



Cramming Ethernet into FPGAs

William Kamp, Ph.D
High Performance Computing Research Lab
Auckland University of Technology

14th February 2019 - C4SKA @ AUT

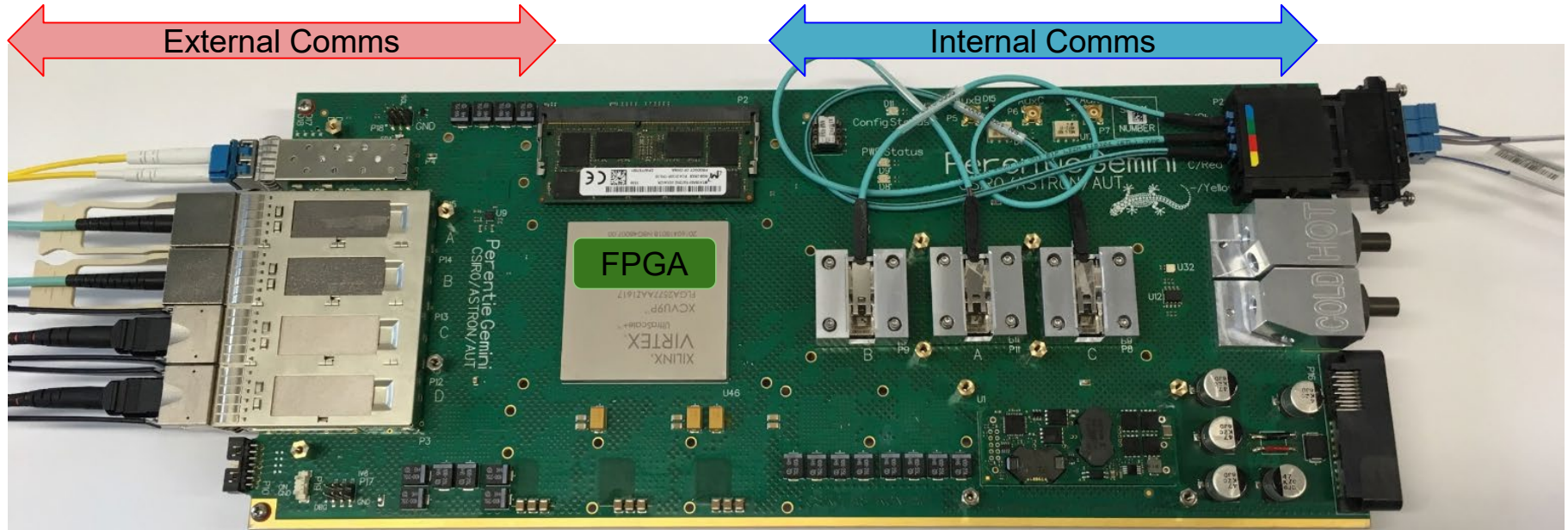


Low Correlator BeamFormer (Low.CBF)

Low CBF has 288 FPGA.

Each FPGA must communicate data to every other FPGA.

Per FPGA bandwidth required of ~400 Gbps.



COTS is King. Lets use Ethernet!

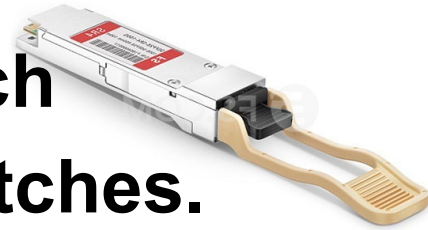


Requires a 1152 Port 100GbE switch.

About 15,000 NZD per 32x100GbE switch

Need 72 of them in the first layer of switches.

You'd make a network vendor sales person's year!



And we're done ...



THE END.

A screenshot of the FS.COM website. The main product page is for the "N8500-32C (32*100Gb) 100Gb Spine/Core Layer Switch with Cumulus® Linux® OS". The price is listed as NZ\$18,594.33. A modal window titled "Added to Cart" is overlaid on the page, showing the product name, price (NZ\$1,338,791.76), and a quantity of 72. Below the modal, a red button labeled "Proceed to Checkout" is highlighted with a blue mouse cursor. The background page shows navigation menus, a search bar, and a description of the switch's capabilities.

FS.COM All Categories Shop by Subject Solutions Resources About Us Search Products...

Home Enterprise Network Ethernet Switches 100G Switches Item ID: 75877

CUMULUS

N8500-32C (32*100Gb) 100Gb Spine/Core Layer Switch with Cumulus® Linux® OS #75877

★★★★★ 1 Review 2 Questions Get a quote

NZ\$18,594.33

FS P/N: N8500-32C

Software Support: 1 Year (Default) 3 Years 5 Years

Available, AU Warehouse

✓ Added to Cart

N8500-32C (32*100Gb) 100Gb Spine/Core Layer Switch with Cumulus® Linux® OS NZ\$1,338,791.76

Software Support - 5 Years 72

#75877

Cart Subtotal (72 Items) NZ\$ 1,338,791.76 — Continue to Cart

Keep Shopping Proceed to Checkout

N8500-32C (32*100GbE) 100GbE Open Networking Switch Preloaded with Cumulus Linux

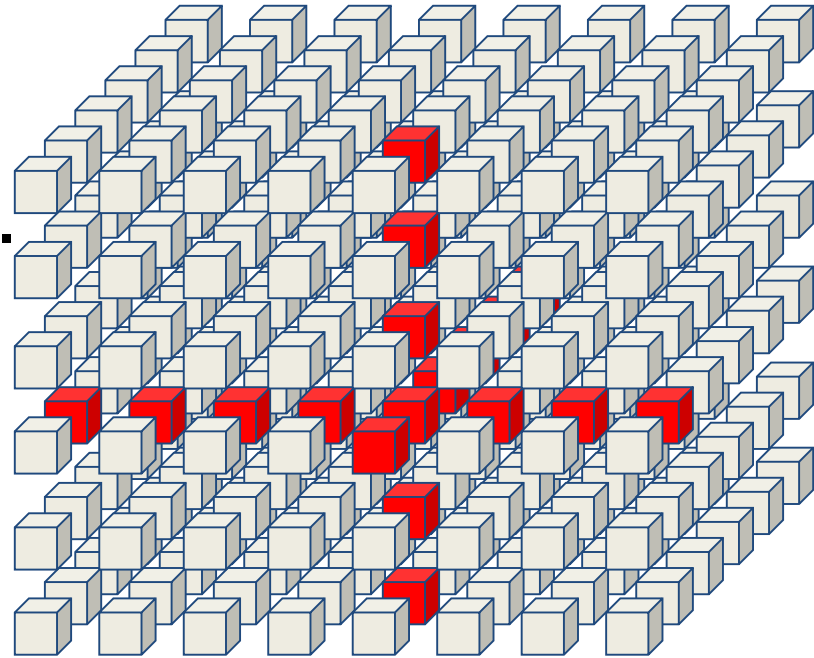
N8500-32C is a Top-of-Rack (TOR) or Spine switch in a compact 1U form factor, ideally suited for high performance and programmable data centre environments. N8500-32C is a 32 ports 100GbE switch designed for carrier/enterprise aggregation and data centre top-of-rack/spine. Designed with top performance in mind, N8500-32C spine/core switch provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Redundant power and fans along with numerous high availability features ensure that N8500-32C is always available for business-sensitive traffic. Combined with Cumulus Linux network operating system, FS.COM N series switches allow customers to deploy fast, high-capacity fabrics, simplified network automation and consistent tools, and help lower operational and capital expenditures. With support for advanced features, including MLAG, VXLAN, SFLOW, SNMP etc, this layer 3 switch is ideal for traditional or fully virtualised data centre.

N8500-32C data switch supports current and future data centre requirements, including a x86-based control plane for easier integration of automation tools. Additionally, N8500-32C Ethernet switch also supports the advanced hardware based VXLAN feature to support over 16M virtual networks.

Ports	32*100Gb	Operating System	Cumulus® Linux® OS
Max. 100Gb Ports	32	CPU	Intel Rangeley C2538 2.4Ghz 4-core
Max. 25Gb Ports	128	Switching Chip	Tomahawk BCM56960
Switch Capability	6.4Tbps Full-duplex	Packet Buffer Memory	16M

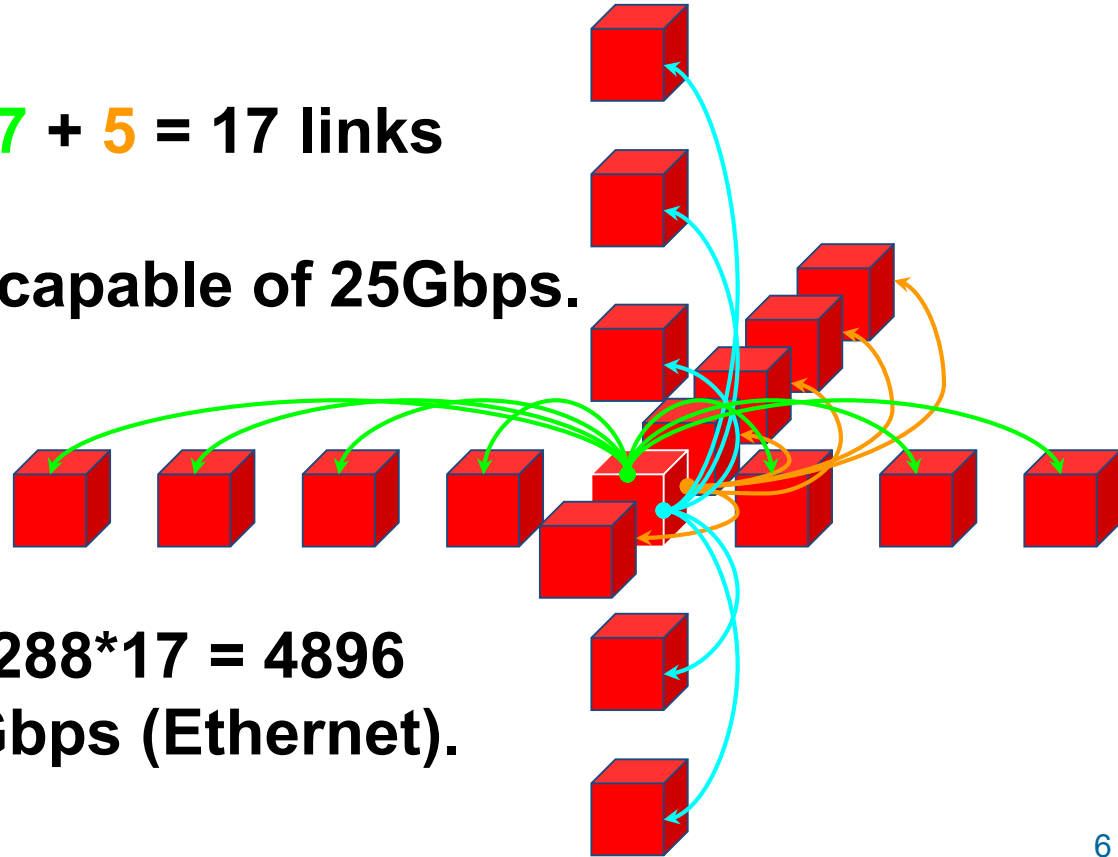
OK, So that didn't scale well

- Let's do the switching ourselves - we do have FPGAs after all.
- Arrange FPGA in a $6 \times 8 \times 6$ cube.
- Each FPGA communicates directly with every FPGA that has two common indexes in X, Y, or Z.
- Can get to any other FPGA in two or three hops.



Low.CBF solution

- Each FPGA has $5 + 7 + 5 = 17$ links (plus 3 loop-back)
- Serial Transceivers capable of 25Gbps.



- Therefore a total of $288 * 17 = 4896$ simplex links of 25Gbps (Ethernet).

Mid.CBF Solution



VCC (Very Coarse Channeliser) with 200 FPGA
to 26 FSPs (Frequency Slice Processors) each of 20 FPGA
and back again.

Each FSP FPGA has

- 10 duplex links to the VCCs
- 19 simplex links to every other FPGA in the FSP.

In total, $(26*20)*(10*2+19)$
= 20,280 links.




Ethernet's Backwards Legacy



Why has Ethernet owned the World?

- Because each new speed is backwards compatible.
- Only the Physical Layer has changed.
 - Coax -> Twisted Pair -> Optical Fibre.

Ethernet packet has accumulated a lot of cruft in its four decades.



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction

[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#)
[Log in](#)

Article [Talk](#) Read [More](#)

Cruft

From Wikipedia, the free encyclopedia

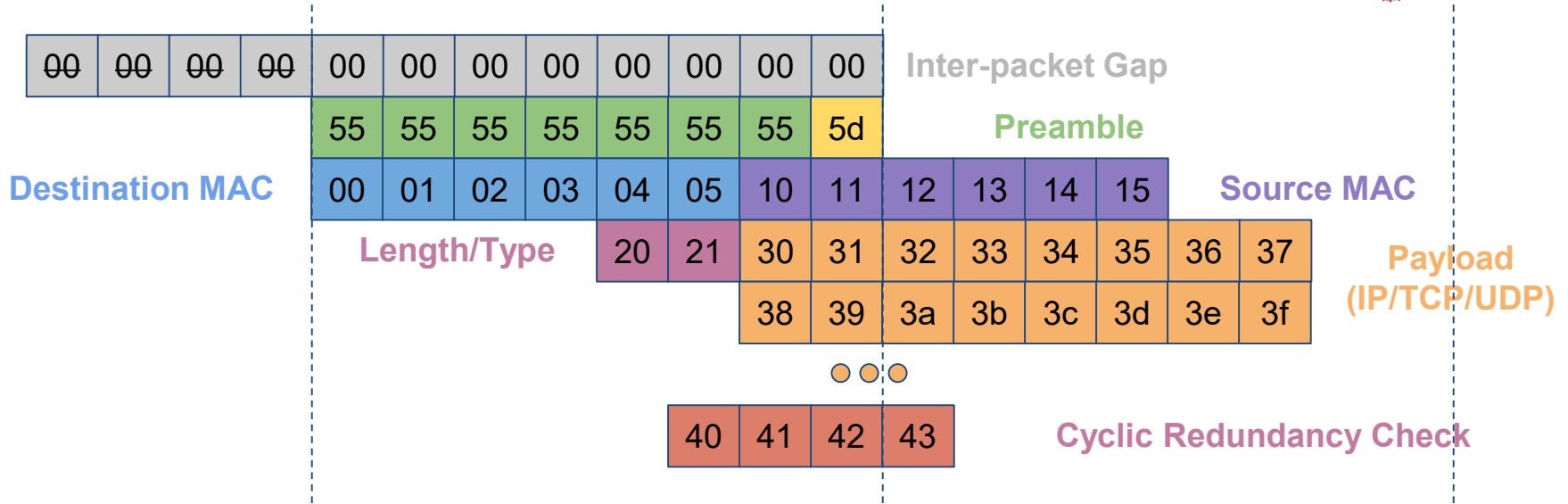
This article is about a computing term. For other uses, see [Cruft \(disambiguation\)](#).

Cruft is a [jargon](#) word for anything that is left over, redundant and getting in the way. It is used particularly for defective, superseded, useless, superfluous, or dysfunctional elements in [computer software](#).

Contents [hide]

- 1 History
- 2 Software
- 3 Computer hardware
- 4 See also

Ethernet's Cruft



- IPG - min of 12 bytes. Insert or delete bytes to match clock rates. ❌
- Preamble - To train your receiving PLL on shared coaxial cable. ❌
- MAC addresses - 6 Bytes each - not a power of 2! ❌
- Type - 2 Bytes - still not back to a nice alignment! ❌
- Packet length can be any integer number of bytes. ❌

IP, TCP and UDP

- All designed with a 4 byte (32 bit) alignment.
- Both have checksums to detect relay errors.
 - Checksums stored in the header. ❌
 - Must store the whole packet to insert checksum.
 - Increases the RAM requirements.
 - Makes streaming difficult.
- Ethernet Frame's CRC
 - Stored at the end of the packet ✓
 - No RAM required to append.
 - Can have any 1-byte alignment.
 - Need to encode its location. ❌

TCP Segment Header Format									
Bit #	0	7	8	15	16	23	24	31	
0	Source Port				Destination Port				
32	Sequence Number								
64	Acknowledgment Number								
96	Data Offset	Res	Flags			Window Size			
128	Header and Data Checksum					Urgent Pointer			
160...	Options								

UDP Datagram Header Format									
Bit #	0	7	8	15	16	23	24	31	
0	Source Port				Destination Port				
32	Length					Header and Data Checksum			

Do we need all that stuff?



Well, not really.

- **Internal data busses are 64-bit wide (not 8).**
- **Links are streaming point to point - Source MAC address not required.**
- **Optical links are continuously signaled - preamble not required.**
- **Inter-packet gap (IPG). No. Allow intra-packet gaps instead.**

Being Backwards is Expensive

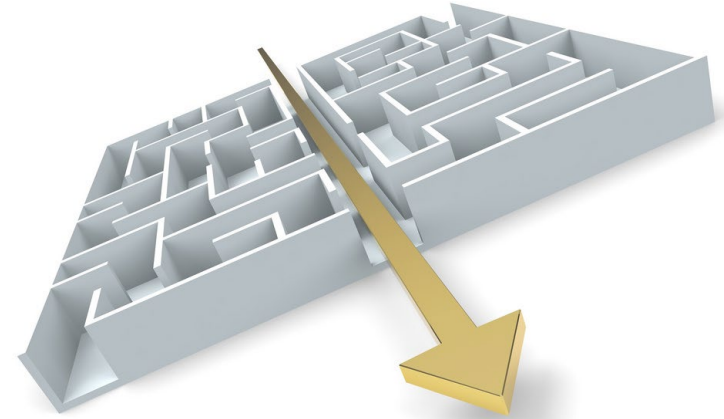


Supporting all the cruft (variation and backwards compatibility) costs significant FPGA resources.

We have full control over the packets.

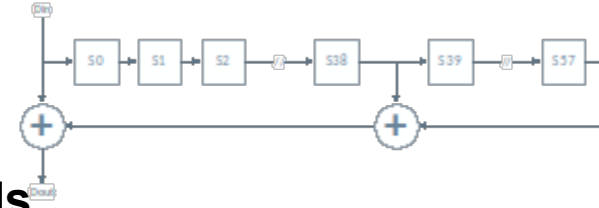
We don't *need* interoperability with other equipment.

Let's simplify down to what we do *need*.



Custom Packet Framing

- **Use the 25Gb Ethernet Physical Layer (PHY).**
 - **64b/66b Encoding.**
 - Encodes 64b words plus 2b to say it is a control or data word.
 - Provides a mechanism to find alignment.
 - **Multiplicative Scrambling**
 - Provides a DC balance - statistically.
 - Easy to decode, no synchronisation words.
 - **However, it multiplies bit errors - bad. Turns 1 bit error into 3.**



FPGAs provide some or all of this functionality in the transceiver hard logic.

Custom Packet Framing



Simplified Packet Framing (MAC)

- **Minimum unit of transmission is 8 bytes (64b) matching the PHY.**
 - Maintains simple alignment.
 - Packet lengths a multiple of 8 bytes.
- **Control words have 4 flavours (2 bits encoding)**
 - **00 = Idle/Invalid** - Inter/Intra-Packet gap,
 - **10 = Header** - Start of Packet,
 - **01 = Trailer** - End of Packet,
 - **11 = Header & Trailer** - End and Start.

00	00	00	00	00	10	11	12
30	31	32	33	34	35	36	37
38	39	3a	3b	3c	3d	3e	3f



30	31	32	33	34	35	36	37
00	00	00	00	00	00	00	00



38	39	3a	3b	3c	3d	3e	3f
40	41	42	43	44	00	00	00

Collapse Headers and Trailers



- **Header & Trailer bits are mutually exclusive.**
 - Header + trailer = 62b + 2b flags.
 - Can occupy the same word.
 - Halves the minimum overhead
 - To 64b per packet.
 - Minimum packet size = 8 bytes.
 - Just Header and Trailer.
 - Compare to 84 byte minimum Ethernet Frame (including IPG and preamble).

○ ○ ○

30	31	32	33	34	35	36	37
38	39	3a	3b	3c	3d	3e	3f
40	41	42	43	44	00	00	00
00	00	00	00	00	10	11	12
30	31	32	33	34	35	36	37
38	39	3a	3b	3c	3d	3e	3f
40	41	42	43	44	10	11	12
30	31	32	33	34	35	36	37
38	39	3a	3b	3c	3d	3e	3f

○ ○ ○

Idle Words



- **Idle words anywhere**
 - No limit on maximum packet length.
 - However, prefer idle between packets to help correcting bit errors in sync bits.
- **Custom Idle words supported**
 - Could be used to uniquely identify each link.
 - Detect miswiring
 - Detect misconfiguration
- **Used to monitor and report bit error rate (BER)**

00	00	00	00	00	00	00	00
00	00	00	00	00	10	11	12
30	31	32	33	34	35	36	37
38	39	3a	3b	3c	3d	3e	3f



30	31	32	33	34	35	36	37
00	00	00	00	00	00	00	00



38	39	3a	3b	3c	3d	3e	3f
40	41	42	43	44	00	00	00
00	00	00	00	00	00	00	00

What are we missing?

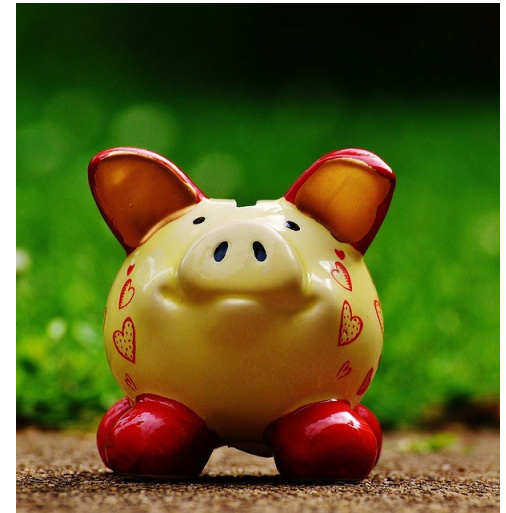


- **Warm fuzzies that comes with proven, industry standardised techniques.**
- **The bloat that comes with Ethernet's Cruft.**

Size Comparison

Intel Stratix10, with 29 instances.

- **25Gb Ethernet IP**
 - **Smallest PHY+PCS+MAC = 3850 ALMs each**
 - **Totals 111,650 ALMs = 16% of the FPGA.**
- **New Serial Interconnect**
 - **Full Bells and Whistles = 650 ALMs**
 - **Total of 18,850 ALMs ~ 2.8% of FPGA.**
- **13% less power and cost OR**
- **15% more science!**



Questions / Discussion

Thank you!