MODULE 1 · STUDY NOTES

Data Center Basics

What a data center is, the four core elements and four types, the basic design requirements, and the power, cooling, redundancy, and maintenance that keep it available

A data center is the physical facility behind every information system — servers, storage, networking, and communications gathered in one place and wrapped in the power, cooling, fire protection, and security that keep them running. This module builds the vocabulary. It defines what a data center is and the four core elements that make one up (core components, the facility, support infrastructure, and operations staff); it lays out the four types organizations choose among (enterprise, colocation, cloud-based, and managed services); it works through the basic design requirements that deliver fault tolerance (support infrastructure, plan and design, location, and architecture); and it closes on the two performance measures every operator lives by — availability (how much of the time the facility is up) and resilience (how fast it recovers and how little data it loses) — along with the redundancy levels and maintenance practices that hold both of them high.

> **LEARNING OBJECTIVES**
>
> - Recognize the role of a data center
> - Identify the core elements of a data center
> - Identify the four types of data centers
> - Identify the basic data center design requirements
> - Recognize two key performance measures commonly tracked by data centers
> - Identify best practices a data center should follow to achieve these performance measures

1.1 What a data center is, and its role

A **data center** is a physical facility that houses servers, storage equipment, network equipment, and various communication equipment used to manage the IT operations of one or more organizations. It **centralizes** an organization's IT operations and equipment in order to store, process, and distribute data, and to provide access to a huge amount of data and applications. In short, it is the physical home of the systems an organization runs on.

> **THE ROLE A DATA CENTER PLAYS**
>
> - Houses one of an organization's most **critical and proprietary assets** — its data.
> - Supports the organization's **business applications**.
> - Provides services to **manage** that data, such as backup and recovery.

Organizations do not all obtain a data center the same way. They can **build and maintain their own**, **share** one with other organizations, use a data center **managed by a third party** or rent servers at a **colocation** facility, or consume **public cloud** services from a provider such as Amazon Web Services (AWS). Those options become the four data center *types* covered later in this module.

1.2 The four core elements

Every data center is built from four core elements. Two are physical — the equipment and the building that holds it — one is the web of systems that keeps the building alive, and one is the team that runs it.

TABLE 1.1 The four core elements

ELEMENT	WHAT IT COVERS
Core components	The hardware- and software-related equipment that stores the data — servers, routers, racks, network and firewall infrastructure, and storage. Because the data is sensitive, everything is kept under highly restricted access controls.
Facility	The building that services and optimizes space to hold the equipment in a temperature-controlled environment, backed by good infrastructure, continuous large amounts of power, and fail-safe backup for uninterrupted service.
Support infrastructure	Redundant power (generators, UPSs, other sources), physical security (CCTV, access and authorization controls), and environmental controls (air handling units and automated cooling).
Operations staff	Trained, qualified personnel who oversee and monitor operations, keep the infrastructure and facility maintained, and troubleshoot faults to minimize downtime to as close to zero as possible.

The four elements answer four different questions: core components are **what stores the data**, the facility is **where it lives**, support infrastructure is **what keeps it powered, cooled, and secure**, and operations staff are **who keeps all of that working**.

1.3 The four types of data centers

Data centers are **classified** two ways: by the **technology** used for compute power and storage, and by **ownership** — whether one organization or many share the facility. Those distinctions produce four main types: **enterprise**, **colocation**, **cloud-based**, and **managed services**.

TABLE 1.2 The four types compared

TYPE	WHAT IT IS	WHO MAINTAINS THE FACILITY
Enterprise	Built by an organization for its own application deployment, often on its office campus; the org selects the site, customizes the design, and manages daily operations	The organization itself
Colocation	The organization rents part of the physical space or a few racks and manages its own servers and storage in that space	Provider maintains the building and utilities
Cloud-based	Pay-as-you-go infrastructure from a CSP (AWS, Google Cloud, Microsoft Azure, IBM Cloud), reachable globally; you pay only for what you use	The cloud service provider (CSP)
Managed services	Also called outsourced — the org leases physical access and outsources data center operations to a third party, with dedicated infrastructure for its apps	A third-party vendor

TABLE 1.3 Advantages and trade-offs by type

TYPE	ADVANTAGES	WATCH-OUTS
Enterprise	Dedicated infrastructure available anytime; full control over the data; easy on-campus access for troubleshooting	Must size for peak load; extra resources for power and cooling; consumes office space; higher cost for space
Colocation	Cost-effective (tenants share space and utilities); quickly scalable; strong reliability from redundant utilities, security, and ISPs	Access can be impeded by the facility's own security rules and procedures
Cloud-based	Servers deploy and scale up or down in minutes; global reach places servers near clients; the CSP handles infrastructure maintenance	Depends on internet connectivity; data-residency rules can create legal issues; recurring monthly cost can exceed a one-time build
Managed services	No need to allocate in-house resources to maintain it; the provider handles infrastructure maintenance	Still must size for peak load and supply extra resources for continuous power and cooling

COMMON PITFALL

Colocation and cloud-based are not the same as “hands-off.” In **colocation** you still manage your own servers and storage — only the building and utilities are the provider's job. In a **cloud-based** data center you still own application security, the operating system, firewalls, accounts, and the network, even though the CSP owns the hardware.

1.4 Designing for fault tolerance: critical support infrastructure

A data center's defining requirement is that it be **available 24/7 —** hardware, software, and power all functioning continuously, even through disruptions such as power outages, server crashes, or surges in users and data. **Fault tolerance** is the ability of a system to keep working uninterrupted, without breaking down, even when a **major component fails**: if a primary server fails, a backup server takes over; if power fails, backup power components carry the load. Fault tolerance is not limited to systems **—** it applies to **human resources** as well.

Four basic design requirements deliver that fault tolerance: **support infrastructure**, **plan and design**, **location**, and **architecture**. Support infrastructure is the largest of the four, built from five critical systems.

FIVE CRITICAL SUPPORT SYSTEMS

- **Power source** — distributes constant, uninterrupted, surge-safe power using UPSs, diesel generators, PDUs, and integrated data center units (IDUs).
- **HVAC** — a localized air-cooling system that controls temperature, humidity, and airflow and removes air pollutants by cooling or heating the space.
- **Fire protection** — alarm and notification systems that detect fire early and help prevent it from spreading.
- **Security** — systems that monitor, authorize, and restrict access to the data center.
- **Raised floor** — provides space for cold airflow and secures power and data cables.

Power distribution — the electrical backbone. An efficient power plant and distribution system are essential because servers consume power in proportion to the work they process. The backbone elements are the **maintenance and emergency shutdown switches** at all site entry points, a **standard grounding system**, a **signal reference grid (SRG)**

to reduce high-frequency impedance, **higher-gauge wire** for scalable electrical expansion, and **PDUs** that integrate circuit breakers along with the equipment connections.

Power supply — UPS and generators. Multiple power sources keep every system fed: redundant feeds from the public utility, a **UPS** with backup battery capacity, and **generators** for longer outages. A UPS should be sized for **peak load** to supply all electrical equipment for at least **5–20 minutes** during an outage, and it runs continuously to filter and condition the power. When an outage runs past the UPS run-time limit, a generator takes over — so generator maintenance, fuel contracts, and scheduled tests all have to be planned during the design phase.

TABLE 1.4 Air-conditioning targets

CONDITION	TARGET
Ambient temperature	64–81°F (18–27°C)
Relative humidity	45–60%
Static pressure under the raised floor	5% greater than room pressure
Design approach	Ensure thorough airflow across racks with no leakage; use spot cooling; install multiple HVAC units for redundancy; avoid centralized systems

FIRE DETECTION AND SUPPRESSION

- **Detection** — smoke alarms and heat detectors catch fire at an early stage.
- **Suppression** — automated firefighting systems deploy fire retardants and sprinkler systems to control and put out the fire.

TABLE 1.5 Raised-floor height by data center area

RAISED-FLOOR HEIGHT	DATA CENTER AREA
12 inches	1,000 square feet or less
18 inches	Between 1,000 and 5,000 square feet
24 inches	Greater than 5,000 square feet

Perforated tiles in the raised floor facilitate airflow and keep equipment cool, and a raised floor is more cost-competitive than vinyl-tile installations — which is why sizing the floor height to the room area is part of the basic design.

1.5 Plan and design, location, and architecture

The **plan and design** of a data center must be carried out by **qualified resources**. The design has to reserve **spare space** for core infrastructure — electrical systems, cable trenches, cooling equipment, piping, and ducts — plan for the **expansion** of raised floors and spare areas, and maximize the space available for core components such as servers, storage, and network equipment. Throughout, the goal is reliable infrastructure with **redundancy** that delivers maximum availability.

CHOOSING A LOCATION

- Lower-cost real estate market and adequate labor supply
- Feasibility of reliable power and communication infrastructure
- Proximity to the parent organization, customers, employees, and support vendors
- Highway access and adequate parking for vehicles
- Water and fuel storage on site
- Minimal risk of severe weather and natural disasters, plus public safety
- Avoid hazardous areas where fire or machine vibration is likely — cafeterias, machine shops, and wet labs

ARCHITECTURAL PARAMETERS

- A **single story** with large floor plates, high ceilings, and few windows
- **Redundant** (more than one) power and communication services
- Highly secure areas throughout the building

1.6 Availability and redundancy

Information systems depend on data centers, and cloud computing has only deepened that dependency — customers expect their data to be retrievable at any time, regardless of disruptions. **Availability** captures that: a data center is available when it is **operational and its data is reachable**. Disruptions come from power failures, equipment failures, network outages, and natural disasters. As a performance metric, availability is the amount of time the data center is **not in failure mode and not under maintenance**.

MEASURING AVAILABILITY

Availability is calculated from two quantities. **Uptime** is the total time the data center is operational; **downtime** is the total time it was unavailable for any reason, scheduled or unscheduled. Availability is then **uptime ÷ (total uptime + total downtime)**. Cutting downtime — planned or unplanned — is what moves the number upward.

Redundancy is a design in which systems or components are **duplicated** so that when one fails, the backup keeps things running. A data center with more redundant power avoids costly downtime — an emergency diesel generator, for instance, can maintain server availability while main power is restored, and the more backup power available, the stronger the redundancy. Because **power and cooling** are the **primary infrastructure elements**, they have the biggest impact on availability: without redundant cooling a single failed air-conditioning unit can overheat the servers, and without redundant power there is no way to maintain the IT load during power-system maintenance.

TABLE 1.6 Degrees of redundancy

LEVEL	MEANING	10-UNIT EXAMPLE
N	Exactly the infrastructure needed to run at full IT load, with no spare	10 generators
N+1	The N components plus one independent shared backup, covering a single failure or maintenance	11 generators
2N	A fully independent duplicate for every component — maintenance can take the entire primary set offline at once	20 generators
2N+1	A full duplicate set plus one more flexible spare that can back up any pair	21 generators

In this notation, **N** represents the infrastructure needed to operate a facility at full IT load. It is typically applied to cooling units, diesel generators, and UPSs, but the same idea extends to other parts of the data center infrastructure.

> COMMON PITFALL
>
> Don't confuse **2N** with **2N+1**. 2N is a complete second set — one backup per component. 2N+1 keeps that full duplicate set **and** adds a single flexible spare that can stand in for any pair, buying still more protection during overlapping failures or maintenance.

1.7 Maintenance and resilience

The right level of redundancy is only part of stable availability — **proper maintenance** matters just as much. Data centers rely on three types.

TABLE 1.7 Three types of maintenance

TYPE	WHAT IT IS	GOAL
Preventive	Planned maintenance actions — equipment checks, partial or complete overhauls, oil changes, lubrication — that replace worn components before they fail	Prevent equipment failure before it occurs
Predictive	Periodic (offline) or continuous (online) condition monitoring of equipment	Maintain at the most cost-effective moment, just before breakdown
Scheduled	Repair and upkeep performed within a set time frame, detailing what is done and by whom, often at repeating intervals	Keep recurring tasks on a fixed calendar (e.g., filters twice a year)

Resilience is a data center's ability to **recover quickly and continue operating** after a failure such as a power outage or equipment failure. It is measured with two parameters.

KEY TERMS

RTO (recovery time objective)
The defined duration within which the data center must be restored after a failure to minimize business loss. A lower RTO means faster recovery — which requires redundant equipment for business-critical systems, balanced against the setup and maintenance cost of that redundancy.

RPO (recovery point objective)
The maximum targeted or acceptable period in which data might be lost from hosted data center applications.

> **COMMON PITFALL**
>
> A lower RTO is not automatically the right target. Because faster recovery demands more redundant equipment, RTO has to be **balanced against cost** — you identify the **business-critical** equipment and plan redundancy there, rather than duplicating everything in the facility.

1.8 Best practices for availability and resilience

Five best practices keep availability and resilience high. They build on one another — find the risks, protect against them, watch for trouble, keep the gear healthy, and prove the plan works.

TABLE 1.8 Five best practices

PRACTICE	WHAT IT INVOLVES
Evaluate equipment and assess risks	Rank equipment by the applications it runs and the data it stores to find high-risk items, then assess each one's risk of failure by probability of occurrence × impact of occurrence.
Plan redundancy	Plan redundancy for high-risk items only — low- and medium-risk items are left out to avoid needless cost — and revisit the plan at regular intervals, increasing or decreasing redundancy as needed.
Create systems infrastructure and processes	Stand up a data center information management system to monitor all high-risk items, set proper alarms for their failures, and define preventive and predictive maintenance processes.
Perform proper maintenance	Run preventive and predictive maintenance so equipment stays up, and staff the facility with a consistent team of trained, qualified personnel plus a training program for new hires.
Conduct simulated tests	Run periodic simulated load tests to confirm capacity, and test the resilience and availability plans regularly to find and close gaps.

> **COMMON PITFALL**
>
> Redundancy is not a blanket policy. You **always** plan redundancy for **high-risk** items, but extending it to low- and medium-risk equipment just raises infrastructure cost and overloads operations for little benefit.

Module summary

- A data center is the physical facility that houses servers, storage, network, and communication equipment and centralizes an organization's IT operations to store, process, and distribute its most critical asset — data.
- Four core elements make up any data center: core components (the equipment), the facility (temperature-controlled space), support infrastructure (power, cooling, security), and operations staff.
- The four types are enterprise (in-house), colocation (rent space, manage your own servers), cloud-based (pay-as-you-go, CSP-owned), and managed services (outsourced) — classified by technology and ownership.
- In colocation the provider maintains only the building and utilities; in a cloud-based data center the customer still owns application security, the OS, firewalls, accounts, and the network.
- Fault tolerance keeps a data center available 24/7 through four design requirements: support infrastructure, plan and design, location, and architecture.
- Critical support infrastructure is five systems — power source, HVAC, fire protection, security, and raised floor — with a UPS sized for 5–20 minutes and generators for longer outages, and air kept at 64–81°F and 45–60% humidity.
- Availability is uptime ÷ (uptime + downtime); power and cooling have the biggest impact, and redundancy levels run N+1, 2N, and 2N+1.
- Resilience is fast recovery, measured by RTO (time to restore) and RPO (acceptable data loss), and supported by preventive, predictive, and scheduled maintenance.
- Best practices: assess risks, plan redundancy for high-risk items only, build monitoring and alarms, maintain properly with trained staff, and run simulated tests regularly.

Review questions

1. **What is a data center, and what is its primary role?**

   Answer: A physical facility that houses servers, storage, network, and communication equipment and centralizes an organization's IT operations to store, process, and distribute data. Its role is to keep the organization's most critical asset — its data — available, support business applications, and provide data-management services such as backup and recovery.

2. **List the four core elements of a data center and one thing each covers.**

   Answer: Core components (servers, racks, network/firewall, storage under restricted access); facility (temperature-controlled space and supporting infrastructure); support infrastructure (redundant power, physical security, environmental controls); operations staff (monitor, maintain, and troubleshoot to keep downtime near zero).

3. **How does colocation differ from a managed services data center?**

   Answer: In colocation the organization rents space or racks and manages its own servers and storage while the provider maintains the building and utilities. A managed services (outsourced) data center hands the operation of the data center itself to a third party, which handles infrastructure maintenance.

4. **Name the five critical support systems and the four basic design requirements.**

   Answer: Five support systems: power source, HVAC, fire protection, security, and raised floor. Four design requirements: support infrastructure, plan and design, location, and architecture.

5. **How long should a UPS carry the load, and what takes over after that?**

   Answer: A UPS is sized for peak load to supply power for about 5–20 minutes; when an outage exceeds the UPS run-time limit, a generator takes over for the longer term.

6. **How is availability calculated, and which infrastructure elements affect it most?**

   Answer: Availability = uptime ÷ (total uptime + total downtime). Power and cooling are the primary infrastructure elements and have the biggest impact on availability.

7. **Explain N+1, 2N, and 2N+1 using a 10-generator load.**

   Answer: N+1 adds one shared backup, so 11 generators. 2N provides a full duplicate set, so 20 generators, allowing maintenance on the whole primary set at once. 2N+1 keeps the full duplicate set and adds one flexible shared spare, so 21 generators.

8. **What do RTO and RPO measure, and which equipment should you always make redundant?**

   Answer: RTO is the time within which the data center must be restored after a failure; RPO is the maximum acceptable amount of data that might be lost. Always plan redundancy for high-risk items — not low- or medium-risk equipment, which would only raise cost.