

Can trading transition to 25G?

Arista Networks

David Snowdon, Q2 2023
daves@arista.com



The obligatory marketing slide

- OpenFDK: An Open-source FPGA developer's kit for Arista switches
 - <https://github.com/aristanetworks/openfdk>
- SwitchApp is Arista's ~132 ns Layer 3 switch*
 - Released in mainline EOS.
 - PTP boundary clock is now supported.
 - Dynamic NAT coming in Q3.
- 40G line-side timestamps
 - Timestamps on 40G are now properly precise.



* Not a STAC benchmark
Source: Arista



Introduction

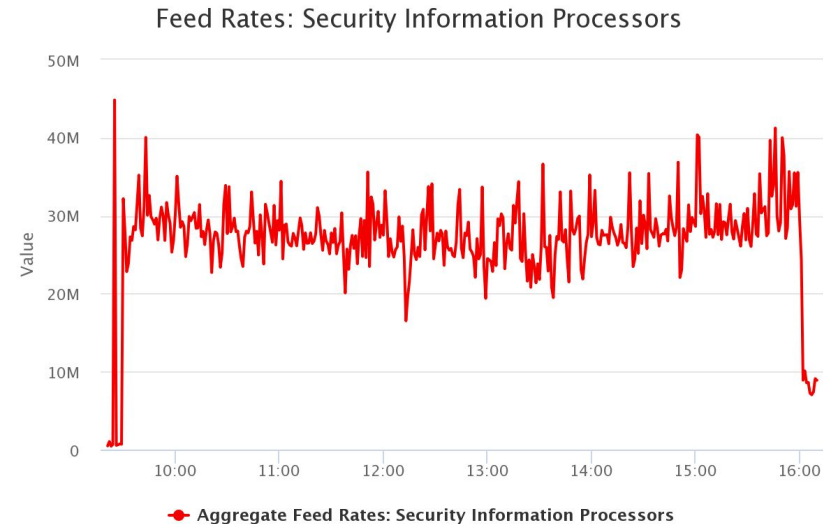
Introduction

- 10G Ethernet is the predominant standard for connecting exchanges.
 - 100M, 1G, and 40G are also used.
- Arista's early success in 2008 came from 10G.
 - First low-latency 10G switch announced in November, 2007 (7124S)
 - A series of low latency switches transformed trading: 7124S -> 7124SX -> 7150
 - 16 years at 10G.
- Datacentre switching has moved on.
 - 25G/100G is the new 10G. 400G is the new 40G. 800G is the bleeding edge.
 - 10G switches are now end of life – Exchanges buy 25G and limit to 10G!



Why change? – Background

- More market activity means more messages, means more bandwidth.
- More markets/venues means more messages, means, more bandwidth.
- Market events tend to be correlated:
 - so the peak bandwidth is very high;
 - and the average bandwidth is moderate.



NOT STAC BENCHMARK

Source: <https://marketdatapeaks.net/>
1ms burst rate max, 30th March, 2023

Aside: what is bandwidth, anyway?!

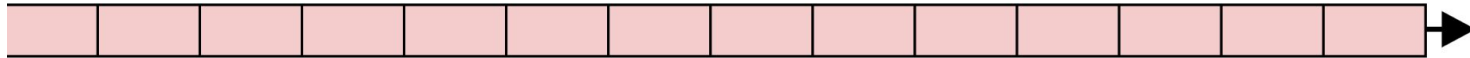
- Bandwidth is always measured as an average over a period of time.
 - When a packet is being sent, the line is 100% utilised.
 - When a packet is not being sent, the line is 0% utilised.
 - Over a period of time, the line utilisation is an average.
- At some level of granularity we always hit “line rate”
 - Any time a packet is queued, the rate at which the egress link operates dramatically affects the latency.

Aside: what is bandwidth, anyway?!

- Assuming 10G, this is ~3.3Gbps:



- And this is 10Gbps:

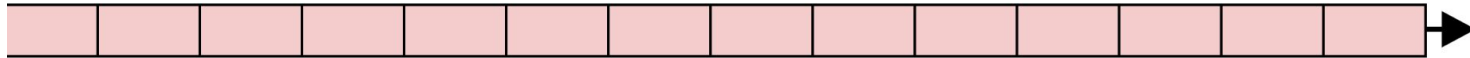


Aside: what is bandwidth, anyway?!

- Assuming 10G, this is ~3.3Gbps:



- And this is 10Gbps:



- So what is this?

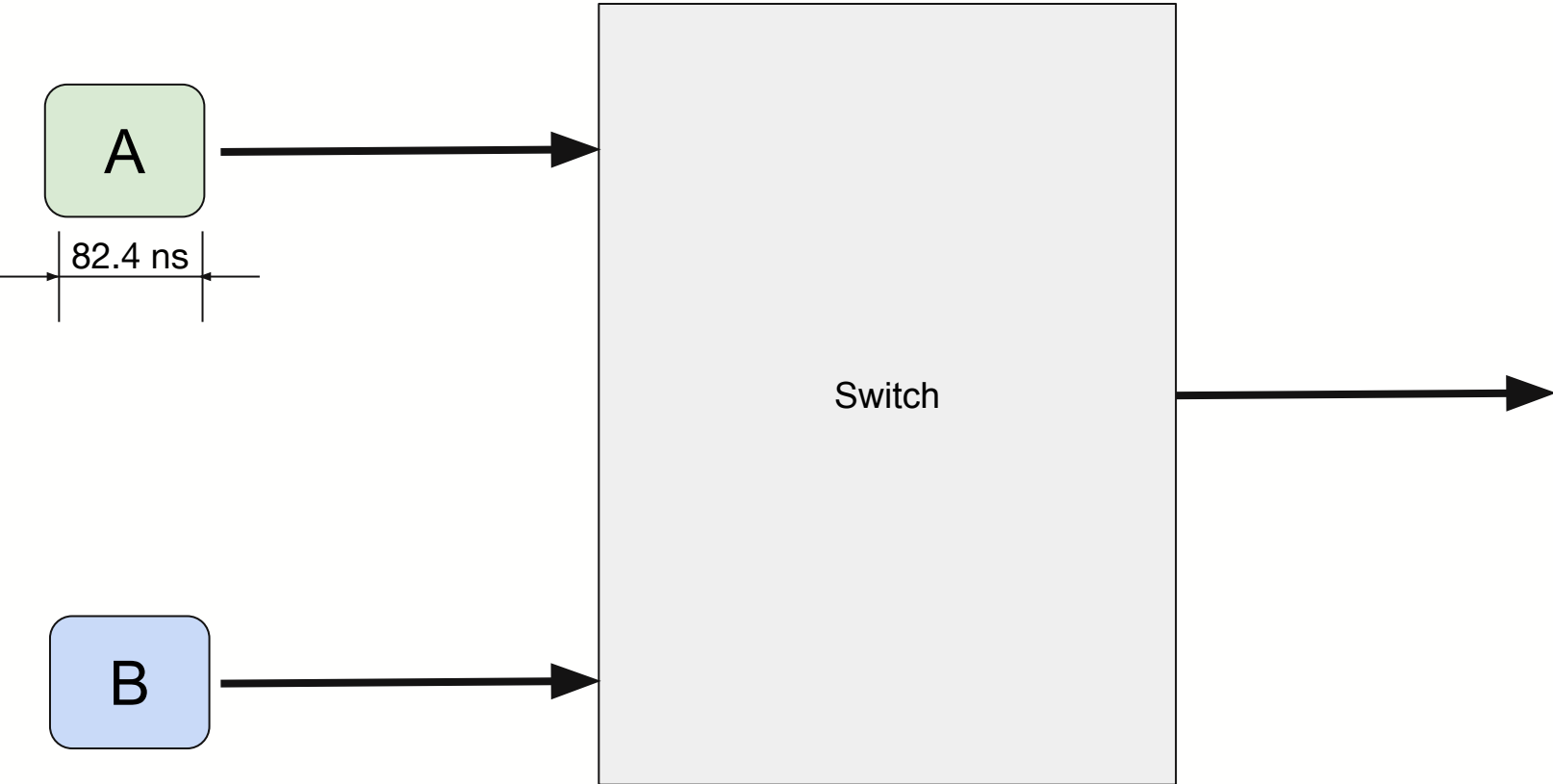


Aside: what is bandwidth, anyway?!

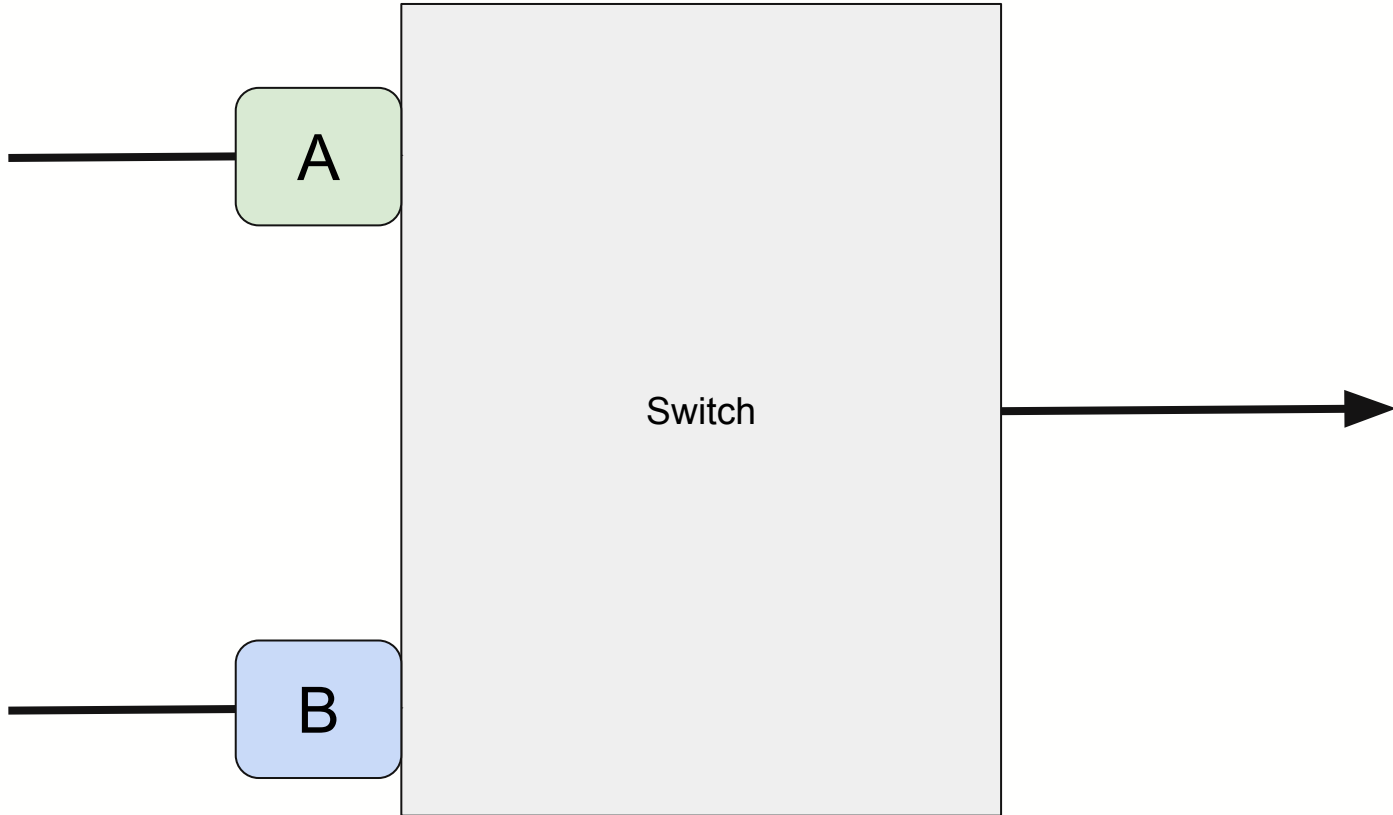
- `Gigabits per second` only makes sense as an average over a time period.
 - Every packet is sent at line rate.
- What period of time do we care about?
 - How many packets in a row, before the latency is enough to care?

Why bandwidth matters: queueing

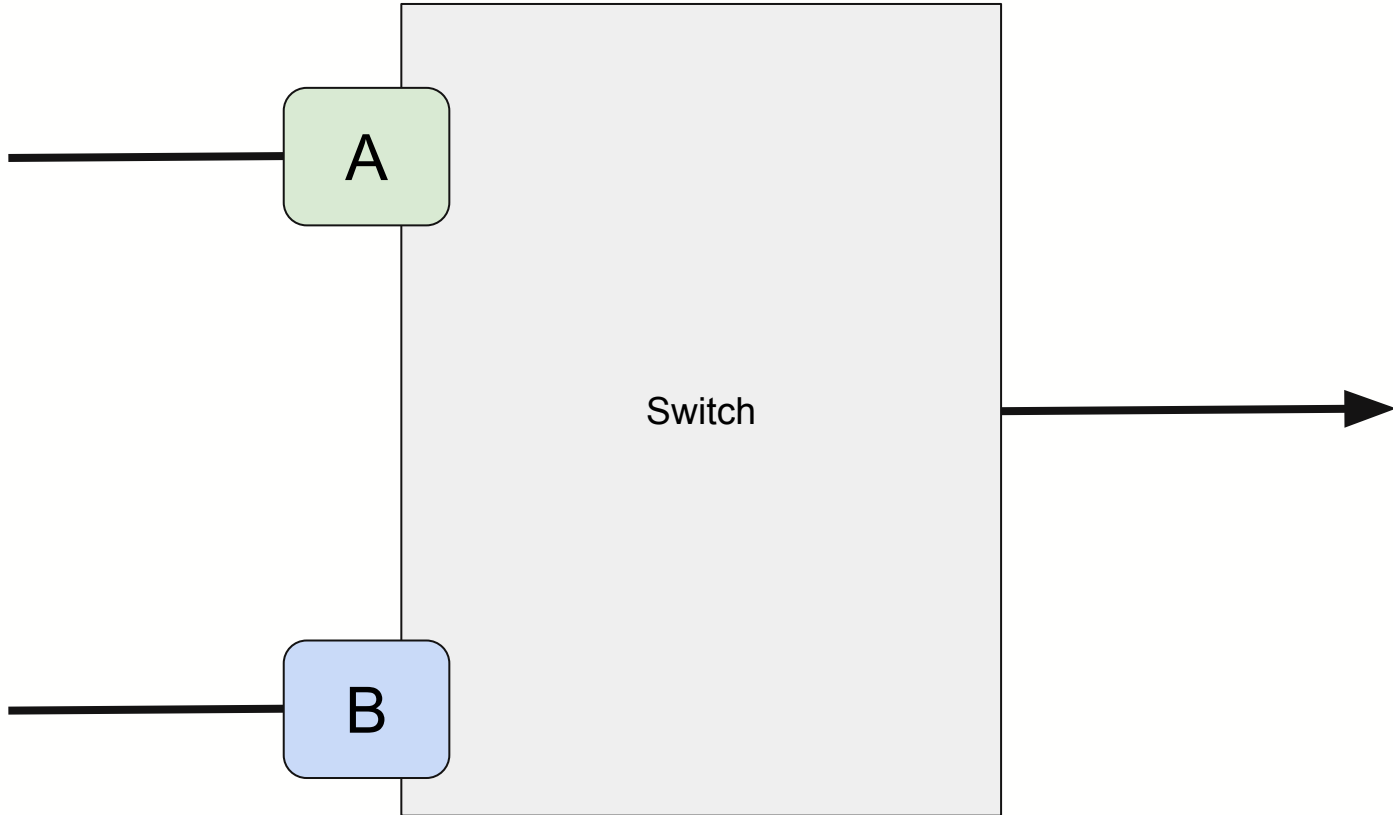
- If two packets are sent at the same time, they are serialised
 - One packet is sent first, then the other.
 - In the meantime, the other packet sits in a queue buffer – in the host or a switch.
 - If many packets need to be sent at the same time, they queue for longer.
- The serialisation delay of packets directly affects latency:
 - Minimum 64 byte packet nanos: at 10G – ~60 ns, at 25G – ~24 ns.
 - Realistic 256 byte packet nanos: at 10G – ~250 ns, at 25G – 100 ns.
- Worst-case KRX in 2012 running on aggregated T1/E1 links – each packet occupies *milliseconds*. Queuing delays may be **seconds**.



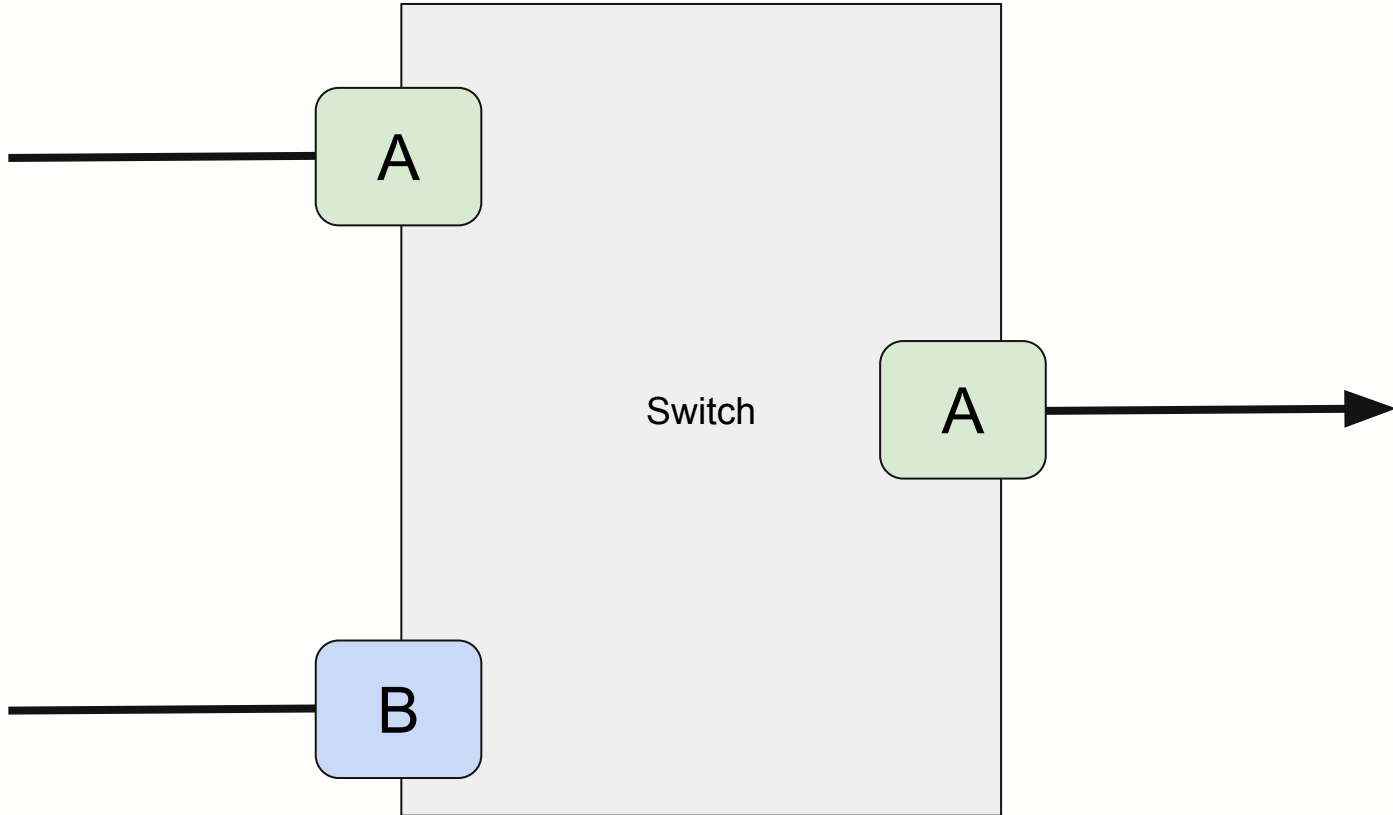
$t = 0$ ns



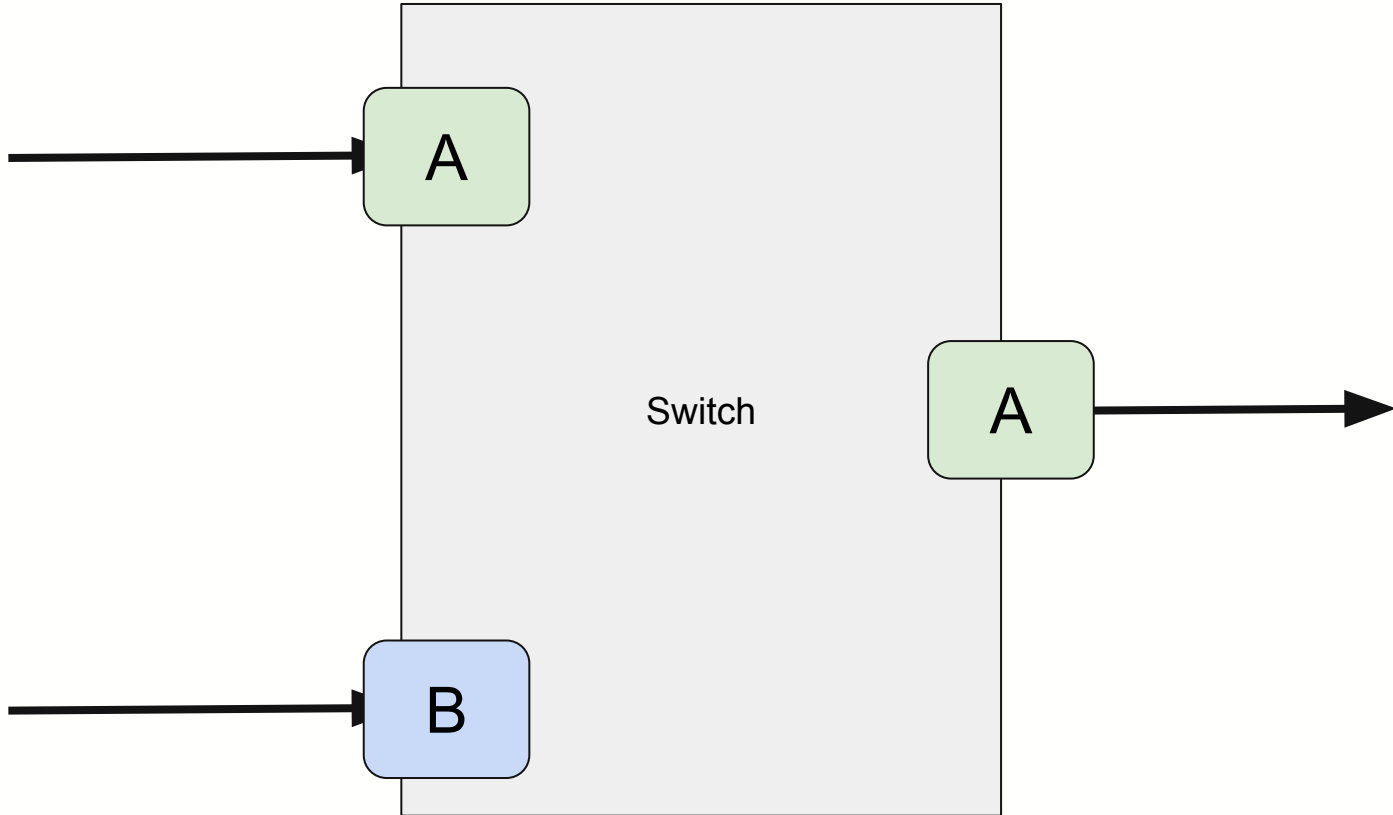
t = 300 ns



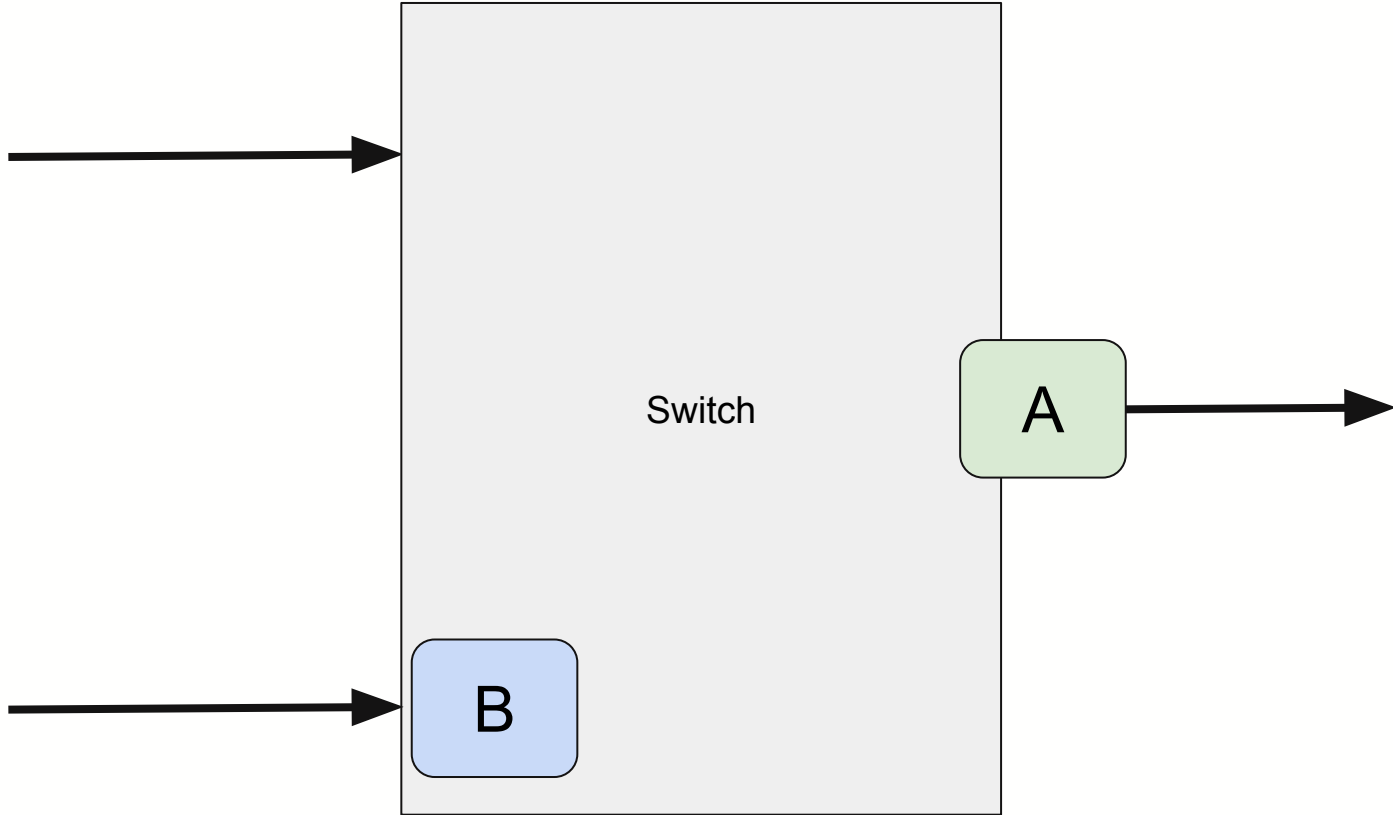
t = 330 ns



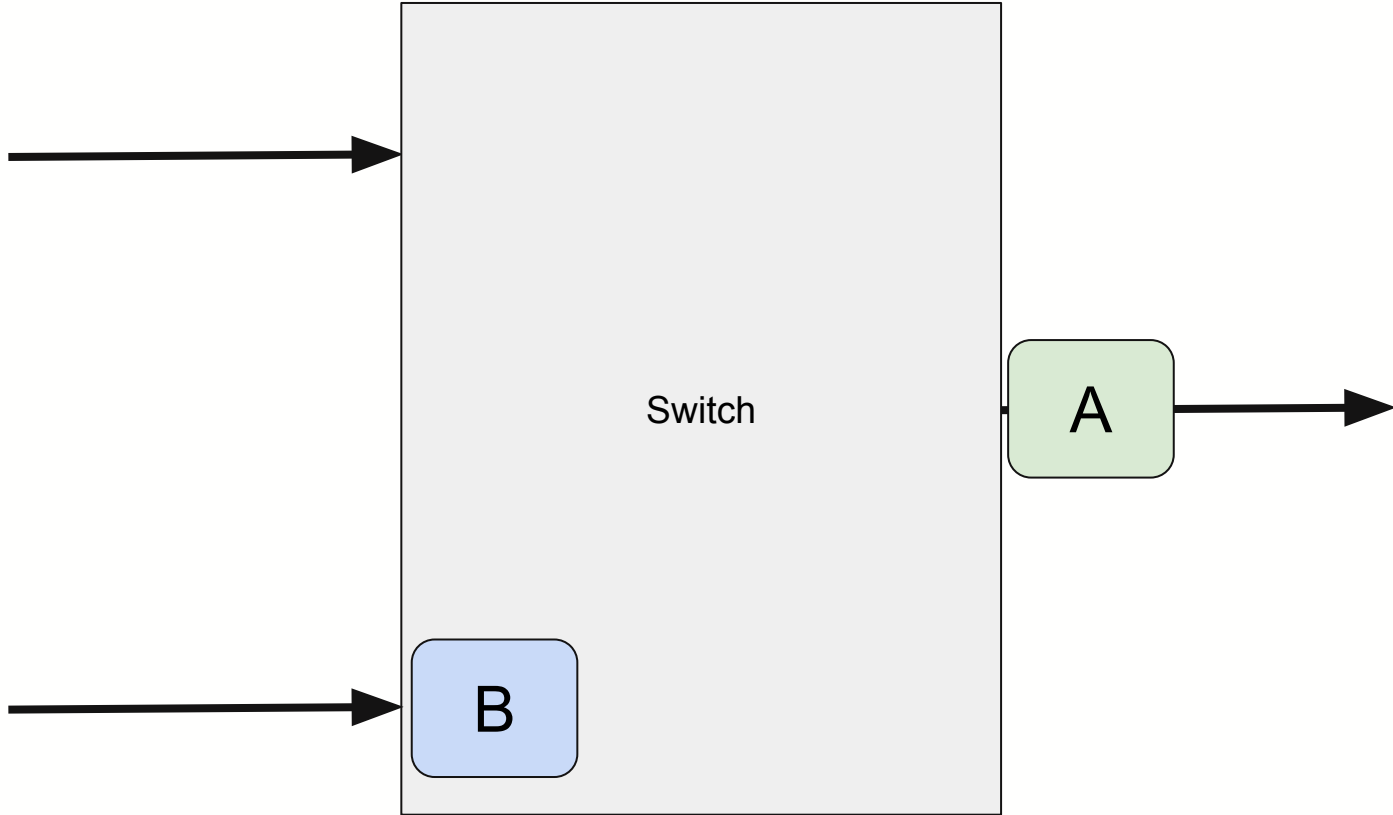
t = 350 ns



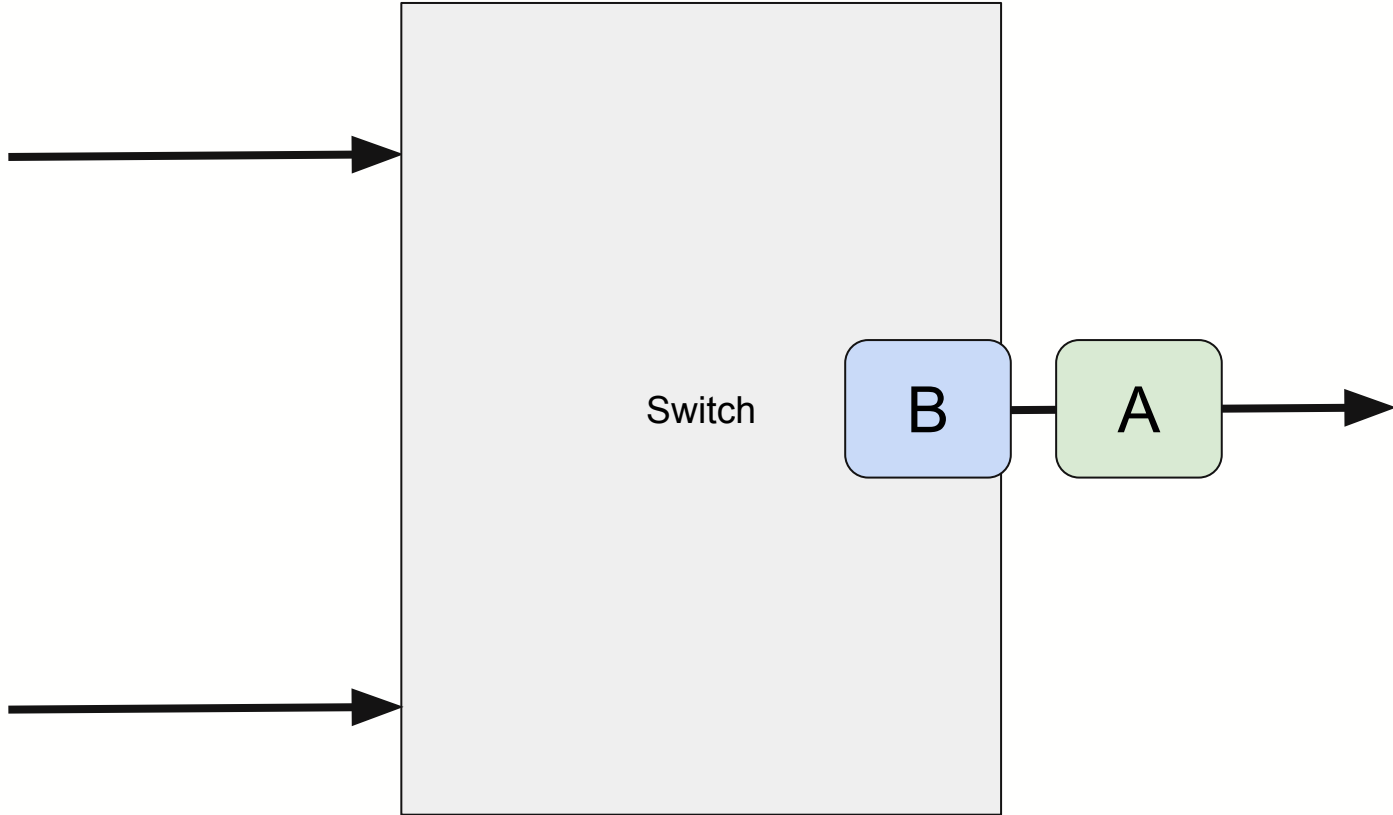
t = 370 ns



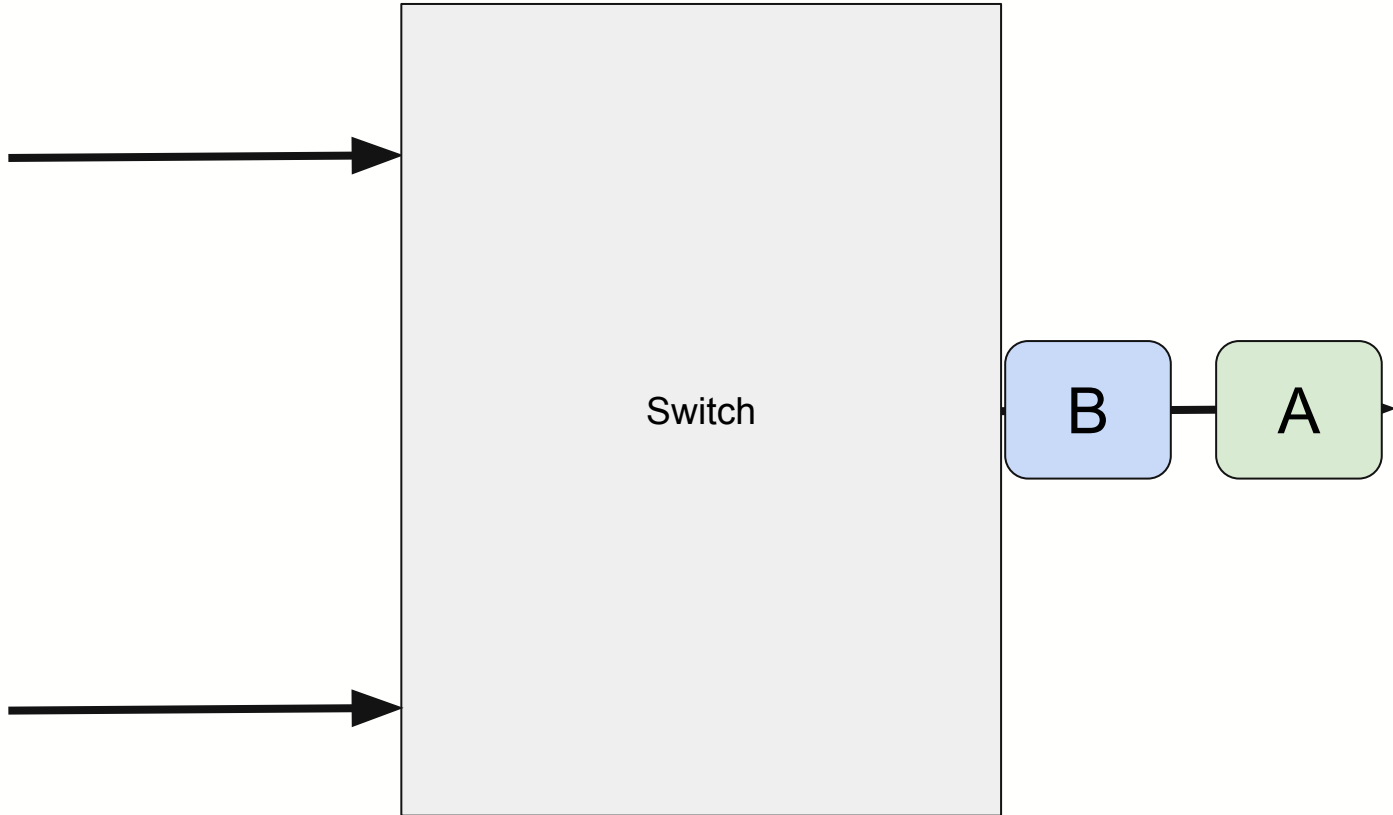
$t = 382.4 \text{ ns}$



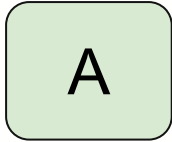
t = 412 ns



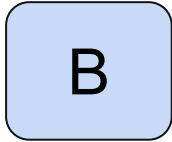
t = 430 ns



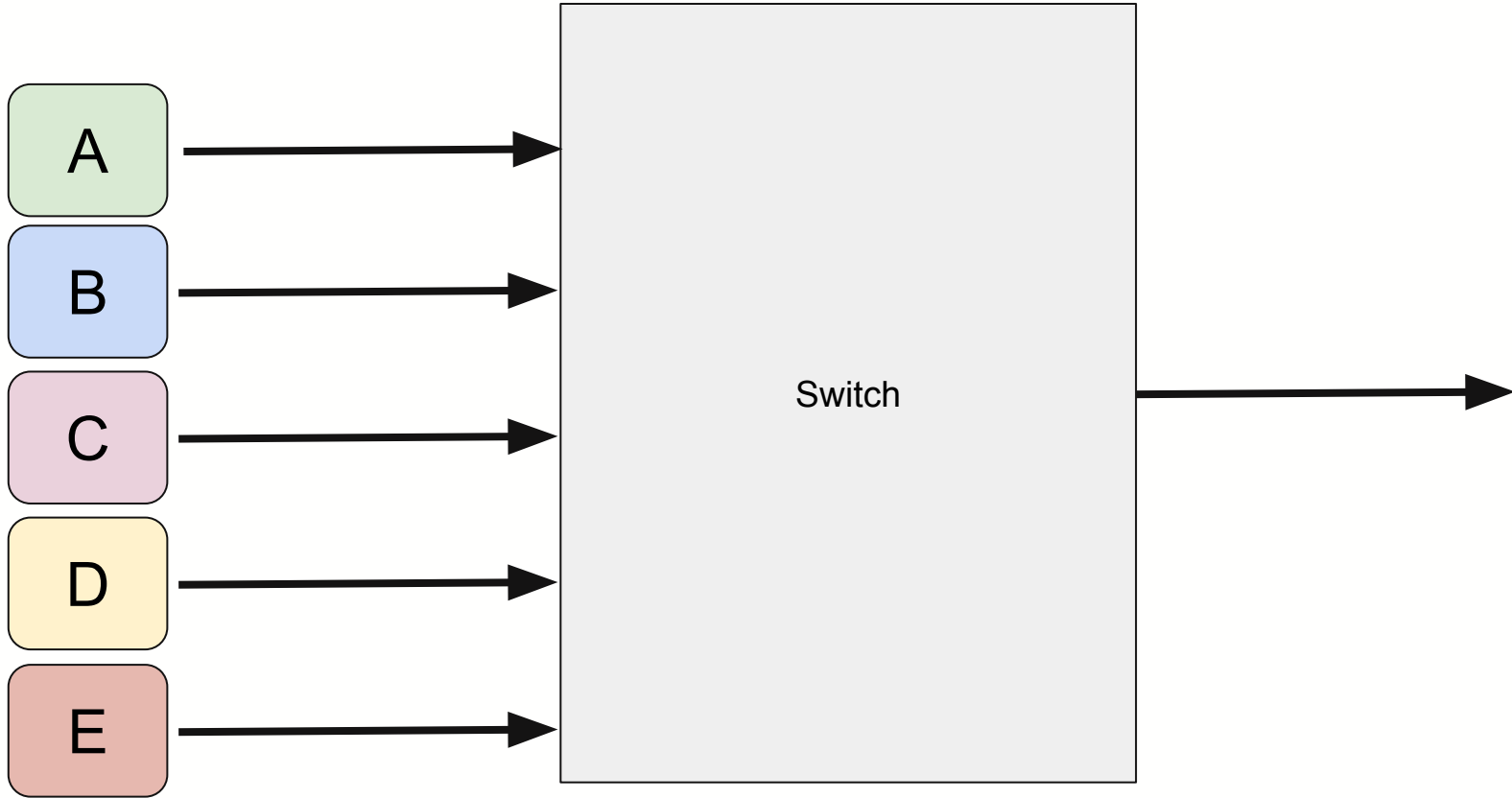
t = 494.4 ns

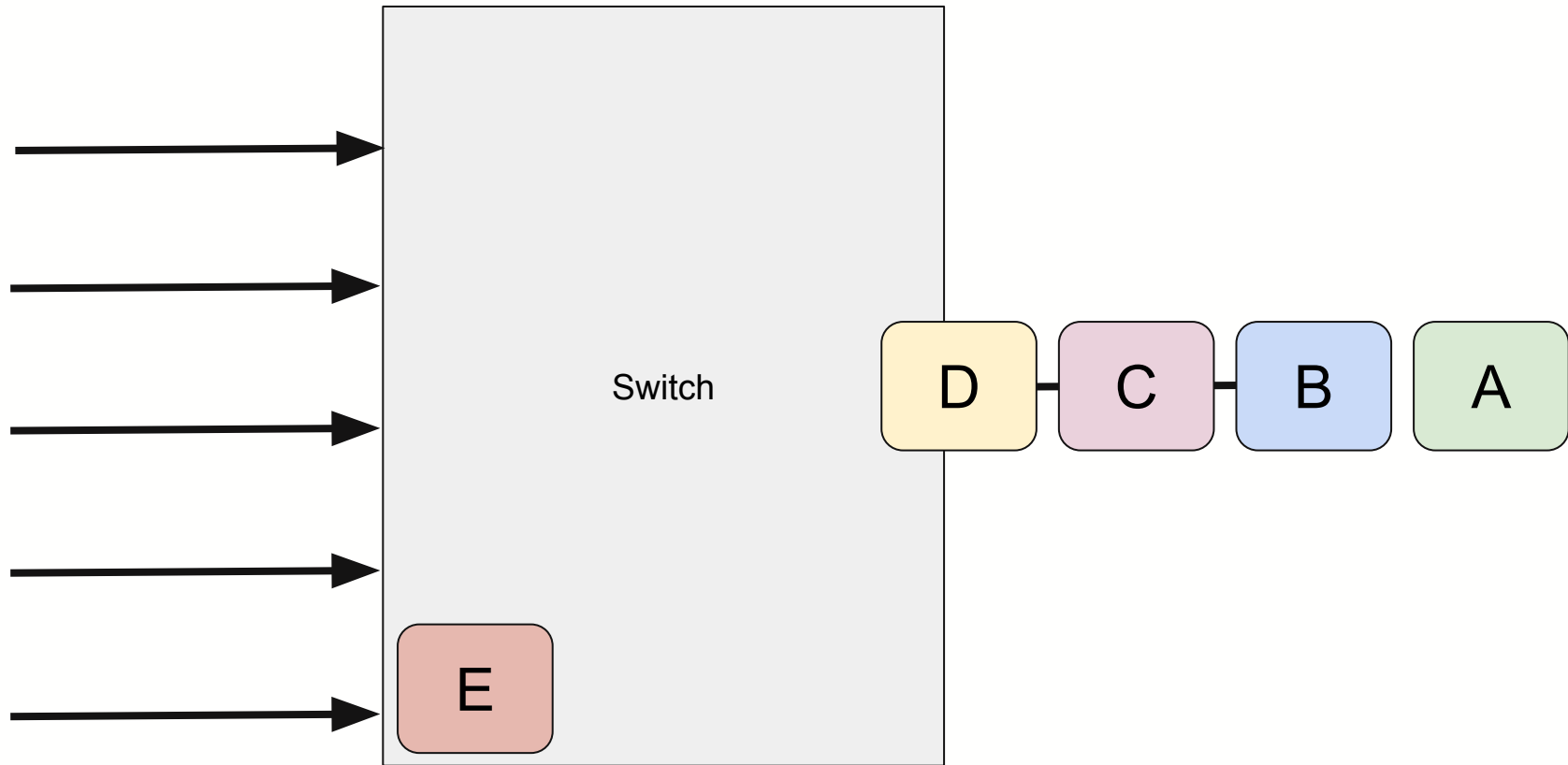


Starts to arrive: 380 ns
Finishes arriving: 462 ns



Starts to arrive: 480 ns
Finishes arriving: 562 ns





A	Starts to arrive: 380 ns Finishes arriving: 462 ns
B	Starts to arrive: 480 ns Finishes arriving: 562 ns
C	Starts to arrive: 580 ns Finishes arriving: 662 ns
D	Starts to arrive: 680 ns Finishes arriving: 762 ns
E	Starts to arrive: 780 ns Finishes arriving: 862 ns

So, why change?

- Higher serialisation rate (i.e. 25 vs. 10) means:
 - Reduced likelihood of collisions/queuing.
 - Lower queuing delays during periods of line contention.
 - More reliable packet delivery if we run out of buffer space.
- Traders like better predictability, lower risk.
 - So, exchanges and traders want to reduce queuing latency or drops.
 - Waiting for packets to be serialised sucks.
- Arista don't really sell 10G switches any more (for data centres).
 - Many exchanges use traditional network switches.
 - Exchanges buying networks are artificially limiting to 10G.



Where to next?

The path forward

- Finance connectivity is defined by venues.
 - Changing rates means store-and-forward switching.
 - Venues are the gatekeepers to higher bandwidth connections.
- Venues are influenced by their customers – traders, brokers.

Core Proposition: trading should move to 25Gb Ethernet

25G Ethernet



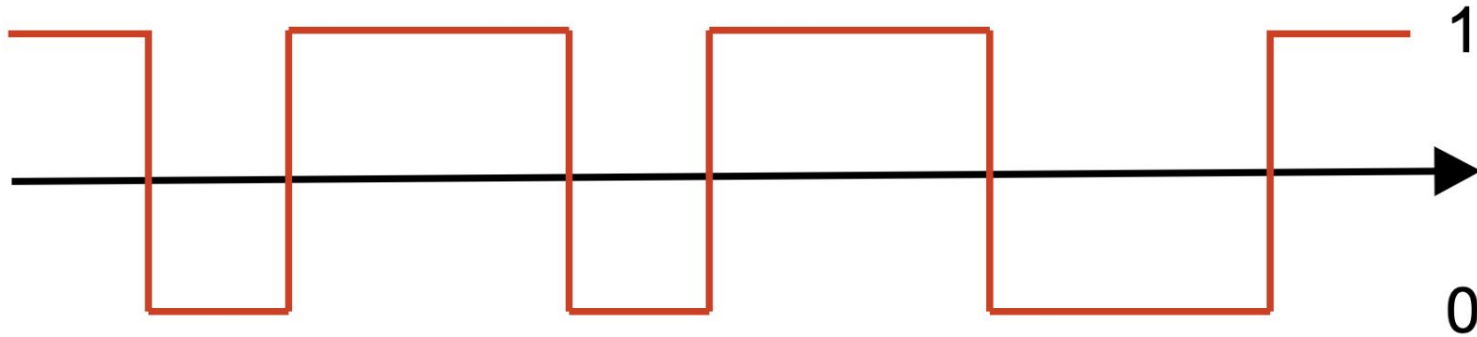
- 25G Ethernet:
 - IEEE 802.3by was approved in June, 2016.
 - 25.75GHz line rate.
 - 64b/66b encoding.
 - Very similar to 10GbE, but 2.5x the rate.
 - Similar array of physical media: copper, SR, LR, ER optics. Similar ranges.
- FEC
 - Higher signal rates over long distances may require embedded error correction.
 - But for the *right media*, or *shorter distances*, with *quality hardware*, it's **not** required.
- Implications:
 - 2.5x the bandwidth.
 - Seamless upgrade path – install 10/25G switches and upgrade via config.
 - Avoid recabling.

Why not something else?

- Why not 40G?
 - 40G is 4x 10G links, with each packet spread across all four links.
 - Some exchanges (e.g. NYSE) have used 40G to resolve congestion on data feeds.
 - Some firms use 40G (or higher) internally – but this means store-and-forward switching to 10GbE exchange links.
 - 40G requires four fibres per connection and often not software configurable to 10G.
 - 40G is more complicated and slower in FPGA/ASIC due to encode/decode.
- Why not 100G?
 - 100G is 4x25G links. See 40G.
 - Requires FEC per the standard.
- Why not 400G?
 - 400G uses PAM4 encoding – four different levels represents two bits per clock.
 - This requires more complicated (much slower) decoding than NRZ.
 - FEC is a requirement for all media.
- Why not 50G?
 - It uses PAM4 – See 400G.

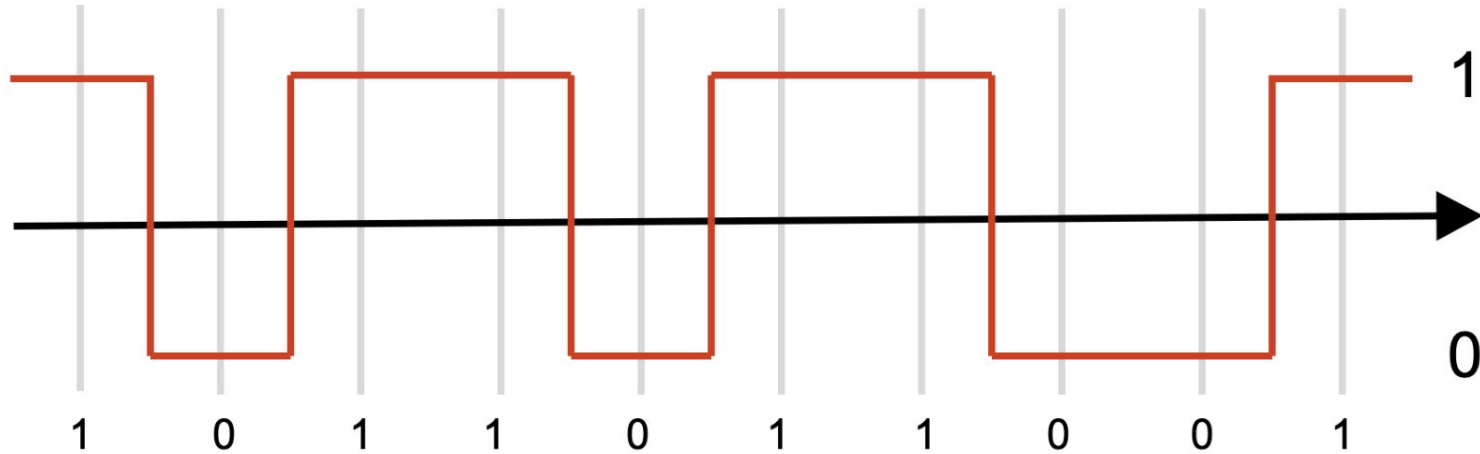
Aside: What is PAM4?

- Non-return-to-Zero (NRZ) signalling uses two states on the fibre/cable.
 - Each state represents one or zero.



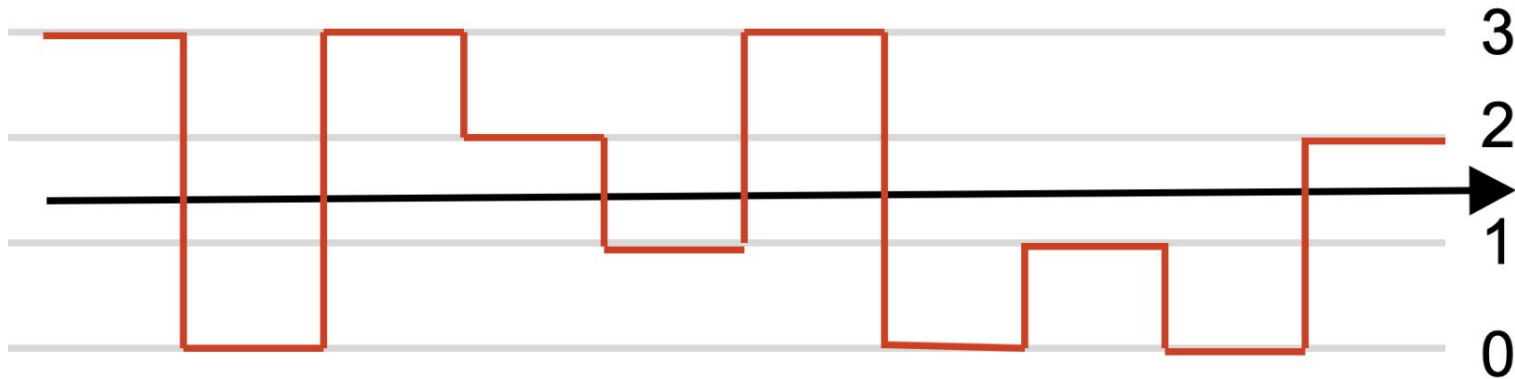
Aside: What is PAM4?

- Non-return-to-Zero (NRZ) signalling uses two states on the fibre/cable.
 - Each state represents one or zero.



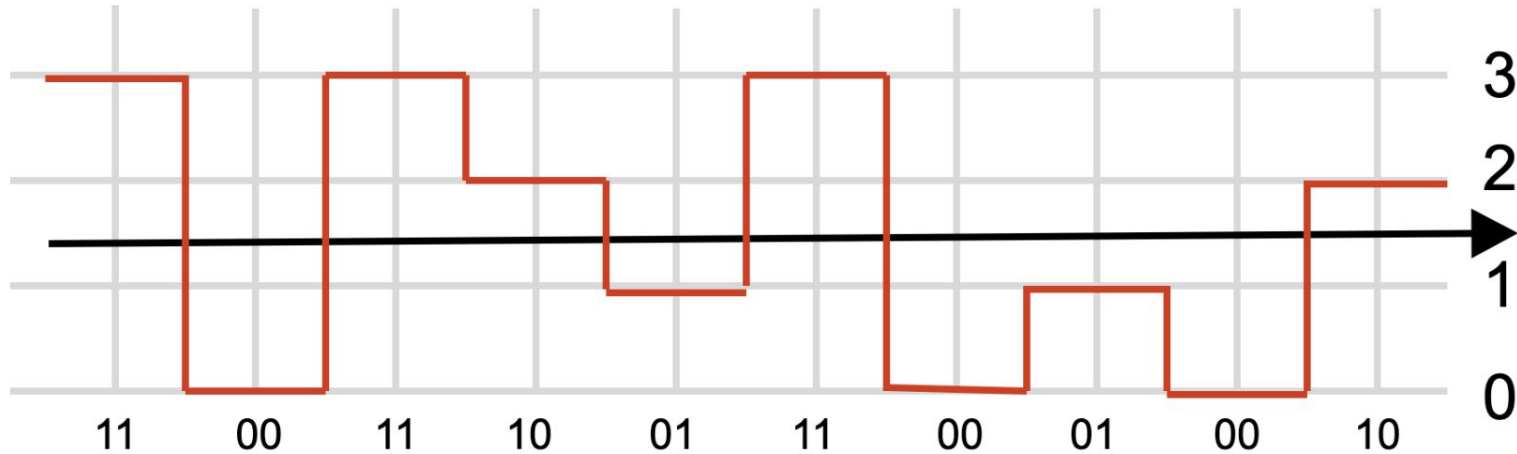
Aside: What is PAM4?

- PAM4 signalling uses four states on the fibre/cable.



Aside: What is PAM4?

- So it encodes two bits per clock period.



Why not something else?

- PAM4 means:
 - High latency serdes (complex hardware to receive a complex signal)
 - FEC
- 25G is a **magic combination** of
 - single fibre;
 - NRZ (i.e. low latency);
 - optional FEC.

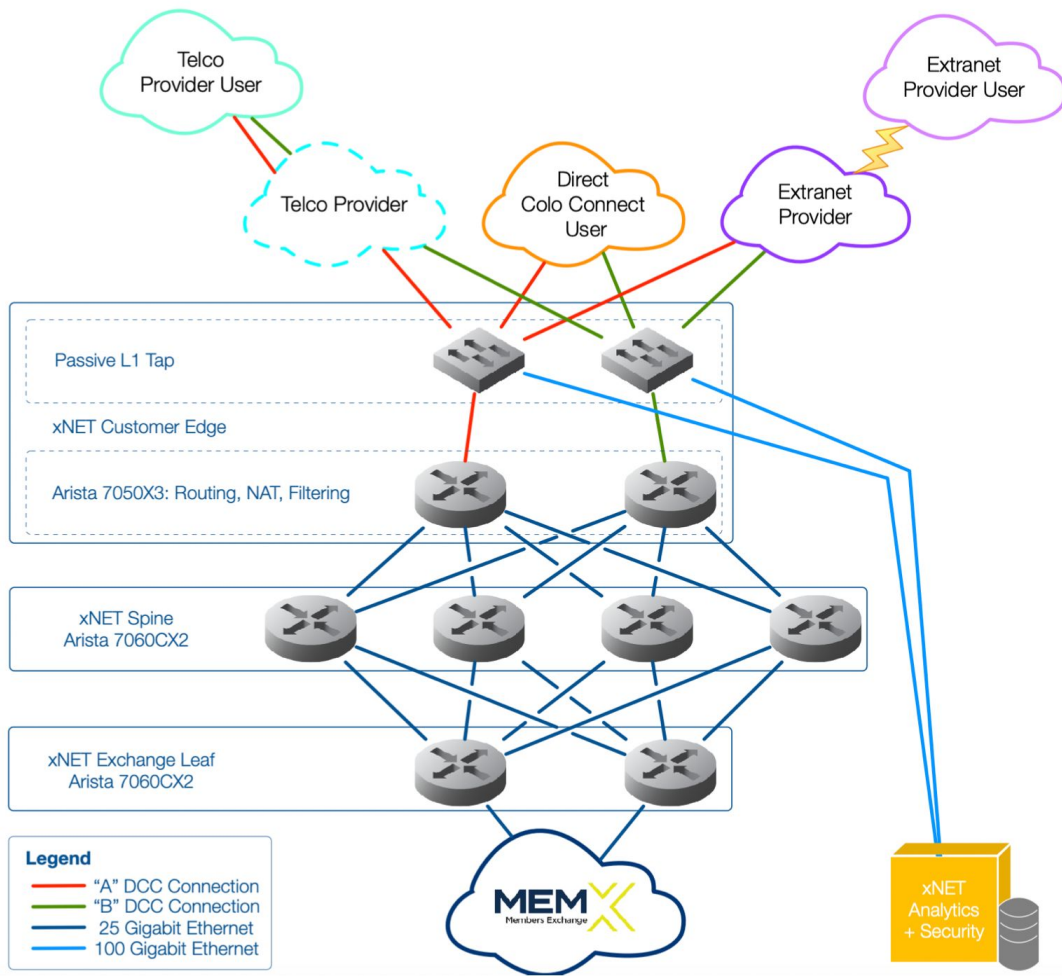


Why hasn't it happened yet?

- Exchanges have tried...
 - E.g. MEMX – <https://www.memxtrading.com/connect>
- But customers have said “no”
 - there aren't any suitable L1 switches
 - L2/3 switches are higher latency than 10G equivalents.
 - it is going to require work for traders to upgrade (but maybe that's OK).
 - it is going to be expensive for traders to upgrade (but maybe that's OK).
- The key reason is an absence of network devices or ecosystem at 25G.
 - We can help!

Exchanges will enable 25G once there are devices to use.

MemX



Source: <https://www.memxtrading.com/connect>



How to get there...

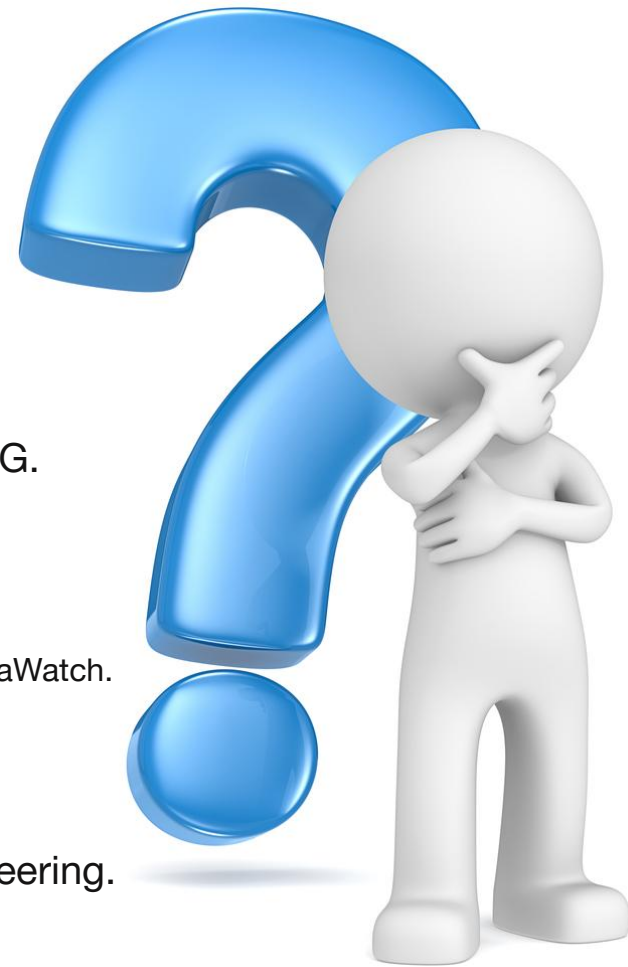
25G – The existing ecosystem



- Standard datacentre networking products
 - 25G is a standard server interconnect, gradually giving way to 100G.
 - Optics, NICs, fibre, etc, are well established.
- Network cards
 - Several low latency vendors (e.g. Nvidia, AMD, Cisco)
- Switches
 - 25G options – e.g. Arista (Tomahawk 2, Trident 3, Intel Tofino), Nvidia Spectrum.
- Network capture
 - Optical taps, capture cards, standard switch-based capture. Less precision.

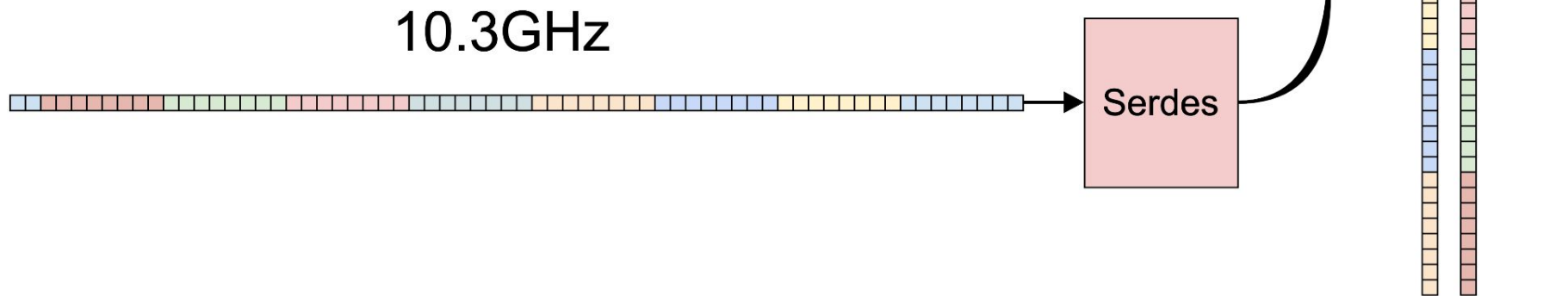
25G – The missing pieces

- Layer 1 switching
 - Optical tapping works well at 25G.
 - But large-scale tapping, broadcast, patching.
- Switching
 - Current L2/3 switches have latencies of 600 ns+ at 25G.
 - Low latency muxing, filtering, firewalls.
- Capture/Timestamping
 - Highly accurate, large-scale timestamping, analogous to Arista MetaWatch.
- FPGA application platforms
 - FPGAs have been 25G capable for a long time.
 - But FPGA applications may need substantial re-engineering.
 - But creating a basic interface is trivial.



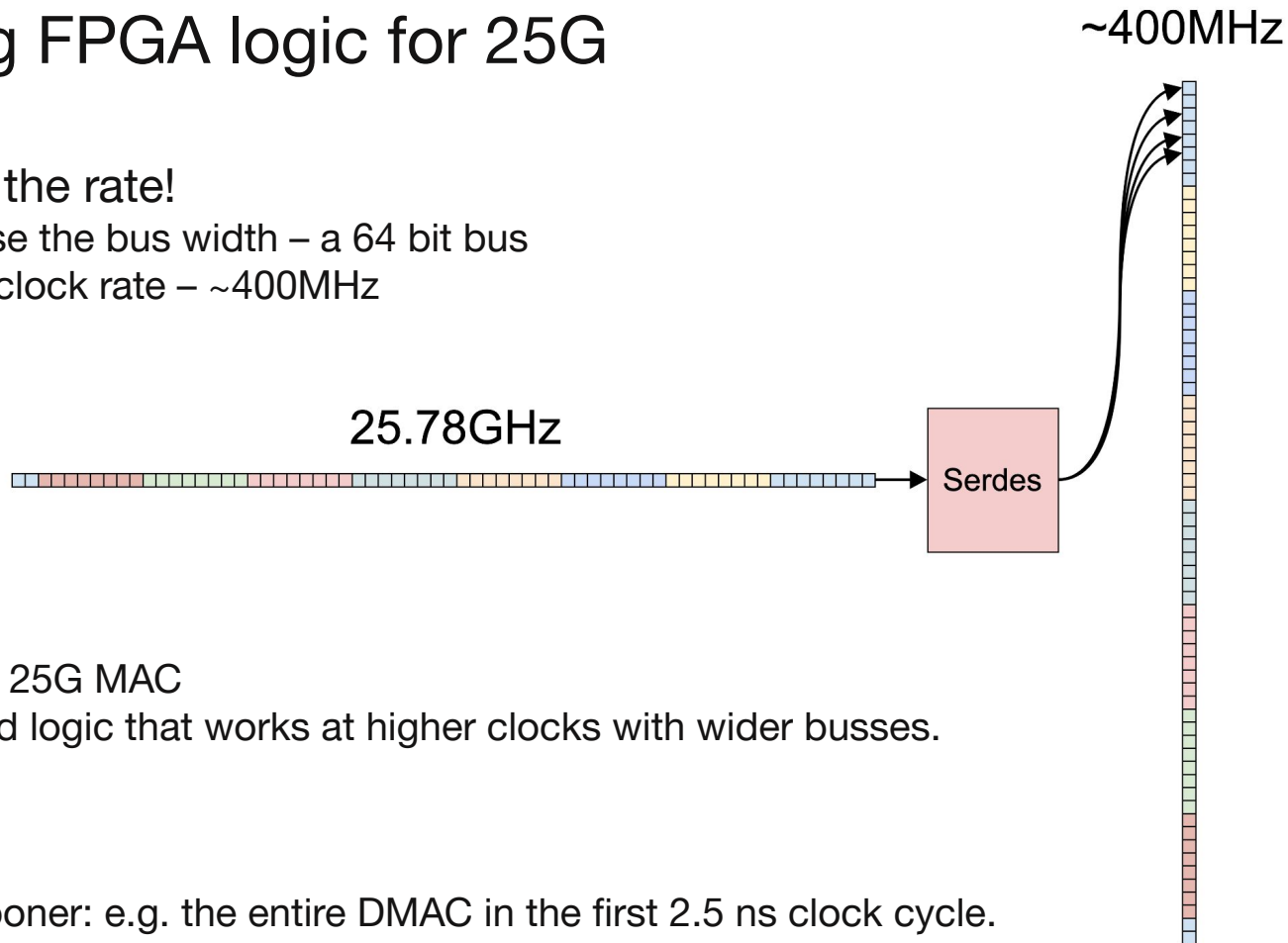
Re-architecting FPGA logic for 25G

- FPGAs operate at max freq ~700MHz
 - But that's really hard, since a high proportion of each cycle is overhead
 - ~300MHz is much more comfortable.
- SERDES reduce serial to parallel:
 - 10G MACs often use a 32-bit interface – 64 bits in two cycles



Re-architecting FPGA logic for 25G

- But 25G is 2.5x the rate!
 - So we increase the bus width – a 64 bit bus
 - And a higher clock rate – ~400MHz



- So we need:
 - A low-latency 25G MAC
 - Re-engineered logic that works at higher clocks with wider busses.
- And we get:
 - More data, sooner: e.g. the entire DMAC in the first 2.5 ns clock cycle.



FEC

But what about FEC?

- Problem: at 25G the signal integrity is less good than at 10G.
- So IEEE specifies optional Forward Error Correction
 - Two types – RS-FEC or Clause 74 FEC. (~200 ns vs. ~80 ns latency).
- Finance applications won't like the latency
 - So we need to be able to run without!
 - IEEE standard makes FEC optional for some cable types.
- Running without FEC means:
 - Maintaining high signal integrity, shorter cable runs, higher quality optics.
 - System level testing and qualification of error rates for different cables.
 - Validation and testing.

Some numbers

Property	Units	Clause 91 RS-FEC 4 lanes (100 Gb/s)	Clause 91 RS-FEC 1 lane (25 Gb/s)	Clause 74 FEC 1 lane (25 Gb/s)
Block size	Bits	5280	5280	2112
Block time	ns	51	205	82
Latency for error correction (marking bypassed)	# Blocks	~2	~1.25	1
	ns	~100	~250	82
	Equivalent m of cable	20	50	16
Latency for only error marking (correction bypassed)	# Blocks	1	1	1
	ns	51	205	82
	Equivalent m of cable	10	40	16
Input BER for FLR≈6e-10		1e-5	1e-5	1e-8
Supported cable length (26 AWG)	m	5	5	3

Some NIC vendors look at this...

- Solarflare (now AMD) at STAC in 2018:
 - Note: 4.5 years ago!
 - <https://stacresearch.com/system/files/resource/files/STAC-Summit-15-Nov-2018-Solarflare.pdf>
- IEEE spec says FEC is optional – we need to make sure we can disable FEC wherever we can.
 - FEC can be off within the rack.
- Qualification and validation are key
 - Will exchanges require FEC for ~300m 25GBase-LR single mode fibres?

But will there be side effects?

- Does using FEC create an uneven playing field – those who are willing to take the risk, vs those who are not?
 - Is this different to those willing to trade before receiving an Ethernet checksum?
- Or will FEC-arb become a new advantage?
 - One line with FEC, and one without?
- 25G products tend to be faster than 10G products.
 - When processing data, having more of it, earlier, makes things faster.



Conclusions

Concluding...

- When the ecosystem can manage, exchanges will upgrade.
 - 25G L1 and FPGA switches
 - Qualification, validation.
 - Trading firm upgrades.
- There are huge benefits to moving to 25G.
 - 2.5x the sustained bandwidth.
 - Reduced queueing latency, packet drops.
 - Lower latency products.
- It's time to take actions:
 - Get comfortable with 25G – deploy it for internal infrastructure.
 - Test and buy 25G-ready hardware.
 - Engineer your FPGA and network systems assuming a 25G future.





Thank You

www.arista.com