belden.com

redundancy-availability-reliability-related-but-not-the-same

Michael Salvador

5-7 minutes

Understanding and determining data center redundancy, availability and reliability is a critical part of the holistic design process, and one that can significantly impact cost. However, there is often confusion surrounding these three factors.

Some believe that redundancy, availability and reliability are one in the same, but that is not the case at all. Let's take a closer look.

Redundancy vs. Availability

Redundancy is an operational requirement of the data center that refers to the duplication of certain components or functions of a system so that if they fail or need to be taken down for maintenance, others can take over.

Redundant components can exist in any data center system, including cabling, servers, switches, fans, power and cooling. It is often based on the "N" approach, where "N" is the base load or number of components needed to function. N+1 means having one more component than is actually needed to function, 2N means having double the amount of total components, and 2N+1 is having double the amount plus one.

For example, consider that you have 3 cars available for 3 people

to take a trip. In this scenario, N = 3. However, in case one car breaks down, you might want to go with N+1, or 4 cars. This essentially means that you'll always have one extra car if one should break down.

If you were to plan for 2N redundancy, you would need 6 cars. In this case, if all 3 cars break down, everyone would still have an available car take the trip. If you relate that to power, a 2N system is considered fully redundant because the entire system could fail or be taken offline for an extended period of time and another can fully take over.

In the car scenario, 2N+1 would involve having 7 cars available —which is often overkill for almost any scenario or data center. Furthermore, hyper data centers and new technologies like software defined networking (SDN) are changing the value of redundancy. If software can immediately sense and alert a failure and strategically fail over to back up systems, N+1 redundancy might be all that is needed since meantime to repair is much faster.

While redundancy can be viewed as a "planned" operational function, availability is based on "unplanned" downtime and relates specifically to how many minutes or hours can be tolerated. For example, with 8,760 hours in a year, an availability of 99.9% indicates the ability to tolerate 8.76 hours of downtime per year.

Determining Reliability

It's really the combination of redundancy and availability that go into determining reliability, and there are a variety of industry ratings that indicate reliability. The Uptime Institute's tier ratings is one example that is commonly used. While more detailed information can be found by visiting the Uptime Institute's website at www.uptimeinstitute.com, following is a summary.

- **Tier 1:** Non-redundant capacity components with an availability of

~99.671% and 28.8 hours of downtime per year.

- **Tier 2:** Tier 1 + redundant capacity components with an availability of ~99.741% and 22 hours of downtime per year.
- **Tier 3:** Tier 1 + Tier 2 + dual-powered equipment and multiple uplinks with an availability of ~99.982% and 1.6 hours of downtime per year.
- **Tier 4:** Tier 1 + Tier 2 + Tier 3 + all components fully fault-tolerant including uplinks, storage, chillers, HVAC systems and servers with everything dual-powered and an availability of ~99.995% with 0.4 hours of downtime per year.

  There are additional data center design standards available, which can cause some confusion on where to turn when determining reliability based on availability and redundancy.

  The TIA-942 data center standard calls out various levels of availability, based heavily on the Uptime Institute's tiers, but the requirements are more specific and include a checklist and component count. TIA-942 also covers physical construction, security and other factors outside the scope of the Uptime Institute.

  BICSI-002-2010 defines five classes of availability for data centers – F0 through F5. Classes F1 through F5 appear rather similar to the Uptime Institute tiers, but F0 indicates a data center with no redundancy and is primarily reserved to define a single system or telecommunications room rather than an entire data center.

  It's important to note that metrics like power usage effectiveness (PUE) that are geared towards energy efficiency are not related to the availability and redundancy factors that go into determining a data center's level of reliability.

A Right Sized Approach

Regardless of which availability classification you use, there needs

to be a strategy involved and full awareness regarding the impact of downtime. Many businesses believe they need 100% uptime, 365 days a year, 24 hours per day. But the initial cost to build a data center at each tier, level or class must be considered.

For example, the cost to design a Tier 3 data center rather than Tier 2 can require a 50% increase in capital expenditure, while designing a Tier 4 rather than a Tier 3 typically only requires a 22% increase in expenditure. That's why big business with enormous data centers often create multiple data center "halls" that meet different classifications for "right size" availability and redundancy. This tactic can save on construction costs and allows a business to achieve lower operational cost.