Maximizing QoS for Video Streams

Ahmed Redha Mahlous

Computer Science Department, Prince Sultan University, Riyadh, KSA

Summary
In computer networks, quality of service (QoS) is considered an important user demand especially in real-time multimedia systems where transmission delay is not tolerated. An important aspect of such networks is maximizing bandwidth usage through minimizing the amount of unused bandwidth in the channel. In this paper, we present an application of a novel QoS bandwidth management algorithm, MaxBalanced Spanning Trees Packing (MBSTP) for video streaming that minimizes the amount of unused bandwidth and allows parallel video streaming services in networks with a series-parallel graph topology.

Key words:
Computers networks, Quality-of-Service routing, Virtual sub-networks, Series parallel graphs, Packing of spanning trees.

1. Introduction

Multimedia applications that modern telecommunication networks are supporting, such as video streaming, video conferencing and online interactive games require high-quality combined with quality of service (QoS) support such as guaranteed bandwidth, delay, jitter and error rate. These requirements push Internet Service providers (ISPs) to prolong their range of services to allow users to utilize these multimedia applications with a certain level of QoS. As a result, users are required to have a certain service level agreement (SLA) with ISPs via a contract in order to guarantee the required level of QoS. With many SLAs received by the ISPs which are then required to determine which request to accept and which route to choose for forwarding the requests in order to reach their respective destinations, bearing in mind that the total revenue generated from the services should be maximized.
In addition, transmission of high quality requests such as video streaming requires high bandwidth, which is difficult to guarantee because of the resource constraints in current networks. Thus, to ensure an efficient provision of such multimedia application requests, ISPs need to have a well-managed bandwidth allocation strategy that guarantees to service the maximum number of requests by minimizing the total unused bandwidth.
With this paradigm, a lot of unused bandwidth is lost while other applications are waiting their turn to be served. One solution to this problem is to use the unused bandwidth effectively for simultaneously servicing other applications in a parallel fashion where possible (Fig.1), hence providing better QoS. This is especially necessary for applications that do not tolerate any form of delay, such as video streaming and multimedia applications and is known as the bandwidth packing problem.
In this paper we extend the classical bandwidth packing problem, which serves one to one connections, to the more practical one in networks with a series-parallel graph topology, by proposing a new algorithm where requests can be served in parallel in a one-to-many or many-to-many connections fashion, thus minimizing the unused bandwidth.

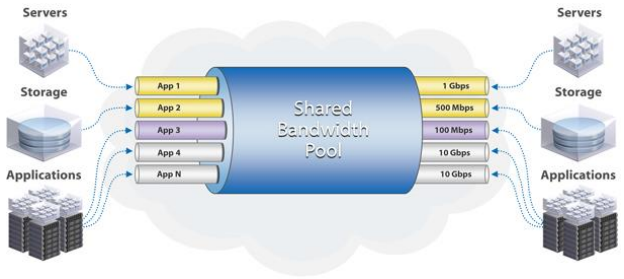

Fig. 1 Parallel services

The remainder of the paper is organized as follows: Section 2 contains a literature review of related works, while section 3 contains a description of the algorithm and its mathematical notations. Section 4 presents the simulation results and analysis, followed by the conclusion in Section 5.

2. Literature Review

With the rapid development of Internet networks in recent years, several research efforts have tackled the bandwidth allocation problem with different bandwidth request models [1-3]. However, the studies focused only on the pricing strategy designed specifically for bandwidth resources of a cloud based system, neglecting the quality of service resources, the performance guarantees and capacity right sizing in this system.
In telecommunication networks with limited bandwidth, a non-linear integer-programing model for a combinatorial bandwidth-packing problem (CBPP) has been proposed [4]. To get a linear mixed integer problem, authors did some transformation and piecewise outer-approximation, followed by model linearization. Then, to solve the

Manuscript received October 5, 2018
Manuscript revised October 20, 2018

resulting linear mixed integer problem, they proposed a cutting plane approach. After extensive computation experiments applied to a variety of network sizes with a number of requests/calls, and call bandwidth requirements, they showed that the proposed solution produces a near optimal solution within a reasonable time.

In another research within the telecommunications networks domain, authors [5] studied the combinatorial bandwidth packing problem where the objective was to decide which requests to accept in order to maximize total revenue subject to bandwidth capacities. Using the Lagrangean relaxation technique, the authors mathematically formulated the problem as an integer programming problem and developed an effective solution algorithm. Through experiments over many different problem structures, they showed that the proposed algorithm provides feasible solutions to the problem within acceptable computational times. However, they did not extend their study to explore the impact of the queuing delay.

In wireless networks, bandwidth study has also been a topic of interest. Under heterogeneous environments consisting of 2G and 3G, authors [6] considered available bandwidth as a dynamic parameter. They used a bootstrap approximation based technique for the estimation of available bandwidth, and then compared it with a hidden Markov model based estimation for accuracy checking. To verify the robustness of the algorithm, they implemented it in both temporal and spatial domains, where the numerical results showed an improvement of the proposed algorithm in terms of error estimation, reliability and overhead, compared to some existing algorithms. The authors limited their study to multi-access network environments where the bandwidth has a dynamic aspect, neglecting environments where the bandwidth is static.

In [7], authors observed that prioritization of data traffic can affect the overall performance of the network. This was also confirmed by authors in [8] who showed that segmenting packets during transmission between Ethernet nodes impacts the network's performance. Therefore, they presented a dynamic QoS aware bandwidth allocation scheme for multihop WiLD networks, which addresses the congestion problem and facilitates QoS support for real-time traffic. They proposed a dynamic slot scheduling mechanism, which distributes the unused bandwidth among the needy nodes in an efficient manner. Simulation results showed that the proposed protocol achieves a significant performance improvement in terms of throughput and delay of real-time traffic. The proposed studies [7] and [8] did not consider the impact of packet sizes larger than MTU and how that impacts the overall network QoS such as throughput, bit error and delay.

Tackling the problem of bandwidth fluctuation in a mobile environment was the focus of [9]. The authors proposed a portioning method where bandwidth is considered as a variable in order to alleviate the problem induced with static portioning and avoid high costs due to the dynamic partitioning. Two bandwidth adaptive algorithms for small and scaled networks were proposed. Even though the research did not directly address offloading techniques, the experiment results showed that both algorithms reduce energy consumption, execution time of mobile applications and adapt well to bandwidth fluctuation.

In a cloud-computing environment, authors [10] studied resource allocation problems for differentiated multimedia services. They proposed a Queuing model to characterize the service process in the cloud center. Based on the proposed Queuing model, they investigated the resource allocation in an FCFS scenario and a priority scenario, respectively. For each scenario they formulated and solved the optimal resource allocation problem to minimize resource costs under the response time constraints. Simulation results demonstrated that the proposed resource allocation schemes can optimally utilize cloud resources to provide satisfactory services for different classes of requests at a minimal resource cost.

The researchers in [11] explored the optimization of bandwidth allocation in hybrid cloud peer to peer (P2P) networks. The authors studied the relationship between the cloud uploading bandwidth allocation and maximizing user's quality of experience (QoE). Additionally, they provided a heuristic algorithm in order to approximate the optimized solution. Through simulation, they showed that it outperformed two classic bandwidth allocation algorithms, namely allocations by arrival rate and by demand.

In a more generalized network, authors in [12] tried to find maximum flow applied to a class of series-parallel graphs. They showed that this type of problem can be solved using a greedy approach, and presented a combinatorial algorithm that runs in O(nm+mlogm) time where m is the number of arcs, and a dynamic programming algorithm with running time O(mlogm). Then, they presented a pseudo-polynomial algorithm for an integral version of the problem which is known to be NP-complete.

For video streaming, there are a number of researches that discussed the QoS for video streaming. In a highly variable bandwidth conditions, authors in [13] studied the streaming videos encoded using scalable video coding (SVC). Under the variable bandwidth constraints, they formulated the quality decisions of video chunks, and then they presented a layered bin packing (LBP) adaptation algorithm as an optimization problem along with its corresponding solution.

In another study, authors in [14] and [15], propose QoS-aware VM provisioning policy for on-demand video transcoding while [16] studied the reliability of Quality of Service (QoS) mechanism over IPv6 for video streaming.

Authors in [17] proposed a Cloud-based Video Streaming Services (CVS2) architecture that faces the challenges

endured by transcoding videos streams in order to match the characteristics of viewers' devices. The studied showed that the proposed scheme provided a decrease of up to 85% of the cost of streaming videos while keeping the robustness of QoS.

From the related works reviewed above we observe that in all of the studies on bandwidth allocation in telecommunication links where QoS has an important impact of network traffic especially for videos streaming, none of them have used the spanning tree approach to minimize the unused bandwidth. Furthermore, none of them have considered this type of approach in series parallel networks, or its application for videos streaming. It is known that using the sum of spanning tree networks is an open problem; however, in this study we propose an application of polynomial algorithm that gives an optimal solution for minimum unused bandwidth in series parallel networks [18] and [19] for video streaming. Such a solution has an important impact on ISPs, where some client applications do not tolerate delay and required a continuous a highly QoS vides streaming.

3. Definitions and Notations

A computer network is presented as an undirected graph G=(V, E) where V is the set of vertices (computers) and E is the set of edges (links) (Fig. 2). We refer to n as the number of vertices of G and m as its number of edges. The metric bandwidth is represented by a vector b on the links set E.

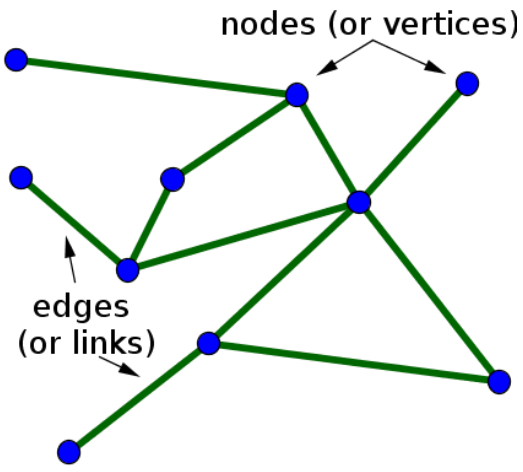

Fig. 2 Network representation graph

3.1 Problem Formulation

The parallel services needed to be run on the network should have two characteristics:

Cover all nodes in the graph and optimize resources usage, thus a minimum number of spanning trees noted T of graph G need to be calculated.

All links of T should have a fixed bandwidth noted bT.

So the problem of Minimizing Unused Bandwidths in Computers Networks can be defined as follows:

Given a network G=(V, E) and an integer bandwidths vector b on E, find spanning tree sub-networks (pre-computed parallel services) T1, …, Tk and corresponding integer bandwidths b1, …, bk such that:

Minimizing unused bandwidth :

$$\text{Min}(b - \sum_{i=1}^{k} b_i T_i)$$

Maximizing bandwidth (bj) in each of the sub-network (Tj):

$$\text{Max } bj(Tj) \qquad (1)$$

Note: (b) can be seen as minimizing the number of sub-networks (parallel services) with maximum bandwidth for each.

We denote by bMax(G, b) the maximum total bandwidths for parallel services and by bMin(G, b) the minimum total unused bandwidths which are represented as the following notations bMax and bMin respectively in the paper.

Note that: $x(E)=[bMax (|V|-1)]+bMin$.

The problem of finding a sum of spanning tree sub-networks with a total minimum unused bandwidth equal to zero is an open problem [20] and known as the integer cover conjecture for spanning trees problem. However, it is a polynomial problem for a special kind of topology network called series parallel networks (Figure 3).

3.2 Definitions:

A parallel extension to a graph G is adding a parallel edge to one edge of G. A series extension of G is subdividing an edge and creating a new vertex and two new adjacent edges.

A series parallel network is a graph obtained by starting from one link and applying recursively series and/or parallel extensions (Fig. 3).

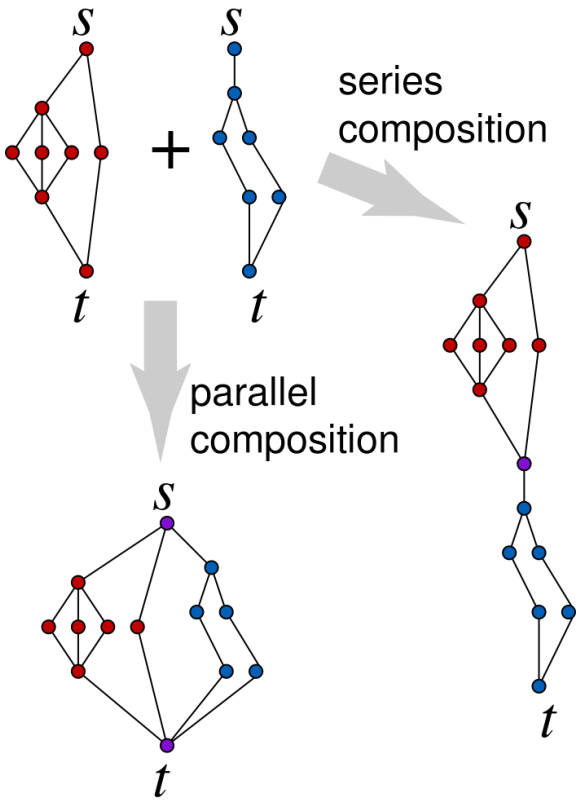

Fig. 3 Series and parallel composition operations for series-parallel graphs

For any series parallel graph G, we denote by s(G) the number of series extensions that we have used to create G. For any vector x defined on the links set E and any subset of links H,

$$x(H) = \sum_{e \in H} x(e) \qquad (2)$$

The x-maximum spanning tree is a spanning tree T of G such that x(T) is maximum
This problem is polynomial [21] and we will use Prim algorithm [22] to solve it when needed.
The x-maxBalanced spanning tree (MBST) is a spanning tree T of G such that the difference between the maximum node degree and the minimum node degree is minimum. If the network is Hamiltonian the balanced spanning tree is a Hamiltonian path.
The main result of this paper is a polynomial algorithm, which gives, using simulation, an exact solution and solves the problem for series parallel networks and any integer bandwidths vector.

3.3 Description of the algorithm (MaxBalanced Spanning Trees Packing: MBSTP)

MBSTP Phases:
Input: A computer network G=(V, E) and a bandwidths vector b.

Phase 1: Find T=b-maxBalanced spanning tree in G using the below algorithm. If there is no spanning tree then Go to Phase 4.

Phase 2: Find bT = b–minimum link in T.

Phase 3: Compute new bandwidths vector b = b – bTT.

Go to Phase 1.

Phase 4: Compute unused bandwidths bMin = b(E).

Algorithm's complexity:

Let m = |E| and n=|V|

Phase 1: The complexity of MBSTP in Phase 1 is at most the complexity of finding a maximum spanning tree in the network. The best algorithm gives a complexity of O(m).
Phase 2: Finding a minimum number from a list of n-1 elements has the complexity of O(n).
Phase 3 has complexity O(n).
Phase 4 has complexity O(m).
As we have to repeat phase 1 through phase 3 at most n times, so the complexity would be of n(O(m)+O(n)+O(n)) = n O(m) = O(n m).
Phase 4 is used only once so the whole complexity is:

O(n m) + O(m) = O(n m).

Since we used Prim's algorithm for this purpose, the worst case of the heuristic is O(|n| log |n|) [22] and [15]. The memory complexity is O(|m|) because we can store series parallel networks by the way they are constructed starting from one edge and indicating the nature of each extension (parallel or series: 0 or 1 for example).

Theorem: If the network is series parallel then MBSTP solves Maximum unused bandwidth (MUB) in a polynomial time.
We will give now a polynomial time algorithm to find a b-maxBalanced spanning tree:

Given a graph G=(V, E) and a weight function w defined on E.
We suppose that the edges are sorted according to their decreasing weights:
w(e1)≥w(e2)≥ … ≥w(em).
E1=E;
For j=1, 2, …, m
While Gj=(V, Ej) is connected do
Find Tj the w-maximum spanning tree in Gj (using Prim's algorithm)
dj=w(ej)-min{w(e) : e belongs to Tj}
Ej+1=Ej-{ej}
End While
End For

MBSTP Pseudo-code

The w-maximum balanced spanning tree is the best balanced from the Tj's, j=1, …, m.
Theorem: The previous algorithm gives the w-maximum balanced spanning tree.
Proof:
Let T be the w-maximum balanced spanning tree, e=ek and f=et be respectively the maximum and the minimum w-edges of T, d=w(e)-w(f) and F={ek, ek+1, …, et}.
It is clear that T is the w-maximum spanning tree in (V, F). So it is also the w-maximum spanning tree in Gk. Then T=Tk and we are done.

4. Simulation Results and Analysis

To demonstrate the effectiveness of our algorithm, we ran two scenarios using an NS2 simulator. The first scenario comprises two sub-scenarios: one, which uses our algorithm and a second, which doesn't. The aim of these sub scenarios is to show that using our algorithm in a series parallel network achieves a higher video stream delivery ratio. However, the aim of the second scenario is to show that using our algorithm in different network topologies minimizes unused bandwidth, thus servicing many sources in parallel by maximizing the bandwidth usage and providing QoS, which is heavily sought in video stream networks.

4.1 Scenario 1:

In the first sub-scenario, the simulation involves a single receiver (sink node) receiving simultaneous video streams from a multimedia server which needs to have a low delay transmission. The receiving node (R1) should create virtual parallel services to maximize the bandwidth usage and minimize the unused bandwidth using our algorithm (MBSTP) before forwarding packets to the sink node (R4). The link capacities (Mbps) are shown in Fig. 4. The multimedia server streams videos (MPEG-1 video sequence) to the receiver (R4).

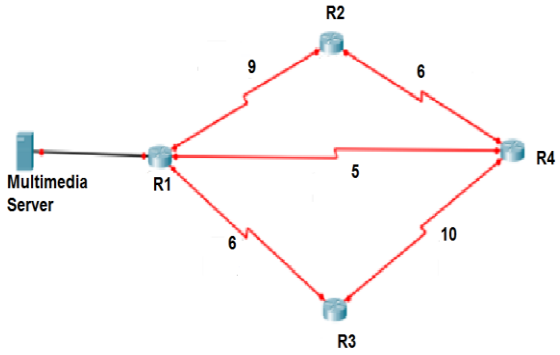

Fig. 4 Network topology

Using MBSTP, Router (R1) starts calculating the different possible virtual spanning trees with total unused bandwidth equal to zero before forwarding any video streams received from the multimedia server. Once done, it will forward them as parallel services based on the size of the video and the remaining bandwidth on the link. Fig. 5 shows the resulting spanning trees with their corresponding bandwidth after running our algorithm. The dotted lines are those links that are not used by the corresponding spanning tree.

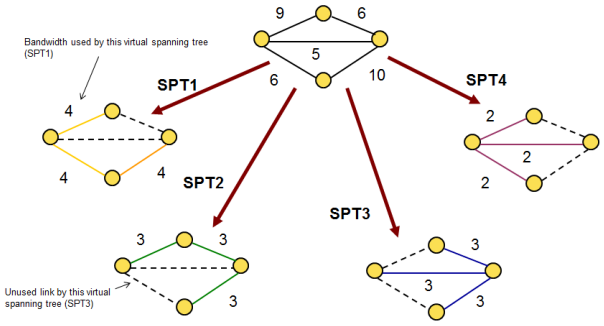

Fig. 5 MBSTP algorithm's spanning trees results

In the second sub-scenario, and using the same topology in Figure 4, the shortest path (SP) approach using Dijkstra's algorithm [23] is used to forward video streams from the server to the sink (R4). Fig. 6 shows both sub scenarios' packet delivery ratio (the number of packets received by the sink divided by the number of packets originated by the source.). Because MBSTP uses maximum available bandwidth through virtual spanning trees to forward packets in a virtual parallel services fashion, it has a higher packet delivery ratio compared to SP that uses only one

shortest path, leaving unused bandwidth in other paths, thus resulting in a lower packet delivery ratio.

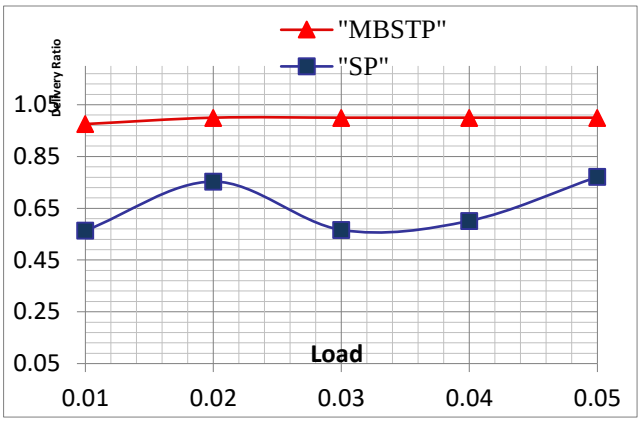

Fig. 6 Packet delivery ratio

4.2 Scenario 2:

The second scenario tests the effectiveness of our developed algorithm (MBSTP) by comparing it with another algorithm, which is the Maximum Spanning tree Packing (MSTP) algorithm [24]. To do so, we run both algorithms on nine different series parallel networks and non-series parallel network generated using NS2, with node numbers varying from 5 to 26 nodes and link numbers from 5 to 38 links. In all simulations, the bandwidth of each link was randomly generated between 20Kbps-40Kbps and the packet size was 500 bytes. The average interval was 1 second, while the latency was set to 500ms.

For each network topology generated (series parallel and non-series parallel) we ran two scenarios using the same simulation parameters. The first one sends a video stream to a receiver using MBSTP in order to minimize unused bandwidth, thus minimizing delay; while the second scenario uses MSTP.

The simulation results are shown in Table 1 and Table 2 respectively.

Table 1: Unused bandwidth in non-series parallel network

Number of Nodes	Number of Links	MBSTP-Total unused bandwidth	MSTP-Total unused bandwidth
6	10	5	6
8	13	2	10
11	19	12	25
14	18	4	6
17	24	10	12
16	27	1	6
21	29	6	35
24	35	23	26
26	38	15	24

Table 2: Unused bandwidth in series parallel network

Number of nodes	Number of links	MBSTP-unused bandwidth	MSTP-unused bandwidth
6	10	0	0
8	13	0	5
11	19	0	4
14	18	0	2
17	24	0	10
16	27	0	5
21	29	0	14
24	35	0	1
26	38	0	2

From Fig. 7 it is clearly shown that the total unused bandwidth for MBSTP is far less than that of MSTP in non-series parallel network. Furthermore, and as expected, in series parallel network, our algorithm has an optimal total unused bandwidth (Unused bandwidth=0) as shown in Fig. 8.

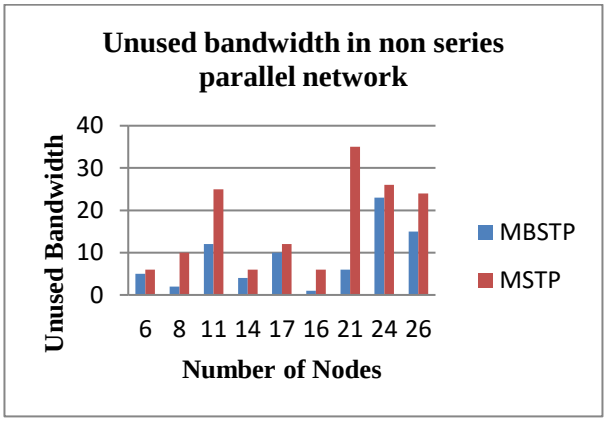

Fig. 7 Unused bandwidth in the non series parallel network

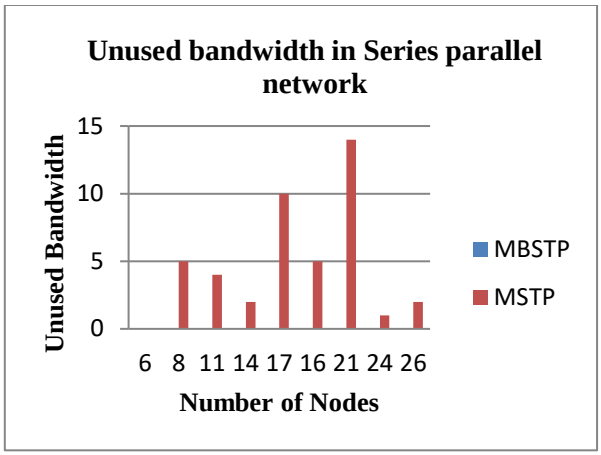

Fig. 8 Unused bandwidth in series parallel network

5. Conclusion

We have shown that our algorithm yields an optimal solution in a series parallel network and outperforms MSTP in a non-series parallel network for a videos streaming. Additionally, the algorithm can be used in series parallel extensions of a network for which an optimal solution is known, even if the initial network is not one unique edge as in series parallel networks. This can be considered when designing a computer network or an extension of a current network into a bigger one. In terms of limitation, the current algorithm is limited by its use only in series parallel network, thus as future work, we can extend it to a more general type of network.
Furthermore, other directions can be investigated involving the mathematical assessment of the optimality or near optimality of the algorithm, considering heuristics for general cases based on MBSTP and considering other types than the spanning tree for virtual sub-networks.

References

[1] D. Niu, C. Feng, and B. Li. Pricing Cloud Bandwidth Reservations under Demand Uncertainty, In ACM SIGMETRICS'12, London, UK, pp. 151–162, Jun 2012

[2] L. Popa, G. Kumar, M. Chowdhury, A. Krishnamurthy, S. Ratnasamy, and I. Stoica, FairCloud: Sharing the Network in Cloud Computing, in ACM SIGCOMM12, Helsinki, Finland, Aug. 2012

[3] J. Guo, F. Liu, X. Huang, J. Lui, M. Hu, Q. Gao, and H. Jin. On Efficient Bandwidth Allocation for Traffic Variability in Datacenters, In IEEE INFOCOM 2014, Toronto, Canada, pp. 1572–1580, Apr. 2014

[4] J. Sachin, V. Navneet, D. Sagnik. An Efficient Solution Approach for Combinatorial Bandwidth Packing Problem with Queuing Delays, Ahmedabad : IIMA, W.P. No. 2014-12-05

[5] A. Amiri , R. Barkhi . The combinatorial bandwidth packing problem. European. Journal of Operation Research, Vol. 208, Issue: 1, pp37–45, 2011

[6] K. Ahuja, B.Singh, and R.Khanna. Network selection based on available link bandwidth in multi-access networks, Communications and Networks, vol. 2, no. 1, pp. 15–23, 2016.

[7] M. Carmo, J. Sá Silva, E. Monteiro, P. Simões, and F. Boavida. Ethernet QoS Modeling in Emerging Scenarios. Proceedings of 3rd International Workshop on Internet Performance, Simulation, Monitoring and Measurement (IPS-MoMe 2005), Warsaw, pp 90-96, 14-15 March 2005

[8] I. Hussain, Z.I. Ahmed, D.K. Saikia, and N. Sarma. A QoS-Aware Dynamic Bandwidth Allocation Scheme for Multi-Hop WiFi-Based Long Distance Networks. EURASIP Journal on Wireless Communications and Networking, vol.1, pp 1-18, 2015 http://dx.doi.org/10.1186/s13638-015-0352-z

[9] J. Niu, Wenfang Song, Mohammed Atiquzzaman. Bandwidth-adaptive partitioning for distributed execution optimization of mobile application, Journal of Network and Computer Applications, vol. 37, pp 334–347, January 2014

[10] N. Xiaoming, He. Yifeng, and G. Ling. Towards Optimal Resource Allocation for Differentiated Multimedia Services in Cloud Computing Environment, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 684 - 688, 4-9 May 2014, DOI:10.1109/ICASSP.2014.6853683

[11] Y. Zhang, Y. Guo, and Y.Chen. Optimized bandwidth allocation for maximizing user's QoE in hybrid cloud P2P content distribution, The Journal of China Universities of Posts and Telecommunications, vol. 15, Issue 3, Jun 2015, DOI: 10.1016/S1005-8885(15)606562

[12] S. O. Krumke and C. Zeck. Generalized max flow in series-parallel graphs. Discrete Optimization, vol. 10, issue 2, pp 155–162, 2013

[13] Anis Elgabli , Vaneet Aggarwal , Shuai Hao, Feng Qian, and Subhabrata Sen, LBP: Robust Rate Adaptation Algorithm for SVC Video Streaming, IEEE/ACM Transactions on Networking, Vol. 26, No. 4, August 2018.

[14] X. Li, M. A. Salehi, M. Bayoumi, and R. Buyya, CVSS: A cost- efficient and QoS-aware video streaming using cloud services, in Proc. 16th IEEE/ACM Int. Conf. Cluster Cloud Grid Computing, May 2016, pp. 106–115.

[15] X. Li, M. A. Salehi, and M. Bayoumi, High perform on-demand video transcoding using cloud services, in Proc. 16th IEEE/ACM Int. Conf. Cluster Cloud Grid Computing, ser. CCGrid'16, May 2016, vol. 16, pp. 600–603.

[16] Rosilah Hassan, Rana Jabbar, End-to-End (e2e) Quality of Service (QoS) For IPv6 Video Streaming. 19th International Conference on Advanced Communication Technology (ICACT), 2017

[17] Xiangbo Li , Mohsen Amini Salehi, Magdy Bayoumi, Transcoding Using Heterogeneous Cloud Services, IEEE Transactions On Parallel And Distributed Systems, Vol. 29, No. 3, March 2018

[18] Chaourar, B., and Mahlous, A. R., Minimizing Unused Bandwidths in Computers Networks, Applied Computing and Informatics 6 (2), 2008, 15-22.

[19] Chaourar, B., Mahlous, A. R., and Fretwell, R. J., On Minimizing Unused Bandwidths in Series Parallel Networks, Journal of King Saud University - Science 21 (Special Issue), 2009, 21-24

[20] J. C. De Pina, and Soares, J. Improved Bound for the Carathéodory Rank of the Bases of a Matroid, Journal of

Combinatorial Theory, Series B, vol. 88, Issue 2, pp. 323-329, July 2003

[21] J. B. Kruskal. On the shortest spanning tree of a graph and the traveling salesman problem, Proceedings of the American Mathematical Society, vol. 7, pp. 48-50, 1956

[22] R. C. Prim. Shortest connection networks And some generalizations, Bell System Technical Journal, vol. 36, Issue 6, pp 1389–1401, November 1957, doi:10.1002/j.1538-7305.1957.tb01515.x

[23] E. W. Dijkstra and C. S. Scholten. Termination Detection for Diffusing Computations, Information Processing Letters, vol. 11, No. 1, pp. 1-4, August 1980

[24] F. Barahona. Packing spanning trees, Journal Mathematics of Operations Research archive, vol. 20, Issue 1, pp 104-115, Feb. 1995