June 1990/Revised September 1990 LIDS–P–1986

ON THE AVERAGE COMMUNICATION COMPLEXITY OF ASYNCHRONOUS DISTRIBUTED ALGORITHMS[1]

John N. Tsitsiklis[2] and George D. Stamoulis[2]

Abstract

We study the communication complexity of asynchronous distributed algorithms, such as the distributed Bellman–Ford algorithm for the shortest path problem. Such algorithms can generate excessively many messages in the worst case. Nevertheless, we show that, under certain probabilistic assumptions, the expected number of messages generated per time unit is bounded by a polynomial function of the number of processors under a very general model of distributed computation. Furthermore, for constant-degree processor graphs, the expected number of generated messages is only $O(nT)$, where n is the number of processors and T is the running time. We also argue that our bounds are tight in certain cases. We conclude that (under our model) any asynchronous algorithm with good time complexity will also have good communication complexity, on the average.

Keywords: Distributed algorithms, asynchronous algorithms, communication complexity, shortest paths, Bellman–Ford algorithm, branching processes.

---

1. Research supported by the National Science Foundation under Grant ECS–8552419, with matching funds from Bellcore Inc. and the Draper Laboratory, and by the ARO under Grant DAAL03–86–K–0171.

2. Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139; e-mail: jnt@lids.mit.edu, stamoulis@lids.mit.edu

1. INTRODUCTION

In recent years, there has been considerable research on the subject of asynchronous distributed algorithms (see [BT] for a comprehensive set of references). Such algorithms have been explored both in the context of distributed numerical computation, as well as for the purpose of controlling the operation of a distributed computing system (e.g., finding shortest paths, keeping track of the systems's topology etc. [BG]). Some of their potential advantages are faster convergence, absence of any synchronization overhead, graceful degradation in the face of bottlenecks or long communication delays, and easy adaptation to topological changes such as link failures.

In the simplest version of an asynchronous distributed algorithm, each processor i maintains in its memory a vector y^i consisting of a variable x_i, together with an estimate x_j^i of the variable x_j maintained by every neighboring processor j. Every processor j updates once in a while its own variable x_j on the basis of the information available to it, according to some mapping f_j. In particular, x_j is replaced by $f_j(y^j)$. Furthermore, if the new value of x_j is different from the old one, processor j eventually transmits a message containing the new value to all of its neighbors. When a neighbor i receives (in general, with some delay) the new value of x_j, it can use it to update its own estimate x_j^i of x_j.

A standard example is the asynchronous Bellman–Ford algorithm for the shortest path problem. Here, there is a special processor designated by 0, and for each pair (i,j) of processors, we are given a scalar c_{ij} describing the length of a link joining i to j. One version of the algorithm is initialized with $x_i = c_{i0}$, $i \neq 0$, and is described by the update rule

$$x_i := \min\Big\{x_i, \min_{j \notin \{i,0\}} \{c_{ij} + x_j^i\}\Big\}, \qquad i \neq 0.$$

Under reasonable assumptions, the distributed asynchronous implementation of this algorithm terminates in finite time and the final value of each x_i is equal to the length of a shortest path from i to 0 [B].

In general, whenever some processor i receives a message from another processor j, there is a change in the vector y^i and, consequently, a subsequent update by processor i may lead to a new value for x_i that has to be eventually transmitted to the neighbors of processor i. Thus, if each processor has d neighbors, each message reception can trigger the transmission of d messages, and there is a clear potential for an exponential explosion of the messages being transmitted. Indeed, there are simple examples, due to E. Gafni and R. Gallager (see [BT, p. 450]), showing that the asynchronous Bellman–Ford algorithm for an n–node shortest path problem is capable of generating $\Omega(2^n)$ messages, in the worst case. These examples, however, rely on a large number of "unhappy coincidences": the communication delays of the different messages have to be chosen in a very special way. It is then reasonable to inquire whether excessive amounts of communication are to be expected under a probabilistic model in which the communication delays are modeled as random variables.

In the main model studied in this paper, we assume that the communication delays of the transmitted messages are independent and identically distributed random variables, and show that the expected number of messages transmitted during a time interval of duration T is at most of the order of $nd^{2+\frac{1}{m}}(\ln d)^{1+\frac{1}{m}}T$, where n is the number of processors, d is a bound on the number of neighbors of each processor, and m is a positive integer depending on some qualitative properties of the delay distribution; in particular, $m = 1$ for an exponential or a uniform distribution, while, for a Gamma distribution, m equals the corresponding number of degrees of freedom.[3] Note that this estimate corresponds to $O\left(d^{1+\frac{1}{m}}(\ln d)^{1+\frac{1}{m}}\right)$ messages per unit time on each link, which is quite favorable if d is of the order of 1 (i.e., when the interprocessor connections are very sparse). We also argue that this bound on the expected total number of messages generated is tight (or is close to tight) in certain cases, such as that of a complete graph. Our result is derived under practically no assumptions on the detailed operation of the asynchronous algorithm (with one exception discussed in the next paragraph). Furthermore, the result is valid for a very broad class of probability distributions for the message delays, including the Gamma distributions as special cases.

Since we are assuming that the delays of different messages are independent, messages can arrive out of order. Suppose that a message l carrying a value x_j is transmitted (by processor j) before but is received (by processor i) later than another message l' carrying a value x'_j. Suppose that l is the last message to be ever received by i. Then, processor i could be left believing that x_j is the result of the final update by processor j (instead of the correct x'_j). Under such circumstances, it is possible that the algorithm terminate at an inconsistent state, producing incorrect results. To avoid such a situation, it is essential that a receiving processor be able to recognize whether a message just received was transmitted earlier than any other already received messages and, if so, discard the newly arrived message. This can be accomplished by adding a timestamp to each message, on the basis of which old messages are discarded. There are also special classes of algorithms in which timestamps are unnecessary. For example, in the Bellman–Ford algorithm described earlier, the value of x_j is nonincreasing with time, for every j. Thus, a receiving processor i need only check that the value x_j in a newly received message is smaller than the previously stored value x_j^i, and discard the message if this is not the case.

The above described process of discarding "outdated" messages turns out to be a very effective mechanism for controlling the number of messages generated by an asynchronous algorithm. In particular, whenever the number of messages in transit tends to increase, then there are many messages that are overtaken by others, and therefore discarded.

Outline of the paper

In Section 2, we present our main model and assumptions, and state the main results. In Section

3. In fact, it will be seen that, for $m = 1$, the logarithmic factor in the upper bound can be removed.

3, we prove these results and we argue that they constitute fairly tight bounds. In Section 4, we discuss issues related to the average time complexity of an asynchronous algorithm, under the same probabilistic model. Finally, in Section 5, we provide a brief discussion of alternative (possibly, more realistic) probabilistic models of interprocessor communication, and argue that under reasonable models, there will exist some mechanism that can keep the number of transmitted messages under control.

2. THE MODEL AND THE MAIN RESULTS

There are n processors, numbered $1, \ldots, n$, and each processor i has a set $A(i)$ of neighboring processors.[4] Let $d = \max_i |A(i)|$. The process starts at time $t = 0$, with processor 1 transmitting a message to its neighbors.

Whenever processor i receives a message, it can either ignore it, or it can (possibly, after some waiting time) transmit a message to some of its neighbors. Suppose that a message l is transmitted from i to j and, at some later time, another message l' is transmitted from i to j. If l' is received by j before l, we say that l has been *overtaken* by l', and that l is *discardable.* Our main assumption is:

Assumption 2.1: (a) Every discardable message is *ignored* by the receiving processor.
(b) Every nondiscardable message can trigger *at most* one transmission to each one of the neighbors of the receiving processor.

Assumption 2.1(b) allows a processor to ignore messages that are not discardable. In practical terms, this could correspond to a situation where a processor i receives a message, updates its value of y^i, evaluates $x_i = f_i(y^i)$ and finds that the new value of x_i is the same as the old one, in which case there is nothing to be communicated to the neighbors of i.

Our first assumption on the communication delays is the following:

Assumption 2.2: The communication delays of the different messages are independent, identically distributed, positive random variables.

Let D be a random variable distributed according to the common probability distribution of the communication delays. Let F be its cumulative probability distribution function; that is, $F(t) = \Pr[D \leq t]$. We will be using the following technical assumption on F:

Assumption 2.3: There exists some positive integer m and some $\epsilon_0 > 0$ [with $F(\epsilon_0) < 1$] such that F is m times differentiable in the interval $(0, 2\epsilon_0]$ and satisfies

$$\lim_{t \downarrow 0} F(t) = \lim_{t \downarrow 0} \frac{dF}{dt}(t) = \cdots = \lim_{t \downarrow 0} \frac{d^{m-1}F}{dt^{m-1}}(t) = 0 \qquad \text{and} \qquad \lim_{t \downarrow 0} \frac{d^m F}{dt^m}(t) > 0\,;$$

4. To simplify language, we make the assumption that $i \in A(j)$ if and only if $j \in A(i)$. Our subsequent results remain valid in the absence of this assumption.

moreover, there exist $c_1, c_2 > 0$ such that the mth derivative of F satisfies

$$c_1 \leq \frac{d^m F}{dt^m}(t) \leq c_2, \qquad \forall t \in (0, 2\epsilon_0].$$

Our assumption on the distribution of the delays is satisfied, in particular, in the case of a probability density function f that is right–continuous and infinitely differentiable at 0. Of course, the assumption also holds under rather milder conditions, such as right–continuity of f at 0 together with $\lim_{t \downarrow 0} f(t) > 0$; in this case, we have $m = 1$. (Various important distributions satisfy these properties; e.g., the exponential and the uniform distributions.) Assumption 2.3 is also satisfied by the Gamma distribution with m degrees of freedom.

Assumption 2.2 does not fully capture the intuitive notion of "completely random and independent" communication delays. For example, the way that Assumption 2.2 stands, it allows a processor to "know" ahead of time the communication delay of each one of the messages to be transmitted, and then act maliciously: choose the waiting time before sending each message so as to ensure that as few messages are discarded as possible. Such malicious behavior is more difficult to analyze, and also very unnatural. Our next assumption essentially states that as long as a message is in transit, there is no available information on the delay of that message, beyond the prior information captured by F.

Note that if a message has been in the air for some time $s > 0$, and only the prior information is available on the delay of that message, then its total delay D is a random variable with cumulative distribution function

$$G(r \mid s) = \Pr[D \leq r \mid D > s] = \frac{F(r) - F(s)}{1 - F(s)}, \qquad r \geq s. \tag{2.1}$$

[Of course, $G(r \mid s) = 0$ if $r < s$.]

Assumption 2.4: For every $s > 0$, $t \geq 0$, and every i, j, k, the following holds. The conditional distribution of the delay of the kth message transmitted from i to j, conditioned on this message having being sent at time t and not being received within s time units, and also conditioned on any other events that have occured up to time $t + s$, has the cumulative probability distribution function $G(\cdot \mid s)$.

Our main results are given by the following two theorems. In particular, Theorem 2.1 corresponds to the case where Assumption 2.3 is satisfied with $m = 1$, while Theorem 2.2 corresponds to $m > 1$.

Theorem 2.1: Assume that $T \geq 1$ and that $m = 1$. Then, there exists a constant A (depending only on the constants c_1, c_2 and ϵ_0 of Assumption 2.3), such that the expected total number of messages transmitted during the time interval $[0, T]$ is bounded by And^3T. ■

Theorem 2.2: Assume that $T \geq 1$ and that $m > 1$. Then, there exists a constant A' (depending only on the constants m, c_1, c_2 and ϵ_0 of Assumption 2.3), such that the expected total number of messages transmitted during the time interval $[0, T]$ is bounded by $A'nd^{2+\frac{1}{m}}(\ln d)^{1+\frac{1}{m}}T$. ■

Notice that the difference between Theorems 2.1 and 2.2 lies on the logarithmic factor; some more discussion on this point is presented in Subsection 3.3.

In the remainder of this section, we motivate Theorem 2.1, by considering the following special case:[5]

(i) The message delays have exponential probability distributions, with mean 1.

(ii) Each processor transmits a message to every other processor, immediately upon receipt of a nondiscardable message. (That is, the underlying graph is assumed to be complete.)

Let $m_{ij}(t)$ be the number of messages in transit from i to j at time t, that have not been overtaken; that is, no later transmitted message from i to j has already reached its destination. [The notation $m_{ij}(t)$ should not be confused with the constant m involved in Assumption 2.3.] At time t, the rate at which messages arrive to j along the link (i,j) is $m_{ij}(t)$. Since any such arrival triggers a message transmission by j, the rate of increase of $m_{jk}(t)$ is $\sum_{i\neq j} m_{ij}(t)$. On the other hand, an arrival of a message travelling along the link (i,j) overtakes (on the average) half of the other messages in transit across that link. Thus,

$$\frac{d}{dt}E[m_{ij}(t)] = \sum_{k\neq i} E[m_{ki}(t)] - E[m_{ij}(t)] - \frac{E[(m_{ij}(t)-1)m_{ij}(t)]}{2}$$

$$\leq \sum_{k\neq i} E[m_{ki}(t)] - \frac{1}{2}E[m_{ij}(t)]^2. \tag{2.2}$$

Let $M(t) = \sum_{i=1}^{n}\sum_{j\neq i} E[m_{ij}(t)]$. The Schwartz inequality gives $\frac{1}{n^2}M^2(t) \leq \sum_{i=1}^{n}\sum_{j\neq i}E[m_{ij}(t)]^2$ and Eq. (2.2) becomes

$$\frac{d}{dt}M(t) \leq nM(t) - \frac{1}{2n^2}M^2(t).$$

Using the fact $\lim_{t\downarrow 0} M(t) = 1$ (because only one message initiates the execution of the algorithm), we obtain $M(t) \leq 2n^3$, for all $t > 0$. Thus, the rate of reception of nondiscardable messages, summed over all links, is $O(n^3)$. Since each such message reception generates $O(n)$ message transmissions, messages are generated at a rate of $O(n^4)$. We conclude that the expected number of messages generated during a time interval $[0,T]$ is $O(n^4T)$.

We can now provide some intuition for the validity of Theorem 2.1 for the case $m = 1$: messages with communication delay above ϵ_0 have very little probability of not being overtaken and can be ignored; messages with communication delay below ϵ_0 have approximately uniform distribution (cf. Assumption 2.3 with $m = 1$), which is approximately the same as the lower tail of an exponential distribution, for ϵ_0 small. Thus, we expect that the analysis for the case of exponential distributions should be representative of any distribution satisfying Assumption 2.3 with $m = 1$. In fact, the proof of Theorem 2.1 is based on the argument outlined above. The proof of Theorem 2.2 is based on a somewhat different idea and is more involved.

5. This calculation is due to David Aldous.

3. THE PROOFS OF THE MAIN RESULTS

We start by considering the transmissions along a particular link, say the link from i to j. Let M_i be the (random) number of messages transmitted by processor i along that link during the time interval $[0, T]$. Any such message is called *successful* if it arrives at j no later than time T and if it is not discarded upon arrival, that is, if that message has not been overtaken by a later transmitted message along the same link. Let S_{ij} be the number of successful messages sent from i to j. Since only successful messages can trigger a transmission by the receiving processor, we have

$$M_j \leq 1 + \sum_{i \in A(j)} S_{ij}, \tag{3.1}$$

which leads to

$$E[M_j] \leq 1 + \sum_{i \in A(j)} E[S_{ij}]. \tag{3.2}$$

Note that the term "+1" captures the possibility that processor j is the one that starts the process by transmitting a message, not triggered by the arrival of another message.

In order to establish Theorems 2.1 and 2.2, we upper bound $E[S_{ij}]$ by an appropriate function of $E[M_j]$. This is done in a different way for each of the two theorems.

3.1 The Proof of Theorem 2.1

In the present subsection, we assume that Assumption 2.3 is satisfied with $m = 1$. Prior to proving Theorem 2.1, we establish the following result:

Lemma 3.1: There exist constants B, B', depending only on the constants c_1, c_2 and ϵ_0 of Assumption 2.3, such that

$$E[S_{ij}] \leq B\sqrt{T\,E[M_i]} + B'T. \tag{3.3}$$

■

Once Lemma 3.1 is established, the proof of Theorem 2.1 is easily completed by the following argument. Let $Q = \max_i E[M_i]$. Then, Eq. (3.3) yields $E[S_{ij}] \leq B\sqrt{TQ} + B'T$. Using Eq. (3.2), we obtain $E[M_j] \leq 1 + dB\sqrt{TQ} + dB'T$. Taking the maximum over all j, and using the fact $dT \geq 1$ (for $T \geq 1$), we obtain $Q \leq dB\sqrt{TQ} + d(B'+1)T$. Suppose that $Q \geq T$. Then, $Q \leq d(B+B'+1)\sqrt{TQ}$, which yields $Q \leq (B+B'+1)^2 d^2 T$. If $Q < T$, this last inequality is again valid. We conclude that there exists a constant A such that $Q \leq Ad^2T$. Each processor sends M_i messages along every link. Since $E[M_i] \leq Ad^2T$ and since there are at most nd links, the expected value of the total number of transmitted messages is bounded above by And^3T, which is the desired result.

It now remains to prove Lemma 3.1.

Proof of Lemma 3.1: For the purposes of the lemma, we only need to consider a fixed pair of processors i and j. We may thus simplify notation and use M and S instead of M_i and S_{ij}, respectively.

Note that if $E[M] \leq T/\epsilon_0{}^2$, then $E[S] \leq T/\epsilon_0{}^2$ (because $S \leq M$ with probability 1) and Eq. (3.3) holds, as long as B' is chosen larger than $1/\epsilon_0{}^2$. Thus, we only need to consider the case $E[M] > T/\epsilon_0{}^2$, which we henceforth assume.

Successful messages can be of two types:

(i) Those that reach their destination with a delay of at least ϵ_0; we call them *slow* messages.

(ii) Those that reach their destination with a delay smaller than ϵ_0; we call them *fast* messages.

Let S_s and S_f be the number of slow and fast successful messages, respectively. We will bound their respective expectations using two somewhat different arguments, starting with $E[S_f]$.

We split $[0, T]$ into disjoint time intervals of length

$$\delta \stackrel{\text{def}}{=} \sqrt{\frac{T}{E[M]}}.$$

To simplify notation, we assume that $\sqrt{T\,E[M]}$ is an integer. (Without this assumption, only some very minor modifications would be needed in the argument that follows.) Thus, the number of intervals is $T/\delta = \sqrt{T\,E[M]}$. Note also that $\delta < \epsilon_0$, due to our assumption $E[M] > T/\epsilon_0{}^2$.

Let $t_k = (k-1)\delta$ be the starting time of the kth interval. Let $\mathcal{I}_k$ be the set of messages transmitted during the kth interval, and let I_k be the cardinality of $\mathcal{I}_k$. Let $\mathcal{N}_k$ be the set of messages with the following properties:

(a) The time t at which the message was transmitted satisfies $t_k - \epsilon_0 < t \leq t_k$.

(b) At time t_k the message has not yet reached its destination.

(c) The message has not been overtaken by another message that has reached its destination by time t_k.

Thus, the set $\mathcal{N}_k$ contains the messages that are in transit at time t_k, that still have a hope of being successful (not yet overtaken), and that have not been in the air for "too long". Let N_k be the cardinality of $\mathcal{N}_k$.

Consider now a message in the set $\mathcal{N}_k$ and suppose that it was transmitted at time $t_k - s$, where $0 \leq s < \epsilon_0$. Such a message reaches its destination during the time interval $(t_k, t_{k+1}]$ with probability

$$G(\delta + s \mid s) = \frac{F(\delta + s) - F(s)}{1 - F(s)}.$$

[See Eq. (2.1) and Assumption 2.4.] Furthermore, Assumption 2.3 (which was taken to hold with $m = 1$) implies that

$$c_1\delta \leq F(\delta + s) - F(s) \leq c_2\delta, \qquad \forall \delta, s \in [0, \epsilon_0];$$

also, for $s \in [0, \epsilon_0]$, we have $0 < 1 - F(\epsilon_0) \leq 1 - F(s) \leq 1$. [Recall that $F(\epsilon_0) < 1$ by Assumption 2.3.] Thus, it follows that

$$c_1\delta \leq G(\delta + s \mid s) \leq \alpha_2\delta, \qquad \forall \delta, s \in [0, \epsilon_0], \tag{3.4}$$

where $\alpha_2 = c_2/[1 - F(\epsilon_0)]$. Therefore, the probability that a message in the set $\mathcal{N}_k$ reaches its destination during $(t_k, t_{k+1}]$ lies between $c_1\delta$ and $\alpha_2\delta$. Similarly, for any message in the set $\mathcal{I}_k$, the probability that it reaches its destination during the time interval $(t_k, t_{k+1}]$ is at most $F(\delta)$, which does not exceed $\alpha_2\delta$. [To see this, apply Eq. (3.4) with $s = 0$.]

For a message to be received during the time interval $(t_k, t_{k+1}]$ and for it to be successful and fast, it is necessary that it belong to the set $\mathcal{N}_k \cup \mathcal{I}_k$. Using the bounds of the preceding paragraph, the expected number of such successful fast messages is bounded above by $\alpha_2\delta(E[N_k + I_k])$. Adding over all k, we see that the expected number of successful fast messages satisfies

$$E[S_f] \leq \alpha_2\delta \sum_{k=1}^{T/\delta} E[N_k + I_k]. \tag{3.5}$$

Next, we estimate the number of messages in the set $\mathcal{N}_k$ that also belong to $\mathcal{N}_{k+1}$. (Notice that these two sets may possibly intersect, because $t_{k+1} - \epsilon_0 < t_k$ due to the assumption $\delta < \epsilon_0$.) Let us number the messages in the set $\mathcal{N}_k$ according to the times that they were transmitted, with earlier transmitted messages being assigned a smaller number. Note that the ℓth message in $\mathcal{N}_k$ belongs to $\mathcal{N}_{k+1}$ only if none of the messages $1, \ldots, \ell$ has been received during the time interval $(t_k, t_{k+1}]$. Using our earlier calculations, each message in $\mathcal{N}_k$ has a probability of at least $c_1\delta$ of being received during $(t_k, t_{k+1}]$. Using the independence of the delays of different messages (Assumption 2.4), the ℓth message in $\mathcal{N}_k$ makes it into $\mathcal{N}_{k+1}$ with probability no larger than $(1 - c_1\delta)^\ell$. Summing over all ℓ, the expected number of elements of $\mathcal{N}_k$ that make it into $\mathcal{N}_{k+1}$ is bounded above by $1/(c_1\delta)$. The set $\mathcal{N}_{k+1}$ consists of such messages together possibly with some of the elements of $\mathcal{I}_k$. We thus have

$$E[N_{k+1}] \leq \frac{1}{c_1\delta} + E[I_k]. \tag{3.6}$$

Combining Eqs. (3.5) and (3.6), and using the property $\sum_{k=1}^{T/\delta} E[I_k] = E[M]$, we obtain

$$\begin{aligned} E[S_f] &\leq \frac{\alpha_2 T}{c_1\delta} + \alpha_2\delta \sum_{k=1}^{T/\delta} E[I_{k-1} + I_k] \\ &\leq \frac{\alpha_2 T}{c_1\delta} + 2\alpha_2\delta\, E[M] \\ &= \left(\frac{\alpha_2}{c_1} + 2\alpha_2\right) \sqrt{T\, E[M]}. \end{aligned} \tag{3.7}$$

We now derive an upper bound for the expected number of successful "slow" messages. For the purposes of this argument, we split $[0, T]$ into intervals of length $\epsilon_0/2$. (The last such interval might have length smaller than $\epsilon_0/2$ if $2T/\epsilon_0$ is not an integer.) The total number of such intervals is $\lceil 2T/\epsilon_0 \rceil$. Let $t_k = (k-1)\epsilon_0/2$. Let us number the messages transmitted during $[t_k, t_{k+1}]$, in order of increasing time that they were transmitted. Clearly, a message generated at time $t_{k+1} - s$, with $0 \leq s \leq \epsilon_0/2$, is received during the time interval $[t_{k+1}, t_{k+2}]$ with probability $F(s + \epsilon_0/2) - F(s)$;

reasoning similarly as in previous cases, it is seen that this probability is at least $c_1(\epsilon_0/2)$. Notice now that for the ℓth message transmitted during $[t_k, t_{k+1}]$ to be a slow and successful message, it is necessary that none of the messages $1, \ldots, \ell$ transmitted during that same interval is received during the time interval $[t_{k+1}, t_{k+2}]$; the probability of this event is at most $\left(1 - c_1(\epsilon_0/2)\right)^{\ell}$. Thus, the expected number of messages that are transmitted during $[t_k, t_{k+1}]$ and are slow and successful is bounded above by $\frac{1}{c_1(\epsilon_0/2)}$. Adding over all k, we obtain

$$E[S_s] \leq \left\lceil \frac{2T}{\epsilon_0} \right\rceil \cdot \frac{1}{c_1(\epsilon_0/2)} \leq B'T, \tag{3.8}$$

where B' is a suitable constant.

Since $E[S] = E[S_f] + E[S_s]$, Eqs. (3.7) and (3.8) establish the lemma and the proof of Theorem 2.1 is complete. **Q.E.D.**

3.2 The Proof of Theorem 2.2

In the present subsection, we assume that Assumption 2.3 is satisfied with $m > 1$. Prior to proving Theorem 2.2, we establish the following result:

Lemma 3.2: There exists a constant $\hat{B}$, depending only on the constants m, c_1, c_2 and ϵ_0 of Assumption 2.3, such that

$$E[S_{ij}] \leq \hat{B} T^{\frac{m}{m+1}} \left(E[M_i]\right)^{\frac{1}{m+1}} \max\left\{1, \ln(E[M_i]/T)\right\}. \tag{3.9}$$

■

Once Lemma 3.2 is established, the proof of the Theorem 2.2 is completed by the following argument. Let $Q = \max_i E[M_i]$. Then, Eq. (3.9) yields $E[S_{ij}] \leq \hat{B} T^{\frac{m}{m+1}} Q^{\frac{1}{m+1}} \max\{1, \ln(Q/T)\}$. Using Eq. (3.2), we obtain $E[M_j] \leq 1 + d\hat{B} T^{\frac{m}{m+1}} Q^{\frac{1}{m+1}} \max\{1, \ln(Q/T)\}$. Taking the maximum over all j, and using the fact $dT \geq 1$ (for $T \geq 1$) we obtain $Q \leq dT + d\hat{B} T^{\frac{m}{m+1}} Q^{\frac{1}{m+1}} \max\{1, \ln(Q/T)\}$. Suppose that $Q > e^{\frac{m+1}{m}} T$. Then, $Q \leq d(\hat{B}+1) T^{\frac{m}{m+1}} Q^{\frac{1}{m+1}} \ln(Q/T)$, which yields

$$\frac{(Q/T)^{\frac{m}{m+1}}}{\ln\left[(Q/T)^{\frac{m}{m+1}}\right]} \leq \overline{B} d, \tag{3.10}$$

where $\overline{B} = \frac{m+1}{m}(\hat{B}+1)$.

Next, we prove the following auxiliary result: if $x > e$ and $\frac{x}{\ln x} \leq y$, then $x \leq 2y \ln y$. Indeed, since $\frac{x}{\ln x}$ is an increasing function of x for $x > e$, it is sufficient to show that if $\frac{x}{\ln x} = y$ then $x \leq 2y \ln y$. Thus, it is enough to show that $x \leq 2\frac{x}{\ln x} \ln\left(\frac{x}{\ln x}\right)$ or $x \leq 2x - 2x\frac{\ln\ln x}{\ln x}$; equivalently $2\ln\ln x \leq \ln x$ or $\ln x \leq \sqrt{x}$, which is true for all $x > e$.

Due to Eq. (3.10) and the assumption $Q > e^{\frac{m+1}{m}} T$, we can apply the above result with $x = (Q/T)^{\frac{m}{m+1}}$ and $y = \overline{B}d$; thus, it follows that

$$\left(\frac{Q}{T}\right)^{\frac{m}{m+1}} \leq 2\bar{B} d \ln(\bar{B} d),$$

which gives

$$Q \leq A' d^{1+\frac{1}{m}} (\ln d)^{1+\frac{1}{m}} T,$$

where A' is a suitable constant. If $Q \leq e^{\frac{m+1}{m}} T$, this last inequality is again valid. We conclude that there exists a constant A' such that $Q \leq A' d^{1+\frac{1}{m}} (\ln d)^{1+\frac{1}{m}} T$. Each processor sends M_i messages along every link. Since $E[M_i] \leq A' d^{1+\frac{1}{m}} (\ln d)^{1+\frac{1}{m}} T$ and since there are at most nd links, the expected value of the total number of transmitted messages is bounded above by $A' n d^{2+\frac{1}{m}} (\ln d)^{1+\frac{1}{m}} T$, which is the desired result.

It now remains to prove Lemma 3.2.

Proof of Lemma 3.2: For the purposes of the lemma, we only need to consider a fixed pair of processors i and j. We may thus simplify notation and use M and S instead of M_i and S_{ij}, respectively.

Let δ be defined as follows:

$$\delta \stackrel{\text{def}}{=} \left(\frac{T}{E[M]} \right)^{\frac{1}{m+1}}. \tag{3.11}$$

Note that if $\delta \geq \epsilon_0$, then $E[M] \leq T/{\epsilon_0}^{m+1}$, which implies that $E[M] \leq \left(\frac{1}{\epsilon_0}\right)^m T^{\frac{m}{m+1}} (E[M])^{\frac{1}{m+1}}$; therefore, Eq. (3.9) holds as long as $\hat{B}$ is chosen larger than $1/{\epsilon_0}^m$. Thus, we only need to consider the case $\delta < \epsilon_0$, which we henceforth assume.

We split the interval $[0, T]$ into disjoint intervals of length δ. To simplify notation, we assume that T/δ is an integer. (Without this assumption, only some very minor modifications would be needed in the arguments to follow.) For definiteness, let the qth interval be $\mathcal{I}_q = [(q-1)\delta, q\delta)$, with the exception of $\mathcal{I}_{T/\delta} = [T - \delta, T]$. Let M_q denote the number of messages generated during $\mathcal{I}_q$. Clearly, we have

$$\sum_{q=1}^{T/\delta} E[M_q] = E[M]. \tag{3.12}$$

Let S_q be the number of nondiscardable messages generated during $\mathcal{I}_q$. We have

$$\sum_{q=1}^{T/\delta} E[S_q] = E[S]. \tag{3.13}$$

Henceforth, we fix some $q \in \{1, \ldots, T/\delta\}$ and we concentrate on bounding $E[S_q]$.

Let $\hat{N}_q$ be the number of messages that are generated during the interval $\mathcal{I}_q$ and arrive *no later* than time $q\delta$. We will now bound $E[\hat{N}_q]$. Let $t_1, \ldots, t_{M_q}$ be the times in $\mathcal{I}_q$, in increasing order, at which messages are generated. Let $D_1, \ldots, D_{M_q}$ be the respective delays of these messages. We have

$$
\begin{aligned}
E[\hat{N}_q] &= \sum_{\ell=1}^{\infty} \Pr[M_q = \ell] \left\{ \sum_{k=1}^{\ell} \Pr[t_k + D_k \le q\delta \mid M_q = \ell] \right\} \\
&= \sum_{k=1}^{\infty} \sum_{\ell=k}^{\infty} \Pr[M_q = \ell \text{ and } t_k + D_k \le q\delta] \\
&= \sum_{k=1}^{\infty} \Pr[M_q \ge k \text{ and } t_k + D_k \le q\delta] \\
&= \sum_{k=1}^{\infty} \Pr[M_q \ge k] \Pr[D_k \le q\delta - t_k \mid M_q \ge k] \\
&\le \sum_{k=1}^{\infty} \Pr[M_q \ge k] \Pr[D_k \le \delta \mid M_q \ge k], \qquad (3.14)
\end{aligned}
$$

where the last inequality follows from the fact $t_k \ge (q-1)\delta$. By Assumption 2.4, the delay of a message is independent of all events that occured until the time of its generation; hence, we have

$$\Pr[D_k \le \delta \mid M_q \ge k] = F(\delta), \qquad (3.15)$$

because, at time t_k, the event $M_q \ge k$ is *known* to have occured. Furthermore, it is an immediate consequence of Assumption 2.3 that there exist constants $\alpha_1, \alpha_2 > 0$ such that

$$\alpha_1(x^m - y^m) \le F(x) - F(y) \le \alpha_2(x^m - y^m), \qquad \text{for } 0 \le y \le x \le 2\epsilon_0. \qquad (3.16)$$

(In particular, $\alpha_1 = \frac{\epsilon_1}{m!}$ and $\alpha_2 = \frac{\epsilon_2}{m!}$.) Applying Eq. (3.16) with $x = \delta$ and $y = 0$, we have $F(\delta) \le \alpha_2 \delta^m$; combining this with Eqs. (3.14) and (3.15), we obtain

$$E[\hat{N}_q] \le \alpha_2 \delta^m \sum_{k=1}^{\infty} \Pr[M_q \ge k] = \alpha_2 \delta^m E[M_q]. \qquad (3.17)$$

Let $\tilde{S}_q$ be the number of nondiscardable messages that are generated during $\mathcal{I}_q$ and arrive *after* time $q\delta$. Recalling that $\hat{N}_q$ is the number of messages that are generated during $\mathcal{I}_q$ and arrive no later than $q\delta$, we have

$$E[S_q] \le E[\hat{N}_q] + E[\tilde{S}_q]. \qquad (3.18)$$

Eq. (3.17) provides a bound for $E[\hat{N}_q]$; thus, it only remains to upper bound $\tilde{S}_q$.

Let $\mathcal{F}$ be a σ-field describing the history of the process up to and including time $q\delta$. Let N_q be the number of messages that were transmitted during $\mathcal{I}_q$ and have not been received by time $q\delta$; note that $N_q = M_q - \hat{N}_q$. We will be referring to the aforementioned N_q messages as $P_1, \ldots, P_{N_q}$. In particular, message P_k is taken to be generated at time t_k, where $(q-1)\delta \le t_1 \le t_2 \le \cdots \le t_{N_q} < q\delta$. The delay of P_k is denoted by D_k; there holds $D_k \ge q\delta - t_k$, by assumption. Note that N_q and

$(t_1, \ldots, t_{N_q})$ are $\mathcal{F}$-measurable. Also, Assumption 2.4 implies that, conditioned on $\mathcal{F}$, the random variables $D_1, \ldots, D_{N_q}$ are independent, with the conditional cumulative distribution of D_k being $G(\cdot \mid q\delta - t_k)$.

In the analysis to follow, we assume that $N_q \geq 2$; the trivial cases $N_q = 0$ and $N_q = 1$ will be considered at the end. At time $q\delta$, message P_k has been in the air for $s_k \stackrel{\text{def}}{=} q\delta - t_k$ time units; notice that $s_k \leq \delta$. Let R_k denote the random variable $D_k - s_k$; that is, R_k is the residual time (after $q\delta$) for which message P_k will remain in the air. As argued above, conditioned on $\mathcal{F}$, the random variables $R_1, \ldots, R_{N_q}$ are independent; moreover, the conditional cumulative distribution function of R_k is given by

$$H_k(r) \stackrel{\text{def}}{=} \Pr[R_k \leq r \mid \mathcal{F}] = G(r + s_k \mid s_k) = \frac{F(r + s_k) - F(s_k)}{1 - F(s_k)}. \tag{3.19}$$

Let $f(r) = (dF/dr)(r)$ and $h_k(r) = (dH_k/dr)(r)$; both derivatives are guaranteed to exist in the interval $(0, \epsilon_0]$ due to Assumption 2.3 and the fact $s_k < \delta < \epsilon_0$. Clearly, if $k \neq N_q$, then for P_k not to be discardable it is necessary that messages $P_{k+1}, \ldots, P_{N_q}$ arrive later than P_k. Therefore, we have

$$\begin{aligned}
\Pr[P_k \text{ is nondiscardable} \mid \mathcal{F}] &\leq \Pr[R_k \leq R_\ell \text{ for } \ell = k+1, \ldots, N_q \mid \mathcal{F}] \\
&= \int_0^\infty \Pr[r \leq R_\ell \text{ for } \ell = k+1, \ldots, N_q \mid \mathcal{F}]\, dH_k(r) \\
&= \int_0^\infty \left(\prod_{\ell=k+1}^{N_q} \Pr[R_\ell \geq r \mid \mathcal{F}] \right) dH_k(r) \\
&= \int_0^\infty \left(\prod_{\ell=k+1}^{N_q} [1 - H_\ell(r)] \right) dH_k(r) \\
&\leq H_k(\delta) + \int_\delta^{\epsilon_0} \left(\prod_{\ell=k+1}^{N_q} [1 - H_\ell(r)] \right) dH_k(r) + \prod_{\ell=k+1}^{N_q} [1 - H_\ell(\epsilon_0)].
\end{aligned} \tag{3.20}$$

In what follows, we derive an upper bound for each of the three terms in the lower part of the above equation.

Starting with $H_k(\delta)$, we have

$$H_k(\delta) \leq \frac{\alpha_2[(\delta + s_k)^m - s_k^m]}{1 - F(s_k)},$$

due to Eqs. (3.19) and (3.16). Since $s_k \leq \delta$, we have $(s_k + \delta)^m - \delta^m \leq (2^m - 1)\delta^m$; moreover, there holds $0 < 1 - F(\epsilon_0) \leq 1 - F(s_\ell)$, because $s_\ell \leq \delta < \epsilon_0$ and $F(\epsilon_0) < 1$ (see Assumption 2.3). Combining these facts, it follows that

$$H_k(\delta) \leq \frac{\alpha_2(2^m - 1)}{1 - F(\epsilon_0)} \delta^m = \beta_1 \delta^m. \tag{3.21}$$

Furthermore, let Δ be a small positive real number; by Eq. (3.19), we have

$$H_\ell(r+\Delta) - H_\ell(r) = \frac{F(r+s_\ell+\Delta) - F(r+s_\ell)}{1-F(s_\ell)}.$$

Since $s_\ell \le \delta < \epsilon_0$, it follows from Eq. (3.16) that

$$\frac{\alpha_1}{1-F(s_\ell)}[(r+s_\ell+\Delta)^m - (r+s_\ell)^m] \le H_\ell(r+\Delta) - H_\ell(r) \le \frac{\alpha_2}{1-F(s_\ell)}[(r+s_\ell+\Delta)^m - (r+s_\ell)^m], \quad \forall r \in [0,\epsilon_0].$$

Reasoning similarly as in the case of Eq. (3.21), it follows (after some algebra) that

$$\alpha_1[(r+\Delta)^m - r^m] \le H_\ell(r+\Delta) - H_\ell(r) \le \frac{\alpha_2(2^m-1)}{1-F(\epsilon_0)}[(r+\Delta)^m - r^m], \qquad \forall r \in [0,\epsilon_0]. \quad (3.22)$$

On the other hand, using Eq. (3.16), we have

$$\alpha_1[(r+\Delta)^m - r^m] \le F(r+\Delta) - F(r) \le \alpha_2[(r+\Delta)^m - r^m], \qquad \forall r \in [0,\epsilon_0];$$

this together with Eq. (3.22) implies that there exist constants $\beta_2, \beta_3 > 0$, which do *not* depend on ℓ, such that

$$\beta_3[F(r+\Delta) - F(r)] \le H_\ell(r+\Delta) - H_\ell(r) \le \beta_2[F(r+\Delta) - F(r)], \qquad \forall r \in [0,\epsilon_0].$$

Using this, it follows easily that

$$h_\ell(r) \le \beta_2 f(r), \qquad \forall r \in [0,\epsilon_0], \quad (3.23)$$

and

$$H_\ell(r) \ge \beta_3 F(r), \qquad \forall r \in [0,\epsilon_0]. \quad (3.24)$$

Combining Eqs. (3.23) and (3.24), we have

$$\begin{aligned}
\int_\delta^{\epsilon_0} \left(\prod_{\ell=k+1}^{N_q} [1 - H_\ell(r)] \right) dH_k(r) &\le \beta_2 \int_\delta^{\epsilon_0} [1-\beta_3 F(r)]^{N_q-k} f(r) dr \\
&= \frac{\beta_2}{\beta_3} \int_\delta^{\epsilon_0} [1-\beta_3 F(r)]^{N_q-k} d\big(\beta_3 F(r)\big) \\
&\le \frac{\beta_2}{\beta_3} \int_0^1 (1-y)^{N_q-k} dy \\
&= \frac{\beta_2}{\beta_3} \frac{1}{N_q - k + 1}, \qquad (3.25)
\end{aligned}$$

where we have also used the fact $\beta_3 F(\epsilon_0) \le H_k(\epsilon_0) \le 1$ [see Eq. (3.22) with $r = \epsilon_0$ and $\ell = k$]. Similarly, by Eq. (3.24), we have

$$\prod_{\ell=k+1}^{N_q} [1 - H_\ell(\epsilon_0)] \le [1-\beta_3 F(\epsilon_0)]^{N_q-k} = \gamma^{N_q-k}, \quad (3.26)$$

where γ is constant and satisfies $0 \leq \gamma < 1$.

Combining Eqs. (3.20), (3.21), (3.25) and (3.26), we obtain

$$\Pr[P_k \text{ is nondiscardable} \mid \mathcal{F}] \leq \beta_1 \delta^m + \frac{\beta_2}{\beta_3} \frac{1}{N_q - k + 1} + \gamma^{N_q - k}.$$

The above result holds for $k = 1, \ldots, N_q - 1$; adding over all those k, we have

$$\sum_{k=1}^{N_q - 1} \Pr[P_k \text{ is nondiscardable} \mid \mathcal{F}] \leq \beta_1 \delta^m (N_q - 1) + \frac{\beta_2}{\beta_3} \sum_{k=1}^{N_q - 1} \frac{1}{N_q - k + 1} + \sum_{k=1}^{N_q - 1} \gamma^{N_q - k}. \quad (3.27)$$

Notice that

$$\sum_{k=1}^{N_q - 1} \frac{1}{N_q - k + 1} = \sum_{k=2}^{N_q} \frac{1}{k} \leq \ln(N_q + 1),$$

and

$$\sum_{k=1}^{N_q - 1} \gamma^{N_q - k} \leq \sum_{k=1}^{\infty} \gamma^k = \frac{\gamma}{1 - \gamma},$$

because $0 \leq \gamma < 1$. Thus, it follows from Eq. (3.27) that

$$E[\tilde{S}_q \mid \mathcal{F}] = \sum_{k=1}^{N_q} \Pr[P_k \text{ is nondiscardable} \mid \mathcal{F}] \leq \beta_1 \delta^m N_q + \frac{\beta_2}{\beta_3} \ln(N_q + 1) + \frac{\gamma}{1 - \gamma} + 1,$$

where the term "+1" bounds the probability that P_{N_q} is nondiscardable. The above result was established for all $N_q \geq 2$; it is straightforward to see that it also holds for $N_q = 1$ and for $N_q = 0$. Thus, relaxing the conditioning on $\mathcal{F}$, we obtain

$$E[\tilde{S}_q] \leq \beta_1 \delta^m E[N_q] + \frac{\beta_2}{\beta_3} E[\ln(N_q + 1)] + \frac{1}{1 - \gamma}.$$

This together with Eqs. (3.18) and (3.17) implies that

$$E[S_q] \leq (\alpha_2 + \beta_1) \delta^m E[M_q] + \frac{\beta_2}{\beta_3} E[\ln(M_q + 1)] + \frac{1}{1 - \gamma},$$

where we have also used the fact $N_q \leq M_q$ with probability 1. The above inequality holds for all $q \in \{1, \ldots, T/\delta\}$; adding over all q, and using Eq. (3.13), we obtain

$$E[S] = \sum_{q=1}^{T/\delta} E[S_q] \leq (\alpha_2 + \beta_1) \delta^m \sum_{q=1}^{T/\delta} E[M_q] + \frac{\beta_2}{\beta_3} \sum_{q=1}^{T/\delta} E[\ln(M_q + 1)] + \frac{1}{1 - \gamma} \frac{T}{\delta}. \quad (3.28)$$

By concavity of the logarithmic function, we have $E[\ln(M_q + 1)] \leq \ln(E[M_q + 1])$ (due to Jensen's inequality); hence, there holds

$$\sum_{q=1}^{T/\delta} E[\ln(M_q + 1)] \leq \sum_{q=1}^{T/\delta} \ln(E[M_q] + 1) \leq \frac{T}{\delta} \ln \left(\frac{\delta}{T} \sum_{q=1}^{T/\delta} E[M_q] + 1 \right),$$

where we have again used the concavity property. This together with Eqs. (3.12) and (3.28) implies that

$$E[S] \leq (\alpha_2 + \beta_1)\delta^m E[M] + \frac{\beta_2}{\beta_3}\frac{T}{\delta}\ln\left(\frac{\delta}{T}E[M]+1\right) + \frac{1}{1-\gamma}\frac{T}{\delta}. \tag{3.29}$$

By Eq. (3.11), we have $\delta^m E[M] = \frac{T}{\delta} = T^{\frac{m}{m+1}}(E[M])^{\frac{1}{m+1}}$ and $\frac{\delta}{T}E[M] = 1/\delta^m = (E[M]/T)^{\frac{m}{m+1}}$; since $\delta < \epsilon_0$, we have $\frac{\delta}{T}E[M] > 1/{\epsilon_0}^m$, which gives (after some algebra) that

$$\ln\left(\frac{\delta}{T}E[M]+1\right) \leq \ln\left(\frac{\delta}{T}E[M]\right) + \ln({\epsilon_0}^m + 1).$$

Using these facts, it follows from Eq. (3.29) that

$$E[S] \leq [\alpha_2+\beta_1+\frac{1}{1-\gamma}+\ln({\epsilon_0}^m+1)]T^{\frac{m}{m+1}}(E[M])^{\frac{1}{m+1}} + \frac{m\beta_2}{(m+1)\beta_3}T^{\frac{m}{m+1}}(E[M])^{\frac{1}{m+1}}\ln(E[M]/T);$$

this proves the lemma for the case $\delta < \epsilon_0$. **Q.E.D.**

3.3 Some Further Results

First, we discuss a generalization of Theorems 2.1 and 2.2. Let us suppose that the distribution of the delays is as described by Assumption 2.3, except that it is shifted to the right by a positive amount. (For example, the delay could be the sum of a positive constant and an exponentially distributed random variable.) As far as a particular link is concerned, this change of the probability distribution is equivalent to delaying the time that each message is transmitted by a positive constant. Such a change does not affect the number of overtakings that occur on any given link. Thus, Lemmas 3.1 and 3.2 remain valid, and Theorems 2.1 and 2.2 still hold.

Next, we examine cases for which the bounds of Theorems 2.1 and 2.2 are *tight*. In particular, for $m = 1$, we are looking for cases where the expected total number of messages per interval of duration $T \geq 1$ is $\Omega(nd^3T)$. Obviously, this is true whenever $d = O(1)$. Moreover, the calculation presented in the last paragraph of Section 2 suggests that the bound is tight for the special case considered therein: namely, for a complete graph and exponential delay distribution.

As far as the bound of Theorem 2.2 is concerned, it is again tight whenever $d = O(1)$. Next, we argue that this bound is close to being tight in the case where all nodes have roughly the *same degree* [which is $\Omega(1)$]. Even though we have not completely established this claim, we provide some strong evidence for it; in particular, we show that, if messages are generated at *constant* pace, then the bound of Lemma 3.2 is tight within a factor of $(\ln d)^{1+\frac{1}{m}}$. Recall that this bound is interesting for $E[M] \gg T$. Thus, the lemma to follow pertains to a case where "many" messages are transmitted.

Lemma 3.3: Suppose that a fixed number M of messages are transmitted along a link, with the kth message being generated at time $(k-1)\frac{T}{M}$; let S be the corresponding (random) number of

nondiscardable messages. Assume that

$$M \geq T \max\left\{ \left(\frac{\beta}{2\epsilon_0}\right)^{m+1}, \left(\frac{2}{\beta}\right)^{1+\frac{1}{m}} \right\}, \tag{3.30}$$

where the constant β satisfies

$$\frac{1}{m+1}\alpha_2 \beta^{m+1} = \frac{1}{2}. \tag{3.31}$$

[α_2 is the constant appearing in Eq. (3.16).] Then there exists an absolute constant $\tilde{B}$ such that

$$E[S] \geq \tilde{B} T^{\frac{m}{m+1}} M^{\frac{1}{m+1}}. \qquad \blacksquare$$

Proof: The main idea for proving this result is to show that any message with delay not exceeding $\beta\delta$ [where β is the constant defined in Eq. (3.31) and δ is given by Eq. (3.11)] is *not* discarded with probability *at least* $\frac{1}{2}$.

Let Δ be defined as follows:

$$\Delta \stackrel{\text{def}}{=} \frac{T}{M}; \tag{3.32}$$

also, let L be defined as follows:

$$L \stackrel{\text{def}}{=} \frac{\beta\delta}{\Delta}; \tag{3.33}$$

Using Eqs. (3.11) and (3.32) we have $L = \beta(\frac{M}{T})^{\frac{m}{m+1}}$. Thus, by Eq. (3.30), we have $L \geq 2$; for simplicity, we assume that L is integer. We denote by P_k the message generated at time $(k-1)\Delta$, for $k = 1, \ldots, M$; let D_k denote the delay of this message. Assume that $D_k \leq \beta\delta$; that is, message P_k arrives prior to time $(k-1)\Delta + \beta\delta = (k+L-1)\Delta$ [see also Eq. (3.33)]. Then, P_k may be discarded *only* if at least one of the following events occurs: either P_{k+1} arrives prior to time $(k+L-1)\Delta$ or P_{k+2} arrives prior to time $(k+L-1)\Delta$ or ... or P_{k+L-1} arrives prior to time $(k+L-1)\Delta$. (Notice that messages $P_{k+L}, P_{k+L+1}, \ldots, P_M$ cannot cause the discarding of P_k, because they are generated after the latter has arrived.) Applying the union bound, it follows that

$$\Pr[P_k \text{ is discarded} \mid D_k \leq \beta\delta] \leq F(L\Delta - \Delta) + F(L\Delta - 2\Delta) + \cdots + F\big(L\Delta - (L-1)\Delta\big) = \sum_{\ell=1}^{L-1} F(\ell\Delta). \tag{3.34}$$

In the argument presented above, it was implicitly assumed that $k + L - 1 \leq M$, in order that all messages $P_{k+1}, \cdots, P_{k+L-1}$ actually exist; however, it is straightforward that Eq. (3.34) holds even if this is not the case.

Notice now that, by Eqs. (3.11) and (3.30) , we have $\beta\delta \leq 2\epsilon_0$; this together with Eq. (3.33) implies that $\ell\Delta \leq 2\epsilon_0$ for $\ell = 1, \ldots, L-1$. Thus, applying Eq. (3.16) with $x = \ell\Delta$ and $y = 0$, we have $F(\ell\Delta) \leq \alpha_2(\ell\Delta)^m$ for $\ell = 1, \ldots, L-1$. Combining this with Eq. (3.34), we obtain

$$\Pr[P_k \text{ is discarded} \mid D_k \leq \beta\delta] \leq \alpha_2 \Delta^m \sum_{\ell=1}^{L-1} \ell^m;$$

furthermore, there holds

$$\sum_{\ell=1}^{L-1} \ell^m \leq \int_0^L x^m dx = \frac{L^{m+1}}{m+1}.$$

Thus, it follows that

$$\Pr[P_k \text{ is discarded} \mid D_k \leq \beta\delta] \leq \alpha_2 \Delta^m \frac{L^{m+1}}{m+1}. \tag{3.35}$$

Notice now that, by Eq. (3.33), we have $\Delta^m L^{m+1} = \beta^{m+1} \frac{\delta^{m+1}}{\Delta}$; since $\delta^{m+1} = \Delta$ [due to Eqs. (3.11) and (3.32)], it follows that $\Delta^m L^{m+1} = \beta^{m+1}$. This together with Eq. (3.35) implies that

$$\Pr[P_k \text{ is discarded} \mid D_k \leq \beta\delta] \leq \frac{1}{m+1} \alpha_2 \beta^{m+1} = \frac{1}{2},$$

where we have also used the definition of β [see Eq. (3.31)]. Thus, each message with delay not exceeding $\beta\delta$ is nondiscardable with probability at least $\frac{1}{2}$. Therefore, we have

$$E[S] \geq \frac{1}{2} F(\beta\delta) M. \tag{3.36}$$

Recall now that $\beta\delta \leq 2\epsilon_0$; thus, it follows from Eq. (3.16) that $F(\beta\delta) \geq \alpha_1 (\beta\delta)^m$; this together with Eqs. (3.36) and (3.11) implies that

$$E[S] \geq \frac{1}{2} \alpha_1 (\beta\delta)^m M = \frac{1}{2} \alpha_1 \beta^m T^{\frac{m}{m+1}} M^{\frac{1}{m+1}},$$

which proves the lemma. **Q.E.D.**

The lower bound of Lemma 3.3 differs from the upper bound of Lemma 3.2 by a factor of $\ln(E[M]/T)$. We conjecture that Lemma 3.3 is closer to the truth; that is, we believe that $E[S_{ij}] \leq B T^{\frac{m}{m+1}} (E[M_i])^{\frac{1}{m+1}}$, for some constant B. Some evidence is provided by Lemma 3.1, which shows that the conjecture is true for $m = 1$. Furthermore, we can also prove the conjecture if the number M_i of transmitted messages is deterministic and the generation times of the M_i messages are also deterministic. If our conjecture is true, then the logarithmic factor in Theorem 2.2 is redundant.

4. SOME REMARKS ON THE TIME COMPLEXITY

In this section, we still assume that the model of Section 2 is in effect. Furthermore, to simplify the discussion, let us assume that if a message reception triggers the transmission of messages by the receiving processor, these latter messages are transmitted without any waiting time.

Consider the asynchronous Bellman–Ford algorithm and consider a path $(i_k, i_{k-1}, \ldots, i_1, 0)$ from a node i_k to the destination node 0. We say that this path has been *traced* by the algorithm if there exist times $t_1, t_2, \ldots, t_k$ such that a message is transmitted by processor i_j at time t_j and this message is received by processor i_{j+1} at time t_{j+1}, $j = 1, \ldots, k-1$. Under the initial conditions introduced in Section 1, it is easily shown [BT] that the shortest distance estimate x_{i_k} of processor

i_k becomes equal to the true shortest distance as soon as there exists a shortest path from i_k to 0 that has been traced by the algorithm.

It is easily seen that under the model of Section 2, the time until a path is traced is bounded by the sum of (at most n) i.i.d. random variables. Assuming that the delay distribution has an exponentially decreasing tail, we can apply large deviations bounds on sums of independent random variables (e.g., the Chernoff bound [C]). We then see that the time until the termination of the asynchronous Bellman–Ford algorithm is $O(n)$, with overwhelming probability. Furthermore, the expected duration of the algorithm is also $O(n)$.

From the above discussion and Theorem 2.1, we can conclude that, for $m = 1$, the number of messages until termination of the asynchronous Bellman–Ford is $O(n^2d^3)$, with overwhelming probability.[6] Similarly, for $m > 1$, the corresponding upper bound is $O(n^2d^{2+\frac{1}{m}}(\ln d)^{1+\frac{1}{m}})$. We note that for sparse graphs [i.e., when $d = O(1)$], the asynchronous Bellman–Ford has very good communication complexity, equal to the communication complexity of its synchronous counterpart.

It should be clear at this point that the above argument is not specific to the Bellman–Ford algorithm. In particular, any asynchronous algorithm with polynomial average time complexity will also have polynomial communication complexity, on the average.

5. DIFFERENT MODELS

We have established so far that (under the assumption of i.i.d. message delays) the average communication complexity of asynchronous distributed algorithms is quite reasonable. In particular, discarding messages that are overtaken by others is a very effective mechanism for keeping the number of messages under control.

Modeling message delays as i.i.d. random variables seems reasonable when a "general mail facility" is used for message transmissions, and the messages corresponding to the algorithm are only a small part of the facility's load. On the other hand, for many realistic multiprocessor systems, the i.i.d. assumption could be unrealistic. For example, any system that is guaranteed to deliver messages in the order that they are transmitted (FIFO links) will violate the i.i.d. assumption (unless the delays have zero variance). This raises the issue of constructing a meaningful probabilistic model of FIFO links. In our opinion, in any such model (and, furthermore, in any physical implementation of such a model) the links have to be modeled by servers preceded by

6. For $m = 1$, the formal argument goes as follows. If T is the random time until termination and $C(t)$ is the number of messages transmitted until time t, then

$$\Pr[C(\infty) \geq A_1A_2n^2d^3] \leq \Pr[T \geq A_1n] + \Pr[C(A_1n) \geq A_1A_2n^2d^3].$$

We bound $\Pr[T \geq A_1n]$ using the Chernoff bound, and we bound $\Pr[C(A_1n) \geq A_1A_2n^2d^3]$ using Theorem 2.1 and the Markov inequality.

buffers, in the usual queueing–theoretic fashion. We discuss such a model below.

Let us assume, for concreteness, that each link consists of an infinite buffer followed by a server with i.i.d., exponentially distributed, service times. In this setup, the following modification to the algorithm makes most sense: whenever there is a new arrival to a buffer, every message that has been placed earlier in that same buffer, but has not yet been "served" by the server, should be deleted. This modification has no negative effects on the correctness and termination of an asynchronous distributed algorithm. Furthermore, the rate at which a processor receives messages from its neighbors is $O(d)$. This is because there are at most d incoming links and the arrival rate along each link is constrained by the service rate of the server corresponding to each link. Each message arrival triggers $O(d)$ message transmissions. We conclude that the expected communication complexity of the algorithm will be $O(nd^2T)$, where T is the running time of the algorithm.

We have once more reached the conclusion that asynchronous algorithms with good time complexity T will also have a good communication complexity.

Let us conclude by mentioning that an alternative mechanism for reducing the communication complexity of an asynchronous algorithm is obtained by introducing a "synchronizer" [A]. A synchronizer could result in a communication complexity which is even better than the one predicted by Theorem 2.1 or by the calculation in this section. On the other hand, our results indicate that acceptable communication complexity is possible even without a synchronizer.

REFERENCES

[A] Awerbuch, B., "Complexity of network synchronization", *Journal of the ACM*, 32, 1985, pp. 804–823.

[B] Bertsekas, D.P., "Distributed dynamic programming", *IEEE Transactions on Automatic Control*, AC–27, 1982, pp. 610–616.

[BG] Bertsekas, D.P., and R.G. Gallager, *Data Networks*, Prentice Hall, Englewood Cliffs, NJ, 1987.

[BT] Bertsekas, D.P., and J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[C] Chernoff, H., "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on a Sum of Observations", *Annals of Mathematical Statistics*, 23, 1952, pp. 493–507.

Acknowledgement: We are grateful to David Aldous for carrying out the calculation in the end of Section 2, which suggested that a nice result should be possible for the general case. Furthermore, he suggested that the correct power of d in Theorem 2.2 is $d^{2+\frac{1}{m}}$, by deriving, for the case of a single arc, the exact expression for the rate of arrivals of nondiscardable messages when the message transmission times are Poisson.