On token protocols for high-speed multiple ring networks

Włodek Dobosiewicz* Paweł Gburzyński†

To be presented at ICNP'93

Abstract

A token ring protocol which restricts the transmission rights to the station possessing the token is not a good candidate for driving a gigabit network. This is due to the fact that the periods while the token is "in transit" are wasted: they must be subtracted from the effective throughput achievable by the network. As the network's transmission rate becomes higher, the impact of the token transition time becomes more pronounced and the effective throughput of the network becomes a smaller and smaller fraction of the nominal channel capacity. In this paper, we present a collection of token passing protocols for ring networks which are devoid of this unpleasant property. The protocols are simple and inexpensive, yet they possess a number of advantageous properties as high capacity (similar to METARING and independent of the propagation length of the ring), fairness, and natural accommodation of synchronous and isochronous traffic. Our protocols operate on multiple rings. As these rings can be virtual, i.e., obtained by a logical division of a single physical ring, the proposed solutions can be naturally combined with wave division multiplexing techniques.

1 Introduction

To focus our discussion on the logical aspects of the investigated protocols, we will use the *normalized propagation delay* (expressed in bits) to measure time, distance, and packet length [9]. We also abstract from coding issues, which may otherwise obscure the discussion, and assume that signals transmitted through the channel (ring) consist of bits. In particular, an FDDI symbol (encoded using the 4B5B code—[6, 7]) is composed of 4 bits, not 5. This way we are not concerned here with the actual implementation of the physical layer, although we will assume that this layer offers us a collection of reasonable prerequisites.

Let L denote the round-trip propagation delay of the ring (expressed in bits) and l_p be the average length of a packet transmitted through the network. The ratio L/l_p is commonly denoted by a. A large value of L, and consequently a, may result from a very high transmission rate of the network, a

*Supported in part by NSERC Grant No. OGP9110. Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2H1. email: dobo@cs.ualberta.ca.

†Supported in part by NSERC Grant No. OGP9183. Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2H1. email: pawel@cs.ualberta.ca.

very long geographic length of the ring, or a combination of these two properties. The performance of many protocols (not only ring protocols) deteriorates when a becomes significantly bigger than one. For example, consider the following simple formula expressing the maximum throughput achievable by FDDI:

$$T_{FDDI} = \frac{\Sigma THT}{\Sigma THT + L} \tag{1}$$

where ΣTHT represents the sum of the token holding times at all stations (and can be viewed as a representative of the packet length l_p). Although one may naively try to increase the maximum throughput of FDDI by increasing ΣTHT, this approach has obvious limitations: it assumes that each station has sufficiently many ready packets to fill its token holding window; moreover, it increases the average ring access time.

Ideally, the formula expressing the maximum throughput of a ring network should be independent of L (or at least its value should not be inversely proportional to L). Protocols with this property are called *capacity–1* protocols: they are able to utilize a fixed portion of the channel bandwidth, irrespective of the propagation length of the network. To demonstrate that designing protocols with this property is not trivial let us consider METARING [2] which is one of the best protocols for ring networks known to the authors. The maximum normalized throughput of "pure" METARING (without the SAT mechanism) does not depend on L. Assuming a uniform distribution of traffic, this throughput is equal to 8 (4 per each of the two rings), which is the upper bound on the throughput of any network based on two counter-rotating rings. Unfortunately, in its pure version, METARING is starvation prone. If starvation is defined as a situation in which a station may be forced to let pass $\mathcal{S}$ packets before it can transmit one packet (in a given ring), the probability of starvation equals $\left(1 - \frac{2}{N-1}\right)^{\mathcal{S}}$, assuming uniform traffic, a heavy load and an odd number of stations N. This is a quickly diminishing function of $\mathcal{S}$, but the potential of starvation may cause problems in time-sensitive applications, such as real-time or synchronous traffic. The probability of starvation may, of course, be much larger for nonuniform traffic, especially for biased traffic. This issue is partially addressed in [1] and [12]; in this paper, we propose a completely different approach.

To eliminate starvation in METARING, a special mechanism (SAT) is introduced imposing a limit k on the number of packets that a station can transmit before it is obliged to yield the bandwidth to other backlogged stations. With this mechanism, the maximum throughput of METARING is given by:

$$T_{META} = min\left(\frac{2Nk}{L}, 8\right) \tag{2}$$

where N is the number of stations in the network and k is the above-mentioned limit. Clearly, this formula depends on L in the "unpleasant" way. Although one may observe that by increasing k, the impact of L can be formally eliminated, the protocol's fairness and its ability to respond to varying load patterns is impaired in direct proportion to k. When $k = \infty$, the SAT mechanism is turned off and the network becomes starvation-prone. On the other hand, any criticism of METARING in this respect should be put into a proper perspective: at the level of throughput that its main competitors (today) can handle, METARING guarantees a fair and starvation-free access to the media; however, it happens at a much greater hardware expense per station. It would be nice to be able to produce a formula for this tradeoff, but in the fast-pace world of electronics, it is totally impossible.

The above example is rather typical. The trade-off between fairness and maximum effective throughput (and, implicitly, cost) becomes clearly visible for higher values of L. Intuitively, fairness is often achieved by a direct or indirect form of negotiating medium access across the network. Such negotiations incur access delays proportional to L. If stations are not allowed to use the bandwidth before it has been negotiated, the maximum effective throughput must deteriorate with increasing L. On the other hand, stations trying to take advantage of idle negotiation periods—to improve throughput—run the risk of preempting other needy stations in an uncontrolled (unnegotiated) way.

Unfair protocols with a formal *capacity–1* property are not difficult to conceive. Formally, a protocol in which the entire bandwidth remains constantly allocated to one permanently ready station (e.g., FDDI with $TTRT = \infty$) is a *capacity–1* protocol: its heavy-load behavior does not depend on the propagation length of the network. Of course, nobody serious will consider this approach in reality, but less explicit trade-offs may pass unnoticed and blur the comparison of protocols with different merits. To avoid this confusion, we formulate below a list of properties that we would expect from an ideal protocol.

1. The protocol must be simple enough to be implemented directly in hardware and to operate at a speed that can exploit the bandwidth offered by optical fiber channels.

2. It must be fair and starvation-free. Under uniform load, both the mean packet delay and the maximum throughput of each station must be independent of the location of that station with respect to other stations. Under non-uniform load, the contribution of each station to the global throughput of the network should be proportional to the station's load. The protocol may define various priority schemes to bias access rights of different stations, but there should be no bias if

all stations have the same priority.

3. The total throughput of the network (measured in the ratio of bits received to the time elapsed, expressed in bits) must be independent of the network size and of the transmission rate.

4. The average medium access delay must approach 0 as the global load of the network approaches 0.

5. Consider the following traffic scenario:

   > In an alternating sequence, (1) a selected station A has a burst of messages to send to a selected receiver B, while all the other stations generate a low intensity uniform traffic; (2) all the stations generate a low intensity uniform traffic. These two phases have a duration such that the average total load offered is at saturation point and the duration of phase (1) is equal to $L/2$.

   The protocol should be fair in this scenario with respect to the selection of stations A and B, i.e., the performance of A should neither depend on the selection of A nor on the selection of B. This fairness should be achieved without a centralized processing of feedback from stations, as such processing does introduce delays proportional to L.

6. The protocol must be able to accommodate heterogeneous traffic demands, guaranteeing simultaneously a finite maximum packet delay for synchronous traffic and a sustainable throughput for asynchronous traffic.

7. It must be able to carry synchronous traffic of variable intensity, up to using the whole bandwidth of the network.

8. It must be self-synchronizing, so that jitter remains negligible.

9. It must be predictable, so that a critical failure can be recognized by at least one station in time not exceeding L.

None of the existing protocols for ring networks is perfect; therefore, no protocol fulfills all the above-listed postulates. For example, FDDI violates postulates 3, 4, 7, 8, and 9 whereas METARING does not fulfill postulates 2 or 3, depending on the version, and also may violate postulates 8 and 9.

In this paper, we propose a number of protocols for ring networks which come very close to possessing all the properties of an ideal protocol. We focus on the *capacity–1* property, fairness, and accommodation of synchronous traffic. These properties are most important from the viewpoint of gigabit applications. The proposed protocols are also simple, flexible, and predictable.

2 Protocols

2.1 Strong token versus tokenless protocols

Any MAC-layer protocol for a ring network has two basic responsibilities:

- To determine when a given station is granted medium access (i.e., it is allowed to transmit).
- To specify how the ring is cleaned, i.e., how packets are removed from the ring.

For example, in FDDI, the transmission rights are restricted to the station currently possessing the token. Since at most one station at a time has these rights, the transmitting station has exclusive active access to the ring.[1] A token protocol with such transmission rules will be called a *strong token* protocol. The station holding the token also disconnects the ring. All traffic arriving at the disconnected end of the ring disappears from the network. FDDI specifies other cleaning rules. A station recognizing the header of its own transmitted packet is supposed to strip the remainder of the packet, i.e., remove it from the ring. This feature has no impact on network performance: the periods of silence in the medium resulting from stripped packets cannot be reused until they hit the token holding station.

In METARING, any station can start transmission at any moment, provided that certain criteria are met. As these criteria are typically fulfilled by many stations at the same time, multiple stations can transmit packets simultaneously. Cleaning is done by receivers. To be able to erase its packet completely from the ring, the receiver must buffer at least some initial portion of the packet before it can determine whether the packet should be relayed. Additionally, METARING includes the possibility of preemption where a station forces a packet of its own in front of an incoming packet that it is supposed to relay.[2]

[1] An almost obvious idea: to have more than one token in FDDI does not work, since the location of a token is unpredictable in FDDI. See, however, [8].

[2] We do not discuss preemption in this paper, since this feature may be combined with many other protocols as well.

Our protocols attempt to combine the best features of the strong token approach and token-less solutions as METARING. Let us begin with listing the advantages and weak spots of the two concepts:

Strong token

Advantages:

- Both transmission and cleaning rules are very simple and packets don't have to be buffered at intermediate stations (postulate 1).
- The circular movement of the token provides a natural means for enforcing fairness (postulate 2).
- If each station always holds the token for the same amount of time, the protocol is self-synchronizing and predictable (postulates 8 and 9). Note that FDDI, with its insisting on variable token-holding times, does not have this property.

Problems:

- Wasted bandwidth due to token transition periods (postulate 3).
- The average medium access time is no less than $L/2$, even for very light load (postulate 4).
- Trade-off between flexible and rigid allocation of token holding time. If the token holding time is flexible, the network is better utilized, but less synchronized and less predictable (postulates 8 and 9). On the other hand, with a rigid allocation of token holding time, the network is better synchronized and more predictable, but the bandwidth unused by one station cannot be reclaimed by another one (postulates 6 and 7).

Token-less approach with destination cleaning

Advantages:

- Better utilization of bandwidth, possibly independent of L (postulate 3).
- Zero access delay under light traffic conditions (postulate 4).
- Potential flexibility of bandwidth allocation (postulates 6 and 7).

Problems:

- Trade-off between high bandwidth utilization and fairness (postulates 2 and 5).
- Complexity (postulate 1).
- Poor predictability and synchronization (postulates 8 and 9).

2.2 Weak token protocols

Our protocols are based on token passing. The role of the token is to guarantee that the protocol has all the advantages of the strong token approach, even those that most actual strong token protocols do not have, i.e., good predictability and synchronization. In contrast to strong token protocols, our token is *weak* and its possession is not a necessary condition for transmission.

The target version of the protocol is driven by multiple tokens. Therefore, we will call it the *multiple weak token* protocol, or MWT for short (multiple tokens were introduced in a number of papers, notably, from the perspective of this paper, in [8] and [3]). MWT demonstrates how to achieve the benefits of destination cleaning (i.e., high throughput, low access delays) and rigid strong token (fairness, predictability) without actual destination cleaning. With MWT, packets are not buffered at intermediate stations and the cleaning rules are extremely simple (restricted to cleaning while holding the token). The protocol exhibits very good characteristics for synchronous traffic and allocates bandwidth in a flexible way (there is no need to set aside a fixed portion of the bandwidth reserved for synchronous traffic).

In this paper, the protocols are presented in their slotted versions. As in METARING, our protocols can be unslotted or semi-slotted (fixed length packets without explicit slot markers). The unslotted versions may be slightly trickier to implement than the slotted ones, due to more stringent requirements on the clock accuracy. The slotted versions are conceptually simpler; they are also easier to present.

2.2.1 Single weak token

Assume that the network consists of a single ring and there is exactly one token circulating in the network. The token can be held by a station for a prescribed amount of time (number of slots). A station holding the token ignores the incoming traffic. It is also responsible for generating empty slots and inserting them into the ring.

The slot header includes two special bits. The *full* bit indicates whether the slot carries a packet[3] (the *full* bit is 1) or is empty and can be used to insert a packet (*full*= 0). The other bit is used to pass the token. Normally, the *token* bit is set to 1 by the slot generator (the station currently holding the token), unless the station decides to pass the token to its successor, in which case it sets the token bit to 0.

To acquire an empty slot for transmission, a backlogged station monitors all slot headers passing by and sets the *full* bit to 1, simultaneously reading its previous value.[4] If the previous value was zero, the station knows that it has acquired the slot. Then, it fills the slot with a segment. The same mechanism is used for token acquisition. A station receiving a slot sets the *token* bit to one and, at the same time, reads its previous value. If the previous value happens to be 0, the station knows that it has acquired the token.

With this approach, slots are not buffered at stations. The smallest possible repeater delay is sufficient to perform the simple acquisition operations described above. There is no explicit destination cleaning; thus, a station has no need to recognize the destination address in the segment header before relaying a slot.

Having passed the token, the station has to resynchronize to the slots coming from the input port of the ring. Thus, the station should delay the actual re-connection until it sees the beginning of a slot arriving on the input port. An alternative solution is possible in which a token holding station does not insert new slots and does not disconnect the ring. With this approach, the network is filled with slots during the initialization phase. From then on, the slots permanently circulate in the network. A token-holding station does not disconnect the ring—it just clears (unconditionally) the *full* bits of the incoming slots. The network may need a special "manager" station responsible for the initial "loading" of the network with slots and later monitoring the network—to detect failures requiring slot insertions.

The single-token version of our protocol (dubbed SWT for *single weak token*) operates as follows.

The length of the ring L expressed in slots is known to all stations; this parameter may easily be calculated by each station at initialization time. The current number of active[5] stations in the

[3]Following the terminology of DQDB [5], we will call it a segment.

[4]The same mechanism is used in DQDB.

[5]Relaying slots and holding the token when their turn comes, as opposed to stations in bypass mode, which do not interfere with the signal in the medium.

network N is also known; while this number varies in time, stations are made aware of changes by an *early token* phenomenon (not to be confused with the *early token* in FDDI).

Every station holds the token for a fixed amount of time, irrespective whether it has a segment awaiting transmission or not. We assume that the token holding time (expressed in slots) is the same for all stations. We will denote it by THT. It is possible to assign priorities by using different (but fixed) token holding intervals for different stations.

Every station counts the time (in slots) elapsed since the moment the station last released the token. To accomplish this the station maintains a counter called TT_i, where i is the station index. This counter is set to zero at the moment when the station passes the token (at the beginning of the slot in which the token is passed) and then incremented by 1 whenever the station receives a new slot.

There is only one simple transmission rule. A station i ready to transmit waits for the moment when the following condition is satisfied:

$$TT_i >= (N-1) \times THT \tag{3}$$

Starting from this moment, the station transmits the segment in the first free slot.

Note that the above condition is satisfied for a station holding the token. Since TT_i continuously increases while a station is waiting for the token, the station is guaranteed to transmit its packet after a bounded delay. Condition (3) ensures that the transmitted segment will not be absorbed by a token holding station, before it has reached the most distant destination with respect to station i.

The cleaning rules are similar to FDDI: besides token cleaning, a station transmitting in a given slot is responsible for emptying that slot when it arrives back at the station, L slots after it was filled with a segment. Note that the cleaning station can determine which slot should be emptied by simply counting slots since its transmission. As a station may transmit several segments before the first of those segments arrives back at the station, it should maintain a FIFO queue of slot numbers (modulo L) identifying the slots to be cleaned. The cleaning operation is trivial: it consists in resetting unconditionally the *full* bit in the slot header to 0. Instead of emptying its slot, a backlogged station may reuse it immediately to transmit another segment, provided that condition (3) is satisfied.

The simple protocol described above has a number of interesting properties. First of all, it is a *capacity–1* protocol. Let $TRT = N \times THT + L$. The maximum throughput of SWT depends on the relationship between the values of L and TRT. A maximum throughput of 1 can be achieved by tuning the network in such a way that L is a multiple of $TRT - L$. In such case, under heavy

load, stations reusing their own segments will always find condition (3) to hold. If the network is not perfectly tuned, the maximum throughput of the protocol is bounded from below by the following expression:

$$T_{SWT} \geq max(\frac{L}{TRT}, 1 - \frac{L}{TRT}) \tag{4}$$

This lower bound can be arrived at by observing that the worst that can happen is when stations only transmit while possessing the token. Thus, the throughput cannot be worse than $(TRT - L)/TRT$. On the other hand, token-less transmissions are allowed with a gap of $(N-1) \times THT < TRT - L$ following the token. This gives a throughput of L/TRT.

Although the lower bound on T_{SWT} depends on L, it is never less than $1/2$ and even increases with increasing L. Note that the lower bound is reached only when $L = N \times THT - \varepsilon$, for a very small, but positive, ε. Thus, when high throughput is critical, decreasing THT or inserting a small segment of additional fibre (whichever is more practical) will bring the maximum throughput arbitrarily close to 1.

To estimate the average access delay under light load, assume that the network is idle and a randomly selected station gets a segment to transmit. With probability $(TRT - L)/TRT$ the station is not allowed to transmit the segment immediately. Then the average waiting time is equal to $(TRT - L)/2$. Thus, the expected waiting time is equal to:

$$A_{SWT} = \frac{(TRT - L)^2}{2TRT} \tag{5}$$

which tends to 0 as L becomes bigger.

The maximum throughput of SWT can be improved by weakening the transmission condition (3). In its original version, this condition guarantees that the segment will make a full circle through the ring and reach its sender before reaching the token (therefore making *token cleaning* redundant). If the sender knows the number of stations separating it from the recipient, it can start the transmission as soon as the segment is guaranteed to reach the recipient. Then condition 3 can be transformed into:

$$TT_i >= (D-1) \times THT \tag{6}$$

where D is the number of forward "hops" separating the recipient station from the sender.[6] With this modification, the maximum throughput of SWT may slightly exceed 1, as some segments are

[6]Immediate neighbors are separated by one hop.

absorbed by the token-holding station before they circle the ring and reach their senders. Similarly, the average access delay under light load will decrease by a factor of 4 (assuming a uniform distribution of recipients).

Whether condition (3) or (6) is used for transmission, movement of the token is very regular and each station gets a guaranteed share of the network bandwidth at very regular intervals of TRT slots. Similarly, stations can reuse their segments at regular intervals of L slots (if $L \leq TRT - L$). Thus the network can handle synchronous and isochronous traffic with very low jitter (postulates 7 and 8). Moreover, the network is well synchronized. In fact, since every station knows when it is going to receive the token, stations could emulate token transition by counting slots. When a token arrives at a station at an unexpected moment, a failure must have occurred (this *early token* signals that the upstream neighbor of the station died). Likewise, when a station expecting a token does not receive it at the expected time, the token must have been lost. The station experiencing this *late token* signals the failure and immediately claims the token. This way the network easily recovers from failures (postulate 9).

One can follow up on the idea of weakening the transmission condition (6) in a direction that gets us closer to our target solution. Assume that each station is able to rearrange the queue of segments awaiting transmission in such a way that segments addressed to closer forward neighbors (i.e., separated from the sender by fewer hops) are processed first. This way condition (6) is fulfilled more often and the silent gap following the token is reduced. Apparently, this approach has the flavor of unfairness or even starvation, if no precaution is made to process segments addressed to distant destinations, even if there is a continuous supply of "local" segments. Some researchers postulate that in a MAN environment, preference should be given to local traffic (e.g., see [11]). However, this idea does not quite work in single-ring network, since for any two neighbors $\mathcal{A}$ and $\mathcal{B}$ either the distance $\mathcal{A}$ to $\mathcal{B}$ or *vice versa* is not less than $L/2$. To make sure that the notion of geographic proximity is well represented by the propagation distance along the ring, we should switch to a double-ring topology and assume that the rings are counter-rotating. As in METARING, the ring offering the shorter path to the destination will be used to transmit the segment.

Note that the procedure of selecting the "right" segment for transmission need not pervasively discriminate against distant destinations. At the moment when the station is about to fill a free slot with a segment, it may choose the segment whose destination "best fits" the current distance from the token: the most distant destination for which condition (6) still holds.

2.2.2 Multiple weak tokens

Assume that there are multiple tokens circulating in the ring. If each token is always held for the same amount of time at every station, the distance between the tokens (measured in slots) is fixed. If this distance is small, it may happen that a station holding a token receives another one. The second token is then held by the station independently of the first one—by the prescribed amount of time, so that the two tokens depart from the station separated by the same interval of slots as upon their arrival. Note that it is not absolutely necessary that all stations hold all tokens by the same amount of time. Different stations may use different token holding intervals as long as the same interval is applied to all tokens held by a given station. For simplicity, we will assume that all stations observe the same holding time; it should be clear how this assumption can be relaxed. The transmission rule is as before (using condition (6)) and the only cleaning rule is the token rule. Transmitters do not empty the slots they filled. In fact, we would like to make sure that such a slot will never arrive back at its transmitter without being first emptied by a token holding station.

Assume that a station has a segment to transmit to a destination located D hops down the ring. The station waits until condition (6) is fulfilled. This requires a gap of at least $(D-1) \times THT$ slots between the last token departure from the station and the departure of the next token. Note that possession of a token is neither a sufficient nor a necessary condition for transmission. Consider a dual counter-rotating ring configuration with N stations. The maximum number of hops to be traveled by a segment is $D_{max} = \lfloor N/2 \rfloor$. The two rings are independent and following the simple ring selection operation, the fate of a segment is confined to the selected ring.[7] In fact, the sole purpose of the other ring (besides providing additional bandwidth) is to reduce D_{max}. The minimum condition to make such a network operable is the existence of at least one pair of adjacent tokens separated by at least $(D_{max}-1) \times THT$ slots. Otherwise, it would never be possible to send a segment to the most distant destination.

Although a station holding a token may not be able to transmit a packet to a distant receiver, it still enjoys the privilege of acquiring empty slots for transmission. This, however, brings the issue of **fairness**, which we will define here in the following way: *a medium access strategy is fair, if it does not discriminate against any destinations.*[8]

[7] In contrast to METARING, where backward feedback information for one ring travels along the opposite ring.

[8] This is not the only way to define fairness; see [4]. Note that in our case "source" fairness is guaranteed by token circulation.

The idea behind the multiple tokens is to provide a mechanism for spatial reuse: segments should be removed before they reach their senders, preferably, immediately after they have reached their destinations. This brings us to the following postulates:

- The maximum space between two consecutive tokens should not be greater than $(D_{max} - 1) \times THT$. Larger token spacing brings no new transmission opportunities, but reduces the number of tokens which has a detrimental impact on the maximum throughput achievable by the network (see below).

- To maximize throughput, a station with several segments awaiting transmission should select a segment addressed to the farthest destination reachable at the current moment. This way, the amount of bandwidth wasted by the segment after is has passed its recipient will be minimized. Ideally, if each station has a variety of segments to choose from, the protocol may emulate destination cleaning without buffering the segments at the destinations.

- Under heavy traffic (all stations constantly backlogged), all transmissions are done by token holding stations. To avoid starvation, we postulate the following informal condition (which will subsequently be formalized):

  > For each D, $1 \leq D \leq D_{max}$, and some positive constant $\mathcal{K}$, there should exist exactly $\mathcal{K}$ tokens T_D such that the maximum transmission distance for a station holding a token T_D is exactly D.

  If this condition is not fulfilled, then, when the traffic is heavy, packets to some destinations may be starved. The requirement that there are exactly $\mathcal{K}$ tokens of each given reach is meant to ensure fairness; it is unlikely that it can be met completely (since L and N are given in advance), so we settle for a slightly weaker requirement: the number of situations in which a station can reach a destination located D hops down the ring should be **approximately** the same for all D, $1 \leq D \leq D_{max}$.

The maximum throughput of MWT is very simple to calculate. Under heavy load, all transmissions are performed by token holding stations. Assuming that every station has a constant backlog of segments to transmit, every station transmits through its entire THT interval, and only then. Thus, the maximum throughput of MWT is given by the following formula:

$$T_{MWT} = M \times \frac{TRT - L}{TRT} \tag{7}$$

where M is the combined number of tokens in both counter-rotating rings. Although the second factor decreases with increasing L (this part describes the throughput of a strong token protocol), we will shortly see that the number of tokens M (constrained by our postulates) is proportional to L. Thus, MWT is a *capacity–1* protocol.

2.2.3 Token allocation

We assume that we are given three parameters: the ring length L in slots, the number of stations N, and the value of THT (token holding time) per station and per token. The last parameter can be flexible, however, in most cases its value is constrained by the application profile. It is always reasonable to keep the token holding time small. Based on these parameters we would like to determine the sequence $V_0, \ldots, V_{M-1}$ of integer numbers describing the configuration of tokens to be inserted into the ring upon initialization. V_i is the interval (in slots) between a pair of adjacent tokens. We have:

$$\sum_{i=0}^{M-1} V_i = TRT = L + N \times THT \tag{8}$$

The problem of determining the best configuration of tokens for a particular network can be solved in two steps. First, we determine the collection of values $V_0, \ldots, V_{M-1}$ treated as a set, i.e., without assuming any specific ordering of these values. Note that from the point of view of fairness under heavy load, the permutation of tokens is irrelevant. The only important point is that all distances to destinations are included in the collection of token intervals in approximately the same multitude. By permuting the token intervals we influence the medium access time for light load; this is a separate problem.

Assume that the load is heavy and a station S just receives a token T_i. Let the interval between T_{i-1} (the previous token seen by S) and T_i be V_i. Since traffic is heavy, S can only transmit while it holds the token—for THT slots. To be able to transmit to a station located at distance D, the number of slots elapsed since the departure of T_{i-1} must be at least $(D-1) \times THT$ (see condition (6)). Thus, in the first slot of its token holding interval S can transmit at distance $D_i(0) = \lfloor \frac{V_i - THT}{THT} \rfloor + 1$. In general, during the j-th slot of the token holding interval, the maximum distance at which S can transmit is given by the following formula:

$$D_i(j) = \lfloor \frac{V_i - THT + j}{THT} \rfloor + 1$$

While holding token T_i, station S may transmit THT slots. These slots can be grouped in sets based

on the distance they will travel before reaching token T_{i-1}. Let $K_D(V_i)$ be defined as the set of all slots that will travel a distance D before reaching token T_{i-1}, for $1 \leq D \leq D_{max}$. Thus,

$$K_D(V_i) = \{j \in [0, THT - 1] : D_i(j) = D\}$$

Note that $K_D(V_i)$ is empty if D never happens to be the maximum transmission distance while the station is holding token T_i. If $THT = 1$, there is exactly one nonempty set K_D for a given V_i.

Considering all the tokens,

$$r_D = \sum_{i=0}^{M-1} \overline{K_D(V_i)}$$

gives the total number of slots during which a token holding station can transmit at the maximum distance D. We define the unfairness of a given token allocation scheme as:

$$U = \max |r_A - r_B|, \quad 1 \leq A, B \leq D_{max} \tag{9}$$

The problem of ensuring fairness is equivalent to minimizing U. Although this integer optimization problem is hard to solve by a closed formula (due to its discrete nature), it can be successfully attacked by approximate methods. We have devised a genetic algorithm for finding solutions with small U, which operates along the lines described below.

A starting configuration of token intervals is built according to the following scheme (described in C):

$D = D_{max}$; $i = 0$; $left = TRT$;
`while` $(left \neq 0)$ {
  $V_i = (D - 1) \times THT + \lfloor THT/2 \rfloor$;
  `if` $(V_i < THT)$ $V_i = THT$;
  `if` $(V_i > left$ `||` $left - V_i < THT)$ $V_i = left$;
  $left = left - V_i$;
  $i = i + 1$;
  $D = D - 1$;
  `if` $(D == 0)$ $D = D_{max}$;
}
$M = i$;

The initialization procedure continues until the entire TRT interval is exhausted (see equation 8). In each turn, it services one distance between 1 and D_{max}: it tries to build a token interval for which the given distance is the longest transmission distance in the middle slot of the token holding interval. All distances are served cyclically, starting from D_{max} down to 1, then again proceeding from D_{max}, and so on.

Then, the algorithm calculates the unfairness of the initial configuration and performs a number of iterations trying to improve this unfairness by adjusting boundaries between adjacent intervals. The intervals are permuted in a randomized way by grouping together the intervals whose adjustments resulted in the biggest improvement. The algorithm may decide to add a new interval (by inserting a token in the middle of an existing interval) or to combine two intervals into one. In all cases, the resulting number of tokens M ends up very close to the initial number produced by the initialization procedure listed above. To estimate this number, note that the average length of a token interval after initialization is of order $(D_{max}/2-1)\times THT + THT/2$ which translates into $(N-2)\times THT/4$, assuming the number of stations N is even. The total number of tokens can be estimated by dividing $TRT = L + N \times THT$ by the average length of a token interval which yields:

$$M \approx \frac{4L}{(N-2)\times THT} + \frac{N}{N-2} > \frac{4L}{TRT-L} + 1 \tag{10}$$

In combination with formula 7 and assuming two symmetric counter-rotating rings, this gives us the following estimate on the maximum throughput achieved by MWT:

$$T_{MWT} > \frac{8}{1+\frac{TRT-L}{L}} \tag{11}$$

Somewhat paradoxically, and contrary to most medium access protocols, the maximum throughput of MWT tends to improve with increasing L. Asymptotically, it approaches 8 which equals the throughput of pure, dual-ring, buffer-insertion METARING without SAT. MWT achieves this throughput without explicit destination cleaning and without buffering segments at intermediate stations. Moreover, MWT is fair and starvation-free without sacrificing any portion of the bandwidth to implement this feature. It is also well synchronized (tokens arrive at stations at very regular intervals) which makes it well suited for synchronous and isochronous applications.

Let us now devote some attention to the medium access delay under light traffic conditions. In MWT, a station having a segment ready to transmit may have to wait before the destination becomes reachable, even if there is no contention from other stations.

Suppose we are given an allocation of token intervals $V_0, \ldots, V_{M-1}$. This time the ordering of these intervals is important so we assume that the tokens circulate in the ring in the listed order. Consider a random station S and the slots that visit S between two consecutive arrivals of a token (say, T_0) to S. There are TRT such slots; let them be numbered from 0 to $TRT - 1$, according to the order in which they visit S. Consider a given distance D, $1 \leq D \leq D_{max}$. Let $t^D_0, \ldots, t^D_{N_D-1}$ be the ordered sequence of slot numbers, induced by $V_0, \ldots, V_{M-1}$, identifying the slots within which a transmission at distance D is possible. These are all the slots satisfying condition (6). Assume that the network is idle and S gets a segment addressed D stations down the ring. The probability that the packet arrives at S while the station is between slots t^D_i and t^D_{i+1} is given by the following formula:

$$P^D_i = \frac{t^D_{i+1} - t^D_i}{TRT}$$

To avoid discussing the special case when S is between $t^D_{N_D-1}$ and t^D_0, we put $t^D_{N_D} = t^D_0 + TRT$. Then, the expected waiting time is $(t^D_{i+1} - t^D_i)/2$. Consequently, the expected access delay for transmission at distance D is:

$$A_D = \sum_{i=0}^{N_D-1} \frac{(t^D_{i+1} - t^D_i)^2}{2\,TRT}$$

If all destinations are equally likely, the expected global access delay under light load is obtained by averaging A_D over all D, i.e.:

$$A_{MWT} = \frac{\sum_{D=1}^{D_{max}} A_D}{D_{max}} \tag{12}$$

For a biased distribution of destinations, the contribution of particular A_D's in formula 12 should be weighted properly.

The problem of minimizing the access delay under light load boils down to finding the permutation of the token intervals $V_0, \ldots, V_{M-1}$ that minimizes A_{MWT}. Again, we are generally able to find good solutions to this problem with a genetic algorithm that locates at random permutations with a "reasonable" A_{MWT} and then tries to improve them in a methodological way.

While the above discussion was based on using two physical counter-rotating rings, it may be extended to any number of rings. For example, with $2\mathcal{R}$ rings ($\mathcal{R}$ clockwise and $\mathcal{R}$ counterclockwise), the same multiple-token protocol works correctly and yields a maximum throughput exceeding $\frac{8\mathcal{R}}{1+\frac{TRT-L}{L}}$. If the tokens are staggered appropriately,[9] adding more rings will reduce the average access delay under light load. Although building a network of a large number of physical rings may be an

[9]Similar to disk striping.

uninteresting proposition, the above extension also applies to *logical rings*, e.g. WDM-based networks [10].

2.3 Strengths and Weaknesses of MWTP

Since the Multiple Weak Token protocol may be seen as an alternative to METARING, it is natural to compare the two. In comparison to METARING, the use of multiple tokens has the following properties:

Advantages

- No need for an insertion buffer.
- A much simpler mechanism for spacial reuse, one that does not require fast reaction of stations.
- Starvation free.
- A simple mechanism guaranteeing each station its share of the total bandwidth, when needed and only then.
- Lesser vulnerability to station failures. Packets cannot be "orphaned."
- Predictability of network events, which allows automatic healing of failures on the fly.

Disadvantages

- Unsuitable for small networks, especially small networks with a large number of stations.
- Slightly smaller maximum throughput for uniform traffic.
- A greater average access delay, especially for light loads.
- Problems with a heavily biased traffic pattern in which the receiver is located exactly $N/2$ stations away from the sender. The protocol is not able to assign more than $2/N$ of the total bandwidth to such a traffic pattern, unless the presence of such pattern is known in advance (then, a different token allocation scheme will give such pattern no less bandwidth than METARING).

Altogether, MWT satisfies performance postulates 1, 2, 3, 6, 7, 8, and 9,is close to satisfying 4, and fails to satisfy 5 (since the bandwidth available to the sender of a burst depends on the location

of its receiver). It is an improvement over METARING in the way it satisfies postulates 1, 2 (or 3), 6, 8, and 9 (and probably 7, although this is debatable). Additionally, it is cheaper, since there is no need for an insertion buffer nor for a fast circuitry responsible for spatial reuse. This improvement is counterbalanced by a loss of performance (in comparison to METARING) with respect to postulates 4 and 5, and marginally 3 (if SAT is not used).

Whether the net balance is positive depends on the applications using the network.

3 Simulated environment

We used simulation to compare the performance of MWT, FDDI, and METARING. The number of messages transmitted during a single experiment varied as a function of the network size, the offered load, and the token rotation time; its range was between 200000 and 2000000 messages.

Two lengths of a ring were considered: 10^5 bits and 10^6 bits. Assuming a 200 km ring, the two propagation lengths represent transmission rates of $100Mb/s$ and $1Gb/s$. Note that FDDI was designed for smaller lengths and is not expected to work efficiently beyond the 10^5 bit range.[10]

The number of stations was the same for all networks and equal 33. Since both METARING and MWT assign packets to rings based on the distance between the sender and the receiver, it is convenient to use an odd number of stations (stations are equidistant in the experiments).

The traffic was uniform in the sense that the probability of every pair (*sender*, *receiver*) was the same. For simplicity, networks were assumed to be **slotted** and the slot format of DQDB was assumed: each slot consisted of a 384-bit payload and 40-bit header. This assumption, while inconsistent with the formal definitions of FDDI and METARING, does not alter the validity of the results: firstly, all the protocols were affected in the same way; secondly, none of the protocols takes advantage of a fixed packet size. Additionally, FDDI was assumed to operate on both counterrotating rings, instead of using one of them as a standby ring. For all the protocols, messages were assigned to transmitter queues at the moment of arrival and there were not moved from queue to queue.

In MWT, THT was the same for each station and equal to 1 slot. In FDDI, the target token rotation times $TTRT$ were taken to be:

- 1.5×10^6 bits for the 10^5 bit rings, which is equivalent to 15 ms.

[10] The comparison to FDDI is not intended to belittle this protocol, but rather to show that the adoption of the "weak token" concept allows projecting its use into the gigabit range.

- 2.4×10^6 bits and 15×10^6 bits for the 10^6 bit rings. These values are equivalent to 24 ms and 150 ms, respectively.

The message length was fixed and equal to the segment payload size.

In METARING, the SAT mechanism was not used (ie, k was assumed to be ∞).

4 Performance measures

Since we assumed that messages are of a length equal to a DQDB segment, there is no difference between packet access delay and message access delay. Thus, we measured the mean message access delay (A in the figures), defined as the mean time elapsing from the moment when message is enqueued for transmission and the moment when its first bit is successfully transmitted. The delay is measured in bits. Note that a meaningful comparison of the access delay of FDDI and MWT with that of METARING should take into account that in METARING, packets are buffered at each station (at least their headers), which, on average, delays the message delivery by at least 340 bits in a 33 stations network (this delay is not included in the performance measurements).

The message access delay is a function of the effective throughput, as well as of the properties of the protocol (and the network's size).

The maximum effective throughput of FDDI operating on a single ring is given in 1; this value should be doubled for a two-ring version. Likewise, the maximum throughput of METARING is given in 2. Since we assumed $k = \infty$, this maximum equals 8.

The maximum throughput of MWT depends on the number of tokens used and the values of L and $TTRT$, as discussed in a previous section. In theory, a maximum of 8 could be reached (equation 9).

5 Comparison

The **throughput** *vs.* **mean access delay** graphs serve as witnesses to both the ability to handle congestion and the quality of service offered under normal (i.e., light) traffic conditions. An ideal network protocol should induce a mean access delay close to half a slot length[11] when traffic is very light; conversely, when traffic is extremely heavy, it should reach a throughput as high as possible.

[11] Since the networks are assumed slotted; otherwise, it should approach 0.

Figures 1 and 2 show the message access delay versus throughput. METARING performs best, which is not surprising, since it uses more powerful hardware: an insertion buffer (for destination reuse of slots). In these figures, "`mwt,r`" represents a random permutation of the best token allocation obtained using the algorithm of section 2.2.3, while "`mwt,m`'" represents the permutation that is supposed to minimize delay under light load.

Figure 3 shows how the delay incurred by MWT depends on the number of stations. Note the strange behavior of the 129-stations network near saturation point; the number of stations is too large for a good token allocation in a ring of this size and packets to be sent to distant receivers are starved.

The performance of METARING, while excellent, is marred by the possibility of starvation for heavy load. This starvation potential is difficult to measure experimentally; we took a random subinterval of simulation and measured the message access delays for all the stations **during that interval**. Then, we took the ratio U of the highest delay observed to the lowest delay observed and used this ratio as a measure of temporary protocol unfairness. The measurements are reported in figure 4 for the 10^5 network, with an observation interval of 40,000 slots of time (each station generated some 9,000 messages during this time interval).

Since in MWT, one can allocate tokens in different ways, figure 5 demonstrates the performance of a sample of token allocations. Note that in many applications, throughput is not the only important measure to be optimized. Many applications place a greater emphasis on the regularity of transmission opportunities (for isochronous traffic) or on a greater predictability (for failure recovery). In the case of MWT, the tradeoff between throughput and other considerations is particularly eminent. Figure 5 compares four different token allocation schemes in a 10^5 bit network with 33 stations: the scheme that maximizes fairness under heavy traffic conditions (32 tokens—see figure 1), 17 equally spaced tokens with the maximum reach of 16 stations each, 22 tokens with alternating reach of 8 and 16 stations, and the single-token scheme (SWT). Recall that SWT is absolutely fair for all traffic conditions. One can clearly see a tradeoff between the maximum throughput achievable by the network and the access delay for light load. It is conceivable to make the number of tokens vary in response to changing load patterns. More research is needed to devise procedures suitable for this approach.

6 Summary

We presented a MAC–level protocol for a ring network. The protocol is based on passing multiple tokens. Its performance is similar to that of METARING, but it requires less hardware support, supports isochronous applications better and is self-healing after either a loss of station or a loss of token.

The protocol is of particular use in high-speed networks, since its performance actually improves with the increasing transmission rate or the size of the network.

References

[1] J. Chen, I. Cidon, and Y. Ofek. A local fairness algorithm for gigabit lans/mans with spatial reuse. *IBM Technical Report*, RC 18114, 1992.

[2] I. Cidon and Y. Ofek. A full-duplex ring with fairness and spatial reuse. In *IEEE INFOCOM'90*, pages 969–981, 1990.

[3] R. Cohen and A. Segall. Multiple logical token rings in a single high-speed ring. *Technion Technical Report #738*, 1992.

[4] W. Dobosiewicz, P. Gburzyński, and V. Maciejewski. A classification of fairness measures for local and metropolitan area networks. *Computer Communications*, 15:295–304, 1992.

[5] Distributed Queue Dual Bus Subnetwork of a Metropolitan Area Network. IEEE Std 802.6–1990, July 1991.

[6] Fiber Distributed Data Interface (FDDI) – Token Ring Media Access Control (MAC). American National Standard for Information Systems, Doc. No. X3, 139–1987, Nov. 1987.

[7] Fiber distributed data interface, System Level Description. Digital Equipment Corporation, June 1990.

[8] A. Kamal. On the use of multiple tokens on ring networks. In *Proceedings of IEEE INFOCOM'90*, pages 15–22, San Francisco, CA, June 1990.

[9] L. Kleinrock. *Queuing Systems*. John Wiley & Sons, Inc., 1975.

[10] B. Mukherjee. Wdm-based local lightwave networks part I: Single-hop systems. *IEEE Network*, 6(3):12–27, May 1992.

[11] A. Pach, S. Palazzo, and D. Panno. Improving DQDB throughput by a slot preuse technique. In *ICC'92*, Chicago, USA, June 1992.

[12] H. Wu, Y. Ofek, and K. Sohraby. Integration of synchronous and asynchronous traffic on the metaring architecture and its analysis. *IBM Technical Report*, RC 17718, 1992.

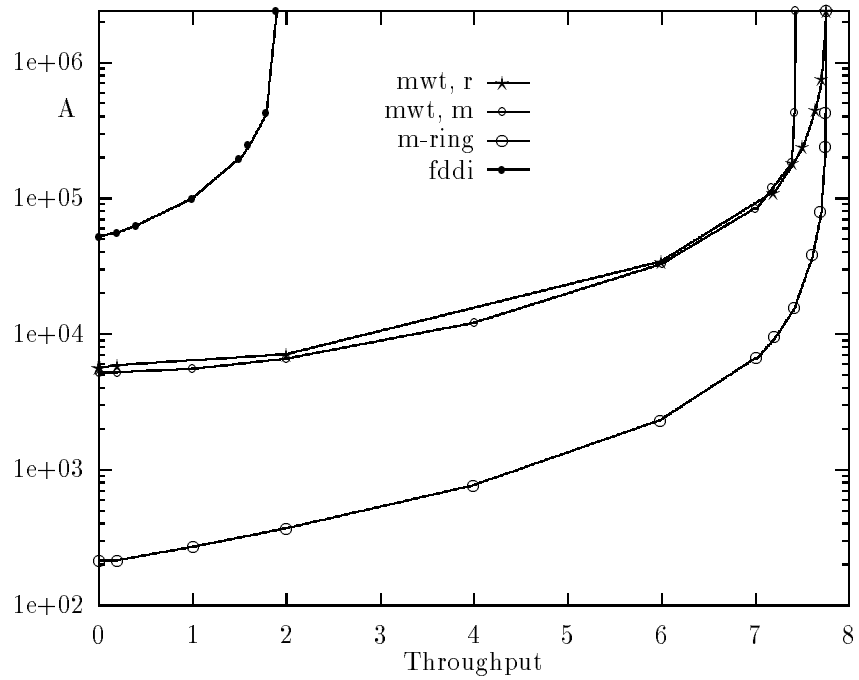

Figure 1: MWT versus METARING and FDDI, 100kb ring, 33 stations.

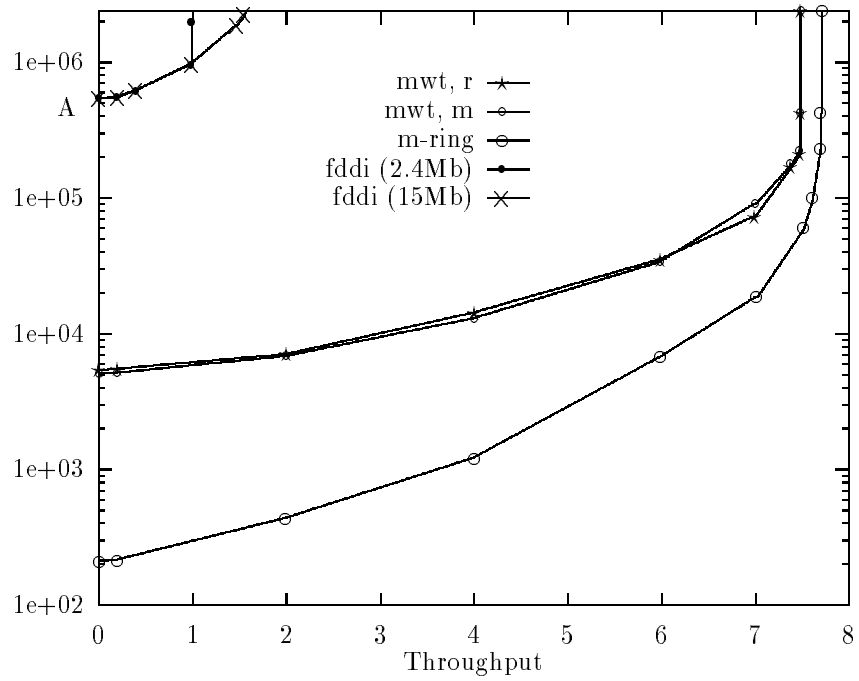

Figure 2: MWT versus METARING and FDDI, 1mb ring, 33 stations.

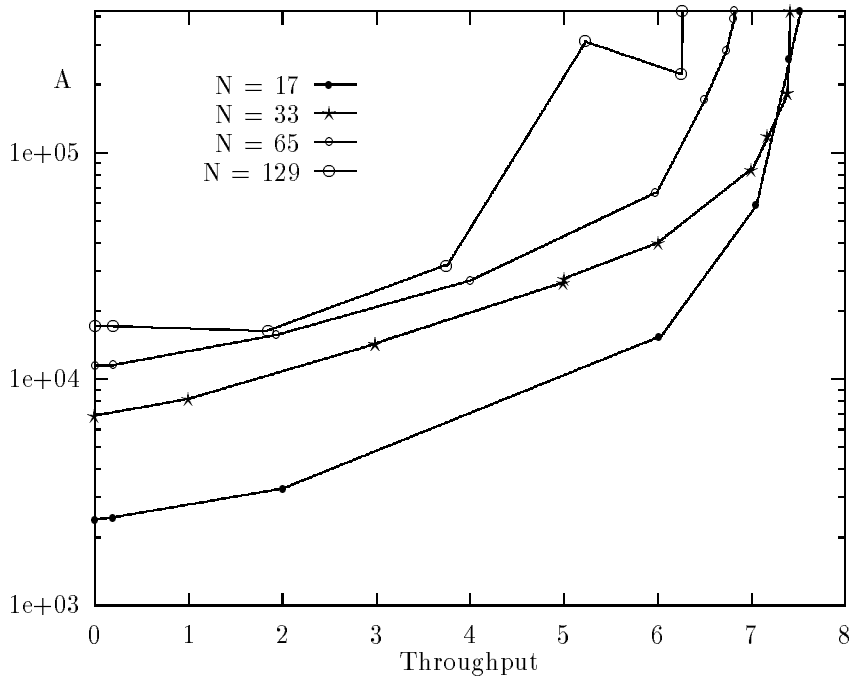

Figure 3: MWT (100kb) for different numbers of stations.

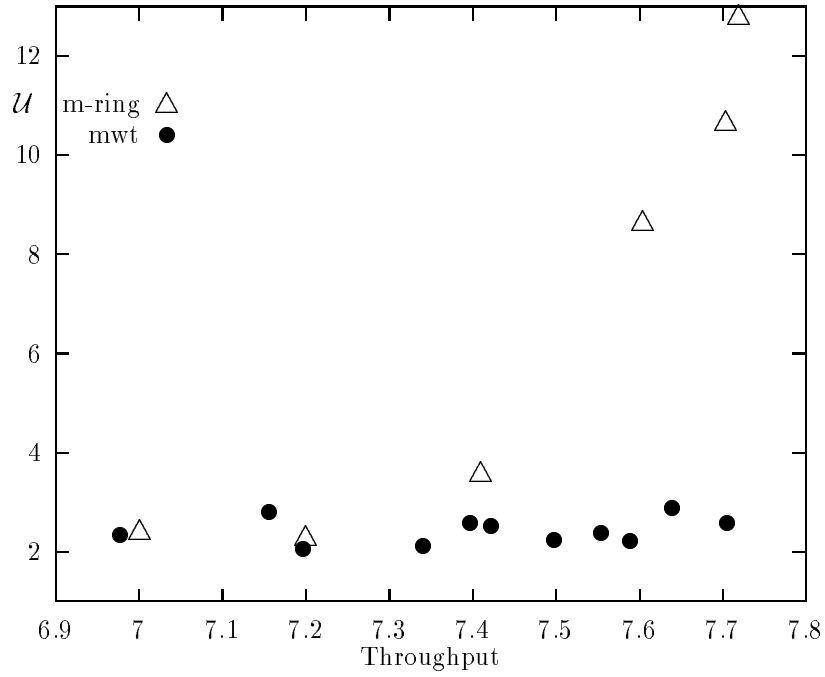

Figure 4: Starvation potential: MWT vs. METARING (no SAT), 100kb network.

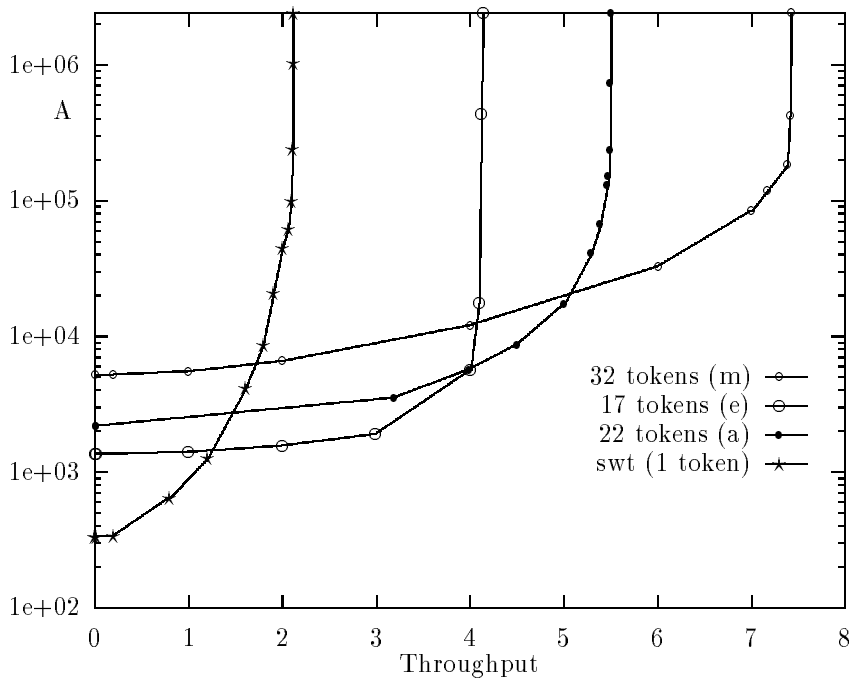

Figure 5: MWT/SWT for different token allocation strategies.