Advanced Computer Networks

CS740

Physical Connectivity Beyond the Data Center

https://pages.cs.wisc.edu/~mgliu/CS740/F26/index.html

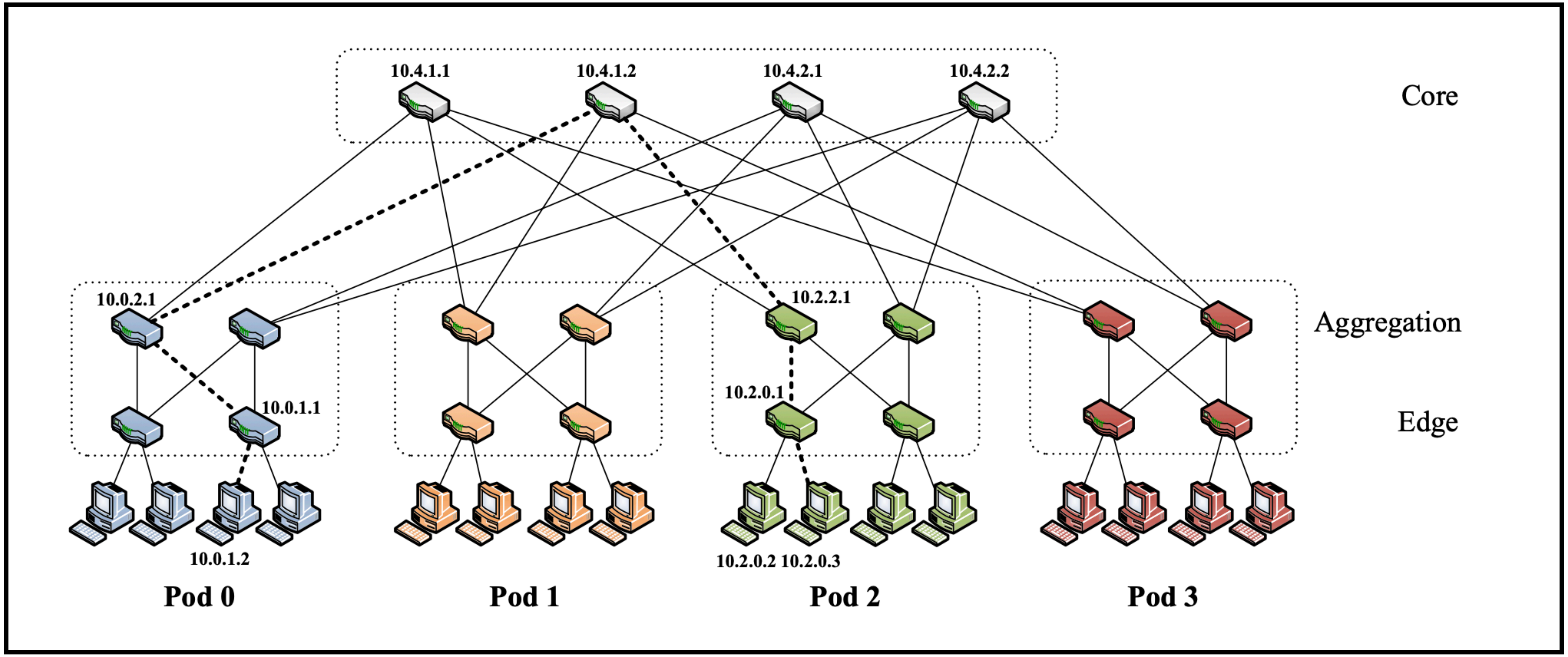
Ming Liu
mgliu@cs.wisc.edu

Outline

- Last lecture
  - Physical connectivity at the rack/cluster scale
- Today
  - Physical connectivity beyond the data center
- Announcements
  - Lab1 due 10/09/2026 11:59 PM

Physical Connectivity Within a Data Center

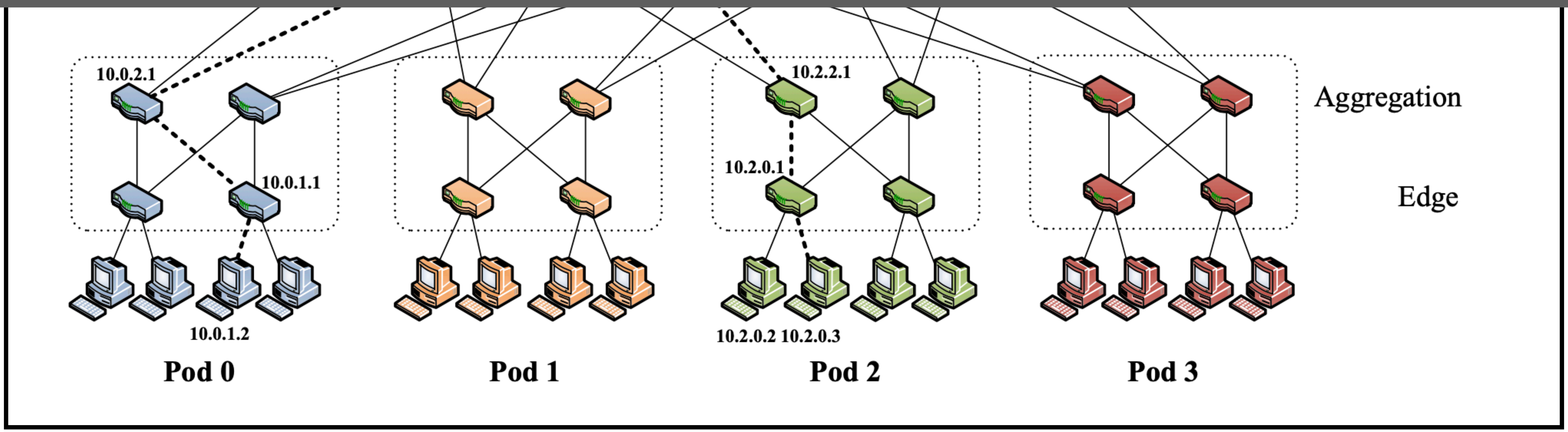
- Multistage switching network
  - Fat-tree topology



Physical Connectivity Within a Data Center

- Multistage switching network
  - Fat-tree topology

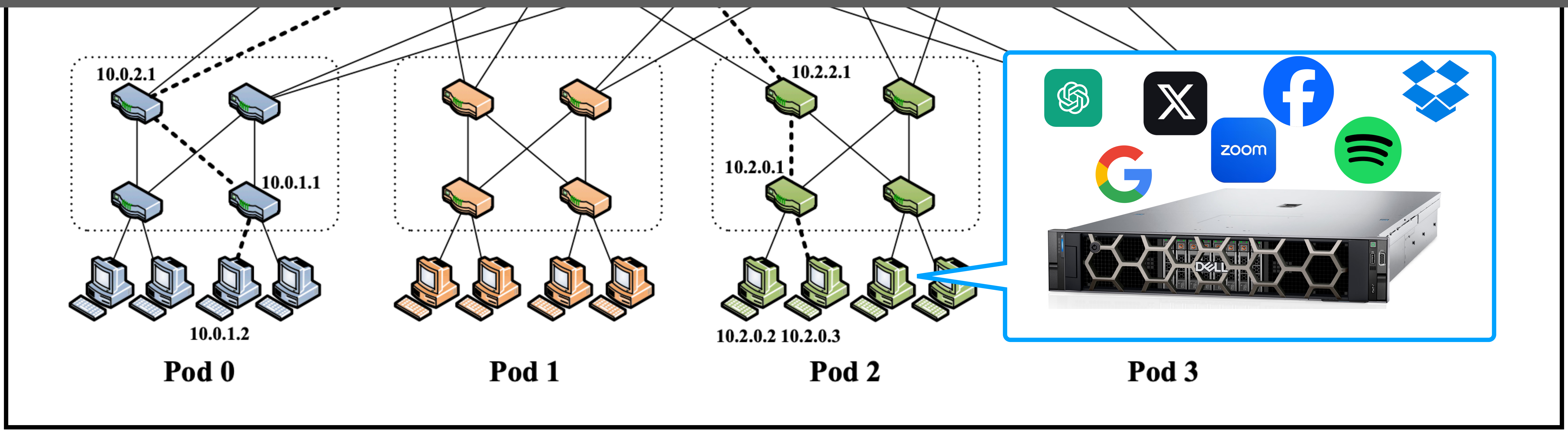
Where do data center services (applications) run?



Physical Connectivity Within a Data Center

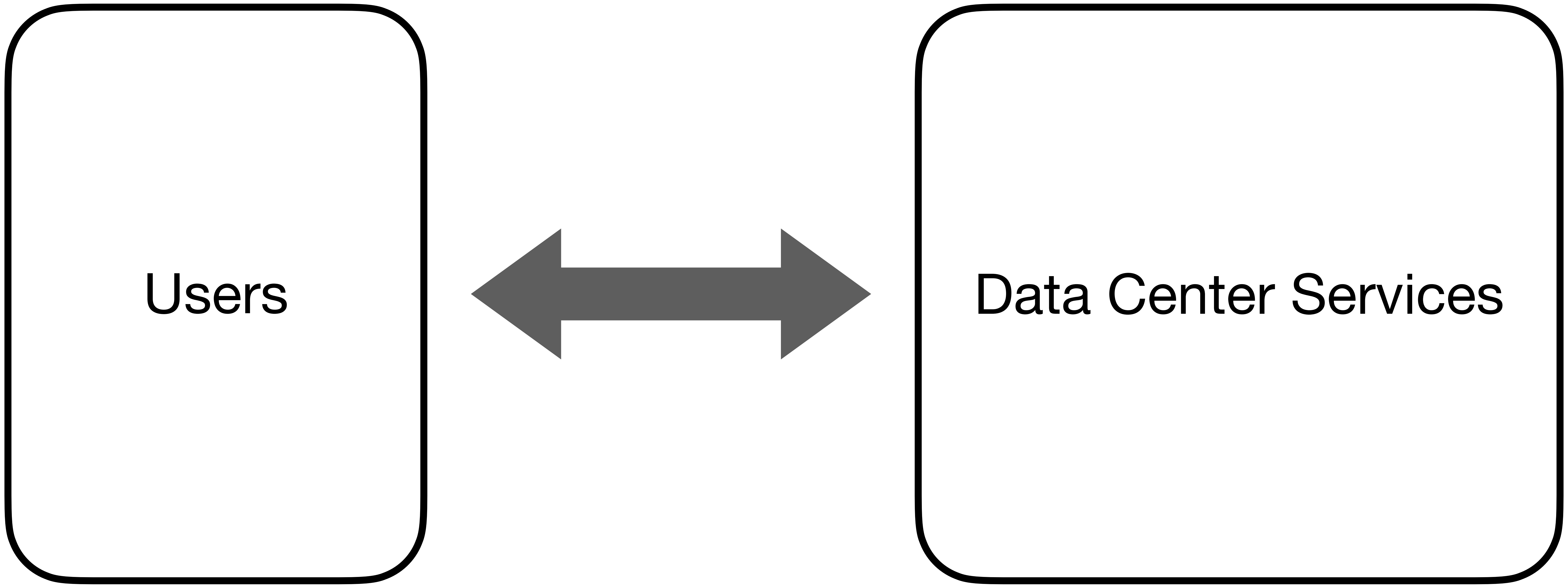
- Multistage switching network
  - Fat-tree topology

Where do data center services (applications) run?



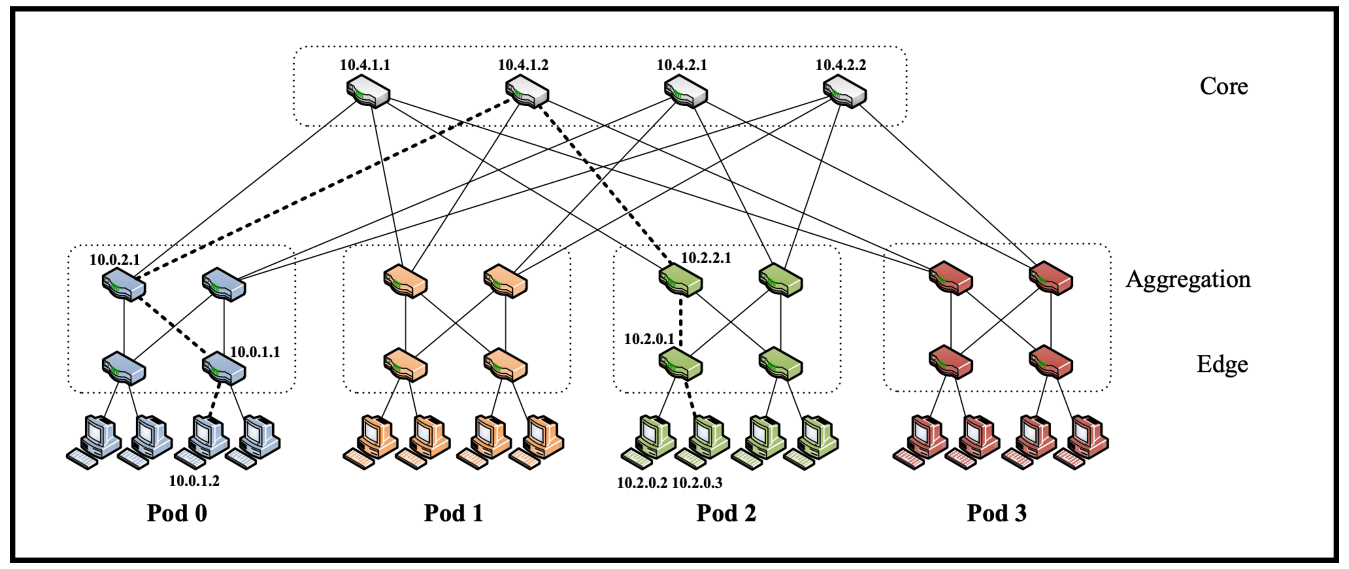
How can we send requests to data center services (applications)?

How can we send requests to data center services (applications)?



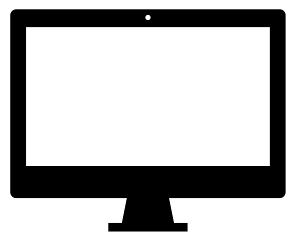
How can we send requests to data center services (applications)?

Data Center Services



Suppose you are doing a Google search on a desktop on Campus, how does the search request enter Google's data center?

Request Communication Path



Request Communication Path

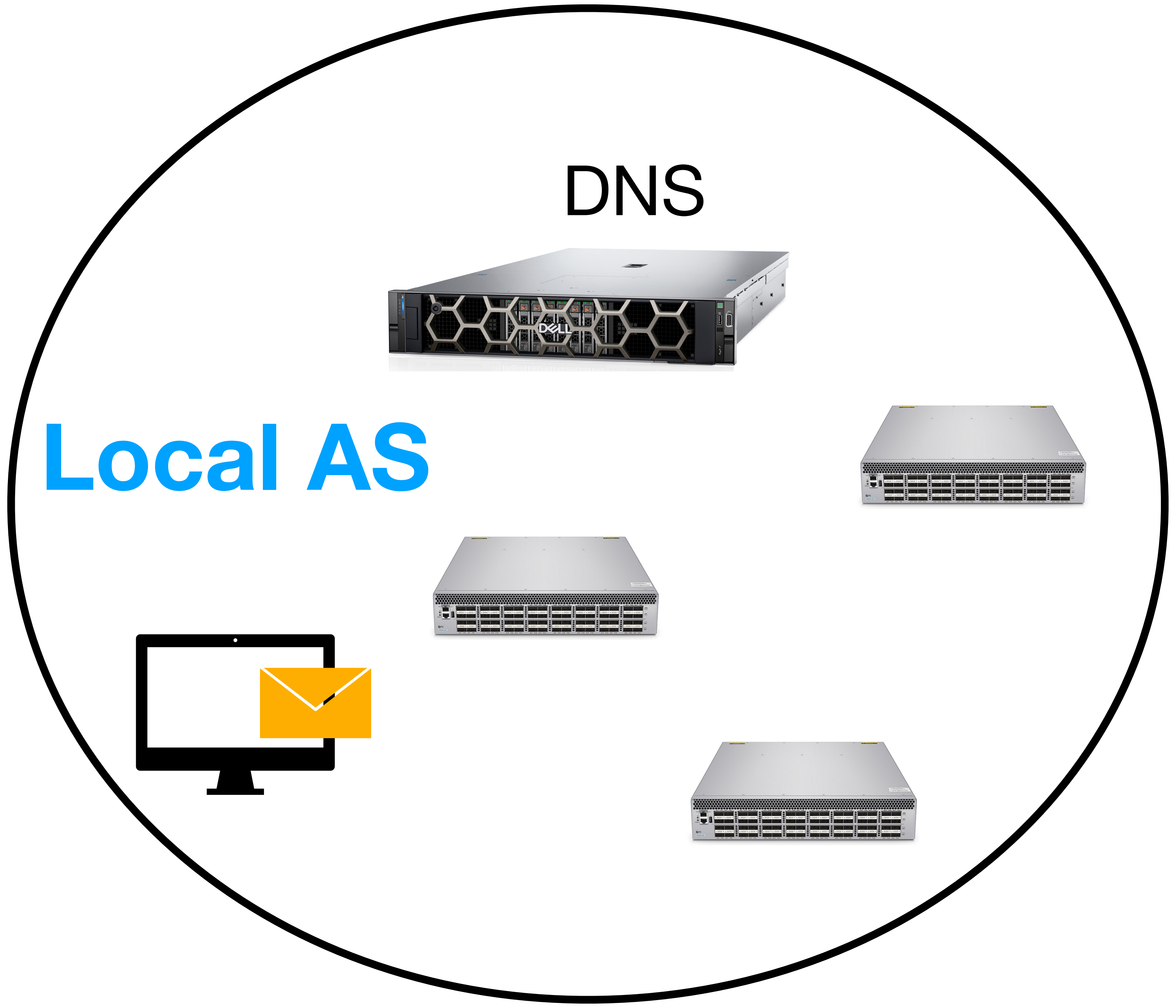
DNS

Request Communication Path

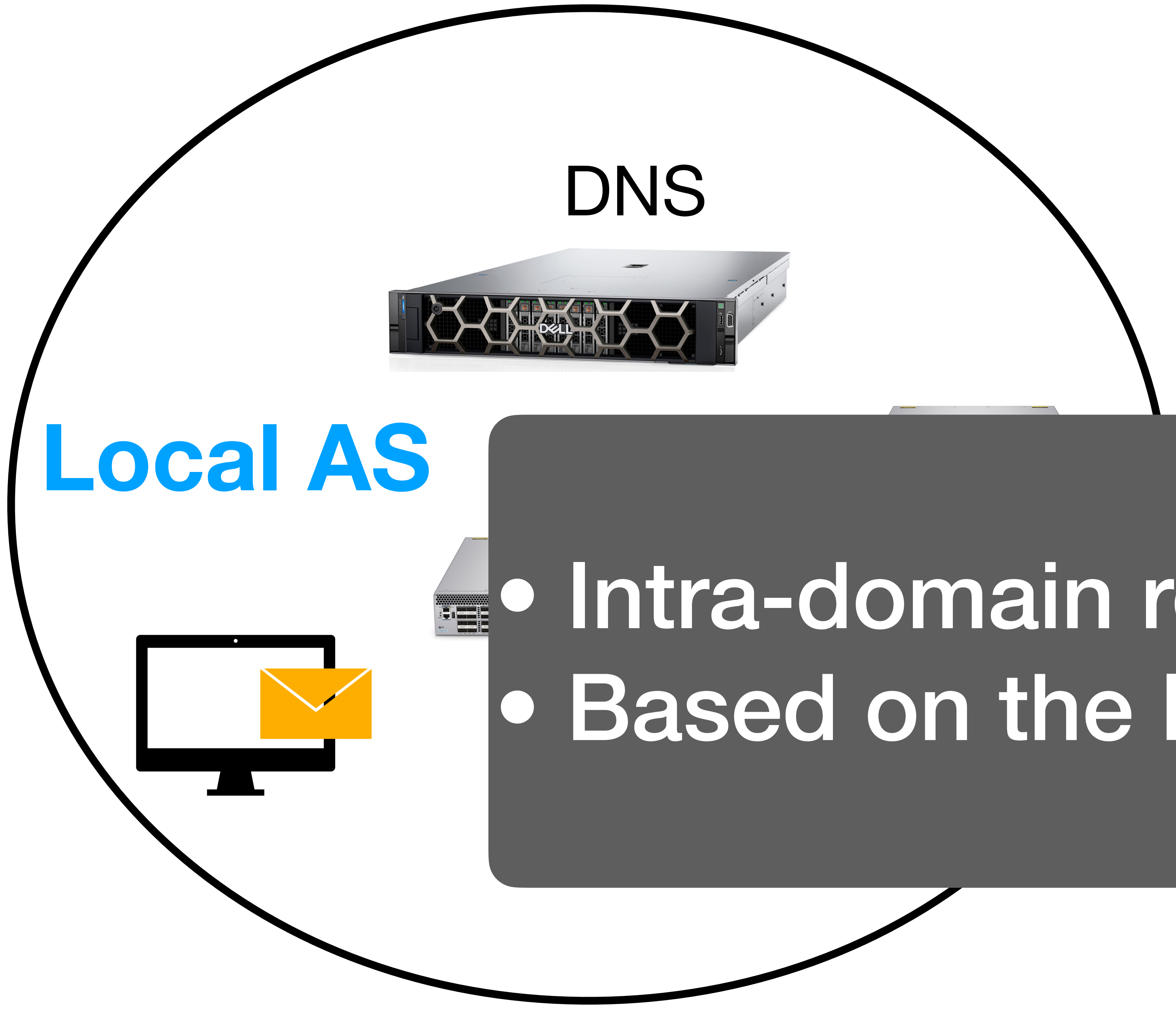
DNS

- A distributed name resolution system
- Organize domain names hierarchically

Request Communication Path

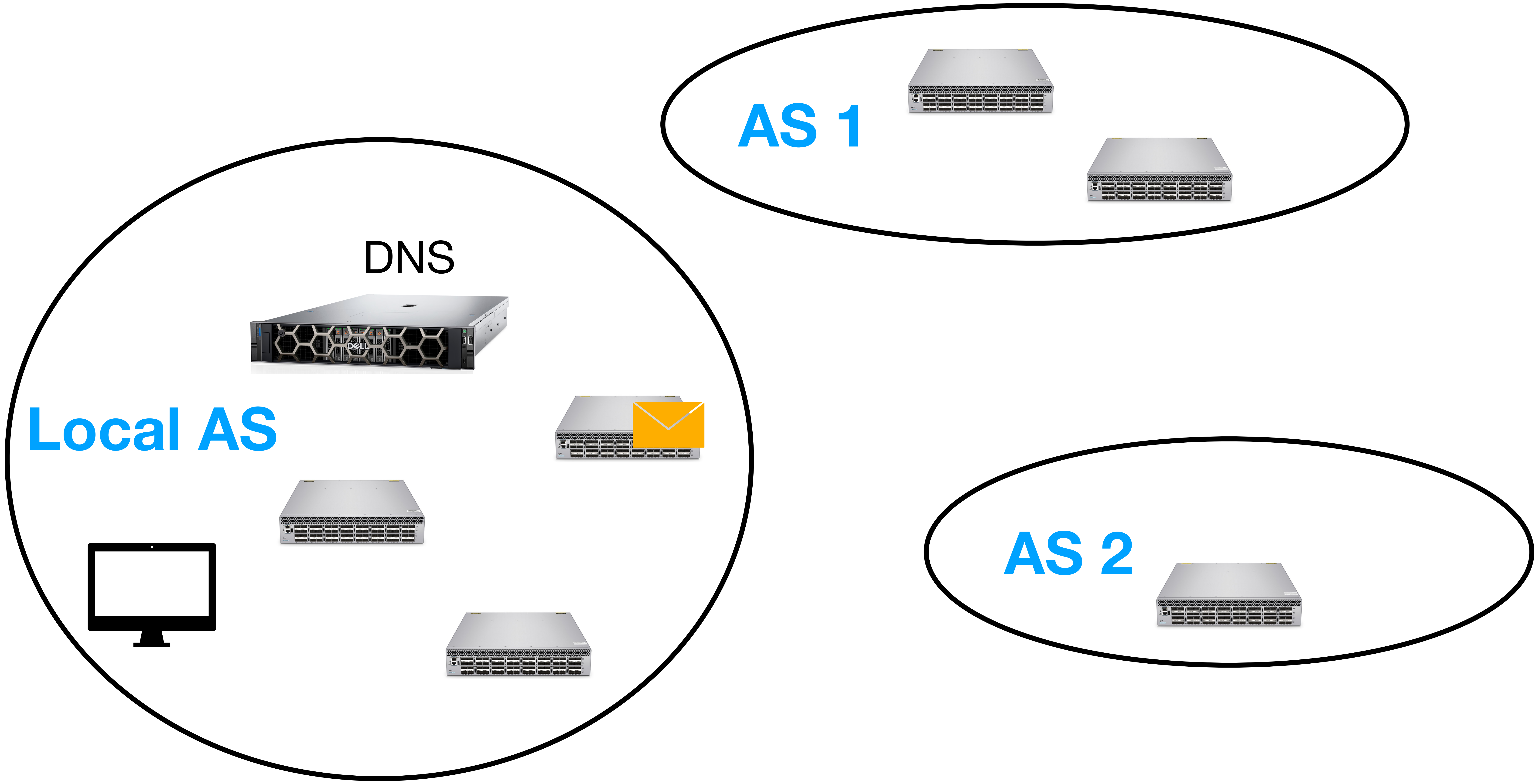


Request Communication Path

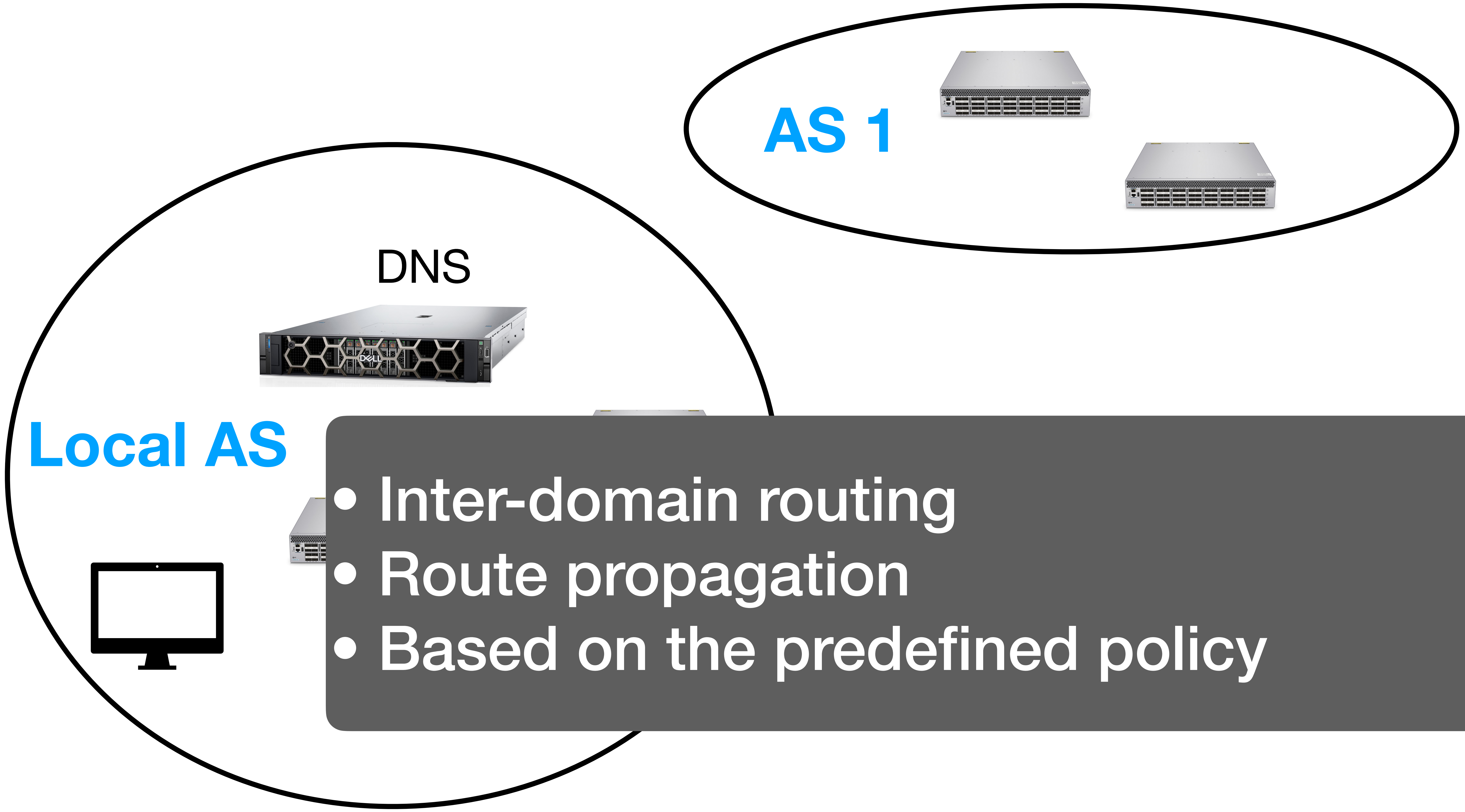


- Intra-domain routing
- Based on the IP address

Request Communication Path

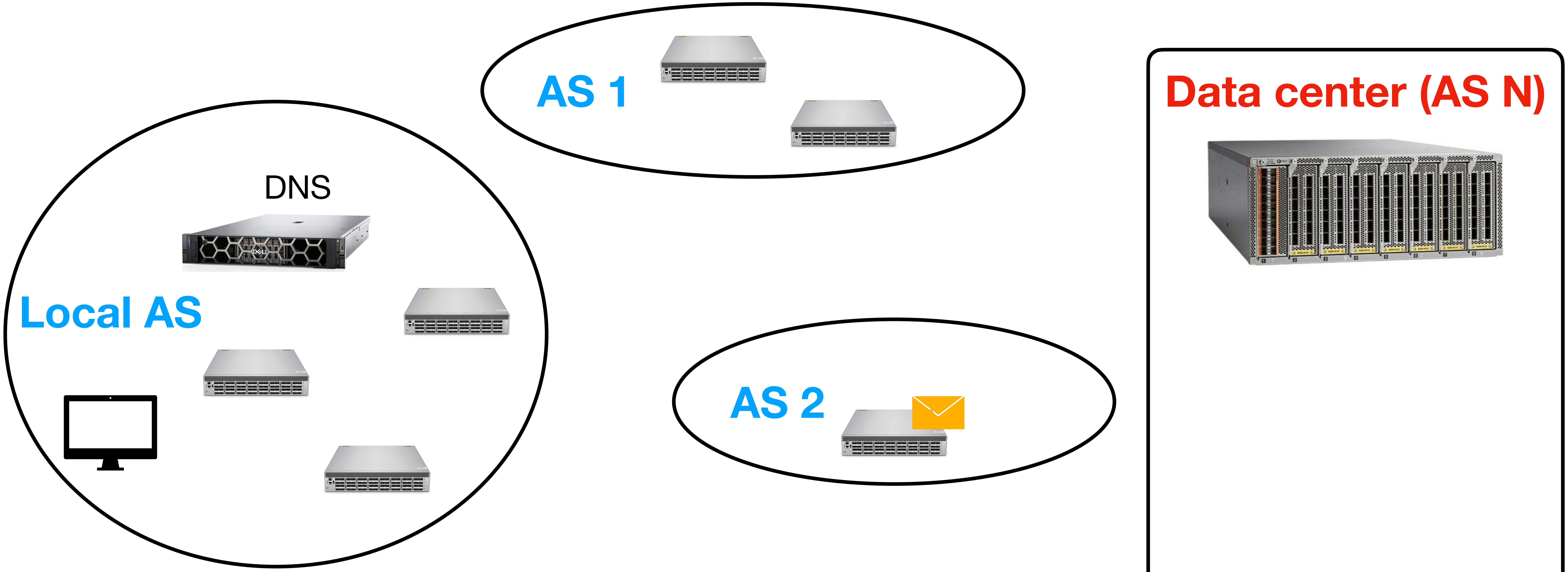


Request Communication Path

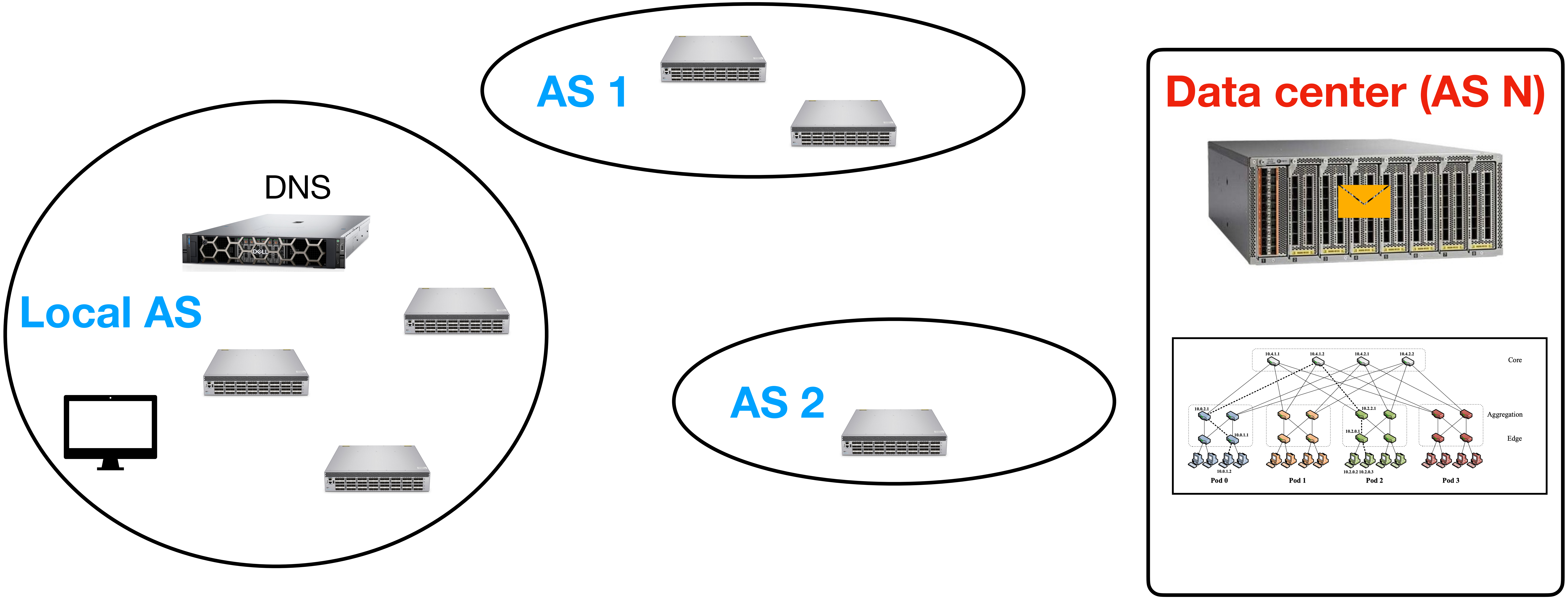


- Inter-domain routing
- Route propagation
- Based on the predefined policy

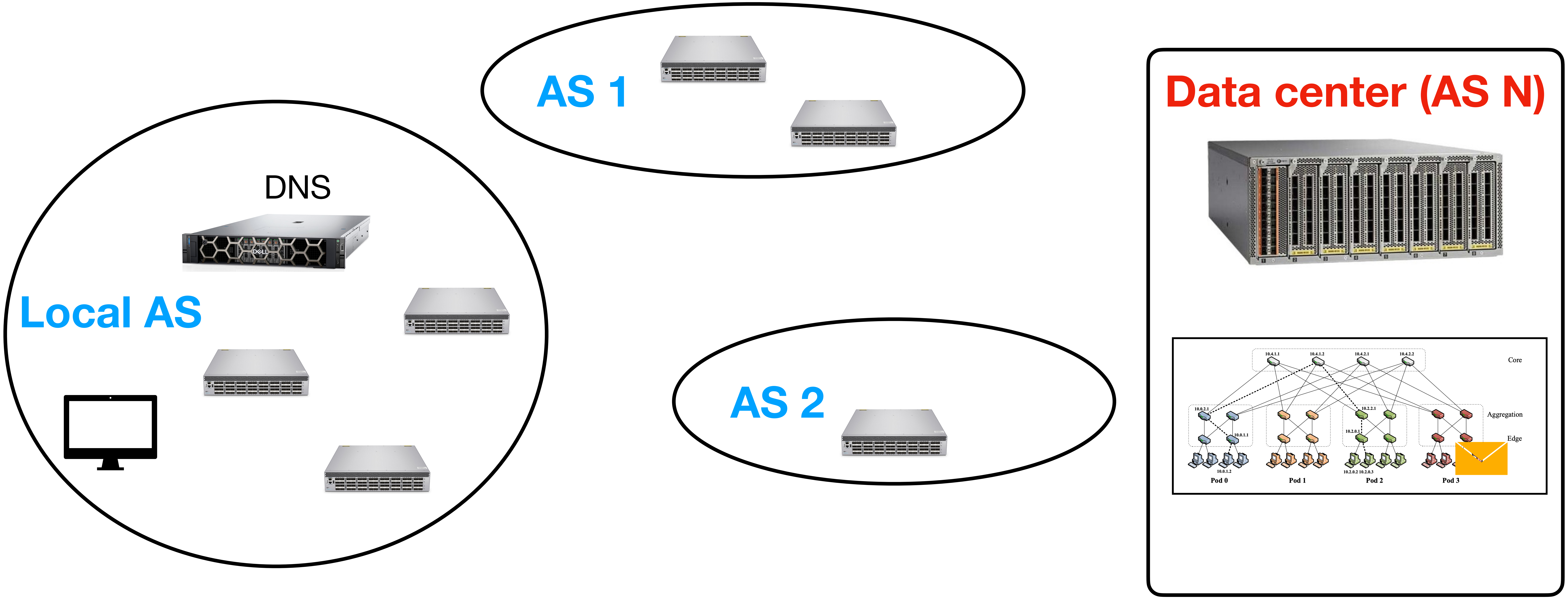
Request Communication Path



Request Communication Path



Request Communication Path

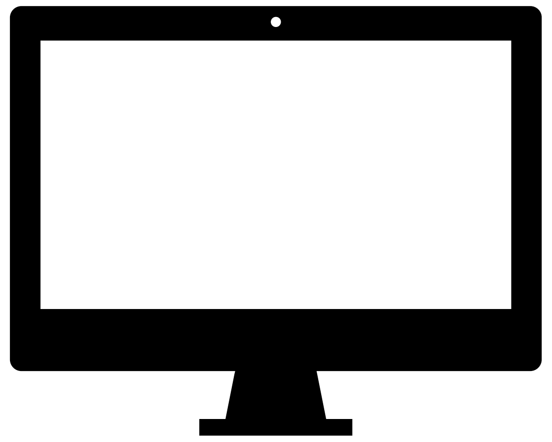


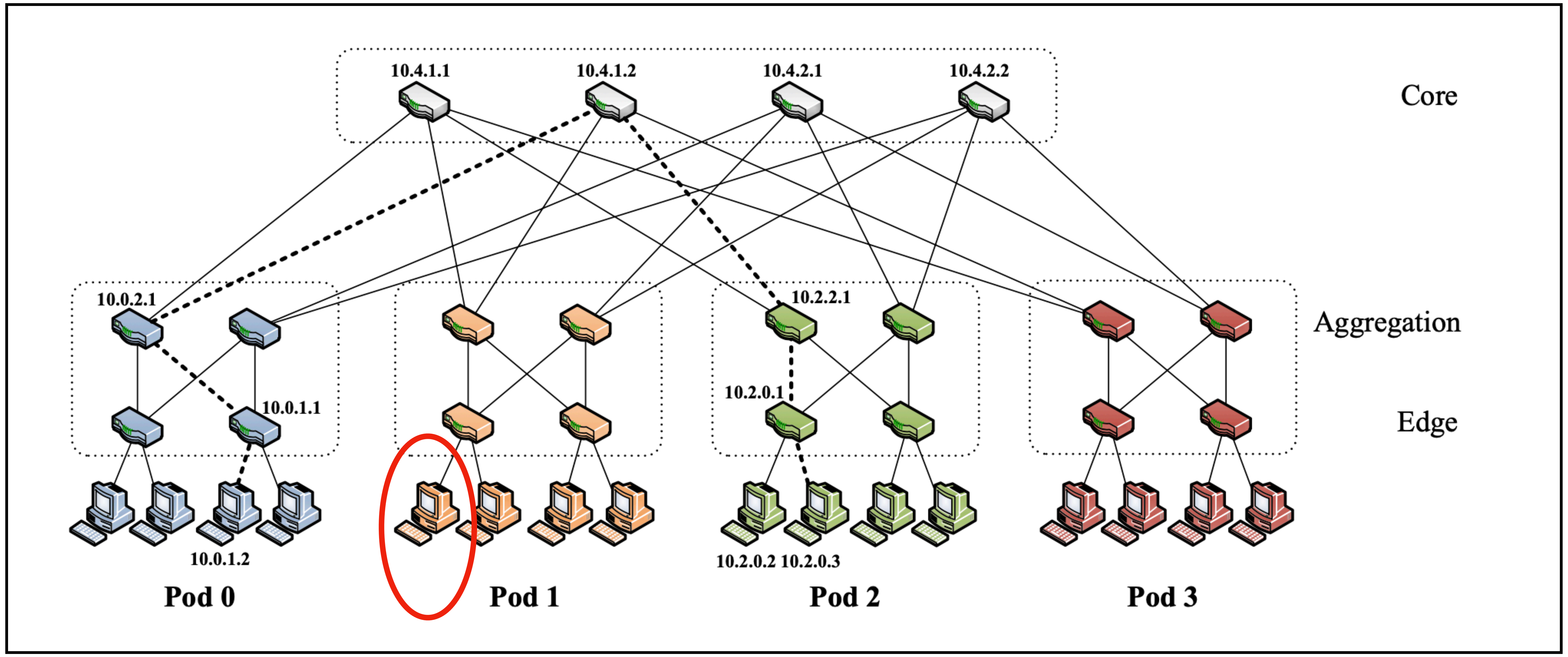
Basic Request Flow

- #1: DNS lookup
  - Local/public DNS server
- #2: Intra-domain routing in an autonomous system (AS)
  - Distance vector, e.g., Routing Information Protocol (RIP)
  - Link state, e.g., Open Shortest Path First (OSPF)
- #3: Inter-domain routing across autonomous systems (ASes)
  - Boarder Gateway Protocol (BGP)
- #4: Enter data center network
  - Through the data center gateway switch
- #5: Routing inside the data center
  - Will be discussed later

Suppose you are issuing a Google search request on a desktop on Campus, what is the source and destination entity?

Any Issues?



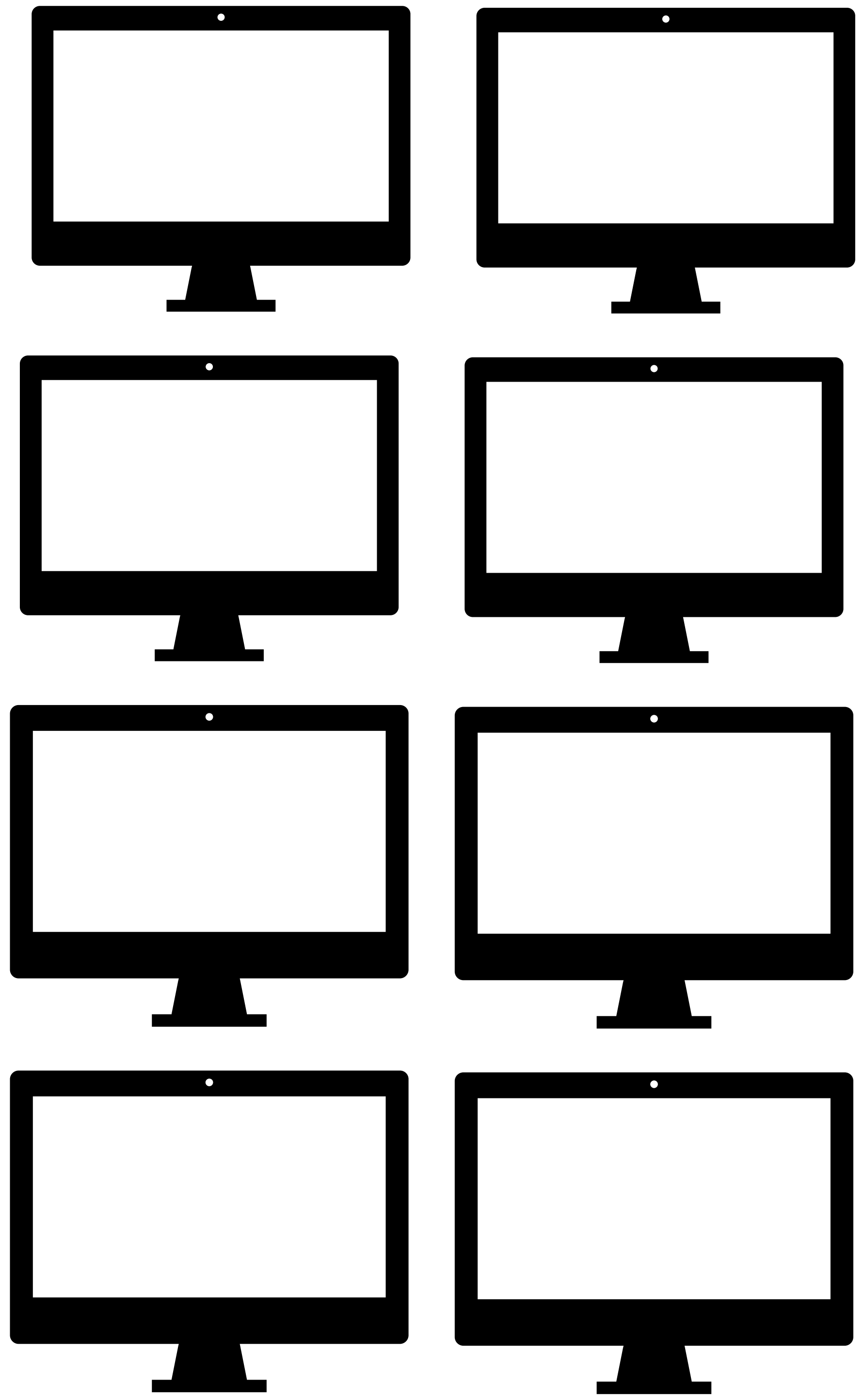


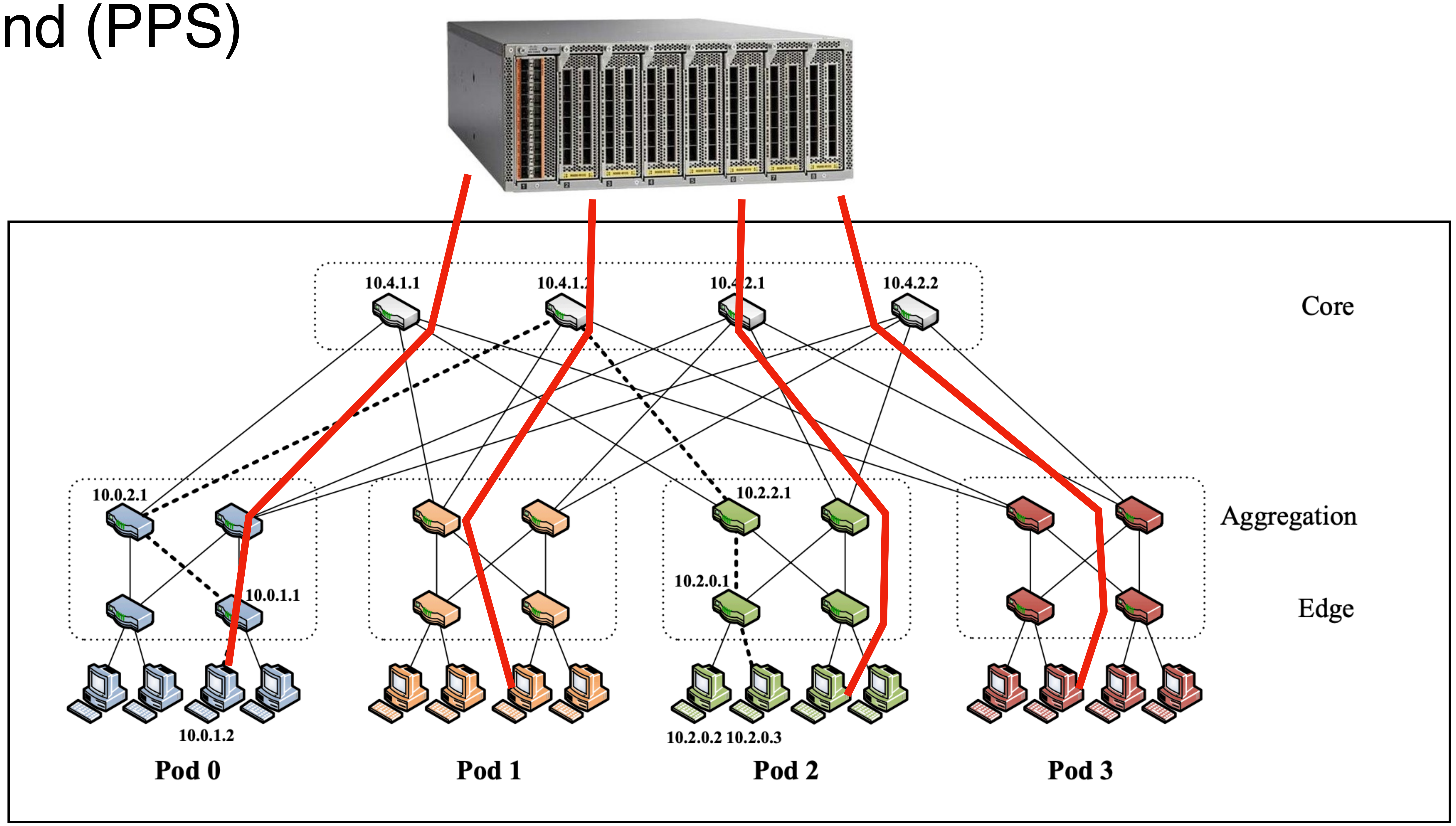
Source: Desktop

Destination: One server in Pod1

Bandwidth Bottleneck

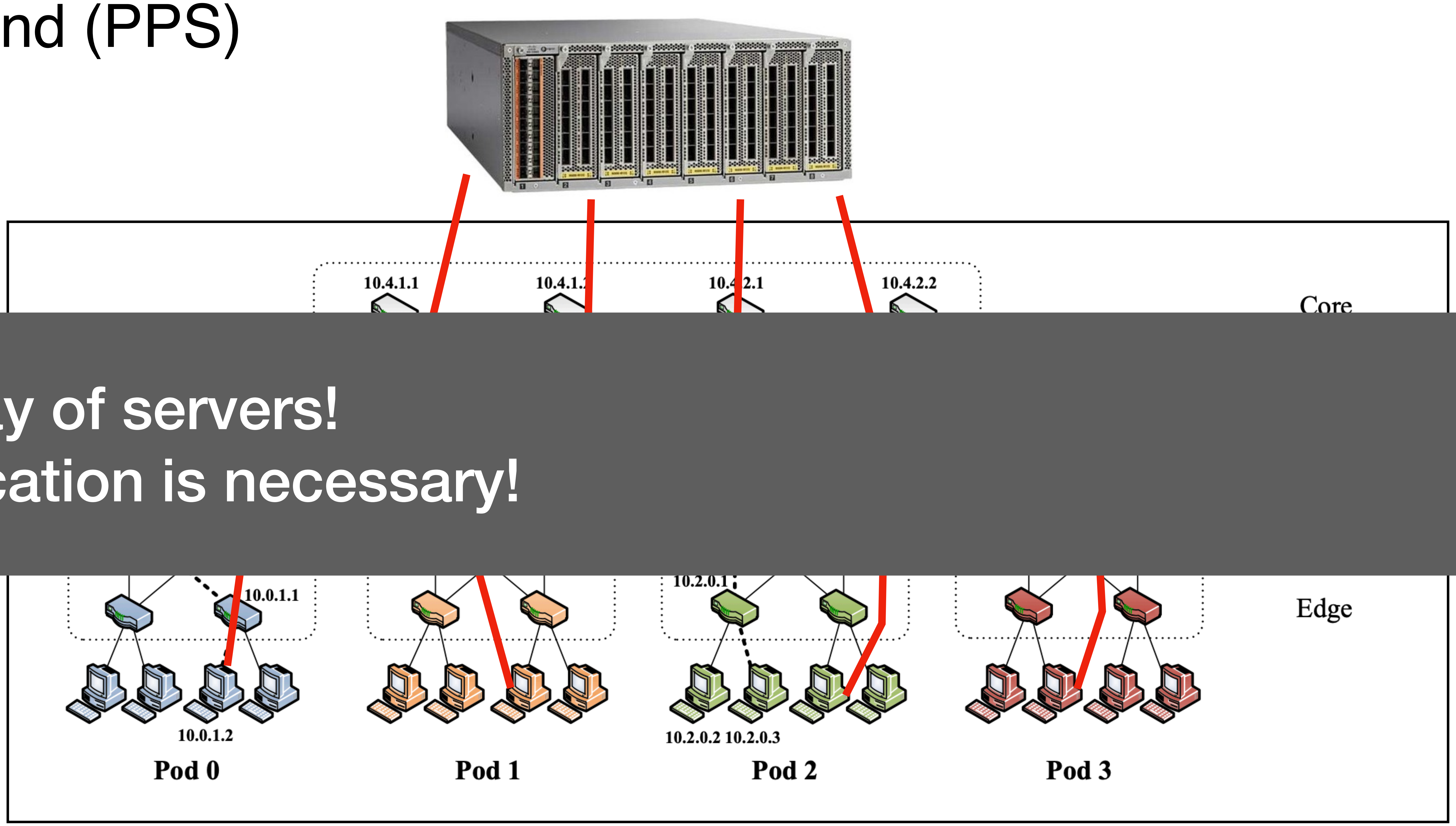
- Bandwidth scaling requires multiple communication paths
  - A NIC port has fixed bandwidth!
  - Packets per second (PPS)





Bandwidth Bottleneck

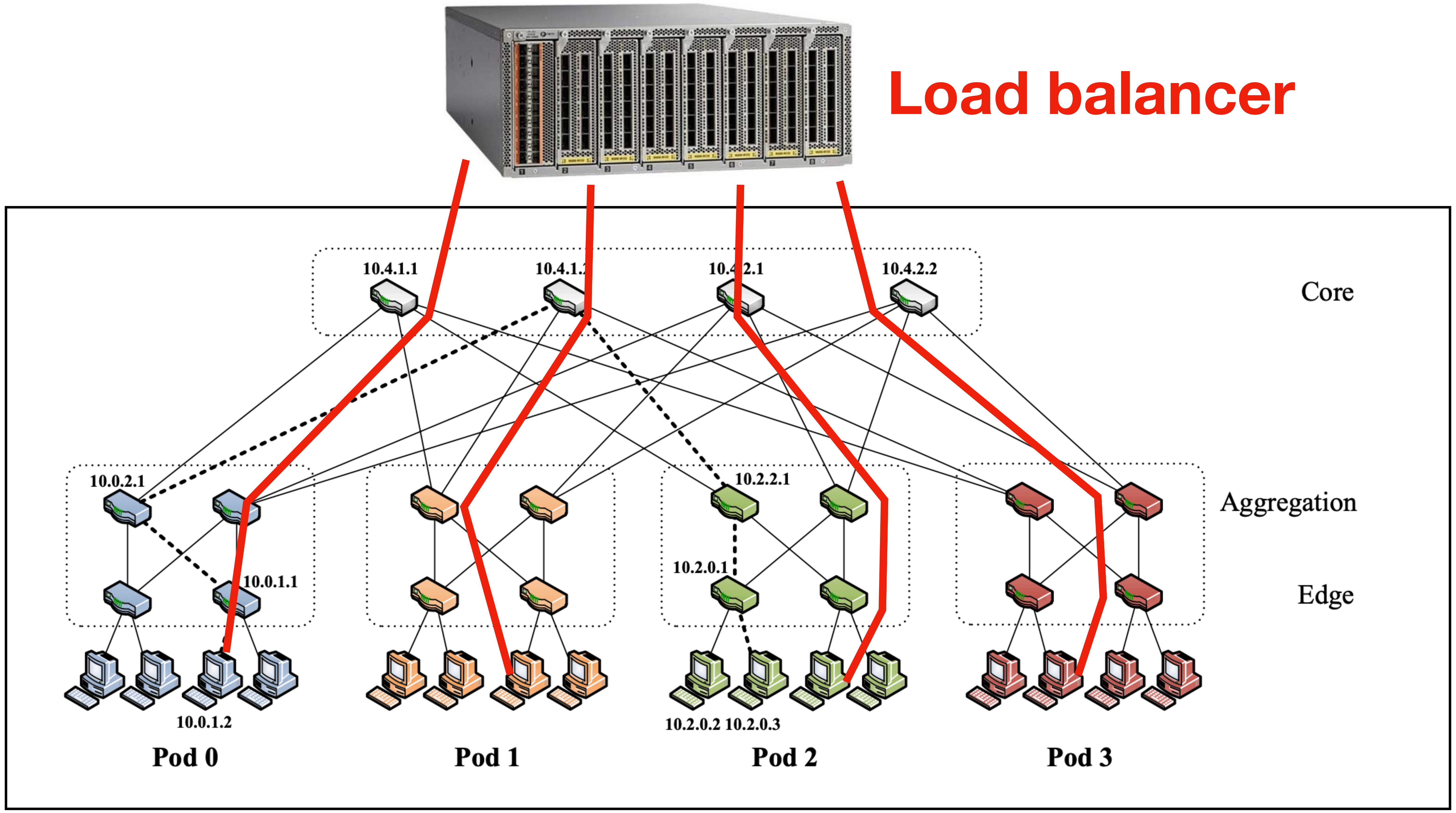
- Bandwidth scaling requires multiple communication paths
  - A NIC port has fixed bandwidth!
  - Packets per second (PPS)



- Destination is an array of servers!
- Multi-path communication is necessary!

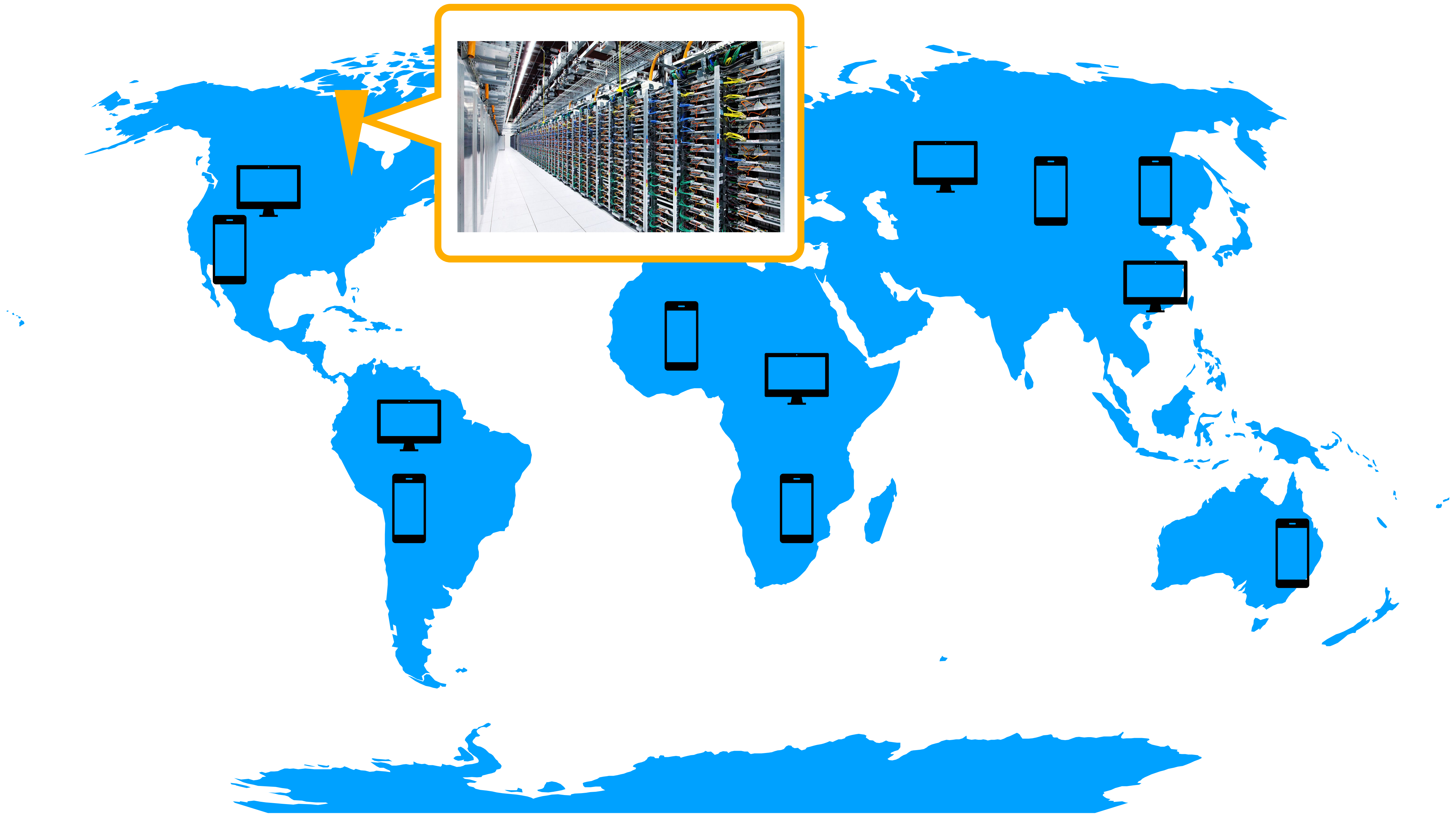
Physical Connectivity Cross Data Centers

- User-facing data center services need multiple physical paths.
  - A load balancer colocated with the gateway



Suppose a data center is located in Madison, how can we achieve consistently low access delay?

Suppose a data center is located in Madison, how can we achieve consistently low access delay?



What does the delay include?

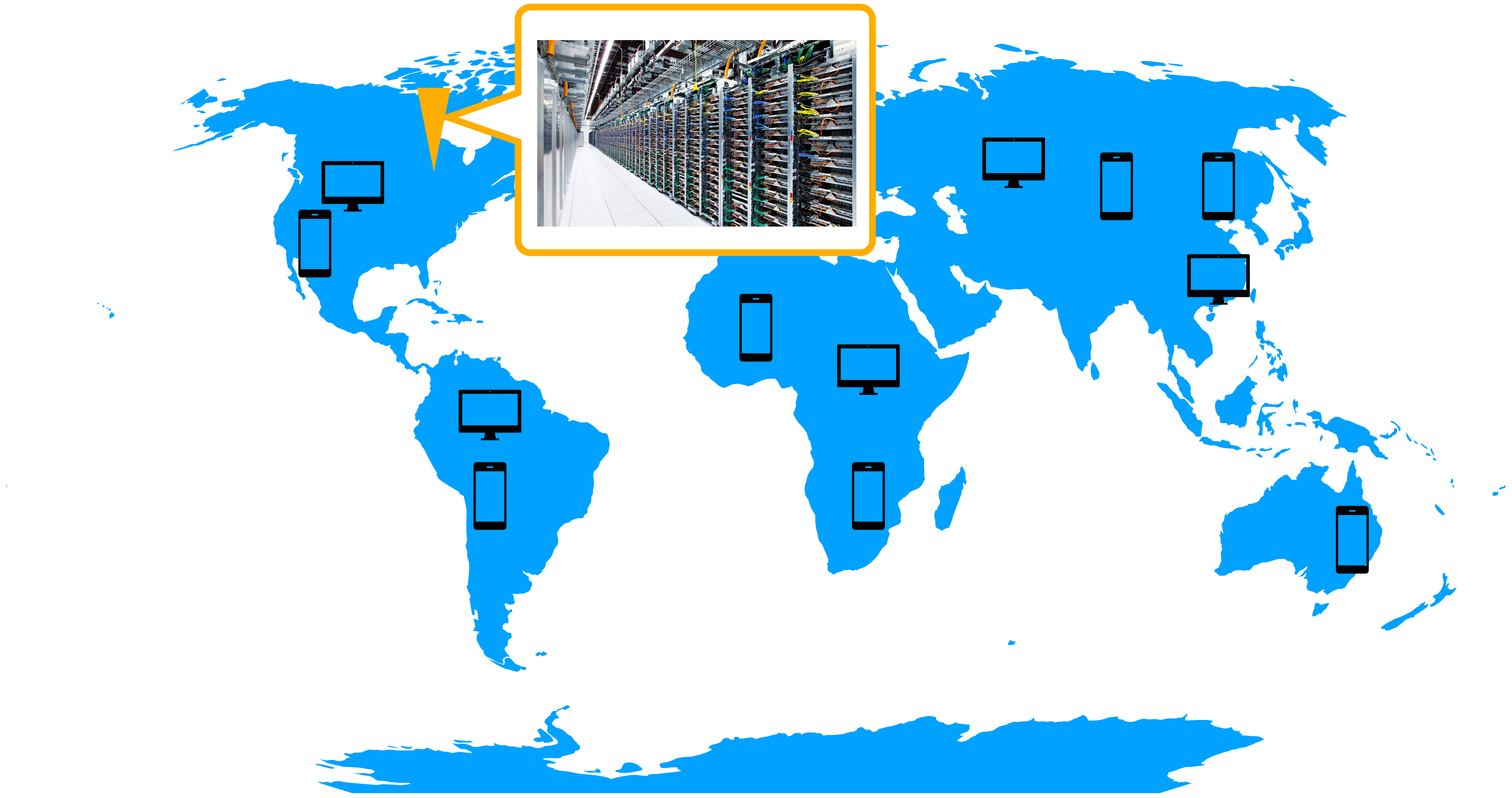
Network Delay

- #1: Transmit delay
  - The total amount of time required to move data between wire and port
- #2: Propagation delay
  - The amount of time required to propagate bits from one point to another
- #3: Queuing delay
  - The amount of time required to stay in the switch/router queue

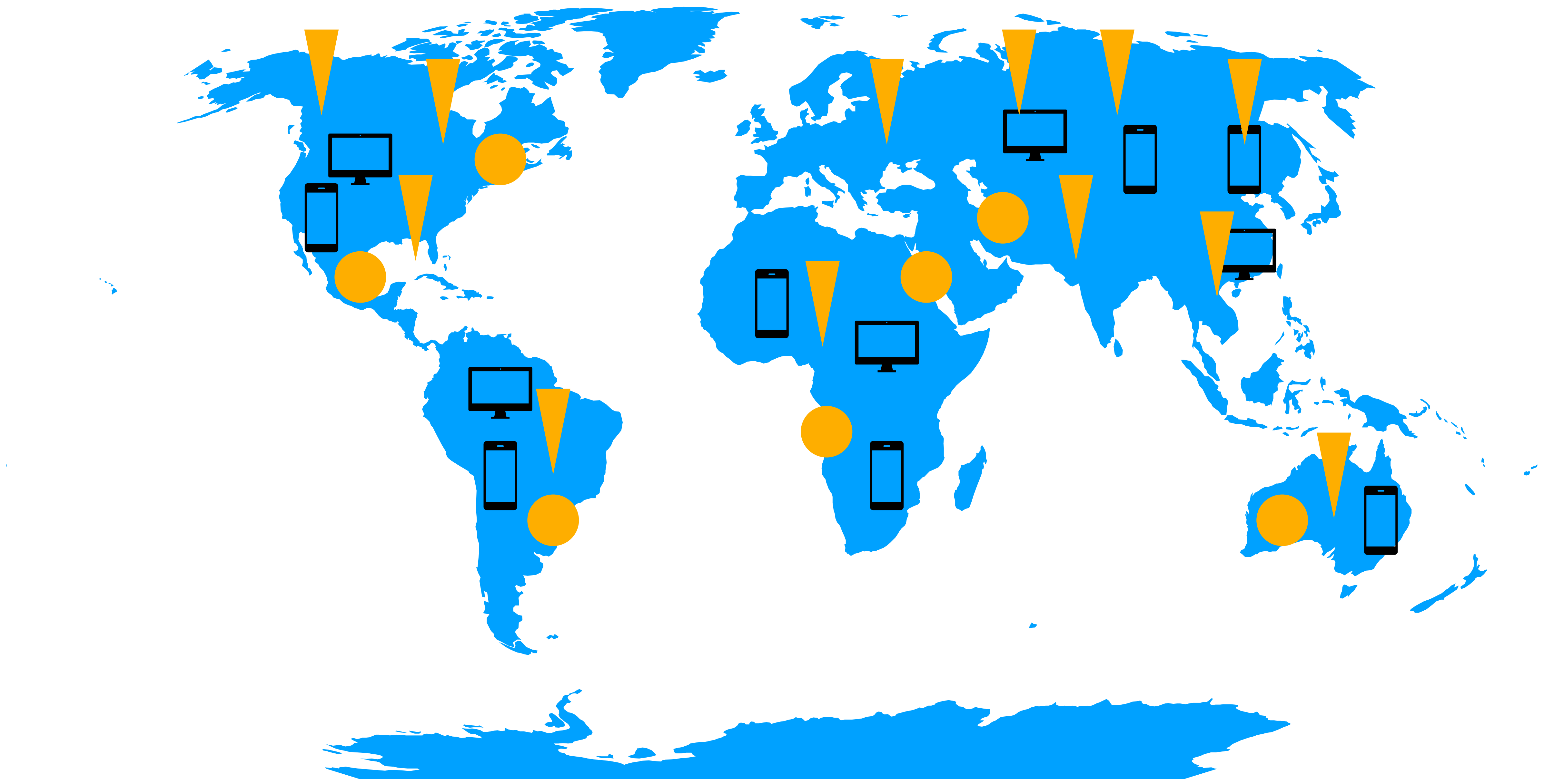
Network Delay

- #1: Transmit delay
  - The total amount of time required to move data between wire and port
- #2: Propagation delay
  - The amount of time required to propagate bits from one point to another
- #3: Queuing delay
  - The amount of time required to stay in the switch/router queue

Suppose a data center is located in Madison, how can we achieve consistently low access delay?

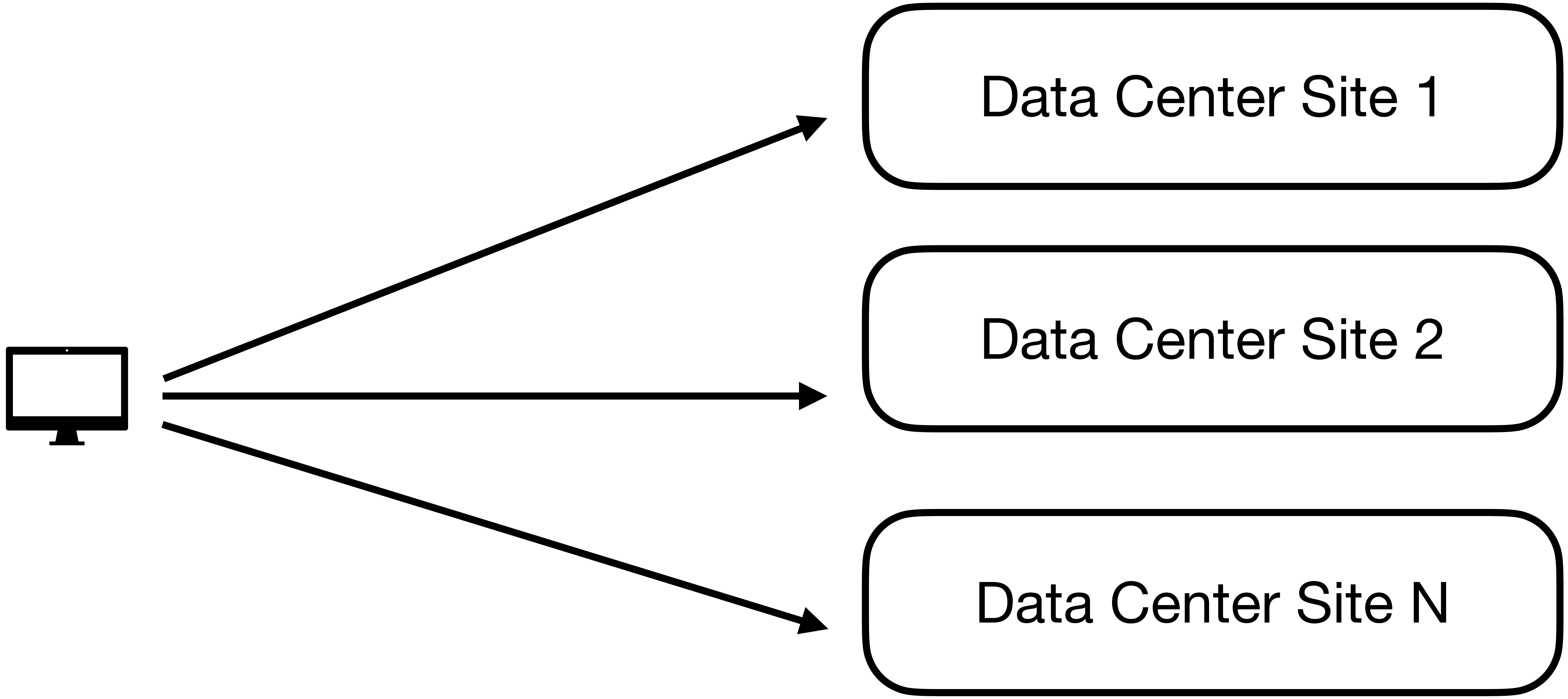


Replication and Caching!



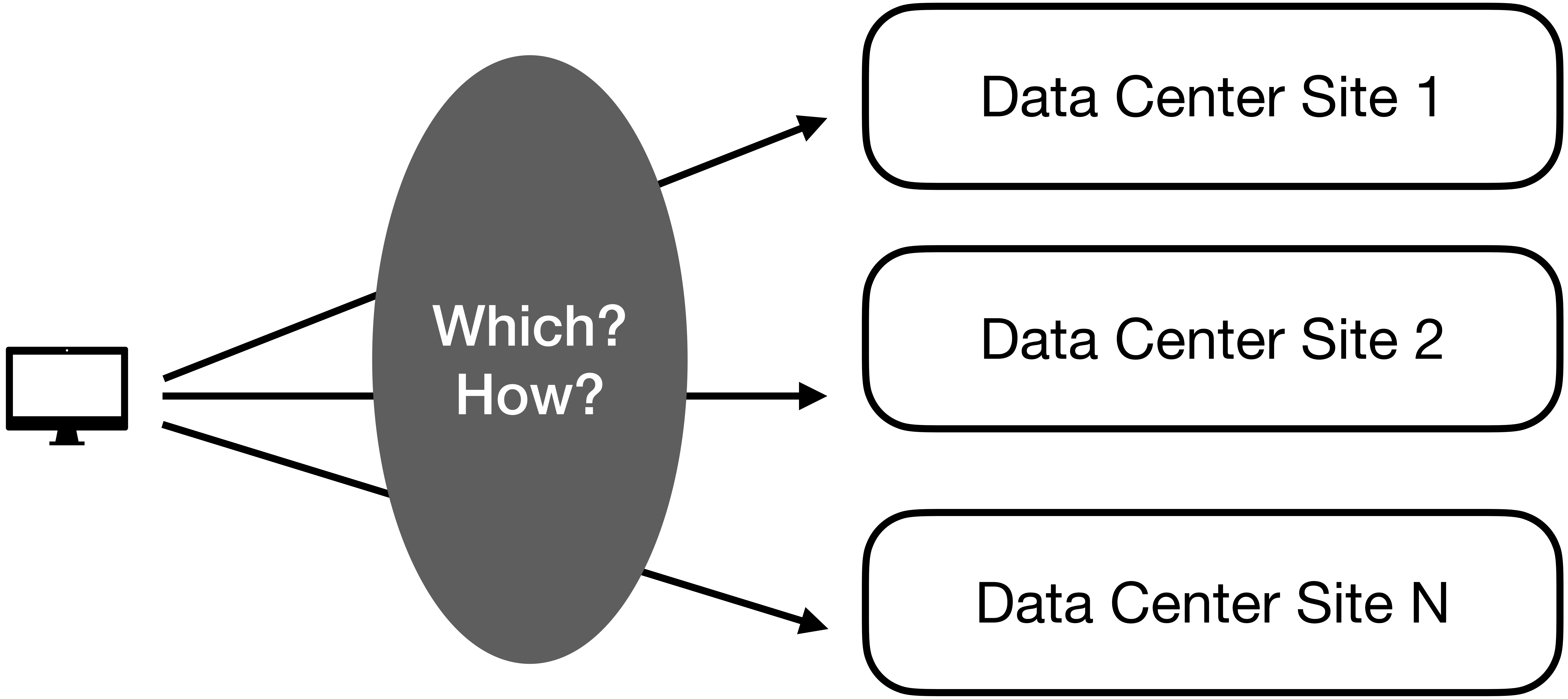
Approach #1: Add Multiple Data Center Sites

- Expand the network, compute, and storage capacity
  - On-demand replication



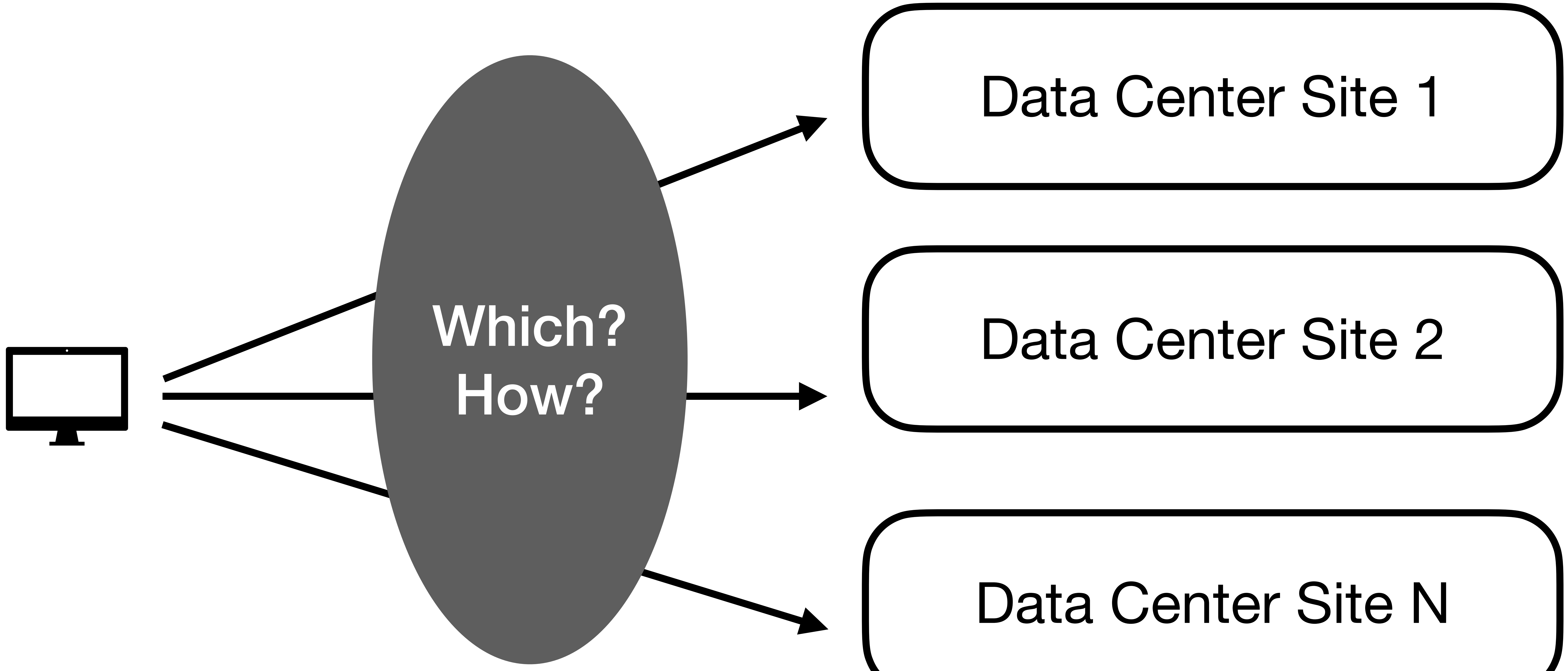
Approach #1: Add Multiple Data Center Sites

- Expand the network, compute, and storage capacity
  - On-demand replication



Approach #1: Add Multiple Data Center Sites

- Expand the network, compute, and storage capacity
  - On-demand replication

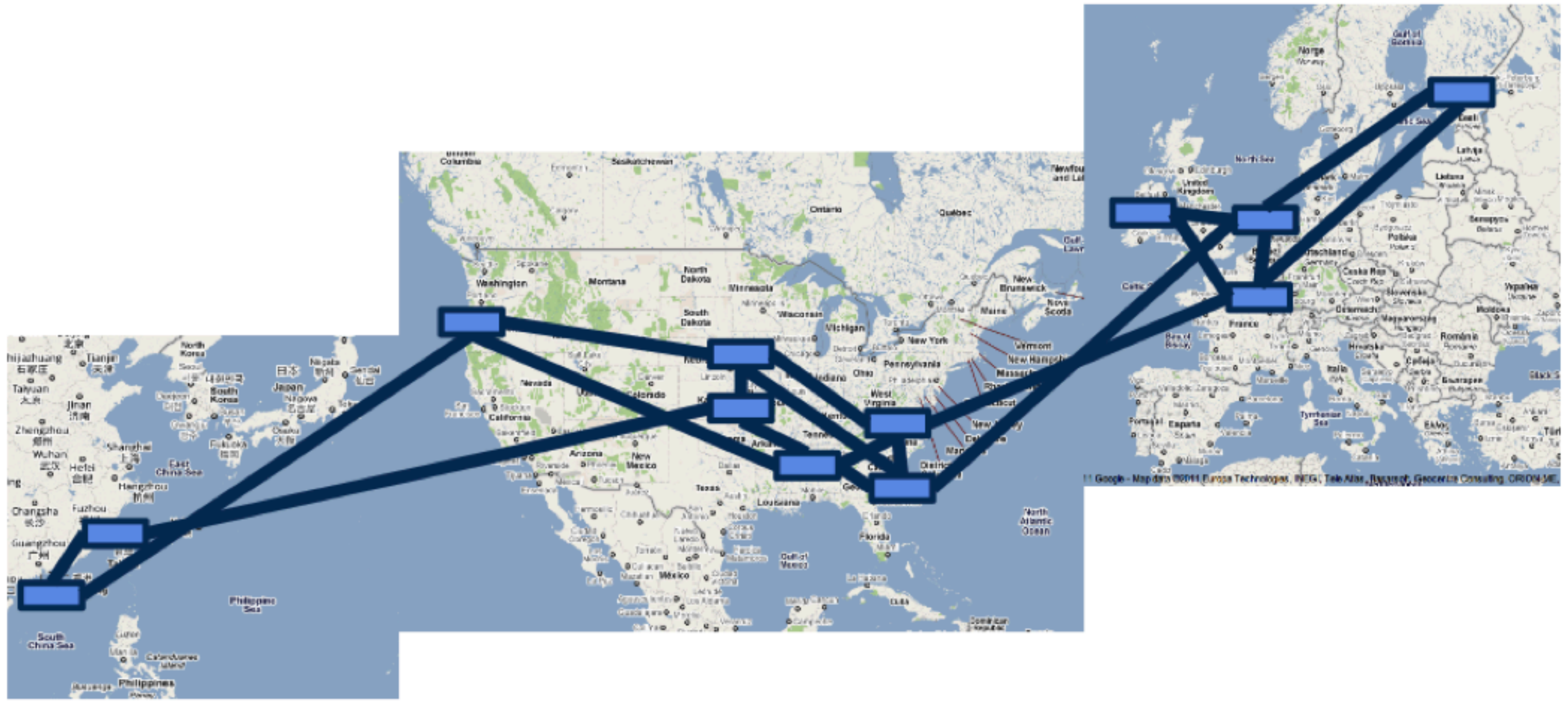


Load balancer again!

Do we connect multiple data center sites or keep them isolated?

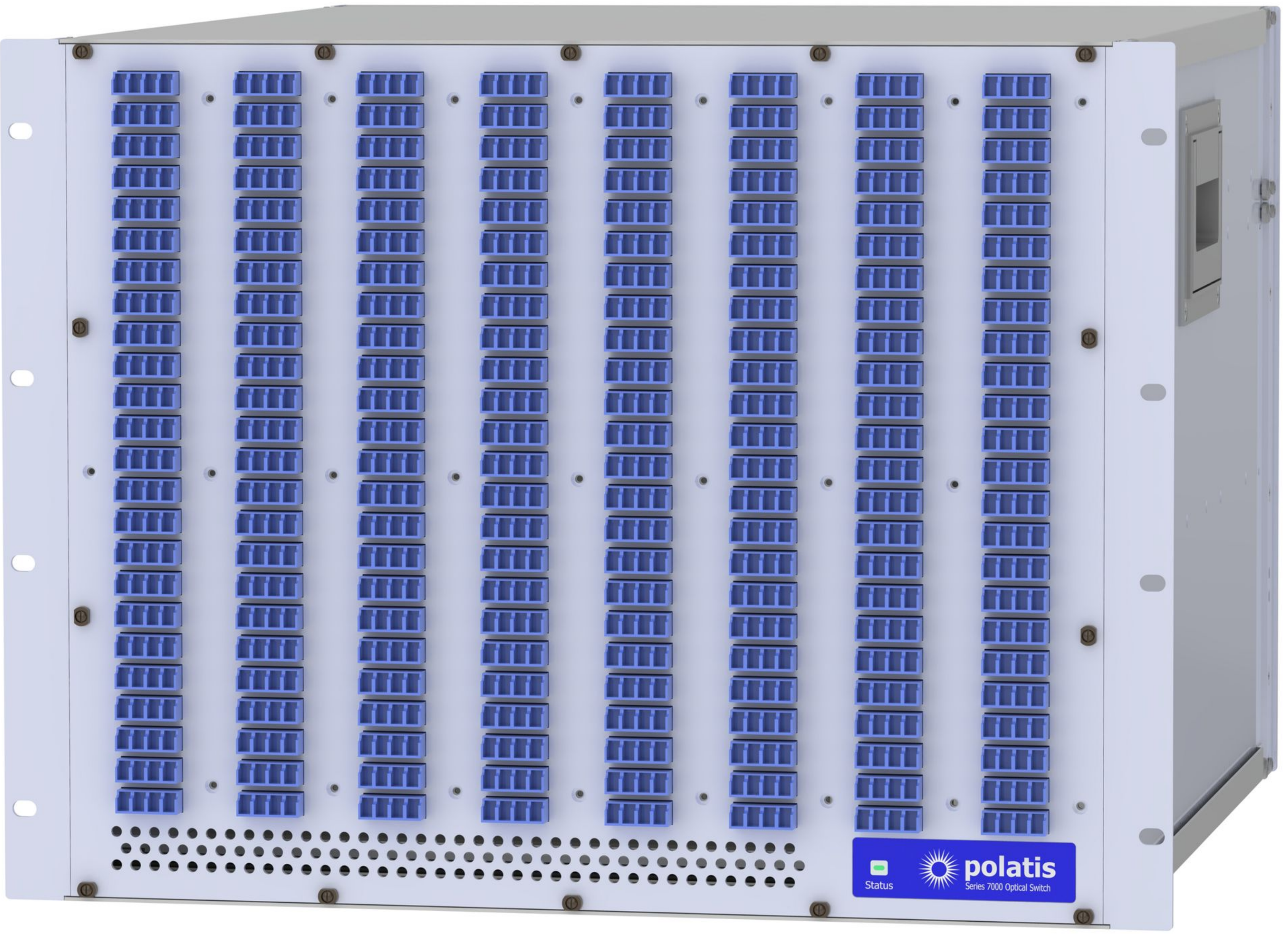
Wide Area Network (WAN) for Data Centers

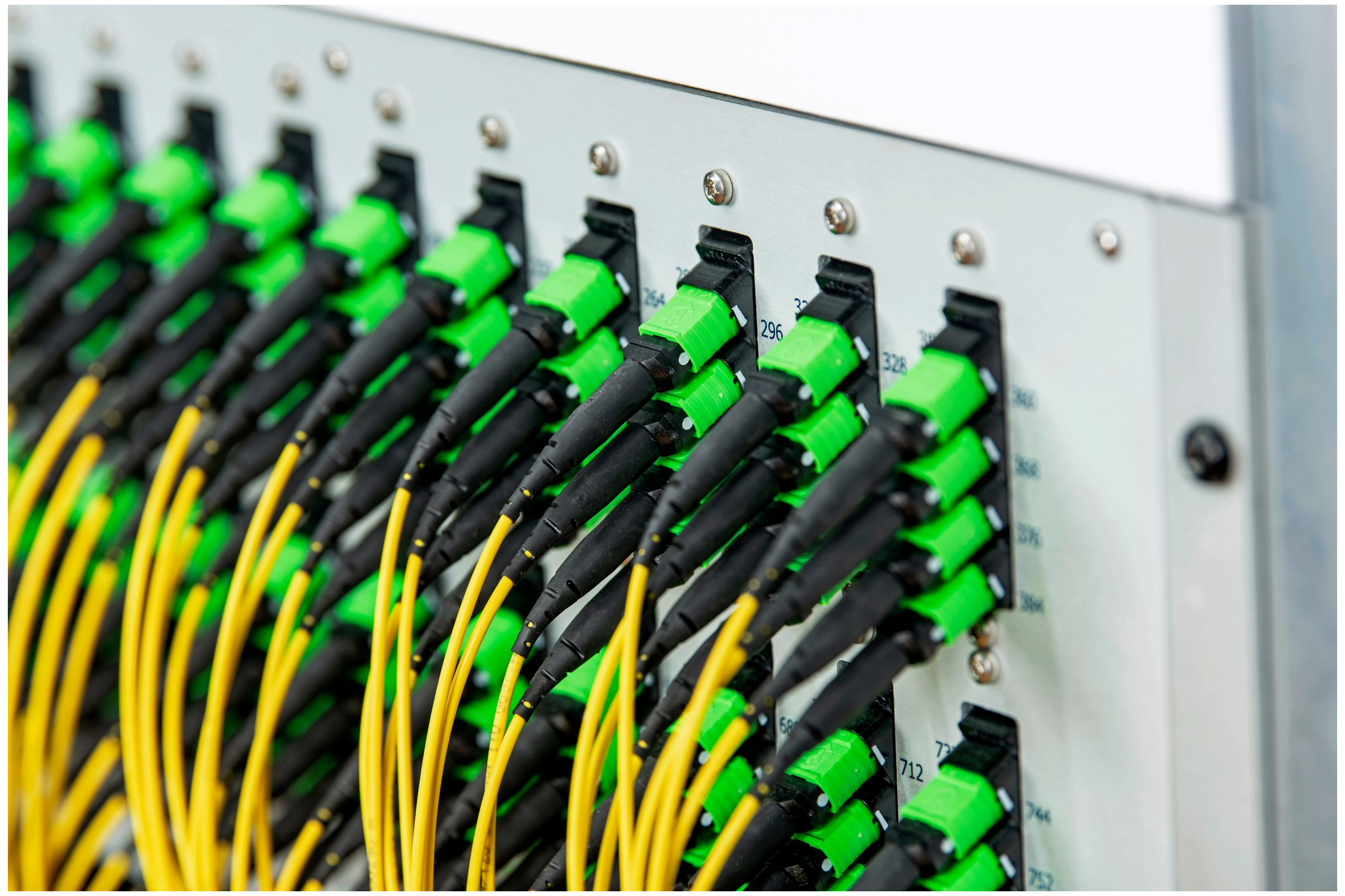
- Why?
  - #1: Data storage is not replicated.
  - #2: Computing load is skewed.
  - #3: Applications are not distributed at each site.
  - #4: Power failures happen.
  - #5: Inter-domain cross-AS communications are down.
  - ……



Data Center WAN

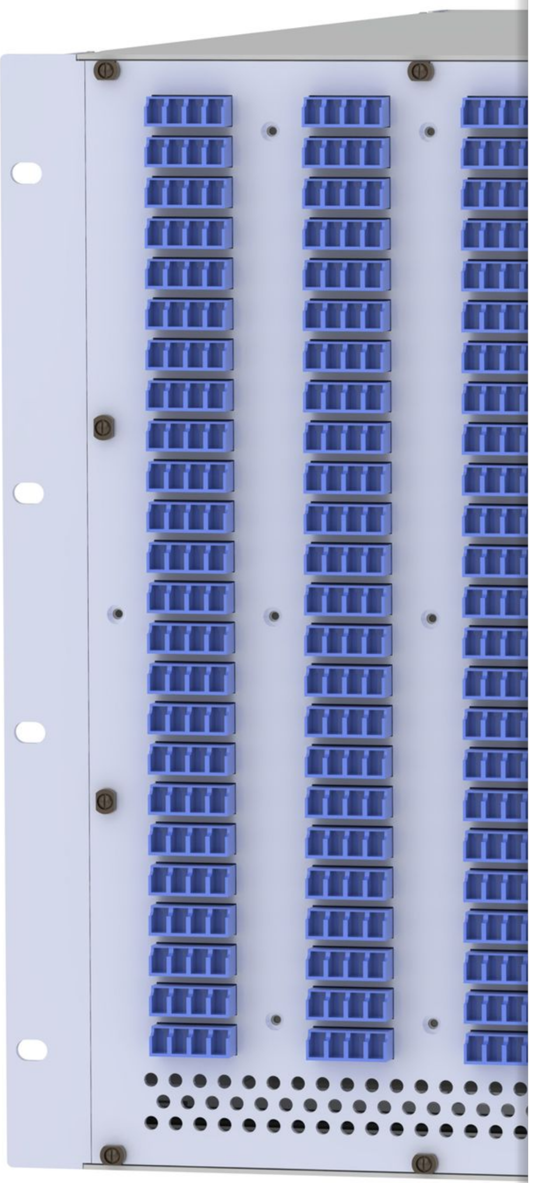
- High-bandwidth long-distance networking
  - Optical circuit switching
  - Fiber optic cables

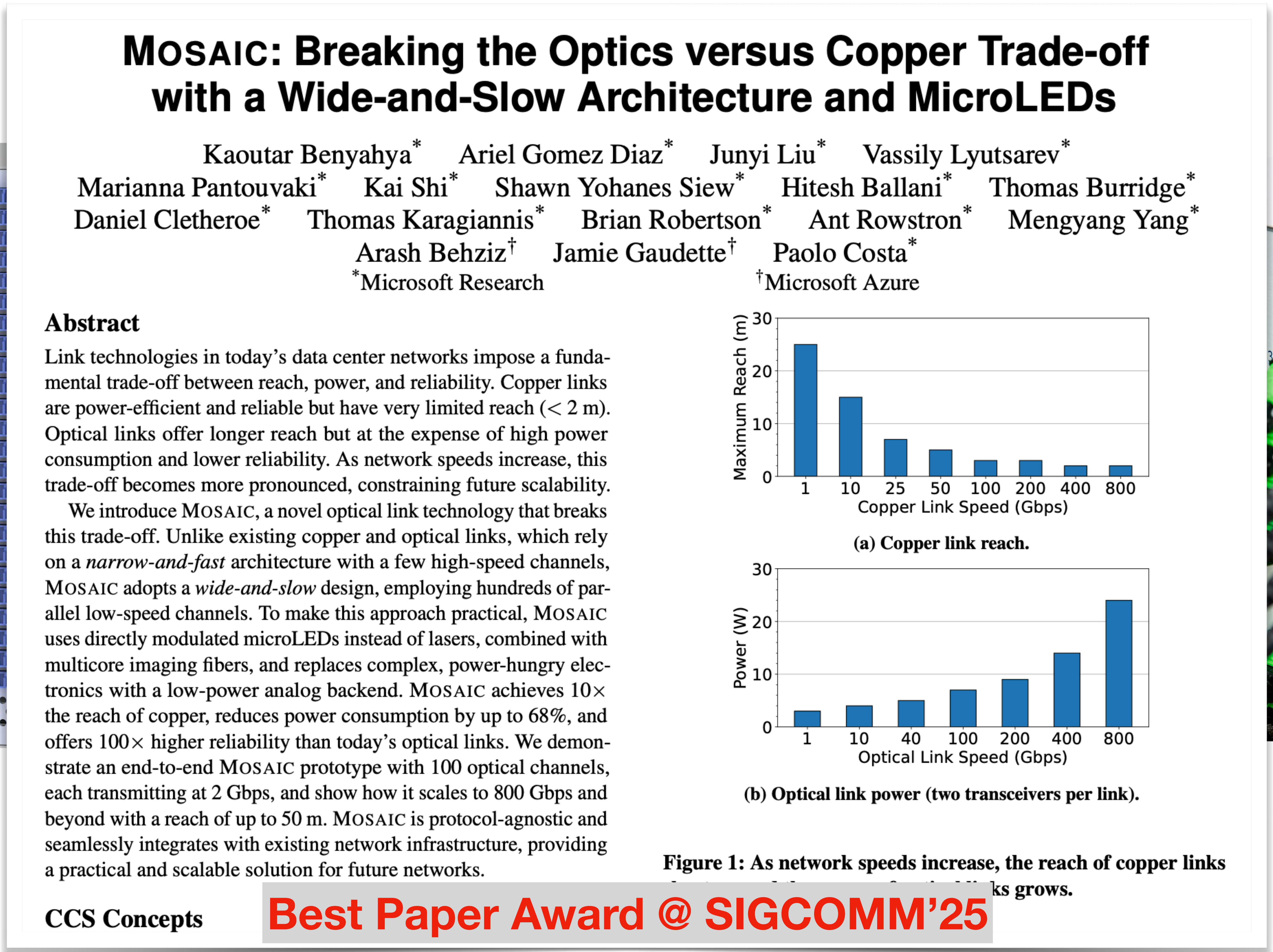




Data Center WAN

- High-bandwidth long-distance networking
  - Optical circuit switching
  - Fiber optic cables





MOSAIC: Breaking the Optics versus Copper Trade-off with a Wide-and-Slow Architecture and MicroLEDs

Kaoutar Benyahya* Ariel Gomez Diaz* Junyi Liu* Vassily Lyutsarev*
Marianna Pantouvaki* Kai Shi* Shawn Yohanes Siew* Hitesh Ballani* Thomas Burridge*
Daniel Cletheroe* Thomas Karagiannis* Brian Robertson* Ant Rowstron* Mengyang Yang*
Arash Behziz† Jamie Gaudette† Paolo Costa*
*Microsoft Research †Microsoft Azure

Abstract

Link technologies in today's data center networks impose a fundamental trade-off between reach, power, and reliability. Copper links are power-efficient and reliable but have very limited reach (< 2 m). Optical links offer longer reach but at the expense of high power consumption and lower reliability. As network speeds increase, this trade-off becomes more pronounced, constraining future scalability.

We introduce MOSAIC, a novel optical link technology that breaks this trade-off. Unlike existing copper and optical links, which rely on a *narrow-and-fast* architecture with a few high-speed channels, MOSAIC adopts a *wide-and-slow* design, employing hundreds of parallel low-speed channels. To make this approach practical, MOSAIC uses directly modulated microLEDs instead of lasers, combined with multicore imaging fibers, and replaces complex, power-hungry electronics with a low-power analog backend. MOSAIC achieves 10× the reach of copper, reduces power consumption by up to 68%, and offers 100× higher reliability than today's optical links. We demonstrate an end-to-end MOSAIC prototype with 100 optical channels, each transmitting at 2 Gbps, and show how it scales to 800 Gbps and beyond with a reach of up to 50 m. MOSAIC is protocol-agnostic and seamlessly integrates with existing network infrastructure, providing a practical and scalable solution for future networks.

(a) Copper link reach.

(b) Optical link power (two transceivers per link).

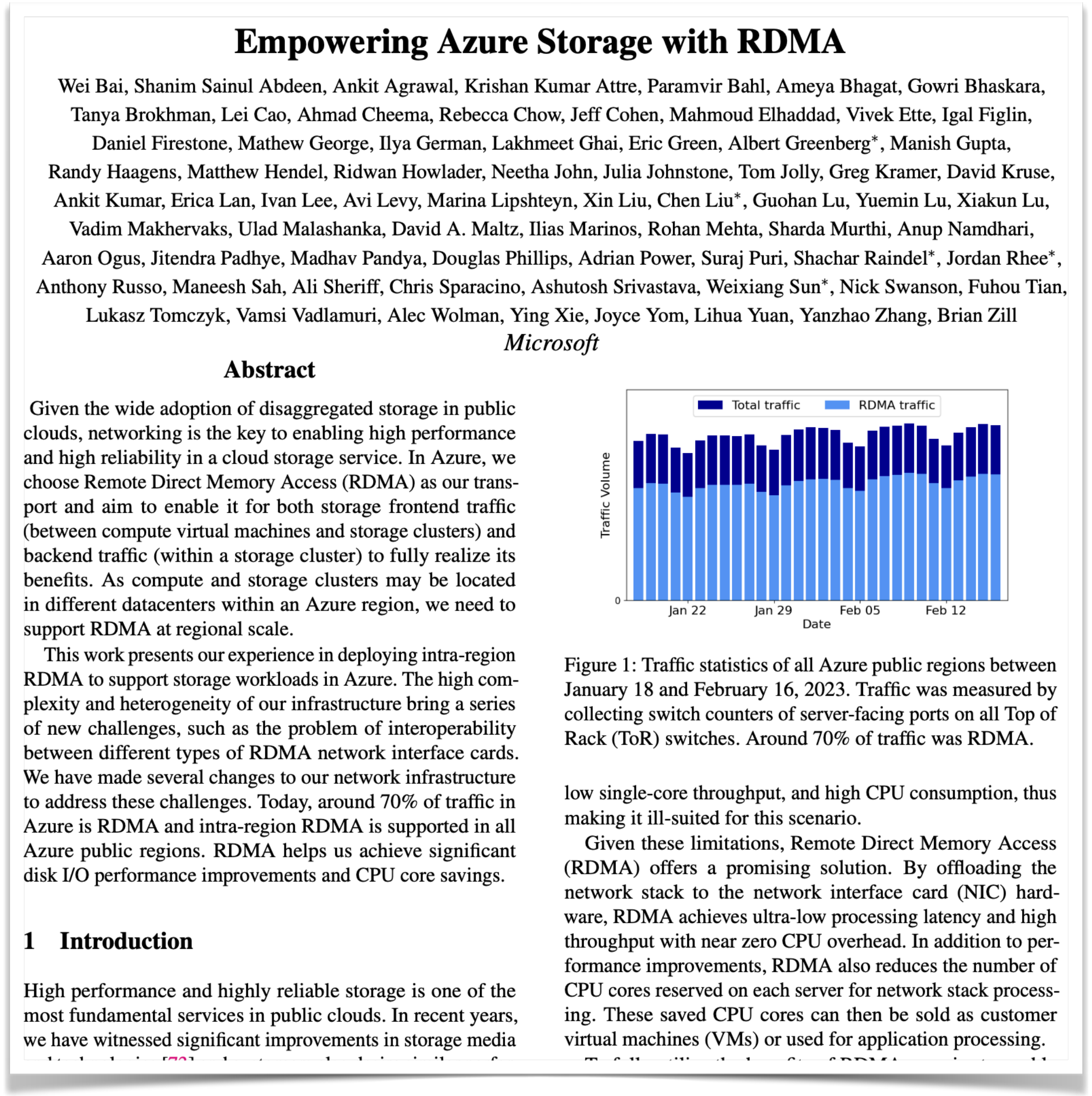
Figure 1: As network speeds increase, the reach of copper links ... ks grows.

CCS Concepts

Best Paper Award @ SIGCOMM'25

A Special Data Center WAN

- Each data center site (region) is decoupled into several subsites.
  - An engineering-driven solution
  - Scale out the data center at a “busy” area incrementally
  - Long-haul links can be as long as tens of kilometers.



Empowering Azure Storage with RDMA

Wei Bai, Shanim Sainul Abdeen, Ankit Agrawal, Krishan Kumar Attre, Paramvir Bahl, Ameya Bhagat, Gowri Bhaskara, Tanya Brokhman, Lei Cao, Ahmad Cheema, Rebecca Chow, Jeff Cohen, Mahmoud Elhaddad, Vivek Ette, Igal Figlin, Daniel Firestone, Mathew George, Ilya German, Lakhmeet Ghai, Eric Green, Albert Greenberg*, Manish Gupta, Randy Haagens, Matthew Hendel, Ridwan Howlader, Neetha John, Julia Johnstone, Tom Jolly, Greg Kramer, David Kruse, Ankit Kumar, Erica Lan, Ivan Lee, Avi Levy, Marina Lipshteyn, Xin Liu, Chen Liu*, Guohan Lu, Yuemin Lu, Xiakun Lu, Vadim Makhervaks, Ulad Malashanka, David A. Maltz, Ilias Marinos, Rohan Mehta, Sharda Murthi, Anup Namdhari, Aaron Ogus, Jitendra Padhye, Madhav Pandya, Douglas Phillips, Adrian Power, Suraj Puri, Shachar Raindel*, Jordan Rhee*, Anthony Russo, Maneesh Sah, Ali Sheriff, Chris Sparacino, Ashutosh Srivastava, Weixiang Sun*, Nick Swanson, Fuhou Tian, Lukasz Tomczyk, Vamsi Vadlamuri, Alec Wolman, Ying Xie, Joyce Yom, Lihua Yuan, Yanzhao Zhang, Brian Zill
Microsoft

Abstract

Given the wide adoption of disaggregated storage in public clouds, networking is the key to enabling high performance and high reliability in a cloud storage service. In Azure, we choose Remote Direct Memory Access (RDMA) as our transport and aim to enable it for both storage frontend traffic (between compute virtual machines and storage clusters) and backend traffic (within a storage cluster) to fully realize its benefits. As compute and storage clusters may be located in different datacenters within an Azure region, we need to support RDMA at regional scale.

This work presents our experience in deploying intra-region RDMA to support storage workloads in Azure. The high complexity and heterogeneity of our infrastructure bring a series of new challenges, such as the problem of interoperability between different types of RDMA network interface cards. We have made several changes to our network infrastructure to address these challenges. Today, around 70% of traffic in Azure is RDMA and intra-region RDMA is supported in all Azure public regions. RDMA helps us achieve significant disk I/O performance improvements and CPU core savings.

1 Introduction

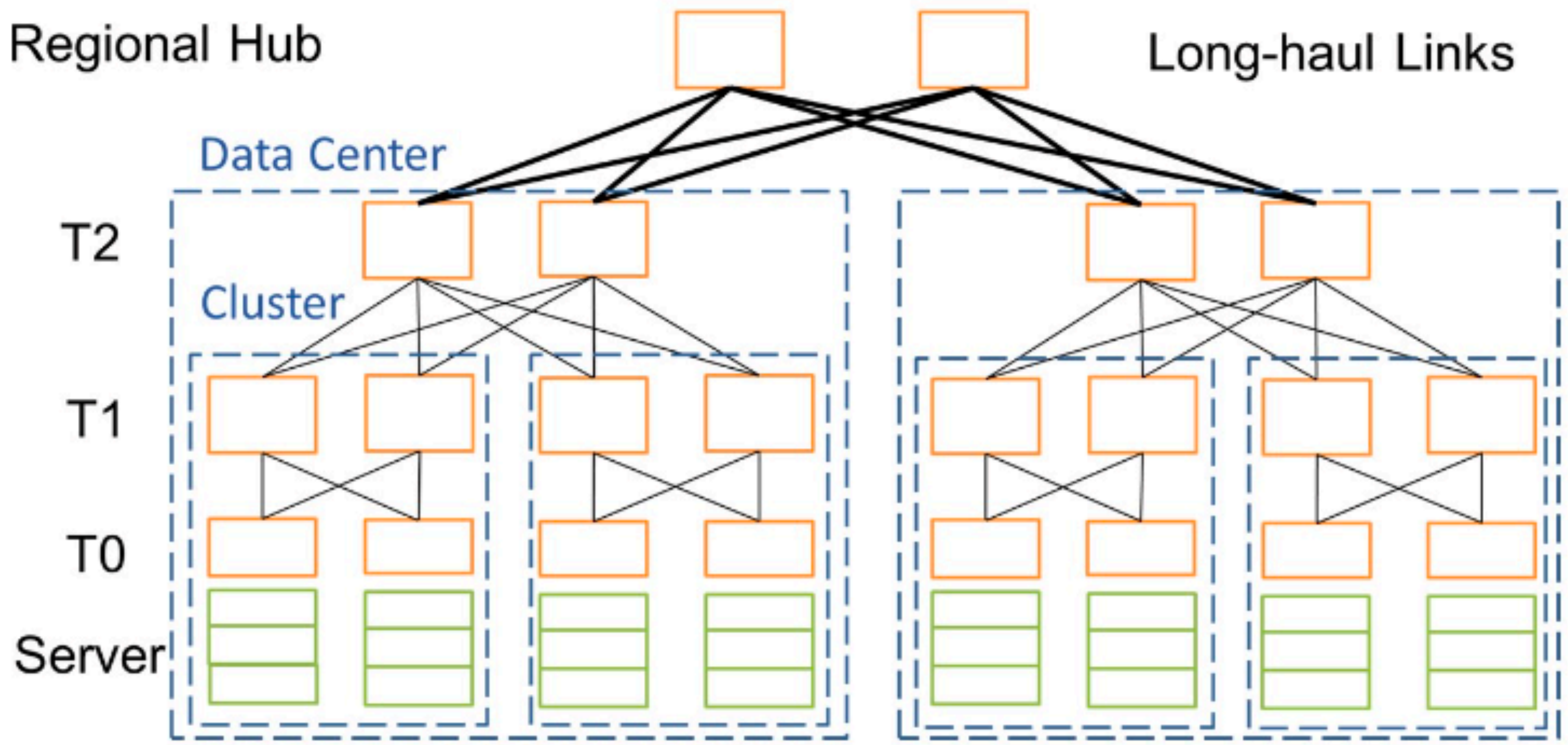
High performance and highly reliable storage is one of the most fundamental services in public clouds. In recent years, we have witnessed significant improvements in storage media

Total traffic
RDMA traffic
Traffic Volume
0
Jan 22
Jan 29
Feb 05
Feb 12
Date

Figure 1: Traffic statistics of all Azure public regions between January 18 and February 16, 2023. Traffic was measured by collecting switch counters of server-facing ports on all Top of Rack (ToR) switches. Around 70% of traffic was RDMA.

low single-core throughput, and high CPU consumption, thus making it ill-suited for this scenario.

Given these limitations, Remote Direct Memory Access (RDMA) offers a promising solution. By offloading the network stack to the network interface card (NIC) hardware, RDMA achieves ultra-low processing latency and high throughput with near zero CPU overhead. In addition to performance improvements, RDMA also reduces the number of CPU cores reserved on each server for network stack processing. These saved CPU cores can then be sold as customer virtual machines (VMs) or used for application processing.

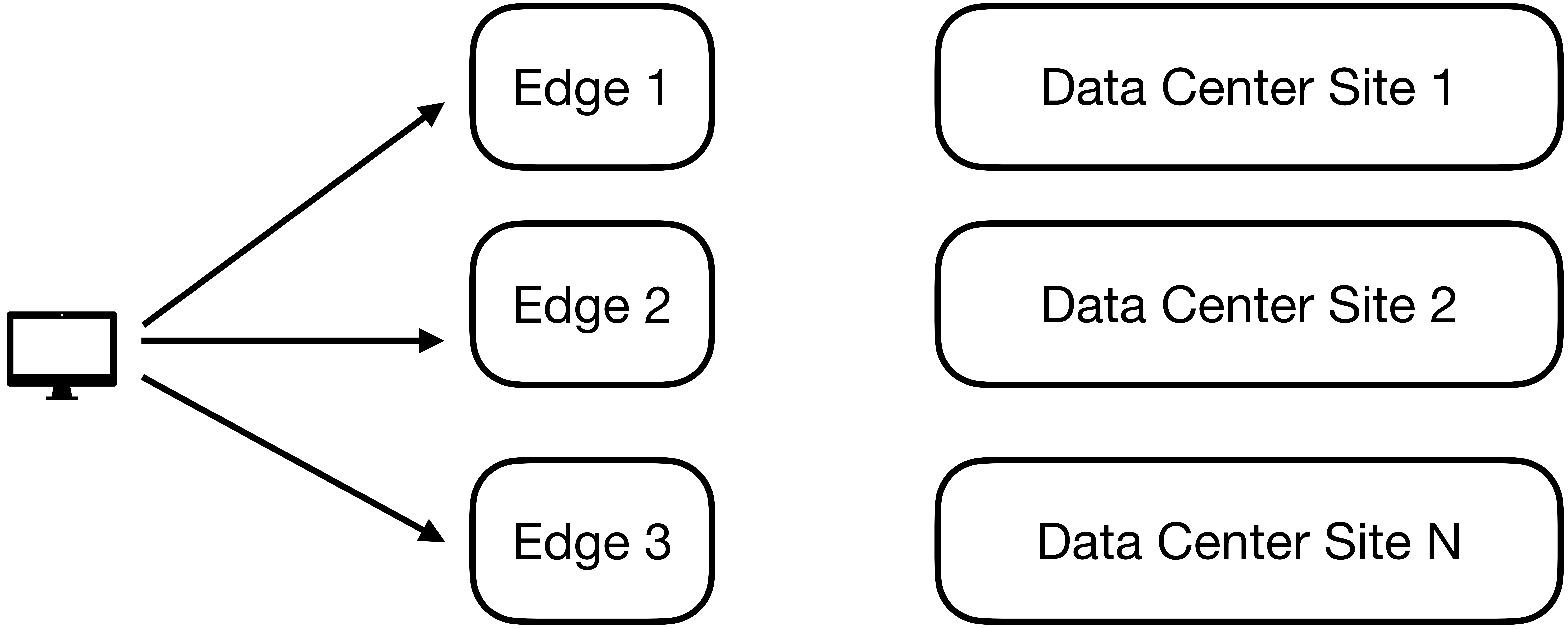


NSDI’23

Intra-Region RDMA

Approach #2: Add Edge Sites

- But, "Edge Site" is not well-defined
  - A small-scale data center, dominated by networking capability
  - Equipped with some compute and storage resources



Approach #2: Add Edge Sites

- But, “Edge Site” is not well-defined
  - A small-scale data center, dominated by networking capability
  - Equipped with some compute and storage resources

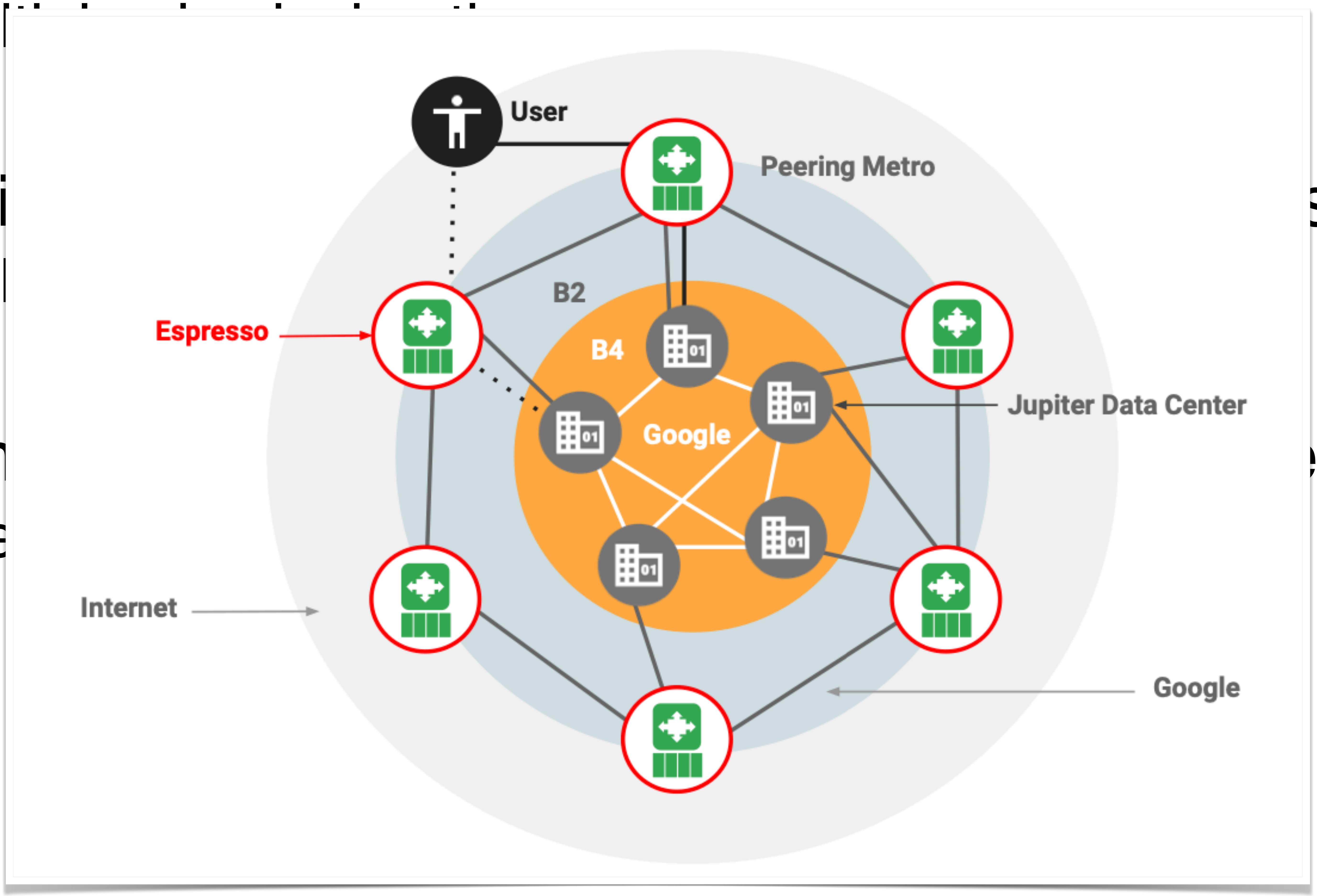
- Peering or Points-of-Presence (PoP)
  - Efficient routing
  - Reduced communication costs
  - Fast
  - Better QoS
  - …

Combine Everything Together

- #1: Expand the bandwidth capability to serve user traffic demand
  - Add multiple physical paths
- #2: Replicate the data center sites to reduce access latency
  - Add multiple physical data center sites
- #3: Cache traffic control (and app.) to reduce access latency
  - Add peering edge points

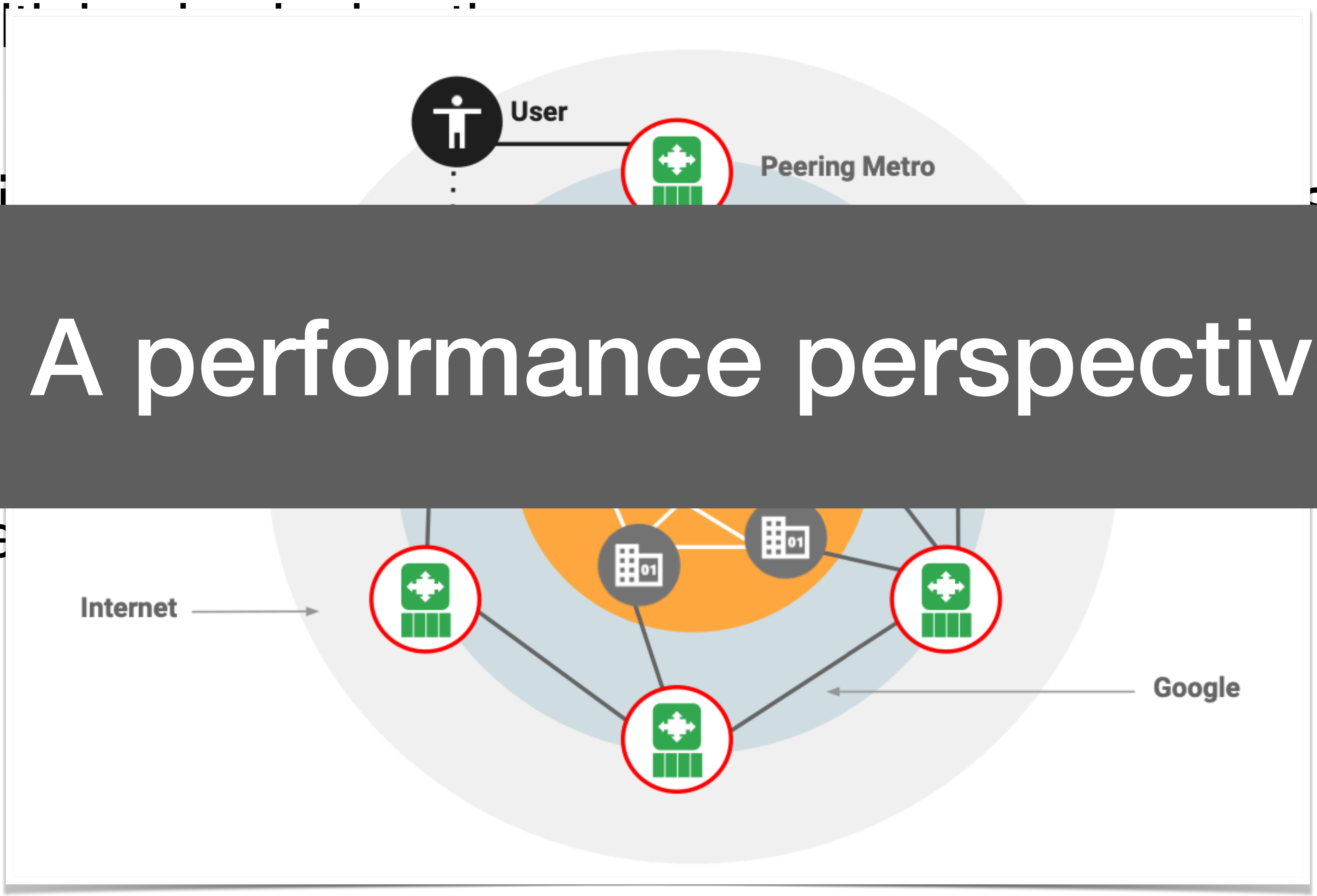
Combine Everything Together

- #1: Expand the bandwidth capability to serve user traffic demand
  - Add mul
- #2: Repli ss latency
  - Add mul
- #3: Cach ess latency
  - Add pee



Combine Everything Together

- #1: Expand the bandwidth capability to serve user traffic demand
  - Add mul
- #2: Repli ... ss latency
  - Add pee



A performance perspective

However, reliability is sometimes the first quest in reality!

Summary

- Today
  - Physical connectivity beyond the data center

- Next lecture
  - Addressing inside and outside the data center network
  - No new readings