



THE TECH BEHIND IMMERSION:

Technical learnings from the 2026 IBC Accelerator “IFeL: Immersive Festival Live - Remote Presence at Scale”

2026 TEAM CONSORTIUM:

IFeL Champions: BBC R&D · ITV Studios UK · Reality Hack at MIT · King’s College London · University of Galway · SMPTE (Society of Motion Picture and Television Engineers) · EBU (European Broadcasting Union) · University of Southampton · XPLOR (Academy of Live Technology)

IFeL Participants: Shure · Magic Beans · The Third Orchestra · Surround Sync · RePro Stream · Beam Immersive

INTRODUCTION

To create immersive experiences that are convincing and enjoyable, technology needs to be emphasised. This spans the entire “glass-to-glass” broadcast pipeline, and each step must be reconsidered for live broadcasts of these immersive experiences. Understanding the desired experiences drove the technology requirements and developments; however, our multi-disciplinary IBC Accelerator consortium of global broadcasters, academic institutions, and technology vendors learned several technical lessons along the way while trying to create those experiences, which we will apply to future cutting-edge POC explorations and real-world immersive experiences and broadcasts.

01 AUDIO CAPTURE

Audio capture learnings focused on capturing and reproducing spatial audio alongside close-mic or direct-line audio to create a variety of novel immersive experiences while still delivering traditionally expected audio content.

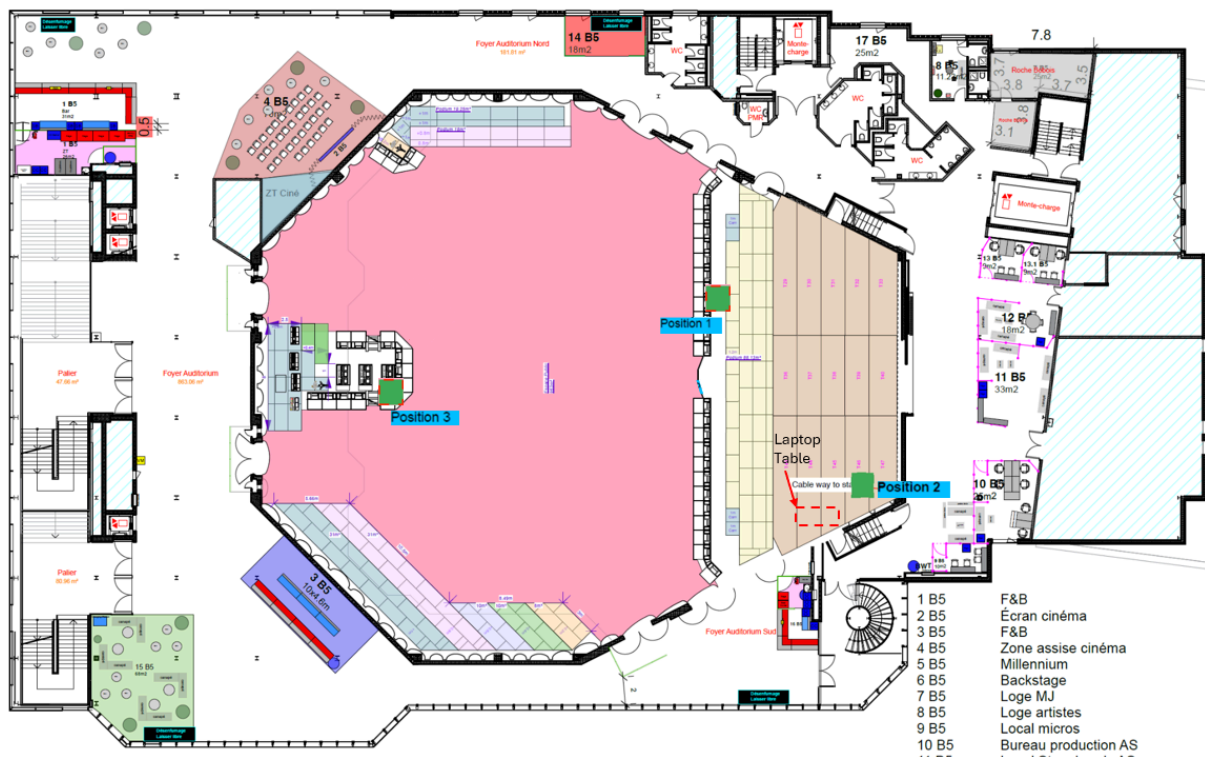
The first full system test was at the Montreux Jazz Festival, where we captured spatial audio and mixed it with Front of House (FOH) instrument/vocal microphone audio. Spatial audio capture devices were placed throughout the venue in 3 positions: (1) On-stage, (2) in the Press Pit, and (3) mounted on the ceiling. FOH live tracks were spatialized and combined with the spatial audio to create a full spatial broadcast.

From there, we produced and mixed 2 audio formats with video for broadcast: a 5.1 surround sound output and a binaural output, presented via an OPSODIS sound bar. The broadcast spatial audio was mixed from the perspective of the On-Stage device.

The on-stage device was paired with the on-stage POV that was broadcast. This captured crowd and venue ambience. We mixed the ambience as a “room mic” with the FOH tracks to convey a sense of reality and venue characteristics.

The Press Pit device was recording right in front of the audience and in front of an on-stage monitor wedge. This device picked up crowd noise close to other concert-goers, giving the auditory impression of being entrenched in the crowd.

Image: Layout of Stravinski Hall, at the Montreux Jazz Festival with each camera position shown.



The ceiling-mounted device captured a blanket of crowd noise but also picked up the line-array sound reinforcement. This is a common issue when capturing/broadcasting music festivals or sporting events, but it still needs to be considered for future work. Listening back, this was

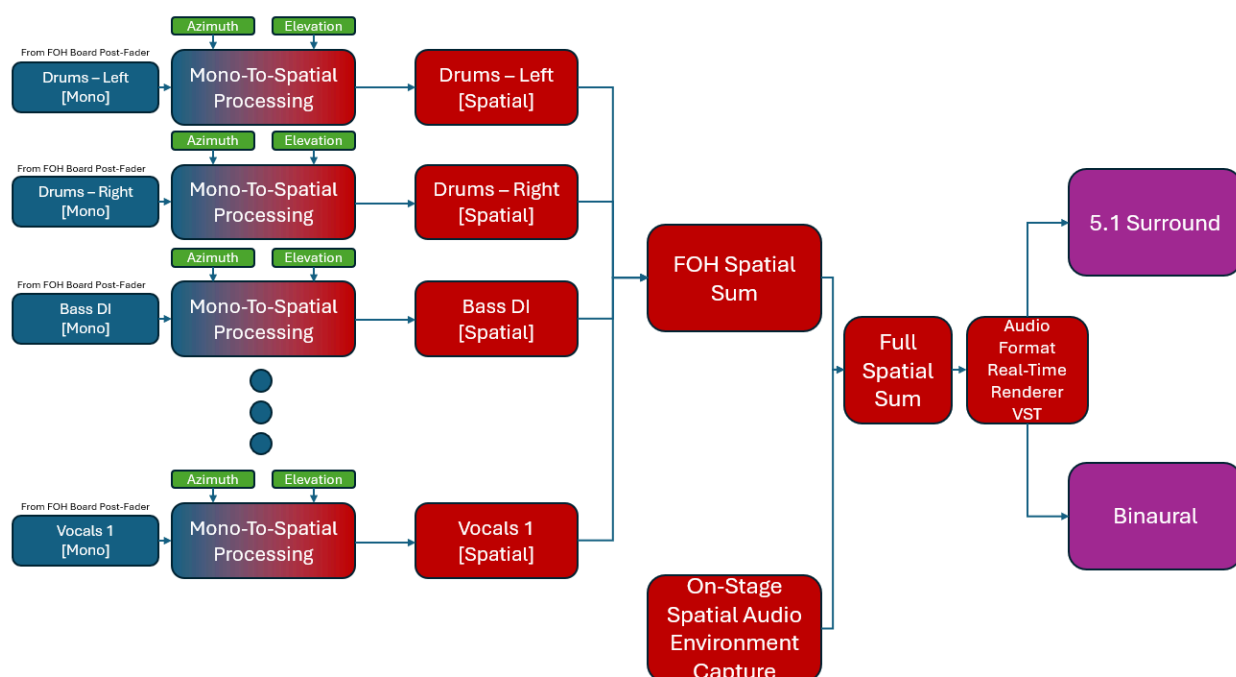
picking up fantastic crowd pop sounds, particularly on “Dean Town”, where the crowd was singing with the band.

We pulled FOH board tracks post-fader (post-FX) and spatialized them to combine with the On-Stage spatial audio capture device. We spatialized the tracks to roughly reflect instrument and musician placement according to the On-Stage Camera POV. This was imperative for a pleasant and grounded audio experience.

It was important to match the perceived positions of each individual player’s audio source in virtual space to their real positions captured by spatial audio environment capture. This helped align the “mixed” audio to the “real” audio in space. For example, if a player/audio source is to the left of the spatial audio capture device, their close-mic signal must be perceived from the left as well; otherwise, the system would create 2 copies of the same audio source in 2 locations, even though there is only 1 player/audio source.

The spatial capture audio worked well as “context glue,” enhancing the close-mic audio in the broadcast. It captured the sounds of that specific venue (Stravinski Hall) and gave context to the performance, creating the perception of being there.

Image: Audio signal and processing path for capture of Vulfpeck at the Montreux Jazz Festival



While a singular point-of-view mix was used for the Montreux Jazz Festival live broadcast proof-of-concept, a spatial mix was created for each camera position to reflect the changing audio scene. This was used in the post-production proof-of-concept, where all captured audio and video were put together to create a multi-POV experience consumed as a VOD (video on demand).

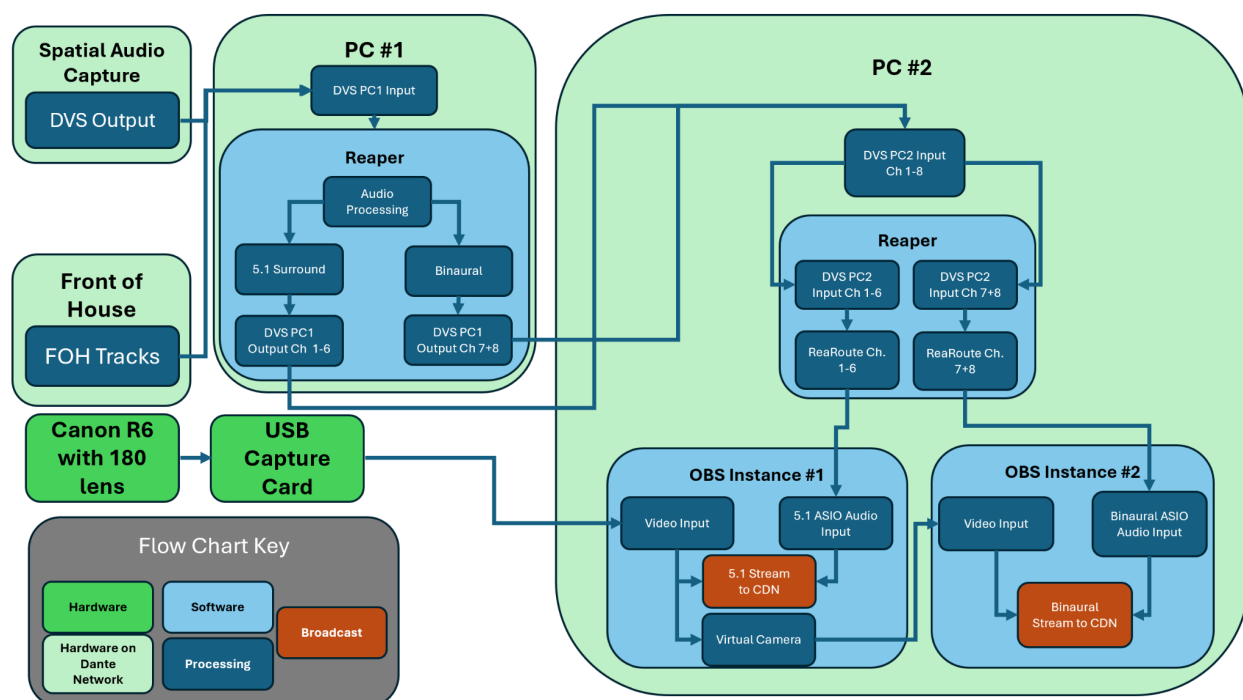
At King's College, the Third Orchestra performance, while not broadcast live, was captured in a similar way - spatial audio was captured for venue "context glue" and was mixed with FOH audio. One major difference was the redundant capture of individual performers.

Along with Instruments and performers being close-miked (FOH audio), we used a Shure DCA901 Broadcast Microphone Array as a reduced-setup capture solution. The DCA901 used array-based lobe steering to isolate musicians into their own tracks, without the need to set up multiple XLR microphones. These lobes from the DCA901 microphone could create FOH audio for smaller events, replacing or supplementing close mics while providing an easy-to-use setup workflow that required only a single physical connection to the device but delivered multiple output channels. Various audio captures were used in visual POCs within the project.

02 BROADCAST AND PLAYBACK

On-site at the Montreux Jazz Festival, a spatial audio mix was constructed and splashed out to desired broadcast audio formats. This mix was sent to another on-site computer responsible for muxing and broadcasting the respective audio formats with a stereoscopic video feed from the On-Stage camera. It is worth noting that there was a camera also placed in the Press Pit and at the FOH position, but these cameras were for capture only, as the team did not have the broadcast infrastructure alone to use these together to create a multi-POV experience live; that capture was used later to show what a multi-POV broadcast could look like given the appropriate resources.

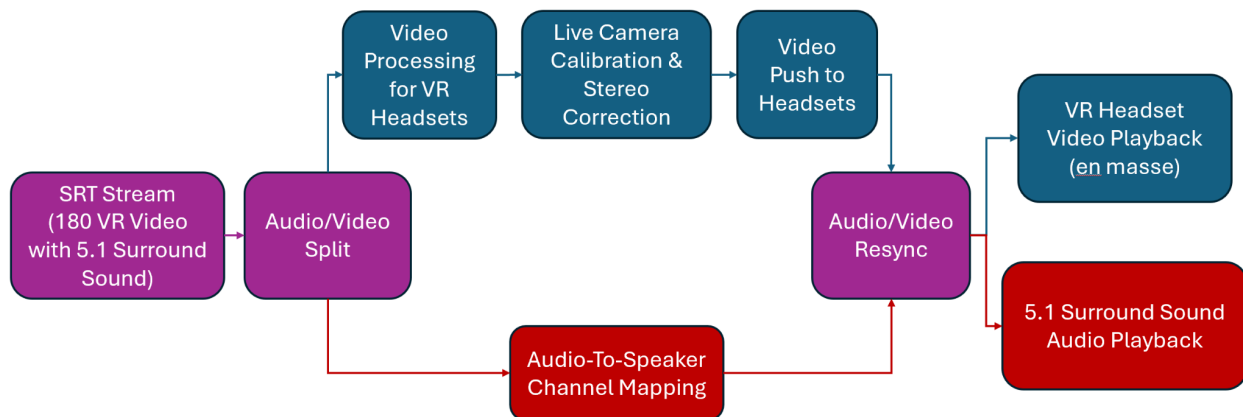
Image: The basic broadcast workflow IFeL used to send an immersive broadcast to a CDN.



The broadcast was constructed using a crude combination of Dante Virtual Soundcard (DVS), Reaper, and Reaper's ASIO driver "ReaRoute" to handle the audio flow into two instances of Open Broadcast Software (OBS) (OBS can only broadcast one audio format at a time, and the team wanted to avoid channel allocation confusion on the broadcast's receiving end) to create two SRT streams (one for the Binaural audio format and one for the 5.1 Surround Sound audio format) to a Content Delivery Network (CDN). This workflow is not representative of an actual broadcast, but it allowed video and audio to be pushed from Stravinski Hall to a CDN for processing and consumption in London, England, and Perth, Australia.

Playback occurred in rooms set up with multi-channel audio outputs and VR headsets. Each attendee received a VR headset that played back immersive video, while the room's surround sound system played back the audio. This would be one way to watch the broadcasts at scale, an alternative to personal viewing on one's VR headset at home or in a shared space like Cosm (theatres designed for large-scale immersive broadcast viewing of sporting events). Additional video processing was needed to convert the stereoscopic video into fully unwrapped video for viewing on VR headsets. In order to do this in near real-time for live viewing, the audio stream and video stream were separated, the video stream was processed - which included camera calibration and stereo correction - and then resynced with the audio before finally being played in the headsets and over the surround sound system in the room.

Image: Broadcast conditioning for at-scale VR playback in a shared space.



The final experience let many people view the broadcast at once, allowing them to look wherever they wanted while maintaining synchronised spatial audio playback. As stated before, this would work as a to-the-masses experience, but the equipment available to the team hindered it.

On-site in London and Perth, synchronization between the audio and video was incredibly important. Although the audio and video were received together, video processing introduced latency that would have broken the experience's immersion. We noted this in London, where we

used an older set of playback hardware. The older hardware could not correctly synchronize the audio and video without deviation from the setup in Perth. As a result, we learned that careful selection of playback hardware is needed to achieve this type of playback experience.

While the Montreux Jazz Festival live proof-of-concept was a singular, fixed point of view, the post-production proof-of-concept showed that several points of view could be used (closer to a typical broadcast), each one accompanied by its own spatial audio mix to accurately represent audio source locations.

Because our team was dispersed across the globe, the initial ambition was to have an Insta360 camera *streaming* the *Third Orchestra* setup and PoC days. For Day 1, we quickly learned that you need more than this to embed auxiliary audio effectively - you need software such as OBS. We took an analogue/phono output from the sound desk for the mix and a USB-C link from the Insta360 for video, then used OBS on a MacBook Pro to embed both and stream to a private YouTube channel with several seconds of latency.

This provided a 180-degree video and immersive live stream for our remote participants to view, but it was clear that this was not going to be what would be needed for Montreux - so that is where SurroundSync stepped in - who do this type of live, ultra-low-latency streaming for large public music events/festivals.

Several lessons: 360 video cameras don't natively handle spatial audio (or even an aux audio source on some occasions). They come in at various price points for each use-case - from a few hundred pounds, to middle-tier several thousands (£10k-12k), to the really high-end professional grade (£20-30k). Lighting levels are a key consideration and can and will often be a challenge at concerts - so the bigger the lens/lenses, the better, as is connectivity - also something that can be contended at the concert and is also an unknown/uncontrollable variable at the recipient's location. Pre-recorded/post-produced 360 experiences remove much of this - but our ambition was always to evolve the Live capability.

For the Third Orchestra proof-of-concept at Bush House, the ensemble performed live in the main auditorium while the microphone signals from the stage were carried over Dante to a separate part of the building. From this matrix of microphone signals, the performance was reconstructed as a real-time free-roam spatial soundfield for a remote audience wearing tracked headphones. Each participant carried an Android phone running a real-time audio simulation, with their position and head orientation continuously driving their individual binaural render. As listeners moved freely through the playback space, the reconstructed soundfield was mapped onto the physical environment, with the musicians and other sound sources remaining fixed in space while the listener's perspective changed continuously with their movement.

This form of 6DOF playback places particular requirements on both capture and reconstruction. Close source capture is valuable because it provides sufficiently direct material for individual

performers to be positioned and rendered dynamically for each listener. At the same time, spatial and ambient capture of the performance space retains important information about the original event's acoustic character. Together, these elements contribute to the whole reconstructed soundfield. The direct source material can be positioned and rendered according to the listener's location, while captured ambience and real-time acoustic modelling contribute to the surrounding spatial and reverberant field.

King's College London's real-time acoustic simulation generated listener-dependent reflections and reverberation within the reconstructed environment. This highlighted an important capture consideration: acoustic information already present in the microphone signals has to remain compatible with acoustics subsequently introduced during rendering. Recordings with a high proportion of room sound give less freedom to alter source position, distance or acoustic environment convincingly. Conversely, sufficiently direct source capture provides greater control over reconstruction, while room and ambient capture can retain the character and spatial impression of the original venue.

Multi-microphone capture also introduces important inter-channel relationships. Each microphone inevitably contains some contribution from sources other than its primary source, and across the microphone array those contributions have relative level, phase and arrival-time relationships determined by the layout of the original performance. The spatial information therefore exists not simply within individual microphone feeds, but in the relationships between them. When those feeds are repositioned within a reconstructed soundfield, their spatial placement and timing cannot be considered independently. Significant inconsistencies between the original capture layout and the reconstructed layout can reduce localisation accuracy, introduce colouration or produce a less coherent auditory scene. Maintaining plausible spatial and temporal relationships helps the listener's natural auditory system resolve the reconstructed environment as they move through it.

Tracking quality is equally important. Because virtual sources are intended to remain fixed within the physical playback space, errors or latency in either position or head orientation are immediately translated into movement of the perceived soundfield. Stable absolute tracking and low end-to-end latency are therefore fundamental to maintaining a convincing spatial illusion.

Four listeners experienced the Bush House system simultaneously. The source audio and scene description can be shared, but the final binaural rendering must be listener-specific because each participant occupies a different position and orientation. Scaling therefore requires additional tracking and rendering resources for each participant rather than simply distributing a single pre-rendered spatial audio stream.

03 VIDEO CAPTURE AND VIRTUAL RECREATION OF SPACES

We captured and used various video types across the proof of concepts, including 180-degree immersive stereoscopic, 360-degree, and multi-camera array video. While 180- and 360-degree video is simpler to work with, we learned lessons about using it to create immersive experiences.

These cameras lack features most other broadcast cameras have, including zoom. Because of this, you must carefully consider camera placement. Too close and filmed subjects will feel uncomfortably close, large, and warped. Too far, and the subjects will feel distant. There is a “sweet spot” at which these cameras create ideal experiences when played back. These issues cannot be fixed in post-production either, so finding the “sweet spot” is critical when planning an experience. Because placement matters so much, ideal camera placement can be inconvenient for the event being captured.

For example, the Press Pit camera in the Montreux Jazz Festival test was too close to the performers when they approached the lip of the stage. As a result, the musicians looked unrealistically large and warped. The ideal placement of this camera would have been about 0.5 to 1 meter back from its original position, which would have put the camera in the audience. This is not practical because it disrupts the experience of those attending the event in person. Immersive cameras may need additional hardware upgrades to be placed in reasonable locations while maintaining correct perspectives.

Another issue was working with large amounts of data. The real-time Montreux Jazz Festival broadcast used lower video quality to keep the data transmission package size reasonable. Unfortunately, immersive video needs very high resolution because it is unwrapped and stretched into an equirectangular VR format. While the team's resources were limited, any serious broadcast will need a way to reliably handle large data packages. While this is not an uncommon problem to solve, it deserves extra emphasis given how the video data is processed when the final consumer endpoint is a VR headset.

As we look towards the future of real-time volumetric capture, we're finding a whole new set of challenges. Rather than placing 2D or stereoscopic 180°/360° video into a VR environment, volumetric techniques let us capture the real world in three dimensions as it changes over time. The aim is to let people move through an event as it happens, getting ever closer to the experience of “being there”.

We recently had the chance to put this to the test on the set of *The Voice* in the Netherlands. Access to the set was a real privilege, and we greatly appreciated the effort the ITV team put into making this R&D capture possible.

Faithfully recreating an event takes many cameras. The set, other performers and even the audience can block our view, leaving gaps that a handful of cameras simply can't cover. AI techniques can help fill those gaps through prediction, averaging or generative fill, but there's a fundamental limitation: they're guessing what's missing, rather than capturing what's actually there.

That matters in live events, where authenticity is such a big part of the appeal: a real thing happening right now, in front of you. Artists also go out of their way to surprise their audiences—to be unpredictable. If we want to capture that experience, we need to represent what really happened.

Authenticity is critical to us for the future, so we're committed to a faithful capture that needs lots of different viewpoints. It's essential so we can see everything. We're not inventing details. This is certainly one of the biggest challenges in volumetric video. It's why so many demos focus on small objects or limited scenes, or they lean heavily into generative AI. The cost and computing power needed to capture something larger are significant, even when the processing happens afterwards.

But technology is changing what's possible. We now have a deployable data centre cluster, and we felt it was worth taking one of our prototype servers and a large camera array into a live environment. We wanted to move beyond lab conditions and start understanding the practical challenges. This was R&D in the field: minimal setup time, limited preparation and calibration and having to cope with the real world.

Most importantly, it reminded us that flexibility is critical. Data centre computing systems aren't designed to be deployed quickly in this kind of environment. Simply switching the servers on can take 15 minutes, turning a five-metre rack move into a 25-minute job. Small delays like these quickly added up, limiting the time we had to deploy and properly calibrate the recording system.

Real-world conditions are a great teacher. We'd already found that our original approach to real-time point-cloud generation using lidar didn't work with lighting beams in haze, so we're now using only computer vision to ensure accuracy. What surprised us was the lens flare: it created different artefacts in different cameras, leading to errors in the 3D reconstruction. We'd always controlled the lighting ourselves before, so we hadn't encountered this.

Vibration from the sound system also affected the continuous calibration, and we had problems with the remote preview system during setup and focusing. We realised we need to automate more of the camera setup and calibration for locally supplied broadcast cameras, so we can integrate them more quickly alongside our pre-calibrated cameras.

The automatic segmentation didn't behave as expected with this specific content either, which prevented the real-time processing we'd hoped for. We're now processing the recordings afterwards, addressing some of these issues for next time and continuing to learn how to improve both quality and performance. It's an exciting time, even when the lessons are hard-earned.

The footage from this experiment will need significant post-processing to become good content. But the test has shown us where we need to improve to make this a truly groundbreaking technique for broadcast and immersive experiences. We suspect we'll need a few more rounds of this kind of R&D before the system is ready to be used—and relied on—in the wild.

We're not content with developing a system that simply works. We're pursuing authentic cinematic quality, built from what was actually present. Those details make the difference between a generic experience and something that really is 'like being there'.

MOVING FORWARD

Each technology block discussed here is part of a greater system, and combining these blocks is integral to creating immersive experiences. These blocks need more development, and future efforts will be needed to create a seamless, consumable experience. That said, many types of immersive experiences can branch off these blocks, each with its own requirements.

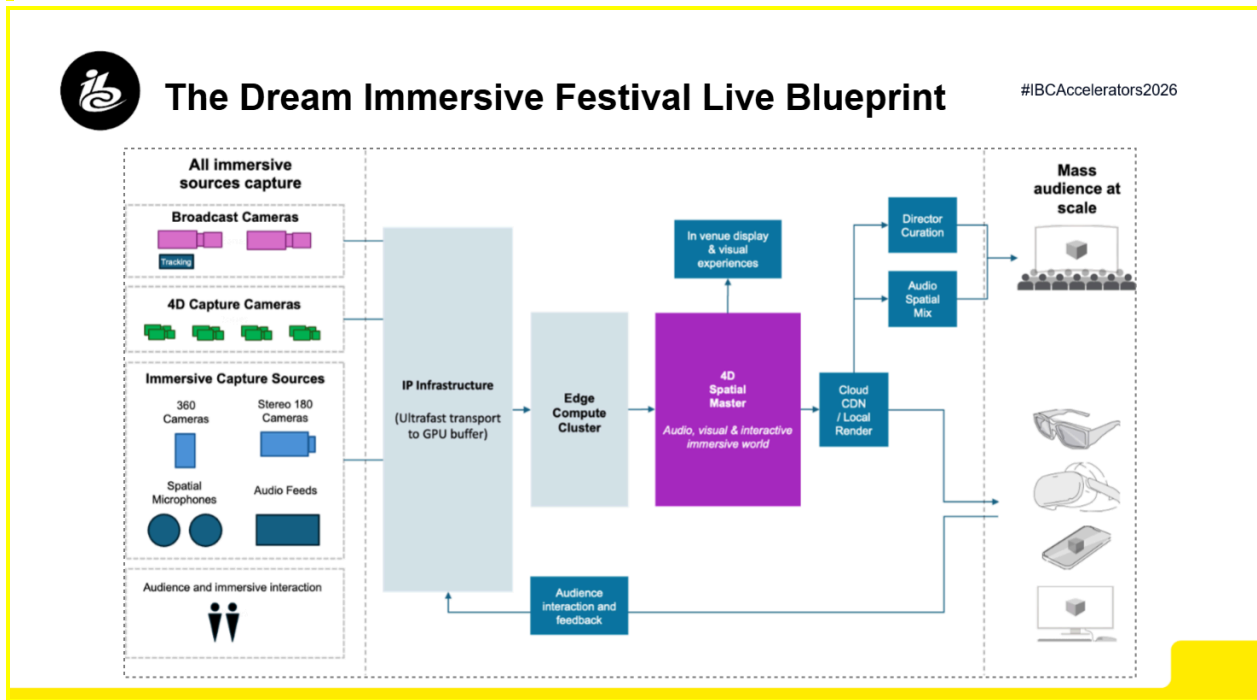
Moving forward, we will need to develop and make accessible additional hardware and decoding software focused on high data transfer, immersive capture, and mixing/muxing immersive video and audio to create and curate full experiences. This would also include creating dedicated spaces to work with live immersive media and prepare it for broadcast. While many solutions exist, we must emphasise how they help create immersive experiences for remote audiences to consume, either as individual consumers or as consumers in shared spaces.

We will also need to expand on-site compute capabilities for processing and conditioning before broadcasting. Edge compute is essential to achieve the bandwidth and latency needed for real-time immersive world construction. Then we can broadcast the much smaller constructed spatial master. Additional processing would allow much more complex, viewer-led experiences to be created and hosted for events. The easiest adoption will be at stationary events, i.e., events that are not regularly torn down, moved, and set up in new locations. In event spaces like studios, additional space can be allocated for computing power to condition media for broadcast before that media is used to create the actual broadcast.

Tying technology blocks together will require additional support and input from partners and event venues. In the scope of the Accelerator projects, a natural limit will be reached regarding

how far each tech block can be proven out before it becomes more of a full installation or product development. Regardless, this will be needed to drive the broadcast industry forward towards creating live immersive experiences for remote participants.

Image: IBC IFeL consortium's future-facing workflow utilising multiple 2026 POC results.



IFeL Champions: BBC R&D · ITV Studios UK · Reality Hack at MIT · King's College London · University of Galway · SMPTE (Society of Motion Picture and Television Engineers) · EBU (European Broadcasting Union) · University of Southampton · XPLOR (Academy of Live Technology)

IFeL Participants: Shure · Magic Beans · The Third Orchestra · Surround Sync · RePro Stream · Beam Immersive