

Resource Optimization Subject to a Percentile Response Time SLA for Enterprise Computing

Kaiqi Xiong, Harry Perros
Department of Computer Science
NC State University, Raleigh, NC 27965-7534
{xiong, hp}@csc.ncsu.edu

Abstract—We consider a set of computer resources used by a service provider to host enterprise applications subject to service level agreements. We present an approach for resource optimization in such an environment that minimizes the total cost of computer resources used by a service provider for an enterprise application while satisfying the QoS metric that the response time for executing service requests is statistically bounded. That is, $\gamma\%$ of the time the response time is less than a pre-defined value. This QoS metric is more realistic than the mean response time typically used in the literature. Numerical results show the applicability of the approach and validate its accuracy.

I. INTRODUCTION

The increasing pervasiveness of network connectivity and the proliferation of on demand e-business applications and services in public domains, corporate networks, as well as home environments give rise to the need for the design of appropriate service management solutions.

In this paper, we consider a collection of computer resources used by a service provider to host enterprise applications for business customers. An enterprise application running in such a computing environment is associated with a service level agreement (SLA) [8], [9] and [17]. That is, the service provider is required to execute service requests from a customer within negotiated QoS requirements for a given price. Figure 1 depicts a scenario for such an environment. A customer represents a business that generates service requests at a given rate to be processed by the service provider's resource stations according to QoS requirements and for a given fee. As shown in Figure 1 a service request is transmitted to the service provider over a network provider. After it is processed at the various resource stations of the service provider the final result is sent back to the customer. For presentation purposes, we assume that each resource station has only one type of server associated with cost c_j . If they have multiple types of servers, we can decompose each resource station into several individual stations so that each one only contains one type of server with the same cost.

Let N_j be the number of servers at station j ($j = 1, 2, \dots, m$). Thus, the resource allocation is quantified by solving for n_j ($n_j = 1, 2, \dots, N_j$) in the following optimization problem:

$$I = \min_{n_1, \dots, n_m} (n_1 c_1 + \dots + n_m c_m) \quad (1)$$

subject to SLA constraints. Performance and price are the two most important components for a variety of SLAs for business applications. Hence, in this paper, we consider the minimization of the overall cost of the service provider's computing resources allocated to the business customer so that $\gamma\%$ of the time the response time, i.e., the time to execute a service request, is less than a pre-defined value. Typically, in the literature the response time is taken into account through its mean [1], [9], [11] and [14]. However, this may not be a meaningful quality of service as far as the customer is concerned, who may be more interested in a statistical bound of the response time.

Resource optimization problems subject to performance metrics such as response time, bandwidth, and link utilization have been extensively studied. It was studied for multiple packet networks in [3], IP networks in [10], wireless networks in [2], grid computing in [11] and optical networks in [13]. The calculation of the response time often becomes critical in solving the resource optimization problem. Martin and Nilsson [10] measured the average response time of a service request. A framework for service management in grid computing was defined in [11], but they did not provide a method for calculating the probability distribution of the response time.

In order to compute a percentile of the response time one has to first find the probability distribution function (pdf) of the response time. This is not an easy task in a complex computing environment involving many computing nodes. The calculation of the pdf of the response time is relatively simple for overtake-free paths in Jackson and Gordon-Newell networks. Walrand and Varaiya [18] showed that in any open Jackson network, the response times of a customer at the various nodes of overtake-free path are all mutually independent. Daduna [5] further proved that the same result is valid for overtake-free paths in Gordon-Newell networks. The pdf of the response time was derived for a closed queueing network in [12], and the passing time distribution was calculated for large Markov chains in [7].

In this paper, we present an approach for the resource optimization that minimizes the total cost of computer resources required while preserving a given percentile of the response time. We calculate the number of servers in each resource station that minimize a cost function that reflects operational costs. We first analyze an overtake-free open tandem queueing network, and then we extend our work

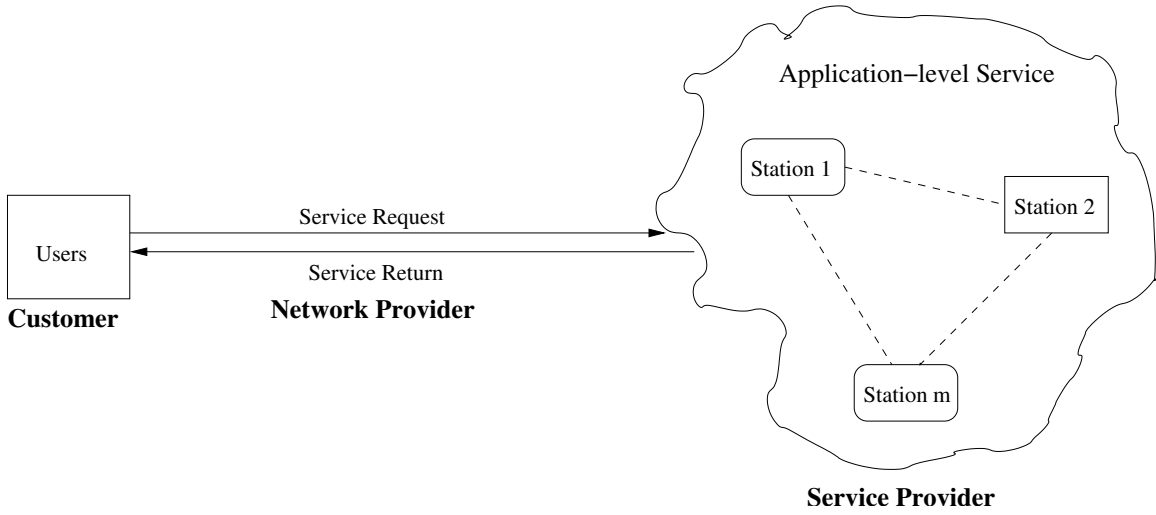


Fig. 1. Execution of Service Requests

to an open tandem queueing network with feedback. We note that the proposed approach can be also applied to any queueing network consisting of nodes arbitrarily linked. To the best of our knowledge, this is the first work that provides an analytical solution of the resource optimization problem subject to the constraints of a percentile response time and a price.

The rest of the paper is organized as follows. In section II we define the SLA performance metric considered in this paper and formulate the resource optimization problem. In section III, we present two typical real-life models and propose an approach for solving the optimization problem. In section IV, numerical simulations demonstrate the applicability and validity of the proposed approach. Finally, the conclusions are given in section V.

II. THE SLA PERFORMANCE METRIC AND THE RESOURCE OPTIMIZATION PROBLEM

A SLA is a contract between a customer and a service provider that defines all aspects of the service that is to be provided. In this paper the SLA consists of service performance and a fee. We consider the percentile of the response time as the performance metric. This is the time it takes for a service request to be executed on the service provider's multiple resource stations. Let T be a random variable representing the response time, and let $f_T(t)$ and $F_T(t)$ be its probability and cumulative distributions respectively. Also, let T^D be the desired target response time that a customer requests and agrees with its service provider based on a fee paid by the customer. The SLA performance metric is as follows.

$$F_T(t)|_{t=T^D} = \int_0^{T^D} f_T(t) dt \geq \gamma\% \quad (2)$$

That is, $\gamma\%$ of the time a service request will be executed in less than T^D .

As an example let us consider an $M/M/1$ queue with an arrival rate λ and a service rate μ . The service discipline is

FIFO. The steady-state probability of the system is

$$p_0 = 1 - \rho, \quad \text{and} \quad p_k = (1 - \rho)\rho^k, \quad k > 0,$$

where $\rho = \frac{\lambda}{\mu}$ (see [15]). The response time T is exponentially distributed with the parameter $\mu(1 - \rho)$, i.e., its probability distribution is given by

$$f_T(t) = \mu(1 - \rho)e^{-\mu(1 - \rho)t}$$

Using the definition given in (2), we have that

$$F_T(t)|_{t=T^D} = 1 - e^{-\mu(1 - \rho)T^D} \geq \gamma\% \quad (3)$$

or

$$\mu \geq \frac{-\ln(1 - \gamma\%)}{T^D} + \lambda$$

This means that in order to guarantee higher SLA service levels, μ has to increase when T^D decreases. Similarly, for a given arrival rate λ and service rate μ , we can use (3) to find the percentile of γ .

Then, the resource optimization problem can be formulated as the following optimization problem:

Resource Optimization Problem:

Find integers n_j ($1 \leq n_j \leq N_j$; $j = 1, 2, \dots, m$) in the m -dimensional integer optimization problem (1) under the constraint of a percentile response time as expressed by (2), and the constraint: $I \leq C^D$, where C^D is a fee negotiated and agreed upon between a customer and the service provider.

III. APPROACHES FOR RESOURCE OPTIMIZATION

In this section, we study two queueing network models that depict the path that service requests have to follow through the service provider's resource stations. These two models are shown in Figures 2 and 3. We refer to these two queueing models as service models since they depict the resources used to provide a service to a customer.

The first service model consists of a single infinite server and m stations numbered sequentially from 1 to m as shown

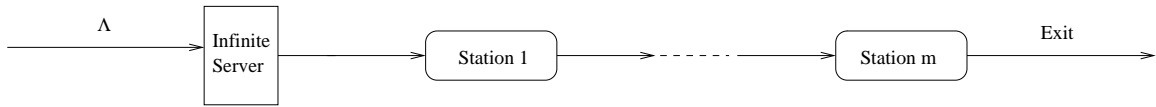


Fig. 2. A Tandem-station Service Model

in Figure 2. Each station j is modeled as a single FIFO queue served by n_j identical servers, each providing a service at the rate μ_j . Let Λ be the external arrival rate to the infinite server, and let λ and λ_j be the effective arrival rates to the infinite server and station j ($j = 1, 2, \dots, m$). We assume that all service times are exponentially distributed and the external arrival to the infinite server occurs in a Poisson fashion.

The infinite server represents the total propagation delay from the user to the service provider and back and also from station 1 to m . Each station carries out a particular function. For instance, it could be a database server, a file server, a web server, a group of CPUs and local disks, etc. We only consider a single class of customer in this paper.

In the following discussion each station is modeled as a single $M/M/1$ queue with arrival rate λ_j and service rate $\psi(n_j)\mu_j$, where $\psi(n_j)$ is a known function of n_j and depends on the configuration of servers at each station. It is non-decreasing and can be inverted, i.e., ψ^{-1} exists. For instance, suppose that a station represents a group of CPUs. Then, $\psi(n)$ can be seen as a CPU scaling factor for the number of CPUs from 1 to n . According to [4], $\psi(n) = \xi^{\log_2 n}$, where ξ is a basic scaling factor from 1 CPU to 2. So, $\psi^{-1}(n) = \xi^{-\log_2 n}$.

Since the queueing network is overtake-free, the waiting time of a customer at a station is independent of its waiting times at other stations (see [5] and [18]). Let X be the service time at the infinite server and X_j be the time elapsed from the moment a customer arriving at station j to the moment it departs from the station. Then, the total response time is

$$T = X + X_1 + X_2 + \dots + X_m,$$

and hence the LST (Laplace-Stieltjes transform) of the response time T is

$$L_T(s) = L_X(s)L_{X_1}(s) \cdots L_{X_m}(s) \quad (4)$$

where $L_X(s)$ is the LST of the service time X given by

$$L_X(s) = \frac{\lambda}{s + \lambda} \quad (5)$$

and $L_{X_j}(s)$ is the LST of the response time X_j at the j -th station given by

$$L_{X_j}(s) = \frac{\psi(n_j)\mu_j(1 - \rho_j)}{s + \psi(n_j)\mu_j(1 - \rho_j)}, \quad (6)$$

where $\rho_j = \frac{\lambda_j}{\psi(n_j)\mu_j}$ ($j = 1, 2, \dots, m$).

From (4), (5) and (6) we have that

$$L_T(s) = \frac{\lambda}{s + \lambda} \prod_{j=1}^m \frac{\psi(n_j)\mu_j(1 - \rho_j)}{s + \psi(n_j)\mu_j(1 - \rho_j)}$$

We observe that $f_T(t)$ and $F_T(t)$ are usually nonlinear functions of t and n_j . Hence, the resource optimization

problem is an m -dimensional linear optimization problem subject to nonlinear constraints. In general, it is not easy to solve this problem. However, the complexity of the problem can be significantly reduced by requiring that the service rates of the queues in the service model in Figure 2 are all equal. That is, we find the optimum value of n_1, \dots, n_m such that

$$\psi(n_1)\mu_1 = \dots = \psi(n_m)\mu_m$$

(We note that in production lines, it is commonly assumed that the service stations are balanced.)

From the traffic equations:

$$\lambda = \lambda_j = \Lambda$$

for $j = 1, 2, \dots, m$, we have that the utilization of each station

$$\rho_j = \frac{\lambda_j}{\psi(n_j)\mu_j} = \frac{\Lambda}{\psi(n_j)\mu_j}$$

Thus we have

$$\hat{a}_i = \psi(n_i)\mu_i(1 - \rho_i) = \psi(n_j)\mu_j(1 - \rho_j) = \hat{a}_j \triangleq \hat{a},$$

which implies $n_j = \psi^{-1}(\frac{\hat{a} + \lambda_j}{\mu_j})$ ($i, j = 1, 2, \dots, m$). Hence, from (4) we have

$$f_T(t) = L^{-1}\left\{\frac{\lambda}{s + \lambda} \cdot \frac{\hat{a}^m}{(s + \hat{a})^m}\right\},$$

and subsequently obtain that

$$F_T(t) = L^{-1}\left\{\frac{\lambda}{s(s + \lambda)} \cdot \frac{\hat{a}^m}{(s + \hat{a})^m}\right\} \quad (7)$$

Consequently, $\sum_{j=1}^m n_j c_j$ reduces to a function of variable \hat{a} due to $n_j = \lceil \psi^{-1}(\frac{\hat{a} + \lambda_j}{\mu_j}) \rceil$. Thus we have the following algorithm for the resource optimization problem.

Algorithm 1:

- 1) Find \hat{a} in the one dimensional optimization problem:

$$\hat{a}^{min} \leftarrow \arg \min_{\hat{a}} F_T(t)|_{t=T^D}$$

subject to the constraint $F_T(t)|_{t=T^D} \geq \gamma\%$ at $\hat{a} = \hat{a}^{min}$, where $F_T(t)$ is given by (7).

- 2) Compute integers n_j by using

$$n_j = \lceil \psi^{-1}(\frac{\hat{a}^{min}}{\mu_j(1 - \rho_j)}) \rceil,$$

and check if $1 \leq n_j \leq N_j$ ($j = 1, 2, \dots, m$) and $I \leq C^D$ are satisfied. If yes, the obtained n_j is the number of servers required at each station. Otherwise, print “the problem cannot be solved.”

In order to allow for a more complex execution path of a service request, we extended the above model to a service model with feedback, as shown in Figure 3. Infinite server

1 represents the total propagation delay within a network provider and infinite server 2 represents the propagation delay within the service provider, i.e. among stations 1 to m . In this figure, a customer upon completion of its service at the m -th station, exits the system with probability α , or returns to the beginning of the system with probability $1 - \alpha$.

We note that the model shown in Figure 3 can be easily extended to a network of queues arbitrarily connected. We reuse the notation in the first model shown in Figure 2: Λ as the external arrival rate, λ_d , λ and λ_j as the effective arrival rates to the second infinite server and station j , and μ_j as the service rate at station j , where $j = 1, 2, \dots, m$.

The main difficulty of this resource optimization problem is to find $f_T(t)$, the probability distribution function of T . We obtain this probability distribution function assuming that the waiting time of a customer at a station is independent of its waiting time at other stations, and each visit at the same station j is independent of the others. (We note that this assumption of independence does not hold in queueing networks with feedback. However, as will be discussed in the validation section the solution obtained has a good accuracy.) We first have the traffic equations:

$$\lambda_d = \Lambda, \quad \lambda = \Lambda + (1 - \alpha)\lambda_m, \quad \text{and} \quad \lambda_j = \lambda,$$

which implies $\lambda_j = \lambda = \frac{\Lambda}{\alpha}$, and the utilization of each station is

$$\rho_j = \frac{\lambda_j}{\psi(n_j)\mu_j} = \frac{\Lambda}{\alpha\mu_j\psi(n_j)} \quad (j = 1, 2, \dots, m)$$

Furthermore, the response time of the k -th pass at the infinite station and the j -th station is considered as the sum of $m + 2$ random variables

$$T(k) = D + X + X_1 + \dots + X_m,$$

where we assume that the waiting time of a customer at a station is independent of its waiting times in other visits to the same station. D and X are the service times at the first and second infinite servers respectively, and X_j is the time elapsed from the moment a customer arrives at station j to the moment it departs from it. Then, the total response time is

$$T = \sum_{k=0}^{\infty} p(k)T(k),$$

where $p(k)$ is the steady state probability that a request will circulate k times at the infinite station and the j -th station through the computing system. $p(k)$ is determined by

$$p(k) = \alpha(1 - \alpha)^{k-1}$$

Thus the LST of the response time T is

$$L_T(s) = L_D(s) \sum_{k=0}^{\infty} p(k) L_X^k(s) L_{X_1}^k(s) \cdots L_{X_m}^k(s),$$

which can be re-written as follows:

$$L_T(s) = \frac{\alpha L_D(s) L_X(s) \prod_{j=1}^m L_{X_j}(s)}{1 - (1 - \alpha) L_X(s) \prod_{j=1}^m L_{X_j}(s)} \quad (8)$$

where $L_D(s)$ is the LST of the service time D given by

$$L_D(s) = \frac{\Lambda}{s + \Lambda},$$

and replacing $L_X(s)$ and $L_{X_j}(s)$ ($j = 1, 2, \dots, m$) by (5) and (6) in (8), we have that

$$L_T(s) = \frac{\Lambda^2 \prod_{j=1}^m \hat{a}_j}{(s + \Lambda)[(s + \lambda) \prod_{j=1}^m (s + \hat{a}_j) - (1 - \alpha) \lambda \prod_{j=1}^m \hat{a}_j]} \quad (9)$$

To find the response time distribution $f_T(t)$, we are required to invert the above LST using partial fraction decomposition of a rational function. However, the partial fraction decomposition of the rational function requires searching for roots of a high-order polynomial. It is usually not an easy task when the order of the polynomial is more than 5. Instead, in this paper the LST is inverted numerically.

Similarly, we want to find n_1, \dots, n_m such that the best utilization of these stations is achieved, which implies that each station has the same maximal service capacity. That is,

$$\hat{a}_i = \hat{a}_j = \hat{a} \quad (i, j = 1, \dots, m)$$

Then, from equation (9) and $F_T(t) = L^{-1}\{L_T(s)/s\}$ we have

$$F_T(t) = L^{-1}\left\{\frac{\Lambda^2 \hat{a}^m}{s(s + \Lambda)[(s + \lambda)(s + \hat{a})^m - (1 - \alpha)\lambda \hat{a}^m]}\right\} \quad (10)$$

Thus we have the following algorithm for the resource optimization problem in the model shown in Figure 3.

Algorithm 2:

Steps 1) and 2) are the same as Steps 1) and 2) in Algorithm 1 except $F_T(t)$ given by (10).

Note that if we cannot get a solution for the resource optimization problem using Algorithm 1 (or 2), then the service provider cannot execute the service request for the service model 1 (or 2) due to at least one of the following reasons:

- 1) The service provider has an insufficient resource (i.e., N_j is too small).
- 2) A pre-specific fee is too low (i.e., $I > C^D$).
- 3) A network connection is either too slow or has a problem so that (2) cannot be satisfied.

Using these information, we may detect and debug either a network problem or a service provider's capacity problem, or the SLA needs to be re-negotiated.

In Algorithms 1 and 2, the run-time for Step 2) is $O(m)$. Thus, the run-time of either Algorithm 1 or 2 is a sum of $O(m)$, the run-time for inverting the LST of the response time, and the run-time for finding the maxima of the resulting function (or $F(t)$). While an efficient approach to finding the maxima of $F(t)$ can be found in [16], inverting the LST of the response time can be easily done by using the methods presented in [6].

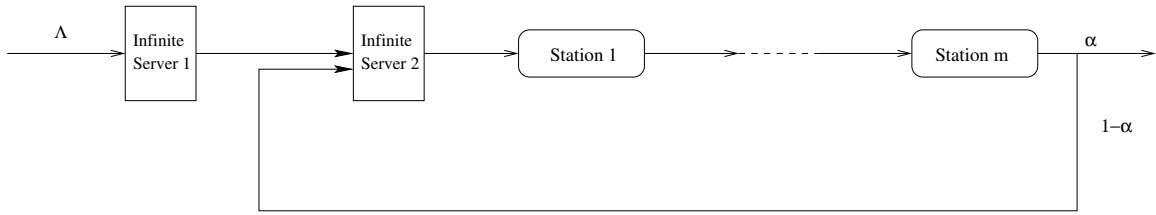


Fig. 3. A Service Model With Feedback

TABLE I
THE SERVICE RATES OF THE EIGHT STATIONS UNDER STUDY

Service Rates	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
Values	52	18	80	35	41	15	25	35

IV. NUMERICAL VALIDATIONS

In this section we demonstrate the accuracy and applicability of our proposed approximation method.

Two types of errors are introduced in our proposed approximation method. The first, hereafter referred to as Class I error, comes from numerically inverting the Laplace transform. The other, hereafter referred to as Class II error, is due to the assumptions that the waiting time of a customer at each station is independent of the waiting times at the other stations, and it is also independent of its waiting times in other visits to the same station.

The relative error % is used to measure the accuracy of the approximate results compared to model simulation results, and it is defined as follows

$$\text{Relative error \%} = \frac{\text{Approximate Result} - \text{Simulation Result}}{\text{Simulation Result}} \times 100$$

We study the accuracy of our proposed approximation method using two examples below.

We first verify the accuracy of our approach for the first service model shown in Figure 2. Let $m = 8$, $\lambda = 100$, $N_j = 100$, $c_j = 2$, $\psi(n_j) = 1.5^{\log_2 n_j}$ and $C^D = 400$ ($j = 1, \dots, 8$). The service rates of these eight stations are listed in Table I.

We simulated the queueing network using Arena and the analytical method was implemented in Mathematica. The simulation results are considered as “exact” since the simulation model is an exact representation of the queueing network under study.

Table II shows the simulated and approximate cumulative distribution of the response time. In the table, the column labeled “Simul” gives the simulation result, the column labeled “Approx” gives the approximate result, and the column labeled “R-Err %” gives their relative errors. The same abbreviations are also used in Table VI. It appears that the results obtained by Algorithm 1 are very accurate. The optimal number of servers required for 97.5% of the response time to be less than $T^D = 0.16$ is shown in Table III. The exact optimal number of servers, obtained by exhaustive search using the simulation model, and assuming that each

TABLE II
THE CUMULATIVE DISTRIBUTION OF THE RESPONSE TIME

Response Time	Simul	Approx	R-Err %
0.04	0.0213	0.0214	0.4393
0.06	0.1517	0.1528	0.7004
0.08	0.4070	0.4075	0.1112
0.10	0.6681	0.6672	-0.1377
0.12	0.8468	0.8450	-0.2158
0.14	0.9398	0.9379	-0.1974
0.16	0.9785	0.9780	-0.0498
0.18	0.9931	0.9929	-0.0157
0.20	0.9979	0.9979	0.0000
0.22	0.9995	0.9994	-0.0077
0.24	0.9999	0.9998	-0.0051
0.26	1.0000	1.0000	0.0000

station has the same utilization, or balanced utilization, is consistent with the ones shown in Table III. Thus, $I = 382 < C^D$. We point out that the relative errors shown in Table II are only due to the Class I error since the Class II error is not present for this service model.

TABLE III
THE OPTIMAL NUMBER OF SERVERS

Station	1	2	3	4	5	6	7	8
#Servers	11	62	5	20	16	84	35	20

Let us now consider an example of the service model 2 shown in Figure 3. We choose $m = 8$, $\Lambda = 100$, $\alpha = 0.67$, $N_j = 250$, $c_j = 1$, $\psi(n_j) = 1.5^{\log_2 n_j}$, and $C^D = 580$ ($j = 1, \dots, 8$). The service rates of these eight stations are listed in Table IV. Thus it follows from equation $\lambda = \frac{\Lambda}{\alpha}$ that $\lambda = 149.25$.

We obtained the cumulative distribution of the response time by solving (10) using the package of the inverse Laplace transforms in Graf [6]. Table V shows the number of servers in the eight stations necessary to ensure the 95% SLA guarantee for $T^D \leq 0.6$. We also simulated the tandem queueing network and validated using the brute-force approach that these numbers of servers obtained by our approximate method are in fact optimal, provided that each station has balanced utilization. The optimal number of

TABLE IV
THE SERVICE RATES OF THE EIGHT STATIONS UNDER STUDY

Service Rates	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
Values	10	45	100	20	32	18	8	85

TABLE V
THE OPTIMAL NUMBER OF SERVERS

Station	1	2	3	4	5	6	7	8
#Servers	168	13	4	52	23	62	246	5

TABLE VI
THE CUMULATIVE DISTRIBUTION OF THE RESPONSE TIME

Response Time	Simul	Approx	R-Err %
0.20	0.4865	0.4781	-1.7284
0.30	0.7418	0.7267	-2.0336
0.40	0.8551	0.8541	-0.1201
0.50	0.9189	0.9226	0.4067
0.60	0.9538	0.9589	0.5327
0.70	0.9733	0.9782	0.5000
0.80	0.9845	0.9884	0.3967
0.90	0.9908	0.9938	0.3070
1.00	0.9946	0.9967	0.2136
1.10	0.9967	0.9983	0.1563
1.20	0.9980	0.9991	0.1079
1.30	0.9988	0.9995	0.0714
1.40	0.9997	0.9997	0.0000
1.60	0.9999	0.9999	0.0026
1.80	0.9999	1.0000	0.0079
2.00	1.0000	1.0000	0.0000

servers is given in Table V. It derives that $I = 560 < C^D$, i.e., Step 2) in Algorithm 2 is met. Table VI gives the cumulative distribution of the response time obtained using the approximate method and the simulation method, and the relative error %. The relative error comes from both Classes I and II error. We note that our approximate method has a very good accuracy.

In both two examples as above, the run-time for the approximate method is less than 1 second when Algorithm 1 or 2 was implemented in Mathematica, and the run-time for the simulation result is less than 1 minute when Arena was used.

Extensive numerical results (not reported here due to lack of space) point to the fact that the independence assumption has little impact on the accuracy of the results when the number of nodes is large. A contributing factor is that typically we are interested in values of the cumulative distribution of the response time that correspond to very high percentiles for which the approximate results seem to have a very good accuracy through a comparison with simulation results that are considered as "exact."

Additionally, to the best of our knowledge, this is the first work that provides an analytical solution of the resource optimization problem subject to the constraints of a percentile response time and a price. Hence, we do not give a comparison of our proposed method with other methods in this paper.

V. CONCLUSIONS

We proposed an approach for resource optimization in a service provider's computing environment, whereby we minimize the total cost of computer resources allocated to a customer so that satisfies a given percentile of the response time. We have formulated the resource optimization

problem as an optimization subject to SLA constraints for a service model with or without feedback. In the model without feedback, the obtained LST of a customer's response time is exact, and in the case of the model with feedback, it is approximate. We also developed an efficient and accurate numerical solution for inverting the LST of a customer's response time numerically. Validation testes showed that our approach has a very good accuracy.

In this paper we assumed that service requests are served in a queue in a FIFO manner. Priority service disciplines will be discussed in another paper.

REFERENCES

- [1] M. Allman and V. Paxson, "On estimating end-to-end network path properties," In *Proceedings of the ACM SIGCOMM*, pp. 263-274, August-September 1999.
- [2] I. Aib, N. Agoulmine, and G. Pujolle, "The generalized service level agreement model and its application to the SLA driven management of wireless environments," In *Proceedings of the International Arab Conference on Information Technology (ACIT)*, December 2004.
- [3] E. Bouillet, D. Mitra, and K. Ramakrishnan, "The structure and management of service level agreements in networks," *IEEE Journal on Selected Areas in Communications*, 20(4), pp. 691-699, 2002.
- [4] J. Chang, "Processor performance, Update 1," In *SQL-Server-Performance.com*.
- [5] H. Daduna, "Burke's theorem on passage times in Gordon-Newell networks," *Adv. Appl. Prob.*, 16, 1984.
- [6] U. Graf, *Applied Laplace Transforms and z-Transforms for Scientists and Engineers*. Birkhauser Verlag, Basel-Boston-Berlin, 2004.
- [7] P. Harrison and W. Knottenbelt, "Passage time distributions in large Markov chains," In *Proceedings of the ACM SIGMETRICS*, pp. 77-85, June 2002.
- [8] J. Lee and R. Ben-Natan, *Integrating Service Level Agreements : Optimizing Your OSS for SLA Delivery*, Wiley Publisher, 2002.
- [9] C. Matthys, P. Bari, E. Lieurain, D. Salomon, L. Winkelbauer, B. Jacob, S. Mui, J. Pannu, S. Park, H. Raguette, J. Schneider, and L. Vanel, *On Demand Operating Environment: Managing the Infrastructure (Virtualization Engine Update)*, IBM Redbooks, June, 2005.
- [10] J. Martin and A. Nilsson, "On service level agreements for IP networks," In *Proceedings of the IEEE INFOCOM*, June 2002.
- [11] D. Menasce and E. Casalicchio, "A framework for resource allocation in grid computing," In *Proceedings of the 12th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (IEEE MASCOTS'04)*, pp. 259-267, October 2004.
- [12] J. Muppala, K. Trivedi, V. Mainkar, and V. Kulkarni, "Numerical computation of response time distributions using stochastic reward nets," *Annals of Oper. Res.*, 48, 1994.
- [13] D. Nowak, P. Perry, and J. Murphy, "Bandwidth allocation for service level agreement aware Ethernet passive optical networks," In *Proceedings of the IEEE Globecom*, pp. 1953-1957, November-December 2004.
- [14] V. Paxson, "End-to-end Internet packet dynamics," In *Proceedings of the ACM SIGCOMM*, pp. 139-152, September 1997.
- [15] H. Perros, *Queueing Network with Blocking, Exact and Approximate Solutions*, Oxford University Press, 1994.
- [16] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in Fortran*, Cambridge University Press, 1997.
- [17] Sun Microsystems, "Service level agreement in the data center," <http://www.sun.com/blueprints/0402/sla.pdf>.
- [18] J. Walrand and P. Varaiya, "Sojourn times and the overtaking condition in Jacksonian networks," *Adv. Appl. Prob.*, 12, 1980.