A High-Speed Stateful Packet Processing Approach for Tbps Programmable Switches

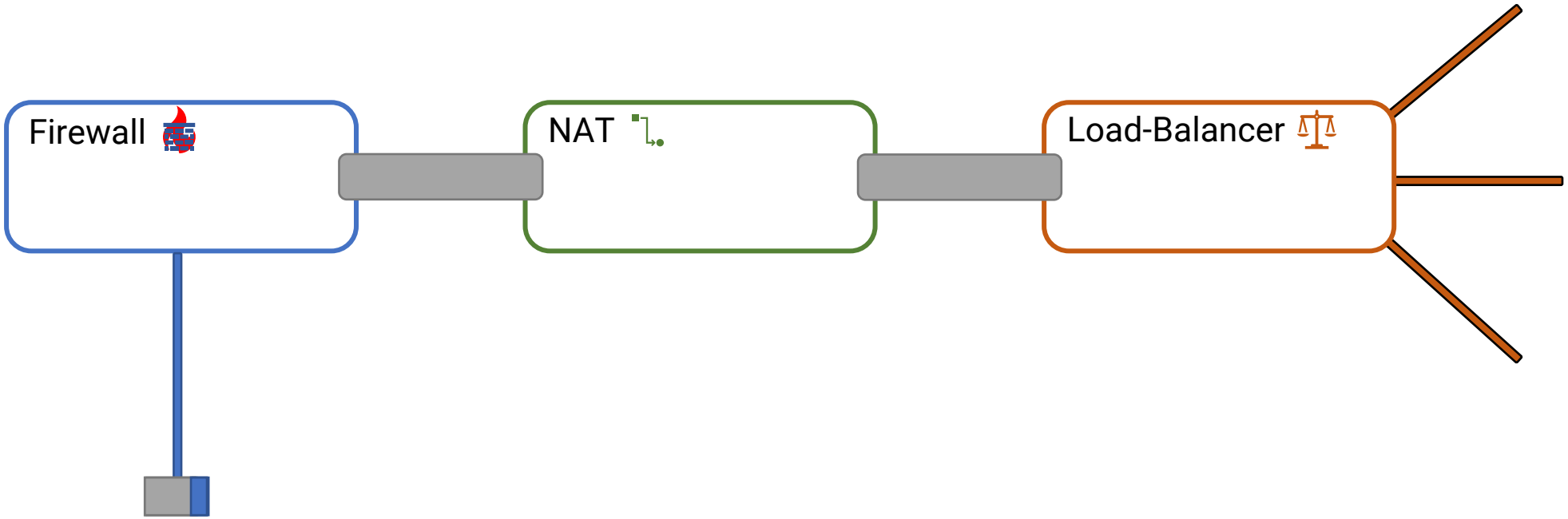
Mariano Scazzariello* †, **Tommaso Caiazzi*** †, Hamid Ghasemirahni †,
Tom Barbette ‡, Dejan Kostić †, Marco Chiesa †

*Roma Tre University, Italy – † KTH Royal Institute of Technology, Sweden - ‡UCLouvain, Belgium

Network Functions Are Pervasive

Network Functions Are Pervasive

Network Functions Virtualization is an essential architectural paradigm of today's networks



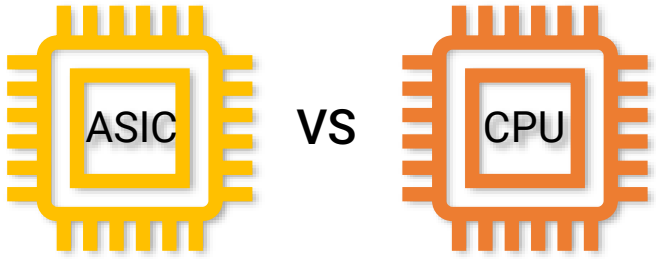
Network Functions Deployment

Cost

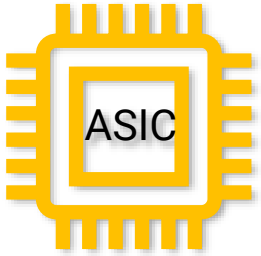
Energy footprint

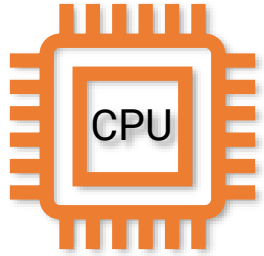
Network Functions Deployment

Network Functions Deployment

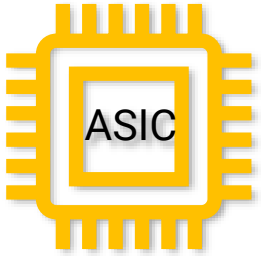


Network Functions Deployment

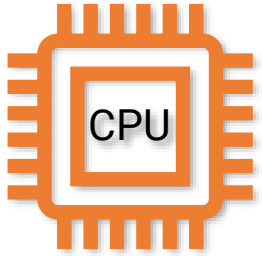


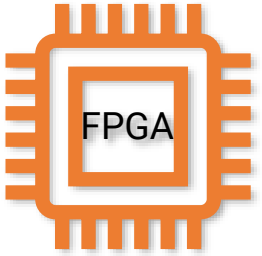


Network Functions Deployment

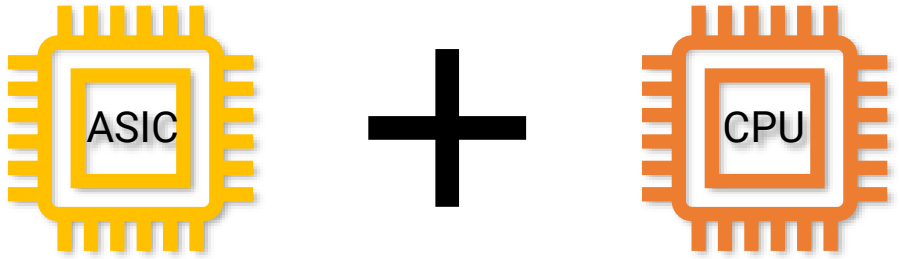


- Throughput in Tbps
- Lower energy footprint
- Cost-effective
- Scarce memory (10-100MB)
- Not Expressive





- Throughput in Gbps
- Higher energy footprint
- Not cost-effective
- Enough memory
- Expressive



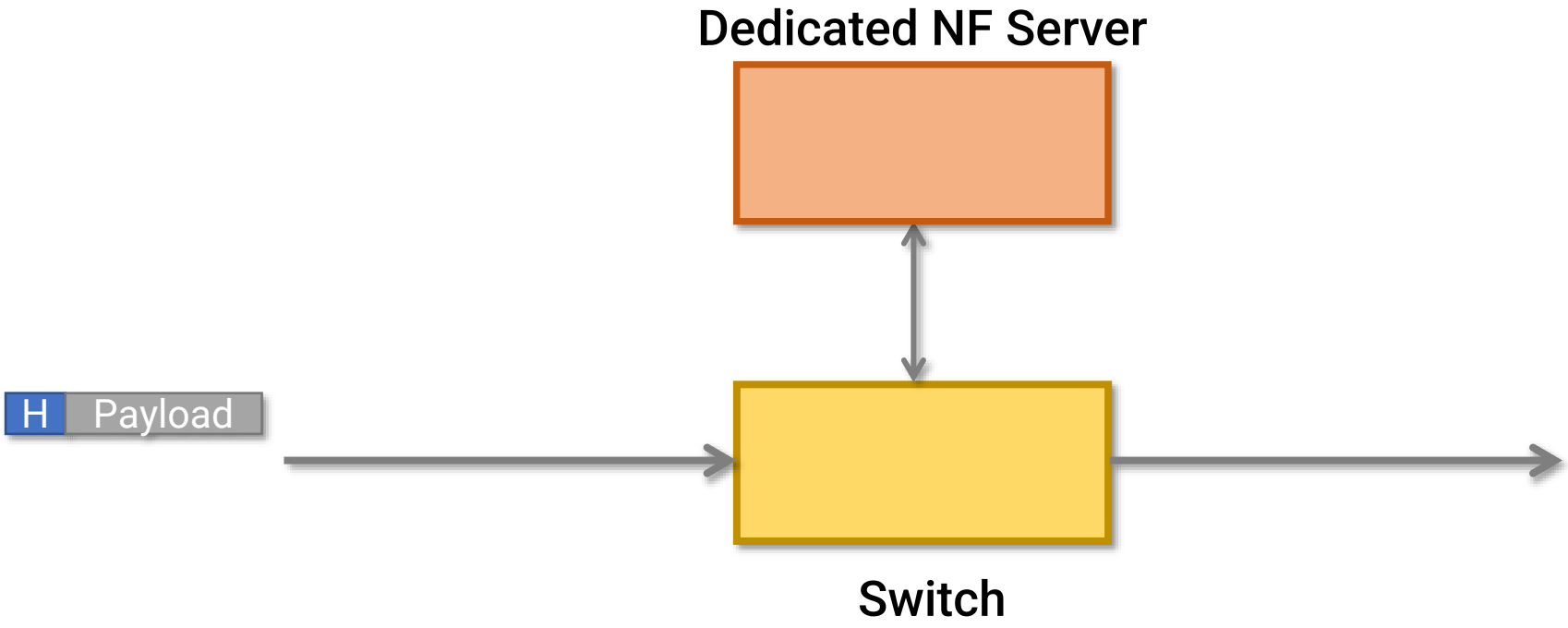
✓ Throughput in Tbps

✓ Support complex functions

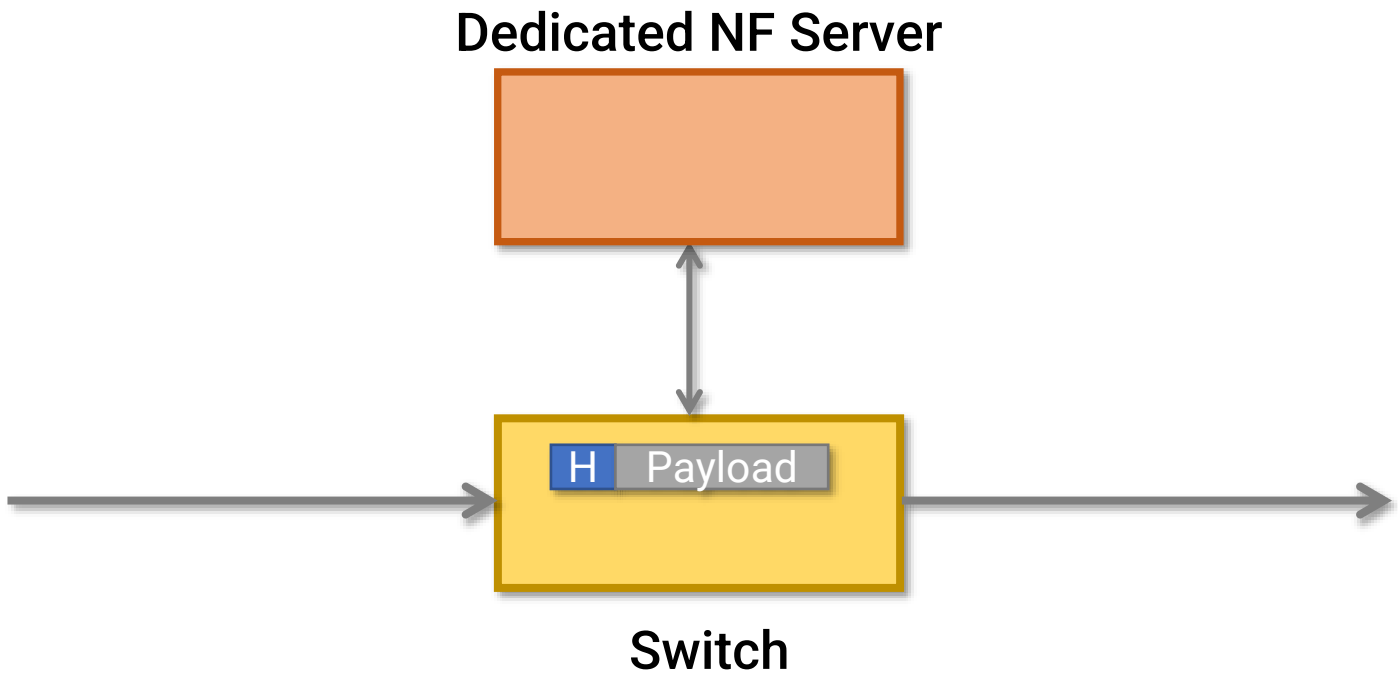
✓ Cost-effective

Can we design a packet processing pipeline that handles
one terabit per second of traffic
on a single dedicated device?

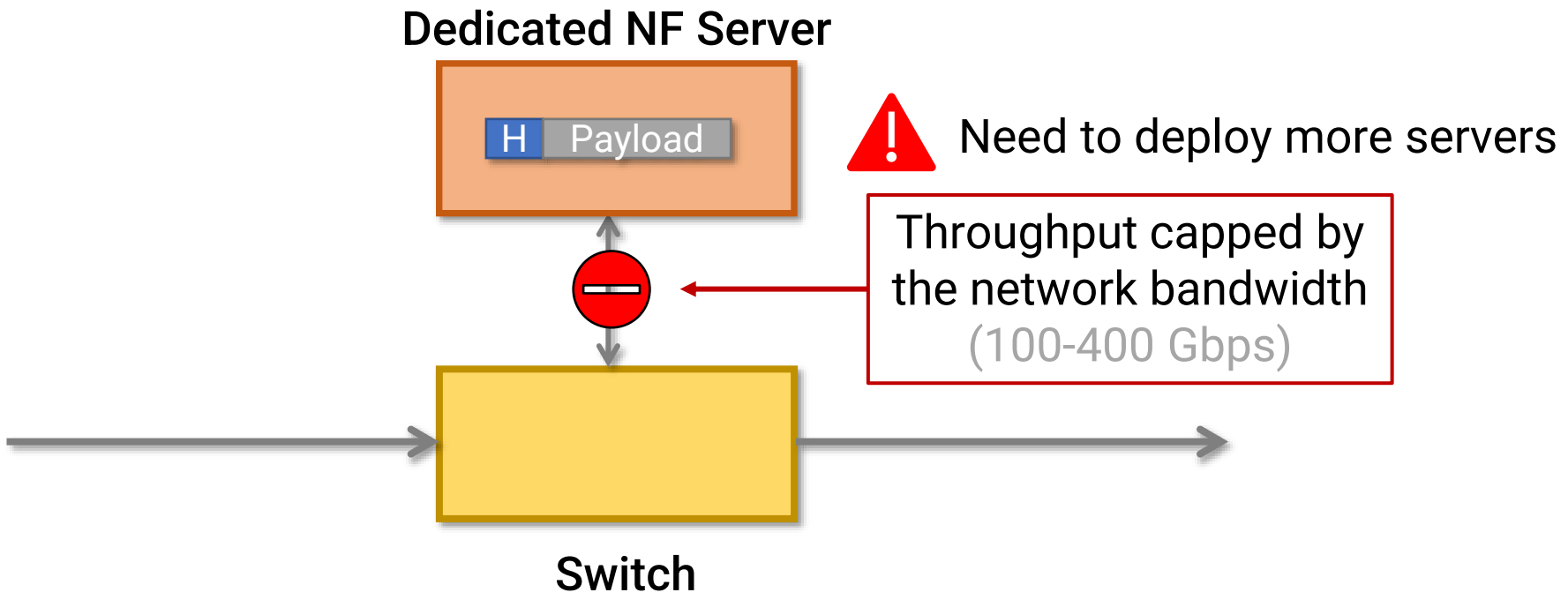
The bandwidth limit



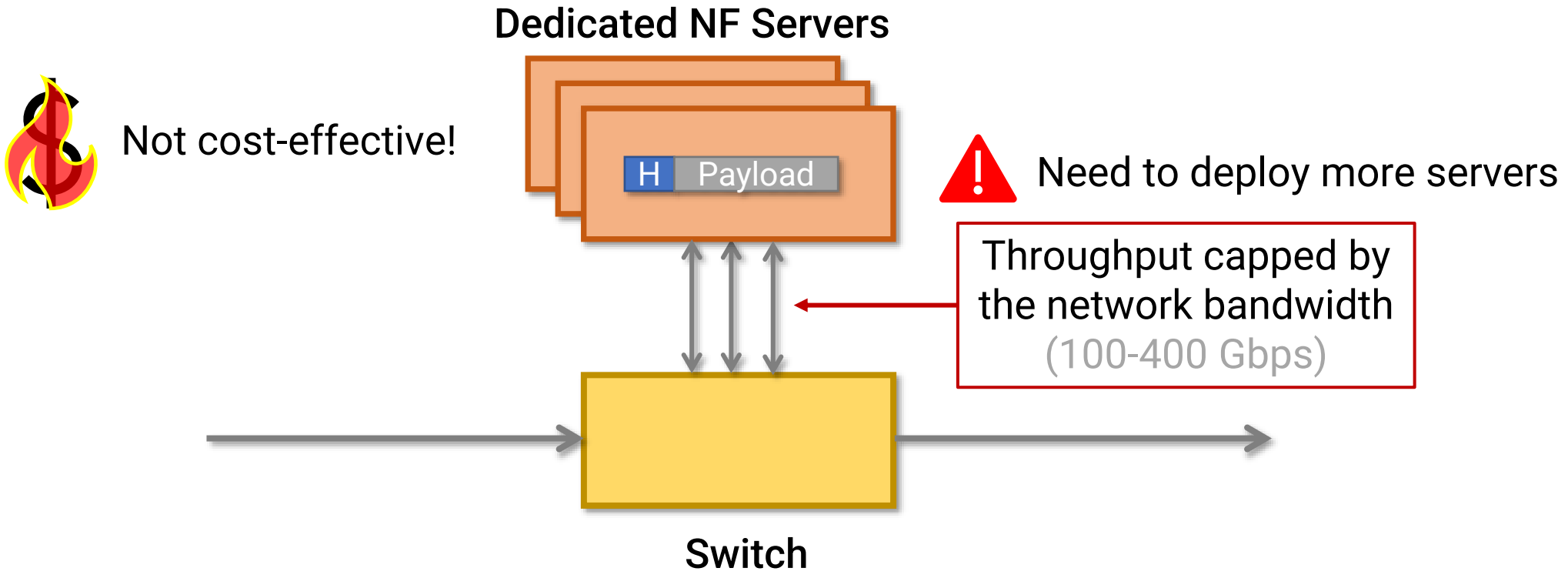
The bandwidth limit



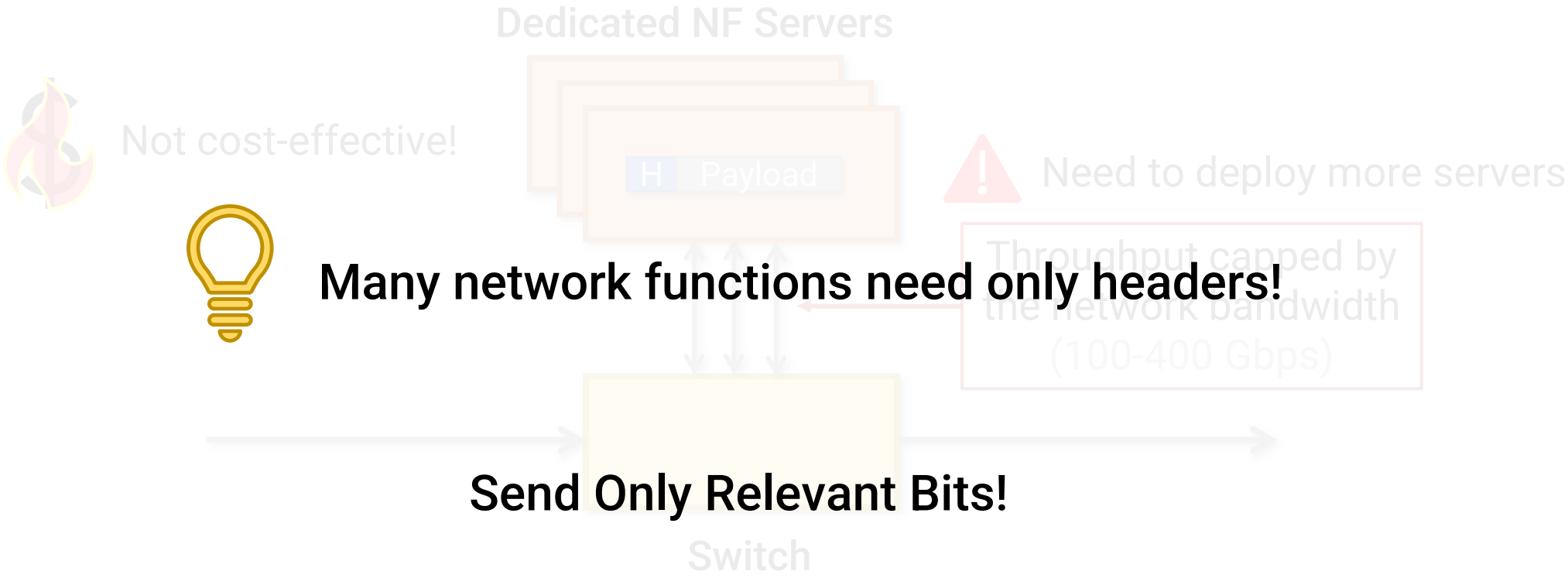
The bandwidth limit



The bandwidth limit



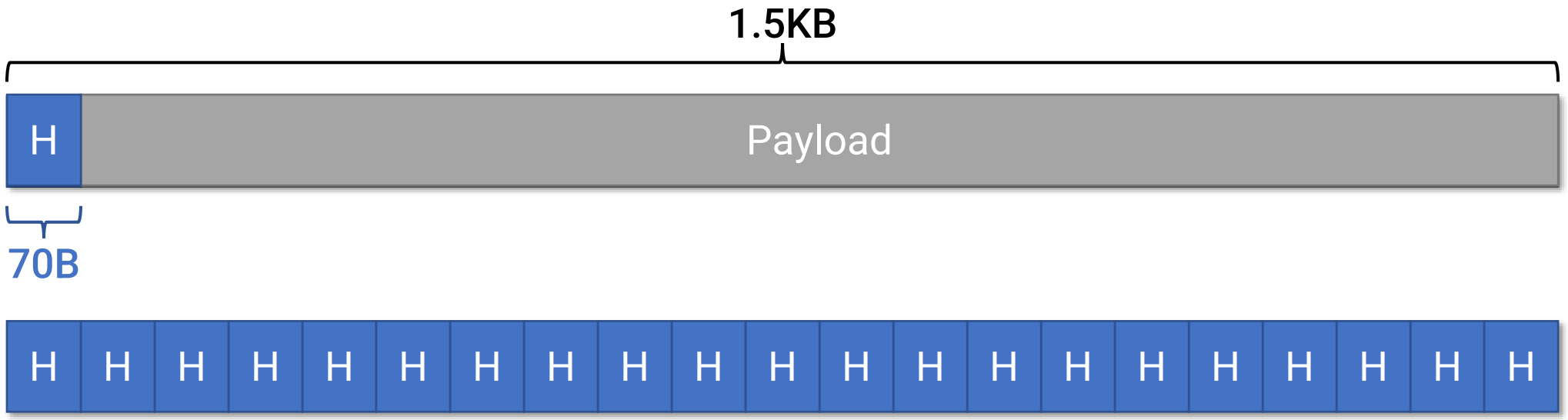
The bandwidth limit



Many network functions need only headers!

Send Only Relevant Bits!

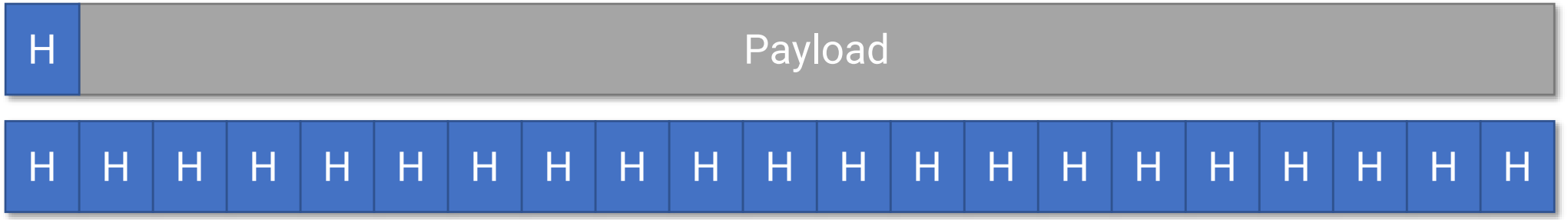
Send Only Relevant Bits!

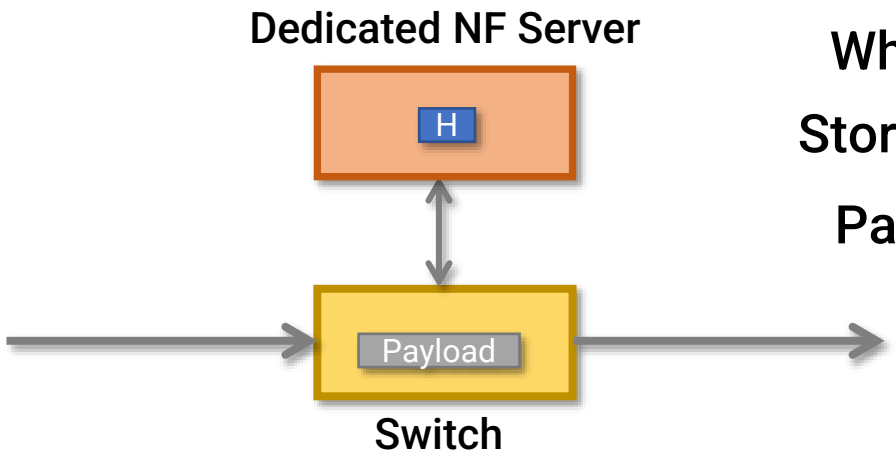


✓ Free up bandwidth

✓ Higher cache-hit ratio

Send Only Relevant Bits!



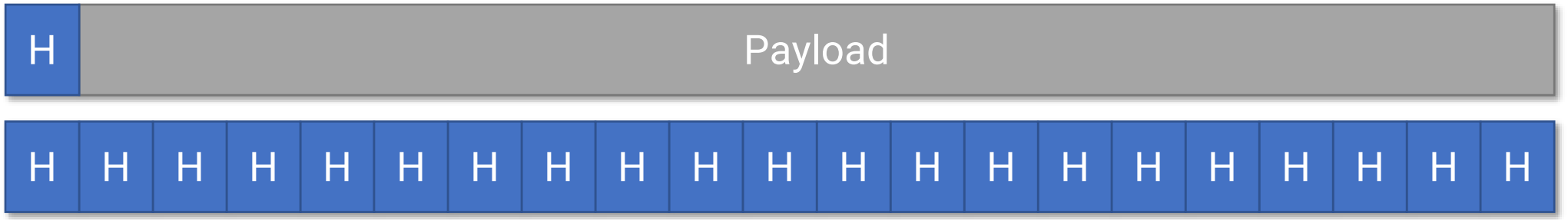


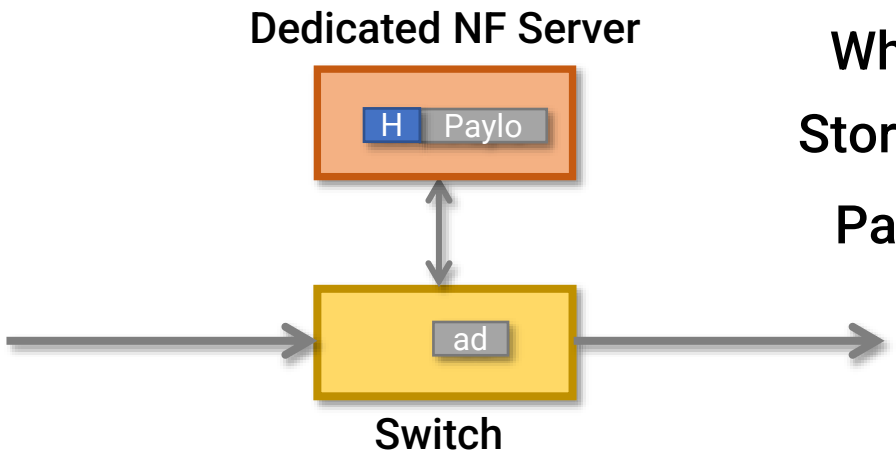
Where to store payloads?

Store Payloads on the Switch

PayloadPark [CoNEXT '20]

Send Only Relevant Bits!





Where to store payloads?

Store Payloads on the Switch

PayloadPark [CoNEXT '20]

Store Payloads on the Switch

What is the impact?

Store Payloads on the Switch

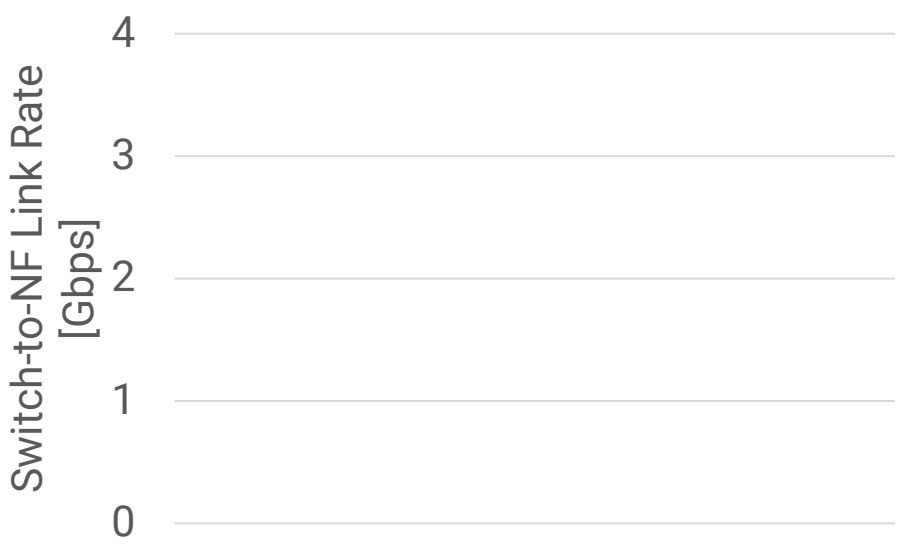
What is the impact?

Let’s examine a CAIDA trace

Store Payloads on the Switch

What is the impact?

Let’s examine a CAIDA trace

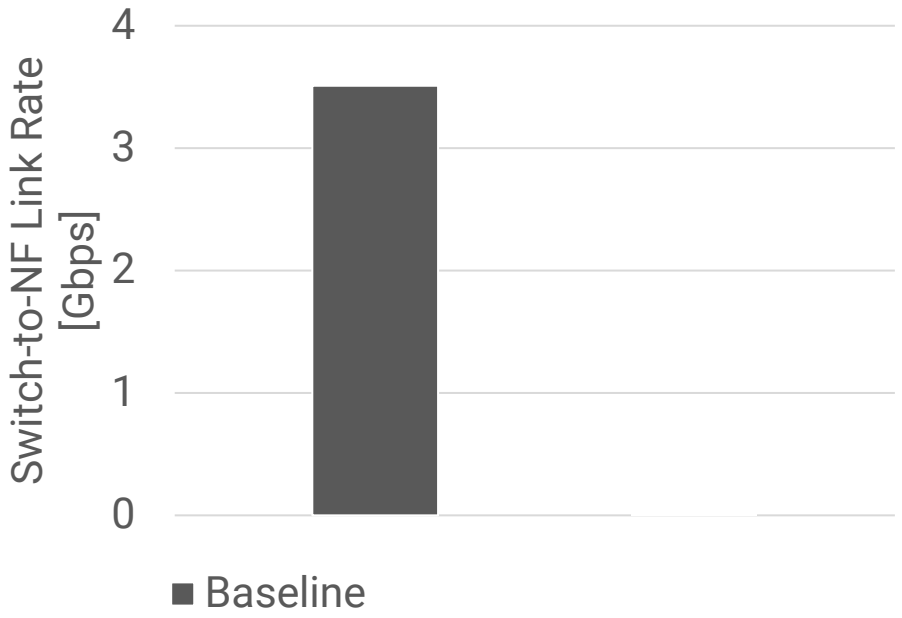


Packets sent to the NF

Store Payloads on the Switch

What is the impact?

Let's examine a CAIDA trace



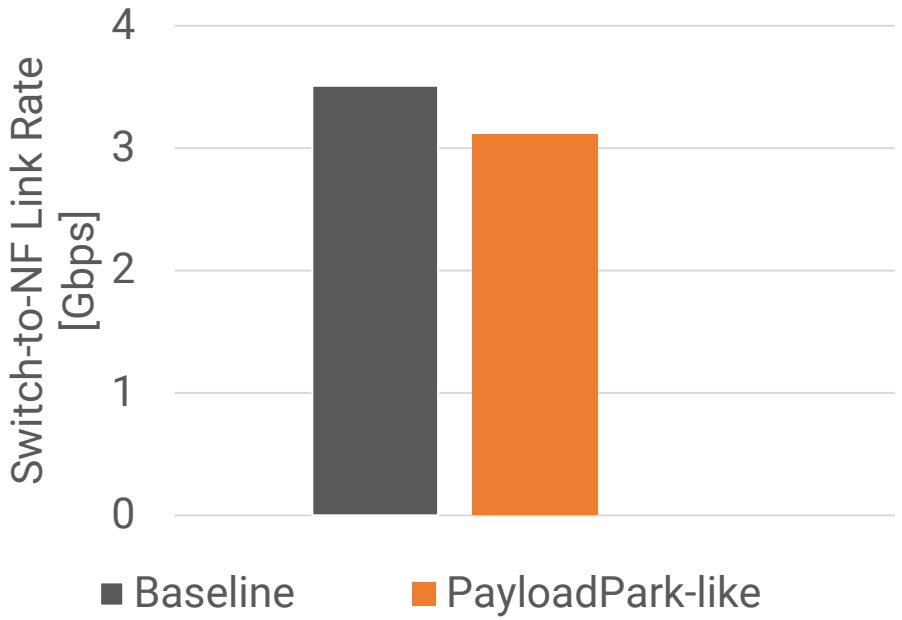
Packets sent to the NF

Baseline

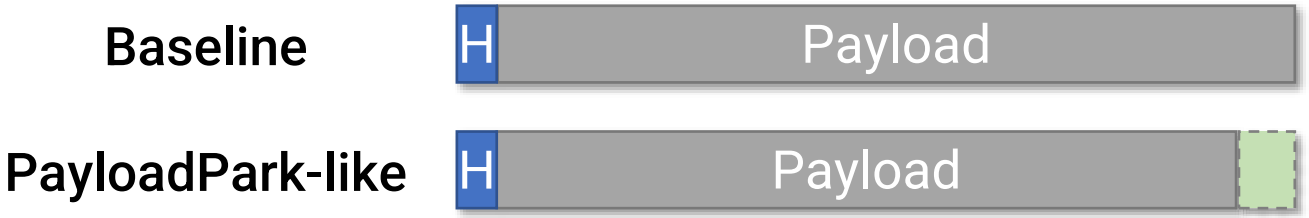
Store Payloads on the Switch

What is the impact?

Let's examine a CAIDA trace



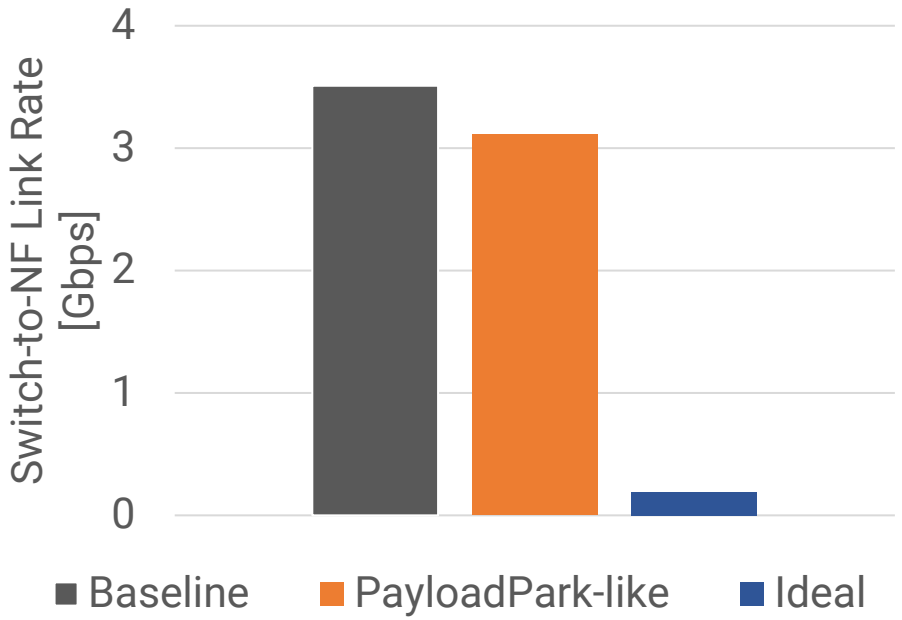
Packets sent to the NF



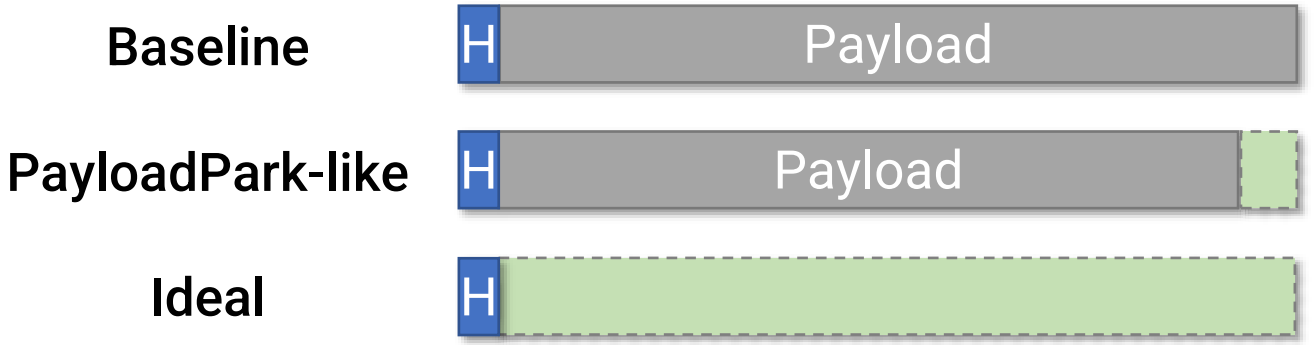
Store Payloads on the Switch

What is the impact?

Let's examine a CAIDA trace



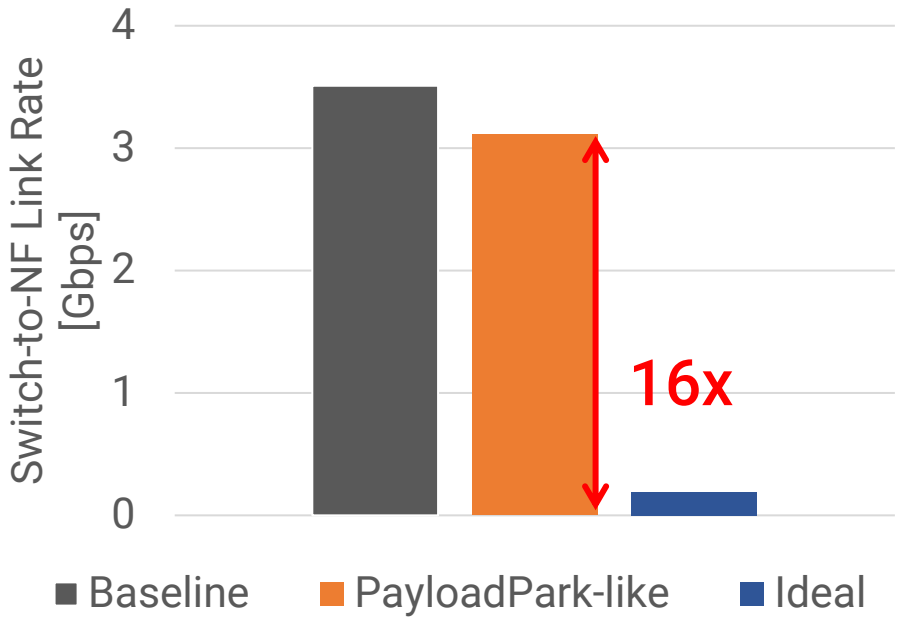
Packets sent to the NF



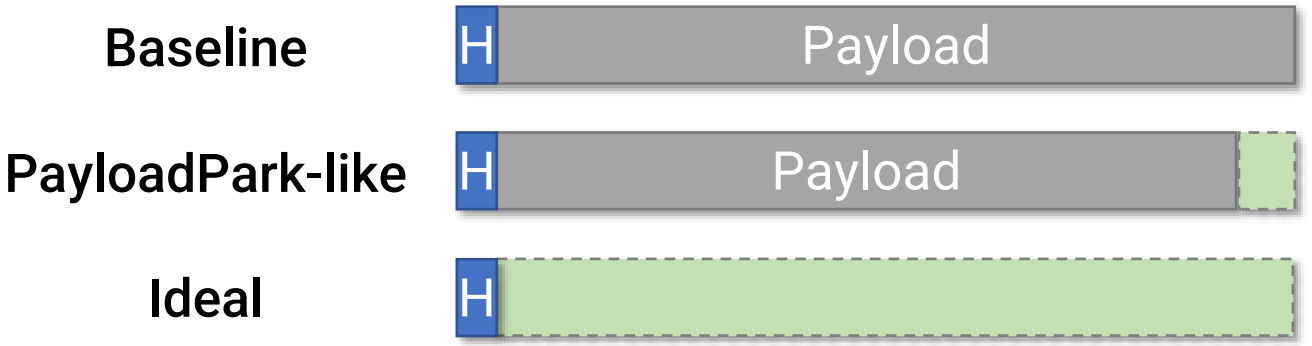
Store Payloads on the Switch

What is the impact?

Let's examine a CAIDA trace



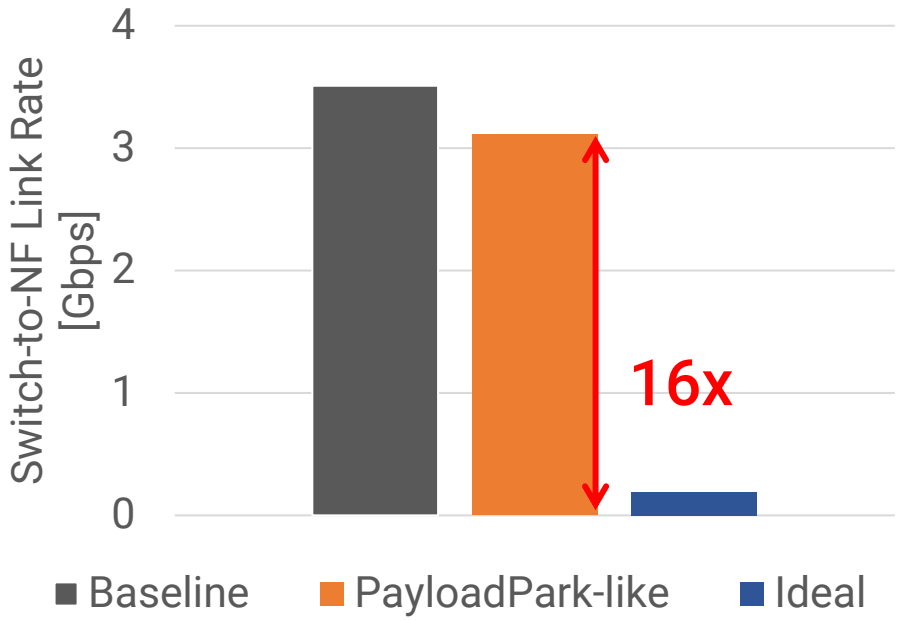
Packets sent to the NF



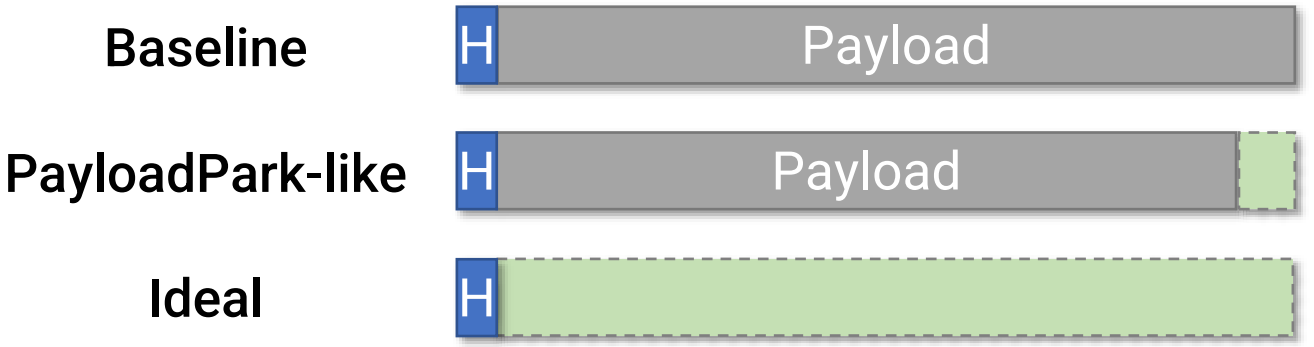
Store Payloads on the Switch

What is the impact?

Let's examine a CAIDA trace



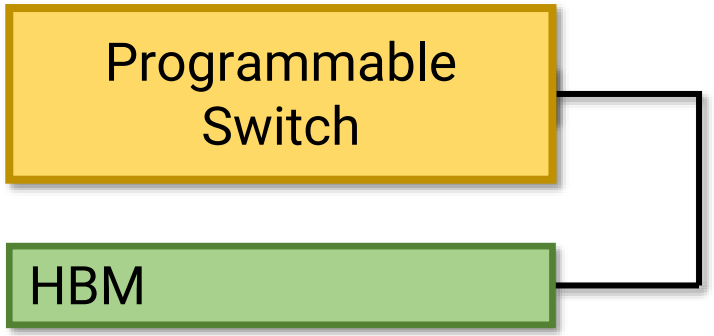
Packets sent to the NF



How to extend the switch memory?

How to extend the switch memory?

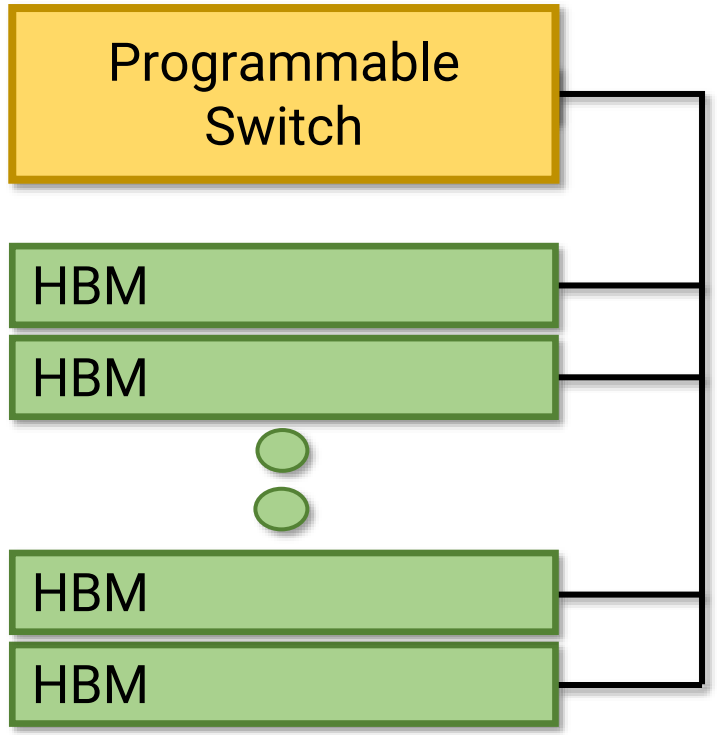
- Using a dedicated external memory (*e.g.,* HBM)
  - ✓ Simple solution



How to extend the switch memory?

Not cost-effective

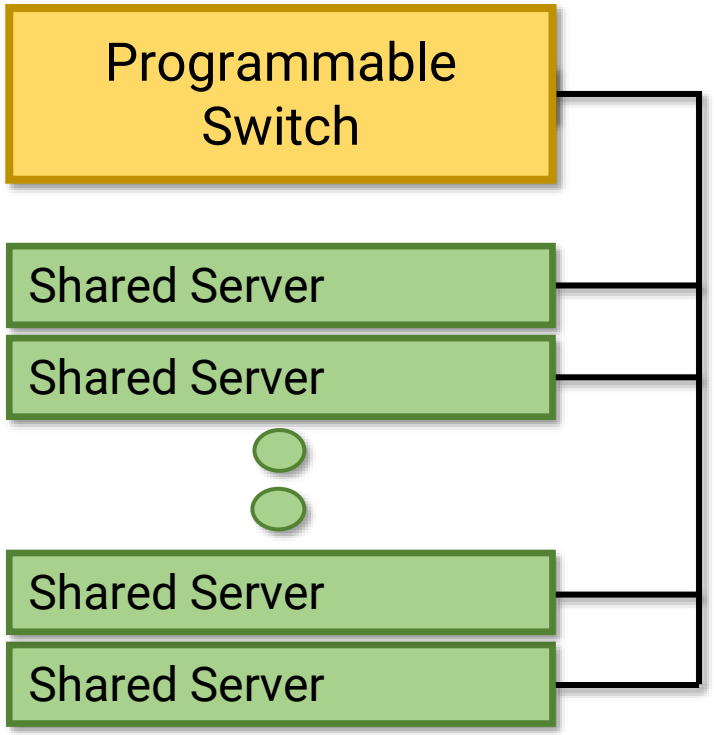
- Using a dedicated external memory (*e.g.,* HBM)
  - ✓ Simple solution
  - ↓ Higher energy footprint
  - ↓ High-cost
  - ! Wastes some ports on the switch



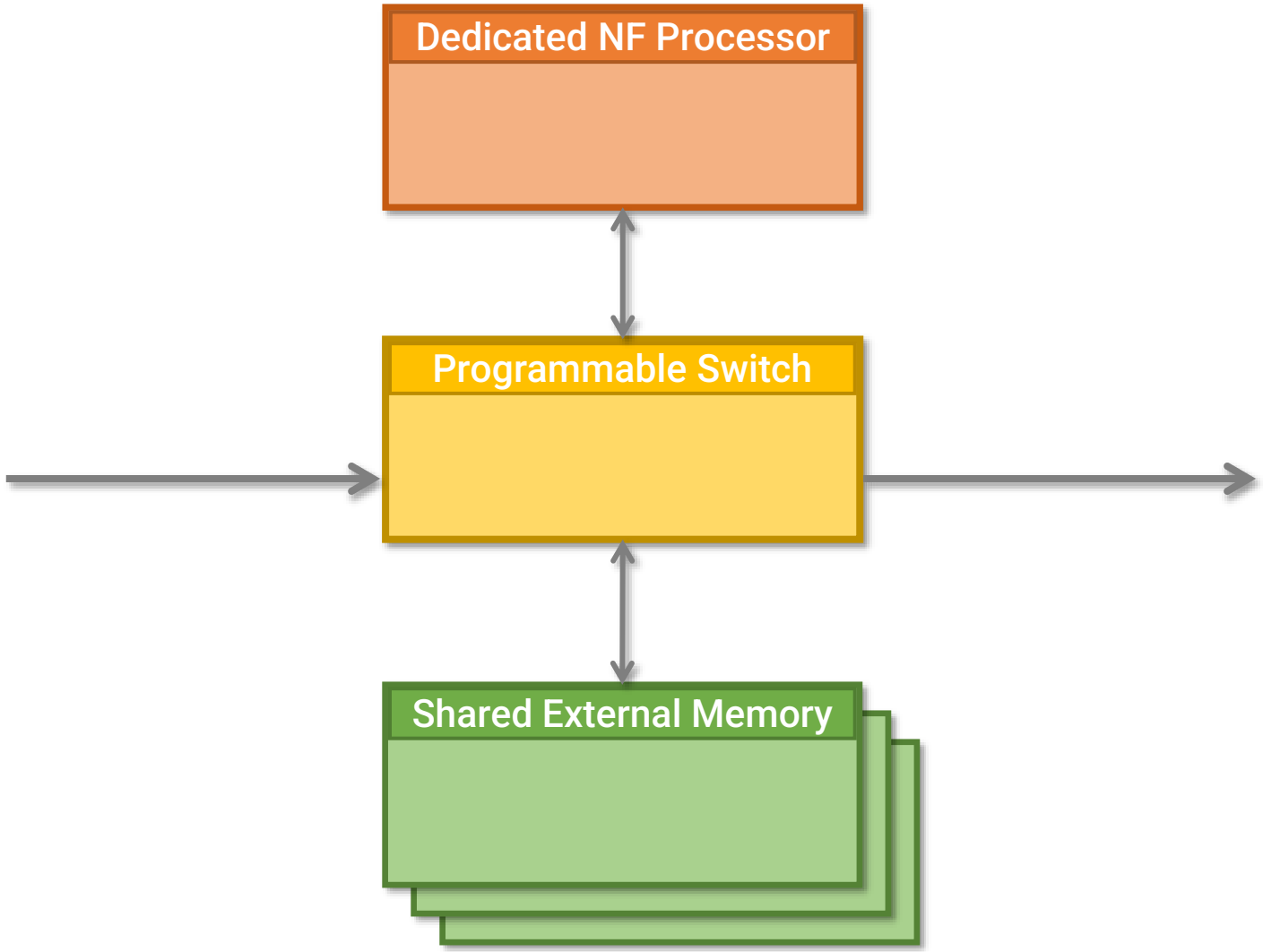
How to extend the switch memory?

The Ribosome Approach

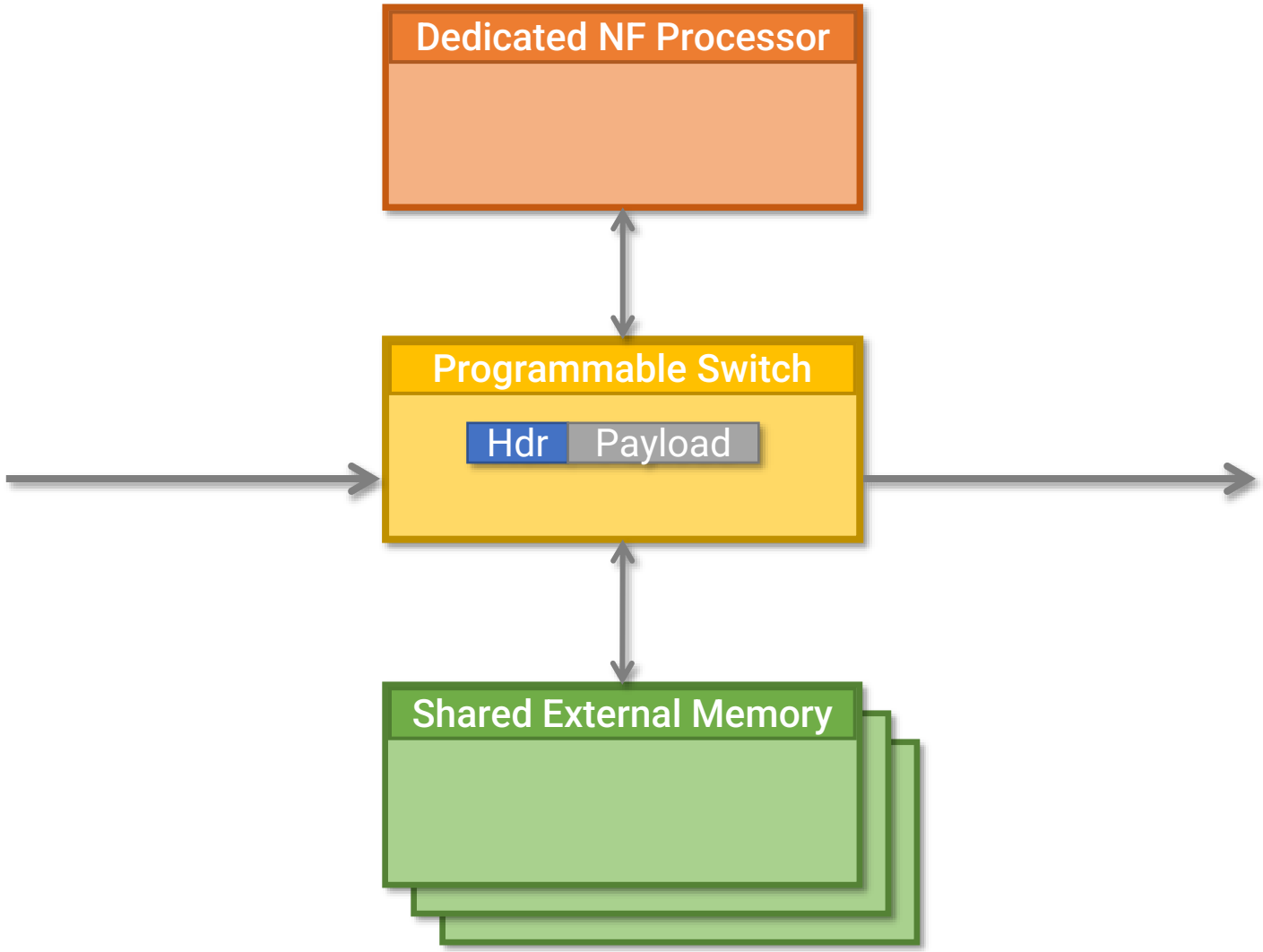
- Exploiting a disaggregated pipeline on shared servers
  - ✓ Many spare resources in the datacenter
  - ↑ Better resources usage
  - ↑ Low-cost



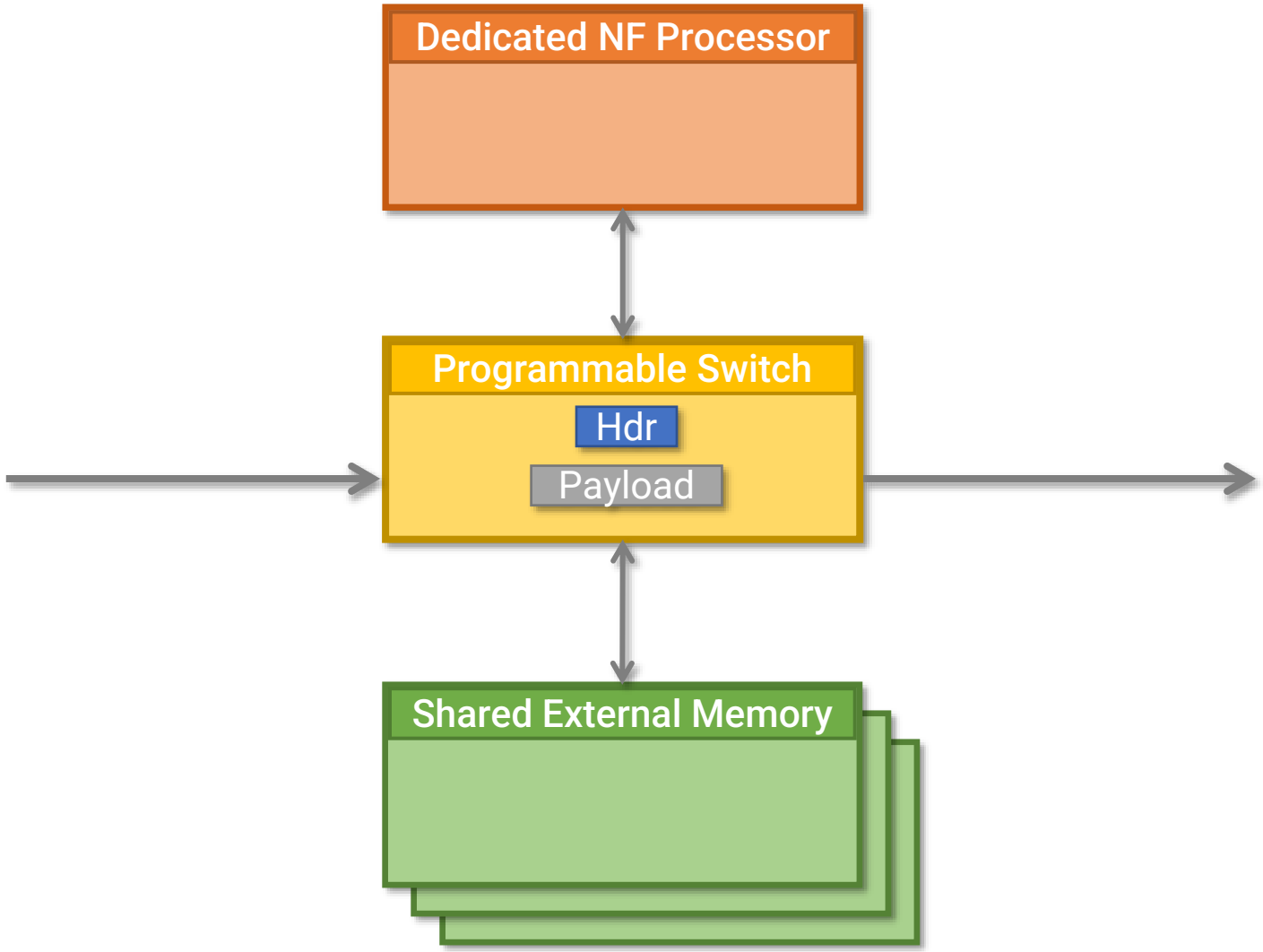
The Ribosome Approach



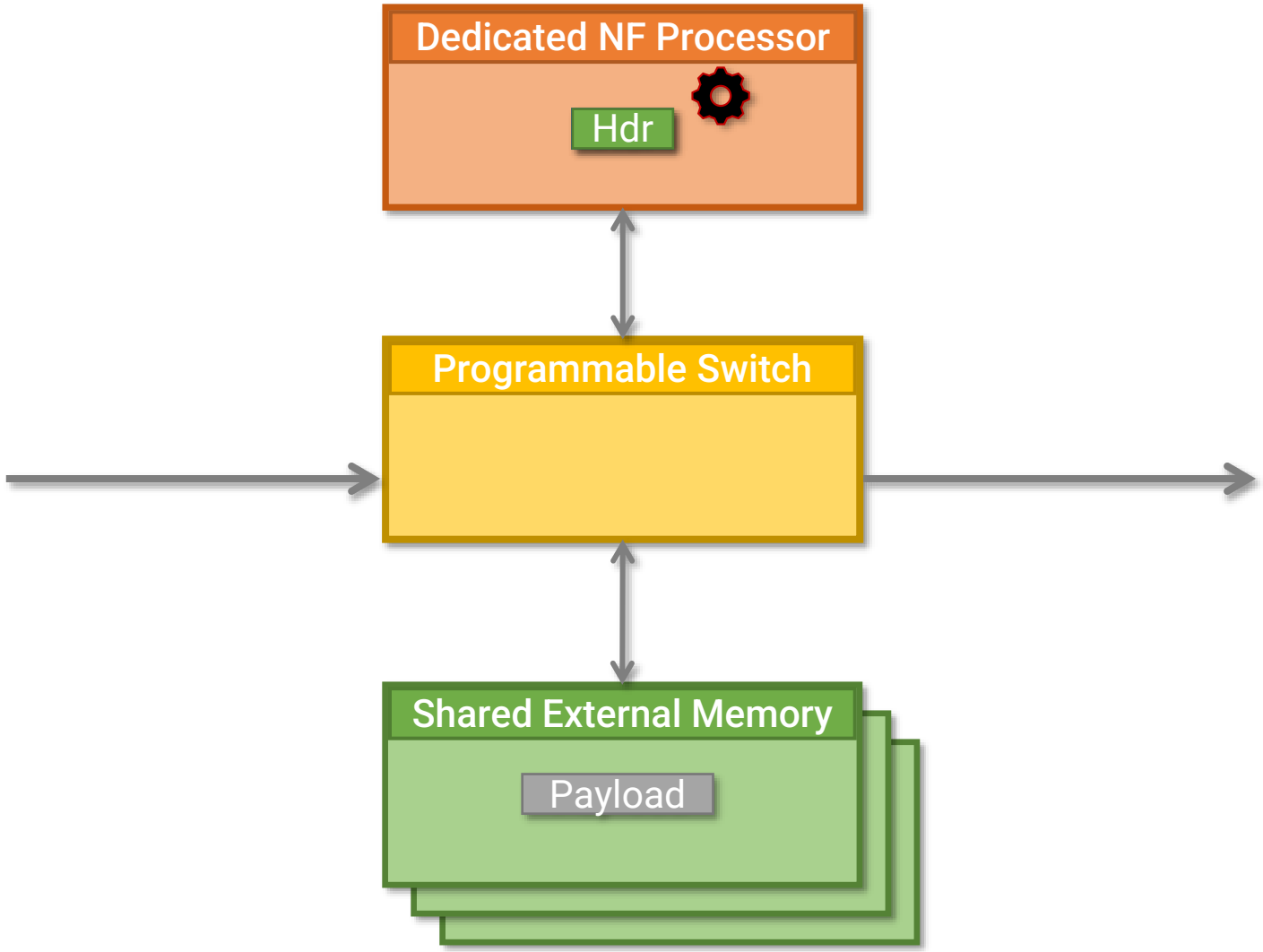
The Ribosome Approach



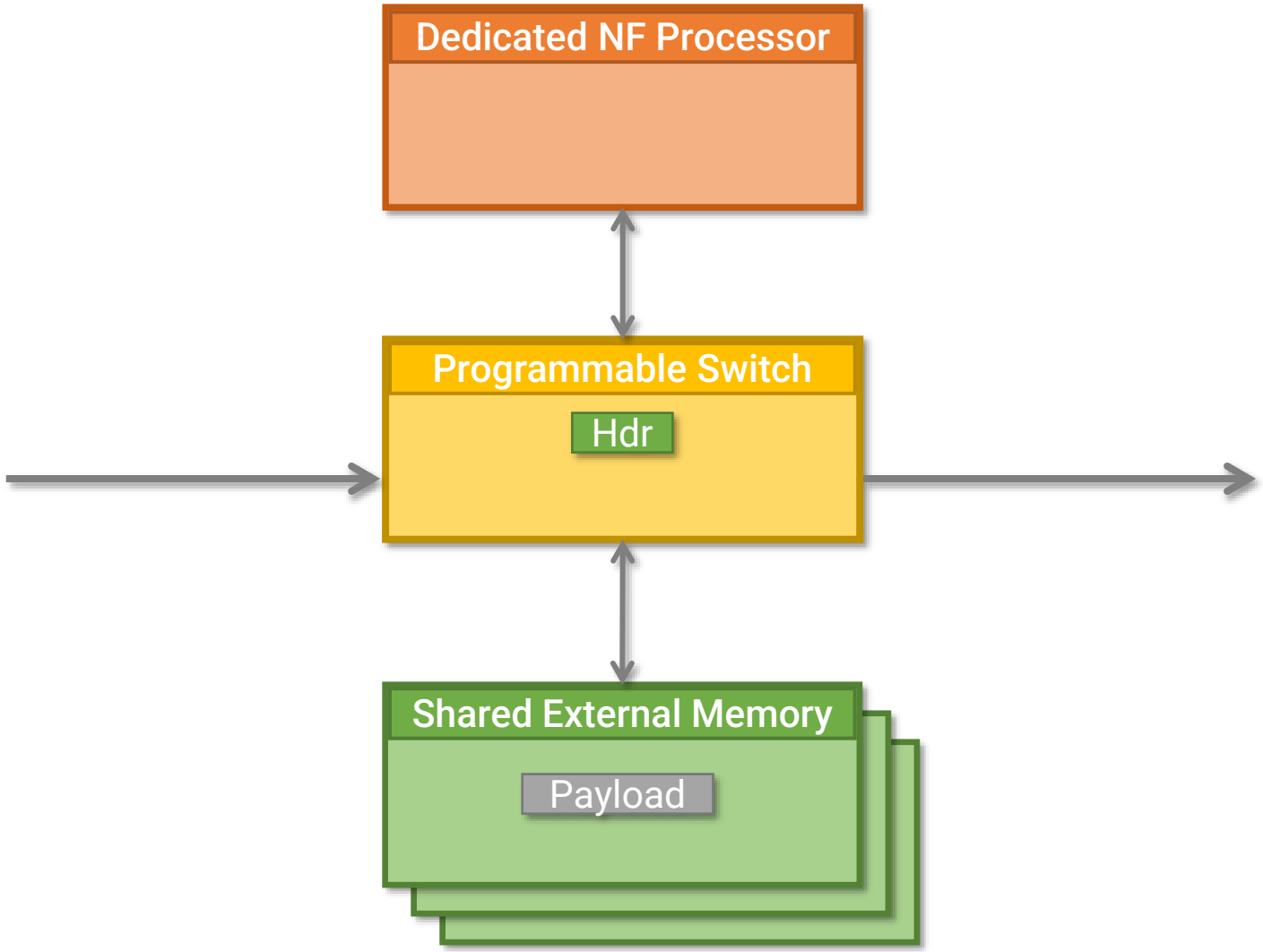
The Ribosome Approach



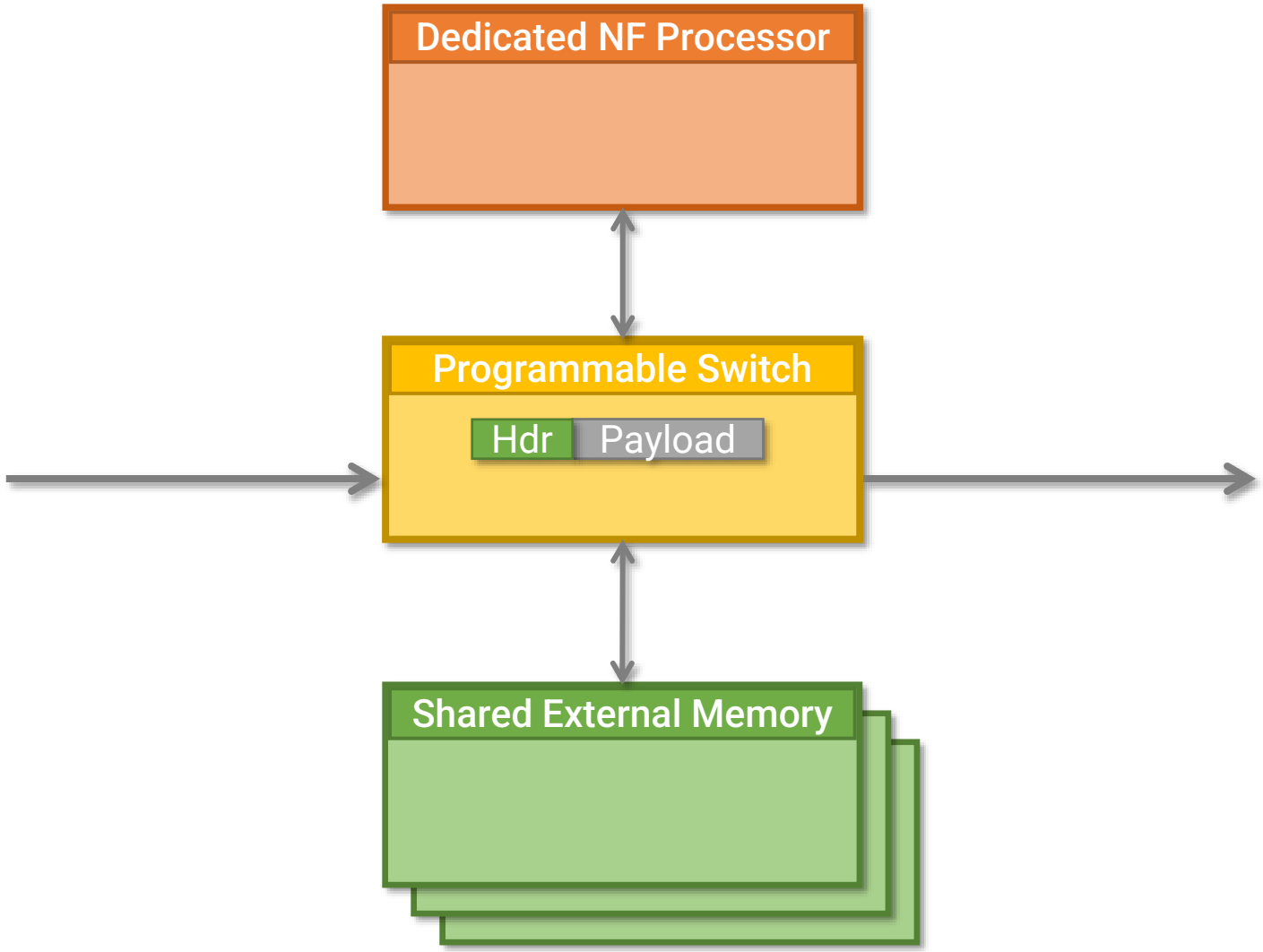
The Ribosome Approach



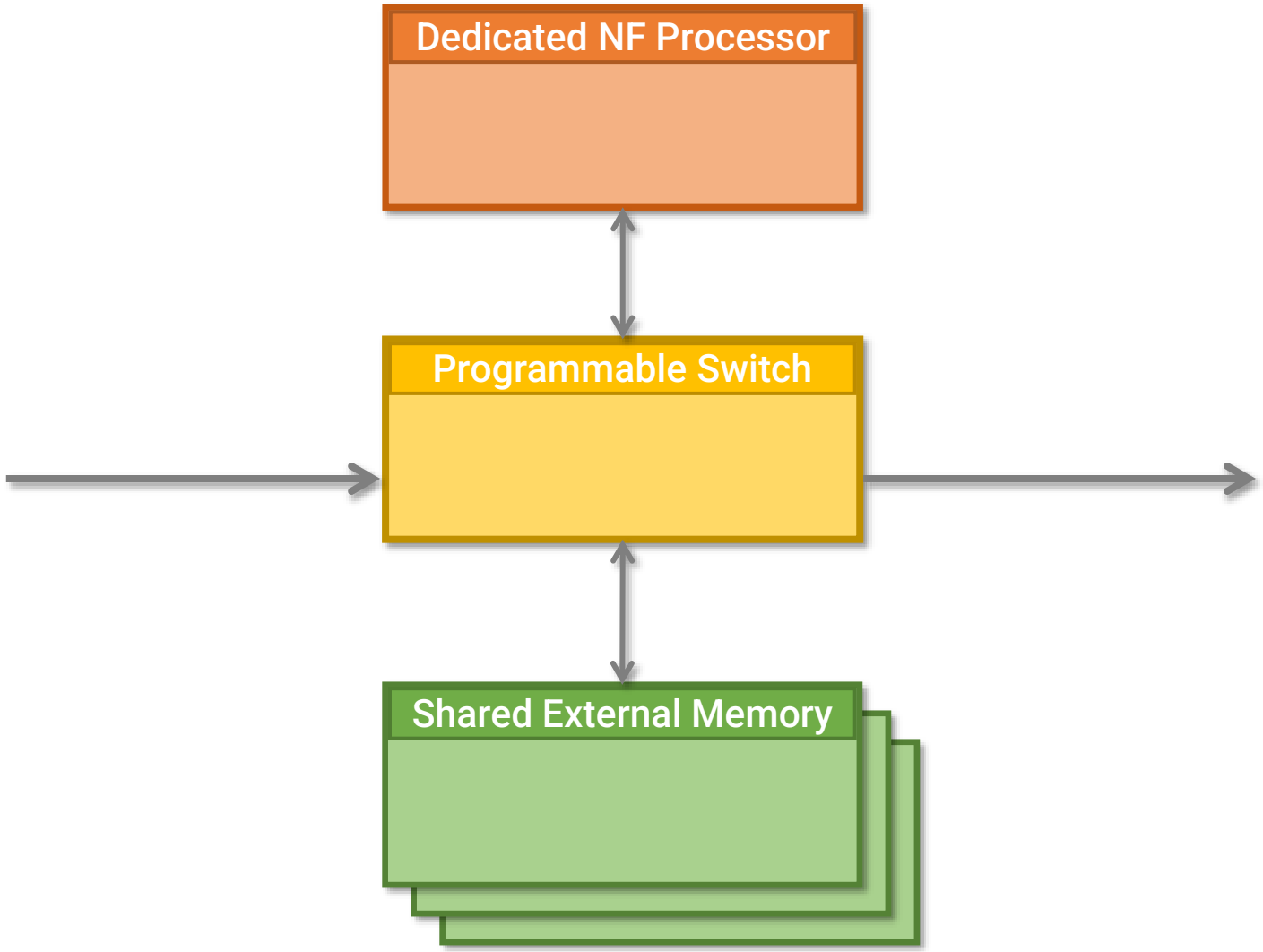
The Ribosome Approach



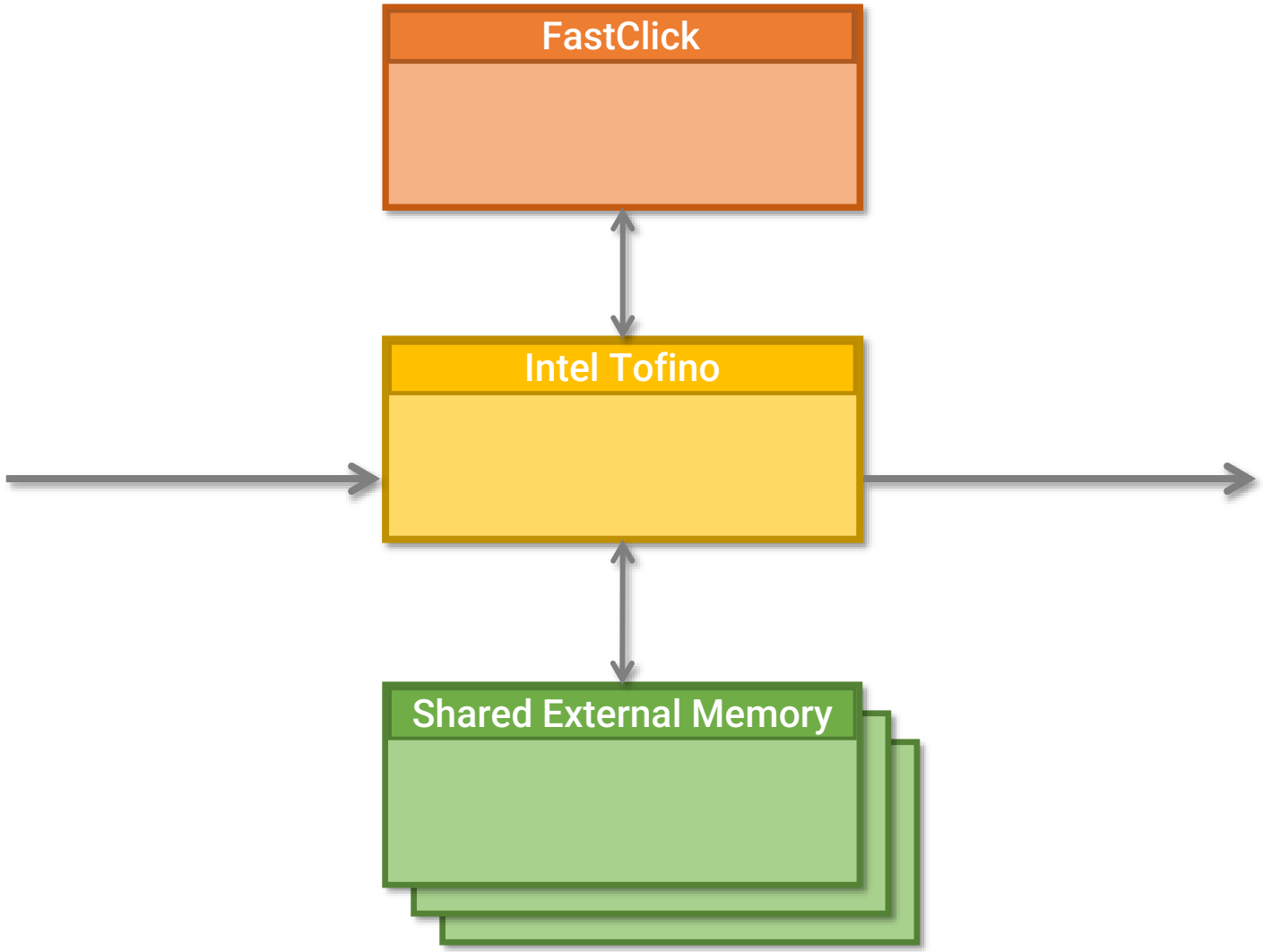
The Ribosome Approach



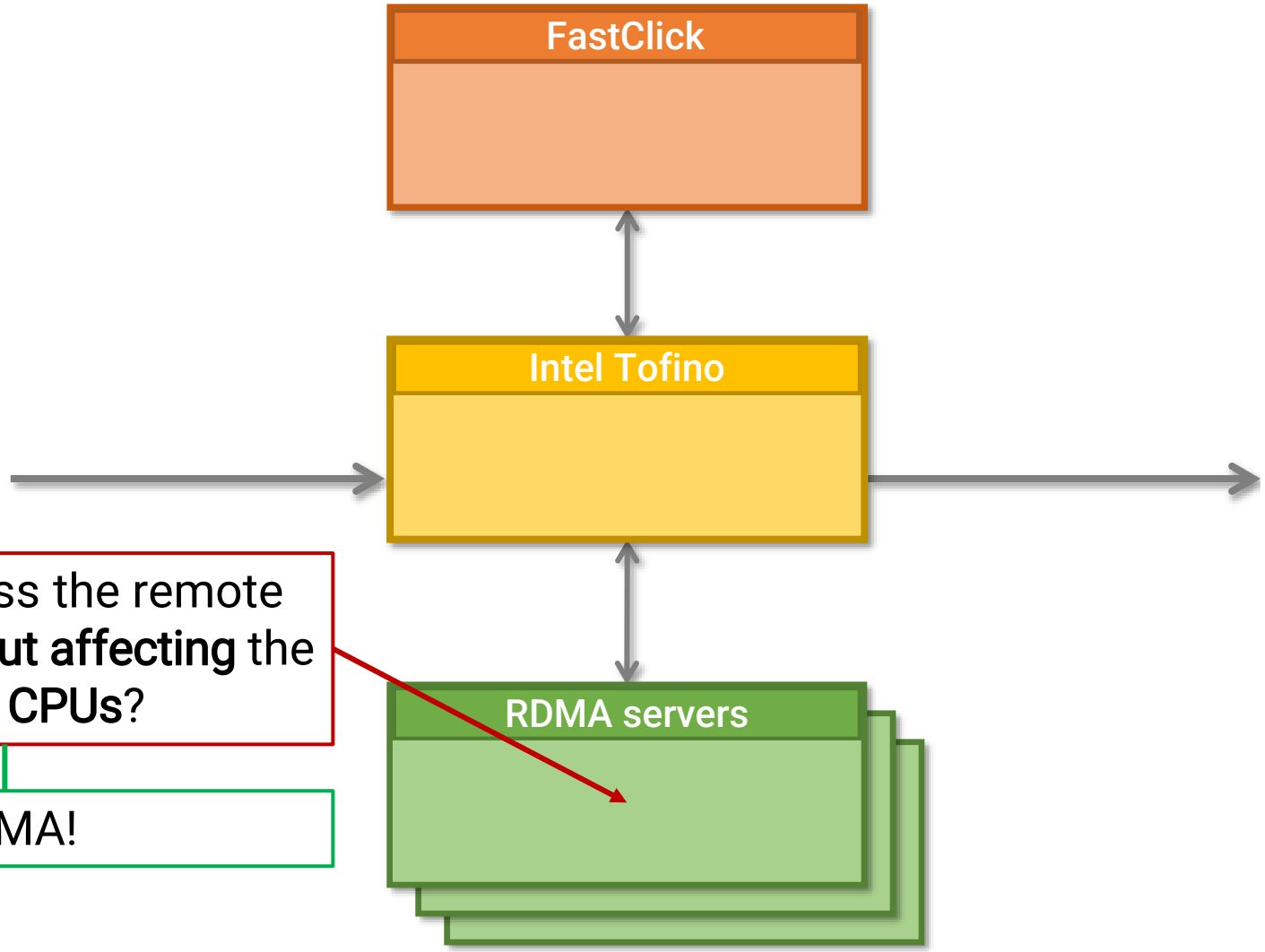
The Ribosome Approach



Implementation



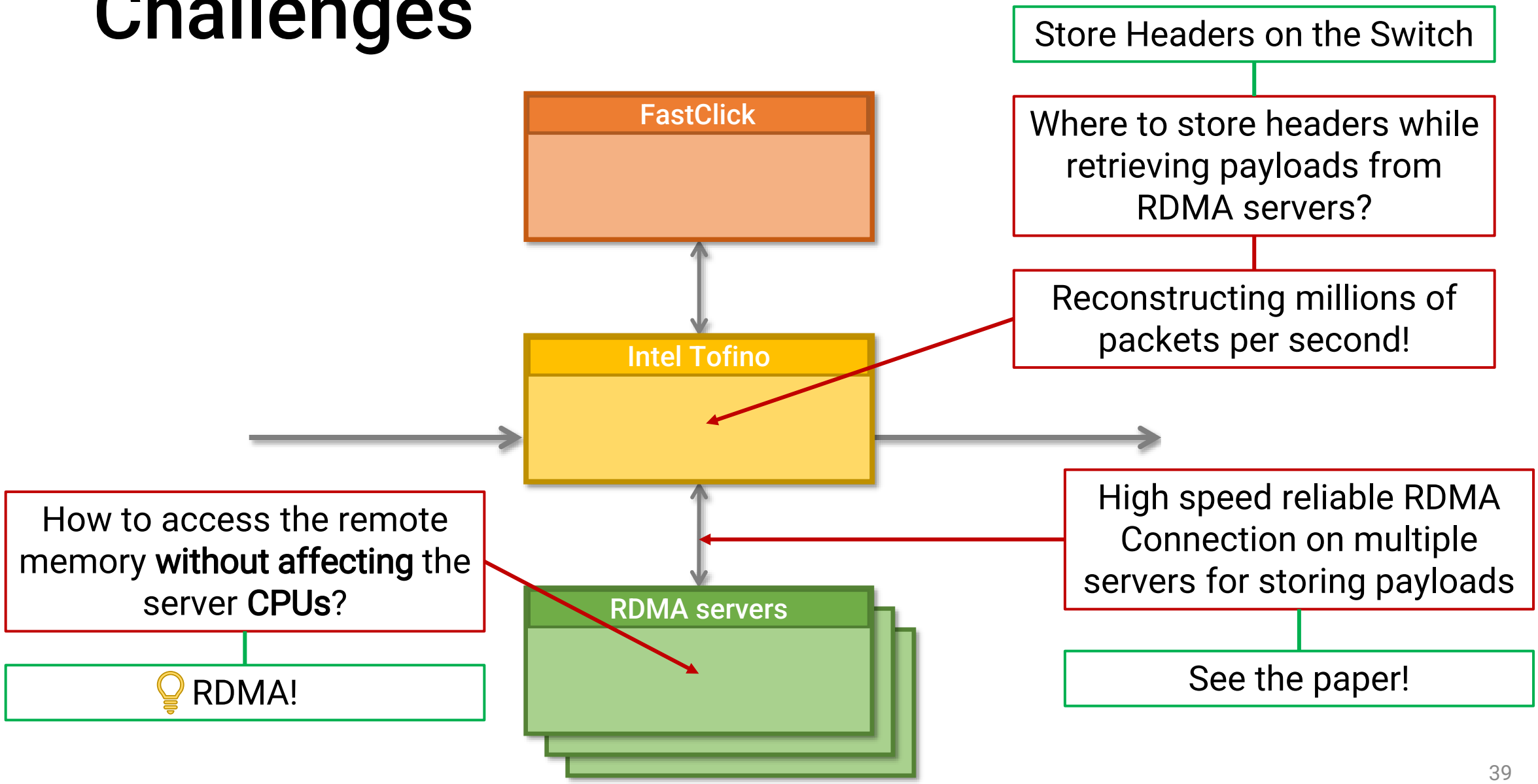
Implementation



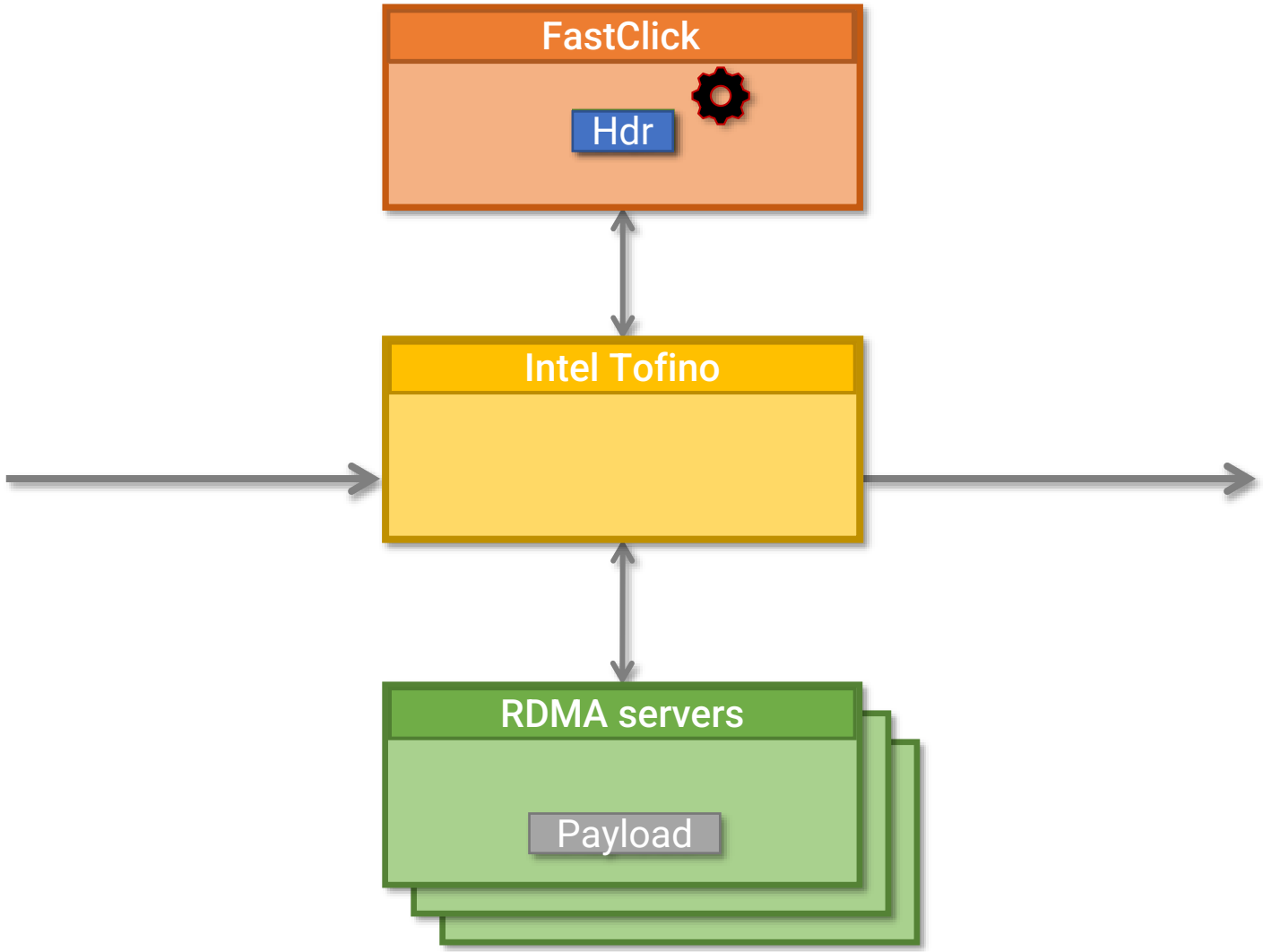
How to access the remote memory **without affecting** the server **CPUs**?

RDMA!

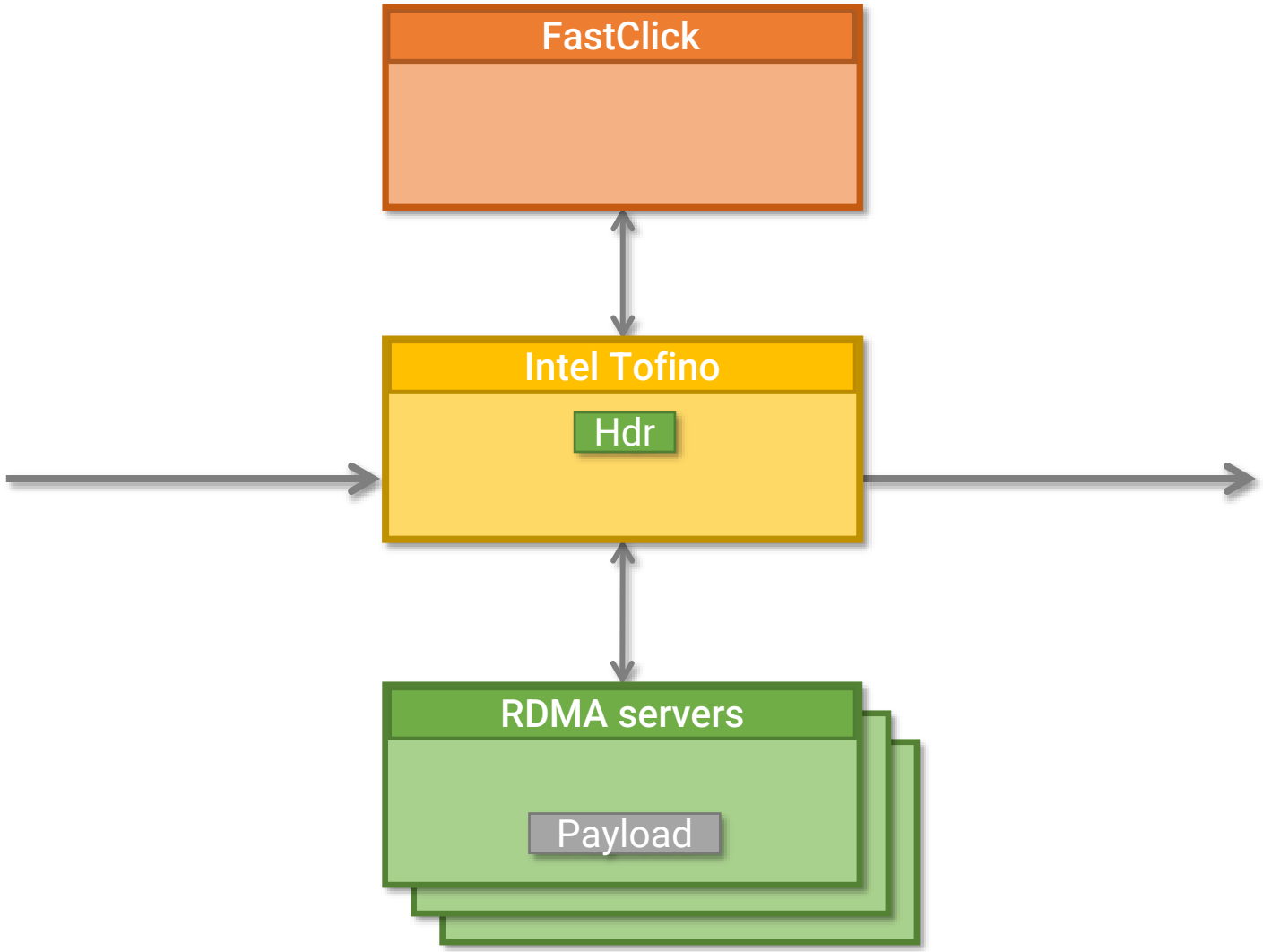
Challenges


Store Headers on the Switch
Where to store headers while retrieving payloads from RDMA servers?
Reconstructing millions of packets per second!
FastClick
Intel Tofino
RDMA servers
How to access the remote memory **without affecting** the server **CPUs**?
RDMA!
High speed reliable RDMA Connection on multiple servers for storing payloads
See the paper!

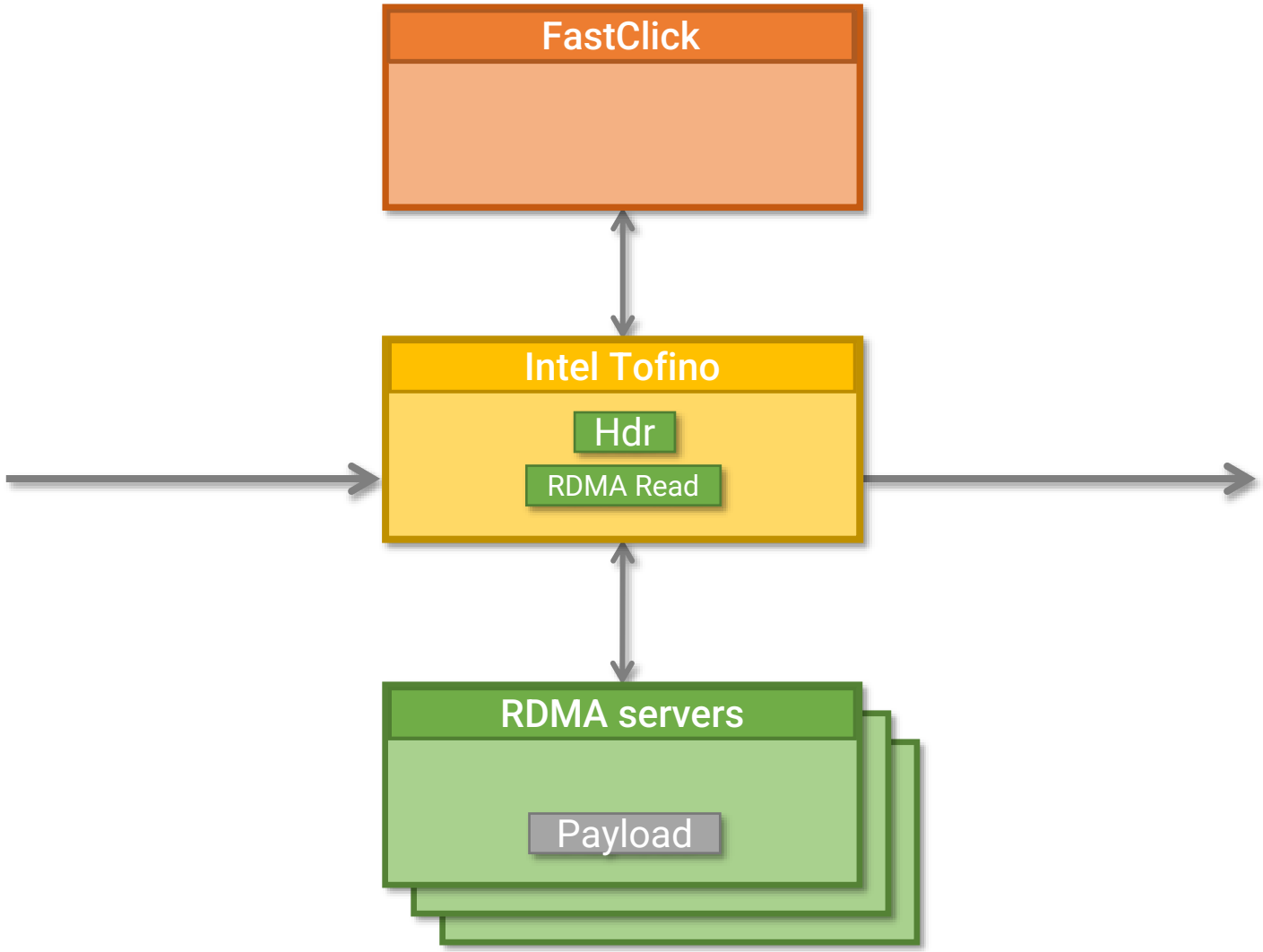
Store Headers on the Switch



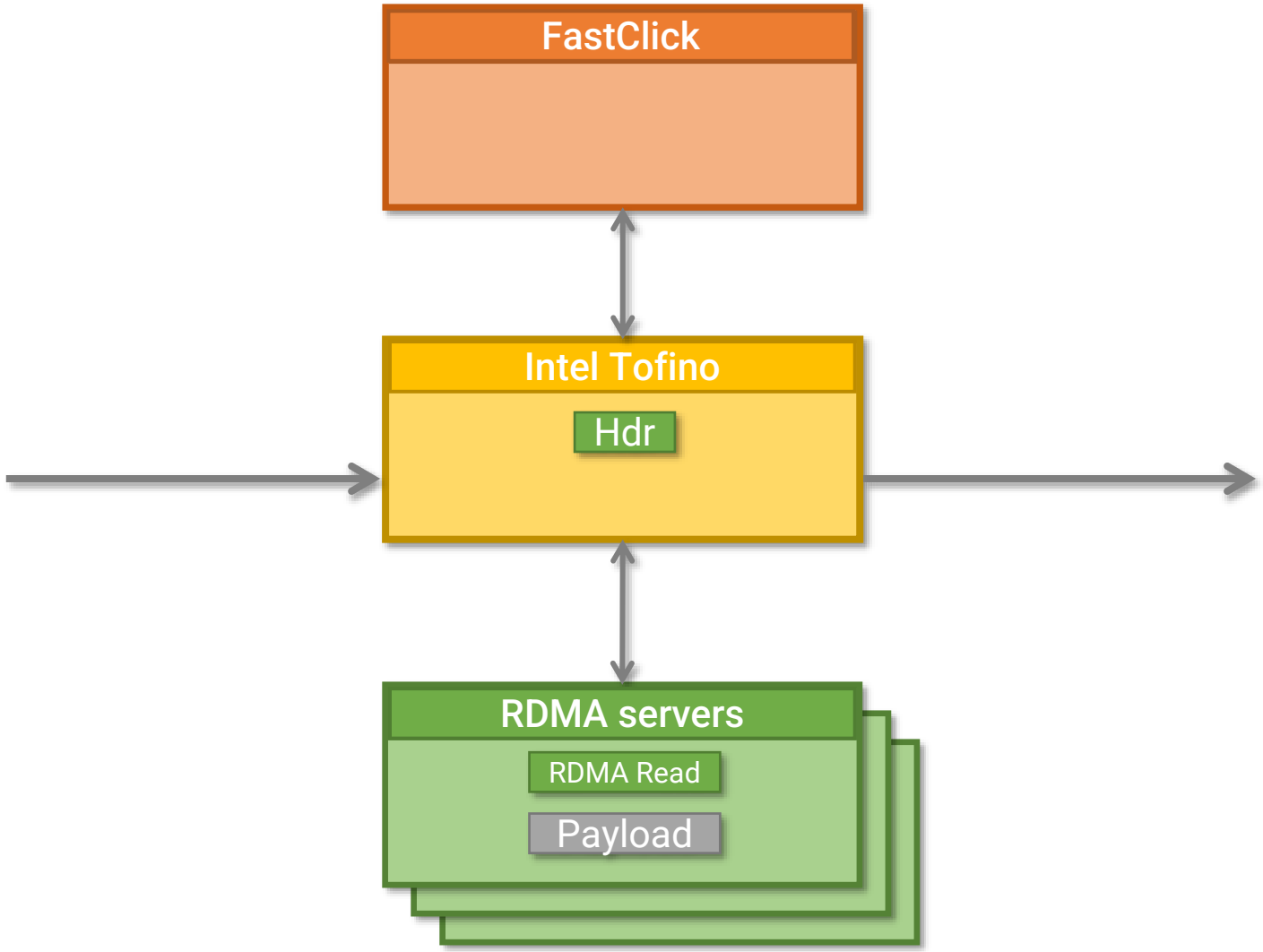
Store Headers on the Switch



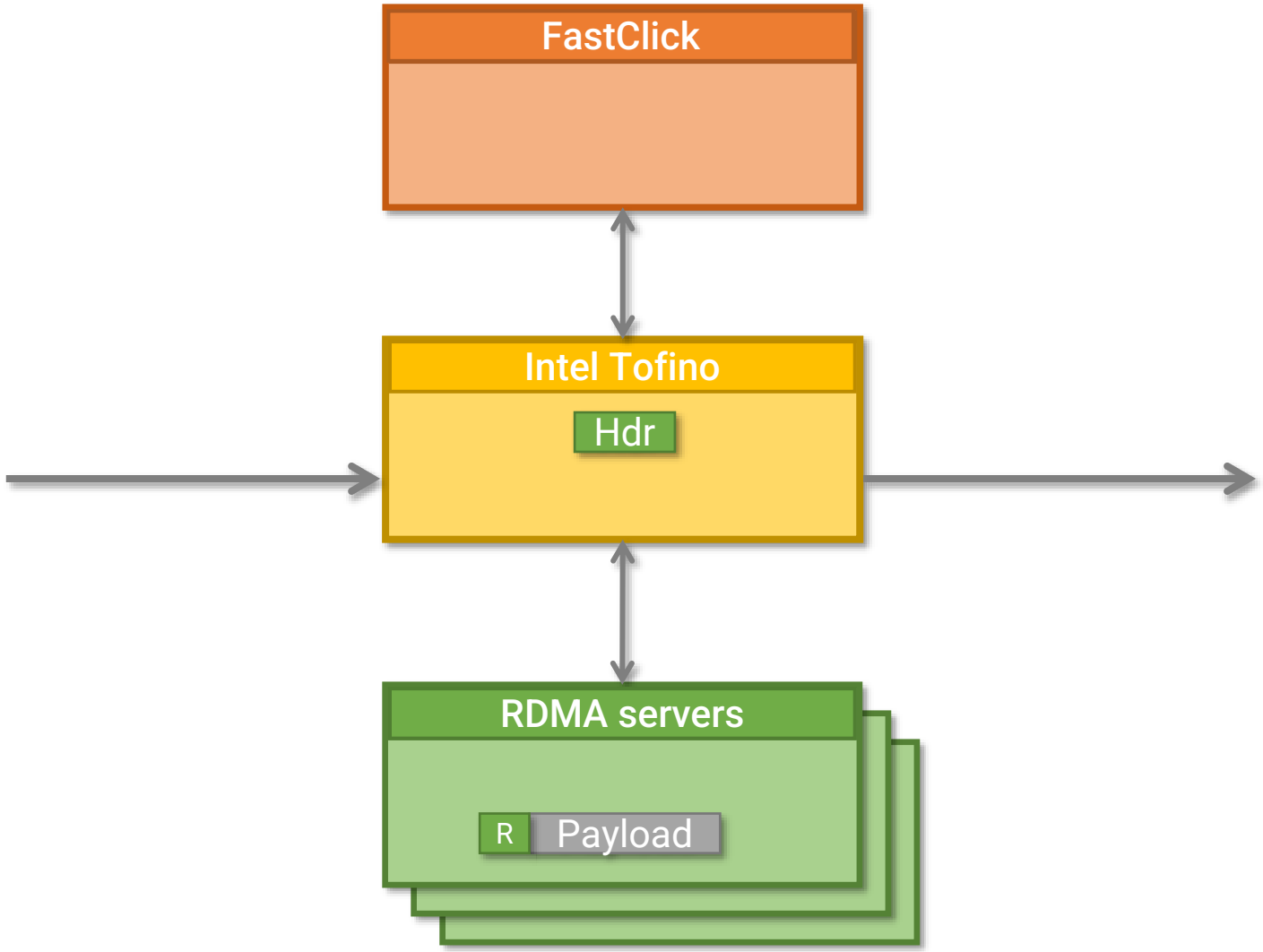
Store Headers on the Switch



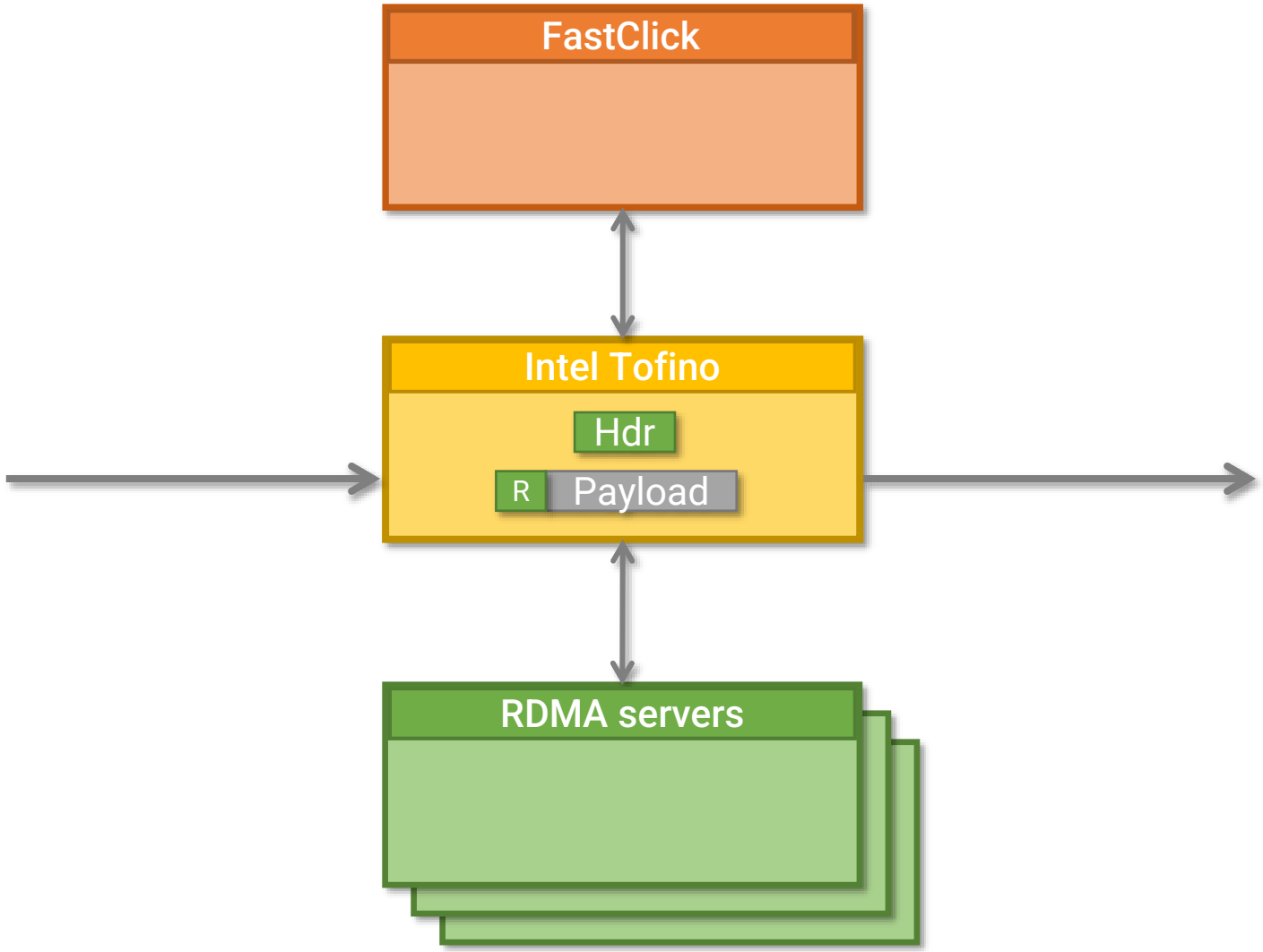
Store Headers on the Switch



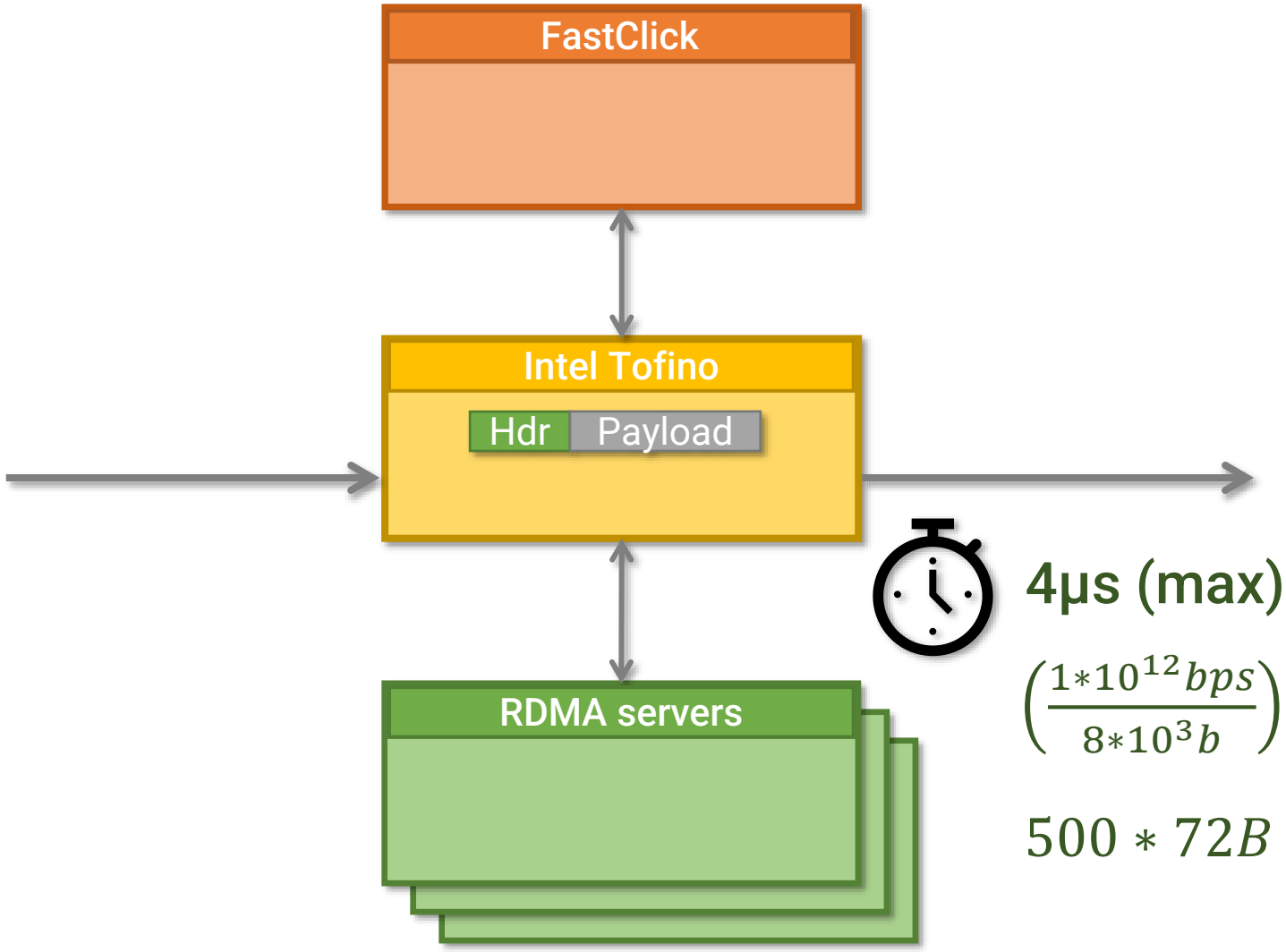
Store Headers on the Switch



Store Headers on the Switch



Store Headers on the Switch



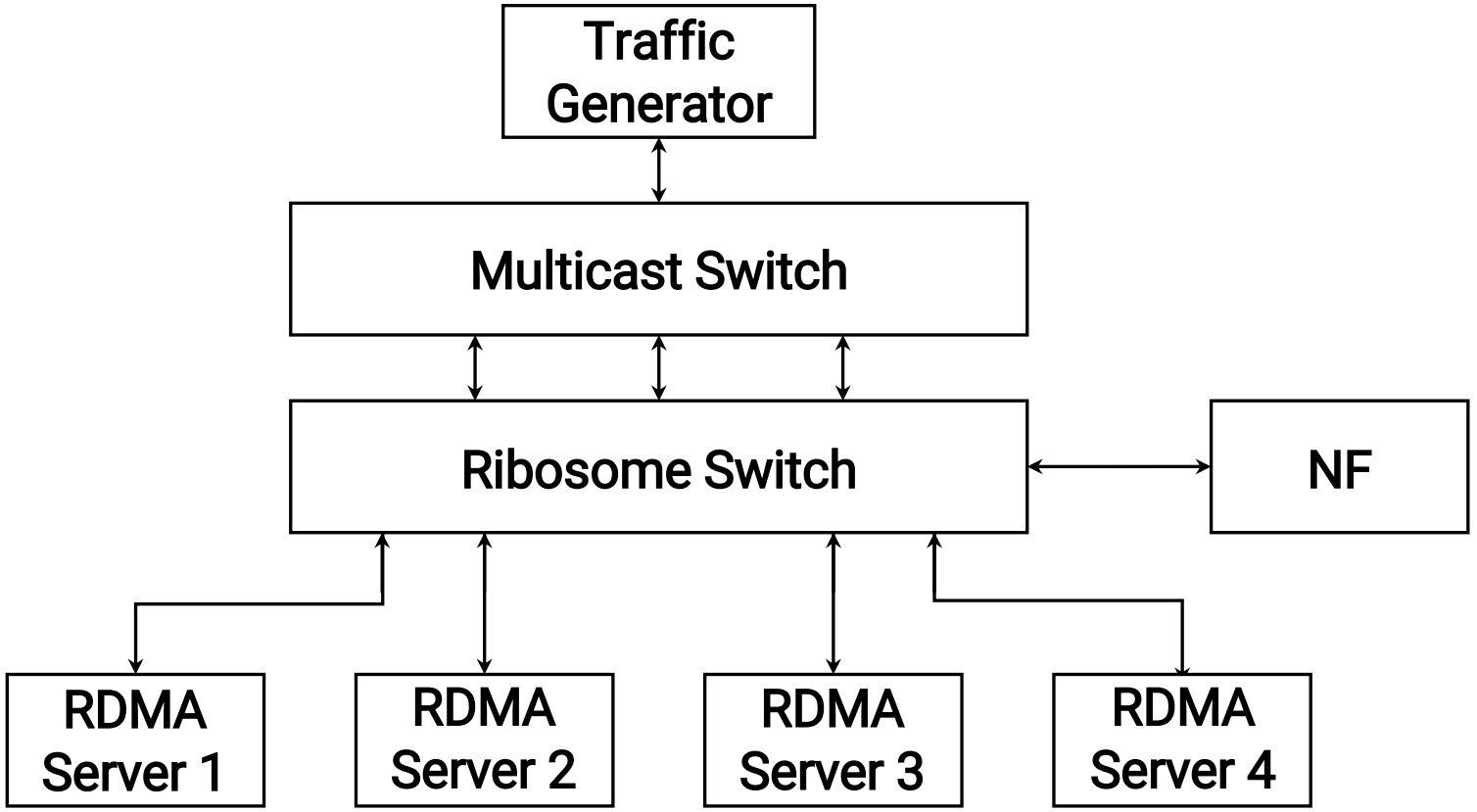
4µs (max)

$\left(\frac{1*10^{12}bps}{8*10^{3}b}\right) * 4\mu s = 500$ headers

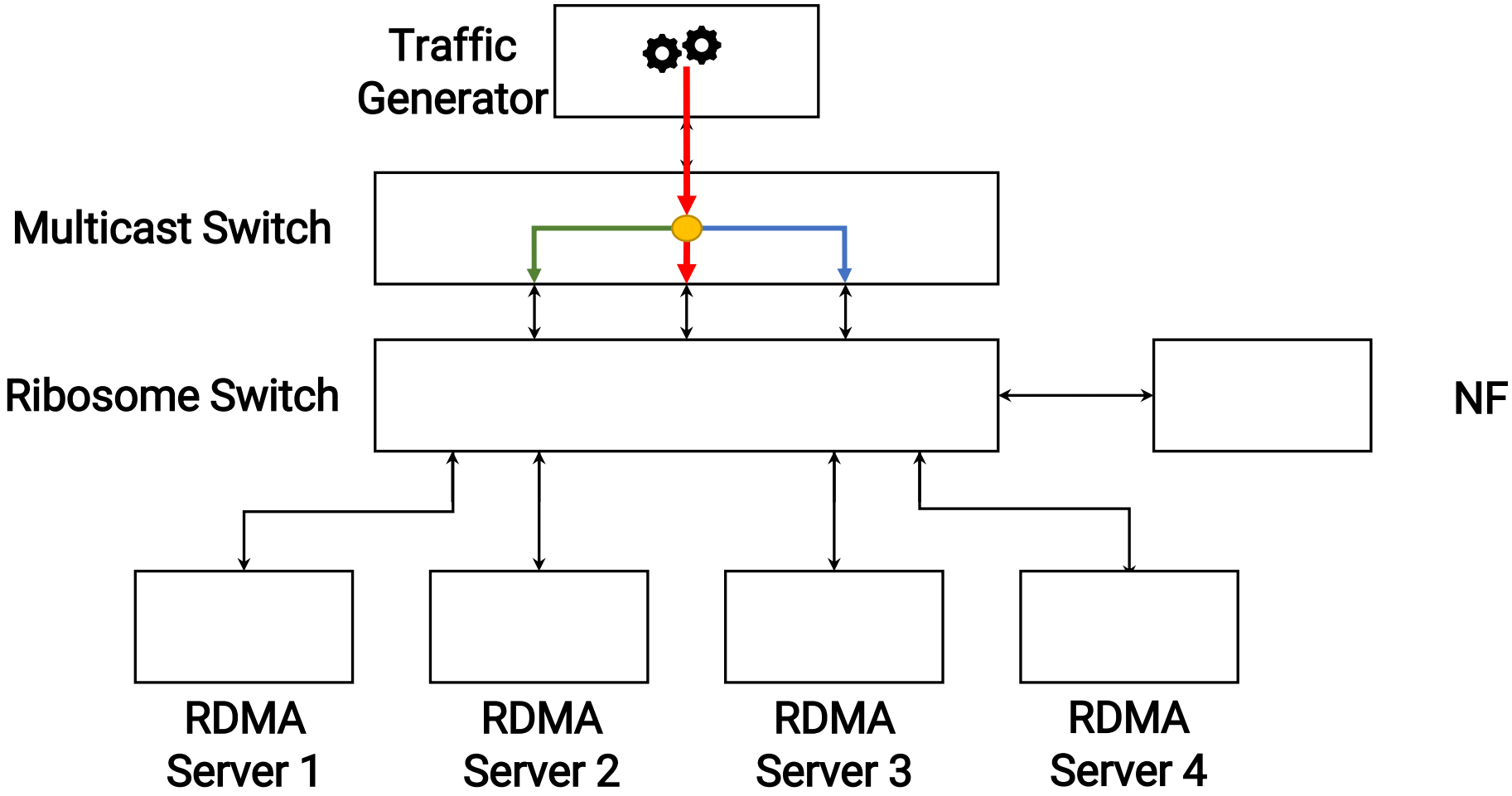
$500 * 72B = 36000B$

Evaluation

Testbed and Workload Generation

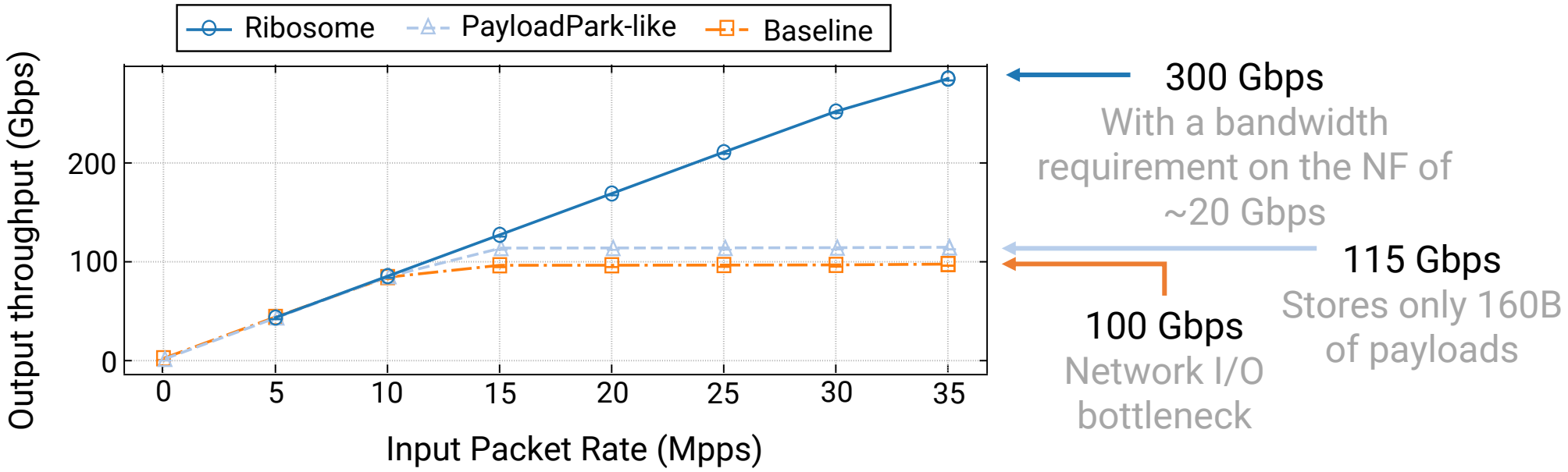


Testbed and Workload Generation



Throughput Gain

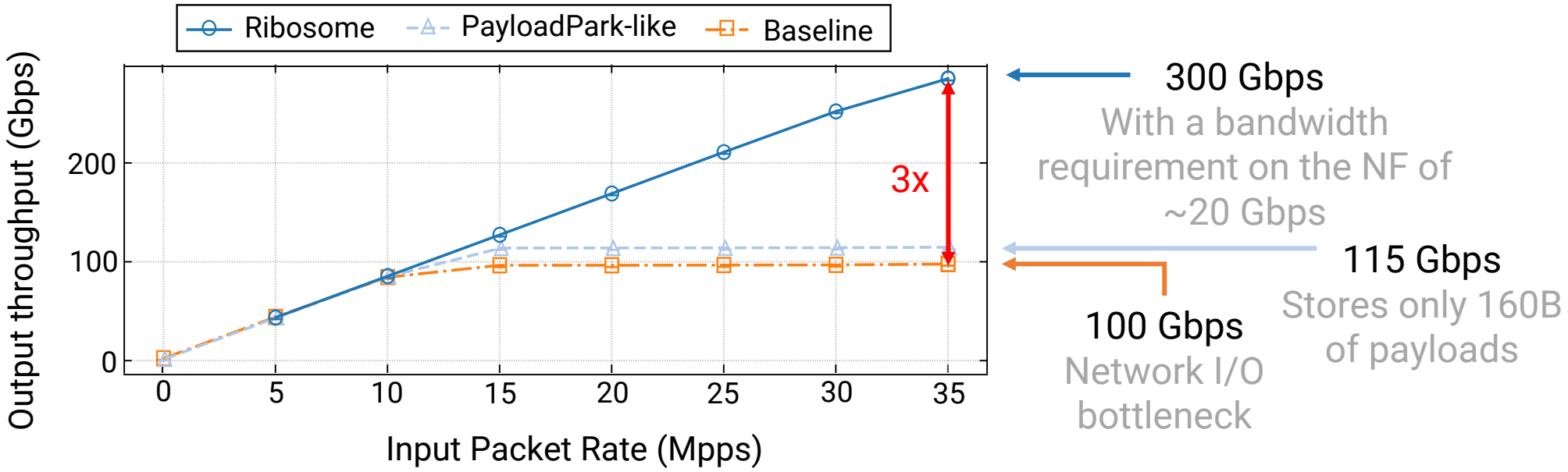
How much Ribosome improves the per-packet throughput on the NF server?



Tested NF: Forwarder

Throughput Gain

How much Ribosome improves the per-packet throughput on the NF server?

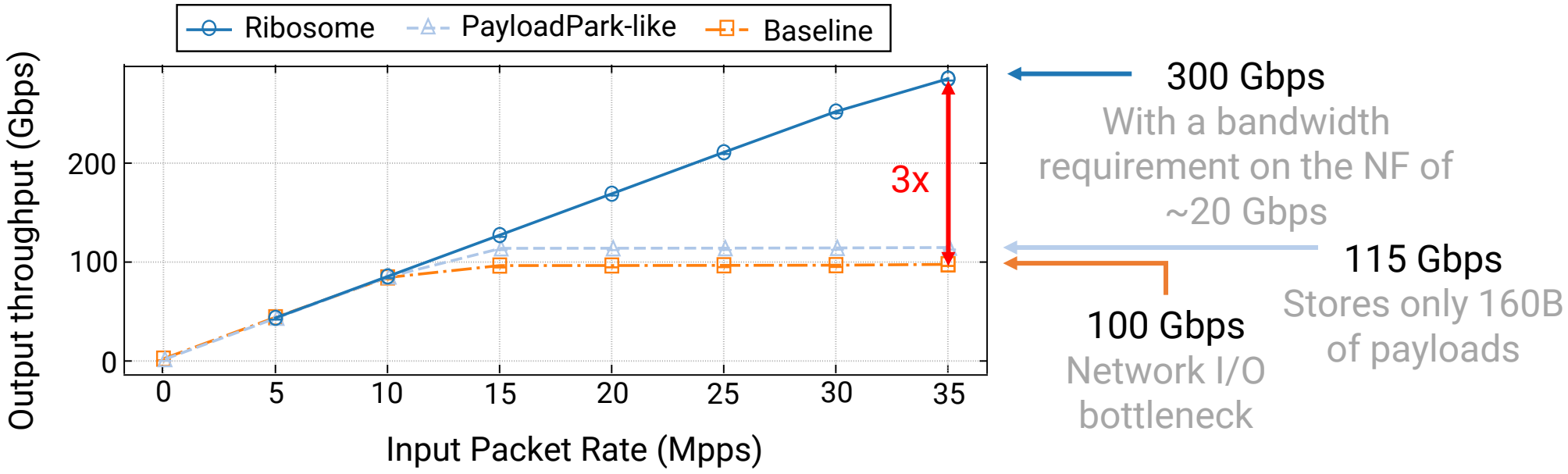


Tested NF: Forwarder

Throughput Gain

How much Ribosome improves the per-packet throughput on the NF server?

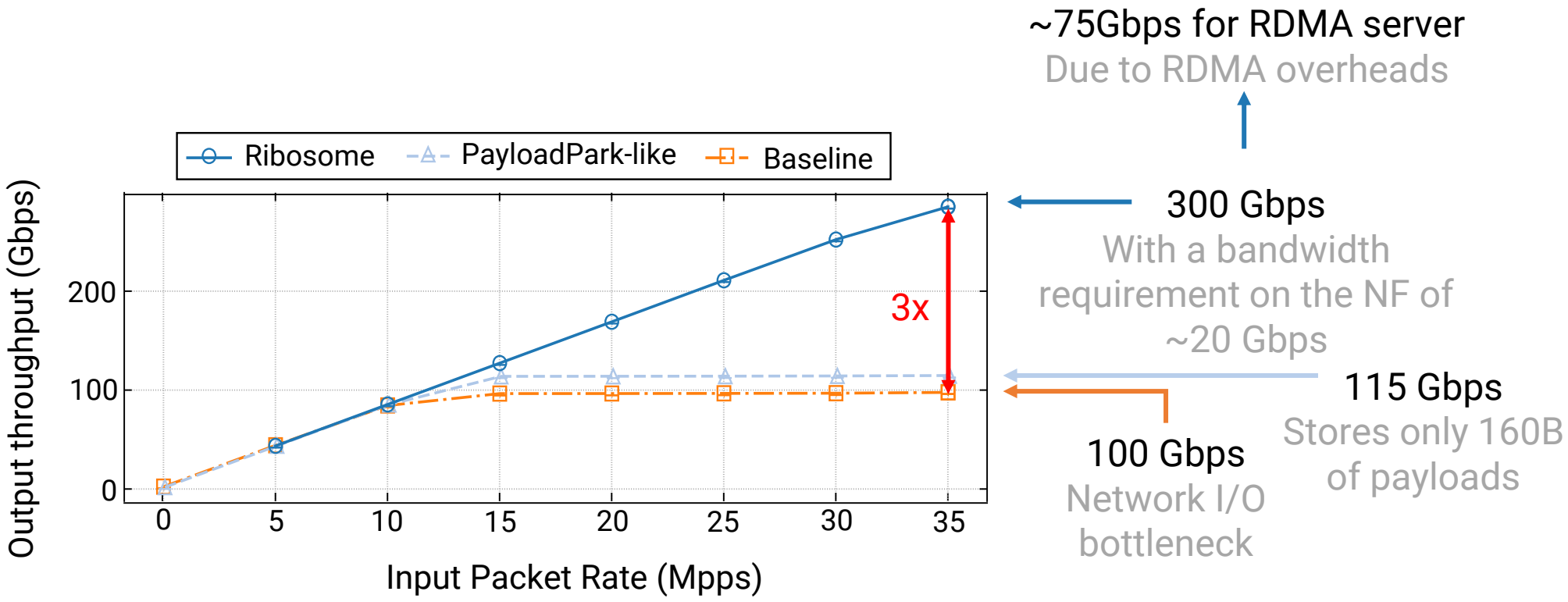
Ribosome enables multi-100Gbps packet processing!



Throughput Gain

How much Ribosome improves the per-packet throughput on the NF server?

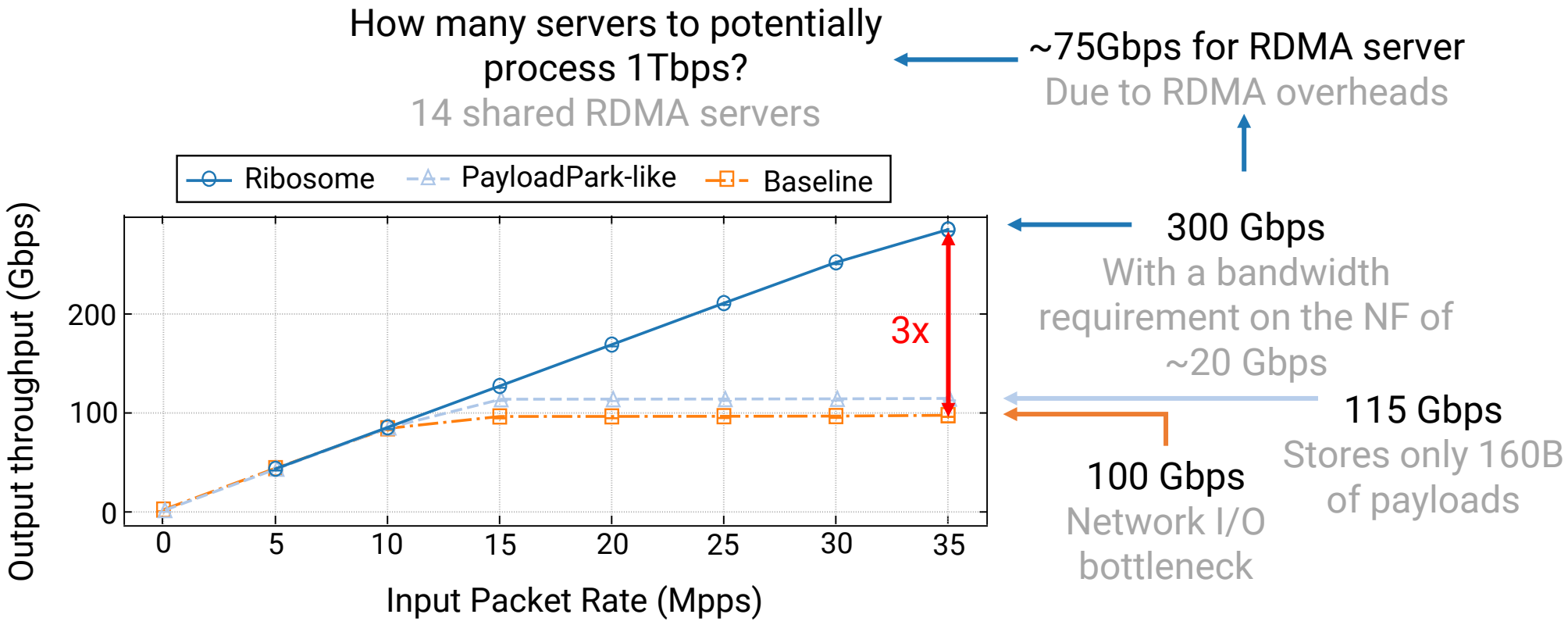
Ribosome enables multi-100Gbps packet processing!



Throughput Gain

How much Ribosome improves the per-packet throughput on the NF server?

Ribosome enables multi-100Gbps packet processing!



Throughput Gain

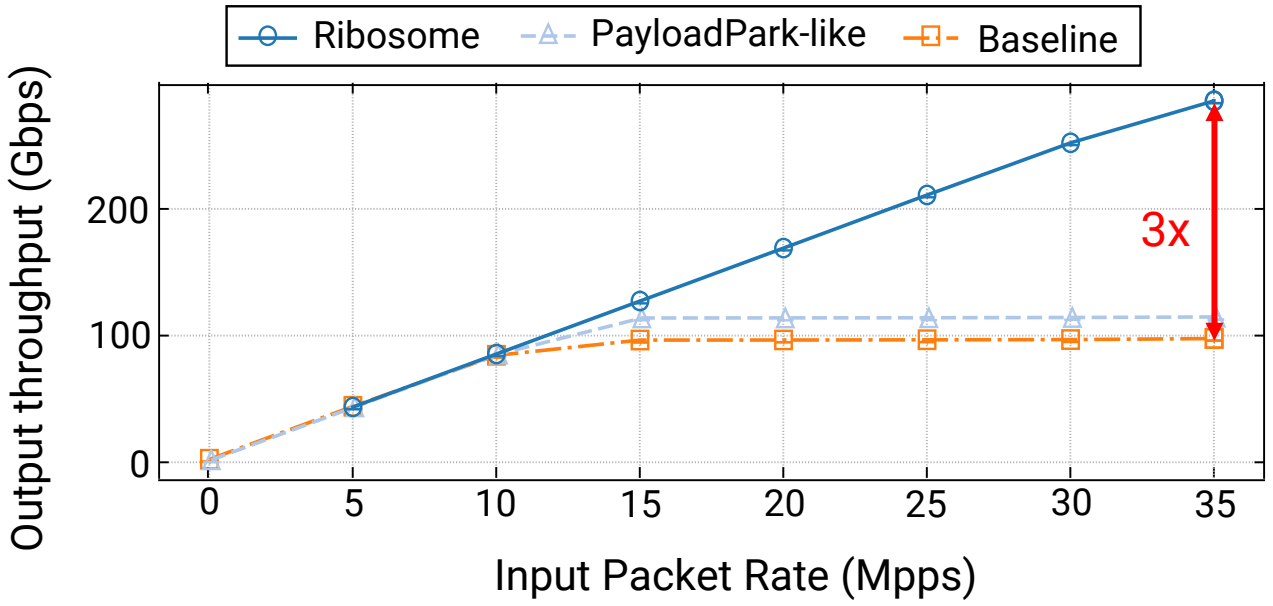
How much Ribosome improves the per-packet throughput on the NF server?

Ribosome enables multi-100Gbps packet processing!

A datacenter has thousands of servers!

How many servers to potentially process 1Tbps?
14 shared RDMA servers

~75Gbps for RDMA server
Due to RDMA overheads



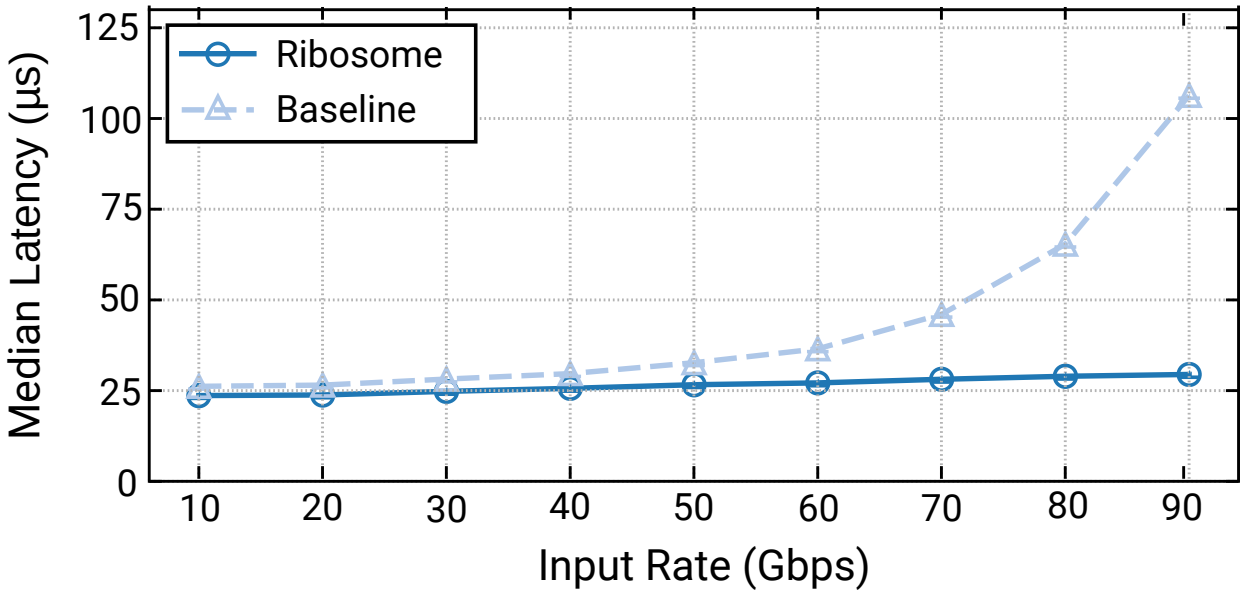
300 Gbps
With a bandwidth requirement on the NF of ~20 Gbps

115 Gbps
Stores only 160B of payloads

100 Gbps
Network I/O bottleneck

Latency Gain

How much Ribosome improves the latency gain on the NF server?



~ 110 µs

Due to **congestioned queues** and the **high trasmission rate**

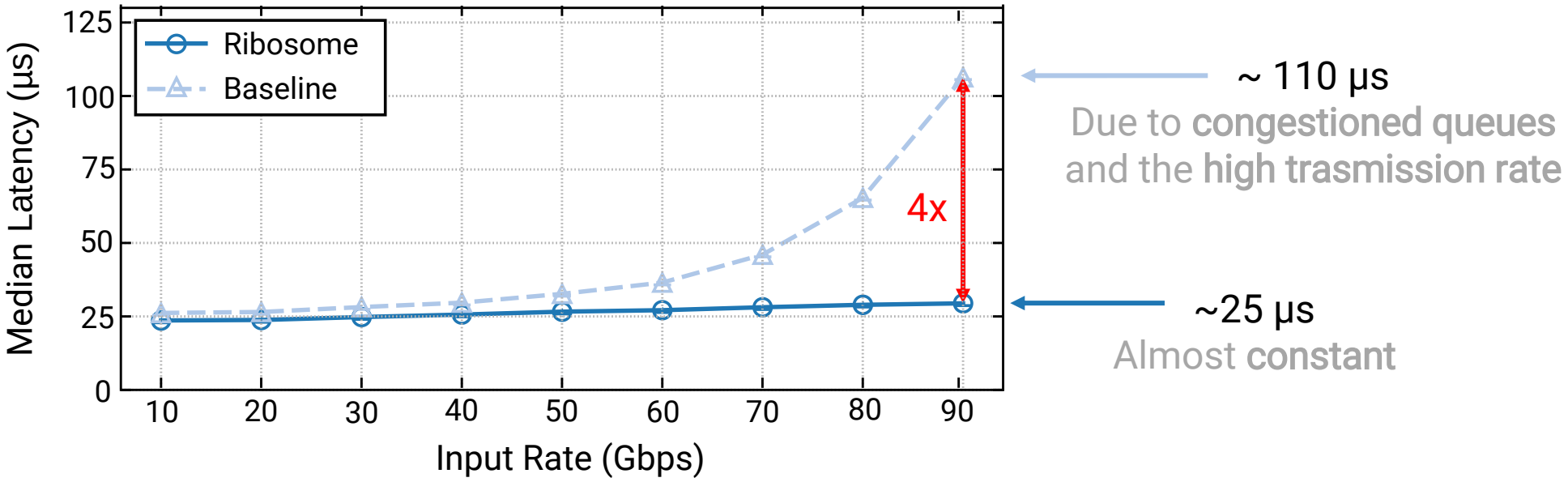
~25 µs

Almost **constant**

Tested NF: Forwarder

Latency Gain

How much Ribosome improves the latency gain on the NF server?

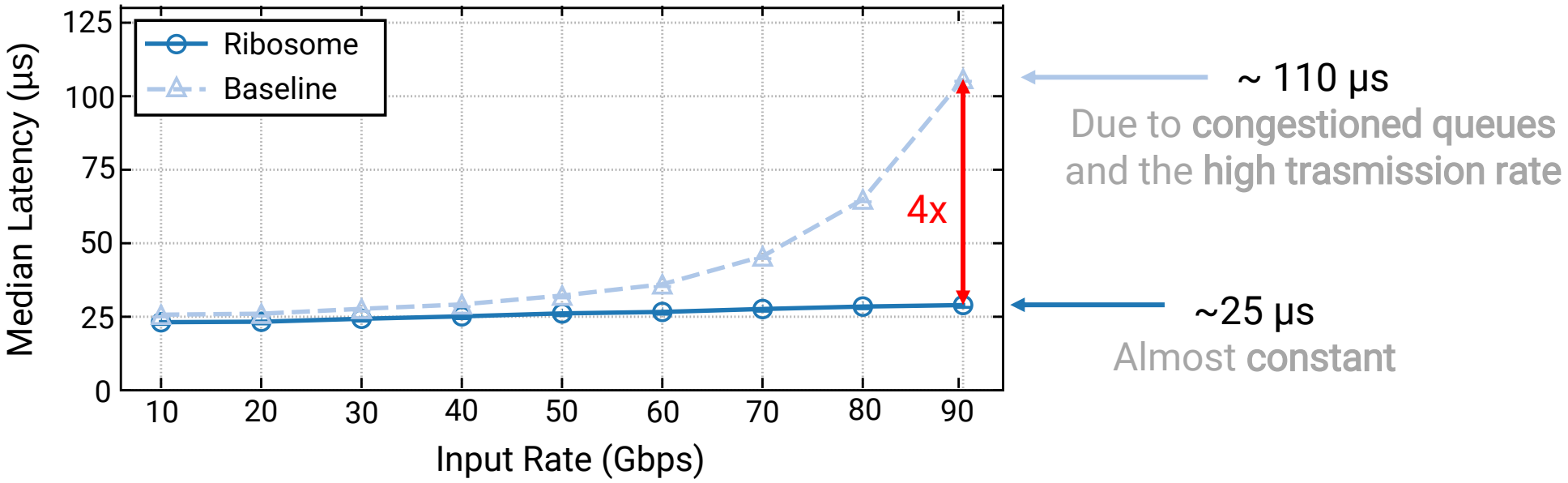


Tested NF: Forwarder

Latency Gain

How much Ribosome improves the latency gain on the NF server?

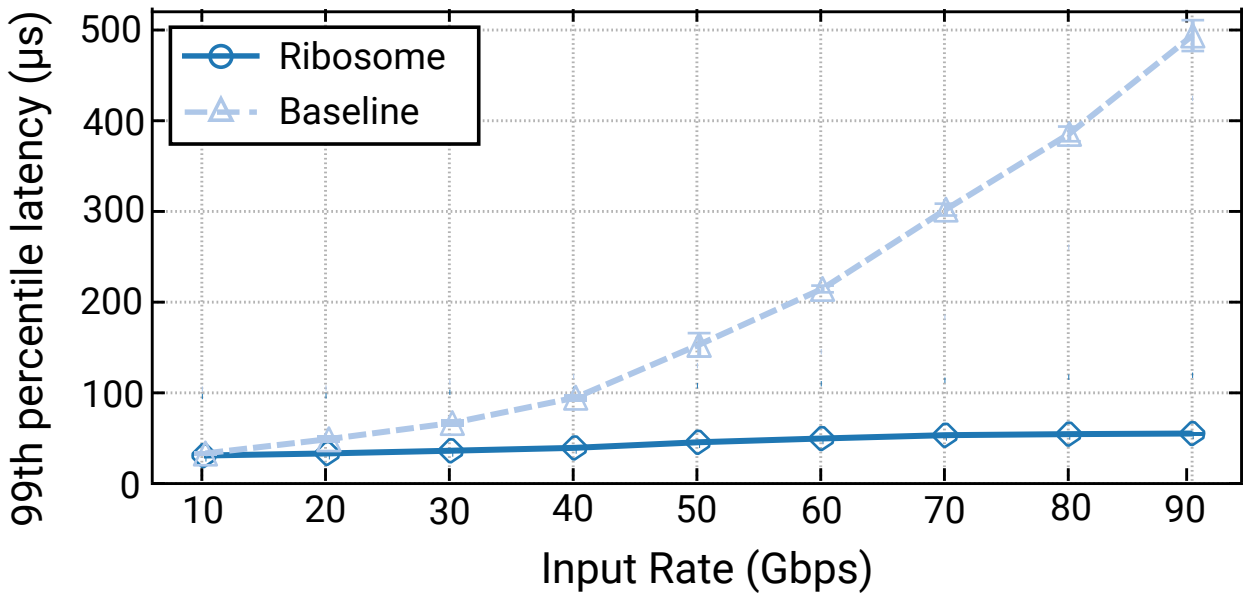
And the tail latency? ⟶ Similar trend!



Latency Gain

How much Ribosome improves the latency gain on the NF server?

And the tail latency? ⟶ Similar trend!



~ 500 µs

Due to **congestioned queues** and the **high trasmission rate**

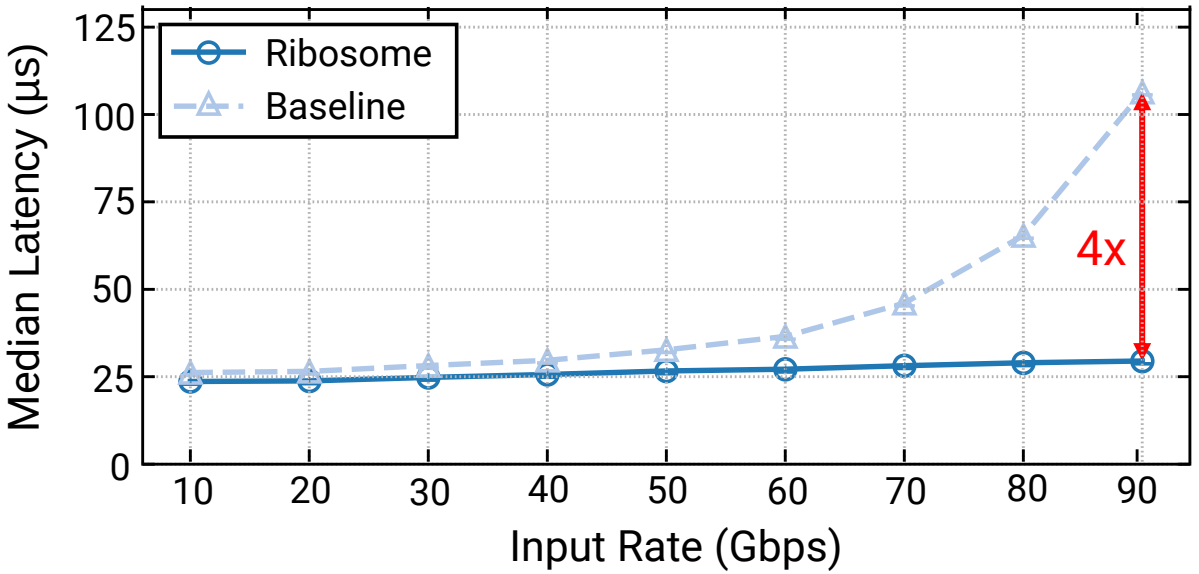
~60 µs

Almost **constant**

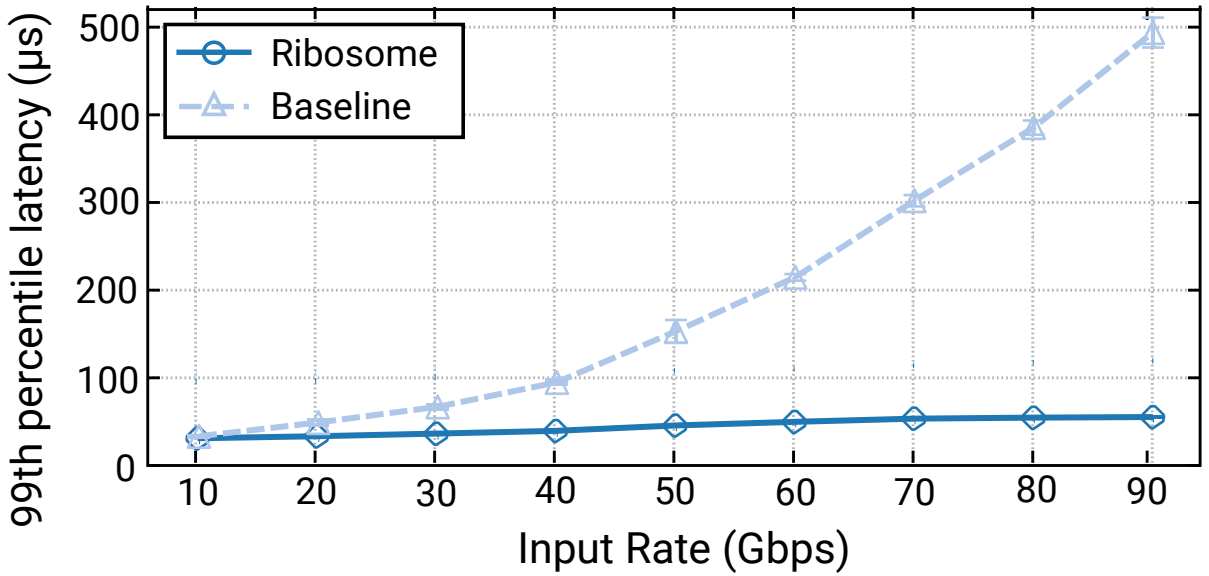
Latency Gain

How much Ribosome improves the latency gain on the NF server?

Reducing queue sizes and the input throughput on the NF reduce latency!

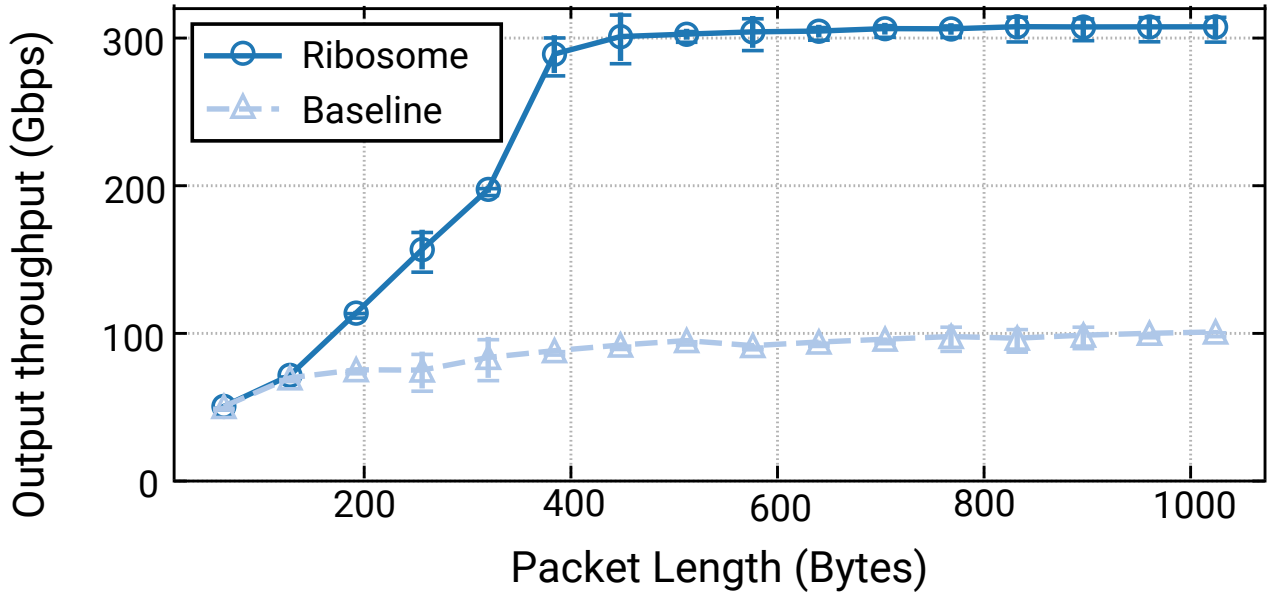


And the tail latency? → Similar trend!



Packet Size Impact

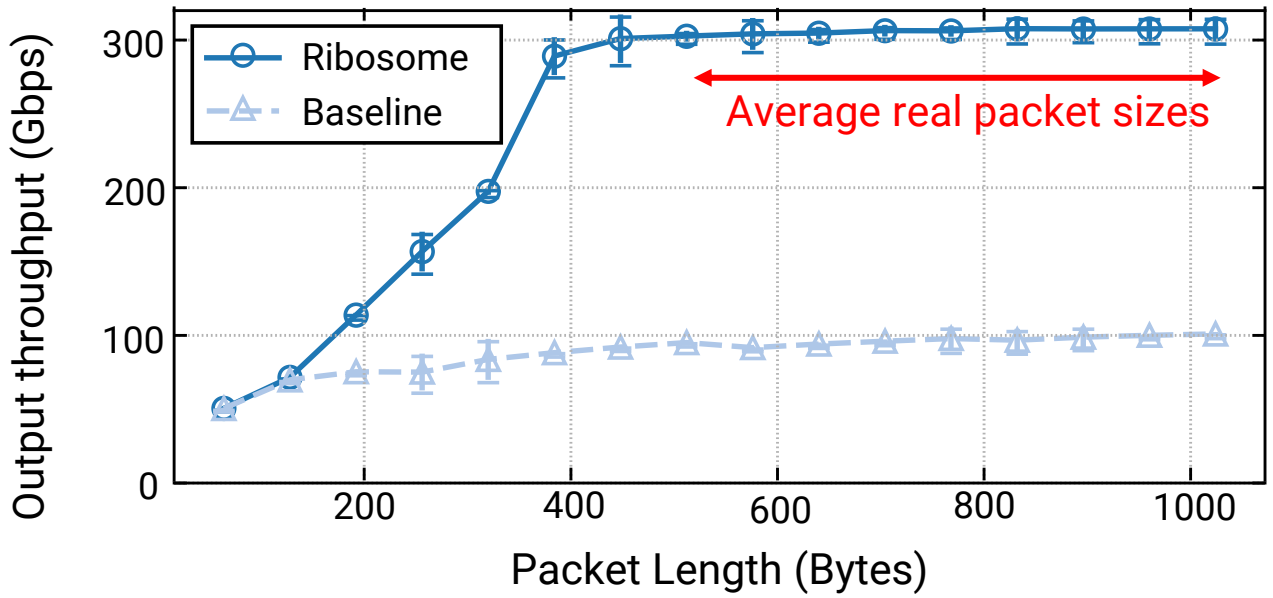
How does the packet size impact the throughput gains?



Tested NF: Forwarder

Packet Size Impact

How does the packet size impact the throughput gains?

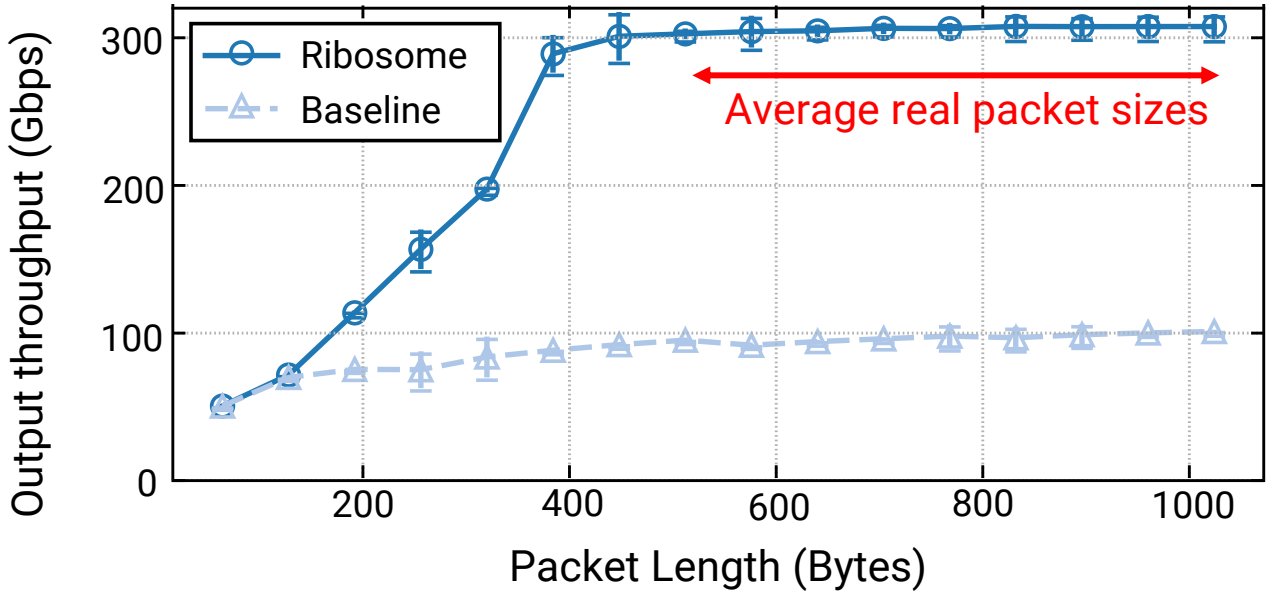


Tested NF: Forwarder

Packet Size Impact

How does the packet size impact the throughput gains?

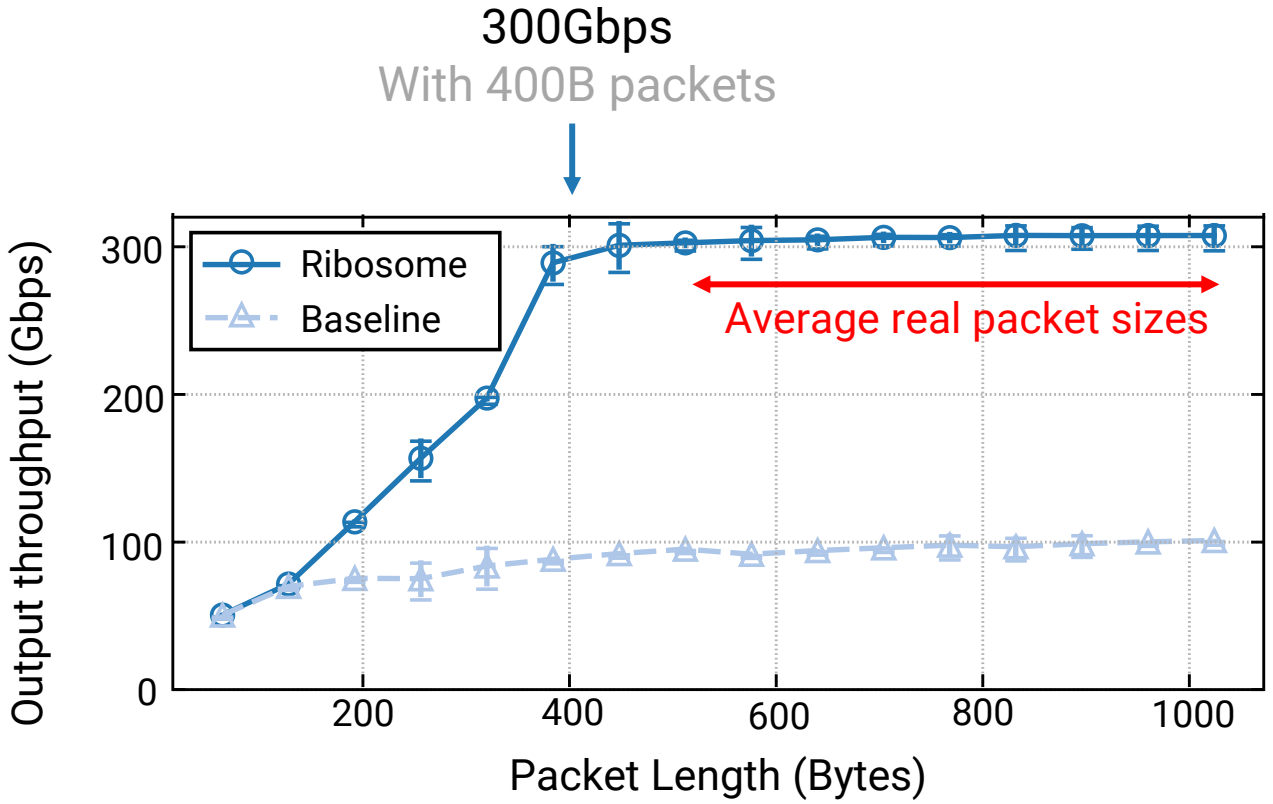
Highly effective for relevant real-world scenarios!



Packet Size Impact

How does the packet size impact the throughput gains?

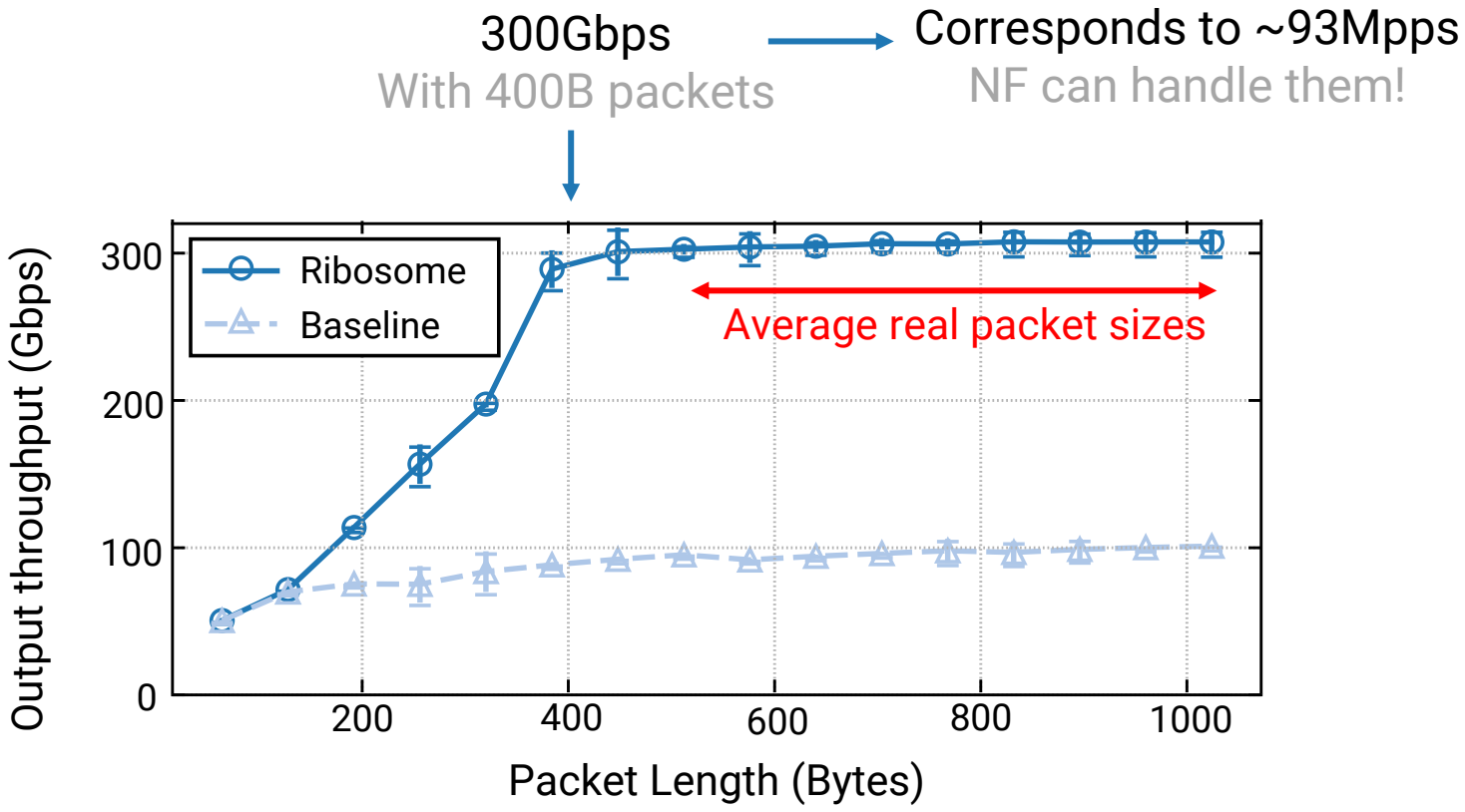
Highly effective for relevant real-world scenarios!



Packet Size Impact

How does the packet size impact the throughput gains?

Highly effective for relevant real-world scenarios!



Packet Size Impact

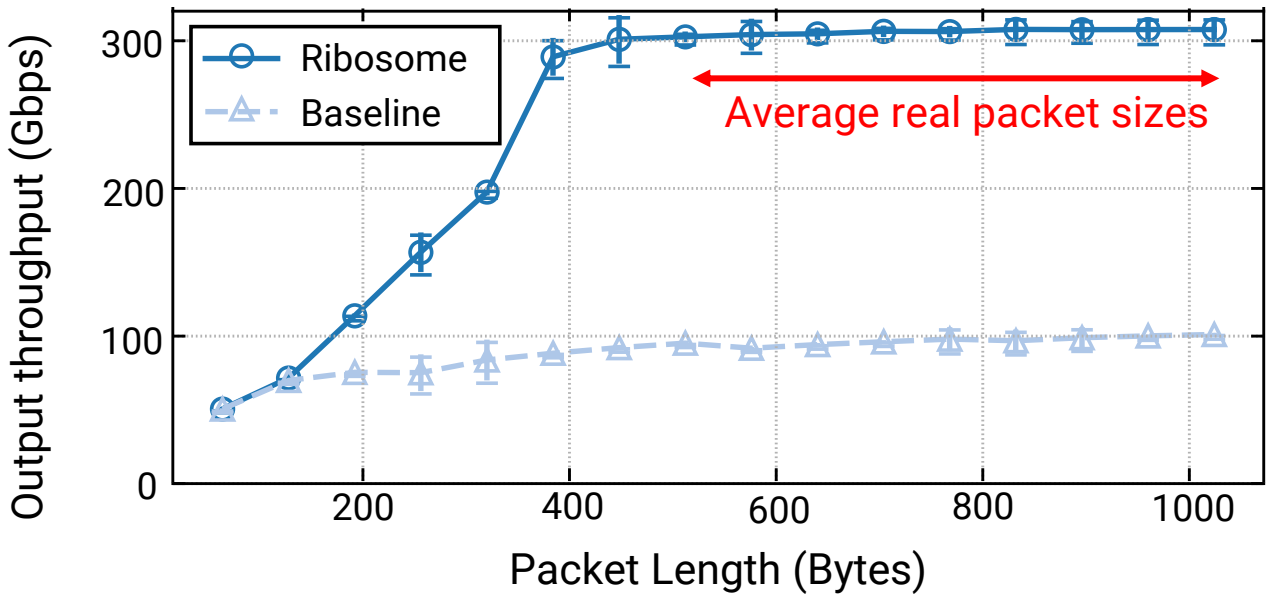
How does the packet size impact the throughput gains?

Highly effective for relevant real-world scenarios!

300Gbps
With 400B packets

→ Corresponds to ~93Mpps
NF can handle them!

→ With 93M packets of 1.5KB
We process 1Tbps!



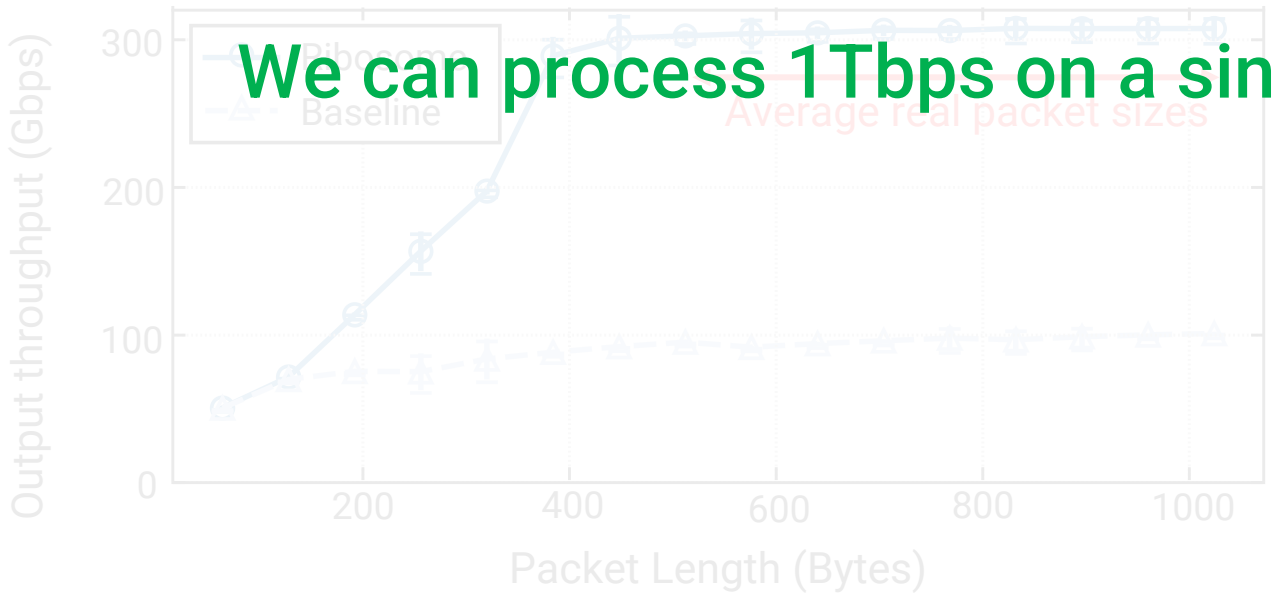
Packet Size Impact

How does the packet size impact the throughput gains?

Highly effective for relevant real-world scenarios!

300Gbps → Corresponds to ~93Mpps → With 93M packets of 1.5KB

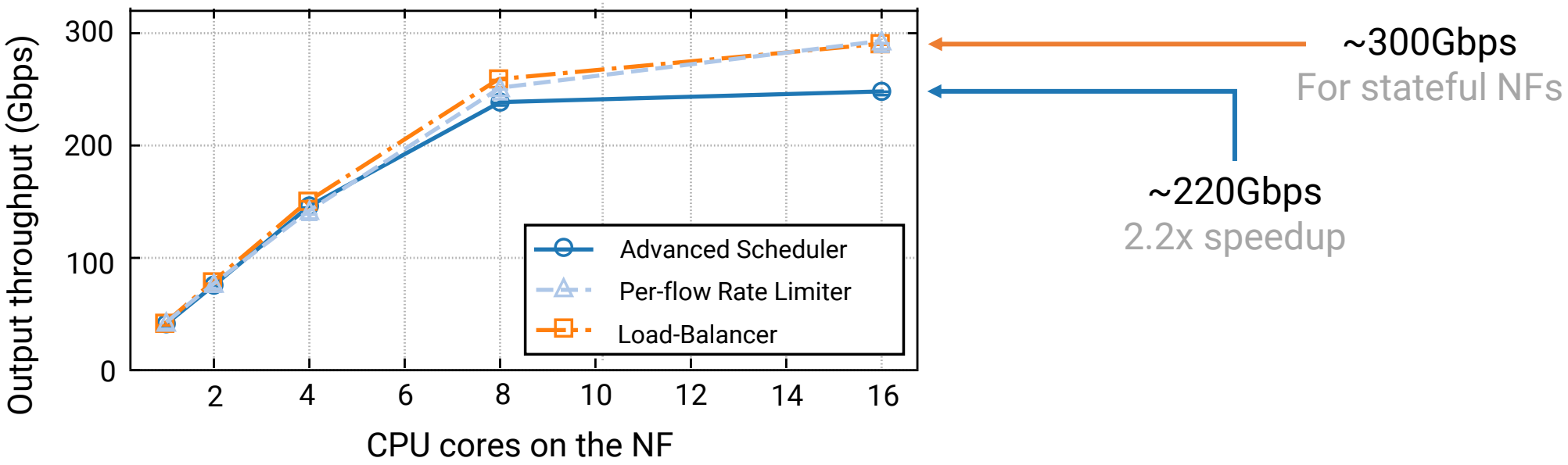
With 400B packets — NF can handle them! — We process 1Tbps!



We can process 1Tbps on a single dedicated CPU!

Advanced Network Functions

Can we build advanced NFs on top of Ribosome?



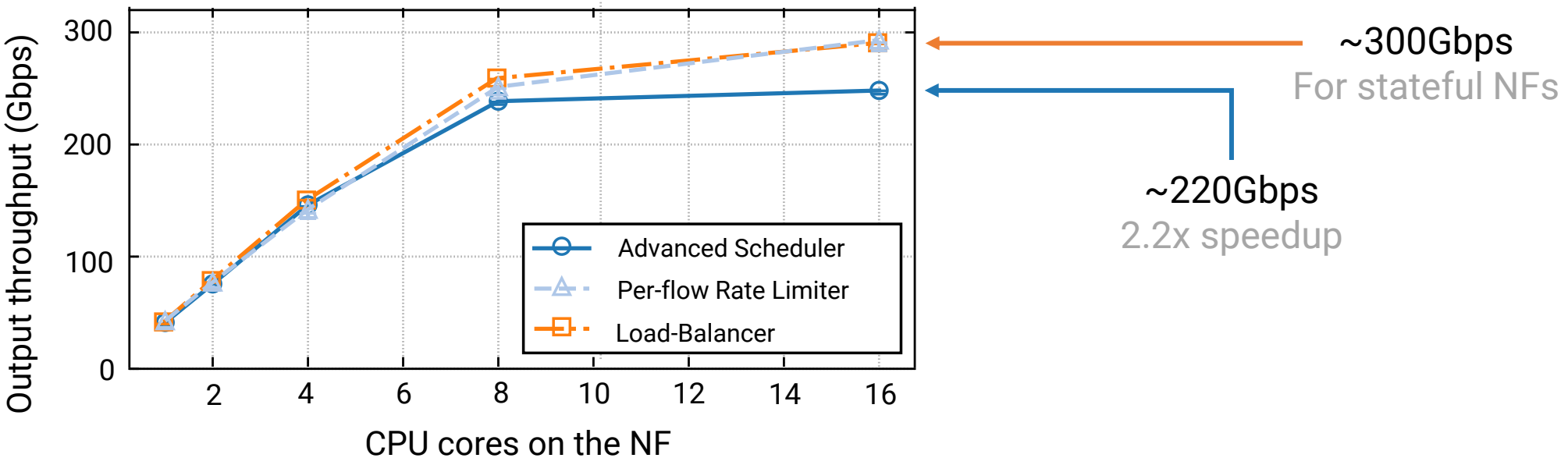
Advanced Scheduler → Reframer [NSDI'22]

Advanced Network Functions

Can we build advanced NFs on top of Ribosome?

Ribosome supports advanced NFs!

Ribosome moves the NF bottleneck on the CPU! → Work in Progress!



Conclusion

The **Ribosome** system:

- Reduce the amount of dedicated NF processors by carefully sending only relevant bits
- Improve the throughput and latency gains on the NF
- Support complex NFs
- Process 1Tbps on a single dedicated device

Ribosome-Packet-Processor