CHAPTER 5.

THE DESIGN AND OPERATION OF A SINGLE NET

Described in this chapter the design and operational principles of a single net. The net is a multiple–access medium constructed of a centralized, passive, optical star and single mode fiber optic links (see Section 2.3). The communication over the net merges to one point in space, and then broadcasts back to all the net's nodes. The transmission is divided into fixed length–time slots, with one slot counter at each node, to keep track of time using a slot as its basic time step.

The net's operational principles and timing are derived from the centralized star topology, which enables simple synchronization of the net's operation, and also the conservative coding scheme, which enables the periodic exchange of state information, via very short control messages.

5.1. The Net's Operational Principles and Timing

5.1.1. Net timing

The optical star is constructed in a very small area – a few square inches. Thus, this center is used as the time reference for the net's nodes. Each time slot has a fixed time duration T_s, which is measured in bit periods, i.e., the maximum number of bits which can be transmitted within one time slot at a given baud rate. The one–way delay of a node i from the star center is Δ_i. Thus, the n nodes of each net may be regarded as lying on the circumference of an imaginary circle of radius R, such that

$T_R \geq \max\{\Delta_i$, such that $n \geq i \geq 1\}$, as shown in Figure 5.1. T_R is an upper bound of the node–to–center delay. This uniform arrangement is used for the net synchronization mechanism. Because of this circular symmetry the nodes are **indistinguishable** with respect to the center point, and the symmetry is later used for modeling a partial hypergraph as a centralized switch.

The slots are grouped into frames of duration T_f, with f slots per frame and $T_f = fT_s$, as shown in Figure 5.2. For synchronization purposes, the frame duration is equal to $2T_R$, i.e., T_f is greater than the delay from any node to any other. The slot duration is then $T_s = \frac{2}{f}T_R$. Although T_f depends on the physical size of the network, it should be noted that T_s can be chosen according to other system considera-

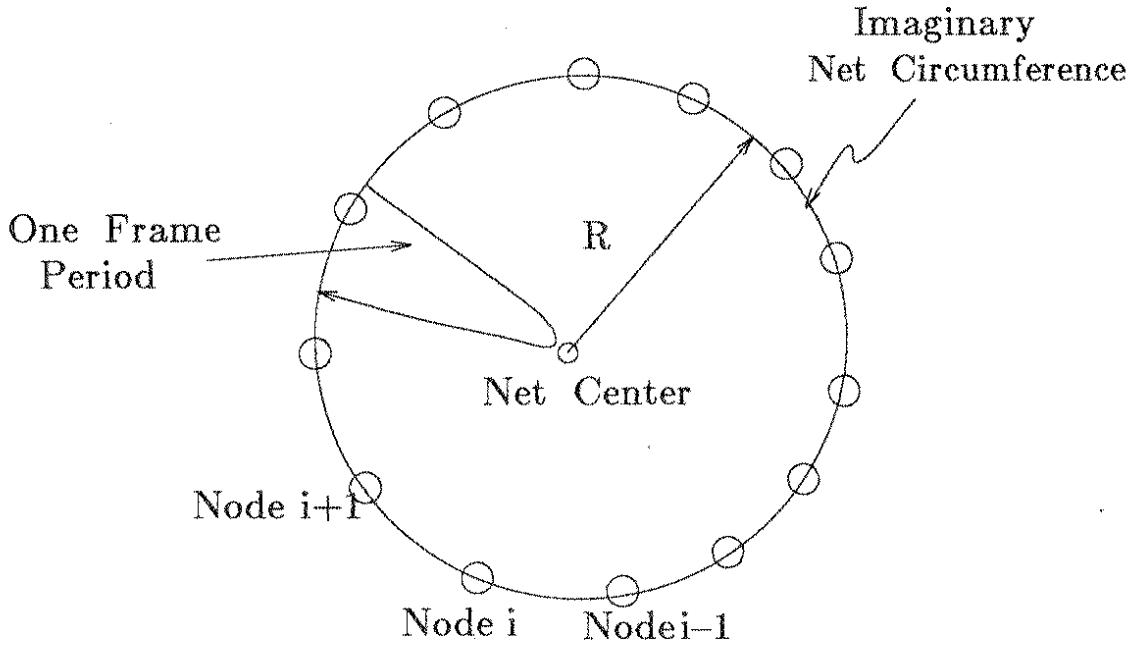


Figure 5.1: The Net Uniform Arrangement

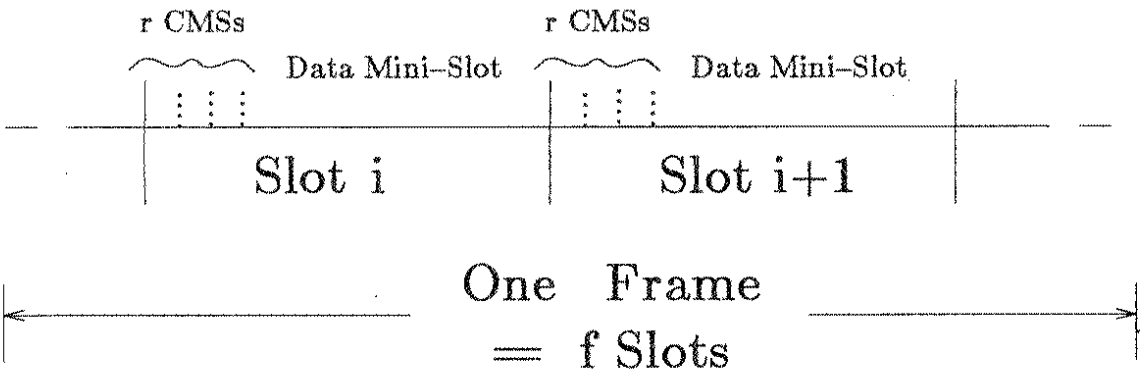
tions. Once it is chosen, however, the net synchronization condition depends on f.

$T_s = \frac{2}{f} T_R$.

5.1.2. Periodic exchange of state information

The basic operational principle of the net is the periodic exchange of state information. The major reason for having it is for making the integration of various distributed functions more uniform and efficient. In general, the control mechanism and the actual data communication and computation are independent. Therefore, it is reasonable to assume two basic requirements for the control mechanism: (i) periodic exchange, via (ii) short control messages.

The access control, synchronization mechanism, and the integration of other functions are based on the periodic exchange of timing and state information. To this end, each slot is divided into $r+1$ minislots, as shown in Figure 5.2. The first r are



CMS – Control Mini–Slot

Figure 5.2: The Frame and Slot Description

very short control minislots (CMSs), used for control. The last minislot, which occupies most of the slot space, is the data minislot (DMS), for transmitting one packet of data. The $r+1$ minislots of each slot are used by $r+1$ different nodes. The asymmetric partitioning of the slot between CMSs and one DMS makes it possible for one node to send "useful" data, and for r nodes to send very short messages, consisting of state information for the control of the communication and computation.

A message, sent during a CMS, is broadcast over the net from a **known origin** but without a **specific destination**. The access to the CMSs on each net is deterministic. For example, under a **uniform scheme** the set of n nodes is partitioned into r **disjoint subsets** of sizes either $\left\lceil \frac{n}{r} \right\rceil$ or $\left\lfloor \frac{n}{r} \right\rfloor$ nodes. The sum of the sizes of all these subsets is n, and each node belongs to exactly one of the subsets. The nodes in a subset use the corresponding CMS in a round–robin fashion.

For the following analysis it is important to note that each node can use its CMS: every $l = \left\lceil \frac{n}{r} \right\rceil$ time slots. As a result, each node can broadcast its *view* over the net every l time slots.

5.2. Hybrid Multiple Access Control Algorithm

The access control algorithm uses the load information which is exchanged periodically during the control minislots. The access control algorithm operates using two basic modes: **random** for light net load, and **deterministic** for high net load. The net dynamically switches its operational mode. Another hybrid scheme that operates in deterministic and random modes is proposed and analyzed in [GoWo85]. The random

algorithm is a simple version of a nonpersistent (or p–persistent) access control scheme [see Tane81]. The deterministic algorithm is distributed, and is performed by each node's interface. The algorithm considers only the load information currently **known** by all the nodes on the net. As will be shown, the deterministic algorithm is adaptive and guarantees high utilization of the net's communication capacity.

An access control scheme is analyzed on the basis of how efficiently it utilizes its communication capacity. For slotted communication, two events should be minimized or avoided: (i) having empty slots while some ports have full buffers, and (ii) having collisions of two or more packets, which result in a wasted slot. The first event **is avoided** by the random access mechanism, and the second **is minimized** by the deterministic algorithm. Hence, in order to maximize the utilization, the deterministic algorithm will be used whenever there is a **known** load, and the random access mechanism will be used whenever there is **no known** load.

There is a reciprocal relationship between the network capacity utilization and the average network delay. For a given load pattern, *maximizing the utilization* is equivalent to *minimizing the average delay* (and vice versa).

5.2.1. The deterministic adaptive access control scheme

Let $\{node_1, node_2, ..., node_n\}$ be the set of n nodes of a single net. The set of nodes is divided into r subsets, each subset uses one of the r CMSs for updating its state information. Each node belongs to **at least one** of the subsets. In general, a node can be either **busy** – with one or more full buffers to send, or **empty** – with no packet to send. A node which is **empty** and then receives a new packet to send will

change its state to **busy** only after its new load becomes **common knowledge**, i.e., after the new state information has been broadcast over the net, and has reached the imaginary circumference. It is assumed, in general, that there is an additional one–time slot delay in order to incorporate the new state information into the next state. The delay from the time a CMS message is sent until the state transition, which uses this information, is $f+1$ time slots. The state information sent during each CMS may include

(i) the node identification (its physical and logical addresses)

(ii) the number of packets which are ready to be sent from this port

(iii) the number of empty buffers at this port

(iv) the time stamp – the current local slot counter reading

(v) the average node's load during the past s time slots

(vi) global synchronization information

5.2.2. Priority of the control minislots

The way a net is partitioned into control minislots can implement a priority scheme. The priority is viewed here as the average delay of a node for changing its own state information, which depends on how often the node gets a CMS. Two basic partitioning methods can be used:

(i) different–sized subsets nodes with high priority would be in a smaller subset, and therefore, would receive the CMS more often.

(ii) nondisjoint subsets nodes with high priority would be in more than one CMS subset.

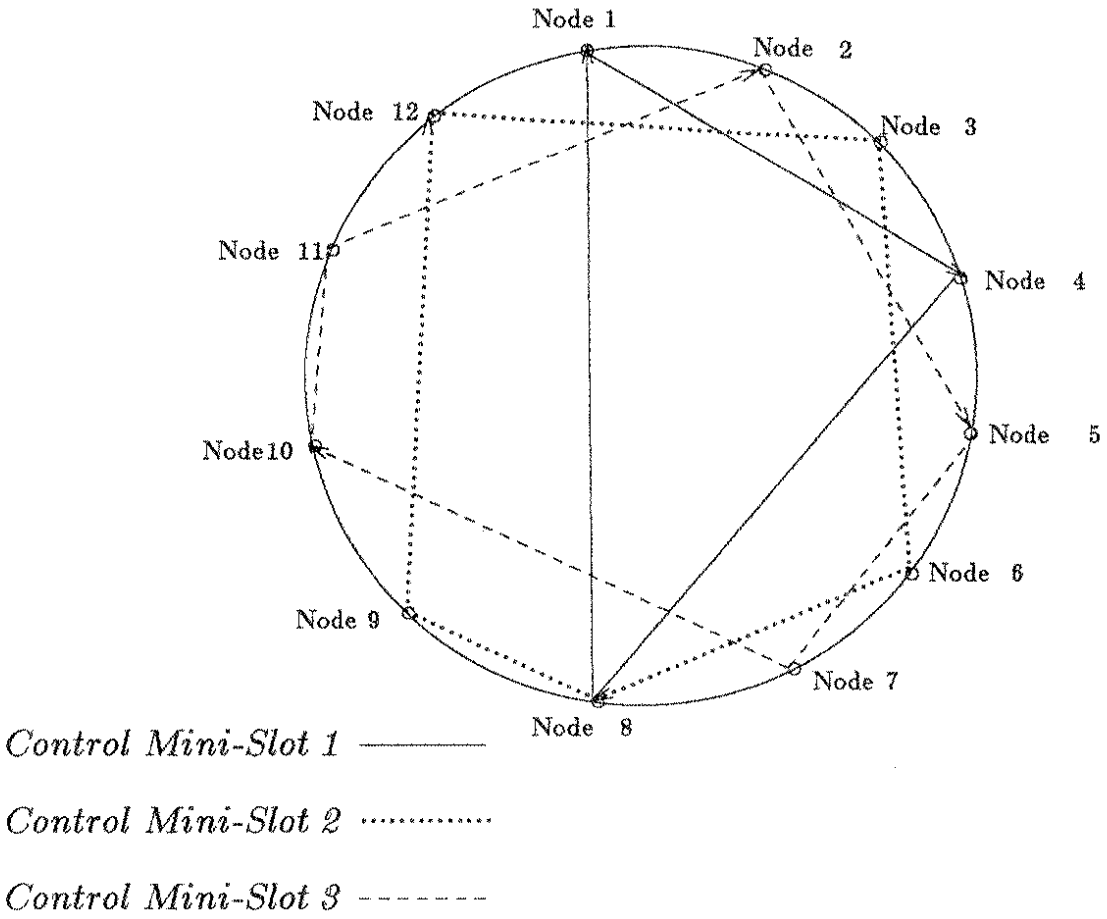


Figure 5.3: An Example for CMSs Partitioning

Figure 5.3 is a possible partitioning of 12 nodes of a net into three subsets. The three subsets are not of the same size, and a node can belong to more than one subset (e.g., node 8). In this example there three priority levels: level 1 – $\{node_8\}$, level 2 – $\{node_1, node_4\}$, and level 3 – $\{node_2, node_3, node_5, node_6, node_7, node_9, node_{10}, node_{11}, node_{12}\}$. Nodes which generate more traffic and act as "masters" are likely to have a higher priority in order to improve the system's response time, and hence its efficiency. In a real time application, a node on a critical timing path is likely to get a higher priority in order to improve its response time.

5.2.3. The scheduling during the data minislot

In the deterministic access control scheme, one of the nodes that is **known** to be in the busy state sends a packet during the data minislot (DMS). This access rotates among the other known busy nodes. Several different scheduling schemes can be implemented:

(i) round–robin scheduling

(ii) scheduling with a fixed priority among the busy nodes

(iii) scheduling as a function of the number of full packets of the busy nodes

(iv) scheduling as a function of the number of empty buffers of the busy nodes

(v) some combination of the above

The desired priority scheme should have two basic properties:

(1) Fairness – a packet from any node under worst–case scenario would be sent after finite number of time slots.

(2) Decreased overflow probability – the number of buffers at each node is finite, and the transfer rate from the interface to the host storage is slower than the net bandwidth. Note that these two requirements are in conflict; if the fairness is increased then the overflow probability is also increased.

For example, assume that the priority scheme gives the highest priority to the nodes with more than ten full buffers. As a result, the probability of overflow is likely to be low, but there is a possibility for starvation of nodes with only a few packets. On the other hand, the round–robin scheduling is very fair, but the overflow probability can be relatively high. As will be shown in Chapter 7, overflow can be prevented

by using a special flag bit in the CMS control message.

In general, in the following analysis and examples a uniform round–robin scheduling is assumed during CMSs. Thus, each node updates its state information every $l = \left\lceil \frac{n}{r} \right\rceil$ time slots.

5.2.4. The random access scheme

The load is considered to be **light** whenever there are no **known busy** nodes, and then the net's nodes switch their operational mode, during the data minislot (DMS), to the random accessing mode. Note that the accessing mechanism during the CMSs remains unchanged.

In the random access scheme, all nodes which have a full buffer (not yet **commonly known**) will transmit it in the next DMS with the probability p. If more than one node transmits at the same time, a collision occurs. To resolve this conflict, the nodes involved will retry to transmit with the probability p' ($p' \leq 0.5$) after $f+1$ slots (after the collision becomes commonly known). The access control will continue in the random mode as long as there are no **known** nodes in the **busy** state.

5.3. The Analysis of a Single Net

The following terms will be used in the analysis:

(i) μ_{net} – the efficiency of the use of the communication capacity for data transfers, or the ratio of the duration of data minislots to the total duration of the slot. Note that under this definition it is assumed that the messages during the CMSs are not "useful"

data transfers.

(ii) d_{state} – the delay of a node changing its state from **empty** to **busy**, using its CMS message. This time is measured from the time an **empty** receives a new packet until the node becomes **busy**.

(iii) d_{access} – the delay in accessing the net (the time interval from the moment the node changes its state from **empty** to **busy** until the time it gets the first access right), and

(iv) $d_{src-dst}$ – the total time delay of a packet from source to destination.

In most cases, only the mean time analysis will be performed. This is sufficient, since the operation of the net is based mainly on the deterministic access control algorithms and round–robin scheduling. The objective of this analysis is to get an indication of the expected performance of such a system.

In the discussion, the following **basic model** is assumed:

(i) the basic unit of time is a slot

(ii) n identical nodes on each net

(iii) f slots in every frame

(iv) r control minislots (CMSs) in every slot

(v) the set of n nodes are partitioned into r **disjoint subsets** of sizes either $\left\lceil \frac{n}{r} \right\rceil$ or $\left\lfloor \frac{n}{r} \right\rfloor$ nodes, such that the sum of the subset sizes is n

(vi) one data minislot (DMS) in every slot

(vii) only the CMSs are used for exchanging state information; the DMS is used only for data transfers

(viii) the basic scheduling scheme during the DMS is round–robin

(ix) one packet at a time is sent by each node, i.e., the node becomes **busy**, sends one packet, then becomes **empty**, and so on. This last assumption presents the net performance as poorer than in actuality, since each node could have several packets to send each time it becomes **busy**.

Definition – Heavy Load:–

the net is in the heavy load state during the period of time (in slots) in which the **known** number of full buffers is one or more, or when at least one node is in the **busy** state. Thus, when the load is heavy, the access control is **deterministic.**

Definition – Light Load:–

the net is in the light load state during the period of time in which the **known** number of full buffers is *zero,* or when all the net's nodes are in the **empty** state. Thus, when the load is light, the **random mode** is used.

5.4. Communication Efficiency

The net synchronous organization guarantees, in principle, that in each instance a bit of information can be transmitted into the net from one of the net's interfaces. However, the communication efficiency is reduced synchronization timing error (Δ_s), discussed separately in Chapter 6, and by the presence of control minislots, which constitute a communication and computation management overhead. Let T_{CMS} be the duration of the CMS, and T_{DMS} be the duration of the DMS, then the **maximum communication efficiency** is

$$\mu_{com_{\max}} = \frac{T_{DMS}}{T_{DMS} + rT_{CMS}} 100\ (\%).$$

In this formula it is assumed that all the DMSs have one packet of data. For example, if r=4, CMS duration is a 256 bits, and the DMS duration is 4Kbytes, then the communication efficiency is 97%.

5.5. Delay Analysis

5.5.1. State change delay

The delay in changing the node's state information (e.g., changing its state from **empty** to **busy**), is the time it takes for **locally known** information to become **commonly known**. On the average, this time is

$$d_{state_{ave}} = \left\lceil \frac{1}{2}\frac{n}{r} \right\rceil + f + 1 \ slots.$$

The expression $\left\lceil \frac{1}{2}\frac{n}{r} \right\rceil$ is the average time until the node has its next turn to send a control message during its CMS. The $f + 1$ slots is the time it takes the control message to be propagated and decoded by all the destinations. The upper–bound (worst case) on this time is

$$d_{state_{max}} = \left\lceil \frac{n}{r} \right\rceil + f + 1 \ slots.$$

5.5.2. Access delay

The access delay is the time for a node, in the **busy** state, to send a packet into the net. If on the average, in the heavy load case, there are m nodes in the busy state, then under round–robin scheduling the average time for a packet to be transmitted into the net is

$$d_{access_{ave}} = \left\lceil \frac{m}{2} \right\rceil slots.$$

The upper–bound on this delay is

$$d_{access_{max}} = n-1 \ slots.$$

5.5.3. Source–destination delay

The total time to send a packet from a node which is in the empty state to its destination, is on the average

$$d_{src-dst_{ave}} = \left\lceil \frac{1}{2} \frac{n}{r} \right\rceil + \left\lceil \frac{m}{2} \right\rceil + 2(f+1) \ slots.$$

The $(f+1)$ is added twice: for the transmission delay of the control message, and for the transmission delay of the packet.

The upper–bound on the total delay is

$$d_{src-dst_{max}} = \left\lceil \frac{n}{r} \right\rceil + n - 1 + 2(f+1) \ slots.$$

5.5.4. Source–destination delay with reservation or look–ahead

In real–time applications, the time required for transferring a packet from its source to destination can be significantly reduced. If a slot counter is used, a specific time slot can be reserved, and then the delay will be only $f+1$ slots.

Without a slot counter, the node can look ahead by changing its state from **empty** to **busy**, and then the delay will be $\left\lceil \frac{m}{2} \right\rceil + (f+1)$ slots, so the time to change the state is saved.

5.5.5. The delay as a function of the bandwidth and l

The actual time delay over the net depends on three basic factors: the bandwidth, the ratio $l = \left\lceil \frac{n}{r} \right\rceil$, and the traffic model. The first two are determined by the optical architecture, and the third one is determined by the way the system is used. In this section the delay is computed as a function of the first two parameters.

Assume that

(i) the slot duration is 4Kbytes or 32Kbit periods,

(iii) the bandwidth – BW is expressed in kilohertz,

(iii) the R is 10,000 meters,

(iv) the T_R is 0.05 millisecond, and

(v) the number of slots in a frame $f = \left\lceil \frac{2(0.05)}{\frac{32{,}768}{BW}} \right\rceil = \left\lceil \frac{0.1BW}{32{,}768} \right\rceil$

Thus, the maximum delay for the propagation of state information is

$$d_{state_{max}} = (l+f+1)\frac{32{,}768}{BW} = (l+1)\frac{32{,}768}{BW} + \left\lceil \frac{0.1BW}{32{,}768} \right\rceil \frac{32{,}768}{BW} \approx$$

$$\approx (l+1)\frac{32{,}768}{BW} + 0.1 \; millisecond.$$

The diagram in Figure 5.4 describes the delay as a function of the BW. Three cases are shown l=12, l=17, and l=22.

It is clear from Figure 5.4 that the state delay propagation is not more than a millisecond in a high bandwidth. This time is much faster than the typical access time to a magnetic disk. Note that the delay which is due to the physical size of the network is relatively insignificant.

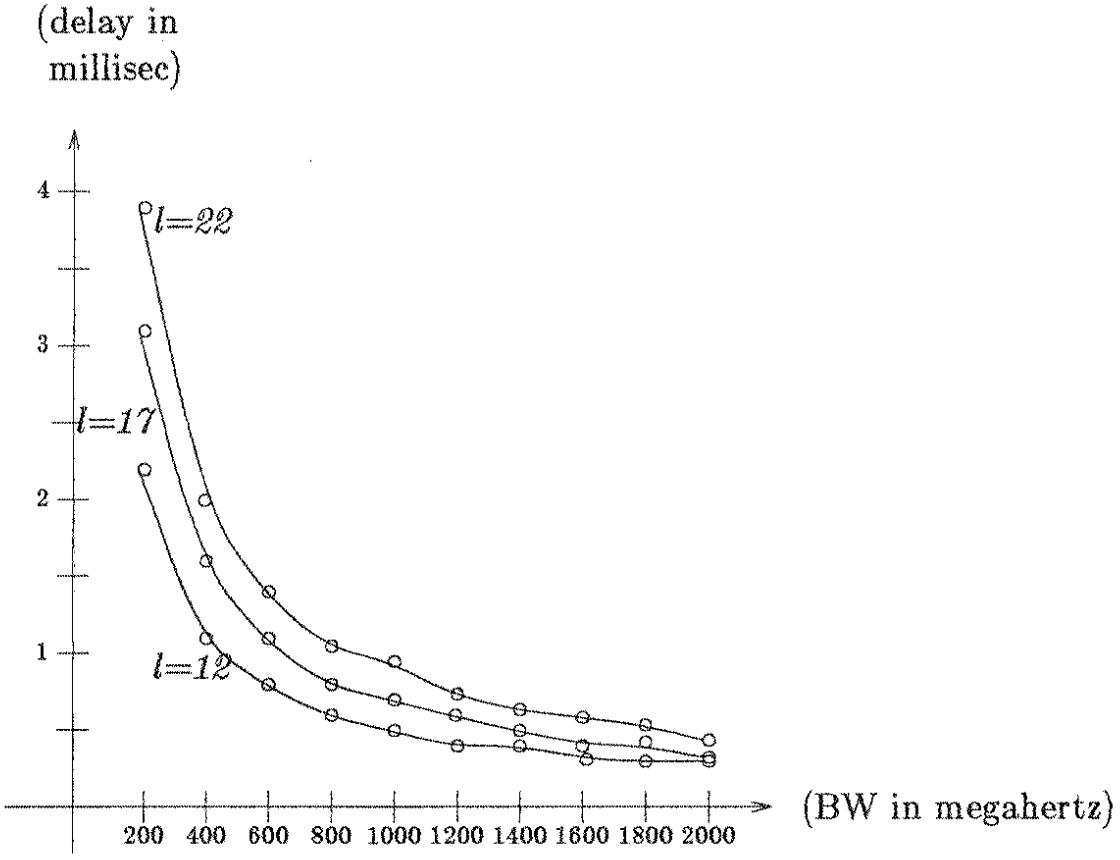


Figure 5.4: Maximum State Delay

5.6. Light Load Analysis

5.6.1. The model

Light load is analyzed on the basis of the efficiency of the communication over the net when the **known load is zero**. During this time the access control is random. It reverts to deterministic access, after new load information is received via the CMSs.

For the analysis the following model is assumed:

(i) the net has n identical nodes

(ii) the packet arrival at each of the n nodes follows Poisson statistics, with a mean

arrival rate of λ per time slot (such that $n\lambda < 1$)

(iii) the Poisson process has the following characteristics:

- the numbers of arrivals in disjoint time intervals are **independent** random variables
- the number of arrivals is proportional to the duration of the observation time interval
- the superposition of Poisson processes is Poisson, with the mean arrival rate as the sum of the individual arrival rates
- the randomized thinning of a Poisson process is a Poisson process

(iv) $f = 1$ (frame of one slot)

(v) the set of n nodes is partitioned into r **disjoint subsets** of sizes either $\left\lceil \frac{n}{r} \right\rceil$ or $\left\lfloor \frac{n}{r} \right\rfloor$ nodes, with the sum of the subsets sizes equal to n

(vi) the result of the transmission (success, collision, or new **commonly known** load) is **known** by the end of each slot (thus, a node can retry to transmit in the subsequent time slot); also the control messages are decoded by the end of each time slot, and their information is incorporated into the next state transition

(vii) a node with a full buffer will try to transmit with the probability p in the first time slot of the random access mode, and with probability p' after a collision ($p' < 0.5$). Note that after a collision only the nodes involved will try to transmit.

(viii) state information is exchanged only via the r CMSs

(ix) in the random access mode, if a node has exactly one buffer full and can use its CMS, then it will transmit the data packet at random and declare its load using its CMS.

Let $l = \left\lceil \frac{n}{r} \right\rceil$, and divide the set of n nodes into r disjoint subsets $\{A_1, A_2, ..., A_r\}$. The nodes in each subset are arranged in a temporal order in which they last use their CMS

$A_i = \{a^i{}_{1T_s}, a^i{}_{2T_s}, ... , a^i{}_{lT_s}\}$; i.e., the $a^i{}_{lT_s}$ did not use its CMS for l time slots.

The probability that a node a_j has no packets after j time slots is

$$P_{j_{no-packet}} = e^{-\lambda j T_s},$$

and the probability that a node a_j has at least one packet after j time slots is

$$P_{j_{packet}} = 1 - e^{-\lambda j T_s}.$$

5.6.2. The probability for successful transmission

The probability for successful transmission during the **first** time slot after the **known load** becomes zero (all the nodes are empty), is computed in this section.

The probability that a node a_j will not transmit (idle) is

$$P_{j_{idle}} = 1 - pP_{j_{packet}},$$

this probability includes the cases (i) the node being empty, and (ii) the node being not empty but not transmitting $(1-p)$.

The probability that none of the nodes on the net will transmit (all the net's nodes are IDLE) is

$$P_{IDLE} = \left[\prod_{j=1}^{l} P_{j_{idle}} \right]^r$$

The probability that only the node a_j will transmit (node j ACTIVE) is

$$P_{j_{ACTIVE}} = pP_{j_{packet}} \frac{P_{IDLE}}{P_{j_{idle}}}$$

Finally, the probability for successful transmission in the first time slot after the net has changed its mode of operation to random access, is

$$P_{SUCCESS} = r \sum_{j=1}^{l} P_{j_{ACTIVE}}.$$

The probability for not having a collision is

$$P_{NO-COLLISION} = P_{SUCCESS} + P_{IDLE} = r \sum_{j=1}^{l} P_{j_{ACTIVE}} + \left[\prod_{j=1}^{l} P_{j_{idle}} \right]^r.$$

Thus, the probability for a collision is

$$P_{COLLISION} = 1 - P_{NO-COLLISION}.$$

5.6.3. The probability for a wasted slot

A slot is wasted when one or more of the nodes have packets to transmit, and there was no successful transmission, either because of a collision or because the net was idle and not empty, i.e., a slot is wasted unless there was a successful transmission or the net was empty.

The probability for an empty net during the first time slot of the random access mode is

$$P_{EMPTY} = \left[\prod_{j=1}^{l} P_{j_{no-packet}} \right]^r$$

Thus, the probability that the first slot will be wasted is

$$P_{WASTE1} = 1 - P_{EMPTY} - P_{SUCCESS} = 1 - \left[\prod_{j=1}^{l} P_{j_{no-packet}} \right]^r - r \sum_{j=1}^{l} P_{j_{ACTIVE}}$$

In the design of the net the objective is to minimize the probability of a wasted slot during the random access mode, and by this to maximize the utilization (during

the deterministic mode there are no wasted slots). If λ, r, and l are some system parameters, then p should be selected by simulation such that the probability for a wasted slot is minimized.

5.6.4. The second time slot

Three cases are possible in the second slot of the random access mode.

Case 1 – the transmission in the first time slot was successful, and the load of the node decreases (since $n\lambda < 1$). Thus, the probability for a wasted second slot is less than in the first time slot ($P_{WASTE2} < P_{WASTE1}$).

Case 2 – the net was idle in the first time slot. The probability for a wasted slot remains the same ($P_{WASTE2} = P_{WASTE1}$).

Case 3 – there was a collision during the first slot. Since p'–persistent algorithm is used, only the nodes which were involved in the collision will try again in the second time slot with probability p'. If m nodes were involved in the collision, then the probability for a wasted slot is

$$P_{WASTE2} = 1 - P_{SUCCESS2} - P_{IDLE2} = 1 - m\,(p'(1-p')^{m-1}) - (1-p')^m\,.$$

5.7. Discussion

The criterion for switching the access control between deterministic mode and random modes is optimal, if the access control objective is to minimize the number of wasted slots. In fact, it is not hard to prove the theorem that operating under the random access mode **only** while there is **no known load** is optimal.

Proof: the theorem is proved by negation; there are two cases.

Case 1 – the net switches its operation to random while there are one or more **known** full buffers. Since some nodes might also have full buffers which are **not known**, the probability for collision is greater than zero. Thus, a slot might be wasted.

Case 2 – the net switches its operation to random mode one or more slots after the last **known** full buffer was sent. Clearly, in this case one or more slots will be wasted. Thus, the switching criterion is optimal.

Similar arguments are valid for switching back from the random mode to the deterministic mode. In the random mode the p–persistent protocol is not the optimal protocol. It is, however, reasonable to use, since the protocol is very simple to implement and the load is light.

The maximum efficiency of the hybrid access control algorithm is almost 100%, since under heavy load the access control is deterministic.