Ultra-Reliable and Low-Latency Vehicular Transmission: An Extreme Value Theory Approach

Chen-Feng Liu, *Student Member, IEEE*, and Mehdi Bennis, *Senior Member, IEEE*

Abstract—Considering a Manhattan mobility model in vehicle-to-vehicle networks, this work studies a power minimization problem subject to second-order statistical constraints on latency and reliability, captured by a network-wide maximal data queue length. We invoke results in *extreme value theory* to characterize statistics of extreme events in terms of the maximal queue length. Subsequently, leveraging Lyapunov stochastic optimization to deal with network dynamics, we propose two queue-aware power allocation solutions. In contrast with the baseline, our approaches achieve lower mean and variance of the maximal queue length.

Index Terms—5G, ultra-reliable low latency communications (URLLC), vehicular communications, finite blocklength, extreme value theory.

I. INTRODUCTION

VEHICLE-TO-VEHICLE (V2V) communication is one of the most promising enablers for intelligent transportation systems in which latency and reliability are prime concerns. Nevertheless, the vast majority of the existing V2V literature does not address latency and reliability while some others focus on the coverage probability of radio signal transmission [1]. To ensure ultra-reliable low latency communication (URLLC), queuing latency plays a pivotal role when the traffic arrival and service rates are dynamic and non-deterministic. Particularly in V2V communication, the quality of wireless links varies significantly due to vehicles' high mobility. The authors in [2] take into account the dynamics of queue length and aim at bounding the average queue length within a finite value. While interesting, focusing only on average performance metrics (e.g., average queue length and average delay) is not sufficient to enable URLLC, which instead requires looking into the higher-order statistics or the tail behavior of the distribution. To this end, we define a new reliability measure in terms of maximal queue length among all vehicle pairs and characterize its statistics. Analyzing the statistics of the network-wide maximal queue length provides key insight for the URLLC system design. The studied problem is cast as a power minimization problem subject to statistical constraints on the network-wide maximal queue length. However, to get the network-wide maximal queue length, all vehicles and the roadside unit (RSU) need to exchange queue state information (QSI) which can incur significant signaling overhead in V2V communication. To alleviate this issue, we leverage principles of *extreme value theory* (EVT) [3] to locally characterize

the maximal queue length, which is incorporated as a constraint into the stochastic optimization problem. Our proposed solutions include one semi-centralized and one distributed extreme queue-aware power allocation approaches for V2V communication. Numerical results show the effectiveness of using EVT for the study of ultra-reliable and low-latency vehicular communication.

II. SYSTEM MODEL

We consider a Manhattan mobility model (i.e., grid road topology in urban areas) in which a set $\mathcal{K}$ of K vehicular user equipment (VUE) transmitter-receiver pairs transmits over a set $\mathcal{N}$ of N resource blocks (RBs) with equal bandwidth W. In each pair, the transmitter-receiver association is fixed during the communication lifetime. One RSU is deployed to coordinate the network. We further assume that the communication timeline is slotted and indexed by t. The instantaneous channel gain, including path loss and channel fading, from the transmitter of pair k to the receiver of pair k' over RB n in slot t is denoted by $h_{kk'}^n(t)$. Thus, given VUE pair k's transmit power $P_k^n(t)$ over RB n in slot t with $\sum_{n \in \mathcal{N}} P_k^n(t) \leq N P_{\max}$, the VUE pair k's transmission rate in time slot t is expressed as $R_k(t) = \sum_{n \in \mathcal{N}} W \log_2 \left(1 + \frac{P_k^n(t) h_{kk}^n(t)}{N_0 W + \sum_{k' \in \mathcal{K} \setminus k} P_{k'}^n(t) h_{k'k}^n(t)} \right)$. Here, $P_{\max}$ and N_0 are the power budget per RB and the power spectral density of the additive white Gaussian noise, respectively. Moreover, each VUE transmitter has a queue buffer to store the data destined to its VUE receiver. Denoting VUE pair k's queue length in slot t as $Q_k(t)$, the queue dynamics is given by $Q_k(t+1) = \max \{ Q_k(t) + \lambda_k(t) - T_c R_k(t), 0 \}$, where T_c is the time slot length, and $\lambda_k(t)$ is the traffic arrival at the transmitter of VUE pair k in slot t with the average arrival rate $\lambda_{\text{avg}} = \mathbb{E}[\lambda_k(t)]/T_c$. We also assume that traffic arrivals are independent and identically distributed *(i.i.d.)* among VUE pairs. In order to mitigate interference coming from simultaneous transmissions on the same RB, the RSU clusters vehicles into $g > 1$ disjoint groups based on their geographic locations in which nearby VUE pairs are grouped together, and all RBs are orthogonally allocated within each group. Note that the vehicles' geographic locations vary slowly with respect to the slotted time length (i.e., coherence time of fading channels). Therefore, the RSU clusters VUE pairs and allocates RBs in a long timescale, i.e., every $T_0 > 1$ time slots. Vehicle grouping is done by means of *spectral clustering* [4]. In this regard, firstly denoting $\mathbf{v}_k \in \mathbb{R}^2$ as the midpoint Euclidean coordinate of the VUE transmitter-receiver pair k, we use the distance-based Gaussian similarity matrix $\mathbf{S}$ to represent the geographic proximity information, in which the (k, k')-th element is defined as $s_{kk'} := e^{-\|\mathbf{v}_k - \mathbf{v}_{k'}\|^2/\zeta^2}$ if $\|\mathbf{v}_k - \mathbf{v}_{k'}\| \leq \phi$, and $s_{kk'} := 0$ otherwise. Here, ϕ captures the neighborhood size while ζ controls the impact of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2018.2828407, IEEE Communications Letters

2

Algorithm 1 Spectral Clustering for VUE Grouping

1: Calculate matrix $\mathbf{S}$ and the diagonal matrix $\mathbf{D}$ with the i-th diagonal element $d_{ii} = \sum_{j=1}^{K} s_{ij}$.

2: Let $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_g]$ in which $\mathbf{u}_g$ is the eigenvector of the g-th smallest eigenvalue of $\mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$.

3: Numerically, e.g., by Matlab, use the k-means clustering approach to cluster K normalized row vectors (which represent K VUE pairs) of matrix $\mathbf{U}$ into g groups.

---

neighborhood size. Subsequently, $\mathbf{S}$ is used to group VUE pairs using spectral clustering as shown in Algorithm 1. After forming the groups, the RSU orthogonally allocates all RBs to the VUE pairs in each group. Herein, we further denote VUE pair k's available RBs as a set $\mathcal{N}_k$ which implicitly imposes $P_k^n(t) = 0, \forall n \notin \mathcal{N}_k$, and modify the power constraints as

$$\sum_{n \in \mathcal{N}_k} P_k^n(t) \leq NP_{\max} \text{ and } P_k^n(t) \geq 0, \ \forall t, n \in \mathcal{N}_k, \quad (1)$$

for all VUE pairs $k \in \mathcal{K}$. Additionally, since the RBs are reused by distant VUE transmitters in multiple groups, we treat the aggregate interference power as a constant term I and approximate the transmission rate as $R_k(t) \approx \sum_{n \in \mathcal{N}_k} W \log_2\left(1 + \frac{P_k^n(t)h_{kk}^n(t)}{N_0 W + I}\right)$.

III. EXTREME QUEUE-AWARE POWER ALLOCATION

A. RSU-Aided Power Allocation

As motivated in Section I, this work is concerned about the maximal queue length among all VUE pairs which is mathematically defined as $M(t) := \max_{k \in \mathcal{K}}\{Q_k(t)\}$ in slot t. The network-wide maximal queue length also reflects the worst-case sustained queuing delay. As a reliability measure, we leverage the notion of *risk* in financial mathematics, where risk is synonymous with gaining or losing something valuable. In our considered V2V communication, higher delay (or queue length) can result in an urgent-message loss undermining traffic safety. Therefore, to ensure reliable V2V communication, we aim at minimizing the "risk". To do that, we use the entropic risk measure $\ln(\mathbb{E}[e^{\delta M(t)}])/\delta$ with a risk-sensitivity parameter $\delta > 0$ as our reliability metric [5]. Imposing a threshold κ on the the entropic risk measure, i.e., $\lim_{t \to \infty} \ln(\mathbb{E}[e^{\delta M(t)}])/\delta \leq \kappa$, we aim at minimizing the VUEs' long-term transmit power consumption. By taking the Maclaurin series expansion, we get $\ln(\mathbb{E}[e^{\delta M(t)}])/\delta = \mathbb{E}[M(t)] + \frac{\delta}{2}\text{Var}(M(t)) + \mathcal{O}(\delta^2)$. Next, we focus on the mean and variance of $M(t)$ by considering $0 < \delta \ll 1$, and leave the studies of other high-order statistics, e.g., skewness, for future works. Thus, the studied problem is formulated as

$$\underset{\mathbf{P}(t)}{\text{minimize}} \quad \lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{k \in \mathcal{K}}\sum_{n \in \mathcal{N}_k} P_k^n(t) \quad (2a)$$

$$\text{subject to} \quad \lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[M(t)] \leq \bar{M}_{\text{th}}, \quad (2b)$$

$$\lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[(M(t))^2] \leq \bar{B}_{\text{th}}, \quad (2c)$$

with $\mathbf{P}(t) = (P_k^n(t), k \in \mathcal{K}, n \in \mathcal{N}_k)$ satisfying (1) and $\bar{B}_{\text{th}} = (\bar{M}_{\text{th}})^2 + 2(\kappa - \bar{M}_{\text{th}})/\delta$. To solve problem (2), we use tools from Lyapunov stochastic optimization to dynamically allocate VUEs' transmit power. In order to ensure (2b) and

(2c), we respectively introduce two virtual queues which evolve as follows:

$$Q^{(M)}(t+1) = \max\left\{Q^{(M)}(t) + M(t+1) - \bar{M}_{\text{th}}, 0\right\}, \quad (3)$$

$$Q^{(B)}(t+1) = \max\left\{Q^{(B)}(t) + [M(t+1)]^2 - \bar{B}_{\text{th}}, 0\right\}. \quad (4)$$

Due to space limitations, we skip the rest of the derivations related to the Lyapunov optimization. The interested readers please refer to [6] for the details. Here, we directly show the results after applying Lyapunov optimization. In each slot t, each VUE pair $k \in \mathcal{K}$ solves the convex optimization problem,

$$\underset{P_k^n(t)}{\text{minimize}} \sum_{n \in \mathcal{N}_k}\left[VP_k^n(t) - J_k(t)\log_2\left(1 + \frac{P_k^n(t)h_{kk}^n(t)}{N_0 W + I}\right)\right] \quad (5)$$

with $P_k^n(t)$ satisfying (1) and $J_k(t) = WT_c\big[Q^{(M)}(t) + \big(2Q^{(B)}(t) + 1\big)\big(Q_k(t) + \lambda_k(t)\big) + 2\big(Q_k(t) + \lambda_k(t)\big)^3\big]$. Here, the parameter $V \geq 0$ trades off the power cost optimality and queue length reduction of (2). Applying the Karush-Kuhn-Tucker (KKT) conditions to (5), the VUE transmitter finds a transmit power $P_k^{n*}(t) > 0, \forall n \in \mathcal{N}_k$, which satisfies $\frac{J_k(t)h_{kk}^n(t)}{(N_0 W + I + P_k^{n*}(t)h_{kk}^n(t))\ln 2} = V + \eta$, if $\frac{J_k(t)h_{kk}^n(t)}{(N_0 W + I)\ln 2} > V + \eta$. Otherwise, $P_k^{n*}(t) = 0$. Moreover, the Lagrange multiplier η is 0 if $\sum_{n \in \mathcal{N}_k} P_k^{n*}(t) < NP_{\max}$, and we have $\sum_{n \in \mathcal{N}_k} P_k^{n*}(t) = NP_{\max}$ when $\eta > 0$. Note that given a small value of V, the derived power $P_k^{n*}(t)$ provides a sub-optimal solution to problem (2) whose optimal solution is asymptotically obtained by increasing V. After sending data, the VUE pair k updates $Q_k(t+1)$ for the next time slot $t+1$. Note that to obtain $J_k(t)$ at the VUE, the RSU requires all VUEs' QSI in each time slot to calculate $M(t)$, update (3) and (4), and feed $Q^{(M)}(t)$ and $Q^{(B)}(t)$ back to all VUE pairs. However, frequent information exchange between the RSU and VUEs incurs significant overhead. To address this issue, we propose a solution based on EVT to locally characterize the distribution of the network-wide maximal queue length.

B. EVT-Based Power Allocation

Theorem 1 (Fisher–Tippett–Gnedenko theorem [3]**).** *Given K i.i.d. random variables (RVs), $Q_1, \cdots, Q_K$, and defining $M := \max\{Q_1, \cdots, Q_K\}$, as $K \to \infty$, we can approximate M as a generalized extreme value (GEV) distributed RV which is characterized by three parameters $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$. The support of M is $\{m: 1 + \xi(m - \mu)/\sigma \geq 0\}$.*

Considering that VUE pairs are uniformly distributed on the lanes, we can assume that VUEs' transmission rates are i.i.d. since $R_k(t)$, approximately, does not vary with the other VUEs' transmit power. The traffic arrivals are also i.i.d. among VUE pairs. Thus, we deduce that $Q_1(t), \cdots, Q_K(t)$ are i.i.d., and $M(t)$ converges to a GEV distributed RV as $K \to \infty$. Referring to the support of $M(t)$, we focus on VUE pair k's queue length conditioned on $1 + \xi(Q_k(t) - \mu)/\sigma \geq 0$. In other words, we consider the situation in which VUE pair k is likely to achieve the largest queue length in the network. Subsequently, imposing the constraints on the mean and second moment of the conditional queue length, i.e.,

$$\lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[Q_k(t)|\mathbb{1}_{\{1+\xi(Q_k(t)-\mu)/\sigma \geq 0\}}\big] \leq \bar{M}_{\text{th}}, \quad (6)$$

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[(Q_k(t))^2 | \mathbb{1}_{\{1+\xi(Q_k(t)-\mu)/\sigma\geq 0\}}\big] \leq \bar{B}_{\text{th}}, \quad (7)$$

each VUE pair k locally focuses on the power minimization problem which is modeled as follows:

$$\underset{P_k^n(t)}{\text{minimize}} \quad \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{n\in\mathcal{N}_k} P_k^n(t) \qquad (8)$$
$$\text{subject to} \quad \lim_{t\to\infty}\mathbb{E}[|Q_k(t)|] < \infty, (1), (6), \text{and } (7).$$

In (6) and (7), the VUE requires the parameters μ, σ, and ξ of the network-wide maximal queue length $M(t)$, which are unknown beforehand. To deal with this, we introduce the following Theorem and then specify a local and empirical estimation mechanism for these parameters.

Theorem 2 (**Pickands–Balkema–de Haan theorem** [3]). *Consider any RV Q_k of Theorem 1 and a high threshold d. As $d \to F_{Q_k}^{-1}(1)$, we can approximately characterize the excess value $S = Q_k - d > 0$ by a generalized Pareto distribution (GPD) with two parameters $\tilde{\sigma} = \frac{\mathbb{E}[S^2]\mathbb{E}[S]}{2\mathbb{E}[S^2]-2\mathbb{E}[S]^2} > 0$ and $\xi = \frac{\mathbb{E}[S^2]-2\mathbb{E}[S]^2}{2\mathbb{E}[S^2]-2\mathbb{E}[S]^2} \in \mathbb{R}.$*

In Theorems 1 and 2, ξ is identical while $\sigma = \tilde{\sigma} + \xi(\mu - d)$. From von Mises conditions [3], we can asymptotically find $\mu = \lim_{K\to\infty} F_{Q_k}^{-1}(1 - \frac{1}{K})$. Based on the above results, VUE pair k empirically estimates μ, σ, and ξ of (6) and (7) as per

$$\begin{cases} d_k(t) = \hat{F}_{Q_k}^{-1}(1-\psi), \\ c_k^{\text{m}}(t) = \frac{\sum_{\tau=1}^{t}(Q_k(\tau)-d_k(t))\cdot\mathbb{1}_{\{Q_k(\tau)-d_k(t)>0\}}}{\sum_{\tau=1}^{t}\mathbb{1}_{\{Q_k(\tau)-d_k(t)>0\}}}, \\ c_k^{\text{v}}(t) = \frac{\sum_{\tau=1}^{t}(Q_k(\tau)-d_k(t))^2\cdot\mathbb{1}_{\{Q_k(\tau)-d_k(t)>0\}}}{\sum_{\tau=1}^{t}\mathbb{1}_{\{Q_k(\tau)-d_k(t)>0\}}}, \\ \hat{\mu}_k(t) = \hat{F}_{Q_k}^{-1}(1-\frac{1}{K}), \quad \hat{\xi}_k(t) = \frac{c_k^{\text{v}}(t)-2[c_k^{\text{m}}(t)]^2}{2c_k^{\text{v}}(t)-2[c_k^{\text{m}}(t)]^2}, \\ \hat{\sigma}_k(t) = \frac{c_k^{\text{v}}(t)c_k^{\text{m}}(t)+(c_k^{\text{v}}(t)-2[c_k^{\text{m}}(t)]^2)(\hat{\mu}_k(t)-d_k(t))}{2c_k^{\text{v}}(t)-2[c_k^{\text{m}}(t)]^2}, \end{cases} \quad (9)$$

with $\psi \approx 0$, and $\hat{F}_{Q_k}$ is the empirically estimated cumulative distribution function (CDF) of Q_k. Analogously to Section III-A, we solve problem (8) using the Lyapunov optimization by introducing two virtual queues,

$$Q_k^{(M)}(t+1) = \max\big\{Q_k^{(M)}(t) + \big(Q_k(t+1) - \bar{M}_{\text{th}}\big)$$
$$\times \mathbb{1}_{\{1+\hat{\xi}_k(t)(Q_k(t+1)-\hat{\mu}_k(t))/\hat{\sigma}_k(t)\geq 0\}}, 0\big\}, \quad (10)$$

$$Q_k^{(B)}(t+1) = \max\big\{Q_k^{(B)}(t) + \big([Q_k(t+1)]^2 - \bar{B}_{\text{th}}\big)$$
$$\times \mathbb{1}_{\{1+\hat{\xi}_k(t)(Q_k(t+1)-\hat{\mu}_k(t))/\hat{\sigma}_k(t)\geq 0\}}, 0\big\}, \quad (11)$$

for constraints (6) and (7), respectively. VUE pair k then finds its transmit power by solving the optimization problem (5) with $J_k(t) = WT_{\text{c}}\big(Q_k(t) + \lambda_k(t)\big) + WT_{\text{c}}\big[Q_k^{(M)}(t) + \big(2Q_k^{(B)}(t) + 1\big)\big(Q_k(t) + \lambda_k(t)\big) + 2\big(Q_k(t) + \lambda_k(t)\big)^3\big] \cdot \mathbb{1}_{\{1+\hat{\xi}_k(t)(Q_k(t)+\lambda_k(t)-\hat{\mu}_k(t))/\hat{\sigma}_k(t)\geq 0\}}$ in each time slot t. After sending data, VUE pair k locally updates $Q_k(t+1)$, (9), (10), and (11). In the EVT-based solution, the VUE pair can locally estimate the statistics of the network-wide maximal queue length. In other words, the RSU is not needed to track the network-wide maximal queue length and exchange QSI for the VUEs. This mechanism remarkably alleviates signaling overhead for the high-mobility V2V communication.

Table I
SIMULATION PARAMETERS [1], [2], [7], [8]

Para.	Value	Para.	Value	Para.	Value
K	$\{20, 40, 60, 80\}$	W	180 kHz	T_{c}	3 ms
N_0	-174 dBm/Hz	$P_{\max}$	10 dBm	N	20
λ_{avg}	0.5 Mbps	ψ	10^{-2}	T_0	100
ζ	30 m	ϕ	150 m	g	10
M_{th}	225 kbit	l_0'	-54.5 dB	α	1.61
B_{th}	6×10^{10} bit^2	l_0	-68.5 dB	$\triangle$	15 m

IV. NUMERICAL RESULTS

We simulate a $250\times250\,\text{m}^2$-area Manhattan mobility model as in [2]. The average vehicle speed is $60\,\text{km/h}$, and the distance between the transmitter and receiver of each VUE pair is $15\,\text{m}$. Assuming the $5.9\,\text{GHz}$ carrier frequency and expressing $\mathbf{x} = (x_i, x_j) \in \mathbb{R}^2$ and $\mathbf{y} = (y_i, y_j) \in \mathbb{R}^2$ as the transmitter's and receiver's Euclidean coordinates, respectively, we consider the path loss model for the urban areas [1]. When the transmitter and receiver are on the same lane, we have the line-of-sight path loss value $l_0\|\mathbf{x} - \mathbf{y}\|^{-\alpha}$. Provided that the transmitter and receiver are separately located on the perpendicular lanes, we consider the weak-line-of-sight path loss model $l_0(|x_i - y_i| + |x_j - y_j|)^{-\alpha}$ if, at least, one is near the intersection within the distance $\triangle$. Otherwise, we have the non-line-of-sight path loss value $l_0'(|x_i - y_i|\cdot|x_j - y_j|)^{-\alpha}$ with $l_0' < l_0(\frac{\triangle}{2})^\alpha$. Finally, if the transmitter and receiver are not located on the same lane nor on the perpendicular lanes, we assume no signal propagation. Moreover, all wireless channels experience Rayleigh fading with unit variance, and Poisson traffic arrivals are considered. The remaining parameters are listed in Table I. For performance comparison, we consider a **baseline** in which the VUE transmits with a constant rate. From [9], we know that given a constant service rate R_{c}, the complementary cumulative distribution function (CCDF) of the queue length can be approximately written as $\bar{F}_Q(q) \approx \Pr(Q > 0)\cdot e^{-\theta q}$, where exponent θ can be found by equating the effective bandwidth function $\beta(\theta)$ [9] to the constant service rate, i.e., $\theta = \beta^{-1}(R_{\text{c}})$. Furthermore, applying $\bar{F}_Q(q)$ to Theorems 1 and 2, we obtain the corresponding GEV distribution, with $\mathbb{E}[M] \approx [\ln(K \cdot \Pr(Q > 0)) + 0.57721]/\theta$ and $\text{Var}(M) \approx \pi^2/(6\theta^2)$, of the baseline.

Let us first verify the accuracy of using EVT to characterize the network-wide maximal queue length M in the EVT-based scheme. Specifically, in Fig. 1, we plot the CCDFs of M obtained numerically in the EVT-based scheme as well as theoretically using Theorem 1. When $K = 20$, there is a gap since the number of VUE pairs is not sufficient to have a converged GEV approximation. However, when $K \geq 40$, numerical values match well with the theoretical approximation. Thus, even though the number of VUE pairs is moderate, EVT still provides a powerful framework to characterize the network-wide metric without resorting to $K \to \infty$. If there are more VUEs sharing resources, the incurred lower rate results in higher queue length. Next, we consider $K = 80$ in the following simulations. In Fig. 2, we show the throughput-latency (i.e., power-delay since throughput increases with transmit power) tradeoffs of our proposed queue-aware approaches and the baseline. At $V = 0$, the VUE aims to boost the transmission rate as per (5), yielding the highest
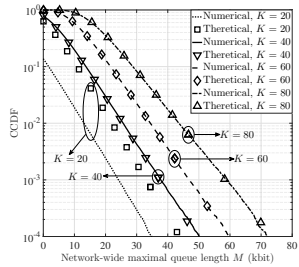
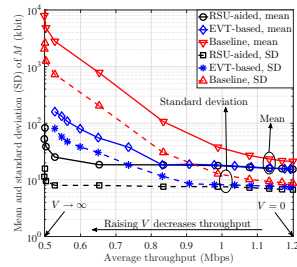Figure 1. Accuracy of the theoretical approximation using EVT, $V = 0$.

Figure 2. Tradeoff between the VUE' average throughput and the statistics of the network-wide maximal queue length.

Figure 3. Average throughput and queuing latency versus blocklength with 15 m VUE pair distance, $\lambda_{\mathrm{avg}} = 0.5$ Mbps.

Figure 4. Average throughput versus blocklength with 100 m VUE pair distance, $\lambda_{\mathrm{avg}} = 0.01$ Mbps.

average throughput with lowest maximal queue length. On the other hand, the optimal solutions to the power minimization problems (2) and (8) are asymptotically achieved by increasing V in (5). Since the average throughput to maintain system stability is minimized as $V \to \infty$ (via power minimization), the queue length increases dramatically. Additionally, given that the VUE can increase its transmit power with a tighter requirement on $\bar{M}$ and $\mathrm{Var}(M)$, the VUE can estimate the statistics of M locally and find the transmit power without global QSI exchange with the RSU. If the VUE has lower power budget, using the RSU for exchanging the global QSI helps to alleviate the maximal queue length albeit increasing signaling overhead. In contrast with the baseline, our two proposed approaches achieve performance enhancement since the former is oblivious to the queue value. At low average throughput whereby higher gains are attained, resource scheduling helps to deliver data efficiently. Subsequently, we consider the RSU-aided scheme with $V = 0$ owing to its highest throughput and lowest queue length performance. Note that due to the high mobility feature in V2V communication, the small time slot length T_c (i.e., coherence time) restricts the codeword length (or blocklength) in each transmission. This hinders vehicles from achieving the Shannon rate with an infinitesimal decoding error probability. Taking into account this practical concern in **finite blocklength** transmission, we consider the transmission rate $R_f = \log_2(1 + \gamma) - \frac{\sqrt{2\gamma(\gamma+2)}\mathrm{erfc}^{-1}(2\epsilon)}{\sqrt{L}(1+\gamma)\ln 2}$ which incorporates the blocklength $L \ll \infty$ and a block error probability $\epsilon > 0$ with the inverse error function $\mathrm{erfc}^{-1}(\cdot)$ [10]. Additionally, the performance of the system design in Section III can be generalized by letting $\epsilon = 0.5$. Based on R_f, we investigate the average throughput, denoted by $\bar{R}(L, \epsilon)$, and average queuing latency versus the blocklength for various block error probabilities in Figs. 3 and 4, where L is varied by changing the coherence time T_c (i.e., vehicle speed [8]). For a given L, decreasing the average throughput allows for more reliable communication, i.e., lower ϵ, as per R_f. On the other hand, lower throughput increases the queue length, resulting in longer average queuing latency. Next we vary L while fixing ϵ. Although decreasing L lowers the transmission rate, the average queuing latency can be further alleviated due to the smaller transmission time period T_c. At $\epsilon = 0.5$, $R_f = \log_2(1 + \gamma)$ is not explicitly affected by L. However, as L (or T_c) is increased, more traffic arrivals require higher power (i.e., higher throughput) whereas the average latency increases with L. As the blocklength increases, the average throughout curves converge to the capacity-achieving bound,
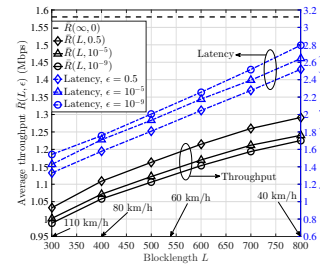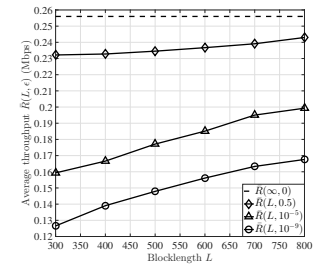
i.e., $L \to \infty$ (unbounded latency) and $\epsilon \to 0$. Furthermore, using the Shannon rate-based design in the finite blocklength transmission, i.e., $\epsilon = 0.5$, reliable communication is obtained at the expense of significant throughput loss in the low signal-to-noise ratio case (i.e., large VUE pair distance). Finally, Table II shows throughput ratios as a function of different VUE pair distances.

V. Conclusions

This letter has studied the problem of transmit power minimization subject to high-order constraints on the maximal queue length among all vehicles. We have proposed a semi-centralized and a distributed dynamic power allocation solutions by marrying tools from Lyapunov stochastic optimization and EVT. Simulation results have shown the effectiveness of extreme value theory in designing URLLC systems as well as the performance improvements of our proposed approaches.

TABLE II
THROUGHPUT RATIO IN THE FINITE BLOCKLENGTH TRANSMISSION

Distance	$\frac{\bar{R}(300,10^{-9})}{\bar{R}(300,0.5)}$	$\frac{\bar{R}(800,10^{-9})}{\bar{R}(800,0.5)}$	$\frac{\bar{R}(300,10^{-5})}{\bar{R}(300,0.5)}$	$\frac{\bar{R}(800,10^{-5})}{\bar{R}(800,0.5)}$
15 m	95%	95%	96%	96%
100 m	55%	69%	69%	83%

References

[1] M. Abdulla and H. Wymeersch, "Fine-grained vs. average reliability for V2V communications around intersections," in *Proc. IEEE Global Commun. Conf. Workshops*, Dec. 2017, pp. 1–5.

[2] M. I. Ashraf, C.-F. Liu, M. Bennis, and W. Saad, "Towards low-latency and ultra-reliable vehicle-to-vehicle communication," in *Proc. European Conf. Netw. Commun.*, Jun. 2017, pp. 1–5.

[3] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction*. Springer, 2006.

[4] U. von Luxburg, "A tutorial on spectral clustering," *Statistics Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *CoRR*, vol. abs/1801.01270, pp. 1–26, Jan. 2018.

[6] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan and Claypool Publishers, Jun. 2010.

[7] J. Kunisch and J. Pamp, "Wideband car-to-car radio channel measurements and model at 5.9 GHz," in *Proc. IEEE 68th Veh. Technol. Conf.*, Sep. 2008, pp. 1–5.

[8] Z. Pi and F. Khan, "System design and network architecture for a millimeter-wave mobile broadband (MMB) system," in *Proc. 34th IEEE Sarnoff Symp.*, May 2011, pp. 1–6.

[9] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[10] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.