# Improving Resource Utilization By Over Admission

Yao-Nan Lien, and Kuo-Chan Huang
*Computer Science Department*
*National Chengchi University*
*Taipei, Taiwan, R.O.C.*
*lien@cs.nccu.edu.tw*

## ABSTRACT

*BBQ, Budget-Based management infrastructure, is designed to offer end-to-end QoS assurance for All-IP networks. Each edge router in a core network is pre-allocated with some short-path resources, which are the paths from edges to edges of the core network. Short-paths are then allocated to individual incoming service requests on a reservation basis when they are admitted to the network. However, some types of traffics such as telephony traffics may not be persistently active during their life time leaving part of reserved resources unused. To enhance resource utilization, we propose an over-admission method in which each edge router can grant more resources to the incoming requests than it actually acquired. We developed an analytical model that can be used for estimating the optimal level of over-admission without incurring too much overflow. Simulations show that our analytical model is accurate and this method can enhance resource utilization effectively. For regular telephony traffics, near 100% improvement can be reached when their active ratios are near 50%.*

**Index Terms:** All-IP Network, QoS, admission control.

## 1. INTRODUCTION

### 1.1 All-IP Networks

An *All-IP Network* uses IP based packet-switched networks to carry all types of network traffics [6]. This revolutionary All-IP network not only reduces network deployment and management costs, but also offers a great opportunity opening to various new services that are not possible on the conventional separated networks. However, running time-sensitive services such as VoIP on packet-switched networks may suffer from poor quality problem due to long delay time, large jitter, and high packet loss rate. To make All-IP networks possible, QoS is a critical problem yet to be overcome [3].

Without loss of generality, we assume the following simplified All-IP network architecture. A worldwide All-IP network consists of several core networks interconnected together through some interconnection links (e.g. undersea fiber optic cables) and some number of stub networks (also named *access networks*) connected to the core networks. A *core network* consists of some *Interior Routers* (*IR*) and some *edge routers*. The network architecture is depicted in Fig. 1.

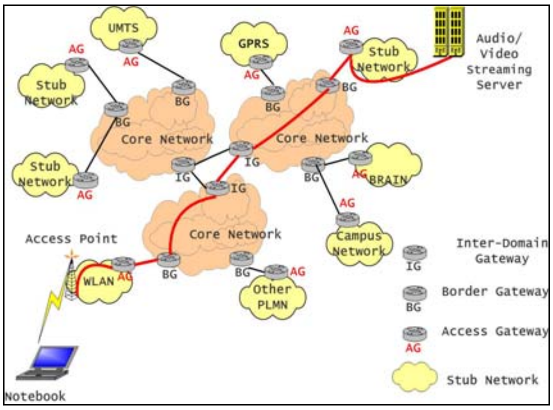

Fig. 1. Simplified All-IP Network Architecture

### 1.2. Related Work

The two most popular QoS technologies are Differentiated Services (DiffServ) and Integrated Services (IntServ) [1,4,5]. The heart of IntServ is RSVP (Resource Reservation Protocol) [4,5]. Before admitting a service request, IntServ first reserves demanded resources along the path selected for the request. It can provide end-to-end QoS assurance with a very high confidence, but it suffers from scalability due to its tremendous management overhead. On the contrast, DiffServ is more scalable, but has lower confidence on QoS assurance. DiffServ is a mechanism for specifying and controlling network traffics by their classes so that certain types of traffics, such as voice, get precedence [1]. The major advantage of DiffServ is its simplicity, scalability, as well as easy to implement. However, it can only control class-based per-hop behavior. Extra mechanisms are needed to guarantee per-flow end-to-end QoS.

### 1.3. BBQ

To support end-to-end QoS for All-IP networks, we proposed Budget-Based End-to-end QoS management infrastructure, *BBQ*, in which the quality bound of each network component is controlled based on a calculated budget plan. End-to-end QoS will be assured by a global QoS management agent. The objective of this infrastructure is to facilitate network operators to tune their networks with a great flexibility and scalability to achieve their own operational objectives.

BBQ takes resource reservation approach to ensure QoS. However, the resource utilization of reservation approach may not be very high due to the reasons such as "no show" effect and busty nature of some services. Thus, one of our design challenges is to maximize the resource utilization without incurring too much real-time overhead.

## 1.4. Over-Booking and Over-Admission

In order to reduce real-time response time, many of the management mechanisms in BBQ, such as resource allocation and reservation, take pre-planning approach, instead of real-time on-demand approach. Each edge router in a core network is pre-allocated with some short-path resources, which are paths from edges to edges of the core network. Short-paths are then allocated to individual incoming service requests on a reservation basis when they are admitted to the network.

In order to compensate the "no show" effects caused by forecasting error, BBQ uses *over-booking* technology to improve the resource utilization [7]. On the other hand, some types of traffics such as telephony traffics may not be persistently active during its life time leaving part of reserved resources unused. To solve this problem, we propose an *over-admission* method in which each edge router can admit more traffics than the capacity it actually acquired.

## 2. OVER-ADMISSION

In BBQ, each edge router controls some short-paths which will be used by incoming traffics on a reservation basis. It executes admission control to allocate/deallocate short-paths to/from incoming requests. The admission procedure is shown in Fig. 2. When a service request enters the network, the Global Admission Control Agent (GACA) at the access network performs the admission control to select an end-to-end path from the source to the destination and then submit the reservation requests to the admission controllers of all the short-paths that are included in the selected end-to-end path. The admission controller of a core network will grant the admission if the requested short-path is available. An end-to-end path will be established if reservation requests are all successful.

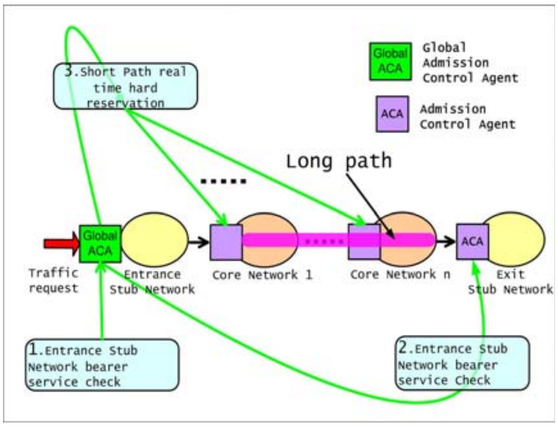
3.Short Path real time hard reservation
Global ACA
Global Admission Control Agent
ACA
Admission Control Agent
Long path
.....
Global ACA
Traffic request
Entrance Stub Network
Core Network 1
Core Network n
ACA
Exit Stub Network
1.Entrance Stub Network bearer service check
2.Entrance Stub Network bearer service Check

Fig. 2 Admission Procedure

Some types of traffics such as telephony traffics may not be persistently active during its life time leaving part of reserved resource unused. In the conversations over a normal phone call, users usually talk alternatively. It is unusual that both users talk simultaneously. Therefore, the resource utilization of telephony services is usually lower than 50% [2].

Assuming a voice terminal doesn't generate any packet during silence periods, BBQ takes this advantage to enhance resource utilization by *over admission*, which allows edge routers admit more traffics than whatever resources they have. The challenge is to find the maximum level of over admission without incurring too much traffic overflow.

Because telephony service will continue to be the most popular venue-generation services in the near future, we study telephony services first for simplicity. We assume that each telephony session is on-off exponentially distributed. Followings are the notations used in our analytic model:

| | |
|---|---|
| $Z$ | number of admitted telephony sessions |
| $y$ | number of active telephony sessions |
| $p$ | probability of a telephony session being active |
| $C_1$ | per unit time profit of an admitted telephony session |
| $C_2$ | per unit time penalty of an overflowed telephony session |
| $T$ | active time of a telephony session, an exponentially distributed random variable with a mean value of 1/λ |
| $T'$ | inactive time of a telephony session, an exponentially distributed random variable with a mean value of 1/λ' |

The objective of optimal admission control is to find the optimal admitted amount, *Z*, such that the average net profit is maximized. For simplicity, we assume the statistical characteristics of voice traffics is time independent so that the problem can be easily solved if the time-independent probability of overflow as a function of *Z* is known. We also assume that the holding time of each individual voice session is longer than the concerned time period. This assumption makes the calculation of the earned profit ($C_1$ times *Z*) for an admitted session and the overflow probability much easier. Further, the penalty for the overflow is assumed linearly proportional to the number of overflowed voice sessions.

Given the parameters and assumptions described above, the overflow probability can be easily calculated using Eq. 1.

$$P(Z,y)=\binom{Z}{y}(p)^{y}(1-p)^{Z-y}, p=\frac{1}{\left(1+\lambda/\lambda'\right)}, p>0 \qquad (1)$$

Likewise, the net profit can be calculated using Eq. 2.

$$C1*Z-C2*\sum_{y=C+1}^{Z}\left\{(y-C)*\binom{Z}{y}(p)^{y}(1-p)^{Z-y}\right\}, p=\frac{1}{\left(1+\lambda/\lambda'\right)} \qquad (2)$$

The optimal number of sessions to be admitted can then be easily calculated in linear time.

## 3. PERFORMANCE EVALUATION

The proposed over-admission method was evaluated by simulation using the UCB NS2 network simulator. The evaluation metrics are the net profit and the ratio of over-admission, which is the ratio of the over-admitted voice sessions to the capacity. The objectives of the experiments are three folds: (1) verification of analytical model; (2) performance evaluation of over-admission; and (3) assessment of robustness.

When a network manager aggressively adopts both over-booking and over-admission methods, it is possible that total incoming traffics exceeds the total capacity. The is equivalent to that parts of resources acquired by edge routers are revoked by the Bandwidth Broker implicitly without notifying edge routers. As a consequence, these edge routers my over-admit too many traffic sessions. In order to encourage the system manager to use over-admission method, it is important to assess the impact of over-admission when the assumed capacity is not fully available.

## 3.1. Experimental Environment

Three experiments were conducted in the simulation study to fulfill the objectives. The robustness is assessed by evaluating the performance when the actually allocated resources are less than the assumed capacity. Fig. 3 shows the basic topology used in the simulation. Experimental parameters are listed in Table 1.

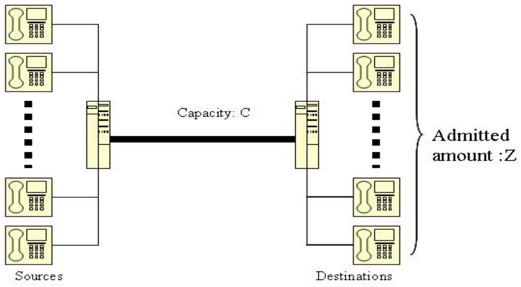

Fig. 3. Basic Topology in Simulation

Table 1. Simulation Parameters

| Parameter | Value Range |
|---|---|
| **Active Ratio** | 0.1-0.9 |
| **Mean Holding Time** | 2,5 min |
| **Capacity** | 10-30 |
| **$C_1$: $C_2$** | 1:2, 1:3, 1:4 |
| **Admitted Amount** | 20-150 |

Table 2. Calculation of Profit and Penalty

| | | |
|---|---|---|
| **In Analytical Model** | Profit | *$C_1$* * number of admitted sessions * mean holding time * PPS |
| | Penalty | *$C_2$* * number of overflowed sessions * mean holding time * PPS |
| **In Simulation** | Profit | *$C_1$* * accumulated holding time * PPS |
| | Penalty | *$C_2$* * total number of lost packets |

The calculation of profit and penalty are shown in Table 2, where PPS stands for packet per second.

## 3.2. Experiment Results - Analytical Model Verification

The discrepancy between the analytical model and the simulation results are shown in Table 3 and Fig. 4. As we can see that the errors are nominal especially at the optimal over-admission levels and below.

Table 3. Analytical Errors

| **Active Ratio** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** |
|---|---|---|---|---|---|---|---|
| **Avg error in net-profit** | 0.09% | 1% | 2% | 5% | 5% | 6% | 8% |
| **Max error in net-profit** | 0.5% | 4% | 6% | 8% | 13% | 12% | 12% |
| **Error in optimal Z** | 3% | 3% | 2% | 2% | 2% | 1% | 1% |
| **Error in max profit** | 6% | 5% | 5% | 4% | 2% | 1% | 0% |

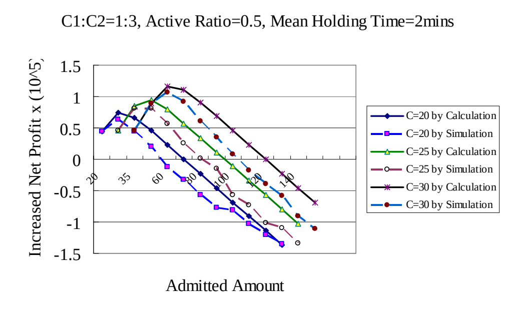

Fig. 4. Increased Profit Ratio at 0.5 Active Ratio

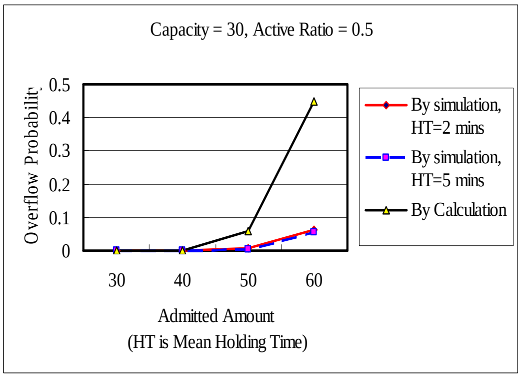

Fig. 5. Overflow Probability

Fig. 5 shows the overflow probability at 0.5 active ratio. As we can see from the figure that the analytical model tends to overly estimate the overflow probability when the over-admission level is high. Thus estimation errors won't hurt the system quality but only discourage the level of over-admission.

## 3.3. Experiment Results - Performance Evaluation

The results of performance evaluation are shown in Fig. 6 and 7. As we anticipated, the lower the active ratio, the higher the over-admission level. Performance improved by approximately 100% at 0.5 active ratio and by 600% at 0.1 active ratio.

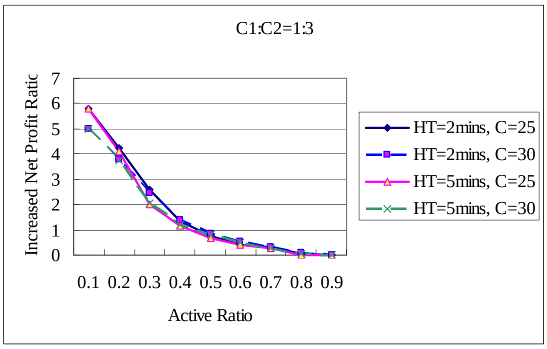

Fig. 6. Increased Admission Ratio

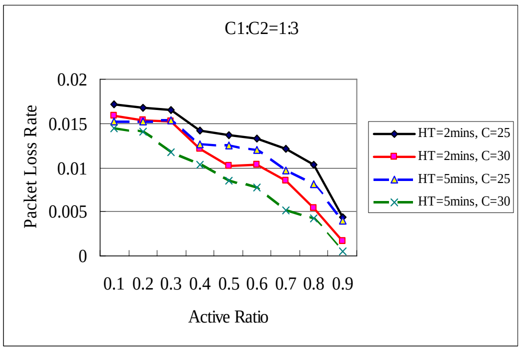

Fig. 7. Increased Net-Profit Ratio

## 3.4. Experiment Results – Robustness Assessment

The results of robustness assessment are shown in Fig. 8 and 9. As we can see that over-admission can still offer performance improvement even if the acquired resources are 6(8) units less than what it expected when the assumed capacity is 20(30).

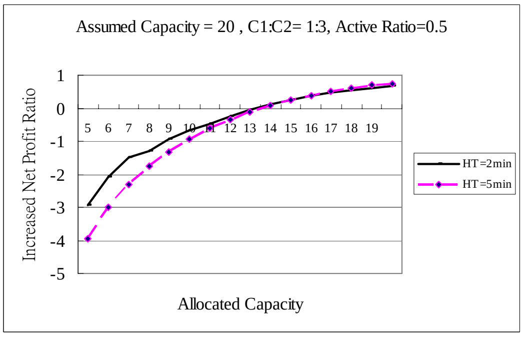

Fig. 8. Results of Robustness Assessment (C=20)

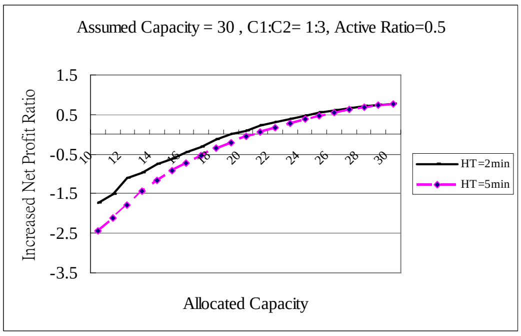

Fig. 9. Results of Robustness Assessment (C=30)

## 4. CONCLUDING REMARKS

Because some network traffics such as telephony traffics may not be persistently active during its life time leaving part of reserved resources unused, we propose an *over-admission* method to enhance resource utilization by admitting more traffics than the capacity an admission controller actually acquired. We developed an analytical model that can be used for estimating the optimal level of over-admission without incurring too much overflow. Simulations show that our analytical model is accurate and this method can enhance resource utilization effectively. For regular voice traffics, near 100% improvement can be reached when their active ratios are near 50%. Assuming silence depression is adopted in encoding a voice stream, a rule of thumb in applying over-admission technology is to admit two voice sessions for each unit of acquired resource for voice. Simulations also show that over-admission can still offer performance improvement even if the acquired resources are not actually fully available. Network managers can aggressively adopt our method without worrying too much about the possible penalty caused by miscalculation.

Although this research was done under BBQ environment, the developed method is applicable in many different recourse reservation environments. In fact, a similar technique, concentration, has been used in local loops.

## REFERENCES

1. D. Black, M. Carlson, E. Davies, Z. Wang, "An Architecture for Differentiated Services", RFC 2475, Dec. 1998.
2. P. Brady, "A Model for On-Off Speech Patterns in Two-Way Conversation," *Bell System Technical Journal*, Sep. 1696.
3. Janusz Gozdecki, Andrzej Jajszczyk, and Rafal Stankiewicz, "Quality of Service Terminology in IP Networks", *IEEE Communications*, Mar. 2003.
4. IETF RFC 1633, Integrated Service Framework (IntServ).
5. IETF RFC 2205, Resource reSerVation Protocol (RSVP).
6. Yao-Nan Lien, Hung-Chin Jang, Tse-Chieh Tsai and Hsing Luh, "BBQ: A QoS Management Infrastructure for All-IP Networks", *Communications of IICM*, vol. 7, no. 1, Mar. 2004, pp. 89-115.
7. Yao-Nan Lien, Yi-Min Chen, "Forecasting Error Tolerable Resource Allocation in Budget-Based QoS Management for All-IP Core Networks", Proc. of the 3rd IEEE International Conference on Information Technology: Research and Education, June 27-30, 2005.