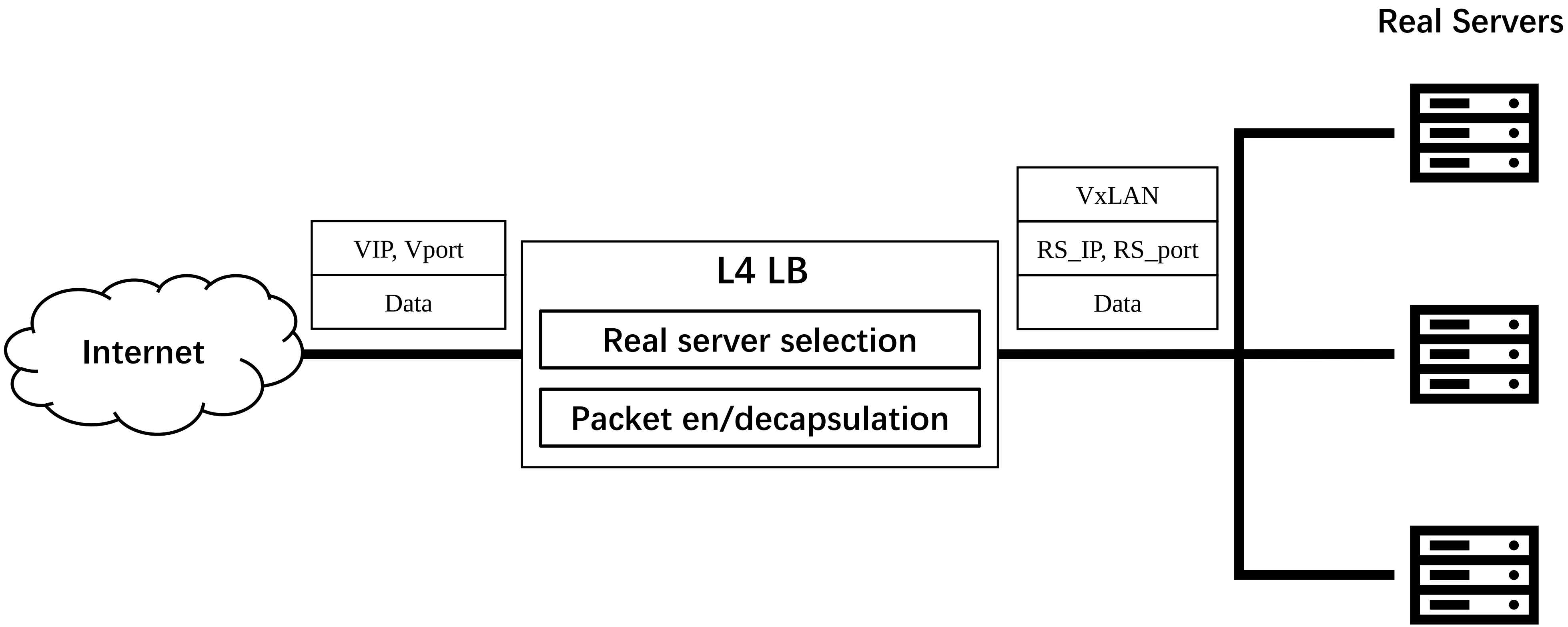
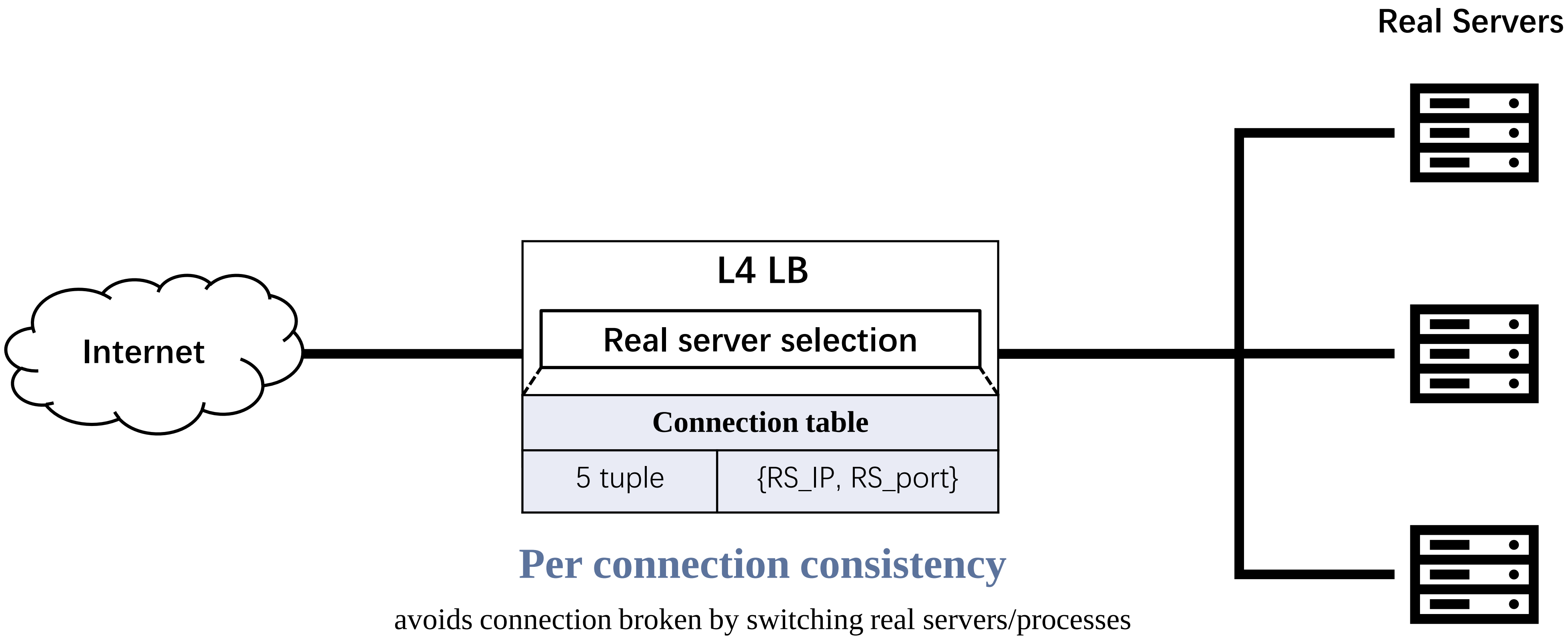
Tiara: A Scalable and Efficient Hardware Acceleration Architecture for Stateful Layer-4 Load Balancing

Chaoliang Zeng, Layong Luo, Teng Zhang, Zilong Wang, Luyang Li,
Wenchen Han, Nan Chen, Lebing Wan, Lichao Liu, Zhipeng Ding, Xiongfei Geng,
Tao Feng, Feng Ning, Kai Chen, Chuanxiong Guo

L4 LB at datacenter boundary



Being stateful



Per connection consistency

avoids connection broken by switching real servers/processes

Stateful L4 LB requirements

Driven by exponentially increased content delivery and cloud computing demands, a typical LB in large service providers usually supports

- Terabits per second of Internet traffic
- Tens of millions of concurrent flows
- Millions of new connections per second (CPS)

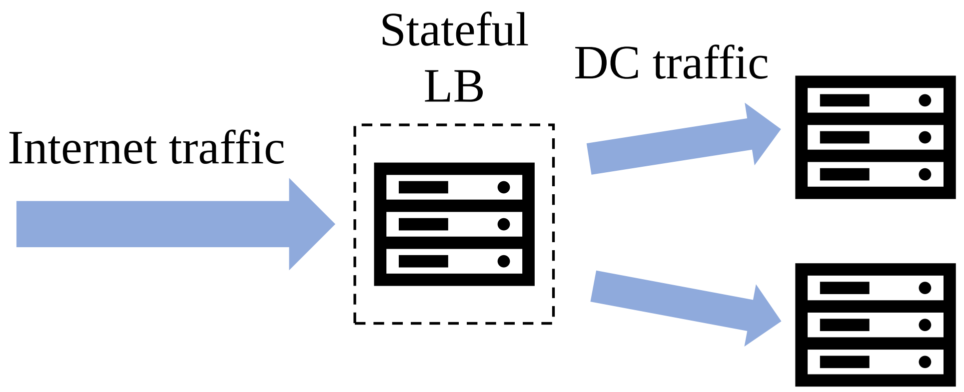
Stateful L4 LB requirements

Driven by exponentially increased content delivery and cloud computing demands, a typical LB in large service providers usually supports

- Terabits per second of Internet traffic
- Tens of millions of concurrent flows
- Millions of new connections per second (CPS)

Existing LBs fail to meet these requirements in a scalable and efficient way

Existing solution: software-based LB



Software-based LB can scale out to support high throughput

Ananta [SIGCOMM'13]
Maglev [NSDI'16]

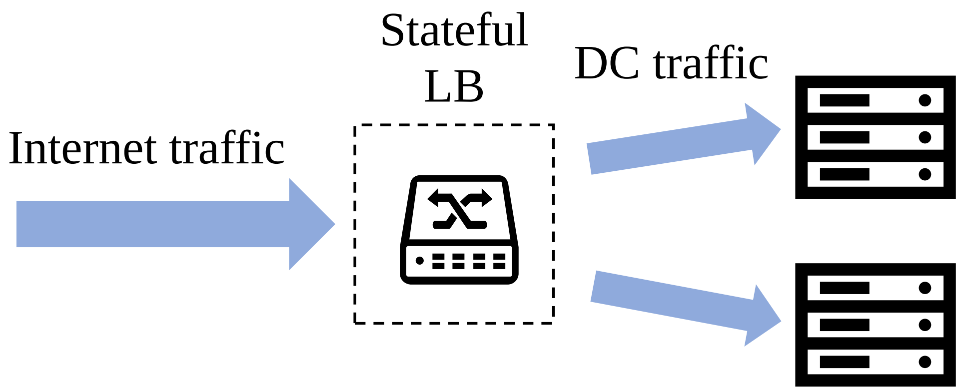
Low (cost, energy and space) efficiency
- 10 Gbps/server or 2 Mpps/core
- 100 servers to support 1 Tbps

- Expensive
- Sometimes undeployable in resource-constrained PoPs or edge DCs

High latency and jitter
- 10 us average latency
- up to ms tail latency/jitter

- Sometimes comparable to Internet latency when CPU utilization is high

Existing solution: switch-based LB



Leveraging programmable switches can improve efficiency and latency

Silkroad [SIGCOMM'17]

Cheetah [NSDI'20]

Scalability issue on data plane

- 50-100 MB on-chip memory

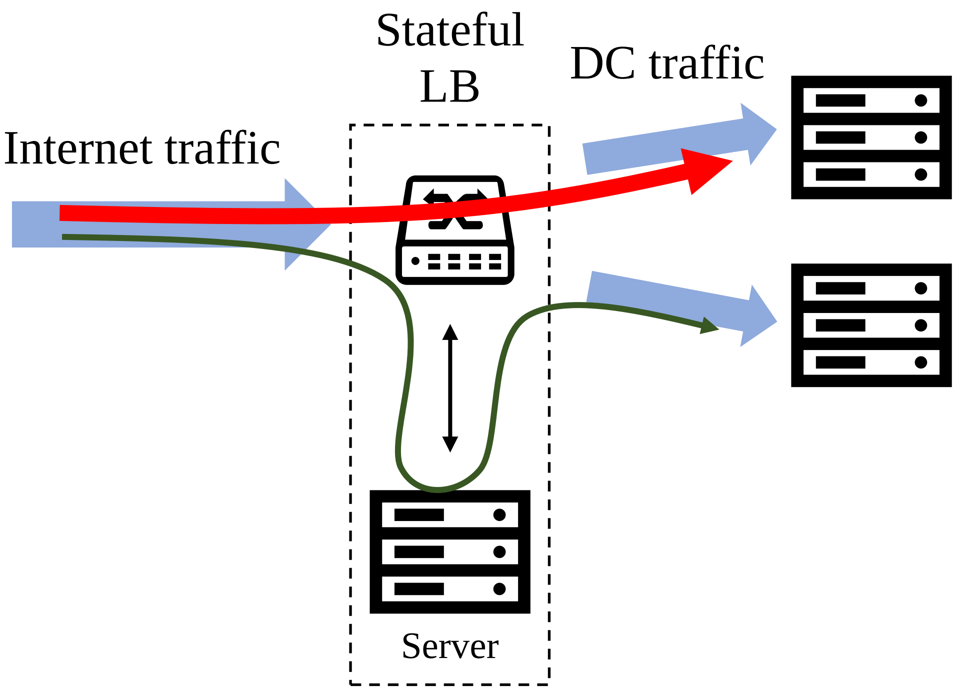
➢ Fail to support a large number of concurrent connections

Scalability issue on control plane

- 100K entry insertions per second
  - low-end SoC
  - slow PCIe interconnect
  - Cuckoo hash

➢ Fail to support high CPS

Strawman solution: switch-server LB



Leveraging traffic locality can address scalability issues of switches

Serving only a few elephant flows in the switch
Serving the rest traffic in the server

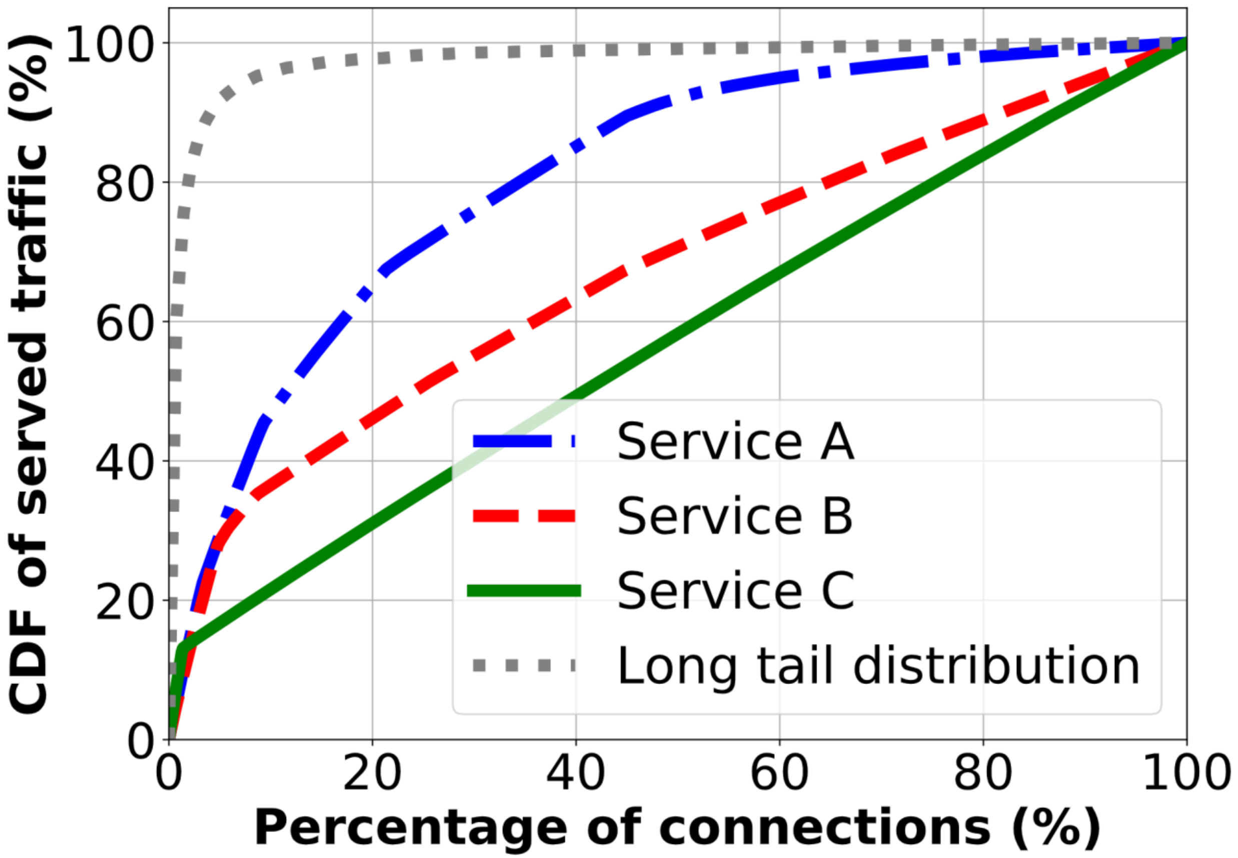
Traffic locality assumption →

- ➢ Traffic do not necessarily follow a long-tail distribution.
- ➢ It is dynamic and unpredictable!

Traffic at datacenter boundary

The flow distribution of individual services varies

- Top 10% connections carry 46.3%, 35.5%, and 19.6% traffic in three traces respectively.



Traffic distribution may not be long-tail!

➢ Limited memory in switch cannot hold enough connections to serve the majority of traffic

Traffic at datacenter boundary

The flow distribution of individual services varies

- Top 10% connections carry 46.3%, 35.5%, and 19.6% traffic in three traces respectively.

The traffic volume of a service can dynamically change

- Tidal traffic in a single day.
- Uncertainty in long-term due to change of users' interests.

The number of VIPs at a datacenter boundary can change over time

- A cluster can increase 3.2x VIPs in 6 months.

No assumption on traffic distribution at datacenter boundary!

System goals

Scalable – 10M concurrent connections and 1M CPS

Efficient – high cost, energy, and space efficiency

Generic – no assumption on traffic patterns

Our answer: Tiara

Tiara idea

LB Functionalities

Real server selection

Stateful
memory-intensive

Packet en/decapsulation

Stateless
throughput-intensive

Tiara idea

Perfect mapping

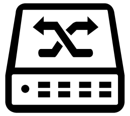
LB Functionalities → Hardware Components

Real server selection → 

FPGA-based SmartNIC

Stateful
memory-intensive

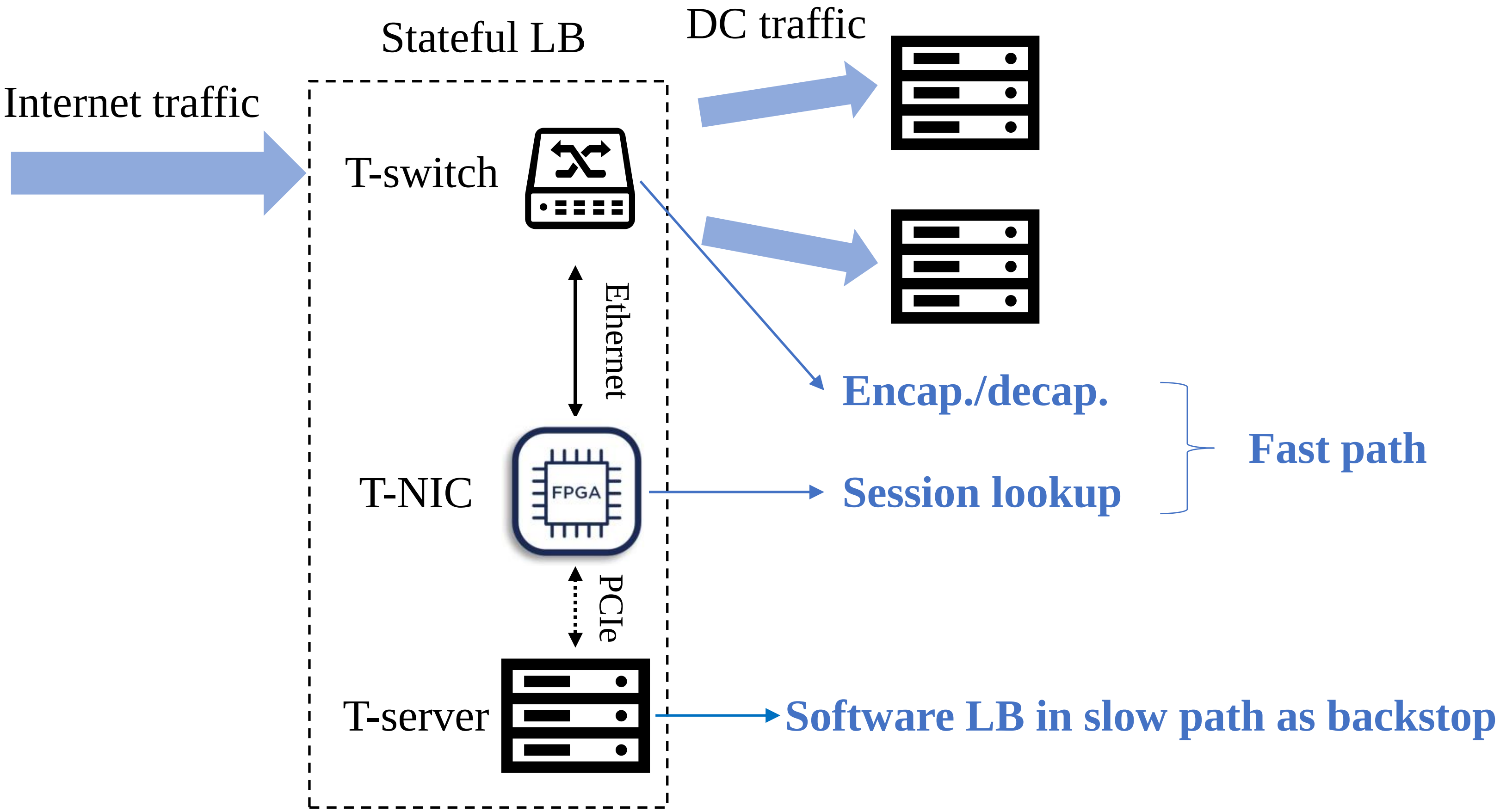
Large and fast HBM

Packet en/decapsulation → 
Programmable Switch

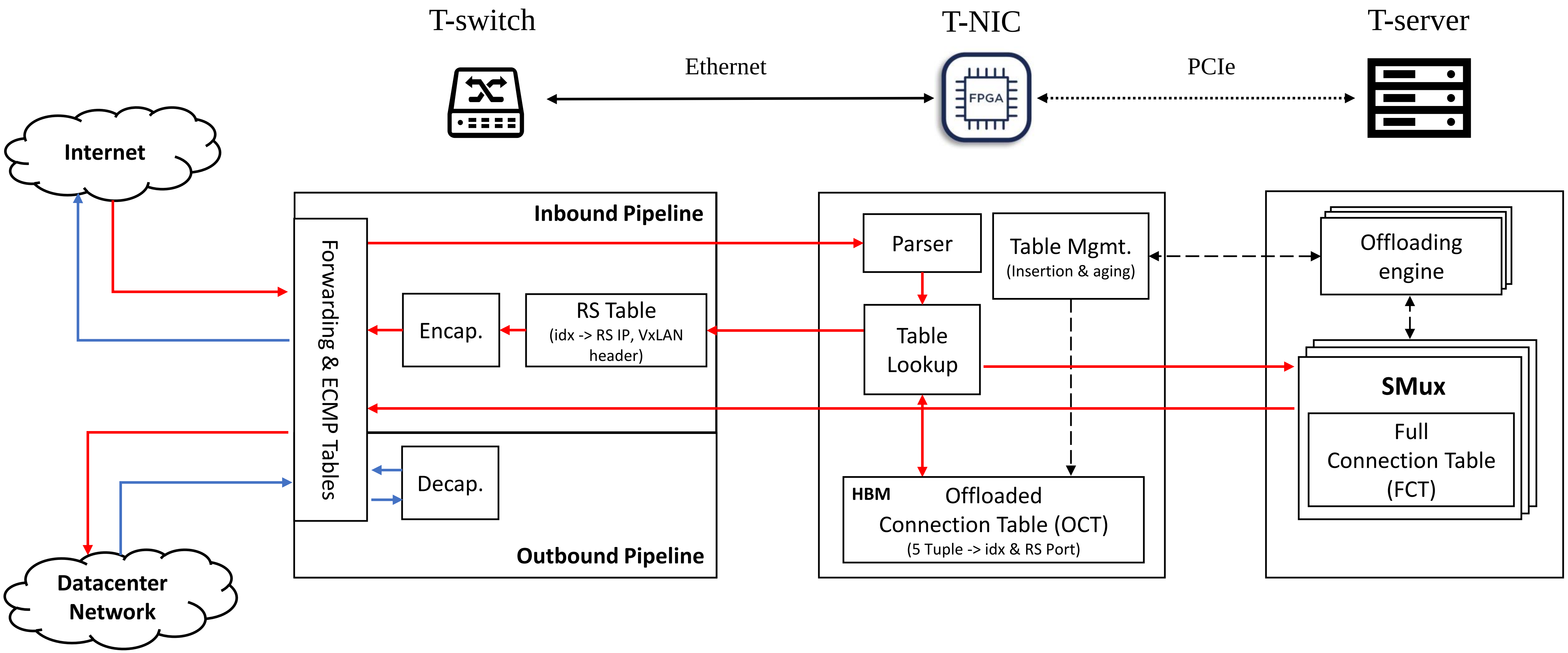
Stateless
throughput-intensive

Limited memory
High throughput

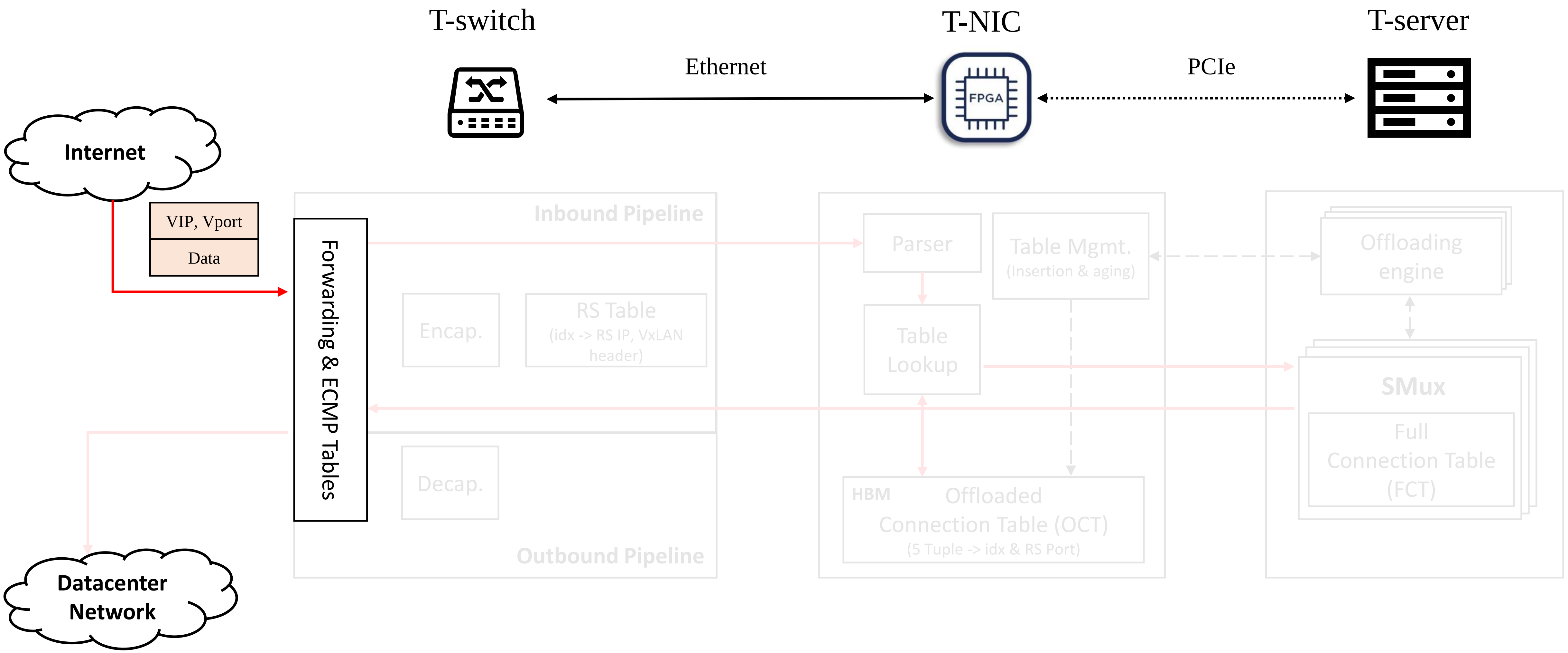
Tiara three-tier architecture



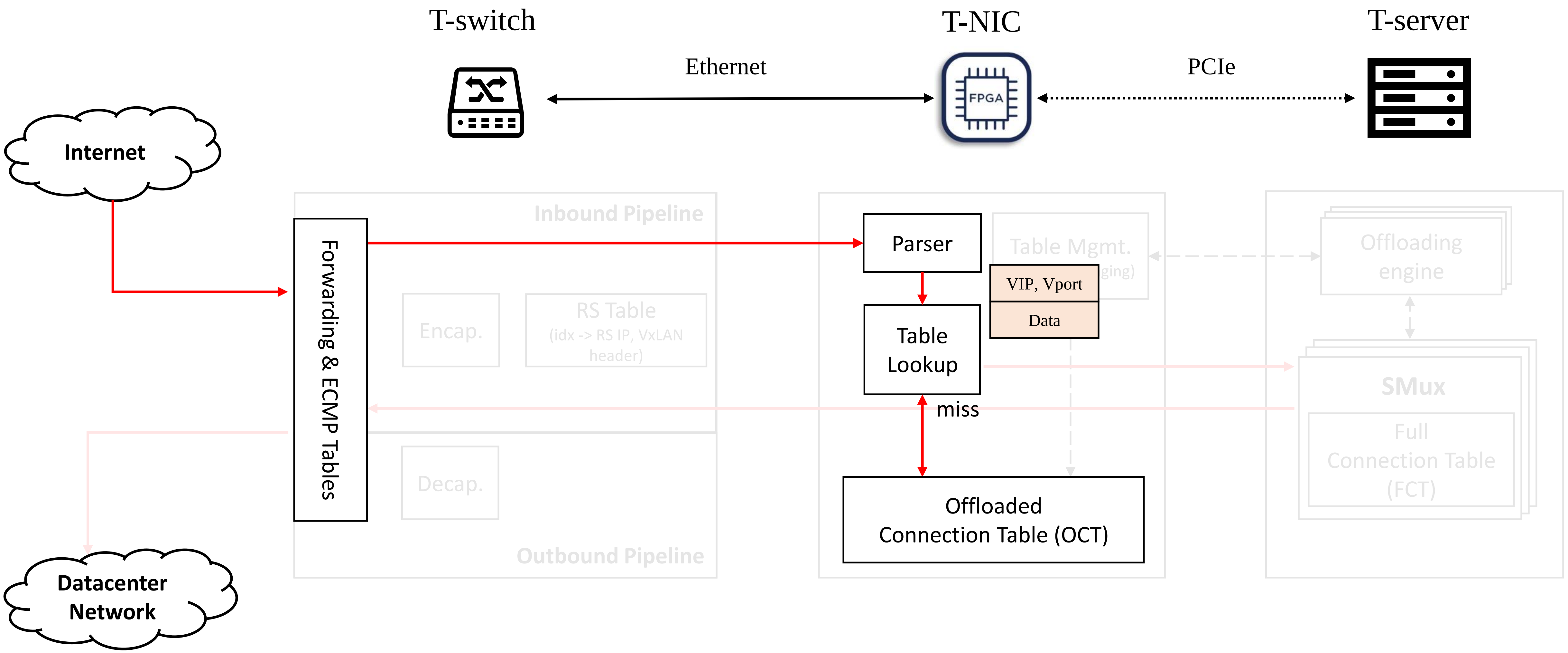
Tiara architecture in details



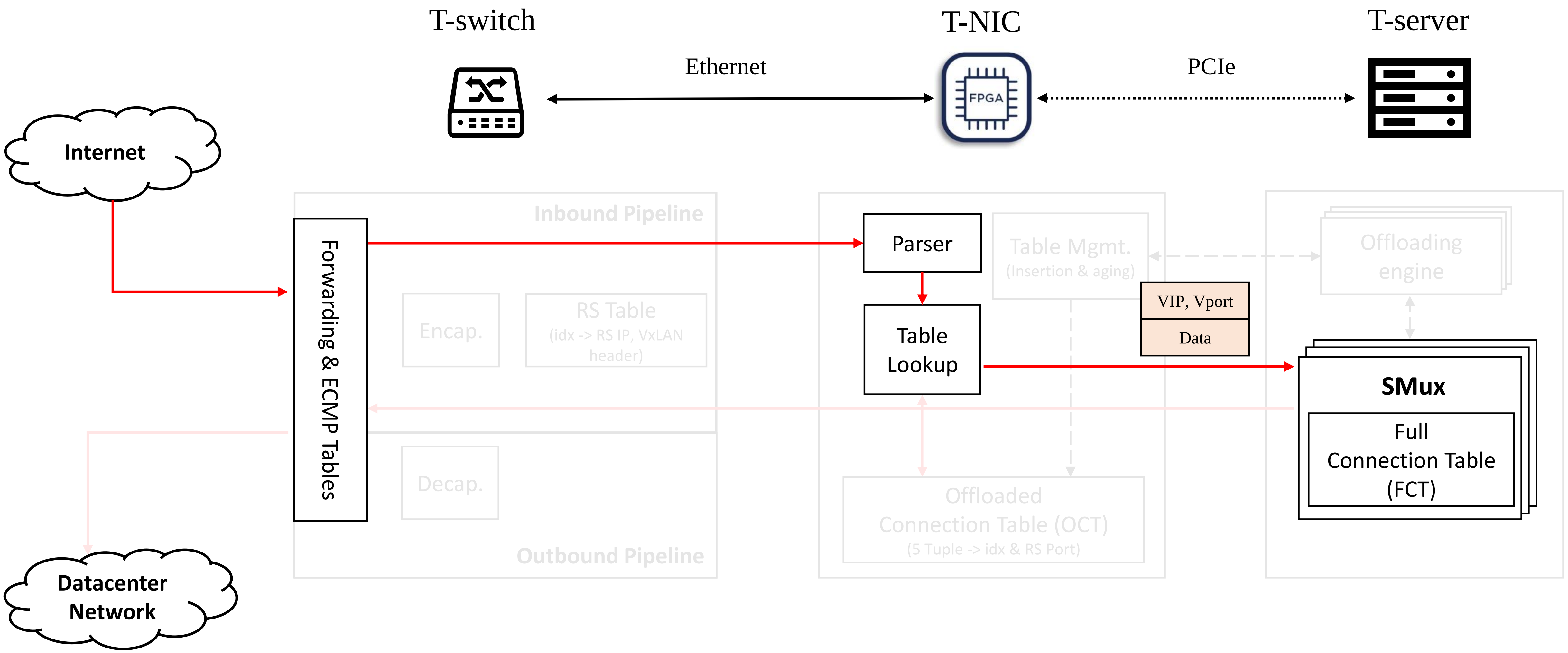
Inbound traffic



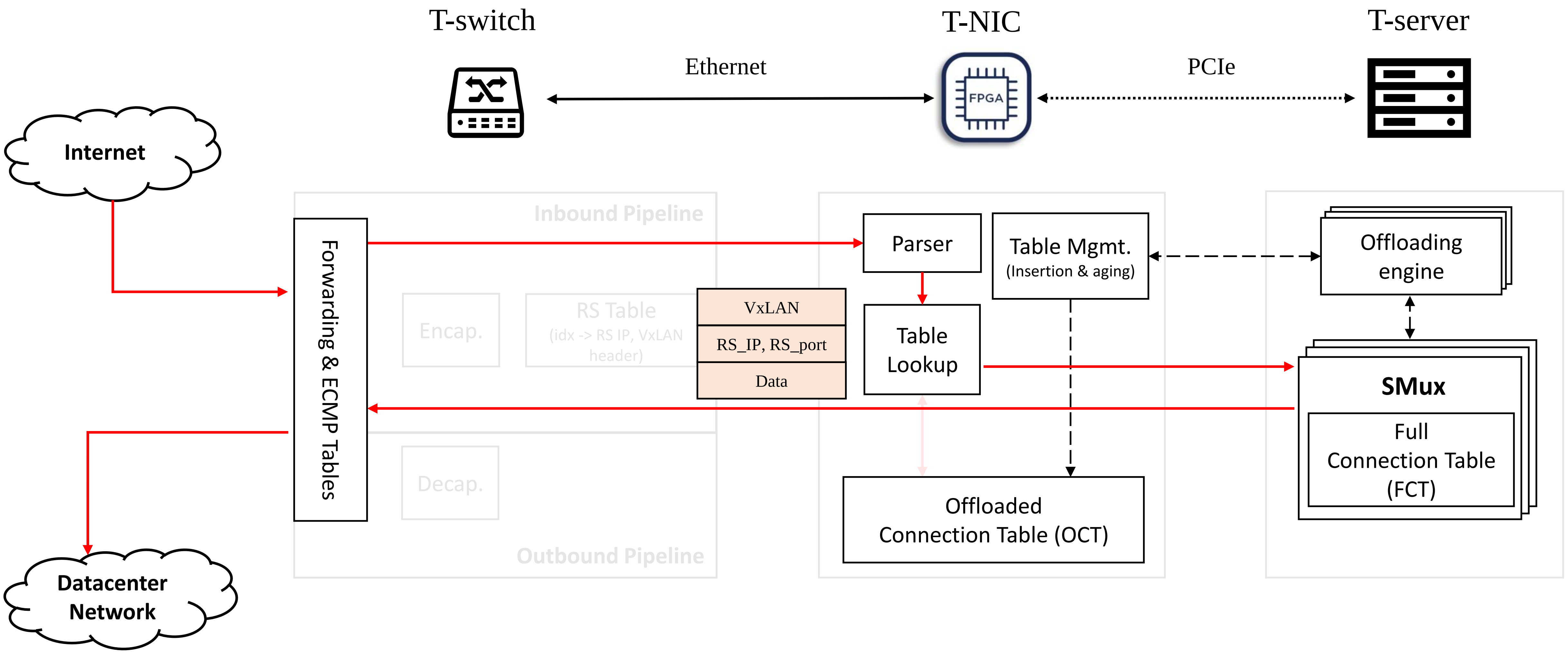
Inbound traffic: the first packet



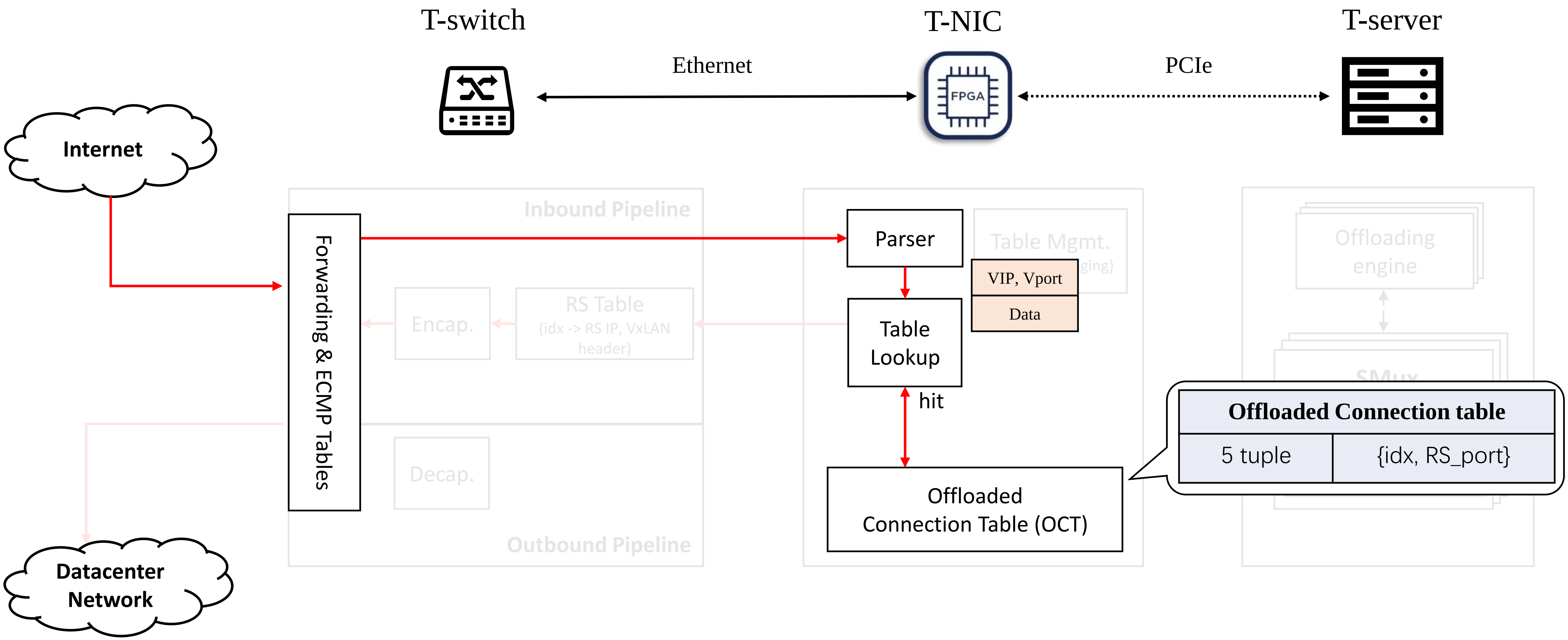
Inbound traffic: the first packet



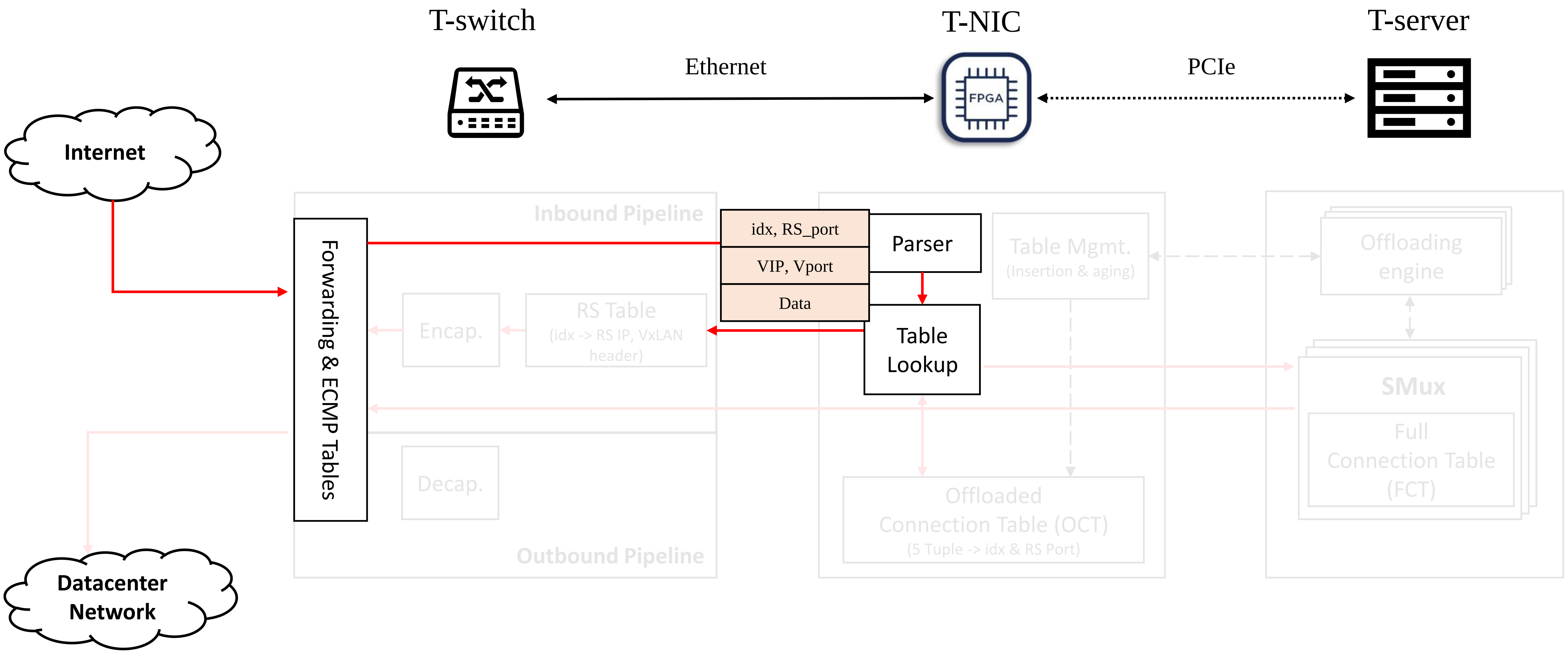
Inbound traffic: the first packet



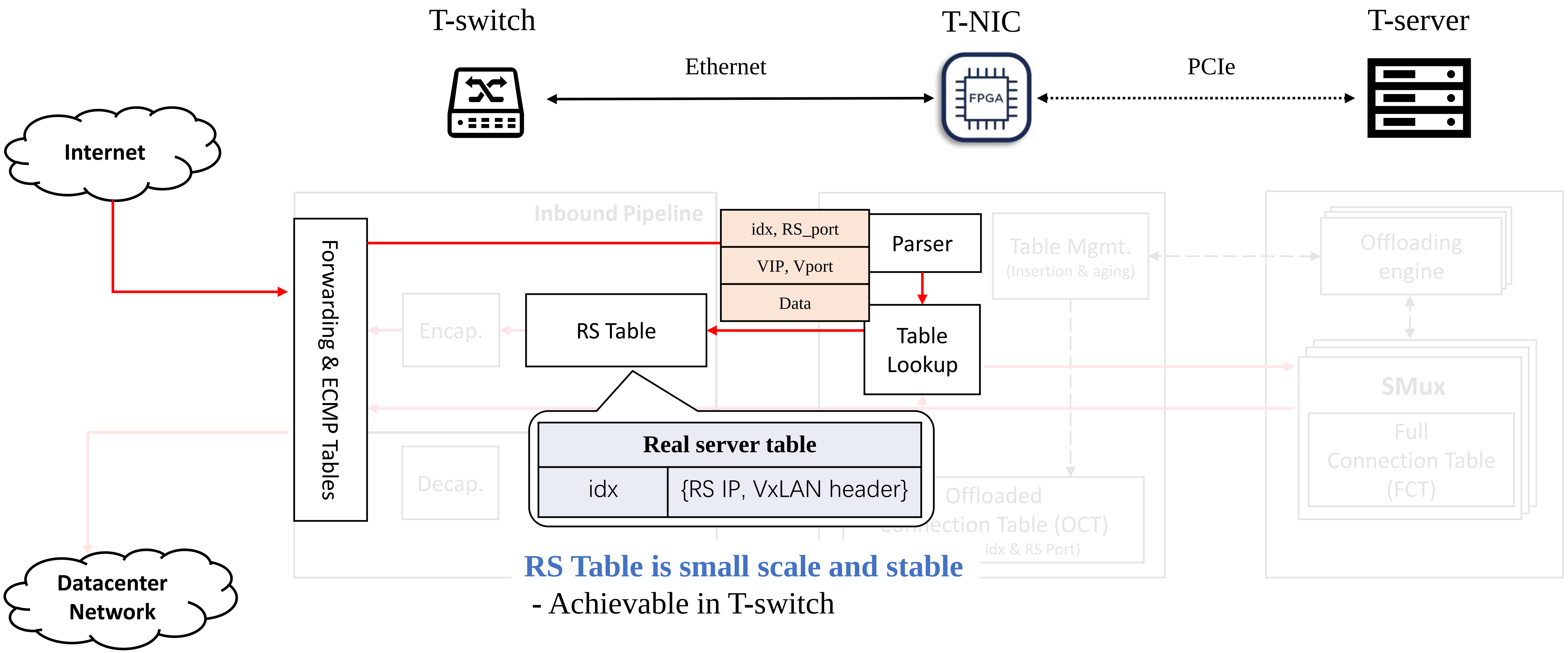
Inbound traffic: the rest packets



Inbound traffic: the rest packets



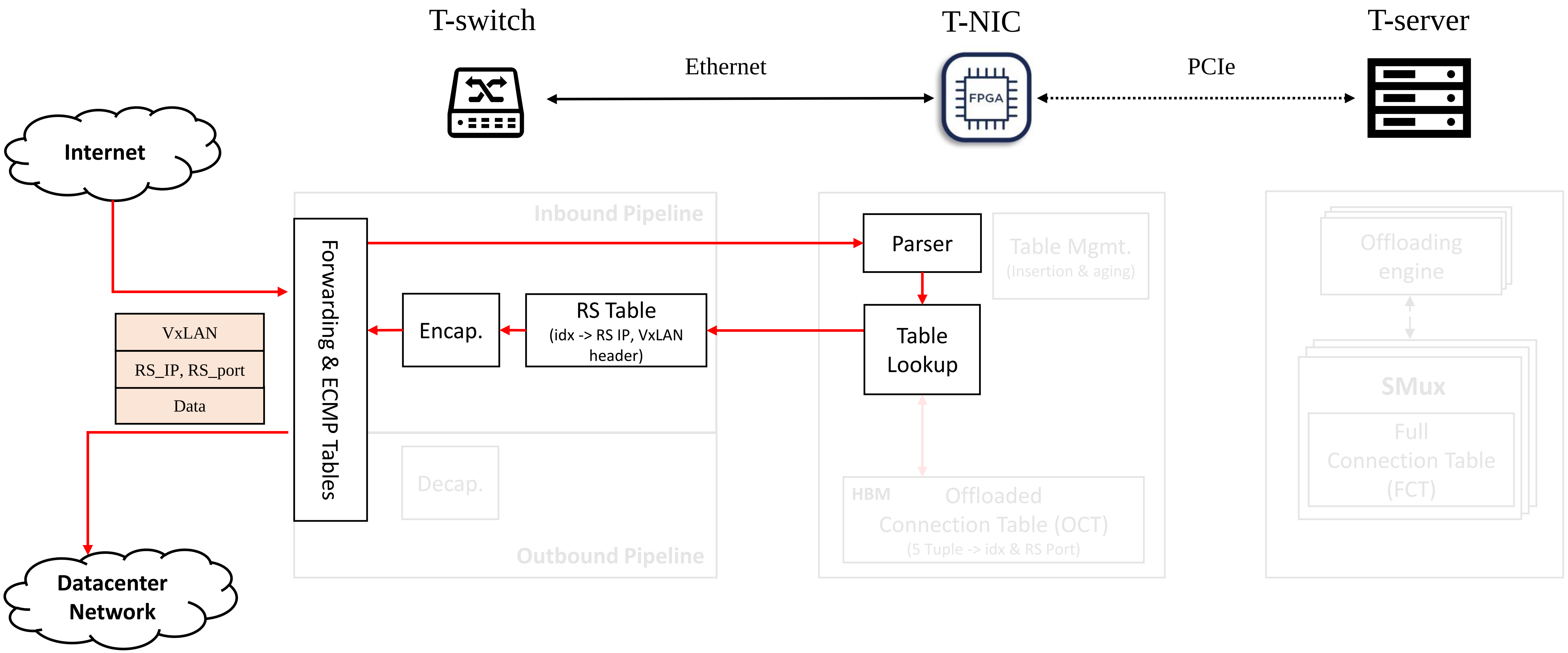
Inbound traffic: the rest packets



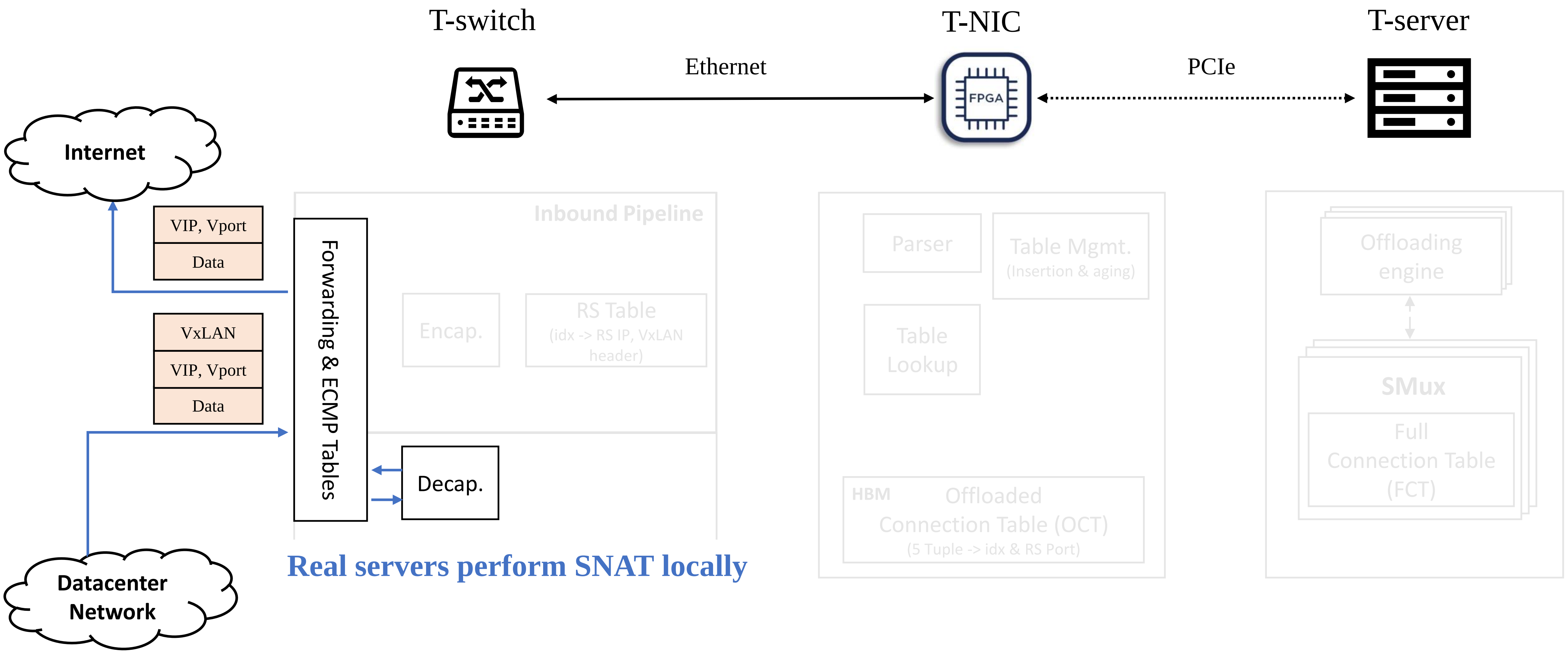
RS Table is small scale and stable

- Achievable in T-switch

Inbound traffic: the rest packets



Outbound traffic



Real servers perform SNAT locally

Optimizations

- Efficient hash table structure
  - To enable both fast lookup in T-NIC and fast entry insertion in T-server
  - Optimization for throughput, concurrent flow number, and CPS
- Lock-free offloading approach
  - To enable millions of flow offloading operations per second
  - Optimization for CPS
- Lightweight aging mechanism
  - To recycle outdated entries in FPGA HBM
  - Optimization for efficiency

Prototype implementation

T-switch: Barefoot Tofino switch

- RS Table: 64K entries

T-NIC: Xilinx FPGA-based SmartNIC with two 100GE ports & one HBM stack

- Connection table: 32M entries

T-server: Server with two Intel Xeon Platinum 8260 CPUs running a production SMux

- SMux CPS: 1.8M

Resource Untilization

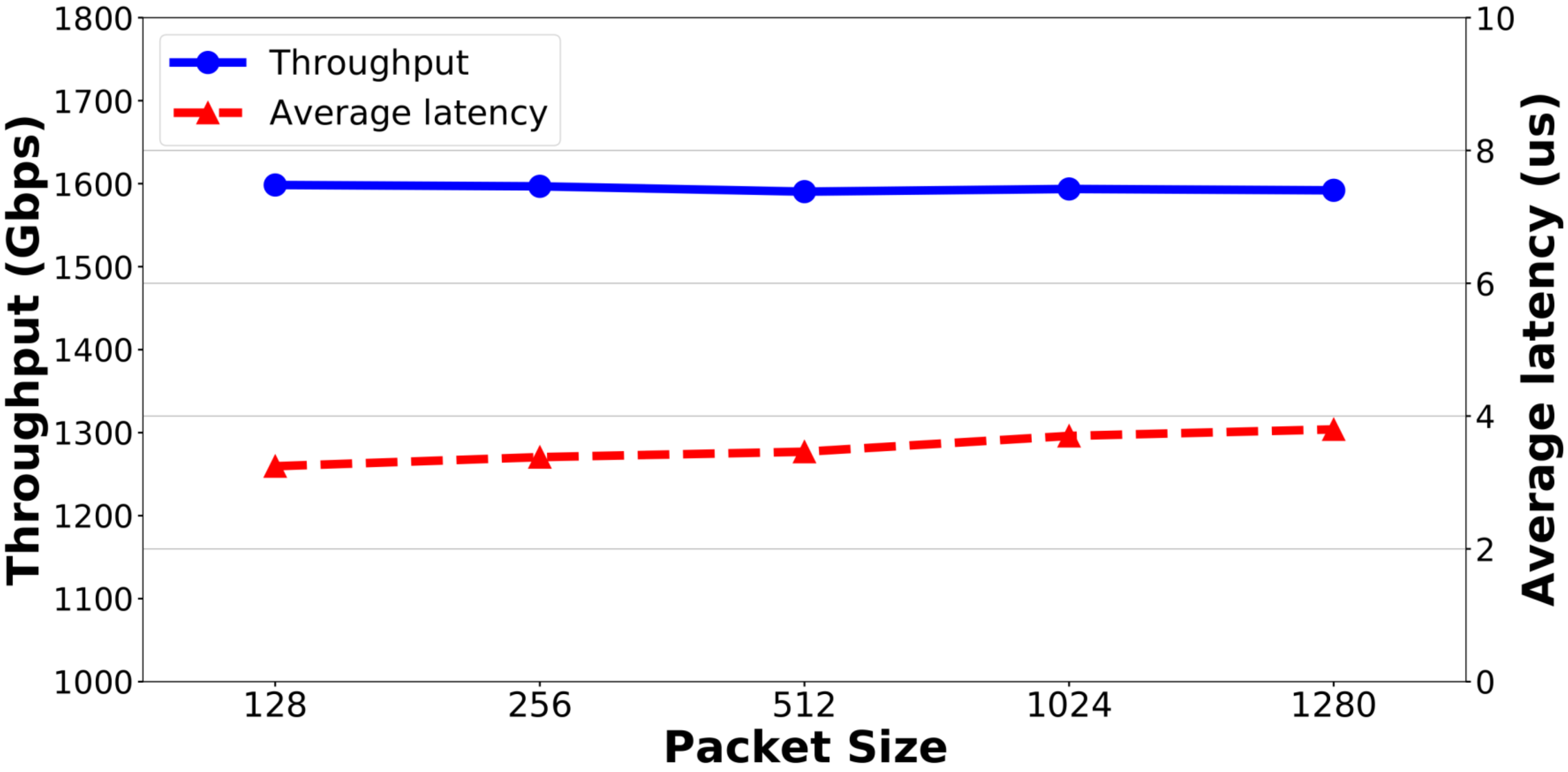
T-switch

SRAM	53.85%
TCAM	13.19%

T-NIC

LUT	33.22%
FF	28.46%
BRAM	50.93%
URAM	36.72%

System performance

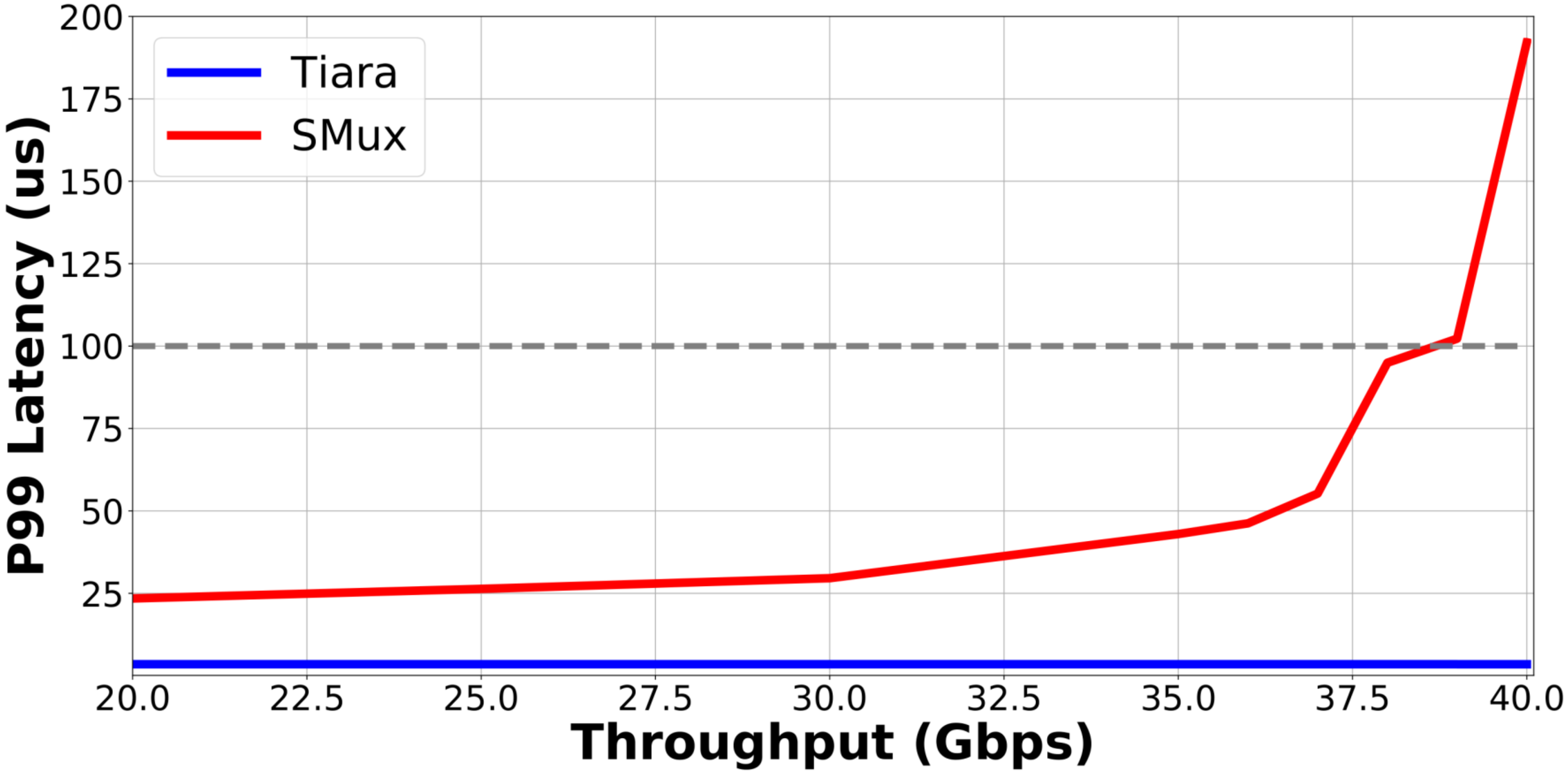


10M concurrent connections
(all offloaded in FPGA HBM)

1 T-server with 8 T-NICs
= 1.6 Tbps & less than 4 us

1.8M CPS for HTTP transactions
(bounded by SMux)

Latency-bounded throughput



P99 latency ≤ 100 us:
Tiara = 200 Gbps (1 T-NIC)
& 1.6 Tbps (8 T-NIC)
SMux = 38 Gbps

Switch-based solutions share a similar result as Tiara

Tiara vs. existing approaches

	Throughput	P99 lat.	CPS	CT size*	Cost efficiency	Energy efficiency	Space efficiency
SMux	38 Gbps	100 us	1.8M	~100 GB	4.75 Gbps/(cost unit)	76 Mbps/Watt	19 Gbps/U
Silkroad**	1.6 Tbps	< 2 us	200K	100 MB	457.14 Gbps/(cost unit)	2909.1 Mbps/Watt	1600 Gbps/U
Tiara	1.6 Tbps	< 4 us	1.8M	4 GB	82.05 Gbps/(cost unit)	969.7 Mbps/Watt	320 Gbps/U

17.4x higher cost efficiency, 12.8x higher energy efficiency, and 16.8x higher space efficiency than server-based solution

9x higher CPS and 40x larger connection table size than switch-based solution

Conclusion

Tiara is a three-tier hardware architecture for stateful L4 LB

- T-switch for stateless packet encap./decap.
- T-NIC for stateful real server selection
- T-server as slow path and make offloading decision

Tiara meets all design goals with high performance

- Scalable
  - Large HBM and efficient hash table for 10M concurrent flows
  - Fast PCIe DMA and lock-free offloading for 1M CPS
- Efficient
  - Specialized hardware for fast path
- Generic
  - No assumption on traffic patterns and fully programmable architecture

Contact: czengaf@connect.ust.hk, luo@bytedance.com