We really move YOUR tail for you

David Lucey - Salesforce

HPTS 2024

Networks *weren't* optimizing for you.

- **Networks maximized bandwidth and number of talkers**
  - Latency improvements were a distant second

- **Latency *did* improve as 'speeds and feeds' got better**
  - And most said 'that should be good enough', except....

GPUs are changing that - and you will benefit from it

How the needs changed

Network's Use-case

- High Entropy - Many talkers
- Smooth transmission rates - the law of large numbers takes over
- Loss Tolerant, as long as it is rare
- Latency 'tolerant', within reason

GPUs' Use-case

- Low Entropy - Few talkers
- Bursty - from zero to wire speed and back
- Loss Intolerant
- Latency Intolerant

How the needs changed

Network's Use-case

- High Entropy - Many talkers
- Smooth transmission rates - the law of large numbers takes over
- Loss Tolerant, as long as it is rare
- Latency 'tolerant', within reason

GPUs' Use-case

- Low Entropy - Few talkers
- Bursty - from zero to wire speed and back
- Loss Intolerant
- Latency Intolerant

- GPU farms are Very Expensive
  - Idle hardware is expensive

- Training failure is Ridiculously Expensive

I’m a Trendy Guy

Moore’s “Law” drove competition and investment

Networking bought in too:

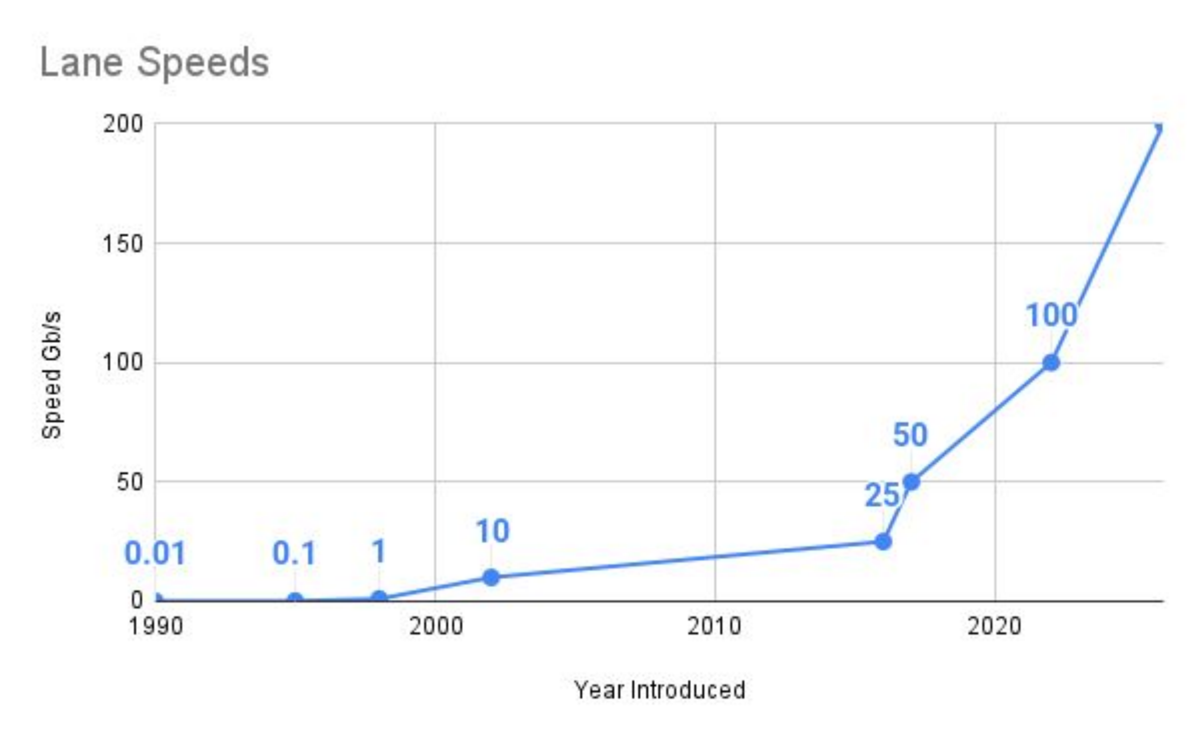
- Faster Speeds
- Fatter Pipes (with Parallel Lanes)
- Faster ASICs

These are the trends

I'm a Trendy Guy

Faster Speeds

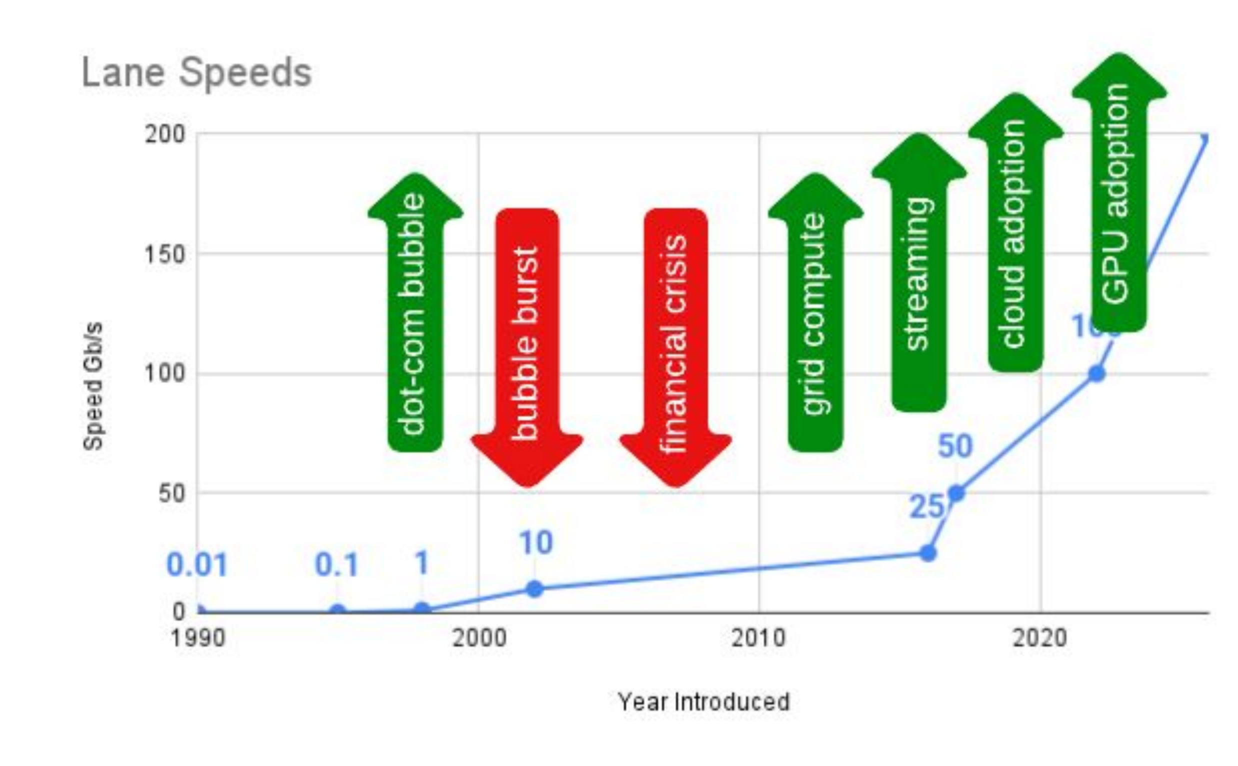
Lane Speed	Intro Date
10Mb/s	1990
100Mb/s	1995
1Gb/s	1998
10Gb/s	2002
25Gb/s	2016
50Gb/s	2017
100Gb/s	2022
200Gb/s	2026?



I'm a Trendy Guy

Faster Speeds

Lane Speed	Intro Date
10Mb/s	1990
100Mb/s	1995
1Gb/s	1998
10Gb/s	2002
25Gb/s	2016
50Gb/s	2017
100Gb/s	2022
200Gb/s	2026?



I’m a Trendy Guy

Fatter Pipes

Lane speeds define host NIC adoption

Multi-lane NICs are pricey

Multi-lane is mainly in network links

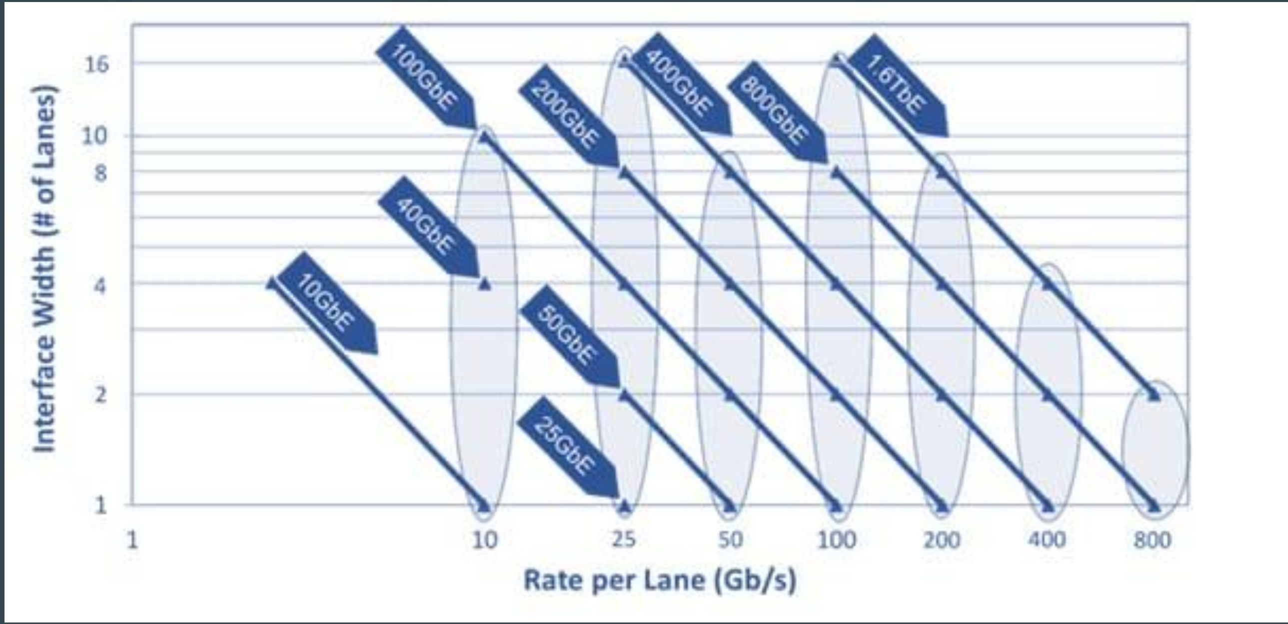


Image credit: IEEE SA

I’m a Trendy Guy

Fatter Pipes

Too many lanes inhibit adoption

4 Lanes is the common pattern

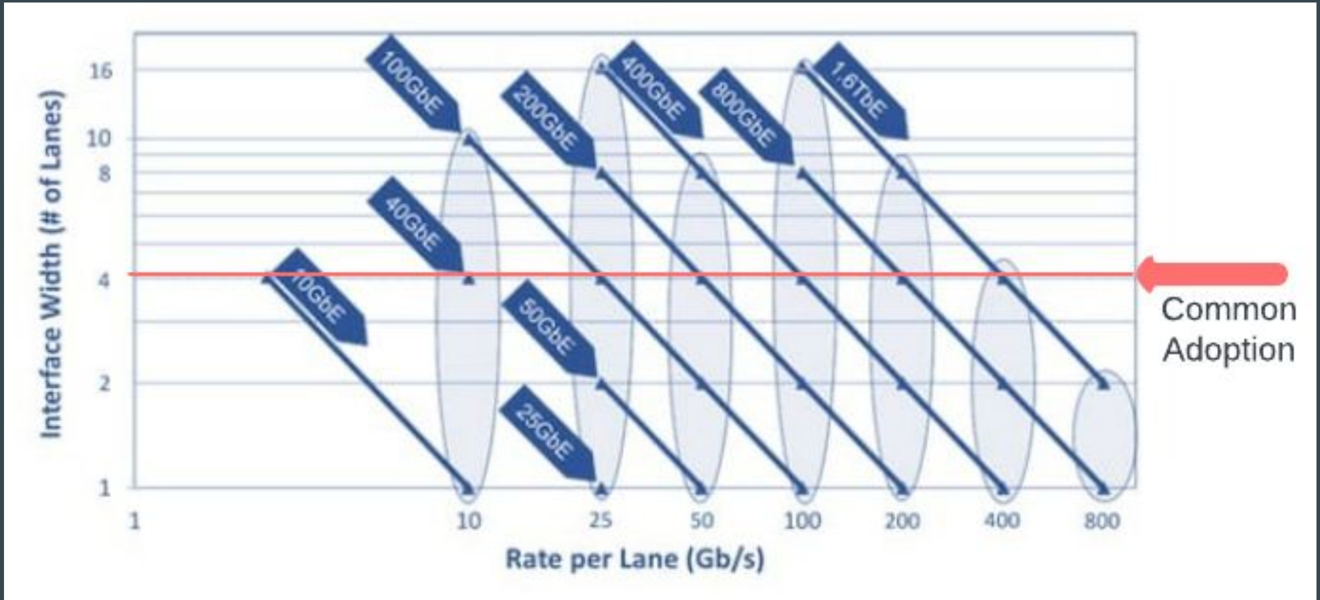


Image credit: IEEE SA

I’m a Trendy Guy

Fatter Pipes

100G NICs readily available

400G network links are common

800G is available, not widely used yet

In 2026, all of that doubles again

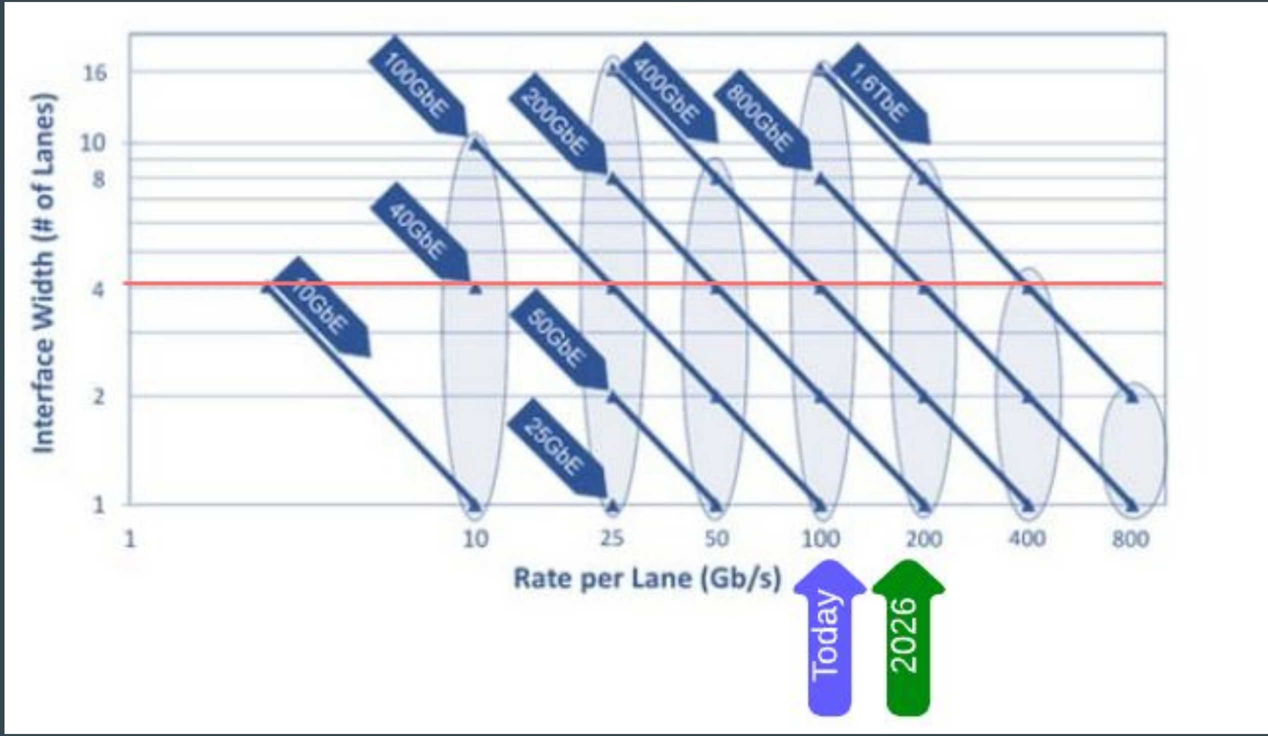
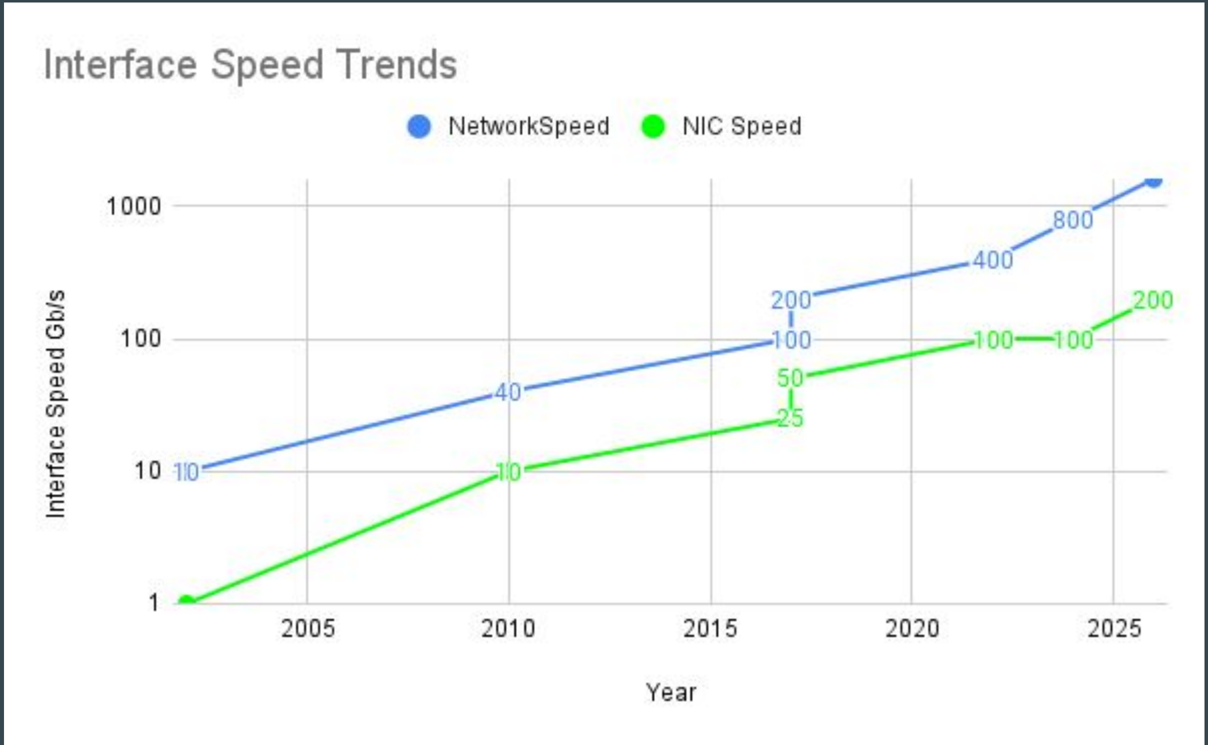


Image credit: IEEE SA

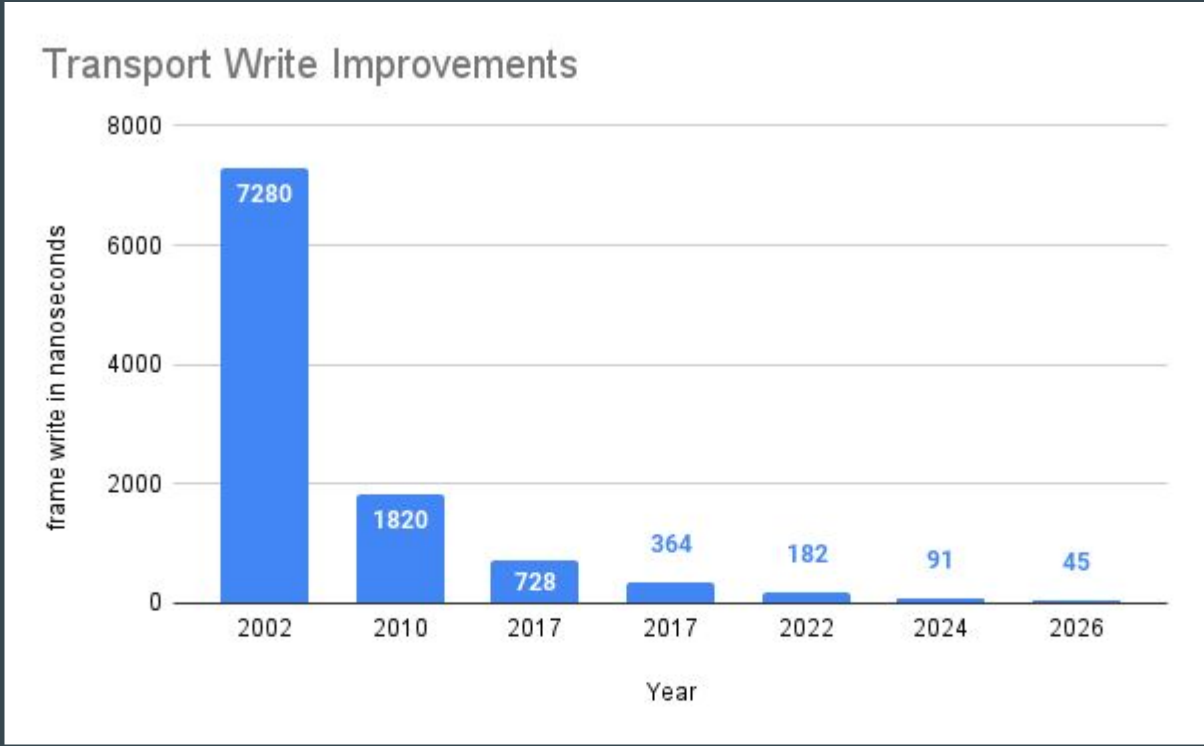
I'm a Trendy Guy

Giving us a clean trend line

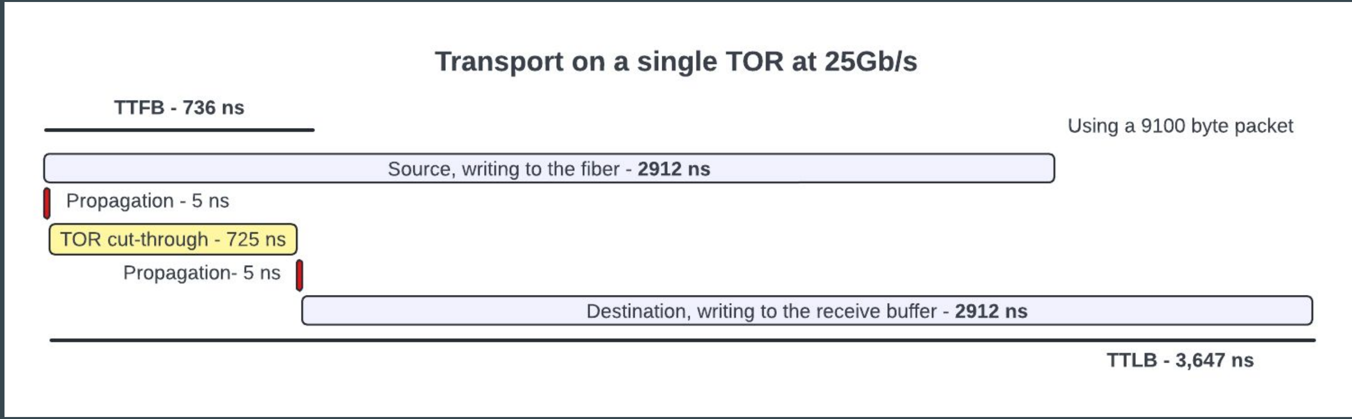


I’m a Trendy Guy

Speed	Date	Time to write 9100 Byte frame	Multiplier from 100Gb
10Mb/s	1990	7,280,000 ns	10000
100Mb/s	1995	728,000 ns	1000
1Gb/s	1998	72,800 ns	100
10Gb/s	2002	7,280 ns	10
40Gb/s	2010	1,820 ns	2.5
25Gb/s	2016	2,910 ns	4
100Gb/s	2017	728 ns	1
200Gb/s	2017	364 ns	0.5
400Gbs	2022	182 ns	0.25
800Gbs	2024	91 ns	0.125
1.6Tb/s	2026?	45 ns	0.0625



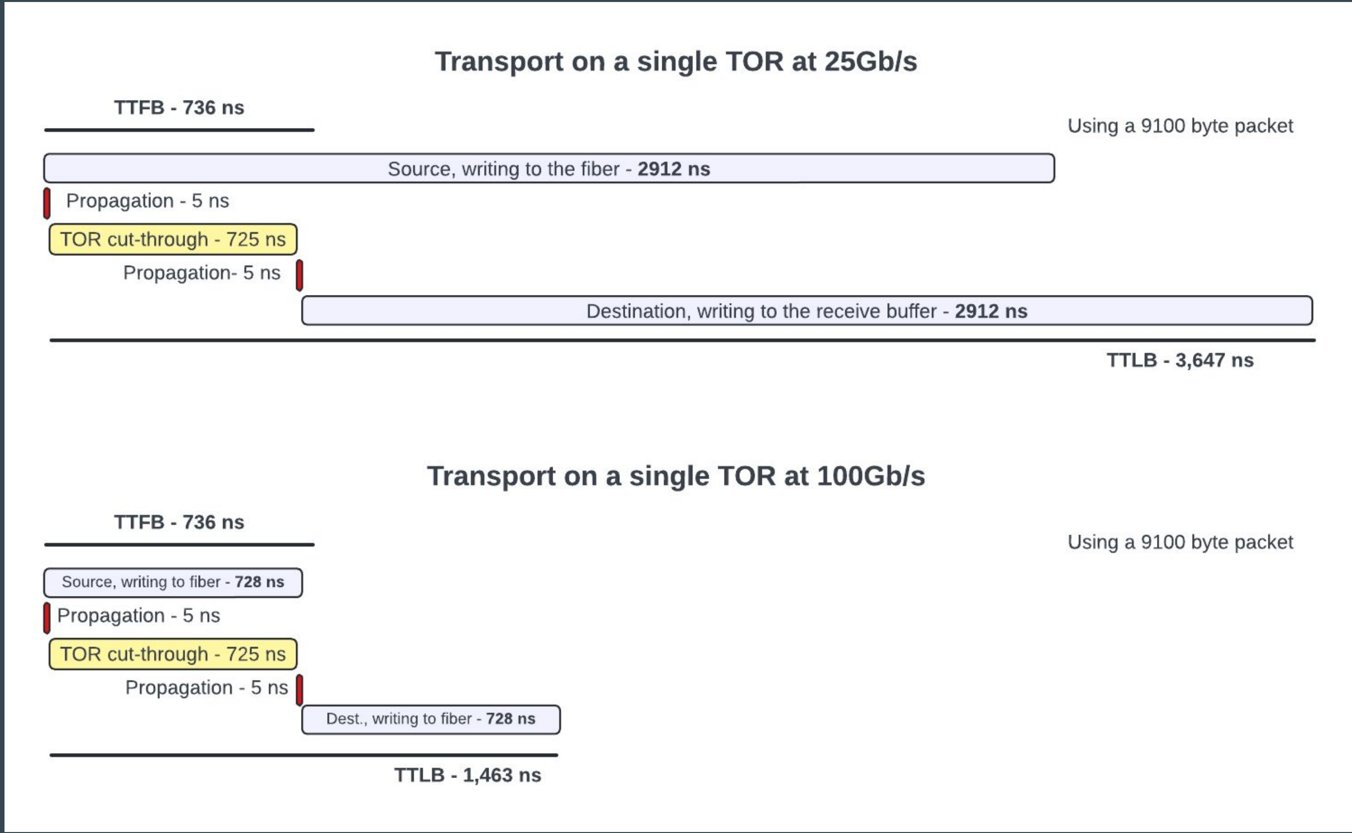
So what does this mean to me?!?



TTFB: 736 ns

TTLB: 3,647 ns

So what does this mean to me?!?



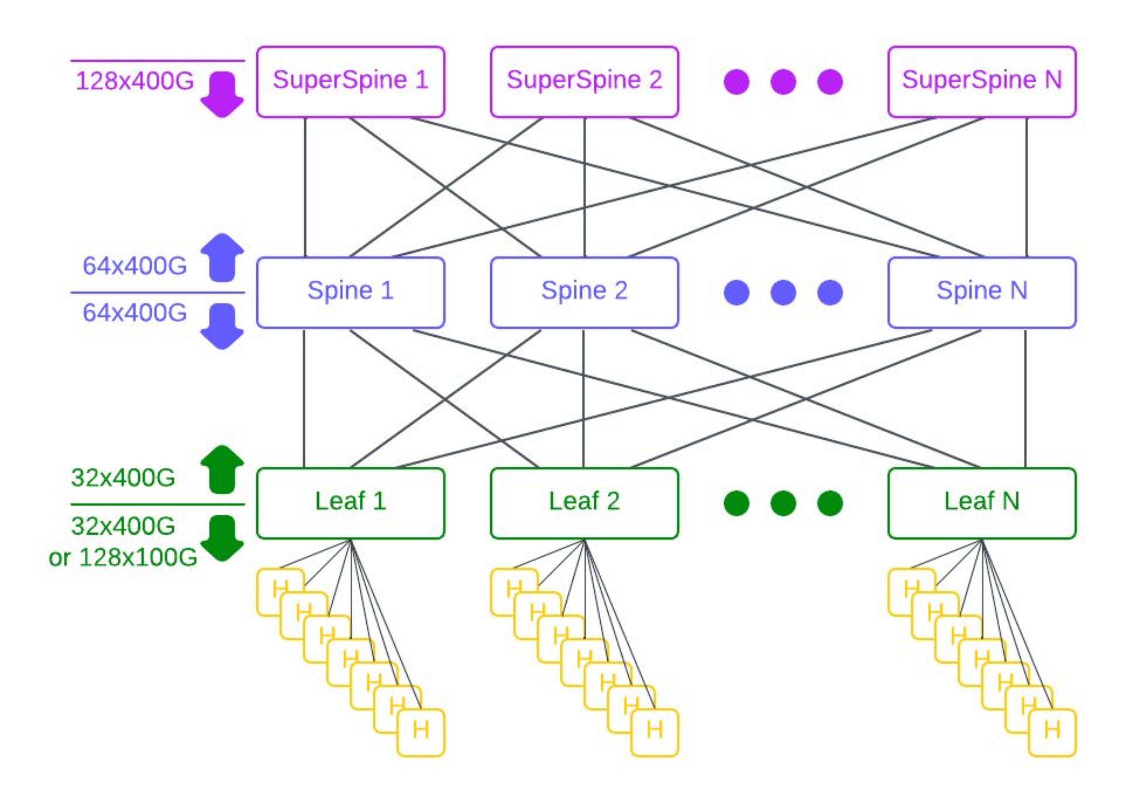
TTFB: 736 ns

TTLB: 3,647 ns

TTFB: 736 ns

TTLB: 1,463 ns

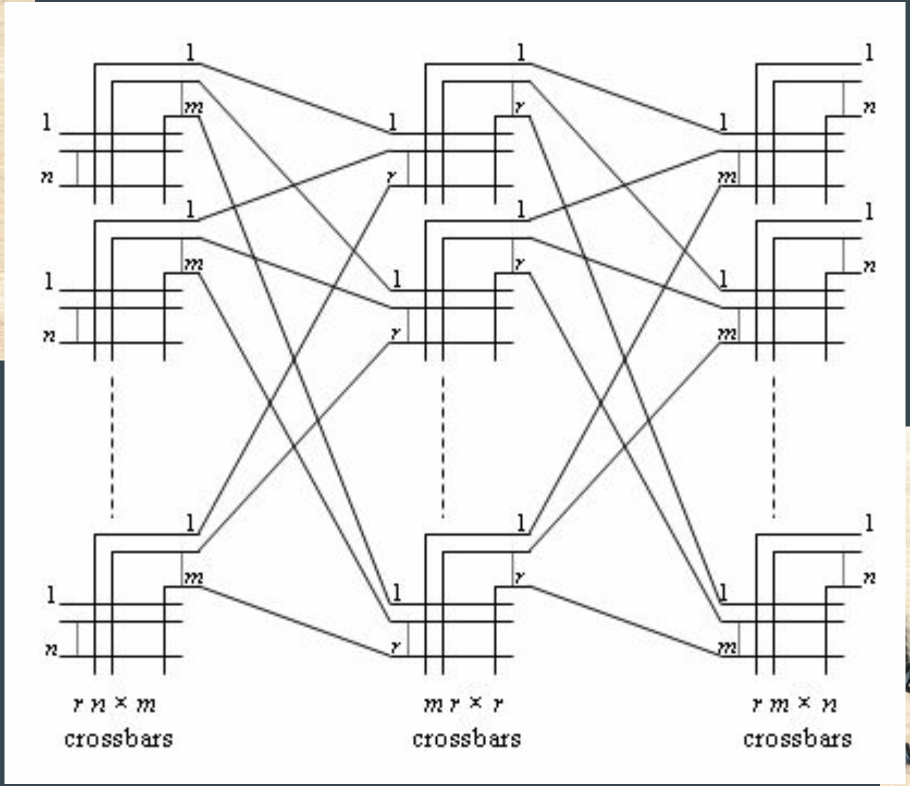
What we *did* do



The Clos Network

- Non-oversubscribed
- Minimized buffers (usec not msec)
- Session hashed path selection
- Fan-in is still an issue
- It only approximates a Clos Network

All that is old is new again



The secret was deploying enough capacity in the fabric to carry all circuits

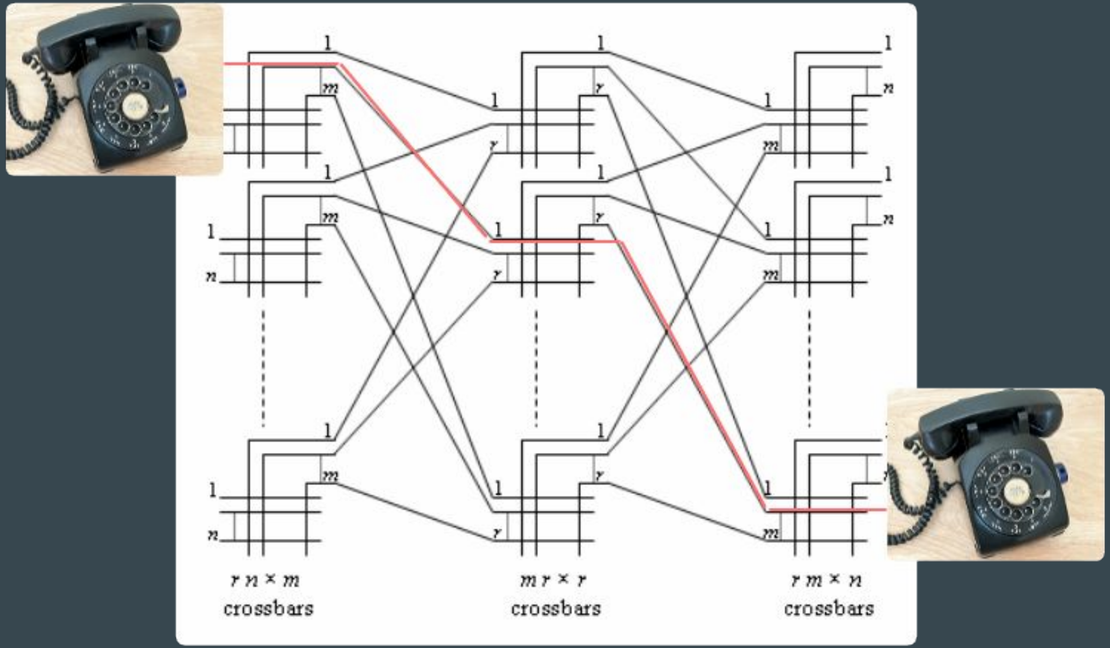
In 1953, the Clos Network was designed for phone circuit switching

All that is old is new again

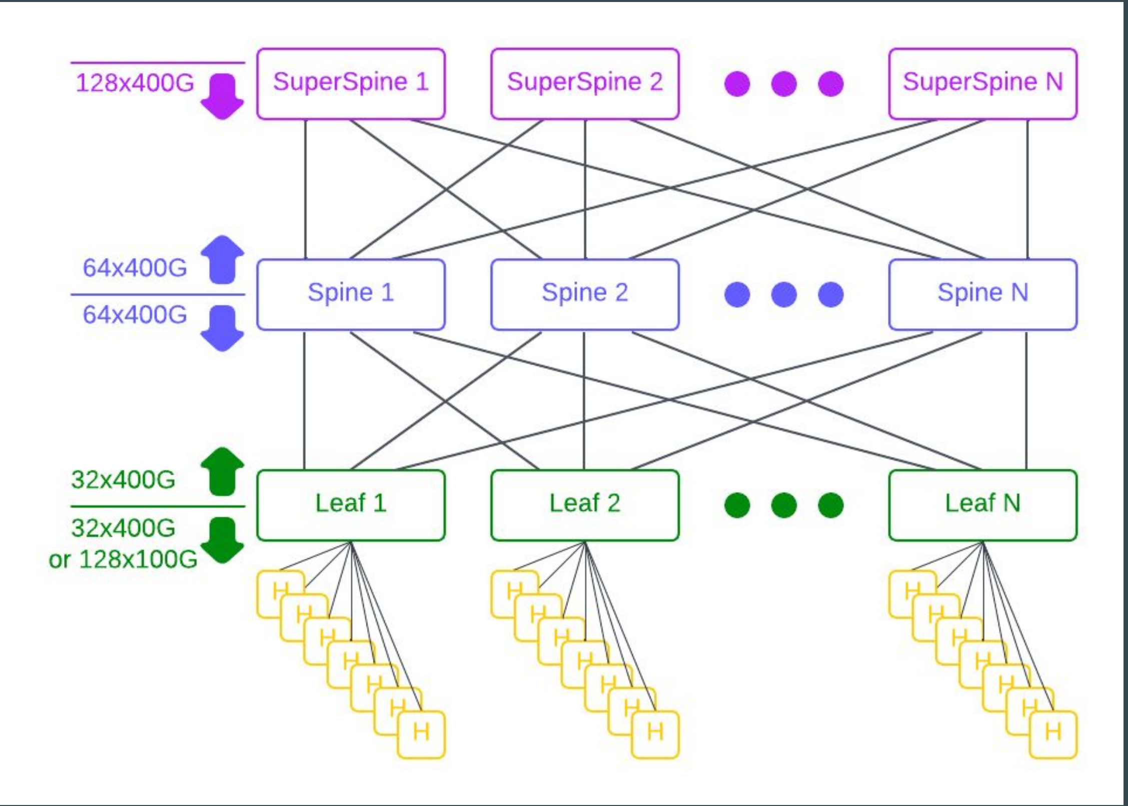
No buffering, a circuit was available or engaged

It was able to guarantee behavior

This is what enabled the phone system to offer 99.999%



What we *did* do

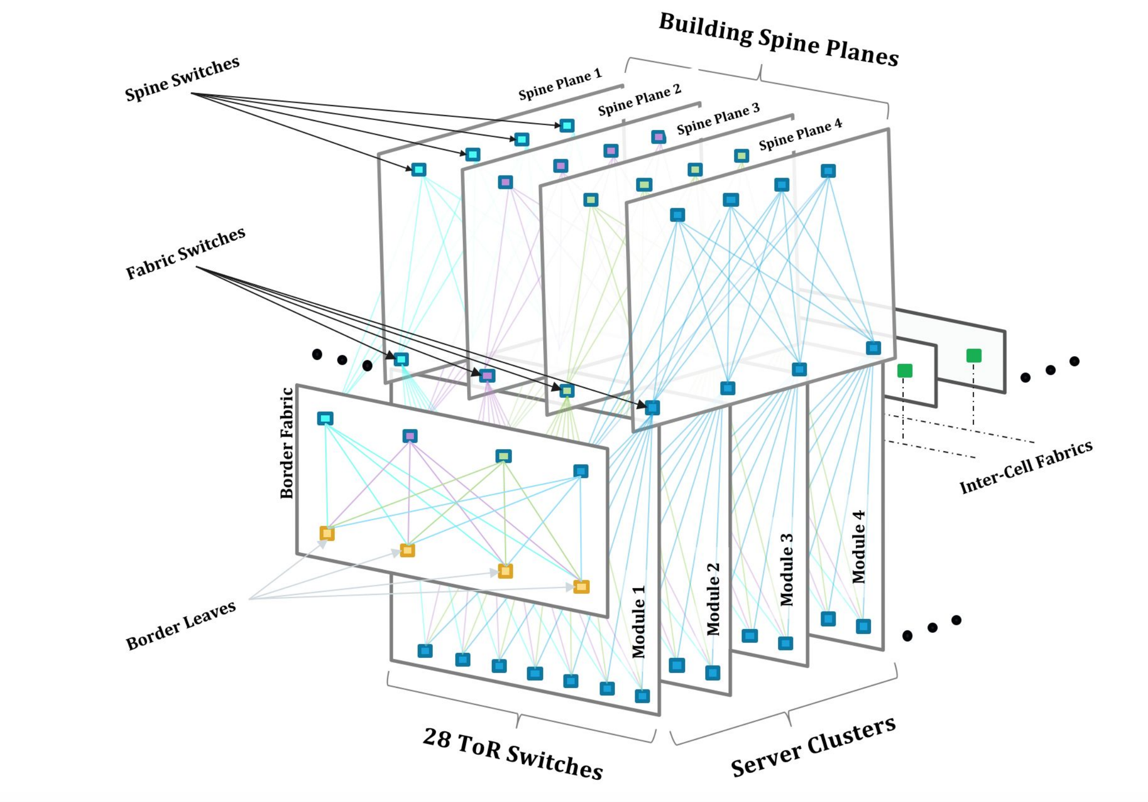


The Difference

- Non-oversubscribed
- Not circuit switched
- Giving Probabalistic Guarantees
- Fan-in is still an issue
- Contention and buffering is seen, but infrequent “enough”

What we *did* do

- We maximize fabric size
- A multi-planar network can easily be 80,000+ hosts and 30+MW of power
- All non-oversubscribed
- Size defined by ASIC Radix



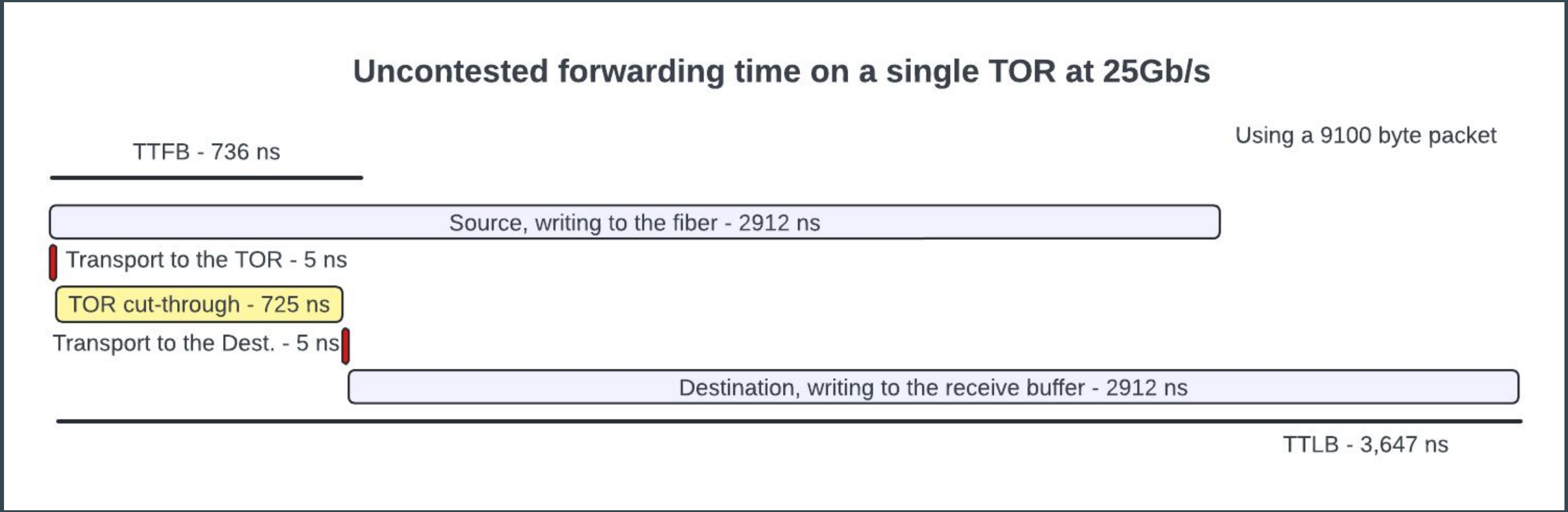
I’m still a Trendy Guy

- Radix keeps up with lane speeds
- It doubles radix as fab improves
- Lanes combined for interfaces
- The Radix defines how large a Clos fabric can be built

Chipset	Throughput	Radix	Date Launched
Trident	640G	64x10G	2010
Trident2	1.28T	128x10G	2012
Tomahawk	3.2T	128x25G	2014
Tomahawk2	6.4T	64x100G	2016
Tomahawk3	12.8T	128x100G	2017
Tomahawk4	25.6T	256x100G	2019
Tomahawk5	51.2T	512x100G	2023
Jericho3	28.8T	288x100G	2025?
Ramon3	51.2T	512x100G	2025?

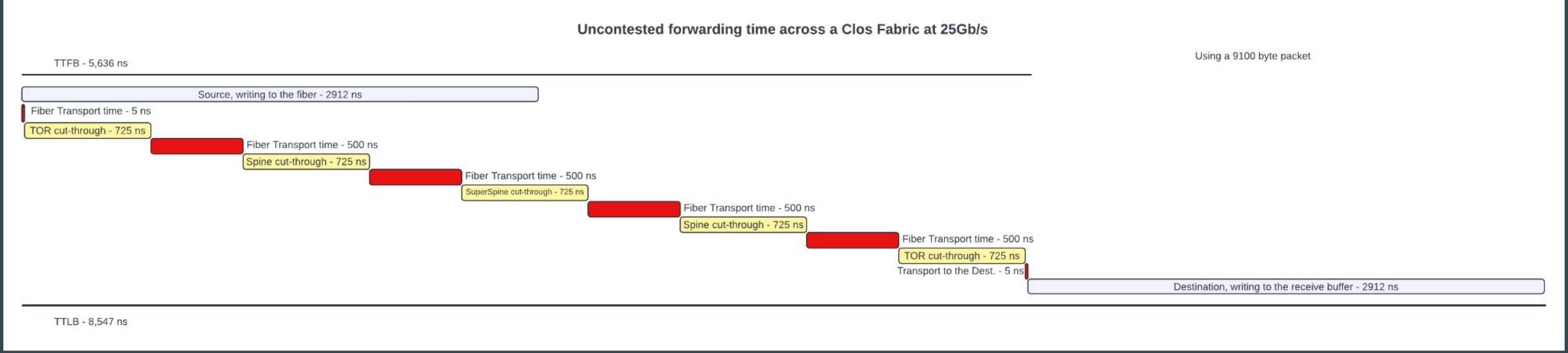
Moving YOUR tail

- Now, think back to the waterfall timing in a single TOR



Moving YOUR tail

- This eye chart is showing the same packet crossing a large Clos network
- Important bits: TTFB - 5.64 usec, TTLB - 8.55 usec



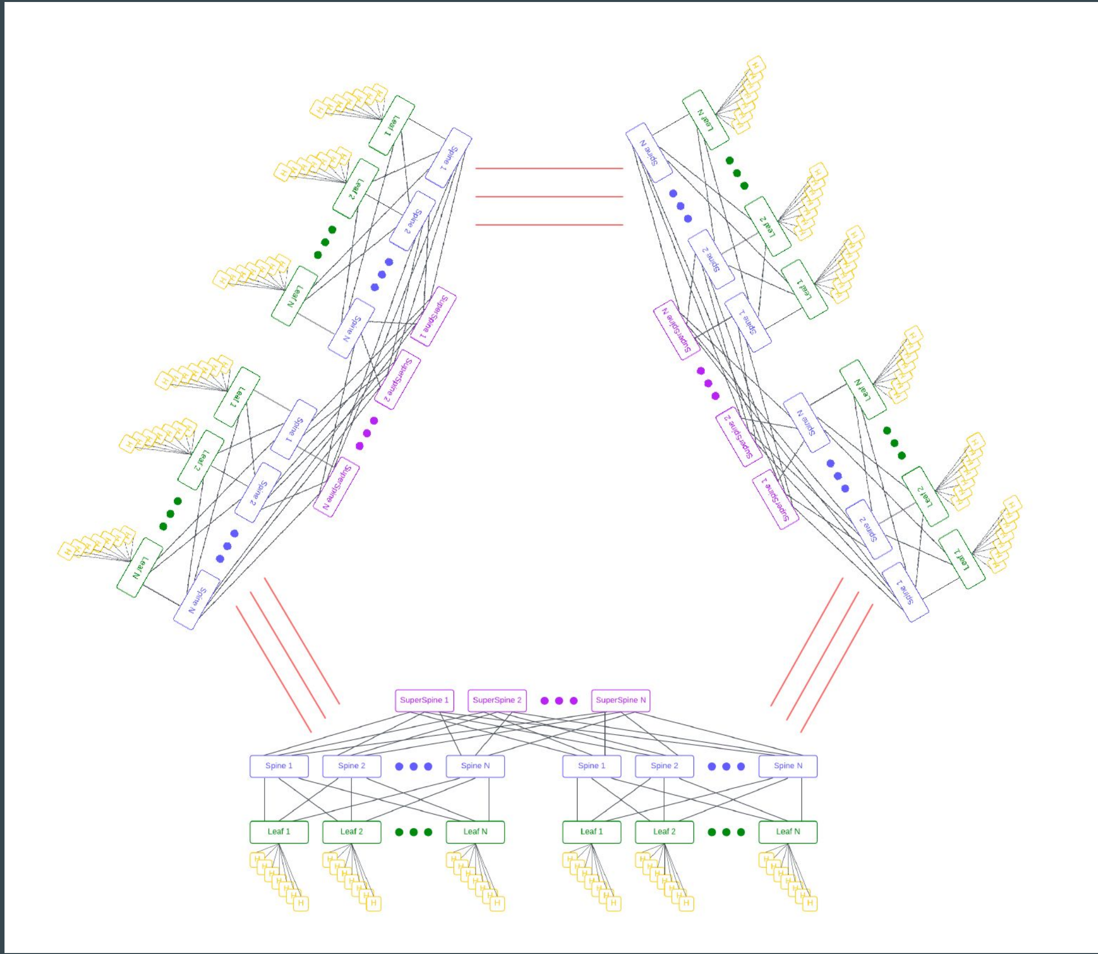
Moving YOUR tail

- You're still sharing the Network
  - ASICs hate buffering too
- Switches have ~53 usec of buffer
  - Worst case across a Rack - 53 usec
  - Worst case across a Clos - 274 usec

- Absolute worst case

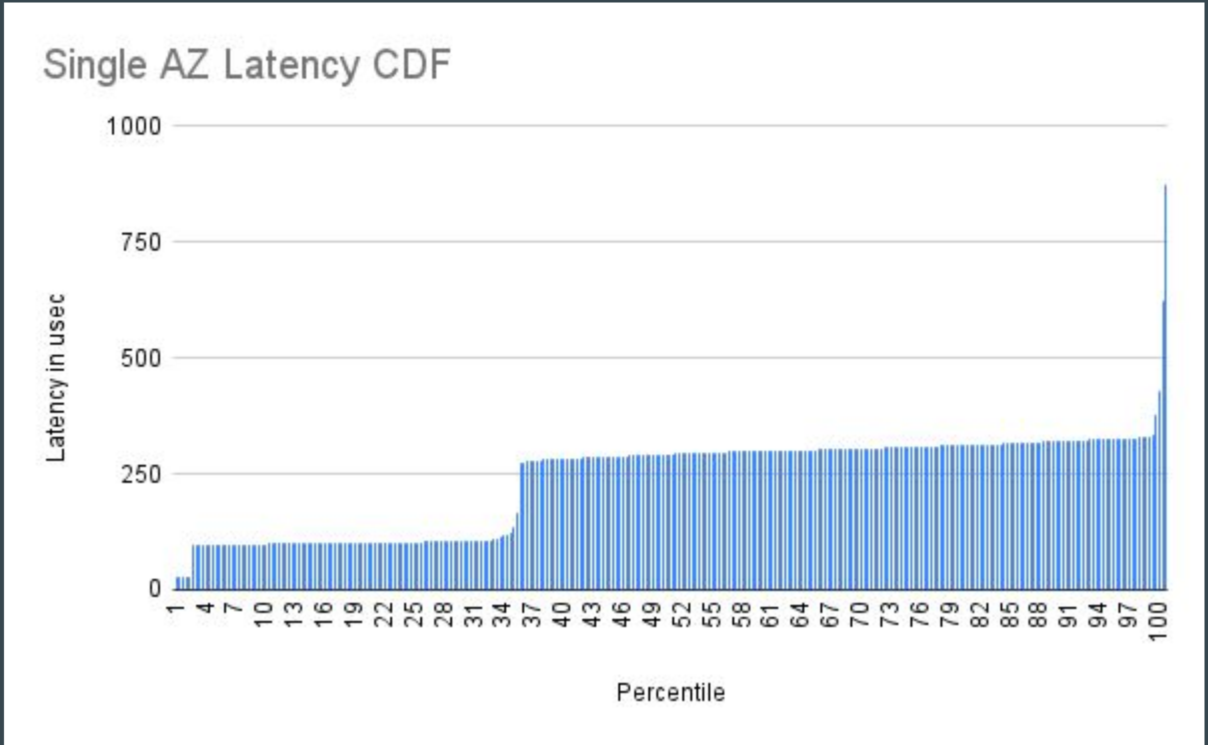
A Note about Public Clouds

- An AZ is one or more Clos networks
- Inter-Clos capacity is demand-signaled for augmentation



A Note about Public Clouds

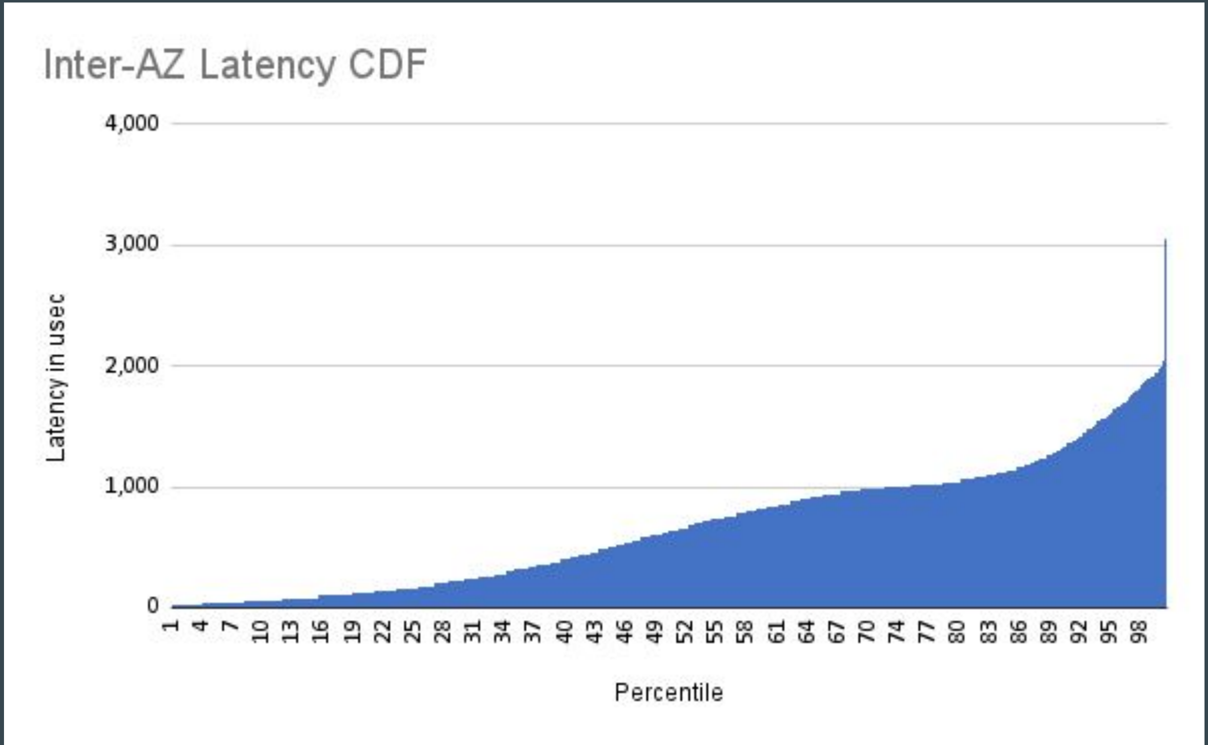
- We infer intimacy as categories of latency
- In this test we observed
    - Same rack ~30 usec
    - Single clos ~100 usec
    - Inter-clos ~300 usec
- Through a Hypervisor and SDN



A Note about Public Clouds

- Crossing AZs is much less intimate

- It also has far less consistency



What we are doing about it now

In the last month:

- Azure announced they're building 24k GPU fabrics (Network@Scale'24)
- Meta announced their 100k GPU fabrics (Network@Scale'24)
- Alibaba published their GPU Network Design (SIGCOMM'24)

All focusing on fabric 'intimacy' and lack of contention

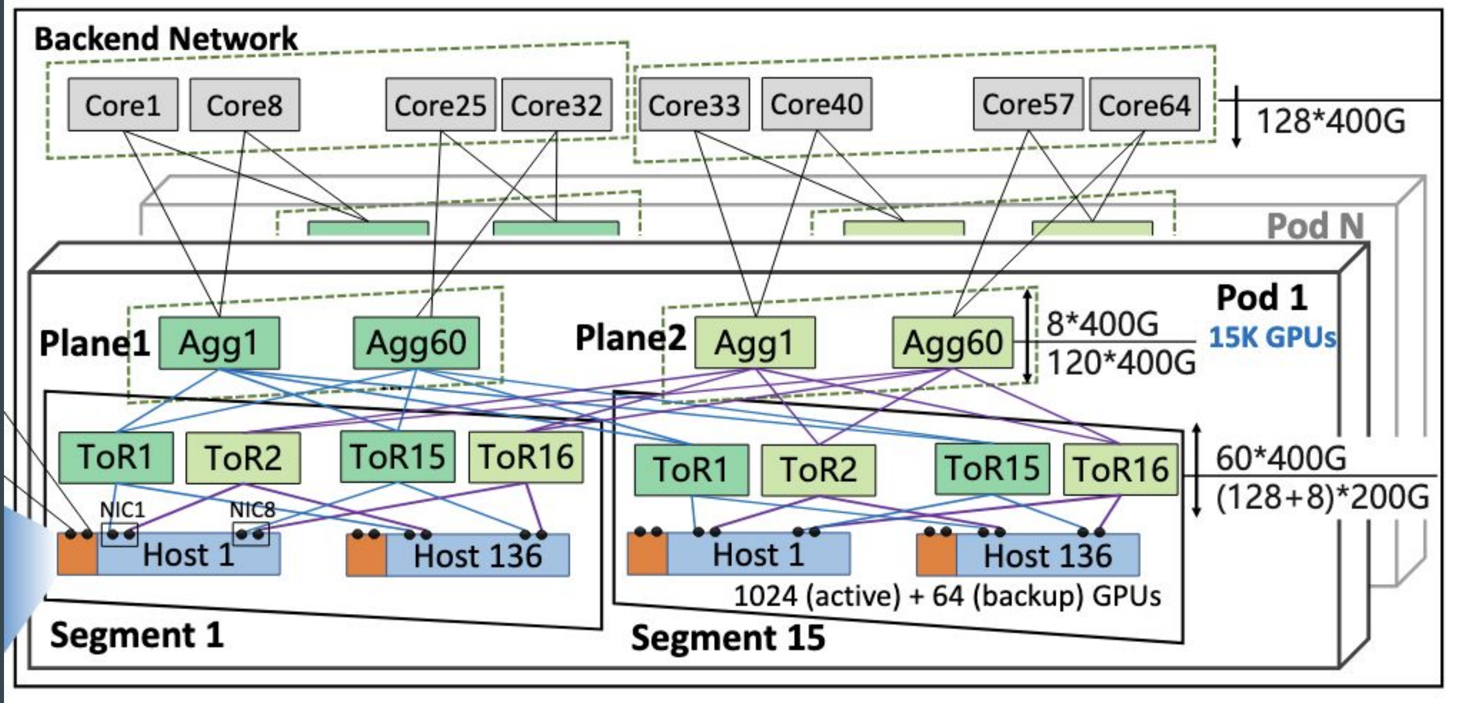
We solve the problem with the tool we have - Throwing bandwidth at it.

What are we doing about GPUs now?

In a dedicated GPU Network

1,024 GPUs per ‘segment’

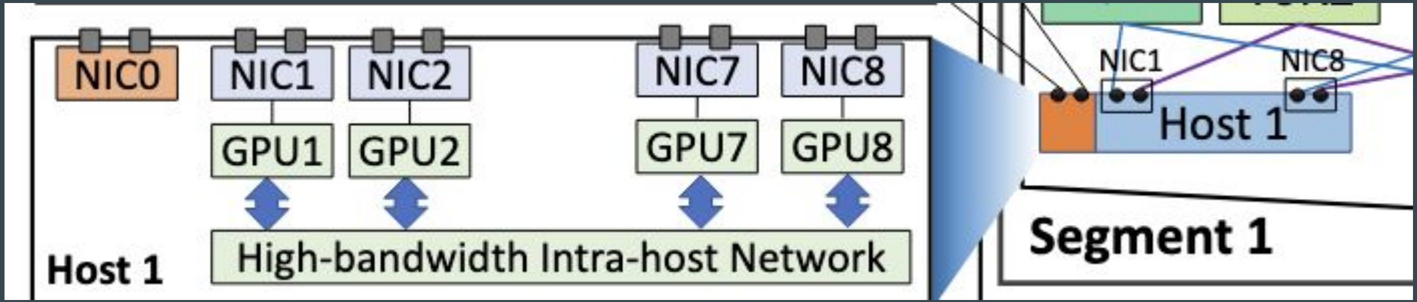
15k GPUs per ‘pod’



Credit: [Kun Qian, et. al. SIGGCOMM’24]()

What are we doing about GPUs now?

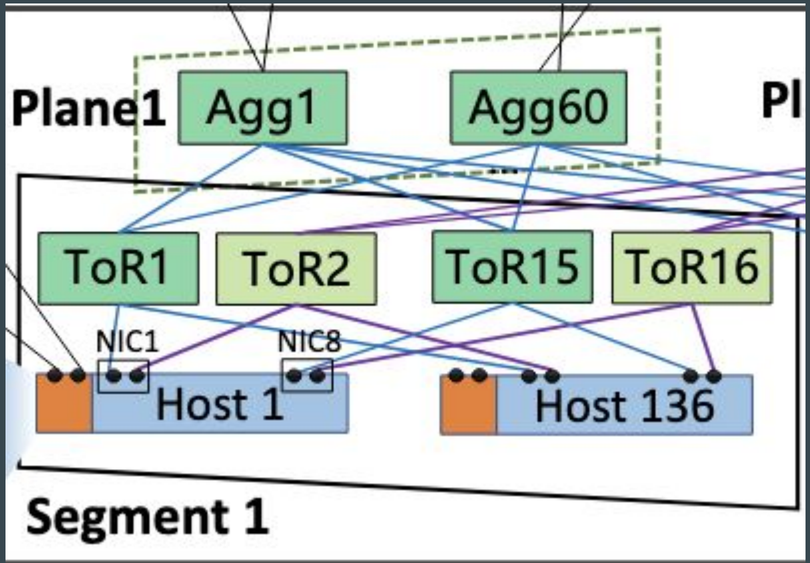
- Each GPU gets its own 400G NIC
- Split to 2x200G interfaces to different switches
- 8:1 outnumbering the front-end interfaces.



Credit: Kun Qian, et. al. SIGGCOMM'24

What are we doing about GPUs now?

- Striping across many TORs increases paths
- Without session entropy, we increase the interfaces to avoid conflicts



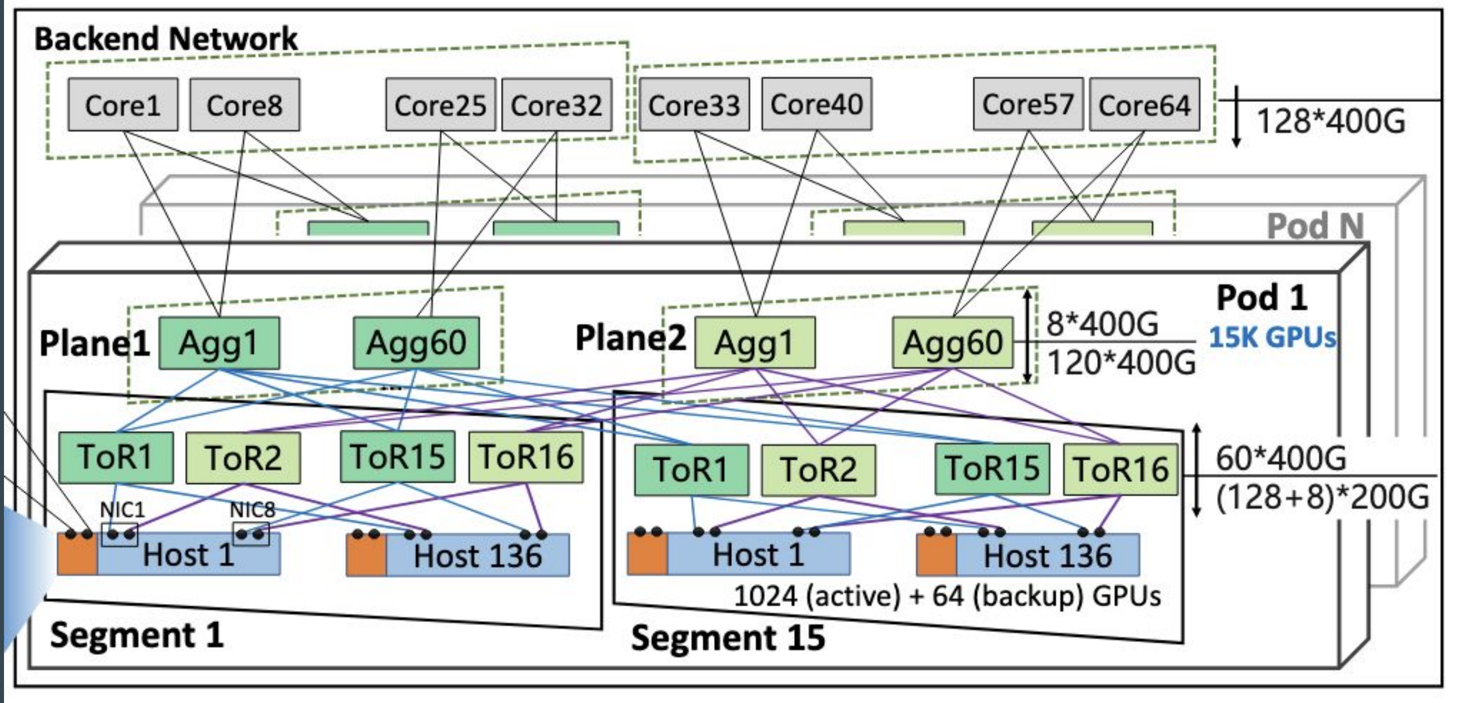
Credit: [Kun Qian, et. al. SIGGCOMM'24]

What are we doing about GPUs now?

Within a segment, no oversubscription
- 410 Tb per segment

Between segments, slight oversub. (16:15)
- 5.76 Pb per pod

Between pods, heavy oversub. (8:1)

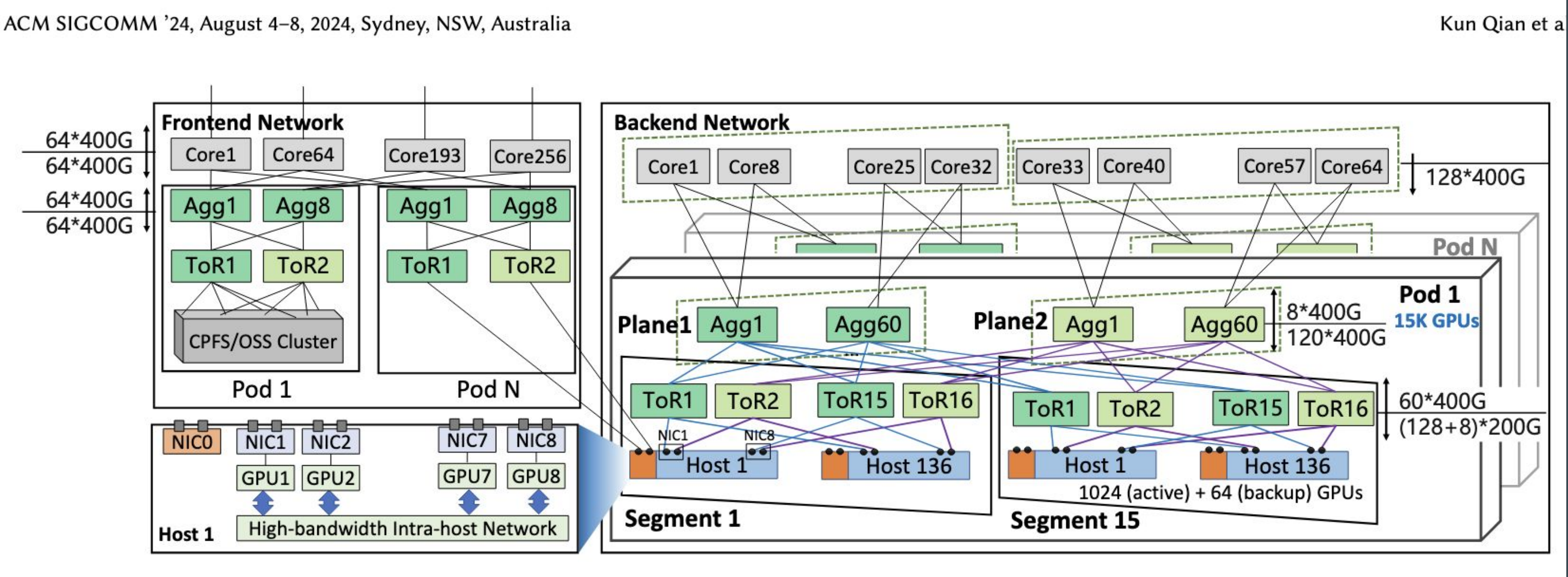


Credit: Kun Qian, et. al. SIGGCOMM'24

Credit: Kun Qian, et. al. SIGGCOMM'24

What are we doing about GPUs now?

Putting it all together, the Clos we'd normally be talking about is a footnote at the top right.



What will we do about it tomorrow?

Hardware Design changes - E.g. Jericho3/Ramon3 from Broadcom

- ‘AI focused’ ASICs still suffer from the ‘elephant flow’ problem
- Newer chips will determine path in hardware and reserve bandwidth
  - Giving predictable latency and order
  - Using Time-Division Multiplexing (TDM)
- Not a new idea, but it is new to the LAN
  - Circuit Reservation was in mechanical phone switches, ATM-SMDS, and MPLS-RSVP.

What will we do about it tomorrow?

Standards changes - E.g. Ultra Ethernet Consortium

- UEC has a number of ideas, but the exciting one is a return to packet-based load balancing
  - Session-based load balancing has dominated for >20 years - to preserve packet order
  - Reassemble before the kernel to preserve packet order
  - This will rethink the concept of ‘flows’ and make RoCE smoke!
- Packet spraying across all paths, combined with forward error correction, and re-assembly prior to the destination application is a game changer.

Where this will put us… 5 Years from Now

- We will squander the abundant (bandwidth) to save the precious (latency)

- Critical communications will make reservations across the fabric
  - Return to circuit switching
  - Locking in a latency guarantee and minimizing jitter

- Other sessions will spray packets across all paths to scavenge the remaining bandwidth

- …and speeds and feeds will be faster too…

Takeaways

- Networks got good at making huge non-oversubscribed fabrics
  - But these were probabilistic guarantees
  - Crossing fabrics is still very oversubscribed

Takeaways

- Networks got good at making huge non-oversubscribed fabrics
  - But these were probabilistic guarantees
  - Crossing fabrics is still very oversubscribed
- GPUs shifted how we think about Networks
  - Latency and Jitter are in the spotlight to keep GPUs fed
  - Intimacy between systems is critical
  - All latency dependent systems will benefit

Takeaways

- Networks got good at making huge non-oversubscribed fabrics
  - But these were probabilistic guarantees
  - Crossing fabrics is still very oversubscribed
- GPUs shifted how we think about Networks
  - Latency and Jitter are now in the spotlight to keep GPUs fed
  - Intimacy between systems is critical
  - All latency dependent systems will benefit
- Innovation will squander the abundant (bandwidth) to preserve the precious (latency)
  - Critical Communications will make path and bandwidth reservations to guarantee latency and jitter
  - Other sessions will spray packets across all paths to scavenge the remaining bandwidth

Want to talk networks?

Reach out: dlucey@salesforce.com