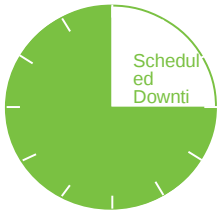
Giant VM:
I just want to use one vm

Liang Yan (lyan)
SUSE Labs

Lee Martin
SAP Alliance



Outline

1. Background

2. User Cases and Challenges
   VDI
   SAP HANA
   Edge computing

3. Brain Storm for Optimization

4. Q&A

Background

What is Giant VM

Giant is relative to HOST

Instead of using a mount of VMs, Only use quite a limit number in Host, sometimes may just one.

This VM will take full usage of Host in every way, CPU, Memory, IO devices.

Why do we want it?

Virtualization still has advantage on management even we do not think about scalability here.

VDI

VM should be able to do anything bare-metal can do, works like one.

SAP HANA

VM workload is so heavy that it could not allow so many VMs running at the same time.

Edge Computing

The host resource is limited that it could not support so many VMs running at the same time.

User Cases and Challenges

VDI + decent performance

This is an visualize application, has graphic performance and 3D requirement.

For some reasons, customer still wants to keep their legacy code which was 11sp3, so we need to come up with a solution that .

Host 12SP4
Guest 11SP3

VDI + decent performance

"Passthrough" everything inside VM at boot time.

1. Usb(3.x) redirection /keyborad&Mouse

-device usb-host,hostbus=3,hostaddr=11

2. Physical DVD/CD/BD
hde and virtio-scsi

-drive file=/dev/sr2,if=none,id=scsicd
-device
virtio-blk,drive=scsicd,logical_block_size=2048,physical_block_size=2048

3. Serial port redirection
-chardev tty,path=/dev/ttyUSB0,id=hostusbserial

VDI + decent performance

4. Network

SR-IOV

Vhost-net + multi-queue

5. Disk

A little bit tricky, could not use LV and passthrough.

Did a lot test on disk cache mode and iothreads setup

6 VGA

Dual monitor support

VGA PCI-passthrogh denied

Spice + QXL

Xserver + Xclient

SAP HANA KVM

https://etherpad.nue.suse.com/p/SAP-KVM-network-experiment

- Single-VM, Multi-VM, SAP HANA scale-out

- SAP Certification process (OLTP, OLAP, IO, Memory, CPU...)

- NUMA, network latency

- Cross functional/cross team project

- May 2018 Certification [1] for Single-VM, 2TB, Haswell CPU on SLES12 SP2

- KVM != KVM

- Customer Demand

Footnote:

https://documentation.suse.com/sbp/all/pdf/SBP-HANAonKVM-SLES12SP2_color_en.pdf

SAP HANA KVM

https://etherpad.nue.suse.com/p/SAP-KVM-network-experiment

Some test ideas:

multiple queue numbers with different values

pin vhost-pids

isolcpu vs libvirt-pin LPcores for VM

different vm setup: mem size, disable virtio-ballon, vcpus

vhost, nr-queues == nr-vcpus, pin IRQs to vcpus

MTU size 1500 vs 9000

OVS-DPDK

SRIOV

Edge Computing/IoT Devices

Computing is remotely and close to end user, it has higher security requirement.

Hardware resource is limited, need less and lighter VMs

Distribute management, sometimes it may need boot up and shut down frequently.

Real time requirement, at least low latency.

Edge Computing/IoT Devices

VNF in NFV
CDN
Auto Drive

Is virtualization necessary? Yes and No
Does people want it? Sure.

We just think how could make virtualization fit Edge Compute better.

Lighter, Real time, Boot up faster ...

Optimization Brain Storm

Object Analysis

What does hardware looks like today?(X86-Intel)

NUMA(Non-Uniform Memory Access)

North Bridge/South Bridge
PCIe Controller and Memory Controller are moved into CPU, per-node
LAPIC, IOAPIC

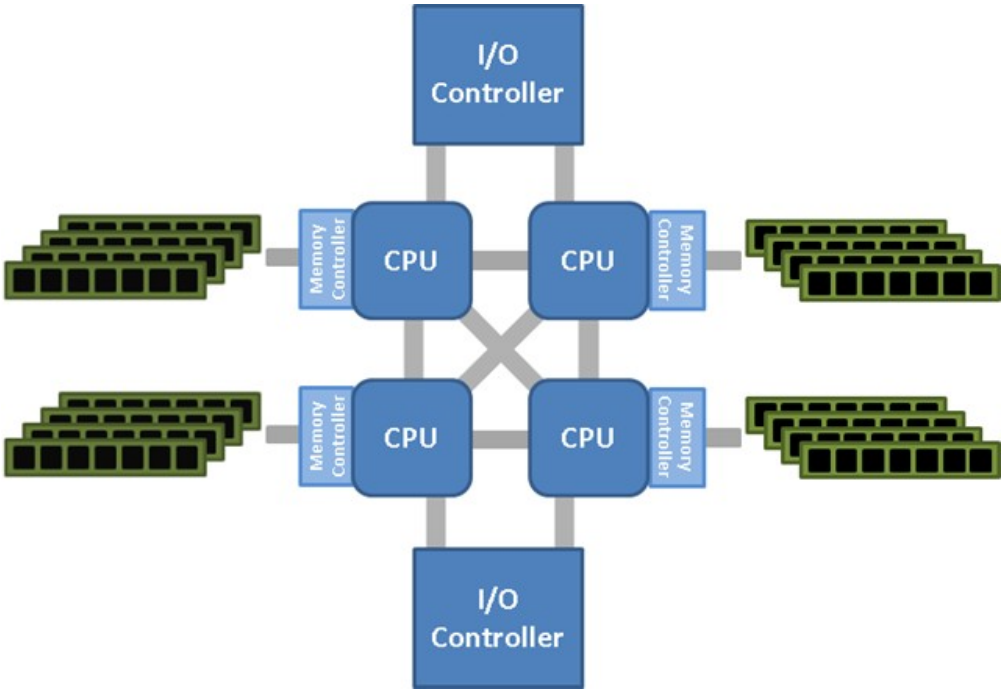
Hardware Implementation for Virtualization

VT-x/EPT
IOMMU

IO: GPU/NIC/Disk

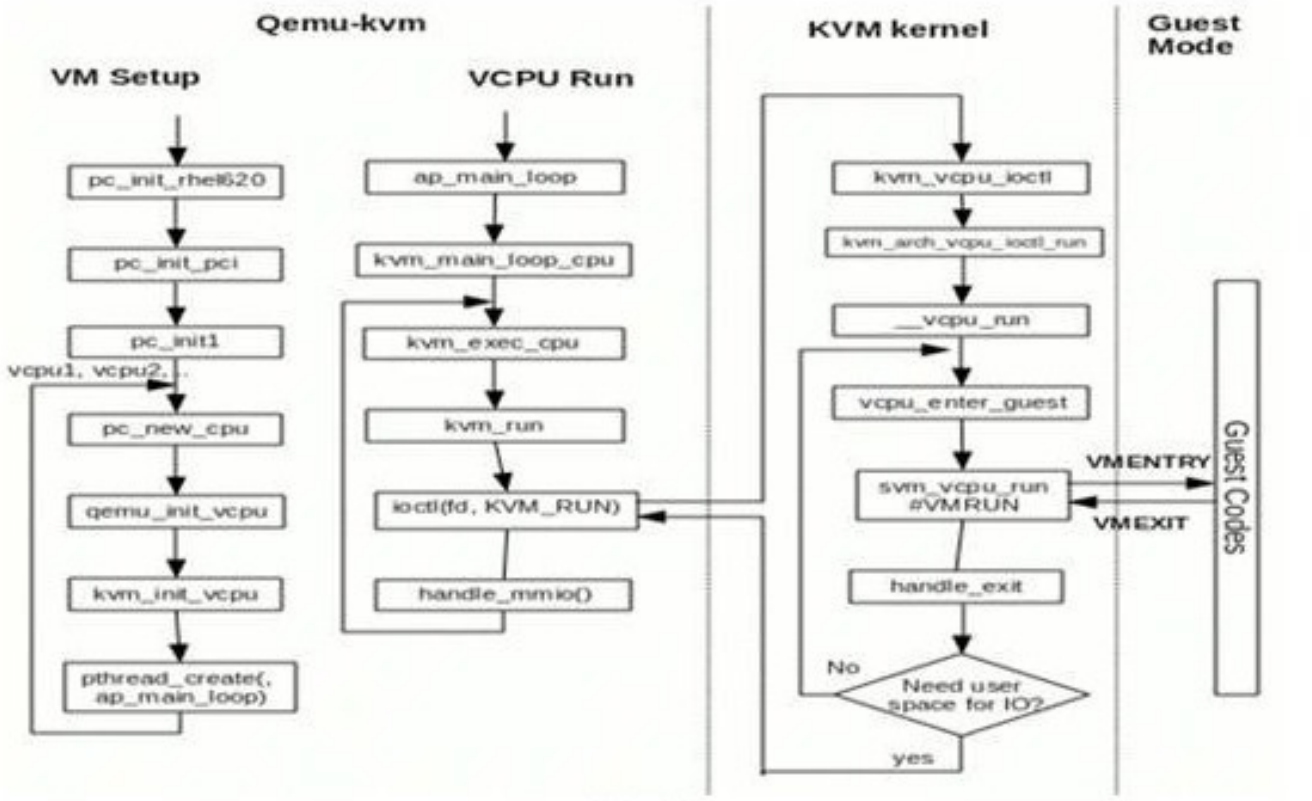
SRIOV
MultipQueue

Object Analysis



https://www.sqlskills.com/blogs/jonathan/understanding-non-uniform-memory-accessarchitectures-numa/

Object Analysis



Object Analysis

```
[root@localhost]# perf stat -e 'kvm:*' -a sleep 1h
^Csleep: Interrupt
Performance counter stats for 'sleep 1h':
  1,880 kvm:kvm_entry                                             [100.00%]
     20 kvm:kvm_pio                                                [100.00%]
      0 kvm:kvm_cpuid                                              [100.00%]
    596 kvm:kvm_apic                                               [100.00%]
  1,934 kvm:kvm_exit                                               [100.00%]
    284 kvm:kvm_inj_virq                                           [100.00%]
      3 kvm:kvm_inj_exception                                     [100.00%]
    602 kvm:kvm_page_fault                                       [100.00%]
      0 kvm:kvm_msr                                                 [100.00%]
    102 kvm:kvm_cr                                                  [100.00%]
    260 kvm:kvm_pic_set_irq                                      [100.00%]
    156 kvm:kvm_apic_ipi                                            [100.00%]
    292 kvm:kvm_apic_accept_irq                                [100.00%]
    292 kvm:kvm_eoi                                                 [100.00%]
      0 kvm:kvm_pv_eoi                                            [100.00%]
      0 kvm:kvm_invlpga                                             [100.00%]
      0 kvm:kvm_skinit                                                [100.00%]
    979 kvm:kvm_emulate_insn                                     [100.00%]
    618 kvm:vcpu_match_mmio                                    [100.00%]
    635 kvm:kvm_userspace_exit                                 [100.00%]
    276 kvm:kvm_set_irq                                             [100.00%]
    276 kvm:kvm_ioapic_set_irq                                   [100.00%]
      4 kvm:kvm_msi_set_irq                                         [100.00%]
    277 kvm:kvm_ack_irq                                             [100.00%]
  1,627 kvm:kvm_mmio                                              [100.00%]
    762 kvm:kvm_fpu                                                   [100.00%]
      0 kvm:kvm_age_page                                          [100.00%]
      0 kvm:kvm_try_async_get_page                           [100.00%]
      0 kvm:kvm_async_pf_doublefault                         [100.00%]
      0 kvm:kvm_async_pf_not_present                       [100.00%]
      0 kvm:kvm_async_pf_ready                                   [100.00%]
      0 kvm:kvm_async_pf_completed
1.895712367 seconds time elapsed
```

Object Analysis

```
# ./perf kvm stat report --event=vmexit
Analyze events for all VCPUs:
         VM-EXIT    Samples    Samples%     Time%    Avg time
       APIC_ACCESS      65381    66.58%     5.95%     37.72us ( +-   6.54% )
EXTERNAL_INTERRUPT      16031    16.32%     3.06%     79.11us ( +-   7.34% )
             CPUID       5360     5.46%     0.06%      4.50us ( +-  35.07% )
               HLT       4496     4.58%    90.75%   8360.34us ( +-   5.22% )
     EPT_VIOLATION       2667     2.72%     0.04%      5.49us ( +-   5.05% )
 PENDING_INTERRUPT       2242     2.28%     0.03%      5.25us ( +-   2.96% )
     EXCEPTION_NMI       1332     1.36%     0.02%      6.53us ( +-   6.51% )
    IO_INSTRUCTION        383     0.39%     0.09%     93.39us ( +-  40.92% )
         CR_ACCESS        310     0.32%     0.00%      6.10us ( +-   3.95% )
Total Samples:98202, Total events handled time:41419293.63us.
```

One Vs Scale

No overcommits:

Disable overcommit for vcpu and memory

Less competition between qemu processes, but same situation for kernel vcpu.

Memory:

Disable switch

Disable virtio-ballon

Disable KSM
systemctl disable ksm
systemctl disable ksmtuned

NUMA

It is complicated with very large VMs, where vCPUs and memory cannot be held in single NUMA node.

The number of NUMA nodes inside the VM directly effects the VM's performance

– Kernel compile on 120-vcpu VM on host:
– 1 NUMA node: 192 seconds
  4 NUMA nodes: 169 seconds

This is because many locks are per-node, and more nodes means more choices for guest kernel and less lock.

VM_EXIT

Kernel main-loop

1. CPU-PIN
2. Huge-Page 1G
3. Pass-through/SRIOV
4. virtio/vhost device
5. irq affinity for host device, try to bind the cpus that not bind vcpu
6. idle=poll for HLT vm_exit

iothread

QEMU main-loop and global mutex

Iothread
QEMU thread + VCPU threads + worker threads

Big qemu lock for thread Sync

qemu -object iothread,id=iothread0

QEMU block layer used inside of IOThread
IOThread runs an AioContext event loop

iothread

1. Add iothreads for virtio devices

2. Enable multiqueue

3. Lock-holder Preemption issue
PLE(pause-loop-exit), good for small vcpu numbers

linux-lyan:~/code/test # rmmod kvm-intel
linux-lyan:~/code/test # modprobe kvm-intel ple_gap=128
linux-lyan:~/code/test # cat /sys/module/kvm_intel/parameters/ple_gap
0

Others?

1. Userspace driver

2. Real Time Virtualziation

Q&A

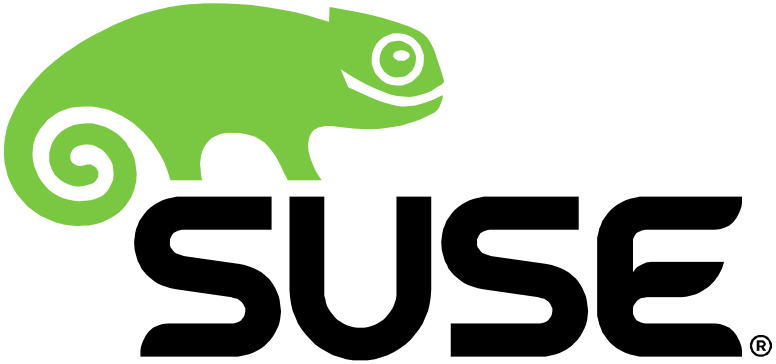
Question?

Thank you.

REFERENCE

https://documentation.suse.com/sles/12-SP4/html/SLES-all/article-vt-best-practices.html,

https://www.spice-space.org/usbredir.html



Corporate Headquarters
Maxfeldstrasse 5
90409 Nuremberg
Germany

+49 911 740 53 0 (Worldwide)
www.suse.com

Join us on:
www.opensuse.org