

13th ANNUAL WORKSHOP 2017

USE CASES FOR RAW ETHERNET QUEUE PAIRS

Christoph Lameter, Ph.D.

Jump Trading LLC

[March 28th , 2017]



OVERVIEW RAW ETH QP USE CASES

Doing Amazing things with Ethernet Frames

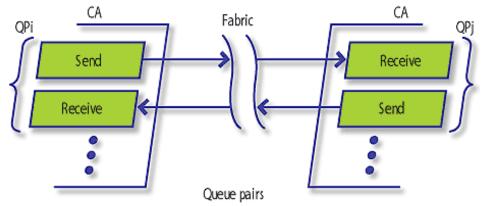
- •What is a Raw Ethernet Queue Pair
- Packet Capture
- Direct User space packet
- •Offloading Multicast streams to user space

WHAT IS A RAW ETHERNET QUEUE PAIR

- A queue pair like available for normal Infiniband traffic.
 - UD only

Ability to send and receive Raw Ethernet Frames

- No IP stack. Just the raw hardware and the acceleration functionality provided by the Ethernet portion of the NIC.
- No software overhead through kernel processing
- Ethernet does not support connections in hardware like Infiniband
 - Hardware support for connections missing
 - User space code must provide additional logic to compensate

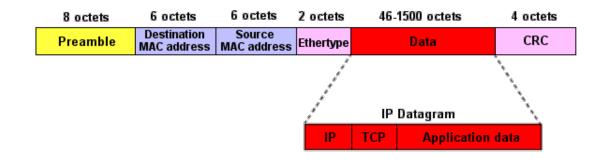




RAW OPS FOR PACKET CAPTURE

RAW PACKET RECEIVE PROCESSING

- Queue pair is opened for receice only.
- Interface in promiscuous mode
- Ability to audit all packets on a port
- Timestamping
- FCS checksum
- Low overhead reception in user space. Ability to capture at line rate for 10G and 1G.
- Ability to inspect packets that violate checksums and/or other requirements that are usually rejected by the network stack or hardware.
- Can be the basis for network diagnostic functionality.





SENDING RAW FRAMES

CUSTOM ETHERNET FRAMES

- Ability to format a frame in user space
- Allows specialized frames with unusual fields set.
- Helps debugging network devices.
- Full control...
- Finally the OS is no longer in the way.

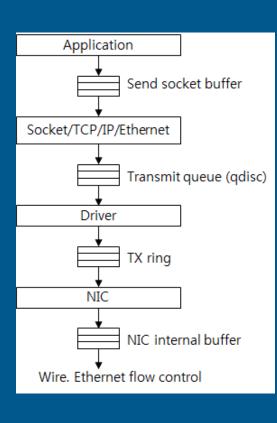
				1518 bytes -					
remble :	DA	SA	Leigh	Payload	FCS				
byte e +	⊨ 6 byte -	🕳 6 byte 🛥	- 2 byte -						
+	t= 1518 bytus *t								
Premble	DA	SA	Тура	Payload	FCS	Ethernet II Frame			
8 byte -+	- 6 byte -	6 byte -	4-2 byte -4	46-1500 byte	ala- 4 byte -al				
				15.22 bytes					
				13.22 0,188					
Premble	DA	SA	0x8100	MD Type	Payload	FCS			
					45.1500 hver				
8 byte -	- 6 byte -	e- 6 byte -	- 2 byte -	a 2 byte 2 byte		- 4 byta -			
8 byte	■ 6 byte —	e 6 byte -	i 2 byte -	• 2 byte		ala- 4 byta -al		-+	
+	-			1528	bytas		6/2	-+ □	
8 byte	e 6 byte e DA	e⊷ 6 byte → SA		1528		eje-4 byta ej Paylaad	FCS	-	
Premble	DA	SA	0,000 40	1528	C Type	Payload			
Premble	DA	SA	0,000 40	15.26 ND 0x81.00	C Type	Payload			
hemble	DA	SA	0,000 40	15.26 ND 0x81.00	bytas VD Type yla → 2 byta →	Payload		Ţ	-



MULTI CAST KERNEL BYPASS

COMBINED NETWORK IP STACK AS WELL AS MULTICAST BYPASS

The Wizard solution



- Regular network stack of Linux does its job (and handles the traffic with its regular slowness)
- Select multicast groups are forwarded to user space bypassing kernel processing.
- Redirection of select multicast groups to queue pairs created by the RDMA application in user space.



13th ANNUAL WORKSHOP 2017

THANK YOU

Christoph Lameter Jump Trading LLC