MODULE 13 · STUDY NOTES

Data Center Availability

How a data center defines availability, expresses it as the nines and in availability formulas, separates it from reliability and maintainability, and defends it against the threats that cause downtime

A data center earns its keep only when it is *available* — ready to perform every function it is supposed to perform, whenever it is needed. This module builds the vocabulary and the math behind that idea. It starts with operational availability and the levels that describe how much yearly time a facility can spare for maintenance, then shows how service-level agreements turn availability into a contract measured in the 9s — where five nines (99.999%) is the near-perfect target and every hour of downtime carries a steep price. From there it separates three terms that are constantly confused: availability (the percentage of time equipment is operational), reliability (whether equipment does its job), and maintainability (how fast it can be repaired) — and shows that availability is a function of the other two, expressed through MTBF and MTTR. It closes with the threats that erode availability, from power failures and human error to natural disasters and a family of cyberattacks, along with the defenses that hold them off.

LEARNING OBJECTIVES

- Define availability of a data center
- Identify key terms associated with availability
- Differentiate between availability and reliability
- Recognize the relationship between availability, reliability, and maintainability
- Recognize key elements of availability and reliability calculations
- Recognize threats to data center availability

13.1 Operational availability and its levels

Operational availability means a data center is able to perform all of the functions it is supposed to perform at one time. It is decided by the **total data center uptime** the facility should be available to support without any interference. A data center can be paused for a *preplanned* maintenance process without disturbing ongoing performance, based on the availability throughout the year, or the level of operation. The idea is to keep the data center running smoothly and to identify the correct time for maintenance — and the main objective is to figure out the **maximum yearly permissible downtime** and the related operational level the data center must cater to. The broader term, **availability**, refers simply to the readiness of a data center to be accessible for use.

TABLE 13.1 Levels of operational availability

LEVEL	YEARLY MAINTENANCE HOURS	CHARACTER
Level 0	More than 400 hours	Less operational; can pause for maintenance during off or working hours
Level 1	100 to 400 hours	More operational than Level 0, but still has adequate time to pause for maintenance
Level 2	50 to 99 hours	Operational most of the time; maintenance can still be done during off or working hours
Level 3	0 to 49 hours	Highly operational; cannot spare time for preplanned maintenance
Level 4	0 hours	Highly operational; no time at all for prescheduled maintenance

FACTORS FOR CALCULATING OPERATIONAL AVAILABILITY

- **Knowledge of the operational needs** — identify weekly hours of operation, annual hours of production, level of operation, and permissible maintenance-cease time.
- **Grade of operational availability** — determined by the production hours.
- **Level of downtime risk** — classified as remote, minor, major, acute, and fatal.
- **Design decisions of the data center** — the design is finalized to reduce risk and improve reliability.

COMMON PITFALL

The level numbering runs opposite to intuition. A *higher* level number means a *more* operational data center with *less* maintenance time, not more. Level 0 sits idle enough for 400+ maintenance hours a year; Level 4 is so heavily used it can spare zero.

13.2 Service-level agreements and the 9s of availability

Whenever a business hires data center space or a service, the data center provides a **service-level agreement (SLA)**. In the SLA the data center specifies the aggregate time it will stay up and running through the year — there are SLAs for both uptime and downtime. The SLA holds the data center **accountable** for delivering the agreed-upon service, including the promised availability and reliability levels. To mark those required levels, SLAs typically use a key term: the **9s of availability**.

Five nines is common terminology that refers to **99.999 percent**, indicating a near-perfect level of performance. The rule is simple — *more 9s mean better availability and less downtime*. The stakes are high: according to a **Gartner** survey, an estimated loss of about **$300,000 is suffered per hour of downtime**, a cost that has climbed by roughly 30 percent in recent years. Reducing downtime and increasing availability is therefore a top priority.

TABLE 13.2 The 9s of availability and annual SLA downtime

UPTIME	NUMBER OF 9S	SLA DOWNTIME ANNUALLY
99.9%	Three	525.6 minutes
99.99%	Four	52.56 minutes
99.999%	Five	5.256 minutes
99.9999%	Six	0.525 minute
99.99999%	Seven	0.0525 minute

COMMON PITFALL

Read the table in the right direction: adding a nine cuts the *allowed downtime* by about a factor of ten. Three nines still permits 525.6 minutes (roughly 8.76 hours) a year, while five nines permits only 5.256 minutes. Do not mistake more 9s for more permitted downtime.

13.3 Uptime, downtime, MTBF, and MTTR

Business domains constantly seek services that add to their performance, and a key data center aspect that ensures performance is its availability, or **uptime**, throughout the year. **Uptime** is the time when a data center is up and running, aiming for **99.999 percent or higher** availability across a year. It is determined annually by the *uptime reliability measurement*, which is calculated as a percentage and stays near perfection.

Downtime is the time when a data center is going through some sort of failure or is not available. It falls into two categories. **Scheduled downtime** is preplanned downtime for maintenance, upgrades, or internal testing. **Unscheduled downtime** cannot be predetermined and can occur due to failures or delays in maintenance.

KEY TERMS

Mean time between failures (MTBF)
The average time between the start and end of failures caused by underlying mechanical or electronic conditions — the uptime that falls between consecutive failures of repairable equipment. Calculated as the difference between the beginning of downtime and the beginning of uptime, divided by the total number of failures.

Mean time to repair (MTTR)
The average time a device requires to recover from a failure. Calculated as the total time for maintenance divided by the total number of repairs.

WHY MTBF AND MTTR MATTER

- They are the two parameters used to calculate both availability and reliability of a data center facility.
- Both are provided by the component manufacturer, obtained from historical data, and denoted in **hours**.
- The measurement **excludes human errors** and human factors associated with failures.

13.4 Availability formulas

Optimal data center availability is calculated using standard formulas, and the right formula depends on the scenario being assessed. Three scenarios recur in the module, each using uptime (**UT**), downtime (**DT**), MTBF, and MTTR.

TABLE 13.3 Availability formulas by scenario

SCENARIO	FORMULA	VARIABLES
A hardware manufacturer checks a server's availability	MTBF ÷ (MTBF + MTTR)	MTBF, MTTR
Supporting prescheduled system downtime for maintenance	(UT + prescheduled maintenance) ÷ (DT + UT + prescheduled maintenance)	UT, DT, maintenance
Assessing around-the-clock (24/7) business availability	UT ÷ (UT + DT)	UT, DT

WORKED EXAMPLE – APPLYING THE SERVER FORMULA

The server-availability formula is **MTBF ÷ (MTBF + MTTR)**. Suppose a manufacturer's historical data reports an MTBF of 1,000 hours and an MTTR of 1 hour. Then availability = 1,000 ÷ (1,000 + 1) = 1,000 ÷ 1,001 ≈ **0.999**, or about **99.9%** — three nines. The lesson the arithmetic teaches: raising availability means either increasing MTBF (fewer failures) or decreasing MTTR (faster repairs). The numbers here are illustrative; the formula and its inputs come straight from the module.

The reliability and availability of an element can also be captured in equations. **Availability = MTBF ÷ (MTBF + MTTR)** and **Reliability = e ^ (−t ÷ MTBF)**, where *e* is the exponential function (the inverse of a logarithm), *t* is the time duration in hours, and MTBF and MTTR are both in hours. Availability behaves as a *functional* parameter — if equipment is available 85 percent of the time, it is said to be operating at 85 percent of its technical limit, which typically maps to its financial performance.

13.5 Availability versus reliability

The terms **reliability** and **availability** are frequently confused in the context of a data center. The measure of the *percentage of time* equipment is in an operational state is called **availability**. The measure of the *time taken* for equipment to achieve its intended functionality is called **reliability**. These definitions can be refined further.

AVAILABILITY – RELATED TO FAULT-TOLERANT SYSTEMS	RELIABILITY – RELATED TO DURABILITY AND QUALITY
stays up even when parts fail ▪ The system stays operational even during component failures ▪ Provides a more accurate and meaningful measure of performance than reliability alone ▪ Example: a UPS breaks down from an unreliable inverter, but the overall system does not halt because an **alternate UPS takes over** ▪ That backup component **increases the system's availability**	*does the component do its job?* ▪ Probability that a structure or component does its job under given operational states and ecological conditions over a period ▪ Expressed as a percentage of **success ÷ failure** of the component over that period ▪ Example: a UPS with a common inverter failure has **low reliability** ▪ It does not perform its intended purpose

The UPS pair captures the distinction precisely: the *same* unreliable inverter drives *low reliability*, yet the *presence of a backup* keeps *availability high*. That is why, for critical data center systems, availability is treated as the more useful figure — it reflects what the system delivers, not just how sturdy any single part is.

13.6 Availability, reliability, and maintainability

Three relationships must be considered together in a data center: availability, reliability, and maintainability. **Availability** is the most critical measurement — it gives the most accurate and meaningful measurement of a data center's performance. **Reliability** of equipment directly affects its availability. And **maintainability** is a third factor that must be considered when measuring the other two.

Maintainability refers to the ease and speed of performing repair activities on equipment to bring it back up and running after a failure occurs. **High maintainability** indicates high speed and ease of repair, resulting in the minimum time needed to get equipment running again. **Low maintainability** indicates complex repair work that requires more time to complete.

TABLE 13.4 How the three terms move together

AVAILABILITY	RELIABILITY	MAINTAINABILITY
Increase	Constant	Increase
Decrease	Constant	Decrease
Increase	Increase	Constant
Decrease	Decrease	Constant

Reading the table: when reliability is held constant, an increase in maintainability increases availability, and a decrease in maintainability decreases it. When maintainability is held constant, an increase in reliability increases availability, and a decrease in reliability decreases it. In short, **availability is a function of reliability and maintainability** — which is exactly why it provides a more accurate measure than reliability alone, especially for critical data center systems.

> **COMMON PITFALL**
>
> Two statements from the module are the correct ones to remember: *reliability directly affects availability*, and *availability is a function of reliability and maintainability*. A decrease in availability comes from a **decrease** (not an increase) in reliability and maintainability, and if reliability increases while maintainability stays constant, availability **increases**.

13.7 Calculating availability and reliability

The two parameters that drive both calculations are **MTBF** and **MTTR**. As the name suggests, MTBF is the predicted duration between two consecutive failures of a component during normal operation, calculated by dividing the **total service duration** of the component by the **total number of failures**. MTTR is the average time required to react to and restore the failed components. Both parameters come from the component manufacturer, are drawn from historical data, and are denoted in hours; the measurement excludes human errors and human factors.

TABLE 13.5 Reliability and availability equations

QUANTITY	EQUATION	TERMS
Availability	MTBF ÷ (MTBF + MTTR)	MTBF and MTTR in hours
Reliability	e ^ (−t ÷ MTBF)	e = exponential function (inverse of log); t = time in hours

The **five 9s** standard restates availability in everyday units. Five 9s denotes the preferred availability of a service or equipment: it is common terminology for **99.999 percent** and is considered optimal because it results in the lowest downtime per day. The higher the uptime percentage, the lower the downtime — five nines is the desired percentage of availability for a given system or service.

TABLE 13.6 Five 9s – uptime and downtime per day

UPTIME	DOWNTIME PER DAY
90%	2.4 hours
99%	14 minutes
99.9%	86 seconds
99.99%	8.6 seconds
99.999%	0.86 second

COMMON PITFALL

The module shows two different downtime tables. The **9s of availability** table lists SLA downtime *per year* (five nines = 5.256 minutes/year); the **five 9s** table lists downtime *per day* (five nines = 0.86 second/day). Match the figure to the time frame the question asks about.

13.8 Threats to data center availability

Data center availability is crucial for business continuity, but many factors can put it at risk. These threats not only interrupt the continuous availability of the system, they can cause long-term damage to the organization. The module groups the common threats and then examines each.

COMMON THREATS TO DATA CENTERS

- **Power failure**
- **Human error**
- **Data center-level failure** (natural disasters)
- **DDoS attacks** on Domain Name System (DNS) servers
- **DDoS attacks** on web applications
- **DDoS attacks** on Secure Sockets Layer (SSL) traffic
- **Brute force and fragile authentication**

Power failure. A steady, reliable power source is essential across all computing systems. Although data centers are equipped with power generators, unforeseen events still cause power failure: the fuel in the backup generators may be exhausted, the batteries may be nonfunctional, or the backup power supply itself may break down.

> **HUMAN ERROR – COMMON FORMS**
>
> - Accidental activation of the **safety shutdown button**, shutting down all equipment.
> - Accidental **spills of liquids** on sensitive electronic equipment, causing damage.
> - **Improper configuration of failover clusters** of servers, leading to imbalanced workload allocation and downtime.
> - **Failure to conduct periodic testing and replacement** of equipment, causing malfunction, lower performance, or complete failure.

Data center-level failure. This does not happen often, but when it does it leads to significant complications for operational availability. To keep critical workloads running, hardware redundancy is ensured in all data centers — yet natural disasters such as **earthquakes, hurricanes, and floods** can still cause data center failure and make critical systems fail or stop functioning.

13.9 Cyberattacks and defenses

DDoS attacks on DNS servers. Distributed denial of service threats are often aimed at prime servers, disturbing and deactivating essential internet services. Attackers manipulate defenseless web applications to hack web servers, DNS servers, and SSL traffic, then use those controlled servers to attack other vulnerable websites. DNS servers are highly susceptible: disconnecting them keeps large numbers of internet subscribers from internet access, and **reflection attacks** imposed on DNS servers can cause significant downtime and even power outages in the data center facility.

DDoS attacks on web applications. Web application attacks happen through web servers and turn the compromised servers into sources for further DDoS attacks. In **SQL (structured query language) injection**, the attacker injects code to retrieve data stored in a database. In **cross-site scripting (XSS)**, the attacker adds malicious code to a website that users are accessing to retrieve their personal information. In **cross-site request forgery (CSRF)**, the attacker induces users to perform actions they do not intend, such as changing passwords.

DDoS attacks on SSL traffic. Applications use SSL to establish an encrypted link between the server and users, and attacks increasingly focus on encryption to dodge recognition. As more applications rely on SSL, more malicious activity can hide inside encrypted traffic, causing availability issues. Although many firewall, intrusion-deterrence, and threat-avoidance products can decrypt SSL traffic, keeping up with emerging SSL encryption demands is hard — so organizations need efficient means to intercept and decrypt SSL traffic while minimizing the heavy SSL handling load on their security equipment and servers.

BRUTE FORCE AND FRAGILE AUTHENTICATION – THREATS	SOLUTIONS
single-factor passwords are the weak point ▪ **Password predicting** (guessing) ▪ **Stolen identities** ▪ **Computerized brute force attacks** from password-cracking tools ▪ Reusing one password across accounts risks them all if one is compromised	*raise the cost of an attack* ▪ Link passwords with **out-of-band verification**, such as SMS notifications to a mobile or other device ▪ Use **two-factor authentication** to sharply reduce password compromise ▪ Handle authentication services **centrally** to avoid extra cost and effort ▪ Adopt an integrated authentication solution across applications

Application delivery controllers (ADCs). To reduce a multitude of threat factors, organizations deploy ADCs — a network component typically placed in the data center **between the firewall and one or more application servers**. An ADC can block threats, cut off and scrutinize encrypted data movement, and prevent unauthorized admission to data center applications; next-generation ADCs add further resistances against emergent threats. Organizations must assess an ADC's safety features carefully and make sure it protects servers and applications **without hindering performance**.

COMMON PITFALL

Match the attack to how it interrupts availability. **Web application attacks** breach applications to steal confidential data; **DNS server attacks** deny subscribers access to domain names and internet services; **SSL attacks** focus on encryption to dodge recognition. On the knowledge check, web application attacks are the ones that interrupt availability *by breaching applications to steal confidential data*.

Module summary

- Operational availability means a data center performs all expected functions; it is set by total uptime and by the maximum yearly permissible downtime, and levels 0–4 describe how many yearly maintenance hours a facility can spare (Level 0 = 400+ hours/least operational; Level 4 = 0 hours/most operational).
- Availability is the readiness of a data center to be accessible for use; the factors for calculating operational availability are knowledge of operational needs, grade of operational availability, level of downtime risk (remote, minor, major, acute, fatal), and design decisions.
- A service-level agreement (SLA) contractually fixes availability and reliability levels and is usually stated in the 9s of availability — more 9s = better availability, less downtime; five nines = 99.999% ≈ 5.256 minutes downtime per year; Gartner cites about $300,000 lost per hour of downtime.
- Uptime is time the data center is running (target 99.999%+ across a year); downtime is scheduled (planned maintenance, upgrades, testing) or unscheduled (unpredictable failures or delays).
- MTBF (mean time between failures = total service duration ÷ number of failures) and MTTR (mean time to repair = total maintenance time ÷ number of repairs) come from the manufacturer's historical data, are in hours, and exclude human factors.
- Availability formulas: server = MTBF ÷ (MTBF + MTTR); prescheduled maintenance = (UT + prescheduled) ÷ (DT + UT + prescheduled); around-the-clock = UT ÷ (UT + DT); reliability of an element = e ^ (−t ÷ MTBF).

- Availability (percentage of time operational, tied to fault tolerance) is not reliability (durability/quality, the probability a component does its job); availability is the more accurate and meaningful measure, illustrated by a backup UPS keeping a system available despite an unreliable inverter.
- Availability is a function of reliability and maintainability (ease and speed of repair): holding one constant, raising the other raises availability; reliability directly affects availability.
- Threats to availability include power failure, human error, data center-level failure (natural disasters), and DDoS attacks on DNS servers, web applications (SQL injection, XSS, CSRF), and SSL traffic, plus brute force / fragile authentication — countered with out-of-band and centralized authentication and application delivery controllers (ADCs).

Review questions

1. **Define operational availability, and state its main objective.**

   Answer: Operational availability means a data center can perform all of its expected functions; it is set by the total uptime the facility must support without interference. Its main objective is to figure out the maximum yearly permissible downtime and the operational level the data center must meet, so maintenance can be scheduled without disturbing performance.

2. **What is an SLA, and how does it express availability requirements?**

   Answer: A service-level agreement is the contract provided when a business hires data center space or service; it fixes agreed availability and reliability levels and holds the data center accountable. It usually states the requirement in the 9s of availability — the number of 9s in the uptime percentage.

3. **Convert five nines to a percentage and to annual downtime, and explain why more 9s is better.**

   Answer: Five nines = 99.999% uptime ≈ 5.256 minutes of downtime per year. More 9s means a higher uptime percentage and therefore less allowed downtime — better availability.

4. **Distinguish availability, reliability, and maintainability, and give the relationship among them.**

   Answer: Availability is the percentage of time equipment is operational (tied to fault tolerance); reliability is whether a component performs its intended function (durability and quality); maintainability is the ease and speed of repair after a failure. Availability is a function of reliability and maintainability — holding one constant, increasing the other increases availability.

5. **A server has MTBF = 1,000 hours and MTTR = 1 hour. What is its availability, and which two levers change it?**

   Answer: Availability = MTBF ÷ (MTBF + MTTR) = 1,000 ÷ 1,001 ≈ 99.9%. Increase availability by raising MTBF (fewer failures) or lowering MTTR (faster repairs).

6. **Name the categories of downtime and give an example of each.**

   Answer: Scheduled downtime is preplanned — for maintenance, upgrades, or internal testing. Unscheduled downtime cannot be predetermined and occurs from failures or delays in maintenance.

7. **List the common threats to data center availability, and name two defenses against weak authentication.**

   Answer: Power failure, human error, data center-level failure (natural disasters), DDoS attacks on DNS servers, web applications, and SSL traffic, and brute force / fragile authentication. Against weak authentication: link passwords with out-of-band verification such as SMS (two-factor authentication) and handle authentication centrally; application delivery controllers (ADCs) provide additional protection.