Simple, Efficient Routing Schemes for All-Optical Networks*
(Extended Abstract)

Michele Flammini†
Department of Mathematics
University of L'Aquila, Via Vetoio loc. Coppito
I-67100 L'Aquila, Italy

Christian Scheideler‡
Department of Applied Mathematics and Computer Science
The Weizmann Institute of Science
76100 Rehovot, Israel

Abstract

All-optical networks promise data transmission rates several orders of magnitudes higher than current networks. The key to high transmission rates in these networks is to maintain the signal in optical form, thereby avoiding the prohibitive overhead of conversion to and from the electrical form, and to exploit the large bandwidth of optical fibers by sending many signals at different frequencies along the same optical link. Optical technology, however, is not as mature as electronic technology. Hence it is important to understand, how efficiently simple routing elements can be used for all-optical communication. In this paper, we consider two types of routing elements. Both types can move messages at different wavelengths to different directions. If in the first type a message wants to use an outgoing link that is already occupied by another message using the same wavelength, the arriving message is eliminated (and therefore has to be rerouted). The second type can evaluate priorities of messages. If more than one message wants to use the same wavelength at the same time then the message with highest priority wins. We prove nearly matching upper and lower bounds for the runtime of a simple and efficient protocol for both types of routing elements, and apply our results to meshes, butterflies, and node-symmetric networks.

* Authors supported in part by the EU ESPRIT Long Term Research Project 20244 (ALCOM-IT).

† email: flammini@univaq.it. Supported in part by the EU TMR Research Training Grant N. ERBFMBICT960861, the Project SLOOP I3S-CNRS/INRIA/Université de Nice–Sophia Antipolis, France, and the Italian MURST 40% project "Algoritmi, Modelli di Calcolo e Strutture Informative".

‡ email: chrsch@wisdom.weizmann.ac.il, supported by a scholarship of the MINERVA foundation. Work was done while staying at Paderborn University, supported in part by DFG-Sonderforschungsbereich 376 "Massive Parallelität: Algorithmen, Entwurfsmethoden, Anwendungen" and by DFG Leibniz Grant Me872/6-1.

SPAA '97 Newport, Rhode Island USA

1 Introduction

The subject of this paper is to present and analyze a simple protocol for sending messages in an emerging generation of networks known as *all-optical networks* [6, 8, 12, 16, 23, 26]. These networks promise data transmission rates several orders of magnitudes higher than current networks. The key to high speeds in these networks is to maintain the signal in optical form, thereby avoiding the prohibitive overhead of conversion to and from the electrical form. (Traditional networks use the electrical form to switch signals along routes, and to restore signal strength. Signals can be modulated electronically at a maximum bit rate of about 50 Gbit/s, while the optical fiber bandwidth is about 25 THz [7].) The high bandwidth of the optical fiber is utilized through *wavelength-division multiplexing*: two signals connecting different source-destination pairs may share a link, provided they are transmitted on carriers having different frequencies (i.e., wavelengths) of light.

The major applications for such networks are in video conferencing, scientific visualization and real-time medical imaging, high-speed supercomputing and distributed computing [12, 26, 10]. We consider routing elements that are capable of directing messages at different wavelengths to different destinations and detecting collisions of messages. A routing element (or *router* in short) consists of wavelength-selective *switches* and *couplers*.

The task of the switches is to direct different wavelengths to different directions. Several types of optical switches have already been developed [15, 5].

The task of the couplers is to combine the signals from many incoming optical fibers into one outgoing optical fiber. Since we do not want to rely on central control, collisions might occur, that is, two or more signals from different incoming fibers use the same wavelength. In our design of protocols we will consider two different strategies to avoid collisions:

- If a message that arrives at a coupler uses a wavelength already used by another message traversing the coupler, the new message is eliminated. This can be realized with the help of detector arrays that tell the electronic control of the coupler which wavelengths are currently used, and wavelength-selective filters at each

incoming fiber.

- If a message that arrives at a coupler uses a wavelength already used by another message traversing the coupler, the message with higher priority is forwarded and the other suspended. This can be realized by using receiver-arrays at the incoming fibers that can read headers of messages on the fly, and wavelength-selective filters.

We call a coupler using the first rule *serve-first coupler* and *priority coupler* otherwise.

The following picture illustrates how a 2×2 router can be built by switches and couplers.

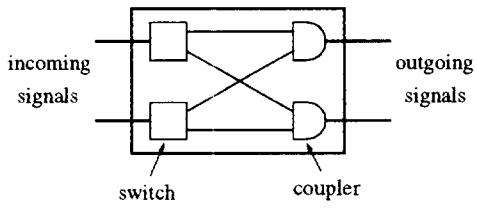

Figure 1: A 2×2 router.

1.1 The Model

We model the topology of an optical network as an undirected graph $G = (V, E)$ where each node in V represents a router (that is connected to a processor) and each edge in E represents two optical links, one in each direction. Each node in V contains an *injection buffer* and a *delivery buffer*. Initially, each message is stored in the injection buffer of its source. Once a message reaches its destination, it is stored in the destination's delivery buffer. During the routing a message cannot be buffered and therefore has to be moved forward or eliminated. Since we do not want to convert messages to and from the electrical form, we do not require the nodes to operate in discrete, synchronous time steps. Instead, we just need to assume that the nodes are fast enough to operate correctly according to one of the collision rules defined above. Hence one time step within our model is defined as the time one bit of the message needs to traverse a link.

The routing problem will be defined by specifying a path collection $\mathcal{P}$, which is a multiset of paths in G. A path is called

- *shortcut-free* if there is no piece of a path in it that is shortcut by any combination of pieces of paths in $\mathcal{P}$, and
- *leveled* if levels can be assigned to the nodes in $\mathcal{P}$ such that for every path in $\mathcal{P}$ every edge leads from a node in level i to a node in level $i+1$ for some $i \geq 0$.

The routing problem consists of routing one message along each of the paths in $\mathcal{P}$. We measure the routing performance of our protocols by

- the number n of paths in $\mathcal{P}$,
- the *dilation* D of $\mathcal{P}$, that is, the length of the longest path in $\mathcal{P}$, and
- the *path congestion* $\tilde{C}$ of $\mathcal{P}$, that is, the maximum number of paths that share a link with a path in $\mathcal{P}$.

A major problem in all-optical networks is to interpret the address header of messages arriving at optical switches, since their switching time is still slow compared with the transmission speed in optical fibers. An approach investigated by AT&T [11, 13] and elsewhere employs a low bit-rate header which is read on the fly by a photodiode or a contact on a semiconductor amplifier. These electrical bits are fed to a controller that operates an optical switch that sends the unconverted optical data bits along the proper path. A message might occupy several links on its way through the network. We therefore model the messages as *worms*, each of which consists of a sequence of fixed size units called *flits*. We assume that it takes one time step to send a flit along a link. The *length* of a worm is defined as the number of flits it contains. The first flit is called the *head* and the remaining flits are called the *body* of the worm. During the routing, a worm occupies a contiguous sequence of links along its path, one flit per link.

The number of wavelengths a router can handle is called the *bandwidth* of the router and denoted by B. As defined for the coupler above, we distinguish between two rules for the router: the *serve-first rule* and the *priority rule*.

1.2 Previous Results

Barry and Humblet [3, 4], Pieris and Sasaki [24] and Pankaj [22] have given lower bounds on the number of wavelengths required for permutation routing in any network, independent of the topology, with a given number of wavelength-selective switches. Pankaj [22] went on to consider lower and upper bounds for a few specific networks; for example, he gives an upper bound of $O(\log^2 n)$ wavelengths for permutation routing on the hypercube. In addition, a number of papers in the communication literature [2, 8, 20] have formulated the routing problem for both switches that can and cannot direct different wavelengths to different directions as combinatorial optimization problems. Aggarwal, Bar-Noy, Coppersmith, Ramaswami, Schieber and Sudan [1] gave bounds on the number of switches required without taking into account the network topology, as a function of the number of wavelengths available. In addition, they proved results on routing in non-blocking permutation networks. Raghavan and Upfal [25] prove results that establish a connection between the expansion of a network and the number of wavelengths required for routing on it considering both switches that can and cannot direct different wavelengths to different directions. In [27], Ramaswami and Sivarajan present a lower bound on the blocking probability for any so-called routing and wavelength assignment (RWA) algorithm if requests and terminations of connections arrive at random, and wavelength-selective switches are used. They study both the case that wavelength conversion is allowed and not allowed at the routers.

To our knowledge, nothing has been found out so far about the maximum number of trials to send a message to its destination given an arbitrary path collection and a fixed bandwidth, if wavelength conversion is not allowed. In case that wavelength conversion is allowed at every router, Cypher *et al* [9] presented an online protocol that routes messages of length L along any simple path collection with congestion C and dilation D in time $O((L \cdot C \cdot D^{1/B} + (D +$

$L)\log n)/B)$, w.h.p.*. However, all-optical devices for wavelength conversion are still a topic in research and might significantly increase the cost of a router. Therefore we want to show in this paper how far one can get without wavelength conversion.

1.3 New Results

In this paper we investigate how much time is necessary to route messages to their destinations given an arbitrary shortcut-free path collection with some fixed bandwidth in case that wavelength conversion is not allowed.

Since the communication time is usually much higher than the calculation time of processors, it is very important to have routing protocols that are as simple as possible. Hence the processors should use strategies that do not need any coordination. Since messages cannot be buffered during the routing along their path there are basically two types of strategies to handle such a situation: starting messages with random delays, or assigning priorities to messages. Clearly, the most simple protocol that can be thought of for sending worms along a fixed path collection using routers with bandwidth B is the following.

Trial-and-Failure Protocol:
all n worms are declared active
for $t = 1$ to T do:

- each active worm is sent out from its source with random startup delay in some suitably chosen range $[\Delta_t]$ using a random wavelength in $[B]$
- for every worm that completely reaches its destination, an acknowledgement is sent back to the source immediately afterwards
- every source that gets back an acknowledgement declares its worm as inactive

Let us call the execution of one for-loop one *round*. Clearly, round t requires at most $\Delta_t + 2(D + L)$ steps to be sure that either an acknowledgement of a successful worm reaches its source, or the worm or its acknowledgement has been (partly) discarded. (Note that if we use priority routers it can happen that worms are only partly discarded.)

Previously, only delay sequence arguments were used to analyze such protocols (see, e.g., [9, 28]). In this paper we use delay tree arguments that yield much more accurate upper bounds on the runtime. In particular, we are able to prove the following three results depending on the contention resolution rule. Their proofs can be found in Section 2 and Section 3. Let $\alpha = \tilde{C} + B(\frac{D}{L} + 1) + 2$ and $\beta = \alpha/\tilde{C} + 2$. The first theorem presents a nearly tight analysis of the protocol above for leveled path collections.

Main Theorem 1.1 *For any leveled path collection of size n with dilation D and path congestion $\tilde{C}$ using serve-first routers with bandwidth B the protocol above routes a worm of length L along each of these paths in time*

$$O\left(\frac{L\cdot\tilde{C}}{B} + \left(\sqrt{\log_\alpha n} + \log\log_\beta n\right)\left(\frac{L\log n}{B} + D + L\right)\right),$$

w.h.p. Furthermore there exists a leveled path collection such that, for any $L \geq 2$, the expected runtime is bounded by

$$\Omega\left(\frac{L\cdot\tilde{C}}{B} + \left(\sqrt{\log_\alpha n} + \log\log_\beta n\right)(D + L)\right).$$

Since in contrast to leveled path collections it can happen in some shortcut-free path collections that worms prevent each other from reaching their destinations, we get a slightly worse result for arbitrary shortcut-free path collections.

Main Theorem 1.2 *For any shortcut-free path collection of size n with dilation D and path congestion $\tilde{C}$ using serve-first routers with bandwidth B the protocol above routes a worm of length L along each of these paths in time*

$$O\left(\frac{L\cdot\tilde{C}}{B} + (\log_\alpha n + \log\log_\beta n)\left(\frac{L\log^{3/2} n}{B} + D + L\right)\right),$$

w.h.p. Furthermore there exists a shortcut-free path collection such that, for any $L \geq 2$, the expected runtime is bounded by

$$\Omega\left(\frac{L\cdot\tilde{C}}{B} + (\log_\alpha n + \log\log_\beta n)(D + L)\right).$$

Note that for the case $L = 1$ (i.e., worms cannot prevent each other from reaching their destinations) or there are no directed loops in the path collection of length below $\sqrt{\log_\alpha n}$, the upper bound in Main Theorem 1.2 can be reduced to the upper bound in Main Theorem 1.1. For any other situation, we also obtain this bound if we replace the serve-first routers by priority routers.

Main Theorem 1.3 *For any collection of n shortcut-free paths with dilation D and path congestion $\tilde{C}$ using priority routers with bandwidth B the protocol above routes a worm of length L along each of these paths in time*

$$O\left(\frac{L\cdot\tilde{C}}{B} + \left(\sqrt{\log_\alpha n} + \log\log_\beta n\right)\left(\frac{L\log n}{B} + D + L\right)\right),$$

w.h.p. Furthermore there is a shortcut-free path collection and a strategy for assigning priorities to the worms such that, for any $L \geq 2$, the expected runtime is bounded by

$$\Omega\left(\frac{L\cdot\tilde{C}}{B} + \left(\sqrt{\log_\alpha n} + \log\log_\beta n\right)(D + L)\right).$$

Note that the upper bound holds for *any* assignment of priorities to the worms such that no two worms with the same priority can meet in one round, whether these priorities are changed from round to round, chosen randomly, or deterministically.

The main theorems indicate that for shortcut-free path collections the priority rule is more powerful than the serve-first rule. Often, $\Omega(\frac{L\cdot\tilde{C}}{B} + D + L)$ is a lower bound for any protocol using serve-first or priority routers. In this case the runtime of our protocol can get optimal if $\tilde{C}$ is large enough compared to D and L. Note that, for instance, for the butterfly network of size N the average path congestion of permutation routing problems is $\Theta(\log^2 N)$, whereas its diameter is $O(\log N)$.

The upper and lower bounds in Main Theorems 1.1 and 1.3 will be proved in Section 2, and the upper and lower bound in Main Theorem 1.2 will be given in Section 3. In the following, we describe some applications of the trial-and-failure protocol.

*By "with high probability" (or w.h.p. for short) we mean a probability of at least $1 - 1/n^k$ for any constant $k > 0$.

1.4 Applications

The results presented above can be applied, e.g., to node-symmetric networks [9]. Note that node-symmetric networks form a very general class and include most of the standard networks such as the d-dimensional torus, the wrap-around butterfly, the hypercube, etc. Furthermore, the best expanders that have an explicit construction are all node-symmetric (see, e.g., [19]).

Theorem 1.4 *For any bounded degree node-symmetric network of size n with diameter D using priority routers with bandwidth B there is an online protocol for routing a randomly chosen function in time*

$$O\left(\frac{L\cdot D^2}{B}+\left(\sqrt{\log_D n}+\log\log n\right)(D+L)\right),$$

w.h.p.

Proof. In [21] it is shown that, for every node-symmetric network of size n with diameter D, a shortcut-free system of paths can be chosen such that a collection of paths chosen out of this system for routing a randomly chosen function has a path congestion of $O(D^2+\log n)$, w.h.p. Using this in the time bound of Main Theorem 1.3 yields the theorem. ∎

The previous best time bound for the case $B=1$ was $O(L\cdot D^2+(D+L)\log n)$ [9]. (Note that for $B>1$ the protocols in [9] allow wavelength conversion which we do not allow here.) The result in Theorem 1.4 can be improved for d-dimensional meshes and tori.

Theorem 1.5 *For any d-dimensional mesh of side length n using priority routers with bandwidth B there is an online protocol for routing a randomly chosen function in time*

$$O\left(\frac{L\cdot d\cdot n}{B}+(\sqrt{d}+\log\log n)\left(\frac{L\cdot d\log n}{B}+d\cdot n+L\right)\right),$$

w.h.p.

Proof. Using techniques in [9], it is easy to show that there exists a routing strategy for routing a randomly chosen function that has a path congestion of $O(d\cdot n)$, w.h.p. Since the size N of a d-dimensional mesh with side length n is equal to n^d, it follows that

$$\sqrt{\log_\alpha N}=O\left(\sqrt{d\log_{dn} n}\right)=O\left(\sqrt{d}\right),$$

where α is chosen as in the main theorems. In case that $\sqrt{d}\le\log\log N$ we have that $n\ge N^{1/\log\log N}$ and therefore $\log\log N=O(\log\log n)$. This concludes the proof. ∎

Note that the previous best time bound for the case $B=1$ was $O(L\cdot d\cdot n+(d\cdot n+L)\log n)$ [9]. In case that we use butterfly networks, we can use more simple serve-first routers to obtain the following result.

Theorem 1.6 *For any $\log n$-dimensional butterfly using serve-first routers with bandwidth B there is a leveled path system such that a randomly chosen q-function can be routed from the inputs to the outputs in time*

$$O\left(\frac{L\cdot q\log n}{B}+\sqrt{\frac{\log n}{\log(q\log n)}}\left(\frac{L\log n}{B}+L+\log n\right)\right),$$

w.h.p.

For $B=1$, this improves for some cases the previous best time bound of $O(L\cdot q\log n+(L+\log n)\log n)$ [9].

2 Proof of Main Theorems 1.1 and 1.3

In this section we prove upper and lower bounds on the runtime of our protocol using serve-first routers in leveled path collections, or priority routers in shortcut-free path collections. In order to simplify the presentation, we will concentrate on serve-first routers in leveled path collections, and note the analogy to routing with priority routers in shortcut-free path collections whenever it is necessary.

Hence suppose we want to route worms of length L along a collection of n leveled paths with path congestion $\tilde{C}$ and dilation D, using serve-first routers with bandwidth B. (In order to simplify the analysis we assume that $\tilde{C}$ covers both messages and acknowledgements.)

2.1 The Upper Bound

In this section we want to prove an upper bound for the number T of rounds that is necessary to route all worms using the trial-and-failure protocol with some suitable values of Δ_t. We first want to find a structure that witnesses a long runtime of the protocol.

Assume that a worm w_0 is still active after t rounds. Then there must have been a worm w_1 that prevented it from moving forward in round t. But if w_0 and w_1 have been active at round t there must have been (not necessarily different) worms w_2 and w_3 which prevented w_0 and w_1 from moving forward in round $t-1$. Continuing with this argumentation until round 1 we find:

If worm w_0 is still active after t rounds then the following tree can be constructed such that the nodes represent worms and two nodes with a common father a collision event.

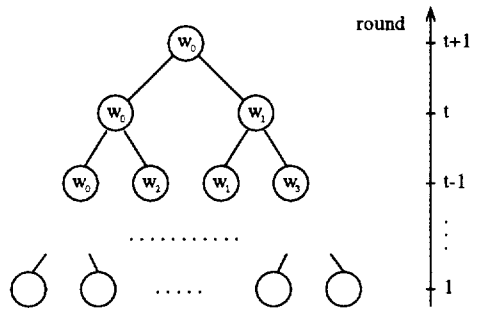

Figure 2: The witness tree of depth t.

Let us call this tree a *witness tree of depth t*, and denote it by $\mathcal{W}(t)$. The following definition formalizes what kind of embeddings of worms into the nodes of $\mathcal{W}(t)$ we only have to consider.

Definition 2.1 *Let φ be an embedding of worms into the nodes of $\mathcal{W}(t)$. A pair of worms (w,w') is called* collision pair *if $w\neq w'$, w is embedded in the left son, and w' is embedded in the right son of a common father in $\mathcal{W}(t)$. We call φ* valid *if for every collision pair (w,w') embedded at level i of $\mathcal{W}(t)$ it holds that*

- *w is also embedded in the father of w and w',*
- *there is no collision pair (w,w'') at level i with $w'\neq w''$, and*

- *the paths of w and w' share an edge.*

A valid embedding is called active *if for any collision pair (w, w') embedded at level i of $\mathcal{W}(t)$ it holds that w and w' use the same wavelength and w' prevents w from moving forward in round $t - i + 1$.*

Following the discussion above, we can state the following lemma.

Lemma 2.2 *If worm w_0 is still active after t rounds then there is an active embedding φ of worms into $\mathcal{W}(t)$ that maps w_0 to the root of $\mathcal{W}(t)$.*

The above lemma implies that it suffices to find a suitable upper bound for the probability (w.r.t. random choices for the delays and wavelengths used by the worms) that there is an active embedding φ for any worm w_0 in order to prove the upper bound in Main Theorem 1.1.

In order to count the number of valid embeddings we introduce the following graph.

Definition 2.3 *Let φ be a valid embedding. For each level $i \in \{1, \ldots, t\}$ of $\mathcal{W}(t)$, let $G_i = (V_i, E_i)$ be a directed graph whose nodes represent the set of worms embedded in level i and whose edges (w, w') represent the collision pairs (w, w') in level i. We call the worms in V_{i-1}* old *and the worms in $V_i \setminus V_{i-1}$* new *w.r.t. G_i.*

We assume G_0 to be the graph consisting only of a single node. Let the set of graphs $G_0, \ldots, G_t$ be called *valid* if they represent a valid embedding into $\mathcal{W}(t)$. Clearly, each valid embedding into $\mathcal{W}(t)$ has a unique valid set of graphs $G_0, \ldots, G_t$, and vice versa. Thus we can switch between either considering valid sets of graphs $G_0, \ldots, G_t$ or considering valid embeddings into $\mathcal{W}(t)$ in an arbitrary way.

For any valid embedding φ into the witness tree $\mathcal{W}(t)$, let $m_i = |V_i|$ denote the total number of worms and $\ell_i = m_i - m_{i-1}$ denote the number of new worms at level i. Let $\tilde{C}_j$ be an upper bound for the path congestion that holds at round j w.h.p. using the protocol above for suitably chosen $\Delta_1, \ldots, \Delta_j$ (determined later). Then it holds for the number $V(t,k)$ of valid embeddings in $\mathcal{W}(t)$ using k worms:

$$V(t,k) \le n \sum_{\substack{\ell_1,\ldots,\ell_t \ge 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} \binom{m_{i-1}}{\ell_i} \tilde{C}_{t-i+1}^{\ell_i} (\ell_i + m_{i-1})^{m_{i-1}-\ell_i}$$

w.h.p. This formula is derived as follows.

- There are n ways to choose the worm that is embedded in the root of $\mathcal{W}(t)$.
- There are $\binom{m_{i-1}}{\ell_i}$ possibilities to choose ℓ_i old worms that collide with (and therefore narrow down the choices for) each of the ℓ_i new worms. Hence afterwards there are at most $\tilde{C}_{t-i+1}^{\ell_i}$ ways w.h.p. to choose the ℓ_i new worms.
- For the remaining $m_{i-1} - \ell_i$ old worms there are at most $\ell_i + m_{i-1}$ possibilities to choose the worm that prevents it from moving forward.

Before we can proceed with our calculation, we need an upper bound that holds for the path congestion after every round w.h.p., and need an upper bound for the probability that the embeddings counted in $V(t,k)$ are active.

Lemma 2.4 *For all $t \ge 2$ it holds that, if $\Delta_i \ge 8e\frac{L\tilde{C}}{B2^{i-1}}$ for all $i \in \{1, \ldots, t-1\}$, then the path congestion $\tilde{C}_t$ at round t is at most $\max\{\frac{\tilde{C}}{2^{t-1}}, O(\log n)\}$, w.h.p.*

Proof. The proof will be done by induction. Suppose, the path congestion at the beginning of round t is bounded by $\frac{\tilde{C}}{2^{t-1}} \ge 2\alpha \log n$ for some arbitrary constant $\alpha > 1$. Let $\Delta_t \ge 8e\frac{L\tilde{C}}{B2^{t-1}}$ be the delay range in round t. Consider any fixed worm w. Let $w_1, \ldots, w_k$ be the worms participating in round t whose paths share a link with the path of w, $k \le \frac{\tilde{C}}{2^{t-1}}$. Further let the binary random variable $X_i = 1$ if and only if w_i fails to reach its destination in round t. Then $X = \sum_{i=1}^{k} X_i$ is a random variable denoting the path congestion of w after round t.

Since the delays and wavelengths are chosen independently and we only consider shortcut-free paths, it holds for every pair of worms w_i and w_j that

$$\text{Prob}(w_i \text{ is (partly) discarded by } w_j) \le \frac{2L}{B\Delta_t} .$$

Therefore,

$$\text{Prob}(X_i = 1) \le \frac{\tilde{C}_t \cdot 2L}{B\Delta_t} \le \frac{1}{4e}$$

independently from the other worms and hence $E(X) \le \frac{\tilde{C}}{4e \cdot 2^{t-1}}$. Let $\mu = \frac{\tilde{C}}{4e \cdot 2^{t-1}}$. Then we can use Chernoff bounds (see [14]) to prove that, for $\epsilon = 2e - 1$,

$$\begin{aligned} \text{Prob}(X \ge (1+\epsilon)\mu) &\le \left(\frac{e}{1+\epsilon}\right)^{(1+\epsilon)\mu} = \left(\frac{1}{2}\right)^{2e\frac{\tilde{C}}{4e \cdot 2^{t-1}}} \\ &\le \left(\frac{1}{2}\right)^{\alpha \log n} = \left(\frac{1}{n}\right)^{\alpha} . \end{aligned}$$

For $\alpha > 1$, this yields the lemma. ∎

Hence in the following we assume that $\tilde{C}_i = \max\{\frac{\tilde{C}}{2^{i-1}}, O(\log n)\}$ for all $i \in \{1, \ldots, t\}$.

Next we bound the probability that any of the embeddings counted in $V(t,k)$ is active. As noted above, the probability that a collision pair (w, w') in level i of $\mathcal{W}(t)$ is active is at most $\frac{2L}{B\Delta_{t-i+1}}$. Let a node in G_i be called *sink* if it has outdegree 0. Then we can prove the following nice property.

Lemma 2.5 *For every level i, the connected components in G_i are directed trees with new worms as sinks.*

Proof. Every old worm needs a witness for its collision in round i and therefore can not be a sink like the new worms, that have no witness since they are just introduced as witnesses in round i. Further a connected component can not have a cycle since

- in leveled path collections using the serve-first rule this would mean that worms prevent each other from moving forward. This however, is not possible in a leveled path collection.

- in shortcut-free path collections using the priority rule this would mean that a worm w_1 is discarded by a worm w_2 that has a higher priority than w_1, and w_2 is discarded by a worm w_3 that has a higher priority than w_2, and so on, until we arrive at a worm w_i that is discarded by w_1, since it has a higher priority than w_i. This, however, is not possible as long as no two worms with the same rank can meet in a round.

∎

Since every directed tree in G_i of size s implies a probability of $\leq (\frac{2L}{B\Delta_{t-i+1}})^{s-1}$ that its edges correspond to collisions of worms, and since there are exactly ℓ_i trees in G_i, we obtain a probability of at most

$$\left(\frac{2L}{B\Delta_{t-i+1}}\right)^{(m_{i-1}+\ell_i)-\ell_i} = \left(\frac{2L}{B\Delta_{t-i+1}}\right)^{m_{i-1}}$$

that the collisions in level i are active. Therefore the probability $P(t,k)$ that there exists an active embedding in $\mathcal{W}(t)$ is at most

$$n \sum_{\substack{\ell_1,\ldots,\ell_t\geq 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} \binom{m_{i-1}}{\ell_i} \tilde{C}_{t-i+1}^{\ell_i} (\ell_i + m_{i-1})^{m_{i-1}-\ell_i} \left(\frac{L}{B\Delta_{t-i+1}}\right)^{m_{i-1}}$$

In case that $\ell_i \leq m_{i-1}/2$, we get

$$\binom{m_{i-1}}{\ell_i} (\ell_i + m_{i-1})^{m_{i-1}-\ell_i} \leq (3em_{i-1})^{m_{i-1}-\ell_i},$$

and otherwise (that is, $m_{i-1}/2 < \ell_i \leq m_{i-1}$)

$$\binom{m_{i-1}}{\ell_i} (\ell_i + m_{i-1})^{m_{i-1}-\ell_i} \leq 2^{2\ell_i} (2m_{i-1})^{m_{i-1}-\ell_i}.$$

Therefore, $P(t,k)$ is at most

$$n \sum_{\substack{\ell_1,\ldots,\ell_t\geq 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} 2^{2\ell_i} (3em_{i-1})^{m_{i-1}-\ell_i} \tilde{C}_{t-i+1}^{\ell_i} \left(\frac{2L}{B\Delta_{t-i+1}}\right)^{m_{i-1}}$$

$$\leq n \cdot \left(\frac{8L\cdot\tilde{C}}{B\Delta_1}\right)^{k-1} \sum_{\substack{\ell_1,\ldots,\ell_t\geq 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} \left(\frac{6eLm_{i-1}}{B\Delta_{t-i+1}}\right)^{m_{i-1}-\ell_i}$$

for Δ_i chosen such that $\frac{\tilde{C}}{\Delta_1} \geq \frac{\tilde{C}_i}{\Delta_i}$. Furthermore, the following lemma holds. Its proof is omitted here.

Lemma 2.6 *If $\Delta_i \geq \frac{40e^2Lk}{B}$ and $\Delta_{i+1} \leq \Delta_i$ for all $i \in \{1,\ldots,t-1\}$ then*

$$\max_{\substack{\ell_1,\ldots,\ell_t\geq 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} \left(\frac{6eLm_{i-1}}{B\Delta_{t-i+1}}\right)^{m_{i-1}-\ell_i} \leq \left(\frac{6eLt}{B\Delta_t}\right)^{\frac{1}{2}(t-\lceil\log k\rceil)^2}.$$

Clearly, there are $\binom{t+k-1}{t} \leq 2^{t+k-1}$ possibilities for choosing the $\ell_1,\ldots,\ell_t$ such that $\sum_{i=1}^{t} \ell_i = k-1$. Thus we get for $\Delta_i \geq \frac{40e^2Lk}{B}$ that

$$\begin{aligned} P(t,k) &\leq n \left(\frac{8L\cdot\tilde{C}}{B\Delta_1}\right)^{k-1} 2^{t+k-1} \left(\frac{6eLt}{B\Delta_t}\right)^{\frac{1}{2}(t-\lceil\log k\rceil)^2} \\ &= n\cdot 2^t \left(\frac{16L\cdot\tilde{C}}{B\Delta_1}\right)^{k-1} \left(\frac{6eLt}{B\Delta_t}\right)^{\frac{1}{2}(t-\lceil\log k\rceil)^2}. \end{aligned}$$

For any constant $\gamma > 0$, let

$$k_0 = \frac{(2+\gamma)\log n}{\log\left(2 + \frac{B}{16\tilde{C}}\left(\frac{D}{L}+1\right)\right)} + 1$$

and

$$T \geq \sqrt{\frac{2(2+\gamma)\log n}{\log\left(\left(\frac{\tilde{C}}{\log^{3/2} n} + \sqrt{\log n} + \frac{B(D/L+1)}{\sqrt{\log n}}\right)\right)}} + \lceil\log k_0\rceil.$$

If the routing takes more than T rounds then one of the following two cases must be true:

(1) There must exist an active embedding into a witness tree $\mathcal{W}(t)$ with $t \leq T$ and $k \in \{k_0,\ldots,2k_0\}$ different worms.

(2) There must exist an active embedding into a witness tree $\mathcal{W}(T)$ with $k \leq k_0$ different worms.

Suppose that $\Delta_i \geq \max\{\frac{32L\cdot\tilde{C}_i}{B}, \frac{32L\cdot\tilde{C}}{B\log n}, \frac{40e^2Lk_0}{B}\} + D + L$. Then we get:

$$\begin{aligned} &\text{Prob(The routing takes more than } T \text{ rounds)} \\ &\leq \text{Prob(Case (1) holds)} + \text{Prob(Case (2) holds)} \\ &\leq \sum_{t=\log k_0}^{T} \sum_{k=k_0}^{2k_0} P(t,k) + \sum_{k=T}^{k_0} P(T,k) \\ &\leq \sum_{t=\log k_0}^{T} \sum_{k=k_0}^{2k_0} n\cdot 2^t \left(\frac{16L\cdot\tilde{C}}{B\Delta_1}\right)^{k-1} + \\ &\quad \sum_{k=T}^{k_0} n\cdot 2^T \left(\frac{16L\cdot\tilde{C}}{B\Delta_1}\right)^{k-1} \left(\frac{6eLT}{B\Delta_T}\right)^{\frac{1}{2}(T-\lceil\log k\rceil)^2} \\ &\leq \ldots \leq \frac{n^{-\gamma}}{2} + \frac{n^{-\gamma}}{2} \leq n^{-\gamma} \end{aligned}$$

Therefore the overall runtime is

$$\begin{aligned} &\sum_{t=1}^{T} (\Delta_t + 2(D+L)) \\ &= O\left(\sum_{t=1}^{T}\left(D + L + \frac{L}{B}\left(\frac{\tilde{C}}{2^{t-1}} + \frac{\tilde{C}}{\log n} + \log n\right)\right)\right), \end{aligned}$$

w.h.p., which is bounded by

$$O\left(\frac{L\cdot\tilde{C}}{B} + \left(\sqrt{\log_\alpha n} + \log\log_\beta n\right)\left(\frac{L\log n}{B} + D + L\right)\right),$$

where $\alpha = \tilde{C} + B(\frac{D}{L}+1)+2$ and $\beta = 2+\frac{B}{\tilde{C}}(\frac{D}{L}+1)$. This completes the proof of the upper bound of Main Theorems 1.1 and 1.3.

2.2 The Lower Bound

In this section we will prove the lower bound in Main Theorems 1.1 and 1.3. We use a path collection that consists of the following two types of subcollections.

- Let $d = \lfloor \frac{L-1}{2} \rfloor + 1$. The first type consists of $n/(2\sqrt{\log n})$ structures consisting of $\sqrt{\log n}$ paths of length D that are connected as shown in Figure 3.

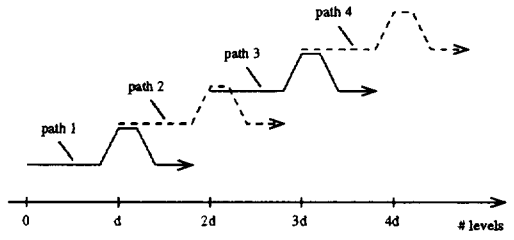

Figure 3: A type-1 structure.

  In general, the ith path starts in level $(i-1)d$ for all $i \geq 0$. Paths i and $i+1$ have a common edge from level $i \cdot d$ to level $i \cdot d + 1$.

- The second type consists of $n/(2\tilde{C})$ structures each consisting of $\tilde{C}$ identical paths of length D.

We assume that along each of these paths one worm of length $L \geq 2$ has to be sent.

We first want to compute how long it takes to route all worms in a type-1 structure. In case of routing along shortcut-free paths using priority routers, we assume that the worm traversing path i has rank i, and in case of conflicts worms with higher ranks are preferred. In order to bound the number of ways to assign delays and wavelengths to the worms such that conflicts occur, we need the following lemma. Its proof is easy and therefore omitted here.

Lemma 2.7 *Consider an arbitrary round of the trial-and-failure protocol with delay range $\Delta \geq L$. Suppose that the worms traversing the first $i+1$ paths are still active at the beginning of this round. Then with probability at least $\left(\frac{L-1}{2B\Delta}\right)^i$ the worms traversing the first i paths are discarded.*

Consider now the situation that it takes $t+1$ rounds to route the worms traversing the first $t+1$ paths in a type-1 structure. This could happen, e.g., if in round i only w_{t-i+2} is able to reach its destination, and the worms $w_1, \ldots, w_{t-i+1}$ are discarded. According to the lemma above, for $L \geq 2$ the probability of such an event is at least

$$\prod_{i=1}^{t+1} \frac{B\Delta_i \left(\frac{L-1}{2}\right)^{t-i+1}}{(B(\Delta_i + L))^{t-i+2}} = \prod_{i=1}^{t} \left(\frac{L-1}{2B(\Delta_i + L)}\right)^{t-i+1}, \quad (1)$$

where $\Delta_i \geq 1$ is the delay range for round i. Clearly, the number of time steps necessary for the t rounds is at least $\Omega(\sum_{i=1}^{t} (\Delta_i + D + L))$. Given a fixed $\Delta = \sum_{i=1}^{t} \Delta_i$, the product in (1) gets minimal if $\Delta_i + L = (t-i+1)(\Delta + t \cdot L)/\binom{t+1}{2}$ for all $i \in \{1, \ldots, t\}$. This is shown in the following lemma.

Lemma 2.8 *Consider $x_1, \ldots, x_n \in \mathbb{R}_+$ with $y = \sum_{i=1}^{n} x_i$. Then, for every $b \in [0, y]$, $\prod_{i=1}^{n} (x_i + b)^i$ gets maximal if $x_i + b = i(y + n \cdot b)/\binom{n+1}{2}$ for all $i \in \{1, \ldots, n\}$.*

The proof of the lemma is a simple induction argument. Let $\tilde{\Delta} = \Delta/t$. Since there are $n/(2\sqrt{\log n})$ type-1 structures, and each structure has a probability of at least

$$\prod_{i=1}^{t} \left(\frac{(L-1)(t+1)}{2B \cdot 2(t-i+1)(\tilde{\Delta} + L)}\right)^{t-i+1} \geq \left(\frac{L-1}{4B(\tilde{\Delta} + L)}\right)^{t^2}$$

to have active worms after t rounds, the expected number of type-1 structures that have active worms after t rounds is at least

$$\frac{n}{2\sqrt{\log n}} \left(\frac{L-1}{4B(\tilde{\Delta}+L)}\right)^{t^2} < 1 \quad \Leftrightarrow \quad t \geq \sqrt{\frac{\log\left(\frac{n}{2\sqrt{\log n}}\right)}{\log\left(\frac{4B(\tilde{\Delta}+L)}{L-1}\right)}}$$

Hence the expected number of rounds that are needed to route all worms in all type-1 structures is at least

$$\Omega\left(\sqrt{\log_{B(\tilde{\Delta}/L+2)} n}\right) .$$

In order to bound the time needed to route all worms in the type-2 structures, we distinguish between the cases $\tilde{C} \geq 2^{\sqrt{\log n}}$ and $\tilde{C} \leq 2^{\sqrt{\log n}}$.

Case $\tilde{C} \leq 2^{\sqrt{\log n}}$:
Note that any routing protocol needs at least $\Omega(\frac{L\tilde{C}}{B} + D + L)$ steps to route all worms in a type-2 structure. Therefore the expected number of steps the protocol needs to route all worms is at least

$$\Omega\left(\frac{L\tilde{C}}{B} + \sqrt{\log_{B(\tilde{\Delta}/L+2)} n}(\tilde{\Delta} + D + L)\right) .$$

Since the runtime bound gets minimal for some $\tilde{\Delta}$ chosen in $O(\frac{L\tilde{C}}{B} + D + L)$, the expected runtime of the protocol is at least

$$\Omega\left(\frac{L\tilde{C}}{B} + \sqrt{\log_{\alpha} n}(D + L)\right) ,$$

where $\alpha = \tilde{C} + B(\frac{D}{L} + 1) + 2$. Let $\beta = \alpha/\tilde{C} + 2$. Since $\tilde{C} \leq 2^{\sqrt{\log n}}$, it holds that $\sqrt{\log_{\alpha} n} \leq \log\log n$ only if $B(\frac{D}{L} + 1) \geq 2^{\log n/(\log\log n)^2} \gg \tilde{C}$. In this case, however, $\log\beta = \Theta(\log\alpha)$, that is, $\sqrt{\log_{\alpha} n} \geq \log\log_{\beta} n$. Therefore we arrive at an expected runtime of the protocol of at least

$$\Omega\left(\frac{L\tilde{C}}{B} + \left(\sqrt{\log_{\alpha} n} + \log\log_{\beta} n\right)(D + L)\right)$$

time steps.

Case $\tilde{C} \geq 2^{\sqrt{\log n}}$:
Let $\tilde{C}_i$ be the minimum over all type-2 structures P of the number of worms that are still active in P after i rounds. Then the following lemma holds.

Lemma 2.9 *For every $t \geq 2$ and $L(\frac{\tilde{C}}{B}+2) \leq \Delta_1, \ldots, \Delta_{t-1} \leq \hat{\Delta}$ with $\tilde{C}/(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}})^{2^{t-1}-1} \geq 7 \ln n$ it holds that*

$$\tilde{C}_t \geq \frac{\tilde{C}}{\left(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}}\right)^{2^{t-1}-1}}$$

w.h.p.

Proof. We only sketch the proof. For $t = 1$, the bound on $\tilde{C}_t$ trivially holds. Suppose that the bound above for $\tilde{C}_t$ is true for some $t \geq 1$. Then we want to show that, if $\Delta_t \leq \hat{\Delta}$ and $\tilde{C}/(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}})^{2^t-1} \geq 7 \ln n$, then we get $\tilde{C}_{t+1} \geq \tilde{C}/\left(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}}\right)^{2^t-1}$, w.h.p.

Assume in the following that $\tilde{C}/(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}})^{2^t-1} \geq 6\lceil\alpha \ln n\rceil$ for some fixed constant $\alpha > 1$. Consider any fixed type-2 structure P. Let $w_1, \ldots, w_c$ be the worms participating in round t that use this type-2 structure, $c \geq \tilde{C}/(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}})^{2^{t-1}-1}$. Further let the binary random variable $X_i = 1$ if and only if w_i fails to reach its destination in round t, and $X = \sum_{i=1}^{c} X_i$. Then it can be shown that $X_1, \ldots, X_{c/2}$ can be considered as $(2\lceil\alpha \ln n\rceil)$-wise independent with probability at least $\frac{\Delta_t - L}{\Delta_t} \cdot \frac{(L-1)\tilde{C}_t}{4B\hat{\Delta}} \geq \frac{(L-1)\tilde{C}_t}{8B\hat{\Delta}}$ if $\Delta_t \geq 2L$. In this case we get an expected path congestion after round t of $E(X) \geq \frac{\tilde{C}_t}{2} \cdot \frac{(L-1)\tilde{C}_t}{8B\hat{\Delta}}$. Let $\mu = \frac{\tilde{C}_t}{2} \cdot \frac{(L-1)\tilde{C}_t}{8B\hat{\Delta}}$. Then $\mu \geq 2\tilde{C}/\left(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}}\right)^{2^t-1}$. Using this together with a result shown in [29] (see Theorem 5), we get:

$$\begin{aligned}\text{Prob}\left(X \leq \frac{\tilde{C}}{\left(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}}\right)^{2^t-1}}\right) &\leq \text{Prob}\,(X \leq \mu/2) \\ &\leq e^{-\alpha \ln n} = \left(\frac{1}{n}\right)^{\alpha}.\end{aligned}$$

Hence for $\alpha > 1$ the path congestion after round t is bounded by $\tilde{C}/(\frac{32B\hat{\Delta}}{(L-1)\tilde{C}})^{2^t-1}$ for all type-2 structures, w.h.p. ∎

Thus for any $L \geq 2$ and $\hat{\Delta} \geq 1$ it holds for the expected number t of rounds to route all worms in type-2 structures that

$$\frac{\tilde{C}}{\left(\frac{32B(\hat{\Delta}+L(\tilde{C}/B+2))}{(L-1)\tilde{C}}\right)^{2^t-1}} \leq 7 \ln n$$

$$\Leftrightarrow \quad t \geq \log\left(1 + \log_\gamma \frac{\tilde{C}}{7 \ln n}\right),$$

where $\gamma = \frac{32B(\hat{\Delta}+L(\tilde{C}/B+2))}{(L-1)\tilde{C}}$. Since $\tilde{C} \geq 2^{\sqrt{\log n}}$, the expected runtime of the protocol is at least

$$\Omega\left(\frac{L\tilde{C}}{B} + \left(\sqrt{\log_{\frac{B\hat{\Delta}}{L}+2} n} + \log\log_\gamma n\right)(\hat{\Delta} + D + L)\right).$$

Since this bound gets minimal for $\hat{\Delta} = O(\frac{L\tilde{C}}{B} + D + L)$, we get an expected runtime of at least

$$\Omega\left(\frac{L\tilde{C}}{B} + \left(\sqrt{\log_\alpha n} + \log\log_\beta n\right)(D + L)\right)$$

time steps, where $\alpha = \tilde{C} + B(\frac{D}{L} + 1) + 2$ and $\beta = \alpha/\tilde{C} + 2$.

3 Proof of Main Theorem 1.2

In this section we prove upper and lower bounds on the runtime of our protocol for shortcut-free path collections using serve-first routers. Hence suppose we want to route worms of length L along a collection of n shortcut-free paths with path congestion $\tilde{C}$ and dilation D, using serve-first routers with bandwidth B. (We again assume that $\tilde{C}$ covers both messages and acknowledgments.)

3.1 The Upper Bound

In this section we want to prove the upper bound in Main Theorem 1.2. Let the witness tree $\mathcal{W}(t)$ be defined as in Section 2. For any valid embedding φ into $\mathcal{W}(t)$, let $m_i = |V_i|$ denote the total number of worms and $\ell_i = m_i - m_{i-1}$ denote the number of new worms at level i. Furthermore let c_i denote the number of old worms that are in a connected component in G_i with a new worm. Let $\tilde{C}_j$ be an upper bound for the path congestion that holds w.h.p. after round j using the trial-and-failure protocol for suitably chosen $\Delta_1, \ldots, \Delta_j$ (determined later). Then it holds for the number $V(t,k)$ of valid embeddings in $\mathcal{W}(t)$ using k worms:

$$\begin{aligned}V(t,k) \leq\ & n \sum_{\substack{\ell_1,\ldots,\ell_t \geq 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} \sum_{c_i=\ell_i}^{m_{i-1}} \binom{m_{i-1}}{c_i}\binom{c_i}{\ell_i} \cdot \tilde{C}_{t-i+1}^{\ell_i} \cdot \\ & (\ell_i + c_i)^{c_i-\ell_i} \cdot (m_{i-1} - c_i)^{m_{i-1}-c_i},\end{aligned}$$

w.h.p. This formula is derived as follows.

- There are n ways to choose the worm that is embedded in the root of $\mathcal{W}(t)$.
- There are $\binom{m_{i-1}}{c_i}$ possibilities to choose c_i old worms that lie in a connected component in G_i with a new worm, and $\binom{c_i}{\ell_i}$ possibilities to choose ℓ_i old worms that collide with (and therefore narrow down the choices for) each of the ℓ_i new worms. Therefore afterwards there are at most $\tilde{C}_{t-i+1}^{\ell_i}$ ways w.h.p. to choose the ℓ_i new worms. For the remaining $c_i - \ell_i$ old worms there are at most $\ell_i + c_i$ possibilities to choose the worm that prevents it from moving forward.
- For each of the remaining $m_{i-1} - c_i$ old worms there are at most $m_{i-1} - c_i$ ways to determine the old worm which prevents it from moving forward.

Before we can proceed with our calculation, we need an upper bound that holds for the path congestion w.h.p., and need an upper bound for the probability that the embeddings counted in $V(t,k)$ are active.

Since the delays and wavelengths are chosen independently and we only consider shortcut-free paths, it holds for every pair of worms w_i and w_j at round t that

$$\text{Prob}(w_i \text{ is blocked by } w_j) \leq \frac{L}{B\Delta_t}.$$

Therefore we get analogous to Lemma 2.4 that, if $\Delta_i \geq 4e\frac{L\tilde{C}}{B2^{t-i}}$ for all $i \in \{1, \ldots, t-1\}$, then the path congestion $\tilde{C}_t$ at round t is at most $\max\{\frac{\tilde{C}}{2^{t-1}}, O(\log n)\}$, w.h.p.

Next we bound the probability that the embeddings counted in $V(t,k)$ are active. As noted above, the probability

that a collision pair (w, w') in level i of $\mathcal{W}(t)$ is active is at most $\frac{L}{B\Delta_{t-i+1}}$. For every level i, each connected component in G_i that contains no new worms has a size of at least three. This is true since we only allow the worms to be routed along shortcut-free paths and therefore two worms can not block each other. Hence there are at most $g_i \leq \frac{m_i - c_i}{3}$ components with no new worms. Since every connected component of size s implies a probability of at most $(\frac{L}{B\Delta_{t-i+1}})^{s-1}$ that its edges represent collisions of worms we obtain a probability of at most

$$\left(\frac{L}{B\Delta_{t-i+1}}\right)^{((m_{i-1}-c_i)-g_i)} \leq \left(\frac{L}{B\Delta_{t-i+1}}\right)^{\frac{2(m_{i-1}-c_i)}{3}}$$

that these components are active. Note that we can improve this bound if we know that at least $k > 3$ pieces of paths in the collection are needed to obtain a directed cycle in G_i.

According to Section 2, each connected component in G_i that contains a new worm forms a tree. Furthermore each new worm lies in a different connected component. Therefore the probability that their edges represent collisions of worms is at most

$$\left(\frac{L}{B\Delta_{t-i+1}}\right)^{(\ell_i+c_i)-\ell_i}.$$

Altogether the probability that all collision pairs in level i are active given m_{i-1} and c_i is at most

$$\left(\frac{L}{B\Delta_{t-i+1}}\right)^{c_i+\frac{2(m_{i-1}-c_i)}{3}}.$$

Therefore the probability $P(t,k)$ that there exists an active embedding in $\mathcal{W}(t)$ is at most

$$n \sum_{\substack{\ell_1,\ldots,\ell_t \geq 0,\\ \sum_i \ell_i = k-1}} \prod_{i=1}^{t} \sum_{c_i=\ell_i}^{m_{i-1}} \binom{m_{i-1}}{c_i} \binom{c_i}{\ell_i} \tilde{C}^{\ell_i}_{t-i+1} (\ell_i + c_i)^{c_i-\ell_i} \cdot$$

$$(m_{i-1} - c_i)^{m_{i-1}-c_i} \left(\frac{L}{B\Delta_{t-i+1}}\right)^{c_i+\frac{2(m_{i-1}-c_i)}{3}}$$

The rest of the proof is similar to that in Section 2.1 with the difference that, for any constant $\gamma > 0$, we choose

$$k_0 = \frac{(2+\gamma)\log n}{\log\left(2 + \frac{B}{8\tilde{C}}\left(\frac{D}{L}+1\right)\right)} + 1$$

and

$$T \geq \frac{(2+\gamma)\log n}{\log\left(\frac{\tilde{C}}{\log n} + \log^{3/2} n + \frac{B}{2}\left(\frac{D}{L}+1\right)\right)} + \lceil \log k_0 \rceil .$$

Therefore the overall runtime is

$$\sum_{t=1}^{T} (\Delta_t + 2(D+L))$$

$$= O\left(\sum_{t=1}^{T}\left(D + L + \frac{L}{B}\left(\frac{\tilde{C}}{2^t} + \frac{\tilde{C}}{\log n} + \log^{3/2} n\right)\right)\right)$$

w.h.p., which is bounded by

$$O\left(\frac{L\cdot\tilde{C}}{B} + (\log_\alpha n + \log\log_\beta n)\left(\frac{L\log^{3/2} n}{B} + D + L\right)\right),$$

where $\alpha = \tilde{C} + B(\frac{D}{L}+1) + 2$ and $\beta = \alpha/\tilde{C} + 2$.

3.2 The Lower Bound

In this section we will prove the lower bound in Main Theorem 1.2. We use a path collection that consists of the following two types of subcollections.

- The first type consists of $n/6$ structures consisting of three paths of length D that are connected as shown in Figure 4.

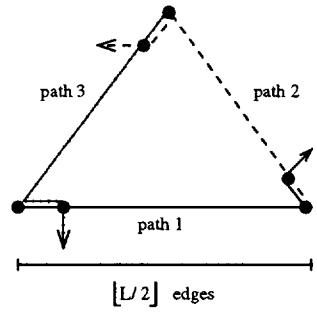

Figure 4: A type-1 structure.

- The second type consists of $n/(2\tilde{C})$ structures each consisting of $\tilde{C}$ identical paths of length D.

We assume that along each of these paths one worm of length $L \geq 2$ has to be sent. (Note that in case of $L = 1$ no cycles of colliding worms can occur, that is, we are in a situation of Main Theorems 1.1 and 1.3.)

We first want to compute how long it takes to route all worms in a type-1 structure. Consider an arbitrary round i of the trial-and-failure protocol. Suppose that in a given type-1 structure all three worms are still active. Then we want to calculate the probability that these three worms block each other in round i.

Suppose that $\Delta_i \geq L$. Let the worm traveling along path $j \in \{1,2,3\}$ be called w_j. Then it is easy to show that the probability that w_1, w_2, and w_3 collide at round i is at least $(\lfloor \frac{L}{2} \rfloor / (B\Delta_i))^2$. Therefore the probability that w_1, w_2, and w_3 collide for t rounds is at least

$$\prod_{i=1}^{t} \left(\frac{\lfloor L/2 \rfloor}{B(\Delta_i + L)}\right)^2$$

for any choice of $\Delta_1, \ldots, \Delta_t \geq 1$. Given a fixed $\Delta = \sum_{i=1}^{t} \Delta_i$ this product yields the smallest probability if $\Delta_i = \Delta/t$ for all $i \in \{1,\ldots,t\}$. Hence assume that all delay ranges are equal to $\bar{\Delta} = \Delta/t$. Since there are $n/6$ type-1 structures, and each structure has a probability of at least $(\frac{L}{3B(\bar{\Delta}+L)})^{2t}$ to have active worms after t rounds, the expected number

of type-1 structures that have active worms after t rounds is at least

$$\frac{n}{6}\left(\frac{L}{3B(\bar{\Delta}+L)}\right)^{2t} < 1 \quad \Leftrightarrow \quad t \geq \frac{\log(n/6)}{2\log\left(\frac{3B(\bar{\Delta}+L)}{L}\right)} .$$

Hence the expected number of rounds that are needed to route all worms is $\Omega(\log_{B(\bar{\Delta}/L+1)} n)$. In order to bound the time needed to route worms in the type-2 structures, we distinguish between the cases $\tilde{C} \geq 2^{\sqrt{\log n}}$ and $\tilde{C} \leq 2^{\sqrt{\log n}}$.

Case $\tilde{C} \leq 2^{\sqrt{\log n}}$:
Note that any routing protocol needs at least $\Omega(\frac{L\tilde{C}}{B}+D+L)$ steps to route all worms in a type-2 structure. Therefore the expected runtime of the protocol is at least

$$\Omega\left(\frac{L\tilde{C}}{B}+\log_{\frac{B\bar{\Delta}}{L}+2} n \cdot (\bar{\Delta}+D+L)\right)$$
$$= \Omega\left(\frac{L\tilde{C}}{B}+(\log_\alpha n+\log\log_\beta n)(D+L)\right),$$

where $\alpha = \tilde{C}+B(\frac{D}{L}+1)+2$ and $\beta = \alpha/\tilde{C}+2$.

Case $\tilde{C} \geq 2^{\sqrt{\log n}}$:
This case follows analogous to Section 2.

4 Summary and Open Problems

In case that wavelength conversion is not allowed we presented a very accurate analysis of the performance of a simple routing protocol for two types of all-optical routing elements. The question is, what the exact time bound for the runtime of the trial-and-failure protocol is if wavelength conversion is allowed. (The bound presented in [9] seems to be too weak compared to the bounds obtained in this paper.) Furthermore it would be interesting to consider cases in which only a few routers can convert wavelengths (see, e.g., [18]).

References

[1] A. Aggarwal, A. Bar-Noy, D. Coppersmith, R. Ramaswami, B. Schieber, M. Sudan. Efficient routing and scheduling algorithms for optical networks. In *Proc. of the 5th Ann. ACM-SIAM Symp. on Discrete Algorithms*, pp. 412-423, 1994.

[2] K. Bala, T.E. Stern. Algorithms for routing in a linear lightwave network. In *Proc. of INFOCOM*, pp. 1-9, 1991.

[3] R.A. Barry, P.A. Humblet. Bounds on the number of wavelengths needed in WDM networks. In *LEOS'92 Summer Topical Mtg. Digest*, pp. 114-127, 1992.

[4] R.A. Barry, P.A. Humblet. On the number of wavelengths and switches in all-optical networks. In *IEEE Trans. on Communications* **42**, pp. 583-591, 1994.

[5] T.A. Birks, D.O. Culverhouse, S.G. Farwell, P.St.J. Russel. 2 x 2 single-mode fiber routing switch. *Optics Letters* **21**, pp. 722-724, 1996.

[6] C. Brackett. Dense wavelength division multiplexing networks: Principles and applications. *IEEE J. Selected Areas in Comm.* **8**, pp. 373-380, August 1990.

[7] V.W.S. Chan. All-optical networks. *Scientific American*, pp. 56-59, Sept. 1995.

[8] N.K. Chung, K. Nosu, G. Winzer. *IEEE JSAC: Special Issue on Dense WDM Networks* **8**, 1990.

[9] R. Cypher, F. Meyer auf der Heide, C. Scheideler, B. Vöcking. Universal algorithms for store-and-forward and wormhole routing. In *28th Ann. ACM Symp. on Theory of Computing*, pp. 356-365, 1996.

[10] D.H.C. Du, R.J. Vetter. Distributed computing with high-speed optical networks. *IEEE Computer* **26**, pp. 8-18, 1993.

[11] R.D. Gitlin, Z. Haas. Field coding: A high-speed 'almost-all' optical interconnect. *25th Ann. Conf. Information Sci. Syst. CISS* (Baltimore, MD), March 20-22, 1991.

[12] P.E. Green. *Fiber-Optic Communication Networks*. Prentice Hall, 1992.

[13] Z. Haas. The 'staggering switch': an 'almost-all' optical packet switch. In *Tech. Dig. Optical Fiber Commun. Conf. OFC'92* (San José, CA), paper WH6.

[14] T. Hagerup, C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters* **33**, pp. 305-308, 1989/90.

[15] G.R. Hill, P.J. Chidgey et al. A transport network layer based on optical network elements. *J. of Lightwave Techn.* **11**, pp. 667-677, 1993.

[16] H.S. Hinton. Architectural considerations for photonic switching networks. *IEEE J. Selected Areas in Comm.* **6**, pp. 1209-1226, August 1988.

[17] M. Hofri. *Probabilistic Analysis of Algorithms: On Computing Methodologies for Computer Algorithms Performance Evaluation*, Springer Verlag, 1987.

[18] K.C. Lee, V.O.K. Li. A wavelength-convertible optical network. *J. of Lightwave Techn.* **11**, pp. 962-970, 1993.

[19] A. Lubotzky, R. Phillips, R. Sarnak. Ramanujan graphs. *Combinatorica* **8**(3), pp. 261-277, 1988.

[20] F. Matera, S. Settembre. All optical implementations of high capacity TDMA networks. *Fiber and Integrated Optics* **12**, pp. 173-186, 1993.

[21] F. Meyer auf der Heide, C. Scheideler. Communication in parallel systems. To appear at *SOFSEM'96*.

[22] R.K. Pankaj. *Architectures for Linear Lightwave Networks*. PhD thesis, MIT, 1992.

[23] S. Personick. Review of fundamentals of optical fiber systems. *IEEE J. Selected Areas in Comm.* **3**, pp. 373-380, April 1983.

[24] G.R.Pieris, G.H. Sasaki. A linear lightwave Benes network. *IEEE/ACM Trans. on Networking* **1**, 1993.

[25] P. Raghavan, E. Upfal. Efficient routing in all-optical networks. In *Proc. of the 26th Ann. ACM Symp. on Theory of Computing*, 1994.

[26] R. Ramaswami. Multi-wavelength lightwave networks for computer communication. *IEEE Communications Magazine* **31**, pp. 78-88, 1993.

[27] R. Ramaswami, K.N. Sivarajan. Routing and wavelength assignment in all-optical networks. *IEEE/ACM Trans. on Networking* **3**(5), pp. 489-500, 1995.

[28] C. Scheideler, B. Vöcking. Universal continuous routing strategies. In *Proc. of the 8th Ann. ACM Symp. on Parallel Algorithms and Architectures*, pp. 142-151, 1996.

[29] J.P. Schmidt, A. Siegel, A. Srinivasan. Chernoff-Hoeffding bounds for applications with limited independence. *SIAM J. Disc. Math.* **8**(2), pp. 223-250, 1995.