POLITECNICO DI TORINO
Repository ISTITUZIONALE

Towards a Networking Architecture for the Musical Metaverse: The MUSMET Vision

Original
Towards a Networking Architecture for the Musical Metaverse: The MUSMET Vision / Turchet, L., Boem, A., Ayoub, O., Malandrino, F., Llorca, J., Fischione, C., Rottondi, C.. - ELETTRONICO. - (2025), pp. 1-8. (2025 IEEE 6th International Symposium on the Internet of Sounds (IS2) L'Aquila (Ita) 29-31 October 2025) [10.1109/is264627.2025.11284568].

Availability:
This version is available at: 11583/3006481 since: 2026-01-12T20:37:50Z

Publisher:
IEEE

Published
DOI:10.1109/is264627.2025.11284568

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)

15 October 2026

Towards a Networking Architecture for the Musical Metaverse: The MUSMET Vision

Luca Turchet*, Alberto Boem*, Omran Ayoub†, Francesco Malandrino‡,
Jaime Llorca*§, Carlo Fischione§, and Cristina Rottondi¶
*Dep. of Information Engineering and Computer Science, University of Trento, Trento, Italy
†Dep. of Innovative Technologies University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland
‡ Centro Nazionale di Ricerca, CNR-IEIIT, Turin, Italy
§School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
¶Dep. of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
§ Centre Tecnologic de Telecomunicacions de Catalunya (CTTC/CERCA), Castelldefels, Spain
Emails: luca.turchet@unitn.it, alberto.boem@unitn.it, omran.ayoub@supsi.ch,
francesco.malandrino@cnr.it, jaime.llorca@unitn.it, carlofi@kth.se, cristina.rottondi@polito.it

***Abstract*—The Musical Metaverse (MM) relates to a vision of an immersive environment that enables and augments social musical activities involving distant musicians and audience members. Currently, this emerging field of research faces several technical challenges that prevent its realization. This paper identifies such challenges and proposes a research agenda to address them, which falls within the research program of the MUSMET project, recently funded by the European Innovation Council. A cornerstone of such an agenda is a networking architecture specific for the MM, which makes extensive use of 5G and beyond 5G technologies to realize the project vision. We present the main hardware and software components of such an architecture, including the end nodes, the edge devices, and the cloud, considering the strict quality of service and quality of experience constraints imposed by MM applications.**

***Index Terms*—Musical Metaverse, Internet of Musical Things, Musical XR, 5G**

I. Introduction

Throughout history, music composition, performance, and experience have continually transformed due to technological advancements and changing preferences of musicians and audiences. Today, the Metaverse represents an emerging medium with the great but yet unexplored potential for creating and performing music in novel ways, as well as for creating radically new musical experiences for audiences. This has led researchers to propose the concept of the "Musical Metaverse" (MM), a term that describes a version of the Metaverse dedicated to musical activities, including attending concerts, participating in collaborative composition, or delivering immersive remote education (e.g., [1]–[8]). The MM will make possible to rethink musical activities such as live performance, composition, and education by enabling a rich repertoire of remote interactions between stakeholders, exploits the understanding of users' emotional states, and providing compelling multisensory experiences. In the MM musicians not only will be able to conduct musical activities using their real, physical musical instruments, but can also avail themselves with purely virtual instruments. All this has the unexplored potential to revolutionize the way music is collaboratively composed, performed, taught, learned, and experienced in Extended Reality (XR) environments. Such environments are built leveraging both Virtual Reality (VR) and Mixed Reality (MR) technologies.

As shown by recent user studies, musicians need enhanced ways of musical expression and interaction with music and their audience [9], [10]. This trend is evident not only from the increasing use, of XR technologies to create virtual concerts [11], but also from panel discussions at music fairs (such as Slush Music) and conferences (such as the 1st IEEE International Workshop on the Musical Metaverse[1]), as well as recent academic publications [10], [12]–[15]. Meanwhile, today's audiences seek more engaging musical experiences and a closer interaction with performers and other audience members, as indicated by a growing body of scientific research [16]–[18]. End-users need real-time technological solutions to support the sense of immersion, the feeling of being together in the same environment in order to share the same experience with the others [19].

However, the technologies that are at the core of the Metaverse are insufficient to satisfactorily address not only the demands of audience members, but also and most importantly of musicians [20]. This is due to the lack of market-ready technical solutions tailored for musical experiences in the Metaverse. As a consequence, musicians and audiences, along with the music industry, have not yet fully exploited the enormous artistic and economic possibilities that the Metaverse can offer. First, the Metaverse is not ready for musicians who decide to take advantage of the distinctive features of social immersive environments to express themselves and create music together. This is mainly due to the lack of appropriate tools for music-making in shared immersive environments, as

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379) and the Swiss State Secretariat for Education, Research and Innovation.

[1]https://internetofsounds2024.ieee-is2.org/workshop/ieee-iwmm-workshop

well as the lack of systems supporting real-time interactions between distant users [20]. Second, the Metaverse has several limitations in terms of truly immersive user experiences for audiences attending virtual performances. This is mainly due to the lack of methods that enable audience members to feel fully immersed in the performance, and the lack of tools to make audience members satisfactorily interact with the performers as well as with each other. Third, being the Metaverse a new framework for gathering data, including biometric data, and where users' identity is virtualized, ethical issues such as privacy and security still need to be properly addressed to ensure trustworthy experiences for music stakeholders [1], [21].

The design of collaborative systems able to deliver radically new musical experiences in the Metaverse requires the integration of fundamental social cognitive aspects (including emotions, flow, togetherness), which necessitates a deep understanding of human interaction with music [22]. The real-time monitoring and repurposing of neurophysiological user states represent a potentially successful avenue for this purpose. However, to the best of our knowledge, thus far no research effort has been conducted to inform MM systems with such features. Most of commercially available Head-Mounted Displays (HMDs) are not equipped with technologies that can either understand the cognitive states of the user or retrieve information about the music played. Enhancing XR devices with such intelligent capabilities has the concrete potential to enable completely new kinds of interaction between musical stakeholders as well as new ways of experiencing music performances.

The integration of contextual information into MM systems entails the processing of vast amounts of multimodal data, distributed over a large number of geographically displaced systems. To properly handle this data, we need an optimization of current data modeling methods, advanced techniques for real-time distributed Machine Learning (ML) over communication networks, and comprehensive and temporally accurate datasets specific to MM applications. All these features are expected to enhance XR devices with unprecedented context-awareness and proactivity capabilities, which will be able to support interactions between various stakeholders not possible beforehand.

To address all the aspects above, there is a need for an interdisciplinary endeavor at the confluence of academic, industrial, and artistic research, merging competences from the field of XR, networking, artificial intelligence, affective neuroscience, and music. This is the ambition of the recently started MUSMET project (2025–2029)[2] funded under the auspices of the European Innovation Council and the Swiss State Secretariat for Education, Research and Innovation. MUSMET proposes a vision and cutting-edge technological innovation for the future classes of MM devices, networking systems, and services, capable of catering to the needs and expectations of musicians and audiences. The implementation of this vision will spur the creation of radically new ecosystems of interoperable devices and communities utilizing them, which holds significant benefits for society, economy, and art. The approach proposed in the project aspires to effect a step-change in the design of musical interfaces and systems to musically interact online, resulting in a potentially high economic impact on the music industry: progressing the technologies at the core of musical interfaces, telecommunication, and services not only entails the advancement of musical devices manufacturing processes, artistic practices, and the way musical stakeholders interact, but also paves the way for new monetization opportunities, the creation of new markets, and the definition of novel business models.

In this paper, we identify some of the existing technical challenges preventing the realization of a Metaverse tailored for musical experiences and propose a network architecture for the MM, which makes extensive use of 5G and beyond 5G technologies. We also discuss the technical changes that will be faced and the goals to be achieved during the MUSMET project. More in detail, in Section II, the related scientific literature and the shortcomings of state-of-the-art solutions are presented, whereas an agenda to address the identified issues is illustrated in Section III. The envisioned networking architecture is described in Section IV. A final discussion is offered in the last section.

II. Technical Challenges

A. XR audio development tools

A crucial aspect hindering the development of distributed interactive applications such as those envisioned in the MM is the inadequacy of audio tools currently offered by game engines used to create immersive applications (e.g., Unity3D) [20]. Current frameworks used in commercial multiuser applications (e.g., Photon Fusion) only support microphones as audio input and are specifically designed for speech broadcasting. Instead, the MM requires a music-oriented, multidirectional, and low-latency audio streaming. This entails the integration of Networked Music Performance (NMP) systems [23], which allow distant musicians to play together over the Internet. However, NMP systems have neither been primarily conceived for the MM nor integrated within it, which poses significant technical challenges such as delivering a consistent immersive experiences that combine spatial audio, visuo-spatial awareness, embodied and social interactions, as well the integration of biometric monitoring.

Moreover, achieving immersive audio experiences requires seamless integration of low-latency spatial audio in MM platforms and NMP systems, minimizing their latency contribution while preserving the offered audio quality [24]. Additionally, identifying the most effective spatialization for multi-user MM is essential, both from the perceptual and processing latency perspective.

B. Music-oriented XR hardware

Current HMDs have not been specifically designed for music-making applications: current hardware needs to be ex-

[2]https://musmet.eu/

tended with music-specific input/output capabilities to support the usage of tools such as acoustic, electric, and digital musical instruments [20]. Especially standalone HMDs are primarily designed as visual-first devices, and audio is relegated to be an ancilliary part of the experience, both in terms of input and output capabilities, as well as in terms of dedicated processing. Moreover, they typically rely on wireless connectivity via Wi-Fi or Bluetooth communication technologies, which do not meet the requirements of low latency and high bandwidth imposed by MM applications. Similarly, tethered HMDs (i.e., those that only work in conjunction with an external computer) rely on the connection with a multitude of external audio hardware, which is cumbersome, complex to set up, and may also introduce unacceptable latency.

C. MM Platforms

Existing platforms supporting performances in the Metaverse (e.g., Somnium Space, PatchWorld, VRChat) need to be improved at multiple levels to become suitable for the MM [11]. Most of live concerts that occur in the Metaverse are typically limited to small audiences (divided in instances typically of 50 users). Except from pre-recorded performances, live concerts have been performed by a single musician (mostly DJs) or a group of co-located musicians. Current platforms used for music are mostly based on frameworks for Social and Collaborative XR, where the tolerated latency is much larger the one for audio-centric applications, usually around 150 ms [25], [26].

Another significant limitation to the widespread deployment of the MM is that current platforms are not interoperable: each virtual world has been conceived as self-sufficient, without the ability to exchange information with other worlds. This fragmentation represents a limitation not only for performers (who cannot use their virtual musical instruments across different worlds) and audiences (who cannot move from a world to another preserving their profile and avatar), but also for MM platforms providers (who cannot monetize the opportunities resulting from a wider user base).

D. Social Music Cognition for the MM

At present, the assessment of users' cognitive and emotional states while interacting with MM platforms and other involved users has remained completely unexplored. Knowledge about such states is pivotal for unveiling new understandings of social aspects in music making and consumption [27]–[29]. The assessment of users' mental states can be conducted through EEG, audio, and bodily expressions captured by portable Brain-Computer Interfaces (BCIs) in fast, direct, and minimally invasive ways. However, currently compatibility between BCIs and XR systems (especially HMDs) is limited, which prevents the exploitation of emotional and mental states in MM contexts.

E. Embedded Real-Time Music Information Retrieval for MM applications

Opportunities to explore novel artistic formats arise from the real-time analysis of the content of the music signal produced by musicians in the MM, and from the re-purposing of the extracted information to control parameters of the XR environment. As of today, most Music Information Retrieval (MIR) research has focused on offline analysis of large audio datasets, such as for automatic transcription. While real-time MIR holds great potential for MM applications, the majority of existing methods lack robustness, especially in embedded systems [30]. This is due to strict processing time constraints and the absence of pre- and post-processing steps used in offline methods. Additionally, current music datasets suffer from temporal inaccuracies. To advance real-time MIR for the case of XR environments, novel methods and precisely annotated datasets at the sub-millisecond level are essential.

F. Real-Time ML for distributed MM Services

Existing communication and ML processing solutions for remote audio-only musical interactions already struggle with high latency, jitter, and poor sound quality. Musicians require ultra-low latency (below 30 ms [23]) and minimal packet loss (on the order of 10^{-4} [31]) for proper synchronization without audible signal deterioration, but current wireless networks, including 5G, do not meet these demands, as recently shown in [32]. Moreover, MM devices (e.g., HMDs, BCIs, musical instruments) generate vast amounts of data, requiring real-time transmission and processing, which existing networks and ML methods cannot efficiently handle: the key challenges include the lack of wireless protocols supporting real-time ML for music services and ML methods optimized for embedded networks. Current ML approaches rely on large datasets and substantial computing power, but in MM applications, datasets of any size will be distributed among several nodes that might not be able to share them in real-time due to bandwidth and privacy constraints, delays, and jitters. Furthermore, embedded devices that generate user and music-related data may lack the bandwidth necessary to transmit real-time data for remote analysis.

To address the aforementioned limitations, a shift towards federated and decentralized ML architectures could alleviate data transmission bottlenecks while maintaining data privacy across MM nodes. Incorporating lightweight inference models into XR Music Boxes could enable on-device real-time analysis and decision making, mitigating the latency of cloud-bound processing. Cross-layer optimization frameworks that jointly adapt ML processing and communication parameters based on musical context awareness can further enhance responsiveness and accuracy. Furthermore, real-time feedback loops between ML models and user interactions can support adaptive experiences that respond to evolving user states and performance conditions. Ultimately, benchmarking frameworks specific to MM applications should be developed to evaluate ML pipelines not only in terms of accuracy and latency but also musical expressiveness and perceptual quality. Innovative communication protocols tailored for real-time ML in MM applications are essential to overcome current limitations and enable seamless remote musical collaboration.

G. 5G and beyond 5G networks for the MM

Currently, most systems designed to interconnect remote musicians have been developed for wired networks and to support audio-only communications [33] or audiovisual communication utilizing 2D video streams. Recently, researchers have started to focus on the multimodal stream of information, but are still relying on wired networks [34], [35]. Existing deployments of audio-only NMP systems over 5G networks have consistently shown that 5G is a promising technology to satisfy the strict latency and perceptual quality demands needed by musicians [32], [36]. Nevertheless, those results also indicated that there is still a large margin of improvement, especially when dealing with a high number of connected nodes. As a complicating factor, the MM entails a large amount of multimodal data to be exchanged among musicians and audience members, well beyond conventional audio-visual communications. Therefore, there is a need to advance the state-of-the-art of wireless technologies, especially pursuing directions that go beyond current 5G architectures, as highlighted in [37]. However, at present, there are no guidelines on how to design NextG (5G and beyond) system architectures for supporting MM applications.

To realize the low-latency vision of MM, dynamic network slicing strategies must support variable QoS profiles tailored to session type (e.g., rehearsal vs. live performance). Moreover, the exploitation of Multi-Access Edge Computing (MEC) capabilities of 5G infrastructures for the placement of virtual functions and/or containerized microservices tailored for MM applications requires further investigation, to efficiently tackle the peculiar requirements and constraints imposed by our envisioned MM scenarios. In particular, the optimal selection of edge nodes to host MM functionality (e.g., MM microservices), in order to guarantee minimal cost and latency, and the definition of procedures for seamless migration between nodes in dynamic scenarios, while ensuring load balancing among available infrastructures, are yet open research directions [38]. Cross-domain orchestration of MEC nodes and cloud backends can facilitate the distributed processing of multimodal data streams, supporting scenarios with fluctuating audience sizes and geographic dispersions. Integrating semantic-aware routing mechanisms can improve the prioritization of mission-critical musical data packets (e.g., tempo cues, solo channels) over less time-sensitive background streams.

Advances in 6G research, such as Reconfigurable Intelligent Surfaces and THz communications, could significantly expand the MM's capabilities for ultra-high fidelity remote music making. Finally, robust Quality of Experience (QoE) models grounded in perceptual musicology must inform the design of network protocols to ensure that technical performance aligns with user expectations.

III. Research agenda

To tackle the challenges enumerated in the previous section, MUSMET sets a pathway for the realization of the following goals, organized in the following categories:

Hardware

- Design and development of an "XR Music Box", i.e., a smart portable device to be connected to both standalone and tethered HMDs that will handle all audio processing tasks and ultra-reliable low-latency wireless connectivity, integrating a 5G module into an audio-specific embedded system (see Fig. 1). Such custom hardware will be interfaced with a hard real-time operating system prioritizing audio-related tasks;
- Integration of haptic technology for musical use [39] which, up to now, has been largely neglected within MM platforms. Notably, haptic wearable devices can enhance the listening experience of audiences of live music concerts while ensuring inclusivity for the hearing-impaired population, who may leverage the sense of touch to compensate their listening deficits [40], [41];
- Integration of BCIs with XR environments and hardware, with the scope of identifying mental states relating to engagement, emotion and flow in social interactions and use this information to i) adapt musical experiences accordingly, with dynamic visualizations and sound based on audiences' collective mental states; ii) identify periods of high or low flow in multi-user musical performance; and iii) monitor increased engagement in audience members.

Software

- Integration of immersive NMP systems, with high-quality spatial audio rendering optimized for low-latency processing, into MM platforms;
- Acquisition and exploitation of the emotional states of audience members, detected via a multimodal stream of voice (e.g., cheering), gestures (e.g., clapping), and brain signals, to create radically new visual, sonic and haptic experiences for both musicians and audiences that can increase the levels of presence and immersion, as well as the sense of social togetherness;
- Design and implementation of spatial audio rendering techniques tailored for large audiences, by defining a fluent communication model where only the audio generated by the user's nearest participants are transmitted, whereas audio flows from farther audience members will be processed by means of several multichannel submixes, rendered in the edge, which will be streamed to the user's XR music box where the final audio rendering will occur;
- Definition of a solution for ensuring interoperability across the MM platforms, which will allow users to maintain their identity across them and, in the case of musicians, also their musical setup;
- Development of methods for the synchronization among the concurrent multimodal streams generated by MM devices (which involve diverse data types such as EEG, sound, haptics, and avatar control signals like gestures), including: (i) time-stamping and alignment of heterogeneous data streams at capture, (ii) latency estimation

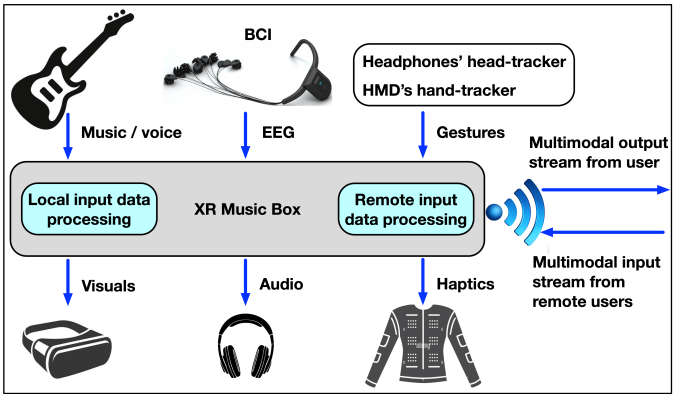

Fig. 1: The XR music box and its interactions.

and compensation across modalities, (iii) buffering and adaptive resampling to handle jitter or missing data, and (iv) psycho-perceptive calibration procedures to ensure perceptual coherence between auditory, visual, neural, and haptic feedback in real-time.

Network

- Establishment of fundamentally new wireless communication methods to support real-time ML over networks for music services, including new encoding formats and new access schemes for small-sized packets;
- Enhancement of the current generation of MM systems to support real-time, multidirectional, large-scale, and multisensory remote interactions among musicians, as well as among musicians and audiences;
- Development of adaptive ML-based traffic forecasting models that anticipate multimodal data flows, including spatial audio, haptic signals, BCI inputs, and emotional feedback, to enable proactive orchestration and ultra-low-latency resource allocation across edge and cloud infrastructures in the MM;
- Development of end-to-end orchestration solutions capable of dynamically optimizing 1) the placement of MM stream-processing functions or microservices over edge and cloud platforms, 2) the routing of MM multimodal data streams, and 3) the joint allocation of computation, communication, and storage resources.
- Investigation of scalability and cost-aware deployment strategies for MM networking architectures, with a focus on quantifying how communication, computation, and storage demands grow when supporting large ensembles (e.g., hundreds of simultaneous performers). This includes the analysis of bandwidth partitioning, edge–cloud load balancing, and latency–jitter control mechanisms under dense user scenarios, as well as the evaluation of economic trade-offs between centralized and distributed deployments to ensure sustainable and accessible large-scale MM services.

Applications

It is important to distinguish three main categories of MM applications:

1) those involving only VR, where musicians play using instruments that exist only in a virtual space, while audience members attend the performance within the same virtual environment where musicians are placed;
2) those involving only MR, where musicians play using conventional musical instruments and audience members attend the performance in spaces that superimpose virtual elements to the real scene;
3) those involving both VR and MR, where the cases above can be combined.

MUSMET will focus on these three categories, which have an increasing level of complexity and different requirements.

IV. Proposed Network Architecture

First and foremost, to achieve the goals set in the research agenda above, there is a strong need for designing innovative mobile networking architectures specific for the MM vertical. Fig. 2 illustrates the architecture that will be implemented in the MUSMET project to interconnect the various MM stakeholders.

Ad hoc deployment solutions of the 5G User Plane Function (5GUPF) need to be devised to support edge computing in the MM framework, adapting to our scenario the 5G architectural variants proposed by 3GPP in [42]. Instantiating 5GUPFs at the edge of the network is a key enabler of distributed edge computing (MEC). This minimizes data exchange latencies both within 5G-connected players and between the players and the MEC nodes where the musical information is processed.

MEC nodes are envisioned to host part of the MM platform functionality, such as zero-delay error recovery tasks, allowing offloading expensive computations from embedded end-user devices while ensuring low latency requirements. Simultaneously, if needed, traffic will still be directable from the end devices to a centralized cloud or to the Internet via the 5GUPF. It is necessary that dedicated IP addressing classification mechanisms are put into place at the 5GUPF level to guarantee all the required routing options.

MUSMET will define how 5G MM devices or a centralized MM application controller can leverage the features of the control plane of the 5G Core network (5GC) to request guaranteed Quality of Service (QoS). To implement such a QoS management, we will first investigate a static solution, based on the definition of a "musical User Equipment default profile" within the Unified Data Management function of the 5GC. Then, we will target a context-aware MM device-to-network QoS optimization loop. Depending on the perceived service quality, a MM application will be able to solicit the Policy Control Function of the 5GC and demand dedicated QoS (implemented at a 5GUPF packet filtering level) to support its requirements. The combination of these solutions will lead to the creation of an "MM slice" to accommodate MM services with predefined non-best-effort quality in public or private networks.

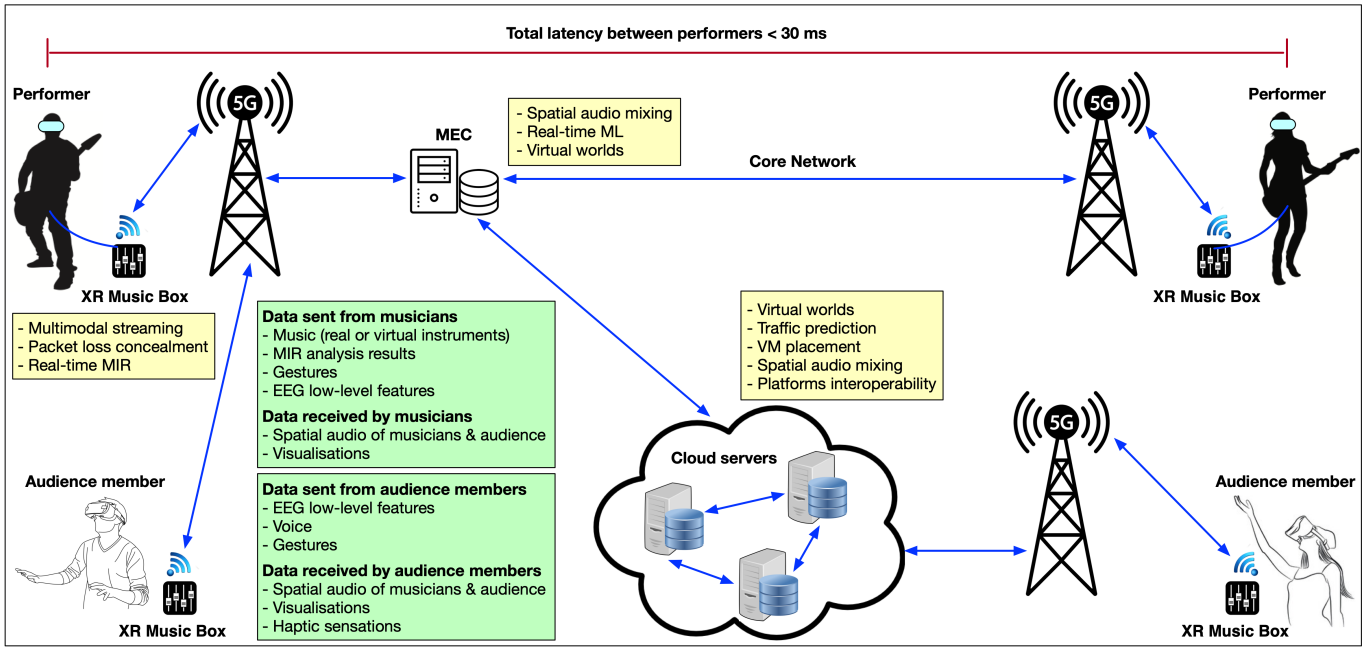

Fig. 2: The MUSMET general architecture.

The key necessity for enhancing the real-time delivery of multimodal data is represented by the ability to swiftly adapt the data delivery to changing network conditions. In this context, AI and ML are both a challenge and a resource. The challenge lies in the fact that MM services will require significant amounts of learning and inference, which in turn will necessitate vast quantities of data and computational resources. At the same time, AI and ML can significantly enhance the management of the network and improve its performance.

Specifically, we will opt for methodologies combining highly adaptive online ML models to develop highly adaptable real-time ML-based models for traffic prediction tailored for MM slices. The models will operate at various time scales and proactively trigger adjustments in the multimodal streaming parameters to improve the Quality of Experience (QoE).

Furthermore, we will develop optimization algorithms for the automatic placement and online migration of virtual functions and/or containerized microservices, with running function instances to operate as backend infrastructure for the support of client/server-oriented multimodal streaming. The virtual functions will be allocated in Edge Computing nodes located in the proximity of the participants in MM-based sessions. The optimization algorithms will permit to reduce the network delay component by exploiting the above-mentioned ML-based traffic prediction techniques, which will enable them to operate proactively and to dynamically adjust virtual functions locations depending on the current and forecast network congestion conditions. Especially useful for the orchestration of real-time resource-intensive MM applications will be the use of recently proposed frameworks for the end-to-end optimization of next-generation media applications over the edge-cloud continuum, which go beyond virtual network embedding formulations and allow jointly optimizing function placement, mixed-cast flow routing, and multidimensional resource allocation [38].

In real-time multimodal streaming, jitter and packet losses due to transmission errors may cause data portions to arrive too late – or not arrive at all – to be reproduced in due time at the receiver's side. In turn, this generates audible glitches (in audio streams), blurring (in video streams), deterioration of the rendered gestures (in motion data streams) or of the perceived haptic feedback (in tactile data streams). We will devise novel Packet Loss Concealment (PLC) techniques specifically tailored for the MM multimodal streams [43] while ensuring compatibility with the stringent low-latency requirements of MM interactions.

A significant unresolved issue is the synchronization among the concurrent multimodal streams generated by MM devices, which is crucial for maintaining a seamless and immersive experience. The challenge is compounded by the diversity of data types involved, including EEG, sound, haptics, and avatar control signals like gestures, each requiring unique processing approaches. Our proposal is to develop a novel protocol specifically designed to synchronize these varied data types efficiently. The development of this protocol will be informed by perceptual research aimed at determining the acceptable limits of temporal discrepancies across different sensory modalities (e.g., the tolerable delay between a sound and the corresponding visual action of an avatar).

It is necessary to address issues of data ownership to ensure that users feel comfortable when participating in MM-enabled activities, especially when biometric signals are tracked. MUSMET will adopt a privacy-by-design approach

[44], incorporate privacy impact assessments into the design stage of MM devices and platforms, and rely on privacy-preserving federated learning methods. MM ecosystems must ensure that data is safely transferred between nodes and cloud infrastructure. We will secure IP-based communication with Software Defined Wide Area Networks encryption capabilities while simplifying also the network management layer.

V. Discussion and Conclusions

The plethora and diversity of open challenges identified in Section II highlight that only a highly interdisciplinary approach will enable advancing the state-of-the-art in social immersive musical interactions. To the best of our knowledge, MUSMET represents the first concerted effort to concretely address the existing barriers preventing the implementation of a Metaverse supporting musical activities.

Nevertheless, the technical challenges need to be grounded in the actual user needs and ethical concerns for the MM to be successfully implemented. For this purpose, it is paramount that the attention of researchers is directed towards activities that place the end user at the center of the design process. This will allow to define models for the QoE, and in turn, for the QoS.

While there are a number of collaborative musical activities that can be performed in the MM, MUSMET will focus on performance. This is arguably the most challenging case study, as it requires adhering to the strict latency and QoS constraints needed by musicians, where audience members can also interact among each other in immersive spaces.

A significant limitation of current MM implementations relates to the visual components of the experience. While our architecture addresses multimodal data streaming including audio, haptics, and biometric signals, high-fidelity visual capture and transmission remains currently constrained by fundamental technological limitations. This is especially true in scenarios that require a realistic visualization of audience and musicians, for example in the context of MR. Previous work explored technologies and methods developed for telepresence applications such as meetings and office-based collaborative work. This include simplified embodied avatars (e.g., [8], [45]) and real-time point clouds (e.g., [14], [46]). While the first is the most lightweight it makes difficult to reproduce the interaction with physical musical instruments. Moreover, depending on the style and appearance of avatars, it might not be accepted in all context. On the other hand, real-time capture and transmission of volumetric point clouds, while appreciated by musicians [46], [47], cannot currently meet the ultra-low latency requirements essential for musical synchronization, as already showed in preliminary explorations in the musical field [48].

Therefore, MUSMET will adopt an audio-first approach that will try to leverage current solutions for video capture and transmission (from avatars to volumetric video) to understand the limits and capabilities of different approaches. This technical constraint reinforces that truly immersive visual experiences in the MM await breakthrough technologies in ultra-low latency video processing and next-generation network capabilities that MUSMET's foundational research aims to enable.

It is worth mentioning that, due to its social nature and stringent requirements (i.e., coordination, tight timing, balance between precision and expression), music can help tackle technological problems related to the development of the Metaverse in general (e.g., low-latency, scalability, interoperability). Looking at the Metaverse from the point of view of musical technologies will provide practitioners in academia and industry with key insights into what is needed to achieve true real-time activities and support human expression for different Metaverse domains well beyond the musical one. By providing technical solutions to users' concerns and adopting open hardware and open software practices, the MUSMET project aims to contribute to the recent endeavors of steering the XR technological transition and ensuring an open, secure, trustworthy, fair, and inclusive digital environment.

Acknowledgment

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379) and by the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

References

[1] L. Turchet, "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, pp. 1–17, 2023.

[2] C. Rinaldi and C. Centofanti, "The musical metaverse: Advancements and applications in networked immersive audio," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–7.

[3] R. Hupke, S. Preihs, and J. Peissig, "Immersive networked music performance: Impact of extended reality on the quality of experience," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–9.

[4] R. Vieira, S. Wei, T. Röggla, D. C. Muchaluat-Saade, and P. César, "Immersive Io3MT Environments: Design Guidelines, Use Cases and Future Directions," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[5] L. Severi, M. Sacchetto, A. Bianco, C. Rottondi, G. Abbate, A. Paolillo, and A. Giusti, "Remote orchestral conduction via a virtual reality system," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–6.

[6] A. Boem and L. Turchet, "Musical metaverse playgrounds: exploring the design of shared virtual sonic experiences on web browsers," in *2023 4th International Symposium on the Internet of Sounds*. IEEE, 2023, pp. 1–9.

[7] M. Buffa, D. Girard, and A. Hofr, "Using web audio modules for immersive audio collaboration in the musical metaverse," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[8] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, "Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, 2023.

[9] A. Boem, M. Tomasetti, L. Turchet *et al.*, "Harmonizing the musical metaverse: unveiling needs, tools, and challenges from experts' point of view," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2024, pp. 206–214.

[10] A. Boem, M. Tomasetti, A. Gabriele, A. Di Scipio, and L. Turchet, "User needs in the musical metaverse: a case study with electroacoustic musicians," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. International Conference on New Interfaces for Musical Expression, 2024.

[11] J. Park, Y. Choi, and K. M. Lee, "Research trends in virtual reality music concert technology: A systematic literature review," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[12] S. Ppali, V. Lalioti, B. Branch, C. S. Ang, A. J. Thomas, B. S. Wohl, and A. Covaci, "Keep the vrhythm going: A musician-centred study investigating how virtual reality can support creative musical practice," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–19.

[13] R. Schlagowski, F. Wildgrube, S. Mertes, C. George, and E. André, "Flow with the beat! human-centered design of virtual environments for musical creativity support in vr," in *Proceedings of the 14th Conference on Creativity and Cognition*, 2022, pp. 428–442.

[14] N. Amarie, G. Fadda, V. Popescu, J.-M. Ghenta, M. Murroni, and A. A. Simiscuka, "Live feedback for immersive music performances-a case study," in *ACM International Conference on Interactive Media Experiences Workshops (IMXw)*, 2025, pp. 26–29.

[15] A. A. Simiscuka, G. Fadda, V. Popescu, M. Murroni, and G.-M. Muntean, "Immersive live concert: A multi-sensory experience based on real-time lyrics detection from spatial audio data," in *ACM International Conference on Interactive Media Experiences Workshops (IMXw)*, 2025, pp. 7–11.

[16] N. K. Baym, *Playing to the crowd: Musicians, audiences, and the intimate work of connection*. New York University Press, 2018.

[17] G. W. Young, N. O'Dwyer, M. Moynihan, and A. Smolic, "Audience experiences of a volumetric virtual reality music video," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2022, pp. 775–781.

[18] S. Ppali, M. Scorer, E. Ppali, B. Branch, and A. Covaci, "Remote rhythms: Audience-informed insights for designing remote music performances," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 2675–2690.

[19] S. Ppali, E. Ppali, S. Scorer, A. Boem, B. Branch, L. Turchet, C. Siang Ang, and C. A, "The virtual concert-goer: Audience perspectives on remote music performances," *Proceedings of the ACM on Human-Computer Interaction*, 2025.

[20] A. Boem, M. Tomasetti, and L. Turchet, "Issues and challenges in audio technologies for the musical metaverse," *Journal of the Audio Engineering Society*, 2025.

[21] G. W. Young, "Ethical considerations in the production and consumption of music in the metaverse," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–9.

[22] M. Lesaffre, P.-J. Maes, and M. Leman, *The Routledge companion to embodied music interaction*. Routledge New York, NY, 2017.

[23] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[24] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-Enabled Internet of Musical Things Architectures for Remote Immersive Musical Practices," *IEEE Open Journal of the Communications Society*, 2024.

[25] R. Cheng, N. Wu, M. Varvello, S. Chen, and B. Han, "Are we ready for metaverse? a measurement study of social virtual reality platforms," in *Proceedings of the 22nd ACM internet measurement conference*, 2022, pp. 504–518.

[26] C. Cortés, P. Pérez, and N. García, "Understanding latency and qoe in social xr," *IEEE Consumer Electronics Magazine*, vol. 13, no. 3, pp. 61–72, 2024.

[27] 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*, 2024.

[28] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[29] M. Leman and P.-J. Maes, "Music perception and embodied music cognition," in *The Routledge handbook of embodied cognition*. Routledge, 2014, pp. 81–89.

[30] D. Stefani, S. Peroni, and L. Turchet, "A comparison of deep learning inference engines for embedded real-time audio classification," in *Proceedings of the Digital Audio Effects Conference*, 2022.

[31] J. Dürre, N. Werner, S. Hämäläinen, O. Lindfors, J. Koistinen, M. Saarenmaa, and R. Hupke, "In-depth latency and reliability analysis of a networked music performance over public 5G infrastructure," in *Audio Engineering Society Convention 153*. Audio Engineering Society, 2022.

[32] L. Turchet and P. Casari, "On the impact of 5g slicing on an internet of musical things system," *IEEE Internet of Things Journal*, vol. 11, no. 19, pp. 32 079–32 088, 2024.

[33] J. Cáceres and C. Chafe, "Jacktrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010.

[34] A. F. Genovese, Z. Nguyen, M. Gospodarek, R. Pahle, C. Brenner, and A. Roginska, "Holodeck: A research framework for distributed multimedia concert performances," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[35] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally adapted immersive environments for distributed music performances in mixed reality," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[36] A. Carôt, M. Dohler, S. Saunders, F. Sardis, R. Cornock, and N. Uniyal, "The world's first interactive 5G music concert: Professional quality networked music over a commodity network infrastructure," in *Proc.17th Sound and Music Computing Conf.*, 2020, pp. 407–412.

[37] L. Mohjazi, B. Selim, M. Tatipamula, and M. A. Imran, "The journey toward 6G: A digital and societal revolution in the making," *IEEE Internet of Things Magazine*, vol. 7, no. 2, pp. 119–128, 2024.

[38] A. Mauro, A. M. Tulino, and J. Llorca, "End-to-end orchestration of nextg media services over the distributed compute continuum," *arXiv preprint arXiv:2407.08710*, 2024. [Online]. Available: https://arxiv.org/abs/2407.08710

[39] L. Turchet, T. West, and M. M. Wanderley, "Touching the audience: musical haptic wearables for augmented and participatory live music performances," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 749–769, 2021.

[40] C. Rottondi, "Inclusiveness in remote music teaching and networked music performances: Vision, technological enablers and design strategies," *IEEE Communications Magazine*, vol. 62, no. 12, pp. 34–40, 2024.

[41] M. D. Fletcher, "Can haptic stimulation enhance music perception in hearing-impaired listeners?" *Frontiers in Neuroscience*, vol. 15, p. 723877, 2021.

[42] 3rd Generation Partnership Project (3GPP), "5G system enhancements for edge computing," ETSI, Technical Specification TS 23.548, 2024.

[43] X. Wei, Y. Yao, H. Wang, and L. Zhou, "Perception-aware cross-modal signal reconstruction: From audio-haptic to visual," *IEEE Transactions on Multimedia*, vol. 25, pp. 5527–5538, 2022.

[44] S. Gürses, C. Troncoso, and C. Diaz, "Engineering privacy by design," *Computers, Privacy & Data Protection*, vol. 14, no. 3, p. 25, 2011.

[45] L. Bruns, B. Saurbier, T. M. Voong, and M. Oehler, "Presence and flow in virtual and mixed realities for music-related educational settings," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–7.

[46] R. Schlagowski, D. Nazarenko, Y. Can, K. Gupta, S. Mertes, M. Billinghurst, and E. André, "Wish you were here: Mental and physiological effects of remote music collaboration in mixed reality," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–16.

[47] A. Boem, M. Tomasetti, and L. Turchet, "Between immersion and usability: A comparative study of 2d and mixed reality interfaces for remote music making," *International Journal of Human-Computer Studies*, p. 103586, 2025.

[48] L. Turchet, N. Garau, and N. Conci, "Networked musical xr: where's the limit? a preliminary investigation on the joint use of point clouds and low-latency audio communication," in *Proceedings of the 17th International Audio Mostly Conference*, 2022, pp. 226–230.