NSDI
April 10, 2018

Andromeda

Performance, Isolation, and Velocity at Scale in Cloud Network Virtualization

Google Cloud

Andromeda Goals

Performance and Isolation

High throughput and low latency, regardless of the actions of other tenants

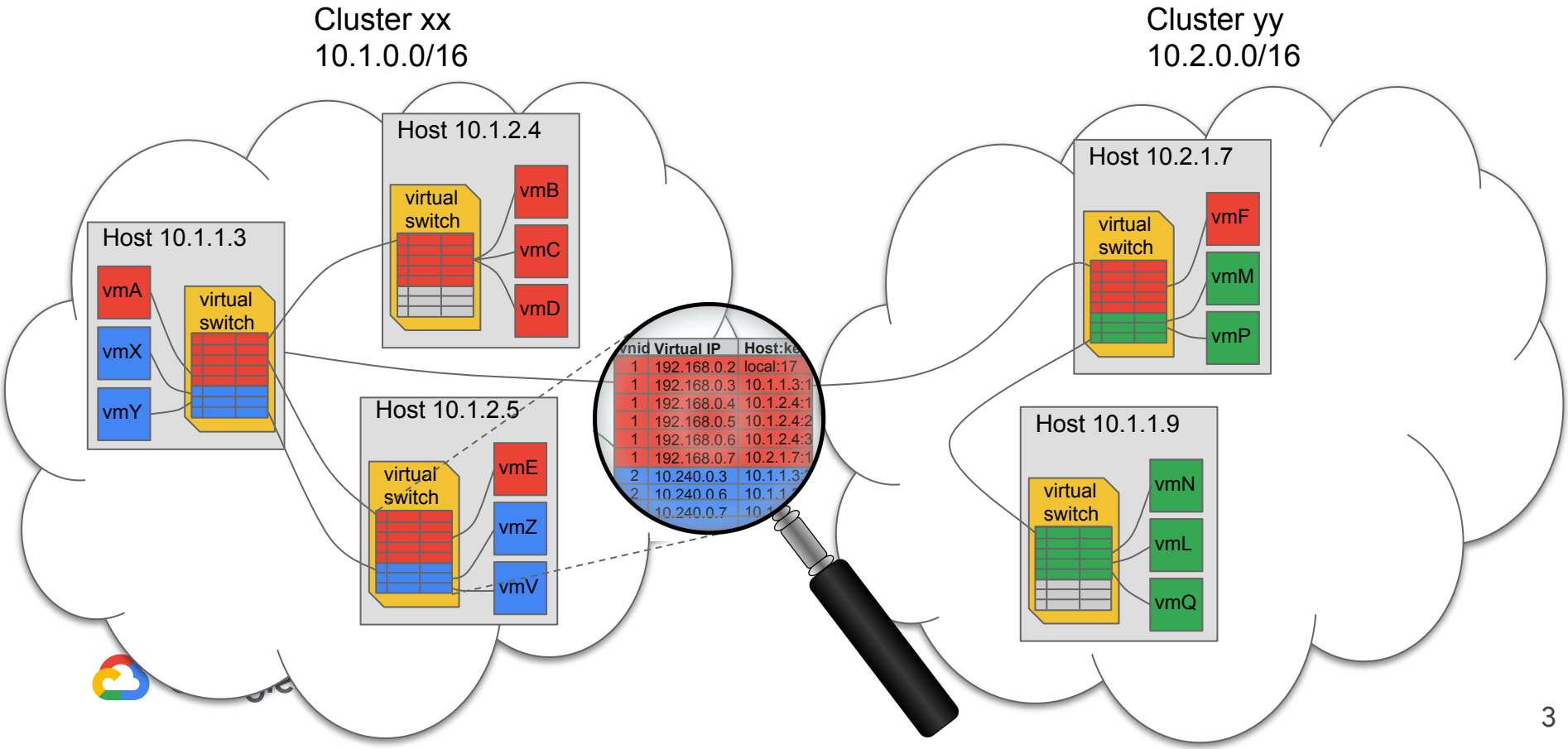
Velocity

Quickly develop and deploy new features and performance improvements

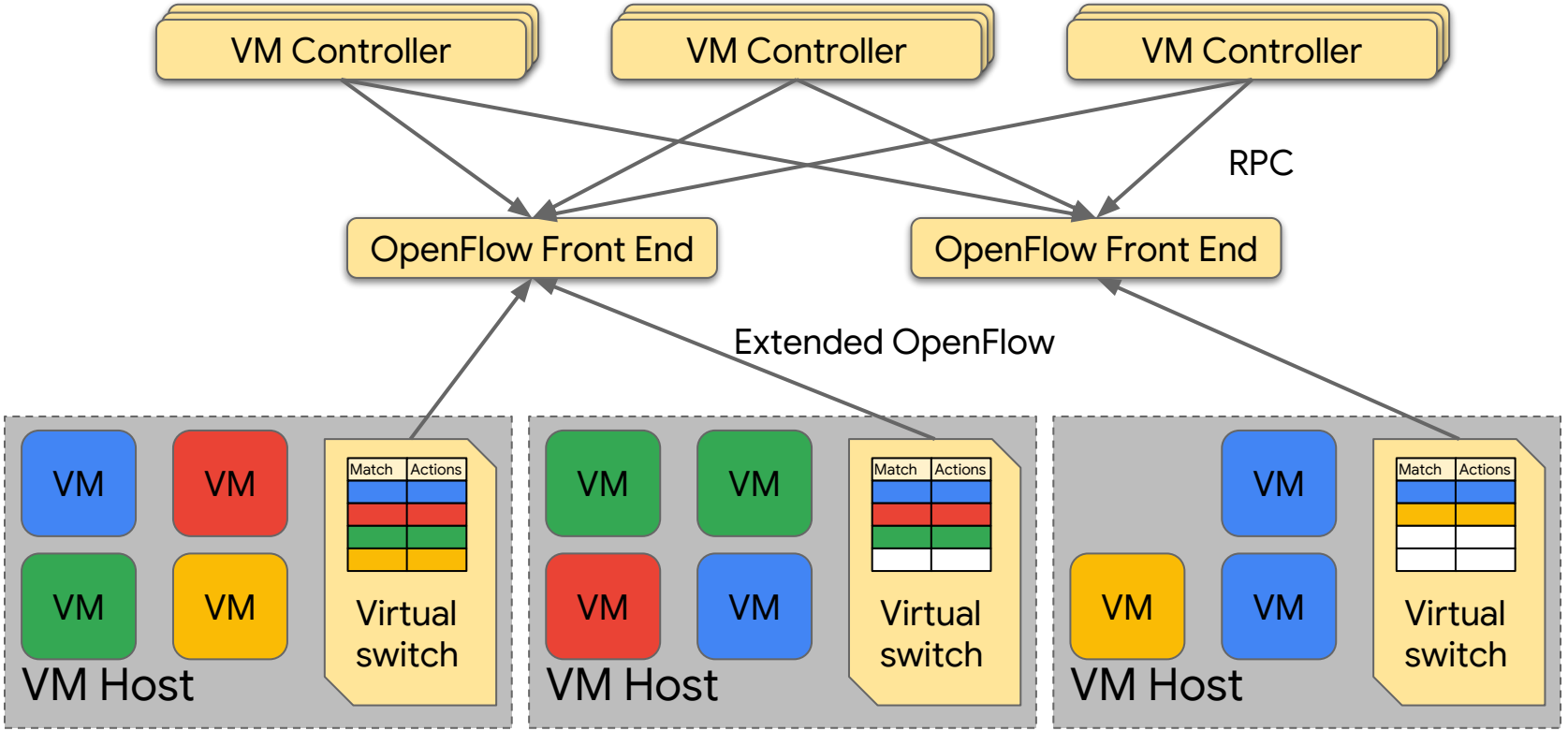
Scalability

Large networks, many tenants, rapid provisioning

Network Virtualization

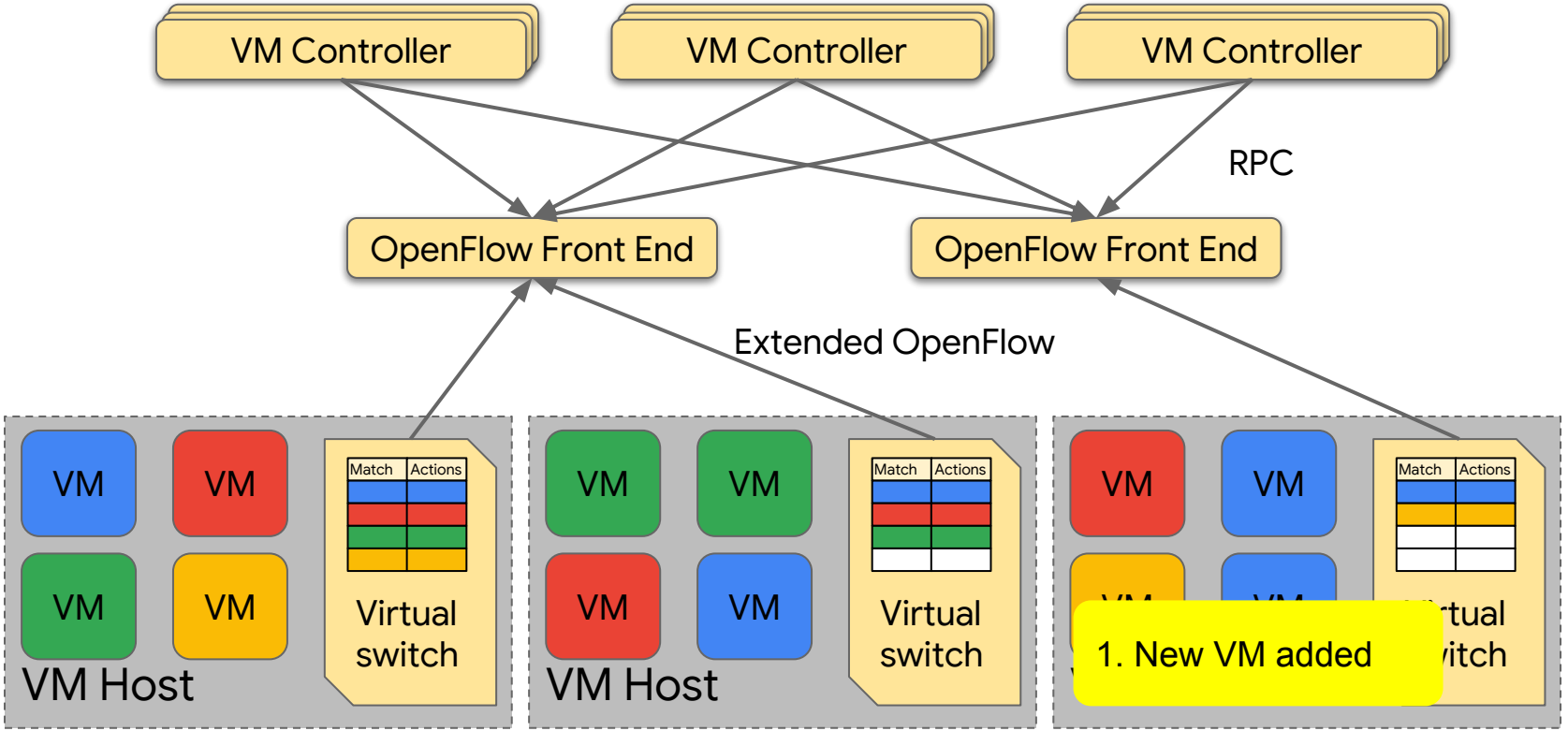


Andromeda Architecture



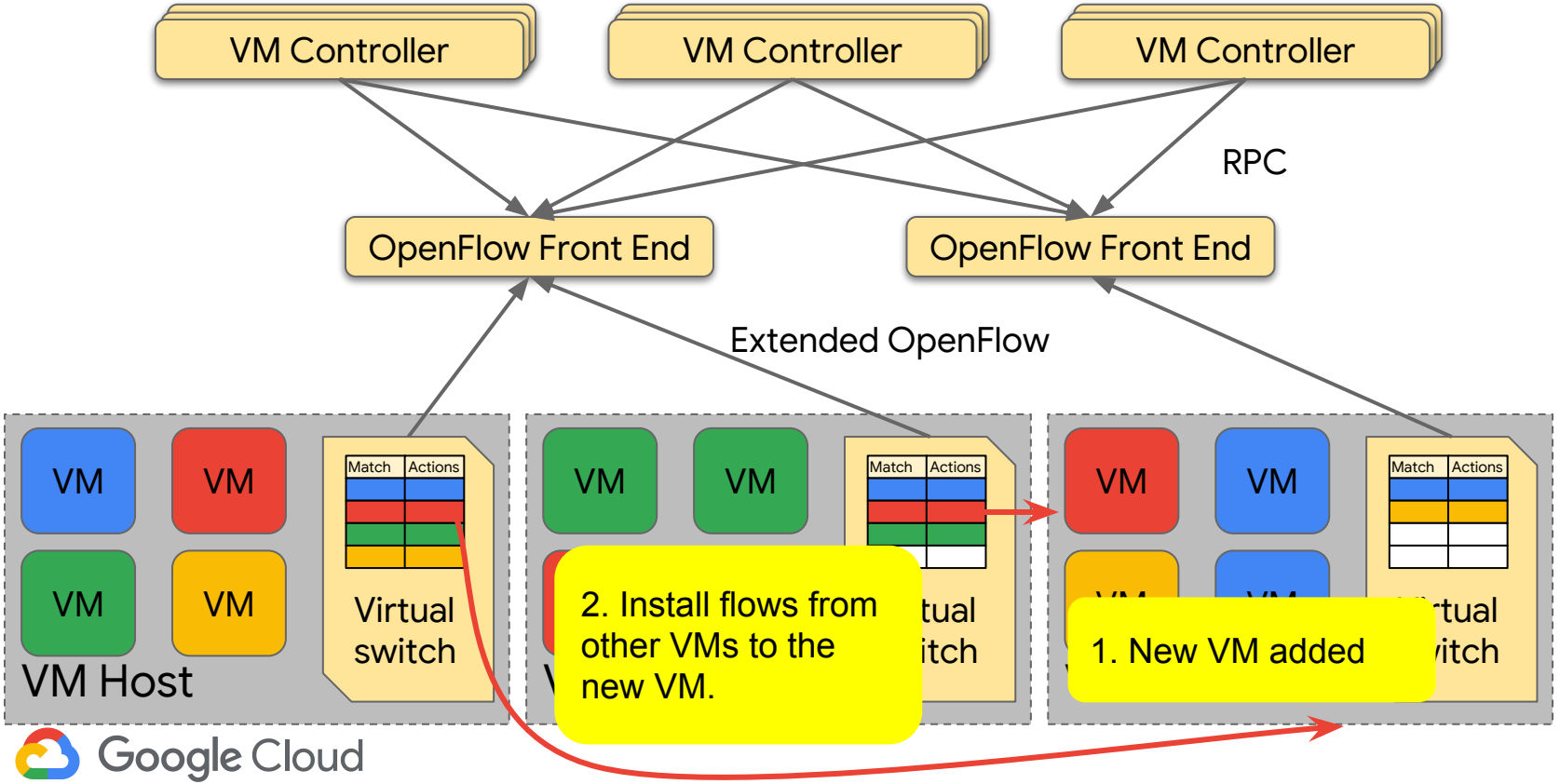
Google Cloud

Andromeda Architecture



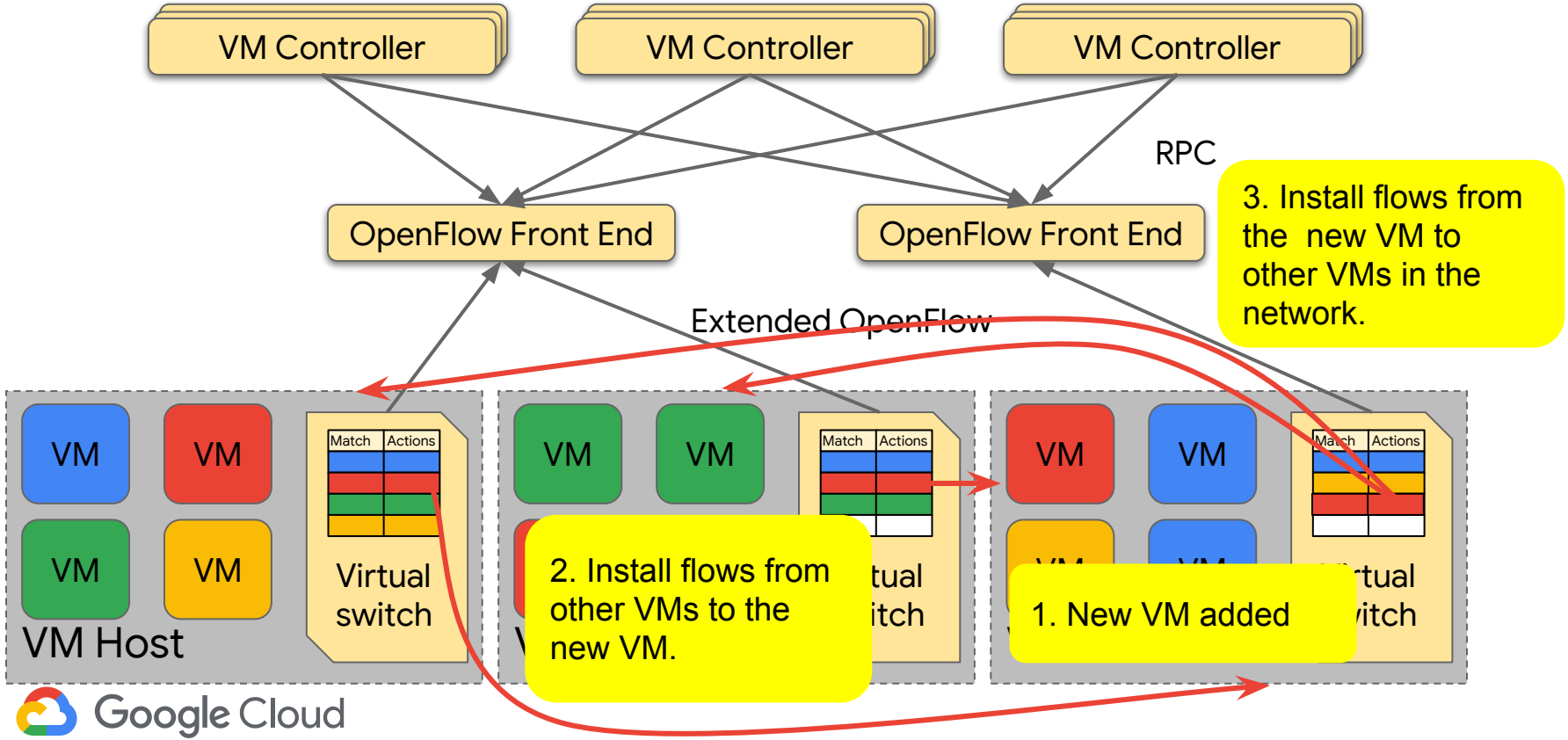
Google Cloud

Andromeda Architecture



Google Cloud

Andromeda Architecture



Google Cloud

Scaling Goals

Global connectivity

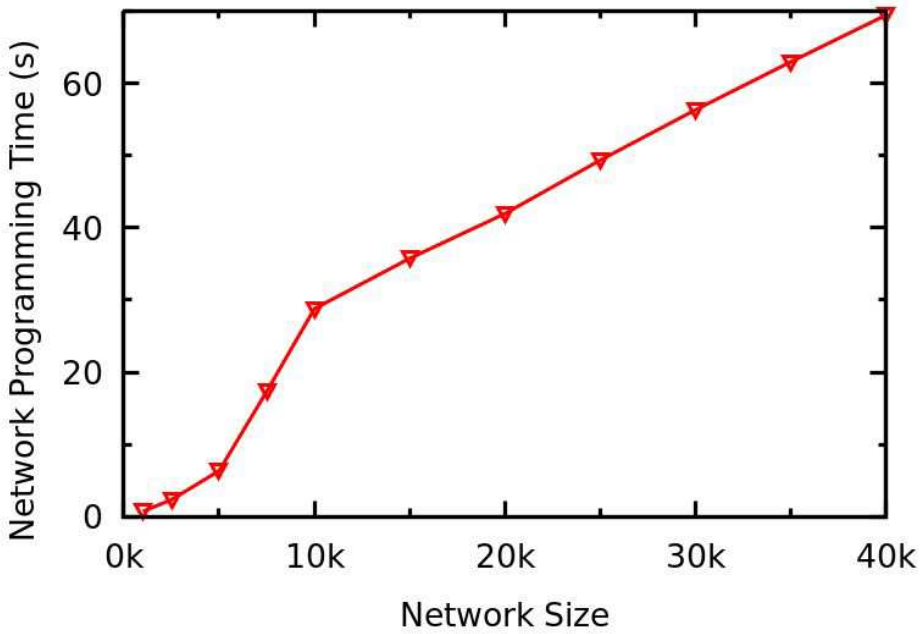
Large virtual networks (100k+ VMs)

Rapid provisioning

Enable on-demand workloads

Google Cloud

Programming Time for Large Networks



Setup:

- VMs are placed on 10,000 hosts
- 30 VM Controller partitions

Programming time is $O(n \times H)$

n = number of VMs

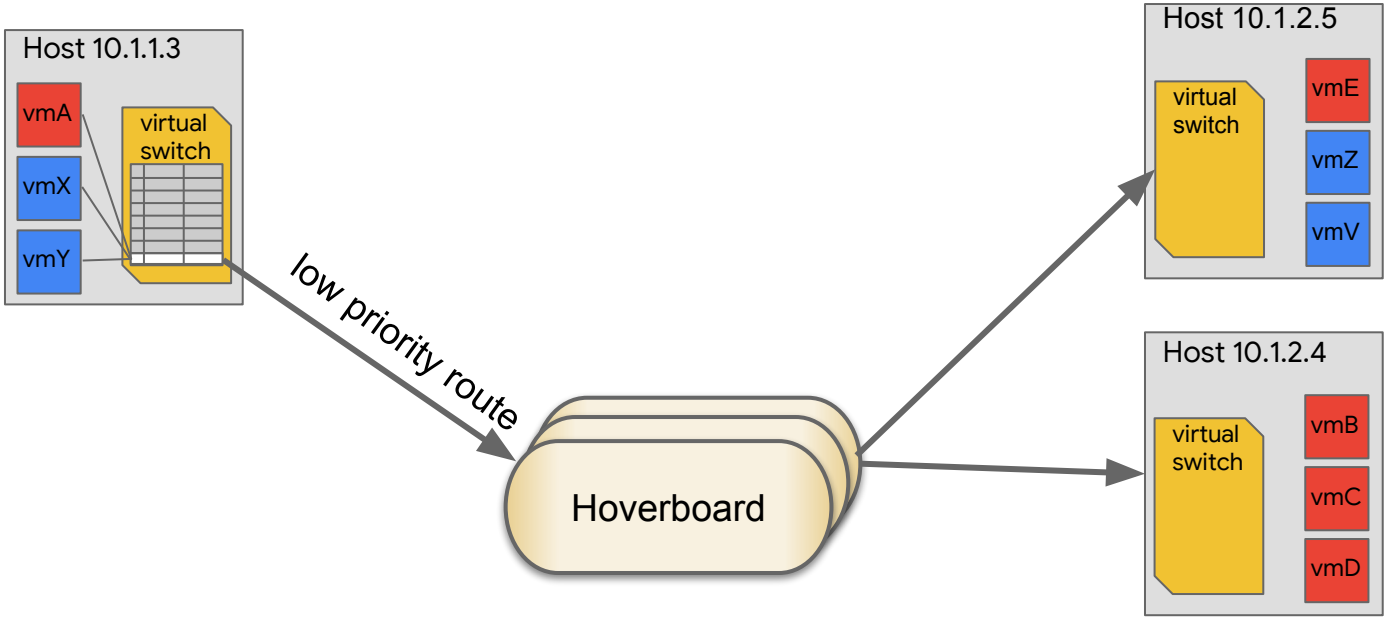
H = number of hosts

Quadratic scaling leads to provisioning challenges

- Control plane CPU and memory
- Dataplane memory

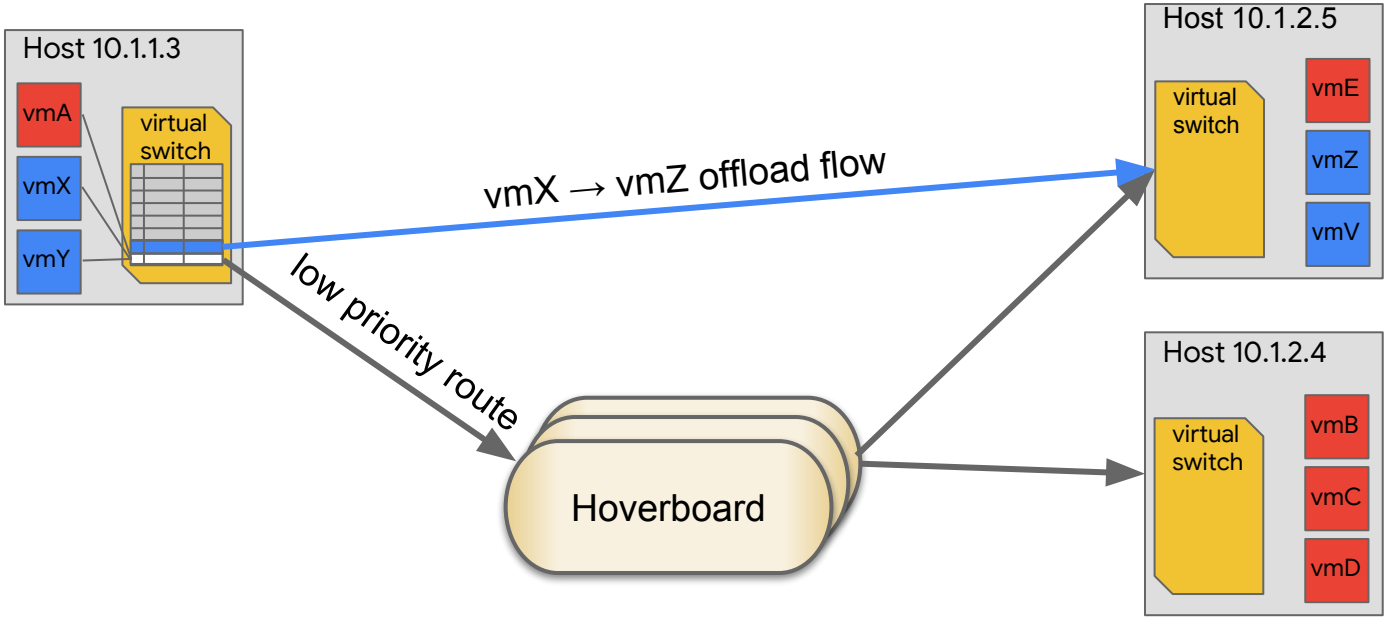
Google Cloud

Scaling with Hoverboards



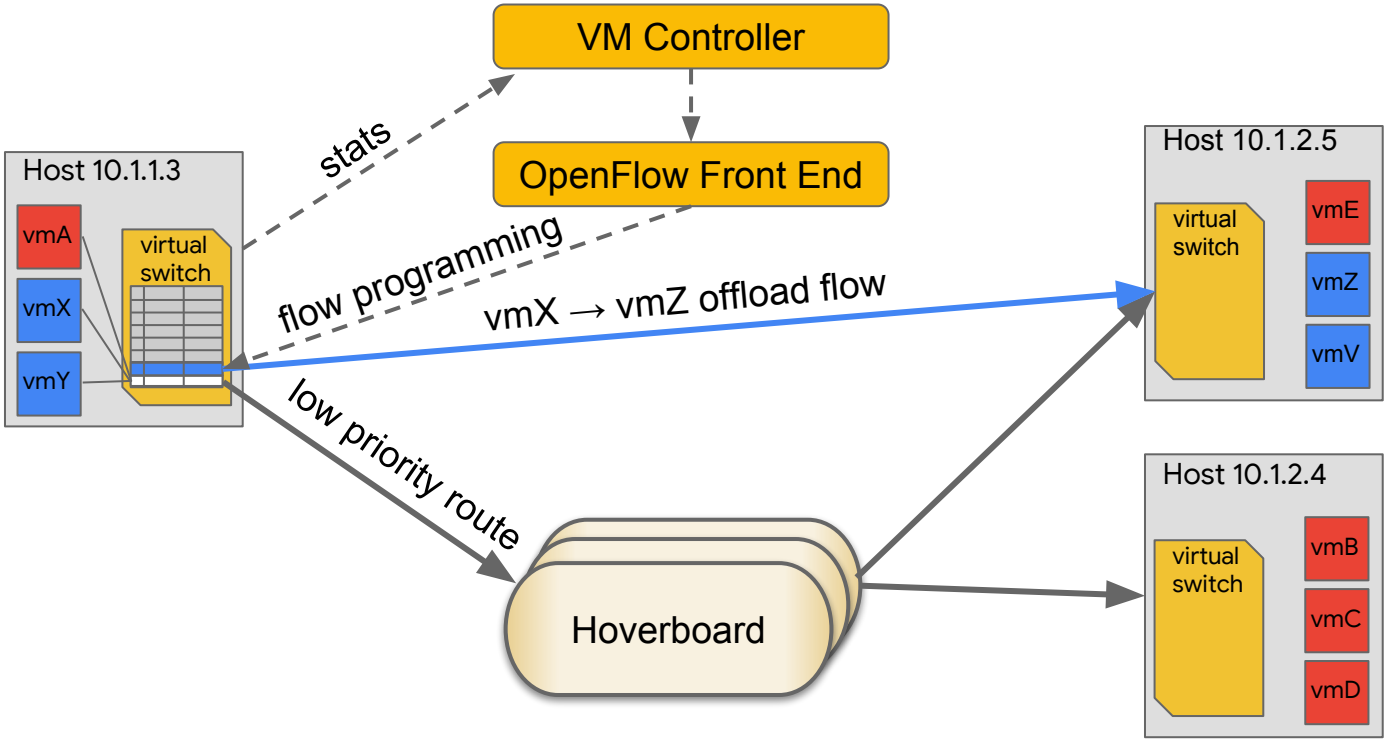
Google Cloud

Hoverboard Offloading



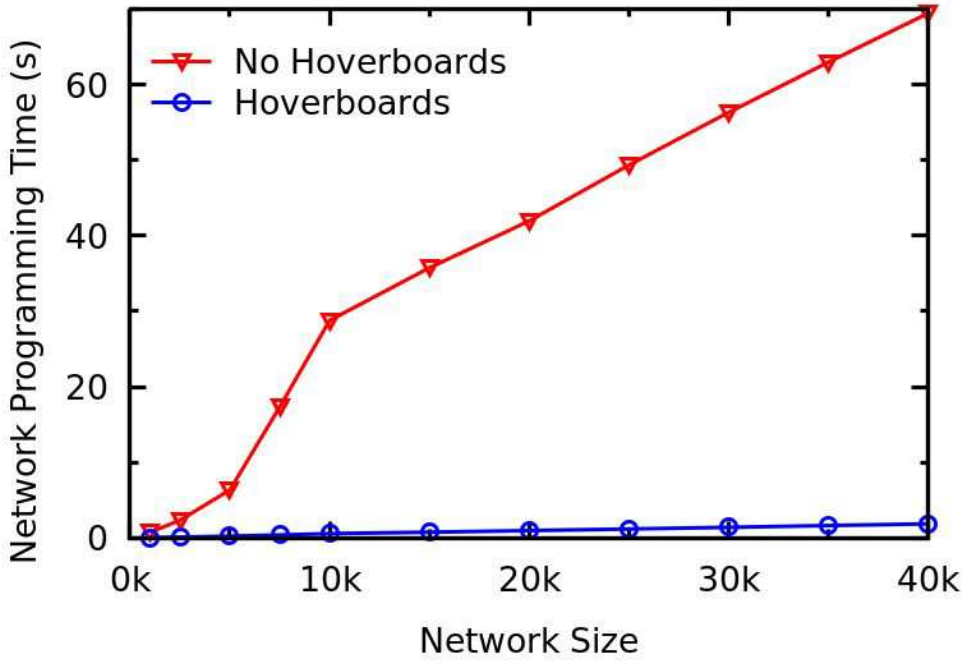
Google Cloud

Hoverboard Offloading



Google Cloud

Programming Time for Large Networks



Hoverboards reduce time to program network connectivity for large networks

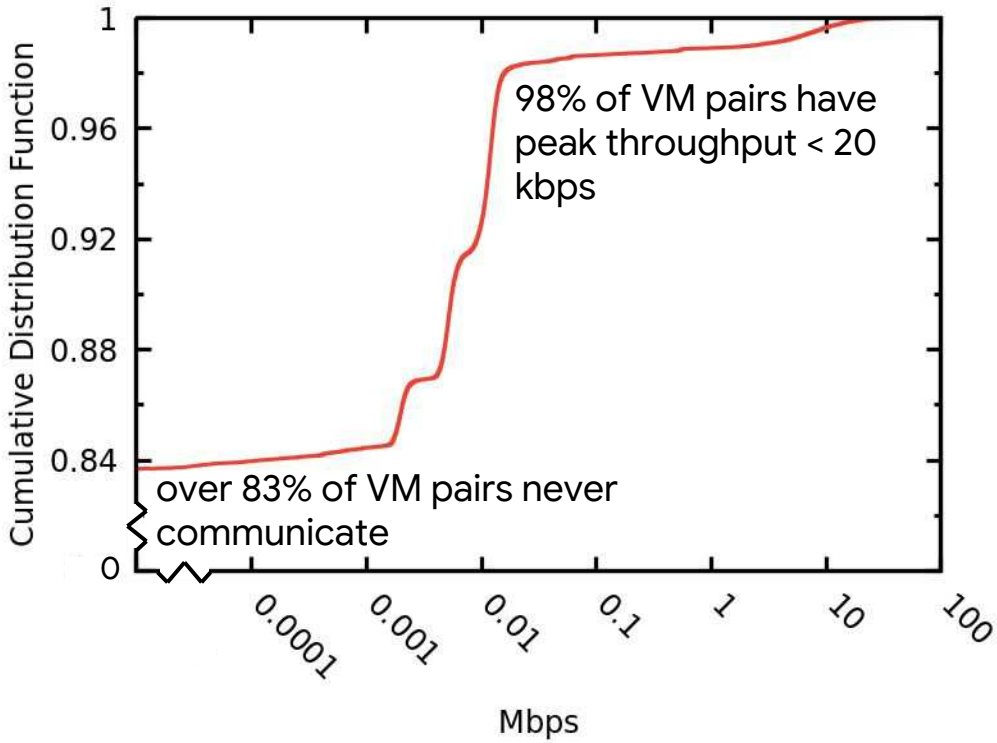
- 37X faster for a 40,000-VM network

Google Cloud

Why Hoverboards Are Effective

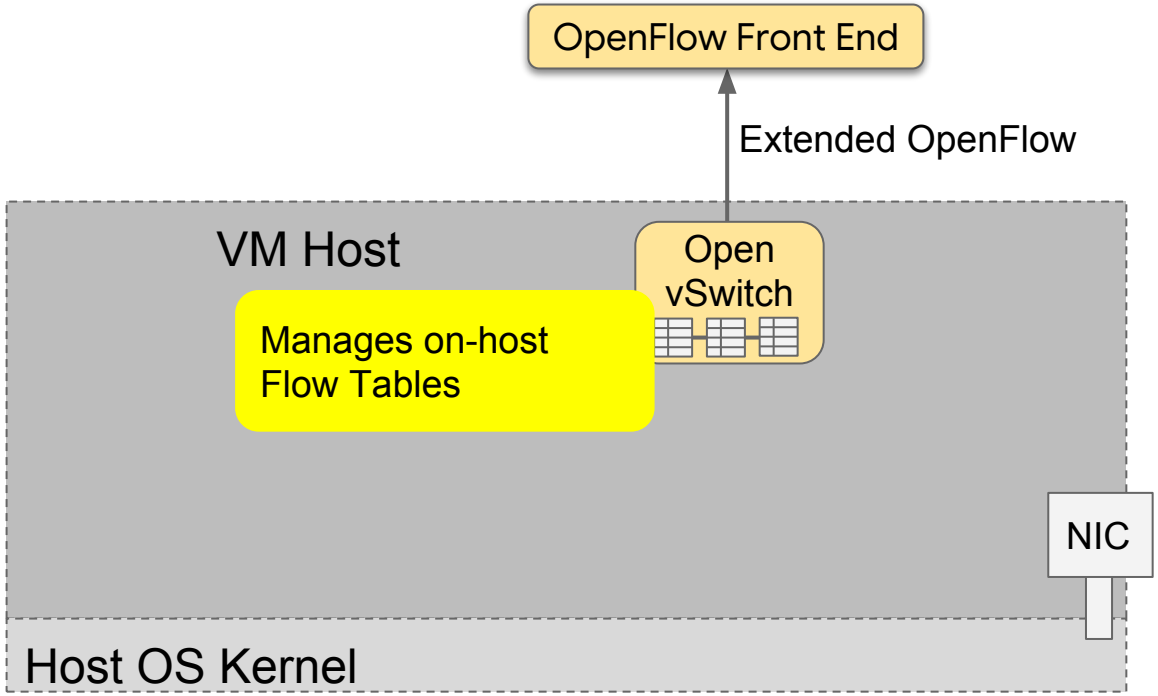
Peak throughput for all VM pairs in all virtual networks in one cluster over a 30-minute interval

Today, **more than 99.5% of traffic is offloaded**.



Google Cloud

Andromeda Data Plane

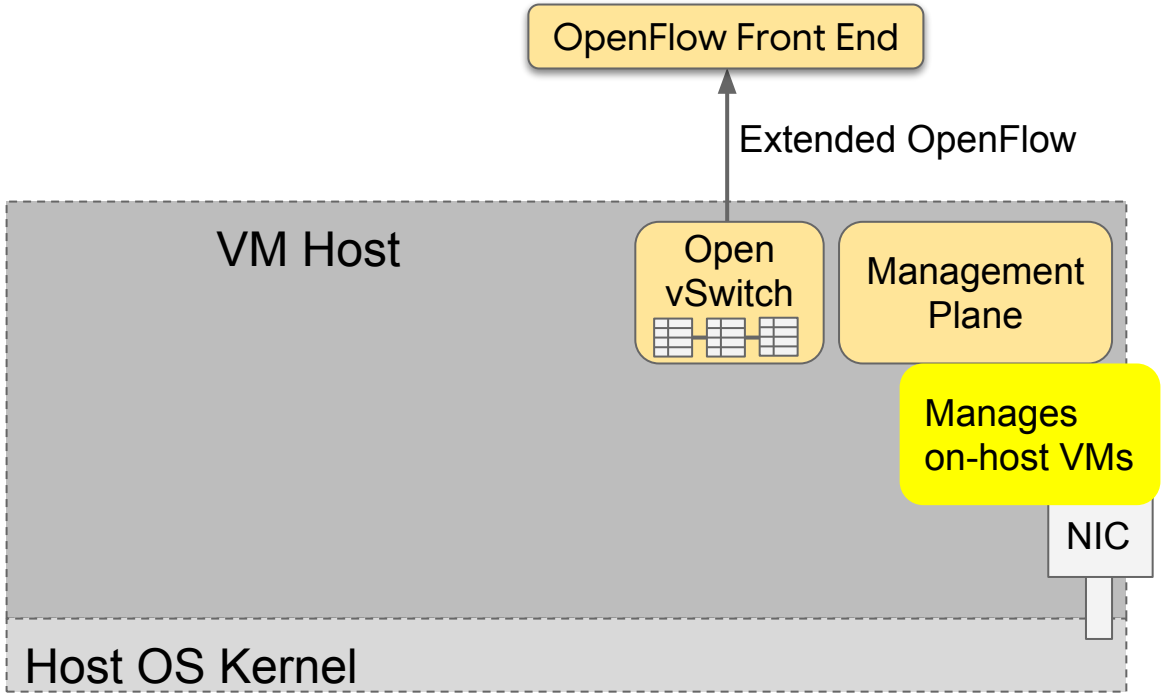


OS bypass, busy polling dedicated CPU Fast Path for **high performance**

Userspace dataplane, live migration, and hitless upgrades for **feature velocity**

Google Cloud

Andromeda Data Plane

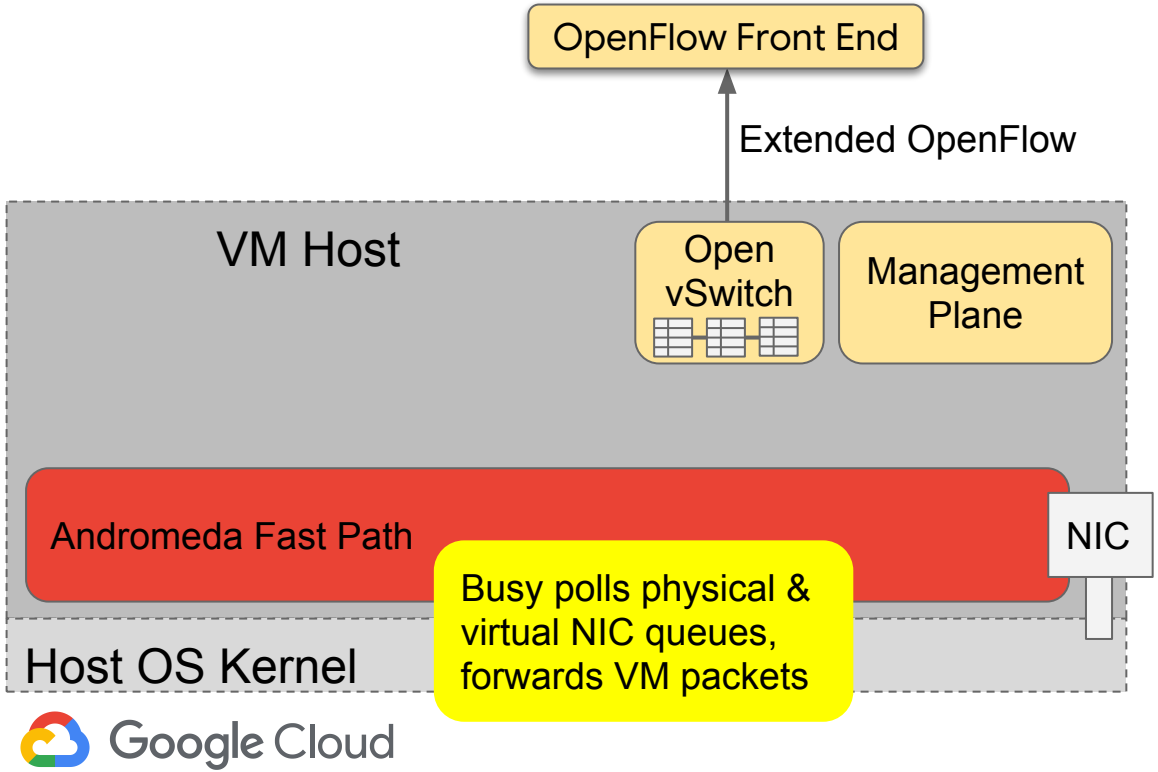


OS bypass, busy polling dedicated CPU Fast Path for **high performance**

Userspace dataplane, live migration, and hitless upgrades for **feature velocity**

Google Cloud

Andromeda Data Plane

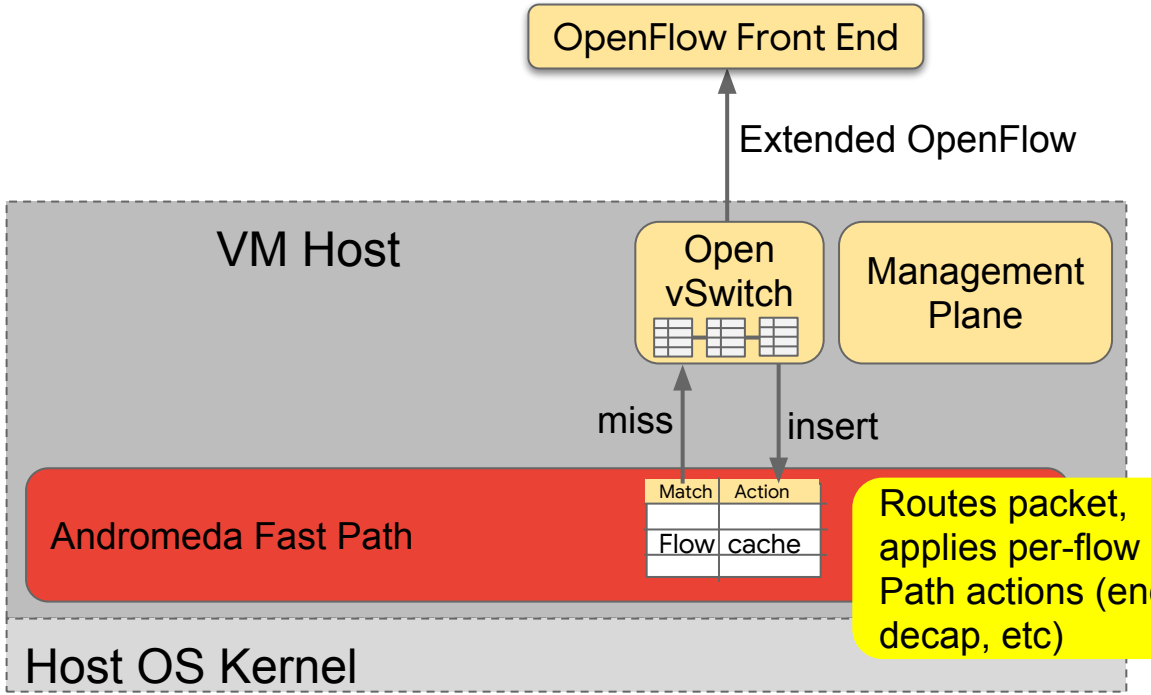


OS bypass, busy polling dedicated CPU Fast Path for **high performance**

Userspace dataplane, live migration, and hitless upgrades for **feature velocity**

Google Cloud

Andromeda Data Plane

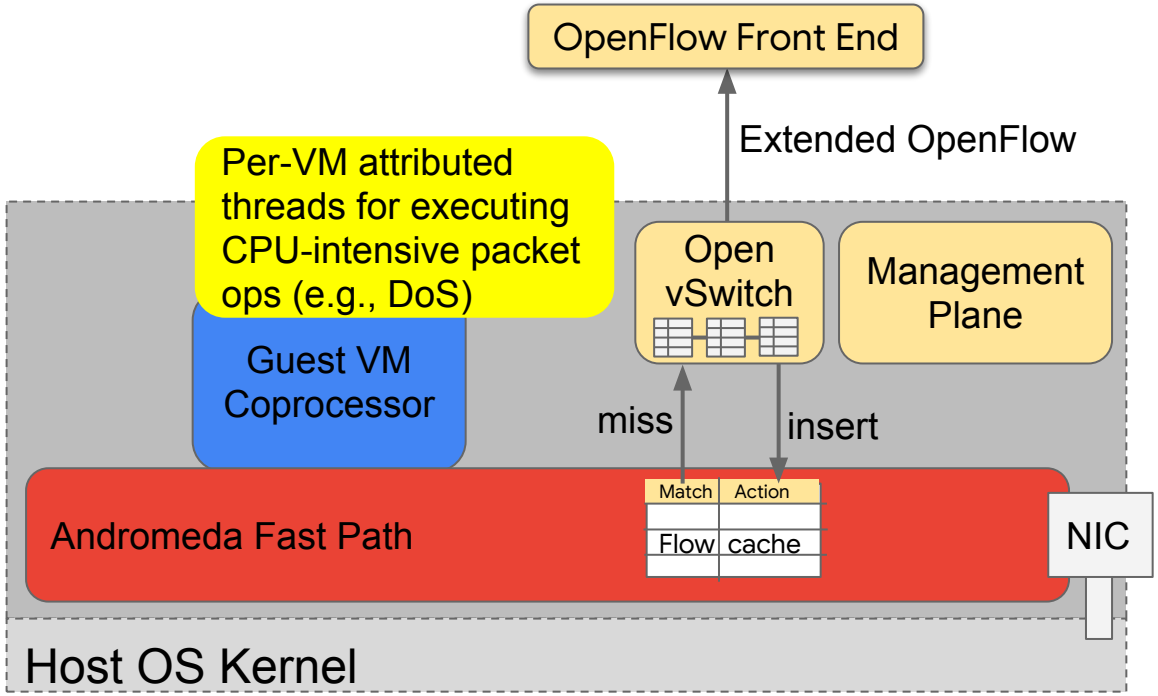


OS bypass, busy polling dedicated CPU Fast Path for **high performance**

Userspace dataplane, live migration, and bitless upgrades for **feature velocity**

Google Cloud

Andromeda Data Plane



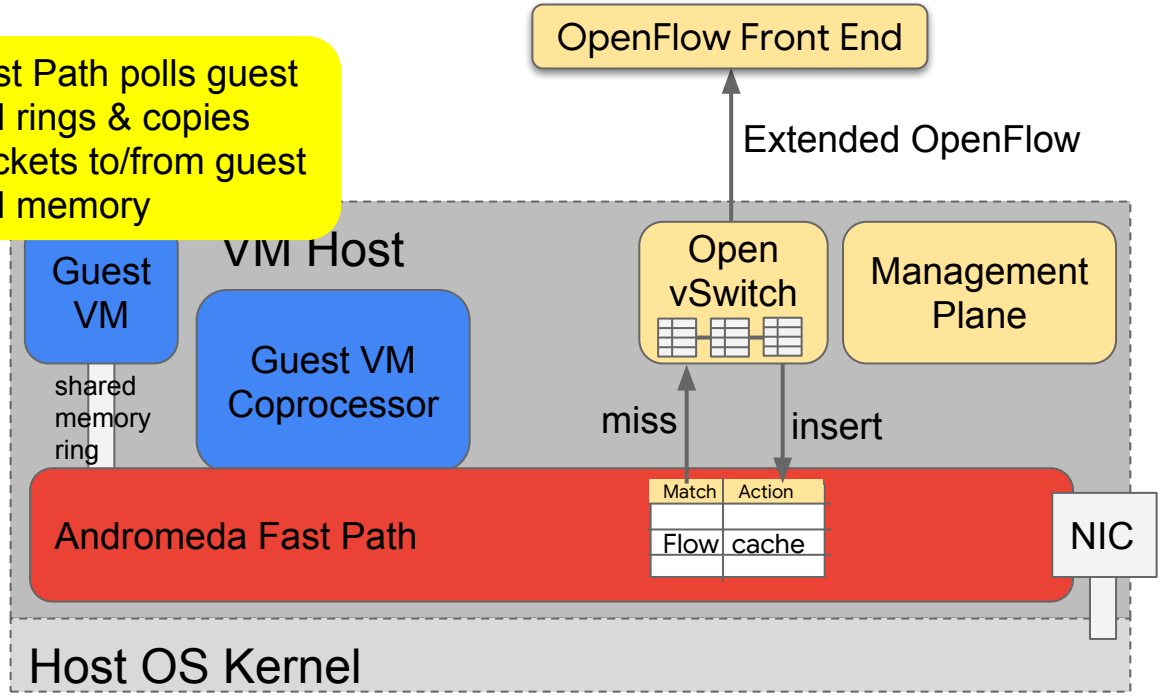
OS bypass, busy polling dedicated CPU Fast Path for **high performance**

Userspace dataplane, live migration, and hitless upgrades for **feature velocity**

Google Cloud

Andromeda Data Plane

Fast Path polls guest VM rings & copies packets to/from guest VM memory

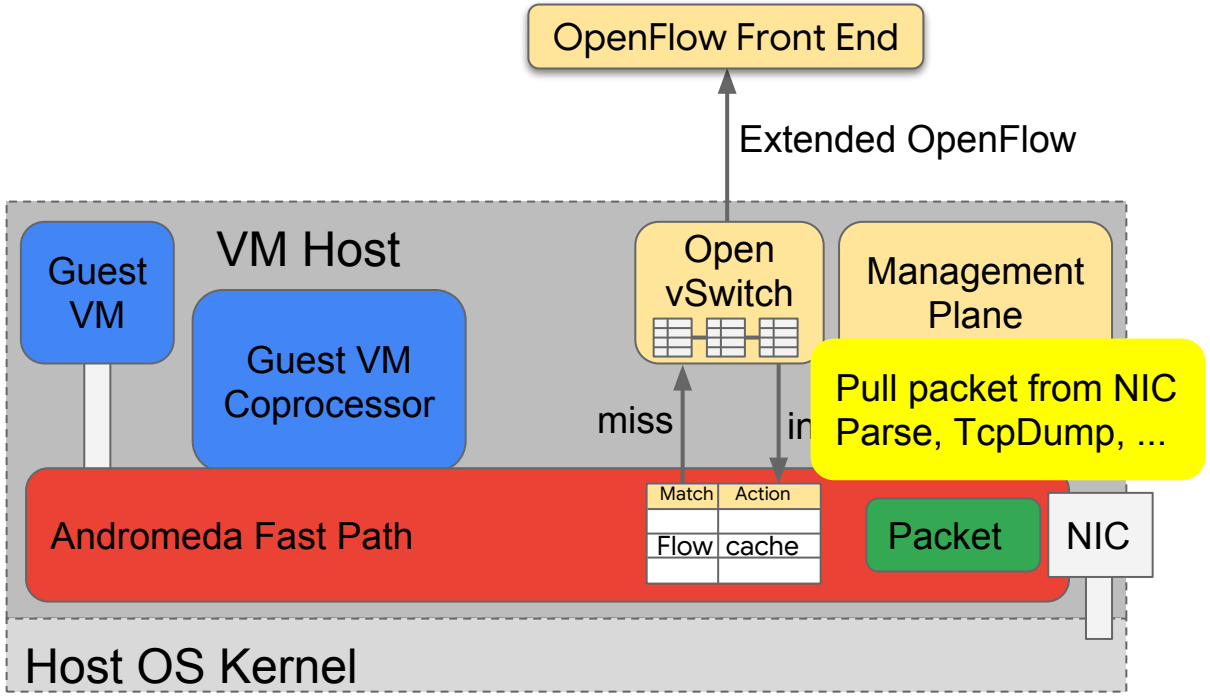


OS bypass, busy polling dedicated CPU Fast Path for **high performance**

Userspace dataplane, live migration, and hitless upgrades for **feature velocity**

Google Cloud

Data Plane - Fast Path



High performance traffic processed end-to-end on **Fast Path**

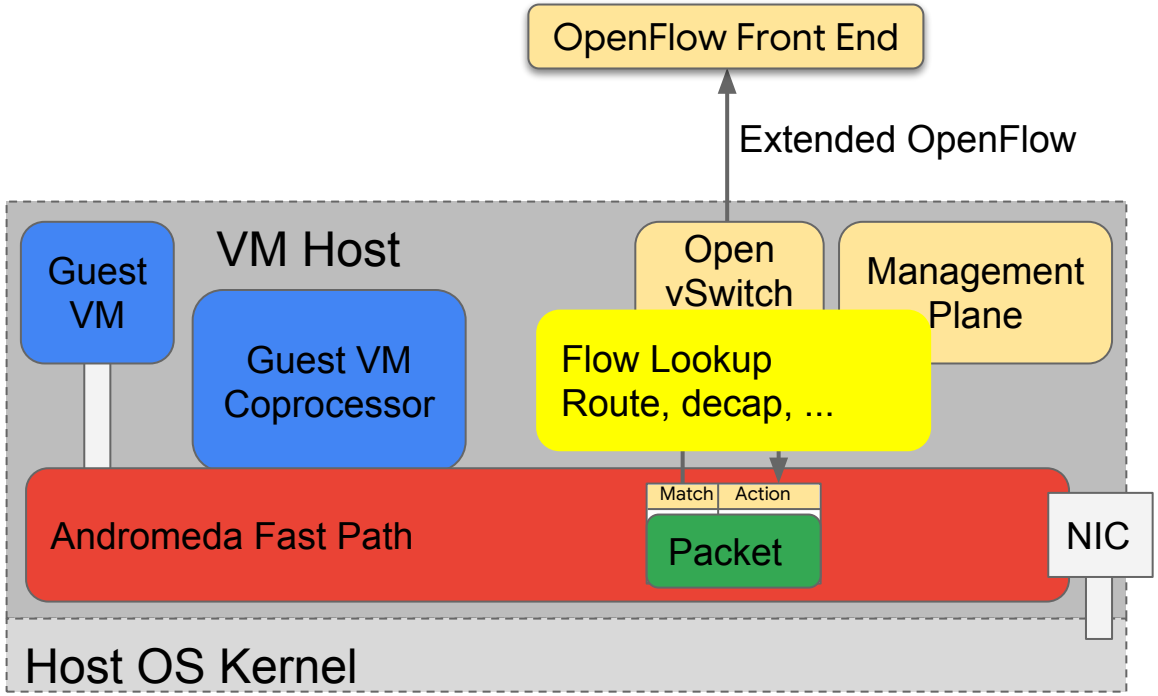
> 30Gb/s throughput &
> 3M pps on one core

Flow Table performs routing, encap/decap, etc.

Fast Path polls virtual & physical NIC rings

Google Cloud

Data Plane - Fast Path



Google Cloud

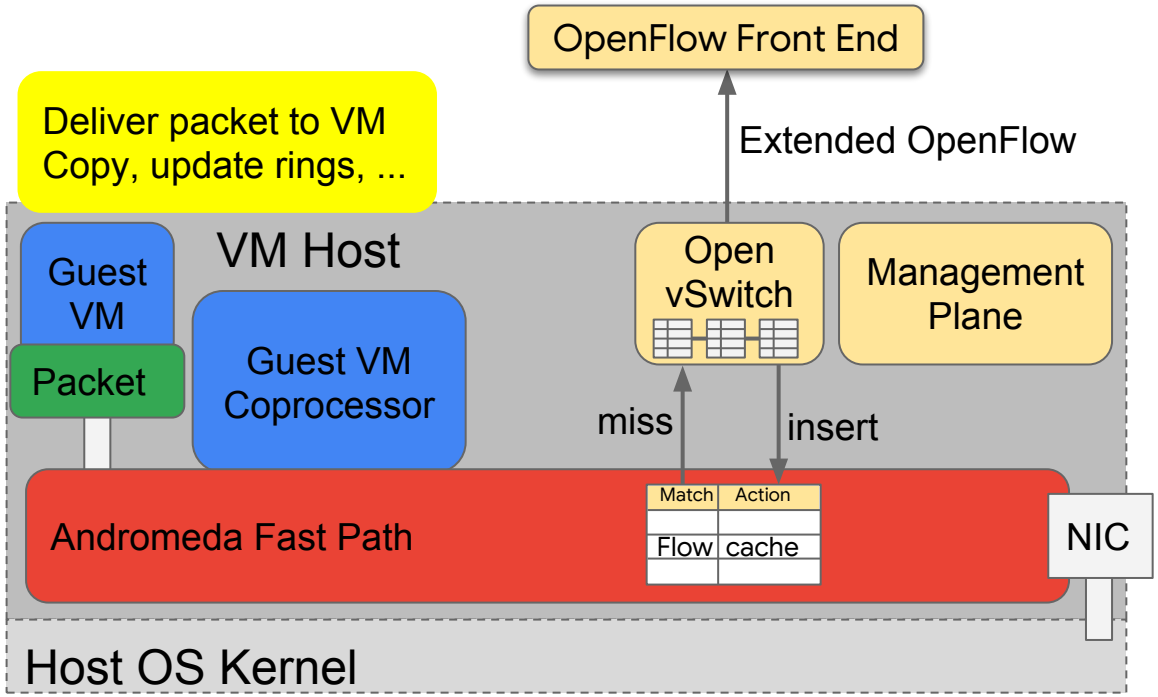
High performance traffic processed end-to-end on **Fast Path**

> 30Gb/s throughput &
> 3M pps on one core

Flow Table performs routing, encap/decap, etc.

Fast Path polls virtual & physical NIC rings

Data Plane - Fast Path



Google Cloud

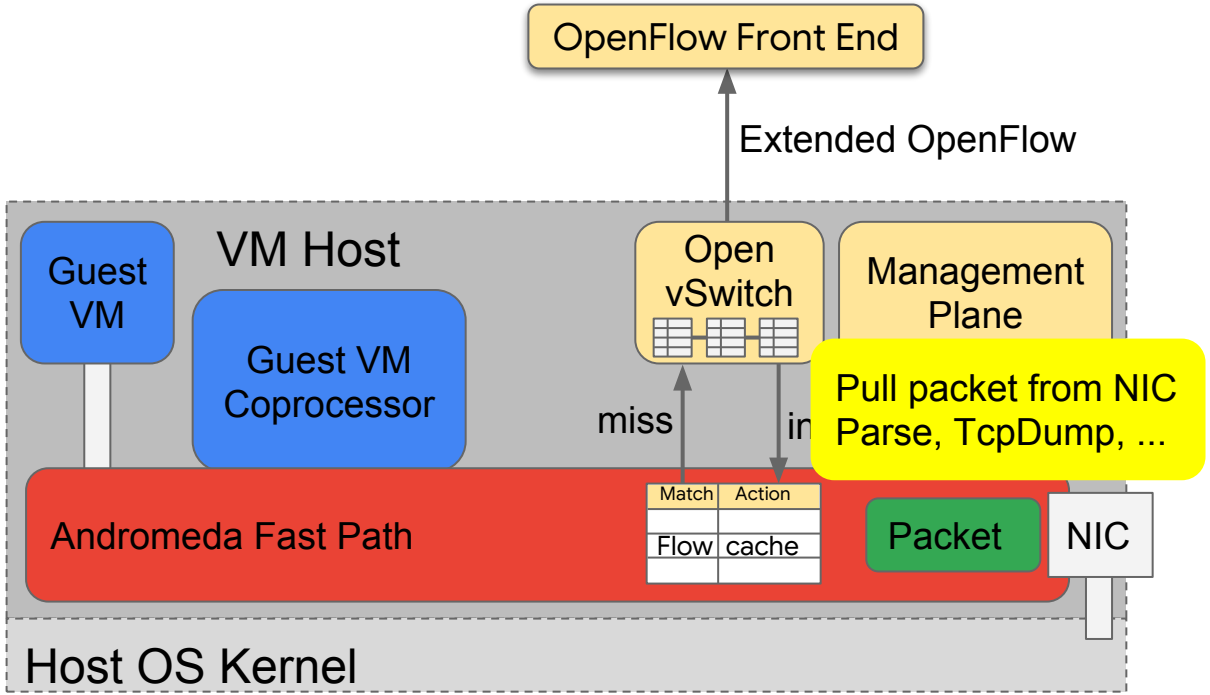
High performance traffic processed end-to-end on **Fast Path**

> 30Gb/s throughput &
> 3M pps on one core

Flow Table performs routing, encap/decap, etc.

Fast Path polls virtual & physical NIC rings

Data Plane - Coprocessor Path



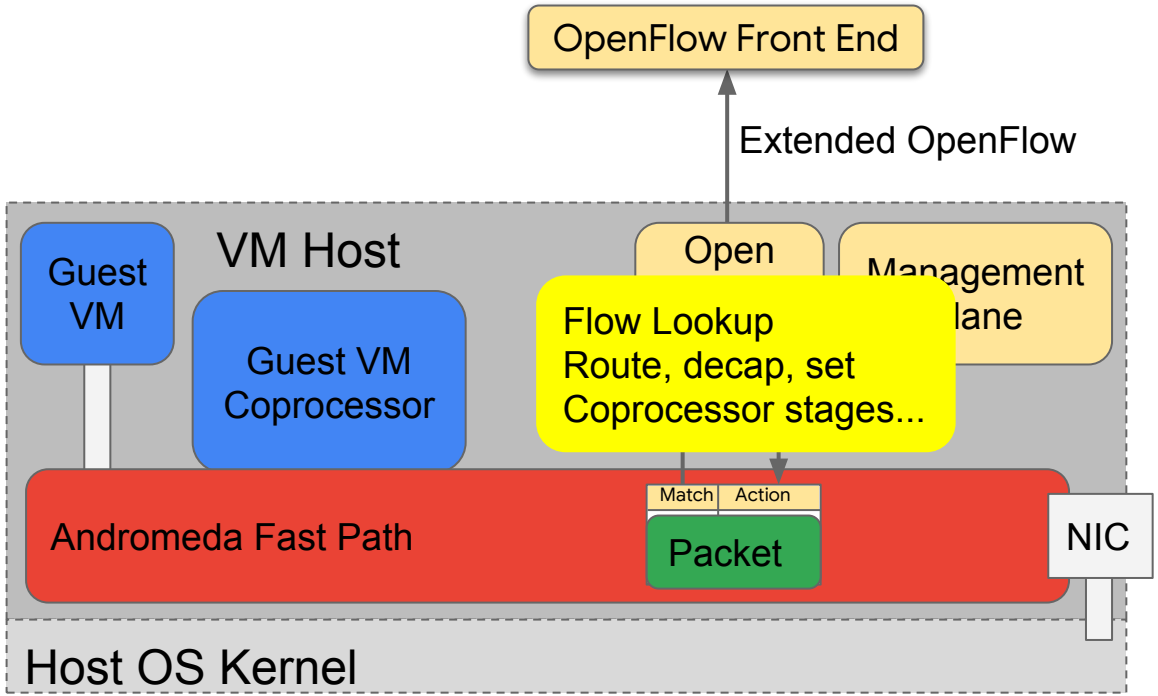
Coprocessors are per-VM threads CPU attributed to VM container

Coprocessors execute CPU-intensive packet ops such as DoS

Decouples feature growth from Fast Path speed

Google Cloud

Data Plane - Coprocessor Path



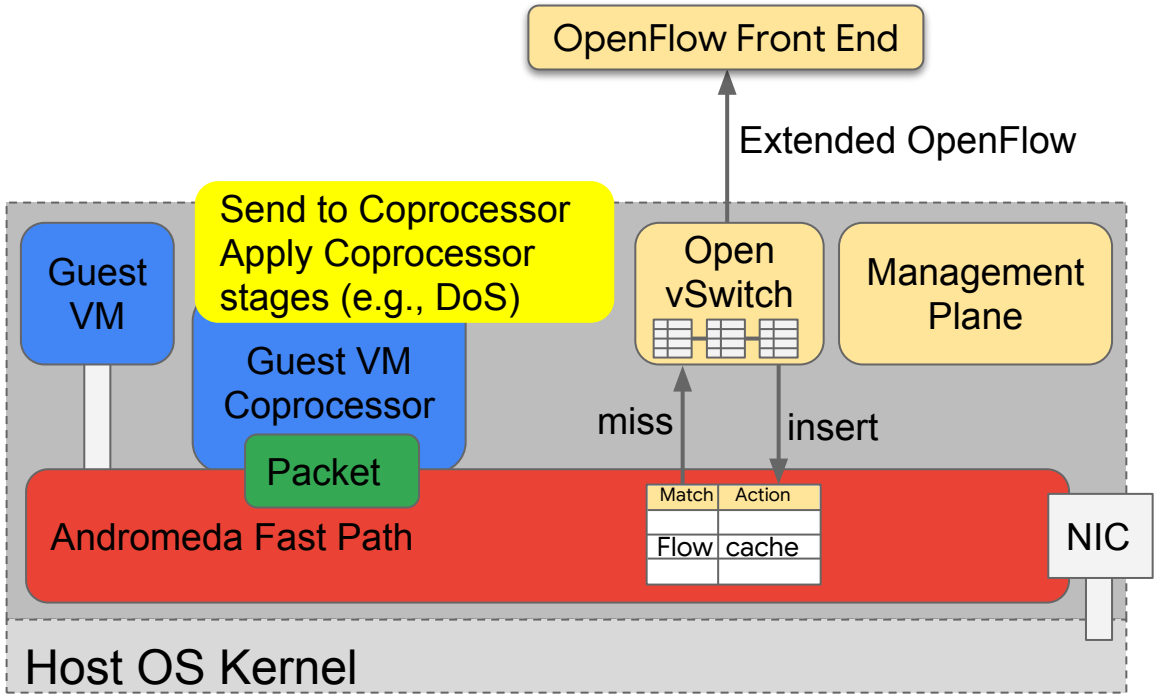
Coprocessors are per-VM threads CPU attributed to VM container

Coprocessors execute CPU-intensive packet ops such as DoS

Decouples feature growth from Fast Path speed

Google Cloud

Data Plane - Coprocessor Path



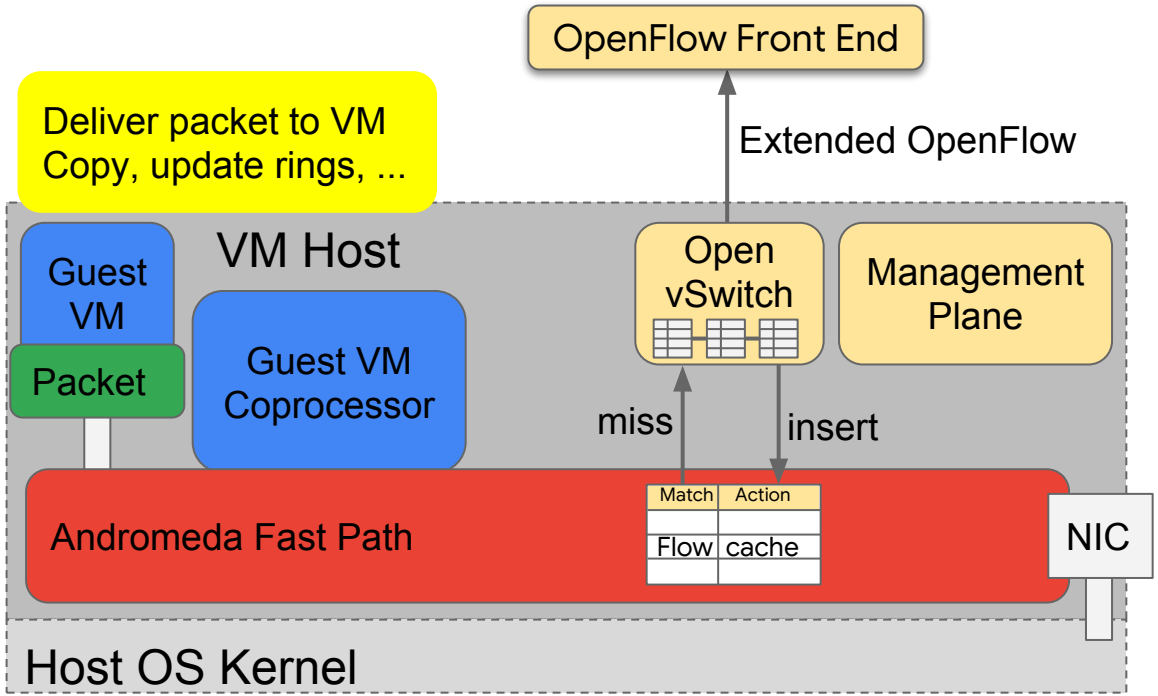
Coprocessors are per-VM threads CPU attributed to VM container

Coprocessors execute CPU-intensive packet ops such as DoS

Decouples feature growth from Fast Path speed

Google Cloud

Data Plane - Coprocessor Path



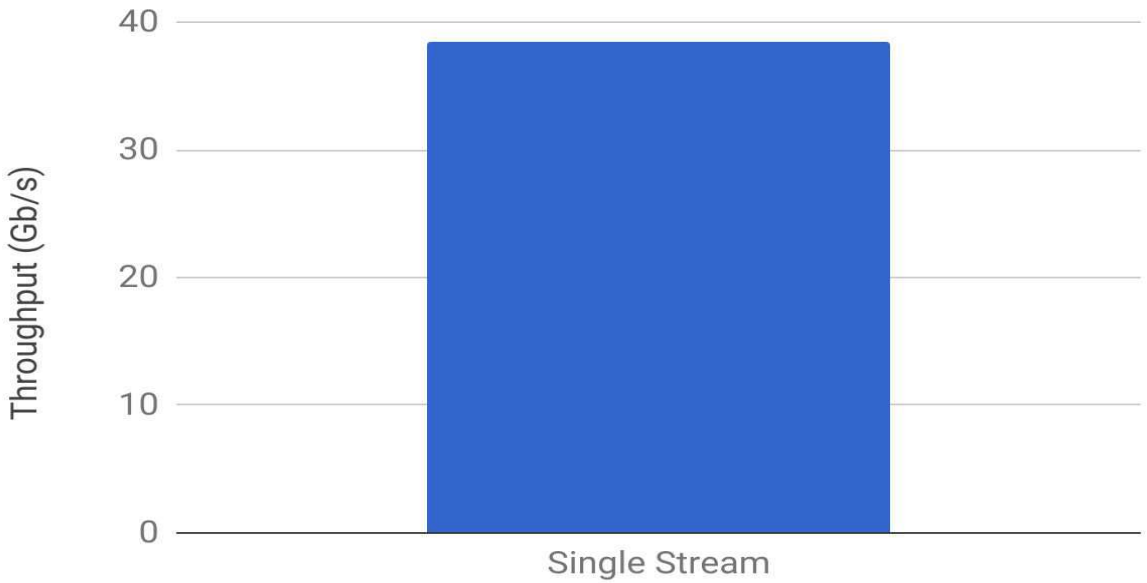
Coprocessors are per-VM threads CPU attributed to VM container

Coprocessors execute CPU-intensive packet ops such as DoS

Decouples feature growth from Fast Path speed

Google Cloud

VM-VM Throughput

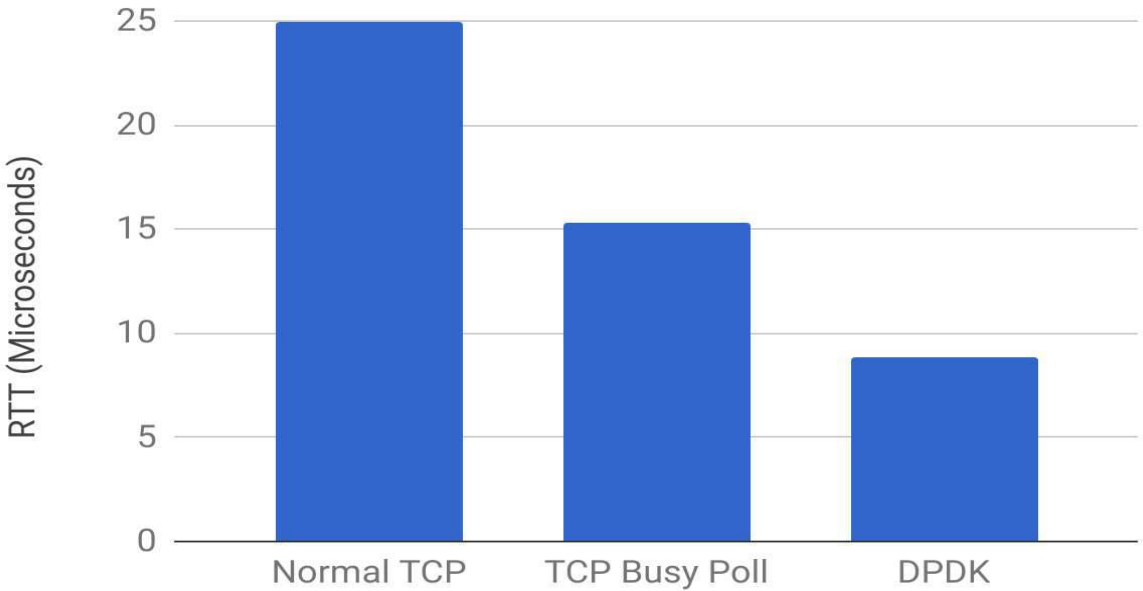


Single core per host for dataplane Fast Path. Skylake testbed hosts.

Both hosts connected to same Top of Rack switch.

Google Cloud

VM-VM Round Trip Latency



Single core per host for dataplane Fast Path. Skylake testbed hosts.

Both hosts connected to same Top of Rack switch.

Google Cloud

CPU Efficiency

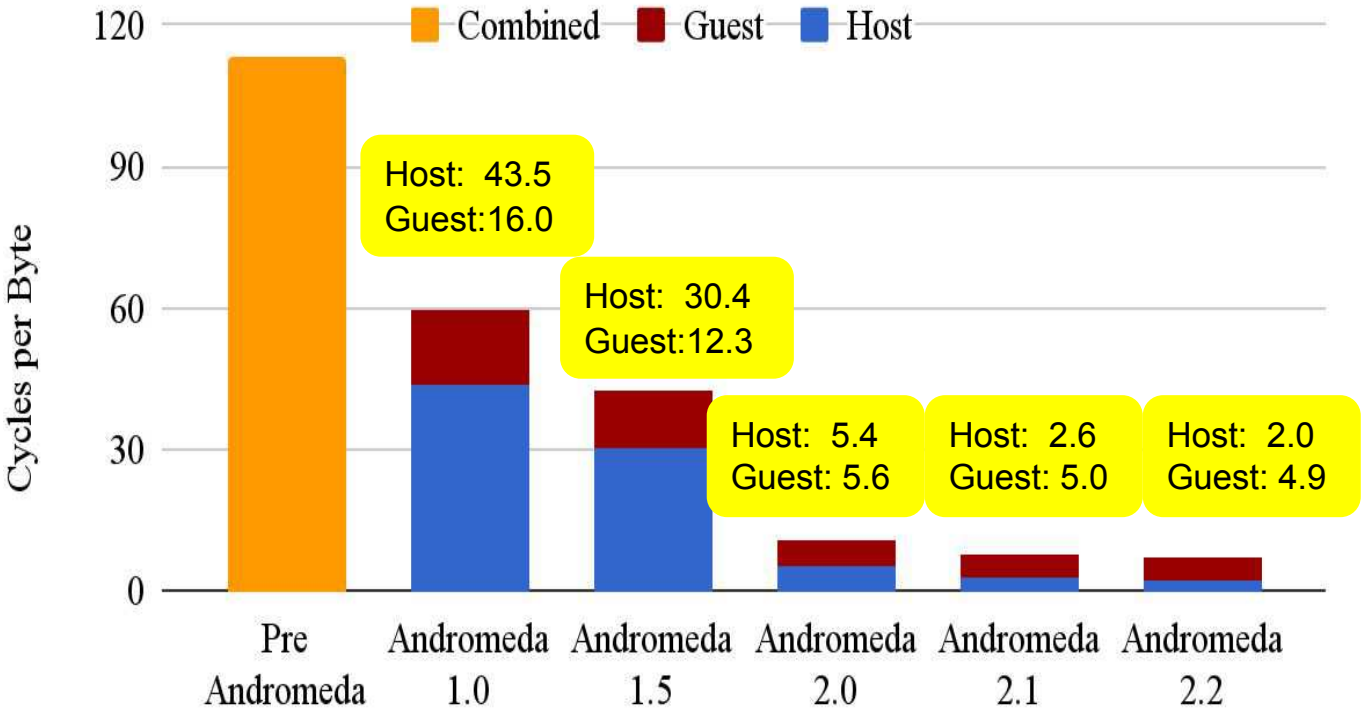
Minimizing host (and guest) network CPU cycles per byte (CPB) is critical

Since initial production release, we have improved CPB by **> 16x** as measured on sender + receiver host during a multi-stream benchmark.

Andromeda 2.0+ use a **single** core per host for the dataplane Fast Path. Results from Sandybridge testbed hosts connected to same ToR switch.

Google Cloud

CPU Efficiency Evolution



Andromeda 1.0
Kernel datapath

Andromeda 1.5
Optimize pipeline

Andromeda 2.0
OS bypass, 1 thread hop

Andromeda 2.1
Remove thread hop

Andromeda 2.2
Memory copy offload

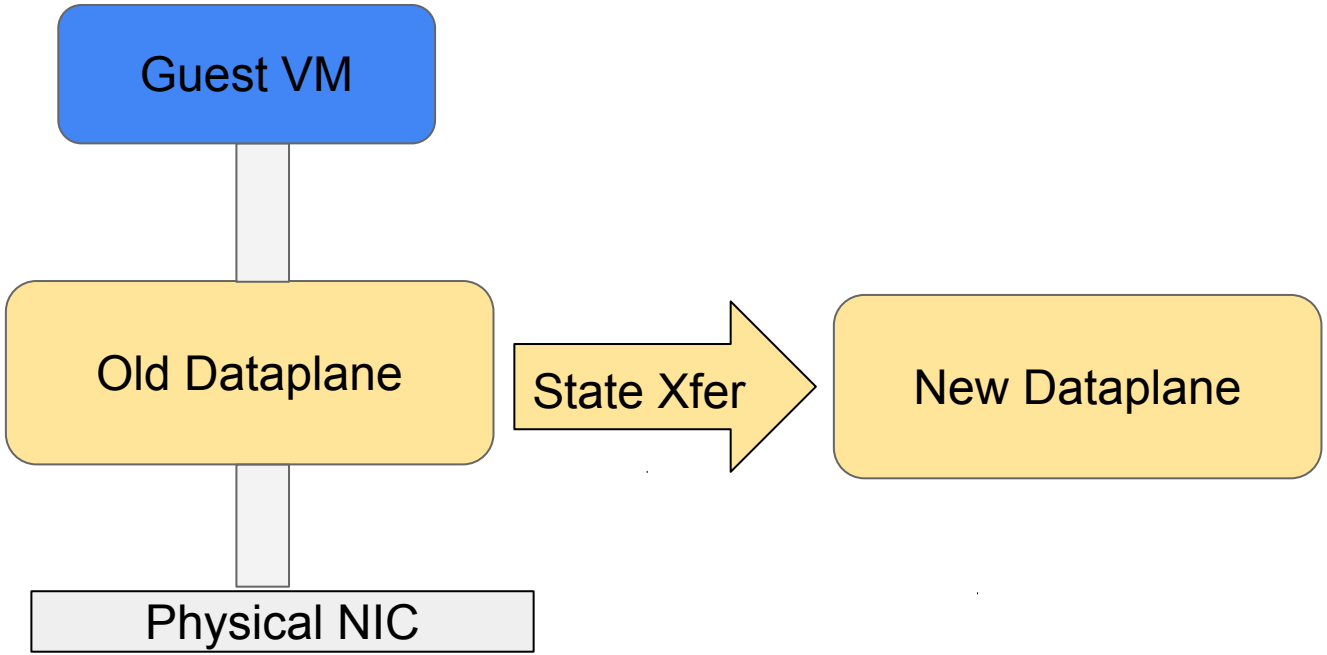
Velocity

A rapid release cycle enables swift deployment of features & bug fixes.

Our dataplane has **weekly** rollouts via non-disruptive upgrades.

Live migration allows VMs to be migrated between physical host without disruption, enabling transparent host maintenance.

Dataplane Hitless Upgrade 1 / 3



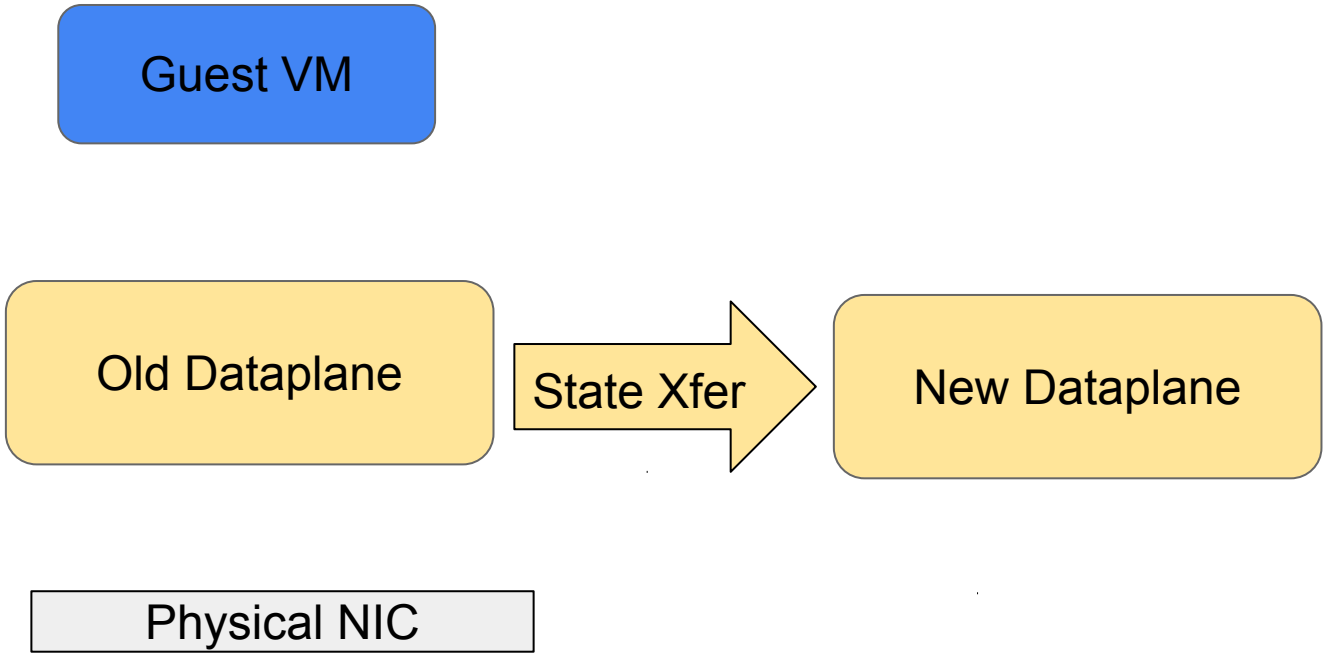
Upgrade Brownout

Old Dataplane state is transferred to New Dataplane in the background

Old Dataplane continues serving physical NIC & virtual NIC queues

Google Cloud

Dataplane Hitless Upgrade 2 / 3



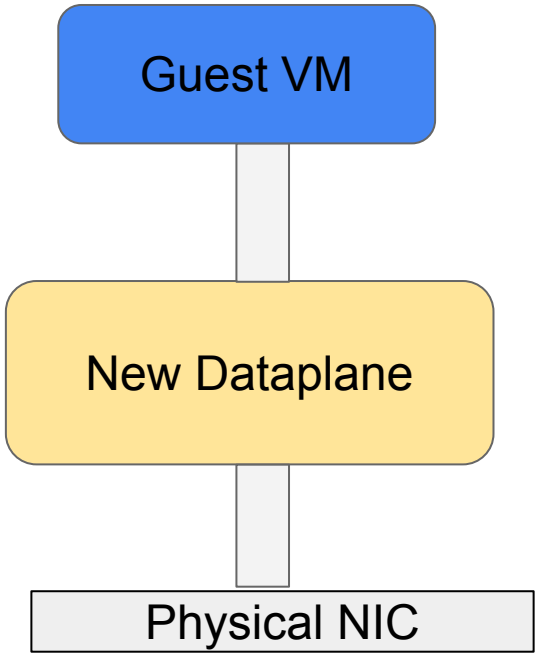
Upgrade Blackout

Old Dataplane stops serving virtual & physical NIC queues

Then, any updated (delta) Old Dataplane state is transferred to New Dataplane

Google Cloud

Dataplane Hitless Upgrade 3 / 3



Google Cloud

Upgrade Complete

State xfer done. Median blackout time is **270ms**.

New Dataplane starts serving VM virtual NIC & physical NIC queues

Old dataplane terminated

Conclusion

We have discussed the design and evolution of Andromeda

Control plane scalability & Rapid provisioning

- Hoverboard model avoids programming long tail of mostly idle flows on VM host. Scales to 100k VMs/network

High performance & Feature velocity

- OS Bypass dedicated CPU dataplane provides high performance (**> 30Gb/s**, **> 3M pps** with 1 core) & **weekly** non-disruptive updates