IBIMA Publishing
Journal of Cloud Computing
http://www.ibimapublishing.com/journals/JCC/jcc.html
Vol. 2014 (2014), Article ID 586109, 8 pages
DOI: 10.5171/2014.586109

Research Article

Selling the Cloud to Smaller Business Organisations

D.J. Collins and K.P. Lam

School of Computing and Mathematics, Keele University, UK

Received date: 5 September 2013; Accepted date: 30 January 2014; Published date: 21 March 2014

Academic Editor: K.P. Lam

Abstract

We examine the problems faced by an IT Sector SME in planning the migration of existing Web Applications to the Cloud. Potential performance evaluation proved sufficiently complex, costly and imprecise as to result in postponement or cancellation of the migration, and plans for a more attractive alternative. We argue that traditional benchmarking techniques, even those being developed for cloud use, cannot provide potential cloud users with the information required for sound business planning. We propose a system of vendor performance agreements based upon the automated characterisation of customer applications.

Keywords: cloud computing, benchmarking, cloud resource allocation, transaction processing

Introduction

Cloud service provision is rare amongst technological products and services in that it is sold without meaningful performance guarantees (explicit or implied). Service Level Agreements (SLAs) generally relate solely to availability and the description of resource units is vague at best (we will justify this assertion later in the document).

Statistics on adoption rates for cloud computing vary considerably. Part of this variation is accounted for by poor definition of what constitutes the cloud – for example, the simple use of ISP hosting may be considered to be encompassed by the term. However, across most surveys, adoption rates by small to medium sized enterprises (SMEs) fall behind those of larger organisations – see, for example, Microsoft's SMB Cloud Adoption Study (Microsoft, 2011). Additionally, smaller organisations are more likely to be using free rather than paid for services than is the case with their larger counterparts (Microsoft, 2011).

Theoretically, SMEs are the most likely organisations to benefit from cloud provision. Internal IT departments are likely to be small or non-existent and do not enjoy the economies of scale available to larger companies. Furthermore, staff and infrastructure upgrades will require significant investment in relative terms and carry higher levels of risk. In a small organisation, investments in infrastructure,

Cite this Article as: D.J. Collins and K.P. Lam (2014)," Selling the Cloud to Smaller Business Organisations", Journal of Cloud Computing , Vol. 2014 (2014), Article ID 586109, DOI: 10.5171/2014.586109

mail servers, generic business applications etc. are substantial costs in relation to turnover.

The problem is more profound for SMEs whose business is primarily delivered by IT. Here we present a case study based upon our close involvement with a single SME[1] providing employee vetting solutions as they migrated from a largely manual paper-based system to a highly virtualised online solution. Thereafter we observed their strategy in planning maximal cloud deployment (IaaS and elements of PaaS).

The Problem

The company acts on behalf of customers (employers) to manage the process of vetting potential employees (applicants). This involves providing a variety of checks, most of which can now be conducted online using official sources (such has the Driver Vehicle Licensing Authority) and third-party information processors such as credit-checking agencies. Essentially, vetting forms are purchased by *customers* and released to *applicants*. Applicants complete them and following authentication checks by the customers, sub-sets of the gathered information are sent to the aforementioned external sources for vetting. Some third-party checks are asynchronous, others synchronous. The results are aggregated and presented to the customer in a form that supports their subsequent decision making process regarding the applicant.

It should be evident from this description that the process can be near fully automated. Following the construction of the initial online system, further development effort is largely restricted to incorporating additional checks (with some horizontal diversification) and responding to the changing interface specifications of the external agencies involved. In consequence, if the application is capable of scaling linearly in terms of volumes, costs are highly predictable and the basis of a sound business model exists.

To this end, the company engaged in a great deal of virtualisation (although certain external factors placed some minor constraints on their ability to do this). Much of this virtualisation was in-house, and some at two major Internet Service Providers utilised by the company.

However, the company continued to experience a number of business problems. Trading volumes expanded fairly rapidly leading to repeated requirements for increased staffing and infrastructure. Naturally, these were staged events which initially had a high impact on unit costs and overall profitability and thus presented high levels of risk for a small company. The company CEO was aware of the "Cloud" and the claims made for it, notably 'The ability to scale-up IT capacity on demand' and 'The ability to align IT resources with cost', which are generic claims made by most cloud service providers. If such claims were true, risk could be mitigated and the business could be better focussed on expansion.

Evaluating Potential Solutions

From the CEO's perspective, the decision to be made was one of finding the cloud service provider capable of delivering a desired level of service at the lowest cost whilst ensuring scalability of the applications. This, of course, had to be achieved with minimal development effort mitigating against significant changes to the established applications. The project thus sought to:

a) Characterise the existing applications

b) Establish resource requirements necessary to maintain SLAs at:

(i) Current Volumes

(ii) 2 X current Volumes

(iii) 3 X current volumes

c) Compare the costs/benefits of selected suppliers

a) **Characterising the Existing Application**

The company had very clear service level agreements in place regarding availability and latencies. In order to achieve this, there was a high level of log analysis – response times, query-times, I/O times, CPU usage,

storage loadings etc. Given three major categories of users (administrative, employers and applicants), all of the analysis was performed against usage categories – the ratio between which had historically been extremely stable. They were thus armed with a fairly comprehensive but voluminous characterisation of the application under different usage patterns and loads. The process of analysis for potential cloud deployment produced some excitement with the realisation that usage for two categories of user was largely restricted to between the hours of 09.00 and 17.00 for five days a week with massive under-utilisation of resources outside of these periods (See Figure 1). Furthermore, within a typical day there were quite narrowly defined morning and afternoon peaks. Of course, derivation of any benefits from this usage pattern would depend upon the granularity of the elasticity of resource provisioning offered by prospective suppliers.

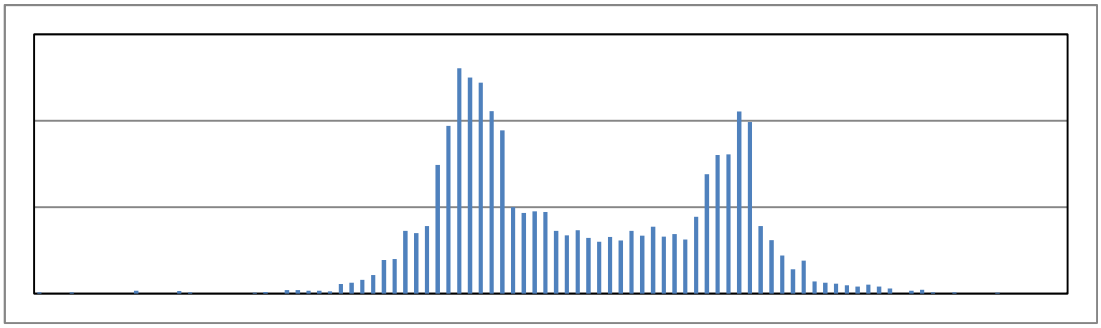

Figure 1. Total Sessions (00.00 – 24.00hrs) User Category 1 and 2 Average over one month period

***b*) Establishing Resource Requirements**

Leaving aside the system database and associated servers, resource requirements could easily be established in terms of physical CPU capability, memory, storage and I/O. Furthermore this could be achieved for different anticipated usage levels accommodating the desired 2X and 3X expansion. The database provision was more problematic, but based upon past infrastructure expansion decisions; server specifications were produced in which the technical team had confidence, albeit with a possible requirement for further query optimisation and minor index/table restructuring. Ultimately, there was a database sharding plan which provided sufficient insurance as to allow a high level of confidence in the resultant specification. The existing system maintained a high level of redundancy in order to provide for both expansion of 50% and disaster recovery. The requirements specifications made similar provisions for the three target volumes. Certain assumptions were made regarding the stability of third-party APIs and the ratios and relative activities of the three user categories.

c) Comparing Costs/Benefits of Selected Suppliers

In order to address this problem, it was first necessary to identify potential suppliers. There is now a huge range of cloud providers and aggregators and an objective means was sought of determining the likely best candidates. This proved to be a far more difficult problem than anticipated. Gartner is usually regarded as a reliable source of information within the industry with assessments based upon routinely conducted surveys. However, there was imperfect correspondence between Gartner's results (Gartner, 2012) and those of other surveys. There have been some attempts to more objectively and empirically assess cloud providers. For example, *Compuware* (http://www.compuware.com/) provides a service which continuously monitors a reference application running in each of the 'major' cloud service providers worldwide

(Molyneaux, 2009). Results are available in terms of availability and response-time statistics updated in real time. This revealed that at least one provider had failed to meet its SLA targets within the previous week and also suggested considerable response-time variability – both within and between providers. The company considered the Compuware tool to be the most objective measure and used it to select an initial list of potential providers.

Based upon the analysis performed with the Compuware tool, the following providers were identified as those providing the best response times in the seven days prior to the analysis - UMBEE, ElasticHosts, BlueSquare, Qube, Netcetera, Rackspace UK, BT Global Services, VOXEL (EU Netherlands), Windows AZURE, LUNACloud (EU France), Dimension Data (EU Netherlands), Amazon EC2 (EU Ireland) and CloudSigma (EU Switzerland). Based upon the Compuware data, there was considerable variation, with the worst average response time of this elite group being some 300% of the fastest for the simple HTTP based reference application accessed from the London spine. The study company was UK based with customer access almost exclusively from the UK.

In gathering information on all aspects of the services of each provider, their on-line promotions were utilised, an approach adopted by the University of Surrey in their *Fair Benchmarking for Cloud Computing Systems* study (Gilam, 2013). At this stage, consultations with sales and/or technical staff were considered likely to be too time consuming. The information was gathered over a one month period during early 2013. It is possible that there were commercial and/or technical changes occurring during this period that are not reflected in subsequent narrative.

There has been some improvement in the transparency of charges made by the major hosting providers over the past few years. Most employ some form of interactive cost calculator (which were utilised for this study). However, there were variations in practice, with some companies for example, having a fixed inclusive element - say for outgoing bandwidth - and others having only a unit charge. Some companies had a 'base system' with an associated cost – to which unit priced resources could be added. All quoted hourly rates for incremental resource units. Some companies continued to represent systems in terms of relative capabilities (x small, small, medium, large, x large etc), often with reference to a 'base' machine. For example Amazon cites a 2006 Xeon running at 1.7GHz as an ECU which they further equate to a 2007 Opteron or Xeon processor running at between 1.0 and 1.2 GHz (note the introduction of ambiguity). These variations made exact comparisons extremely difficult.

The company, with limited time, produced a best estimate of the costs to meet the resource requirements for each of the above providers. The figures are not presented here – they are undoubtedly inaccurate and would likely be unfair to one or more providers. However, they did represent the best efforts of the company to perform a task whose complexity was magnified by the non-uniform promotional strategies of cloud providers. The company further considered the complexities that might be involved in migration and the capabilities of each provider to expand and contract resource provisioning with the finest level of granularity possible. Security, data integrity, data privacy and other regulatory issues were also concerns.

Provisioning costs across the companies varied considerably: the highest estimate being considerably more than double the lowest estimate assuming provisioning based upon peak demands within a 24 hour cycle. However, even the highest estimate was some 20% less than the 'in house' cost based on a three year capital depreciation cycle. The cost of software licences was included and an assumption was made that staffing would continue at existing levels and cost. At face value, the case for cloud migration thus seemed compelling.

However, the complexities of migration had to be considered, together with possible effects on SLAs and other issues. The most attractive solution, based upon both overt cost and the level of elasticity granulation

was Amazon EC2. The ability to provision instances 'in minutes' rather than 'hours' was particularly attractive. The company therefore prioritised EC2 in its further evaluation. Based upon the literature available and costing in development time necessary to overcome perceived constraints in the Amazon offering (notably persistence problems, MySql hosting limitations on EBS (Elastic Block Store) and the necessity to use software based RAID), EC2 was excluded. Observationally, this may have been misguided but the business sought certainty and the migration to EC2 afforded unacceptable risks from the perspective of the CEO based upon the published information available. Limited literature searches supported the perception that EBS I/O performance was too variable, particularly with live replication of the master database. Refer, for example, to (Robertson, 2011).

Testing

Two further providers were selected for consideration, having met the 'minimal development cost' constraint, meeting regulatory requirements regarding security and data privacy, and falling at the lower end of the cost estimates. We will refer to these as CPA and CPB. Initial literature searches suggested that for both providers' performance variation was less than 5% which could be 'budgeted in' if necessary. Testing was destined to be difficult given the asynchronous nature of the applications and the involvement of third party services. However, the timings involved for such services were known and hence could be simulated.

Images were created for the two selected services. This was not trivial, given the need for simulation, but the VMware based virtualisation already employed by the company provided an advanced starting point. The critical latencies in the application centred on a number of complex queries. The database comprised some 90 tables and certain critical sessions could involve staged queries involving between 15 and 20 tables. These were thus targeted for testing together with more mundane but high volume transactional sessions. At this stage the MySql server was not migrated, a decision based upon the fact that live replication to a separate provider was a resilience requirement and hence the I/O capabilities were crucial.

During the first eight hours of testing, both providers were failing to meet targeted response times with load levels at 50% of peak. Furthermore, there was high variability in performance (as measured by response times) which was most marked with low instance levels. One of the providers reported a problem during testing and provided notification of the necessity for a machine reboot. Both providers were contacted through online messaging. CPA took a look at the configuration and reported nothing amiss, although did admit to a 'temporary high volumes of traffic'. CPB insisted that a support ticket be raised which would be dealt with within 24 hours. The experience produced such a lack of confidence in both providers that the testing (which was costly) was abandoned.

Lessons Learnt

The evaluation of services offered by competing cloud service providers is not trivial. Marketing information published by providers highlights potential benefits but offers little information which might assist potential customers in exercising informed judgements. Whilst all providers offer uptime SLAs (often merely guaranteeing to not to charge for hours which fall below the SLA figure), none give any performance guarantees for CPU and I/O. There is a range of comparative studies available, for example CloudSpectator, 2013, but they are often associated with one or more providers, or are based upon a single reference application/test suite which may inadequately represent the optimal ratio of resource allocation suited to a target application. With the exception of Amazon's EC2, the granularity of resource allocation in all of the providers considered is not only coarse but generally also requires manual intervention, additional configuration input information and significant time. Most of the research into ensuring elasticity of response is largely to the benefit of providers rather than consumers of cloud services; see, for

example, (Bennani, 2005) and (Menasc'e, 2009).

Resource allocation in the physical world is problematic enough and is often responsive – in the sense that it is a reaction to a perceived impending problem. For example, a lengthening of response latency produced by long query times might be addressed by a combination of query optimisation, additional indices, expansion of database server RAM, increased SSD storage and bandwidth expansion. In the virtual world, where the performance of these elements (or their substitutes) is ambiguous, the problem is exacerbated. In deciding whether to migrate to the cloud, organisations are in forward-planning rather than responsive mode and will therefore be more cautious and less experimental.

In selecting infrastructure and platforms in the physical world, organisations have the benefit of an extensive range of benchmarking services. They are particularly well established in high performance computing, for example the HPC Challenge Benchmark Suite (Luszczek, 2006). The need for Cloud benchmarking is well established (Luszczek, 2011) and the Transaction Processing Council is working on a new benchmark for assessing transaction based applications on Cloud infrastructure and platforms (Nambier, 2013). Although there have been some experimental tools (Calheiros, 2010), these are complex and are beyond the reach of most human resource challenged SMEs.

In the embedded computing world the characterisation of applications for optimal hardware configuration is reasonably well established (Sanna, 2009). The problem in doing so for web-based applications differs little conceptually. What we need as a starting point is a means of characterising applications and their usage in terms of the optimal combinations of elements for differing scales of processing, subject to automatically identified upper (infrastructure or platform) limits. With adequate benchmarking of cloud IaaS and PaaS services coupled such automated characterisation of web applications, we would have the possibility of tools permitting automatic configuration of cloud resources which would not only assist migrators but also serve as a basis for vendor provided resource planning and performance guarantees.

Conclusion

At present, the development effort required to evaluate cloud service providers for SMEs with established web-based applications is too high. Furthermore, on the basis of the evidence that we have reviewed (over a period of 12 months), large-volume applications with large numbers of machine instances are assured less variability in performance levels and consequently smaller organisations are necessarily disadvantaged. Variability in performance coupled with lack of clarity in instance performance specification and realisation make it extremely difficult to plan resource usage. The level of (timely) granularity in resource responsiveness is not sufficiently fine to allow smaller companies to benefit significantly from cloud economies. Much of the work on improving resource elasticity is for the benefit of service providers rather than consumers who continue to be offered rather quaint 'compute units' with (with the exception of EC2) impractical methods of adjusting resource allocation. More work needs to be done on the automatic characterisation of applications and thereby the automatic determination of cloud resource allocation with associated performance guarantees within defined limits.For the company concerned in the study, the revelation that resource usage was concentrated in an eight hour block offered possibilities of far clearer benefit than cloud migration. They sought an agreement with a co-located company dealing with consumer sales in Asia. Perhaps it is time to add the term *Cooperative Cloud* to the ever expanding cloud glossary. We are presently engaged in work on the characterisation of Web 2.0 applications to permit the automatic generation of reference models for comparative evaluation across cloud service providers.

Note

[1]Earlier collaboration had attracted two funding awards: i) "*Digital-media Authenticated Electronic Disclosure Application System*", UWSP Advantage Proof of Concept (POC), 2009-10. ii) "*A Practical framework for the development and Evaluation of Multi-factored Authentication Schemes for Secure Distributed Systems*", EPSRC/UK CASE Studentship Award (EP/H501320/1), 2009-2012.

References

1. Bennani, M.N. and Menasc´e, D.A.(2005), 'Resource Allocationfor Autonomic Data Centers Using Analytic Performance Models', Proceedings of 2005 IEEE International Conference on Autonomic Computing, Seattle, WA,

2. Calheiros, R.N.; Ranjan, R.; Beloglazo, A.; De Rose, C.A.F.; Buyya, R., (2010), '*CloudSim*: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms', Software Practice and Experience, 2011; 41:23–50, Published online 24 August 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.995.

3. Cloud Spectator, (2013), 'Cloud Server Performance - A Comparative Analysis of 5 Large Cloud IaaS Providers', Online, (at CloudSpectator,last), last accessed on 02-08-2013, available at ,http://www.cloudspectator.com/wp-content/uploads/2013/06/Cloud-Computing-Performance-A-Comparative-Analysis-of-5-Large-Cloud-IaaS-Providers.pdf,

4. Gartner, (2012), 'Critical Capabilities for Public Cloud Storage Services', Online, Gartner,Last Accessed 02/08/2013, Corrected 10/01/2013, http://www.gartner.com/technology/reprints.do?id=1-1D9C6ZM&ct=121216&st=sg.

5. Gillam L., Li B., O'Loughlin J., Tomar A.P.S (2013), 'Fair Benchmarking for Cloud Computing Systems', Journal of Cloud Computing: Advances, Systems and Applications 2013, 2:6 7th March, 2013

6. Maatta, S. ; Indrusiak, L.S. ; Ost, L. ; Moller, L. ; Nurmi, J. ; Glesner, M. ; Moraes, F, (2009) 'Characterising embedded applications using a UML profile', Proceedings of the 11th international conference on System-on-chip (SOC'09), IEEE Press, Piscataway, NJ, USA, 172-175.

7. Microsoft, (2010), 'SMB Cloud Adoption Study Dec 2010', Online,Microsoft, Last accessed on 02/09/2013,http://www.microsoft.com/en-us/news/presskits/telecom/docs/smbstudy_032011.pdf.

8. Molyneaux, I., (2009), The Art of Application Performance Testing: Help for Programmers and Quality Assurance, 1st edition, O'Reily Media.

9. Robertson, K, (2011), 'Our Pain Points with EC2 and how our Move solved them, available online at InvalidLogic, last accessed on 15/08/2013, http://invalidlogic.com/2011/02/16/our-pain-points-with-ec2/ .

10. Menasc´e D.A. and Ngo P. (2009), 'Understanding Cloud Computing: Experimentation and Capacity Planning', CMG, Proceedings of 2009 Computer Measurement Group, Dallas, Texas, December 2009

11. Luszczek, P., Bailey, D., Dongarra, J., Kepner, J., Lucas, R., Rabenseifner, R., Takahashi, D. (2006), 'The HPC Challenge (HPCC) Benchmark Suite', SC06 Conference Tutorial, ACM/IEEE SC2006 Conference on High Performance Networking and Computing, Tampa, Florida, November 12, 2006.

12. Luszczek, P., Meek, E., Moore, S., Terpstra, D., Weaver, V., Dongarra, J. (2011), 'Evaluation of the HPC Challenge Benchmarks in Virtualized Environments', Proceedings of 6th Workshop on Virtualization in High-Performance Cloud Computing, Bordeaux, France, August 30, 2011.

13. Nambiar, R. et al, (2012), 'TPC Benchmark Roadmap 2012', Springer, Selected Topics in Performance Evaluation and Benchmarking, Lecture Notes in Computer Science Volume 7755, 2013, pp 1-20