Description - IT infrastructure building depends on - what kind of availability is required - 99.9, 99.99 or 99.995 uptime? Netmagic’s Mandar Kulkarni shares insights into key questions on, what these numbers mean & the impact it has on the infrastructure, cost & the customer.

Keywords: High Availability of IT Infrastructure, Paradox of the 9s, IT Infrastructure, infrastructure support, article

The Paradox of the 9s - What High Availability actually means

Most IT Managers think that having more 9s is actually better - but what is to be considered is the cost associated with having more 9s. The question to be answered here is how many 9s are actually needed? Mandar Kulkarni (Senior Vice President - Solutions Engineering and Private Cloud Practice at Netmagic Solutions) shares insights and answers some key questions on High Availability of IT infrastructure.

Let's get down to brass tacks. What is high availability and what does it mean to a CIO? And how does a CIO look at HA?

High availability is the co-efficient of the uptime required to run your IT Infrastructure - simply put, for a CIO it translates into two things - peace of mind and growth for his business. If a CIO is assured of the availability (read uptime) of his business applications, he is sure of his business running optimally at all times. HA is a co-efficient of this assurance.

Typically, at the CIO level, IT is looked at in a tiered manner - within the purview of High Availability. A CIO looks at things from an application perspective. He would look at the entire...
IT into multiple tiers based on applications - infrastructure's priority depends on what resources do these applications run on - the backbone if you may! Broadly, there are 3 tiers - what applications fall under these tiers depends on what the business is and which industry does the organization belong to.

Tier 1 applications are those on which the core business runs. For example, for an ecommerce business, the portal, the e-shop or the online marketplace would be the tier 1 applications. Likewise, for a manufacturing business, it is the machine floor applications on which the production runs, for a bank it is the core banking, Internet banking and similar applications, and for a telco the billing software, etc. so on and so forth.

Tier 2 applications are those, which are critical but not what runs the core of the business. Examples of these would be the infrastructure support, mail and messaging applications, portal, etc.

Finally, tier 3 applications are what we call periphery ones such as file servers, and some of the good-to-have applications.

So when it comes to High Availability of infrastructures in particular, CIOs base high availability based on applications. So a CIO will put highest availability and performance ask for his tier 1 application, medium availability and performance for his tier 2 and least for tier 3 applications.

Now, the approach extends to a scenario of outsourcing equally. Typically tier 1 applications are what they put into a data center. And in most cases their tier 2 applications as well. But for tier 3 applications, they would not worry too much - they could run out of their own data centers / server rooms or local offices, or in some cases the cloud too.

Entire IT infrastructure building depends on one thing - what kind of availability is required - 99.9, 99.99 or 99.995 uptime? What do these numbers actually mean? And what impact does it have on the infrastructure, cost and the customer?

There is a fundamental change needed in what people think of availability. It is not negotiated or bought - it is not a commercial discussion. HA is designed and architected and not negotiated - just because one is able to pay a little more does not buy him high availability.

The table below highlights the acceptable downtime - from an availability perspective - for various 9s.

<table>
<thead>
<tr>
<th>SLA</th>
<th>Minutes of acceptable downtime per month</th>
<th>Minutes of downtime Annualized</th>
</tr>
</thead>
<tbody>
<tr>
<td>99%</td>
<td>432</td>
<td>5184</td>
</tr>
<tr>
<td>99.9%</td>
<td>43.2</td>
<td>518.4</td>
</tr>
<tr>
<td>99.99%</td>
<td>4.32</td>
<td>51.84</td>
</tr>
<tr>
<td>99.995%</td>
<td>0.21</td>
<td>25.92</td>
</tr>
</tbody>
</table>

There is a standard approach to go from 99 to 99.9 to 99.99.

Now, moving up the value chain from 99 to 99.9 and so on impacts multiple aspects of the infrastructure. There are 2 key impacts when you add a 9 to the availability figure. One is that obviously the availability goes high - availability commitment goes high - this is the visible part of the impact.

The invisible impact is that the infrastructure architecture goes through a change. When you add a 9 to availability, you are adding a level of redundancy to the infrastructure - resistance of the infrastructure increases, and so does cost. So infrastructure should be built keeping in mind the availability that is needed - and not the other way around.

How are various levels of High Availability achieved and what is the impact on the architecture of the underlying infrastructure and cost?

Explaining High Availability architecture for infrastructure is a complex process - especially when you go up the value chain.

Achieving 99% Availability:

Broadly and briefly, when you want to achieve 99% availability, it essentially means the infrastructure is allowed 432 minutes of downtime each month, which is roughly little over 7 hours.

This essentially means that you can easily recover from a complete failure - if you look at most of professionally run IT setup the recovery time is roughly 2-4 hours depending on what kind of IT operations are run. What this means is that you are allowed to create single points of failure in your architecture.

Look at the diagram 1.2(a) below. All components are reliable and high performing, but are not redundant.
Diagram 1.2 (a): Setup for 99% availability

For example if you look at the power factor - there is a generator and an UPS that increases the reliability of the infrastructure because you are not at the mercy of the local power supply. So if there is a power failure, your data center is not going to go down. UPS gives certain levels of resiliency to the infrastructure and the generator adds further more. But there is no redundancy built into the system.

Here you are allowed to have a single point of failure - at the power, IT systems level, bandwidth or air conditioning of the DC levels - and you have enough time to recover from the failure. Diagrams 1.2 (b), (c) and (d) show the outage of the architecture built for 99% availability.

Even though new and disruptive technologies have changed the technology landscape, organizations continue to look for ways to reduce the percentage of their IT budget dedicated to running infrastructure and maintaining “status quo”. Data Center (DC) takes up the maximum of this maintenance cost.

Achieving 99.9% Availability:

When moving from 99% to 99.9% means that the acceptable downtime is only 43.2 minutes per month. Here, the target recovery time as mentioned above is more than this acceptable downtime. What this means is that if you have a severity 1, i.e. end user outage, you are invariably breaching the SLA of 99.9% - basically you cannot afford to have an outage.

So how does one avoid outage? Isolate all single points of failure and create redundancies. As shown in diagram 1.3(a), there are 2 ISPs, power inputs, servers and network setup (clustered server and network device setup), and redundant aircon, so on and so forth.

Diagram 1.3 (a): Setup for 99.9% availability

In this architecture, failure at any single point does not result in end user outage. Failure at any single point also allows ample time to switch to the redundancy that is build into the architecture. The 43.2 minutes is enough to only shift to the redundant device or setup - and not recover from a failure.

Impact on the cost moving from 99% to 99.9%:

The move from 99% to 99.9%, though it is only a single 9 addition to the availability matrix, significantly increases the cost. It almost doubles the cost of the setup - about 100% incremental cost over the infra setup of 99% availability.

Achieving 99.99% Availability:

99.99% of availability means 4.32 minutes of outage is acceptable per month. This essentially means that one does not even have time to move workloads or shift to redundant infra pieces.

So how does one achieve 99.99% availability? The
infrastructure architecture for achieving 99.99% availability is to have a combination of 99% and 99.9% infra architectures.

A high available architecture as shown in diagram 1.3 (a) of 99% availability with an active-passive disaster recovery site, which looks like diagram 1.2 (a) of 99% availability - see figure 1.4 (a). It is a fault tolerant architecture that can sustain 2 system failures in parallel. The need here is to have replication and DR management tools in place like Sanovi / SRM to invoke DR in case of multiple system failure or site level disaster.

Diagram 1.4 (a): Setup for 99.99% availability

Impact on the cost moving from 99.9% to 99.99%:

Cost impact of moving from 99.9% to 99.99% availability infrastructure is incremental of about 60-70% more. It involves a cost of adding a DR site - which is half of the cost of primary site. Also there is an additional cost of the overlay for replication, bandwidth and DR management.

Achieving 99.995% Availability:

Achieving 99.995% availability at the infrastructure level is a different ball game altogether. While it means adding an automation and virtualization layer over the 99.99% architecture, the implications and complexity involved are quite different.
The complexity is not just limited to having a global load balancer at the primary site, but applications will have to support this architecture - there is a need to architect the infrastructure as well as the application layer to achieve the 99.995% availability. See diagram 1.5 (a)

Diagram 1.5 (a): Setup for 99.995% availability

Applications should have the ability also to be able to run out of both sites and seamlessly so. Users are not directly accessing the application but to an abstraction layer, which then connects to the application layer - this is the only way the two sites can be seamlessly interchanged without impacting the users. Typically a global load balancer is used to achieve this.
Impact on the cost moving from 99.99% to 99.995%:

There is only a 10% additional cost for the abstraction layer and virtualization. But impact on the infrastructure here is not just in terms of cost - there is the management of complexity and architecting the application layer that has to be taken into consideration here.

In conclusion, what kind of companies uses the various availability requirements?

99.995% and 99.99% availability
- Financial Systems with very high transaction volumes - millions of transactions per minutes - stock markets, core banking systems, billing applications at telcos, etc.
- Organizations with mandate of regulatory requirement for this level of resilience
- In organizations doing business transactions where a single minute of outage causes significant financial losses

99.9% availability
- Most tier 1 applications run by companies - those running 24x7 operations
- eCommerce websites, SAP environments, Internet Banking websites, international companies that are accessed 24x7, etc.
About Netmagic Solutions (An NTT Communications Company)

Netmagic, an NTT Communications company, is India’s leading Managed Hosting and Cloud Service Provider, with 8 carrier-neutral, state-of-the-art datacenters across the country. Established in 1998, Netmagic has been a pioneer in the Indian IT Infrastructure services space as it was the first to launch services such as cloud computing, managed security, Disaster Recovery-as-a-Service and software-defined storage. An IT Infrastructure Partner to more than 1400 enterprises globally, Netmagic, also delivers Remote Infrastructure Management services to NTT Communications’ customers across Americas, Europe and Asia-Pacific region. Netmagic is the recipient of several industry accolades and was recently chosen by India’s CIO community as the best service provider of Datacenter Managed Services and Cloud services, at the CIO Choice 2015 Awards. NTT Communications has subsidiaries and offices in 43 countries / regions, with over 21,600 employees worldwide. It had total operating revenues of 1230 billion JPY for year ending March 31, 2014 and has infrastructure worldwide, including leading global tier-1 IP network, the Arcstar Universal One™ VPN network reaching 196 countries/regions, and 130 secure datacenters.

NTT Communications is a part of NTT Corporation, Japan – one of world’s largest telecommunications companies and ranked 53 on Fortune Global 500 list (2014).

For further details please log into www.netmagicsolutions.com | Twitter: @netmagic | LinkedIn: @Netmagic Solutions

About NTT Communications Corporation

NTT Communications provides consultancy, architecture, security and cloud services to optimize the information and communications technology (ICT) environments of enterprises. These offerings are backed by the company’s worldwide infrastructure, including leading global tier-1 IP network, Arcstar Universal One™ VPN network reaching 196 countries/regions, and over 150 secure data centers. NTT Communications’ solutions leverage the global resources of NTT Group companies including Dimension Data, NTT DOCOMO and NTT DATA.