云联万维
YUNION

虚拟机迁移与主备机简介

目录 /Contents

01

迁移使用场景

场景一：宿主机宕机

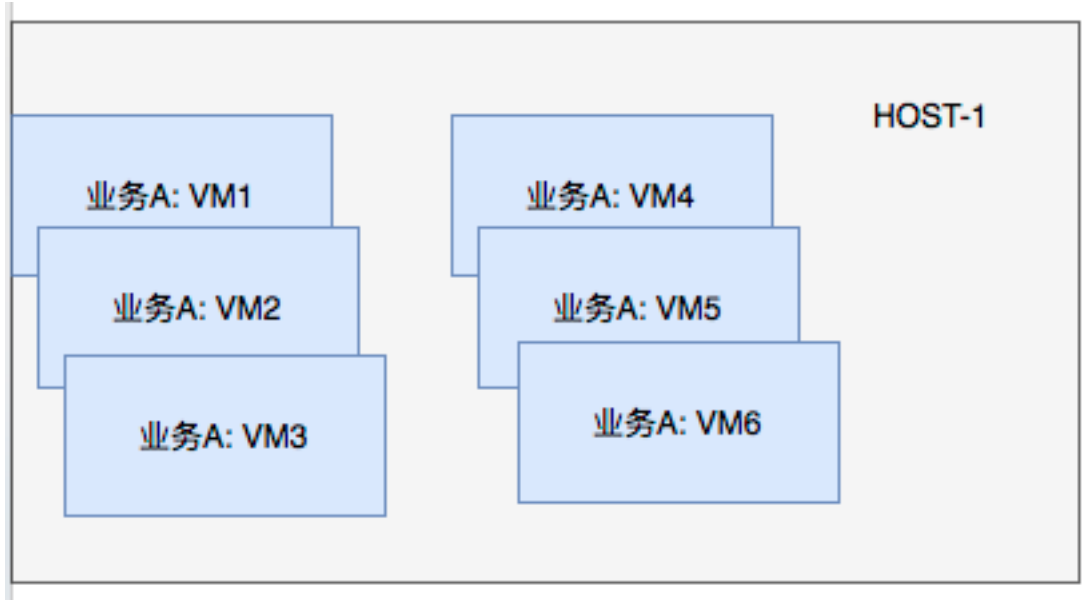
Q: 现实世界中宿主机总会由于各种各样的原因宕机，
如何在宿主机宕机的情况下恢复业务？

A: 对于共享存储可以进行紧急模式下的冷迁移，
在另外一台宿主机上启动该虚拟机。

Q: 那本地存储的虚拟机该如何处理？

场景二：业务集群优化

- 某个业务的虚拟机集中在一台宿主机上，宿主机宕机对业务的影响非常大



- 某业务不支持横向扩容，做虚拟机热扩容时宿主机的资源不够

场景三：宿主机维护

- 生产环境下运行一定时长的宿主机可能需要维护或者下线，进行硬件的检修

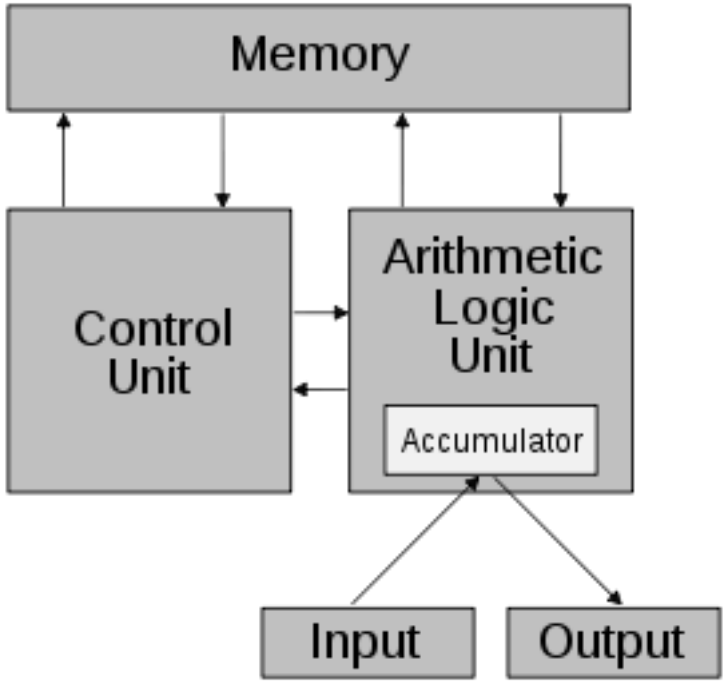
⭐: 我们在迁移的基础上开发了宿主机维护模式的功能，将宿主机禁用
并将宿主机上的虚拟机全部迁移

02

迁移的原理介绍

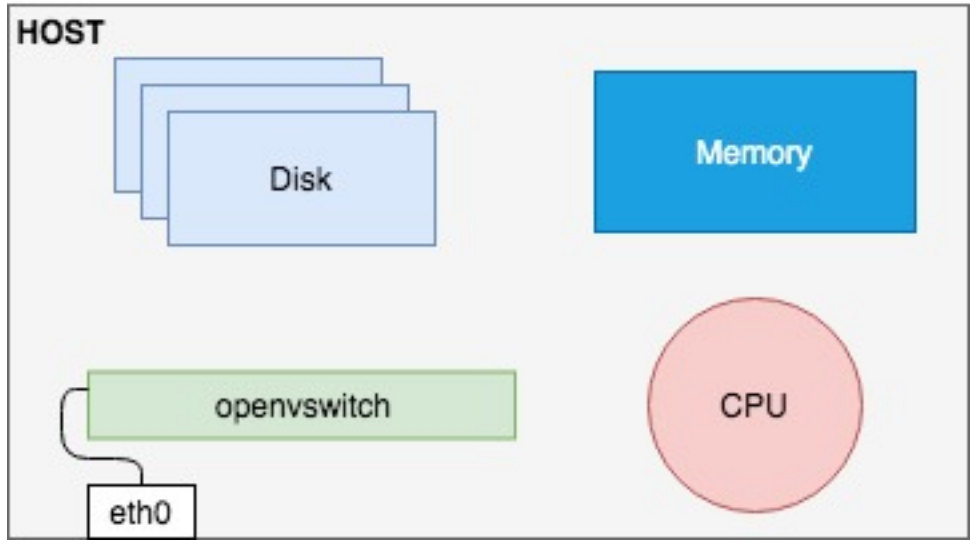
虚拟机的组成

冯诺依曼体系结构中计算机由运算器、控制器、存储器、输入和输出设备组成，分别对应CPU、内存、磁盘和网络等设备



虚拟机的组成

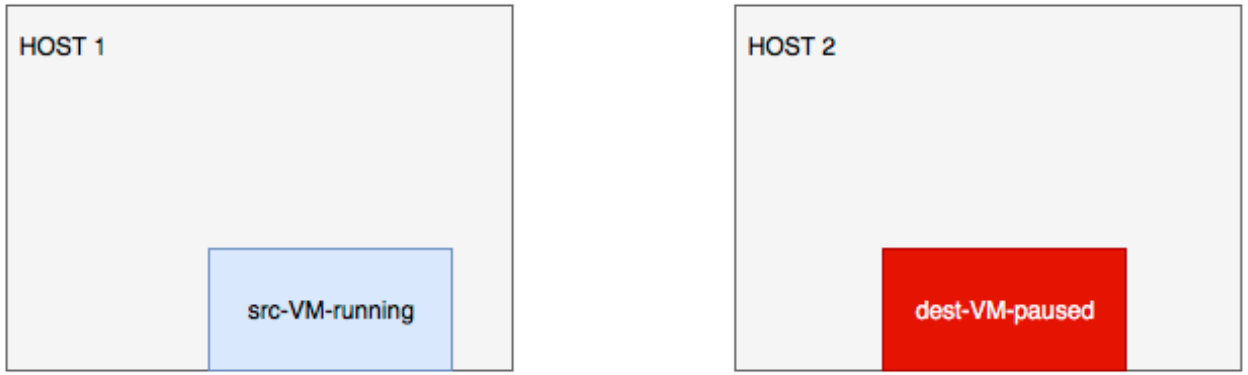
虚拟机也是如此，由Hypervisor为虚拟机模拟CPU、内存、磁盘和网络等设备，而Hypervisor运行在宿主机之上，我们从CPU、内存、磁盘和网络的角度来看看虚拟机迁移是怎么实现的




```
root     2262665       1  0 Sep29 ?        02:16:51 /usr/local/qemu-2.12.1/bin/qemu-system-x
86_64 -enable-kvm -cpu host,kvm=off -chardev socket,id=hmqmondev,port=55937,host=127.0.0.1,n
odelay,server,nowait -mon chardev=hmqmondev,id=hmqmon,mode=readline -chardev socket,id=qmqmo
ndev,port=56137,host=127.0.0.1,nodelay,server,nowait -mon chardev=qmqmondev,id=qmqmon,mode=c
ontrol -rtc base=utc,clock=host,driftfix=none -daemonize -nodefaults -nodefconfig -no-kvm-pi
t-reinjection -global kvm-pit.lost_tick_policy=discard -machine pc,accel=kvm -k en-us -smp 2
,maxcpus=128 -name iso-test -m 2048M,slots=4,maxmem=262144M -boot order=cdn -device virtio-s
erial -usb -device usb-kbd -device usb-tablet -vga std -vnc :37,password -device virtio-scsi
-pci,id=scsi -drive file=/opt/cloud/workspace/disks/e027faa8-9a34-46d4-80e7-e6eb23f0c2f5,if=
none,id=drive_0,cache=none,aio=native -device scsi-hd,drive=drive_0,bus=scsi.0,id=drive_0 -d
evice ide-cd,drive=ide0-cd0,bus=ide.1,unit=1 -drive id=ide0-cd0,media=cdrom,if=none -netdev
type=tap,id=vnet222-216,ifname=vnet222-216,vhost=on,vhostforce=off,script=/opt/cloud/workspa
ce/servers/f35d077a-5910-4504-88ab-e6cb22a50194/if-up-br0-vnet222-216.sh,downscript=/opt/clo
ud/workspace/servers/f35d077a-5910-4504-88ab-e6cb22a50194/if-down-br0-vnet222-216.sh -device
 virtio-net-pci,netdev=vnet222-216,mac=00:22:f8:7e:08:c5,addr=0xf,speed=1000 -pidfile /opt/c
loud/workspace/servers/f35d077a-5910-4504-88ab-e6cb22a50194/pid -chardev socket,path=/opt/cl
oud/workspace/servers/f35d077a-5910-4504-88ab-e6cb22a50194/qga.sock,server,nowait,id=qga0 -d
evice virtserialport,chardev=qga0,name=org.qemu.guest_agent.0 -object rng-random,filename=/d
ev/random,id=rng0 -device virtio-rng-pci,rng=rng0,max-bytes=1024,period=1000 -chardev pty,id
=charserial0 -device isa-serial,chardev=charserial0,id=serial0 -device pvpanic
```

准备目标虚拟机

- 在开始迁移之前，需要在目标宿主机上创建一台虚拟机，该虚拟机处于暂停状态



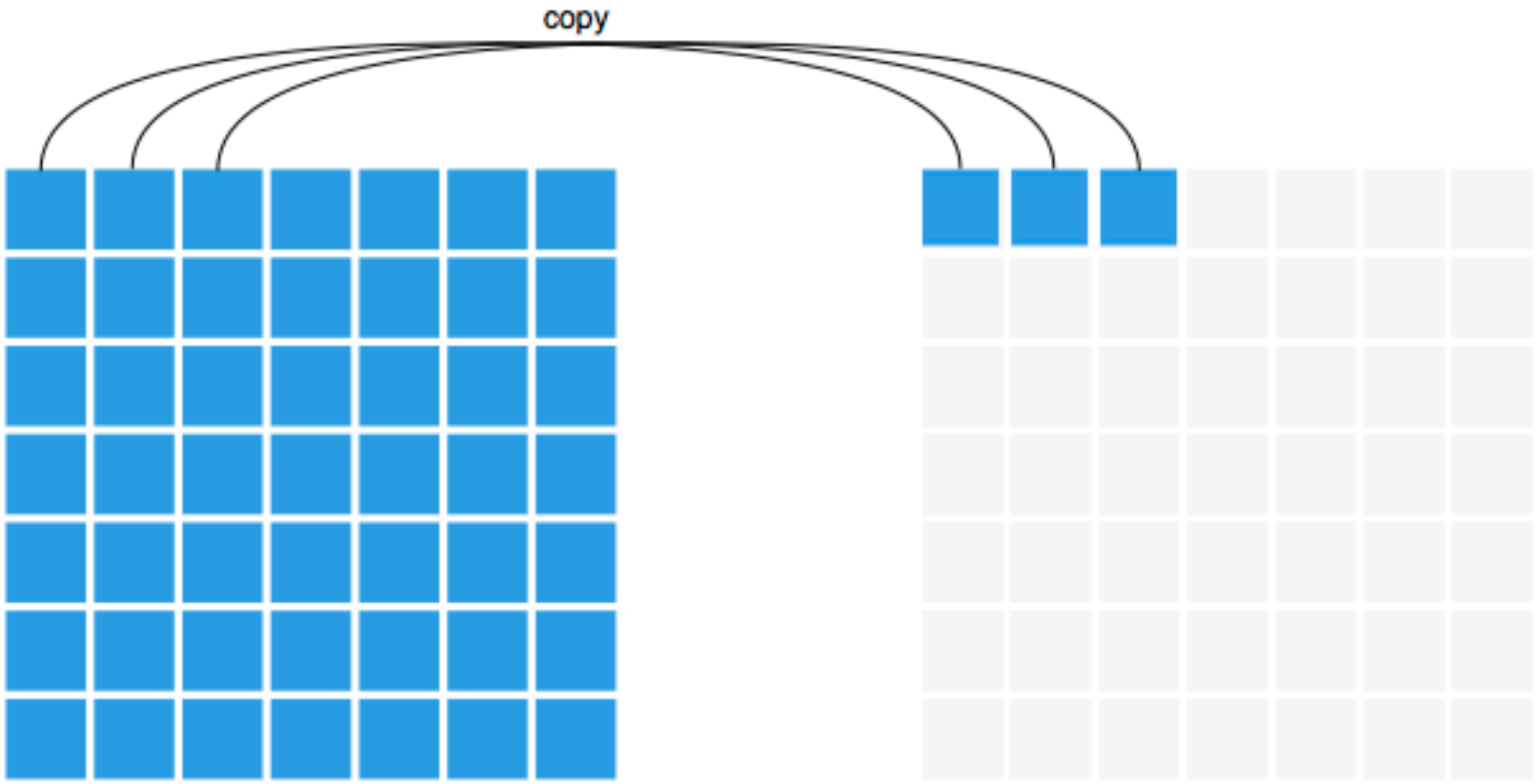
- 目标宿主机上创建虚拟机的时候需要指定好虚拟机CPU的数量和内存的大小，磁盘和网络也需要先准备好，虚拟机本地磁盘一般采用qcow2文件

```
root disks # qemu-img info 8c0184dc-189a-4b4a-890c-d7285a739ac1
image: 8c0184dc-189a-4b4a-890c-d7285a739ac1
file format: qcow2
virtual size: 30G (32212254720 bytes)
disk size: 288M
cluster_size: 2097152
backing file: /opt/cloud/workspace/disks/image_cache/e15647c8-c996-4fab-a0b8-3fa27b8cb5c8
Format specific information:
    compat: 1.1
    lazy refcounts: false
    refcount bits: 16
    corrupt: false
```

- 启动后目标宿主机就会为虚拟机分配CPU和内存，然后就可以开始迁移工作了

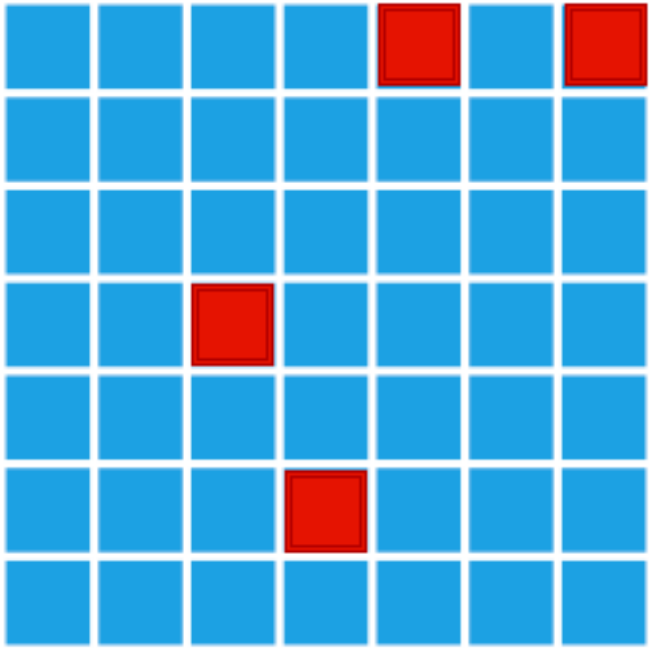
内存迁移

- 内存的迁移采用的预拷贝的方式，也就是把源虚拟机的内存按页的方式拷贝到目的虚拟机，直到迁移完成

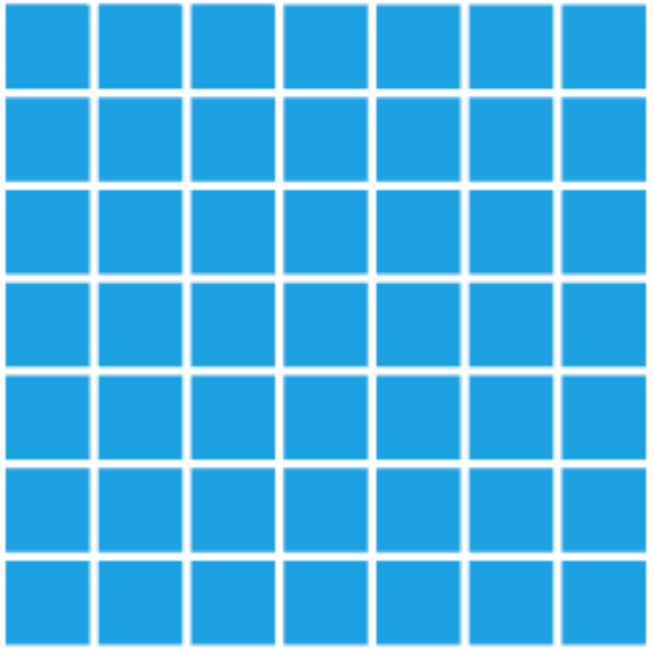


源虚拟机　　目标虚拟机

内存迁移



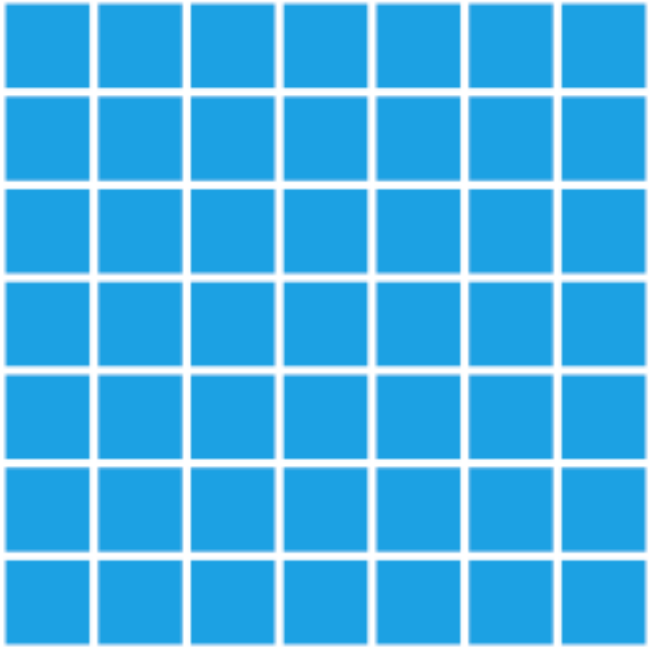
源虚拟机



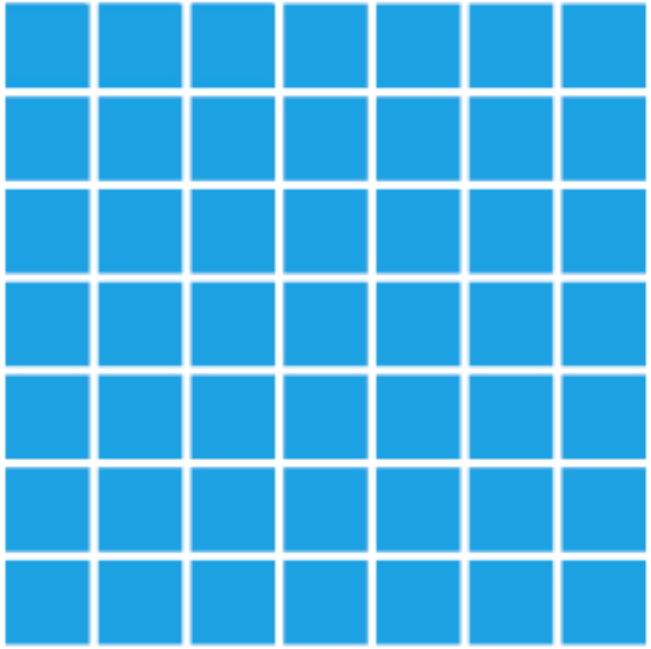
目标虚拟机

如果一个内存在拷贝之后再次被修改了，将会置为dirty状态，那么那将会在下一轮拷贝中发送到目的节点

内存迁移



源虚拟机



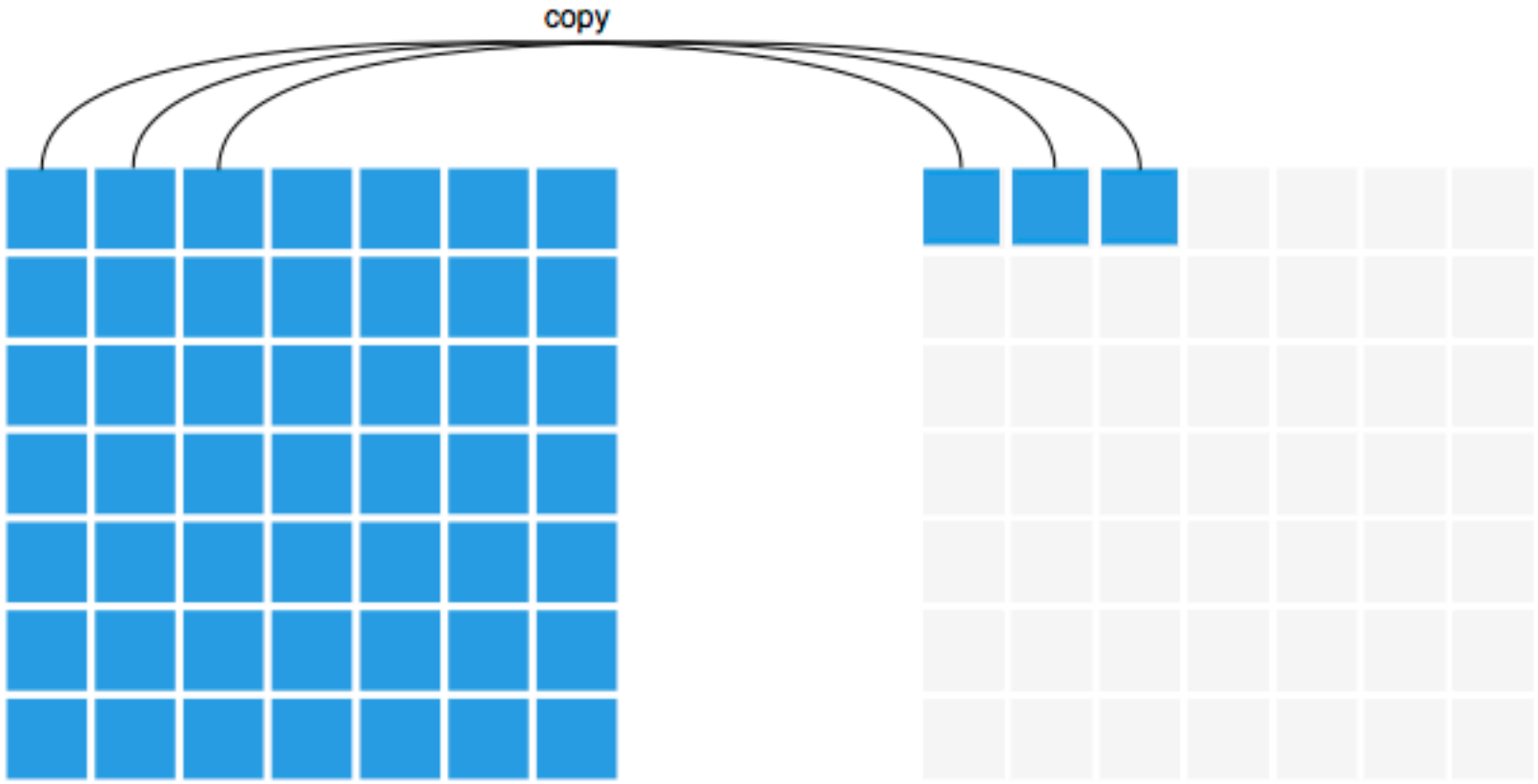
目标虚拟机

这个迭代的过程是收敛的，直到能够一次性迁移完所有数据，那么源虚拟机就会暂停，从而避免产生新的脏数据，以便进行迁移收尾工作。

磁盘迁移

本地存储的磁盘才需要迁移，共享存储的磁盘不需要迁移

磁盘的迁移与内存的迁移过程是类似的，磁盘以块为单位向目标宿主机的磁盘拷贝，直到迁移结束



源虚拟机　　目标虚拟机

CPU

- 在虚拟化环境中，CPU的状态也是保存在内存中， CPU状态的拷贝几乎可以瞬间完成。

值得注意的是由于创建虚拟机的时候CPU的类型默认使用的是宿主机的CPU类型 -cpu host ,

在调度的时候需要要求目标宿主机的CPU型号和源宿主机的CPU型号相等，避免迁移失败

如果看host服务的日志的话, 可以看到迁移完成时里面有个这样一条日志

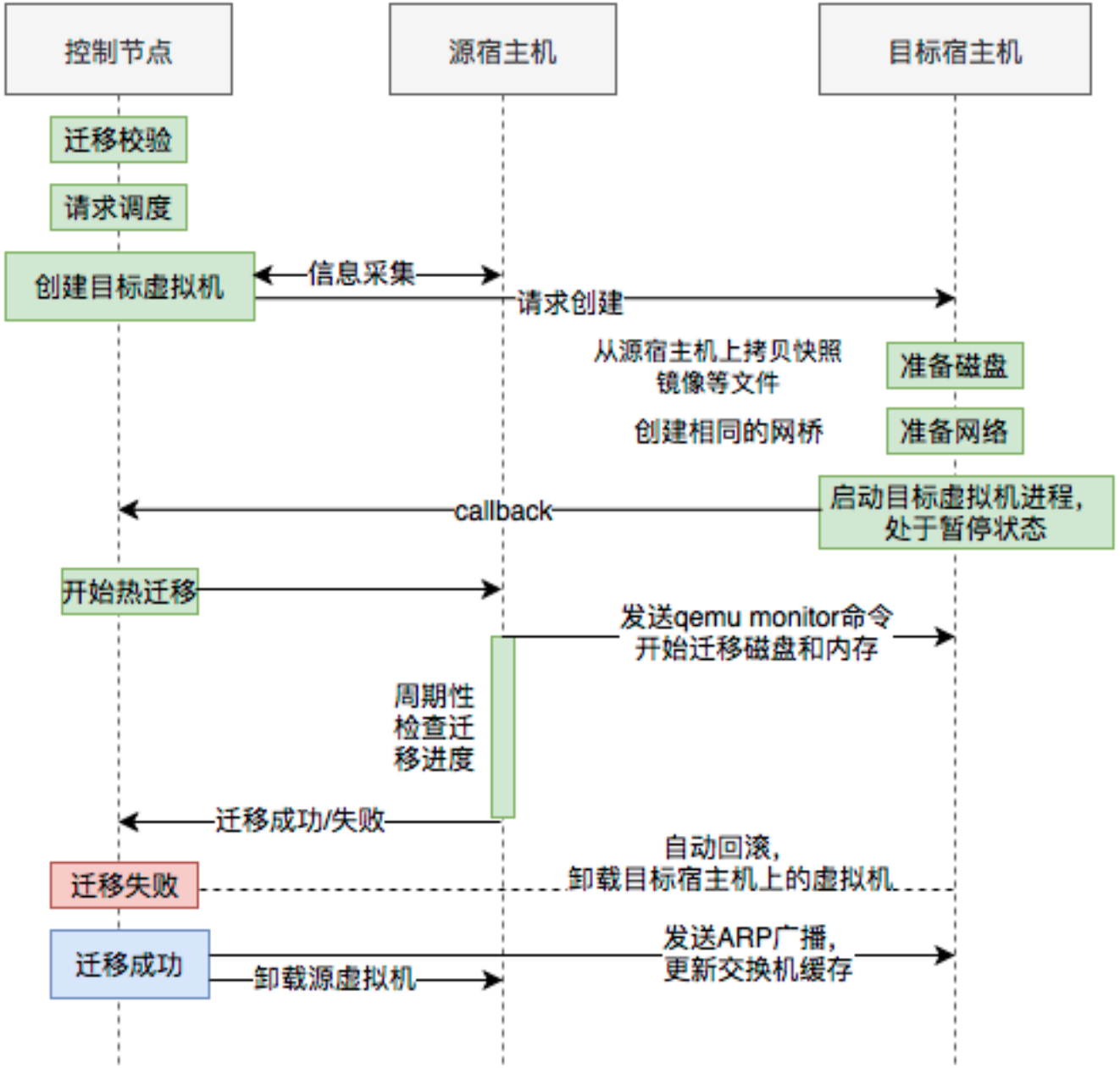

```
[I 191023 17:40:58 monitor.(*QmpMonitor).GetMigrateStatus.func1(qmp.go:628)] Query migrate status: {"downtime":58,"ram":{"dirty-pages-rate":0,"dirty-sync-count":5,"duplicate":129743,"mbps":84.471684,"normal":150836,"normal-bytes":617824256,"page-size":4096,"postcopy-requests":0,"remaining":0,"skipped":0,"total":1091379200,"transferred":620862440},"setup-time":20,"status":"completed","total-time":84473}
```

网络

- 目标宿主机上复制相同的网络接口，虚拟机的MAC和IP保持一致

- 在迁移完成后，宿主机主动发ARP广播通告交换机

云平台迁移流程


控制节点
源宿主机
目标宿主机
迁移校验
请求调度
创建目标虚拟机
信息采集
请求创建
从源宿主机上拷贝快照
镜像等文件
准备磁盘
创建相同的网桥
准备网络
callback
启动目标虚拟机进程，
处于暂停状态
开始热迁移
发送qemu monitor命令
开始迁移磁盘和内存
周期性
检查迁
移进度
迁移成功/失败
自动回滚，
卸载目标宿主机上的虚拟机
迁移失败
迁移成功
卸载源虚拟机
发送ARP广播，
更新交换机缓存

迁移总结

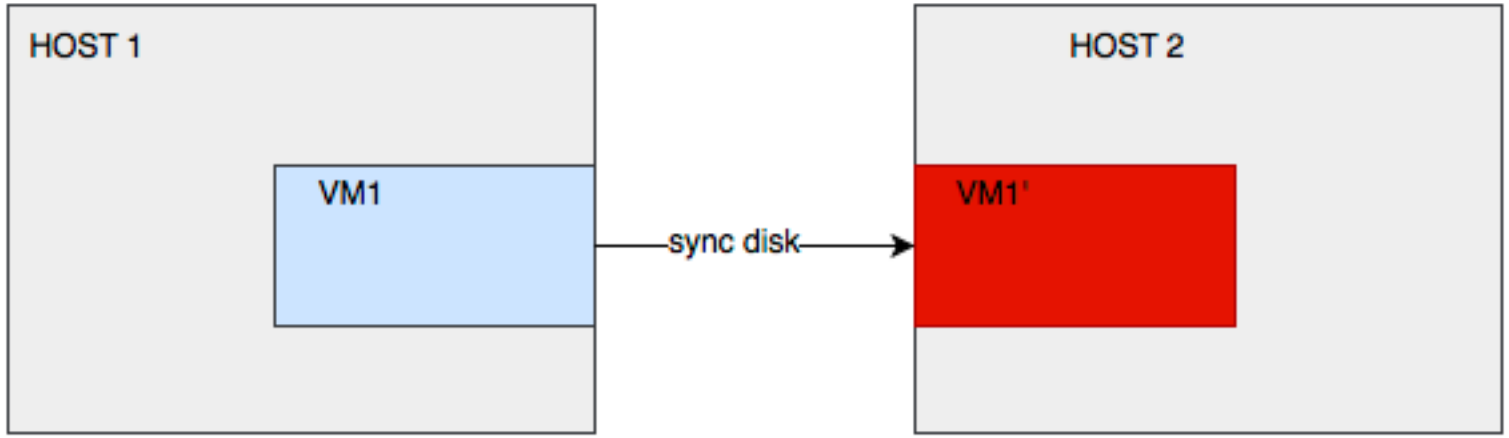
- 虚拟机有GPU等透传设备不能进行热迁移
- 源宿主机和目的宿主机要在一个广播域内
- 源宿主机和目的宿主机CPU型号一致
- 虚拟机在热迁移的切换过程中有downtime，取决于当时的环境

03

主备机介绍

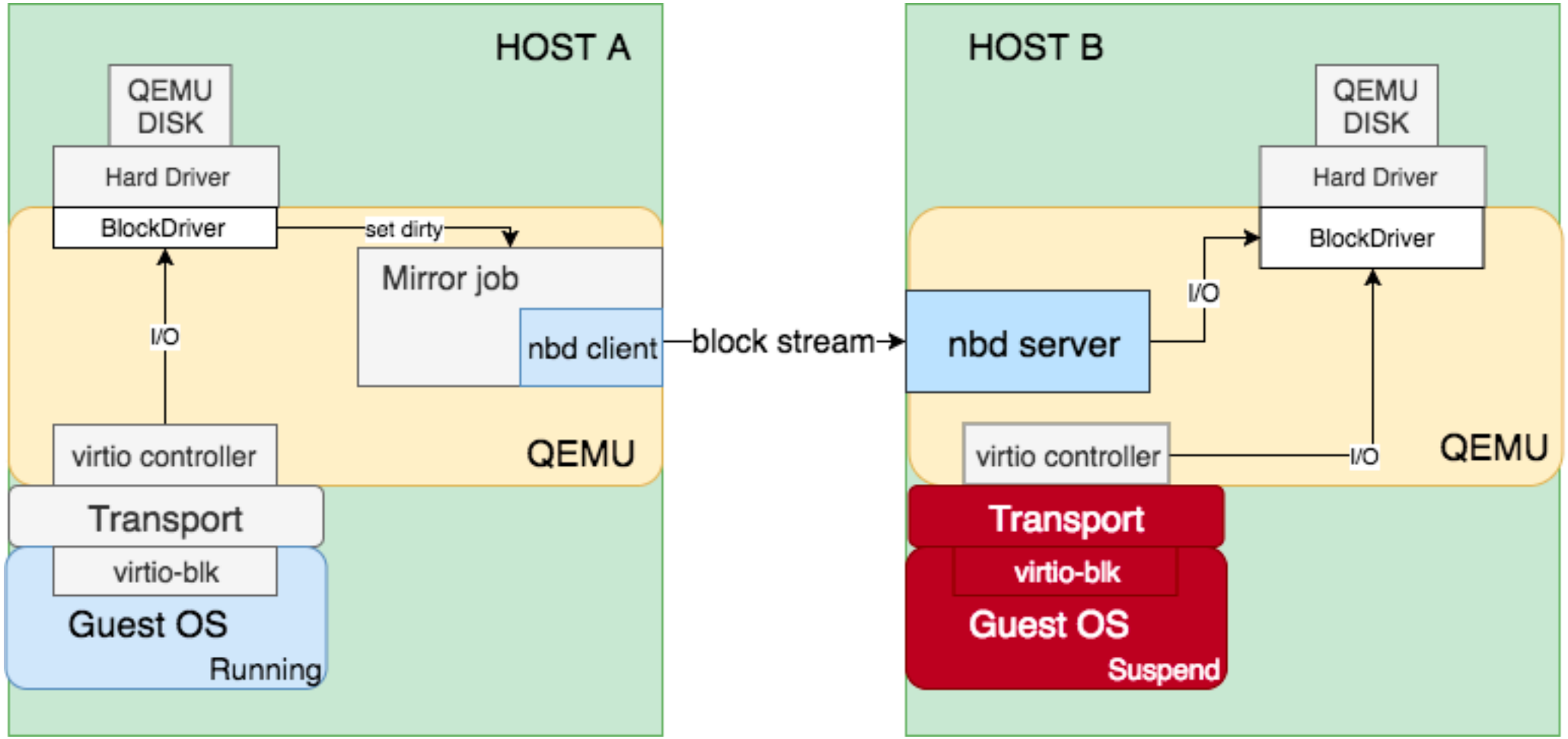
主备机介绍

- 通过为一台基于本地存储的虚拟机在另外一台宿主机上创建一台备份机的方式实现本地磁盘的高可用。主备两台虚拟机之间磁盘数据保持实时同步。当主虚拟机所在宿主机宕机之后，可以切换到备份虚拟机，从而实现本地存储虚拟机的高可用



主备机原理介绍

利用了qemu dirver mirror的功能让数据保持同步状态



主备机注意事项

- 主备机的虚拟机会同时占用两份资源

- 添加备份机有一段时间的数据同步期，期间内不能进行主备切换

- 开启主备机对虚拟机I/O的性能有一定的损耗

- 有本地盘快照的虚拟机不允许创建主备机

云联万维
YUNION

Thanks
Q&A

携手同行 共赢未来

云联万维 YUNION | 智能多云领导者