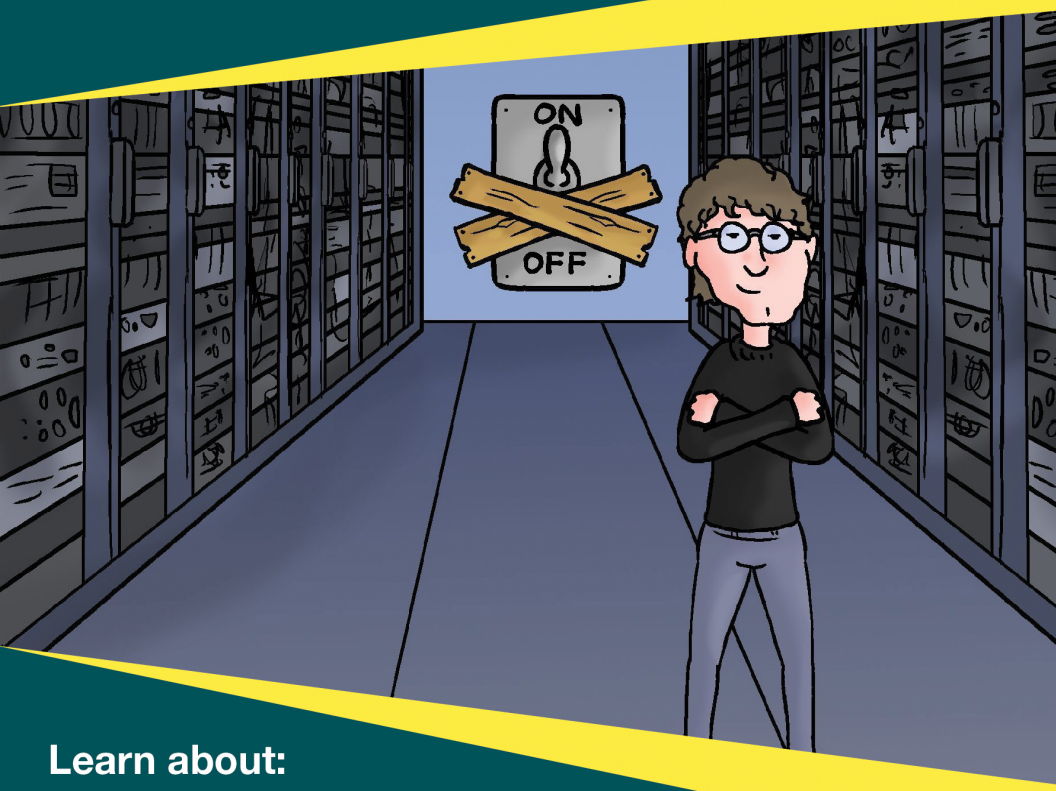
Conversational
Availability for the Always-On Enterprise

A ConversationalGeek™ Book



Learn about:

- What's causing the demand for an "always-on" Enterprise
- The difference between backup and availability
- What steps are necessary to achieve availability in the enterprise

By Nick Cavalancia (Techvangelist)

Conversational Availability for the Always-On Enterprise

By Nick Cavalancia

ConversationalGeek™

Conversational Availability for the Always-On Enterprise

Published by Conversational Geek Inc.

www.conversationalgeek.com

Trademarks

Warning and Disclaimer

Additional Information

For general information on our other products and services, or how to create a custom Conversational Geek book for your business or organization, please visit our website at ConversationalGeek.com

Publisher Acknowledgments

All of the folks responsible for the creation of this guide:

Author:	Nick Cavalancia
Project Editor:	J. Peter Bruzzese
Copy Editor:	John Rugh
Content Reviewers:	Shanna G. Giarrano

Note from the Author

We all know the simple truth of up is good, down is bad. But, applications crash, hardware fails, and data becomes corrupt – it's one of the reasons we have a job in the first place!

So, is it even possible to get operations to an always-on state where availability is the norm?

I'll try to provide the best answer possible, but as you can see from the lack of thickness of this book, it's not meant to be an exhaustive reference on establishing availability in the enterprise. It was written with the intent of walking through the need for availability and of reviewing the improvements in technology and strategy that assist in getting to availability. It then aims to show you that it's really, really possible today to actually achieve both availability of mission-critical applications, and to become an *always-on* enterprise.

So, grab a comfortable chair, and let's dive into what it's going to take to get to the always-on enterprise.

-Nick Cavalancia

The “Conversational” Method

We have two objectives when we create a “Conversational” book: First, to make sure it’s written in a conversational tone so it’s fun and easy to read. Second, to make sure you, the reader, can immediately take what you read and include it in your own conversations (personal or business-focused) with confidence.

These books are meant to increase your understanding of the subject. Terminology, conceptual ideas, trends in the market, and even fringe subject matter are brought together to ensure you can engage your customer, team, co-worker, friend and even the know-it-all Best Buy geek on a level playing field.

“Geek in the Mirror” Boxes

We infuse humor into our books through both cartoons and light banter from the author. When you see one of these boxes, it’s the author stepping outside the dialog to speak directly to you. It might be an anecdote, it might be a personal experience or gut reaction and analysis, it might just be a sarcastic quip, but please don’t skip these “geek in the mirror” boxes.

Greetings. I’m the author’s inner geek. Within these boxes I can share just about anything on the subject at hand. Read ’em!

Demand for the “Always-On” Enterprise

How much downtime can your business really handle? Of course, you have an immediate answer. But go ask your CEO – do they have a different answer? Perhaps even one that’s a bit unreasonable? I’m assuming you’d come up with a well-thought-out number of minutes or hours, based on your ability to recover, Whereas, your CEO will likely just tell you “none.”

Sound about right?

There's a reason why your CEO is thinking that way (and it's not like they're crazy, or have unrealistic demands, right? That never happens…). The reason they're thinking about zero downtime is because the expectation of customers, partners, and even employees is that services you provide are *always* available.

While it used to be acceptable for even enterprise companies to be "open for business" from 9 to 5, today's customer demands the ability to access and interact with a company 24/7 via their website, telephone, email, and even chat. Purchases, account updates, access to data, support, and more anytime you want is now the norm.

Now, you might think it's just large companies like Target, Bank of America, or IKEA that have to be always on – but you'd be wrong. Think about it – the same person who is experiencing this great anytime, anywhere access to information and services from the larger companies isn't going to excuse you just because you're not as large. They'd just wonder why you can't provide the same level of service in this day and age… and, perhaps, move on if you don't.

So, every organization needs to be up and running consistently – *and that's no easy task*!

An "always-on" enterprise, regardless of size, has invested in a modernized data center, where hardware and operating system have been kept up to date (so no Windows Server 2003 for you!), and new technologies have been embraced.

Veeam recently published the results of their *2014 Data Center Availability Report*, which reveals only 25% of organizations already consider their data center to be "modern". Look for more stats from this report throughout!

What's in a Modern Data Center?

Given the stat above, it's likely you either don't have a modern data center, are working towards one, or aren't sure if you do. The list below gives you an idea of the more common initiatives involved in modernizing the data center.

- Server virtualization
- Upgrades to storage and operating systems
- Data protection and disaster recovery
- Network virtualization
- Virtual desktop Infrastructure (VDI)
- Private cloud
- Infrastructure as a Service (IaaS)
- Software as a Service (SaaS)
- Disaster Recovery as a Service (DRaaS)

It's a pretty tall list, so don't think you need to implement *every* initiative shown. This is more a guideline to let you know if you're organization is moving in the direction of "modern."

Does a modern data center equate to being always on? Not necessarily. In Veeam's *Availability* report, about a third of organizations **did not** see being *always on* as a key business driver for modernizing their data center.

If your organization is one of those that are either not quite there, or haven't started, given the somewhat daunting list above, the question does need to be raised "*Do you even need a modern data center?"*

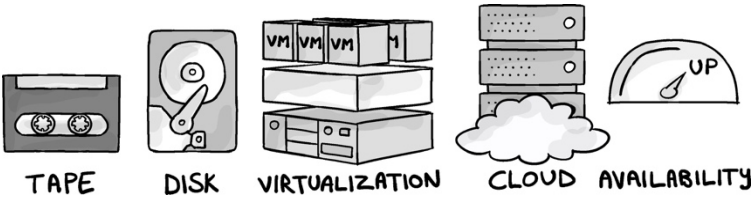
There are more reasons than just availability that are driving the move to a modern data center. Organizations that have modernized their data center are focusing efforts on these

newer technologies to lower operational costs, improve security, and provide more efficient services. But, in the end, it all goes back to customer and employee expectations around accessing applications and data necessary for them both, and therefore the organization, to be productive and successful.

End-user demand is a powerful motivator for moving to always on. In Veeam's *Availability* report, 63% of organizations cited "*more frequent, real-time interactions between customers, partners, suppliers and employees"* as a key driver for minimizing downtime.

That being said, if you strive to provide an always-on infrastructure, the modern data center is somewhat a requirement. As you'll see in this next section, the evolution from just backup and restore to true availability relies heavily on the existence (and implementation) of new technologies.

The Evolution to Availability



Availability today means organizations can recover from the loss of data, applications, systems, or locations with minimal or zero downtime. This is a reality that is somewhat new. While it's possible today to come really close to this definition (the *zero* downtime is the sticking point), to achieve availability we've had to evolve a bit from backup technologies and methods of yesteryear to the modern capabilities we enjoy today.

If you go far enough back in IT like I do, you can remember a time when simply having daily, weekly, and monthly backups were considered enough to protect the business. Backups ran at night after business hours, and you used a simple grandfather-father-son tape rotation method to establish a long-term ability to recover.

If you go back as far in IT as I do, you can answer the following question:

Q: What was the main difference between DOS v3.2 and v3.3?

A: v3.3 broke the 33MB partition barrier!

When you stop and think about what kind of protection a nightly, weekly, or monthly backup really afforded your organization, you realize, at best, it involved the loss of at least

a day's worth of work. Fast-forward to today and this, obviously, wouldn't cut it for an Always-on Enterprise.

Recovering back to *yesterday*? Seriously?

So, what's kept you from getting to "always-on"?

There have been two sets of constraints over the years stopping organizations from achieving true availability (that, thankfully, have been slowly eroding away):

The first is the available technology. Travel in the 'Wayback Machine' a few years and the only way to get high availability was using a clustered server. Today, we have so many more options to get data, services, and applications moved to, and running on, other systems in alternate locations.

The second is the cost of implementing the availability strategy. The strategy you use depends heavily on the tech available to you. And, as you'll see, our thinking around what's possible with regard to availability has changed through the years.

Let's look at each of these and see how they benefitted the evolution from backup to availability.

The evolution of backup technology

Backup technology used to limit us to a simple once a day backup because of the *constraints in the physical world*. Tape systems only go so fast, hard drives can only pull data so quickly, backup processes impact application performance and user productivity, and scheduling downtime during production usually wasn't an option.

Enter direct-to-disk backups, the cloud as a backup target, and even hybrid-cloud backup, along with more advanced methods of detecting changes, deduplication, and compression – and the technology began to empower the backing up of far more data much more quickly while, somehow, still maintaining the

ability to recover the same amount that used to take an entire night to back up. You begin to see how improvements in technology start to overcome the limitation of the physical world of servers, data, and backup.

A backup tape drive in the 90's had a 2:1 compression ratio and often utilized further software-based compression to shrink the backup size. Today's change identification, compression and deduplication methods can result in as little as .1% of the total data size being used to complete an incremental backup.

Then add in virtualization – a complete game changer for availability, and you could now back up changes made to an entire system, rather than a file, application, or database. Snapshot technology, along with replication capabilities took the concept of needing to back up a subset of data at a particular time and gave IT the ability to have access to a backup that protected the entire machine, supplied more backups at faster intervals, and facilitated the copying of those backups to another server. This had a real "WOW" factor to it!

And as technology evolved, getting us closer to availability at each turn, so did backup strategies.

The evolution of the cost of a protection strategy

I previously mentioned the Grandfather-Father-Son tape rotation – that was a backup and recovery strategy that, at the time, was cutting edge and had withstood the test of time. Its cost was predictable (as you only needed a fixed number of tapes). But, like technology, recovery strategies have evolved, each one successively taking advantage of the latest technology and finding a lower availability price-point.

The following list, in more-or-less chronological order, demonstrates the changes in thinking from backup to availability, as technology has improved and cost has reduced.

- Backup and Restore – this was a truly viable method when file-level backups actually represented the entirety of an operating system. The focus here is simply file level. Even today when backing up virtual machine images, the focus of backup and restore is to simply restore that image file.
- Backup and Recovery – this is a shift in the goal while utilizing largely the same technology as backup & restore. It's no longer about just restoring the data in the backup, but really concerning itself with whether you can get the intended application or system back into operation.
- Business Continuity / Disaster Recovery (BCDR) – when this concept was introduced, we came close to availability's intent with this line of thinking. There was a shift from just recovering a system or application to truly "how do we get everything operational again." The problem? It was seen as costly, requiring expensive redundant hardware, use of cold or hot sites lying in wait, comparatively slow delivery mechanisms, such as UPS and FedEx, getting on a plane yourself to work on the remote system in person, or accessing a remote system via slow WAN links to get copies of data to off-site locations.
- Use of Virtualization – by abstracting the need for actual hardware, virtualization created the ability to instantly allow redundancy through image and snapshot backups and replication to offsite locations.

Virtualization also meant far fewer physical servers, thereby reducing the cost of maintaining a secondary site. This advanced the BCDR movement, giving it a new means to replicate entire environments far more quickly and cheaply.

- Use of the Cloud – the cloud as a storage medium has brought simple, inexpensive, and yet secure offsite backups to the masses. However, using it for spinning up compute is still relatively new, and it brings new possibilities for having on-demand compute as a resource, leveraging the use of virtualization and replication.
- Disaster Recovery as a Service (DRaaS) – Combining the use of virtualization and the trust in the cloud, building on the simple concept of a hot site, DRaaS provides (usually) manual failover services to a replicated, hosted copy of anything from a single server to an entire network. For tier 1 applications, even this level of service can be justified, despite the high cost.
- Recovery as a Service (RaaS) – while DRaaS tends to focus on the replication and then failover of services in the event of a disaster, you can think of RaaS as Managed DRaaS. That is, instead of simply providing a failover environment, RaaS involves testing of replicated backups to ensure viability, continuous recovery of data at very small time windows, and automated failover, allowing you to make very complex environments live in a very fast timeframe.

By looking through this evolution, you can see some common themes:

1) It's less about data & more about continuity – By the end of the evolution to availability, no one is really concerned about which files are being recovered; only that the critical systems and applications are available.

2) A shift toward protecting larger data sets – This creates the need for identifying smaller incremental changes to be protected, rather than the entire data set as a whole.

3) Smaller backup and recovery windows – This is a by-product of the smaller incremental changes, but is a key enabler of availability.

4) Move from protecting just on-premises to some degree of off-premise thinking – Sure, you are still protecting what's on-prem, but use of the cloud, IaaS, SaaS, etc. are all part of the availability plan.

5) An improved cost-to-benefit ratio – While the last option (RaaS) probably comes across as the most expensive (in a list that is supposed to be getting less costly as we go), consider what the cost would be to achieve this using simple backup and restore… and then add on the cost the lack of availability of backup and restore brings. Even though the check written each month is far more than the cost of a few backup tapes, this option guarantees you are recoverable in a very short timeframe.

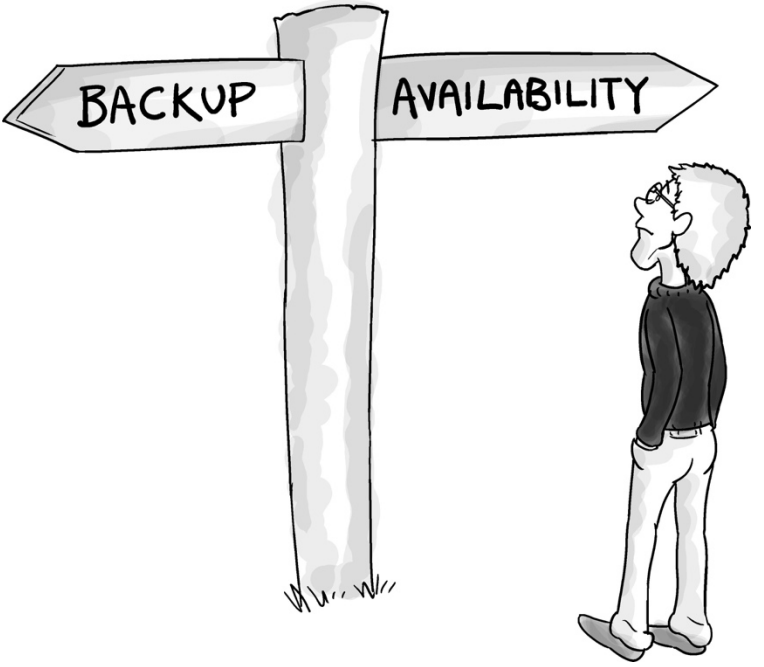
All of this puts us squarely where we are today – thinking about continual recovery processes utilizing both on-prem and cloud-based infrastructures together in the most cost-effective way to empower true availability.

While you are likely investing in some of the technologies that make up a modern data center, you probably still find your backup and recovery efforts somewhere in the middle of the evolution to availability.

So, how do you make the transition from backup to availability?

The answer lies in first defining the difference between them.

Backup vs. Availability



Everyone, at some point, still performs backups. They might go by another name, such as snapshots or replication. But at their core, they serve the same purpose.

Regardless of where you are in the strategy or technology spectrums previously listed, when it comes down to it, you still focus, at some level, on backups.

It may seem that availability is merely an advanced form of backup and recovery (probably because you're thinking about it in terms of backups). In reality, nothing could be farther from the truth.

You see, backups, at their core, define a data set that's being protected and can be recovered, and then have you work *forward* to see if what you've backed up will actually get you

operational. For example, backups of entire servers and system states in Windows may still require an additional recovery step, such as restoring and bringing databases in Exchange or SQL Server into a consistent state.

So a single backup doesn't necessarily equate to a recovered application, server, or environment. On the flip side, availability demands that you begin with what needs to be kept running and work *backwards* to identify backup technologies that meet the need and the backup and recovery strategy that supports the end goal.

So, just how far apart is your backup strategy from availability?

There is some difference that exists between the requirements your business has around keeping systems and applications available, and the backup/recovery capabilities you currently can provide. You likely have identified applications that have specific SLAs, and you also know how quickly your backup solution can recover said system. Then you multiply that same scenario by the number of critical applications you need to recover. More likely than not, you *can* recover some of them in the time allowed, but not necessarily all of them.

Others of you may not have defined the availability requirements, so you don't know quite how large the gap between backup and availability really is.

82% of organizations identify themselves as having a gap between backups and availability in the *Availability* report. If you're not sure, you probably fit into that group.

So let's give availability some level of definition.

Because you're already familiar with backups, let's borrow some common terms from backups you're familiar with – after

all, even with availability, you will still work backwards all the way to a backup of some kind at some point in the process.

Availability itself is an objective – a goal to be up and running as close to always as possible. Backups, too, have a number of objectives that specify operating parameters for your backups. Two specific backups objectives can be used to help define *availability*: *Recovery Time Objective* and *Recovery Point Objective*.

Recovery Time Objective (RTO) is the amount of time allowed for recovery to take place. This usually is defined on a per-application or per-server basis, as each application, system, or data set has its own level of criticality to the organization.

Recovery Point Objective (RPO) is the point in time the recovery should bring the business back to. Or, in other words, the amount of data (in terms of time) the organization is willing to lose.

So, if the RTO is an hour and the RPO is 4 hours for a given application, The RTO dictates you have an hour to recover the given data set, whereas the RPO dictates the IT organization needs to be planning for backups of that data set every 4 hours, at a minimum.

So, what should the values be for true availability?

While there is no industry-defined value, a good rule of thumb is to consider both an RTO and RPO of just *15 minutes*. I know, I know… sounds like I'm a crazy person, right? But we're talking about applications so important to the organization they need to be truly *available.* To paraphrase Miracle Max, "*There's a big difference between mostly available and available. Mostly available is slightly dead.*"

No IT pro ever wants to tell the executive team their critical app is "slightly dead." By using the 15 minute RTO/RPO values, you'd be up and running in 15 minutes only having lost 15

minutes of data. That's a pretty tight window, but you need to consider each application in your environment, its criticality to the business, and how long operations can really go without it.

Availability has been around for years, but the cost of achieving a 15-minute RTO/RPO has been too high to make it cost-effective.

Regardless of the specific values you come up with, the goal should be to have very small values to ensure both operations are up very quickly and very little data has been lost. And, remember, a short RTO/RPO *is* possible (more on that later).

So, how do you achieve availability? Is it just by buying some new technology? Coming up with a new strategy? Or by doing something different, instead?

Achieving Availability

Hopefully by now, you recognize the importance of getting your organization to a high degree of availability. The sheer number of identifiable mission critical applications and systems within your organization alone would dictate how important availability truly is. So, how many is that? What percentage of your application and system workloads would you consider mission critical to the organization?

Given the growth in the need for availability, would it surprise you at all that the number of mission critical applications and systems in your organizations will only grow, increasing the need for availability? The time to get this right is *now*.

According to Veeam's *Availability* report, almost half of all workloads are considered to be "mission-critical" today. Than number will increase by 15% over the next two years.

The really good news here is you're no longer constrained by technology. The ability exists today to meet very tight RTOs and RPOs by simply taking advantage of technologies that are part of a modern data center strategy. Now, it may cost you more to implement, but given the importance of a given application, that cost may be justified.

Since the technology exists, achieving availability is a lot more about an execution strategy and choosing the right technology to get there. Here are 5 steps to help you make progress toward achieving availability.

1) Understand the business requirements – While I've discussed using backup objectives, don't start there (those will come in a moment); begin with defining what the business needs to operate. List out the applications that are both internally and customer facing and can be thought of as mission critical.

2) Establish tiers of Criticality – while every application and system in the current list will be "critical", we all know you can weasel out different tiers of importance from management if pressed hard enough. After all, if you were in a situation where you could only get one of the mission critical systems up, I'm sure the executive team would be able to tell you which one they want (and then, of course, they'd fire you for only being able to recover one… but I digress). The tiers of

importance may very well be a list of applications in descending order of priority, or you may be able to put them in groups based on equal importance.

3) Develop per-application recovery objectives – Here's where you put some actual numbers to quantify what it means to be both critical and at a given tier. Ask the question "How long can we be without this application, and how much data can we afford to lose?" to develop RTOs and RPOs. Any answer, including the seemingly unreasonable "None and none" is an acceptable answer. Why? Because it gives you a goal to hit (that is why they're called *objectives*) as well as a starting point with which to begin building an availability strategy.

4) Build a strategy around available technology – So, let's say your ERP system needs to hit that 15-minute RTO/RPO. How are you going to do that? Look at what's available (and I don't mean just what you have today, but what is available in the industry) to create a method by which you can achieve the objectives. With a 15-minute RPO, you're most likely going to need to look at image-based backups of block changes replicated to a mirrored virtual server. If you're looking to protect against natural disasters, that server's location would need to be remote. The 15-minute RTO means whatever redundant infrastructure you have in place will need to be lying in wait, ready to be failed over to at a moment's notice. It's this kind of thinking that's just the *start* of planning your strategy.

5) Test, test, test – This is the often- overlooked, most critical step. I mean, c'mon – we're talking about your

most critical applications here with a 15- minute margin of error! You need to be continuously (not periodically… *continuously*) testing each and every backup that is recovered to the alternate site to ensure – without a doubt – that when needed, the application *will* recover.

Somewhere between steps 4 and 5 is, of course, implementing the technology. I purposely left that out, as it's more a tactical aspect, rather than a strategic one.

You can't just buy something, switch it on and… BAM – Availability! It's going to take careful assessment and planning to craft a strategy and execution plan that will ensure the highest levels of availability possible.

Before I end this section, I want to cover a few considerations that are a mix of final thoughts and notes that didn't really fit neatly anywhere else in the book, in an effort to give you the best chance of making availability happen.

Availability-enabling considerations

- The Cloud and the 3-2-1 rule – A common rule of thumb for backups is to have 3 copies on 2 different media types with 1 of the copies stored offsite. The cloud provides an extremely cost-effective means to very quickly move a backed-up data set (whether it be files, a VM image, etc.) offsite. Remember, disasters take many forms, including the loss of location, in which case having an offsite copy will not only be helpful, but necessary.

- Going offsite doesn't guarantee availability – While having an offsite option (whether public or private cloud) helps from a DR and backup perspective,

remember that option is still only as good as your cloud provider. Be sure you know who you're signing up with, and get both their SLA history and a remediation plan should something happen.

- Backups still have a role – Remember, availability is a goal; backups – even if they change their spots a bit moving from file-based to VM image-based – are still a strategic part of the tactical execution of the availability strategy.

- Even local storage-based snapshots can help – These backups have less impact on production systems and can be created frequently. While they have no real place in the 3-2-1 rule (as they sit in the same place your VMs are), they still can help bring about availability, depending on the type of failure.

Making Availability a Reality

If the demand to be always on is low for your organization, don't kid yourself – it's coming. Eventually, every organization will find some level of need for availability, and treating this matter in a proactive fashion now will help ensure your success.

Today, there is an abundance of technology able to support the availability of mission critical applications and systems. By implementing some or all of the modern data center initiatives, you'll also be putting in place the pieces necessary to build a proper availability strategy that takes advantage of backups while not making them the focus.

By identifying the business requirements around availability, establishing relevant recovery objectives, and working backwards to take advantage of those parts of the modern

data center you've implemented, you'll be well on your way to achieving a state of availability.

Don't take it all on at once; start with something… but start *now*.

Easily “converse” about availability and the modern data center.

Never before in the history of IT have we been at a time when availability of mission critical applications and systems has been both in such high demand and is actually possible to achieve. Advancements in the modern data center have led to developments in backup and recovery strategy that facilitate recovery with even the smallest of timeframes. If you’re organization has critical applications that employees and customers alike depend on, this book will allow you to talk about the needs and requirements of availability.

About Nick Cavalancia

Nick Cavalancia has nearly 20 years of enterprise IT experience, is an accomplished consultant, speaker, trainer, writer, and columnist, achieving industry certifications including MCSE, MCT, Master CNE and Master CNI. He has authored, co-authored and contributed to over a dozen books on Microsoft technologies. Follow Nick on Twitter @nickcavalancia and @Techvangelism.

ConversationalGeek

Visit conversationalgeek.com for more books on topics geeks love.