Smart (Programmable) NICs

ECE/CS598HPN

Radhika Mittal

Microsoft Case Study

Azure Accelerated Networking: SmartNICs in the Public Cloud

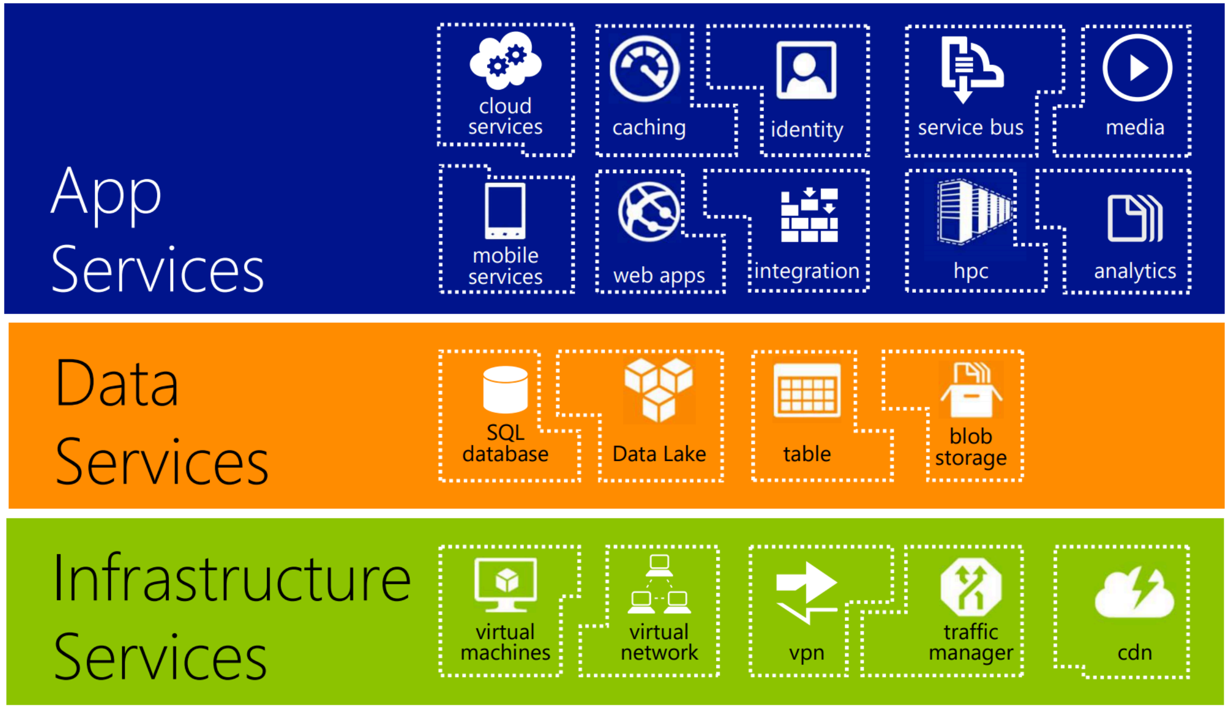
NSDI'18

Slides borrowed from the NSDI talk

Overview

- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

Microsoft Azure



Microsoft

Azure Scale & Momentum

>85%
Fortune 500 using
Microsoft Cloud

>120,000
New Azure customers a month

>60 TRILLION
Azure storage
objects

>900 TRILLION
requests/day

>9 MILLION
Azure Active
Directory Orgs

>18 BILLION
Azure Active Directory
authentications/week

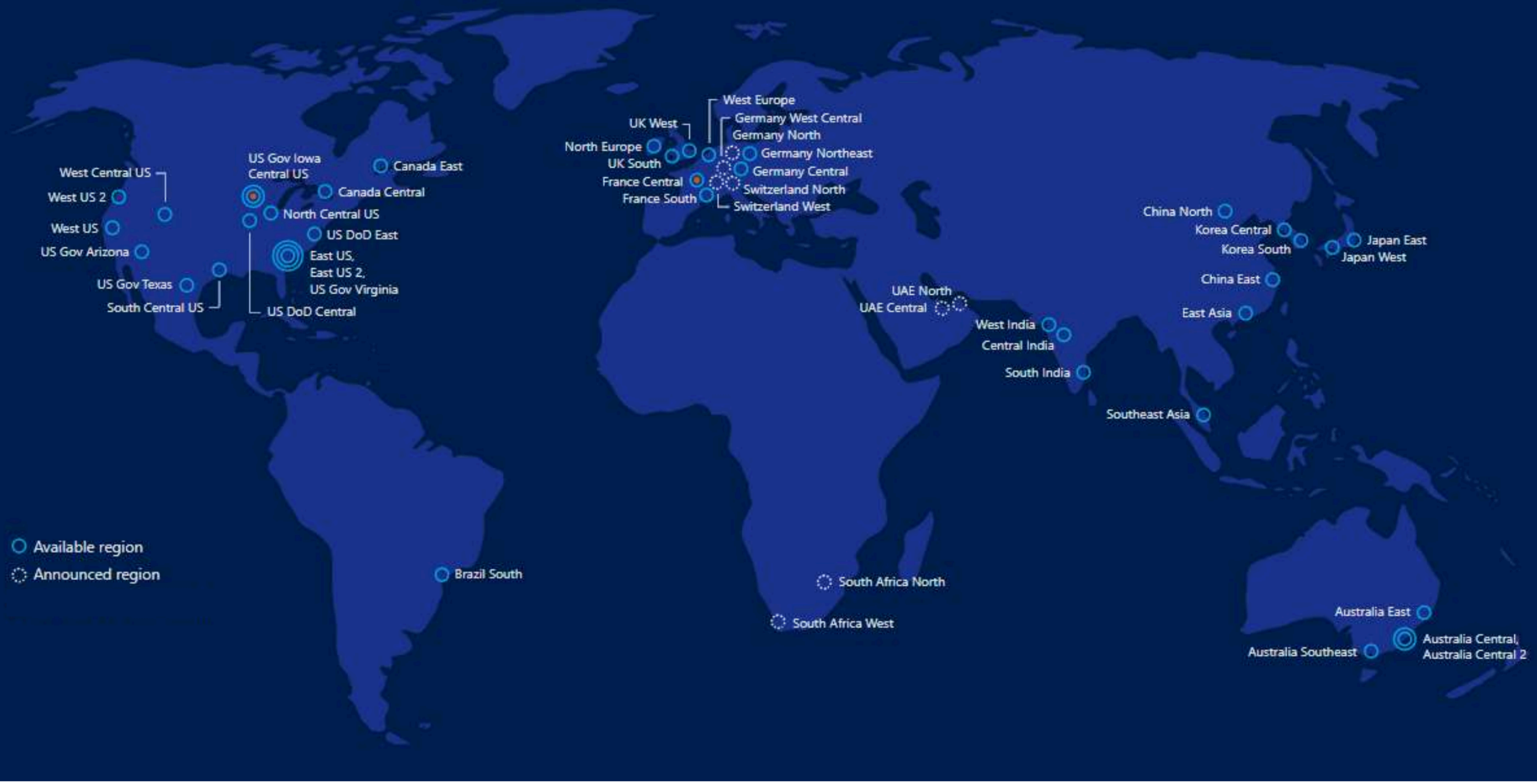
>50% of
Azure VMs
are Linux VMs

> 3 TRILLION
Azure Event Hubs
events/week

>110 BILLION
Azure DB requests/day

Microsoft

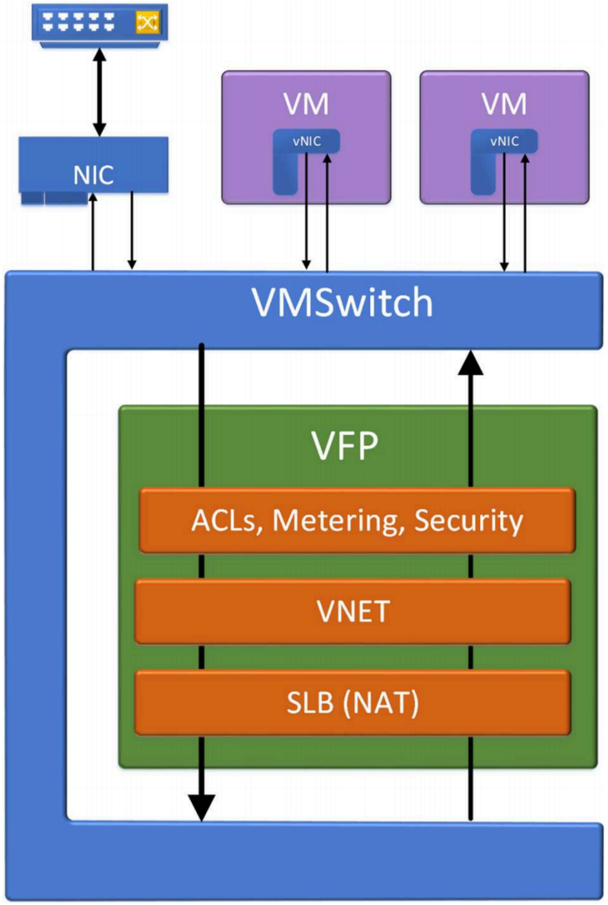
50 Global Regions, Hundreds of DCs, Millions of Servers



Overview

- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

Virtual Filtering Platform (VFP) Azure's SDN Dataplane

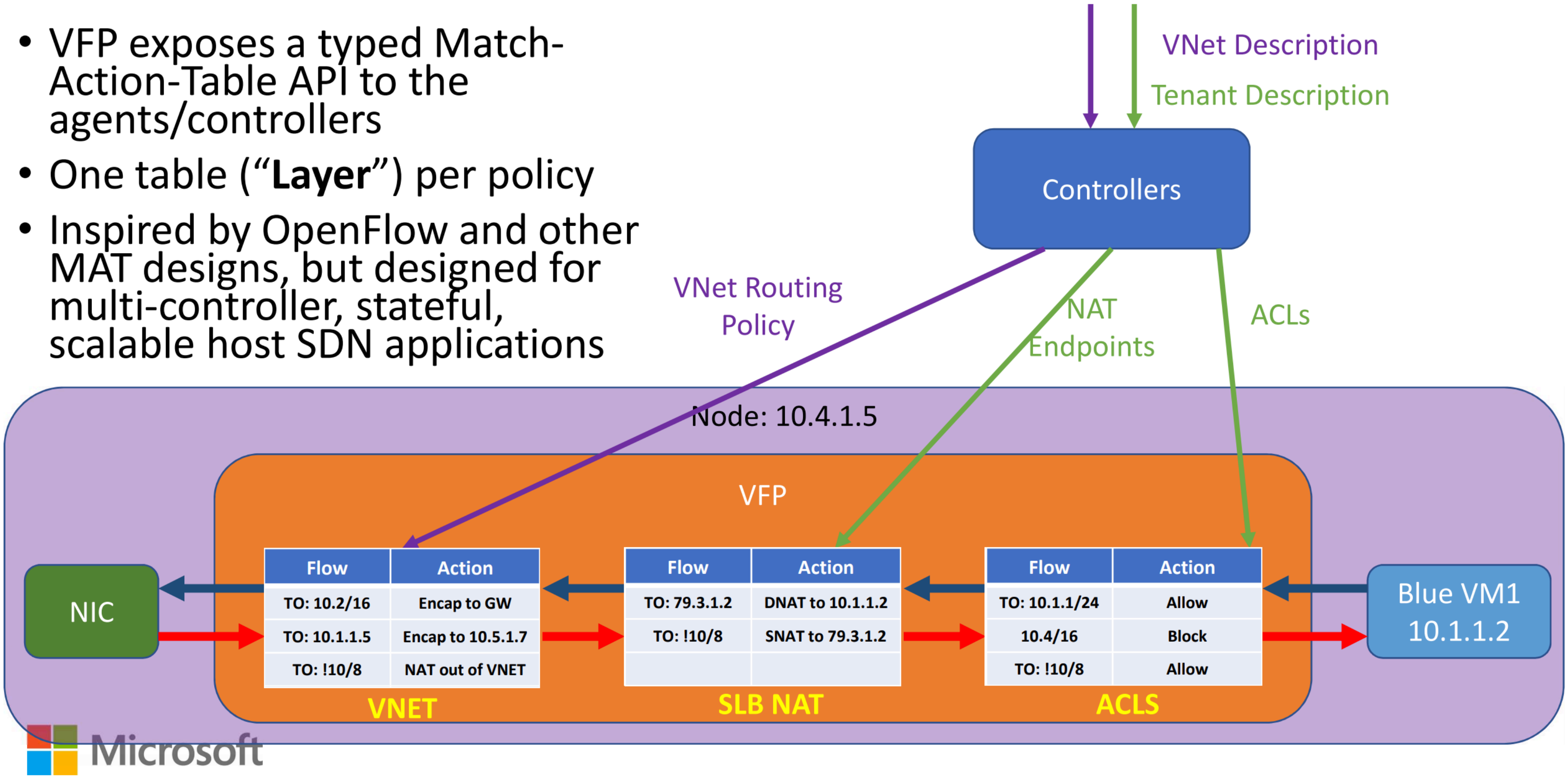


Microsoft

- Virtual switch for Hyper-V / Azure

Key Primitive: Match Action Tables

- VFP exposes a typed Match-Action-Table API to the agents/controllers
- One table (“**Layer**”) per policy
- Inspired by OpenFlow and other MAT designs, but designed for multi-controller, stateful, scalable host SDN applications



VNET

Flow	Action
TO: 10.2/16	Encap to GW
TO: 10.1.1.5	Encap to 10.5.1.7
TO: !10/8	NAT out of VNET

SLB NAT

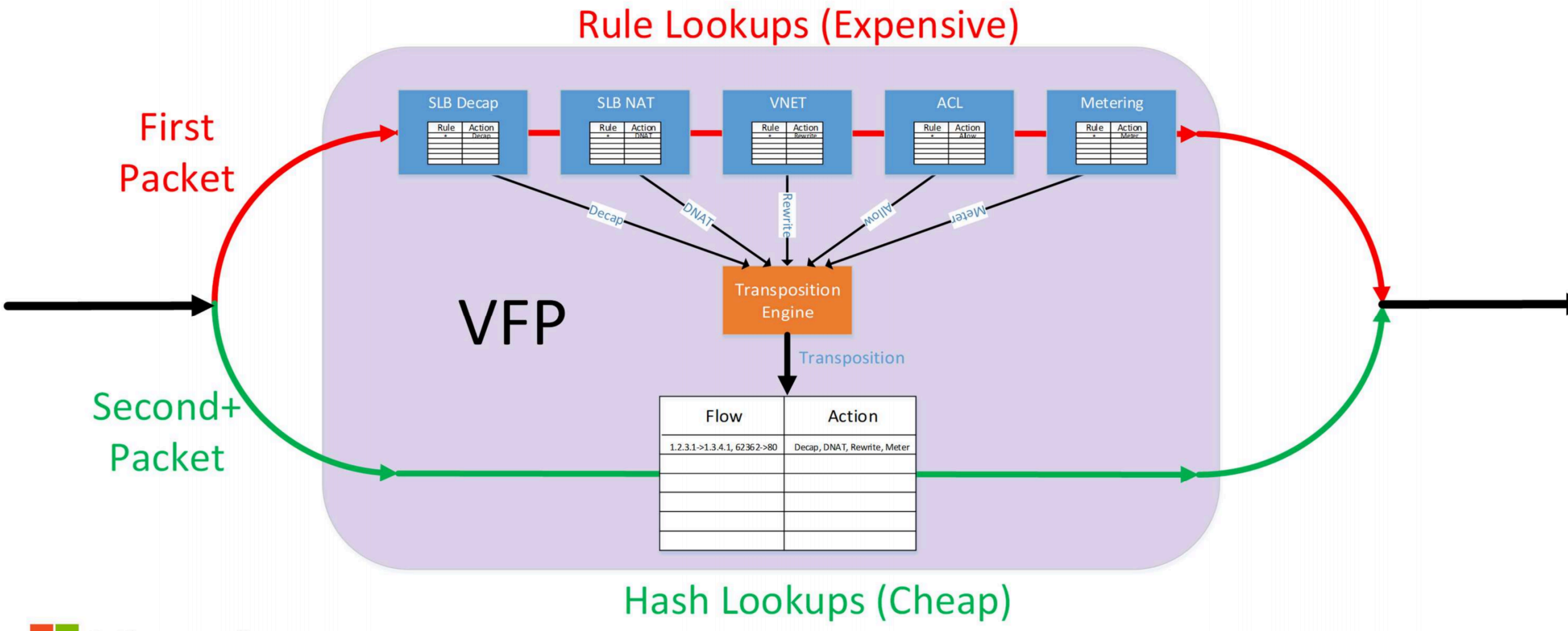
Flow	Action
TO: 79.3.1.2	DNAT to 10.1.1.2
TO: !10/8	SNAT to 79.3.1.2

ACLS

Flow	Action
TO: 10.1.1/24	Allow
10.4/16	Block
TO: !10/8	Allow

Microsoft

Unified Flow Tables – A Fastpath Through VFP



Flow	Action
1.2.3.1->1.3.4.1, 62362->80	Decap, DNAT, Rewrite, Meter

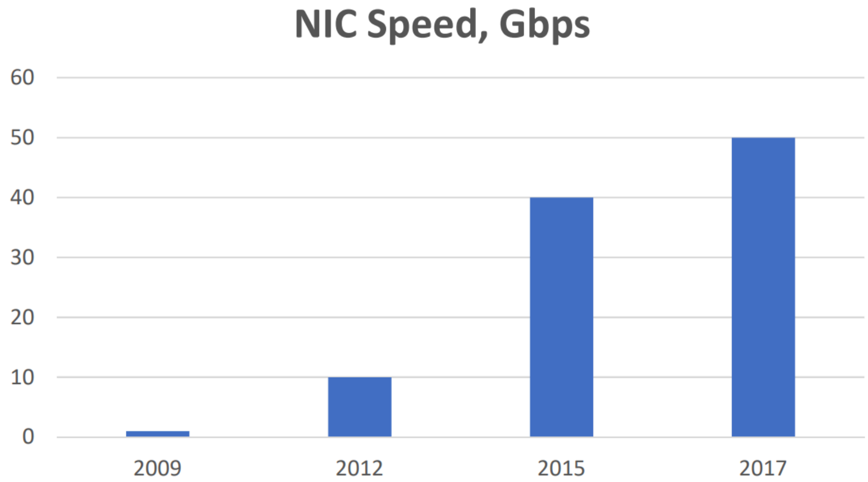
Microsoft

Overview

- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

Scaling Up SDN: NIC Speeds in Azure

- 2009: 1Gbps
- 2012: 10Gbps
- 2015: 40Gbps
- 2017: 50Gbps
- Soon: 100Gbps?



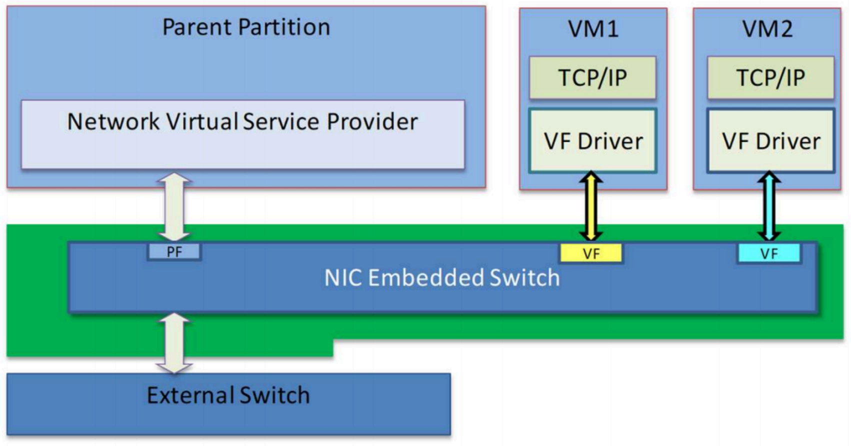
We got a 50x improvement in network throughput, but not a 50x improvement in CPU power!

Microsoft

Host SDN worked well at 1GbE, ok at 10GbE… what about 40GbE+?

Traditional Approach to Scale: ASICs

Example ASIC Solution: Single Root IO Virtualization (SR-IOV) gives native performance for virtualized workloads



Microsoft

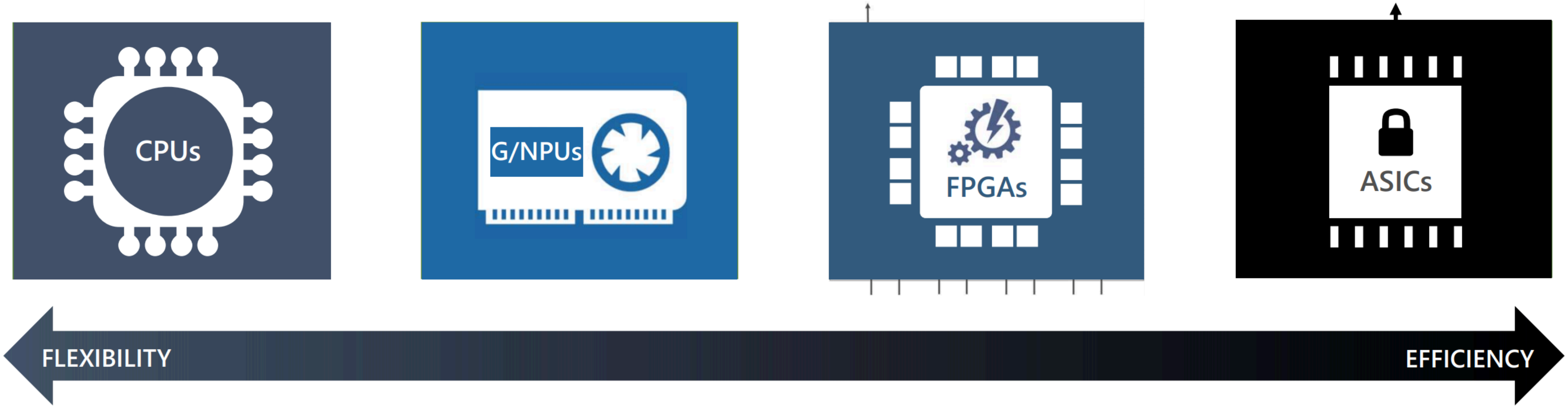
Hardware or Bust

- SR-IOV is a classic example of an “all or nothing” offload – its latency, jitter, CPU, performance benefits come from skipping the host entirely
- If even one widely-used action isn’t supported in hardware, have to fall back to software path and most of the benefit is lost even if hardware can do 99% of the work
- Other examples: RDMA, DPDK, … a common pattern
- This means we need to consider carefully how we will add new functionality to our hardware as needed over time

Overview

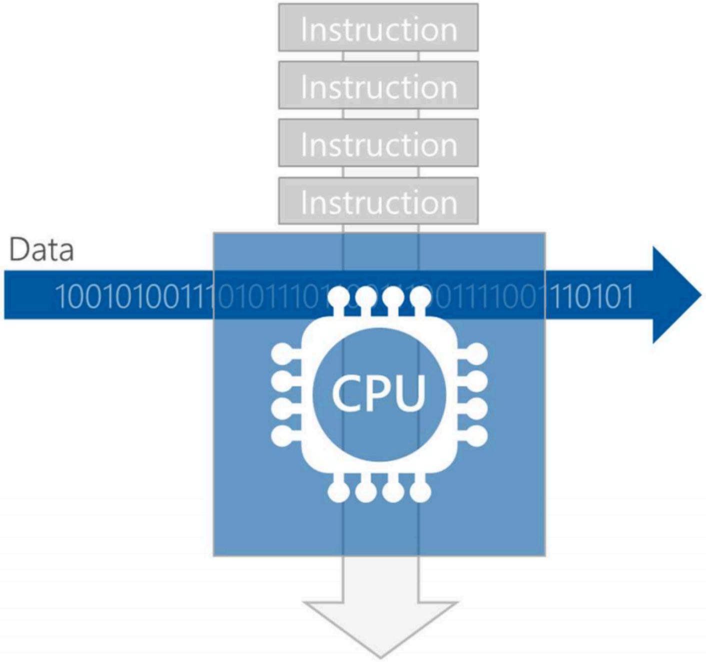
- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

Silicon alternatives



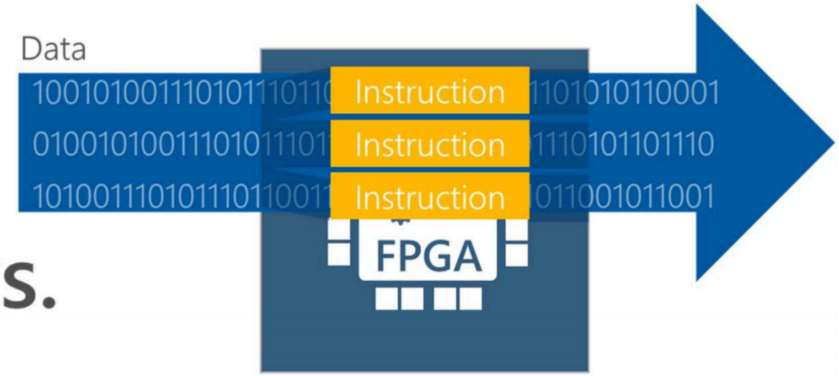
Option 5: Don’t offload at all, instead make SDN more efficient with e.g. poll-mode DPDK

CPU vs. FPGA



CPU: temporal compute

vs.

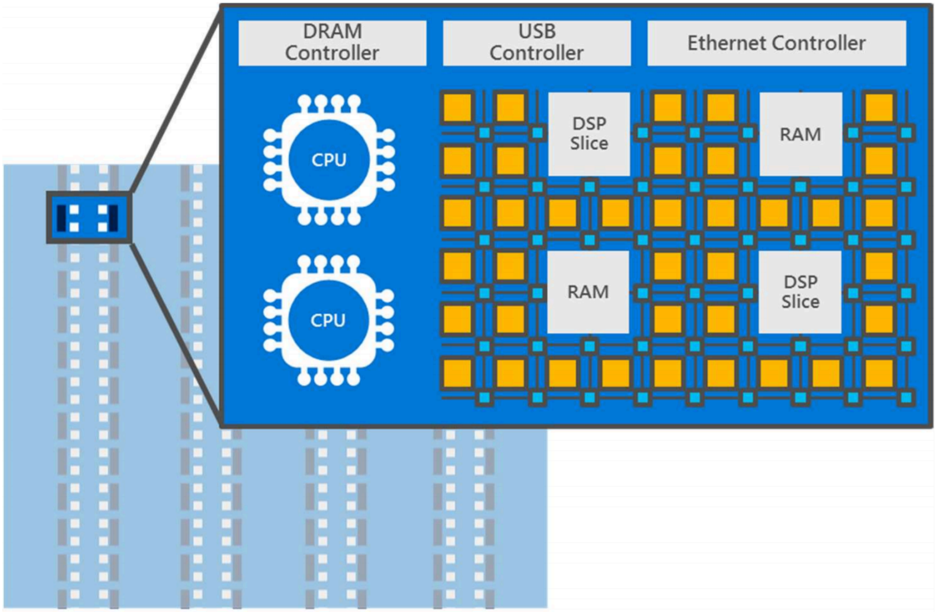


FPGA: spatial compute

Microsoft

What is an FPGA, Really?

- Field Programmable Gate Array
- Chip has large quantities of programmable gates – highly parallel
- Program specialized circuits that communicate directly
- Two kinds of parallelism:
  - Thread-level parallelism (stamp out multiple pipelines)
  - Pipeline parallelism (create one long pipeline storing many packets at different stages)

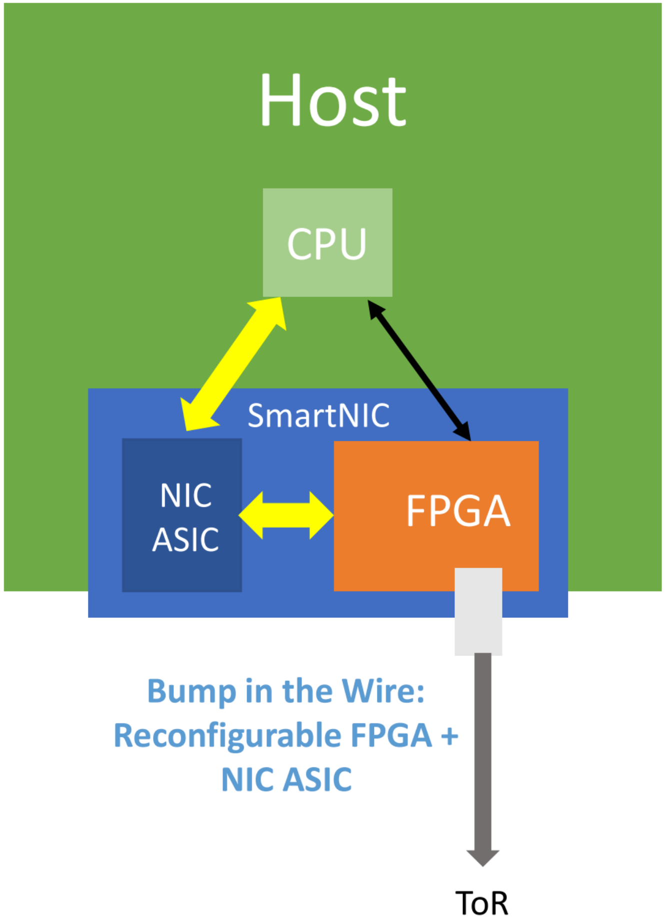


Microsoft

Our Solution:
Azure SmartNIC (FPGA)

- HW is needed for scale, perf, and COGS at 40G+
- 12-18 month ASIC cycle + time to roll new HW is too slow
- To compete and react to new needs, we need agility – SDN
- Programmed using Generic Flow Tables
  - Language for programming SDN to hardware
  - Uses connections and structured actions as primitives

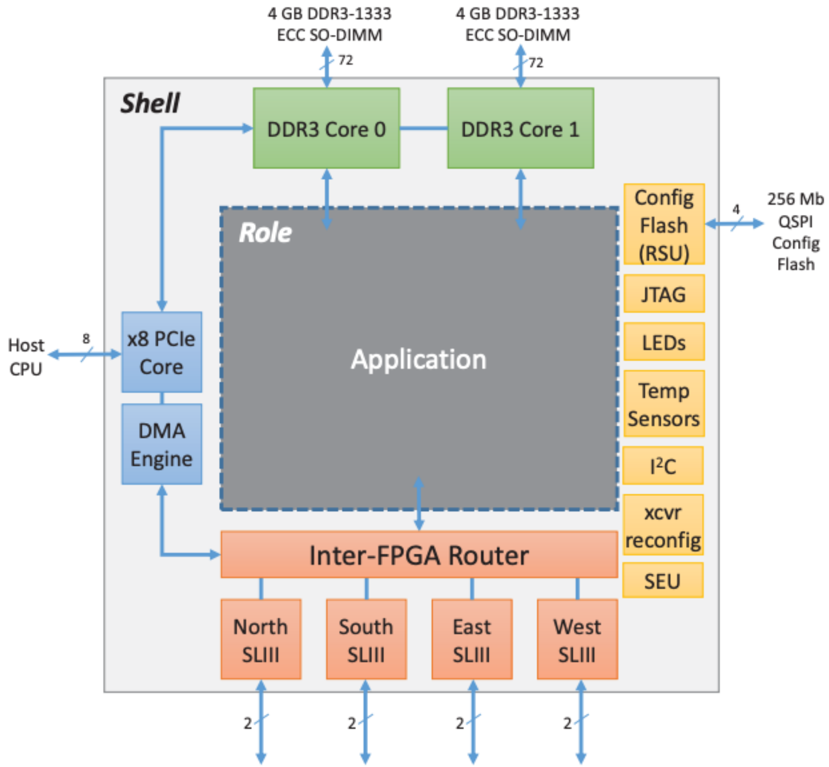
Microsoft



Bump in the Wire: Reconfigurable FPGA + NIC ASIC

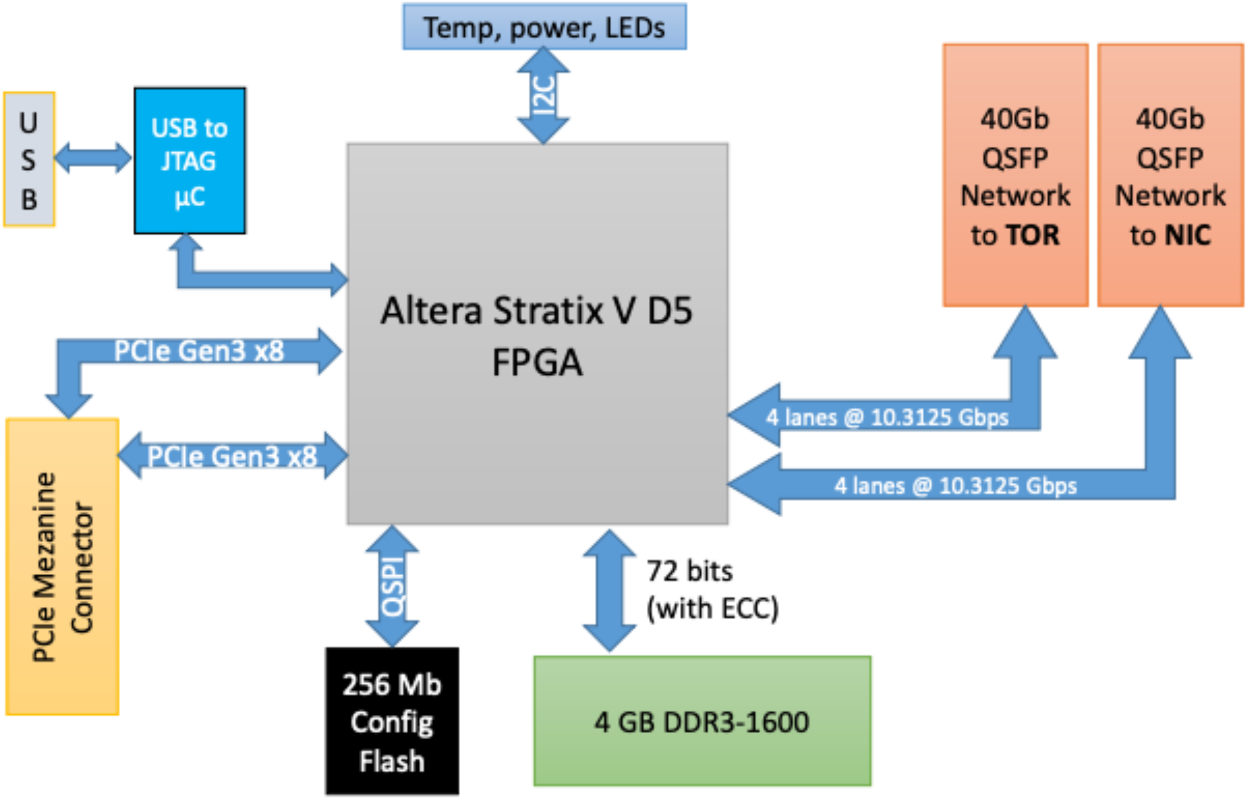
Detour: Project Catapult

Original design: FPGA was not in the NIC's path.



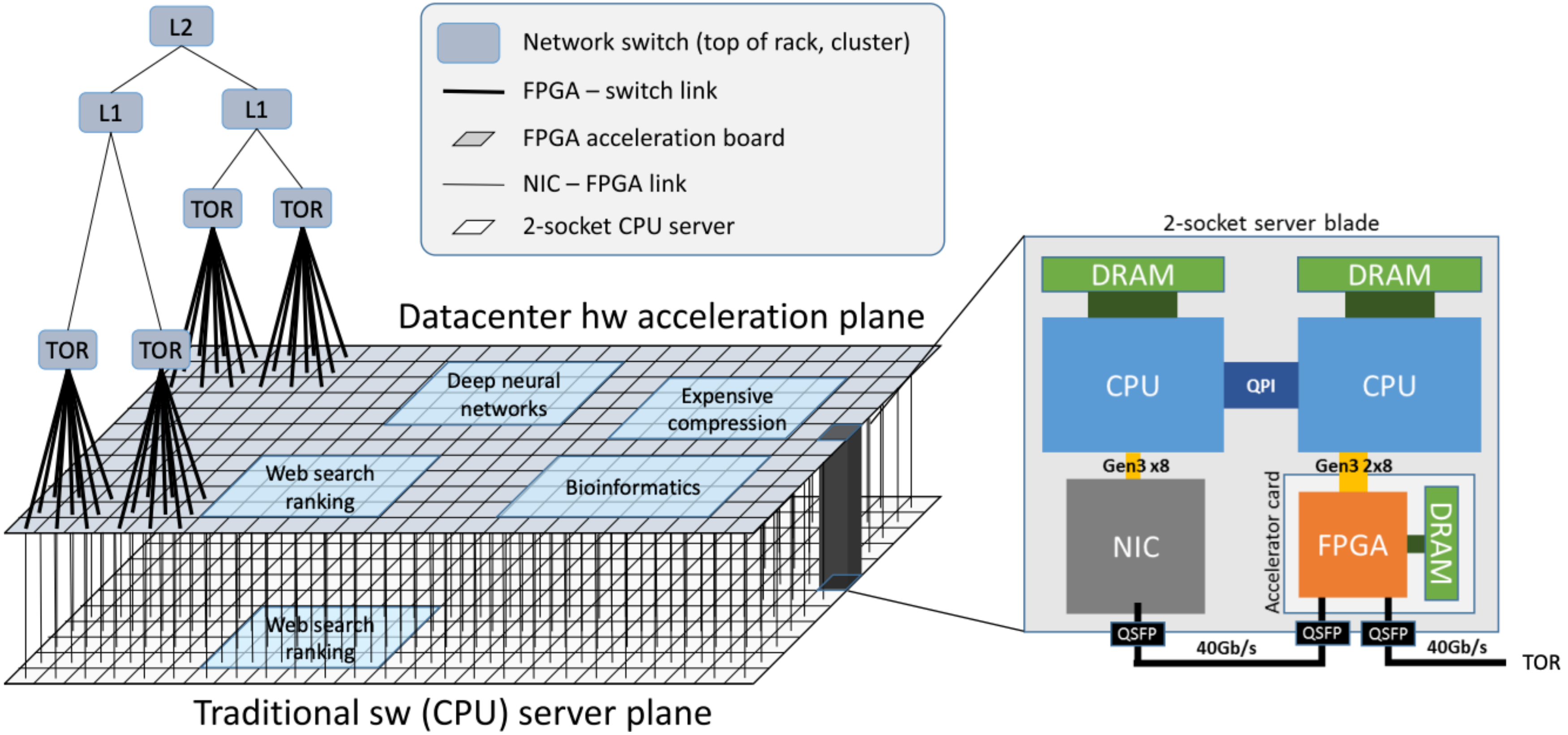
A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services, ISCA 2014

Detour: Project Catapult



A Cloud-Scale Acceleration Architecture, Micro 2016

Detour: Project Catapult



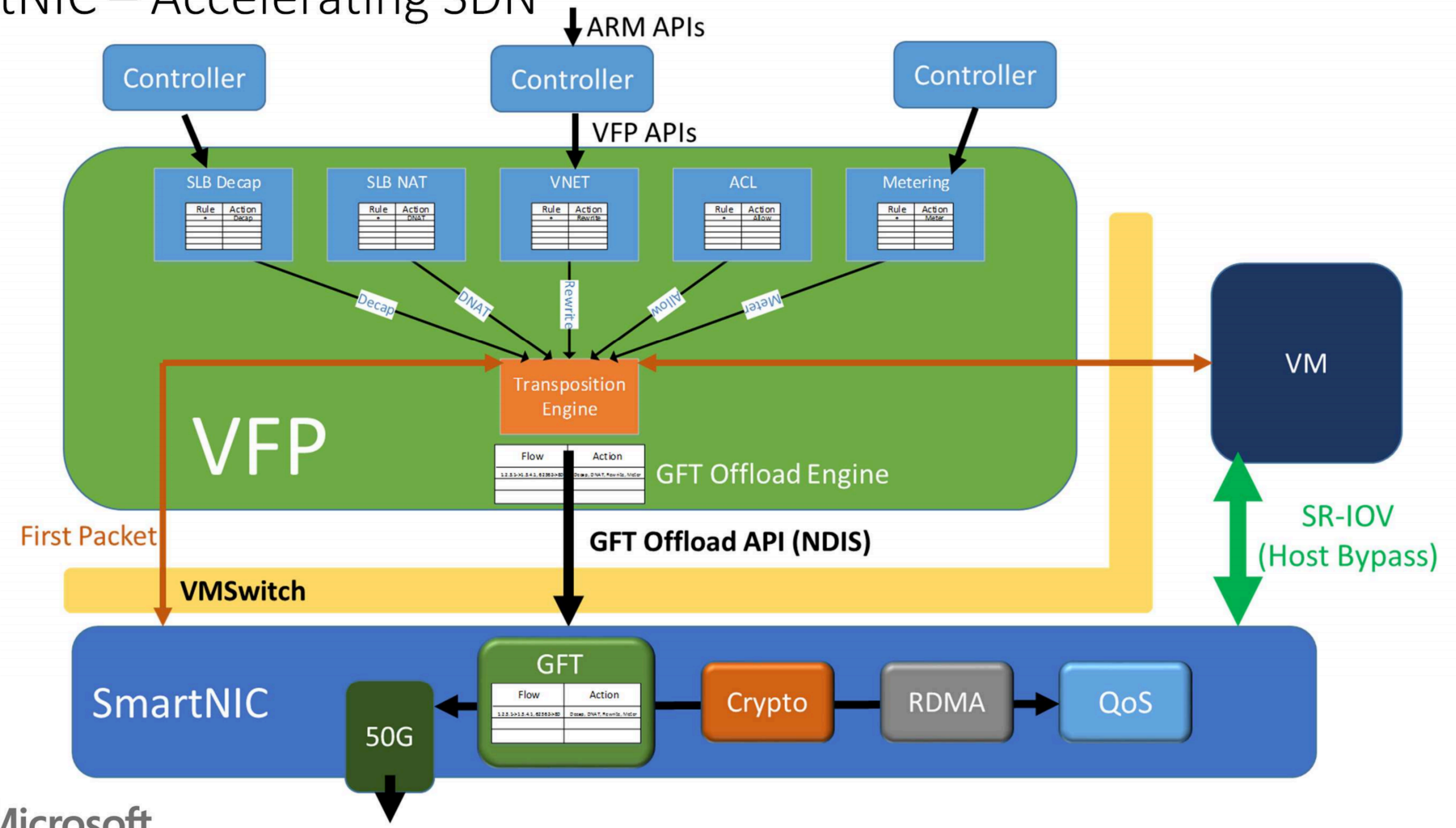
Lightweight Transport Layer for direct inter-FPGA communication.

A Cloud-Scale Acceleration Architecture, Micro 2016

FPGAs: Internal Q&A

1. Aren’t FPGAs much bigger than ASICs?
2. Aren’t FPGAs very expensive?
3. Aren’t FPGAs hard to program?
4. Isn’t my code locked in to a single FPGA vendor?
5. Can FPGAs be deployed at hyperscale? Are they DC-ready?

SmartNIC – Accelerating SDN



Microsoft

Overview

- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

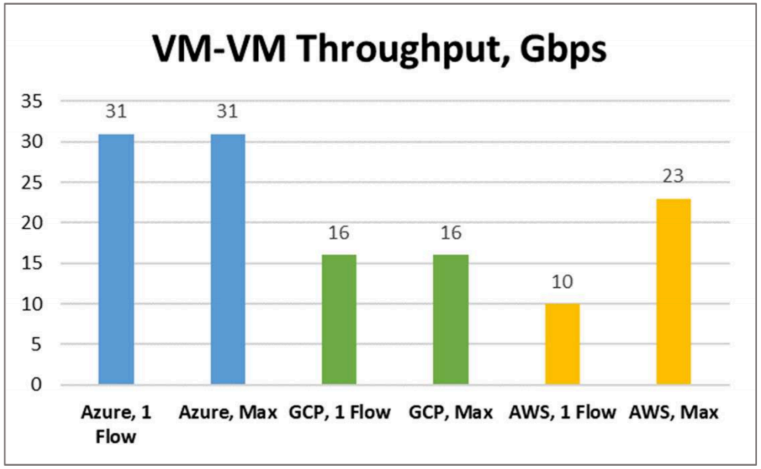
Microsoft

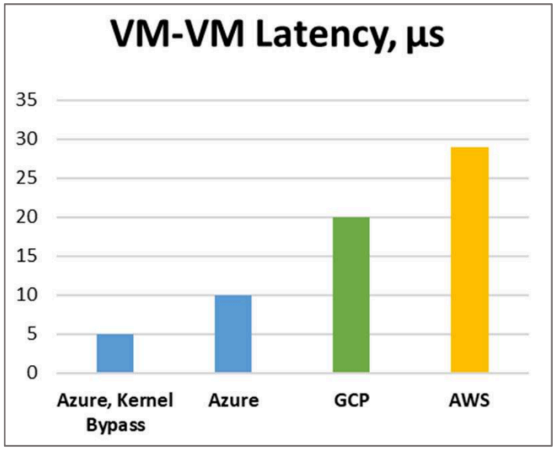
Azure Accelerated Networking

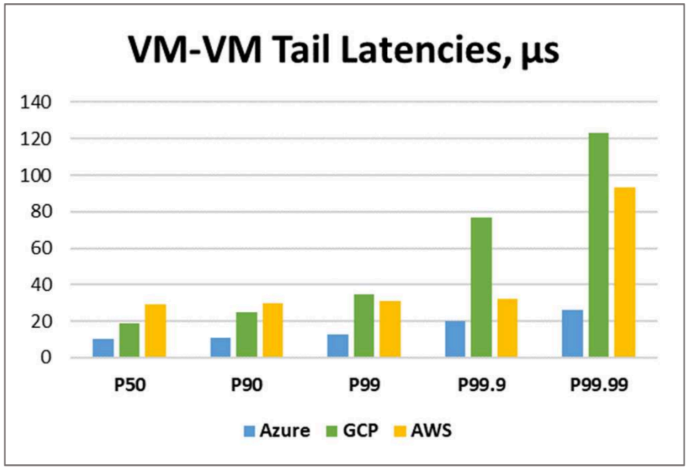
- Highest bandwidth VMs of any cloud so far…
  - Standard compute VMs get up to 32Gbps
  - Stock Linux VM with CUBIC gets 30+Gbps on a single connection
- Consistent low latency network performance
  - Provides SR-IOV to the VM
  - 5x+ latency improvement – sub 15us within tenants
  - Increased packets per second – Up to 25M PPS (12M forwarding) for DPDK VMs
  - Reduced jitter means more consistency in workloads
- Enables workloads requiring native performance to run in cloud VMs
  - >2x improvement for many DB and OLTP applications

Microsoft

AccelNet Comparative Results

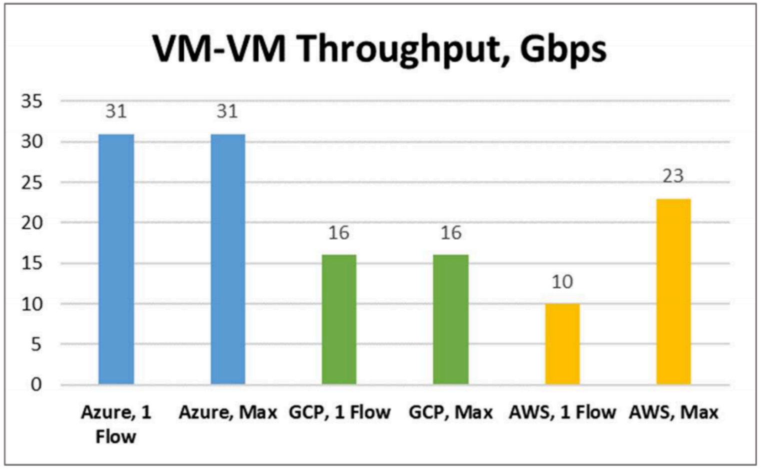


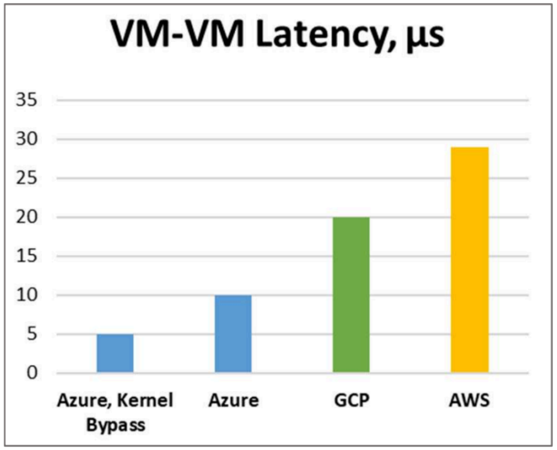


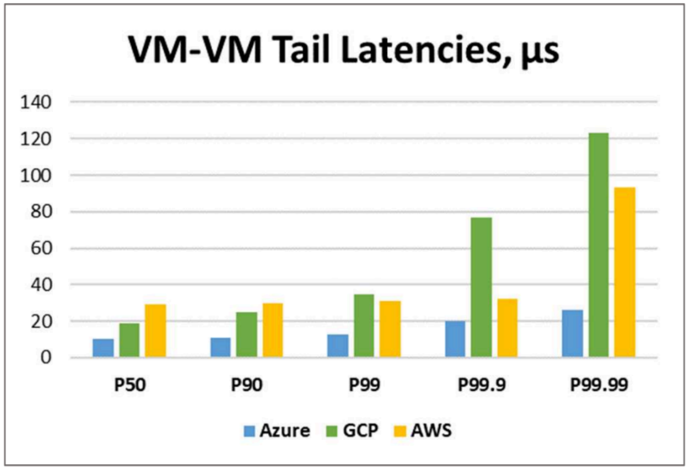


Microsoft

AccelNet Comparative Results







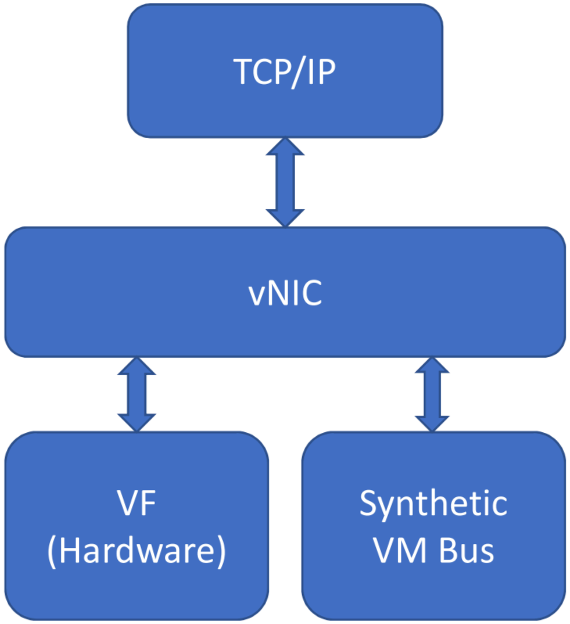
Microsoft

Overview

- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

Serviceability is Key

- All parts of this system can be updated, any of which require us to take out the hardware path – or VM can be live migrated
  - FPGA image, driver, GFT layer, Vswitch/VFP, NIC PF driver



Changes, Changes, Changes

A few examples of many…

- TCP and protocol state machines
- Complex packet forwarding and duplication actions
- New SDN actions
- Accelerating the offload path
- Line rate diagnostics and monitoring

Changes, Changes, Changes

A few examples of many…

- TCP and protocol state machines
- Complex packet forwarding and duplication actions
- New SDN actions
- Accelerating the offload path
- Line rate diagnostics and monitoring

Lessons Learned

- Design for serviceability upfront
- Use a unified development team
- Use software development techniques for FPGAs
- Better perf means better reliability
- HW/SW co-design is best when iterative
- Failure rates remained low – FPGAs in the DC were reasonably reliable
- Upper layers should be agnostic of offloads
- Mitigating Spectre performance impact

Your Opinions

- What did you like about the paper?
- What are its limitations?

Your Opinions

Software or Hardware?

Upcoming tasks

- Warm-up assignment 3 will be released today. Due on 11/16.
- Second progress report due on Monday, 11/07
  - Please identify the delta over your first report.
  - e.g. tagging a paragraph/section as [new]/[updated]
  - or, a separate paragraph at the beginning/end that lists changes over last report.
- List the paper you will present by 11/11
  - "TBDs" not allowed!