

## **POLICY-BASED BILLING ARCHITECTURE FOR INTERNET DIFFERENTIATED SERVICES**

Felix Hartanto and Georg Carle

*GMD FOKUS, Kaiserin-Augusta-Allee 31, D-10589 Berlin, Germany  
{hartanto, carle} @fokus.gmd.de*

**Abstract** The Differentiated Services architecture allows a service provider to configure new services dynamically using a policy protocol. This benefit, however, may not be fully realized if the service provider can not charge for the services, or need a high effort to update its billing system to charge for the services. Thus, there is a real need for a flexible billing architecture to complement the flexibility offered by the differentiated service model. To meet this need, a policy-based billing architecture is proposed in this paper. This architecture allows a service provider to define policies for configuring various processes of a billing system based on the charging and pricing schemes used for individual services. It is demonstrated how the architecture supports flat-rate, duration-based and volume-based charging, and also both static and dynamic pricing. Definitions of policies for various charging and pricing schemes are discussed and the potential complexity of each of them is analyzed. Based on the complexity analysis we recommend the least complex charging schemes for four classes of differentiated services, which have been proposed for meeting different application requirements.

**Keywords:** Policy-based Architecture, Charging, Pricing, Differentiated Services.

### **1. INTRODUCTION**

The current Internet supports a single level of best-effort service, where all data packets have equal access to the network bandwidth. In case of congestion, every packet has the same probability of being delayed or discarded. Any lost packets can be recovered by using additional higher layer protocols (e.g. TCP), which incorporate an acknowledgement procedure. However, this mechanism degrades the achievable throughput and incurs additional delay, making it less suitable for emerging real-time applications, which have strict delay requirements.

In an attempt to enrich this service model, the Internet Engineering Task Force (IETF) is considering a number of architectural extensions that permit the allocation of different service levels to different users. One of the outcomes of this effort is the Integrated Services (IntServ) architecture that integrates guaranteed and predictive service quality with the best-effort service of the Internet. This service model provides service discrimination through explicit allocation and scheduling of resources in the network using RSVP (Resource Reservation Setup Protocol) [8]. RSVP keeps state for each reservation in all intervening routers. The potential number of reservations is very large and may exceed the number of all pairs of communicating computers, because reservations can be specified for individual flows between any application on any computer. Thus, the feasibility of supporting the potentially huge number of reservations aggregated near the core of the Internet is questioned [24]. Because of the scalability problems of the IntServ architecture, the Differentiated Services (DiffServ) architecture [3][28] has been proposed to provide a means of offering a spectrum of services without having to maintain per-flow state in every router.

The DiffServ architecture is based on an interconnecting of administrative domains. A bilateral service level agreement (SLA) is negotiated between neighboring domains to provide differential service contracts for different traffic aggregates. By strictly enforcing the aggregate traffic contracts between domains, the

DiffServ architecture provides a well-defined end-to-end service over chains of separately administered domains.

Within each domain most of the resource management complexity is pushed to the edge of the domain. On domain ingress, incoming traffic is classified by the per-hop behavior (PHB) bits into aggregates. The aggregated traffic is forwarded and policed within the domain according to the aggregate profiles in place. A given PHB is configured using the DSCP (Differential Services Code Point) field, which is the first six bits of the DS byte in the header of IP packets [28].

The definitions of the PHB within a domain define the different services that can be provided by the DiffServ architecture. Policy protocols, such as COPS (Common Open Policy Service) [5], have been suggested to provide dynamic and automatic configuration of various network elements in implementing the PHB. This offers high flexibility for a domain administrator or service provider to define a wide variety of services to meet market needs. This benefit, however, may not be fully realized if the service provider can not charge for the new services, or if an update of their billing system to charge for the services requires a high effort. Thus, a flexible billing architecture is needed to complement the flexibility offered by the differentiated service model. To meet this need, a policy-based billing architecture is proposed in this paper. Prior to describing the architecture, a structure review of existing charging and pricing schemes is presented in the next section.

## 2. CHARGING MODELS

### 2.1 Charging Structure

As charging for other telecommunication services, e.g. telephony service, the Internet charging is also structured into *subscription charge* and *session charge*. Each of these in turns has a setup component and a recurring or usage part.

The subscription-setup charge is sometimes termed the "joining fee", for setting up the user account and provision of software or hardware required for connection to the Internet. The subscription-recurring charge is often termed "access or rental". It is often a simple flat charge.

The session-setup charge is often termed "session-access" charge. It is often a simple flat charge, such as for setting up a multicast session. The session-usage charge usually varies according to the amount of resources reserved or consumed. The session duration differs between connection-oriented and connectionless service. For connection-oriented services session duration can be defined as the time resource is reserved till the time it is freed. On the other hand, for connectionless services session duration can be defined as the time the first packet of a flow is observed till a time out period, or based on accounting session, which can be from the time last user bill is produced till the next one.

The time scale for the subscription charge is normally longer than or equal to the time scale of the session charge. Thus, the subscription charge can be used to reduce the need for a session charge. In view of this, we can differentiate the charging into two categories, i.e.

1. *Flat rate charging*, where the session charge is zero and the session-usage charge is absorbed by the subscription-recurring charge. This charge allows the user to receive unlimited network access, regardless of the amount of time connected to the network and the amount of traffic sent or received.
2. *Usage sensitive charging*, where the session-usage charge is non-zero and varies according to the level of resources used or reserved by the customers.

The usage sensitive charge can be defined by a formula which expresses the usage charge  $UC$  as a function of setup charge  $SC$ , pricing  $p$  and usage  $U$  parameters, i.e.

$$UC = SC + \sum_i p_i U_i \quad (1)$$

The usage parameters quantify the number of units of usage. The possible parameters used are duration ( $D$ ) and volume ( $V$ ). Based on the usage parameters, we can differentiate two categories of usage-sensitive charging, namely *duration-based* and *volume-based charging*. Duration-based charging is commonly used for charging reserved resources, while the volume-based charging is commonly used for charging consumed resources. A combination of both charging categories has also been suggested, for example, to charge out-profile traffic differently from in-profile traffic [10,[33], or to charge consumed resources on top of the reserved resources [15]. The combined or two-tier charging can be expressed as:

$$UC = SC + p_1 D_1 + p_2 V_2 \quad (2)$$

The benefits of such a combined pricing is to allow the user to lower the 'per unit time' cost at the cost of raising the 'per unit volume' cost.

The pricing parameters, sometimes referred to as tariff parameters, define the price per unit of reserved or consumed resources. The reserved resources can be expressed in terms of bandwidth and buffer ( $B$ ), or token bucket filter ( $F$ ) parameters (e.g. leaky rate and bucket size). The reserved bandwidth can be specified explicitly by the users, measured [11], or derived from the source parameters (e.g. mean rate, peak rate, and burstiness) and the required quality of service usage, for example, using the equivalent bandwidth concept [17],[22]. On the other hand, the consumed resources can be given in terms of packets, octets, or bits. The volume can be measured directly using a traffic meter, e.g. IETF real-time flow measurement [6] or derived from bandwidth utilization over a given time period [14].

In this paper, we assume the bandwidth based pricing for reserved resources and packet based pricing for consumed resources. The packet can be of different priority or precedence level ( $P$ ), where different levels of resources are allocated implicitly to meet specific quality of service.

In general we can write the pricing parameters as a function of allocated resources ( $R$ ), i.e.  $p = f(R)$ , where  $R = B$  or  $P$ . For example, the function  $f(B)$  can simply be a linear function, e.g.  $f(B) = (B/B_u) p_{Bu}$ , where  $B_u$  is unit of bandwidth and  $p_{Bu}$  is the price per unit of bandwidth.

The price per unit bandwidth  $p_{Bu}$  can be static or dynamically varied depending on the current demand on the network resources, which gives rise to the two pricing categories, namely *static pricing* and *dynamic pricing*.

### 2.1.1 Static Pricing

The price in this category is set in the contract between the user and the service provider. The time scale of this price change is much longer than the session duration. Prices normally do not change simply because of instantaneous congestion within the network, but rather due to long term observation of network usage and market conditions. For example, the service provider will give advance notice to lower prices to stay competitive, or raise prices to meet increasing cost.

The static pricing parameters can be modified by other session characteristics. Three price modifiers, which are commonly used in the telephony pricing, are *time-of-day* ( $T$ ), *destination* or *end-points* ( $E$ ), and *usage* ( $U$ ).

With the *time-of-day* modifier, the price depends on the calendar time, such as the time of day (peak or off-peak), day of the week (weekday or weekend), or public holiday. It is a form of congestion pricing based on long-term observations. This factor has been suggested in [31] for Internet pricing.

With the *destination* modifier, the price depends on the distance between the sender and the receiver. It can be based on the number of hops traversed or simply the location of the receiver [13]. Unlike in POTS where the distance to the destination can be derived directly from the telephone number, in the Internet no such standard numbering is available. While hop count and domain name may reveal information on distance and location, no ubiquitous method is available for accurately determining distance or location. With the diminishing distance pricing in the telephony industry, it is also expected that the distance pricing in the Internet will play a less significant part. Thus, this argues for a distance independence pricing or other simple alternatives, such as a zoning approach. With this approach user traffic is categorized as in-zone if it is addressed to a receiver within the same service provider domain and out-zone otherwise. A surcharge is applied to the out-zone traffic which accounts for the interconnecting charge incurred by the traffic.

With the *usage* modifier, the price depends on the amount of usage, i.e. the duration of a session or the level of traffic volume. This is to encourage long session in order to minimize the overhead of session setup process or to encourage high volume customers [9].

Taking into account these three factors as surcharges or discounts to the base charges  $f(R)$ , we can write the pricing formula as:

$$p = f(R, T, E, U) = f(R) f(T) f(E) f(U), \text{ where } U = D \text{ or } V \quad (3)$$

### 2.1.2 Dynamic Pricing

The price in this category varies depending on the demand on the network resources or congestion level within the network. The price changes instantaneously or on the spot (thus, the name spot-pricing). The intention is that the price should be zero when the network is uncongested, but when there is congestion the price should reflect the incremental social cost determined by the marginal delay cost to other users and the willingness of the user to pay for the cost. Price adjusting can be performed by auction pricing and by feedback pricing.

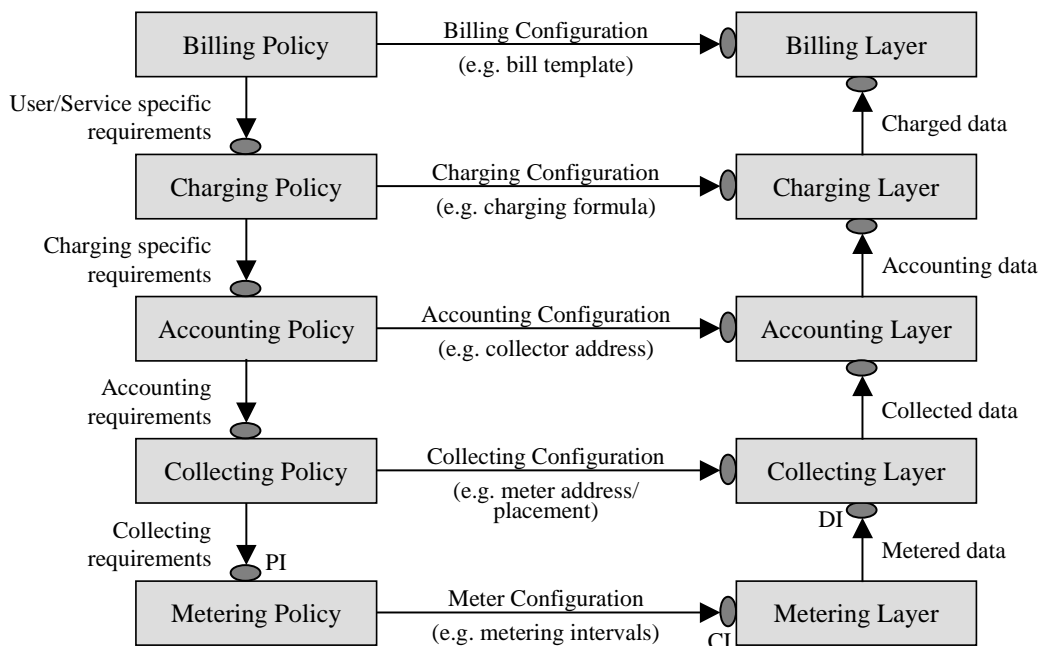
With *auction pricing* or a "smart market" approach [25], prices are determined based on consumers bids. Users include a bid in each packet, or based on a larger granularity [26]. At congested routers, packets are prioritized based on these bids. In case of congestion, packets of users offering the lowest bid are discarded first, and accepted packets are priced at a rate determined by the highest bid among the rejected packets. The cost of carrying each packet is thus related to the marginal value (represented by the bid) of the traffic which has been pushed out. At the equilibrium price the user's willingness to pay for additional data packets equals the marginal increase in delay cost generated by those packets [25]. Auctioning of bandwidth, rather than auctioning based on individual packets, has been considered in [16][23]. In this approach, bandwidth is split into small units, and users bid for the required bandwidth at each auction period.

With *feedback pricing*, the prices are calculated by the provider dynamically according to the load on the network. For example, in [27], the price is calculated based on the instantaneous filling level of the buffers at network nodes. Price feedback can be initiated by a customer query, e.g. by sending a request to convey the source demand followed by the network feeding back the price [32], or by a load threshold within the network [16]. The basic unit for pricing can be sent packet, or units of bandwidth reserved over a fixed period of time [27]. The users decide whether or not to send packets or to reserve bandwidth based on the prices given.

## 3. POLICY-BASED BILLING ARCHITECTURE

### 3.1 Billing System Framework

From capturing the usage to creating a bill to be sent to a customer, a billing system goes through processes, which can be modeled by a layering framework as shown in Figure 1. Each layer represents a specific basic functionality of a generic billing system.



PI = policy interface; CI = configuration interface; DI = data interface

Figure 1. Billing System Framework

The layer functionality is configurable using parameters supplied by specific policy definitions. The policy hierarchy indicates that the higher-level policy requirements trigger the selection and execution of the lower level policy. For example, the charging specific requirements, such as the need for measuring duration of the resource reservation in duration-based charging triggers the selection and execution of appropriate accounting policy. The policy, which may contain duration definition, i.e. between what events, in turns triggers the associated collection policy to collect/observe the events.

The *metering layer* tracks and records usage of resources by observing the traffic flows. The metering policy, used for configuring the metering layer, specifies the attributes of the traffic flows to be observed. In a connectionless network, such as Internet, where it is difficult to locate the end-point of a flow, the metering policy can also be used to define the flow duration.

The *collecting layer* accesses data provided by metering entities as well as collecting charged related events and forward them for further processing to accounting layer. This layer can collect information from multiple meters, as for multicast and distribute to home domains, as for user roaming. For this reason, the efforts in standardizing data exchange format and protocol at this layer will be beneficial. The meters from where to collect the data, the type of data and the frequency in collecting them are defined by the accounting policy.

The *accounting layer* consolidates the collected information from the collecting layer either within the same provider domain or from other provider domains and creates *network accounting data sets or records* which are passed to the charging layer for the assignment of prices. For supporting multicast charging, the multicast topology including splitting points can be reconstructed by entities of this layer (see [9] for further information on multicast charging).

The *charging layer* derives session charges for the accounting records based on service specific charging and pricing schemes, which are specified by the charging policy. This layer basically translates technical values (i.e. measured resource reservation and consumption) into monetary units using a charging formula as per equation (1).

The *billing layer* collects the charging information for a customer over a time period, e.g. one month, and include subscription charges and possible discounts into a bill. Billing policy can be used to specify the bill details.

Not all components of the framework will appear in every billing system. For example, a service provider who only provides a single service and charges the customers flat-rate will only implement the functionality of the billing layer. On the other hand, a service provider offering multiple services may implement the policy-based architecture to allow different charging schemes to be used for different services or customers without having to hard-code the charging formula into the billing system.

## 3.2 Implementation Architecture

*Figure 2* shows a possible architecture for implementing the framework introduced in the previous section. The architecture is being implemented in the ACTS project SUSIE. The network accounting part of the architecture is implemented using IP technology. The charging and billing part of the architecture supports TINA reference points [12] and is implemented using CORBA technology [30]. A policy gateway and an accounting data gateway provide interfaces between these two parts.

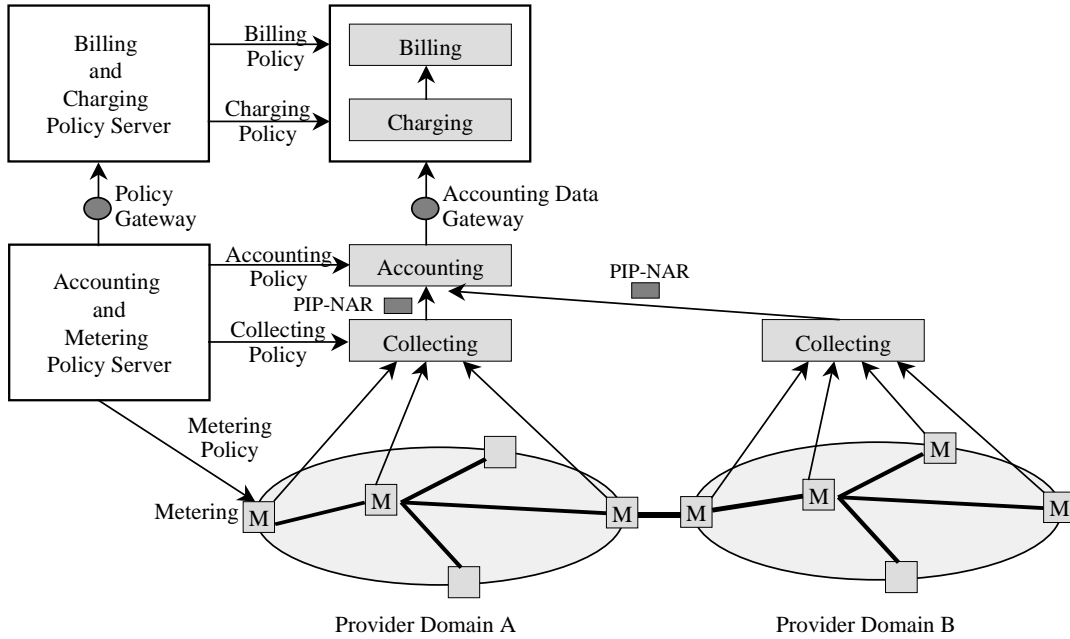


Figure 2. Billing Architecture

The architecture shows that in order to support multicast charging, metering may take place in the edge routers as well as at the multicast routers. In order to charge, for example the sender, the collected information needs to be fed back to the sender domain (e.g. provider domain A), so that the provider can reconstruct the multicast tree structure in calculating the overall multicast costs.

The architecture uses metering conformant to the IETF Real Time Flow Measurement Architecture (RTFM) [6]. For configuring a meter and for collecting accounting data, SNMP and a special meter MIB [7] are used.

The collecting entities at each domain collect the data from the meters and fill the PIP-NAR (Premium IP Network Accounting Record) data structure, which is used for transporting usage information within one provider domain and also for exchanging usage information in a multi-provider scenario. The data structure contains flow description for unique identification of a flow and reserved and used resources for the flow. It also contains a measurement point identifier and a record description for supporting different styles of PIP-NARs (uni-directional/bi-directional, DiffServ/IntServ style, and others). A detailed specification of the PIP-NAR can be found in [29].

### 3.3 Policy Setup

To illustrate the policy definitions and their relationship to the user subscription and session profile, we consider a subscriber with a default service and being allowed to use two other services. For each service, there are associated charging, accounting, collecting and metering policies which are set up based on the subscriber's service contract or service level agreement (SLA). We assume that a combined charging scheme is applied to one of the service, where duration-based charging is applied to the reserved resources and volume-based charging is applied to the consumed resources or out-profile traffic. A sample of policy setup for this customer is shown in Figure 2.

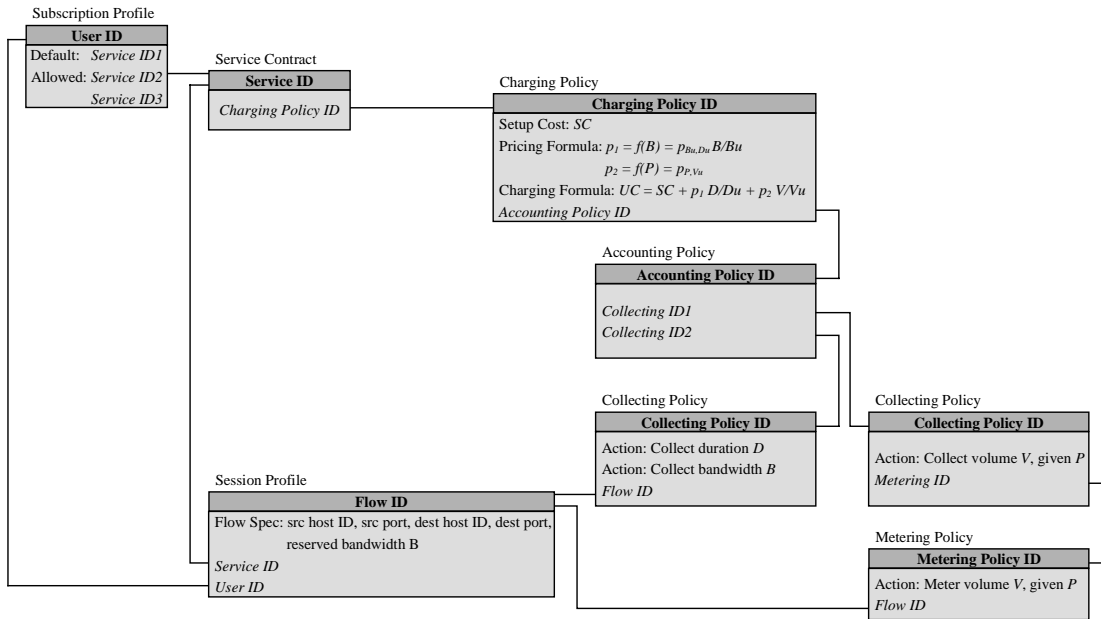


Figure 3. Relationship between Profiles and Policies

In the figure the identifications appearing in italic indicate the link between the records. For example, *Accounting Policy ID* links a charging policy to its corresponding accounting policy. In the charging policy a combined charging formula is used for Service ID2. Based on the charging formula, an accounting policy is defined to collect the required usage (duration and volume for packets with priority P) and pricing (reserved bandwidth) parameters for a specific Flow ID. The figure demonstrates the use of separate collecting policy elements to meet the accounting requirements. One of the collecting policy specifies the need for metering the usage and thus, a metering policy is defined to meter the required traffic volume.

The policies are used by authentication server [1] in conjunction with admission control or bandwidth broker. The authentication process checks the subscriber profile based on User ID and Service ID from the user service request and decide whether the request is admissible or not. Once a request is admitted the process will create a session profile and retrieve the accounting, collecting and metering policy from policy database and distribute them to the accounting, collecting and metering layer, respectively. The metering policy, for example, can be translated into rulesets used in configuring NeTraMet, a traffic meter defined by the IETF RTFM working group [6].

### 3.3.1 Applicability for Various Charging Schemes

In discussing the policy setup in the previous section a combined charging has been used. This charging scheme represents a general scheme as the three charging schemes, flat-rate, duration-based and volume-based, can be treated as special cases of this scheme. For the flat-rate charging no charging policy, and hence accounting policy and metering policy, need to be defined. For the duration-based charging no metering policy needs to be defined. Assuming that the bandwidth is static throughout the session, then the accounting policy just needs to obtain the bandwidth from the session profile and the duration from the flow start time and flow stop time. This means that the scheme can be considered simple. For the volume-based, all three types of policies are needed and the metering can be quite complex and resource consuming [4]. This means that the volume-based charging is the most complex among the three charging schemes. Moreover, as the accounting and metering policy definitions are based on the charging policy, which in turns based on the charging formula used, we can conclude that the complexity of a billing system depend on the charging formula.

### 3.3.2 Applicability for Various Pricing Schemes

The pricing strategies used in the above example is based on static resource pricing, where only bandwidth has been used for pricing the reserved resources and priority packet pricing has been used for pricing the consumed resources.

The use of price modifiers, such as time-of-day, destination and usage, in conjunction with the resource pricing may require additional policy definitions. For example, let us consider a flow, which starts at 17:55 and lasts until 18:30 and the peak rate to off-peak rate transition is at 18:00. The simplest alternative for the service provider is to charge the peak rate for the entire flow duration at 1 minute granularity, i.e.  $UC = 35 f(B, peak)$ . In this case no additional policy is needed. However, in order to be competitive and fairer to the users, the service provider may use an alternative charging scheme where the peak rate is charged up to 18:00 and off-peak rate is charged for the rest of the flow duration, i.e.  $UC = \sum_i U_i f_i(B, T) = 5 f(B, peak) + 30 f(B, off-peak)$ . In this case additional accounting policy is needed to specify the creation of an accounting record at 18:00. This increases the complexity of the charging scheme.

Additional accounting policy will also be needed for dynamic pricing, which can be viewed as an extension of the time of day resource pricing where the price changes on the spot depending on the network condition. Unlike time-of-day, it is imperative for the dynamic pricing to generate accounting records when the price changes. These records form a dynamic contract between the customer and the service provider and thus, can be used as evidence in any disputes with the customers. The resulting number of accounting records can be large depending on the price update period. For example, let us consider a flow, which lasts for three minutes and generate an average of 250 packets. Using a bandwidth auction pricing with an auction period of 30 seconds (same as default refresh RSVP's soft-state period) six accounting records needs to be created. On the other hand, if the auction is based on the packet, then up to 250 accounting records need to be generated if all packets successfully gained access to the network. This number is obviously much larger than one or two accounting records created for static resource pricing. In addition to the potential large overhead, the dynamic pricing has been considered complex due to the need for modifying the bandwidth reservation protocol or packet transfer protocol to include the price information [31].

Table 1 summarizes the policies required by different charging and pricing schemes along with the complexity of the schemes.

Table 1. Complexity of Charging and Pricing Schemes

| Charging and Pricing Schemes                              | Charging Policy  | Accounting Policy   | Metering Policy                      | Complexity/ Overhead  |
|---|--|---|--------------------------------------|---|
| Flat rate charging  | None   | None  | None                                 | Simple  |
| Duration charging, bandwidth pricing                      | $p = f(B) = p_{Bu} B/Bu$<br>$UC = SC + p D/Du$<br>( $Bu$ is bandwidth unit, $Du$ is duration unit)                         | Collect duration $D$ and bandwidth $B$  | None                                 | Simple<br>Require additional policies if price modifiers are used   |
| Duration charging, dynamic pricing (auction and feedback) | $p_i = p_{ci} Bi/Bu$<br>$UC = SC + \sum_i p_i D_i/Du$<br>( $Bu$ is unit of bid bandwidth)                                  | Collect duration $D_i$ , reserved bandwidth $B_i$ and price per unit bandwidth $p_{ci}$ within $i$ -th auction/ feedback period | None                                 | Complex<br>Increasing session duration or auction/ feedback period increases number of accounting records             |
| Volume charging, packet pricing                           | $p_{DS} = f(P) = p_{DS, Vu}$<br>$UC = SC + \sum_{DS} p_{DS} V_{DS}/Vu$<br>( $Vu$ is volume unit, $P$ indicated by DS-byte) | Collect number of packets $V_{DS}$ per priority $P$   | Count packets per priority $P$       | Medium<br>Large overhead in counting packets.   |
| Volume charging, auction pricing                          | $p_i = p_{ci}$<br>$UC = SC + \sum_i p_i$   | Collect current price per packet $p_{ci}$   | None                                 | Very complex<br>Increasing number of packets generated within session duration increases number of accounting records |
| Volume charging, feedback pricing                         | $p_i = p_{ci}$<br>$UC = SC + \sum_i p_i V_i$   | Collect volume $V_i$ and current price per packet $p_{ci}$ within $i$ -th feedback period                                       | Count packets within feedback period | Complex<br>Increasing session or feedback period increases number of accounting records                               |

#### 4. CONCLUSTIONS AND RECOMMENDATIONS FOR DIFFSERV CHARGING AND SERVICES

The proposed policy-based architecture has been demonstrated to support flat-rate, duration-based and volume-based charging, which are the three charging schemes identified through a structure review in Section 2. The architecture has also been shown to support static and dynamic pricing schemes. The analysis in the previous section has shown that some schemes are more complex than others, especially the dynamic pricing schemes. Based on this complexity analysis we recommend the least complex charging schemes for four classes of differentiated services (see Table 2), which have been proposed for meeting different application requirements.

Table 2. Proposed Service Classes

| Service              | Description  | Examples   | Recommended Charging Scheme                                   |
|----------------------|--|--|---|
| Premium service [21] | Peak bandwidth reservation, strict delay, jitter and loss guarantee, highest delay and loss priority   | Real-time applications that concern about jitter and cannot tolerate loss, e.g. CBR video transmission | Duration charging, bandwidth pricing                          |
| Assured service [19] | Expected bandwidth reservation, guarantee delay and jitter, but tolerate some loss<br>Marking of out-profile traffic to user default service | Real-time applications that concern about jitter but tolerate loss, e.g. VBR video transmission        | Subscription charging or Duration charging, bandwidth pricing |
| Priority service [2] | Relative delay priority with loss priority option as per default services  | Applications that require some delay or throughput guarantee, e.g. priority data transmission          | Volume charging, priority packet pricing                      |
| Default services     | Relative loss priority, best effort and above best effort  | Non-real time applications, e.g. normal data transfer  | Flat rate charging  |

By providing various *default services* with the subscription charge varied according to the loss priority level, users can select the most suitable default service based on their long-term requirements. The users can use an adaptive traffic control, such as packet marking engine [14], to monitor their traffic and select to transmit at higher priority than their chosen default service (i.e. using the *priority service*) if the observed service rate falls below the minimum target rate. With the network dropping lower priority packets first, this approach approximates the dynamic pricing where the high priority indicates the willingness of the users to offer higher bids in order to ensure the delivery of their packets. It avoids the complexity of the dynamic pricing scheme as the charge for the priority service is simply volume-based with static priority packet pricing.

*Assured service* assures bandwidth requirements to the user. This bandwidth assurance can be specified statically or dynamically as static or dynamic SLA, respectively. With static SLA no per flow bandwidth reservation is made and thus, duration based charging will not be appropriate, but instead subscription charging which takes the SLA into account can be used. Duration-based charging can be used for dynamic SLA where bandwidth reservation is made per flow. In both static and dynamic SLA at least one token bucket filter is placed along the path of the flow for bandwidth restriction. In regard to marking of out-profile traffic in this service, out-profile packets can be marked down to the default service of the user, which is not necessary best-effort, since users value their packets as low as their default service. The default marking can be conveyed to the routers using AAA policy protocol [1] during authentication process. Since the default service is based on the flat-rate charging, no usage metering is needed for the out-profile traffic. This approach avoids the debate over how to charge marked packets which are dropped within the network.

*Premium Service* is also called a Virtual Leased Line service. Two different ways for handling a VLL service are defined in [21], i.e. peak bandwidth reservation and statistical bandwidth reservation. In both cases duration-based charging based on the reserved bandwidth can be used.

The four service classes can be implemented using a multi-queue with protective buffer policies [19] in order to ensure that the traffic from one class does not affect the traffic from other classes during overload, while making full use of network resources during light load.

## 5. ACKNOWLEDGEMENT

This research is carried out within the framework of European Commission ACTS SUSIE project. The authors would like to thank Mikhail Smirnov and Tanja Zseby for their valuable discussions and anonymous reviewers for their useful comments.

## 6. REFERENCES

- [1] IETF Authentication, Authorization and Accounting (AAA) Working Group.  
<http://www.ietf.org/html.charters/aaa-charter.html>.
- [2] Y. Bernet, et al. A Framework for Differentiated Services. *IETF Internet Draft <draft-ietf-diffserv-framework-02.txt>*, Feb. 1999.
- [3] S. Blake, et al. An Architecture for Differentiated Services. *IETF Request for Comments, RFC2475*, December 1998.
- [4] M. S. Borella, V. Upadhyay and I. Sidhu. Pricing Framework for a Differential Services Internet. *European Transactions on Telecommunications*, Vol. 10(2), March/April 1999.
- [5] J. Boyle, et al. The COPS (Common Open Policy Service) Protocol. *IETF Internet draft <draft-ietf-rap-cops-06.txt>*, Feb. 1999.
- [6] N. Brownlee, C. Mills and G. Ruth. Traffic Flow Measurement: Architecture. *IETF Request for Comments, RFC2063*, Jan. 1997.
- [7] N. Brownlee. Traffic Flow Measurement: Meter MIB. IETF Internet Draft <draft-ietf-rtfm-meter-mib-09.txt>, June 1999.
- [8] R. Braden, et al. Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification. *IETF Request for Comments RFC2205*, September 1997.
- [9] G. Carle, F. Hartanto, M. Smirnov and T. Zseby. Charging and Accounting for QoS-Enhanced IP Multicast. *Proc. of Protocols for High-Speed Networks (PfHSN)*, Salem, MA, August 1999.
- [10] D. Clark. A Model for Cost Allocation and Pricing in the Internet. *Technical Report*, Laboratory for Computer Sciences, MIT, Cambridge, MA, August 1995.
- [11] C. Courcoubetis, F. Kelly and R. Weber. Measurement-based Charging in Communication Networks. *Statistical Laboratory Research Report 1997-19*, Univ. of Cambridge, 1997.
- [12] F. Dupuy, G. Nilsson and Y. Inoue. The TINA Consortium: Toward Networking Telecommunications Information Services. *IEEE Communication Magazine*, Vol. 33(11), November 1995, pp. 78-83.
- [13] H. Einsiedler, P. Hurley, B. Stiller and T. Braun. Charging Multicast Communications Based on A Tree Metric. *Proc. of Multicast Workshop*, Braunschweig, Germany, 1999.
- [14] W. Feng, D. Kandlur, D. Saha and K.G. Shin. Adaptive Packet Marking for Providing Differentiated Services in the Internet. *Proc. of ICNP*, Austin, TX, October 1998.
- [15] D. Ferrari and L. Delgrossi. Charging for QoS. *Keynote Paper at the IWQoS'98*, Napa, CA, May 1998.
- [16] G. Fankhauser, B. Stiller, C. Vögli and B. Plattner. Reservation based Charging in an Integrated Services Network. *Proc. of 4th INFORMS Telecommunications Conference*, Boca Raton, FL, March 1998.
- [17] R. Guerin, H. Ahmadi and M. Naghshineh. Equivalent Bandwidth and Its Application to Bandwidth Allocation in High-Speed Networks. *IEEE J. Select. Areas Commun.*, Vol. 9(7), September 1991, pp. 968-981.
- [18] G. Guthrie and M. Carter. User Charges for Internet: the New Zealand Experience. *Telecommunication Systems*, Vol. 6, September 1996.
- [19] F. Hartanto, H. Sirisena and K. Pawlikowski. Protective buffer policies at ATM switches. *Proc. IEEE ICC*, Seattle, WA, June 1995, pp. 960-4.
- [20] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski. Assured Forwarding PHB Group. *IETF Request for Comments, RFC 2597*, June 1999
- [21] V. Jacobson, K. Nichols and K. Poduri. An Expedited Forwarding PHB. *IETF Request for Comments, RFC2598*, June 1999.
- [22] F.P. Kelly. Tariffs and Effective Bandwidths in Multiservice Networks. *Proc. of ITC-14*, Antibes, France, June 1994, pp. 401-410.
- [23] A. Lazar and N. Semret. Auctions for Network Resource Sharing. *CTR Technical Report*, Columbia University, February 1997.
- [24] A. Mankin, et al. Resource ReSerVation Protocol (RSVP) Version 1 Applicability Statement Some Guidelines on Deployment. *IETF RFC2208*, September 1997.
- [25] J. MacKie-Mason and H. Varian. Pricing the Internet. In *Public Access to the Internet (Kahin B. and Keller J., Eds.)*, Prentice Hall 1995.
- [26] J. MacKie-Mason. A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network. *Technical Report*, University of Michigan, September 1997.
- [27] J. Murphy, L. Murphy and E. Posner. Distributed Pricing for Embedded ATM Networks. *Proc. of ITC-14*, Antibes, France, June 1994.
- [28] K. Nichols, S. Blake, F. Baker and D. Black. Definition of the Differentiated Services Field (DS Field) in the IP v4 and IP v6 Headers. *IETF RFC2474*, December 1998.
- [29] H. Orlamünder (Editor). Parameters and Mechanisms for Charging in IP based Networks [Network Aspects]. TR/NA-080301 V1.0.7 (1999-06), ETSI Working Group NA8 Technical Document, 1999.
- [30] OMG, The Common Object Request Broker: Architecture and Specification, Revision 2.1, August 1997.
- [31] S. Shenker, D. Clark, D. Estrin and S. Herzog. Pricing in Computer Networks: Reshaping the Research Agenda. *Communications Policy*, Vol. 20(3), 1996, pp. 183-201.
- [32] V.A. Siris, C. Courcoubetis and G.D. Stamoulis. Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks. *Proc. of IEEE GLOBECOM*, London, U.K., November 1996.
- [33] D. Songhurst and F. Kelly. Charging Schemes for Multiservice Networks. *Proc. of ITC-15 (V. Ramaswami and P.E. Wirth, editors)*, Elsevier, June 1997.