A Constant Approximation Algorithm for Scheduling Packets on Line Networks

Guy Even[1], Moti Medina[2], and Adi Rosén[3]

1 **Tel Aviv University**
guy@eng.tau.ac.il
2 **MPI for Informatics**
medinamo@mpi-inf.mpg.de
3 **CNRS and Université Paris Diderot**
adiro@liafa.univ-paris-diderot.fr

Abstract

In this paper we improve the approximation ratio for the problem of maximizing the throughput of line networks with bounded buffers. Each node in the network has a local buffer of bounded size B, and each edge (or link) can transmit a limited number c of packets in every time unit. The input to the problem consists of a set of packet requests, each defined by a source node, a destination node, and a release times. We denote by n the size of the network, by B the size of a node buffer, and by c the capacity of the links. A solution for this problem is a schedule that delivers (some of the) packets to their destinations without violating the capacity constraints of the network (buffers or edges). Our goal is to design an algorithm that computes a schedule that maximizes the number of packets that arrive to their respective destinations.

We give a randomized approximation algorithm with constant approximation ratio for the case where the buffer-size to link-capacity ratio, B/c, is constant. This improves over the previously best result of $O(\log^* n)$ [**?**]. Our improvement is based on a new combinatorial lemma that we prove, stating, roughly speaking, that if packets are allowed to stay put in buffers only a limited number of time steps, $2d$, where d is the longest source-destination distance, then the optimal solution is decreased by only a constant factor. This claim was not previously known in the integral (unsplitable, zero-one) case, and may find additional applications for routing and scheduling algorithms.

While we are not able to give the same improvement for the related problem when packets have hard deadlines, our algorithm does support "soft deadlines". That is, if packets have deadlines, we achieve a constant approximation ratio when the produced solution is allowed to miss deadlines by at most $\log n$ time units.

1998 ACM Subject Classification F.2 ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY

Keywords and phrases Approximation algorithms, linear programming, randomized rounding, Admission control, Packet scheduling

1 Introduction

In this paper we give an approximation algorithm with an improved approximation ratio for a network-scheduling problem which has been studied in numerous previous works in a number of variants (cf. [**?**, **?**, **?**, **?**, **?**, **?**]). The problem consists of a directed line network over nodes $\{0, \ldots, n-1\}$, where each node i can send packets to node $i+1$, and can also store packets in a local buffer. The maximum number of packets that can be sent in a single time unit over a given link is denoted by c, and the number of packets each node can store in any given time is denoted by B. An instance for the problem is further defined by a set

M of packets $r_i = (a_i, b_i, t_i)$, $1 \leq i \leq M$, where a_i is the source node of the packet, b_i is its destination node, and $t_i \geq 1$ is the release time of the packet at vertex a_i. The goal is to maximize the number of packets that reach their respective destinations without violating the links or the buffers capacities. See Section **??** for a formal definition of the problem.

We present a randomized approximation algorithm for that problem, which has a *constant* approximation ratio for the case that the ratio B/c is a constant., which improves upon the previous $O(\log^* n)$ approximation ratio given in [**?**, Theorem 3]. While this constant approximation result does not hold for the variant of the problem where packets have deadlines, our algorithm does provide a constant-approximation solution that abides to "soft deadlines". That is, in that solution each packet is delivered at most $\log n$ time units past its deadline.

Our algorithm is based on a novel combinatorial lemma (Lemma **??**) which states the following. Consider a set of packets such that all source-destination distances are bounded from above by some d. The throughput of an optimal solution in which every packet r_i must reach its destination no later than time $t_i + 2d$ is an $\Omega(B/c)$-fraction of the unrestricted optimal throughput. This lemma plays a crucial role in our algorithm, and we believe that it may find additional application for scheduling and routing algorithms in networks. We emphasize that the fractional version of a similar property, i.e., when packets are splitable and one accrues a benefit also from the delivery of partial packets, presented first in [**?**], does not imply the integral version that we prove here.

AR: check is this paragraph is correct at all and if it is interesting, and if this was not done before, and why we do it this way

Another interesting ingredient of our algorithm and its analysis is a somewhat different usual way to performa and analyze randomized rounding for multicommodity fractional solutions, applicable to the case that we consider in the present paper. Usually one rounds the fractional solutions and shows that with high probability no edge is overloaded in the rounded solution, hence we obtain with high probability a feasible integral solution. Here we instead filter out any packet (demand) that passes through a link that is overloaded. We show that the probability that any given packet remains in the solution is high, and hence our integral feasible solution is of cardinality close to the fractional one.

AR: ——————————————————————————————

We emphasize that the problem studied here, namely, maximizing the throughput on a network with bounded buffers, has resisted substantial efforts in its (more realistic) distributed, online setting, even for the simple network of a directed line. Indeed, even the question whether or not there exists a constant competitive online distributed algorithm for that problem on the line network remains unanswered at this point. We therefore study here the less realistic, offline, setting , with the hope that results and ideas from this setting will contribute to progress on the online problem.

AR: verify, check, complete

Furthermore, to the best of our knowledge, there is only a single problem known to be $O(\log^* n)$-approximable but not $O(1)$-approximable (unless ....), namely ..... The state of the art for the problem we consider here was since a number of years now that its best efficient approximation ratio was $O(\log^* n)$ It is therefore of interest to resolve, as we do in the present paper, the question whether the problem we consider here has a "barrier" at $\log^* n$. We note that to the best of our knowledge it is not known if the problem is NP-hard or not. **AR**: ——
——————————————————————————————

Related Work. The problem of scheduling packets so as to maximize the throughput (i.e., maximize the number of packets that reach their destinations) in a network with bounded buffers was first considered in [**?**], where this problem is studied for various types of networks in the distributed (and, hence, online) setting. The results in this paper, even for the simple network of a directed line, were far from tight but no substantial progress has been made since on the realistic, distributed and online, setting. This has motivated the study of this problem in easier settings, as a first step towards solving the realistic, possibly applicable, scneario.

Angelov et al. [] give centralized online randomized algorithms for the line network, achieving an $O(\log^3)$-competitive ratio. Azar and Zachut [] improved the randomized competitive ratio to $O(\log^2 n)$ which was later improved by Even and Medina [**?**] to $O(\log n)$. A deterministic $O(\log^5 n)$-competitive algorithm was given in [], which was later improved in [] to $O(\log n)$ if buffer and link capacities are not very small (not smaller than 5). Somewhat better results are known for the related problem of information gathering, i.e., scheduling on the line when all destinations are the end-node of the line.

To the best of our knowledge, it is not known if the problem we consider here is NP-hard. The related problem of maximizing the throughput when packets have deadlines (i.e., a packet is counted towards the quality of the solution only if it arrives to its destination before a known deadline) is known to be NP-hard [**?**]. This problem, in the exact setting that we consider here, was studied in [**?**] where it is shown to have a $O(\log^* n)$-approximation randomized algorithm. This result immediately gives an $O(\log^* n)$-approximation randomized algorithm for the problem we study in the present paper [**?**].

2 Preliminaries

2.1 Model and problem statement

We consider the standard model of synchronous store-and-forward packet routing networks [**?**, **?**, **?**]. The network is modeled by a directed path over n vertices. Namely, the network is a directed graph $G = (V, E)$, where $V = \{0, \ldots, (n-1)\}$ and there is a directed edge from vertex u to vertex v if $v = u + 1$. The network resources are specified by two positive integer parameters B and c that describe, respectively, the local buffer capacity of every vertex and the capacity of every edge. In every time step, at most B packets can be stored in the local buffer of each vertex, and at most c packets can be transmitted along each edge.

The input consists of a set of M packet requests $R = \{r_i\}_{i=1}^{M}$. A packet request is specified by a 3-tuple $r_i = (a_i, b_i, t_i)$, where $a_i \in V$ is the *source node* of the packet, $b_i \in V$ is its *destination node*, and $t_i \in \mathbb{N}$ is the time of arrival of the request. Note that $b_i > a_i$, and r_i is ready to leave a_i in time step t_i.

A solution is a schedule S. For each request r_i, the schedule S specifies a sequence s_i of transitions that packet r_i undergoes. A *rejected* request r_i is simply discarded at time t_i, and no further treatment is required (i.e., $s_i = \{reject\}$). An *accepted* request r_i is delivered from a_i to b_i by a sequence s_i of actions, where each action is either "store" or "forward". Consider the packet of request r_i. Suppose that in time t the packet is in vertex v. A store action means that the packet is stored in the buffer of v, and will still be in vertex v in time step $t + 1$. A forward action means that the packet is transmitted to vertex $v + 1$, and will be in vertex $v + 1$ in time step $t + 1$. The packet of request r_i reaches its destination b_i after exactly $b_i - a_i$ forward steps. Once a packet reaches its destination, it is removed from the network and it no longer consumes any of the network's resources.

A schedule must satisfy the following constraints:

1. The *buffer capacity constraint* asserts that at any time step t, and in every vertex v, at most B packets are stored in v's buffer.
2. The *link capacity constraint* asserts that at any step t, at most c packets can be transmitted along each edge.

The *throughput* of a schedule S is the number of accepted requests. We denote the throughput of a schedule S by $|S|$. As opposed to online algorithms, there is no point in injecting a packet to the network unless it reaches its destination. Namely, a packet that is not rejected and does not reach its destination only consumes network resources without any benefit. Hence, without loss of we assume that every packet that is dropped before reaching its designation is rejected at its source node at its release time.

We consider the offline optimization problem of finding a schedule that maximizes the throughput. We propose a centralized constant-ratio approximation algorithm. By *offline* we mean that the algorithm receives all requests in advance[1]. By *centralized* we mean that all the requests are known in one location where the algorithm is executed. Let $\mathsf{opt}(R)$ denote a schedule of maximum throughput for the set of requests R. Let $\mathsf{alg}(R)$ denote the schedule computed by alg on input R. We say that the approximation ratio of a scheduling algorithm alg is c if

$$\forall R : |\mathsf{alg}(R)| \geq c \cdot |\mathsf{opt}(R)|.$$

The Max-Pkt-Line Problem. The problem of maximum throughput schedule of packet requests on directed line (Max-Pkt-Line) is defined as follows. The input of consists of: n - the size of the network, B - node buffer capacities, c - link capacities, and M packet requests $\{r_i\}_{i=1}^{M}$. The output is a schedule S. The goal is to maximize the throughput of S.

2.2 Path Packing in a uni-directed 2D-Grid

In this section we define a problem of maximum cardinality path packing in a two-dimensional uni-directed grid (Max-Path-Grid). This problem is equivalent to maximum throughput scheduling of packet requests on a directed line, and was used for that purpose in previous work, where the formal reduction is also presented [**?**, **?**, **?**, **?**]. For completeness, this reduction is give in Appendix **??**. As the two problems problems are equivalent, we use in the sequel terminology from both problems interchangeably.

The grid, denoted by $G^{st} = (V^{st}, E^{st})$, is an infinite directed acyclic graph. The vertex set V^{st} equals $V \times \mathbb{N}$, where $V = \{0, 1, \ldots, (n-1)\}$. Note that we use the first coordinate (that corresponds to vertices in V) for the y-axis and the second coordinate (that corresponds to time steps) for the x-axis. The edge set consists of horizontal edges (also called store edges) directed to the right and vertical edge (also called forward edges) directed upwards. The capacity of vertical edges is c and the capacity of horizontal edges is B. We often refer to G^{st} as the space-time grid because the x-axis is related to time and the y-axis corresponds to the vertices in V.

A *path request* in the grid is a tuple $r^{st} = (a_i, t_i, b_i)$, where $a_i, b_i \in V$ and $t_i \in \mathbb{N}$. The request is for a path that starts in node (a_i, t_i) and ends in any node in the row of b_i (i.e., the end of the path can be any node (b_i, t), where $t \geq t_i$).

[1] The number of requests M is finite and known in the offline setting. This is not the case in the online setting in which the number of requests is not known in advance and may be unbounded.

A *packing* is a set of paths S^{st} that abides the capacity constraints. For every grid edge e, the number of paths in S^{st} that contain e is not greater than the capacity of e.

Given a set of path requests $R^{st} = \{r_i^{st}\}_{i=1}^{M}$, the goal in the Max-Path-Grid problem is to find a packing S^{st} with the largest cardinality. (Each path in S^{st} serves a distinct path request.)

Multi-Commodity Flows (MCFs). Our use of path packing problems gives rise to *fractional* relaxations of that problem, namely to multi-commodity flows (MCFs) with unit demands on un-directional grids. The definitions and terminology of MCFs appears in Appendix **??**.

2.3 Tiling, Classification, and Sketch Graphs

To define our algorithm we make use of partiniongs of the space-time grid described above into sub-grids. We define here the notions we use for this purpose. In this section we focus on the case of unit capacities, namely, $B = c = 1$. An extension to other values of B and c can be found [**?**].

Tiling. *Tiling* is a partitioning of the two-dimensional space-time grid (in short, grid) into squares, called *tiles*. Two parameters specify a tiling: the side length k of the squares and the shifting (φ_x, φ_y) of the squares. The shifting refers to the x- and y-coordinates of the bottom left corner of the tiles modulo k. Thus, the tile $T_{i,j}$ is the subset of the grid vertices defined by

$$T_{i,j} \triangleq \{(v,t) \in V \times \mathbb{N} \mid ik \leq v - \varphi_y < (i+1)k \text{ and } jk \leq t - \varphi_x < (j+1)k\},$$

where φ_x and φ_y denote the horizontal and vertical shifting, respectively. We consider two possible shifts for each axis, namely, $\varphi_x, \varphi_y \in \{0, k/2\}$.

Quadrants and Classification. Consider a tile T. Let (x', y') denote the lower left corner (i.e., south-west corner) of T. The *south-west quadrant* of T is the set of vertices (x, y) such that $x' \leq x \leq x' + k/2$ and $y' \leq y \leq y' + k/2$.

For every vertex (x, y) in the grid, there exists exactly one shifting $(\varphi_x, \varphi_y) \in \{0, k/2\}^2$ such that (x, y) falls in the south-west (SW) quadrant of a tile. Fix the tile side length k. We define a *class* for every shifting (φ_x, φ_y). The class that corresponds to the shifting (φ_x, φ_y) consists of all the path requests r_i^{st} whose origin (a_i, t_i) belongs to a SW quadrant of a tile in the tiling that uses the shifting (φ_x, φ_y).

Sketch graph and paths. Consider a fixed tiling. The *sketch graph* is the graph obtained from the grid after coalescing each tile into a single node. There is a directed edge (s_1, s_2) between two tiles s_1, s_2 in the sketch graph if there is a directed edge $(\alpha, \beta) \in E^{st}$ such that $\alpha \in s_1$ and $\beta \in s_2$. Let p^s denote the projection of a path p in the grid to the sketch graph. We refer to p^s as the *sketch path* corresponding to p. Note that the length of p^s is at most $\lceil |p|/k \rceil + 1$.

3 Outline of our Algorithm

For the sake of simplicity we focus hereafter on the case of unit capacities, namely $B = c = 1$. Extension to non-unit capacities are discussed in Section **??**.

Packet requests are categorized into three categories: short, medium, and long, according to the source-destination distance of each packet. A separate approximation algorithm is executed for each category. The algorithm returns a highest throughput solution among the solutions computed for the three categories.

Notation. Two thresholds are used for defining short, medium, and long requests.

$$\ell_M \triangleq 3\ln n, \qquad \ell_S \triangleq 3 \cdot \ln(\ell_M) = 3 \cdot \ln(3\ln n)$$

▶ **Definition 1.** A request r_i is a *short* request if $b_i - a_i \leq \ell_S$. A request r_i is a *medium* request if $\ell_S < b_i - a_i \leq \ell_M$. A request r_i is a *long* request if $b_i - a_i > \ell_M$.

We use a deterministic algorithm for the class of short packets, and in Theorem **??** we prove that this deterministic algorithm achieves a constant approximation ratio. We use a randomized algorithm for each of the classes of medium and long packets; in Theorem **??** we prove that this randomized algorithm achieves a constant approximation ratio in expectation for each of these classes. Thus, we obtain the following corollary.

▶ **Corollary 2** (Main Result). *If $B = c = 1$, then there exists a randomized approximation algorithm for the Max-Pkt-Line problem that achieves a constant approximation ratio in expectation.*

In Section **??** we discuss non-unit capacities, give the approximation ratio for this case and show that we achieve a constant approximation ratio as long as the ratio c/B is constant.

4 Approximation Algorithm for Short Packets

In this section we present a constant ratio deterministic approximation algorithm for short packets. This algorithm, which is key to achieving the results of the present paper, makes use of a new combinatorial lemma that we prove in the next subsection, stating, roughly speaking, that if packets from a given set of packets are allowed to stay put in buffers (i.e., use horizontal edges in the grid) only a limited number of time steps, $2d$ (where d is the longest source-destination distance in the set of packets), then the optimal solution is decreased by only a constant factor. We believe that this lemma may find additional application is future work on routing and scheduling problems.

4.1 Bounding Path Lengths in the Grid

In this section we prove that bounding from above the slack (i.e., the number of horizontal edges along a path) incurs only a small reduction in the throughput. Previously known bounds along these lines hold only for fractional solutions [**?**], while we present here the first such claim for integral schedules.

Let R_d denote a set of packet requests r_i, $i \geq 1$, such that $b_i - a_i \leq d$ for any i. Consider the paths in the space-time grid that are allocated to the accepted requests. We prove that restricting the path lengths to $2d$ decreases both the optimal fractional and the optimal *integral* throughput only by a multiplicative factor of $O(c/B)$. We note that if the ratio B/c is a constant, then we are guarantees an optimal solution which is only a (different) constant away from the unrestricted optimal solution.

Notation. For a single commodity acyclic flow f_i, let $p_{\max}(f_i)$ denote the diameter of the support of f_i (i.e., length of longest path[2]). For an MCF $F = \{f_i\}_{i \in I}$, let $p_{\max}(F) \triangleq \max_{i \in I} p_{\max}(f_i)$. Let $F^*_{frac}(R)$ (respectively, $F^*_{int}(R)$) denote a maximum throughput fractional (resp., integral) MCF with respect to the set of requests R. Similarly, let $F^*_{frac}(R \mid p_{\max} < d')$ (respectively, $F^*_{int}(R \mid p_{\max} < d')$) denote a maximum throughput fractional (resp., integral) MCF with respect to the set of requests R subject to the additional constraint that the maximum path length is at most d'.

▶ **Lemma 3.**

$$F^*_{frac}(R_d \mid p_{\max} \leq 2d) \geq \frac{c}{B+2c} \cdot F^*_{frac}(R_d)$$
$$F^*_{int}(R_d \mid p_{\max} \leq 2d) \geq \frac{c}{2(B+c)} \cdot F^*_{int}(R_d).$$

Proof. Partition the space-time grid into *slabs* S_j of "width" d. Slab S_j contains the vertices (v,k), where $k \in [(j-1) \cdot d, j \cdot d]$, $j \geq 1$. We refer to vertices of the form (v, jd) as the *boundary* of S_j. Note that if $v - u \leq d$, then the forward-only vertical path from (u, jd) to $(v, jd + (u - v))$ is contained in slab S_{j+1}.

We begin with the fractional case. Let $f^* = F^*_{frac}(R_d)$ denote an optimal fractional solution for R_d. Consider request r_i and the corresponding single commodity flow f^*_i in f^*. Decompose f^*_i to flow-paths $\{p_\ell\}_\ell$. For each flow-path p_ℓ in f^*_i, let p'_ℓ denote the prefix of p_ℓ till it reaches the boundary of a slab. Note that $p'_\ell = p_\ell$ if p_ℓ is confined to a single slab. If $p'_\ell \subsetneq p_\ell$, then let (v, jd) denote the last vertex of p'_ℓ. Namely, the path p'_ℓ begins in $(a_i, t_i) \in S_j$ and ends in (v, jd). Let q''_ℓ denote the forward-only path from (v, jd) to $(b_i, jd + (b_i - v))$. (If $p'_\ell = p_\ell$, then q''_ℓ is an empty path.) Note that q''_ℓ is confined to the slab S_{j+1}. We refer to the vertex (v, jd) in the intersection of p'_ℓ and q''_ℓ as the *boundary* vertex. Let g_i denote the fractional single commodity flow for request r_i obtained by adding the concatenated flow-paths $q_\ell \triangleq p'_\ell \circ q''_\ell$ each with the flow amount of f^*_i along p_ℓ. Define the MCF g by $g(e) \triangleq \sum_{i \in I} g_i(e)$. For every edge e, part of the flow $g(e)$ is due to prefixes p'_ℓ, and the remaining flow is due to suffixes q''_ℓ. We denote the part due to prefixes by $g_{pre}(e)$ and refer to it as the prefix-flow. We denote the part due to suffixes by $g_{suf}(e)$ and refer to it as the suffix-flow. By definition, $g(e) = g_{pre}(e) + g_{suf}(e)$.

The support of g_i is contained in the union of two consecutive slabs. Hence, the diameter of the support of g_i is bounded by $2d$. Hence $p_{\max}(g) \leq 2d$.

Clearly, $|g_i| = |f^*_i|$ and hence $|g| = |f^*|$. Set $\rho = c/(B+2c)$. To complete the proof, it suffices to prove that $\rho \cdot g$ satisfies the capacity constraints. Indeed, for a "store" edge $e = (v,t) \to (v, t+1)$, we have $g_{suf}(e) = 0$ and $g_{pre}(e) \leq f^*(e) \leq B$. For a "forward" edge $e = (v,t) \to (v+1, t+1)$ we have: $g_{pre}(e) \leq f^*(e) \leq c$. On the other hand, $g_{suf}(e) \leq B + c$. The reason is as follows. All the suffix-flow along e starts in the same boundary vertex (u, jd) below e. The amount of flow forwarded by (u, jd) is bounded by the amount of incoming flow, which is bounded by $B + c$. This completes the proof of the fractional case.

We now prove the integral case. The proof is a variation of the proof for the fractional case in which the supports of pre-flows and suffix-flows are disjoint. Namely, one alternates between slabs that support prefix-flow and slabs that support suffix-flow.

In the integral case, each accepted request r_i is allocated a single path p_i, and the allocated paths satisfy the capacity constraints. As in the fractional case, let $q_i \triangleq p'_i \circ q''_i$, where p'_i is the prefix of p_i till a boundary vertex (v, jd), and q''_i is a forward-only path. We need to

[2] Without loss of generality, we may assume that each single commodity flow f_i is acyclic.

prove that there exists a subset of at least $c/(2(B+c))$ of the paths $\{q_i\}_i$ that satisfy the capacity constraints. This subset is constructed in two steps.

First, partition the requests into "even" and "odd" requests according to the parity of the slab that contains their origin (a_i, t_i). (The parity of request r_i is simply the parity of $\lceil t_i/d \rceil$.) Pick a part that has at least half of the accepted requests in $F^*_{int}(R_d)$. We only keep accepted requests whose origin belong to even slabs.

In the second step, we consider all boundary vertices $(v, j \cdot d)$. For each boundary vertex, we keep up to c paths that traverse it, and delete the remaining paths if such paths exist. In the second step, again, at least a $c/(B+c)$ fraction of the paths survive. It follows that altogether at least $c/(2(B+c))$ of the paths survive.

We claim that the remaining paths satisfy the capacity constraints. Note that prefixes are restricted to even slabs, and suffixes are restricted to odd slabs. Thus, intersections, if any, are between two prefixes or two suffixes. Prefixes satisfy the capacity constraints because they are prefixes of $F^*_{int}(R_d)$. Suffixes satisfy the capacity constraints because if two suffixes intersect, then they start in the same boundary vertex. However, at most c paths emanating from every boundary vertex survive. Hence, the surviving paths satisfy the capacity constraints, as required. This completes the proof of the lemma. ◀

We note that if the ratio B/c is a constant, then Lemma **??** guarantees an optimal solution which is only a (different) constant away from the unrestricted optimal solution.

▶ Remark.

- One may compute a maximum throughput fractional solution with bounded diameter using linear programming. This is true because the constraint $p_{\max}(f_i) \le d'$ is a linear constraint and can be imposed by a polynomial number of inequalities (i.e, polynomial in n and d'). For example, one can construct a product network with $(d'+1)$ layers, and solve the MCF problem over this product graph.
- Note that in the proof of Lemma **??**, $|q_\ell| \le |p_\ell|$, namely, the length of each new path q_ℓ that replaces an original flow-path p_ℓ is not greater than the length of p_ℓ.
- The proof of Lemma **??** is algorithmic. Given any MCF F, one can compute in polynomial time a diameter-bounded MCF F' such that $|F'|$ is at least a fraction of $|F|$ (the value of the fraction depends on whether the MCF is integral or fractional).
- In [**?**] it is proved that a constant fraction of the fractional throughput can be obtained with diameter $2d(1+\frac{B}{c})$. Formally,

$$F^*_{frac}\left(R_d \mid p_{\max} \le 2d\left(1+\frac{2B}{c}\right)\right) \ge \frac{1}{2}\cdot\left(1-\frac{1}{e}\right)\cdot F^*_{frac}(R_d).$$

4.2 The Algorithm for Short Packets

Short requests are further partitioned into four classes, defined as follows. Consider four tilings each with side length $k \triangleq 4\ell_S$ and horizontal and vertical shifts in $\varphi_x, \varphi_y \in \{0, k/2\}$ [3]. The four possible shiftings define four classes; in each class, all the sources nodes reside in the SW quadrant of tiles. We say that a path p_i from (a_i, t_i) to the row of b_i is *confined to a tile* if p_i is contained in one tile. We bound the path lengths by $2\ell_S$ so that, within each class, every path is confined to the tile that contains its origin. **AR**: It is no good to talk about $c = B = 1$ here since a few lines later we talk about general c, B, so it does not combine. On the other hand I think that k above is defined given $B = c = 1$... So it does not

[3] Recall that $\ell_M \triangleq 3\ln n$ and $\ell_S \triangleq 3 \cdot \ln(\ell_M) = 3 \cdot \ln(3\ln n)$.

work. I would delete the following lines, but this needs to be taken care of at some point. Please comment out this comment, but don't delete from the file so that we deal with it for a future version. — Lemma ref(lemma:bounded path length) used with $B = c = 1$, implies that an optimal solution.—

If we restrict the paths lengths to be at most $2\ell_S$, then by exhaustive search, it is possible to efficiently compute a maximum throughput solution for each class. The algorithm computes an optimal (bounded path length) solution for each class, and returns a highest throughput solution among the four solutions.

The polynomial running time of the exhaustive search algorithm per class is based on two observations.

▶ **Observation 4.** *A path of length at most $k/2 = 2\ell_S$ that begins in the SW quadrant of tile T is confined to T.*

Proof. The tile side length equals $k = 4\ell_S$. If the origin of a request is in the SW quadrant of a tile and the path length is at most $2\ell_S = k/2$, then the end of the path belongs to the same tile. ◀

▶ **Observation 5.** *An optimal solution for each class in which paths are confined to their origin tile is computable in time polynomial in n and M.*

Proof. **AR**: it is a mess here, since above we apply the path length lemma for capacities 1 but here we talk about general c and B. I didn't touch. Please comment out this comment, but don't delete from the file so that we deal with it for a future version. Fix a tile T. Let X denote the set of short requests in T. Let Y denote the set of paths in T. It suffices for exhaustive search to consider all functions $f : X \to Y$. The number of short requests that originate in the SW quadrant of T is bounded by $(B + c)k^2$ (there are less than k^2 possible origins in T, and each origin can serve as the origin of at most $(B + c)$ requests). Hence $|X| < 2k^2$. For each request there are less than $\binom{2k}{k} < 2^{2k}$ possible paths within the tile. Hence, $|Y| < 2^{2k}$. This implies that the number of combinations of paths that exhaustive search needs to consider is bounded by

$$|Y|^{|X|} < \left(2^{2k}\right)^{(2k^2)} = o(n).$$

The number of tiles that contain a request is bounded by the number of requests M. Hence the running time of the algorithm for short requests is polynomial in n and M. ◀

▶ **Theorem 6.** *The approximation ratio of the algorithm for short requests is $\frac{c}{8(B+c)} = \frac{1}{16}$.*

Proof. The short requests are partitioned to 4 classes. For each class and tile, the exhaustive algorithm computes a $c/(2(B + c))$-approximation (by the integral version of Lemma **??**) because it bounds the path length by no less than twice the source-destination vertical distance. ◀

▶ Remark. The algorithm for short requests can handle requests with deadlines, and achieve the same performance while respecting hard deadlines, because it employs exhaustive search.

5 Approximation Algorithm for Medium & Long Requests

We use the same algorithm for the two classes of medium and long requests, the only difference being some parameters of the algorithm. As indicated above, we consider at this point the case of unit capacities ($B = c = 1$). We further note that the approximation ratio of the algorithm for these classes is with respect to the optimal *fractional* solution.

Notation. Let $R_{d_{\min},d_{\max}}$ denote the set of packet requests whose source-to-destination distance greater than $d_{\min}$ and at most $d_{\max}$. Formally, $R_{d_{\min},d_{\max}} \triangleq \{r_i \mid d_{\min} < b_i - a_i \leq d_{\max}\}$.

Parametrization. When applied to medium requests we use the parameter $d_{\max} = \ell_M$ and $d_{\min} = \ell_S$. When applied to long requests the parameters are $d_{\max} = n$ and $d_{\min} = \ell_M$. Note that these parameters satisfy $d_{\min} = 3 \cdot \ln d_{\max}$.

5.1 Outline of the Algorithm for $R_{d_{\min},d_{\max}}$

The algorithm for $R_{d_{\min},d_{\max}}$ proceeds as follows. (To simplify notation, we abbreviate $R_{d_{\min},d_{\max}}$ by R.)

1. Reduce the packet requests in R to path requests R^{st} over the space-time graph G^{st}.
2. Compute a maximum throughput fractional MCF $F \triangleq \{f_i\}_{r_i \in R^{st}}$ with edge capacities $\tilde{c}(e) = \lambda$ (for $\lambda = \beta(1)/6 \approx 1/15.54$) [4] and bounded diameter $p_{\max}(F) \leq 2d_{\max}$. We remark that this MCF can be computed in time polynomial in n - the number of nodes and M - the number of requests. The reason is that $d_{\max} \leq n$. Hence, for every request one needs to consider at most an $n \times n$ subgrid.
3. Partition R to 4 classes $\{R^j\}_{j=1}^4$ according to the quadrant that contains the source node in a $k \times k$ tiling, where $k \triangleq 2d_{\min} = 6 \ln d_{\max}$. Pick a class R^j such that the throughput of F restricted to R^j is at least a quarter of the throughput of F, i.e., $|F(R^j)| \geq |F|/4$.
4. For each request $r_i \in R^j$, apply randomized rounding independently to f_i as described in Appendix **??**. The outcome of randomized rounding per request $r_i \in R^j$ is either "reject" or a path p_i in G^{st}. Let $R_{rnd} \subseteq R^j$ denote the subset of requests r_i that are assigned a path p_i by the randomized rounding procedure.
5. Let $R_{fltr} \subseteq R_{rnd}$ denote the requests that remain after applying filtering (described in Section **??**).
6. Let $R_{quad} \subseteq R_{fltr}$ denote the requests for which routing in first quadrant is successful (as described in Section **??**).
7. Complete the path of each request in R_{quad} by applying crossbar routing (as described in Section **??**).

5.2 Filtering

Notation. Let e denote an edge in the space-time grid G^{st}. Let e^s denote an edge in the sketch graph. We view e^s also as the set of edges in G^{st} that cross the tile edge that corresponds to the sketch graph edge e^s. The path p_i is a random variable that denotes the path, if any, that is chosen for request r_i by the randomized rounding procedure. For a path p and an edge e let $\mathbb{1}_p(e)$ denote the 0-1 indicator function that equals 1 iff $e \in p$.

The set of filtered requests R_{fltr} is defined as follows (recall that $\lambda = \beta(1)/6$).

▶ **Definition 7.** A request $r_i \in R_{fltr}$ if and only if r_i is accepted by the randomized rounding procedure, and for every sketch-edge e^s in the sketch-path p_i^s it holds that $\sum_i \mathbb{1}_{p_i^s}(e^s) \leq 2\lambda \cdot k$.

▶ **Claim 8** (Proof in Appendix **??**). $\mathbf{E}[|R_{fltr}|] \geq \left(1 - O(\frac{1}{k})\right) \cdot \mathbf{E}[|R_{rnd}|]$.

[4] The function β is defined in Definition **??**.

5.3 Routing in the First Quadrant

In this section, we deal with the problem of evicting as many requests as possible from their origin quadrant to the boundary of the origin quadrant.

▶ Remark. Because $k/2 \leq d_{\min}$ every request that starts in a SW quadrant of a tile must reach the boundary (i.e. top or right side) of the quadrant before it can reach its destination.

The maximum flow algorithm. Consider a tile T. Let X denote set of requests r_i whose source (a_i, t_i) is in the south-west quadrant of T. We say that a subset $X' \subseteq X$ is *quadrant feasible* (in short, feasible) if it satisfies the following condition: There exists a set of edge disjoint paths $\{q_i \mid r_i \in X'\}$, where each path q_i starts in the source (a_i, t_i) of r_i and ends in the top or right side of the SW quadrant of T.

We employ a maximum-flow algorithm to solve the following problem.
Input: A set of requests X whose source is in the SW quadrant of T.
Goal: Compute a maximum cardinality quadrant-feasible subset $X' \subseteq X$.

The algorithm is simply a maximum-flow algorithm over the following network, denoted by $N(X)$. Augment the quadrant with a super source $\tilde{s}$, a super sink $\tilde{t}$. The super source $\tilde{s}$ is connected to every source (a_i, t_i) (of a request $r_i \in X$) with a unit capacity directed edge. (If γ requests share the same source, then the capacity of the edge is γ.) There is a unit capacity edge from every vertex in the top side and right side of the SW quadrant of T to the super sink $\tilde{t}$. All the grid edges are assigned unit capacities. Compute an integral maximum flow in the network. Decompose the flow to unit flow paths. These flow paths are the paths that are allocated to the requests in X'.

Analysis. Fix a tile T and let $R_T \subseteq R_{\mathit{fltr}}$ denote the set of requests in R_{fltr} whose source vertex is in the SW quadrant of T. Let $R'_T \subseteq R_T$ denote the quadrant-feasible subset of maximum cardinality computed by the max-flow algorithm. Let $R_{quad} = \bigcup_T R'_T$.

We now prove the following theorem that relates $|R'_T|$ to $|R_T|$.

▶ **Theorem 9.** *[**?**, **?**] Let τ denote the probability space induced by the randomized rounding procedure.* $\mathbf{E}_\tau[|R_{quad}|] \geq 0.93 \cdot \mathbf{E}_\tau[|R_{\mathit{fltr}}|]$.

Routing within the first tile (see Section **??**) requires a stronger limit on the number of requests that emanate from each side of the quadrant.

▶ **Corollary 10.** *One can further limit the quadrant-feasible set of requests in R_{quad} so that at most $k/3$ paths reach each side of the quadrant. This restriction further reduces the expected throughput by a factor of* $2/3$.

Proof. The sum of the capacities of the edges emanating from a side of the quadrant is $k/2$. Limiting the number of paths to $k/3$ reduces the throughput by at most a factor of $2/3$. ◀

5.4 Detailed Routing

In this section we deal with computing paths for requests $r_i \in R_{quad}$ starting from the boundary of the SW quadrant that contains the source (a_i, t_i) till the destination row b_i. These paths are concatenated to the paths computed in the first quadrant to obtain the *final* paths of the accepted requests. Detailed routing is based on the following components: (1) The projections of the final path and the path p_i to the sketch graph must coincide. (2) Each

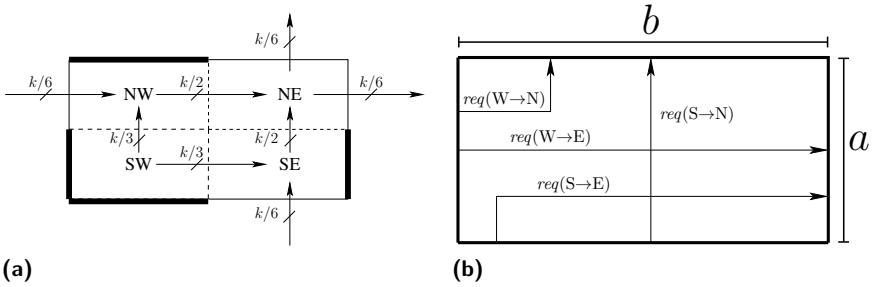

Figure 1 (a) Partitioning of a tile to quadrants [**?**]. Thick lines represent "walls" that cannot be crosses by paths. Sources may reside only in the SW quadrant of a tile. Maximum flow amounts crossing quadrant sides appears next to each side. Final destinations of paths are assumed (pessimistically) to be in the top row of the NE quadrant. (b) Crossbar routing: flow crossing an $a \times b$ grid [**?**].

tile is partitioned to quadrants and routing rules within a tile are defined. (3) Crossbar routing within each quadrant is applied to determine the final paths (except for routing in SW quadrants in which paths are already assigned).

Sketch paths and routing between tiles. Each path p_i computed by the randomized rounding procedure is projected to a sketch path p_i^s in the sketch graph. The final path $\hat{p}_i$ assigned to request r_i traverses the same sequence of tiles, namely, the projection of $\hat{p}_i$ is also p_i^s.

Routing rules within a tile [?]. Each tile is partitioned to quadrants as depicted in Figure **??**. The bold sides (i.e., "walls") of the quadrants indicate that final paths may not cross these walls. The classification of the requests ensures that source vertices of requests reside only in SW quadrants of tiles. Final paths may not enter the SW quadrants; they may only emanate from them. If the endpoint of a sketch path p_i^s ends in tile T, then the path $\hat{p}_i$ must reach a copy of its destination b_i in T. Reaching the destination is guaranteed by having $\hat{p}_i$ reach the top row of the NE quadrant of T (and thus it must reach the row of b_i along the way).

Crossbar routing. [?]. Routing in each quadrant is simply an instance of routing in a 2D grid where requests enter from two adjacent sides and exit from the opposite sides. Figure **??** depicts such an instance in which requests arrive from the left and bottom sides and exit from the top and right side. The following claim characterizes when crossbar routing succeeds.

▶ **Claim 11.** *[?] Consider a 2-dimensional directed $a \times b$ grid. A set of requests can be routed from the bottom and left boundaries of the grid to the opposite boundaries if and only if the number of requests that should exit each side is at most the length of the corresponding side.*

We conclude with the following claim.

▶ **Claim 12.** *Detailed routing succeeds in routing all the requests in R_{quad}.*

Proof sketch. The sketch graph is a directed acyclic graph. Sort the tiles in topological ordering. Within each tile, order the quadrants also in topological order: SW, NW, SE, NE. Prove by induction on the position of the quadrant in the topological ordering that detailed routing in the quadrant succeeds. The induction basis, for all SW quadrants, follows because

this routing is performed by the routing in the first quadrant. Note that filtering ensures that the number of paths between tiles is at most $2\lambda k = k/6$. Routing in the first quadrant ensures that the number of paths emanating from each side of a SW quadrant is at most $k/3$. The induction step follows by applying Claim **??**. ◀

5.5 Approximation Ratio

AR: As we discussed this section should be revised: (1) all above talks about $c = B = 1$, see the top of the section (it was there before my pass), so the proof cannot use B and c. (2) there should be a claim for $B = c = 1$ which is the main proof for this section. (3) then a corollary for arbitrary B and c with some argument why even of we have arbitrary $B = c$ (i.e., they are equal, but anything) we get something good (what do we get ?) (4) a few words that explain why the remark is correct.

▶ **Theorem 13.** *The approximation ratio of the algorithm for packet requests in $R_{d_{\min},d_{\max}}$ is constant in expectation.*

Proof. By Lemma **??**, bounding path lengths in the MCF incurs a factor of $c/(B + 2c)$. The scaling of the capacities in the space-time grid incurs a factor of λ. The classification into 4 classes incurs a factor of $1/4$. By Claim **??**, the filtering stage incurs a factor of $1 - O(1/k)$ in expectation. By Theorem **??**, routing in the first quadrant incurs factor of 0.93 in expectation. By Corollary **??**, limiting the number of paths that cross the sides of SW quadrants incurs an additional factor of $2/3$. ◀

▶ Remark. If deadlines are present, then they are missed by most k time steps.

6 Extension to Non-unit Capacities & Buffer Sizes

Our results extend to arbitrary values of B and c. If the ratio of B/c is a constant, our results will give a randomized approximation algorithm with a (different) constant approximation ratio We outline the required modifications to handle this case.

1. **AR**: I have no idea what this item means, and in particular I don't see in its text anything that talks about c and B. So it is not clear to me (or any other reader) why if B or c are not 1, there should be an additional class. Exhaustive search can be executed in polynomial time provided that the distance of each packet request is at most $\ln(\ell_S)$ (see [**?**, Lemma 7]). This means that there is an additional category of requests, called *very short* requests, in addition to short, medium, and long requests. The distance of short requests is lower bounded by $\ln(\ell_S)$. Hence we can apply the same algorithm to long, medium, and short requests with the proper parametrization.
2. We can run the algorithm of Section **??** for the short, medium, and long requests using as both the the buffer size and the link capacity $\min\{B, c\}$. By Corollary **??** we get an ....-approximation with respect to a *fractional* optimal that uses the same capacities. But, because we consider the the fractional optimal we can use the fact that the fractional Optimum with this capacities is only at most a c/B fraction away from the real fractional optimum. Hence, because, we assume that c/B is a constant we remain with a constant approximation algorithm.

▶ **Theorem 14.** *There exists a randomized algorithm for the Max-Pkt-Line such that if B/c is constant, then its approximation ratio is a (different) constant.*

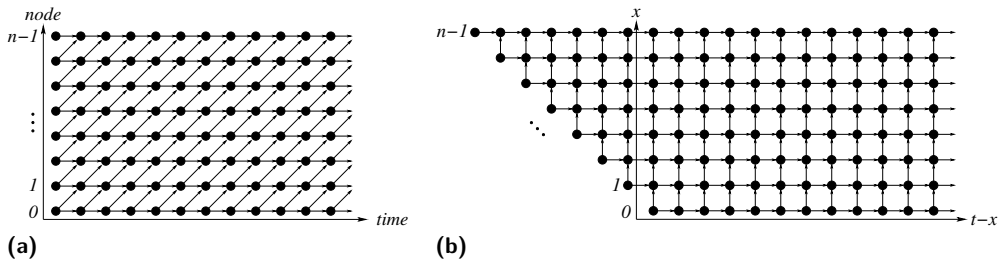

Figure 2 The space-time graph G^{st} before and after untilting [**?**].

References

A Reduction of Packet-Routing to Path Packing

A.1 Space-Time Transformation

A *space-time transformation* is a method to map schedules in a directed graph over time into paths in a directed acyclic graph [**?**, **?**, **?**, **?**]. Let $G = (V, E)$ denote a directed graph. The space-time transformation of G is the acyclic directed infinite graph $G^{st} = (V^{st}, E^{st})$, where: (i) $V^{st} \triangleq V \times \mathbb{N}$. We refer to every vertex (v, t) as a *copy* of v. Namely, each vertex has a copy for every time step. We often refer to the copies of v as the *row* of v. (ii) $E^{st} \triangleq E_0 \cup E_1$ where the set of forward edges is defined by $E_0 \triangleq \{(u, t) \to (v, t + 1) \ : \ (u, v) \in E \ , \ t \in \mathbb{N}\}$ and the set of store edges is defined by $E_1 \triangleq \{(u, t) \to (u, t + 1) \ : \ u \in V, t \in \mathbb{N}\}$. (iii) The capacity of every forward edge is c, and the capacity of every store edge is B. Figure **??** depicts the space-time graph G^{st} for a directed path over n vertices. Note that we refer to a space-time vertex as (v, t) even though the x-axis corresponds to time and the y-axis corresponds to the nodes. We often refer to G^{st} as the space-time grid.

A.2 Untilting

The forward edges of the space-time graph G^{st} are depicted in Fig. **??** by diagonal segments. We prefer the drawing of G^{st} in which the edges are depicted by axis-parallel segments [**?**]. Indeed, the drawing is rectified by mapping the space-time vertex (v, t) to the point $(v, t - v)$ so that store edges are horizontal and forward edges are vertical. Untilting simplifies the definition of tiles and the description of the routing. Figure **??** depicts the untilted space-time graph G^{st} (e.g., the node $(2, 1)$ is mapped to $(2, -1)$.).

A.3 The Reduction

A schedule s_i for a packet request r_i specifies a path p_i in G^{st} as follows. The path starts at (a_i, t_i) and ends in a copy of b_i. The edges of p_i are determined by the actions in s_i; a store action is mapped to a store edge, and a forward action is mapped to a forward edge. We conclude that a schedule S induces a *packing* of paths such that at most B paths cross every store edge, and at most c paths cross every forward edge. Note that the length of the path p_i equals the length of the schedule s_i. Hence we can reduce each packet request r_i to a path request r_i^{st} over the space-time graph. Vice versa, a packing of paths $\{p_i\}_{i \in I}$, where

p_i begins in (a_i, t_i) and ends in a copy of b_i induces a schedule[5]. We conclude that there is a one-to-one correspondence between schedules and path packings.

B Multi-Commodity Flow Terminology

Network. A network N is a directed graph[6] $G = (V, E)$, where edges have non-negative capacities $c(e)$. For a vertex $u \in V$, let $\mathsf{out}(u)$ denote the outward neighbors, namely the set $\{y \in V \mid (u, y) \in E\}$. Similarly, $\mathsf{in}(u) \triangleq \{x \in V \mid (x, u) \in E\}$.

Grid Network. A grid network N is a directed graph $G = (V, E)$ where $V = [n] \times \mathbb{N}$ and $(i, t_1) \to (j, t_2)$ is an edge in E if $t_2 = t_1 + 1$ and $0 \leq j - i \leq 1$.

Commodities/Requests. A *request* r_i is a pair (a_i, b_i), where $a_i \in V$ is the *source* and $b_i \in V$ is the *destination*. We often refer to a request r_i as *commodity* i. The request r_i is to ship commodity i from a_i to b_i. All commodities have unit demand.

In the case of space-time grids, a request is a triple (a_i, b_i, t_i) where $a_i, b_i \in [n]$ are the source and destination and t_i is the time of arrival. The source in the grid is the node (a_i, t_i). The destination in the grid is any copy of b_i, namely, vertex (b_i, t), where $t \in \mathbb{N}$.

Single commodity flow. Consider commodity i. A *single-commodity flow* from a_i to b_i is a function $f_i : E \to \mathbb{R}^{\geq 0}$ that satisfies the following conditions:

(i) Capacity constraints: for every edge $(u, v) \in E$, $0 \leq f_i(u, v) \leq c(u, v)$.
(ii) Flow conservation: for every vertex $u \in V \setminus \{a_i, b_i\}$

$$\sum_{x \in \mathsf{in}(u)} f_i(x, u) = \sum_{y \in \mathsf{out}(u)} f_i(u, y).$$

(iii) Demand constraint: $|f_i| \leq 1$ (amount of flow $|f_i|$ defined below).

The *amount* of flow delivered by the flow f is defined by

$$|f_i| \triangleq \sum_{y \in \mathsf{out}(a_i)} f_i(a_i, y) - \sum_{x \in \mathsf{in}(a_i)} f_i(x, a_i).$$

The *support* of a flow f_i is the set of edges (u, v) such that $f_i(u, v) > 0$. As cycles in the support of f_i can be removed without decreasing $|f_i|$, one may assume that the support of f_i is acyclic.

Multi-commodity flow (MCF). In a multi-commodity flow (MCF) there is a set of commodities I, and, for each commodity $i \in I$, we have a source-destination pair denoted by (a_i, b_i). Consider a sequence $F \triangleq \{f_i\}_{i \in I}$ of single-commodity flows, where each f_i is a single commodity flow from the source vertex a_i to the destination vertex b_i. We abuse notation, and let F denote also the sum of the flows, namely $F : E \to \mathbb{R}$, where $F(e) \triangleq \sum_{i \in I} f_i(e)$, for

[5] In [?], super-sinks are added to the space-time grid so that the destination of each path request is single vertex rather than a row.
[6] The graph G is this section is an arbitrary graph, not a directed path. In fact, we use MCF over the space-time graph of the directed graph with super sinks for copies of each vertex.

every edge e. A sequence F is a *multi-commodity flow* if, in addition it satisfies *cumulative capacity constraints* defined by:

for every edge $(u, v) \in E$: $F(u, v) \leq c(u, v)$.

The *throughput* of an MCF $F \triangleq \{f_i\}_{i \in I}$ is defined to be $\sum_{i \in I} |f_i|$. In the maximum throughput MCF problem, the goal is to find an MCF F that maximized the throughput.

An MCF is *all-or-nothing* if $|f_i| \in \{0, d_i\}$, for every commodity $i \in I$. An MCF is *unsplittable* if the support of each flow is a simple path. The support of each single commodity flow f_i is a simple path if $F = \{f_i\}_{i \in I}$ is an unsplittable MCF. An MCF is *integral* if it is both all-or-nothing and unsplittable. An MCF that is not integral is called a *fractional* MCF.

C Randomized Rounding Procedure

In this section we present material from Raghavan [**?**] about randomized rounding. The proof of the Chernoff bound is also based on Young [**?**].

Given an instance $F = \{f_i\}_{i \in I}$ of a fractional multi-commodity flow, we are interested in finding an integral (i.e., all-or-nothing and unsplittable) multi-commodity flow $F' = \{f'_i\}_{i \in I}$ such that the throughput of F' is as close to the benefit of F as possible.

▶ **Observation 15.** *As flows along cycles are easy to eliminate, we assume that the support of every flow $f_i \in F$ is acyclic.*

We employ a randomized procedure, called *randomized rounding*, to obtain F' from F. We emphasize that all the random variables used in the procedure are independent. The procedure is divided into two parts. First, we flip random coins to decide which commodities are supplied. Next, we perform a random walk along the support of the supplied commodities. Each such walk is a simple path along which the supplied commodity is delivered. We describe the two parts in details below.

Deciding which commodities are supplied. For each commodity, we first decide if $|f'_i| = 1$ or $|f'_i| = 0$. This decision is made by tossing a biased coin $b_i \in \{0, 1\}$ such that

$$\mathbf{Pr}[b_i = 1] \triangleq |f_i| \leq 1.$$

If $b_i = 1$, then we decide that $|f'_i| = 1$ (i.e., the packet is accepted). Otherwise, if $b_i = 0$, then we decide that $|f'_i| = 0$ (i.e., the packet is rejected).

Assigning paths to the supplied commodities. For each commodity i that we decided to fully supply (i.e., $b_i = 1$), we assign a simple path P_i from its source s_i to its destination t_i by following a random walk along the support of f_i. At each node, the random walk proceeds by rolling a dice. The probabilities of the sides of the dice are proportional to the flow amounts. A detailed description of the computation of the path P_i is given in Algorithm **??**.

Definition of F'. Each flow $f'_i \in F'$ is defined as follows. If $b_i = 0$, then f'_i is identically zero. If $b_i = 1$, then f'_i is defined by

$$f'_i(u, v) \triangleq \begin{cases} 1 & \text{if } (u, v) \in P_i, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $F' = \{f'_i \mid b_i = 1\}$ is an all-or-nothing unsplittable flow, as required.

Algorithm 1 Algorithm for assigning a path P_i to flow f_i.

$P_i \leftarrow \{s_i\}$.
$u \leftarrow s_i$
while $u \neq t_i$ **do** ▷ did not reach t_i yet
    $v \leftarrow$ *choose-next-vertex*(u).
    Append v to P_i
    $u \leftarrow v$
end while
return (P_i).
procedure *choose-next-vertex*(u, f_i)
    Let $\mathsf{out}(u, f_i)$ denote the set of edges in the support of f_i that emanate from u.
    Consider a dice $C(u, f_i)$ with $|\mathsf{out}(u, f_i)|$ sides. The side corresponding to an edge $(u, v) \in \mathsf{out}(u, f_i)$ has probability $f_i(u, v)/(\sum_{(u,v')\in\mathsf{out}(u,f_i)} f_i(u, v'))$.
    Let v denote the outcome of a random roll of the dice $C(u, f_i)$.
    return (v)
end procedure

C.1 Expected flow per edge

The following claim can be proved by induction on the position of an edge in a topological ordering of the support of f_i.

▶ **Claim 16.** *For every commodity i and every edge $(u, v) \in E$:*

$$\mathbf{Pr}\left[(u, v) \in P_i\right] = f_i(u, v),$$
$$\mathbf{E}\left[f'_i(u, v)\right] = f_i(u, v).$$

D Chernoff Bound

▶ **Definition 17.** The function $\beta : (-1, \infty) \to \mathbb{R}$ is defined by $\beta(\varepsilon) \triangleq (1 + \varepsilon)\ln(1 + \varepsilon) - \varepsilon$.

▶ **Observation 18.** *For ε such that $-1 < \varepsilon < 1$ it holds that $\frac{\varepsilon^2}{2} \geq \beta(\varepsilon) \geq \frac{2\varepsilon^2}{4.2+\varepsilon}$. Hence, $\beta(\varepsilon) = \Theta(\varepsilon^2)$.*

▶ **Theorem 19** (Chernoff Bound [?, ?])**.** *Let $\{X_i\}_i$ denote a sequence of independent random variables attaining values in $[0, 1]$. Assume that $\mathbf{E}[X_i] \leq \mu_i$. Let $X \triangleq \sum_i X_i$ and $\mu \triangleq \sum_i \mu_i$. Then, for $\varepsilon > 0$,*

$$\mathbf{Pr}\left[X \geq (1 + \varepsilon) \cdot \mu\right] \leq e^{-\beta(\varepsilon)\cdot\mu}.$$

▶ **Corollary 20.** *Under the same conditions as in Theorem* **??***,*

$$\mathbf{Pr}\left[X \geq \alpha \cdot \mu\right] \leq \left(\frac{e}{\alpha}\right)^{\alpha\cdot\mu}.$$

E Proofs

Proof of Claim ??

Proof. We begin by bounding the probability that at least $2\lambda k$ sketch paths cross a single sketch edge.

▶ **Lemma 21** (Chernoff Bound). *For every edge e^s in the sketch graph,*[7]

$$\mathbf{Pr}\left[\sum_i \mathbb{1}_{p_i^s}(e^s) > 2\lambda k\right] \leq e^{-k/6}\,. \tag{1}$$

Proof of lemma. Recall that the edge capacities in the MCF F are λ. The capacity constraint $\sum_i f_i(e) \leq \lambda$ implies that $f_i(e) \leq \lambda$. Each sketch edge e^s corresponds to the grid edges between adjacent tiles. Since the demand of each request is 1, it follows that $f_i(e^s) \leq 1$.

For every edge e and request r_i, we have

$$\mathbf{E}\left[\mathbb{1}_{p_i^s}(e^s)\right] = \mathbf{Pr}\left[\mathbb{1}_{p_i^s}(e^s) = 1\right] = f_i(e^s) \leq 1.$$

Fix a sketch edge e^s. The random variables $\{\mathbb{1}_{p_i^s}(e^s)\}_i$ are independent 0-1 variables. Moreover, $\sum_i \mathbf{E}\left[\mathbb{1}_{p_i^s}(e^s)\right] = \sum_i f_i(e^s) = \sum_{e\in e^s}\sum_i f_i(e) \leq \lambda \cdot k$. By Chernoff bound (Theorem **??**),

$$\mathbf{Pr}\left[\sum_i \mathbb{1}_{p_i^s}(e^s) > 2\cdot\sum_i \mathbf{E}\left[\mathbb{1}_{p_i^s}(e^s)\right]\right] < e^{-\beta(1)\cdot\lambda k} = e^{-k/6}.$$

◀

A request $r_i \in R_{rnd}$ is not in R_{fltr} iff at least one of the edges $e^s \in p_i^s$ is over-congested. Hence, by a union bound,

$$\begin{aligned}\mathbf{Pr}\left[r_i \notin R_{fltr} \mid r_i \in R_{rnd}\right] &\leq |p_i^s| \cdot e^{-k/6}\\ &\leq \left(\left\lceil\frac{d_{\max}}{k}\right\rceil + 1\right)\cdot e^{-\ln d_{\max}}\\ &= O\left(\frac{1}{k}\right).\end{aligned}$$

◀

▶ Remark. Note that filtering uses the fact that $k \geq 6\ln d_{\max}$.

Proof of Theorem ??

Proof. By linearity of expectation, it suffices to prove that $\mathbf{E}_\tau\left[|R'_T|\right] \geq 0.93\cdot\mathbf{E}_\tau\left[|R_T|\right]$

Let $\hat{R}_T \subseteq R_T$ denote the subset of requests that do not violate the capacity constraints of "rectangles" described below. We prove that: (1) The cuts induced by rectangles are critical in the sense that if none of these cuts is violated, then the set of requests is feasible. (2) $\mathbf{E}_\tau\left[|\hat{R}_T|\right] \geq 0.93\cdot\mathbf{E}_\tau\left[|R_T|\right]$. By the construction, R'_T is of maximum cardinality, therefore, $|R'_T| \geq |\hat{R}_T|$, and the theorem follows.

We now describe how the feasible subset $\hat{R}_T$ is constructed. Consider a subset S of the vertices in the SW quadrant of T. Let $\mathsf{dem}(S)$ denote the number of requests whose origin is in S. Let $\mathsf{cap}(S)$ denote the capacity of the edges in the network $N(R_T)$ that emanate from S. By the min-cut max-flow theorem, a set of requests $X \subseteq R_T$ is feasible if and only if $\mathsf{dem}(S) \leq \mathsf{cap}(S)$ for every cut $S \cup \{\tilde{s}\}$ in the network $N(X)$.

In fact, it is not necessary to consider all the cuts. It suffices to consider only axis parallel rectangles contained in the quadrant T. The reason is that, without loss of generality, the set

[7] The e in the RHS is the base of the natural logarithm.

S is connected in the underlying undirected graph of the grid (i.e., consider each connected components of S separately). Every "connected" set S can be replaced by the smallest rectangle $Z(S)$ that contains S. We claim that $\mathsf{cap}(S) \geq \mathsf{cap}(Z(S))$ and $\mathsf{dem}(S) \leq \mathsf{dem}(Z(S))$. Indeed, there is an injection from the edges in the cut of $Z(S)$ to the edges in the cut of S. For example, a vertical edge e in the cut of $Z(S)$ is mapped to the topmost edge e' in the cut of S that is in the column of e. Hence, $\mathsf{cap}(Z(S)) \leq \mathsf{cap}(S)$. On the other hand, as $S \subseteq Z(S)$, it follows that $\mathsf{dem}(S) \leq \mathsf{dem}(Z(S))$. Hence if $\mathsf{dem}(S) > \mathsf{cap}(S)$, then $\mathsf{dem}(Z(S)) > \mathsf{cap}(Z(S))$.

We say that a rectangle Z is *overloaded* if $\mathsf{dem}(Z) > \mathsf{cap}(Z)$. The set $\hat{R}_T$ is defined to be the set of requests $r_i \in R_T$ such that the source of r_i is not included in an overloaded rectangle. Namely,

$$\hat{R}_T \triangleq \{r_i \in R_T \mid \forall \text{ rectangles } Z : Z \text{ is overloaded } \Rightarrow (a_i, t_i) \notin Z\}$$

Consider an $x \times y$ rectangle Z. We wish to bound the probability that $\mathsf{dem}(Z) > \mathsf{cap}(Z)$. Note that $\mathsf{cap}(Z) = x + y$. Since requests that start in Z must exit the quadrant, it follows that $\mathsf{dem}(Z)$ is bounded by the number of paths in R_T that cross the top or right side of Z (there might be additional paths that do not start in Z but cross Z.). The amount of flow that emanates from Z is bounded by $\lambda \cdot (x + y)$ (the initial capacities are λ and there are $x + y$ edges in the cut). By Claim **??**, $\mathbf{Pr}[e \in p_i] = f_i(e)$. Summing up over all the edges in the cut of Z and the requests in R_T, the expected number of paths in R_T that cross the cut of Z equals the flow which, in turn, is bounded by the capacity $\lambda \cdot (x + y)$. As the paths of the requests are independent random variables, we apply Coro. **??** to obtain

$$\begin{aligned}\mathbf{Pr}[\mathsf{dem}(Z) > \mathsf{cap}(Z)] &\leq \mathbf{Pr}\left[\sum_{i \in R_T} |p_i \cap \mathit{cut}(Z)| > (x + y)\right] \\ &\leq (\lambda \cdot e)^{x+y}\end{aligned}$$

Each source (a_i, t_i) is contained in at most $x \cdot y$ rectangles with side lengths $x \times y$. By applying a union bound, the probability that (a_i, t_i) is contained in an overloaded rectangle is bounded by

$$\begin{aligned}\mathbf{Pr}[\exists \text{ overloaded rectangle } Z : (a_i, t_i) \in Z] &\leq \sum_{x=1}^{\infty} \sum_{y=1}^{\infty} xy \cdot (\lambda \cdot e)^{x+y} \\ &\leq \frac{(\lambda \cdot e)^2}{(1 - \lambda \cdot e)^4} \leq 0.07,\end{aligned}$$

and the theorem follows. ◀