Inside a Datacenter
Hardware and Software Architecture

Wei Bai

Microsoft Research Redmond

Guest Lecture, Computer Networks @ Xiamen University

Disclaimer

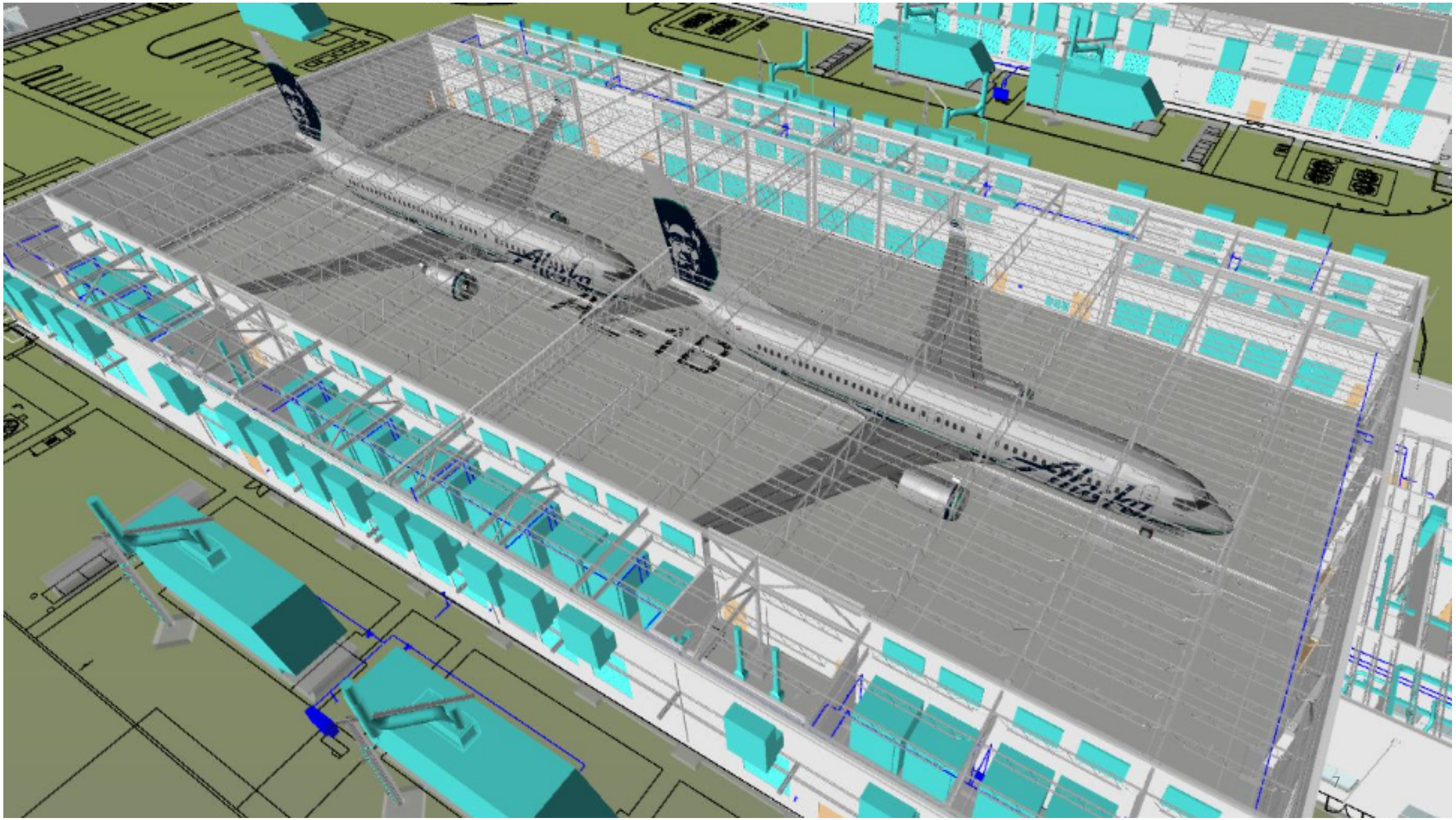
- The opinions in this talk are my own.
- Datacenter architecture is a very large topic. In this talk, I will focus on **networking** and briefly discuss storage and compute.
- I will use many Microsoft products and technologies as examples in this talk. But general concepts should be able to be applied to other companies as well.
- Material with thanks to Mark Russinovich, Albert Greenberg, Dave Maltz, Marcus Fontoura, Daniel Firestone, Costin Raiciu and others.

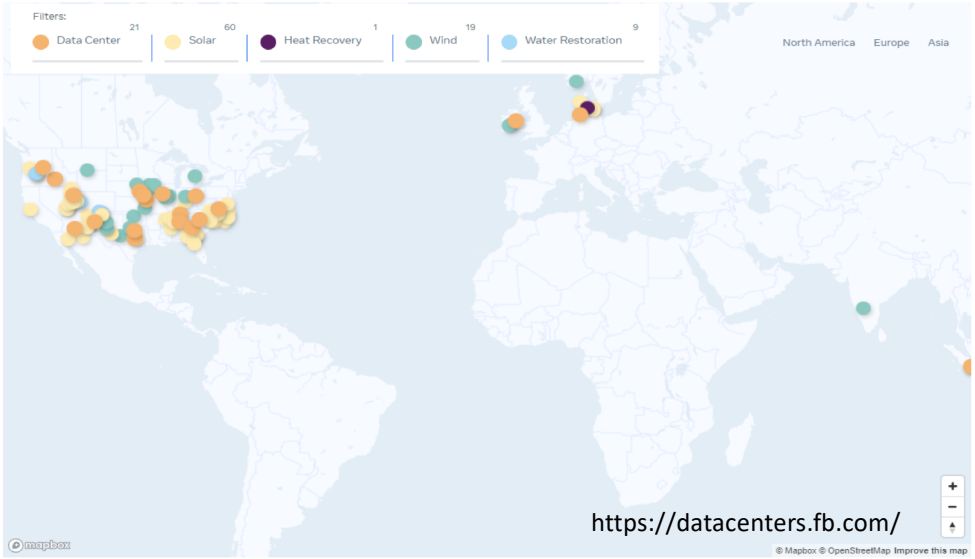
What is a datacenter?


A datacenter is a building, or a dedicated space used to house computer systems and associated components, such as telecommunications and storage systems.

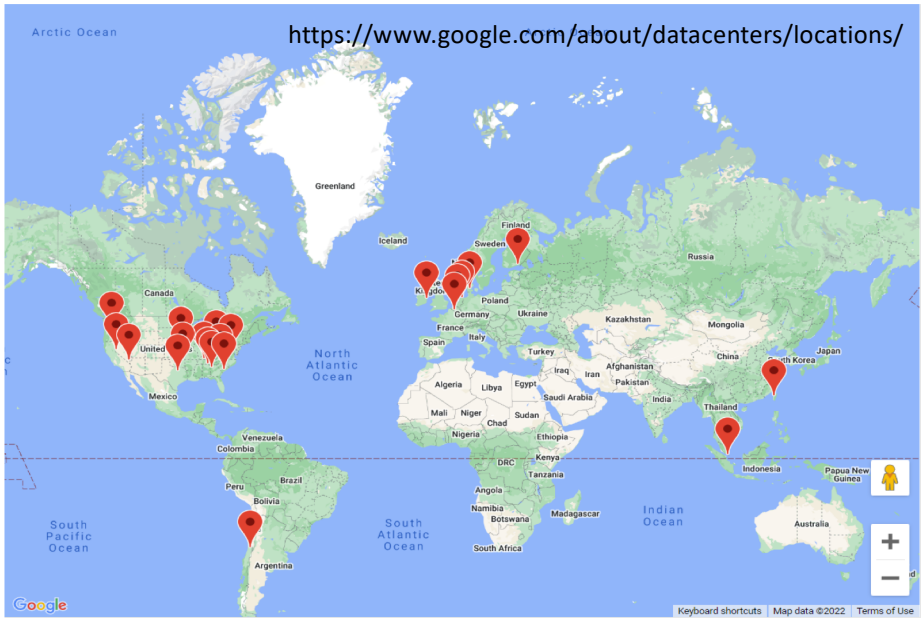
Microsoft @ Amsterdam

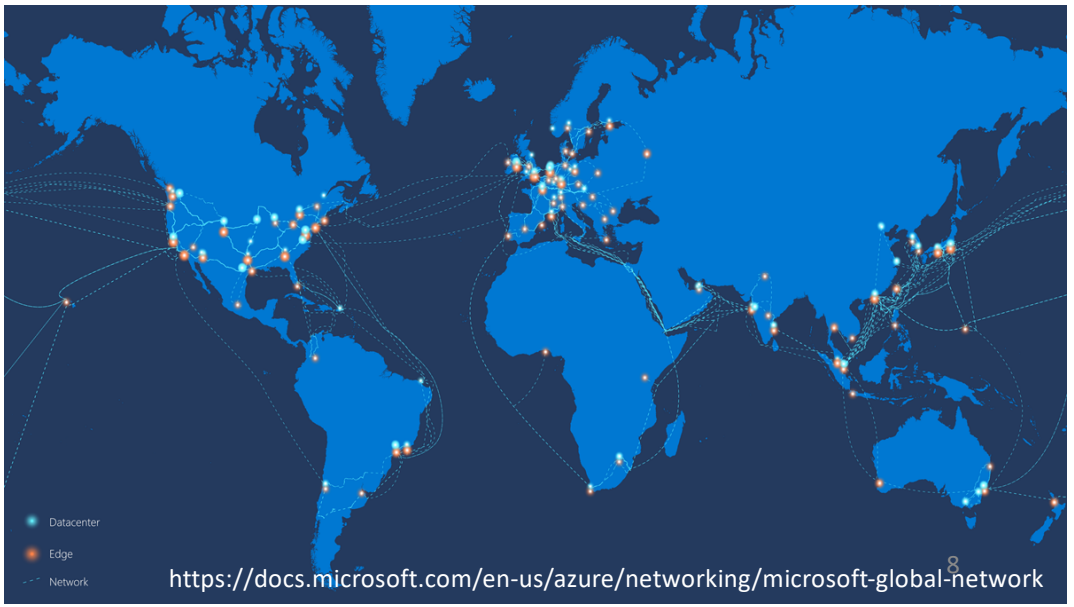

Microsoft @ Quincy


Alaska
Alaska


Filters:
Data Center 21
Solar 60
Heat Recovery 1
Wind 19
Water Restoration 9
North America
Europe
Asia
https://datacenters.fb.com/
mapbox
© Mapbox © OpenStreetMap Improve this map


Regions
Coming soon
https://aws.amazon.com/about-aws/global-infrastructure/
17 Local Zones
28 Wavelength Zones
for ultralow latency applications
245 Countries and Territories Served
115 Direct Connect Locations

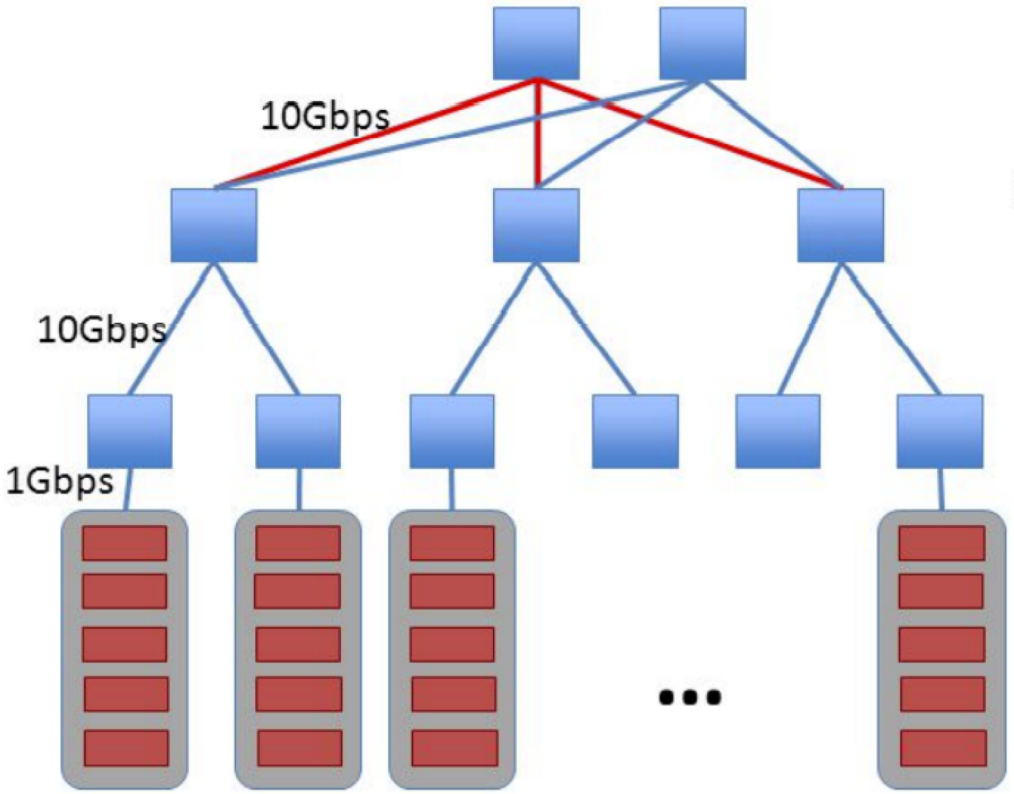

https://www.google.com/about/datacenters/locations/
Arctic Ocean
Greenland
Iceland
Finland
Sweden
Russia
Canada
Poland
Germany
Ukraine
Kazakhstan
Mongolia
France
Spain
Italy
Turkey
China
Japan
United States
Mexico
North Atlantic Ocean
Iraq
Iran
Afghanistan
Pakistan
Algeria
Libya
Egypt
Saudi Arabia
India
Thailand
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Venezuela
Colombia
Brazil
Peru
Bolivia
DRC
Kenya
Tanzania
Angola
Namibia
Botswana
Madagascar
South Africa
Indonesia
Papua New Guinea
Australia
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Argentina
Google
Keyboard shortcuts
Map data ©2022
Terms of Use


Datacenter
Edge
Network
https://docs.microsoft.com/en-us/azure/networking/microsoft-global-network

Physical Networking

- How to connect servers within a datacenters?
- How to connect different datacenters and Internet?

Traditional Data Center Topology (2008)



Core Switches

Aggregation Switches

Top of Rack Switches

Racks of Servers

Problems in Traditional Solutions

- Lack robustness
  - Aggregation switch failures wipe out several racks
- Poor performance

  Oversubscription = max_host_throughput / worst_case_throughput
  - Typical oversubscription ratio 4:1, 8:1, 10:1
- Expensive
  - 7K for 48-port Gigabit switch
  - 700K for 128-port 10Gigabit switch

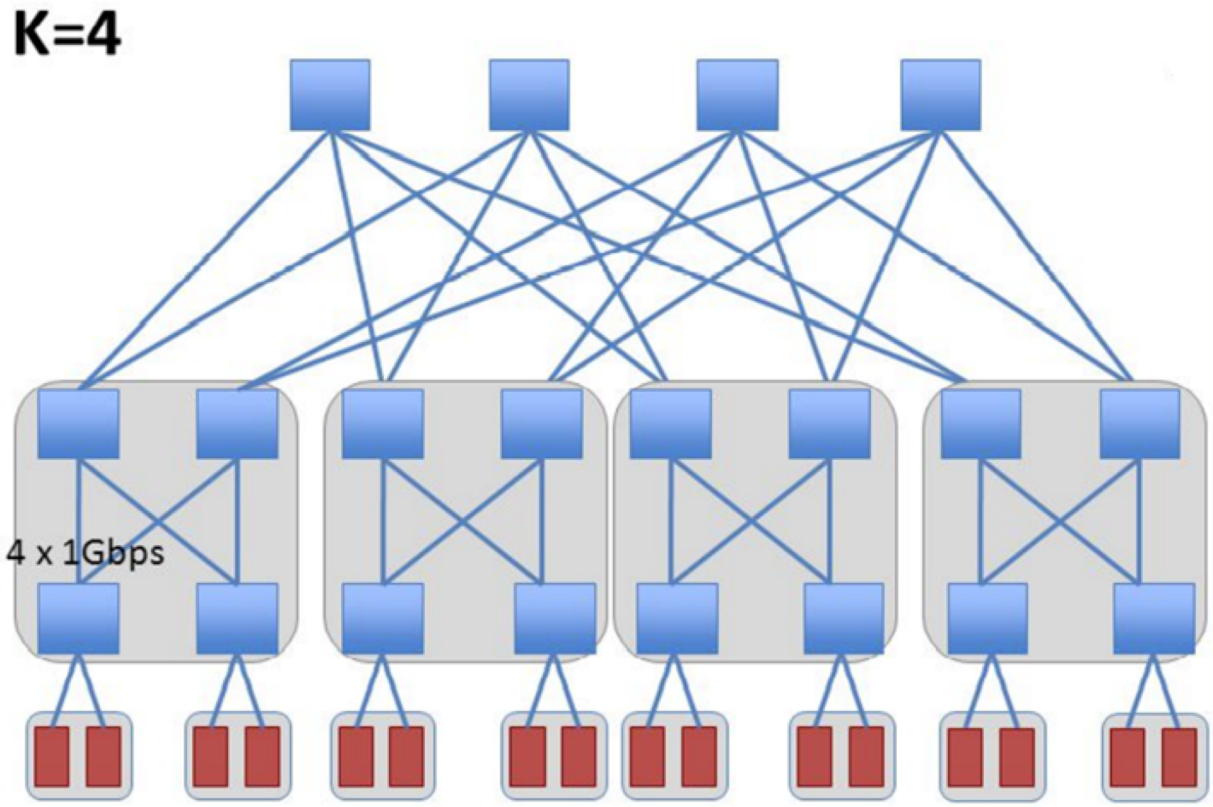
Ideal Datacenter Network Topology

- Offer **full-bisection bandwidth**
  - Over-subscription ratio of 1:1
  - Worst case: every host can talk to every host at line rate!
- Is **fault tolerant**
- Is **cheap**

The Fat Tree [Al-Fares et al, SIGCOMM 2008]

- Inspired from the telephone networks of the 50's – Clos networks
- Use cheap, commodity switches – all switches are the same
- Lots of redundancy
- Single parameter to describe the topology:
  - **K** – the number of ports in a switch

Fat Tree Topology



Core Switches

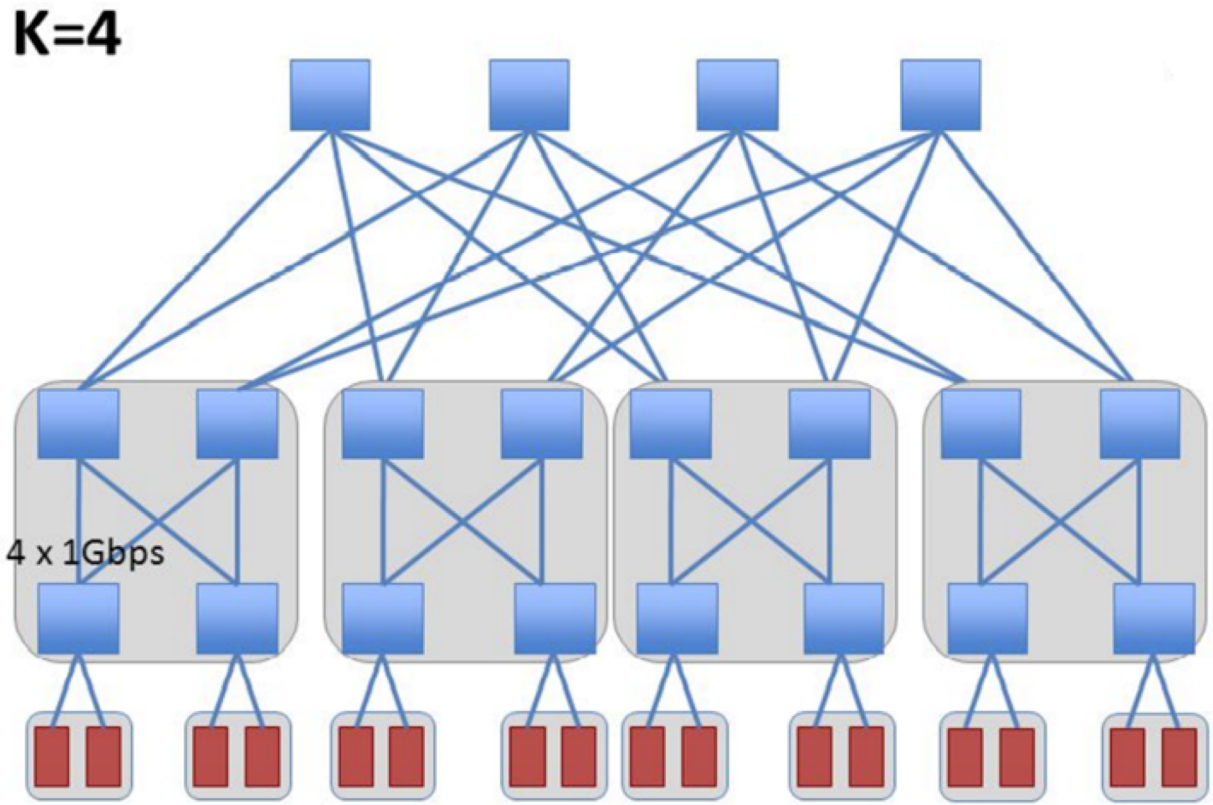
Aggregation Switches

Edge Switches

Servers

K^3/4

Fat Tree Topology



K^3/4

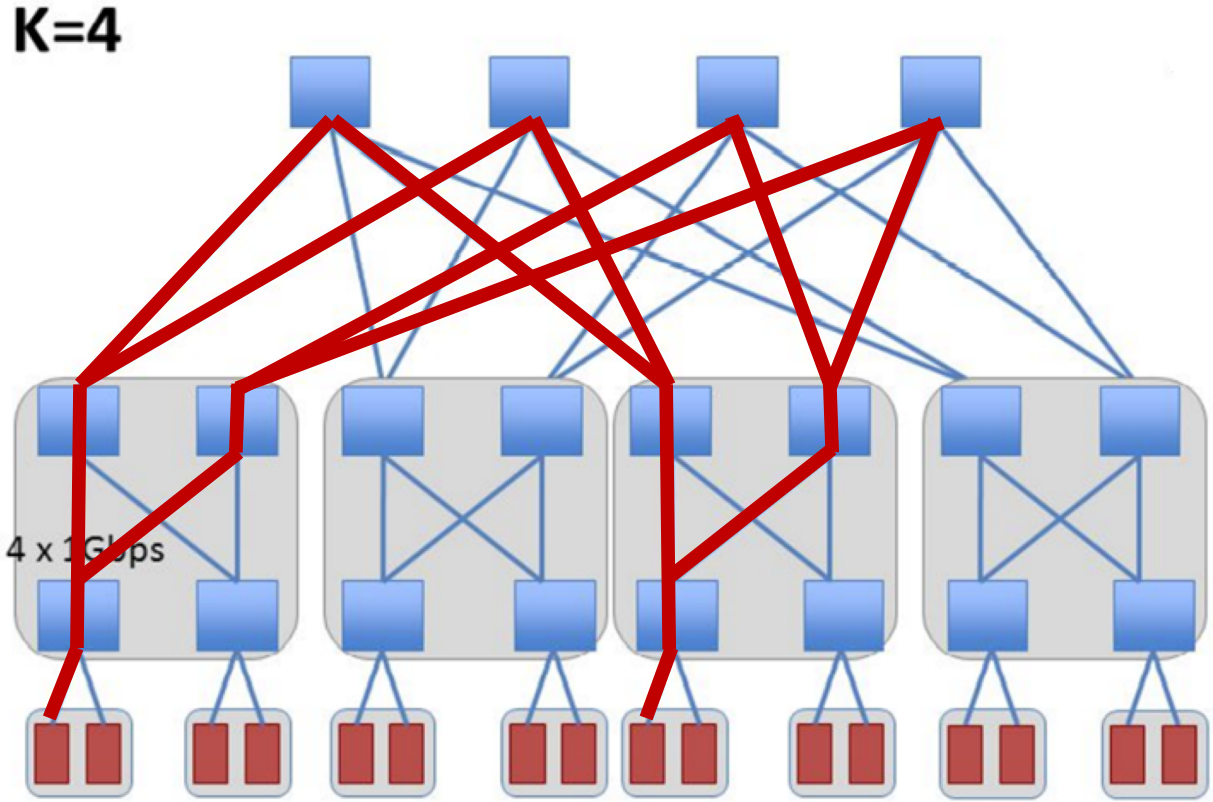
Core Switches $\frac{K^2}{4}$

Aggregation Switches $\frac{K^2}{2}$

Edge Switches $\frac{K^2}{2}$

Servers $\frac{K^3}{4}$

The Fat Tree Topology has $\frac{K^2}{4}$ paths between any two endpoints



Core Switches $\frac{K^2}{4}$

Aggregation Switches $\frac{K^2}{2}$

Edge Switches $\frac{K^2}{2}$

Servers $\frac{K^3}{4}$

Routing: how to use these paths

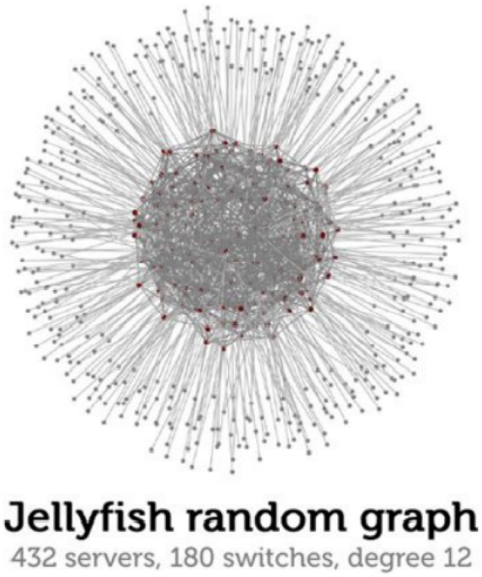
- **Layer 2**: Spanning Tree Protocol (STP)
  - STP does not offer multiple paths
  - What about L2 broadcasts? (e.g., ARP)

- **Layer 3**: **Equal Cost Multipath Routing** (ECMP)
  - When there are multiple shortest paths, pick one "randomly"
  - Hash packet header to choose a path
  - All packets of the same flow go on the same path

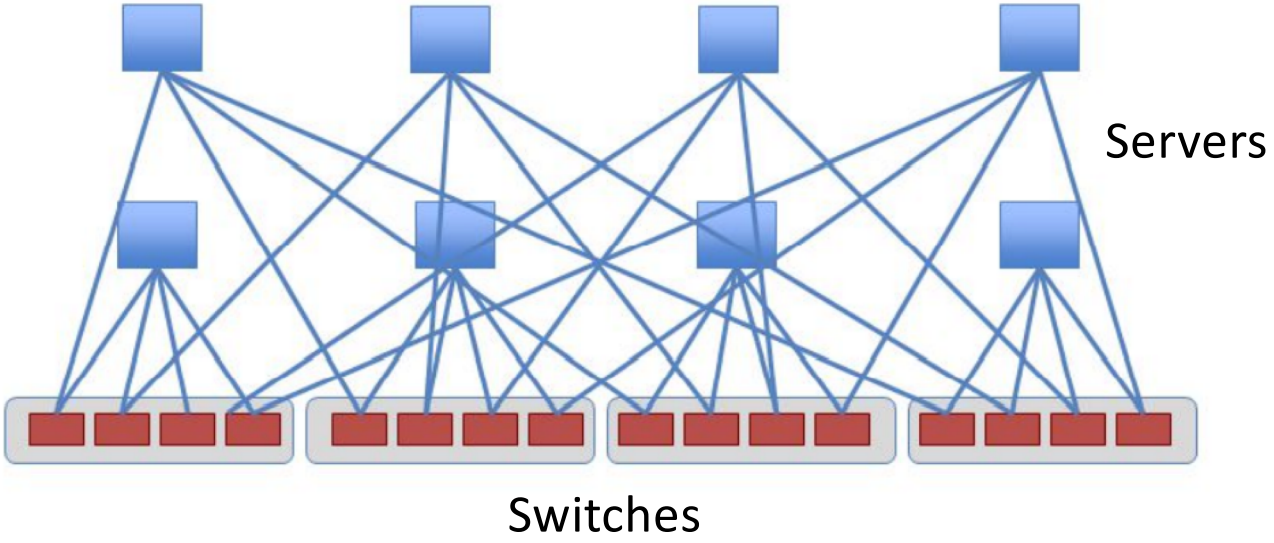
Why not use per-packet ECMP?

Other Topologies

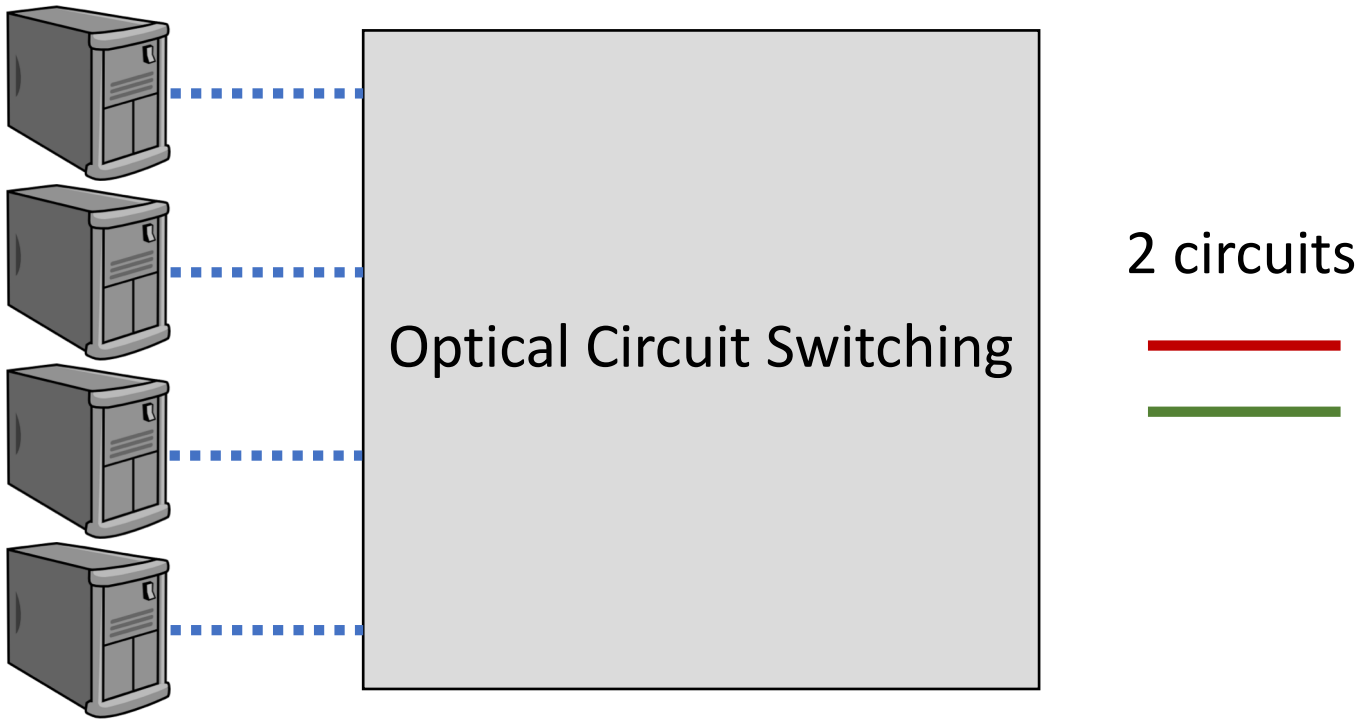
Jellyfish [Singla et al, NSDI 2012]



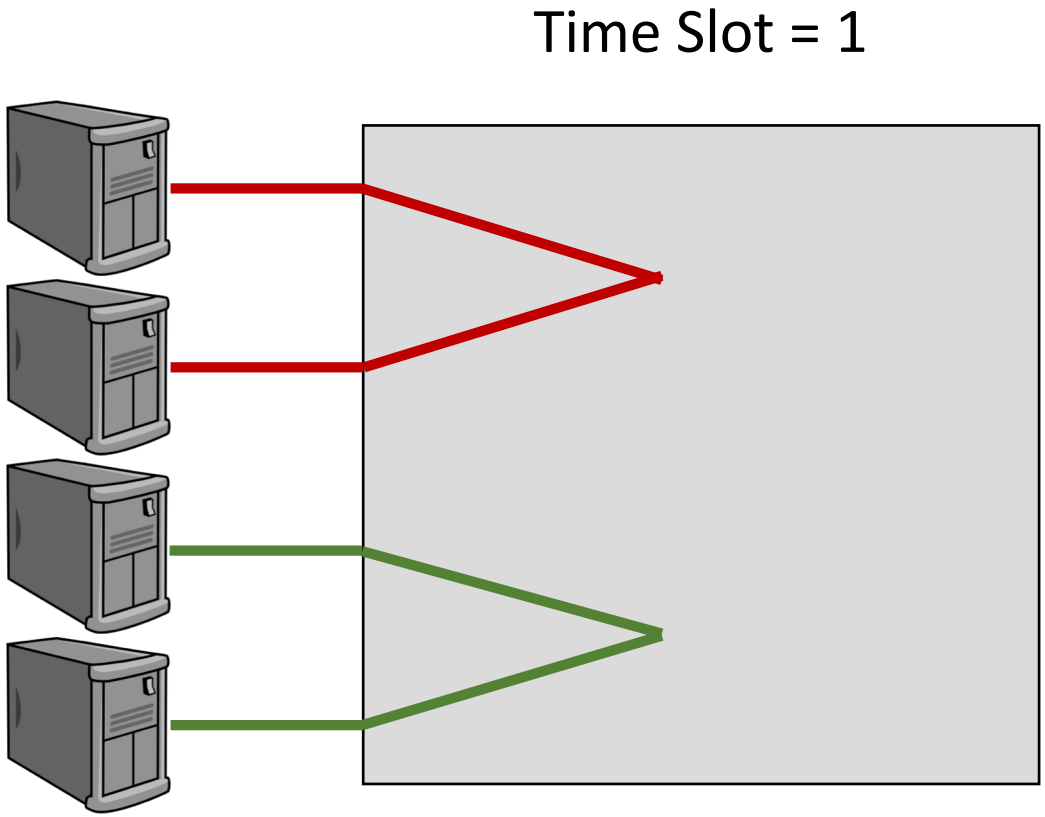
BCube [Guo et al, SIGCOMM 2009]



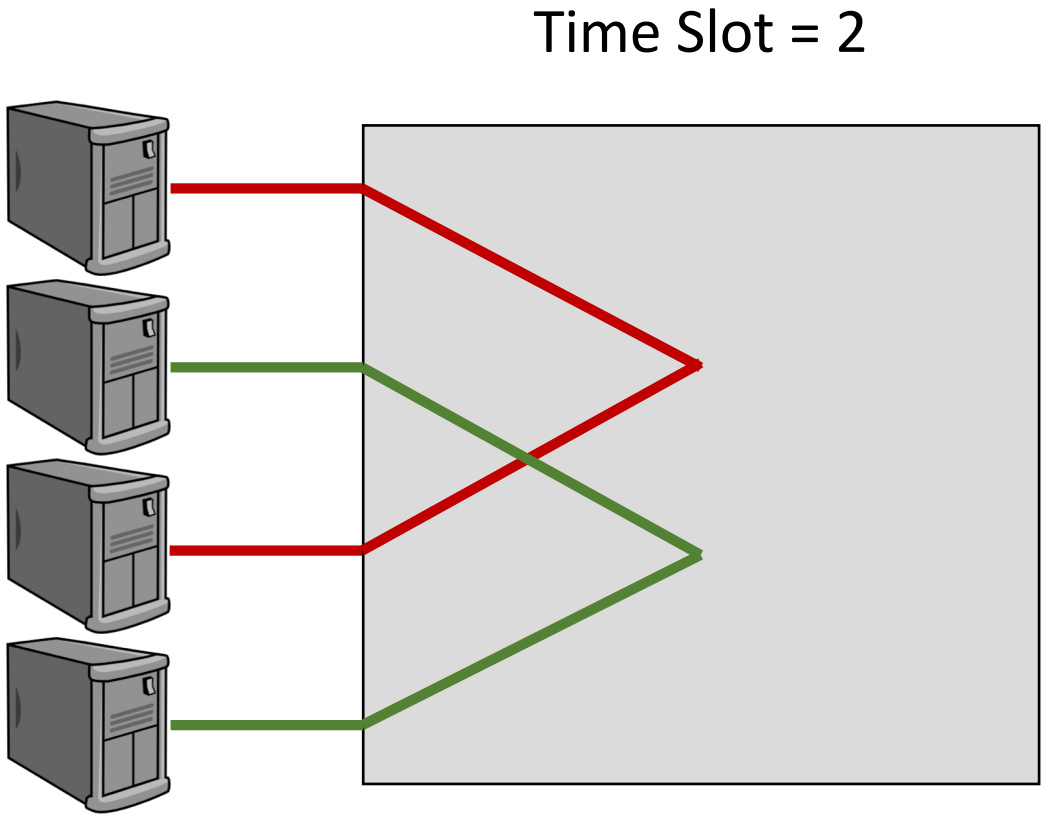
Reconfigurable Topologies



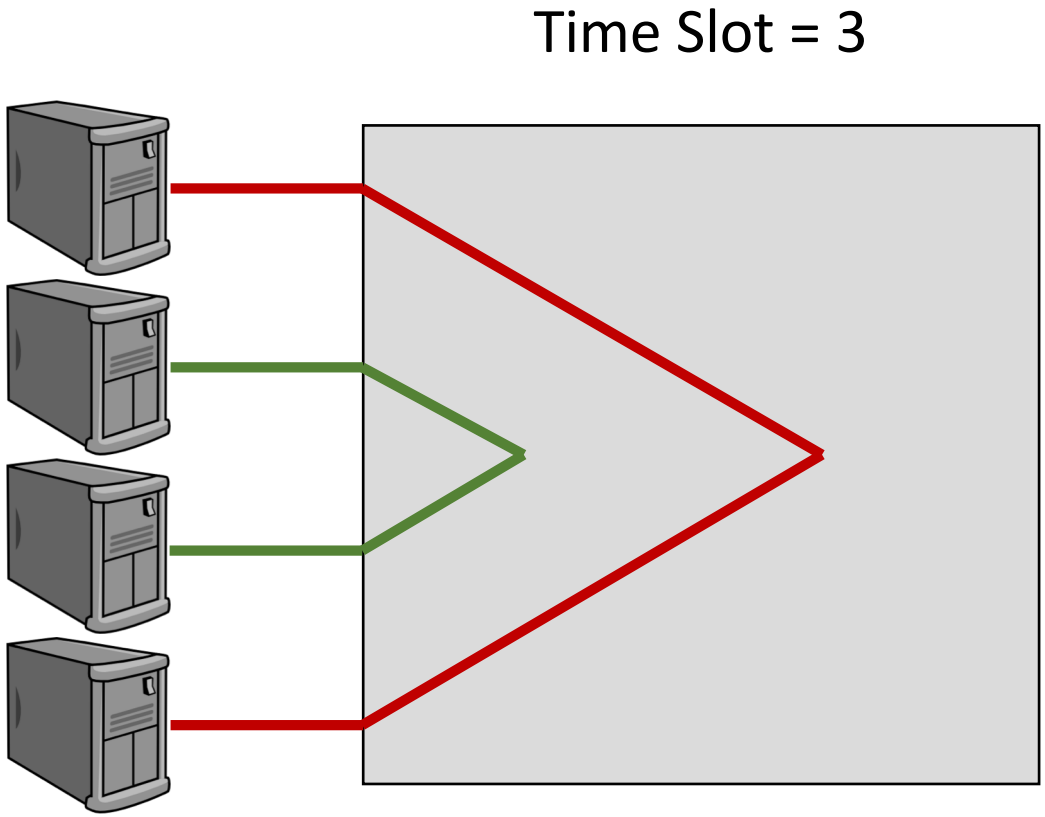
Reconfigurable Topologies



Reconfigurable Topologies

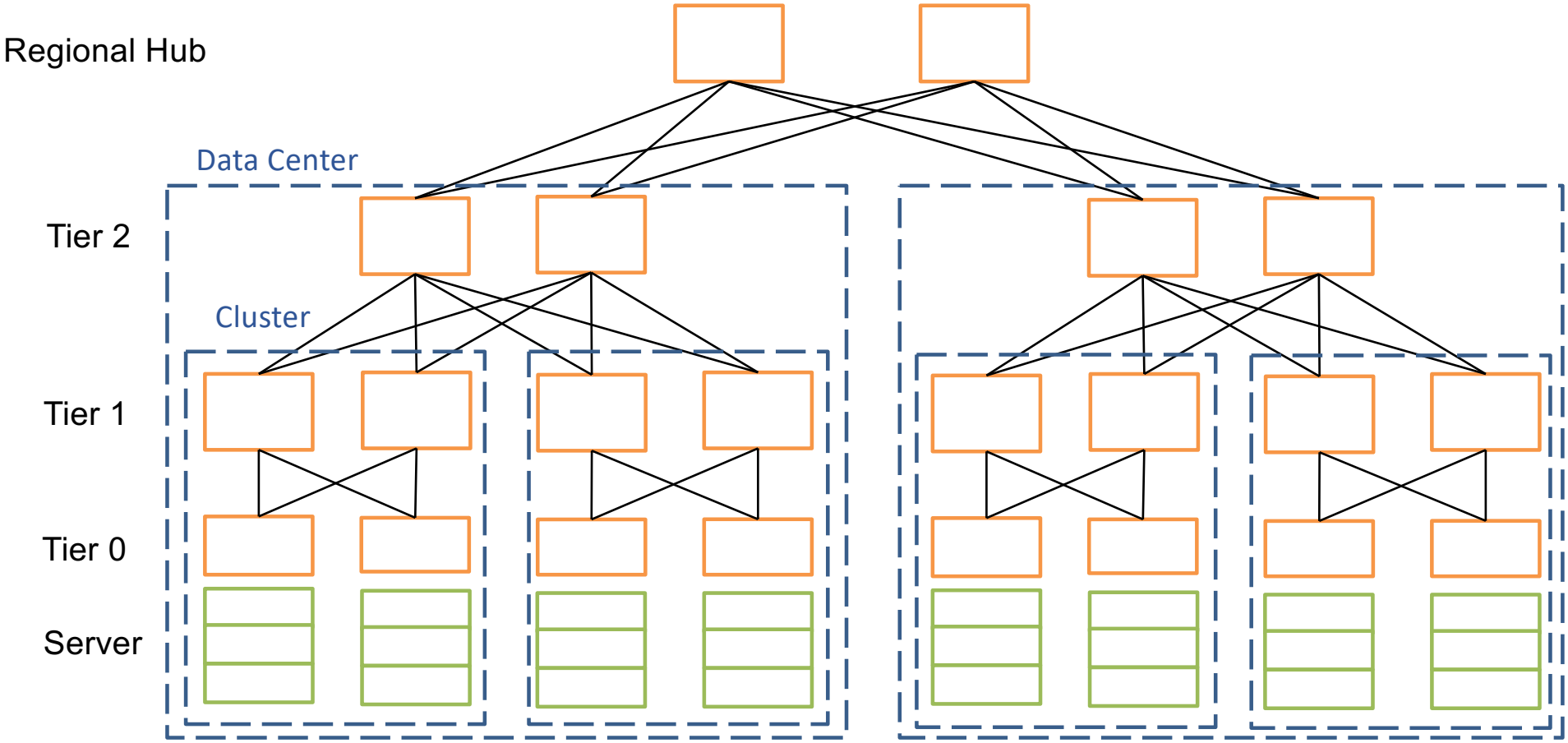


Reconfigurable Topologies



What topologies are used in practice?

Network Architecture of an Azure Region



Regional Spine Architecture

Customers want the Cloud to be infinitely scalable – how to create that illusion?

Introducing: Regions

Contiguous geographical area up to roughly 100km in diameter **(2.0ms RTT)**

Regional network gateway

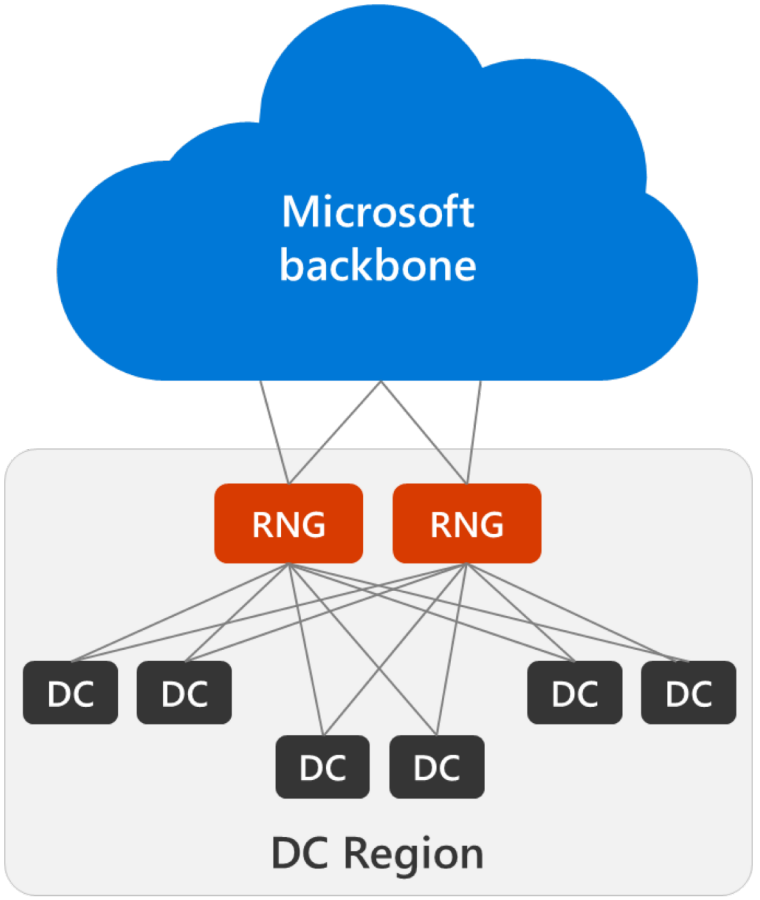
Massively parallel, hyper scale

Space and power protected

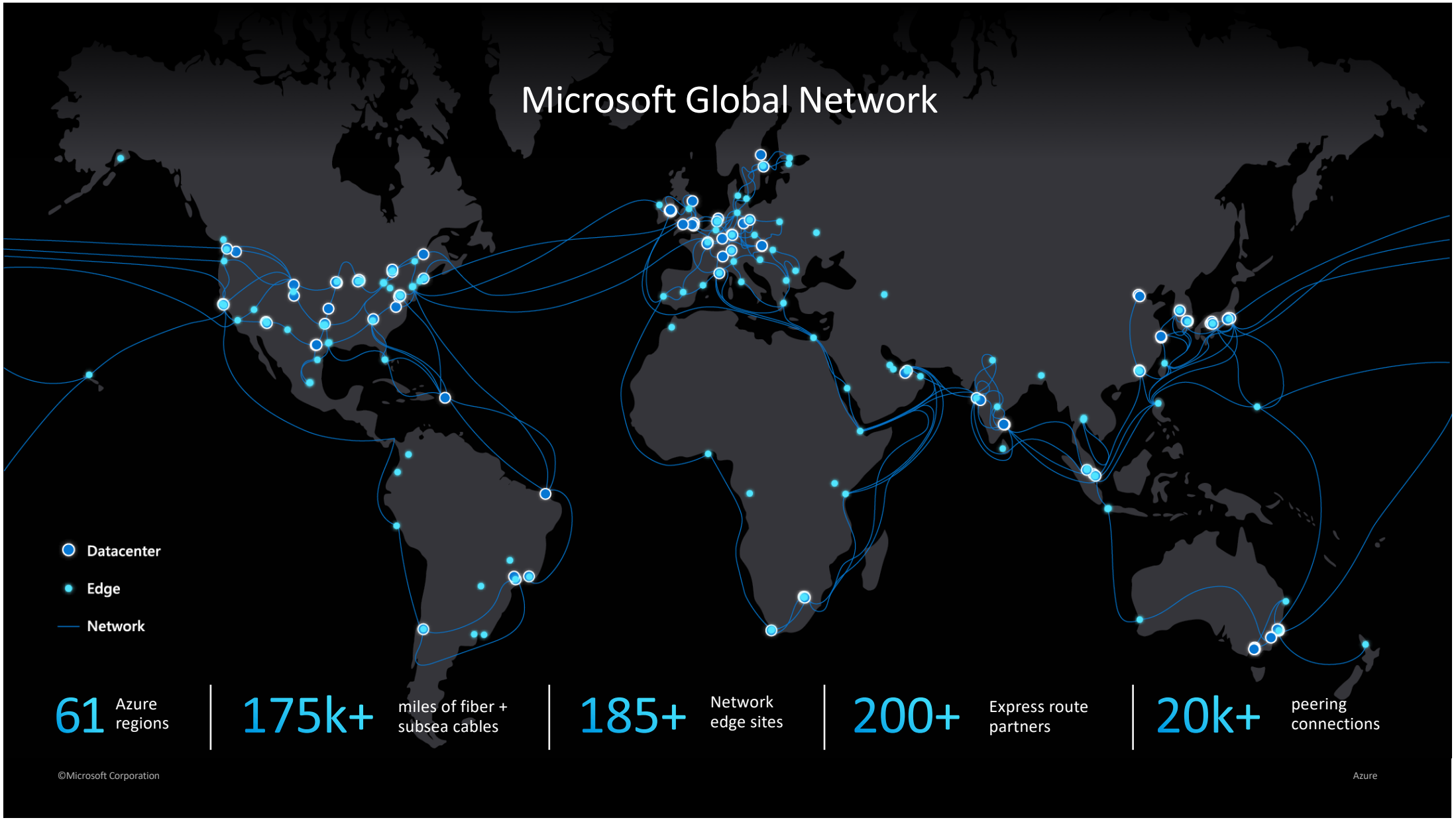
Data centers

Only contains server racks, DC network

RNGs are sized to support growing the region by adding data centers



Microsoft Global Network



61 Azure regions

175k+ miles of fiber + subsea cables

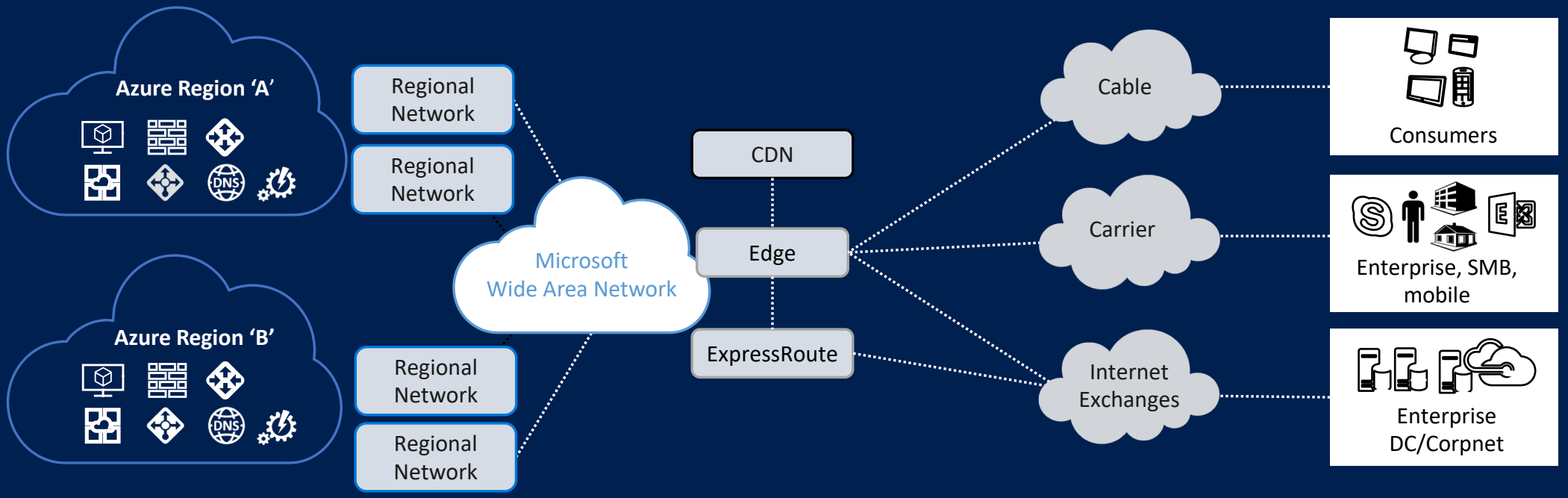
185+ Network edge sites

200+ Express route partners

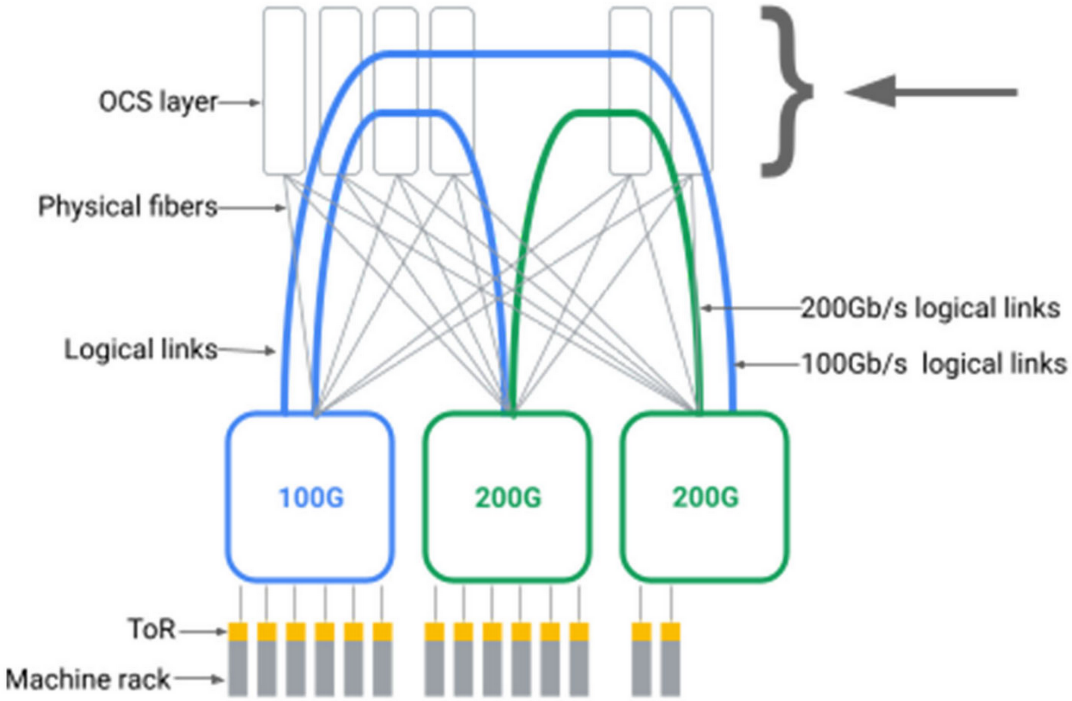
20k+ peering connections

Azure

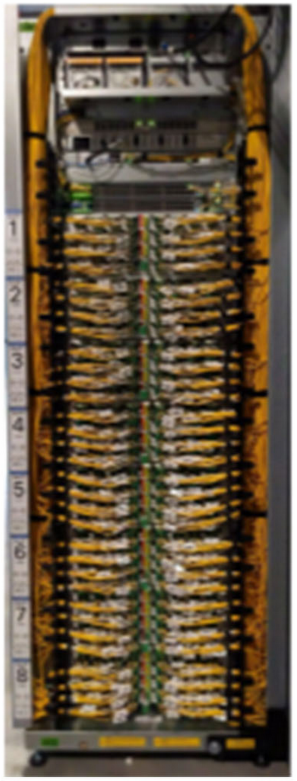
Microsoft Azure Networking



Google uses optical circuit switching (OCS) switches to **reconfigure** networks between aggregation blocks



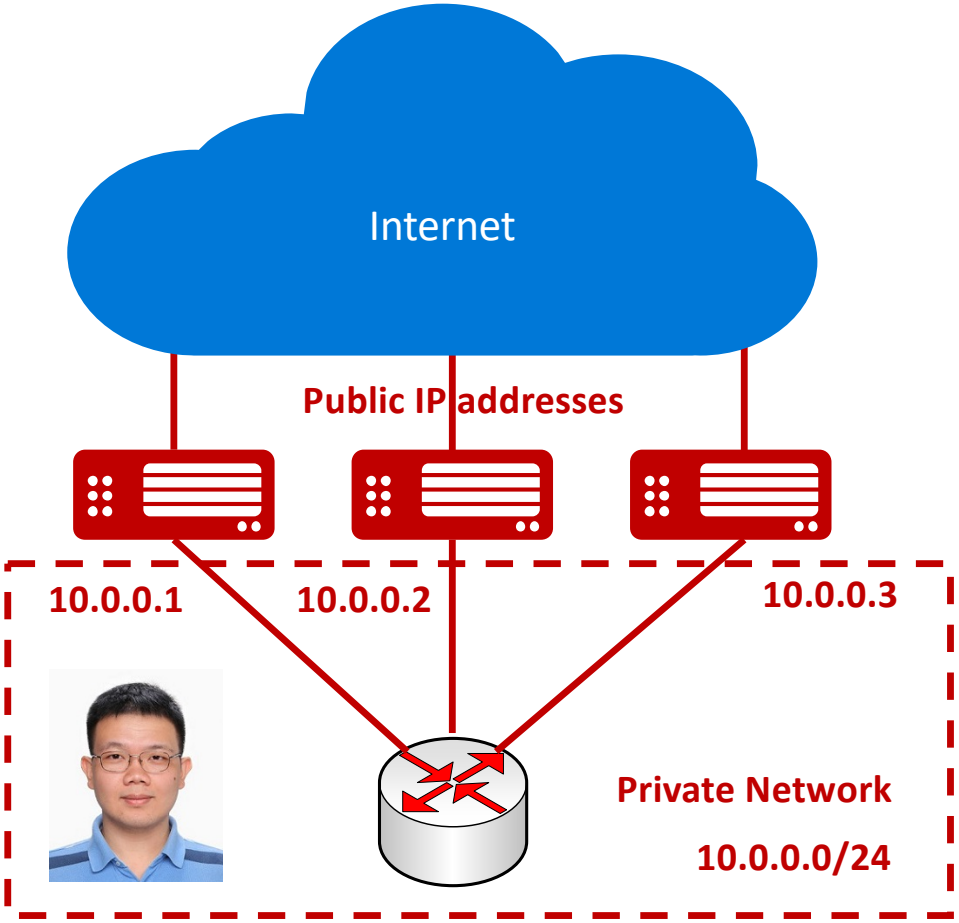
Rack with 8 OCS switches

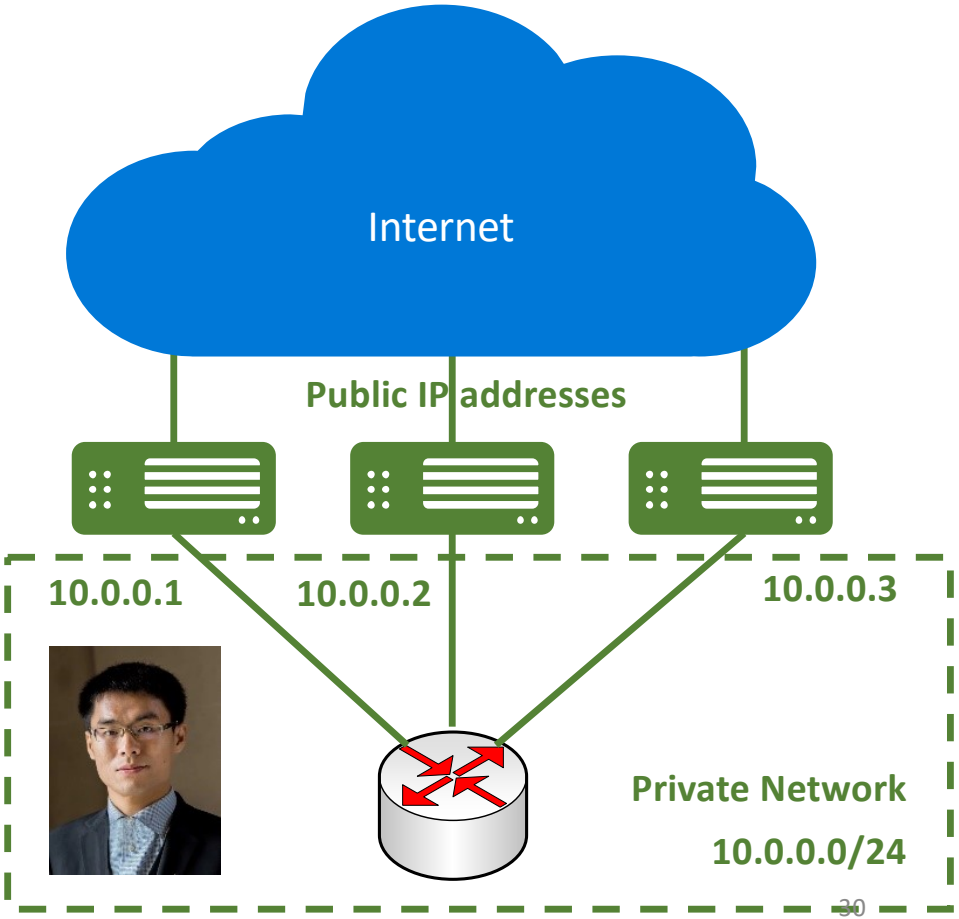


Virtual Networking

How to enable multi-tenancy on a shared network fabric?

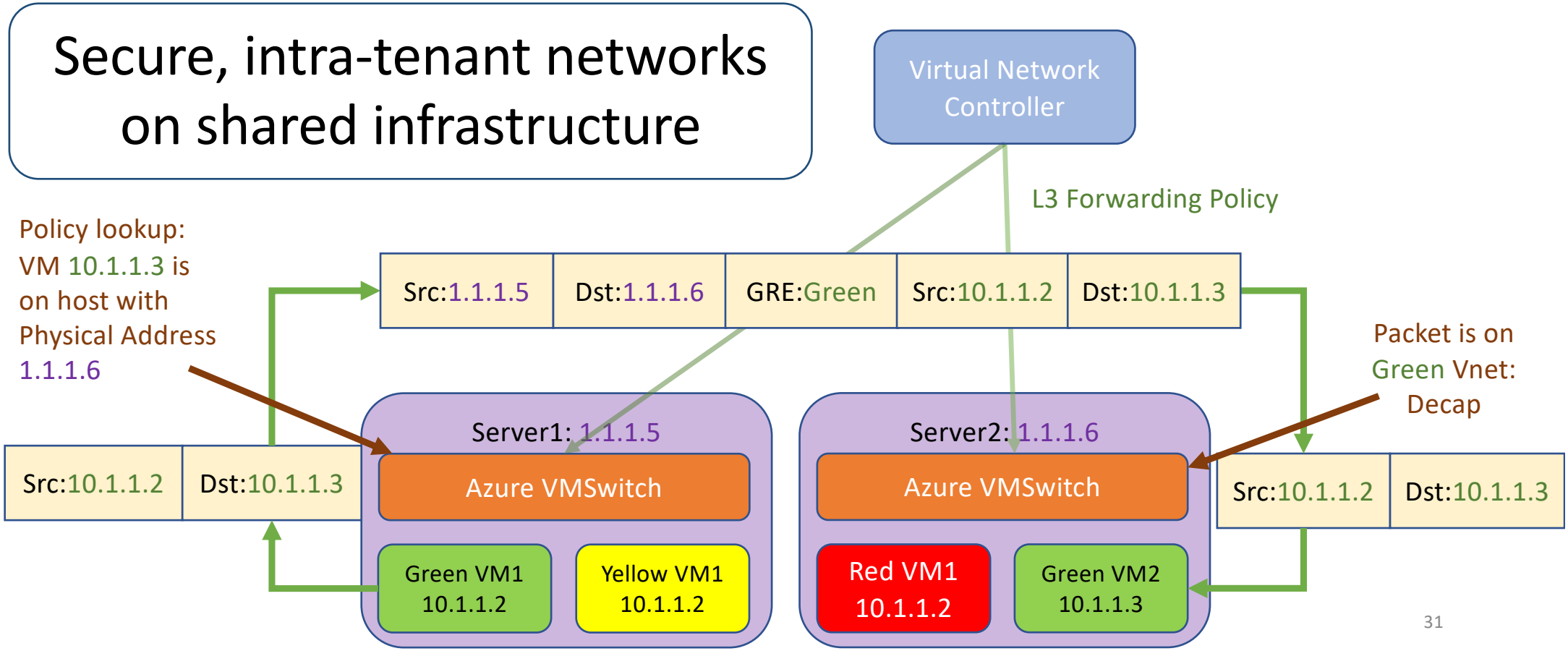
Customers only see Virtual Networks (VNets)



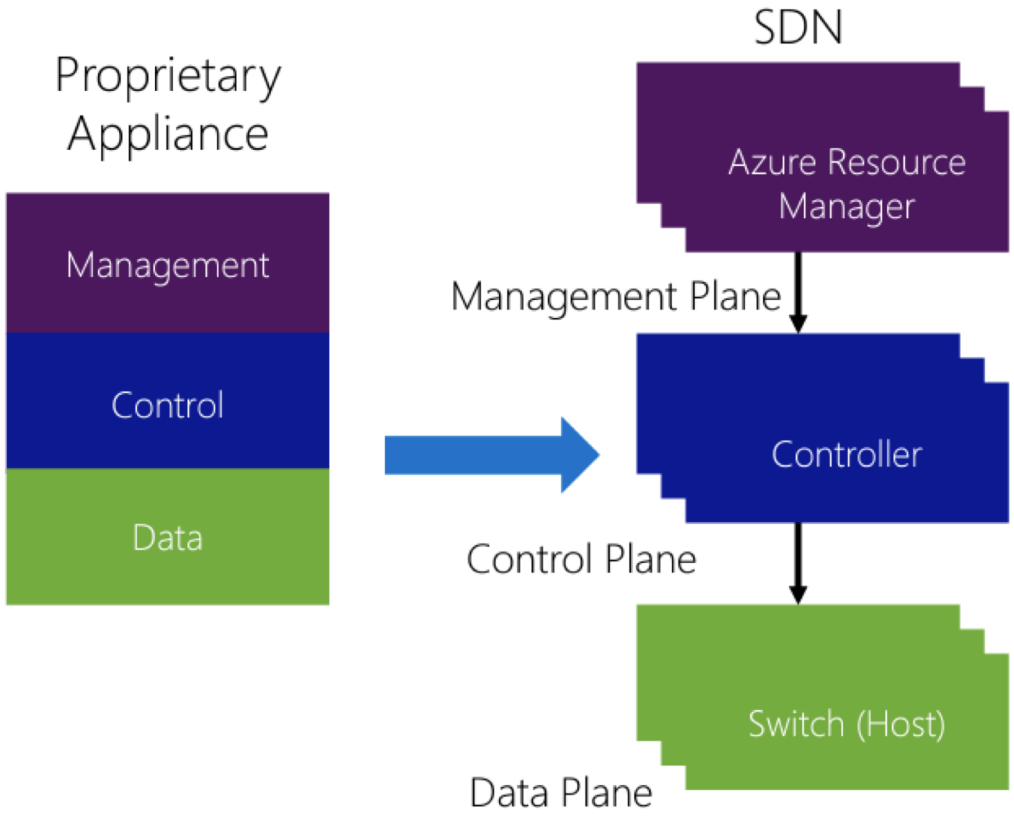


Key Mechanism: Packet Encapsulation

Secure, intra-tenant networks on shared infrastructure



Azure Software Defined Networking (SDN)

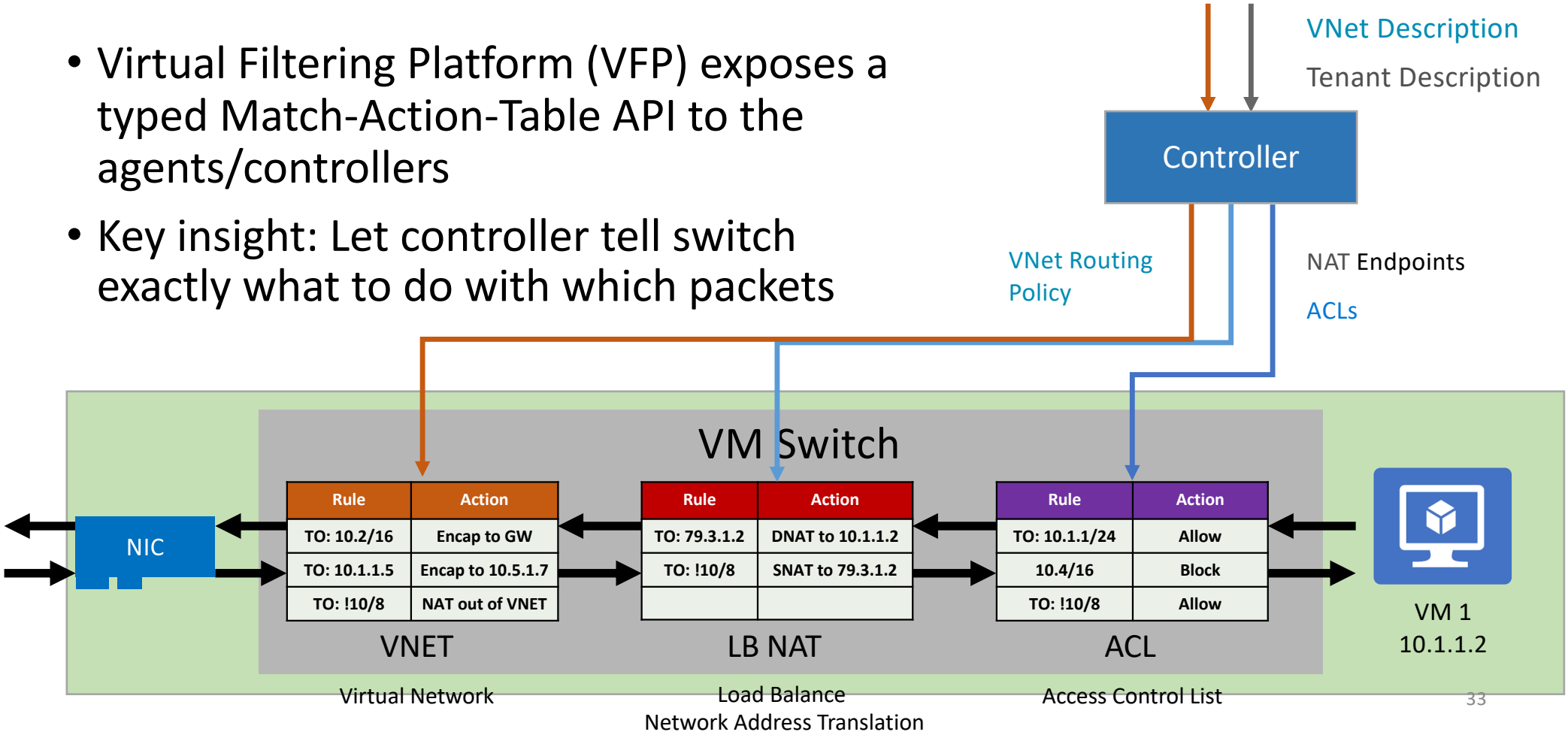


Management plane	→	Create a tenant
Control plane	→	Plumb tenant policies to VM switches
Data plane	→	VM switches apply policies to flows

Key to flexibility and scale is Host SDN

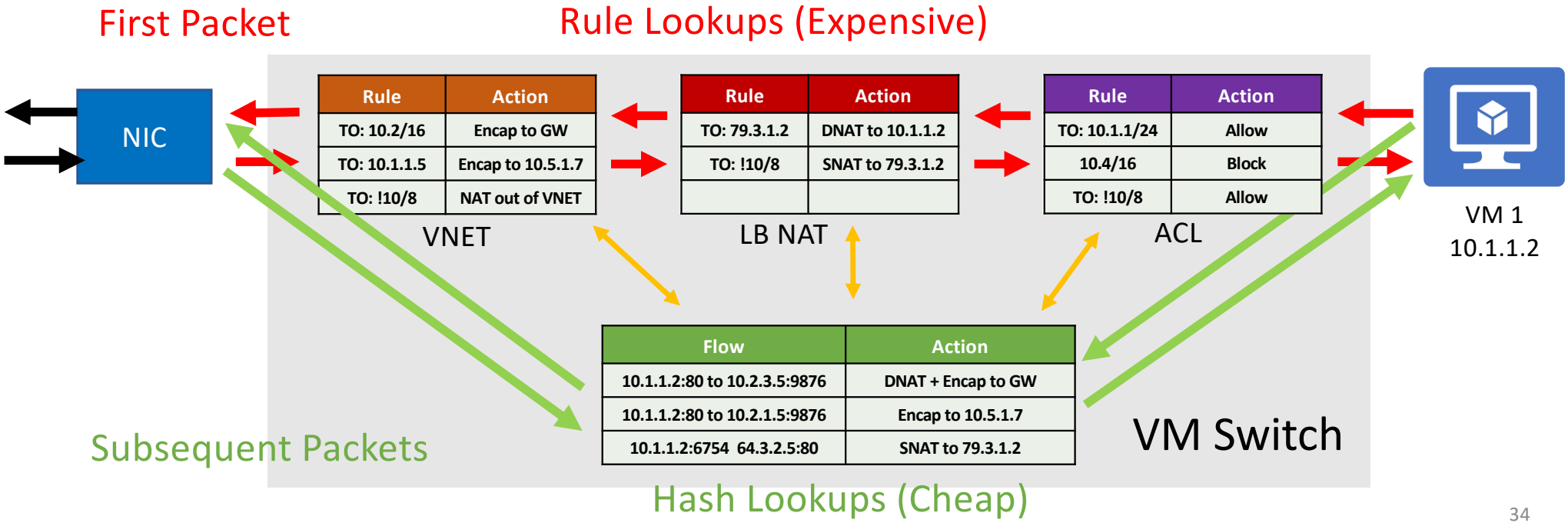
Azure VM Switch: Virtual Filtering Platform

- Virtual Filtering Platform (VFP) exposes a typed Match-Action-Table API to the agents/controllers
- Key insight: Let controller tell switch exactly what to do with which packets



Generic Flow Table (GFT) – A Fastpath

- First-packet actions can be complex
- Established-flow matches must be typed, predictable, and simple hash lookups

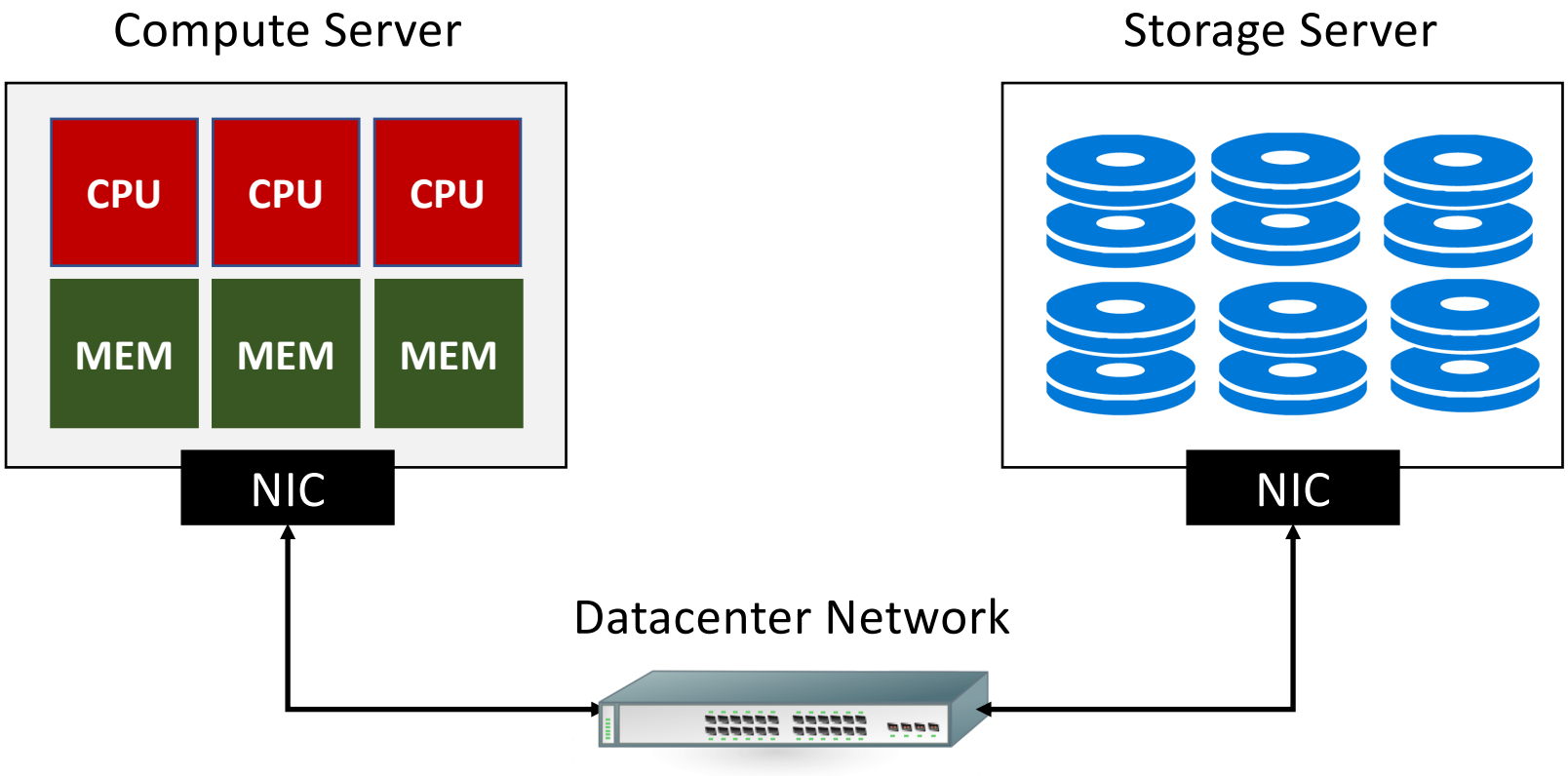


Compute and Storage

Virtual Machines in Azure

Type	Sizes	Description
General Purpose	B, Dsv3, Dv3, Dasv4, Dav4, DSv2, Dv2, Av2, DC, DCv2, Dpdsv5, Dpldsv5, Dpsv5, Dplsv5, Dv4, Dsv4, Ddv4, Ddsv4, Dv5, Dsv5, Ddv5, Ddsv5, Dasv5, Dadsv5	Balanced CPU-to-memory ratio. Ideal for testing and development, small to medium databases, and low to medium traffic web servers.
Compute optimized	F, Fs, Fsv2, FX	High CPU-to-memory ratio. Good for medium traffic web servers, network appliances, batch processes, and application servers.
Memory optimized	Esv3, Ev3, Easv4, Eav4, Epdsv5, Epsv5, Ev4, Esv4, Edv4, Edsv4, Ev5, Esv5, Edv5, Edsv5, Easv5, Eadsv5, Mv2, M, DSv2, Dv2	High memory-to-CPU ratio. Great for relational database servers, medium to large caches, and in-memory analytics.
Storage optimized	Lsv2, Lsv3, Lasv3	High disk throughput and IO ideal for Big Data, SQL, NoSQL databases, data warehousing and large transactional databases.
GPU	NC, NCv2, NCv3, NCasT4_v3, ND, NDv2, NV, NVv3, NVv4, NDasrA100_v4, NDm_A100_v4	Specialized virtual machines targeted for heavy graphic rendering and video editing, as well as model training and inferencing (ND) with deep learning. Available with single or multiple GPUs.
High performance compute	HB, HBv2, HBv3, HC, H	Our fastest and most powerful CPU virtual machines with optional high-throughput network interfaces (RDMA).

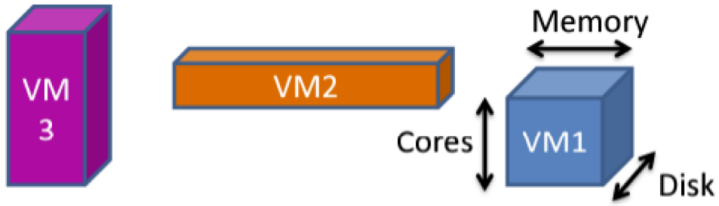
Compute-Storage Disaggregation

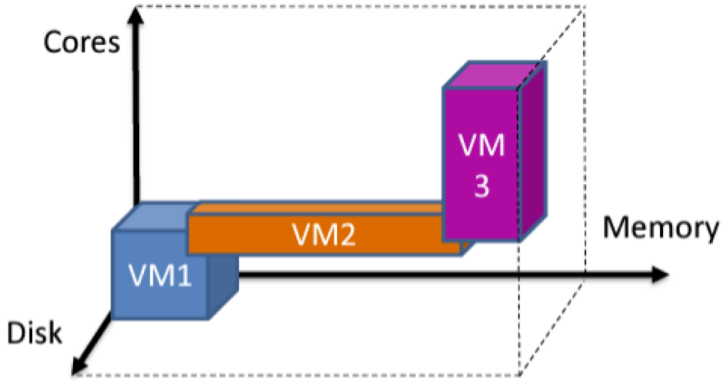


What are benefits of resource disaggregation?

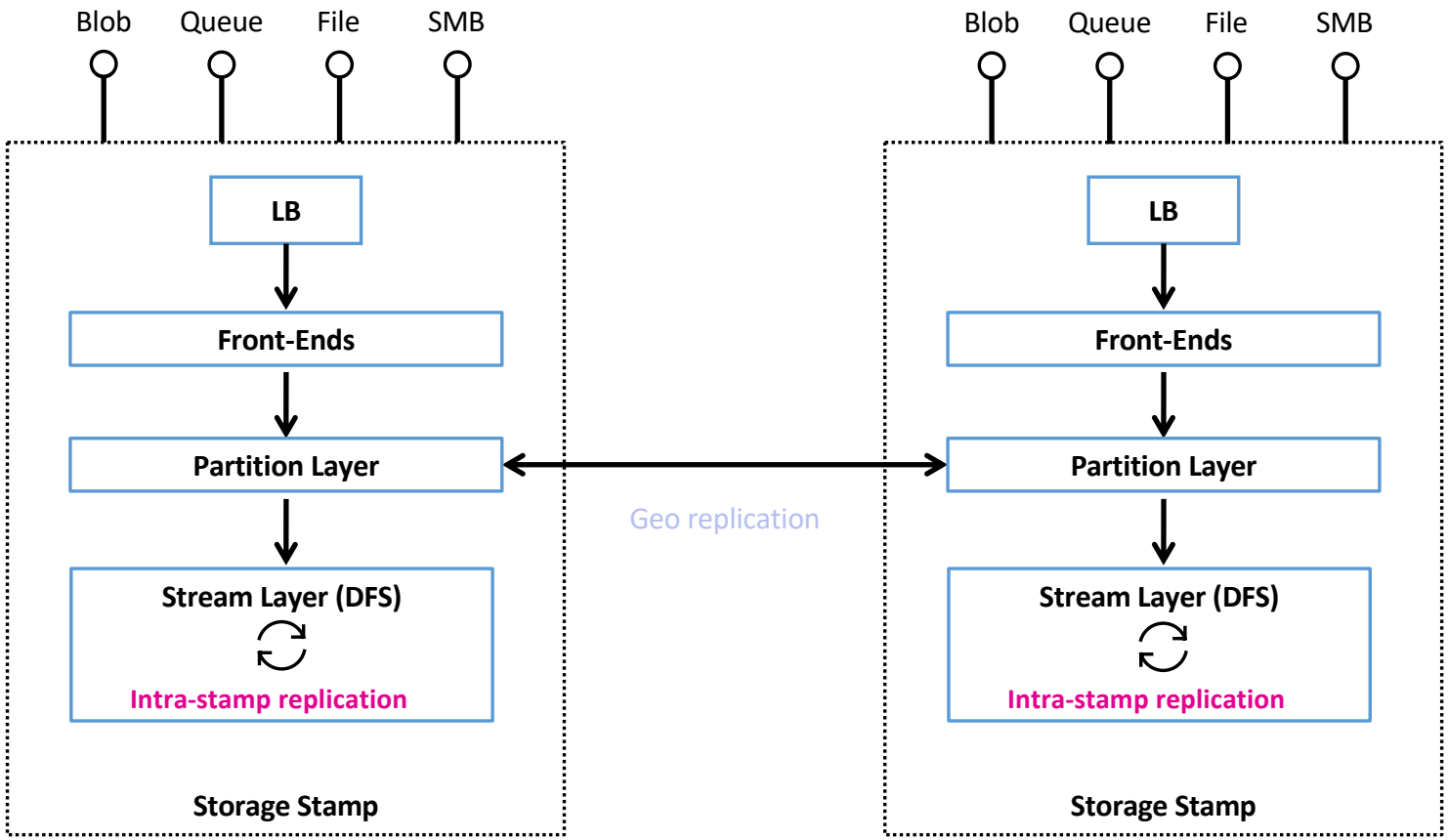
Virtual Machine (VM) Allocation

- Resource constraints
  - Sum of recourses of all VMs on a node cannot exceed server resources (CPU, memory, disk, network, …) -> **Bin-Packing**
- Failure domain constraints
  - VMs of the same tenant must be spread across many failure domains
- Co-location constraints
  - Certain types of VMs cannot be co-located together





Azure Storage Architecture

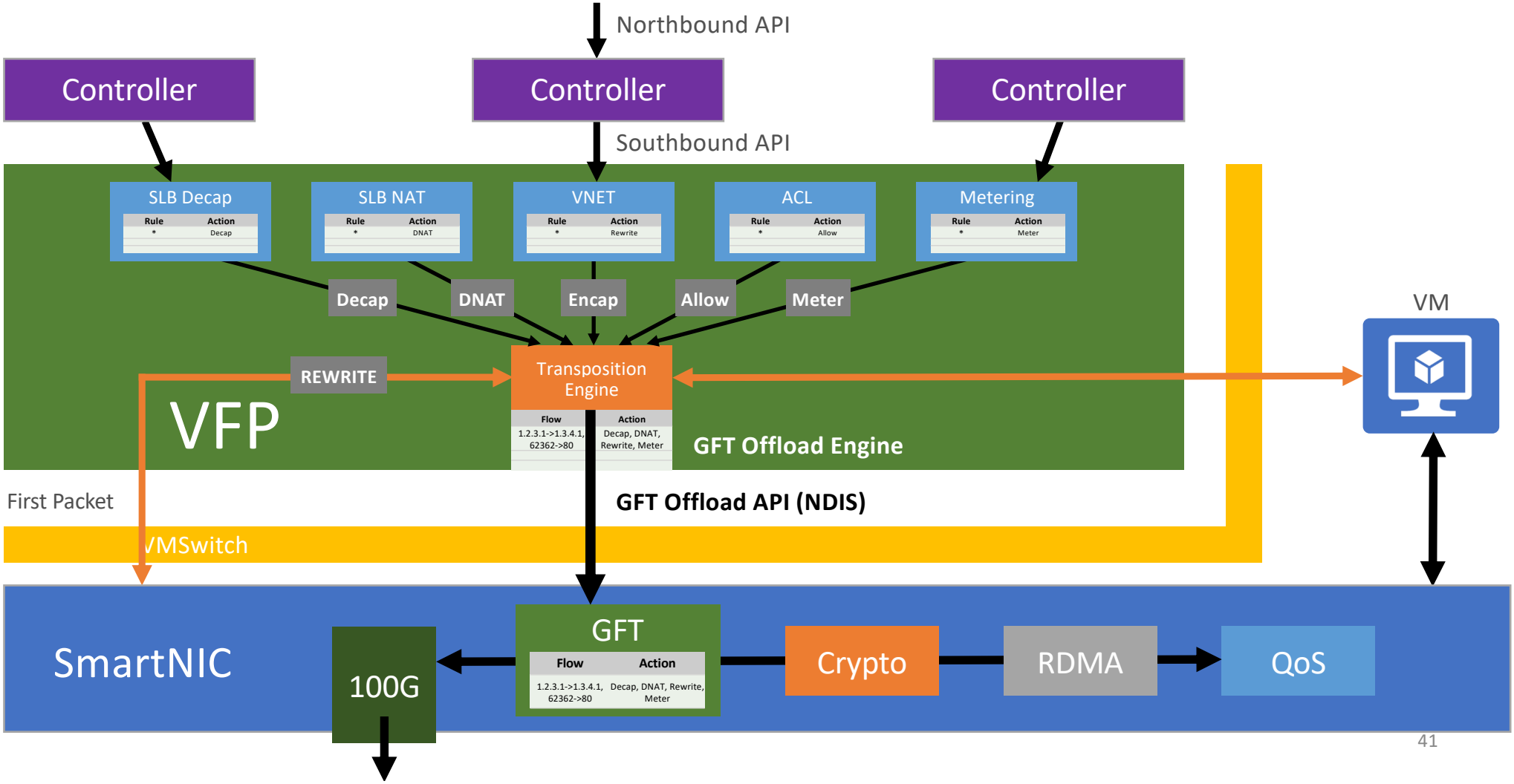


Datacenter Innovations

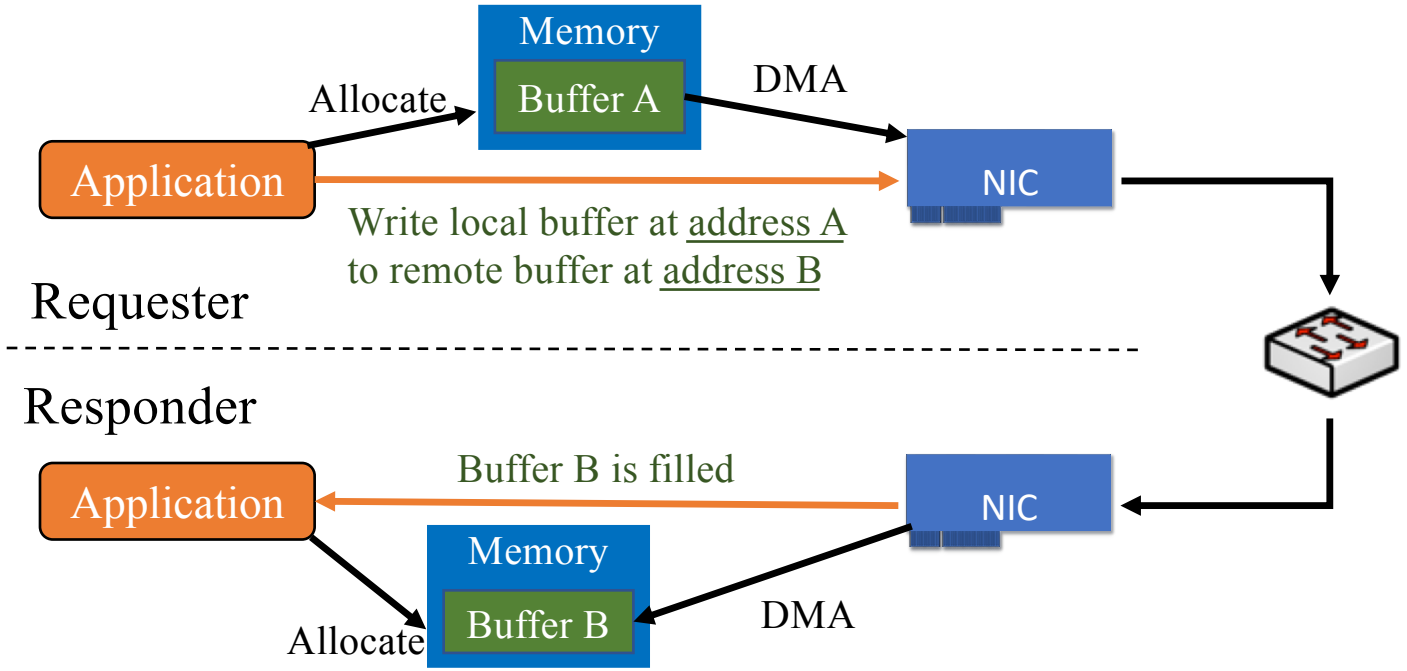
1% saving or improvement means a lot at cloud scale

0.01% downtime also means a lot for a cloud

Azure SmartNIC: accelerate VM switch

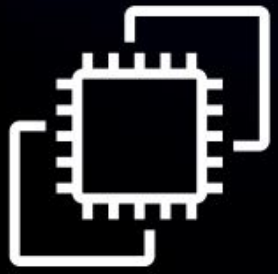


Remote Direct Memory Access (RDMA)

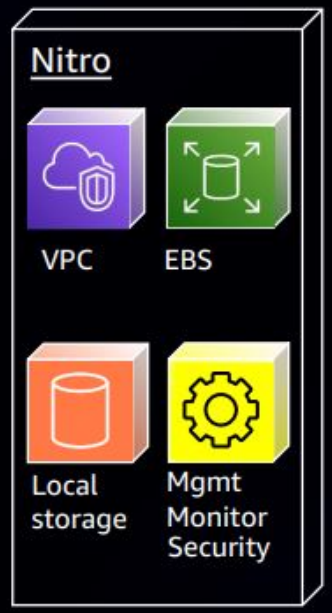


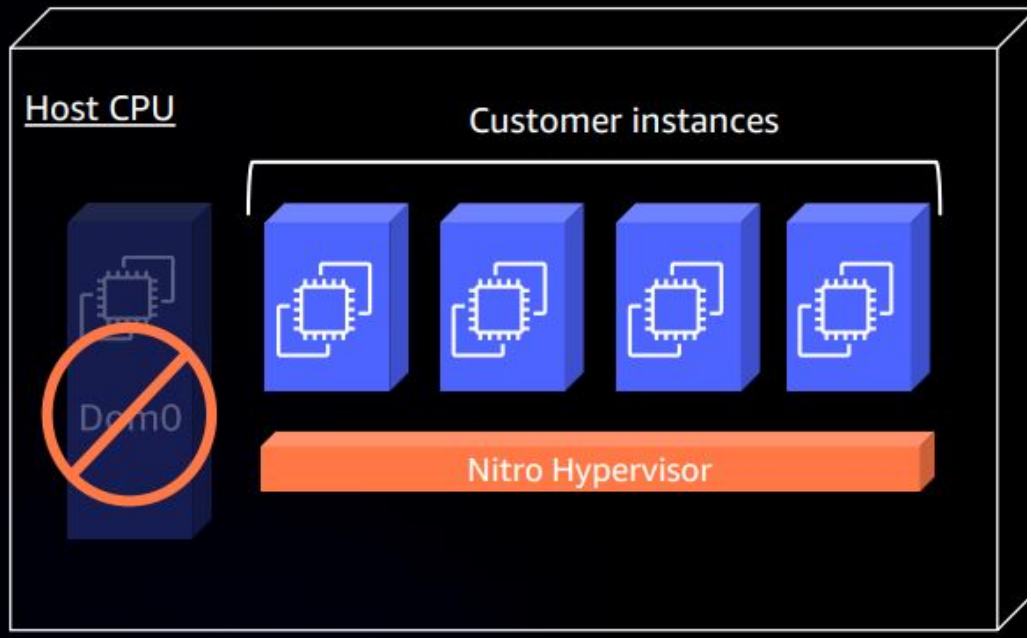
RDMA bypasses host OS stack
→ frees host CPU, lowers latency

AWS Nitro System

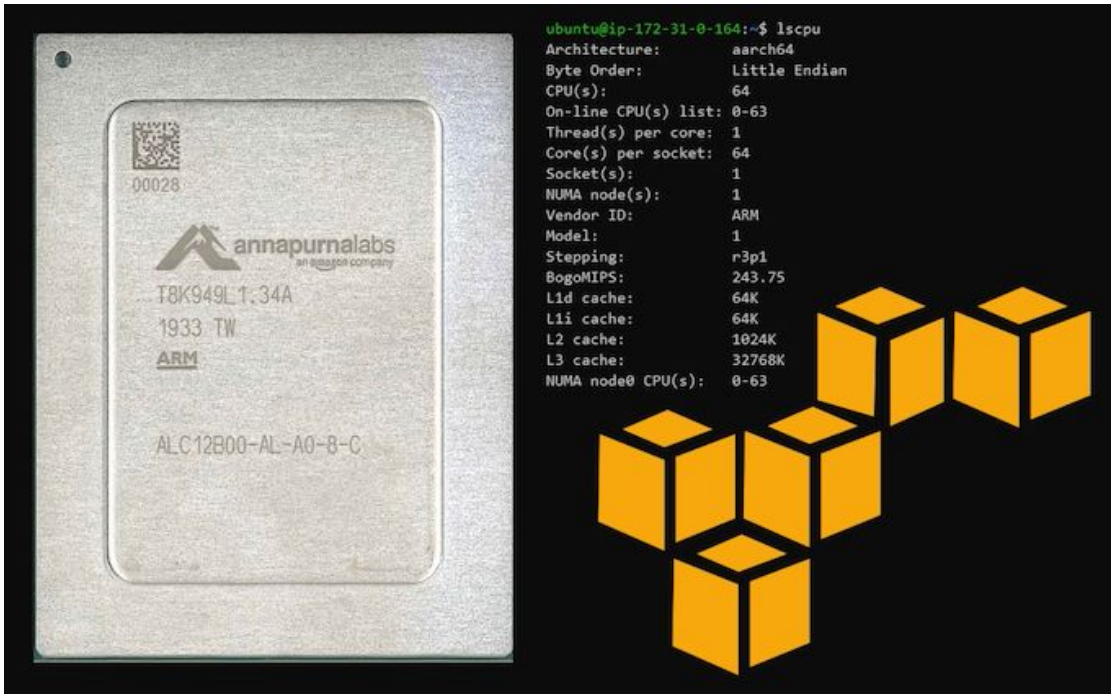


AWS Nitro System



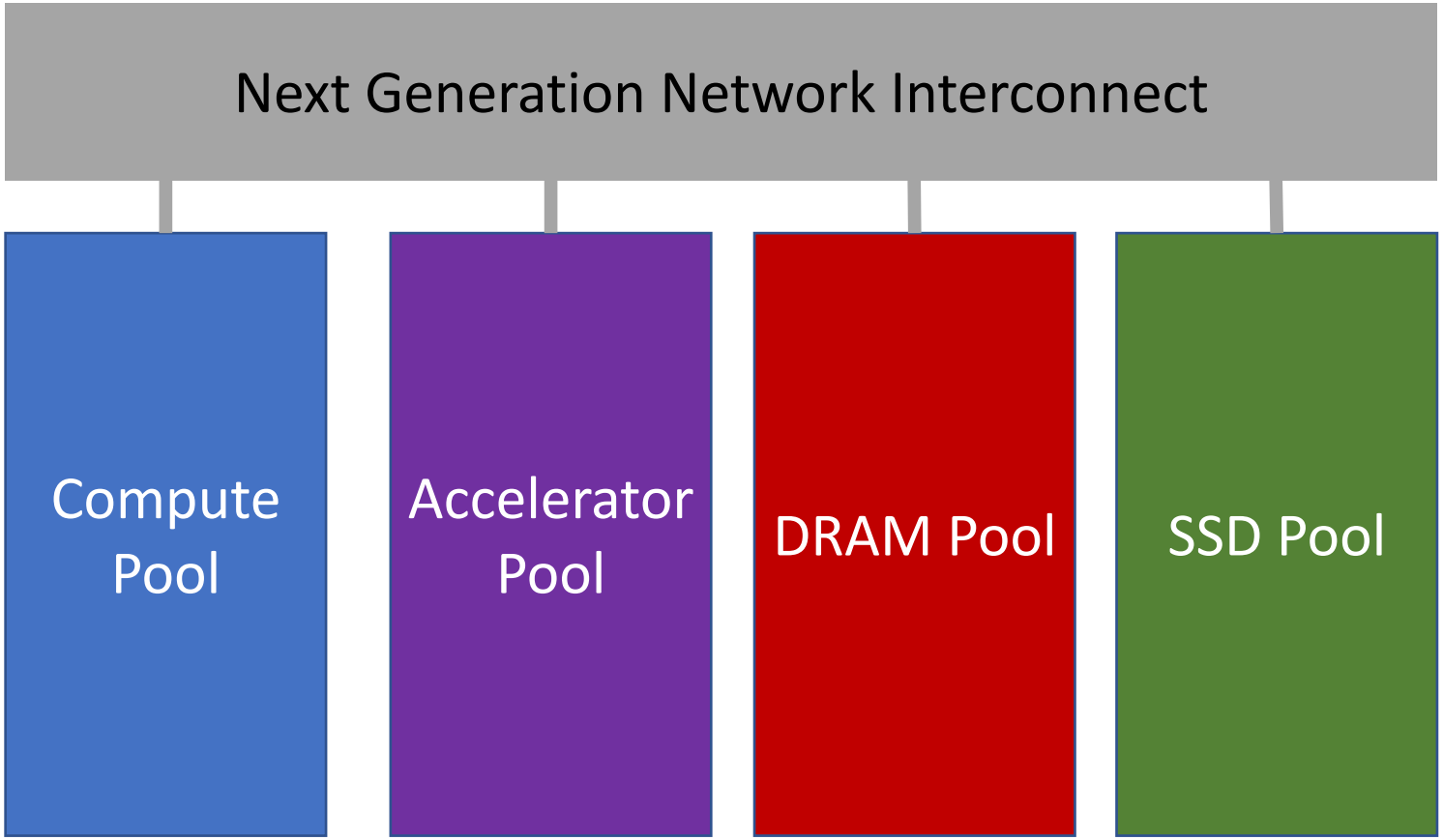


New CPUs

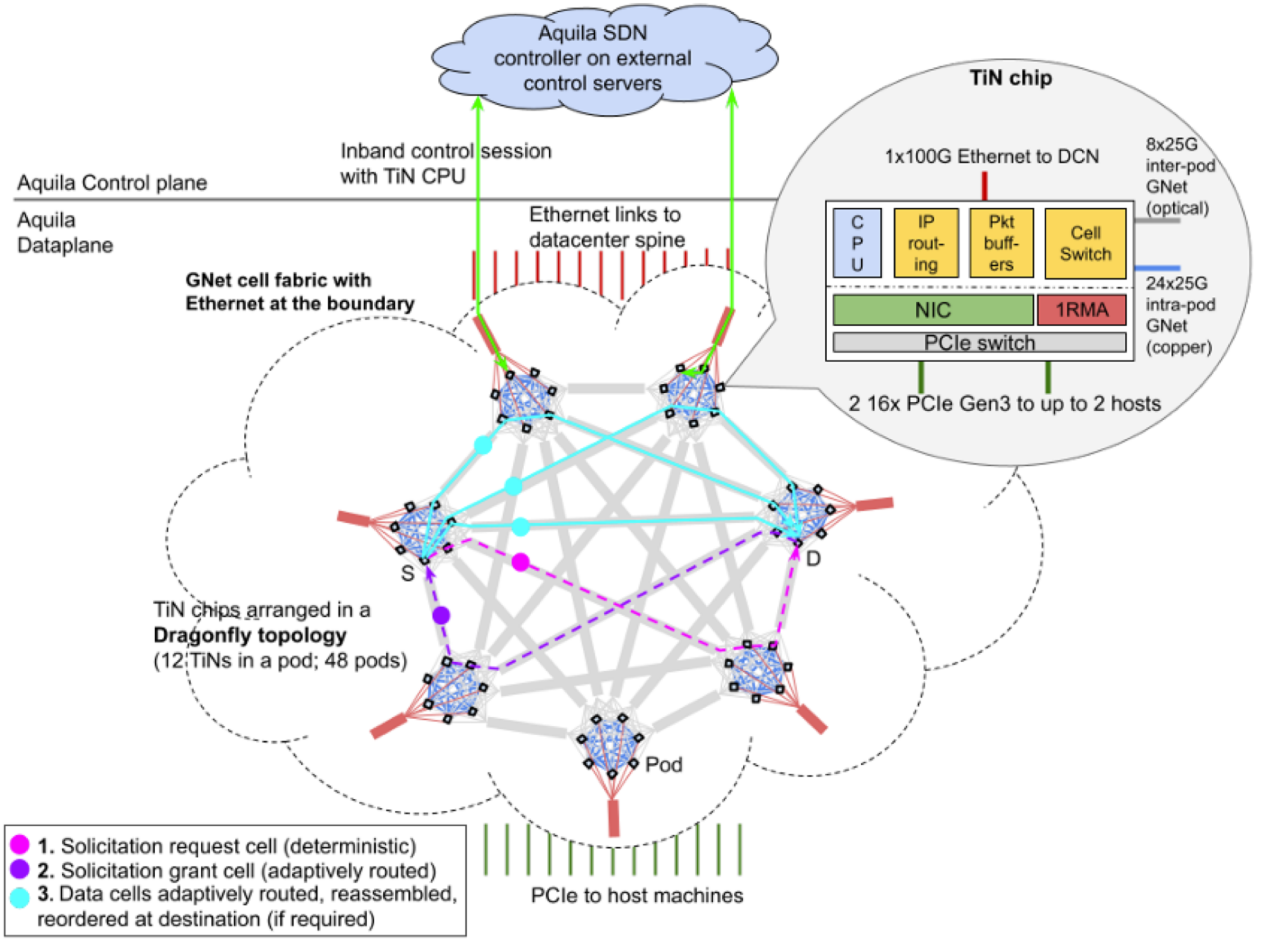


AWS Arm-based Graviton2 CPU

Next Generation Datacenter Architectures



Aquila: A Unified, Low-latency Datacenter Network



Sustainable Datacenters

An on-site solar energy array at the Google data center campus in Belgium

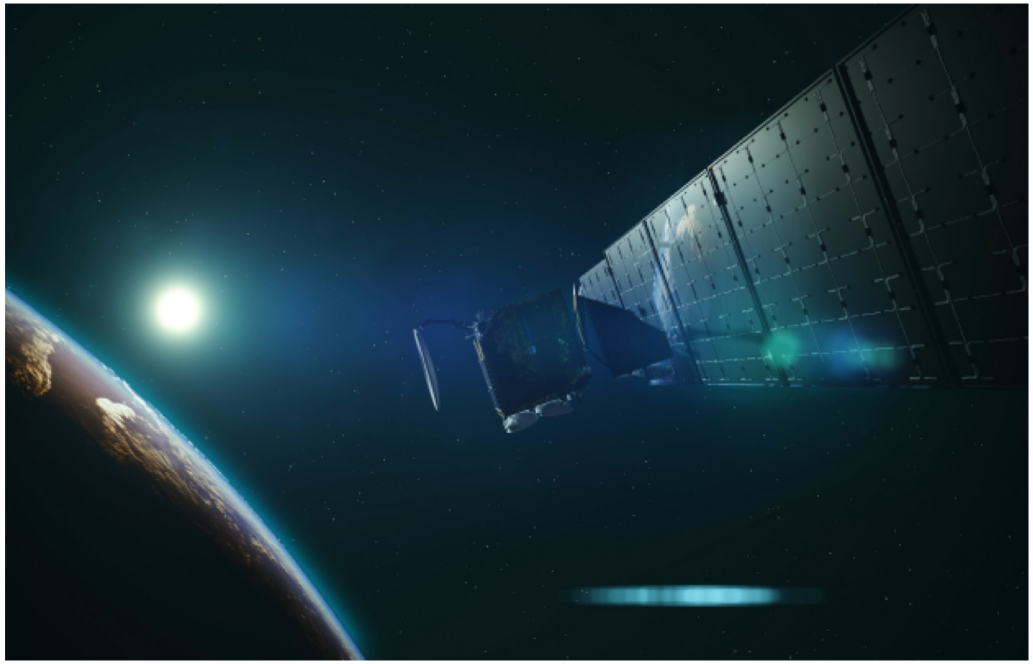
Boiling liquid carries away heat generated by computer servers at a Microsoft datacenter

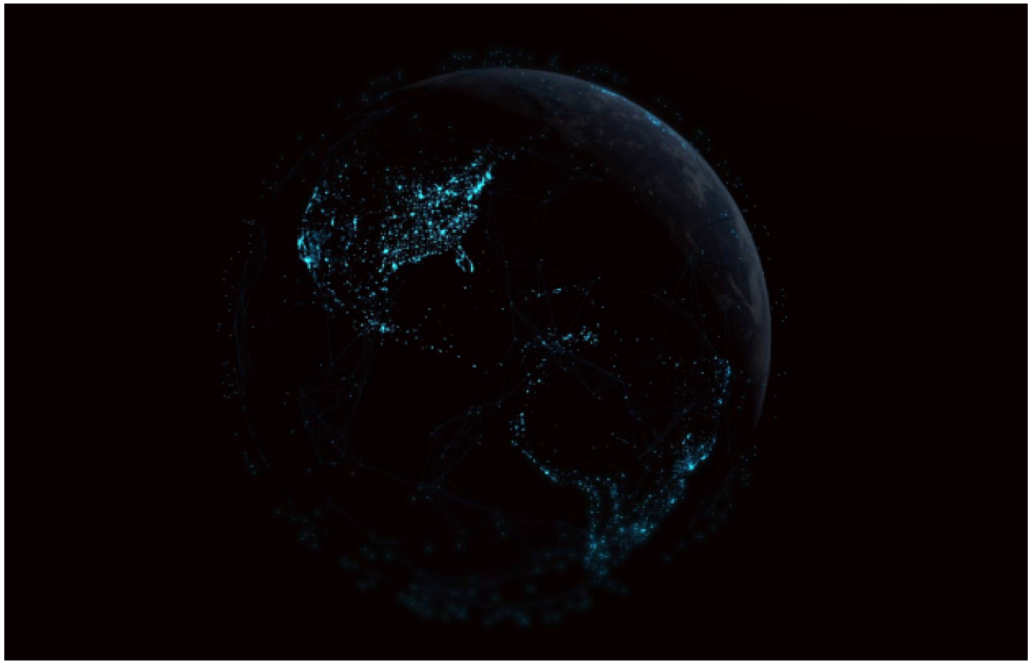
Project Natick: Underwater Datacenters

Customer Benefits

- Reduce the cost of cooling by leveraging the natural sea temperature.
- Benefit from offshore renewable energy sources (e.g., wave and tidal power) that could be truly zero emission.
- Made from fully recyclable materials.
- Reduce the latency by locating datacenters closer to customers.

Datacenters in Space





Thank You!

Interested in opportunities at Microsoft?

Please drop me an email webai@microsoft.com