1 Ping-pong Performance

Working with Solarflare we measured performance of a simple ping-pong test using a Solarflare 7122 NIC in many configurations including using vNICs & vSwitches, PCI pass-through, and combining pass-through with Solarflare's Onload kernel stack bypass. The measurements were taken with generic RHEL distributions with no special tuning other than as noted below. Systems were connected back-to-back with no intervening Ethernet switch. Though measurements were taken using UDP, similar performance can be had with TCP albeit with slightly more overhead.

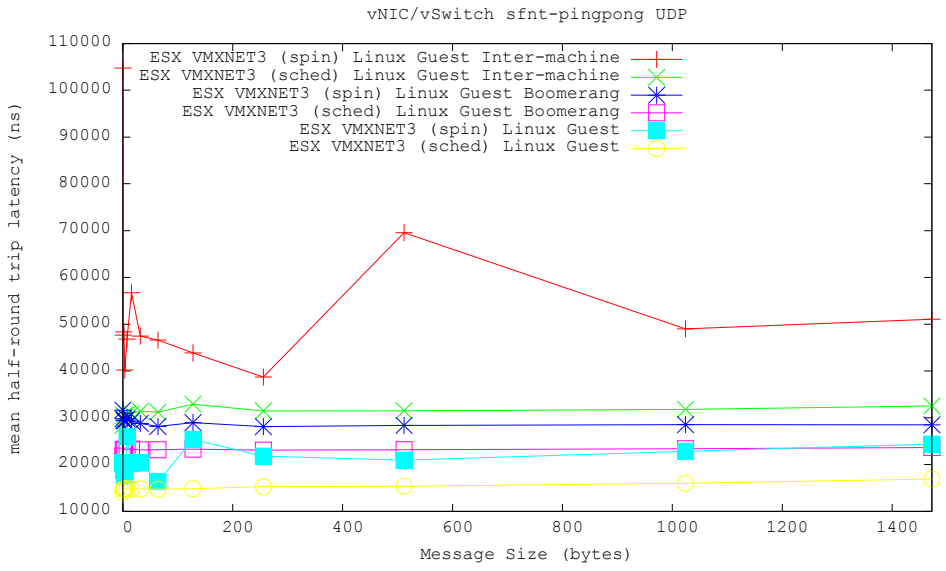
For apples-to-apples comparisons we ran the client against both a native Linux server with stack bypass for the best possible result and a server using the same configuration as the client (aka symmetric). This gives a feel for possible performance when virtualizing either one side or both sides of a client/server application.

The figure-of-merit is "half round-trip" time in nanoseconds; a round trip is twice the time shown.

1.1 vNIC/vSwitch Results

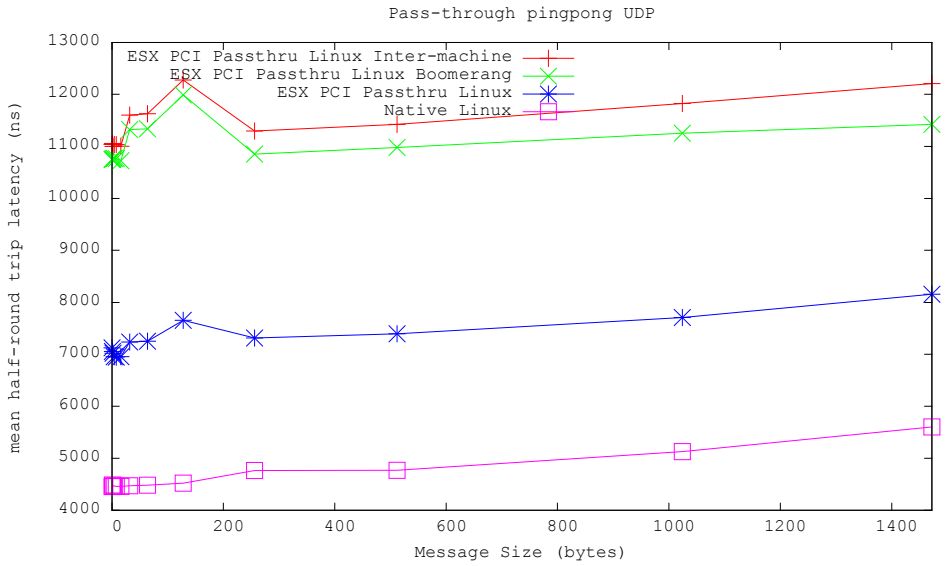
Good performance from a standard virtual machine using a virtual NIC connected to a virtual switch with a physical NIC (pNIC) has been a focus of VMware since ESX was created. By default ESXi is tuned for throughput, so for these tests NIC interrupt coalescing was disabled with "`ethtool -C vmnic4 adaptive-rx off rx-usecs 0`".

Recently many customers have tried to improve performance by preventing the vmkernel from rescheduling a virtual machine's vCPUs when idle by setting "`monitor_control.halt_desched=false`", noted as "(spin)" in the graphs. As shown, this setting actually degrades latency and introduces additional jitter.



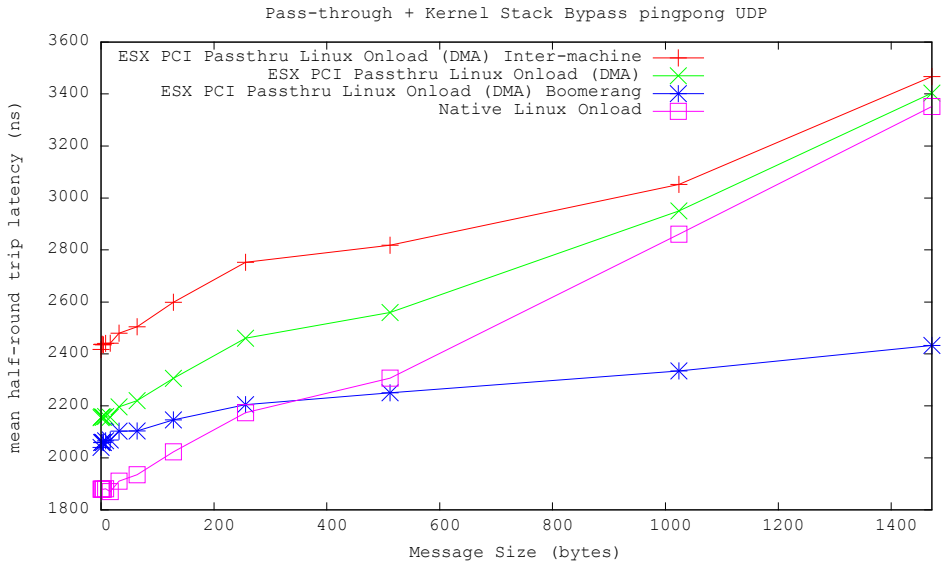
1.2 Pass-through Performance

Using pass-through reduces round-trip times significantly. This configuration represents a standard physical use case (the red line) and virtualized use-cases where kernel stack bypass isn't available, such as on Windows guests.



1.3 Kernel Stack Bypass Performance

As expected, using pass-through plus kernel stack bypass produces best results. For small messages 5µs round-trips are achievable between virtual machines on different servers. A possible surprise is how well bypass between virtual machines on the same system (boomerang) perform, with sub-5µs round-trips up to the Ethernet MTU. At larger packet sizes virtual machine performance is almost identical to physical configurations.



As measured, for symmetric virtual machine configurations expected round-trips are roughly 60µs between standard vNICs, 25µs with pass-through, and 5-7µs using pass-through and bypass.

2 Appendix – Solarflare Testing Recipe

- System Hardware:
  Dell PowerEdge R720
  CPUs (2 sockets): Intel E5-2643v2 CPUs @ 3.5GHz
  Cores per package: 6
  Hyperthreading: Disabled
  Bios Version: 2.2.2 (1/16/2014)
  32 GB RAM
- Solarflare Hardware:
  SFN7122 2-spigot 10GbE which reports to ESXi as a SFC9120
  Card is configured with 8 PFs
  Machines are connected back-to-back with
- Solarflare Firmware:
  `driver: sfc`
  `version: 41.0.6734A`
  `firmware-version: 4.2.2.1003 rx0 tx0`
  `bus-info: 0000:42:00.0`
- Solarflare Driver Tuning: (disable interrupt moderation & coalescing)
  `ethtool –C vmnic4 adaptive-rx off rx-usecs 0`
- VM Configuration:
  `2vCPUS`
  `numa.nodeAffinity: 1` (the NIC's PCI was on CPU package 1)
  `monitor_control.halt_desched: false` (when using pass-through)
  `monitor_control.halt_desched: true & false` (w/ vNIC depending on test)

- Pass-through: Solarflare 10GbE
  vNIC: vmxnet3 to a vSwitch with Solarflare 10GbE
- Linux Configuration
  RHEL 6.5 physical – `ethtool –C eth0 rx-usecs 0 adaptive-rx off`
  RHEL 6.4 virtual – `ethtool –C eth0 rx-usecs 0 adaptive-rx off`
- Test program
  sfnt-pingpong version 1.5.0 (from openonload.org)
  EF_PIO=0 (use DMA instead of PIO w/onload)