

METHOD AND SYSTEM FOR ALLOCATING RESOURCES

The invention relates to a method and a system for allocating resources. It
5 applies in a communications network, particularly to communications in a
wireless network.

The present invention finds its application in the field of wireless networks
operating on the principle of Time Division Multiple Access or "TDMA".

10

Amongst other things, and in a nonexclusive manner, it applies to the
communications in an "ad hoc" distributed network and in particular in a network
of the ad hoc type whose data has an urgently nature.

15 There are several techniques making it possible to share resources between
the various users or services of a network. Frequency-Division Multiple Access
or FDMA, for example, consists in dividing the bandwidth of a communication
bearer into distinct frequency bands. Time division multiple access (TDMA)
consists in dividing the time into small intervals or time windows sometimes
20 called "slots" and in allocating a given time slot to each service or user. The
users then transmit data on different time slots.

In an ad hoc network, operating on the TDMA principle, a station or a network
node plays the role of network coordinator by allocating the transmission
25 moments to the various users or services according to their requirements. This
node may be called the central node, master node, "scheduler" or "cluster
head". Each network node, for example, may be associated with a user or a
service. Ad hoc networks have the advantage of allowing high bit rate transfers
and the possibility of interconnection with existing structures.

30

The use of this type of network requires very precise and fair allocation
management between the users or the services. In addition, in such a network,
the data to be transmitted do not all have the same requirements. For example,
voice "times out" rapidly (in approximately 250 ms) and the transmission delays

must therefore be minimized. On the other hand, the loss of a few packets of this data type does not significantly harm the quality of transmission. On the contrary, for the background traffic class which corresponds to secondary traffic with no real time constraints, such as for example email, the data may be subjected to delays but it is imperative that all the packets are transmitted. The requirements expressed by the nodes may be associated with the quality of service (QoS) of the associated service. The quality of service is specified by QoS parameters expressing features provided for data transfer such as the bit rate, transit time, precision and reliability.

10

Various queue management methods for the allocation of resources are currently known in the prior art, for example:

- First in First out or "FIFO";
- static scheduling with absolute priorities or "priority queuing";
- 15 • fair scheduling of packets according to the detected flows or "weighted fair queuing" or "flow based queuing";
- scheduling according to a determined traffic class or "class-based queuing or CBQ". The classes may be based on a number of parameters such as the priority, interface or originating program; and
- 20 • static scheduling with relative priorities computed on a bandwidth guarantee request or "Custom queuing".

US patent 2003/0198204 discloses a method and system for allocating resources for the transmission of data subjected to quality of service (QoS) constraints. This system makes it possible to schedule traffic initially based on static allocation parameters then, by adapting this order thanks to compensation coefficients computed according to the flow, the weighting time in the queue and the priority.

30 This approach is well suited to traffic of the isochronous type such as video. However, in the case of non deterministic sources (voice for example depends on speakers), compensation does not seem sufficient to make the approach usable. The adaptation is carried out little by little and cannot respond

immediately to greatly variable traffic.

Furthermore, in this system, the concept of timing out of data or of delay is not absolute and therefore does not make it possible to guarantee delays imposed
5 by certain applications. For example, voice which "times out" rapidly.

Although they are effective, the various resource-allocation techniques in the prior art have a certain number of disadvantages:

- 10 - they do not guarantee scheduling according to the real urgency of the data;
- they do not support absolute classification according to the concepts of delay or timing out of data and therefore they do not make it possible to guarantee delays imposed by certain applications; and
- 15 - they do not solve the problems of scheduling for highly variable and non deterministic traffic.

In the description, "node" or "station" means any network entity implemented in hardware or software that may be the input point or output point of an interchange or of a transmission of data or information between stations.

20 A request characterizes a requirement for resource, such as for example a slot, associated with an application or a class of service expressed by a network node to an allocation mechanism (the scheduler). A requirement may indicate notably a specification of end-to-end quality of service.

25 The invention relates to a method of allocating resources in a communication network comprising several nodes adapted to transmit and receive data and a central node adapted to allocated resources, characterized in that it comprises at least the following steps:

- 30 - at least one node transmits one or more requests R_{nqt} to the central node, each request R_{nqt} being associated with a requirement q_i ;
- the central node allocates a lifetime Δt_{q_i} to each request R_{nqt} according to the requirement q_i of the request, and

- each request R_{nqt} is scheduled in a queue according to the lifetime Δt_{qi} and the moment of receipt t_R of the request,
- the first m requests R_{nqt} in the queue are retrieved for m available transmission moments,
- 5 - the requests retrieved are scheduled per node n ,
- the distribution of the transmission moments is written in a signaling flag and then transmitted from the scheduler to the nodes.

10 In one embodiment of the invention, the lifetime may be allocated to a request according to the application associated with the request. So the requests associated with urgent or important applications may be placed at the beginning of the queue.

15 The lifetime may be allocated to a request according to a quality of service (QoS) associated with the request. This allows a classification of requests according to the quality of service of the associated application and makes it possible to comply with the priorities between the various quality of service levels. Therefore, high priority requests may be favored in the classification of
20 the queue.

The lifetime may be allocated according to the time constraints of the application associated with the request or according to the time allowed for the transfer of the data associated with the request. This makes it possible to satisfy
25 the most pressing requests and to supply scheduling reflecting the real urgency of the packets.

The network is for example an ad hoc network. The lifetime allocated to a request may depend on the number of network nodes that the data to be
30 transmitted must pass through. This makes it possible to ensure transmission of the data from end to end of a network.

It is possible to use a TDMA protocol. Therefore, the resources to be allocated may be the moments of transmission.

In one embodiment for equal moments of expiration, the requests in the queue are scheduled according to the quality of service of the requests, a moment of expiration for a request being defined based on the lifetime allocated to the request and the moment of receipt of the request.

The method may comprise a step consisting in discriminating between the perishable requests and the nonperishable requests. The method may, for example, also comprise a step in which it is verified whether the requests are timed out, and the timed-out request(s) is (are) removed from the queue. This allows the management of timed-out data and hence makes it possible to respond to time constraints imposed by the services associated with the requests.

The checking of the requests may, for example, consist in comparing the moment of expiration of each request with the current moment and if the moment of expiration of a request is less than the current moment, the request is removed.

In another embodiment of the invention, each request is allocated a lifetime at the output and a comparison is made between a moment of expiration at the output, defined based on the lifetime at the output of the request, and the moment of receipt of the request, and the current moment. If the moment of expiration at the output of a request is less than the current moment, the request is removed.

The network is for example a wireless network.

In one embodiment, it is possible for the lifetime allocated to a request to be adjusted according to the priority or the time constraints associated with the request. This makes it possible to provide these requests with a competitive advantage in access to the resource since they will be systematically favored in the classification in the queue.

The invention also relates to a system for allocating resources in a communication network comprising several nodes suitable for transmitting and receiving data and a central node suitable for allocating resources, characterized in that several nodes are fitted with a processor capable of:

- 5 - transmitting one or more requests R_{nqt} to the central node, each request R_{nqt} being associated with a requirement q_i , and in that the central node is fitted with a processor capable of:
 - allocating a lifetime Δt_Q to each request R_{nqt} received by the central node according to the requirement q_i of said request R_{nqt} ;
 - 10 and
 - scheduling each request R_{nqt} received in a queue according to the lifetime Δt_{q_i} and the moment of receipt t_R of said request R_{nqt} ,
 - the first m requests R_{nqt} in the queue are retrieved for m available transmission moments,
 - 15 the retrieved requests are scheduled per node n ,
 - the distribution of the transmission moments is written in a signaling flag and then transmitted from the scheduler to the nodes.
- 20 The method and the system for allocating resources according to the invention notably has the following advantages:
 - they make it possible to optimize the allocation of resources;
 - they allow an allocation that is fair between the nodes of a network;
 - they make it possible to satisfy the priorities of the services associated
 - 25 with the nodes;
 - they make it possible to provide guarantees of service in terms of bit rate and delay, whether it involves point-to-point or end-to-end transfers; and
 - they make it possible to take account of the time constraints of the data to be transmitted and ensure scheduling according to the real urgency of
 - 30 the data.

In addition, the data are not interlaced by priority or by flow but by real time constraints and therefore by their real degree of urgency. This interlacing is dependent on the data and not the flows, and is therefore naturally suitable to

the "burst" and non-deterministic character of traffic such as voice, since it acts packet by packet.

Other features and advantages of the present invention will appear more clearly on reading the following description given as an illustration and being in no way limiting, with reference to the appended figures in which:

- figure 1 schematizes an example of an ad hoc network with several nodes and a central node;
- figures 2A to 2C illustrate data interchanges in the network of figure 1;
- 10 - figure 3 details the steps of the method according to one embodiment of the invention;
- figures 4A to E illustrate the method detailed in figure 3;
- figure 5 illustrates a data checking step according to a variant of the invention; and
- 15 - figure 6 illustrates a data checking step according to another variant of the invention.

In order to make the subject of the invention more comprehensible, the following example, given as an illustration, relates to an ad hoc communications network.

Figure 1 schematizes an example of an ad hoc network 1 operating on the TDMA (time division multiple access) principle. The network 1 comprises a central node 2 and a plurality of nodes 3 connected together and to the central node 2 by communication links. Each node 3 is fitted with transmit/receive means and is associated with a service or a user. The central node 2 is also fitted with transmit/receive means and plays the role of coordinator or "scheduler" of the network by allocating transmission moments to the nodes 3 according to their requirements q. The requirement q may correspond to a quality of service.

The nodes 3 may be mobile or non mobile, for example, microcomputers, wireless telephones, digital television sets. The communication links may be either wire links, or wireless links, for example, radio links.

It will be noted that the method according to the invention also applies for any other communication protocol.

- 5 Figures 2A to 2C illustrate in a general manner the data interchanges in the network of figure 1.

10 In a first step, illustrated in figure 2A, each node 3 transmits at least one request, in the form of a data packet containing the requests for transmission moments, to the scheduler 2. A request corresponds to a data packet that contains a request for a moment or some moments of transmission for a given node 3.

15 The requirements for transmission moments are specified in a two-dimensional space (Time and Importance, quality of service) originating from the applications associated with the nodes via a system = known per se for specifying the characteristics of a data stream, such as the use of the TOS "type of service" field of the IP packet header, the interpretation of the port of their TCP/UDP headers which indicates the corresponding application or the
20 use of multiple queues.

Based on the requests received, the scheduler 2 determines the distribution of the transmission moments as will be explained below.

- 25 The scheduler 2 transmits to the nodes 3 the transmission moment distribution that has been defined, as shown in figure 2B.

30 The nodes 3 use the transmission moments that have been allocated to them for transmitting data packets to their intended recipient as illustrated in figure 2C.

Figures 3, 4A to 4C illustrate an example of the method of allocating transmission moments for the various nodes. This method comprises the following steps:

In a first step 31, the scheduler 2 receives requests for resources or requests R_{nqt} issued by the nodes 3 where n represents the node requesting the resource, q the quality of service (QoS) associated with this request R_{nqt} and t the moment of receipt of the request R_{nqt} by the scheduler 2. Figure 4A illustrates a first series of requests R_{130} , R_{120} , R_{120} , R_{100} , R_{230} , R_{230} , R_{220} and R_{210} from two network nodes N1 and N2, the requests being classified according to their quality of service q_0 to q_3 .

In a second step 32, the scheduler 2 allocates a lifetime Δt_{qi} to each request R_{nqt} , according to the quality of service q_i associated with the request R_{nqt} . The lifetime Δt_{qi} allocated to each request R_{nqt} is chosen, for example, according to the time constraints of the application associated with the request. Figure 4A illustrates the allocation of the lifetime Δt_{qi} to the first series of requests received by the scheduler at a moment T_0 . To the requests associated with quality of service q_3 , a lifetime $\Delta q_3=0$ is allocated, and to the requests associated with quality of service q_2 , q_1 and q_0 a lifetime Δq_2 , Δq_1 and Δq_0 is allocated respectively where $\Delta q_0 > \Delta q_1 > \Delta q_2$. Figures 4B and 4C illustrate the allocation of the lifetimes associated with a second series of requests R_{111} , R_{231} , R_{221} , R_{201} and a third series of requests R_{132} , R_{122} , R_{232} , R_{212} , R_{202} , received respectively by the scheduler at the times T_1 and T_2 .

In a third step 33a, 33b, the scheduler 2 retrieves the requests R_{nqt} starting from the most pressing to the least pressing. It then schedules the requests R_{nqt} in a queue observing the moments of expiration of the requests as shown in figure 4D. The moment of expiration for a request R_{nqt} is defined by $t + \Delta t_{qi}$.

The order for two requests R_{n1q1t1} and R_{n2q2t2} having respectively a lifetime Δt_{q1} and Δt_{q2} is imposed by the following relation:

30

$$t_1 + \Delta t_{q1} < t_2 + \Delta t_{q2} \Leftrightarrow R_{n1,q1,t1} < R_{n2,q2,t2}$$

For the equal moments of expiration $t + \Delta t_{qi}$, the requests are discriminated by

qualities of service q associated with the requests. That is to say that it gives:

$$t_1 + \Delta t_{q1} = t_2 + \Delta t_{q2}, q_1 > q_2 \Leftrightarrow R_{n1,q1,t1} < R_{n2,q2,t2}$$

- 5 The highest qualities of service q (priorities increasing according to their importance) are then favored and placed in front, during the scheduling in the queue when the moments of expiration $t + \Delta t_{qi}$ of the requests are equal. The method therefore provides a queue in which the high priority requests are favored, giving an order satisfying the real urgency of the requests.

10

In a fourth step 34a and 34b, a check is made that the requests R_{nqt} in the queue are not timed out. This checking step consists in comparing the current moment T_c with the moment of expiration $t + \Delta t_{qi}$ of the request R_{nqt} . If the moment of expiration $t + \Delta t_{qi}$ of the request R_{nqt} is less than the current moment

- 15 T_c , the request R_{nqt} is removed from the queue as shown in figure 4E.

In a fifth step 35, the first m requests R_{nqt} in the queue are retrieved for m available transmission moments as illustrated in figure 4E.

- 20 In a sixth step 36, the retrieved requests are scheduled per node n as shown in figure 4E.

In a seventh step 37, the distribution of the transmission moments is written in a signaling flag and then transmitted from the scheduler to the nodes.

25

The method according to the invention described above may be carried out according to many variants. Therefore, according to certain variants, the method does not comprise the fourth step 34a and 34b of checking and removing the timed-out requests.

30

According to a variant of the invention, it is possible to refine or adjust the lifetime Δt_{qi} allocated to the requests. Accordingly, it is possible to reduce, or even nullify a lifetime Δt_{qi} for the applications with very high time constraints or

high priority, for example voice. This makes it possible to give these requests a competitive advantage in access to the resource since they will be systematically favored in the classification in the queue.

- 5 According to another variant of the invention, the checking step consists in defining distinct constraints at the output from the constraints imposed for the scheduling of the list. For this, it is possible to allocate a lifetime at the output Δt_{si} to each request. Since the lifetime at the input Δt_{qi} aims, depending on the case, only at giving preference to the requests of such or such a priority or QoS,
10 it does not always correspond to the real acceptable constraints. It is possible, for example, for the requests with very strong time constraints for which zero lifetimes have been imposed for scheduling (voice, for example), to authorize a margin at the output corresponding to the real acceptable lifetime in order to allow the allocation. In this case, the checking of the timed-out requests
15 consists in comparing the moment of expiration at the output, defined based on the time of life at the output and the moment of receipt of the request, with the current moment. Therefore, the requests with a moment of expiration at the output that is less than the current moment will be removed from the queue.
- 20 In the case of data that must be transmitted whatever happens, such as for example the background traffic, the method may also comprise a step of discrimination between perishable requests and non-perishable requests.

Figure 5 illustrates a variant of the invention in which the method comprises a
25 step 52 of discrimination between perishable requests and nonperishable requests and a checking step 53 in which the perishable requests are then checked, according to one of the checking procedures described above, in order to check whether they are timed out or not. The non-timed-out and nonperishable requests, such as for example the requests associated with
30 email, are kept in the queue 54 for the transmission of the allocation to the nodes.

Figure 6 illustrates another variant of the invention in which the method comprises a checking step 63 in which the nonperishable requests are

rearranged in the queue 61 and the timed-out requests are removed if there are no more transmission moments to be allocated. This makes it possible to manage queue congestion and to ensure the highest quality of service requests.

5

The step of checking the perishable and nonperishable requests may consist in allocating an artificial lifetime to the nonperishable requests. This lifetime is chosen, for example, according to the priority of the traffic, making it more or less competitive, but is not taken into account by the output check. The nonperishable requests are added to the requests allocated according to the order of time urgency without there being any check on their possible timing-out.

10

According to yet other variant embodiments, the lifetime may be allocated to each request or refined according to one or more of the following criteria: the application associated with the request, the time constraints associated with the application and/or the number of network nodes that the data must pass through.

15

This method may be interfaced with a "sessions" mechanism of the "connected mode type" MAC level or of the "RSVP type" network level that is capable of providing the characteristics of urgency and importance of the data to be transmitted.

20

Note that the method according to the invention may, without departing from the context of the invention, be applied to any other type of network and to propagation channels of any nature.

25

1. A method of allocating resources in a communication network (1) comprising several nodes (3) suitable for transmitting and receiving data and a central node (2) suitable for allocating resources, characterized in that it comprises at least the following steps:

First step: at least one node (3) transmits one or more requests (R_{nqt}) to the central node (2), each request (R_{nqt}) being associated with a requirement (q_i); where n represents the node requesting the resource, q the quality of service (QoS) associated with this request R_{nqt} and t the moment of receipt of the request R_{nqt}

Second step : the central node (2) allocates a lifetime (Δt_{qi}) to each request (R_{nqt}) according to the requirement (q_i) of the request, and

Third step : each request (R_{nqt}) is scheduled in a queue according to the lifetime (Δt_{qi}) and the moment of receipt (t_R) of the request, starting from the most pressing to the least pressing, and it then schedules the requests R_{nqt} in a queue observing the moments of expiration for a request R_{nqt} defined by $t + \Delta t_{qi}$; for equal moments of expiration $t + \Delta t_{qi}$, the requests are discriminated by qualities of service q associated with the requests

Fourth step : a check is made that the requests R_{nqt} in the queue are not timed out, consisting in comparing the current moment T_c with the moment of expiration $t + \Delta t_{qi}$ of the request R_{nqt} , if the moment of expiration $t + \Delta t_{qi}$ of the request R_{nqt} is less than the current moment T_c , the request R_{nqt} is removed from the queue,

Fifth step: the first m requests R_{nqt} in the queue are retrieved for m available transmission moments

Sixth step : the requests retrieved are scheduled per node n , the distribution of the transmission moments is written in a signalling flag and then transmitted from the scheduler to the nodes.

2. The method as claimed in claim 1, characterized in that the lifetime (Δt_{ql}) is allocated to a request (R_{nqt}) according to the application associated with the request (R_{nqt}).
- 25 3. The method as claimed in either of claims 1 and 2, characterized in that the lifetime (Δt_{ql}) is allocated to a request (R_{nqt}) according to a quality of service (QoS) associated with the request (R_{nqt}).
- 30 4. The method as claimed in claim 1, characterized in that the lifetime (Δt_{ql}) allocated to a request (R_{nqt}) depends on the number of network nodes that the data to be transmitted must pass through.

5. The method as claimed in any one of the preceding claims, characterized in that a TDMA protocol is used and the resources to be allocated are the moments of transmission.

5 6. The method as claimed in any one of the preceding claims, characterized in that, for equal moments of expiration, the requests (R_{nqt}) in the queue are scheduled according to the quality of service (QoS) of the requests (R_{nqt}), a moment of expiration for a request being defined based on the lifetime (Δt_{ql}) allocated to the request (R_{nqt}) and the moment of receipt (t_R) of the request
10 (R_{nqt}).

7. The method as claimed in any one of the preceding claims, characterized in that it comprises a step consisting in discriminating between the perishable requests and the nonperishable requests.

15

8. The method according to any one of the preceding claims, characterized in that it comprises a step in which it is verified whether the requests (R_{nqt}) are timed out, and the timed-out request(s) is (are) removed from the queue.

20 9. The method as claimed in either one of claims 7 and 8, characterized in that the verification of the requests (R_{nqt}) consists in comparing the moment of expiration of each request with the current moment and if the moment of expiration of a request is less than the current moment, the request (R_{nqt}) is removed.

25

10. The method as claimed in one of claims 7 to 9, characterized in that a lifetime is allocated to each request (R_{nqt}) at the output and a comparison is made between a moment of expiration at the output, defined based on the time of life at the output of the request (R_{nqt}), and the moment of receipt of the
30 request (R_{nqt}) and the current moment and, if the moment of expiration at the output of a request (R_{nqt}) is less than the current moment, the request (R_{nqt}) is removed.

11. The method as claimed in any one of the preceding claims, characterized

in that the network (1) is a wireless network.

12. The method as claimed in any one of the preceding claims, characterized in that the lifetime (Δt_{qi}) allocated to a request (R_{nqt}) can be adjusted according to the priority or the time constraints associated with the request (R_{nqt}).

13. A system for allocating resources in a communication network (1) comprising several nodes (3) suitable for transmitting and receiving data and a central node (2) suitable for allocating resources, characterized in that several nodes (3) are fitted with a processor capable of:

- transmitting one or more requests (R_{nqt}) to the central node (2), each request (R_{nqt}) being associated with a requirement (q_i), and in that the central node is fitted with a processor capable of:
- allocating a lifetime (Δt_Q) to each request (R_{nqt}) received by the central node (3) according to the requirement (q_i) of said request (R_{nqt}); and
- scheduling each request (R_{nqt}) received in a queue according to the lifetime (Δt_{qi}) and the moment of receipt (t_R) of said request (R_{nqt}),
- the first m requests (R_{nqt}) in the queue are retrieved for m available transmission moments, the retrieved requests are scheduled per node n ,
- the distribution of the transmission moments is written in a signaling flag and then transmitted from the scheduler to the nodes.

14. The system as claimed in claim 13, characterized in that the network (1) is a wireless network.