Hybrid Modular Switch (HyMoS)

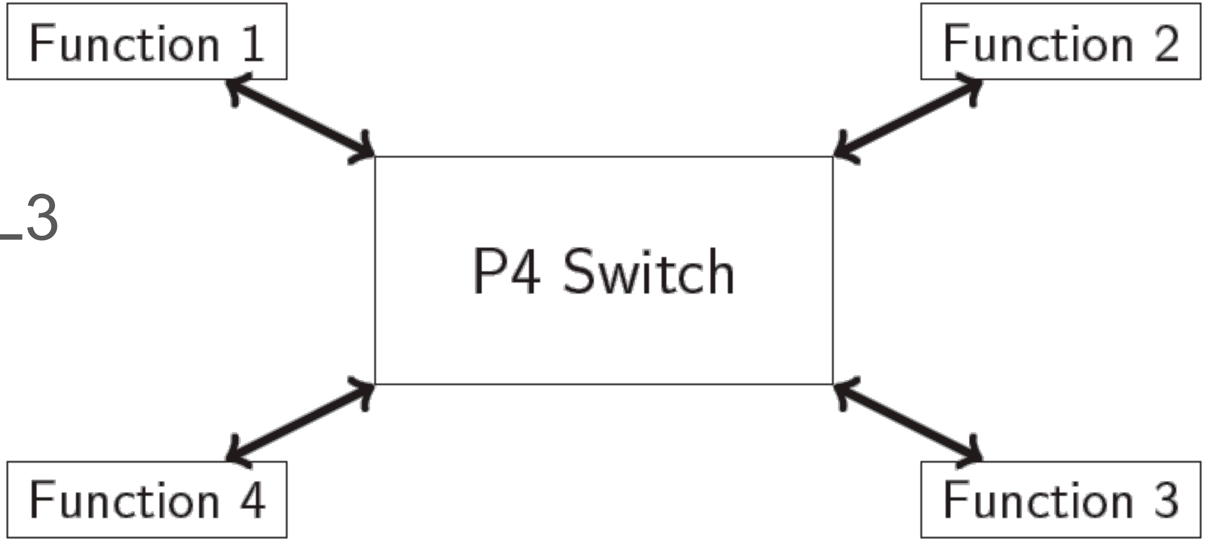
Ashkan Aghdai, Yang Xu, H. Jonathan Chao

NYU | TANDON SCHOOL OF ENGINEERING

Implementing NFV Functions

- Complex functions are offloaded to servers

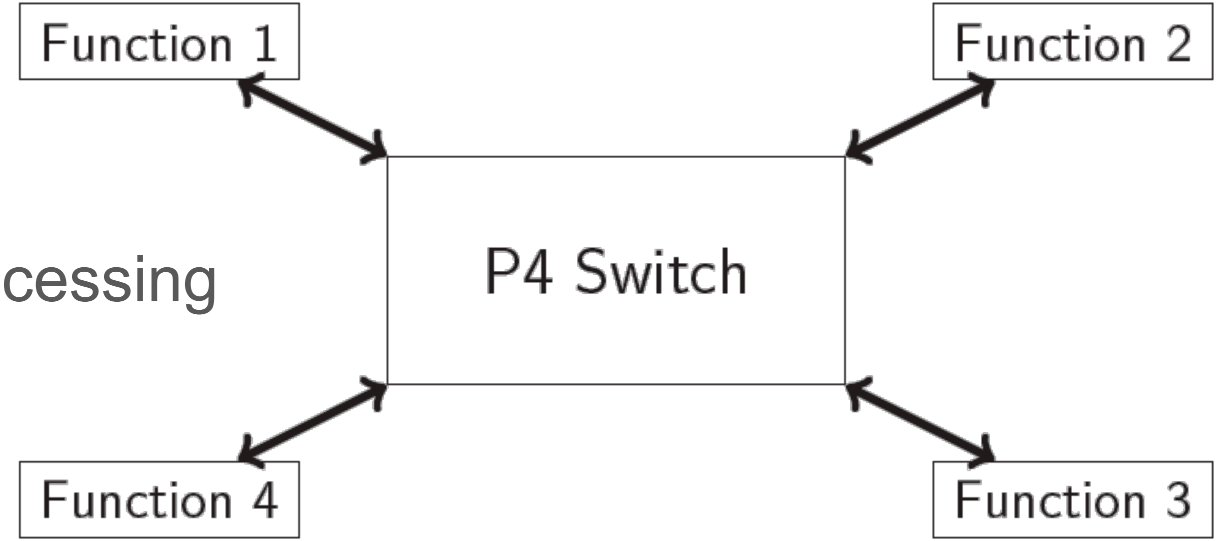
- P4 Switch: L2 and L3
- Servers: L4 and up



Offloading is not always the answer

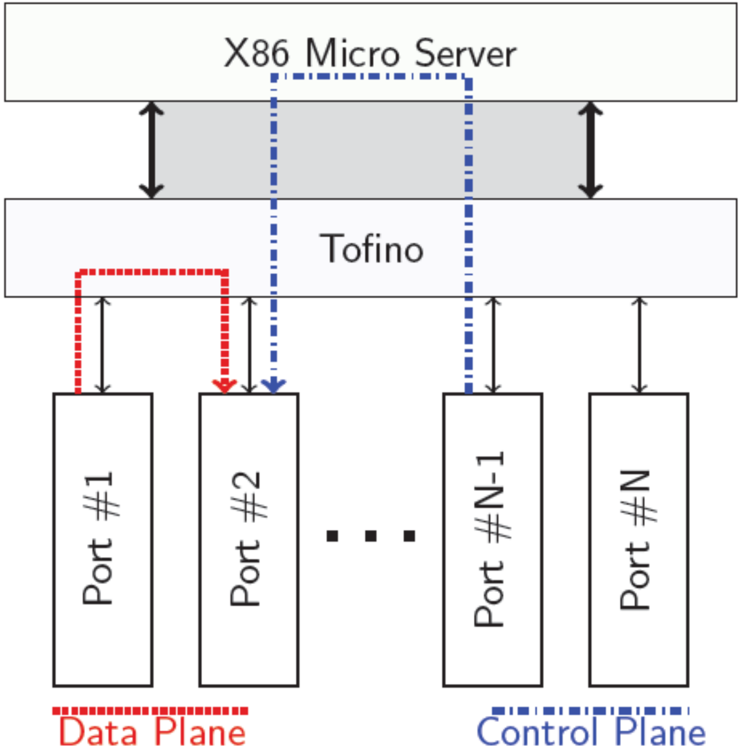
- QoS functions
  - SLAs
  - Priority Classes

- Stateful packet processing



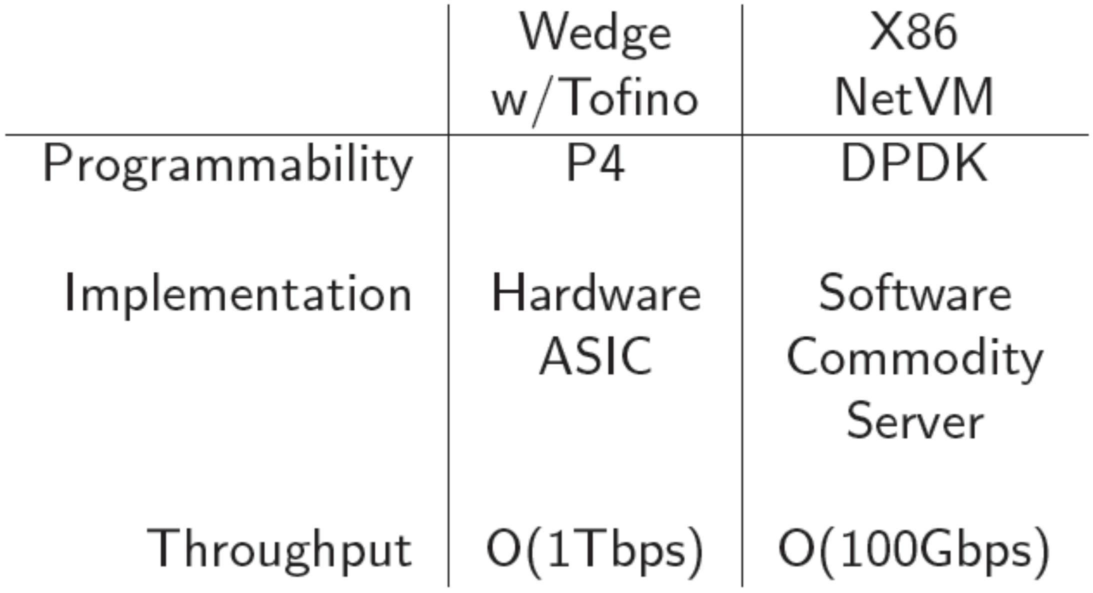
The Wedge Switch

- P4-Compatible/ Wedge
  - Micro-server for additional programmability
    - Control Plane
    - Data Plane
  - Scheduling is not programmable



©2017 Open-NFP

Implementing NFV Applications



	Wedge w/Tofino	X86 NetVM
Programmability	P4	DPDK
Implementation	Hardware ASIC	Software Commodity Server
Throughput	O(1Tbps)	O(100Gbps)

Can we get the best of both worlds?

- P4 as DSL
- DPDK for Network Functions
- Modularity

Compromise on the Throughput

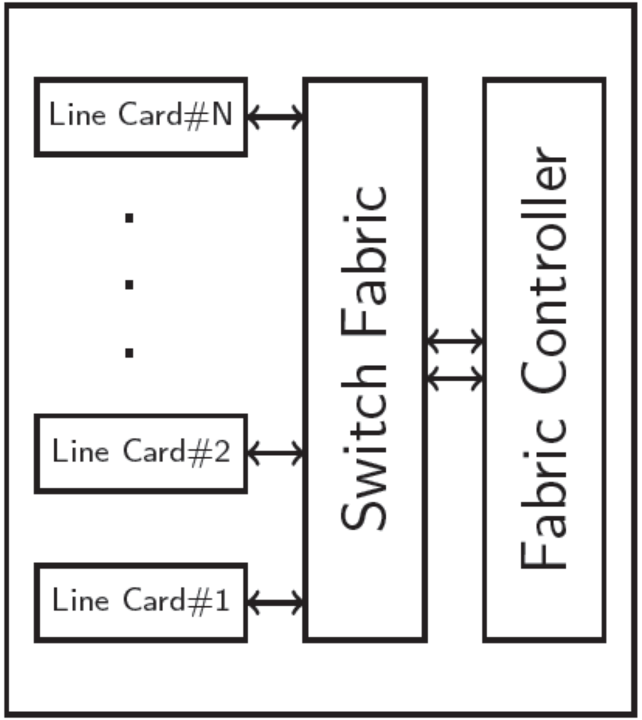
- O(100Gbps) for P4 and DPDK paths

Programmable Packet Forwarding

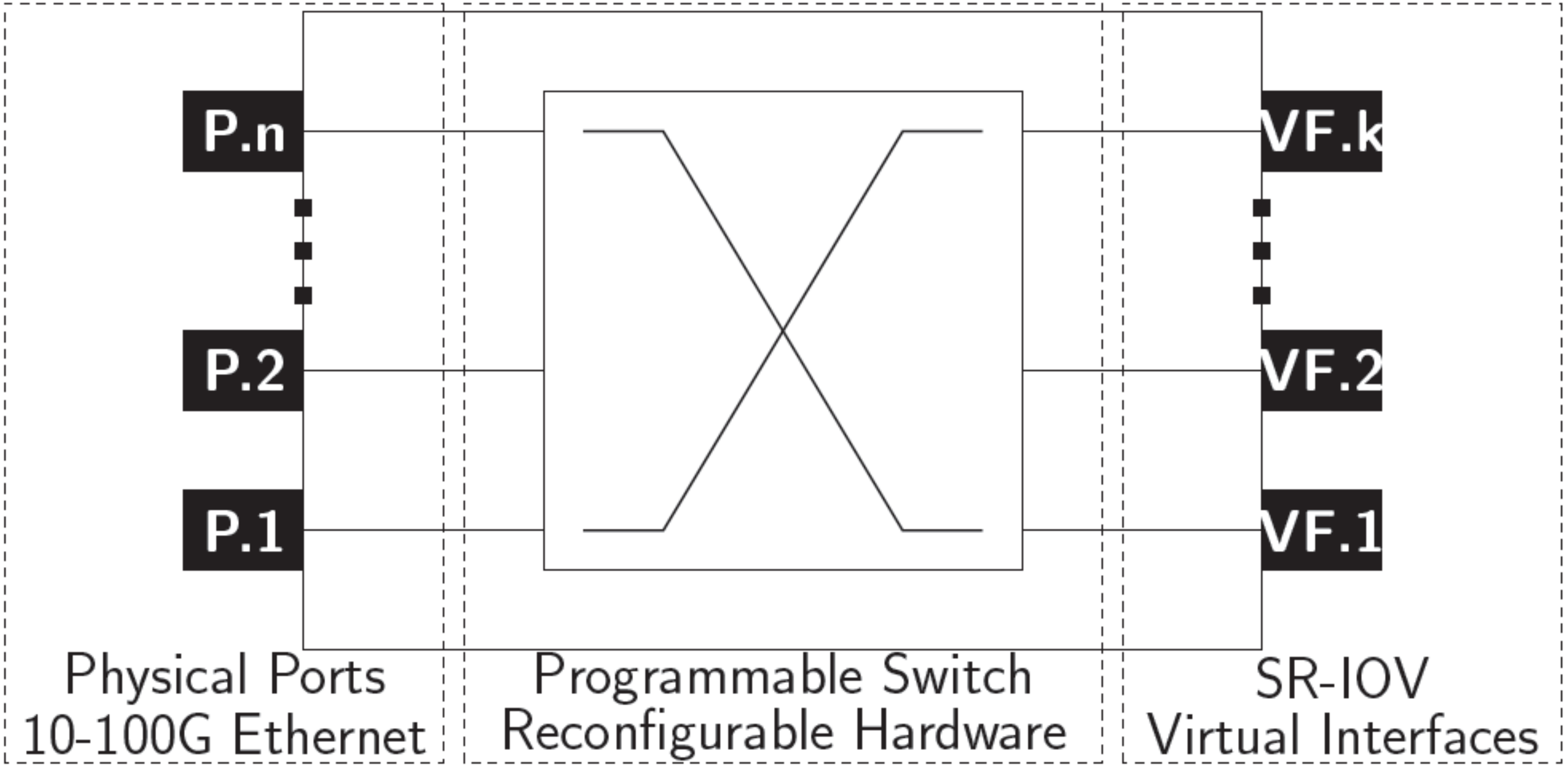
Let's make a programmable input-buffered switch

- Packet Processing
  - Table look-ups and header updates
  - Programmable Match+Action Tables
- Packet Switching
  - Copy from ingress to egress port/s
- Packet Scheduling
  - Orchestrate packet transfers

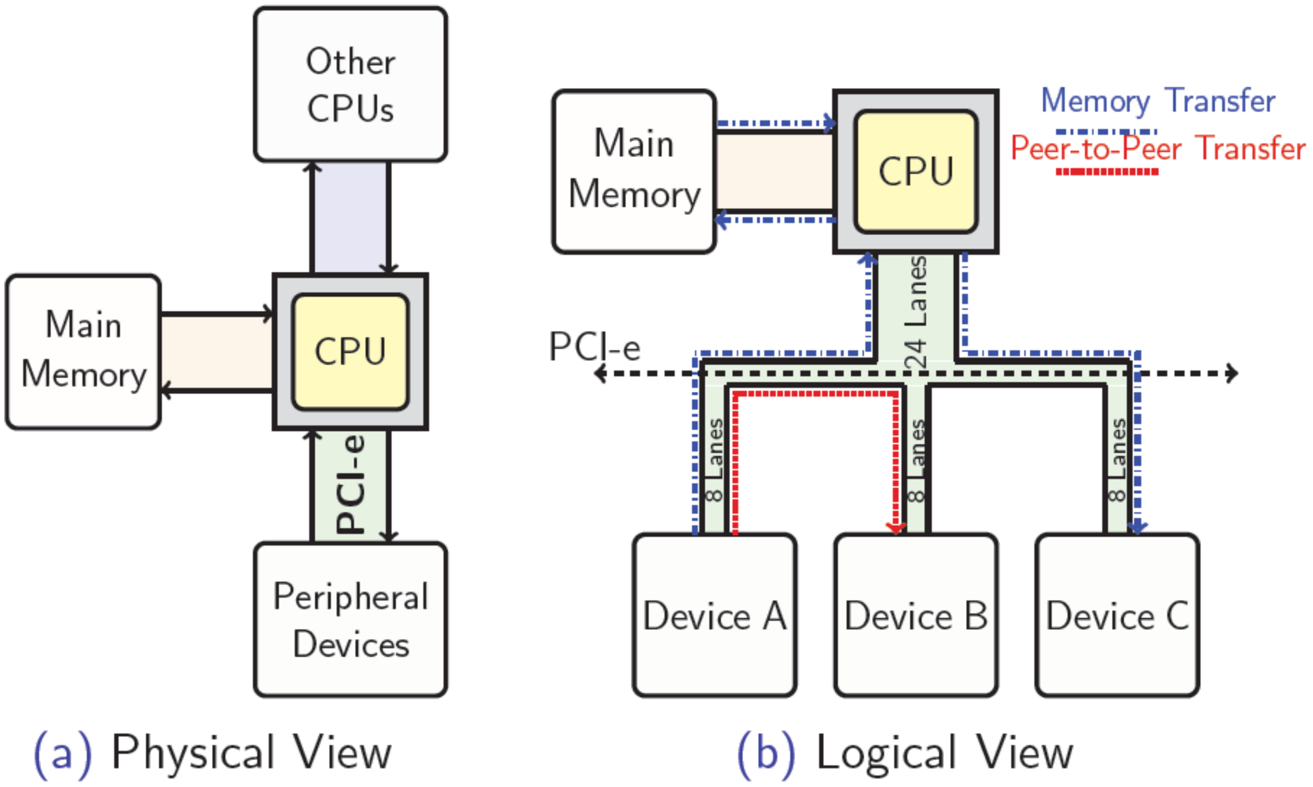
Programmable Switch



Smart NICs as Line Cards



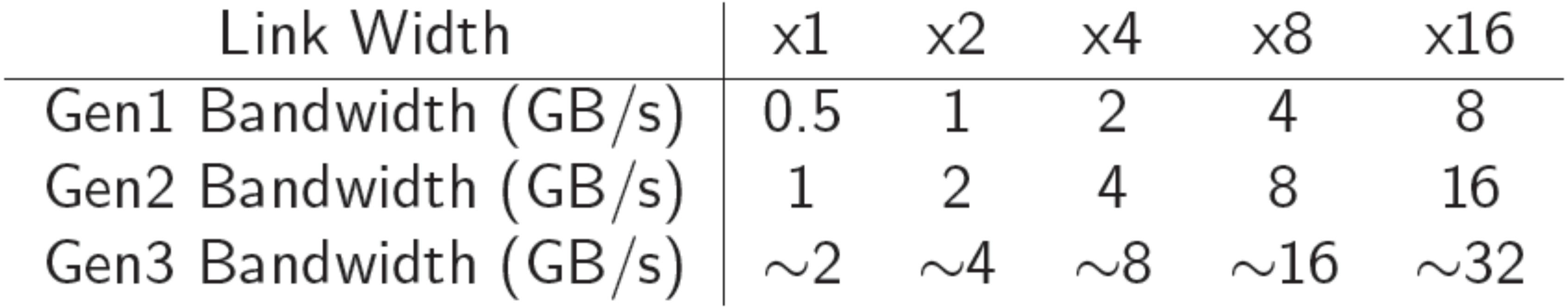
PCI-e as the Switch Fabric



(a) Physical View

(b) Logical View

PCI-e as the Switch Fabric

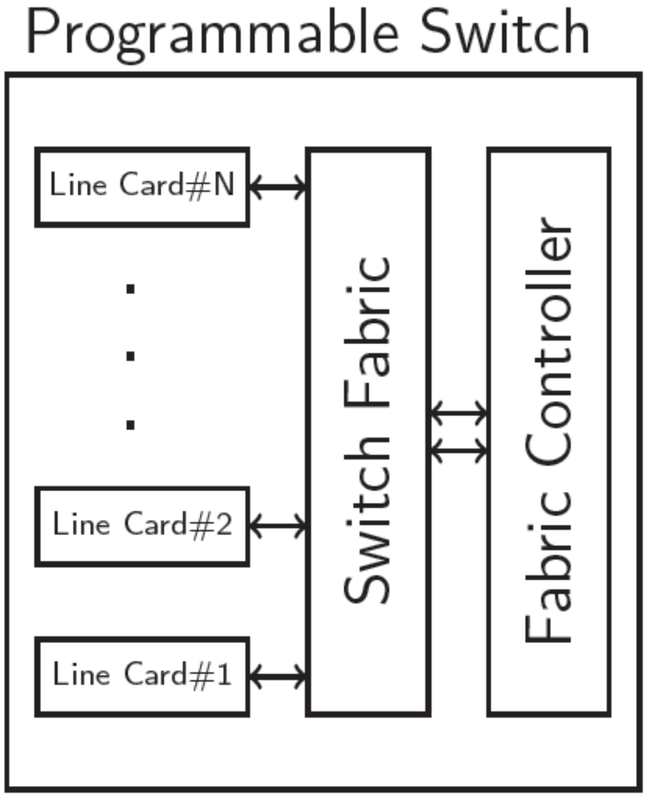


Link Width	x1	x2	x4	x8	x16
Gen1 Bandwidth (GB/s)	0.5	1	2	4	8
Gen2 Bandwidth (GB/s)	1	2	4	8	16
Gen3 Bandwidth (GB/s)	∼2	∼4	∼8	∼16	∼32

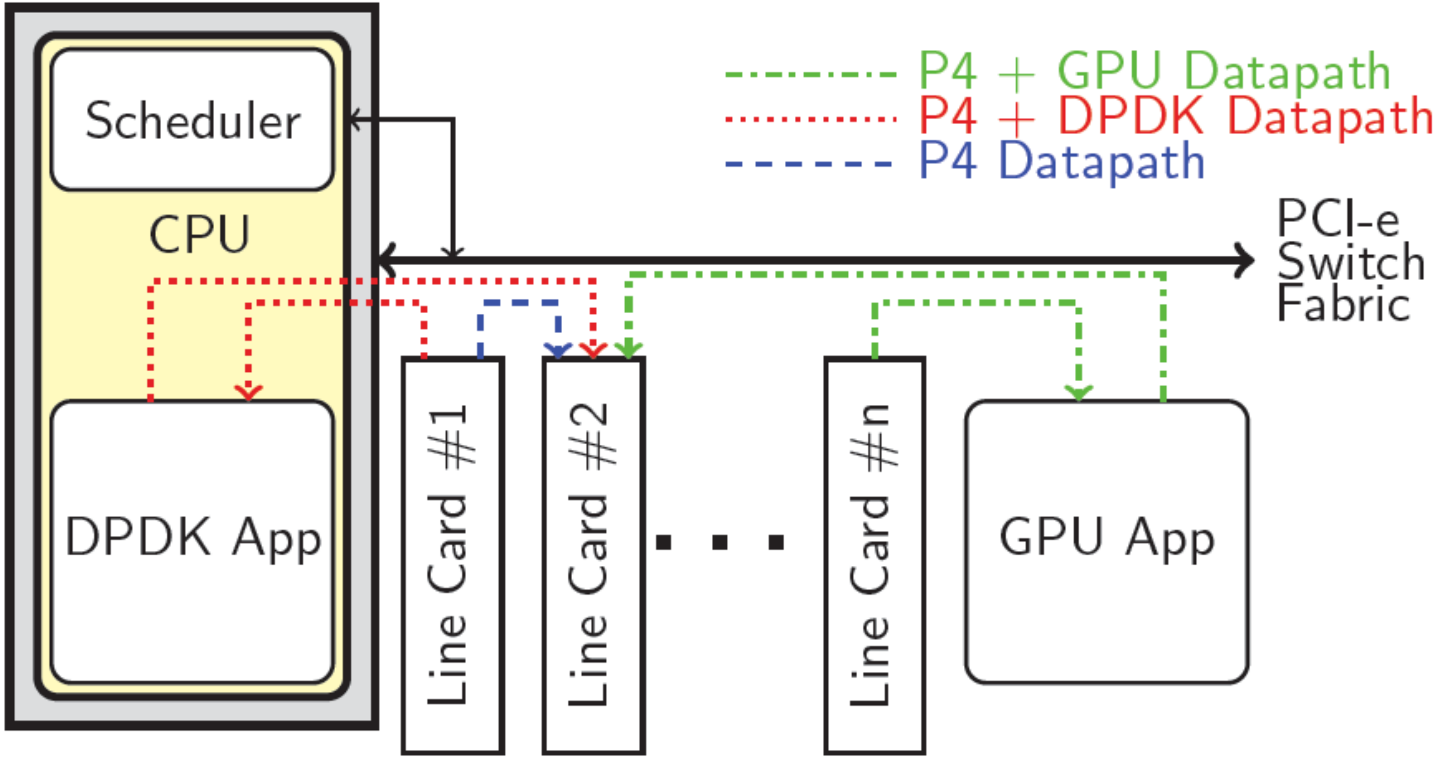
Proposed Architecture

Let's make a programmable input-buffered switch

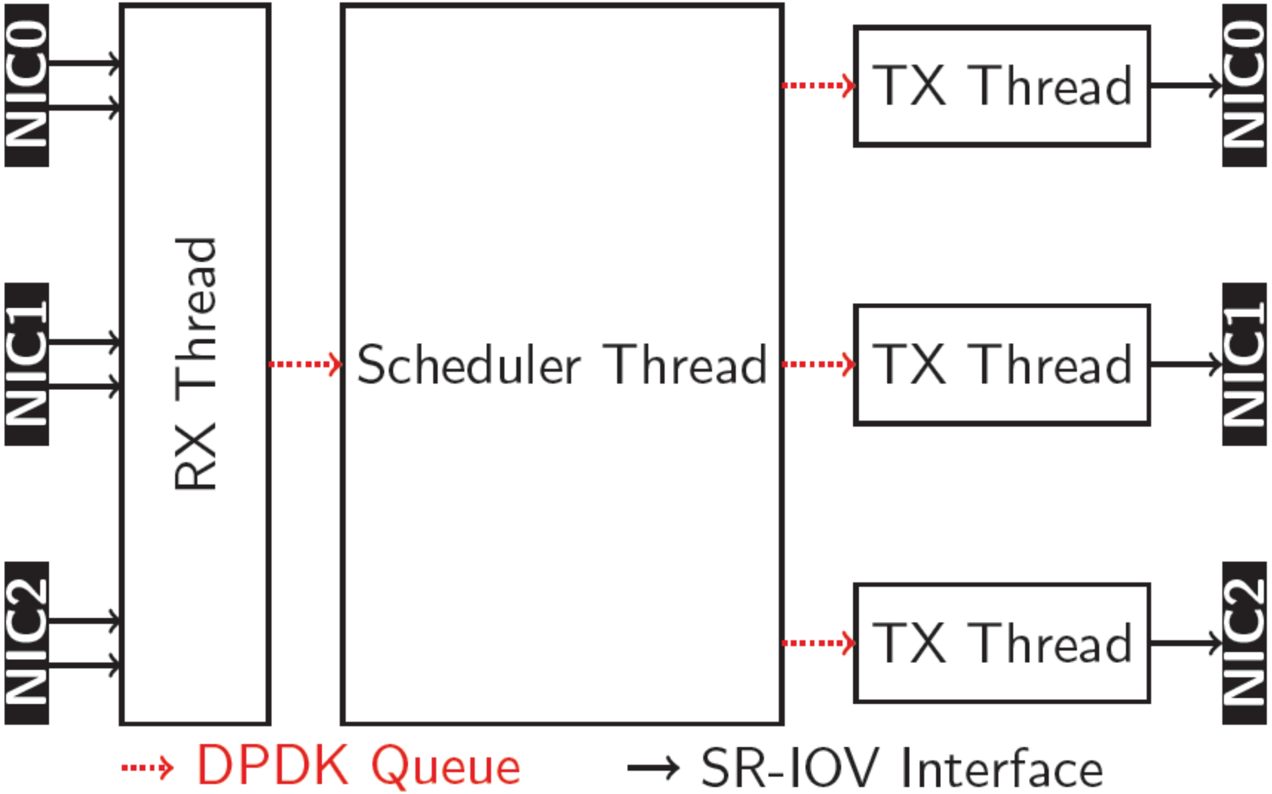
- Line Cards
  - Smart NICs
- Switch Fabric
  - PCI Express
- Fabric Controller
  - CPU orchestrates packet transfers
  - Small bi-partite matching problem



Flexible Datapath



Real-world Implementation



Hybrid Packet Switching I

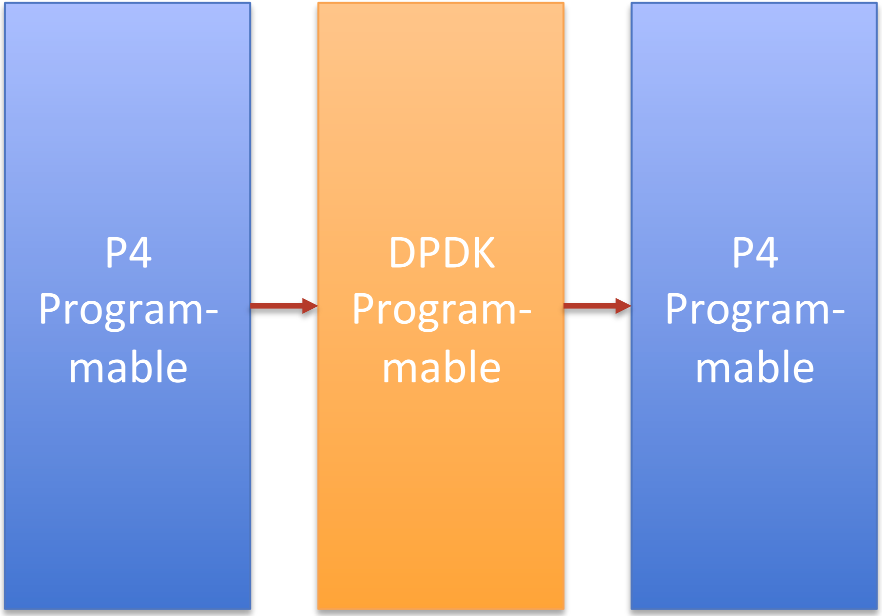
Transfer Metadata

- To be efficient in an architecture with multiple programmable stages we need be able to transfer metadata between stages
- Example:

Output port is determined in the first stage and the second stage needs to know the location of egress port in order to forward it.

Location: <Card#, Port#>

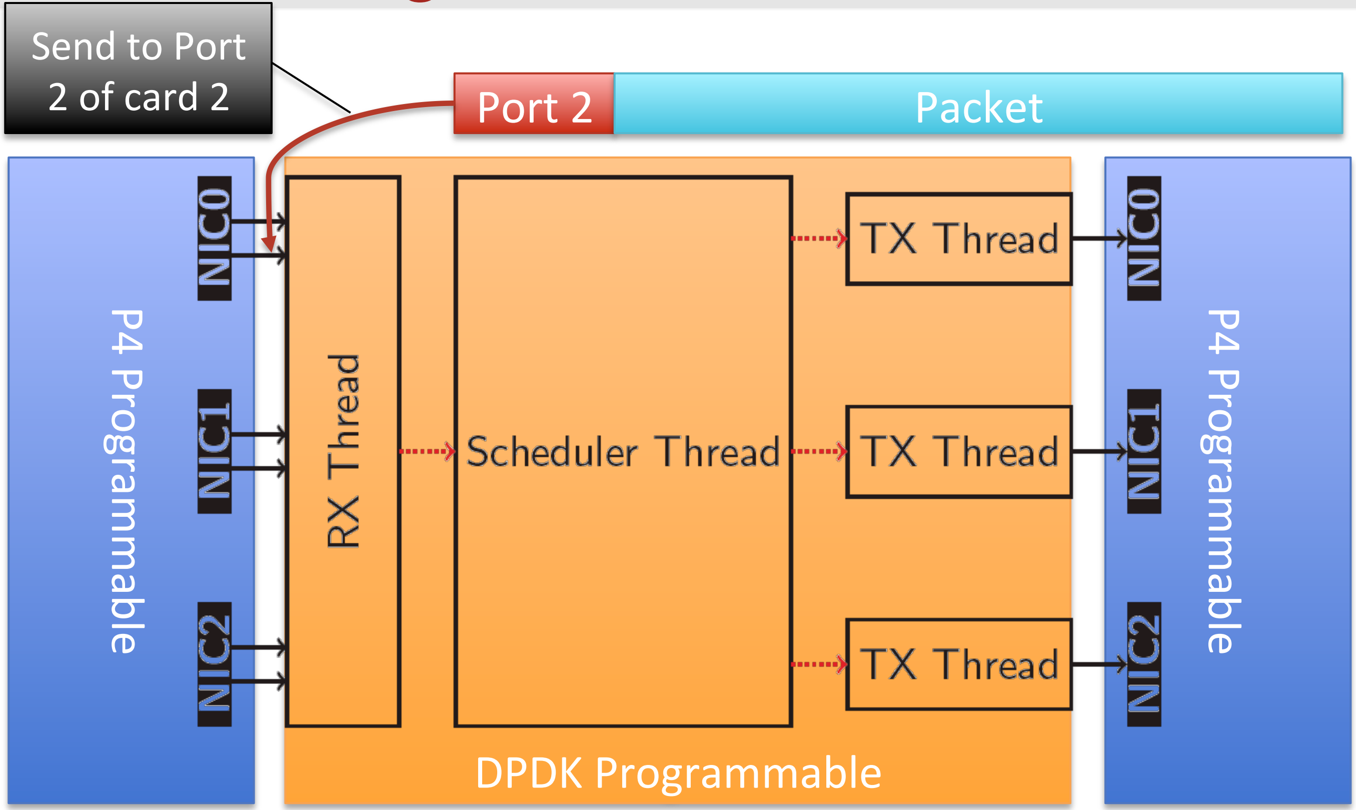
- Card# is associated with queue address
- Port# is added to the packet



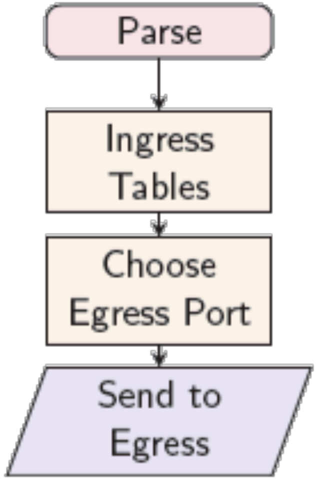
Hybrid Packet Switching II

Transfer Metadata

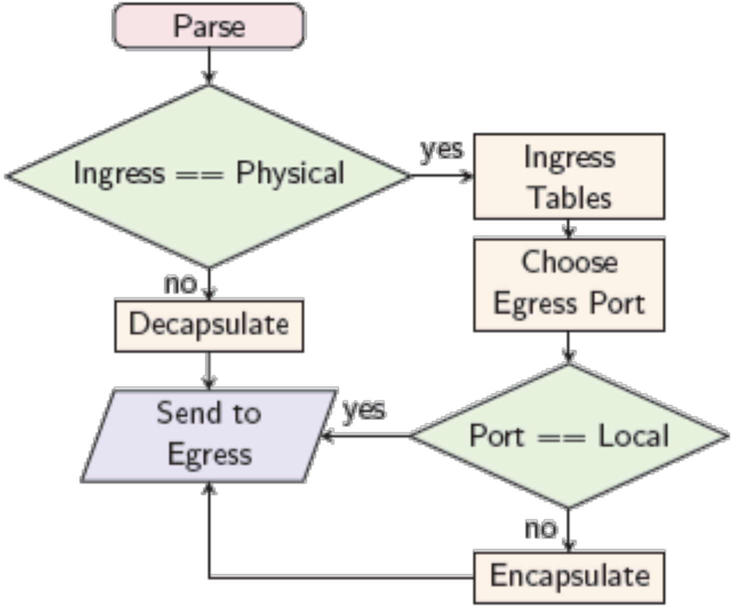
- Metadata is transferred between the stages by adding additional headers to the packet at source stage. The destination stage parses and removes the custom header.



P4 Translation I

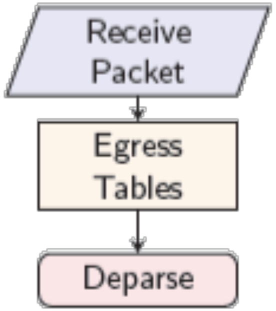


(a) Switch ingress

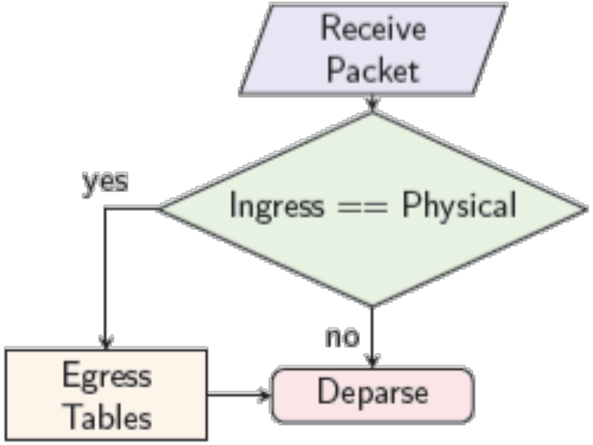


(b) Cards ingress

P4 Translation II

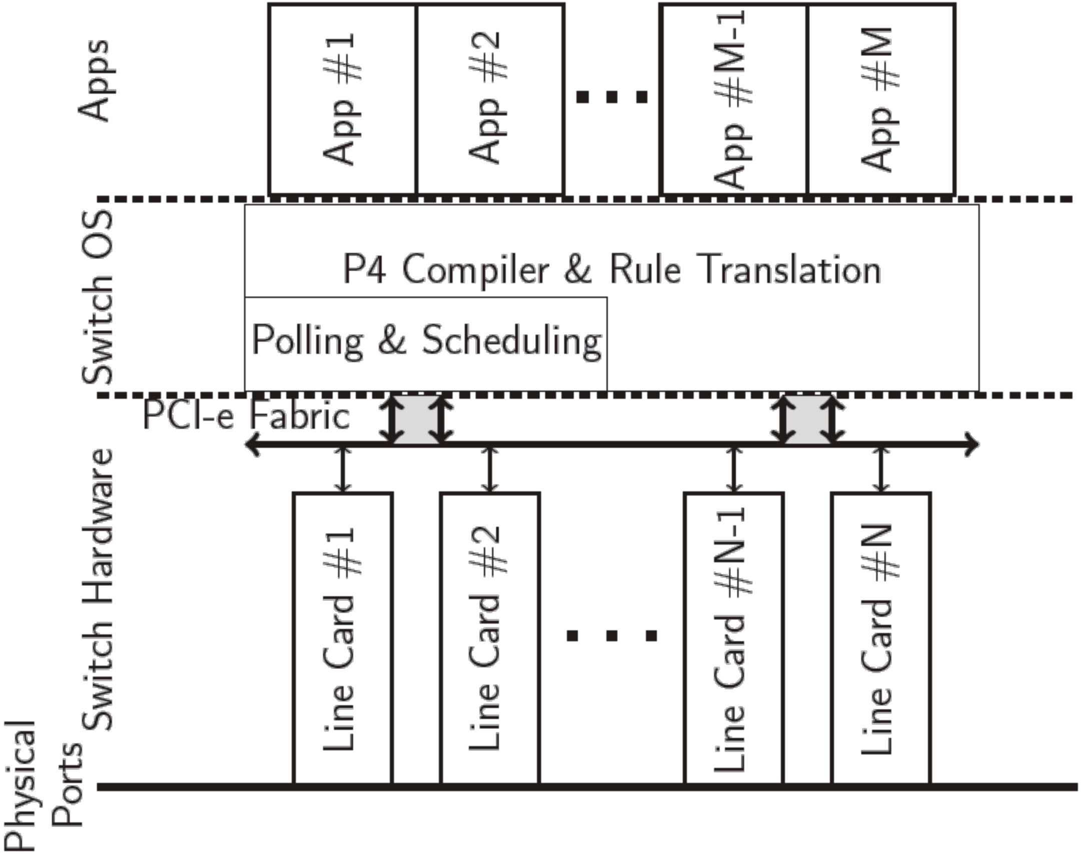


(c) Switch egress



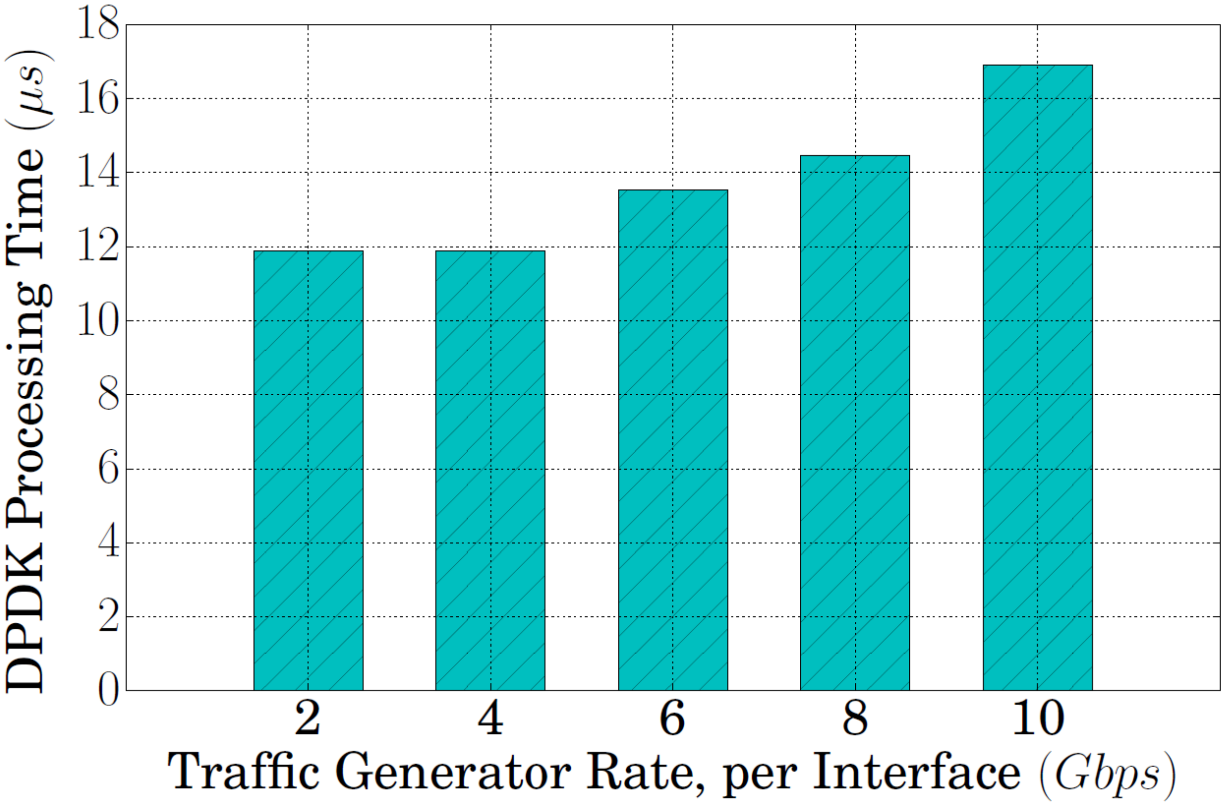
(d) Cards egress

Modularity

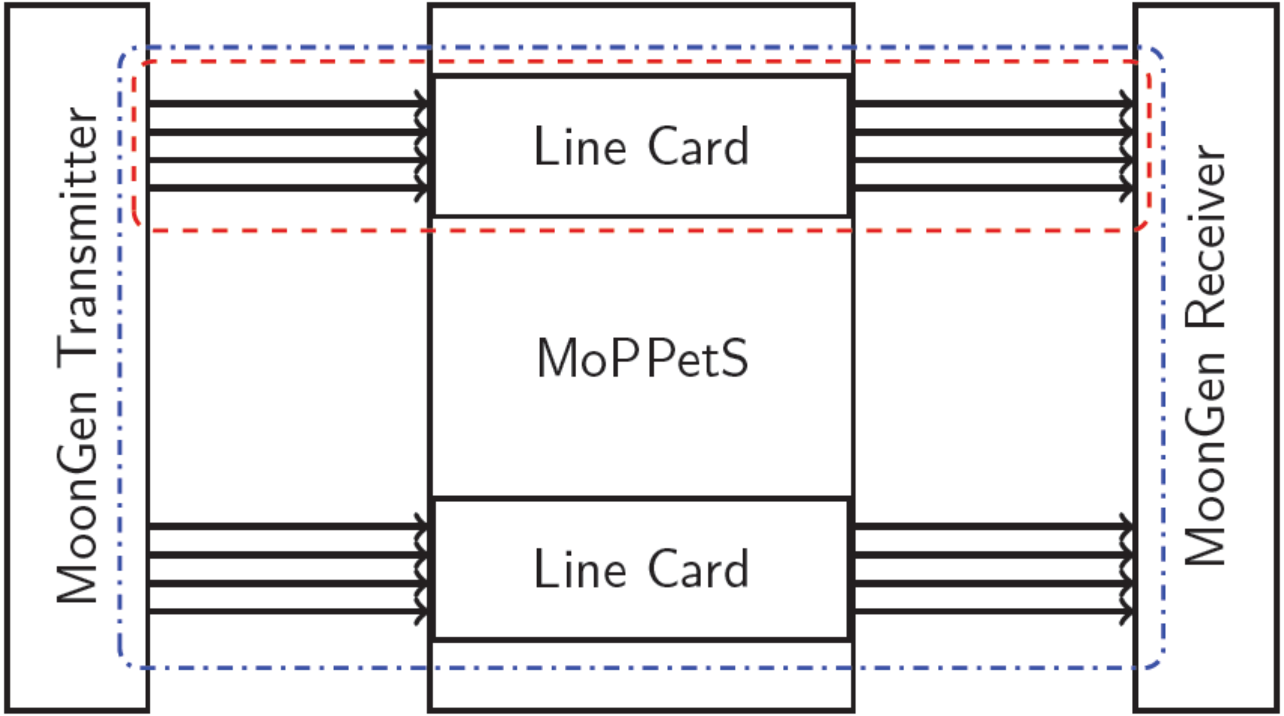


Apps
App #1
App #2
. . .
App #M-1
App #M
Switch OS
P4 Compiler & Rule Translation
Polling & Scheduling
PCI-e Fabric
Switch Hardware
Line Card #1
Line Card #2
. . .
Line Card #N-1
Line Card #N
Physical Ports

Preliminary Results



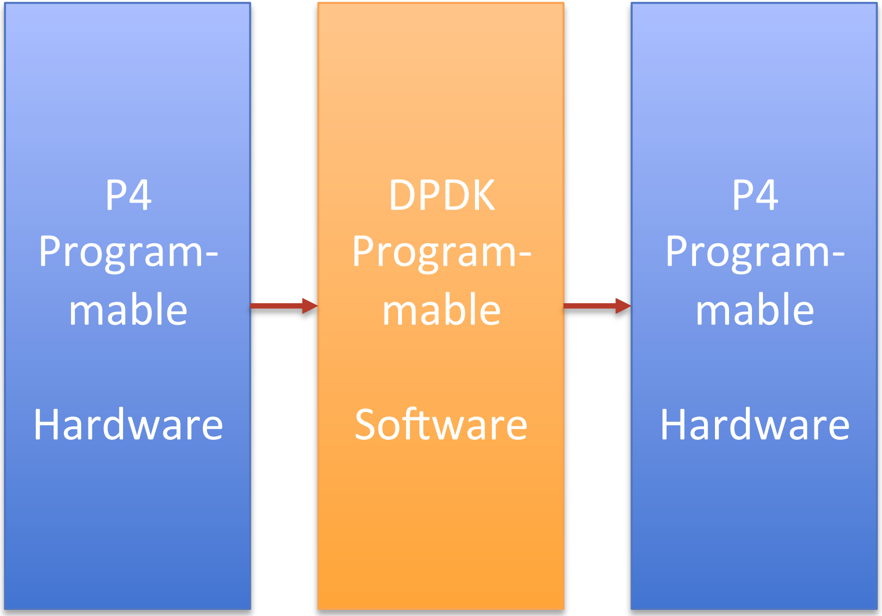
Demo


MoonGen Transmitter
Line Card
MoPPetS
Line Card
MoonGen Receiver

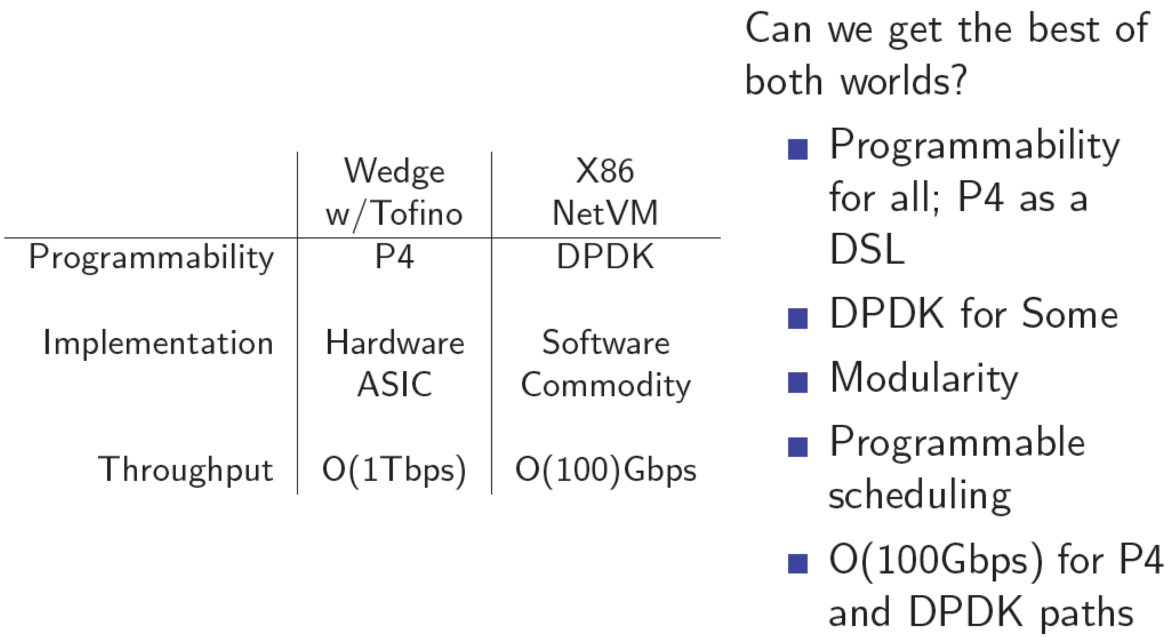
Implications of Hybrid Switching

Network functions can be offloaded to a suitable stage

- Hardware
  - Parsing packets
  - Table operations
- Software
  - Programmable scheduling
    - VLAN priority queues[aghdai17design]
    - PIFO[sivaraman16programmable]
  - Complex monitoring operations
  - Domino atoms[sivaraman16packet]



Conclusion



	Wedge w/Tofino	X86 NetVM
Programmability	P4	DPDK
Implementation	Hardware ASIC	Software Commodity
Throughput	O(1Tbps)	O(100)Gbps

Can we get the best of both worlds?

- Programmability for all; P4 as a DSL
- DPDK for Some
- Modularity
- Programmable scheduling
- O(100Gbps) for P4 and DPDK paths

Full citation:

Aghdai, Ashkan, Yang Xu, and H. Jonathan Chao. "Design of a Hybrid Modular Switch." In *Network Function Virtualization and Software Defined Networks (NFV-SDN), 2017 IEEE Conference on*. IEEE, 2017.

Preprint: aghdai17design