10GigE optimizations

Kip Macy
16 May 2007
FreeBSD Developer Summit

Lessons learned from cxgb

- Ifnet is showing its age
- Fairness and hardware functionality not well supported
  - if_start needs to yield to avoid starving out potential waiters
  - Newer hardware supports multiple queues
    - Should driver be allowed to manage its own queues?
  - No notion of connection to cpu affinity
- Scheduler can play a large (> 2Gbps) role
- Cache awareness is very important

Implemented optimizations

- Yielding after an empirically determined number of descriptors and pushing remainder to a taskqueue eliminates starvation issues
- Defer mbuf allocation until after rx
  - yielded ~500Mbps improvement in peak throughput

Implemented optimizations II

- Mbuf iovec optionally used on tx
  - Reduces variance in throughput and cpu usage
- Mbuf iovec used on rx
  - No measured gain/loss, will be more useful for LRO
- Receive Side Steering support added to driver
  - Packets for a given TCP connection are all rx'ed on the same cpu
  - Not currently a win due to the absence of any notion of affinity

Planned optimizations

- Near-term - enable LRO in driver, long term push into ifnet
- Either generalize queue management in ifnet or move into driver
- Move mbuf iovec usage into the rest of the stack
  - Short term use for LRO
  - Mid-term use for SCTP
  - Long-term move into TCP and sockets layer