社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－ＭＢＬ－33（10）
2005／5／25

Evaluation of User-centric Performances to Select an Optimal Access Network in Heterogeneous Systems

Ved P. KAFLE †, Eiji KAMIOKA ‡ and Shigeki YAMADA ‡

† Department of Informatics, The Graduate University for Advanced Studies
‡ National Institute of Informatics
Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo, 101-8430 Japan
E-mail: † kafle@grad.nii.ac.jp, ‡ {kamioka,shigeki}@nii.ac.jp

Abstract - In recent years, wide varieties of wireless access networks supporting multimedia services are emerging with different characteristics. The service areas of many of these networks are overlapping so that a mobile user from an overlapped service area can access any network that supports the user application. A mobile user can take advantages of the availability of such heterogeneous multimedia networks only when the user terminal is equipped with a mechanism that can select an optimal network for the application. In this report, we present an analytical framework of such a network selection mechanism and discuss its implementation issues.

Key words: User-centric performance-cost analysis, optimal network selection mechanism, heterogeneous wireless networks.

1. Introduction

The next generation mobile communication system is expected to be a heterogeneous system comprising of widely different radio access networks. Each access network may possess some advantages over others in terms of network characteristics such as bandwidth, coverage, cost, reliability, and so on. To exploit these advantages, the heterogeneous system appears in an overlay form [1], [2]; one access network (e.g. wireless LAN) overlapping the service area of other access networks (e.g. 3G networks). In such an environment, a mobile host with multi-mode network interfaces should be capable of carrying out the following two functions: (1) select an optimal access network (2) transfer connection from one access network to others when previous one becomes sub-optimal or unavailable. In this report, we focus on the first function, i.e. the selection of an optimal network. For this purpose, we (a) define user-centric performance, (b) define user-centric cost, and (c) present a mechanism for selecting an optimal network based on the performance-cost analysis.

There have been a large number of researches on the evaluation of the network-centric performance or network-level quality of service (QoS), which are concerned with optimizing the network characteristics. The network-centric evaluation indicates, for instance, the larger the bandwidth (or the smaller the latency and loss rate) is the better the network performs. However, it cannot answer the question: how large bandwidth and smaller latency or loss rate are appropriate for a user's application. To answer this question, we need to evaluate user-centric performances. The user-centric performance, which is also a measure of the user-perceived QoS, relates user application requirements with the network service characteristics or quality. Note that a user requires network resources just sufficient to satisfy its application's requirements. Any extra resource beyond the requirement may not give any additional benefit to the user. In such a case, a user may not opt for a network that has highest resources; it may rather select a network that provides

just optimal performance at lowest user-centric cost.

This report is organized as follows. In Section 2 we present an optimal network selection mechanism by analyzing the user-centric performances and user-centric costs. In Section 3 we describe the implementation issues of the presented mechanism. Section 4 concludes this report.

2. Performances and Costs Analyses

2.1 User-centric Performances

We define the user-centric performance (UcPerf) as the degree of the fulfillment of user requirements by the network characteristics. There are the following two issues associated with UcPerf analysis.

(a) *Relationship between the UcPerf and network characteristic*: how the UcPerf can be expressed as the function of network characteristic. For instance, how the UcPerf increases as the bandwidth increases or delay decreases.

(b) *Combining the UcPerf of characteristics to get an overall UcPerf*: how to combine the performance of each characteristic to get an overall UcPerf.

As a first step towards these issues, we define a UcPerf as a continuous function of network characteristic because the continuous functions are easier to generate and manipulate than other types of functions. We take the weighted sum of the UcPerf of characteristics to get an overall UcPerf, because the weights enable us to control the contribution of the individual characteristic on overall performance.

Suppose that there are N wireless networks, each having some different characteristics from others. Let $\varphi_x^a(x_k)$ be the UcPerf component of characteristic x for an application a in a network k ($k \in N$). Then the overall UcPerf of the network, $UPerf_k^a$, is given as:

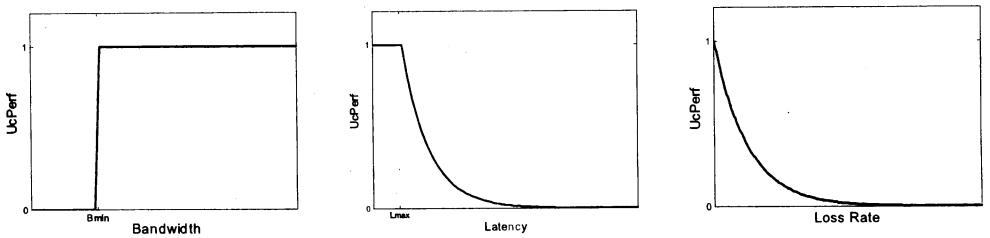
$$UPerf_k^a = \sum_F w_x \varphi_x^a(x_k), \qquad (1)$$

where w_x is a weighing factor for characteristic x and F is the set of network characteristics that are considered to evaluate the overall UcPerf.

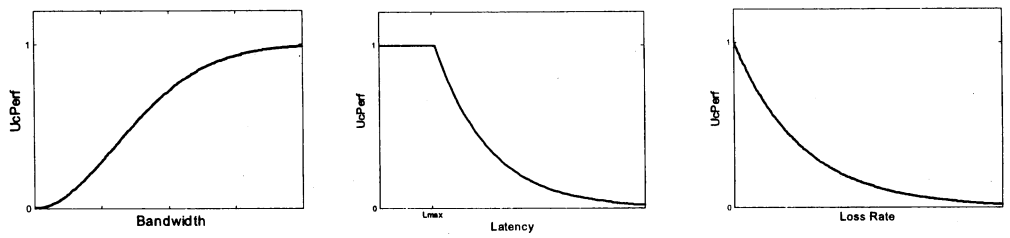
We define the $\varphi_x^a(x_k)$ as a continuous function whose value ranges between 0 and 1; 0 indicates the worst performance and 1 indicates the best performance. The value of a weighing factor w_x determines the contribution of the network characteristic x on the overall UcPerf. Considering $\sum_F w_x = 1$ guarantees $0 \le UPerf_k^a \le 1$. The shape (or nature) of $\varphi_x^a(x_k)$ function depends on both application type and network characteristic under consideration. We consider three types of applications: rigid, adaptive and elastic. These application types correspond to the real-time conversational service, real-time streaming service, and non real-time interactive and background services, respectively, defined in 3GPP specifications [3].

Similarly, although UcPerf depends on many networks characteristics such as bandwidth, latency, loss rate, reliability, availability, coverage, service provider's reputation, for simplicity, we consider only the first three characteristics, i.e. bandwidth, delay and loss rate because these are the most commonly used network characteristics to assess the network quality. Let $\varphi_b^a(b_k)$, $\varphi_l^a(l_k)$ and $\varphi_r^a(r_k)$ be the respective UcPerf components derived from bandwidth (b_k), latency (l_k) and loss rate (r_k) in a network k for an application a, where $a \in$ {*rigid, elastic, adaptive*}. To define $\varphi_b^a(b_k)$ we take bandwidth utility functions defined in papers [4], [5] as references. In addition, we provide the novel definitions of $\varphi_l^a(l_k)$ and $\varphi_r^a(r_k)$ based on the specifications of different types of applications.

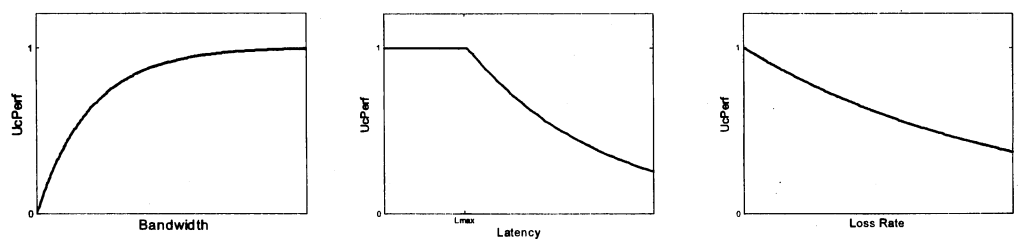
Rigid application: The UcPerf curves of a rigid application are shown in Fig. 1(a). There can be different sets of functions that can generate these curves. One set of such functions is shown by Equations (2)–(4). As a rigid application requires a fixed amount of bandwidth (say B_{min}) to support it, its performance is zero when



(a) Rigid application



(b) Adaptive application



(c) Elastic application

Fig. 1 User-centric performance (UcPerf) of bandwidth, latency and loss rate for (a) rigid, (b) adaptive and (c) elastic applications.

bandwidth is less than B_{min}. Similarly, it can tolerate very low network latency up to $L_{\max}^{rigid}$ without affecting the performance. However, when the delay increases beyond $L_{\max}^{rigid}$ the performance degrades exponentially at a rate of δ_l^{rigid}. Since these applications are loss intolerant, the performance decreases exponentially at a rate of δ_r^{rigid} as the loss increases.

$$\varphi_b^{rigid}(b_k) = 0 \quad \text{if } b_k < \mathrm{B_{min}} \qquad (2)$$
$$= 1 \quad \text{otherwise.}$$

$$\varphi_l^{rigid}(l_k) = 1 \quad \text{if } l_k \le L_{\max}^{rigid} \qquad (3)$$
$$= e^{-(l_k - L_{\max}^{rigid})\delta_l^{rigid}} \quad \text{otherwise.}$$

$$\varphi_r^{rigid}(r_k) = e^{-r_k \delta_r^{rigid}} \qquad (4)$$

Adaptive application: The UcPerf curves of an adaptive application are shown in Fig. 1(b), and the corresponding functions are given by Equations (5)–(7). Adaptive applications adapt their data rates to the available bandwidths in the network and can tolerate occasional delay bound violations and packet losses. However, they have intrinsic bandwidth requirements, as they must maintain the data rate at least some minimum level, below which performance suffers badly. This is accounted by a constant C_b in Eq. (5); larger the value of C_b, the larger the bandwidth required for better performance. Similar to the rigid application, the latency and loss rate performances of adaptive applications decrease

exponentially at the rate of $\delta_l^{adaptive}$ and $\delta_r^{adaptive}$, respectively.

$$\varphi_b^{adaptive}(b_k) = 1 - e^{-\left(\frac{b_k^2}{b_k + C_b}\right)} \quad (5)$$

$$\varphi_l^{adaptive}(l_k) = 1 \quad \text{if } l_k \le L_{\max}^{adaptive} \quad (6)$$

$$= e^{-(l_k - L_{\max}^{adaptive})\delta_l^{adaptive}} \quad \text{otherwise.}$$

$$\varphi_r^{adaptive}(r_k) = e^{-r_k \delta_r^{adaptive}} \quad (7)$$

Elastic application: The UcPerf curves of an elastic application are as shown in Fig. 1(c), and the corresponding functions are given by Equations (8)–(10). Elastic applications follow the diminishing marginal rate of performance improvement as bandwidth increases [4]. This means, when the bandwidth is low, an increment in the bandwidth increases the performance higher than the same increment does when the bandwidth is high. In Eq. (8) $\delta_b^{elastic}$ is performance increment rate, and B_{max} is the maximum bandwidth the application can utilize to improve its performance. Similarly, the performances of the latency and loss rate decrease exponentially at the rate of $\delta_l^{elastic}$ and $\delta_r^{elastic}$, respectively.

$$\varphi_b^{elastic}(b_k) = 1 - e^{-\left(\frac{\delta_b^{elastic} b_k}{B_{\max}}\right)} \quad (8)$$

$$\varphi_l^{elastic}(l_k) = 1 \quad \text{if } l_k \le L_{\max}^{elastic} \quad (9)$$

$$= e^{-(l_k - L_{\max}^{elastic})\delta_l^{elastic}} \quad \text{otherwise.}$$

$$\varphi_r^{elastic}(r_k) = e^{-r_k \delta_r^{elastic}} \quad (10)$$

Note that since the delay and loss tolerant capacities increase from the rigid to elastic applications, the following relations hold: $L_{\max}^{rigid} \le L_{\max}^{adaptive} \le L_{\max}^{elastic}$, $\delta_l^{rigid} \ge \delta_l^{adaptive} \ge \delta_l^{elastic}$ and $\delta_r^{rigid} \ge \delta_r^{adaptive} \ge \delta_r^{elastic}$.

2.2 User-centric Costs

After evaluating the user-centric performances, now we evaluate the user-centric costs (UcCost). We consider two types of costs: monetary cost and resource cost. The monetary cost includes the price that users have to pay, and the resource cost includes the resources of user terminal that have to be used for accessing the network services. From users' point of view, the battery power of a mobile terminal is the most precious resource. For instance, a user with low battery power prefers 3G networks to wireless LAN as the former consume lower power. Similar to the UcPerf, the UcCost is a normalized cost whose value ranges from 0 to 1; 0 indicates the service cost is trivial, and 1 indicates the service cost is the highest of what users can afford to pay.

Let $\vartheta_p^a(p_k)$ and $\vartheta_e^a(e_k)$ be the UcCost components of the network service price and battery power (energy) consumption, respectively, in a network k for an application a. Then the overall UcCost of the network, $UcCost_k^a$, is given as:

$$UcCost_k^a = w_p \vartheta_p^a(p_k) + w_e \vartheta_e^a(e_k) \quad (11)$$

where w_p and w_e are the weighing factors. ($w_p + w_e = 1$)

To derive the UcCost, we use the well-known principle of the demand function of economics. The theory of economics states that the quantity of goods/services demanded (q) increases as price (p) decreases [6]. As shown by the empirical results in [7], [8], the demand function of a communication service is:

$$q = A\, p^{-E}, \quad (12)$$

where A is the scaling constant which is equal to the value of q when p = 1, and E is the constant elasticity of demand for the given service. E is defined as the negative ratio of the relative change in demand to the relative change in price, that is:

$$E = -\frac{\Delta q / q}{\Delta p / p}, \quad (13)$$

where Δq is the change in demand and Δp is the change in price. Since the demand increases (decreases) as prices decreases (increases), Δp and Δq have opposite signs. Therefore, the value of E is always positive. As estimated by the France Telecom the elasticity, E, for voice service is 1.337 [8].

An important property of the constant elasticity in Equation (12) is that it creates a demand curve that has

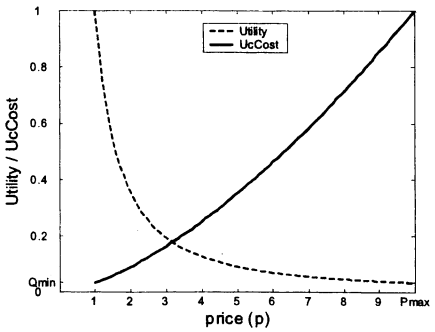


Fig. 2 Utility and user-centric cost versus network price.

different slopes in different price regions. When the price is low, small change in price creates larger change in the quantity demanded. On the other hand, when the price is high, even a large change in price creates only a small change in demand. Demand function can also be interpreted in terms of utility function. A utility function measures the willingness of users to pay an amount of money for a service with certain performance or QoS guarantees [9]. When the price is low, users think that the utility of the service is higher than the price paid, so they demand more quantity of the service. On the other hand, when the price is large the utility of the service becomes smaller than the price paid, so that users demand less quantity of the service. Based on this ground, we can express the utility function (u) of the service in terms of the value of quantity demanded (q) as: u = q when A = 1, so that $0 \le u \le 1$. Large utility indicates that the user is more satisfied hence the user-centric cost is low. Similarly, small utility indicates that the user is less satisfied hence the user-centric cost is high. Therefore, the user-centric cost of the price can simply be expressed as inversely proportional to the service utility. That is,

$$\vartheta_p^a(p) = \frac{C}{u}, \qquad (14)$$

where C is the proportionality constant. To keep the value of $\vartheta_p^a(p)$ within the range of 0 to 1, we take C = q_{min}, where q_{min} is the minimum amount of service demanded when the price is maximum (p_{max}). When we plot the utility and $\vartheta_p^a(p)$ in y-axis, and price in x-axis we get curves as shown in Fig. 2.

Now we define $\vartheta_e^a(e)$ as a function of the battery power consumption of the mobile terminal. As there are no references available on how users behave for different levels of power consumption, for simplicity we assume that $\vartheta_e^a(e)$ varies with battery power consumption in the same way as $\vartheta_p^a(p)$ does with price. This means, we suppose that the elasticity of power consumption is the same as that of network service price.

2.3 Optimal Network Selection

We use UcPerf and UcCost to estimate a performance-cost ratio (PCR) as given by Eq. (15). The network selection mechanism estimates UcPerf, UcCost and PCR, and selects an access network that has the highest value of PCR for the given application a.

$$PCR_k^a = UcPerf_k^a / UcCost_k^a, \quad \text{for } k = 1,2,\ldots,N \qquad (15)$$

where N is the number of access networks available from the location of mobile user.

3. Implementation Issues

The optimal network selection mechanism can be implemented as a module comprising of different profiles, as shown in Fig. 3. These profiles are described below.

Network profile: The network profile includes values of network characteristics, e.g. bandwidth, latency, loss rate, prices, etc. The mobile terminal monitors all the available networks in its surroundings and interrogate with network entities such as the base stations or access points to get the values of these characteristics.

Application profile: The application profile maintains the values of parameters, such as B_{min}, L_{max} and δ_l, and δ_r, related to the application requirements. These parameters of an application can be defined while the application is designed or developed by observing the effect of these parameter changes on the application quality. These parameters are provided to the application profile when the application is installed in the mobile

terminal.

User preference profile: The user preference profile maintains the values of weighing factors used in the evaluation of the user-centric performance and cost. These factors give the notion of the relative importance of the components of UcPerf and UcCost. The update of a user preference profile can be done by users themselves through user interfaces or can be done automatically (by mobile terminals) by interacting with the application profile, network profile, and user and device contexts.

Architecture of network selection module is shown in Fig. 3. In this figure, arrows with numbers in parentheses show the interaction between different units of the module. We describe these interactions one by one. (1) The network detection unit probes the available networks through network interfaces and stores the collected network attributes in the network profile. (2) User applications are interacted to develop the application profile. (3) Network profile, (4) application profile and (5) user interfaces are consulted to maintain the user preference profile. The UcPerf/UcCost estimator gets information from the (6) network profile, (7) application profile, and (8) user preference profile to compute the UcPerf and UcCost, which are used by the network selection algorithm to select the best access network. Arrows (9) and (10) are to provide the information of selected wireless network to the concerned network interface card. As this module interacts with the user interfaces and applications as well as network interfaces, it can efficiently be located as a middleware between the application and transport layers.

4. Conclusion

In this report, we presented a theoretical framework of a mechanism for carrying out an optimal network selection decision based on user-centric performance and user-centric cost analyses. In future work, we will investigate the issues of graceful transfer of connections from one network to another when the previous one becomes sub-optimal or unavailable.

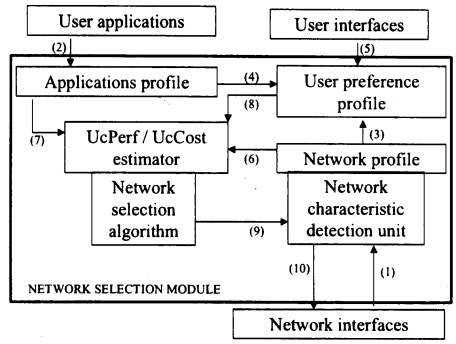


Fig. 3 Architecture of network selection module.

References

[1] E. A. Brewer et al., "A network architecture for heterogeneous mobile computing," IEEE Pers. Commun., pp.8-24, Oct. 1998.

[2] G. Wu, M. Mizuno and P. J. Havinga, "MIRAI architecture for heterogeneous networks," IEEE Commun. Mag., pp.126-134, Feb. 2002.

[3] 3GPP TS 23.107, "Quality of service (QoS) concept and architecture," Version 6.1.0, April 2004.

[4] S. Shenker, "Fundamental of design issues for the future Internet," IEEE J. Sel. Areas Telecommun., vol.13, no.7, pp.1176-1188, Sept. 1995.

[5] V. P. Kafle, E. Kamioka, S. Yamada, "Mobility and its impact on user satisfaction function in heterogeneous wireless networks," Proc. Int'l Conf. on Commun. and Computer Networks (CCN2004), Nov. 2004.

[6] N. G. Markiw, Essential of Economics, Second Edition, South-western, Thomson Learning, 2001, pp.67-98.

[7] S. G. Lanning, D. Mitra, Q. Wang and M. H. Wright, "Optimal planning for optical transport networks," Philosoph. Trans. Royal Soc. London A, vol.358, no.1773, pp.2183-2196, Aug. 2000.

[8] M. Aldebert, M. Ivaldi, C. Roucolle, "Telecommunication demand and pricing structure: an econometric analysis," Telecommun. System, Kluwer, 25:1,2, 89-115, 2004.

[9] L. A. DaSilva, "Pricing for QoS-enabled networks: a survey," IEEE Commun. Surveys, Second Quarter 2000.