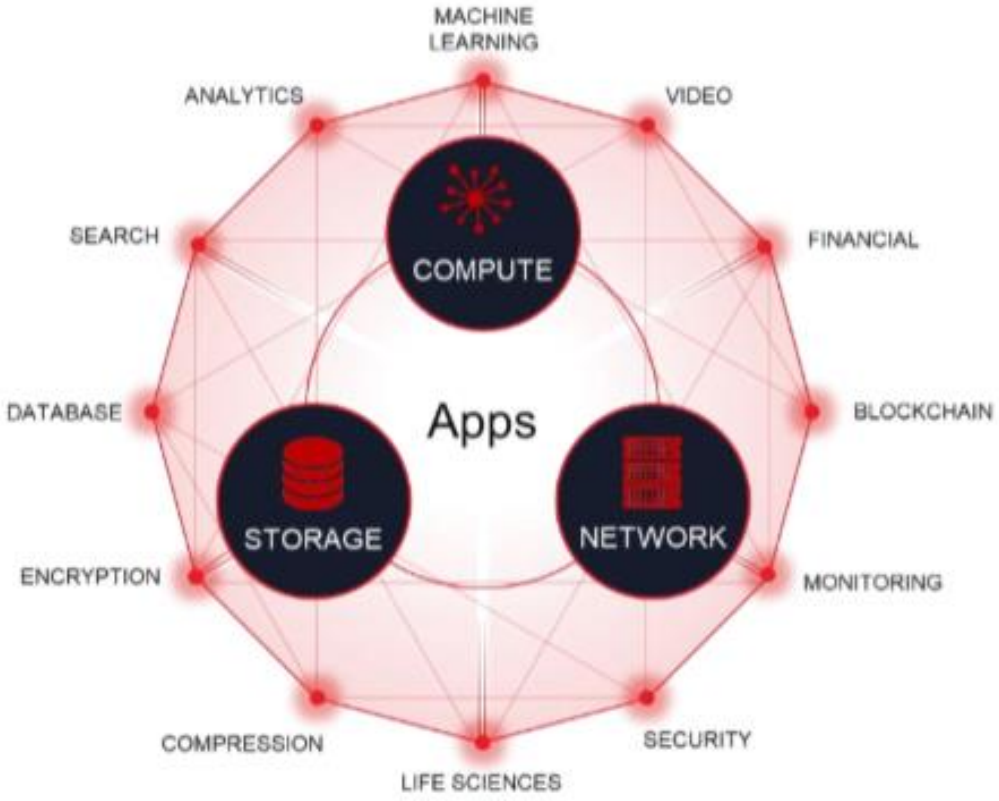
XILINX

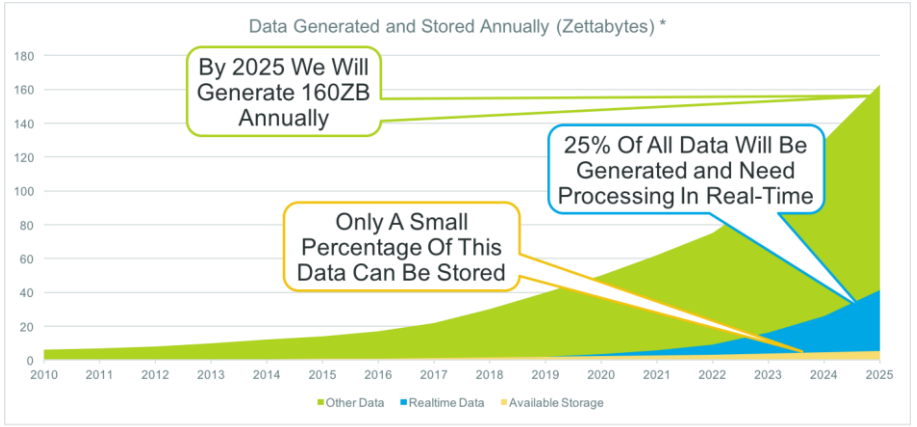
Towards Distributed Adaptive Computing

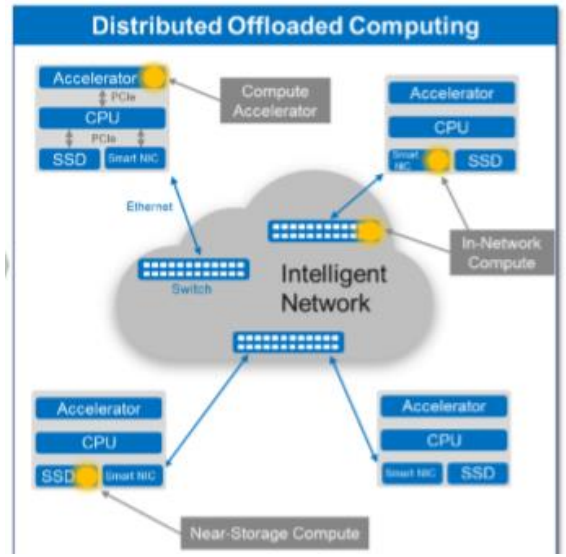
Chengchen Hu

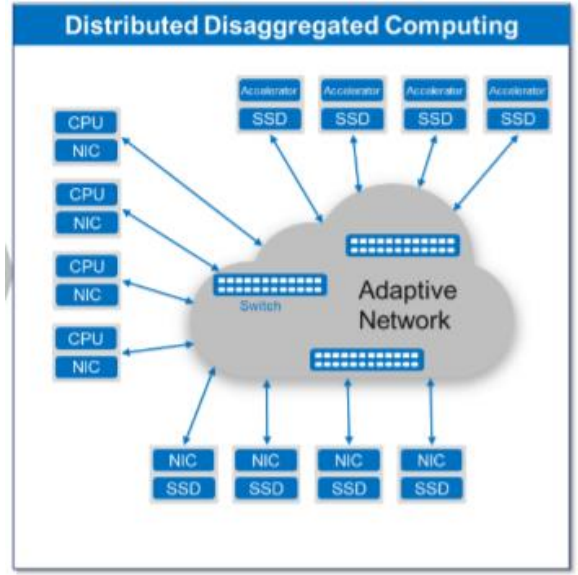
Xilinx Labs
July 24, 2020

Trends in DCs



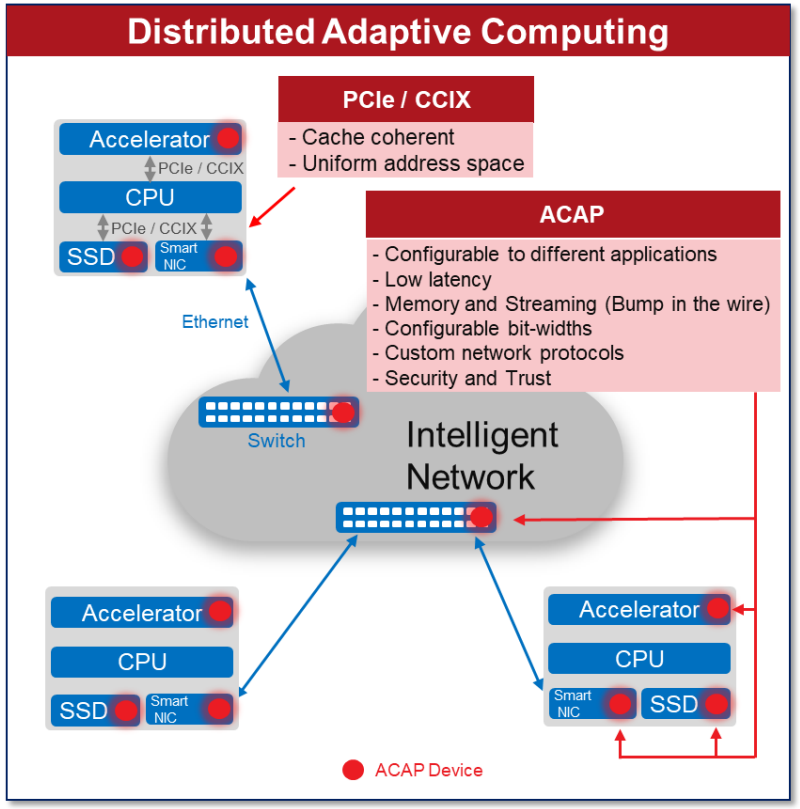






XILINX.

Concept of Distributed Adaptive Computing

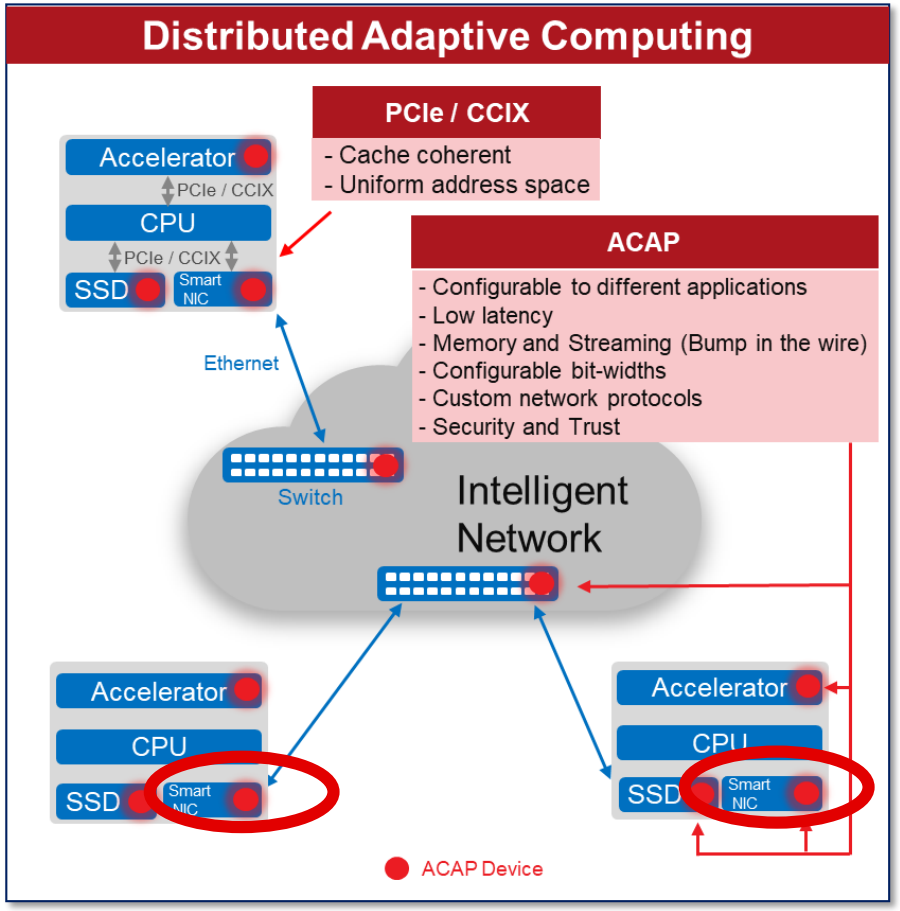


Adaptable NIC:
a new type of SmartNIC as host network device

Adaptable Switch:
Programmability in intermediate network nodes

XILINX.

Adaptable NIC



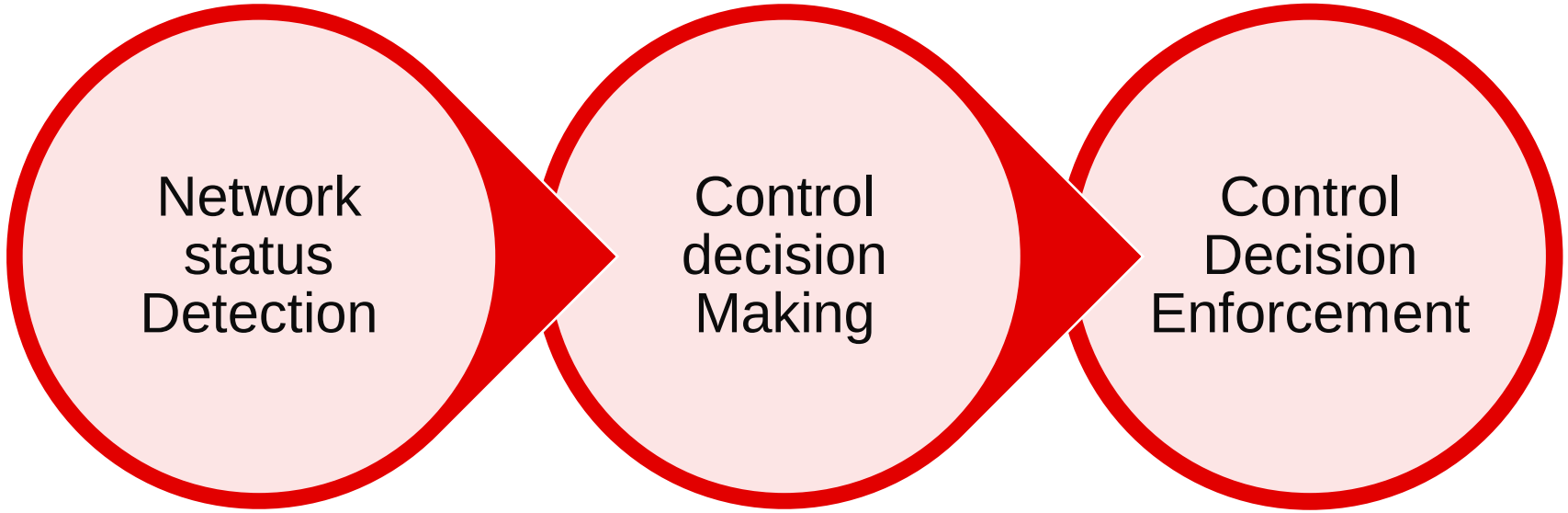
	Description	Features
Type 1	Basic Connectivity NIC	• Basic offloads, simple virtualization
Type 2	SmartNIC for Network Acceleration	• Crypto, vSwitch, custom tunneling
Type 3	SmartNIC for Network + Compute + Storage Acceleration	• Machine Learning, video transcoding • Database Analytics, smart storage
Adaptable NIC	SmartNIC as standalone managed network node	• Domain Specific Programmable Engines offloaded

XILINX.

Why it is adaptable: HW+SW Abstraction

Hardware function

Software function



PPM: Programmable Passive Measurement Engine

PCC: Programmable Congestion control Engine

PTM: Programmable Traffic Management Engine

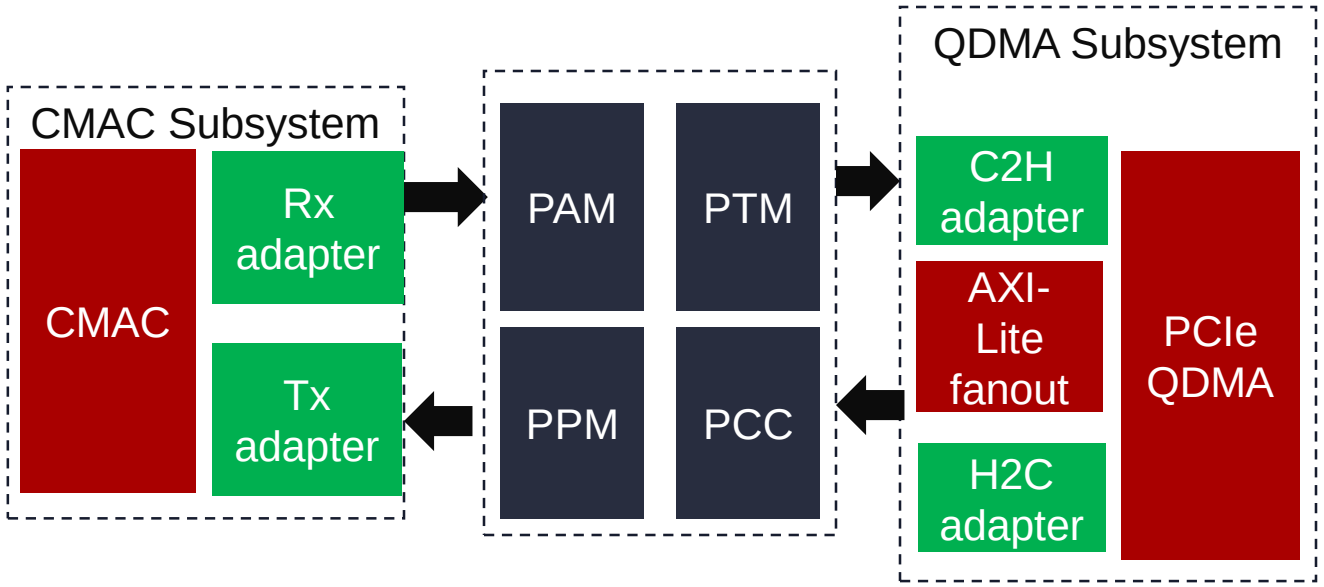
PAM: Programmable Active Measurement Engine

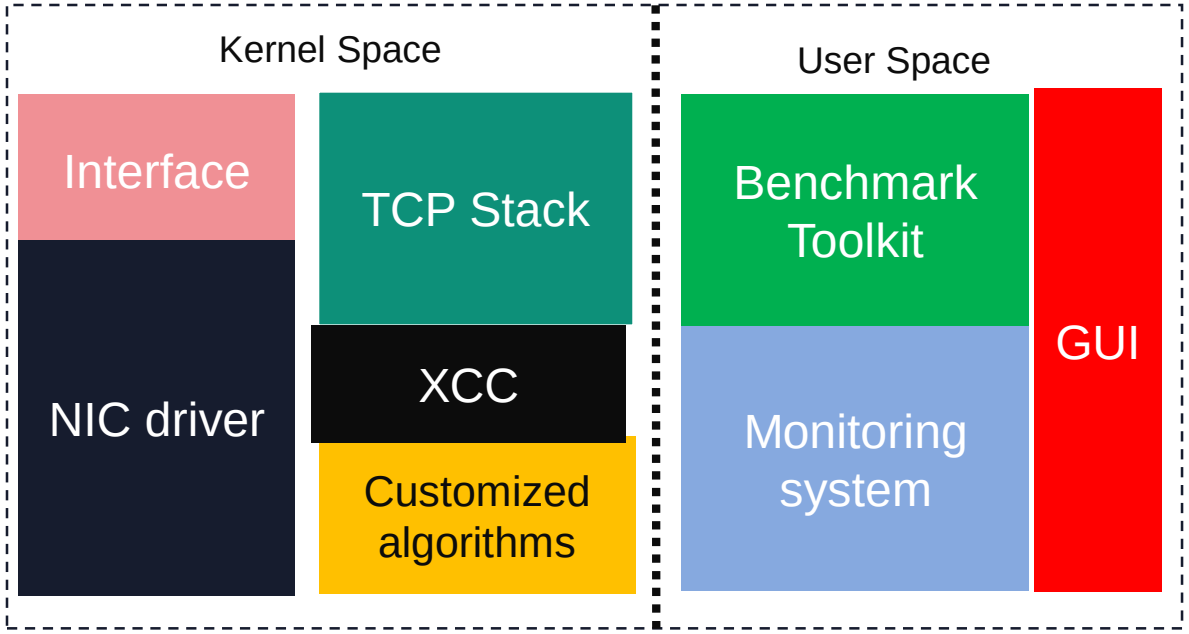
Customized Algorithm in Software

XILINX.

Adaptable NIC prototype

Alveo U250





- **NIC Hardware**
- Network measurement: PPM(passive), PAM(active)
- Network Congestion Control: PCC
- Network Traffic Shaping: PTM

- **NIC Software at host server**
  - Network Congestion Control: XCC
  - Network Monitoring: Monitoring System, GUI
  - Network Benchmark Generation Toolkit

XILINX.

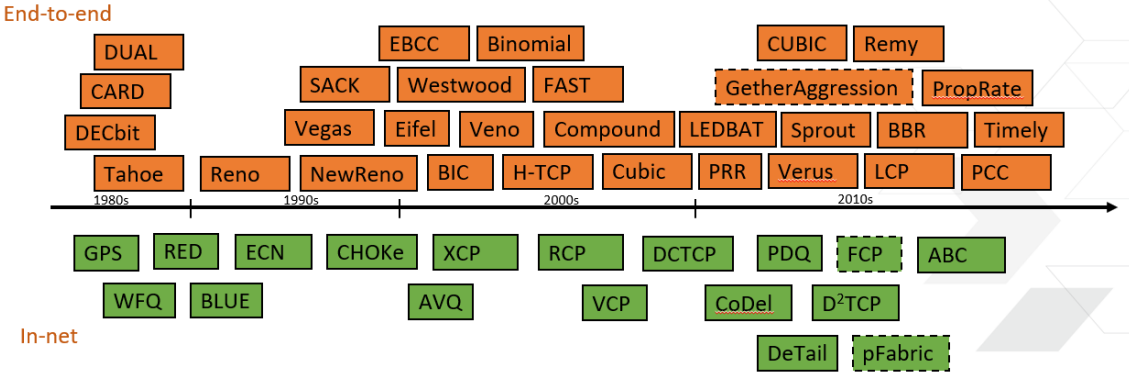
Story of Transport Congestion Control

- Congestion Control on HW
  - TCP variants on HW (Chelsio).
  - Socket API (TCP offload engine).
  - RMDA(iWARP).
  - RoCE using DCQCN (Microsoft).
  - HPCC (Alibaba),FPGA based.

> “needed to modify the data delivery algorithm to avoid livelocks in their network but had to rely on the NIC vendor to make that change.”

- Congestion Control keeps evolving
  - Environment changes
  - Application requirement changes

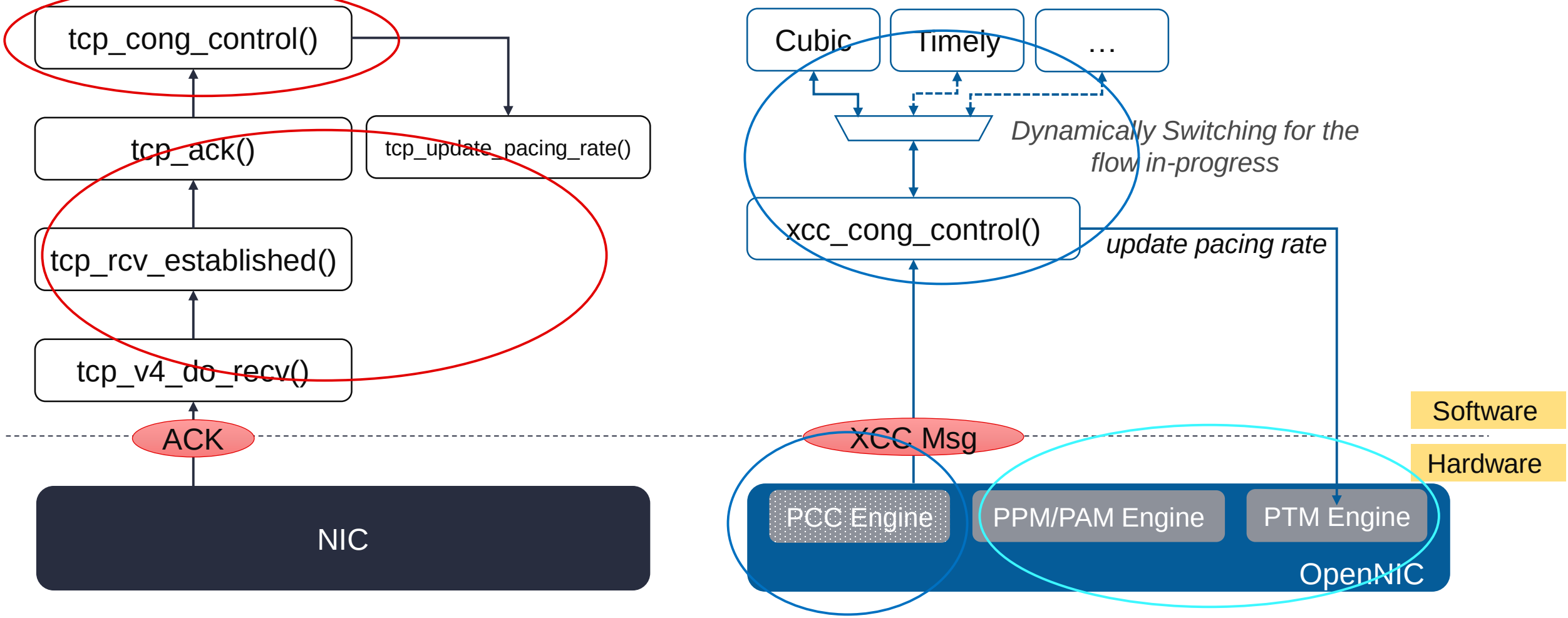
- Programable CC on FPGA
  - TONIC (Princeton & MIT), NSDI 2020
  - Limited programmability provided



XILINX.

Mapping CC to Adaptable NIC

The CC algorithm is fixed for the flow in-progress

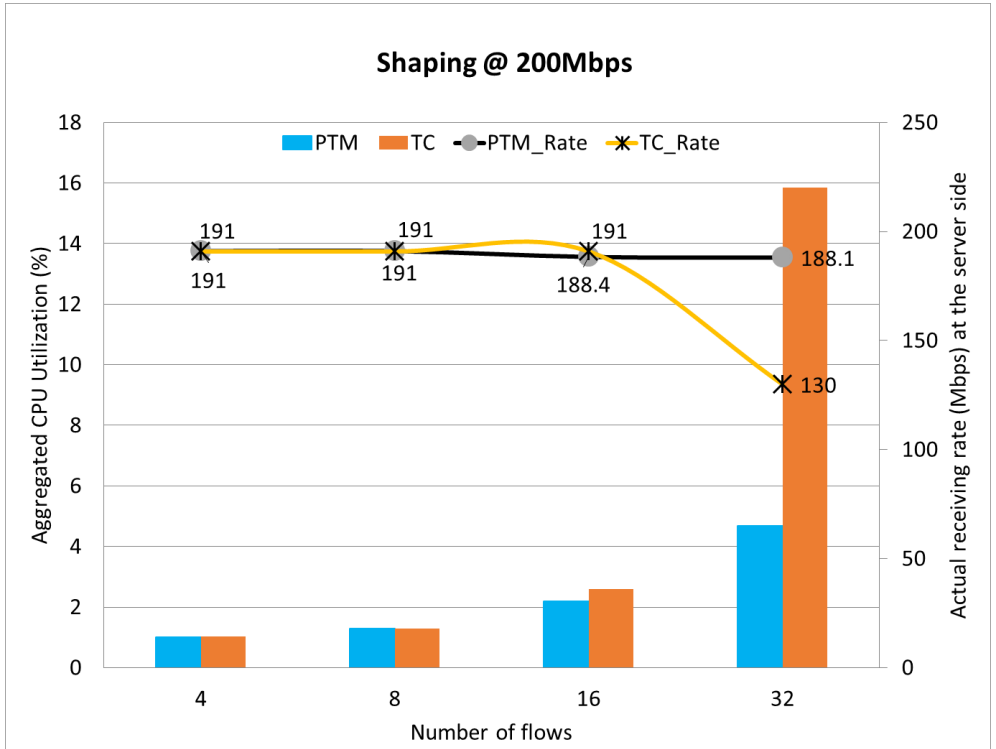


*Parts of the **xcc_cong_control()** could be further moved to the **PCC Engine***

XILINX.

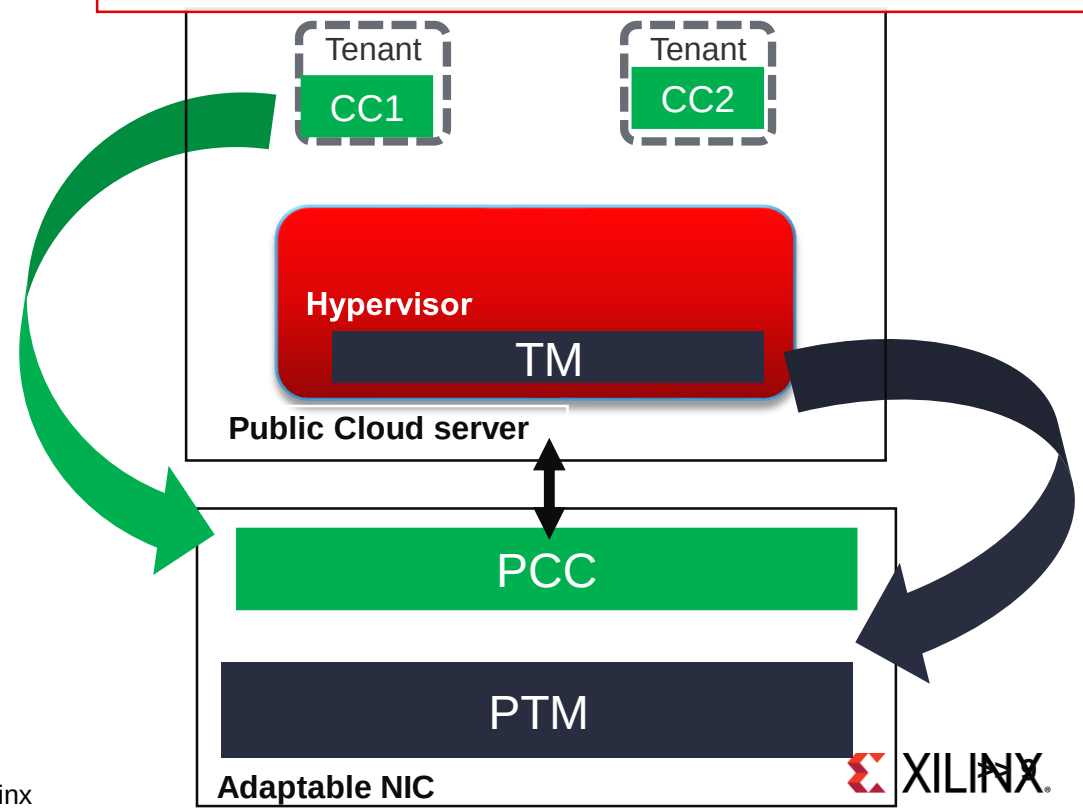
Use Case: Public Cloud

- **Tenants Share Machine in Public Cloud**
- **Each tenant can customize CC policy for service optimization, with the support of PCC in Adaptable NIC**
- **Cloud operator define scheduling/policing/shaping in hypervisor supported by PTM in Adaptable NIC**
- **QoS guarantee on both ingress and egress**

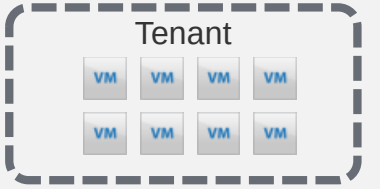


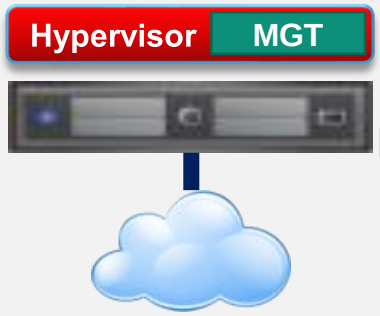
Management in Adaptable NIC

- Less CPU, more income
- Fine-grained bandwidth control and guarantee



Use Case: Bare Metal Cloud

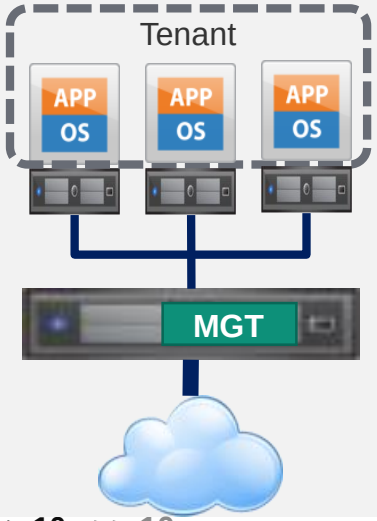




VM cloud issues

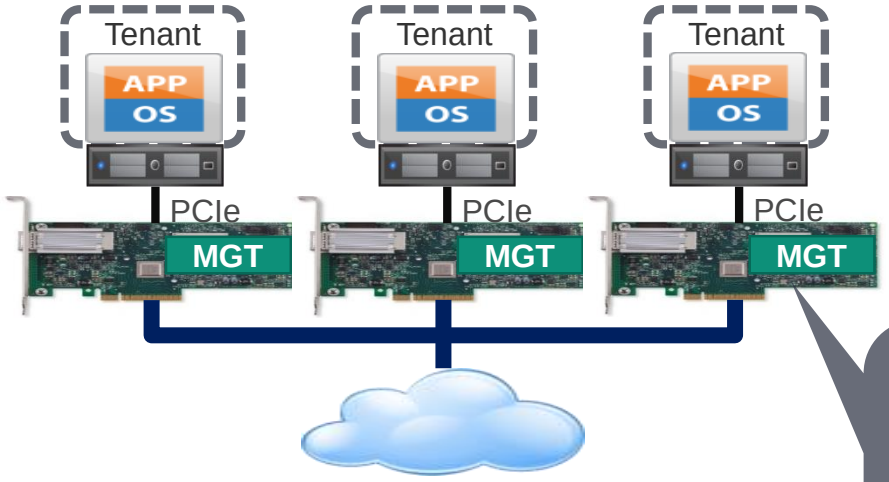
- Hardware resource
- Performance
- Mitigation and licenses

Dedicated bare metal machine(s) for one tenant



Dedicated gateway

- Performance bottleneck
- Hard to scale out and scale up



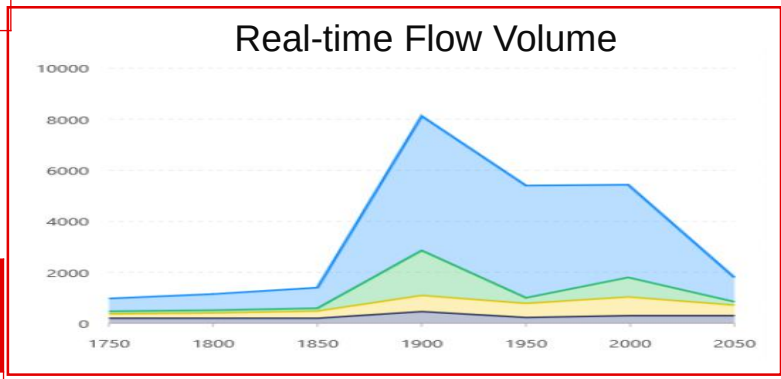
Measurements

- Usage Statistics
- Anomaly Detection
- Heavy jitter
- Active Probes
- ……

IP Address	Port	Alarm
192.168.1.12	22	True
192.168.1.17	1433	True
192.168.1.18	80	True

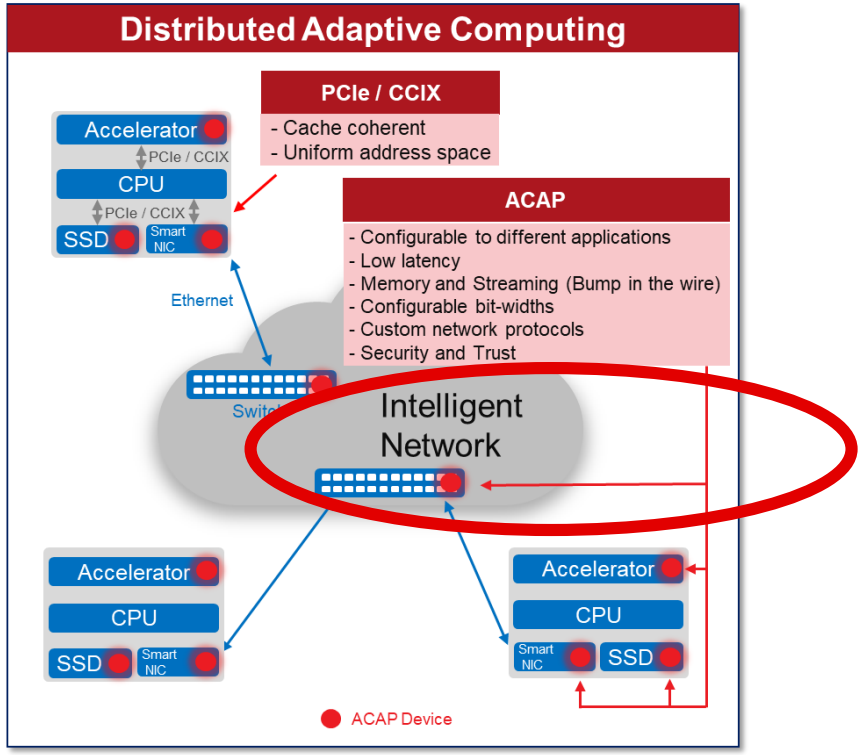
Flow ID	Heavy-changer
2	True

IP Address	Port Scanner
192.168.1.8	False
192.168.1.5	True



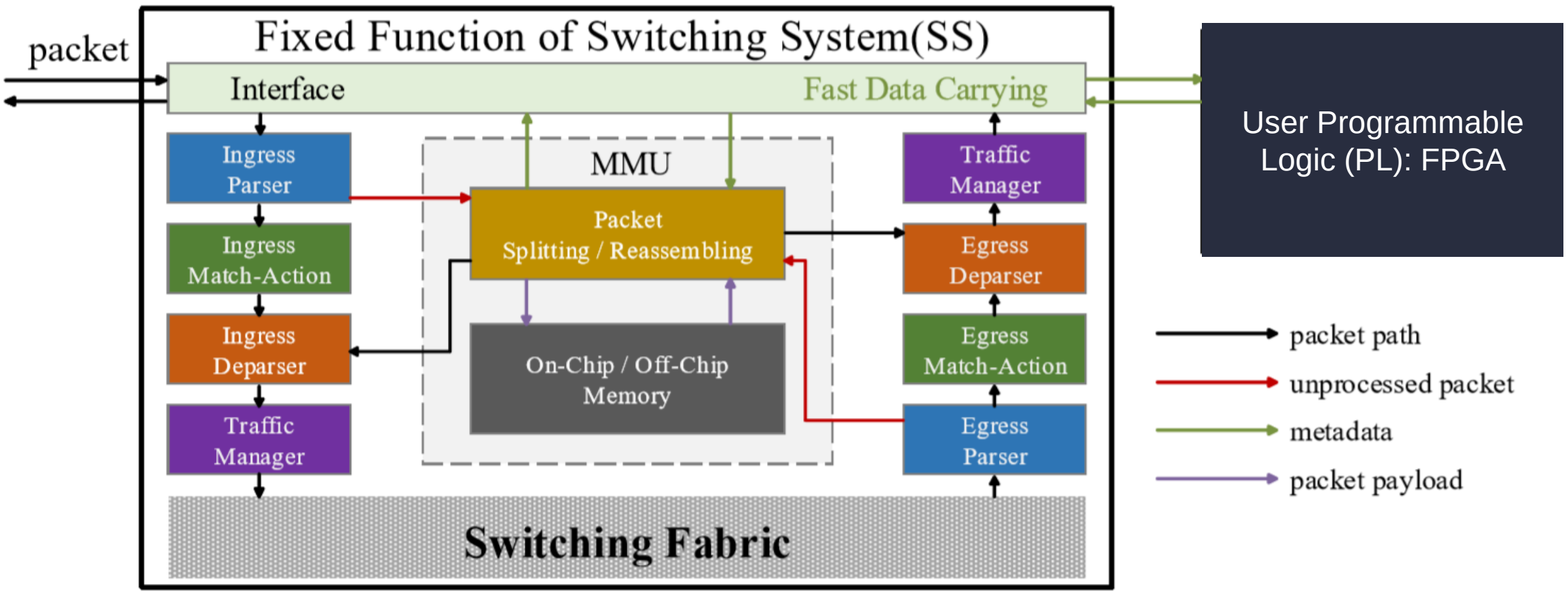
XILINX

Adaptable Switch



- Push adaptivity to switch for new opportunities.
  - In-network compute, e.g. distributed ML
  - In-network storage, e.g., cached streaming video
- Step 1: Two-chip approach
  - Combined software stacks
  - ‘Flow-through’ processing architecture
  - Switch chip and Xilinx chip on mother board
- Step 2: Integrated chip
  - Integrated software stack
  - ‘Lookaside’ processing architecture
  - Integrated switching and processing

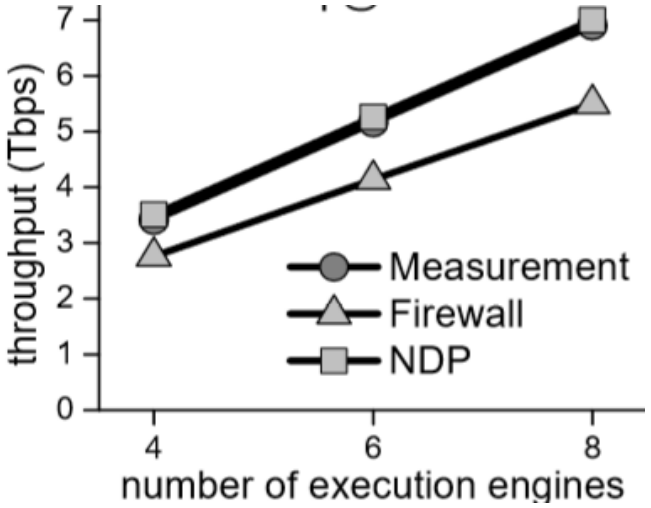
Architecture of Adaptable Switch



XILINX.

Use case: Programmable Networking

- Event Driven processing
  - NDP – trigger congestion control logic by tracking buffer occupancy
- Stateful processing
  - Firewall
- Algorithmic capacity
  - Measurement – Statistical calculation



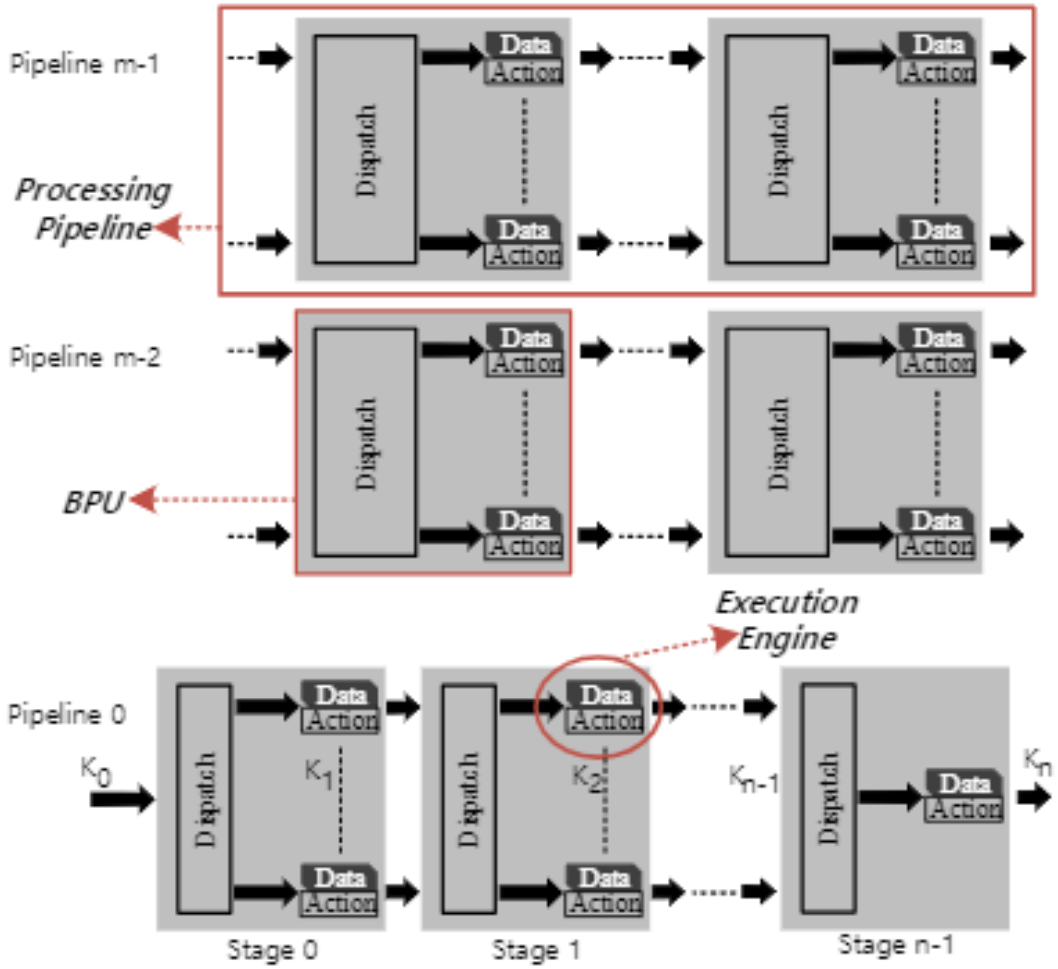
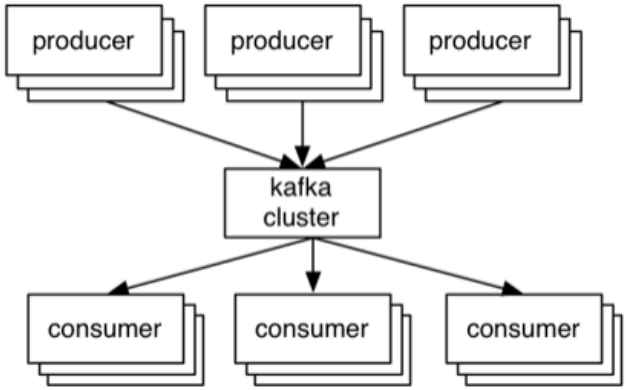


Fig. 3. A General Model of Parallel processing Pipeline in PL.

Use case: In-Networking computing

- Kafka is a data streaming platform to manage/process data from multiple input streams
- How to enable a Kafka cluster in an Adaptable Switch
  - Add disks to Adaptable Switch attached to FPGA/MPSoC
  - FPGA/MPSoC enables brokers and zookeeper in the Kafka cluster
  - Filters out date to be kept in Kafka and offload to ACAP/FPGA/MPSoC for data



Summary

- Distributed Adaptive Computing is flexible to enable user specific functions
  - Requirements vary user-by-user
  - Adaptable NIC/Switch provide new programming ways to deploy application accelerations

- User cases show value for cloud data center
  - Free server CPU for more users/tenants to increase revenue
  - Increase scalability (e.g., remove gateway for Bare Metal cloud)
  - Innovate with new functions, proprietary protocols and customized processing.

Thank You