Azure Accelerated Networking: SmartNICs in the Public Cloud

Daniel Firestone, Andrew Putnam, Sambhrama Mundkur, Derek Chiou, Alireza Dabagh, Mike Andrewartha, Hari Angepat, Vivek Bhanu, Adrian Caulfield, Eric Chung, Harish Kumar Chandrappa, Somesh Chaturmohta, Matt Humphrey, Jack Lavier, Norman Lam, Fengfen Liu, Kalin Ovtcharov, Jitu Padhye, Gautham Popuri, Shachar Raindel, Tejas Sapre, Mark Shaw, Gabriel Silva, Madhan Sivakumar, Nisheeth Srivastava, Anshuman Verma, Qasim Zuhair, Deepak Bansal, Doug Burger, Kushagra Vaid, David A. Maltz, and Albert Greenberg

Daniel Firestone

Tech Lead and Group Manager, Azure Host Networking Team

Overview

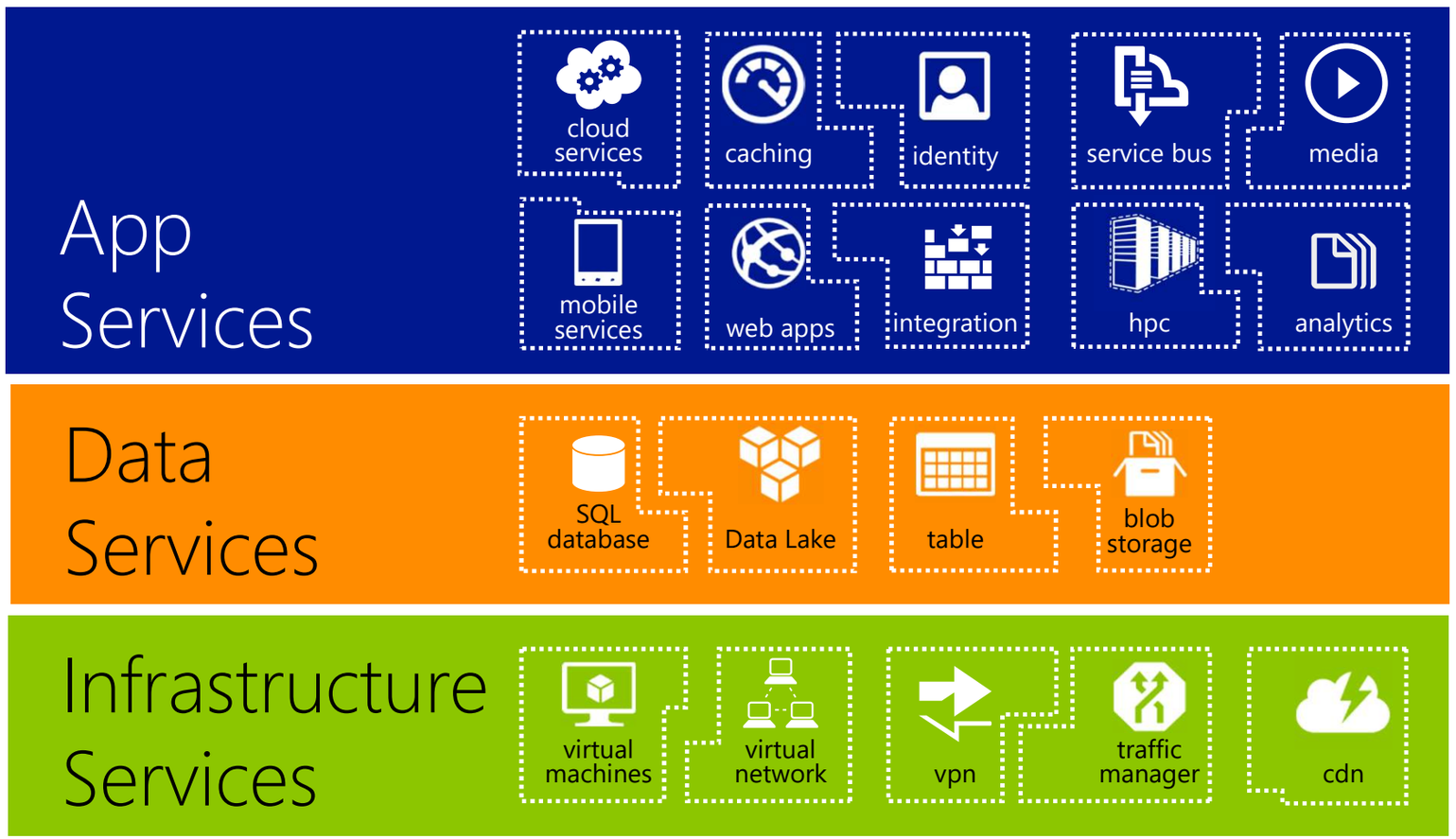
- Azure and Scale
- Recap: Virtual Filtering Platform and Host SDN
- Why Accelerated Networking? Scaling up SDN
- Hardware Choices
- Azure SmartNIC
- Accelerated Networking in Azure: Results
- Experiences and Lessons Learned
- Conclusion and Future

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

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

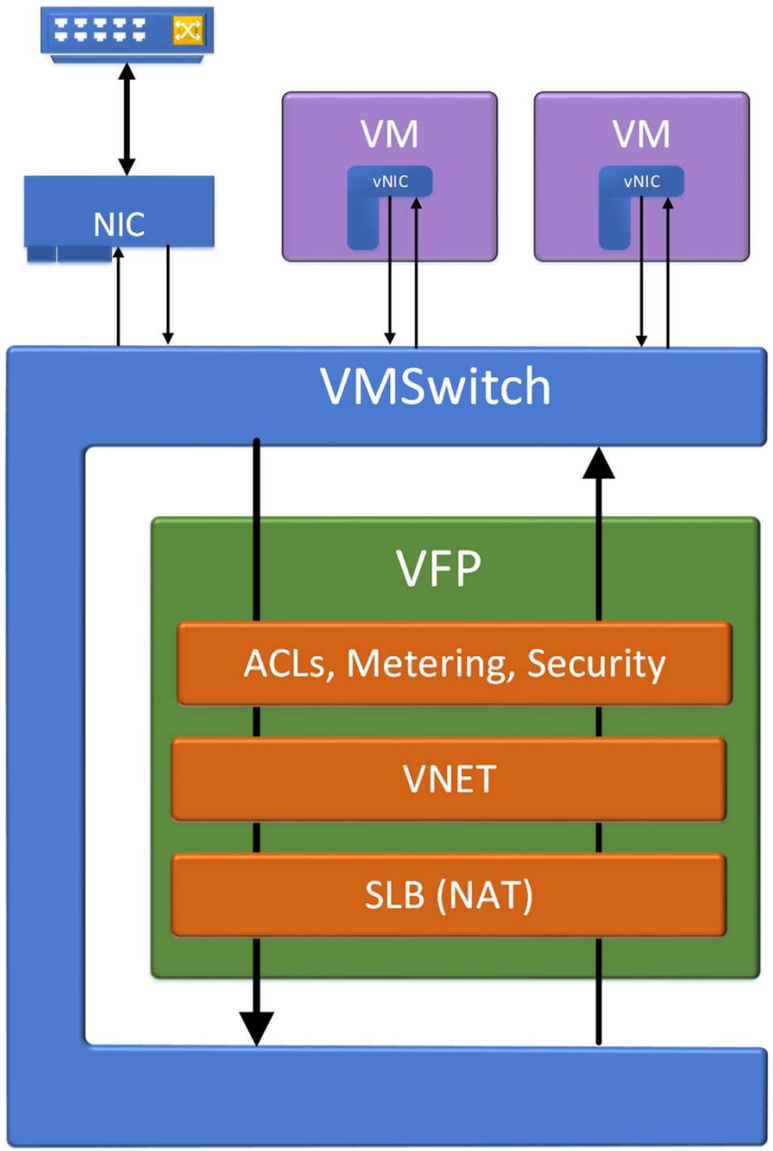
Microsoft

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

Microsoft



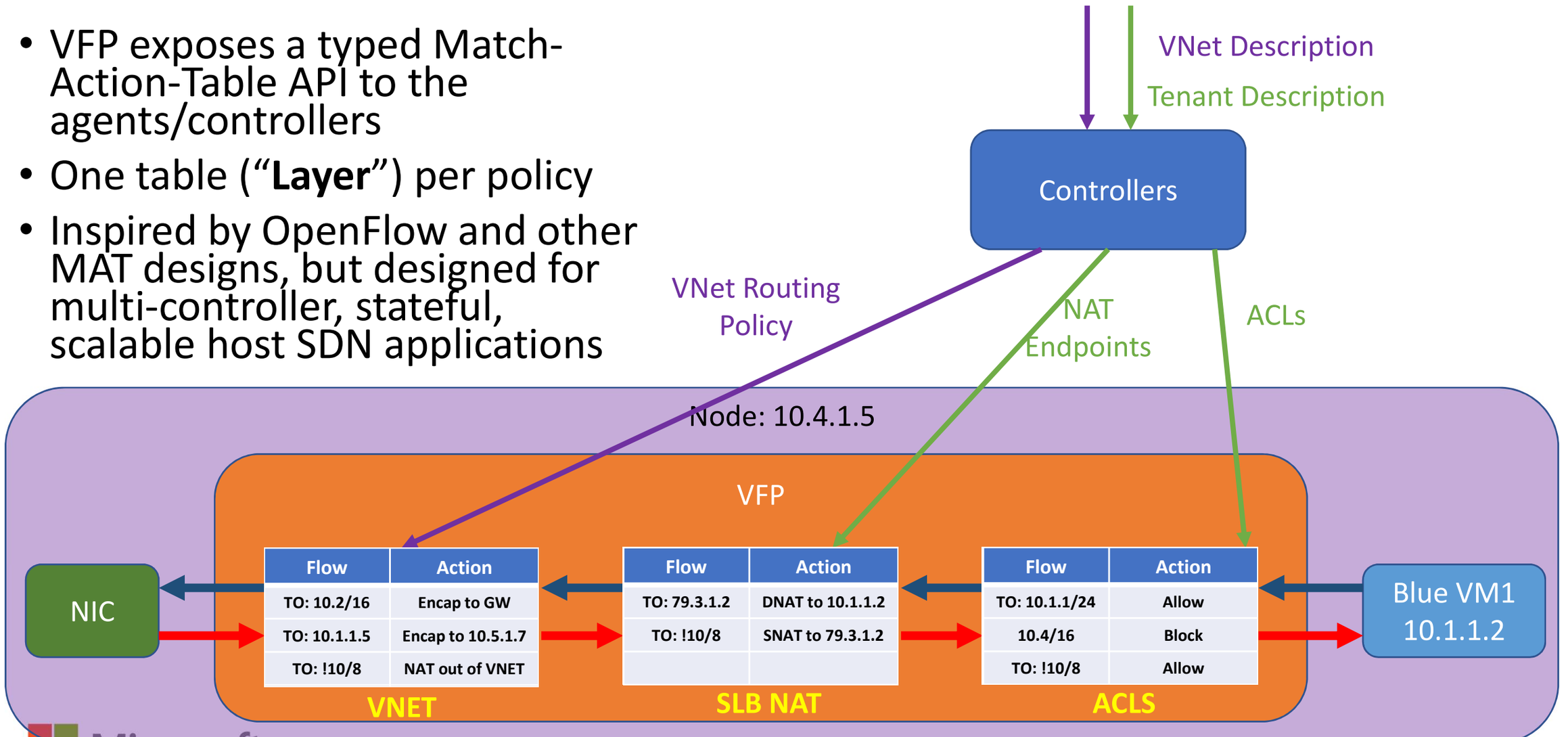
Virtual Filtering Platform (VFP) Azure's SDN Dataplane

- Virtual switch for Hyper-V / Azure
- Provides core SDN functionality for Azure networking services, including:
  - Address Virtualization for VNET
  - VIP -> DIP Translation for SLB
  - ACLs, Metering, and Security Guards
- Uses programmable rule/flow tables to perform per-packet actions
- Programmed by multiple Azure SDN controllers, supports all dataplane policy at line rate with offloads

Microsoft

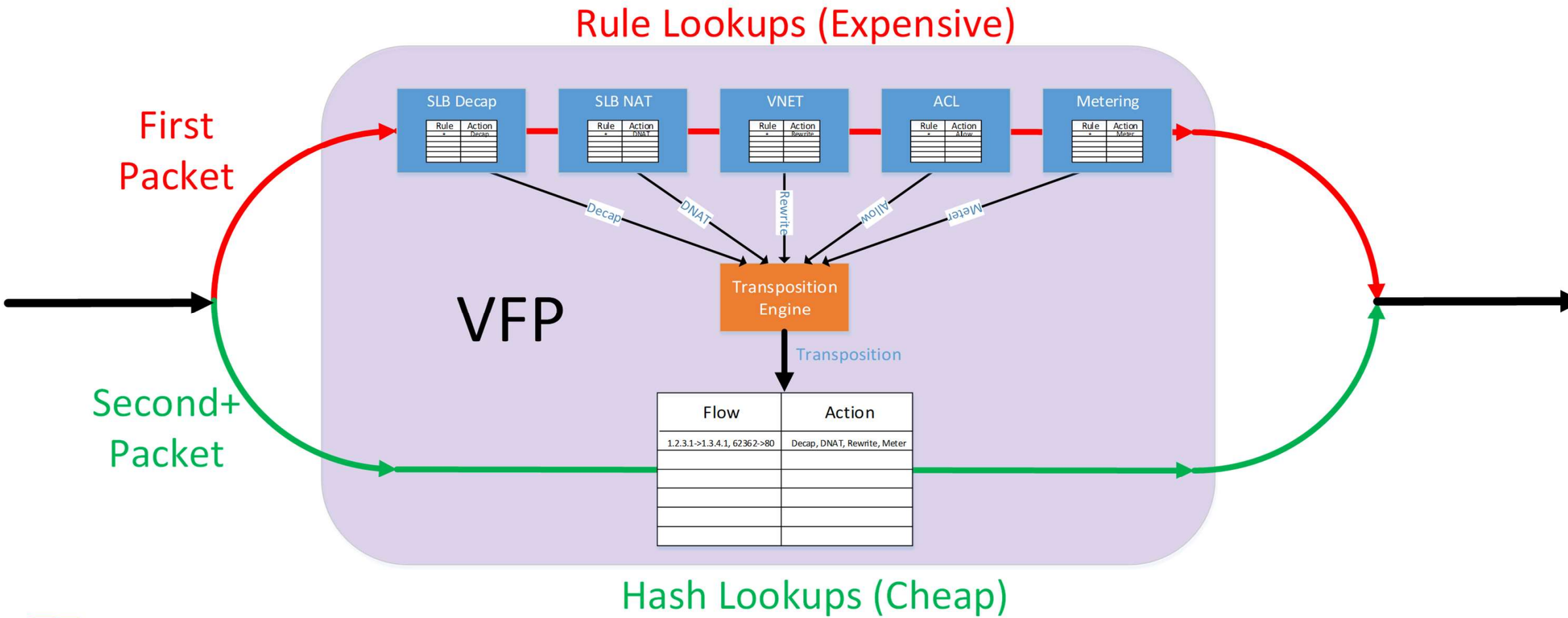
Key Primitive: Match Action Tables

- 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



Microsoft

Unified Flow Tables – A Fastpath Through VFP



Microsoft

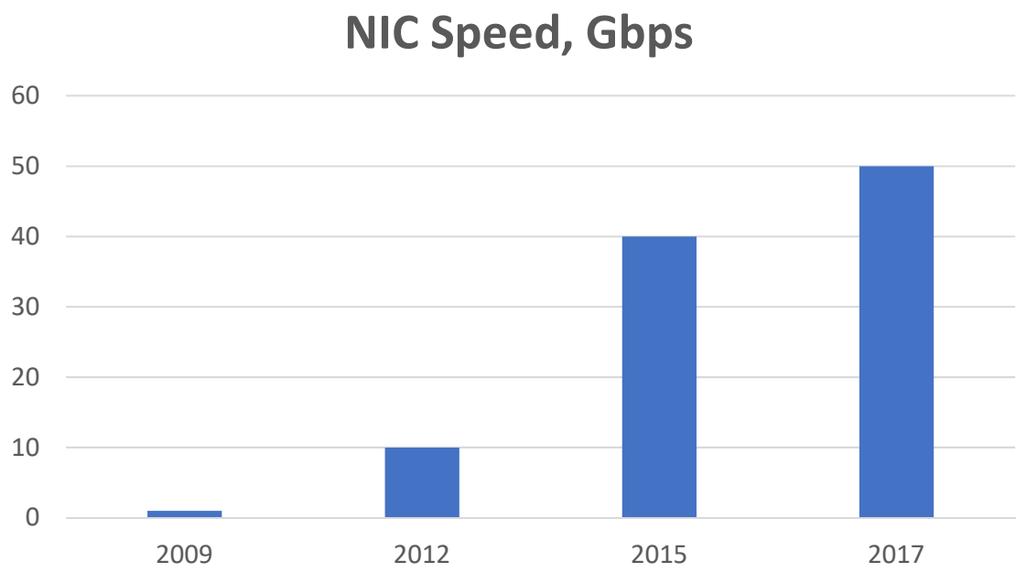
Overview

- 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

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!

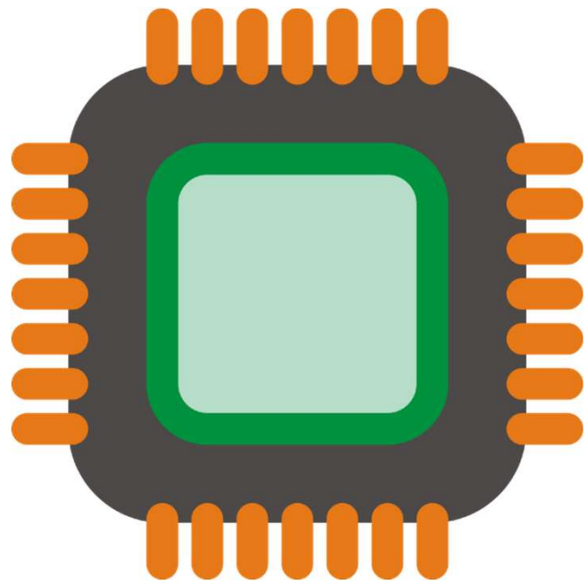
Microsoft

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

Microsoft

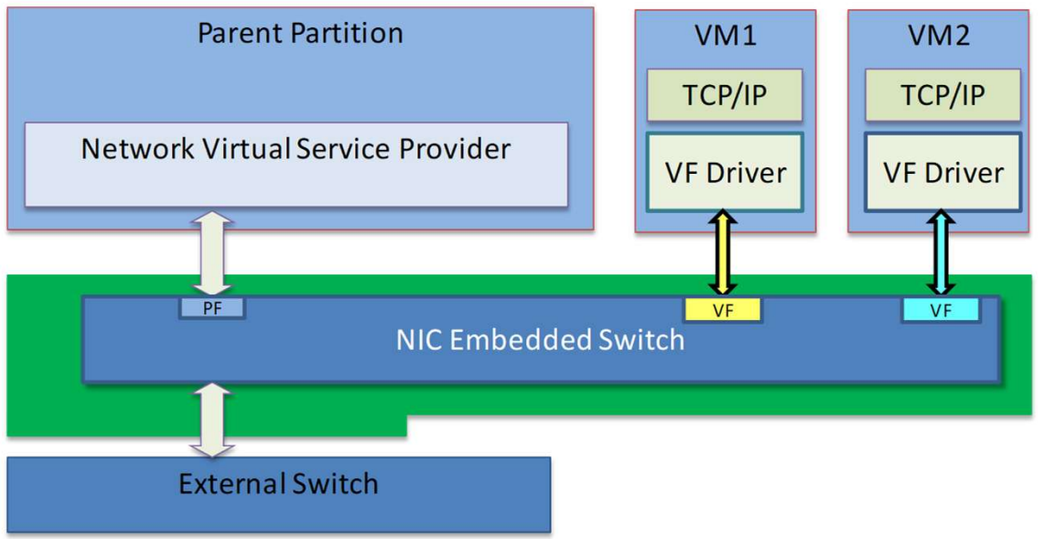
Traditional Approach to Scale: ASICs

- We've worked with network ASIC vendors over the years to accelerate many functions, including:
  - TCP offloads: Segmentation, checksum, …
  - Steering: VMQ, RSS, …
  - Encapsulation: NVGRE, VXLAN, …
  - Direct NIC Access: DPDK, PacketDirect, …
  - RDMA
- Is this a long term solution?



Microsoft

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



But where is the SDN Policy?

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

How do we get the performance of hardware with programmability of software?

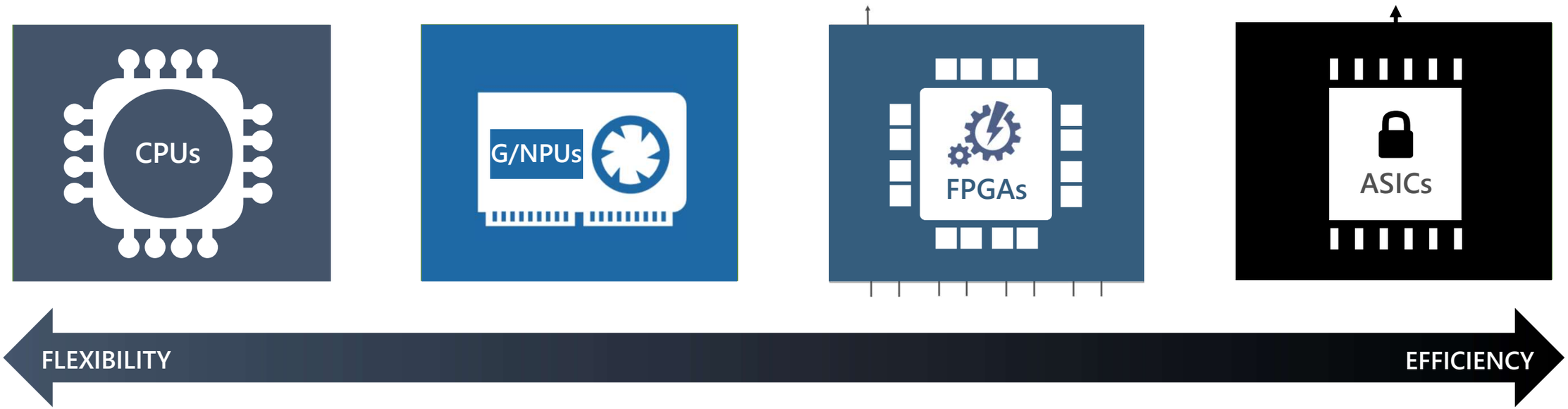
Microsoft

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

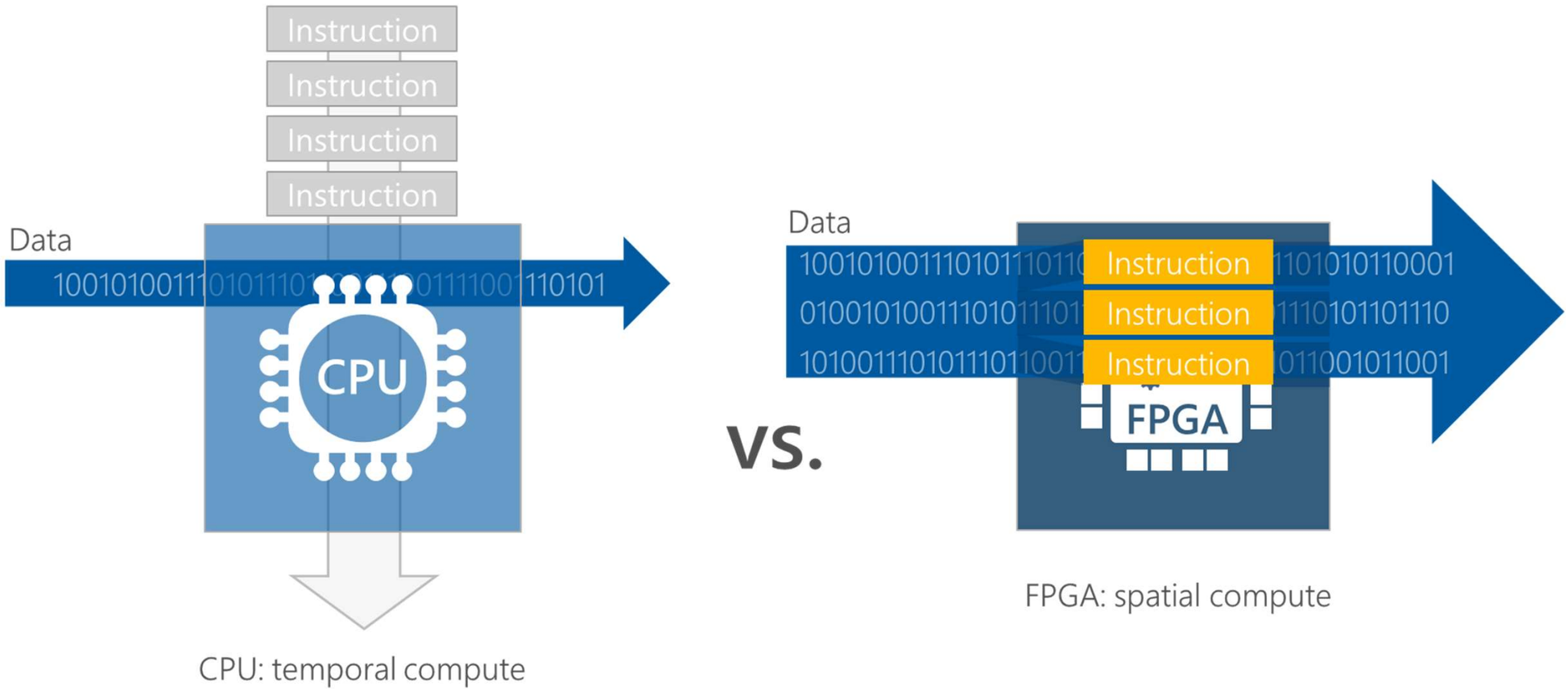
Silicon alternatives



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

Microsoft

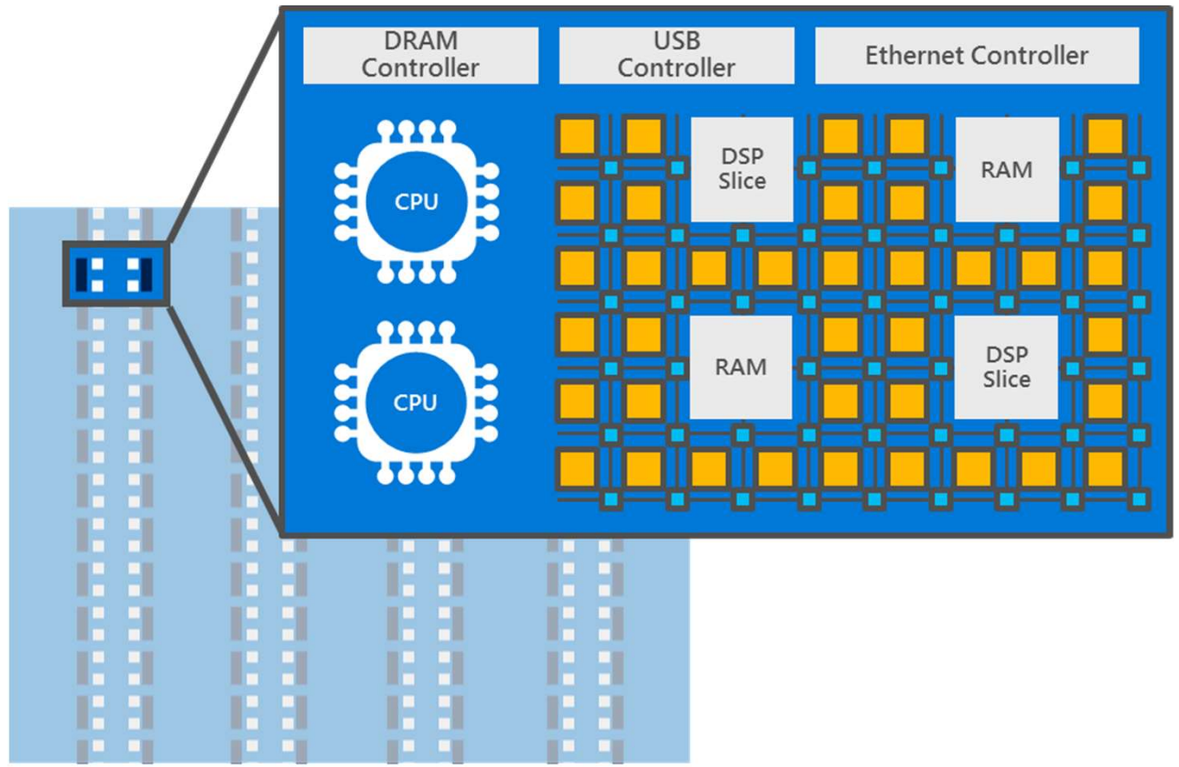
CPU vs. FPGA



Microsoft

What is an FPGA, Really?

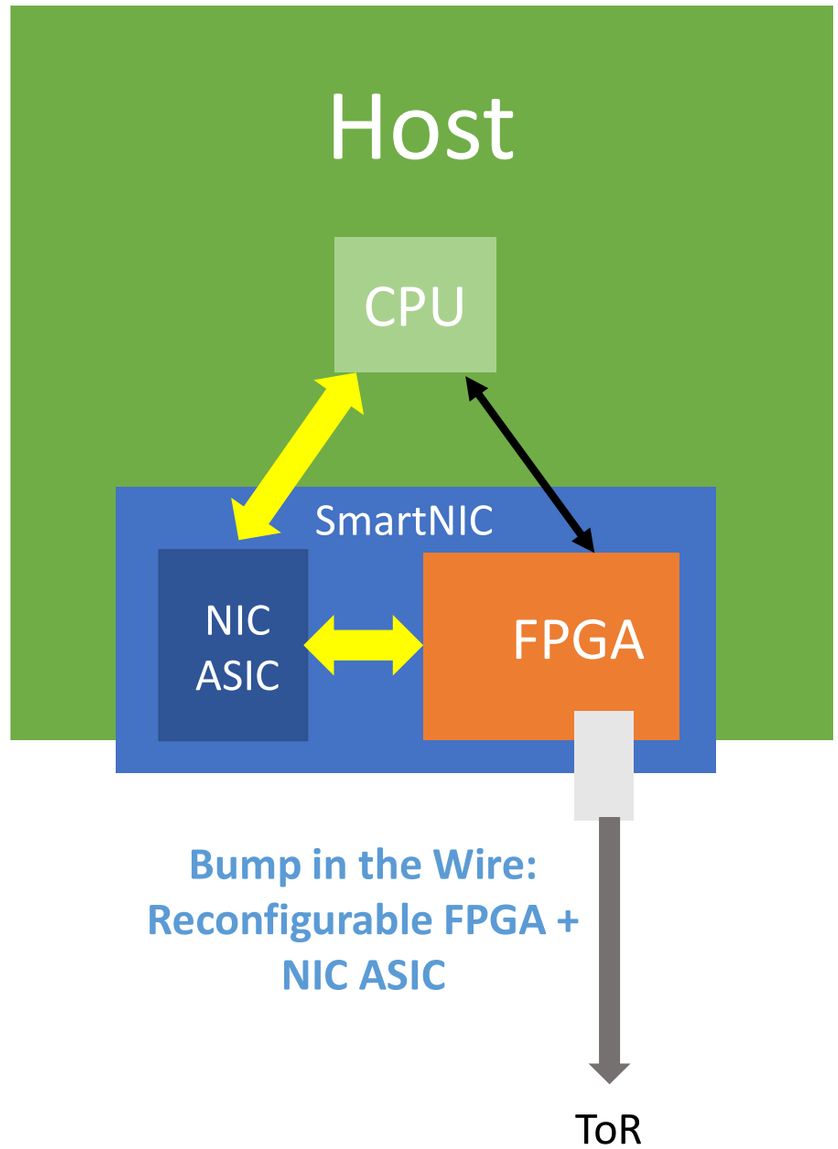
- Field Programmable Gate Array
- Chip has large quantities of programmable gates – highly parallel
- Program specialized circuits that communicate directly
- Two kinds of parallelism:
  - Thread-level parallelism (stamp out multiple pipelines)
  - Pipeline parallelism (create one long pipeline storing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



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



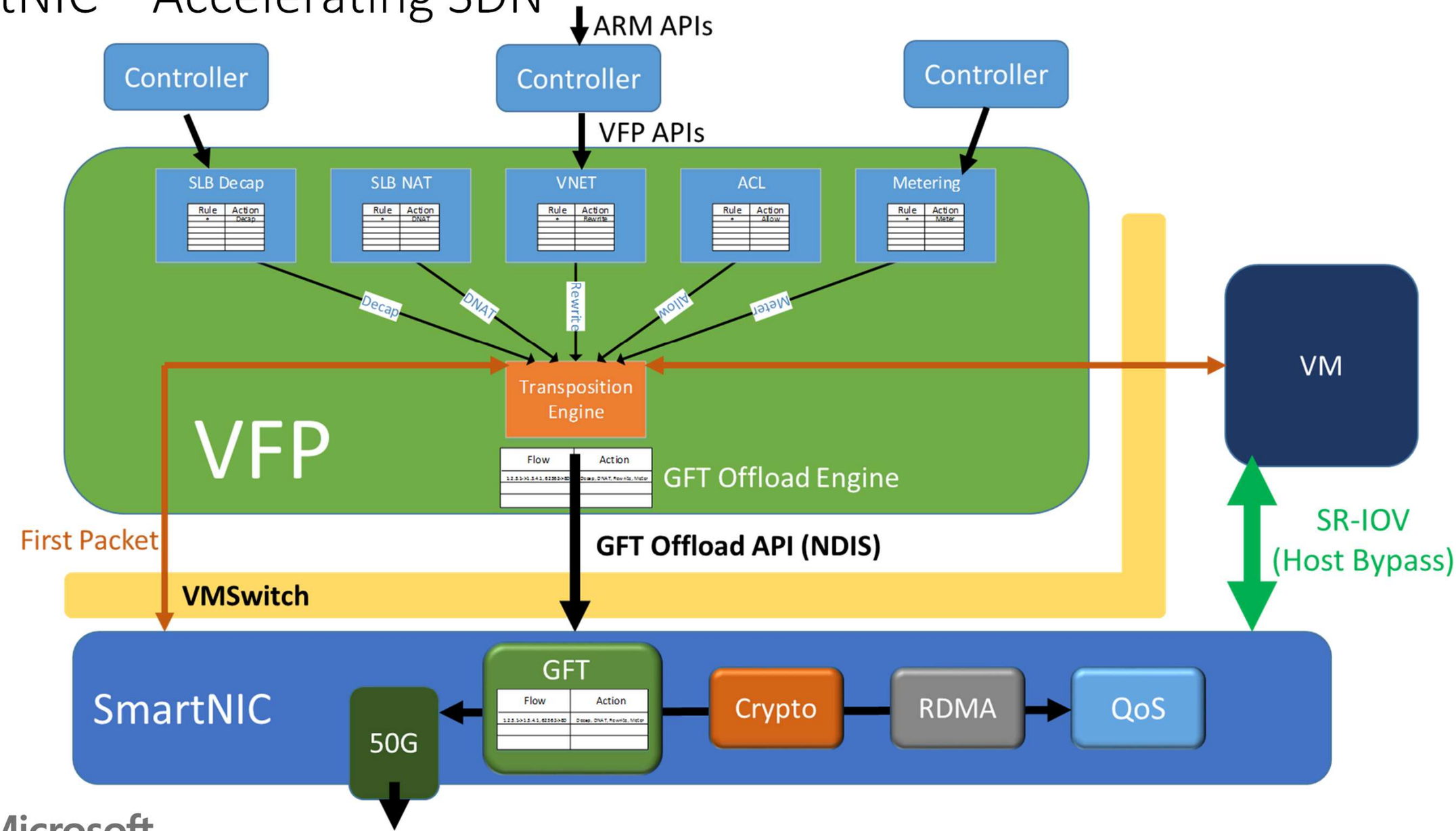
Microsoft

FPGAs: Rude 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?

Microsoft

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

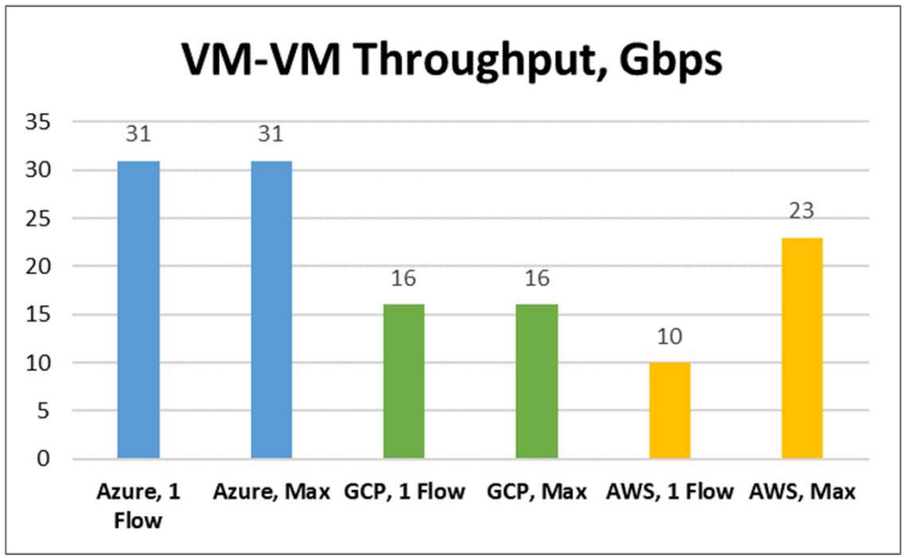
Microsoft

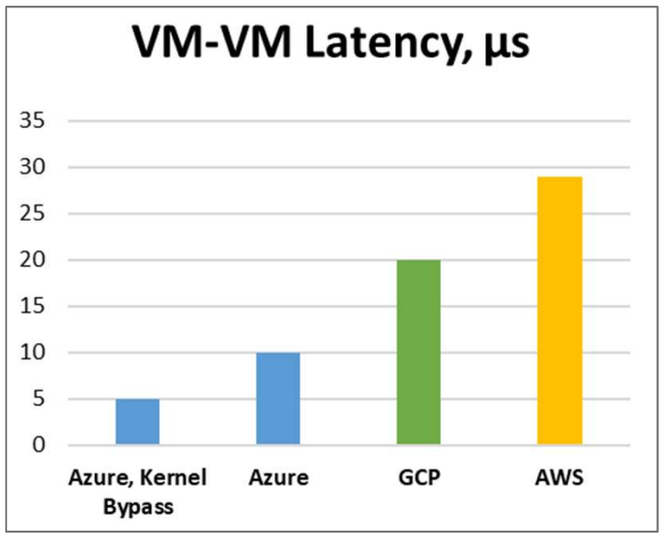
Azure Accelerated Networking

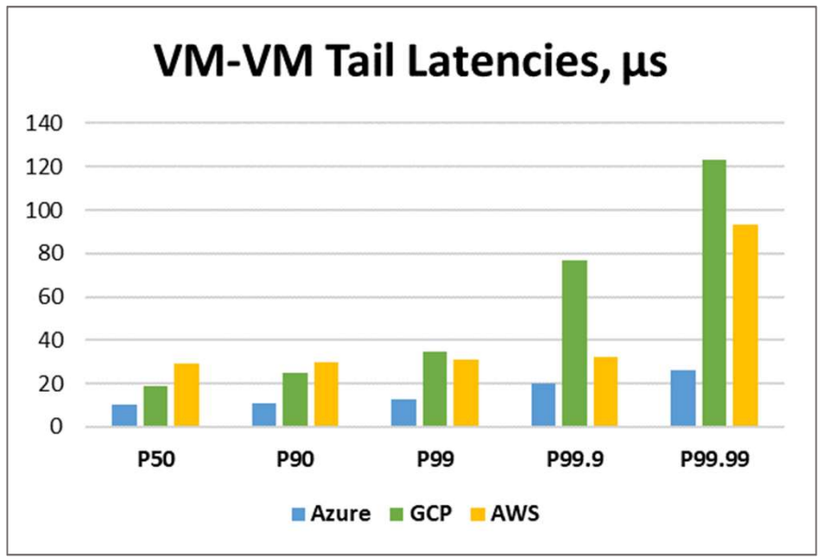
- Highest bandwidth VMs of any cloud so far…
  - Standard compute VMs get up to 32Gbps
  - Stock Linux VM with CUBIC gets 30+Gbps on a single connection
- Consistent low latency network performance
  - Provides SR-IOV to the VM
  - 5x+ latency improvement – sub 15us within tenants
  - Increased packets 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

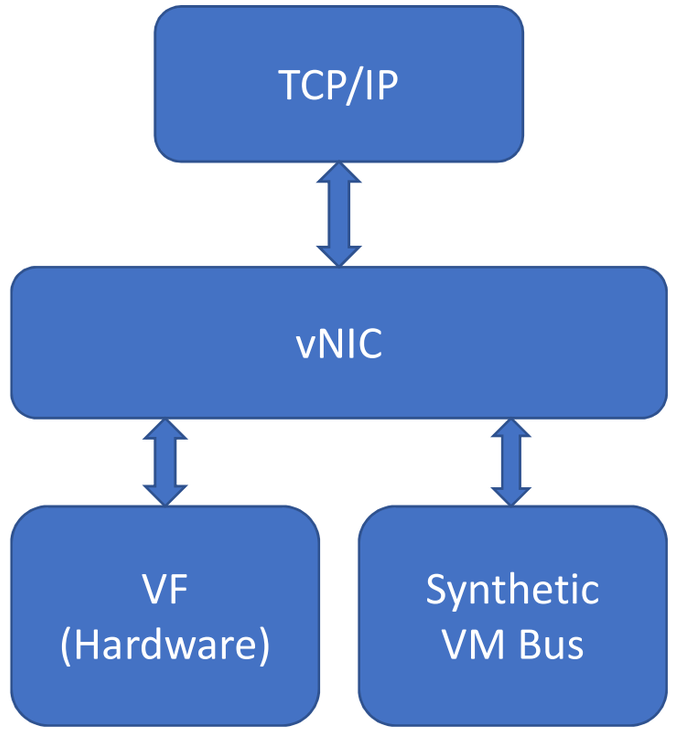
Overview

- 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

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
- IaaS requires high uptime and low disruption – can't take away the NIC device from under the app, and can't reboot the VM / app
- Instead, we keep the synthetic vNIC and support transparent failover between the vNIC and VF
  - Added support to Windows, Linux, DPDK (all upstream)



Lesson: A huge amount of the effort to deploy AccelNet was in making all parts of this path rebootlessly serviceable without impact

Timeline

- 2009: Azure!
- 2011-2012: VFP
- 2013-2014: VFPv2
- 2H’2015: Begin deploying SmartNICs
- Summer 2016: AccelNet Preview
- Summer 2017: AccelNet Global GA
- Late 2017-Early 2018: DPDK

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

Microsoft

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

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

Future Work:

Everything here is a hardware acceleration of a software SDN function… what new functions can we build with programmable hardware as a base primitive?

Conclusion:
Software or Hardware?

Each has its place…

Microsoft

Want to come build the next generation of scalable cloud networks? We're hiring!

fstone@microsoft.com