Obscured by the cloud: A resource allocation framework to model cloud outage events☆

Jonathan Dunne[a,*], David Malone[a]

[a]*Hamilton Institute, Maynooth University, Kildare, Ireland*

Abstract

As SME's adopt cloud technologies and rapid delivery models as a means to provide high value customer offers, there is a clear focus on uptime. Cloud outages represent a challenge to an SME's to deliver and maintain a services platform. If a Cloud platform suffers from downtime this can have a negative on business revenue. Additionally outages can divert resources from product development/delivery tasks to reactive remediation. These challenges are more immediate to an SME's with a small pool of resources at their disposal. Therefore it is necessary to develop a framework which can be used to model the arrival of cloud outage events. A framework which can be used by cloud Operations teams to manage their scarce pool of resources to resolve outages, therby minimising impact to service delivery. This article considers existing modelling techniques such as the M/M/1 queue, and proposes a more accurate approach. We first calculate the inter arrival and service distributions. Next we formally test for dependence between event arrivals. We then model a series of outage events in a G/G/1 queue with a Monte Carlo simulation to determine queue busy time. Finally we compare the precision of our framework against an M/M/1 simulation and real outage event data. The results demonstrated that our framework can improve the estimation of cloud outage events and aid DevOps resource planning.

Keywords: Outage simulation, Resource allocation model, Queuing Theory

1. Introduction

Cloud outage prediction and resolution is an important activity in the management of a cloud service. Recent media reports have documented cases of cloud outages from high profile cloud service providers [1]. During 2016 alone the CRN website has documented ten highest profile cloud outages to have occurred so far. Due to the increasing complex nature of the data centre infrastructure coupled with the rapid continuous delivery of incremental software updates it seems that cloud outages are with us for the time being.

For operations teams that maintain a cloud infrastructure, they rely on state of the art monitoring and alert systems to determine when an outage occurs. Examples of monitoring solutions include: New Relic, IBM Predictive Insights and Ruxit. Once a new outage is observed, depending on the outage type (e.g. Software component, infrastructure, hardware etc) additional relevant experts may be called to remediate the issue. The time taken to resolve the issue may depend on a number of factors: ability to find the relevant expert, swift problem diagnosis and velocity of the pushing a fix to production systems.

Both SME's and micro teams within large organisation's face a number of challenges when adopting a cloud platform and a mechanism to deliver products and services. A number of recent studies have outlined that both frequency and duration of outage events are key challenges. Almost all European SME's (93%) employ less than ten people [2]. Ensuring that adequate skills and resources are available to accommodate incoming outage events is highly desirable.

Downtime is bad for business. Whether the company provides a hosting platform, more commonly known as Platform as a Service (PaaS) or for a company that consumes such a platform to deliver their own services, more commonly known as Software as a Service (SaaS). The end result is the same: Business disruption, lost revenue, recovery/remediation costs etc. A recent US study which looked at the cost of data centre downtimes, calculated the mean cost to be $5617 per minute of downtime [3].

In preparation for this study, a thorough search of current literature was conducted. No frameworks to model cloud outage events were found. This study observed that outage events arrive over a period of time, which require fixing to return a system to a steady state. With these attributes in mind, our literature search focuses on queuing theory and distribution fitting for repairable systems.

Another consideration is the idea of event dependence. Typical off the shelf single-server queue models such as M/M/1 and G/G/1 assume that the inter-arrival and service times between events are independent [4]. However if some form of dependence is found between events how

*Corresponding author

Email address: jonathan.dunne.2015@mumail.ie (Jonathan Dunne)

useful would a queuing model which assumes independence compare against that of a queuing model with dependence properties.

In this paper we propose a framework that a micro team or SME can leverage to best manage their existing resource pool. The core idea of this framework is for operations teams to use a special case of the G/G/1 queue to model the inter arrival, service times of outage events. Our study shows that the special case of G/G/1 delivers a high degree of accuracy compared to off the shelf queuing models like M/M/1. Additionally our framework also considers independence between successive outage events. This article consists of a study of outage event data from a large enterprise dataset. By analysis of this outage event data this study shows the efficacy of the G/G/1 special case and how it can be used to reasonable model cloud outage event data. Additionally we highlight the shortcomings of the M/M/1 queue specifically in the realm of service times of repairable systems, such as cloud based platforms. Finally this study highlights how independence/dependence between cloud outage's plays a role in the frameworks precision.

To help researchers reproduce and extend the work conducted as part of this study, pseudo-code of the queue modelling framework is provided . By utilising this queue framework, researches will have the ability to test their preferred case of the G/G/1 model against the M/M/1 model.

The rest of this article is divided up into the following sections: Section 2 introduces background and related work. Section 3 discusses the data set collected (and associated study terminology), outlines the research questions that are answered by this study and the limitations of the dataset. Section 4 outlines the experimental approach and associated results. Section 5 discusses the results of our experiments. Finally in Section 6, we conclude this paper and discuss future work.

2. Background and related work

The following section provides some background information on both Software & Platform as a service, followed by a section on high profile cloud outages that have made the media headlines. Finally this section concludes with a in-depth look at related studies in the field of repairable systems modelling, queuing theory and cloud outage studies.

2.1. 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 [5].

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 [6].

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 [7].

2.2. Platform as a Service

PaaS is defined as a delivery and platform management model. This model allows customers to develop and maintain cloud based software and services without the need of building and managing a complex cloud based infrastructure.

The main attraction of PaaS is that it allows SME's to rapidly develop and deliver could based software and services. While focusing on their core products and services SMEs are less distracted by having to design, build and service a large complex cloud based infrastructure.

However one drawback of PaaS is that an SME may not have a full view of the wider infrastructure. Therefore if an outage event occurs at an infrastructure level (e.g. Network, Loadbalancer) an SME may be unaware of the problem until the problem is reported by a customer.

Many companies now offer PaaS as their core service. Once seen as the preserve of a large organisation (e.g. Amazon EC2, Google Apps and IBM Bluemix) a number of smaller dedicated companies also offer PaaS (e.g. Dokku, OpenShift and Kubernetes) [8]

2.3. High profile 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 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. [9]

Dropbox experienced two widespread outages during 2013 [10, 11]. The first in January, users were unable to connect to the service. It took Dropbox 15 hours to restore a full service. No official explanation as to the nature of the outage was given. The second occurred in May, again users were unable to connect to the service. This outage lasted a mere 90 minutes. Again no official explanation was provided.

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 2016 a number of high profile outages were catalogued by the CRN website. [1] Table I provides a summary.

2.4. Other related studies

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

Yuan et al. [12] 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. [13] 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 [14] 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.

A number of recent studies have focused on how network reliability and hardware failures contribute to cloud outages, these are discussed briefly.

Sedaghat et al [15] modelled correlated failures caused by both network and power failures. As part of the study the authors developed a reliability model and an approximation technique for assessing a services reliability in the presence of correlated failures.

Potharaju and Navendu [16] conducted a similar study in relation to network outages, with focus on categorising intra and inter data centre network failures. Two key findings include: Network redundancy is most effective at inter-datacentre level and interface errors, hardware failures and unexpected reboots dominate root cause determination.

Bodik et al [17] analysed the network communication of a large scale web application. Then proposed a framework which achieves high fault tolerance with reduced bandwidth usage in outage conditions.

Synder et al [18] conducted a study on the reliability of cloud based systems.The authors developed an algorithm based on a non-sequential Monte Carlo Simulation to evaluate the reliability of large scale cloud systems. The authors found that by intelligently allocating the correct types of virtual machine instances, overall cloud reliability can be maintained with a high degree of precision.

Kenney [19] proposes a model to estimate the arrival of field defects based on the number of software defects found during in-house testing. The model is based on the Weibull distribution which arises from the assumption that field usage of commercial software increases as a power function of time. If we think of cloud outages as a form of field defect, there is much to consider in this model.

Kleyner and O'Connor [20] 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 [21] 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.

Adedigba [22] analysed the service times from a help desk call centre. Her study showed that the exponential distribution did not provide a reasonable fit for call centre service times. However a log-normal distribution was a reasonable fit for overall service times. Her study also showed that a phase-type distribution with three phases provided a reasonable fit for service times for specific jobs within the call centre job queue.

John [23] discusses dependencies between inter-arrival and services times within a queue system as the assumption of independence between the two times are not always valid.

Carcary et al. [24] 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.

Gholami et al. [25] provided a detailed review of current cloud migration processes. One of the main migration concerns mentioned was the unpredictability of a cloud environment. Factors that led to this unpredictability included: Network outages and middleware failures. The study concluded that a fixed migration approach is not possible to cover all migration scenarios due to architecture heterogeneity.

Table 1: Summary of high profile Cloud outages in 2016

Company	Duration	Date	Outage Details
Office 365	Several days	18th Jan	Users reported issues being able to access their cloud based mail services. The issue was identified and a software fix was applied. This fix proved unsuccessful, thereafter a secondary fix was developed and applied which was successful.
Twitter	8 hours	19th Jan	Users experienced general operational problems after an internal software update was added to the production system with faulty code. It took Twitter 8 hours to debug and remediate the defective code.
Salesforce	10 hours	3rd March	European Salesforce users had their services disrupted due to a storage problem in their EU Data Centre. After the storage issue was resolved, users reported performance degradation.
Symantec	24 Hours	11th April	A portal to allow customers to manage their cloud security services became unavailable. The exact nature of the outage was undisclosed. Symantec were require to restore and configure a database to bring the system back online.
Amazon	10 hours	4th June	Local storms in Australia caused Amazon Web Services to lose power. This resulted in a number of EC2 instances to fail, which affected both SaaS and PaaS customers.

3. Data set and research methodology

4. Results

5. Discussion

6. Conclusion

References

[1] The 10 biggest cloud outages of 2016 (2016).
URL http://bit.ly/2bjsPGL

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

[3] Calculating the cost of data center outages (2011).
URL http://bit.ly/2bInDgh

[4] M/M/1 Queue (2015).
URL https://en.wikipedia.org/wiki/M/M/1_queue

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

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

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

[8] Best platform as a service (PaaS) (2016).
URL http://bit.ly/2bavsb5

[9] The 10 worst cloud outages (2015).
URL http://bit.ly/1ISiawO

[10] Dropbox outage represents first major cloud outage of 2013 (2013).
URL http://bit.ly/2bjFdla

[11] Dropbox currently experiencing widespread service outage (2013).
URL http://tcrn.ch/2bDEyM5

[12] D. Yuan, Y. Luo, X. Zhuang, G. R. Rodrigues, X. Zhao, Y. Zhang, P. U. Jain, 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.

[13] S. Hagen, M. Seibold, 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.

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

[15] M. Sedaghat, E. Wadbro, J. Wilkes, S. De Luna, O. Seleznjev, E. Elmroth, Die-hard: Reliable scheduling to survive correlated failures in cloud data centers.

[16] R. Potharaju, N. Jain, When the network crumbles: An empirical study of cloud network failures and their impact on services, in: Proceedings of the 4th annual Symposium on Cloud Computing, ACM, 2013, p. 15.

[17] P. Bodík, I. Menache, M. Chowdhury, P. Mani, D. A. Maltz, I. Stoica, Surviving failures in bandwidth-constrained datacenters, in: Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication, ACM, 2012, pp. 431–442.

[18] B. Snyder, J. Ringenberg, R. Green, V. Devabhaktuni, M. Alam, Evaluation and design of highly reliable and highly utilized cloud computing systems, Journal of Cloud Computing 4 (1) (2015) 1.

[19] G. Q. Kenny, Estimating defects in commercial software during operational use, IEEE Transactions on Reliability 42 (1) (1993) 107–115.

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

[21] R. Almog, A study of the application of the lognormal distribution to corrective maintenance repair time, Ph.D. thesis, Monterey, California. Naval Postgraduate School (1979).

[22] A. Adedigba, Statistical distributions for service times, Ph.D. thesis, Citeseer (2005).

[23] F. I. John, Single server queues with dependent service and inter-arrival times, Journal of the Society for Industrial and Applied Mathematics 11 (3) (1963) 526–534.

[24] M. Carcary, E. Doherty, G. Conway, The adoption of cloud computing by Irish SMEs–an exploratory study, Electronic Journal Information Systems Evaluation Volume 17.

[25] M. F. Gholami, F. Daneshgar, G. Low, G. Beydoun, Cloud migration processa survey, evaluation framework, and open challenges, Journal of Systems and Software 120 (2016) 31–69.