

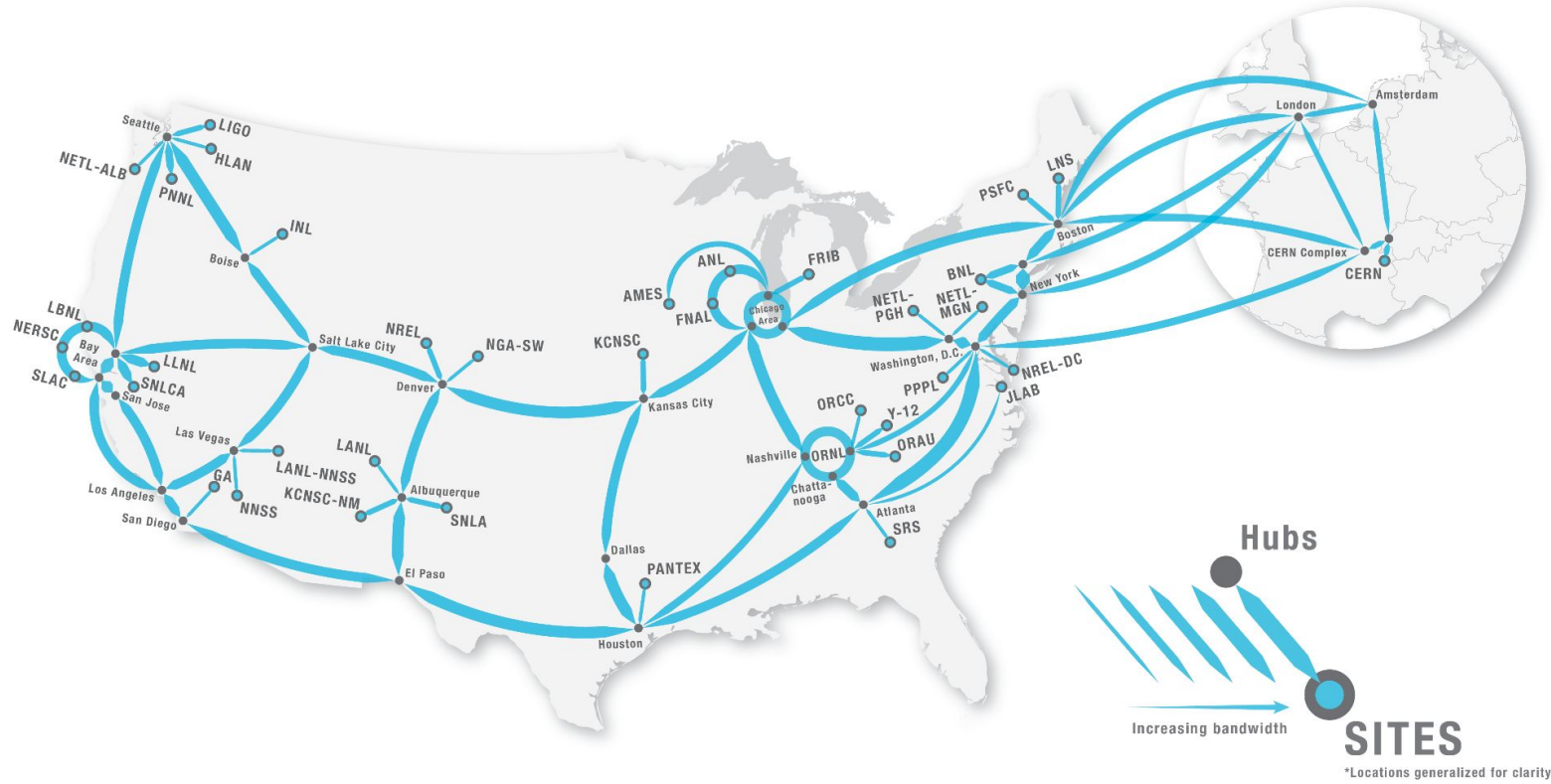
DPDK as an Offload Engine for P4 SmartNIC Applications

Chris Cummings
Network Automation Software Engineer
Sean Cummings
Student Assistant

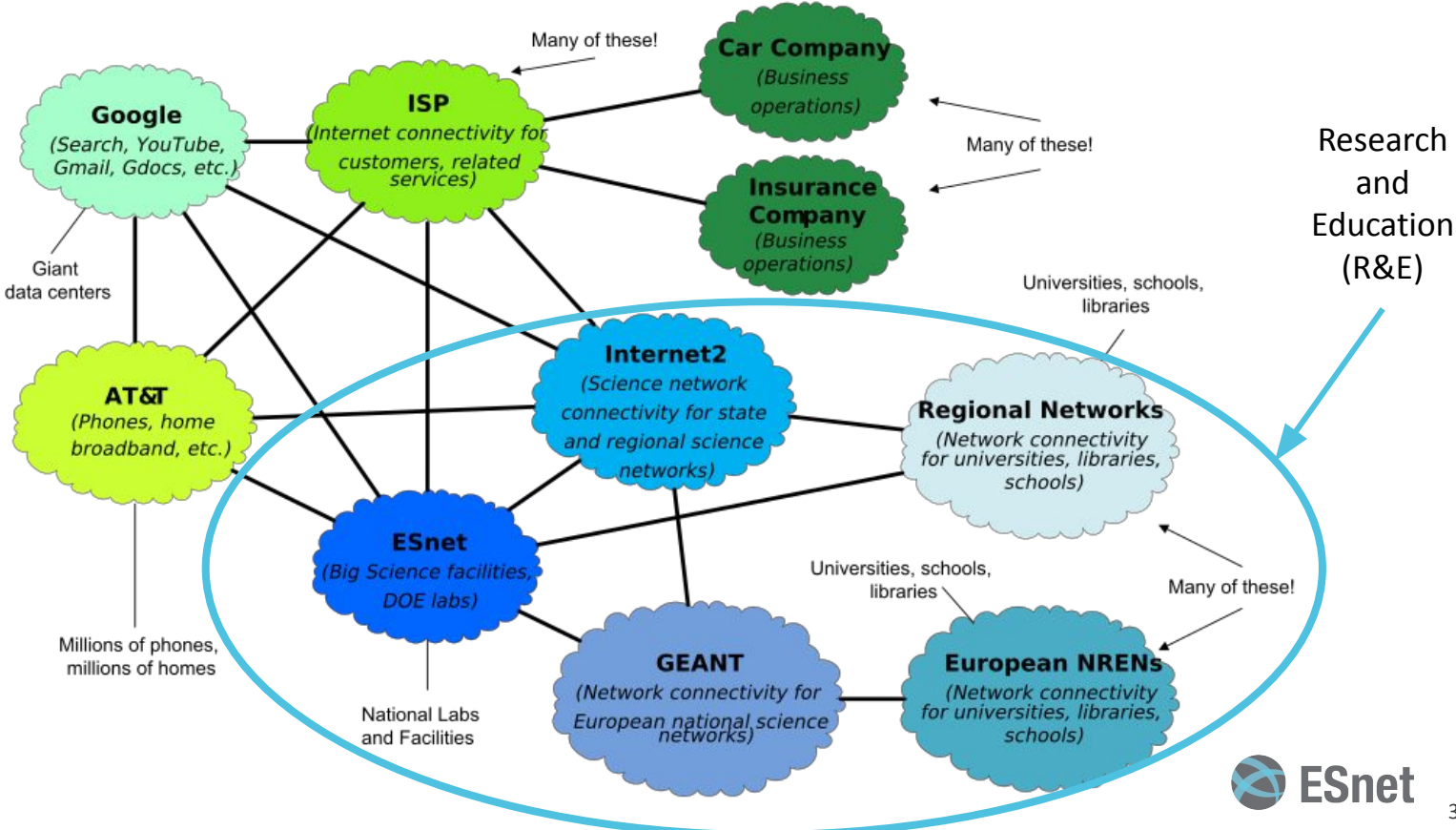
Energy Sciences Network (ESnet)
Lawrence Berkeley National Laboratory
U.S. Department of Energy

DPDK Summit

September 2023



What is an R&E Network?



Commercial ISPs vs. R&E Networks:

Normal ISP:[1]



ESnet:[2]

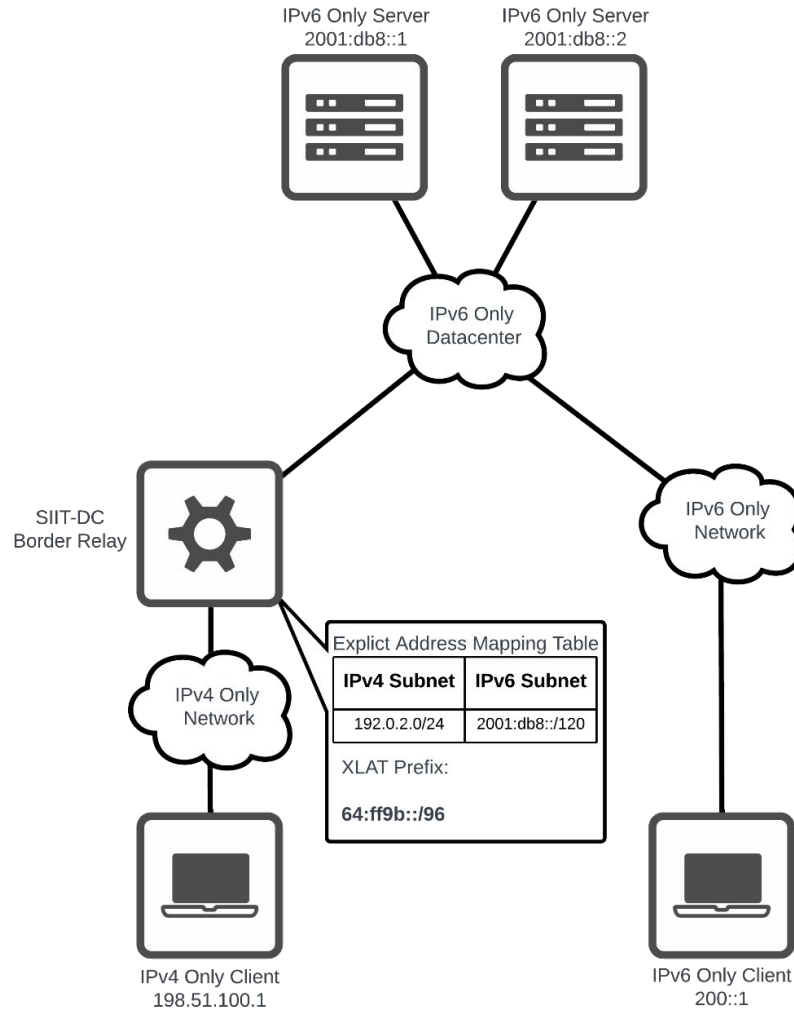


Project Background

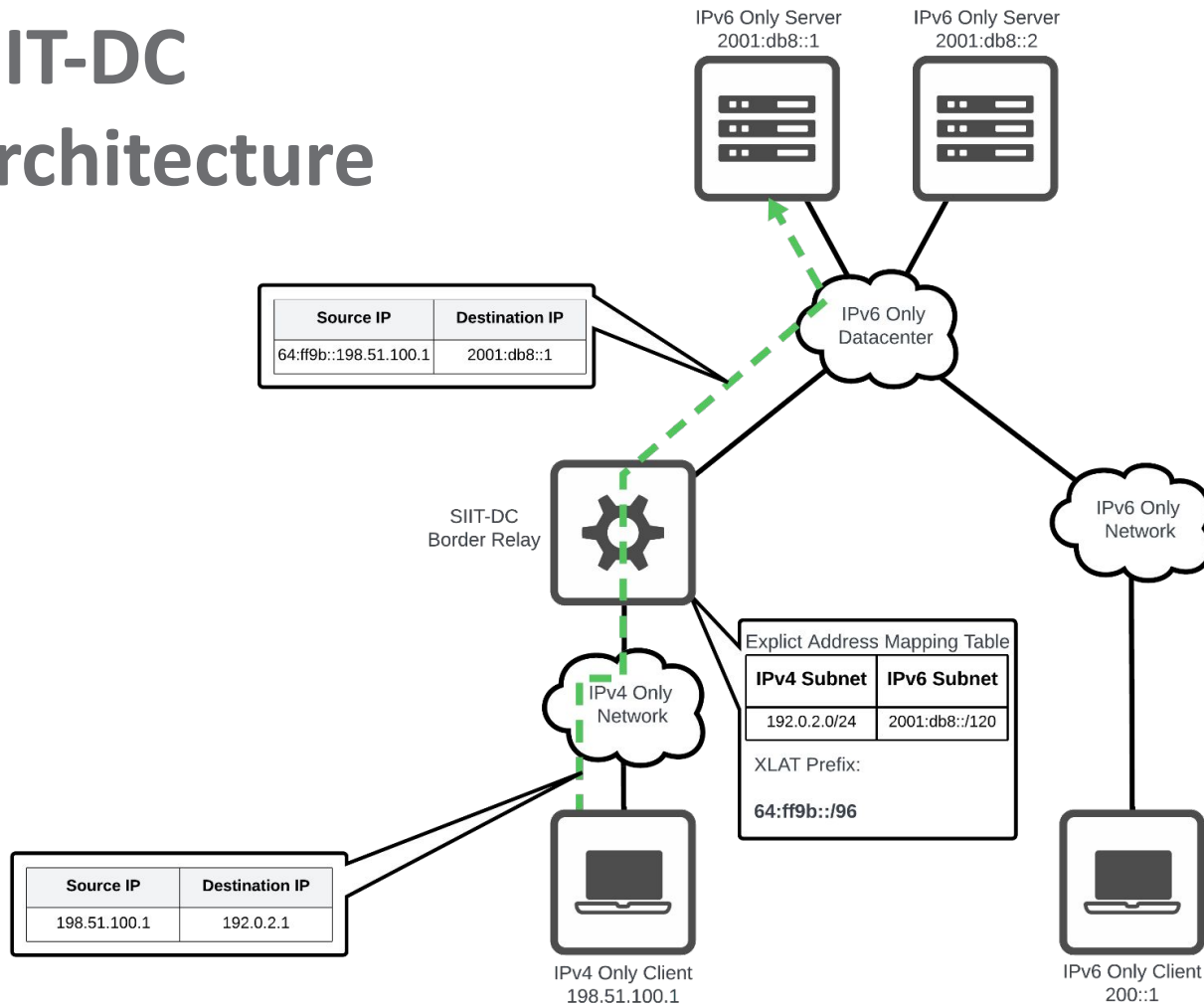
- IPv6: The future is here, but not *all* the way here
- Some things can't be upgraded for IPv6 (expensive scientific instruments)
- Dual-stacking isn't necessarily the right solution anymore (OMB M-21-07)
- v4-only and v6-only segments still need to communicate... How?
- There is a tried-and-true solution...



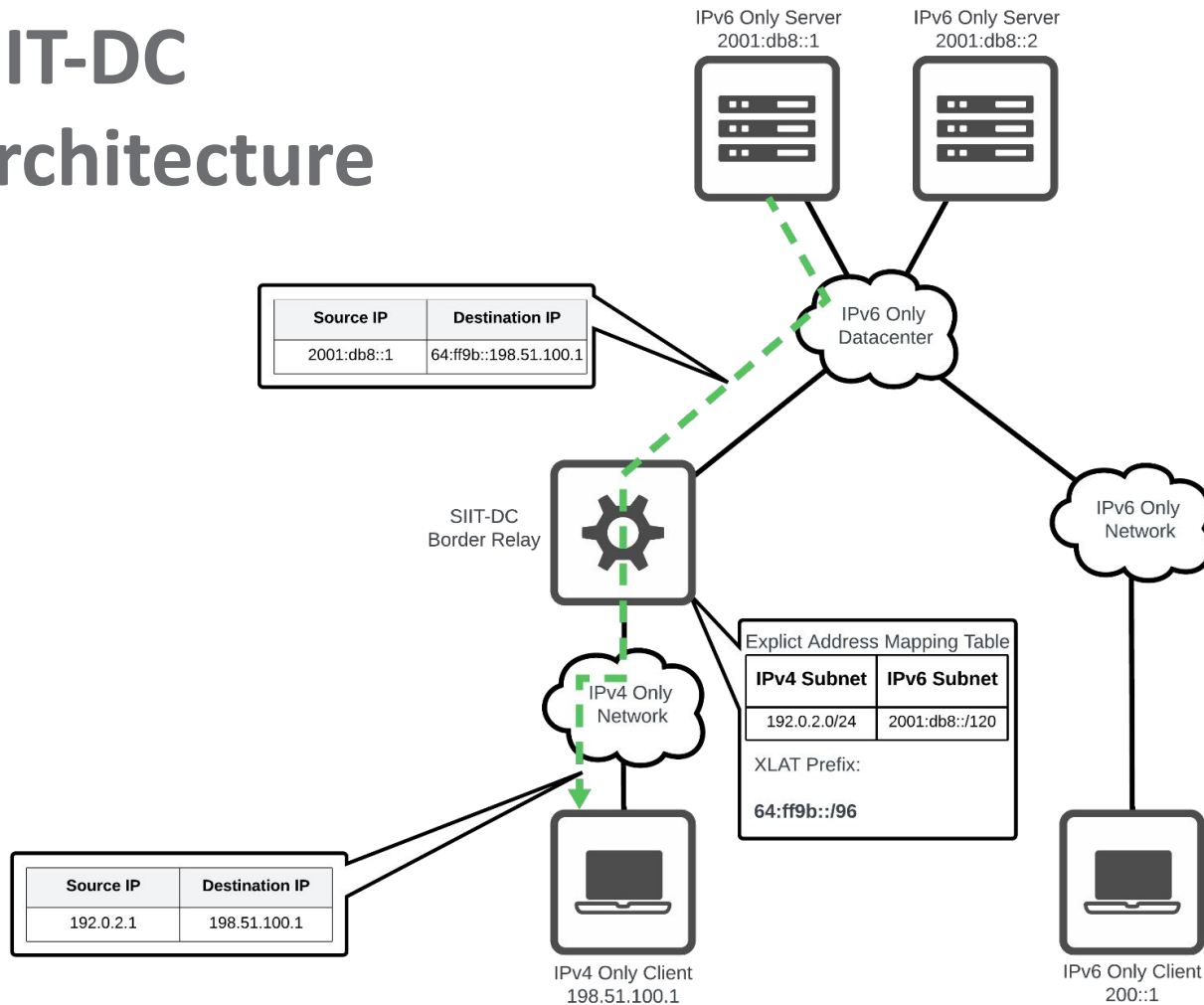
SIIT-DC Architecture



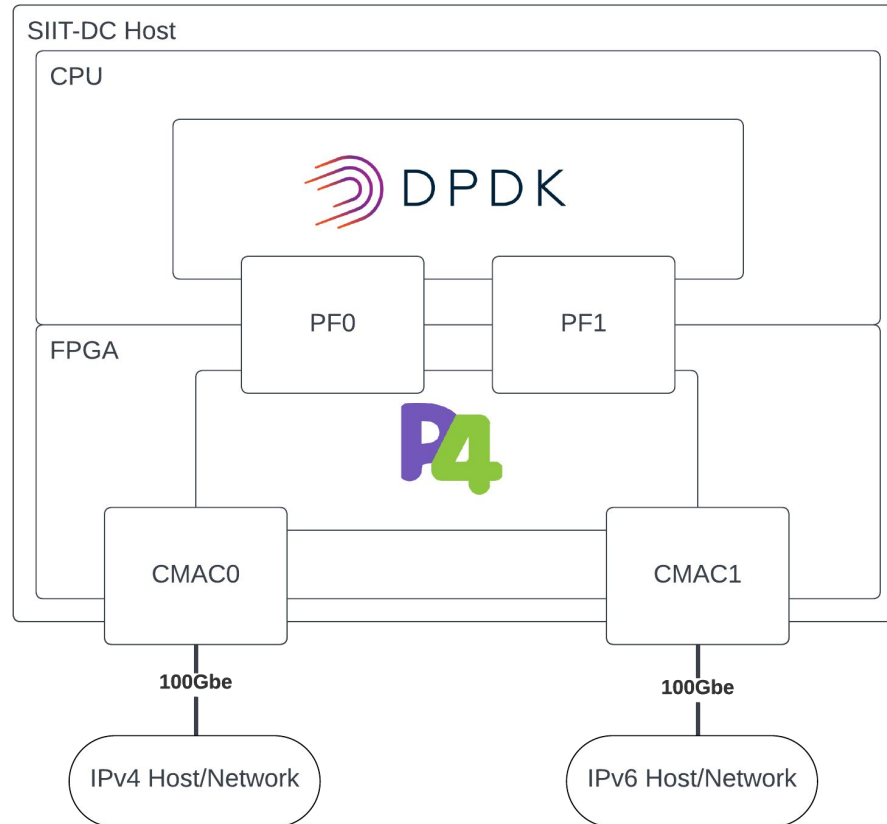
SIIT-DC Architecture



SIIT-DC Architecture



Architecture



Hardware

- Xilinx Alveo U280 FPGA^[3]
- 2x 100GbE Ethernet CMAC interfaces
- 8x PCIe Gen4 Lanes
- Intel Xeon E5-2670 CPU in Host Server

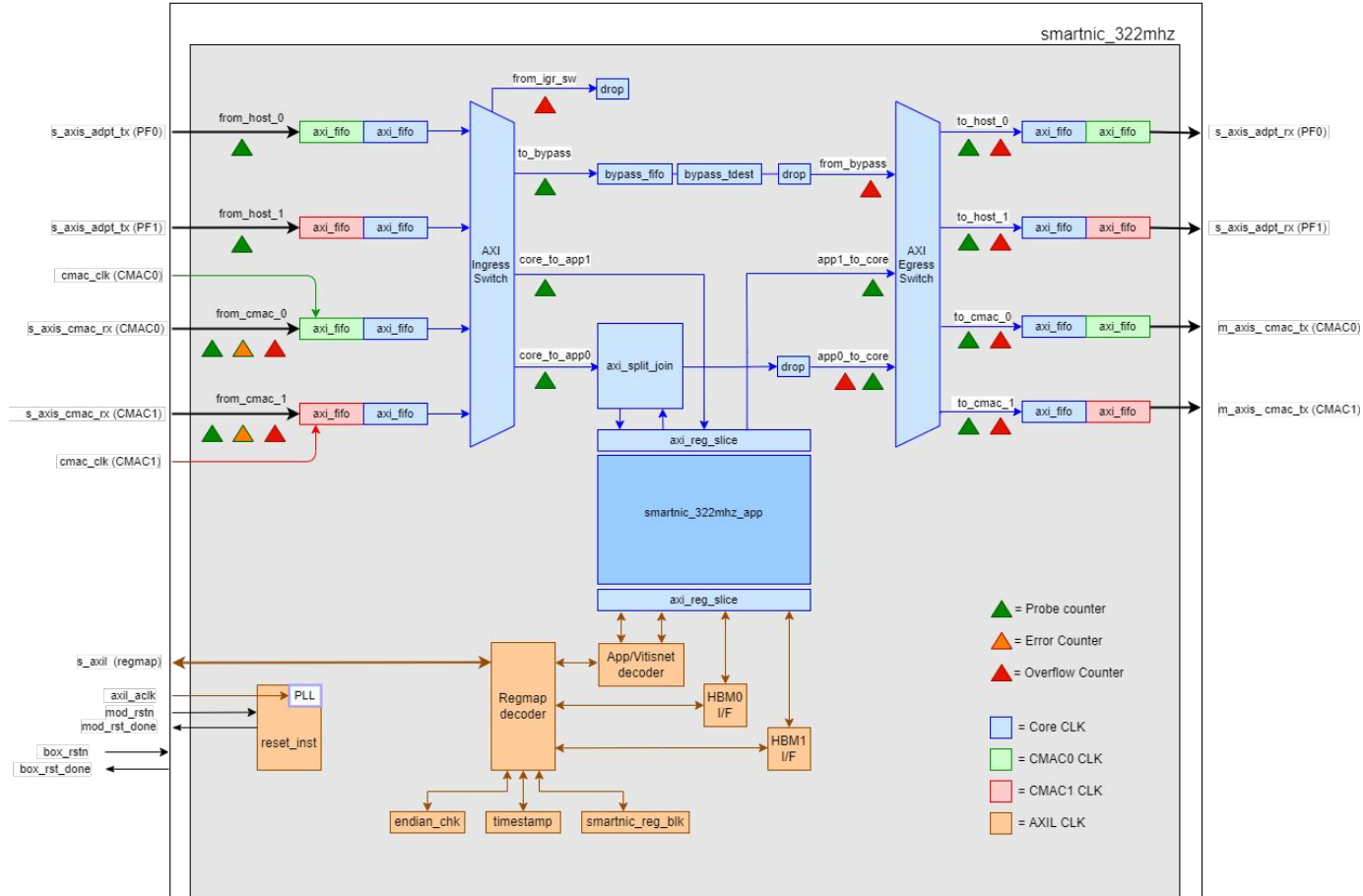


What is ESnet SmartNIC?

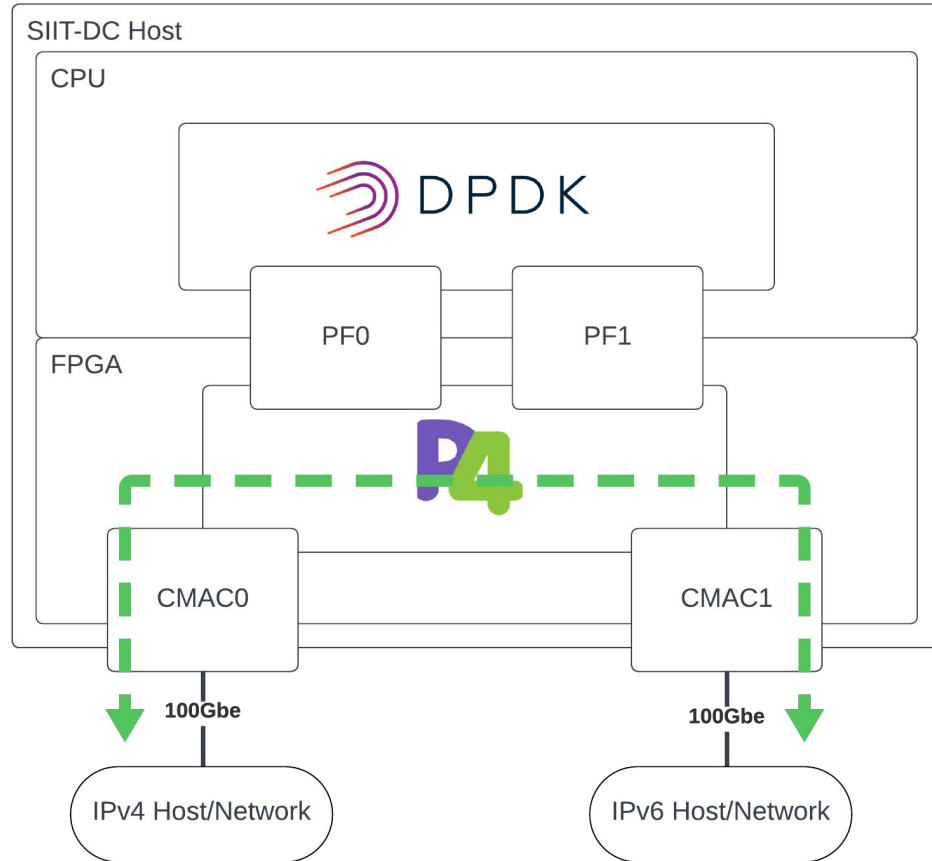
- Framework for developing FPGA-based applications
- Helps you deploy your FPGA-based applications
- Uses docker compose
- Repos Links:
 - [esnet-smartnic-fw](#)
 - [esnet-smartnic-hw](#)

ESnet SmartNIC HW block diagram

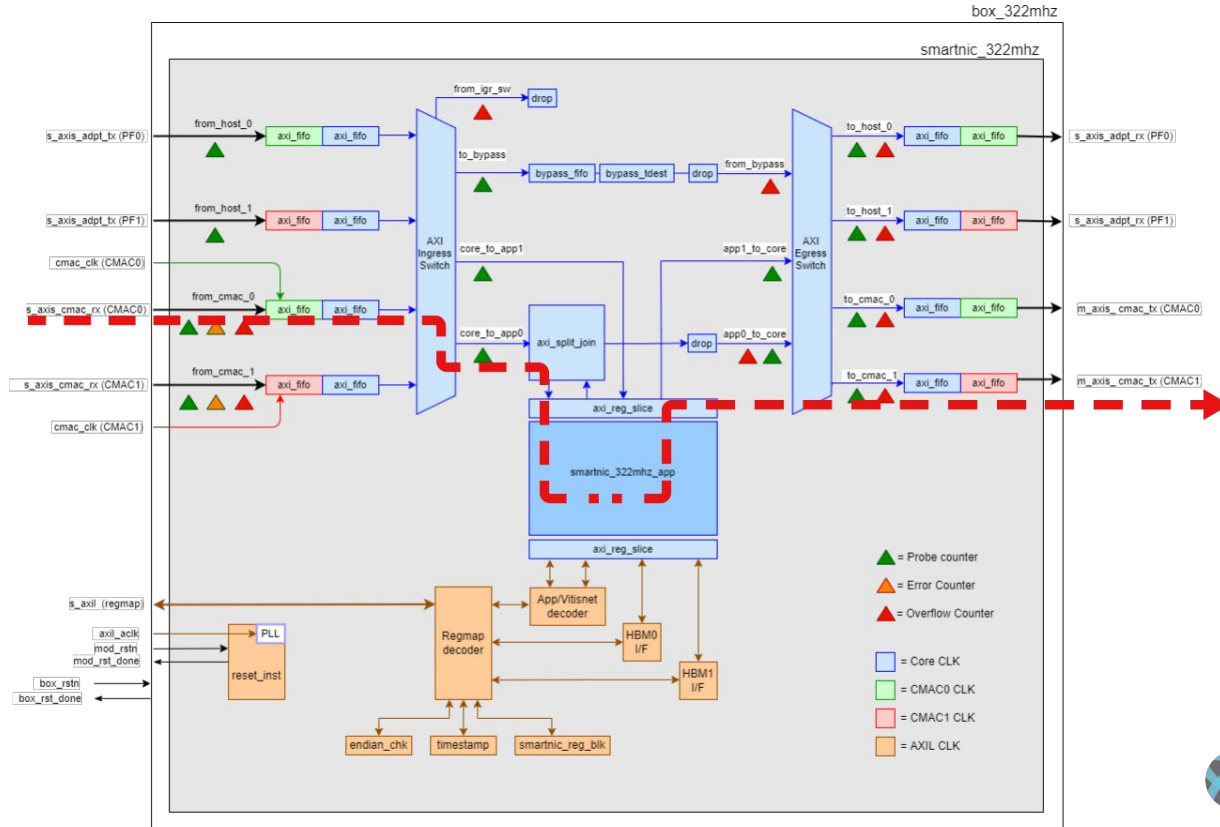
box_322mhz



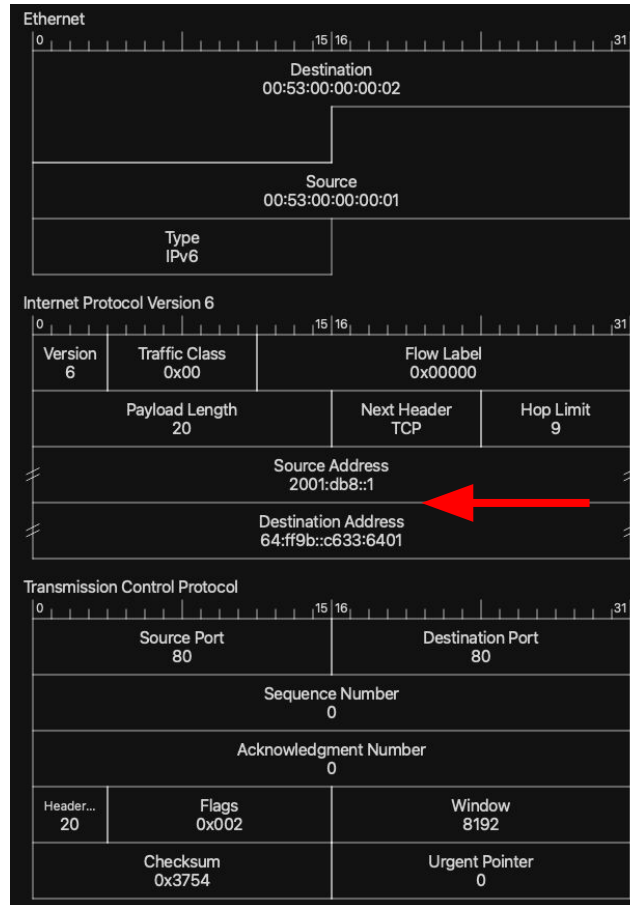
Fast Path Overview



Fast Path Detail



Untranslated IPv6/TCP



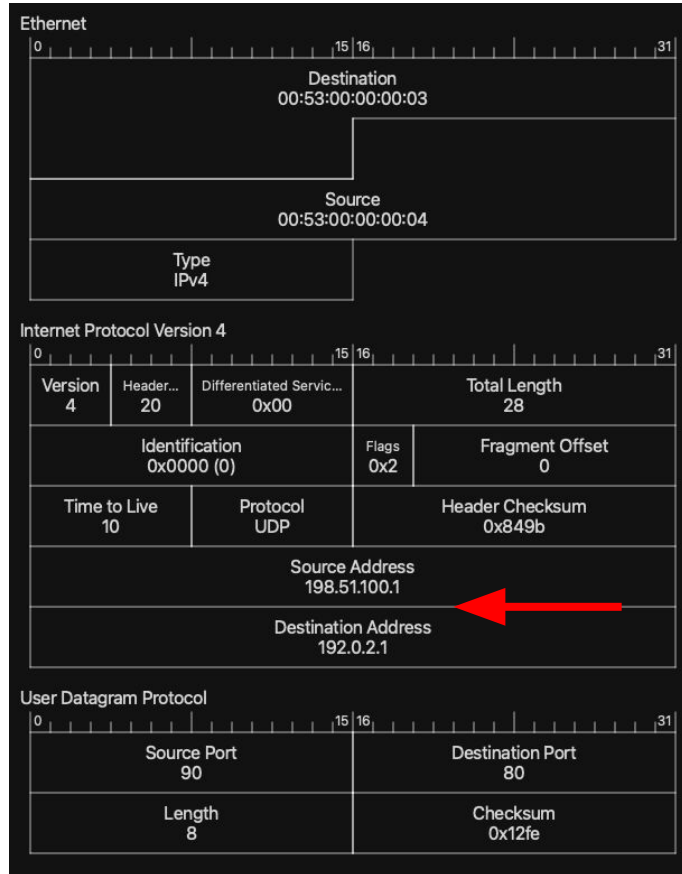
Translated IPv4/TCP

Ethernet					
0		15 16		31	
Destination 00:53:00:00:00:04					
Source 00:53:00:00:00:03					
Type IPv4					

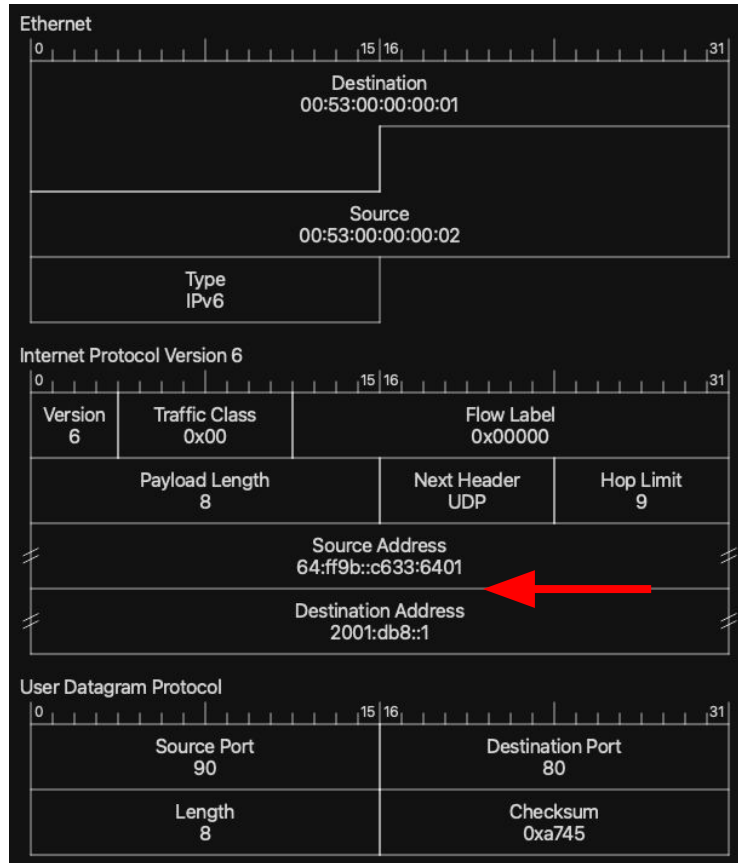
Internet Protocol Version 4					
0		15 16		31	
Version 4	Header... 20	Differentiated Servic... 0x00	Total Length 40		
Identification 0x0000 (0)			Flags 0x2	Fragment Offset 0	
Time to Live 8	Protocol TCP		Header Checksum 0x869a		
Source Address 192.0.2.1					
Destination Address 198.51.100.1					

Transmission Control Protocol					
0		15 16		31	
Source Port 80		Destination Port 80			
Sequence Number 0					
Acknowledgment Number 0					
Header... 20	Flags 0x002		Window 8192		
Checksum 0xa30c			Urgent Pointer 0		

Untranslated IPv4/UDP Packet

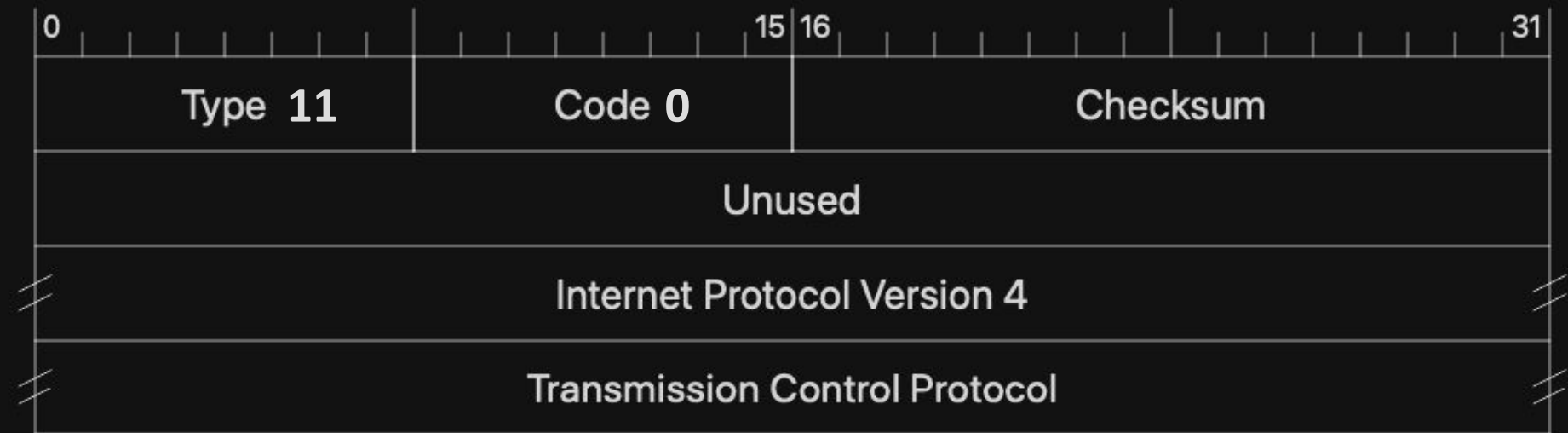


Translated IPv6/UDP Packet



ICMPv4 Error

Internet Control Message Protocol

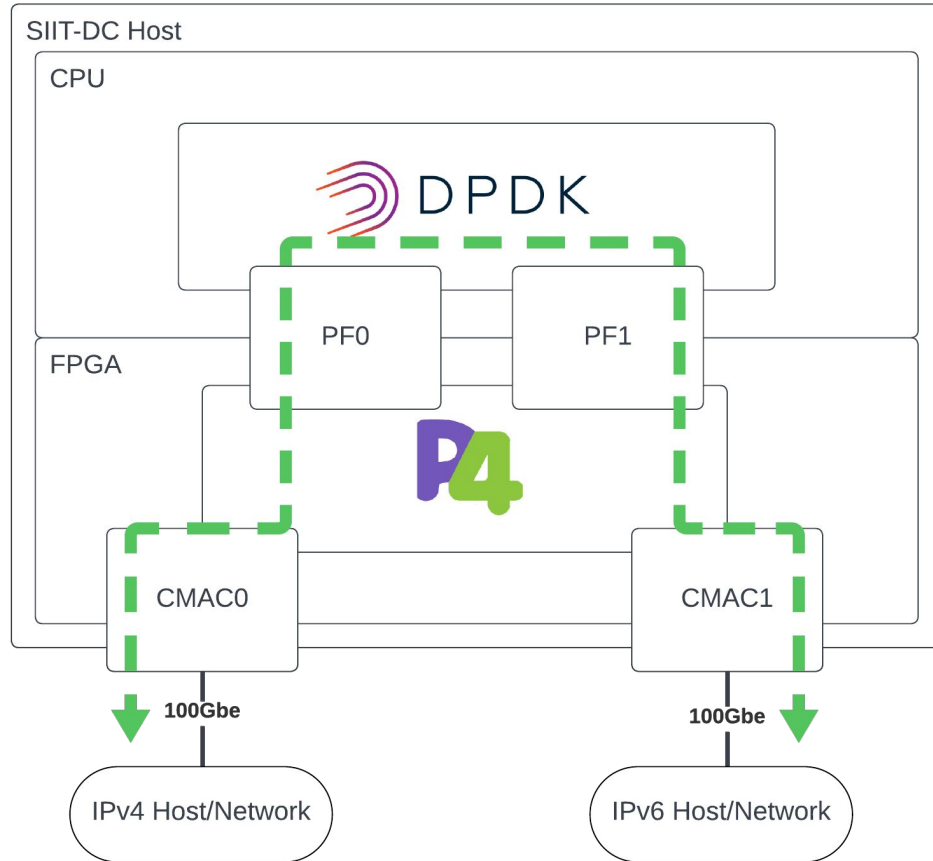


ICMPv6 Error (Variable length body!)

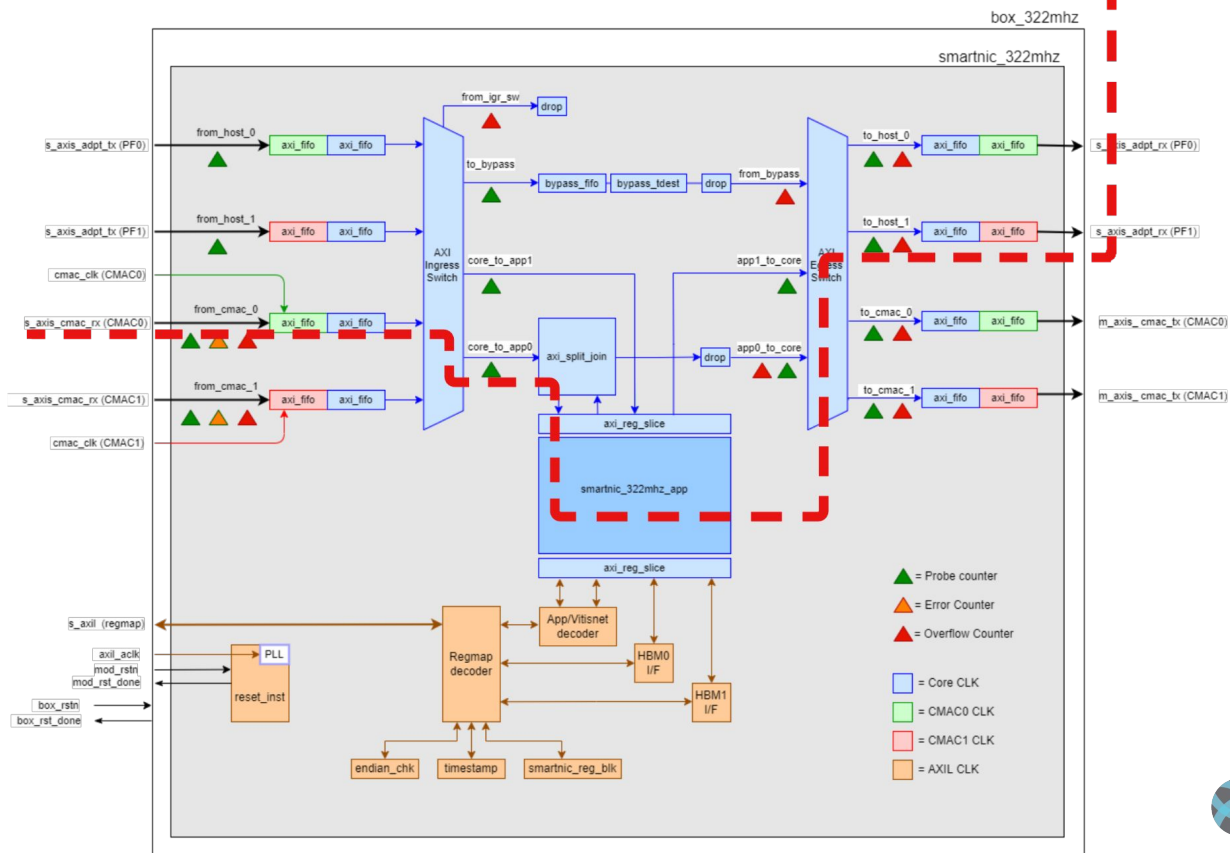
Time Exceeded

Bit offset	0-7	8-15	16-31
0	3	Code	Checksum
32	Unused		
64	Message body (Variable Size)		

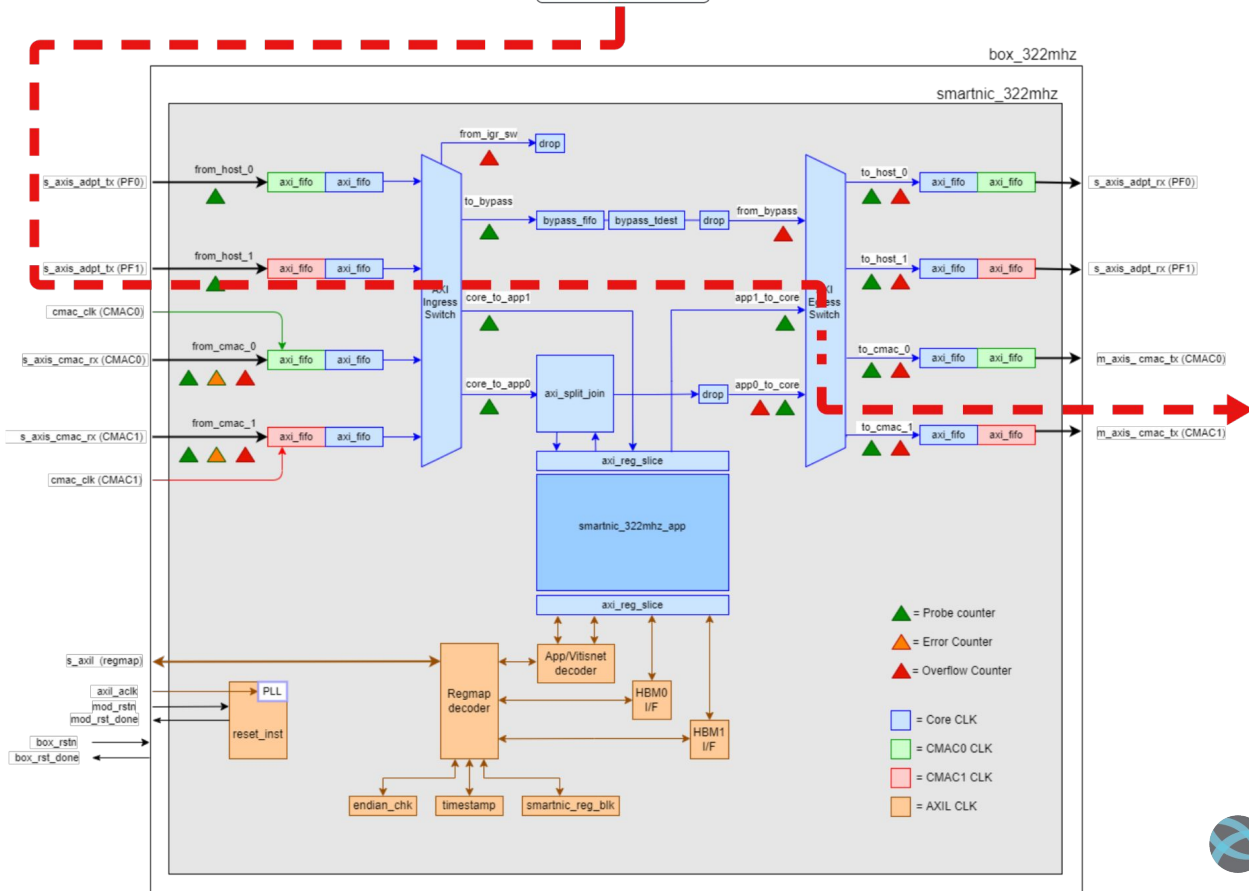
“Slow” Path Overview



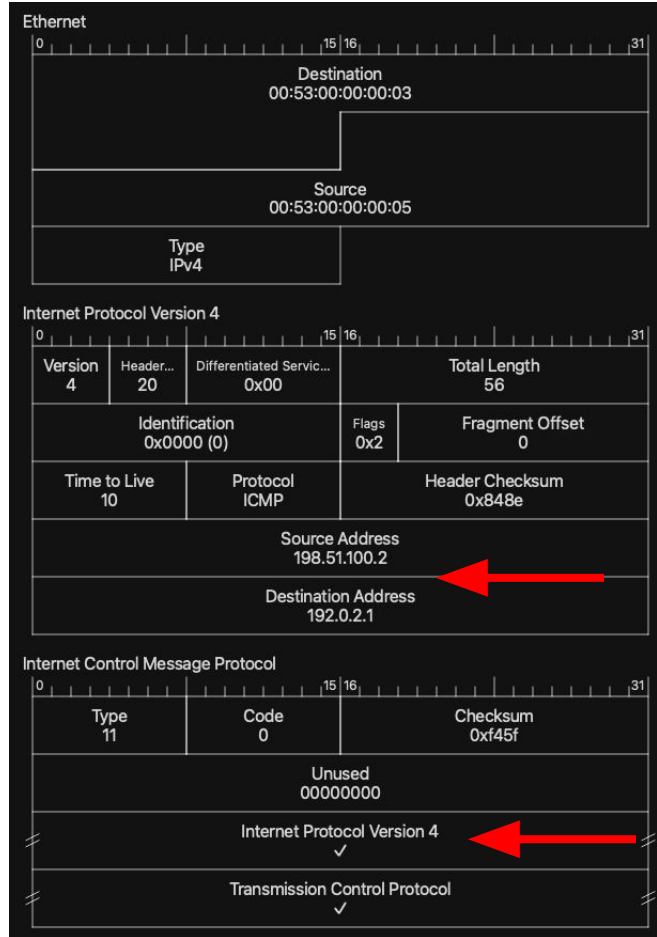
To DPDK



From DPDK



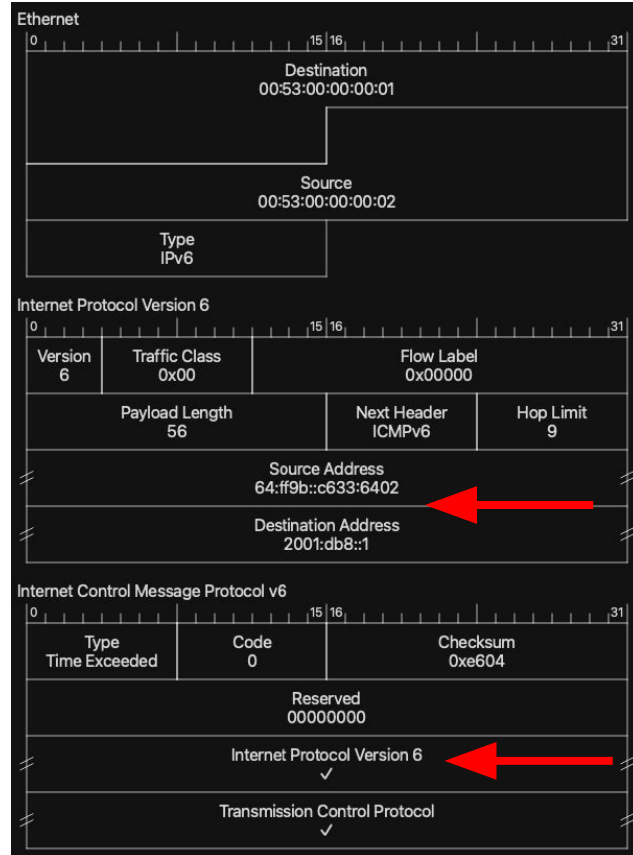
Untranslated IPv4/ICMP4 Error



Untranslated IPv4/ICMP4 Error

```
> Frame 1: 70 bytes on wire (560 bits), 70 bytes captured (560 bits)
> Ethernet II, Src: 00:53:00:00:00:05 (00:53:00:00:00:05), Dst: 00:53:00:00:00:03 (00:53:00:00:00:03)
> Internet Protocol Version 4, Src: 198.51.100.2, Dst: 192.0.2.1
  0100 ... = Version: 4
  ... 0101 = Header Length: 20 bytes (5)
> Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
  Total Length: 56
  Identification: 0x0000 (0)
> 010. ... = Flags: 0x2, Don't fragment
  ...0 0000 0000 0000 = Fragment Offset: 0
  Time to Live: 10
  Protocol: ICMP (1)
  Header Checksum: 0x848e [correct]
  [Header checksum status: Good]
  [Calculated Checksum: 0x848e]
  Source Address: 198.51.100.2
  Destination Address: 192.0.2.1
> Internet Control Message Protocol
  Type: 11 (Time-to-live exceeded)
  Code: 0 (Time to live exceeded in transit)
  Checksum: 0xf45f [correct]
  [Checksum Status: Good]
  Unused: 00000000
> Internet Protocol Version 4, Src: 192.0.2.1, Dst: 198.51.100.1
  0100 ... = Version: 4
  ... 0101 = Header Length: 20 bytes (5)
> Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
  Total Length: 28
  Identification: 0x0000 (0)
> 010. ... = Flags: 0x2, Don't fragment
  ...0 0000 0000 0000 = Fragment Offset: 0
> Time to Live: 0
  Protocol: TCP (6)
  Header Checksum: 0x8ea6 [correct]
  [Header checksum status: Good]
  [Calculated Checksum: 0x8ea6]
  Source Address: 192.0.2.1
  Destination Address: 198.51.100.1
> Transmission Control Protocol, Src Port: 80, Dst Port: 80
  Source Port: 80
  Destination Port: 80
  Sequence Number: 0
```

Translated ICMPv6 Error



Translated ICMPv6 Error

```
> Frame 1: 110 bytes on wire (880 bits), 110 bytes captured (880 bits)
> Ethernet II, Src: 00:53:00:00:00:02 (00:53:00:00:00:02), Dst: 00:53:00:00:00:01 (00:53:00:00:00:01)
v Internet Protocol Version 6, Src: 64:ff9b::c633:6402, Dst: 2001:db8::1
  0110 .... = Version: 6
  > .... 0000 0000 .... .... = Traffic Class: 0x00 (DSCP: CS0, ECN: Not-ECT)
  .... 0000 0000 0000 0000 0000 = Flow Label: 0x000000
  Payload Length: 56
  Next Header: ICMPv6 (58)
  Hop Limit: 9
  Source Address: 64:ff9b::c633:6402
  Destination Address: 2001:db8::1
  [Embedded IPv4 Prefix: 0064ff9b0000000000000000]
  [Source Embedded IPv4: 198.51.100.2]
  [Embedded IPv4: 198.51.100.2]
v Internet Control Message Protocol v6
  Type: Time Exceeded (3)
  Code: 0 (hop limit exceeded in transit)
  Checksum: 0xe604 [correct]
  [Checksum Status: Good]
  Reserved: 00000000
v Internet Protocol Version 6, Src: 2001:db8::1, Dst: 64:ff9b::c633:6401 ←
  0110 .... = Version: 6
  > .... 0000 0000 .... .... = Traffic Class: 0x00 (DSCP: CS0, ECN: Not-ECT)
  .... 0000 0000 0000 0000 0000 = Flow Label: 0x000000
  Payload Length: 8
  Next Header: TCP (6)
  Hop Limit: 1
  Source Address: 2001:db8::1
  Destination Address: 64:ff9b::c633:6401
  [Embedded IPv4 Prefix: 0064ff9b0000000000000000]
  [Destination Embedded IPv4: 198.51.100.1]
  [Embedded IPv4: 198.51.100.1]
v Transmission Control Protocol, Src Port: 80, Dst Port: 80
  Source Port: 80
  Destination Port: 80
```

Overcoming P4 Limitations

- P4 is limited by design - this is good!
- No loops or recursion
- Variable length packet processing is complex
- Reliance on externs for complex functionality
- Limited support for maintaining state

Offload possibilities

- Performing complex packet processing operations
- Prototyping externs without Verilog/VHDL
- Developing stateful applications
- Make quick table updates at runtime

Challenges Encountered

- Bit offsets are hard
- Compiling P4 app takes a long time (but bmv2 helps!)
- DPDK is very verbose compared to P4
- DPDK with C has less “syntactic sugar” compared to P4

Special Thanks!

Dr. Nik Sultana, IIT
Mohammad Firas Sada, IIT
Yatish Kumar, ESnet
Stacey Sheldon, ESnet
Scott Richmond, ESnet
Peter Bengough, ESnet

You!

Questions?

- Contact Us:
 - Sean Cummings
 - scummings@hawk.iit.edu
 - <https://www.linkedin.com/in/sean-cummings-6968061b4/>
 - Chris Cummings
 - chriscummings@es.net
 - <https://www.linkedin.com/in/chriscummingsak/>

References

- [1] Title: “Shinkansen N700 with Mount Fuji”, Author: [tansaisuketti](#), Source: [WikiMedia Commons](#), License: [CC BY-SA 3.0](#)
- [2] Title: “High Five”, Author: [austrini](#), Source: [WikiMedia Commons](#), License: [CC BY 2.0](#)
- [3] Title: “Alveo U280 Data Center Accelerator Card”, Author: AMD Xilinx, Source: [AMD Xilinx U280 Product Page](#)