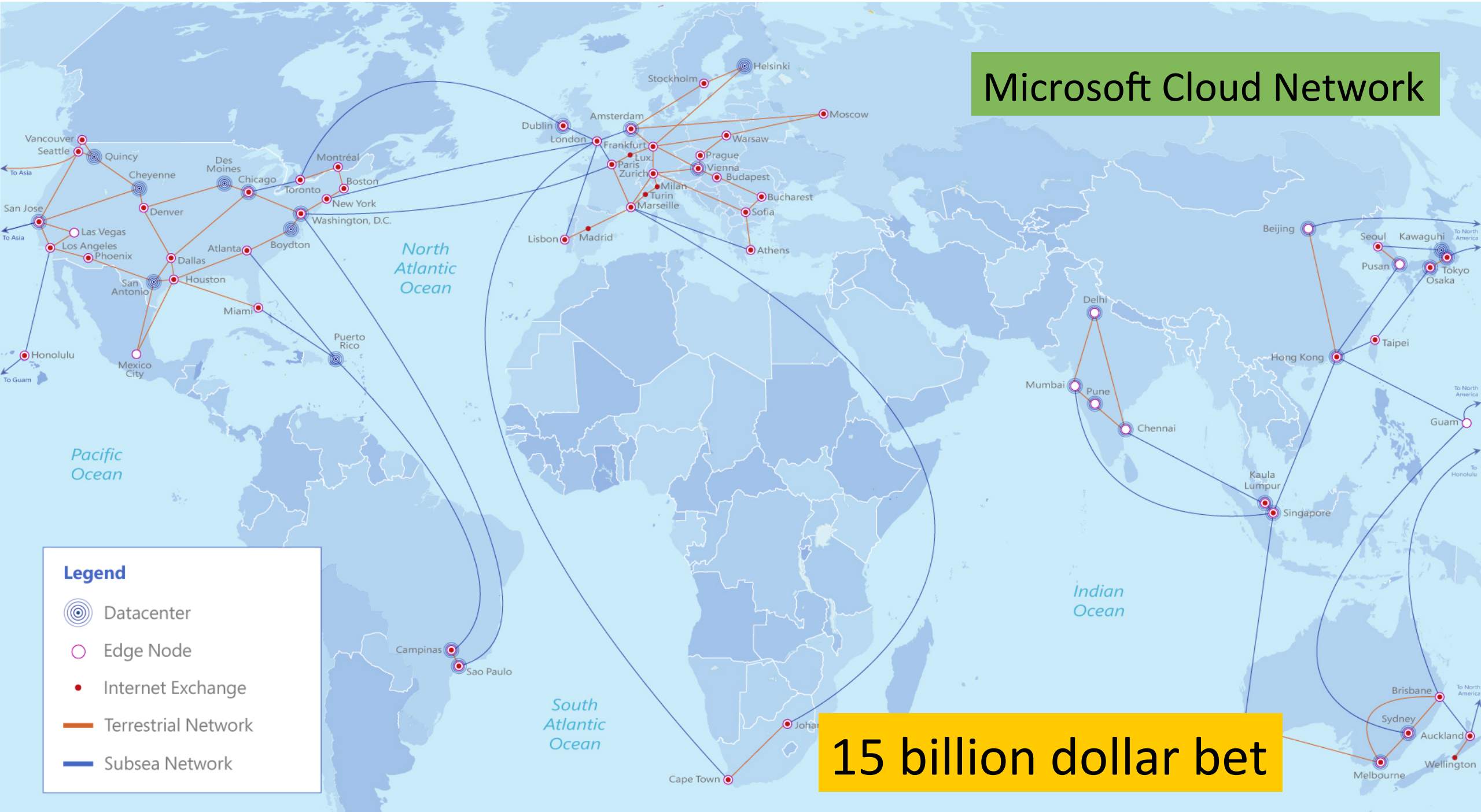
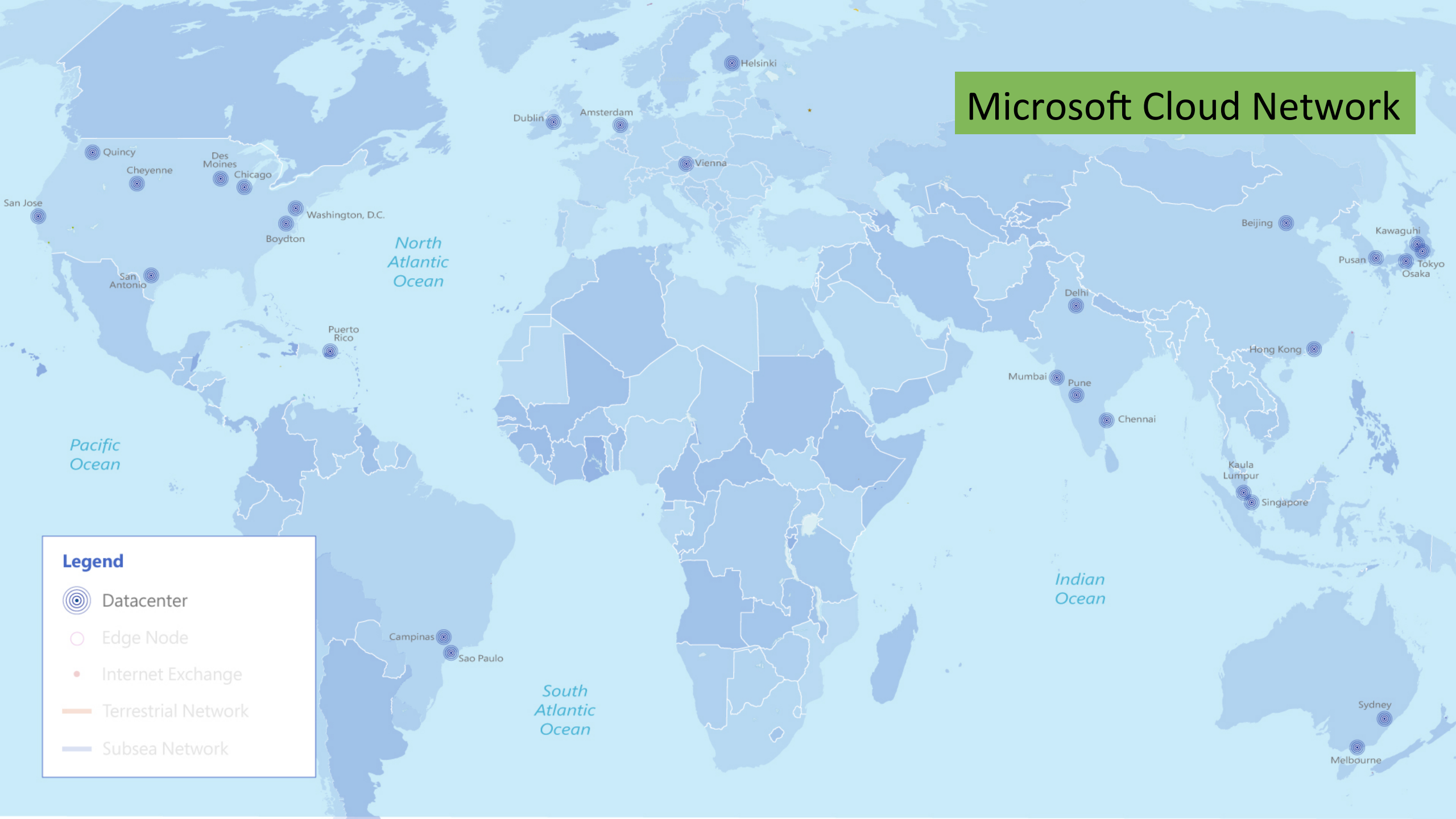
NFV for the Cloud

Parveen Patel

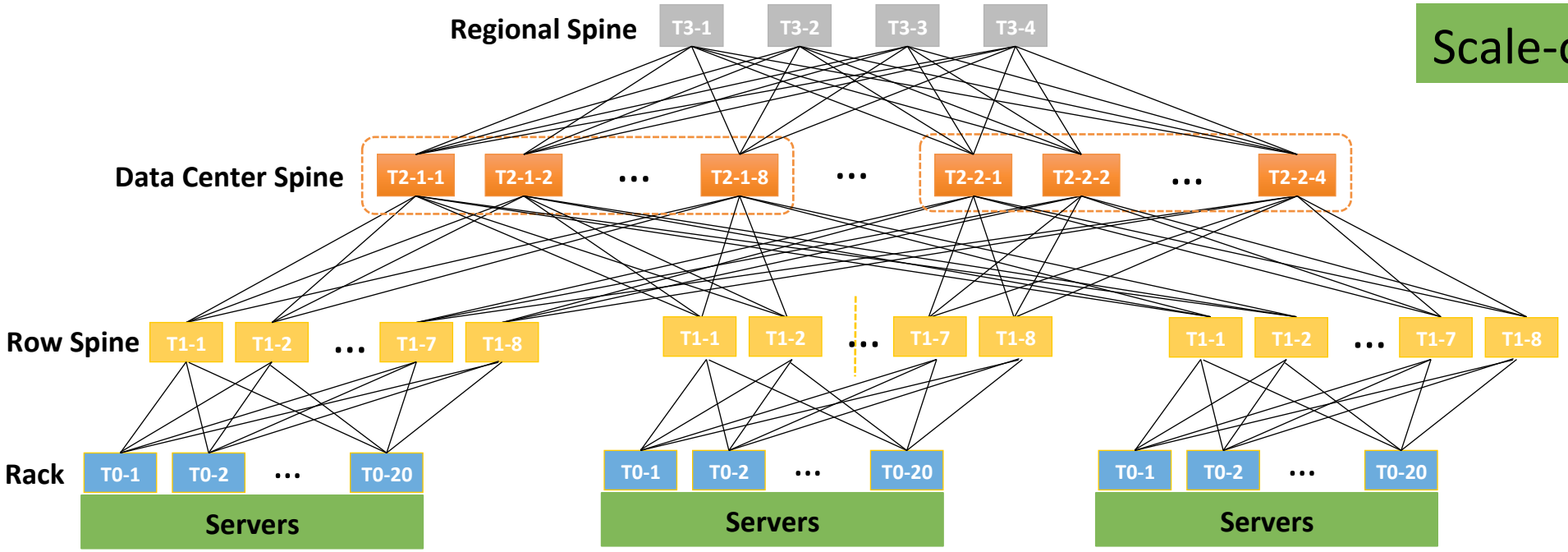
Group Engineering Manager @ Microsoft Azure

Parveen.patel@microsoft.com


Microsoft Cloud Network
15 billion dollar bet
Legend
Datacenter
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network
North Atlantic Ocean
Pacific Ocean
South Atlantic Ocean
Indian Ocean
Vancouver
Seattle
Quincy
To Asia
Cheyenne
Des Moines
Chicago
Montréal
Toronto
Boston
New York
Washington, D.C.
San Jose
Denver
Las Vegas
Los Angeles
Phoenix
Dallas
Atlanta
Boydton
San Antonio
Houston
Miami
Puerto Rico
Honolulu
To Guam
Mexico City
Campinas
Sao Paulo
Cape Town
Dublin
London
Amsterdam
Helsinki
Stockholm
Moscow
Warsaw
Frankfurt
Lux.
Paris
Zurich
Prague
Vienna
Budapest
Milan
Turin
Marseille
Bucharest
Sofia
Athens
Lisbon
Madrid
Beijing
Seoul
Kawaguhi
Pusan
Tokyo
Osaka
Delhi
Taipei
Hong Kong
Mumbai
Pune
Chennai
Guam
Kaula Lumpur
Singapore
Brisbane
Sydney
Auckland
Melbourne
Wellington
To North America
To Honolulu

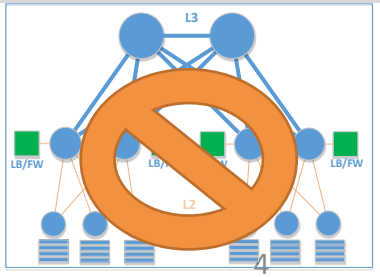

Microsoft Cloud Network
Helsinki
Dublin
Amsterdam
Vienna
Quincy
Cheyenne
Des Moines
Chicago
San Jose
Washington, D.C.
Boydton
San Antonio
Puerto Rico
North Atlantic Ocean
Pacific Ocean
Beijing
Kawaguhi
Pusan
Tokyo
Osaka
Delhi
Hong Kong
Mumbai
Pune
Chennai
Kaula Lumpur
Singapore
Indian Ocean
Campinas
Sao Paulo
South Atlantic Ocean
Sydney
Melbourne
Legend
Datacenter
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network

Physical Network: Azure Clos Fabrics (VL2)



Scale-out, active-active

Scale-up, active-passive



Outcome of >10 years of history, with major revisions every six months

Challenge of Multi-tenant Cloud Networking

- Provide on-demand network services to customer networks
  - Routing
  - Segmentation

  } Virtual Networking

  - Load Balancer
  - NAT
  - VPN
  - WAN Connectivity
  - Application Security
  - DDoS Protection
  - …

  } Virtualized Network Functions (VNFs)

The job of Network Function Virtualization (NFV) is to enable VNFs

Agenda

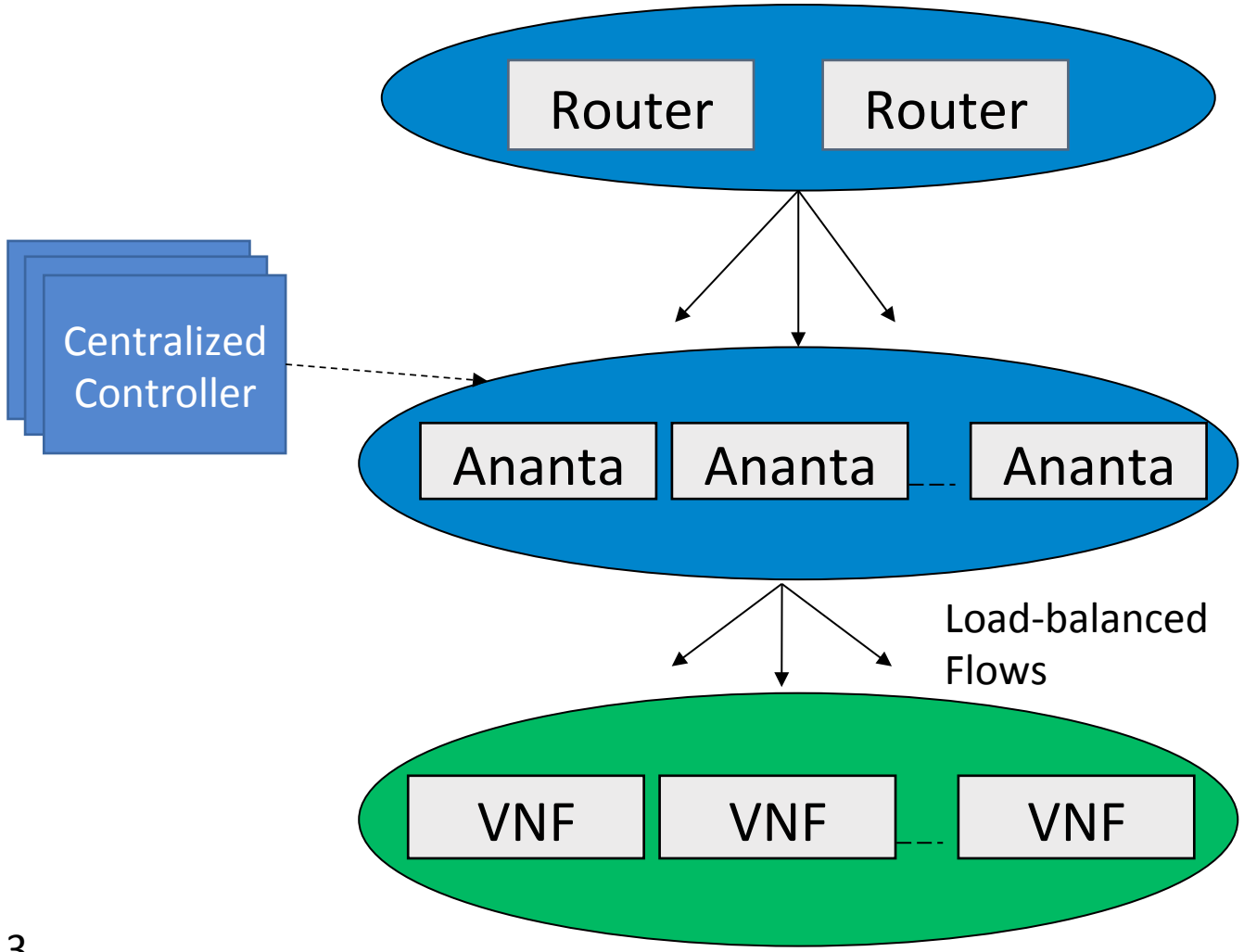
- Requirements for NFV
- Approach
- Example VNFs: VPN, Express Route, Application Gateway
- Lessons learnt
- Call for action for P4 community

Requirements for NFV Infrastructure

- No specialized, fixed-function hardware
  - VNFs need to be implemented on programmable hardware
  - Evolve at the speed of software
- On-demand provisioning
  - 1000s of instances in seconds
- Scale-out to meet availability and capacity
  - Once the data center is built, reprogram hardware to meet customers' needs
- Support both single-tenant and multi-tenant VNFs
  - Support legacy as well as cloud-scale VNFs

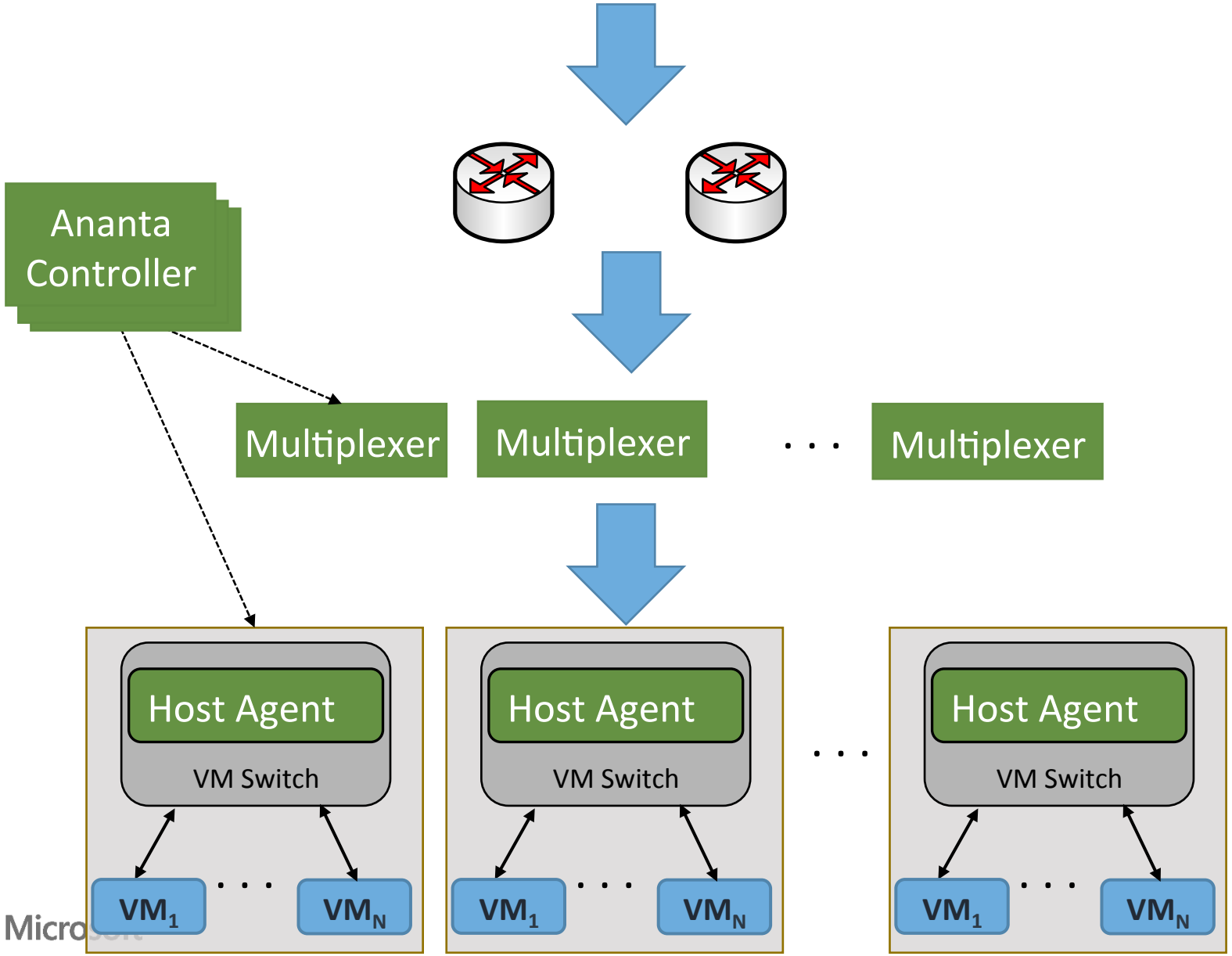
Approach: Scale VNFs via Layer-4 LB

- **Key Ideas**
  - Ananta*: Scale-out Layer-4 load balancing via SDN
  - Ananta provides load balancing and high availability for individual VNFs
  - Centralized control plane and scale-out data plane for each VNF, e.g., VPN, Application Security, DPI



*Ananta: Cloud Scale Load Balancing, SIGCOMM 2013

Ananta: Cloud Scale Load Balancing



Idea: decompose load balancing into tiers.

1st Tier: Provides packet-level (layer-3) load spreading, implemented in routers via ECMP.

2nd Tier: Provides **connection-level** (Layer-4) load spreading and health monitoring

3rd Tier: Provides stateful NAT implemented in the virtual switch in every server.

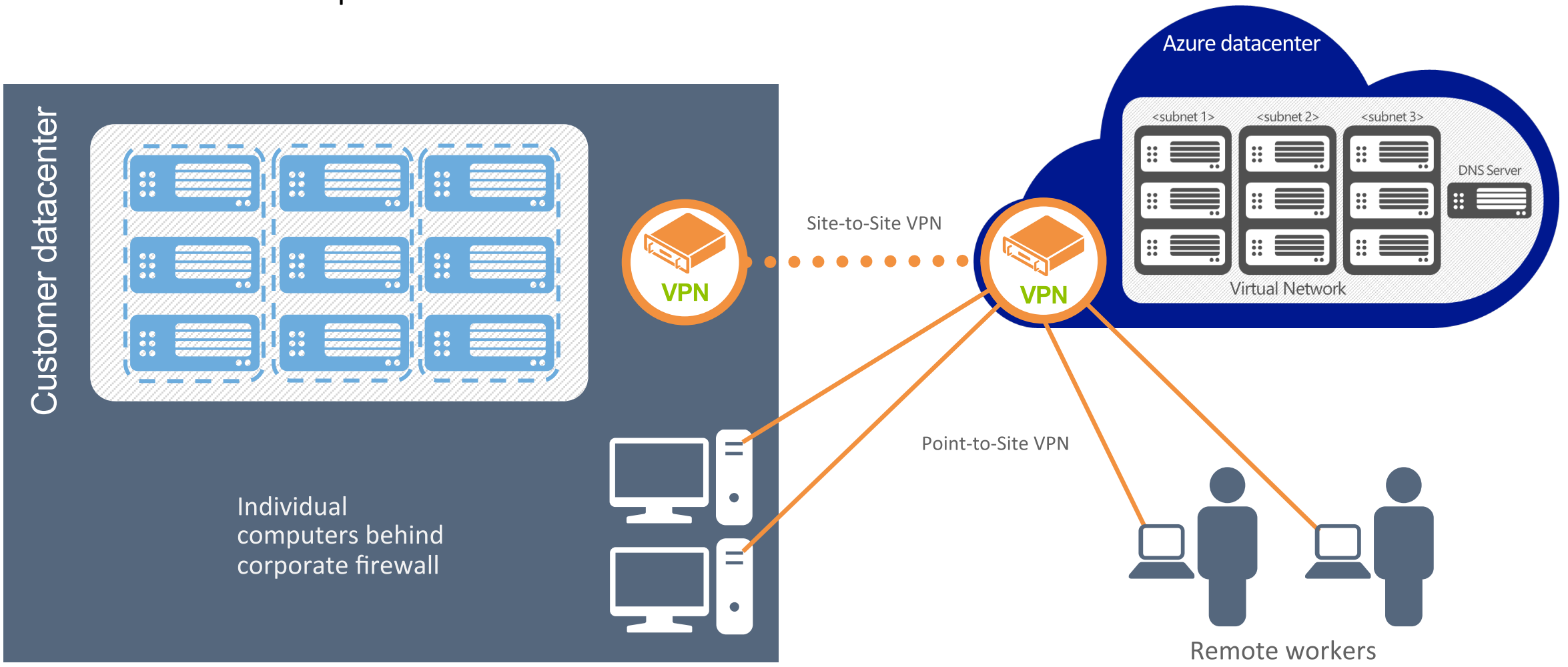
Example VNFs

- VPN
- Express Route
- Application Gateway

VPN – Hybrid Cloud over the Internet

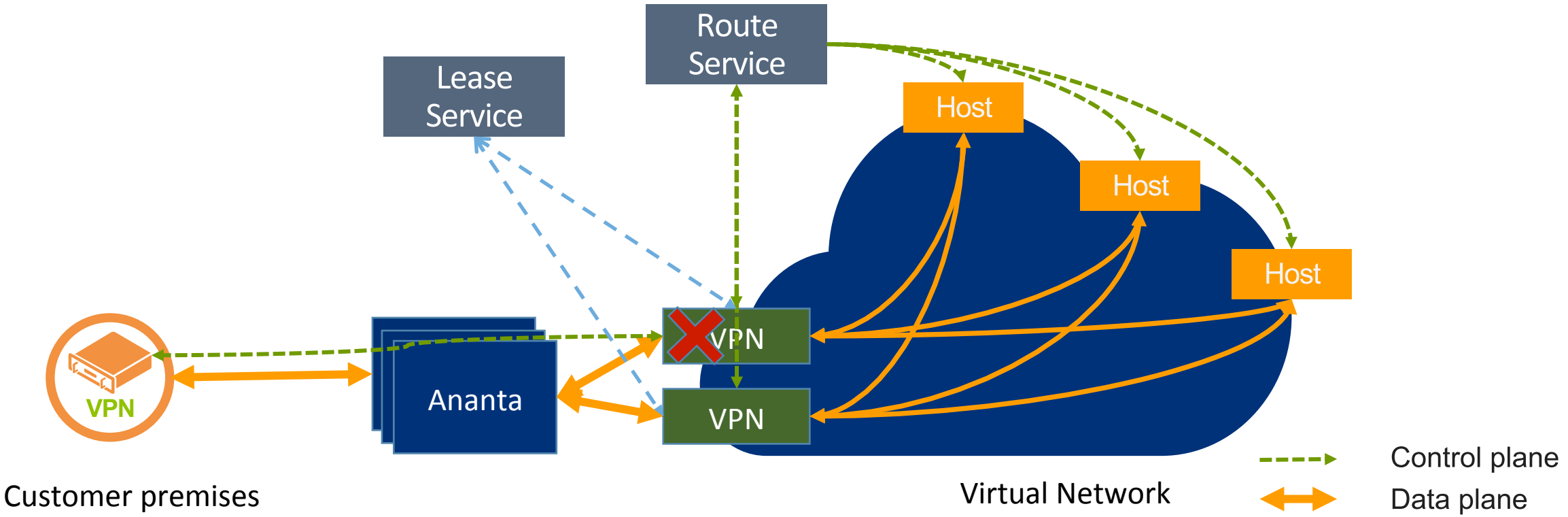
Securely connect to Virtual Network over the Internet

Traverses firewalls and proxies



IPSEC VPN High Availability

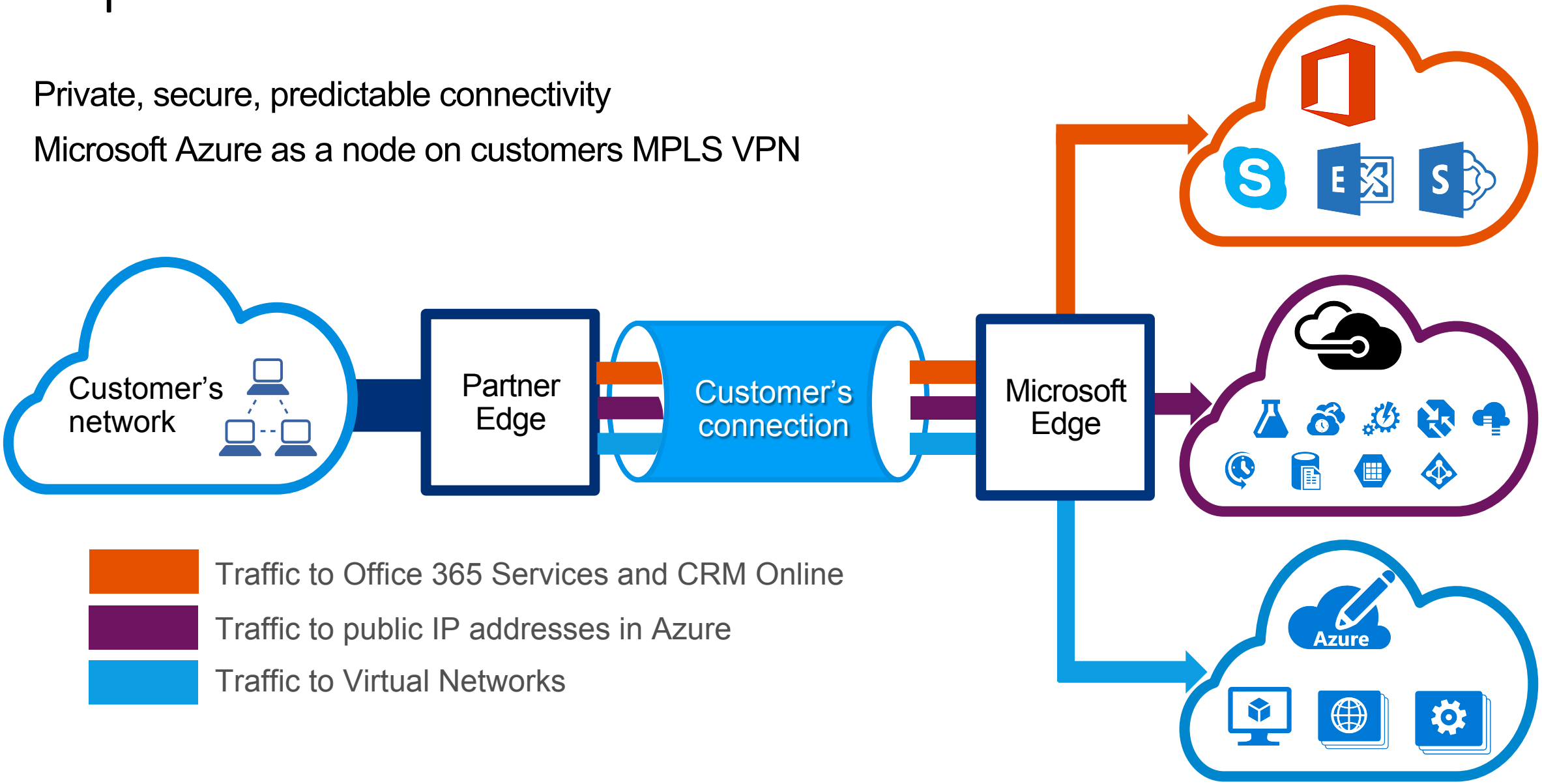
- Active-passive instances, deployed in a virtual network
- Leader election picks active instance
- LB Probe to route traffic to the active instance



Express Route – Software Defined WAN

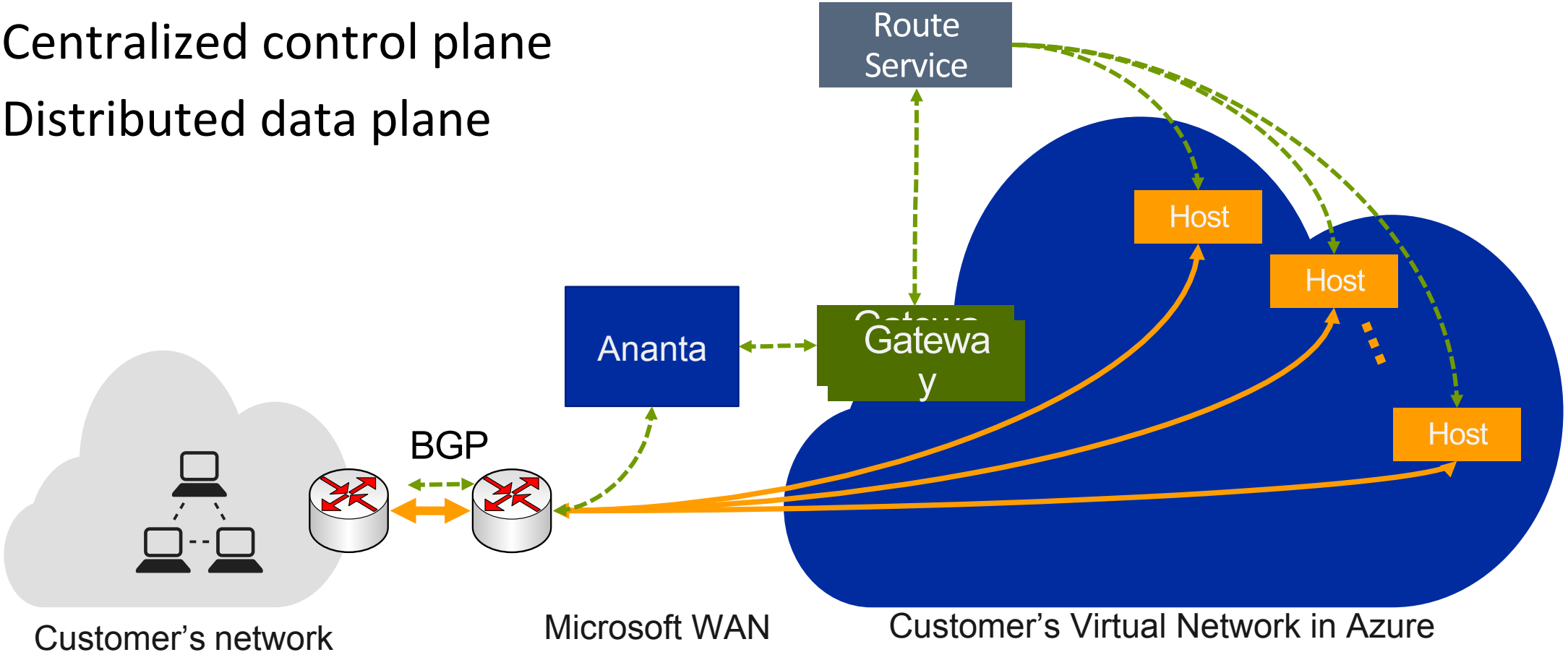
Private, secure, predictable connectivity

Microsoft Azure as a node on customers MPLS VPN



Data Center-Scale Distributed Router

- Active-active instances
- Centralized control plane
- Distributed data plane



Microsoft Cloud Network

ExpressRoute Partners

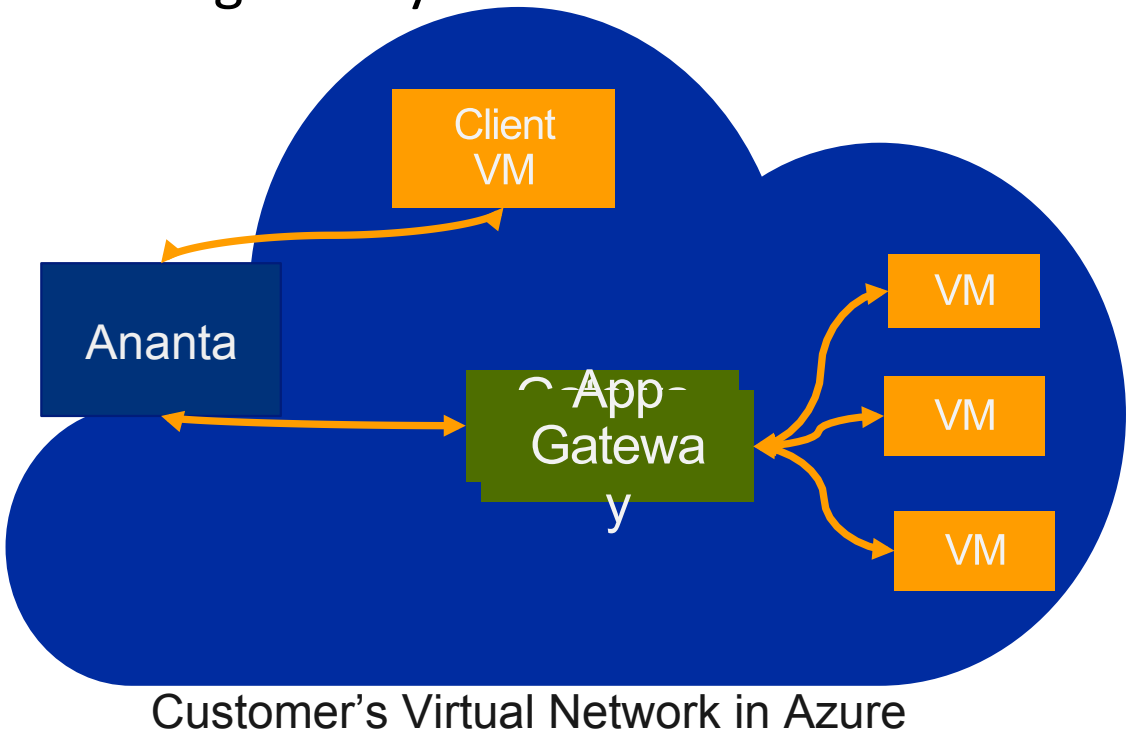
EQUINIX · TelecityGroup (where content meets connectivity) · Level(3) · BT · AT&T · verizon · OPTUS yes · colt · intercloud systems · Aryaka · zayo · orange · PCCW 電訊盈科 · dimension data · NEXTDC · NTT Communications · SingTel · inxn · IIJ Internet Initiative Japan · TATA COMMUNICATIONS · xo communications · Nimbo · Inc. 500

Legend

- Datacenter
- Edge Node
- Internet Exchange
- Terrestrial Network
- Subsea Network

Azure Application Gateway

- Active-active instances
- SSL offload, session affinity, URL rewrite, visibility, …
- NFV Infrastructure synchronizes configuration, monitors performance and scales gateways



Azure NFV Ecosystem

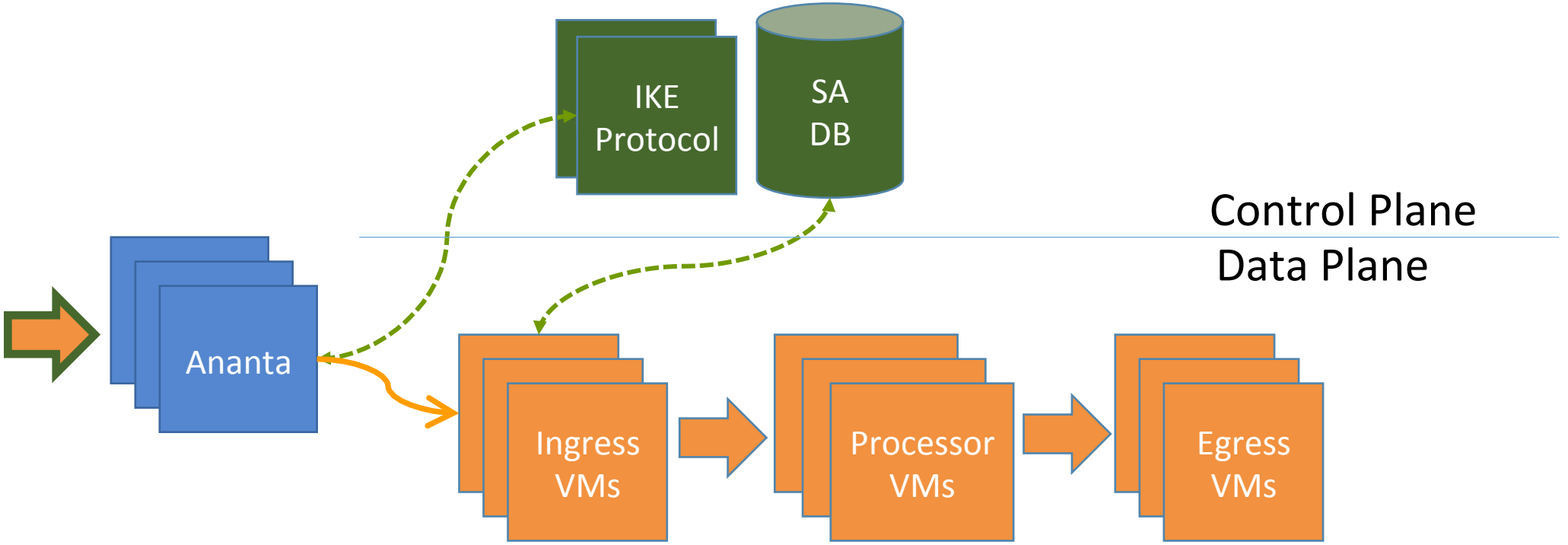
- The same LB and HA features are available to third-parties as well
- Third-party VNFs available now:
  - F5, Citrix, Barracuda, Cisco, A10, Riverbed, Brocade, …

Lessons learnt

- Active-passive does not work
  - Maintenance windows too painful for customers
  - Example: VOIP server in the cloud
  - **Conclusion**: Decompose and scale-out each VNF

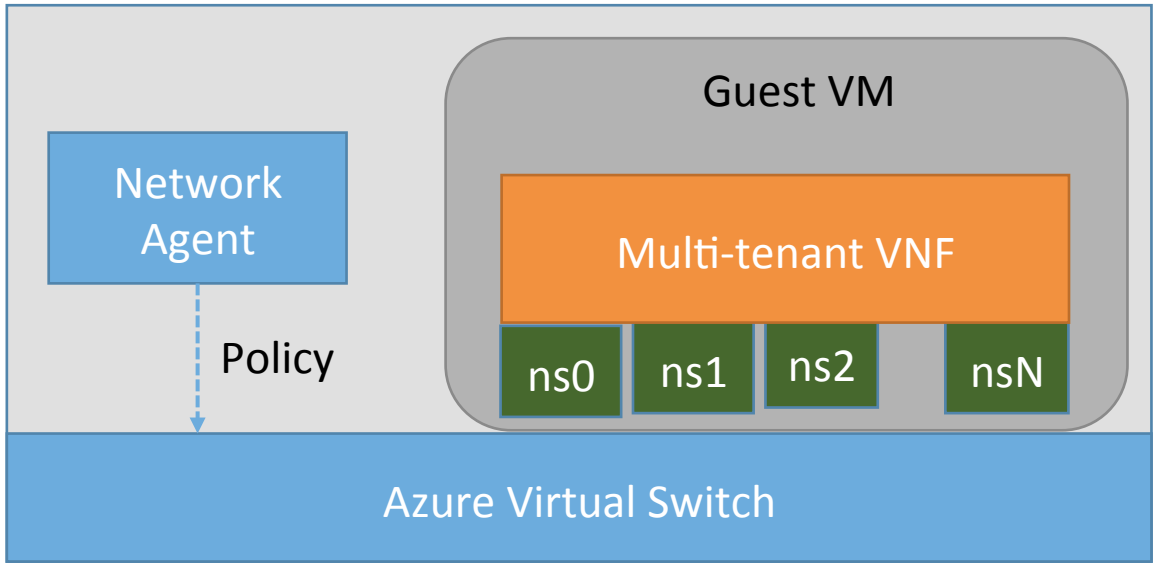
- Single-tenant services are too expensive
  - Multi-tenancy in VMs does not provide sufficient performance isolation
  - **Conclusion**: network needs to provide isolation support

Applying SDN to IPSEC



Multi-tenant virtual networking

- Leverage networking namespaces (Windows compartments)
- Azure Virtual Switch delivers namespace metadata to the VM
- Virtual Switch supports *cut-through forwarding* for high-performance



Call to action for P4

- In-network performance isolation
  - By the time packets hit a server, it's already too late!
- Support stateful packet processing functions
  - Most flows in the data center create state at the virtual network edge
  - Support functions that create per-flow state from data path, e.g., LB, FW
- Support on different hardware and form factors
  - ASIC, NPU, FPGA; NIC, Router
  - Currently, most networking is done on x86 due to its programmability
  - Enable best price/performance for specific needs
- Visibility and diagnostics key to at scale network operation
  - EverFlow, PingMesh examples of visibility tools

Appendix