Neyman-Pearson Criterion Driven NFV-SDN Architectures and Optimal Resource-Allocations for Statistical-QoS Based mURLLC Over Next-Generation Metaverse Mobile Networks Using FBC

Xi Zhang, *Fellow, IEEE*, Qixuan Zhu, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*— *Metaverse streaming*, as one of the key wireless services over 6G mobile networks, generates the delay/error-sensitive and bandwidth-intensive wireless traffics with stringent quality-of-service (QoS) requirements. Consequently, metaverse streaming can be modeled as a new type of *massive ultra-reliable low-latency communications* (mURLLC) traffic over 6G mobile networks. However, how to efficiently support metaverse streaming with constrained wireless resources and dynamic network conditions has imposed many new challenges not encountered before. To conquer these difficulties, in this paper we propose the *Neyman-Pearson criterion* driven network functions virtualization (NFV) and software-defined network (SDN) architectures and optimal resource-allocations for *statistical-QoS theory* based mURLLC streaming over 6G metaverse mobile networks using *finite blocklength coding* (FBC). First, we use Neyman-Pearson hypothesis tests for characterizing metaverse streaming requests' distribution profiles to predict their future accessing frequencies/patterns. Second, our formulated NFV/SDN architectures and virtual-network slices are assigned to the designated metaverse mobile users with the same predicted data request distributions, categories, and statistical-QoS requirements. Third, integrating the statistical QoS theory with FBC, we develop metaverse-streaming schemes by maximizing aggregate ϵ-*effective capacity* and deriving optimal transmit power allocations. Finally, we use numerical analyses to validate and evaluate our proposed schemes over 6G mobile networks.**

***Index Terms*— 6G, metaverse, Neyman-Pearson test, m-MIMO, ϵ-effective capacity, statistical delay/error-rate bounded QoS.**

I. Introduction

THE 6G wireless networks are envisioned to provide various advanced wireless services featuring massive access/connection, ultra-reliability, low latency, intelligence, and security while maximizing spectral and power efficiencies. Widely recognized as a new wave of wireless technologies, *metaverse streaming* provides *immersive environments*, including digital twins (DTs), in the virtual space generated by computers for communications. Metaverse streaming is considered to be a type of *massive ultra-reliable and low-latency communications* (mURLLC) service, which is expected to be the dominant traffic type in 6G networks supporting *massive numbers* of mobile users (MUs) demanding stringent *quality of service* (QoS) requirements with one-way radio latency less than 1 ms and successful-transmission probability higher than 99.99999% [1], [2].

The integration of widely envisioned 6G key wireless techniques including *massive multiple-input and multiple-output* (massive-MIMO) [3], [4], [5], statistical QoS theory [6], [7], [8], [9], [10], [11], finite blocklength coding (FBC) [12], etc., is expected to provide an efficient solution to implement mURLLC transmissions [3], [4], [13]. Using massive numbers of antennas on the base station (BS) or multiple antennas on the access point (AP), it is possible to direct the main beam of signal waves towards the targeted MUs, serve more users through the spatial multiplexing, and mitigate the multipath effect via the spatial diversity provided by different MIMO antennas. The conventional Shannon theory based analysis is usually not applicable for simultaneously guaranteeing *both* the *low-latency* and the *high-reliability* of wireless transmissions, because the traditional Shannon formalism is based on the assumption that the coding block length tends to infinity to achieve arbitrarily-high or even perfect reliability. Towards this end, the *statistical QoS theory* has been proposed to support both statistical delay and error-rate bounded QoS provisioning for wireless transmissions over time-varying wireless fading channels. Furthermore, the FBC technique has been developed to enable *small packet communications* for adaptive error-control and real-time transmissions, where senders encode their messages into short packets (i.e., packets with small numbers of bits) to reduce the transmission latency while constraining and controlling the decoding error probability.

There have been various studies conducted on metaverse streaming transmissions and techniques. The authors of [14] proposed a distributed collaborative computing framework for vehicular metaverse streaming by employing coded

Manuscript received 15 March 2023; revised 1 August 2023; accepted 31 August 2023. Date of publication 25 December 2023; date of current version 1 March 2024. The work of Xi Zhang and Qixuan Zhu was supported in part by the U.S. National Science Foundation under Grant CCF-2142890, Grant CCF-2008975, Grant ECCS-1408601, and Grant CNS-1205726; and in part by the U.S. Air Force under Grant FA9453-15-C-0423. The work of H. Vincent Poor was supported in part by the U.S. National Science Foundation under Grant CNS-2128448 and Grant ECCS-2335876. *(Corresponding author: Xi Zhang.)*

Xi Zhang and Qixuan Zhu are with the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xizhang@ece.tamu.edu; qixuan@tamu.edu).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2023.3345428.

Digital Object Identifier 10.1109/JSAC.2023.3345428

distributed computing and blockchain schemes. The work of [15] proposed a novel digital twin scheme to support metaverse communications by jointly integrating communications, computations, and storage techniques through the applications of mobile edge computing and mURLLC. To study the interaction between the metaverse service provider and the network infrastructure provider, the work of [16] designed an optimal framework to maximize the quality of experience for metaverse users. An edge computing-assisted metaverse system is designed by [17] to partially offload the sensing data collected from physical objects to the edge computing platform, ensuring the promptness of metaverse services and satisfying latency requirements of metaverse users. However, in addition to the stringent QoS provisioning for general mURLLC streaming, metaverse streaming also need to take into account humans' activities, e.g, data items' accessing/requesting, frequencies/patterns, heterogeneous QoS requirements, and MU's mobility, etc., which impose a number of new challenges. One of the main challenges is how to predict humans' behaviors in terms of data-access frequencies and patterns so that we can optimize network architectures and resource allocations [18] to best fit and improve the statistical QoS in human-centric metaverse mobile wireless networks.

To efficiently integrate the above described 6G techniques and architectures and optimize the wireless resource-allocation for supporting human-centric metaverse streaming, in this paper we propose to develop the Neyman-Pearson criterion driven network functions virtualization (NFV) and software-defined network (SDN) [19], [20] architectures and optimal resource-allocations for statistical-QoS based mURLLC-streaming over 6G human-centric metaverse mobile networks using FBC. We apply the Neyman-Pearson criterion based *sequential hypothesis testing technique* [21], [22], [23] to *predict humans' activities*, in terms of data-content items' requests probability distribution profile. The sequential hypothesis testing technique has been broadly applied in a large number of detection applications because it provides efficient rules to make a decision by accepting one of multiple hypotheses at any stage in a sequence of observations. In the setting of this paper, the hypothesis represents the metaverse mobile networks' claim/statement on the distribution for a sequence of stochastic observations regarding that an MU requests/accesses the designated metaverse data item. Here, in order to apply sequential testing, we observe a set of data items' requests and calculate the likelihood ratio of these data requests. We accept one of a set of appropriate hypotheses accordingly or take another observation, depending on the threshold regions of the likelihood ratio. We repeat these procedures until a hypothesis is selected. In this paper, we model the probability distributions of MUs' requests for metaverse data items as a series of Zipf distributions, and each hypothesis is characterized by an individual Zipf distribution with a specific parameter.

In particular, first our proposed schemes conduct a Neyman-Pearson criterion based sequential hypothesis test, the outcome of which is the prediction of an MU's requesting probability distribution for metaverse streaming. Using the obtained prediction of metaverse request distributions, we formulate and select the NFV and SDN architectures to implement the metaverse streaming while satisfying data items' statistical delay and error-rate bounded QoS requirements, through adaptively designing the optimal network architectures and allocating the necessary wireless resources. Second, our schemes map MUs that are estimated to have the same data request probability distribution into one virtual network slice, since each data request probability distribution represents a specific category of metaverse streaming (e.g., virtual-reality (VR) online gaming, e-health care, etc.) and metaverse data in the same category share the same delay and error-rate bounded QoS requirements. Finally, combining the statistical QoS theory and FBC, we derive a number of closed-form expressions to accurately model and analyze our newly defined metaverse streaming performance metrics and controlling functions, including the optimal transmit power allocation policies and the corresponding maximum ϵ-effective capacity functions, etc.

The rest of this paper is organized as follows. Section II establishes systems models for our proposed Neyman-Pearson criterion hypothesis testing driven NFV/SDN architectures for statistical-QoS based mURLLC-streaming over 6G human-centric metaverse mobile networks using FBC. Section III develops the decision making schemes to derive the optimal sequential hypothesis testing. Section IV creates the modeling framework and a set of performance metrics to characterize and analyze our proposed metaverse streaming schemes by deriving the aggregate ϵ-effective capacity and the optimal transmit power allocations to maximize the aggregate ϵ-effective capacity. Section V validates and evaluates our developed schemes and our derived analytical results for supporting the statistical delay and error-rate bounded QoS based metaverse streaming in the non-asymptotic regime. This paper concludes with Section VI.

II. The System Models for Our Proposed Metaverse-Streaming Schemes

A. Neyman-Pearson Criterion Driven NFV/SDN Architectures Over 6G Metaverse Wireless Networks

Figure 1 shows our proposed Neyman-Pearson criterion driven NFV/SDN architectural system models to support metaverse streaming with mURLLC traffic requirements over 6G mobile networks. Since the 6G metaverse streaming includes different types of traffic, such as VR-based online gaming, high-resolution video streaming, digital twins, e-health-care, conference, and education, etc., MUs that request different categories of metaverse streaming demand diverse statistical QoS requirements. In Fig. 1, we map each category of metaverse streaming into one virtual network slice, enabling the sharing of all wireless network functionalities among multiple metaverse service providers and transmitting different metaverse services under diverse QoS requirements and logical architectures through the same infrastructure. The SDN-based computing algorithm dynamically allocates wireless resources (i.e., transmit power, sub-channels, etc.) to different network slices, maximizing the overall networks performance metrics. Physical devices function as simple packet forwarding devices (*data plane*). The intelligent control logic functions

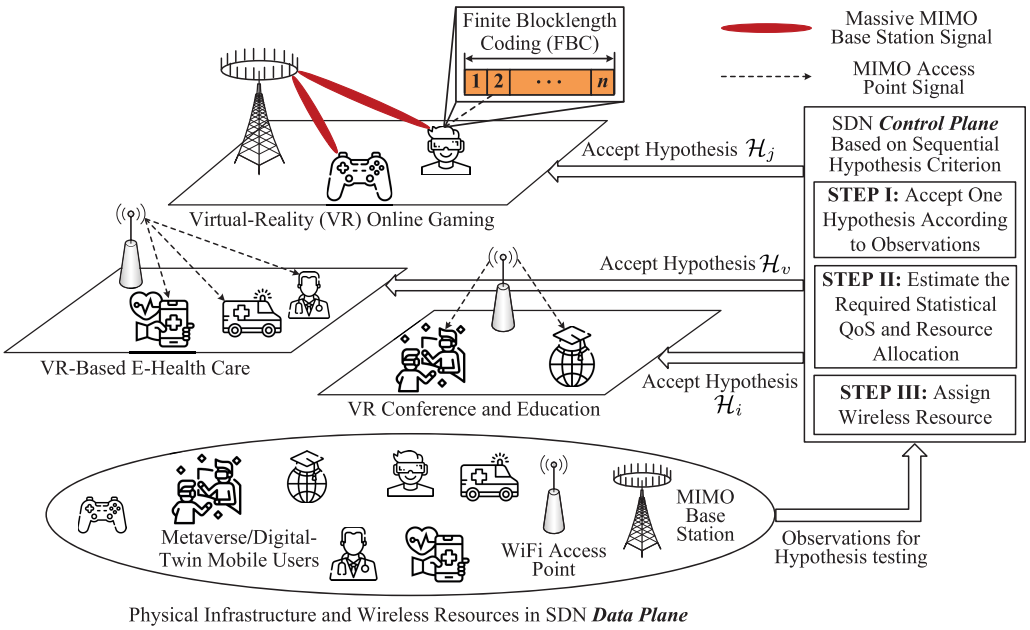


Fig. 1. The system models of our proposed Neyman-Pearson hypothesis testing driven network-functions virtualization and software-defined architectures and optimal resource allocations for statistical-QoS based mURLLC-streaming over 6G human-centric metaverse (virtual-reality) mobile networks using FBC, where hypotheses $\mathcal{H}_j$, $\mathcal{H}_i$, and $\mathcal{H}_v$, with $i, j, v \in \mathcal{J}$ and $i \neq j \neq v$, represent three different metaverse data request probability-distributions profiles.

are performed by the *control plane* that assigns wireless resources based on the MU's metaverse service types and their corresponding statistical QoS requirements. As shown in Fig. 1, each metaverse MU communicates (down-stream/up-stream) with the massive-MIMO BS or MIMO AP by using the FBC technique with a finite blocklength equal to n.

B. Our Proposed Neyman-Pearson Criterion Based Sequential Hypothesis Test for Metaverse Streaming

Assume that we take an observation for an MU's requested metaverse data item at each time slot. Let q be the index of the time slot/stage, where $q \in \{1, 2, 3, \cdots\}$. Define $\{X_q\}$ as a sequence of random variables taking values on a set of all metaverse data-content items $\mathcal{D} = \{1, 2, \ldots, D\}$ with $|\mathcal{D}| = D$ denoting the total number of different metaverse data content items, where X_q is the observation random variable for the requested data item by an MU at the time slot q. Assume that elements in $\{X_q\}, \forall q$, are independent and identically distributed (i.i.d.) random variables. We consider multiple hypotheses $\{\mathcal{H}_j\}, \forall j \in \mathcal{J} = \{0, 1, 2, \ldots, J\}$. Hypothesis $\mathcal{H}_j$ implies that X_q follows the probability distribution P_j (i.e., metaverse data request probability distribution at the time slot q is P_j), which is a probability measure as follows:

$$\mathcal{H}_j : X_q \sim P_j = \text{Zipf}(r_j, D), \quad \forall\, r_j \in (\xi_j, \xi_{j+1}], \tag{1}$$

where $\text{Zipf}(r_j, D)$, $\forall j$, denotes the Zipf distribution with the exponent r_j; and we also define $\xi_0 = 0$; $\xi_j > 0, \forall j \geq 1$; and $\xi_{j+1} > \xi_j$. The probability mass function (pmf) for the dth data content item, $\forall d \in \{1, 2, \ldots, D\}$, of a Zipf distribution is given by

$$f_{r_j}(d) = \frac{d^{-r_j}}{\sum_{k=1}^{D} k^{-r_j}}. \tag{2}$$

Assume $f_{r_j}(d) \neq f_{r_i}(d)$ if $j \neq i$. Let t be the total number of time slots for observing an MU's data requests. Define $z_{j,i}^{(t)}$ as the likelihood ratio of the hypothesis $\mathcal{H}_j, \forall j \in \mathcal{J}$, to the hypothesis $\mathcal{H}_i, \forall i \in \mathcal{J}$, by the time slot t, which is given as follows [24]:

$$z_{j,i}^{(t)} \triangleq \prod_{q=1}^{t} \frac{f_{r_j}(X_q)}{f_{r_i}(X_q)} \tag{3}$$

where $X_q \in \mathcal{D}$. According to [25], we also define the log likelihood ratio of the conditional density functions, denoted by $Z_{j,i}(q)$, as follows:

$$Z_{j,i}(q) \triangleq \log\left(\frac{f_{r_j}(X_q|X_1, X_2, \ldots, X_{q-1})}{f_{r_i}(X_q|X_1, X_2, \ldots, X_{q-1})}\right). \tag{4}$$

Based on Eq. (3) and Eq. (4), we define the sequential hypothesis test [25], [26] by two sequences of decision functions $(\psi_t)_{t\geq 0}$ and $(\delta_t)_{t\geq 0}$, where ψ_t: $\mathcal{D}^t \mapsto T$ is the stopping rule, denoted by ψ_t, that maps the current observations to a decision that stops the testing at time slot/stage T; and δ_t: $\mathcal{D}^t \mapsto \{0, 1, \ldots, J\}$, where $\delta_t = j$ corresponds to a decision for accepting $\mathcal{H}_j$ upon stopping. For the rest of this paper, we formally denote our proposed sequential hypothesis test by the decision-functions pair: (ψ_t, δ_t).

We propose to estimate an MU's future metaverse data requests, in terms of a Zipf distribution with a specific parameter r_j, by using the Neyman-Pearson criterion [27] based sequential hypothesis testing according to the observations for this MU's past data requests. We define two types of errors: ***Type I error*** is defined as "Reject $\mathcal{H}_j$ (i.e., accept $\mathcal{H}_i, i \neq j$) when $\mathcal{H}_j$ is the actual MU's data request distribution" and ***Type II error*** is defined as "Accept $\mathcal{H}_j$ when $\mathcal{H}_i, i \neq j$, is the actual MU's data request distribution". Type I error can

be considered as the false alarm and Type II error can be considered as the miss detection. We denote the probabilities of Type I and Type II errors as $P^{\mathrm{F}}(j,i)$ and $P^{\mathrm{M}}(j,i)$, respectively. Then, the hypothesis hitting probability, denoted by $P^{\mathrm{H}}(j,j)$, can be written as $P^{\mathrm{H}}(j,j) = 1 - \sum_i P^{\mathrm{F}}(j,i)$. For a decision rule δ_t, we define the error probability $P^{\mathrm{F}}(j,i)$ as follows:

$$P^{\mathrm{F}}(j,i) = P_j\{\delta_t = i\}, \quad \text{if} \quad j \neq i \tag{5}$$

where $P_j\{\cdot\}$ represents the probability of an event under the hypothesis $\mathcal{H}_j$ being true, and define the hypothesis hitting probability $P^{\mathrm{H}}(j,j)$ as follows:

$$P^{\mathrm{H}}(j,j) = P_j\{\delta_t = j\}, \quad \forall j. \tag{6}$$

We apply Neyman-Pearson criterion based hypothesis testing scheme to maximize the hypothesis hitting probability subject to constraints of upper-bounding the Type I and Type II error probabilities as follows:

$$\begin{aligned} \max_{(\psi_t,\delta_t)} \quad & P^{\mathrm{H}}(j,j) \\ \textbf{s.t.: C1:} \quad & \mathbb{E}_{i,j}\big[P^{\mathrm{F}}(j,i)\big] \leq \alpha; \\ \textbf{C2:} \quad & \mathbb{E}_{i,j}\big[P^{\mathrm{M}}(j,i)\big] \leq \beta, \end{aligned} \tag{7}$$

where $\mathbb{E}_{i,j}[\cdot]$ denotes the expectation for all i and j, $\forall i,j \in \mathcal{J}$; and α and β are the upper bounds on the Type I and Type II error probabilities, respectively.

C. Integrating Effective Capacity With FBC to Support Metaverse Streaming

The statistical delay-bounded QoS guarantees [6], [7], [9], [10], [28], [29] have been shown to be powerful in analyzing queuing behavior for the stochastic arrival and service processes over the time-varying wireless fading channels. The key statistical-QoS performance metric is the *effective capacity* which measures the maximum packet's constant arrival rate such that the given statistical *delay-bounded* QoS can be guaranteed. Based on the large deviation principle (LDP) [7], under sufficient conditions, the queue-length process $Q_d(t)$ for the metaverse data-content item d converges in distribution to a random variable $Q_d(\infty)$ such that

$$-\lim_{Q_{\mathrm{th},d}\to\infty} \frac{\log(\Pr\{Q_d(\infty) > Q_{\mathrm{th},d}\})}{Q_{\mathrm{th},d}} = \theta_d \tag{8}$$

where $Q_{\mathrm{th},d}$ is the queue length threshold (bound) and $\theta_d > 0$ is defined as the *QoS exponent* for the data item d. The insights of Eq. (8) reveal that the probability of the queueing process exceeding a certain threshold $Q_{\mathrm{th},d}$ decays exponentially fast at the rate of θ_d as the threshold $Q_{\mathrm{th},d}$ increases and tends to infinity. A smaller θ_d corresponds to a slower decay rate, which implies that the system can only provide a looser QoS guarantee, while a larger θ_d leads to a faster decay rate, which implies that a more stringent QoS can be supported. In particular, when $\theta_d \to 0$, the system can tolerate a long delay; when $\theta_d \to \infty$, the system cannot tolerate any delay.

However, the conventional statistical-QoS theory modeled by Eq. (8) focuses only on the statistical delay-bounded QoS without considering the transmission reliability, which is thus not feasible to support mURLLC in our proposed metaverse wireless networks. To support the stringent both statistical *delay* and *error-rate* bounded QoS provisioning for mURLLC, we propose to integrate the effective capacity theory with the *FBC scheme*, which is an emerging and powerful solution in wireless networks, to encode the message. Using the FBC scheme, terminals send messages using packets with small numbers of bits to achieve low latency transmissions while mitigating the packet's decoding error probability for reliable transmissions. We define an FBC scheme in the following definition.

Definition 1: Consider a fading channel which uses input blockcode set $\mathcal{A}$ and output blockcode set $\mathcal{B}$. We define that an (n, W, ϵ)-code, for a memoryless channel consists of [12]

- A message set $\mathcal{W} = \{c_1, \cdots, c_W\}$ with the cardinality W and the message length equal to $\log_2 W$.
- An encoder is a function: $\mathcal{W} \mapsto \mathcal{A}^n$, where $\mathcal{A}^n$ is the set of codewords with length n. At the receiver end, a decoder produces an estimate of the original message by observing the channel output, according to a function: $\mathcal{B}^n \mapsto \widehat{\mathcal{W}}$, where $\mathcal{B}^n$ is the set of received codewords with length n and $\widehat{\mathcal{W}}$ is the estimation of $\mathcal{W}$.
- The decoding error probability, denoted by ϵ, is defined as $\epsilon \triangleq (1/W)\sum_{w=1}^{W} \Pr\{c_w \neq \widehat{c}_w\}$, with $c_w \in \mathcal{W}$, $\widehat{c}_w \in \widehat{\mathcal{W}}$.

where usually $\epsilon > 0$ if $n < \infty$. ■

Thus, the triple-tuple (n, W, ϵ) represents that a source with the cardinality W can successfully transmit messages with a probability of success $(1-\epsilon)$ over n channel uses.

To integrate the effective capacity with the FBC scheme, we propose to employ the *ϵ-effective capacity*, which measures the maximum packet's arrival rate that a wireless channel can support under a given QoS exponent and a given decoding error-rate. Let k be the index of an MU. Let ϵ_d and $\mathcal{P}_{d,k,j}$ be the decoding error probability requirement of the data item d and the transmit power allocation for transmitting the data item d to the kth MU if accepting the hypothesis $\mathcal{H}_j$, respectively. Denote by $EC_k(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ the ϵ-effective capacity for the kth MU, which characterizes both statistical delay and error-rate bounded QoS provisionings under the power allocation $\mathcal{P}_{d,k,j}$. We define the ϵ-effective capacity $EC_k(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ as follows [13, Definition 5]:

Definition 2: Using the (n, W, ϵ)-code for the metaverse data-content item d, the ϵ-effective capacity $EC_k(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ for the kth MU is defined as the maximum packet's constant arrival rate for a given service process considering the delay QoS exponent θ_d and the non-vanishing decoding error probability ϵ_d, subject to statistical *delay* and *error-rate* bounded QoS constraints, respectively, which is given as follows:

$$\begin{aligned} &EC_k(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}) \\ &\triangleq -\frac{1}{n\theta_d} \log\left\{\epsilon_d + (1-\epsilon_d)\,\mathbb{E}_{\gamma_k}\Big[e^{-\theta_d n R(\gamma_k(\mathcal{P}_{d,k,j}))}\Big]\right\} \end{aligned} \tag{9}$$

where $\gamma_k(\mathcal{P}_{d,k,j})$ is the signal-to-noise ratio (SNR) of the kth MU under the transmit power allocation $\mathcal{P}_{d,k,j}$, $\mathbb{E}_{\gamma_k}[\cdot]$ denotes the expectation with respect to the SNR $\gamma_k(\mathcal{P}_{d,k,j})$, and $R\left(\gamma_k(\mathcal{P}_{d,k,j})\right)$ is the data rate (bits/sec/Hz) under the SNR $\gamma_k(\mathcal{P}_{d,k,j})$. ■

D. Hypothesis Testing Based NFV/SDN Network Architectures for Metaverse Streaming

We propose an optimal hypothesis testing to solve Eq. (7) and accept one hypothesis $\mathcal{H}_j, \forall j \in \mathcal{J}$, as the estimated MU's data request pmf profile for its metaverse streaming. Then, using the hypothesis testing outcome, we propose to map all MUs, which have the same metaverse data requests pmf profile Zipf(r_j, D), into the same virtual network slice and let them exchange metaverse data with other MUs in this virtual network slice. This is because MUs having the same metaverse data requests pmf profile are predicted to request the same metaverse data-content items with the same probability and uniform statistical QoS requirements, including the QoS exponent θ_d and decoding error probability ϵ_d, for each metaverse data-content item d. Then, we develop **Algorithm 1** given in TABLE I which shows our proposed hypothesis testing based NFV/SDN network architectures and resources allocation schemes.

III. Optimal Decision and Bounds of Sequential Hypothesis Testing

A. Optimal Decision of Sequential Hypothesis Testing

In order to determine the optimal sequential hypothesis testing, our proposed decision schemes consist of the following two steps.

Step 1: *Selecting the Optimal Subset of Hypotheses*

Around ξ_j, there exists the interval $(\underline{\xi}_j, \overline{\xi}_j)$, where $\underline{\xi}_j \triangleq \xi_j - \sigma/2$ and $\overline{\xi}_j \triangleq \xi_j + \sigma/2$, for $\sigma > 0$ [30], and $\sigma = \overline{\xi}_j - \underline{\xi}_j$ is a constant, $\forall j$. Let $\mathcal{R}_j$ be the sequential hypothesis test for testing the hypothesis $\mathcal{H}_{\underline{\xi}_j}$: "accepting $\underline{\xi}_j$" against $\mathcal{H}_{\overline{\xi}_j}$: "accepting $\overline{\xi}_j$". Similar to the two types of errors in Section II-B, we define another two types of errors: *Type I error* is defined as "Reject $\underline{\xi}_j$ when $\underline{\xi}_j$ is true", whose probability is upper bounded by α_j, and *Type II error* is defined as "Accept $\underline{\xi}_j$ when $\overline{\xi}_j$ is true", whose probability is upper bounded by β_j. According to [30] and [31], we select two positive constants A_j and B_j and set the values of A_j and B_j, respectively, as

$$A_j \propto \frac{1-\beta_j}{\alpha_j}, \quad B_j \propto \frac{\beta_j}{1-\alpha_j}, \text{ and } B_j < A_j \tag{10}$$

where $\varphi(x) \propto g(x)$ implies $\lim_{x\to\infty} \varphi(x)/g(x) = 1$. Define the likelihood ratio of $\mathcal{R}_j$ by time t as $z_{\underline{\xi}_j}^{(t)}$, which is given by

$$z_{\underline{\xi}_j}^{(t)} \triangleq \prod_{q=1}^{t} \frac{f_{\overline{\xi}_j}(X_q)}{f_{\underline{\xi}_j}(X_q)} = \underbrace{\left(\frac{\sum_{d=1}^{D} d^{-(\underline{\xi}_j+\sigma)}}{\sum_{d=1}^{D} d^{-\underline{\xi}_j}}\right)^t}_{g(\underline{\xi}_j)} \left(\prod_{q=1}^{t} X_q\right)^{\sigma}. \tag{11}$$

We accept $\mathcal{H}_{\overline{\xi}_j}$ if $z_{\underline{\xi}_j}^{(t)} > A_j$, accept $\mathcal{H}_{\underline{\xi}_j}$ if $z_{\underline{\xi}_j}^{(t)} < B_j$, and take an additional observation if $A_j < z_{\underline{\xi}_j}^{(t)} < B_j$. The decision procedure of this multiple sequential hypothesis testing is: (1) All $\mathcal{R}_j$'s are testing simultaneously at each stage until each $z_{\underline{\xi}_j}^{(t)}$ leads to the hypothesis testing $\mathcal{R}_j$ stopping. (2) Based on the result of each $\mathcal{R}_j$, we select an optimal subset of hypotheses from multiple hypotheses $\{\mathcal{H}_j\}, \forall j \in \mathcal{J}$.

TABLE I

Algorithm 1 Neyman-Pearson Criterion Driven NFV-SDN Architectures and Optimal Resource-Allocations for Statistical-QoS Based mURLLC Over 6G Metaverse Mobile Networks Using FBC

Input: The BS/AP set and MU set; all hypotheses $\{\mathcal{H}_j\}, \forall j$; each metaverse streaming's delay-bounded QoS exponent θ_d and error probability $\epsilon_d, \forall d$; and each MU's channel fading $h_k, \forall k$.
for each BS/AP **do**
  Assign each BS/AP to a virtual network slice that supports one type of metaverse streaming.
end for
for each MU **do**
  Observe its requested data content items and accept an optimal hypothesis $\mathcal{H}_j$, i.e., the pmf $f_{r_j}(d)$.
  According to the accepted hypothesis $\mathcal{H}_j$, assign the MU to the corresponding virtual network slice that supports this metaverse streaming.
end for
for Each MU in each network slice **do**
  Derive the optimal wireless resources (i.e., transmit power) for each MU using $f_{r_j}(d)$ to maximize the average aggregate ϵ-*effective capacity* over the entire network slice based on the required delay QoS exponent θ_d and error probability ϵ_d for the metaverse data and MU's wireless channel fading h_k.
end for
Output: optimal wireless network slicing, optimal MU's mapping and wireless resource allocations, and maximum average aggregate ϵ-*effective capacity* of each virtual network slice.

The optimal subset selection needs to consider the following **Case 1** and **Case 2**, respectively.

Case 1. All A_j's are equal to each other, i.e., $A_j = A, \forall j$, and all B_j's are equal to each other, i.e., $B_j = B, \forall j$. For this cases, if $z_{\underline{\xi}_j}^{(t)} > A$, i.e., accepting $\overline{\xi}_j$, we must reject all $\underline{\xi}_{\widetilde{j}}$, where $\widetilde{j} > j$; if $z_{\underline{\xi}_j}^{(t)} < B$, i.e., accepting $\underline{\xi}_j$, we must reject all $\overline{\xi}_{\widetilde{j}}$'s, where $\widetilde{j} < j$. This is because of the following derivations. When $z_{\underline{\xi}_j}^{(t)} > A$, we have $g(\underline{\xi}_j) > A\left(\prod_{q=1}^{t} X_q\right)^{-\sigma}$. Observing that $g(\underline{\xi}_j)$ in Eq. (11) is an increasing function of $\underline{\xi}_j$, we must have $g(\underline{\xi}_{\widetilde{j}}) > g(\underline{\xi}_j)$ when $\widetilde{j} > j$, and thus, $g(\underline{\xi}_{\widetilde{j}}) > A\left(\prod_{q=1}^{t} X_q\right)^{-\sigma}$. Therefore, $g(\underline{\xi}_{\widetilde{j}}) < B\left(\prod_{q=1}^{t} X_q\right)^{-\sigma}$ (i.e., accepting $\underline{\xi}_{\widetilde{j}}$) must not hold, due to $B < A$ as shown in Eq. (10). Using the similar derivation, we obtain that if $z_{\underline{\xi}_j}^{(t)} < B$, i.e., accepting $\underline{\xi}_j$, we must reject all $\overline{\xi}_{\widetilde{j}}$'s, where $\widetilde{j} < j$.

At one stage, if there exist multiple j's which result in $z_{\underline{\xi}_j}^{(t)} > A$, we make the decision to accept $\overline{\xi}_j$ such that $\overline{\xi}_j$ is the smallest number satisfying $z_{\underline{\xi}_j}^{(t)} > j$. Similarly, at another stage if there exist multiple j's which result in $z_{\underline{\xi}_j}^{(\widetilde{t})} < B$,

we make the decision to accept $\underline{\xi}_j$ such that $\underline{\xi}_j$ is the largest number satisfying $z_{\underline{\xi}_j}^{(\widetilde{t})} < B$, where $\widetilde{t} \neq t$. We continue with this procedure until all tests stop.

Case 2. $A_j \neq A_{\widetilde{j}}$ and $B_j \neq B_{\widetilde{j}}$ if $j \neq \widetilde{j}$. To satisfy both $z_{\underline{\xi}_j}^{(t)} > A_j$ and $z_{\underline{\xi}_{\widetilde{j}}}^{(t)} > A_{\widetilde{j}}$, let

$$A_j[g(\underline{\xi}_j)]^{-1} \geq A_{\widetilde{j}}[g(\underline{\xi}_{\widetilde{j}})]^{-1} \quad \text{if} \quad j > \widetilde{j}. \tag{12}$$

Since $g(\underline{\xi}_j) > g(\underline{\xi}_{\widetilde{j}})$ for $j > \widetilde{j}$, we must design the sequential hypothesis testing such that $A_j > A_{\widetilde{j}}$ to let Eq. (12) hold. Similarly, we design the testing such that $B_j > B_{\widetilde{j}}$ for $j > \widetilde{j}$.

Based on the conditions in **Case 1** and **Case 2**, if we happen to accept two hypotheses $\mathcal{H}_{\underline{\xi}_j}$ and $\mathcal{H}_{\underline{\xi}_{j+1}}$, then we will accept the hypothesis $\mathcal{H}_j$ as the optimal distribution estimation for the future data request probability profile; otherwise, we obtain a set of accepted hypotheses, denoted by $\{\mathcal{H}_s\}, \forall s \in \mathcal{S}$, which is the optimal subset of $\{\mathcal{H}_j\}$ and $\mathcal{S} \subset \mathcal{J}$.

Step 2: *Deriving the Optimal Hypothesis from the Subset*

We further derive the optimal hypothesis from the above obtained subset of hypotheses. The optimizing problem in Eq. (7) is converted to the equivalent problem of minimizing the cost of stopping a sequential procedure at the time slot t, denoted by $C_t(X_1, X_2, \ldots, X_t)$, based on the observations $(X_1, X_2, \ldots, X_t)$, as follows:

$$\begin{aligned} \min_{(\psi_t, \delta_t)} \quad & \mathbb{E}_{P_j}[C_t(X_1, X_2, \ldots, X_t)] \\ \textbf{s.t.: C1:} \quad & \mathbb{E}_{i,j}\big[P^{\mathrm{F}}(j,i)\big] \leq \alpha; \\ \textbf{C2:} \quad & \mathbb{E}_{i,j}\big[P^{\mathrm{M}}(j,i)\big] \leq \beta, \end{aligned} \tag{13}$$

where $\mathbb{E}_{P_j}[\cdot]$ denotes the expectation under the condition that $\mathcal{H}_j$ is the actual data request probability distribution.

B. Exponentially Bounded Stopping Time of Zipf Sequential Hypotheses Testing

Suppose that the sequential hypothesis testing stops at stage T, and the test stops when T is the smallest number of t such that $l^{\text{low}} < l_t(j,i) < l^{\text{up}}$ is violated, where $l_t(j,i) \triangleq \log\left(z_{j,i}^{(t)}\right)$ with $z_{j,i}^{(t)}$ defined by Eq. (3), and $l^{\text{low}} \triangleq \log(B_j)$ and $l^{\text{up}} \triangleq \log(A_j)$ are stopping bounds of $l_t(j,i)$, $j \neq i$, with A_j and B_j defined by Eq. (10). Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables. Define $Y_q \triangleq y(X_q), q = 1, 2, \ldots$, as a function $y(\cdot)$ that maps the range of X_q into an Euclidean k-space, $k \geq 1$, and define $\overline{Y}_t \triangleq \frac{1}{t}\sum_{q=1}^{t} Y_q$. Then, we have the following Theorem 1.

Theorem 1: **If** a sequential hypothesis test is testing the Zipf exponent under a Zipf distribution, **then** the following three claims hold.

Claim 1. $\mathbb{E}[Y_1] = \varepsilon$ exists and is finite.

Claim 2. There exist, respectively, a neighborhood $\mathcal{N}$ of ε, a real-valued continuous function $\Phi_{j,i}(\cdot)$ on $\mathcal{N}$, and a finite bound constant $I_{j,i} > 0$ such that if $\overline{Y}_t \in \mathcal{N}$, $t = 1, 2, \ldots$, then

$$\left|l_t(j,i) - t\Phi_{j,i}(\overline{Y}_t)\right| < I_{j,i} \tag{14}$$

where $|\cdot|$ is the absolute value, $I_{j,i}$ given by Eq. (14) is the Kullback-Leibler divergence measurement between distributions P_j and P_i which is specified as follows:

$$I_{j,i} = \left|\mathbb{E}_{P_j}\left[\log\frac{f_{r_j}(X_1)}{f_{r_i}(X_1)}\right]\right|, \tag{15}$$

and

$$\begin{aligned} l_t(j,i) \triangleq \log\left(z_{j,i}^{(t)}\right) &= \log\left[\prod_{q=1}^{t}\frac{f_{r_j}(X_q)}{f_{r_i}(X_q)}\right] \\ &= \log\left[\prod_{q=1}^{t}\left(X_q^{r_i - r_j}\frac{\sum_{d=1}^{D} d^{-r_i}}{\sum_{d=1}^{D} d^{-r_j}}\right)\right] \\ &= (r_i - r_j)\sum_{q=1}^{t}\log X_q + tM_{j,i} \end{aligned} \tag{16}$$

where $z_{j,i}^{(t)}$ is given by Eq. (3) and

$$M_{j,i} \triangleq \log\left(\frac{\sum_{d=1}^{D} d^{-r_i}}{\sum_{d=1}^{D} d^{-r_j}}\right). \tag{17}$$

Claim 3. $\Phi_{j,i}(\varepsilon) \neq 0$ or the first-order derivative of function $\Phi_{j,i}(\cdot)$ satisfies the following equation:

$$\Pr\left\{\frac{\partial\Phi_{j,i}(Y_1 - \varepsilon)}{\partial X_1} = 0\right\} < 1. \tag{18}$$

Proof: The proof is provided in Appendix A. ■

Remarks on Theorem 1: Theorem 1 ensures that our proposed Neyman-Pearson based sequential hypotheses test will eventually stop and accept a hypothesis as the optimal data request pmf profile. In addition, Theorem 1 shows that $I_{j,i} = \left|\mathbb{E}_{P_j}\left[\log\left\{f_{r_j}(X_1)/f_{r_i}(X_1)\right\}\right]\right|$, which is the Kullback–Leibler divergence measurement.

Theorem 2: If a sequential hypothesis test is testing the Zipf exponent under a Zipf distribution, then the stopping time of this sequential hypothesis test is *exponentially bounded* [23], namely, for some $c < \infty$ and $0 < \rho < 1$, the following two equations hold:

$$\begin{cases} \Pr\{T < \infty\} = 1, & (19) \\ \Pr\{T > t\} < c\rho^t, \quad t = 1, 2, \ldots & (20) \end{cases}$$

Proof: The proof is provided in Appendix B. ■

Remarks on Theorem 2: Theorem 2 reveals that the stopping time of our proposed sequential hypothesis test is exponentially bounded, implying that the convergency speed is an exponentially decaying function.

C. Bounds and Convergency on Stopping Time for Neyman-Pearson Hypothesis Test

Simplify the notation of $P^{\mathrm{F}}(j,i)$ as $\alpha_{j,i}$. For every $0 < \eta < 1$, there exists $\widetilde{\eta} > 1$ such that $\eta\widetilde{\eta} < 1$, and let t^* be the greatest integer such that

$$t^* \leq \eta \min_{j,i,j\neq i}\left\{\frac{|\log\alpha_{j,i}|}{I_{j,i}}\right\}. \tag{21}$$

Since P_j is the σ-finite probability measure as defined by Eq. (1) and Eq. (2) we use $\int \phi dP_j$ to represent $P_j\{\phi\}$ if ϕ is

a measurable function. Then, for the decision-functions pair (ψ_t, δ_t), we can derive the probability $\alpha_{j,i}$ as follows:

$$\begin{aligned}\alpha_{j,i} &\triangleq P^{\mathrm{F}}(j,i) = \int_{\{j\neq i\}} \exp\left\{-\sum_{q=1}^{T} Z_{j,i}(q)\right\} dP_j \\ &\overset{(a)}{=} \int_{\{j\neq i\}} \exp\left\{-l_T(j,i)\right\} dP_j \\ &\geq \int_{\{j\neq i, T\leq t^*, l_T(j,i)\leq \widetilde{\eta} t^* I_{j,i}\}} \exp\left\{-l_T(j,i)\right\} dP_j \\ &\geq \exp\{-\widetilde{\eta} t^* I_{j,i}\} P_j\{j\neq i, T\leq t^*, l_T(j,i)\leq \widetilde{\eta} t^* I_{j,i}\} \quad (22)\end{aligned}$$

where $j, i \in \mathcal{J}$, (a) holds due to Eq. (4) with $\{X_q\}, \forall q$, being i.i.d. and Eq. (16), and $Z_{j,i}(q)$ is specified in Eq. (4). From Eq. (21), we can get

$$\widetilde{\eta} t^* I_{j,i} \leq \eta\widetilde{\eta} \left|\log \alpha_{j,i}\right| \quad (23)$$

and thus, it follows from Eq. (22) and Eq. (23) that

$$\begin{aligned}\alpha_{j,i} &\geq \exp\left\{-\eta\widetilde{\eta}\left|\log\alpha_{j,i}\right|\right\} P_j\{j\neq i, T\leq t^*, l_T(j,i)\leq \widetilde{\eta} t^* I_{j,i}\} \\ &= \exp\left\{\eta\widetilde{\eta}\log\alpha_{j,i}\right\} P_j\left\{j\neq i, T\leq t^*, l_T(j,i)\leq \widetilde{\eta} t^* I_{j,i}\right\} \\ &= \left(\alpha_{j,i}\right)^{\eta\widetilde{\eta}} P_j\left\{j\neq i, T\leq t^*, l_T(j,i)\leq \widetilde{\eta} t^* I_{j,i}\right\}. \quad (24)\end{aligned}$$

From Eq. (24), we can further obtain the followings:

$$\left(\alpha_{j,i}\right)^{1-\eta\widetilde{\eta}} \geq P_j\left\{j\neq i, T\leq t^*, l_T(j,i)\leq \widetilde{\eta} t^* I_{j,i}\right\}. \quad (25)$$

Using Eq. (25), we can get

$$\begin{aligned}&P_j\left\{j\neq i, T\leq t^*\right\} \\ &\leq \left(\alpha_{j,i}\right)^{1-\eta\widetilde{\eta}} + P_j\left\{T\leq t^*, l_T(j,i)\geq \widetilde{\eta} t^* I_{j,i}\right\} \\ &\leq \left(\alpha_{j,i}\right)^{1-\eta\widetilde{\eta}} + P_j\left\{\max_{T\leq t^*} l_T(j,i)\geq \widetilde{\eta} t^* I_{j,i}\right\}. \quad (26)\end{aligned}$$

Since Eq. (26) holds for each j, summing up Eq. (26)'s for all j with $j\neq i$ and then taking the supremum over (ψ_t, δ_t), we obtain:

$$\begin{aligned}\sup_{(\psi_t,\delta_t)} P_j\left\{T\leq t^*\right\} &\leq \sum_{j\in\mathcal{J}, j\neq i} \left(\alpha_{j,i}\right)^{1-\eta\widetilde{\eta}} \\ &+ \sum_{j\in\mathcal{J}, j\neq i} P_j\left\{\max_{T\leq t^*} \log\frac{f_{r_j}(X_1,\ldots,X_T)}{f_{r_i}(X_1,\ldots,X_T)} \geq \widetilde{\eta} t^* I_{j,i}\right\}. \quad (27)\end{aligned}$$

Since $\sum_j \alpha_{j,i} \to 0$, $\widetilde{\eta} > 1$, and using Eq. (23) and

$$\frac{1}{T}\log\frac{f_{r_j}(X_1,\ldots,X_T)}{f_{r_i}(X_1,\ldots,X_T)} \to I_{j,i}, \text{ a.s. } [P_j] \quad (28)$$

which is due to Eq. (14), we can further derive Eq. (27) as follows:

$$\sup_{(\psi_t,\delta_t)} P_j\left\{T\leq t^*\right\} \leq \sum_{j\in\mathcal{J}, j\neq i} P_j\left\{\max_{T\leq t^*} T \geq \eta\frac{\left|\log\alpha_{j,i}\right|}{I_{j,i}}\right\}. \quad (29)$$

Thus, for every $0<\eta<1$, we have

$$\inf_{(\psi_t,\delta_t)} P_j\left\{T\geq \eta \min_{j\neq i}\left\{\frac{\left|\log\alpha_{j,i}\right|}{I_{j,i}}\right\}\right\} \to 1. \quad (30)$$

For each $\alpha_{j,i}$, let $C_{j,i}$ be a positive constant such that

$$\log C_{j,i} \propto \left|\log\alpha_{j,i}\right|, \quad \text{as } \sum_{j,i}\alpha_{j,i}\to 0 \quad (31)$$

where $\propto$ is defined in the text following Eq. (10). Define U as

$$U \triangleq \inf\left\{t\geq 1 : \prod_{q=1}^{t}\frac{f_{r_j}(X_q)}{f_{r_i}(X_q)} \geq C_{j,i}\right\}, \quad (32)$$

implying that the test stops sampling at time slot U and accepts $\mathcal{H}_j$ if $z_{j,i}^{(U)} \geq C_{j,i}$. As $\sum_{j,i}\alpha_{j,i}\to 0$, using Eq. (28) and Eq. (31), we get

$$\frac{U}{\min_{j\neq i}\{\left|\log\alpha_{j,i}\right|/I_{j,i}\}} \to 1, \text{ a.s. } [P_j]. \quad (33)$$

Thus, for every $0<\eta<1$, we get

$$\inf_{(\psi_t,\delta_t)} \Pr\left\{T\geq \eta U\right\} \to 1 \quad (34)$$

and the error probability of the test as follows:

$$P_j\{(\psi_t,\delta_t) \text{ rejects } \mathcal{H}_j\} \leq \frac{1}{C_{j,i}} P_i\{(\psi_t,\delta_t) \text{ rejects } \mathcal{H}_i\}. \quad (35)$$

We also extend the results of Eq. (34) into the notion of *κ-quick convergency*, which is defined as follows: for $\kappa>0$, a sequence $\{\Psi_t\}$ of random variables is said to *converge κ-quickly* [32], to a constant λ if $\mathbb{E}[(L_a)^\kappa] < \infty$ for all $a>0$, where $L_a = \sup_t\{t\geq 1 : |\Psi_t - \lambda| \geq a\}$. Therefore, we can obtain from Eq. (30) that

$$\inf_{(\psi_t,\delta_t)} \mathbb{E}\left[T^\kappa\right] \geq \eta^\kappa \left(\min_{j\neq i}\left\{\frac{\left|\log\alpha_{j,i}\right|}{I_{j,i}}\right\}\right)^\kappa (1+o(1)) \quad (36)$$

where $\varphi(x) = o(g(x))$ if $\lim_{x\to\infty}\varphi(x)/g(x) = 0$. Using the definition of κ-quick convergency given above, let $0<a<\min_{j\neq i}\{I_{j,i}\}$, and define

$$L_a \triangleq \sup_t\left\{t\geq 1 : \max_{\substack{i=0,\ldots,J\\ j\neq i}}\left|\frac{1}{t}\log\frac{f_{r_j}(X_1,\ldots,X_t)}{f_{r_i}(X_1,\ldots,X_t)} - I_{j,i}\right| > a\right\}. \quad (37)$$

When $U-1\geq L_a$, we have

$$\left|\frac{1}{U-1}\log\frac{f_{r_j}(X_1,\ldots,X_T)}{f_{r_i}(X_1,\ldots,X_T)} - I_{j,i}\right| \leq a. \quad (38)$$

Suppose $I_{j,i}>0$, Eq. (38) is equivalent to

$$(I_{j,i}-a)(U-1) < \log\frac{f_{r_j}(X_1,\ldots,X_{U-1})}{f_{r_i}(X_1,\ldots,X_{U-1})} < \log C_{j,i}. \quad (39)$$

Then, for $U\leq L_a+1$, we have the complementary event of Eq. (39) as follows:

$$U \geq 1 + \min_{j\neq i}\left\{\frac{\log C_{j,i}}{I_{j,i}-a}\right\}. \quad (40)$$

For some positive constant κ, Eq. (28) can be further enhanced into

$$\frac{1}{T}\log\frac{f_{r_j}(X_1,\ldots,X_T)}{f_{r_i}(X_1,\ldots,X_T)} \to I_{j,i}, \quad \kappa\text{-quickly under } P_j. \quad (41)$$

Thus, combing Eq. (40) and Eq. (41) we obtain the lower-bound on the stopping time as follows:

$$\inf_{(\psi_t,\delta_t)} \mathbb{E}\left[T^\kappa\right] \propto \mathbb{E}\left[U^\kappa\right] \geq \left(\min_{j\neq i}\left\{\frac{\left|\log\alpha_{j,i}\right|}{I_{j,i}}\right\}\right)^\kappa. \quad (42)$$

IV. NFV/SDN Architecture to Maximize ϵ-Effective Capacity for Each Metaverse Streaming Virtual Slice

After observing all MUs' data requests and determining an optimal hypothesis for each MU, we apply the NFV/SDN architectures to map all MUs which are predicted to have the same metaverse streaming request probability profile, according to their corresponding accepted hypotheses, into one virtual network slice and allocate the optimal wireless resources, i.e., transmit power, to maximize their *average aggregate ϵ-effective capacity* [3], [4]. We derive the average aggregate ϵ-effective capacity by first obtaining the aggregate ϵ-effective capacity over all MUs, and then calculating its average using each metaverse data request probability.

A. The ϵ-Effective Capacity Over Nakagami-m Fading for Metaverse Streaming Through Single Antenna Transmission

We employ the Nakagami-m fading channel model, since it is a more practical model to characterize the multipath scattering. Let h_k be the fading amplitude of the kth MU following the Nakagami-m distribution, and let N_0 be the power of additive white Gaussian noise (AWGN). The SNR of the kth MU is given by $\gamma_k(\mathcal{P}_{d,k,j}) = (h_k^2\mathcal{P}_{d,k,j})/N_0$. Under the Nakagami-$m$ fading wireless channel, the probability density function (pdf) of the SNR $\gamma_k(\mathcal{P}_{d,k,j})$, denoted by $P_\Gamma(\gamma_k)$, is given by

$$P_\Gamma(\gamma_k) = \frac{\gamma_k^{m-1}}{\Gamma(m)}\left(\frac{m}{\overline{\gamma}_k}\right)^m \exp\left(-\frac{m}{\overline{\gamma}_k}\gamma_k\right) \tag{43}$$

where m is the fading parameter of the Nakagami-m distribution, $\overline{\gamma}_k = (\mathbb{E}[h_k^2]\mathcal{P}_{d,k,j})/N_0$ is the average SNR, $\forall k$, and $\Gamma(\cdot)$ is the gamma function. Under the Nakagami-m fading channel, we give a closed-form expression for the ϵ-effective capacity, defined in Eq. (9), in the following theorem.

Theorem 3: Under our proposed NFV/SDN architectures for metaverse streaming, **if** the metaverse streaming is transmitted through a single antenna wireless channel experiencing the Nakagami-m fading with $m > 1$, where the pdf of the received SNR is characterized by Eq. (43), **then** the closed-form expression for the ϵ-effective capacity defined by Eq. (9) for the kth metaverse MU using the (n, W, ϵ)-code in the non-asymptotic regime is determined by

$$EC_k(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}) = -\frac{1}{n\theta_d}\left\{\log\left[\epsilon_d + (1-\epsilon_d)\left(\frac{1+\overline{\gamma}_k}{2^{\widetilde{\epsilon}_d\sqrt{V(\overline{\gamma}_k)}}}\right)^{-\widetilde{\theta}_d}\right]\right\} \tag{44}$$

where $\widetilde{\theta}_d \triangleq (\log_2 e)\theta_d n$, $\widetilde{\epsilon}_d \triangleq Q^{-1}(\epsilon_d)/\sqrt{n}$, and $V(\overline{\gamma}_k) \approx 1 - [1/(1+\overline{\gamma}_k)^2]$ is the channel dispersion of the AWGN channel [12, Eq. (293)], where $Q^{-1}(\cdot)$ is the inverse of the Q-function.

Proof: The proof is provided in Appendix C. ■

Remarks on Theorem 3: The closed-form expression derived in Eq. (44) of Theorem 3 identifies the explicit relationships between the ϵ-effective capacity and other important control variables over a single-input-single-output channel, which are to be used to derive the ϵ-effective capacity over the massive MIMO channel in Theorem 5 of this paper.

B. Channel Estimations for Massive MIMO Communications for Metaverse Mobile Users

Suppose that there are M_T antennas on the BS/AP and there are M_R antennas for each MU, where $M_T \gg M_R$. Denote by $\mathbf{g}_{k,\nu} \in \mathbb{C}^{M_R\times 1}$ the channel gain between the kth MU and the νth antenna on the massive antenna equipped BS/AP, where $\mathbb{C}^{M_R\times 1}$ denotes a set of elements each consisting of a complex-valued matrix with M_R rows and one column. Denote by R_k the distance between the kth MU and antennas of the BS/AP, assuming that the distance between two antennas on the BS/AP is small comparing with the distance between an MU and the BS/AP. We give $\mathbf{g}_{k,\nu}$ as follows [5, Eq. (2.19)]:

$$\mathbf{g}_{k,\nu} = \sqrt{\beta_k}\mathbf{h}_{k,\nu} \tag{45}$$

where $\beta_k \approx [\lambda/(4\pi R_k)]^2$ is the large-scale fading coefficient, where λ is the wavelength, and $\mathbf{h}_{k,\nu} \in \mathbb{C}^{M_R\times 1}$ indicates the effect of small-scale fading between all antennas on the kth MU and the νth antenna on the BS/AP. We consider that each coherence interval is divided into two phases: (1) ***uplink training phase*** to estimate the channel gain and (2) ***downlink payload data transmission phase*** to download the data.

*1) **Uplink Training Phase:*** Denote by $\tau_{\text{ul,p}}$ the number of samples for the uplink pilot signal, where we assume that $\tau_{\text{ul,p}} \geq M_R$. Define $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_{\tau_{\text{ul,p}}}] \in \mathbb{C}^{1\times\tau_{\text{ul,p}}}$ as an orthogonal pilot training sequence satisfying $\|\boldsymbol{\phi}\|^2 = 1$, where $\|\cdot\|$ is the Euclidean norm. The pilot signal sending from the kth MU to the BS/AP is denoted by $\mathbf{x}_k^{(\text{p})} = \sqrt{\tau_{\text{ul,p}}}\boldsymbol{\phi} \in \mathbb{C}^{1\times\tau_{\text{ul,p}}}$. In the training phase, we assign M_R orthogonal pilot sequences to M_R antennas of the MU k, and both the MU k and the BS/AP know these pilot sequences. Let ρ_{ul} be the transmit power over uplink and $\mathbf{W}^{(\text{p})} \in \mathbb{C}^{M_R\times\tau_{\text{ul,p}}}$ be the AWGN matrix, whose elements are i.i.d., following the complex Gaussian distribution $\mathcal{CN}(0,1)$. The received pilot signal, denoted by $\mathbf{Y}_{k,\nu}^{(\text{p})} \in \mathbb{C}^{M_R\times\tau_{\text{ul,p}}}$, at the νth antenna of the BS/AP, is given by

$$\mathbf{Y}_{k,\nu}^{(\text{p})} = \sqrt{\rho_{\text{ul}}}\mathbf{g}_{k,\nu}\mathbf{x}_k^{(\text{p})} + \mathbf{W}^{(\text{p})} = \sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\mathbf{g}_{k,\nu}\boldsymbol{\phi} + \mathbf{W}^{(\text{p})}. \tag{46}$$

Applying the de-spreading scheme [5, Section 3.1.2] to the received pilot signal, the BS/AP performs a de-spreading operation by correlating its received signals with the pilot signal. Denote by $\overline{\mathbf{y}}_{k,\nu}^{(\text{p})} \in \mathbb{C}^{M_R\times 1}$ the received signal after the de-spreading operation, which is given by

$$\overline{\mathbf{y}}_{k,\nu}^{(\text{p})} = \mathbf{Y}_{k,\nu}^{(\text{p})}\boldsymbol{\phi}^H = \sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\mathbf{g}_{k,\nu} + \overline{\mathbf{w}}^{(\text{p})} \tag{47}$$

where $(\cdot)^H$ denotes the Hermitian transpose, $\overline{\mathbf{w}}^{(\text{p})} \triangleq \mathbf{W}^{(\text{p})}\boldsymbol{\phi}^H \in \mathbb{C}^{M_R\times 1}$ is the AWGN after de-spreading, and each element of $\overline{\mathbf{w}}^{(\text{p})}$ follows $\mathcal{CN}(0,1)$. Let $\mathbf{G}_k = [\mathbf{g}_{k,1}, \mathbf{g}_{k,2}, \cdots, \mathbf{g}_{k,M_T}] \in \mathbb{C}^{M_R\times M_T}$ be the channel gain matrix between all antennas on the kth MU and all antennas on the BS/AP. Let $\widehat{\mathbf{G}}_k = [\widehat{\mathbf{g}}_{k,1}, \widehat{\mathbf{g}}_{k,2}, \cdots, \widehat{\mathbf{g}}_{k,M_T}] \in \mathbb{C}^{M_R\times M_T}$ be the estimated channel gain matrix, indicating the estimation of $\mathbf{G}_k$. Using the minimum mean-square error (MMSE) estimation, we obtain the estimated channel gain $\widehat{g}_{k,\nu}^{(l)}$ between the lth antenna (with $\forall l \in \{1, \cdots, M_R\}$) on the kth MU and the νth antenna (with $\forall \nu \in \{1, \cdots, M_T\}$) on the BS/AP as

follows [5, Eq. (3.7)] [33, Eq. (4)]:

$$\widehat{g}_{k,\nu}^{(l)}=\mathbb{E}\left[g_{k,\nu}^{(l)}\middle|\overline{y}_{k,\nu}^{(\mathrm{p},l)}\right]=\frac{\mathbb{E}\left[\overline{y}_{k,\nu}^{(\mathrm{p},l)*}g_{k,\nu}^{(l)}\right]}{\mathbb{E}\left[\left|\overline{y}_{k,\nu}^{(\mathrm{p},l)}\right|^2\right]}\overline{y}_{k,\nu}^{(\mathrm{p},l)}=\frac{\sqrt{\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}}\beta_k}{1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_k}\overline{y}_{k,\nu}^{(\mathrm{p},l)} \tag{48}$$

where $\mathbb{E}[\cdot|\cdot]$ denotes the conditional expectation, $(\cdot)^*$ denotes the conjugate, and $\widehat{g}_{k,\nu}^{(l)}$ and $\overline{y}_{k,\nu}^{(\mathrm{p},l)}$ are the lth element of $\widehat{\mathbf{g}}_{k,\nu}$ and $\overline{\mathbf{y}}_{k,\nu}^{(\mathrm{p})}$, respectively. Substituting each element of $\overline{\mathbf{y}}_{k,\nu}^{(\mathrm{p})}$ given by Eq. (47) into Eq. (48), the channel estimation $\widehat{\mathbf{g}}_{k,\nu}$ is given by

$$\widehat{\mathbf{g}}_{k,\nu}=\frac{\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_k}{1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_k}\mathbf{g}_{k,\nu}+\frac{\sqrt{\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}}\beta_k}{1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_k}\overline{\mathbf{w}}^{(\mathrm{p})}. \tag{49}$$

*2) **Downlink Payload Data Transmission Phase:*** During the downlink payload data transmission phase, the BS/AP treats the channel estimation $\widehat{\mathbf{g}}_{k,\nu}$ as the true channel to transmit the data packet to the MU k. Let b_k be symbol intended to the MU k, satisfying $\mathbb{E}[|b_k|^2]=1$. Let $\mathbf{x}=[x_1,x_2,\cdots,x_{M_T}]^\intercal\in\mathbb{C}^{M_T\times 1}$ be the weighted symbol transmitted from all antennas of the BS/AP, where $(\cdot)^\intercal$ is the transpose. Let K_d be the total number of MUs in a virtual network slice requesting the same metaverse data item d. Let $P_{\max}$ be the total transmit power from the BS/AP to all MUs for transmitting the same metaverse data. Using the maximum ratio transmission (MRT) precoding as the beamforming scheme to focus the signal of the payload metaverse data towards the kth MU, each element $x_\nu,\forall\nu$, of $\mathbf{x}$, which is the transmit signal on the νth antenna of the BS/AP, is given by

$$x_\nu=\sum_{k=1}^{K_d}\sqrt{\mathcal{P}_{d,k,j}}\left(\boldsymbol{\eta}_{k,\nu}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{k,\nu}^{*}b_k, \tag{50}$$

where $\mathcal{P}_{d,k,j}$ is the downlink transmit power for transmitting the data item d to the kth MU if accepting $\mathcal{H}_j$ as its data request pmf profile and $\sum_{k=1}^{K_d}\mathcal{P}_{d,k,j}=P_{\max},\forall d,j$, $\boldsymbol{\eta}_{k,\nu}\in\mathbb{R}^{1\times M_R}$ is the power control coefficient for the signal from the νth antenna of the BS/AP to the kth MU, each element of $\boldsymbol{\eta}_{k,\nu}$, denoted by $\eta_{k,\nu}^{(l)}$, satisfies $\eta_{k,\nu}^{(l)}\in[0,1],\forall l$, and $(\cdot)^{\frac{1}{2}}$ is taking square root for each element. The received signal at the kth MU, denoted by $\mathbf{y}_k\in\mathbb{C}^{M_R\times 1}$, is given by $\mathbf{y}_k=\mathbf{G}_k\mathbf{x}+\mathbf{w}_k$, where $\mathbf{w}_k\in\mathbb{C}^{M_R\times 1}$ is the AWGN on all antennas of the kth MU, whose element is denoted by $w_k^{(l)}$ following $\mathcal{CN}(0,1)$. Each element $y_k^{(l)}$ of $\mathbf{y}_k$, $\forall l\in\{1,2\cdots,M_R\}$, representing the received signal by the lth antenna of MU k, is given by

$$\begin{aligned}y_k^{(l)}=&\sum_{\nu=1}^{M_T}g_{k,\nu}^{(l)}x_\nu+w_k^{(l)}=\sqrt{\mathcal{P}_{d,k,j}}\sum_{\nu=1}^{M_T}g_{k,\nu}^{(l)}\left(\boldsymbol{\eta}_{k,\nu}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{k,\nu}^{*}b_k\\&\underbrace{+\sum_{\nu=1}^{M_T}\sum_{u=1,u\neq k}^{K}\sqrt{\mathcal{P}_{d,u,j}}g_{k,\nu}^{(l)}\left(\boldsymbol{\eta}_{u,\nu}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,\nu}^{*}b_u+w_k^{(l)}}_{\text{effective additive noise }N_k^{(l)}}\end{aligned} \tag{51}$$

where b_u is the symbol intended to the MU $u,u\neq k$, and $N_k^{(l)}$ is the *effective additive noise* of the kth MU on its lth antenna.

C. The Average SNR Over Nakagami-m Channels for Massive MIMO Metaverse Streaming

Denote the SNR of the lth antenna on the kth MU under the BS/AP power allocation $\mathcal{P}_{d,k,j}$ by $\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})$ over the massive MIMO channel. The key step to derive ϵ-effective capacity of massive MIMO communications is deriving the expression for $\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})$. Using Eq. (51), the SNR $\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})$, $\forall l$, for the massive MIMO channel is given by [34] and [35]

$$\begin{aligned}&\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})\\&\triangleq\frac{\mathrm{Var}\left[\sqrt{\mathcal{P}_{d,k,j}}b_k\sum_{\nu=1}^{M_T}g_{k,\nu}^{(l)}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}\widehat{g}_{k,\nu}^{(i)*}\right]}{\mathrm{Var}\left[\sum_{\nu=1}^{M_T}\sum_{u=1,u\neq k}^{K}\sqrt{\mathcal{P}_{d,u,j}}g_{k,\nu}^{(l)}\left(\boldsymbol{\eta}_{u,\nu}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,\nu}^{*}b_u\right]+1}\end{aligned} \tag{52}$$

where $\mathrm{Var}[\cdot]$ is derived with respect to both the random distance R_k and the random $h_{k,\nu}^{(l)}$, which is the lth element of $\mathbf{h}_{k,\nu}$, representing the small-scale fading amplitude between the lth antenna on MU k and the νth antenna of the BS/AP. We assume that all $h_{k,\nu}^{(l)},\forall l,\nu$, are i.i.d., following the Nakagami-m distribution. We can derive a closed-form expression for $\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})$ specified by Eq. (52) in the following theorem.

Theorem 4: For our proposed NFV/SDN architectures to support metaverse streaming, **if** the metaverse streaming is transmitted over a massive MIMO channel experiencing the Nakagami-m fading and assume that all metaverse MUs are uniformly distributed within a wireless cell with the inner radius $R_{\min}$ and the outer radius $R_{\max}$, **then** we can derive a closed-form expression for SNR $\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})$, which is defined by Eq. (52), as follows:

$$\gamma_{k,(l)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})=\frac{\mathcal{P}_{d,k,j}N_{k,1}}{(P_{\max}-\mathcal{P}_{d,k,j})N_{k,2}+1} \tag{53}$$

where

$$N_{k,1}=\frac{\beta_k^3\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\overline{\eta}M_TM_R\overline{h}^2}{(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_k)^2}\left(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_kM_TM_R\overline{h}^2\right) \tag{54}$$

and

$$\begin{aligned}N_{k,2}=&\frac{\beta_k\overline{\eta}M_TM_R\overline{h}^2\lambda^2}{(4\pi)^2(R_{\max}^2-R_{\min}^2)}\left\{\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\left(\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\overline{h}^2+1\right)(X_{\max}-X_{\min})\right.\\&\left.+\frac{M_T(M_R-1)\lambda^2}{(4\pi)^2(R_{\max}^2-R_{\min}^2)}\left[\log\left(\frac{X_{\max}}{X_{\min}}\right)\right]^2\right\},\end{aligned} \tag{55}$$

where $\overline{\eta}\triangleq\mathbb{E}\left[\eta_{k,\alpha}^{(i)}\right]$, $\overline{h}\triangleq\mathbb{E}\left[h_{k,\nu}^{(l)}\right]$, and

$$\begin{cases}X_{\max}=\dfrac{\lambda^2}{16\pi^2(R_{\min}^2+\iota^2)+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\lambda^2},\\X_{\min}=\dfrac{\lambda^2}{16\pi^2(R_{\max}^2+\iota^2)+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\lambda^2},\end{cases} \tag{56}$$

where we define ι as the height of a BS/AP.

Proof: The proof is provided in Appendix D. ■

Remarks on Theorem 4: Theorem 4 reveals that all $\gamma_{k,(l)}^{\text{MIMO}}(\mathcal{P}_{d,k,j}), \forall l$, are the same, since random variables $h_{k,\nu}^{(l)}, \forall l, \nu$, are i.i.d.

According to Eq. (44), the ϵ-effective capacity is a function of the average SNR for a wireless channel. Using *Remarks on Theorem 4*, the average SNR for the massive MIMO channel, denoted by $\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})$, over all elements $\gamma_{k,(l)}^{\text{MIMO}}(\mathcal{P}_{d,k,j})$, $\forall l$, is given by

$$\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}) = \gamma_{k,(l)}^{\text{MIMO}}(\mathcal{P}_{d,k,j}), \quad \forall l. \tag{57}$$

D. The Optimal Transmit Power Allocation for Maximizing Aggregate ϵ-Effective Capacity Over Massive-MIMO Channels for Each Virtual Network Slice

Define $\boldsymbol{\mathcal{P}}_{k,j} \triangleq [\mathcal{P}_{1,k,j}, \cdots, \mathcal{P}_{D,k,j}]^{\intercal}$ as the transmit power allocations for transmitting all metaverse data items to the kth MU if accepting $\mathcal{H}_j$ as the data requests pmf profile, and define $\boldsymbol{\mathcal{P}}_{k,j}^*$ as the optimal value of $\boldsymbol{\mathcal{P}}_{k,j}$ that maximizes the average aggregate ϵ-effective capacity. Defining $EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ as the ϵ-effective capacity for the kth MU over the massive MIMO channel, we formulate the average aggregate ϵ-effective capacity maximization problem for a virtual network slice as follows:

$$\begin{aligned}
&\left[\boldsymbol{\mathcal{P}}_{1,j}^*, \cdots, \boldsymbol{\mathcal{P}}_{K,j}^*\right] \\
&= \arg \max_{[\boldsymbol{\mathcal{P}}_{1,j}, \cdots, \boldsymbol{\mathcal{P}}_{K,j}]} \sum_{d=1}^{D} f_{r_j}(d) \sum_{k=1}^{K_d} EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}) \qquad (58)\\
&\quad \textbf{s.t.: C1: } \sum_{k=1}^{K_d} \mathcal{P}_{d,k,j} \le P_{\max}, \\
&\quad\quad \textbf{C2: } \mathcal{P}_{d,k,j} \ge 0, \ \forall k.
\end{aligned}$$

To solve the maximization problem specified by Eq. (58), we give closed-form expressions for the maximum average aggregate ϵ-effective capacity and the optimal transmit power allocation $\boldsymbol{\mathcal{P}}_{k,j}^*$, respectively, in the following theorem, by extending the results of Theorem 3 over the single antenna channel into the massive MIMO channel version.

Theorem 5: **If** the metaverse streaming for each virtual network slice is dictated under our proposed Neyman-Pearson hypothesis testing driven massive MIMO NFV/SDN architectures and optimal resource allocation schemes, **then** the following two claims hold true.

Claim 1. The closed-form expression for the maximum ϵ-effective capacity, denoted by $EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}^*)$, for the kth MU over the massive-MIMO channel in its corresponding virtual network slice is determined by

$$\begin{aligned}
&EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}^*) \\
&= -\frac{1}{n\theta_d} \left\{ \log \left[\epsilon_d + (1-\epsilon_d) \left(\frac{1 + \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*)}{2^{\widetilde{\epsilon}_d \sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*))}}} \right)^{-\widetilde{\theta}_d M_R} \right] \right\}
\end{aligned} \tag{59}$$

where $\mathcal{P}_{d,k,j}^*$ is the optimal $\mathcal{P}_{d,k,j}$ to be specified by **Claim 2** of this theorem, and $\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*)$ is the average SNR under the optimal transmit power allocation $\mathcal{P}_{d,k,j}^*$ given by

$$\begin{aligned}
&\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*) \\
&= \begin{cases}
\frac{1}{N_{k,2}} \left[\frac{\varrho_{\gamma,1}^* 2^{\widetilde{\epsilon}_k} N_{k,1}}{f_{r_j}(d)} (1 + N_{k,2} P_{\max}) \right]^{\frac{1}{2}} - \frac{N_{k,1}}{N_{k,2}}, \\
\quad \text{if } \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*) \gg 1, \\
\frac{1}{N_{k,2}} \left[\frac{\varrho_{\gamma,2}^* N_{k,1}}{f_{r_j}(d)} (1 + N_{k,2} P_{\max}) \right]^{\frac{1}{2}} - \frac{N_{k,1}}{N_{k,2}}, \\
\quad \text{if } 0 < \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*) < 1,
\end{cases}
\end{aligned} \tag{60}$$

where $N_{k,1}$ and $N_{k,2}$ are given by Eq. (54) and Eq. (55), respectively, and $\varrho_{\gamma,1}^*$ and $\varrho_{\gamma,2}^*$ are given, respectively, by the following equations:

$$\begin{cases}
\varrho_{\gamma,1}^* = f_{r_j}(d) \left(\dfrac{\sum_{k=1}^{K_d} \frac{1}{N_{k,2}} \left[\frac{N_{k,1}}{2^{\widetilde{\epsilon}_k}} (1 + N_{k,2} P_{\max}) \right]^{\frac{1}{2}}}{\sum_{k=1}^{K_d} \frac{1 + N_{k,2} P_{\max}}{N_{k,2}} - P_{\max}} \right)^2, \\
\varrho_{\gamma,2}^* = f_{r_j}(d) \left(\dfrac{\sum_{k=1}^{K_d} \frac{1}{N_{k,2}} \left[N_{k,1} (1 + N_{k,2} P_{\max}) \right]^{\frac{1}{2}}}{\sum_{k=1}^{K_d} \frac{1 + N_{k,2} P_{\max}}{N_{k,2}} - P_{\max}} \right)^2.
\end{cases} \tag{61}$$

Claim 2. The closed-form expression for the optimal solution $\left[\boldsymbol{\mathcal{P}}_{1,j}^*, \cdots, \boldsymbol{\mathcal{P}}_{K,j}^*\right]$ for the maximization problem given by Eq. (58) can be obtained by deriving each of its element, denoted by $\mathcal{P}_{d,k,j}^*, \forall k, d$, which is the optimal transmit power allocation for sending the data item d to the kth MU using the estimated data request pmf profile $f_{r_j}(d)$ obtained from Section III-A. The closed-form expression for $\mathcal{P}_{d,k,j}^*, \forall k, d$, is determined as follows:

$$\begin{aligned}
\mathcal{P}_{d,k,j}^* = &\begin{cases}
\frac{1 + N_{k,2} P_{\max}}{N_{k,2}} - \frac{1}{N_{k,2}} \left[f_{r_j}(d) N_{k,1} \frac{1 + N_{k,2} P_{\max}}{\varrho_{\gamma,1}^* 2^{\widehat{\epsilon}_k}} \right]^{\frac{1}{2}}, \\
\quad \text{if } \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*) \gg 1, \\
\frac{1 + N_{k,2} P_{\max}}{N_{k,2}} - \frac{1}{N_{k,2}} \left[f_{r_j}(d) N_{k,1} \frac{1 + N_{k,2} P_{\max}}{\varrho_{\gamma,2}^*} \right]^{\frac{1}{2}}, \\
\quad \text{if } 0 < \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}^*) < 1.
\end{cases}
\end{aligned} \tag{62}$$

Proof: The proof is provided in Appendix E. ■

V. Performance Evaluations

We conduct the extensive numerical analyses to validate and evaluate our developed schemes. Figure 2 shows a case study of testing a set of metaverse data items between two hypotheses with Zipf exponents $r_j = 0.3$ and $r_i = 0.1$. In this case study, suppose that there are totally 99 metaverse data items. We set that the e-health care data is at the top-ranked data range, the educational data is at the medium-ranked data range, and the online gaming data is at the low-ranked data range. For sake of the fairness, we also set that each type of metaverse data evenly shares the total 99 data items, and thus,

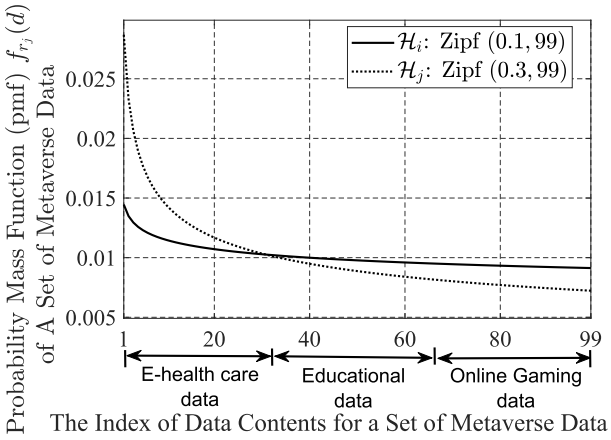


Fig. 2. A case study of request probability profile for three categories of metaverse data items with r_j = 0.1 and r_i = 0.3.

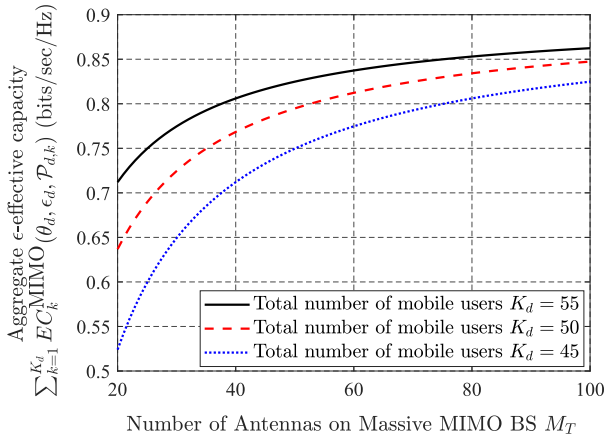


Fig. 4. The function of aggregate ϵ-effective capacity $\sum_{k=1}^{K_d} EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ with different numbers of antenna M_T on the massive-MIMO BS/AP.

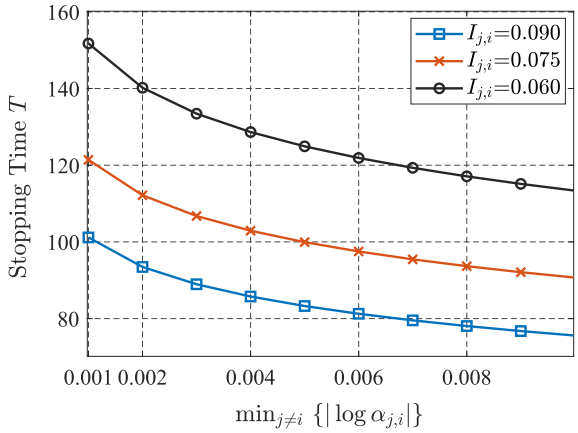


Fig. 3. The stopping time T of the sequential hypothesis testing against the minimum value of hypothesis testing error probability under different values of the Zipf distributions difference $I_{j,i}$.

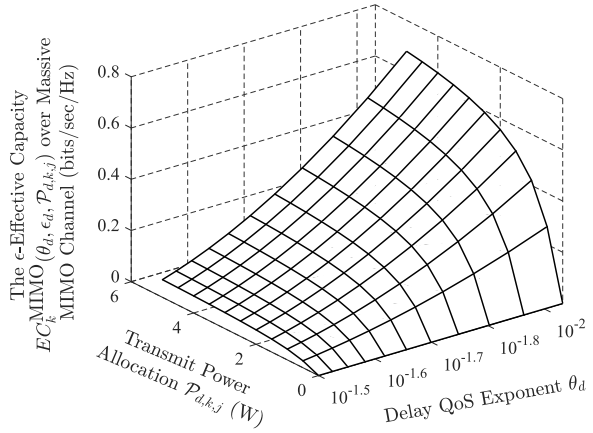


Fig. 5. The ϵ-effective capacity $EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ under different values of the delay QoS exponent θ_d and the transmit power allocation $\mathcal{P}_{d,k,j}$.

the number of each type of metaverse data items is 33. We take t observations for an MU and select a pmf profile between $\mathcal{H}_i$ and $\mathcal{H}_j$. If our proposed sequential hypothesis testing selects $\mathcal{H}_j$, this MU is more likely to request more e-health care metaverse streaming with a distribution Zipf$(0.3, 99)$ in the future and we can map it into a virtual network slice where other MUs also request the e-health care data with the distribution Zipf$(0.3, 99)$. Otherwise, if accepting $\mathcal{H}_i$, the MU is more likely to request e-health care data with a distribution Zipf$(0.1, 99)$ and we map it into a virtual network slice that other MUs also request the e-health care data with the distribution Zipf$(0.1, 99)$.

Figure 3 plots the stopping time T versus the minimum value of error probability $\min_{j\neq i}\{|\log \alpha_{j,i}|\}$ of the sequential hypotheses testing under three values for the difference of two Zipf distributions $I_{j,i} = \left|\mathbb{E}_{P_j}\left[\log\left\{f_{r_j}(X_1)/f_{r_i}(X_1)\right\}\right]\right|$. Figure 3 shows that the stopping time T is a decreasing function of $\min_{j\neq i}\{|\log \alpha_{j,i}|\}$, because a larger number of observations yields a hypothesis testing with a lower error probability. Figure 3 also shows that the stopping time T is a decreasing function of the Zipf distributions difference $I_{j,i}$, because a smaller difference of Zipf distributions also requires a larger number of observations to distinguish two hypotheses.

We show the aggregate ϵ-effective capacity $\sum_{k=1}^{K_d} EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$, i.e., the sum of ϵ-effective capacity over all MUs on a virtual network slice, under different numbers of antennas M_T of the massive-MIMO BS/AP in Fig. 4. We set the number of antennas for each MU as $M_R = 2$; the maximum and minimum distances to the massive-MIMO BS/AP antennas as $R_{\max} = 30$m and $R_{\min} = 5$m, respectively; and $\lambda = 10$m, $\tau_{\text{ul,p}} = 16$, $\rho_{\text{ul}} = 1$W, $P_{\max} = 10$W. Observing Fig. 4, we obtain that the aggregate ϵ-effective capacity monotonically increases as the number of antennas M_T increases, because a larger number of antennas can improve the massive-MIMO channel performance. We can also observe that the increasing rate of the aggregate ϵ-effective capacity decreases as M_T increases, since these antennas also results in the interference to each other.

In Fig. 5, we show the ϵ-effective capacity $EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ of the kth MU under different values of the delay QoS exponent θ_d and the transmit power allocation $\mathcal{P}_{d,k,j}$. The parameters are the same as in Fig. 4. We observe from Fig. 5 that the ϵ-effective capacity is a monotonically increasing function of the transmit power allocation. We also observe that in a high

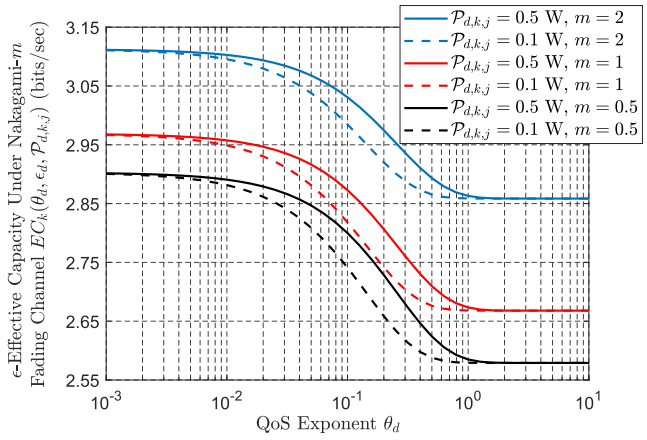


Fig. 6. The function of ϵ-effective capacity with respect to the delay QoS exponent θ_d under different values of the transmit power $\mathcal{P}_{d,k,j}$ and the Nakagami-m fading parameter m.

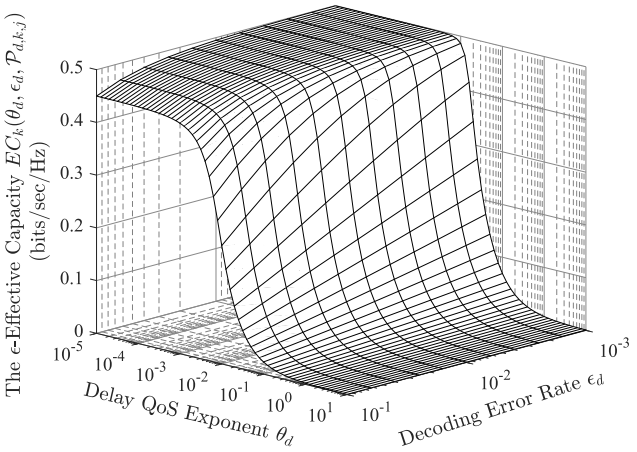


Fig. 7. The function of ϵ-effective capacity under different values of the delay QoS exponent θ_d and the decoding error probability ϵ_d.

power regime ($\mathcal{P}_{d,k,j} > 3$ W), the increasing speed of $EC_k^{\text{MIMO}}(\theta_d,\epsilon_d,\mathcal{P}_{d,k,j})$ decreases as the transmit power increases, showing that a large power allocation cannot produce a large benefit in this regime. Thus, there exists an optimal power $\mathcal{P}^*_{d,k,j}$ that maximizes the average aggregate ϵ-effective capacity under a maximum transmit power constraint.

Figure 6 plots the function of ϵ-effective capacity $EC_k(\theta_d,\epsilon_d,\mathcal{P}_{d,k,j})$ for a single antenna channel with respect to the delay QoS exponent θ_d under different values of the transmit power $\mathcal{P}_{d,k,j}$ and the Nakagami-m fading parameter m. We set $\mathcal{P}_{d,k,j} = 0.1, 0.5$ and $m = 0.5, 1, 2$, respectively. We set other parameters as follows: the length of a codeword $n = 1000$, the average SNR $\overline{\gamma}_k = 20$ dB, and the fading power range $h_k^2 = [0.1, 1]$. Figure 6 shows that for the same Nakagami-m fading parameter, the ϵ-effective capacity increases as the transmit power allocation $\mathcal{P}_{d,k,j}$ increases, since a larger value of $\mathcal{P}_{d,k,j}$ yields a larger value of SNR for the same channel fading h_k. We can also observe from Fig. 6 that for the same transmit power, ϵ-effective capacity increases as the fading parameter m increases, because a larger m represents a better channel quality. For each value of m and $\mathcal{P}_{d,k,j}$, ϵ-effective capacity monotonically decreases as the delay QoS exponent θ_d increases. This is because θ_d indicates the stringency of the statistical delay QoS, and thus, a channel with a less stringent delay QoS requirement can support a larger data arrival rate.

Figure 7 shows the ϵ-effective capacity $EC_k(\theta_d,\epsilon_d,\mathcal{P}_{d,k,j})$ under different values of the delay QoS exponent θ_d and decoding error probability ϵ_d. Similar to Fig. 6, the ϵ-effective capacity is a monotonically decreasing function of the delay QoS exponent θ_d. Figure 7 reveals that the ϵ-effective capacity is also a monotonically decreasing function of the decoding error probability ϵ_d. This is because a smaller decoding error probability indicates a better channel quality and a larger achievable data rate, which yield a larger ϵ-effective capacity.

VI. Conclusion

As metaverse streaming in 6G wireless networks is expected to demand stringent QoS provisionings on delay and decoding error probability and will need to be transmitted among massive MUs, we have proposed to use the mURLLC technique to support metaverse traffic, by integrating massive MIMO, FBC, statistical QoS theory, and NFV/SDN architectures. To estimate the MU's future metaverse data request probability profile, we have proposed a Neyman-Pearson hypothesis testing based human-centric data prediction scheme and have shown that the stopping time for the hypothesis testing is bounded and converges. According to the estimated data request probability profile, we have proposed to dynamically map MUs that request the same set of metaverse data items into the same virtual network slice using NFV/SDN architectures and have derived optimal transmit power allocations to maximize the average aggregate ϵ-effective capacity, which guarantees both statistical delay and error-rate bounded QoS, for this virtual network slice.

Appendix A
Proof of Theorem 1

Proof: We proceed with the proof by showing **Claim 1**, **Claim 2**, and **Claim 3**, respectively. For presentation convenience, we prove **Claim 2** first.

Claim 2: We construct the functions: $y(\cdot)$ and $\Phi_{j,i}(\cdot)$, respectively, as follows:

$$\begin{cases} Y_q = y(X_q) \triangleq \log X_q - \frac{1}{t}\mathbb{E}_{P_j}[\log X_1] \\ \Phi_{j,i}(x) \triangleq (r_i - r_j)x + \frac{t-1}{t}M_{j,i} \end{cases} \tag{63}$$

where $M_{j,i}$ is defined in Eq. (17) and $\mathbb{E}_{P_j}[\cdot]$ is defined in the text following Eq. (13). Thus, we can derive $t\Phi_{j,i}(\overline{Y}_t)$ given in Eq. (14) as follows [23, Eqs. (3.2) and (3.9)]:

$$\begin{aligned} t\Phi_{j,i}(\overline{Y}_t) &= t(r_i - r_j)\overline{Y}_t + (t-1)M_{j,i} \\ &= (r_i - r_j)\left[\left(\sum_{q=1}^{t}\log X_q\right) - \mathbb{E}_{P_j}[\log X_1]\right] \\ &\quad + (t-1)M_{j,i}. \end{aligned} \tag{64}$$

Substituting Eqs. (16) and (64) into $\left|l_t(j,i) - t\Phi_{j,i}(\overline{Y}_t)\right|$ of Eq. (14), we obtain

$$\left|l_t(j,i) - t\Phi_{j,i}(\overline{Y}_t)\right| = \left|\mathbb{E}_{P_j}[(r_i - r_j)\log X_1 + M_{j,i}]\right|. \tag{65}$$

Comparing Eq. (65) with Eq. (14), we can show that Eq. (14) holds true which is detailed as follows:

$$I_{j,i} = \left|\mathbb{E}_{P_j}\left[(r_i - r_j)\log X_1 + M_{j,i}\right]\right| \overset{(b)}{=} \left|\mathbb{E}_{P_j}\left[\log \frac{f_{r_j}(X_1)}{f_{r_i}(X_1)}\right]\right| \tag{66}$$

where (b) follows by using Eq. (17) and Eq. (2). Thus, Eq. (66) shows that Eq. (15) holds true, implying that $I_{j,i}$ is the Kullback-Leibler divergence between P_j and P_i, regardless whatever neighborhood $\mathcal{N}$ is chosen, which completes the proof for **Claim 2**.

Claim 1: Using Eq. (63), we can obtain ε as follows:

$$\begin{aligned}\varepsilon = \mathbb{E}[Y_1] &= \mathbb{E}_{P_j}\left[\log X_1 - \frac{1}{t}\mathbb{E}_{P_j}[\log X_1]\right] \\ &= \frac{t-1}{t}\sum_{d=1}^{D}\frac{(\log d)d^{-r_j}}{\sum_{d=1}^{D} d^{-r_j}}\end{aligned} \tag{67}$$

and thus **Claim 1** holds.

Claim 3: Substituting Eq. (67) into Eq. (63), we can derive $\Phi_{j,i}(\varepsilon)$ as follows:

$$\Phi_{j,i}(\varepsilon) = \frac{t-1}{t}\left[\frac{(r_i - r_j)\sum_{d=1}^{D}[(\log d)d^{-r_j}]}{\sum_{d=1}^{D} d^{-r_j}} + M_{j,i}\right] \tag{68}$$

If r_i and r_j satisfy the following condition:

$$\frac{\sum_{d=1}^{D}[(\log d)d^{-r_j}]}{\sum_{d=1}^{D} d^{-r_j}} \neq -\frac{M_{j,i}}{r_i - r_j} \tag{69}$$

and also applying the condition of Eq. (69) into Eq. (68), then we have $\Phi_{j,i}(\varepsilon) \neq 0$ for **Claim 3**. On the other hand, if $\Phi_{j,i}(\varepsilon) = 0$, using Eq. (63), we can obtain the following:

$$\frac{\partial \Phi_{j,i}(Y_1 - \varepsilon)}{\partial X_1} = (r_i - r_j)\frac{1}{X_1}. \tag{70}$$

Thus, Eq. (70) implies that $\Pr\{(r_i - r_j)(1/X_1) = 0\} < 1$, which is equivalent to Eq. (18). Then, since $r_j \neq r_i$, $X_1 \in \{1, 2, \ldots, D\}$, and $D \leq \infty$, we obtain Eq. (18), completing the proof of **Claim 3**. Therefore, the proof for Theorem 1 follows. ■

Appendix B
Proof of Theorem 2

Proof: According to [23, Theorem 2.1], Eq. (19) follows due to **Claim 1**, **Claim 2**, and **Claim 3** stated in Theorem 1. To show Eq. (20), we need first to show that there exists a $\varsigma > 0$ such that:

$$\Pr\left\{l_{t+1}(j,i) > l^{\text{up}} | X_1, \ldots, X_t, l^{\text{low}} < l_t(j,i) < l^{\text{up}}\right\} \geq \varsigma \tag{71}$$

which is equivalent to

$$\Pr\left\{l_{t+1}(j,i) - l_t(j,i) > \Delta | X_1, \ldots, X_t, l^{\text{low}} < l_t(j,i) < l^{\text{up}}\right\} \geq \varsigma \tag{72}$$

where $\Delta = l^{\text{up}} - l^{\text{low}}$. We derive $l_{t+1}(j,i) - l_t(j,i)$ as

$$l_{t+1}(j,i) - l_t(j,i) = (r_i - r_j)\log X_{t+1} + M_{j,i} \tag{73}$$

when $0 < r_j < r_i < 1$ or $r_j > r_i > 1$. Therefore, we have $l_{t+1}(j,i) - l_t(j,i) > 0$, and then, there exists a $\varsigma > 0$ such that

$$\Pr\left\{l_{t+1}(j,i) - l_t(j,i) > \Delta\right\} \geq \varsigma. \tag{74}$$

Thus, Eq. (72) holds. Based on [23, pp. 1864], Eq. (72) implies that Eq. (20) holds for every $-\infty < l^{\text{low}} < l^{\text{up}} < \infty$, completing the proof of Theorem 2. ■

Appendix C
Proof of Theorem 3

Proof: Using the FBC scheme, $R\left(\gamma_k(\mathcal{P}_{d,k,j})\right)$ in Eq. (9) is given by [12, Eq. (1)]

$$\begin{aligned}&R(\gamma_k(\mathcal{P}_{d,k,j})) \\ &\quad= \log_2\left(1 + \gamma_k(\mathcal{P}_{d,k,j})\right) - \sqrt{\frac{V(\gamma_k(\mathcal{P}_{d,k,j}))}{n}}Q^{-1}(\epsilon_d) \\ &\quad= \log_2\left(\frac{1 + \gamma_k(\mathcal{P}_{d,k,j})}{2^{\widetilde{\epsilon}_d\sqrt{V(\gamma_k(\mathcal{P}_{d,k,j}))}}}\right)\end{aligned} \tag{75}$$

where $\widetilde{\epsilon}_d$ is defined in the text following Eq. (44). To simplify the notation, we replace $\gamma_k(\mathcal{P}_{d,k,j})$ by γ_k in this proof. We derive $\mathbb{E}_{\gamma_k}\left[e^{-\theta_d n R(\gamma_k)}\right]$ in Eq. (9) as follows:

$$\mathbb{E}_{\gamma_k}\left[e^{-\theta_d n R(\gamma_k)}\right] = \mathbb{E}_{\gamma_k}\left[\left(\frac{1+\gamma_k}{2^{\widetilde{\epsilon}_d\sqrt{V(\gamma_k)}}}\right)^{-(\log_2 e)\theta_d n}\right]. \tag{76}$$

Using Taylor-series expansion over $\sqrt{V(\gamma_k)}$, we obtain [36, Eq. (34)]

$$\sqrt{V(\gamma_k)} = -\sum_{i=0}^{\infty} B_i(1+\gamma_k)^{-2i} \tag{77}$$

where $B_0 = -1$ and

$$B_i = \left|\binom{\frac{1}{2}}{i}\right| = \left|\frac{\left(\frac{1}{2}\right)\left(\frac{1}{2}-1\right)\cdots\left(\frac{1}{2}-i+1\right)}{i!}\right|, \ \forall i \geq 1. \tag{78}$$

Defining $\widetilde{\theta}_d \triangleq (\log_2 e)\theta_d n$ and substituting Eq. (77) into Eq. (76), Eq. (76) can be rewritten as

$$\begin{aligned}&\mathbb{E}_{\gamma_k}\left[\left(\frac{1+\gamma_k}{2^{\widetilde{\epsilon}_d\sum_{i=0}^{\infty} B_i(1+\gamma_k)^{-2i}}}\right)^{-\widetilde{\theta}_d}\right] \\ &= \int (1+\gamma_k)^{-\widetilde{\theta}_d} 2^{-\widetilde{\theta}_d\widetilde{\epsilon}_d\sum_{i=0}^{\infty}\left[B_i(1+\gamma_k)^{-2i}\right]} P_\Gamma(\gamma_k)d\gamma_k \\ &\overset{(c)}{=} \sum_{j=0}^{\infty}\frac{(-\widetilde{\theta}_d\widetilde{\epsilon}_d\log 2)^j}{j!}\int (1+\gamma_k)^{-\widetilde{\theta}_d}\left(\sum_{i=0}^{\infty}\left[B_i(1+\gamma_k)^{-2i}\right]\right)^j \\ &\quad\times P_\Gamma(\gamma_k)d\gamma_k \\ &= \sum_{j=0}^{\infty}\frac{(-\widetilde{\theta}_d\widetilde{\epsilon}_d\log 2)^j}{j!}\int\left(\sum_{i=0}^{\infty}\left[B_i(1+\gamma_k)^{-\left(2i+\frac{\widetilde{\theta}_d}{j}\right)}\right]\right)^j \\ &\quad\times P_\Gamma(\gamma_k)d\gamma_k\end{aligned} \tag{79}$$

where $P_\Gamma(\gamma_k)$ is given by Eq. (43), (c) holds by using Taylor-series for $2^x, \forall x$. Defining $\Omega_j \triangleq \sum_{i=0}^{\infty} H_i$, where

$$H_i \triangleq B_i(1+\gamma_k)^{-\left(2i+\frac{\widetilde{\theta}_d}{j}\right)}, \tag{80}$$

the jth moment of Ω_j, denoted by $\mathbb{E}\left[(\Omega_j)^j\right]$, can be derived using the jth order derivative of the moment generating function of Ω_j, denoted by $\mathbb{E}\left[e^{-s\Omega_j}\right]$, with parameter s [37, Eq. (4.89)]. We derive $\mathbb{E}\left[e^{-s\Omega_j}\right]$ as follows:

$$\mathbb{E}\left[e^{-s\Omega_j}\right]=\mathbb{E}\left[e^{-s\sum_{i=0}^{\infty}H_i}\right]=\sum_{\mu=0}^{\infty}\frac{(-s)^{\mu}}{\mu!}\mathbb{E}\left[\left(\sum_{i=0}^{\infty}H_i\right)^{\mu}\right]. \tag{81}$$

We then expand $\left(\sum_{i=0}^{\infty}H_i\right)^{\mu}$ in Eq. (81) using the multinomial theorem as follows:

$$\left(\sum_{i=0}^{\infty}H_i\right)^{\mu}=\sum_{a_1+a_2+\cdots+a_\imath=\mu}\binom{\mu}{a_1,a_2,\cdots,a_\imath}\prod_{\jmath=1}^{\imath}H_{\jmath}^{a_\jmath} \tag{82}$$

where $\{a_1,a_2,\cdots,a_\imath\}$ are all combinations of nonnegative integers such that the sum of all $a_\jmath,\forall\jmath\in\{1,2,\cdots,\imath\}$, is μ, and

$$\binom{\mu}{a_1,a_2,\cdots,a_\imath}=\frac{\mu!}{a_1!a_2!\cdots a_\imath!}. \tag{83}$$

Substituting Eq. (82) into Eq. (81), we can derive $\mathbb{E}\left[\left(\sum_{i=0}^{\infty}H_i\right)^{\mu}\right]$ in Eq. (81) as follows:

$$\mathbb{E}\left[\left(\sum_{i=0}^{\infty}H_i\right)^{\mu}\right]=\sum_{a_1+\cdots+a_\imath=\mu}\binom{\mu}{a_1,a_2,\cdots,a_\imath}\times\prod_{i}B_i\mathbb{E}_{\gamma_k}\left[(1+\gamma_k)^{\ell}\right]. \tag{84}$$

where $\ell\triangleq-(2+\frac{\widetilde{\theta}_d}{\jmath})-(4+\frac{\widetilde{\theta}_d}{\jmath})-\cdots$. Using the binomial theorem, we further derive $\mathbb{E}_{\gamma_k}\left[(1+\gamma_k)^{\ell}\right]$ in Eq. (84) as follows:

$$\mathbb{E}_{\gamma_k}\left[(1+\gamma_k)^{\ell}\right]=\sum_{v=0}^{\infty}\binom{\ell}{v}\mathbb{E}_{\gamma_k}\left[(\gamma_k)^v\right]\overset{\text{(d)}}{\approx}\sum_{v=0}^{\infty}\binom{\ell}{v}\overline{\gamma}_k^{v}=(1+\overline{\gamma}_k)^{\ell} \tag{85}$$

where (d) is obtained by applying Eq. (43) into the derivation of $\mathbb{E}_{\gamma_k}\left[(\gamma_k)^v\right]$ as follows:

$$\begin{aligned}&\mathbb{E}_{\gamma_k}\left[(\gamma_k)^v\right]\\&=\int_0^{\infty}(\gamma_k)^v\frac{\gamma_k^{m-1}}{\Gamma(m)}\left(\frac{m}{\overline{\gamma}_k}\right)^m\exp\left(-\frac{m}{\overline{\gamma}_k}\gamma_k\right)d\gamma_k\\&=\left(\frac{m}{\overline{\gamma}_k}\right)^{-v}\frac{1}{\Gamma(m)}\int_0^{\infty}\left(\frac{m}{\overline{\gamma}_k}\gamma_k\right)^{m+v-1}\exp\left(-\frac{m}{\overline{\gamma}_k}\gamma_k\right)d\left(\frac{m}{\overline{\gamma}_k}\gamma_k\right)\\&=\left(\frac{m}{\overline{\gamma}_k}\right)^{-v}\frac{\Gamma(m+v)}{\Gamma(m)}\overset{\text{(e)}}{\approx}\left(\frac{m}{\overline{\gamma}_k}\right)^{-v}\frac{m^v\Gamma(m)}{\Gamma(m)}=\overline{\gamma}_k^{v}\end{aligned} \tag{86}$$

where (e) holds true when $m>1$. Applying the results of Eq. (85) into Eq. (84), we further derive Eq. (84) as

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=0}^{\infty}H_i\right)^{\mu}\right]&\approx\sum_{a_1+\cdots+a_\imath=\mu}\binom{\mu}{a_1,a_2,\cdots,a_\imath}\prod_{i}B_i(1+\overline{\gamma}_k)^{\ell}\\&=\sum_{i=0}^{\infty}\left[\mathbb{E}\left(H_i\right)^{\mu}\right].\end{aligned} \tag{87}$$

Substituting Eq. (87) into Eq. (84) and then substituting Eq. (84) into Eq. (81), we further derive Eq. (81) as follows:

$$\mathbb{E}\left[e^{-s\Omega_j}\right]\approx e^{-s\sum_{i=0}^{\infty}\mathbb{E}[H_i]}=e^{-s\sum_{i=0}^{\infty}\left[B_i(1+\overline{\gamma}_k)^{-\left(2i+\frac{\widetilde{\theta}_d}{\jmath}\right)}\right]} \tag{88}$$

and thus we can derive the integral in Eq. (79) as

$$\begin{aligned}&\int\left(\sum_{i=0}^{\infty}\left[B_i(1+\gamma_k)^{-\left(2i+\frac{\widetilde{\theta}_d}{\jmath}\right)}\right]\right)^{\jmath}P_{\Gamma}(\gamma_k)d\gamma_k\\&=\mathbb{E}\left[(\Omega_j)^j\right]=(-1)^j\left.\frac{\partial^j\left(\mathbb{E}\left[e^{-s\Omega_j}\right]\right)}{\partial s^j}\right|_{s=0}\\&=\left(\sum_{i=0}^{\infty}\left[B_i(1+\overline{\gamma}_k)^{-\left(2i+\frac{\widetilde{\theta}_d}{\jmath}\right)}\right]\right)^{\jmath}.\end{aligned} \tag{89}$$

Substituting Eq. (89) into Eq. (79) and then substituting Eq. (79) into Eq. (76), we have

$$\begin{aligned}&\mathbb{E}_{\gamma_k}\left[e^{-\theta_d nR(\gamma_k)}\right]\\&=\mathbb{E}_{\gamma_k}\left[\left(\frac{1+\gamma_k}{2^{\widetilde{\epsilon}_d\sum_{i=0}^{\infty}B_i(1+\gamma_k)^{-2i}}}\right)^{-\widetilde{\theta}_d}\right]\\&=\sum_{\jmath=0}^{\infty}\frac{(-\widetilde{\theta}_d\widetilde{\epsilon}_d\log 2)^{\jmath}}{\jmath!}\left(\sum_{i=0}^{\infty}\left[B_i(1+\overline{\gamma}_k)^{-\left(2i+\frac{\widetilde{\theta}_d}{\jmath}\right)}\right]\right)^{\jmath}\\&=(1+\overline{\gamma}_k)^{-\widetilde{\theta}_d}2^{-\widetilde{\theta}_d\widetilde{\epsilon}_d\sum_{i=0}^{\infty}B_i(1+\gamma_k)^{-2i}}=\left(\frac{1+\overline{\gamma}_k}{2^{\widetilde{\epsilon}_d}\sqrt{V(\overline{\gamma}_k)}}\right)^{-\widetilde{\theta}_d},\end{aligned} \tag{90}$$

which yields the main term in Eq. (44), completing the proof of Theorem 3. ■

APPENDIX D
PROOF OF THEOREM 4

Proof: In order to obtain a closed-form expression for $\gamma_{k,(l)}^{\text{MIMO}}(\mathcal{P}_{d,k,j})$, we further derive the numerator of Eq. (52) as

$$\begin{aligned}&\text{Var}\left[\sqrt{\mathcal{P}_{d,k,j}}b_k\sum_{t=1}^{M_T}g_{k,\nu}^{(l)}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}\widehat{g}_{k,\nu}^{(i)*}\right]\\&=\mathcal{P}_{d,k,j}\mathbb{E}\left[\left|\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}g_{k,\nu}^{(l)}\sqrt{\eta_{k,\nu}^{(i)}}\widehat{g}_{k,\nu}^{(i)*}\right|^2\right]\\&\overset{\text{(f)}}{=}\mathcal{P}_{d,k,j}\left(\frac{\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}{1+\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}\right)^2\mathbb{E}\left[\left|\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}g_{k,\nu}^{(l)}g_{k,\nu}^{(i)*}\right|^2\right]\\&+\mathcal{P}_{d,k,j}\left(\frac{\sqrt{\tau_{\text{ul,p}}\rho_{\text{ul}}}\beta_k}{1+\tau_{\text{ul,p}}\rho_{\text{ul}}\beta_k}\right)^2\mathbb{E}\left[\left|\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}g_{k,\nu}^{(l)}\overline{w}_{\text{p}}^{(i)*}\right|^2\right]\end{aligned} \tag{91}$$

where $\overline{w}_{\text{p}}^{(i)}$ is the ith element of $\overline{\mathbf{w}}^{(\text{p})}$, and (f) follows by applying Eq. (49) and using the identity of $\mathbb{E}[|X+Y|^2]=$

$\mathbb{E}[|X|^2]+\mathbb{E}[|Y|^2]$ when X and Y are two independent random variables and $\mathbb{E}[Y]=0$. We further derive

$$\mathbb{E}\left[\left|\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}g_{k,\nu}^{(l)}g_{k,\nu}^{(i)*}\right|^2\right]=\beta_k^2\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\mathbb{E}\left[\left(h_{k,\nu}^{(l)}\right)^2\eta_{k,\nu}^{(i)}\left(h_{k,\nu}^{(i)}\right)^2\right]$$

$$+\beta_k^2\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\mathbb{E}\left[h_{k,\nu}^{(l)}\sqrt{\eta_{k,\nu}^{(i)}}h_{k,\nu}^{(i)}\sum_{(q,p)\neq(t,i)}h_{k,q}^{(l)}\sqrt{\eta_{k,q}^{(p)}}h_{k,q}^{(p)}\right]. \quad (92)$$

According to Section IV-A that the small-scale fading follows the Nakagami-m fading, we have $\mathbb{E}\left[h_{k,\nu}^{(i)}\right]=\overline{h}$ and $\mathbb{E}\left[\left(h_{k,\nu}^{(i)}\right)^2\right]=\overline{h}^2$. We further derive Eq. (92) by using $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are uncorrelated random variables as follows:

$$\mathbb{E}\left[\left|\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}g_{k,\nu}^{(l)}g_{k,\nu}^{(i)*}\right|^2\right]=\beta_k^2\overline{\eta}M_T^2M_R^2\overline{h}^4. \quad (93)$$

Similarly, we also derive the expectation in the last part of Eq. (91) as

$$\mathbb{E}\left[\left|\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}g_{k,\nu}^{(l)}\overline{w}_{\mathrm{p}}^{(i)*}\right|^2\right]=\beta_k\overline{\eta}\sum_{t=1}^{M_T}\sum_{i=1}^{M_R}\mathbb{E}\left[\left(h_{k,\nu}^{(l)}\right)^2\right]$$
$$=\beta_k\overline{\eta}M_TM_R\overline{h}^2. \quad (94)$$

Thus, substituting Eq. (93) and Eq. (94) into Eq. (91), we can rewrite Eq. (91) as follows:

$$\mathrm{Var}\left[\sqrt{\mathcal{P}_{d,k,j}}b_k\sum_{t=1}^{M_T}g_{k,\nu}^{(l)}\sum_{i=1}^{M_R}\sqrt{\eta_{k,\nu}^{(i)}}\widehat{g}_{k,\nu}^{(i)*}\right]=\mathcal{P}_{d,k,j}N_{k,1} \quad (95)$$

where $N_{k,1}$ is given by Eq. (54). We also derive the first term of the denominator in Eq. (52) as

$$\mathrm{Var}\left[\sum_{t=1}^{M_T}\sum_{u=1,u\neq k}^{K_d}\sqrt{\mathcal{P}_{d,u,j}}g_{k,\nu}^{(l)}\left(\boldsymbol{\eta}_{u,t}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,t}^*b_u\right]$$

$$=\left(\sum_{u=1,u\neq k}^{k_d}\mathcal{P}_{d,u,j}\right)\overline{\eta}M_TM_R\mathbb{E}\left[\left(g_{k,\nu}^{(l)}\right)^2\left(\widehat{g}_{u,t}^{(i)*}\right)^2\right]$$

$$+\left(\sum_{u=1,u\neq k}^{k_d}\mathcal{P}_{d,u,j}\right)\overline{\eta}M_T^2M_R(M_R-1)\mathbb{E}\left[\left(g_{k,\nu}^{(l)}\right)^2\right]\left(\mathbb{E}\left[\widehat{g}_{u,t}^{(i)*}\right]\right)^2$$

$$=(P_{\max}-\mathcal{P}_{d,k,j})\overline{\eta}\beta_k\overline{h}^2M_TM_R\left\{\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}(\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\overline{h}^2+1)\right.$$

$$\left.\times\mathbb{E}\left[\frac{\beta_u^2}{(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_u)^2}\right]+M_T(M_R-1)\left(\mathbb{E}\left[\frac{\beta_u^{\frac{3}{2}}}{1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_u}\right]\right)^2\right\}. \quad (96)$$

Using the assumption that MUs are uniformly distributed within a wireless cell with inner radius $R_{\min}$ and outer radius $R_{\max}$ and defining the random variable of an MU's distance to the BS/AP as R, the pdf of the distance R, denoted by $p_R(r)$, is given by:

$$p_R(r)=\frac{2r}{R_{\max}^2-R_{\min}^2}. \quad (97)$$

Defining $X\triangleq\beta_u/(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_u)$, we derive the cumulative distribution function (cdf) of X, denoted by $P_X(x)$, as:

$$P_X(x)=\Pr\left\{\frac{\lambda^2}{(4\pi)^2\left(R^2+\iota^2\right)+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\lambda^2}\leq x\right\}$$
$$\overset{\text{(g)}}{=}1-\frac{1}{R_{\max}^2-R_{\min}^2}\left(\frac{\lambda^2(1-\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}x)}{(4\pi)^2x}-\iota^2\right) \quad (98)$$

where (g) is obtained by using Eq. (97) and ι is the height of a BS/AP. Then, using Eq. (98), we have the pdf of X, denoted by $p_X(x)$, as follows:

$$p_X(x)=\frac{\partial P_X(x)}{\partial x}=\frac{\lambda^2}{16\pi^2x^2(R_{\max}^2-R_{\min}^2)}. \quad (99)$$

Therefore, we have:

$$\mathbb{E}\left[\frac{\beta_u^2}{(1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_u)^2}\right]=\mathbb{E}\left[X^2\right]=\int_{X_{\min}}^{X_{\max}}x^2p_X(x)dx$$
$$=\frac{\lambda^2}{16\pi^2(R_{\max}^2-R_{\min}^2)}\left(X_{\max}-X_{\min}\right) \quad (100)$$

where $X_{\max}$ and $X_{\min}$ are given by Eq. (56). Similarly, we also have

$$\mathbb{E}\left[\frac{\beta_u^{\frac{3}{2}}}{1+\tau_{\mathrm{ul,p}}\rho_{\mathrm{ul}}\beta_u}\right]\approx\mathbb{E}[X]=\int_{X_{\min}}^{X_{\max}}xp_X(x)dx$$
$$=\frac{\lambda^2}{16\pi^2(R_{\max}^2-R_{\min}^2)}\log\left(\frac{X_{\max}}{X_{\min}}\right). \quad (101)$$

Substituting Eq. (100) and Eq. (101) into Eq. (96), we further derive Eq. (96) as

$$\mathrm{Var}\left[\sum_{t=1}^{M_T}\sum_{u=1,u\neq k}^{K}\sqrt{\mathcal{P}_{d,u,j}}g_{k,\nu}^{(l)}\left(\boldsymbol{\eta}_{u,t}\right)^{\frac{1}{2}}\widehat{\mathbf{g}}_{u,t}^*b_u\right]$$
$$=(P_{\max}-\mathcal{P}_{d,k,j})N_{k,2}, \quad (102)$$

where $N_{k,2}$ is given by Eq. (55). Substituting Eq. (95) and Eq. (102) into Eq. (52), we obtain Eq. (53), completing the proof for Theorem 4. ■

APPENDIX E
PROOF OF THEOREM 5

Proof: We proceed with the proof by showing the **Claim 1** and **Claim 2**, respectively.

Claim 1: Denote the received SNR of the kth MU on all antennas by the vector $\boldsymbol{\gamma}_k^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})\triangleq\left[\gamma_{k,(1)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j}),\cdots,\gamma_{k,(M_R)}^{\mathrm{MIMO}}(\mathcal{P}_{d,k,j})\right]$. First, extending the derivations for the data rate given in Eq. (75) over the single antenna channel into its massive-MIMO-channel version, we can obtain the data rate for the massive-MIMO

channel, denoted by $R(\boldsymbol{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))$, which is a function of the SNR vector $\boldsymbol{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})$, as follows:

$$R\left(\boldsymbol{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})\right) = \sum_{l=1}^{M_R} \log_2\left(\frac{1+\gamma_{k,(l)}^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_d}{\sqrt{V(\gamma_{k,(l)}^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}\right) \overset{\text{(h)}}{=} M_R \log_2\left(\frac{1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_d}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}\right) \tag{103}$$

where (h) is due to Eq. (57). Replacing $R(\gamma_k)$ in Eq. (90) by $R\left(\boldsymbol{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})\right)$ derived in Eq. (103), we can obtain

$$\mathbb{E}_{\gamma_k}\left[e^{-\theta_d n R\left(\boldsymbol{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})\right)}\right] = \left(\frac{1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_d}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}\right)^{-\widetilde{\theta}_d M_R}. \tag{104}$$

Replacing $\mathbb{E}_{\gamma_k}\left[e^{-\theta_d n R(\gamma_k(\mathcal{P}_{d,k,j}))}\right]$ in Eq. (9) by the expression in Eq. (104), we can obtain a closed-form expression for ϵ-effective capacity of the kth MU $EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ over the massive-MIMO channel as follows

$$EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}) = -\frac{1}{n\theta_d}\left\{\log\left[\epsilon_d + (1-\epsilon_d)\left(\frac{1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_d}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}\right)^{-\widetilde{\theta}_d M_R}\right]\right\} \tag{105}$$

Employing the optimal transmit power allocation $\mathcal{P}^*_{d,k,j}$ for Eq. (105), we obtain Eq. (59). Second, taking the summation $\sum_{k=1}^{K_d} EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ for all MUs, which request the same data item d, and using the estimated data request pmf $f_{r_j}(d)$ of each metaverse data item d by the selected optimal hypothesis $\mathcal{H}_j$, we obtain the average for the aggregate ϵ-effective capacity given by

$$\sum_{d=1}^{D} f_{r_j}(d) \sum_{k=1}^{K_d} EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j}) = \sum_{d=1}^{D} f_{r_j}(d) \sum_{k=1}^{K_d} \left[-\frac{1}{n\theta_d} \times \left\{\log\left[\epsilon_d + (1-\epsilon_d)\left(\frac{1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_d}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}\right)^{-\widetilde{\theta}_d M_R}\right]\right\}\right]. \tag{106}$$

The optimization for the aggregate ϵ-effective capacity given by Eq. (106) is equivalent to individually optimizing each term $f_{r_j}(d)\sum_{k=1}^{K_d} EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ in Eq. (106). Moreover, observing Eq. (44), we obtain that $EC_k^{\text{MIMO}}(\theta_d, \epsilon_d, \mathcal{P}_{d,k,j})$ is a monotonically increasing function of $(1+\overline{\gamma}_k)/2^{\frac{\widetilde{\epsilon}_k}{\sqrt{V(\overline{\gamma}_k)}}}$. Thus, we can convert the optimization problem in Eq. (58) into the following optimization problem:

$$\left(\mathcal{P}^*_{d,1,j}, \cdots, \mathcal{P}^*_{d,K_d,j}\right) = \arg \max_{\left(\mathcal{P}_{d,1,j}, \cdots, \mathcal{P}_{d,K_d,j}\right)} f_{r_j}(d) \sum_{k=1}^{K_d} \frac{1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_k}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}, \quad \forall d$$

$$\textbf{s.t.: C1: } \sum_{k=1}^{K_d} \frac{\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})(1+N_{k,2}P_{\max})}{N_{k,1}+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})N_{k,2}} \le P_{\max},$$

$$\textbf{C2: } \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}) \ge 0, \ \forall k. \tag{107}$$

To solve Eq. (107), we can formulate a Lagrangian function $\mathcal{L}$ as follows:

$$\mathcal{L} = f_{r_j}(d)\frac{1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})}{2^{\frac{\widetilde{\epsilon}_k}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}} - \varrho_\gamma\left(\sum_{k=1}^{K_d}\frac{\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})(1+N_{k,2}P_{\max})}{N_{k,1}+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})N_{k,2}} - P_{\max}\right) \tag{108}$$

where ϱ_γ is the Lagrangian multiplier for the constraint C1 of Eq. (107). Then, using Karush-Kuhn-Tucker (KKT) conditions, we can get the following equations, respectively:

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})} = \dfrac{f_{r_j}(d)}{2^{\frac{\widetilde{\epsilon}_k}{\sqrt{V(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}))}}}}\Big[1-(\log 2)\widetilde{\epsilon}_k \times\left(1+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})\right)^{-2}\left[V\left(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})\right)\right]^{-\frac{1}{2}}\Big] \\ \qquad -\dfrac{\varrho_\gamma N_{k,1}(1+N_{k,2}P_{\max})}{\left[N_{k,1}+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})N_{k,2}\right]^2} = 0, \ \forall k, \quad (109) \\ \sum_{k=1}^{K_d}\dfrac{\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})(1+N_{k,2}P_{\max})}{N_{k,1}+\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j})N_{k,2}} - P_{\max} = 0, \quad (110) \\ \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}_{d,k,j}) \ge 0, \ \forall k. \quad (111) \end{cases}$$

Solving Eq. (109), we can obtain the average SNR $\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j})$ under the optimal transmit power allocation $\mathcal{P}^*_{d,k,j}$. Depending on the average SNR $\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j})$ falling in the high-regime or low-regime, we need to consider the following two cases, respectively:

Case 1. If $\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j}) \gg 1$, then we have $V\left(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j})\right) \approx 1$ and by solving Eq. (109), we can obtain

$$\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j}) \approx \frac{1}{N_{k,2}}\left[\frac{\varrho^*_{\gamma,1}2^{\widetilde{\epsilon}_k}N_{k,1}}{f_{r_j}(d)}(1+N_{k,2}P_{\max})\right]^{\frac{1}{2}} - \frac{N_{k,1}}{N_{k,2}}, \tag{112}$$

where $N_{k,1}$ and $N_{k,2}$ are given in Eqs. (54) and (55), respectively.

Case 2. If $0 < \overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j}) < 1$, then we have $V\left(\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j})\right) \approx 0$ and by solving Eq. (109), we can obtain

$$\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j}) \approx \frac{1}{N_{k,2}}\left[\frac{\varrho^*_{\gamma,2}N_{k,1}}{f_{r_j}(d)}(1+N_{k,2}P_{\max})\right]^{\frac{1}{2}} - \frac{N_{k,1}}{N_{k,2}}, \tag{113}$$

where $N_{k,1}$ and $N_{k,2}$ are given in Eqs. (54) and (55), respectively. Combining Eq. (112) and Eq. (113), Eq. (60) holds. Substituting $\overline{\gamma}_k^{\text{MIMO}}(\mathcal{P}^*_{d,k,j})$ specified by Eq. (112) and Eq. (113), respectively, into Eq. (110), we can obtain Eq. (61), completing the proof for **Claim 1** of Theorem 5.

Claim 2: Substituting Eq. (60) into Eq. (57) and substituting Eq. (57) into Eq. (53), we obtain the optimal power allocation as shown in Eq. (62), completing the proof for **Claim 2**. This completes the proof for Theorem 5. ■

REFERENCES

[1] *3GPP TSG RAN WG1 Meeting #87*, AccelerComm, Reno, NV, USA, Nov. 2016.

[2] F. Chen, P. Li, D. Zeng, and S. Guo, "Edge-assisted short video sharing with guaranteed quality-of-experience," *IEEE Trans. Cloud Comput.*, vol. 11, no. 1, pp. 13–24, Jan. 2023.

[3] X. Zhang and Q. Zhu, "Statistical delay and error-rate bounded QoS provisioning over massive-MIMO based 6G mobile wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2022, pp. 353–358.

[4] X. Zhang, Q. Zhu, and H. V. Poor, "Massive-MIMO channel capacity modeling for mURLLC over 6G UAV mobile wireless networks," in *Proc. 56th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2022, pp. 49–54.

[5] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[6] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 118–129, Jan. 2008.

[7] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[8] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.

[9] H. Su and X. Zhang, "Clustering-based multichannel MAC protocols for QoS provisionings over vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 56, no. 6, pp. 3309–3323, Nov. 2007.

[10] J. Tang and X. Zhang, "Cross-layer resource allocation over wireless relay networks for quality of service provisioning," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 645–656, May 2007.

[11] X. Zhang and Q. Zhu, "AI-enabled network-functions virtualization and software-defined architectures for customized statistical QoS over 6G massive MIMO mobile wireless networks," *IEEE Netw.*, vol. 37, no. 2, pp. 30–37, Mar. 2023.

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

[13] X. Zhang, J. Wang, and H. V. Poor, "AoI-driven statistical delay and error-rate bounded QoS provisioning for mURLLC over UAV-multimedia 6G mobile networks using FBC," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 11, pp. 3425–3443, Nov. 2021.

[14] Y. Jiang et al., "Reliable distributed computing for metaverse: A hierarchical game-theoretic approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1084–1100, Jan. 2023.

[15] D. Van Huynh, S. R. Khosravirad, A. Masaracchia, O. A. Dobre, and T. Q. Duong, "Edge intelligence-based ultra-reliable and low-latency communications for digital twin-enabled metaverse," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1733–1737, Aug. 2022.

[16] H. Du et al., "Attention-aware resource allocation and QoE analysis for metaverse xURLLC services," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2158–2175, Jul. 2023.

[17] N. T. Hoa, L. V. Huy, B. D. Son, N. C. Luong, and D. Niyato, "Dynamic offloading for edge computing-assisted metaverse systems," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1749–1753, Jul. 2023.

[18] W. Feng, S. Lin, N. Zhang, G. Wang, B. Ai, and L. Cai, "Joint C-V2X based offloading and resource allocation in multi-tier vehicular edge computing system," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 2, pp. 432–445, Feb. 2023.

[19] Z. Ding, L. Shen, H. Chen, F. Yan, and N. Ansari, "Energy-efficient topology control mechanism for IoT-oriented software-defined WSNs," *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13138–13154, Aug. 2023.

[20] Y. Zeng et al., "Mobility-aware proactive flow setup in software-defined mobile edge networks," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1549–1563, Mar. 2023.

[21] X. Zhang, Q. Zhu, and H. V. Poor, "Neyman-pearson criterion based optimal hierarchical caching over D2D wireless ad-hoc networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[22] X. Zhang, Q. Zhu, and H. V. Poor, "Sequential hypothesis criterion based optimal caching schemes over mobile wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1254–1258.

[23] R. A. Wijsman, "Exponentially bounded stopping time of sequential probability ratio tests for composite hypotheses," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1859–1869, Dec. 1971.

[24] M. Fauß, A. M. Zoubir, and H. V. Poor, "Minimax optimal sequential tests for multiple hypotheses," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 1044–1046.

[25] T. Leung Lai, "Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 595–608, Mar. 2000.

[26] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer, 1994.

[27] A. Wald, *Sequential Analysis*. London, U.K.: Wiley, 1947.

[28] X. Zhang, J. Tang, H.-H. Chen, S. Ci, and M. Guizani, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Commun. Mag.*, vol. 44, no. 1, pp. 100–106, Jan. 2006.

[29] X. Zhang and Q. Zhu, "Scalable virtualization and offloading-based software-defined architecture for heterogeneous statistical QoS provisioning over 5G multimedia mobile wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 12, pp. 2787–2804, Dec. 2018.

[30] M. Sobel and A. Wald, "A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution," *Ann. Math. Statist.*, vol. 20, no. 4, pp. 502–522, Dec. 1949.

[31] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, Jun. 1945.

[32] T. L. Lai, "Asymptotic optimality of invariant sequential probability ratio tests," *Ann. Statist.*, vol. 9, no. 2, pp. 318–333, Mar. 1981.

[33] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[34] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[35] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Guest editorial special issue on multiple antenna technologies for beyond 5G-Part II," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 1941–1944, Sep. 2020.

[36] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 13–22.

[37] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 3rd ed. London, U.K.: Pearson, 2008.

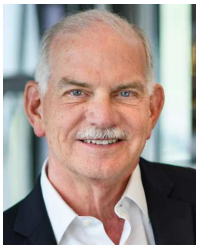
Xi Zhang (Fellow, IEEE) received the B.S. and M.S. degrees from Xidian University, Xi'an, China, the M.S. degree from Lehigh University, Bethlehem, PA, USA, all in electrical engineering and computer science, and the Ph.D. degree in electrical engineering and computer science (Electrical Engineering-Systems) from the University of Michigan, Ann Arbor, MI, USA.

He is currently a Full Professor and the Founding Director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is a Fellow of the IEEE for contributions to quality of service (QoS) theory in mobile wireless networks. He was with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hill, NJ, USA, and AT&T Laboratories Research, Florham Park, NJ, USA, in 1997. He was a

Research Fellow with the School of Electrical Engineering, University of Technology, Sydney Australia, and the Department of Electrical and Computer Engineering, James Cook University, Australia. He has published more than 400 research articles on wireless networks and communications systems, network protocol design and modeling, statistical communications, random signal processing, information theory, and control theory and systems. He received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He received six Best Paper Awards at IEEE GLOBECOM 2020, IEEE ICC 2018, IEEE GLOBECOM 2014, IEEE GLOBECOM 2009, IEEE GLOBECOM 2007, and IEEE WCNC 2010, respectively. One of his IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS papers has been listed as the IEEE Best Readings Paper (receiving the highest citation rate among all IEEE TRANSACTIONS/JOURNAL articles in the area) on wireless cognitive radio networks and statistical QoS provisioning over mobile wireless networking. He is an IEEE Distinguished Lecturer of both IEEE Communications Society and IEEE Vehicular Technology Society. He received the TEES Select Young Faculty Award for Excellence in Research Performance from the College of Engineering, Texas A&M University, in 2006, and the Outstanding Faculty Award from Texas A&M University, in 2020.

Prof. Zhang is serving or has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, twice as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for two special issues on "Broadband Wireless Communications for High Speed Vehicles" and "Wireless Video Transmissions," an Associate Editor for IEEE COMMUNICATIONS LETTERS, twice as the Lead Guest Editor for IEEE *Communications Magazine* for two special issues on "Advances in Cooperative Wireless Networking" and "Underwater Wireless Communications and Networks: Theory and Applications," a Guest Editor for IEEE *Wireless Communications Magazine* for special issue on "Next Generation CDMA vs. OFDMA for 4G Wireless Applications," an Editor for Wiley's *Journal on Wireless Communications and Mobile Computing*, *Journal of Computer Systems, Networking, and Communications*, and Wiley's *Journal on Security and Communications Networks*, and an Area Editor for Elsevier's *Journal on Computer Communications*, among many others. He is serving or has served as the TPC Chair for IEEE GLOBECOM 2011, TPC Vice-Chair for IEEE INFOCOM 2010, TPC Area Chair for IEEE INFOCOM 2012, Panel/Demo/Poster Chair for ACM MobiCom 2011, General Chair for IEEE ICDCS 2024 Workshop on "Digital Twin-Enabled 6G Multi-tier Distributed Computing Systems," General Chair for IEEE WCNC 2013, TPC Chair for IEEE INFOCOM 2017–2019 Workshops on "Integrating Edge Computing, Caching, and Offloading in Next Generation Networks," etc.

Qixuan Zhu received the B.S. degree from Tianjin University of Technology and Education, Tianjin, China, and the M.S. degree from The George Washington University, Washington, DC, USA, all in electrical and computer engineering. She is currently pursuing the Ph.D. degree under the supervision of Professor Xi Zhang in the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. She received the Best Paper Award from IEEE ICC 2018. She also received the Hagler Institute for Advanced Study Heep Graduate Fellowship Award and the Dr. Christa U. Pandey '84 Fellowship from Texas A&M University.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of with the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he was the Dean of the Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests include information theory, machine learning, and network science and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He is a member of the National Academy of Engineering and the National Academy of Sciences and a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.