

Multiqueue NIC Support in Linux

Herbert Xu

Background

- Ethernet speed vs. CPU frequency:

– 10Mb/s	10-40Mhz
– 100Mb/s	100-600Mhz
– 1Gb/s	1Ghz-4Ghz
– 10Gb/s	1Ghz-4Ghz
- A single core can no longer fill a 10Gb/s NIC.
- Multi-core or SMP is required.

Multi-core and Multiqueue

- Multi-core is similar to SMP, needs locking.
- High lock contention reduces CPU efficiency.
- 10Gb/s cannot afford reduced efficiency.
- Solution: multiqueue NICs
 - Each core has its own queue and interrupt.
 - Transmit: CPU chooses queue.
 - Receive: NIC chooses queue.

Support for Multiqueue Receive

- NIC decides which queue to use.
- Usually done by hashing the packet header.
- Only needs to modify driver to support this.
- Multiqueue NAPI requires stack modifications.
- Oct 07: Multiqueue NAPI support added.

Support for Multiqueue Transmit

- Original design:
 - Each netdev corresponds to a qdisc.
 - Each qdisc corresponds to a hardware queue.
- July 07 (Intel):
 - Each qdisc corresponds to many hardware queues.
- Hardware queues no longer point of contention.
- The qdisc lock becomes bottleneck.

Support for Multiqueue Transmit

- July 08 (David S. Miller):
 - Default qdisc (pfifo_fast):
 - Each netdev corresponds to many qdiscs.
 - Each qdisc corresponds to a hardware queue.
 - All other qdiscs remain as before.
- Resolves qdisc lock contention for default qdisc.

TODO

- Queue coordination :
 - Transmit queue needs to align with receive queue.
 - Queue distribution needs to align with processes.
- Support for qdiscs other than default:
 - Need to add multiple queues to each qdisc.
- Virtualisation and user-space networking.

Questions