



Improving the performance of a Gigabit Ethernet driver under Linux

DataTAG project meeting
UCL

26th February 2003

Tom Kelly

`ctk21@cam.ac.uk`

CERN

and

Laboratory for Communication Engineering

University of Cambridge

- ⑥ Driver basics
- ⑥ TX interrupt moderation
- ⑥ TCP slow start improvements
- ⑥ Linux NAPI model
- ⑥ Early NAPI results

Linux driver basics - TX

- ⑥ Application system call
- ⑥ Encapsulation in UDP/TCP and IP headers
- ⑥ Enqueue on device send queue
- ⑥ Driver places information in DMA descriptor ring
- ⑥ NIC reads data from main memory via DMA and sends on wire
- ⑥ NIC signals to processor that TX descriptor sent

Linux driver basics - RX

- ⑥ NIC receives packet onto card
- ⑥ NIC places data in main memory via DMA to a free RX descriptor
- ⑥ NIC signals RX descriptor has data
- ⑥ Driver passes frame to IP layer and cleans RX descriptor
- ⑥ IP layer passes data to application

SysKonnnect driver

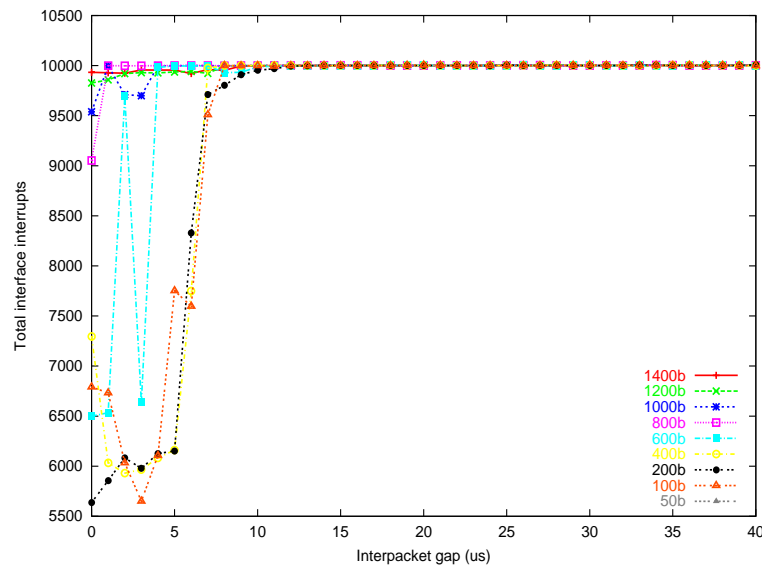
- ⑥ Distributed driver in Linux kernel is old
 - △ splits from SysKonnnect distribution in May 2001
- ⑥ Default is to have TX interrupts on every packet
- ⑥ Linux releases before 2.4.19 have no mechanism to moderate RX interrupts
- ⑥ Driver patch from SysKonnnect better but still has bugs
- ⑥ Stale TX ring possible, but some SMP races fixed

TX Interrupt moderation

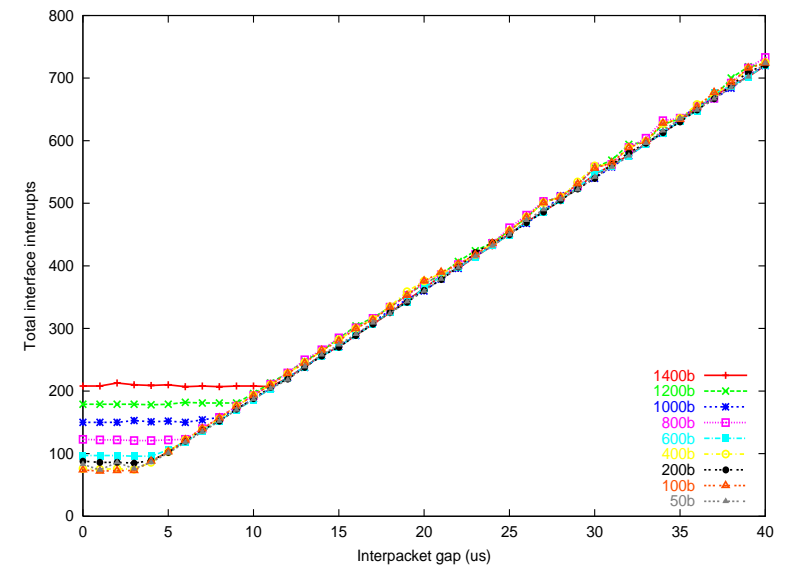
- ⑥ Interrupt with TX completion at most 1800 times a second
- ⑥ Use `netif_wake_queue` to ensure no stale TX queue
- ⑥ Should improve send performance and reduce CPU utilisation
- ⑥ If TX ring fills attempts to clean it are done when frames are transmitted

TX Interrupt moderation - results

- 6 Using TX complete moderation dramatically reduces the number of interrupts



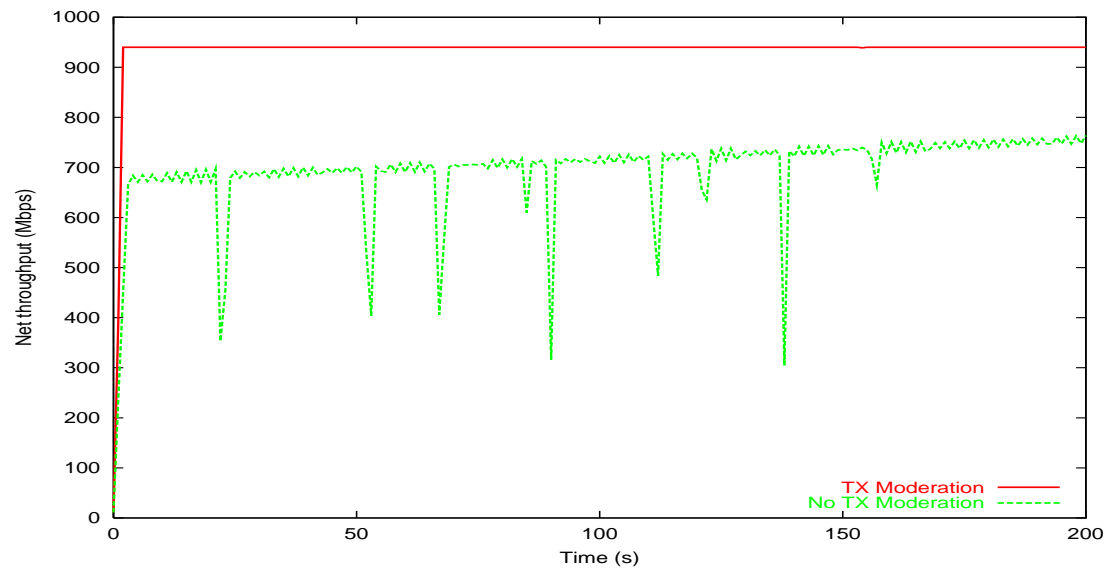
No TX interrupt moderation



TX interrupt moderation (1800i/s)

TCP slow start results

- Without TX interrupt moderation the connection leaves slow start early



- 2.4Ghz machines over DataTAG link
- Sender uses standard TCP with interrupt moderation, BW delay send buffer and 10000 txqueue

Linux NAPI driver model

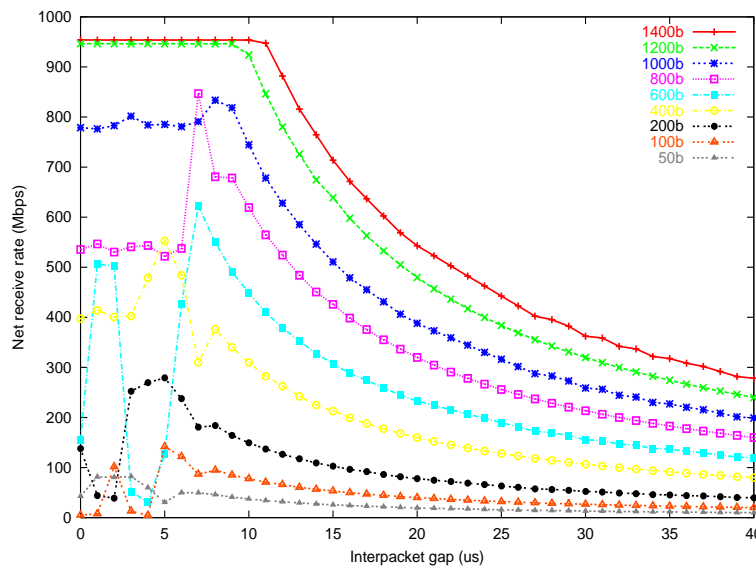
- ⑥ Around for some time in 2.5.x and incorporated in 2.4.20
- ⑥ On receiving a packet, NIC raises interrupt
- ⑥ Driver switches off RX interrupts and schedules RX DMA ring poll
- ⑥ Frames are pulled off DMA ring and is processed up to application
- ⑥ When all frames are processed RX interrupts are re-enabled
- ⑥ Dramatic reduction in RX interrupts under load

Experimental SysKonnnect NAPI driver implemented

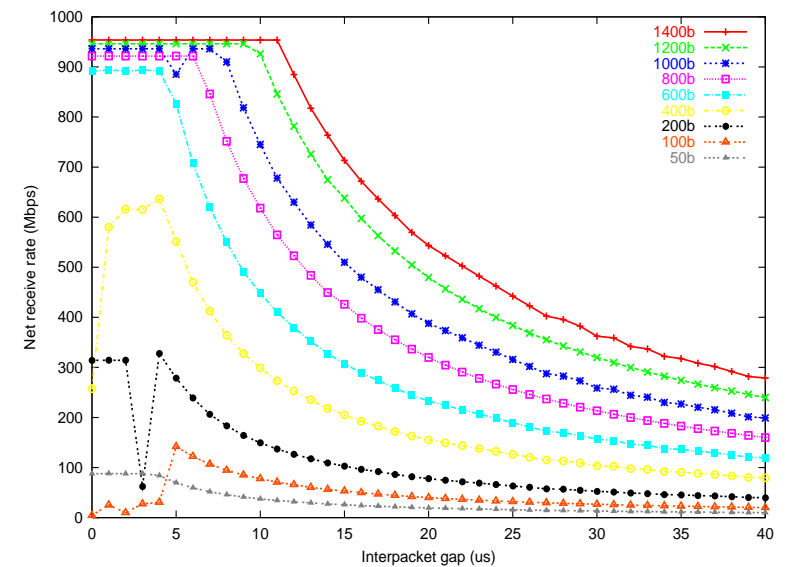
- ⑥ No spec sheet for PCI card ASIC since SysKonnnect was bought by Marvell
- ⑥ Still some RX flagged interrupts appearing; appears benign but makes me suspect there is a bug somewhere
- ⑥ Bottom line is improved performance under heavy load

NAPI receiver results

- ⑥ 2.4Ghz machines connected through router with 2.4.20 sender using TX interrupt moderation



2.4.19 non-NAPI receiver



2.4.20 NAPI receiver

- ⑥ Better throughput for NAPI receiver under load
- ⑥ Some strange behavior with 100b and 50b packets...

Conclusion

- ⑥ SysKonnnect driver not perfect but tangible performance and stability improvements gained
- ⑥ Sending and receiving small packets still hurts
- ⑥ TCP performance much improved with good device driver
- ⑥ Looking to release NAPI SysKonnnect driver soon...