RJ 10242 (A0204-028) April 22, 2002
Computer Science

IBM Research Report

An Optimization Model for Storage Service Based on Quality of Service

T. Paul Lee
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

An Optimization Model for Storage Service based on Quality of Service

T. Paul Lee

IBM Research Division
IBM Almaden Research Center
San Jose, CA 95120, USA
tpl@almaden.ibm.com

Abstract

In this paper, we present an optimization model for storage service based on quality of service. The quality measure is the sojourn time of an individual I/O request. The service provider gets a premium P for each I/O request served, commits to serve certain I/O request rate λ, and in return, the service provider guarantees its service quantitatively using a rebate model. Intuitively, the service guarantee is quantified as a rebate to the customer depending on how well the service is rendered. The overall objective for the storage service provider is to find an optimal I/O rate it should commit to serve to maximize its net gain, that is, the difference between the premium received and the rebate back to the customer.

This model has a clear and quantifiable measure of net gain. The provider has every reason not to over-commit its service capacity while the customer has a way to verify the guaranteed quality of service in monetary terms. To gain insight on the structure of the model, we show some analytic results for this optimization problem with a simple M/M/1 FCFS queuing model. Two scenarios are examined. One uses queuing time of the I/O request as the measure of quality of service and the other uses response time.

We have a few suggestions on how to apply this model in the real-life world. To operate at an optimal level, storage service providers have to quantitatively characterize the customer workload and their own storage subsystems. On the operational side, the storage service providers need to continuously monitor and shape the arrival stream for conformance and deliver the guaranteed level of service. Postmortem analysis of I/O traces and measurements should help in determining and projecting sustainable and profitable I/O rate.

We could anticipate the existence of a third-party measuring and reporting agent which is between customer applications and storage service provider. With this arrangement, full accountability shall prevail, and the proposed model would make even better business sense.

1 Introduction

As technology evolves, enterprise computing has been based more and more on the client-server model. Data repositories are now commonly stored and served by storage servers over the network; powerful clients on the desktops perform relatively simple data computation and transformation before they are displayed on large monitor screens. The storage service providers are now becoming

separately accountable entities for their services, and the quality of their services is of much recent research interest [1] [2]. Today, it is quite feasible to measure and record the quality of service at a very small granularity, that is, at an individual I/O request level. Given this trend, we present an optimization model for this storage service. Customers pay premiums for the agreed-upon service volume and get quality of service guarantee from the provider. The guarantee is in the form of rebate or credit back to the customer as a function of quality of service. Thus, the service provider has to seriously engineer its infrastructure to lower cost while understanding quantitatively its commitment to meet the requirement and achieve its financial goals. In a way, the service provider has to continue to monitor and look for ways to achieve maximum net gain.

After the model is presented in Section 2, we use a simple M/M/1 queuing model with FCFS discipline [3] to show a few analytic results for the optimal I/O arrival rate the service provider could commit to realize maximum net gain or profit. Two scenarios are discussed in Section 3. The first one uses response time of the I/O request as the measure of quality of service and the second one uses queuing or waiting time.

While the analysis is only tractable for certain simple mathematical models, we offer a few suggestions in Section 4 on applying this optimization model to real-life situations. To operate at an optimal level, storage service providers have to quantitatively characterize the customer workload and their own storage subsystems. On the operational side, the storage service providers need to continuously monitor and shape the arrival stream for conformance and deliver the guaranteed level of service. Postmortem analysis of I/O traces and measurements should help in determining and projecting sustainable and profitable I/O rate. We conclude this paper with a few observations in Section 5.

2 An Optimization Model for Storage Service based on Quality of Service

Our optimization model for the storage service provider is based on an agreed-upon I/O service level or arrival rate, the premium or service charge per I/O request, and a rebate model as a function of quality of service. Specifically, customers pay certain premium P for each I/O request, and the storage service provider commits to serve at an agreed-upon rate λ (number of I/Os per second). The product of P and λ is the income (per unit time) for the service provider. In return, the storage service provider guarantees certain quality of service for the customer in a quantitative manner. In this model, the quality measure is the sojourn time of the I/O request in the system. This guarantee is represented by a rebate model as a function of sojourn time for each I/O request. The customer gets a rebate (or credit) as a function of sojourn time. In essence, the service provider gives more rebate as the quality of service gets worse.

This rebate is represented by $C(t)$ where t is the sojourn time for the I/O request in question. To get the average rebate for the I/O request in question, we need to sum over all possible sojourn

time based on the probability density function $f(t)$ where t is used here to denote generically the quality of service in question.

The overall objectives for the storage service provider is to find an optimal I/O rate it should commit to serve to maximize its net gain, that is, the difference between the premium received and the rebate back to the customer. In general, $G(\lambda)$ can be represented as follows:

$$G(\lambda) = P\lambda \ - \ \lambda \int_0^\infty C(t) \ f(t)dt \tag{1}$$

3 Analytic Results based on M/M/1 FCFS Queuing Model

In addition to the arrival rate, the net gain formula (1) is obviously a function of arrival structure, service time requirement, and internal processing strategy of the storage system. To gain some insight on the structure of this optimization problem, we make some simplifying assumptions to solve the problem analytically.

Here we use the classical M/M/1 queuing model where the arrival process is characterized by a Poisson process with rate λ, and the service requirement is captured by an independently and identically distributed exponential service rate μ of a single aggregated storage server. The utilization ρ is thus defined to be λ/μ. The appropriate measure for quality of service for individual I/O request would be its sojourn time. We'll investigate both the response time and the queuing time in the following Sections. To obtain explicit probability density function for this model, we assume further that the queuing discipline is FCFS, i.e., first come first serve. In our model, the sojourn time is always non-negative and we'll do our analysis accordingly.

Under realistic circumstances, the optimal I/O arrival rate λ_{opt} should be strictly less than μ. Although some mathematically unreasonable set of parameters might indicate otherwise, we shall observe this constraint.

3.1 Using Response Time as the Quality Measure

The response time is the total time an I/O request stays in the system including both queuing time and its own service time. The response time distribution function for M/M/1 FCFS queue $F_T(t)$ and its probability density function $f_T(t)$ are respectively as follows:

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

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

3.1.1 Fixed Rebate if Quality Measure is not Met

In this rebate model, $C(t)$ is a step function. That is, $C(t)$ is 0 when response time is less or equal to the service guarantee w_0 and is constant R when response time exceeds the prescribed limit

w_0. Specifically,

$$C(t) = \begin{cases} 0 & \text{if } w \leq w_0 \\ R & \text{otherwise} \end{cases}$$

Thus, our net gain

$$G(\lambda) = P\lambda \ - \lambda \int_{w_0}^{\infty} R \ \mu(1-\rho)e^{-\mu(1-\rho)t}dt$$

This reduces to

$$G(\lambda) = P\lambda \ - \lambda R e^{-(\mu-\lambda)t}|_{\infty}^{w_0}$$

or

$$G(\lambda) = P\lambda \ - Re^{-(\mu-\lambda)w_0}\lambda \tag{2}$$

The first and second derivative of $G(\lambda)$ can be easily derived now as

$$G'(\lambda) = P - R(1 + w_0\lambda)e^{-(\mu-\lambda)w_0}$$

$$G''(\lambda) = -Rw_0(2 + w_0\lambda)e^{-(\mu-\lambda)w_0}$$

Setting the first derivative to 0 gives us the following implicit equation for optimal λ

$$(1 + w_0\lambda)e^{-(\mu-\lambda)w_0} = \frac{P}{R} \tag{3}$$

which maximizes the net gain $G(\lambda)$ because $G''(\lambda) < 0$ is always true. Numerical method is required to solve for λ explicitly.

3.1.2 Linear Rebate Model

A more interesting rebate function for bad service is the linear function of response time. That is, the form of the rebate function is $C(t) = Rt$ where R the slope of this linear function. Substituting this into the general equation (1) for $G(\lambda)$, we have

$$G(\lambda) = P\lambda - \lambda \int_{0}^{\infty} Rt \ \mu(1-\rho)e^{-\mu(1-\rho)t}dt$$

This reduction is equivalent to that of summing the tail of the distribution function to get the first moment of the the response time random variable t. That is, the first moment can be computed as follows:

$$E[X] = \int_{0}^{\infty} (1 - F_X(x))dx \tag{4}$$

With this technique, we have

$$G(\lambda) = P\lambda - R\frac{\lambda}{\mu - \lambda} \tag{5}$$

And itss derivatives are readily derived as follows:

$$G'(\lambda) = P - R\frac{\mu}{(\mu-\lambda)^2}$$

$$G''(\lambda) = -2R\frac{\mu}{(\mu-\lambda)^3}$$

Since $G''(\lambda)$ is always less than 0, we solve λ_{opt} by setting the first derivative to 0. Thus,

$$\lambda_{opt} = \mu - \sqrt{\frac{R\mu}{P}}$$

or, in a more interesting form,

$$\lambda_{opt} = \mu \left(1 - \sqrt{\frac{R}{P\mu}} \right) \tag{6}$$

This result provides some good insight and intuition that the slack, i.e., the difference between committed arrival rate and service rate, is related by the square root relationship of the ratio $R/P\mu$. The high penalty rate R or the low premium P would reduce the I/O rate the service provider should commit to serve. Also, given these two are fixed, improved μ for the storage service system would encourage serving more I/O traffic closer to the capacity μ.

3.1.3 Quadratic Rebate Model

Now, we investigate a more stringent rebate function for bad service such as the quadratic function. That is, the $C(t)$ is of the form $C(t) = Rt^2$. Substituting this into the general equation (1) for $G(\lambda)$, we have

$$G(\lambda) = P\lambda - \lambda \int_0^\infty Rt^2 \ \mu(1-\rho)e^{-\mu(1-\rho)t}dt$$

This reduction is equivalent to compute the second moment of the response time random variable t. Note that the computation of second moment can be done through the following result:

$$E[X^2] = 2 \int_0^\infty x(1 - F_X(x))dx \tag{7}$$

With this formula (7) and $F_T(t)$, we can integrate by parts to obtain

$$G(\lambda) = P\lambda - 2\,R\frac{\lambda}{(\mu-\lambda)^2} \tag{8}$$

Its derivatives are readily derived as follows:

$$G'(\lambda) = P - 2R\,\frac{\mu+\lambda}{(\mu-\lambda)^3}$$

$$G''(\lambda) = -4R\,\frac{2\mu+\lambda}{(\mu-\lambda)^4}$$

Since $G''(\lambda)$ is always less than 0, we solve for optimal λ by setting the first derivative to 0. That is,

$$\lambda^3 - 3\mu\lambda^2 + (3\mu^2 - \frac{2R}{P})\lambda - (\mu^3 + \frac{2R}{P}\mu) = 0 \tag{9}$$

Numerical method can be used to solve for λ explicitly.

3.2 Using Queuing Time as the Quality Measure

It can be argued that the queuing time is a better quality measure than the response time measure since it excludes the service time requirement from the particular I/O request. Some I/O requests

are inherently longer since it requires more data to be moved. Although it is generally more difficult to define and separate between the service time portion and the queuing time portion in a complex storage subsystem, we'll investigate this in the simple framework of M/M/1 queue with FCFS service discipline.

The queuing or waiting time distribution function for M/M/1 FCFS queue is $F_W(t)$ and its probability density function $f_W(t)$ are respectively as follows:

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

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

The analysis is very similar to that of the response time counterpart in the previous Sections.

3.2.1 Fixed Rebate if Quality Measure is not Met

In this rebate model, $C(t)$ is the same step function defined previously. Thus, our net gain is

$$G(\lambda) = P\lambda \ - \lambda \int_{w_0}^{\infty} R \ \rho\mu(1-\rho)e^{-\mu(1-\rho)t}dt$$

This reduces to

$$G(\lambda) = P\lambda \ - R\lambda\rho e^{-(\mu-\lambda)t}|_{\infty}^{w_0}$$

and, we obtain

$$G(\lambda) = P\lambda \ - R\frac{\lambda^2}{\mu}e^{-(\mu-\lambda)w_0} \tag{10}$$

The first and second derivative of $G(\lambda)$ can be easily derived now as

$$G'(\lambda) = P - R \ \frac{\lambda(2+w_0\lambda)}{\mu} \ e^{-(\mu-\lambda)w_0}$$

$$G''(\lambda) = -R \ \frac{2+4w_0\lambda+w_0^2\lambda^2}{\mu} \ e^{-(\mu-\lambda)w_0}$$

Setting the first derivative to 0 gives us the following implicit equation for optimal λ

$$\lambda(2+w_0\lambda)e^{-(\mu-\lambda)w_0} = \frac{P\mu}{R} \tag{11}$$

which maximizes the net gain $G(\lambda)$ because $G''(\lambda) < 0$ is always true. Numerical method is required to solve for λ explicitly.

3.2.2 Linear Rebate Model

In this case, the form of the rebate function is $C(t) = Rt$ where R is the slope of the linear function. Substituting this into the general equation (1) for $G(\lambda)$, we have

$$G(\lambda) = P\lambda - \lambda \int_{0}^{\infty} Rt \ \mu\rho(1-\rho)e^{-\mu(1-\rho)t}dt$$

This reduction is equivalent to that of summing the tail of the distribution function to get first moment of the queuing time random variable t. Using the result in equation (4), we obtain

$$G(\lambda) = P\lambda - R\frac{\lambda^2}{\mu(\mu-\lambda)} \tag{12}$$

Its derivatives are readily derived as follows:

$$G'(\lambda) = P - R\frac{\lambda(2\mu-\lambda)}{\mu(\mu-\lambda)^2}$$

$$G''(\lambda) = -2R\frac{\mu}{(\mu-\lambda)^3}$$

Since $G''(\lambda)$ is always less than 0, we solve the λ_{opt} by setting the first derivative to 0. Thus,

$$\lambda_{opt} = \mu - \sqrt{\frac{R}{R+P\mu}}\ \mu$$

or, in a more interesting form,

$$\lambda_{opt} = \mu\ (\ 1 - \sqrt{\frac{R}{R+P\mu}}\ \ ) \tag{13}$$

This result provides similar insight and intuition for the slack, i.e., the difference between committed arrival rate and service rate, observed in Section 3.1.2.

3.2.3 Quadratic Rebate Model

Now, we analyze a more stringent rebate function for bad service such as the quadratic function of the queuing time. That is, the $C(t)$ is of the form $C(t) = Rt^2$. Substituting this into the general equation (1) for $G(\lambda)$, we have

$$G(\lambda) = P\lambda - \lambda \int_0^\infty Rt^2\ \mu\rho(1-\rho)e^{-\mu(1-\rho)t}dt$$

This reduction is equivalent to compute the second moment of the queuing time random variable t. Similar to reduction in Section 3.1.3, we obtain

$$G(\lambda) = P\lambda - 2R\ \frac{\lambda^2}{\mu(\mu-\lambda)^2} \tag{14}$$

Its derivatives are readily derived as follows:

$$G'(\lambda) = P - 4R\frac{\lambda}{(\mu-\lambda)^3}$$

$$G''(\lambda) = -4R\frac{\mu+2\lambda}{(\mu-\lambda)^4}$$

Since $G''(\lambda)$ is always less than 0, we solve for optimal λ by setting the first derivative to 0. That is,

$$\lambda^3 - 3\mu\lambda^2 + (3\mu^2 - \frac{4R}{P})\lambda - \mu^3 = 0 \tag{15}$$

Numerical method can be used to solve for λ explicitly.

4 Applications and Practical Concerns

Although analytic result provides good insight on how various parameters in the model might interact, it is always necessary to examine the real-life workload characteristics of I/O requests [4]. Thus, it is important that the storage service provider has the tools and methodology to do workload characterization, and to parameterize its storage subsystem in order to benefit from the proposed optimization model. The storage system needs to record traces of I/O requests, response times, queuing distribution, and utilizations of various system components. Based on these measurement, storage subsystem can be characterized and verified models can be built to do prediction as required. Synthetic workload can be used often to test various capacities and parameters of the system as well [5].

Operationally, the service provider needs a comprehensive set of monitoring and control tools to shape and serve the incoming I/O traffic. This includes measuring the actual I/O request rates for compliance. The leaky bucket algorithm [6] is one way to enforce the agreed-upon rate while still allowing certain fluctuation and burstiness in the arrival stream. In this approach, we maintain certain service volume for I/O arrivals in the larger time scale; smaller time scale is used to measure the arrival rate and shape the arrival stream.

If I/O traces for a particular customer are available, storage service provider can also employ the technique of trace playback to obtain response time distribution for modeling purposes. The model can be used to estimate and predict optimal I/O arrival rate for maximum net gain. Efficient algorithms for this task are interesting areas of future research.

5 Conclusion

Traditionally, customers and service providers rely on largely qualitative terms in their contract and agreement. It is difficult to define service quality and even harder to enforce ambiguous measures. In part, business relationship has to be based on mutual trust. Some service providers might even be able to over-commit service capacity in the hope that bad services can be masked and, thus, cannot be accountable for.

In this paper, we present an optimization model for storage service based on quality of service. The quality measure is the sojourn time of an individual I/O request. The service provider gets a premium P for each I/O request served, commits to serve certain I/O request rate λ, and in return, the service provider guarantees its service quantitatively using a rebate model. The service provider has every reason not to over-commit its service capacity while the customer has a way to verify the guaranteed quality of service in monetary terms. Although we are focusing on figuring out optimal I/O rate it should commit to serve, we should not exclude the opportunities to improve internal processing algorithms for better response time distribution. Improve and lower the fixed cost of storage infrastructure should be a constant struggle to stay competitive with respect to other service providers in the market place.

To gain some better understanding on the structure of this optimization problem, we show some analytic results with a simple M/M/1 FCFS queuing model. For the linear rebate model, the inter-relationship among premium, rebate rate, and service capacity provides excellent insight for this optimization model in formula (6) and (13).

An interesting aspect of the optimization model is not explored here in this paper. The subscription model and premium structure is fixed at certain agreed-upon volume and certain constant charge per I/O for that volume. The flow-shaping mechanism of some sort for the incoming traffic is assumed in this model since we cannot allow additional traffic to affect the response times of the I/O requests being served. Even we have extra capacity at the server, we would not be able to take advantage of this excess capacity. The customers have tendency to under-subscribe to lower the fixed cost of the service contract. To compensate for this, the model could allow for limited additional traffic at a higher premium. This would coerce the customer not to under-subscribe, and would allow the service provider to use excess capacity. This flexibility allows for sudden surge of service demands.

Note that we formulate this model from the viewpoints of a storage service provider. Although not explicitly addressed here, the customers of storage service providers search for their own "optimality" in the storage service marketplace. If market forces are working, customers should be able to find the lowest cost for the service, that is, the lowest premium for the desired service volume while keep the same rebate model as service guarantee.

It is highly likely that there can be a third-party measuring and reporting agent who sits between customer applications and storage service provider. With this arrangement, full accountability can exist and the proposed model would be very practical indeed.

Acknowledgement

The author likes to thank Honesty Young and Windsor Hsu for their reviews and comments on the early drafts of this paper.

References

[1] C. Aurrecoechea, A. Campbell, and L. Hauw, "A survey of QoS architecture," *Multimedia Systems* 6: 138-151, 1998.

[2] E. Borowsky and et. al., "Using attribute-managed storage to achieve QoS," *Proceedings of 5th Intl. Workshop on Quality of Service*, June 1997.

[3] H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology.* Addison-Wesley, Reading, Massachusetts, 1978.

[4] C. Ruemmler and J. Wilkes, "Unix disk access patterns," *Proceedings of Winter'93 USENIX Conference*, pp.405-420, 1993.

[5] G. Alvarez and et. al., "MINERVA: An Automated Resource Provisioning Tool for Large-scale Storage Systems," *HP Labs Technical Reports*, HPL-2001-139, 2001.

[6] J. Turner, "New directions in communication, or Which way to the information age?" *IEEE Communication Magazine*, vol. 24, pp. 8-15, October 1986.