

MuMeNet: A Network Simulator for Musical Metaverse Communications

*Original*

MuMeNet: A Network Simulator for Musical Metaverse Communications / Al Housseini, Ali; Llorca, Jaime; Turchet, Luca; Leidi, Tiziano; Rottondi, Cristina; Ayoub, Omran. - (2025), pp. 1-10. ( 2025 IEEE 6th International Symposium on the Internet of Sounds (IS2) L'Aquila (Ita) 29-31 October 2025) [10.1109/is264627.2025.11284646].

*Availability:*

This version is available at: 11583/3006484 since: 2026-01-12T20:51:24Z

*Publisher:*

IEEE

*Published*

DOI:10.1109/is264627.2025.11284646

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

IEEE postprint/Author's Accepted Manuscript

©2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

# MuMeNet: A Network Simulator for Musical Metaverse Communications

Ali Al Housseini<sup>\*‡</sup>, Jaime Llorca<sup>‡§</sup>, Luca Turchet<sup>†</sup>, Tiziano Leidi<sup>‡</sup>, Cristina Rottondi<sup>\*</sup>, Omran Ayoub<sup>‡</sup>

<sup>\*</sup> Politecnico di Torino, Turin, Italy, <sup>†</sup> University of Trento, Trento, Italy

<sup>‡</sup> University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland

<sup>§</sup> Centre Tecnologic de Telecomunicacions de Catalunya (CTTC/CERCA), Castelldefels, Spain.

ali.alhousseini@supsi.ch

**Abstract**—The Metaverse, a shared and spatially organized digital continuum, is transforming various industries, with music emerging as a leading use case. Live concerts, collaborative composition, and interactive experiences are driving the Musical Metaverse (MM), but the requirements of the underlying network and service infrastructures hinder its growth. These challenges underscore the need for a novel modeling and simulation paradigm tailored to the unique characteristics of MM sessions, along with specialized service provisioning strategies capable of capturing their interactive, heterogeneous, and multicast-oriented nature. To this end, we make a first attempt to formally model and analyze the problem of service provisioning for MM sessions in 5G/6G networks. We first formalize service and network graph models for the MM, using “live audience interaction in a virtual concert” as a reference scenario. We then present *MuMeNet*, a novel discrete-event network simulator specifically tailored to the requirements and the traffic dynamics of the MM. We showcase the effectiveness of *MuMeNet* by running a linear programming based orchestration policy on the reference scenario and providing performance analysis under realistic MM workloads.

**Index Terms**—Musical Metaverse, Network Simulator, Graph Modeling, Resource Allocation, Service Orchestration.

## I. INTRODUCTION

The *Metaverse* refers to a vision of a shared, spatially organized digital continuum in which individuals experience an embodied presence and interaction across interconnected and heterogeneous virtual, augmented and physical environments [1]. Early deployments of the Metaverse have centered on gaming, retail, and digital twins for industrial monitoring, simulation, and optimization, serving as testbeds for immersive interaction and real-time collaboration applications [2].

More recently, new forms of the Metaverse are emerging, among them, the *Musical Metaverse* (MM), which aims to enable real-time, interactive experiences such as live virtual concerts, collaborative music creation, and shared listening sessions [3]. From a technical perspective, the MM represents an interoperable, persistent network of multi-user environments that merge physical reality with digital reality. It is based upon the convergence of Musical eXtended Reality (XR) [4] and Internet of Musical Things (IoMusT) [5] technologies that enable multisensory, networked musical interactions, not only among musicians, audiences, and performers, but also with virtual environments and intelligent instruments [3].

Realizing such deeply immersive and time-sensitive experiences places unprecedented demands on the network infrastructure. Unlike conventional media applications, MM sessions involve synchronized delivery of high-fidelity audio and video alongside real-time control signals [6], gestural and sensor data [7], and haptic feedback [8]. Achieving musical copresence, i.e., the perception of synchronous musical interplay, requires ultra-low end-to-end latency (in the order of a few milliseconds), near-zero jitter, and bidirectional data exchange [4], [9]. These stringent requirements challenge existing communication frameworks and necessitate a distributed, responsive infrastructure encompassing 5G/6G technologies, edge computing, and adaptive bitrate control [3], [10].

To meet Quality of Service (QoS) and Quality of Experience (QoE) expectations, MM sessions must *adapt* in real time to changing network conditions and explicitly *account* for the concurrent multimodal traffic [3]. These challenges underscore the need for novel modeling and simulation methods tailored to the unique aforementioned characteristics of MM sessions, along with specialized service provisioning and orchestration strategies capable of capturing their real-time, interactive, and multicast-oriented nature [11].

To model, simulate, and analyze the problem of service provisioning in the MM, several aspects should be considered:

- *Scale and Heterogeneity*: A realistic MM session spans thousands of geographically dispersed users, multiple edge tiers, backbone transport, and cloud data centers. Recreating that hierarchy in hardware would demand a large number of servers, equipments, and investments.
- *Observability and repeatability*: Field measurements can reveal end-to-end delays, but make it difficult to expose what happens *inside* each compute node (e.g., buffers); nor can they be replayed under identical load. A simulation environment, by contrast, allows one to have deterministic control over every link, timestamp, and flow, enabling fair comparisons and statistically significant ablation studies.
- *Risk and iteration speed*: Early-stage policies often violate QoS constraints; running them live would produce audible glitches or broken sessions. Simulation environment offers a low-risk sandbox in which unsafe ideas can fail fast and guide the next design cycle.

In this work, we make a first attempt to formally model, analyze, and simulate the problem of service provisioning for MM sessions in 5G/6G networks. To this end, we offer the following contributions:

- We design a graph-based model for MM services that describes the interaction between MM components, and illustrate it for the specific scenario of live audience interaction in a virtual concert;
- We leverage the CNFlow framework to build *service graph* (SG) and *cloud-network graph* (CNG) mathematical models for the MM that reconceptualizes how services are mapped on the network infrastructure;
- We introduce *MuMeNet*, a discrete-event network simulator specifically tailored to the requirements and traffic dynamics of the MM that allows embedding (i.e., allocation of resources) dynamic SGs while analyzing the cross-stream synchronization.
- Using the proposed modeling, we formulate the joint placement and routing of MM SGs onto a CNG as a Mixed-Integer Linear Program (MILP) and demonstrate its effectiveness across diverse network instances and traffic scenarios simulated via MuMeNet.

The remainder of this paper is organized as follows. Section II discusses related works. Section III models the MM scenario. Section IV describes the architecture of MUMENET. Section V presents the MILP. Section VI discusses numerical results, and Section VII concludes the paper.

## II. RELATED WORK AND TECHNICAL GAPS

### A. Network Communication and Resource Allocation in the Metaverse

Early visions depict the Metaverse as the Internet’s next evolutionary stage, enabling users to work, play, and socialize in persistent, fully immersive 3D spaces [12]. Delivering such experiences at scale places unprecedented pressure on telecommunication networks as they must support thousands of concurrent users, stream multi-sensory XR content at gigabit rates, and keep motion-to-photon delays below 20 ms to sustain presence [10]. Tang *et al.* outline a 6G roadmap that combines intelligent sensing, integrated space–air–ground networking and edge computing to achieve the Ultra-Reliable Low-Latency Communication (URLLC) essential for the Metaverse [12]. In [13], authors conducted empirical studies that confirm that even modest latency spikes or packet loss degrade QoE, consequently producing motion lag and user dizziness in networked Virtual Reality (VR). It is thus envisioned that next-generation infrastructures will be re-engineered around 5G/6G slices and edge intelligence to satisfy the extreme bandwidth, latency, and reliability demands of immersive, real-time Metaverse applications. Complementing this network focus, Van *et al.* propose a digital-twin-enabled architecture that offloads compute-intensive tasks (e.g., physics updates, rendering) to multi-access edge servers and caches popular results, jointly optimizing communication, computation, and storage to trim end-to-end delay [14].

Beyond architectural enablers, a significant body of recent research addresses resource allocation algorithms for efficiently managing network and computing resources in Metaverse environments. In [15], the authors introduce a human-centric resource allocation framework for wireless Metaverse applications that explicitly optimizes perceived user utility relative to the cost of resources used. Their system jointly allocates communication bandwidth, computing power (for rendering), and even adjusts the streaming video resolution to maximize a utility-cost ratio, essentially delivering the best experience per unit of cost in energy and latency.

A recent line of work by Chu. *et al.* tackles generic Metaverse resource management through a two-stage evolution. Their main contribution, *MetaSlicing* [16], decomposes each application into fine-grained functions and clusters applications that share those functions into *MetaInstances*, so that one instantiation can be transparently reused across slices.

While such contributions are notable, the comprehensive orchestration of resource-intensive Metaverse applications necessitates novel modeling approaches that can accurately capture their intricate processing graph structures, multimodal flow scaling behavior, and efficient sharing and replication of real-time data streams [11]. For this reason, we leverage the Cloud Network Flow (CNFlow) modeling and optimization framework [11], [17], [18], which offers a robust abstraction for modeling and optimizing the end-to-end orchestration of next-generation media services over distributed cloud-integrated networks.

Despite the rapid progress in research on Metaverse communications and service provisioning, most studies to date validate their proposals via analysis or simulation (e.g., custom network simulators or trace-driven experiments). There is a lack of standardized design and modeling of Metaverse slices, as well as of testbeds and emulation platforms for Metaverse networking. This raises questions about how solutions will perform under real network conditions, with unpredictable user behaviors and interference. Developing realistic prototypes and common simulation benchmarks for Metaverse scenarios remains an open challenge.

### B. Comparison between Existing Network Simulators

*Packet-level network simulators:* Several works extend classical packet simulators to Metaverse use-cases. Wang *et al.* [19] connect a cloud renderer to a HoloLens 2 client in a “Virtual City” built on ns-3 [20], assessing end-to-end delay and throughput for remote AR rendering. Lecci *et al.* [21] contribute an ns-3 XR traffic generator that replays head-movement traces as bursty UDP flows, emulating the high data rate of VR video, while audio support is left as future work. Recent research prototypes even couple network simulators (OMNET++ [22]) with 3D engines (e.g., Unity) to create holistic simulations aimed at analyzing how dense user motion affects 5G links in a virtual event arena. While these attempts enable a faithful evaluation of congestion control and scheduling, their primary limitation is architectural: these tools assume *static* network topologies and treat packets as

homogeneous byte streams, lacking the abstractions necessary to represent SGs or modality-specific latency constraints.

*Cloud/edge simulators:* To explore service placement and resource allocation policies, the research community has widely adopted the CloudSim family of simulators, whose internal models follow the classic Virtual Network Embedding (VNE) abstraction [23]. CloudSim Plus [24] focuses on data-center virtual machine scheduling, while fog/edge variants reuse the same VNE kernel in distributed settings. Although these frameworks excel at graph-layer optimization, they cannot express information-centered replication, mixed-cast flows, or millisecond-level latency variance, capabilities that MM workloads require. In addition, critical network dynamics, such as packet jitter, burst loss, and congestion effects (the primary causes of musical desynchronization), are rendered invisible. As a result, a placement strategy that appears optimal in simulation may prove unusable once the dynamics of a real-world network are considered.

The work in [25] provides an initial solution to incorporate the CNFlow framework into ns-3 and allow service orchestration at different timescales. However, it focused on network service chains, without capturing all the intricate requirements and complex nature of MM sessions.

*MuMeNet* fills this gap by (i) embedding latency-annotated SGs onto a live packet topology, (ii) tracking audio, video, and haptic objects separately to compute inter-stream skew at micro-second resolution, and (iii) injecting bursty audience events that reshape traffic in real time. These capabilities enable faithful evaluation of orchestration algorithms under the exact timing constraints of musical co-presence.

### III. REFERENCE SCENARIO: AUDIENCE INTERACTION IN PRE-RECORDED LIVE CONCERTS

#### A. Reference Scenario Description

We consider a reference MM scenario that entails a VR concert platform for distributed audiences using pre-recorded music and musicians' avatars on stage. The pre-recorded concert encompasses multitrack studio stems (audio), synchronized high-resolution video, lighting cues, and stage metadata, forming a passive immersive scene. Users, represented as audience avatars, can then participate and interact in real time, without any geographical constraints.

We model this scenario as a SG  $\mathcal{M} = (\mathcal{N}, \mathcal{K})$ , where vertices identify the required components for real-time interactive communications in the virtual concert, and edges their interconnections. A component in the SG belongs to one of these sets:

- *Producers* ( $\mathcal{N}^s$ ): The entities generating continuous flows of data;
- *Processors* ( $\mathcal{N}^p$ ): The service functions responsible for processing the data generated by the producers, possibly receiving information also from databases;
- *Consumers* ( $\mathcal{N}^d$ ): The entities receiving and consuming data computed by the processors.

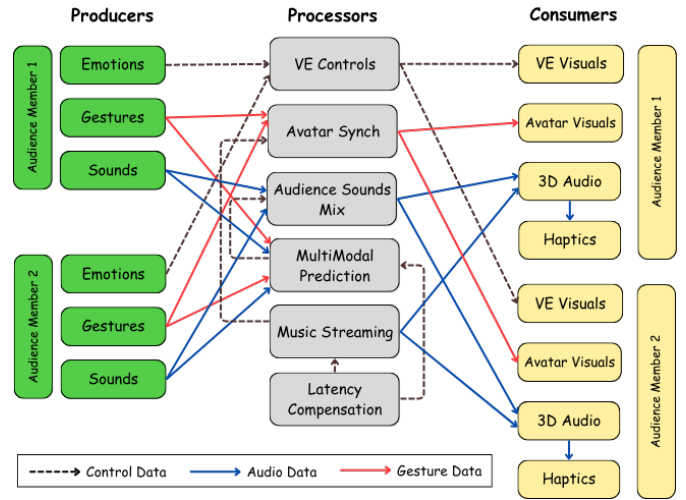


Fig. 1. Graph representation of the modeled components and their connections of the reference SG given two users.

Each user is represented in the graph by its *producer* and *consumer* components. Fig. 1 illustrates the SG for the described scenario, assuming two users.

The producers are the following:

- *Sounds:* This component tracks, via the microphone embedded in the Head-Mounted Display (HMD), the voice and contextual sounds generated by the audience member (e.g., while cheering, clapping), and transmits them onto the network;
- *Gestures:* This component tracks the movements of the audience member (typically the position and rotation of head and hands), and transmits them onto the network at fixed intervals;
- *Emotions:* This component retrieves in real-time, from biometric signals (e.g., EEG, heart rate), the emotional state of the audience member (as a class of 4 basic emotions as reported in [7]), and transmits it onto the network at fixed intervals.

The processors are as follows:

- *Virtual Environment (VE) Controls:* This processor takes as input the emotional state of each audience member and produces as output (e.g., via majority voting schemes) a set of control messages for changing the parameters of the VE (e.g., brightness, contrast);
- *Avatars Synchronizer:* This processor takes as input the gestures of audience members and uses them to control the embodiment of their corresponding avatar, synchronized across all instances of the multi-player VR application. This processor also receives as input the data from the Multimodal Prediction module;
- *Audience Sounds Mix:* This processor ingests the audio streams from every audience member. It blends the sounds of avatars positioned at a considerable distance from the listener's own avatar into one mixed track, while sending to the listener the individual sound streams of up to ten nearby avatars (if their distance from the user is

compatible with perceptual constraints). This processor also receives as input the data from the Multimodal Prediction module;

- *Multimodal Prediction*: This module takes as input the data related to gestures, sounds and delay compensation, and produces predictions for gestures and sounds (via ML methods based on previous data) in order to cope with the heterogeneous latencies that would prevent audio-visual synchronization;
- *Music Streaming*: This processor takes as input a database of compensation delays and uses them to stream music content to the connected audience members, each with a different delay. This ensures all geographically displaced audience members receive the music stream at the same time, regardless of the latency introduced by the network;
- *Latency Compensation*: This processor implements a network controller that updates a database of network-related information used to compute compensation delays for the Music Streaming and Multimodal Prediction modules.

The consumers are as follows:

- *VE Visuals*: This component runs the multi-player VR application that updates the parameters of the VE based on the control messages received from the VE Controls;
- *Avatars Visuals*: This component runs the multi-player VR application that updates the movements of the audience avatars based on the control messages received from the Avatars Synch;
- *3D Audio*: This component receives as input the position of the closest avatars surrounding the avatar of the audience member and delivers to them their sounds spatialized according to 3D audio methods, along with the mix of the sounds of the farther avatars;
- *Haptics*: This component provides a tactile representation of the music and other sounds, based on the audio signal produced by the 3D Audio device.

### B. Graph-Based Service and Network Modeling

We leverage the CNFlow modeling and optimization framework [17] to propose a graph-based model for MM services and networks. Unlike VNE, which views a service as a collection of point-to-point demands without information awareness, i.e., without actual knowledge of the information they carry, CNFlow allows modeling information-aware service placement, routing, and resource allocation problem as a single, unified information-flow problem on an augmented cloud-network graph (CNG).

The CNG modeling the physical infrastructure not only includes communication links and nodes, but also available compute and storage resources at those nodes. Network nodes (e.g., user device, edge server, cloud data center) are characterized by their computational capacity to host and run service functions and possibly storing content, and network links are characterized by their communication capacity and propagation delay.

On the service side, the SG is an information-aware graph that captures the structure of the application’s data flows and

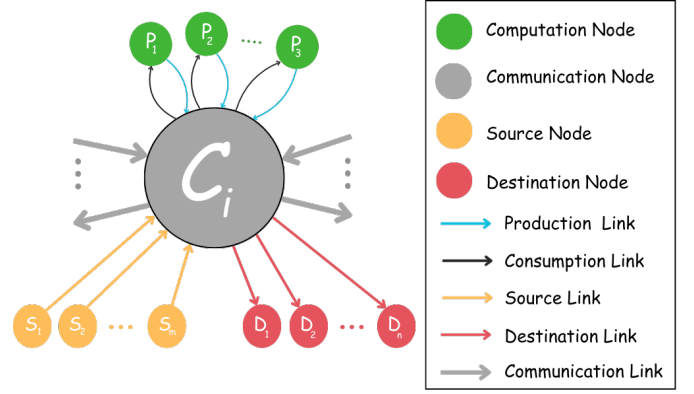


Fig. 2. Augmented cloud-network graph. Gray edges represent network links indicating the availability of communication resources for transmitting information between nodes, yellow for producing data (source), red for consuming data (destination), blue and black are computation links representing producing and consuming capabilities, respectively.

processing requirements. A SG is typically a DAG where *vertices* represent service functions, e.g., audio analysis, mixing, spatial rendering, sensor processing, and *edges* represent data streams, e.g., audio signals, control messages, flowing between functions or destined to end-users. Importantly, SGs in this graph modeling, include information attributes for each stream: for example, an audio stream might be one information object with a certain content, bitrate, and latency requirement. In particular, we differentiate between commodities (edges in the SG) and information objects (attributes that determine the information carried by each commodity). A given *information object* (say, a particular musician’s audio stream) may be carried by multiple commodities, each indicating the need to receive that same information by multiple destinations (all other performers and audience members who should receive that audio).

*Cloud-Network Graph Model*: We model the physical infrastructure as a CNG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where vertices represent network nodes (e.g., core nodes, edge cloud nodes) and edges represent links between computing locations. Each node  $i \in \mathcal{V}$  is augmented as illustrated in Fig. 2, to take into account their heterogeneous capabilities. In the figure,  $\mathcal{V}^c, \mathcal{V}^s, \mathcal{V}^p$ , and  $\mathcal{V}^d$ , are used to model the *communication* (where messages are transmitted over the network), *source* (where data is generated), *computation* (where data is processed), and *destination* (where data reaches its final destination) nodes, respectively. The resulting sets of nodes and links are denoted as:  $\mathcal{V} = \mathcal{V}^c \cup \mathcal{V}^s \cup \mathcal{V}^p \cup \mathcal{V}^d$ , and  $\mathcal{E} = \mathcal{E}^c \cup \mathcal{E}^s \cup \mathcal{E}^p \cup \mathcal{E}^d$ , denoting the set of communication links, source links, computation links, and destination links, respectively. In addition, computation links can either be a *computation in* link  $\mathcal{E}^{pin} \subset \mathcal{E}^p$  (shown in black) or a *computation out* link  $\mathcal{E}^{pout} \subset \mathcal{E}^p$  (shown in blue) [11].

Each link  $(i, j) \in \mathcal{E}$  is characterized by its capacity  $c_{ij}$  and cost  $w_{ij}$  parameters. In particular, for each  $(i, j) \in \mathcal{E}^c$ ,  $c_{ij}$  and  $w_{ij}$  denote the capacity in communication flow units (e.g., bps) and the cost per unit flow, respectively. Analogously, for

each computation link  $(i, j) \in \mathcal{E}^p$ ,  $c_{ij}$  denotes the capacity in computation flow units. However, the unit of this value depends on the type of link: for  $(i, j) \in \mathcal{E}^{pin}$ ,  $c_{ij}$  represents memory resources (e.g., RAM), hence, the corresponding value is expressed in bits. On the other hand, if  $(i, j) \in \mathcal{E}^{pout}$ ,  $c_{ij}$  represents processing resources (e.g., CPU), it is measured in FLOPs. Source and destination links  $\mathcal{E}^s$ ,  $\mathcal{E}^d$  are assumed to have zero cost and high-enough capacity, acting as network ingress and egress points, respectively.

*Service Graph model:* A SG for a MM application is a graph  $\mathcal{M} = (\mathcal{N}, \mathcal{K})$ , where vertices represent service functions (e.g., a user sounds decoder) and edges corresponding data streams (or commodities), as shown in Fig. 3.

A function  $n \in \mathcal{N}$  can either be a *source node* (a producer of data), a *processing node* (a processor data), or a *destination node* (a consumer of data). Similarly, a commodity in  $\mathcal{K}$  can either be a *source commodity*, *processing commodity*, or *destination commodity*. Analogously,  $\mathcal{N} = \mathcal{N}^s \cup \mathcal{N}^p \cup \mathcal{N}^d$ , and consequently  $\mathcal{K} = \mathcal{K}^s \cup \mathcal{K}^p \cup \mathcal{K}^d$ . An commodity  $k \equiv (i, j) \in \mathcal{K}^p \cup \mathcal{K}^p$  represents a commodity produced by function  $i \in \mathcal{N}^s \cup \mathcal{N}^p$ , and consumed by function  $j \in \mathcal{N}^p \cup \mathcal{N}^d$ . We denote by  $\mathcal{X}(k)$  the set of required input commodities to produce commodity  $k$ . For example, for the commodity  $(i, j)$  shown in Fig. 3,  $\mathcal{X}((i, j))$  is the set of the two incoming source commodities  $\mathcal{K}^s$ . On the other hand,  $\mathcal{X}(z)$  consists only of commodity  $(i, j)$ . Moreover, we also denote by  $s(k) \in \mathcal{V}^s$  the node hosting the function producing source commodity  $k \in \mathcal{K}^s$ ,  $d(k) \in \mathcal{V}^d$  the node hosting the function consuming destination commodity  $k \in \mathcal{K}^d$ .

In  $\mathcal{M}$ , each commodity is characterized by a rate requirement  $R_{ij}^k$  denoting the required rate of commodity  $k \in \mathcal{K}$  when it goes over link  $(i, j) \in \mathcal{E}$ . Note that the rate of  $k$  will depend on the type of link  $(i, j)$  it traverses. We use  $R_{prod}^k$ ,  $R_{comm}^k$ , and  $R_{cons}^k$  to denote the rate of commodity  $k$  when it goes over a production link, communication link, or consumption link, respectively.

Finally, as described in [11], one of the most important aspects for efficient representation of real-time data-intensive applications is the concept of information awareness. We define the set of information objects  $\mathcal{O}$ , and the surjective *information mapping function*  $g : \mathcal{K} \rightarrow \mathcal{O}$  to indicate the object identifier  $o \in \mathcal{O}$  associated with each commodity  $k$ . This mapping function is key to allow the overlapping of commodity flows carrying the same information.

#### IV. MUMENET: MUSICAL METAVERSE NETWORK SIMULATOR

We now present MuMeNet, our proposed discrete-event simulator for the MM. MuMeNet allows researchers to (i) develop flexible algorithm to embed MM dynamic SGs on arbitrary network topologies with real-world parameters, (ii) stream fine-grained traffic commodities through those embeddings under realistic link-level impairments, and (iii) measure, study, and analyze end-to-end musical-quality, QoE, and QoS for users with a large set of extensible evaluation metrics.

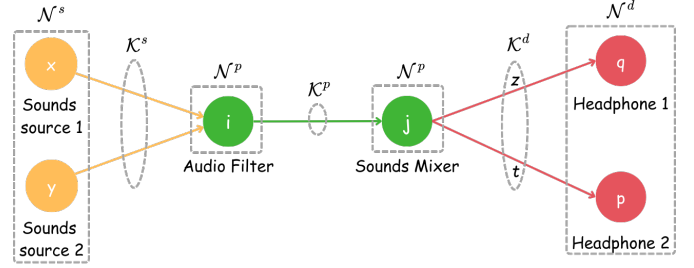


Fig. 3. Example of a SG, where edges represent data streams (commodities) and vertices service functions.

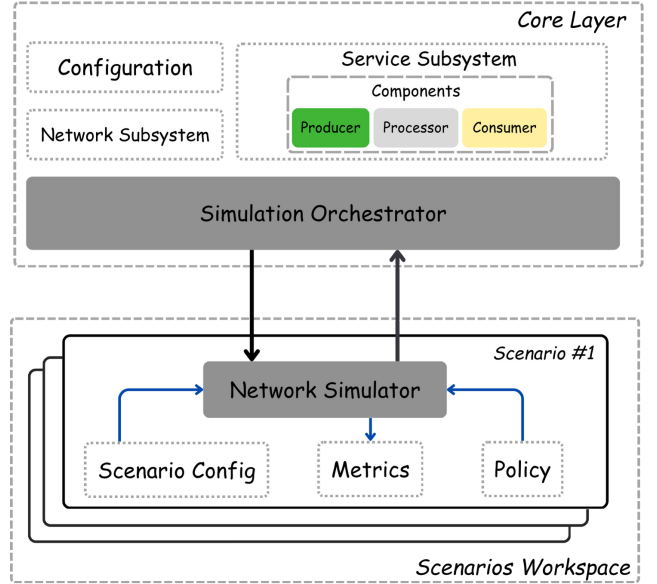


Fig. 4. MuMeNet Design Architecture: The top part shows the *Core Layer* responsible for general orchestration and management, while the bottom part shows *Scenarios Workspace* where event-driven experiments are performed.

Fig. 4 depicts the architectural design of MuMeNet, which consists of two main parts: the *core layer* and the *scenarios workspace*.

##### A. Core Layer

The Core Layer is the invariant runtime backbone of MuMeNet. It collects every (simulation) service that must remain stable across experiments (e.g., configuration management, network modeling, component life-cycle control, and the discrete-event orchestrator) into a single, version-controlled module. It bundles the following four subsystems:

*Simulator Configuration:* This subsystem provides the authoritative source for all runtime parameters needed by the simulator. At start-up, the engine instantiates a *BaseConfig* object that pulls a version-controlled YAML/JSON file with default values (e.g., simulation horizon, time step, random seeds, logging levels, and directories), validates every field against a schema and reports any mismatch, then archives the resolved configuration alongside the run outputs to preserve full reproducibility.

This architectural approach facilitates extensibility through inheritance-based configuration customization. Researchers can derive domain-specific configuration classes from the *BaseConfig* base class within their respective scenario directories. Such derived classes retain the capability to override existing attributes or introduce additional parameters while preserving the underlying validation and persistence mechanisms. Upon initialization, the simulation engine maintains an immutable reference to the configuration object, which serves as the canonical parameter source for all core subsystems, including the network model, metrics collection framework, and policy plug-in architecture. This design pattern ensures centralized parameter management with strong type safety guarantees while simultaneously enabling scenario-specific customization without requiring modifications to the core system implementation.

*Network Subsystem:* This subsystem is the bridge between the abstract network graph template and the event traffic that flows through the simulator at run time. It performs two high-level roles:

- **Graph Construction:** Assembling a topology whose nodes and links carry detailed resource attributes (e.g., packet loss, jitter, propagation delay). Moreover, a user can easily add some other dimensions.
- **Topological Consistency Checking:** Ensuring that the constructed topology obeys domain-specific constraints so that no unrealistic artifacts bias the results.

Note that at the start of every run, the network subsystem ingests either a ready-made template derived from real deployments or a synthetic topology.

*Service Subsystem:* A SG is realized as an ordered *set of components* (see Sec. III). Each component encapsulates the internal logic of a single service-graph node and is instantiated as one of three canonical roles (Fig. 1 shows how these roles combine to form the pre-recorded-concert template). Custom behavior is added by subclassing the role-specific bases (Producer, Processor, or Consumer), while the shared interface keeps every extension drop-in compatible with the framework. A component in the simulator passes through a set of different phases. First, the developer instantiates the object, during this phase the simulator validates the provided metadata (e.g., I/O rates, state size, latency budget) against domain constraints. Second, the component is added to the global registry, making it discoverable by other elements. Finally, after having all required components successfully registered, the simulator connects them via provisional data-flows and exchanges lightweight “dummy” messages to verify type compatibility, buffer sizes, and back-pressure signals. Any mismatch triggers a descriptive error before the actual experiment begins.

*Simulation Orchestrator:* The Orchestrator constitutes the control plane of the simulator. Its design is deliberately centered on a discrete-event scheduling paradigm, because audience interaction workloads exhibit highly uneven traffic bursts and tight latency budgets that are best captured by event-driven, rather than fixed-timestep execution.

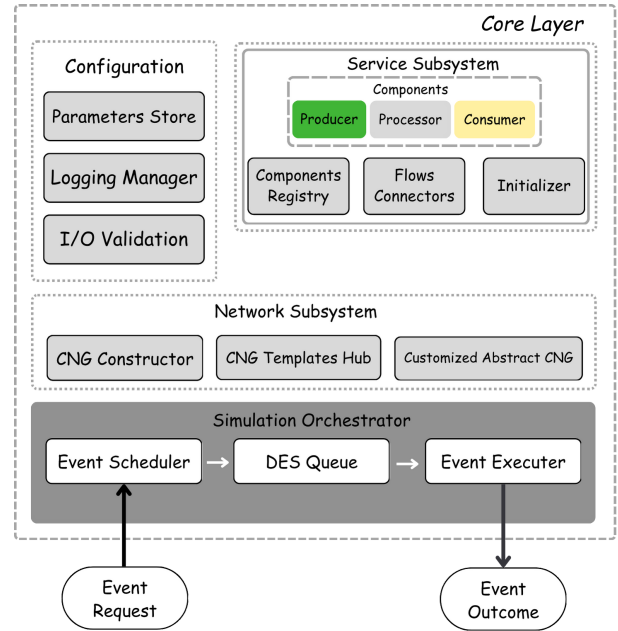


Fig. 5. Core Engine Architecture of MuMeNet.

The orchestrator fulfills the following responsibilities:

- **Global Time Monitoring:** It moves the simulation clock to the timestamp of the next event in the queue.
- **Event Dispatcher:** It invokes the *callback* attached to each event, thereby triggering computation, communication, or rendering inside the appropriate component. The event also accepts a set of optional arguments if required.
- **State Monitor and Update:** It updates link capacities, queue lengths, and component states after every callback, guaranteeing that subsequent events observe a consistent world view.
- **Variability Engine:** It injects stochastic effects (propagation-delay distributions, packet-loss bursts, device jitter) at the precise moment they materialize, this information is initially provided by the user in the configuration files.

Every occurrence in the simulator such as, e.g., sending a packet, completing a CPU task, finishing an audio render, is encoded as an *event object* with three immutable fields: *unique ID*, *timestamp* (at which the event must start executing), and *callback* (the function to execute).

Fig. 5 provides a more in-depth view of the core engine, where each module/subsystem is shown along with its parts.

### B. Scenarios Workspace

This workspace allows users to create their customized scenarios. Fig. 4 shows the general architectural design of a scenario inside the workspace. Four modules compose the workspace, as detailed below.

*Scenario Configuration Module:* This module can extend the set of initial configuration parameters defined in the *core* module. The reason behind this design is to ensure separability: only the necessary parameters are called, otherwise they

remain set to default. In addition to this, we store the necessary scenario configuration in this module. We share with the user the set of possible parameters to consider for the pre-defined scenarios; however, if the user wants to create a completely customized scenario out of the provided template, this may require additional steps as values have to be defined (and validated) first in the core part.

*Policy Module:* This module contains a set of abstract classes and methods to be extended for easy design of a customized policy. The policy's objective is to determine placement, routing, and resource allocation for the incoming SGs over the CNG, following static and dynamic approaches.

*Metrics Module:* This module collects real-time metrics of all components, monitors the changes of their statuses, and updates the list of metrics. This list depends on the use case; however, users can easily extend the base classes provided in this module to create any customized metric along with the interval of measuring or the set of components to use for collecting results.

*Network Simulation Module:* The scenario's key subsystem bridges the scenario with the core layer: it runs the orchestrator's network-level callbacks and returns updates or event requests, acting as the simulator's central communication hub, where all connections inside the simulator meet, from the CNG to the SG, and where metrics are collected before being sent to the "Metrics" module.

## V. PROBLEM FORMULATION AND MILP MODEL

The processors embedding problem encompasses the placement of processors and routing traffic flows within the SG over the CNG, given user locations. It is formulated as follows: given a CNG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with a set of compute-capable nodes interconnected by links with bandwidth and propagation delay; a set of users connected to specific access nodes; and a SG  $\mathcal{M} = (\mathcal{N}, \mathcal{K})$  representing the MM service pipeline, *decide* the placement of each processor in  $\mathcal{N}^p$  and the route for each commodity in  $\mathcal{K}$ , with the *objective* of either minimizing the sum of costs over utilized links, or minimizing the maximum link utilization, *constrained* by flow conservation, flow-chaining, source initialization, destination collection, commodity to information mapping, capacity bounds, and end-to-end latency bounds. Sets, list of variables, and list of parameters are reported in Table I.

*Variables:* We define two groups of variables:

- 1) *Virtual Commodity Flows*  $\{f_{ij}^k\}$ : Adimensional binary variables indicating whether commodity  $k \in \mathcal{K}$  goes (i.e., is transmitted, processed, or stored) over link  $(i, j) \in \mathcal{E}$ .
- 2) *Actual Information Flows*  $\{\mu_{ij}^o\}$  and  $\{\mu_{ij}\}$ : Real variables indicating the *amount* of information flow associated with object  $o \in \mathcal{O}$  and the *total* information flow, respectively, going over link  $(i, j) \in \mathcal{E}$ .

*Objectives:* We consider two different objective functions in our analysis:

- 1) *Minimum-Cost (MinCost):* This objective seeks the minimum cost feasible embedding by minimizing the sum

of costs (total cloud-network resource cost) over utilized link resources, as shown in Eq.1.

$$\sum_{(i,j) \in \mathcal{E}} \mu_{ij} w_{ij} \quad (1)$$

- 2) *Load Balancing (LB):* This objective minimizes the maximum utilization over links (i.e., maximum saturated link), with the aim of ensuring a balanced resources allocation in the network. This can be modeled as:

$$\min \left[ \max_{(i,j) \in \mathcal{E}} U_{ij} \right]; U_{ij} = \frac{\mu_{ij}}{c_{ij}} \quad (2)$$

where  $U_{ij}$  is the utilization (or load) of link  $(i, j) \in \mathcal{E}$ . Note that, since the nested min-max breaks linearity, we linearize it using the standard epigraph trick. Specifically, we use a single auxiliary variable  $\mathbf{Z}$ , resulting in an additional constraint. The final form becomes:

$$\min \mathbf{Z} \quad (3)$$

$$\text{s.t. } \mu_{ij} \leq \mathbf{Z} \cdot c_{ij} \quad (4)$$

$$\mathbf{Z} \in \mathbb{R}^+ \quad (5)$$

$$\text{Constraints (7)-(17)} \quad (6)$$

*Constraints:*

$$\text{s.t. } \sum_{j \in \delta^-(i)} f_{ji}^k = \sum_{j \in \delta^+(i)} f_{ij}^k \quad \forall i \in \mathcal{V}, k \in \mathcal{K} \quad (7)$$

$$f_{ij}^k = \begin{cases} f_{ji}^l & \forall k \in \mathcal{K}^p \\ 0 & \text{otherwise} \end{cases} \quad \forall l \in \mathcal{X}(k), i \in \mathcal{V}^p, j \in \delta^+(i) \quad (8)$$

$$f_{ij}^k = \begin{cases} 1 & \forall k \in \mathcal{K}^s \\ 0 & \text{otherwise} \end{cases} \quad \forall i = s(k), j \in \delta^+(i) \quad (9)$$

$$f_{ij}^k = \begin{cases} 1 & \forall k \in \mathcal{K}^d \\ 0 & \text{otherwise} \end{cases} \quad \forall j = d(k), i \in \delta^-(j) \quad (10)$$

$$f_{ij}^k R_{ij}^k \leq \mu_{ij}^o \quad \forall (i, j) \in \mathcal{E}, k \in \mathcal{K}, o = g(k) \quad (11)$$

$$\sum_{o \in \mathcal{O}} \mu_{ij}^o \leq \mu_{ij} \leq c_{ij} \quad \forall (i, j) \in \mathcal{E} \quad (12)$$

$$l^k = \sum_{(i,j) \in \mathcal{E}} l_{ij}^k f_{ij}^k \quad \forall k \in \mathcal{K} \quad (13)$$

$$l_T^k = l^k \quad \forall k \in \mathcal{K}^s \quad (14)$$

$$l_T^k \geq l^k + l_T^l \quad \forall k \in \mathcal{K}^p \cap \mathcal{K}^d, l \in \mathcal{X}(k) \quad (15)$$

$$l_T^k \leq L^k \quad \forall k \in \mathcal{K}^d \quad (16)$$

$$f_{ij}^k \in \{0, 1\}, \mu_{ij}^o \in \mathbb{R}^+, \mu_{ij} \in \mathbb{R}^+, l^k \in \mathbb{R}^+, l_T^k \in \mathbb{R}^+ \quad \forall (i, j) \in \mathcal{E}, k \in \mathcal{K}, o \in \mathcal{O} \quad (17)$$

Constr. 7 generalizes communication, computation, and storage flow conservation, requiring the total incoming flow to a given communication node  $i \in \mathcal{V}$  for a given commodity  $k \in \mathcal{K}$  to be equal to the total outgoing flow from node  $i$  for commodity  $k$ . Constr. 8 enforces the flow chaining condition, i.e., a computation node  $i \in \mathcal{V}^p$  can produce commodity  $k \in \mathcal{K}$  only if all required input commodities  $l \in \mathcal{X}(k)$  are available at its input. For instance, generating an enhanced video ( $k$ ) requires the decoded video ( $l$ ) as input. Constr. 9 and 10 define the source and destination conditions, ensuring that each commodity  $k \in \mathcal{K}^s$  is generated at its unique source node  $s(k)$  and delivered to its destination  $d(k)$ . Constr. 11 ensures that overlapping flows carrying the same information can share

TABLE I  
SYSTEM MODEL SETS, PARAMETERS AND VARIABLES

Sets	Description
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Cloud-Network graph
$\mathcal{V}^c; \mathcal{V}^s; \mathcal{V}^p; \mathcal{V}^d$	Communication nodes; source nodes; computation nodes; destination nodes
$\mathcal{E}^c; \mathcal{E}^s; \mathcal{E}^p; \mathcal{E}^d$	Communication links; source links; computation links; destination links
$\mathcal{E}^{pin}; \mathcal{E}^{pout}$	Computation in links (storage resources); Computation out links (processing resources)
$\delta^+(i); \delta^-(i)$	Incoming and outgoing neighbors of node $i \in \mathcal{V}$
$\mathcal{M} = (\mathcal{N}, \mathcal{K})$	Service graph DAG
$\mathcal{N}^s; \mathcal{N}^d; \mathcal{N}^p$	Source functions; destination functions; processing functions
$\mathcal{K}^s; \mathcal{K}^d; \mathcal{K}^p$	Source commodities; destination commodities; processing commodities
$\mathcal{X}(k)$	Set of input commodities required to produce commodity $k$
$s(k); d(k)$	Source node hosting the function producing commodity $k \in \mathcal{K}^s$ ; Destination node hosting the function consuming commodity $k \in \mathcal{K}^d$ .
Parameters	Description
$c_{ij} \in \mathbb{R}^+$	Capacity of link $(i, j)$
$w_{ij} \in \mathbb{R}^+$	Cost of link $(i, j)$
$U_{ij} \in [0, 1]$	Load of link $(i, j)$
$R_{ij}^k \in \mathbb{R}^+$	Rate of $k \in \mathcal{K}$ when it goes over link $(i, j)$
$l_{ij}^k \in \mathbb{R}^+$	Latency to transmit or process a unit of $k$ over link $(i, j)$
$l^k \in \mathbb{R}^+$	Local latency of commodity $k$
$L_T^k \in \mathbb{R}^+$	Cumulative latency of commodity $k$
$L^k \in \mathbb{R}^+$	Maximum service latency associated with destination commodity $k$
$\mathbf{Z} \in [0, 1]$	Utilization upper bound
Variables	Description
$f_{ij}^k \in \{0, 1\}$	Virtual commodity flow
$\mu_{ij}^o \in \mathbb{R}^+$	Actual information flow
$\mu_{ij} \in \mathbb{R}^+$	Total information flow

link capacity, reflecting the multicast nature of MM services. Recall that MM services involve real-time data streams required by multiple processing and/or destination functions. To capture this, the formulation allows virtual commodity flows corresponding to the same information object to overlap on a link  $(i, j) \in \mathcal{E}$ . This is achieved by weighting each flow variable with its rate requirement and summing only distinct information objects, not individual commodities. Constr. 12 limits the total information flow on each link  $(i, j) \in \mathcal{E}$  to its available capacity. This total flow is computed by summing the sizes of all distinct information flows traversing the link. Constr. 13-16 ensure end-to-end service latency constraints. Specifically, Constr. 13 enforces that the local latency of commodity  $k$ ,  $l^k$ , (i.e., the time taken to produce, deliver and consume a unit of commodity  $k$ ) is the sum, on the links carrying commodity  $k$ , of the latency to transmit or process a unit of commodity  $k$  over the given link, indicated by  $l_{ij}^k$ . Constr. 14-15 computes the cumulative latency of commodity  $k$ ,  $L_T^k$ , which represents the service latency that has been accumulated until the consumption of commodity  $k$ . Constr. 14 ensures that the cumulative latency is equal or less than the local latency for all source commodities. Constr. 15 computes the cumulative latency for all remaining commodities recursively by setting the cumulative latency of commodity  $k$  to be larger than or equal to the local latency of

TABLE II  
CONFIGURATION SETUP FOR CLOUD NETWORK TOPOLOGIES

Parameter	Range of values
Communication Links capacity	$\mathcal{U}(1000, 3000)$
Communication Links cost	$\mathcal{U}(10, 30)$
Computation in capacity (RAM)	$\mathcal{U}(1000, 3000)$
Computation in cost	$\mathcal{U}(10, 25)$
Computation out capacity (CPU)	$\mathcal{U}(1000, 2000)$
Computation out cost	$\mathcal{U}(10, 25)$
Computation Percentage	0.7
Number of computation nodes	$\mathcal{U}(1, 3)$
Density	0.7

commodity  $k$  plus the cumulative latency of input commodity  $l$ , for all input commodities in  $\mathcal{X}(k)$ . Lastly, Constr. 16 imposes the cumulative latency at each destination commodity to be no greater than the maximum allowed service latency  $L^k$ , while Constr. 17 imposes the binary nature of commodity flow variables and the real positive nature of information flow and latency variables.

## VI. EXPERIMENTAL SETTINGS AND RESULTS

### A. Experimental Settings

We perform our evaluations considering two distinct synthetic cloud network graphs (CNG) generated as a binomial topology  $G(n, \rho)$  with  $n \in \{8, 16\}$  with an edge-existence probability  $\rho = 0.7$  [26]. Link capacities and costs are drawn from a uniform distribution  $\mathcal{U}(a, b)$ , whose range of values is reported in Table II. We consider the *computation-node percentage*, i.e., the per-communication-node probability of hosting at least one computation node, to be equal to 0.7 to reflect heterogeneous edge-core deployments. We also consider  $N = 19$  SGs (each SG has an audience number equal to  $|\mathcal{U}| = N+1$  users). Table III reports the parameters used to generate the 19 SGs. Each SG follows the modeling shown in Fig. 1. We evaluate the MILP using the two objective functions described in Sec V. Although our experiments highlight the trade-off between monetary cost and network resilience under different load conditions, the primary objective is to demonstrate the types of analyses that can be performed using MuMeNet. The reported values are obtained by aggregating results from 100 repeated runs for each parameter combination, with confidence intervals calculated at a 95% confidence level ( $\alpha = 0.05$ ).

### B. Numerical Results

Figures 6 and 7 show the cost per user achieved by the MILP when employing the two different objective functions for varying numbers of users per SG for 8-node and 16-node CNG, respectively. In both cases, and for each objective, the cost per user shows a general downward trend, i.e., it decreases with more users per SG, with *MinCost* achieving between 10% to 15% and between 10% to 33% less cost than *LB* for the 8-node CNG and 16-node CNG, respectively<sup>1</sup>. This indicates that when more users share the same MM session, traffic can

<sup>1</sup>The *LB* shows a fluctuating curve for the 16-node CNG case. This is due to the fact that the objective is insensitive to small increases in total price as long as peak utilization is reduced.

TABLE III  
CONSIDERED VALUES OF PRODUCTION, COMMUNICATION AND CONSUMPTION RATES ASSIGNED TO EACH SYNTHETIC SG EDGE.

Pattern (src → dst)	$R_{\text{prod}}$	$R_{\text{comm}}$	$R_{\text{cons}}$
<b>Source</b>			
_ → VEControls	$\mathcal{U}(3, 10)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$
_ → AvatarSynch	$\mathcal{U}(50, 70)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$
_ → AudienceMix	$\mathcal{U}(80, 120)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$
MMPred <sub>gest</sub> → AvatarSynch	$\mathcal{U}(50, 70)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$
MMPred <sub>sound</sub> → AudienceMix	$\mathcal{U}(80, 120)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$
<b>Destination</b>			
VEControls → _	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(3, 10)$
AvatarSynch → _	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(50, 70)$
AudienceMix → _	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(120, 200)$
Streaming → _	$\mathcal{U}(10, 50)$	$\mathcal{U}(10, 50)$	$\mathcal{U}(120, 200)$
<b>Processing</b>			
MMPred → AvatarSynch	$\mathcal{U}(50, 70)$	$\mathcal{U}(30, 70)$	$\mathcal{U}(50, 70)$
MMPred → AudienceMix	$\mathcal{U}(80, 120)$	$\mathcal{U}(30, 40)$	$\mathcal{U}(80, 120)$
LatComp → Streaming	$\mathcal{U}(3, 10)$	$\mathcal{U}(3, 10)$	$\mathcal{U}(3, 10)$
LatComp → MMPred	$\mathcal{U}(3, 10)$	$\mathcal{U}(3, 10)$	$\mathcal{U}(3, 10)$

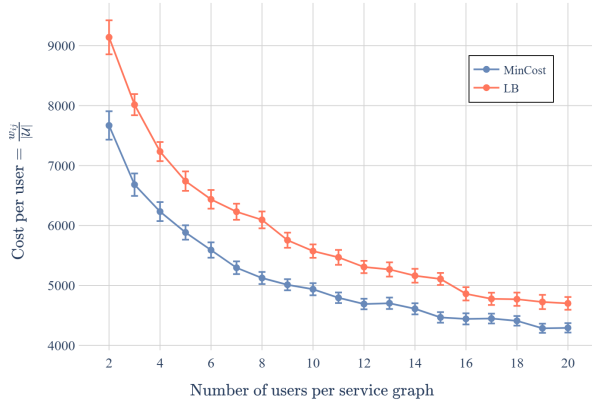


Fig. 6. Cost per number of users for varying number of users in the SG for 8-node binomial network topology.

be aggregated and common data streams can be shared, driving the per-user cost down.

More specifically, with 2 users per SG, *MinCost* achieves a cost of roughly 7700 units per user, whereas the *LB* achieves a cost of user 9200 (18% more). This surcharge arises as the *LB* deliberately routes traffic to minimize peak utilization, even

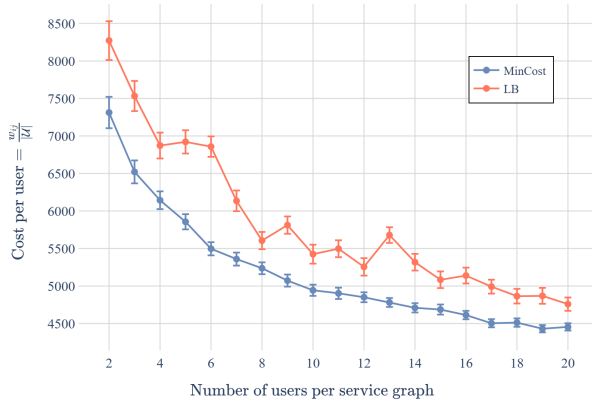


Fig. 7. Cost per number of users for varying number of users in the SG for 16-node binomial network topology.

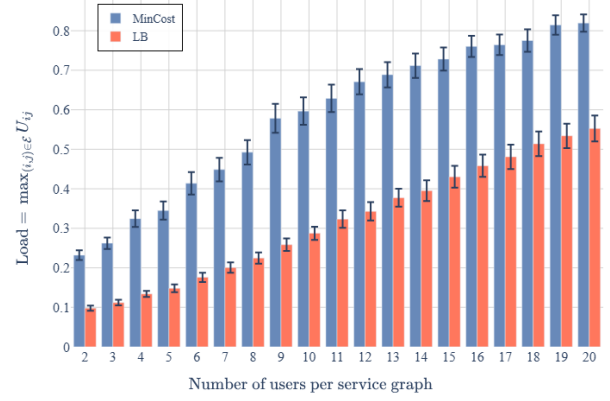


Fig. 8. Peak link utilization versus number of users per service graph on an 8-node binomial cloud network.

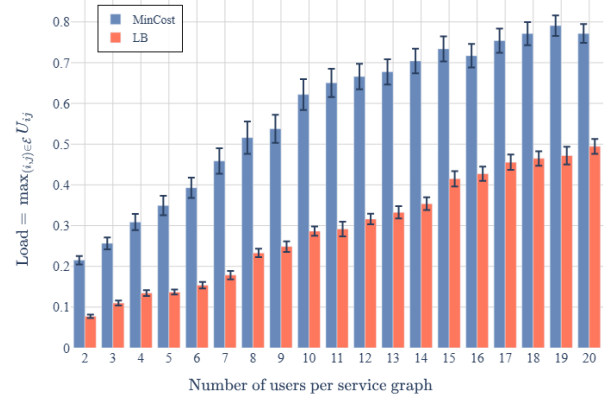


Fig. 9. Peak link utilization versus number of users per service graph on a 16-node binomial cloud network.

if it incurs additional monetary costs (i.e., even if it utilizes costlier links). Moreover, we further note that starting from 15 users per SG for both 8 and 16-node CNGs, the two curves begin to align more closely (the difference at 20 users is  $\approx 350$  units, or  $\approx 10\%$ ). This indicates that with more concurrent users, multicast-style traffic replication and the information-aware placement model force both formulations to reuse the same high-utility paths; consequently, balancing the load no longer carries a significant additional price tag.

We now focus on load balancing. Figures 8 and 9 report the maximum link-utilization ratio achieved by the two MILP formulations on the 8- and 16-node CNGs, respectively. In both settings, and as expected, *LB* keeps the maximum-utilized link markedly below that of *MinCost*. Specifically, with 2 to 7 users, *LB* achieves a load between 0.16 and 0.2, whereas the *MinCost* achieves a cost between 0.2 and 0.45 (25% to 55% more). We also observe that these percentages increase significantly up to 63% (0.47-0.5 for *LB* against 0.76-0.79 for *MinCost*) when the number of users in the SG is between 18-20 users. This demonstrates that while *LB* may incur an additional cost of roughly 700 units per user on average, that cost can be compensated by realizing a more balanced load across the network.

In summary, the analysis highlights a clear trade-off between cost efficiency and load balancing. While the *MinCost* formulation yields lower per-user costs, especially at smaller scales, the *LB* formulation achieves significantly better load distribution across the network. As the number of users increases, the gap in cost narrows, making *LB* a compelling choice in scenarios where network stability and QoE are critical.

## VII. CONCLUSION

In this paper, we study the end-to-end service provisioning for time-critical MM sessions over 5G/6G infrastructures. Starting from a detailed analysis of interactive virtual-concert workflows, we (i) distill their functional dependencies into a service graph (SG), (ii) formalize the MM processor placement and routing problem as a CNFlow problem, (iii) present MuMeNet, a discrete-event simulator that reproduces sub-millisecond cross-stream synchronization under realistic network impairments, and (iv) formulate a Mixed-Integer Linear Program (MILP) that embeds MM SGs onto arbitrary CNGs under two different objectives: Minimizing overall cost, and minimizing maximum resource utilization. On the experimental side, integrating MuMeNet with commercial cloud/edge platforms will allow empirical validation of the simulator assumptions and extend the impairment models. In summary, this study lays the theoretical and methodological groundwork for delivering latency-critical, multisensory musical interactions at Metaverse scale. We believe that MuMeNet will accelerate reproducible research on ultra-low-latency multimedia systems, bridging the methodological gap between networking, edge computing, and musical-interaction communities.

## ACKNOWLEDGMENT

This work is supported by MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379), and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

## REFERENCES

- [1] S. Mystakidis, "Metaverse," *Encyclopedia*, vol. 2, no. 1, pp. 486–497, 2022.
- [2] Y. Cai, J. Llorca, A. M. Tulino, and A. F. Molisch, "Compute-and data-intensive networks: The key to the metaverse," in *2022 1st international conference on 6G networking (6GNet)*. IEEE, 2022, pp. 1–8.
- [3] L. Turchet, "Musical metaverse: vision, opportunities, and challenges," *Personal Ubiquitous Comput.*, vol. 27, no. 5, p. 1811–1827, Jan. 2023.
- [4] L. Turchet, R. Hamilton, and A. Çamci, "Music in extended realities," *IEEE Access*, vol. 9, pp. 15 810–15 832, 2021.
- [5] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.
- [6] L. Turchet, C. Sassi, D. Vecchia, and G. P. Picco, "Real-time musical haptics with ultra-wideband: A study on latency, reliability, and perception," *IEEE Transactions on Haptics*, vol. 18, no. 1, pp. 269–280, 2025.
- [7] L. Turchet, B. O'Sullivan, R. Ortner, and C. Guger, "Emotion recognition of playing musicians from eeg, ecg, and acoustic signals," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 5, pp. 619–629, 2024.

- [8] L. Turchet, T. West, and M. M. Wanderley, "Touching the audience: musical haptic wearables for augmented and participatory live music performances," *Personal Ubiquitous Comput.*, vol. 25, no. 4, p. 749–769, Mar. 2020.
- [9] L. Turchet and M. Tomasetti, "Immersive networked music performance systems: identifying latency factors," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2023, pp. 1–6.
- [10] M. Ali, F. Naeem, G. Kaddoum, and E. Hossain, "Metaverse communications, networking, security, and applications: Research issues, state-of-the-art, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 1238–1278, 2024.
- [11] A. Mauro, A. M. Tulino, and J. Llorca, "End-to-end orchestration of nextg media services over the distributed compute continuum," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2025.
- [12] F. Tang, X. Chen, M. Zhao, and N. Kato, "The roadmap of communication and networking in 6g for the metaverse," *IEEE Wireless Communications*, vol. 30, no. 4, pp. 72–81, 2023.
- [13] R. D. Tripathi, M. Lyu, and V. Sivaraman, "Assessing the impact of network quality-of-service on metaverse virtual reality user experience," in *2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*. IEEE, 2024, pp. 206–213.
- [14] 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 Communications Letters*, vol. 11, no. 8, pp. 1733–1737, 2022.
- [15] J. Zhao, L. Qian, and W. Yu, "Human-centric resource allocation in the metaverse over wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 3, pp. 514–537, 2023.
- [16] N. H. Chu, D. T. Hoang, D. N. Nguyen, K. T. Phan, E. Dutkiewicz, D. Niyato, and T. Shu, "Metaslicing: A novel resource allocation framework for metaverse," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 4145–4162, 2023.
- [17] M. Barcelo, A. Correa, J. Llorca, A. M. Tulino, J. L. Vicario, and A. Morell, "Iot-cloud service optimization in next generation smart environments," *IEEE J.Sel. A. Commun.*, vol. 34, no. 12, p. 4077–4090, Dec. 2016.
- [18] M. Michael, J. Llorca, and A. Tulino, "Approximation algorithms for the optimal distribution of real-time stream-processing services," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.
- [19] H. Wang, R. Martinez-Velazquez, H. Dong, and A. El Saddik, "Experimental studies of metaverse streaming," *IEEE Consumer Electronics Magazine*, vol. 14, no. 1, pp. 26–36, 2024.
- [20] T. R. Henderson, M. Lacage, G. F. Riley, C. Dowell, and J. Kopena, "Network simulations with the ns-3 simulator," *SIGCOMM demonstration*, vol. 14, no. 14, p. 527, 2008.
- [21] M. Lecci, M. Drago, A. Zanella, and M. Zorzi, "An open framework for analyzing and modeling xr network traffic," *IEEE Access*, vol. 9, pp. 129 782–129 795, 2021.
- [22] A. Varga and R. Hornig, "An overview of the omnet++ simulation environment," in *Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops*, 2008, pp. 1–10.
- [23] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1888–1906, 2013.
- [24] M. C. Silva Filho, R. L. Oliveira, C. C. Monteiro, P. R. M. Inácio, and M. M. Freire, "Cloudsim plus: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2017, pp. 400–406.
- [25] Q. Pagliuca, L. J. Chaves, P. Imputato, A. Tulino, and J. Llorca, "Dual timescale orchestration system for elastic control of nextg cloud-integrated networks," in *2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN)*, 2024, pp. 234–241.
- [26] O. ERD and A. Renyi, "On random graphs," *Publ. Math.*, vol. 6, pp. 290–297, 1959.