



TCP/Generic Segmentation Offload and Its Application in Xen

Herbert Xu

Principal Software Engineer

Red Hat Asia Pacific

What is TSO?

- Faster Ethernet (Gigabit) => higher CPU load:
 - 1500-byte Ethernet MTU set in 70's.
 - Amount of data per second 100 times higher.
 - CPU load per second 100 times higher.
 - Jumbo frames (9000-byte MTU or higher) help.
 - Hard to deploy due to PMTU issues.

What is TSO?

- Solution: Offload segmentation to NIC:
 - Effectively increases local MTU to ~64KB.
 - Stateless offload => easy OS support (cf TOE).
 - Supported by major NIC vendors and OSes.
 - Complements checksum offload.
 - Greatly reduces MTU-related CPU load.

TSO in Linux

- Added in August 2002 by Alexey Kuznetsov.
- Original version was incredibly fast.
- It ignored congestion window requirements :)
- Stable since 2.6.16.10/2.6.17.
- Enabled by default on TG3 and E1000.

Xen Paravirtual Networking

- Simulates a NIC in software.
- dom0/netback \Leftrightarrow domU/netfront.
- Operates on Ethernet packets.
- Uses ring buffer like a real NIC.
- Uses page flipping unlike a real NIC.
- Checksum offload support.

Xen Paravirtual Networking

- Performance problem with TCP:
 - loopback in domU: 5543.91Mb/s
 - domU => dom0: 1228.08Mb/s
 - domU => domU: 323.91Mb/s
- Loopback performance on par with baremetal.
- domU/domU performance less than wire speed.

Xen Paravirtual Networking

- $\text{MTU}(\text{lo}) = 16436$, $\text{MTU}(\text{eth0}) = 1500$.
- Change $\text{MTU}(\text{eth0})$ to 16436?
- 16436 bytes > 4KB (page size), requires SG.
- Change $\text{MTU}(\text{lo})$ to 1500.
- Throughput down to 2178.86Mb/s.

Xen Paravirtual Networking

- Solution: Implement SG for Xen.
- domU => dom0 throughput reaches 3097.49Mb/s with MTU of 16436.
- Comparable with 5543.91Mb/s on lo.
- Remaining difference due to longer code path.
- Higher MTU unrealistic due to PMTU.

Xen Paravirtual Networking

- Solution: TCP Segmentation Offload.
- No segmentation at all within Xen.
- Effective MTU of ~64KB within Xen.
- domU => dom0 throughput: 3208.41Mb/s.
- domU => domU throughput: 1678.52Mb/s.
- domU/domU has extra copy for memory protection. Can be removed with MMU help.

Generic Segmentation Offload

- Problem: Fails if packet hits non-TSO NIC.
- Solution: Simulate TSO in dom0.
- Add TCP knowledge to generic path?
- Add GSO infrastructure first.
- Packet => GSO => IPv4 => TCP.

Generic Segmentation Offload

- Reality check: Solves more than one problem?
 - TCP/ECN support with TSO.
 - TSO over IPv6.
 - Share code with UFO (UDP Fragment Offload).
 - Potential to enable TSO on all NICs.
 - Support more protocols, e.g., DCCP.

Gory Details of GSO/SG in Xen

- Extend ring buffers as real NICs do.
- Chaining descriptors for SG.
- Add extra descriptor for GSO.
- Same strategy can be used fo TX checksum.
- Maintain compatibility with feature negotiation.

Questions