Poster: Network Utility Maximization under Maximum Delay Constraints and Throughput Requirements

Qingyu Liu, Haibo Zeng
Electrical and Computer Engineering
Virginia Tech

Minghua Chen
Information Engineering
The Chinese University of Hong Kong

ABSTRACT

We consider a multiple-unicast network flow problem of maximizing aggregate user utilities under link capacity constraints, maximum delay constraints, and user throughput requirements. A user's utility is a concave function of the achieved throughput or the experienced maximum delay. We first prove that it is NP-complete either (i) to construct a feasible solution meeting all constraints, or (ii) to obtain an optimal solution after we relax maximum delay constraints or throughput requirements. We then leverage a novel understanding between non-convex maximum-delay-aware problems and their convex average-delay-aware counterparts, and design a polynomial-time approximation algorithm named PASS. PASS achieves constant or problem-dependent approximation ratios, at the cost of violating maximum delay constraints or throughput requirements by up to constant or problem-dependent ratios, under realistic conditions. We empirically evaluate our solutions using simulations of supporting video-conferencing traffic across Amazon EC2 datacenters. Compared to conceivable baselines, PASS obtains up to 100% improvement of utilities, meeting throughput requirements but relaxing maximum delay constraints that are acceptable for video conferencing applications.

CCS CONCEPTS

• **Mathematics of computing → Network flows**;

KEYWORDS

Network utility maximization, delay-aware network flow

ACM Reference Format:
Qingyu Liu, Haibo Zeng and Minghua Chen. 2019. Poster: Network Utility Maximization under Maximum Delay Constraints and Throughput Requirements. In *The Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobihoc '19), July 2–5, 2019, Catania, Italy.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3323679.3326628

1 INTRODUCTION

We consider a multi-hop network with multiple users, where each user requires to stream a flow from a source a destination. We study the problem of maximizing aggregate user utilities under link capacity constraints, maximum end-to-end delay constraints, and user throughput requirements. A user's utility is a concave function of the achieved throughput or the experienced maximum delay. We consider a delay model where transmission over a link experiences a constant delay if the link aggregate flow rate is within a constant capacity, and an unbounded delay otherwise. This model fits many practical applications, particularly the routing of delay-critical video conferencing traffic over inter-datacenter networks.

Many existing studies maximize utilities with throughput concerns, but less of them consider maximum delay which is non-convex. Misra *et al.* [3] minimize maximum delay under a throughput requirement, and design a Fully-Polynomial-Time Approximation Scheme (FPTAS). It solves flow problems iteratively in time-expanded networks using binary search that is applicable only in the single-unicast setting. Cao *et al.* [1] develop an FPTAS to maximize throughput under maximum delay constraints in a multiple-unicast setting. Their FPTAS also solves problems iteratively in time-expanded networks, which is time-consuming. Besides, it leverages the primal-dual algorithm to solve the problem casted as a linear program. It is unclear how to extend their technique to the scenario where a user's utility can be a general concave function.

Our studied problem cover existing problems as special cases. It is a challenging problem, as we prove that it is NP-complete either (i) to construct a feasible solution meeting all constraints, or (ii) to obtain an optimal solution after we relax maximum delay constraints or throughput requirements. We design a polynomial-time approximation algorithm PASS. Different from existing techniques based on time-expanded networks, we leverage a novel understanding between the non-convex maximum delay and the convex average delay, and suggest a new avenue for optimizing maximum-delay-aware network communications. PASS achieves constant or problem-dependent approximation ratios, at the cost of violating maximum delay constraints or throughput requirements by up to constant or problem-dependent ratios, under realistic conditions. We empirically evaluate our solutions by simulating video-conferencing traffic across Amazon EC2 datacenters.

2 PROBLEM DEFINITION

We model a multi-hop network as a directed graph $G \triangleq (V, E)$. Each link $e \in E$ has a constant capacity $c_e \geq 0$ and a constant delay $d_e \geq 0$. Data streamed to each $e \in E$ experiences a delay of d_e to pass it, and the streaming rate must be within c_e. We are given K users, where each user i requires a source $s_i \in V$ to stream a flow to a destination $t_i \in V \backslash \{s_i\}$, possibly using multiple paths.

We denote a multiple-unicast flow as $f \triangleq \{f_i, i = 1, 2, ..., K\}$, with f_i to be a single-unicast flow from s_i to t_i. We denote the *throughput* of f_i by $|f_i|$, which is the flow rate sent from s_i to t_i following f_i. The *maximum delay* (resp. *average delay*) of f_i is denoted by $\mathcal{M}(f_i)$ (resp. $\mathcal{A}(f_i)$), which is the maximum (resp. average) delay experienced by all flow units from s_i to t_i.

Mobihoc '19, July 2–5, 2019, Catania, Italy

ACM ISBN 978-1-4503-6764-6/19/07.
https://doi.org/10.1145/3323679.3326628

Algorithm 1 Our Proposed Algorithm PASS

procedure
  Replace each $\mathcal{M}(f_i)$ in problem (1) by $\mathcal{A}(f_i)$
  Get the solution $f = \{f_i, i = 1, 2, ..., K\}$ to the problem
  $x_i^{\text{delete}} = \epsilon \cdot |f_i|, \forall i = 1, 2, ..., K$
  for $i = 1, 2, ..., K$ **do**
    while $x_i^{\text{delete}} > 0$ **do**
      Find the slowest flow-carrying path $p_i \in P_i$
      if $x^{p_i} > x_i^{\text{delete}}$ **then**
        $x^{p_i} = x^{p_i} - x_i^{\text{delete}},\ x_i^{\text{delete}} = 0$
      else
        $x_i^{\text{delete}} = x_i^{\text{delete}} - x^{p_i},\ x^{p_i} = 0$
  return the remaining flow $f = \{f_i, i = 1, 2, ..., K\}$

For each f_i, (i) we denote its *throughput-based utility* as $\mathcal{U}_i^t(|f_i|)$, which rewards f_i based on the achieved throughput; (ii) we denote its *maximum-delay-based utility* as $-\mathcal{U}_i^d(\mathcal{M}(f_i))$, where $\mathcal{U}_i^d(\mathcal{M}(f_i))$ penalizes f_i based on the experienced maximum delay. We consider the following fundamental network utility maximization problem

$$\text{obj:} \quad \text{either max} \sum_{i=1}^{K} \mathcal{U}_i^t(|f_i|) \text{ or } \max -\sum_{i=1}^{K} \mathcal{U}_i^d(\mathcal{M}(f_i)) \quad (1a)$$

$$\text{s.t.} \quad |f_i| \geq R_i \text{ and } \mathcal{M}(f_i) \leq D_i,\ \forall i = 1, 2, ..., K, \quad (1b)$$

where $f = \{f_i, i = 1, ..., K\}$ is a feasible multiple-unicast flow meeting flow conservation constraints and link capacity constraints.

3 OUR RESULTS

Theorem 3.1. *For our problem, it is NP-complete (i) to construct a feasible solution that meets all constraints, or (ii) to obtain an optimal solution that meets throughput requirements but relaxes maximum delay constraints, or (iii) to obtain an optimal solution that meets maximum delay constraints but relaxes throughput requirements.*

PASS (Polynomial-time Algorithm Supporting utility-maximal flows Subject to throughput/delay constraints) is presented in Algorithm 1, where P_i is the set of paths from s_i to t_i. PASS first solves the average-delay-aware counterpart of our problem and obtains the solution $f = \{f_i, i = 1, 2, ..., K\}$. It then deletes a rate of $\epsilon \cdot |f_i|$ iteratively from the slowest flow-carrying paths of f_i, for each $i = 1, ..., K$. The remaining flow after deleting is the solution.

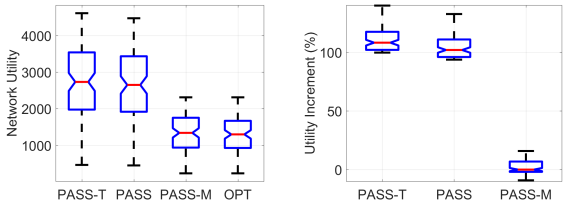
Theorem 3.2. *Suppose we use PASS to solve problem* (1) *with an arbitrary* $\epsilon \in (0, 1)$. *If the problem is feasible, meeting following conditions*

(1) *for each* $i = 1, 2, ..., K$ *and any* $a \geq 0$, $\mathcal{U}_i^t(a)$ *is concave, non-decreasing, and non-negative with a, and* $\mathcal{U}_i^d(a)$ *is convex, non-decreasing, and non-negative with a,*
(2) *for an arbitrary* $a \geq 0$, *the following holds given any* $\sigma \geq 1$

$$\mathcal{U}_i^d(\sigma \cdot a) \leq \sigma \cdot \mathcal{U}_i^d(a),\ \forall i = 1, 2, ..., K,$$

then PASS must return a solution $\bar{f} = \{\bar{f}_i, i = 1, ..., K\}$ *in a polynomial time, meeting the following relaxed constraints*

$$\left|\bar{f}_i\right| \geq (1 - \epsilon) \cdot R_i \text{ and } \mathcal{M}\left(\bar{f}_i\right) \leq D_i/\epsilon,\ \forall i = 1, 2, ..., K.$$



(a) Network utility results. (b) Utility improvement to optimal.

Figure 1: Simulation results of our algorithms, with $\epsilon = 3\%$.

Suppose $f^* = \{f_i^*, i = 1, 2, ..., K\}$ *is the optimal solution. If our objective is to maximize throughput-based utility, we have*

$$\sum_{i=1}^{K} \mathcal{U}_i^t\left(\left|\bar{f}_i\right|\right) \geq (1 - \epsilon) \cdot \sum_{i=1}^{K} \mathcal{U}_i^t\left(\left|f_i^*\right|\right).$$

If our objective is to maximize maximum-delay-based utility, we have

$$\sum_{i=1}^{K} \mathcal{U}_i^d\left(\mathcal{M}\left(\bar{f}_i\right)\right) \leq \frac{1}{\epsilon} \cdot \sum_{i=1}^{K} \mathcal{U}_i^d\left(\mathcal{M}\left(f_i^*\right)\right).$$

PASS may violate maximum delay constraints and throughput requirements. We can modify PASS and get anther two algorithms PASS-M and PASS-T. PASS-M (resp. PASS-T) meet maximum delay constraints (resp. throughput requirements), but may violate throughput requirements (resp. maximum delay constraints). PASS-M, PASS-T, and proofs to our theorems refer to our report [2].

4 PERFORMANCE EVALUATION

We simulate video conferencing traffic over a real-world inter-datacenter network of 6 Amazon EC2 datacenters. We set link delays and capacities according to practical evaluations, and assume two unicasts. We assume the utility is $\mathcal{U}_i^t(|f_i|) = w_i \cdot |f_i|, i = 1, 2$. We set $R_1 = R_2 = 80$ and $D_1 = D_2 = 150$, as for video conferencing, (i) a delay less than 150ms can provide a transparent interactivity, while (ii) a delay larger than 150ms but within 400ms is still acceptable.

We vary w_1 (resp. w_2) from 1 to 10 with a step of 1, and give simulation results in Fig. 1. PASS and PASS-T obtain a huge (over 100%) utility improvement compared to optimal. Besides in average, (i) the throughput of PASS is 138 (resp. 302) for the first unicast (resp. second unicast), satisfying requirements $R_1 = R_2 = 80$; (ii) the maximum delay of PASS is 195 (resp. 301) for the first unicast (resp. second unicast), violating constraints $D_1 = D_2 = 150$ but still within 400ms; (iii) the throughput of PASS-M is 71 (resp. 154) for the first unicast (resp. second unicast); (iv) the maximum delay of PASS-T is 222 (resp. 322) for the first unicast (resp. second unicast).

REFERENCES

[1] Zizhong Cao, Paul Claisse, René-Jean Essiambre, Murali Kodialam, and TV Lakshman. 2017. Optimizing throughput in optical networks: The joint routing and power control problem. *IEEE/ACM Trans. Networking* 25, 1 (2017), 199–209.
[2] Qingyu Liu, Haibo Zeng, and Minghua Chen. 2018. Network Utility Maximization under Maximum Delay Constraints and Throughput Requirements. *arXiv preprint arXiv:1812.06169* (2018). Available at https://arxiv.org/abs/1812.06169.
[3] Satyajayant Misra, Guoliang Xue, and Dejun Yang. 2009. Polynomial time approximations for multi-path routing with bandwidth and delay constraints. In *Proc. IEEE Int'l Conf. Computer Communications*. IEEE, 558–566.