Greedy Dynamic Routing on Arrays

Nabil Kahale* Tom Leighton †

Abstract

We study the problem of dynamic routing on arrays. We prove that a large class of greedy algorithms perform very well on average. In the dynamic case, when the arrival rate of packets in an $N \times N$ array is at most 99% of network capacity, we establish an exponential bound on the tail of the delay distribution. Moreover, we show that, in any window of T steps, the maximum queue-size is $O(1 + \log T / \log N)$ with high probability. We extend these results to the case of bit-serial routing, and to the static case. We also calculate the exact value of the ergodic expected delay and queue-sizes under the farthest-first protocol for the one dimensional array, and for the ring when the arrivals are Poisson.

1 Introduction

Many parallel machines, such as the MPP, Ametek and Intel Touchstone are configured as a low-dimensional array containing a large number of processors. These machines generally route packets using simple greedy algorithms. While these algorithms tend to behave well experimentally, their analysis from the theoretical point of view can be challenging.

In this paper, we analyse rigorously the problem of dynamic routing on arrays for a large class of greedy protocols. At each step and at each node, a packet is generated with a fixed probability and its destination is uniformly distributed among all nodes of the network. Packets are routed along edges of increasing dimension. This model has been considered by Leighton [9] with the additional assumption that packets with the furthest destination are routed first. Under this assumption he showed that, if the bisection width condition is met, the system is stable. Moreover, the tails of the distributions of the delay (i.e. number of steps a packet does not

*XEROX Palo Alto Research Center, 3333 Coyote Hill Road, CA 94304. Partially supported by the NSF under grant CCR–9404113. Email: kahale@parc.xerox.com. This work was mostly done while the author was at the Massachusetts Institute of Technology, and partly when he was at DIMACS and at XEROX Palo Alto Research Center.

†Mathematics Department and Laboratory for Computer Science, Massachsetts Institute of Technology, Cambridge MA 02139. Supported in part by AFOSR Grant F49620-92-J-0125 and ARPA Grants N00014-91-J-1698 and N00014-92-1799.

move) and of the queue-sizes at each node are exponentially decreasing, with mean bounded by a constant *independent of the size of the array.* He also showed that the same results hold for arbitrary protocols if the load is less than a half.

Recently, Mitzenmacher [10] established a tight bound (up to a multiplicative constant) on the expected time that packets spend in the system under the first-in first-out (FIFO) protocol, when the load is less than 1. His analysis follows the work of Tsitsiklis and Stamoulis [11] who analysed the problem of dynamic routing on hypercubes and butterflies under the FIFO protocol using techniques from queuing theory. Dynamic routing on arrays has also been studied in [5] under a different model, where the transmission times are exponentially distributed with mean one, instead of being constant. An exact analysis [5] of various parameters of the system is derived under that model for the FIFO protocol. The expected delay under that model gives an upper bound on the expected delay under the unit transmission model as was shown very recently (for the FIFO protocol and a large class of networks) in [6]. However, the bound on the expected delay in [5, 6, 10] is linear in N (for the FIFO protocol in an $N \times N$ array), whereas it is constant in [9] (for the farthest-first protocol.) Also, the approach in [6, 10, 11] does not seem to yield any non-trivial bounds on the tails of the distributions of the queue-sizes or on the time packets spend in system. The problem of dynamic routing on rings under special protocols has also been studied in [3].

In this paper, we apply successfully generating functions techniques to analyse the problem of dynamic greedy routing on arrays. Generating functions are a standard tool in probabilistic analysis [2, 4], but we are not aware of their use in routing problems, except in a few trivial cases. Our first main result applies to a wide class of greedy algorithms that route along edges of increasing dimension. We show that, under an *arbitrary greedy protocol*, if the arrival rate of packets is at most 99% of network capacity, an exponential bound holds on the tail of the distribution of the delay. Moreover, in any window of T steps, the maximum queue-size is $O(1 + \log T/\log N)$ with high probability. If the load is fixed and the number of nodes is sufficiently large, our bound on the expected delay improves upon the bound in [10] for the FIFO protocol. We extend these results to the case of bit-serial routing, and to the static case. Our application for bit-serial routing is an additional motivation for studying greedy protocols other than the farthest-first protocol: whereas the farthest-first protocol does not translate into a purely greedy protocol for bit-serial routing, other greedy protocols do. Some of the techniques we use in this part are based on [9].

Second, we calculate the exact value of the ergodic expected delay and queue-sizes under the farthest-first protocol for the one dimensional array, and for the ring when the arrivals are Poisson. If the load $\rho < 1$ is fixed and the number of nodes goes to infinity, we find that the expected delay on the ring (for either arrival model) converges towards $\frac{1}{2(1-\rho)^2} - \frac{1}{2}$ as N goes to infinity. This shows that the behavior of the ring under the farthest-first protocol is fundamentally different from the behavior of classical queuing systems, where the expected delay under heavy traffic [8] is asymptotically proportional to $1/(1-\rho)$.

The rest of the paper is organized as follows. In Section 2, we define the model more precisely.

In Section 3, we give some probabilistic lemmas to be used in later sections. In Section 4, we study the case of the one-dimensional array under arbitrary protocols. In Section 5, we consider the case of higher dimensional arrays, and also the static case. In Section 6, we analyse the expected delay for the farthest-first protocol on the one-dimensional array and on the ring. In Section 7, we extend some of our results to the case of cut-through routing.

2 Model and definitions

Consider an $\underbrace{N \times N \times \cdots \times N}_{k}$ array of fixed dimension k. At the beginning of each step and at each node, a packet is generated with probability $p = 4\rho/N$ and its destination is uniformly distributed among all nodes of the array. Packets arrivals and packets destinations are mutually independent over time. A packet generated at the beginning of a step can move during that step. At each step, at most one packet can cross a given oriented edge, and it reaches the other end at the beginning of next step. A packet that is not sent along an edge it wishes to traverse is stored in a queue along that edge. Packets are routed along edges of increasing dimension. For example, in the case of arrays of dimension two, packets are routed first to the right column and then to the right row. The protocol is greedy, that is, a nonempty queue must send a packet to its neighbor. Unless otherwise specified, the packet to be sent is chosen according to an arbitrary protocol, but the choice is based only on the state of the queue when the packet is sent. (The state of a queue consists of the destinations of the packets it contains together with their order of arrival, and may include any finite information that the packets hold.) All queues are empty at the beginning. No limit is set on the number of packets that a queue can hold.

We say that a packet is *delayed* at a given step if it does not move during this step. The delay of a packet is the total number of steps during which it does not move. If V is an integral random variable and $z > 0$, we denote by $\widehat{V}(z) = \mathrm{E}[z^V] = \sum_{i=-\infty}^{\infty} \Pr[V = i] z^i$ the generating function associated with V. We say that an exponential bound holds on the tail of the distribution of V if there is a positive constant c such that $\Pr[V \geq v] \leq e^{-cv}$ for any integer v. Unless otherwise specified, the decay parameter c is assumed to be function of ρ, and does not depend on N or on the protocol.

3 Probabilistic lemmas

This section contains some probabilistic results that will be used in later sections. The following classical results show that, under general conditions, the tail of a distribution is exponentially bounded if and only if the radius of convergence of its generating function is greater than 1.

Proposition 3.1 *If V is a nonnegative integral random variable and $z > 1$ is such that $\widehat{V}(z)$ is finite, then $\Pr[V \geq v] \leq \widehat{V}(z) z^{-v}$.*

Proof Since $z > 1$, we have

$$\Pr[V \geq v] = \Pr[z^V \geq z^v] \leq \frac{\mathrm{E}[z^V]}{z^v} = \widehat{V}(z) z^{-v}.$$

■

Proposition 3.2 *If V is a nonnegative integral random variable is such that* $\Pr[V \geq v] \leq e^{-cv}$*, for some $c > 0$, then $\widehat{V}(e^{c/2})$ is finite.*

Proof

$$\widehat{V}(e^{c/2}) = \sum_{i=0}^{\infty} \Pr[V = i] e^{ci/2} \leq \sum_{i=0}^{\infty} e^{-ci} e^{ci/2} = \frac{1}{1 - e^{-c/2}}.$$

■

Lemma 3.3 (Chernoff) *If S is the sum of Bernoulli random variables, $\beta > 1$ and $\Delta \geq \mathrm{E}[S]$, then*

$$\Pr[S \geq \beta\Delta] \leq e^{(\beta - 1 - \beta\log\beta)\Delta}.$$

Proofs of various versions of the Chernoff bound can be found in, e.g., [1].

Proposition 3.4 *If a random variable V is the sum of a finite number of independent Bernoulli variables, then $\widehat{V}(z) \leq e^{\mathrm{E}[V](z-1)}$ for $z \geq 0$.*

Proof Since the generating function of the sum of independent random variables is equal to the product of the generating functions of these variables and since the expectation of the sum is equal to the sum of expectations, it suffices to prove the proposition when V is a Bernoulli variable, which is straightforward since $\Pr[V = 1] = 1 - \Pr[V = 0] = \mathrm{E}[V]$ and so $\widehat{V}(z) = 1 - \mathrm{E}[V] + \mathrm{E}[V]z \leq e^{\mathrm{E}[V](z-1)}$. ■

It is known (see, for example, Hajek [4]) that under general conditions, if a sequence Y_t of random variables is such that $\mathrm{E}[Y_{t+1} - Y_t | Y_t > a] < -\epsilon$, for some $\epsilon > 0$ and a constant a, then the tail of Y_t is uniformly exponentially decreasing. We show below a version of this result where a is replaced by a random variable X_t with an exponentially decreasing tail.

Lemma 3.5 *Let R_t, $t \geq 0$, be a sequence of integral random variables such that $R_t + 1$ is the sum of Bernoulli independent random variables, with $\mathrm{E}[R_t + 1] \leq r < 1$. Let $z_0 = 2 - r$. Let X_t and Y_t, $t \geq 0$, be two sequences of nonnegative integral random variables such that $Y_{t+1} \leq \max(Y_t + R_t, X_t)$, and Y_0 is identically 0. Assume that Y_t and R_t are independent. If $\widehat{X}_t(z_0) \leq M$ for all $t \geq 0$, then $\widehat{Y}_t(z_0) \leq M/(1-\delta)$ for all $t \geq 0$, where $\delta = e^{r(z_0 - 1)}/z_0$.*

Proof By Proposition 3.4, $z\widehat{R}_t(z) \leq e^{r(z-1)}$. The two functions $e^{r(z-1)}$ and z coincide when $z = 1$. On the other hand, the derivative of $e^{r(z-1)}$ at $z = 1$ is $r < 1$. Since the derivative of the function z is equal to 1, it follows that $e^{r(z-1)} < z$, for some $z > 1$. In fact, an elementary calculation shows that z_0 satisfies the above inequality. Thus $R_t(z_0) \leq \delta < 1$. Since $z_0 > 1$, we have ${z_0}^{Y_{t+1}} \leq {z_0}^{Y_t+R_t} + {z_0}^{X_t}$. Since Y_t and R_t are independent, the generating function of their sum is equal to the product of their generating functions. Thus, by taking the expectations of the two sides of the above equation, it follows that

$$\begin{aligned} \widehat{Y}_{t+1}(z_0) &\leq \widehat{Y}_t(z_0)\widehat{R}_t(z_0) + \widehat{X}_t(z_0) \\ &\leq \delta\widehat{Y}_t(z_0) + M. \end{aligned}$$

Since $\widehat{Y}_0(z_0) = 1 \leq M$, it follows by induction on t that $\widehat{Y}_t(z_0) \leq M/(1-\delta)$. ■

4 The one dimensional array

We first show that the linear array is stable if the bisection condition is met. Then we establish bounds on the tails of the delay and the queue-sizes. Nodes are labeled from 1 to N. Throughout this section, we only consider packets whose destination is to their right.

4.1 Stability

Let $A_{t,k}$ be the set of packets generated at step t in the first k processors and whose destination is in the last $N-k$ processors. Since at most one packet per step can traverse the edge $(k, k+1)$, a necessary condition for stability (assuming $N > 2$) is that $\mathrm{E}[|A_{t,k}|] < 1$ for all k. Since $\mathrm{E}[|A_{t,k}|] = \frac{4\rho}{N} k \frac{N-k}{N}$, this is equivalent (when N is even) to $\rho < 1$, a condition that we assume throughout the paper. We show below that this condition is sufficient.

Theorem 4.1 *If the arrival rate of packets in a linear N-array is at most 99% of network capacity, then at any particular step, the probability that the maximum queue-size exceeds q_0 is at most $e^{-\alpha q_0}$, where $\alpha > 0$ is a function of N and of the arrival rate.*

Proof Let $P_{t,k}$ be the set of packets located at step t in the first k processors and whose destination is in the last $N-k$ processors, and let $U_{t,k}$ be the packets in $P_{t,k}$ located at processor k at step t. The proof is based on the following proposition.

Proposition 4.2 *If $|U_{t,k}| \neq 0$, then $|P_{t+1,k}| = |P_{t,k}| + |A_{t+1,k}| - 1$. If $|U_{t,k}| = 0$, then $|P_{t+1,k}| \leq |P_{t,k-1}| + |A_{t+1,k}|$, where $P_{t,0} = \emptyset$.*

Proof If $|U_{t,k}| \neq 0$, then there is at least one packet at node k at step t. Since the protocol is greedy, one of these packets will move to node $k+1$ during step t, and therefore will not belong to $P_{t+1,k}$. But, since $P_{t+1,k}$ is contained is the disjoint union of the sets $P_{t,k}$ and $A_{t+1,k}$, it follows that $|P_{t+1,k}| = |P_{t,k}| + |A_{t+1,k}| - 1$.

If $|U_{t,k}| = 0$, then $P_{t+1,k}$ is equal to the disjoint union of the sets $P_{t,k}$ and $A_{t+1,k}$, hence $|P_{t+1,k}| = |P_{t,k}| + |A_{t+1,k}|$. On the other hand, $P_{t,k} \subset P_{t,k-1}$ in this case since the queue at node k is empty, and so $|P_{t,k}| \leq |P_{t,k-1}|$. Thus $|P_{t+1,k}| \leq |P_{t,k-1}| + |A_{t+1,k}|$, as desired. ■

Since $\mathrm{E}[|A_{t+1,k}|] < 1$, Proposition 4.2 shows that for fixed k, the sequence $|P_{t,k}|$ has a negative drift when $|U_{t,k}| \neq 0$ (the drift here depends on t.) Hajek's result does not seem to apply here, however. (It applies in the trivial case $k = 1$ with $a = 0$ since, when $|U_{t,1}| = 0$, $|P_{t,1}| = 0$. For the same reason, it would have applied for a given k if, somehow, we could show that $|P_{t,k-1}|$ is uniformly bounded in t. But this is generally not the case.) Our strategy is to show by induction on k that the tail of the distribution of $P_{t,k}$ is exponentially decreasing, with a parameter depending only on N and on the arrival rate, not on t. We do so by using Proposition 3.1 and exhibiting a real $z_0 > 1$ such that $\widehat{P}_{t,k}(z_0)$ is bounded by a number independent of t. The intuition behind the proof is that, when $|U_{t,k}| \neq 0$, the sequence $|P_{t,k}|$ behaves well since it has a negative drift: $\mathrm{E}[|P_{t+1,k}| - |P_{t,k}| \mid |U_{t,k}| \neq 0] < 0$. If $|U_{t,k}| = 0$, the sequence $|P_{t,k}|$ behaves well also since $|P_{t+1,k}| \leq |P_{t,k-1}| + |A_{t+1,k}|$ in this case, and we know by the induction hypothesis that the distribution of $|P_{t,k-1}|$ has an exponentially decreasing tail.

We now prove by induction on k that, for all integers t,

$$\widehat{P}_{t,k}(z_0) \leq \gamma^k, \tag{1}$$

where $z_0 = 2 - \rho$ and $\gamma = \gamma(\rho)$ is a constant to be determined later. Eq. 1 holds for $k = 0$ since $P_{t,0} = \emptyset$ by definition. Assume now that it holds for $k-1$ and all $t \geq 0$. We show that the conditions of Lemma 3.5 hold for $Y_t = |P_{t,k}|$, $X_t = |P_{t,k-1}| + |A_{t+1,k}|$, $R_t = |A_{t+1,k}| - 1$ and $r = \rho$. Indeed,

$$\mathrm{E}[A_{t+1,k}] = \frac{4\rho}{N} k \frac{N-k}{N} \leq \rho.$$

The condition $Y_{t+1} \leq \max(Y_t + R_t, X_t)$ follows from Proposition 4.2. Finally, Y_t and R_t are clearly independent.

Let $\gamma = e^{\rho(z_0-1)}/(1-\delta)$. Since $\widehat{P}_{t,k-1}(z_0) \leq \gamma^{k-1}$ by the induction hypothesis and since $\widehat{A}_{t+1,k}(z_0) \leq e^{\rho(z_0-1)}$ by Proposition 3.4, $\widehat{X}_t(z_0) \leq \gamma^{k-1} e^{\rho(z_0-1)}$. Thus Lemma 3.5 yields

$$\begin{aligned} \widehat{P}_{t,k}(z_0) &\leq \gamma^{k-1} e^{\rho(z_0-1)}/(1-\delta) \\ &= \gamma^k, \end{aligned}$$

as desired.

Hence, the generating function of the maximum queue-size, when evaluated at z_0, is upper bounded by $N\gamma^N$. This is because, the generating function of the maximum is upper bounded by the sum of the generating functions. We conclude the proof of Theorem 4.1 using Proposition 3.1. ■

Corollary 4.3 *For an arbitrary greedy time-independent protocol, the system is ergodic. In particular, all the queues will empty with probability 1 and in a finite expected number of steps.*

Proof This follows easily from Theorem 4.1 and general properties of Markov chains (see, for example, [8, pages 29–30].) ■

4.2 Bounds on the queue-sizes and the delays

The bound on the tails of the queue-sizes that follows from the proof of Theorem 4.1 depends on N. This is because Lemma 3.5 has been applied N times, and each time the upper bound on the generating function (evaluated at z_0) is multiplied by a constant factor. In order to get bounds independent of N, we will introduce the concept of *holdover packet* that will enable us to divide the set of packets into a constant number of subsets, and analyse one subset at a time. We will then apply Lemma 3.5 a number of times equal to the number of these subsets, thus getting bounds on the tails of the queue-sizes independent of N.

We say that a packet at a given node and a given step t is holdover if and only if it has not moved at step $t-1$. This means that either it was at the same node at step $t-1$ or it was generated at step t. Let M_s denote the set of holdover packets at node k and step $k+s$ for some $1 \leq k \leq N$ (see Fig. 1). Note that a packet cannot be holdover at $(k, k+s)$ and $(k', k'+s)$ for $k \neq k'$ since, if $k > k'$ for example, this would imply that the packet has been moving from k' to k at steps $k'+s, k'+s+1, \ldots, k+s-1$, contradicting the fact that it was holdover at step $k+s$. Therefore each packet in M_s is counted only once. If a packet is at node k at step $k+s$, we say that it has *relative* time s. We observe that any packet at relative time s belongs to M_s. Indeed, let k' be the smallest number such that the packet was at k' at step $k'+s$. Then the packet is holdover at $(k', k'+s)$. As a consequence, a packet holdover at node k at relative time $s+1$ either belongs to M_s or has been created at step $k+s+1$. To see this, we note that if the packet has not been generated at step $k+s+1$, then it was at node k at step $k+s$, and so it belongs to M_s. Finally, we observe that the relative time of a packet cannot decrease in time. This shows that the set of states of the queues at nodes $(k, k+s)$ is independent of the set of packets created at relative time $s+1$.

The intuition for the introduction of the concept of holdover packets is that, while regular packets travel at most one unit of distance in one step, holdover packets can travel long distances in one step, in some sense. Moreover, while the expected number of regular packets at any given time is linear in N, the expected number of holdover packets at any given relative time is constant, as shown in the following lemma. Thus we expect to get tighter bounds by analysing the evolution of holdover packets rather than that of regular packets.

Lemma 4.4 *There exists a constant $c' = c'(\rho)$ such that, for any step s, the probability that there are more than U packets in M_s is at most $e^{-c'U}$.*

Proof For simplicity, assume that $\kappa = \frac{8}{1-\rho}$ divides N. We divide the array into κ consecutive intervals of same size. For $1 \leq h \leq \kappa$, let $M_{s,h}$ denote the set of packets holdover at relative time s at some node in the first h intervals and whose destination is in the last $\kappa - h + 1$ intervals (note that the h-th interval belongs to both sets of intervals. This fact is crucial to our

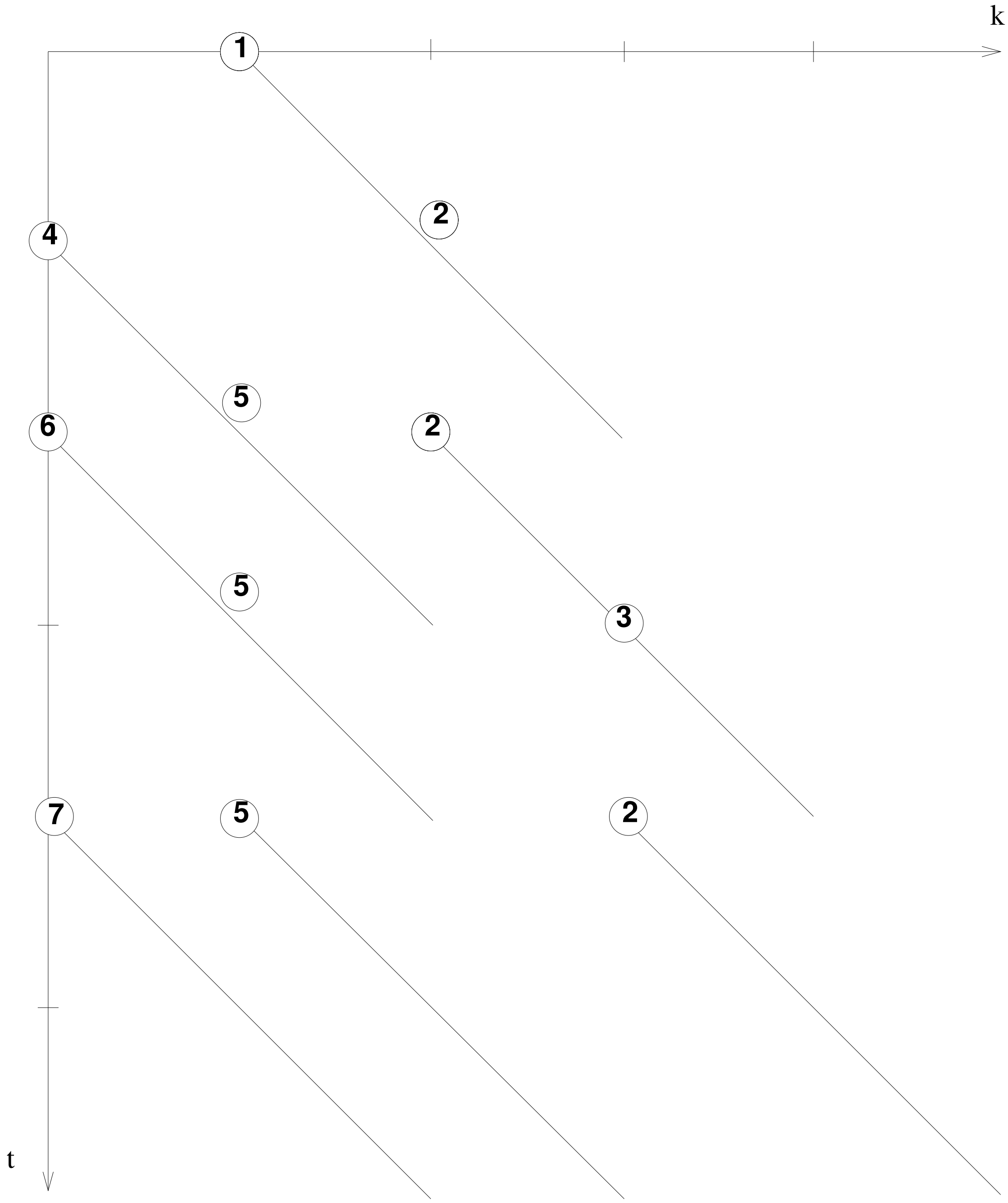

Figure 1: The farthest-last protocol on a linear array. Diagonal lines represent the relative time. Packets have been labeled in increasing order according to the relative time of their birth. Holdover packets at a given time and location are circled. Thus, $M_{-1} = \{1, 2\}$, $M_0 = \{2, 3\}$, $M_1 = \{4, 5, 2\}$, $M_2 = \{6, 5\}$, $M_3 = \{5\}$, $M_4 = \{7\}$.

analysis. In particular, the following proposition would not be valid if the sets of intervals were not overlapping.)

Proposition 4.5 *If there is a packet at relative time s in the h-th interval, then $|M_{s+1,h}| \leq |M_{s,h}| + |B_{s+1,h}| - 1$, where $B_{s+1,h}$ is the set of packets that were generated at relative time $s+1$ in the first h intervals and whose destination is to the right of their origin and in the last $\kappa - h + 1$ intervals.*

Proof By the above discussion, the set $M_{s+1,h}$ is contained in the disjoint union of the sets $M_{s,h}$ and $B_{s+1,h}$. To prove the proposition, we exhibit an element of $M_{s,h}$ not belonging to $M_{s+1,h}$. Consider the largest k in the h-th interval such that the queue at node k at step $k+s$ is non-empty. It is clear that any packet in this queue was holdover at some node in the first h intervals at relative time s, and thus belongs to $M_{s,h}$. Since the algorithm is greedy, a packet will leave node k at step $k+s$. By the definition of k, this means that either the destination of the packet is $k+1$ or that k is the last node in the h-th interval. In both cases, the packet will not belong to $M_{s+1,h}$. ∎

It follows from Proposition 4.5 that the $B_{s,h}$'s satisfy an inequality analogue to those in Proposition 4.2.

Proposition 4.6 *For $1 \leq h \leq \kappa$ and $s \geq -N$,*

$$|M_{s+1,h}| \leq \max(|M_{s,h}| + |B_{s+1,h}| - 1, |M_{s,h-1}| + |B_{s+1,h}|), \tag{2}$$

where $M_{s,0} = \emptyset$.

Proof We distinguish two cases:

1. There is no holdover packet at relative time s in the h-th interval. Then any element of $M_{s,h}$ is holdover at relative time s in the first $h-1$ intervals, and so $M_{s,h} \subset M_{s,h-1}$. Therefore, $|M_{s,h}| \leq |M_{s,h-1}|$. Eq. 2 follows since $M_{s+1,h}$ is contained in the union of the sets $M_{s,h}$ and $B_{s+1,h}$, and so $|M_{s+1,h}| \leq |M_{s,h}| + |B_{s+1,h}|$.
2. There is a holdover packet at relative time s in the h-th interval. In this case, Eq. 2 follows from Proposition 4.5.

∎

For each node in the first h intervals, the probability that a packet is generated at this node at a given step and that its destination is in the last $\kappa - h + 1$ intervals is at most $\frac{4\rho}{N}\frac{\kappa-h+1}{\kappa}$. Therefore,

$$\mathrm{E}[B_{s,h}] \leq \frac{4\rho}{N}\frac{\kappa - h + 1}{\kappa}\frac{hN}{\kappa} < \rho' = \frac{1+\rho}{2}. \tag{3}$$

Let $z_0' = 2 - \rho'$, $\delta' = e^{\rho'(z_0'-1)}/z_0'$ and $\gamma' = e^{\rho'(z_0'-1)}/(1-\delta')$. By Eqs. 2 and 3, we can apply Lemma 3.5 with $Y_s = |M_{s,h}|$, $X_s = |M_{s,h-1}| + |B_{s+1,h}|$, $r = \rho'$ and $R_s = |B_{s+1,h}| - 1$. (A technical detail: in this case, s can be as small as $-N$, and Y_0 is not identically 0, but Y_{-N} is. This statement of Lemma 3.5 can be modified in an obvious manner to handle this case, without affecting the conclusion. We will ignore this issue in the rest of the paper.) It follows from an argument similar to the one in the proof of Theorem 4.1 that

$$\widehat{M}_{s,h}(z_0') \leq \gamma'^h, \tag{4}$$

for $1 \leq h \leq \kappa$. Hence the tail of the distribution of $M_{s,h}$ is exponentially decreasing, with a parameter depending only on ρ, and so is the tail of M_s. This concludes the proof of Lemma 4.4. ■

Theorem 4.7 *If the arrival rate of packets in a linear N-array is at most 99% of network capacity, then the probability that any particular packet is delayed Δ steps is at most $e^{-c\Delta}$ for some constant c that does not depend on N or on the time at which the packet was generated or on the origin or destination of the packet or on the protocol. Moreover, in any window of T steps, the maximum delay incurred by any packet is $O(\log T + \log N)$ with high probability, and the maximum observed queue-size is $O(1 + \frac{\log T}{\log N})$ with high probability.*

Proof We first show that the tail of the distribution of the delay of a packet in the window in which it is generated is exponentially decreasing. Consider a packet generated in the h-th window at relative time s, and let W_i, $1 \leq i \leq \kappa - h + 1$, be the number of steps it is delayed in the $h + i - 1$-st window. If $W_1 \geq \Delta_1$, then a packet has to leave each of the sets $M_{s,h}, M_{s+1,h}, \ldots, M_{s+\Delta_1-1,h}$ by the proof of Proposition 4.5. In other words, $|M_{s+\Delta_1,h}| \leq |M_{s,h}| + |B_{s+1,h}| + |B_{s+2,h}| + \ldots |B_{s+\Delta_1,h}| - \Delta_1$, as can be easily seen by induction on Δ_1. This implies that either $|M_{s,h}| \geq \frac{1-\rho'}{2}\Delta_1$ or $|B_{s+1,h}| + |B_{s+2,h}| + \ldots |B_{s+\Delta_1,h}| \geq \frac{1+\rho'}{2}\Delta_1$. By Lemma 4.4, the probability of the first event is $e^{-\Omega(\Delta_1)}$. On the other hand, $\mathrm{E}[|B_{s,h}|] \leq \rho'$ and so, by Lemma 3.3, the probability of the second event is at most $\left(\frac{e^{\beta-1}}{\beta^\beta}\right)^{\rho'\Delta_1}$, where $\beta = \frac{1+\rho'}{2\rho'} > 1$, and it is also exponentially decreasing in Δ_1.

Next, we have to show that the tail of the distribution of the delay in the subsequent windows is exponentially decreasing. This is not as obvious as it first appears because, when the packet arrives to a subsequent window, the distribution of the packets in the system is not necessarily the same as the ergodic distribution. More precisely, $|M_{s+W_1,h+1}|$ and $|M_{s,h+1}|$ do not necessarily have the same distribution, even as s goes to infinity. However, since the delay of the packet in the h-th window is exponentially decreasing and since the expected number of packets that are created in the system at each relative time is constant (less than 4), the tail of the distribution of $|M_{s+W_1}|$ is exponentially decreasing. Indeed, if $|M_{s+W_1}| \geq \Delta_2$, then either $|M_s| \geq \Delta_2/2$ or at least $\Delta_2/2$ packets have been created at relative time $s+1, s+2, \ldots, s+W_1$. The probability of the first event is $e^{-\Omega(\Delta_2)}$, by Lemma 4.4. For the second event to happen, either $W_1 \geq \Delta_2/10$, or at least $\Delta_2/2$ packets have been created at relative time $s+1, s+2, \ldots, s+\Delta_2/10$. The probability of the first event is $e^{-\Omega(\Delta_2)}$ by the argument above. The probability of the second event is $e^{-\Omega(\Delta_2)}$ by Lemma 3.3. Thus $\Pr[|M_{s+W_1}| \geq \Delta_2] = e^{-\Omega(\Delta_2)}$. Hence we can use the same argument as in the previous paragraph to prove that the tail of W_2 is exponentially decreasing. A similar argument applies to the remaining windows. This shows that the total delay is exponentially decreasing. Indeed, the total delay exceeds Δ only if W_i exceeds Δ/κ for some $1 \leq i \leq \kappa - h + 1$.

Next, we bound the probability that the queue-size at a given step and a given node exceeds q. This can happen only in the following two cases:

1. The queue at this node was non-empty throughout the last $q \log N$ steps. If the node is in the h-th interval, this implies that there was a packet in the h-th interval for $q \log n$ consecutive "relative steps". By Proposition 4.5 and in the same way we proved that the tail of W_1 is exponentially decreasing, the probability of this event is $e^{-\Omega(q \log N)}$.
2. The queue has been empty at least once during the last $q \log N$ steps. Since the queue-size can increase by at most one when the queue is non-empty and only if a packet is generated at the corresponding node, this implies that at least q packets have been generated at this node in the last $q \log N$ steps. By Lemma 3.3, the probability of this event is at most $e^{(\beta-1-\beta\log\beta)4q\log N/N}$, where $\beta = \frac{N}{4\log N}$. Again, this probability is $e^{-\Omega(q \log N)}$.

Hence the probability that the queue-size at a given step and a given node exceeds q is $e^{-\Omega(q \log N)}$.

As a consequence, in any window of T steps, the maximum delay incurred by any packet is $O(\log T + \log N)$ and the maximum observed queue-size is $O(1 + \frac{\log T}{\log N})$ with probability $1 - O(\frac{1}{TN})$. ■

5 Higher dimensional arrays

In this section, we analyse the behavior of greedy routing protocols on higher dimensional arrays. The class of protocols we consider has been defined in Section 2. First, we analyse the case of dynamic routing, i.e. when a packet is generated at each node and at each step with probability $4\rho/N$. Then we extend the results to the static case where each node contains one packet in the beginning and no packets are generated at a later time.

Theorem 5.1 *If the arrival rate of packets in an* $\underbrace{N \times N \times \cdots \times N}_{k}$ *array of fixed dimension is at most* 99% *of network capacity, then the probability that any particular packet is delayed* Δ *steps is at most* $e^{-c\Delta}$ *for some constant* c *that does not depend on* N *or on the time at which the packet was generated or on the origin or destination of the packet or on the protocol. Moreover, in any window of* T *steps, the maximum delay incurred by any packet is* $O(\log T + \log N)$ *with high probability, and the maximum observed queue-size is* $O(1 + \frac{\log T}{\log N})$ *with high probability.*

Sketch of Proof For simplicity, consider the case of a two dimensional array. The stability can be shown by considering the set of packets $Q_{t,k,l}$ in the entire network at time t that eventually want to traverse the directed edge $((k,l),(k+1,l))$. Note that the expected size of the set $A_{t,k,l}$ of packets that are generated at step t and that eventually want to traverse the directed edge

$((k,l),(k+1,l))$ is $k(N-k)4\rho/N^2 \le \rho < 1$. We fix l and show by induction on k that the tail of $Q_{t,k,l}$ is exponentially decreasing (with a decay parameter depending on N.) We use a proof similar to the one dimensional case. Indeed, $Q_{t,k,l}$ satisfies recurrence relations similar to those in Proposition 4.2, except that packets in row k must be taken into account. The contribution of these packets can be bounded using the stability of row k. More precisely, if there is a packet at node (k,l) at time t that wants to traverse the directed edge $((k,l),(k+1,l))$, then $|Q_{t+1,k,l}| = |Q_{t,k,l}| + |A_{t+1,k,l}| - 1$. Otherwise, $|Q_{t+1,k,l}| \le |Q_{t,k-1,l}| + |A_{t+1,k,l}| + |S_{t,k,l}|$, where $S_{t,k,l}$ is the set of packets in row k (at a horizontal queue) at time t that eventually want to traverse the directed edge $((k,l),(k+1,l))$. Thus,

$$|Q_{t+1,k,l}| \le \max(|Q_{t,k,l}| + |A_{t+1,k,l}| - 1, |Q_{t,k-1,l}| + |A_{t+1,k,l}| + |S_{t,k,l}|).$$

As in the proof of Theorem 4.1, we recursively apply Lemma 3.5 with $Y_t = |Q_{t,k,l}|$, $X_t = |Q_{t,k-1}| + |A_{t+1,k,l}| + |S_{t,k,l}|$, $r = \rho$ and $R_t = |A_{t+1,k,l}| - 1$. We need, however, to take into account the term $|S_{t,k,l}|$ when we bound $\widehat{X}_t(z_0)$ in our induction step. Since the generating function of X_t is equal to the product of the generating functions of its components, all we need to check is that $\widehat{S}_{t,k,l}(z_0)$ is uniformly bounded in t. But $|S_{t,k,l}|$ itself is the sum of two independent terms, one analogue to $|P_{t,l-1}|$ and the other one to $|P_{t,N-l}|$. By the proof of Theorem 4.1, the generating functions of these terms, when evaluated at z_0, are uniformly bounded in t. Hence $\widehat{S}_{t,k,l}(z_0)$ is uniformly bounded in t, as desired.

The bounds on the delays and on the queue-sizes are also derived in a similar manner to the one dimensional case, by changing the definitions in a suitable manner. We will consider only the packets whose destination row has no smaller index than their origin row. The relative time of a packet at time t is the difference between t and the distance between the node $(1,1)$ and the packet current location. Divide the array into κ horizontal bands of same size. Let $V_{s,h,l}$ be the set of packets holdover in the first h horizontal bands at relative time s and whose destination is in the last $\kappa - h + 1$ horizontal bands and in column l. As in Lemma 4.4, we show by induction on h that the tail of $V_{s,h,l}$ is exponentially bounded. Let $B_{s,h,l}$ be the set of packets generated at relative time s in the first h bands and whose destination is in the last $\kappa - h + 1$ bands. Then

$$|V_{s+1,h,l}| \le \max(|V_{s,h,l}| + |B_{s+1,h,l}| - 1, |V_{s,h-1,l}| + |B_{s+1,h,l}| + |E_{s,h,l}|),$$

where $E_{s,h,l}$ is the set of packets holdover in band h (at a horizontal queue) at relative time s and whose destination is in row l. We can now apply the same method as in the previous paragraph. However, in order to get a decay parameter c independent of N, we need to show an exponential bound on the tail of $|E_{s,h,l}|$. Such a bound can be shown by using the fact that the tails of the delays in each row are exponentially decreasing. More precisely, for $|E_{s,h,l}|$ to exceed Δ, either one of the packets in $E_{s,h,l}$ was generated before relative time $s - \Delta/5 + 1$, or more than Δ packets were generated between relative time $s - \Delta/5 + 1$ and s. Since the expected number of packets that are generated at a given relative time and whose destination is in column l is at most 4, the probability of the second event is $e^{-\Omega(\Delta)}$ by Lemma 3.3. We now bound the first event. If a packet generated at relative time $s - \delta$ is holdover at relative time s

at some horizontal queue, it must have been delayed δ steps in the row where it was generated. By Theorem 4.7, the probability of this event is at most $e^{-c\delta}$, where $c = c(\rho)$. The probability that some packet is generated at relative time $s - \delta$ and is holdover at relative time s at an horizontal queue and that its destination is in column l is thus at most $N^2 \frac{4\rho}{N} e^{-c\delta}/N \leq 4e^{-c\delta}$. Hence the probability of the first event is at most

$$4 \sum_{\delta \geq \Delta} e^{-c\delta} = O(e^{-c\Delta}).$$

By the proof of Proposition 3.2, we conclude that $\widehat{E}_{s,h,l}(e^{c/2}) = O(1)$. The rest of the proof is similar to Theorem 4.7. ■

Theorem 5.2 *In the static case, where each node of a* $\underbrace{N \times N \times \cdots \times N}_{k}$ *array of fixed dimension node contains one packet in the beginning and the destinations are uniformly distributed among all nodes of the array, the probability that any particular packet is delayed* Δ *steps is at most* $e^{-c\Delta}$ *for some constant* c *that does not depend on* N *or on the time at which the packet was generated or on the origin or destination of the packet or on the protocol. Moreover, the maximum delay incurred by any packet is* $O(\log N)$ *with high probability, and the maximum observed queue-size is* $O(1)$ *with high probability.*

Sketch of Proof First, consider the case of a two dimensional array. No delays are incurred in the rows. At each node of a given column and at each step, the expected number of packets that arrive to this node and whose destination is in this column is at most $2/N$. For a given column, these arrivals are independent over time. Thus any column can be analysed in a way similar to the one dimensional array (with the minor difference that two packets may arrive at the same time; this does not affect the proof since the variables $A_{t,k}$ are still the sum of independent Bernoulli random variables.) This implies that the delay distribution has an exponentially bounded tail. In particular, the maximum delay is $O(\log N)$ with high probability. The bound on the maximum queue-size follows by replacing T with $3N$ in Theorem 4.7. The case of higher dimensional arrays can be similarly reduced to the proof of Theorem 5.1. ■

6 Farthest-first protocol on the linear array and the ring

In this section, we use generating functions to calculate exactly the ergodic expected delay on the one dimensional array and (as N goes to infinity) on the ring. We also derive a simple expression for the expected delay on an odd ring under Poisson arrivals.

Lemma 6.1 *Assume that on each node on an* N*-array, a packet is generated at each step in* $[1, k]$ *with some probability independent of time and that its destination is to the right of* k*. Assume that packets arrivals are independent over time, and that* $\mathrm{E}[|A|] = \widehat{A}'(1) < 1$*, where*

A is the set of packets that are generated at each step in the first k nodes. Then the system consisting of the first k nodes is ergodic and the steady-state expected number of delayed packets in the system is

$$\frac{\widehat{A}''(1)}{2(1-\mathrm{E}[|A|])} = \frac{\mathrm{Var}[|A|]}{2(1-\mathrm{E}[|A|])} - \frac{\mathrm{E}[|A|]}{2},$$

where $\widehat{A}''(1)$ is the second derivative of $\widehat{A}(z)$ at $z = 1$.

Proof Throughout this proof, only packets in the first k nodes will be counted. Let M_s be the set of holdover packets at relative time s, and A_s the set of packets generated at relative time s. By the same proof as Proposition 4.5, we know that

$$|M_{s+1}| = \begin{cases} |M_s| + |A_{s+1}| - 1 & \text{if } M_s \neq \emptyset, \\ |A_{s+1}| & \text{otherwise.} \end{cases}$$

Therefore,

$$\widehat{M}_{s+1}(z) = \mathrm{E}[z^{|M_s|+|A_{s+1}|-1}\chi_{M_s\neq\emptyset}] + \mathrm{E}[z^{|A_{s+1}|}\chi_{M_s=\emptyset}].$$

But the first term is equal to $\mathrm{E}[z^{|M_s|-1}\chi_{M_s\neq\emptyset}]\widehat{A}(z)$ by the independence hypothesis, where $\widehat{A} = \widehat{A}_{s+1}$. Since $\widehat{M}_s(z) = \mathrm{E}[z^{|M_s|}\chi_{M_s\neq\emptyset}] + \Pr[M_s = \emptyset]$, we have

$$\widehat{M}_{s+1}(z) = \widehat{A}(z)\frac{\widehat{M}_s(z) - \Pr[M_s = \emptyset]}{z} + \widehat{A}(z)\Pr[M_s = \emptyset].$$

When s goes to infinity, $\widehat{M}_s$ converges to $\widehat{M}$, where M is the steady-state distribution, and $\Pr[M_s \neq \emptyset]$ goes to $\mathrm{E}[|A|]$ by standard queuing theory, and so

$$\widehat{M}(z) = \widehat{A}(z)\frac{\widehat{M}(z) - 1 + \mathrm{E}[|A|]}{z} + (1 - \mathrm{E}[|A|])\widehat{A}(z).$$

Hence

$$\widehat{M}(z) = (1 - \mathrm{E}[|A|])\frac{\widehat{A}(z)}{1 - \frac{1-\widehat{A}(z)}{1-z}}.$$

Therefore, $\mathrm{E}[\text{holdover packets}] = \widehat{M}'(1) = \mathrm{E}[|A|] + \frac{\widehat{A}''(1)}{2(1-\mathrm{E}[|A|])}$. Since the set of holdover packets at relative time s is the disjoint union of the set of delayed packets, and of the set of packets that were generated at relative time s, and since the expected number of packets generated at relative time s is $\mathrm{E}[|A|]$, we conclude that $\mathrm{E}[\text{delayed packets}] = \frac{\widehat{A}''(1)}{2(1-\mathrm{E}[|A|])}$. The second equality follows from $\widehat{A}''(1) = \mathrm{Var}[|A|] + \mathrm{E}[|A|]^2 - \mathrm{E}[|A|]$. ■

Corollary 6.2 *The steady-state expected number of packets delayed at node k is no more than $\frac{p}{2(1-\mathrm{E}[|A|])^2} - \frac{p}{2}$, where p is the probability that a packet is generated at node k.*

Proof Let A_* (resp. a) be the set of packets generated in the first $k-1$ nodes (resp. node k). The expected number of packets delayed at node k is equal to

$$\frac{\mathrm{Var}[|A|]}{2(1-\mathrm{E}[|A|])} - \frac{\mathrm{Var}[|A_*|]}{2(1-\mathrm{E}[|A_*|])} - (\frac{\mathrm{E}[|A|]}{2} - \frac{\mathrm{E}[|A_*|]}{2}).$$

The second term is equal to $p/2$. Since the variance of the sum of independent random variables is equal to the sum of their variances and since the variance of a Bernoulli random variable is at most its expectation, we can rewrite the first term as

$$\begin{aligned}
&\frac{\mathrm{Var}[|A|]}{2(1-\mathrm{E}[|A|])} - \frac{\mathrm{Var}[|A|]}{2(1-\mathrm{E}[|A_*|])} + \frac{\mathrm{Var}[|A|]-\mathrm{Var}[|A_*|]}{2(1-\mathrm{E}[|A_*|])} \\
&= \frac{p\mathrm{Var}[|A|]}{2(1-\mathrm{E}[|A|])(1-\mathrm{E}[|A_*|])} + \frac{\mathrm{Var}[|a|]}{2(1-\mathrm{E}[|A_*|])} \\
&\leq \frac{p\mathrm{E}[|A|]}{2(1-\mathrm{E}[|A|])^2} + \frac{p}{2(1-\mathrm{E}[|A|])} \\
&= \frac{p}{2(1-\mathrm{E}[|A|])^2}.
\end{aligned}$$

■

Theorem 6.3 *For the farthest-first protocol on an N-array, the steady-state expected delay tends to $-\frac{1}{2}+\frac{1}{4(1-\rho)}(\frac{1}{\sqrt{\rho(1-\rho)}}\arcsin(\sqrt{\rho})+1)$ as N goes to infinity. On the ring, the steady-state expected delay is upper bounded by and tends to $\frac{1}{2(1-\rho)^2}-\frac{1}{2}$ as N goes to infinity. For Poisson arrivals on a ring with $N=2l+1$ processors where a packet's destination is uniformly distributed among the* remaining *processors, the steady-state expected delay is*

$$\frac{g(\lambda(l+1)/4)-g(\lambda(l-1)/4)}{\lambda},$$

where λ is the arrival rate at a given node, and $g(x)=x^2/(1-x)$.

Proof Under the farthest-first protocol, packets in the first k nodes whose destination is to the right of a node k do not interact with the packets whose destination is in the first k nodes. So we can apply Lemma 6.1 to the first set of packets with $p=\frac{4\rho}{N}\frac{N-k}{N}$ and $\widehat{A}(z)=(1-p+pz)^k$ to calculate the expected number of delayed packets in the first k nodes whose destination is to the right of k. Similarly, we can calculate the expected number of delayed packets in the first $k-1$ nodes whose destination is to the right of k. The difference of these two quantities is the expected number w_k of delayed packets at node k:

$$w_k = W_{k,p}-W_{k-1,p} = \frac{(2-kp)}{2(1-kp)(1-kp+p)}(k-1)p^2,$$

where $p=\frac{4\rho}{N}\frac{N-k}{N}$. By Little's law, the expected delay per packet is $\mathrm{E}[\mathrm{Delay}]=\frac{1}{2\rho}\sum_{k=1}^N w_k$. An easy calculation shows that $w_k=2\rho\frac{f(k/N)}{N}+O(\frac{1}{N^2})$, where

$$f(x)=8\rho x(1-x)^2\frac{1-2\rho x(1-x)}{(1-4\rho x(1-x))^2}.$$

Hence, $\mathrm{E}[\mathrm{Delay}]=\int_0^1 f(x)\,dx+O(\frac{1}{N})$. But

$$8\rho\int_0^1 x(1-x)^2\frac{1-2\rho x(1-x)}{(1-4\rho x(1-x))^2}\,dx \quad = \quad 4\rho\int_0^1 x(1-x)\frac{1-2\rho x(1-x)}{(1-4\rho x(1-x))^2}\,dx$$

$$
\begin{aligned}
&= \frac{\rho}{4}\int_{-1}^{1}(1-y^2)\frac{2-\rho(1-y^2)}{(1-\rho(1-y^2))^2}\,dy \\
&= \frac{1}{2}\int_{0}^{1}(-1+\frac{1}{(1-\rho(1-y^2))^2})\,dy \\
&= -\frac{1}{2}+\frac{1}{4(1-\rho)}(\frac{1}{\sqrt{\rho(1-\rho)}}\arctan\sqrt{\frac{\rho}{1-\rho}}+1).
\end{aligned}
$$

The second equality follows by setting $y = 2x - 1$, and the last one follows from

$$2(1-\rho)\int\frac{dy}{(1-\rho(1-y^2))^2} = \frac{1}{2(1-\rho)}\frac{1}{\sqrt{\rho(1-\rho)}}\arctan\left(\sqrt{\frac{\rho}{1-\rho}}y\right) + \frac{y}{1-\rho+\rho y^2}.$$

Finally, we have $\arctan\left(\sqrt{\frac{\rho}{1-\rho}}\right) = \arcsin\sqrt{\rho}$.

In the case of the ring, we assume that the arrival rate is $\frac{8\rho}{N}$, with $\rho < 1$, and that packets are routed to their destinations along a shortest path. As in the case of the one dimensional array, we can calculate the exact expression of the expected delay. To obtain an upper bound on the delay, we apply Corollary 6.2 by considering only packets going clockwise. We have $p < \frac{4\rho}{N}$ and $\mathrm{E}[|A|] \leq \rho$, and so the expected number of delayed packets per node is at most $\frac{4\rho}{N}(\frac{1}{2(1-\rho)^2} - \frac{1}{2})$. Thus, the upper bound on the expected delay follows from Little's Law. Moreover, it is not hard to see that the bound $\frac{1}{2(1-\rho)^2} - \frac{1}{2}$ is equal to the limit of the expected delay on the ring as N goes to infinity.

The case of the ring under Poisson arrivals follows from Lemma 6.1 and similar calculations. Moreover, if λN tends to 8ρ as N goes to infinity, the ergodic expected delay tends to $g'(\rho)/2 = \frac{1}{2(1-\rho)^2} - \frac{1}{2}$. ∎

7 Cut-through routing

In cut-through routing, each packet consists of b flits and moves through the network like a worm, occupying at most b consecutive nodes.

Theorem 7.1 *If the arrival rate of packets in an $N \times N$ array is at most 99% of network capacity, and if packets have length b, then the probability that any particular packet is delayed $b\Delta$ steps is at most $e^{-c\Delta}$ for some constant c that does not depend on N or on the time at which the packet was generated or on the origin and destination of the packet or on the protocol. Moreover, in any window of T steps, the maximum delay incurred by any packet is $O(b\log T + b\log N)$ with high probability, and the maximum observed queue-size is $O(b + b\frac{\log T}{\log N})$ with high probability.*

Sketch of Proof First note that a necessary condition for stability is that the probability that a packet is generated is $p = 4\rho/(bN)$, with $\rho < 1$. An arbitrary greedy protocol as defined in Section 2 can be tranformed into a protocol for cut-through routing as follows: when two heads of packets wish to traverse the same edge at the same time, ties are broken according to the original protocol. Otherwise, when two flits of different packets wish to traverse the same edge at the same time, priority is given to the packet whose head has already traversed this edge. Thus cut-through routing can be implemented by a greedy protocol. We can use the same methods as in Section 5 to prove the theorem. The main difference in the proof is that we analyse the evolution of the system every b steps. For example, in the one dimensional case, $M_{s,h}$ will denote the set of packets whose head is holdover at relative time bs in the first h intervals, and whose destination is to the right of their origin and in the last $\kappa - h + 1$ intervals. Similarly, $B_{s,h}$ is the set of packets that are generated at relative time $bs - b + 1, bs - b + 2, \ldots, bs$ in the first h intervals and whose destination is to the right of their origin and in the last $\kappa - h + 1$ intervals. The variables $B_{s,h}$ and $M_{s,h}$ statisfy the same relation as in Propositions 4.6. As before, $B_{s,h}$ is the sum of independent Bernoulli random variables, with total expectation less than ρ'. Hence the tail of the distribution of $M_{s,h}$ is exponentially decreasing. As in Theorem 4.7, we can show that a packet's head generated at relative time bs in the h-th window is delayed $b\Delta_1$ steps in its window of origin only if either $|M_{s,h}| \geq \frac{1-\rho'}{2}\Delta_1$ or $|B_{s+1,h}| + |B_{s+2,h}| + \ldots |B_{s+\Delta_1,h}| \geq \frac{1+\rho'}{2}\Delta_1$. Again, the probability of either event is $e^{-\Omega(\Delta_1)}$. The rest of the analysis of the delay distribution can be carried out as before.

Similarly, the probability that a given queue contains more than q packet heads is shown to be $e^{-\Omega(q \log N)}$ by distinguishing two cases depending on whether the queue has been empty in the last $bq \log N$ steps. The analysis can be easily extended to higher dimensions. Details are omitted. ■

8 Concluding remarks

1. All our results easily generalize to the case when the arrivals obey a Poisson distribution. This is because the generating function of a Poisson distribution obeys the inequality stated in Proposition 3.4 (for a Poisson distribution, it is in fact an equality.)

2. An analogue of Theorem 4.7 is shown in [9] for the ring and the torus when the arrival rate is less than 49% of network capacity. Whether this can be extended to the case when the arrival rate is less than 99% of network capacity remains an intriguing open question.

3. An analogue of Theorem 7.1 holds for the $N \times N$ torus using a synchronized version of the farthest-first protocol. We describe this protocol when N is a multiple of b: a packet head moving in increasing order in i and j can move at location (i, j) at time t only when t is congruent to $i + j$ modulo b. Similar rules apply to packets heading in other directions. When two packets wish to traverse the same edge, the one heading further in that direction has priority.

4. We have shown that the ergodic expected delay for the farthest-first protocol on an N-array and on the ring is asymptotically proportional to a negative power of $1 - \rho$ (when N goes to infinity and, then, ρ goes to 1), and proved that the expected delay is bounded by a constant for any greedy protocol as long as $\rho < 1$. A drawback of our approach is

that our bound on the expected delay for arbitrary protocols is at least exponential in $1/(1-\rho)$, as ρ approaches 1. Thus, it seems unlikely that its dependence on ρ is tight. In our extended abstract [7], we have obtained slightly better bounds by using a less intuitive approach and tighter relations betweeen generating functions. How the expected delay for the farthest-first protocol compares with the expected delay for other protocols is a question that deserves further investigation.

References

[1] N. Alon and J. H. Spencer. *The Probabilistic Method.* Wiley Verlag, 1991.

[2] H. Chernoff. A measure for asymptotic efficiency of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

[3] E. G. Coffman, E. N. Gilbert, A. G. Greenberg, F. T. Leighton, P. Robert, and A. L. Stolyar. Queues served by a rotating ring. Manuscript, 1993.

[4] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Prob.*, 14:502–525, 1982.

[5] M. Harcol-Balter and P. Black. Queueing analysis of oblivious packet routing networks. In *Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 583–592, January 1994.

[6] M. Harcol-Balter and D. Wolfe. Bouding delays in packet-routing networks, January 1995. Accepted for publication in the Proceedings of the 36th Annual Symposium on Foundations of Computer Science.

[7] N. Kahale and T. Leighton. Greedy dynamic routing on arrays. In *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 558–566. SIAM Press, 1995.

[8] L. Kleinrock. *Queueing systems*, volume I. Wiley, New York, 1975.

[9] T. Leighton. Average case analysis of greedy routing algorithms on arrays. In *Proceedings of the Second Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 2–10, July 1990.

[10] M. Mitzenmacher. Bounds on the greedy algorithm for array networks. In *Proceedings of the Sixth Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 346–353, June 1994.

[11] G. D. Stamoulis and J. N. Tsitsiklis. The efficiency of greedy routing in hypercubes and butterflies. In *Proceedings of the Second Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 248–259, July 1991.