Microsoft Azure

Clermont'ech – API Hour #57
La plomberie dans les Datacenters Azure : comment ça marche ?

David SANTIAGO & Vincent MISSON

19 décembre 2023

~~$ whoami~~ $ whoarewe

Vincent Misson

Cloud Solution Architect @Microsoft

blog.cloud63.fr

vmisson

David Santiago

Cloud Solution Architect @Microsoft

davidsantiago.fr

davidsntg

Agenda

- Azure Global Infrastructure
- Inside an Azure Datacenter
- The life of a packet

Inside Azure

Global Infrastructure

Microsoft global network

65 Azure regions

160K+ miles of fiber and subsea cables

180 Edge PoPs

200+ ExpressRoute partners

Azure region architecture

Geography

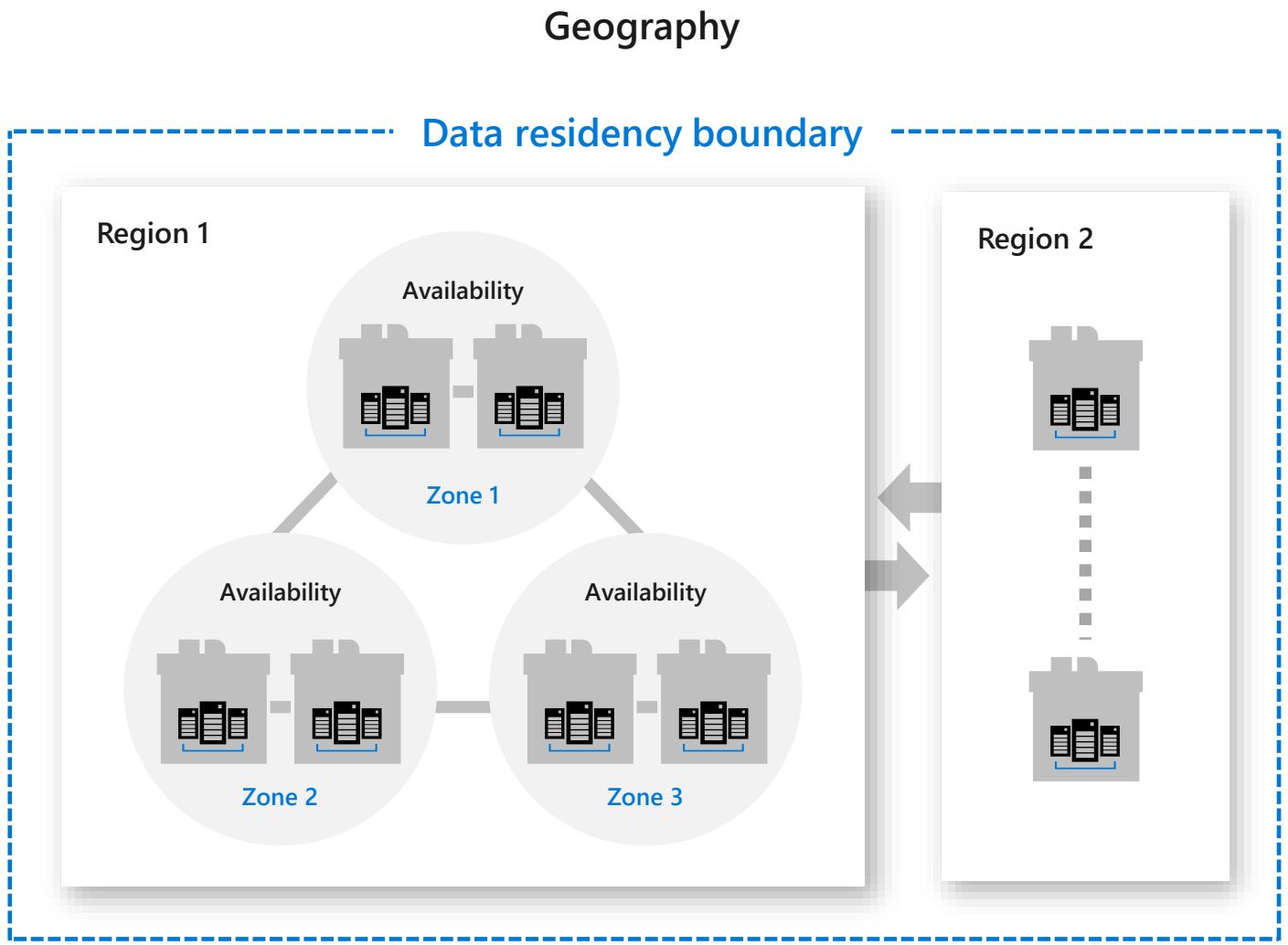
- Discrete market with two or more regions
- Meets data residency and compliance requirements
- Fault-tolerant to protect from complete region failure

Region

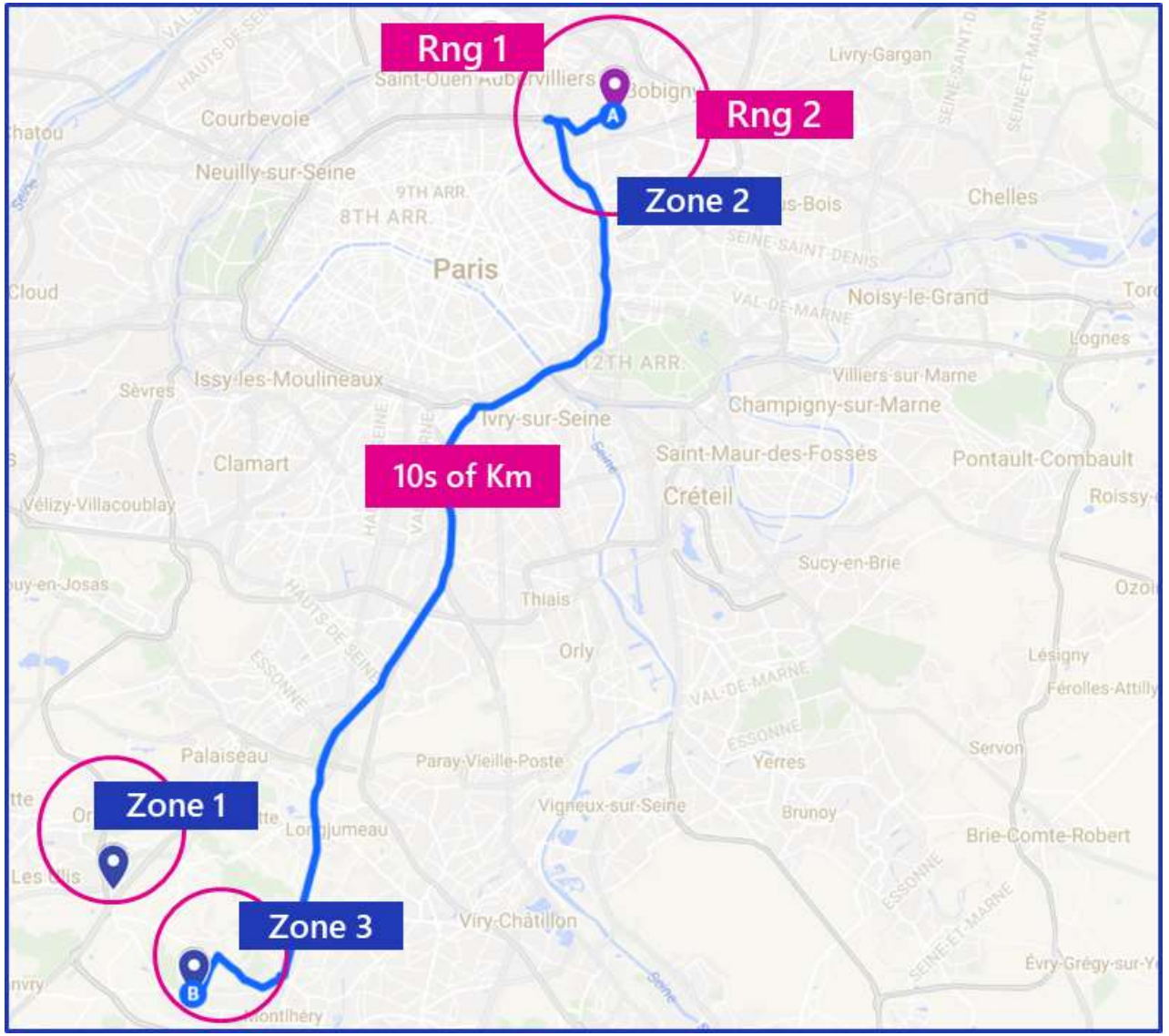
- Set of datacenters within a metropolitan area
- Network latency perimeter <2ms

Availability Zones

- Unique physical locations within an Azure region
- Each zone is made up of one or more DCs
- Independent power, cooling and networking
- Inter-AZ network latency <2ms
- Fault-tolerant to protect from datacenter failure



Region: France Central



Regional Networks

Edge

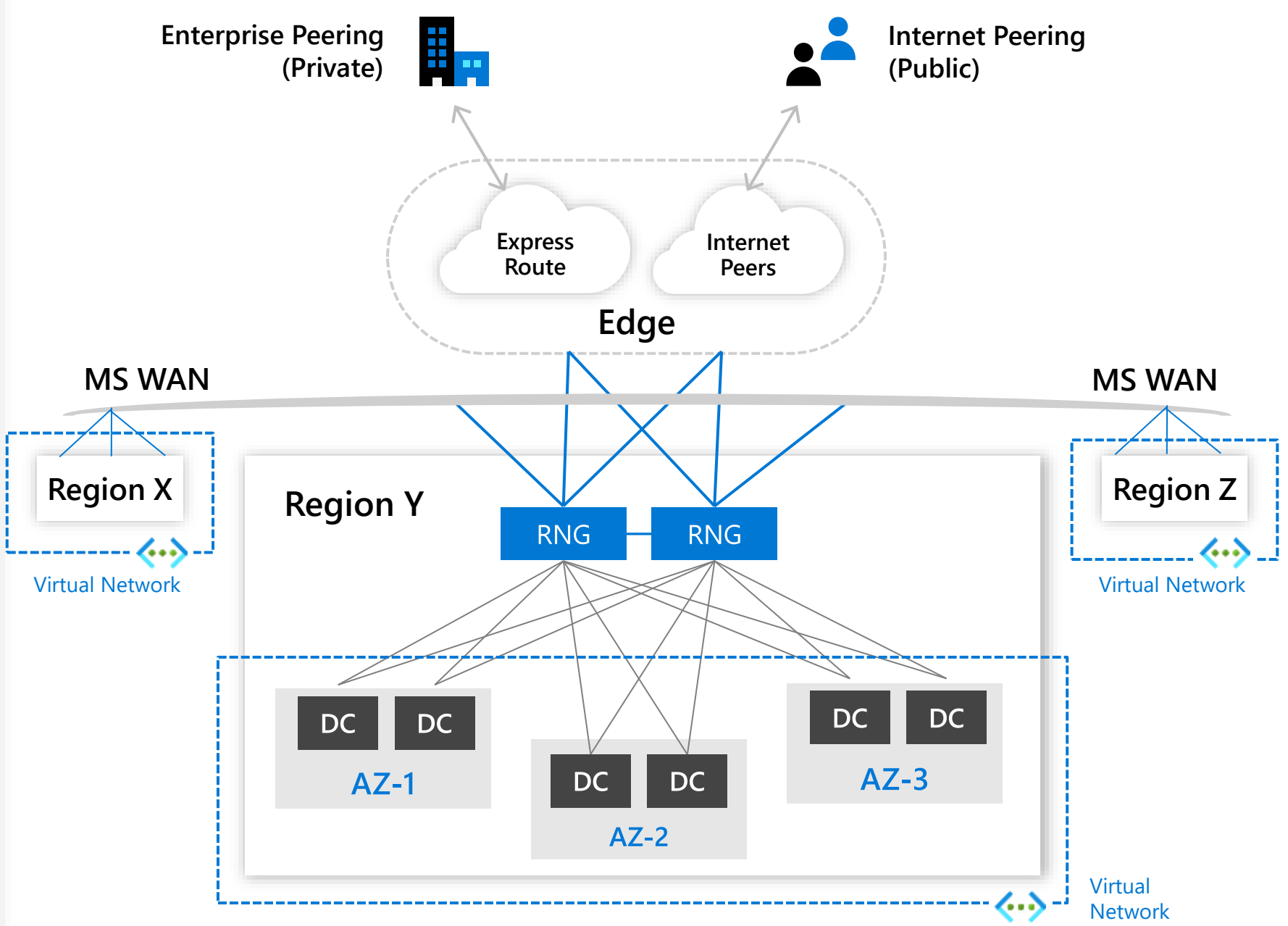
- Connects Region to Internet and Enterprise peers

Regional Network Gateway

- Massively parallel, hyper scale DC interconnect
- Space and power protected

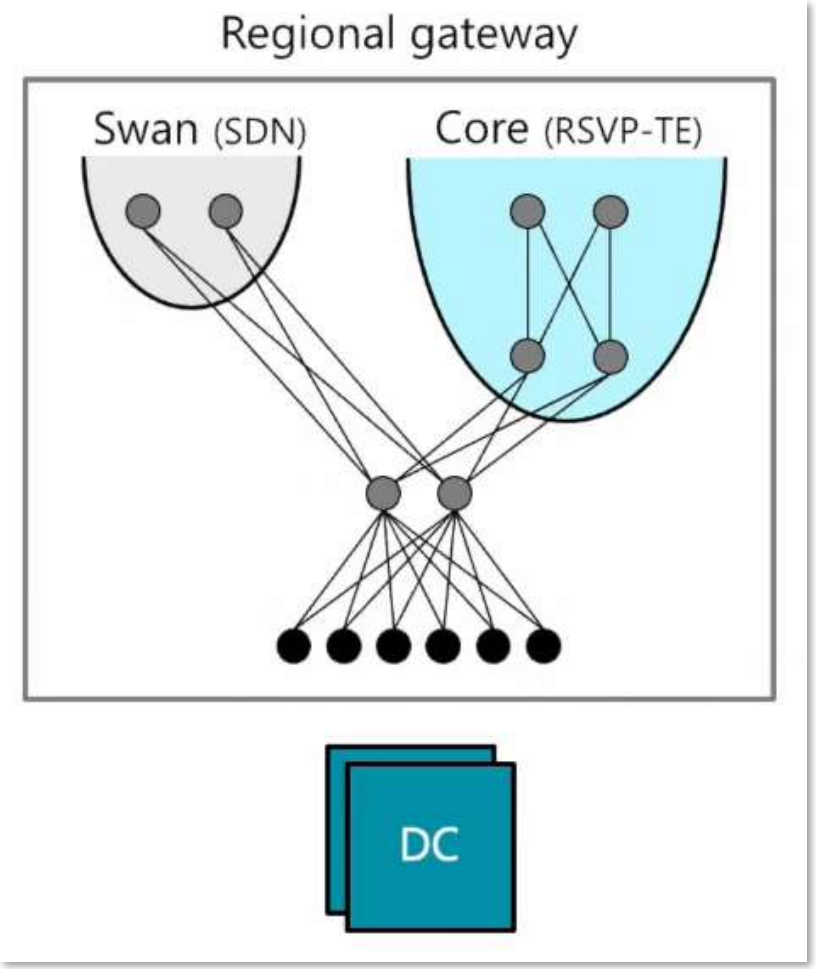
Datacenters

- Small, Medium or Large (T-shirt sizes)
- Only contains server racks, DC network
- RNGs are sized to support growing the region by adding data centers



OneWAN is better than two: Unifying a split WAN architecture

- 8074 network (SWAN)
  - Inter-datacenter
- 8075 network (Core)
  - User facing traffic



https://www.microsoft.com/en-us/research/publication/onewan-is-better-than-two-unifying-a-split-wan-architecture/

Azure Networking overview

DC hardware

- SmartNIC/FPGA
- SONiC

Services

- Virtual Networks
- Load Balancing
- VPN Services
- Firewall
- DDoS Protection
- DNS and Traffic Management

Intra-region

- DC Networks
- Regional Networks
- Optical Modules

WAN backbone

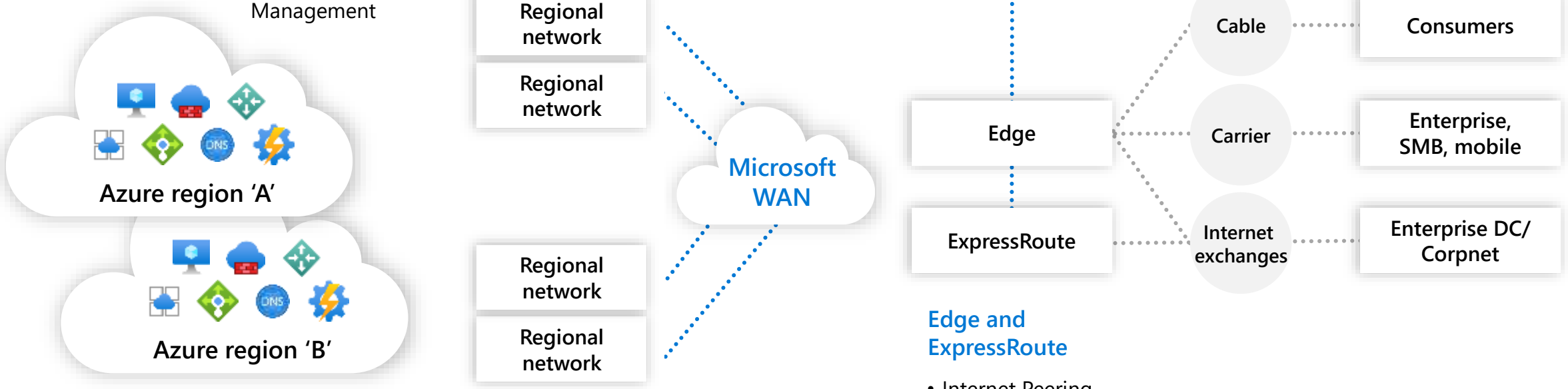
- Software WAN
- Subsea Cables
- Terrestrial Fiber
- National Clouds

CDN

- Acceleration for applications and content

Last mile

- E2E monitoring (Network Watcher, Network Performance Monitoring)



Edge and ExpressRoute

- Internet Peering
- ExpressRoute

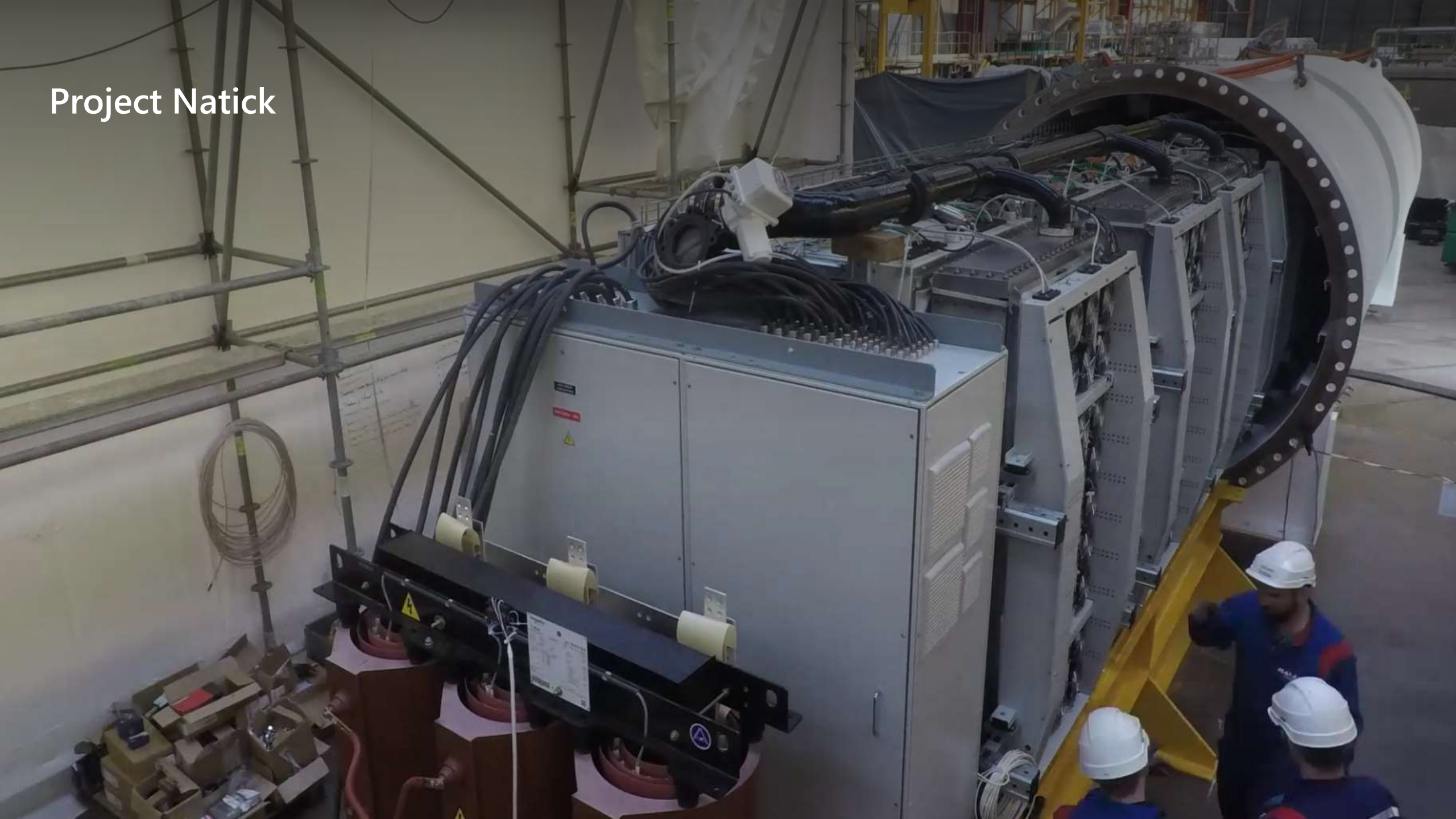
Inside Azure

Inside Datacenter

Quincy, WA

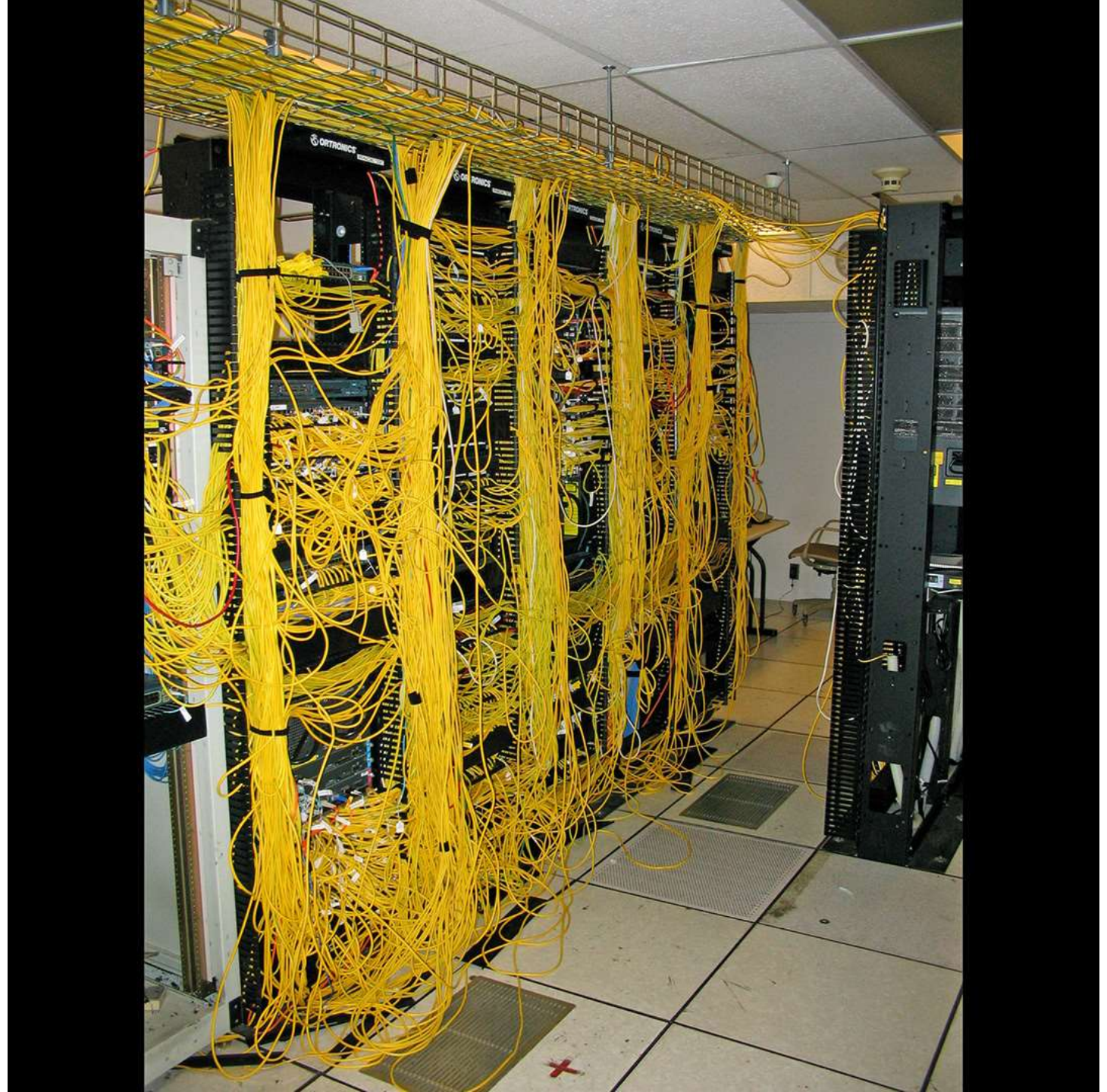

Dublin, Ireland

Project Natick



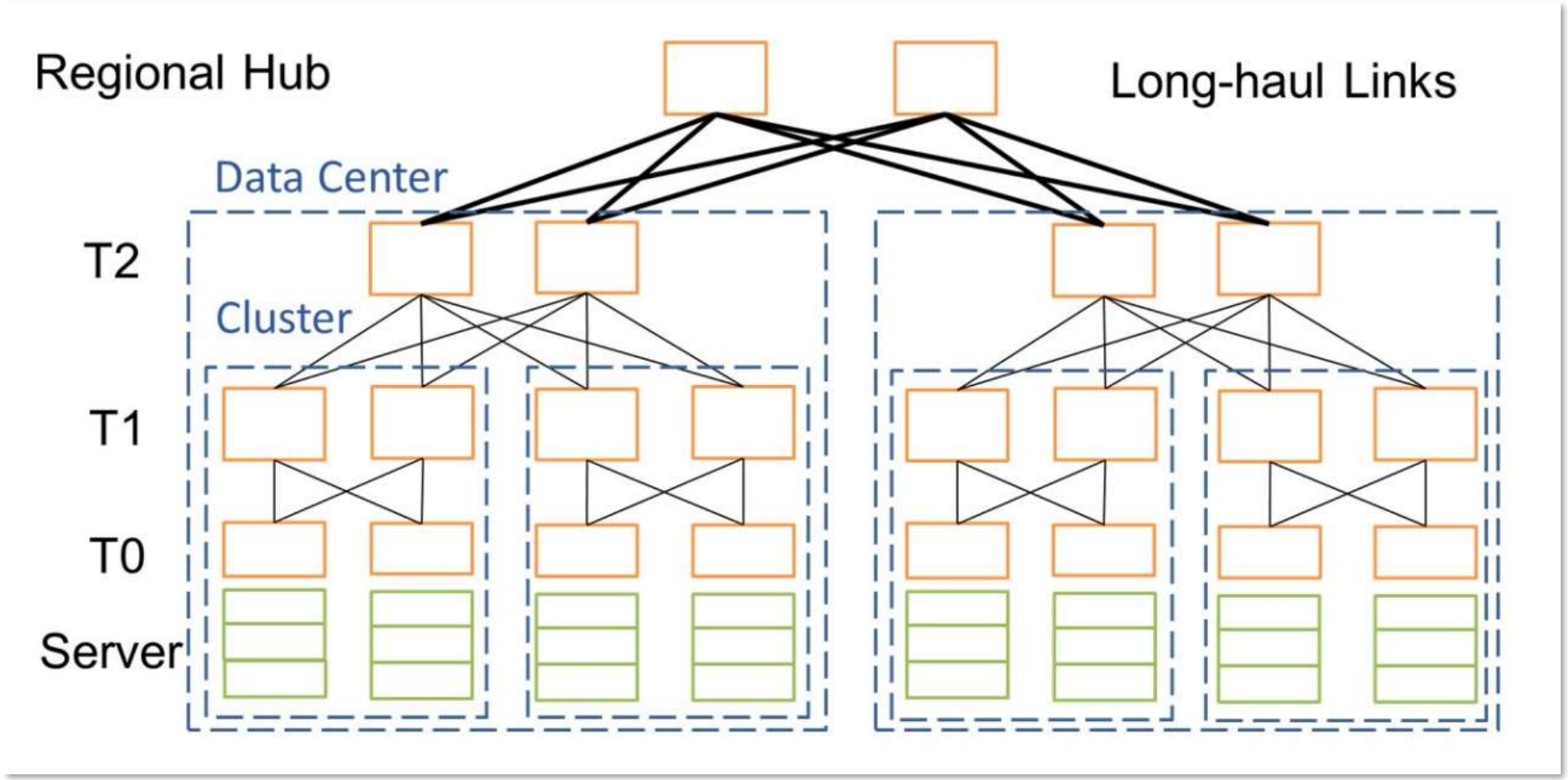
Amsterdam

Amsterdam


ORTRONICS


Cheyenne, WY

Azure Networking Physical Layout



https://www.microsoft.com/en-us/research/publication/empowering-azure-storage-with-rdma/

Software for Open Networking in the Cloud (SONIC)

Native Linux

- Native Linux kernel
- Leverage Unix networking stack

Containerized

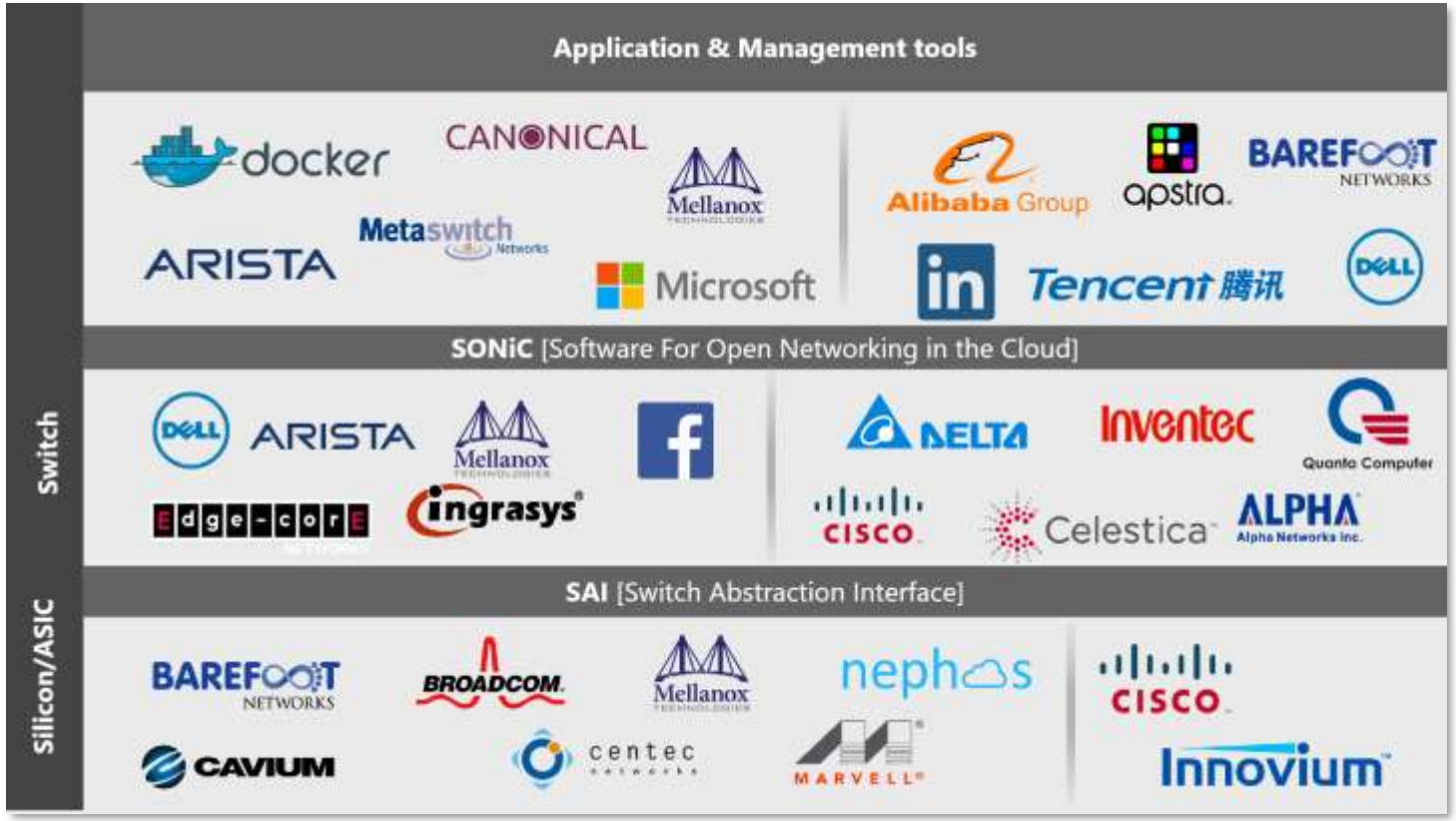
- Quick feature release and bug fix
- Hitless upgrade without customer impact

Open Sourced

- On GitHub with Apache License
- Build on top of SAI

Rich ecosystem

- Platform/ASIC Agnostic



https://github.com/sonic-net/SONiC

Inside Azure

Life of a packet

Azure servers: General purpose

RAM ⟷ Cores

3x Beast

Gen 2	
Processor	2 x 6 Core 2.1 GHz
Memory	32 GiB
Hard Drive	6 x 500 GB
SSD	None
NIC	1 Gb/s

Godzilla	
Processor	2 x 16 Core 2.0 GHz
Memory	512 GiB
Hard Drive	None
SSD	9 x 800 GB
NIC	40 Gb/s

Intel Gen 6	
Processor	2 x Skylake 24 Core 2.7GHz
Memory	768GiB DDR4
Hard Drive	None
SSD	4 x 960 GB M.2 SSDs and 1 x 960 GB SATA
NIC	40 Gb/s + FPGA

Optimized Gen 6	
Processor	2 x 24 core Skylake Lake
Memory	192 GB DDR4
Hard Drive	None
SSD	5 x 960 GB M.2 NVMe
NIC	40 Gb/s + FPGA

AMD Gen 6	
Processor	2 x 32 core Naples
Memory	512 GB DDR4
Hard Drive	None
SSD	7 x 960 GB M.2 NVMe
NIC	50 Gb/s + FPGA

Intel Gen 7	
Processor	2 x 26 core Cascade Lake
Memory	576 GB DDR4
Hard Drive	None
SSD	7 x 960 GB M.2 NVMe
NIC	50 Gb/s + FPGA

AMD Gen 7	
Processor	2 x 32 core Rome
Memory	768 GB DDR4
Hard Drive	None
SSD	7 x 960 GB M.2 NVMe
NIC	50 Gb/s + FPGA

Beast	
Processor	4 x 18 Core 2.5 GHz
Memory	4096 GiB
Hard Drive	None
SSD	4 x 2 TB NVMe, 1 x 960 GB SATA
NIC	40 Gb/s

Beast v2	
Processor	8 x 28 Core 2.5 GHz
Memory	12 TiB
Hard Drive	None
SSD	4 x 2 TB NVMe, 1 x 960 GB SATA
NIC	50 Gb/s

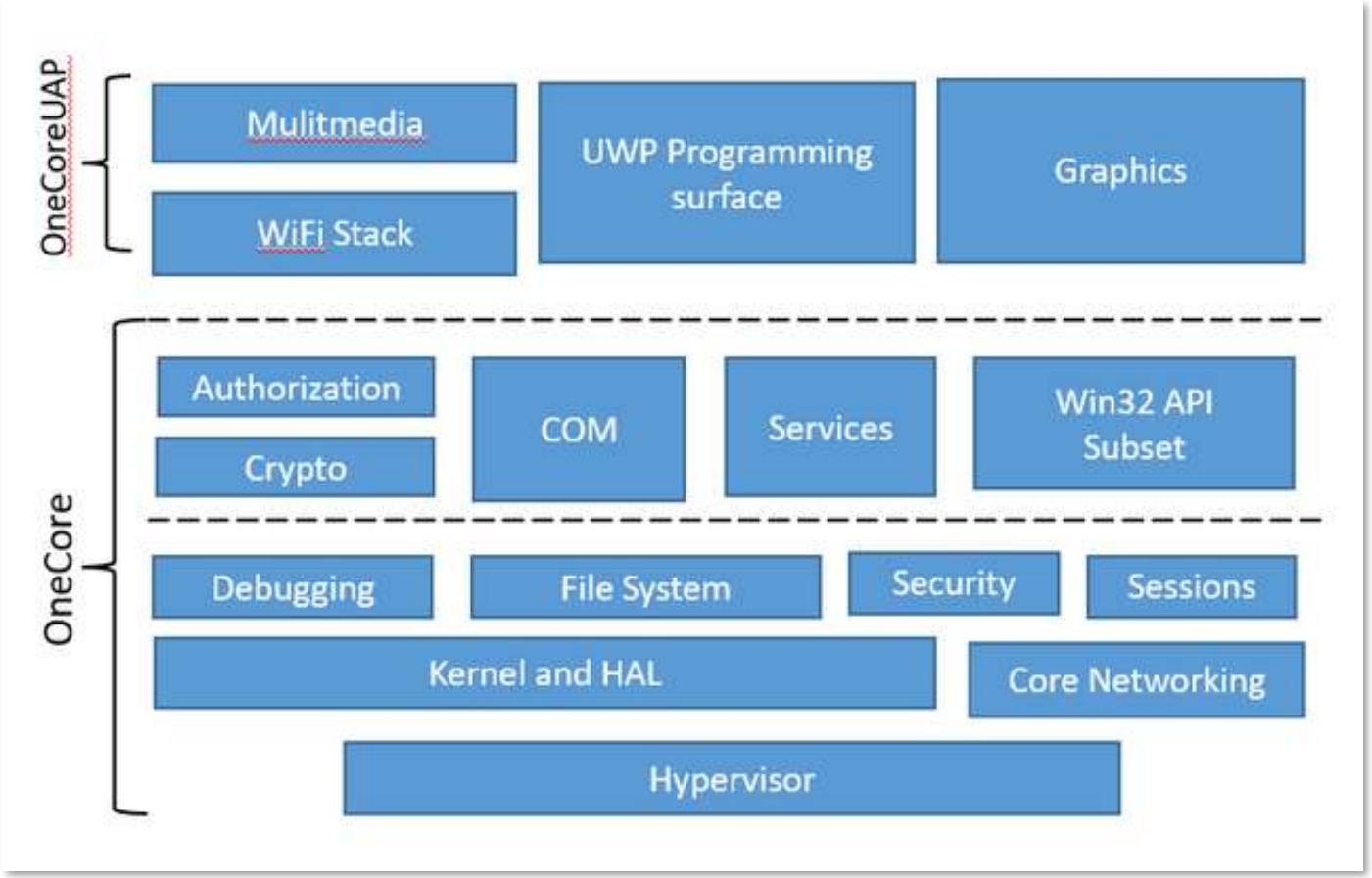
Azure Host OS

Long story short

- Headless (no GUI) – console only
- Supports x86, x64 & ARM architectures
- Mainly C++, Python & Rust
- ~280 MB
- Disable unused drivers and features such a print spooler, etc.
- Minimal set of server roles

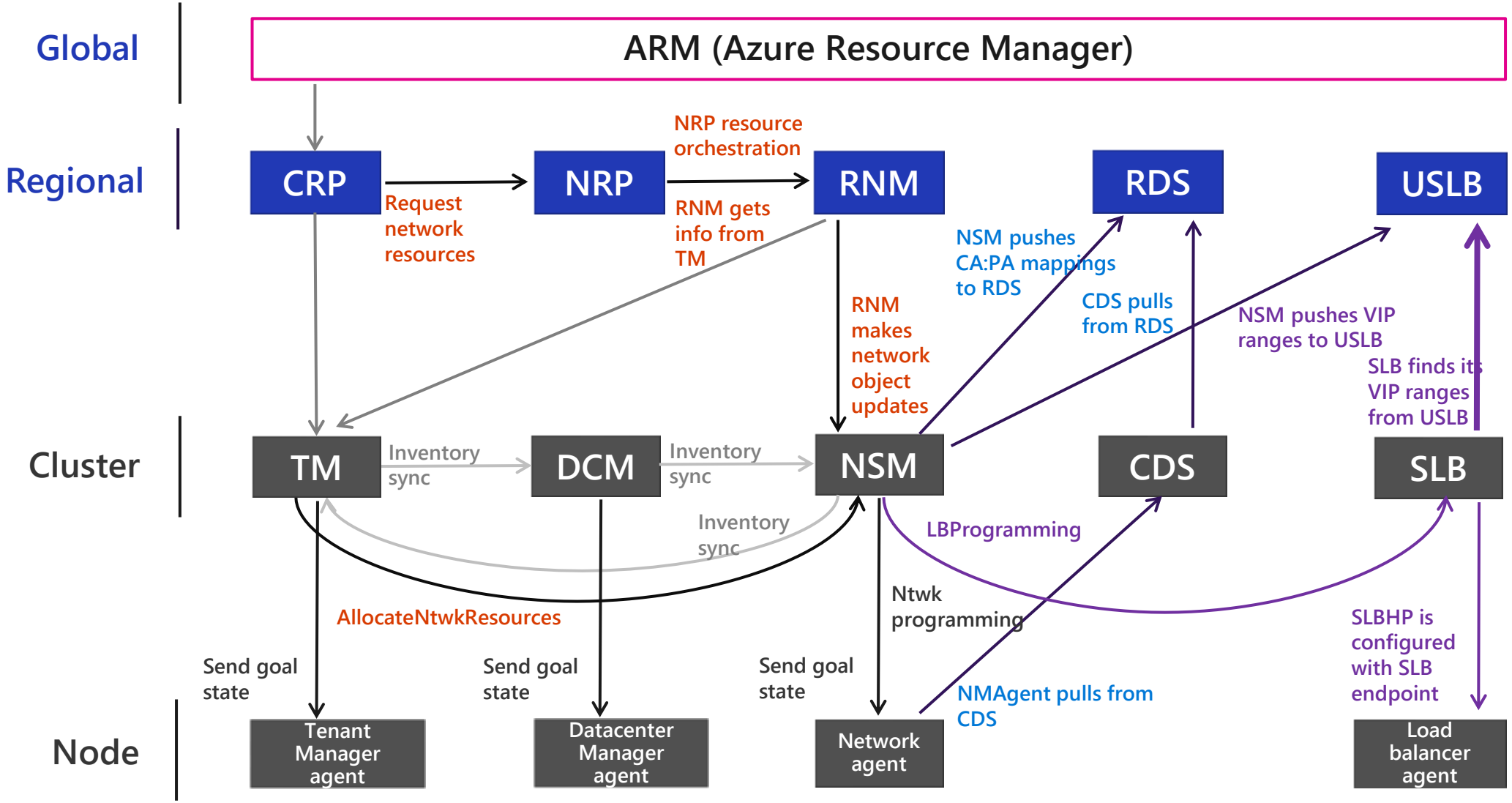
Patching

- Hot Patching – in ms
- VM PHU (~30s)
- Live Migration (few s)
- Hypervisor Hot Restart < 1s
- Details

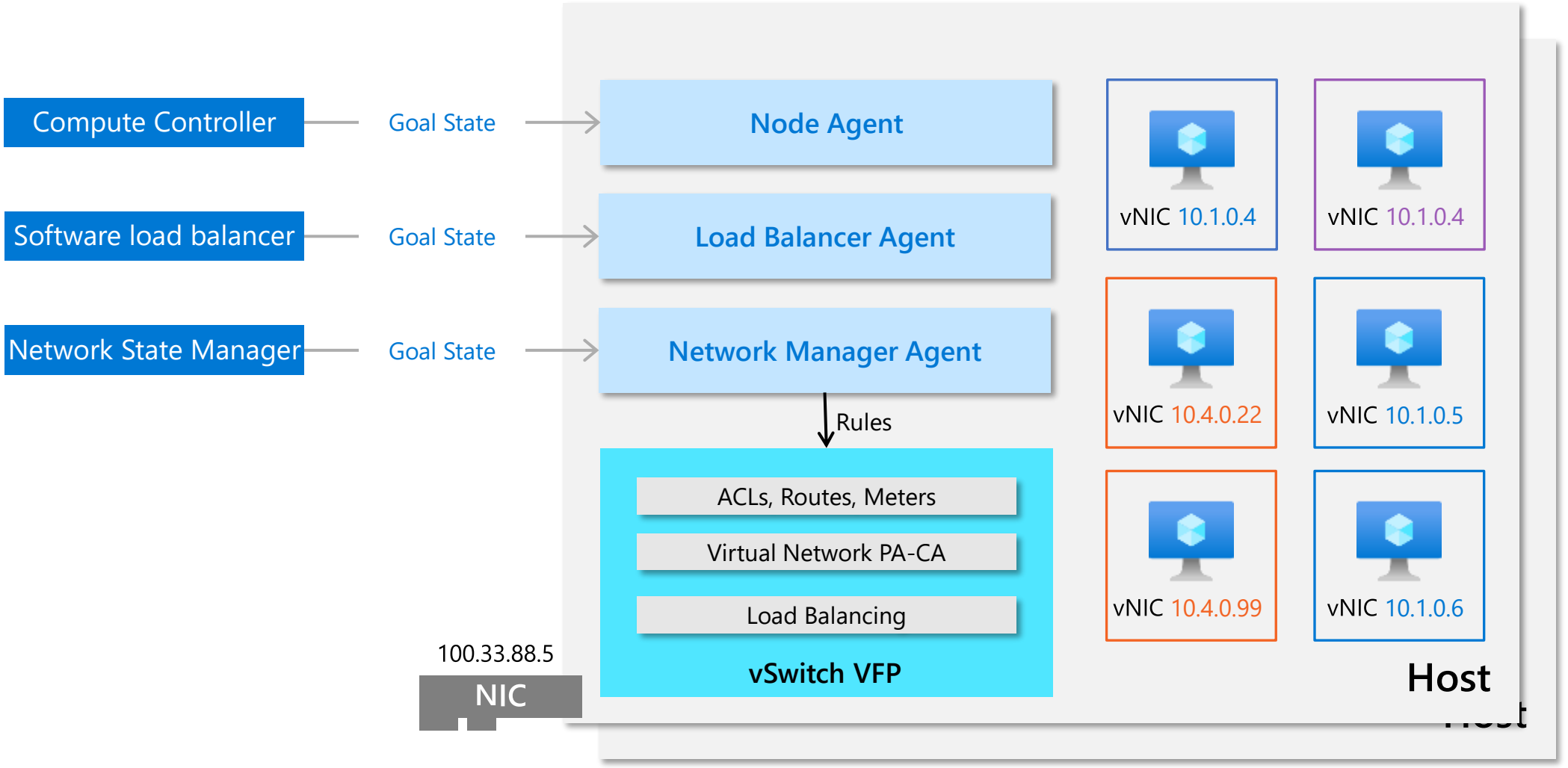


https://techcommunity.microsoft.com/t5/windows-os-platform-blog/azure-host-os-cloud-host/ba-p/3709528

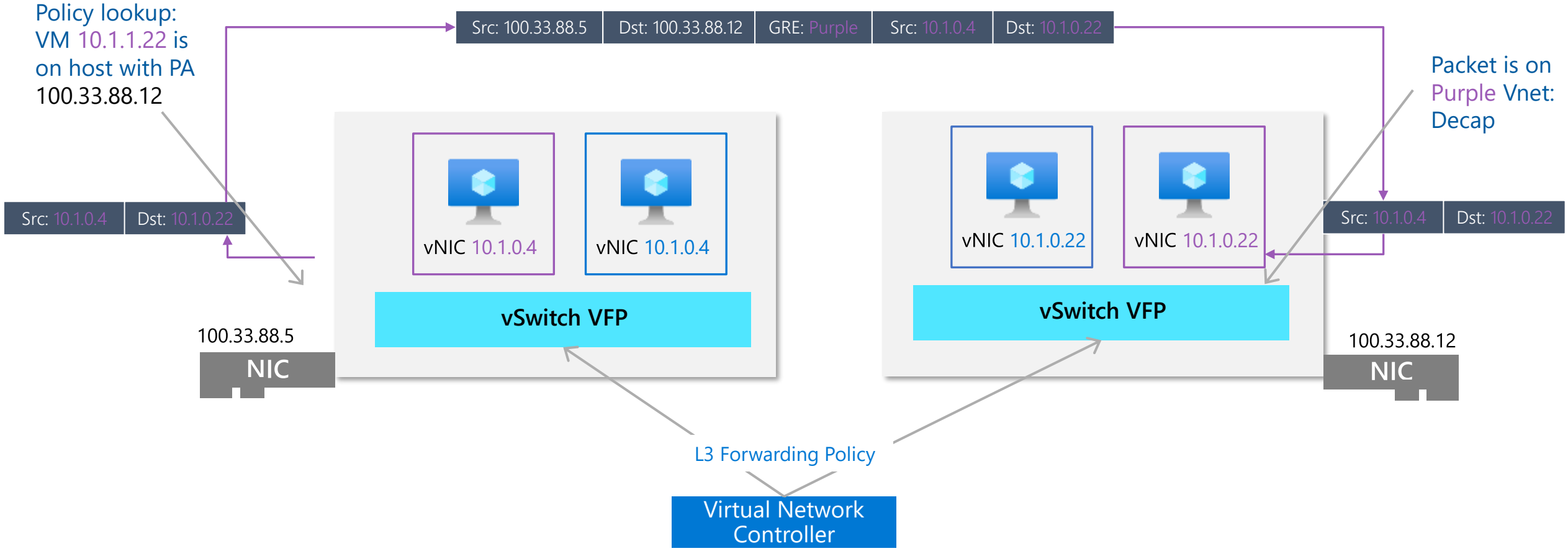
Azure compute architecture



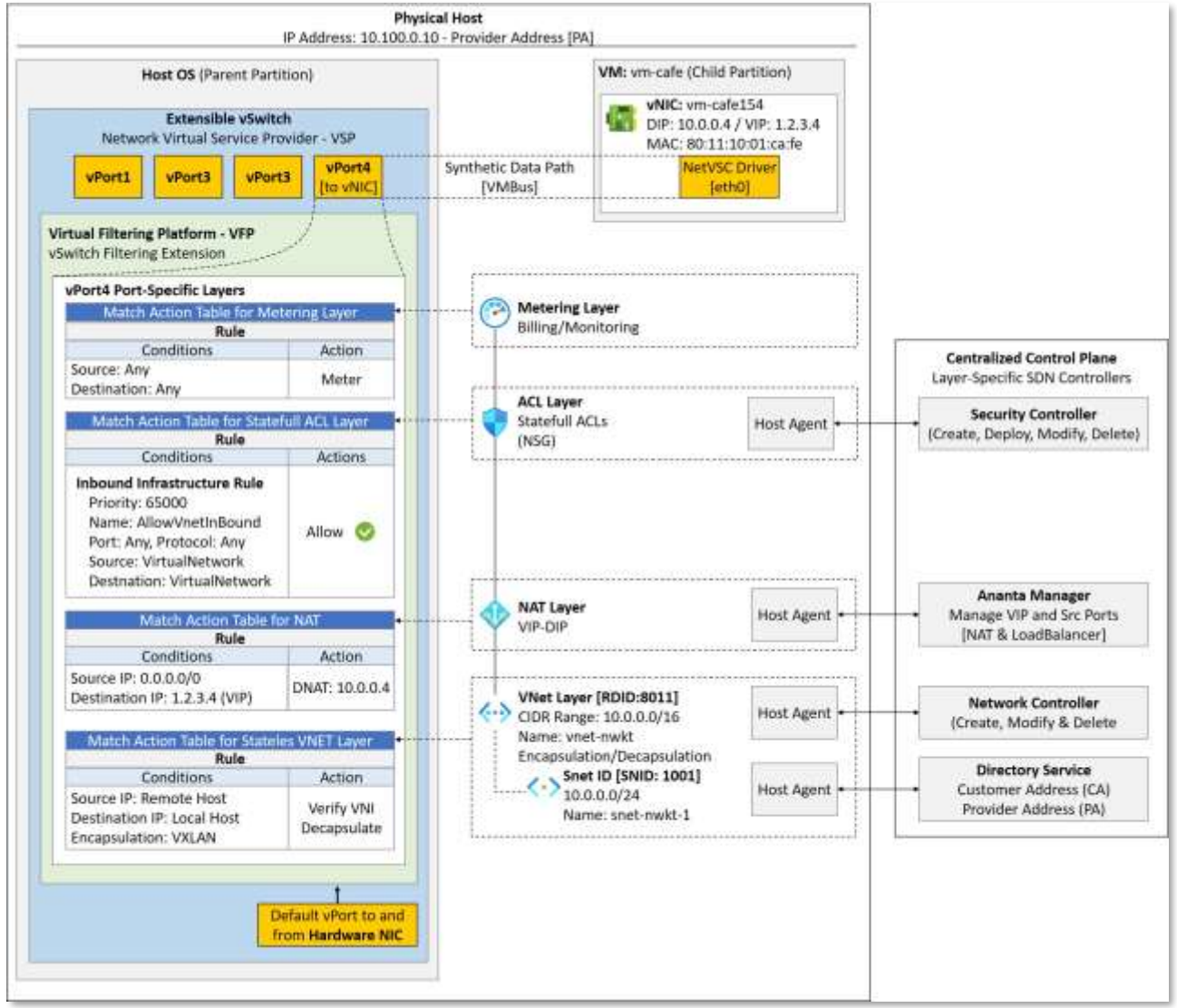
Node / Host


Compute Controller
Goal State
Node Agent
Software load balancer
Goal State
Load Balancer Agent
Network State Manager
Goal State
Network Manager Agent
Rules
ACLs, Routes, Meters
Virtual Network PA-CA
Load Balancing
vSwitch VFP
100.33.88.5
NIC
vNIC 10.1.0.4
vNIC 10.1.0.4
vNIC 10.4.0.22
vNIC 10.1.0.5
vNIC 10.4.0.99
vNIC 10.1.0.6
Host

Packet Encapsulation

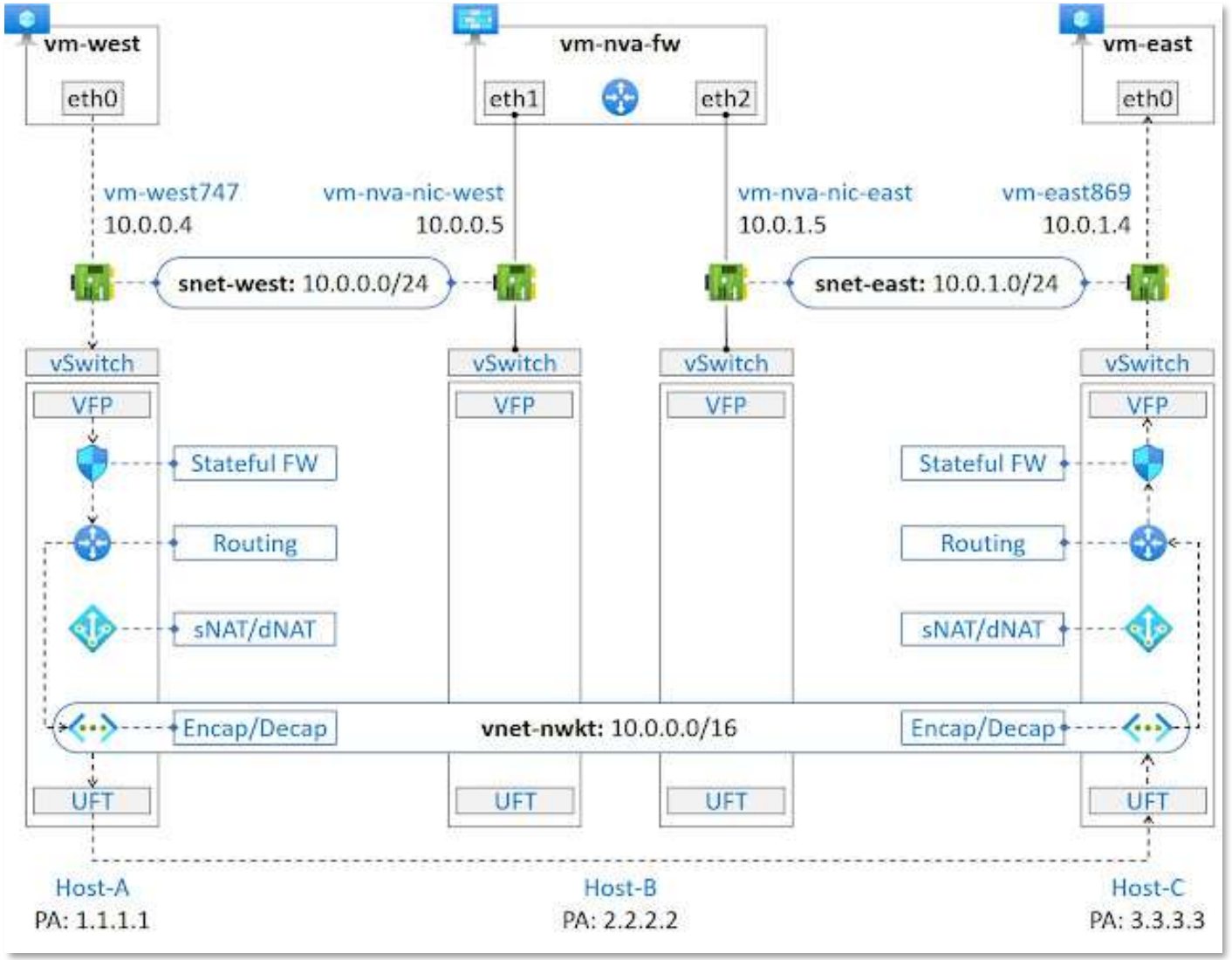


VFP – Virtual Filtering Platform



https://nwktimes.blogspot.com/2023/01/azure-host-based-sdn-part-1-vfp.html

Routing



https://nwktimes.blogspot.com/2023/04/routing-in-azure-subnets.html

More information

- The Cost of a Cloud: Research Problems in Data Centers Networks
  - http://research.microsoft.com/~dmaltz/papers/DC-Costs-CCR-editorial.pdf
- VL2: A Scalable and Flexible Data Center Network
  - https://www.microsoft.com/en-us/research/publication/vl2-a-scalable-and-flexible-data-center-network/
- Towards a Next Generation Data Center Architecture: Scalability and Commoditization
  - http://research.microsoft.com/~dmaltz/papers/monsoon-presto08.pdf
- DCTCP: Efficient Packet Transport for the Commoditized Data Center
  - https://www.microsoft.com/en-us/research/publication/data-center-tcp-dctcp/
- The Nature of Datacenter Traffic: Measurements and Analysis
  - https://www.microsoft.com/en-us/research/publication/the-nature-of-data-center-traffic-measurements-and-analysis/
- What Goes into a Data Center ?
  - https://www.microsoft.com/en-us/research/publication/what-goes-into-a-data-center-sigmetrics-2009-tutorial/

Thank you!