Gbps Open Source Routing

Bengt Gördén
bengan@resilans.se

Resilans AB (Ltd)

- Routing and infrastructure
- Registry
- Open source router
- Filtering software
- System development for web directory services
- Health care system
- .....

KTH

- Royal Institute of Technology in Stockholm
- KTHNOC
  - Operation center for
    - SUNET
    - Nordunet

Links

- http://www.nada.kth.se/~olofh/
- http://www.herjulf.se/
- http://www.linux-kongress.org/2010/slides/lk-2010-10G.pdf
- http://data.guug.de/slides/lk2008/10G_preso_lk2008.pdf
- http://www.iis.se/internet-for-alla/internetfonden/uppdrag-2009#kallkodsroutrar
- bengan@resilans.se

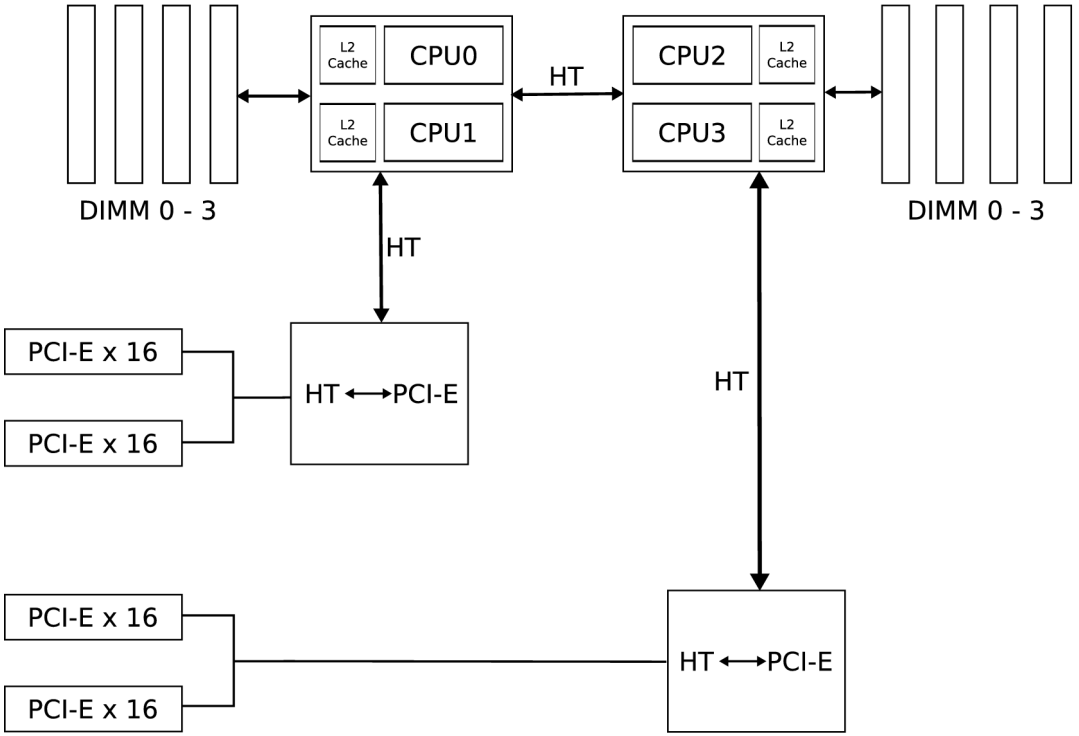
Three different projects

- Sponsored by IIS and Intel
- Project 1
  - Open-source routing at 10Gb/s
- Project 2
  - Multiqueue
- Project 3
  - Separation

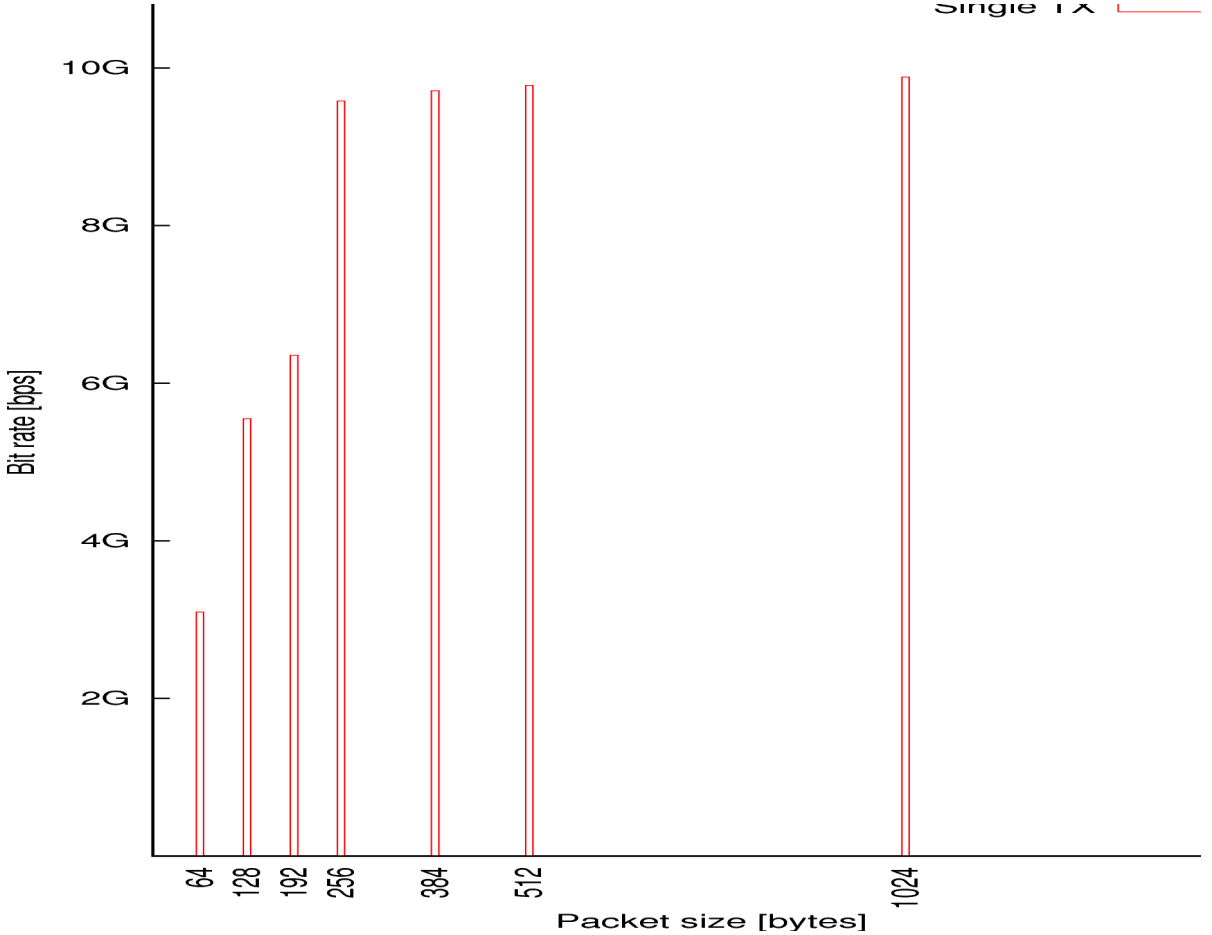
Hardware

- Hardware used
  - XEON 2 x E5630, TYAN S7025 Motherboard
  - AMD 2x2222, Tyan 2915 board
  - Intel cards
    - chipset 82598 and 82599

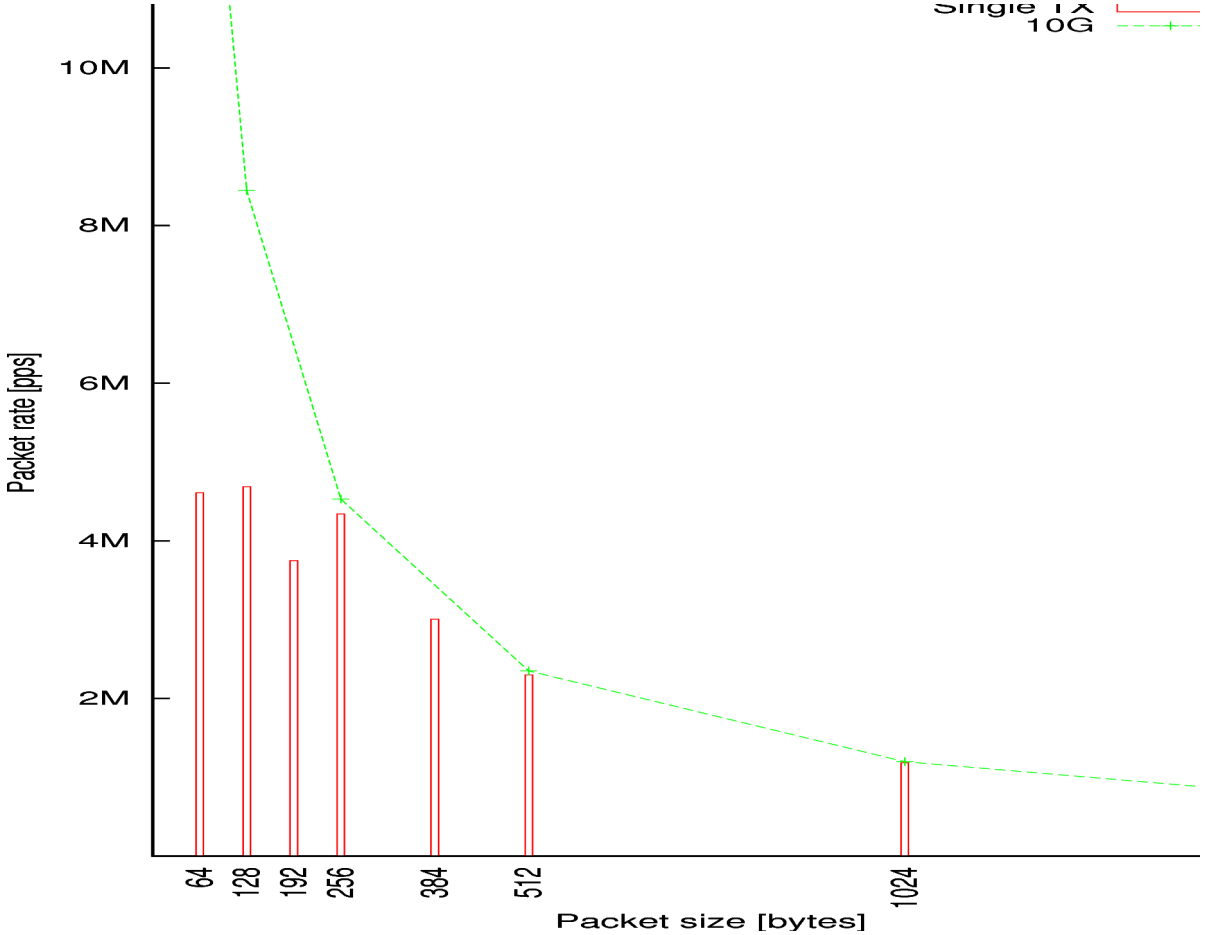
Project 1
Open-source routing at 10Gb/s



bps


Single TX
10G
8G
6G
4G
2G
Bit rate [bps]
64
128
192
256
384
512
1024
Packet size [bytes]

pps

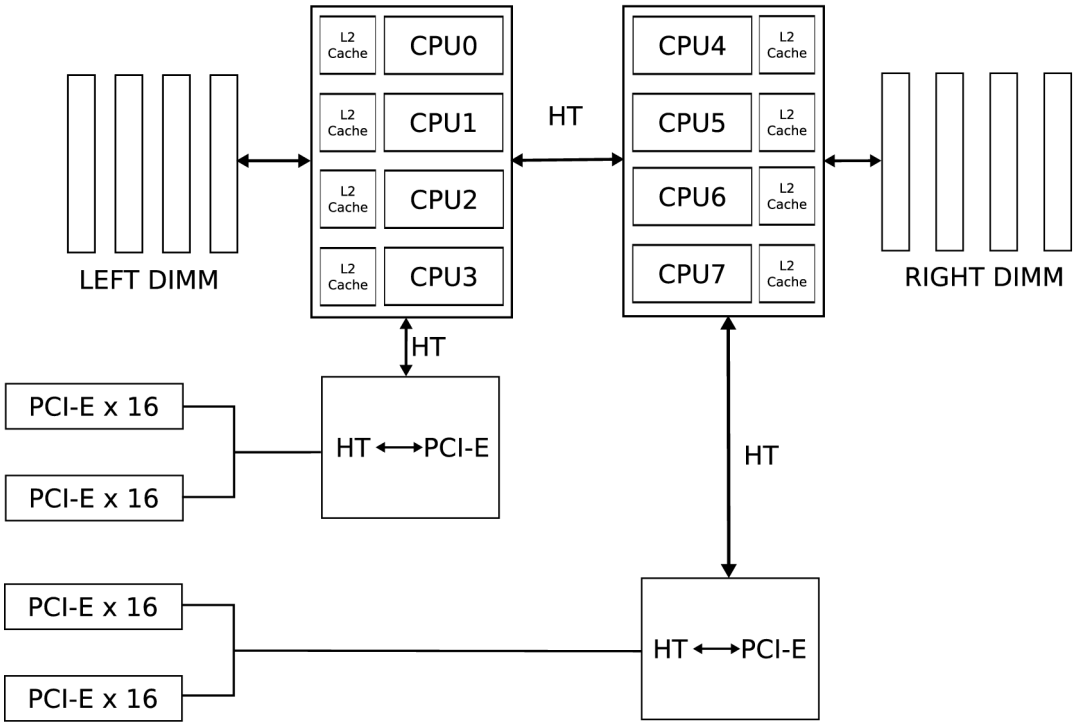

Single TX
10G
10M
8M
6M
4M
2M
Packet rate [pps]
64
128
192
256
384
512
1024
Packet size [bytes]

Single CPU and Multiple CPU

%	symbol name
14.8714	kfree
14.6510	dev_kfree_skb_irq
11.2902	skb_release_data
5.8686	eth_type_trans
5.3384	ip_rcv
4.0116	__alloc_skb
2.8413	raise_softirq_irqoff
2.6253	nf_hook_slow
2.6111	kmem_cache_free
2.5622	netif_receive_skb

%	symbol name
22.0815	dev_queue_xmit
13.2333	*__qdisc_run*
4.7552	eth_type_trans
4.1455	dev_kfree_skb_irq
3.5418	kfree
3.2220	netif_receive_skb
3.0430	pfifo_fast_enqueue
2.3723	ip_finish_output
2.3253	__netdev_alloc_skb
2.2434	cache_alloc_refill

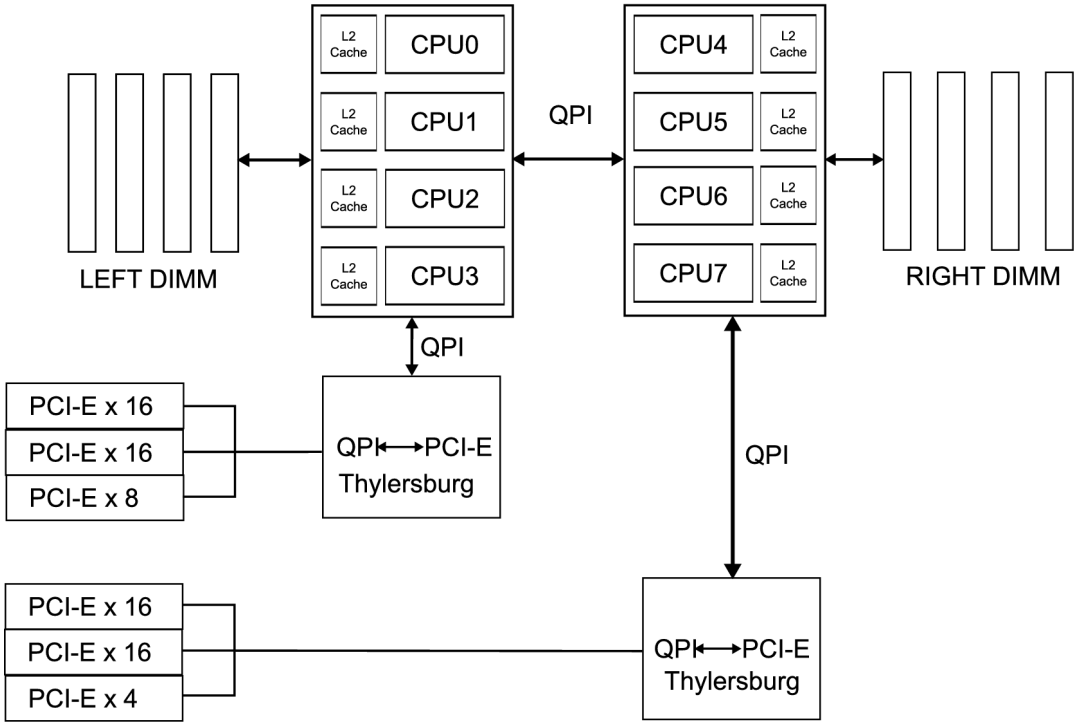
Project 2
Multique



Multique

- Try to separate the flows, and send them to different cores
- Multicore CPU
- Multiqueu on NIC

Project 3
Control and Forwarding plane separation



Control plane

- Control plane:
  - Routing
    - bgp
    - ospf
    - .....
  - ssh
  - Statistics
  - .....
- This goes to CPU0

Forwarding plane

- General forwarding is done on CPU1...CPU(n)
- Multi core CPUs
- Hardware classifiers on NIC
- Fast buses ie QPI / PCIe (2.0)

Classification on 82599

- RSS
  - Microsoft NDIS spec
- N-tuples (Peter P Waskiewicz)
- Flow director, RPS (netdev)

Links

- http://www.nada.kth.se/~olofh/
- http://www.herjulf.se/
- http://www.linux-kongress.org/2010/slides/lk-2010-10G.pdf
- http://data.guug.de/slides/lk2008/10G_preso_lk2008.pdf
- http://www.iis.se/internet-for-alla/internetfonden/uppdrag-2009#kallkodsroutrar
- bengan@resilans.se

RSS can be programmed

- Jens Laas and Robert Olsson found a way to fill the redirection table but skip index 0, which means don't do RSS for CPU0

Hard figures

- The hard limits for the platform
  - >90Gbit/s fan out
  - 25,8Gbit/s forwarding
  - 3.5 Mpps for 1 NIC

Conclusions

- It is possible to do:
  - Forwarding in 10Gbit/s and above on a PC platform
  - Use hardware selection of packets
  - Flow separation

Links

- http://www.nada.kth.se/~olofh/
- http://www.herjulf.se/
- http://www.linux-kongress.org/2010/slides/lk-2010-10G.pdf
- http://data.guug.de/slides/lk2008/10G_preso_lk2008.pdf
- http://www.iis.se/internet-for-alla/internetfonden/uppdrag-2009#kallkodsroutrar
- bengan@resilans.se