KVM Performance – Vhost Scalability

Shirley Ma, IBM
John Fastabend, Intel
- Per virtio_net device vhost thread
• The number of vhost threads depend on the number of virtio_net device per VM.
• The number of vhost threads depend on the number of VMs.

• **Pros:**
  - Easy VM based cgroup control

• **Cons:**
  - Scheduling:
    - None numa-awareness
  - Performance
    - When increasing the number of virtio_net devices per VM, or increasing the number of VMs, the performance does not scale
    - When the number of vhost threads are larger than the number of host cpus, there are lots of context switch overhead
    - vhost TX and RX are shared, so TX and RX work can’t be processed simultaneously
- Per CPU vhost thread
• Share vhost thread among VMs when number of VM/virtio_net devices greater than the number of CPUs to avoid scheduling overhead

• Split vhost TX and RX work based on the workload

• NUMA-awareness scheduling
  – The vhost thread is picked up based on idelest allowed cpu in local numa node

• Cgroup control
  – The vhost thread is attached to the cgroup on the VM which the work comes from
  – When idle/need_sched, the vhost thread is detached from the previous cgroup.
TCP_RRs Results

Remote Host to Guest

TCP_RRs Efficiency

Remote Host to Guest

Number of instances

Number of instances

MQ Baseline
Per-cpu-v host
MQ-2

MQ Baseline
Per-cpu-v host
MQ-2
TCP_RRs Results

Remote Host to 4 Guest

MQ Baseline
Per-cpu-v-host
MQ-2

TCP_RRs Efficiency

Remote Host to 4 Guest

MQ Baseline
Per-cpu-v-host
MQ-2
TCP_RRs Results

Remote Host to 12 Guest

TCP_RRs Efficiency

Remote Host to 12 Guest
TCP_RRs Results

Remote Host to 24 Guest

TCP_RRs Efficiency

Remote Host to 24 Guest
• Per CPU Vhost thread shows win with many Vms
• Crossover occurs at about

• Similar trends for
  – TCP_STREAM
  – UDP_STREAM, UDP_RR
  – Local (east-west) traffic
Thanks to Tom Lendacky for perf data!

Questions/Comments?

http://github.com/jrfastab/
• IBM x3650-M2
  - Intel E5530 2.4 Ghz Nehalem processors
  - dual socket with 4 cores/socket
  - Hyperthreading disabled
  - NIC X540-SR1 10GbE (ixgbe)