Advanced Computer Networks

CS740

Introduction

https://pages.cs.wisc.edu/~mgliu/CS740/F25/index.html

Ming Liu
mgliu@cs.wisc.edu

Today

- Course Logistics
- Schedule Overview
- Computer Network Design History
- Data Center Network Basics
- Data Center Network Design Requirements

Administrative Details

- Instructor: Ming Liu
  - mgliu@cs.wisc.edu
  - Office: #7673@Morgride Hall
  - Research Interest: Networking and Systems
  - Theme: Network-based Hardware Acceleration (HW4NET and NET4HW)

- Research Focus:
  - Load-Store Interconnects
  - Disaggregated Storage
  - Programmable Networks

Course Communication Channels

- #1. Course website: slides, readings, and schedule
  - https://pages.cs.wisc.edu/~mgliu/CS740/F25/index.html
- #2. Course mailing list: announcements
  - compsci740-1-f25@g-groups.wisc.edu
- #3. Piazza: Q&A
  - https://piazza.com/wisc/fall2025/cs740
- #4. Submission
  - Canvas: labs, projects, and grades

Goals

- #1. Understand core networking concepts and principles
  - Best-effort, soft states, layering
- #2. Explore the philosophy of building networked systems
  - Networking architecture, protocols, and algorithms
- #3. Engage in networking research
  - Reading, discussion, implementation, and evaluation

Course Format

- **Before class:** read a mix of classic and recent papers

- **In-class:** in-depth discussion

- **After class:** labs, projects, and exam

Grade Breakdown

- Programming Labs (30%)
  - Lab1: Sliding window over DPDK (15%)
  - Lab2: Flow scheduling and load balancing (15%)
  - Teams of 1-3 people
- In-class Exam (30%)
  - Open-book, open-notes
  - 11/20/2025 (Thursday)
- Course Project (40%)
  - Proposal report: 2.5%
  - Midterm report: 2.5%
  - Code, demo, final report, and presentation: 35%
  - Teams of 1-3 people

Grading Option #1

- **A: [90, 100]**
- **AB: [85, 90)**
- **B: [80, 85)**
- **BC: [75, 80)**
- **C: [70, 75)**
- **D: [60, 70)**
- **F: [0, 60)**

Grading Option #2 (Curving)

- **A: [0, 35%]**
- **AB: [35%, 60%)**
- **B: [60%, 80%)**
- **BC: [80%, 90%)**
- **C: [,)**
- **D: [,)**
- **F: [,)**

Your final grade = Max (op#1, op#2)

- E.g., (AB, A) = A

Reading and References

- No textbook
- ~20 research papers
  - Even less
  - We care about depth not breadth
- Additional references (mentioned in the lecture)

Collaboration & Late Submission

- Working together is encouraged.

- But the final submission must be your own work!

- Participate
  - Ask and answer questions in class and after class

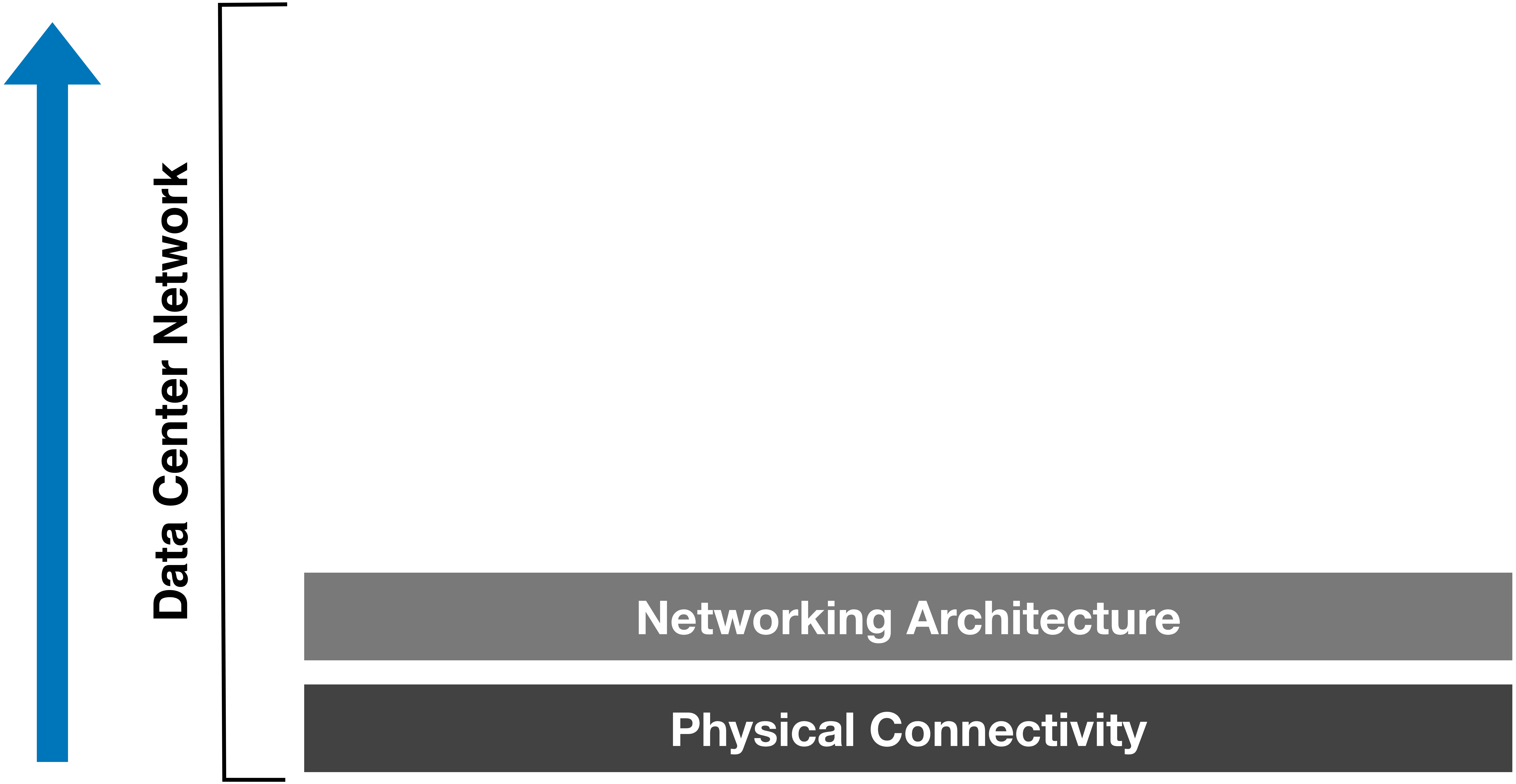
- Meet deadlines
  - Turn assignments in on time
  - Late penalty (see each assignment)

Today

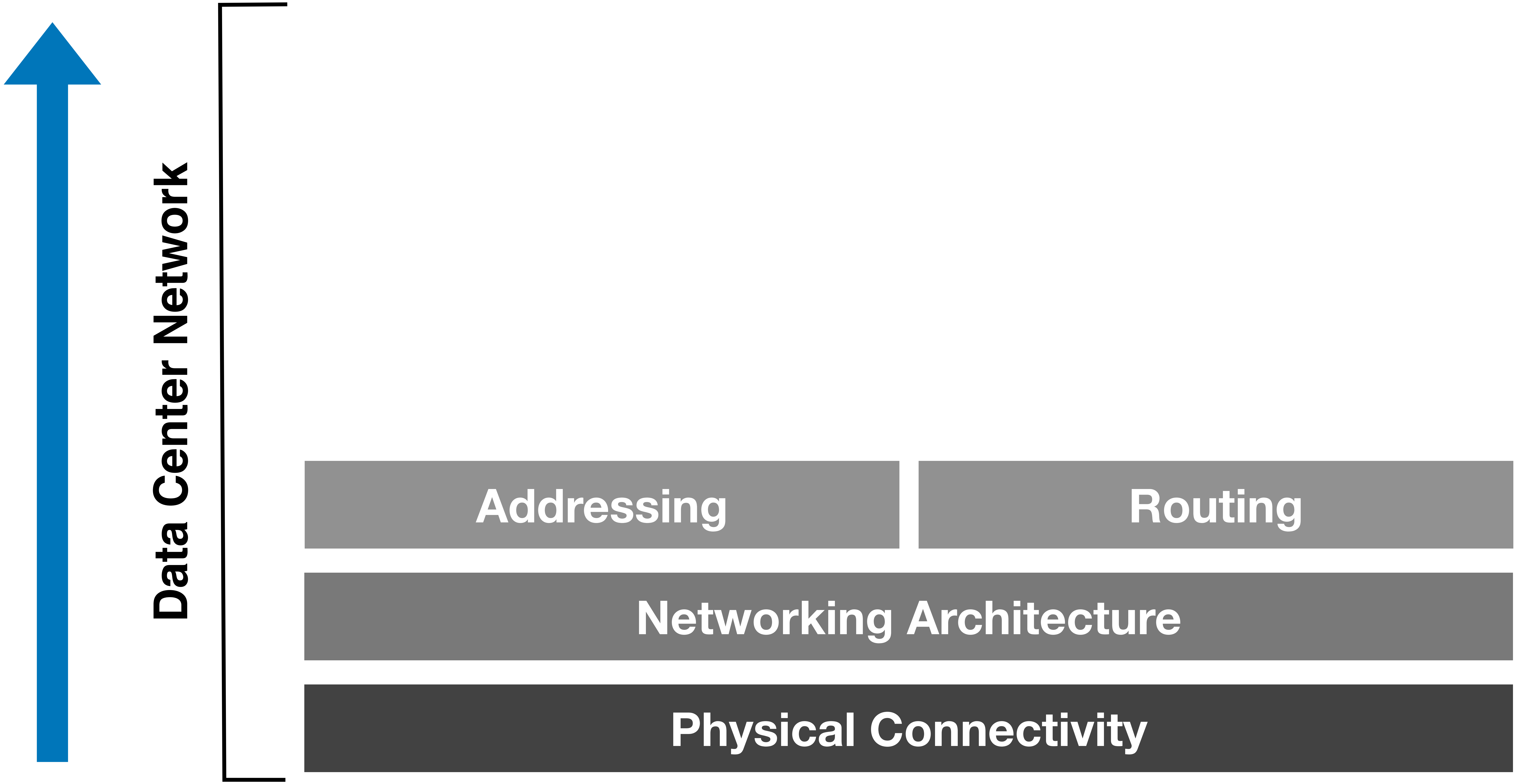
- *Course Logistics*
- Schedule Overview
- Computer Network Design History
- Data Center Network Basics
- Data Center Network Design Requirements

Reading Topics — Bottom-up

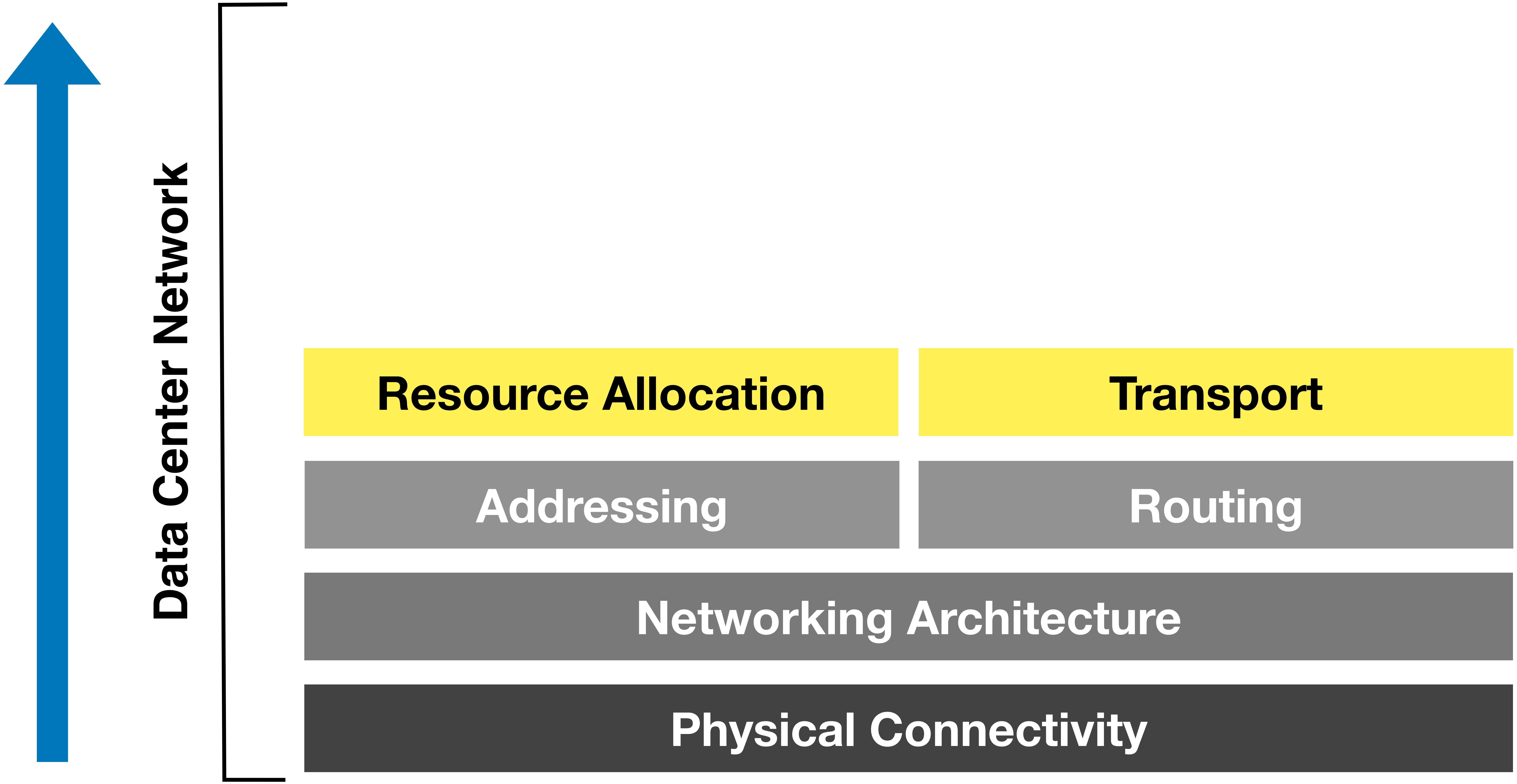
Reading Topics — Bottom-up



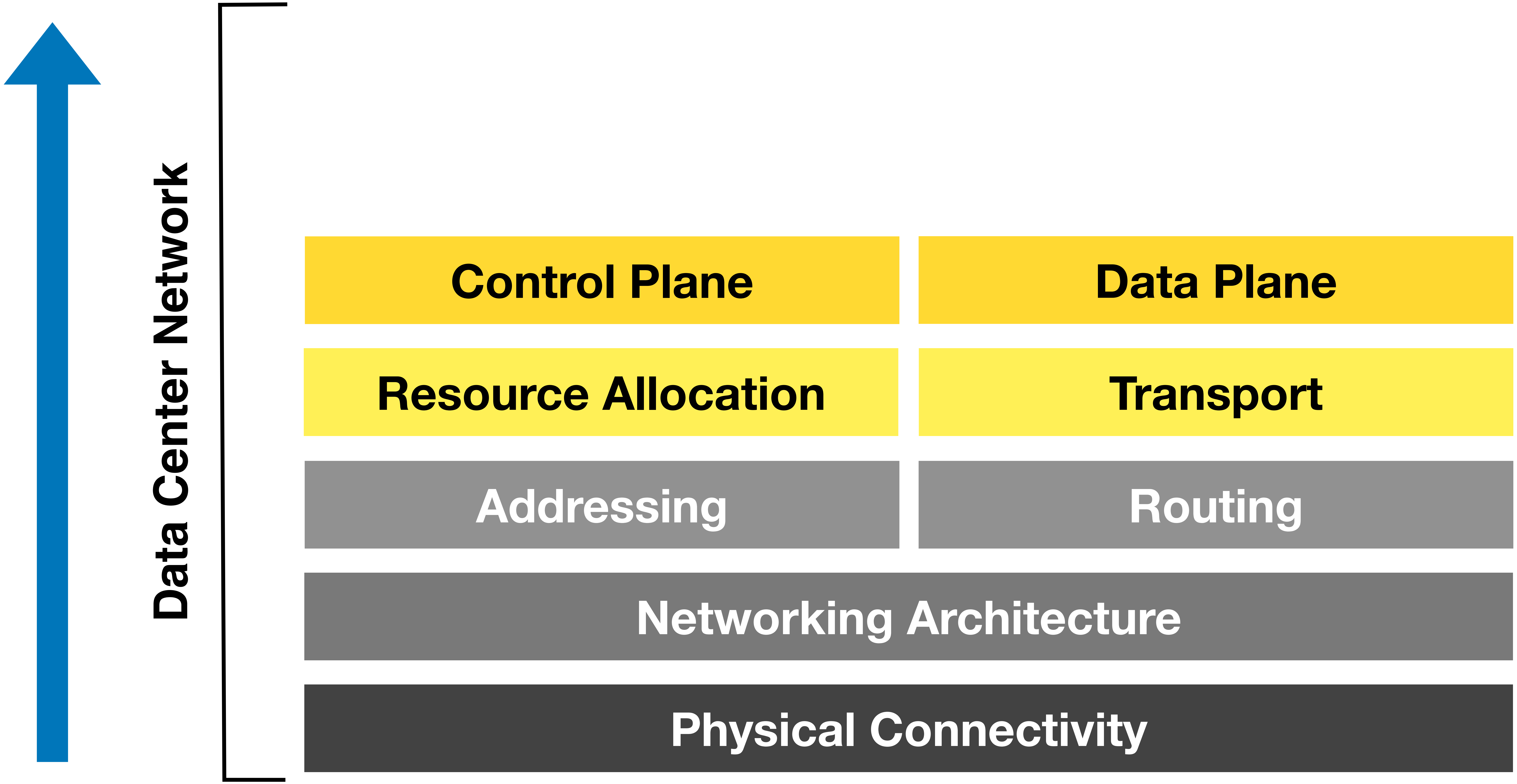
Reading Topics — Bottom-up



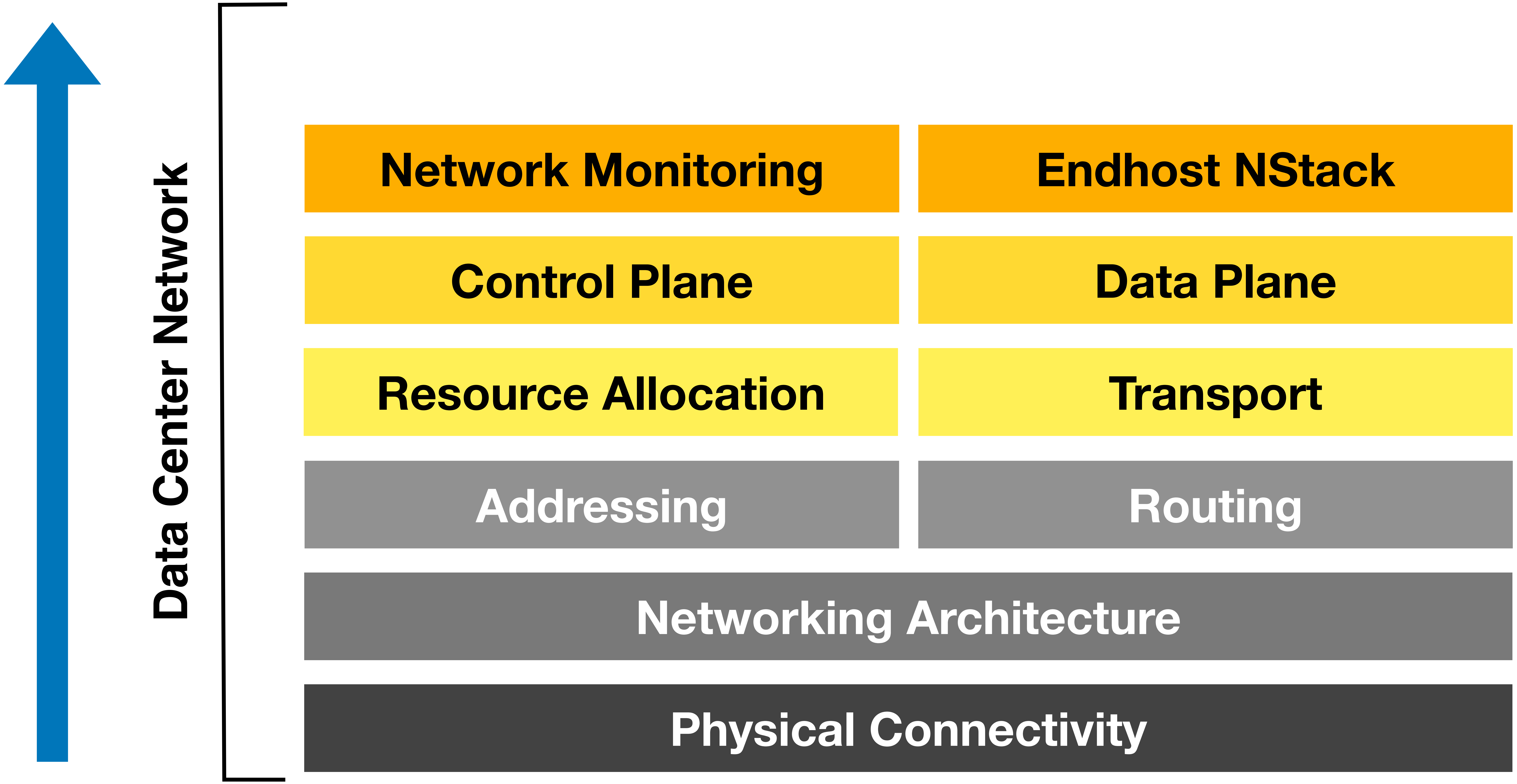
Reading Topics — Bottom-up



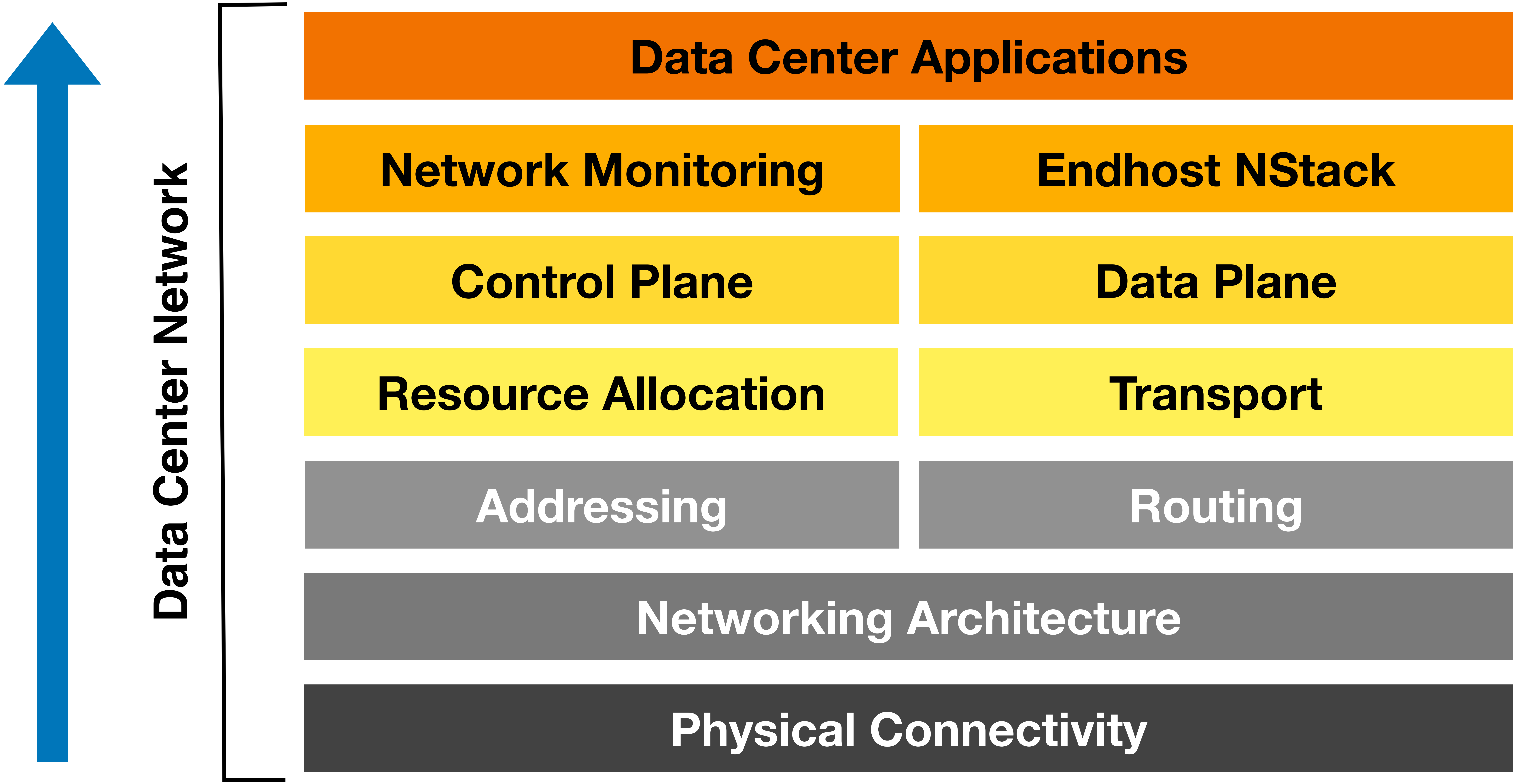
Reading Topics — Bottom-up



Reading Topics — Bottom-up



Reading Topics — Bottom-up



Reading Topics — Bottom-up

Data Center Applications

Network Monitoring | Endhost NStack

- We will not cover every networking topic!

Networking Architecture

Physical Connectivity

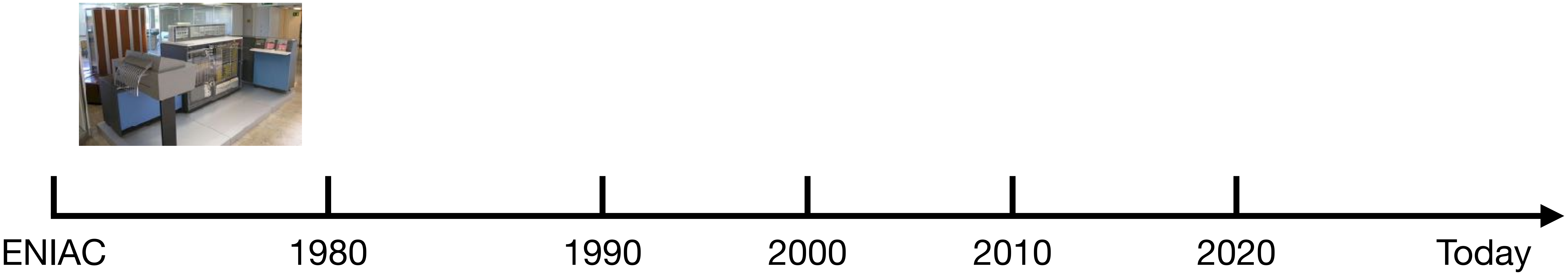
Today

- *Course Logistics*
- *Schedule Overview*
- Computer Network Design History
- Data Center Network Basics
- Data Center Network Design Requirements

Computer Network Design for <?>

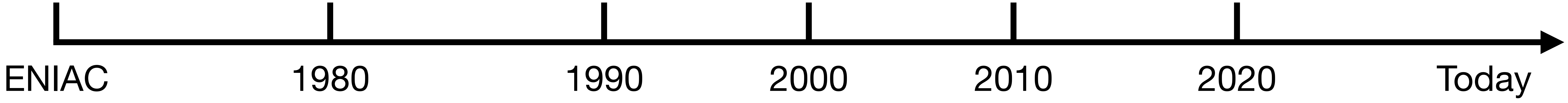
Computer Network Design for <?>

- Before 1980s: Mainframes



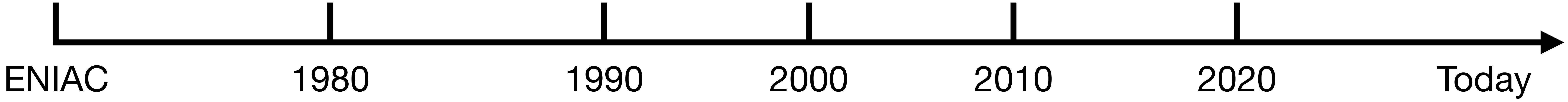
Computer Network Design for <?>

- Before 1980s: Mainframes
- 1980s-1990s: Workstations



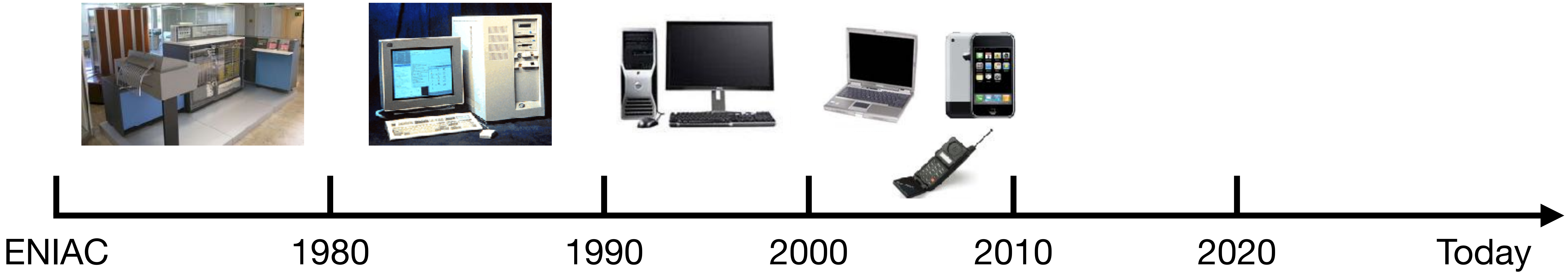
Computer Network Design for <?>

- Before 1980s: Mainframes
- 1980s-1990s: Workstations
- 1990s-2000s: PC



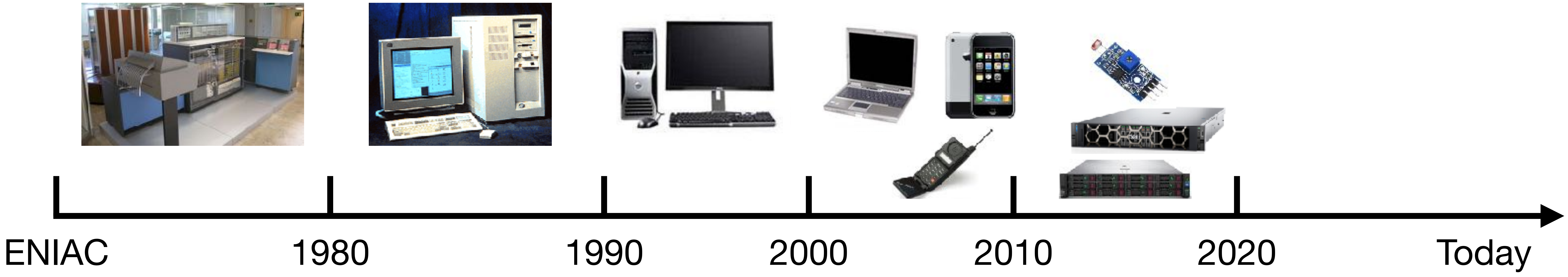
Computer Network Design for <?>

- Before 1980s: Mainframes
- 1980s-1990s: Workstations
- 1990s-2000s: PC
- 2000s-2010s: Wireless devices, e.g., laptops, phones, etc.



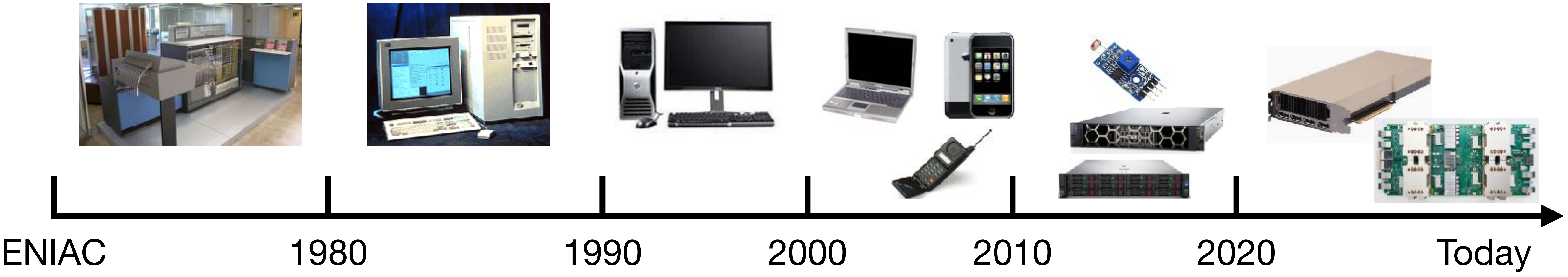
Computer Network Design for <?>

- Before 1980s: Mainframes
- 1980s-1990s: Workstations
- 1990s-2000s: PC
- 2000s-2010s: Wireless devices, e.g., laptops, phones, etc.
- 2010s-2020s: Servers and IoT devices

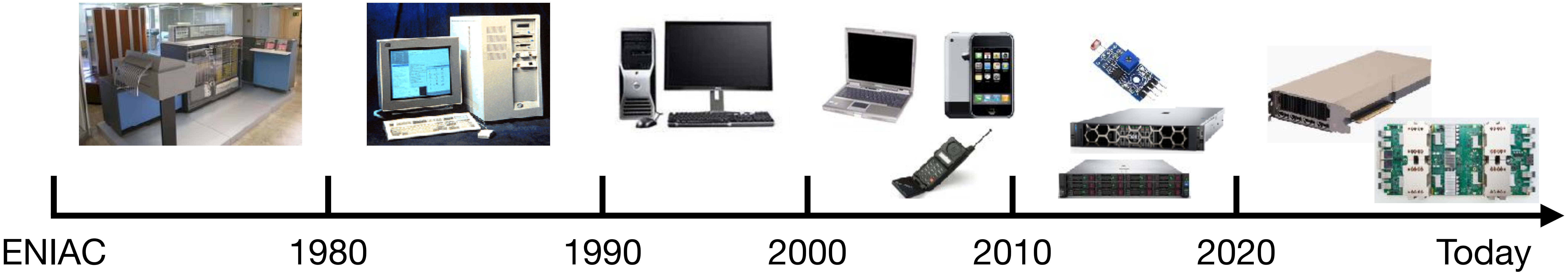


Computer Network Design for <?>

- Before 1980s: Mainframes
- 1980s-1990s: Workstations
- 1990s-2000s: PC
- 2000s-2010s: Wireless devices, e.g., laptops, phones, etc.
- 2010s-2020s: Servers and IoT devices
- 2020s-present: Accelerators, e.g., GPUs, TPUs, etc.



What kinds of networks have we built?



What kinds of networks have we built?

Rank	System	Cores	Rmax (GFlop/s)	Rpeak (GFlop/s)	Power (kW)
1	Numerical Wind Tunnel, Fujitsu National Aerospace Laboratory of Japan Japan	140	124.00	235.79	
2	CM-5/1024, Thinking Machines Corporation Los Alamos National Laboratory United States	1,024	59.70	131.00	
3	CM-5/544, Thinking Machines Corporation Minnesota Supercomputer Center United States	544	30.40	69.63	

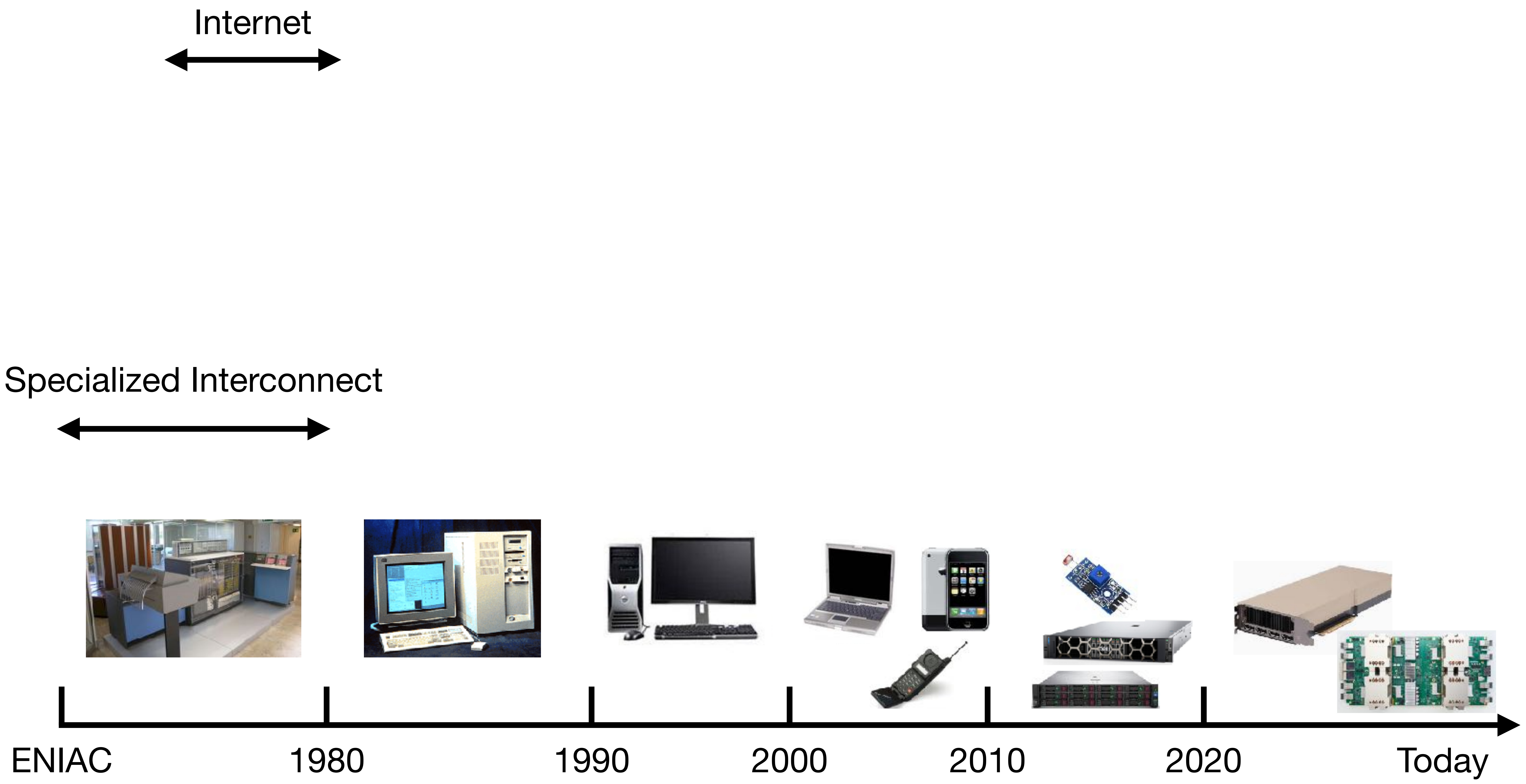
Specialized Interconnect






ENIAC 1980 1990 2000 2010 2020 Today

What kinds of networks have we built?

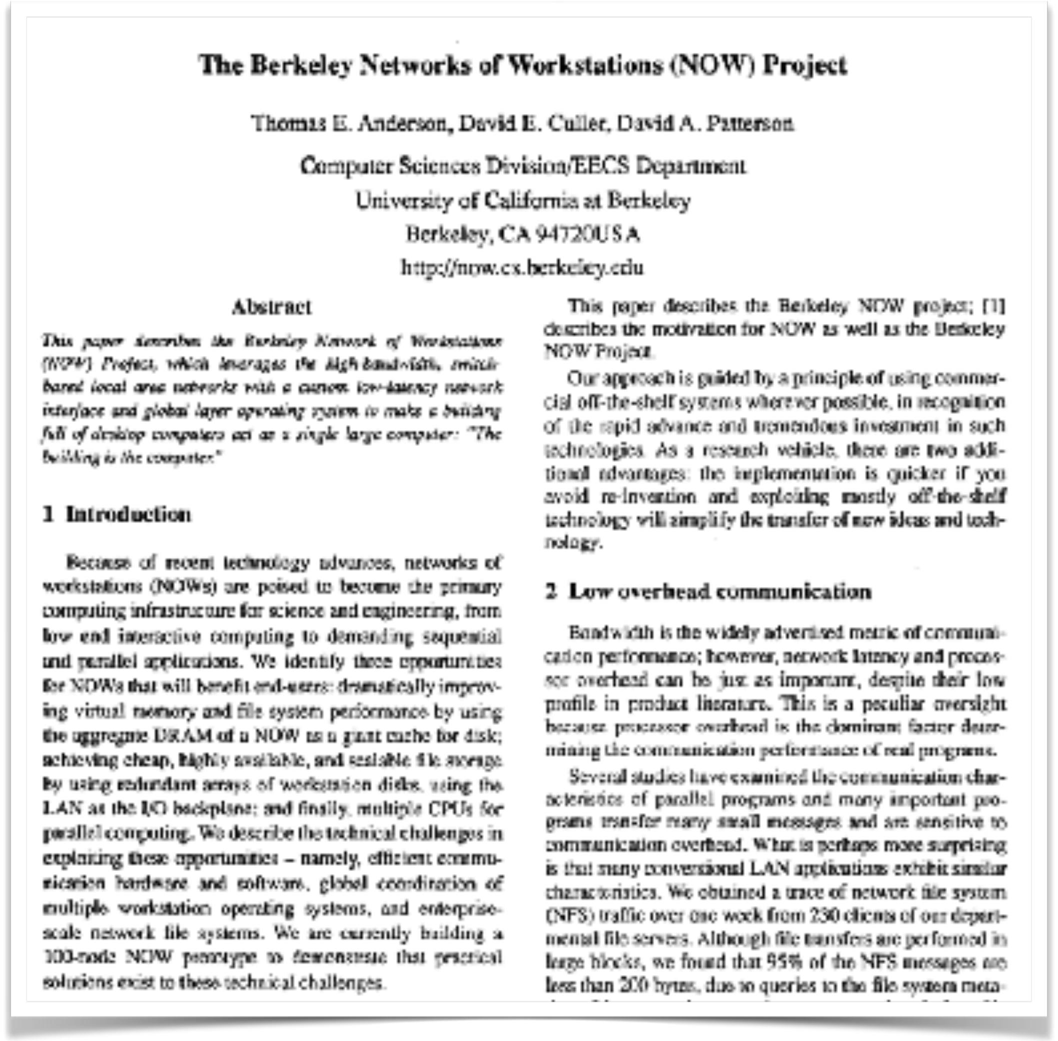


What kinds of networks have we built?

Internet

Local Area Network

Specialized Interconnect



The Berkeley Networks of Workstations (NOW) Project

Thomas E. Anderson, David E. Culler, David A. Patterson

Computer Sciences Division/EECS Department

University of California at Berkeley

Berkeley, CA 94720USA

http://now.cs.berkeley.edu

Abstract

This paper describes the Berkeley Network of Workstations (NOW) Project, which leverages the high bandwidth, switch-based local area networks with a custom low-latency network interface and global layer operating system to make a building full of desktop computers act as a single large computer: "The building is the computer."

1 Introduction

Because of recent technology advances, networks of workstations (NOWs) are poised to become the primary computing infrastructure for science and engineering, from low end interactive computing to demanding sequential and parallel applications. We identify three opportunities for NOWs that will benefit end-users: dramatically improving virtual memory and file system performance by using the aggregate DRAM of a NOW as a giant cache for disk; achieving cheap, highly available, and scalable file storage by using redundant arrays of workstation disks, using the LAN as the I/O backplane; and finally, multiple CPUs for parallel computing. We describe the technical challenges in exploiting these opportunities – namely, efficient communication hardware and software, global coordination of multiple workstation operating systems, and enterprise-scale network file systems. We are currently building a 100-node NOW prototype to demonstrate that practical solutions exist to these technical challenges.

This paper describes the Berkeley NOW project; [1] describes the motivation for NOW as well as the Berkeley NOW Project.

Our approach is guided by a principle of using commercial off-the-shelf systems wherever possible, in recognition of the rapid advance and tremendous investment in such technologies. As a research vehicle, there are two additional advantages: the implementation is quicker if you avoid re-invention and exploiting mostly off-the-shelf technology will simplify the transfer of new ideas and technology.

2 Low overhead communication

Bandwidth is the widely advertised metric of communication performance; however, network latency and processor overhead can be just as important, despite their low profile in product literature. This is a peculiar oversight because processor overhead is the dominant factor determining the communication performance of real programs.

Several studies have examined the communication characteristics of parallel programs and many important programs transfer many small messages and are sensitive to communication overhead. What is perhaps more surprising is that many conventional LAN applications exhibit similar characteristics. We obtained a trace of network file system (NFS) traffic over one week from 230 clients of our departmental file servers. Although file transfers are performed in large blocks, we found that 95% of the NFS messages are less than 200 bytes, due to queries to the file system meta-










ENIAC 1980 1990 2000 2010 2020 Today

What kinds of networks have we built?

Internet

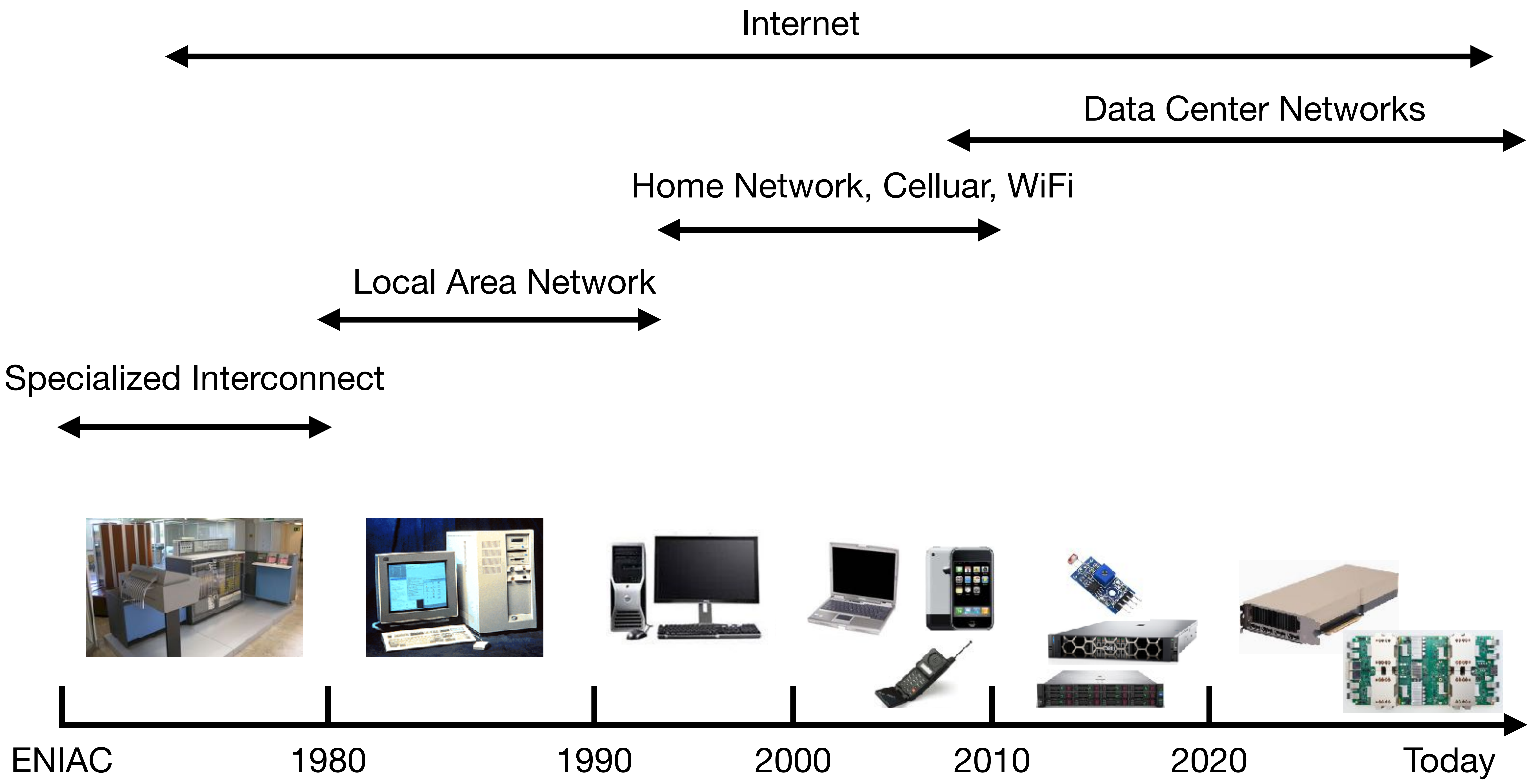
Home Network, Celluar, WiFi

Local Area Network

Specialized Interconnect

ENIAC 1980 1990 2000 2010 2020 Today

What kinds of networks have we built?



What kinds of networks have we built?

Internet

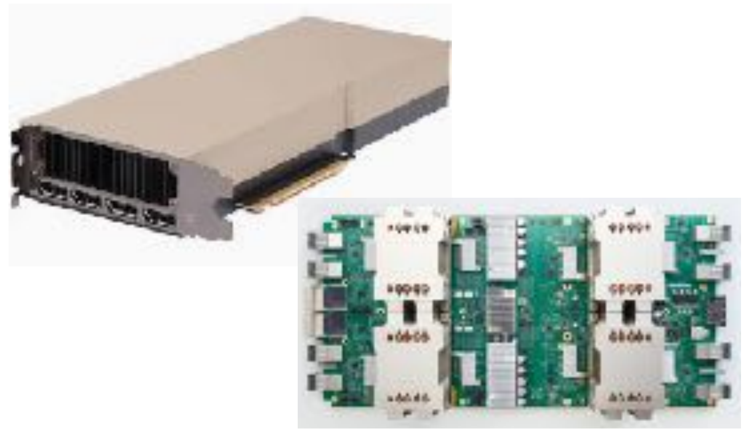
Data Center Networks

Our focus

Home Network, Celluar, WiFi

Local Area Network

Specialized Interconnect



ENIAC 1980 1990 2000 2010 2020 Today

Today

- *Course Logistics*
- *Schedule Overview*
- *Computer Network Design History*
- Data Center Network Basics
- Data Center Network Design Requirements

What are data centers?

Data Centers

- Data centers are an agglomeration of compute, memory, and storage resources.

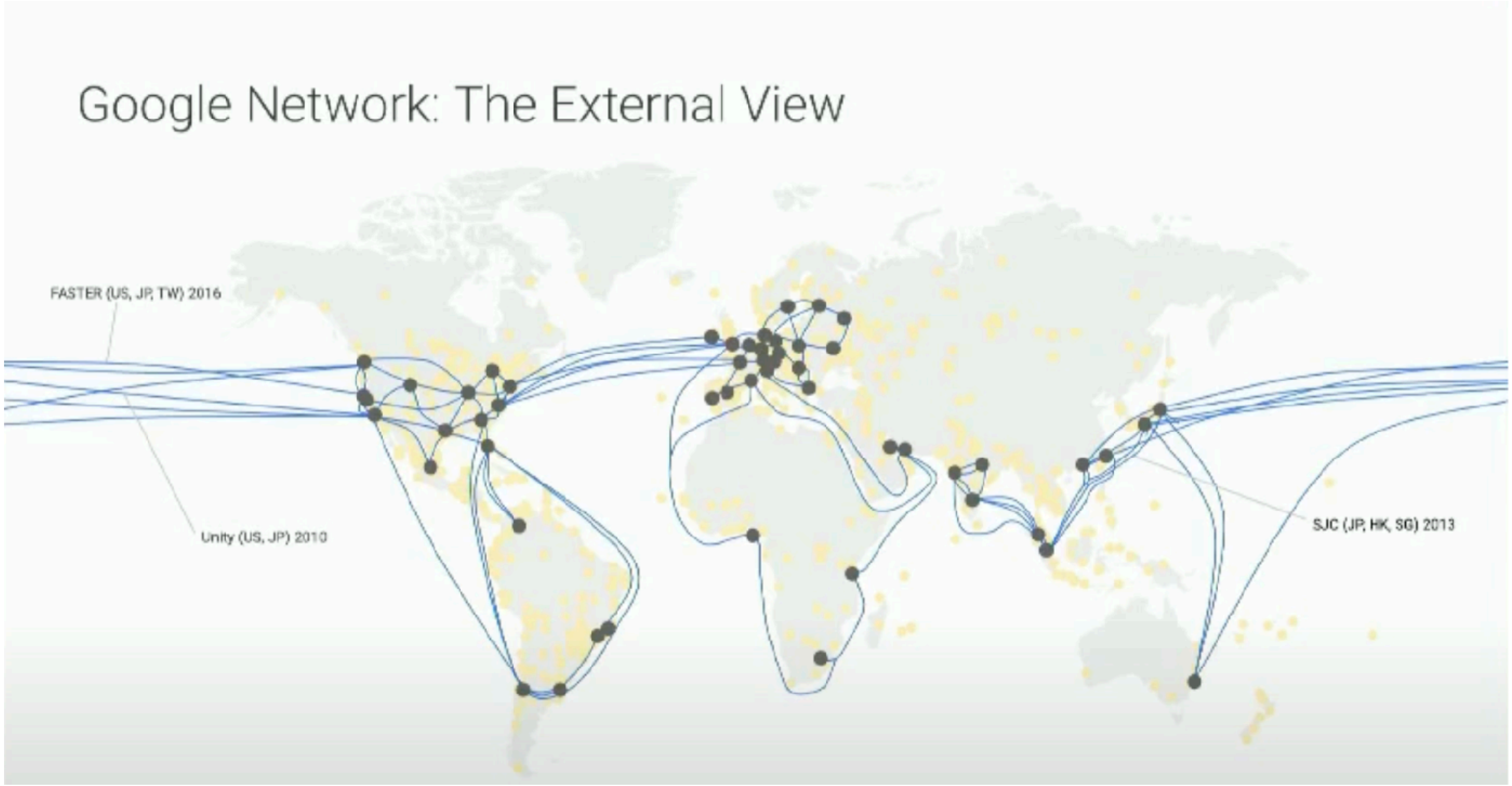
What are data center networks?

Data Center Networks

- Data center networks connect <?>

Data Center Networks

- Data center networks connect <?>

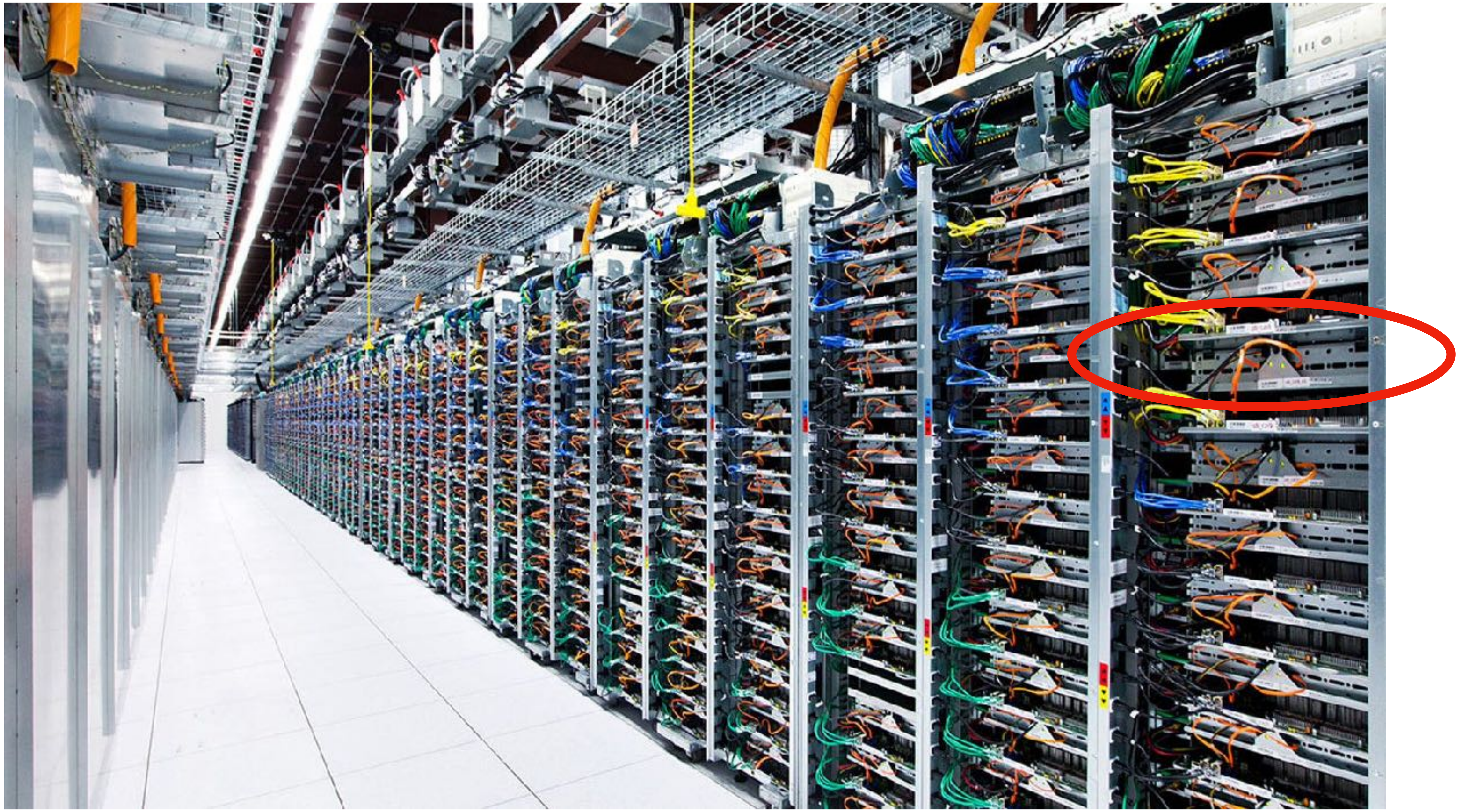


Data Center Networks

- Data center networks connect <?>

Data Center Networks

- Data center networks connect **servers**.



What is a server?

Server and System Architecture

- A server is a computing machine under a predefined form factor.



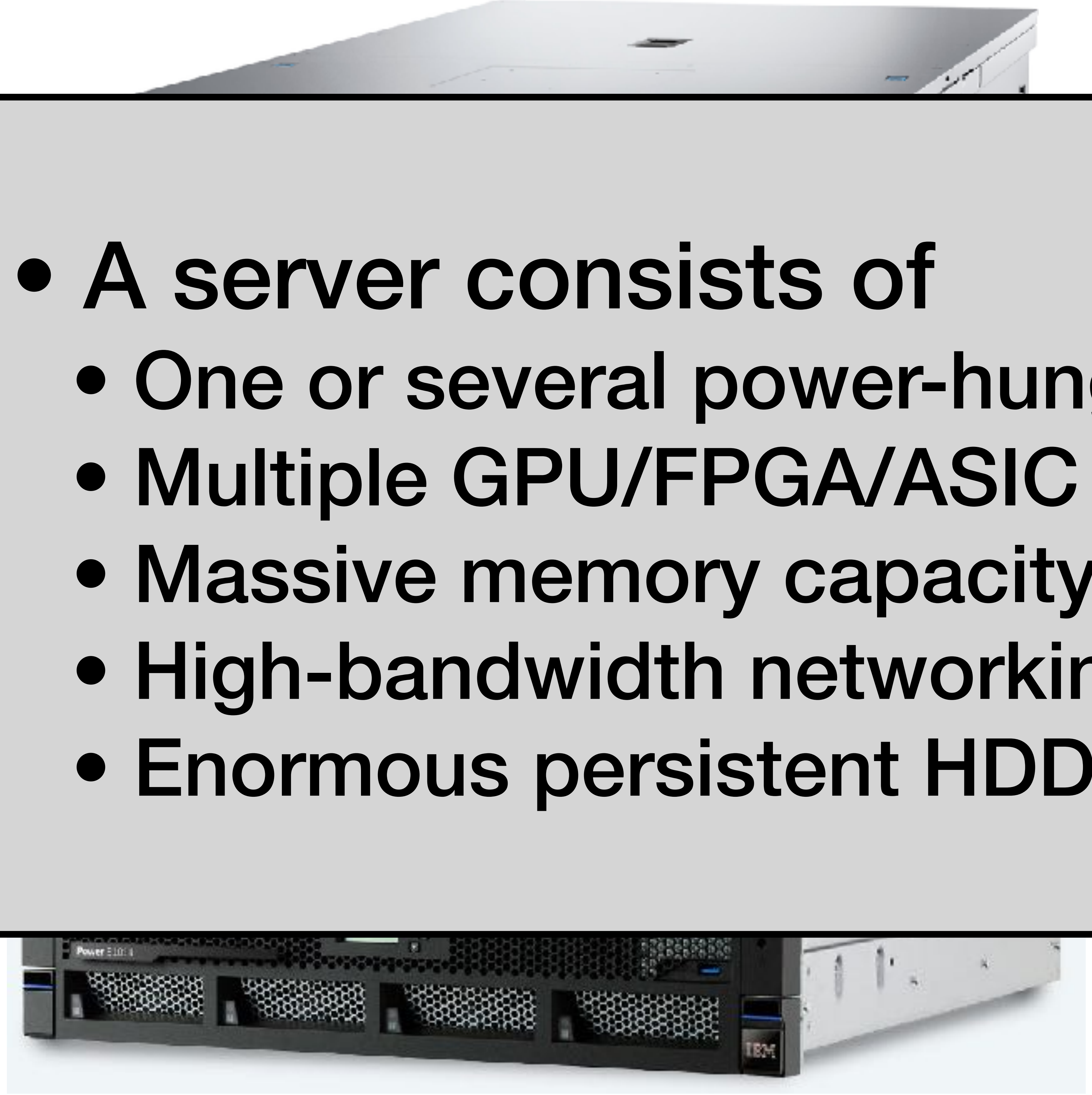

Server and System Architecture

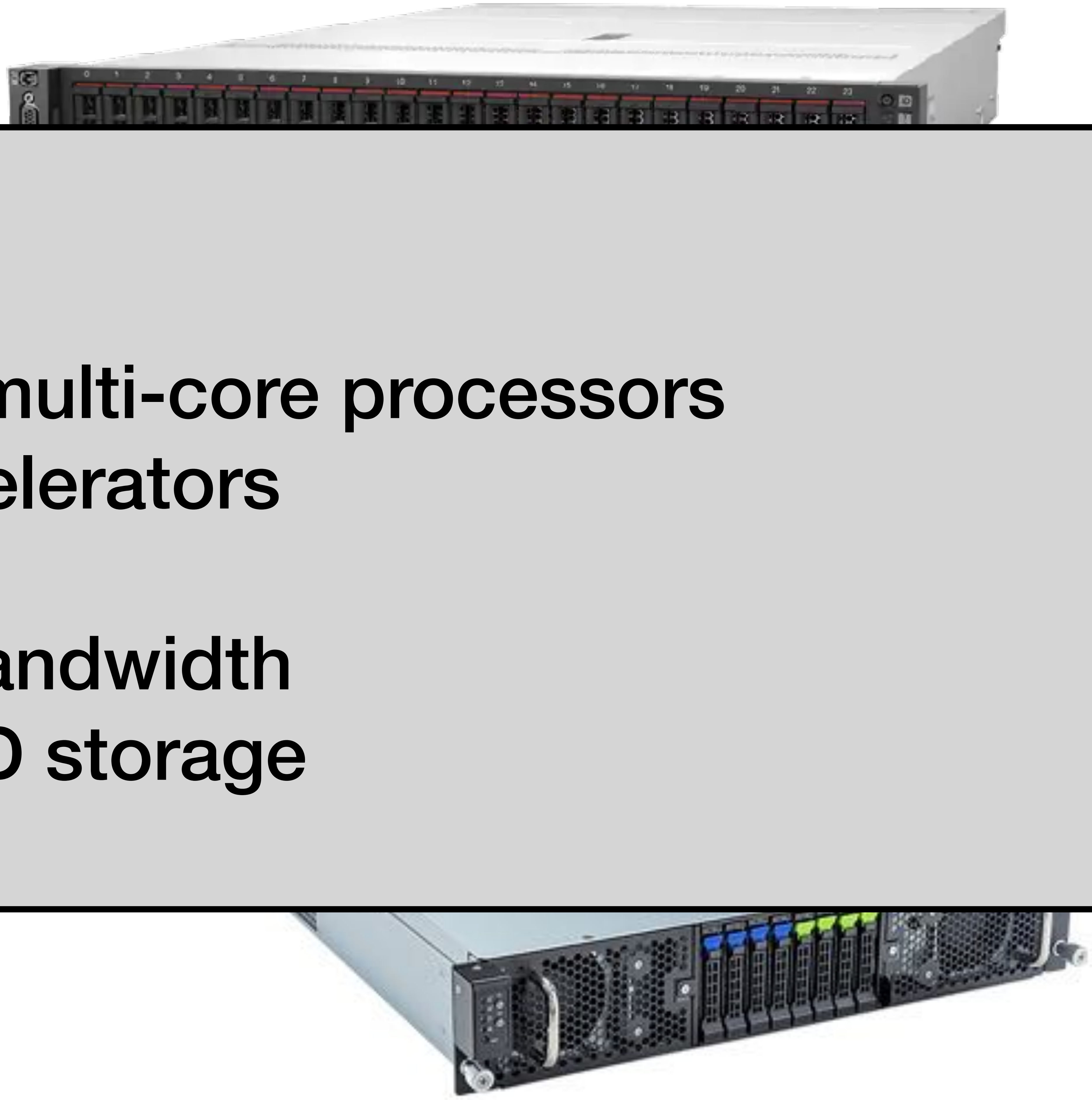
- A server is a computing machine under a predefined form factor.

- **Form factors:**
  - Size: 1U, 2U, 4U, and 5U **1U = 1+3/4 inches (44.45mm)**
  - Shape: rack, tower, and blade
  - Packaging: specifications, supported files, and system firmware

Server and System Architecture

- A server is a computing machine under a predefined form factor.





- **A server consists of**
  - One or several power-hungry multi-core processors
  - Multiple GPU/FPGA/ASIC accelerators
  - Massive memory capacity
  - High-bandwidth networking bandwidth
  - Enormous persistent HDD/SSD storage

Take the Dell PowerEdge R760 as an Example

- Officially released in Feb. 2023
- 2U rack server

Zoom in a Dell R760

- Server front view

24x 2.5-inch drives

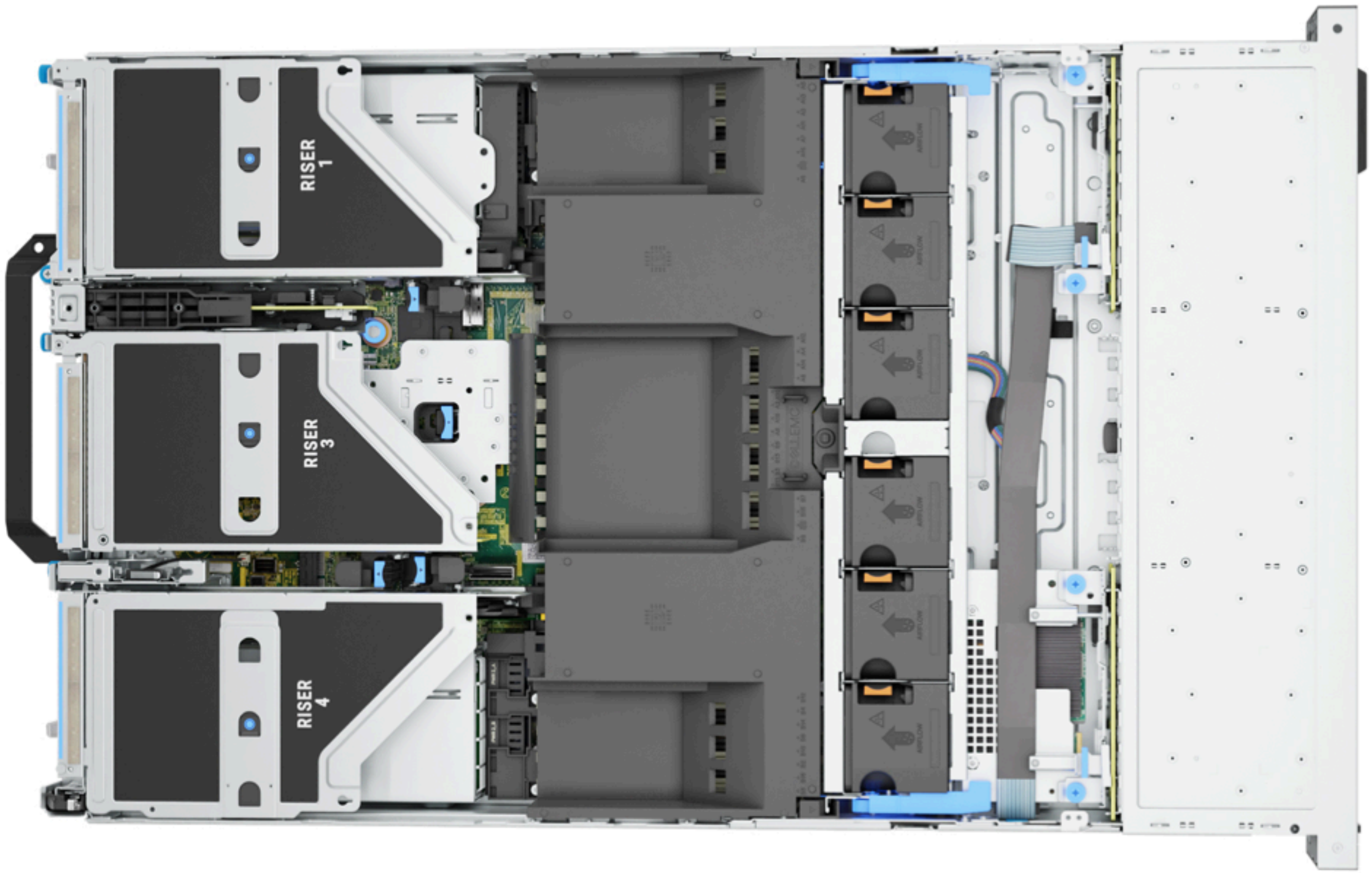
16x EDSFF E3.S
Gen5 NVMe drives

Zoom in a Dell R760

- Server rear view

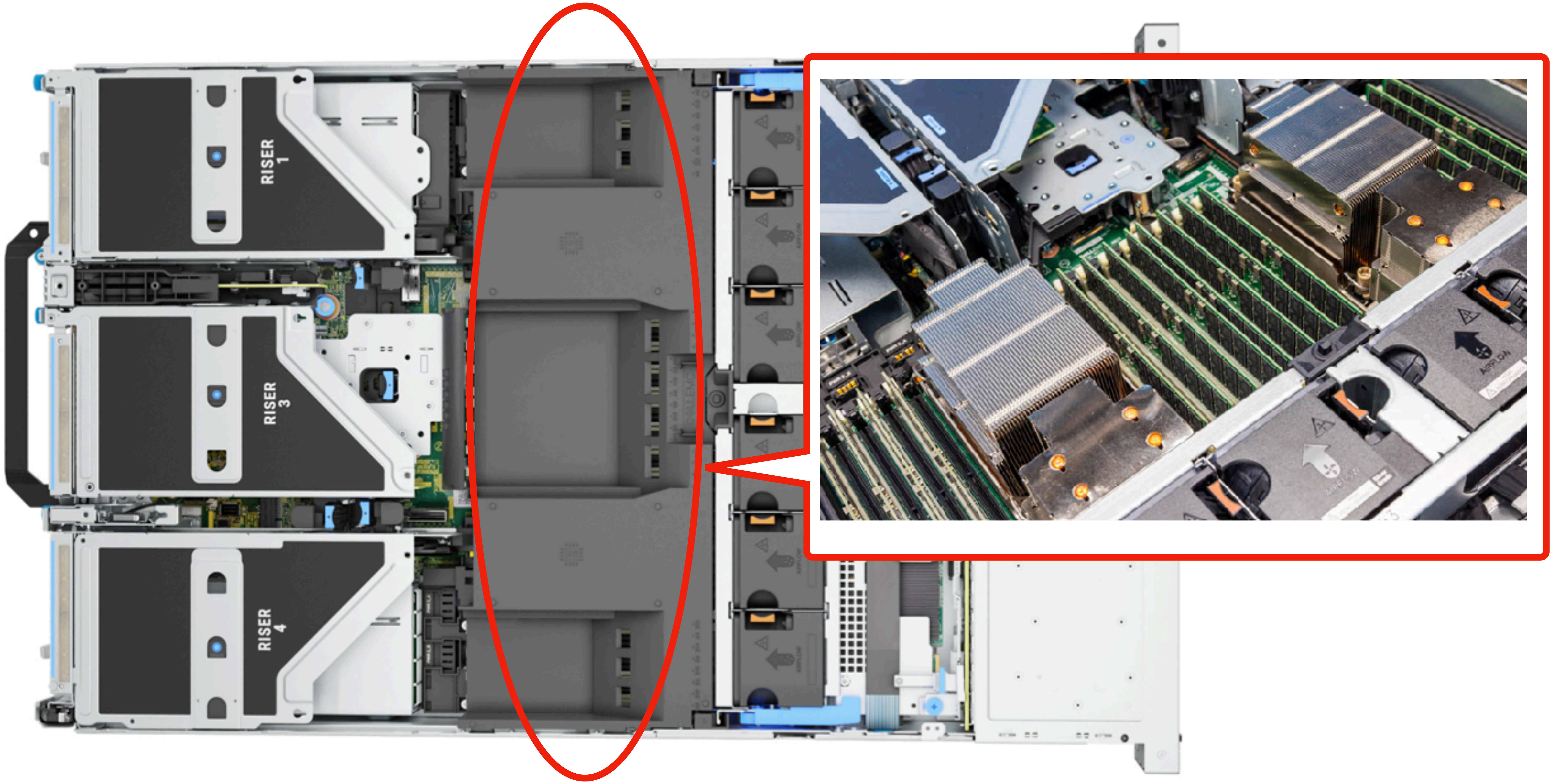
Zoom in a Dell R760

- Inside the server



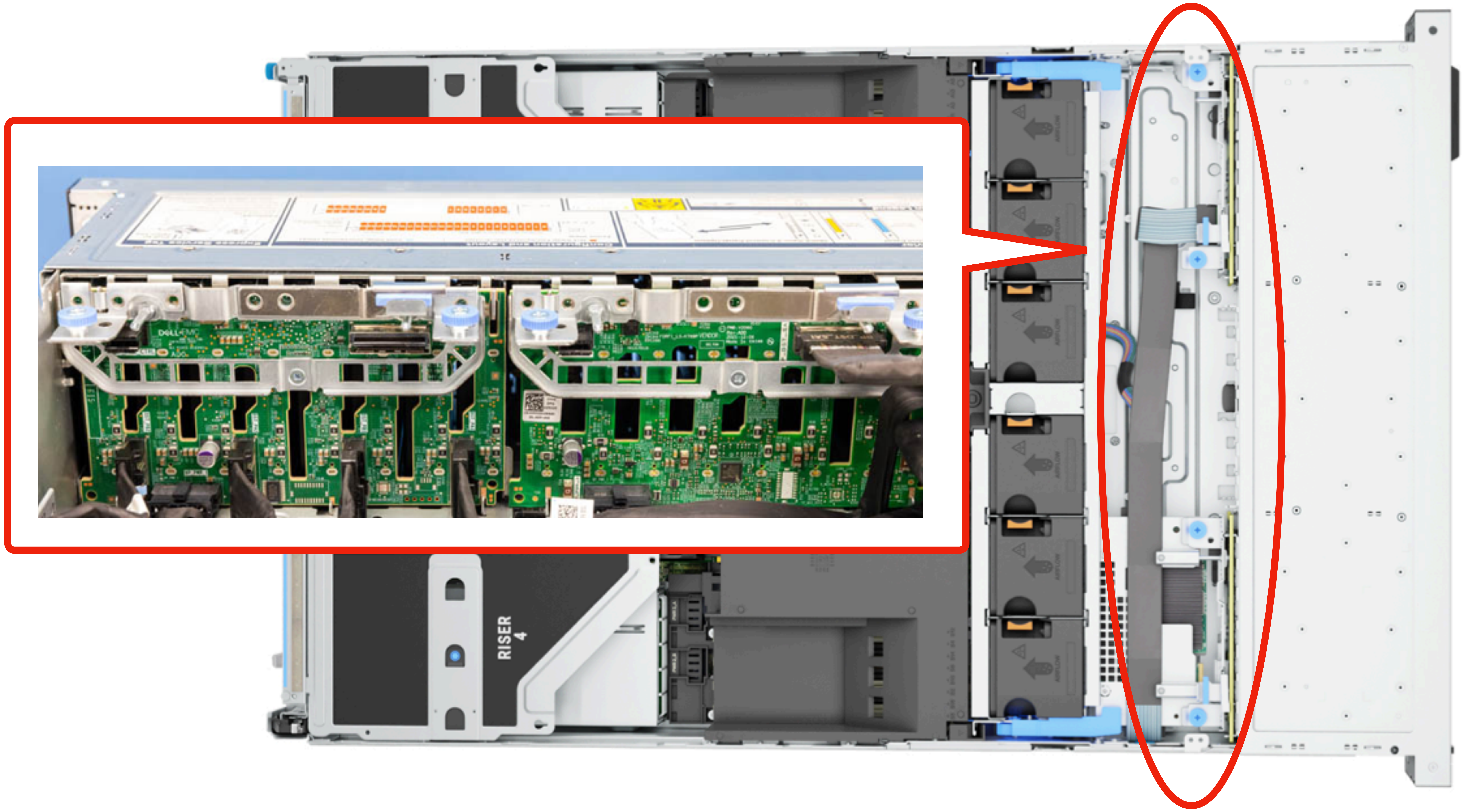
Zoom in a Dell R760

- Inside the server: CPU and Memory



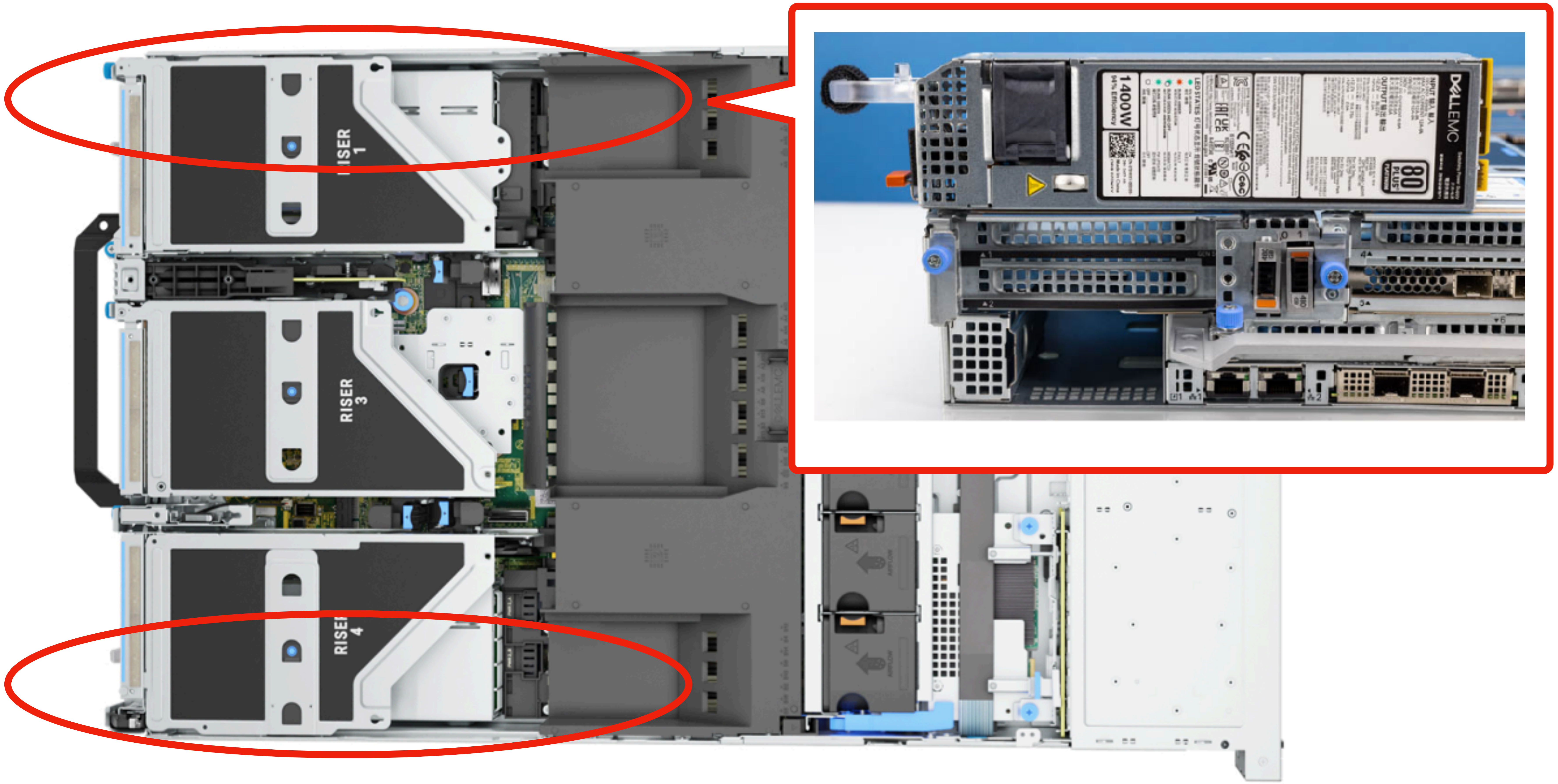
Zoom in a Dell R760

- Inside the server: Storage Backplane



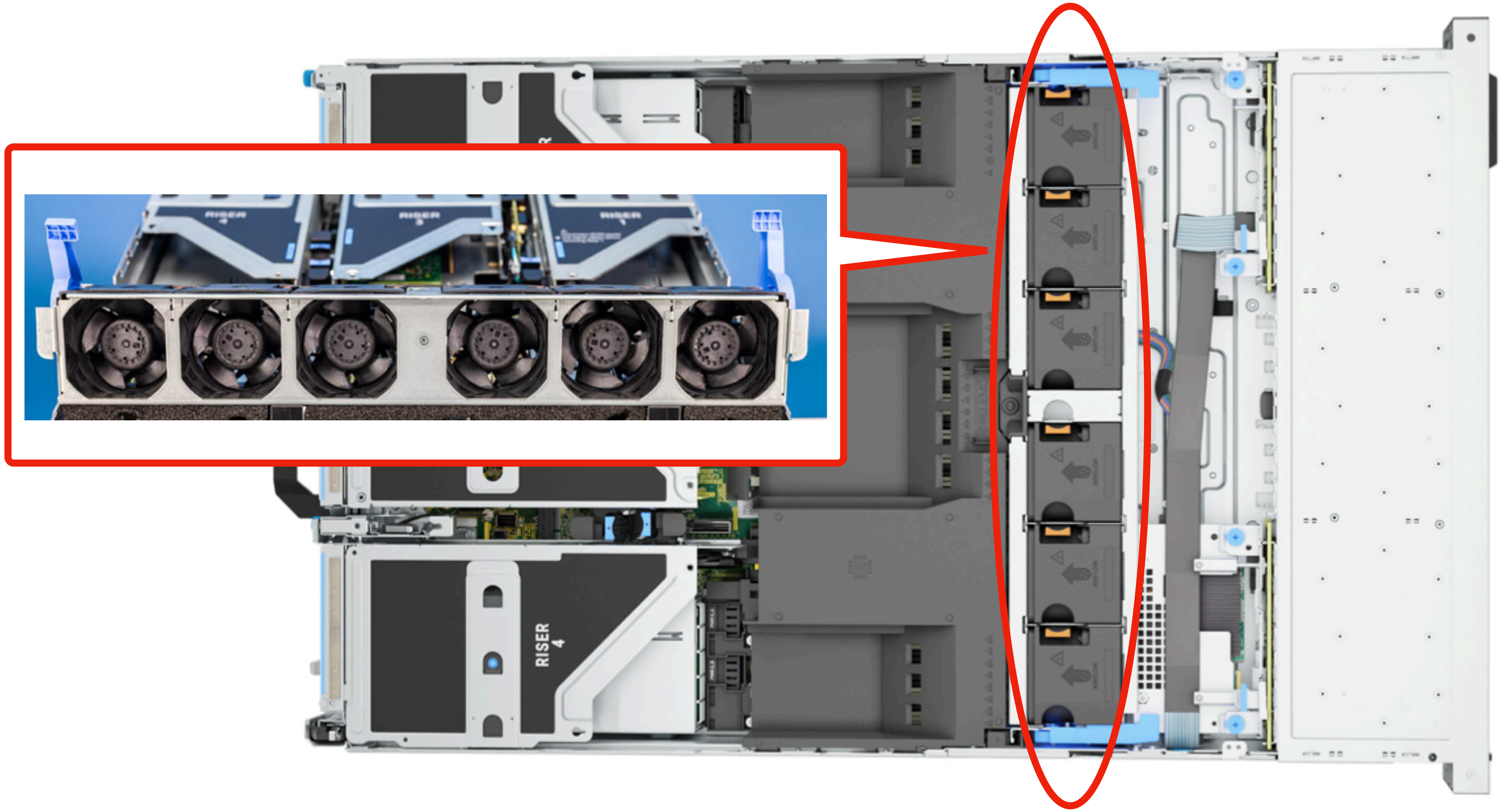
Zoom in a Dell R760

- Inside the server: Power Supply



Zoom in a Dell R760

- Inside the server: Fan

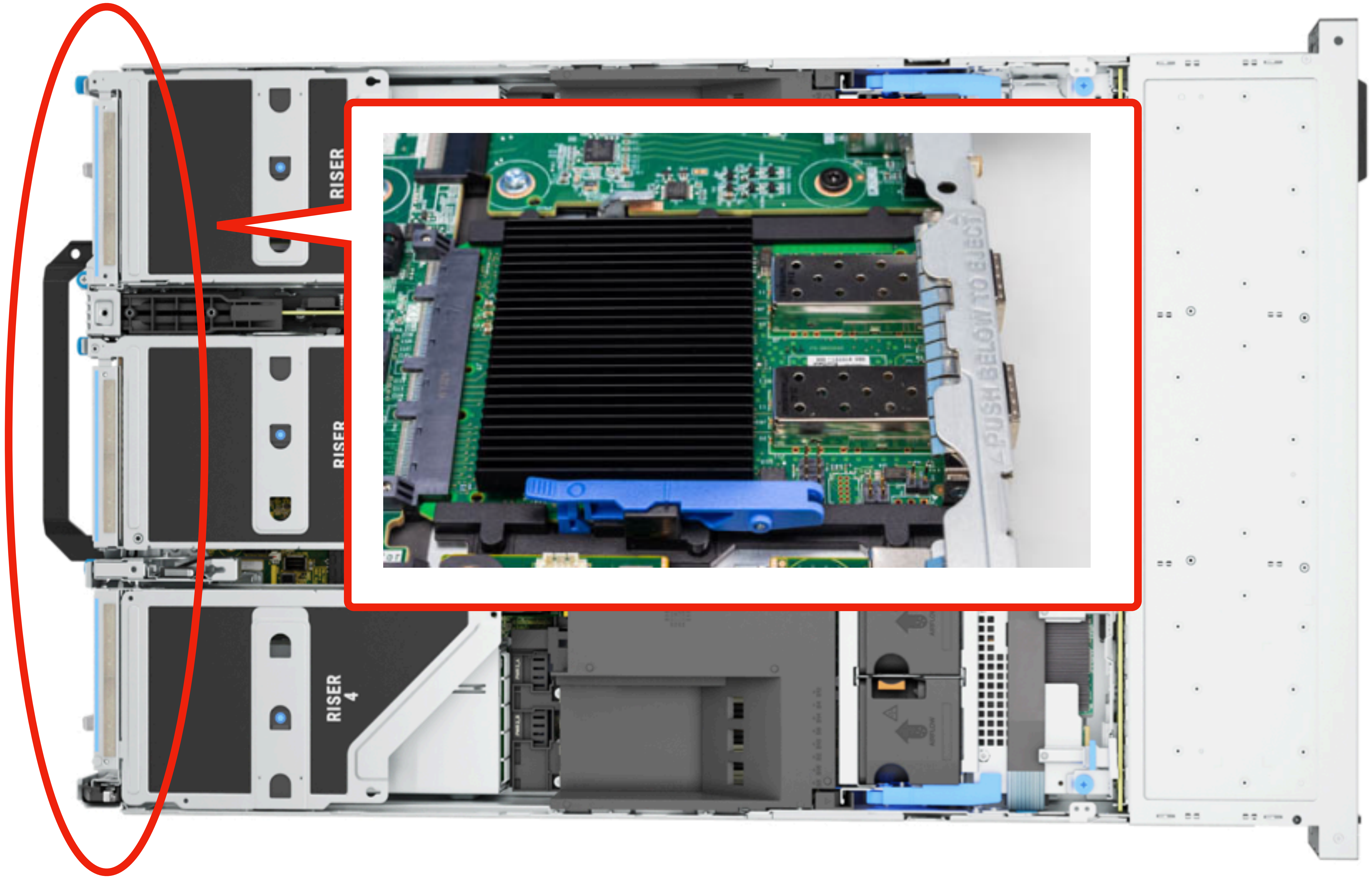


Zoom in a Dell R760

- Inside the server: PCIe cards (GPU, FPGA, and ASIC)

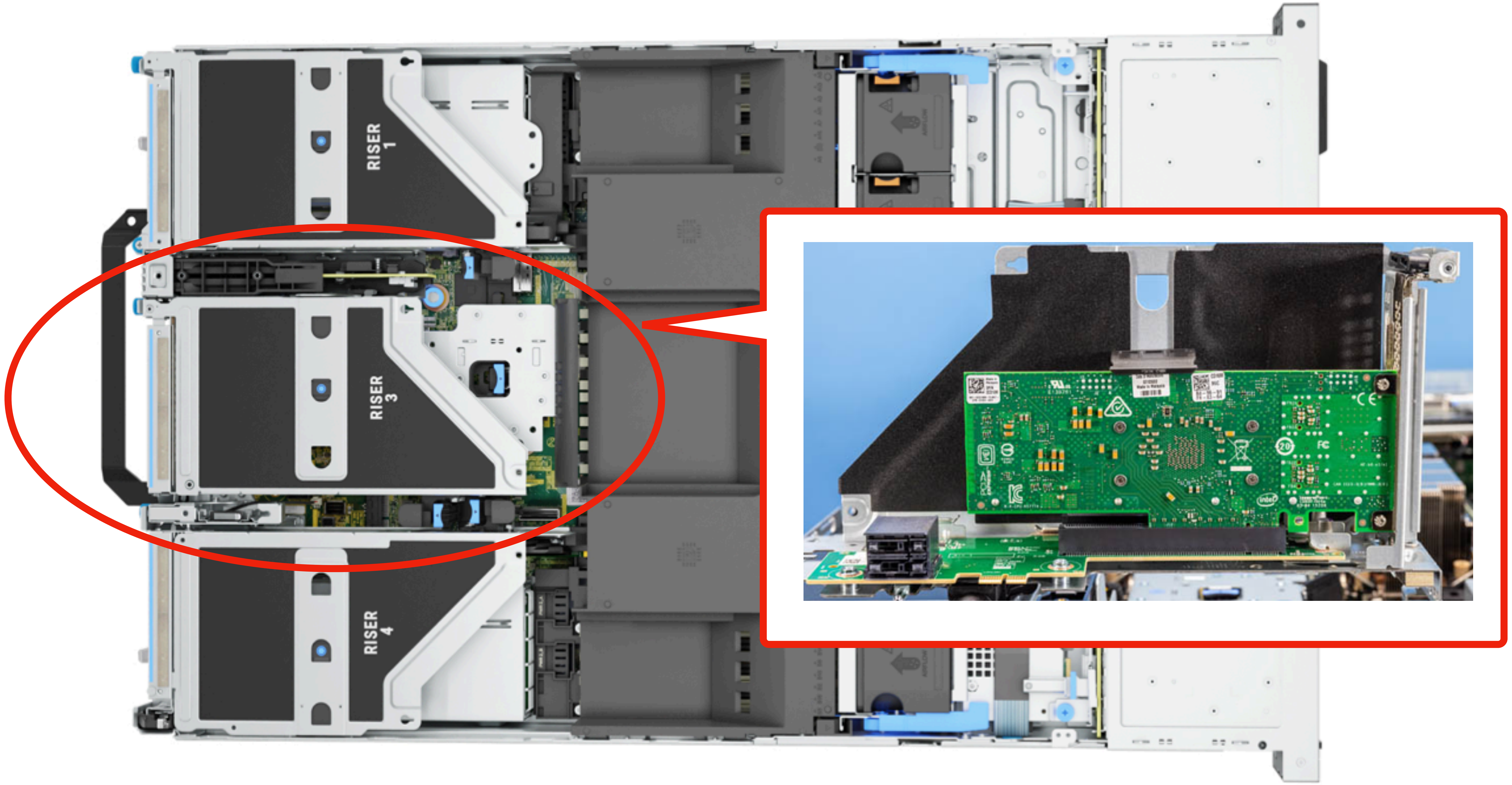
Zoom in a Dell R760

- Inside the server: OCP NIC, others have onboard NIC



Zoom in a Dell R760

- Inside the server: PCIe NIC



How do data center networks connect servers?

Data center networks connect servers via NIC ports!

Servers Connected via NIC ports

Servers Connected via NIC ports

- Management port:
  - Configure and manage a baseboard management controller (BMC)

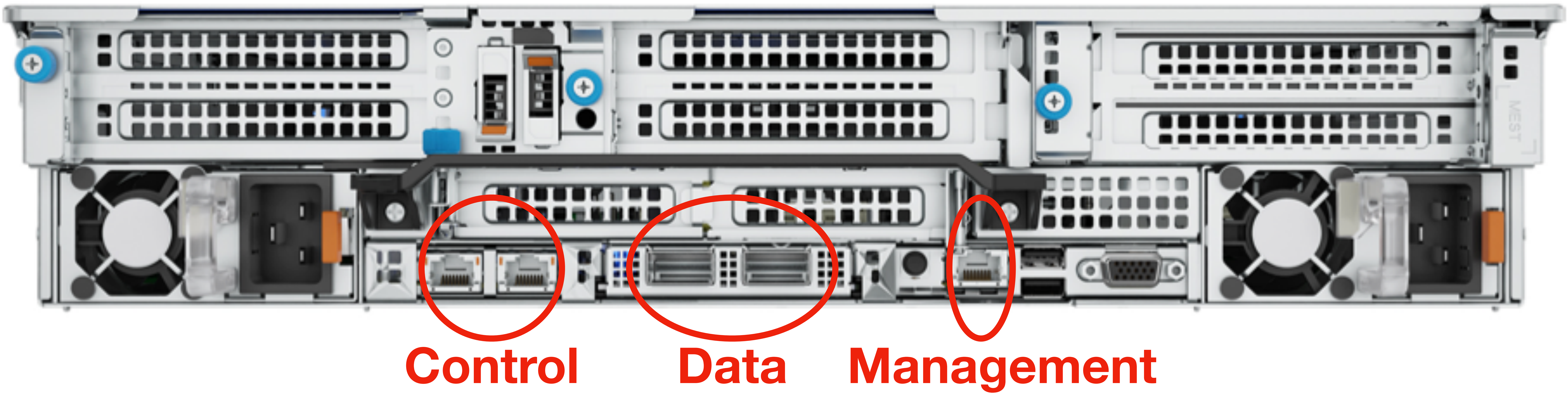
Management

Servers Connected via NIC ports

- Management port:
  - Configure and manage a baseboard management controller (BMC)
- Control network port:
  - Load the server OS and set up the network
  - Initiate and upgrade the server configurations

Servers Connected via NIC ports

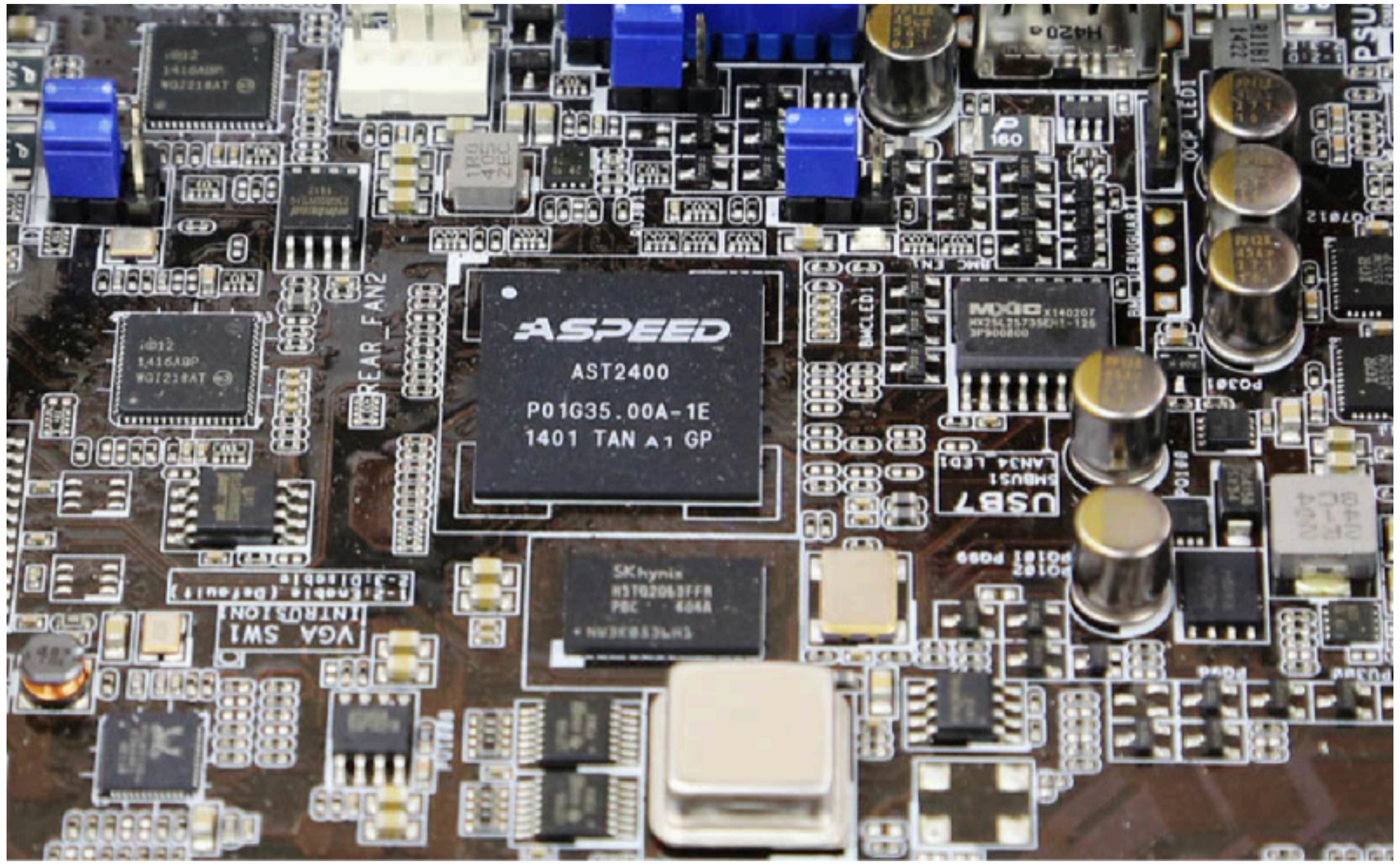
- Management port:
  - Configure and manage a baseboard management controller (BMC)
- Control network port:
  - Load the server OS and set up the network
  - Initiate and upgrade the server configurations
- Data network port:
  - Carry application traffic

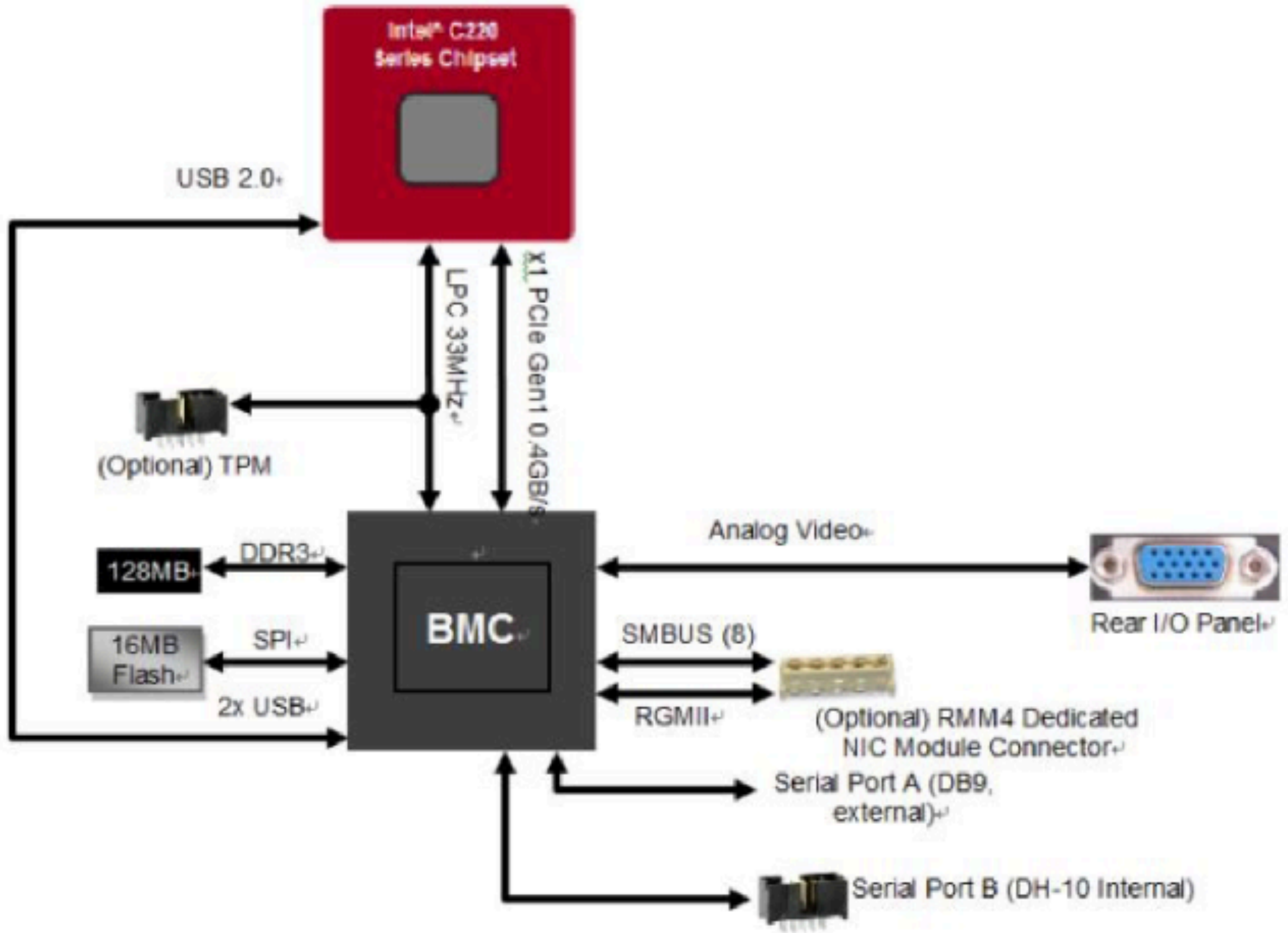


Data center networks provide connectivity for all three ports!

Type #1: Management Traffic

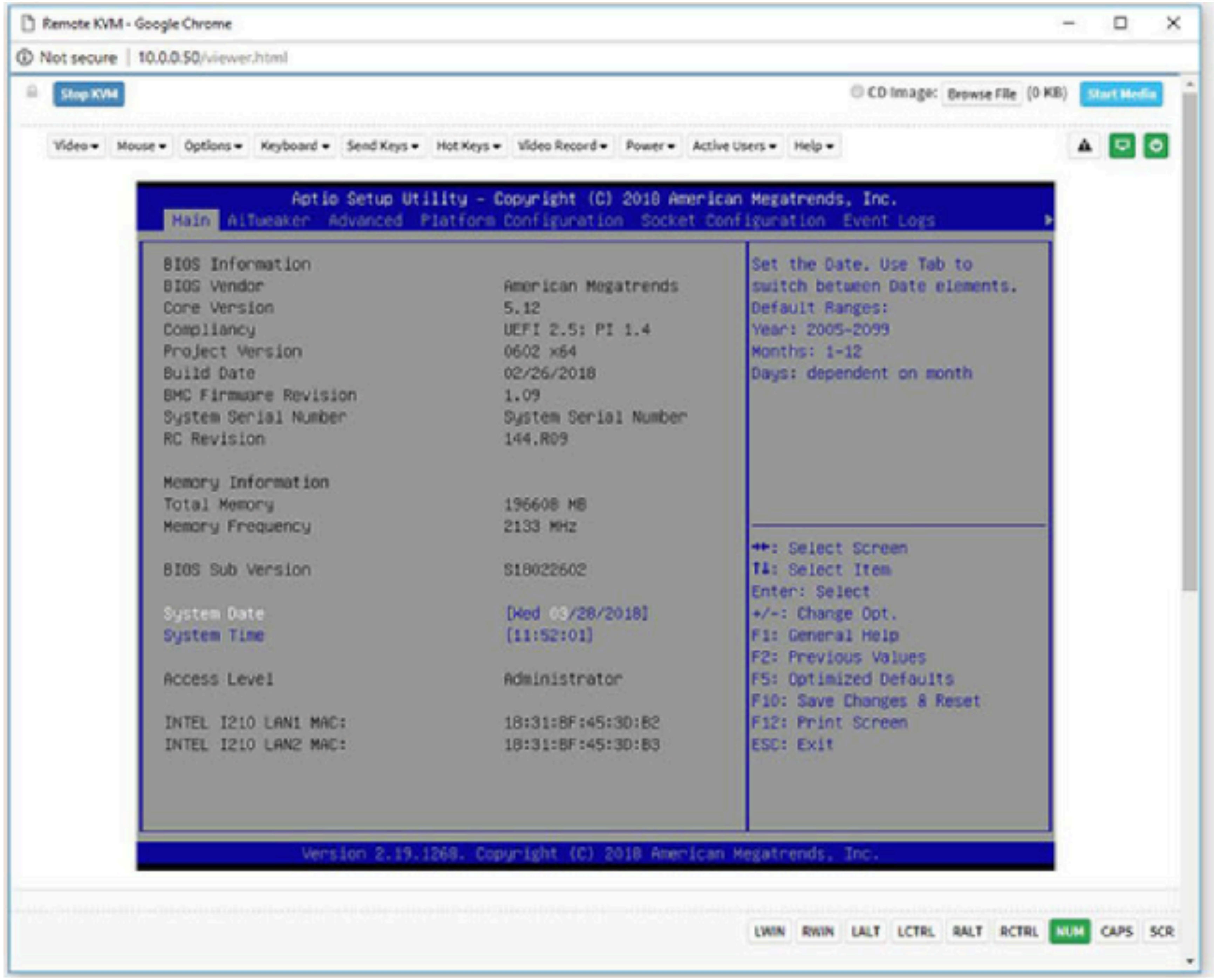
- BMC: an ARM-based SoC with graphics and control logic built-in
  - Vendor: ASPEED
  - There are non-ARM solutions

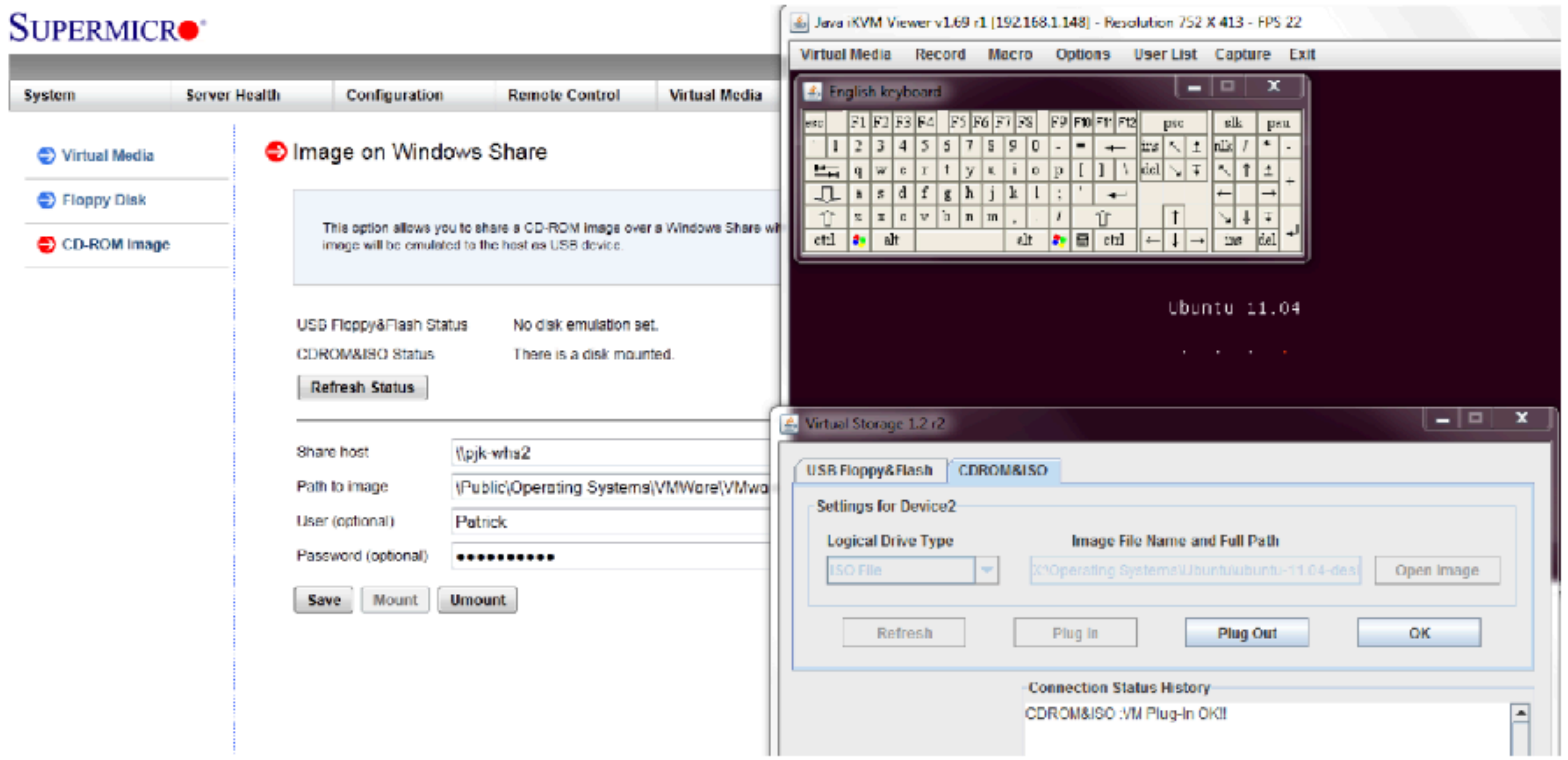




Why do we need this?

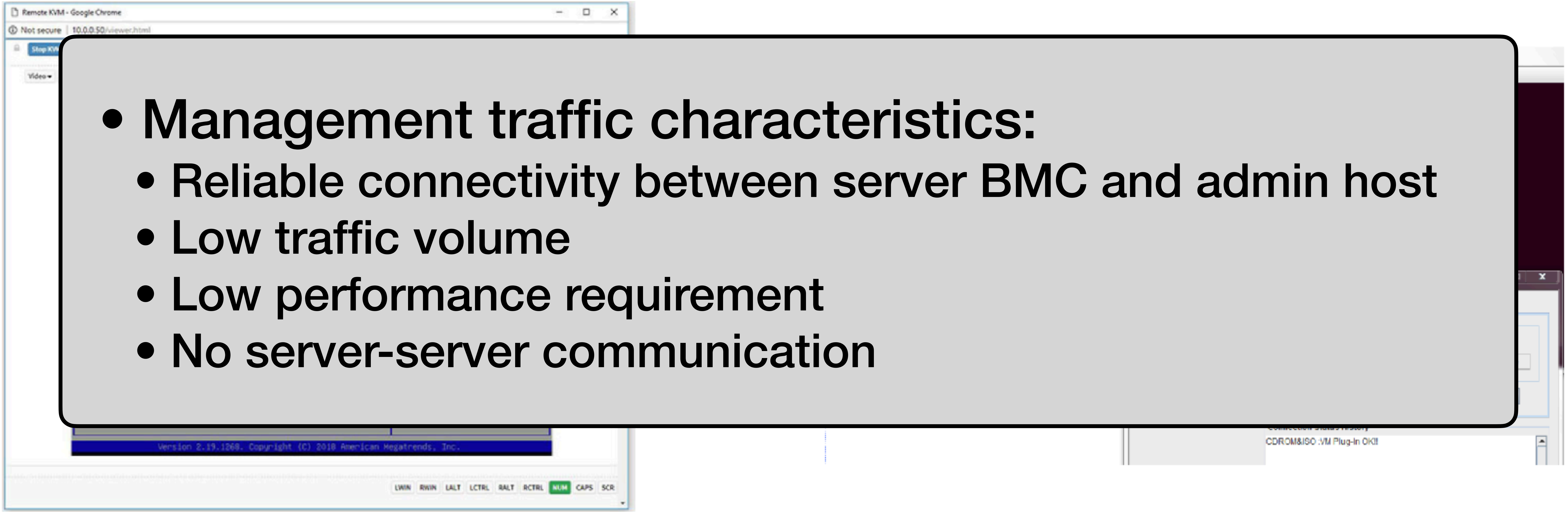
- Target user: server administrator
  - No need to enter the server room physically to perform tasks
  - Virtual KVM





Why do we need this?

- Target user: server administrator
  - No need to enter the server room physically to perform tasks
  - Virtual KVM



- **Management traffic characteristics:**
  - Reliable connectivity between server BMC and admin host
  - Low traffic volume
  - Low performance requirement
  - No server-server communication

Type #2: Control Traffic

- An out-of-band network for server configuration and maintenance
  - A 1G/10G/25G onboard integrated NIC or external one
  - OS load/upgrade, system setup, health monitoring, SSH, …
- Target user: server user or infrastructure engineer
  - Responsible for refactoring servers in a way that satisfies applications
  - Becoming increasingly important today

Type #2: Control Traffic

- An out-of-band network for server configuration and maintenance
  - A 1G/10G/25G onboard integrated NIC or external one
  - OS load/upgrade, system setup, health monitoring, SSH, …

- Target user: server user or infrastructure engineer
  - Responsible for refactoring servers in a way that satisfies applications

- **Control traffic characteristics:**
  - **Reliable connectivity control port and control host**
  - **Modest traffic volume**
  - **Modest performance requrirement**
  - **Partial server-server communication**

Type #3: Data Traffic

- An in-band network for data center applications
  - Dominated by external PCIe-based NICs
  - Vendor: NVIDIA/Mellanox, Broadcom, Intel, …

- Target user: application developers
  - Bandwidth keeps increasing: 1G->10G->25/40G->100G->200G->400G
  - More data ports are added for different use cases.

Type #3: Data Traffic

- An in-band network for data center applications
  - Dominated by external PCIe-based NICs
  - Vendor: NVIDIA/Mellanox, Broadcom, Intel, …

- Target user: application developers
  - Bandwidth keeps increasing: 1G->10G->25/40G->100G->200G->400G

- **Data traffic characteristics:**
  - **Reliable connectivity among servers**
  - **High traffic volume**
  - **High performance requrirement**

We focus on data traffic!

Today

- *Course Logistics*
- *Schedule Overview*
- *Computer Network Design History*
- *Data Center Network Basics*
- Data Center Network Design Requirements

What are the requirements of data center networks?

Top-3 Data Center Networks Requirements

- #1: Highly available connectivity (Metric = MTTF/…)
  - Ideally, always-on connectivity between any endpoint servers
  - In reality, 99.999% (or even six 9s) is challenging
  - Failures always happen in software or hardware

Top-3 Data Center Networks Requirements

- #1: Highly available connectivity (Metric = MTTF/…)
  - Ideally, always-on connectivity between any endpoint servers
  - In reality, 99.999% (or even six 9s) is challenging
  - Failures always happen in software or hardware
- #2: High-performance (Metric = Latency and Bandwidth)
  - Absolute performance: approach physical limits of communication media
  - Performance under multi-tenancy: networking is shared
  - Performance under scalability: networking keeps expanding

Top-3 Data Center Networks Requirements

- #1: Highly available connectivity (Metric = MTTF/…)
  - Ideally, always-on connectivity between any endpoint servers
  - In reality, 99.999% (or even six 9s) is challenging
  - Failures always happen in software or hardware
- #2: High-performance (Metric = Latency and Bandwidth)
  - Absolute performance: approach physical limits of communication media
  - Performance under multi-tenancy: networking is shared
  - Performance under scalability: networking keeps expanding
- #3: Cost-efficiency (Metric = Gbps/$)
  - Capital cost, including NICs, switches, cables, gateways, …
  - Operational cost: configuration, diagnostic, verification,…
  - Nothing is easy under scale!

How can we build a data center network to connect X ports with high availability at Y bandwidth under cost efficiency?

- Connectivity at scale
- X ranging from 1 to millions of ports

How can we build a data center network to connect X ports with high availability at Y bandwidth under cost efficiency?

- Connectivity at scale
- X ranging from 1 to millions of ports

How can we build a data center network to connect X ports with high availability at Y bandwidth under cost efficiency?

- Performance at scale
- Y ranging from Mbps to Tbps

- Connectivity at scale
- X ranging from 1 to millions of ports

How can we build a data center network to connect X ports with high availability at Y bandwidth under cost efficiency?

- Performance at scale
- Y ranging from Mbps to Tbps

- Cost at scale
- The lower the better

Summary

- Today
  - Computer network design history
  - Data center network basics
  - Data center network design requirements

- Next lecture
  - Physical connectivity @ rack and cluster scale