Dedicated vCPU Guest，Overcommit cpu-pm Feature Test Results

1. 结论，限制，有效性

对于企业云主机（绑核机器，isolate cpu），使能 cpu-pm 特性，会提升企业主机的性能稳定性和性能，并降低功耗，这些提升是小幅度的改进，并不会产生性能的 boost；此特性不能应用与非绑核的主机，否则会导致 vm 的 latence 增高，并降低整体的性能；使能该特性后，跨版本（向低版本）迁移变得不可能；

建议对小批量 x6 机器的增量企业云主机灰度上线，并长期观察效果；后期再根据结果进一步决策；

2. 兼容性

Host + guest	**boot up**	**analyze**
kernel xxx + qemuxxx（support overcommit and hint dedicated）+ libvirt xxx + guest OS centos7.6	Success, warning:qemu-kvm: kvm: guest stopping CPU not supported: Invalid argument	KVM_CAP_X86_DISABLE_EXITS 使能失败，代码依然会 return 0， 不会影响其他执行路径； VM 会以等同 cpu-pm=off 的配置启动
kernel xxx + QEMU emulator version xxx(qemu-kvmxxxx) + libvirt-xxxx + vm guest xxx	Fail，error：qemu-kvm: -overcommit: invalid option；unsupported configuration:	xml 配置不能保存；即使保存，也不能启动；
Src Host	**Dst**	**Result**
kernelxxx + qemuxxx（support overcommit and hint dedicated and ON）+ libvirt xxx + guest OS centosxxx	xxx qemuxxx（support overcommit and hint dedicated and ON）libvirtxxx	qemu-kvm: kvm: guest stopping CPU not supported: Invalid argument 回退到内核不支持的模式，系统正常，feature disabled
kernel xxx + qemuxxx（support overcommit and hint dedicated and ON）+ libvirt xxx + guest OS centosxxx	Xxx qemu-kvmxxx-xksyun.el7 libvirtxxxx	qemu-kvm: -overcommit: invalid option,迁移失败
cpu-pm 不支持 host-passthrough 类型的虚机	**如要支持，需 backport 相关 patch，并测试**	**无支持计划**

3.Performance 收益， Unix Bench

core：72， frequency：3300.158 ，all VM 8G memory with 2M hugepage backend， isolcpus=4-17,22-35,40-53,58-71 nohz_full=4-17,22-35,40-53,58-71 rcu_ncbs=4-17,22-35,40-53,58-71 16 cores reserved for host no isolating， 4 core reserved for spdk；

13 Virtual Machine, 12 with same configure; 1 VM with no cpu bindings

VM name	**vcpu num=4**	**单核 cpu-pm=on**	**多核 cpu-pm=on**	**单核 cpu-pm=off**	**多核 cpu-pm=off**	**Result**
Vm1	<vcpupin vcpu='0' cpuset='10'/> <vcpupin vcpu='1' cpuset='46'/> <vcpupin vcpu='2' cpuset='11'/> <vcpupin vcpu='3' cpuset='47'/>					**1. 性能有提升** CPU-pm=on 单核（cpu-pm=on）Ave1=xxx 多核（cpu-pm=on）Ave2=xxx cpu-pm=off 单核（cpu-pm=off）Ave3=xxx 多核（cpu-pm=off）Ave4=xxx Ave1/Ave3 = 104.0% Ave2/Ave4 = 102.40% **2. 使能 cpu-pm 后，性能更稳定** 标准方差： cpu-pm=on STDEV1 = 34.50 STDEV2 = 35.11 cpu-pm=off STDEV3 = 61.06 STDEV4 = 259.28 **3. 在特定场景下，预期功耗更低，但受限于工具，无实际数字** **4. 每轮对比测试性能上整体会有波动，但是上述结论不变，性能提升和方差可能会有差异，趋势不变；受限**
Vm2	<vcpupin vcpu='0' cpuset='12'/> <vcpupin vcpu='1' cpuset='48'/> <vcpupin vcpu='2' cpuset='13'/> <vcpupin vcpu='3' cpuset='49'/>					
Vm2	<vcpupin vcpu='0' cpuset='14'/> <vcpupin vcpu='1' cpuset='50'/> <vcpupin vcpu='2' cpuset='15'/> <vcpupin vcpu='3' cpuset='51'/>					
Vm4	<vcpupin vcpu='0' cpuset='7'/> <vcpupin vcpu='1' cpuset='43'/> <vcpupin vcpu='2' cpuset='8'/> <vcpupin vcpu='3' cpuset='44'/>					
Vm4	<vcpupin vcpu='0' cpuset='23'/> <vcpupin vcpu='1' cpuset='59'/> <vcpupin vcpu='2' cpuset='24'/> <vcpupin vcpu='3' cpuset='60'/>					
Vm6	<vcpupin vcpu='0' cpuset='25'/> <vcpupin vcpu='1' cpuset='61'/> <vcpupin vcpu='2' cpuset='26'/> <vcpupin vcpu='3' cpuset='62'/>					
Vm7	<vcpupin vcpu='0' cpuset='27'/> <vcpupin vcpu='1' cpuset='63'/> <vcpupin vcpu='2' cpuset='28'/> <vcpupin vcpu='3' cpuset='64'/>					
Vm8	<vcpupin vcpu='0' cpuset='29'/> <vcpupin vcpu='1' cpuset='65'/> <vcpupin vcpu='2' cpuset='30'/> <vcpupin vcpu='3' cpuset='66'/>					

Vm9	<vcpupin vcpu='0' cpuset='31'/> <vcpupin vcpu='1' cpuset='67'/> <vcpupin vcpu='2' cpuset='32'/> <vcpupin vcpu='3' cpuset='68'/>					**于时间，所列数字为其中一轮对比数字。**
Vm10	<vcpupin vcpu='0' cpuset='33'/> <vcpupin vcpu='1' cpuset='69'/> <vcpupin vcpu='2' cpuset='34'/> <vcpupin vcpu='3' cpuset='70'/>					
Vm11	<vcpupin vcpu='0' cpuset='16'/> <vcpupin vcpu='1' cpuset='52'/> <vcpupin vcpu='2' cpuset='17'/> <vcpupin vcpu='3' cpuset='53'/>					
Vm12	<vcpupin vcpu='0' cpuset='5'/> <vcpupin vcpu='1' cpuset='41'/> <vcpupin vcpu='2' cpuset='6'/> <vcpupin vcpu='3' cpuset='42'/>					
Vm13	4 vcpu No bingings					

4.Patches

kvm-support-overcommit-cpu-pm-on-off	0001-kvm-support-overcommit-cpu-pm-on-off.patch
kvm-add-call-to-qemu_add_opts-for-overcommit-option	0002-kvm-add-call-to-qemu_add_opts-for-overcommit-option.patch
qemu-update-linux-header-to-kernel-c82	0003-qemu-update-linux-header-to-kernel-c82.patch
target-i386-kvm.c-Handle-renaming-of-KVM_HINTS_DEDIC	0004-target-i386-kvm.c-Handle-renaming-of-KVM_HINTS_DEDIC.patch
support guest access CORE cstate	0001-i386-kvm-support-guest-access-CORE-cstate.patch
Kernel 4.18	Already support; 766d3571d8e50d3a73b77043dc632226f9e6b389 b31c114b82b2b55913d2cf744e6a665c2ca090ac caa057a2cad647fb368a12c8e6c410ac4c28e063 4d5422cea3b61f158d58924cbb43feada456ba5c b51700632e0e53254733ff706e5bdca22d19dbe5 6c6a2ab962af8f197984c45d585814f9839e86d5
Kernel 3.10-327	No support

5. XML changes

```
<domain type='kvm' xmlns:qemu='http://libvirt.org/schemas/domain/qemu/1.0'>
  ......
  <qemu:commandline>
    <qemu:arg value='-overcommit'/>
    <qemu:arg value='cpu-pm=on'/>
  </qemu:commandline>
</domain>
```

6.内核对 over-commit 不支持，qemu 支持，qemu-kvm: kvm: guest stopping CPU not supported: Invalid argument;

```
kernel xxxx + qemu xxx        （support overcommit and hint dedicated）
htop查看，cpu-pm未被设置：

2021-08-31T02:39:54.168673Z qemu-kvm: kvm: guest stopping CPU not supported: Invalid argument
2021-08-31T02:39:55.816548Z qemu-kvm: -device cirrus-vga,id=video0,bus=pci.0,addr=0x2: warning: 'cirrus-vga' is deprecated, please use a diffe
qmp cont start 1630377595877

int kvm_arch_init(MachineState *ms, KVMState *s)
{
    uint64_t identity_base = 0xfffbc000;
    uint64_t shadow_mem;
    int ret;
    struct utsname utsname;
    ........

    if (enable_cpu_pm) {
        int disable_exits = kvm_check_extension(s, KVM_CAP_X86_DISABLE_EXITS);
        int ret;

        if (disable_exits) {
            disable_exits &= (KVM_X86_DISABLE_EXITS_MWAIT |
                              KVM_X86_DISABLE_EXITS_HLT |
                              KVM_X86_DISABLE_EXITS_PAUSE);
        }

        ret = kvm_vm_enable_cap(s, KVM_CAP_X86_DISABLE_EXITS, 0,
                                disable_exits);
        if (ret < 0) {
            error_report("kvm: guest stopping CPU not supported: %s",
                         strerror(-ret));
        }
    }

    return 0;
↵

KVM_CAP_X86_DISABLE_EXITS使能失败，代码依然会return 0， 不会影响其他执行路径； guest成功启动；
```

7. kernel CPU-PM

```
内核包含四个配置，
KVM_X86_DISABLE_EXITS_MWAIT  --- 金山bios关闭，不支持；
KVM_X86_DISABLE_EXITS_HLT    --- 支持
KVM_X86_DISABLE_EXITS_PAUSE  --- 默认开启（和之前无变化）
KVM_X86_DISABLE_EXITS_CSTATE --- 默认不开启，使能cpu-pm后，开启

qemu kvm_vm_enable_cap预期使能KVM_X86_DISABLE_EXITS_HLT，KVM_X86_DISABLE_EXITS_PAUSE，KVM_X86_DISABLE_EXITS_CSTATE
```

8. 上线规划