



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI



Università di Cagliari

UNICA IRIS Institutional Research Information System

This is the Author's *accepted* manuscript version of the following contribution:

N. Amarie, G. Fadda, M. Murroni, M. Alexandru and V. Popescu, "An Architecture for Real-Time Interactive Music Performances," *2025 23rd Mediterranean Communication and Computer Networking Conference (MedComNet)*, Cagliari, Italy, 2025, pp. 1-3.

© 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 collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

The publisher's version is available at:

<http://dx.doi.org/10.1109/MedComNet65822.2025.11103534>

When citing, please refer to the published version.

An Architecture for Real-Time Interactive Music Performances

Nicușor Amarie, Gianluca Fadda, Maurizio Murrone, Marian Alexandru and Vlad Popescu

Index Terms—XR Audio, Networked Music Performance, Real-Time Interaction.

Abstract—We present a system architecture that delivers real-time Extended Reality (XR) music performances. Live shows are captured using multichannel audio and depth-enhanced video, then processed and streamed through a distributed media platform that supports low-latency audio rendering and dynamic scene composition. This approach balances performance with scalability and is suitable for deployment on standard infrastructure.

Our findings demonstrate that the system can bridge physical distances, support audience feedback, and provide an engaging, interactive concert experience that is scalable across diverse environments.

I. INTRODUCTION

The 2020 pandemic served as a significant inflection point in the evolution of remote and interactive audio production workflows, accelerating the development of distributed performance systems under real-world constraints [1], [2]. In response to physical distancing requirements, artists and engineers turned to low-latency networked solutions to preserve ensemble coherence and audience engagement, prompting renewed interest from both industry and academia in the design of resilient, high-performance remote collaboration architectures.

This surge in activity has given rise to a substantial body of scholarly research that examines the multidimensional challenges of networked performance platforms—including audio transport stability under fluctuating network conditions, end-to-end latency thresholds for perceptual synchrony, jitter buffering strategies, and the integration of spatial audio within interactive 3D environments [3]. These studies often adopt a cross-layer perspective, evaluating the performance trade-offs between transport protocols (e.g., RTP/UDP, QUIC, or WebRTC), audio codecs (e.g., OPUS, AAC-LD), and adaptive buffering mechanisms.

Simultaneously, field experiments such as Networked Music Performances (NMP) and large-scale interactive concert streams have expanded in both frequency and scale, facilitated by advances in real-time media pipelines capable of achieving round-trip latencies below 30 ms in ideal conditions [4]. These experiments have been further enabled by widespread access to high-throughput, low-latency broadband infrastructure—including fiber-optic links. 5G SA networks

with URLLC (Ultra-Reliable Low Latency Communication) capabilities, and emerging edge computing frameworks that minimize packet traversal distances [5].

Together, these developments signal a broader technological shift toward persistent, infrastructure-agnostic musical collaboration. The convergence of networking, signal processing, and immersive media design is fostering new paradigms of real-time musical interaction that are no longer bound by the spatial and temporal constraints of traditional venues or broadcast workflows. Such systems pave the way for scalable, low-latency, geographically distributed performance environments that support bidirectional interactivity and immersive audio-visual fidelity across heterogeneous endpoints.

The present study proposes an architecture that supports real-time streaming of live music concerts, enabling users equipped with appropriate hardware—such as Smart TVs with integrated cameras and microphones, personal computers, smartphones, or XR headsets—to actively engage with performers within a shared virtual environment. The key innovation of this approach lies in its bidirectional feedback mechanism, which enhances the overall user experience for both audiences and musicians. End-users can hear and view the performance while simultaneously transmitting real-time reactions, which are spatially rendered and delivered back to the performers. This continuous exchange is further strengthened by the system’s capability to generate spatialized audio in real time, contributing to a heightened sense of immersion and mutual presence across the virtual concert space.

II. EXTENDED REALITY AND LIVE MUSIC

Extended Reality (XR) technologies have begun to transform the way music performances are delivered and experienced. Building on the foundations of Networked Music Performance (NMP), artists in different locations can now participate in shared, synchronous concerts that extend beyond physical venues. This shift has been driven by a growing demand for immersive and interactive formats, accelerated by advancements in network infrastructure, media processing, and real-time rendering platforms.

While the concept of remote performance is not new, achieving a convincing sense of co-presence remains a key challenge. Latency, audio-visual misalignment, and the absence of spatial cues can break the illusion of a shared space. Current research focuses on three main aspects: maintaining low delay across the media pipeline, delivering accurate spatial sound, and supporting interaction loops that allow audiences to respond in real time.

Spatial audio plays an essential role in enhancing immersion. Techniques like head-related transfer functions (HRTF) and Ambisonic rendering help simulate realistic sound fields that change based on listener orientation [6]. These methods

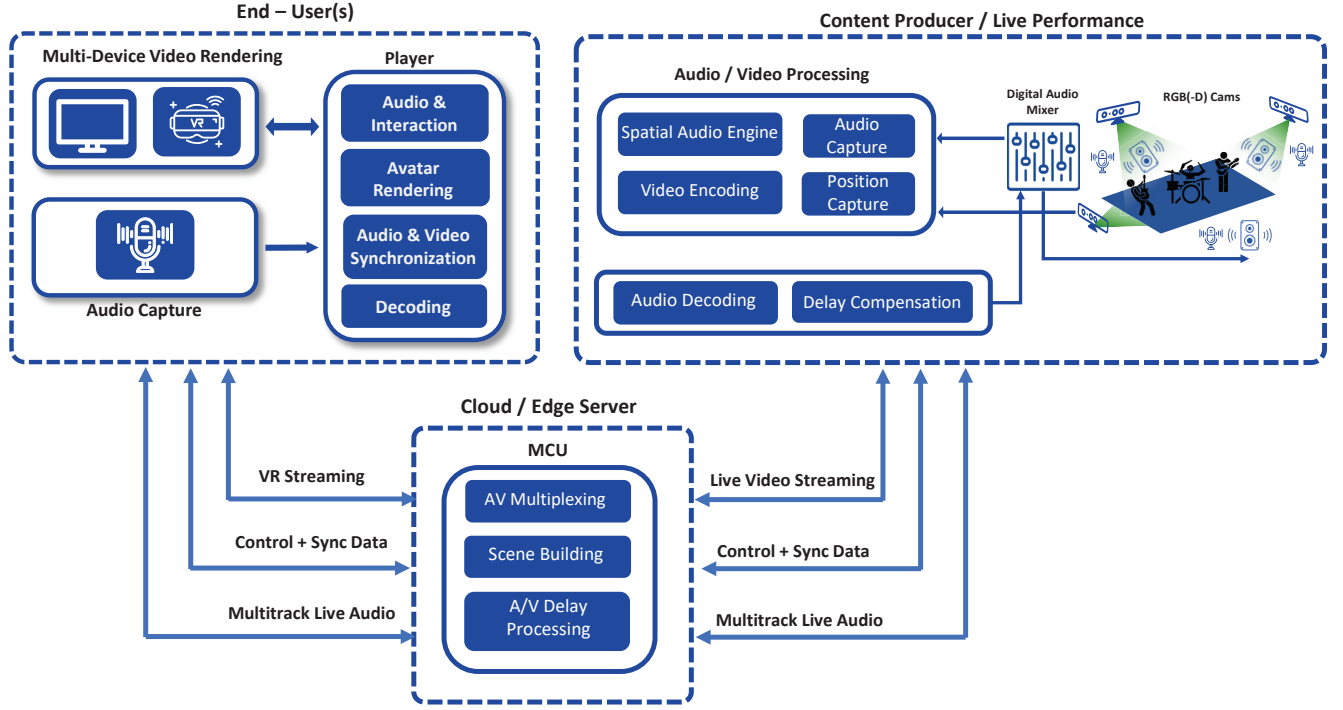


Fig. 1. Architecture for interactive music performances

are integrated into engines that power XR platforms and enable more natural musical interactions across virtual stages.

Latency, however, remains the most critical constraint. When end-to-end delays exceed approximately 30 milliseconds, musical tightness degrades noticeably [7]. To address this, edge computing and 5G networks have introduced new pathways for low-latency streaming by bringing processing closer to both performers and audiences [5]. Combined with lightweight media pipelines and efficient transport protocols, these technologies support smoother synchronisation and faster feedback.

Recent systems also incorporate audience interaction, where feedback like applause or cheering is spatially rendered and sent back to performers with minimal delay. This closes the loop between stage and crowd, strengthening the feeling of shared presence. Such architectures point toward a future in which concerts are not just streamed, but lived together-synchronously and interactively-in fully virtual or hybrid formats.

III. PROPOSED ARCHITECTURE

A. System overview

The architecture supports real-time Extended Reality (XR) performances by structuring the system into three functional layers: *Content Producer*, *Cloud/Edge Processing*, and *End User*. Each layer plays a distinct role in capturing, processing, and rendering the interactive concert experience.

The *Content Producer* tier captures multichannel audio and depth-enhanced video from performers using various sensors and audio inputs. These streams are processed in real time to extract spatial and temporal information, including performer position and direction, which are crucial for immersive rendering.

At the *Cloud/Edge* layer, these media streams are synchronized, encoded, and composed into a lightweight virtual scene. The edge infrastructure ensures low-latency delivery by colocating critical services close to both the performers and audience members. Adaptive scene management and resource orchestration help balance performance and scalability during live events.

The *End User* layer receives the real-time stream and reconstructs the scene using a platform-agnostic rendering engine. This enables users on headsets, TVs, or mobile devices to access an immersive experience tailored to their device. Visual and audio assets are dynamically aligned with spatial metadata to preserve presence and coherence.

B. Audience interaction and feedback loop

A core feature of the system is its support for bi-directional, low-latency interaction between the audience and performers. Users are represented as avatars or interactive entities within the virtual venue and can express reactions through voice input, gestures, or gaze. These signals are captured, spatially processed, and routed back to the performers in near real time.

Different client types contribute interaction data in various forms. Head-mounted displays stream continuous pose and motion information, while flat-screen devices use camera-based gaze detection to infer attention and intent [8], [9]. Audio feedback, such as clapping or cheering, is captured through device microphones, tagged with positional metadata, and delivered to the performer mix with minimal delay.

The interaction loop closes as performers hear spatialized crowd reactions within their monitoring system, creating the impression of a live audience present in a shared virtual space. This feedback mechanism enhances performer engagement and enables more dynamic, audience-aware shows. By simulating the emotional and auditory cues typical of a physical

concert environment—such as applause, cheers, or laughter—delivered in real time and mapped to the virtual space, performers can adapt their delivery, pacing, and intensity in response to perceived audience energy. This not only fosters a stronger sense of presence and immersion but also bridges the psychological gap often encountered in remote or virtual performances, reinforcing the authenticity of the experience for both performers and viewers alike.

C. Implementation challenges and solutions

Maintaining real-time performance across the system introduces several key challenges. First, *latency* must be minimized to preserve musical synchronicity. Studies show that delays beyond 30–40 ms negatively impact ensemble timing [4], [7]. To address this, media processing is anchored at the edge, and low-delay codecs such as Opus are employed [10], [11].

A second challenge is *bandwidth efficiency*, especially when supporting volumetric video and high-fidelity spatial audio. To mitigate overload, the system adopts object-based audio encoding and scalable video delivery. Ambisonic rendering [6] and MPEG-H 3D Audio [12] help compress complex sound scenes without compromising spatial accuracy.

Finally, *synchronization* of audio, video, and interaction cues is vital for sustaining immersion. Timestamping and centralized clocking ensure that crowd responses are injected into the performance stream within the same perceptual window. Optional haptic cues further expand sensory feedback while remaining within the system’s timing constraints.

Together, these solutions enable a robust and interactive XR concert system that supports multi-user engagement with minimal latency, adaptable quality, and spatial coherence.

IV. CONCLUSIONS

This paper outlines the design and initial implementation of an edge-native architecture that enables real-time streaming of live music performances into a shared extended reality (XR) environment. The system allows remote audiences to not only observe but also actively interact with on-stage artists, fostering a sense of co-presence and mutual engagement. We introduced a proof-of-concept pipeline that integrates multitrack audio capture and a bidirectional feedback loop between performers and audience members. Our implementation demonstrates that spatially accurate audio streams can be delivered with low end-to-end latency, maintaining both bandwidth efficiency and scalability for high numbers of concurrent users.

One of the key contributions of this work is the realization of audience-to-stage audio loops—such as localized applause, cheering, and collective singing—which are rendered in spatial context and fed back to the performers in their monitoring system. This interaction paradigm significantly enhances performer responsiveness and deepens the overall emotional resonance of the virtual concert experience. The system’s architecture addresses complex technical challenges, including low-latency audio delivery, efficient bandwidth utilization, and real-time synchronization of multisensory data across distributed participants.

By addressing these core issues, our research contributes to the evolving field of immersive networked music performance, offering a scalable and inclusive model for future XR-based concerts. Looking ahead, the future work will

concentrate on comprehensive evaluation of subjective Quality of Experience (QoE) metrics from both audience and performer standpoints. We are also refining the orchestration and resource management layers to support multi-tenant scenarios, such as simultaneous concerts across different stages or venues, thereby paving the way toward richly interactive and socially connected live performance ecosystems.

ACKNOWLEDGMENT

The research activities described in this paper were carried out within the “HEAT – Hybrid Extended reALiTy” Project GA 101135637 funded by the EU Horizon Europe Framework Programme (HORIZON).

REFERENCES

- [1] R. Hamilton, “Real-time musical performance across and within extended reality environments,” *The Journal of the Acoustical Society of America*, vol. 153, pp. A35–A35, 03 2023.
- [2] S. Giacomelli, C. Centofanti, J. Santos, M. Galbiati, T. Salvi, F. Graziosi, and C. Rinaldi, “Remote immersive audio production: State of the art implementation, challenges, and improvements,” 09 2024, pp. 1–10.
- [3] L. Comanducci, *Intelligent Networked Music Performance Experiences*. Cham: Springer International Publishing, 2023, pp. 119–130. [Online]. Available: https://doi.org/10.1007/978-3-031-15374-7_10
- [4] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, “An overview on networked music performance technologies,” *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
- [5] 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*, vol. 5, pp. 4691–4709, 2024.
- [6] Wikipedia contributors, “Ambisonics,” <https://en.wikipedia.org/wiki/Ambisonics>, 2024, accessed: 2024-12-11.
- [7] A. Carôt and C. Werner, “Network music performance-problems, approaches and perspectives,” in *Proceedings of the “Music in the Global Village”-Conference, Budapest, Hungary*, vol. 162, 2007, pp. 23–10.
- [8] S. Porcu, A. Floris, M. Anedda, V. Popescu, M. Fadda, and L. Atzori, “Quality of experience eye gaze analysis on hbbtv smart home notification system,” in *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2020, pp. 1–6.
- [9] S. Porcu, A. Floris, and L. Atzori, “Towards the prediction of the quality of experience from facial expression and gaze direction,” in *2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 2019, pp. 82–87.
- [10] J. Spittka, K. Vos, and J.-M. Valin, “RTP Payload Format for the Opus Speech and Audio Codec,” RFC 7587, Jun. 2015. [Online]. Available: <https://www.rfc-editor.org/info/rfc7587>
- [11] Chris Chafe and Juan Pablo Cáceres, “Jacktrip: A system for real-time network audio streaming,” <https://ccrma.stanford.edu/groups/soundwire/software/jacktrip/>, 2023, accessed: 2024-12-11.
- [12] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “Mpeg-h 3d audio—the new standard for coding of immersive spatial audio,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, 2015.