Rethinking networks in finance

Dr David Snowdon
Director, Engineering

ARISTA

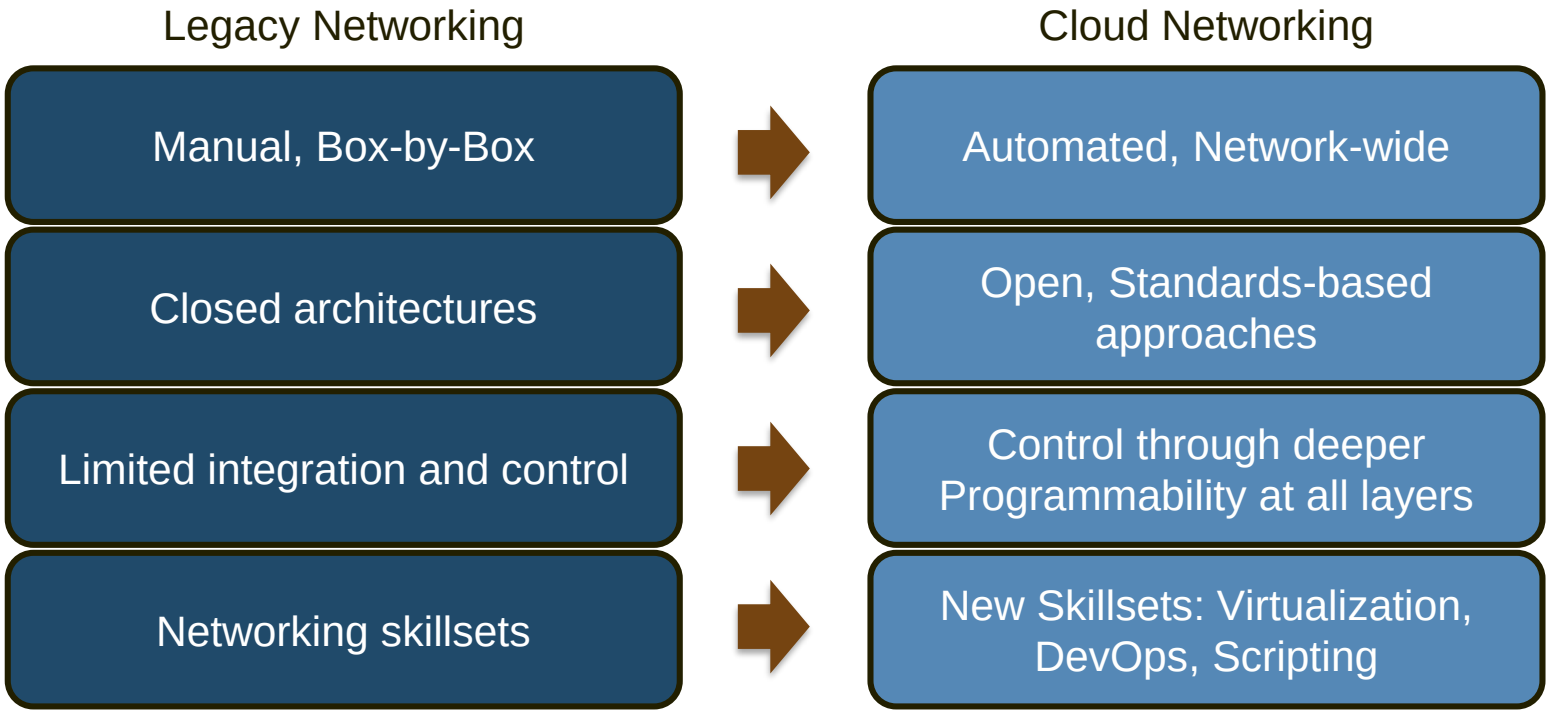
Arista Acquire Metamako

September, 2018

- 5 Countries
- 57 Employees
- Most low latency firms

- Three things:
  - Low Latency
  - Network Visibility (Tap/Agg)
  - FPGA Apps

The Transition to Cloud Networking



3

Metamako is now Arista

- Metamako HQ is now Arista's Sydney office
- Engineering team remains, dedicated to financial services
- Operational integration underway (finance, marketing, manufacturing, sales)
- Scale is good.

- Going forward…
  - Consolidating on EOS for next-gen platforms
  - Releasing new products as per the existing roadmap
  - Looking for input from everyone here

- Introducing the Arista 7130 series

What next?

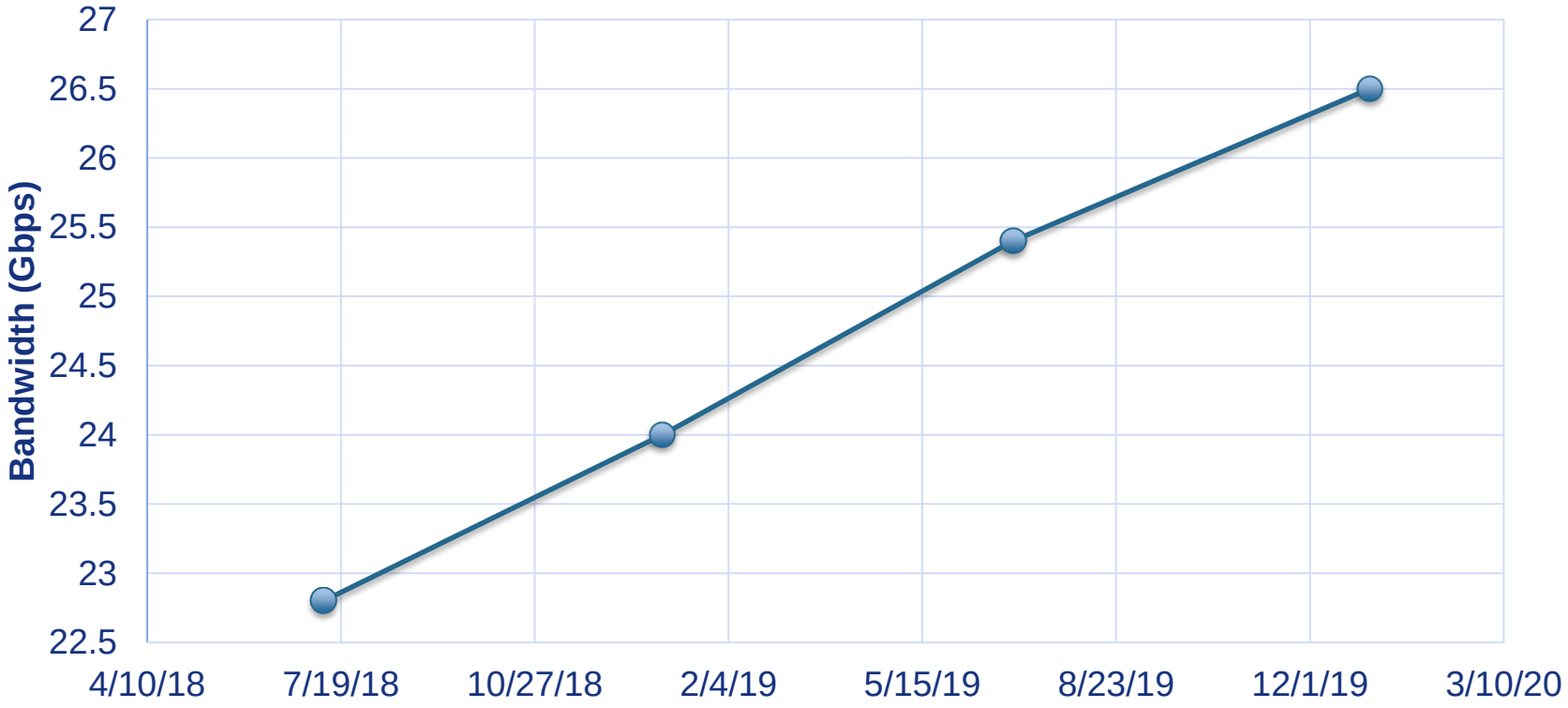
Financial Services Networks are strange…

- We don’t care much about average bandwidth, but we care a lot about latency
- Gigantic bursts, but low average bandwidth
- We often care about consistency and determinism more than we care about peak average performance…

Bandwidth

- Much of the push in the data plane is toward higher bandwidth
- Gigantic and growing East-West traffic in cloud applications drivers higher bandwidth for intra-DC links
  - 10G -> 40G -> 25G -> 100G -> 200G -> 400G -> 800G
  - 1.25G NRZ -> 10.3G NRZ -> 25.8G NRZ -> 25.6G PAM4 -> 53.1G PAM4
- Evolving optical module standards:
  - SFP -> SFP+ -> QSFP+ -> SFP28 -> QSFP28 -> QSFP-DD -> OSFP
- Most up-to-date financial services deployments use 10G over SFP+
  - Bandwidth has not been a driver
- Some experiments with 40G with varied results

Opra Projections



Source: https://www.opradata.com

Bandwidth – The tradeoffs

- Higher rate can mean lower serialization delay
- Higher rate can mean less queueing
- And it can also mean lower latency in the surrounding logic
  - 1G -> 10G resulted in much lower latency equipment
- But it can result in more complex logic
  - Complex logic generally means higher latency
  - Multi-lane protocols have some extra overhead inherent to the striping/reassembly
  - Unless they're used as lower-rate links
  - Like: 40G with 4 lanes
- Higher signaling rates are less reliable – require error correction
  - Like: 25G with FEC enabled (about 80ns or 250ns extra latency)

There are some complexities as we move to higher bandwidth connections

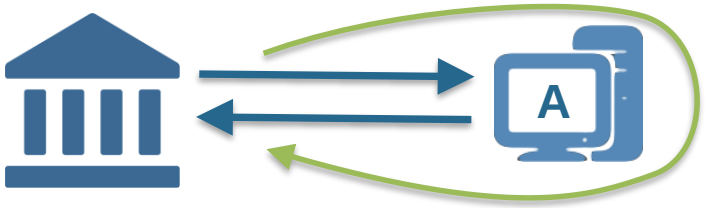
Bandwidth conclusions

- So we tend to care about bandwidth, but only when it affects our:
  - Latency
  - Determinism
  - Reliability

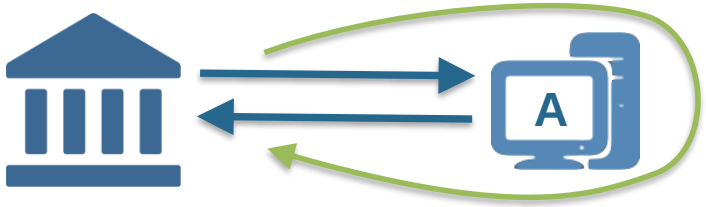
- Higher bandwidth links are in our future.

Latency

- Traders (you) care a lot about latency, but no-one (including me) is sure what those figures are.
- A key indicative latency metric is the tick-to-trade latency – how long does it take to send an order after a market event:



Latency



Some recent public reference points. Difficult to compare due to varying functionality:

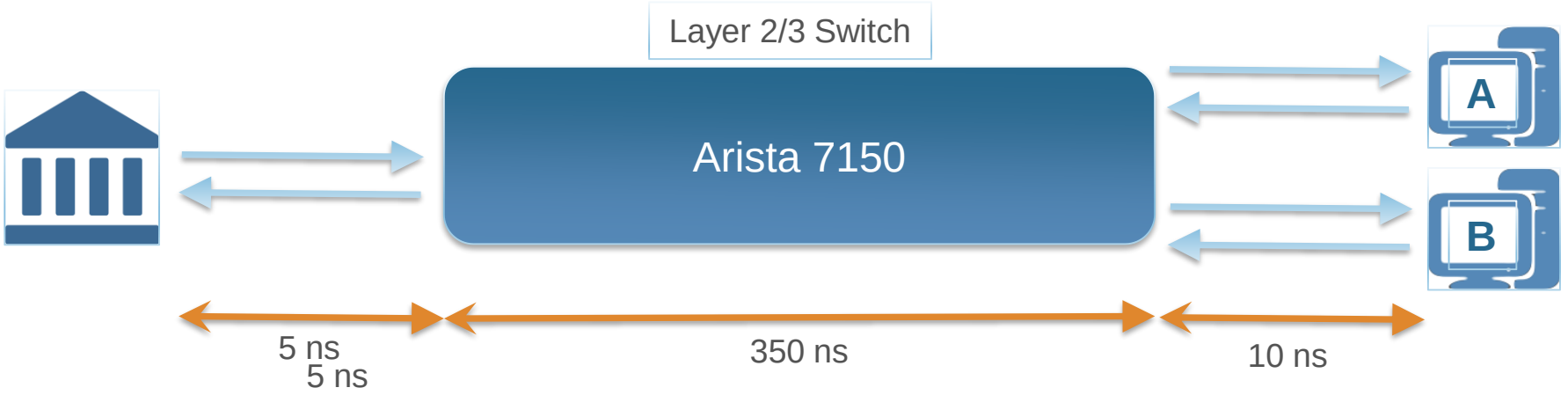
- FPGA vendor X web site: “Sub microsecond”
- LDA/Solarflare/Penguin: 98 nanosecond tick-to-trade measured using STAC-T0
- Arista: 45 ns multiplexing

NOT STAC BENCHMARK

Latency

- Most market participants want to share their expensive connections
- Traditional networking means a Layer2/Layer3 switch to:
  - Route packets
  - Create peering sessions

Circa 2012



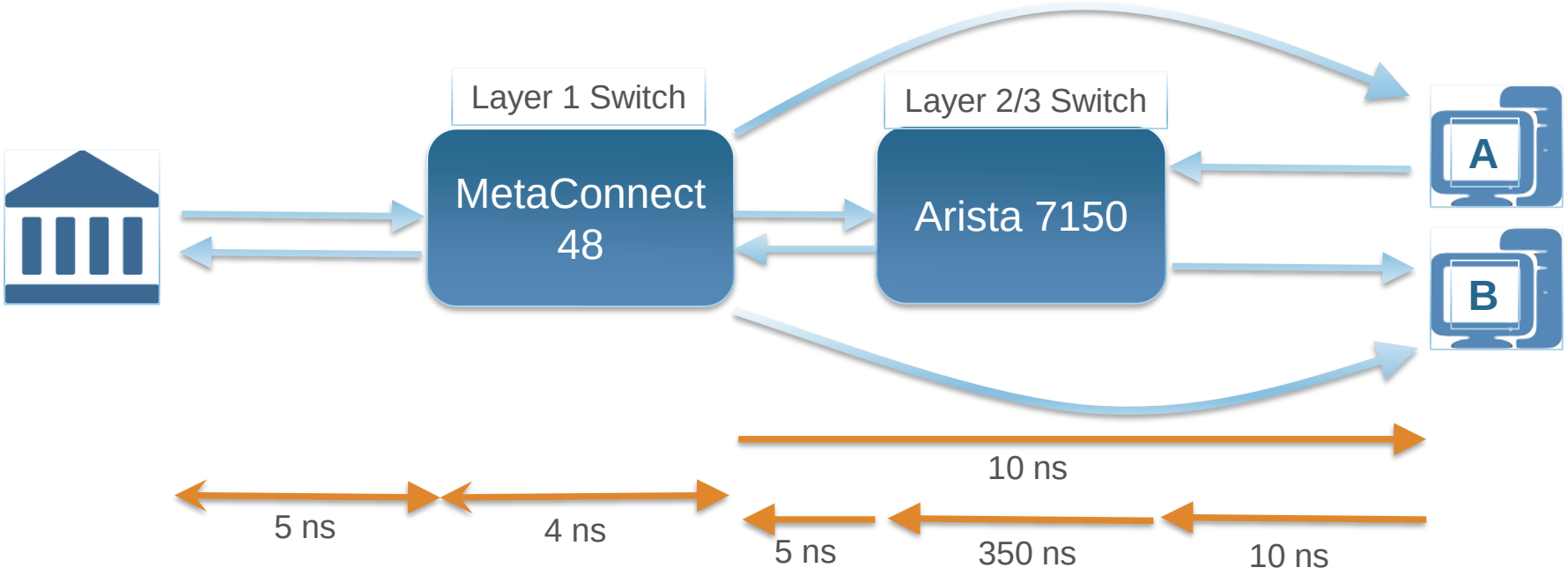
Total: 720 ns

NOT STAC BENCHMARK

Latency

- Layer 1 switches gave dramatically lower market data fan out

Circa 2014

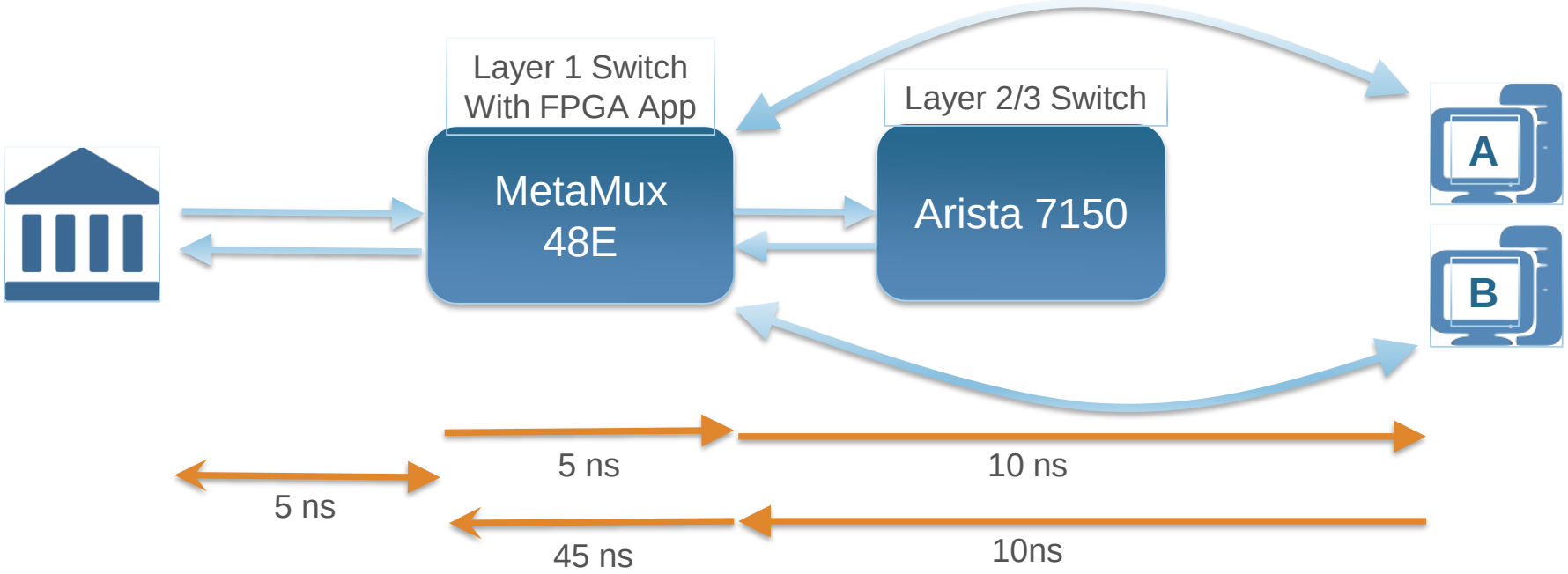


Total: 393 ns

Latency

- FPGA aggregation sped up fan-in

Circa 2016



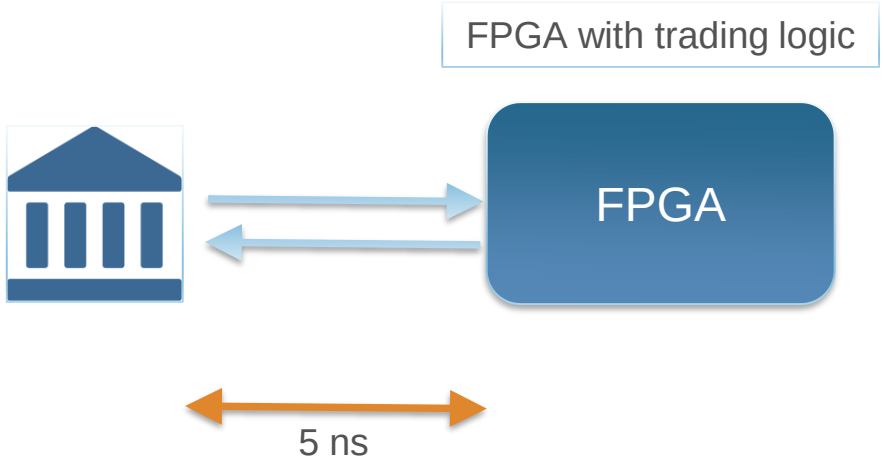
Total: 75 ns

NOT STAC
BENCHMARK

Latency

- Directly connected FPGA
- Lacking network visibility, counters, manageability

Circa 2017



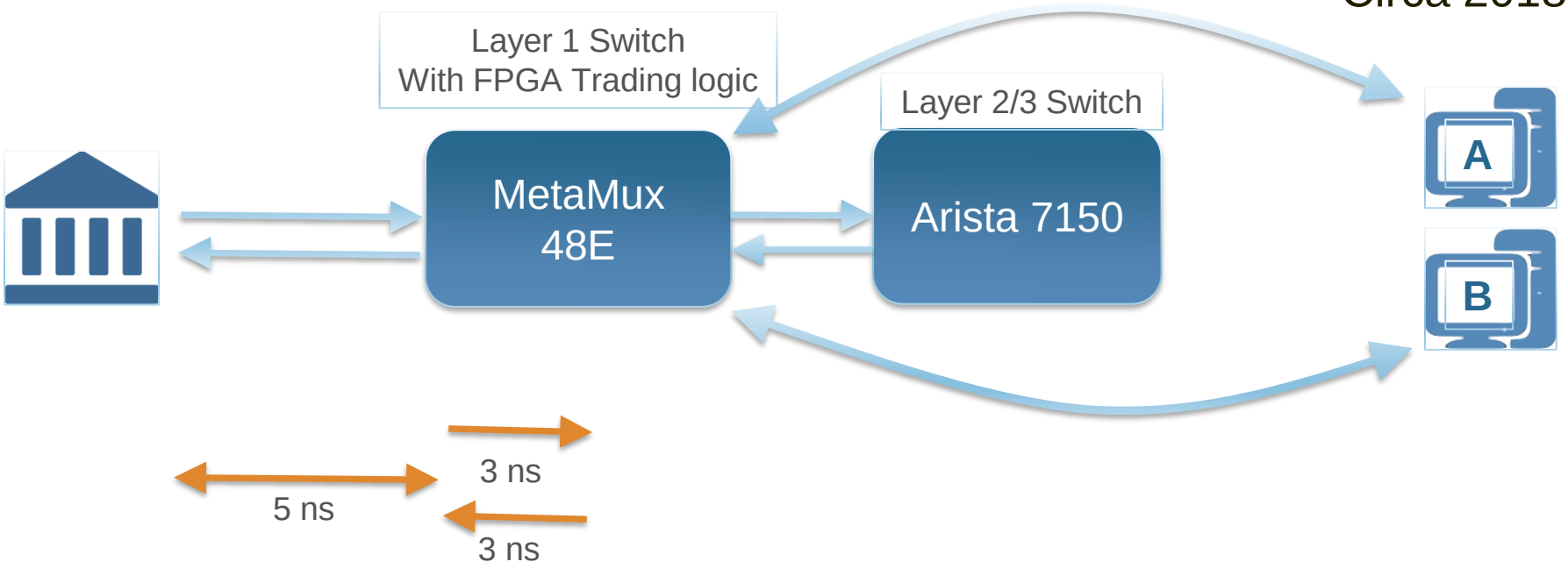
Total: 10 ns

NOT STAC BENCHMARK

Latency

- So we enabled FPGA trading applications that reside in the switch

Circa 2018



Total: 16 ns

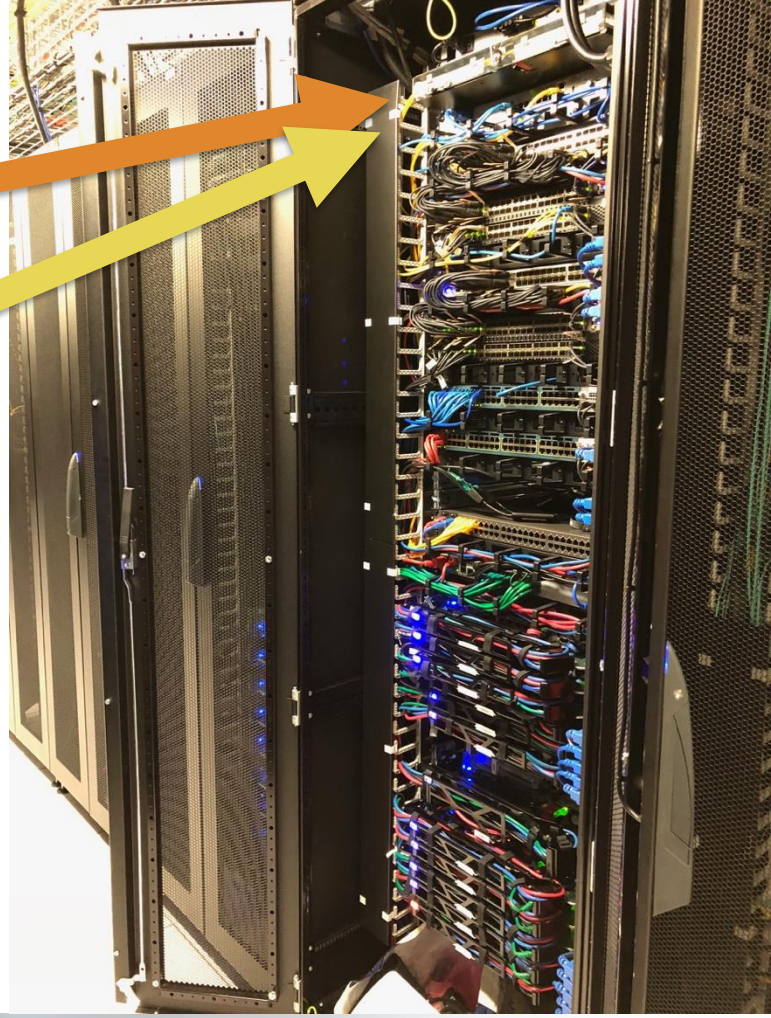
NOT STAC
BENCHMARK

Latency

Exchange fibers

Best location for algo FPGAs

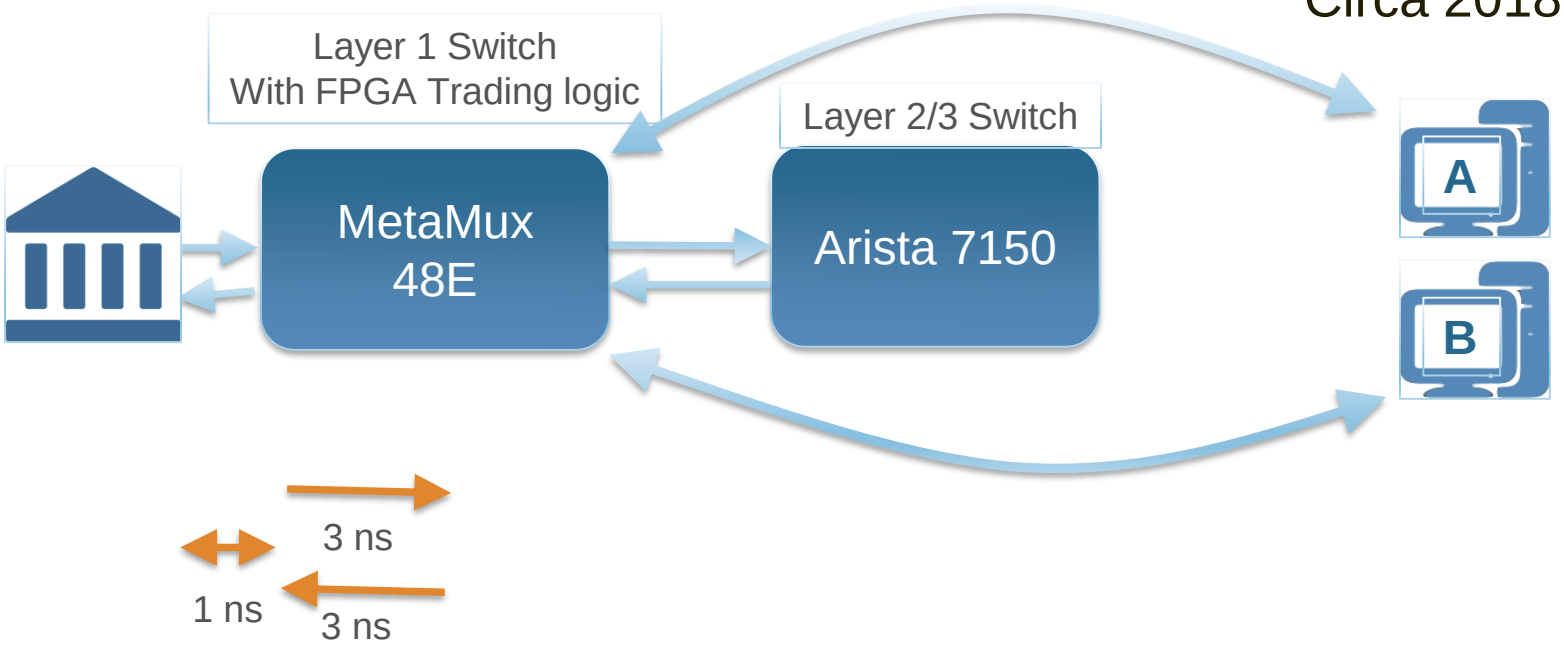
- Physical distance becomes the bottleneck
- FPGAs densely located with patch panel
- High bandwidth back-end networks



Latency

- Reducing the length of the cables now has a significant effect on total latency

Circa 2018

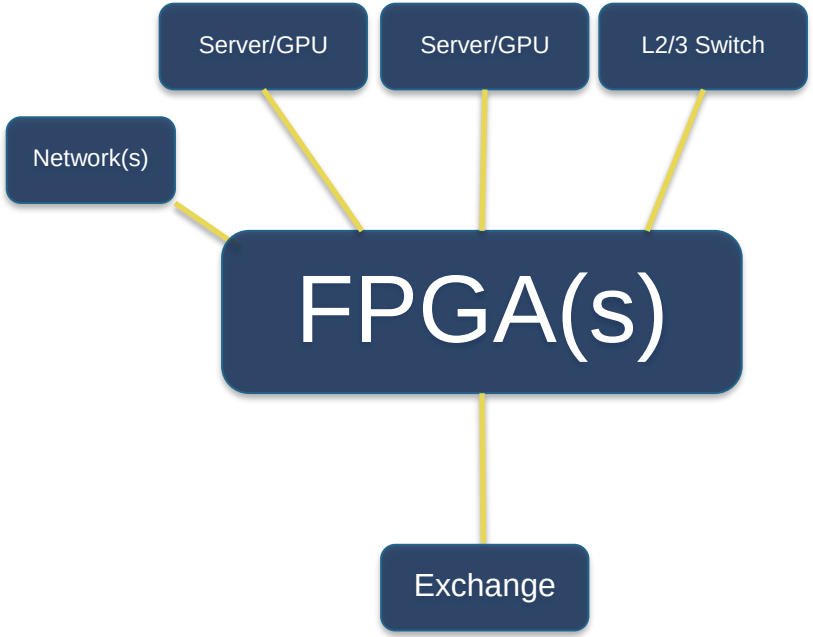


Total: 8 ns

NOT STAC
BENCHMARK

Latency

- FPGAs can respond in < 200 ns
- For FPGA-based trading, single-digit nanoseconds matter. *Any* switching latency is too high.
- Fiber latency *is* significant
- Still need monitoring and timestamping
- Solution: deploy custom trading algorithms to the FPGAs inside the switch – connect direct to the market.
- Result: minimal fiber latency in adjacent rack units, no switching latency, monitoring features.
- Pass through for time-insensitive traffic.



Latency: Where next?

- Hyper-accurate time synchronization via network protocols like PTP, White Rabbit
- Network visibility becomes increasingly important as response times becomes smaller and determinism improves.
- Network bandwidth upgrade via moves to 25GbE

Arista Innovations – Low Latency, APIs and IP Cores

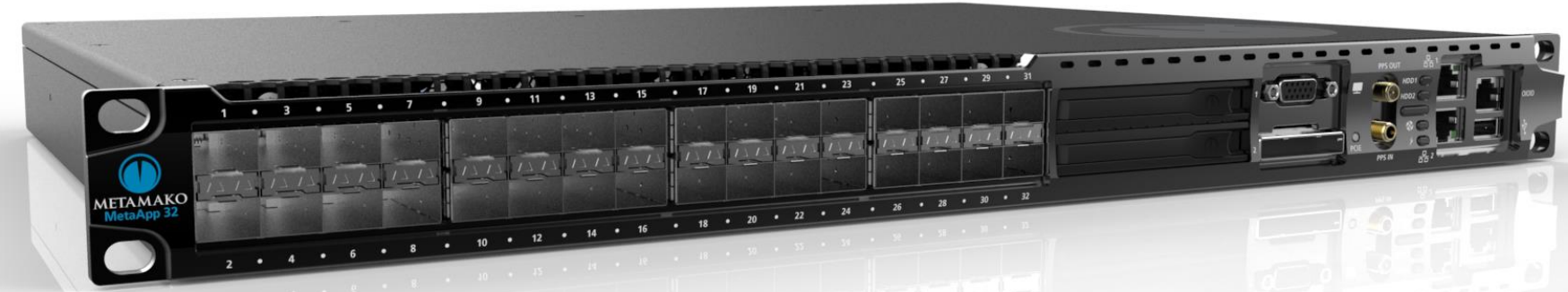
- MOSAPI – Write your own apps for MOS
  - A software framework for integrating apps with MOS
  - Used by most Arista MOS apps
  - Available on the web site now.
  - Take advantage of MOS features to write full featured applications on the switch.

- New dev-kits and IP cores
  - MMP IP core – get an AXI stream between FPGAs in 8 nanoseconds
  - Mux IP core – build our mux into your FPGA image
  - Revised/refreshed E-series dev kit

- MetaMux 3.0.0
  - < 45 ns average latency (< 39 ns minimum)

Arista Innovations – A32EH

- EPCIE combined with one or three VU9P FPGAs
- MMP support between FPGAs (8 ns latency)

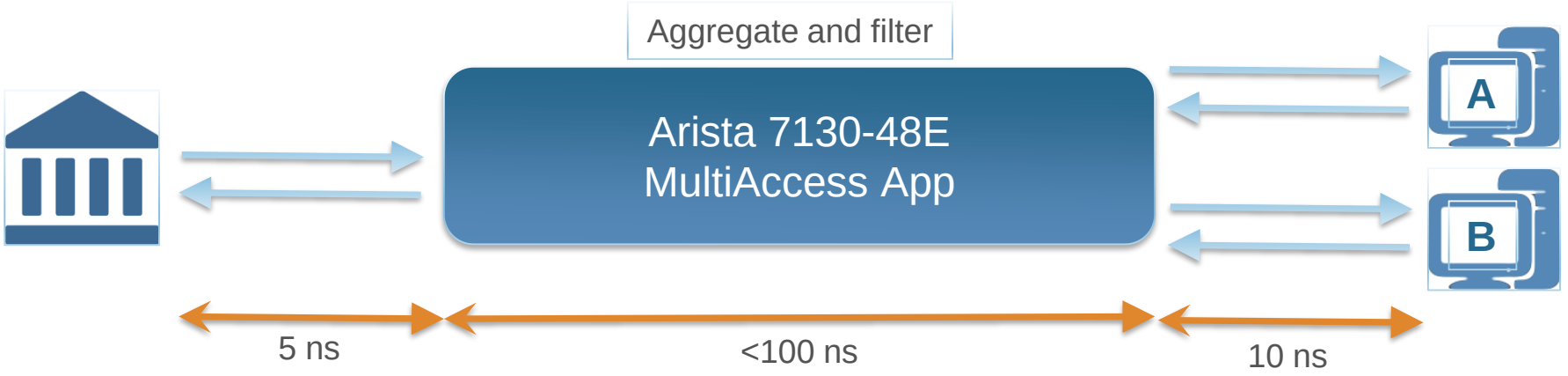


Arista Innovations – MetaWatch for Tap/Agg

- In production for two years and stable with major critical clients
- Metawatch-0.9.0 out this week:
  - Sub-nanosecond precision timing on all modes -- metawatch-0.9.0
  - 40G capture ports enabled
  - Support for running on triple-FPGA M48EP -- *no deep buffer*
  - Aggregation in time order from many-to-one output

Arista Innovations – MultiAccess

- MultiAccess is a new app targeting managed service providers.
- Mux on the way into a service, and filter on the way out.
- Avoids privacy issues with true Layer 1 solutions.



Total: < 200 ns

NOT STAC BENCHMARK

A few crazy ideas…

- Higher bandwidths
- Co-location within a machine?
  - Within an FPGA?
- The rise and fall of co-location?
  - Cloud based trading? Rent resources within the exchange?
- Fairness and determinism within an exchange?
- Timestamps with precision better than Nyquist?
  - Better than 49 ps for 10GbE
- Avoiding queueing? 100G? 400G aggregation?

ARISTA color Palette


16325B


58595A


25678D


5588B7


221F20


929598


BBBDBE


E0872D

E8D754


B39E2E


ACC37B


6A8236


ACD13B


FD661B


FAAF4B

Alternate Text Colour Hex 112346

ARISTA