



November, 2019
daves@arista.com

Why accuracy-driven markets will transform trading.

“That’s when I realized the markets are rigged. And I knew it had to do with the technology.”

–Brad Katsuyama, Quoted by Michael Lewis, Flash Boys

What this talk is *NOT* about...

- Market micro-structure.
- Policy
 - Policy vs. Mechanism.

“The markets aren’t perfect. And technology can help.”

-David Snowden, STAC Summit Chicago, October 2019

A customer's view of an exchange...

Exchange fibres



What does it mean to be fair?

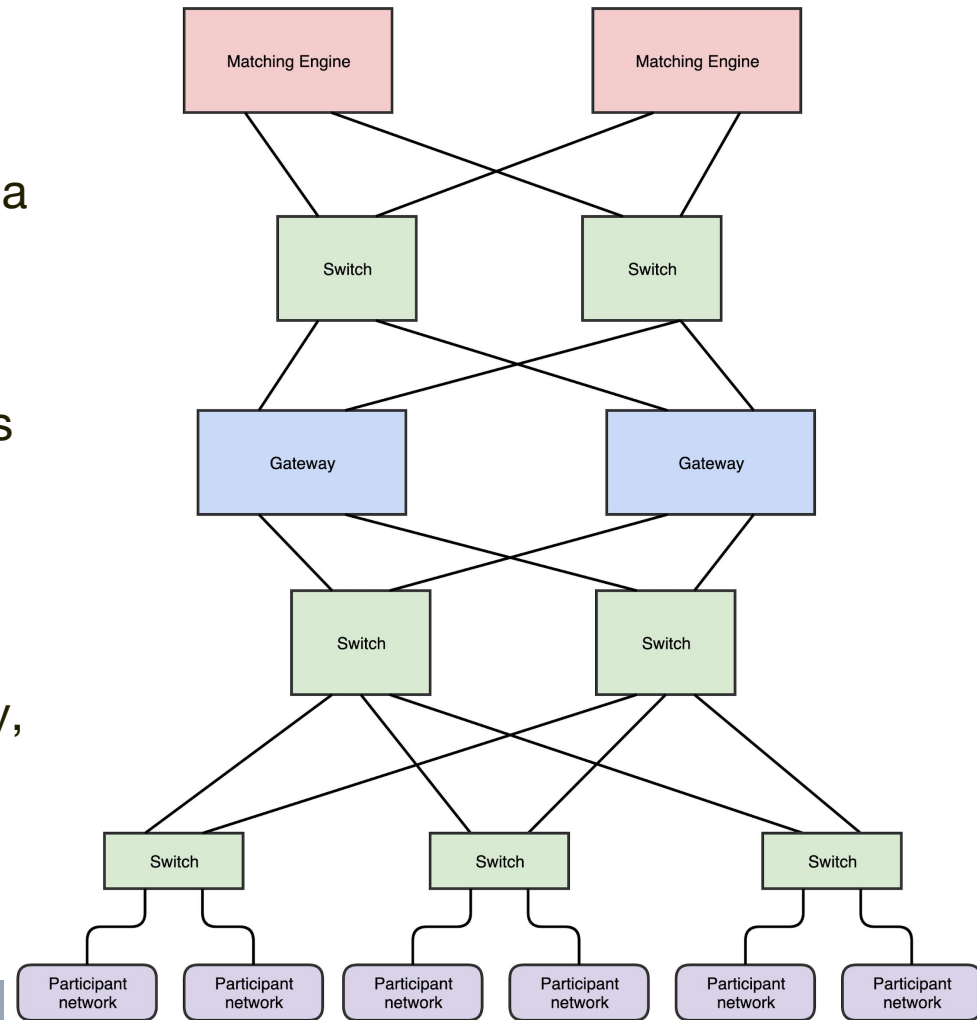
- Venues should do what it says on the tin. e.g...

ASX Trade automatically matches all trades in Price/Time Priority on a continuous basis.
<https://www.asx.com.au/about/trading-hours.htm>

- TX: Two key “time” properties: Two equal orders, placed in a sequence, should be executed in that sequence.
 - Other policies might make more sense?
 - Traders should be isolated from one another.
- RX: Two market data feeds deemed to be equal by the venue should deliver the same information at the same time (to within stated tolerances).

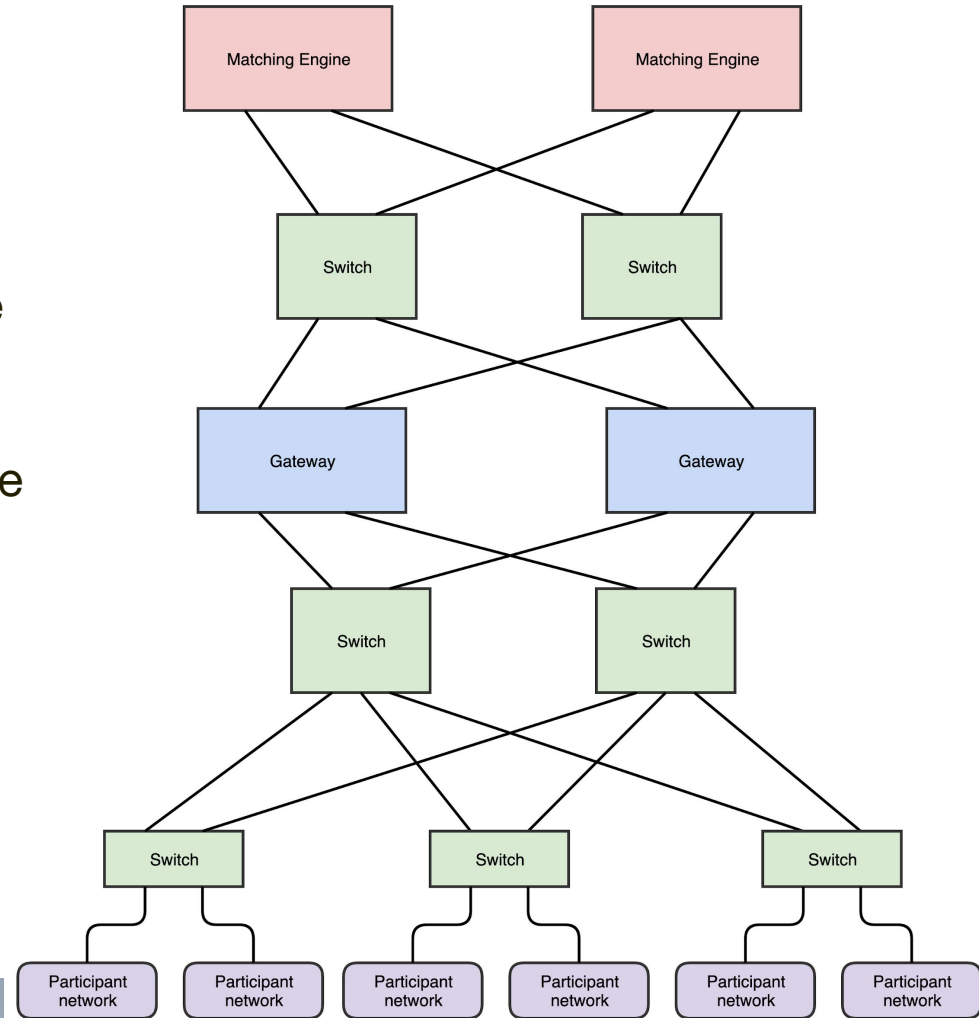
A typical exchange...

- Participants are co-located with a top-of-rack cross-connect.
- Matched-length fibres deliver messages to the exchange.
- An L3 network with BGP delivers unicast traffic for orders.
- Market data is delivered via sparse multicast (PIM).
- Gateways process this, manage sessions, check for order validity, route to matching engines.
- Matching engines manage the books, match orders, and generate data.



Typical issues

- Multiple gateways -- each gateway gets uneven load, so takes more or less time -- Venue moves to a single gateway?
- Individual switches may get more load, and therefore more contention queuing -- Trader connects to all the switches, monitors load, places orders? Avoids adverse selection?
- Sources of load -- different traders, sparse multicast.

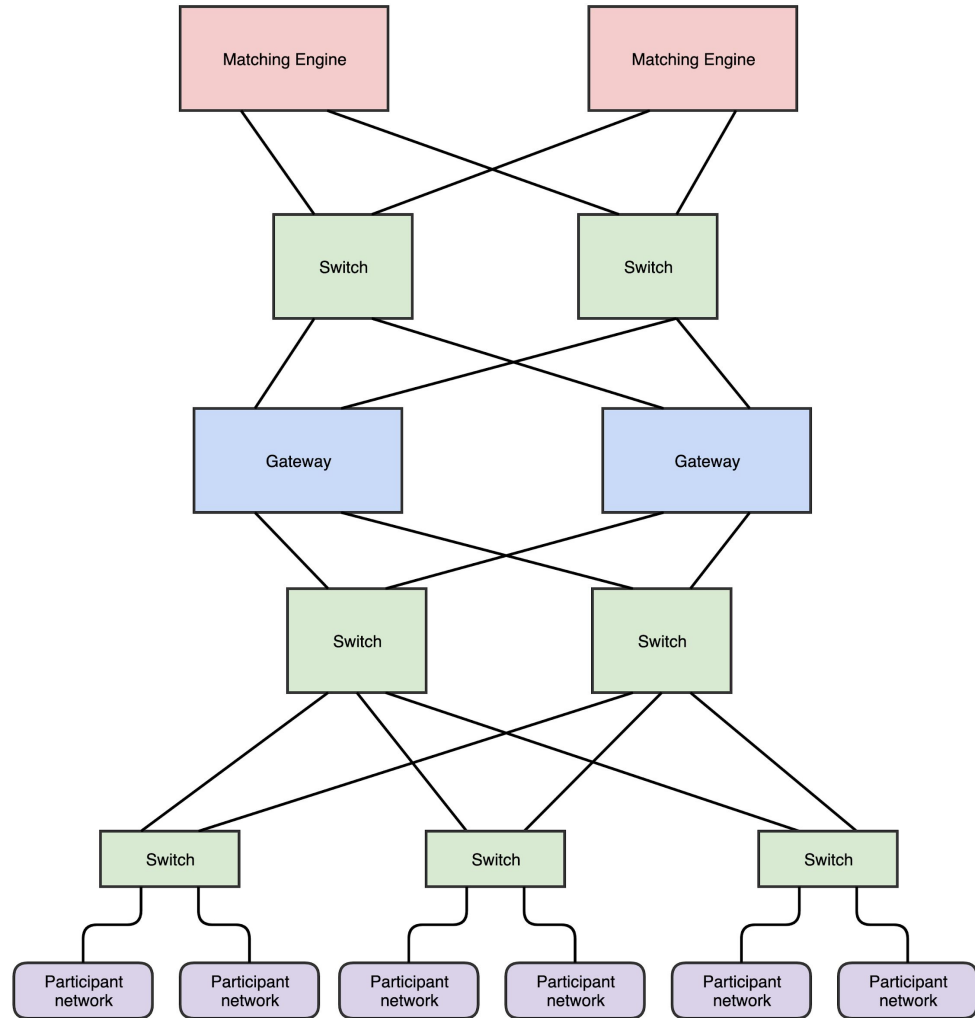


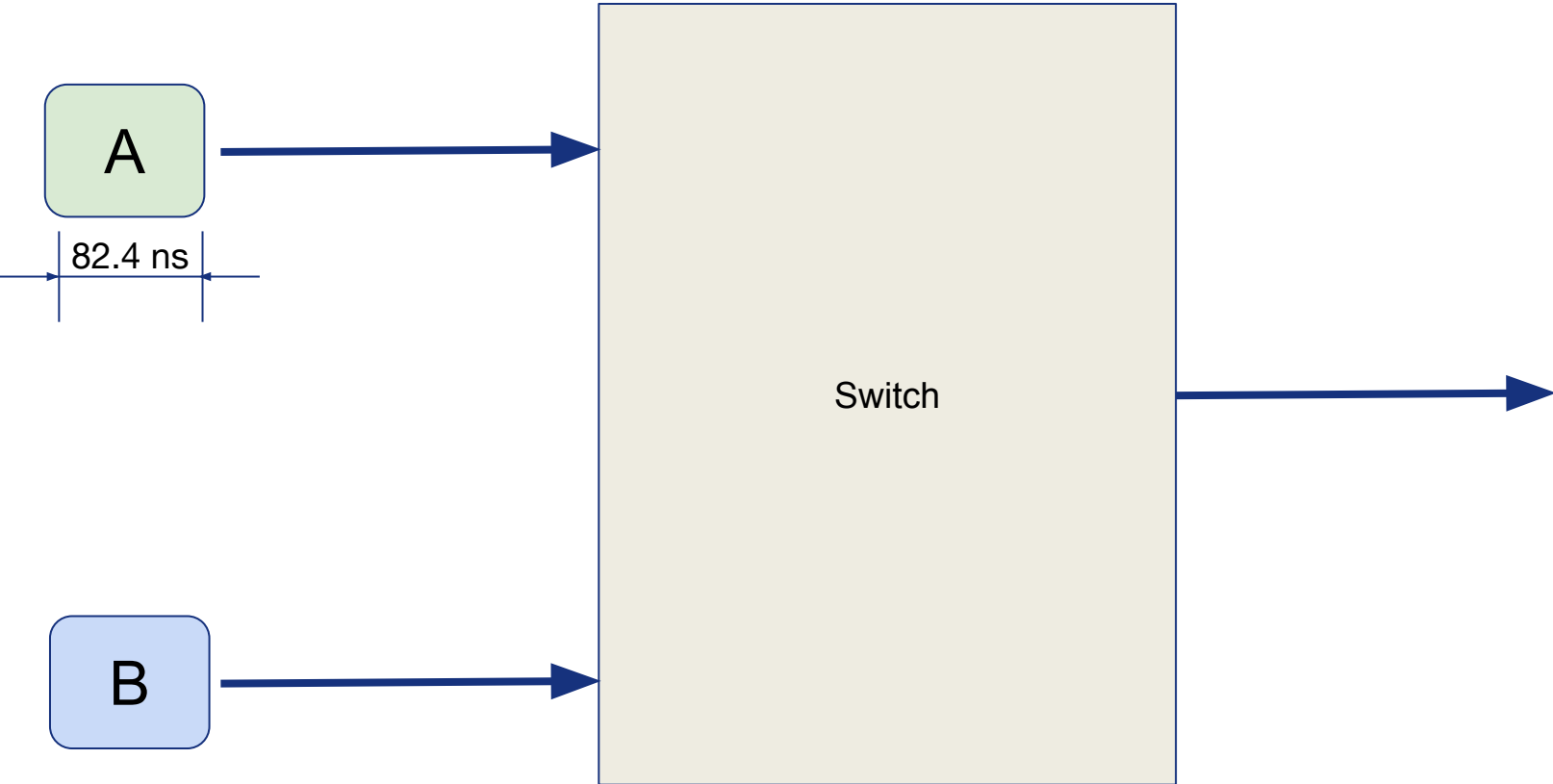
Sequencing point

- Sequencing point -- the place where the execution order is fixed.
- By default, the sequencing point is usually at the hand-off to the matching engine software.
 - Note: this is, by definition, store-and-forward.
 - Software usually executes once it's received the *entire* order message.
- Going with a single gateway (can) move the sequencing point
 - Dependent on gateway implementation

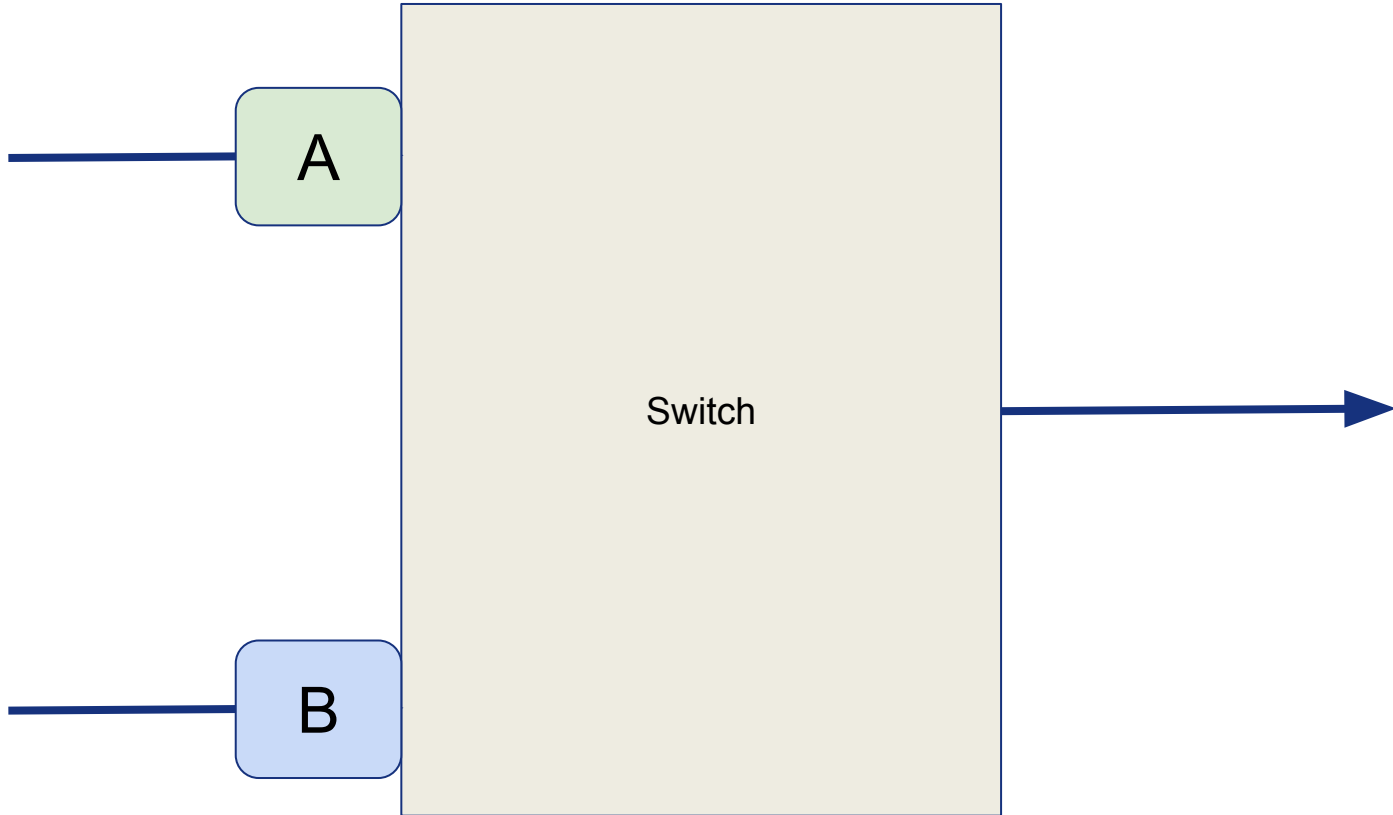
Typical issues

- Switches and gateways are not deterministic.
- Gateway network cards and network stacks do not deliver packets to software in-order.
 - Sometimes it's faster not to.
 - Use a specialised network stack like Solarflare WODA.
- Some racks are closer than others -- Equalise the length of the cross-connect fibres.

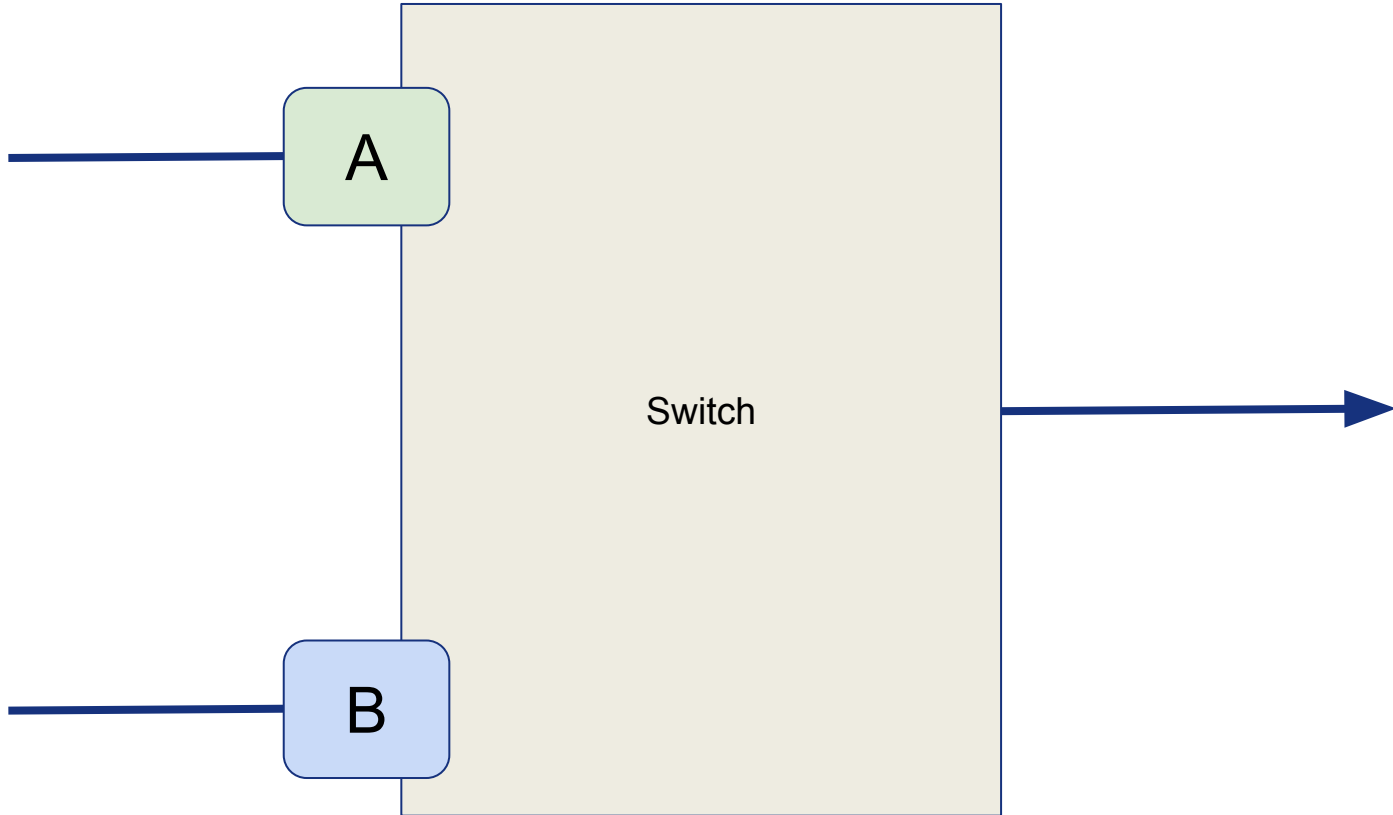




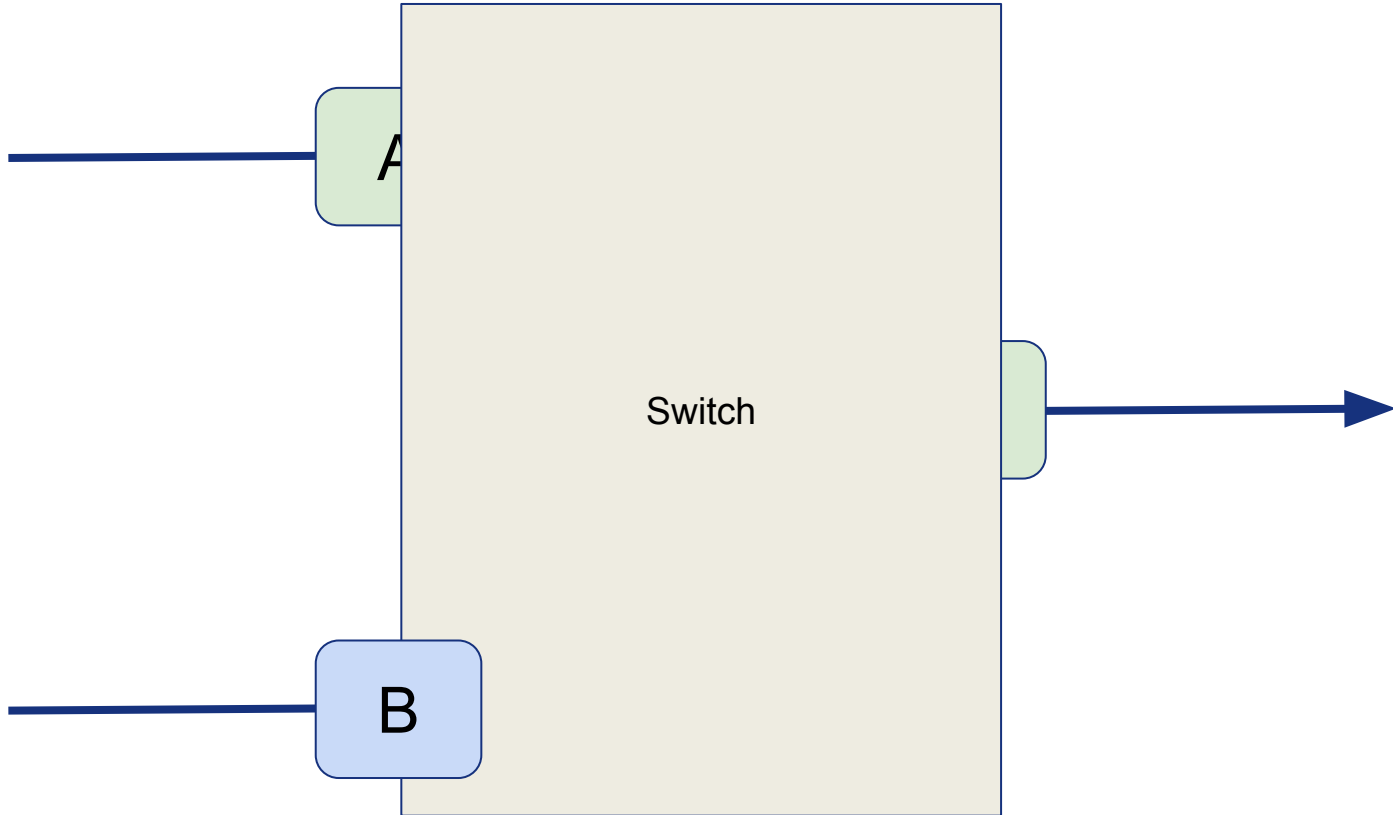
$t = 0$ ns



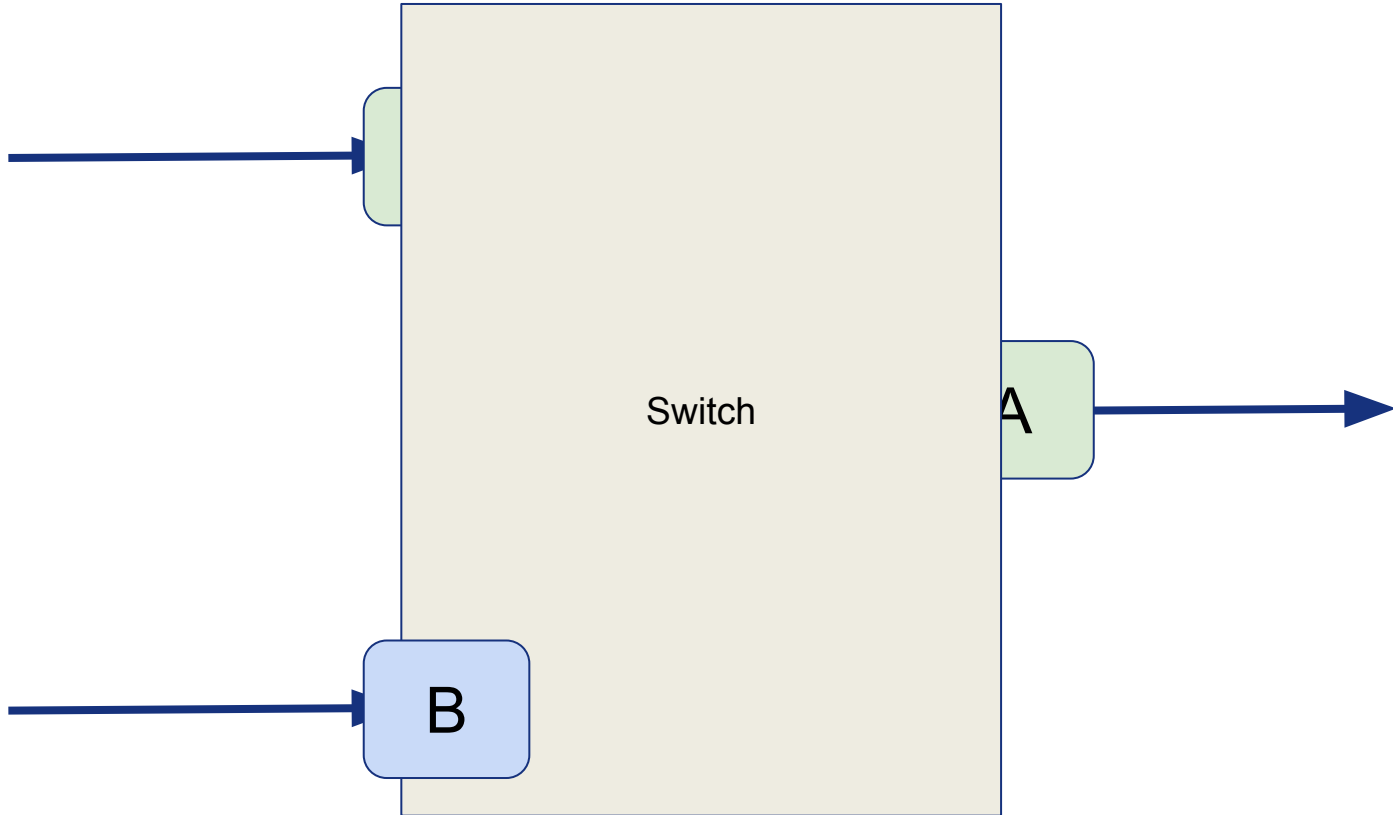
t = 300 ns



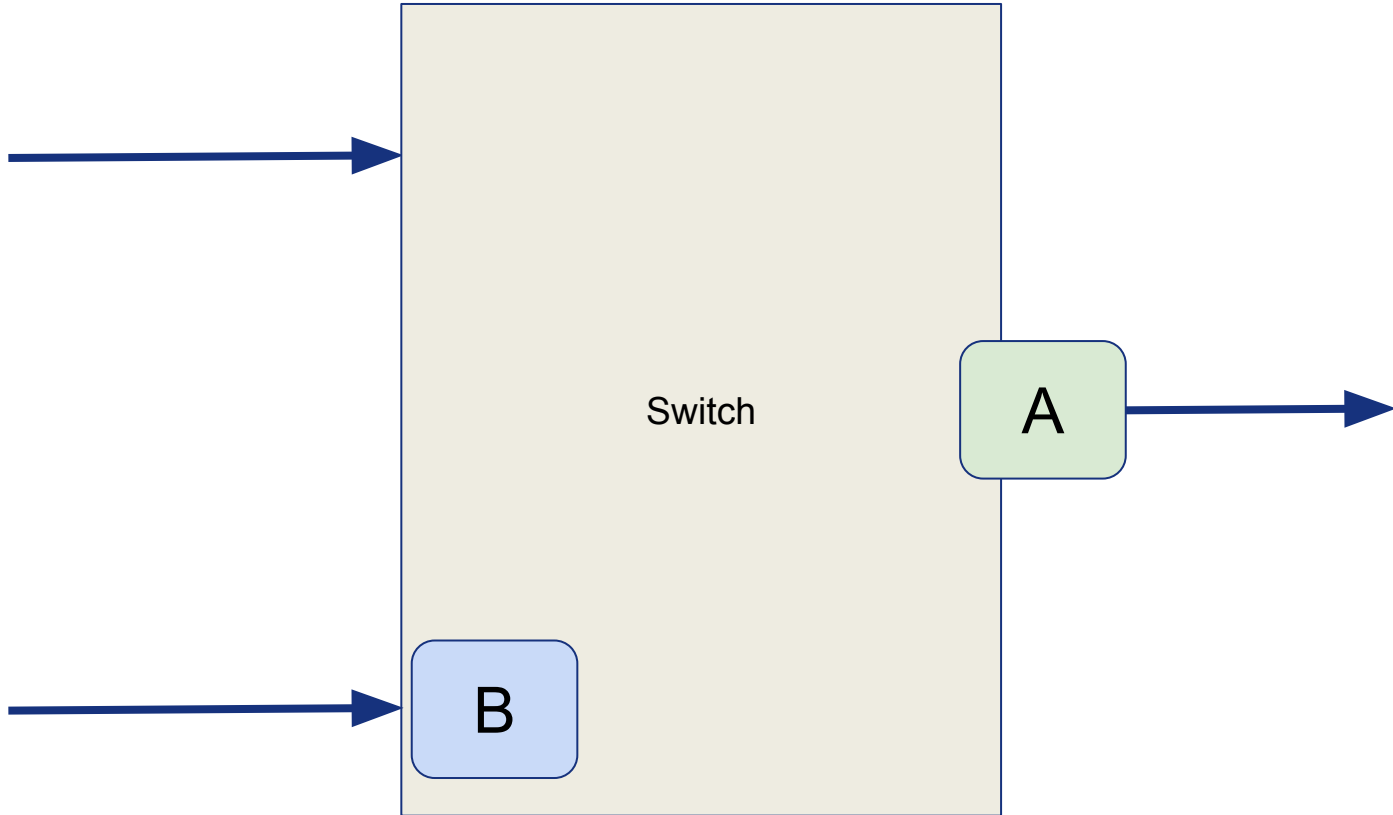
t = 330 ns



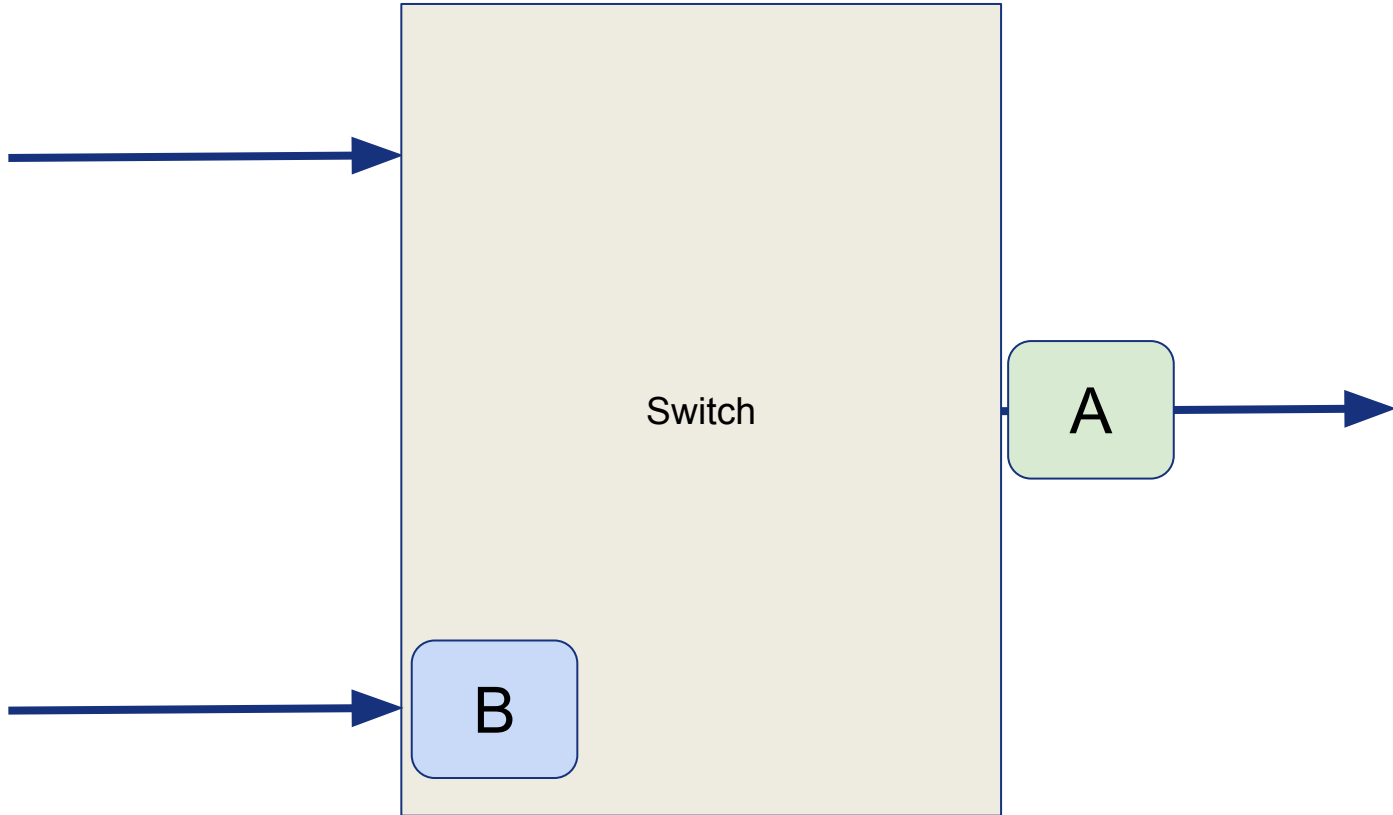
t = 350 ns



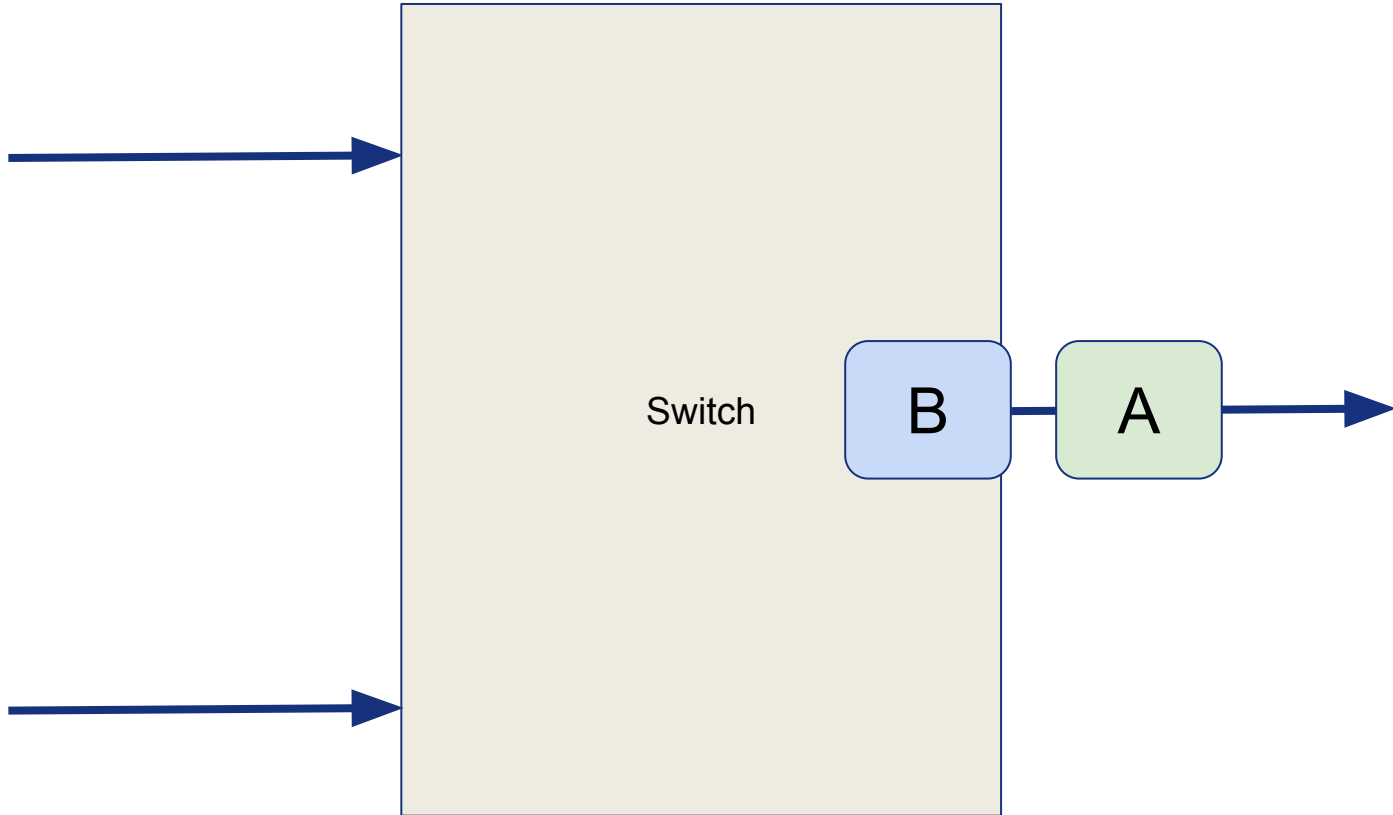
t = 370 ns



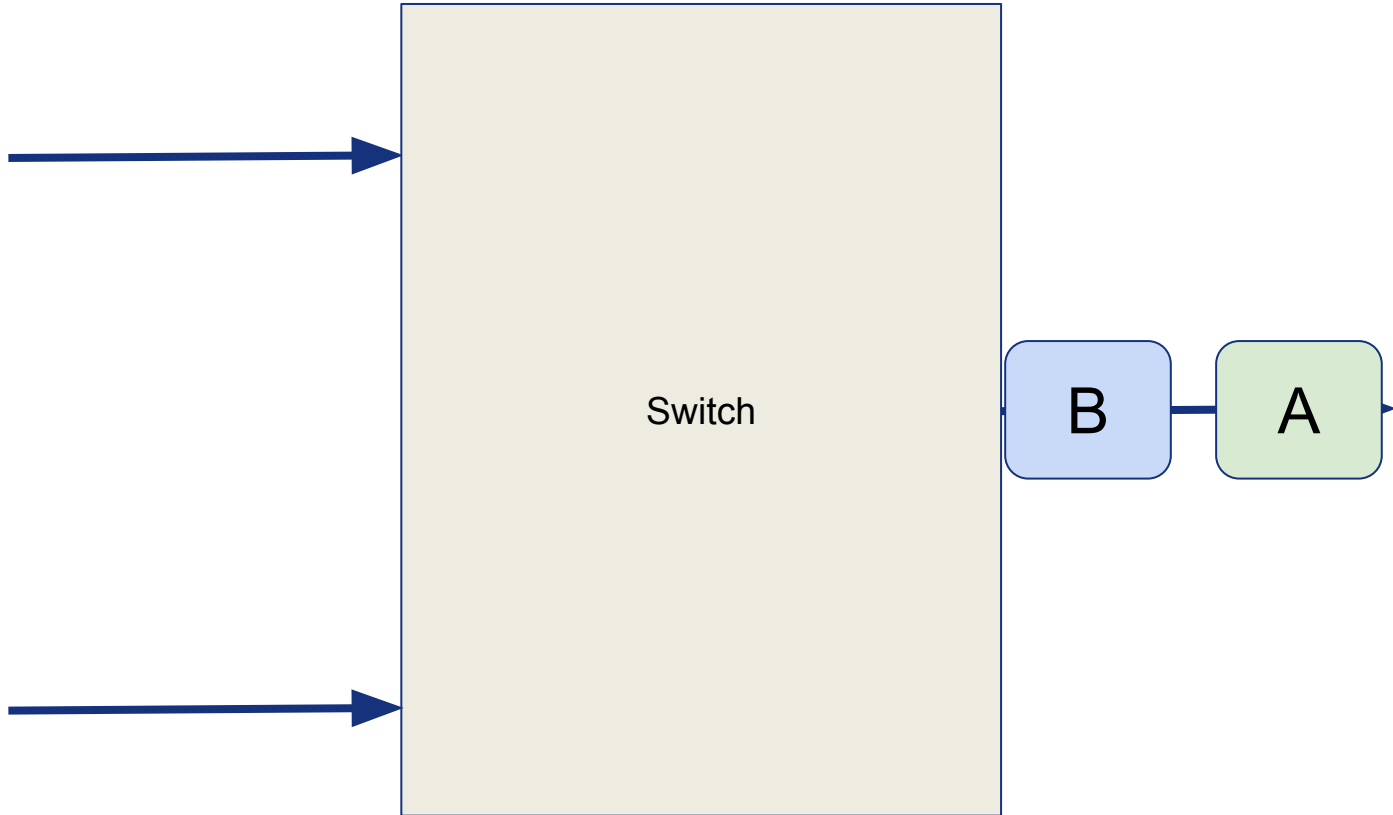
t = 382.4 ns



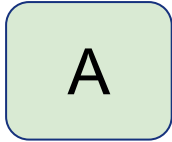
t = 412 ns



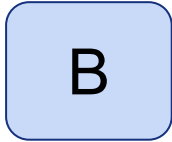
t = 430 ns



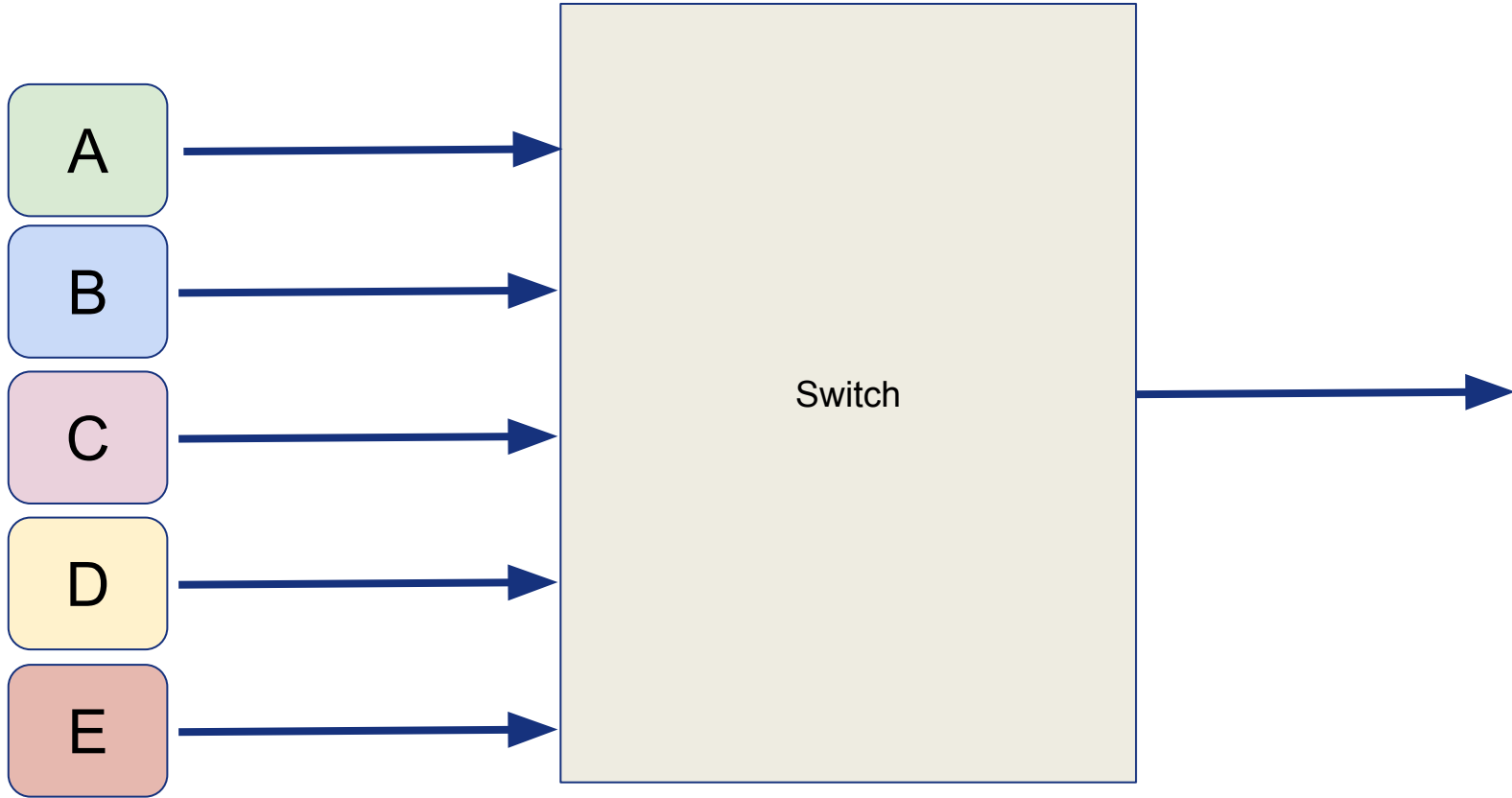
$t = 494.4 \text{ ns}$

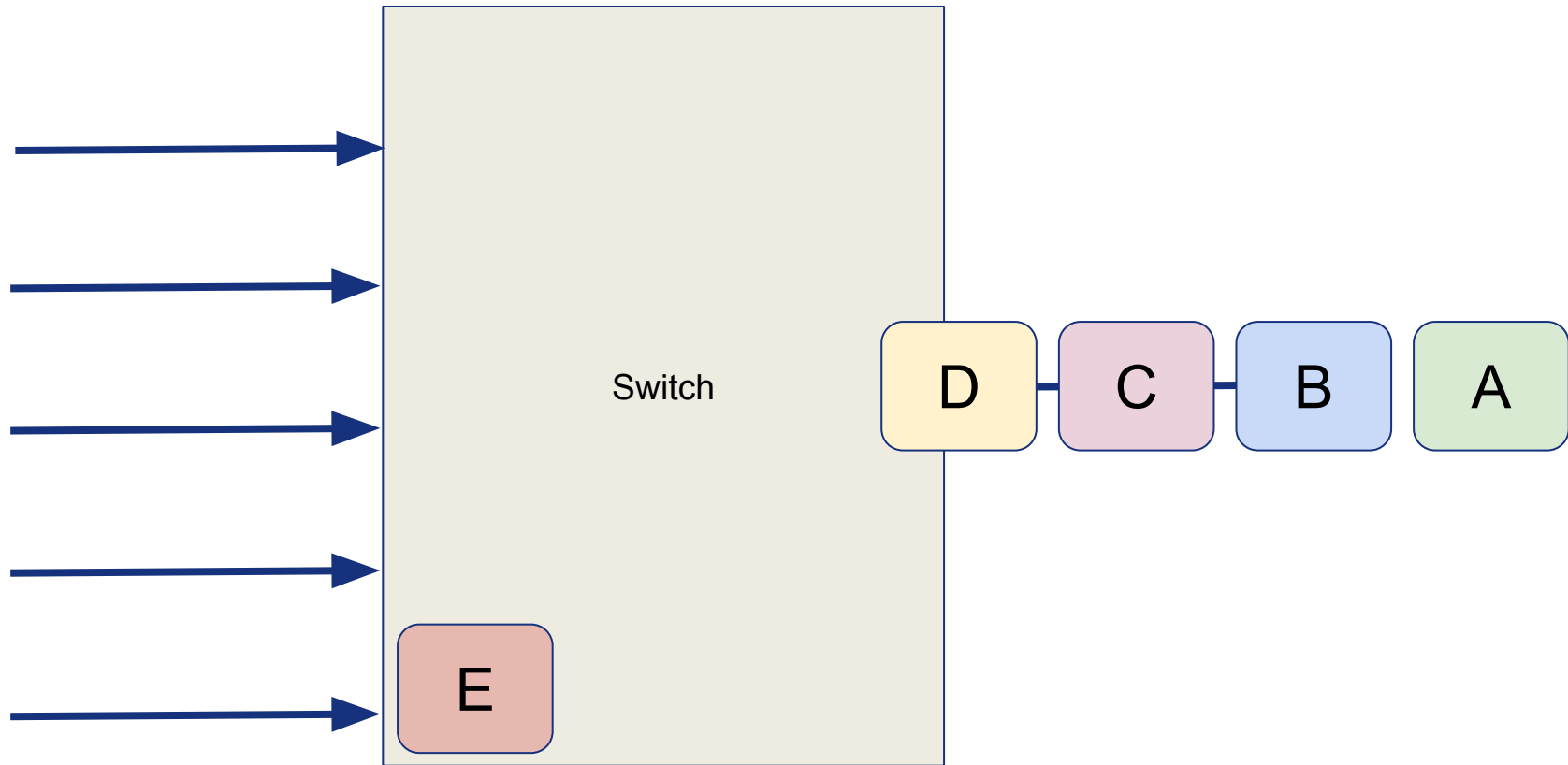


Starts to arrive: 380 ns
Finishes arriving: 462 ns

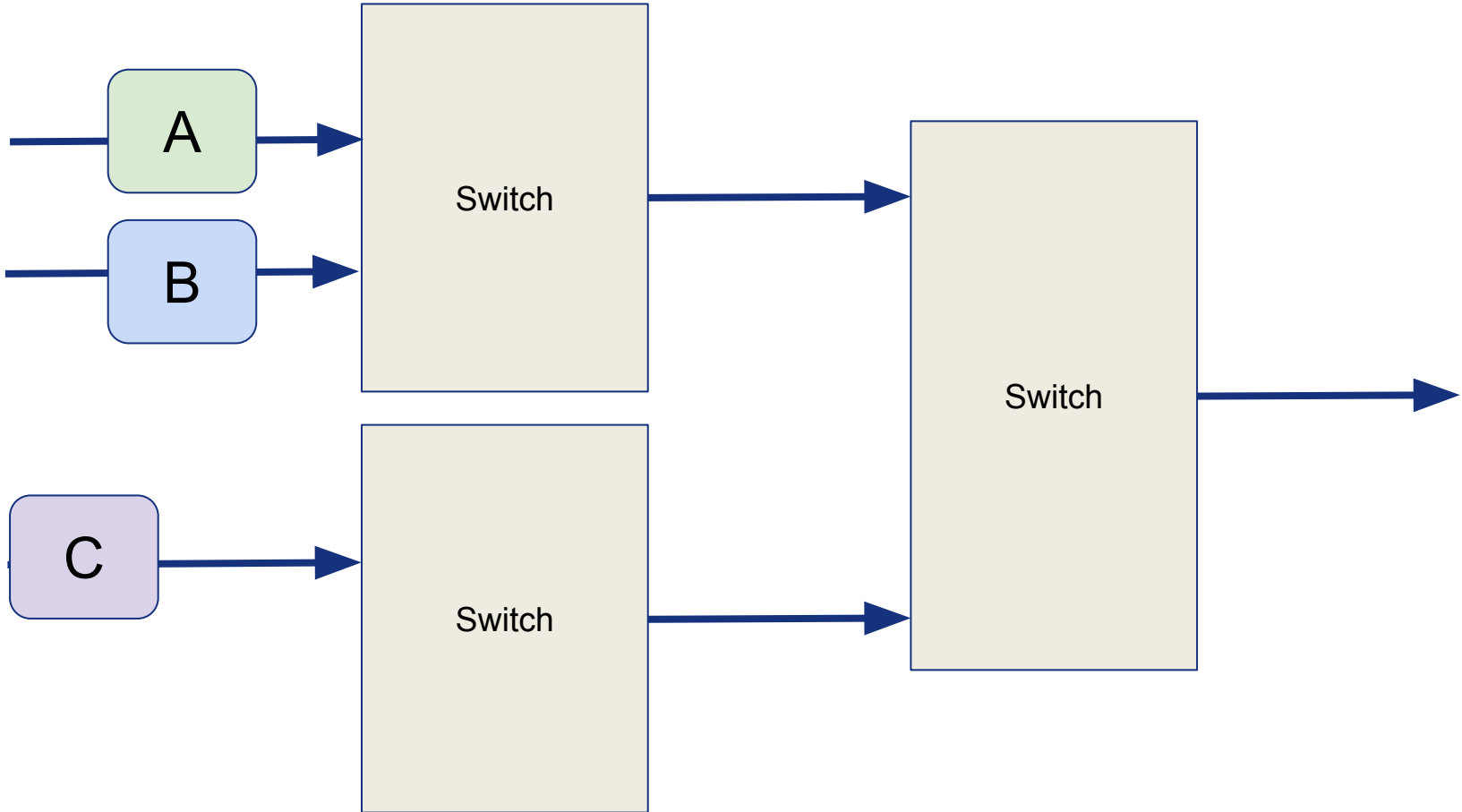


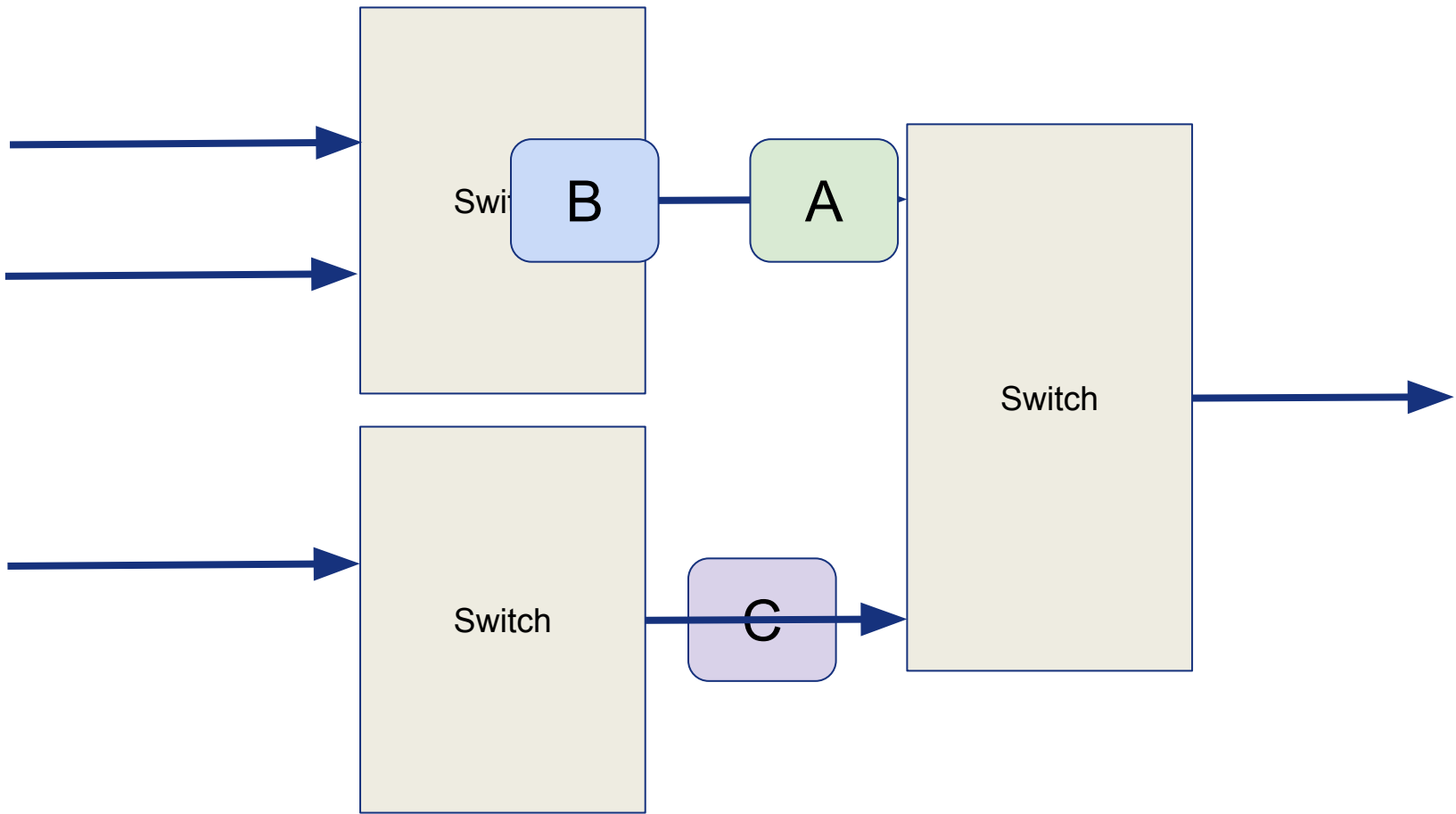
Starts to arrive: 480 ns
Finishes arriving: 562 ns

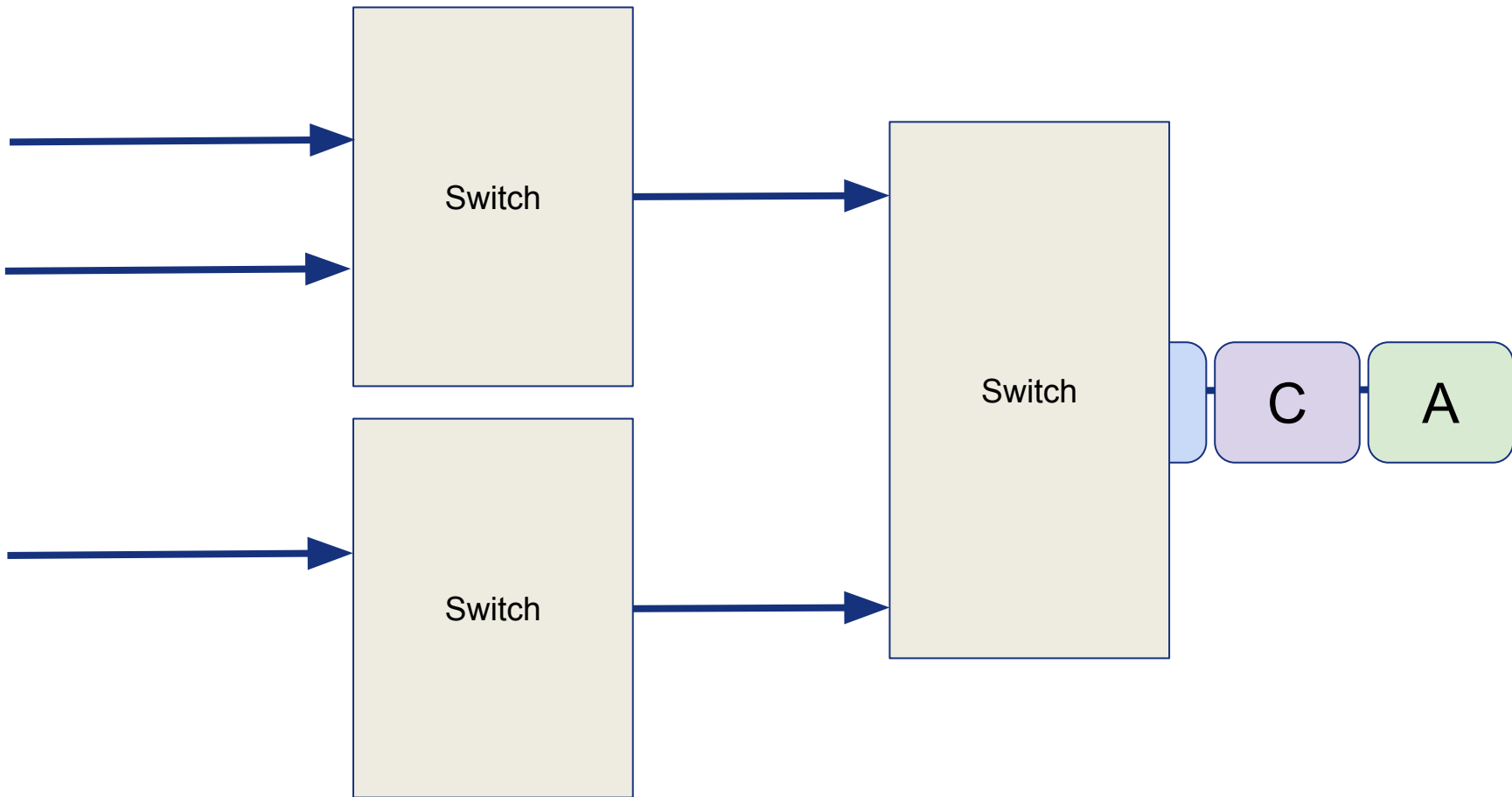




A	Starts to arrive: 380 ns Finishes arriving: 462 ns
B	Starts to arrive: 480 ns Finishes arriving: 562 ns
C	Starts to arrive: 580 ns Finishes arriving: 662 ns
D	Starts to arrive: 680 ns Finishes arriving: 762 ns
E	Starts to arrive: 780 ns Finishes arriving: 862 ns

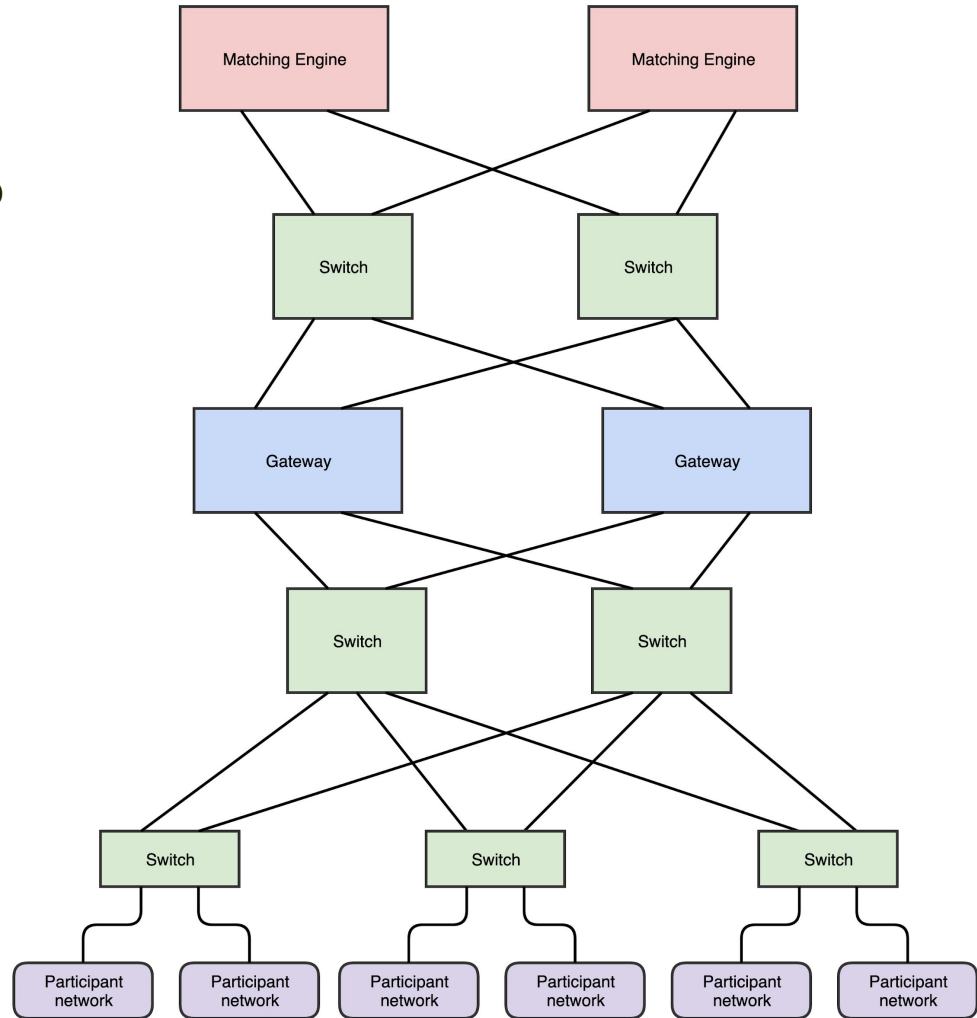






Typical issues

- Equal-length fibres that aren't so equal.
 - Fibre-lengths inside the venue



Getting to the nub of it...

- Few venues really understand their imperfections.
- As always, there's a trade-off between latency/determinism and bandwidth

Some things we've talked about...

- *We want our response time to be better*
 - Latency at exchanges is usually pursued to gain determinism.
- *The HFTs are going to hate you for this*
 - Most HFT firms that I talk to would be happier to spend their time, energy and money on something more productive.
- *We like the randomness... It helps stop the HFTs.*
 - If you want randomness, get to perfection, and add *true* randomness, not a predictable systematic error.

Latency vs. determinism

- Latency optimisation is about reducing some measure of delay
 - minimum, median, average, worst case?
- Determinism is about consistency
 - Consistency is required for fairness -- the exchange should do what it says on the tin.
- Bandwidth is about volume
 - How many orders per second can we handle?

For fairness, venues need to optimise for determinism at the participant interface.

Some other questions...

- What's the event that matters?
 - First bit of the first or last network packet containing the order?
 - Out of order TCP segments, IP fragments?
 - The last bit of the order network message?

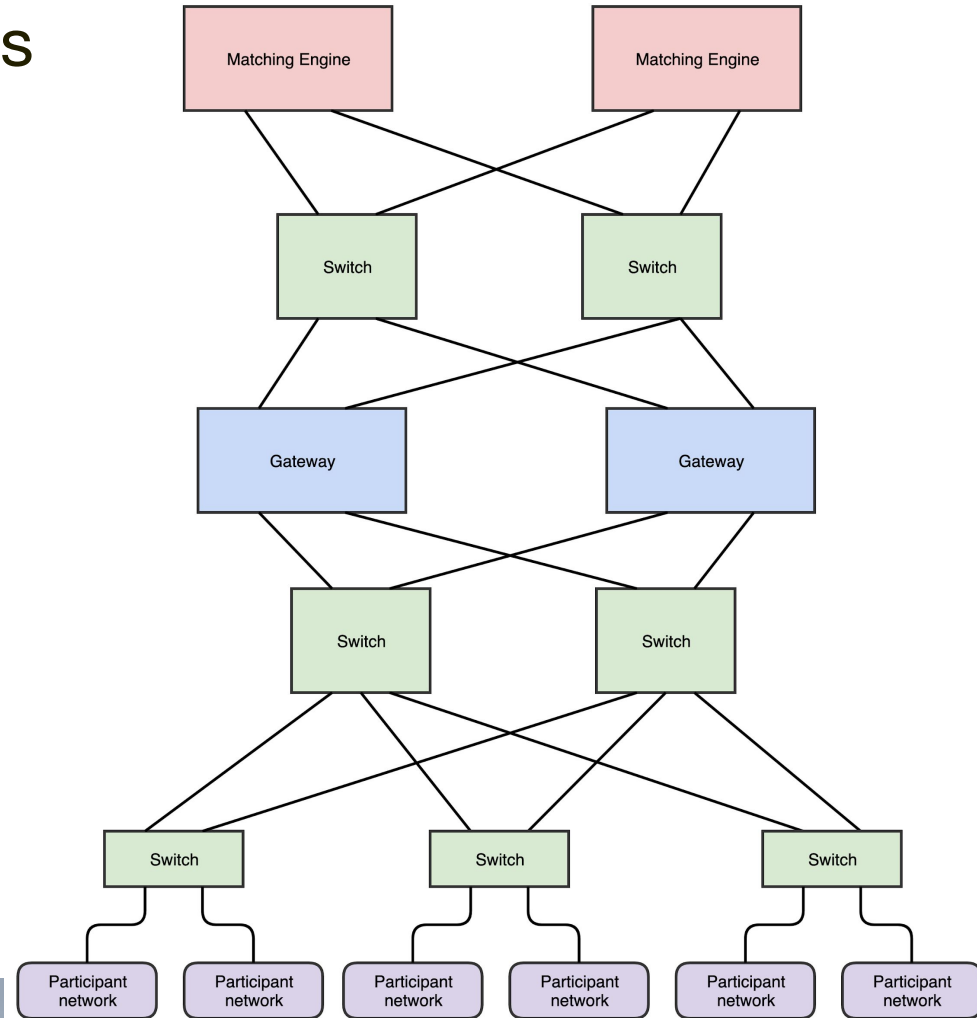
Typical venue optimisations

Do:

- Co-location, matched fibres
- Low-latency switches
- Single threaded software
- Single gateway architecture
- Single data rate networks (10G)
- Cut through switching
- Specialist network stacks (e.g. Solarflare's Onload, WODA)
- FPGA implementations

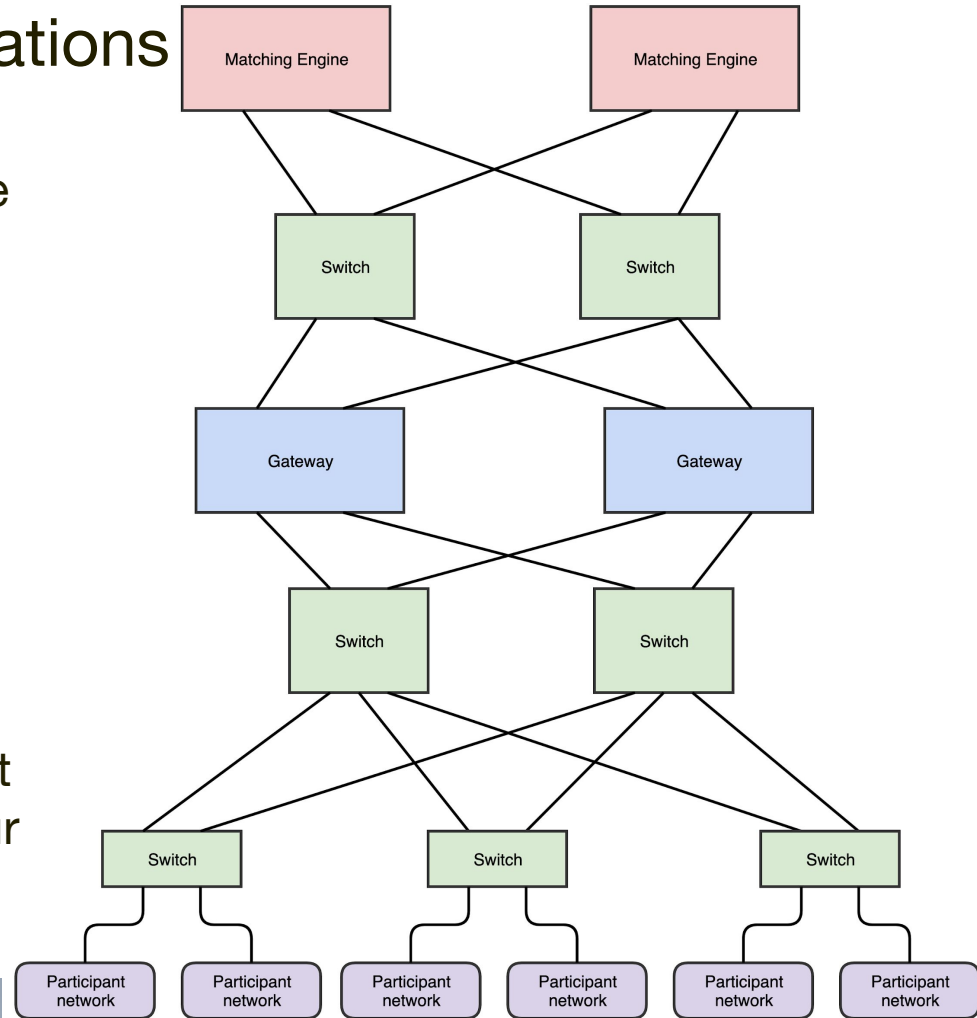
Don't:

- Use VMs, cloud
- Use new high bandwidth links



Typical participant optimisations

- Connect to every switch and use the earliest copy of each message.
- Model switch congestion and queueing to avoid adverse selection.
- Model/monitor the gateways.
- Game the network stack to reduce serialisation delay
- Start transmitting early to “reserve” the line -- pad the front of the packet until you know your order info.



What can we do?

- Step 1: Measure what needs to be measured, as accurately as it needs to be measured.
 - Understanding the symptoms will indicate the problems.

“To know thyself is the beginning of wisdom.”

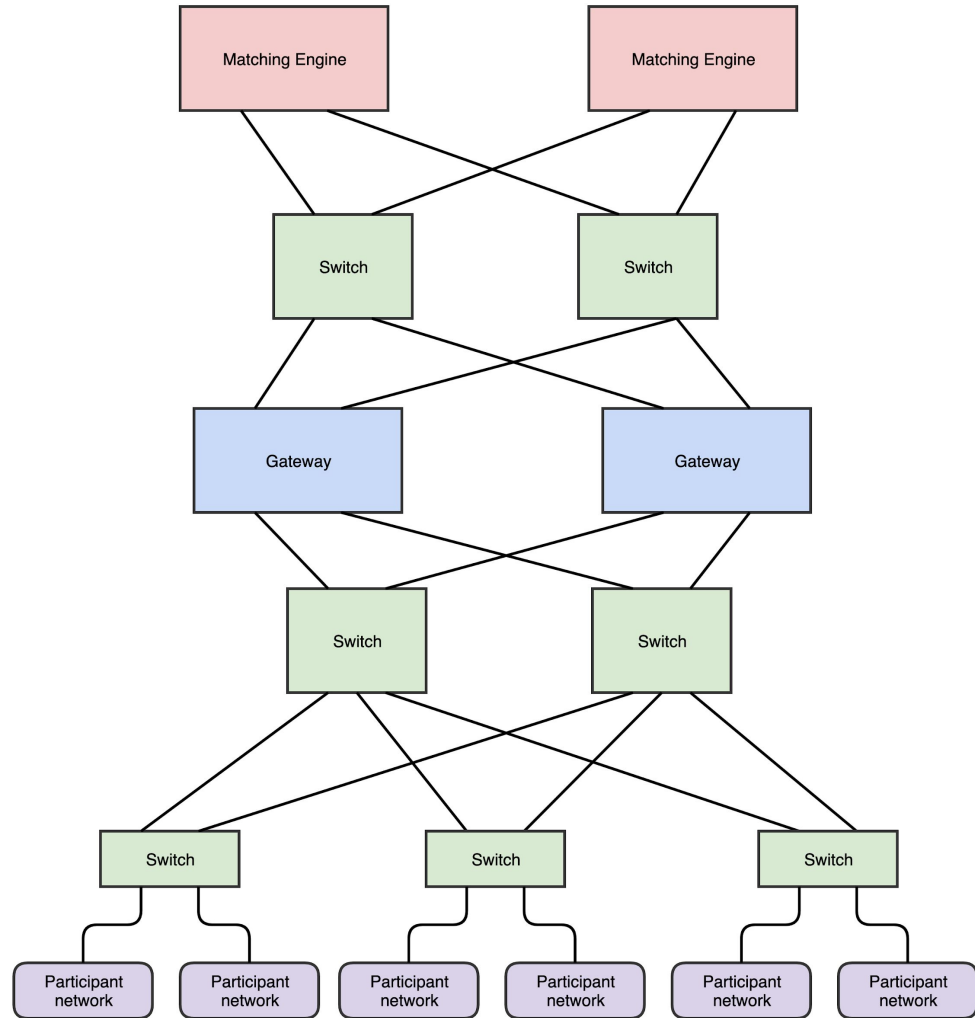
— Socrates

How to measure event timing in an exchange?

- Software timestamps
- Network packet capture
- In the future:
 - virtual network interface?
 - In the NIC?

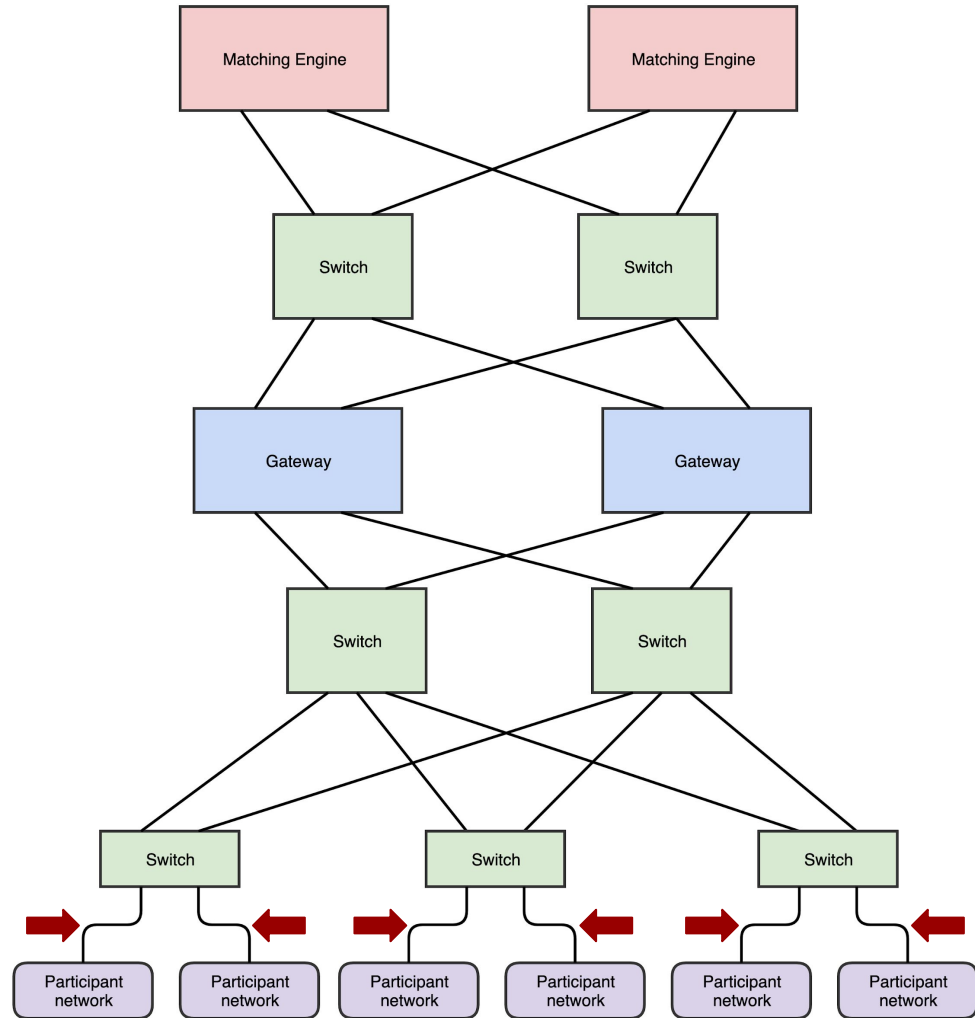
- Disclaimer (sort of): we have a product that does this.

Where to measure?



Where to measure?

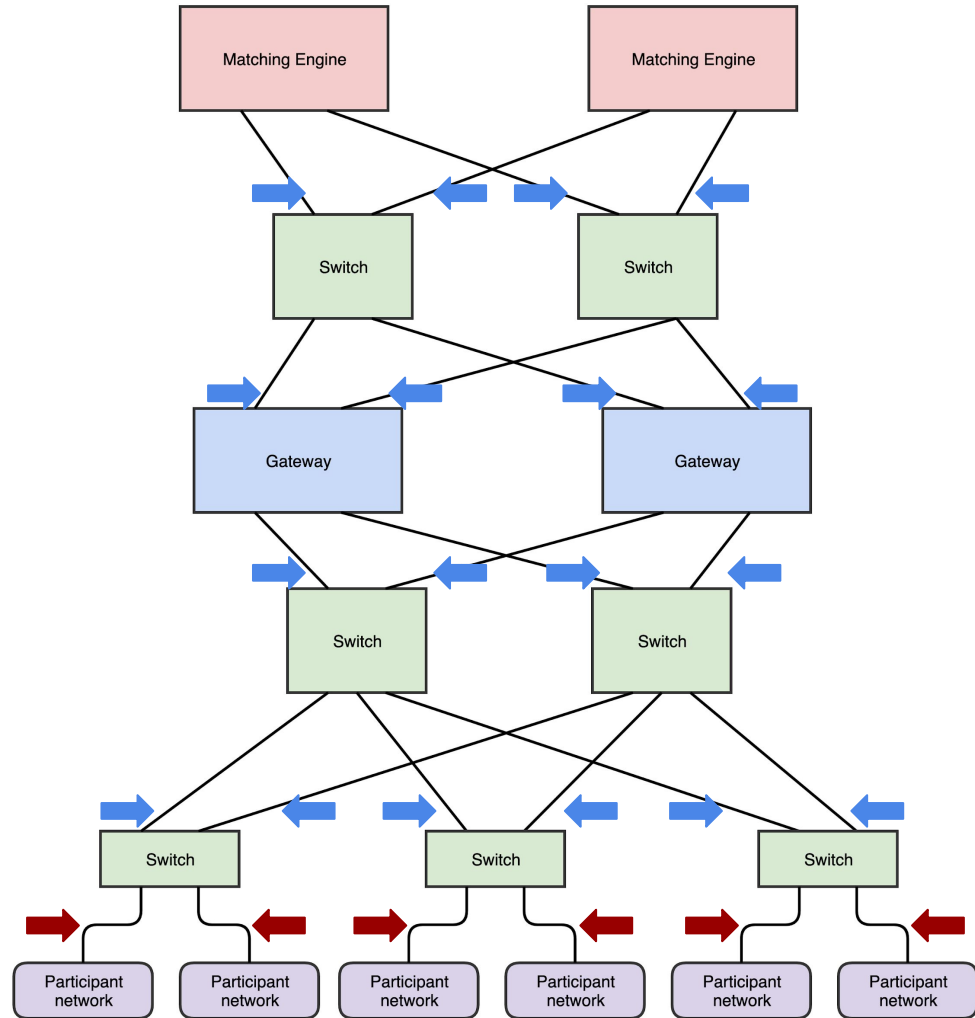
➔ Measuring means measuring at the interface that matters.



Where to measure?

➔ Measuring means measuring at the interface that matters.

➔ But there are lots of places that useful to know about.



But how?

- Technology to the rescue. This is a **solved** problem.
- Software
 - Clocks can be synchronised to substantially sub-micro errors.
 - e.g. PTP, Ticktock Huygens, FSM Timekeeper
- Network capture:
 - We can measure across thousands of network links to sub-nano resolution. (Not STAC benchmark). e.g. Arista.
- STAC-TS is a great tool to measure this area.

NOT STAC BENCHMARK

But how?

- Deutsche Boerse have instrumented to better than 10 ns.
 - Seven Solutions + Arista
 - Very wide ranging instrumentation
 - Great example of transparency!
 - https://www.eurexchange.com/resource/blob/48918/09570713f62b7719635742e52d525d87/data/presentation_insights-into-trading-system-dynamics_en.pdf

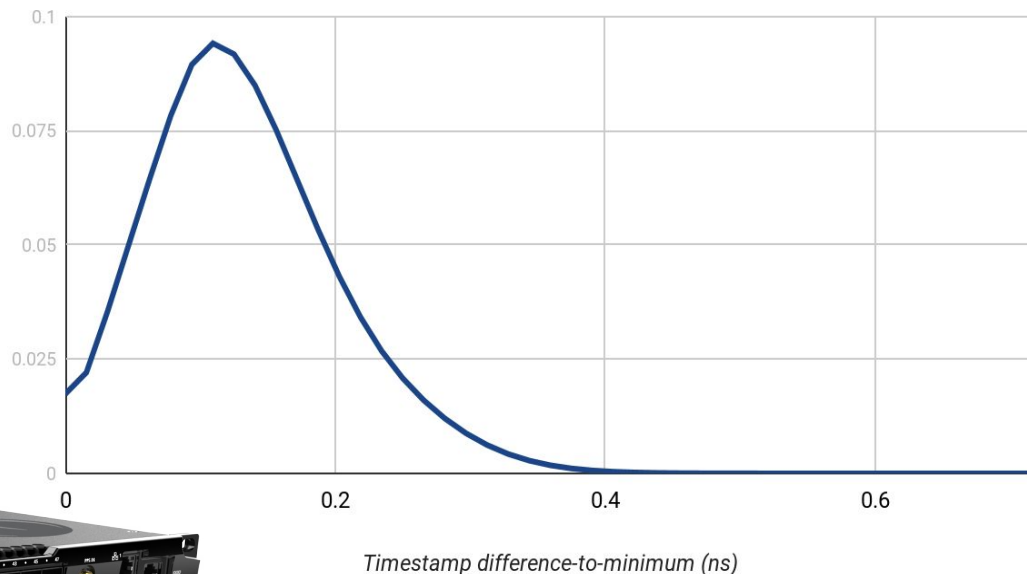
NOT STAC BENCHMARK

Some not-STAC benchmarks...

- Two devices, sync'd via PPS.
- Replicate a packet to all 96 ports.
- For each packet, measure the difference to the min timestamp.
- **Max difference: 0.719 ns**

Arista 7130L - Two switches, PPS sync

Timestamp PDF, 2 Devices, 48 ports per device, 10G



NOT STAC BENCHMARK

ARISTA

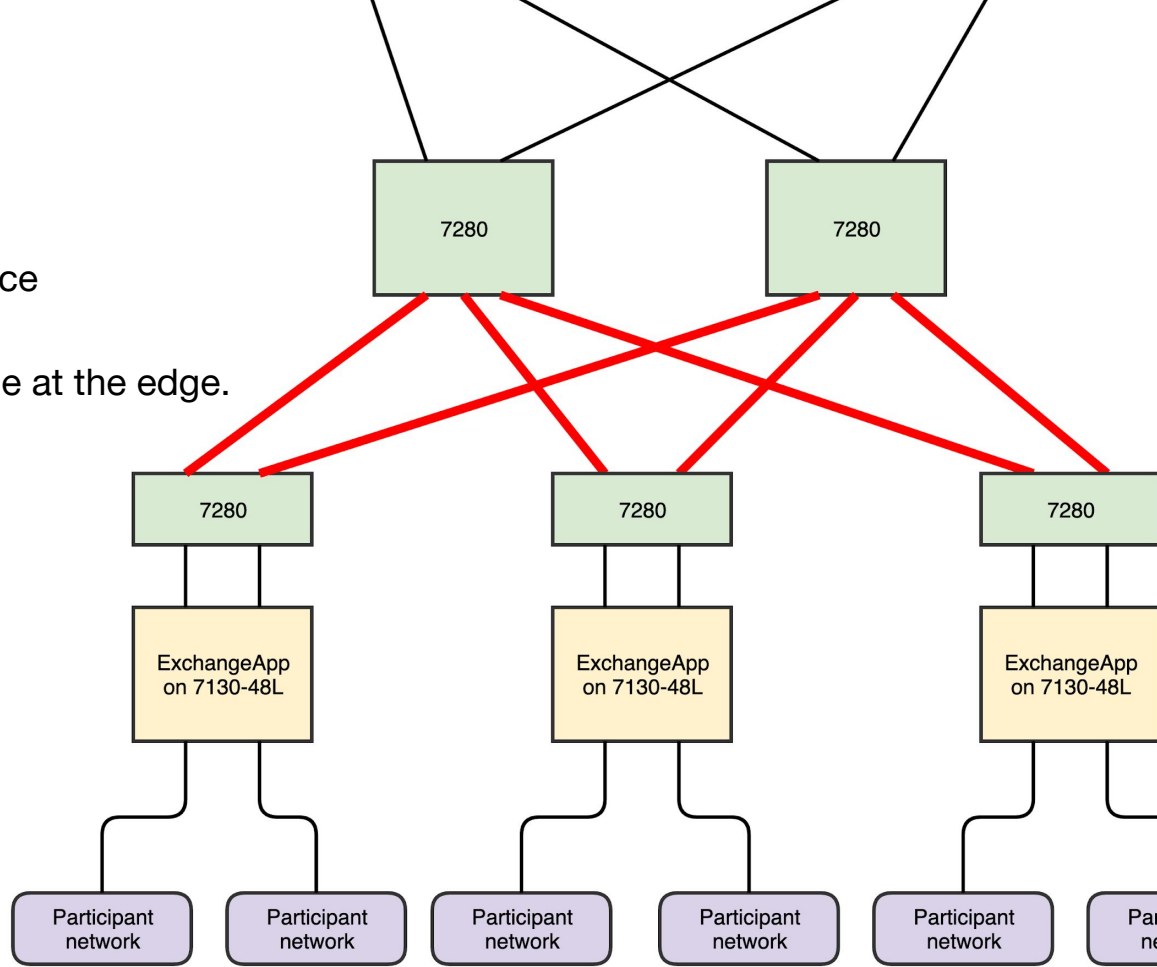


What can we do?

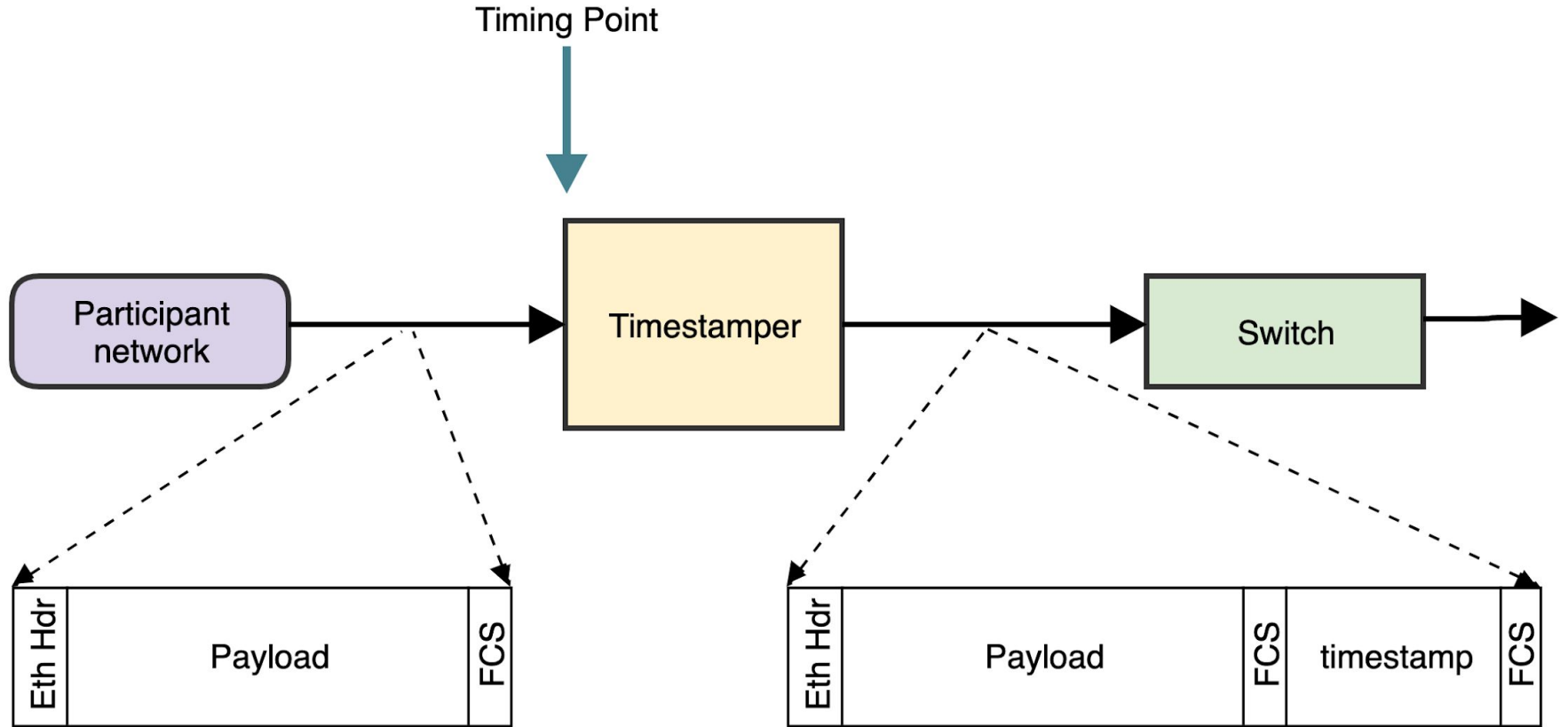
- Step 2: Solve what can be solved.

What can we do?

- Step 3: Change the paradigm.
 - Make time as explicit as price
 - Measure event timing
- Arista's solution: timestamp in-line at the edge.



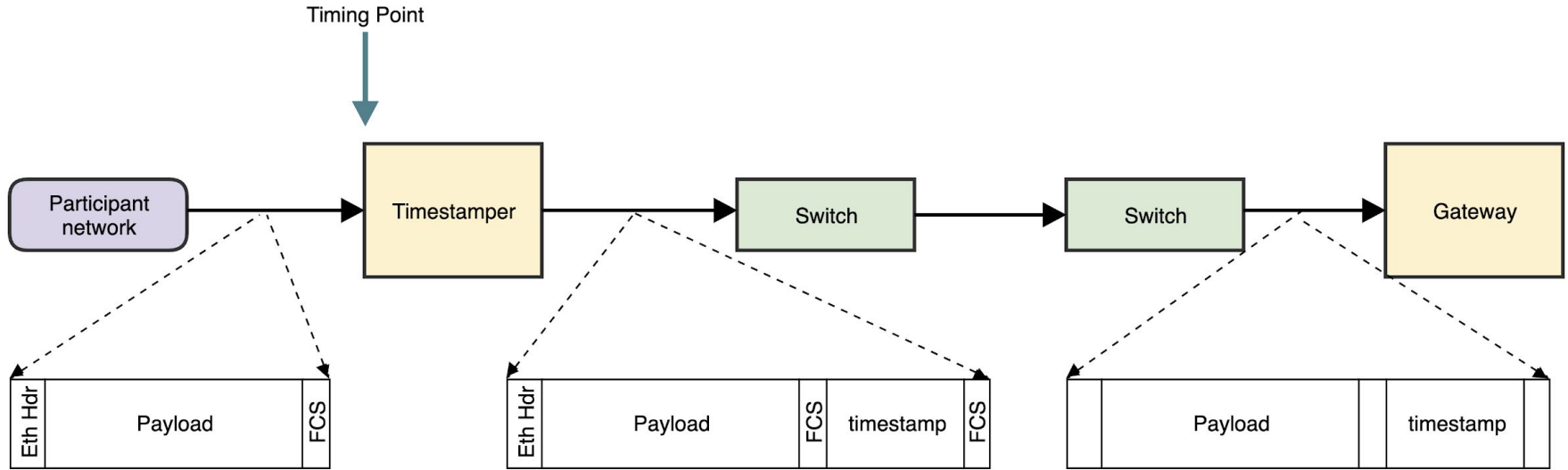
Timestamping at the edge...



ExchangeApp

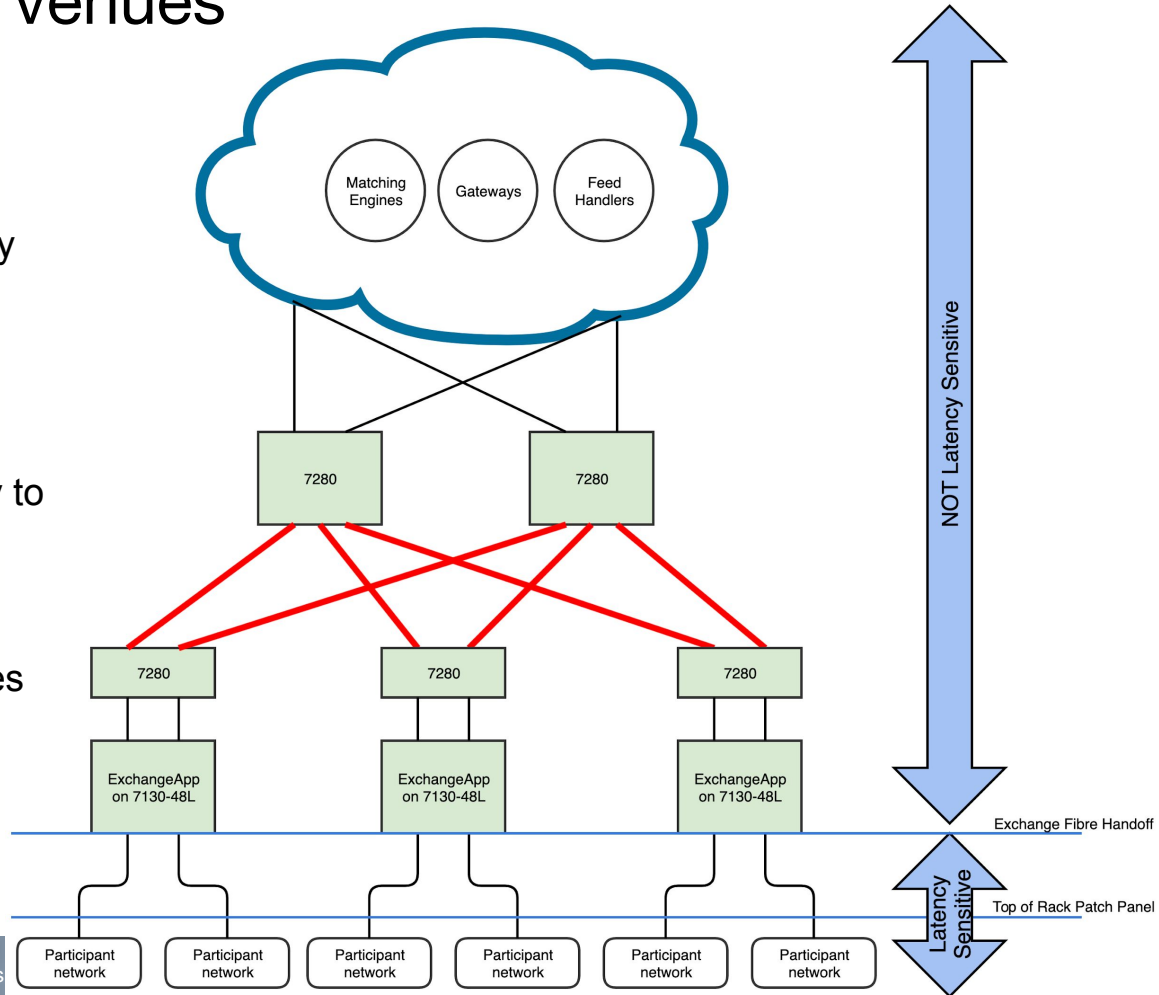
- It turns out that switches only modify packet headers, so trailers make it to the gateway, even with an L3+ network.
- Venue operators now don't have to worry about latency. This is nice!
 - Reduce contention using store/forward, fat core pipes.
 - Reduce loss using deep buffer switches, dynamic routing.
 - Use multiple gateways, switches, etc.
- The next challenge is how to deliver responses. Watch this space.

Timestamping at the edge...



A better way to build venues

- Enabled by reliable accuracy.
- Timestamp at the edge.
- Stop worrying about time/latency in the venue implementation.
- Reason about time in software instead.
- Use jittery-but-better technology to implement your venue:
 - Deep buffers
 - Store-and-forward switches
 - VMs/Containers
 - Cloud
 - Garbage collection



Implications

- Venue operators don't have to be quite so paranoid about latency, because we already have the determinism.
- That means we can build a better venue -- focus on avoiding queuing and improving reliability by building the network *right*.
- Software can be implemented using best practices, without restriction.
- Fibre matching can be achieved using fixed timestamp offsets.
- Will sub-nano determinism result in a new arms race?
 - Re-ordering based on time-of-arrival is only one policy.
- I can't currently work out how to game this without a good workaround.
 - Let me know if you can.

For participants

- Will this result in an accuracy arms race?
 - Winner-takes-all is probably not in anyone's interest.
 - Probably not -- executing in order-of-timestamp enables randomisation, discrete auctions, or other policies.
- Could accurate markets mean the end of co-location?
 - Probably not.
- Are there better policies that we could implement?
 - Almost certainly.

The markets aren't perfect. And technology can help.



Thank You

www.arista.com