WHAT IS HIGH AVAILABILITY? CONCEPTS & BEST PRACTICES

Before the internet, ecommerce, and online services, the concept of availability was restricted to the business hours of the brick and mortar shops. The services were available only as long as the lights were on and the doors open.

In the digital age, however, the consumer perspective changed. Users expect Internet services to be up and running 24/7. Online shopping stores are expected to sell products regardless of time zone, business hours, and holidays—the last is even the source of the largest revenue streams globally. Social media outlets keep users engaged because their friends and connections are online and available for communication any time of the day.

So, what does high availability mean for your business? What is realistic and how do you determine that? We'll explain it all in this article.

What does high availability mean for IT?

Internet services run on complex IT systems and environments. Large data center networks power consisting of hundreds of thousands of hardware components. Software systems must all function correctly to deliver the necessary IT functionality at all times.

Despite multiple layers, protections, and redundancy, systems and components fail. This makes 100% availability virtually impossible in any IT environment, large or small. So, what number can be realistically considered as a highly available service?

Before answering this question, let's define availability:

> *Availability refers to the percentage of time that the IT service and its underlying systems remain operational under normal circumstances to deliver on the expected purpose.*

The mathematical formula for calculating *Availability* is as follows:

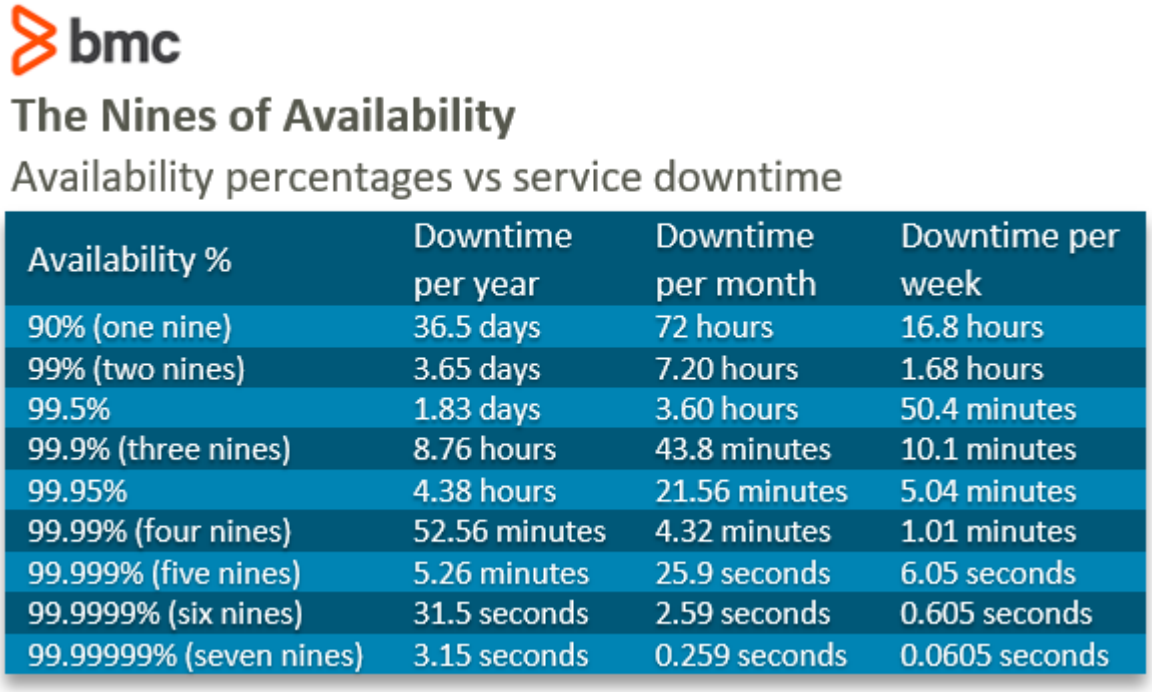
Percentage of availability = (total elapsed time – sum of downtime)/total elapsed time

Availability percentage is calculated over a significant duration where typically at least one downtime incident has occurred. This can be a few hours, days, or even months, especially since IT incidents can occur for a variety of distinct causes. It then gives the duration of downtime that can be expected with a particular percentage of Availability.

Technology vendors offer Service Level Agreements (SLAs) that guarantee the minimum availability levels. In the tech world, these numbers are referred to as '-Nines'. For example:

- 90% availability is a one-nine
- 9% is three-nines
- The most popular of the lot, 99.999%, is the five-nines of availability

This chart describes how each Availability percentage correspond to the yearly, monthly, and weekly downtime:



Availability %	Downtime per year	Downtime per month	Downtime per week
90% (one nine)	36.5 days	72 hours	16.8 hours
99% (two nines)	3.65 days	7.20 hours	1.68 hours
99.5%	1.83 days	3.60 hours	50.4 minutes
99.9% (three nines)	8.76 hours	43.8 minutes	10.1 minutes
99.95%	4.38 hours	21.56 minutes	5.04 minutes
99.99% (four nines)	52.56 minutes	4.32 minutes	1.01 minutes
99.999% (five nines)	5.26 minutes	25.9 seconds	6.05 seconds
99.9999% (six nines)	31.5 seconds	2.59 seconds	0.605 seconds
99.99999% (seven nines)	3.15 seconds	0.259 seconds	0.0605 seconds

How do you achieve high availability?

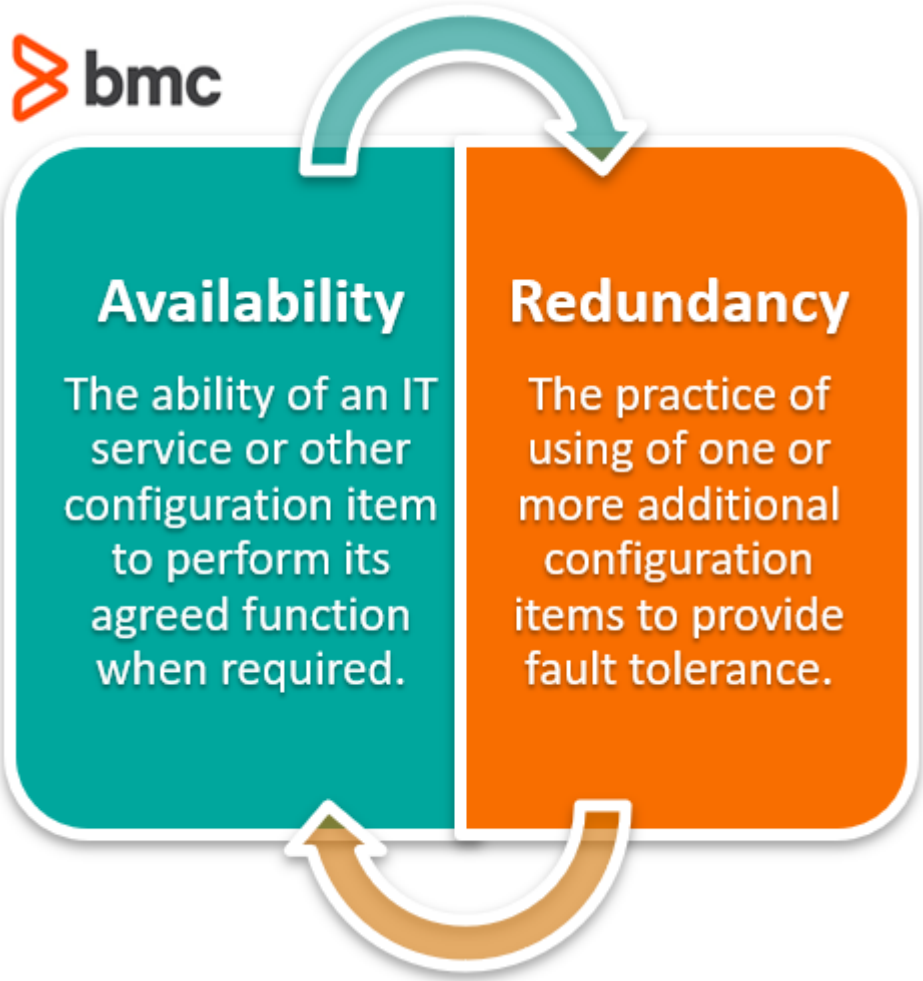
So, what is High Availability for your organization? Determining a specific number requires you to thoroughly analyze your business needs for availability—and the costs required to achieve those goals.

For critical infrastructure, such as hospital emergency rooms or power supply to nuclear power cooling plants, even the six-nines could potentially risk human lives. For such specific use cases, several redundant layers of IT system and utility power infrastructure are deployed to reach High Availability figures close to 100%, such as nine-nines or perhaps, even better.

In simple terms, it all comes down to redundancy. In engineering, redundancy is the act of duplicating components, systems, or functions in order to:

- Increase the overall system reliability and availability.
- Cause a safe system failure or termination.

Duplication, then, refers to the trustworthiness of the system. The redundancy model can follow several design principles, usually described as N-Modular redundancy.



Best practices for high availability

From an enterprise IT perspective, downtime maps directly to lost revenue streams and dissatisfied users. While the accurate calculations vary vastly between organizations, the average may cost $9,000 per minute according to recent research by Ponemon Institute. For organizations as large as Amazon, the cost of downtime is as high as $13.22 million per hour.

So how can you maximize availability of your IT services for lowest financial and business risk? Follow these industry-proven best practices:

- **Assess your business requirements.** Evaluate IT from a business and user perspective.
- **Know the true cost of downtime.** Account for dissatisfied customers and lost user base.
- **Understand your SLAs.** Does the availability correspond to desired metrics?
- **Set Recovery Point Objective (RPO) and Recovery Time Objective (RTO) based on your expected availability percentage.** The system should be capable of recovering to a state with

prior to the longest expected duration of downtime. For example, for five-nines, the RTO should be less than 30 seconds.
- **Prepare a thorough disaster recovery program.** Even with the SLAs in place, vendors only reimburse the IT service cost during the period of downtime. The lost business opportunity and revenue are not calculated.
- **Introduce redundancy strategically.** Mission-critical IT workloads are more in need of redundancy than other operational IT workloads that may not be frequently accessed. The cost of making every workload redundancy might not provide ROI.
- **Understand the metrics.** Make sure you and relevant stakeholders understand the differences between and purposes of reliability, availability, and uptime.

Related reading

- BMC IT Operations Blog
- Availability Management & the Role of the Availability Manager
- MTBF vs. MTTF vs. MTTR: Defining IT Failure
- What is MTTA? Mean Time to Acknowledge Explained
- Error Budgets Explained: Risk & Reliability in One Metric