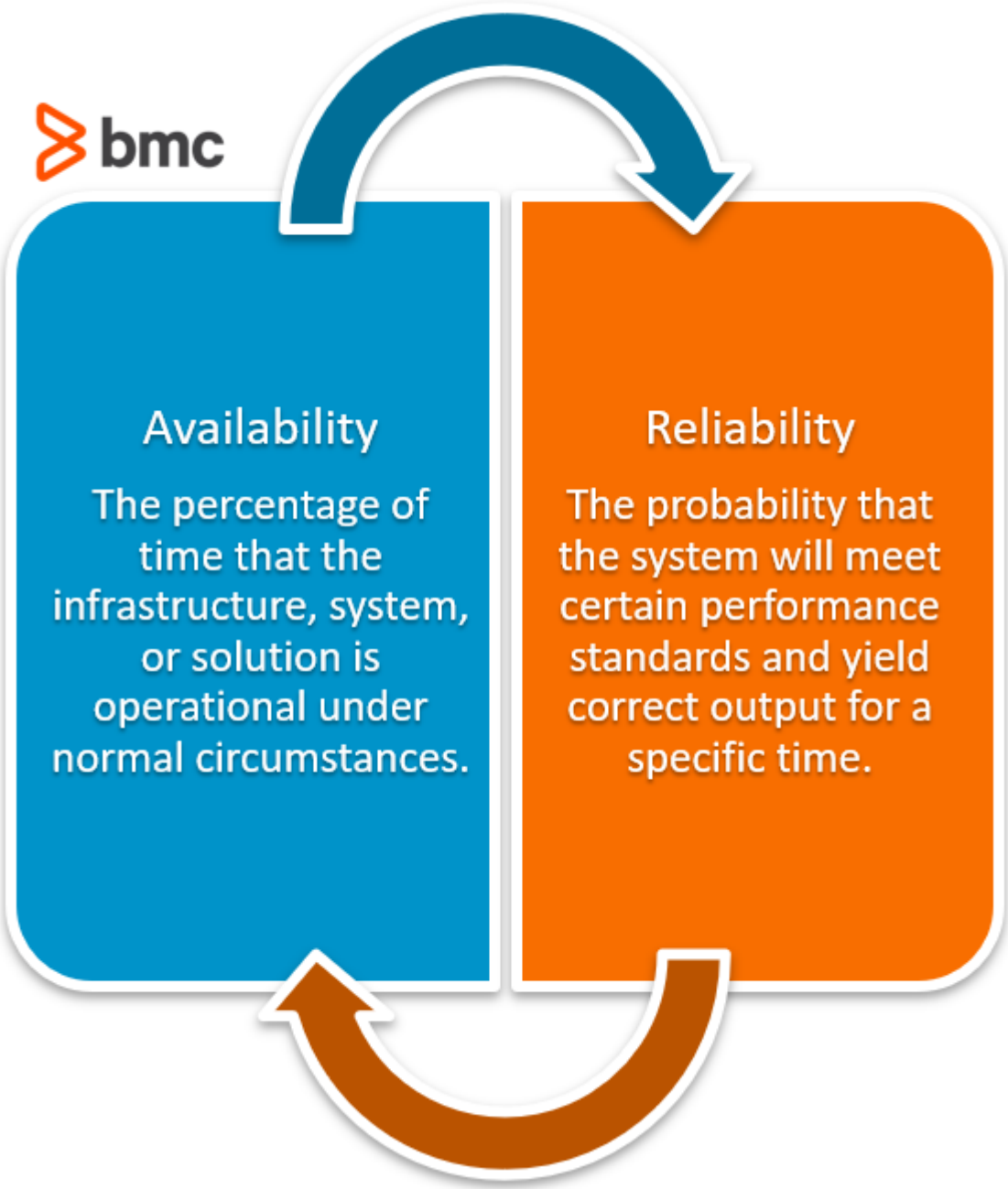
RELIABILITY VS AVAILABILITY: WHAT'S THE DIFFERENCE?

When you pay for a service or invest in the underlying technology infrastructure, you expect the service to be delivered and accessible at all times, ideally. In the real world of enterprise IT however, ideal service levels are virtually impossible to guarantee. For this reason, organizations evaluate the IT service levels necessary to run business operations smoothly, to ensure minimal disruptions in event of IT service outages.

Two meaningful metrics used in this evaluation are **Reliability** and **Availability**. Often mistakenly used interchangeably, both terms have different meanings, serve different purposes, and can incur different cost to maintain desired standards of service levels.



Both reliability and availability serve as key decision factors in your IT strategy. Make sure you understand these concepts before planning and implementing IT infrastructure solutions.

Let's take a look.

What is availability?

Availability refers to the percentage of time that the infrastructure, system, or solution remains operational under normal circumstances in order to serve its intended purpose. For cloud infrastructure solutions, availability relates to the time that the data center is accessible or delivers the intend IT service as a proportion of the duration for which the service is purchased.

The mathematical formula for Availability is :

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

For instance, if an IT service is purchased at a 90 percent service level agreement for its availability,

the yearly service downtime could be as much as 876 hours. For an SLA of 99.999 percent availability (the famous five nines), the yearly service downtime could be as much as 5.256 minutes.

bmc

The 9s of Availability

Availability percentages vs service downtime

Availability %	Downtime per year	Downtime per month	Downtime per week
90% (1 nine)	36.5 days	72 hours	16.8 hours
99% (2 nines)	3.65 days	7.20 hours	1.68 hours
99.5%	1.83 days	3.60 hours	50.4 minutes
99.9% (3 nines)	8.76 hours	43.8 minutes	10.1 minutes
99.95%	4.38 hours	21.56 minutes	5.04 minutes
99.99% (4 nines)	52.56 minutes	4.32 minutes	1.01 minutes
99.999% (5 nines)	5.26 minutes	25.9 seconds	6.05 seconds
99.9999% (6 nines)	31.5 seconds	2.59 seconds	0.605 seconds
99.99999% (7 nines)	3.15 seconds	0.259 seconds	0.0605 seconds

(*Source*)

The numbers portray a precise image of the system availability, allowing organizations to understand exactly how much service uptime they should expect from IT service providers. However, measuring availability is a challenging task.

Organizations aim to measure and track availability of the most impactful functionality of the IT service. In the real world, it may be difficult to understand exactly which metric of the service performance corresponds best to this requirement. For instance:

- An organization may consider service outage to occur only when a certain percentage of users have been affected.
- Another organization may consider service outage to occur when certain server instances are not accessible regardless of the users affected.

Additionally, organizations may want to invest in different SLA agreements for different types of workloads:

- A mission-critical cloud infrastructure service may require 'six 9s' of availability to ensure the core app functionality is always up and running.
- Low-priority workloads may run reasonably well at low SLA performance in terms of service availability.

Merely having a service available isn't sufficient. When an IT service is available, it should actually serve the intended purpose under varying and unexpected conditions. One way to measure this performance is to evaluate the reliability of the service that is available to consume. Organizations depend on different functionality and features of the IT service to perform business operations. As a result, they need to measure how well the service fulfils the necessary business performance needs.

What is reliability?

Reliability refers to the probability that the system will meet certain performance standards in yielding correct output for a desired time duration.

Reliability can be used to understand how well the service will be available in context of different real-world conditions. For instance, a cloud solution may be available with an SLA commitment of 99.999 percent, but vulnerabilities to sophisticated cyber-attacks may cause IT outages beyond the control of the vendor. As a result, the service may be compromised for several days, thereby reducing the effective availability of the IT service.

Similar to Availability, the Reliability of a system is equality challenging to measure. There may be several ways to measure the probability of failure of system components that impact the availability of the system. A common metric is to calculate the Mean Time Between Failures (MTBF).

MTBF = (total elapsed time - sum of downtime)/number of failures

MTBF represents the time duration between a component failure of the system. Similarly, organizations may also evaluate the Mean Time To Repair (MTTR), a metric that represents the time duration to repair a failed system component such that the overall system is available as per the agreed SLA commitment.

Other ways to measure reliability may include metrics such as fault tolerance levels of the system. Greater the fault tolerance of a given system component, lower is the susceptibility of the overall system to be disrupted under changing real-world conditions.

Using availability & reliability

The measurement of Availability is driven by **time loss** whereas the measurement of Reliability is driven by the **frequency** and **impact** of failures. Mathematically, the Availability of a system can be treated as a function of its Reliability. In other words, Reliability can be considered a subset of Availability.

For either metric, organizations need to make decisions on how much time loss and frequency of failures they can bear without disrupting the overall system performance for end-users. Similarly, they need to decide how much they can afford to spend on the service, infrastructure and support to meet certain standards of availability and reliability of the system.

An important consideration in evaluating SLAs is to understand how well it aligns with business goals. The resulting strategy is often a tradeoff between cost and service levels in context of the

business value, impact, and requirements for maintaining a reliable and available service.

With the traditional IT service delivery models, organizations are in full control of the system and have to make extra efforts internally or through external consultants to fix failures or service outages. For cloud-based technology solutions, organizations rely on vendors to meet SLA standards. Vendors are responsible for:

- Infrastructure management
- Troubleshooting and repair
- Security
- Other associated operations that make the service adequately reliable and available

While vendors work to promise and deliver upon SLA commitments, certain real-world circumstances may prevent them from doing so. In that case, vendors typically don't compensate for the business losses, but only reimburses credits for the extra downtime incurred to the customer. Additionally, vendors only promise "commercially reasonable" efforts to meet certain SLA objectives. As such, customers are expected to leverage adequately redundant and failover systems to guarantee availability and reliability of the service in response to disruptions caused by impactful natural disasters such as Hurricane Sandy.

Related reading

- BMC Service Management Blog
- BMC IT Operations Blog
- System Reliability & Availability Calculations
- Impact of Redundancy on Availability
- What Is High Availability? Concepts & Best Practices