

3. Datacenter Hardware

3.1 Servers and storage systems

3.2 Network infrastructure

3.3 IT apparatus inside the datacenter





Summary

Overview: This module covers the essential hardware and IT infrastructure that power datacenters.

01

Servers

Learn about different types of servers and their components.

02

Storage options

Explore various storage options like SAN, NAS, and object storage.

03

Network infrastructure

Understand key components such as switches, routers, and firewalls, and how external connectivity is managed.

04

IT apparatus

Discover important IT apparatus such as racks, enclosures, and application delivery controllers that optimize performance and organization within the datacenter.

By the end of this module, you will have a solid understanding of the hardware that supports the data storage, processing, and networking needs of a modern datacenter.

Learning Objectives

Understand the different types of servers and storage systems (SAN, NAS, Object Storage) used in datacenters.

Learn about the key components of network infrastructure within a datacenter (switches, routers, firewalls, and external connectivity).

Explore the role of IT apparatus in a datacenter, such as racks, enclosures, and application delivery controllers.

Understand the integration of storage and network systems to support the operational needs of a datacenter.





3.1 Servers and storage systems

- What are a couple of examples of servers and recent innovations for each?
- What are key server components and recent innovations for each?

What are some types of servers and recent innovations for each?

Blade servers: Compact, power-efficient, modular servers

Innovations: Integrated fabric switching, increased density, energy efficiency, AI workload support.

Rack servers: Scalable, flexible, versatile servers

Innovations: Advanced cooling, composable infrastructure, accelerator integration, high-speed connectivity.



What are key server components and recent innovations for each?

Components	Processors (CPUs)	Memory (RAM)	Storage (HDDs/SSDs)	Power Supplies	Cooling Systems
Innovations	Multi-core, AI acceleration, chiplet technology.	DDR5, ECC (Error-Correcting Code) memory, persistent memory.	NVMe (non-volatile memory express) SSDs, zoned storage, high-capacity HDDs.	High-efficiency units, redundancy and failover.	Liquid cooling, smart cooling. Ex. Google's AI-driven cooling systems in datacenters.

What are a few storage options and recent innovations for each?

Storage area networks (SANs): High-speed networks offering block-level access

Innovations: NVMe over Fabrics (NVMe-oF), all-flash arrays (AFAs), automated tiering, AI-powered management.

NAS: File-level storage ideal for shared file storage

Innovations: Scale-out network attached storage (NAS), hybrid NAS, cloud-integrated NAS, data deduplication and compression.

Object storage: Stores data as objects, optimized for unstructured data

Innovations: Cloud-native, on-premises solutions, erasure coding, AI and ML integration, data lifecycle management.



3.2 Network infrastructure

- What are a couple of examples of network infrastructure and recent innovations for each?
- What are key trends shaping network infrastructure?

What are some recent innovations in network infrastructure?

Switches

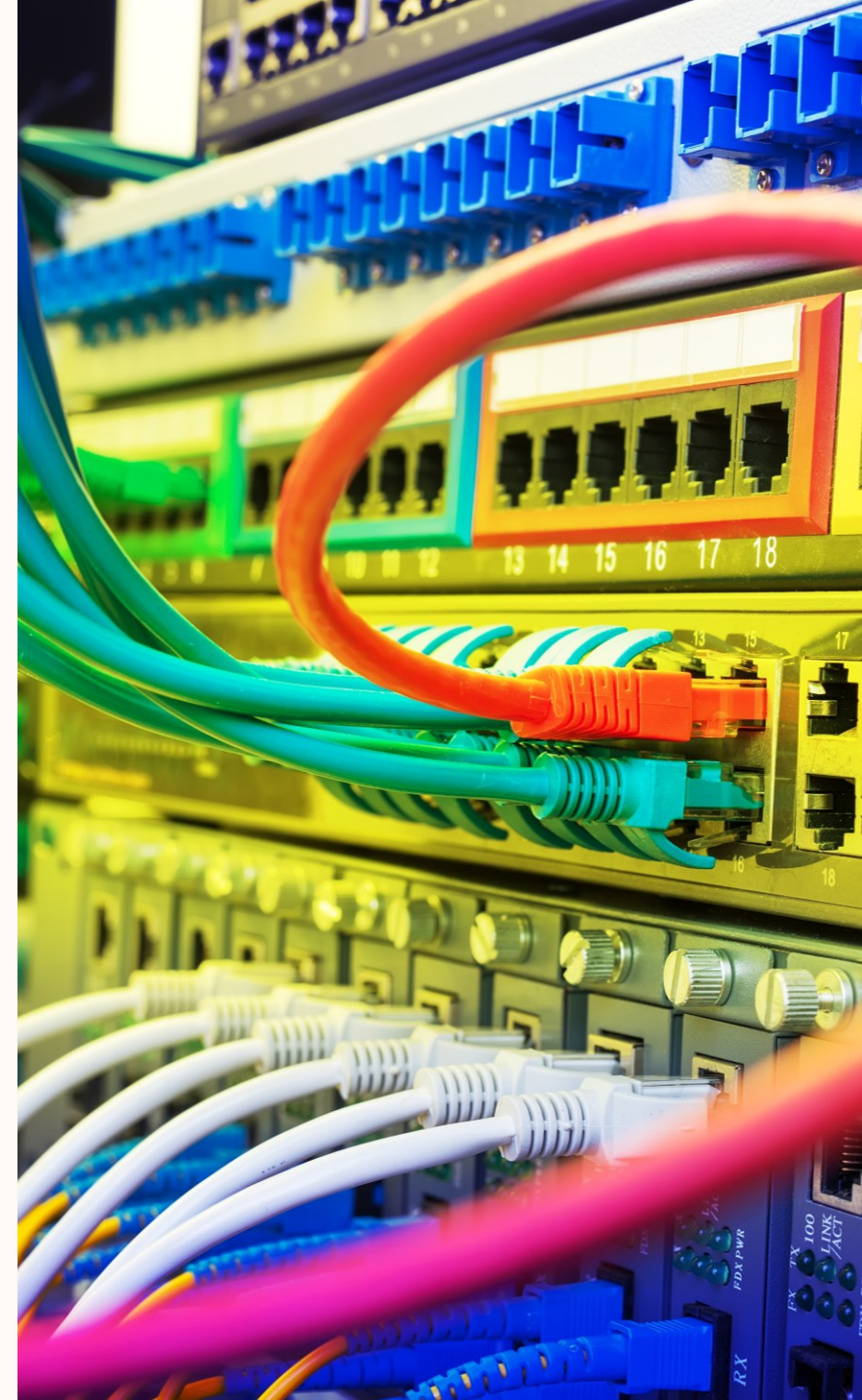
What they do: Interconnect DC devices organized in a hierarchical architecture.

Innovations: High-speed Ethernet, smart switching, zero-touch provisioning, energy-efficient design, SDN integration.

Routers

What they do: Enable communication between local and wide-area networks.

Innovations: High-capacity routing, segment routing (SR), edge routers for multicloud, AI-driven routing, IPv6 support.



What are some recent innovations in network infrastructure?

Firewalls

What they do

Monitor and control traffic based on predefined security rules.

Innovations

Next-generation firewalls (NGFWs), firewall as a service (FWaaS), zero trust network access (ZTNA), decryption and inspection, microsegmentation.

External connectivity

What they do

Connect DC to external networks.

Innovations

Direct cloud connectivity, high-bandwidth optical links, 5G integration, submarine cable systems, software-defined area networking (SD-WAN).

Key trends shaping network infrastructure



Network automation

Automation tools (e.g., Ansible, Terraform) enable efficient network provisioning, configuration, and management.

Energy efficiency

Green networking focuses on reducing power consumption through energy-efficient devices and cooling technologies.

Artificial intelligence and analytics

AI-powered tools analyze network traffic in real-time to detect anomalies, optimize paths, and predict failures.

Hybrid and multicloud networking

Seamless integration of private and public clouds through advanced connectivity solutions is a major focus for modern datacenters.

Disaggregated networking

Separating network hardware and software gives datacenters flexibility to choose best-in-class components and reduce costs.



3.3 IT Apparatus inside the datacenter

- What are a couple of IT apparatus examples and recent innovations for each?
- What are key trends shaping IT apparatuses?

Racks and enclosures

What they do

House hardware, manage cooling, and ensure security.

Recent Innovations

High-density racks: Support higher power loads and cooling for modern hardware like GPUs used in AI and machine learning.

Integrated cooling solutions: Built-in mechanisms such as liquid cooling, rear-door heat exchangers, and airflow management systems for high-density deployments.

Intelligent rack management: Sensors for monitoring temperature, humidity, and power consumption, enabling better management and reduced downtime.

Physical security enhancements: Biometric locks, RFID access control, and video monitoring to prevent unauthorized access.

Cable management innovations: Improved cable trays, harnesses, and labels simplify cabling in high-density racks, reducing maintenance time and risk of errors.



Application delivery controllers (ADCs)



What they do: Optimize performance, availability, and security for applications.

Multi-cloud and hybrid cloud support: Seamlessly manage application traffic across on-premises, cloud, and hybrid environments (e.g., F5 BIG-IP, Citrix ADC).

AI-powered optimization: AI and ML analyze traffic patterns, detect anomalies, and optimize traffic distribution dynamically.

Kubernetes and container integration: Support microservices and containerized applications, enabling efficient traffic management in environments like Kubernetes clusters.

TLS 1.3 and QUIC support: Ensure faster, more secure application delivery, improving user experiences.

Web Application Firewall (WAF) integration: Protect applications from common threats like SQL injection and cross-site scripting (XSS).

API gateway features: Facilitate secure and efficient API traffic management for modern applications.

Edge computing and CDN support: Integrate with edge platforms and Content Delivery Networks (CDNs) to reduce latency and enhance performance for geographically distributed users.

Key considerations and trends



Efficiency and sustainability

As datacenters focus on energy efficiency, racks and application delivery controllers (ADCs) are being designed with sustainability in mind, such as energy-efficient power distribution and intelligent workload balancing.

Security integration

Both physical (racks) and logical (ADCs) elements now include advanced security features to address growing cyber security threats.

Automation and orchestration

ADCs and rack management tools are increasingly integrated into automation frameworks, simplifying operations and enhancing responsiveness.

Edge datacenters

Compact racks and lightweight ADCs are enabling the deployment of edge datacenters closer to end-users for ultra-low latency applications.

Open standards

Open rack standards (OCP) are gaining traction, promoting interoperability and reducing costs by standardizing rack dimensions and features.

Review: Datacenter hardware



Identify different types of servers (blade servers, rack servers) and recent innovations (e.g., liquid cooling, AI accelerators).



Understand key server components (central processing units (CPUs), memory, storage, power supplies, cooling systems) and how advancements improve efficiency and performance.



Differentiate between storage options:

- Storage Area Network (SAN): High-speed, block-level storage for mission-critical applications.
- Network Attached Storage (NAS): File-based storage for shared access over a standard network.
- Object Storage: Scalable storage optimized for unstructured data, used in cloud and large-scale environments.



Recognize the role of network infrastructure in datacenters, including:

- Switches (enable device interconnectivity, support high-speed Ethernet).
- Routers (connect datacenters to external networks and manage traffic).
- Firewalls (provide security through intrusion prevention, encryption, and segmentation).
- External Connectivity (ensures high-speed, redundant links to cloud providers and other datacenters).



Understand IT apparatus like racks, enclosures, and application delivery controllers that optimize datacenter organization and efficiency.

4. Datacenter Operations

4.1 Monitoring

4.2 Maintenance

4.3 Physical security

4.4 Low latency operations



Summary

Overview: This module covers the essential operations that ensure the efficiency, reliability, and security of datacenters.

01

Monitoring

Learn who is responsible for datacenter monitoring.

02

KPIs

Understand how Key Performance Indicators (KPIs) track datacenter health.

03

Maintenance

Explore the importance of regular maintenance in hardware, software, power systems, cooling systems, and fire protection systems.

04

Disaster recovery and low latency

Discover how datacenters recover after unexpected disruptions and how to minimize delays in data transmission.

By the end of this module, you will have a solid understanding of maintenance and monitoring best practices.

Learning Objectives

Gain knowledge of how datacenters are monitored in real-time, including environmental, performance, security, and infrastructure monitoring.

Learn about the key performance indicators (KPIs) used to track datacenter health and operational efficiency.

Understand the role of automation and predictive monitoring in preventing downtime and ensuring reliability.

Explore the importance of regular maintenance in hardware, software, power systems, cooling systems, and fire protection systems.

Learn about strategies for disaster recovery and low latency operations.





4.1 Monitoring

- What is datacenter monitoring?
- Who is responsible for datacenter monitoring?
- What are key aspects of datacenter monitoring?

What is datacenter monitoring?

Datacenter monitoring ensures all systems are functioning optimally.

It ensures servers, storage devices, cooling systems, power units, and networks operate within safe and efficient parameters.

Through proper datacenter monitoring, reliable and high-performance datacenter operations can be maintained.



Who is responsible for datacenter monitoring?

By type of datacenter:

Enterprise datacenters

Owned and operated by large corporations (e.g., banks, hospitals). Monitoring is typically handled by **in-house teams or specialized operations staff**.

Colocation datacenters

Typically, **datacenter providers** handle physical infrastructure monitoring while **clients** monitor their own equipment and software (e.g., servers, storage).

Hyperscale datacenters

They are usually operated by large cloud providers (e.g., Amazon, Microsoft). Monitoring is highly automated, and **specialized teams** oversee environmental and performance monitoring.

Edge datacenters

These are smaller facilities located closer to end-users. **Local teams** and **automated monitoring systems** are responsible for maintaining uptime and optimal performance.

Managed services datacenters

The **provider** is responsible for both infrastructure and service-level monitoring. These datacenters offer end-to-end managed services.

What are key aspects of datacenter monitoring and why are they important?



Environmental: Important to maintain optimal conditions that prevent equipment failure and improve efficiency.

Performance: Helps ensure all IT sources are performing as expected.

Security: Protects data and infrastructure from unauthorized virtual and in person access.

Infrastructure: Monitors physical systems that provide power, cooling, and fire suppression.

Network and connectivity: Ensures datacenter network connections are always online and functioning efficiently.

Compliance: Ensures the datacenter is operating in accordance with industry standards, regulations, and customer contracts.



What are **key elements** for each of the datacenter monitoring categories?

Environmental: Temperature, humidity, airflow, water leakage.

Performance: Server, storage, network, and application performance.

Security: Physical, digital, cyber, and access control security.

Infrastructure: Power, cooling, fire suppression, generators and battery backups.

Networks and connectivity: Bandwidth usage, latency, packet loss, redundancy and failover systems.

Compliance: Regulatory compliance, audit trails, data retention.

What **tools** can be used for each of the datacenter monitoring categories?

Environment

Sensors that measure temperature, humidity, airflow, and water leakage.

Performance

Software and other tools that monitor server, application, storage, and network performance.

Security

Physical security, cybersecurity, access control systems, and video surveillance systems.

Infrastructure

Devices that monitor uninterruptible power supplies (UPS), cooling systems, and fire suppression systems.

Network and connectivity

Bandwidth and latency monitoring tools, packet loss detection, and redundancy and failover monitoring.

Compliance

Compliance management software, audit and logging systems, and data retention monitoring tools.

What **KPIs** can be used for each of the datacenter monitoring categories?



Environment

Temperature, humidity, airflow (CFM – cubic feet per minute), water leakage detection.

Infrastructure

Power usage effectiveness (PUE), UPS battery health, energy and cooling system efficiency, fire suppression system readiness.

Performance

CPU utilization, memory usage, disk health and usage, disk read/write latency, network throughput, latency, mean time between failures (MTBF), mean time to repair (MTTR), availability.

Network and connectivity

Network bandwidth utilization, latency, packet loss, network downtime, jitter.

Security

Intrusion detection, security incident response time, firewall logs.

Compliance

Audit results, incident reporting, data retention.



4.2 Maintenance

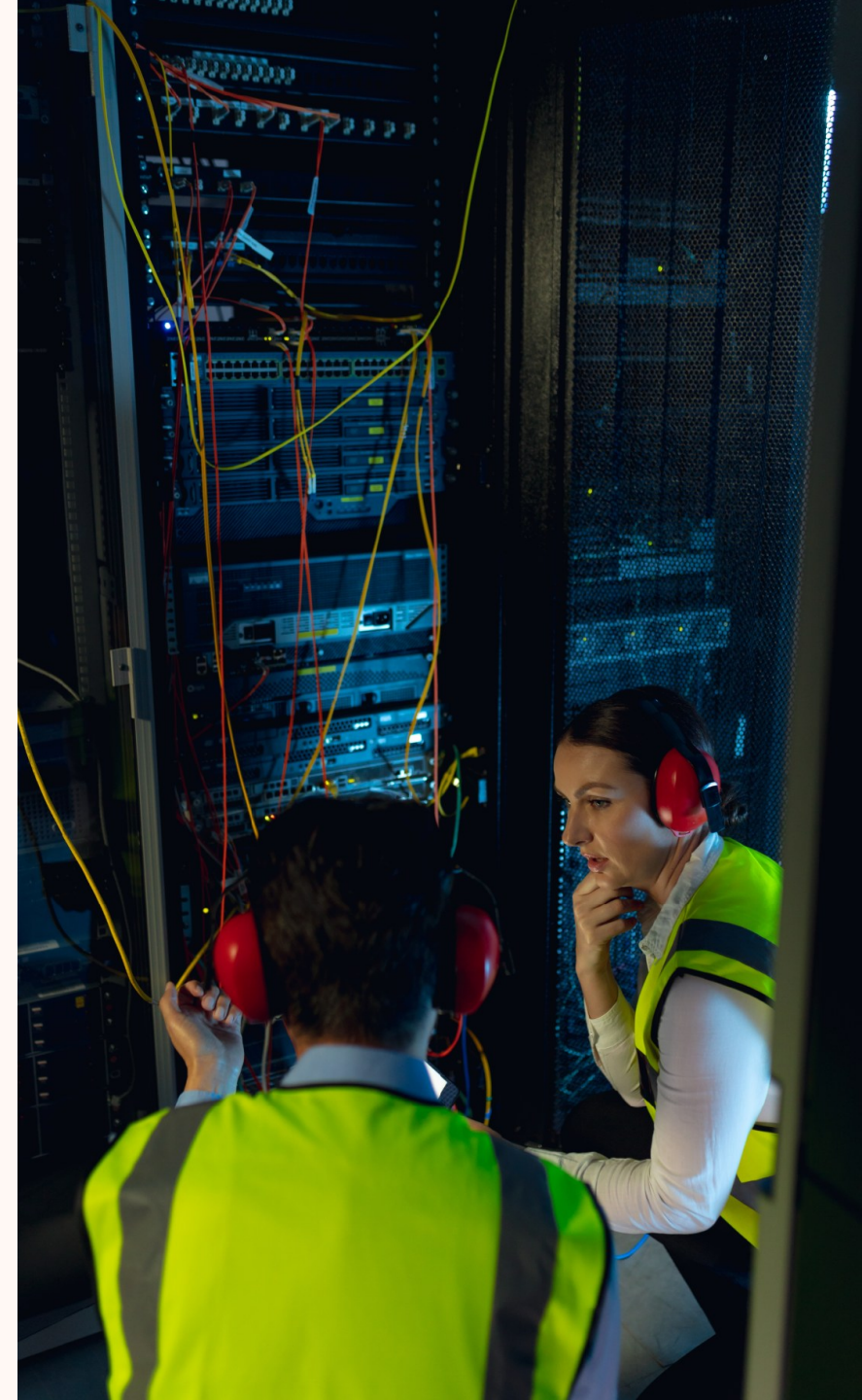
- What is datacenter maintenance?
- Who is responsible for datacenter maintenance?
- What are key aspects of datacenter maintenance?
- What are the general concepts of datacenter maintenance?
- What are key concepts of reliability and redundancy in datacenters?

What is datacenter maintenance?

Datacenter maintenance focuses on keeping all infrastructure and equipment running smoothly and efficiently.

It involves regular inspection, servicing and updating of systems (e.g., servers, power supplies).

Through proactive datacenter maintenance, datacenters can address potential issues before they lead to failures which can reduce downtime.



Who is responsible for datacenter maintenance?

By type of datacenter:

Enterprise datacenters

The organization that owns the datacenter is responsible for its maintenance. **In-house IT staff** or **specialized datacenter operations teams** manage maintenance.

Colocation datacenters

The responsibility for maintenance is shared between **services provider** and **the client (tenant)**.

Hyperscale datacenters

Maintenance is fully managed by the provider, and **specialized teams** handle all maintenance tasks and issues.

Edge datacenters

If an edge datacenter is owned by a company or service provider, then the **provider** is responsible for maintenance. If it is owned by a third-party provider, the **third-party provider** is typically responsible for maintenance while the **customer** manages their own equipment.

Managed services datacenters

The **service provider** is primarily responsible for maintaining the datacenter infrastructure and the customer's hardware.

What are key aspects of datacenter maintenance?



Hardware

Servers and storage devices, network equipment, power systems.

Environmental

Temperature and humidity control, water leakage detection, smoke, and fire detection.

Cooling system

VAC (Heating, ventilation, air conditioning), chillers and cooling units, airflow management.

Software and firmware

Operating system and software patches, firmware updates, monitoring, and alerting systems.

Power and electrical

Electrical distribution, backup power, energy efficiency.

Data backup and disaster recovery

Backup systems, disaster recovery testing.

Security

Physical security, cybersecurity, surveillance.

Documentation and record keeping

Maintenance logs, compliance, and audits.



What are the general concepts of datacenter maintenance and a few examples of each?

Preventative: Replacing aging hardware components, cleaning equipment, performing inspections of cooling and power systems.

Corrective/Reactive: Fixing/replacing failed hardware, repairing malfunctioning cooling systems, restoring services after an outage or disruption.

Routine or operational: Monitoring system logs, performing routine backups, inspecting physical and digital security, checking environmental conditions.

Predictive: Power systems that monitor the health of generators, monitoring disk drives for increasing read/write errors, collecting real-time data on humidity or air quality.

What are reliability and resiliency in datacenters?

Reliability refers to the ability of the datacenter to consistently perform its functions without failure.

Example metrics or strategies:

- Uptime percentage
- Mean time between failures (MTBF)
- Mean time to repair (MTTR)

Resiliency refers to the ability of the datacenter to continue functioning during and after a failure, disruption, or disaster.

Example metrics or strategies:

- Redundancy in power, cooling, and network systems
- Disaster recovery and backup systems
- Fault-tolerant infrastructure



What are key concepts of reliability and resiliency in datacenters?

Redundancy

Multipath network connections for uninterrupted connectivity, cooling system backups.

Disaster recovery (DR)

Offsite backups for data recovery, warm or cold datacenter sites.

Environmental controls

HVAC (Heating, ventilation, and air conditioning) systems, water leak detection systems.

Fault tolerance

RAID (Redundant array of independent disks) for data storage, load balancing.

High availability (HA)

Active-active server configurations, failover mechanisms for critical applications.



4.3 Disaster recovery in datacenters

- What is datacenter disaster recovery?
- What are examples of disaster scenarios?
- What are key concepts of disaster recovery?

What is datacenter disaster recovery (DR)?

Datacenter disaster recovery ensures that systems, applications, and data can be restored after an unexpected disruption.

It is a structured approach to maintaining business continuity after events like natural disasters and cyberattacks.



What are a few disaster scenarios?



Natural disasters

Hurricanes, earthquakes, or floods.

Cyberattacks

Ransomware, Distributed Denial of Service (DDoS).

Human error

Accidental deletion of important files, misconfigurations.

Hardware error

Server crashes, power outages, or cooling system breakdowns.

What are key concepts of disaster recovery and a few examples of each?



Recovery time objective (RTO)

A 4-hour RTO means systems must be operational within 4 hours of an outage.

Backup and restore

Full, incremental, or differential backups.

Recovery point objective (RPO)

A 1-hour RPO means no more than 1 hours' worth of data can be lost.

Redundancy

Dual power sources, replicated servers, and geographically distributed datacenters.

Disaster recovery plans (DRPs)

A documented strategy to respond to disasters.

Tiered recovery approach

Restoring critical systems first and non-critical systems once the business is stable.



4.4 Low latency operations in datacenters

- What are latency and low latency?
- What are the types of latency?
- Why is low latency important?
- What are factors influencing latency?
- What are techniques for achieving low latency in datacenters?
- How can we measure latency?
- What sectors or services need low latency?
- What are challenges to achieving low latency?

What are latency and low latency?

Low latency refers to minimizing the delay in data transmission between systems, applications, and users.

Latency is the delay between a request for data and the delivery of that data.



What methods or tools can be used to monitor each key aspect of a datacenter?



Network latency

Time taken for data to travel across a network.

Processing latency

Time taken for servers or systems to process a request.

Storage latency

Time taken to read/write data from storage devices.

Why is low latency important?



Enhanced user experience: Applications like video streaming, online gaming, and virtual reality require seamless performance.

Critical business operations: In financial trading or healthcare, delays can lead to significant losses.

Support for real-time applications: Applications like IoT and autonomous vehicles depend on low latency for optimal functionality.



What are key factors influencing latency and a few solutions for each?

Distance: Data transmission speed can be constrained by the physical distance between users and the datacenter.

Network infrastructure: The quality of network hardware can significantly impact latency.

Datacenter architecture: Server design, processors, and software can impact processing and queuing delays.

Storage systems: Older storage technologies have higher latency compared to modern alternatives like solid state drives (SSDs) or non-volatile memory express (NVMe).

Congestion: High traffic can slow down data transfer rates.

What are techniques for achieving low latency in datacenters?

Edge computing

Move processing closer to the end-user reduces network latency.

High-speed networking

Use of advanced technologies like fiber-optic networks and software-defined networking (SDN).

Optimized hardware

Utilizing cutting-edge server hardware and faster storage systems.

Automation and AI

AI-driven systems reduce unnecessary delays.

Data prioritization

Traffic shaping and quality of service (QoS) policies reduce delays.

How can we measure latency?



Ping time: Measures round-trip time for data packets.

Throughput: Amount of data successfully delivered within a specified timeframe.

Jitter: Variability in latency over time, which can disrupt real-time applications.

What sectors or services need low latency?



Financial
services

Stock trading
platforms.

Online
gaming

Lag-free gameplay.

Video streaming and
virtual reality (VR)

Streaming and
immersive VR
experiences.

Autonomous
vehicles

Real-time
communication
between vehicle
sensors and
control sensors.

Internet of things
(IoT)

Thermostats and
industrial sensors.

What are challenges to achieving low latency?



Global reach: Serving users across the world requires creative solutions like global edge datacenters and content delivery networks (CDNs).

Scalability: As user bases grow, maintaining low latency while scaling infrastructure is a challenge.

Legacy systems: Older infrastructure may not support modern low latency demands and may require upgrades.

Review: Datacenter maintenance

- ✓ Gained knowledge of how the key elements of environmental, performance, security, infrastructure, network & connectivity, and compliance operations are monitored.
- ✓ Reviewed key performance indicators (KPIs) such as temperature, memory usage, access control, UPS battery health, packet loss, and incident reporting.
- ✓ Reviewed the role of automation and predictive monitoring in preventing downtime and ensuring reliability (e.g., break-fixing and monitoring UPS battery health).
- ✓ Understand key aspects of datacenter reliability and redundancy (e.g., RAID, warm or cold datacenters).
- ✓ Learn about strategies for disaster recovery (e.g., RTO and DRPs) and challenges to low latency operations (e.g., global reach and scalability).