Are you being served: A Framework to manage Cloud outage repair times for Small Medium Enterprises.

Jonathan Dunne
Hamilton Institute
Maynooth University
Email: jonathan.dunne.2015@mumail.com

David Malone
Hamilton Institute
Maynooth University
Email: david.malone@nuim.ie

***Abstract*—Hosting software applications in a Cloud based infrastructure represents challenges for Small Medium Enterprises (SMEs), due to the variety of ways in which production outages can occur. We consider repair times for outage events in a framework where these downtimes are used to re-focus Systems Operations resources. Using an enterprise dataset, we address the question of how outage events are distributed and what relationship these events have with different types of failures that can occur in a cloud data centre. The proposed framework can aid SMEs to maintain a highly available On-Demand service infrastructure, with limited resources.**

I. Introduction

SMEs have seen significant growth in recent years; a 71% increase in employment (excluding financial sector) was recorded in 2014. Moreover SMEs employed almost 90 million people in Europe. [1]. As the European economy continues to recover, both businesses and clients are looking for new avenues to drive growth across the EU and beyond.

One way to provide services with an elevated market reach is through a Software as a Service (SaaS) model. This cloud-based approach is seen as a shift away from highly complex bespoke solutions, to more focused and cost effective solutions [2]. As customers demand highly effective services to solve their business problems, a cloud platform can help keep pace with these needs. A single delivery platform is used to host multiple software solutions and services.

However SMEs face a number of key challenges when embracing a cloud service model, especially in the area of reliability and maintainability. Recent work has highlighted a number of challenges, which include: outage frequency and duration. Almost all SMEs (93%) employ less than 10 people [1], therefore for this study, we analyse the factors that may impede reliability especially for businesses with low levels of resources.

In this paper we describe a framework, that the SME can use to best manage their limited pool of resources. The core idea of this framework is for cloud operations teams to focus on areas with high outage times (typically areas with high manual processes) to reduce the overall outage time. This paper contains a study of software outage data from a large enterprise dataset. Through study of this outage event data we show which types of outage events take the longest to resolve, why having standardised homogeneous data centres are key to reducing outage times, and how application types play a role in the duration of outage remediation.

For businesses who provide their cloud platform to allow companies to host services or solutions, this is known as Platform as a Service (PaaS). These providers allow for multi-tenancy. It is proposed that high-level outage data could be shared between organisations to triangulate cross application outage events.

The rest of the paper is structured in five Sections: Section II gives some description of study background and related works. Section III describes the enterprise dataset. Section IV discusses the analysis and method and it is followed by section V that explains the result. Finally, the conclusion and future work is described in Section VI.

II. Background and related research

A. Software as a Service

SaaS is defined as a delivery and licensing model in which software is used on a subscription basis (e.g. monthly, quarterly or yearly) and where applications or services are hosted centrally.

The key benefits for software vendors are the ability for software to be available on a continuous basis (on-demand) and for a single deployment pattern to be used. It is this single deployment pattern that can greatly reduce code validation times in pre-release testing, due to the homogeneous architecture. Central hosting also allows for rapid release of new features and updates through automated delivery processes [3].

SaaS is now ubiquitous, while initially adopted by the large software vendors (e.g. Amazon, Microsoft, IBM, Google and Salesforce) many SMEs are now using the cloud as their delivery platform of choice [4].

B. Cloud Outages

A cloud outage is the amount of time that a service is unavailable to the customer. While the benefits of cloud systems are well known, a key disadvantage is that when a cloud environment becomes unavailable it can take a significant

978-1-5090-3409-3/16/$31.00 ©2016 IEEE

TABLE I

Company	Outage Time	Outage Details
Verizon	40 hours	Scheduled maintenance to improve overall reliability.
Apple iCloud	12 hours	A DNS error meant that users were unable to make purchases.
Apple iCloud	7 hours	iCloud unavailable / poor performance affected 200 million users.
Windows Azure	2 hours	A network infrastructure outage resulted in loss of service for all central US users.
Starbucks	Unspecified	Scheduled maintenance resulted in the tilling system going off-line.

amount of time to diagnose and resolve the problem. During this time the platform can be unavailable for all customers.

One of the first cloud outages to make the headlines in recent times was the Amazon outage in April 2011. In summary, the Amazon cloud experienced an outage that lasted 47 hours, the root cause of the issue was a configuration change made as part of a network upgrade. While this issue would be damaging enough for Amazon alone, a number of consumers of Amazon's cloud platform (Reddit, Foresquare) were also affected. [5]

While great improvements have been made in relation to redundancy, disaster recovery and ring fencing of key critical services, the big players in cloud computing are not immune to outages. As of mid 2015 a number of high profile outages were catalogued by the CRN website. [6] Table I provides a summary.

C. Other related studies

A number of studies have been conducted in relation to cloud outages and the time observed to resolve problems in repairable systems.

Yuan et al. [7] performed a comprehensive study of distributed system failures. Their study found that almost all failures could be reproduced on reduced node architecture and that performing tests on error handling code could have prevented the majority of failures. They conclude by discussing the efficacy of their own static code checker as a way to check error-handling routines.

Hagen et al. [8] conducted a study into the root cause of the Amazon cloud outage on April 21st 2011. Their study concluded that a configuration change was made to route traffic from one router to another, while a network upgrade was conducted. The backup router did not have sufficient capacity to handle the required load. They developed a verification technique to detect change conflicts and safety constraints, within a network infrastructure prior to execution.

Li et al [9] conducted a systematic survey of public Cloud outage events. Their findings generated a framework, which classified outage root causes. Of the 78 outage events surveyed they found that the most common causes for outages included: System issues i.e. (human error, contention) and power outages being the primary root cause.

Kleyner and O'Connor [10] propose an important thesis regarding reliability engineering. While emphasis is placed on measuring reliability for both mechanical and electrical/electronic systems, the authors do broaden their scope to discuss reliability of computer software. One aspect of interest is their discussion of the lognormal distribution and its application in modelling for system reliability with wear out characteristics and for modelling the repair times of a maintained systems.

Almog [11] analysed repair data from twenty maintainable electronic systems to validate whether either the lognormal or exponential distribution would be a suitable candidate distribution to model repair times. His results showed that in 67% of datasets the lognormal distribution was a suitable fit, while the exponential was unsuitable in 62% all of datasets.

Carcary et al. [12] conducted a study into Cloud Computing adoption by Irish SMEs. The key findings of the study were as follows: Almost half the 95 SMEs surveyed had not migrated their services to the cloud. Of those SMEs that had migrated they had not assessed their readiness to adopt cloud computing. Finally the study noted that the main constraints for SMEs adoption of Cloud computing were: Security/compliance concerns, lack of IT skills and data protection concerns.

III. Data set

Cloud outage studies have been shown to provide an effective way to highlight the distribution of failure events. These studies can be leveraged by enterprises to pre-empt common failure patterns [5] [6].

The study presented in this paper examines approximately 250 field outage events from a large cloud based system. The data was collected over a 12-month period (Jan – Dec) and is comprised of four main components: E-mail, Collaboration, Social and Business Support System (BSS). Additionally the type of failure events have been categorised into the following main categories: Configuration/Manual Process, Contention/Concurrency, Disaster Recovery, Network and Hardware/Other. The systems have been deployed within three data centres and are used by customers globally. The software is developed in Java and runs on Linux.

Product development follows a Continuous delivery (CD) model whereby small amounts of functionality are released to the public on a monthly basis. For each outage event we have access to the full outage report, but we particularly focus on the time taken to resolve the outage with additional focus on the software component and the type of error, which was the root cause of the outage. The following terminology will now be defined to provide clear context. These definitions are referenced from wikipedia as no formal IEEE definitions could be obtained. [13].

- Downtime (Outage): The term downtime is used to refer to periods when a system is unavailable. Downtime or outage duration refers to a period of time that a system fails to provide or perform its primary function.
- Maintenance window: In information technology and systems management, a maintenance window is a period

of time designated in advance by the technical staff, during which preventive maintenance that could cause disruption of service may be performed.
- Tiger Team: A tiger team is a group of experts assigned to investigate and/or solve technical or systemic problems.

This study aims to answer a number of questions. First, How are the times of cloud outage events distributed? Second, does the distribution vary by component? Third, does the distribution differ by failure category? Fourth, does the relationship differ by data centre? In order to answer these four questions, this study is broken down into the following attributes: outage distribution, outage component, outage failure category and data centre location.

A. Outage Distribution

Probability distributions are used in statistics to assign a likelihood of an event-taking place. In the case of cloud outage events, by analysing the distribution of all events, it may be possible to fit a known distribution to our dataset. If a distribution can be fitted, these distribution properties can be used to infer the most likely outcome of an outage event. For example a probability distribution could be used to infer the likelihood of an outage event taking a specific period of time to resolve. An outage distribution is plotted for the complete set of outages. For distribution validation we used the R library ADGofTest [14] against a number of likely distributions types. (e.g. Exponential, Gamma, lognormal and Weibull)

B. Outage Component

Recognising the location of an outage event at a component level gives an understanding of a) which components are more likely to contribute to an outage event and b) the relative duration to detect and resolve an outage with respect to a component. For example, operations teams may have various probes to determine if an event is likely to cause a failure. Development and test teams may have a suite of test cases to find a certain class of issue. Outage events can provide operations teams with an understanding of potential gaps in their probes and monitoring solutions. Likewise for development and test teams outage events can provide both teams with either weaknesses in feature implementation and gaps in test coverage. Depending on the nature of these test gaps and the size of the test organisation, they may be difficult to close. In each data centre there are four main specific components: BSS, collaboration, e-mail and social. For this study we categorised our software components as follows: BSS/social, collaboration, e-mail and mixed (where multiple components where involved).

C. Outage Type

Over the course of our study, we found a variety of outage events. To give clarity to these different types of outage event, we divided the outages into five main categories: Configuration/Manual, Contention/Concurrency, Disaster Recovery, Network and Hardware/Other

Configuration/Manual errors involve situations where a configuration change is made from one value to another which causes a piece of infrastructure to behave abnormally. For example a Load Balancer setting could be changed manually which reduces the throughput from Gigabits to Megabits which could greatly reduce the infrastructures's ability to manage incoming traffic.

Contention/Concurrency outages refer to a class of issue, which is triggered through normal operations on the underlying server component code. These issues are triggered due to the inability of the code to handle either concurrent or parallel usage. Software defects may include issues related to contention under load (e.g. memory leaks, high Disk I/O, CPU usage), concurrency (e.g. deadlocks) or miscellaneous race conditions.

Disaster Recovery errors typically involve scenarios where system load was required to move from one application server or database to another. In some situations the session data may not transfer correctly and cause a piece of infrastructure to become unavailable.

A network error relates to a class of failure outside of misconfiguration or a hardware failure within the network infrastructure. Network failures can typically present themselves as intermittent temporary network outages, high latency/packet loss conditions or congestion based on overloading of available bandwidth. As cloud data centres contain a number of distributed systems, having a reliable network infrastructure is highly desirable.

A Hardware/Other failure relates to a class of problem, which causes a piece of hardware to fail. These failures relate to a malfunction within the electronic circuits or electromechanical components (disks, tapes) of a computer system. Recovery from a hardware failure requires repair or replacement of the offending part. Additionally the error may relate to some miscellaneous type of error that is not part of the four main failure categories.

D. Data Centre Location

Understanding the measure of outage events at a data centre level can highlight whether a specific data centre is a factor in the duration and distribution of outage events raised. There are three data centres in our dataset: data centre A (High usage), data centre B (Low usage) and data centre C (Medium usage). Having a correlation between outage duration can be a useful data point for cloud operations teams.

E. Limitations of dataset

The dataset has a number of practical limitations, which are now discussed. While the outage event tracking system allows for a granular categorisation system, whereby outages can be mapped to a subcomponent, there are a number of outages, which due to their severe nature can affect more than one component and subsystem. The authors reviewed the functional location of each defect to ensure precision across the analysis of the dataset. In a number of limited cases outages affected a more than one component and data centre at a time.

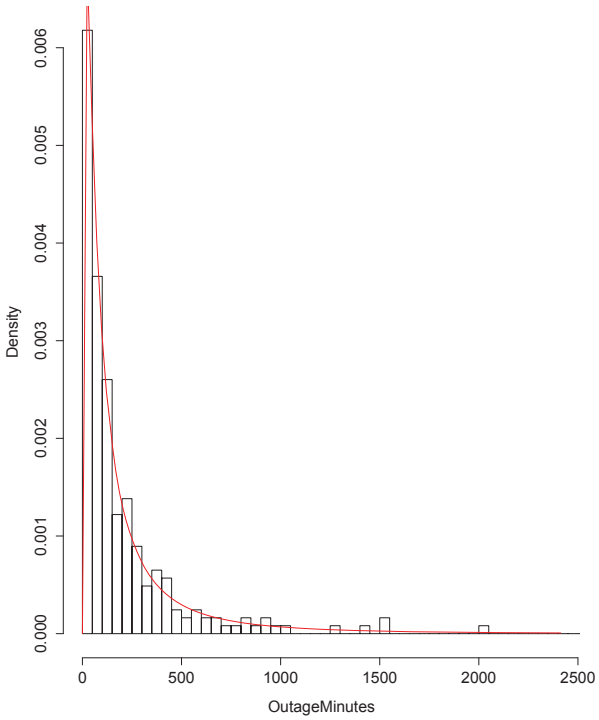

Fig. 1. Histogram of Outage Times (In Minutes) with fitted lognormal Curve

In the case of mixed component outages, summary analysis was performed. However due to the borderline number of samples, in the case of mixed data centre outages, analysis was not performed.

The outage events that form part of this study are from an enterprise cloud system. The outage events are applicable to the software domain of BSS, Collaboration, Email and social. Additionally the outage events are applicable to the field of Configuration/Manual, Contention/Concurrency, Disaster Recovery, Network and Hardware types. As a consequence the analysis may not be relevant outside of these fields.

IV. Results

We now explore the attributes of outage events observed.

A. Outage Distribution

Fig. 1 shows a probability density function histogram for all 246 outage events with a fitted lognormal curve. Table II lists the measure of location (i.e. mean, standard deviation, median and skew) for all outage events along with the distribution type and an Anderson-Darling goodness of fit p-value.

B. Outage Component

Table III lists the summary statistics of outage events broken down by Component. E-Mail recorded the highest proportion of all outages. BSS/social recorded had the lowest. Outages are most likely to happen in the E-mail component.

TABLE II

Statistic	Value
Samples	246
Mean	314
Std Dev	1414
Median	105
Skew	13.80
Distribution	lognormal
AD GoF Test (p)	0.95

TABLE III
SUMMARY STATISTICS FOR OUTAGES BY COMPONENT WITH LOGNORMAL GOF

Statistic	BSS/Social	Collaboration	Email	Mixed
Samples	16	34	152	43
% Samples	7	14	62	17
Mean	274	189	258	626
Std Dev	639	379	423	3261
Median	45	61.5	126.5	85
Skew	3.56	3.83	5.45	6.30
AD GoF (p)	0.69	0.62	0.99	0.64

Due to the small number of samples (16) recorded for the BSS/social category, the goodness of fit value should be treated with caution.

C. Outage Type

Table IV lists the summary statistics of outage events broken down by type. Configuration/Manual and Contention/Concurrency recorded the highest proportion of outages while Hardware/Other had the lowest. Outages are most likely to be either Configuration/Manual or Contention/Concurrency.

Due to the smallnumber of samples (23) for the Hardware/Other category, the goodness of fit value should be treated with caution.

D. Data Centre Location

Table V lists the summary statistics of outage events broken down by data centre. Data Centre A recorded the highest proportion of outages, while Data centre B had the lowest. The remaining 3% were from outages found in all thee data centres. Outages are most likely to happen within Data Centre A.

TABLE IV
SUMMARY STATISTICS FOR OUTAGES BY TYPE WITH LOGNORMAL GOF

Statistic	Configuration Manual	Contention Concurrency	Disaster Recovery	Network	Hardware Other
# Samples	74	64	36	49	23
% Samples	30	26	15	20	9
Mean	488	239	134	315	243
Std Dev	2488	469	161	591	358
Median	114.5	86	72	145	91
Skew	8.28	3.69	2.33	5.30	2.11
AD GoF	0.91	0.97	0.94	0.75	0.96

TABLE V
SUMMARY STATISTICS FOR OUTAGES BY DATA CENTRE WITH LOGNORMAL GOF

Statistic	Data Centre A	Data Centre B	Data Centre C
Samples	160	24	54
% Samples	65	10	22
Mean	224	188	645
Std Dev	313	280	2961
Median	113.5	89.5	79.5
Skew	2.93	2.89	6.67
AD GoF (p)	0.99	0.99	0.31

Eight outages were found in all three data centres. Due to the small number of samples detailed analysis was not performed. Furthermore the researches felt it in was inappropriate to merge these eight samples into one of the existing data centre pools as this may confound analysis and results from a single data centre category.

V. DISCUSSION

Section IV provided an outline of outage events that were studied as part of our overall dataset, including distribution, component, outage type and data centre location. The following section provides deeper analysis of the results. In each section references will be made to each research question asked in section III.

A. Outage Distribution

To answer the question how are the times of cloud outage events distributed, an Anderson Darling goodness of fit test was conducted against a number of distributions; Exponential, Gamma, lognormal and Weibull. With the exception of lognormal, the p values were very low, which indicated that these distribution types were a poor fit. For lognormal a p value was found to be 0.95.

In this case the hypothesis that the outage times are Log normally distributed is a surprisingly good fit. Fig. 1 and Table II clearly show that the distribution type is lognormal. This finding further expands the applicability of the use of the lognormal distribution of model repair times. It is known that repairable systems typically refer to mechanical, electric and electronic systems. However given the above results we can now include software systems as another subtype.

It is also worth noting that the mean outage time is approximately 314 minutes, which indicates that resolution of an outage in complex system architecture is a non-trivial task. Additionally with a standard deviation found to be approximately 1414 minutes and a skew value of 13.80, clearly indicates that there is a high level of dispersion within the dataset.

Given the nature of cloud computing, new code updates and configuration changes are made on a regular basis. It is not uncommon for an enterprise to introduce changes on a bi-weekly or monthly basis. Therefore with this high level of system activity it is not unsurprising that outages can occur frquently. If a state of the art outage tracking system were introduced, it would be interesting to determine overall as both process improvements were made coupled with underlying code stability to observe the overall affect on both the distribution type and shape. This would provide a concrete answer to questions such as: what impact do specific process improvements make to overall outage times? As a business where do resources need to be deployed to improve platform stability: Development, Operations or Quality Assurance?

B. Outage Component

Examining outages by component can give insight as to which component are likely to exhibit outages and whether these times vary by component.

Table III provides summary details of outages by component. It was noted that mixed components had the highest mean outage time with 627 minutes, followed by BSS/social, Mail with 274 & 258 respectively and Collaboration with 189 minutes. Consequently mixed components also has the highest standard deviation and skew. In all cases, each component class had a good fit to a lognormal distribution, with the e-mail category fitting best with a p value of 0.99. However the BBS-Social category has a low number of samples, therefore the goodness of fit assessment should be treated with some caution.

Based on these results it is clear that the e-mail component has the highest proportion of outages. Tiger teams should review the root cause of each outage related to the e-mail component. This will gain understanding as to the what types of failures contribute most to e-mail outage events. Triangulating each outage event against the failure type and data centre location can help business and operations teams resource their crisis teams on a per component basis.

Secondly the results show that mixed components have the highest mean outage time. This result seems logical, given that when an outage occurs across common infrastructure and/or multiple components that the repair time is greater. There are many systems to check and repair as part of the remediation process. To verify, the individual reports were checked for mixed component failures. It was found that one outage took multiple days to resolve. Hence skewing the overall mean time. While this data point may be considered an outlier in the classic sense, given this was a real fault, it must be included as part of analysis. Tiger teams should determine the the root cause of each outage to intersect failure type and data centre to understand common failure patterns.

C. Outage Type

Examining outages by type gives a deeper understanding of what types of problems are likely to cause an outage within a cloud infrastructure. Table IV provides this insight.

Significantly Configuration/Manual had the highest expected outage time with 489 minutes, with Network next highest with 315 minutes. Contention/Concurrency, Hardware and Disaster recovery had expected outages times of 239, 243 and 134 minutes respectively. Finally the outage times of each category were fitted with a lognormal distribution. In each case the hypothesis of whether a lognormal distribution

was a suitable distribution could not be rejected. However one caveat is that the Hardware/Other category had a low number of samples, so this result must be treated with caution.

It is clear that issues related to Configuration/Manual contribute most to the overall number of outages but also take the longest to resolve. Given the relative complexity of the overall cloud architecture it is apparent that a system of managed configuration changes is required. Firstly to ensure that for all configuration changes made, that there is a commit and rollback feature to ensure that that harmful (extreme) configuration settings can be reversed if required. Additionally tiger teams should implement a system, which can monitor real-time configuration changes across all data centres.

With any distributed system the network health plays an important role in system stability. The network issues studied fell into two main categories: network congestion and temporary network outages. For congestion issues, business and operations teams need to define clear bandwidth capacity requirements to ensure that their infrastructure has the bandwidth to meet the demands of their existing user base and future subscription signings. The underlying application code and middleware stack should have additional resiliency added to ensure that temporary outages do not cause cascade failures.

D. Data Centre Location

Table V provides summary details of outages by data centre. As discussed previously in section III, user concurrency varies by data centre. Data Centre C had the highest mean outage time with 645 minutes, while data centre's A & B had mean outages times of 224 and 188 respectively. All three data centre outage times were modelled with a lognormal distribution and both data centre A & B were an excellent fit. Each had a p value of 0.99. Data centre C faired worst in terms of fit with a p value of 0.31. Even with this value the hypothesis of whether a log normal distribution is a suitable fit cannot be rejected.

In some ways the above results are expected, it seems intuitive that a high use data centre would incur the most outages due to the high level of customer activity, however even with all these outages the mean outage time is 224 minutes, which is approximately 90 minutes less than the overall mean. What appears somewhat counter intuitive is that data centre C has the second highest number of outages and the highest mean outage time. From closer inspection the mean outage time of data centre C is due to a small number of outages with high durations. Finally it is worth noting that Data centre B has the lowest number of outage events and the lowest mean outage time.

In the context of software delivery to multiple data centres, the same code is released to each system. Clearly customers are impacted in different ways depending on which data centre is used. With this knowledge, tiger teams can investigate in two areas. Firstly does the underlying customer use case of each data centre vary? Secondly a root and branch investigation of each data centre configuration should be conducted and compared for discrepancies, with specific focus on the configuration of the e-mail component.

VI. Conclusion

The purpose of this study was to examine the duration of outage events within a Cloud based application platform. It was found that the lognormal distribution is a useful distribution for modelling repair times of SaaS Outages. The findings of this study support previous research particularly in the field of system reliability and repair times.

Previous studies have shown that shown that Cloud Outages are an infrequent occurrence. Additionally we show that the lognormal distribution is a useful tool for modelling repair times in mechanical and electronic maintainable systems.

This work provides a more comprehensive study in relation to how outage times can vary between failure type, component and the data centre used at the time of an outage.

In future SMEs can assess their outage data to understand the core issues that effect their underlying service platform. A specific operations framework can then be developed to allow SMEs to focus on specific areas of their architecture or business process, which impede reliability. Likewise, by usage of this framework on an iterative basis, an SME can then set realistic remediation targets.

In future work we shall assess our framework in conjunction with the time between outage events to understand how best operations teams can be deployed where parallel outage events occur.

References

[1] P. Muller, C. Caliandro, V. Peycheva, D. Gagliardi, C. Marzocchi, R. Ramlogan, and D. Cox. (2015) SME performance review European SME's. [Online]. Available: http://bit.ly/23NnKIx

[2] (2015) Why multi-tenancy is key to successful and sustainable software-as-a-service (SaaS). [Online]. Available: http://bit.ly/1vyAKDb

[3] (2015) From Google to Amazon - the rise of the cloud catalog. [Online]. Available: http://bit.ly/1S5elbl

[4] (2015) Pole position: Ranking the top 5 IaaS, PaaS and private cloud providers. [Online]. Available: http://bit.ly/1UQCaSf

[5] (2015) The 10 worst cloud outages. [Online]. Available: http://bit.ly/1ISiawO

[6] (2015) The 10 biggest cloud outages of 2015. [Online]. Available: http://bit.ly/1QhMiiR

[7] D. Yuan, Y. Luo, X. Zhuang, G. R. Rodrigues, X. Zhao, Y. Zhang, P. U. Jain, and M. Stumm, "Simple testing can prevent most critical failures: An analysis of production failures in distributed data-intensive systems," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 249–265.

[8] S. Hagen, M. Seibold, and A. Kemper, "Efficient verification of it change operations or: How we could have prevented amazon's cloud outage," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*. IEEE, 2012, pp. 368–376.

[9] Z. Li, M. Liang, L. O'Brien, and H. Zhang, "The cloud's cloudy moment: A systematic survey of public cloud service outage," *arXiv preprint arXiv:1312.6485*, 2013.

[10] P. O'Connor and A. Kleyner, *Practical reliability engineering*. John Wiley & Sons, 2011.

[11] R. Almog, "A study of the application of the lognormal distribution to corrective maintenance repair time," Ph.D. dissertation, Monterey, California. Naval Postgraduate School, 1979.

[12] M. Carcary, E. Doherty, and G. Conway, "The adoption of cloud computing by Irish SMEs–an exploratory study," *Electronic Journal Information Systems Evaluation Volume*, vol. 17, no. 1, 2014.

[13] (2015) Wikipedia - the free encyclopedia. [Online]. Available: https://en.wikipedia.org/wiki

[14] C. J. G. Bellosta. Package adgoftest. [Online]. Available: http://bit.ly/1NU3c5y