CHAPTER 4

A SURVEY OF CHARGING INTERNET SERVICES

BURKHARD STILLER

4.1 INTRODUCTION

Today's information society has a stringent need for advanced communication services and content, which are provided currently by a packet-switched networking technology instead of traditional circuit-switching. This is driven by the fact that provisioning of different applications on dedicated network infrastructures are inefficient [106], such as seen, for example, with the telephone network, cable TV networks, radio broadcast, and dedicated leased-line services for mission-critical data transfers. Deploying a single multiservice network infrastructure to create an integrated services network, offering services that range, e.g., from managed bandwidth services and VPNs to interactive voice, video, messaging, and eCommerce services, promises the potential for possible cost reductions, which in total is much larger than tuning networking technologies for different types of applications.

Considering the particular application support, the Internet as it exists today is designed for elastic data applications [107]. Experiences on a large scale basically are not available on reliably integrating multiple services in a single network in a commercial way. The commercial environment of tomorrow's Internet needs to revise the assumption that service customers rely on cooperation among them. Therefore, the TCP's fairness, provided by its congestion control algorithm, needs to be addressed, and, e.g., may be enforced by alternative protocol implementations in the network or may be provided by suitable economic incentives for the use of a service [61]. Because changes of transport protocols networkwide are economically difficult, an adequate distributed approach to

Managing IP Networks. Edited by Saleh Aidarous and Thomas Plevyak
ISBN 0-471-39299-5 © 2003 Institute of Electrical and Electronics Engineers

congestion control, service provisioning, and service usage is charging. In addition, due to the commercialization of the Internet, the provisioning of its services will form an open market requesting by definition an integrated charging system for services, transport, and content. For that reason, appropriate pricing schemes for communication services will provide incentives for reasonable resource usage, traditional network capacity, or bandwidth. Although solutions for methods of charging packet-switched, single-service class networks exist and are applied successful, packet-switched integrated- and differentiated-services networks have another dimension of complexity and require a completely different approach [118]. Consequently, charging for the future Internet, as the most prominent example of a packet-switched network, remains an unsolved problem even though a number of proposed approaches exist. Mainly, these problems are have three distinct causes: (1) technical reasons, (2) economic factors, and (3) operational constituents.

First, since a variety of service characterizations by QoS concepts exist [113], heterogeneous and advanced networking technology alternatives, such as ATM or Frame Relay, have been developed over time, the shape of the Internet of the future based on the IntServ [7] and the DiffServ [4] is still not fully defined. It remains unclear, if the RSVP [8] can be applied in a large backbone Internet without any loss of scalability, or if the current state of work on signaling between DiffServ domains will be able to cope with all QoS demands and traffic characterizations known from existing applications [59]. In addition, standardization work on DiffServ seems to be limited to a single provider domain [88], explicitly excluding any hint of definitions considered business-relevant, such as SLAs.

Second, an ISP today is faced with huge opportunities for growth [102] and it is challenged to increase profitability. In this highly competitive telecommunications service provider and ISP market, strategic differentiation and advanced pricing schemes for integrated multiservice networks are required in order to deal with efficient schemes, basic bandwidth allocation, or advanced QoS services to reap financial benefits of new services and to gain a competitive edge [5, 53].[1] The ability to be flexible for different and value-added services is the key factor. An efficiency gain, which is achieved in competitive markets, has a theoretical Pareto efficiency foundation, where no player can be better off without hurting any other [124]. But in a globally distributed system, such competitive markets can only be approximated. Nevertheless, a gain in efficiency in the telecommunications services market means a distributed surplus. Because the effects of multiservice provisioning onto incentive-compatible and efficient pricing schemes have not been studied in great detail, knowing how customers utilize services and therefore service demand as well as traffic profiles becomes crucial for viable multiservice pricing schemes. In addition, recent Internet service offerings and future advanced services lack another crucial component for businesses: adequate charging methods for differentiated services. For example, since the funding of transport services with revenue from separate services, such as content and entertainment offerings or advertising, is not transparent for open markets to cover network costs of transport services, different technologies for avoiding cross-subsidizing approaches are essential. In addition, high capacities provisioned in large backbone Internet networks, by separate players in the ISP market, will cross-subsidize as a viable business model.

[1] The number of ISPs in North America had reached about 7000 operational companies by August 1999, starting from about 1500 in February 1996.

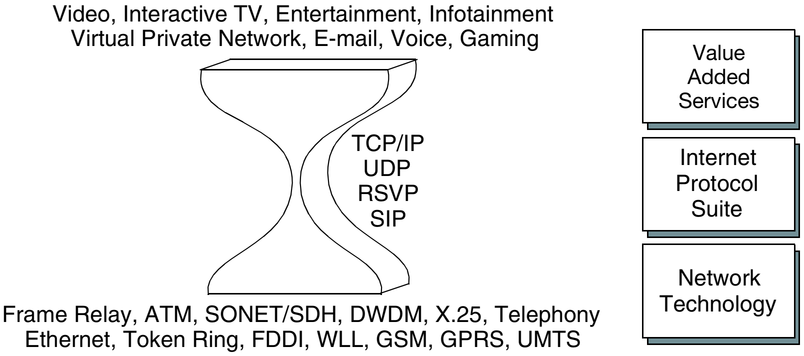
Video, Interactive TV, Entertainment, Infotainment
Virtual Private Network, E-mail, Voice, Gaming
TCP/IP
UDP
RSVP
SIP
Frame Relay, ATM, SONET/SDH, DWDM, X.25, Telephony
Ethernet, Token Ring, FDDI, WLL, GSM, GPRS, UMTS
Value Added Services
Internet Protocol Suite
Network Technology

Figure 4.1 Internet hourglass model: value-added services, protocols, and network technology.

New telecommunication services, in particular Internet services, impose a third degree of complexity to charging support systems (CSS).[2] Different services and content, and need to be service provider overlapping. This determines the stringent need to integrate concepts for interoperable and standardized charging solutions between providers for interoperator agreements, including content delivery services and provisioning pure data transport. Finally, the performance and scalability of CSSs are major factors in their suitability for Internet services charging. While the handling of many different Internet services, millions of customers, and various technology choices in use have to be optimized on the lower levels of a CSS, the upper levels for controlling the subscriber database and for managing the money collection process perform major functions for a competitive ISP infrastructure as well. The per-customer, responsive, and on-demand functionality of a modern CSS determines its optimal characteristics, which traditional billing systems or customer care systems do not provide.

Technical, economic, and operational prerequisites are closely related. Besides its emerging popularity, even for mobile users [90], and its increasing level of interconnectivity [35], the Internet offers the central possibility of accessing different types of usage information for many services at a single network layer. This is due to the fact that most services will be transported by IP which is independent of the underlying medium-access technology. For technical and operational reasons, this eases network deployment and maintenance, since a single network protocol needs only to be supervised.[3] For commercial reasons, this allows for straightforward and interesting product offerings, where single services, each of which resides on top of the IP, are bundled and offered as a value-added service or product. Figure 4.1 shows the hourglass-model [118], which describes the relationship between network technology, Internet protocols, and value-added services. Clearly, the Internet protocol suite is the “bottleneck” for provisioning value-added services based on network technology. However, based on the Internet’s years of success,

[2] Charging support systems are utilized to manage all economic tasks related to the commercial deployment of communication services, since operation support systems (OSS) deal with the management of technology issues and business support systems (BSS) handle customer care, customer support, and business model issues. While customer management and billing systems (CM&B) traditionally have been used for a back-office approach of handling batch-based monthly billing periods, CSS have to show a responsive, customer-oriented functionality.

[3] Different network layer protocols, such as IPx or Appletalk, become more obsolete, as estimations postulate [97].

developing appropriate means and mechanisms for this provisioning and charging task forms a challenge.

4.1.1 Outline

This overview on charging for commercial, integrated services, and packet-switched networks focuses explicitly on the case of the Internet and surveys the current state of the art and future steps for designing a flexible CSS for the Internet. Since charging for telecommunications services is not a new area as such, approaches required for charging packet-based networks show significantly different technical principles. The methodology chosen at this stage will combine known experience with solutions for the new technology.

While Section 4.2 at first discusses a scenario, and important customer and provider viewpoints, charging terms are introduced and appropriate terminology is defined to allow for an unambiguous discussion. Section 4.3 establishes an overview of related work in charging projects for the Internet as well as on traditional telecommunications systems, and summarizes metering, accounting, and mediation technology available today. To obtain a central understanding of various Internet services and technology choices, Section 4.4 elaborates on QoS methods, Internet network architecture, SLAs, and standardized data formats, which provide “hooks” for charging purposes. The relevant economic dimension is discussed in terms of pricing models for the Internet presented in Section 4.5 and ISP cost models in Section 4.6. The design of a CSS presented in Section 4.7 integrates all technical and economic issues discussed so far. Finally, major business model aspects for ISPs are outlined in Section 4.8, including provider segmentation issues as well as content charging. Section 4.9 summarizes this chapter and draws conclusions.

4.2 MOTIVATION AND TERMINOLOGY

The interest in a single multiservice network and its efficiency gains in terms of operation, utilization, and cost recovery, leads to the essential demand of appropriate technology alternatives, which will allow for the inherent differentiation of services. Let us compare this situation with a traditional business case in the toy railroad market. A store that cannot distinguish between expensive brand X model engines and cheap brand Y model engines for freight trains will have to charge all customers the same price for every engine sold. Even though these engines have extremely different values and features, they haul model trains only. Choices, if they exist at all and make sense, need to be distinguished by a set of parameters, such as degree of detail, correct coloring, or reliable technology, and each choice needs to be priced differently. With communications services, the situation is the same. Initially, communications services provide the possibility to communicate between two or multiple locations, independent of the service quality observed. However, on the one hand, customers want service choices and their explicit selection, such as services for Web surfing, fast downloads and file-transfer capabilities, or interactive gaming. On the other hand, and as a sort of causal derivation, ISPs need operational, efficient, and scalable technology, termed CSS, to charge for these service alternatives.

To provide a first overview of charging-relevant responsibilities [119], the tasks of a CSS are to be refined. In existing CSSs, sometimes called charging and accounting systems [115, 85], billing systems [22, 110], or simple charging systems [120], with different points of view and functionality included, many different tasks are subsumed. Basically, on the networking level, every provider maintains a network consisting of routers with network links between them. As illustrated in Figure 4.2, necessary functions for a CSS are metering, mediation, accounting, pricing and tariffing schemes, charge calculation, identification, and billing. While metering functions can be provided by components inside the networks on the wire and are integrated into routers, servers, or computers (active probing), they can form independent devices as well (passive probing). In either case, they generate raw accounting information (base accounting records), which show vast amounts of information. Mediation gathers, preprocesses, filters, merges, and correlates this information into accounting records, which are maintained in accounting systems. Call detail records (CDRs) [58] and Internet protocol detail records (IPDR) [21] are two examples of accounting records as agreed upon data structures (for more details, cf. Section 4.4.5). These accounting systems, in turn, forward all accumulated and perhaps abstracted accounting information through a charge calculation function toward the billing system. The charge calculation, which receives pricing and tariffing information as well as customer identification information from outside, translates the accounting information into charging records, hence, it maps the resource-oriented information into monetary, financial values. Within the charge calculation, discounting strategies, marketing-driven pricing schemes, or simply fixed prices which have been expressed in service- or customer-dependent tariff structures, can be applied on a per-customer basis. Finally, billing uses these values to prepare the invoice to be sent to the customer.

Depending on the particular system observed, setting prices and tariffs, calculating charges, or performing bill processing are integrated into a CSS. Additionally, they combine the maintenance of service classes, user profiles, customer data, identities, banking account data, and billing functions. Although these tasks can be and need to be distinguished clearly (cf. terminology as defined in Section 4.2.5), they are heavily centralized in today's CSSs. Future CSSs need to be able to integrate a variety of different charging

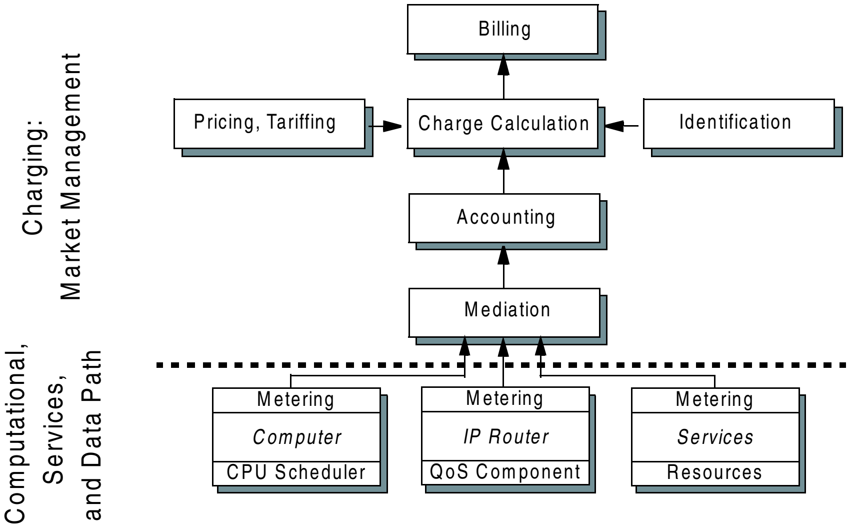

Figure 4.2 Concept of charging communications, services, transport, or even computation.

and accounting records from different communications providers or content providers, since customers' requirements are determined by the so-called "one-stop billing" approach [118]. This strongly suggests dividing existing monolithic charging systems into several components with clearly defined interfaces and interface data units. By doing so, it will become possible to exchange individual components and to integrate different components supporting different technologies without having to adapt the entire system.

Exact definitions of components and tasks of a CSS, as well as their applied nomenclature, form the basis for further comparisons and discussions. To allow for a precise specification, all charging-relevant terms are introduced below, based on an illustrating example scenario.

4.2.1 Scenario

A typical advanced communication scenario for the home[4] could include charging for communications in terms of specific transport, services, and content characteristics. Assume the presence of a single asymmetrical digital subscriber line (ADSL) at the home of a supervisory board member. To access his company, he needs a high-speed, secured, and remote VPN access. His wife organizes various activities for the local community and requires phone-conferencing facilities. The children like low-latency access to interactive gaming sites and music Web sites. These different user roles are represented contractually by a single customer, renting the ADSL line that runs from the home to an ISP. But how is charging applied?

Metering is performed on this single ADSL line. However, accounting is applied on a per-user basis to the metered and mediated data, which are correlated to the services utilized. The resulting accounting records contain, e.g., the duration of each data transfer, the obtained QoS characteristics (such as bandwidth consumed, delay encountered, and error rates experienced), and additional resource and device usage (such as the phone or a gaming device). The content may be indicated by different games played. These accounting records are fed into the charge calculation, which happens, e.g., in an administrative domain of the game provider. Pricing schemes for the ADSL service have been defined by the network provider and are adjusted to the special group of customers, which finally determines the tariff to be applied, e.g., a flat fee per month for 2 Mbit/s, proportional excess charges for bursts, and weekend reductions for VPN services. The tariff is based on the measurable QoS characteristics of a particular service and, in case of content, on the type of content distributed. All charges for services and content are calculated, possibly at different locations, and collected in multiple charging records from the billing system. A number of charging records for a certain prearranged period of time are accumulated and billed to the customer, who has been identified in advance and represents all user flows metered so far. Finally, the customer may decide at this point, or in a predefined manner, how the bill will get paid, applying traditional payment schemes or electronic systems, e.g., by credit card payments using secure electronic transactions (SET) [105] or electronic money.

4.2.2 Viewpoints

As we can see from the preceding scenario, different points of view on the CSS and its tasks exist. Concerning the customer, after service provisioning, the bill for the service

[4] A similar scenario for a customer premises network scenario can be defined as well.

utilized summarizes a week or month of communication activities and content. Generally speaking, there is no incentive to disconnect from the dial-up ADSL Internet connection, when there is no charge per time or volume. Therefore, only a flat fee for basic services without special requirements will be incentive-compatible. However, an always existing spending cap will limit the amount of services utilized over time. As stated in Ferrari and Delgrossi [37], the following charging properties exist from a customer's perspective: comprehensibility, controllability, predictability, stability, and fairness. With a commercial ISP, a high revenue maximization and best cost-recovery strategy should be achieved [122]. Therefore, the pricing model applied and the service bundle offered in a given market situation determines the optimization dimensions for customers and providers. The provider's charging properties need to cover settlements and should allow for a high probability of cost recovery, the competitiveness of prices, the encouragement of service usage, low implementation costs, and low usage costs. In summary and based on Stillos *et al.* [118] and Farten *et al.* [61], the customer and provider view points are clarified, which are complemented by further requests.

4.2.2.1 Customer Point of View On the one hand, customers' budget constraints and spending strategies are to get as much service for as little money as possible. The underlying economic principle can be formulated as "users buy the best service bundle they can afford." With software agents and brokers, automated or optimal spending strategies for finding telecommunications service and service bundles can be used to achieve this goal. Targeted at services include the following: phone, fax, Internet access, value-added services, video-on-demand (VoD), VPN, gaming, or conferencing. These services can be chosen by software agents, working on behalf of customers in services markets, which become more and more competitive. Users only need to express their preference and budget. On the other hand, the opposite strategy may be chosen, where the quality of the service is given and the current market price returned. An implicit customer segment may be drawn here, where business customers mainly will determine their needs and will accept or reject a presented price in favor of the requested service, and where the private users will specify their budget constraints and accept or reject the service offered accordingly. This leads to explicit customer requirements:

- *Predictability of Charges* Users want to be able to predict all of the costs of using a particular application, which include expenditures for communication services induced by this application. Therefore, an exact a priori specification of communication charges would be desirable. However, if this requirement cannot be fulfilled, a set of weaker demands can be sufficient. First, a user should be able to roughly estimate the charges. Such an estimation does not need to be exact, but should give at least a rough feeling to the user—similar to the knowledge that an international phone call of some minutes duration costs more than a dollar and not just a few cents. Second, a worst-case price should be known. Finally, a user must not be charged a higher price than previously announced, without the user's explicit approval.
- *Transparency and Accuracy of Charging* To find out how much is spent for which application and what the reasons are for this, users need to be able to determine the costs of a particular session, e.g., if an application uses several flows, costs for each of these should be stated explicitly. Furthermore, for some users it might also be of interest to see what it is inside the network that causes major

charges. This may give them information to switch to a different provider in the future. Detailed per-session information about charges can also be used to decide whether a certain service and its quality offers good value for price. Since not all users are interested in such details, each user must be able to decide how much information should be given.

- *Convenience* Charging components should not make using the communication services much more difficult. Charging mechanisms themselves, as well as the final bill based on the information gathered by the charging system, must be convenient for users. Hence, it must be possible for users to define "standard charging behavior" for their applications so they are not bothered with details during the start-up of an often-used application. On the other hand, they should be able to change such a description easily to have control over their expenditures, e.g., changing spending caps. Furthermore, most users want to have as few separate bills as possible, i.e., have contracts and thus business procedures with only one provider.

4.2.2.2 Provider Point of View From a business point of view, the costs of providing telecommunication services must be recovered in order to guarantee the stable, long-term existence of a provider. While pricing for traditional telecommunications services is well understood by companies, large as well as small ISPs still struggle to make profit [81]. Furthermore, providers want to maximize revenues, particularly in the open market of Internet services provisioning. This leads to explicit provider requirements:

- *Technical Feasibility* The charging approach and its mechanisms must be implementable and operable with little effort. Otherwise, if it becomes too complex, costs for the charging mechanisms might be higher than their gains. A set of real-life user trials needs to be performed to assure any of these statements. The added overhead for communication due to additional information transmitted between senders, network nodes, and receivers, and also for processing and storage purposes, especially in network nodes, e.g., to keep and manipulate charging information, must be as low as possible [36]. In addition, the introduction of scalable and low-effort security mechanisms is essential for any type of counterfeit-proof charging records and billing data.
- *Variety of Business Models* The business of providing network service over packet-switched networks must be sustainable and profitable to attract necessary investments into the infrastructure. It is not likely that all service providers will adopt exactly the same business model and strategies. Therefore, charging mechanisms must be flexible enough to support a large variety of business models and interoperate between multiple network domains employing different models. In addition, a charging system must be flexible enough to handle different pricing strategies, for example, during peak and off-peak times.

4.2.2.3 Further Requirements A CSS and the operation of a network with respect to charging require a set of additional requirements to be fulfilled.

- *Flexibility* When information is transmitted from a sender to one or several receivers, the flow of value associated with this information can be (1) in the same direction as that of the data flow, (2) in the opposite direction, or (3) a mixture of both, because both sides benefit from the information exchange [11]. For example, in the first case, the sender transmits a product advertisement, in the second case, the receiver retrieves a movie for playback; and in the third case, both sides hold a project meeting via a video-conferencing system. To support these different scenarios, a charging architecture must provide flexible mechanisms to allow participants in a communication session to specify their willingness to pay for charges in a variety of ways. Senders must be able to state that they will pay for some percentage of the overall communication costs or up to a specified total amount. Similarly, receivers may state what amount of costs they will cover. Additionally, charging mechanisms must allow flexibility in the distribution of communication charges among members of a multicast group [51].

- *Fraud Protection and Legal Security* One of the most important issues demanded by participants is protection against fraud, i.e., that they do not have to pay for costs they have not incurred and that no one can misuse the system. The fear of users is that a provider may cheat or that other users may use their identity or derogate from them in any other way. Providers want to be sure users indeed pay for the used service. A prerequisite against fraud is technical security, such that users cannot damage, misuse, or intrude on the provider's communications systems. Legal security denotes the demand that in case of a failure, there is enough information to determine responsibility for it.
- *Stability of Service* When a particular service with a certain quality has been agreed upon by the user and the provider, it must be ensured that the service indeed is delivered to the user. Hence, an exact definition of "quality assurance is met" is needed. On the other hand, users must be able to estimate the impact of such quality goals on their applications, so the definition must not be too complex. For example, multiple users want a video conference application, so they will likely request a communication service with a specified bandwidth and delay. If the provider promises delivery of this service, users expect no quality degradation and a very low probability of service disruption during the conference. Should quality degradation or service disruption occur, an appropriate refund mechanism must be applied which largely depends on the type of application, and hence, should be negotiated during setup of the communication service.
- *Reliability of Service* In order to provide the infrastructure for an integrated packet switched network, service availability must be very reliable. Current telephone networks are designed to keep the blocking probability on the order of 10^{-4}. Similar requirements are likely to apply to integrated services networks. To assure such a low blocking probability, even during peak hours, significant effort in network and traffic engineering is necessary, which in turn must be accompanied by appropriate business calculation. A slightly different situation exists in the case of per-packet QoS guarantees without explicit flow admission control. In that case, the notion of blocking probability might be replaced by reliability of service measured in terms of probability that the promised level of QoS is violated.

4.2.3 Fairness and Utility

Assuming higher market efficiency and the views of users and providers described earlier, *fairness* defines to what extent each party profits from improved efficiency. Furthermore, fairness requires that all customers pay the same price for the same telecommunications services at the same time. In networks that do not charge for usage, i.e., the currently used Internet, fairness is defined in technical terms. Therefore, the introduction of charging needs to define a new notion of fairness, since a common sense of fairness as "an allocation where no person in the economy prefers anyone else's consumption bundle over his own" [39] does not reflect technical requirements. Historically, fairness has been considered in economics and at a later stage in resource-sharing work.

Traditionally, the notion of the fairness of network operations has two instances. One is fairness of the TCP, based on TCP congestion control algorithm and binary exponential back-off strategy in case of heavy traffic, where the protocol tries to serve all connections with the same throughput. This is a type of proportional fairness per unit bandwidth. Unfortunately, this works only for regional access without high delay variations between competing connections [25]. The other type is max-min fairness, where everyone is equal everywhere [3], i.e., max-min fairness maximizes the minimum share of sources, whose demand is not fully satisfied. For example, sources S_n demand x_i each of a single resource with capacity C. After ordering all demands x_i in $x_1 \leq \cdots \leq x_n$, S_1 is assigned a share of C/n. The reminder of $C/n - x_1$ is evenly distributed to $n - 1$ sources: $C/n + (C/n - x_1)/(n - 1)$. As long as there are leftovers, this process is iterated until the resource C is fully shared.

The concept of maximized welfare defines an additional fairness notion. In this case, a welfare function aggregates a number of utility functions, each of which increases in all of its arguments. According to [27], microfairness determines a fair distribution of network resources at a much finer granularity to applications. This affects basic data delivery, advanced data delivery, QoS, and application quality. Instead, macro-fairness is related to network mechanisms, which consider flows in a given network, such as mechanisms for fair queuing to perform a flow protection, e.g., by the WFQ scheduling strategy.

In the case of charging services, these notions have to change to so-called proportional fairness per unit charge [41, 64], which is relative to the number of charges a customer is willing to pay for. On the other hand, a market and competition mechanism is useful in providing users with the best and most inexpensive level of service, while creating incentives for network providers to supply more resources when there is sufficient demand.

However, the basic prerequisite for defining the notion of fairness requires the existence of a mechanism to express preferences of service usage. In economic terms these preferences are expressed by utility functions, where basically an outline of network services is applied to the perception of performance [26, 42]. Theoretically, utility functions are required to find an optimal resource allocation, while maximizing the utilization of the resource. This case is relevant, e.g., for ISPs to optimize their service provisioning. Therefore, service users can be modeled by their required utility or degree of satisfaction, since users value their perceived quality over price. For example, the utility function for a file-transfer application in Figure 4.3 shows that the user's satisfaction, S, is infinite, if the file can be transferred instantaneously. However, this utility decreases at a rate proportional to αt, where α defines a constant, e.g., depending reciprocally on the speed of the access link, and t depicts the transfer time. In the case where $t > S/\alpha$, a negative utility determines a file transfer that lasts too long, which changes the user's original valuation. Clearly, the larger α is, the longer the file transfer is accepted by the customer. Applying different values of α to different access speeds and assuming a 1-Mbyte file to be transferred, Table 3.1 indicates different utilities for a range of different user profiles.

In general, finding out about utility functions for all the different applications and services becomes awkward. However, two generally accepted classes of Internet applications

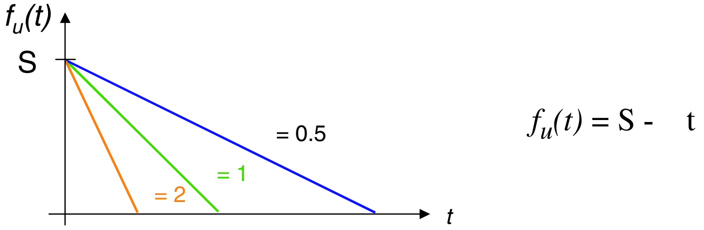

Figure 4.3 Example utility function $f_u(t)$, for a file-transfer application.

show a significantly different type of utility curve, which has been investigated by Cocchi *et al.* [26]. These characteristic curves are depicted in Figure 3.4 for elastic and real-time applications.

Assuming advance knowledge of these utility functions, they can be used as a means for distinguishing performance on the provider side; however, nonindividual or aggregated utility functions, which characterize the aggregated backbone Internet traffic or aggregated customer premises network traffic, are still an open issue for research. In addition, it remains unclear if the sound economic theory of utilities will be compliant with the user's privacy requirements. Even when an ISP knows about handling aggregated utilities, it still requires advance knowledge on a per-application basis of single input utility functions of its users and according timing information, when an application will be utilized. Even when users do not object to giving their per-application utilities to their ISP, predetermining the exact demand on an hourly or even per-minute basis is almost impossible. In addition, extensions of utility to different applications of QoS parameters, e.g., ranging from throughput to delay and jitter, are required. Finally, users evaluate charges for communication services in terms of value, which is dependent on the perceived utility of the task performed. But due to network status and congestions, the described utility may deviate from the utility perceived by the customer, therefore, investigations of the user's perception of QoS and the service started just recently, such as performed in [6], are needed.

4.2.4 Charging Tasks and Terms

Charging refers to the number of main activities to be performed for packet-switched networking and for other communications services' provisioning, if financial values are to be mapped onto resource usage or consumption. A fully operational CSS is required to accept, answer, and offer certain messages and information being exchanged over the net-

Table 4.1 Example Utilities for a File-Transfer Application

User Profile	Utility	Transfer Time (s)	α	Minimal Access Speed (Kbit/s)
Executive board member	Good	1	125.0	8000
	Acceptable	10	12.5	800
	Unacceptable	63	2.0	128
Residential user	Very well	63	2.0	128
	Good	143	1.0	56

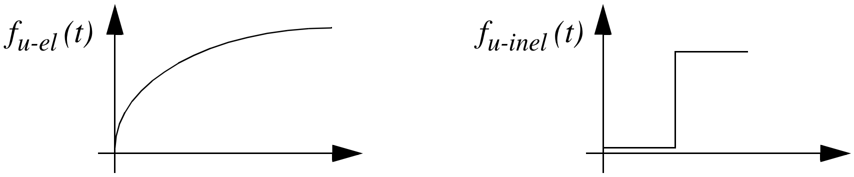

Figure 4.4 Characteristic utility functions for elastic and inelastic applications.

work. Based on the cited literature, there is a range of different, sometimes contradictory task descriptions and definitions. Therefore, the terms *metering, pricing, tariffing,* and *billing* are discussed at this stage and a unique definition is given in Section 4.2.6.

4.2.4.1 Charging and Charge Calculation Charging is one of the most important terms in the domain considered. Based on Webster's Dictionary [128] "to charge" is defined as "to impose or record a financial obligation."

Therefore, "charging determines the process of calculating the cost of a resource by using the price for a given accounting record, which determines a particular resource consumption" [118]. Thus, charging is defined as a function that translates technical values into monetary units. The monetary charging information is included in charging records. Prices already may be available for particular resources in the accounting record or any suitable resource combination depending on the network technology or the application. Standards and research work tend to show quite a close understanding of tasks and definitions for charging.

The European Telecommunications Standardization Institute (ETSI) [33] defines charging as follows: "Charging is the determination of the charge units to be assigned to the service utilization (i.e., the usage of chargeable related elements)." In Karsten *et al.* [61] a full process point of view is defined: "Once these accounting records are collected and prices are determined in full pricing schemes on unit service, e.g., encompassing different quality levels for services or service bundles, the data for an invoice need to be calculated. The process of this calculation is termed charge calculation, performing the application of prices of unit services onto accounting records determining resource consumption. Thus, the charging function transforms mathematically unequivocal technical parameter values into monetary units. These units need to be collected, if they appear at different locations in the given networking environment, and are stored in charging records. Of course, accounting as well as charging records determine a critical set of data which need to be secured to ensure its integrity when applied to calculate monetary values or when used to compute an invoice's total." The charging process for business models offering ATM services is also called a "rating and discounting process" [112] and is "responsible for the charge calculation according to a specific pricing policy using the collected usage data." Therefore, charging mechanisms correlate service usage and calculate the charge the customer is faced with after the service utilization. Finally, according to [16]:

> Charging is the process of evaluating costs for usage of resources. Different cost metrics may be applied to the same usage of resources, and may be allocated in parallel. An example would be a detailed evaluation of resource consumption for further processing by the service

> provider, and a simple evaluation of resource usage for online display of current costs. A detailed evaluation of the resource consumption can be used for generating bills to the customer, or for internal analysis by the service provider. A simple evaluation of current costs can be used for displaying an estimation of accumulated costs for the service user, or for control purposes by the customer organization or by the provider. Cost allocation assigns costs to specific endpoints, such as sender and receivers of a multicast group.

4.2.4.2 Accounting Accounting is the second of the most important terms frequently used. Based on Webster's dictionary [128] "accounting" is defined as "the system of recording and summarizing business and financial transactions and analyzing, verifying, and reporting the results."

Accounting considers two different points of view. The first one is related to economic theory, where accounting relates to business processes, including profits and benefits. The second one relates to technical aspects, where technical parameters are collected. Therefore, applied to the networking environment, accounting "determines the collection of information in relation to a customer's service utilization being expressed in resource usage or consumption" [118]. Thus, accounting describes a mapping function from a particular resource usage into technical values. The information to be collected is determined by a parameter set included within an accounting record. This record depends on (1) the network infrastructure, which supports the service, e.g., IP, narrowband integrated services digital network (NISDN), ATM, or frame relay, and (2) the service provided. The content of an accounting record is of a technical nature, such as a parameter describing the duration of a phone call, the distance of a high-speed network link utilized, or the number of market transactions done. This accounting record forms the basis for further charging steps.

ETSI [33] defines accounting as "revenue sharing amongst operators." The ITU-T [54] defines terms in an economic sense: "Accounting revenue division procedure: the procedure whereby accounting revenue is shared between terminal administrations and, as appropriate, between the administrations of transit countries; Accounting rate: the rate agreed between administrations in a given relation that is used for the establishment of international accounts; Accounting rate share: the part of the accounting rate corresponding to the facilities made available in each country. This share is fixed by agreement among administrations."

The Network Working Group of the IETF has published an informal Request for Comment [14] on summarized Accounting Attributes and Record Formats. It defines an accounting server as "A network element that accepts usage events from service elements. It acts as an interface to back-end rating, billing, and operations support systems." Within this context, usage events refer to "the description of an instance of service usage," and service elements include all types of service provisioning devices, such as application bridges or gateways.

A networking technology-oriented explanation of the tasks and interfaces for accounting is presented in [61]:

> [. . .], these units need to be accounted for, traditionally performed on a percall basis over time. However, in packet-switched networks, the accounted for information may encounter a huge number of different parameters, e.g., number of packets sent, duration of a communication, number of transactions performed, distance of the communication peer, number of hops traversed, or bandwidth used. Depending on the protocol layer applied for this accounting task, only a subset of accounted for parameters are useful. In general the accounting record

> determines the container for collecting this information. These records and their special appearances depend on the networking technology used, such as N-ISDN, ATM, Frame Relay, or IP. They can also be created for application services, for example, the call data record is being used for this purposes in H.323 IP telephony. Further, the Real-time Flow Measurement working group within the IETF investigates appropriate accounting mechanisms.

The accounting process applied to ATM services is defined in [112] and complies with the ITU-T process definitions just summarized from [54]. Additionally, the IETF's Authentication, Authorization, and Accounting (AAA) working group has applied the protocol Diameter (cf. Section 4.5) to specific accounting purposes. Finally, [16] states: "the process of accounting involves the following functions: collection of usage data by usage meters, creation of accounting records (data structures, or protocol data units of an accounting protocol), transport of accounting records, and collection of usage data by an accounting server."

4.2.4.3 Metering Broad commonality and conformance can be observed for the definition of metering. ETSI [33] defines metering as ". . . the measurement of 'components' which can be used for charging such as the duration of the call . . . named also 'collection of charging information'." A full task and term definition for metering is included in [61]: ". . . there remains a single technical prerequisite for identifying and collecting accounting data. This process is called metering." Based on existing technical equipment in operation, metering tasks identify the technical value of a given parameter or resource and determine their current usage. If possible, metering can be tied to signaling events. Otherwise, it can be performed regularly, e.g., every ten seconds or every hour, it can be stimulated on other external events, such as polling requests, or it can be performed according to some statistical sampling scheme. In that case, it is closely related to network monitoring. The IETF's management information base (MIB) for switched networks and the simple network-management protocol (SNMP) [18] architectural framework can provide a means of keeping monitored data." New Zealand's approach on metering resulted in network traffic meter (NeTraMet) specifications [13] (cf. Section 4.3.2). Also for the ATM approach, network element usage metering functions are described as being responsible for the generation and reporting of accountable resource information [112].

4.2.4.4 Mediation Mediation is concerned with the full communication service and its information level of data, such as the "begin" and "end" of communication sequences [22]. It filters, aggregates, and correlates raw data to yield different views on current network activity, and enforces business rules to package data into the form that is known to billing systems. Therefore, mediation controls a (traditionally deterministic) preprocessing of metered data according to policies and rules that have been set up by providers to minimize the amount of data collected and stored. Based on the application scenario, the final accounted-for data have to reflect the intended level of detail, e.g., guided by legal restrictions or business policies and pricing models to be applied.

4.2.4.5 Pricing and Tariffing Pricing is the process of setting a price on a service, a product, or on content [118]. This process is an integral and critical part of business and is closely related to marketing. Prices can be calculated on a cost/profit basis or on the current market situation. For businesses selling telecommunication services, prices are set on predefined services, where the quantity used is measured, e.g., in units, time, distance, bandwidth, volume, or any combination thereof. These basic quantities to be priced are obtained from accounting devices and depend on the network type. Tariffing is a special case of pricing, normally regulated by governmental and political economic impacts [118]. Tariffs have been applied to the traditional telephone network. Karsten et al. [62] elaborate on pricing as "setting a certain price for a unit service used. Appropriate pricing of network communication provides incentives for reasonable usage of resources." In addition, a pricing scheme "describes a particular choice [. . .] and is applied to unit services offered from a communication service provider." Furthermore, a document filed with the Federal Communication Commission (FCC) or a state public utility commission by a (regulated) telephone service provider that details services, equipment, and pricing of services they provide, e.g., calling plans, is termed tariff. So-called telemanagement software uses these tariffs to determine the charge of a telephone call. For auditing purposes, telemanagement reports should match the end-of-month bill a company receives from its provider, because the same tariffs are being used to charge for calls made.

A quite similar use of pricing has been observed with respect to other related work. ETSI [33] defines pricing as "[. . .] the correlation between 'money' and 'goods' or 'service'," while it is noted that "the term is not generally used in telecommunications, the usual term being 'tariffing'." A clear distinction between price and tariff has been drawn within the M3I project (cf. [85, 119]). While "the price determines the monetary value the user owes a provider for his service provisioned and utilized, in particular it is the price per unit service. It may be based on charges and costs or it may be determined by other marketing means," price setting is defined as "the specification and the setting of prices for goods, specifically networking resources as well as services in an open market situation." Hereby the tariff "[. . .] determines [the algorithm used to] charge for a service usage. It is applied in the charge calculation for a given customer and service he utilizes to calculate the charges." The process of tariffing decides upon the algorithm used to determine a tariff for a given service and/or customer.

4.2.4.6 Cost Based on [128], "cost" covers many different explanations: "The amount or equivalent paid or charged for something: Price." Obviously this usage of cost confuses mainly in technical areas. Therefore, the definition "the outlay or expenditure made to achieve an object" heads in the right direction. However, in economic accounting, various different sorts of costs are distinguished, such as general costs, capita costs, joint costs, opportunity costs, or marginal costs. Details can be obtained from, e.g., [124]. The M3I project states [119]: "Costs determine the monetary equivalent on equipment, installation, maintenance, management, operation of networks, network entities, and service provisioning. Many different types of costs can occur but it is important to note that [. . .] only costs in terms of money are of interest."

In particular, the business area of network services provisioning is considered in [61]. Since real variable costs basically do not exist or are extremely limited in nature, the term "cost per service invocation" characterizes opportunity costs (lost revenue), because resources are bound and cannot be sold otherwise. Consequently, resource usage and consumption is considered as the main cost factor. Networking is characterized by the following aspects [61]:

- High fixed costs (installation and maintenance of infrastructure)
- Low variable costs

- Fixed capacity
- Nonstorable resources and products

Business economics calls an appropriate management theory for such a business field as yield management [75]. Based on these characteristics, it is appropriate to differentiate prices according to variations in demand, instead of performing a full cost calculation for prices. The marginal return, which is given by the difference between prices for sale and variable costs, is considered to be the primary variable. The goal of a yield-management approach is to optimize the sum of marginal returns over a certain investment cycle. This sum has to exceed investments in order to allow for a profitable business.

4.2.4.7 Service Level Agreements and Interconnection While agreements between providers are used for a long period of time in traditional telecommunications, they become important for the Internet market as well. SLAs for the Internet define formal aspects of a contract between an ISP and a customer as well as the type of traffic to be forwarded into the global Internet to allow for consistent delivery and measurement of services (cf. Section 4.4.4). SLAs are the basis for (network) interconnection between potential business partners. Traditionally, (1) it includes the collection of performance and traffic data, including availability, delay (round-trip time), and throughput measures; (2) it provides the basis for a fee calculation; (3) it offers mechanisms for data reporting and presentation; (4) it defines measurement points, such as end-to-end or switch-to-switch metrics; and (5) allows for the comparison of contracted thresholds with measured data. SLA management systems will form the key for commercial operation and management of Internet services. This covers rules applied to measurements taken inside the network, in an end-to-end fashion, or in the local loop.

According to Kilkki [69], an SLA includes, among others, definitions of the bit rate of the access link, delay information to the provider's domain, information on the network's availability including compensation and reporting activities, and timing data for repairing and installing new services. In addition, the classification rule for the stream, the specification of the traffic the customer ISP is allowed to send to the selling ISP, and the service level to be applied, e.g., best effort or premium service with bandwidth and delay, will be included. SLAs may show a (semi-)static or a dynamic behavior. While static SLAs are negotiated between the provider and a customer, dynamic SLAs change without human interactions in an automated fashion [2]. In addition, a second differentiation is made on the service quality; i.e., relative service quality and constant or variable service quality [69]. Further SLA developments address those questions of an SLA's scope, defining roles of possible interactions, the quality of privacy to be applied, e.g., encrypted/ not-encrypted, secure/insecure transmission, agreements on price, the payment method, the definition of noncompliance and reimbursement schemes, a monitoring method for service compliance, and the SLA duration [28].

Above this technical level of interest, financial interactions between providers and customers need to be considered as well [52]. While provider–provider interactions are termed "financial settlements," provider–customer interactions often are referred to as "payment." Settlements are defined as the "payment or adjustment of an account" [128], where the account in economic terms reflects the monetary equivalent for service provisioning between two business roles, such as providers and customers. These definitions lead directly to the discussion of billing and payment in the next section.

Different interconnection schemes determine various alternatives for offering, retailing, reselling, or wholesaling business schemes. However, no clear distinction is possible, since a recursive structure of the ISP market exists and relative roles in this market vary over time. In addition, the deregulated environment proposes no external entity that would be able to decide on the relationships between ISPs, customers, and their roles. Therefore, interconnection is the technical prerequisite for exchanging data. This exchange is guided and legally guarded by an SLA.

4.2.4.8 Billing and Payment Billing "denotes the process of transforming the collected charging information for a customer to his/her bill" [118]. It includes the process of listing for a customer all charging information being contained in charging records that were collected over a period of time, e.g., one month. The bill summarizes all charges and indicates the amount to be paid. The bill may identify the method of payment chosen or selected, and it is transferred to customers electronically or on paper. The method of payment defines a scheme, how money is exchanged between participants, e.g., between customers and retailers or service users and providers. In general, electronic payment systems or traditional systems as utilized for traditional payment transactions are applicable.

Finally, this is similar to "the process of consolidating charging records on a per customer basis and delivering a certain aggregate of these records to a customer is termed billing" [61, 16]. The collection of these charging records requires adequate protocol support, e.g., including authentication, to allow for counterfeit-proof computation of invoices. The aggregation of monetary values (billing data) can be performed on a daily, weekly, monthly basis, or some other accepted period of time. The bill or invoice, summarizes a number of contracted details and parameters originally collected in the accounting records. Songhurst [112] distinguishes between various billing mechanisms and options based on the form of the bill (e.g., itemized or aggregated) or the time of delivery (e.g., periodic, per-call, or prepaid). The three-tiered billing architecture of Cisco systems [22] defines the process of collecting usage and accounting as billing, and refers to the sending of bills as the invoicing system.

Additionally, all bills show the amount of money to be paid by the customer to the service provider. This money may be delivered traditionally (termed payment) on paper or in an electronic-funds transfer fashion. Because new methods of payment exist, the method of how the exchange of money between buyers and sellers will be performed may include advanced electronic payments schemes.

4.2.5 Definitions of Terminology

Based on these observations, the following definitions (expanded from [121]) are utilized in the following sections:

- *Accounting* Summarized information (accounting records) in relation to a customer's service utilization. It is expressed in metered resource consumption, e.g., for the end-system, applications, middleware, calls, or any type of connections.
- *Accounting Record* An accounting record includes all relevant information acquired during the accounting process. Its internal definition can rely on Call Detail Records, Internet Protocol Detail Records, or similar standardization proposals.
- *Billing* Collecting charging records, summarizing their charging content, and delivering a bill or invoice including an optional list of detailed charges, to a user.

- *Billing Record* A billing record includes all relevant information acquired during the billing process. Its internal definition should match proposed billing-system interface standards.
- *Charge Calculation* Completing the calculation of a price for a given accounting record and its consolidation into a charging record, while mapping technical values into monetary units. Therefore, charge calculation applies a given tariff to the data accounted for.
- *Charges* Charges determine what is owed for a particular resource utilization. It is contained in a charging record.
- *Charging* The overall term "charging" is utilized as a summary word for the overall process of metering resources, enumerating their details, setting appropriate prices, calculating charges, and providing the fine-grained set of details required for billing. Note, that billing as such is not included in this definition.
- *Charging Record* A charging record includes all relevant information acquired during the charge calculation process. Its internal definition is for further definition, but may correspond to CDRs or IPDRs.
- *Charging Support System* Based on the definition of charging, the CSS implements all required tasks and interfaces that are essential and sometimes optional for managing charging-relevant data. It is concerned only with the collection of technical services data, which are mapped onto financial values, while an OSS deals with technical management tasks only.
- *Costs* Costs determine the monetary equivalent for equipment, installation, maintenance, management, operation of networks, network entities, and service provisioning. Many different types of costs can occur, but it is important to note that in the case of CSS only costs in terms of money are of interest.
- *Mediation* The task of mediation includes the filtering, aggregation, and correlation of raw, metered data. Mediation reconstructs sessions, matches measured IP addresses with users, if possible, and perform reconciliation.
- *Metering* The task of metering determines the particular usage of resources within end-systems (hosts) or intermediate systems (routers) on a technical level, including QoS, management, and networking parameters.
- *Payment* The task of payment defines the manner in which money is transferred between commercial partners to settle a rendered bill.
- *Price* The price determines the monetary value the user owes a provider for his or her service provisioned and utilized; in particular, it is the price per service unit. It may be based on charges, costs, or profits, or it may be determined by other marketing means.
- *Pricing* The specification and the setting of prices for goods, specifically networking resources and services in an open market situation. This process may combine technical considerations, e.g., resource consumption, and economical ones, e.g., applying tariffing theory or marketing, and it is part of the enterprise policy layer and requires that appropriate means of communication be in place.
- *Quality-of-Service* QoS defines the quality of a service provided. It contains technical application-level as well as network-level views and definitions. Although particular specializations exist, a commonly agreed upon definition will be used. Here, the definition from ITU-T, E. 800 [57], is applied for QoS: "The collective effect of

service performance which determines the degree of satisfaction of a user of the service." Refinements are applied where necessary.

- *Service Level Agreements* An SLA defines the level of interconnection for service provisioning between providers or between a provider and a customer. The types of parameters and attributes ranges from technical values to payment schemes applied, and is still in flux.
- *Settlements* Settlements refer to the payment or adjustment of an account that needs to be enforced due to an agreed upon SLA between business roles. While important technical issues are handled by the SLA, the settlement is concerned with the financial level of interaction only.
- *Tariff* The algorithm used to determine a charge for the use of a service. It is applied in calculating the charge for a given customer and the service utilized to by that customer. Tariffs can contain, e.g., discount strategies, rebate schemes, or marketing information.
- *Tariffing* The process of deciding upon the algorithm used to determine a tariff.

4.2.6 Networking Terms

Finally, just as with the preceding charging terminology, the main networking terms and their definitions in the context of packet-switched networking are introduced, based on [61]. Resources of interest for a packet-switched communication infrastructure are given by the computing power and the buffer space of switching systems, as well as the by capacity of transmission lines (links). Computing power is a major determining factor of the number of packets that can be serviced and the amount of flow that can be handled in the case of flow-based service models. Consequently, it determines all possible computational levels of service differentiations. It should be noted that the capacity of a transmission line is useless if it cannot be fully utilized by a feeding switching system. Therefore, important resources that need to be assessed for a pricing scheme are given by the flow setup overhead (if it exists), the packet rate (and its schedulability), the traffic bandwidth (not necessarily constant), and the buffer space.

In addition, the terms packet rate and traffic bandwidth are listed separately to distinguish between two distinct units of measurement for the transmission capacity. The number of packets handled is mainly determined by the computing power of a switching system, whereas the traffic bandwidth is limited by the overall link bandwidth and throughput of a switching system. Furthermore, the input parameter for network capacity dimensioning and for pricing schemes is given by the access bandwidth of customers and adjacent providers. Eventually, this parameter defines the maximum resource QoS requests the local link.

4.3 RELATED WORK

Charging for communication services has been on the research-and-development agenda for years. However, service profiles and networking technologies changed rapidly and different communication paradigms, e.g., the shift from a circuit-switched to a packet-switched communication model, evolved. Therefore, new approaches for Internet transport, services, and content charging become necessary. This section on related work discusses major steps and projects taken for handling Internet charging. It adds a short description of accounting and mediation technologies for the Internet environment, and finishes the discussion with a brief review of traditional billing of telephony and ATM services.

4.3.1 Charging Projects

The number of projects on the Internet concerned with charging has increased quite significantly. Only a small number of recent and charging-centric work dealing with system's design and modeling is summarized below. Another overview can be found in [114]. Many projects dealing with charging and accounting functionality at the network level try to achieve a high independence from pricing models [118]. However, it has been noted that pricing in general and usage-based pricing in particular can impose a high overhead on telecommunication systems [78, 108]. Any form of usage-based pricing for various telecommunication services is interesting, because underlying resources (such as satellites, frequencies, cables, routers/switches, and most notably operating personnel) are scarce and very costly. The traditional Internet pricing model has been critiqued constantly in the past for its economic drawbacks of not being incentive-compatible [108, 24, 47]. Furthermore, it is inflexible—for example, it does not allow for combined sender/receiver payments—and does not provide economic signals, which are needed for network planning and expansion. But most importantly, the current model is based on the assumption of a single-service best effort network that provides a similar service to all customers. Therefore, the multiservice paradigm needs to be investigated with respect to heterogeneous networking infrastructures and technologies of the Internet. An early per-flow billing system for TCP flows and initial ideas on a billing service design are presented in [29] and [110], respectively. Advanced per-flow charging and accounting approaches based on reservations have been tackled in [36, 59, 62]. For the case of an integrated-services packet-switched network, the approach in [37] defines a service-dependent charging policy. Based on a list of charging properties and a cascading queuing station model, a charging formula is presented and discussed, which includes a reservation and usage portion.

4.3.1.1 Charging and Accounting Technology for the Internet The objectives of the Swiss National Science Foundation project charging and accounting technology for the Internet (CATI) [115] included the design, implementation, and evaluation of charging and accounting mechanisms for Internet services and VPN. This covered the enabling technology support for open, Internet-based EC platforms in terms of usage-based transport service charging as well as high quality Internet transport services and its advanced and flexible configurations for VPNs. In addition, security-relevant and trust-related issues in charging, accounting, and billing processes have been investigated. Important application scenarios, such as an Internet telephony application, demonstrated the applicability and efficiency of the developed approaches [116]. This work was complemented by investigations of cost recovery for ISPs, including various investigations of suitable usage-sensitive pricing models for end-to-end communications based on reservations [36, 122], as well as SLAs between service providers [28].

4.3.1.2 Market Managed Multiservice Internet The 5th Framework European IST project, Market-Managed Multiservice Internet (M3I) [85], aims at the design and implementation of a next-generation system that will enable Internet resource management through market forces, specifically by enabling differential charging for multiple levels of service. This novel approach, offering a charging system, will increase the value of Internet services to customers through a greater choice in price and quality and better QoS through reduced congestion. Flexibility will be improved for the ISP, management complexity reduced, and the potential for increasing revenues is great. Price-based resource management pushes intelligence and, hence, complexity to the edges of the Internet, ensuring similar scalability and simplicity of the current network. It is intended to design a trial system, which will enable players in the Internet services market to explore sophisticated charging options and business models with their customers.

4.3.1.3 Internet Demand Experiment another highly important question concerns the issue of user acceptance of pricing schemes. The Index Demand Experiment (INDEX) project was started in order to investigate user reaction when exposed to various pricing schemes for different qualities of Internet access [20, 30]. It turned out that users were not opposed to flexible pricing models. Moreover, the widespread flat-rate model, at least in its pure form, proved to tend toward waste of resources, unfairness among users, and revenue losses for ISPs. Therefore, an "alternative" ISP has been proposed, offering differentiated services with dynamic volume-based pricing and suitable feedback mechanisms to inform the users on their own patterns of consumption. These project results have become a stimulus for efforts to shift Internet pricing schemes away from the simple flat rate model.

4.3.1.4 Lightweight Policing and Charging The main assumption of this work is that a multiservice packet network can be achieved by adding classification and scheduling to routers, but not policing [12]. Therefore, a lightweight, packet-granularity charging system has been investigated emulating a highly open policing function that is separated from the data path. The number of charging functions required depends on the customer's selection of services and is operated on the customer's platform. The proposed architecture includes a set of functions distributed to customers, which can include metering, accounting, and billing as well as per-packet or per-flow policing and admission control. The proposal concludes that lower cost is achieved through simplicity without sacrificing commercial flexibility or security. Different reasons for charging, such as interprovider charging, multicast charging, and open bundling of network charges with those for higher class services, are considered within the same design. A discussion of value flows in such an environment can be found in Briscoe [11].

4.3.1.5 Edge Pricing The fundamental idea of edge pricing concerns pricing decisions, which are made at the edge of an ISP locally [26, 108]. Therefore, no standardized pricing models are necessary, since ISP interconnections involve bilateral agreements only, e.g., in DiffServ this will be a major part of SLAs between ISPs. This decentralized approach allows for different edges of the Internet to support differing pricing models at the same time. Furthermore, edge pricing's characteristic of transparency enables ISPs to use, adapt, and evolve pricing policies independently. In a basic approach, customers define the maximal total price they are willing to pay as a sender or a receiver of data, respectively, as well as an upper limit for the maximal number of hops. This charging information can be transmitted as part of a signaling protocol, e.g., in the RSVP header [36, 62].

4.3.1.6 Congestion Pricing Congestion is a problem for today's packet-switched networks, particularly the current Internet. The central question is: Are there means to possibly encourage users to cooperate with the network and at the same time allow for differential QoS provisioning in the network? Emerging ways of controlling network resources use traffic-control mechanisms in terms of congestion pricing approaches to achieve differential QoS. If there is no congestion, the price for utilizing the network is zero or at a minimal value, but it increases with increasing congestion. This scheme is incentive compatible, as it gives users the choice of backing off when the network is overloaded, and allows those willing to pay more to get more. In this case, feedback signals from the network to the customer are related to shadow prices and the marginal cost of congestion. All customers are free to react as they chose, but will have to pay charges when resources are congested. Such behavior between users and the network can be considered self-management [63]. In addition, this model complies with gaming theory, since users play a game with the network [66]. Within the Internet, different algorithms can be linked to TCP or flow control schemes. In addition, the Explicit Congestion Notification (ECN) proposal [99] offers a feedback mechanism to users, which is applied as one possible example in the M3I project [85] as well as proposed in [67, 68]. Finally, the congestion pricing approach has an impact on networking infrastructure investments [46].

4.3.1.7 Cumulus Pricing Scheme Pricing models form a scalable approach for network management. This alternative view on pricing as an economic traffic-control scheme is based on eliciting user information about expected usage patterns. Based on this information, ISPs are enabled to optimize, e.g., network configuration or admission control, with respect to objectives such as the maximization of utilization or revenue. The cumulus pricing scheme (CPS) defines a flat-rate scheme (but rates may vary over long time-scales). It provides a feedback mechanism to bring market forces into play (where this feedback is not an immediate one, but requires the accumulation of a sufficient number of discrete "flags" indicating user behavior), and it allows for a wide flexibility in terms of the technical prerequisites, especially concerning the measuring and accounting mechanisms of the required data records [121]. CPS has been developed with respect to three main dimensions—(1) customer-oriented, (2) ISP economic, and (3) ISP technical—which define the Internet pricing "feasibility problem," i.e., an optimal trade-off between the ISP's technical, the ISP's economic, and customer-oriented requirements [101]. The fundamental decision between static and dynamic schemes touches customers' desires concerning price stability, e.g., highly fluctuating auctions, whereas orienting a pricing scheme strictly according to the forces of the market induces technical infeasibility. The key to the solution proposed lies in building a contract between customer and ISP upon suitable information about the expected usage pattern of the service and influencing the actual customer behavior by a new type of feedback mechanism that is specific in terms of its relation to different time scales. Measurements take place over a short time period and allow evidence about user behavior on a medium time scale. This evidence is expressed in terms of discrete flags, so-called cumulus points, yet not triggering a reaction by themselves, but only as a result of their accumulation over a long time period. Reichl and Stiller [101] propose a framework for tariff descriptions that identifies that existing tariffs seldomly consider a time-scale mapping. It demonstrates that the design of CPS eventually even solves the feasibility problem mentioned.

4.3.1.8 Charging for Premium IP Services The ACTS project SUSIE is focused on the examination, design, implementation, and testing of solutions for charging and accounting of QoS-enhanced IP services. Driven by the need to support a wide range of possible tariff schemes and prices, the proposed charging and accounting architecture shall support a flexible set of metering solutions, the exchange of accounting information between providers and customers, and a means to provision tariffing information. Due to the fact that a wide range of application requirements and user value are in place, a tariff- dependent service selection is desirable. A charging and accounting architecture has been developed and applied to premium IP services [17]. In particular, a tariff formula language defines the description options for the charging formulas and utility curves. The Charging and Information Protocol allows for user information on current tariffs in a push and pull mode. The tariff- dependent service selection is supported by a utility-price optimizer based on user preferences, offered service classes, and applied tariffs.

4.3.2 Metering, Accounting, and Mediation Technology

A major input for CSSs is defined by the set of parameters and their values, which are measurable from the underlying networking infrastructure, including hardware and equipment, and software and communication protocols that are in place. While the lowest-layer task is defined as metering (cf. Figure 4.2 and Section 4.2.4.3), two example technology choices are mentioned at this stage.

One example is the NeTraMet [15, 13] as the first implementation of the Internet AAA architecture [80]. The NeTraMet implementation defines a meter that runs on various platforms. It collects IP packets and byte counts for traffic flows that are defined by their addresses. Addresses can be Ethernet addresses, various protocol addresses, e.g., IP or IPX, transport address information, such as IP port numbers, or combinations thereof. While the traffic flows to be observed are specified by a set of rules that are downloaded to NeTraMet by a manager, traffic flow data are collected via SNMP by a collector. Within NeTraMet's newest version, DSCPs and IPv6 implementations are supported.

A second example is given by Cisco's NetFlow product, which provides usage-based data collection to be integrated into a so-called three-tier billing architecture, where instrumentation, mediation, and billing are combined [22]. Within the instrumentation level, raw data measured at devices are collected in different data formats. The following level, "mediation," is concerned with the service and information levels of data, which begin and end the communication sequences. It filters, aggregates, and correlates raw data to yield different views on current network activity, and enforces business rules to package data into the form that is common to billing systems. The nonvolatile collection and storage platform is based technically on an accounting adjunct processor, which is necessary to relieve router memory from vast amounts of accounting data. Finally, a third-party billing system collects these mediated and packaged data. It matches these formatted data with rating systems, prices the resource usage, and outputs the record details to existing invoicing systems.

A third example covers Hewlett-Packard's Smart Internet Usage (SIU) product [111], which determines a distributed usage management system with open interfaces to a wide range of applications and data sources. Its distributed architecture was designed to scale according to growing demands with respect to numbers of services and customers. SIU collects, aggregates, and correlates data obtained from the technical infrastructure, including hardware and software components, relying on a metering tool that is in place. The configurable presentation and transformation of usage data provide the basic step to obtain information, which can be utilized for any sort of charging approach for the various services an ISP might envision.

Among various other technologies, several mediation, accounting, and billing systems exist in the market. A close similarity to various data format definitions can be observed (cf. Section 4.4.5).

4.3.3 Traditional Telecommunication Billing

Charging is not a new area, since related areas of significance for charging for data communications exist. This is due to, e.g., the length of time it took the telephone network to determine traditional telecommunications billing. Before the days of deregulation, the handling of interconnections between national telecommunication operators was based on dealing with telephone calls. As standardized in the ITU-T D Recommendations series [54], Accounting-rate systems (ARS) were devised to simplify operator-to-operator interfaces and sharing of revenues between originating and terminating operators. The accounting-revenue procedure division defines how the accounting revenue is shared between terminal administrations, and, as appropriate, between administrations of transit countries [54]. It includes the accounting-rate share as part of the accounting rate, corresponding to facilities made available in each country. This share is fixed by agreement among operators.

Thus, the ARS had to be negotiated between every separate interconnect partner. The advantage of this system is based on the fact that only bilateral agreements had to be negotiated. However, its drawback came with the liberalization of the telecommunications industry, since an ARS does not reflect the real costs of service provisioning, traffic analysis between operators is limited, and data interrogation in cases of disputes is almost impossible [74]. In addition, on the technical side, all charging capabilities (including interfaces to billing systems) were hardwired down to the switching fabric, where all data collected were tied inextricably to devices, such as line cards, crossbars, or ATM switches [22, 103]. This tight model does not offer a single degree of flexibility, particularly for different interoperable equipment, as is required for today's and tomorrow's Internet networking devices. Finally, Internet charging and SLA need to integrate multiservice agreements between providers. While telephony bearer services basically consist of a single end-to-end service class, the telephony circuit of 64-kbit/s equivalents and some predetermined delay, today's Internet services show a much larger variety of technical parameters and resource usage within the network itself. The main difference in telephony compared to the Internet is visible in the set of fixed QoS characteristics per telephone connection. The style of packet-switched networks shows major technical differences and requires different handling of charging. Thus, new concepts for Internet transport, services, and content charging are essential to develop a similar level of operation and reliability for Internet CSS.

Work on charging in the ATM environment shows commonalities, but is still significantly different from the Internet, due to at least the virtual connection principle applied. For ATM-based B-ISDN the tasks of accounting, charge calculation, and billing are required to complete a commercial services offer of integrated services. ATM charging can be expected to serve as an embracing network functionality capable of supporting the needs of service providers, retail customers, value-added service providers, and other businesses in a differentiated services market situation. VPNs offer the possibility of satisfying special enterprise needs on a closed networking environment, where an ATM-based solution is highly qualified to obtain the bandwidth and guaranteed QoS required. It guarantees maximum flexibility for a variety of different applications requiring multimedia services, it eases management overhead, and it reduces costs of operating the VPN. However, ATM-based Intranets are only affordable for medium and larger enterprises, because tariffing structures slightly favor high-volume customers.

The ATM view of accounting, charging, and billing has been preliminarily defined in [32]. The basic charging for ATM, termed "three tier charging" [72], includes the setup fee, all duration fees, and all volume fees. In contrast, two basic components of ATM tariffs are commonly identified by ETSI [31]. Charges of an access component are typically fixed per installation, and they remain constant over billing periods, which does not require any on-line measurements. However, this scheme should still allow providers to compensate and recover costs for required facilities of a service subscriber to access a service or services, e.g., facilities specifically provided to that service subscriber. In addition, these charges are independent of the utilization and are related mainly to the type of access, such as capacity provided, maintenance, or redundancy. Charges of the utilization component should be in accordance with the service requested by the service subscriber. The measurement of this utilization component usually has to be carried out. Most ATM utilization charging schemes are based on saving parameters received through the ATM signaling, e.g., including traffic contract, source and destination addresses, counting ATM cells during the ongoing call, and saving the setup time and duration of the call. Since ATM technology in the wide-area environment used to be controlled by Post, Telephone, Telegraph (PTTs), tariffing schemes defined initial approaches for public ATM networks. Legacy ATM networks still rely on conventional tariff models as applied to telephone services. Current implementations on ATM pricing models are based either on a flat rate, as for legacy leased-line tariffs, or on a two-part pricing scheme that includes a monthly access and a usage-based fee, as it has been for legacy switched-circuits tariffs. Research results on pricing ATM services have been obtained, e.g., by several ACTS projects, such as Charging and Accounting Schemes in Multi-Service ATM Networks (CA$HMAN) [112] and Contract Negotiation and Charging in ATM Networks (CANCAN) [72], as well as another Swiss project [103]. These proposals suggested different ATM pricing models to take into account various service classes offered by ATM.

4.4 INTERNET SERVICES AND TECHNOLOGY CHOICES

The different technologies for Internet services can be distinguished by their trade-off between features and complexity. While a larger set of features, such as QoS support or advanced security functionality, has the potential for a better services differentiation, the complexity of this particular technology increases. However, in order to combine these technical service differentiation methods with components for adding economic control mechanisms, these different technologies require adequate interfaces for dealing with charging-relevant information and tasks. For that reason, major QoS methods and service differentiation methods are described, Internet technology choices are discussed, and interprovider agreements, as well as suitable accounting and charging data formats, are presented.

4.4.1 Quality of Service Methods

Three major groups of QoS methods are concerned with the control of data flows in a network [113], the Internet in particular. The first group deals in the shortest amount of time with per-packet or per-flow issues, once the flow has been set up or data are transmitted. The second group handles procedures in medium time periods to signal QoS requirements to appropriate network elements. Finally, the third group provides in longest amount of time to engineer multiple traffic streams and networks as a whole.

Based on Karsten *et al.* [61], these different groups include in particular the following methods. Packet scheduling impacts the QoS experienced by a packet, since the queuing delay constitutes a portion of the total end-to-end transfer delay. Therefore, scheduling is concerned with the decision of which packet to send next on a given link. Examples include FIFO, WFQ and RED. Traffic policing and shaping deal with the task shaping traffic to either a negotiated or advertised level of service at the edges of networks or between network elements. Example mechanisms include leaky or token-bucket traffic shapers in order to ensure a controllable network load. Finally, adaptiveness determines the capability of end-systems to react to congestion in the network by evaluating signals from the network. These signals can be implicit, e.g., loss of packets, or explicit, e.g., by an ECN [99]. Dynamic and congestion-based pricing of network services are also a form of network signals proposed for managing QoS (cf. Section 4.3.1.6).

Signaling and admission control are a major representative of the second group of QoS methods. This integrated set of mechanisms builds on a session or call paradigm, where users of the network signal their requirements explicitly and the network consults local admission-control modules to accept or reject those requests. While per-flow admission control allows for statistical QoS guarantees on a per-packet basis only, admission-control procedures are either parameter- or measurement-based. An example of a proposed signaling protocol for the RSVP [8], while another one is an inter-bandwidth broker protocol [115].

Finally, long-term methods include traffic engineering, and are concerned with the distribution of traffic for a given network by mechanisms such as explicit or QoS-based routing schemes. Network design and engineering, called provisioning as well, deal with the set up and maintenance of network equipment and the design of particular instances of QoS methods based on experience, expert knowledge, heuristics, or formal optimization methods.

4.4.2 Service Differentiation Methods

Until recently, the Internet has performed on a noncommercial basis and service differentiation has not been necessary. However, with commercialization of the Internet, considered to be a commercially operated networking infrastructure and its offered services, this point of view changes. In particular, once an end-customer has to choose from, say, two different service classes, a best effort one and another one delivering some sort of bandwidth guarantees, the purely technical solution of providing these classes is not sufficient any more. The reason for this is obscured by the greedy nature of almost every, certainly the majority of, end-customers—they will always choose that service class with the best QoS. Of course, if this is the case, the service class with less QoS will become obsolete, since it is not used. In turn, users encounter similar problems as before within the better class of service due to its potential for being heavily congested. This situation will remain unchanged as long as no financial incentives for choosing a service class that is perfectly suited for the end-customer's needs are provided by the Internet.

Today's Internet does not offer service differentiation mechanisms, since the best effort type of service still dominates. In addition, the basic Internet protocol is defined by the IP, which is currently used in its Version 4 and does not provide any service class differentiation features besides the ToS field. Nevertheless, this field is only optionally used on a wide scale within IPv4. Enhancements and changes, including a flow label field [96], determine the new version IPv6 being prepared by the IETF.

An important way to make an effective service differentiation is the definition of a QoS model for services offered. According to Bradner [9], the following macroscopic facets exist. The scope defines the logical distance over which a service model is provided. The granularity identifies the smallest service unit, which is treated individually by the service model. The time period specifies the granularity in time for which services are being provided. And the control model formulates those entities, which perform the control over the network and the traffic. They can be located exclusively in the network or in end-systems, with a continuum of hybrid forms in between. However, since two distinct Internet protocol architectures are used today as approaches for a service differentiated Internet, distinct QoS models and their corresponding technology choices exist.

Another challenge is due to the fact that the application of a differentiated pricing model for differentiated Internet services gives network operators a substantial gain in efficiency. It has been shown theoretically and by simulation that this increase in efficiency depends on the traffic characteristics of applications [26, 106]. Another important factor is the degree of competition allowed by regulators. Global Internet services usually cross many different provider networks, and providers overlap each other geographically. This development not only increases competition, but also increases the choice of service offerings and the efficiency of network operation. Therefore, provisioning of differentiated services requires charging and accounting services in the Internet. The offer of multiple service classes and the precision used in appropriate pricing models depends on the way packet-based communication is handled. For example, in the medium time-scale phase of QoS methods, existing signaling protocols like RSVP [8] can provide the basis for collecting charging data [117, 59].

4.4.3 Internet Technology Choices

Starting from the presentation of the best effort model with overprovisioning, continuing to a price-based best effort, the integrated services and the differentiated services models, a combination of these approaches is discussed.

4.4.3.1 Best Effort Assuming that overprovisioning of network resources, basically bandwidth, is both possible and sufficient to sustain the single-service nature of the current Internet, an end-to-end communication is possible, where all control entities are located in end-systems [61]. Therefore, no state exists in the network and all traffic is treated at the same granularity with longer time periods, essentially equal to the length of a capacity planning cycle. The QoS method applied to this model is the network design and engineering model to provide for a superabundance of network resources. In periods of resource scarcity this model relies on the adaptiveness of end-systems. Based on another assumption that pure overprovisioning is not sufficient without an additional means of signaling besides packet loss, this additional signal is a per-packet price, depending on the internal state of the network, e.g., its congestion level. The time period of interest for this model is related to the period of price announcements and the ability to set them from the provider's side. Again, network design and engineering methods are applied, but end-systems' or users' sensitivity to pricing signals has be estimated.

Neither of these two models has been proven to be optimal or totally unsuited. In addition, the combination of technical means and an economic-driven control strategy integrates a set of not yet fully understood factors (such as packet-switched, connectionless networking technology, and the extremely high pace of network and customer growth) of a new and rapidly emerging Internet services market.

4.4.3.2 Integrated Services Internet IntServ defines a framework in support of unidirectional end-to-end flows [7]. These flows can request certain QoS and can use a controlled-load service [129] or a guaranteed service [109]. As shown in Figure 4.5, available service classes in IntServ distinguish between best effort and guaranteed services.

However in every case, flows need to establish a context between the sender, the receiver, and intermediate nodes. Therefore, RSVP [8] has been defined as a protocol for reserving network resources for single flows, which are specified by the sender and receiver in terms of desired traffic classes, including bandwidth, delay, and loss characteristics. RSVP relies on the existence of an admission control, resource allocation, and packet-forwarding mechanism in each router to ensure that the requested QoS parameters can be guaranteed. Advantageous features of the IntServ and RSVP approach encompass a receiver-driven QoS specification, the support of multicast traffic and its merging for reservations, and the soft state approach for maintaining the context data of a flow. However, the support on a per-flow basis shows a linear scalability with respect to the number of flows and states to be kept in large backbone routers [59]. The per-flow granularity imposes overhead that may not be necessary for a certain number of situations.

4.4.3.3 Differentiated Services Internet Due to these assumed scalability problems when handling single flows, a different framework was developed. Instead of treating a single flow as the entity of interest, the DiffServ handles Internet traffic based on the notion of aggregated, unidirectional flows and fixed numbers of service levels in terms of service profiles [4]. This approach minimizes the state to be maintained in the routers. In addition, this is supported by a domain concept, where a group of routers implements a similar number of service levels and the appropriate set of policies. This DiffServ domain is defined by a fixed boundary, consisting of ingress and egress routers. However, traffic

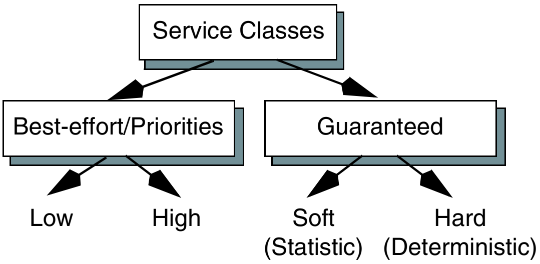

Figure 4.5 Service classes in IntServ (based on Braden et al. [7]).

traversing such a DiffServ domain is required to be marked. This marking happens on a per-IP-packet basis at the ingress routers and utilizes the DS field in an IP packet [87]. This DS field replaces the ToS field from the IPv4 protocol and accepts the definition of PHB, which in turn determine the service level such a packet will be treated by. Once the DSCP as part of the DS field has been set, the packet travels through the DS domain, and are treated equally in every interior router [89].

Therefore, since single end-to-end flows are bundled to aggregated flows with a similar behavior within a DS domain, the DS approach requires less overhead. However, the need to mark IP packets at DS borders remains. In addition, a longer termed service contract may be required between different DS domains, since a certain service level may be required. This type of flow aggregation, in conjunction with service guarantees, requires some sort of admission control, since an overutilization can lead to service degradations. SLAs are regularly set up between interconnecting ISPs in order to maintain the desired service level for the aggregated flows. An initial SLA needs to be set up between interconnected ISPs before any service is exchanged. SLAs can also be adjusted dynamically. Further details on SLAs are provided in Section 4.4.4 below.

4.4.3.4 Comparison and Combination (IntServ over DiffServ) As presented earlier, IntServ as well as DiffServ have a number of advantages and drawbacks. Based on six classification criteria, Table 4.2 summarizes these differences for the IntServ, DiffServ, and best effort traffic architectures of the Internet.

One possible combination of IntServ and DiffServ advantages could apply to IntServ in the access and DiffServ in the core network. Local area networks (LANs) tend to show an overprovisioning of bandwidth, which does not require sophisticated resource management and signaling, if certain topology and traffic considerations are taken into account. The access network, however, utilizes RSVP to signal flow requirements from LAN-based hosts to the core's edge routers. They perform a mapping of these requirements onto particular flow aggregation types available in the DiffServ core represented by a dedicated SLA. Since core routers perform traffic forwarding based purely on PHBs, they are able to cope with many aggregated flows. Therefore, only edge routers need to keep the state of flows from their local domain.

4.4.4 Interprovider Agreements

Interprovider agreements are required between ISPs, to define terms and conditions of services for traffic exchange. Such an agreement represents a contract-like relationship,

Table 4.2 Comparison of Internet Network Architectures

Criteria	Best Effort	IntServ	DiffServ
QoS Guarantees	No	Per data stream	Aggregated
Configuration	none	Per session (dynamic)	Long-term (static)
Zone	Entire network	End-to-end	Domain-oriented
State information	None	Per data stream (in router)	(None, in BB, in edge router)
Protocols	None	Signaling (RSVP)	Bit field (DS byte)
Status	Operational	Matured	Being worked on

indicating all characteristics of the service and its implied financial settlements. Traditionally, within the Internet, ISPs engage in interconnection agreements to assure each other pervasive Internet connectivity. This allows ISPs to exchange traffic at an interconnection point, either with or without financial settlements [1]. Traditionally, these agreements do not consider QoS-related issues. More recently, SLAs are used as a means of establishing a provider/customer type of relationship between two providers. While the exact content of an SLA often depends on the specific business and technology context, the core of an SLA always includes a description of the traffic covered by the SLA and the service level that is to be applied. SLAs provide the contractual envelope for QoS-based assurances. As discussed in Kneer et al. [71] and Stiller et al. [120], the following types of SLAs can be described.

4.4.4.1 DiffServ SLAs The DiffServ approach is based on the notion of "network domains," which can be operated by different ISPs. In a pure DiffServ world, both access and core domains would use DiffServ technology to transfer data within their domains as well as between domains. In order to indicate service commitments between domains, SLAs are employed. However, the IETF working group on DiffServ does not consider SLAs as their area of interest.

DiffServ SLAs are defined on the contract level, according to the type of underlying physical network connectivity. According to the contemplated DiffServ support for multiple service classes, an SLA also specifies a selected service class. Thus, a DiffServ SLA includes at least:

- Description of the aggregated flow to which the SLA is applicable.
- Corresponding throughput values.
- Corresponding service class, e.g., expedited/assured forwarding, determining QoS, delay or loss characteristics.

In principle, DiffServ SLAs can be defined at any granularity level, including the level of application flows. As DiffServ pursues the goal of high scalability, only a limited set of SLAs is likely to exist between any two domains, so that explicit support of individual flows through DiffServ SLAs will be an exception. Further work on SLAs and service-level specifications can be found in Salsano et al. [104].

4.4.4.2 Commercial SLAs SLAs have emerged in the commercial domain as a result of increasing customer demand to understand what kind of service they can expect from an ISP. Such SLAs are typically tied to the provision of a network service on a long-term basis and service provision, which includes both the installation of physical equipment, e.g., routers or access lines, and the provision of a data transporting service, e.g., based on IP protocols.

In order to capture QoS aspects, SLAs include parameters similar to those considered in the DiffServ case. As an example, the SLA employed by UUNET foresees the following attributes:

- Throughput offered to customers, e.g., T1
- Round-trip latency across the provider's network
- Availability of the service

- Outage notification duration
- Duration between order and installation
- Reimbursement procedure in case of noncompliance

These SLA approaches tend to serve the same purpose: establish longer-term service relationships between adjacent ISPs, and provide assurances for traffic aggregates exchanged between them. This avoids the overhead implied by a high number of concurrent SLAs and frequent changes in terms of agreements.

4.4.4.3 Flow-based SLAs In principle, SLAs can be applied at the service interface between ISPs, providing IP access and their end-customers. However, given that in access domains per-flow handling of traffic is a viable option, a different approach is feasible, offering superior flexibility in applying charging schemes to QoS-based traffic.

First, it is likely that QoS support within the Internet will not be pervasive. Whenever a QoS-based flow is requested by an application, the availability of such support has to be established, depending on source and destination endpoints as well as the sequence of ISPs involved. Providing SLAs dynamically on a per-flow basis allows such dependencies to be considered and automatically adapt to increasing Internet support for QoS. Second, considering QoS automatically implies a strong differentiation among Internet services. QoS can be provided at multiple levels, e.g., to support various audio qualities. Consequently, there is a need for differential and QoS-dependent pricing in order to prevent users from making use of the best QoS level only. Similarly, applying differential pricing is required in order to reflect the communication path, i.e., ISPs traversed, including source and destination locations. Both aspects lead to service costs, which can be established only if characteristics of a requested flow are known, i.e., on a per-flow basis. Third, there is an ongoing discussion on the pricing scheme that should be applied to Internet traffic. Dynamic pricing of offered services based on the current level of network usage was shown to significantly improve service characteristics a network can provide, for instance, reduce congestion and smooth traffic. Assuming such an approach, both the implied QoS level and the price provided are to be established dynamically for each new flow.

Providing flow-based SLAs captures all mentioned issues. Such SLAs are in contrast to the ones considered earlier. They directly concern application-level flows and not aggregations of traffic. Furthermore, they are likely to be set up for the required duration of communications only. Flow-based SLAs are in line with the service model proposed by IntServ, as far as the end-customer's point of view is concerned, since they want to have selectable QoS on a per-flow basis. In contrast, the arguments mentioned earlier are driven by economic considerations. ISPs want to provide incentives for end-customers to make use of more or fewer resources in the network and, in the case of dynamic pricing, ISPs want to consider the availability of free resources when setting prices. These aspects are best considered in the context of individual demand units, i.e., flows for end-customers or aggregates for enterprises.

4.4.5 Standardized Data Formats

CDRs [58], sometimes termed call detail reporting, call data records, or station message detail record (SMDR), are the most commonly known records for call-specific data, originating from telephony-based telecommunication systems and developed over many years in an environment with quite static services portfolios. Such a record defines the fundamental unit of data to be exchanged in the circuit-switched voice world. It contains data about each call made, e.g., dialed digits, the phone number dialed from, call direction, service type, associated inverse multiplexing session and port, date, time, off-hook time, on-hook time, how long the call lasted, and a circuit identifier. Virtually all telephony switches, private branch exchanges (PBXs), and ATM switches [103] produce CDRs. However, each switch product tends to produce CDRs in different formats, which means that data fields of each record may appear in a different order from one switch to another. Therefore, performance-intensive software needs to convert various CDR formats into a standard format usable by a charging system. Because the network provider can charge for bandwidth on an as-used basis, the CDR can be used to understand and manage bandwidth usage. However, they are different in that an SMDR is focused on the station (terminal) and the CDR is focused on the call itself. Therefore, the two terms should not be used interchangeably, since the formats of the records will be different. Usually, a single device, say a PBX, will produce one or the other, but not both.

To cope with networking characteristics of the Internet, mainly the packet-switched characteristic compared to the telephone network's circuit-switched system, a corresponding data specification is required. In addition, the Internet market trend to develop and deploy new services frequently raises a second dimension of complexity for an "Internet CDR." Therefore, the initiative "IPDR.org" decided to develop a basic framework for a usage specification, called IPDR, which allows different companies to develop dedicated code within the framework, support interoperability, and the usefulness of the specification [21]. It refers (1) to a functional operation, where an NDM function collects data from devices and services in a provider's network, and (2) to usage, the type of data, which shows an open, extensible, and flexible record format (the IPDR record) for exchanging usage information of essential parameters of IP-related transactions. A repository for defined IPDRs is envisioned, including the variety of services, such as email services, as well as real-time services. These definitions will form the essential elements of data exchange between network elements, OSSs, CSSs, and BSSs. The framework will provide the foundation for the development of open, carrier-grade Internet services enabling next-generation IP networks to operate efficiently and cost effectively.

The recently published informal RFC on accounting attributes and record formats [14] summarizes existing IETF and ITU-T work and discusses advantages as well as drawbacks in close detail. With respect to the Internet, the remote-access dial-in user service (RADIUS) accounting record (RAC), the DIAMETER attributes, and real-time flow measurement (RTFM) architecture are important. While RADIUS, among others, deals with start, stop, and activity data including various accounting, tunneling, and general attributes, DIAMETER being part of the AAA architecture (cf. Section 4.3.2) inherits all of them and defines a secure protocol to transfer these accounting attributes. Finally, the RTFM architecture supports flow measurements via RTFM meter readers, which read data from MIB to be stored in RTFM attributes, such as source and destination information, as well as packets and byte counts.

Additional data formats are available, but mainly with respect to a particular protocol or application. The domain name system (DNS) and the dynamic-host configuration protocol (DHCP) maintain customer profile data, which form a type of standardized data format. In addition, the lightweight directory-access protocol (LDAP) offers mechanisms with transfer capabilities for customer-profile data. However, these data formats are not generally used for the purpose of accounting.

4.4.6 Electronic Payment Systems

Aside from communication protocol relevant issues and Internet networking technology choices, a particular area of interest arises. With respect to fully integrated electronic service delivery, electronic payments for various kinds of transport and content services determine the clear necessity of pico- or micropayments. Since existing electronic payment systems are not well suited for this task, solutions have to be researched, including efficient cryptographic protocols for secure transmission of payments [98]. To implement a complete billing system successfully, legal contracts are needed that are based today mostly on verification of customers' identity by letter or telephone. This is due to the absence of efficient electronic authentication mechanisms and certification authorities. Once a contract has been established, traditional invoicing or credit card billing is the most popular way to collect money. However, electronic payment systems that provide anonymity [19] and/or small amounts [79] are still not accepted with ISPs. Today, it is not clear, whether micropayments or anonymous e-cash provide a real advantage to service providers offering usage-based pricing for their services.

4.5 PRICING MODELS

The increasing deregulation of the telecommunications market and the emerging business orientation of Internet services drive the need for appropriate pricing models for packet-based communications, which are independent of regulated aspects. Well-known and widely accepted pricing models for communication networks offering a single network service, e.g., telephony or X.25, are provider-centric, i.e., they are set to fixed values and reissued, whenever provider costs or regulations change. However, in an increasingly competitive environment, this approach of changing models is too slow. Furthermore, the deregulation opens the field of pricing in particular communication services within an open market approach [86].

Besides these changes in the overall networking environment and market situation, the selection of a set of suitable pricing models for the Internet remains an open problem. Many projects covering charging functionality on the network level, including services and sometimes content, intend to achieve complete independence from pricing models. Therefore, this section investigates pricing model constituents and then discusses existing models.

4.5.1 Pricing Model Constituents

In general, the components of pricing include three basic constituents as illustrated in Figure 4.6, each of which may be empty. First, an access fee is usually a time-periodic charge, e.g., weekly or monthly, for using an access link to the network. The price for this link depends, e.g., on the capacity of that link or its length to the provider's point of origin. This constituent does not require any on-line measurements (metering), since the data required for the price model to be defined are of a static nature, based on the installed configuration. However, metering for traditional network management purposes, such as utilization and load balancing, can be performed. Second, a per-call or connection/reservation setup fee may be included. In connection-oriented, circuit-switched networks or connectionless, packet-switched networks with reservations, different mechanisms setting up the connection and the reservation, respectively, can be charged separately. This constituent unavoidably requires on-line metering on a per-call or per-reservation basis, including corre-

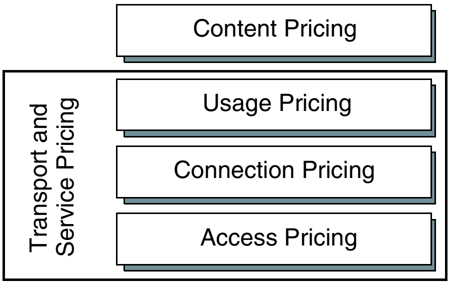

Figure 4.6 Layered components of Internet pricing.

sponding data formats for accounting for this information (cf. Section 4.4.5). Third, a usage fee may be used to charge for data transport and services on time, volume, or any particular QoS basis. This fee reflects the actual resource usage, e.g., users consume capacity. Based on economic principles of marginal cost, market mechanisms, congestion-driven schemes, or marketing policies, details of this usage fee and its corresponding pricing model have to be defined explicitly. Finally, independent of the basic transport and service fees, a content fee may be introduced. Depending on the particular application and its content, this fee may be omitted explicitly (e.g., telephony, fax, e-mail services where the "content" is provided by the user himself), billed separately (e.g., the *Wall Street Journal* on-line edition), or indistinguishably integrated into the transport and services charging (e.g., 1-900 numbers).

The traditional telecommunication services approach, as defined in [55], follows a similar approach, defining the elements "access," "invocation," and "usage," but considering a single service only and neglecting all content issues. While the access shows a subscription form and uniform periodic charge, the actual bill is based on a call or service setup attempt, which is measured in units of uniform charges or successful connections. Finally, the usage element shows the form of the call duration or the volume transmitted, which are measured in time, pulses, packets, or segments as well. Different combinations of these constituents and approaches to pricing of telecommunication services are classified in [40]. For instance, where the service is a single-voice service, traditional voice services show all three transport and services components.

The price constituents in Figure 4.6 are fully or partially reflected in Internet pricing models. An ISP usually used to charge for access and optionally for usage on a connect-time basis or on a flat rate. However, the most important pricing models for Internet services include flat fee, usage-based, reservation-based, volume-based, service class-based, and connect time-based methods, as discussed below.

Edge pricing (cf. Section 4.3.1.5) deals with another aspect of complexity reduction, only in terms of locality, concentrating the distributed nature of pricing decisions by shifting them to the edge of the ISP [108]. This concept is preferred for its simplicity, decoupling complex price negotiations between customers and various ISPs into a series of bilateral ones on different time scales, as well as for its transparence toward the customer.

4.5.2 General Characteristics of Pricing Models

Now that we know about the pricing model constituents, we can see that the targets of pricing are at least twofold. On the one hand, it acts as a means for allowing fair, financially driven resource sharing of services, and it provides an economically driven tool for traffic control functions, such as bandwidth management. On the other hand, pricing determines the approach through which providers to recover costs or increase their revenue. In general, Internet pricing needs to comply with customer demands as well as provider demands, which creates an inherent problem, since the time scales of interest to a single customer are significantly different compared to provider time scales. Therefore, to enable a comparison of particular pricing models for the Internet, time scales for Internet pricing are introduced, properties of pricing models are summarized, and relevant pricing model dimensions are derived.

4.5.2.1 Time Scales Time scales define the major criteria for distributed systems to operate with feedbacks. Since, according to Section 4.2.4.1, charges are derived from prices, and since they reflect a financial feedback to utilization of a service, existing management time scales are, according to Hegering et al. [50]: short-term in minutes, medium-term in hours, and long-term in weeks or months. These scales are extended for charging purposes by an atomic scale for ultrashort times in seconds and below. As illustrated in Table 4.3, intervals and units of measurement show the relevant timing and information to be accounted for. The type of feedback is identified as well.

Applying these time scales onto pricing-controlled activities results in [120]:

1. The atomic monitoring and control level involves sending packets, round-trip times, and managing feedback between sender(s) and receiver(s).
2. The short-term intervention level is concerned with the usual duration of applications like file transfer, video conferencing, or IP phone calls. The accounting and metering tasks are closely related to these activities.
3. The medium-term service provisioning level performs billing actions and depends strongly on the usual human lifestyle habits of humans, e.g., monthly payments of rents, phone charges, or newspaper bills.
4. The long-term business/strategic level in this context determines the duration of contracts between customers and ISPs, which usually varies from several months to years. Note that contracts between ISPs may be shorter.

Therefore, as depicted in Figure 4.7, the proposed charging system operates as a management system that is capable of supporting various pricing schemes. Depending on the specific pricing model applied, different time-scales are effected. Models include usage-based and congestion pricing as a means of mediating the current network utilization, or time-of-day pricing, which is part of the strategic level, since different business models

Table 4.3 Time Scales, Measurement, and Feedback Content [121]

Time-scale Name	Measurement Intervals	Measurement Units	Feedback Content
Atomic	Milliseconds, round trip-times	Packets	Communication-relevant data
Short-term	Minutes	Flows/sessions	Application data
Medium-term	Hours/days	Billing periods	Billing data
Long-term	Weeks/months	Contract periods	Contract data

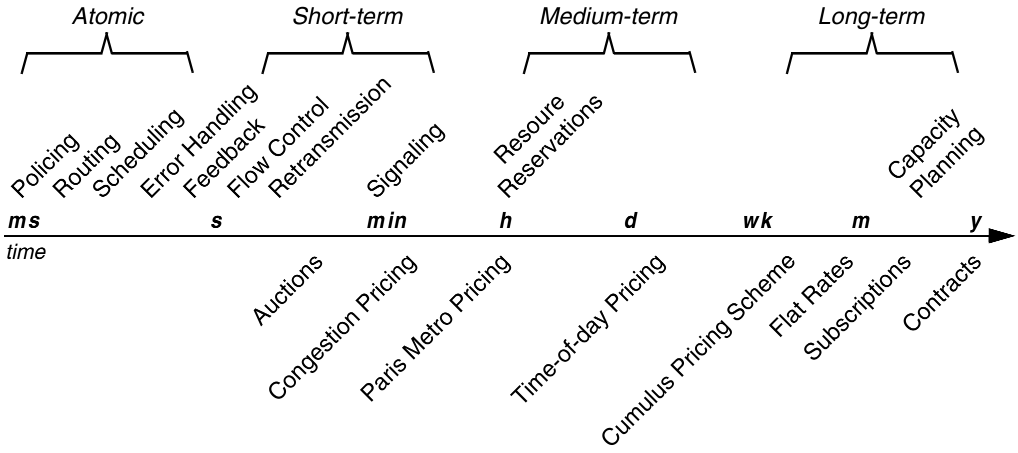

Figure 4.7 Types of feedback and relation to pricing.

are developed for different user demand and user segments.[5] An important result from this mapping of management time scales and their mechanisms in place on pricing time scales is the fact that atomic time-scale mechanisms are technically in operation, just not in an economic model. As described in the following sections, this dilemma of operating a network in smaller time scales than the customer is able to respond to single events (feedback), has been termed the "Feasibility Problem" of Internet Pricing (cf. Section 4.3.1.7 and [101]).

4.5.2.2 Pricing Model Properties Pricing models determine the price of a communication service or application referred to. On the one hand, the price for, e.g., data transport, a service, a service invocation, a certain quality level of the service, or in general for a unit service, depends on costs, demand, and marketing considerations. They allow for service providers to recover costs or maximize their revenues. On the other hand, prices provide feedback signals to users and influence demand and usage.

Therefore, as presented in [37] and extended here, a set of pricing model properties are considered relevant, mainly based on the provider's point of view in the beginning and to the customer's point of the view afterwards (cf. Section 4.2.2). These properties encompass:

- High probability of cost recovery
- Competitiveness of prices
- Encouragement (discouragement) of customers
- Implementation and usage costs
- Comprehensibility
- Controllability, auditability, and trust
- Predictability
- Responsiveness
- Fairness

[5] For technical terms, Figure 4.7 is based on [65] extended by pricing approaches, and completed by time scales.

From the provider's point of view, to operate a network economically and efficiently, a minimal degree of cost recovery is essential and the production of revenues is intended by an applied pricing model. In addition, prices set by the pricing model shall be competitive to remain in the corresponding services market. Pricing of services needs to include incentives for customers to reward a technological, efficient use of network services and to discourage the abundant waste of network resources. Finally, the investment in the implementation of appropriate metering and accounting equipment shall be minimized.

Turning to the customer's point of view, the use of a pricing model and its necessary equipment needs to show a low usage cost limit. Comprehensibility illustrates the ease of understanding a pricing model and the simplicity of the model, particularly with respect to the tariff applied, important parameters used, and the suitability for a given communication scenario. This covers the quality of the service that has to be provided at a minimal level, and this level needs to be related to the price setting. In addition, mainly as a subjective point of view the perceived quality at the service's interface (which is not measurable today) will become equally even more important. The less a user has to think about an applied special pricing model, the better. Pricing models need a certain degree of controllability. On the one hand, this assumes that due to a user's basic amount of trust in the service provider, a price for a utilized service has to be calculable based on locally accessible information. On the other hand, the service provider has to establish means and mechanisms to ensure that the application of a published price on a utilized service remains controllable, auditable, and supervisory.

While private access to Internet services may show a predictable price of the service (flat fee) than a predictable QoS (guaranteed service with time-dependent pricing), commercial users tend to value a service that depends on the dedicated task to be performed, not the application currently used. ISPs have gained practical experience in this matter and found that residential customers prefer the option of predictable prices rather than predictable QoS (cf. [69, page 18]). However, due to the change to different and new Internet services, such as multicasting and video conferencing, it is not obvious that these assumptions will hold. In addition, the change of prices, if necessary for dynamic pricing models, should not happen too frequently, as stable prices for a reasonable period of time show a greater predictability. It always must be considered in close relation to the user group and the area of application. Responsiveness defines the feature of a pricing model to be able to inform the user on the current price for the actual service usage. Since users want feedback about charges (their calculation is based on the definitions of the price model) for their service used, it needs to be ensured that this feedback does not interfere with the task being performed currently. Therefore, the selection of an option of a push model, where feedback is given periodically, but not too often, or a pull model, where the feedback is given on request only, is left to the user. Summarizing, predictability and responsiveness for pricing models are very important and by themselves a premium service option. Finally, fairness issues in general were discussed in Section 4.2.3. Certainly, the effect of a positive willingness to pay paved the way for an economic-driven fairness notion. However, this feature is a principal disagreement with the welfare situation, where a "social fair share" of resources is intended.

Any of these properties affects one of the three dimensions of economic and social aspects (efficiency, marketing, user requirements), technical aspects (technology, services classes, parameters), and research aspects (applications, theory).

4.5.2.3 Pricing Model Dimensions Pricing models have been proposed for almost every technical protocol architecture in an attempt to satisfy QoS demands. Thus, different and usually incompatible pricing models exist, reflecting no standardization efforts. However, from the pool of existing pricing models an important distinction must be made on whether prices are set ahead of time as a fixed (static) price or determined and potentially changed as service is provided, showing a variable (dynamic) price. As further distinctions reveal, a basic classification and differentiation has been developed [120].

The classification is based on five attributes and a number of well-defined parameters, all of which are summarized in Table 4.4. These attributes include time, space, quality (i.e., class quality characterization), technological requirements, and volume.

For the "Time" attribute, the following parameter semantics have been applied:

- *Duration* defines the elapsed time between the start and end of service usage, e.g., duration of a video conference.
- *Period* determines the committed length measured in time, which is per se independent, i.e., decoupled, of the service appliance. This commitment is usually set up in advance, e.g., a leasing period.
- *Time of day* defines the sensitivity of service usage to a given time of day. The influence of the time of day may be known in advance by the customer, e.g., weekend tariffs, or they as well may change dynamically, e.g., based on congestion in an auction system.
- *Not applicable* means that the attribute *Time, Space, Class Quality Characterization, Technological Requirement,* or *Volume* is not relevant, e.g., for the time attribute *not applicable* means, that the time has no significance at all, which can be reasonable, e.g., in a volume-based system.

For the attribute "Space" the following parameters are distinguished:

- *Distance* defines the length of the (virtual circuit) from the sender to the receiver, which is passed by messages. Its length in meters is not relevant, but rather how much infrastructure has been used to enforce the service provision.
- In contrast to distance, *route/path,* describes the relevance of where the message flow passes through, i.e., through which, how many, and what kind of routers. The route/path attribute plays in important role when particular associations are made between the chosen circuit and the service.
- The *location, distance,* and *route/path* parameters are not sufficient by themselves to describe all cases that occur for pricing models. Suppose edge-pricing has to be expressed. Saying that the distance and the route/path are not relevant implies a

Table 4.4 Pricing Model Attributes and Parameters [121]

Time	Space	Class Quality Characterization	Technological Requirements	Volume
Duration	Distance	ISP	Flow-based	Linear cumulation
period	Route/path	Customer	Class specified	Nonlinear cumulation
time of day	Location	Self-adjusting	Not applicable	Not applicable
Not applicable	Not applicable	Indifferent		
		Not applicable		

transparent network from the point of view of space. Indeed a transparent network (cloud) does not imply local importance of service provisioning. Thus *location* allows places/entities in the network that have particular importance for the pricing model, e.g., just like edge-pricing, to be considered.

The "Class Quality Characterization" attribute describes the sensitivity of pricing models to be quality classes. It mainly explores how a differentiation of quality is made and who is influencing the selection or creation of quality classes, i.e., the ISP, the customer, or both. It has to be noted that a differentiation of quality does not imply that only a fix number of classes exists.

- *ISP* sets up quality classes. Often the ISP will have a limited set of quality classes, which it may slightly adapt and distribute among customers.
- *Customer* initiates and defines quality class specification, e.g., with a signaling protocol such as RSVP.
- *Self-adjusting.* The class quality may change with network state, where the correction toward the new stable state is performed in a system-inherent manner, e.g., as with the Paris Metro Pricing approach [92].
- *Indifferent,* where no different quality classes are available.

For the attribute "Technological Requirements," the following parameters exist:

- *Flow-based:* the supporting network technology offers a clear technology for maintaining flows within the network, such as with the integrated services architecture.
- *Class-based:* the network supplies a set of discrete classes, where classes are not necessarily associated with particular technologies or QoS commitments.

Finally, the "Volume" attribute defines:

- *Linear cumulation* as the amount of data accumulated linearly, determining that every single data unit measured has the same weight.
- *Nonlinear cumulation,* which covers all other cases, where the volume of a pricing model is taken into account.

Obviously, the combination of all parameters allows for a large number of different pricing models to be identified. It is up to the designers to agree upon the most reasonable ones. As already seen with the parameter *not applicable,* pricing models are not required to be precise on all attributes. In case that a pricing model has to choose just a single parameter per attribute, this approach is inappropriate. Therefore, a supplementary notation is introduced. The two variables, x and X, describe alternatives of the importance of a given parameter on a per-attribute basis:

- x*:* exactly one, but an arbitrary parameter of an attribute needs to be set, e.g., for the attribute *Volume:* x = [linear cumulation | nonlinear cumulation | not applicable].
- X*:* at least one, but an arbitrary number of parameters of the attribute need to be set. This is required, if a combination of parameters is utilized to precisely define the scope of the pricing model, e.g., for the attribute *Space:* X = distance & route/path.

These pricing model dimensions are applied during the process of classifying existing pricing models in Section 4.5.4. However, two distinct classes of pricing models are discussed before we do that.

4.5.3 Static and Dynamic Pricing Models

Static and dynamic pricing models reflect the two general alternatives a provider needs to choose from for its service offers. Since Falkner et al. [34] discuss particular pricing models in detail, both categories of static and dynamic pricing will be evaluated and compared in the following sections. While flat-fee models show a long-term static characteristic, dynamic models may frequently change the price over time or may cover a usage-sensitive component.

4.5.3.1 Flat-Fee Models Static models reflect the fact that no charges for calls or service usage within a specified geographical area are raised. Within the telecommunications industry, a flat fee (comparable to an "all-you-can-eat" buffet offer), or even free local telephony calls, as with some U.S.-based telephone providers, is quite common. Therefore, the adaptation of this model to a new Internet services model seemed to be straightforward. A fixed fee for the IP access is independent of the bandwidth utilized, the QoS perceived or requested, the congestion state of the network, the transmitted information, or the customer's valuation of the service. Actually, the most traditional pricing model that has been implemented for Internet services is a flat rate model, where the customer paid a flat fee for unlimited usage of the service provided.

The major advantage of static models and flat-rate schemes in particular is that charging is easy and simple to apply, since usage-based metering tasks and call timing equipment are avoided, at least for charging purposes. In consequence, accounting tasks basically do not exist, and the charge calculation is reduced to the flat fee, as defined by the pricing and tariff chosen in advance. Therefore, the resulting billing task becomes simple as well, and users know exactly what their bill will be. Billing complaints are avoided. In addition, the predictability of expected charges is deterministic, since the flat fee is known in advance and the user's budget is predictable, which reduces the risk and simplifies financial budgeting. For the provider, the revenue and cash flow estimating is simplified. Consequently, customer and provider need make only a minimal effort.

However, problems with respect to the resource utilization exist with static pricing models. Usage tends to be high, because no per-call charges are due and incentives to save resource usage are missing. Staying on-line for long periods of time from a residential location, even while not using the service at all, blocks the resource of the local phone line and switching equipment, which other users would like to use explicitly. Power users effectively penalize low usage users and could discourage demand. To solve this problem of higher provider costs, the provider could time users out after *n* minutes of inactivity. However, exactly determining *n* is difficult and will result in inefficiencies for long *n*'s and, for short *n*'s, in a larger overhead for new connection setups. Bandwidth assignments are made by time and by price. Even worse, bandwidth assignments are based on the customer's patience and not (social) customer valuation of the service. The major drawback of a flat-rate scheme is, however, that services differentiation is not possible at all, since the access to an ISP's point of presence will not allow for the technical differentiation of packets belonging to different services. In addition, a user differentiation, e.g., business and residential, becomes difficult, since the variations of usage are not reflected. For a profitable continuation of commercial service provisioning, a flat-rate scheme is insufficient, since they lack revenue sources. An adaptation of flat fees on a shorter time scale could be envisioned.

Strict flat-fee pricing has proven to be difficult in practice, for the U.S. as well as Europe [120]. For example, America-on-Line (AOL) and other ISPs changed from a usage-sensitive scheme (9.95 US$ monthly fee, including 5 hours and 2.95 US$ for each additional hour) to this flat-fee pricing scheme (19.95 US$ per month for unlimited access) on December 1, 1996. This led an explosion in demand that AOL initially could not cope with [84], and to blocked dial-up phone lines at the regional Bell operating companies, since they do not charge for local calls. From a customer's point of view, there is no incentive to hang up a dial-up Internet connection when there is no charge per time or volume. Hence, due to lawsuits, AOL was forced to offer extensive refunding to unsatisfied users, but finally coped with the situation and succeeded in becoming an important player in the ISP world.

European ISPs followed a more differentiated pricing model with free hours and a charge for additional hours for using the on-line service. Additionally, local phone calls have a significant price in Europe. However, Breathe Freely, a UK-based ISP, had severe problems with a flat-fee scheme. After introducing in May 2000 an unlimited access scheme for an on–off payment of 50£, the apparent lack of capacity forced the ISP to skim off the most Internet-intensive 1% of its users, because they caused real problems for the rest [82]. Finally, Breathe Freely went bankrupt at the end of 2000.

These examples clearly indicate difficulties in offering free and unmetered access to the Internet. However, without any doubt, flat-rate pricing schemes are still the most popular ones, and, hence, a reason for designing new pricing schemes. An important topic to add, in order to prevent disasters like the ones mentioned, is a defined concept of feedback to customers on their current usage patterns and their compliance to the overall network situation.

The ITU-T defines two additional variants of a flat-rate charging scheme for telephony [56]. A partial flat-rate scheme is one where a specified number of calls or call units can be made at no charge. Usage may be stimulated; however, it will be quite common that users will restrict their service usage to the exact amount of service allowed for free. A message-rate scheme is one where the metered, but untimed, call to or within a geographical region is charged at a fixed amount, independent on their duration. It is argued in [56] that users "can make long duration calls reasonably cheaply. This increases or expedites the requirement for additional equipment [. . .] resulting in additional costs," and continues, "because calls are charged at a common rate per call there may be a cost saving on equipment, there being no need for periodic pulse metering."

4.5.3.2 Overprovisioning Flat fees are quite common, since the concept of overprovisioning is adopted. This identifies the fact that "sufficient" bandwidth is always available. This situation seems to be viable in principle, since small and further decreasing costs per bandwidth deployed in the public networking sector are observed, even though regional differences exist, e.g., between a transatlantic cable and a citywide ring. The major advantage of such an approach can be expressed by the statement: "Larger bandwidth for the same amount of money." However, since human beings are greedy in nature, no natural limit for bandwidth usage can be set. Therefore, even smaller costs are not limited by an upper bound. The major drawback is the fact that even in the case of overprovisioning and the lack of QoS and traffic control mechanisms within the network, no deterministically reliable service differentiation will be feasible. For example, real-time application support may not be granted, once the full bandwidth is used by all other customers, even at a very small probability.

For those reasons, flat fees and overprovisioning will not work for a services differentiated Internet, even though basic IP access and best effort services may show a strong static pricing component in their place.

4.5.3.3 Usage-Sensitive Models

Within dynamic models, all services are chargeable on a metered basis, e.g., on duration, volume, distance, or time of day. An important prerequisite for all dynamic models includes a suitable metering and accounting system, which allows for the detection of per-flow usage data, different QoS, time-of-day, and further service-specific information, which are part of the price and tariff applied. In addition, this requires resource allocation mechanisms for managing delineated resources. Usage-based pricing for telecommunication services is especially interesting, the because underlying resources used (satellites, spectrum, cables, routers/switches, and, most notably, operating personnel) are scarce and very costly. Bandwidth scarcity could be solved by installing more fibers or multiplexing on an existing fiber; however, as discussed in Section 4.5.3.2, this approach holds for certain links only, regularly for the local enterprise area, but not for the residential customer access loop or for the whole Internet. Operation of an entire and global network, and providing high quality end-to-end service, is still an expensive venture.

Therefore, major advantages of usage-based models include the fact that this system offers a wide range of service selectivity and an incentive to chose the service class really required for a current task. This allows for a service differentiation based on customer service valuation as well as reaching network efficiency (optimization of network resource utilization) and economic efficiency (Pareto efficiency), since congestion within the network can be avoided by raising prices and thus reducing customer demand. Due to potentially lower demand, the service provisioning can be maintained at a lower level and the quality can be maintained at a higher level than with flat-rate models. In this sense, potentially greedy human behavior is, at least, restricted by financial means. In addition, users are able to control to a large extent their bills for service usage, mainly based on feedback received with charging information. However, this approach requires a metered system. But insufficiencies in the performance of Internet metering, mediation, and accounting systems are fading, due to the appearance of high-performance technology solutions, including the ability to collect service-specific data at a high frequency, as is described in Section 4.3.2, and according charging systems are developed [85, 120]. Besides these dynamic technical aspects, service usage varies and, hence, the price for this usage may be influenced by the variable prices for a similar service usage at different times or locations. A good example for a dynamic pricing model is an auction with continuous price variability considered as repeated incarnations of the auctioning process [36, 124, 125].

On the problem side for usage-based schemes, it is often mentioned that pricing in general and usage-based pricing in particular, can impose a high overhead on telecommunication systems [78, 108]. However, [117] and [59] show two approaches for the Internet, which result in a manageable effort for the implementation of usage-based pricing schemes for integrated services. Besides any implementation, however, there exists a fundamental problem with usage-based pricing. This is caused by the type and precision (granularity) of the collected accounting information, which is used as the basis for pricing. For example, collecting information about connection times to an ISP rounded to 10 seconds means much less overhead than counting IP packets at each interconnection point. With current pricing models in single-service networks there is also implicit information, which can be used in the pricing process by exploiting an implicit traffic specification. However, if one aims at a more efficiently operating multiple-services network for applications with varying requirements [106], this implicit knowledge is lost and must be recovered from the information made available by the protocols employed. Billing is relatively complicated, since accounting information needs to be stored and aggregated accordingly. For the ISP, additional capital cost is required initially to provide accounting systems and user meters. Finally, revenue forecasting and budget planning will become more complicated, since demand estimations are required.

Volume-based approaches determine the form of a usage-sensitive model, since they reduce the accounting task to the traffic volume as an important parameter in terms of resource usage, but still require rather accurate monitoring of the amount of data traveling through the network. Note that delay is an equally important parameter, but even harder to be accounted for, as it is valid for further QoS parameters as well.

4.5.4 Pricing Model Classification Approaches

While flat-fee models show a fixed price and variable QoS due to unpredictable service usage, dynamic price models offer dynamic prices and variable QoS due to unpredictable service usage. But in the latter case, willingness to pay allows for the assurance of a guaranteed QoS. Usage-based service examples encompass user-defined VPNs, subscriber-activated VoD, or any value-added service, which utilizes “more” resources than regular services would require. Flat-fee examples include an email service, without multimedia attachments and chat functions, as well as a basic IP access service without any QoS requirements. These two categories outline two extreme ends of a spectrum of possible pricing schemes. Many combinations and a variety of approaches are possible.

Based on these considerations of price model components, characteristics, and types, relevant pricing models are classified and their parameter dimensions are defined (cf. Section 4.5.2.3). The following paragraphs contain an overview of important Internet pricing models that have been investigated over the last few years and have turned out to be of special importance from a practical and economic point of view.

Consider the example of a flat-rate pricing scheme. Over a fixed time, i.e., described by the attribute *time* and parameter *period,* customers can send as many packets as they like, i.e., the attribute volume is not applicable. No metering and charging entities are needed, since the volume is irrelevant in this classic flat-rate example, the *space* attribute is set to *no relevance* as well. The *quality* attribute instead may have some influence to the initial price set as the flat rate, but it is not a necessity for flat-rate pricing, so it can be set to an arbitrary parameter. In Table 4.5 the latter fact is expressed by an *x*. The volume-based (static and edge) pricing, the Paris Metro pricing (PMP) scheme [92], Vickrey Auctions [78], congestion-based pricing, and CPS [120] are also classified in the table.

Concerning the goals targeted by pricing models, a clear focus on two concurrent topics can be recognized. The first set of pricing models targets congestion in networks, i.e., congestion control and avoidance. This is a global approach covering the entire scope of an ISP’s domain, representing ISP’s desires in the management of limited network resources by the deployment of appropriate network technologies, e.g., ECN [99].

In contrast, the second goal to be achieved aims at individual QoS provisioning. It has to meet dedicated customer requirements, where service differentiation is available and

Table 4.5 Pricing Model Attributes and Parameters [121]

Example Pricing Model	Time	Space	Class Quality Characterization	Technological Requirements	Volume
Flat rate	Period	Not applicable	ISP	*x*	Not applicable
Volume-based (static and edge pricing)	Duration	Location	*X*	Not applicable	(Non)linear cumulation
PMP	Duration	Not applicable	Self-adjusting	Class-specified	*x*
Vickrey Auction	Time-of-day	*X*	Self-adjusting	*x*	*x*
Congestion pricing					Not applicable
CPS	Period	Location	*X* (ISP and customer)	Not applicable	(Non)linear cumulation

Table 4.6 Combinations of QoS Models and Pricing Schemes [61]

QoS Model	Flat Fees	Static Prices	Dynamic Prices
Overprovisioned best effort	Good fit	Likely not viable	Undecided
Price-controlled best effort	Likely not viable	Likely not viable	Good fit
Differentiated services	Likely not viable	Good fit	Good fit
Integrated services	Likely not viable	Good fit	Good fit

suitable pricing models are likewise necessary. Pricing models are strongly influenced by these goals and cannot be decoupled from the objectives of the network provider, i.e., to satisfy customer QoS requirements or to avoid and control congestion. With respect to QoS provisioning and in reflection of the networking technology discussion in Section 4.4, Table 4.6 shows an overview of pricing scheme fittings to different QoS models [61]. While flat fee denotes the current access-based pricing scheme of the Internet as described in Section 4.5.3.1, static and dynamic prices correspond to the description of Section 4.5.3.3. This includes in particular a usage-sensitive component, where a pricing scheme is based individually on the amount of resources used for a service invocation or service usage.

As discussed in Karsten et al. [61], the nature of overprovisioned best effort service is such that no service discrimination is possible, and hence, price discrimination is not appropriate.[6] Although best-effort services have been used in combination with fixed per-packet prices, this cannot be considered a useful alternative, since fixed prices do not represent the resource consumption of best-effort communication. When best-effort services are combined with resource-sensitive pricing and variable prices, it basically resembles price-controlled best-effort service. In general, it seems doubtful, whether this QoS model is capable of providing the kind of service that is needed for differentiated application demands. Even the assumption of an ever-increasing amount of transmission resources at constantly decreasing prices (overprovisioning, cf. Section 4.5.3.2), a situation of superabundance can only exist in relation to a certain amount of aggregated demand. To attract

[6] The price can vary according to the customer access bandwidth, but still, this determines a flat fee for the customer on longer time scales (cf. Section 4.5.2.1).

widespread usage, such a system must be kept flexible with regard to requests from customers. Nevertheless, for reliable operation, it must be ensured that aggregated demand does not exceed an acceptable level. To combine both requirements, some kind of dynamic access control is needed (1) to ensure proper and controllable consumption of resources, and (2) to account for any premium service usage.

For a price-controlled best effort service, appropriate pricing and responsiveness of end-systems to price signals is the crucial management aspect. Because of this responsiveness of end-systems, per-packet charges provide a mechanism for dynamic access control. Under the assumption of stable price-demand patterns, it is possible to proportion capacity such that reliable operation and QoS assurances can be met statistically. However, since performance predictability can only be given under certain restrictions [94], such a service cannot provide the exclusive technology for an overall network infrastructure. Furthermore, prices are inherently variable in order to fulfill their functionality as congestion signals. It has been suggested that such a basic service be combined with higher-level entities, which act as trading or insurance brokers to remove price fluctuations or improve QoS predictability [85]. However, it may add a significant complexity to the overall system to implement such brokers and fine-grained interactions between them, if the frequency of these interactions reaches a certain limit. Future investigations need to design, simulate, and implement such systems carefully to provide evidence for their feasibility. For that reason, price-controlled best effort serves as an alternative implementation choice for certain service classes that do not specify hard QoS guarantees, e.g., similar to the controlled load service class [129].

In the differentiated service [4] and integrated services [7] models, resources are engineered or reserved according to requested service offerings. Independent of actual service implementation, some kind of admission control has to be executed on service requests in order to guarantee reliable and predictable transmission quality, as specified in the respective service classes. Since resources are allocated (more or less exclusively) to service requests and are therefore not available to others, charging has to be resource-based in order to keep the demand at a sound level and to avoid the tragedy of the commons phenomenon [48]. While these technologies gain relatively high complexity at the technical level of service provision in the network (IntServ higher than DiffServ), they also provide the most sophisticated interfaces to network management both and users (again, IntServ more than DiffServ). Consequently, the additional complexity of providing a wide range of different application services and pricing and charges for these services is lower than for price-controlled best effort approaches. Proposals for appropriate pricing models for these technology choices can be obtained from [36, 62, 77, 101, 112, 127].

4.6 ISP COST MODELS

At present, the ISP market is characterized by a set of new Internet services and interactions that differ significantly from the traditional telecommunications market. Mainly, this is due to the fact that basic IP access is extended in the case of the Internet with service offerings and content provisioning. Balancing these developments, the cost model for ISPs requires a fundamental reshaping, since traditional ways of modeling costs in a communication network do not hold any more. Therefore, the main focus of an ISP cost model is (1) to identify all relevant parameters, (2) to list their mutual relationships that contribute to the cost for providing network services to a variety of users, and (3) to include

Internet service to be considered explicitly. In a competitive market, cost modeling is helpful in two ways: first, a suitable cost model serves the ISP with respect to its internal cost management, as it helps to understand its own costs, which may have crucial influence on marketing decisions as well as operational processes. Second, a cost model provides a solid basis for calculating and determining prices, tariffs, and charges for value-added Internet services.

The earliest relevant work on ISP cost models investigates costs for Internet access using cable modems compared to the ISDNs [44]. Advances are made in Leida [75], where observations from yield management techniques are included and the set of access technologies is refined. In a first step, Leida [75] assumes that a limited set of services access classes (different bandwidths of the local loop) offered by an ISP, including dial-in analog access, dial-in ISDN access (128 kbit/s), 56 kbit/s leased-line access, and T1 leased-line access. This does not distinguish further among value-added Internet services. In a second step, a customer segmentation is performed, resulting in residential dial-in subscribers, business dial-in subscribers, business ISDN subscribers, business leased-line subscribers (56 kbit/s), and business leased-line subscribers (T1). While various assumptions about residential and business access are made, a mix of product-related (bandwidth of local loop) and customer type-related segmentation (business vs. residential customers) was achieved. The value-added service case includes investigations of two situations only: one with IP telephony in place and a second without. A different perspective was taken in the OPTIMUM project, which developed a tool for investment calculation; hence, the main goal is the investigation of investment in the telecommunication business [93]. Further features of the tool include especially network aportioning.

One characteristic of all relevant approaches of ISP cost models is their focus on concrete cases instead of developing abstract models. But concrete models lack flexibility and independence. A second characteristic of these approaches is their depth in technical and economical details, e.g., Leida [75] eventually uses more than 300 parameters for describing the model, which has major consequences with respect to the transparency of the model. In contrast to these approaches, the model ICOMO [100] is abstract, flexible, independent, and can be adapted to a set of services of interest. A trade-off has been made, e.g., between abstraction and simplicity, as it is possible to develop a purely formal model for all possible types of providers and services, but only by using an large number of parameters in contrast to the requirement of simplicity. In order to cope with them, the model starts from a purely formal and abstract view, but aims at concrete cases, which are used to feed the model and reduce its complexity. It turns out that this goal is reached by a subtle mixture of classic accounting and abstraction.

4.7 CHARGING SUPPORT SYSTEMS

Earlier work on charging and accounting in telecommunication systems has been focused on connection-oriented networks, such as the telephony network, ATM-based networks, or leased lines. However, the Internet provides a connectionless network layer, including an IP-based network service. While the set of traffic modeling parameters and service parameters for connection-oriented networks are quite well understood and agreed upon, these parameters remain heavily debated for the Internet. For example, the interpacket arrival time for an Internet service makes a significant difference for this service. However, how should a future system account for this parameter? In addition, as for connection-oriented networks, the call-blocking rate determines the level of utilization for a given topology and the potential sender is blocked in sending data into the network, while connectionless networks suffer from the problem of congestion, since in general there is no admission control available. Of course, a set of newly defined Internet services proposes the existence of such an admission control; however, a commonly agreed upon architecture has not been developed up to now. Once congestion situations can occur in a network, congestion control mechanisms are required. Traditionally, these mechanisms have operated in the purely technical domain, e.g., by dropping packets, but left out incentives to evaluate the requested service by economic measures. As discussed, sensible pricing of services in an open services market is also required [118]. However, this approach depends on the technical ability to collect and account for those data necessary to charge the customer. Therefore, a CSS provides the technology required to support price-based mechanisms for charging tasks, including the potential to perform a market-driven congestion control in the Internet along with interfacing billing systems and proving customer feedback signals on resource usage.

4.7.1 CSS Components

The CSS approach taken, provides (1) a generic and modular Internet charging system in support of various pricing schemes applicable to different communication technologies, and (2) an interface for billing systems (cf. terminology in Section 4.2.5). The goal is to identify relevant components and their relations to each other, and to create an open and complete system structure, which allows for the integration of charging support technologies available today (cf. Section 4.3.2 and Section 4.4), ranging from the data-orientated tasks (such as metering) to the money-related tasks (such as a billing interface).

Several components are needed for an Internet charging system, which have to interact to provide all offered functionalities to customers. As illustrated in Figure 4.8, a general

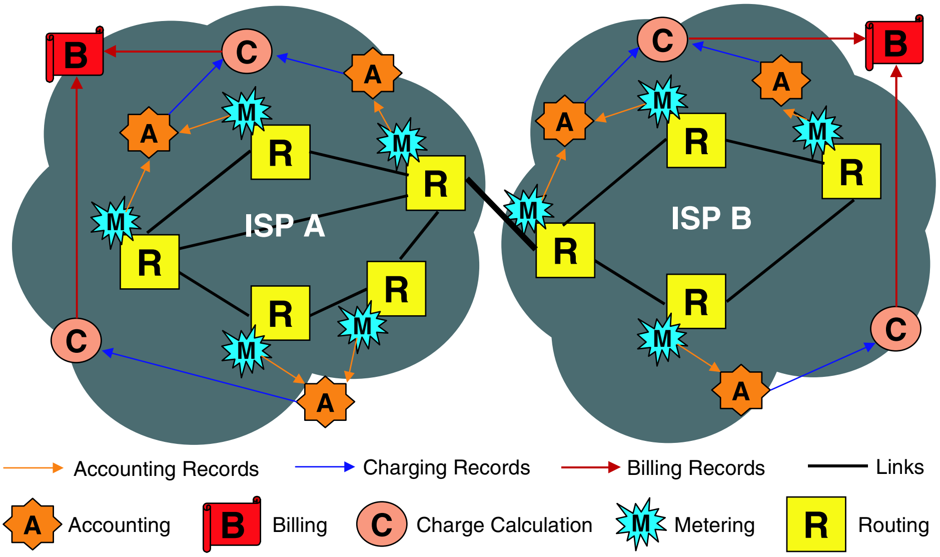

Figure 4.8 Location and replication of charging components in provider networks [120].

scenario contains interconnected communication service providers. Each provider operates a network consisting of routers and network links between them, accounting components, charge calculation components, and billing systems. Metering components are located inside the router or attached to them as separate devices. In either case, they generate accounting information, which are gathered and accumulated in an accounting component following some type of data format (cf. Section 4.4.5). In turn, it forwards the accumulated accounting information through a charge calculation function toward the billing system. The charge calculation translates the accounting information into charging records; hence, it maps resource-oriented information from accounting components into monetary values. The billing system uses these values to prepare bills that are to be sent to customers. Within the charge calculation, discounts and rebate strategies, marketing-driven pricing schemes, or simply fixed prices can be applied in terms of a selected tariff.

In the existing billing systems of today's providers, the setting of prices, the function of charge calculation, and the billing itself is integrated, even additionally combining the maintenance of service classes, user profiles, customer data, identities, and bank account data. Although these steps still can be seen clearly, they are almost completely centralized within a monolithic system. Future billing systems need to be able to integrate a variety of different charging records, even from different communication providers or content providers, since customer's demand is one-stop billing [118]. Future CSSs need to react on user requests in a short time scale, provisioning a soft real-time feedback. The move of tomorrow's CSSs from a back-end and batch-based to a front-desk and real-time-capable system are clearly visible. This strongly suggests dividing the existing monolithic systems into several components with clearly defined interfaces. By doing this, it will become possible to exchange individual components and to integrate different components supporting different technologies without having to adapt the entire system. Additionally, interfaces to metering, accounting, and other components also have to be defined, based on available technology choices (cf. Section 4.3.1.1). The goal is to define components and their relations to each other and to create an open Internet charging system architecture, which allows the charging task to be performed for various different technologies.

4.7.2 CSS DIMENSIONS

Based on these component identifications, it must be determined how these components are deployed in a particular distributed scenario with potentially several different ISPs [121]. A charging system can vary with respect to four essential dimensions, driven by the scenario (cf. Figure 4.9) and the ISP type, defining a set of different choices based on the different roles for access ISPs and core ISPs (cf. Section 4.8.3 and [71]). Depending on the ISP type, the location as well as replication of components will determine suitable and less useful components combinations. However, today there is no general set of criteria available depicting the optimal location and replication of components for a given scenario. It is expected that future work on ISP cost modeling may determine suitable design input [100].

The "Location" dimension defines where components are located. In particular, an inhouse location refers to an ISP, hosting this component and providing the corresponding functionality internally. The outsourced location defines that this component and functionality are being performed outside the scope and administrative domain of the ISP. Mainly business case assumptions and the size of the ISP considered will determine the

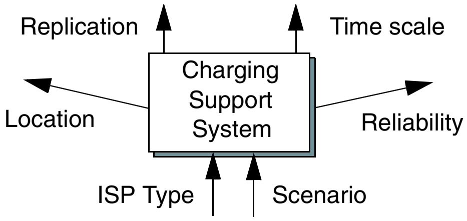

Figure 4.9 Charging support system dimensions.

final location of components in a given ISP infrastructure. In addition, security-relevant questions may arise, once the outsourcing of financial activities is intended.

The "Replication" dimension defines how many of the components considered exist in a given environment. Mainly the number of clients served by an ISP and the number of interconnection points with peering ISPs will determine the number of replicated components required. However, besides the pure replication an important issue is the interaction between these replicated components. Appropriate protocols (open, ISP-specific, or vendor-specific) need to be selected for a suitable and correct design as well as the implementation. Open interface specifications are required to provide a chance for replication.

The "Reliability" dimension defines how reliable these components have to be. The required degree of reliability depends only indirectly on the ISP type. Rather it depends on the other dimensions of location and replication previously mentioned, and heavily on the component type itself. Nevertheless, the required reliability of components is a dimension in which a specific charging system can differ from others.

Finally, "Time scales" define the last important dimension. As discussed in Section 4.5.2.1, four network management time scales are distinguished and applied to pricing-controlled mechanisms for feedback purposes.

4.7.3 Basic CSS Tasks

The CSS is supposed to support the following tasks:

- Perform accounting tasks according to service definitions. Data gathered from the physical infrastructure and mediated according to policies needs to be accounted for. This requires the knowledge of sessions, durations, or flows. Mainly, this information is derived from metered data as well, such as "begin-of-session" or "end-of-flow." If such starting and endpoints cannot be determined explicitly, heuristics need to be applied for session- or flow-detection purposes. In any case, the "length" of a communication relation will be recorded, if any usage-based charging approaches are to be supported.
- Perform multiservice accounting. The accounting task for a single service that is well known is performed by an algorithm, which utilizes a clear service specification. In the case of multiservice provisioning, these service specifications must exist and need to be maintained concurrently. Therefore, the separation of incoming data and their mapping onto the particular service in operation is essential.
- Support transport, service, and content charging. The optimal design for a CSS includes a combined approach for the three different levels of charging. Transport

charging, sometimes termed network charging or network access charging, forms the basis for providing a system to deal with the transfer of data based mainly on the network infrastructure, such as the Internet.

Service charging located above this level allows for the clear distinction of different services, including different QoS requirements and resource consumptions. Services include the ones provided by a variety of service providers (cf. Kneer [71] for their different definitions and distinctions) that are offered in an open-market situation. This charging task needs to be service-independent as far as possible, to ensure future extensions and adaptations to yet unknown services. Transport charging will be integrated into this concept and may even be hidden completely.

Content charging includes the accounting tasks for information that is specifically monetary-sensitive and needs to be paid for by reading, using, or copying it. Based on the level of business interactions, it might be useful to apply content charges for certain services only, integrating invisibly by customers the underlying transport and services charging.

- Support different levels of security for charging and accounting information. All data and information related to monetary equivalents contain a certain degree of sensitivity. However, due to the dedicated level of interest, a single accounting record, a metered routing datum, or a charging record may not be a security problem, since their lifetime and validity, and therefore asset, are short. But other combinations of aggregated data, e.g., flow-related information in terms of usage information, duration, and customer identification, form critical information.
- Support auditing. Communication services offered in a market environment need mechanisms that support the proof of service delivery under well-defined circumstances. Therefore, an auditing functionality will be based on accounted for data, which may be specifically restricted, structured, or stored, depending on legal aspects, such as telecommunications acts.

4.7.4 CSS Architecture

Based on these charging system tasks, their clear separation, and design dimensions, an overall architecture for a CSS is driven by the mapping of the conceptual view onto certain components [121]. Adding their interactions and interfaces results in the CSS architecture depicted in Figure 4.10, where interactions between two neighboring providers take place on two levels. The first one is on the data path, since providers must exchange data between their networks. Interprovider information exchange happens as part of the specific protocol processing as defined in the QoS model applied, e.g., for resource reservations using the RSVP or inter-Bandwidth Broker communication, where messages are exchanged between the border routers of neighboring providers. In these cases, a type of signaling or consolidation protocol has to take care of distributed information scattered around in the network.

Since the transport of these data is not for free, ISPs will charge each other for data transported. This leads to the second level of interaction. Each provider collects information on the amount of data transported and calculates a charge for it.[7] The provider issues inter-provider invoices through a billing system to the responsible neighboring ISP's entity. Thus, information exchange between providers occurs on the level of billing systems.

[7] Accounting rate regimes have been applied in the traditional telephony system, however, differentiated services in the Internet require an SLA-based distinction of services exchanged and offered (cf. Section 4.4.4).

Instead of performing absolute billing between interconnected providers, they can also offset their claims against each other. A set of peering agreements and settlement schemes exist for today's ISPs; however, (1) they are defined in a quite static manner, (2) they do not allow for immediate responses to bandwidth bottlenecks or further customer and user demands, and (3) they cannot support differentiated services effectively. Besides this interprovider billing, providers bill their single customers as well.

4.7.4.1 External Components For describing the CSS completely, an outside-first approach is taken to illustrate all components external to CSS's central component, the Internet Charge Calculation and Accounting System (ICCAS). As shown in Figure 4.10, metering is integrated in the IP router. Alternatively, it could be placed directly on the wire. Such a solution introduces supplementary expenditures, e.g., an entity needs its own IP address or requires special protocols. Furthermore, it can only monitor the actual usage of the link and has no knowledge of usage of any critical resources relevant to congestion control within the router. Therefore, the interconnection of several metering units to reconstruct the current router status is not feasible. Finally, it would be necessary, in spite of having metering units on the wire, to know the state of the router, so explicit interaction of the charging system and the router would be required. The purpose of the mediation entity is to transform metered data (of each single meter), to merge data of different meters, and to reduce the amount of data metered.

Prices are important for calculating charges of transmitted data. Since there are many different ways to set prices, a separate pricing component performs this task. It can make use of economic models or just use fixed prices set by hand (cf. Section 4.5 and Karsten [60]). For dynamic pricing models most often an input from metering is needed, since the amount of data traffic influences prices. These prices are applied in the charge-calculation component within the ICCAS. An interface to a billing system exists and is used to perform interprovider charging concerns as well as customer billing interfacing.

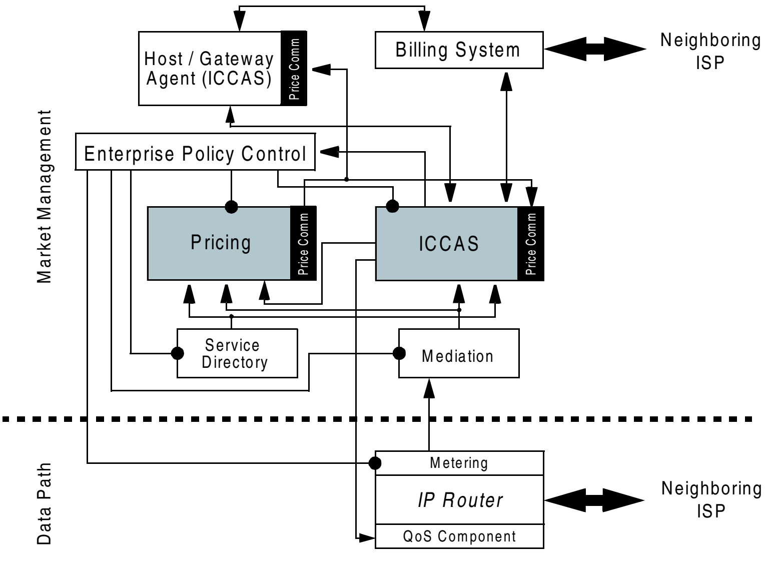

Figure 4.10 Architecture of the charging support system [120].

The policy control entity of the enterprise represents the ISP's interface for the management and supervision of all ICCAS-related entities except for the billing system. It covers and controls the ISP's business strategy with respect to its implementation and configuration of the given networking equipment and services catalog. The host/gateway agent performs two different functions. The first one is to communicate charges to hosts (users) and gateways to provide an optional feedback channel for their service usage. In this case, the host agent acts on behalf of the user. This can include the negotiation of services with the ICCAS, an automatic reaction to communicated charges, or even payment information. A host agent can also restrict the user's options, when the customer in control of these users wants to restrict the behavior of the users it pays for. In particular, this is the case for companies where the company constitutes the customer to an IPS and its employees act as users of the services offered.

Since in general a users tend to lack a complete understanding of QoS in technical terms, they will be unable to specify detailed requirements in a way that can be used as a direct input to the QoS component within the router. Instead the users have a higher, application level view of quality. This view must be translated into technical values, which can be used for setting parameters in QoS components and for charging according to technical usage data. Therefore, this translation takes place in the Service Directory.

4.7.4.2 ICCAS Architecture and Internal Components The internal entities of the ICCAS include a charge calculation, an accounting, a customer support, and a user support component. The separation of the ICCAS into these components increases the required degree of flexibility, since these components can be physically distributed as discussed in Section 4.7.2. Embedding the ICCAS into the CSS is accomplished through eight distinct interfaces. Concerning the flow of data internally, the ICCAS has been divided into two logical paths as shown in Figure 4.11. The first, the Accounting Information Path (AIP), depicts the flow of pure charging-relevant data. On the other hand, the Control Policy Path (CPP) is used to manage and configure the ICCAS, especially all entities involved with the processing of charging data. These two paths differ mainly in the order and direction that they process data. The AIP starts from the bottom of the graph (taking raw technical data) by processing metered and mediated data as well as pricing information. It ends at the top of the graph, where complete charging records are handed over to the billing system. In contrast, the CPP starts from the top of the graph (taking business-related data) by receiving enterprise policy control information and processes down to the bottom of the graph, resulting in QoS control data to be handed over to the underlying router or optionally an agent.

The accounting component receives all metered and mediated usage data and is responsible for storing it. It must provide these stored data to other ICCAS components and interfaces for further processing, feedback, or statistic evaluation. Accounting is the central usage data storage component. The charge calculation component processes the accounted-for usage data. It calculates appropriate charges for resource usage by applying a tariff, communicated by the pricing component. To be able to determine the charges fully, it needs input from the user support component, e.g., user identifications, to apply further contract specialities.

An ISP may have many different customers. Additionally, a customer is not the same as a user, e.g., one customer might pay the bills of several users. Therefore, a customer is the one who negotiated a contract with the ISP. The content of this contract, e.g., number of users covered by the contract and their names and accounts, are managed within the customer support component. While the customer support component is responsible for

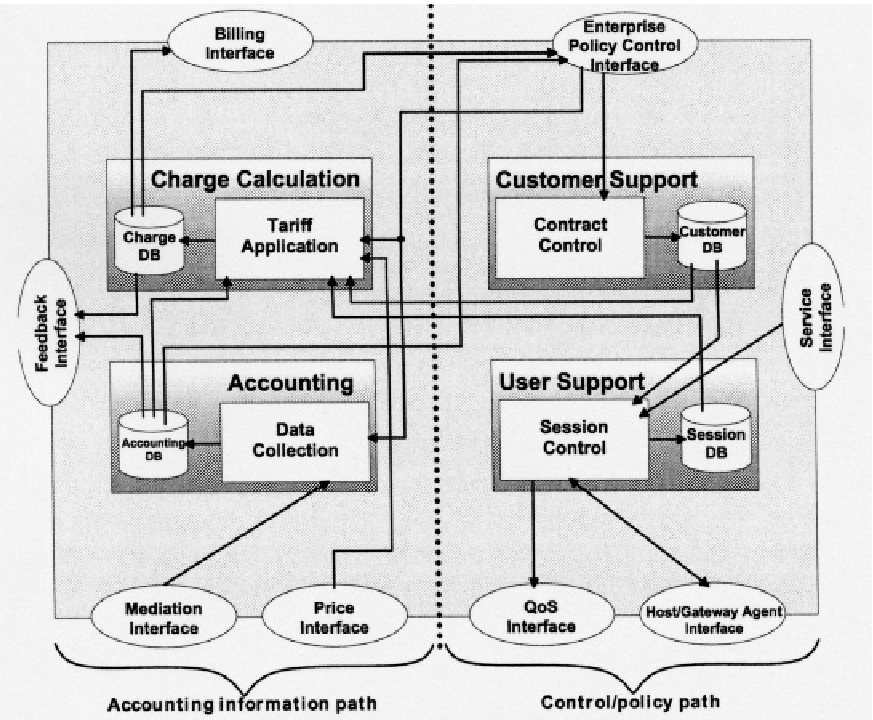

Figure 4.11 ICCAS architecture in detail [120].

keeping all contract information, the user support component is responsible for making sure that those contracts are kept safely. On the one hand, this means that it blocks any user requests that are not covered by the contract the customer, who pays for the user, has. On the other hand, is can make sure that a service requested by a user is delivered, if the contract allows it.

4.7.4.3 Open ICCAS Interfaces All ICCAS interfaces between the components just described are designed to act (1) as protocols, allowing for open communication between two remote entities of components, or (2) as software interfaces, reflecting the clear architectural decision, that the interaction between those components happens within a common address space [120].

The QoS Interface (CPP) offers the control of routers' QoS components, where services to the customer are provided. It can be used to set QoS parameters of routers, depending on the technology in place. The Mediation Interface (AIP) is responsible for collecting data from several mediation entities, possibly even from mediation entities of different types. The usage data, which are mediated after the data gathering takes place, need to be transferred to the ICCAS. Therefore, a protocol is designed that defines rules and transmission units for transferring mediated data to the accounting component. Since the anticipated load for this interface will be high, the protocol must be highly efficient, yet extensible. The data exchanged across this interface include one of the following alternatives, which depend on the particular scenario: (1) a simple handover of data gathered by metering, or (2) a handover of data mediated based on the particular inputs from the enterprise policy control. This may result in a dedicated specification of specialized data to be required for the ICCAS, some special aggregation of these data, or even the neglect of data resulting from the gathering process. The Enterprise Policy Control Interface (CPP) is intended for changing parameters of the ICCAS after the system has been deployed. By using this interface, the enterprise policy control can install new services or request and receive charging or accounting data. The Service Interface (CPP) can be used to read service definitions out of the service directory. The Billing Interface (AIP) is responsible for sending calculated charging records to the billing system. Pricing is responsible for setting the prices used by the charge calculation component, therefore, the Pricing Interface (AIP) is used to send calculated prices to the charge calculation component. To set suitable prices, the pricing component uses price models with various input variables. Some price models need usage or charge information as input variables, hence, these variables can be communicated to the pricing component via the Feedback Interface (AIP). Finally, the Host/Gateway Agent Interface (CPP) is responsible for optional communication with the customer. Mainly, this includes the selection of services the customer can use or the transfer of a feedback signal from the service provider to a user. This interface is open for future enhancements. Further details on interfaces can be obtained from Stiller et al. [120].

4.8 BUSINESS MODEL ASPECTS

Commercially operated networks, in particular subnetworks of the Internet, follow a set of requirements that have been discussed in [37, 61, 62, 83]. They cover the fact that the business task of running a communication network must be sustainable and profitable to attract necessary investments in the infrastructure. This means that for charging systems, e.g., objectives in terms of flexibility and efficiency have to be met, which are completed by a number of criteria arising from current practice and user expectations in the areas of product liability and consumer protection in general. This is behind the need to introduce new services and charging for them.

For those reasons, a three-tiered model is introduced that interfaces with the CSS presented earlier. Based on its three layers, ISP segmentation in the lower two layers is performed and application service providers (ASP) are introduced. Because electronic content in the Internet forms an extremely important area of concern, a valuation of content versus transport is discussed and a combination of content with transport charging is proposed.

4.8.1 Market-Managed Three-Tiered Model

The infrastructure of a market managed multiservice Internet is based on an overall three-tiered model [120], as shown in Figure 4.12. This model outlines basic states and sources of information within three distinct layers as well as its interfaces to the CSS. Starting from the topmost layer, customers and providers within an Internet services market interact for any type of business based on business models as defined within the Enterprise Policy Layer (EPL). This layer defines, among others, products to be sold, i.e., services, models of business interactions between customers and providers, pricing mechanisms, and agreements on an offer. Details of relevance to the CSS can encompass, e.g., rebate systems, discounting schemes, service plans, or service pricing models. These details

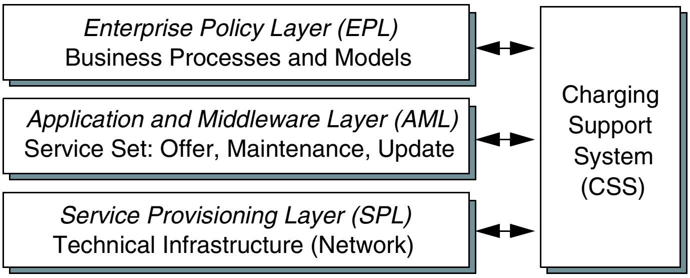
Enterprise Policy Layer (EPL)
Business Processes and Models
Application and Middleware Layer (AML)
Service Set: Offer, Maintenance, Update
Service Provisioning Layer (SPL)
Technical Infrastructure (Network)
Charging Support System (CSS)

Figure 4.12 Market-managed three-tiered model.

form the business-dependent and business-central policy, which may not be completely published, but is required to provide the CSS with operational dimensions. However, to perform any type of market-driven enterprise policy, the CSS needs to offer a set of service descriptions that are applicable to all areas of enterprise policies.

Besides these business perspectives, the technical view of the market-managed approach is found in the Application and Middleware Layer (AML). It provides functions or policies that are (1) initiated due to a predetermined application or (2) acting on the application's behalf, such as a given enterprise policy. Within this layer, a set of (value-added) communication services is provided, utilized, and charged for according to customer demand. Middleware takes from the details of the technical infrastructure of the network itself. The middleware is able to provide a generic service set for offering, maintaining, and updating all types of communication services. Therefore, the CSS interface includes application-centric configuration options for particular session and service descriptions as well as more generic service descriptions according to the middleware layer functionality.

Finally, the task of service provisioning as identified in the Service Provisioning Layer (SPL) offers interfaces for the lower layer to the CSS. Therefore, data and information on the technical infrastructure (the "network") are collected and maintained depending on the service offered by the middleware layer.

4.8.2 Internet Service Provider Segmentation

In regard to the SPL of the three-tiered model, the type of service provider considered needs to offer a wider range of services from the networking domain. Hence, the focus for segmentation has to be directed to network service providers. In the current market situation, many network providers already have completed forward integration, since this allows them to reach the end-customer directly. Others remain within the backbone of the Internet. Therefore, from the position of end-customers, the distinction between two types of providers is straightforward. The first type is named telecommunications Internet service provider (TISP), and the second type pure Internet service provider (PISP) [100]. In the simplest case, the organization of these service providers consists of two levels: the lower level represents the TISP offering their services to PISPs, which are located on the upper level of Figure 4.13. Some TISPs may participate as hybrid providers in the market, i.e., they reach end-customers directly, but offer their services to other PISPs, too, as in the case for $TISP_4$.

A TISP is an enterprise owning and operating a network as well as an infrastructure of

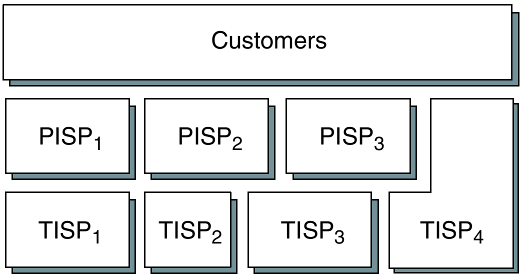

Figure 4.13 Vertical provider market structure [100].

its own, including routers, switches, and network management software, among others. This network and infrastructure includes virtually "every" necessity, i.e., the local loop, backbone links, and peerings to other backbones. Hence, the TISP segment includes former monopolists, as they become the only ones within a long-haul transmission network after the market has been deregulated. Additionally, providers exist that have taken the chance of setting up and building new networks of their own. These providers usually offer various IP services, such as voice or phone. TISPs offer services to PISPs, but they reach the end-customer directly as well, by means of forward integration. Typically, these providers possess a great deal of stamina in the market. They currently possess the largest customer base and are able to deal relatively easy with their telephone subscribers as potential IP customers.

PISPs basically do not own networking infrastructure at all, since they require TIPS or other PISP business interactions. In particular, they have to rent access from TISPs or buy services from other PISPs. Instead, their equipment is service-specific, e.g., web, multicasting, or name servers, and their portfolio offers value-added services. Moreover, these providers usually operate call centers, help desks, and hot lines in order to care for customers. The customer ought to be confident in the quality of the offered service and wants to be assisted promptly in case of failures. Current market conditions allow for the customer to choose among different offers or to change the provider, even if the transparency of the market has decreased significantly due to the deregulation and hard pricing battles among providers. Mapping these segmented provider markets into the tree-tiered model, TISPs operate on the SPL and probably on the AML. PISPs run their business in the AML, focusing on the middleware and services set. Finally, ASP operate on the AML, and are discussed in the following section.

4.8.3 Application Service Provider

ASPs deliver applications and services from distributed data centers to end-customers. The economics considered define a single business relationship between a customer and an ASP. In this case, a business model includes the series of business events and reflects all processes within an economic system [71]. In terms of electronic business systems, the ASP can be modeled by one instance and the end-customer by another, which means that these instances form the two endpoints of a communication.

ASPs offer products, content, and services on-line, e.g., via the WWW, and represent the merchant or the seller of a good. Products offered by the ASP include (1) physical and (2) nontangible goods (content). Physical goods could be purchased in stores, e.g., books, cars, or CDs. However, there are several incentives for customers to purchase products from home, such as 24-hours order service and delivery, no waiting queues at cash desks, time saving since no parking space has to be found, and electronic catalogs and navigation functions. Content offered by ASPs includes digital information in the form of bits, including the digital content of an on-line book or the digital version of a CD that can be downloaded into endcustomers' local disks. Furthermore, there exists a broad spectrum of multimedia contents and services, e.g., audio-on-demand or VoD, offered to customers on-line by the ASP.

The basic idea of an economic system established between end-customers and ASPs is well understood and feasible. Problems arise with the introduction of network performance and QoS for the delivery of Internet services. TISPs provide the physical infrastructure for the Internet and form the basic foundation of electronic commerce activities. PISPs and TISPs provide the transport of data packets over the Internet between the end-customer and the ASP. However, traffic congestion and unreliable connections on the best effort Internet are a problem for sending data, especially if wide bandwidth and reliable throughput are required, e.g., for multimedia applications or business transactions. ISPs could overcome this problem by offering services with QoS guarantees and by charging Internet users according to the service they request.

A business model for a general eCommerce scenario is shown in Figure 4.14 with an end-customer and ASP at the ends, and several intermediate ISPs. Business relationships between all parties involved are depicted, and ISPs will charge and account for Internet transportation services as well as higher-level eCommerce services. There might be a payment provider that is responsible for the financial clearing between parties. The payment provider can be a bank, a credit institution, or a trusted third party.

For the design of a business model, ISPs are differentiated into two different roles according to their scope of duties: Access ISPs (mainly PISPs and some TISPs) and Core ISPs (TISPs only). This differentiation is orthogonal to the ISP segmentation, since the definition of a particular roles of an ISP is a must for business models. Access ISPs support local access networks and provide Internet connections to the end-customer, be it directly or through a Customer Premises Network (CPN). Core ISPs increase the reach of Access ISPs to a global extent and form the backbone of the Internet. They perform data transportation by interconnecting Access ISPs. There may be more than one core ISP involved in a communicating connection between end-customer and ASP, depending on the

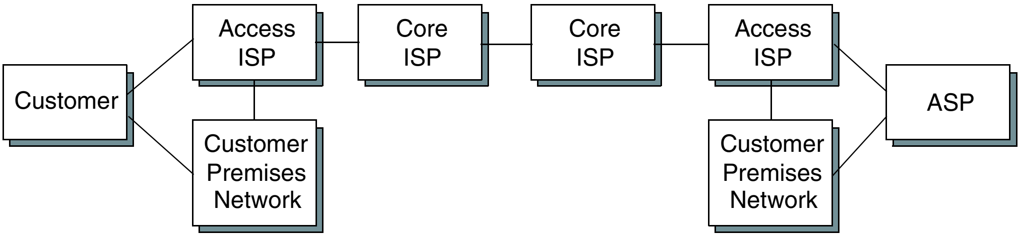

Figure 4.14 Relationships between involved parties of an eCommerce scenario.

connectivity of ISPs and their local distances. When there is only one ISP located between end-customer and ASP, it acts both as Access ISP and Core ISP, providing data transport. Thus, Access ISPs and Core ISPs can be physically identical.

An end-customer can be connected either directly to the CPN or to an Access ISP, if the end-customer is a residential user. In the first scenario, end-customers may be affiliated to a CPN, e.g., a LAN of an enterprise or a university campus, which groups end-customers and establishes the connectivity to the Access ISP. A CPN may offer additional private applications or extra conditions for data transportation to their end-customers. Therefore, a CPN represents a group of users in terms of a common policy and conceals an individual end-customer from the Access ISP.

4.8.4 Charging Content

Although, the focus of this overview has been on charging of transport services and their technological implications on the Internet networking technology, charging for content is important and becomes a substantial factor in the case of value-added services. The charge for the content of an ASP can be combined with the charge for the transportation service of the ISP, since the transport may be only a very small fraction of the overall charge. Furthermore, as mentioned earlier, the generalization of charging for services will be an important basis for the market-driven positioning of service differentiation of ISPs and ASPs. It should be noted that changes in the networking infrastructure, which are mainly due to upcoming mobile providers, will lead to new business models, since the customers in the future will dial up a connection to their mobile providers, and not a traditional ISP. This means that typical ISPs have to differentiate their services to obtain a future source of revenue due to decreasing income from stationary or mobile dial-up clients. Current investigations on including content in a wholesale approach [45] and electronic business models for content delivery and charging for them [76] lead initially to emerging areas of concern.

This section describes relevant factors that allow for the decision, whether or not it is favorable to charging content along with the transport service [71]. Figure 4.15 shows

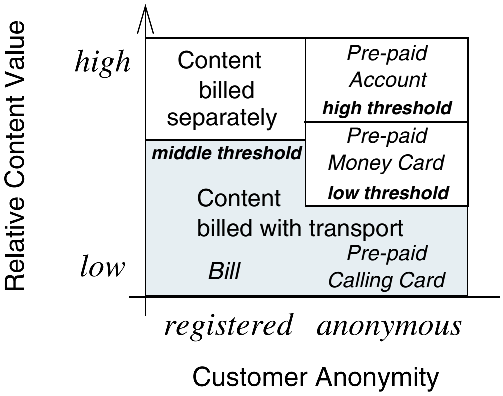

Figure 4.15 Primary influencing factors for billing content separately or including transport [71].

the primary influencing factors, namely, the relative value of content (defined as the value of content divided by the cost of transport) and required end-customer anonymity. Note that the number of bytes does not determine the value of the content as it does the current market price. For low values of content, it is more efficient to charge and bill content along with the transport service either on a single bill or through a prepaid calling card. A separate bill for the content will be provided for registered end-customers, if the relative value of content is higher than the middle threshold. If end-customers want to be anonymous and the relative value of content is between the lower and the higher threshold, prepaid calling cards can be used to pay for that content. Extremely valuable content above the high threshold requires prepaid accounts on the ASP's side with secured access rights. Content with middle as well as high values should be charged only when the transfer of content has been successful, e.g., the entire data file has been received correctly.

In addition to these primary factors, Figure 4.16 shows secondary influencing factors for deciding whether to bill separately for content or along with the transportation service. It can be assumed that there is a measure for the number of logically independent content blocks per time and the value of the full content per time.

Logically independent blocks of content are discrete and contain useful information for the customer, such as the reply on a railroad schedule request. However, financial information, e.g., stock rates, consists of a sequence of independent data packets, which are billed continuously. A backup file or a high-resolution image file can be seen as a large block of content. If only a single intermediate packet gets lost, the transmission will fail and no billing for that content should be performed. Certainly, boundaries for these loss numbers and QoS parameters are not fixed, but vary according to the application in use. For example, watching VoD can show an acceptable ratio of lost packets as long as the audio and video quality do not decrease below a certain limit. The transmission of a high number of logically independent blocks of content with a low value (lower right corner of Figure 4.16) is billed together with the transport service, if the QoS for the transport service is below a certain threshold. If QoS for the transport service increases, the QoS threshold moves to the upper left corner in Figure 4.16, the rel-

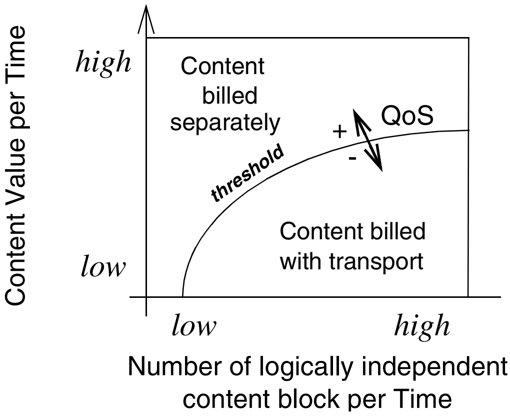
Content Value per Time
high
low
Content billed separately
QoS
+
-
threshold
Content billed with transport
low
high
Number of logically independent content block per Time

Figure 4.16 Secondary influencing factors for billing content separately or including transport [71].

ative value of content decreases, and billing of content is performed along with the transport service.

4.9 SUMMARY AND CONCLUSIONS

Today's Internet shows a clear move toward the support of differentiated services. However, purely technical mechanisms to control the access, congestion, and QoS disclose a major drawback—the lack of highly scalable mechanisms—mainly for millions of users. Therefore, economic control mechanisms, considered network management functions of the second generation, such as charging and its service-oriented derivative pricing, bridge this gap.

This chapter provided an overview of charging, the application of pricing mechanisms to Internet services, and related technology in support of these mechanisms. Determining such an approach to charging for packet-switched networks, the solution presented integrates economic mechanisms into the area of service provisioning. This extension is mainly driven by the commercialization in networking and the demand for QoS-based services, but minimized technical effort for QoS provisioning in networks.

In regard to the historic development of the Internet, two phases where economic principles have not been regarded as the major driving force can be distinguished. In the initial phase, networks were run by universities, backbones were centrally funded, and commercial use did not exist. Basically, pricing for service usage was hidden from the user and customer, and private use was almost unknown. The second phase included commercial and private users at the same time, the global deployment of the network, and its continuous extension with respect to technical mechanisms and geographical space. Simple pricing models have been implemented to support the cost recovery process of ISPs recently starting business in order to provide commercial Internet services. Limited by technical capacities of Internet protocols, driven by the run for a quick and large market share, and minor knowledge on charging packet-switched networks, flat-rate pricing schemes dominated this phase. Even though economic theory stated that an all-you-can-eat mentality is an inefficient way to price Internet services, ISPs are still offering monthly rates. Flat-fee schemes seem to offer unlimited access, but in reality they prompt an offer for a low-end QoS, at least on average.

Today the Internet is in transition between phases three and four and economic incentives play a major role. Newly developed network technology for the last mile, such as digital subscriber line technology, emerging QoS-capable Internet protocols, and an ever increasing Internet backbone capacity allows for the provisioning of new Internet services, service bundles, and applications, such as IP telephony, radio over IP, interactive TV, and other time-sensitive multimedia applications. This shift from a single-service class, best effort network to an integrated services Internet providing multiple service classes cannot be safely supported by flat-fee pricing models, since the particular resource on a per-application basis varies widely. Therefore, a completely integrated, efficient, and purely technical solution will form the immediate starting point for the fourth phase in Internet charging. All findings for an optimal balance between economic and engineering efficiency as well as user transparency and its large-scale deployment in a market managed multiservice Internet will guide commercial service suppliers and the research community. The variety of pricing schemes will be consolidated based on accounting technology, user transparency, and achievable economic efficiency. By offering flat-fee pricing for low-quality services, such as email or best effort traffic, and providing usage-based schemes for resource-consuming services, a distinction between residential user Internet access and commercial enterprise access will become a reality. High-quality end-to-end connections between the customer and service provider as well as between customers are a must. More expensive services will require a finer charging granularity than cheaper services do. Technologywise, least-cost routing functions or intelligent agents will act as price and resource brokers. Trading resources on bandwidth spot-markets has been established already for the large backbone capacity in order to provide flexibly adapted interconnectivity. Since an increasingly competitive market will exist, a trend toward dynamic pricing models can be assumed for enterprise access as well as value-added service provisioning.

Predicting the outcome of technological choices is difficult, and is specifically impossible for the Internet, but the major trend of the Internet development from a technically focused network to an economically controlled, efficient global network and distributed system is on its way. Modern accounting technology, emerging rapidly, e.g., vendors plan to ship OC-12 active measurement components and passive probes at Gigabit Ethernet [22], supports this trend. CSSs and their internal mechanisms provide the technical foundation for integrating those developments and making it happen, including front-end functionality or soft real-time feedback to customers. New economic models, implemented in terms of pricing and tariffing models, will ensure an efficient allocation of network resources. Therefore, the development of discrete, predictable, transparent, and technically feasible pricing schemes, one of them called CPS, and an ICCAS, illustrate a feasible technical solution in support of pricing tomorrow's differentiated services in the Internet. This solution will maintain parameters for customizing subscriber preferences and profiles, offer granular service control and service management features, and provide mechanisms to correlate user profiles with service profiles.

Business models for traditional ISPs will fade out as soon as providers become mobile themselves. This will be obtained by the third-generation Universal Mobile Telecommunications System (UMTS), where the typical role of an Access ISP can be performed by any mobile provider. However, due to the fact that frequencies and the currently available spectrum for sending data will be limited always, which is in clear contrast to the wired Internet backbone, efficient and user-based charging approaches have an emerging need for next-generation, packet-switched, mobile service providers. Barriers for usage-based approaches by appropriate, efficient, and technically manageable technology are being phased out. In addition, the ability to aggregate massive amounts of data, and filter, mediate, and correlate them with user and customer data eases per-customer and per-user charge calculation.

ACKNOWLEDGMENTS

The author expresses his gratitude to his senior researchers and Ph.D. students at TIK, ETH Zürich, namely P. Reichl (Sections 4.5.2.3, 4.5.4, and 4.8.2), P. Flury (Section 4.5.2.3), and J. Gerke and Hasan (Section 4.7), and his research partners T. Braun, G. Dermler, and M. Günter, (Section 4.4.4), B. Briscoe and C. Courcoubetis (Section 4.3.1.2), M. Karsten, J. Schmitt, and L. Wolf (Sections 4.2.2, 4.4.1, and 4.4.3), H. Kneer and U. Zurfluh (Sections 4.8.3 and 4.8.4), and S. Leinen. Finally, many thanks are addressed to B. Plattner for his generous support.

REFERENCES

1. J. P. Bailey, “The Economics of Internet Interconnection Agreements,” in *Internet Economics,* L. McKnight and J. P. Bailey, Eds., MIT Press, Cambridge, Massachusetts, pp. 155–168, 1997.
2. Y. Bernet, J. Binder, S. Blake, M. Carlson, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, and W. Weiss, “A Framework for Differentiated Services,” Internet Draft, October 1998.
3. D. Bertsekas and R. Gallager, *Data Networks,* Prentice Hall, Englewood Cliffs, New Jersey, 1987.
4. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, “An Architecture for Differentiated Services,” Internet Engineering Task Force, RFC 2475, December 1998.
5. J. Borland, “Rivalry in ISP Market Heats Up,” CMP’s Tech Web, *Technology News,* http://www.techweb.com/wire/story/TWB19980101S0006, January 1998.
6. A. Bouch and M. A. Sasse, “It Ain’t What You Charge, It’s the Way that You Do It, A User Perspective of Network QoS and Pricing,” in *Proceedings of 6th IFIP/IEEE International Symposium on Integrated Network Management (IM’99),* Boston, Massachusetts, May, pp. 639–654, 1999.
7. R. Braden, D. D. Clark, and S. Shenker, “Integrated Services in the Internet Architecture, An Overview,” Internet Engineering Task Force, RFC 1633, June 1994.
8. R. Braden, S. Berson, S. Herzog, and S. Jamin, “Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification,” Internet Engineering Task Force, RFC 2205, September 1997.
9. S. Bradner, “Internet Protocol Quality of Service Problem Statement,” Internet Draft, September 1997.
10. B. Briscoe, Ed., “Architecture, Part I Primitives & Compositions,” M3I Deliverable 2, Version 1, June 2000.
11. B. Briscoe, “The Direction of Value Flow in Connectionless Networks,” 1st International Workshop on Networked Group Communication (NGC’99), Pisa, Italy, November 1999.
12. B. Briscoe, M. Rizzo, J. Tassel, and K. Damianakis, “Lightweight Policing and Charging for Packet Networks,” 3rd IEEE Conference on Open Architectures and Network Programming (OpenArch 2000), Tel Aviv, Israel, March 2000.
13. N. Brownlee, “Traffic Flow Measurement, Meter MIB,” RFC 2064, Internet Engineering Task Force, January 1997.
14. N. Brownlee and A. Blount, “Accounting Attributes and Record Formats,” RFC 2924, Internet Engineering Task Force, September 2000.
15. N. Brownlee, C. Mills, and G. Ruth, “Traffic Flow Measurement, Architecture,” RFC 2063, Internet Engineering Task Force, January 1997.
16. G. Carle, M. Smirnov, and T. Zseby, “Charging and Accounting Architectures for IP Multicast Integrated Services over ATM,” 4th International Symposium on Interworking (Interworking’ 98), Ottawa, Canada, July 1998.
17. G. Carle and M. Smirnov, “Charging and Accounting for Value-added IP Services,” QoS Summit’99, Paris, France, November 1999.
18. J. Case, M. Fedor, M. Schoffstall, and J. Davin, “A Simple Network Management Protocol (SNMP),” RFC 1157, *Internet Engineering Task Force,* May 1990.
19. D. Chaum, “David Chaum on Electronic Commerce, How Do You Trust Big Brother?” *IEEE Internet Computing,* vol. 1, no. 6, pp. 8–16, November/December 1997,.
20. K. Chu, “User Reactions to Flat Rate Options under Time Charges with Differentiated Quality of Access, Preliminary Results from INDEX,” International Workshop on Internet Service

Quality Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts, December 1999.

21. S. A. Cotton, Ed., “Network Data Management—Usage (NDM-U) for IP-Based Services,” *IPDR Specification Version 1.1,* June 2000.
22. Cisco Systems, *Service Providers New World News—Pricing for Profitability,* Packet, Cisco Systems, pp. 35–39, 3rd Quarter 2000.
23. Cisco Systems, “Accounting and Billing Technology,” URL, http://www.cisco.com/warp/ public/cc/so/cuso/sp/sms/acct/index.shtml, January 2001.
24. D. Clark, “Combining Sender and Receiver Payments in the Internet,” 24th Telecommunications Research Policy Conference, Solomon’s Island, Maryland, October 1996.
25. D. Clark and W. Fang, “Explicit Allocation of Best Effort Packet Delivery Service,” Tech. Rep., Massachusetts Institute of Technology, Cambridge, Massachusetts, 1997.
26. R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, “Pricing in Computer Networks, Motivation, Formulation and Example,” *IEEE/ACM Transactions on Networking,* vol. 1, no. 6, pp. 614–627, December 1993.
27. R. Denda, A. Banchs, and W. Effelsberg, “The Fairness Challenge in Computer Networks,” LNCS, Springer-Verlag, Heidelberg, Germany, vol. 1922, 2000, pp. 208–220.
28. G. Dermler, G. Fankhauser, B. Stiller, T. Braun, M. Günter, H. Kneer, “Approaches for the Integration of Charging and Accounting into Internet Models and VPN,” CATI Project Deliverable, CATI-IBM-DN-P-004-1.0, July 1999.
29. R. Edell, N. McKeown, and P. P. Varaiya, “Billing Users and Pricing for TCP,” *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 7, pp. 1162–1175, July 1995.
30. R. Edell and P. P. Varaiya, “Providing Internet Access, What We Learn from the INDEX Trial,” Keynote Talk, Infocom’99, New York, March 1999.
31. European Telecommunications Standardization Institute (ETSI), “Parameters and Mechanisms Provided by the Network Relevant for Charging in B-ISDN,” ETR 123 Rev. 1, October 1995.
32. European Telecommunications Standardization Institute (ETSI), “Considerations on Network Mechanisms for Charging and Revenue Sharing,” Draft DTR/NA 010040, Version 10, October 1997.
33. European Telecommunications Standardization Institute (ETSI), “Internet Protocol (IP) based Networks,” Parameters and Mechanisms for Charging,” ETSI TR 101 734 V.1.1.1, Sophia Antipolis, France, September 1999.
34. M. Falkner, M. Devetsikiotis, and I. Lamdadaris, “An Overview of Pricing Concepts for Broadband IP Networks,” *IEEE Communications Surveys,* pp. 2–13, 2nd Quarter 2000.
35. M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On Power-Law Relationships of the Internet Topology,” *ACM Computer Communication Review (SIGGCOMM’99),* vol. 29, no. 4, pp. 251–261, 1999.
36. G. Fankhauser, B. Stiller, C. Vögtli, and B. Plattner, “Reservation-based Charging in an Integrated Services Network,” 4th INFORMS Telecommunications Conference, Boca Raton, Florida, March 1998.
37. D. Ferrari and L. Delgrossi, “Charging for QoS,” 6th IEEE/IFIP International Workshop on Quality-of-Service (IWQoS’98), Napa, California, May 1998.
38. P. Flury, P. Reichl, J. Gerke, and B. Stiller, “Charging Considerations for Virtual Private DiffServ Networks,” TIK Report No. 94, Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland, August 2000.
39. D. K. Foley, “Resource Allocation in the Public Sector,” *Yale Economic Essays,* vol. 7, 1967, pp. 73–76.
40. C. Gadecki, “Usage Bills, Easier Said Than Done,” *tele.com Magazine,* November 1997.

41. R. Gardner, *Games for Business and Economics,* John Wiley & Sons, New York, 1995.

42. R. Gibbens and P. Key, "Distributed Control and Resouce Pricing," ACM SIGCOM, Stockholm, Sweden, August 2000.

43. M. A. Gibney and N. R. Jennings, "Dynamic Resource Allocation by Market Based Routing in Telecommunication Networks," in *Proceedings of 2nd International Workshop on Intelligent Agents for Telecommunication Applications (IATA'98),* vol. 1437, S. Albayrak and F. S. Garijo, Eds., Springer-Verlag, Heidelberg, Germany, pp. 102–117, 1998.

44. S. E. Gillett, "Connecting Homes to the Internet, An Engineering Cost Model of Cable vs. ISDN," Massachusetts Institute of Technology, Cambridge, Massachusetts, 1995.

45. P.-L. de Guillebon, "Wholesaling for Survival and Growth," *Telecommunications, International Edition,* vol. 34, no. 2, February 2000.

46. A. Gupta, B. Jukic, D. O. Stahl, and A. B. Whinston, "Impact of Congestion Pricing on Network Infrastructure Investment," International Workshop on Internet Service Quality Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts, December 1999.

47. A. Gupta, D. O. Stahl, and A. B. Whinston, "Managing the Internet as an Economic System," CISM, University of Texas at Austin, July 1994.

48. G. Hardin, "The Tragedy of the Commons," *Science,* no. 162, pp. 1243–1247, 1968.

49. A.L.G. Hayzelden and J. Bigham, "Heterogeneous Multi-Agent Architecture for ATM Virtual Path Network Resource Configuration," in *Proceedings of 2nd International Workshop on Intelligent Agents for Telecommunication Applications (IATA'98),* Vol. 1437, S. Albayrak and F. S. Garijo, Eds., Springer-Verlag, Heidelberg, Germany.

50. H.-G. Hegering, S. Abeck, and B. Neumair, *Integrated Management of Networked Systems,* Morgan Kaufmann Publishers, San Francisco, California, 1999.

51. S. Herzog, S. Shenker, and D. Estrin, "Sharing the 'Cost' of Multicast Trees, An Axiomatic Analysis," *IEEE/ACM Transactions on Networking,* vol. 5, no. 6, pp. 847–860, December 1997.

52. G. Huston, *ISP Survival Guide,* Wiley, New York, 1999.

53. G. Huston, "Interconnection, Peering, and Settlements," The Internet Summit (INET'99), San Jose, California, 1999.

54. ITU-T D-Series Recommendations, "General Tariff Principles—Definitions," D.000, Geneva, Switzerland, March 1993.

55. ITU-D.260, "General Tariff Principles, Charging and Acounting International Telecommunications Services—Charging and Accounting Capabilities to be Applied on the ISDN," Recommendation D.260, Geneva, Switzerland, March 1991.

56. ITU-D Suppl3, "General Tariff Principles, Supplement 3, Handbook on the Methodology for Determining Costs and Establishing National Tariffs," Recommendation D-Series, Supplement 3, Geneva, Switzerland, March 1993.

57. ITU-T E.800, "Quality-of-Service and Dependability Vocabulary," Recommendation E.800, Geneva, Switzerland, November 1994.

58. ITU-T Q.825, "Specification of TMN Applications at the Q3 Interface, Call Detail Recording," Recommendation Q.825, Geneva, Switzerland, 1998.

59. M. Karsten, *QoS Signaling and Charging in a Multi-service Internet Using RSVP,* Ph.D. Thesis, University of Technology, Darmstadt, Germany, July 2000.

60. M. Karsten, Ed., *Pricing Mechanisms Design (PM),* M3I Deliverable 3, Version 1.0, June 2000.

61. M. Karsten, J. Schmitt, B. Stiller, and L. Wolf, "Charging for Packet-switched Network Communications—Motivation and Overview," *Computer Communications,* vol. 23, no. 3, pp. 290–302, February 2000.

62. M. Karsten, J. Schmitt, L. Wolf, and R. Steinmetz, "An Embedded Charging Approach for

RSVP," 6th IEEE/IFIP International Workshop on Quality of Service (IWQoS'98), Napa, California, May 1998.

63. F. Kelly, "Models for a Self-managed Internet," Presented at the Department of Economics, University of Southhampton, U.K., Network Modelling in the 21st Century, The Royal Society, London, U.K., December 1999.
64. F. Kelly, A. Maulloo, and D. Tan, "Rate Control for Communication Networks—Shadow Prices, Proportional Fairness, and Stability," *Journal of the Operational Research Society,* vol. 49, pp. 237–252, 1998.
65. S. Keshav, *An Engineering Approach to Computer Networking,* Addison-Wesley, Reading Massachusetts, 1997.
66. P. Key and R. Gibbens, "The Use of Games to Assess User Strategies for Differential Quality of Service in the Internet," International Workshop on Internet Service Quality Economics (ISQE'99), Massachusetts Institute of Technology, Cambridge, Massachusetts, December 1999.
67. P. Key and L. Massoulié, "User Policies in a Network Implementing Congestion Pricing," International Workshop on Internet Service Quality Economics (ISQE'99), Massachusetts Institute of Technology, Cambridge, Massachusetts, December 1999.
68. P. Key, D. McAuley, P. Barham, K. Laevens, "Congestion Pricing for Congestion Avoidance," MSR-TR-99-15, *Microsoft Research, Cambridge, England,* February 1999.
69. K. Kilkki, *Differentiated Services for the Internet,* Macmillan Technology Series, Indianapolis, Indiana, 1999.
70. H. Kneer, U. Zurfluh, and C. Matt, "Business Model (RSVP)," Public CATI Deliverable CATIUZH-DN-P-001-2.0, March 1999.
71. H. Kneer, U. Zurfluh, G. Dermler, and B. Stiller, "A Business Model for Charging and Accounting of Internet Services," 1st International Conference on Electronic Commerce and Web Technologies, Greenwich, U.K., September 42000.
72. A. Kuiper, "Charging Methodologies for ATM, An Introduction," Cap Gemini, The Netherlands, August 1997.
73. K. Kuwabara, T. Ishida, Y. Nishibe, and T. Suda, "An Equilibratory Market-based Approach for Distributed Resource Allocation and its Application to Communication Network Control," in *Market-Based Control, A Paradigm for Distributed Resource Allocation,* E. Clearwater, Ed., pp. 53–73, 1996.
74. K. S. Lee and D. Ng, "Billing for Interconnect," *Telecommunications, International Edition,* vol. 33, no. 11, pp. 81–82, November 1999.
75. B. A. Leida, *Cost Model of Internet Service Providers, Implications for Internet Telephony and Yield Management,* Master Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, February 1998.
76. B. Liver and G. Dermler, "The E-Business of Content Delivery," International Workshop on Internet Service Quality Economics (ISQE'99), Massachusetts Institute of Technology, Cambridge, Massachusetts, December 1999.
77. J. MacKie-Mason, "A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network," Tech. Rep., University of Michigan, Ann Arbor, Michigan, September 1997.
78. J. MacKie-Mason and H. Varian, "Pricing Congestible Network Resources," *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 7, pp. 1141–1149, 1995.
79. J. MacKie-Mason, and K. White, "Evaluating and Selecting Digital Payment Mechanisms," 24th Telecommunications Policy Research Conference, Solomon's Island, Maryland, October 1996.
80. D. Mitton, M. St. Johns, S. Barkley, D. Nelson, B. Patil, M. Stevens, and B. Wolff,

"Authentication, Authorization, and Accounting, Protocol Evaluation," Internet Draft, October 2000.

81. L. W. McKnight and B. A. Leida, "Internet Telephony, Costs, Pricing, and Policy," 25th Telecommunications Policy Research Conference, Alexandria, Virginia, September 1997.
82. N. McIntosh, "Heavy Surfers Pay the Price," *The Guardian,* London, August 3, 2000.
83. D. Morris and V. Pronk, "Charging for ATM Services," *IEEE Communications Magazine,* vol. 37, no. 5, pp. 133–139, May 1999.
84. A. J. Mund, "AOL's Breakdown, A Harbinger for the Internet's Future?" *Cable TV and New Media—Law and Finance,* vol. 14, no. 12, February 1997.
85. M3I, "Market Managed Multi-Service Internet," EU 5th Framework ISP Project 11429, URL, http://www.m3i.org, January 2001.
86. D. N. Newbery, *Privatization, Restructuring, and Regulation of Network Utilities,* The MIT Press, Cambridge, Massachusetts, 1999.
87. K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," RFC 2474, Internet Engineering Task Force, December 1998.
88. K. Nichols and B. Carpenter, "Definition of Differentiated Services Per-Domain-Behaviors and Rules for their Specification," Internet Draft, June 2000.
89. K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," RFC 2638, Internet Engineering Task Force, July 1999.
90. M. Nilsson, "Third-Generation Radio Access Standards," *Ericsson Review,* no. 3, pp. 110–121, 1999.
91. A. M. Noll, "Internet Pricing Versus Reality," *Communications of the ACM,* vol. 40, no. 8, pp. 118–121, August 1997.
92. A. Odlyzko, "Paris Metro Pricing, The Minimalist Differentiated Services Solution," in *Proceedings of 7th IEEE/IFIP International Workshop on QoS (IWQoS'99),* London, pp. 159–161, June 1999.
93. "Optimum, Broadband Access Networks," *Telektronikk,* vol. 95, no. 2/3, Telenor, Norway, 1999.
94. K. Park, M. Sitharam, and S. Chen, "Quality-of-Service Provision in Noncooperative Networks, Heterogeneous Preferences, Multi-Dimensional QoS Vectors, and Burstiness," in *Proceedings of 1st International Conference on Information and Computation Economies (ICE-98),* pp. 111–127, 1998.
95. C. Partridge, "A Proposed Flow Specification," RFC 1363, Internet Engineering Task Force, September 1992.
96. C. Partridge, "Using the Flow Label Field in IPv6," RFC 1809, Internet Engineering Task Force, June 1995.
97. B. Plattner, B. Stiller, and D. Bauer, "High-Performance Networks—Applications and Technology," Industry Course, Zürich, Switzerland, June 2000.
98. H. Petersen, D. Konstantas, M. Michels, D. Som, G. Fankhauser, D. Schweikert, B. Stiller, N. Weiler, R. Cantini, and F. Baessler, "MicPay, Micro-payments for Correlated Payments," SI Informatik/Informatique, No. 1, Switzerland, January 2000.
99. K. Ramakrishnan and S. Floyd, "A Proposal to add Explicit Congestion Notification (ECN) to IP," RFC 2481, Internet Engineering Task Force, January 1999.
100. P. Reichl, P. Kurtansky, J. Gerke, and B. Stiller, "The Design of a Generic Service-oriented Cost Model for Service Providers in the Internet (COSMOS)," Applied Telecommunication Symposium 2001 (ATS'01), Seattle, Washington, April 2001.
101. P. Reichl and B. Stiller, "Notes on Cumulus Pricing and Time-scale Aspects of Internet Tariff Design," TIK Report No. 97, Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland, November 2000.

102. L. G. Roberts, "Beyond Moore's Law, Internet Growth Trends," *IEEE Computer,* vol. 33, no. 1, pp. 117–119, January 2000.

103. F. C. Rö hmer, "Charging Information Management and its Technical Implication in a Liberalized Broadband Telecommunications Environment," *SI Informatik/Informatique, Switzerland,* no. 3, pp. 37–38, 1999.

104. S. Salsano, F. Ricciato, M. Winter, G. Eichler, A. Thomas, F. Fünfstück, T. Ziegler, and C. Brandauer, "Definition and Usage of SLSs in the AQUILA Consortium," Internet Draft, November 2000.

105. *SET, Secure Electronic Transactions,* URL, http://www.setco.org, January 2001.

106. S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 7, pp. 1176–1188, Sept. 1995.

107. S. Shenker, "Service Models and Pricing Policies for an Integrated Services Internet," in *Public Access to the Internet,* J. Keller B. Kahin, Eds., Prentice Hall, Englewood Cliffs, New Jersey, 1995, pp. 315–337.

108. S. Shenker, D. Clark, D. Estrin, and S. Herzog, "Pricing in Computer Networks, Reshaping the Research Agenda," *ACM Computer Communication Review,* vol. 26, no. 2, pp. 19–43, April 1996.

109. S. Shenker, G. Partridge, and R. Guerin, "Specification of Guaranteed Quality-of-Service," RFC 2212, Internet Engineering Task Force, September 1997

110. A. Sirbu, "Internet Billing Service Design and Prototype Implementation," Carnegie Mellon University, Computer Science Department Tech. Rep., Pittsburgh, Pennsylvania, 1994.

111. SIU, "Smart Internet Usage," Hewlett-Packard, http://communications.hp.com/smartinternet/, January 2001.

112. D. Songhurst, Ed., *Charging Communication Networks—From Theory to Practice,* Elsevier Publisher, Amsterdam, The Netherlands, 1999.

113. B. Stiller, "QoS Methods for Managing Communicating Applications," *Journal on Integrated Computer-Aided Engineering,* vol. 6, no. 2, pp. 159–169, February 1999.

114. B. Stiller, "Pricing and Charging of Packet-based Networks," Tutorial T12, IEEE/IFIP Network Operations and Management Symposium (NOMS 2000), Honolulu, Hawaii, April 2000.

115. B. Stiller, T. Braun, M. Günter, and B. Plattner, "The CATI Project—Charging and Accounting Technology for the Internet," LNCS, Springer-Verlag, Heidelberg, Germany, vol. 1629, 1999, pp. 281–296.

116. B. Stiller, G. Fankhauser, G. Joller, P. Reichl, and N. Weiler, "Open Charging and QoS Interfaces for IP Telephony," The Internet Summit (INET'99), San Jose, California, June 1999.

117. B. Stiller, G. Fankhauser, and B. Plattner, "Charging of Multimedia Flows in an Integrated Services Network," in *Proceedings of 8th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'98),* Cambridge, England, pp. 189–192, July 1998.

118. B. Stiller, G. Fankhauser, N. Weiler, and B. Plattner, "Charging and Accounting for Integrated Internet Services - State of the Art, Problems, and Trends," The Internet Summit (INET '98), Geneva, Switzerland, July 1998.

119. B. Stiller, J. Gerke, and P. Flury, "The Design of a Charging and Accounting System," TIK Rep. No. 94, Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland, July 2000.

120. B. Stiller, J. Gerke, P. Reichl, and P. Flury, "The Cumulus Pricing Scheme and its Integration into a Generic and Modular Internet Charging System for Differentiated Services," TIK Rep. No. 96, Computer Engineering and Networks Laboratory TIK, ETH Zürich, Switzerland, September 2000.

121. B. Stiller, J. Gerke, P. Reichl, and P. Flury, "Management of Differentiated Service Usage by

the Cumulus Pricing Scheme and a Generic Internet Charging System," Seventh IEEE/IFIP Symposium on Integrated Network Management (IM 2001), Seattle, Washington, May 2001.

122. B. Stiller, P. Reichl, and S. Leinen, "A Practical Review of Pricing and Cost Recovery for Internet Services," *Netnomics—Economic Research and Electronic Networking,* vol. 3, no. 1, March 2001.

123. H. Tö bben, *Concept of a Market-based Routing for ATM-Networks on the Basis of Intelligent Agents,* Ph.D. Thesis (in German), TU-Berlin, DAI-Labor, Germany, April 2000.

124. H. R. Varian, *Intermediate Microeconomics—A Modern Approach,* W. W. Norton & Company, New York, 4th edition, 1996.

125. W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *The Journal of Finance,* vol. 16, pp. 8–37, 1961.

126. D. Walker, F. Kelly, and J. Solomon, "Tariffing in the new IP/ATM Environment," *Telecommunications Policy,* vol. 21, pp. 283–295, May 1997.

127. Q. Wang, J. Peha, and M. Sirbu, "Optimal Pricing for Integrated-Services Networks with Guaranteed Quality of Service," in *Internet Economics,* L. McKnight and J. P. Bailey Eds., MIT Press, Cambridge, Massachusetts, pp. 353–376, 1997.

128. M. Webster, *Merriam-Webster's Collegiate Dictionary,* Merriam-Webster, Inc., Springfield, Massachusetts, 10th Edition, 1996.

129. J. Wroclawski, "Specification of the Controlled-Load Network Element Service," RFC 2211, Internet Engineering Task Force, September 1997.

130. L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP, A New Resource Reservation Protocol," *IEEE Networks Magazine,* vol. 31, no. 9, pp. 8–18, September 1993.