Life Is On | Schneider Electric

Why Cloud Computing is Requiring us to Rethink Resiliency at the Edge

White Paper 256

Revision 0

by Kevin Brown and Wendy Torell

Executive summary

Use of cloud computing by enterprise companies is growing rapidly. A greater dependence on cloud-based applications means businesses must rethink the level of redundancy of the physical infrastructure equipment (power, cooling, networking) remaining on-premise, at the “Edge”. In this paper, we describe and critique the common physical infrastructure practices seen today, propose a method of analyzing the resiliency needed, and discuss best practices that will ensure employees remain connected to their business critical applications.

Introduction

The continued growth of Internet of Things (IoT), the rising volume of digital traffic, and the increasing adoption of cloud-based applications are key technology trends that are changing the landscape of data centers.

Large or extra-large cloud data centers now house many of the critical applications for enterprise businesses that once resided in their on-premise data centers. Not all applications have shifted to the cloud, however, and the reasons are varied – including regulations, company culture, proprietary applications, and latency – just to name a few.

As a result, we're left with what we refer to in this paper as a "hybrid data center environment". That is, an environment consisting of a mix of (1) centralized cloud data centers, (2) regional medium to large data centers, and (3) localized, smaller, on-premise data centers. See **Figure 1**. What once was a 1MW data center on-premise at an enterprise branch location may now consist of a couple of racks of IT equipment running critical applications and/or providing the network connectivity to the cloud. The decreased footprint and capacity of the on-premise data center should not be equated to being lower in criticality. In fact, in many cases, what's left on-premise becomes more important.

In this paper, we'll describe the practices commonly seen in the three data center types mentioned above, discuss how the expectations of availability have changed, propose a method for evaluating the required level of resiliency for edge data centers (on-premise) to ensure business objectives are met, and describe best practices for implementing micro data centers on the edge.

Figure 1
Data centers fit into one of three types. This paper focuses on the edge data centers.

Types of data centers

The centralized cloud was conceived originally for certain types of applications – i.e. email, payroll, social media. These were applications where timing wasn't absolutely crucial. But as critical applications shifted to the cloud, it became apparent that latency, bandwidth limitations, security, and other regulatory requirements had to be addressed. Think of the application of self-driving automobiles. There is an extensive amount of compute required for this application to run successfully, and latency can't be tolerated or people get into accidents. Healthcare is another life-critical application; sensors collecting data on patients, or surgical tools providing surgeons with real-time intra-operative feedback. The need to bring the compute closer to the point-of-use became apparent.

High bandwidth content distribution is another application that benefits from bringing the content closer to the point-of-use. Bandwidth costs are reduced and streaming is improved.

For many enterprises, there is often a need (or desire) to keep some business critical applications on-premise. This allows for a greater level of control, including meeting regulatory requirements and availability needs. Sometimes these applications are replicated in the cloud for redundancy.

Schneider Electric White Paper 226, The Drivers and Benefits of Edge Computing, further explains these applications driving us to an ecosystem that includes more regional and localized data centers. In this section, we'll describe each of these data center types and discuss the typical physical infrastructure practices deployed in each.

Centralized data center

Large multi-megawatt centralized data centers, whether they be part of the cloud or owned by an enterprise, are commonly viewed as highly mission-critical, and as such, are designed with availability in mind. There are proven best practices that have been deployed for years to ensure these data centers do not go down. Facilities and IT staff operate these sites with the number one objective of keeping all systems up and running 24x7. In addition, these sites are commonly designed and sometimes certified to Uptime Institute's Tier 3 or Tier 4 standards. Colocation and cloud providers often tout these high availability design attributes as selling points to moving to their data centers.

Common best practices seen include:

- **Redundant critical systems** – critical power and cooling systems are designed with redundancy (often 2N) to avoid downtime due to failure or maintenance activities.
- **High levels of physical security** – it's common to see biometric sensors at doors, man-traps, video surveillance, and security guards around the clock to ensure systems are secure and only accessed by authorized personnel.
- **Organized racks and rows** – in addition to the racks being locked, power and networking cables are organized to reduce opportunities for human error from pulling the wrong cables, plugging dual power supplies into the same power path, etc. Air distribution is planned, and devices like brush strips and blanking panels are used to reduce hot spots.
- **Monitoring** – Sensors and meters are deployed so that Data Center Infrastructure Management (DCIM) and Building Management Systems (BMS) can manage, control, and optimize all data center systems.

Figure 2 illustrates the types of security practices common in these data centers:

Biometric sensors | Man traps | Security guards

Figure 2
Security practices common in centralized cloud and colocation data centers

Regional data center

Regional data centers are closer to the end points (i.e., where data is created and used) and smaller than the large centralized data centers. As described earlier, these data centers exist to bring latency or bandwidth sensitive applications closer to the point-of-use. They are strategically located to address high volume needs. These data centers can be thought of as the "bridge" between central data centers and on-premise localized data centers.

Similar to the large centralized data centers, regional data centers are typically designed with security and availability in mind. It is not uncommon to see Tier 3 designs in facilities like this. Sometimes prefabricated design approaches are deployed here, and reference designs are available as a starting point (see **Figure 3** example).

Figure 3
Example reference design as a starting point for building centralized or regional data centers

Localized data center

A localized data center is one that is co-located with the users of the data center. There are a number of terms used to describe these data centers, including **on-premise data center** or **micro data center**. Localized data centers can range in size from 1-2 MW to as little as 10- 20kW. As enterprises outsource more and more of their business applications to cloud or colocation providers, these data centers are trending towards the smaller end of that range, with sometimes only a couple of racks left in a small room or closet.

In many of these down-sized data centers today, the design practices often equate to a Tier 1 design, with little thought to redundancy or availability. It is not uncommon to see the following in these small on-premise data centers:

- **Lack of security** – Rooms are often unsecured; racks are often open (no doors)
- **Unorganized racks** – Cable management is an after-thought, causing cable clutter, obstructions to airflow within the racks, and increased human error during adds/moves/changes. See **Figure 4**.
- **No redundancy** – Power (UPS, distribution) systems are often 1N, which decreases availability and ability to keep systems up during maintenance.
- **No dedicated cooling** – These small rooms and closets often rely on the building's comfort cooling, which can lead to overheated equipment.
- **No DCIM monitoring** – These rooms are often left unmanaged, with no dedicated staff or software to manage the assets and ensure downtime is avoided.

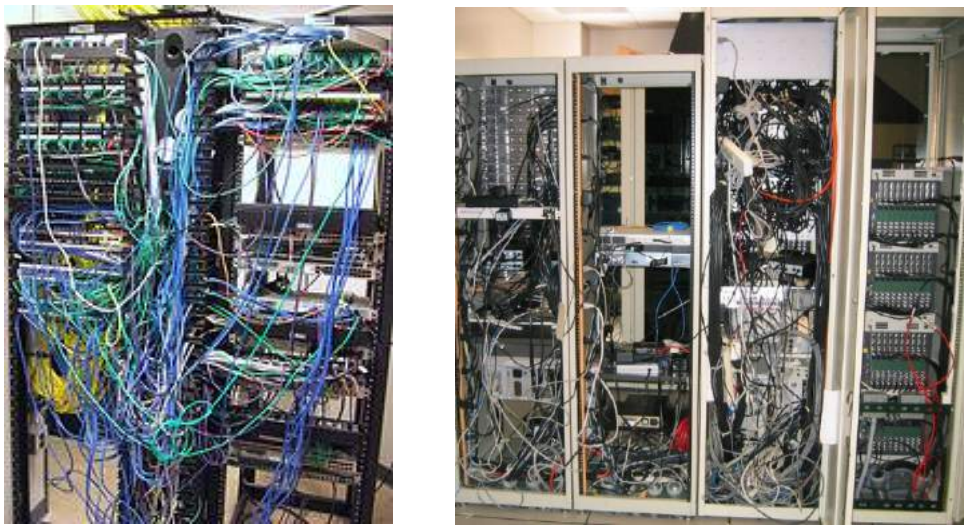

Figure 4
Examples of small on-premise data centers with poor cable management, poor security.

Sites often end up looking like this because as enterprises shift to the cloud or colocation, the few racks that remain are thought of as less important. The focus tends to be on ensuring availability of the bigger data centers. This logic is flawed, however, as often times the remaining racks are equally if not more mission critical.

Consider what's generally left on-premise: (1) proprietary, business critical applications, and (2) network connectivity to the cloud. What are the business productivity implications if I can't access my applications? If we assume the same number of people are still at a particular site, having just a couple of racks remaining on-premise actually **increases the importance on a per rack basis**. The local equipment is critical to connectivity to everyday business applications. With more and more living in the cloud, when that access point is down, employees cannot be productive.

This suggests that a change in how we design these small on-premise data centers is needed. We can no longer focus only on central and regional data centers, and arguably, more focus should be on the localized sites because they are currently the weakest links. Later in this paper, we describe best practices that should be deployed at these sites to ensure a highly connected and productive business.

A more comprehensive availability metric

Given this inter-connected hybrid environment, an important question we must address is: do we need to rethink the way we talk about criticality and redundancy? The tools we use as a data center industry today, focus on how I ensure a single data center is as robust as possible. Tier levels help us design a single site to achieve a particular availability level (number of 9s). Failure is commonly defined as disruption to any IT equipment within a particular data center.

The tools and metrics don't contemplate dependence on multiple data centers, number of users impacted by failure, criticality of business functions impacted, or application (software) fail-over. We believe this is necessary moving forward.

The shift in availability expectations

The expectations of employees today vary from those of past generations. As the workforce ages and shifts towards a greater percentage of millennials, there is an expectation that follows. This generation was raised with an "always on, always connected mentality", where IT devices and systems are expected to work, all the time. Tolerance for disruption in service is low. Technology is important to them in their

daily lives, including in the work place. In fact, 82% of millennials believe that work-place technology would influence them when deciding to accept a new job.[1]

If we anticipate this trend continuing, it is crucial we look at more holistic ways of reporting resiliency of data centers, that provide us with the visibility needed to make the right design changes. As the old saying goes, "you can't manage what you don't measure". Resiliency metrics must evolve to meet today's business needs.

A different approach

A different viewpoint on availability will drive different action. **Table 1** illustrates the comparison of today's (old) paradigm, and the new paradigm we believe is necessary to drive necessary action.

Table 1
Shifting paradigm of data center failure

Old paradigm	New paradigm
Focused on the centralized data center	Focused on the hybrid environment
Failure is when IT equipment in a rack is impacted	Failure is when user experience is impacted
Doesn't comprehend remote sites or people/functions	Criticality is impacted by number of employees impacted and job functions

Think about the utility (power) company, and how they look at availability. They don't just look at their generation plants and HV lines (their "centralized data center"). They trim back tree branches, maintain pole-mounted transformers, and ultimately measure success based on delivery of power to their customers (their "edge" data centers). The data center industry needs to move to this utility model, where the edge is as important (if not more important) as the centralized data centers.

The availability of two systems in series, meaning you depend on both being available is calculated as:

$$Availability_{system} = Availability_1 * Availability_2$$

Let's start by thinking about a single user, who is dependent on their local on-premise data center *and* the central data center to be available or productive. To compute data center availability from their perspective, we would use this formula. If for instance, the central data center had an availability of 99.98% (Tier 3 data center, with 1.6 hours of downtime), and their on-premise data center had an availability of 99.67% (Tier 1 data center, with 28.8 hours of downtime), the total downtime from that user's perspective would be 99.98%*99.67% or 99.65% (30.7 hours of downtime).

If we now take the viewpoint of the CIO, how do I evaluate the impact of my entire eco-system of data centers on business productivity and connectivity? Not every data center is dependent on every other data center being up for employees to

[1] http://www.dell.com/learn/us/en/uscorp1/press-releases/2016-07-18-future-workforce-study-provides-key-insights (Last accessed 10/31/2016)

function. For example, a branch office in London is not dependent on a branch office in California, but they both may be dependent on a central data center in New York.

Not all data centers have the same impact on the business. The number of employees impacted is a factor. For example, an on-premise data center with 1000 employees may be deemed more critical than one with 10 employees. **Table 2** illustrates the amount of people-hours of downtime of an example eco-system with one Tier 3 central data center and 10 Tier 1 localized data centers, each with 100 employees. It is evident from this table that the Tier 1 edge data centers drive the total downtime number. The greater the number of edge sites the fewer the number of hours with no sites down.

Table 2

Availability of 10 edge data centers and 1 central data center, with people impacted factored in

Data Center Availability						
Description	**Availability**	**Downtime (hrs)**	**# Sites**	**# people/site**	**Total people impacted**	**People-hours of downtime/yr**
Tier 1 edge data centers	99.67%	28.82	10	100	1,000	28,820
Tier 3 central datacenter	99.98%	1.58	1	0	1,000	1,580
					Total people-hours of downtime/yr	30,400
					Availability	99.65%

The table above was a simple scenario with 2 tiers of data centers where 1000 people were impacted by both tiers. When more data centers are present, each with different availability levels and numbers of people impacted, the math is not as straight forward. In addition, this is incomplete because it excludes a rating of each site by business function performed at the sites. A site that serves the function of customer service or manufacturing would likely be more critical than one filled with administrators that could work remotely if their network went down.

Criticality analysis

Qualitative criticality analysis is a proven method of evaluating risk and prioritizing corrective actions (also referred to as Failure Modes, Effects and Criticality Analysis (FMECA)). It is well-documented in reliability engineering publications. This analysis includes rating the severity of the effects of a failure with a Risk Priority Number (RPN). The RPN is based on 3 factors: (1) severity of failure, (2) likelihood of occurrence, and (3) detection of the failure.[2]

We propose that the best approach to assessing all sites holistically is with a scorecard, such as the example in **Table 3**. This will aid CIOs and data center managers in identifying the highest priority sites to focus improvements on. The scorecard consists of the availability and associated downtime of every site in the hybrid data center environment (measured ideally), and most importantly a criticality rating for each site. See **sidebar** for more about the science behind criticality ratings[2]. In the case of data centers, each site's "severity of the effects of failure" is based on:

- Number of people impacted
- Function performed

A scale of 1 to 5 is common, where 1 is the lowest in impact on the business if the site goes down, and 5 is the greatest impact. Although a qualitative rating system, this provides a systematic approach to looking at all sites in the business' data center eco-system. Note, that different businesses will have they own preference for how to come up with the values used here. The key is having a consistent method of rating all sites.

[2] http://www.weibull.com/hotwire/issue46/relbasics46.htm (Last accessed 10/31/2016)

In this example, there are five data centers that make up the hypothetical eco-system. The annual downtime of each is multiplied by the defined "severity of the effects of failure" score of each, to get the weighted score.

From this, you can simply sort the sites by score, where the highest score gets the highest priority for data center improvements. You can also calculate the percent of the score for each site (as the example shows for "Site impact on score" and the sites with the highest percentage are the highest priority.

Table 3

Sample scorecard to help prioritize data center improvements

Data Center Scorecard					
Site Name	**Availability**	**Annual Downtime (hours)**	**Severity of Effects of Failure (1-5)***	**Score (weighted for criticality)**	**Site impact on Score**
1	99.98%	1.752	2	3.5	0.4%
2	99.20%	70.08	4	280.3	30.0%
3	99.60%	35.04	1	35.0	3.7%
4	98.60%	122.64	5	613.2	**65.5%**
5	99.98%	1.752	2	3.5	0.4%
			Overall criticality score:	*935.6*	

This is a step and repeat approach. Once the availability of Site 4 in this example is improved, a new site will rise to the top of the list as most important. Through this continuous improvement cycle, the sites with the greatest impact will be improved.

With the right availability reporting method in place, it will become apparent where design improvements are necessary to ensure the greatest productivity and business return on investment. **In the majority of cases, going through this exercise will demonstrate that the edge data centers, which often have a lower availability, have a higher impact on the business.**

Best practices at the edge

With the right metrics and methods in place, the need to rethink the design of the data center systems at the edge will become apparent. The typical design practices at the edge (as described earlier) are inadequate given the mission-critical nature of these sites. Improvements should focus on:

- Physical security
- Monitoring (DCIM), operational practices, remote monitoring
- Redundant power and cooling
- Dual network connectivity

In the following sections, we describe key best practices to deploy at the edge. Schneider Electric White Paper 174, *Practical Options for Deploying Small Server Rooms and Micro Data Centers*, discusses in greater detail how to make realistic improvements to power, cooling, racks, physical security, and monitoring in small server rooms and branch offices with up to 10kW of IT load.

Secure, safe environment

Small local data centers are often placed within a highly accessible room such as a shared office space. There is often no dedicated space, so open racks are unsecured. This presents a security risk, whether it be from malicious or accidental activities.

Best practices to reduce these risks include:

- Move equipment to a locked room or locked enclosure(s).
- Ensure biometric or other access control.
- For harsh environments, secure equipment in an enclosure that protects against fire, flood, humidity, vandalism, and EMF effects.
- Deploy security & environmental monitoring 24x7, and video surveillance

Examples of secured enclosures are shown in **Figure 5**. These are often pre-fabricated with all necessary support infrastructure included.

Figure 5
Examples of micro data centers by Schneider Electric

Data center management

The management and operations protocol often differs from edge site to edge site (if a protocol exists at all). Managing hundreds or thousands of edge sites can be costly and time consuming, and the availability at many sites is dependent on shared infrastructure systems at the facility, like generators, switchgear, and chillers.

Best practices to reduce these risks include:

- Take inventory of existing management methods and systems.
- Consolidate to a centralized monitoring platform of all assets across sites.
- Deploy remote monitoring when resources are constrained. See White Paper 237, Digital Remote Monitoring and How it Changes Data Center Operations and Maintenance, for more on how remote monitoring can help reduce downtime.

Power and cooling

The power and cooling infrastructure systems (like UPSs and air conditioners) are generally deployed at edge sites with no redundancy. This results in single points of failure as well as an inability to concurrently maintain systems. In some cases, no dedicated cooling exists to support the rooms, resulting in over-heated equipment. The infrastructure systems are often shared with the rest of the multi-purpose building, so the availability of the edge data center is dependent on the availability of those shared resources.

Best practices to reduce these risks include:

- Measure temperature and humidity to understand level of cooling needed (i.e. passive airflow, active airflow, or dedicated cooling).
- Consider redundant power paths for concurrent maintainability in critical sites.
- Ensure critical circuits are on emergency generator.

Figure 6 shows an example of a Tier 3 micro data center consisting of a prefabricated integrated solution in a single 42U enclosure, with redundant UPSs and power distribution.

Figure 6
Example of a 1 rack micro data center with redundancy built in

Network connectivity

As discussed earlier, connectivity to the cloud is crucial for the edge sites. Yet, often times, there is a single internet service provider providing that connection. This represents a single point of failure. Cable chaos in the networking closets also breeds human error.

Best practices to reduce these risks include:

- Consider adding a second network provider for critical sites.
- Organize network cables with network management cable devices (raceways, routing systems, ties, etc.).
- Label and color-code network lines to avoid human error.

Conclusion

Cloud adoption is driving more and more enterprises to a hybrid data center environments of cloud-based and on-premise data centers (the edge). Although what's left on-premise may be shrinking in physical size, the equipment remaining is even more critical. This is because:

- With more applications living in the cloud, the connectivity to the cloud is crucial for business operations to continue.
- There is a growing culture of employees that demand "always on" technology and cannot tolerate downtime disruption.

Unfortunately, most edge data centers today are fraught with poor design practices, leading to costly downtime. A systematic approach to evaluating the availability all data centers in a hybrid environment is necessary to ensure investment dollars are spent where they will get the greatest return.

A scorecard approach was presented which allows executives and managers to view their environment holistically, factoring in the number of people and business functions of each data center. This method identifies the most critical sites to invest in.

Prefabricated micro data centers are a simple way to ensure a secure, highly available environment at the edge. Best practices such as redundant UPSs, a secure organized rack, proper cable management and airflow practices, remote monitoring, and dual network connectivity ensure the highest-criticality sites can achieve the availability they require.

About the authors

Kevin Brown is the Chief Technology Officer of the Data Center Division at Schneider Electric. Kevin holds a BS in mechanical engineering from Cornell University. Prior to this position at Schneider Electric, Kevin served as Director of Market Development at Airxchange, a manufacturer of energy recovery ventilation products and components in the HVAC industry. Before joining Airxchange, Kevin held numerous senior management roles at Schneider Electric, including Director, Software Development Group and Sr. VP of Data Center Solutions.

Wendy Torell is a Senior Research Analyst at Schneider Electric's Data Center Science Center. In this role, she researches best practices in data center design and operation, publishes white papers & articles, and develops TradeOff Tools to help clients optimize the availability, efficiency, and cost of their data center environments. She also consults with clients on availability science approaches and design practices to help them meet their data center performance objectives. She received her bachelor's of Mechanical Engineering degree from Union College in Schenectady, NY and her MBA from University of Rhode Island. Wendy is an ASQ Certified Reliability Engineer.

Resources

Cost Advantages of Micro Data Centers
White Paper 223

The Drivers and Benefits of Edge Computing
White Paper226

Digital Remote Monitoring and How it Changes Data Center Operations and Maintenance
White Paper 237

Browse all white papers
whitepapers.apc.com

Browse all TradeOff Tools™
tools.apc.com

Contact us

For feedback and comments about the content of this white paper:

Data Center Science Center
dcsc@schneider-electric.com

If you are a customer and have questions specific to your data center project:

Contact your Schneider Electric representative at
www.apc.com/support/contact/index.cfm