机 (Xen 通常这样称呼)。

IOMMU

输入/输出内存管理单元

IOMMU (AMD* 技术) 是内存管理单元 (MMU)，可将支持直接内存访问 (DMA) 的 I/O 总线连接到主内存。

KSM

内核同页合并

KSM 可用于在 Guest 之间自动共享相同的内存页，以节省主机内存。KVM 经优化后可以使用 KSM (如果已在 VM 主机服务器上启用)。

MMU

内存管理单元

一个计算机硬件组件，负责处理 CPU 的内存访问请求。其功能包括虚拟地址到物理地址的转换 (即虚拟内存管理)、内存保护、缓存控制、总线仲裁，在较为简单的计算机体系结构 (尤其是 8 位系统) 中，负责内存库切换。

PAE

物理地址扩展

32 位 x86 操作系统使用物理地址扩展 (PAE) 模式来实现 4 GB 以上物理内存的寻址。在 PAE 模式下，页表项 (PTE) 的大小为 64 位。

PCID

进程环境标识符

逻辑处理器可通过这些标识符缓存多个线性地址空间的信息，这样当软件切换到其他线性地址空间时，处理器能够保留缓存的信息。INVPID 指令用于进行精细的 TLB 刷新，这对于内核会有所助益。

PCIe

高速外设组件互连

PCIe 用来替代旧式 PCI、PCI-X 和 AGP 总线标准。PCIe 融入了大量改进，包括更高的最大系统总线吞吐量、更低的 I/O 脚数，以及更小的物理占用空间。此外，它还提供更详细的错误检测和报告机制 (AER)，以及本机热插拔功能。它还向后兼容 PCI。

PSE 和 PSE36

扩展页大小

PSE 是指 x86 处理器的一项功能，允许页大于传统的 4 KiB 大小。PSE-36 功能在普通 10 位的基础上再额外提供 4 位，这些额外的位在指向大页的页目录项中使用。这样，便可以将大页放到 36 位地址空间中。

PT

页表

页表是计算机操作系统中的虚拟内存系统用来存储虚拟地址与物理地址之间的映射的数据结构。虚拟地址是访问进程特有的地址。物理地址是硬件 (RAM) 特有的地址。

QXL

QXL 是虚拟化环境的 cirrus VGA 帧缓冲 (8M) 驱动程序。

RVI 或 NPT

快速虚拟化索引、嵌套式页表

适用于处理器内存管理单元 (MMU) 的 AMD 第二代硬件辅助虚拟化技术。

SATA

串行 ATA

SATA 是一个计算机总线接口，可将主机总线适配器连接到硬盘和光驱等大容量存储设备。

SMEP

监督模式执行保护

用于防止 Xen 超级管理程序执行用户模式页，使通过应用程序恶意利用超级管理程序的许多企图更难以实现。

SPICE

适用于独立计算环境的简单协议

SXP

SXP 文件是 Xen 配置文件。

TCG

微代码生成器

模拟指令，而不是由 CPU 执行指令。

THP

透明大页

此功能可让 CPU 使用大于默认 4 KB 的页进行内存寻址。这有助于减少内存消耗量和 CPU 缓存用量。KVM 经过优化，可以通过 `madvise` 和机会性方法使用 THP（如果已在 VM 主机服务器上启用）。

TLB

转换后援缓冲区

TLB 是内存管理硬件用来提升虚拟地址转换速度的缓存。所有最新的台式机、笔记本电脑和服务器处理器都使用 TLB 来映射虚拟和物理地址空间，在任何使用虚拟内存的硬件中，几乎都会用到 TLB。

VCPU

一个调度实体，包含虚拟化 CPU 的每种状态。

VDI

虚拟桌面基础结构

VFIO

从内核 v3.6 开始，推出了一种从用户空间访问 PCI 设备的新方法，该方法称为 VFIO。

VHS

虚拟化主机服务器

VM root

VM 以 VMX root 操作运行，Guest 软件以 VMX 非 root 操作运行。VMX root 操作与 VMX 非 root 操作之间的转换称为 VMX 转换。

VMCS

虚拟机控制结构

VMX 非 root 操作和 VMX 转换由一个称为虚拟机控制结构 (VMCS) 的数据结构来控制。VMCS 访问通过称为 VMCS 指针（每个逻辑处理器一个）的处理器状态组件来管理。VMCS 指针的值是 VMCS 的 64 位地址。通过 VMPTRST 和 VMPTRLD 指令来读取和写入 VMCS 指针。VMM 使用 VMREAD、VMWRITE 和 VMCLEAR 指令来配置 VMCS。VMM 可对它支持的每个虚拟机使用不同的 VMCS。如果虚拟机具有多个逻辑处理器（虚拟处理器），VMM 可对每个虚拟处理器使用不同的 VMCS。

VMDq

虚拟机设备队列

多队列网络适配器可在硬件级别支持多个 VM，它们能使不同的包队列关联到不同的托管 VM（通过 VM 的 IP 地址进行关联）。

VMM

虚拟机监控器（超级管理程序）

当处理器遇到与超级管理程序 (VMM) 相关的指令或事件时，它会退出 Guest 模式并回到 VMM。VMM 以远低于本机的速度模拟该指令或另一事件，然后返回到 Guest 模式。Guest 模式与 VMM 之间的来回转换属于高延迟操作，在此期间，Guest 执行会完全停滞。

VMX

虚拟机扩展

VPID

新的 TLB 软件控制支持（利用 VPID，您只需进行少量的 VMM 开发，就能提升 TLB 性能）。

VT-d

定向 I/O 虚拟化技术

类似于 Intel* (<https://software.intel.com/en-us/articles/intel-virtualization-technology-for-directed-io-vt-d-enhancing-intel-platforms-for-efficient-virtualization-of-io-devices>) 的 IOMMU。

vTPM

用于通过可信计算为 Guest 建立端到端完整性的组件。

基于 Seccomp2 的沙箱

为了增强恶意行为防范能力而只允许使用预先确定的系统调用的沙箱环境。

A 虚拟机驱动程序

利用虚拟化，您可以将工作负载整合到更新、更强大且更节能的硬件上。SUSE® Linux Enterprise Server 等半虚拟化操作系统及其他 Linux 发行套件能够识别底层虚拟化平台，因此可以有效地与它交互。Microsoft Windows* 等未经修改的操作系统无法识别虚拟化平台，需要直接与硬件交互。由于在整合服务器时无法做到这一点，因此必须为操作系统模拟硬件。模拟速度可能很慢，而且在模拟高吞吐量磁盘和网络子系统时尤为麻烦。大部分性能损失都是发生在这一环节。

SUSE Linux Enterprise 虚拟机驱动程序包 (VMDP) 包含适用于多种 Microsoft Windows 操作系统的 32 位和 64 位半虚拟化网络、总线与块驱动程序。这些驱动程序为未经修改的操作系统带来了半虚拟化操作系统具备的诸多性能优势：只有半虚拟化设备驱动程序能够感知到虚拟化平台（其他操作系统都不可以）。例如，对于操作系统而言，半虚拟化磁盘设备驱动程序就像是一个正常的物理磁盘，而设备驱动程序是直接与虚拟化平台交互的（无需模拟）。这有助于有效实现磁盘访问，使磁盘和网络子系统能够在虚拟化环境中以接近本机的速度运行，且无需对现有操作系统进行更改。

SUSE® Linux Enterprise 虚拟机驱动程序包作为 SUSE Linux Enterprise Server 的附加产品提供。有关详细信息，请参见 <https://www.suse.com/products/vmdriverpack/>。

有关详细信息，请参见官方 VMDP Installation Guide (<https://documentation.suse.com/sle-vmdp/2.5/html/vmdp/index.html>)。

B 为 NVIDIA 卡配置 GPU 直通

B1 简介

本文介绍如何将主机计算机上的 NVIDIA GPU 显卡指派给虚拟化 Guest。

B2 先决条件

- 只有 AMD64/Intel 64 体系结构支持 GPU 直通。
- 主机操作系统需是 SLES 12 SP3 或更高版本。
- 本文的内容涉及一组基于 V100/T1000 NVIDIA 卡的指令，仅与 GPU 计算相关。
- 请校验您使用的是否为 NVIDIA Tesla 产品 — Maxwell、Pascal 或 Volta。
- 要管理主机系统，您需要在主机上额外安装一块显卡，以便可以在配置 GPU 直通或 SSH 功能环境时使用它。

B3 配置主机

B3.1 校验主机环境

1. 校验主机操作系统是否为 SLES 12 SP3 或更高版本：

```
> cat /etc/issue
Welcome to SUSE Linux Enterprise Server 15 (x86_64) - Kernel \r (\l).
```

2. 校验主机是否支持 VT-d 技术，并且已在固件设置中启用该技术：

```
> dmesg | grep -e "Directed I/O"
[ 12.819760] DMAR: Intel(R) Virtualization Technology for Directed I/O
```

如果未在固件中启用 VT-d，请启用它并重引导主机。

3. 校验主机是否有额外的 GPU 或 VGA 卡：

```
> lspci | grep -i "vga"
07:00.0 VGA compatible controller: Matrox Electronics Systems Ltd. \
MGA G200e [Pilot] ServerEngines (SEP1) (rev 05)
```

对于 Tesla V100 卡：

```
> lspci | grep -i nvidia
03:00.0 3D controller: NVIDIA Corporation GV100 [Tesla V100 PCIe] (rev a1)
```

对于 T1000 Mobile（可在 Dell 5540 上使用）：

```
> lspci | grep -i nvidia
01:00.0 3D controller: NVIDIA Corporation TU117GLM [Quadro T1000 Mobile]
(rev a1)
```

B3.2 启用 IOMMU

默认已禁用 IOMMU。您需要在系统引导时于 /etc/default/grub 配置文件中启用它。

1. 对于基于 Intel 的主机：

```
GRUB_CMDLINE_LINUX="intel_iommu=on iommu=pt rd.driver.pre=vfio-pci"
```

对于基于 AMD 的主机：

```
GRUB_CMDLINE_LINUX="iommu=pt amd_iommu=on rd.driver.pre=vfio-pci"
```

2. 保存修改后的 /etc/default/grub 文件时，请重新生成主 GRUB 2 配置文件 /boot/grub2/grub.cfg：

```
> sudo grub2-mkconfig -o /boot/grub2/grub.cfg
```

3. 重引导主机并校验是否已启用 IOMMU：

```
> dmesg | grep -e DMAR -e IOMMU
```

B3.3 将 Nouveau 驱动程序加入黑名单

要将 NVIDIA 卡分配给 VM Guest，我们需要防止主机操作系统加载 NVIDIA GPU 的内置 `nouveau` 驱动程序。创建包含以下内容的 `/etc/modprobe.d/60-blacklist-nouveau.conf` 文件：

```
blacklist nouveau
```

B3.4 配置 VFIO 并隔离用于直通的 GPU

1. 查找卡供应商和型号 ID。使用第 B3.1 节 “校验主机环境” 中列出的总线编号（例如 `03:00.0`）进行查找：

```
> lspci -nn | grep 03:00.0
03:00.0 3D controller [0302]: NVIDIA Corporation GV100 [Tesla V100 PCIe]
[10de:1db4] (rev a1)
```

2. 创建包含以下内容的 `/etc/modprobe.d/vfio.conf` 文件：

```
options vfio-pci ids=10de:1db4
```



注意

校验您的卡是否不需要额外的 `ids=` 参数。对于某些卡，还必须指定音频设备，因此还必须将该设备的 ID 添加到列表中，否则无法使用该卡。

B3.5 加载 VFIO 驱动程序

可通过三种方式加载 `VFIO` 驱动程序。

B3.5.1 在 initrd 文件中包含该驱动程序

1. 创建 `/etc/dracut.conf.d/gpu-passthrough.conf` 文件并在其中添加以下内容（请注意前导空格）：

```
add_drivers+=" vfio vfio_iommu_type1 vfio_pci vfio_virqfd"
```

2. 重新生成 initrd 文件：

```
> sudo dracut --force /boot/initrd $(uname -r)
```

B3.5.2 将该驱动程序添加到自动加载的模块的列表

创建 `/etc/modules-load.d/vfio-pci.conf` 文件并在其中添加以下内容：

```
vfio
vfio_iommu_type1
vfio_pci
kvm
kvm_intel
```

B3.5.3 手动加载该驱动程序

要在运行时手动加载该驱动程序，请执行以下命令：

```
> sudo modprobe vfio-pci
```

B3.6 为 Microsoft Windows Guest 禁用 MSR

对于 Microsoft Windows Guest，我们建议禁用 MSR（特定于模型的寄存器），以避免 Guest 崩溃。创建 `/etc/modprobe.d/kvm.conf` 文件并在其中添加以下内容：

```
options kvm ignore_msrs=1
```

B3.7 安装 UEFI 固件

主机需使用 UEFI 固件进行引导（即，不使用旧式 BIOS 引导序列）才能使 GPU 直通功能正常工作。安装 `qemu-ovmf` 软件包（如果尚未安装）：

```
> sudo zypper install qemu-ovmf
```

B3.8 重引导主机计算机

要使上述步骤中的大部分更改生效，需要重引导主机计算机：

```
> sudo shutdown -r now
```

B4 配置 Guest

本节介绍如何配置 Guest 虚拟机，以使其能够使用主机的 NVIDIA GPU。使用虚拟机管理器或 `virt-install` 安装 Guest VM。有关详细信息，请参见 [第 10 章 “Guest 安装”](#)。

B4.1 Guest 配置要求

在安装 Guest VM 期间，选择在安装之前自定义配置并配置以下设备：

- 如果可能，请使用 Q35 芯片组。
- 使用 UEFI 固件安装 Guest VM。
- 添加以下模拟设备：
图形：Spice 或 VNC
设备：qxl、VGA 或 Virtio
有关详细信息，请参见 [第 14.6 节 “视频”](#)。
- 将主机 PCI 设备（在本示例中为 `03:00.0`）添加到 Guest。有关详细信息，请参见 [第 14.12 节 “将主机 PCI 设备分配到 VM Guest”](#)。
- 为获得最佳性能，我们建议对网卡和存储设备使用 virtio 驱动程序。

B4.2 安装显卡驱动程序

B4.2.1 Linux Guest

过程 B1：基于 RPM 的发行套件

1. 从 <https://www.nvidia.com/download/driverResults.aspx/131159/en-us> 下载驱动程序 RPM 软件包。

2. 安装下载的 RPM 软件包：

```
> sudo rpm -i nvidia-diag-driver-local-repo-sles123-390.30-1.0-1.x86_64.rpm
```

3. 刷新储存库并安装 `cuda-drivers`。对于非 SUSE 发行套件，此步骤有所不同：

```
> sudo zypper refresh && zypper install cuda-drivers
```

4. 重引导 Guest VM：

```
> sudo shutdown -r now
```

过程 B2：通用安装程序

1. 由于安装程序需要编译 NVIDIA 驱动程序模块，因此请安装 `gcc-c++` 和 `kernel-devel` 软件包。
2. 在 Guest 上禁用安全引导，因为 NVIDIA 的驱动程序模块未签名。在 SUSE 发行套件上，可以使用 YaST GRUB 2 模块来禁用安全引导。有关详细信息，请参见《管理指南》，第 17 章“UEFI（统一可扩展固件接口）”，第 17.1.1 节“在 SUSE Linux Enterprise Server 上实施”。
3. 从 <https://www.nvidia.com/Download/index.aspx?lang=en-us> 下载驱动程序安装脚本，将此脚本转换为可执行文件，然后运行此可执行文件以完成驱动程序安装：

```
> chmod +x NVIDIA-Linux-x86_64-460.73.01.run  
> sudo ./NVIDIA-Linux-x86_64-460.73.01.run
```

4. 从 [https://developer.nvidia.com/cuda-downloads?](https://developer.nvidia.com/cuda-downloads?target_os=Linux&target_arch=x86_64&target_distro=SLES&target_version=15&target_type=rpm)

`target_os=Linux&target_arch=x86_64&target_distro=SLES&target_version=15&target_type=rpm`

下载 CUDA 驱动程序，并按照屏幕上的说明进行安装。



注意：显示器问题

安装 NVIDIA 驱动程序后，虚拟机管理器显示器将与 Guest 操作系统断开连接。要访问 Guest VM，您必须通过 **ssh** 登录，然后切换到控制台界面，或者在 Guest 中安装专用 VNC 服务器。为避免屏幕闪烁，请停止并禁用显示器管理器：

```
> sudo systemctl stop display-manager && systemctl disable display-manager
```

过程 B3：测试 LINUX 驱动程序安装

1. 将目录切换到 CUDA 示例模板：

```
> cd /usr/local/cuda-9.1/samples/0_Simple/simpleTemplates
```

2. 编译并运行 `simpleTemplates` 文件：

```
> make && ./simpleTemplates
runTest<float,32>
GPU Device 0: "Tesla V100-PCIE-16GB" with compute capability 7.0
CUDA device [Tesla V100-PCIE-16GB] has 80 Multi-Processors
Processing time: 495.006000 (ms)
Compare OK
runTest<int,64>
GPU Device 0: "Tesla V100-PCIE-16GB" with compute capability 7.0
CUDA device [Tesla V100-PCIE-16GB] has 80 Multi-Processors
Processing time: 0.203000 (ms)
Compare OK
[simpleTemplates] -> Test Results: 0 Failures
```


B4.2.2 Microsoft Windows Guest

! 重要

在安装 NVIDIA 驱动程序之前，需要使用 Guest `libvirt` 定义中的 `<hidden state='on' />` 指令对驱动程序隐藏超级管理程序，例如：

```
<features>
  <acpi/>
  <apic/>
  <kvm>
    <hidden state='on' />
  </kvm>
</features>
```

1. 从 <https://www.nvidia.com/Download/index.aspx> 下载并安装 NVIDIA 驱动程序。
2. 从 https://developer.nvidia.com/cuda-downloads?target_os=Windows&target_arch=x86_64 下载并安装 CUDA 工具包。
3. 在 Guest 上的 `Program Files\Nvidia GPU Computing Toolkit\CUDA\v10.2\extras\demo_suite` 目录中可以找到多个 NVIDIA 演示示例。

C XM、XL 工具栈和 libvirt 框架

C1 Xen 工具栈

从早期发行版 Xen 2.x 开始，**xend** 就一直是事实上用于管理 Xen 安装的工具堆栈。Xen 4.1 中引入了一个处于技术预览状态的新工具堆栈 libxenlight（又称为 libxl）。libxl 是以 C 语言编写的小型低级别库，旨在为所有客户端工具堆栈（**XAPI** (<https://wiki.xen.org/wiki/XAPI>)、**libvirt**、**xl**) 提供简单的 API。Xen 4.2 中已将 libxl 提升为受支持状态，并将 **xend** 标记为弃用。Xen 4.3 和 4.4 系列中包含了 **xend**，使用户有充足的时间将其工具过渡到 libxl。从 Xen 4.5 系列和 SUSE Linux Enterprise Server 12 SP1 开始，xend 已从上游 Xen 项目中去除，不再提供。

尽管 SLES 11 SP3 包含了 Xen 4.2，但 SUSE 仍保留了 **xend** 工具堆栈，因为在服务包中进行这种有创性更改会给 SUSE Linux Enterprise 客户造成过大干扰。不过，SLES 12 将提供适当的机会让客户迁移到新的 libxl 工具堆栈，并去除已弃用且不再保留的 **xend** 堆栈。从 SUSE Linux Enterprise Server 12 SP1 开始，**xend** 不再受支持。

xend 与 libxl 之间的主要差别之一是，前者是有状态的，而后者是无状态的。使用 **xend** 时，所有客户端应用程序（例如 **xm** 和 **libvirt**）都会看到相同的系统状态。**xend** 负责维护整个 Xen 主机的状态。在 libxl 中，**xl** 或 **libvirt** 等客户端应用程序必须维护状态。因此，使用 **xl** 创建的域对于 **libvirt** 等其他 libxl 应用程序是不可见或不可知的。一般情况下，我们建议不要混用多个 libxl 应用程序，而是使用单个 libxl 应用程序来管理 Xen 主机。在 SUSE Linux Enterprise Server 中，我们建议使用 **libvirt** 来管理 Xen 主机。这样，便可以通过 **libvirt** 应用程序（如 **virt-manager**、**virt-install**、**virt-viewer**、**libguestfs** 等）来管理 Xen 系统。如果使用 **xl** 管理 Xen 主机，**libvirt** 将无法访问由 **xl** 管理的任何虚拟机。因此，任何 **libvirt** 应用程序也无法访问这些虚拟机。

C1.1 从 xend/xm to xl/libxl 升级

xl 应用程序及其配置格式（请参见 `man xl.cfg`）可以向后兼容 **xm** 应用程序及其配置格式（请参见 `man xm.cfg`）。可以通过 **xl** 来利用现有的 **xm** 配置。由于 libxl 是无状态的，并且 **xl** 不支持受管域的表达法，因此 SUSE 建议使用 **libvirt** 来管理 Xen 主机。SUSE 提供了一个名为 **xen2libvirt** 的工具，用于提供简单的机制来将以前由 **xend** 管理的域导入 **libvirt**。有关 **xen2libvirt** 的详细信息，请参见第 C2 节“将 Xen 域配置导入 libvirt”。

C1.2 XL 设计

每个 **xl** 命令的基本结构如下：

```
xl subcommand OPTIONS DOMAIN
```

DOMAIN 是域 ID 编号或者域名（在内部转换为域 ID），**OPTIONS** 是特定于子命令的选项。

尽管 xl/libxl 可以向后兼容 xm/xend，但您应该注意两者之间的几项差别：

- 受管域或持久域。libvirt 现在提供此项功能。
- xl/libxl 不支持在域配置文件中使用的 Python 代码。
- xl/libxl 不支持基于 SXP 格式配置文件创建域 (`xm create -F`)。
- xl/libxl 不支持通过域配置文件中的 **w!** 在 DomU 之间共享存储。

xl/libxl 是尚在大力开发中的新工具堆栈，因此相比 xm/xend 工具堆栈仍然缺少一些功能：

- SCSI LUN/主机直通 (PVSCSI)
- USB 直通 (PVUSB)
- Xen 全虚拟化 Linux Guest 的直接内核引导

C1.3 升级前的核对清单

在将 SLES 11 SP4 Xen 主机升级到 SLES 15 之前：

- 必须从 `xm` 域配置文件中去除任何 Python 代码。
- 建议使用 `virsh dumpxml DOMAIN_NAME DOMAIN_NAME.xml` 捕获所有现有虚拟机中的 libvirt 域 XML。
- 建议备份 `/etc/xen/xend-config.sxp` 和 `/boot/grub/menu.lst` 文件，以保留以前用于 Xen 的参数的参考信息。



注意

目前不支持将 SLES 11 SP4 Xen 主机上运行的虚拟机实时迁移到 SLES 15 Xen 主机。`xend` 与 libxl 工具堆栈的运行时环境不兼容。需要关闭虚拟机才能进行迁移。

C2 将 Xen 域配置导入 libvirt

`xen2libvirt` 是用于将旧式 Xen 域配置导入 `libvirt` 虚拟化库的命令行工具。有关 `libvirt` 的详细信息，请参见《The Virtualization》（虚拟化）一书。使用 `xen2libvirt` 可以轻松将已弃用的 `xm/xend` 工具堆栈所管理的域导入新的 `libvirt/libxl` 工具堆栈中。使用此工具的 `--recursive mode` 模式可以一次性导入多个域

`xen2libvirt` 包含在 `xen-tools` 软件包中。如果需要，请使用以下命令安装该软件包

```
> sudo zypper install xen-tools
```

`xen2libvirt` 的一般语法为

```
xen2libvirt <options> /path/to/domain/config
```

其中，`options` 可以是：

`-h, --help`

列显有关 `xen2libvirt` 用法的简短信息。

`-c, --convert-only`

将域配置转换为 `libvirt` XML 格式，但不将配置导入 `libvirt`。

-r, --recursive

从指定的路径开始，以递归方式转换并/或导入所有域配置。

-f, --format

指定源域配置的格式。可以是 xm 或 sexpr（S 表达式格式）。

-v, --verbose

列显有关导入过程的更详细的信息。

例 C1：将 XEN 域配置转换为 libvirt

假设您有一个通过 xm 管理的 Xen 域，/etc/xen/sle12.xm 中保存了该域的以下配置：

```
kernel = "/boot/vmlinuz-2.6-xenU"
memory = 128
name = "SLE12"
root = "/dev/hda1 ro"
disk = [ "file:/var/xen/sle12.img,hda1,w" ]
```

将此配置转换为 libvirt XML 而不导入，然后查看其内容：

```
> sudo xen2libvirt -f xm -c /etc/xen/sle12.xm > /etc/libvirt/qemu/
sles12.xml
# cat /etc/libvirt/qemu/sles12.xml
<domain type='xen'>
<name>SLE12</name>
<uuid>43e1863c-8116-469c-a253-83d8be09aa1d</uuid>
<memory unit='KiB'>131072</memory>
<currentMemory unit='KiB'>131072</currentMemory>
<vcpu placement='static'>1</vcpu>
<os>
<type arch='x86_64' machine='xenpv'>linux</type>
<kernel>/boot/vmlinuz-2.6-xenU</kernel>
</os>
<clock offset='utc' adjustment='reset'/>
<on_poweroff>destroy</on_poweroff>
<on_reboot>restart</on_reboot>
<on_crash>restart</on_crash>
<devices>
<disk type='file' device='disk'>
```

```
<driver name='file' />
<source file='/var/xen/sle12.img' />
<target dev='hda1' bus='xen' />
</disk>
<console type='pty'>
<target type='xen' port='0' />
</console>
</devices>
</domain>
```

要将域导入 `libvirt`，可以运行不带 `-c` 选项的相同 `xen2libvirt` 命令，或者使用导出的文件 `/etc/libvirt/qemu/sles12.xml`，并通过 `virsh` 定义新的 Xen 域：

```
> sudo virsh define /etc/libvirt/qemu/sles12.xml
```

C3 xm 与 xl 应用程序之间的差异

本章罗列了 `xm` 与 `xl` 应用程序之间的所有差异。一般情况下，`xl` 与 `xm` 兼容。通常，用户只需在自定义脚本或工具中将 `xm` 替换为 `xl` 即可。

您还可以通过 `virsh` 命令使用 `libvirt` 框架。本文档只会显示 `virsh` 的第一个 `OPTION`。要获取有关此选项的更多帮助，请执行：

```
virsh help OPTION
```

C3.1 表示法约定

为了让您轻松了解 `xl` 与 `xm` 命令之间的差异，本节使用了以下表示法：

表 C1：表示法约定

表示法	含义
(-) minus	选项在 <code>xm</code> 中存在，但未包含在 <code>xl</code> 中。
(+) plus	选项在 <code>xl</code> 中存在，但未包含在 <code>xm</code> 中。

C3.2 新的全局选项

表 C2：新的全局选项

选项	任务
(+) <u>-v</u>	详细，提高输出的详细程度
(+) <u>-N</u>	试运行，不实际执行命令
(+) <u>-f</u>	强制执行。如果 <u>xl</u> 检测到 <u>xend</u> 也在运行，将拒绝运行某些命令。此选项会强制执行这些命令，即使这样做不安全

C3.3 未更改的选项

xl 和 xm 的常用选项列表，及其等效的 libvirt 选项。

表 C3：通用选项

选项	任务	<u>libvirt</u> 等效选项
destroy <u>DOMAIN</u>	立即终止域。	<u>virsh</u> destroy
domid <u>DOMAIN_NAME</u>	将域名转换为 <u>DOMAIN_ID</u> 。	<u>virsh</u> domid
domname <u>DOMAIN_ID</u>	将 <u>DOMAIN_ID</u> 转换为 <u>DOMAIN_NAME</u> 。	<u>virsh</u> domname
help	显示简短的帮助消息（即常用命令）。	<u>virsh</u> help
pause <u>DOMAIN_ID</u>	暂停域。处于暂停状态的域仍会消耗分配的资源（例如内存），但不符合由 Xen 超级管理程序调度的条件。	<u>virsh</u> suspend

选项	任务	libvirt 等效选项
<code>unpause</code> <u>DOMAIN_ID</u>	使域脱离暂停状态。这样以前暂停的域便符合由 Xen 超级管理程序调度的条件。	<code>virsh</code> <code>resume</code>
重命名 <u>DOMAIN_ID</u> <u>NEW_DOMAIN_NAME</u>	将 <u>DOMAIN_ID</u> 的域名更改为 <u>NEW_DOMAIN_NAME</u> 。	<pre>> virsh dumpxml DOMAINNAME > DOMXML</pre> - 修改 <u>DOMXML</u> 中的域名 <pre>> virsh undefine DOMAINNAME</pre> <pre>> virsh define DOMAINNAME</pre>
<code>sysrq</code> <u>DOMAIN</u> <letter>	向域发送魔法系统请求，每种类型的请求由一个不同的字母表示。可以使用此选项向 Linux Guest 发送 SysRq 请求，有关详细信息，请参见 https://www.kernel.org/doc/html/latest/admin-guide/sysrq.html 。需要在 Guest 操作系统中安装 PV 驱动程序。	<code>virsh</code> <code>send-keys</code> 只能针对 KVM 发送魔法系统请求
<code>vncviewer</code> <u>OPTIONS</u> <u>DOMAIN</u>	挂接到域的 VNC 服务器，并派生 <code>vncviewer</code> 进程。	<code>virt-viewer</code> <u>DOMAIN_ID</u> <code>virsh</code> <u>VNCDISPLAY</u>

选项	任务	libvirt 等效选项
<u>vcpu-set</u> <u>DOMAIN_ID</u> <u>VCPUS</u>	为相关的域设置虚拟 CPU 数量。与 <u>mem-set</u> 一样，此命令最多只能分配引导时为域配置的最大虚拟 CPU 数量。	<u>virsh</u> <u>setvcpus</u>
<u>vcpu-list</u> <u>DOMAIN_ID</u>	列出特定域的 VCPU 信息。如果未指定域，将提供所有域的 VCPU 信息。	<u>virsh</u> <u>vcpuinfo</u>
<u>vcpu-pin</u> <u>DOMAIN_ID</u> <VCPU all> <CPUs all>	固定 VCPU，使其仅在特定的 CPU 上运行。可以使用关键字 all 向域中的所有 VCPU 应用 CPU 列表。	<u>virsh</u> <u>vcupin</u>
<u>dmesg</u> [-c]	读取 Xen 消息缓冲区，与 Linux 系统上的 dmesg 类似。该缓冲区包含 Xen 引导过程中创建的信息性、警告和错误消息。	
<u>top</u>	执行 <u>xentop</u> 命令，该命令提供域的实时监控。 <u>xentop</u> 是一个 curses 接口。	<u>virsh</u> <u>nodecpustats</u> <u>virsh</u> <u>nodememstats</u>
<u>uptime</u> [-s] <u>DOMAIN</u>	列显正在运行的域的当前运行时间。使用 <u>xl</u> 命令时，必须指定 <u>DOMAIN</u> 参数。	
<u>debug-keys</u> <u>KEYS</u>	将调试键发送到 Xen。等同于按 Xen <u>conswitch</u> （默认为 Ctrl-A）三次，然后按“keys”。	

选项	任务	libvirt 等效选项
<u>cpupool-migrate</u> <u>DOMAIN</u> <u>CPU_POOL</u>	将 <u>DOMAIN_ID</u> 或 <u>DOMAIN</u> 指定的域移到 <u>CPU_POOL</u> 中。	
<u>cpupool-destroy</u> <u>CPU_POOL</u>	停用 CPU 池。仅当该 CPU 池中没有任何处于活动状态的域时，才可执行此操作。	
<u>block-detach</u> <u>DOMAIN_ID</u> <u>DevId</u>	分离域的虚拟块设备。 devId 可以是 Dom0 为设备指定的符号名称或设备数字 ID。需要运行 xl block-list 来确定该编号。	<u>virsh detach-disk</u>
<u>network-attach</u> <u>DOMAIN_ID</u> <u>NETWORK_DEVICE</u>	在 <u>DOMAIN_ID</u> 所指定的域中创建新网络设备。 network_device 描述要挂接的设备，使用与域配置文件中 vif 字符串相同的格式	<u>virsh attach-interface</u> <u>virsh attach-device</u>
<u>pci-attach</u> <u>DOMAIN</u> <BDF> [Virtual Slot]	将新的直通 PCI 设备热插入到指定的域。 BDF 是要直通的物理设备的 PCI 总线/设备/功能。	<u>virsh attach-device</u>
<u>pci-list</u> <u>DOMAIN_ID</u>	列出域的直通 PCI 设备	
<u>getenforce</u>	确定 FLASK 安全模块是否已加载并在强制实施其策略。	
<u>setenforce</u> <1 0 <u>Enforcing Permissive</u> >	启用或禁用强制实施 FLASK 访问控制的功能。默认值为 permissive，您可以在超	

选项	任务	libvirt 等效选项
	级管理程序的命令行上使用 flask_enforcing 选项更改默认值。	

C3.4 已去除的选项

不再可用于 XL 工具堆栈的 `xm options` 列表，以及替代的解决方法（如果有）。

C3.4.1 域管理

已去除的域管理命令及其替代命令列表。

表 C4：已去除的域管理选项

已去除的域管理选项		
选项	任务	等效选项
(-) <code>log</code>	列显 Xend 日志。	可在 <code>/var/log/xend.log</code> 中找到此日志文件
(-) <code>delete</code>	从 Xend 域管理中去除域。 <code>list</code> 选项显示域名	<code>virsh undefine</code>
(-) <code>new</code>	将域添加到 Xend 域管理	<code>virsh define</code>
(-) <code>start</code>	启动使用 <code>xm new</code> 命令添加的 Xend 受管域	<code>virsh start</code>
(-) <code>dryrun</code>	试运行 - 列显 SXP 中生成的配置，但不创建域	<code>xl -N</code>
(-) <code>reset</code>	重置域	<code>virsh reset</code>

已去除的域管理选项		
选项	任务	等效选项
(-) <u>domstate</u>	显示域状态	<u>virsh domstate</u>
(-) <u>serve</u>	通过 stdio 代理 Xend XMLRPC	
(-) <u>resume DOMAIN OPTIONS</u>	使域脱离挂起状态，并将其移回到内存中	<u>virsh resume</u>
(-) <u>suspend DOMAIN</u>	在状态文件中挂起域，以便稍后可以使用 <u>resume</u> 子命令将其恢复。与 <u>save</u> 子命令类似，但不能指定状态文件	<u>virsh managedsave</u> <u>virsh suspend</u>

C3.4.2 USB 设备

USB options 不可用于 xl/libxl 工具堆栈。virsh 提供 attach-device 和 detach-device 选项，但尚不适用于 USB。

表 C5：已去除的 USB 设备管理选项

已去除的 USB 设备管理选项	
选项	任务
(-) <u>usb-add</u>	将新 USB 物理总线添加到域
(-) <u>usb-del</u>	从域中删除 USB 物理总线
(-) <u>usb-attach</u>	将新 USB 物理总线挂接到域的虚拟端口
(-) <u>usb-detach</u>	从域的虚拟端口分离 USB 物理总线
(-) <u>usb-list</u>	列出域的所有虚拟端口挂接状态

已去除的 USB 设备管理选项	
选项	任务
(-) <u>usb-list-assignable-devices</u>	列出所有可指派的 USB 设备
(-) <u>usb-hc-create</u>	创建域的新虚拟 USB 主机控制器
(-) <u>usb-hc-destroy</u>	销毁域的虚拟 USB 主机控制器

C3.4.3 CPU 管理

CPU 管理选项发生了变化。我们提供了一些新选项，请参见：第 C3.5.10 节 “[xlcpupool-*](#)”

表 C6：已去除的 CPU 管理选项

已去除的 CPU 管理选项	
选项	任务
(-) <u>cpupool-new</u>	将 CPU 池添加到 Xend CPU 池管理
(-) <u>cpupool-start</u>	启动 Xend CPU 池
(-) <u>cpupool-delete</u>	从 Xend 管理中去除 CPU 池

C3.4.4 其他选项

表 C7：其他选项

其他已去除的选项	
选项	任务
(-) <u>shell</u>	启动交互式外壳

其他已去除的选项	
选项	任务
(-) <u>change-vnc-passwd</u>	更改 vnc 口令
(-) <u>vtpm-list</u>	列出虚拟 TPM 设备
(-) <u>block-configure</u>	更改块设备配置

C3.5 已更改的选项

C3.5.1 create

xl create CONFIG_FILE OPTIONS VARs



注意：libvirt 等效选项：

virsh create

表 C8：已更改的 **xlcreate** 选项

已更改的 <u>create</u> 选项	
选项	任务
(*) -f= <u>FILE</u> 、-- defconfig= <u>FILE</u>	使用给定的配置文件

表 C9：已去除的 **xmcreate** 选项

已去除的 <u>create</u> 选项	
选项	任务
(-) -s、-- <u>skipdtd</u>	跳过 DTD 检查 - 创建之前跳过 XML 检查

已去除的 <code>create</code> 选项	
选项	任务
(-) <code>-x</code> 、 <code>--xmldryrun</code>	XML 试运行
(-) <code>-F=FILE</code> 、 <code>--config=FILE</code>	使用给定的 <code>SXP</code> 格式配置脚本
(-) <code>--path</code>	在路径中搜索配置脚本
(-) <code>--help_config</code>	列显配置脚本的可用配置变量 (var)
(-) <code>-n</code> 、 <code>--dryrun</code>	试运行 — 列显 <code>SXP</code> 中的配置，但不创建域
(-) <code>-c</code> 、 <code>--console_autoconnect</code>	创建域后连接到控制台
(-) <code>-q</code> 、 <code>--quiet</code>	安静模式
(-) <code>-p</code> 、 <code>--paused</code>	创建域后使其保持暂停状态

表 C10：添加的 `xlcreate` 选项

添加的 <code>create</code> 选项	
选项	任务
(+) <code>-V</code> 、 <code>--vncviewer</code>	挂接到域的 VNC 服务器，并派生 vncviewer 进程
(+) <code>-A</code> 、 <code>--vncviewer-autopass</code>	通过 stdin 将 VNC 口令传递给 vncviewer

C3.5.2 `console`

`xl console OPTIONS DOMAIN`



注意：libvirt 等效选项

virsh console

表 C11：添加的 **xlconsole** 选项

添加的 <u>console</u> 选项	
选项	任务
(+) <u>-t</u> [pv serial]	连接到 PV 控制台，或连接到模拟的串行控制台。PV 域只能使用 PV 控制台，而 HVM 域可以使用上述两种控制台

C3.5.3 **info**

xl info

表 C12：已去除的 **xminfo** 选项

已去除的 <u>info</u> 选项	
选项	任务
(-) <u>-n</u> 、 <u>--numa</u>	Numa 信息
(-) <u>-c</u> 、 <u>--config</u>	列出 Xend 配置参数

C3.5.4 **dump-core**

xl dump-core DOMAIN FILENAME



注意：libvirt 等效选项

virsh dump

表 C13：已去除的 xmdump-core 选项

已去除的 dump-core 选项	
选项	任务
(-) <u>-L</u> 、 <u>--live</u>	在不暂停域的情况下转储核心
(-) <u>-C</u> 、 <u>--crash</u>	转储核心后使域崩溃
(-) <u>-R</u> 、 <u>--reset</u>	转储核心后重置域

C3.5.5 list

xl list options DOMAIN



注意：libvirt 等效选项

virsh list --all

表 C14：已去除的 xmlist 选项

已去除的 list 选项	
选项	任务
(-) <u>-l</u> 、 <u>--long</u>	<u>xm SXP</u> 的输出以 <u>list</u> 格式呈现数据
(-) <u>--state==STATE</u>	输出处于指定状态的 VM 的信息

表 C15：添加的 xllist 选项

添加的 list 选项	
选项	任务
(+) <u>-Z</u> 、 <u>--context</u>	同时列显安全标签
(+) <u>-v</u> 、 <u>--verbose</u>	同时列显域 UUID、关机原因和安全标签

C3.5.6 mem-*

注意：libvirt 等效选项

virsh setmem

virsh setmaxmem

表 C16：已更改的 xlmem-* 选项

已更改的 <u>mem-*</u> 选项	
选项	任务
<u>mem-max</u> <u>DOMAIN_ID</u> <u>MEM</u>	追加 <u>t</u> （表示 TB）、 <u>g</u> （表示 GB）、 <u>m</u> （表示 MB）、 <u>k</u> （表示 KB）和 <u>b</u> （表示字节）。指定域可以使用的最大内存量。
<u>mem-set</u> <u>DOMAIN_ID</u> <u>MEM</u>	使用气球驱动程序设置域的已用内存

C3.5.7 migrate

xl migrate OPTIONS DOMAIN HOST

注意：libvirt 等效选项

virsh migrate --live hvm-sles11-qcow2 xen+

CONNECTOR://USER@IP_ADDRESS/

表 C17：已去除的 xmmigrate 选项

已去除的 <u>migrate</u> 选项	
选项	任务
(-) <u>-l</u> 、 <u>--live</u>	使用实时迁移。这会在不关闭域的情况下在主机之间迁移域
(-) <u>-r</u> 、 <u>--resource</u> <u>Mbs</u>	设置迁移域时允许达到的最大迁移速度 (Mbs)

已去除的 <u>migrate</u> 选项	
选项	任务
(-) <u>-c</u> 、 <u>--change_home_server</u>	更改受管域的宿主服务器
(-) <u>--max_iters=MAX_ITERS</u>	在最终挂起之前的迭代次数（默认值为 30）
(-) <u>--max_factor=MAX_FACTOR</u>	在最终挂起之前传送的最大内存量（默认值：3*RAM）。
(-) <u>--min_remaining=MIN_REMAINING</u>	在最终挂起之前的脏页数（默认值为 50）
(-) <u>--abort_if_busy</u>	中止迁移，而不是执行最终挂起
(-) <u>--log_progress</u>	在 <u>xend.log</u> 中记录迁移进度
(-) <u>-s</u> 、 <u>--ssl</u>	使用 SSL 连接进行迁移

表 C18：添加的 xlmigrate 选项

添加的 <u>migrate</u> 选项	
选项	任务
(+) <u>-s</u> <u>SSHCOMMAND</u>	使用 <sshcommand> 而不是 <u>ssh</u>
(+) <u>-e</u>	在新主机上，不要在后台（在 <host> 上）等待域死机
(+) <u>-C</u> <u>CONFIG</u>	发送 <config> 而不是创建域时使用的配置文件

C3.5.8 域管理

xl reboot OPTIONS DOMAIN



注意：libvirt 等效选项

virsh reboot

表 C19：已去除的 **xmreboot** 选项

已去除的 <u>reboot</u> 选项	
选项	任务
(-) <u>-a</u> 、 <u>--all</u>	重引导所有域
(-) <u>-w</u> 、 <u>--wait</u>	等待重引导完成后再返回。这可能需要一段时间，因为需要干净地关闭域中的所有服务

表 C20：添加的 **xlreboot** 选项

添加的 <u>reboot</u> 选项	
选项	任务
(+) <u>-F</u>	对于没有 PV 驱动程序的 HVM Guest，回退到 ACPI 重置事件

xl save OPTIONS DOMAIN CHECK_POINT_FILE CONFIG_FILE



注意：libvirt 等效选项

virsh save

表 C21：添加的 **xlsave** 选项

添加的 <u>save</u> 选项	
选项	任务
(+) <u>-c</u>	创建快照后使域保持运行状态

`xl restore` OPTIONS CONFIG_FILE CHECK_POINT_FILE

 **注意：libvirt 等效选项**
`virsh restore`

表 C22：添加的 `xlrestore` 选项

添加的 <code>restore</code> 选项	
选项	任务
(+) <code>-p</code>	恢复域后不将其取消暂停
(+) <code>-e</code>	在新主机上不在后台等待域死机
(+) <code>-d</code>	启用调试消息
(+) <code>-V</code> 、 <code>--vncviewer</code>	挂接到域的 VNC 服务器，并派生 <code>vncviewer</code> 进程
(+) <code>-A</code> 、 <code>--vncviewer-autopass</code>	通过 <code>stdin</code> 将 VNC 口令传递给 <code>vncviewer</code>

`xl shutdown` OPTIONS DOMAIN

 **注意：libvirt 等效选项**
`virsh shutdown`

表 C23：已去除的 `xmshutdown` 选项

已去除的 <code>shutdown</code> 选项	
选项	任务
(-) <code>-w</code> 、 <code>--wait</code>	等待域完成关机后再返回

已去除的 shutdown 选项	
选项	任务
(-) <u>-a</u>	关闭所有 Guest 域
(-) <u>-R</u>	
(-) <u>-H</u>	

表 C24：添加的 xlshutdown 选项

添加的 shutdown 选项	
选项	任务
(+) <u>-F</u>	如果 Guest 不支持 PV 关机控制，则回退为发送 ACPI 电源事件

表 C25：已更改的 xltrigger 选项

已更改的 trigger 选项	
选项	任务
<u>trigger DOMAIN</u> <nmi reset init power sleep s3resume> <u>VCPU</u>	向域发送触发器。仅适用于 HVM 域

C3.5.9 xlsched-*

xl sched-credit OPTIONS



注意：libvirt 等效选项

virsh schedinfo

表 C26：已去除的 `xmsched-credit` 选项

已去除的 <code>sched-credit</code> 选项	
选项	任务
<code>-d DOMAIN、--domain=DOMAIN</code>	域
<code>-w WEIGHT、--weight=WEIGHT</code>	权重为 512 的域将获得的 CPU 是所争用的主机上权重为 256 的域的两倍。合法权重范围为 1 到 65535，默认值为 256
<code>-c CAP、--cap=CAP</code>	CAP 可选择性修复域能够消耗的最大 CPU 数量

表 C27：添加的 `xl sched-credit` 选项

添加的 <code>sched-credit</code> 选项	
选项	任务
<code>(+) -p CPUP00L、--cpupool=CPUP00L</code>	将输出内容限制为指定 CPU 池中的域
<code>(+) -s、--schedparam</code>	指定此选项可列出或设置池范围的调度程序参数
<code>(+) -t TSLICE、--tslice_ms=TSLICE</code>	时间片 (TSLICE) 告知调度程序要允许 VM 运行多长时间后再开始抢占模式
<code>(+) -r RLIMIT、--ratelimit_us=RLIMIT</code>	Ratelimit 会尝试限制每秒调度次数

`xl sched-credit2 OPTIONS`



注意：libvirt status

`virsh` 仅支持 credit 调度程序，不支持 credit2 调度程序

表 C28：已去除的 `xmsched-credit2` 选项

已去除的 <code>sched-credit2</code> 选项	
选项	任务
<code>-d DOMAIN、--domain=DOMAIN</code>	域
<code>-w WEIGHT、--weight=WEIGHT</code>	合法权重范围为 1 到 65535，默认值为 256

表 C29：添加的 `xl sched-credit2` 选项

添加的 <code>sched-credit2</code> 选项	
选项	任务
<code>(+) -p CPUP00L、--cpupool=CPUP00L</code>	将输出内容限制为指定 CPU 池中的域

`xl sched-sedf` OPTIONS

表 C30：已去除的 `xmsched-sedf` 选项

已去除的 <code>sched-sedf</code> 选项	
选项	任务
<code>-p PERIOD、--period=PERIOD</code>	常规 EDF 调度用法，以毫秒为单位
<code>-s SLICE、--slice=SLICE</code>	常规 EDF 调度用法，以毫秒为单位
<code>-l LATENCY、--latency=LATENCY</code>	域执行大量 I/O 时延长的时段

已去除的 <code>sched-sedf</code> 选项	
选项	任务
<code>-e EXTRA</code> 、 <code>--extra=EXTRA</code>	允许域额外运行一段时间的标志（0 或 1）
<code>-w WEIGHT</code> 、 <code>--weight=WEIGHT</code>	另一种设置 CPU 片的方式

表 C31：添加的 `xl sched-sedf` 选项

添加的 <code>sched-sedf</code> 选项	
选项	任务
(+) <code>-c CPUP00L</code> 、 <code>--cpupool=CPUP00L</code>	将输出内容限制为指定 CPU 池中的域
(+) <code>-d DOMAIN</code> 、 <code>--domain=DOMAIN</code>	域

C3.5.10 `xlcpupool - *`

`xl cpupool-cpu-remove CPU_P00L <CPU nr>|node:<node nr>`

`xl cpupool-list [-c|--cpus] CPU_P00L`

表 C32：已去除的 `xmcpupool-list` 选项

已去除的 <code>cpupool - *</code> 选项	
选项	任务
(-) <code>-l</code> 、 <code>--long</code>	以 <code>SXP</code> 格式输出所有 CPU 池细节

`xl cpupool-cpu-add CPU_P00L cpu-nr|node:node-nr`

xl cpupool-create OPTIONS CONFIG_FILE [Variable=Value ...]

表 C33：已去除的 **xmcpupool-create** 选项

已去除的 <u>cpupool-create</u> 选项	
选项	任务
(-) <u>-f</u> <u>FILE</u> , -- <u>defconfig=FILE</u>	使用给定的 Python 配置脚本。处理参数后会加载该配置文件
(-) <u>-n</u> 、-- <u>dryrun</u>	试运行 - 列显 SXP 中生成的配置，但不创建 CPU 池
(-) -- <u>help_config</u>	列显配置脚本的可用配置变量 (var)
(-) -- <u>path=PATH</u>	在路径中搜索配置脚本。PATH 的值是冒号分隔的目录列表
(-) <u>-F=FILE</u> 、-- <u>config=FILE</u>	要使用的 CPU 池配置 (SXP)

C3.5.11 **PCI 和块设备**

xl pci-detach [-f] DOMAIN_ID <BDF>



注意：libvirt 等效选项

virsh detach-device

表 C34：添加的 **xlpci-detach** 选项

添加的 <u>pci-detach</u> 选项	
选项	任务
(+) <u>-f</u>	如果指定了 <u>-f</u> ，即使在没有 Guest 协作的情况下， xl 也会强制去除设备

表 C35：已去除的 `xmblock-list` 选项

已去除的 <code>block-list</code> 选项	
选项	任务
(-) <code>-l</code> 、 <code>--long</code>	列出域的虚拟块设备

表 C36：其他选项

选项	<code>libvirt</code> 等效选项
<code>xl block-attach DOMAIN</code> <code><disk-spec-component(s)></code>	<code>virsh attach-disk/attach-device</code>
<code>xl block-list DOMAIN_ID</code>	<code>virsh domblklist</code>

C3.5.12 网络

表 C37：网络选项

选项	<code>libvirt</code> 等效选项
<code>xl network-list</code> <code>DOMAIN(s)</code>	<code>virsh domiflist</code>
<code>xl network-detach</code> <code>DOMAIN_ID devid mac</code>	<code>virsh detach-interface</code>
<code>xl network-attach</code> <code>DOMAIN(s)</code>	<code>virsh attach-interface/attach-device</code>

表 C38：已去除的 `xlnetwork-attach` 选项

已去除的选项	
选项	任务
(-) <code>-l</code> 、 <code>--long</code>	

C3.6 新选项

表 C39：新选项

选项	任务
<code>config-update DOMAIN</code> <code>CONFIG_FILE OPTIONS</code> <code>VARS</code>	更新为运行中的域保存的配置。此更新不会立即生效，而是在 Guest 下次重新启动时应用。此命令可用于确保在 Guest 重新启动时保留对 Guest 进行的运行时修改
<code>migrate-receive</code>	
<code>sharing DOMAIN</code>	列出共享页的计数。专门列出指定的域的该信息。如果未指定域，则列出所有域的该信息
<code>vm-list</code>	列显有关 Guest 的信息。此列表不包括有关 Dom0 和存根域等服务或辅助域的信息
<code>cpupool-rename</code> <code>CPU_POOL NEWNAME</code>	将 CPU 池重命名为 newname
<code>cpupool-numa-split</code>	将计算机分割为在每个 numa 节点一个 CPU 池
<code>cd-insert DOMAIN</code> <VirtualDevice> <type:path>	将 CD-ROM 插入 Guest 域的现有虚拟 CD 驱动器。该虚拟驱动器必须已存在，但当前可以是空的
<code>cd-eject DOMAIN</code> <VirtualDevice>	从 Guest 的虚拟 CD 驱动器中弹出 CD-ROM。仅适用于 HVM 域
<code>pci-assignable-list</code>	列出所有可分配的 PCI 设备。它们是系统中配置为可直通且已绑定到 Dom0 中的适当 PCI 后端驱动程序（而不是真实驱动程序）的设备
<code>pci-assignable-add</code> <BDF>	使 PCI 总线/设备/功能 BDF 中的设备可分配到 Guest。这会将该设备绑定到 pciback 驱动程序

选项	任务
<code>pci-assignable-remove</code> <code>OPTIONS <BDF></code>	使 PCI 总线/设备/功能 BDF 中的设备可分配到 Guest。这至少会从 pciback 取消绑定该设备
<code>loadpolicy</code> <code>POLICY_FILE</code>	从给定的策略文件加载 FLASK 策略。初始策略将以多重引导模块的形式提供给超级管理程序；此命令允许对策略进行运行时更新。加载新安全策略会重置对设备标签进行的运行时更改

C4 外部链接

有关 Xen 工具堆栈的详细信息，请参见以下联机资源：

Xen 中的 XL

Xen 4.2 中的 XL (https://wiki.xenproject.org/wiki/XL_in_Xen_4.2) ↗

xl 命令

XL (<https://xenbits.xen.org/docs/unstable/man/xl.1.html>) ↗ 命令行。

xl.cfg

xl.cfg (<https://xenbits.xen.org/docs/unstable/man/xl.cfg.5.html>) ↗ 域配置文件语法。

xl disk

xl disk (<https://xenbits.xen.org/docs/4.3-testing/misc/xl-disk-configuration.txt>) ↗ 配置选项。

XL 与 Xend

XL 与 Xend (https://wiki.xenproject.org/wiki/XL_vs_Xend_Feature_Comparison) ↗ 功能比较。

BDF 文档

BDF 文档 (https://wiki.xen.org/wiki/Bus:Device.Function_%28BDF%29_Notation) ↗。

libvirt

virsh (<https://libvirt.org/sources/virshcmdref/html/>) ↗ 命令。

C5 以与 **xm** 兼容的格式保存 Xen Guest 配置

尽管 **xl** 是当前用于管理 Xen Guest 的工具集（此外还有首选的 **libvirt**），但您可能需要将 Guest 配置导出为过去使用的 **xm** 格式。要实现此目的，请执行以下步骤：

1. 首先将 Guest 配置导出到某个文件中：

```
> virsh dumpxml guest_id > guest_cfg.xml
```

2. 然后将配置转换为 **xm** 格式：

```
> virsh domxml-to-native xen-xm guest_cfg.xml > guest_xm_cfg
```

This appendix contains the GNU Free Documentation License version 1.2.

GNU Free Documentation License

Copyright (C) 2000, 2001, 2002 Free Software Foundation, Inc. 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA. Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role

of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document

within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <https://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

```
Copyright (c) YEAR YOUR NAME.
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.2
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU
Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.