*Introducing the new IT availability metric*
*For the Digital Age……*

# Mean Time Between Fiasco

*Why are organisations experiencing long duration, high profile outages despite investing in HA & DR solutions?*

Author: Ian MacDonald FBCS CITP FSM
Edenfield IT Consulting Limited
Date: April 2018

**Section** **Page**

## 1. Executive Summary

There is now an increasing reliance by the business on their IT services to enable the business to operate successfully with a growing dependency on their IT services to manage business risk and protect their marketplace and brand reputation.
Increasing customer expectation and demand for convenient, flexible services together with the ongoing digital transformation of all business processes continues to push the business requirements for High Availability (HA) IT services and Disaster Recovery (DR) solutions.

*In the new digital world, increasingly when IT stops the business stops.*

Investment in HA and DR is now an essential business imperative to avoid outages and provide the ability to recover IT services in a controlled manner in the event of a major failure.

However, despite this, major brand companies increasingly are attracting adverse media coverage for IT issues that are impacting customers for several hours and sometimes days. The business consequences from these outages are often huge in terms of financial loss, marketplace reputation and erosion of customer confidence in the brand longer term.

When reading the media reports relating to these news worthy high profile IT outages then journalists often use words such as:- *failure, disaster, catastrophe*. When you read the Twitter comments from frustrated and angry customers the lexicon often consists of more pointed commentary to express customer feeling: *debacle, shambles, farce, mess, car crash.*

All these words are synonyms for the word **FIASCO.**

Duration is a key factor in how customers, the media and regulators will view IT failure. The longer the outage the greater the customer frustration and anger and the more likely the term 'Fiasco' and its synonyms will be applied by customers and the media.

So, why are the benefits of HA design and DR solutions failing to protect these businesses from experiencing these high profile outages?

In this paper we have undertaken research and analysis to gain insight and understanding on the risks and organisational factors that can result in unplanned outages with an excessive duration. The conclusion being that investment in HA technologies and Disaster Recovery solutions cannot guarantee high availability or fast recovery of IT services.

This research (supported by a number of IT industry surveys) has resulted in some interesting findings and observations on why the benefits from investment in HA technologies and DR solutions are not being fully realised, these include:-

- Five out of 10 organisations rank human error as the top cause of unplanned downtime
- IT and the airline industry share common ground – despite resilience and automation human error is the highest cause of aircraft failure so is there a link?
- Human error, Security and Data corruption are top risks that HA and DR cannot mitigate
- Complex IT failures are occurring where no defined recovery option exists
- There is a lack of confidence in invoking DR/Failover procedures
- 40% of organisations survey admitted to having no documented DR plan
- RTO targets are not being consistently achieved by the DR/Failover solutions

As a result of these findings this paper provides some recommendations to mitigate impact and improve recovery.

Key Message:

*In the new digital age, the ability to recover quickly and confidently is essential if you want to avoid 'Mean Time between Fiasco' being your new IT availability metric.*

# 2 The Increasing Business Reliance & Dependency on IT Service

There is now an increasing reliance by the business on their IT services to enable the business to operate successfully to achieve their business outcomes. Equally, the business now has an increasing dependency on their IT services to manage risk and protect their marketplace and brand reputation.

## 2.1 Business Reliance

Today's business more than ever looks to IT to provide the IT solutions and services that allow them to deliver on their business strategy and achieve the business outcomes necessary to be successful.

Continued investment and innovation within the IT industry has provided the technologies, products and services that have enabled the business to respond to their marketplace challenges, these include:-

- Providing customers with flexibility and choice on how and when they transact (Telephony, Online, Mobile)
- Responding to changing regulatory requirements focused on the consumer, i.e. ease of switching Banks and Utility providers, detailed and accurate billing, same day and immediate payments.
- Improving the 'ease of doing business' by digitalising and automating all aspects of the customer journey to create a positive customer experience.
- Increase revenues from existing customers by analysing patterns and behaviours to provide targeted propositions for additional products and services.

### 2.1.1 Customer Facing Services

Customer facing services, particularly those that are promoted as '24x7' are now critical to customer satisfaction. Not unreasonably, customer expectation is that these services are available when required.

IT failures now directly impact the end customer and many businesses are unable to respond and/or cope to customer requests via their traditional channels, i.e. Branch or Telephone when online services are unavailable.

IT related issues are now accounting for an increasing proportion of formal customer complaints to the business. Retaining customers and attracting new customers may become an issue where there is a negative perception of the organisations services.

### 2.1.2 Business Processes

Digital transformation is now being actively embraced by many organisations as they look to improve the efficiency and effectiveness of all their business processes, (customer facing and back office).

As business processes become fully digitalised then the business dependency on IT is now absolute as manual ways of working become obsolete.

In the 'old days' when IT was down, most businesses could handle unplanned outages as some business processes were not fully reliant on IT and/or that there were manual workaround options, e.g. if the ATM service was down customers could get a limited cash advance within a branch.

**"If we examine IT from a value perspective, then the value of IT to the business is often the total value of the business. This is because many modern businesses would be unable to function without IT"**

Source: ITNOW journal Autumn 2017 (Article by John Mitchell)

In the fully digitalised world we are moving towards, then the situation is now very much evident …….. "*When IT stops the Business stops".*

### 2.2 Business Dependency

The increasing pervasiveness of IT across all critical business processes creates a significant business dependency on IT services with increased business risk in the event of failure.

For many organisations *'When 'IT stops the Business stops'* and this has a wider implication than just the denial of service to customers or halting back office processing. From a corporate perspective the loss of IT services can significantly impact:-

- Costs
- Marketplace reputation
- Brand

Hence, there is a dependency on high availability IT services with fast recovery to manage the corporate organisations level of risk and exposure from the consequences of unplanned downtime to their critical services.

#### 2.2.1 Costs

There's an ITIL saying 'Availability costs, but unavailability isn't for free either'. Time is money, so even an outage of a short duration will have a cost. The more prolonged the outage the greater the costs incurred.

The costs to the business from an outage can be attributed to:-

**Direct costs:-** This would include for example the total revenue loss from customers because of their inability to access core systems during the outage period.

**Indirect costs:** This would include for example the additional economic loss of the outage, including reputational damages (fines, penalties) , loss of company valuation from a drop in share price, customer compensation and potential loss of existing customers.

**Opportunity costs:-** This would include for example the potential loss of revenue from potential new customers and lost business opportunities not only during the outage period but subsequently due to negative media coverage and marketplace reputation.


**British Airways**
A total loss of IT systems in May 2017 grounded flights worldwide and stranded thousands of passengers across a 3 day period.

The Chief Executive reported that this will cost the company £80m

#### 2.2.2 Marketplace Reputation

The marketplace reputation of a business is essential to its long term success. The consequences from major outages to critical services can quickly erode the trust and confidence of their customers.

Organisations often compound the impact of a major failure by a lack of timely and accurate information to inform customers of what's happened, what's being done and what their customers can expect and when. Poor marketplace reputation is often created not from the outage in itself but how well the organisation manages the failure from the perspective of the customer.

Today there is a greater focus across industry sectors on the need for businesses to pay customers compensation for any costs and disruption caused.

For regulated businesses, the impact from IT failures on service and the customer can result in regulatory breaches which can result in fines and penalties levied by the regulatory bodies. An extreme could be the withdrawal of their licence to trade


**Royal Bank of Scotland**
In 2012, RBS, NatWest and Ulster Bank customers were affected for weeks following problems with a software upgrade and inability to recover quickly.

The Financial Conduct Authority (FCA) fined RBS £42m and the Prudent Regulation Authority (PCA) fined the bank £14m – a total of £56m

### 2.2.3 Brand

Brand image is developed over time through advertising campaigns with a consistent theme, and is authenticated through the customer's direct experience.

Brand image now has a direct correlation with the availability and reliability of the organisations IT services as these have a primary role in providing a positive customer experience.

Brand-building is hard, time-consuming work and can quickly become tarnished by IT failures and the consequences these create for customers. With the advent of social media, service failure can become widespread public knowledge, very quickly creating negative media led damage to the company brand image.

In the competitive marketplace that most organisations operate, erosion of brand image can significantly impact customer loyalty with existing and prospective customers moving to alternative products and services. This impacts market share and profitability. In some extreme cases the brand may never recover.


**RIM (Blackberry)**
In 2011, technical issues with their Blackberry service impacted users across the globe for 4 days before a stable service was restored.

Blackberry experienced a four day global outage due to a server failure in the UK. The problem spread "like wildfire" to Africa, the Middle East before hitting the Americas and Asia.

The problem brought all email, BBM and web browsing services to a halt, Blackberry had pioneered wire-less email communications and for many years were the dominant market leader. Blackberry was the corporate organisations mobile device of choice. This outage therefore impacted high value corporate customers as well as individual users.

During the outage RIM was heavily criticised for not communicating and disgruntled customers turned to social media networks and Twitter to voice their frustration. After a 3 day period of silence, RIM responded by stepping up its communication efforts and an offer of free apps. However, by then it was almost impossible for RIM to get ahead of the crisis. The outage was bad enough, but the real damage was done by RIM's handling of the affair.

At the time the crisis caused a 3.5% drop of RIMs share price on the Toronto Stock Exchange and a 2% drop on the Nasdaq. Forecast lawsuits in the US and Canada were estimated to cost RIM in the region of £16M.

In reality the Blackberry brand was already in trouble with Apple and Android smartphones now competitors for market share. However, this outage is cited as a key factor in many corporate organisations moving away from Blackberry and the brand demise continued.

In 2007 Blackberry market share of the smartphone market was 20% by the end of 2016 it was 0%.

# 3 High Availability is now a Business Imperative

In responding to today's competitive marketplace, the business customer will look to their IT service provider to deliver high quality IT services that enable them to consistently achieve their business outcomes and help drive their business forward.

Today's IT infrastructure services are now very much an 'enabler' and critical to most organisations business success.

For business critical services, High Availability solutions are now an essential business imperative as they directly influence:-

- Customer satisfaction (Customer experience)
- Ease of doing business (The customer journey)
- Customer Outcome (Customer fulfilment, Sales, Orders, Income generation)
- Brand reputation (Marketplace, Media, Regulatory)

## 3.1 High Availability

'High Availability' (HA) is typically a solution understood to be the ability to achieve an agreed level of operational performance or 'up time' for a prolonged period of time.

**ITIL Definition - High Availability**

*High Availability solutions are designed to achieve an agreed level of Availability and make use of techniques such as Fault Tolerance, Resilience and fast Recovery to reduce the number of Incidents and the Impact of Incidents.*

The above definition quite rightly infers a strong focus on IT service design to avoid or minimise the impact of ***unplanned*** IT interruptions to service which are highly disruptive to the business and their end customers.

However, in reality for most organisations aiming to deliver 'around the clock' customer facing services the biggest single cause of IT downtime is in fact for ***planned*** outages.

This is becoming a growing issue in the area of 24x7 service operation. Planned outages are required to allow essential maintenance and technology refresh activities to be performed for the varied technology components and hosted applications that underpin the IT service.

Service Design will therefore need an approach that not only avoids or minimises the impact from unplanned IT interruptions but can enable maintenance activity to be performed without impacting the availability of IT services.

## 3.2 High Availability – Characteristics

The following additional definitions help to better define the term 'High Availability' and helps determine which aspects of service design are most important to the business.

**HA = CA + CO**

- **High Availability (HA)**: A characteristic of the design that enables the IT service to run continuously without interruption for a prolonged period of time.

- **Continuous Operation (CO)**: A characteristic of the design to eliminate planned downtime of an IT service for routine changes. *NB that individual components or CIs may be down even though the IT service remains available.*

- **Continuous Availability (CA):** A characteristic of the design to eliminate or mask unplanned downtime to users of the IT service.

### 3.3 High Availability Design Considerations

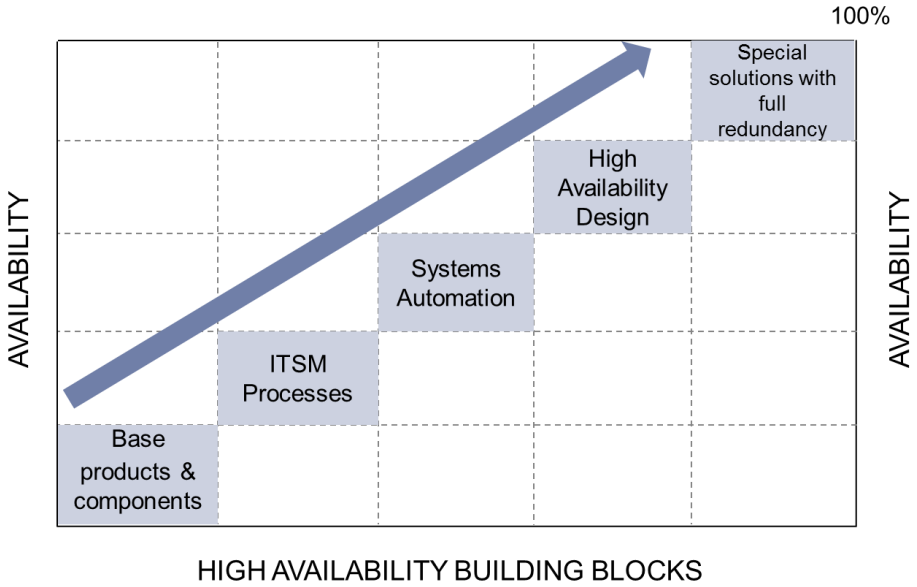

**Book:** ITIL Service Design (Section 4.4 Availability management)

**Author**: Axelos

**ISBN**: 978 0 11 331047 0

When determining how stringent availability requirements from the business can be met there are a number of considerations to be made when approaching the Service Design.

Typically, high availability cannot be delivered on a consistent basis by the base technologies alone. The following illustration highlights the additional capability's (people, process, tools) and HA solutions necessary to achieve business requirements for high availability, reliable IT services.




The level of availability required by the business influences the overall cost of the IT service provided and complexity of the end-end service design. In general, the higher the level of availability required by the business the greater the cost.

### 3.4 High Availability Design Principles

There are three essential service design principles that should underpin all IT service designs where the business states a requirement for high availability. These are:-

#### 3.4.1 Eliminate Single Points of Failure (N+1) [1]

This means adding full redundancy for all critical IT components so that failure of a component does not mean failure of the entire IT service. Redundancy of components also allows for planned maintenance without interruption to IT services.[2]

Alternative and diverse routing within the network design provides resilience to maintain network connectivity from failures within the WAN and LAN networks.

[1] **N+1** redundancy is a form of resilience that ensures system availability in the event of component failure. Components (**N**) have at least **one** independent backup component (+**1**). This is also essential for critical facilities hosting IT systems and functions, i.e. N+1 design of Mechanical + Electrical (M&E) equipment ensures continuous availability of power and cooling services essential for the Data Centre

[2] It also ensures legal requirements for electrical maintenance can be performed without interruption to hosted services by taking offline M&E equipment and utilising the remaining configuration to provide power and cooling

### 3.4.2 Provide Disaster Recovery (DR)

The combination of increasing levels of 'fault tolerance' within software and hardware components and provision of resilience across the End-End infrastructure configuration will avoid or minimise the impact of routine failures.

However, there will be failure scenarios where IT services may need to switch to an alternate configuration, e.g. in response to a catastrophic power failure at the Data Centre.

High Availability design needs to ensure that services can be switched (or failed over) to an alternate configuration quickly and securely.

'Fail over' will typically occur across geographic locations. Data replication and fail over solutions allow IT services to be restored quickly with no loss of business and customer data.

### 3.4.3 Detect Errors

It is important that error conditions are captured and actioned particularly where failures have occurred that have been successfully handled by either fault tolerant features or where resilience has been automatically invoked.

Systems Automation should monitor the End-End infrastructure to detect events (change in status, threshold breached etc.) to provide early warning and/or using predefined responses instigate automated escalation, e g automatically dial out to the vendor's support systems to log a hardware fault for an engineer to attend and replace a failed hardware element.

## 3.5 The Class of Nines

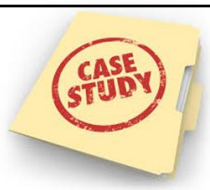
Availability is usually expressed as a percentage of uptime.

In recent years, percentages of a particular order of magnitude are sometimes referred to by the number of nines or "class of nines", e.g. Five nines (99.999%)

This is a common practice by hardware vendors when marketing the availability of their HA technologies. Increasingly IT organisations are now starting to use the 'nines' expression as their (sometimes aspirational) HA target.

**The Class of Nines**

1 'Nine' = 90%
2 'Nines' = 99%
3 'Nines' = 99.9%
4 'Nines' = 99.99%
5 'Nines' = 99.999%




A large UK Financial Services organisation initiated a strategic programme to improve IT service availability. The programme was named the **'Five Nines Programme'**

The programme goal was not to achieve availability levels of 99.999% but to focus the mind across IT on what prevents this target ever being realistic. All outages (Planned and Unplanned) across a 12 month period were analysed with a view to:-

- Could this outage have been avoided?
- Can the duration of this outage be reduced?
- Could this planned outage to apply change have been made dynamically?

A wide range of improvements were delivered that led to higher availability, these included:-

- CO improvements to enable planned changes to made dynamically via technology exploitation
- Reduced the duration of restart and recovery for all critical online transaction processing components
- Established a 24x7 service recovery onsite team to manage all high impact incidents
- Systems automation exploitation to improve event management and overall service health monitoring

Interestingly the programme ended a few years ago but 'Five Nines' remains in the IT lexicon and the thinking still prevails in the IT culture

# 4 High Availability & Disaster Recovery – Situation Appraisal

This Increasing business reliance and dependency on IT services together with ever increasing customer expectation and demand for convenient and flexible services continues to push the business requirements for High Availability (HA) IT services and Disaster Recovery (DR) solutions that enable a controlled recovery from major failures.

Yet, as a customer you probably have experienced inconvenience when a service you wished to use was unavailable and with the growth in social media we are often made aware of major brand companies experiencing IT issues. This would appear to be an interesting paradox worthy of some additional research.

## 4.1 An Industry Perspective

- **The ITIC[3] 2017 Global Reliability Survey**

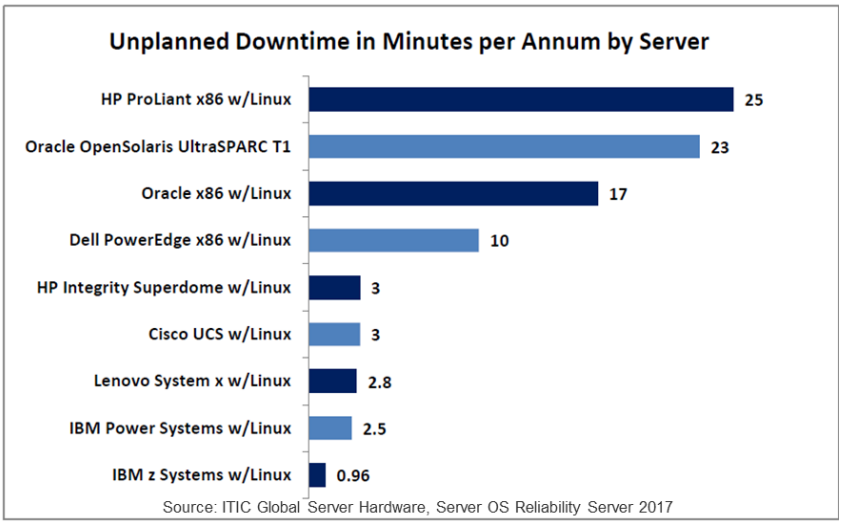
This survey is focused on the availability and reliability of HA server technologies and was completed by 750 organisations worldwide. The following results confirm the increasing trend to aim for high levels of availability and reliability (For comparison the 2008 results are provided in RED)**:-**

- 79% of organisations now require 99.99% (Four Nines) availability of their enterprise servers (2008 = 23%)
- Additionally, 18% of respondents indicated their organisations now require 99.999% (Five Nines) server and operating system uptime (2008 = 7%)
- 0% of survey respondents indicated their organizations could live with just 99% (Two Nines) uptime (2008 = 27%)
- Only 1% said their organisations required just 99.9% (Three Nines) availability (2008 = 40%)

| Availability % | Downtime per year | Downtime per month |
|---|---|---|
| 90% ("one nine") | 36.5 days | 72 hours |
| 99% ("two nines") | 3.65 days | 7.20 hours |
| 99.9% ("three nines") | 8.76 hours | 43.8 minutes |
| 99.99% ("four nines") | 52.56 minutes | 4.38 minutes |
| 99.999% ("five nines") | 5.26 minutes | 25.9 seconds |

The survey also provides information and comparisons on the levels of availability and reliability Server technologies are achieving.

Based on the amount of unplanned downtime incurred per annum this research indicates that today's modern Server technologies (Hardware and Operating Systems) are delivering Four Nines (99.99%) availability. A number of these Server technologies are achieving Five Nines (99.999%) availability.




[3] Information Technology Intelligence Consulting (ITIC) is a research and consulting firm based in Boston, USA. It provides primary research on a wide variety of technology topics for vendors and enterprises.

Whilst the previous research clearly demonstrates that the underpinning server technologies are today capable of delivering Four Nines and Five Nines availability, these of course are only one of many different technologies, products, services and hosted applications that comprise the end-end infrastructure.

Therefore the customer experience is not just reliant on the availability and reliability of the server technologies. To get an industry perspective on whether organisations are consistently achieving their HA availability goals then a good source would be surveys focused on service continuity.

- **IT Service Continuity Surveys**

Research across a number of industry surveys provided some interesting insights on how successful HA technologies and design are:-

- 54% of respondents said they had experienced an outage of >8hrs in the last 5 years[4]
- Only 37% of respondents said they meet their availability targets consistently[5]
- 71% of respondents had experienced an unplanned outage in the previous 12 months[6]
- 58% of respondents experienced issues and delays when a failover was required[7]
- Only 39% of respondents meet their 'fail over' Recovery Time Objective (RTO) consistently


This research provides a balance to the ITIC report and indicates that for a number of organisations, achieving their HA and DR targets consistently can be a challenge.

To further explore this and gain additional insight into how effective HA design is in enabling organisations avoid outages then a good source will be the various media outlets.
.

### 4.2 Media Perspective

#### 4.2.1 Social Media

Today with the growing emergence of social media there is a clear shift in customer behaviour to not accept poor service and an indifferent customer experience and 'share' this using social media channels such as Facebook and Twitter.

This can very quickly create momentum with more and more disgruntled customers adding their experience and frustrations.

The media can pick up and report significant service issues on their online media channels. Customers may also escalate their frustrations to media companies to 'complain'.

IT issues of a prolonged nature can also appear on news reports across TV and radio and in the national press.


[4] State of Disaster Recovery 2016 by Zetta Infographic (403 respondents)
[5] 2017 Disaster Recovery Survey by CloudEndure (270 respondents)
[6] 2017 Disaster Recovery Survey by CloudEndure (270 respondents)
[7] The State of OT Disaster Recovery amongst UK businesses 2016 by iland (250 respondents)


This was extracted from the BBC news website based on technical problems encountered by a recognised UK High Street bank in authorising their customers POS and ATM transactions in February 2017. *(Customer and Bank names removed)*

Nine-months pregnant, Ewa xxxxxxxxx was shopping in Aldi when her card was declined at the till.
The 26-year-old said she was spending the day with her daughter, before she is induced on Tuesday.
"I logged into my account to check and I do have enough money. I tried one more time, but it was declined again. I left Aldi so embarrassed.
"Instead of spending time in the park, going for lunch, buying a costume for Thursday's World Book Day at school, my daughter [and I] went back home.
"Our day was ruined, because I have no cash on me. That's the story of one very sad and disappointed xxxxxxx Bank customer."

Another customer, Chris xxxxxxx, who lives in Camberley, Surrey, said he was worried he would not be able to pay his council tax bill after his card was declined in a local Sainsbury's.
He said he was unable to buy food, fuel and a costume for his daughter for World Book Day.
"I can't do any of those things now. I also want to pay my council tax bill online and I'm worried that I will be fined if I can't pay it. xxxxxxxx cannot tell me whether this will last hours or days. I don't think xxxxxx do enough to keep their customers informed."

### 4.2.2 Media Coverage

Without needing to do any research, there are two high profile IT failures in 2017 that most people would easily recall due to the amount of adverse media coverage that they incurred over several days in May, namely the British Airways loss of all IT services and the WannaCry ransomware attack that had widespread impact to NHS Hospitals, Trusts and GP surgeries.

By performing some random Internet searches we can see from the following table that these high profile outages are not an aberration. Many organisations with a strong brand and marketplace reputation have incurred prolonged and painful IT outages resulting in adverse media coverage on the last 24 months. Given the nature of their business it is a reasonable assumption they have invested in HA and DR solutions. (NB this list is not exhaustive simply a listing of the initial results gained from Internet searches and restricted to 20)

| Company Name | Industry | When | Duration of customer impact |
|---|---|---|---|
| NHS (Wales) | Healthcare | 2018 | 7 hrs |
| British Airways | Airline | 2017 | 3 days |
| NHS | Healthcare | 2017 | 3 days |
| Microsoft Azure | Cloud Service Provider | 2017 | 7 hrs |
| Amadeus | Airline (Reservations) | 2017 | 4 hrs |
| Amazon Web Services | Cloud Service Provider | 2017 | 4 hrs |
| IBM | Cloud Service Provider | 2017 | 36 hrs |
| Barclays | Finance | 2017 | 7 hrs |
| Gitlab | SaaS (App development) | 2017 | 18 hrs |
| CD Baby | Music (online distribution) | 2017 | 4 days |
| Microsoft Skype | SaaS (Collaboration) | 2017 | 24 hrs |
| ASOS | Retail (Clothing) | 2016 | 20 hrs |
| HSBC | Finance | 2016 | 2 days |
| Delta Airlines | Airline | 2016 | 3 days |
| Southwest Airlines | Airline | 2016 | 4 days |
| SSP Worldwide | SaaS (Insurance) | 2016 | 10 days |
| Sainsburys | Retail (Grocery) | 2016 | 2 days |
| Salesforce | SaaS (CRM) | 2016 | 12 hrs |
| RBS, Natwest and Ulster Bank | Finance | 2016 | 8 hrs |
| ASDA | Retail (Grocery) | 2016 | 5 hrs |

# 5 So what's going wrong? (With HA & DR)

Based on the research so far it's a reasoned hypothesis to state that: ***Investment in HA technologies and Disaster Recovery solutions does not guarantee high availability or fast recovery of IT services.***

But why is this?

## 5.1 Causes of Service Interruption

When researching industry reports and surveys on the topics of high availability and disaster recovery, the top issues that are being consistently being reported as the cause of service interruption are:-

- Human Error
- Security (Denial of Service, Malware, Viruses)
- Data Errors and Corruption

So why doesn't HA or DR offer mitigation to these types of service risk?

When researching for some specific examples of the above it becomes clear that that these are risks that HA and DR solutions cannot prevent or provide an immediate recovery option.


***Investment in HA technologies and Disaster Recovery solutions does not guarantee high availability or fast recovery of IT services.***

- **Human Error**

***Company:- GitLab***


In the effort to fix a slowdown on the site, a system administrator accidentally typed the command to delete the primary database. GitLab had to restore a 6-hour-old backup that meant any data created in that six-hour window may have been permanently lost.

**Observation**

HA offers no mitigation to this and most human errors and is more related to the controls that in place to restrict and provide controlled access to system resources, commands etc.

In this example even if the database was being mirrored or replicated to an alternate site, once the database is deleted this is immediately reflected on the alternate site.

The example also infers standard operating procedures may not be able to recover the database back to the to the point of failure

- **Security**

**Company:- NHS**


The WannaCry ransomware attack had widespread impact to NHS Hospitals, Trusts and GP surgeries. It also impacted many businesses across the world with recovery activity taking place over a couple of days.

**Observation**

HA offers no direct mitigation to security threats which rely on the combination of IT security tools for proactive detection and prevention of threats and the IT patching policy to ensure identified vulnerabilities are eliminated. In this attack, organisations impacted had not patched their estate for a previously known and reported vulnerability.

The role of DR in this scenario is also interesting, in this example even if the server estate was being mirrored or replicated to an alternate site the vulnerability would of course still exist and the affected data changed. If the DR solution provides the option to perform point in time recovery to restore servers and application data to a point prior to the attack, to then enable patching. Then this is exactly the same recovery approach required at the primary site and so doesn't offer a quicker recovery with the additional risk of DR specific invocation issues compounding the recovery approach and duration of business impact.

- **Data errors & corruption**

***Company:- Co-operative Group Example***


Due to a processing error, customers who shopped at a Co-op store or used a petrol filling station using a credit or debit card were charged twice due to a processing error. Hundreds of thousands of people were affected across the Co-op's 2,800 stores and 200 petrol stations in the UK.

**Observation**

HA offers no mitigation to incorrect updates being applied to databases or transaction files and again relates to the controls in place for the running of batch. In this example even if the database was being mirrored or replicated to an alternate site, once the erroneous updates are applied this is immediately reflected on the alternate site. This scenario would require a programmatic solution to reverse out the duplicate payments and credit affected customers accounts.

### 5.2 Conclusion

The above I believe begins to provide some insight into why despite significant investment in HA you can still experience excessive periods of unplanned downtime.

HA technologies provide protection from anticipated routine failures through the combinations of fault tolerance, resilience and system automation. However, these do not provide protection from a range of high impact risks such as human errors, application bugs, data processing errors and security issues.

DR will provide the ability to reinstate services at an alternate location in response to a catastrophic failure either within the IT infrastructure or loss of Data Centre (DC) services.

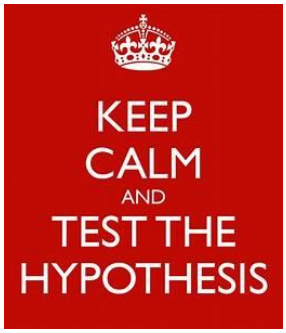
However, with the growing use of data replication and mirroring many 'logical' errors affecting data integrity are immediately reflected at your alternate DR location therefore ruling out DR invocation as an option.

This means DR is perhaps viewed as a solution best suited in responding to the 'physical' impacts to IT services, i.e. multiple hardware failures, power outages, damage to the DC (fire, flood, external collision damage).

However, research indicates that DR invocations for these conditions are often avoided, delayed or take far longer than expected.

So are there other factors that are limiting the benefits of HA and DR? We research and attempt to validate a number of hypotheses that may indicate other potential causes.

# 6 Why are we experiencing long duration outages with HA & DR?




In section 3, we were able to research and provide evidence (20 high profile IT outages) to validate and support the view that investment in HA technologies and Disaster Recovery solutions does not guarantee high availability or fast recovery of IT services.

We have already identified in the previous section a number of scenario's where HA provides no mitigation in preventing an unplanned service outage and where data considerations eliminate DR/Failover from being a viable recovery option.

So whilst there are some specific types of IT failure that may render HA & DR as incapable of preventing an outage of potentially long duration, are there other factors that are preventing better exploitation of your HA technologies and DR solutions?

To explore this further, we have defined four hypotheses (or statements) against which we will undertake research to see if they can be validated and if so provide further insight and understanding of the pitfalls with HA & DR solutions.

## 6.1 Hypotheses

The following hypotheses will be used to provide the basis for additional research and analysis to try and assess the potential reasons why despite HA & DR solutions being implemented, failures within the IT infrastructure can still lead to prolonged and problematic service outages:-

- (Hypothesis 1) - ***Low frequency of failure + increasing dependency on automation is increasing the risk of Human Error***
- (Hypothesis 2) - ***There are recovery scenario's being encountered that standard operating procedures are unable to resolve.***
- (Hypothesis 3) - ***There is limited confidence in the ability to successfully invoke DR/Failover procedures***
- (Hypothesis 4) - ***Recovery Time Objectives (RTO) are not being achieved and contribute to delayed recovery.***

The results of the research and analysis for the above hypotheses are detailed in the following section of this paper.

# 7 (Hypothesis 1) – Low frequency of failure + increasing dependency on automation is increasing the risk of Human Error

**Five out of 10 Enterprises Rank Human Error as the Top Cause of Downtime**

Source: The ITIC 2017 Global Reliability Survey

## 7.1 Potential Consequences

Human error is increasingly being cited as one of the top reasons for unplanned downtime and protracted recovery in all the surveys referenced in this whitepaper.

The paradox is that HA design and systems automation are put in place to avoid human intervention to provide fast and predetermined recovery actions.

For this hypothesis we are looking for human error that occurs when dealing with an event that is not covered by HA design and systems automation or when automation itself fails.

Human error in the context of service delivery can manifest itself in many ways, these are illustrative with the associated consequences:-

- Presented with an error condition, the wrong interpretation and subsequent response creates a major outage which is disproportionate to the impact from the original error condition.
- The implications from a reported event are not immediately recognised or understood and the opportunity to prevent or minimise customer impact is missed.
- An error or omission is made during a recovery process which invalidates the recovery outcome and the recovery sequence has to recommence, further compounding the impact.
- Controls and warnings are ignored and data critical to recovery is erroneously deleted. Standard recovery procedures may now no longer effective.


## 7.2 Possible Causes

In this hypothesis we are attempting to link human error as a consequence of High Availability and systems automation on the basis that it creates an IT environment characterised by the following:-

- Dealing with routine IT failure is a rare exception rather than the norm
- A reliance and trust that system automation will detect, alert and correct all exceptions
- Automation is critical as manual procedures are no longer remembered or documented
- Decision making is no longer influenced by prior experiences, i.e. “Not seen this error before”

The impact of the above on IT personnel is that over time they lose the insight, knowledge and skills associated with event correlation, diagnostic assessment, problem solving and performing regular IT recovery. This lack of familiarity and practice is where human error can occur when dealing with IT failure scenario’s not previously experienced and practiced.

### 7.3 Findings & Observation (using the Airline Industry as a reference)

To help validate this hypothesis I could have called on my 40 years' experience to provide anecdotal examples of how the emergence of highly reliable infrastructure and sophisticated automation systems has over time increased the risk of human error when handling IT failures.

However, to provide a different perspective and independent academic research into the issues raised in the hypothesis I decided to look at the airline industry. The airline industry and IT share some interesting common ground.

| TOP 5 COMMON CAUSES OF AIRCRAFT LOSS |
|---|
| 50% - Human Error (Pilots) |
| 20% - Equipment Failure |
| 10% - Bad Weather |
| 10% - Sabotage |
| 50% - Human Error (Non Pilot) e.g. Air traffic Controllers, Fuellers, Maintenance engineers |

Firstly, there is an obvious commercial imperative for reliability. This is an industry that has embedded the HA design principles of resilience and fault tolerance to every aspect of aircraft design and has developed sophisticated cockpit automation (aka the auto pilot) to remove the need for the majority of manual activities and corrective actions to be performed by pilots.

Secondly, like IT, human error is ranked as the highest cause of airline flight failure[8].

The hypothesis we are exploring from an IT perspective (Low frequency of failure + increasing dependency on automation is increasing the risk of Human Error) can also be applied to the airline industry.

Not surprisingly given this industry's continued focus on air safety there has been significant research on whether automation is impacting the retention of manual flying skills by pilots. The findings of this research I have summarised below which makes interesting reading when we consider this from our original IT perspective.

**Report: The Retention of Manual Flying Skills in the Automated Cockpit[9]**


**Objective:** The aim of this study was to understand how the prolonged use of cockpit automation is affecting pilots' manual flying skills.

**Background:** There is an ongoing concern about a potential deterioration of manual flying skills among pilots who assume a supervisory role while cockpit automation systems carry out tasks that were once performed by human pilots.

**Method:** We asked 16 airline pilots to fly routine and non-routine flight scenarios in a Boeing 747-400 simulator while we systematically varied the level of automation that they used, graded their performance, and probed them about what they were thinking about as they flew.

**Results:** We found pilots' instrument scanning and manual control skills to be mostly intact, even when pilots reported that they were infrequently practiced. However, when pilots were asked to manually perform the cognitive tasks needed for manual flight (e.g., tracking the aircraft's position without the use of a map display, deciding which navigational steps come next, recognizing instrument system failures), we observed more frequent and significant problems.

**Key Finding:**

*"Pilots performed well at detecting failures but often neglected to cross-check other instruments, diagnose the problem, and avoid the consequences of an unresolved failure".*

[8] Statistics compiled from the PlaneCrashInfo.com database, representing 1,104 accidents from 1/1/1960 to 12/31/2015

[9] Report published 2014 – Authors: Stephen M Casner (NASA), Richard W Geven, Matthias P Recker (San Jose University), Jonathon W Schooler (University of California)

### 7.4 Case Study


**Examples of 'Human Error' causing IT downtime**

The following is a small sample of major IT outages attributed to 'human error'

**British Airways**

A data centre outage resulting in the cancellation of over 400 flights, leaving 75,000 passengers stranded on a busy bank holiday weekend. The incident was allegedly traced to a single engineer who disconnected and reconnected a power supply, causing a power surge that severely damaged critical IT equipment. The technician, part of a team operating at the Heathrow facility, was authorized to be on the premises but not to disconnect the power supply in question.

**Amazon Web Services (AWS)**

An Amazon Web Services engineer trying to debug an S3 storage system in the provider's Virginia data centre accidentally typed a command incorrectly, and much of the Internet – including many enterprise platforms like Slack, Quora and Trello – was down for four hours.

**GitLab**

In the effort to fix a slowdown on the site, a system administrator accidentally typed the command to delete the primary database. GitLab had to restore a 6-hour-old backup that meant any data created in that six-hour window may have been permanently lost.

### 7.5 Conclusion

There is a famous saying "To err is human" and various sayings with the sentiment that every mistake is an opportunity to learn and improve.

The term 'human error' can have a broad application and can often be attributed to lack of training, poor or non-existent controls, following outdated procedures and documentation etc.

In this paper we have however explored the hypothesis that well trained and experienced IT personnel over time lose the insight, knowledge and skills associated with event correlation, diagnostic assessment, problem solving and performing regular IT recovery due to diminishing familiarity and practice. This being associated with the growing influence of HA design and systems automation that performs tasks and invokes recovery routines without human intervention.

"Pilots performed well at detecting failures but often neglected to cross-check other instruments, diagnose the problem, and avoid the consequences of an unresolved failure".

In validating this hypothesis we looked at the airline industry where there is common ground with IT with regard to the focus on resilience, fault tolerance and sophisticated automation.

Despite this focus on HA design and automation, pilot error is the highest single cause (50%) of airline flight failures. This is a pattern we are now seeing in IT with human error being cited as a top concern and the biggest cause of unplanned downtime.

The report we used as research into the impact of automation on airline pilots highlighted the erosion of the pilot's cognitive skills, which in exception conditions, where automation is not working correctly or has failed introduces the higher risk of human error.

This I contend is exactly what we can see emerging in the IT industry with IT personnel eroding their cognitive skills, e.g. correlating events, diagnostic assessment, problem solving and performing It recovery.

# 8 (Hypothesis 2) - There are recovery scenario's being encountered that standard operating procedures are unable to resolve.

"Service failure is one of the main determinants for customers changing providers and successful recovery from these failures is seen by some as critical for customer retention. Recovery is especially important for service providers for whom ensuring an error-free service is impossible"

Source: The Service Recovery Paradox – M McDonough, Sundar G Bharadwaj 1992

## 8.1 Potential Consequences

Availability Management and IT Service Continuity are the core activities within Service Design to ensure the business requirements for the availability and reliability of an IT service can be met. However, as we have discussed in this whitepaper, unplanned outages and prolonged service interruptions can still occur, some would say are inevitable.

Where standard operating procedures for backup and recovery are unable to deal with the more complex and 'out of norm' failure scenarios, this can have the following consequences:-


- Outage time is significantly extended as no recovery path is immediately available
- For non-standard recovery there is no reference point to provide customers with a realistic ETA for service restoration
- Recovery options may need 3rd party guidance and expertise due to the complexity of the issues faced
- Recovery may require programmatic solutions to be provided (typically for data processing errors)
- Data loss may be a consequence with commercial and customer implications
- May invalidate DR invocation or force a lengthy DR invocation
- Customer confidence in the IT service provider capability is seriously tarnished

## 8.2 Possible Causes

There are a number of unexpected scenarios where standard operating procedures may not exist or they in themselves cannot perform the required recovery due to other failures in the 'recovery chain'. These include:-

- Backup routines have not secured the required data successfully which only manifests its self when data recovery is required
- Inputs to the recovery process are not available, e.g. files deleted in error
- There has been an over reliance (and trust) in fault tolerant technologies which present 'failures' that technology design should avoid and therefore recovery procedures do not exist
- There is a dependency on availability of other key subsystems to support recoveries that are also unavailable, i.e. Storage subsystem, Auto Tape Libraries, Job schedulers
- Human error indicating a lack of controls to restrict or control access to critical resources

### 8.3 Findings & Observation

The ability to recover from failure is typically predicated on dealing with the expected 'break' points within a design and developing the standard operating procedures necessary to deal with these failure scenarios'.

Within any recovery procedure design there will be assumptions/requirements that all the required elements to enable a successful recovery are available.

Consider this simple analogy of a car. An expected 'break point' is a tyre puncture. A spare tyre is provided and instructions documented on how to perform the tyre change. It assumes the car jack is available. If it isn't, or doesn't work then what should be a straightforward recovery is now a major issue.

In researching this hypothesis I have encountered many examples where the basic assumptions/requirements for what needs to be in place to perform recovery have not been met. As a consequence what started as a standard recovery has changed into a significant recovery challenge.

Some typical themes observed include:-

- Backup routines have not executed successfully and expected data has not been backed up and therefore cannot be restored
- Database recovery to the point of failure cannot be performed as input transaction logs have been deleted in error
- A hardware failure impacts multiple software subsystems necessary to perform individual component/application recoveries, eg. Auto Tape Library control files are corrupted so the location of backup tapes within a tape silo are unknown and cannot be retrieved manually.
- Resilient solutions encounter a failure that the design should avoid and consequently no standard operating procedures exist
- Human error resulting in a failure that standard operating procedures cannot handle

### 8.4 Case Studies


Examples of non-standard recovery scenarios being encountered

**Example:- Backup routines not working correctly but undetected**

Each night, a new trainee operator ran through the backup process: he loaded the tapes, ran the backup, labelled the tapes, boxed the tapes, and had them ready for the courier the next morning. It wasn't until a few weeks later when a user called about a corrupted file that needed to be restored from tape that his superiors realized *all the backup tapes were blank*.

Upon questioning the operator, he explained that he followed the proper procedure. After a bit of digging, it was learned that the backups were finishing in about 5 minutes.

Due to his inexperience, the operator did not understand that this short timeframe indicated an aborted backup due to an error. This company was lucky. They found the problem early on because of a single corrupted file. Another company had a similar situation … and after a major outage, they had to go back 6 months to find a valid copy of their data for the restore.

**Example:- Trusting fault tolerant technology which then fails with no backup plan**

A media company stored TBs of their digital content. They protected their content with a RAID storage subsystem and naively believed that they did not need to back up their files.

Then a drive failed. Normally the failure of a single drive in an array would not result in any data loss. However, the engineer called to replace the faulty drive mistakenly pulled out the good drive to replace it, not the failed one.

The data was lost, and without a backup the company had to use a specialised data recovery service to recover data which proved time consuming and costly.

**Example:- Human error creating a non-standard recovery scenario**

An IT organisation used an external time server for its internal systems.

All of a sudden the external timeserver (NTP) changed its time to +1 year ahead. When someone managing the external timeserver noticed the time difference, he or she adjusted the time and put it one year back.

The customer using the external timeserver was using Active Directory and the time of the AD was synchronized with this time server. All of a sudden objects in AD had a timestamp one year ahead of the now corrected time. This resulted in a wide range of authentication issues. The organisation required Microsoft support to resolve these date inconsistencies. A significant amount of production time was lost fixing this issue.

**Example:- Mainframe Job Scheduler database corrupted and full recovery was not possible**

This computer glitch in 2012 at the Royal Bank of Scotland left millions of customers unable to gain up to date and correct accounts for several days, in some cases weeks.

During a planned software upgrade to their mainframe job scheduler the main database was corrupted and it is understood transaction log files were erased in error. This meant the job scheduler database could not be recovered to the 'point of failure'.

For large mainframe users, and certainly for banking and finance companies the mainframe job scheduler is a key software product used to automate large sequences of batch mainframe work (which are usually referred to as 'jobs'). It will start jobs, wait for them to run, then start other jobs dependent on the first ones completing, and so on. RBS updates customer accounts overnight via thousands of batch jobs.

These batch jobs take transactions from various places, such as ATM withdrawals, bank-to-bank salary payments, credits, debits and so on, and finish by providing an updated customer balance.

The inability to recover the job scheduler database to the point of failure meant that the database only reflected the batch schedules and job status at the time it was backed up. The consequence is that batch jobs run since the backup was taken were not reflected in the schedule and it was unclear which jobs had run or not run. Running jobs that had already run would duplicate entries and invalidate customer balances, not running jobs would result in payments in and out of customer accounts not being applied, e.g. monthly salary was not paid when expected, mortgage payments were not made etc.

The technical challenge of establishing what jobs had and hadn't run led to mistakes and the need for reruns and a backlog of processing to catch up with.

RBS were fined £56M by the regulators as a consequence of this major failure.

**Example:- Unable to access backup tapes for database recovery**

Automated Tape Libraries provide an automated solution for the mounting of tape cartridges to perform backups and to restore data when required for recovery purposes.

If the tape libraries are unavailable due to a software or hardware failure, then backup and restore activities cease as tape cartridges cannot be accessed by the robotic tape handlers. Manual intervention is extremely difficult to locate specific tape cartridges as these are barcoded and the configuration of a single library can contain thousands of cartridges. There are of course Health & Safety considerations and for many organisations these cartridge solutions may be part of a remote operation with the tape libraries being in a 'dark site'

In one example, a tape library failure coincided with the need to recover a critical database. The database recovery required the backup and related log files archived earlier to the tape library. These were not accessible. Furthermore, the offsite copies of these files had not been ejected (this process runs once daily) and was not due to run for several hours. These were therefore also unavailable.

The ETA for the tape library to be operational was 4-6 hours. With no offsite copies available the recovery had to wait incurring an extended service outage.

### 8.5 Conclusion

Manging recovery is all about being able to demonstrate control when failure situations occur. In most cases your standard operating procedures will handle routine restart and recovery activities.

The challenge is of course when your standard operating procedures are in themselves not adequate for the recovery activities required to restore service.

In most cases, IT organisations have not considered or planned for the 'double whammy' effect where a standard recovery option is compromised by another error or condition.

Designing for Availability naturally creates a mind-set of 'Fail safe' looking to avoid routine failures but where break points are anticipated making sure standard operating procedures are provided.

From the examples highlighted to support this hypothesis there are 2 key learns:-

1) Regularly validate that your essential data is being backed up and that scheduled backups are being completed successfully.

2) Change the mind-set and recovery ethos from 'Fail safe' to 'Safe fail'. For each critical component start to understand what assumptions/requirements are necessary to perform a standard recovery. Then pose the 'What if' to understand what approach would be required if one of those assumptions/requirements cannot be met.

E.g. Using the RBS job scheduler failure….ask the 'What if' we can't recover the database to the point in failure? How would we be able to ascertain what batch jobs had run and completed since the last backup. How can we automate and capture this to provide an update to the database.

Let's then develop the procedures and produce the documentation so if in a worse case this situation happens we have a proven recovery option to follow.

Having to react to a non-standard failure scenario requires a degree of out of the box thinking which is not naturally supported in a major incident. New approaches being progressed will be subject to errors and additional risks which may further compound the issue.

# 9 (Hypothesis 3) - There is limited confidence in the ability to successfully invoke DR/Failover procedures

*"We asked the respondents whether they executed a failover when IT issues occurred and assessed how confident they felt about whether the failover would work. Overall 58% have experienced some measure of problems or not enough confidence to press the button on their DR."*

Source: 2016 UK Survey 'The state of Disaster Recovery amongst UK Businesses'

## 9.1 Potential Consequences

A lack of confidence in the ability to quickly and successfully invoke DR/Failover procedures in response to a major failure within the underpinning IT infrastructure or supporting environmental systems, i.e. Power supply, Air Cooling Units, can have the following consequences:-

- Delayed decision making during a crucial stage of the major incident when assessing recovery options
- Perseverance with complex diagnosis of the failure within the impacted configuration when initial cause is unknown
- Significant (and avoidable) delays to the restoration of service
- Business impact is compounded with the risk of negative media coverage and a customer backlash


## 9.2 Possible Causes

The reluctance to consider DR/Failover as an immediate and valid recovery option can be influenced by a number of factors:-

- Documented recovery plans do not exist and rely on tacit awareness of the required actions (and the potential for key man dependencies)
- The documented recovery plans have not been tested or tested recently
- The documented recovery plans have been tested but encountered unresolved issues that remain outstanding.

## 9.3 Findings & Observation

Creating a disaster recovery plan is an essential part of the business continuity planning process to ensure you have an effective backup and recovery solution in the event of a major failure impacting your critical IT services.

This you would assume is a recognised necessity for all IT organisations. However in the Zetta survey *'State of Disaster Recovery 2016'* the response to the question 'Do you have a documented DR plan' reveals a surprising result:-

**40% of companies surveyed admitted to having no documented DR plan to guide them in the event of a major IT failure.**

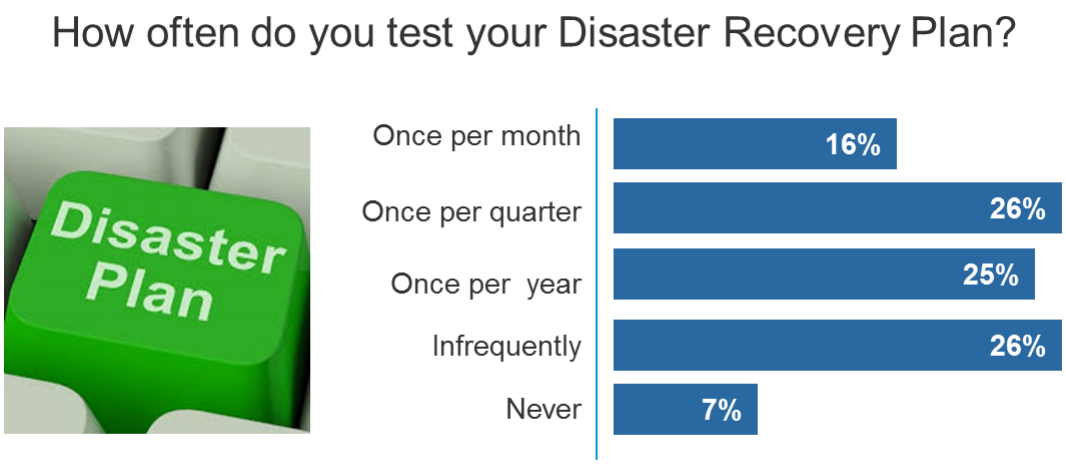

Dilbert cartoon is reproduced under licence

It would of course be no surprise that with no formal documented plans and procedures that there would be little confidence in invoking any failover even though the necessary alternative infrastructure is in place.

If confidence is to be gained, then where a Disaster Recovery plan has been developed it must also be regularly tested. Without the testing and verification of your DR plans, you'll have no idea as to whether or not you'll actually be able to recover from a disaster or extended outage.

Regular testing helps you ensure that all aspects of the DR plan and the associated processes and procedures work as expected to provide you with the confidence to make the invocation decision.

How often should you test your DR plan? The following survey results provide an interesting insight into the importance IT organisations place on DR testing:-




Source:- State of Disaster Recovery 2016 survey by Zetta

**From the above chart you can see that 58% of respondents say they test their DR plan just once a year or less, while 33% of respondents say they test infrequently or never at all.**

Infrequent testing of the DR plan and the processes and procedures used to move to your backup environments impacts confidence and introduces risk.

Confidence is eroded when plans have not been tested successfully within a reasonable period of time as doubts emerge as to how long and how successful invocation will be.

Confidence across the IT support organisation is tempered by a lack of familiarity and awareness with the required invocation procedures for their IT recovery plans for the technologies they support.

The risk of encountering issues and additional delays with the invocation is increased when the DR plan is infrequently tested. The opportunity to identify issues and correct these in regular testing is missed.


Source:- TechTarget storage survey 2017

### 9.4 Case Studies


**Disaster Recovery plan not invoked due to a lack of confidence**

A manufacturing company in America was badly impacted by a major hurricane with the loss of power and localised damage. The company had a DR solution that would allow all critical systems to be restored at a secondary Data Centre (geographically separate location). The Recovery Time Objective to recover all systems was 48 hours.

The decision was made not to fail over their systems to the recovery site. This was because they had no confidence the failover would work or that they could 'failback' cleanly once the primary data centre was stable.

The lack of confidence came from never having tested their recovery plan.

The disruption from the hurricane lasted 6 days significantly more downtime and impact than the 48 hour RTO the failover solution could have provided.

### 9.5 Conclusion

What we have not been able to evidence is that organisations have not made any DR provision so we must conclude that the ability to recover in extremis is recognised as an essential business imperative.

**Famous Saying:**

"If you fail to plan, you are planning to fail"

Quote attributed to Benjamin Franklin – Founding father of the United States (1706-1790)

However, the lack of fully documented plans and/or regular testing to ensure plans are 'proven' would infer that these activities are not being prioritised as essential activities.

Where IT is typically wrestling with the demands on time and resource to deliver business change and 'keep the lights' on the conclusion must be that DR planning and testing is considered sacrificial as it delivers no immediate business benefit.

Where DR activity is not being given appropriate prioritisation, then over time plans are not developed, maintained and regularly tested to increase confidence that they can be executed in anger when required.

# 10 (Hypothesis 4) – Recovery Time Objectives (RTO) are not being achieved and contribute to delayed recovery.

"Only 39% of respondents meet their RTO consistently"

Source: 2017 Disaster Recovery Survey Report (by CloudEndure)

## 10.1 Potential Consequences

The **recovery time objective** (RTO) is the targeted duration of **time** and a service level within which a business process must be restored after a disaster (or major disruption) in order to avoid unacceptable consequences associated with a break in business continuity.

The RTO sets expectation on how quickly services can be restored from the point of invocation of DR/Failover procedures. The inability to meet the stated RTO has the following consequences:-


- Business expectations are not met
- A key SLA target is not met
- Business impact is extended
- The expected time for service restoration may have been communicated by the business to their end customers, the media or industry regulator
- Confidence in the solution is eroded which may delay future use

## 10.2 Possible Causes

In this hypothesis we are making the assumption that there is a documented and periodically tested DR/Failover plan and that as a result the RTO is a realistic target.

The inability to achieve the stated RTO can be due to a number of factors:-

- The procedures required are not automated and require manual intervention
- There is a lack of familiarity with the required invocation procedures creating hesitation and delay.
- The RTO is not validated during DR/Failover testing
- DR/Failover testing is constrained and does not fully replicate the live environment so RTO timings are unlikely to be reflected in the production environment

## 10.3 Findings & Observation

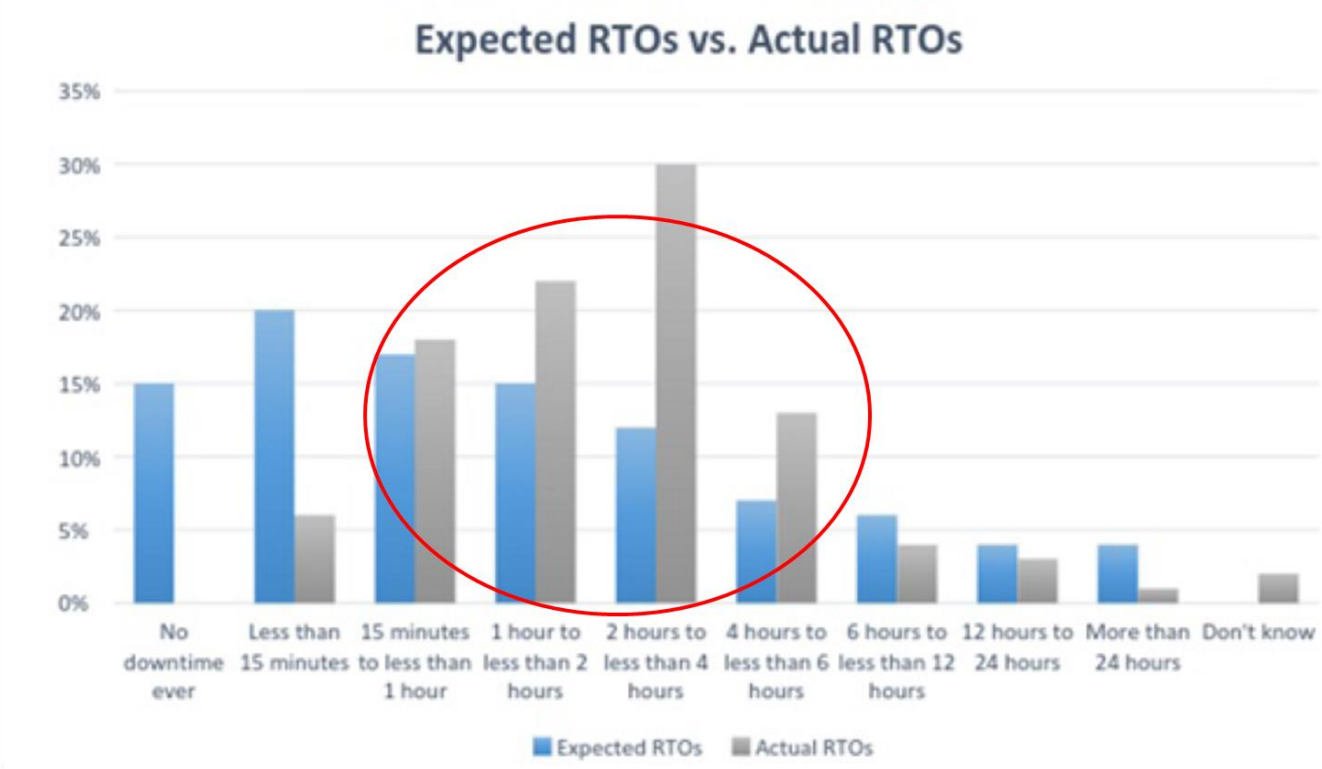
The figure illustrated on the following page below is from an ESG survey taken in 2017.

Here respondents provide feedback on how well their agreed RTO targets are being met.

The findings here would seem to confirm the similar statistic reported earlier from the CloudEndure survey that only 39% respondents consistently meet their RTO targets.

In the ESG survey we can observe an interesting paradox. The RTO targets that are both stringent (<15 mins) and those that are less stringent (> 4hrs) have a higher success rate in the RTO target being achieved.

Solutions designed for an RTO of between 1 hr and 4 hrs highlights where RTO targets are not being consistently achieved.




Source:- ESG research report; The evolving Business Continuity and Disaster Recovery landscape 2016

### 10.4 Conclusion

The above survey results would validate the hypothesis that agreed RTO targets are not being met on a consistent basis and therefore is a key factor in the delay of service recovery.

However, this appears to be an issue more apparent with DR/Failover solutions that target an RTO in the 1hr-4hr timeframe.

Solutions with a stringent RTO of <15 mins have a high level of consistency. It is a reasoned assumption that to achieve this RTO would require no priming of the failover configuration and place a high reliance on automation to invoke and avoid any manual delays. It is also likely that these are well practiced procedures possibly switching between primary and secondary configurations as a routine planned activity.

Solutions with a much less stringent RTO of >4 hrs up to 24hrs, also have a higher success rate. Again it is a reasoned assumption that the RTO infers a need to prime the failover configuration (particularly if this is at a 3rd party recovery services provider location), await data restoration and relies on structured manual procedures with activities that have a degree of parallelism. With these longer RTOs it can also be argued that this maybe a less pressured environment with less likelihood of human error.

So what is the issue with solutions that target an RTO in the 1hr – 4 hr range?

These timings would infer that the DR/Failover configuration is primed and available. However, as the RTO is up to 4 hrs this may not be suited to regular planned switching. So familiarity with the DR/Failover procedure could be a factor. Reliance on automation is less likely for a failover of up to 4 hrs so may in fact require a high degree of manual procedures to be followed in sequence to ensure data integrity is assured. It is here that a lack of familiarity and hesitation can result in delays or open scope for human error.

# 11 Failure Vs Fiasco

Several studies have shown that recovering well from a failure in service can lead to a higher customer satisfaction level than never having a failure at all

Source: The Service Recovery Paradox – M McDonough, Sundar G Bharadwaj 1992

## 11.1 Failure

This may be an extreme view to make a point, but often the business only recognise their dependency on their IT service provider when things go wrong. I use the 'tap' analogy to expand this thinking.

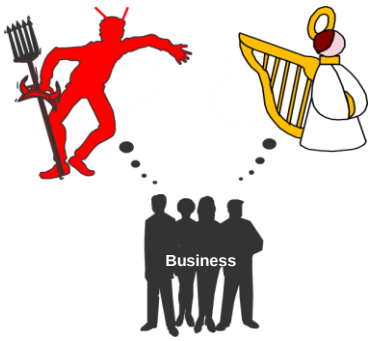
### 11.1.1 The 'Tap' Analogy

A long period of operational stability with continued high levels of availability begins to create the business mind-set of this being the 'expected norm'.

The analogy is that in many ways the business view IT availability in the same way we expect water to flow from a tap when turned on. We expect this to happen every time and cannot remember when the water supply last failed. If the water supply did fail it is likely that we are not prepared to deal with this and will cause some significant disruption to the household.

In reality occasional IT failures do occur, and similar to the analogy above a consequence of this is that dealing with IT failures and invoking business workarounds is something that the business may no longer be adept at performing or may even no longer be viable.

### 11.1.2 The IT Service Provider – Hero or Villain?




There is no doubt that poorly managed IT failures and delayed recovery will erode customer confidence and trust. The reputation of the IT service provider will be damaged and the value of the positive improvements in service quality that have been made is 'lost' as customer perception changes. As a result, the IT service provider is now on trial with their reputation now only as good as how well they manage the next IT failure.

However, IT failures need to be viewed in a different way by the IT service provider. They become *'Moments of Truth'*, windows of opportunity where customer satisfaction can be maintained or even improved based on their response to service failure.

### 11.1.3 Service Recovery Capability

Many IT service providers recognise this and have developed a service recovery capability that provides the ability to differentiate and manage incidents that have or have the potential to cause significant business impact and typically have some or all of the following in place:-

- Incident categorisation defines and prioritises ' Major Incidents'
- A Major Incident Management (MIM) procedure is documented
- The role of 'Major Incident Manager' is assigned (This may be a dedicated role/team)
- A Major Incident Team is invoked
- A facility ( Major Incident Room) is convened to co-ordinate activities and manage technical recovery and provide status updates to stakeholders
- Telephony and collaboration tools are exploited to allow technical groups, vendors and suppliers to participate regardless of location

### 11.2 Business Perception…on how IT manages major incidents

In my experience the business view of their IT service provider's capability in handling major incidents is based simply on how well did they enable the business to manage the impact to their users and end customers.

Their perception is influenced +/- by the following:-


- How quickly the business were engaged by IT
- The quality *(timeliness, openness, fact based content, accuracy of estimated timelines for service restoration)* of verbal and written communications from IT into the business
- Two way communication providing the business with input to key decisions

You are more likely to receive compliments from the business for a 'well managed incident' where the business were able to demonstrate control and confidence to their customers, regulators and occasionally the media because they were proactively engaged and communications were timely and trusted.

This is particularly true today with the growing emergence of social media where there is a clear shift in customer behaviour to not accept poor service and an indifferent customer experience and 'share' this using social media channels such as Facebook and Twitter.

Where the business are slow to respond to their customers and/or poorly manage their expectations an IT failure can quickly get out of control and lead to media interest and wider scrutiny. What starts as a 'failure' ends in being labelled a 'fiasco'

### 11.3 Fiasco

**Definition of Fiasco**

A complete failure, especially a ludicrous or humiliating one.

When reading the media reports relating to news worthy high profile IT outages then journalists often use words such as:- *failure, disaster, catastrophe*. When you read the Twitter comments from frustrated and angry customers the lexicon often consists of more pointed commentary to express customer feeling: *debacle, shambles, farce, mess, car crash, cock-up.*

All these words are synonyms for the word **FIASCO.**

#### 11.3.1 What differentiates a Fiasco from a Failure?


- **Business Perspective**

Earlier in this whitepaper we gave examples of major IT failures that created negative media coverage, brand damage and resulted in significant financial loss (i.e. Fines, penalties, compensation, drop in share price, loss of customers).

Observations of these high profile failures show common themes:-

- Delays in instigating customer centric communications
- Inability to provide realistic expectation on when service will be resumed
- Perceived lack of openness on what happened and why
- Perceived lack of empathy for the impact and consequence of failure on customers
- Excessive duration (more than a couple of hours)

What becomes apparent is that it is more about how badly a company handles an IT outage in the public domain which generates the negative 'Twitter storm' and media interest rather than necessarily the IT issues that caused it.

It could be argued, that this is the responsibility of the corporate entity and how they perform Crisis Management, Manage the media and their approach to PR.

- **IT Perspective**

However, this situation has been caused by an IT failure and more significantly the IT service providers inability to recover and restore services in a timely manner. Characteristics of the IT outages we have seen gain adverse media coverage include:-

- Services unavailable for a significant period of time, i.e. a full business day or longer
- Issues encountered are outside the scope of standard operating procedures, e.g. no immediate recovery path exists
- IT unable to provide realistic estimates on service restoration

Duration is a key factor in how customers, the media and regulators will view this IT failure. The longer the outage the greater the customer frustration and anger and the more likely the term 'Fiasco' and its synonyms will be applied by customers and the media.

### 11.4 How can MIM better support the Business in these scenarios?

In this paper we have seen that HA + DR are not the silver bullet that will prevent major outages.

The working assumption for IT has to be that whilst these solutions can provide a lower frequency of failure and the ability to provide timely recovery from anticipated failures, there are many risks which if materialise can result in a complex major failure that has the potential to create a 'fiasco' situation for the business.

Earlier we mentioned that many IT organisations have invested in Major Incident Management (MIM) to make service recovery a core capability and that the business ability to manage the business consequences of the outage is very dependent on early and continued engagement with MIM and the quality and timeliness of MIM communications.

Over many years I have been involved in the management of high impact incidents (some that would attract the label 'fiasco 'by the media and end customers of the business) and this has helped give me greater insight in terms of how MIM could better support the business.

#### 11.4.1 Observations

The majority of major incidents can be considered 'normal' or 'routine' situations in that a failure has occurred, the underlying issues are quickly identified and standard recovery procedures can be invoked. In these conditions the MIM process works well.

Where MIM comes under stress is when the underlying issues are not immediately apparent and/or a standard recovery procedure doesn't exist.

As the clock ticks the pressure on the MIM process and all who execute within it gets ratchetted up. Stakeholders now include senior business executives across the business, ie, corporate communications, risk management, business continuity and of course senior IT roles such as CIO, IT Director, Heads of Department.

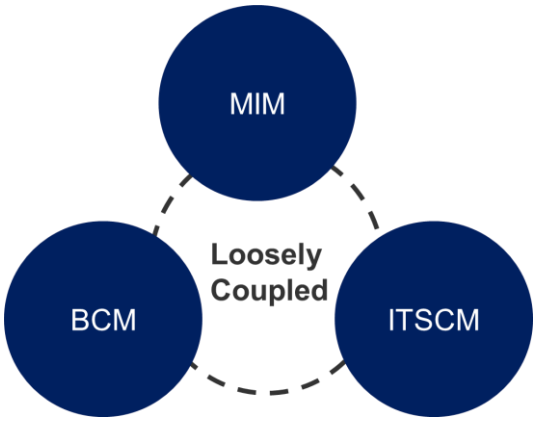
The clamour (and perhaps the word frenzy is not misplaced here) for information can overload the MIM process and actually distract the focus on trying to recover the service.

### 11.4.2 Causes

#### 11.4.2.1 Loosely Coupled Disciplines

When a major incident starts to exhibit the characteristics of the underlying issues not being identified and/or the recovery path is not clear this is typically when the wider stakeholder groups start to get involved.

Often this is an hour or two after the MIM process was invoked. It is only now that the wider organisational disciplines of Business Continuity Management (BCM) and IT Service Continuity Management (ITSCM) get involved.




Often this can be quite unstructured and perhaps unfamiliar and rusty. Whilst there is recognition of the relationship across and between these disciplines it often assumes these only come together for the classic 'Disaster', e.g. an Aeroflot tailfin sticking out of the ground where your Data Centre used to be, with triggers and roles documented in the organisations IT DR invocation plans.

The mind-set is that an IT disaster is when we need to invoke DR, As we have seen in this paper there are many examples of major failures where DR is not a viable option.

Engagement in this context may be considered as originating from 'bau' and this requirement is often not recognised so no triggers exist and consistent escalation points defined.

#### 11.4.2.2 Communication is driven by who shouts loudest

The longer the outage the greater the clamour for information with adhoc demands coming into the MIM process from roles with seniority and gravitas.

The incident manager can be swamped with these demands arriving by phone, SMS or individuals walking into the incident team location. The focus on technical recovery can suffer.

In many cases, the Incident Manager assigned to manage a major incident has a dual role to coordinate technical diagnosis and recoveries as well as ensuring regular status updates are issued to a predefined list of recipients.

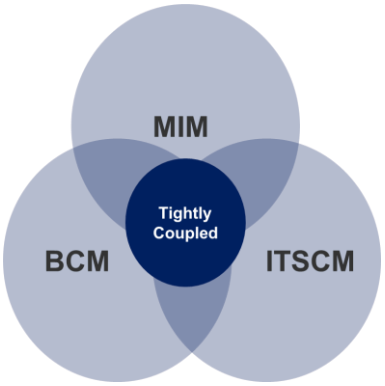
The importance of communication is such that this dual role is ineffective in managing these increasing communication demands.

## 11.5 Making improvements to Major Incident Management

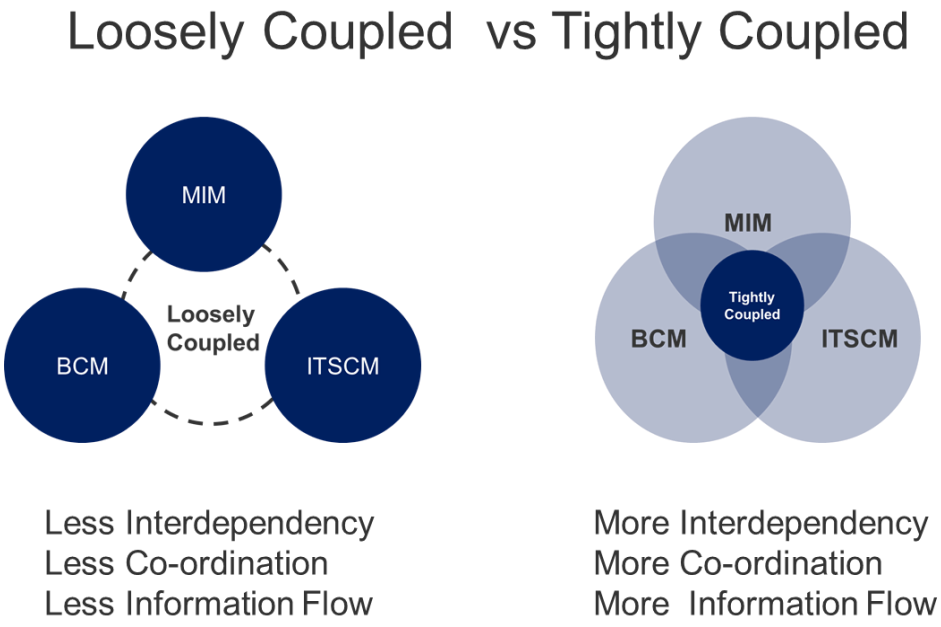
The following are improvements you can make to improve the effectiveness of the MIM process and how it best supports the business in dealing with the consequences of protracted IT outages to the media, regulators and their end customers.

### 11.5.1 Tightly Coupled Disciplines

The MIM, BCM and ITSCM disciplines should be aligned and engaged as soon as a major incident is invoked. Initially this might be 'light touch' with the option to be disengaged as soon as a known recovery option is commenced.




The benefit of this initial 'light touch' is that key roles are primed and if the major incident escalates they are already engaged and ready to commit. Consider, if for every major incident the company's Corporate Communications team were alerted and on an initial conference call to understand the business impact. These teams are typically those contacted by the media and who manage the corporate social media communications, so straight away this area is on the front foot.




### 11.5.2 Assign the role of Communications Manager

Communication throughout the lifecycle of every major incident is so important that it warrants a dedicated Communications Manager.

This divides responsibilities for MIM communications and technical co-ordination and recovery and therefore for every major incident there are two dedicated roles:-

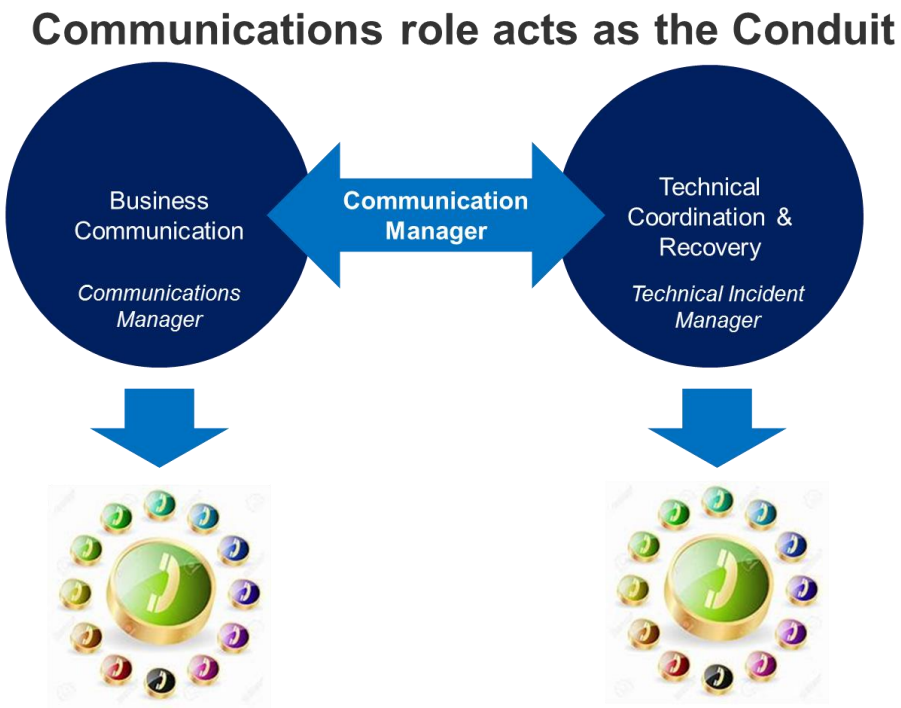
- Technical Incident Manager (aka Incident Manager)
- Communications Manager

The focus of the Communications Manager is to be the conduit between the Business stakeholders and the Technical co-ordination and recovery.

The communications manager ensures the appropriate level of business engagement tales place in a structured and controlled manner. This role takes away any demand for communication from the Technical Incident Manager whose sole focus is technical coordination and recovery.

The Technical Incident manager retains the overall accountability for the major incident but is considered a peer to the Communications Manager with whom they need to define how they will collaborate together for the duration of the major incident and how the activities of technical co-ordination and recovery interface with those of Communications.

The role of Communications Manager can be considered as a ‘hybrid’ role as it requires a range of skills and competencies that can understand complex technical issues but enable these to be explained in business terms and vice versa.




I have seen this approach work well in a number of organisations and the communications process improve and mature because of this focus.

## 12 Recommendations

These are the high level recommendations based on the insight and learning gained from the research and analysis of the 4 hypotheses listed in this whitepaper.

| **Hypothesis 1** | ***Low frequency of failure + increasing dependency on automation is increasing the risk of Human Error*** |
|---|---|
| 1A | Lack of familiarity and practice is where human error can occur when dealing with IT failure scenario's not previously experienced and practiced.<br><br>In the airline industry to counter this, pilots regularly perform routine and non-routine procedures in the flight simulator without the benefits of automation.<br><br>The nearest IT equivalent to the simulator is the test environment. It is recommended:-<br><br>That these environments are scheduled periodically for support staff to practice and perform a various operational activities and recovery procedures with automation disabled.<br><br>• Create 'crash and burn' scenario's to exercise recovery skills, i.e. pull out cables, delete key resources, erroneously edit data or system parameters<br><br>• Schedule a manual close and restart of key software and hardware components to be performed without automation. Operational staff used to do this regularly as a housekeeping routine but advances in OS and Hardware reliability now render this activity as an exception and can be forgotten. This can be an issue if automation fails.<br><br>• Consider the above as part of an individual's personal development plan, This can be two way:- For those who need to devise the crash and burn scenarios and expected outcomes creates a learning opportunity to hone their restart & recovery competance: For those who perform diagnosis and recovery it's an opportunity to learn from errors and mistakes in a 'safe' environment |
| 1B | ITIL recommends the use of 'models' across many of the Service Operation and Service Transition disciplines to document the approach required for specific situations.<br><br>The learning from 'crash and burn' scenario's performed as either BAU training events or as part of pre live testing should be captured and documented to provide:-<br><br>• Incident Models for recovering from specific failure scenario's<br>• Problem Models for aiding problem determination for specific failures<br>• Change Models for infrequent but potentially high risk change implementations *<br><br>** A good example is performing a base rate change in UK banking . After several years of the base rate remaining static it was recognised that many support staff had left or forgotten the process for doing this and indeed IT systems and applications may have changed. A number of Banks instigated the creation of a change model for this process.* |

| | |
|---|---|
| 1C | For mission critical software components and products many vendors provide restart and recovery training courses or consultancy:-<br><br>• Consider sending staff on available restart & recovery training courses to review internal procedures and approaches when they return.<br>• Reinforce the learning from the education by performing test recoveries to validate understanding and any procedural improvements made.<br><br>And/or<br><br>• Consider using vendors to perform an external review of your restart & recovery procedures to highlight, shortcomings, gaps and opportunities to improve.<br>• Test all recovery procedures that have been changed. |

| **Hypothesis 2** | ***There are recovery scenario's being encountered that standard operating procedures are unable to resolve.*** |
|---|---|
| 2A | Implement 'What if' technical workshops.<br><br>The 'What If' approach is based simply on the premise that the standard approach to any recovery doesn't work or is not possible, in other words when considering the recovery approach pose the question 'What if'?<br><br>• Using a workshop approach select a critical technical component and map out the standard recovery approaches.<br><br>• What assumptions are made on the availability of all required inputs and the recovery environment?<br><br>• For each recovery approach remove one of the assumed pre requisite inputs or dependencies within the recovery environment and work out as a group what recovery approach can now be considered.<br><br>• This should highlight the potential for workarounds, manual updates, potential use of vendor support products and offerings that can assist with non-standard recovery etc.<br><br>• Document outputs and look to recreate these in the test environment. Where recovery has been possible document the approach and importantly an estimate of recovery time. This would be important should such an 'out of the norm' recovery scenario occur in the live environment in the future.<br><br>• Validate the approach and rationale with your vendors |
| 2B | Schedule periodic 'scenario planning' exercises for your critical recovery plans. I.e. IT Service Continuity plans, IT Security plans.<br><br>Scenario planning is a structured way for the IT organisation to think about how existing recovery plans would be executed based on a how a particular scenario may unfold and what decisions would need to made.<br><br>This approach can also be used to begin the process of creating plans for scenarios where recovery plans don't currently exist.<br><br>Scenario planning exercises need good design to create the scenario and the unfolding events within the scenario. |

The following provides high level outline on the purpose and objective for your scenario planning exercises:-

**Purpose**
To 'test' existing plans against specific scenario's to ensure plans remain 'fit for purpose'

**Objective**
To identify shortcomings, gaps and opportunities to refine existing plans based on the findings and observations of the current plan in responding to the scenario's presented

**Audience**
They should be attended by a range of IT roles covering leadership, management and practitioner. Ideally they should be facilitated by someone independent of the roles/teams participating.

**Outputs**
- Agreed actions areas for addressing the issues raised
- Ownership assigned for all actions with target dates for completion

**Example – The IT Service Continuity Plan**

This was a scenario I used to assess the ITSCM Plan and deal with potential DR situations

Typically these plans are written from the perspective that a catastrophe has happened, the decision has been made to invoke DR and so the plan documents the service and system priorities to commence invocation and recovery actions.

The scenario was to present a 'potential' threat to the Data Centre (DC) from a local chemical leak and resultant gas cloud.

The events unfolding were…… all staff were ordered to vacate the building by the emergency services…, the gas cloud changes direction and heads towards the DC…..We are informed the cloud contains acid which if drawn into the DC could seriously damage equipment….key vendors are contacted and recommend their equipment is powered down….we are informed the gas cloud will pass over the DC but should disperse within an hour

The issue identified included:-

1. Facilities Management were not part of the plan as in this scenario they needed to stop air intake and how could this be done remotely?
2. Remote access to all devices to perform a pre-emptive power down was not clear.
3. If the DC was powered down there was no 'power up' plan or any view of how long this would take
4. Would DR be invoked in this scenario (view was no) but how would this decision be supported?
5. Should services that could be immediately failed over but are not priority services be performed?

The main observation was that the plan had been predicated on the decision to invoke DR having been had been made and therefore was not geared to deal with potential DR situations and how these would be managed.

| **Hypothesis 3** | ***There is limited confidence in the ability to successfully invoke DR/Failover procedures*** |
|---|---|
| 3A | Define and maintain an awareness campaign for all stakeholders involved in the invocation and execution of activities within the plan. This can include:-<br><br>• Documentation 'page turn' walk thu's with specific groups<br>• Perform scenario planning workshops (see previous recommendation)<br>• Schedule informal 'brown bag' lunch sessions to provide overviews<br>• Post Intranet updates and articles on the importance of the DR plan<br>• Communicate DR testing successes<br>• Create and publish a monthly DR 'dashboard' of relevant measures, test results, highs and lows.<br><br>Ensure your DR plan is regularly reviewed and kept up to date. |
| 3B | The importance of DR testing needs to be established and its status and priority as non-discretionary work agreed by the IT directorate.<br><br>Agree the number and types of DR testing to be performed on an annual basis and consider a 'project management' approach to DR testing to work within resource estimation and allocation processes to agree and lock in resource requirements.<br><br>This approach is to avoid DR tests being viewed as discretionary pieces of work at the mercy of BAU demands for live support and business change. |
| 3C | The DR testing strategy should define and again agreement on the minimum requirements for DR testing of all critical IT services. When considering how often DR plans should be tested the rhetorical answer is probably more than you are currently doing!<br><br>Technical constraints and resource conflicts will always be a challenge but by agreeing minimum requirements this becomes an organisational commitment against which exceptions can be escalated.<br><br>An example of your minimum requirements for an IT service could be as is as follows:-<br><br>• An annual E2E DR test is completed<br>• Individual component recoveries are performed twice per year<br>• Desktop reviews of the recovery plan for the E2E service are performed twice per year |
| 3D | Where the DR solution for an IT service includes backup components with the ability to failover in the event of a failure then consider:-<br><br>• Scheduling regular failover between primary and alternate components so that these become productionised.<br>• Scheduled failover change records should include the RTO for the failover to complete and change success/failure takes account of the RTO being achieved<br>• Consider running production workloads from the alternate components for a defined period of time.<br><br>This approach will ensure familiarity with the failover procedures and increase confidence that failover can be used in anger when required. |

| | |
|---|---|
| 3E | Ensure that issues identified in DR testing are raised and managed within your organisations incident and problem management processes:-<br><br>• There should be no differentiation in the priority levels assigned between production issues and DR issues.<br><br>• There is no reason why a showstopper issue identified in DR testing which would prevent a successful DR invocation cannot be give a Priority 1 status.<br><br>• Support teams should not be allowed to downgrade DR related incidents or problems.<br><br>• This is an important cultural consideration to ensure issues reported from DR testing are not allowed to 'drift' and remain unresolved for an excessive period of time. |
| 3F | Work collaboratively with the IT personnel responsible for Major Incident Management (MIM) to provide an early engagement between MIM and ITSCM for each major incident.<br><br>In the majority of cases ITSCM would be stood down early where clearly DR options are not required. However. By gaining early engagement DR options can be presented and risks assessed quickly. ITSCM presence in the MIM process can provide confidence in the viability of DR/failover plans to restore services asap. |

| **Hypothesis 4** | ***Recovery Time Objectives (RTO) are not being achieved and contribute to delayed recovery.*** |
|---|---|
| 4A | The achievement of the RTO should be a specific objective for every DR test and the overall success of the DR test should factor in the RTO achieved<br><br>• Failure to achieve the RTO in a DR test must be recorded and investigated under the problem management process.<br><br>• Criteria should be agreed to determine if an urgent retest is required where the RTO has been significantly exceeded. |
| 4B | See also 3D<br><br>Where the DR solution for an IT service includes backup components with the ability to failover in the event of a failure then consider:-<br><br>• Scheduling regular failover between primary and alternate components so that these become productionised.<br>• Scheduled failover change records should include the RTO for the failover to complete and change success/failure takes account of the RTO being achieved<br>• Consider running production workloads from the alternate components for a defined period of time.<br><br>This approach will ensure familiarity with the failover procedures and increase confidence that failover can be used in anger when required |

| | |
|---|---|
| 4C | It is not uncommon for the author of specific DR/failover plans to also perform the tests.<br><br>This introduces the risk of creating a key man dependency and an additional risk that the documentation may not have the clarity required for someone not familiar with the procedure to execute this without error or delay.<br><br>• It is recommended that DR/Failover tests are rotated across the IT personnel within the group responsible for this technology.<br><br>• This will increase familiarity and highlight any unclear aspects of the documented procedure. This should ensure over time RTO targets are consistently met.<br><br>• The additional benefit is that by exposure to the wider group, areas for improvement and optimisation may be identified to reduce RTO timings. |

********* End of Report **********

## 13 References

The following are reference materials I have used as input to this whitepaper:-


- Axelos publication:- ITIL Service Design (Sections 4.4 and 4.5)
- CloudEndure publication:- 2017 Disaster Recovery Survey Report
- Zetta publication:- State of Disaster Recovery 2016
- iLand publication:- The State of IT Disaster Recovery amongst UK Businesses

- Research report:- ITIC 2017 Global Reliability Survey Mid-Year Update
- Research report:- ITIC 2015 - 2016 Global Server Hardware, Server OS Reliability Report
- Research report:- The Retention of Manual Flying Skills in the Automated Cockpit
- Research report:- The Evolving Business Continuity and Disaster Recovery Landscape
- Research report:- The Service Recovery Paradox – M McDonough, Sundar G Bharadwaj 1992

- Article:- DataCentre Dynamics - The true cost of downtime
- Article:- SunGard AS:- Tales from the Disaster Recovery (DR) Graveyard
- Article:- Enterprise Features - Disaster Recovery Horror Stories
- Article:- SQLskills - Human nature is a significant hurdle to successful disaster recovery
- Article:- CloudEndure - HA vs DR Why High Availability just isn't good enough
- Article:- CNN – Blackberry service restored after worst outage ever
- Article:- Computerworld UK – The worst software glitches in recent history
- Article:- Worksoft – Top software failures of 2017
- Article:- The 10 biggest cloud outages of 2017
- Article:- 7 painful website outages that kicked off 2017
- Article:- Arcserve – 6 Data Centre outages and lessons learned from 2016
- Article:- CloudTech – How often should you test your disaster recovery plan
- Article:- Mailonline – Why do planes crash?

## 14 Author Contact details

I hope you find this whitepaper informative and useful in reviewing your approach to IT service recovery.

If you require any further information then please do not hesitate to contact me:-

Email:- IKMACDONALD@BTINTERNET.COM

Mobile:- 07809511458

Linkedin:- www.linkedin.com/in/iankeithmacdonald

Edenfield IT Consulting Limited