Global Data Center Engineering

GDCE White Paper Series:

Human Error Management and the Data Center

It is certainly not a requirement to be a data center operator or major business unit to understand the impact of data center outages. In the past 12 months alone, millions globally have been affected by a data center outage. Some are mere nuisance or inconvenience while others can have profound, life altering impacts. Here are just a few of the publicly reported events from recent years:

- May 2017 – British Airways – UK Loss of systems grounding all flights for nearly two days, at an estimated loss between $150 million - $200 million USD, affecting more than 75,000 passengers. Recovery from backlog, about 14 days – Cause – Human Error
- January 2017 – August 2016 – Delta Airlines – US Loss of systems including website and mobile apps, grounding flights for about 5 hours, causing cancellations and other conflicts, at an estimated cost of $150 million USD. Recovery about 2 days – Cause – Human Error
- January 2017, October 2016 – United Airlines – US System outages caused cancellations and delays as manual systems were in place for ACARS systems. No estimates available for cost – Cause – Human Error
- July 2017 – Southwest Airline – Loss of systems, grounding flights. Estimated cost $177 million USD – Cause – Human Error
- September 2016 – June 2016 – Global Switch – UK loses power due to failure in high voltage breaker, impacting rotary UPS systems. The power loss was brief (222 milliseconds), but the impact was wide, and more than 2 hours for power to be restored – Cause – Mechanical Failure
- September 2015 – Fujitsu Cloud – US Cloud Services disrupted after utility power surge. Issues persisted for several days after power was restored – Cause – Human Error
- September 2015 – Amazon AWS – Loss of cloud services for 5 hours effecting Netflix, Airbnb, Reddit, IMDb and others – Cause – Human Error
- August 2014 and again November2014 – Singapore Stock Exchange lost power and systems for up to 4 ½ hours at a cost of over $1 million USD, due to lightning strike and a mismatch of synchronization of the rotary UPS afterward – Cause – Human Error
- April 2014 – Samsung – Massive data center fire resulted in entire loss of facility in Seoul. More than 24 hours all phones, television and other services were unavailable – Cause – Human Error
- 2011 - Microsoft and Amazon – Dublin both impacted by electrical storm, which resulted in loss of power for the full weekend – Cause – Human Error

This list shows that even leading companies with business models oriented around using their own cloud services or selling cloud services to other businesses have also experienced disruptive and costly outages. One study (Ponemon 2016) noted that 16% of Enterprise, Financial and Colocation data centers have had business-impacting outages within the past 12 months. While some of these are the result of component failures, a single component failure is rarely the root cause of an outage. Redundancy within systems, which has created a false sense of security, is not preventing these system outages. It is a cascading series of events – sometimes subtle, and sometimes like a Shakespearian comedy, converging to result in crippling downtime.

Design Standards Role in Data Center Outage

There are several data center design standards in the market, some more widely accepted than others. These design standards do have a role to play, and in many cases, they have helped to bring some common language, and common understanding of data center design intent. The standards may also provide a basis for comparing two or more data centers to each other. The intention of this article is not to delve into the standards, (or what is or is not a standard), but rather to acknowledge the contribution of standards to data center availability at the systems level. Standards have served to facilitate the evolution of a common language in the data center industry, and they allow assessment relative to data center capability metrics. Design standards focus primarily on required redundancy levels of critical systems, and largely all share four-level tiering convention:

1. Basic Components - no redundancy but has a generator
2. Redundant Capacity Components - redundant power and cooling systems
3. Concurrent Maintainability - systems can be maintained, without putting the IT load at risk
4. Fault Tolerance - ability to continue operations after one failure while repairs are undertaken

The fourth tier of any design standard generally provides for the greatest potential variation in how redundancy is achieved and is usually the most controversial as a result.

Regardless of tier, these standards primarily address ***systems*** redundancy. Telecom lines, redundant networks, generators, utility transformers, chillers, with one or two pathways for connectivity, etc. However, mechanical or system failures account for approximately 20% of outages. As further evidence of this, note that among the 10 outages previously listed, only one of these is a system issue. And to that point, of the 10 failure examples only one could have been addressed by design standards which rely on system redundancies as outage prevention.

To put this in perspective, data center failures have been clustered by standards creators into two major groups: Human Error and Non-Human Error. It is an established industry benchmark that 70% of data center outages have been shown to be the result of human error (Uptime Institute 2013, Payton 2015, and others). That would seem to suggest that the existing standards for data center design, address the remaining 30% of outages not caused by human error. But on closer examination these standards are addressing no more than 20% of remaining outage types. Like human error, these areas gain little if any attention from the standards architects. There are three other causes of outage which are: disaster (natural and in some cases, human caused), sabotage, and true accident.

True accident is not to be confused with accidents related to *human error*. True accidents are the unintended impacts of third parties which are not related in any way to the data center or its related supply chains such as utilities (power, gas or water supplies). True accidents are considered separately from human-based hazards preexisting in the operation. This is a very small area of outage, but still must be considered for completeness.

Human Error Types

Human errors come in many forms, and are not just one type of event. Error classifications vary (depending on the expert consulted), but these seven are indented as human error types for the purpose of application to the data center:

1. Decision Errors
2. Skill Errors
3. Perception Errors
4. Impairment Errors
5. Attitudinal Errors
6. Involuntary Errors
7. Violations

Decision Errors – Are mistakes made when making a decision. Budget cuts, design compromises, staffing reductions, are all classified as decision errors. This is not just about deciding to switch off the wrong breaker or enter the wrong set point. These include errors in judgement.

Skill Errors – A skill error occurs when someone makes a mistake due to insufficient capability. Imagine if a newly hired employee is handed a manual for the generator and without prior experience told to "go do the periodic maintenance".

Perception Errors – This type of error is often encountered as a "near miss", which may set the site up for failure later. It may include a perceived level of resiliency that does not exist. When near misses are ignored, or go undetected, the long-term outcome is an error in perception (a failure to recognize the confluence of events that will lead to major outage). It also includes errors in communication (someone says "B" and it is interpreted as "V", for example).

Impairment Errors – Includes difficult conditions, such as high noise area, low visibility area, as well as involve fatigue, distraction, medication or intoxication.

Attitudinal Errors – A poor attitude, usually manifesting in a lack of sufficient focus to insure adherence to procedure, or neglect are all errors of this type. May be the result of poor relationships with leadership, or other working conditions. These also include errors brought on by arrogance and bias.

Involuntary Errors – These errors include events like sneezing, reaction to loud noise, or loss of physical motor control (including the "clumsy drop"), muscle spasm, twitches, medical conditions like epilepsy.

Violations – This does not have to go hand-in-hand with attitude. Sometimes violations occur because it is not known that there is a violation occurring. This may be a willful disregard for protocol, process or procedure.

This is not intended to be an exhaustive list of the various error conditions within a type, rather to provide guidance for what each of these means. It would not be unreasonable to think that it might be possible to compile one hundred or more examples under each type.

What Factors Contribute to Human Error Increase

Human error occurs at a rate likely far greater than might be expected. On average, humans make a mistake every four minutes, according to Duffey and Saull (2008). These are actions that a person did not intend to make, and in most circumstances, are easily correctable. What is the impact of such a mistake? Dropping a pencil, tipping over a glass, misspeaking, or making a typo? These are all very common human errors, and are harmless in typical situations, but can have serious consequences under the wrong conditions.

Complexity

The more complex a task is, the greater the probability for failure. For simple tasks, the failure probability is lower and the probability of it leading to a serious escalation is lower (but not zero). Imagine performing brain surgery. There is an enormous amount that remains unknown about the human brain. Now compare that to making a cup of tea. There are things that can go wrong in tea making, but they are generally do not result in outages.

The Unknown

Complexity can be further described as "the unknown". Trial and error methods are self-descriptively error prone. When entering the unknown the probability for error is nearly 100%. Keep in mind however, the known and the unknown are contextual. An unexperienced engineer will be more likely to commit a skill based error on a complex task than a highly experienced engineer.

Temporal Pressure

The probability of error increases again when placed under time pressure. When the volume of tasks increases (even repeatable task), cognitive activity in the brain can increase exponentially, even when that load increases is small. There are two ways this can happen: 1) complete the same tasks over and over with increase in performance output (i.e. change 6 oil filters in 1 hour instead of 4 in one hour), or 2) reduce the amount of time on the single task (i.e. 15 minutes to change starter batteries instead of 25 minutes).

Both are essentially the same time pressure, but may not be perceived as the same by the person performing the task. Time pressure can be a hidden danger.

A Series of Unfortunate Events

The combination of these error types often creates extended outage; the following is an actual account of a series of events that occurred in 2014:

An incident occurred on a UPS module at a data center belonging to a financial services firm. A breaker on one UPS in a parallel, three UPS configuration (N+1 arrangement), opened on the UPS output side, resulting in the other two UPS increasing in load (mechanical failure). Operations investigated (somewhat inadequately) and could find no faults (skill error). The operations team then decided to turn the breaker back on (decision error), with the UPS running (violation). Within 2 minutes smoke began to pour from the UPS. The operations team returned to the UPS room where they found smoke. They rushed outside and quickly grabbed fire extinguishers (perception error). The extinguishers they picked up were A class (general materials fire) rated, and contained only water. They began to spray the smoking UPS, and in so doing, water made contact with circuits in the next UPS (perception error and impairment error). This caused a short-circuit in the 2nd UPS module, which then had breakers trip, causing all power to be lost to the IT load.

Ironically, the correct fire extinguishers were available, but behind a counter in the outside hall. The team, in a panic grabbed the nearest extinguishers and began "fighting the fire". Another decision error, was to have water based extinguishers in the area at all. ABC class fire extinguishers can be deployed and easily prevent such occurrence.

In the span of less than 5 minutes no less than 8 occurrences of human error are observed, each escalating the confluence of events, until outage occurred. There were other human error issues in this same incident as well, which were present for months prior to it occurring. This particular story is not uncommon, and near identical stories have been described during discussions with other operators.

Why Does One Error Lead to a Bigger Error

Consider the previous scenario, and the temporal elements previously described. One of the great problems of human error is, each error can increase the likelihood of future mistakes by contributing to *cognitive overload*. When an error occurs, usually someone is performing some action to start with.

Now they must manage the error, and still accomplish what they had originally intended. Typically, they feel responsible for their inadvertent contribution to deteriorating the health of the system, amplifying their own stress or experiencing outright panic. Understandably the person may focus on the impact of their mistake on their coworkers or employer’s perception of them, their position, etc. This multi-tasking and lack of focus increases the likelihood of committing errors, as our attention is now split among multiple tasks and issues.   Anxieties can skyrocket. Mental processing increases in both speed and intensity, while deciding what to do next. And when the situations are critical, unless the process has been practiced (to a point of mastery), the pressure to resolve both the error and complete the task assigned can become more intense. The result: Cognitive overload.

What Can Be Done

It is important to resist the urge to believe that there is any scenario where all error can be reduced to zero. It can’t. Humans are too faulty, and there are too many conditions, with combinations of variables that unexpectedly combine, resulting in outcomes that are otherwise unimaginable. Worse still, outcomes that have not yet been observed are in the “we don’t know what we don’t know” risk profile. Continued reduction is possible as new outcomes are observed, making it impossible to eliminate outage producing errors entirely.

What is needed is a new strategy. Yes, systems should be designed that are intuitive, and with safety of personnel and the systems in mind. Yes, error reduction is an important step prior to implementing error management. Once the environment has a reasonable level of reduction of error producing conditions, it is then time to begin an error management program.

Human Error (defect) versus Mechanical Failure (defect)

Human errors cannot be reduced to zero. Not even with automation. There are two reasons for this: 1) The automation is still created by humans and 2) the maintenance and upkeep of the automated and non-automated parts of the system are still carried out by humans. What is also interesting about this is the ratio of human error outages to non-human error will still be 70% - 30%. This is a critical point to understand, because the distinction of percentages is the wrong area to focus. The number of errors which lead to outages are the main point. But it is important to understand that the ratio will not change. In other words, if there are 1,000 outages recorded 700 will be human error triggered while 300 will not. If there are 10 outages, 7 will be human error triggered, while 3 will be other outage causes.

There is an easy explanation for this. The gap between mechanical failures (and other failure types) are greater than the gap between human mistakes (errors). MTBF (mean time between failure) of equipment is measured from tens of thousands to hundreds of thousands of hours, while humans make a mistake on average every four minutes.

If a mechanical device (and for this purpose, electrical/electronic devices are still mechanical), has a 5-year MTBF, that is equal to 2,628,000 minutes. With a 4 minute “defect” rate in humans, that makes human 657,000 times less reliable (more likely to produce an error) than that specific mechanical device (adjust based on MTBF for any mechanical device to work out its human versus mechanical reliability ratio). The result though has obvious ramifications. If a human has a “fault” every 4 minutes, compared to a component with an MTBF one in every thousands of hours, it is not hard to see why 70% of outages then, are caused by human error. This area may not be entirely clear though: human mistakes often have minimal or no impact. But what error is made, and what its impact will be is unpredictable, resulting in some variation but accounting for the 70% - 30% ratio over time.

Enter: Human Error Management

What surprises most people about the concept of Human Error Management (HEM) is that the error is allowed to occur, because error cannot be reduced to zero, (because frequency of an error on average every 4 minutes, does not specifically predict *when* the error will occur) the error must be managed, or intervened to produce a non-negative outcome. This is where the industry needs an entire rethink. This is not just about safety, it is not just about process or procedure. Certainly, some redesign aspects are required for it to be fully effective. While all this might sound crazy, there are already critical industries with highly mature Human Error Management practices in place.

Most obvious, the aviation industry. Because of the high visibility and horrific nature of air crashes involving large planes with large numbers of people aboard, it is no surprise that the aviation industry have pioneered much of the Human Error Management innovation. Consider flight computers. It is not hard to imagine that the cockpit of an aircraft may be very hectic in certain conditions. Weather in particular, or turbulence, or incidents of excessive chatter on the radio, or all of these combined.

Amid that, imagine the pilot must set a 3-digit airport code into the flight computer to tell the plane where to land. And that code is AAB (Arrabury airport in Queensland, Australia). And in the distraction the co-pilot instead enters ABA (Abakan Airport, Russia). This is a pretty big gap between locations. What the “intervention strategy” provides for here, is the pilot to put in the wrong value. But then it considers the plan they will have entered before, and seeing a deviation of several thousand miles will *feedback* to the pilot “That’s a long way from where we are now, are you sure?” This prompts the pilot and copilot to now cross-check the entry and ensure that the right value has been entered. The error still occurred. But now there is an intervention as a checkpoint to resolve the error with virtually no impact.

If that were not in place, and the pilots continued along the path to ABA, at some point they likely will see an airport, and if it is the right distance for their plan, maybe they decide, “That’s it, let’s land there”. (Heard of this happening?) It still happens but less often than it used to. And when it does, there is a lot of explaining to be had.

One advantage that the aviation industry has, is a shared incident reporting system. There are a few around the globe, but two or three that are the concentration of the aviation industry. This is coupled with their strategy of “Crew Resource Management” (CRM) not to be confused with getting more sales or making more business connections. CRM in the aviation world is focused on mitigating issues before they escalate.

Up until the mid-to-late 1970’s, the pilot was the “king of the craft”. This arrogance however, was betrayed by new engine systems (jet engines on commercial airlines) that were supposed to reduce the number of crashes. After a few years in, the incident rates had not gone down as expected. The number of crashes which used to be attributed to engine failures, were the result far more commonly, of human error. Weiner (1993) as well as Duffey and Saull (2008) both note the development of intervention strategies and the enablement of CRM – giving all crew members a voice, were the cause for reduction of fatal and damaging incidents in recent aviation performance.

The aviation industry illustrates that the rate of failure is still not zero, and some are other than non-human error related. But the safety record of aviation remains as the safest method of travel in the world, due largely to the implementation of CRM as an intervention strategy, an extension of Human Error Management.

What the Data Center Industry Needs to Do

First, it needs to begin sharing outage information. That means establishing a forum for the unspecified identity of the provider or owner (though it should be understood that while not publicly available, an identified person needs to lodge the issue, so that follow-up and clarity regarding the incident can be obtained, and incidents validated, while still maintaining the protected identity of those reporting, and their company). This is how the aviation industry handles this point as well.

Second, the data center industry must become self-regulated, or risk becoming regulated as aviation. That means, greater information share. One of the great values of the aviation industry is “near miss” data. That is more obvious in aviation, but just as pertinent in the data center industry. The number of “near fatal crashes” of the data center that are avoided is as much an asset to preventing outages as other observed outages. It further will enhance our understanding of the frequency and causality of data center outage.

Third, devise and distribute a framework for intervention strategy. This requires defining the framework around which interventions should be tested. One crucial issue in intervention strategy is testing strategies to ensure they do not create another problem, or make the situation worse. Interventions that put people at greater risk, even if it means a lower risk to the data center availability, are not accepted.

Fourth, with the framework in hand, engage vendors and suppliers with a focus on human-machine-interfaces that are intuitive, and error trapping. Like the flight destination, UPS, CRAC units, in particular could be made to be “error checking”. That is, simulation within their systems that first determines what the likely outcome is, and compares that with the “current course”. Artificial Intelligence may help expedite this area.

Lastly, train and drill critical operations that may have low probability of occurrence, but have highly compressed timelines, resulting in greater failure rates. One issue that always springs to mind, and again around the UPS, is switching to utility bypass. This is usually a sequence of 15 – 25 steps, and if one step is made out of turn, the IT load will be dropped (one breaker open before another is closed, and the load is lost. One breaker closed while another is still closed, causes an over amperage condition, so downstream breakers trip automatically, and the load is dropped). These types of scenarios should be taught and re-taught and rehearsed through the year. The use of engineering simulation in the near future, coupled with Augmented Reality (AR) or Virtual Reality (VR) may be reasonable approach.

Error management cannot be “trained-in” any more than quality can be “inspect-in”. An error management culture is required for an error management system to work. Aviation, petrochemical, nuclear power all have mature error management programs. It can be done through vigilance, diligence and rigor in the approach to error and outage.

References:

Duffey RB, Saull JW (2008) Managing Risk: The Human Element, *Wiley & Sons Ltd*., Singapore

Payton S (2015) Data Center Infrastructure Management – How DCIM Impacts DC Infrastructure Outage Reduction, *Unpublished MSc Dissertation*. University of Liverpool pp. 44

Ponemon Institute (2016) *Cost of Data Center Outages*, Ponemon Institute Research Report 2016, pp. 14

Uptime Institute (2013) Tier Standard: Operational Sustainability, *Uptime Institute Professional Services*, LLC New York pp. 1-16

Weiner EL (1993) Intervention Strategies for the Management of Human Error, NASA Contractor Report 4547 pp. 1-9

Biography: Scott Payton is technical director for Global Data Center Engineering. He has a critical systems background spanning more than 30 years including engineering in ICBM Missile silos with the US Air Force. He holds a Master of Science in Information Systems Management, and is currently a doctoral candidate, where his thesis topic is Human Error Management in the Data Center. Scott has previously worked for Dell as well as other data center consulting firms, and is a frequent speaker and writer on a range of data center topics including Computational Fluid Dynamics (CFD), Data Center certification, and Data Center Design and Operations standards.

Acknowledgments:

GDCE would like to graciously acknowledge the following people who had contributory input into this white paper, making it exceptional instead of ordinary: Graf Douglas, Gareth Davis, Marc Navabi, Yvette Clark