Scheduling algorithms for 4G wireless networks

Jaume Ramis, Loren Carrasco, Guillem Femenias and Felip Riera-Palou

Mobile Communications Group - Dept. of Mathematics and Informatics
University of the Balearic Islands
07122 Palma, Mallorca (Illes Balears), Spain.
{jaume.ramis,loren.carrasco,guillem.femenias,felip.riera}@uib.es

Abstract. Scheduling algorithms are fundamental components in the process of resource management in mobile communication networks with heterogeneous QoS requirements such as delay, delay jitter, packet loss rate or throughput. The random characteristics of the propagation environment and the use of complex physical layers in order to combat this random behavior further complicates the design of simple, efficient, scalable and fair scheduling algorithms. This paper presents the main criteria used in the design of scheduling algorithms for 3G/4G mobile communications networks and provides a survey of scheduling mechanisms proposed for use in TDMA and CDMA based systems.

Keywords: scheduling, 4G, GPS, utility functions

1 Introduction

One of the most challenging issues for next-generation wireless networks is the provision of Quality-of-Service (QoS) guarantees when dealing with the high number of emergent multimedia applications. This necessitates the development of high-performance physical-layer technologies, as well as powerful resource management strategies to provide high throughput and efficient use of resources. Among these strategies, scheduling algorithms, which distribute the available 'capacity' among existent connections, have been recognized as key components of these QoS aware wireless systems. In order to support the provision of QoS in wireless networks a large number of traffic scheduling algorithms have been proposed in the literature. The knowhow in wireline schedulers has been the basis for the development of these scheduling strategies; however, the service heterogeneity, the scarcity of resources, and the hostility and variability of mobile radio channels, have rendered unavoidable the adaptation of wireline proposals to the more challenging wireless scenario. Scheduling algorithms for wireless networks can be classified into two categories: centralized and distributed algorithms. Distributed proposals are mainly applied in adhoc or uplink operation, where users contend for channel access; these strategies do not achieve the efficiency, fairness and fulfilment of QoS requirements that can be reached with centralized algorithms. In this paper we propose an exhaustive overview of centralized wireless scheduling techniques, which are evaluated in accordance with a set of relevant performance criteria for next generation wireless networks scenarios.

2 Scheduling criteria for 4G wireless networks

The main criteria used to evaluate wireless schedulers in their application to 3G/4G wireless networks are:

- **Efficiency:** In highly loaded 4G scenarios, efficiency (measured in terms of total achieved *throughput*) is one of the most significant performance criteria. In TDMA the maximum efficiency is achieved when at each time instant the user with the highest available throughput is selected for transmission (multiuser diversity). If the physical layer includes a CDMA component the *soft-capacity* phenomena should be considered in the scheduling process in order to increase the efficiency. That is, the scheduler should not only take into account the capacity variations due to changes in the existent interference level, but should also consider the capacity variations provoked when selecting one or another particular combination of services to be transmitted simultaneously.
- **Applicability:** Using this concept we group the issues of algorithm complexity, amount of signalling involved in the scheduling process, parameter settings (i.e. GPS weights determination) and the considered channel model. With respect to this last point, a scheduler for 4G wireless systems should be capable of managing an adaptive channel model with multiple possible states corresponding to the different transmission modes available in 3G-4G physical interfaces. This implies that the scheduler design strongly depends on the physical layer characteristics and that the use of a cross-layer design of both layers is highly desirable.
- **QoS support:** The existing service heterogeneity in new multimedia networks is directly translated into multiple and distinct QoS traffic requirements that should be jointly managed and guaranteed by the scheduler. Three levels of QoS support are considered: best-effort traffic, data traffic with throughput and delay guarantees, and real-time traffic.
- **Fairness:** A fair distribution of resources between connections of the same type is recommended. In wireless scenarios different kinds of fairness could be implemented: fairness in terms of the data rate assigned to each connection or in terms of resource consumption.

3 Classification of scheduling algorithms

Wireless networks schedulers proposed in the literature could be grouped in different families. Figure 1 shows a non exhaustive classification that nevertheless includes the main existing scheduling families.

3.1 GPS-based scheduling algorithms

GPS (*Generalized Processor Sharing*) [4], also known as FFQ (*Fluid-flow Fair Queueing*) [5], is a fair, work-conserving, flexible and efficient algorithm originally devised for error-free wireline networks. The fundamental concept in GPS-based algorithms is that the amount of service session i receives from the switch

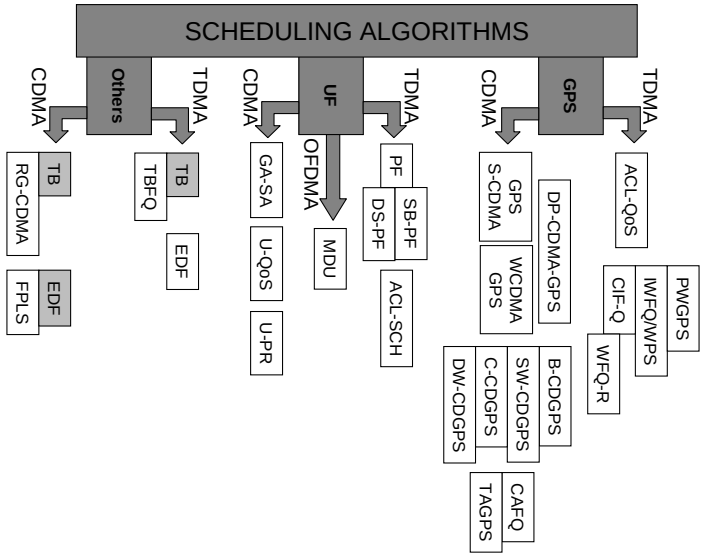

Fig. 1. Wireless schedulers classification.

(in terms of transmitted packets) is proportional to a positive weight also called relative service share ϕ_i, provided that the session is continuously backlogged in this interval. Defining $S_i(t,\tau]$ as the amount of service received by session i in the time interval $(t,\tau]$, then it holds that $\frac{S_i(t,\tau]}{S_j(t,\tau]} = \frac{\phi_i}{\phi_j}$ for all sessions j that have received service in this time interval. It follows that in the worst case, the minimum guaranteed rate r_i given to session i is $r_i = r\phi_i / \sum_{j=1}^{M} \phi_j$, where M is the maximum number of sessions that could be active in the system and r is the available total throughput. Therefore, from a user perspective, GPS guarantees that network resources are allocated to the sessions irrespectively of the behavior of other sessions (isolation property), and that the distribution is performed with perfect fairness, that is, whenever a session i generates traffic at a rate less that r_i, the 'spare' bandwidth is allocated to other sessions proportionally to their respective weights. In addition, if session i is $(\sigma_i,\ \rho_i)$-leaky bucket constrained, where ρ_i is the token generation rate and σ_i the size of the token bucket, and the minimum guaranteed rate is such that $r_i \geq \rho_i$, then the maximum delay is bounded by $D_i^{max} \leq \sigma_i/\rho_i$. The main consequence of this property is the possibility of using GPS for delay constrained traffic. Therefore, session rate and delay guarantees depend on an adequate choice of the service share ϕ, but the determination of sessions weights could be a challenging task. When the sessions have a long term average rate, the ϕ_i allocation appears to be straightforward. The ϕ_i determination is more complex when the traffic is bursty or self-similar. To fix its value, minimum session rate and maximum accepted delay have to be taken into account.

GPS considers traffic as a fluid, therefore ideal GPS cannot be implemented in practical schedulers, since it requires that the scheduler serves multiple flows simultaneously and that the traffic is infinitely divisible. There are many packet-level algorithmic implementations of this model. Essentially, the goal of these algorithms is to serve packets in an order that approximates GPS as closely as possible. The first GPS packet adaptation was the PGPS (*Packet-by-packet GPS*) [4] also called WFQ (*Weighted Fair Queueing*) [5]. This algorithm is based on the 'system virtual time' definition. The virtual time concept is used to track the progress of the system under the GPS discipline and will lead to a practical implementation of PGPS. When a new packet arrives to the PGPS server it is stamped with its 'virtual starting time' t_{vs} and its 'virtual finishing time' t_{vf} defined as the time instant at witch this packet would start/ finish service under the fluid GPS respectively. Packets are served according to virtual finishing time order. Both times could be calculated on the packet arrival as long as the set of active sessions at that moment is known. Obviously, this adaptation of the fluid GPS system to the packetized transmission implies a significant increase of complexity due to the fact that the scheduler has to tag each packet and has to maintain the system virtual time permanently updated. There are several proposals derived from PGPS that either approximate more accurately the GPS behavior [6], [7], [8] or decrease the PGPS complexity [9], [10], [11].

When adapting wireline fair queuing algorithms to wireless channels it may happen that some flows can transmit but other flows cannot due to location dependent channel errors; therefore only a subset of flows can be scheduled on the channel at a given time instant, and this subset is dynamically changing as time evolves. All the wireline algorithms cited above assume that the channel is error-free, or at least that either all flows can be scheduled or none of them can. In order to succeed in maintaining the long term fairness, most wireless GPS-based algorithms introduce compensation models (flows that at a certain time interval receive less channel resources than in an error-free environment will be compensated by receiving extra resources when their channel conditions improve).

GPS-based schedulers for TDMA networks incorporate compensation models to cope with the channel variability, e.g. IWFQ (*Idealized Wireless Fair Queueing*) and WPS (*Wireless Packet Scheduler*) [2]. The compensation model is based on several concepts: the *error free service* of a flow is the service that it would have received at the same time instant if all channels had been error free under identical offered load. A flow is said to be *leading* if it has received an allocation (*lead*) in excess of its error free service, in contrast, a flow is said to be *lagging* if it has received an allocation (*lag*) less than its error free service. The compensation for the service loss suffered by *lagging* sessions will be performed when these connections have a radio channel that can ensure the meeting of their QoS requirements. Obviously the extra service assigned to *lagging* sessions during compensation will be subtracted from *leading* sessions. In IWFQ and WPS strategies, the packet from the subset of flows with a good channel and

the lowest *virtual finishing time* is selected to be transmitted. Since the virtual time of a session increases only when it receives service , this may result in a large difference between the virtual time of an error session i and a error-free session. If session i exits from error it will then have the smallest virtual time among all sessions. The server will select session i exclusively for service until its virtual time catches up with those of other sessions. By artificially bounding the *lag* and *lead*, it can be established a trade off between long-term fairness and the starvation of error-free sessions. To solve these problems the CIF-Q (*Channel Condition Independent Fair Queueing*) [12] algorithm defines an additional parameter called *lag* that keeps track of the difference between the service a session should receive in the error-free reference system and the service it has actually received. The packet selected to be transmitted is the one with the lowest starting virtual time in the error-free system, but if the selected packet corresponds to a leading flow that has already received its guaranteed share of service (according to its weight) the slot will be assigned to the *lagging* flows proportionally to their weights. This mechanism guarantees the delay bound and throughput of error-free sessions and provides long-term and short-term fairness.

Another algorithm that tries to solve the problem of starvation is the PWGPS (*Packet Wireless GPS*) [13]. Packets are served in increasing order of a simplified virtual finishing time. In this case the compensation is achieved increasing the GPS weights ϕ of the error sessions until they have obtained their corresponding service share. This scheduler does not require the maintenance of an error-free system thanks to a redefinition of the virtual finishing time. However, the changes in the weights imply that those virtual times can not be calculated at the packet arrival and should be recalculated at each scheduling interval.

All the above algorithms assume a perfect channel prediction before transmission, but this is a very difficult task and in real networks link layer mechanism as ARQ retransmissions are required. The algorithm WFQ-R (*WFQ with link level Retransmission*) [14] distributes the scarce wireless resources among all flows according to their weights, but considering also the resource consumption of the retransmissions. It combines the CIF-Q algorithm with a compensation system where the share used for retransmissions is regarded as a debt of the retransmitted flow to the others. The compensation could be charged to the retransmitted flow only (an error-prone flow should take responsibility for its own channel condition) or the compensation is distributed over all flows proportionally to their weights. The algorithm ACL-QoS (*Adaptive Cross-Layer scheduler with prescribed QoS guarantees*) [15] is a recent proposal of a TDMA-scheduler with a cross layer design. Instead of an on-off physical layer model it assumes a system with AMC (*Adaptive Modulation and Coding*). The channel is modeled as a finite Markov chain with multiple states, i.e. depending on the channel state a certain modulation and coding scheme will be selected. The scheduler distinguishes between two traffic types: QoS-guaranteed traffic and best-effort traffic. The resource distribution for QoS-guaranteed traffic does not follow a GPS discipline but takes into account the amount of data in the transmission buffer and the channel state, that is, the scheduler uses information from physical and data

link layers. The resources left after this process are distributed among best-effort connections using GPS.

As we already mentioned, the fluid model of the GPS discipline assumes that multiple sessions with different assigned rates can be served simultaneously. The adaptation of this fluid model to TDMA requires the maintenance of a system virtual time and therefore a significant increase in system complexity. Otherwise CDMA offers the possibility of simultaneous transmissions, and the resource distribution could be adjusted by using the spreading factor, the amount of assigned power, etc.

GPS-based schedulers for CDMA networks are derived from the original GPS algorithm, and they consider slotted-CDMA physical layers, i.e. the distribution of resources is performed frame by frame.

B-CDGPS (*Basic CDGPS*) [16] is a simple scheduler that considers an ideal channel with a fixed capacity estimated as a lower bound of the system soft capacity. This implies a severe loss of efficiency and, therefore, the rest of considered algorithms take the variability of the CDMA capacity into account. For instance, GPS S-CDMA (*GPS in Slotted CDMA*) [17], SW-CDGPS (*Static Weighted CDGPS*) [16], C-CDGPS (*Credit based CDGPS*) [16] and the WCDMA GPS [18], consider the capacity variations in an ideal channel. The S-CDMA [17] scheme, derived from PGPS, considers a unicellular system in which the system resources (power and bandwidth) are distributed proportionally to the connections weights ϕ_i. BER and delay requirements of each connection are used to adjust the transmission powers and GPS weights. The algorithm considers that any transmission rate is possible. The SW-CDGPS (*Static Weighted CDGPS*) [16] algorithm introduces the concept of *nominal capacity*, which corresponds to the amount of available resources in a cell, and this can be determined if the intercell interference level is known. The set of SIR values and transmission rates corresponding to all active users in the cell can be derived from the users capacity requests and the nominal cell capacity. The calculation process maximizes the aggregate throughput and satisfies the GPS fairness property. To increase SW-CDGPS efficiency, the C-CDGPS (*Credit based CDGPS*) [16] algorithm sacrifices the short-term fairness of non real time traffic. This is implemented by using a credit counter to track the difference between the received service share of a session and the fair GPS service that would correspond to that session. This counter is bounded in order to achieve a long-term fairness. The WCDMA GPS algorithm [18] also makes use of the *nominal capacity* concept, but instead of following the basic GPS strategy, resources are distributed using the PGPS discipline and considering a discrete set of possible transmission rates.

As we have already mentioned, the described schedulers consider an ideal channel. The introduction of a non-ideal channel implies, as well as in TDMA proposals, the addition of compensation mechanisms to guarantee long-term fairness to those sessions with a 'bad' channel. Some examples of this kind of schemes are DW-CDGPS (*Dynamic Weighted CDGPS*) [16], CAFQ (*Channel Adaptive Fair Queueing*) [19] and TAGPS (*Traffic Aided GPS*) [3]. The proposal

of [16] firstly schedules real time transmissions using SW-CDGPS, and secondly distributes the remaining capacity among the non real time connections according to the DW-CDGPS algorithm: the GPS weights of those connections will be adjusted dynamically every scheduling period depending on their channel conditions in order to increase the data throughput. In this compensation mechanism there is a trade-off between efficiency and short-term fairness. Similarly to many GPS-based TDMA algorithms, the CAFQ [19] algorithm has to keep an error-free system as a reference, in which the SFQ discipline [11] is used to schedule transmissions. In the real system, sessions are arranged in increasing order of *lagging* credits (the bigger the difference of service received in the real system in comparison to the service received in the reference system, the more *lagging* credits), and error-free sessions are chosen to be served. There is an exception: if the reference system selects a lagging session with an error-free channel to be transmitted, the real system will maintain this selection. TAGPS [3] uses a similar mechanism. Similarly to other proposals [17] system resources are represented by the power index g_i. It distinguishes two user types: voice with a constant transmission rate and data with variable bitrate. The distribution of resources consists in determining the g_i values of data users. If a 'lagging' user perceives a 'good' channel, its GPS-weight, ϕ_i (and consequently its power index, g_i) will be increased in order to compensate its lack of service.

The DP-CDMA-GPS (*Dynamic Programming CDMA-GPS*) [20] algorithm tries to avoid delay and bandwidth coupling. To that end it uses variable weights, which are determined by using a cost function that minimizes the queuing delay experienced by active connections. This cost function also takes into account the QoS requirements and the radio channel conditions of the different connections.

3.2 Utility-based scheduling algorithms

This theory has its origins in economics, where the utility functions are used to quantify the benefit of using certain resources. Similarly, utility theory can be used in communications networks to evaluate the degree to which a network satisfies service requirements of user's applications, rather than in terms of system centric quantities like throughput, outage probability, packet drop rate and power [21]. The utility function maps the network resources a user utilizes into a real number that tries to reflect the level of user satisfaction derived from using these resources. Different applications have different utility function curves or even different parameters. For instance, the utility functions of best-effort applications consider throughput, whereas those of delay sensitive traffic applications take into account delay. Finding the most adequate utility functions for the different types of applications is one of the key elements to ensure a correct performance of this kind of techniques. Examples of different utility functions can be found in [22–24].

Once the utility functions of all applications in the system have been defined, the objective is to achieve an optimum scheduler that maximizes the aggregate utility in the system subject to the capacity limit determined by the physical layer techniques. Usually, utility functions are non linear and the problem can

be regarded as a non linear combinatorial optimization of a cost function (the aggregated system utility) with a large dimension. This optimization is a complex process and the algorithms designed for this purpose constitute the other key element of the utility-based scheduling architecture. The optimization algorithms should be efficient (in terms of computational effort) and they have to work with a discrete set of possible solutions (the physical layer offers a discrete set of possible rates and/or delays). For instance, in [25] a set of algorithms for OFDM wireless networks are proposed, and some of them could easily be adapted to CDMA networks.

Proportional fair schedulers. The PF (*Basic Proportional Fair*) [26] is a downlink algorithm well suited for best-effort traffic. The scheduler tries to take advantage of the independent channel variations experienced by different users by scheduling transmissions that perceive strong signal levels. The scheduler sends data to the mobile with the highest DRC_i/R_i ratio, where DRC_i represents the highest feasible rate out of all possible rates under those channel conditions in a given slot, and R_i represents the average rate received by the mobile over a window of appropriate size. This way, each user is served in periods where its requested rate is closer to the peak compared to its recent requests. This channel-aware scheduling can substantially improve network performance through *multiuser diversity*. It can be shown that this system amounts to use a logarithmic utility function for all users $U_i(R_i) = \ln(R_i)$. In this case, the aggregate system utility optimization can be performed using a utility-based gradient scheduling (assigning the resources to the connection with a higher DRC_i/R_i in each scheduling period). PF is an example of an utility-based solution whose aggregate cost function optimization is very simple thanks to the chosen utility function. This algorithm is used in current systems (i.e. HDR channel in the CDMA2000 system).

The main drawback of the PF scheduling stems from the use of a rate-dependent-only utility function that makes the system unable to ensure traffic delay requirements. Several proposals try to solve this limitation [27, 28] maintaining the PF utility function for best-effort traffic and defining new utility functions for delay sensitive traffic flows. In the DS-PF scheduler (*Delay Sensitivity PF*) [27] real-time sessions increase their priority when a certain delay value is trespassed, while in the SB-PF (*Sender Buffer PF*) [28] proposal the objective is to avoid playout buffer starvation. Unfortunately, in these mixed cases with different utility functions, the aggregate system utility optimization is much more complex. To preserve PF simplicity, both Barriac in [27] and Koto in [28] propose sub-optimum optimization algorithms.

Multimedia utility-based schedulers define different utility functions for different traffic types and use optimization algorithms to maximize the aggregate system utility (GA-SA [22], MDU [23], U-PR [24], U-QoS [29], ACL-SCH [30]). Most of these algorithms are designed for CDMA networks [22, 24, 29], but there are also TDMA [30] and OFDMA [23] proposals.

3.3 Other scheduling algorithms

Token Bucket-based schedulers include a data bucket and a token bucket per connection. Each unit of traffic (usually a packet) to be transmitted consumes a token of the token bucket. Tokens are generated at a constant rate derived from the contracted data rate of the connection. The maximum number of tokens in the token bucket limits the allowed peak rate for that connection. The TBFQ (*Token Bucket Fair Queueing*) [31] algorithm was designed for TDMA systems. In real time multimedia applications it is very difficult to predict traffic profiles and out-of-profile degradations may be detrimental to the overall QoS experienced by the end user. The TBFQ tries to adapt to this unpredictable workload by accepting traffic profile violations when excess bandwidth is available, provided the session does not exceed its bandwidth allocation in the long term. To this end, each connection has a sufficiently large input data buffer, and a token bucket holding tokens for one packet transmission only. In addition there is a common token bank and connections can borrow tokens from the bank when its token pool is depleted and there are still packets to be served, or give tokens to the bank during periods when the incoming traffic rate is less than its token generation rate. Each connection has an associated counter keeping track of the number of tokens borrowed from or given to the token bank. The connection with the highest counter value has the highest priority in borrowing tokens. There is also a debt limit to preserve fairness, below which the connection can no longer borrow from the bank.

The RG CDMA (*Rate Guarantee in CDMA*) [32] proposal is adapted to the downlink of a W-CDMA system and it allows connections with variable data rate by using an OVSF (Orthogonal Variable Spreading Factor) code tree. It assigns a guaranteed data rate to each traffic flow and associates a credit counter to it. Every scheduling period the credit counter increases proportionally to the guaranteed data rate and decreases with each transmitted packet. In this way, the credit counter value represents the difference between the guaranteed and the actually executed transmissions. Connections are ordered according to the number of credits and higher rate OVSF codes are assigned to the connections with a higher credit counter value. It should be pointed out that this algorithm does not consider the CDMA soft capacity phenomena and assumes an ideal channel.

EDF-based algorithms. The EDF (*Earliest Deadline First*) [33] is a proposal for TDMA networks. In EDF each packet is tagged with its deadline. This deadline is calculated according to the delay requirements of that traffic flow. Packets are served in the order derived from their deadlines. In mixed TD-CDMA networks the FPLS (*Fair Packet Loss Sharing*) [34] algorithm schedules multimedia packets transmission ensuring that packet losses are distributed among all connections according to their QoS requirements. For this purpose packets are ordered according to their timeout. If the available system capacity is not enough to transmit all backlogged packets, the packets to be discarded (packets reaching their timeout) are selected among packets of all connections taking into account

each connection's BER (Bit Error Rate). The main drawback of this algorithm is the assumption of an ideal channel without intercell interference.

4 Conclusions

An obvious conclusion of this study is the impossibility of designing a wireless scheduler that is simultaneously: fair, simple, efficient and able to ensure real-time delay guarantees. Focusing on the GPS-based algorithms, several aspects have to be highlighted:

- The perfect fairness provided by the GPS discipline is not possible due to channel errors affecting a varying subset of connections. Therefore it is necessary to incorporate compensation mechanisms to maintain the long-term fairness. Currently these mechanisms use too simplistic channel models (usually on-off models) to represent de behavior of adaptive 3G-4G physical layers and they affect the isolation property of GPS resulting in an increased complexity .
- If the traffic is leaky-bucket constrained, the maximum packet delay can be bounded. However, since in GPS-based schedulers rate and delay guarantees are controlled by a single parameter, the session weight ϕ_i, there is a coupling effect between rate and delay requirements. This implies that to ensure low delays in a certain traffic flow it is necessary to assign a high portion of the available bandwidth to that connection (i.e assign a high weight). This behavior suffers from a non efficient use of system resources mainly in multimedia environments. As a consequence, GPS-based schedulers are not suitable to manage heterogeneous multimedia traffic environments.
- One of the most significant drawbacks of GPS-based algorithms is the connection's service share ϕ_i determination, as a correct operation of the algorithm depends on a suitable selection of the sessions weight values. Surprisingly many existent proposals do not treat this problem.
- In addition, the packetized adaptations of the GPS fluid discipline require the introduction of the virtual time concept. The calculation of this virtual time significantly increases the GPS original complexity. The GPS-based proposals for TDMA networks will always require the maintenance of a system virtual time because the simultaneous transmissions are not possible. Some of these proposals include a simplified definition of virtual time although these simplifications compromise the fairness requirements.
- In CDMA systems a discrete set of transmission rates is possible. This fact is not taken into account by many proposals that are solely based on GPS fluid model, where transmission rates can take any value.

Therefore if we apply the criteria defined in section (2) we can conclude that in GPS-based techniques efficiency is sacrificed in favor of fairness and that the applicability of GPS algorithms to 3G/4G wireless networks is quite limited and in any case restricted to best-effort traffic and rate-delay guaranteed data.

The basic idea behind utility-based schedulers is the mapping of the resources use (bandwidth, power, etc.) or performance criteria (data rate, delay, etc.) into the corresponding utility function and its optimization, instead of directly measuring the network performance parameters. For instance, if a small increase in the transmission rate allocated to a multimedia application makes it able to start its transmission, the benefit of this small rate increment is highly valuable, whereas if the same increase of the data rate is allocated to a best-effort application that has already started transmission at a high bit rate it will not imply a significative utility increase. The key is the definition of the most adequate utility function for each of the different application classes. The main drawback of utility-based schedulers is that the required aggregate system utility optimization usually requires the resolution of a nonlinear combinatorial optimization problem of a large dimension. The computational cost of this optimization process can be minimized by using efficient optimization algorithms. Some proposals, as for instante the PF algorithm, use simple utility functions and, therefore, optimization can be performed using a very simple technique. It is worth mentioning that physical layer information can be included by means of restrictions to the aggregate system utility [23]. The inclusion of this information in the definition of the cost function makes it necessary to solve the optimization problem every time the channel state changes [30]. Utility-based techniques are a promising scheduling discipline that is able to cope with heterogeneous QoS requirements, can achieve high efficiency and a certain degree of fairness between connections of the same type. The applicability of these techniques in 3G/4G systems is conditioned by the existence of simple and efficient optimization algorithms.

Acknowledgments

This work has been supported in part by the MEC and FEDER under project MARIMBA (TEC2005-0997), Govern de les Illes Balears under project XISPES and grant PCTIB-2005GC1-09, and a Ramon y Cajal fellowship, Spain.

References

1. H. Zhang, "Service disciplines for guaranteed performance service in packet switching networks," *Proc. IEEE*, vol. 83, no. 10, pp. 1374-1396, 1995.
2. S. Lu et al., "Fair Scheduling in Wireless Packet Networks," *IEEE/ACM Trans. Networking*, vol. 7, no. 4, pp. 473-489, 1999.
3. D. Liao and L. Li, "Traffic aided fair scheduling using compensation scheme in CDMA cellular networks," *ICC'05*, vol. 1, pp. 363-367.
4. A. Parekh, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 344-357, 1993.
5. A. Demers et al., "Analysis and Simulation of a Fair Queueing Algorithm," *SIGCOMM'89*, pp. 1-12, 1989.
6. J. Bennet and H. Zhang, "WF2Q: Worst-case Fair Weighted Fair Queueing," *IEEE INFOCOM 1996*, vol. 1, pp. 120-128, March 1996.

7. J. Lee et al., "WF2Q-M : a worst-case fair weighted fair queueing with maximum rate control," *IEEE GLOBECOM 2002*, vol. 2, pp. 1576-1580, Nov. 2002.
8. J. Gallardo and D. Makrakis, "Dynamic predictive weighted fair queueing for differentiated services," *IEEE ICC 2001*, vol. 8, pp. 2380-2384, June 2001.
9. J. Bennett and H. Zhang, "Hierarchical packet fair queueing algorithms," *IEEE/ACM Trans. Networking*, vol. 5, no. 5, pp. 675-689, 1997.
10. S. Golestani, "A self-clocked fair queueing scheme for broadband applications," *IEEE INFOCOM 1994*, vol. 2, pp. 636-646, June 1994.
11. P. Goyal et al., "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Trans.Networking*,vol.5,no.5, pp.690-704, 1997.
12. T. Ng et al., "Packet fair queueing algorithms for wireless networks with location-dependent errors," *IEEE INFOCOM 1998*, vol. 3, pp. 1103-1111, March-April 1998.
13. M. Jeong et al., "Wireless packet scheduler for fair service allocation," *IEEE APCC/OECC 1999*, vol. 1, pp. 794-797, Oct. 1999.
14. N. Kim and H. Yoon, "Packet fair queueing algorithms for wireless networks with link level retransmission," *IEEE CCNC 2004*, pp. 122-127, Jan. 2004.
15. Q. Liu et al., "Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks," *IEEE JSAC*, vol. 23, no. 5, pp. 1056-1066, 2005.
16. L. Xu et al., "Dynamic Fair Scheduling With QoS Constraints in Multimedia Wideband CDMA Cellular Networks," *IEEE Trans. Wireless Com.*,vol.3,no.1,pp.60-73,2004.
17. M. Arad and A. Leon-Garcia, "A generalized processor sharing approach time to scheduling in hybrid CDMA/TDMA," *IEEE INFOCOM 1998*, vol. 3, pp.1164-1171.
18. X. Wang, "An FDD Wideband CDMA MAC Protocol with Minimum-Power Allocation and GPS-Scheduling for Wireless Wide Area Multimedia Networks," *IEEE Trans. Mobile Comp.*, vol. 4, no. 1, pp. 16-28, 2005.
19. L. Wang et al., "Channel Adaptive Fair Queueing for Scheduling Integrated Voice and Data Services in Multicode CDMA Systems," *IEEE WCNC 2003*, vol. 3, pp. 1651-1656, March 2003.
20. A. Stamoulis et al.,"Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming," *IEEE Trans.Wireless Com.*,vol.3,no.2, pp.512-523,2004.
21. G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks. Part I: Theoretical framework," *IEEE Trans. Wireless Com.*, vol.4, no.2, pp. 614-624, 2005.
22. W. Zhao and M. Lu, "CDMA downlink rate allocation for heterogeneus traffic based on utility function: GA-SA approach," *CNSR'04*, pp. 156-162, May 2004.
23. G. Song, "Utility-based resource allocation and scheduling inOFDM-based wireless broadband networks," *IEEE Commun. Magazine*, pp. 127-134, 2005.
24. X. Duan et al., "A dynamic power and rate joint allocation algorithm for mobile multimedia DS-CDMA networks based on utility functions," *PIMRC'02*, vol. 3, pp. 1107-1111, Sept. 2002.
25. G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks. Part II: algorithm development," *IEEE Trans. Wireless Com.*, vol.4, no.2, pp.625-634,2005.
26. A. Jalali et al., "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *IEEE VTC'00*, vol. 3, pp. 1854-1858, May.
27. G. Barriac and J. Holtzman, "Introducing Delay Sensitivity into the Proportional Fair Algorithm for CDMA Downlink Scheduling," *IEEE ISSSTA'02*,vol.3,pp.652-656.
28. H. Koto et al., "Scheduling Algorithm based on Sender Backlog for Real-Time Application in Mobile Packet Networks," *IEEE WCNC 2005*, vol. 1, pp. 151-157.

29. S. Shen and C. Chang, "A utility-based scheduling algorithm with differentiated QoS provisioning for multimedia CDMA cellular networks," in *VTC'04 Spring*,vol.3,pp. 1421-1425.
30. K. Johnsson and D. Cox, "An adaptive cross-layer scheduler for improved QoS support of multiclass data services on wireless systems," *IEEE JSAC*, vol. 23, no. 2, pp. 334-343, Feb. 2005.
31. W. Wong et al., "Soft QoS provisioning using the token bank fair queuing scheduling algorithm," *IEEE Wireless Commun.*, vol. 10, no. 3, pp. 8-16, 2003.
32. A. Kam et al., "Supporting Rate Guarantee and Fair Access for Bursty Data Traffic in W-CDMA," *IEEE JSAC*, vol. 19, no. 11, pp. 2121-2130, 2001.
33. Q. Pang et al., "Service scheduling for general packet radio service classes," *IEEE WCNC 1999*, vol. 3, pp. 1229-1233, Sept. 1999.
34. V. Huang et al., "QoS-Oriented Packet Scheduling for Wireless Multimedia CDMA Communications," *IEEE Trans. Mobile Comp.*, vol. 3, no. 1, pp. 73-85, 2004.