Business-driven Configuration of IT Services in Public and Hybrid Clouds based on Performance Forecasting

Genady Ya. Grabarnik
Dept. of Math & CS
St. John's University
Queens, NY, USA
grabarng@stjohns.edu

Mauro Tortonesi
Dept. of Engineering
University of Ferrara
Ferrara, Italy
mauro.tortonesi@unife.it

Larisa Shwartz
Operational Innovation
IBM TJ Watson Research Center
NY, USA
lshwart@us.ibm.com

***Abstract*—Modern Cloud computing environments are rapidly evolving, leading to a growing adoption of dynamic pricing for virtual resources and of speedier deployment tools, and to the emergence of hybrid Cloud scenarios. These trends suggest the opportunity to investigate a new generation of Cloud-based IT services, capable of adapting to changes in their operating conditions and deployment environment by dynamically realigning their configuration. This calls for new and more sophisticated management tools, that are capable of evaluating the performance of alternative configurations for Cloud-based IT services and of identifying the one that aligns better to the objectives defined by the business management. This paper presents an optimization tool for Cloud-based IT services, based on queuing theoretic analysis of service workflows and ILP optimization.**

***Keywords*—Resource allocation, Cloud, business processes**

I. Introduction

Cloud computing is evolving from a utility computing paradigm to a more sophisticated market of virtual resources with dynamic pricing schemes. In addition, recent improvements of virtualization technologies are providing new tools, e.g., containers, that enable to substantially decrease the time required for a service component (de)instantiation [1]. Finally, *hybrid Cloud* scenarios are becoming increasingly important, thus presenting even greater challenges from the service management perspective [2].

These trends suggest the opportunity to investigate *highly dynamic and adaptive Cloud based IT services*, that can dynamically realign their configuration to match the ever changing characteristics of modern Cloud environments. This ambitious objective calls for the development of new and sophisticated service management tools, that are capable of evaluating the performance of current IT service configuration as well as forecasting how alternative configurations would perform under the same circumstances.

However, given the complex nature of modern IT services, that implement a large number of workflows on top of many software components of different types, estimating the impact that even a simple reconfiguration of the IT service architecture, e.g., the deployment of a new VM to host a replica of a software component, will have on the entire IT service performance represents a major challenge. In addition, management tools should adopt comprehensive business-driven IT management criteria that enable to identify the IT service configurations that are best suited for the delivering the service quality stipulated with Service Level Agreements (SLAs) with the minimum amount of spending for resource acquisition [3].

This paper presents a business-driven optimization tool for Cloud-based IT services, based on stochastic violation penalty estimation, leverages the $M/\Sigma M/1$ queuing theoretic model that we purposely developed to capture the behavior of IT service workflows, the approximation of the distribution of residual times for $M/\Sigma M/1$ queues through Gamma distribution, and ILP optimization. The approch we followed in this paper represents a different and complementary perspective with respect to our earlier work [4].

II. Stochastic Modeling and Analysis of Cloud-based IT Services

To allow an analytically tractable IT service model, we formulate a few simplifying assumptions. First, we assume that each software component has exponentially distributed service times and that software components operates independently from the other - both practices that are commonly proposed in literature. Then, we assume that requests are processing in a first-in-first-out (FIFO) order and that no arriving request is dropped.

This means that the IT service operates as a $M/G/1$ queueing system where the processing times $B(t)$ are the result of a sequence of steps $B_i(t)$ with exponentially distributed arrival times [5]. This queue is similar to a $M/E_r/1$ queue, a specialization of $M/G/1$, but with different distribution parameters [6]. We could not find a name for the system in literature, so we adopt the $M/\Sigma M/1$ nomenclature, to stress that serving part of the model is sum of exponential random variables.

Since we are interested in estimating the distribution $B(t)$ of the service times for the entire workflow. To sum the contributions of each element to the workflow, we adopt the mechanisms introduced in [7] and [8]. Let us first consider a workflow request traversing a set of software components in a sequential fashion, and requests are inserted in a FIFO queue in front of the component sequence and extracted when the components are idle. The service times for the entire

workflow will therefore be the sum of the service times for each component.

Since we are interested in estimating the distribution $B(t)$ of the service times for the entire workflow, it is convenient to switch to frequency domain analysis. The distribution of a sum of n random variables $S = X_1 + X_2 + \cdots + X_n$ is convolution of the distributions of addends X_i:

$$F_S(t) = (F_{X_1} * F_{X_2} * F_{X_3} * \cdots * F_{X_n})(t). \tag{1}$$

The assumption of exponential distribution for the random variables X_i enables to significantly simplify equation (1), leading to [9]:

Lemma II.1. *Let $X_1, X_2, \ldots, X_n$ be independent exponentially distributed random variables with parameters $0 < \lambda_1, \lambda_2, \ldots, \lambda_n$ respectively. Then their sum $S = \sum_{i=1}^{n} X_i$ has cumulative distribution function $B(t)$ corresponding to the following probability distribution function:*

$$B'(t) = f_S(t) = \sum_{i=1}^{n} \lambda_i \prod_{\substack{j=\overline{1,n} \\ j \neq i}} \frac{\lambda_j}{\lambda_j - \lambda_i} e^{-\lambda_j t} \tag{2}$$

and:

Proposition II.2. *The processing time $B(t)$ of the sequential workflow with components having exponentially distributed response time with parameters $0 < \lambda_1, \lambda_2, \ldots, \lambda_n$ has distribution density described by equation* (2)*.*

We can therefore use well-known results on $M/G/1$ queues to calculate the distribution of response times, that also consider waiting times in addition to the processing times [5].

Proposition II.3. *The distribution of the waiting times for the $M/\Sigma M/1$ queue is:*

$$W(t) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n (R^{(n)} * B)(t), \tag{3}$$

where B is the processing time distribution, R is the residual processing time distribution defined as

$$R(s) = \lambda \int_0^s [1 - B(u)] du$$

and $R^{(n)}(s)$ is n-th convolution of the residual processing time distribution corresponding to the processing of the n requests.

A. Laplace transforms and Gamma approximation

The direct evaluation of equation (3) is usually performed through approximation, by limiting the calculation to the component of order n in the sum. In turn, the single components $R^{(n)} * B$ are calculated either through the direct evaluation of the convolution or through the application of the Laplace transform.

In our attempts to directly evaluate convolutions, we encountered severe oscillations in the calculated density function of $W(t)$, as illustrated by Fig 1. The figure shows how the oscillations start to become evident in the curve of order 9

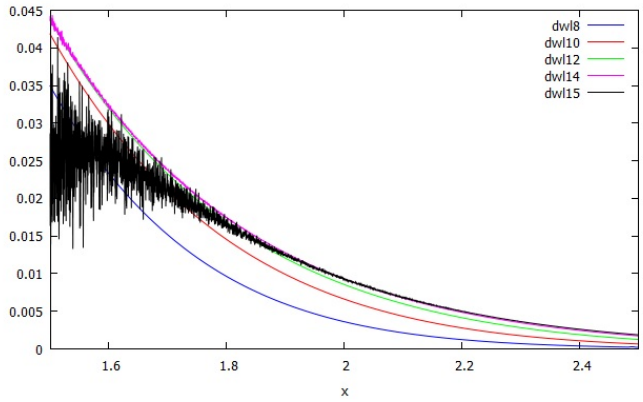

Fig. 1. Approximation of waiting time distribution through direct evaluation of convolutions (*dwl-n* represents the approximation of order *n*).

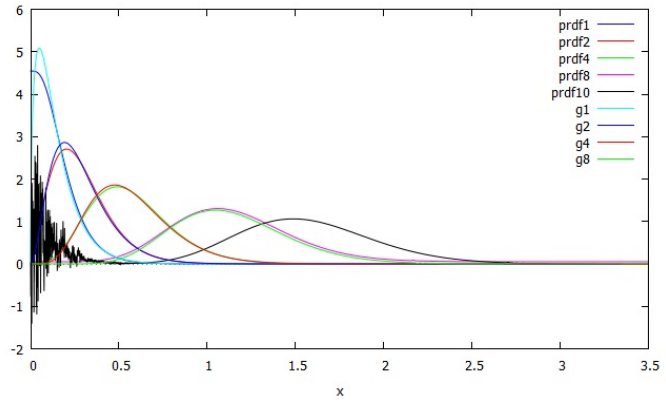

Fig. 2. Numerical calculation of the $R^{(n)}$ component through Laplace transform (*prdf-n*) and Γ distribution based approximation (*g-n*).

and become excessively high for practical application in the curve of order 15.

Similar situation arises in case of evaluating (3) using Laplace transforms. For the convolutions of order $n >= 9$, we encountered the loss of significant digits as mentioned in [10], due to the fact that Laplace transform has an (absolutely continuous) spectrum coinciding with the interval $[-\sqrt{\pi}, \sqrt{\pi}]$, and generalized eigenvalues of the form $\lambda_{\pm s} = \pm\sqrt{\frac{\pi}{cosh(\pi s)}}$. This implies that inverse Laplace transform is unbounded, thus leading to the fast growth of increasing osculations and, hence, loss of significant digits. This may also indirectly explain what we observed while evaluating convolution of order $n > 9$.

To avoid the loss of precision we used a more robust calculation approach based on the approximation of the residual distribution through the Γ (Gamma) distribution. We chose the Γ distribution as it allows a good approximation and has the very convenient property that convolutions of Γ are easy to calculate. Fig. 2 illustrates the quickly improving approximation of the $R^{(n)}$ component by appropriate Γ distributions and high oscillation in the component of order 10 evaluated using Laplace transforms.

We use the method of moments to infer the parameters of the approximating Γ distribution. In a $M/G/1$ queue with

service time distribution B, the moment of order m of the residual service time distribution R is [6]:

$$\mathbb{E}(R^m) = \frac{\mathbb{E}(B^{m+1})}{(m+1)\,\mathbb{E}(B)}. \quad (4)$$

Since the n-th convolution of the residual service time distribution $R^{(n)}$ corresponds to sum of n independent random variables each having distribution R, we have:

$$\mathbb{E}(R^{(n)}) = n\,\mathbb{E}(R) = n\frac{\mathbb{E}(B^2)}{2\,\mathbb{E}(B)}, \quad (5)$$

$$Var(R^{(n)}) = nVar(R) = n\frac{\mathbb{E}(B^3)}{3\,\mathbb{E}(B)} - n(\frac{\mathbb{E}(B^2)}{2\,\mathbb{E}(B)})^2. \quad (6)$$

Formulae (5) and (6) highlight the elegance of the approximation of $R^{(n)}$ with Γ distributions. Once we found the parameters of Γ through the method of moments, then the parameters of the convoluted Γ (which are very easy to calculate) correspond to the convolution of R approximation found by the method of moments. Therefore, we can formulate the following:

Lemma II.4. 1) *The distribution of waiting times for the $M/\Sigma M/1$ queue is approximated by:*

$$\tilde{W}(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n(\Gamma(n\alpha,\beta) * B)(t), \quad (7)$$

where

$$\beta = \frac{\mathbb{E}(B^2)}{\mathbb{E}(B^3)}; \quad \alpha = \frac{(\mathbb{E}(B^2))^2}{\mathbb{E}(B)\,\mathbb{E}(B^3)},$$

2) *Using the method of moments, the α and β parameters in equation* (7) *can be inferred by the following equations:*

$$\beta = \frac{p_2 + p_1^2}{2p_3 + p_1p_2 + p_1^3}; \quad \alpha = \frac{(p_2+p_1^2)^2}{p_1(2p_3 + p_1p_2 + p_1^3)},$$

where

$$p_k = \sum_{i=1}^{m}\frac{1}{\lambda_i^k}$$

and $1/\lambda_i$ is the average processing time for i-th component.

B. Non-sequential workflows and non-exponential service times

The results presented above can be extended to the case of generic, i.e., non-sequential, workflows as defined in [11]. To this end, it is possible to use the composition formulas presented in [12] and approach developed in [7] [8].

In addition, the method described above works even in case the assumption of exponential service times is relaxed. More general distributions and workflows evaluation of the processing time may be considered using simulation, obtaining the resulting response time distribution and then separating the request arrival time and processing time components.

The full exploration of these research directions goes beyond the scope of this paper and is left for future work.

C. Estimation of SLA violation probabilities and costs

We are then interested in devising a method that, given an SLA, estimates the expected amount of violation penalties for a service configuration starting from the approximated waiting time distribution $\tilde{W}$ obtained from equation (3).

For instance, let us consider an SLA S_1 that triggers a violation in case the 95th percentile of service requests is larger than 2 seconds or in case the maximum service request time could be greater than 5 seconds. We then define:

$$p_1 = F_{\tilde{W}}(T > 2secs)$$

and:

$$p_2 = F_{\tilde{W}}(T > 5secs).$$

To evaluate the probability that for $N = 10000$ requests at most a portion $0 < r < 1$ (or rN requests) will have response time greater than p_1, we use the binomial distribution:

$$p_{r,N,p_1} = \Pr(Y > rN) = 1 - \sum_{i=0}^{rN}\binom{n}{i}p_1^i\,(1-p_1)^{n-i}, \quad (8)$$

where Y is the number of requests whose execution requires a time larger than 2 seconds. The expected loss is then the product of SLA violation cost and p_{r,N,p_1}.

D. Constrained Optimization

The results of equations (7) and (8) can be tabulated and used as input in a constrained optimization problem (COP), whose solution represents the optimal configuration for the IT service. We solve the constrained optimization problem using the CPLEX solver, formulating it as follows.

We assume to have a predefined set of data center types DCT and server types ST, and characterize each data center $DC[DCT][ST]$ as a structure containing DCT and a number $SC[DCT][ST]$ of available servers. The cost for VM instantiation is described by $CS[DCT][ST]$ cost per data center and VM type. The different worflow components are described by their type COT. Different workflows are described by their type FlT, which in turn is a sequence of COTs of length $LFlT$.

The constraints for the instantiation of software components in VM types that support their minimum resource requirements are modeled by the boolean matrix $CTR[ST][COT]$. The expected loss per deployment per workflow is tabulated by the matrix $L[FlT][ST][COT]$. The IT service workload is a mixture of workflows $M[FlT]$. A configuration deploys a set of components in the union of the workflow per workload. A deployment is represented by the boolean dynamic variable $x[DCT][ST][SC]$.

The COP can then be formulated as the minimization of workload deployment cost satisfying constraints on VM

instantiation:

$$min \sum_{\substack{i \ in \\ FlT}} M[i] \times \sum_{\substack{(k,l,m) \ in \\ DCT \times ST \times SC}} (x[k][l][m] * CS[k][l] + L[i][l][m]) \tag{9}$$

subject to

$$x[k][l][m] \leq CTR[l][m] \tag{10}$$

$$\sum_{\substack{(k,l,m) \ in \\ DCT \times ST \times SC}} x[k][l][m] == \sum_{i \ in \ FlT} LFlT[i]. \tag{11}$$

As CPLEX allows to start the optimization from a priming point, effectively implementing a continuous optimization process, it is possible to envise the adoption of a component that could monitor the current IT service state and trigger a new optimization if needed. At the same time, it is possible to envise an actuator component that takes in input the result of the optimization process and reconfigures the IT service accordingly.

III. Experimental Results

To demonstrate the effectiveness of our IT service configuation optimization solution, we consider a single sequential 4-step workflow, where requests go through a Web Server (WS), an Application Server (AS), a Financial Transaction Server (FTS), and finally a Persistent/Database Storage layer (PS).

We assume that the processing times for each component are exponentially distributed, with the service rate parameter λ_i varying for each component type and VM type used for its instantiation. We consider a processing response time for the WS component instantiated on "small" VMs of 0.045 seconds, for "medium" VMs of 0.030 seconds, etc. We also assume that on "large" VMs the response times are: for AS - 0.060 seconds, for FTS - 0.090, for PS - 0.025 seconds.

Table I shows a (tiny) portion of the results we obtained, that illustrates the dependency of the expected loss due to SLA violations from the service times, which in turn depend on VM type used to instantiate the components.

TABLE I. Sample of losses per workflow deployment/response times

Service rate (secs)				Expected penalty (USD)
WS	AS	FTS	PS	
0.01	0.045	0.04	0.02	0.0 $
0.01	0.06	0.09	0.025	$4.338*10^{-9}$ $
0.01	0.08	0.09	0.025	1.081365 $
0.030	0.06	0.09	0.025	0.002012 $
0.045	0.06	0.09	0.025	8.718848 $
0.045	0.08	0.09	0.025	6698.311 $

After calculating the distribution function W of waiting times, we set $N = 10000, r = 0.0075$ and use equation (8), meaning that we use binomial distribution to estimate the probability of having more than 0.75% of response times within a given time period (e.g., 1 hour) to be over 2 seconds. The traffic intensity $\rho = 0.65$ corresponds to about 3 requests per second or over 10000 per hour. The probability of having over 75 requests in over 2 seconds category is about 0.000872, and the expected loss due to SLA violation penalties amounting to 10000$ in this case is slightly less than 9$.

We use VM costs from Table III in [4] as VM costs CS and apply COP from section II-D to find statistically optimal deployment of workflow. The optimal configuration we obtained corresponds to the 4th line in Table I {WS - medium, AS - large, FTS - large, PS - large} deployed at US East datacenter with the expected cost of deployment around $0.85$$ per hour.

In this experiment, it is clear that using using more powerful/expensive VMs will just increase IT related spending; at the same time the adoption of less powerful VMs will increase costs related to expected SLA violation penalties.

IV. Conclusions and Future Work

The present paper represents a first exploration in the optimization of Cloud-based IT service configuration through queuing theoretic models and ILP based optimization. The experimental results we presented demonstrated the effectiveness of our solution when applied to sequential service workflows. Future work will focus on the application of the techniques and tools introduced in this paper for the optimization of IT services that implement complex and non-sequential workflows.

References

[1] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An Updated Performance Comparison of Virtual Machines and Linux Containers," IBM Research, Tech. Rep. RC25482 (AUS1407-001), 2014.

[2] K. Bai, N. Ge, H. Jamjoom, E. Ea-Jan, L. Renganarayana, and X. Zhang, "What to Discover Before Migrating to the Cloud," in *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*. IEEE, 2013, pp. 320–327.

[3] A. Moura, J. Sauve, and C. Bartolini, "Business-driven IT management - upping the ante of IT : exploring the linkage between IT and business to improve both IT and business results," *Communications Magazine, IEEE*, vol. 46, no. 10, pp. 148–153, October 2008.

[4] G. Grabarnik, L. Shwartz, and M. Tortonesi, "Business-driven optimization of component placement for complex services in federated Clouds," in *Network Operations and Management Symposium (NOMS 2014), 2014 IEEE/IFIP*. IEEE, 2014, pp. 1–9.

[5] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of queueing theory*. John Wiley & Sons, 2013.

[6] I. Adan and J. Resing, *Queueing Theory*. Eindhoven University of Technology. Department of Mathematics and Computing Science, 2001.

[7] G. Grabarnik, H. Ludwig, and L. Shwartz, "Management of service process qos in a service provider-service supplier environment," in *E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2007. CEC/EEE 2007. The 9th IEEE International Conference on*. IEEE, 2007, pp. 543–550.

[8] ——, "Dynamic management of outsourced service processes qos in a service provider-service supplier environment," in *Business-driven IT Management, 2008. BDIM 2008. 3rd IEEE/IFIP International Workshop on*. IEEE, 2008, pp. 81–88.

[9] M. Bibinger, "Notes on the sum and maximum of independent exponentially distributed random variables with different scale parameters," *arXiv preprint arXiv:1307.3945*, 2013.

[10] C. L. Epstein and J. Schotland, "The bad truth about Laplace's transform," *SIAM review*, vol. 50, no. 3, pp. 504–520, 2008.

[11] W. van der Aalst, "The application of Petri nets to workflow management," *Journal of circuits, systems, and computers*, vol. 8, no. 01, pp. 21–66, 1998.

[12] M. Lomonosov, "Bernoulli scheme with closure," *Problemy Peredachi Informatsii*, vol. 10, no. 1, pp. 91–101, 1974.