QoE-Oriented Resource Allocation Design Coping with Time-Varying Demands in Wireless Communication Networks

Teweldebrhan Mezgebo Kebedew
SnT, University of Luxembourg
Teweldebrhan.Kebedew@uni.lu

Vu Nguyen Ha
SnT, University of Luxembourg
vu-nguyen.ha@uni.lu

Eva Lagunas
SnT, University of Luxembourg
eva.lagunas@uni.lu

Joel Grotz
SES S.A., Luxembourg
joel.grotz@ses.com

Symeon Chatzinotas
SnT, University of Luxembourg
symeon.chatzinotas@uni.lu

***Abstract*—Efficiently utilizing the network resources to minimize the operation costs while satisfying customer's Quality-of-Experience (QoE) related requirement as well as dynamic demands is a challenging task of all network operators. This paper aims to develop a stationary capacity allocation method that anticipates time-varying demand and keeps the network operating under constraints on a stochastic blocking probability. Queuing delay requirement is also regarded as an QoE-oriented practical design. Employing an approximation of time-varying queuing model and continuous time Markov chain (CTMC) for queue length, the technical designs are stated as a convex stochastic optimization based on which a dynamic capacity allocation is proposed by using Lagrangian and gradient descent searching method. Numerical studies confirm that our proposed framework can efficiently and dynamically allocate optimal capacity for a blocking probability of less than 1% and the probability of violating the queuing-delay requirement is less than 5%.**

***Index Terms*—Time-varying queuing, dynamic BW allocation, blocking probability, QoE-based optimization.**

I. Introduction

In the Internet-of-Thing era, a tremendous number of connections due to various applications are expected to access the wireless communication networks simultaneously [1], [2]. The applications can be related to different services such as voice, video streams, web browsing, etc and their traffic demands vary over the time irregularly and unpredictably [3], [4]. Hence, developing an advanced dynamic resource allocation which can cope with such time-varying demand at the customers is essential research topic in both industry and academia. This challenging design is more critical in the Open-RAN platforms in which service providers (SP) can rent a limited network resource from infrastructure providers (InPs) to serve their customers [5]. In such schemes, the operators should carefully consider the QoE at the customers as well as the renting cost to maximize their revenue. In this paper QoE-aware, queue length based dynamic bandwidth allocation model that minimizes service providers' costs while preserving customers maximum queuing delay and minimal blocking probability is presented.

The study of queuing length and queuing delay is not a new idea, and it has previously been addressed for different purposes in both terrestrial and satellite networks. Authors in [4] used M/M/1 queue model to study the queuing and end-to-end delay distributions under time-varying channels for LEO satellite constellations considering infinite buffer size and fixed transmission capacity for processing the arriving packets. Authors in [6] used Markov Modulated Service Process (MMSP) followed by G/G/1 queue model in a two stage queuing process of satellite data relay networks with an objective of buffer design and transmission resource allocation. However, these works considered the fixed transmission and service rate model which is not practical.

To the best of our knowledge a time-varying-related dynamic capacity allocation framework that simultaneously and efficiently reduces the operation cost while satisfying the QoE at customers has not been studied. This paper is aligned to fill this gap. In particular, our work aims to develop a novel BW assignment mechanism which can be employed at the SP to cope with the time-varying stochastic traffic flows going through the wireless network. This proposed focuses on minimizing the BW-renting cost and maintaining the QoE in terms of target blocking probability and queuing-length-related delay at the users. Modelling the time-varying traffic as an $M_t/M_t/1$ queuing framework, we formulate these technical design conditions into a stochastic optimization problem. To deal with this problem, we first employ a continuous time Markov chain (CTMC) to analyze the stochastic queue-length distribution. Based on the analysis results, the stochastic problem is then re-stated as a convex one. Then, the optimal BW allocation solution framework is developed by using the Lagrangian-duality method. Finally, numerical studies are performed to validate the analytical model and demonstrate the efficiency of the proposed design.

II. System Model and Problem Formulation

This work focuses on a typical wireless network in which the carrier bandwidth (BW) in bps is owned by the infrastructure provider (InP) which can be rent to one service providers (SP) for satisfying time-varying user demand, i.e., the data traffic arriving with time-varying rate. In this scheme, the SP aims to minimize the renting cost while maintaining the QoE requirement at the customers. The network operates in a time-

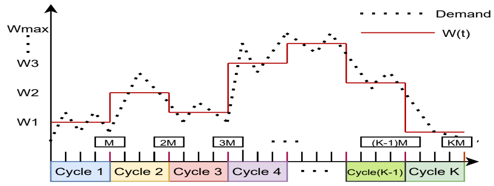

Fig. 1: Bandwidth allocation for time varying demand

slotted manner with a duration of T_p as the transmission time and a time window of T time-slots (TSs) is considered. Let $W(t)$ (bps) be the BW rented by the SP at TS t. Regarding the contract-based and hardware-related requirement, this work assumes that $W(t)$ must be kept unchanged within a cycle duration of M TSs, then it can be updated in the next cycle, i.e illustrated in Fig. 1. Let k denote the cycle index, i.e. $k \in \{1, 2, ..., K\}$ where $K = T/M$. For convenience, hereafter, we sometimes refer $W(t)$ as $W(t) = W_k$ if $t \in ((k-1)M, kM]$. Employing a linear-pricing model, the price of using $W(t)$ (bps) can be described as

$$f(W(t)) = \gamma W(t), \tag{1}$$

where $f(W(t))$ stands for the pricing function corresponding to $W(t)$ and γ is the price in €$/Mbits$. Then, the renting cost that the SP has to pay to InP can be expressed as

$$F_{\mathsf{SP}} = \int_0^{T\times T_p} f(W(t))\, dt = \sum_{\forall k} (MT_p) f(W_k). \tag{2}$$

A. Queuing Model

This work models user demands as data-flows coming from various applications and services, putting them in a queue, and providing them service based on a first-come-first-served (FCFS) discipline. Herein, all applications tending to access the wireless communication platform is considered as one data-flow with a number of arrival packets of L bits. Following the FCFS discipline, the arrived packets are stored in a buffer with maximum length of Q_{max} packets before being propagated thought the operated BW.

Following $M_t/M_t/1$ queuing framework [7]–[9], one assumes that the data packet at TS t following the Poison Process with a time-varying arrival rate $\lambda(t)$, e.g., $\lambda(t)$ changes over the time as shown in Fig. 1. Regarding the BW assigned for this data-flow as $W(t)$ rented from InP, the service rate estimated in number of packets in TS t can be given as

$$\mu(t) = W(t)T_p/L. \tag{3}$$

B. QoE Requirement and Problem Formulation

Let $P_n(t) = Pr\{Q(t) = n\}$ be the probability that there exist n packets in the buffer, so-called queuing length (QL) of n. Given by [7], the probability distribution of QL $Q(t)$ is a function of TS t. This paper aims to maintain the QoE at the users in the manner of transmission delay which can be presented by the QL. In particular, this work aims to keep the probability that the QL is longer than $\bar{Q}$ packets for overall window time $[0, T]$ less than a commitment factor $\bar{P}_{\mathsf{QoE}}$, i.e.

$$(1/T)\int_0^T Pr\{Q(t) \geq \bar{Q}\}dt \leq \bar{P}_{\mathsf{QoE}}. \tag{4}$$

Regarding the network operation requirement, it is worth noting that the network blocks user admission if the QL violates the maximum buffer length. Then, in order to let the network operate smoothly, the QL should not violate the maximum buffer length too frequently. This requirement is transferred to that the blocking probability should be less than a predetermined target in any TS which is expressed as

$$Pr\{Q(t) \geq Q_{\mathsf{max}}\} \leq \bar{p} \quad \forall t, \tag{5}$$

where $\bar{p}$ is the target blocking probability. Taking into account of that $Pr\{Q(t) \geq N\} = 1 - \sum_{n=1}^{N} P_n(t)$, our technical designs can be formulated into a statistical optimization problem as

$$\min_{\mathbf{W}} \sum_{\forall k} MT_p f(W_k) \;\text{ s.t. }\; \sum_{n=0}^{Q_{\mathsf{max}}} P_n(t) \geq 1 - \bar{p}, \forall t, \tag{6a}$$

$$(1/T)\int_0^T \sum_{n=0}^{\bar{Q}} P_n(t)dt \geq 1 - \bar{P}_{\mathsf{QoE}}, \tag{6b}$$

where $\mathbf{W}$ represents the vector of all W_k's.

III. Queuing Stochastic Analysis and Problem Approximation

A. Queuing Stochastic Analysis

Considering the time-varying queuing model, the stochastic QL can be represented as a CTMC where the queue length follows a birth death process [7]. Queue length values can be obtained as a transient solution of the Kolmogorov equation [7]. However, this equation does not provide explicit solutions to the transition probabilities. Therefore, a more appropriate solution is to find approximate solutions [7]–[9] instead, the transient probabilities can be approximated by a cumulative distribution function as, $F(Q, t) = p(Q(t) \leq Q)$. Then, the expected value of the queue distribution yields

$$P_n(t) \approx \int_n^{n+1} F(Q, t)dQ. \tag{7}$$

Regarding the system utilization, which is expressed as $\rho(t) = \lambda(t)/\mu(t)$, the probability that there are n packets in the buffer can be approximated as [8]

$$\bar{P}_n(t) = \int_n^{n+1} F(Q)dQ \approx \rho(t)^n(1 - \rho(t)), \tag{8}$$

where $\rho(t) \leq 1$. It notes that $\rho(t) > 1$ implies that the blocking probability tends to one [8]. The good accuracy of this analysis result is confirmed by the simulation results demonstrated in Section V-A. Thanks to this analysis result, problem (6) can be approximated in the following section.

B. Problem Approximation

Substituting $\mu(t)$ in (3) with regarding of that $W(t) = W_k$ for $t \in ((k-1)M, kM]$ and $k = 1, .., K$, one can rewrite $\bar{P}_n(t)$ as a function of t and W_k as

$$\bar{P}_n(t) = g_n(W_k, t), \forall k \text{ and } t \in ((k-1)M, kM], \tag{9}$$

where $g_n(W_k, t) = -\left(\frac{L\lambda(t)}{T_p W_k}\right)^n \left(1 - \frac{L\lambda(t)}{T_p W_k}\right)$. To ease the presentation, we denote the set $((k-1)M, kM]$ as Ω_k. Employing $\bar{P}_n(t)$, problem (6) can be restated as

$$\min_{\mathbf{W}} \;\; F_{\mathsf{SP}} = MT_p \sum_{k=1}^{K} \gamma W_k \tag{10a}$$

$$\text{s. t.} \quad \sum_{n=0}^{Q_{\max}} g_n(W_k, t) \geq 1 - \bar{p}, \forall k \text{ and } t \in \Omega_k, \tag{10b}$$

$$(1/T)\sum_{k=1}^{K} \int_{(k-1)M}^{kM} \sum_{n=0}^{\bar{Q}} g_n(W_k, t)dt \geq 1 - \bar{P}_{\mathsf{QoE}}, \tag{10c}$$

$$0 \leq \rho(t) \leq 1, \ \forall t, \tag{10d}$$

where constraint (10d) aims to keep the network stable [7]–[9].

IV. Dynamic Resource Allocation Design

A. Problem Convexity Characterization

In order to solve problem (10), we first characterize its convexity. To begin with, we first define the lower bound BW amount required in cycle in which $\mu(t)$ is fixed by considering the following proposition.

Proposition 1. *Constraints (10b) and (10d) in problem (10) can be merged into one constraint as*

$$W_k \geq \alpha_k = \max\left(\alpha_{k,1}, \alpha_{k,2}\right), \forall k, \tag{11}$$

where $\alpha_{k,1} = \max_{t\in\Omega_k} L\lambda(t)/T_p$, $\alpha_{k,2} = \max_{t\in\Omega_k} L g^{-1}_{Q_{\max}}(1-\bar{p}, t)/T_p$, *and* $g^{-1}_{Q_{\max}}(p,t)$ *is the inverse function of* $\sum_{n=0}^{Q_{\max}} g_n(W_k, t)$.

Proof: Denote $\mu(t) = \mu_k$ in cycle k-th. Then, we have $\rho(t) = \lambda(t)/\mu_k$, $\forall t \in \Omega_k$. Then, constraint (10d) can be transferred into the following requirement

$$W_k \geq \alpha_{k,1} = \max_{t\in\Omega_k} L\lambda(t)/T_p. \tag{12}$$

Similarly, the constraint (10b) will be equivalent to

$$W_k \geq \alpha_{k,2} = \max_{t\in\Omega_k} L g^{-1}_{Q_{\max}}(1-\bar{p}, t)/T_p. \tag{13}$$

The results given in (12) and (13) yield the lower bound of W_k as $\alpha_k = \max\left(\alpha_{k,1}, \alpha_{k,2}\right)$, which has completed the proof of Proposition 1. ■

In the next move, we state the convexity of problem (10) based on the result in this proposition in the following theorem.

Theorem 1. *Problem (10) is convex.*

Proof: Proof is given in Appendix A. ■

Thanks to Proposition 1 and Theorem 1, problem (10) can be rewritten as the following convex problem,

$$\min_{\mathbf{W}} \sum_{\forall k} MT_p\gamma W_k \ \text{s.t.} \ (11) \text{ and } \sum_{\forall k} z_k(W_k) \geq 1 - \bar{P}_{\mathsf{QoE}}, \tag{14}$$

where $z_k(W_k)$ is represented in (22). In the following section, a dynamic resource allocation algorithm based on Lagrangian duality optimization approach is proposed to obtain the optimal solution of problem (14).

Remark 1. *It is worth noting that obtaining the optimal solution of problem (6) is very challenging since there is no closed-form formula of $P_n(t)$. Then, instead of dealing with problem (6), this work aims to approximate it to problem (10) by using the queuing stochastic approximation results given in Section III-A. In addition, Theorem 1 only mentions about the convexity of problems (10) (not about problem (6)). Hence, the solution presented in what follows can only achieve the optimal solution of problem (14) (equivalent to problem(10)). How efficient of the queuing stochastic approximation approach will be discussed in the numerical results.*

Algorithm 1 Duality-based Dynamic BW Allocation Algorithm

Initialization:
- Choose an initial dual value $\beta^{[0]}$.
- Select a tolerance ϵ, step size δ, and set $\Delta = 1$, $\ell = 0$.

while $\Delta > \epsilon$ **do**
  For given $\beta^{[\ell]}$, define $W_k^{\star}$'s as in (19).
  Based on $W_k^{\star}$'s, update $\beta^{[\ell+1]}$ as in (17).
  Re-set $\Delta := |\beta^{[\ell+1]} - \beta^{[\ell]}|$.
  Set $\ell := \ell + 1$.
end while
Return $\mathbf{W}^{\star}$.

B. Duality-based Dynamic BW Allocation Algorithm

1) Duality Approach: We first define the Lagrangian function $\mathcal{L}$ associated with (14) as,

$$\mathcal{L}(\mathbf{W}, \beta) = MT_p\sum_{\forall k}\gamma W_k - \beta\left(\sum_{\forall k} z_k(W_k) - 1 + \bar{P}_{\mathsf{QoE}}\right), \tag{15}$$

where β is the Lagrangian multiplier. Then, the dual function of W_k can be defined as

$$\mathsf{g}(\beta) = \min_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \beta) \ \text{s.t.} \ (11). \tag{16}$$

To find the best lower bound that can be obtained from the Lagrange dual function, the dual problem can be written as, $\max_{\beta} \mathsf{g}(\beta)$ s.t. $\beta \geq 0$. Since problem (14) is convex, the dual-gap between primary and dual problem is zero. In the following, one will describe a searching approach to define the optimal solution. In particular, the dual problem is always convex, $\mathsf{g}(\beta)$ can be maximized by using the standard sub-gradient method where the dual variable β can be iteratively updated as follows:

$$\beta^{[\ell+1]} = \left[\beta^{[\ell]} - \delta^{[\ell]}\left(\sum_{\forall k} z_k(W_k) - 1 + \bar{P}_{\mathsf{QoE}}\right)\right]^{+}, \tag{17}$$

where the suffix $[\ell]$ represents the iteration index, $\delta^{[\ell]}$ is the step size, and $[x]^{+}$ is defined as $\max(0, x)$. This sub-gradient method guarantees the convergence for any initial primary point of $\{W_k\}$'s if the step-size $\delta^{[\ell]}$ is chosen appropriately so that $\delta^{[\ell]} \xrightarrow{\ell\to\infty} 0$ such as $\delta^{[\ell]} = 1/\sqrt{\ell}$ [5], [10].

Then, to solve problem (14), one can alternatively solve the problem on the right-hand-side of (16) and updating β as in (17) in each iteration until the convergence. The solution approach is summarized in Algorithm 1 where the remaining task of solving the problem in the right-hand-side of (16) is considered in the following section.

2) Solving the optimization problem related to dual function: This section developed an approach of solving the following problem for given β,

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}, \beta) \ \text{s.t.} \ (11). \tag{18}$$

In particular, the solution of this problem is given in the following proposition.

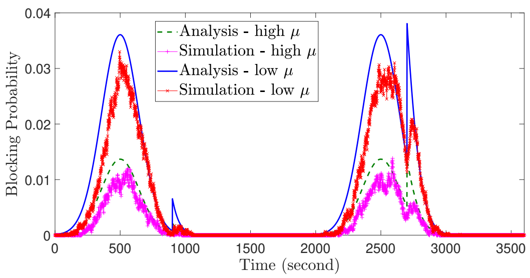

Fig. 2: Stochastic blocking probability value over time.

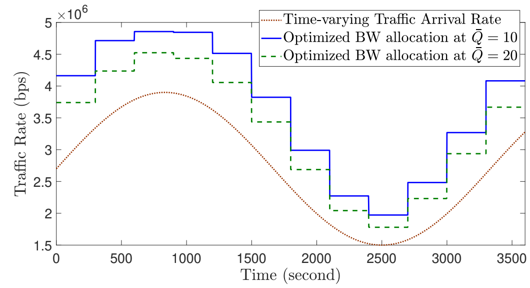

Fig. 3: BW allocation at different values of $\bar{Q}$.

Proposition 2. *The optimal solution of* (18) *is given as*

$$W_k^\star = \max(\alpha_k, \hat{W}_k), \quad (19)$$

where $\hat{W}_k$ *is defined as* $\hat{W}_k = \left[\beta A_k / (M T_p \gamma T \bar{Q})\right]^{1/(\bar{Q}+1)}$.

Proof: As can be observed, the minimum value of $\mathcal{L}(\mathbf{W}, \beta)$ can be defined by equating the partial derivative of $\mathcal{L}(\mathbf{W}, \beta)$ with respect to W_k to zero, i.e.,

$$\partial \mathcal{L}(\mathbf{W}, \beta)/\partial W_k = M T_p \gamma - \beta A_k / (T \bar{Q} W_k^{\bar{Q}+1}) = 0 \quad (20)$$

The solution of this equation can be described as $\hat{W}_k = \left[\beta A_k / (M T_p \gamma T \bar{Q})\right]^{1/(\bar{Q}+1)}$. Then, by considering constraint (11), the optimal value of W_k can be expressed as in (19), which completes the proof of Proposition 2. ■

V. Numerical Results

A. Queue Model Simulation and Analysis Demonstration

In Fig. 2, one illustrates the gap between blocking probabilities obtained by employing the queuing stochastic analysis consequence given in Section III-A and the Monte-Carlo method. In this simulation, we model a time-varying queuing model with the arrival-rate and service-rate functions, $\lambda(t)$ and $\mu(t)$, in a period duration of 3600 secs. In particular, $\lambda(t)$ is set as $\lambda(t) = 9 + 4\sin(0.001\pi t)$ for $t \in \{1, 2, ..., 3600\}$ while $\mu(t)$ is chosen as $\mu(t) = \mu_k$ if $t \in ((k-1)M, kM]$ where $M = 900$ secs, i.e, $\mu(t)$ changes every 15 minutes. Two scenarios are considered for service-rate setting, which are *"low μ"* where $[\mu_1, \mu_2, \mu_3, \mu_4] = [16, 14, 16, 15]$ and *"high μ"* where $[\mu_1, \mu_2, \mu_3, \mu_4] = [17, 15, 17, 16]$. For Monte-Carlo simulation, 10,000 independent data-demand trials following time-varying arrival-rate function $\lambda(t)$ are generated for estimating the average blocking probability. Here, Q_{max} is set to 15, T_p to 20ms and the packet length to 6000bits. Fig. 2 demonstrates the well matching between the analysis and simulation results which confirms the approximation result given in Section III-A as well as the work presented in [8].

B. Dynamic BW Allocation Results and Discussion

This section provides numerical results demonstrating BW allocation, the renting costs as well as the blocking probability corresponding the outcomes of our proposed dynamic BW allocation algorithm. In these simulation, the arrival-rate function is kept the same as presented in previous section while $\mu(t)$ are defined by implementing Algorithm 1 for various values of $\bar{Q}$, Q_{max}, $\bar{p}$, $\bar{P}_{\text{QoE}}$, and K.

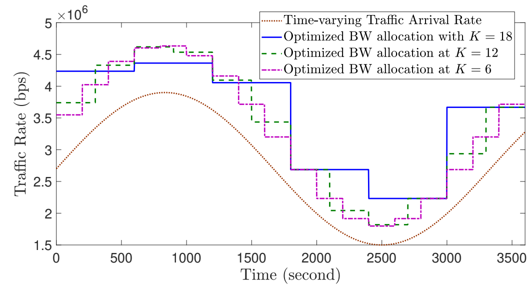

Fig. 4: BW allocation at different values of K.

TABLE I: Renting cost at $Q_{max} = 40$, $\bar{Q} = 20$ for different K.

K	18	12	10	6
Cost per hour (10^4)	1.1828	1.2042	1.22	1.2875

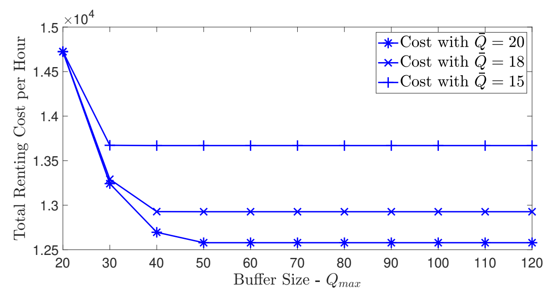

Fig. 5: Total renting cost versus Q_{max} at different values of $\bar{Q}$.

In particular, Fig. 3 illustrates the BW allocation solution of proposed algorithm and the average demand over time for $\bar{Q} = 10$ and 20. As can be observed, the higher $\bar{Q}$, which implies the wider tolerance on QoE requirement, yields the lower required network BW. Fig. 4 shows the required network BW over time for various number of cycles, i.e., $k = 6, 12, 18$. As expected, the higher the number of W_k-updating time within the same time-window we set, the better tracking of BW assignment to traffic demand and the lower renting cost one returns (as shown in Table I). Furthermore, in these both figures, the required BW is always greater than the average demand. This implies the impact of QoE requirement on the BW allocation design since in queuing theory, the change in QL increases when $\rho(t) = \lambda(t)/\mu(t) \to 1^-$.

Figs. 5 and 6 illustrate the renting cost due to different $\bar{Q}$ versus the buffer size (Q_{max}) and target blocking probability ($\bar{P}$), respectively, here $\gamma = 0.1€/Mbits$ is considered. As

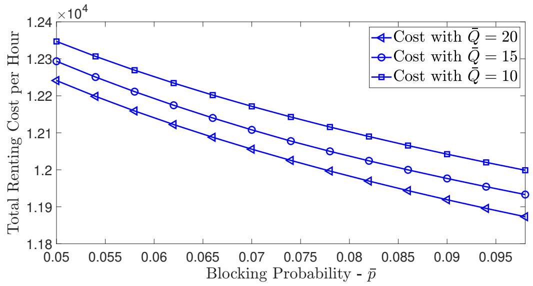

Fig. 6: Total allocated bandwidth cost verses $\bar{p}$ at different values of $\bar{Q}$.

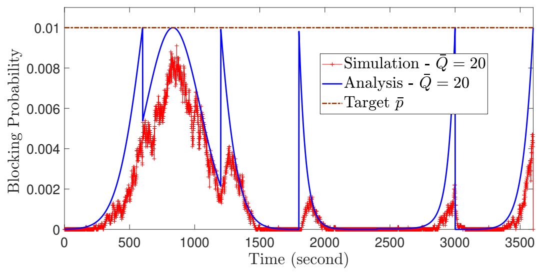

Fig. 7: Blocking probability over time.

TABLE II: Simulated and analysis blocking probability comparison.

$\bar{Q}$	11	15	20	25	30
Simulation	0.0364	0.0081	0.0019	0.00044	0.00008
Analysis	0.0443	0.015	0.0044	0.0013	0.00037

seen in Fig. 5, when the buffer size is increased, the renting cost decreases before saturating. In addition, the higher value of $\bar{Q}$ returns the lower renting cost. This can be explained by that the larger $Q_{\max}$ or $\bar{Q}$ lessens the QoE requirement; hence, the required BW can be set lower. Regarding Fig. 6, as expected, the renting cost can be degraded by increasing the target blocking probability.

Fig. 7 illustrates the blocking probability achieved by analysis result and Monte-Carlo simulation based on the required BW returned by implementing Algorithm 1 for the setting of $K = 6$, $\bar{p} = 1\%$. As can be seen, both analysis and simulation results satisfied the target blocking probability requirement. In addition, Table II also delivers the probability of violating QoE requirements corresponding to various $\bar{Q}$, setting target $\bar{Q} = 10$, $\bar{P}_{\text{QoE}} = 5\%$. Similar to the results given in Section V-A, the well matching between the analysis and simulation results is again confirmed in Fig. 7 and Table II.

VI. Conclusion

This paper proposed a novel dynamic BW allocation minimizing the renting cost at SP which can cope with time-varying demands and the QoE requirement in term of queuing-length delay. $M_t/M_t/1$ queuing model has been used to model the traffic flows in this work. Analysis and simulation results have been proposed to demonstrate the efficiency of our proposed algorithm as well as the relations between required BW allocation, queue length, and target blocking probability.

Appendix A
Proof of Theorem 1

As can be observed, the objective function of (10) is given as a a linear function of W_k. In addition, according to Proposition 1, constraints (10b) and (10d) can be merged into (11) which is also in linear form. Therefore, the proof can be completed by showing the convexity of constraint (10c) as following details.

Let $t_{\bar{Q}}(W_k, t) = \sum_{n=0}^{\bar{Q}} g_n(W_k, t)$. Taking some minor mathematical manipulators, one can express $t_{\bar{Q}}(W_k, t)$ as,

$$t_{\bar{Q}}(W_k, t) = 1 - (C(t)/W_k)^{\bar{Q}+1}. \qquad (21)$$

where $C(t) = L\lambda(t)/T_p$. It is not difficult to see that $t_{\bar{Q}}(W_k, t)$ is a concave function with respect to W_k for any value of $\lambda(t)$ that satisfies $\lambda(t)/\mu_k < 1$. Using notation $t_{\bar{Q}}(W_k, t)$, we further denote $z_k(W_k) = \frac{1}{T}\int_{kM}^{(k+1)M} t_{\bar{Q}}(W_k, t)dt$. Again taking some minor mathematical manipulators, we have

$$z_k(W_k) = M/T - A_k/\left(TW_k^{\bar{Q}+1}\right), \qquad (22)$$

where $A_k = \int_{(k-1)M}^{kM} C(t)^{\bar{Q}+1}dt$. Similar to $t_{\bar{Q}}(W_k, t)$, $z_k(W_k)$ is also a concave function of W_k. Then, constraint (10c) can be rewritten as $\sum_{\forall k} z_k(W_k) \geq 1 - \bar{P}_{\text{QoE}}$, which is convex since it is the form of concave function greater than a constant. Hence, problem (10) must be convex.

References

[1] Y. Xu, F. Yin, W. Xu, J. Lin, and S. Cui, "Wireless traffic prediction with scalable gaussian process: Framework, algorithms, and verification," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1291–1306, 2019.

[2] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[3] Wafanet, "Satellite Bandwidth - Shared or Dedicated," http://www.wafa.ae/en/vsat/faq/Satellite-Internet-faqs-Shared-or-Dedicated.aspx, 2022, [Online; accessed 26-June-2022].

[4] N. J. H. Marcano, L. Diez, R. A. Calvo, and R. H. Jacobsen, "On the queuing delay of time-varying channels in low earth orbit satellite constellations," *IEEE Access*, vol. 9, pp. 87 378–87 390, 2021.

[5] V. N. Ha and L. B. Le, "End-to-end network slicing in virtualized ofdma-based cloud radio access networks," *IEEE Access*, vol. 5, pp. 18 675–18 691, 2017.

[6] Y. Zhu, M. Sheng, J. Li, D. Zhou, and Z. Han, "Modeling and performance analysis for satellite data relay networks using two-dimensional markov-modulated process," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3894–3907, 2020.

[7] W. Whitt, "Time-varying queues," *Queueing models and service management*, vol. 1, no. 2, 2018.

[8] D. Armbruster, S. Göttlich, and S. Knapp, "Continuous approximation of $m\_t/m\_t/1$ distributions with application to production," *arXiv preprint arXiv:1807.07115*, 2018.

[9] G. F. Newell, "Queues with time-dependent arrival rates i—the transition through saturation," *Journal of Applied Probability*, vol. 5, no. 2, pp. 436–451, 1968.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787