Flow-Aware Link Dimensioning for Guaranteed-QoS Services in Broadband Convergence Networks

Hoon Lee and Khosrow Sohraby

Abstract: **In this work,[1] we propose an analytic framework for dimensioning the link capacity of broadband access networks which provide universal broadband access services to a diverse kind of customers such as patient and impatient customers. The proposed framework takes into account the flow-level quality of service (QoS) of a connection as well as the packet-level QoS, via which a simple and systematic provisioning and operation of the network are provided. To that purpose, we first discuss the necessity of flow-aware network dimensioning by reviewing the networking technologies of the current and future access network. Next, we propose an analytic model for dimensioning the link capacity for an access node of broadband convergence networks which takes into account both the flow and packet level QoS requirements. By carrying out extensive numerical experiment for the proposed model assuming typical parameters that represent real network environment, the validity of the proposed method is assessed.**

Index Terms: **Broadband convergence networks (BcN), flow-aware networking, link dimensioning, quality of service (QoS) assurance.**

I. INTRODUCTION

A. Motivation

Recently, we can witness a fast movement in broadband telecommunication networks toward the convergence of broadband wired and wireless access networks via a common IP backbone network as a unified transport platform, which brought forth an era of broadband convergence networks (BcN). Typical examples of broadband wired access network are asymmetric digital subscriber line (ADSL), Ethernet-passive optical network (E-PON), and cable access network (CAN). The packets generated from an ADSL connection follow the same logical and physical paths from customer premises equipment (CPE) to an access node via digital subscriber line access multiplexer (DSLAM). When voice over IP (VoIP) traffic and web traffic share the same link such as an asynchronous transfer mode (ATM) virtual circuit (VC), and if congestion occurs in DSLAM link, the delay of VoIP packet is not guaranteed. This necessitates an adoption of flow-aware allocation of bandwidth for each connection such as a constant bit rate (CBR) VC for a VoIP flow and an available bit rate (ABR) VC for a web flow.

When it comes to a cable access network, a number of connections share the bandwidth of a coaxial cable. Without control for quality of service (QoS) differentiation, packets are served in first in first out (FIFO) manner. However, each customer wants to receive a certain level of QoS in terms of minimum guaranteed bandwidth without interruption from the other connections. To resolve this problem, the data over cable service interface specification (DOCSIS) introduced an enhanced media access control (MAC) protocol with the concept of *service flow*. The concept of *service flow* is not necessarily based on flow, where flow has several different meaning in the context of telecommunication network and we will define it as follows: Flow is defined as an identity such as a connection that has been issued by a customer with a generic source and destination identification. A flow may include a long-duration voice connection, a short-duration data transfer, or a back-to-back transfer of video stream, etc.

An aggregate of flows with the same QoS requirements can receive a certain level of QoS (such as a limited delay) in an aggregate manner if the offered load to that service flow is kept to a certain level [1].

Recently, WiMax, a broadband wireless access network for a metropolitan area network based on the IEEE 802.16 protocol, is about to commence a commercial service. In IEEE 802.16 network, it is assumed that a centralized controller allocates time slots in connection-oriented manner to a large number of customers by use of slotted time division multiple access (TDMA) scheme. Therefore, bandwidth is requested by a connection basis, and so it is allocated to each connection or CPE. This is represented by a *grant* in a distributed coordination function or it is specified as a weight of transfer rate in centralized coordination function using the polling scheme, both of which are based on CPE or connection identification [2].

Sinha *et al.* investigated the traffic behavior of upstream traffic from the operating commercial broadband access networks incorporating wired and wireless user accesses, ADSL for the former and broadband fixed wireless (BFW) for the latter. Even though the upstream channel of ADSL is dedicated and that of BFW is shared, they found that the attributes of user traffic behavior such as the volume of flow in terms of packet count and byte size as well as the duration of flows were strikingly similar between the two types of access networks [3].[2] This indicates that a universal and systematic bandwidth dimensioning method has to be provided in the design of the access network of BcN.

From the findings in the behavior of flows of the three network protocols as well as the user attributes that have been described above represent that a flow-aware bandwidth allocation is required to a customer if a minimum level of QoS (such as delay and throughput) is provided to each connection from a number of competing customers. Even though it is impossible to control the data transfer for each flow separately in a large

Manuscript received March 22, 2006.

H. Lee is with the Changwon National University, Changwon, Gyeong-Nam, 641-773, Korea, email: hoony@changwon.ac.kr.

K. Sohraby is with the University of Missouri-Kansas City, Kansas City, Missouri, 64110-2499, USA, email: sohrabyk@umkc.edu.

[1] This work was carried out while Hoon Lee was a visiting scholar at UMKC.

[2] See Fig. 9 in [3].

backbone network, it would be possible to dimension the bandwidth in flow-aware manner at least at the access network.

Recently, Cerra *et al.* proposed the concept of user-centric broadband services (UCBC), and they advocated that network edge and core is provisioned in service-aware and data-aware manner, respectively [4]. Service and data aware transport implicates the flow-level and packet-level transport, respectively.

On the other hand, the network core can be also provisioned in a flow-aware manner by using the concept of label switched paths (LSPs) defined by generalized multi-protocol label switching-traffic engineering (GMPLS-TE) over wavelength division multiplexing (WDM) path between edges of core network.

The above mentioned research movement motivated our research for the flow-aware network dimensioning in BcN access network which is composed of a wired and/or wireless network interfaces.

B. Related Works

We can find a number of literature concerning the architecture and management for dimensioning the VCs in the world of ATM network (see [5], [6]). The concept of VC or virtual path (VP) had contributed very much in the development of the concept of tunnels and LSPs in GMPLS-TE for the IP network, and the concept of VC in ATM network can be also used in the IP network if we assume the event of a series of user data frame as a flow and if we consider the IP network as a flow-aware network.

On the other hand, in the world of IP network, there exist few results for the development of the analytic model and experiment of the network performance using the concept of VCs, because flow-level model has been regarded as inefficient due to a number of limitations such as low utilization and scalability problem. Low utilization results from the inherent property of connectionless data transfer of IP network, because the IP data lasts a very short time. Scalability problem results from the signaling requirement in the set up of a path for each flow before the transfer of data.

However, it is usual that the user access network operates in high utilization due to over-subscription of customers. In addition, the number of customers in the access network is not as huge as that of core network. Therefore, the concept of flow-awareness can be carefully adopted in the access network.

The concept of flow-aware networking (FAN) proposed by Roberts *et al.* advocates connection admission control (CAC) on the basis of a flow in order that insurance for QoS is provided even in situations of overload. This necessitates a certain degree of over-provisioning in link dimensioning by noting that adequate performance is always assured for services emitting flows at a stable rate but less than the limit on fair rate of a shared network resource if the network is maintained by an appropriate CAC scheme [7].

There exist a few literatures on the concept of flow-aware link dimensioning in IP network. One example is circuit-switched high-speed end-to-end transport architecture (CHEETAH) proposed by Veeraraghavan [8]. In a CHEETAH architecture, multi-service provisioning platforms in enterprises which carry voice and data traffic between geographically distributed buildings use a high-speed end-to-end transport network operated in a circuit-switched manner. CHEETAH assumes that a connection request can be blocked if there is no available bandwidth resource for that flow between an end-to-end path, and an Erlang-B formula is used to determine the call admission.

In the field of IP network, the processor sharing (PS) model has been frequently used as an equivalent flow-aware network model in that a processor is shared by a number of customers in a time-shared or stochastic manner [9]. Recently, an approximate M/G/1-PS queuing model for estimating the loss probability of a customer is given by Boyer *et al.* [10], where loss probability of a customer is estimated by assuming the customer's delay tolerance using the concept of user impatience in the system sojourn time. Boyer *et al.* have shown that the loss probability of a customer obtained by an M/M/1-PS queuing model is the same as that of M/M/1 queue when the threshold of user impatience in an M/M/1-PS queuing model is greater than or equal to the maximum allowable number of customers the system can support.

If we assume that the duration of a flow is assumed to be exponentially distributed,[3] we can use a classical M/M/1 model for the flow-aware network instead of using a complex M/M/1-PS queuing model under a certain network environment where there exists no explicit connection admission control for a flow, which can be seen in the current Internet. From this finding, we can approximate the behavior of IP flows that is usually modeled by an M/M/1-PS queuing model by an equivalent M/M/1 model.

Note that M/M/1 model has been looked upon as a very convenient method in modeling the elastic type services such as the best effort high-speed Internet access at the packet or flow level. However, it is not suited to the modeling of the flow-aware applications of stream type services such as a voice or videophone. If this model can be used in modeling both types of services, we can argue that it is very useful for this model to estimate the flow-level behavior of converged broadband multimedia services. On the other hand, let us remind that M/M/1 model is easily approximated by using the M/M/c model, where c is the number of identical servers, if the server capacity of M/M/1 model is c-times greater than that of M/M/c model.

Recently, Chan *et al.* proposed a new method to solve the waiting time distribution of M/M/c queue by assuming the M/M/c queue as an equivalent M/M/1 queue with server capacity equal to c-times that of M/M/c queue, and proved that his result is in good agreement with that of Takac's result for the waiting time distribution for Palm input and exponential service times for the G/M/c queue when the input is Poisson [11]. Chan's method is the reverse of our method: Chan solved the waiting time distribution of M/M/c queue by assuming an equivalent M/M/1 queue with server capacity equal to c-times that of M/M/c queue. On the other hand, we present the waiting time distribution of M/M/1 queue by assuming an equivalent M/M/c queue with capacity of a single server equal to $1/c$ of M/M/1 queue.

[3] The flow duration of current telephone service is exponentially distributed, and so will be in case of VoIP or videophone. On the other hand, some applications such as web browsing, on-line game, or VOD are not necessarily so. However, let us assume the exponential property of the flow duration for the purpose of mathematical tractability.

C. Contribution and Organization

Our work is in line with the above-mentioned philosophy of connection-oriented network engineering. However, there exist a number of differences in our research compared with the past research. We assume neither resource reservation nor segregation of network resource per each flow in a physical manner. We assume a more realistic network environment which takes into account the BcN access network as a universal broadband service platform by assuming a logical VC as an abstract of a unit of bandwidth granule per flow in a shared big pipe. We also take into account the inherent attributes of user behavior such as the impatience and intolerance of users due to large delays in access network. We also target almost every kind of services composed of connection-oriented and connection-less applications.

In this work, let us propose an analytic method to dimension the link capacity of a subscriber node in BcN. To that purpose, we take a two-step procedure. First, we propose a method to estimate the required VCs of an output port of an L2 switch under the predefined flow-level delay QoS measures for the patient and impatient customers. We derive a formula for the tail distribution of waiting time by use of M/M/c queuing model, especially extending the result of the Erlang-C formula for the patient customers and using the Erlang-A formula for the impatient customers. Second, we use the concept of effective bandwidth of Guerin, and present a method to estimate the required bandwidth of an output port of an L2 switch under the predefined packet-level packet loss measure as well as the flow-level delay measure. The result of former method is incorporated into the latter method, so that our method is a two-step procedure. Via numerical experiments, we will show the following results. First, we present the number of VCs and the total bandwidth capacity of an output port of an L2 switch under the required packet loss probability as well as the required delay tolerance of the flow. Second, we give a result about the average utilization of a link for the proposed dimensioning method, via which we can provide guidance for the link dimensioning of a switch router which accommodates a number of customers in the BcN access network.

The rest of this paper is structured as follows. In Section II, the concept of flow-aware dimensioning in IP network is introduced. In Section III, an analytical model for the dimensioning of flow-aware access network for the BcN is proposed. In Section IV, the performance evaluation and numerical result are presented using the proposed model. In Section V, a brief description on the implementation aspect of the proposed method is given. Finally, in Section VI, we summarize the findings and implication of the work.

II. FLOW-AWARE LINK DIMENSIONING

A. Flow-Awareness in IP Network

Before proceeding with the discussion on the link dimensioning, we need to clarify definitions for the notion of the conventional packet-aware network and the flow-aware network. The conventional packet-aware network allocates the network resource such as bandwidth based on the unit of packet, while flow-aware network operates based on the unit of flow. The notion of packet in IP network is well known, so that we need no further explanation. Concerning the notion of flow, on the other hand, there exist a lot of concepts: To name a few, see [7], [12]–[14].

Let us define a flow in a general manner. Flow is a logical group of packets that have common properties such as a source-destination (SD) host pair with unique addresses or subnets and port numbers for source and destination [15]. Note that this concept is distinguished from the conventional concept of a flow, which is usually defined to be a sequence of packets between individual source-destination applications such as TCP or UDP stream.

When a flow is distinguished by source and destination addresses it can identify an individual customer. On the other hand, when it comes to subnets, we can define a flow as an aggregate of customers from a VPN site. Therefore, the concept of flow is very general, and we can exploit this generality of the concept of flow in dimensioning the link capacity of diverse type of links in IP access network.

In the conventional wired or wireless telephone networks, it has been usual that the network operator estimates the capacity of network resource such as the bandwidth as the number of customers that can be accommodated simultaneously to a link with a certain throughput guaranteed to each user under the constraint of a limited call blocking probability[4] such as 1% or 0.1% [16]. In wireless access network, it is usual that the number of flows that are active simultaneously is recognized as the number of channels, and the number of channels is also computed by taking into account the required call blocking probability. These two examples show that the concept of flow-aware dimensioning of the wired and wireless bandwidth resources in BcN access networks is necessary.

In the world of IP network, especially in the backbone domain, a virtual network (VN) approach which adopts the concept of traffic trunk is proposed as one of a scalable and practical approach to deploy an efficient resource management scheme [17]. In the VN approach, the authors argue that a path-oriented bandwidth dimensioning approach is useful in applying the MPLS-based traffic engineering, which is based on the concept of traffic trunk. Traffic trunk is a logical pipe within an LSP, which allocates a certain amount of link capacity associated with a certain class of services. The concept of VN is born from a composition of the IETF integrated service (IntServ) and differentiated service (DiffServ) concept: A traffic trunk provisioned between two network edges is leased from the concept of IntServ, whereas the class-based traffic aggregation and handling is leased from the concept of DiffServ. By way of traffic trunk, which is calculated independently from different service classes, we can easily estimate the total link capacity of a node if we know the number of concurrent traffic flows that belong to the same service class at a time of network configuration. If we assume that the bandwidth resource required by each flow is represented by the well-defined effective bandwidth, bandwidth usage and left-over link capacity of each traffic trunk can be obtained by tracking the arrival and completion of flows [17].

[4] Call blocking probability was also called grade of service (GoS) in telephone network.

B. Equivalent Flow-Level Network Model

Those two examples described in the previous subsection imply that it is efficient if the network dimensioning is carried out on the basis of flow unit. To that purpose let us utilize the concept of Russian doll model (RDM) of IETF [18]. In RDM, a big bandwidth pipe with capacity C is decomposed into a number of sub-pipes with capacity of B, where $B \ll C$. Each sub-pipe in RDM accommodates packets aggregated from the same type of service class, where service class follows the taxonomy of IETF DiffServ architecture.

For example, a sub-pipe for the services that belong to the expedited forwarding per-hop behavior (EF-PHB) such as VoIP accommodates a number of VoIP flows over a single logical link in FIFO manner, where the VoIP flows are not interrupted by the other classes of flows.

On the other hand, a sub-pipe for the services that belong to the assured forwarding per-hop behavior (AF-PHB) such as a web application accommodates a number of web flows over a single logical link in FIFO manner, where the web flows are not interrupted by the other classes of flows, either. Without loss of generality, let us assume that the traffic characteristics of the sources are homogeneous if they belong to the same class of service. It is known that the nodal performance of end-to-end delay and loss in a network composed of a number of consecutive nodes is equal to each other if the offered load of each node is equal to each other [19]. This implicates that we can implicitly identify the performance of a flow from an aggregation of flows when the traffic attributes of each flow is homogeneous in the node. Then, we can represent the flows and the bandwidth for each class in RDM model as illustrated in Fig. 1 where there exist a number of flows at the input side of a node and a link is accommodating those homogeneous flows at the output side of a node. The model in Fig. 1 can be used to model both the upstream and downstream flows. But, without loss of generality, let us assume that the direction of a flow is upstream.[5]

In Fig. 1, the capacity of an output link is usually provisioned to be greater than the sum of the mean or statistical traffic generation rate of all the input links, so that the network operates in a stable manner. Even in that case the probability of a flow waiting in the queue before it receives service from the output link is greater than zero because the arrival and completion behavior of a flow is random. Let us call it a packet-level aggregate big-pipe (PLAB) model.

In Fig. 2, there exist s output links for a queue, whereas the capacity of each output link B is the same as that of an input link in either statistical or deterministic manner. In this case, the probability of a connection waiting in the queue so that it receives service from the output link is greater than zero, too, because the number of incoming customers will vary with respect to time and there may exist a case when $N > s$. Let us call it an equivalent per-flow link (EPFL) model.

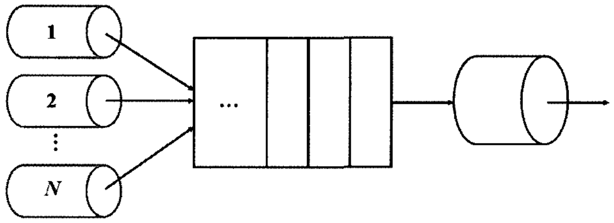


Fig. 1. PLAB model.

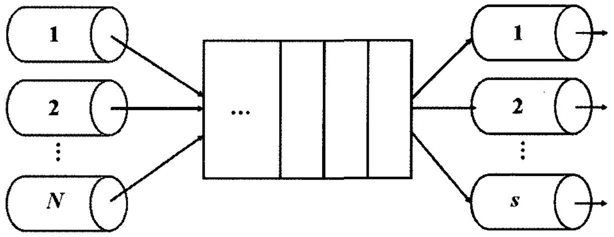


Fig. 2. EPFL model.

Note that Chan's model in [11] is simulating the EPFL model into PLAB model, whereas our model is the reverse of Chan's model. Therefore, note that the following condition holds: $C \approx s \times B$, which indicates that the two quantities in the equation are equivalently equal to each other.

Note that our analysis is focused on the statistical behavior of the delay a flow is expected to experience before it receives service. On the other hand, it is concretely proven in the field of queuing theory that the mean delay performance of M/M/1 single-server queue with service capacity C is slightly better than that of M/M/2 two-server queue with service capacity $C/2$ for each server [20]. Then, we can deduce from this fact that the mean delay performance of M/M/1 queue with service capacity C is slightly better than that of M/M/s queue with service capacity C/s, where s is the greatest integer which is not greater than C/B.

From this fact, one can find that the equivalent EPFL model can be used as an approximated performance estimator of the conventional PLAB model, where the delay performance is estimated from a conservative view point. However, one has to notice here the following fact: *The output link is not actually segregated into s different VCs in EPFL model, rather VC is only a conceptual identity which has bandwidth with capacity equivalent to 1/s of an output link of PLAB model.* It is usually known that there exists a network environment where flow-aware or packet-aware link models are suited. First, flow-aware link model like EPFL model makes sense in the networks that accommodate a large number of simultaneous active users, where users are distributed densely in a region. On the other hand, the PLAB model is more suited in modeling a network that accommodates a small number of active users distributed sparsely [21].

III. Mathematical Model

A. Dimensioning the VCs

Let us present a method to estimate the number of VCs for an EPFL model. In EPFL model the basic unit of network resource is a *logical bandwidth* unit called VC that has the same analogy of circuit in the telephone network as we have shown in Fig. 2.

[5] The bandwidth required by a flow or available by network is asymmetric for upstream and downstream, so that link design has to be carried out separately.

The *'logical bandwidth'* means that each flow is recognized by a flow requiring a certain amount of bandwidth, and it is accepted to the network if there is an available bandwidth in the network. Otherwise, it is rejected from the network. At once a flow is accepted to the network, it grabs a logical bandwidth provided by a VC until the flow is terminated. Therefore, this is analogous to the circuit or VC in conventional circuit switched network such as telephone network or ATM network.

Note that there exist two kinds of customers concerning the tolerance of network delay: Patient and impatient customers. Patient customers can tolerate a certain amount of delay before it receives a service, so that he enters a network if he wants to receive a service irrespective of the state of the network, whereas impatient customers do not tolerate a long delay and may give up the connection to the network if he thinks that the expected waiting time exceeds a certain value.[6] When the impatient customer retries a connection, it is recognized as a new connection request.

A.1 Analysis for Patient Customers

First, let us define the customer's patience in more detail. We call that a customer is patient if he does not mind the status of the network, especially the network congestion, and he enters the network if he wants to receive the service. However, let us extend the concept of patience in more extensive manner such that a customer can abandon the connection if he can not receive the service from the network even though he has been waiting for a certain time, so that a customer is not completely patient. When a network accommodates completely patient customers, the network may not be exempted from the continued congestion. In order to clarify the concept of customer's patience, let us define the customer's patience as follows.

A customer enters the network without minding the status of the link upon his arrival. When his waiting time in the queue is greater than T, then he abandons his connection. Otherwise, the connection is admitted and receives a service.

If we assume that the distribution of the number of flow requests follows a Poisson process and the duration of the flow, which is characterized by the amount of data the users transfer, is distributed exponentially, we can use the result of M/M/s queuing system in modeling the network with patient customers. Via queuing analytic method, we can obtain the probability of a customer waiting in the queue, which can be formulated by $\Pr\{W > 0\}$, where W is the waiting time of a flow. This is a well-known Erlang-C formula [20].

Erlang-C formula is usually represented as follows. Let λ be the average arrival rate of a flow, and let $1/\mu$ be the average duration of a flow. Then, the traffic intensity is given by $\rho = \lambda/\mu$. Let us denote that the occupancy level of customers relative to the number of servers is given by $\varphi = \rho/s$, which is a traffic intensity to a VC among a total of s VCs.

The probability of arriving customer waiting in the queue, which is denoted by $E_C(s, \rho)$ is given by

$$E_C(s,\rho) = \Pr\{W > 0 | \text{traffic load is } \rho \text{ and the number of server is } s\}. \quad (1)$$

Then, (1) is given as follows [20].

$$E_C(s,\rho) = \frac{\rho^s/s!}{\rho^s/s! + (1-\varphi)\sum_{i=0}^{s-1} \rho^i/i!}. \quad (2)$$

Note that it is not an easy job to compute (2) when s becomes very large. Fortunately, it is usual that the number of concurrent flows, which corresponds to s in the current setting, is not so large at the access node of current subscriber network, so that (2) is scalable to computation.

On the other hand, one can easily notice that it is not useful to use $\Pr\{W > 0\}$ in estimating the delay performance of an elastic traffic for IP network, because users can tolerate a certain amount of time greater than zero before they are served. Therefore, it would be more practical if we rewrite the QoS measure such that a flow can wait in the queue for a certain time period, say T, and let us denote it as shown in (3), and call it as a delay tolerance.

$$\Omega(T) = \Pr\{W > T\}. \quad (3)$$

For M/M/s queuing system, the memoryless property of the exponential service time gives us a simple formula for the probability of the waiting time to be greater than T as a function of $\exp[-(s-\rho)\mu T]$ when there is $(s-\rho)$ available spaces in the server. Using this fact, let us introduce a probabilistic measure for the delay performance such that the following relationship exists.

$$\Omega(T) \leq \varepsilon. \quad (4)$$

where ε in (4) is the target value for the probability of QoS violation, which we call a delay tolerance. This implies that the probability that a flow has to wait in the queue at least for T seconds before it is served is bounded by ε.

Using the Erlang-C formula and the memoryless property of the exponential service time we can easily obtain the probability of waiting time greater than a certain threshold T, which is given as follows.

$$\Omega(T) = E_C(s,\rho)\exp(-(s-\rho)\mu T). \quad (5)$$

From (4) and (5), we can obtain a relationship between traffic intensity ρ, number of server s, delay target T, and delay tolerance ε in the network with patient customers, which is given by (6).

$$E_C(s,\rho)\exp(-(s-\rho)\mu T) \leq \varepsilon. \quad (6)$$

Note that it is difficult to draw out an explicit formula for s, the number of servers that is required to satisfy the relationship (6). Therefore, a numerical method has to be used in computing (6). The input parameters are traffic intensity ρ, delay target T, and delay tolerance ε, whereas the output is the number of server s.

[6] We assume that an arriving customer has an information about the expected delay of the network through the process of service level negotiation.

A.2 Analysis for the Impatient Customers

In the real field where service level agreement (SLA) and pricing is negotiated between service provider and customers, it is assumed that a customer gives up a connection to a network when he expects that the guarantee of a certain level of delay from the network can not be met. So, let us assume that a customer is impatient when he abandons the connection to the network for the following cases: First, a customer abandons a connection if a certain time interval, say T seconds, has passed before he receives a response from the network. Second, a customer abandons a connection if he knows that he has to wait T seconds before he receives a response from the network. As to the former case, a customer has to enter the network at once, and he has to see the result by waiting for T seconds in order to know that whether his expectation is satisfied or not. So, this case is similar to the case of patient customers that we have described in the previous subsection. On the other hand, as to the latter case, a customer doesn't have to enter the network before he determines to receive the service. A network operator announces the expected waiting time of a customer upon the customer's arrival to the network, and the customer determines whether he enters the network or not from the information provided by the network. A true impatient customer belongs to the latter case, and we will consider this latter case in this subsection.

If we assume that the distribution of the flow request follows a Poisson process and the duration of the flow is distributed exponentially, we can use the result of M/M/s-M queuing system in modeling the network with impatient customers. In M/M/s-M queuing system, the three first symbols have the same meaning as in the conventional Kendall's notation, whereas the last one specifies *the law of customer's impatience* [22]. So, the last M in M/M/s-M model stands for an exponential impatience of a user, where the degree of impatience of customers grows exponentially with respect to the lapsed time.[7] M/M/s-M queuing system is first proposed by Baccelli *et al.* [22], and it is used in modeling the optimal number of agents for a call (customer service) center in the telephone network [23]. By using approximate analysis for the probability of delay in the queue for the M/M/s-M queuing system, Garnett obtained a formula for the relationship between the offered load and the distribution function of the delay, and he named it as an *Erlang-A formula*, where A stands for *abandonment* (of a call) [23]. Garnett used an intuitive analytic method by assuming a parameter called a safety staffing, which corresponds to an excess capacity needed to achieve the target service level such as the call blocking probability.

On the other hand, it is not an easy job to determine the safety staffing parameter from traffic point of view in IP network. Therefore, we use Baccelli's method to obtain a formula for the probability that the waiting time of a customer is greater than T, so that the impatient customers abandon his connection.

Let us assume that a network operator estimates the expected waiting time of a customer periodically or on demand, so that an arriving customer can be noticed about his expected waiting time upon his arrival to the network. A customer determines whether he enters the network or not based on the following way.

When a customer hears upon his arrival to the network that the expected waiting time of his flow is not greater than T, he enters the network and waits for service. Otherwise, he abandons the connection and leaves from the network, and he retries later as a new flow.

Let us assume that $D(X)$ is the distribution function of a random variable that indicates the time of the customer's impatience, and let T be the first order moment of $D(X)$, which is analogous to the delay parameter for the case of patient customers.

Let $E_A(s,\rho)$ be the equilibrium probability of an arriving customer who decides not to enter the system because his expected waiting time is supposed to be greater than T. Then, from [22], we can obtain the following formula for $E_A(s,\rho)$.

$$E_A(s,\rho) = P_0 \frac{\rho^{s-1}}{(s-1)!}\{1 + (\frac{\rho}{s} - 1)(\frac{1+\alpha}{1+\alpha-\rho/s})\} \qquad (7)$$

where $\alpha = 1/(T\mu s)$ and P_0 satisfies the following normalization condition.

$$P_0 = \frac{1}{1+\rho+\cdots+\frac{\rho^{s-1}}{(s-1)!}+\frac{\rho^s}{s!}\frac{1+\alpha}{1+\alpha-\frac{\rho}{s}}}.$$

Finally, we can obtain a relationship between traffic intensity ρ, number of server s, delay target T, and delay tolerance objectives ϵ in the network with impatient customers, which is given as follows.

$$E_A(s,\rho) \le \varepsilon. \qquad (8)$$

where $E_A(s,\rho)$ is given in (7). Note that the formula (8) is different from the formula (6). This is because the definition of $E_A(s,\rho)$ is different from that of $E_C(s,\rho)$: The former is defined to be $\Pr\{W > T\}$ from the beginning, whereas the latter is defined to be $\Pr\{W > 0\}$.

B. Dimensioning the Bandwidth

Since the appearance of ATM network, the concept of effective bandwidth has been recognized as one of a few effective methods to represent the statistical bandwidth amount required to guarantee the packet-level QoS such as the packet loss and delay for a source with relatively small number of traffic parameters. On the other hand, we argue that one can use the flow-level model of Erlang formula in estimating the number of flows from the required flow blocking probability if one assumes that traffic flows that belong to the same service class are allocated the same bandwidth.

The effective bandwidth of a flow, denoted by BW_{eq}, is calculated by using the well-known result such as Guerin's formula, which is summarized as follows [24]: If the average throughput of each flow is m, then the total throughput from aggregated s flows are $M = s \times m$, because effective bandwidth is additive if the sources are assumed to be homogeneous. If we assume that the variance of the throughput of each flow is τ^2, then the variance of the throughput of aggregate is given by $s \times \tau^2$ from the

[7] At present, we do not have data that indicates the exponential property in customer's impatience, but it is usual that users get impatient with the waiting time in exponential manner.

independence property between flows. Let the standard deviation of the aggregate sources be δ, then, we have $\delta = \sqrt{s \times \tau^2}$.

When s is sufficiently large, we can exploit the central limit theorem and the probability distribution of the aggregated throughput is approximated to the Gaussian distribution. We can find Gaussian distribution in access networks such as the modem pools or switches where access lines from customers have limited capacities [25]. Gaussian approximation gives us accuracy and simplicity in estimating the aggregated traffic from a large pool of users [26]. On the other hand, Gaussian approximation has a possibility of having negative input, which is unrealistic. Therefore, we assume that the input is distributed only in the positive plane of a graph with a mean located at center.

From the assumption of Gaussian distribution, we can represent the total effective bandwidth from a group of independent homogeneous users demanding the same kind of service with a simple formula given as follows.

$$BW_{eq} = M + \alpha\delta. \tag{9}$$

In (9), α is used as a safeguard factor in dimensioning the equivalent bandwidth so that an SLA is not violated: That is, the greater α is, the safer the link is to the connection (which means less packet loss), and vice versa. Note that the introduction of Gaussian approximation provides the network with an exploitation of statistical variation of traffic rate, via which more accurate prediction of traffic rate can be given to the network as compared with either a deterministic estimation provided by network calculus based on a leaky bucket rate predictor or an average-value based estimation provided by moving average [15].

One may notice that Gaussian approximation is prone to unpredictable error due to the period of measurement or the unstable variation of the network load. However, our work assumes a stable network with a very large number of customers, so that the problem is out of the scope of this work, and one can find more information about this problem for the small time scale network dynamics at [15]. When the distribution of bandwidth usage follows Gaussian distribution with mean M and standard deviation δ, the area obtained by summing the probability density function from BW_{eq} to ∞ of right tail is equal to ψ, the packet loss probability due to buffer overflow. Then, we obtain the following formula for α [26].

$$\begin{aligned}\alpha &= \sqrt{2}erfc^{-1}(2\psi)\\ &\cong \sqrt{-2\ln(\psi) - \ln(2\pi) - \ln(-2\ln(\psi) - \ln(2\pi))},\end{aligned}$$

where

$$erfc(x) = \int_x^\infty \frac{2}{\sqrt{\pi}} e^{-y^2} dy.$$

Note that, even though Guerin's effective bandwidth approach assumes bufferless network model, we can use the above model as an approximate tool for the estimation of buffered node if we assume that packets are discarded when the traffic approaches to a certain level of throughput, which we had assumed to be BW_{eq} in the present discussion. If this argument is allowed, the Guerin's model is considered to be a very useful method to estimate an equivalent capacity of link for guaranteeing a certain level of packet loss in the packet network.

Therefore, we can argue that adoption of Guerin's effective bandwidth approach is efficient for taking into account the packet-level QoS measures in the current EPFL model compared to the pure flow-aware network model where no packet level performance can be taken into account in bandwidth dimensioning.

C. Link Utilization

Thus far, we have discussed only about the dimensioning aspect of the link for the L2 switch by focusing only on the guarantee of the flow-level and packet-level QoSs to the customers.

On the other hand, the efficiency of the network usage (in other words, network utilization) is one of hot issues for the network operator, because utilization of a network resource is directly connected to the stable operation of the network as well as the expected revenue of a network operator.

Bearing this fact in mind let us investigate the expected utilization of the network for the proposed EPFL model. The utilization, U, of the node which take into account the objective of flow-level delay performance is defined by (10).

$$U = \frac{(1 - \Omega(T))\rho}{s}. \tag{10}$$

The utilization level is usually used by the network operator as a criterion for the configuration and provisioning of additional bandwidth. It is usual in the world of IP network that a network operator prepares for the provisioning of additional link when the link utilization reaches 50% or so [16].

IV. PERFORMANCE ASSESSMENT

In order to investigate the validity and implication of the proposed model, let us carry out a series of numerical experiments. First, let us compute the number of VCs under the predefined QoS objectives such as the constraint described in (4). Next, let us compute the required bandwidth of an output port of a switch node using the result of the number of VCs. Finally, let us compute the utilization of the network corresponding to the above results.

Before carrying out numerical experiment, let us briefly discuss about a realistic situation of user connection into the network. Let us assume that a number of xDSL[8] users are connected to the network access router via an L2 switch with a port speed of 100 Mbps per input link. It is usual that an output port of an access router has a link capacity of 1 Gbps, so that about 10 output ports of L2 switch are multiplexed to a single input port of an access router and packets multiplexed from that input port is distributed to a specific output port depending on the routing protocol. The output port now transmits multiplexed packets via a specific port for the next node in the core network. Fig. 3 illustrates this situation.

Usually, the speed of user's last mile link is negotiated between the Internet service provider (ISP) and the users in the

[8] x = A for ADSL and x = V for very high-speed digital subscriber line (VDSL).

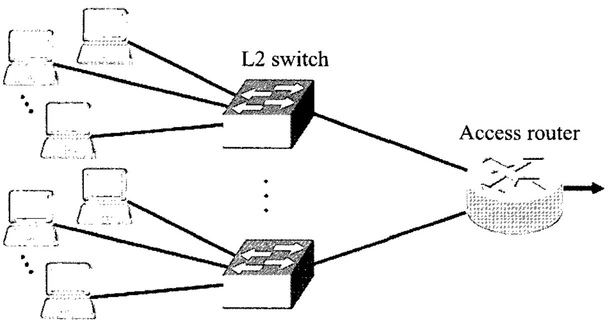


Fig. 3. Architecture of access network.

form of SLA even though the maximum input speed of a customer is 100 Mbps, and there exist a lot of different forms of SLAs in terms of bandwidth and delay. Currently, the typical bandwidth to an individual user provided by an ISP ranges from a few Mbps to tens of Mbps [16]. For simplicity, let us assume that the allowed maximum input rate of xDSL is 10 Mbps per subscriber, so that a maximum number of 10 active customers can be accommodated to a single input link of L2 switch. When 10 active customers simultaneously occupy 10 VCs for an hour, the total traffic will be 10 Erlangs.

Note, however, that a user does not always fully utilize the link capacity of 10 Mbps, and there remains a possibility for the over-booking of the input link of L2 switch for the customers who behave in a statistical manner. Therefore, let us assume that we do not mind the number of active users in the network. Instead, let us assume that the input process for the node is represented by a unit of Erlang, and we will investigate the maximum number of customers (synonymous with flows) that an input link of L2 switch can accommodate simultaneously. The procedure for the link dimensioning is composed of two steps, which is given as follows.

Step1: Flow-level VC dimensioning: For the given parameters from the flow statistics and QoS parameters such as the target delay tolerance, compute the number of VCs for the aggregated customers.

Step2: Packet-level bandwidth dimensioning: For the given target delay tolerance and number of VCs, compute the effective bandwidth for the aggregated customers.

The following subsections represent a detailed discussion on these steps.

A. Dimensioning the VCs

In the flow-level dimensioning, we compute s, the number of VC that has been defined in Section III-A, under the given flow-level QoS constraint of the probability for the delay tolerance.

In order to compute the required number of VCs as a function of traffic load, the following traffic and QoS parameters are assumed. Let us assume the traffic parameter first. The average duration of a flow is assumed to be 180 seconds. The offered load is assumed to vary between 2 and 20 Erlangs.[9] As to the QoS parameters a few typical values are assumed, and they are summarized in Table 1.

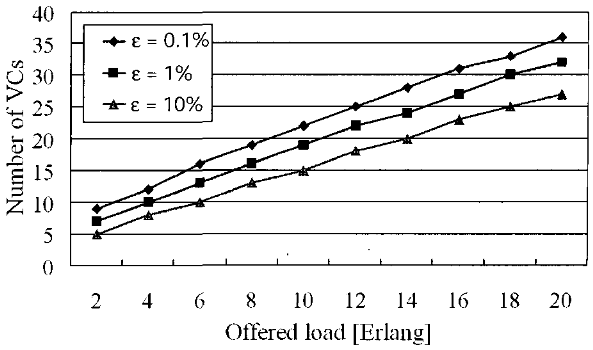


Fig. 4. Number of VCs for patient customers when $T = 1$ second.

Table 1. QoS parameters

QoS parameters	Assumed values
Delay T	1, 10, or 30 seconds
Tolerance ε	0.1, 1, 5, or 10%

In the following experiment, we will investigate the required number of VCs for two cases of patient and impatient customers. We assume the same parameters for the two cases.

A.1 Dimensioning the VCs

Fig. 4 illustrates the number of VCs for the L2 switch which accommodates patient customers as a function of the offered load for different delay tolerance when $T = 1$ second. The x-axis represents the offered load, whereas the y-axis represents the required number of VCs. As one may expect, the number of VCs in the system increases basically as the offered load increases. Note also that it depends on the probability of delay tolerance: The severer the delay tolerance, the greater the number of required VCs.

In order to investigate the effect of T to the number of required VCs, we compute the number of VCs for different values of T. Fig. 5 illustrates the number of VCs for the L2 switch which accommodates patient customers as a function of the offered load for a target delay of $T = 10$ seconds. We assumed three different delay tolerances of 1, 5, and 10% in this case.

Fig. 6 illustrates the number of VCs for the L2 switch which accommodates patient customers as a function of the offered load for $T = 30$ seconds. We assumed the delay tolerance of 1, 5, and 10% in this case.

As we can find from Figs. 4–6, the required number of VCs is sensitive to the values of tolerable delay T and its tolerance ε.

A.2 VCs for Impatient Customers

Fig. 7 illustrates the number of VCs for the L2 switch which accommodates impatient customers as a function of the offered load for different values of delay tolerance when $T = 1$ second.

[9] Note that 20 Erlangs of traffic intensity can be accomplished by 4,000 subscribed customers from a large apartment complex located in the metropolitan city area if we assume that the busy hour connection attempt per customer is 0.1 and the mean connection duration is 180 seconds.

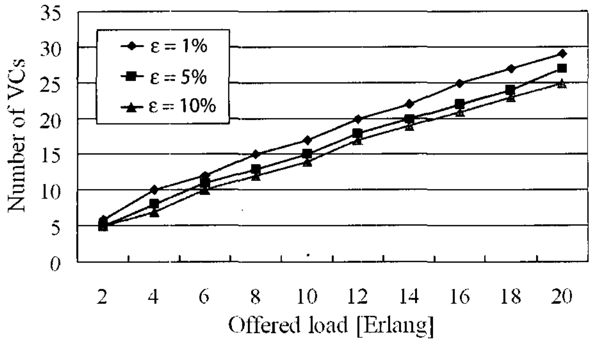


Fig. 5. Number of VCs for patient customers when $T = 10$ seconds.

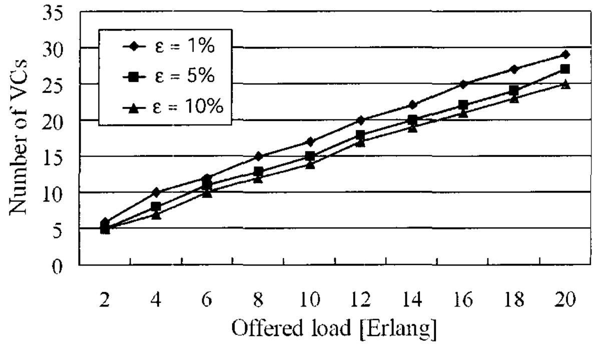


Fig. 6. Number of VCs for patient customers when $T = 30$ seconds.

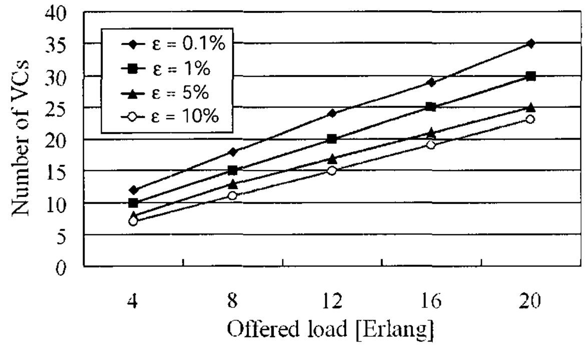


Fig. 7. Number of VCs for impatient customers when $T = 1$ second.

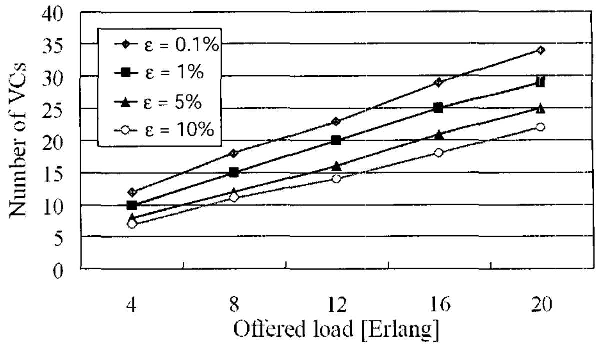


Fig. 8. Number of VCs for impatient customers when $T = 10$ seconds.

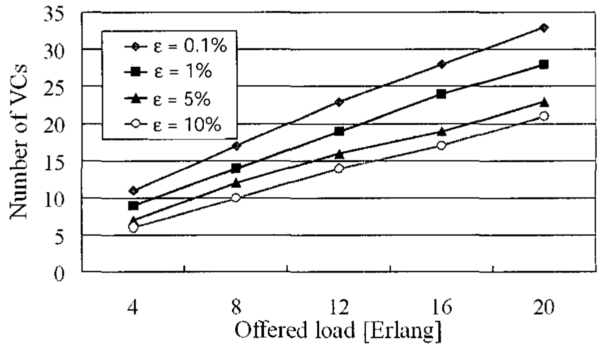


Fig. 9. Number of VCs for impatient customers when $T = 30$ seconds.

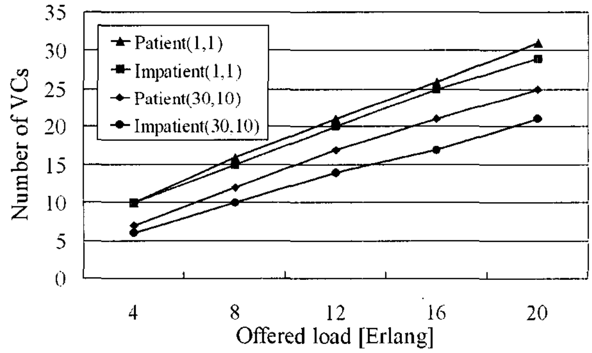


Fig. 10. Number of VCs for two types of customers.

We assumed the delay tolerance of 0.1, 1, 5, and 10% in this case and in the sequel.

Figs. 8 and 9 illustrate the number of VCs for the L2 switch which accommodates impatient customers as a function of the offered load for different delay tolerance when T = 10 and 30 seconds, respectively.

A.3 Comparison of Performance for Two Types of Customers

In order to compare the performance of the EPFL model for different types of customers, the patient and impatient customers, let us compare the required number of VCs for a switch that accommodates patient customers and impatient customers under the same traffic and QoS parameters. The parameters used in this experiment are two extreme cases, which is composed of $(T, \varepsilon) = (1$ second, $1\%)$ and $(T, \varepsilon) = (30$ seconds, $10\%)$.

We presented four graphs for the four cases of *user patience* (*Delay threshold, Tolerance*) in Fig. 10. As we can find from Fig. 10, the number of VCs for two types of customers are almost the same when $(T, \varepsilon) = (1$ second, $1\%)$: However, one can note that the difference between the number of VCs for two types of customers is evident when $(T, \varepsilon) = (30$ seconds, $10\%)$, which becomes more evident as the offered load increases.

From this result, we can argue that the EPFL model assuming the patient customers is safer as a method to estimate the number of VCs in a switch, because it is more conservative than that assuming the impatient customers.

Table 2. Total bandwidth for three cases of delay QoS requirements.

	Bandwidth capacity		
Offered load [Erlang]	Case A [Mbps]	Case B [Mbps]	Case C [Mbps]
4	72	61	44
8	110	94	72
12	142	121	99
16	174	148	121
20	201	174	142

Therefore, we argue that the EPFL model assuming the patient customers is favorable in link dimensioning for the BcN access switch where there may exist impatient customers as well as the patient customers. In the experiment described in the following subsections, we will focus on the experiment of an EPFL model assuming the patient customers.

B. Dimensioning the Link Capacity

Now let us compute the equivalent bandwidth from the results obtained in the previous subsection. Let us assume that the average throughput of each flow is $m = 5$ Mbps, then the total average throughput from aggregated s flows is $M = s \times m = 5s$ Mbps, because the attributes of the traffic sources are assumed to be homogeneous. Let us assume that the variance for the throughput of each flow is $\tau^2 = 0.5$. Then, the variance of the throughput of aggregate is given by $0.5s$. The total bandwidth requirement from s independent homogeneous users demanding the same kind of service is then given by

$$BW_{eq} = M + \alpha\delta = 5s + \sqrt{0.5s}\alpha. \tag{11}$$

Note that α in (11) can be determined by a given target value for the packet loss probability, which is assumed to be no greater than 10^{-6}. We obtained that α is equal to 4.75.

We assumed three cases for tolerable delay T and its tolerance ε, which is summarized as follows

- Case A: $T = 1$ second and $\varepsilon = 0.1\% \rightarrow$ *Tight QoS*
- Case B: $T = 10$ seconds and $\varepsilon = 1\% \rightarrow$ *Moderate QoS*
- Case C: $T = 30$ seconds and $\varepsilon = 10\% \rightarrow$ *Loose QoS*

Table 2 illustrates the total bandwidth of a link required to accommodate the patient customers with three cases of delay QoS requirements.

C. Performance of a Switch Node

Up until now we have focused on the dimensioning of link capacity that is needed to guarantee required QoS parameters in the combined flow and packet level. It is evident that guaranteeing proper level of QoS to users is an important spectrum of the network design.

On the other hand, the performance of the network such as the average utilization of the link is also a hot issue to the network operators. Therefore, let us observe the average utilization of a link for a set of traffic and QoS parameters that have been used in the above experiments. Fig. 11 illustrates the average utilization of an output port of an L2 switch as a function of the offered load for the three cases that have been defined in Table 2.

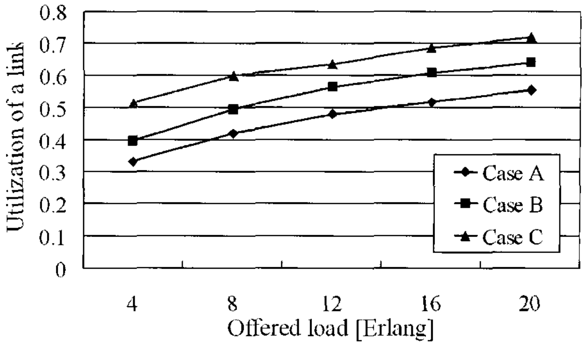


Fig. 11. Average utilization of a switch node.

As one can find from Fig. 11, the average utilization of a link under the defined delay tolerance increases in a concave manner as the offered load increases. It is also found that the severer the delay tolerance, the lower the utilization. This is in accordance with our expectation that the network utilization has to be kept to a certain level so that a certain level of QoS is guaranteed to the customers.

V. IMPLEMENTATION ASPECTS

We can think areas of implementation aspects of the proposed FPFL model at the real operating network in a number of ways. It is known that maintaining the list of flows in the commercial equipment with link capacity up to OC-48 has been realized by Caspian networks [27]. Therefore, the proposed model can be applied to an L2 switch with link speed of 100 Mbps or routers with link speed of a few Gbps located at any node of BcN.

The proposed EPFL model can be used as a resource allocation scheme for either streaming applications or elastic applications. As to the resource allocation scheme for an aggregate of streaming applications, we can associate the peak rate bandwidth allocation to a group of users as provisioning appropriate number of VCs to that user group. As to the resource allocation scheme for elastic applications that share a shared subscriber link, we can associate the number of VCs to the number of flows that share a link in a fair manner,[10] which means that all the elastic flows use the shared bandwidth fairly.

Concerning application of the proposed scheme in the provisioning of network services, the concept of link dimensioning based on EPFL model proposed in this work can be applied to the provisioning of a thicker pipe between nodes in the network. First, we can think of VoIP trunks, where VoIP traffic from multiple flows is aggregated into an LSP along the MPLS tunnel inside a BcN backbone network.

[10] For a router that adopts a fair sharing principle, a link is shared in a fair manner among the competing flows. For more detail, see [3], where a fair share of bandwidth is realized over Intel IXP1200 network processor on the per-flow basis.

VI. CONCLUSIONS

In this work, we have presented a theoretical framework for estimating the link bandwidth for BcN network, especially at the subscriber node. The basic concept in the provisioning of bandwidth at the node is flow-awareness, and thus the notion of logical circuits called VCs is introduced.

Using the flow-level delay measure and packet-level loss measure, we developed a method to dimension the required bandwidth for BcN access network that accommodate customers with patience or without patience. By taking into account the flow-level delay tolerance as well as the packet level packet loss probability, we could obtain a more realistic method to provision the link capacity for the subscriber switch in the network entrance.

The proposed method provides the network operator with a method to an easy estimation of network bandwidth by considering the tangible QoS measure of flow-level delay and the intangible QoS measure of packet loss probability in the design of subscriber access network in a hybrid manner. Via extensive numerical experiments, we illustrated the implication of the proposed work, and we could provide a certain perspective about the design of link capacity for BcN access network.

Therefore, the results illustrated in this work can be effectively used in the estimation of an optimal number of connections in BcN. The result can be also applied to the allocation of bandwidth pipes between gateway routers of neighboring ISPs, because recently developed multi-service provisioning platform can support various kind of network interfaces such as T1, T3, Ethernet, Gigabit Ethernet, etc.

If a logical connection generated by two end hosts for a file transfer, VoIP or a traffic tunnel between edge-to-edge of a core network can keep up with the rate offered to it, the utilization of the link can be improved very much. Thus, the proposed method can be also useful in securing a high utilization if a fare share of the link can be realized by a commercial router.

Our future research area includes the investigation of the performance of current model for more realistic user traffic attributes from specific service models such as VoIP, VPN, etc. One more field for the future study is the investigation of the relationship between the performance of full packet-level link dimensioning model and our hybrid flow-and-packet-level link dimensioning model regarding the required bandwidth and utilization under the same conditions. We hope that we can make it possible if we introduce the concept of CAC to the full packet-level link dimensioning model, where the concept of flow-awareness proposed in this work is incorporated into the estimation of maximum number of flows a node can accommodate in determining the admittance and rejection of a new flow.

ACKNOWLEDGMENTS

The first author would like to thank Tae-Joon Lee for his help in the preliminary numerical experiments. The authors are grateful to the anonymous reviewers for their comments which helped to improve the quality of this paper.

REFERENCES

[1] P. Siripongwutikorn and S. Banerjee, "Per-flow delay performance in traffic aggregates," available at http://www.hpl.hp.com/research/papers/2002/.

[2] M. Hawa and D. W. Petr, "Quality of service scheduling in cable and broadband wireless access systems," in *Proc. IWQoS 2002*, June 2–4, Monterey, CA, USA.

[3] A. Sinha, K. Mitchell, and D. Medhi, "Flow-level upstream traffic behavior in broadband access networks: DSL versus broadband fixed wireless," *IEEE Workshop on IP Operations and Management*, 2003.

[4] A. Cerra and D. Kan, "User centric broadband services: Demand drivers and market opportunities," *Alcatel Telecommun. Rev.*-1st Quarter, 2005.

[5] N. G. Aneroussis and A. Lazar, "Virtual path control for ATM networks with call level quality of service guarantees," available at http://citeseer.ist.psu.edu/aneroussis96virtual.html.

[6] N. G. Aneroussis and A. Lazar, "An architecture for managing virtual circuit and virtual path services on ATM networks," available at http://citeseer.ist.psu.edu/cache/papers/cs/5857/.

[7] S. Oueslati and J. Roberts, "A new direction for quality of service: Flow-aware networking," in *Proc. NGI 2005*, Apr. 18–20, 2005, Rome, Italy.

[8] M. Veeraraghavan, X. Zheng, H. Lee, M. Gardner, and W. Feng, "CHEETAH: Circuit-switched high-speed end-to-end transport arcHitecture," in *Proc. Opticomm 2003*, Dallas, TX, Oct. 13–17, 2003.

[9] A. R. Ward and W. Whitt, "Predicting response times in processor-sharing queues," in *Proc. the Fields Institute Conference on Communication Networks*, 2000.

[10] J. Boyer, F. Guillemin, P. Robert, and B. Zwart, "Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks," in *Proc. IEEE INFOCOM 2003*, San Francisco, Mar. 30–Apr. 3, 2003.

[11] W.-C. Chan and Y.-B. Lin, "Waiting time distribution for the M/M/m queue," *IEE Proc. Commun.*, vol. 150, no. 3, June 2003.

[12] N. Benameur, S. B. Fredj, S. Oueslati-Boulahia, and J. W. Roberts, "Quality of service and flow level admission control in the Internet," *Computer Networks*, vol. 40, 57–71, 2002.

[13] J. W. Roberts and S. Oueslati-Boulahia, "Quality of service by flow aware networking," Article submitted to Royal Society, 2000.

[14] T. Bonald, S. Oueslati-Boulahia and J. Roberts, "IP traffic and QoS control: The need for a flow-aware architecture," *World Telecommunications Congress*, Sept. 2002, Paris, France.

[15] N. G. Duffield, P. Goyal, A. Greenberg, P. Mishra, K. K. Ramakrishnan, and J. E. van der Merwe, "Resource management with hoses: Point-to-cloud services for virtual private networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 5, Oct. 2002.

[16] H. Lee, *et al.*, "Traffic model and link dimensioning for BcN access networks," Technical report for KT, Aug. 2005.

[17] Y. Cheng, A. Tizghadam, M. S. Kim, M. Hashemi, A. Leon-Garcia, and James W.-K. Hong, "Virtual network approach to scalable IP service deployment and efficient resource management," *IEEE Commun. Mag.*, Oct. 2005.

[18] F. le Faucheur, J. Boyle, K. Kompella, W. Townsend, T. D. Nadeau, D. Skalecki, "Russian dolls bandwidth constraints model for Diff-Serv-aware MPLS traffic engineering," draft-ietf-tewg-diff-te-russian-05.txt, Jan. 2004.

[19] Y. Ying, R. Mazumdar, C. Rosenberg, and F. Guillemin, "The burstiness behavior of regulated flows in networks," *Proc. Networking*, 2005.

[20] L. Kleinrock, *Queueing Systems, Theory, Vol. 1*, Wiley, New York, 1975.

[21] L. Noirie, "Mixed TDM and packet technologies as a best compromise solution to ensure a cost-effective bandwidth use with the current traffic evolution," in *Proc. NGI 2005*, Apr. 18–20, 2005, Roma, Italy.

[22] F. Baccelli and G. Hebuterne, "On queues with impatient customers," Rapports de Recherche, no. 94, INRIA, Sept. 1981.

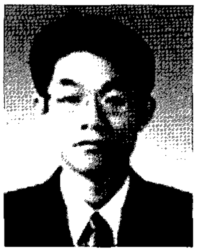
[23] O. Garnett, A. Mandelbaum, and M. Reiman, "Designing a call center with impatient customers," *Manufacturing & Service Operations Management*, vol. 4, no. 3, pp. 208–227, Summer 2002.

[24] R. Guerin and H. Ahmadi, "Equivalent capacity and its applications to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*. vol. 9, no. 7, pp. 968–981, Sept. 1991.

[25] J. Kilpi and I. Norros, "Testing the Gaussian character of access network traffic," in *Proc. ACM SIGCOM Internet Measurement Workshop 2002*, 2002.

[26] L. Noirie, R. Douville, M. Vigoureux, and T. B. de S. Martin, "Statistical multiplexing in data-aware transport networks," *Alcatel Telecommun. Rev.*-3rd Quarter 2005.

[27] Caspian Networks, http://www.caspiannetworks.com/.



Hoon Lee received B.E. and M.E. from Kyungpook National University, Daegu, Korea, and a Ph.D. from Tohoku University, Sendai, Japan. From February 1986 to February 2001, he worked at KT R&D Center as a senior research staff. Currently, he is an associate professor at Changwon National University, Korea. From March 2005 to August 2006, he stayed at UMKC (University of Missouri-Kansas City), USA as a visiting scholar. He has published about 70 papers in the field of QoS control and network design. His research interests include teletraffic engineering, network design and control, performance analysis, provision of QoS, and pricing for broadband convergence networks. Dr. Lee is a member of KICS, Korea.

Khosrow Sohraby is the Curators' professor of Computer Science and Electrical Engineering at the University of Missouri-Kansas City since 2004. He received B.E. and M.E. degrees from McGill University, Montreal, Canada in 1979 and 1981, respectively, and Ph.D. degree in 1985 from the University of Toronto, Canada, all in Electrical Engineering. His current research interests include design and analysis of high speed computer and communications networks, traffic management and analysis, modern queuing theory, large-scale computations in performance analysis, multimedia networks, and networking aspects of wireless and mobile communications. In 1986, he joined AT&T Bell Laboratories, Holmdel, NJ as a member of technical staff in the Teletraffic Theory and System Performance Analysis Department. In 1989, he joined IBM T. J. Watson Research Center, Yorktown Heights, NY. as a research staff member in the Communications Systems Department. He joined the Computer Science Telecommunications Program at the University of Missouri-Kansas City in 1994 as professor. Since June of 2005, he has been serving as the dean of the School of Computing and Engineering. He is a member of IEEE Communications Society and has served as the guest editor of number of special issues of the Journal of Selected Areas in Communications and Communications Magazine. He has served as a member of editorial board of Computer Networks, Wireless Networks Journal, International Journal of Wireless Networks, Network Magazine, and Mobile Computing and Communications Review.