A. Wolisz, "Information Access is Fine, But Who is Going to Pay?", In K. Zielinski, K. Geihs, and A. Laurentowski, editors, New Developments in Distributed Applications and Interoperable Systems, pp. 149-160 Kluwer, September 2001.

Information access is fine, but who is going to pay? The dual charging approach….[1]

(an extended Abstract)

Adam Wolisz
Technical University of Berlin, Telecommunication Networks Group

http://www-tkn.ee.tu-berlin.de

Key words: Internet, Mobile Computing, Wireless Access, Charging

Abstract: Modern telecommunication networks promise new perspective in information access at any time, from any place. On the other hand recent spectrum auctions made it clear, that the transmission capacity (at least for wireless access!) becomes really an economic issue. In this talk we advocate the thesis, that both flat rate and usage based charging of the end-users are structurally wrong, and discuss an alternative dual charging model.

Generally speaking, we are used to the fact that if we use something we have to pay for it. And we are fairly aware that there are different business models driving manufacturers and service providers, so that we can expect different charging principles. With the explosive growth of Internet - which seems to evolve to a universal platform for information access and dissemination, the basic question of how the charging model for internet should look like becomes an essential one.

In the evolution of Internet three phases can be defined as far as charging principles are concerned. In the early phase, due to governmental support for the new, evolving packet network technology, free usage has been natural. In

[1] This is a modified version of a talk given during the Third Berlin Internet Economy Workshop, Mai 2001, Berlin, Germany.

the second phase, the principle of flat rate charging has been introduced: each user subscribing the access has is charged equally, independent from his/her real usage. The reasoning behind this model has been multifold. On one hand, it seemed natural that something which used to be covered in equal shares by all taxpayers, will be in the second phase covered in equal shares by those who use it, and in the US, were the initial Internet deployment took place, there has been a long, good tradition of flat rate charging for information transfer (the local phone calls). On the other hand, the content reached by the internet used to be free, so it would make no sense to differentiate on the basis of the services used. Last, but not least, in the Internet infrastructure there has been no mechanisms which might have supported a different charging schema.

With the rapidly increasing volume of data, but also with the increasing spectrum of services offered via the internet, sincere doubts have been expressed concerning the soundness of flat-rate charging. The arguments for and against flat rate vs. the alternative volume-charging have been mostly of intuitive nature. Sure, nobody cares about economic use of something which is charged usage independent. Sure, if the total costs of the infrastructure increases (because of the strongly increased traffic volume) usage-based charging seems to be more fair. However, it is also possible that the costs of developing an infrastructure for supporting usage-based charging might be higher than the gains of more economical usage. Only recently some studies (eg. [RuEC98] - the Berkeley INDEX project) demonstrated the inefficiency of the flat-rate charging and provided founded arguments for the usage-based charging.

In parallel, recent trends for changing from the single best-effort service model towards some kind of QoS support (Integrated Services or Differentiated Services) will make it anyway necessary to diversify charges in different QoS classes (otherwise there would be no incentive at all for NOT using the highest service class!). So we approach a third period - the usage based charging era.

But - as mentioned before - at the moment there do not exist any efficient mechanisms for accounting and charging in the internet infrastructure. This gave a reason for quite a momentum in research - see eg. [FSVP98], [BrRT99], [CaHS99]. The importance of this topic has been recently recognized also by the IRTF were a special working group [AAAArch] has been commissioned. We believe strongly, that BEFORE real effort will be

invested in development of accounting mechanisms, there is an essential need to discuss the basic philosophy of the future usage based charging model. This paper is intended as a contribution to such discussion. In the following we will first define a simplified view of the Internet used for our reasoning. Afterwards we will discuss a straightforward option in usage based charging: charging for transferred information volume. We will express our criticism towards this approach, and argue for charging per end-user relevant service as the proper model from the end-user perspective. Finally we will formulate and describe our suggested dual model.

Our view of the Internet

Further in this paper we will constrain ourselves to the following, simplified view of the internet, which we believe is a quite realistic one.

There are several **packet transfer providers** (PTPs), $P_{i,}$ I = 1,2, ..M each offering transfer of IP packets in different QoS (Quality of Service) classes. C_{ij} , j= 1,2,...N. Let us stress that the precise semantic of the quality delivered in each of the classes is irrelevant to our discussion, and that the semantic of the classes does NOT have to be identical in the case of different providers - i.e. semantic of the class C_{11} might be different from the semantic of C_{21}!.

We identify, as a special important role, **end users**. For the sake of this paper, we will associate an end-user temporarily with a single device (this association might be modified in time!). Let us also assume that this device has an Internet connectivity, which however does not necessarily imply that it has an own, permanent IP address. In fact, we cannot even assume, that the end-user's device will be sending/receiving IP packets, as usage of special proxies might be quite attractive (see for example [WOL00], this issue will be elaborated in the talk in more detail).

The internet connectivity is assured via an ISP (this might be also an on-campus or company internal network). Information transmission form end-user to the ISP will be assured by the use of some access infrastructure. The owner of access infrastructure might be different from the ISP (although there is a strong tendency for the access infrastructure owners to offer the ISP functionality), thus the end user might have a choice of several ISP over the single access infrastructure. On the other hand, a single device might have a choice of several access infrastructure variants (e.g. telephone line,

TV Cable, and wireless packet access). Each access infrastructure will be characterized usually by the mode of operation (permanent access, switched access) as well as the supported bit-rates, and possibly other quality of transfer parameters. On top of this ISPs might offer also different Quality of Service of Internet Access. In fact ISPs make available to the end-user devices the packet transfer services delivered by PTPs.

End users access Internet in order to use some services and utilities. In order to keep the generality, and avoid the multiple usage of the frequently misused word services, we will introduce the term: **servilities** to describe any kind of service which an end user might trigger via the Internet. For the sake of further discussion we will refer to a single servility, as to something which, from the point of view of the end-user is complete, and satisfies his specific need. A servility might be short or long, in fact we believe that servilities will include, for example such different items as: playing a movie, downloading (or playing!!) an MP3 song, a money transfer, booking a flight, phone call, waking in the morning, permanently observing the courtyard... and many, many others... Servilities are made available to the end-user by **serviders** (servility providers). We introduce this term as a more general than just content providers, or service providers. This should become clear after understanding the term servility.

WHAT DOES IT MEAN FLAT CHARGING? TIME BASED CHARGING?

After having introduced our vision of the internet, we can shortly elaborate on the notion of flat charging. In fact, in general we have to differentiate between charging for the use of access infrastructure and charging for the use of packet transfer service. Under the notion of flat charging both of this charges might be included (say charging for the CATV based Interent access or RICOCHET wireless packet access in the Bay Area). Alternatively, universities usually offer to the students free packet transfer, but charges for long distance modem transmission (the access infrastructure) have to be paid by the user, on a per minute basis. In Europe time dependent charging - namely a fixed charge per time unit for both: access infrastructure and unlimited packet transfer service - is widely deployed for the switched access type (phone or ISDN access) regardless of usage or non usage of any servilities in this time.

Although it has never been stated clearly, flat charging in the case of shared media is reasonable only in the case of servilities which are rather short in time, and elastic (meaning that the offered load is dependent on the feedback). Sure, a wireless packet network (like Ricochet) would break down under a traffic generated, for example by surveillance systems with permanently sending Web cameras.

So, what about charging by volume?

One, rather frequently mentioned alternative to the flat-rate charging is volume based charging ([Brow94], [EdMV95]) meaning that the user will be charged depending on the amount of data being passed to/from end-user device. This approach can be in a very straightforward way extended for the case of different QoS classes (independent of their definition), by differently charging the volume of data passed in each of the priority classes. This is a direct analogy to charging for energy - by counting the amount of electricity taken in the different charging times (day, night). The advantage of this schema is an inherent possibility given to the end-user to verify the measurements (as the flow has to be generated in the end-system. And the ISP could - theoretically - compute the sum of the services obtained from the PTPs by adding the flows of individual end users.

We claim that such simple volume-based charging is not a proper one. Let us give some examples for this- even for the simplest case, when the end-system has a full protocol stack and IP packets are passed directly from the end-user device over the access network. In this case, the most natural approach would be counting the bytes as seen at the IP level.

Example A: Let our servility be the download of medical images. Let us assume that we attempt to download a rather huge image - say an X-Ray image. The volume of data would be dependent on the coding schema (Changing the coding may change the volume of data significantly, thus changing the charge proportionally). Why should the inefficient - or possibly just stupid! selection of a data format by the servider cause unavoidable costs for the user? (sure, your doctor, affiliated with the local hospital will not switch to another hospital only because they have a cleverer data processing guy...he will simply charge you more for your treatment.)

Example B. Electronic Payment. A user is not interested how many bits, bytes or gigabytes are transferred over the internet in order to support his

payment. What he is interested in is security and timeliness of the transfer. In fact it would be highly annoying to the user, if - because of some technical reasons which might influence the amount of data used for individual transfers (like, say repeated authentication !) he would be asked to pay different amount of money for identical transactions. One could in fact imagine even a worse case: if due to errors a financial transaction cannot be completed at all, user would pay for data movement which did not have any value for him!

Example C: It is obvious, that for reliable transmission X Mbytes of USER data we will in fact transmit really some Y>X Mbytes. In fact, the INEFFICIENCY of the reliable transmission can be expressed as (Y-X)/Y. Let us note that charging by Y is the most unfair option: in fact providers with HIGH error/loss rate will have high INEFFICIENCY, thus requesting higher charges. And- frankly - which user does look at the real amount of data being transferred? Who compares the costs of individual transfers?

Combining the observations of the three examples given above, the user would never know what is the reason of the high prize, and furthermore, the prize will not be predictable! The situation becomes even worse if the end-user device is connected to the internet over a kind of proxy - say the WAP architecture. In this case data are converted at some intermediate server, in order to adjust at low-bit-rate access infrastructure. The amount of data downloaded to the end-system is by definition essentially smaller to the amount of data moved by the PTP. So a volume metering function, with measurements per end user! on the „backbone side“ of the converters is needed. This does definitely scale much worse than the measurement at each end system. And in addition the result is unverifiable from the end system.

Let us have a look even one step further. If the end-user device is a mobile one, and uses temporary IP addresses, accounting can no longer be based solely on IP address of the end-device. Instead, additional complexity of the accounting system, and coupling of the accounting with authorization is unavoidable ([AAArch]). This makes the simplicity questionable.

THE DUAL APPROACH TO INTERNET CHARGING[2]

Following these examples we advocate the position, that a new approach to internet services charging is needed. In fact, we argue that the end user is interested in the quality of application services - like the quality of the video which she is going to look at. Or obtaining an MP-3 Song. The end user might be ready to pay for a timely and high quality service, but not for the amount of data being shuffled in the backbone, which - by the way - is not transparent to them, and cannot be influenced by them. We also argue that the user is interested in having fixed, settled IN ADVANCE charges for successfully completed servilities - possibly offered under competitive conditions by several competing serviders. Or - frequently - offered free of charge for a specific class of users (customers, holders of a gold/platinum card, club members). On the other hand, the work done by the PTPs should be charged quite differently

Thus we recognize a duality in the intended charging approaches and suggest a novel approach based on two different charging schemata.:

- End users should be charged by serviders only on the basis of servilities used, independently of the volume of data, because only completing a servility has some value for the user. This is - by the way- consistent with the classical end-to-end argument. Charging schemata must be relevant to the APPLICATION SERVICE SEMANTICS (APS) and charged by actions - servilities - not volume of data, possibly unnecessary, or technology specific - generated at lower layers.

- PTPs (and possibly ISPs) should charge serviders for the amount of data (in each QoS class) transferred on their behalf, independently of the fact, for which servility, and for which customer, the transfer took place. In fact the PTPs are not interested who, and why generates the traffic they carry.

We will refer to this concept as to dual approach to charging. Let us explain this concept in more detail, and discuss the impact of such an approach on the design of serviders, as well as accounting and charging mechanisms.

[2] Recently the author has discovered, that a similar idea has been independently developed by researchers at IBM Zürich [LiDe99]

Serviders - and only serviders!- can develop different, possibly quite complex charging policies for the end-users. One can imagine for this case the following approaches:

- the game machine model - an end user is charged for a couple of minutes of using the game machine)

- the sushi-bar model - one is charged for the retrieving of objects (an MP3-song, a movie)

- the „all you can eat“ approach: for entering a session - eg. chatting, ...until you deregister...

These models may be combined. We can imagine for example a „game model“ charging of the search machine (which you might use to search for, say, recording of theater performances or movies including 10 second samples) combined with the „sushi-bar“ charging for downloading the copies of the downloaded objects (the movies themselves). Which would be, by the way, similar to the pay TV in the hotels....

But charging by servilities makes also very straightforward more complex models, like: members of a society have free access to some subset of the journals and reports, but have to pay for another ones. Or: each servility has a fixed charge. Each member obtains some free allowance per month (year), without limits of the servilities used. After exceeding the allowance further servilities will be charged.

It is obvious that this way of charging per servilities can be in a natural way coupled with granting access right to use the servilities, thus the proper accounting mechanisms can (and should) be tightly coupled with the access and authentication mechanisms. Not necessarily contributing to an essential increase of their complexity. We expect that support for this kind of charging will be also increasingly provided in the WWW browsers, so that a prize of - say downloading an object, or starting an action might be presented in advance to the user.

Quite a different situation is to be considered for charging the packet transfer services by PTP (and possibly ISPs). As those services are to be covered by the serviders, there is no need to care for the high resolution of the carried flows, down to the identification of the end-user and counting IP packets or

even bytes per end- user.. Really meaningful is only the servider responsible for the traffic. An example for this might be the whole traffic of an ISP, or a whole traffic of some electronic commerce service provider. As this traffic is rather highly aggregated, there is no real need to go for a very high resolution of accounting: we will care rather for Megabits or Gigabits exchanged in a given service class. In fact totally different charging models apply. We expect typically the servider to make a contract with a PTP, assuring him the right to put on the PTP in some time frame, up to an agreed amount of traffic following a fixed traffic profile in each of the QoS classes suppoted. The charge might be computed from a component based on the negotiated upper limit as well as a component based on real usage. As for the real usage we conjecture that statistical sampling methods can effectively be applied, rather than precise counting of bytes or packets. We are quite convinced that only usage of statistical sampling methods will is a feasible way to determine the amount of data transferred by the PTPs. In fact development of such methods, which would provide an assesment with some predetermined accuracy, for individual traffic classes seems to be an interesting challenge, which we will investigate in detail separately.

As for the access technology charges, those might be alternatively carried by the servider on behalf of his customers, or covered by a flat rate by the end-user.

Last but not least: the dual charging approach delivers incentives for technological progress and competition both among the PTPS and the serviders. As for the PTPs the situation is clear: by creating competition on the charges for transport of just IP packets (possibly with different Quality of Service classes) without coupling to services, we create a situation similar to the market of long distance calls in US. Which proved to be very advantages for the customers, and has lead to very low tariffs in the US as compared, for example, to the situation in Europe.

As for the serviders the situation is even more convincing. Serviders have to compete only in the quality and charges for the offered servilities (sure, different composed blocks of servilities might be offered - like insurances are offered now). In order to remain competitive, in spite of equal access to the PTPs, serviders will have to optimize the cost of the infrastructure used for providing servilities.

In fact serviders will have to consider technical implication of mapping the user-defined operation into (a possibly complex) sequence of data transfers in different QoS classes, into possible caching, replicating etc. . And, based on how successful a given servider will be, he will be in the position to offer more or less attractive prizes to the end-user. In fact the end-user will most probably prefer obtaining an MP3 song for 50 cents from provider A rather than for one dollar from provider B. And we will not accept the 1 dollar offer with the explanation that this is justified by the higher volume of data used by provider B....because of his less innovative technical solution.

Taking the example of search engines: looking for a keyword with different search engines might cause totally different traffic, depending on the location of the engines themselves (possibly replicated) , the location of the data bases, and- last but not least - their efficiency (if you get the best match as the first one...).

Conclusions

We have presented a concept of charging for Internet services, that distinguishes (A) charging of the end user based on the end-to-end service (denoted the servility offered by a servider), and (B) charging of serviders based on packet transfer volume per QoS class.

The main attraction of this concept of charging, which we call the dual approach, is giving incentives for optimization of the information infrastructure, like considering the tradeoff between information replication vs. transport of information, information compression vs. uncompressed tranport, etc. The dual charging approach opens, however, quite a couple of research issues. We will list here only a few of them:

How should the individual end-user services use the different possible data communication classes? How can the end-user oriented charges for these services be derived from the data communication charges? If the charges structure for data communication services changes, will this have a direct consequence to the end-user-oriented charges, or will there be a large freedom for the provider concerning which structure of charge to use? How to support on-line information about the prizes of individual operations (a user interface extension?) How to organize efficiently the accounting?

Acknowledgements

I would express my gratitude to Dr. Carle, Dr. Smirnov and Tanja Zseby for numerous discussion during the development of this concept.

References

[AAAArch] IRTF Authentication Authorisation Accounting ARCHitecture Research Group (AAAARCH), URL: http://www.irtf.org/charters/aaaarch.html

[Brow94] Nevil Brownlee: New Zealand Experiences with Network Traffic Charging, ConneXions, Volume 8, No. 12, December 1994

[BrRT99] B. Briscoe, M. Rizzo , J. Tassel, K. Damianakis, N. Guba: Lightweight policing and charging for packet networks, BT Research, BT Labs, Martlesham Heath, 1999

[CaHS99] Georg Carle, Felix Hartanto, Michael Smirnow, Tanja Zseby: Charging and Accounting for QoS-enhanced IP Multicast, IFIP Sixth International Workshop on Protocols ForHigh-Speed Networks (PfHSN '99), August 25-27, 1999, Salem, MA, U.S.A.

[EdMV95] R. Edell, N. McKeown, P. Varaiya: Billing Users and Pricing for TCP, IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, September 1995, pp. 1162-1162.

[FSVP98] G. Fankhauser, B. Stiller, C. Vögtli, B. Plattner: Reservation-based Charging in an integrated services Network, 4th INFORMS Telecommunications Conference, Boca Raton, Florida, U.S.A., March 1998.

[LiDe99] B. Liver, G. Dermler: The E-Business of Content Delivery , Talk at Workshop on Internet Service Quality Economics, MIT, Dec. 2-3, 1999. http://www.marengoresearch.com/isqe/agenda_m.htm

[RuEC98] B. Rupp, R. Edell, H. Chand, P. Varaiya. INDEX: A Platform for Determining how People Value the Quality of their Internet Access. Proceedings of the Sixth IEEE/IFIP International Workshop on Quality of Service - IWQoS'98, Napa, CA, May 1998, pp. 85-90.

[Wol00] A. Wolisz. Wireless Internet Architectures: Selected Issues.. (invited paper) in: J.Wozniak, J. Konorski, editors "Personal Wireless Communications" Kluver Academic Publishers, Boston/Dordrecht/London 2000, pages 1-16