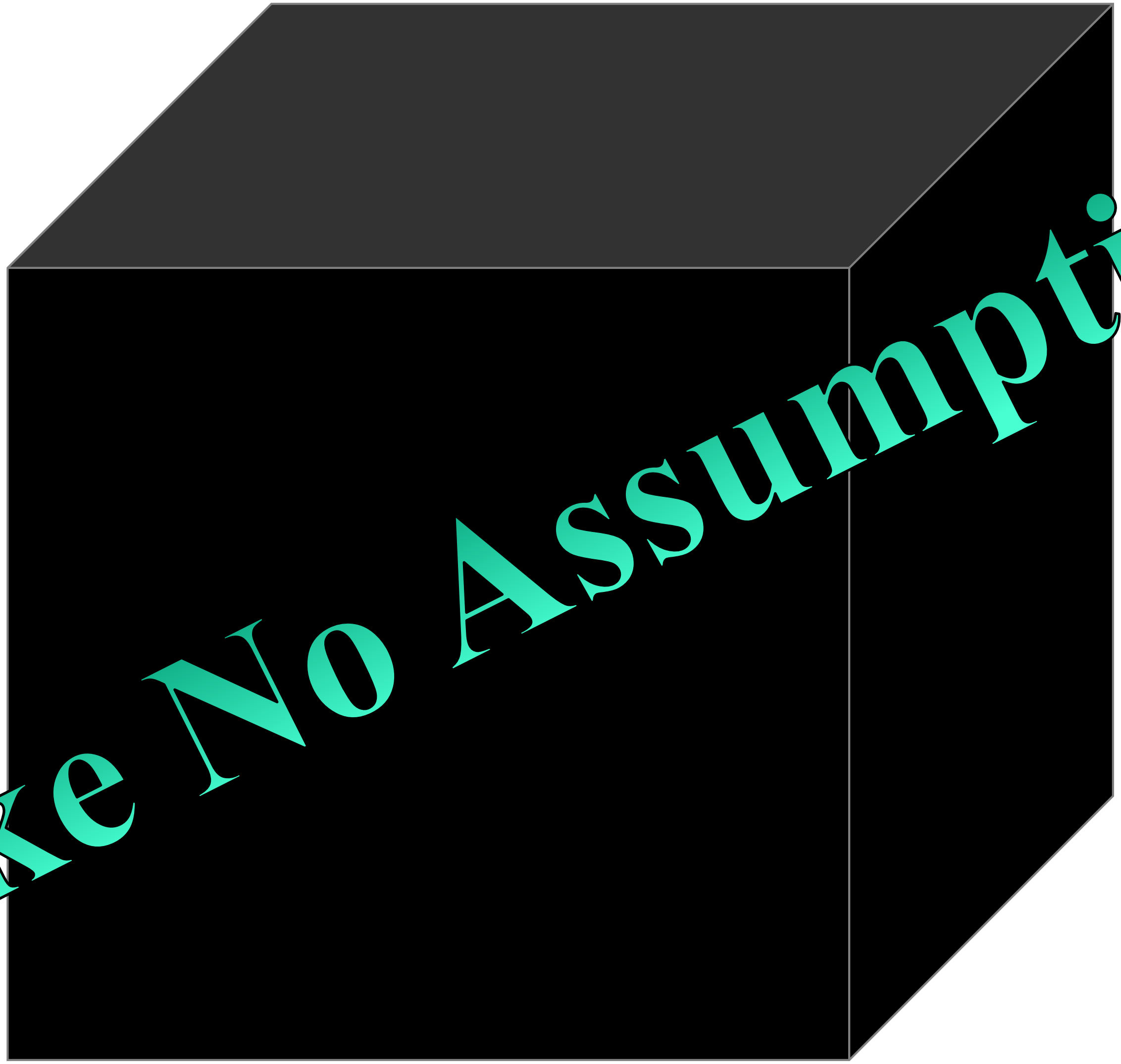
Introduction to Networking and Systems Measurements

Device and System Characterization

Andrew W. Moore

andrew.moore@cl.cam.ac.uk



Make No Assumptions

What is the goal?

- Functional validation?
- Performance testing?
- Characterization?
- Comparison?
- Detecting problems?
- Finding the bottlenecks?

Different goals $\Rightarrow$ different setup + experiments

What is the goal?

- Functional validation, e.g.,:
  - Can we send traffic from port A to port B?
- Performance testing, e.g.,:
  - What is the throughput of sending traffic from port A to port B?

OR

Vantage Points

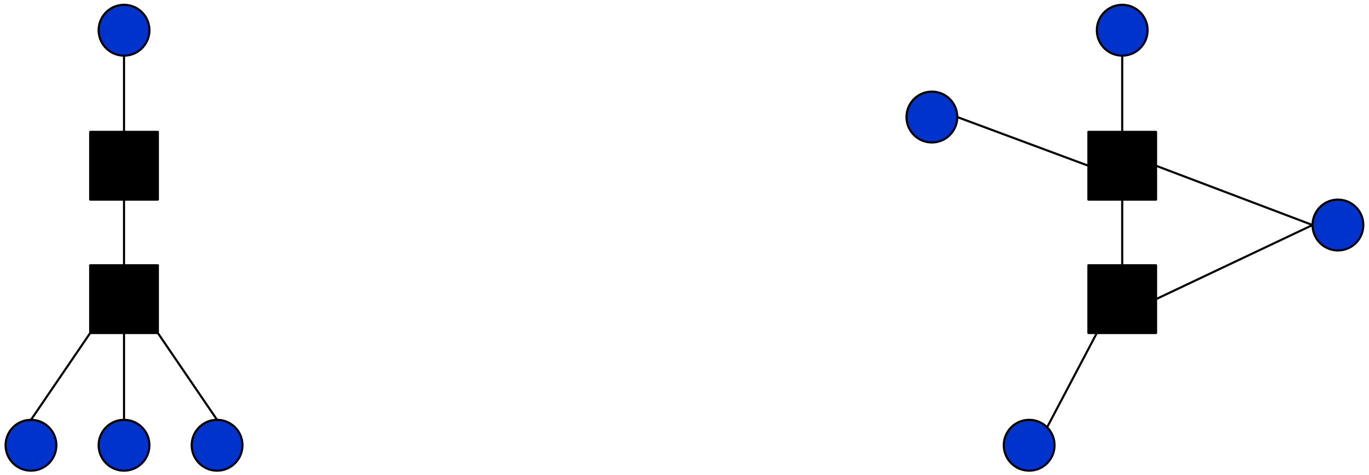
- Characterisation is limited by vantage points
- Single vantage point:
  - Round trip measurements, topology measurements

  OR

  - Passive measurements
- Two vantage points:
  - One way latency measurements, bandwidth measurements + everything a single vantage point can do
- Three vantage points?

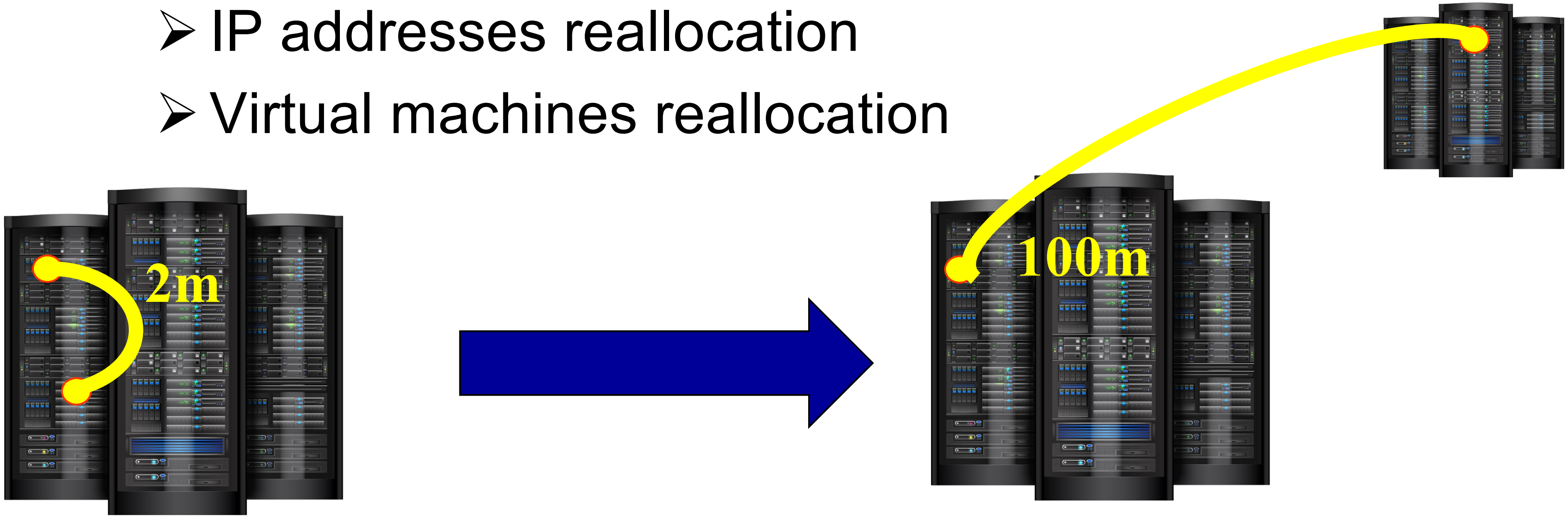
Vantage Points

- <Number> of vantage points is not sufficient
- <Location> of vantage points is important



Vantage Points

- Is your vantage point static?
- Mobile vantage points: Mobile phones, laptops
  - Sometimes good if you seek to increase coverage
- But also (for example):
  - IP addresses reallocation
  - Virtual machines reallocation



What is the workload?

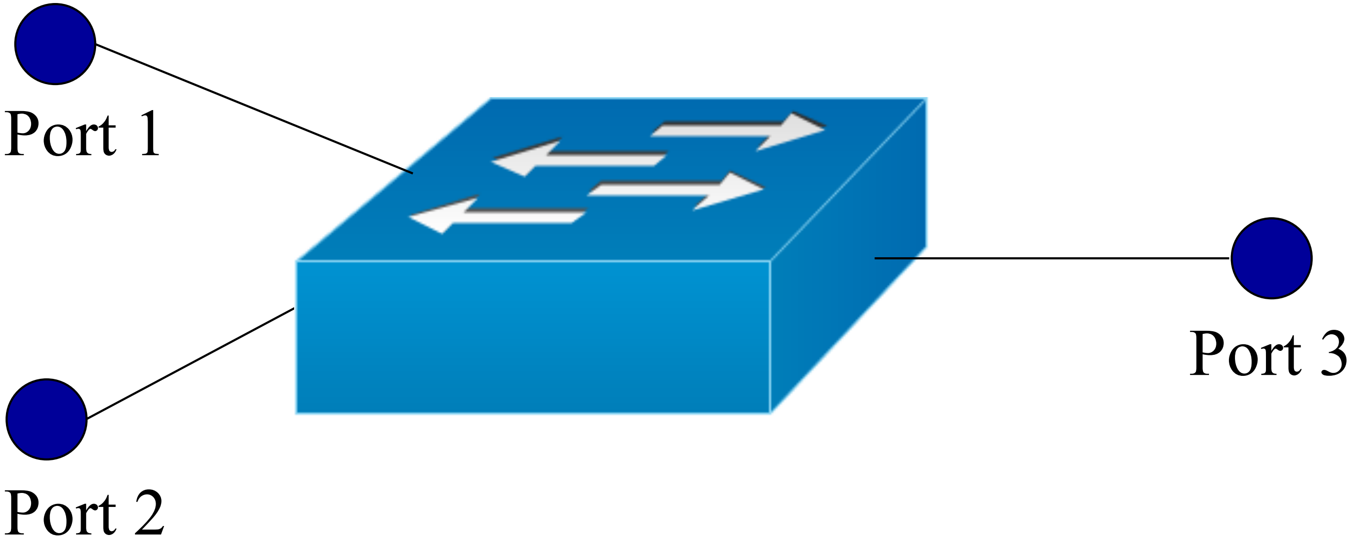
- Synthetically generated, e.g.,
  - 128Byte IPv4 Packets
- Protocol level, e.g.,
  - TCP flows
- Application level, e.g.,
  - Key-value store application

What is the workload?

- Everything matters!
- Packet size distribution
- Traffic rate
  - E.g., Average rate, peak rate,
- Traffic shape
  - E.g. bursts
- Payload
  - Some payloads are more likely to cause errors than others
- Protocol
- ….

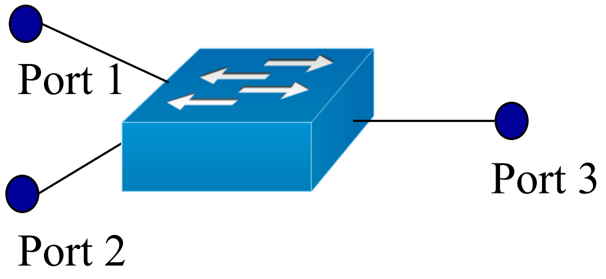
Example

- What can we learn about the internals of a switch using latency measurements and 3 vantage points?
- Assuming a sterile environment



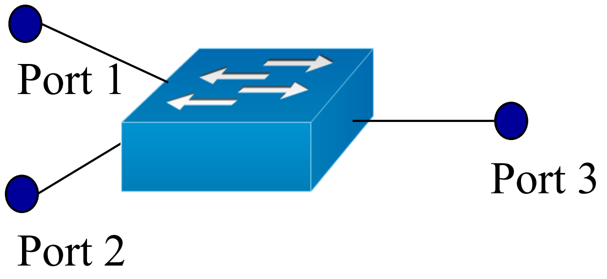
Example

- What is the basic latency of the switch?
  - Send packets from port 1 to port 2, measure the latency
- Is the switch design symmetric?
  - Send packets from port 2 to port 1, measure the latency
- Is the switch design identical for all ports?
  - Send packets from port X to port Y, measure the latency for all combinations



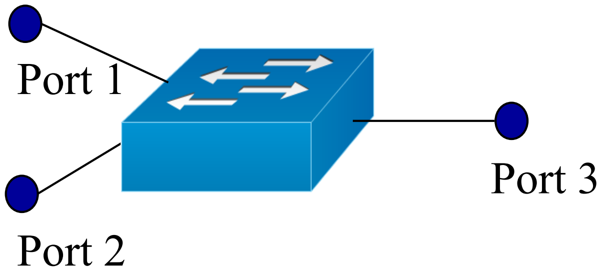
Example

- What type of switch is it?
  - Send packets of various sizes from port 1 to port 2, measure the latency
  - A cut-through switch will have the same latency for all packet sizes, a store-and-forward switch will have a higher latency for bigger packet sizes
- Is the switch sensitive to throughput?
  - Send packets at full line rate from port 1 to port 2, measure the latency
  - Do the results change over time?



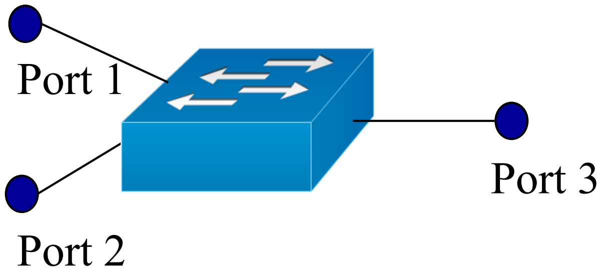
Example

- What can we learn about the output queueing and output scheduling of the switch?
  - ➢ Send packets at port 1 to port 3, measure the latency
    And at the same time
  - ➢ Send packets at port 2 to port 3, measure the latency
  - ➢ Vary the packet rate and discover more….



Example

- What can we learn about the input queueing and input scheduling of the switch?
  - ➢ Send packets at port 1 to port 3, measure the latency
    And at the same time
  - ➢ Send packets at port 2 to port 4
  - ➢ Vary the packet rate and discover more….
  - ➢ Why is sending from port 2 to port 1 a bad idea?

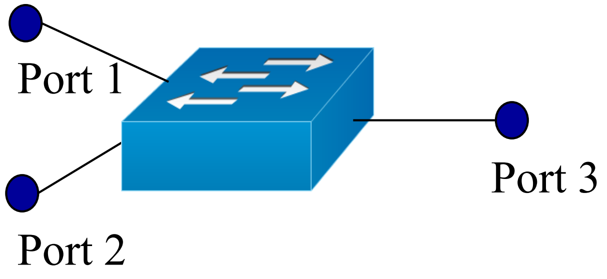


Example

- So….

  What can we learn about the internals of a switch using latency measurements and 3 vantage points?

- A lot!
- This was just a small subset



Example 2

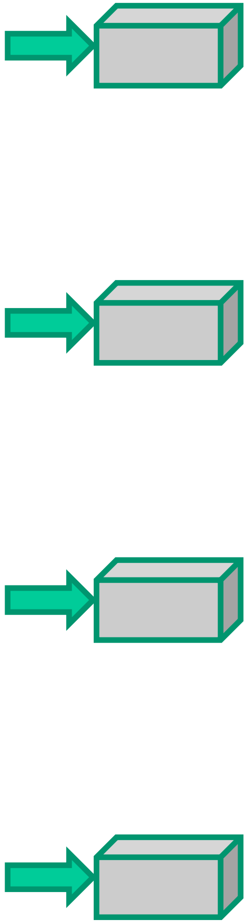
- Mellanox Spectrum vs Broadcom Tomahawk
  - ➢ Tolly report, 2016
    Accessible from the L50 main webpage
- Bandwidth distribution, 3→1 scenario
  - ➢ Source ports 25,26,27, Destination port 31
    33% BW from each port, on both devices
  - ➢ Source ports 24,25,26, Destination port 31
    33% BW from each port, on Spectrum
    25% from ports 25,26, **50**% from port 24 on Tomahawk
- What does it mean?

Switch refresher

Switch Internals 101

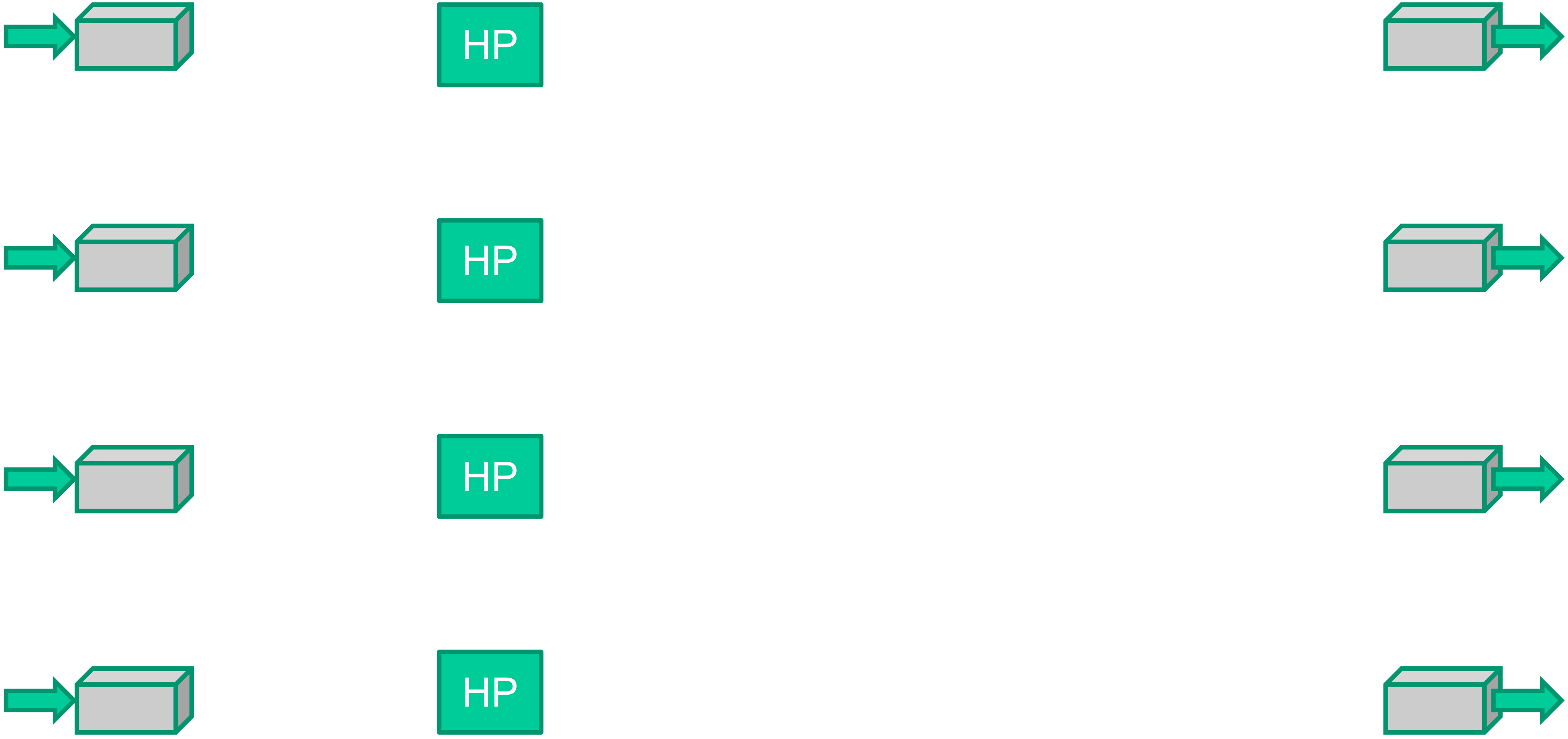
What defines the architecture of a switch?

Input Ports

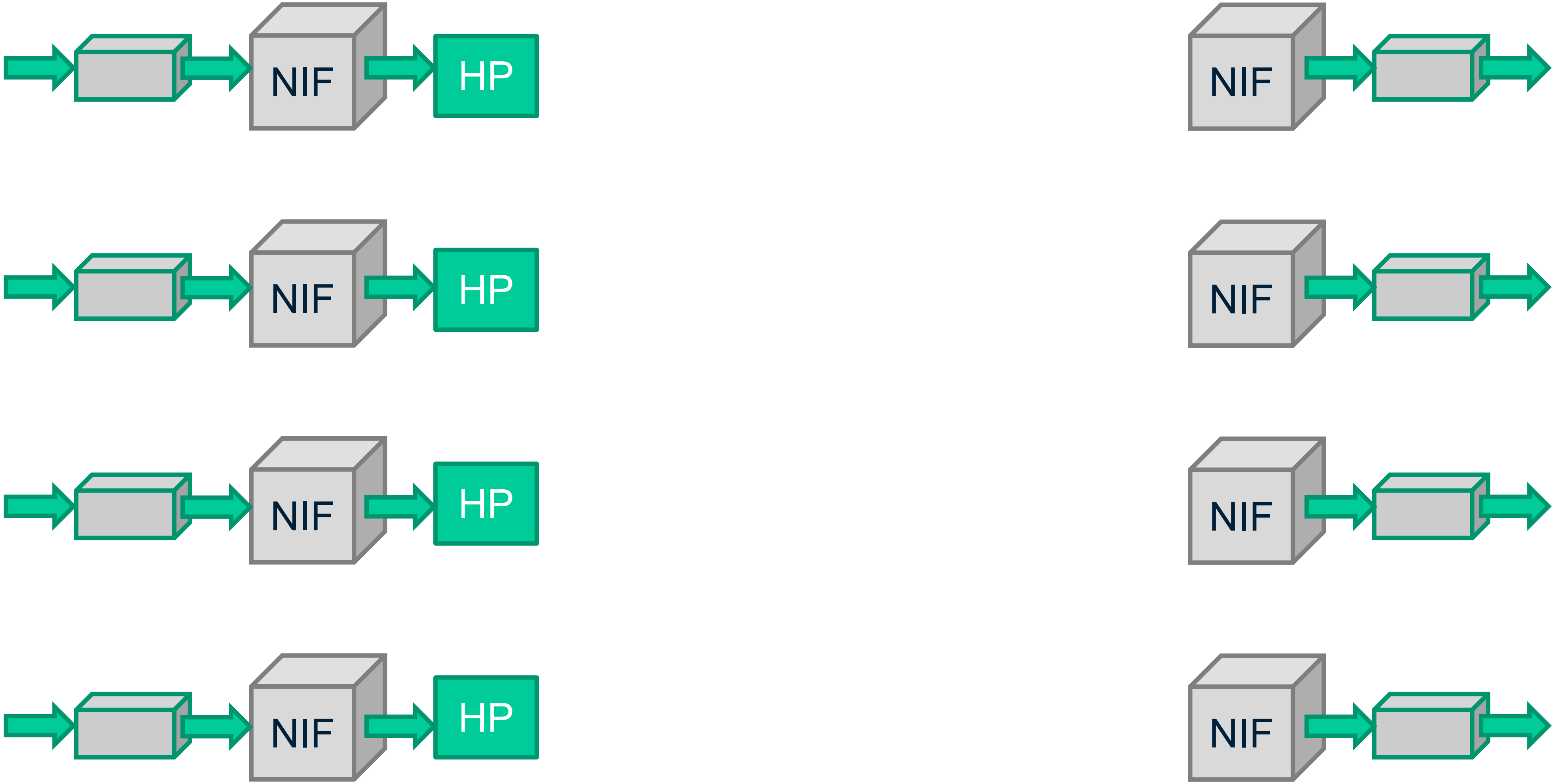


Output Ports

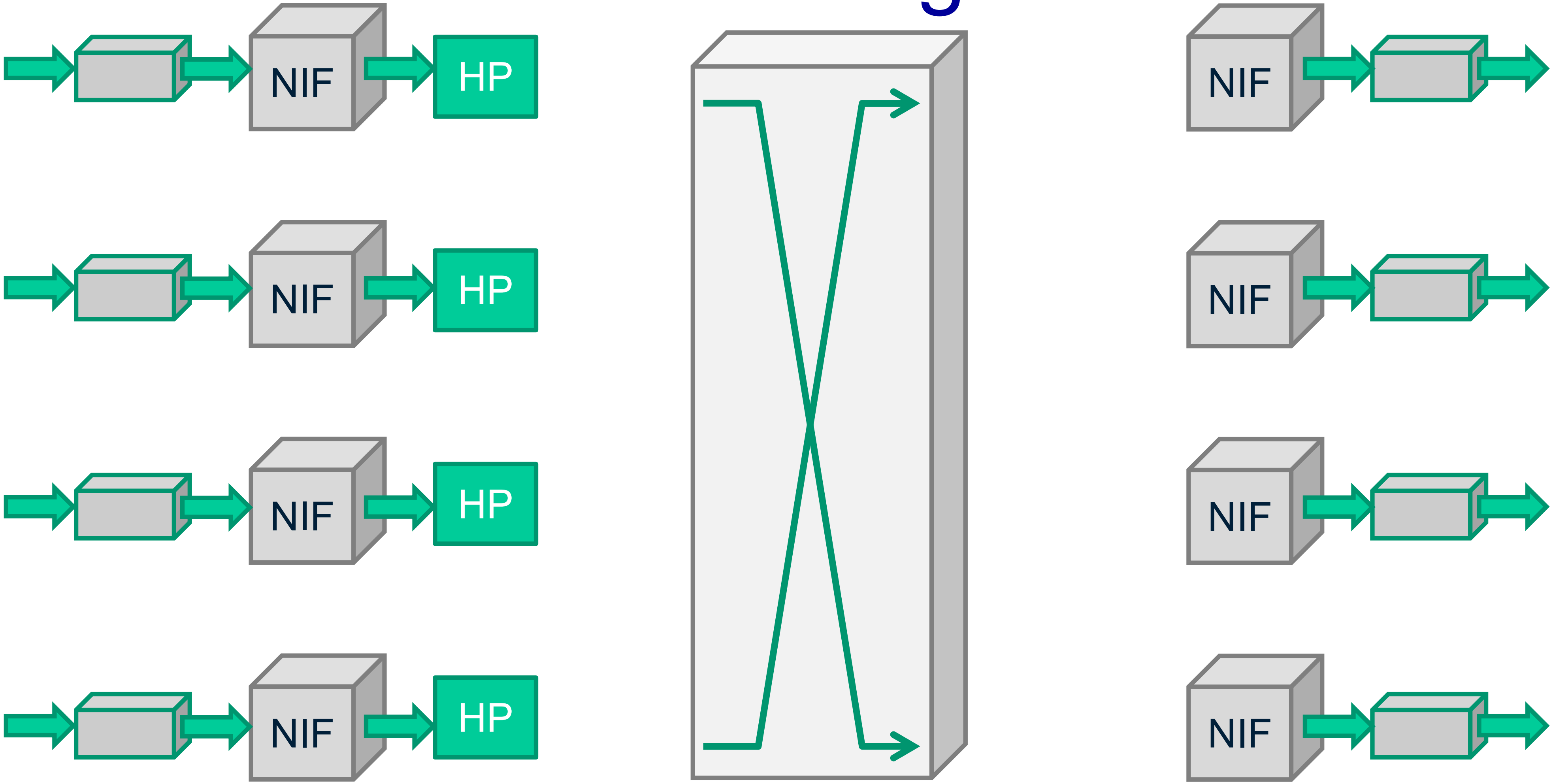
Header Processing



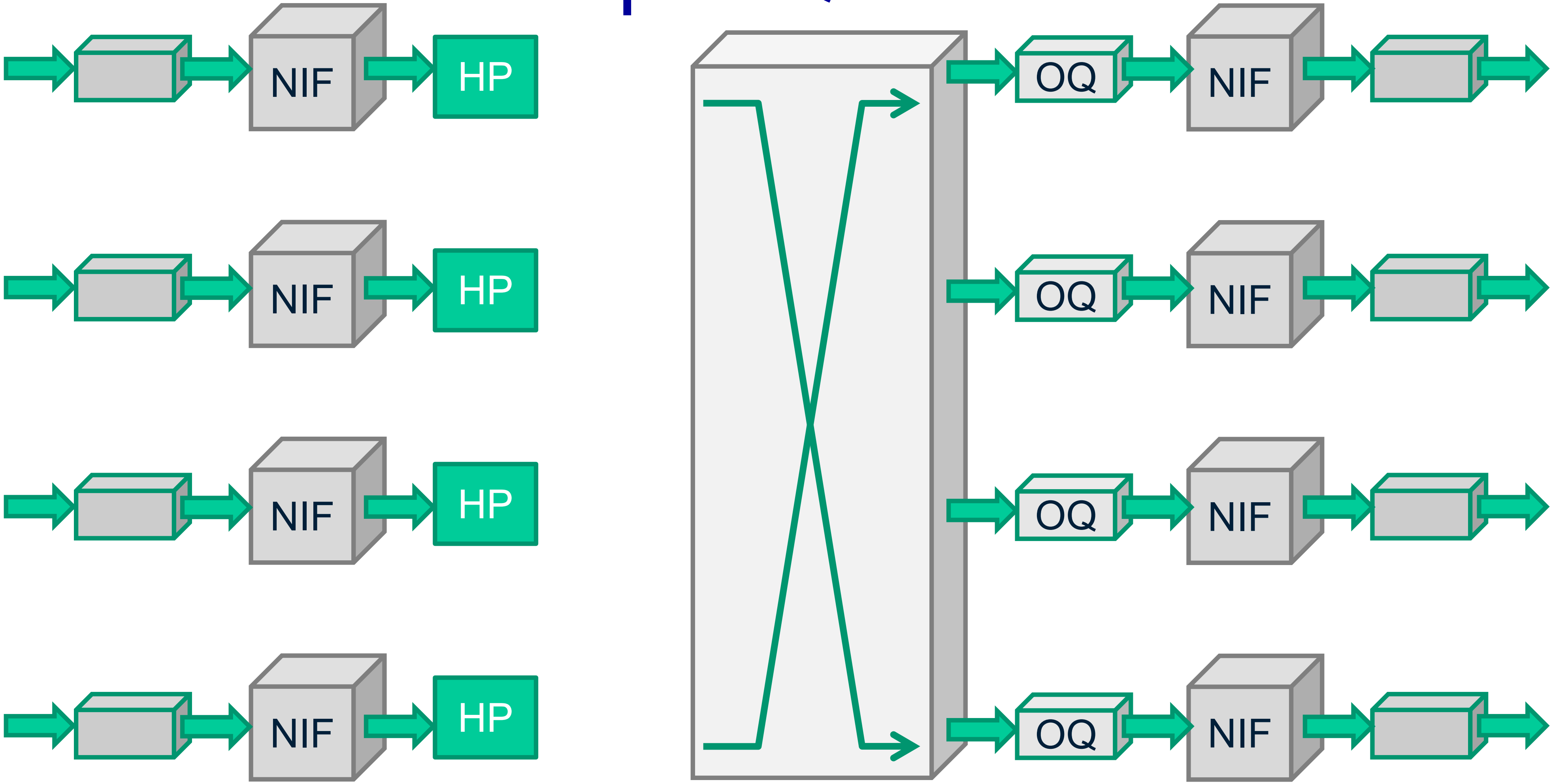
Network Interfaces



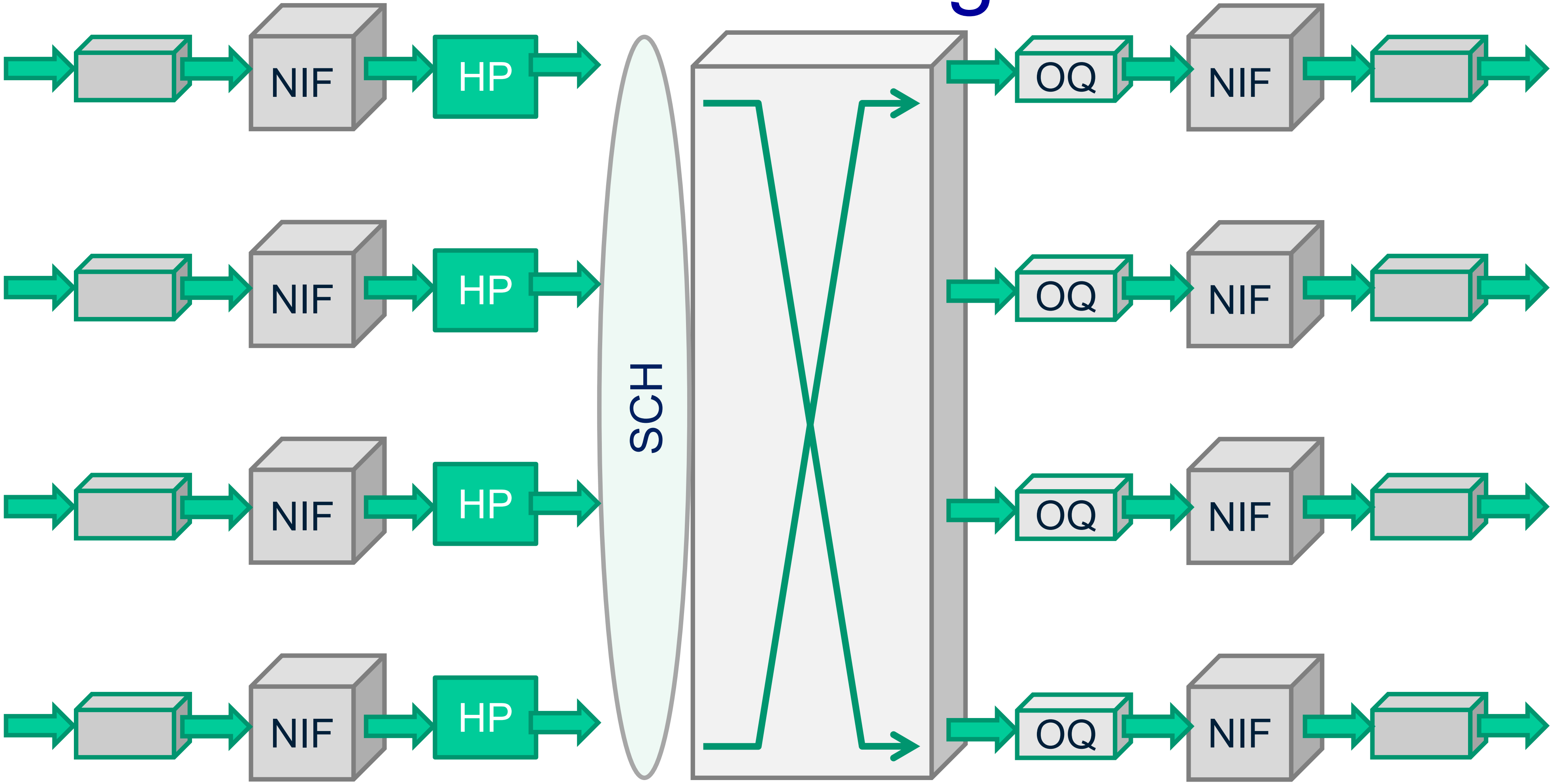
Switching



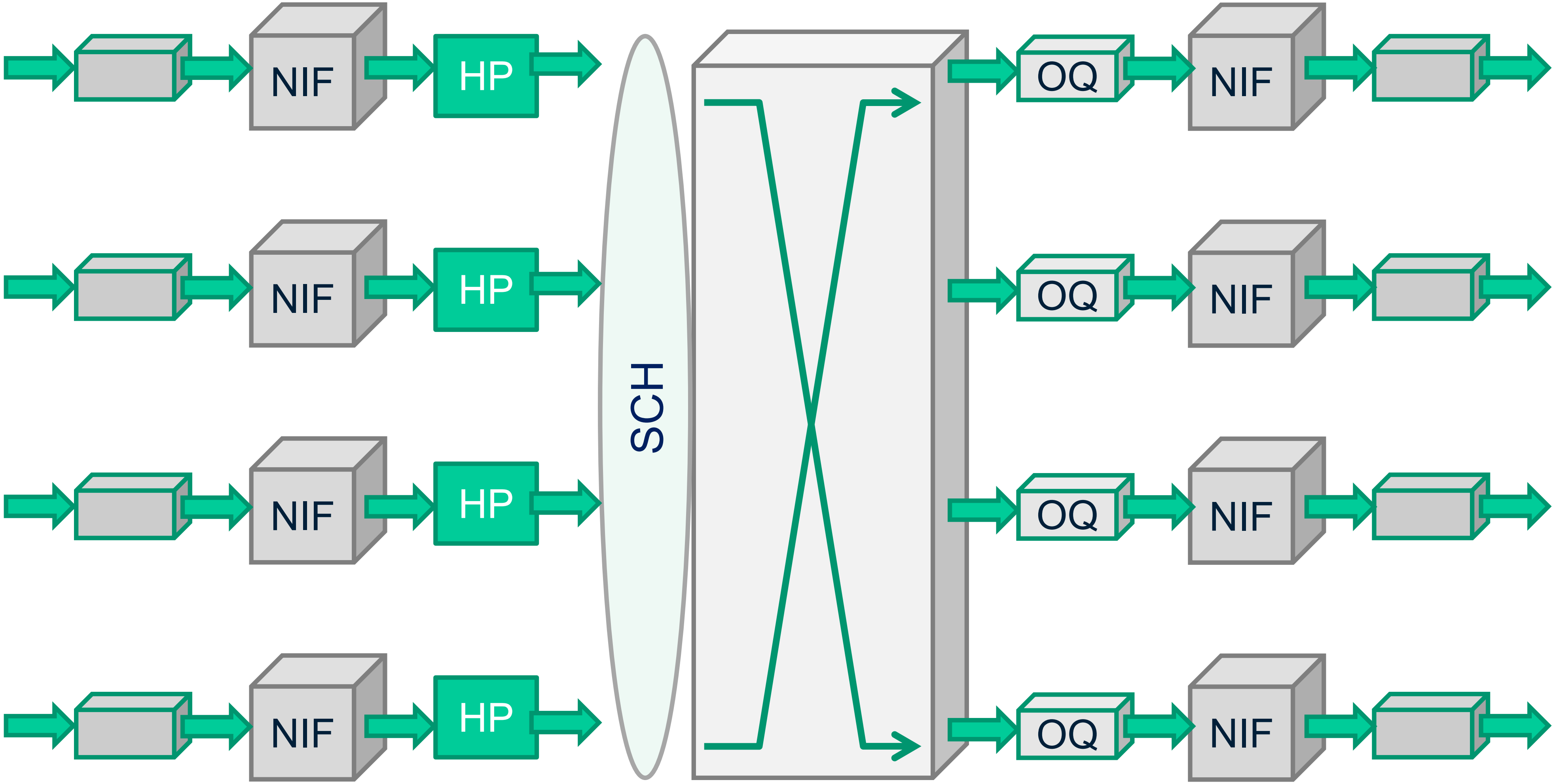
Output Queues



Scheduling



Is This A Real Switch?



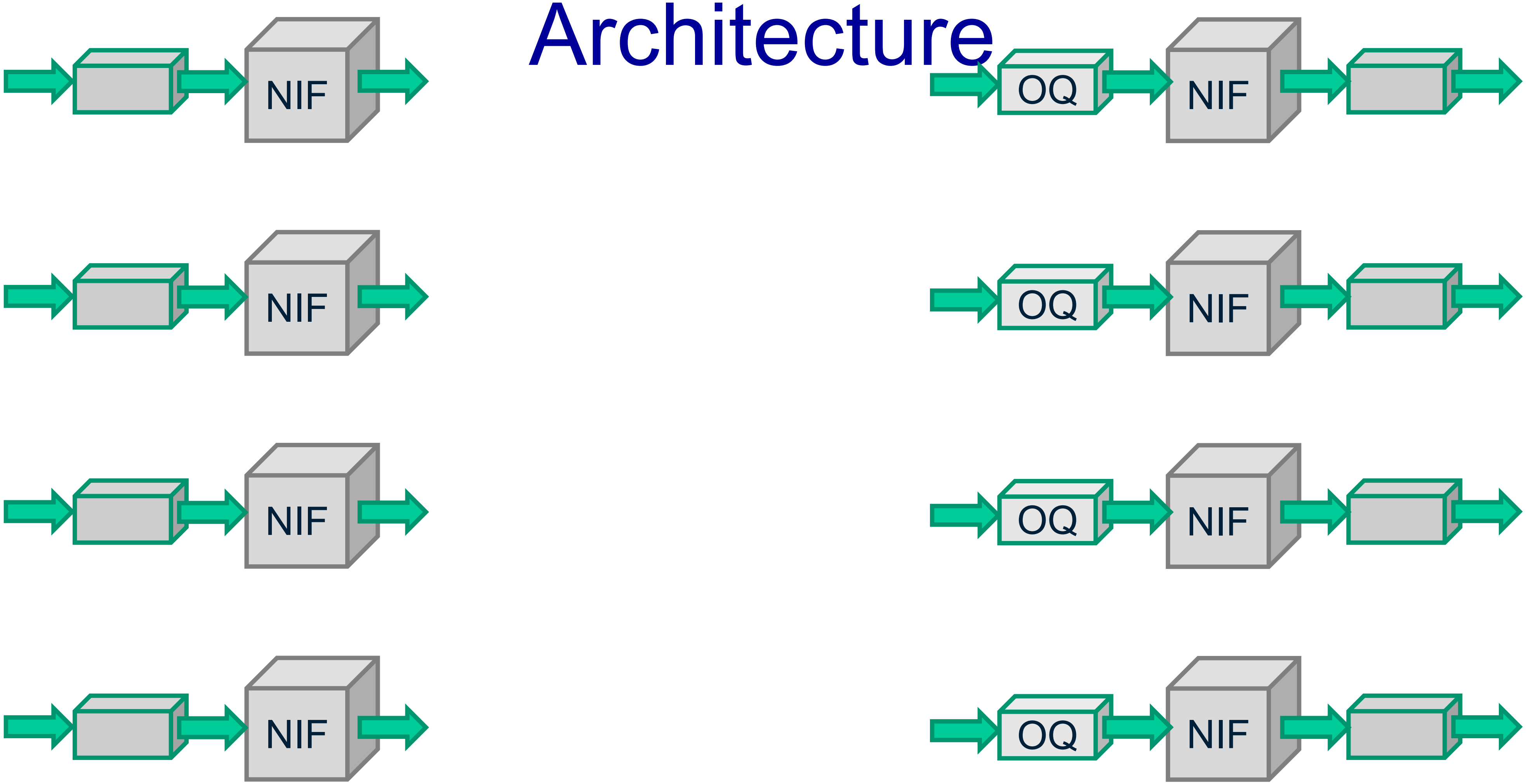
Recall What Drives Real World Switches

- Cost
- Power
- Area

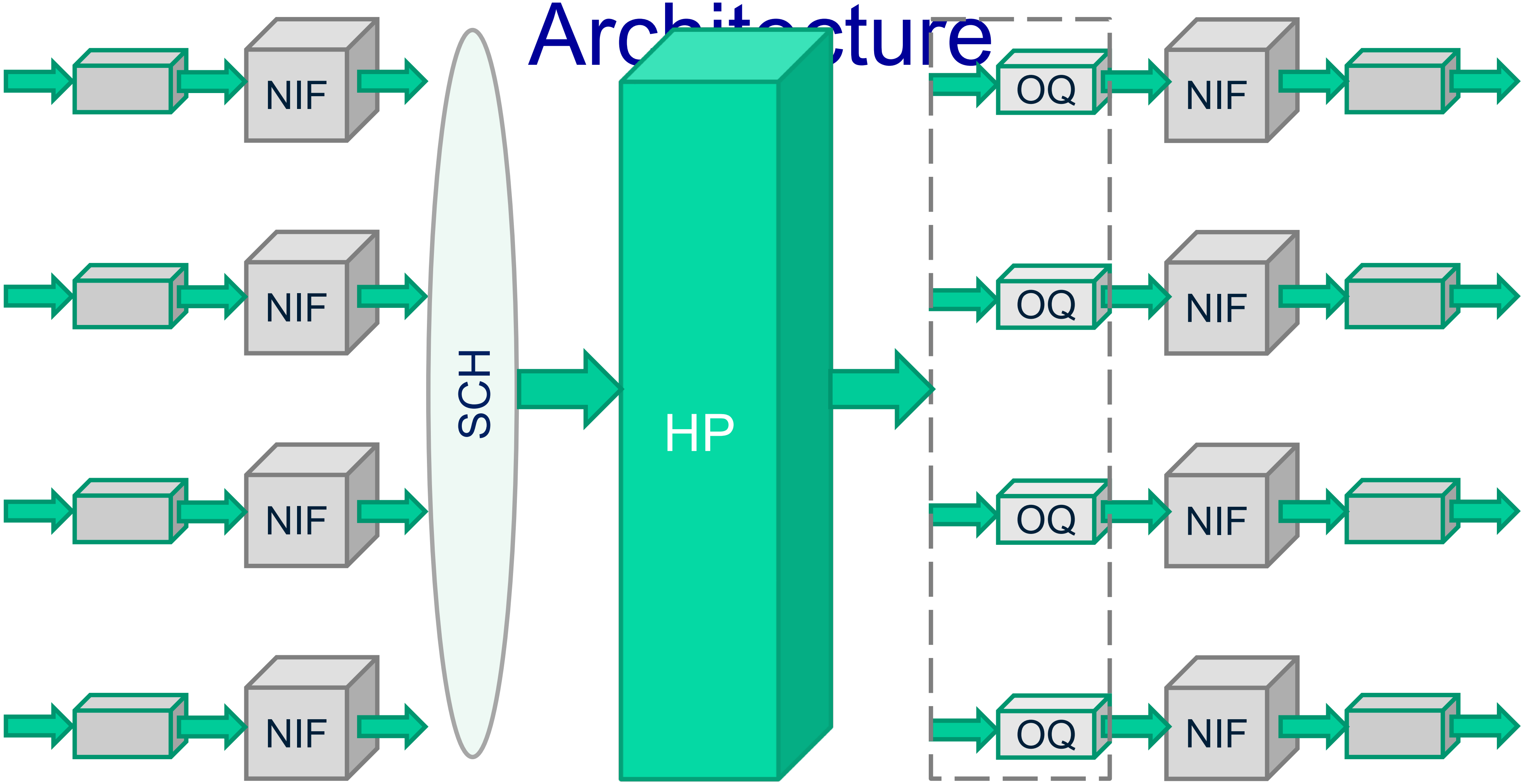
Sharing Resources Is Good!

- Single header processor (if possible)
- Shared memories
- No concurrency problems
  - ➢ Also no need to synchronise tables, no need to send updates, ….

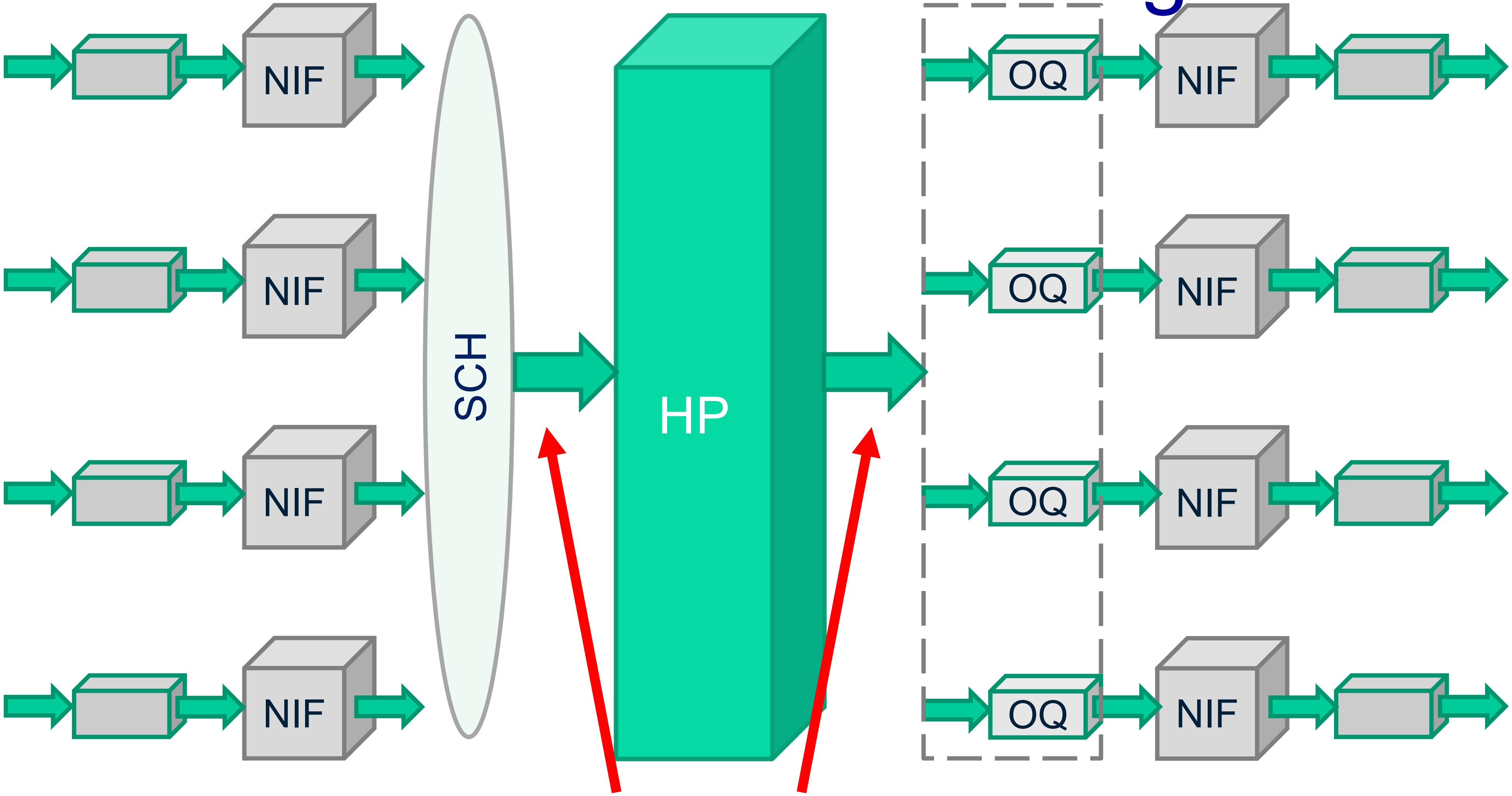
Rethinking The Switch Architecture



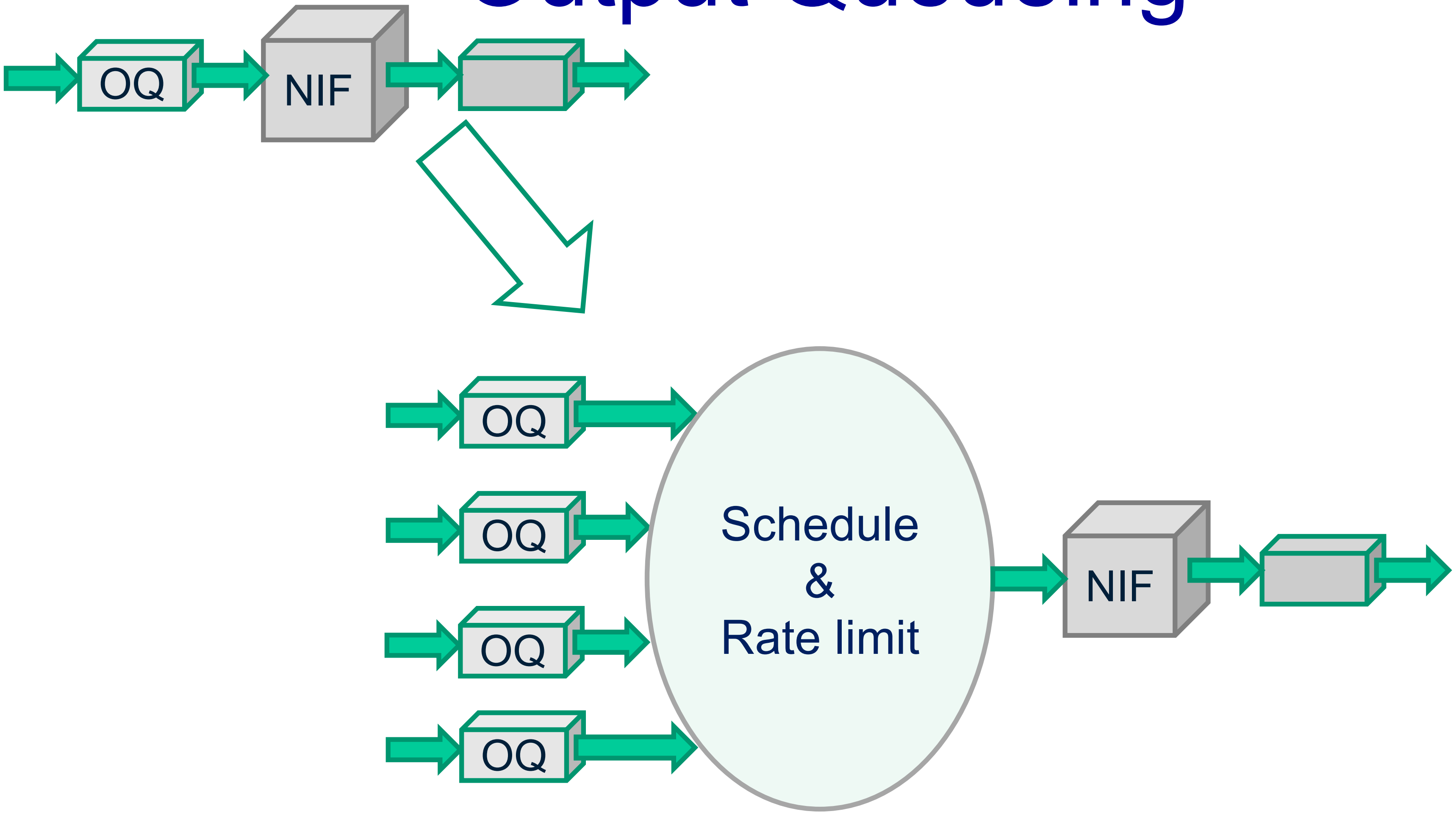
Rethinking The Switch Architecture



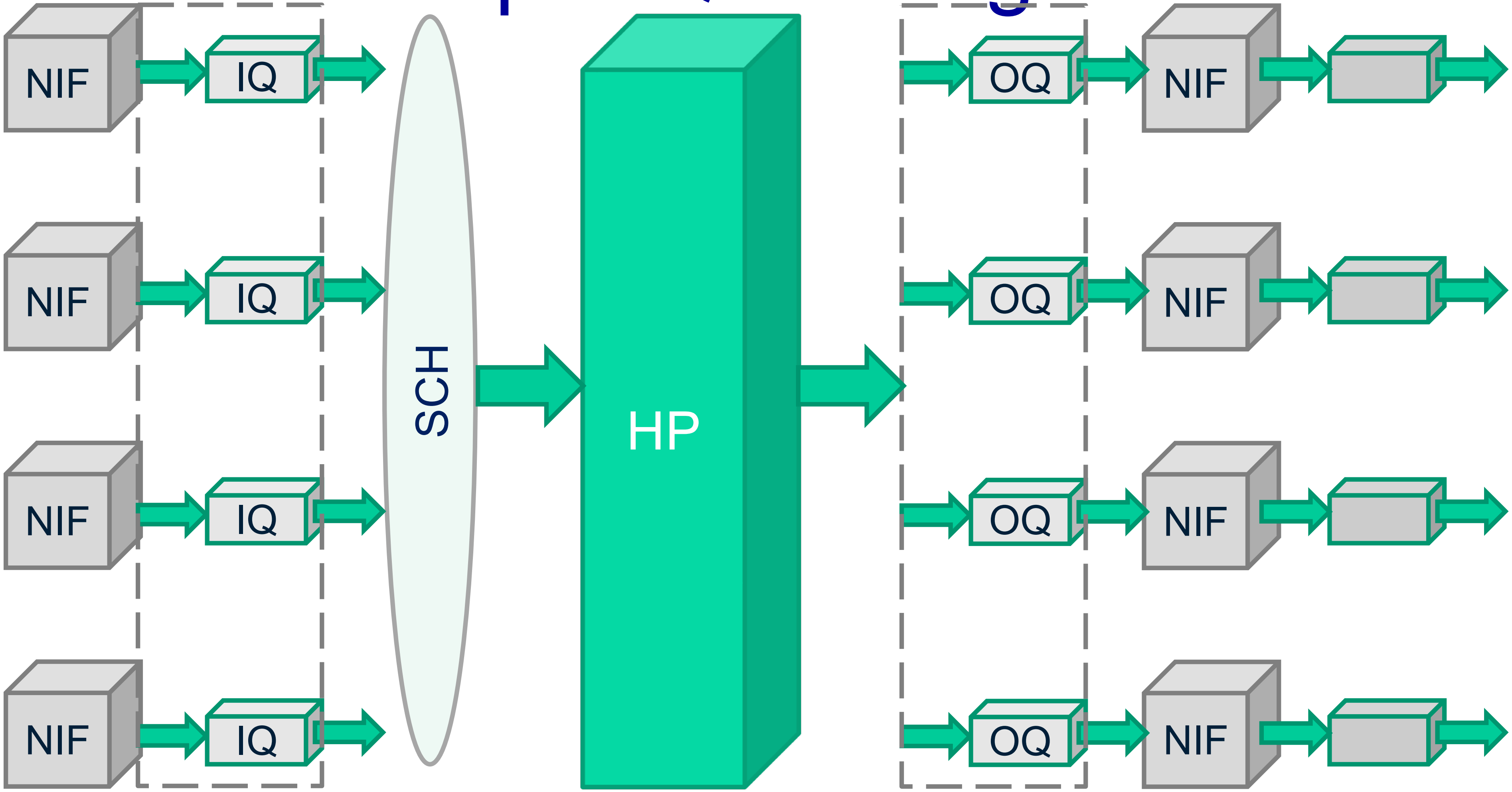
Where Is The Switching?



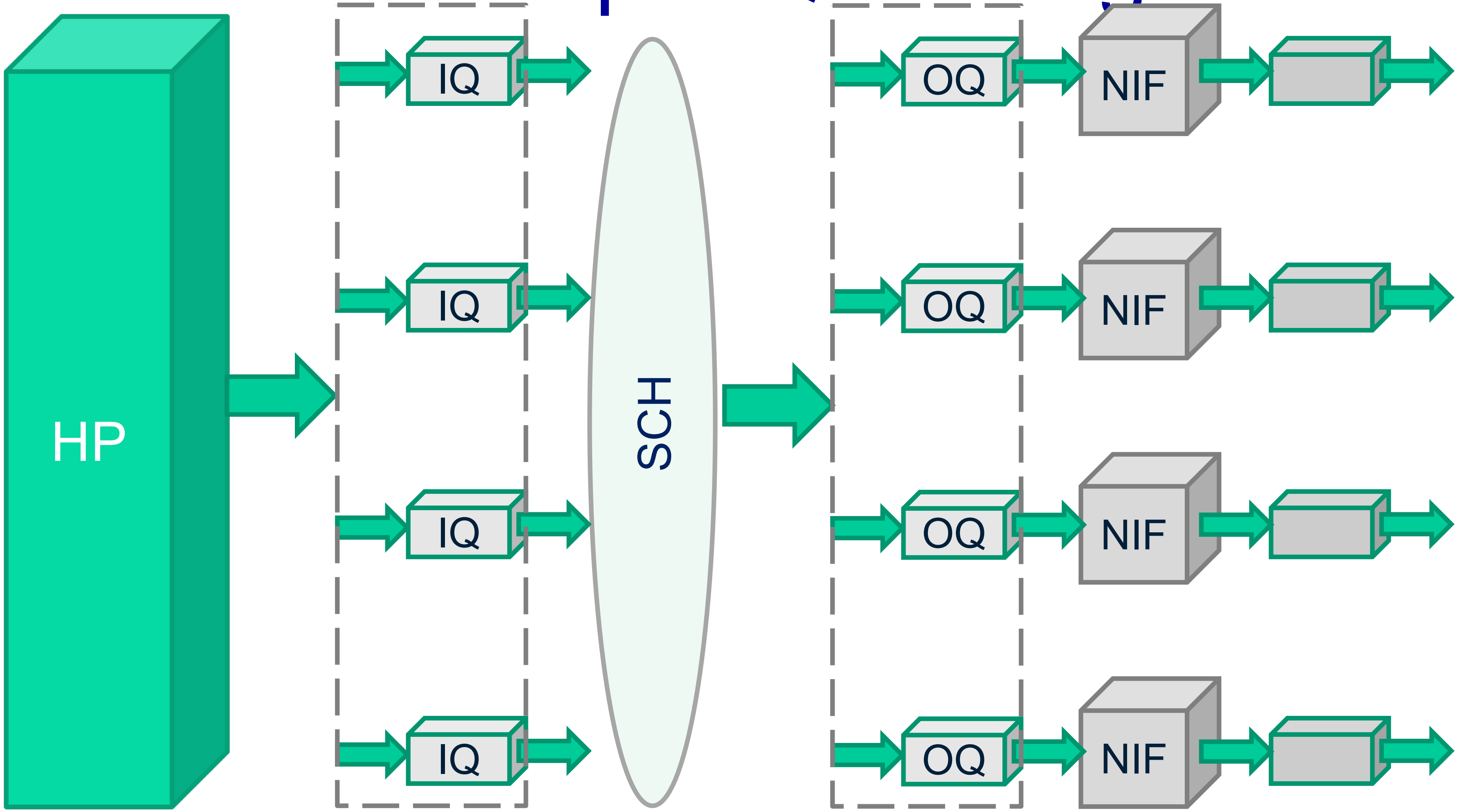
Output Queueing



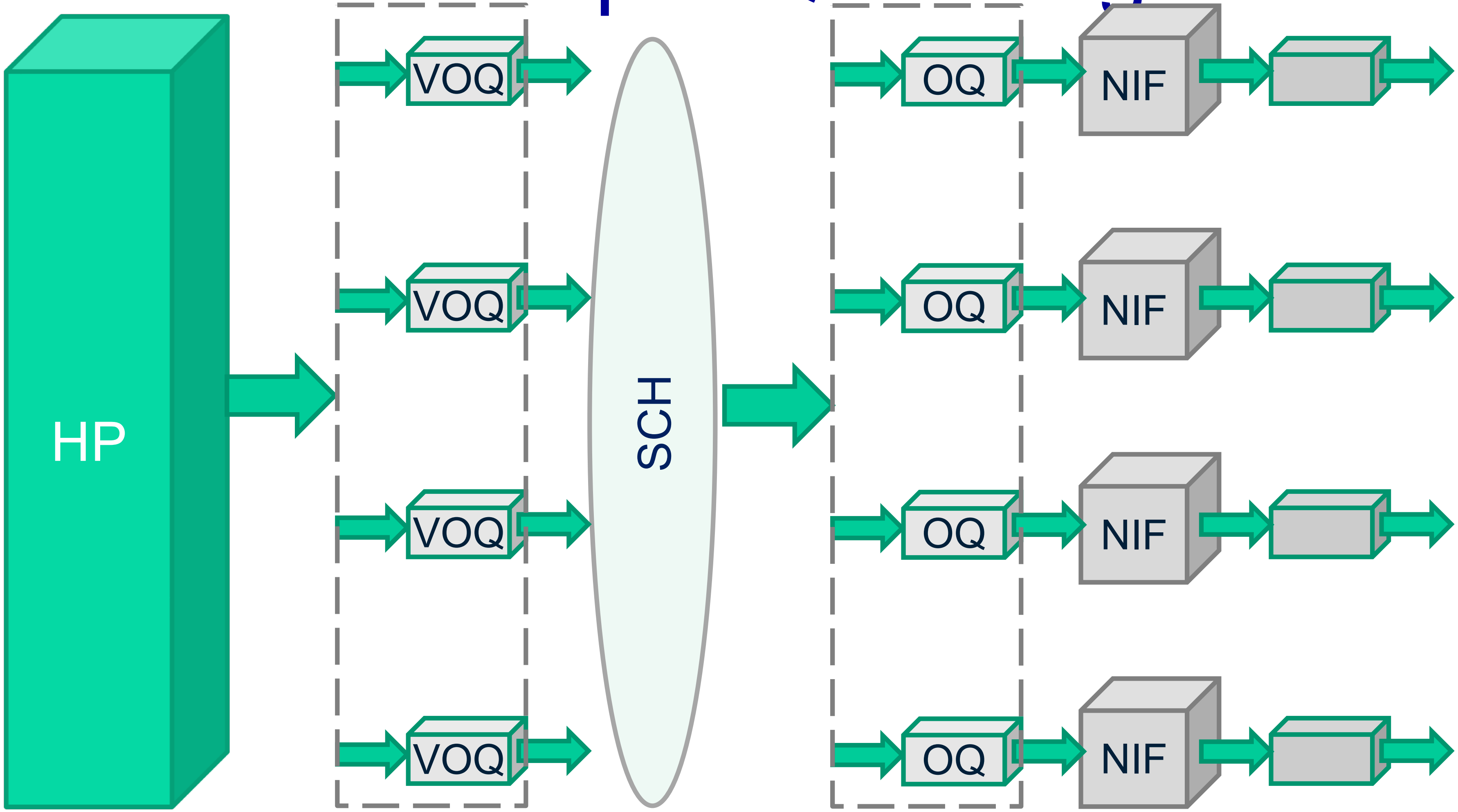
Input Queueing



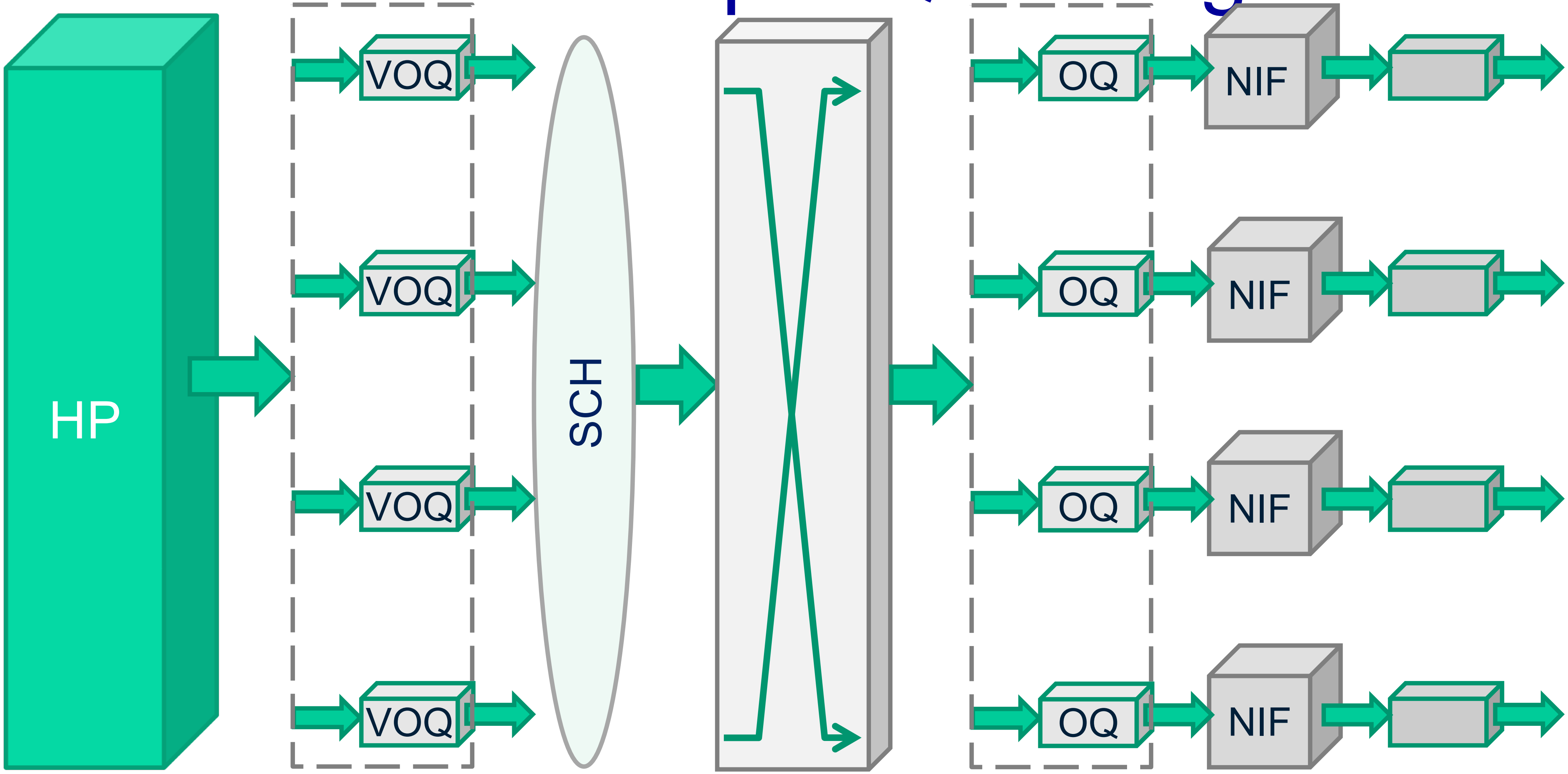
Virtual Output Queueing



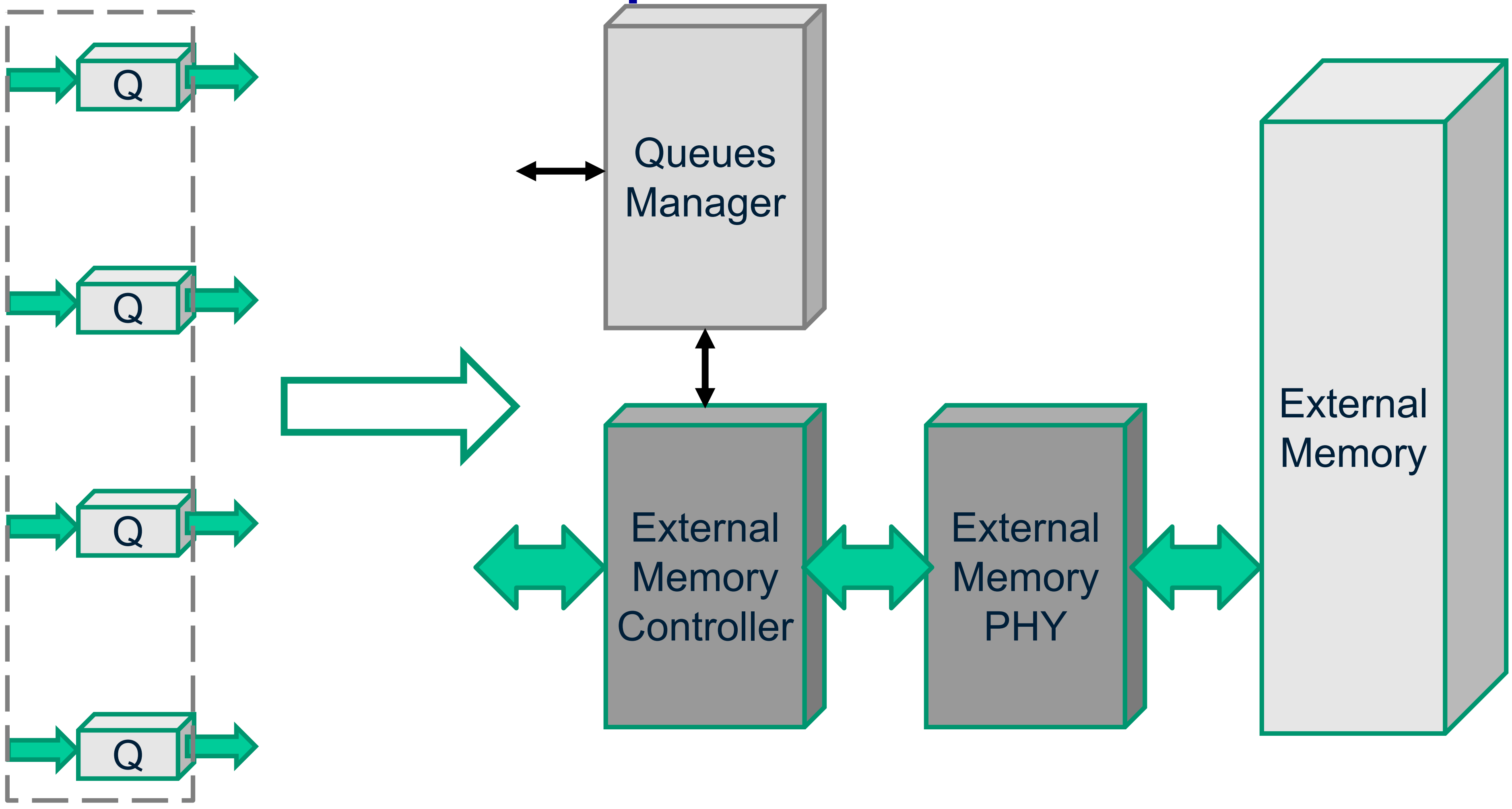
Virtual Output Queueing



Virtual Output Queueing



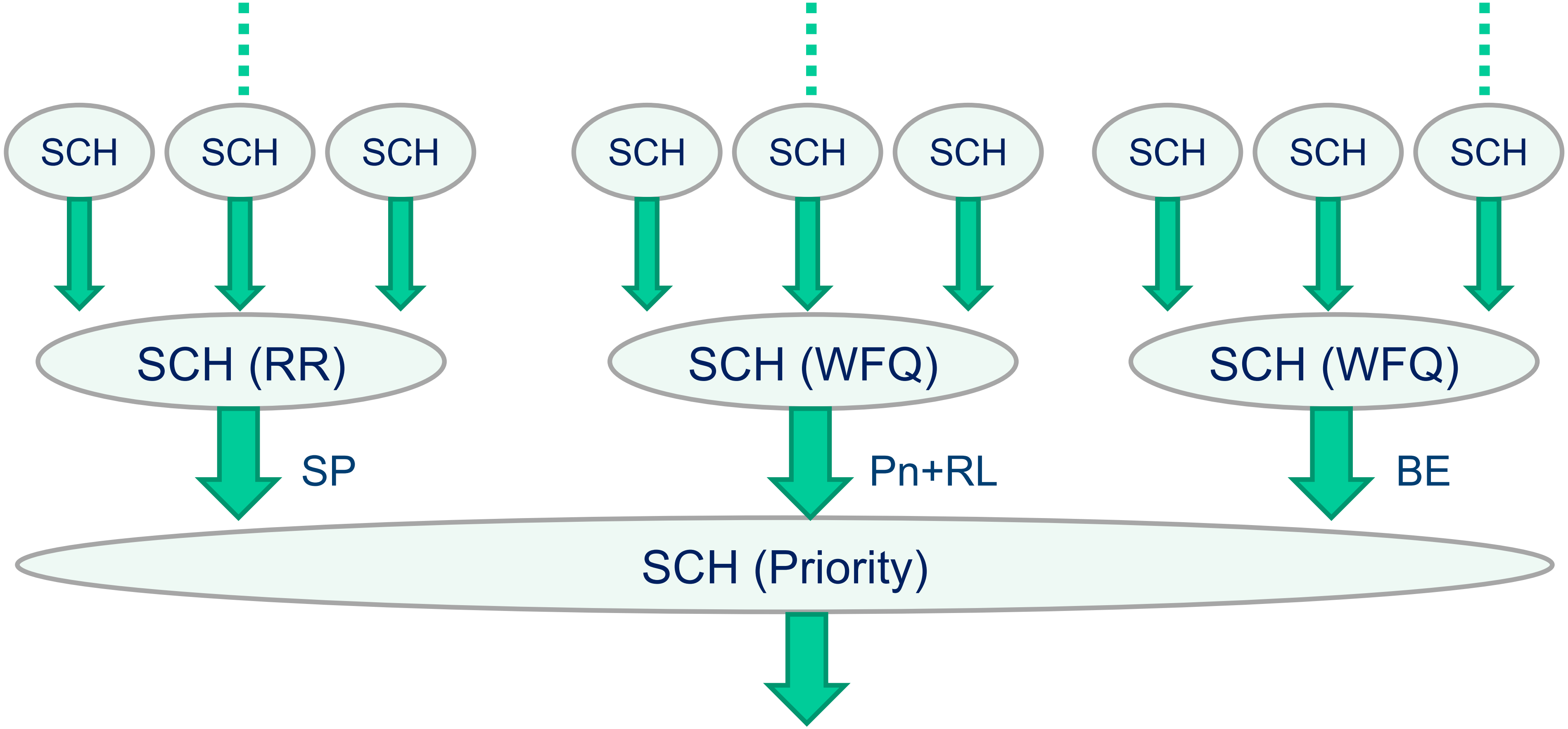
Deep Buffers



Scheduling

- Different operations within the switch:
  - Arbitration
  - Scheduling
  - Rate limiting
  - Shaping
  - Policing
- Many different scheduling algorithms
  - Strict priority, Round robin, weighted round robin, deficit round robin, weighted fair queueing…

Scheduling Hierarchies



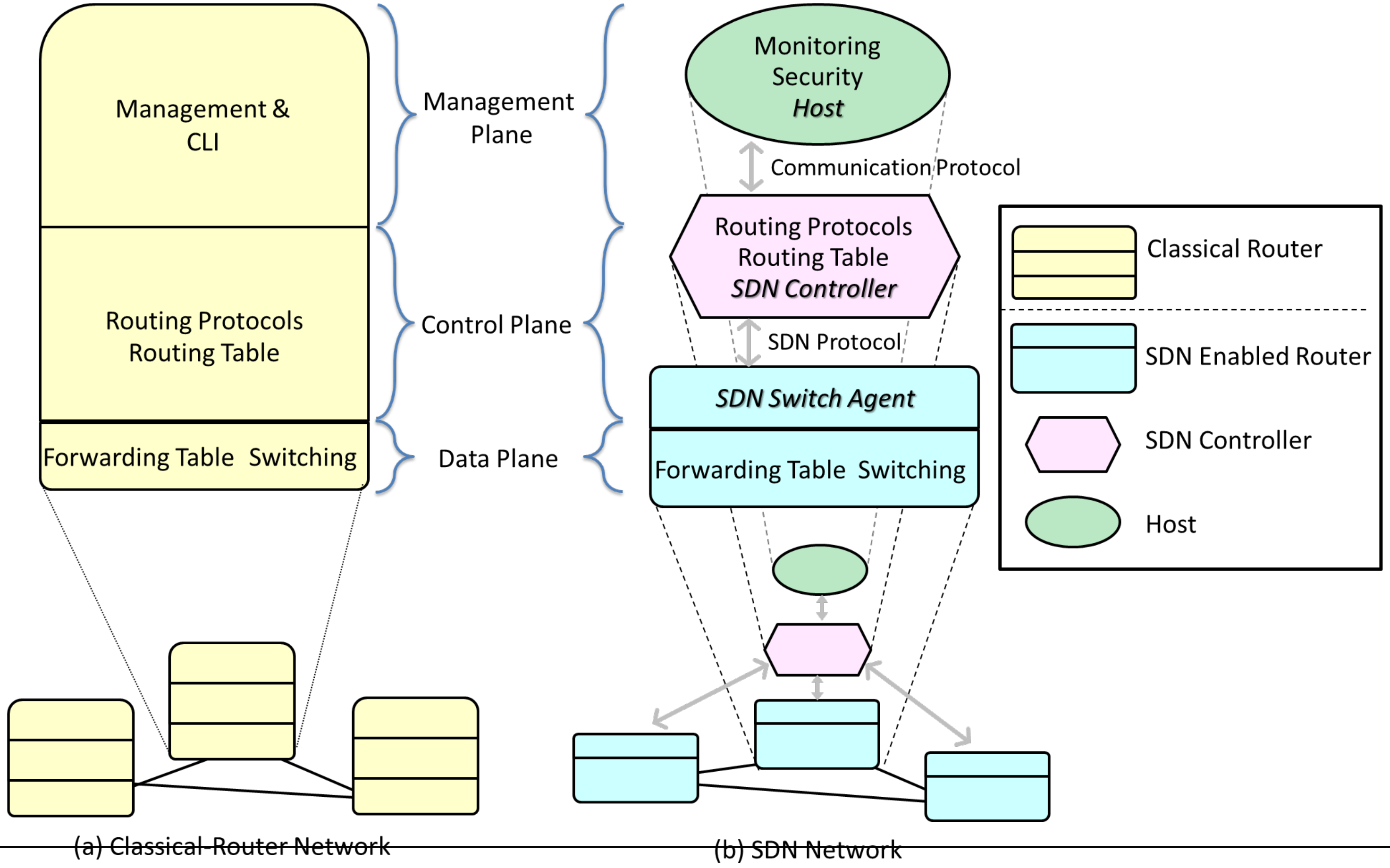
SP – Strict Priority
Pn – Priority <n>

BE – Best Effort
RL – Rate Limiting

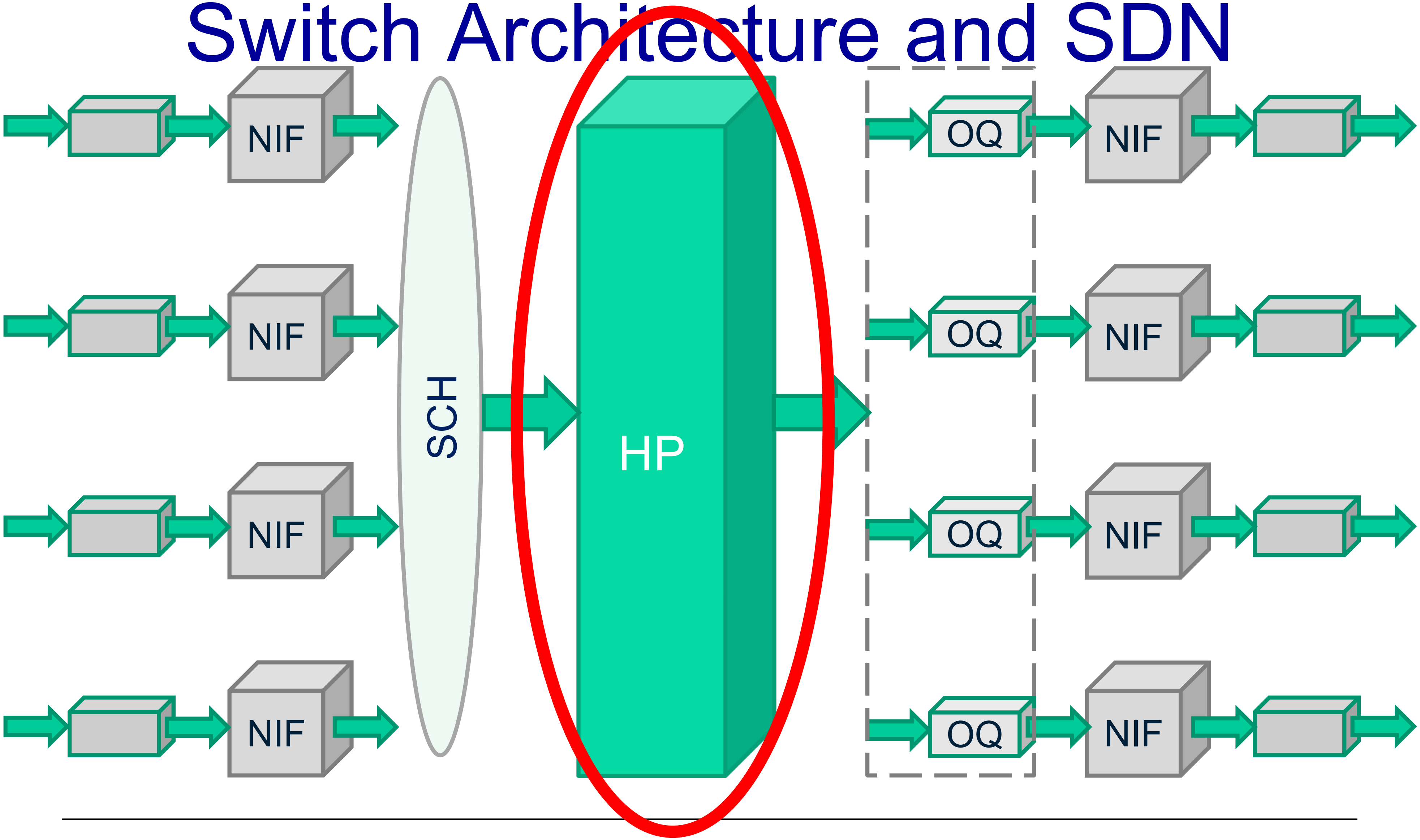
WFQ – Weighted Fair Queueing
RR – Round Robin

Software Defined Networking (SDN)

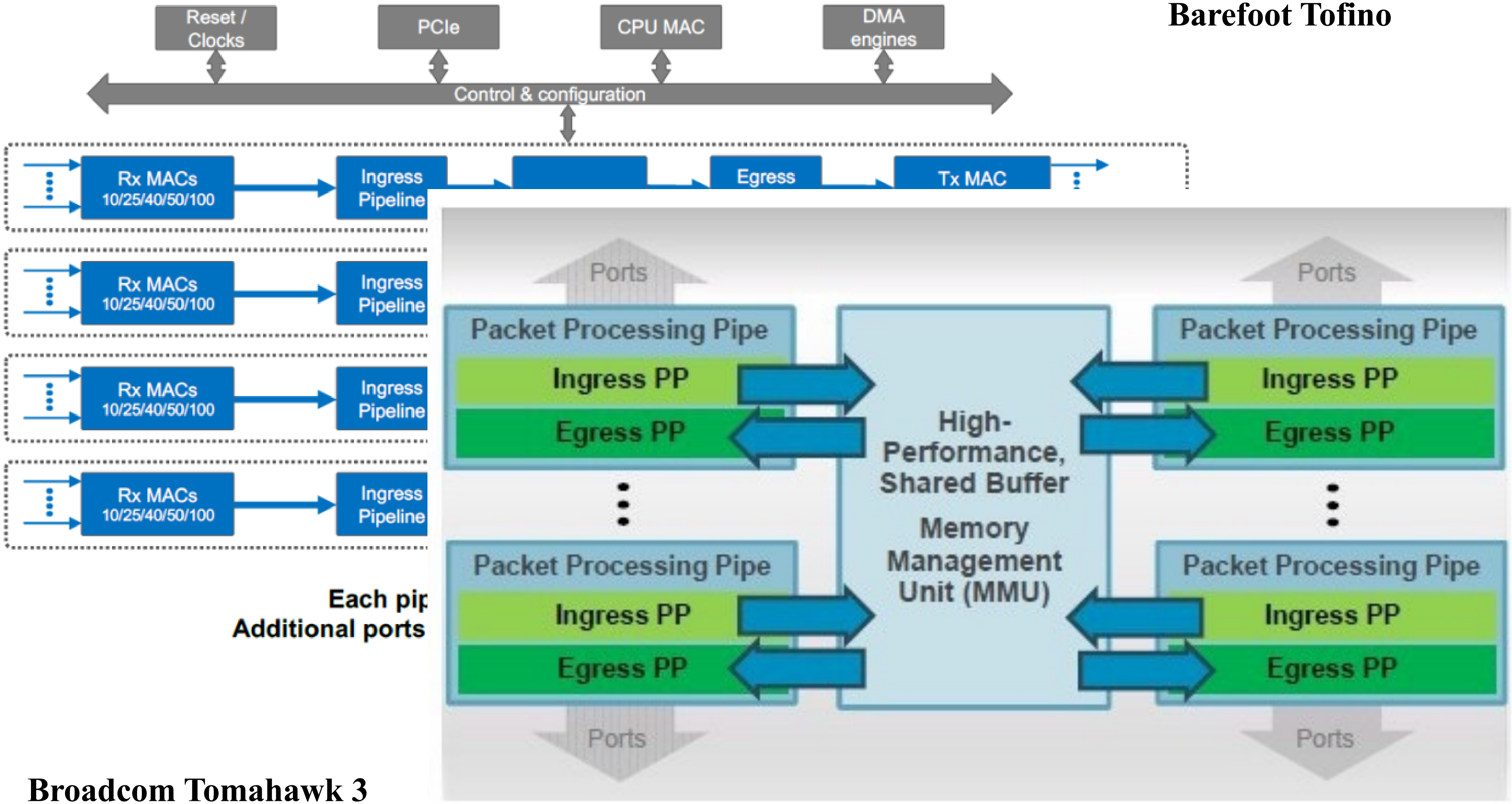
Key Idea: Separation of Data and Control Planes



Switch Architecture and SDN



Multi-Core Switch Design



Barefoot Tofino

Broadcom Tomahawk 3

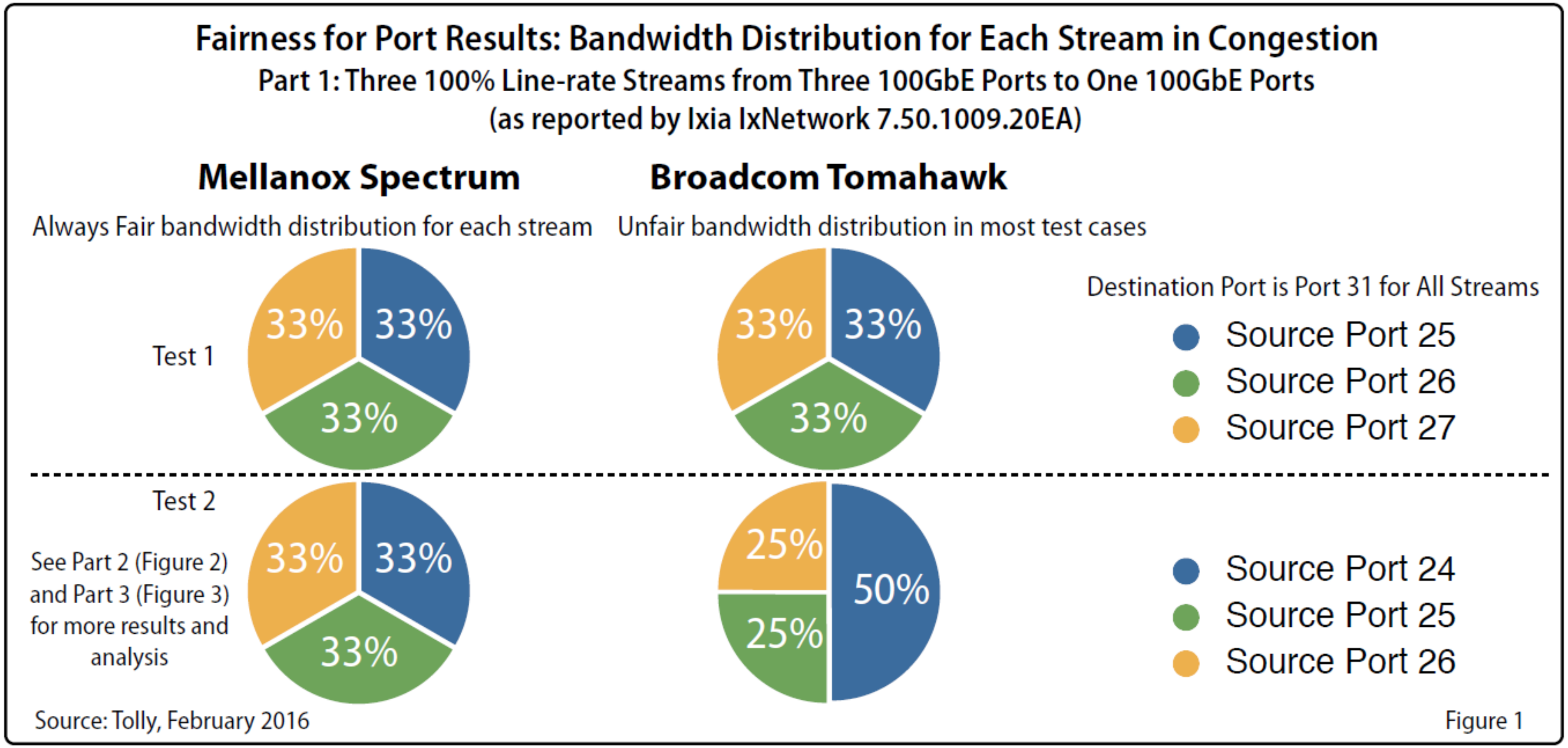
Multi Core Switch Design

- So what? Multi-core in CPUs for over a decade
- Network devices are not like CPUs:
  - CPU: Pipeline - instructions, memory – data
  - Switch: pipeline – data, memory – control
- Network devices have a strong notion of *time*
  - *Must* process the header on cycle X
  - Headers are split across clock cycles
  - Pipelining is the way to achieve performance

Inference and Understanding

All interpretations in the following slides are a *guess*, and not based on internal information – it is taken from careful examination of the Tolly report (and knowledge about switch architecture.)

What makes Mellanox *fairer* than Broadcom Tomahawk?



Fairness for Port Results: Bandwidth Distribution for Each Stream in Congestion

Part 1: Three 100% Line-rate Streams from Three 100GbE Ports to One 100GbE Ports

(as reported by Ixia IxNetwork 7.50.1009.20EA)

Mellanox Spectrum

Always Fair bandwidth distribution for each stream

Broadcom Tomahawk

Unfair bandwidth distribution in most test cases

Destination Port is Port 31 for All Streams

See Part 2 (Figure 2) and Part 3 (Figure 3) for more results and analysis

Source: Tolly, February 2016

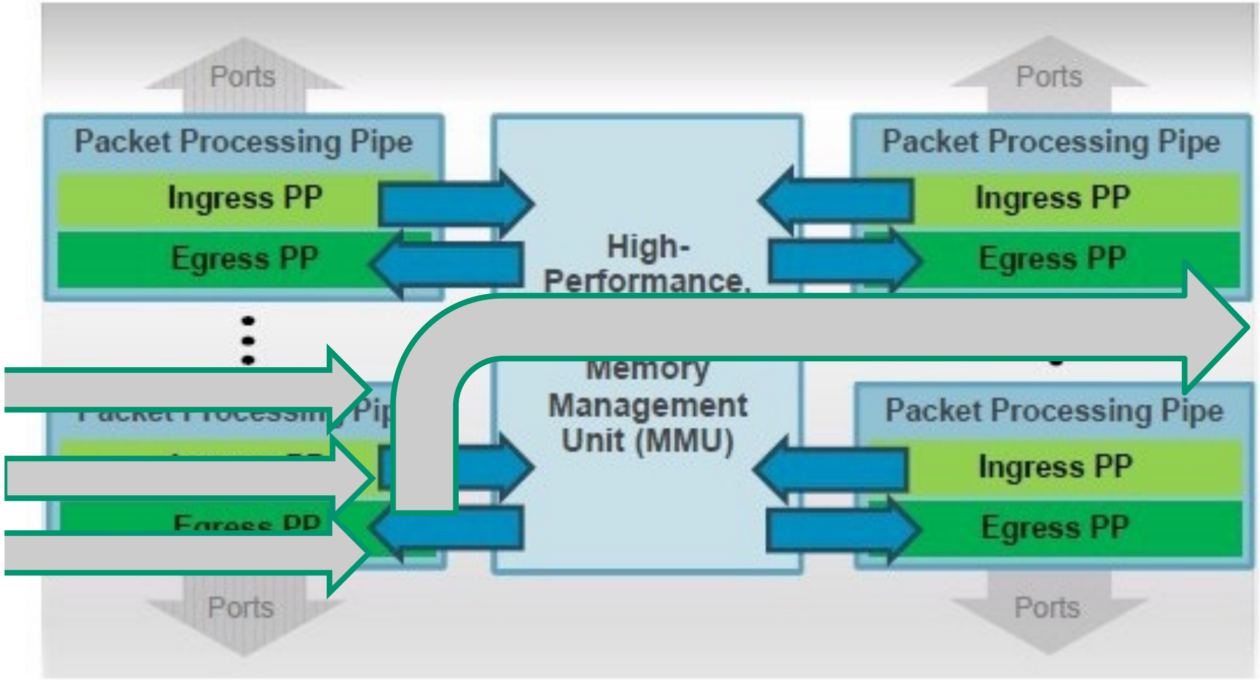
Figure 1

Broadcom Tomahawk

- 32 x 100GE
- In packet rate: 32 x 150Mpps = 4800 Mpps
- Manufacturing process: 28nm
  - Therefore clock frequency likely <1GHz
- More than 7 billion transistor
  - Reference: Intel debut around the same time 18-core Xeon E5-2600 v3 with 5.57 billion transistors
- … now lets think of these experimental results in a multi core switch…

What is weird with Broadcom Tomahawk?

- Let us assume the same architecture as used by Tomahawk 3:



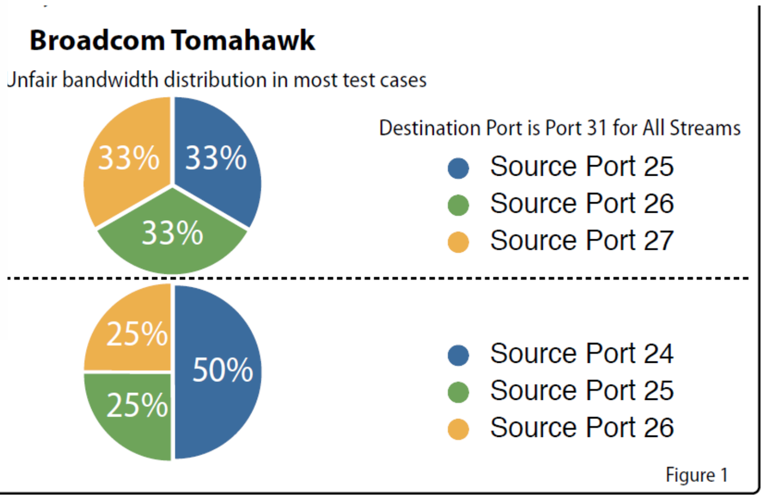
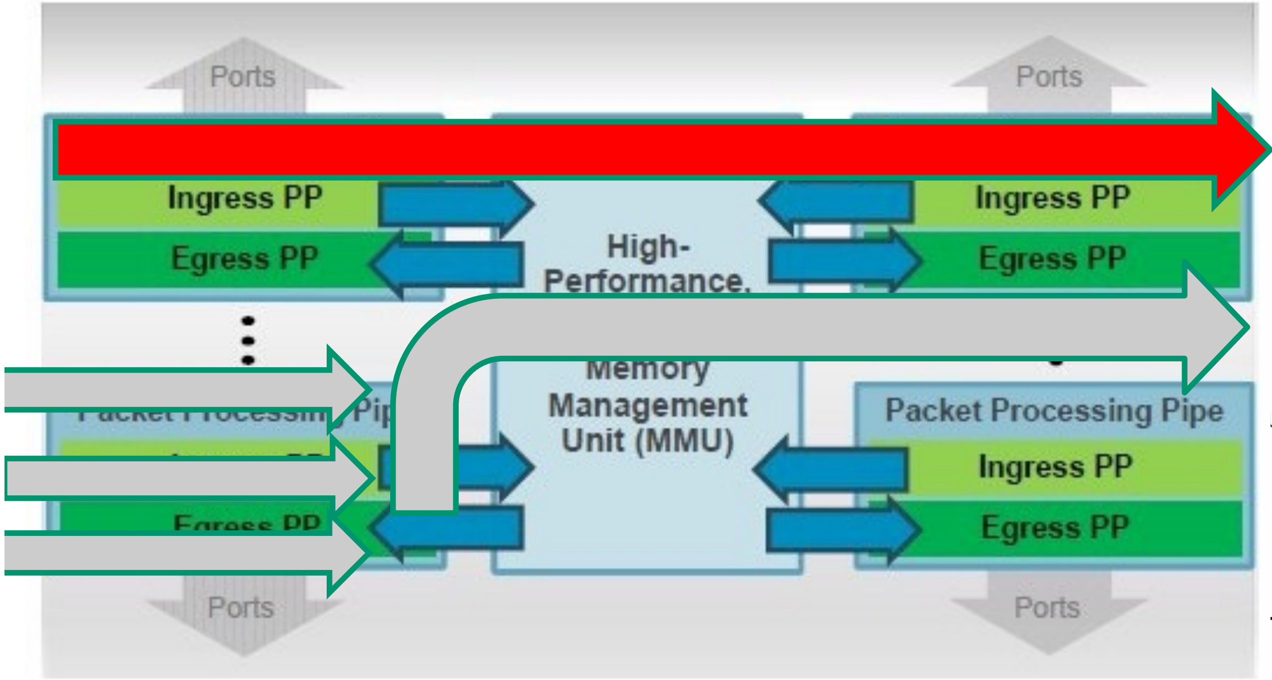


Figure 1

What is different about Broadcom Tomahawk?

- Let us assume the same architecture as used by Tomahawk 3:



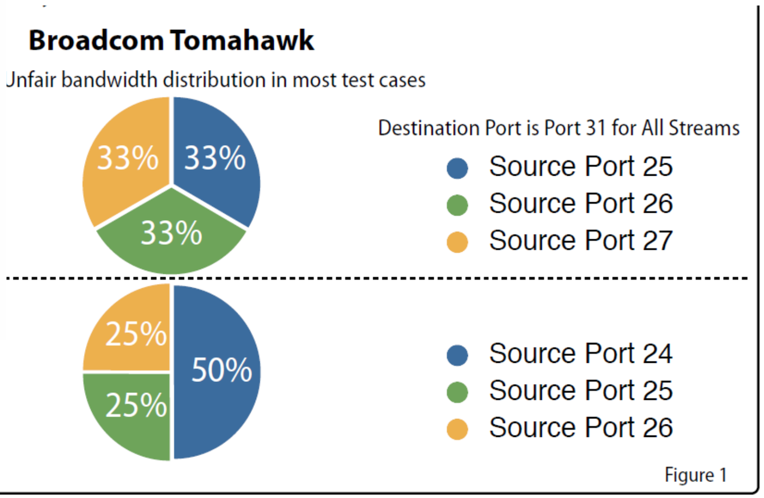
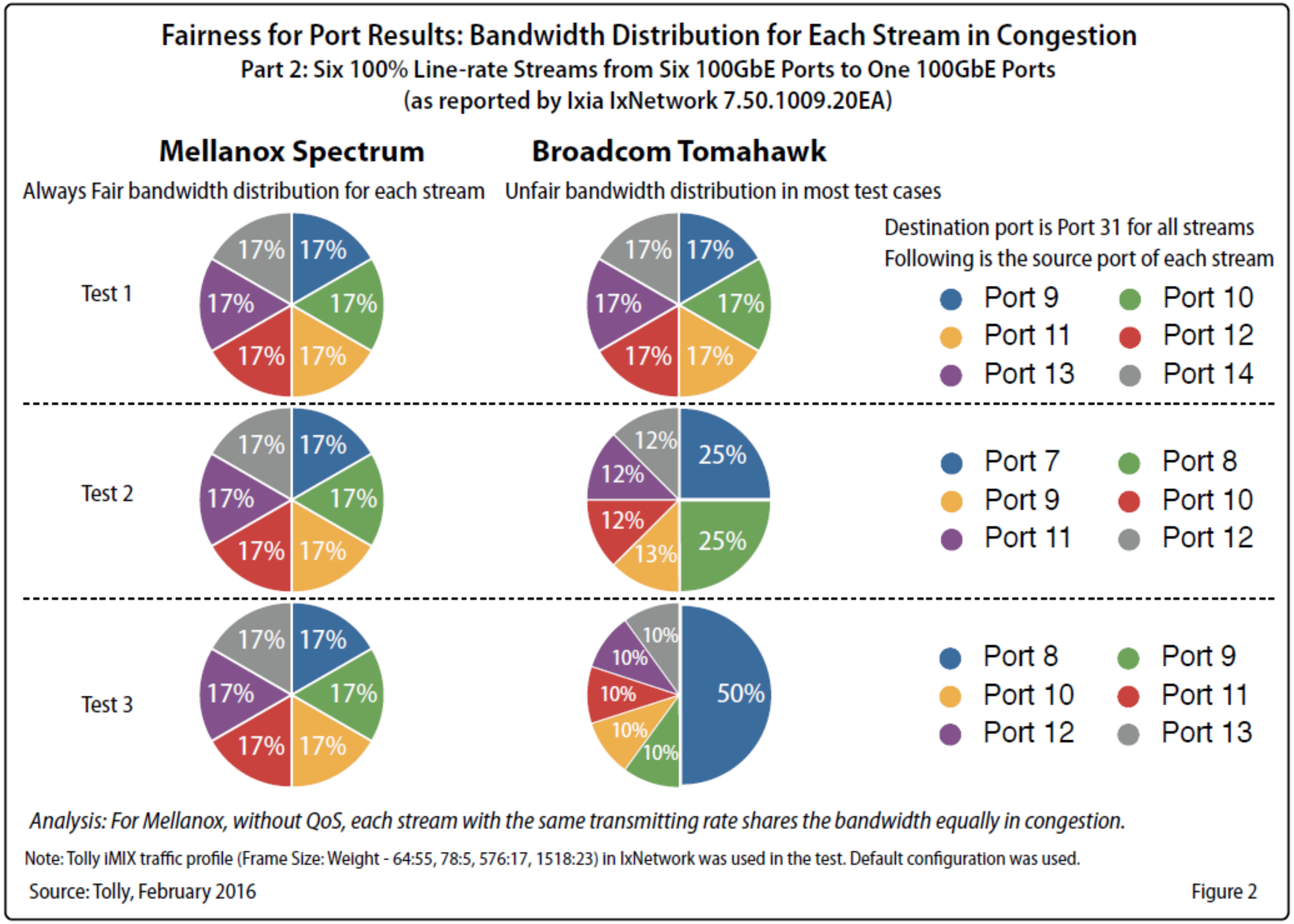


Figure 1

What is different about Broadcom Tomahawk?



Fairness for Port Results: Bandwidth Distribution for Each Stream in Congestion

Part 2: Six 100% Line-rate Streams from Six 100GbE Ports to One 100GbE Ports

(as reported by Ixia IxNetwork 7.50.1009.20EA)

Mellanox Spectrum

Always Fair bandwidth distribution for each stream

Broadcom Tomahawk

Unfair bandwidth distribution in most test cases

Destination port is Port 31 for all streams

Following is the source port of each stream

Analysis: For Mellanox, without QoS, each stream with the same transmitting rate shares the bandwidth equally in congestion.

Note: Tolly iMIX traffic profile (Frame Size: Weight - 64:55, 78:5, 576:17, 1518:23) in IxNetwork was used in the test. Default configuration was used.

Source: Tolly, February 2016

Figure 2

Synchronization

- Recall Lecture 3
- Synchronization of time between multiple machines
  - ➢ E.g., allow one-way latency measurements
- Synchronization of measurements
  - ➢ Can you trigger multiple vantage points to start an experiment at once?
    - E.g. what happens if you measure congestion effects without triggering them simultaneously?

Tools Selection

- When to use hardware tools? When to use software tools?
- You don’t always have omniscient control over resources
  - You may not even have permissions for some basic tools
- What can you do?
  - Similar tools using different protocols
  - Write your own tools
  - Redesign your experiment

So lets start measuring!

- Wait!
- What is your goal?
- What do you know about your experimentation environment?
- Have you collected metadata?
- Are you aware of any limitations to the environment / tests / DUT / usage / …?
- Is your experiment reproducible?

Advice

- Getting measurements right is *HARD*
- More is rarely better
- Prefer:
  - Fewer Measurements and Better methodology
  - Detailed measurements
  - Reproducibility
  - Understanding the results
  - Become an expert of your work