VERTIV.

Reversing the Trend

of Rising Data Center Downtime Costs

Downtime Frequency at the Core and Edge

While speed and capital efficiency are necessities in today's highly competitive data center market, these goals must be put in context of data center availability.

New research from the Ponemon Institute, Data Center Downtime at the Core and the Edge: A Survey of Frequency, Duration and Attitudes, reveals that the 132 core data centers included in the study experienced on average 2.4 total facility shutdowns per year and an additional 10 downtime events isolated to specific racks or servers. In addition, the 1,667 edge locations included in the study experienced an average of 2.7 unplanned shutdowns in a year.

What's particularly alarming about the findings of this report is that the duration of outages rose compared to the last time the study was performed in 2016. The average duration of a total outage in a core data center rose to 138 minutes, an increase of 8 minutes over the previous study. With organizations depending more on their data centers and expanding their edge networks, they are not only experiencing a high frequency of outages but taking longer to recover from those outages.

While the participants in this study were located in the Americas, the results of the study are supported by the Uptime Institute's 2020 Global Data Center Survey. That survey found that, "outages occur with disturbing frequency, that the biggest outages are becoming more damaging and more expensive, and that what has been gained in improved processes and engineering has been partially offset by the challenges of maintaining more complex systems."

While there are many challenges associated with data center management today, including the pressure to deploy capacity with greater speed and cost-efficiency, the core challenge of availability is one that cannot be relegated to a lower priority. This paper proposes strategies organizations can employ to minimize their exposure to downtime, including new approaches to UPS redundancy and scalability, enhanced monitoring and remote access, lithium-ion batteries and high availability power distribution strategies.

Evaluating the Attitudes that Impact Availability

In addition to quantifying downtime frequency and duration at the core and the edge, the Ponemon study also explores the organizational attitudes related to various factors that can impact data center availability (Figure 1).

Across both facility types, cost constraints appear to be a key contributor to downtime. Sixty-nine percent of participants said the risk of unplanned downtime increased in their core data centers as a result of cost constraints, while 62% said the same of their edge facilities. Plus, only half of participants said their senior management fully supports their efforts to prevent downtime at both the core and the edge.

Neither edge nor core facilities were well equipped to recover from an unplanned outage. Only 38% of participants felt they had ample resources at the edge to get the facility up and running if an unplanned outage occurred. This is somewhat expected as these are often remote and unmanned facilities. But it was surprising to see that only 43% of participants felt they had those resources available in their core data centers, potentially contributing to the longer recovery times found in this year's report.

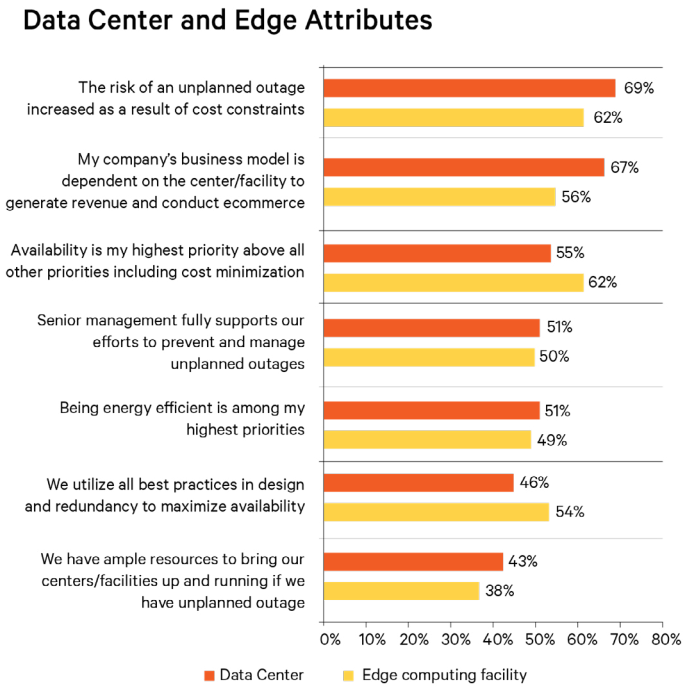

Figure 1: Comparison of edge and core data center attributes.

Finally, edge data centers are more likely to utilize best practices than core data centers, although in neither case are the percentages particularly high. Forty-six percent of participants said they employ best practices in their core data centers compared to 54% in their edge facilities.

These attitudes are showing up in the design of edge data centers. From an availability perspective, we have seen increased redundancy being used at the edge. While core data centers may be shifting to N+1, the edge is perceived as the first line of availability and often deployed as 2N.

Addressing Root Causes

The leading causes of unplanned downtime identified by participants in the Ponemon study included cyber attacks, IT equipment failure, human error, UPS battery failure and UPS equipment failure. When considering these root causes, it's important to reference results from the Uptime Institute 2020 Global Data Center Survey which found that three of four participants said their most recent downtime events were preventable.

Could, for example, many IT equipment failures be prevented through monitoring or replacement prior to failure? The same question can be asked of UPS battery failures. Battery monitoring systems, when properly deployed, can identify potential battery failures before they occur.

Clearly, the cost constraints being imposed on those responsible for facility availability and the corresponding limited use of best practices are playing a role in the relatively high frequency of downtime events revealed by the Ponemon study.

As the Uptime Institute notes in its 2020 Global Data Center Survey: "it is not clear if operators are openly learning from process problems or blaming their managers. It's also possible managers are blaming the operators – or all could be blaming executives for underinvestment. Regardless, the findings point to a clear opportunity: With more investment in management, process and training, outage frequency would almost certainly fall significantly."

Downtime events represent a crisis situation. The focus is always on getting the data center up and running as quickly as possible. But, too often, it appears that the recovery is not followed by sufficient planning and investment to harden the critical data center infrastructure in ways that would reduce the likelihood of future events.

Strategies for Reducing the Frequency and Duration of Data Center Outages

The year of 2020 was a challenge for data center management. Many organizations experienced increased capacity demands due to the global pandemic while simultaneously having to implement new protocols and working with reduced budgets. Yet, these factors cannot be accepted as excuses for increased downtime. Availability of services is more important than ever.

The current situation has also created opportunities to harden infrastructure against future failures. We are seeing more organizations planning for significant infrastructure upgrades, as they prepare their organizations to capitalize on economic recovery. The following strategies can help ensure these upgrades deliver the highest possible availability: infrastructure redundancy, infrastructure monitoring and remote IT management, UPS scalability, lithium-ion batteries, and power distribution design.

Infrastructure Redundancy

Evaluating redundancy and system hardening opportunities is an investment that could provide a positive return by reducing the frequency of downtime events. The challenge is to achieve the right level of UPS redundancy in the simplest and most efficient manner possible. Redundancy needs to be considered in the context of service level agreement (SLA) requirements. There may be a need to increase resiliency to 2N in some cases, or the opportunity to reduce to N in others. System-level analysis and hardening can also reduce the vulnerability to downtime from UPS-related events.

In larger facilities, reserve architectures are increasingly being deployed to reduce the capital costs and increase the efficiency of UPS systems. These architectures fall into two main categories: block reserve and distributed reserve. Block reserve configurations deploy a static transfer switch (STS) and simplify load management. They are generally recommended when SLAs require power to both cords. Distributed reserve architectures increasingly do not deploy an STS and require stricter attention to load management so as not to exceed the redundancy levels. They can be used where SLAs require power at only one cord.

Newer UPS technology, such as that employed by the Vertiv™ Liebert® Trinergy™ Cube, employs internal redundancy to eliminate complexity from multi-module UPS system design. The Liebert Trinergy Cube UPS enables enterprises modernizing their data centers to reduce capital and operating expenses while enhancing availability. By using an internal N+1 configuration, this UPS can shift system-level redundancy to the module level. By integrating multiple power cores within the system, it also provides improved scalability for high-availability 2N or reserve architectures.

Infrastructure Monitoring and Remote IT Management

From telehealth to e-commerce to work from home, the pandemic accelerated the rate of digital transformation. Data center infrastructure monitoring and remote IT management is another example of this. These technologies are not only helping organizations adapt to situations where access to critical facilities is limited due to pandemic restrictions, but are also critical tools in responding faster to outages and protecting against the failure of critical equipment.

By monitoring infrastructure systems in real-time, organizations can often identify early warning signs of impending failure and take corrective action before a failure occurs. These systems also collect the data required to take advantage of predictive analytics and transition to a proactive maintenance strategy. Pairing real-time data with service and maintenance strategies that correlate maintenance with mean time between failures (MTBF) is enabling more effective and efficient equipment service. These capabilities are particularly valuable in providing visibility into remote edge locations and simplifying the management of multiple edge locations.

Figure 2: The Liebert® Trinergy™ Cube features internal redundancy and three-dimensional scalability.

In addition, infrastructure monitoring and management systems can support regular data center health reporting to ensure servers and other equipment are operating in conditions that won't contribute to failures. They also enable modeling to ensure new capacity has the required power and environmental support before it is deployed.

Remote IT management systems, such as serial consoles and KVMs, reduce the need for physical interaction with IT systems, while streamlining management, troubleshooting and recovery. Approximately 80% of IT equipment failures are software or firmware related. In these cases, engineers using remote access tools can typically resolve the situation quickly and remotely to minimize the duration of downtime events.

UPS Scalability

UPS capacity can be a constraint on data center capacity, and, when events like the pandemic create unexpected demand that exceeds UPS capacity, can lead directly to downtime.

Today, there is a solution that enables organizations to minimize their capital investments while maintaining the flexibility to scale the UPS system on the fly. The previously mentioned Liebert Trinergy Cube UPS features a modular, hot-scalable design that allows new capacity to be added without shutting down the unit.

This system also redefines the limits of scalability. It is scalable up to 12.8 megawatts (MW) through its unique three-dimensional modular design. Vertically, the stacked drawers in each core can be individually extracted for service while the UPS continues to protect the load. Horizontally, the system can be scaled up to 1.6 MW by adding up to four individual 400 kilowatt (kW) cores (and optionally a fifth core for 400 kW of redundancy). And orthogonally, up to eight 1.6 MW Liebert Trinergy Cube UPS units can operate in parallel to support a 12.8 MW load.

Lithium-Ion Batteries

Traditional lead-acid batteries are often considered the weak link in the data center's power chain, so it's not surprising batteries are one of the leading causes of downtime. With strings and strings of batteries required to support a modern facility, it can feel as if a possible failure is lurking at any time. These batteries tend to be high maintenance, heavy, and in need of frequent replacement. Advances in monitoring, management, and service have helped to alleviate some of these pains, but not all data centers take advantage of these capabilities.

Lithium-ion batteries have emerged as a viable alternative to lead-acid batteries and should be considered by data center operators seeking to limit their risk of downtime. Lithium-ion batteries have a significantly longer life span than lead-acid batteries, requiring less maintenance and service. Some lithium-ion batteries have also been found to have reduced cooling requirements, resulting in lower operating costs. Perhaps most importantly, when used with a UPS system, these batteries use an integrated battery management system to enhance operation and reduce the risk of system failure and unplanned downtime.

Lithium-ion batteries do come with a higher upfront cost, but their longer life delivers a lower total cost over the life of the battery, even without factoring in downtime costs.

For those organizations not in a position to transition to lithium-ion batteries, implementing a battery monitoring solution for lead-acid batteries provides the visibility into battery performance required to minimize or eliminate outages due to battery failure.

Power Distribution Design

There are multiple options for managing power distribution in the data center, from using large, centralized distribution units to smaller distributed units.

At Vertiv, we've analyzed the impact of various distribution system designs on data center outages. Some operators prefer a "fail small" mentality and have deployed in-the-rack STS units rather than larger centralized STS. Vertiv recognizes the larger STS as a potential single point of failure and has hardened the STS architecture to include redundant power supplies, triple redundant transfer logic, and innovative control algorithms, such as Optimized Transfer, to limit the in-rush due to magnetizing PDU transformers. This has resulted in MTBF an order of magnitude higher than the UPS system.

Investing in Your Future

Making the necessary changes required to minimize downtime events requires shifting from a reactive to proactive approach in which critical infrastructures and the practices for supporting them are evaluated and investments are made to address root causes. In many cases, this will include replacing legacy equipment with new systems and implementing remote monitoring and management systems. While the investment required may be perceived as significant, it should be put into perspective by considering the costs of downtime the organization is incurring every year.

Vertiv.com | **Vertiv Headquarters,** 1050 Dearborn Drive, Columbus, OH, 43085, USA

SL-70705 (R01/21)