IOWN GLOBAL FORUM ™

Reference Implementation Model (RIM) for the Interactive Live Music Entertainment Use Case

Classification: APPROVED REFERENCE DOCUMENT

Confidentiality: PUBLIC

Version 1.0

[2023/02/13]

Legal

Contents

List of Figures

List of Tables

Executive Summary

The IOWN Global Forum (IOWN GF) provided prominent future-looking use cases and identified application-specific service requirements benefiting users from different vertical industries in the use case reports at Cyber-Physical System Use Case Release-1 [IOWN CPS UC] and AI-Integrated Communications Use Case Release-1 [IOWN AIC UC].

A unique approach has been taken in the IOWN GF to create the Reference Implementation Model (RIM) for implementing these future-looking use cases by leveraging Data-Centric Infrastructure (DCI) technologies [IOWN DCI] and Open All-Photonic Network (APN) technologies [IOWN Open APN], such as advancing the necessary technological development using the DevOps methodology.

The purpose of this RIM work is to develop and evaluate the RIMs iteratively. The RIMs provide guidance for the practical implementation of these technologies and their combinations for the overlay solution targeting specific CPS/AIC Use Cases. These are represented by the Benchmark Models the IOWN GF has developed.

This document delves into the RIM for the Interactive Live Music Use Case (ILM UC) as defined in the IOWN AIC UC. This is one of the future-looking VR use cases that involves a live music concert in virtual space with massive number of users. In the early phases, the service would start at a smaller scale for “power users”. Then the service will scale by enabling all kinds of users to subscribe to this ultra-realistic service via consumer-grade devices harnessing the power of distributed pooled computing resources over low-latency connections.

To provide an extreme immersive and interactive user experience, a new network and computing environment with higher capacity and lower latency is needed. Through development of the RIM for this use case, we investigate to what extent these user experiences are realized by IOWN GF technologies and identify the best practice of the system design. This also aims to demonstrate many of the benefits of IOWN GF architecture and technology, which we call the IOWN GF architecture and technology in the remainder of this document, over today’s central-cloud-based implementations.

ILM UC specific requirements include the following features;

- Since the service described in the ILM UC provides the 6 DoF (Degree of Freedom) view for individual Display Devices, it’s possible that some users could experience motion sickness. As a result, a key requirement to providing a superior VR experience must include low motion-to-photon latency between the sensor devices for head axis and eye vectors and the Display Devices in the HMD. This requires a high-bandwidth, ultra-low latency network to process the rendering for a display to the Audience Member between an Audience Member’s site and an edge data center.
- The low end-to-end latency between capturing the Audience Member’s motion and reflecting it to the Audience Member and the Artist’s displays and vice versa is also a key requirement since the service provides Audience Member with the interactive user experiences such as waving hands and shouting each other with other Audience Members and Artist. This requires ultra-high-speed interconnection between edge/cloud data centers and ultra-high-speed interconnection between functional nodes in data centers as well as an ultra-low latency access network.
- Since this service will scale from very small number of Audience Members to massive large-scale events with huge numbers of Audience Members, dynamic network and computing resource allocation are integral to the service’s success. This means that the disaggregation of the functions of the cloud servers and optimized joint co-operation between the distributed edge data centers and the central data center are several essential requirements for ILM UC.

Based on the requirements as defined in the Benchmark Model, we designed the first version of the ILM RIM. We can expect the following benefits.

- The ILM RIM with Open APN technologies delivers the benefits of high bandwidth and reduced power consumption across the core network and access network while collecting massive data from capture devices (cameras, depth sensors, etc.). In addition, this RIM exploits the capabilities of the IOWN GF Open APN to set up ultra-low-latency connections directly between customer premise sites and telco edge/core sites, which are hosting the application computing resources.
- The ILM RIM with DCI technologies obtain benefits for resource allocation flexibility from heterogeneous and disaggregated device resources pool that can allocate an application's functional node at the desired location (i.e. edge data centers) to perform high data performance in hardware rate. This flexible resource allocation in disaggregated infrastructure helps CPU cost reduction and reduced power consumption.

1. Introduction

1.1. Purpose

The IOWN GF is expected to accelerate the development and commercial availability of its architecture and technology in a relatively short period of time. To accomplish this goal for IOWN GF architecture and technology development, the IOWN GF develops and evaluates RIMs, which provide the blueprints for realizing attractive IOWN GF use cases. The RIMs are also helpful for identifying potential technical issues and further improving the IOWN GF architecture and technology specifications.

The RIM utilizes the IOWN GF architecture and technology to describe an end-to-end system that meets the requirements of a specific target use case that has the best metrics for success.

In this document, the ILM UC is a metaverse use case that has been selected as an example of a system architecture designed using IOWN GF technologies. VR services such as a live music concert in a virtual space have begun to be offered to the public in recent years and the market size of these virtual events is expected to grow significantly through 2030. In order to provide an extreme immersive and interactive user experience to distributed Audience Members, the end-to-end networks and distributed computing environment requires much higher capacity and lower latency. Through development of the ILM RIM, we investigate to what extent these user experiences are realized by IOWN GF technology and identify the best practices for such a system design.

This document also aims to demonstrate many of the benefits of the IOWN GF architecture and technology over today's cloud-based implementations. The RIM adopts the IOWN GF's latest architecture and technology and continues to evolve to achieve key requirements such as the 10 [ms] Motion-to-Photon latency which is vital for the ILM UC.

The RIM is expected to evolve repeatedly and achieve larger scalability, lower power consumption and better total cost of ownership performance while achieving the vital key requirements by adopting new/revised IOWN architecture and technology and by adopting the feedback of the PoC.

The remainder of this document is structured as follows: Section 2 details the Benchmark Model for the target use case. Section 3 describes the flows and workloads of data processing in the Benchmark Model. Section 4 explains several major technology gaps and issues between today's cloud-based implementations and use case requirements in the Benchmark Model for the target use case. Section 5 defines the ILM RIM using IOWN GF architecture and technology. Section 6 summarizes our first achievements and describes future studies.

1.2. Scope

This document covers an initial study on the ILM RIM included in AIC Use Case Release 1 [IOWN AIC UC] which was chosen as the target use case. The ILM RIM in this document has been developed with reference to Open APN [IOWN Open APN] and DCI [IOWN DCI], and Data Hub Functional Architecture [IOWN Data Hub].

The IOWN GF continuously revises the ILM RIM through Proof of Concept (PoC) and detailed specification development process.

2. Benchmark Model

This section describes the Benchmark Model for the ILM UC in the AI-Integrated Communications Use Case Release 1 [IOWN AIC UC] which defines the Reference Case in Section 2.1 and Metrics and Evaluation Methods in Section 2.2.

2.1. Reference Case: Interactive Live Music

This subsection develops a Reference Case for the ILM UC. The Reference Case digs deeper into the target use case and specifically defines the conditions for determining functional and non-functional requirements. The aim of defining the Reference Case for the target use case is to make it accurate so it is possible to evaluate implementation models by measuring selected metrics in the specific conditions.

This section describes the basic requirements for the Reference Case.

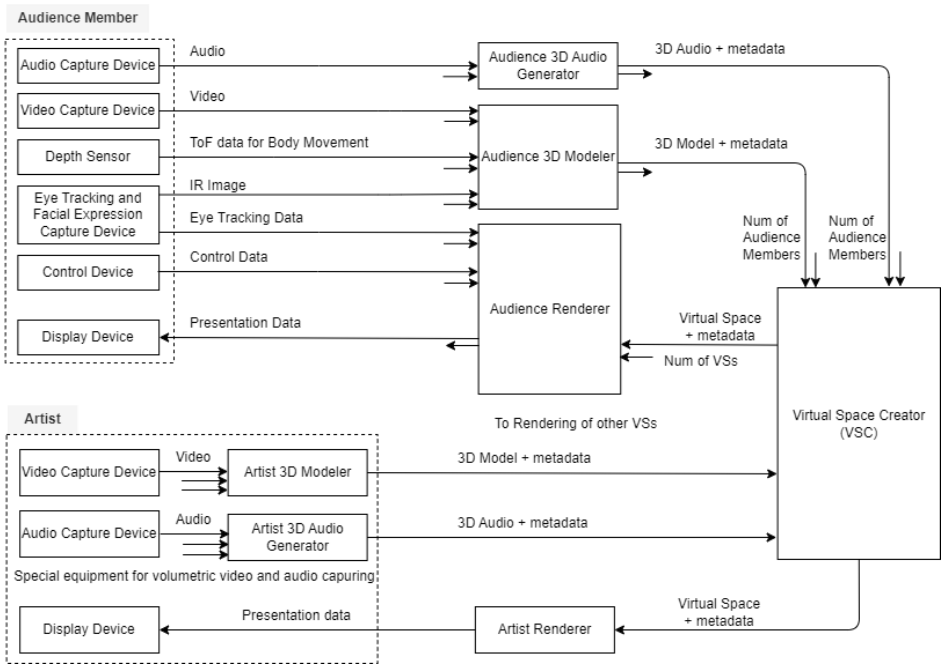

Figure 2.1-1: System Overview

2.1.1. Description

This ILM UC provides a new means to enjoy live music performances through the volumetric video and audio capture of the Artists' real-time performance, video and audio capture of the Audience Members' reaction and activity, and the creation of virtual space to realize these input data. Artists and Audience Members can intimately interact in the created virtual space, such as via the Audience Members' cheering and waving together to the Artists, or the Artists calling out to the Audience Members, as if they are all physically together at a real live performance. Moreover, because this is a Virtual Space, any Audience Members can get as close to the Artist as they wish, watch from any position they want, and even dance with virtual representations of other Audience Members who are miles away, and move around the live concert venue however they wish. As a result, this ILM UC enables new social experiences beyond actual live performance events. Furthermore, conducting mid-sized events in VR would allow even further interactions between Artists and Audience Members. Audience Members can participate from their homes or karaoke boxes and experience an immersive experience. The video and audio feed from the Audience Members can be analyzed to provide appropriate feedback to the Artists of the behavior of the Audience Members.

2.1.2. Artist's Viewpoint

- Artists are not dispersed geographically but are in the same studio that has the video capturing capability. Therefore, the artists see each other directly in real-time.
- In order to build a 3D model, special equipment for volumetric capturing is used to capture video from multiple angles, and the captured data is processed to create a point cloud (not a conventional multi-camera viewpoint toggle)
  - "Volumetric Video" is a technology that can capture an entire real space and freely arrange and stage it in a virtual space, allowing a real world location to be reproduced in the virtual world.
  - Captured data is converted into a point cloud at the shooting site and transmitted to the cloud. (The number of cameras does not necessarily match the data volume of the point cloud data, which means that the number of cameras is not a requirement for the network.)
- Artists can see part of the Audience Members in the Virtual Space just like the Audience Member's point of view. From their own location in Virtual Space, artists can see the same Virtual Space in the same manner as the Audience Members. Artists can also see the individual Audience Members' avatars on a monitor, but the members use HMDs.
  - Selected Audience Members are displayed. The system operator has a function to pick a selected member to be highlighted. Please refer to the Technology outlook [IOWN TECH LOOK] for details.
- Artists only see Audience Members who have "opted in" and giver permission for their faces to be displayed.

*note: Artists would like to see at least an anonymized low-resolution avatar to get a crowd feeling. We will consider this process later.

2.1.3. Audience Member's Viewpoint

- Audience Members are all remote, and the Artists do not see Audience Members in reality.
- Audience Members not only can watch the Artist, but also can watch other Audience Members just like in a live venue.
- The Audience Members can choose to appear in Virtual Space (Appearance OK = Exist in Virtual Space, Appearance NG = Not exist in Virtual Space).
- Audience Members can view the Virtual Space like a live venue, including in third person, where the Audience Members can see themselves
  - Audience Members can see 3DoF footage from where they are
    - Audience Members' avatars are displayed in a non-overlapping manner using techniques such as collision detection; Avatar can be viewed in 3DoF.
  - Audience Members can see 6DoF footage away from where they are
    - Audience Members can move their viewpoint which is independent to the avatar's position.
    - Audience Members can move one's viewpoint freely and view the Artists at their preferred angle. Audience Members has a 360 degree view
- Audience Members using HMD can see 360 degree video from any angle depending on the movement of the head.
- Audience Members can watch content on a large flat panel display. 360 degree video can be viewed from any position and in any direction by moving the remote control or smartphone.
  - e.g.; RCU(Remote Control Unit):
    - To select the direction: up/down/left/right key
    - To select the position: 4-color key etc. (Jump to preset position per color key)

        - ✧ To zoom-in/out: select zoom menu (UI) and up/down key
    - ➢ e.g. Smartphone:
        - ✧ To select the direction: Tilt or move in that direction, or finger operation on the UI
- The Audience Members select an avatar of their choice from the several types provided. The avatar's movements can be reflected in real-time from the Audience Member's own movements, and the Audience Member's facial expression can be applied to the avatar.
- Audience Members can capture whole body movements and facial expression by video capture device and depth sensing device.

2.1.3.1. Number of Audience Members

This ILM UC is based on the geographical environment of Japan. Tokyo is the assumed location of the studio and Audience Members are spread across Japan.

- The maximum distance from the venue to the Audience Members is 1,000 km, covering, for example, Tokyo to Sapporo or Tokyo to Fukuoka.
- 1.2 million concurrent connections (see [IOWN AIC UC] for details)

2.1.3.2. Audience Group

- Audience Group can be formed from Audience Members that are;
    - ➢ in the same geographical location.
    - ➢ geographically dispersed. e.g., gather as a social group, SNS connection or randomly selected by the system.

2.1.4. System Operator's Viewpoint

- System Operator can monitor network parameters such as bandwidth, latency, jitter, and BER.

2.2. Device Types and Specifications

The Reference Case assumes the use of the following devices described in this section.

Artists are volumetric captured. In other words, 3D models of the Artists and other objects taken from multiple angles using special equipment can be directly reproduced in Virtual Space, and the movement of these objects can also be digitalized with high precision, allowing the Audience Members to freely move their viewpoints in the immersive Virtual Space.

Audio is captured with Object Audio. Object Audio is an object-based spatial audio technology that assigns positional information to each sound source, such as vocals, choruses and instruments, and places them in a spherical space. Immersive three-dimensional sound field as if surrounded by the live performance of the Artists can be experienced.

The Reference Case here assumes the following devices that supports volumetric capturing and displaying in the year 2025.

2.2.1. Display Devices

2.2.1.1. HMD for Audience Member

- Video
    - ✧ 3D Stereo Display

- ✧ Resolution : 4K x 2K in total
- ✧ Framerate: 120 [fps]

- Audio
  - ✧ 360 degree sound system

2.2.1.2. Flat Panel Display for Audience Member

- Video
  - ✧ Resolution: 4K
  - ✧ Framerate: 120 [fps]
- Audio
  - ✧ 360 degree sound system
  - ✧ *note: Flat Panel Displays assigns to the actual loudspeaker channel.
  - ✧ *note: Flat Panel Displays are also supported as a use case.

2.2.1.3. Flat Panel Display for Artists

- Video
  - ➢ 4K x 2 Display which can display Audience Members
- Audio
  - ➢ 360 degree sound system

*note: 360 degree sound is assigned to 2 [ch] headphones.

2.2.2. Capturing Device

2.2.2.1. Capturing Device at Audience Member

Refer to Annex C.6 for details.

2.2.2.2. Video Capture Device

- Entire body movement is captured with camera and sent to 3D Modeler.
- The image will be used for facial expression like mouth movement.
- Resolution: 3,840 x 2,160
- Frame rate: 30 [fps]

2.2.2.2.1. Depth Sensing Device

- ToF(Time of Flight) sensor is used to capture body movement. The ToF data is sent to the 3D Modeling
- Resolution: 1,024 x 1,024
- Frame rate: 30 [fps], 8 [bit] Gray Scale

*note: joint data is produced at “Audience Member 3D Modeler“

2.2.2.2.2. Audio Capture Device

- Audience Members are recorded using multiple microphones and the multi-channel audio is sent to Audience Member3D Audio Generator

2.2.2.2.3. Eye Tracking and Facial Expression Capture Device

- Capture eyes and around eye expression with in-camera and send to Audience Member 3D Modeler
  - Resolution: 1920 x 1080
  - Frame rate: 60, 90, 120 [fps]
- Detect eye-tracking information and send to Audience Member Renderer
  - Frame rate: 120 [fps]

2.2.2.2.4. Control Device

- Support 6DoF signals to acquire head motion.
- Support 6DoF signals of a hand-held controller.

2.2.2.3. Capturing Device at Artist

2.2.2.3.1. Volumetric Capture Device

- Artists are captured using multiple cameras and the multi-view video is sent to Artist 3D Modeler and sent to Virtual Space Creator. The output of each camera is processed locally at shooting site.
  - Frame rate: 30, 60 [fps]
- metadata is a kind of scene description (object size, metric, time, 3D model position, etc.)

2.2.2.3.2. Audio Capture Device

- Artists are recorded using multiple microphones and the multi-channel audio is sent to Artist 3D Audio Generator. The output of each microphone is processed locally at shooting site.
- metadata are attributes to render 3D audio (object position, effect, mixing level between audio, etc).

2.3. Interface with Interactive Live Music Service

This section describes the interface to and from the Interactive Live Music Service.

2.3.1. Input to Interactive Live Music Service

2.3.1.1. From Artists

- 3D model + metadata
  - Data rate:
    - 30 [fps]: 56.35 [Gbps]
    - 60 [fps]: 112.74 [Gbps]
    - Refer to Annex C.4 for details
- 3D audio + metadata
  - Mic: 32 [ch]
    - Data rate:
      - Liner PCM 768 [kbps] x 32 [ch] = 24.6 [Mbps]

*note: Artist 3D modeler and Artist 3D audio renderer are not part of the Interactive live music service. Artist 3D Modeler and Artist 3D audio renderer may be operated by a different legal entity than the interactive live music service. Following, computing of modeler and audio renderer might also be physically separated from the computing devices of the interactive live music service, e.g., be located in another data center

2.3.1.2. From Audience Members

- Video
  - Resolution: 3840 x 2160, Frame rate: 30, 60 [fps], Bit Depth: 8 [bit], YCbCr 422
  - Data rate:
    - 30 [fps]: 3840*2160*30*8*(2/3)*3/10^9 = 3.98 [Gbps]
- 3D audio + metadata
  - Data rate: Liner PCM 2ch : 1.5 [Mbps] (16bit x 2 x 48KHz =1.536 [Mbps])
- Eye Tracking and Facial Expression for Audience Member
  - Eye Tracking
    - Data rate: 32bit float value (x, y, z, yaw, pitch, roll) at 120 [fps] /eye, 46 [kbps])
  - Facial Expression
    - Resolution: Width 1920, Height 1080, Frame rate 30, 60, 90 fps, 400 (8bit Gray Scale)/eye
    - Data rate:
      - 30 [fps]: 1920*1080*30*8*2[eyes] Total 1.00 [Gbps]
      - 60 [fps]: 1920*1080*60*8*2[eyes] Total 1.99 [Gbps]
      - 90 [fps]: 1920*1080*90*8*2[eyes] Total 2.99 [Gbps]
- ToF data
  - Resolution: 1,024 x 1,024, Frame rate: 30 [fps], 400 (8bit Gray Scale)
  - Data rate:
    - 30 [fps]: 1024*1024*30*8/10^9 = 0.25 [Gbps]
- Control Device for Audience Member
  - Head Motion
    - 6DoF sensor data (x, y, z, yaw, pitch, roll), 32bit float value
    - Sampling rate: 120 [Hz]
    - Date rate: 32*6*120= 23.04 [kbps]
  - Hand-held controller
    - 6DoF sensor data (x, y, z, yaw, pitch, roll), 32bit float value
    - Sampling rate: 120 [Hz]
    - Date rate: 32*6*120= 23.04/hand, 46.08 [kbps]

*note: See Annex C.4

2.3.2. Output from Interactive Live Music Service

2.3.2.1. To Artists

- Video:
  - Resolution: 3840 x 2160
  - Frame rate: 60 [fps]
  - Bit Depth: 8bit, YCbCr 420
  - Data rate: 3840*2160*60*8*0.5*3/10^9 = 5.97 [Gbps] (Uncompressed)

- Audio: Linear PCM 768 [kbps] x 32 [ch] = 24.6 [Mbps]

2.3.2.2. To Audience Members

- Video: (HMD)
  - Bit Depth: 8bit, 10bit, 12bit
  - Chroma-subsampling: YCbCr 420, 422, 444
  - Data rate: (Uncompressed)
    - e.g.; 8bit YCbCr 420: 4000*2000*120*8*0.5*3/10^9 = 11.52, Total 23.04 [Gbps]
    - e.g., 12bit RGB 444: 4000*2000*120*12*1*3/10^9 = 34.56, Total 69.12 [Gbps]
- Audio: Linear PCM 768 [kbps] x 32 [ch] = 24.6 [Mbps]
- Video: (Flat Panel Display)
  - Bit depth: 8bit
  - Chroma-subsampling: YCbCr 420
  - Data rate: (Uncompressed) 3840*2160*120*8*0.5*3/10^9 = 11.94 [Gbps]
- Audio: Linear PCM 768 [kbps] x 32 [ch] = 24.6 [Mbps]

2.4. Functional Requirements

The Interactive Live Music Service receives continuous Video and Audio streams from Artists and Audience Members and creates Virtual Space to be displayed on Display Devices. Below is a depiction of the data flow and functions between Artists and Audience Members.

2.4.1. Main Data Processing and Relevant Functions

2.4.1.1. Audience Member 3D Modeler

- Creation of a 3D Model (Point Cloud) and associated metadata (object size, metric, etc.) from video, ToF data (body movement), and IR Image (facial expression).
- Avatar expression creation from IR image (eye-tracking and facial expression).
- Audience Member 3D Modeler uses the avatar image for Audience Members to construct 3D Model.

*note: The location of the avatar data will be discussed later.

2.4.1.2. Audience Member 3D Audio Generator

- Integration of multichannel audio and creation of Object Audio and associated metadata
  - note: This 3D Model / 3D Audio and associated metadata could be the element of Virtual Space.

2.4.1.3. Audience Member Renderer

- Generation of Presentation Data (e.g. for HMD, LR image) from Virtual Space, and Audience Members' Eye Tracking Information, and Audience Members' device type

2.4.1.4. Artist Renderer

- Generation of Presentation Data (e.g. 2D image) from Virtual Space to be displayed at Artists

2.4.1.5. Virtual Space Creator

- Creation of Virtual Space from;
  - Artists' 3D Model and associated metadata, and Artists' Object Audio and associated metadata
  - Audience Members' 3D Model and associated metadata, Audience Members' Object Audio and associated metadata, and determined position of Artists in Virtual Space
  - Virtual Space venue composition such as 3D model of stage, seats, light position, speaker position, etc.
  - Addition of CG effects Information including lighting from AI/System Operator's direction
  - Addition of ROI (Region of Interest) Information and provide recommended viewport for Audience Member, as metadata by the System Operator.
  - Selection of Audience Members to be viewed by Artists and addition of ROI Information such as selected Audience Members for Artists, as metadata
- Virtual Space Creator may create multiple Virtual Spaces
  - A Virtual Space contains one Audience Member Group and Artists
  - Virtual Space Creator may combine multiple virtual spaces into an integrated virtual space for presentation to Artists
- System Operator Monitoring
  - Monitoring the network parameters such as bandwidth, latency, jitter, and BER.

2.5. Non-Functional Requirements

This section defines the non-functional requirements for the Reference Case.

- Latency
  - The data captured by reaction sensors such as a head tracking device need to be reflected by the presented view on the Audience Member's presentation device within 10 [ms] (motion-to-photon latency). This is to prevent VR sickness and achieve more realistic experience. For numbers, see [IOWN AIC UC] for more information.
  - Audio and video from Artists and Audience Members must be sent within 70 [ms] as a worst-case number
    - The time between capturing Artists with a set of volumetric video and audio devices, transmitting them, and displaying the volumetric video and audio data on the Audience Member's display device: within 70 [ms]
    - The time between capturing Audience Member with a set of video and audio devices, transmitting them, and displaying the video and audio data on the Artists' display device.: within 70 [ms]
    - The time between capturing Audience Member with a set of video and audio device, transmitting them, and displaying the video and audio data on the other Audience Members' display device.: within 70 [ms]
  - Each individual data such as Artists and distributed Audience Members' video/audio/motion data should be synchronized for body movement analysis. For example, to detect cheering and generate production special effects accordingly: within 100 [ms].

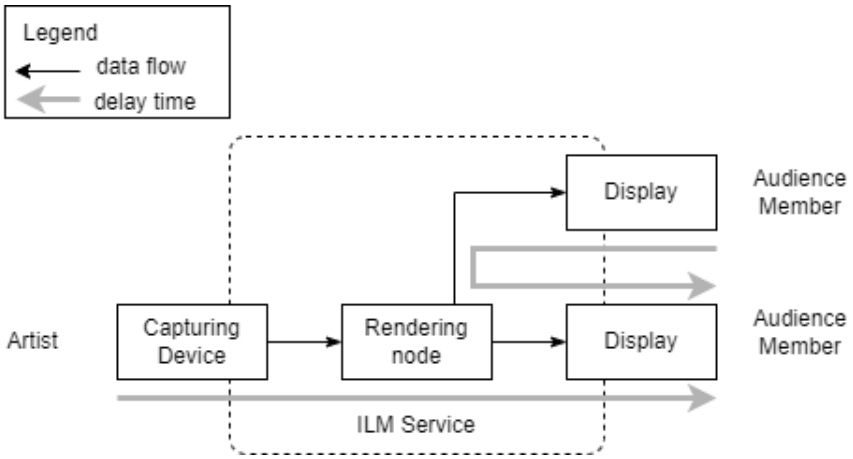

Figure 2.5-1: Data stream toward Audience Members

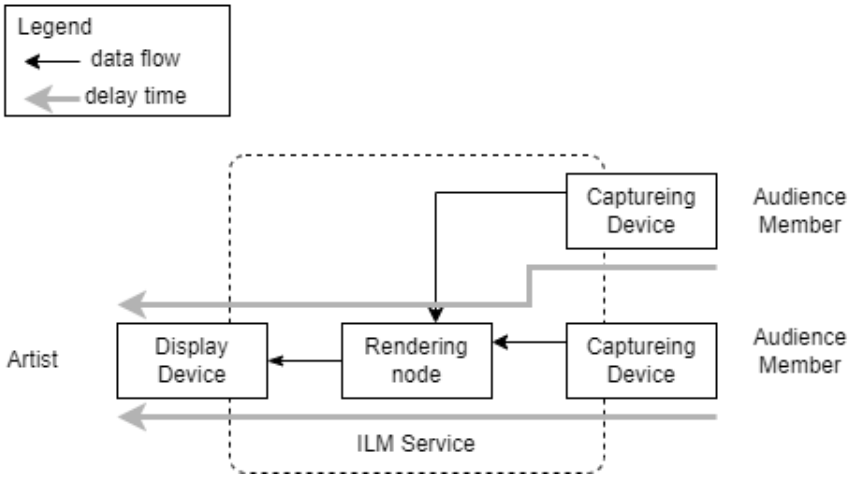

Figure 2.5-2: Data stream toward Artist

*note: The value of the Latency requirements listed here above are that for after starting the service. In addition to these requirements, please note that there exists Waiting Room Time for Audience Member to enter the virtual space after an Audience Member makes the request to join the event. This Waiting Room Time depends on (1) the time to secure all the necessary wavelength and bandwidth path&light-source in APN between the transceivers (APN-Ts) in extra network gateway and DCI-Gateway via APN-G and (2) the time to launch instance of Audience Member Renderer, Display Device and its network attachment device.

*note: In addition to provisioning network resources, additional time might be required to provision (boot up) and add a new rendering server to the application, in case no rendering server with sufficiently low latency from the user is available.

*note: The complexity of placing the rendering servers geographically depends on how dispersed the Audience Members are.

- Scalability
  - Depending on the event, the number of Audience Members may increase or decrease, and the Virtual Space can be configured in a scalable manner.
  - Number of Audience Members in one Audience Group

- ✧ The number of Audience Members in an Audience Group is 3,000. Audience Members can interact among themselves in synchronous manner.
- ✧ The capacity of real concert halls, such as Carnegie Hall, Opera House, Vienna State Opera and La Scala di Milan, is about 2,000 to 3,000. This number is used to define the number of Audience Members in an Audience Group; 3,000.

2.6. Metrics and Evaluation Method

This subsection describes the metrics and evaluation method for the Reference Case.

2.6.1. Overview

This system, which implements the ILM UC, can transmit Artist data to Audience Member in real-time, and can transmit Audience Member's movements to other Audience Members to dance and sing together. Therefore, the IOWN GF infrastructure must transmit the Artist's data and the Audience Member's Animation Data with very low latency.

In addition, this IOWN GF infrastructure is expected to satisfy the functional and non-functional requirements of use cases and to be provided as a system with as low cost and low power consumption as possible.

A certain degree of data and process consolidation in one or several data enters is required to build an efficient system at a low cost and low power consumption. However, it inevitably requires an increase in response time due to the data transfer over the network and the data distribution process in the data center.

2.6.2. Metrics

The ILM RIM needs to realize this Reference Case while meeting the functional and non-functional requirements, especially the end-to-end response time.

The metrics for further evaluation are:

- System Power consumption,
- System cost,
- Latency, and
- Jitter

*note: The capturing system for Artists is located on premise and it has a certain delay.

2.6.3. Evaluation Method

Evaluation of the ILM RIM can be achieved by comparing the above metrics. The power consumption can be obtained by measuring a network device that transmits data through PoC or the like and a computing node that processes the data. In addition, the system cost can be obtained by calculating the cost of the implemented system. Latency can be obtained by measuring the motion-to-photon latency and the end-to-end latency. Jitter can be measured among video, audio, and motion synchronization.

3. Dataflow and Workloads Analysis

This section first analyzes the flows and workloads of data processing in the benchmark model. This section aims to identify a broader range of requirements necessary for system design in addition to the key requirements shown in the IOWN AIC UC. Then this section subdivides data processes into sub-processes and defines the behavior of each sub-process. Finally, this section clarifies the connections and dataflows between sub-processes until they have sufficiently fine granularity for evaluation by the IOWN GF in the context of the technologies it is studying.

Through this analysis, we utilized the Dataflow and Workload Profiling Framework, which the IOWN GF developed to identify services gaps/requirements of use cases accurately and efficiently. Please refer to Annex B for the details of the framework.

3.1. Data Pipeline Diagram

Figure 3.1-1 is the Data Pipeline Diagram (DPD) for the ILM UC to analyze the data processing and dataflow.

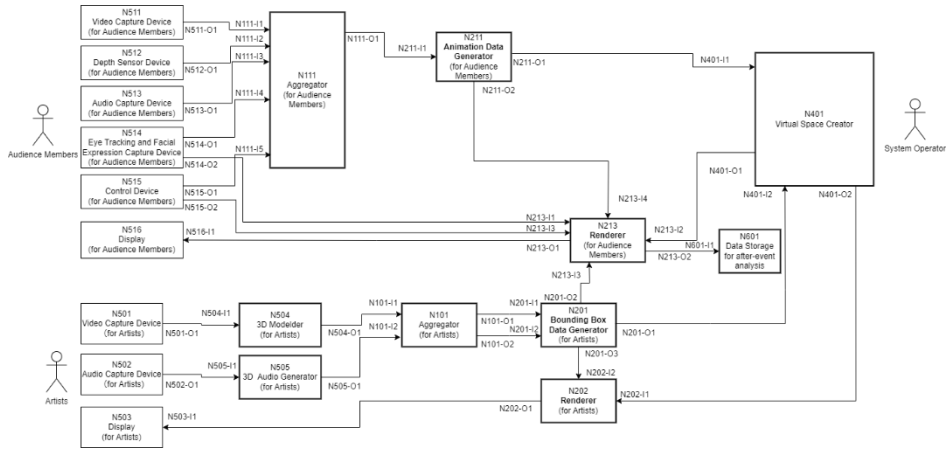

Figure 3.1-1: Data Pipeline Diagram for the Interactive Live Music Use Case

As shown in the diagram, it has seven types of functional nodes:

- Aggregator (N111, N101): Nodes that collect data from the devices and send data to the Animation Data Generator node and Bounding Box Generator node. They may also provide functions for efficient collection, such as consolidating video and audio data and received from multiple devices. N111 collects sensor data from the devices such as Video Capture, Depth Sensor, and Audio Capture Device.
- Animation Data Generator for Audience Members (N211): A node that receives video data, Depth Sensor data, and Eye Tracking and Facial Expression data and generates Animation Data from the movement of the Audience Member.
- Renderer for Audience Members (N213): A node that receives user control data (viewport data), Animation Data, and Avatar data, and audio data from Audience Members, Artists' 3D model data, Artists' audio data, and 3D scene data. Collects Avatars ' positions from Virtual Space Creator (N401), overlays the polygon data on it and renders the scene. Then sends the data to Display. Here, the Avatar is placed and rendered from Avatar data and Animation Data, overwriting the scene created with Virtual Space Creator (N401).
- Renderer for Artists (N202): A node that receives Artists' 3D model data, Artists' audio data and 3D scene data, then performs rendering, and sends the data to Display.
- Data Storage for after event-analysis (N601): A node with a database/storage that stores data and provides these data for further analysis.

- Virtual Space Creator (N401): A node that provides Interactive Live Music services to the Renderer nodes (N213, N202) and is controlled by System Operator. This node creates 3D scenes from various parts, such as venue information, Animation Data, etc., by the System Operator's instructions. Also, it manages the entire position of Avatars and structures in 3D space and detects spatial collisions and corrects Avatars ' positions. This node does not perform rendering; the rendering itself is performed by the Renderer nodes (N213, N202).
- Sensor (N511, N512, N513, N514, N515): Nodes for capturing Audience Members. Video Capture Device (N511) captures Audience Members using RGB camera as image,
  Depth Sensor Device (N512) captures Audience Members' body movement using depth sensor and IR image sensor as depth sensor (ToF) data and an image.
  Audio Capture Device (N513) captures Audience Members' voice using microphones as the multi-channel audio.
  Eye Tracking and Facial Expression Capture Device (N514) captures Audience Members using the camera as image.
  Control Device (N515) captures Audience Members' head motion and hand motion or push buttons operation as control data. Then all the data is sent to the Aggregator (N111). At the same time, Audience Members' Eye Tracking and Control data is sent to the Renderer for Audience Members (N213) as viewport information.
- Sensor (N501, N502): Nodes for Artists that capture Artists as multi-view video and multi-channel audio.
- 3D Modeler (N504) and 3D Audio Generator (N505): Nodes for Artists that capture Artists as volumetric video and multi-channel audio and generate 3D model and 3D audio.
- Display (N503, N516): Nodes that receive video and audio streams from the Renderer nodes (N213, N202) and show them.

Figure 3.1-1 shows the overview of the DPD, and more detailed information on the functional nodes is shown in Annex C.

To meet the requirements of the Benchmark Model defined in the previous section, the DPD contains the following major processing flows in the Artist, the Audience Member, and the System Operator side:

- Artists
  - Transmission of 3D model data, and Bounding Box data to Virtual Space Creator
  - Rendering video using data from Virtual Space Creator
  - Display the rendered video to display for monitoring purposes
- Audience Members
  - Generating Animation Data from Audience Member's sensor devices. Audience Members need send their Avatar Data to the Renderer before the live event
  - Animation Data is sent to the Renderer and Virtual Space Creator
  - Rendering video from the Audience Member's viewpoint using received data.
  - Display the rendered video to display.
- System Operator
  - Creating 3D scenes and sending data to both Audience Members and Artists

3.1.1. Artists

This section describes the data flow from shooting to 3D scene generation, rendering, and display for monitoring purposes. It is not necessary to send volumetric video every frame to create a 3D scene, it is sufficient to send only a Bounding Box.

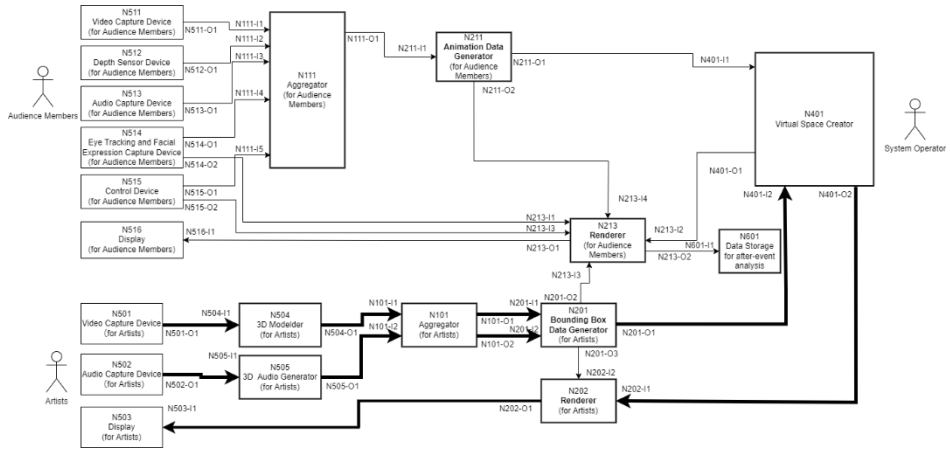

Figure 3.1-2: Transmission of 3D model data and 3D audio data

Figure 3.1-2 shows the transmission flow of Artists' 3D model data and 3D audio data to the Renderers. Also, it shows the transmission flow of Bounding Box data to generate the 3D scene in Virtual Space Creator (N401). After creating the 3D scene in Virtual Space Creator (N401), the 3D scene will be sent to the Renderer (N202). 3D Modeler (N504) is assumed to be on-premise, at the local shooting site. 3D Modeler (N504) uses a 360 degree image from Video Capture Device (N501) and creates a 3D volumetric image using some algorithm.

- Aggregator (N101) aggregates data received from the Video Capture Device (N501) and Audio Capture Device (N501), and synchronizes and associates the 3D model data and 3D audio data on a time and space basis. In addition, each stream may contain embedded metadata, which is transferred to the Bounding Box Data Generator (N201).
- Bounding Box Data Generator for Artists (N201) receives 3D model data and 3D audio, and metadata from the Aggregator (N101). Bounding Box data is generated from the Artists' 3D model. The Bounding Box data and 3D audio data are sent to the Virtual Space Creator (N401). At the same time, the received 3D model data, 3D audio, and metadata are posted to the Renderers (N213, N202).
- The Virtual Space Creator (N401) is the primary functional node for this use case. It receives the Animation Data, 3D audio data, and metadata and decodes both sets of data. A Virtual Space is created from Artists' Bounding Box data and associated metadata, Artists' 3D audio data and associated metadata, Audience Members' Animation Data and associated metadata, Audience Members' 3D audio and associated metadata, the position of Artists and Audience Members, and venue composition data such as 3D models of stage, seats, light position, speaker position, etc. Finally, the Scene Composition data*1 is sent to the Renderer for Artists (N202).
- Renderer for Artists (N202) receives Scene Composition data from Virtual Space Creator (N401) and 3D model data, 3D audio data, and metadata from Bounding Box Data Generator for Artists (N201). Generate presentation data by 3D rendering and send the presentation data to the Display (N503).
- Display for Artists (N503) has a 4K resolution display, which is used by Artists to monitor Audience Members. It also monitors 3D audio. Typically, the 360-degree sound is assigned to two channels of headphones.

3.1.2. Audience Members

This section describes the data flow from shooting of the Audience Member, creating Animation Data, generating 3D scenes, receiving control information from HMD, determining the field of view, rendering, and displaying to Audience Members. It is not necessary to send Avatar data for every frame, Avatar data does not change in every frame, so it may be sent once to the Renderer prior to the live music event. And Animation Data Generator (N211) sends Animation Data to the Renderer (N213) and Virtual Space Creator (N401) to reduce the data rate.

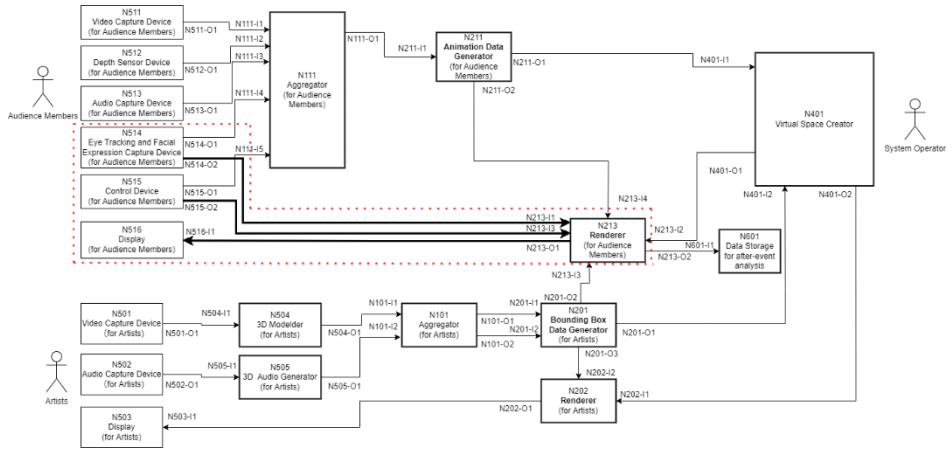

Figure 3.1-3: Transmission of the Animation Data to Virtual Space Creator and Avatar Data to the Renderer

Figure 3.1-3 shows the transmission flow of the Audience Members' Animation Data and 3D audio data to the Renderer. Also, it shows the transmission flow of the Animation Data to generate the 3D scene in the Virtual Space Creator (N401). After creating the 3D scene in Virtual Space Creator (N401), the Scene Composition data will be sent to the Renderer (N213). The presentation data generated at the Renderer (N202) are transmitted to the Display (N516).

- Aggregator (N111) aggregates data received from the Video Capture Device (N511), Depth Sensor Device (N512), Audio Capture Device (N513), Eye Tracking and Facial Expression Capture Device (N514), and Control Device (N515) for Audience Members. This node synchronizes and associates the input data on a time and space basis. Some metadata (e.g., time, position, angle, and owner information in a unified format) can be embedded into each stream so that subsequent nodes can efficiently collect and tie relevant records together for their analysis. The Eye Tracking and Control Device data are then sent to the Renderer for Audience Members (N213).
- Animation Data Generator for Audience Members (N211) receives image and depth sensor (ToF) data, control data, multi-channel audio from the Aggregator (N111). This node generates the Animation Data of the Avatar and sends the Animation Data to the Virtual Space Creator (N401). At the same time, the Animation Data, 3D audio data, and metadata are sent to the Renderer (N213).
- The Virtual Space Creator (N401) is the primary functional node for the ILM UC. For Audience Members' side, it receives the Animation Data, 3D audio, and metadata, and decodes both sets of data. A Virtual Space is created from Artists' Bounding Box data and associated metadata, Artists' 3D audio and associated metadata, Audience Members' Animation Data and associated metadata, Audience Members' 3D audio and associated metadata, the position of Artists and Audience Members, and venue composition data such as the 3D models of stage, seats, light position, speaker position, etc. Finally, Scene Composition Data is sent to the Renderer for Audience Members (N213).
- The Renderer for Audience Members (N213) receives the Scene Composition Data from the Virtual Space Creator (N401). Receives eye tracking data and control data (includes head tracking data) from Eye Tracking and Facial Expression Capture Device for Audience Members (N514) and Control Device for Audience Members (N515). Receives the Audience Member's Animation Data, 3D audio data, and metadata from the Animation Data Generator for Audience Members (N211). Receives Artists' 3D model data, 3D audio data, and metadata from Bounding Box Data Generator (N201). Generates viewport information from eye tracking data and or control data and generates presentation data by 3D rendering using the viewport information. Finally sends the presentation data to the Display (N516).
- Display for Audience Members (N516) has a 4K resolution display with 120 [fps], which is used for Audience Members to see Artists and Audience Members. 360-degree sound is assigned to two channels of headphones.

3.1.3. System Operator

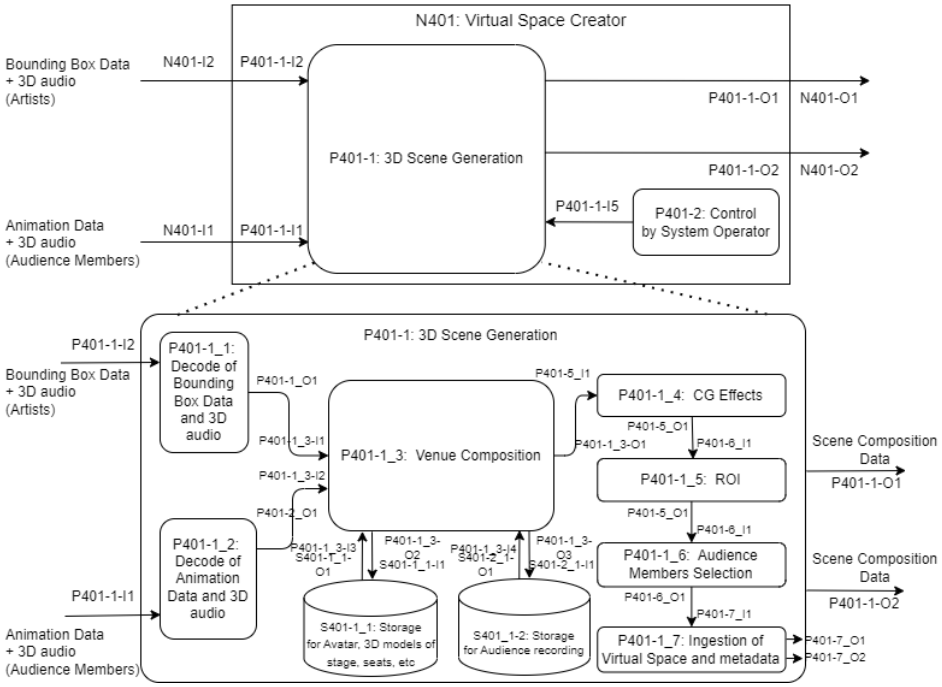

Figure 3.1-4: How System Operator controls the live music event

- The Virtual Space Creator (N401) is the primary functional node for the ILM UC. This node is controlled by the System Operator and can give instructions to how Virtual Space is created. It receives all the information to create the 3D scenes. It creates the Virtual Space from Artists' Bounding Box data and associated metadata, Artists' 3D audio and associated metadata, Audience Members' Animation Data and associated metadata, Audience Members' 3D audio and associated metadata, and determines the position of Artists and Audience Members in the Virtual Space, and the venue composition data such as the 3D models of stage, seats, light position, speaker position, etc. Specifically, the System Operator can add CG effects information including lighting objects for 3D scenes by AI/System or Operator's direction, add Region of Interest (ROI) information and provide recommended viewport for Audience Member as metadata. Also, the System Operator can select Audience Members to be viewed by Artists and add its information as metadata. Finally, Scene Composition Data are sent to the Renderer (N213, N202).
- The Data Storage for after-event analysis (N601) is storage to record the rendered video and audio from the Renderer for Audience Members (N213). The rendered data is compressed and stored Refer to Annex C, N601 for details.

3.2. Dataflow Profiles

This subsection shows dataflow profiles of each data type. The DPD showed that it is necessary to handle different types of data for supporting the given Benchmark Model. Each data type has distinct data flow requirements (e.g., number of sources, data rate, data size, occurrence rate, etc.)

Note that an appropriate communication scheme and compression scheme should be selected for each communication section according to the dataflow requirements. This selection will be discussed in Section 5, ILM RIM. Examples of communication schemes and compression schemes are as follows:

- Communication scheme
  - Shared memory, DMA, RDMA, UDS, RTP over UDP or TCP, etc.

- Compression scheme
  - Video Image: H.264, H.265, H.266, uncompressed, etc.
  - Display: DSC (Display Stream Compression) etc. *1
  - 3D model: geometry/texture compression, point cloud compression (G-PCC, V-PCC), uncompressed, etc.

note *1: HMD could make use of foveated rendering, etc at the application layer.

3.2.1. Summary

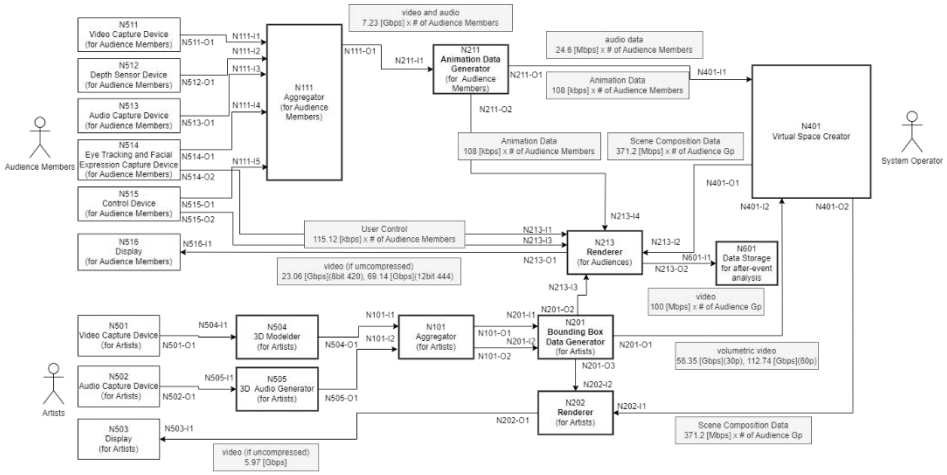

Figure 3.2-1: Summary of the primary dataflow

For Artists' side, the volumetric video data is around 112 [Gbps] at 60 [fps] (if uncompressed), around 56 [Gbps] at 30 [fps]. For Audience Members' side, the Animation Data to move the Avatar is 108 [kbps], and it grows in proportion to the number of people.

3.2.2. Volumetric video data

The following table summarizes the characteristics and significant attributes of the dataflows related to the volumetric video data.

For Artists' side, the volumetric video data drawn in thick lines is not small, around 112 [Gbps] at 60 [fps] (if uncompressed), even 30 [fps] is around 56 [Gbps] and this rate is above 8K 60p baseband data rate for broadcast today. Since real-time compression of volumetric video data is not realistic right now, it may be necessary to be uncompressed data.

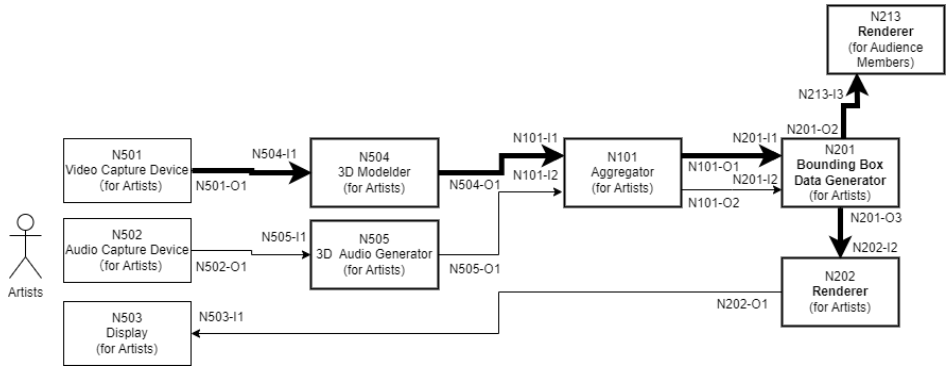

Figure 3.2-2: Sending volumetric video data

ID	Description	Attributes
3D Modeler > Aggregator **N504-O1> N101-I1** **Aggregator > Bounding Box Data Generator** **N101-O1, N201-I1** **Bounding Box Data Generator > Renderers** **N201-O2, N213-I3** **N201-O3, N201-I2**	3D model data from 3D Modeler	# of sources: 1 3D model data stream Compression scheme: uncompressed Data rate 30 [fps]: 56.35 [Gbps] 60 [fps]: 112.74 [Gbps]

3.2.3. Animation Data

For Audience Members' side, gathering sensor data from each Audience Member is not small if the data is uncompressed. Also, the data from the Renderer (N213) to the Display (N516) is not small if the data is the uncompressed video. It may affect the motion-to-photon latency.

Because it is assumed that the number of Audiences Member is assumed up to 1.2 million in the Benchmark Model, the data rate for Animation Data would increase up to around 130 [Gbps] in the central data center if all data is managed in one central data center. Therefore, it is assumed that only the Animation Data of specific Audience Members that are selected for construction of the 3D scene by the System Operator is sent.

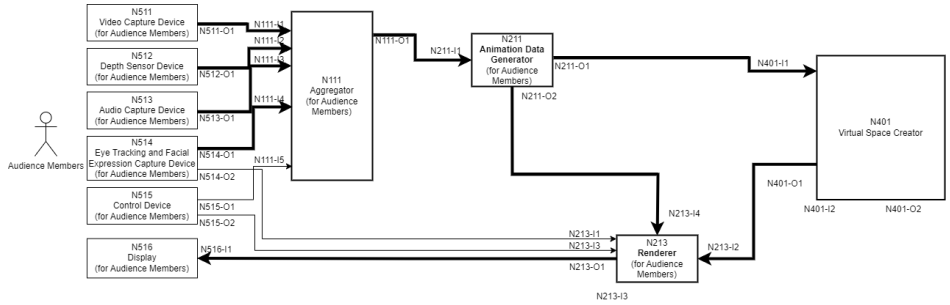

Figure 3.2-3: Sending Animation Data

ID	Description	Attributes

Aggregator > Animation Data Generator **N111-O1> N211-I1**	video data and 3D audio data	# of sources: 1 video data and 3D audio data Compression scheme: uncompressed Data rate 30 [fps]: 5.25 [Gbps] (5.23 [Gbps] video + 24.6 [Mbps] audio) 60 [fps]: 6.26 [Gbps] (6.24 [Gbps] video + 24.6 [Mbps] audio) 90 [fps]: 7.26 [Gbps] (7.24 [Gbps] video + [24.6] Mbps audio)
Animation Data Generator > Virtual Space Creator **N211-O1 > N401-I1** **Animation Data Generator > Renderer** **N211-O2, N213-I1**	Animation Data, 3D audio and metadata	# of sources: 1 Animation Data and 3D audio Compression scheme: uncompressed Data rate: 24.71 [Mbps] Animation Data: 108 [kbps] 3D audio data: Linear PCM 0.768 [Mbps] x 32 [ch] = 24.6 [Mbps]

Virtual Space Creator > Renderer for Audience Members **N401-O1 > N213-I2**	Scene Composition data	Total Data Rate: Per Virtual Space 30 [fps]: 371.2 [Mbps] 60 [fps]: 742.4 [Mbps] Data rate details Bounding Box data 30 [fps]:1,680 [bytes/frame] * 30 [fps] = 13,440 [bits/frame] * 30 [fps] = 403.2 [kbps] Animation Data * (# of Audience Members) 108 [kbps] * 1.2 [million] =129.6 [Gbps] Venue composition data 500 [Mbytes/frame] * 30 [fps] = 4000 [Mbits/frame] * 30 [fps] = 120 [Gbps] 3D audio data Linear PCM 0.768 [Mbps] x 32 [ch] = 24.6 [Mbps] Total Data Rate: (MAX) 30 [fps]: 250 [Gbps] 60 [fps]: 500 [Gbps] Note: The information sent to the Renderer of each Edge Server is not for everyone, but only for the number of people in the Virtual Space. Other Audience Members will be displayed as a grain of rice, so it is not necessary to transmit all the data. In other words, the actual number of Audience Members is not 1.2 million but 3,000 Audience Members. Data rate: (Per Virtual Space) Bounding Box data 1,680 [bytes/frame] * 30 [fps] = 13,440 [bits/frame] * 30 [fps] = 403.2 [kbps] Animation Data * (# of Audience Members) 108 [kbps] * 3,000 = 324 [Mbps] Venue composition data 500 [Mbytes/frame] * 30 [fps] = 4000 [Mbits/frame] * 30 [fps] = 120 [Gbps] 3D audio data Linear PCM 0.768 [Mbps] x 32 [ch] = 24.6 [Mbps] In reality, the venue composition data need not be sent for every frame. Assuming that a song is 3 minutes long and that the Venue composition data is changed for each song, it is sufficient to transmit the data in advance of the performance. In other words, 500 [Mbytes] should be transmitted in 3 minutes. Consequently, the data rate of the venue composition data is: 500 [Mbytes] /(3 * 60 [sec]) = 2.7 [Mbytes/sec] = 22.2 [Mbps] Total Data Rate: Per Virtual Space 30 [fps]: 371.2 [Mbps], 60 [fps]: 742.4 [Mbps] [#N401-O2]
Renderer > Display **N213-O1 > N516-I1**	2D video	# of sources: 1 rendered video Compression scheme: uncompressed Data rate 8bit YCbCr420 23.06 [Gbps] 12bit RGB444 69.14 [Gbps] *note: audio data rate is omitted here.

3.2.4. Audience Member Control Data

One Control Device is a built-in sensor in the HMD to acquire head motion. Another Control device is a hand-held controller to capture hand motion or push buttons operation. These devices generate 6-DoF signals, and they are sent to the Aggregator (N111) (color in purple).

The Renderer (N213) generates the field of view that the user sees based on the 3D scene generated by the Virtual Space Creator (N401) according to the user's line of sight, performs the rendering process, and sends it to the Display (N516).

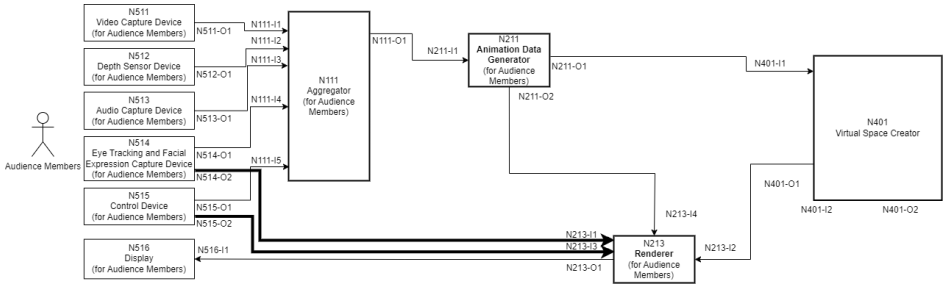

Figure 3.2-4: Sending Audience Member Control Data

ID	Description	Attributes
Eye Tracking and Facial Expression Capture Device > Renderer **N514-O2 > N213-I1** **Control Device > Renderer** **N515-O2 > N213-I3**	Eye Tracking data and Control data	# of sources: 1 data stream Compression scheme: uncompressed Data rate: 115.12 [kbps] eye tracking: 46 [kbps] head motion: 23.04 [kbps] hand-held controller: 46.08 [kbps]

3.2.5. Renderer computing resource estimation

The transmission bandwidth of this ILM UC using high-density volumetric video which is tailored to the characteristics of the human visual eye is extremely high. Also, it also uses high-resolution Avatar, which are tailored to the characteristics of the human visual eye, is extremely high. In practice, it will be necessary to combine APN with various technologies to reduce the processing load.

The estimate from the interpolation of our experiments is that the computational power required to render 100 Audience Members and Artist with the parameters of the Benchmark model is about one order of magnitude higher of what one of the high-end GPU could provide today.

Technically, this can be achieved by evolving GPUs and allocating one GPU per person. However, to make it reasonable from a cost perspective, it is necessary to introduce processing load reduction technologies such as Level Of Detail (LOD), occlusion culling, and view-dependent (i.e., not sending and receiving objects, or loading into memory objects in the viewing area because they are unnecessary for rendering).

Since the target values in the RIM document are envisioned for 2030, it would be difficult to achieve all of them immediately. Therefore, to achieve them step by step over the long term, motion-to-photon latency and fps should be prioritized, and the number of object points that determine image quality should be the last target to be achieved. It is

necessary to determine the performance of display devices, etc. at the time of the PoC. Please refer to Annex H. for details of the estimation.

4. Technology Gaps and Issues

This section describes technical gaps and issues between today's centralized cloud-based implementation and the use case requirements in the given benchmark model. The technical gaps and issues are expected to be addressed by IOWN GF infrastructure.

4.1. Typical Structure of Today's Centralized Cloud-Based Implementations

Figure 4.1-1 shows a typical structure of today's centralized cloud-based implementation. Although there are several variants and more distributed approaches such as MEC (Multi-access Edge Computing), the basic strategy of today's centralized cloud-based implementation is a centralized computing approach. Most of the functional nodes are deployed in a central cloud of a single cloud vendor, and their application traffic is confined within the central cloud as much as possible. This is due to several reasons, such as the cost-effectiveness of a hyper-scale data center, data-transferring costs/overheads to other clouds within/without the cloud vendor, and the cost in time for application developers to educate themselves on a new cloud vendors' managed service.

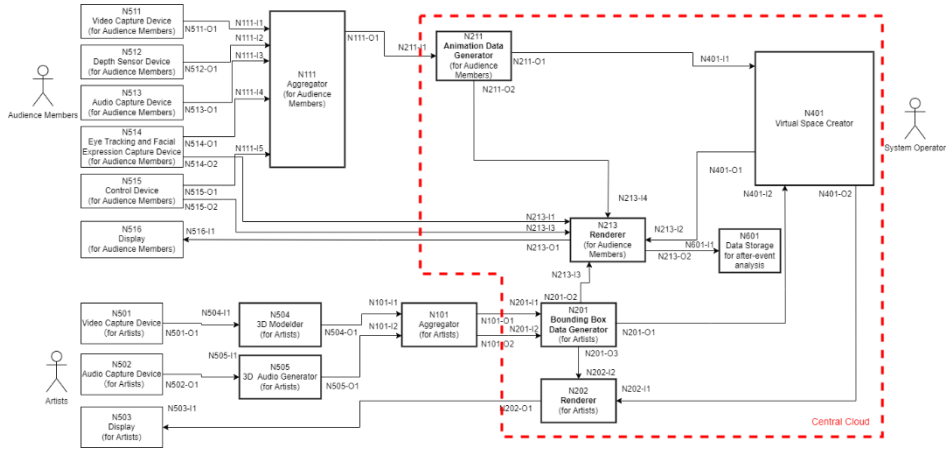

Figure 4.1-1: Data Pipeline Diagram of when all processing is done in Central Cloud

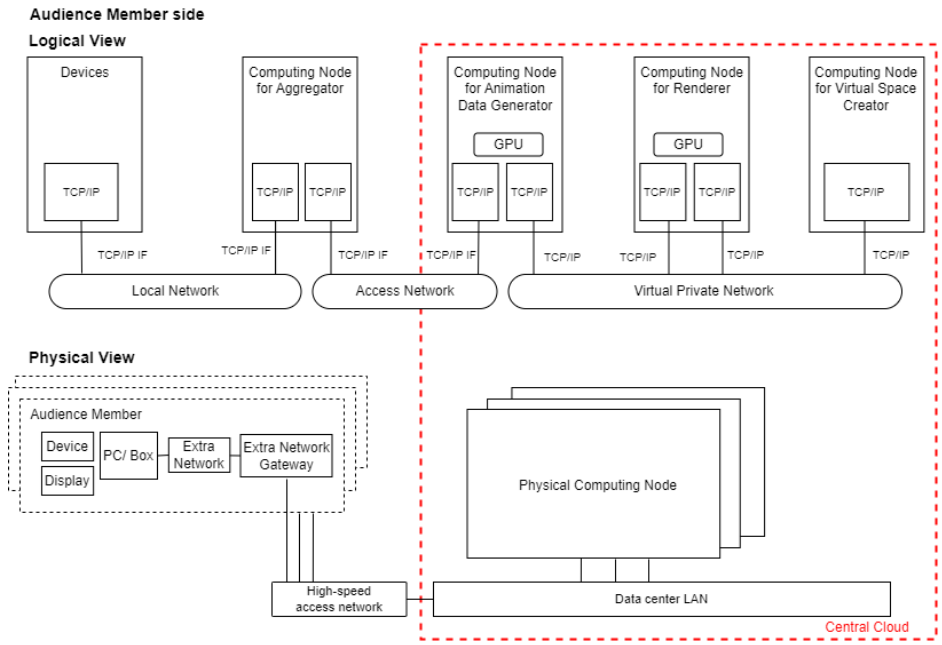

Figure 4.1-2: Today's Centralized Cloud-Based Implementation (Audience Member side)

Concerning the communication between devices (e.g., Control Data, and video and ToF data to generate Animation Data) and functional nodes on the central cloud, the devices are connected with the functional nodes via high-speed access networks and send their captured data to the applications. The TCP protocol is usually used for the transport layer protocol of data transfer such as Control Data. UDP is used to transmit media type data stream such as video images.

Inside a central data center, computing nodes, e.g., virtual machines for Kubernetes nodes, are created with a hardware manager, and microservices for the functional nodes in Linux containers are deployed to computing nodes. Typically, gRPC or REST API is used to exchange data among the microservices, and service mesh is used for controlling the route of the traffic among the computing nodes.

4.2. Issues of Today's Centralized Cloud-Based Technology

Today's centralized cloud-based implementation model has the following technical issues.

4.2.1. High Bandwidth Cost Caused by Centralized Cloud Computing

To build a Interactive Live Music service system in a single central data center, all data needs to be collected in the Central Cloud. The Benchmark Model for the ILM UC assumes that each Audience Member's site sends 7.23 [Gbps] for the transmission of video, ToF (Time of Flight) data, audio and control data to the Central Cloud. For example, if 10,000 Audience Members join the service, the total data traffic is 72.3 [Tbps]. Also, each Audience Member receives more than 23.06 [Gbps] for the presentation data to Audience Member's display and the total data traffic will be 230.6 [Tbps]. The ILM RIM assumes to be realized in the early phase of the service and the total data volume is much smaller than that of the case of the maximum number of Audience Members as defined in the Benchmark Model. Even if we use the low-latency codecs and we can assume the compression ratio is less than 30% for example, such data volume would still be the issue and it would be challenging or very costly.

This issue can be reduced by processing data such as generating Animation Data and rendering process in a more geographically distributed manner. Considering the data flows as shown in the Section 3, transferring all data over the network and aggregating it in the central data center would be wasteful. These data are processed for each Audience

Member and they are not shared with other Audience Members and other functional nodes. Therefore, generating Animation Data and rendering should be done closer to the Audience Member.

4.2.2. Lack of Deterministic Service Quality Caused by Best-Effort Networking

In the key requirements defined in the Benchmark Model, there are some latency requirements like motion-to-photon latency. Motion-to-photon latency is the time delay from the time that Audience Member's head/eye-position movement is captured to the time that the presentation data generated by rendering a Virtual Space are fully reflected to the screen of the Audience Member's display. This latency requirement is specified as 10 [ms] but it is very hard to satisfy the requirement by using the today's best-effort network. With current technology, to manage to deliver the presentation data, some techniques are adopted such as data compression and bitrate adaptation, which is a technology to deliver the different bitrate of video data according to the effective bandwidth. But they sacrifice the video quality and it may cause more delay.

Furthermore, there is another latency requirement that is the end-to-end latency from capturing Artist's volumetric data to display it on an Audience Member's screen. However, it is hard to establish an end-to-end connectivity in today's network with guaranteed network service quality between devices at Audience Members' sites and logical computing nodes (e.g., VMs and containers) at the Centralized Cloud. In many cases, access network, long-distance network and data center network are provided by different service providers. Service level may differ and there is no co-operation mechanism between them to guarantee the network quality. Even though photonic networks are fully deployed and managed within each network domain, packet congestion may still occur at packet-based network equipment at exchange points between the domains. Thus, a mechanism to flexibly deploy end-to-end network connectivity with guaranteed service quality is crucial for satisfying the requirements in the Benchmark Model.

4.2.3. Virtualization Overhead for Tag-Based Multi-Tenancy Operation

The cloud is a vast resource pool consisting of hundreds of thousands of servers, storage, and network equipment, connected through the software-defined network technology. For example, some cloud service providers have built cloud data centers consisting of multiple CELLs (groups of system resources) connected through a clustered network with many routers which control network flow between resources [Cloud DC]. Some cloud service providers have built their data centers based on the Leaf-Spine architecture and deployed many software-based configurable network switches to establish virtual private networks and control packet routing [A. Singh]. This means when the network packet is sent out from any of the user instances, a sort of tag is added to the packet header by the hypervisor or smartNIC of the host machine to establish the virtual private network. When the network switches/routers receive the packet, they must examine the packet header to determine the network path to forward it, apply throttling, and/or perform a security check. Such a networking process increases network latency and consumes a lot of energy. In addition, significant network latency decreases the performance of various user workloads. This is because the CPU cycles may be consumed for no reason while waiting for a network I/O. As a result, a massive amount of energy is also consumed in the user plane.

4.2.4. Non-Orchestration Mechanism for Dynamic Network and Computing Resource Allocation

In this use case, Audience Members join/drop this service dynamically during the event, and this service should be scalable depending on the number of Audience Members. But in today's network service, the application service provider needs to statically have a network that can cover the maximum data traffic for this use case regardless of the scale of the event. Today's network service provider does not provide dynamic network resource allocation between customer premises and edge data centers as well as inter-data center. Regarding computing resources, today's cloud service may dynamically allocate or release the required computing resources requested by the application service provider. But there is no co-operative mechanism that the application service provider can dynamically allocate both

the network and computing resources. Therefore, the application service provider needs to allocate the network and computing resources separately.

4.2.5. Insufficient Resource Utilization Caused by Box-Oriented Computing Platform

Cloud vendors offer a wide variety of selection of computing instances in terms of the number of vCPUs and memory size. But as for instances supporting specific accelerators, the variety of the instance types is often limited, and this limitation lowers the hardware utilization rate. For example, cloud vendors offer an instance supporting a GPU, and its typical configuration has 96 vCPUs and 8 GPUs. But in this use case, functional nodes do not fully utilize all of vCPUs and GPUs provided in instance types. Some of vCPUs and GPUs may be idle.

This inefficiency is because available CPUs and accelerators highly depend on the hardware configuration in CPU centric architecture. Only CPUs and accelerators on the same physical computing node can be available from a single instance. In addition, as data transfer is a significant workload for computing infrastructure, processing big unstructured data in the NIC, i.e., SmartNIC, is becoming a common practice. However, this approach depends on the processor capacity of the smart NIC. IOWN GF technology should support a way to relax these limitations and improve their flexibility to realize composable disaggregated computing.

4.2.6. CPU Overwhelmed by Software-Based Data Transfer

The DPD has many receive-and-forward type processes. For example, Renderer nodes need to receive lots of data such as Scene Composition Data, Artist's volumetric data and send the presentation data to customer premises.

In addition, the service mesh for microservices requires an additional transparent application-level proxy in each computing node. In today's implementation, a socket interface is commonly used for these processes, and CPUs handle TCP/IP data transfer. CPUs execute the following three read and two write tasks:

- Reads the incoming packet from the NIC buffer and writes it to the socket buffer.
    1. Reads the packet in the socket buffer to verify the checksum
    2. Reads the packet in the socket buffer and writes its payload to the application buffer
- As the size of incoming data gets bigger, CPUs get overwhelmed with interruption, context switching, memory copy, and the TCP/IP protocol handling tasks. This has long been an issue in the computing industry. TCP Offloading (TCO) will alleviate this issue to some extent. But, without DMA, CPU resources will be consumed for data transfer from the NIC to the processor. Increased CPU resource consumption results in higher energy consumption.

4.2.7. Increased Energy Consumption and Latency Caused by Data Hub Tier

Considering there are one million Audience Members, and their positions, poses and movements are updated 30 times per second, then the Virtual Space Creation node has to accept data update at the speed of 30 million of record per second. And if real-time interactions need to be supported, of course the latency for data updates should be far below 33 [ms], i.e., a single digit millisecond. It is also required to support massive data queries. The data required for rendering must be extracted from virtual space creation to GPU at high speed and short latency. None of today's cloud-based DB/KVS may not be able to support these requirements.

4.2.8. Too much latency caused by the distance from Customer Premise to Centralized Cloud

In the key requirements defined in the Benchmark Model, there are some latency requirements like motion-to-photon latency. Motion-to-photon latency is the time delay from the time that Audience Member's head/eye-position movement is captured to the time that the presentation data generated by rendering a Virtual Space are fully reflected to the screen of the Audience Member's display. This latency requirements is specified as 10 [ms]. However, light can only travel about 2,000 [km] in 10 [ms] in fibers. This means that in a central cloud model, the round trip time is 10 [ms] from an Audience Member who is 1,000 [km] away to the central cloud. Therefore, motion-to-photon latency can not be satisfied when we assume additional delay time such as input, rendering, and displaying overheads. This model is not feasible for nation-wide deployment.

5. Reference Implementation Model

This section describes the ILM RIM for the given Benchmark Model, leveraging the IOWN GF architecture and technology. To fully realize the merits of Open APN and DCI, wired connectivity is assumed for this RIM. We analyzed what functional nodes are required and how much data rate needs to be transferred throughout the system. Also, we raised technical issues to meet the key requirements, such as 10 [ms] motion-to-photon latency and scalability of the virtual live music events. This section shows the basic strategy to solve these technical issues with IOWN GF technology. Then, it describes a detailed design of a geographically distributed data pipeline and its function nodes for this Use Case. Finally, the last section describes the expected benefits.

5.1. Basic Strategy of IOWN GF Architecture and Technology Adoption

The ILM UC requires high-bandwidth for volumetric video with 6 DoF, very low latency for interactive entertainment between Artists and Audience Members, and scalability from small to very large-scale concerts on the nationwide.

This subsection briefly revisits and summarizes each of these issues and highlights how we can overcome the issues.

5.1.1. Shift from Centralized Cloud Computing to Distributed Cloud Computing

The ILM UC requires high-bandwidth data flows between its functional nodes, but routing and processing these data flows inside a single data center is inefficient and not optimal in terms of latency. The target motion-to-photon latency is less than 10msec.

As a solution, IOWN GF has Open APN and DCI architecture. Open APN allows a single optical path to span multiple segments. This enables end-to-end communications with deterministic quality.

In addition, a DCI is provided on the Open APN to provide a distributed heterogeneous computing and networking environment for end-to-end applications in the cloud, at the edge, and on customer premises.

Thanks to DCI, it is possible to use servers per region that can greatly reduce motion-to-photon latency. This geographically distributed data pipeline will be described in 5.2.

5.1.2. Shift from Best-Effort Networking to Deterministic Quality Networking

This ILM UC is interactive, and simultaneity is important because of the requirement for simultaneous participation and dancing from different regions at various distances. The requirements for this ILM UC are different from typical live music streaming events and are very demanding.

In addition to the geographically distributed deployment challenges described above, the ILM UC requires communication lines with guaranteed quality of service regarding bandwidth, low latency, and reliability between their functional nodes. However, current technology does not provide a standardized or cost-effective way to establish such high QoS links over long distances.

5.1.3. Shift from Static to Dynamic Resource Allocation

For this ILM UC, the number of Audience Members depends on the size of the event, which may be regional or national. In other words, the constant rate approach is wasteful because the bandwidth and processing volume vary greatly depending on the number of participants. It is necessary to make the operation scalable and energy efficient according to the number of participants.

To reach its efficiency and performance goals simultaneously, the ILM UC will need to break away from such constant-rate approaches and instead employ a dynamic resource allocation approach. The dynamic resource allocation approaches reduce the complexity of spreading computation over various accelerators and automatically adjust to varying load conditions expected from ILM UC that help save energy.

5.1.4. Shift from Box-Oriented Computing to Disaggregated Computing

This ILM UC requires extensive resources for processing interactive 3D data. However, there are challenges and requirements within the data center, where the combination and ratio of general-purpose CPU and GPU accelerators needed varies over time depending on the external environment, such as day of the week, time of day, and sudden irregular events, as well as the number of Audience Members participating in this ILM UC. The current fixed hardware configuration of cloud servers leaves a large amount of hardware installed inefficiently utilized.

To solve this problem, the DCI architecture includes disaggregation technology that dynamically adapts to changing load conditions by combining the necessary hardware parts for individual server hardware configurations. This increases overall hardware utilization and reduces the amount of hardware required for this ILM UC.

5.1.5. Shift from Software-Based Data Transfer to Hardware-Based Data Transfer

The global resource management provided by the Open APN and DCI will remove the hurdle of provisioning low-latency and high-bandwidth connections between endpoints in a straightforward manner. With this solution in place on the infrastructure level, the remaining issue from this ILM UC requirement is that network latency must be reduced. This means both latencies due to forwarding data through the network and communication latency due to protocol processing.

The DCI functional architecture proposes to use the following two mechanisms to reduce transmission latency and increase the efficiency of communication in terms of processing time:

- Memory copy reduction: Removing the intermediate step to copy data to kernel memory once before the transmission reduces transfer latency by reducing the total number of required memory accesses for sending and receiving data over the network.
- RDMA-based data transfer protocols: Transfer protocols such as RDMA that lend themselves to hardware offloading contribute to possible link utilization, and latency reduction.

5.1.6. Shift from a Single-Function, Low-Speed Data Hub to a Multi-Function, High-Performance Data Hub

Virtual Space Creator generates 3D Scene Description and sends them to the regional edge data centers. The amount of data is large, and real-time performance is required.

In today's implementation model, each Data Hub service takes a significant amount of processing time, so if the system is built by combining multiple Data Hub services, then the real-time requirement will not be fulfilled. It also creates other issues in system cost and energy consumption because the large amount of data needs to be transferred between Data Hub services. Such Data Hub relevant issues, combined with the inefficiencies of the cloud's internal network mentioned above, make issues, such as energy consumption, more difficult.

As mentioned in the Annex, there are ideas for decentralization processing, which we define as future study items.

5.2. Geographically Distributed Data-Pipeline of IOWN GF-based RIM

Figures 5.2-1, 5.2-2, and 5.2-3 show an overview of an initial form of an IOWN GF-based RIM. In contrast to the centralized approach in today's implementation described in section 4, the RIM adopts a geographically distributed approach leveraging IOWN GF technologies such as Open APN and the DCI subsystem. This approach contributes to network-wide flexible provisioning of large-capacity and application-dedicated networking and computing resources. It significantly improves the end-to-end latency and system efficiency of this ILM UC. For more details of DCI, see subsection 5.2 and section 7 of the DCI document [IOWN DCI].

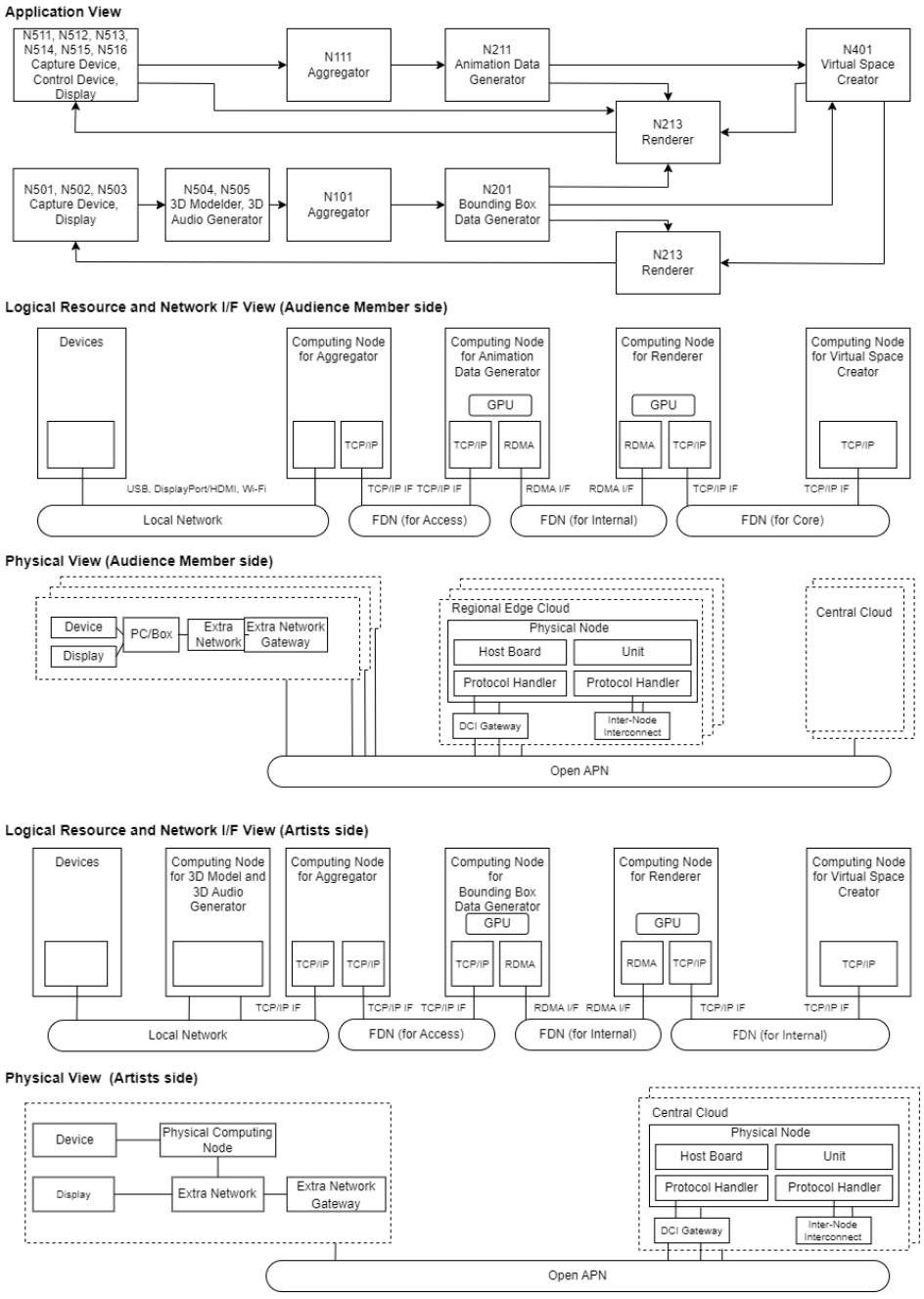

Figure 5.2-1: Overview of an IOWN GF-Based RIM, Application View

As shown in Figure 5.2-1, the RIM assumes three layers of the vertical distribution of computing sites, that is, customer premises, regional edge clouds, and central clouds.

From here, we will explain four measures as methods to improve latency, wide bandwidth, scalability, etc. specific to this UC as follows,

- Geographically distributed Edge servers
- Dynamic resource allocation
- Inter data center multicast over Open APN

- Uncompressed video

*note: The protocol between the display of the customer premise and the Renderer node has not been determined for both Artist side and Audience Member side. It will be discussed in the PoC phase.

5.2.1. Geographically Distributed Edge Servers

Compared to processing everything in one place in the central cloud, the Renderer can be geographically placed near the Audience Member to shorten the transmission distance, and the Open APN can be used to connect between them to reduce motion-to-photon latency by utilizing a network with fewer fluctuations.

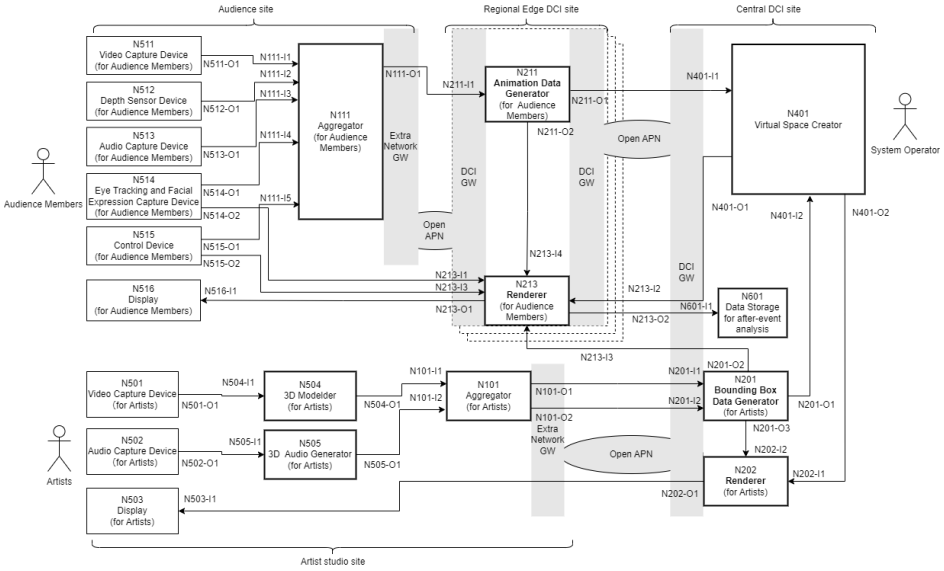

Figure 5.2-2: Overview of an IOWN GF-Based RIM

The points to improve the system are as follows,

- The location for each application function is as follows,
  - The aggregator is located in the customer premise, and the customer should prepare this node unless this is DCI managed service.
  - DCI in the regional edge cloud is located in the Regional site where will install DCI same as IOWN AM-S RIM. For example, Edge DCs could be placed on each network node as defined in JPN48 network model and or Telco central offices in the case of Japan.
  - Animation Data Generator node and Renderer node on LSN (for Audience Members) in regional edge DCI.
  - DCI in central cloud is located in major cities such as Tokyo/Osaka which will install DCI the same as IOWN AM-S RIM in the case of Japan, for example. Virtual Space Creator node, Bounding Box Data Generator node, and Renderer node for Artists are on LSN in central DCI. The data storage node is the data service tier connecting to regional edge DCI.
  - Artists studio systems are located in major cities such as Tokyo/Osaka and nearby central DCI that operate Virtual Space. Aggregator node (for Artists) is on Artists studio site, and 3D Modeler node (for Artists) and 3D Audio Generator node (for Artists) are on LSN in central DCI.

5.2.2. Dynamic Network and Computing Resource Allocation

In ILM UC, the number of Audience Members may vary depending on the type of events (e.g. exclusive and very limited users only, nationwide service with a very large number of Audience Members, etc.). Even during an event, Audience Members may join or drop the event and the number of Audience Members is dynamically changed accordingly. Regarding computing resources, auto-scaling services of virtual machine instances/containers and serverless applications are available in today's cloud services. On the other hand, regarding the network between an Audience Member premise and a data center and inter-data center network, there is no autoscaling and dynamic load-balancing of network links with the allocation of computing resources for applications.

In order to address this issue cost-efficiently, DCI has the Infrastructure Orchestrator that allocates or releases LSNs and APNs necessary for ILM UC dynamically by calling Infrastructure Orchestrator from the application, we can dynamically allocate necessary resource and this contributes to the energy efficiency and operation costs.

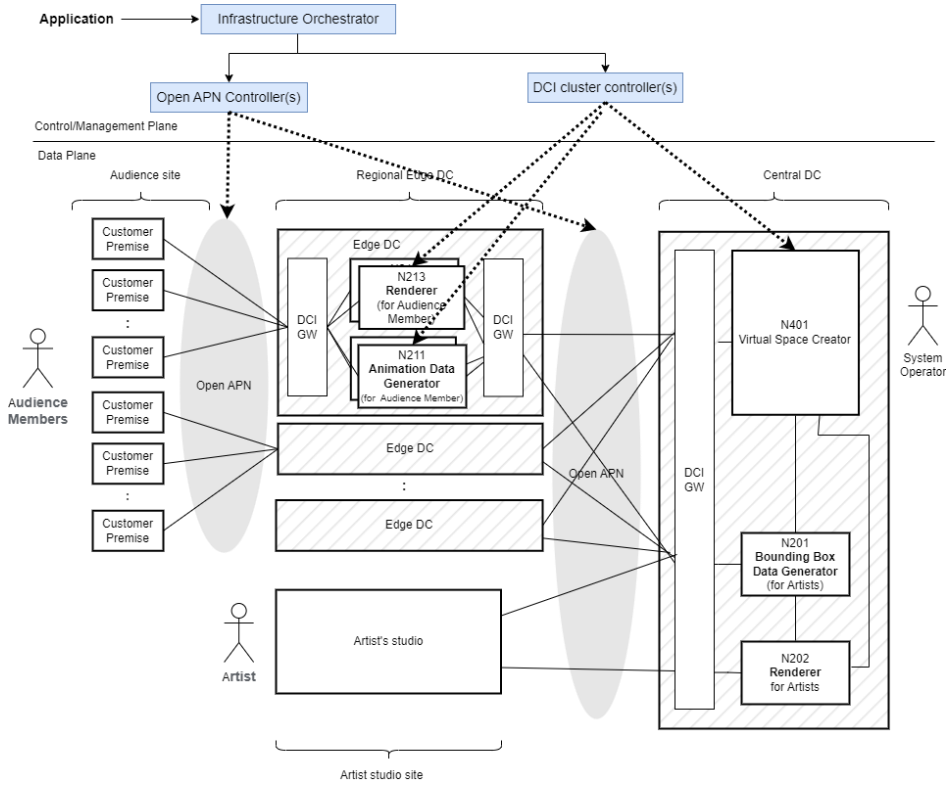

Figure 5.2-3: Dynamic resource allocation by Infrastructure Orchestrator

5.2.3. Inter-data center multicast over APN

In the ILM RIM , there are two types of data that need to multicast from central data center to edge data center as follows:

1. Artist's volumetric video and audio data
2. Scene Composition Data

As the first version of RIM, we take the approach to establish peer-to-peer (p-to-p) connections between data centers in APN layer and multicast the data from a data center to a data center by hop-by-hop transfer.

Figure 5.2-4 shows one possible example of how the data transfer from a central data center to edge data centers could be implemented. Firstly, they establish p-to-p connections from the central data center to regional edge data

centers in each region and establish p-to-p connections from the regional edge data centers to adjacent local edge data centers. Then the multicast packets are generated at an LSN in the central data center and the packets are transferred to the regional edge data centers through APN. The packet copy for multicast could be performed directly by a switch inside the DCI Cluster. The DCI GW of edge data center receives the packets and the DCI GW transfers them to the LSN. Additionally, the packets are further forwarded to the adjacent edge data centers. For this step, too, the required packet copy could also be performed directly by a switch within the DCI Cluster. By this approach, Artist's volumetric video data and Scene Composition Data are transferred to all of the necessary edge data centers.

For other potential options, please refer to the detailed protocol description in Annex E.

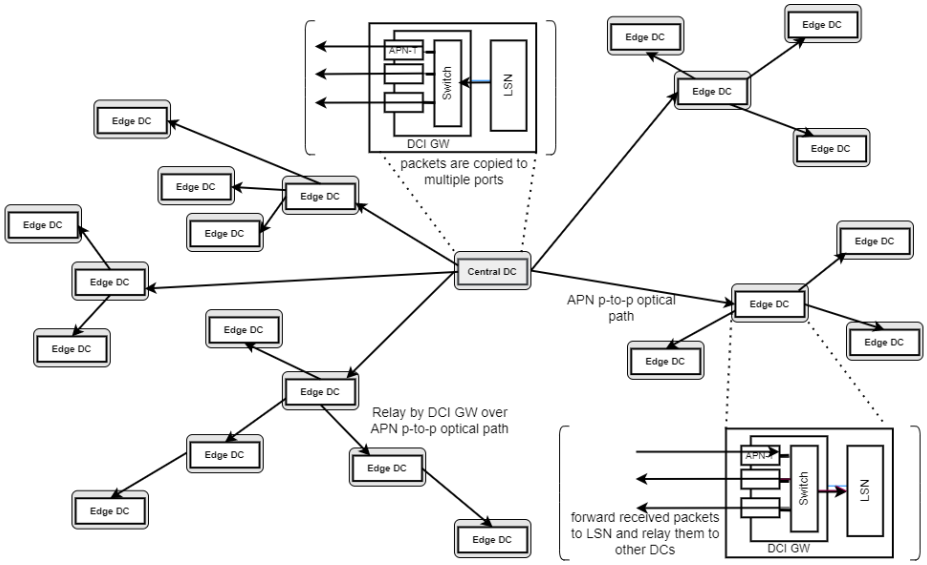

Figure 5.2-4: Example of multicast over APN from central data center to edge data centers

5.2.4. Uncompressed video to the customer premises

There are multiple possible methods for communication between the Renderer node and the Display of Customer premise. An overview of each is given below.

- SDI over IP: Using SMPTE ST2110 and transmitting Serial Digital Interface (SDI) protocol using SmartNIC at the server side. As for the client side, it is possible to convert to HDMI or DisplayPort from SDI protocol.
- RDMA: In order to keep low motion-to-photon latency between the customer site and Functional card in the edge DCI Logical Service Node, RDMA over APN to share uncompressed raw data with low overhead.
- DisplayPort/HDMI over APN: It might be possible to send uncompressed video data more directly with or without Ethernet because APN is protocol-free. There are ways to implement this protocol.

Please refer to the detailed protocol description in Annex F.

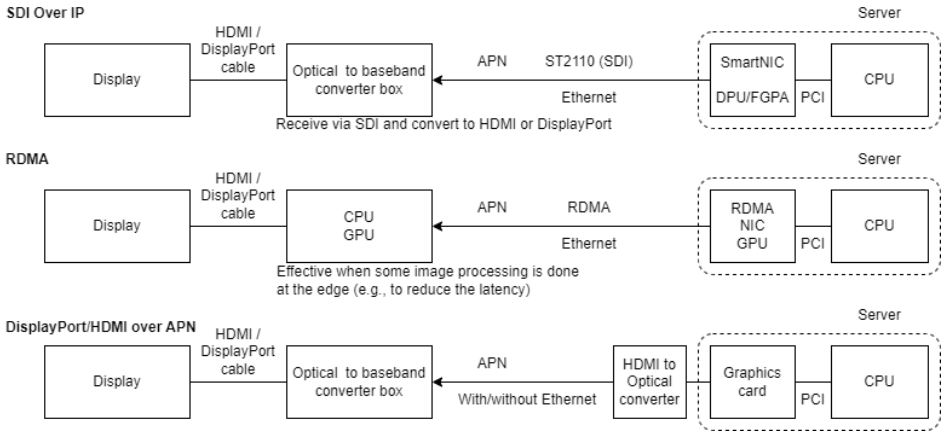

Figure 5.2-5: Communication between the Renderer node and the Display of customer premise

5.3. Application's Functional Node Structure of IOWN GF-based RIM

This subsection describes a concrete design of the major functional nodes in the geographically distributed data pipeline described in Figure 5.2-1.

5.3.1. Renderer node for Audience Members

There are several solutions that locally share raw data (user payload) between two devices such as peer-to-peer DMA via existing PCIe and CXL3.0 supporting CXL device to CXL device/peer-to-peer interconnection or shared memory pool. Figure 5.3.1-1 shows the example of communication between Renderer and Display with sharing the GPU memory. This figure focuses on sharing GPU memory between devices.

In order to keep low motion-to-photon latency between the customer premise and the functional card in the edge DCI Logical Service Node, RDMA over APN to share uncompressed raw data with low overhead is one of the candidate solutions.

The data stored in the functional card can be shared with other functional cards or DRAM / NVMe / pMEM to connect from the Virtual Space Creator node in central data center to the edge Rendering node by using RDMA over APN too. The following diagram illustrates GPU + SmartNIC/DPU/IPU capability that we are exploring.

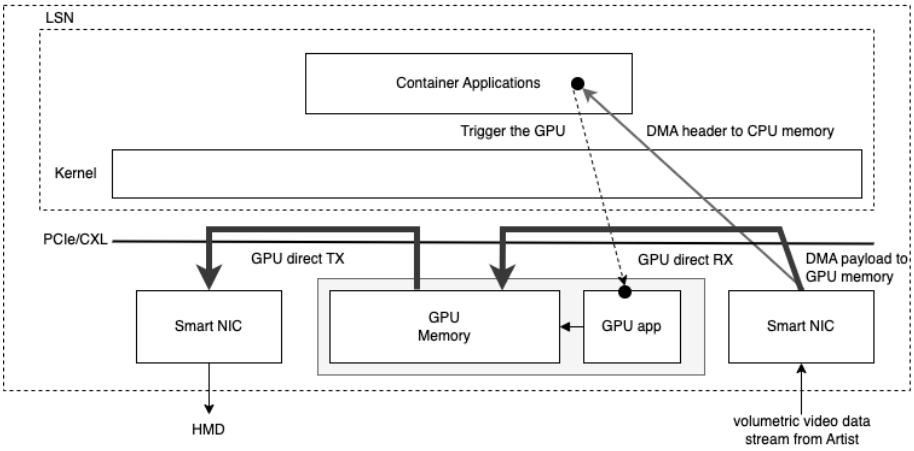

Figure 5.3-1: Example of communication between Renderer and Display

5.3.2. System Topology around Renderer nodes and Virtual Space Creator nodes

Providing a scalable and real-time metaverse service that can accommodate more than thousands of people at a reasonable price is very challenging. Therefore, achieving this goal requires not only more advanced technology such as IOWN GF's technology stack but also a well-crafted system architecture and more advanced technology.

For instance, as discussed so far, by performing rendering on the center side, it is assumed that the GPU resources can be shared, and each Audience Member will be able to experience an immersive Metaverse service simply by owning a cheap VR device. However, if one GPU card is required for each participating Audience Member for rendering, the service price will not be attractive.

It is also necessary to collect data from thousands of participants, map it in virtual space, and detect collisions. In order to build real-time immersive services, it will be necessary to update the animation data representing each user's actions more than 30 times per second. In other words, it is necessary to update the animation data 100,000 times per second and to perform subsequent processes such as collision detection. In order to stably execute such data processing, the data management infrastructure must be designed to scale out well.

Based on such a perspective, please refer to Annex H which discusses data distribution among renderer nodes and Virtual Space Creator nodes, Annex G. which discusses techniques and system configuration required for resolution adjustment to reduce rendering workload, and Annex K. which discusses topology options and their consideration points around renderer nodes and VSC nodes.

5.4. Expected Benefits

This section discusses the benefits of IOWN GF technology in this ILM UC. As a highlight, we compare today's centralized cloud-based implementation model with the IOWN GF-based RIM shown in below.

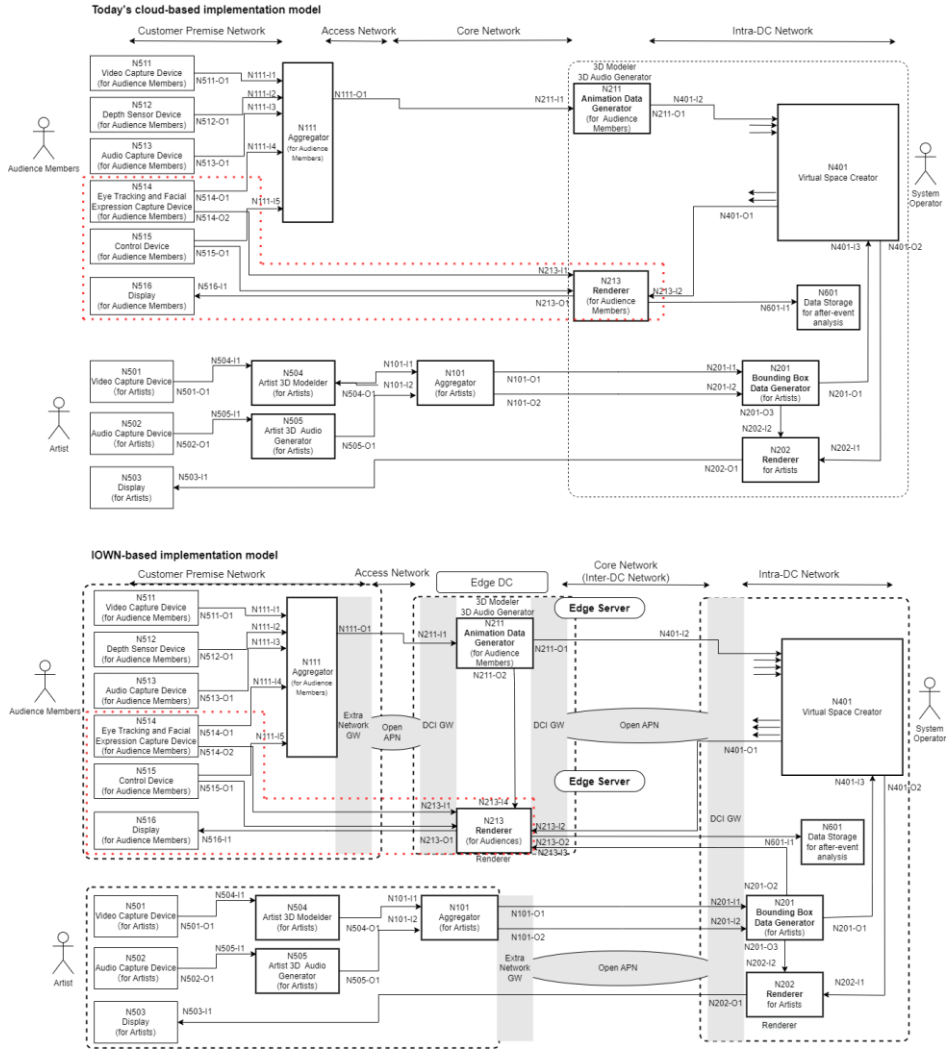

Figure 5.4-1: Today's centralized implementation model and IOWN GF-based RIM

5.4.1. Network

5.4.1.1. Access Network

There is an access network that connects to equipment on customer premises for Audience Members and a network that connects to equipment on the artist's studio premises. Due to the amount of network traffic below, we first considered the uncompressed data to have low latency.

5.4.1.1.1. Workload profile

- Upstream traffic as sensor data
  - Video and Audio: 7.23 [Gbps] * # of Audience Members
  - User control data 11.5.12 [kbps] * # of Audience Members
  - Volumetric Video, 30 [fps]: 56.35 [Gbps], 60 [fps]: 112.74 [Gbps]
- Downstream traffic
  - Uncompressed video, 23.06 [Gbps] (8bit 420), 69.14 [Gbps] (12bit 444)

5.4.1.1.2. Expected Benefits

This ILM UC has very high network traffic due to the use of high-density volumetric video tailored to human visual characteristics proportional to the number of Audience Members. Therefore, the required network equipment and power consumption will equally sizable. IOWN GF Open APN technologies, which provide high bandwidth and reduced power consumption, help mitigate this issue. In addition, the IOWN GF Open APN technologies that shorten the latency by reducing optical-to-electrical conversion and establishing a direct communication path between the customer premise site and the cloud data center will contribute to building the real-time live music event.

5.4.1.2. Core Network

The core network provides wide-area data transfer services and sends 3D scene data from the Virtual Space Creator node to the Renderer nodes, which are the regional servers. RDMA is used in Bounding Box Data Generator (N201) to Renderer for Audiences Member(N213) to transmit Artist's volumetric video data to the Renderer in each Audience Member region.

5.4.1.2.1. Workload profile

The amount of network traffic that comes from or flows into the regional data center over the core network is the data rate per Virtual Space is 371.2 [Mbps] at 30 [fps], 742.4 [Mbps] at 60 [fps] x # of Audience Members and Artist's volumetric video data in total. In addition to that, the IOWN GF-based RIM has the following network flows between the regional edge clouds and the central cloud. It should be noted that there are multiple edge clouds connected to one central cloud.

- From regional edge clouds to central clouds
  - N211-O1 Animation Data Generator to VSC, 24.71 [Mbps]
  - N211-O2 Animation Data Generator to Renderer, 24.71 [Mbps]
- From central cloud to regional edge clouds
  - N401-O1 Scene Composition Data, 371.2 [Mbps] per Audience Group
  - N201-O2 Volumetric video, 30 [fps]: 56.35 [Gbps], 60 [fps]: 112.74 [Gbps]

5.4.1.2.2. Expected Benefits

In today's implementation model, the core network may transfer data between the customer premise site to the cloud over a long distance. To forward network packets over such a long distance, it must have more than several optical transceivers, and the latency will become significant at the end. In addition, there need to be optical-to-electric conversions in the optical transceiver, to control the packet routing and amplify the signal. And routing consumes even more energy and results in a lot of CO_2 emissions.

In the IOWN GF-based RIM, these issues are mitigated with the following perspectives.

- The all-photonic network does not use electrical packet switchers or routers and does not involve optical-to-electrical conversion, so network latency and power consumption can be reduced.
- Regional edge clouds, which have relatively small resources, are used to communicate with Audience Members directly to lower the latency. The single data traffic size is not so large, but it is proportional to the number of Audience Group, e.g., so if the music event is for 300,000 Audience Members, it will be 100 times of the above data size, that is about 37.1 [Gbps]. By adopting Multicast over APN, the Music event will be scalable because both network and computing resources for rendering are flexible.

5.4.1.3. Intra-data center Network

Cloud data center is a huge resource pool, and the role of intra data center network is to establish a virtual network for its user, connect resources for the user flexibly to run the user workload, and mediate communications with an external

system. Since multiple user tenants share the cloud, it is essential to reduce the overhead to control multi-tenancy, to reduce latency, and increase bandwidth.

5.4.1.3.1. Workload Profile

For this ILM UC, the network traffic relevant to cloud services are summarized below:

- For today's implementation model
    - Ingress traffic to the cloud data center
        - It all goes to cloud data center, 5.25 [Gbps] * # of Audience Member will be sent to the Virtual Space Creator.
    - Egress traffic from the cloud data center
        - The data for all Audience Members and the Artist's data are summed up and out, so that the data rate is 5.25 [Gbps] + 56.35 [Gbps] at 30 [fps]

For IOWN GF-based RIM

For example, between the central cloud and the Renderers located at the regional edge, which are connected at high speed by Open APN, large data such as images needed for rendering in the Audience Member are transmitted directly to the Renderer only. Only the Animation Data used for collision detection, etc., is transmitted to the central cloud, thus reducing the transmission capacity and making the structure easier to scale.

- Ingress traffic to the cloud data center
    - Since the Animation Data and the Avatar Data can be separated, it needs to send 371.2 [Mbps] (in case of 30 [fps]), 742.4 [Mbps] (in case of 60 fps) [N401-O1] for all Audience Members.
- Egress traffic from the cloud data center
    - Since only the Artist's Bounding Box data and 3D Scene, and not the Artist's volumetric video, are transmitted, its date rate is 371.2 [Mbps] (in case of 30 [fps]), 742.4 [Mbps] (in case of 60 [fps]) [N213-I2].

5.4.1.3.2. Expected Benefits

In today's implementation model for cloud infrastructure, there are many software-controlled switches/routers to establish virtual private networks for tens of thousands of users and connect hundreds or thousands of resources to build a private environment for the user on top of a large resource pool.

To establish the virtual private network, a sort of tag is assigned to the packet header by hypervisor or Smart NIC of the host machine running user workloads, and these switches/routers read it to control packet routing and apply throttling in order to handle a huge amount of network traffic in a stable and SLA / SLO compliant manner.

Such software-defined network controls increase latency, consume more energy, and become more costly as the size of the resource pool grows.

Typically, the network packets must go through 5-10 switches/routers for internal communications, and each time they go through them, a specific network latency, such as 10 -15 microseconds, will be added. Thus, communications between two resources would be more than 100 microseconds. For external communications, more switches/routers are needed, and additional controls for private connection management and global IP mapping management; thus, the minimum latency accessing the cloud from the outsider is typically 0.2 – 1 milliseconds.

It should be noted that these delays would be accumulated in the actual workload. For instance, when an application instance tries to update data in a DB instance in the cloud, the DB instance will contact a block storage server. The

block storage will contact other block storage servers to replicate data, etc. There, when performing distributed processing with multiple servers in the cloud, effective efficiency is generally lower than on-premises.

In addition, the intra cloud network service is often oversubscribed, resulting in even more uncertain delays during network congestion. There are also the reasons for the limited use of RDMA in the cloud. Because RDMA is a low-level protocol that requires additional controls to ensure the data is delivered and stored properly at the destination site, it is challenging to apply it to a production workload under such a large and fluctuating latency.

In summary, today's centralized cloud implementation models are less efficient at distributed data processing than on-premises and cause higher system costs and energy consumption when assuming constant workloads.

In the IOWN GF-based RIM, these issues will be mitigated by Open APN and DCI technologies, Open APN will establish the direct communication path between user instances, which belongs to the virtual private network of the tenant, by wavelength and mode of optical signals. DCI technologies control the establishment of direct communication paths dedicated to the tenant. That means, from a resource pool consisting of hundreds of thousands of servers, DCI pick up a requested number, such as 100 server for the tenant, and connects them through direct or near-direct communication paths. As a result, it is expected that latency between resources will be reduced, user workloads will be executed more efficiently, and the cost and energy consumption for the inter-cloud network will be reduced. In addition, using the IOWN GF model opens up the possibility of applying RDMA to a production deployment on top of a larger resources pool.

5.4.2. Application's Functional Node

As the first step of the evaluation of applications' functional nodes, we focused on dominant parts of data pipelines placed on Renderer node (N213) and Bounding Box Data Generator (N201). Assumption is we have multiple edge servers for primary region to lower the motion-to-photon latency.

5.4.2.1. Renderer Node

The role of the Renderer node for Audience Members is to continuously receive volumetric video data and process rendering with the User control data of each Audience Members.

5.4.2.1.1. Workload Profile

The Renderer node handles the rendering of the Virtual Space transferred from the "Virtual Space Creator node" in the central data center and with Artist and Audience Member data. The following shows the workload required to support this UC. The details of the DPD are shown in Annex C.

N201-O2: Data rate for volumetric video

30 [fps]: 56.35 [Gbps], 60 [fps]: 112.74 [Gbps]

As for motion-to-photon latency, if the display has a mechanism to display at regular intervals, it is necessary to wait for it, and it depends on the framerate of the display device. So, to achieve a motion-to-photon latency of 10 msec, rendering and transmission etc. must be completed at a rate faster than the framerate to be displayed. See Annex H. for more information.

5.4.2.1.2. Expected Benefits

Utilizing the GPU hardware accelerator, the workload of the rendering process can be largely decreased compared to processing by CPU.

5.4.2.2. Bounding Box Data Generator Node

This node forwards the Artist's volumetric video data through the DMA/RDMA protocol and generates Bounding Box data and sends it to the Virtual Space Creator node. In this node, both data forwarding and data processing is processed in a hardware accelerator.

5.4.2.2.1. Workload Profile

N201-O2: Data rate for volumetric video

30 [fps]: 56.35 [Gbps], 60 [fps]: 112.74 [Gbps]

5.4.2.2.2. Expected Benefits

Utilizing an RDMA-type network protocol overall photonics network, the workload of data forwarding can be largely decreased. By combining the shared memory techniques, the IOWN GF-based RIM can streamline the data plane and significantly eliminate the overheads.

6. Conclusion

We developed the ILM RIM for the given benchmark model to the ILM UC leveraging the Open APN technologies and DCI technologies.

To quickly recap some of the findings:

Using Open APN technologies, this RIM can deliver:

- RIM with Open APN obtains benefits for resource allocation flexibility
- high bandwidth and reduced power consumption across the core network and access network collecting massive data from capture devices (cameras, depth sensors, etc.)
- low latency by reducing optical-to-electrical conversion, as well as establishing a direct communication path in distributed clouds between the customer premise site and the telco edge/core site, which will help to build interactive virtual live event service.
- popularized ILM Services for widely spread Audience Members, which is not limited to "the power users" who have professional class PCs in their premises, by accessing the resource of DCI with low latency.

Using DCI technologies, this RIM can deliver:

- RIM with DCI technologies obtain benefits for resource allocation flexibility from heterogeneous and disaggregated device resources pool that can allocate the application's functional node at the desired location (i.e. edge data centers) to perform high data performance in hardware rate by function dedicated network interface card such RDMA capable Smart NIC, DPU [DPU], and IPU [IPU], in addition to CPU, GPU, and persistent memory. This flexible resource allocation in disaggregated infrastructure helps in the reduction of cost for computation and reduction of power consumption.

In this RIM, we focus our study on the real-time case, where IOWN GF technology can be used more effectively. Analyzing and utilizing the past data is in the scope of our future work.

Among the various use cases of IOWN GF, the realization of ILM UC requires high-bandwidth, low-latency unicast in the layer of Open APN communication in end-to-end networks, including access domain networks between carrier edge data centers and customer premises. To achieve this, the Open APN Architecture (OAA) and its extension to the access domain are being discussed in IOWN GF.

Globally managing resources to enable end-to-end network service quality guarantees is the key enabler to dynamically and cost-effectively configure the optical transmission lines and conversion elements that underlying networks, such as an Open APN [IOWN GF Open APN], will offer. We will need further study through PoC.

Based on this RIM, we are planning to evaluate the feasibility of leveraging IOWN GF technologies through PoC.

Abbreviations

0-9

3D, Three Dimensional

B

bps, bits per second

C

CG, Computer Graphics

D

DB, Data Base

DMA, Direct Memory Access

DC, Data Center

DCI, Data-Centric Infrastructure

DoF, Degrees of Freedom

DPD, Data Pipeline Diagram

DSC, Display Stream Compression

F

fps, frames per second

G

G-PCC, Geometry-based Point Cloud Compression

gRPC, gRPC Remote Procedure Call

H

HMD, Head-Mounted Display

I

ILM UC, Interactive Live Music Use Case

ILM RIM, Reference Implementation Model of the Interactive Live Music Use Case

K

KVS, Key Value Store

R

RDMA, Remote Direct Memory Access

ROI, Region Of Interest

RTP, Real-time Transport Protocol

REST, Representational State Transfer

S

SDI, Serial Digital Interface

SLA, Service Level Agreement

SLO, Service Level Objectives

ST2110, SMPTE ST 2110 Professional Media Over Managed IP Networks

T

TCP, Transmission Control Protocol

ToF, Time of Flight

U

UDP, User Datagram Protocol

UDS, Unix Domain Socket

V

VM, Virtual Machine

Terms and Definitions

Reference Case	Detailed description of the use case with specific conditions for determining functional and non-functional requirements, output/input data flow, system size, and parameters in order to make it accurate to evaluate implementation models by measuring selected metrics in the specific conditions.
Virtual Space	Virtual Space is a virtual live music venue. Audience Members can move freely around the live music venue, and the images seen by each participant are individually generated by rendering from each viewpoint. This is conceptual, and the data is represented by separately defined Scene Composition data and visualized by the rendering process in the renderer.
Interactive Music Service	The overall service on the network side for the Use Case.
Audience Member	Person who participate from their homes, karaoke rooms, and virtual spaces.
Audience Group	A group of Audience Members in the same Virtual Space. Form a group in the same geographical location. Form a geographically dispersed group. e.g., to gather as a hobby group, SNS connection or randomly selection by the system.
Scene Composition Data	Output data from the Virtual Space Creator. It is some part of the elements necessary to construct Virtual Space, which is visualized by the Renderer. The elements of Scene Composition data are as follows; Artists' Bounding Box data(includes determined position and direction) Audience Members' Animation data(includes determined position and direction) Concert Hall component such as 3D model of stage, seats, light position, speaker position, etc. 3D audio data for Virtual Space CG effects including lighting and ROI (Region of Interest) Information and provide recommended view port for Audience Member They are expressed in a scene description language and contains both static and time-varying dynamic information. In the 3D scene generation phase, the 3D models are not necessarily required, but only the position and motion information (Bounding Box data and Animation data: time-varying dynamic information) of the 3D models can be used to construct the 3D scene. In the DPD, multiple inputs arrive at the Renderer. Some of the static information can be sent to the Renderer in advance, before going live. They are then integrated before being used for rendering.
SmartNIC	A programmable accelerator that makes data center networking, security and storage efficient and flexible.
Avatar Data	3D model data without animation.
Animation Data	Time series data of position and direction of joints to move the Avatar data. The Animation data will be sent to the system in real-time during the live event to move the 3D model, Avatar data. This Animation data includes not only body movements but also facial expressions and eye movements.
LOD	LOD stands for "Level of Detail" and means "degree of detail. It is a method of reducing the computational load of a scene by controlling the number of polygons in the model according to the distance from the camera.

References

[IOWN AM-S RIM]	IOWN Global Forum, “Reference Implementation Model (RIM) for the Area Management Security Use Case,” 2022.
[IOWN AIC UC]	IOWN Global Forum, “AI-Integrated Communications Use Case Release-1,” 2021. https://iowngf.org/use-cases/
[IOWN CPS UC]	IOWN Global Forum, “Cyber-Physical System Use Case Release-1,” 2021 https://iowngf.org/use-cases/
[IOWN GF DCI]	IOWN Global Forum, “Data-Centric Infrastructure Functional Architecture,” 2022.
[IOWN Open APN]	IOWN Global Forum, “Open All-Photonic Network Functional Architecture,” 2022.
[DPU]	Data Processing Units (DPUs), https://www.nvidia.com/en-us/networking/products/data-processing-unit/
[IPU]	Infrastructure Processing Units (IPUs), https://www.intel.com/content/www/us/en/products/network-io/smartnic.html
[IOWN Data Hub]	IOWN Global Forum, “Data Hub Functional Architecture,” 2022.
[RoCE]	InfiniBand Trade Association, “RoCE Accelerates Data Center Performance, Cost Efficiency, and Scalability,” 2017. https://www.roceinitiative.org/wp-content/uploads/2017/01/RoCE-Accelerates-DC-performance_Final.pdf
[DFD]	Lucid Software Inc., https://www.lucidchart.com/pages/data-flow-diagram
[IOWN GF Vision]	IOWN Global Forum, “Innovative Optical and Wireless Network Global Forum Vision 2030 and Technical Directions,” 2020. https://iowngf.org/white-papers/
[IOWN TECH LOOK]	IOWN Global Forum, “IOWN Global Forum System and Technology Outlook”, 2021, https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-System_and_Technology_Outlook_1.0-1.pdf
[JPN48]	Technical Committee on Photonic Network, “Japan Photonic Network Model“. https://www.ieice.org/cs/pn/eng/jpnm_en.html
[POPULATION]	Population Census https://www.stat.go.jp/english/index.html, https://www.stat.go.jp/data/nihon/zuhyou/n220200200.xlsx
[LEVIS]	Levi’s® Stadium https://maps.roadtrippers.com/us/santa-clara-ca/sports/levis-stadium-santa-clara
[EYETRACK]	Tobii eye trackers https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/how-do-tobii-eye-trackers-work/
[360RA]	https://biz.musicecosystems.jp/blog/sony-360ra_report_01/

Annex A Development Method of Benchmark Model

This Annex describes the definition of the Benchmark Model in IOWN GF and how to develop the Benchmark Model for target use cases used in Section 2.

A.1 What is the Benchmark Model in IOWN GF?

A “benchmark” is generally defined as an indicator for measuring the performance and operating speed of computer hardware and software. A “model” means a norm that should be the basis for judgment, evaluation, and action.

A Benchmark Model for the IOWN GF defines Reference Case, Metrics, and Evaluation Methods. A Benchmark Model is also developed for a target use case.

Reference Cases dig deeper into target use cases and clearly define the conditions for determining functional and non-functional requirements, output/input data flow, system size, and parameters. In some cases, they also describe features and implementation procedures.

Metrics are the numbers to evaluate the performance and quality of the implementation model. They are assigned to essential requirements when providing services.

Evaluation Methods are ways to evaluate the implementation model. When a new architecture or technology is incorporated into a system, the evaluation uses the same evaluation methods and metrics as before and compares the evaluation results.

A.2 How to Develop the Benchmark Model for the Target Use Case

The Benchmark Model for a target use case will be developed by selecting one use case from the collection of AIC/CPS use cases proposed by the IOWN GF.

First, for the selected use case, detailed functional, non-functional, and performance requirements will be defined, including key requirements in the use case document to define the Reference Case.

Next, capture the characteristics of the selected use case and define the Metrics that will be used as a basis for the Evaluation Method of the implementation model. This is the end of the development of the Benchmark Model.

Use case requirements change over time. In addition, various technologies and products will evolve, including IOWN GF architecture and technology. Therefore, the Benchmark Model needs to be rebuilt and evaluated to suit the changes.

A.3 How to Develop the Evaluation Method and Metrics for the Target Use Case

Determining the correct Evaluation Method is essential for making the right decision since there will be multiple technical options to design the system that satisfies the Reference Case's mandatory requirements. The Evaluation Method is also necessary to demonstrate how advanced the technology developed by IOWN GF is compared to the current technology.

A.3.1 What to Evaluate

According to the ISO/IEC 25000 [ISO/IEC 25000] or “Systems and Software Quality Requirements and Evaluation” (SQuaRE), the quality of the system can be assessed from the perspectives of, Functional Suitability, Performance Efficiency, Compatibility, Usability, Reliability, Security, Maintainability, and Portability. Since the current situation is in the early stages of technological development, we will focus on a limited set of Metrics, i.e., 1) response time, 2) system cost, and 3) energy efficiency, assuming that other mandatory requirements, such as the Minimum Required Response Time, Functional Suitability and Security, are met.

A.3.2 Evaluation Strategy

As discussed above, multiple metrics can be used to evaluate the system, so the question is how to assess the system's quality as a whole. In the IOWN GF, we decided to take the following approach:

- Designing the implementation model so that it satisfies all mandatory requirements, such as:
    - **Functionalities**, as described in the Reference Case
      The functionalities here include data aggregation, collection, recognition, decision making, notification delivery, etc., which are all required to build the system which supports the Reference Case.
    - **End-to-end response time:**
      The End-to-end response time means the total amount of time from the actual occurrence of the target event in the real world to the execution of necessary action by the system.
- Then, analyze the following Metrics of the designed implementation model, with equal weighting:
    - **End-to-end response time**:
      The End-to-end Response time is assumed to meet the number specified by the Reference Case as a mandatory requirement as described above, but on top of that, we think the shorter, the better technology; thus, we will evaluate it, too.
    - **System resources**
      The system resources are defined as metrics required to calculate the hardware and software costs, e.g., the number of CPU cores, GPU cards, and switches. When designing the implementation model, the system resources are determined by considering scale Metrics defined in the Reference Case, e.g., the number of sensors, etc.
    - **System cost**
      The system includes hardware and software costs, power costs, location/facility costs, and labor costs required for the system operation. In the case of today's Cloud solutions, a subscription fee that includes most of them is charged. In the case of today's on-premise model, the labor cost is differs based on the quality of the system design. We have to consider all of these and conduct analysis in detail to estimate the cost; however, for simplicity, we make the following assumptions in this technical paper:
        - Hardware used for the system sizing estimate is standardized, i.e., we will define a set of the standard shapes as described in A.3.3
        - The hardware depreciation period is set to 5 years to calculate the hardware cost
        - Software annual cost (for depreciation and supports) is twice that of the underlying hardware
        - Every 100 physical or virtual server instances will incur a monthly labor cost of $10,000
        - These numbers are defined to be well-aligned with cloud costs. For details, please see A.3.3
    - **Energy consumption**
      Energy consumption is the amount of energy consumed by the system or the system element.
      It is noted that energy consumption is a part of the system cost, but it has independent importance as the sustainability debate progresses. Therefore, we will analyze it in parallel with the system resources analysis.

A.3.3 Structure of Evaluation

The implementation model consists of various data processes and data flows, as described in the next chapter for the Reference Implementation Model. Therefore, it is not possible to immediately analyze the Metrics of the entire system. The IOWN GF has decided to take a solid approach to this challenge. It means the entire system must be broken down into the elements, i.e., the system nodes and the networks that connect the system nodes. Each metric is analyzed for each element and summed up to evaluate the entire system as a whole.

Figure A-1 and Table A-1 show such an evaluation framework to assemble various Metrics.

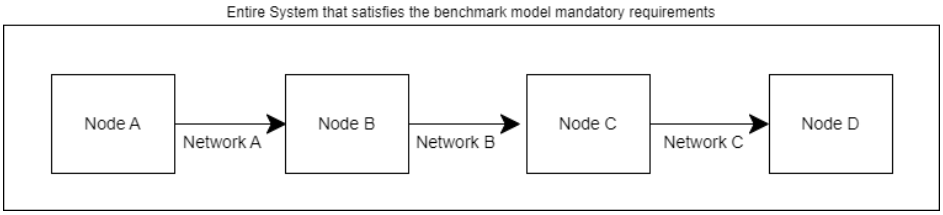

Figure A-1: Implementation Model Breakdown Image

Table A-1: An Example of Implementation Model Evaluation

	Response Time	**System Resources**	**System Cost**	**Energy Consumption**
Node A	20 milliseconds	CPU Core x 36 RAM 512 [GB], etc.	$ 500	1,000 W
Network A	Ten milliseconds	Switch x 4, Etc.	$ 100	50 W
Node B	15 milliseconds	CPU Core x 6 GPU Card x 5 RAM 256 [GB], etc.	$ 1,000	3,000 W
Network B	Five milliseconds	Switch x 2, Etc.	$ 50	25 W
Node C	Five milliseconds	CPU Core x 24 RAM 1024 [GB], etc.	$ 400	500 W
Network C	Eight milliseconds	Switch x 4, Etc.	$ 50	25 W
Node D	Two milliseconds	CPU Core x 48 RAM 512 [GB], etc.	$ 400	1,200 W
Entire System	70 milliseconds	-	$ 2,500	5,800 W

A.3.4 Comparative Evaluation

In the process of technological development, we will find multiple alternatives in determining the implementation model. Also, the current implementation model needs to be compared against them to clarify the benefits of the new IOWN GF solution. It means that we have to evaluate each option one-by-one based on the above framework, and at the end, we will determine the best implementation model.

Table A-2 below shows such a comparison.

Table A-2: An Example of Implementation Model Comparison

	Response Time	**System Cost**	**Energy Consumption**	**Comment**

Current Implementation Model	150 msec	$ 10,000	25,000 W	Mandatory requirement of the response time (100 msec) is not met
Implementation Model - Option A	70 msec	$ 2,500	5,800 W	
Implementation Model - Option B	60 msec	$ 3,000	10,000 W	
Implementation Model - Option C	75 msec	$ 2,200	4,000 W	The best option, as achieving minimum cost and energy consumption while satisfying the mandatory requirements (response time is less than 100msec)

Annex B Dataflow and Workload Profiling Framework

This Annex describes a dataflow and workload profiling framework used in section 3. IOWN GF developed this framework to identify service gaps/requirements of use cases accurately and efficiently.

B.1 Framework Overview

This framework consists of the following two steps:

- Step 1: Develop a data pipeline diagram that depicts the use case with the end-to-end flow of data and processes. A data pipeline diagram comprises functional nodes, processes, dataflows, and database/storage elements as illustrated in B.2.
- Step 2: Develop a profile for each of the functional nodes, processes, and dataflows of the developed data pipeline diagram.

B.2 Data Pipeline Diagram

The Data Pipeline Diagram (DPD) is an extended form of the Data-Flow Diagram (DFD) [DFD]. We have developed DPD with some customizations on DFD to profile dataflows and workload for end-to-end systems that span across data centers, networks, and customer premises. We have introduced the concept of a functional node to represent a group of processes as one node and visualize an end-to-end system in a simpler way. Besides simplification, clarifying how many functional nodes should be accommodated and how distributed they are will help us identify networking challenges for distributed systems.

B.2.1 Elements

A data pipeline diagram is composed of the following elements:

- **Functional Node**: A logical node that executes a set of processes. It is a logical node because multiple functional nodes may be deployed to the same physical node. As to the place of physical deployment, there may be multiple options. For example, some functional nodes may be deployed to either customer premises or edge data centers.
- **Dataflow:** Flow of data transferred from one element to another.

Note: This DPD does not distinguish between “pull” type dataflows and “push” type dataflows.

- **Process**: A set of autonomous operations including getting input data, generating output data, and sending out output data. We describe a process with a set of micro-processes to clarify its scope. We also describe conditions that may affect the behaviors of outgoing dataflows from the process (e.g., conditional branch, relocation of destinations, etc.).
- **Database/Storage**: Database or storage. This element is typically used for realizing asynchronous interaction among processes or consolidating multiple dataflows.

Note: This data pipeline diagram does not distinguish types of databases/storages such as RDBMS, object storage, and so on.

B.2.2 Legends

As illustrated in , elements should be represented with the following notation rules:

- A functional node should be drawn with a rectangle showing its identifier and/or its name in the middle.

Note: You can introduce multiple instances derived from one functional node so that you can describe the specific characteristics of the functional node. The instances, when shown, should be labeled at the top of the rectangle with <>. Examples for the labels are <User 1> and <User 2>, which stand for the personas associated with the instances.

- A dataflow should be drawn with a one-directional arrow showing the direction of flow. Each arrow should be labeled with an identifier for reference.

Note 1: You may assign different identifiers to the start points and endpoints of an arrow if it is necessary to distinguish between them.

Note 2: A functional node may include one or more processes. If you want to clearly show the relation between a dataflow and one of the processes, you can connect the arrow of the dataflow from/to the process. Otherwise, for simplicity, you may connect it just from/to its functional node.

- A process should be drawn with a rounded rectangle (or oval) showing its identifier and/or name in the middle.
- A database/storage should be drawn with a drum showing its identifier and/or name in the middle.

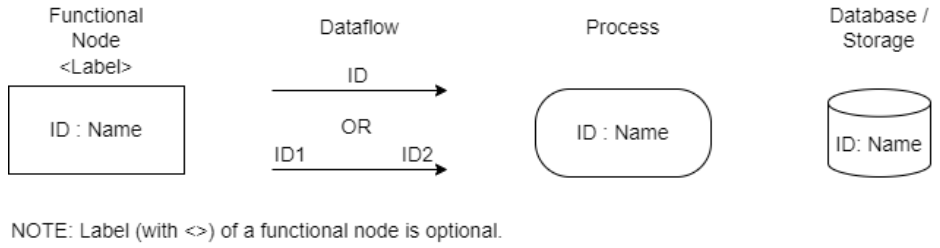

Figure A-2: Elements and Legends of Data Pipeline Diagram

B.2.3 Diagram Example

Shown below is an example of a data pipeline diagram. This example shows an e-mail service system supporting very large attachments.

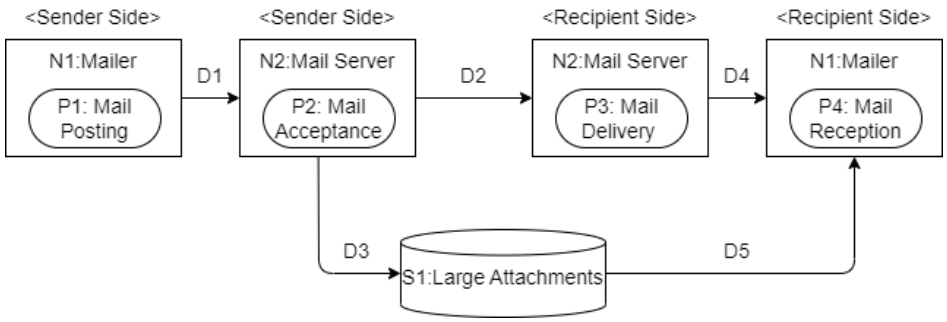

Figure A-3: An Example of Data Pipeline Diagram (Big Mail System)

B.3 Profiling

B.3.1 Profiling Functional Nodes

The following attributes should be clarified for each functional node. Table A-3 is an example of functional node profiles.

- Fixed / Mobile / Semi-Fixed
- Typical places of deployment, e.g., customer premises, regional edge data centers, and centralized cloud
- Total number of nodes

Table A-3: Example of Functional Node Profiles

ID	Name	Description	Attributes

N1	Mailer	Sending/Receiving e-mails	Fixed/Mobile Place: Customer Premises #: billions
N2	Mail Server	Accepting e-mails and delivering them to their recipients.	Fixed/Semi-Fixed Place: Customer Premises/ Carrier's Local Premises #: FFS

B.3.2 Profiling Processes

The following attributes should be clarified for each process. Table A-4 shown below is an example of process profiles.

- The node where the process is implemented
- Micro-processes
- The volume of computations, e.g., OPS (Operations Per Second)
- Energy consumption
- Inherent latency

Table A-4: Example of Process Profiles

ID	Name	Description	Attributes
P1	Mail Posting	Generate MIME Data Send	OPS: FFS Power Consumption: FFS Inherent Latency: FFS
P2	Mail Acceptance	Extract MIME parts Detach large attachments and store them into the database/storage named "Large Attachments" if their total size exceeds 10 MB (dataflow D3) Regenerate MIME data, adding attachments containing the URLs of the original attachments Send the regenerated MIME data to the recipient-side mail server	OPS: FFS Power Consumption: FFS Inherent Latency: FFS
P3	Mail Delivery	Receive the MIME data from the sender-side mail server and deliver it to the recipient-side Mailer	OPS: FFS Power Consumption: FFS Inherent Latency: FFS
P4	Mail Reception	Receive the MIME data Extract MIME parts Fetch the original attachment from the large object storage	OPS: FFS Power Consumption: FFS Inherent Latency: FFS

B.3.3 Profiling Dataflows

The following attributes should be clarified for each Dataflow. Table A-5 shown below is an example of dataflow profiles.

- Data size
- Occurrence rate
- Other requirements: security measures, QoS requirements, etc.

Table A-5: Example of Dataflow Profiles

ID	Description	Attributes
D1	E-mails sent to sender-side Mail Server	Occurrence Rate: varying, up to 10 e-mails per minute Data Size: up to 10 [GB] Other Requirements: mutual authentication, encryption
D2	E-mails sent to recipient-side Mail Server	Occurrence Rate: varying, around 10,000 e-mails per minute Data Size: up to 10 [MB] Other Requirements: encryption, anti domain spoofing
D3	Detached large attachments	Occurrence Rate: varying, around 50,000 puts per minute Data Size: up to 10 [GB] Other Requirements: mutual authentication, encryption
D4	E-mails sent to recipient-side Mailer	Occurrence Rate: varying, up to 10 e-mails per minute Data Size: up to 10 [MB] Other Requirements: mutual authentication, encryption
D5	Large attachments fetched by recipient-side Mailer	Occurrence Rate: varying, up to 50 gets per minute Data Size: up to 10 [GB] Other Requirements: mutual authentication, encryption

Annex C Detailed DPD for the Interactive Live Music Use Case

C.1 Data Pipeline Diagram

Figure A-4 is a DPD used for processing and dataflow analysis on the Interactive Live Music Use Case.

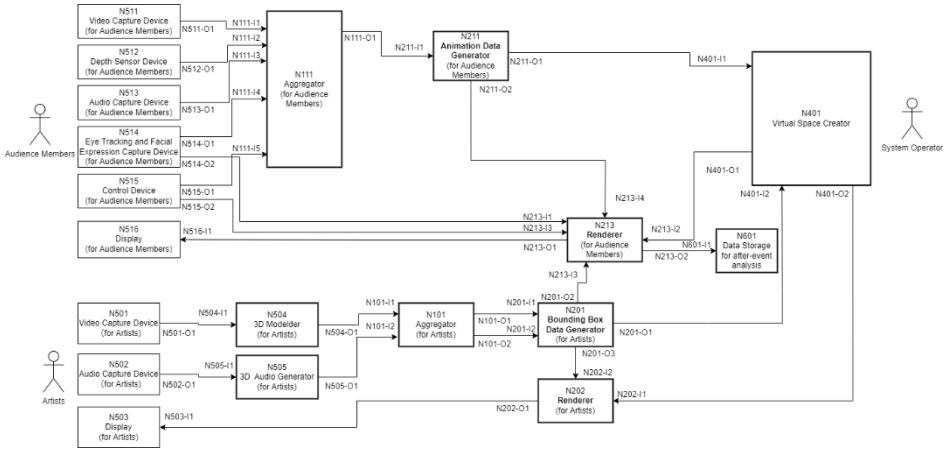

Figure A-4: A Data Pipeline Diagram for the Interactive Music Use Case (same as Figure 3.1-1)

C.2 Functional Node Profiles

The profiles of the functional nodes described in Figure A-4 are shown in Table A-6.

Table A-6 is the description and attribute of each functional node.

Table A-6: Functional Node Profiles

ID	Name	Description	Attributes
N101, N111	Aggregation (ID N1xx)	A node that collects video and audio data and send data to the Ingestion node.	Fixed Network
N201, N202, N211, N213	Bounding Box Data Generator, Renderer Animation Data Generator, Renderer (ID N2xx)	A node receives that receives data from Aggregation (N101) and generates the Bounding Box Data and send to the Renderer for Artists (N202). A node that receives data from Aggregation (N111) and generate Animation Data and send to the Renderer for Audience Members. N213: Receives Scene Composition data from Virtual Space Creator (N401) and Eye Tracking data (N514) and Control data (N515) and send data to the Display (N516).	Fixed Network
N601	Data Storage for after-event analysis (ID 601)	A node that stores the rendered video for after-event analysis.	Fixed Network
N401	Virtual Space Creator (ID N401)	A node that provides “Interactive Live Music” services. It is controlled by System Operator.	Fixed Network

N501, N502, N504, N505	External System (Capturing Device) (ID N50x)	Nodes that capture Artists as volumetric video and 3D audio.	Fixed Network
N511, N512, N513, N514, **N515**	External System (Capturing Device) (ID51x)	Nodes that capture Audience Member's body movement and facial expression with audio. Audience Member is captured using RGB camera, Depth sensor and IR Images (Facial Expression) and sends to Aggregator (N111). Audience Member's audio is captured using microphones and the multi-channel audio and sends to Aggregation (N111). Audience Member's Eye Tracking and Control data is received and sends to Aggregator (N111).	Fixed Network
N503, **N516**	Presentation Device (ID N5xx)	A node that receives video and audio streams from Renderer (N202) and display it. A node that receives video and audio streams from Renderer (N213) and display it.	Fixed Network

C.3 Process and Dataflow Profiles of Each Functional Node

The following subsections describe the internal process of the functional nodes in Figure A-4, the dataflows between them, and their profiles.

An appropriate communication scheme and compression scheme should be selected for each communication section according to the dataflow requirements. This selection will be discussed in section 5. Examples of communication schemes and compression schemes are as follows:

- Communication scheme
  - Shared memory, DMA, RDMA, UDS, RTP over UDP or TCP, etc
- Compression scheme
  - Video Image: H.264, H.265, H.266, Uncompressed etc.
  - Display: DSC (Display Stream Compression) etc. *1
  - 3D model: Geometry/Texture Compression, Point Cloud Compression (G-PCC *2, V-PCC *3), Uncompressed etc.

note *1: HMD could make use of foveated rendering, etc at the application layer.

note *2, *3: These technologies for real-time processing are currently under development. If offline processing is possible, it is also possible to apply ZIP compression, etc.

Symbols

The following symbols are defined and used in the dataflow profiles.

- #_of_Audience Member
  - #_of_Audience Members means the number of Audience Members who join the live music event. The value of the #_of_Audiences_Members will be discussed later.
- #_of_Audience_groups
  - #_of_Audience_group means the number of Audience Groups. A group of people in the same Virtual Space is defined as an "Audience Group". See Terms and Definition for the details. The value of the #_of_Audience Members will be discussed later.

C.3.1 N101: Aggregator Node (for Artists)

Description

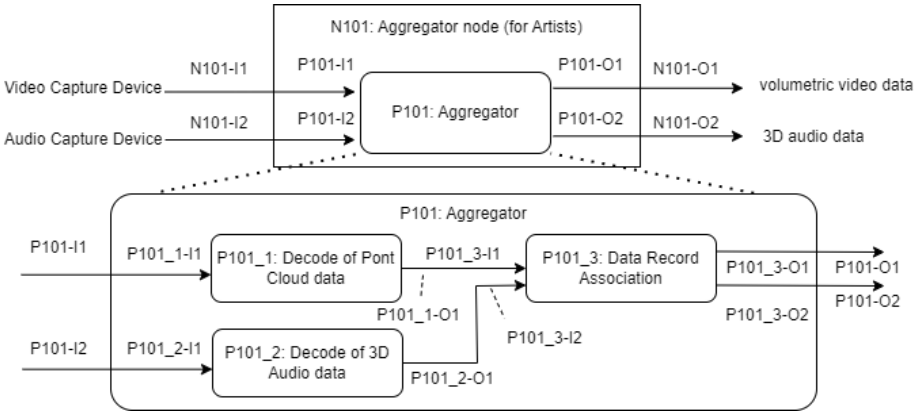

Figure A-5: N101 Aggregator (for Artists)

Process Profiles

The following table shows the detailed description and major attributes of each process.

Table A-7 is the description and attribute of each process.

Table A-7: Process Profiles for N101: Aggregator Node (for Artists)

ID	Name	Description	Attributes
P101_1	Aggregation / Decode of volumetric video data	Receive volumetric video data and metadata from Artist 3D Modeler (Decode the received data into the raw data format)	# of sources: 1 volumetric video data stream Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decoding.
P101_2	Aggregation / Decode of 3D audio data	Receive 3D audio data and metadata from Artist 3D audio Decode the received data into the raw data format	# of sources: 32 ch audio Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decoding.
P101_3	Aggregation / Data record Association	Synchronize and associate the volumetric video data and 3D audio data on a time and space basis. Some metadata may be embedded into each stream. Send the volumetric video and 3D audio data to the Bounding Box Data Generator (N201).	# of sources: camera: 1 volumetric video data stream microphone: 32 ch audio Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, coding.

Dataflow Profiles

The following table shows the detailed description and major attributes of the dataflows.

Table A-8 is the description and attribute of data flow.

Table A-8: Dataflow Profiles for N101: Aggregator Node (for Artists)

ID	Description	Attributes
N101-I1 **P101-I1** **P101_1-I1**	volumetric video at 30/60 [fps] from Artist 3D Modeler	# of sources: 1 volumetric video data stream Compression scheme: uncompressed Data rate: 30 [fps]: 56.35 [Gbps] 60 [fps]: 112.74 [Gbps]
P101_1-O1 **P101_3-I1**	volumetric video data and metadata at 30 [fps] or 60 [fps]	# of sources: 1 volumetric video data stream Compression scheme: uncompressed Data rate: 30 [fps]: 56.35 [Gbps], 60 [fps]: 112.74 [Gbps]
N101-I2 **P101-I2** **P101_2-I1**	3D audio data and metadata from Artist 3D modeler	# of sources: 32 ch 3D audio Data rate: 24.6 [Mbps] (Linear PCM)
P101_2-O1 **P101_3-I2**	3D audio data and metadata	# of sources: 32 ch 3D audio Data rate: 24.6 [Mbps] (Linear PCM)
P101_3-O1 **P101-O1** **N101-O1**	volumetric video data and metadata The size of the metadata is negligible in comparison to the size of volumetric video data.	# of sources: 1 volumetric video data stream Compression scheme: uncompressed Data rate: 30 [fps]: 56.35 [Gbps] 60 [fps]: 112.74 [Gbps]
P101_3-O2 **P101-O2** **N101-O2**	3D audio data and metadata The size of the metadata is negligible in comparison to the size of 3D audio data.	# of sources: 32 ch 3D audio Data rate: 24.6 [Mbps] (Linear PCM)

C.3.2 N201: Bounding Box Data Generator Node (for Artists)

Description

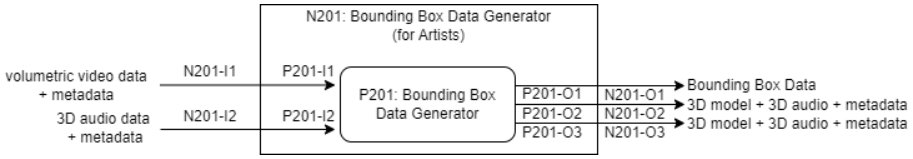

Figure A-6: N201 Bounding Box Data Generator

*note: Sub-Process must be added if the mechanism for synchronizing Video and Audio is to be included here.

Process Profiles

The following table shows the detailed description and major attributes of each process.

Table A-9 is the description and attribute of each process.

Table A-9: Process Profiles for N201: Bounding Box Data Generator Node (for Artists)

ID	Name	Description	Attributes
P201	Bounding Box Data Generator	Receives volumetric video data and metadata from Aggregator (N101). Receives 3D audio data and metadata from Aggregator (N101). Generates Bounding Box Data of the Artist's 3D model. Send Bounding Box Data to Virtual Space Creator (N401). Send the received 3D model, 3D audio and metadata to the Renderers (N213, N202).	# of sources: 2 (1 volumetric video data + metadata stream and 1 3D audio data + metadata stream) Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, encoding, and encryption

Dataflow Profiles

The following table shows the detailed description and major attributes of the dataflows.

Table A-10 is the description and attribute of data flow.

Table A-10: Dataflow Profiles for N201: Bounding Box Data Generator Node (for Artists)

ID	Description	Attributes
N201-I1 **P201-I1**	volumetric video data and metadata	# of sources: 1 volumetric video data stream Compression scheme: uncompressed Data rate: 30 [fps]: 56.35 [Gbps] 60 [fps]: 112.74 [Gbps]
N201-I2 **P201-I2**	3D audio data and metadata	# of sources: 32 ch 3D audio Data rate: 24.6 [Mbps] (Linear PCM)

P201-O1 **N201-O1**	Bounding Box Data	# of sources: 1 Bounding Box Data stream Data rate: 56 [bytes/frame], 30 [fps]:1,680 [bytes], 60 [fps] 3,360 [bytes] Refer to the Annex C.11 for details. [#N201-O1]
P201-O2 **N201-O2** **P201-O3** **N201-O3**	3D model data, 3D audio data and metadata	# of sources: 2 (1 volumetric video data stream and 1 audio stream) 3D model data Data rate: 30 [fps]: 56.35 [Gbps], 60 [fps]: 112.74 [Gbps] 3D audio data Data rate: 24.6 [Mbps] (Linear PCM) [#N201-O2]

C.3.3 N401: Virtual Space Creator Node

Description

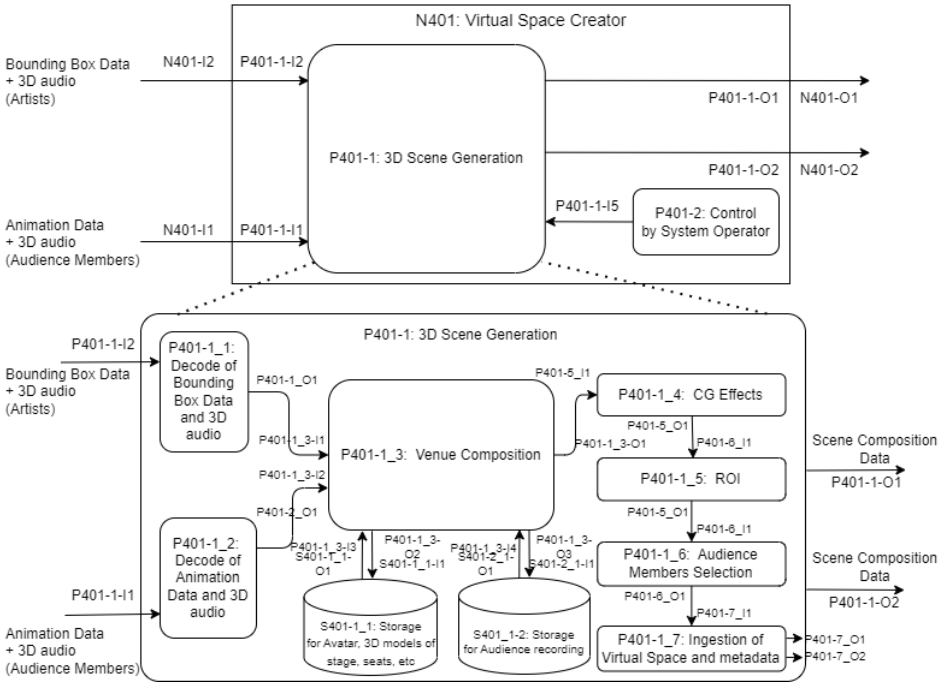

Figure A-7: N401 Virtual Space Creator

Process Profiles

Table A-11 is the description and attribute of each process.

Table A-11: Process Profiles for N401: Virtual Space Creator Node

ID	Name	Description	Attributes
P401-1	3D Scene Generation	Generate 3D Scenes	

P401-2	Control by System Operator	Directs scene composition by System Operator	
P401-1_1	Decode Animation Data and 3D audio data (Audience Member)	Receive Animation Data, 3D audio data and metadata from Animation Data Generator (N211). Decode the received data into the raw data format*1 *1: if the data is uncompressed, this subprocess is not necessary.	# of sources: #_of_Audience_Members Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decoding, generating 3D model data and 3D audio data.
P401-1_2	Decode of Bounding Box Data and 3D audio (Artist)	Receive Bounding Box Data, 3D audio and metadata from Bounding Box Data Generator (for Artist) (N201) Decode the received data into the raw data format*2 *2 if the data is uncompressed, this subprocess is not necessary.	# of sources: 1 (Artists) Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decoding, generating 3D model data and 3D audio data.
P401-1_3	Venue Composition	Create Virtual Space from Bounding Box Data and associated metadata, and Artists' 3D audio and associated metadata Audience Members' Animation Data and associated metadata, Audience Members' 3D audio and associated metadata, and determined position of Artists in Virtual Space Venue composition such as 3D models of stage, seats, light position, speaker position, etc.	# of sources: 1 (Artists) + #_of_Audience_Members Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene, 3D model data and 3D audio data. *note: this process won't process Rendering. Rendering will be processed at Renderer nodes.
S401-1-1	scene data	A database of Parts for composing Audience Members' Avatar data (Selectable faces, clothes, accessories, and so on) Venue composition such as 3D models of stage, seats, light position, speaker position, etc.	# of sources: the number of Avatar models and scene data Occurrence rate: 1 occurrence or on demand. (it depends on Audience Member's and System Operators' control) Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene and 3D model data and 3D audio data.
P401-1_4	CG Effects	Addition of CG effects Information including lighting from AI/System Operator's direction. In N401, Lighting is for Audience Member/Artist/whatever else makes up the Virtual Space. It means adding information such as where on the 3D space the light emitting source is located and how intense it is. This is controlled by System Operator. Based on this information, Renderer will calculate how the Artist or Avatar will be illuminated when the light is emitted and what kind of shadows will be created.	# of sources: 1 (Artists) + #_of_ Audience_Members+ scene data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene and 3D model data and 3D audio data.

P401-1_5	ROI	Addition of ROI (Region of Interest) Information and provide recommended view port for Audience Member, as metadata by the System Operator. This is controlled by System Operator.	# of sources: 1 (Artists) + #_of_ Audience_Members + scene data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene and 3D model data and 3D audio data.
P401-1_6	Audience Member Selection	Selection of Audience Members to be viewed by Artists and addition of ROI Information such as selected Audience Members for Artists, as metadata. This is controlled by System Operator.	# of sources: 1 (Artists) + #_of_Audience_Members+ scene data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene and 3D model data and 3D audio data.
P401-1_7	Ingestion of Scene Composition data	Send Scene Composition data to Renderer nodes	# of sources: 1 (Artists) + #_of_ Audience_Members+ scene data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling.

Dataflow Profiles

Table A-12 is the description and attribute of data flow.

Table A-12: Dataflow Profiles for N401: Virtual Space Creator

ID	Description	Attributes
N401 **P401-1-I1**	Bounding Box Data +3D audio + metadata (for Artist)	Total data rate: 25.0 [Mbps] # of sources: 1 Bounding Box Data Compression scheme: uncompressed Data rate: 30 [fps]: 1,680 [bytes] * 8 * 30 = 403.2 [kbps] + # of sources: 32ch 3D audio data Compression scheme: uncompressed Data rate: 24.6 [Mbps] (Linear PCM) [#N401-I3]

N401-I2 **P401-1-I2**	Animation Data, 3D audio data + metadata (from Audience Members)	Total data rate: 24.7 [Mbps] # of sources: 1 3D model data stream Compression scheme: uncompressed Data rate: Refer to [#N211-O2]; [#Audience_Member_animation_data_size]/person + # of sources: 32ch 3D audio data Compression scheme: uncompressed Data rate: 24.6 [Mbps] (Linear PCM) [#P401-1-I2] *note: Audience_Member_animation_data_size is defined in Annex C.10.
P401-1-I5	Control by System Operator	# of sources: 1 Data rate: 10 KB / control Occurrence rate: 1 control / sec
PP401-1_3-I3 **S401-1_1-O1**	Load Avatar, 3D models of stages and seats etc from Storage	# of sources: # of Audience Members' avatar and scene data Compression scheme: uncompressed Data rate: [#Audience_Member_avatar_data_size] * (#_of_Audience_Members) + [scene data] e.g., 13.54 * 3,000 (1 Audience Group) = 40 [Gbyte]
P401—1_3_O2 **S401-1_1-I1**	Save Avatar, 3D model of stages and seats etc to Storage	# of sources: # of Audience Members' avatar and scene data Compression scheme: uncompressed Data rate: [#Audience_Member_avatar_data_size] * (#_of_Audience_Members) + [scene data]
P401-1_O1 **P401-1_3-I1**	Bounding Box Data, 3D audio and metadata (Artist)	Refer to [#N201-O1]
P401-1_3-I4 **S401-2_1_O1**	Load Animation Data, 3D audio data and metadata (Audience Members)	Refer to [#P401-1-I2]
P401-1_3_O3 **S401-2_1-I1**	Save Animation Data, 3D audio data and metadata (Audience Members)	Refer to [#P401-1-I2]
P401-2_O1 **P401-1_3-I2**	3D model data, 3D audio data and metadata	# of sources: 1 Compression scheme: uncompressed Data rate: 30 [fps]: 56.37 [Gbps], 60 [fps]: 112.76 [Gbps]

P401-3_O1 **P401-4_I1**	CG Effect Artist Bounding Box Data, Audience Member Animation Data, 3D model data (stage, seats etc), 3D audio data and metadata	# of sources: 1 Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene, 3D model data and 3D audio data. Data rate: Refer to [#N201-O1] + [#Audience_Member_animation_data_size]* (#_of_Audience_Members) + [scene data] [#P401-1-I2] *note: The data rate for CG effects is not included because it is very small relative to the other data. *note: Audience_Member_animation_data_size is defined in Annex C.10.
P401-4_O1 **P401-5-I1**	ROI Artist Bounding Box Data, Audience Member Animation Data, 3D model (stage, seats etc), 3D audio data and metadata	# of sources: 1 Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene, 3D model data and 3D audio data. Data rate: Refer to [#N201-O1] + [#Audience_Member_animation_data_size] * (#_of_Audience_Members) + [scene data] *note: Audience_Member_animation_data_size is defined in Annex C.10.
P401-5_O1 **P401-6_I1**	Audience Member Selection Artist Bounding Box Data, Audience Member Animation Data, 3D model data (stage, seats etc), 3D audio data and metadata	# of sources: 1 Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene, 3D model data and 3D audio data. Data rate: Refer to [#N201-O1] + [#Audience_Member_animation_data_size] * (#_of_Audience_Members) + [scene data] *note: Audience_Member_animation_data_size is defined in Annex C.10.
P401-6_O1 **P401-7-I1**	Ingestion of Virtual Space and metadata Artist Bounding Box Data, Audience Member Animation Data, 3D model data(stage, seats etc), 3D audio data and metadata	# of sources: 1 Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, generating metadata for creating scene, 3D Model data and 3D audio data. Data rate: Refer to [#N201-O1] + [#Audience_Member_animation_data_size] * (#_of_Audience_Members) + [Scene Composition data] [#P407-7-I1] *note: Audience_Member_animation_data_size is defined in Annex C.10.

P401-7-O1 **P401-1-O1** **N401-O1** **P401-7-O2** **P401-1-O2** **N401-O2**	Artist Bounding Box Data, Audience Member Animation Data, 3D model data(stage, seats etc), 3D audio data and metadata	Refer to [#P407-7-I1]

C.3.4 N601: Data Storage for after-event analysis

Description

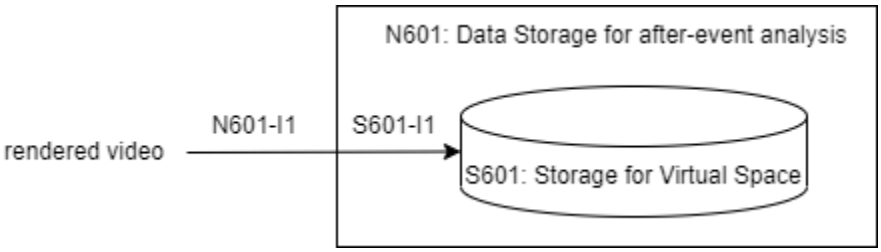

Figure A-8: N601 Data Storage for after-event analysis

Process Profiles

Table A-13 is the description and attribute of each process.

Table A-13: Process Profiles for N601: Data Storage for after-event analysis

ID	Name	Description	Attributes
S601	Storage for Virtual Space (rendered video)	Receives rendered video data from Renderer (N213). Store the received data.	# of sources: 1 rendered video data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decryption and encryption.

Dataflow Profiles

Here, the Rendered Video is recorded, but if you want to save it as Point Cloud, the encoding time is very long with the current technology such as V-PCC, so the issue is whether it is possible to shorten the encoding time with parallel processing.

Table A-14 is the description and attribute of data flow.

Table A-14: Dataflow Profiles for N601: Data Storage for after-event analysis

ID	Description	Attributes

N601-I1 **S601-I1**	Write rendered video data to the Storage	# of sources: 1 rendered video data Data rate: Refer to [#N213-O1] * (#_of_Audience_groups) e.g., 10 Audience Group, 30,000 Audience Members 100Mbps * 10 = 1.0 [Gbps] *note: Store one representative video for each Audience Group.

C.3.5 N202: Renderer for Artists

Description

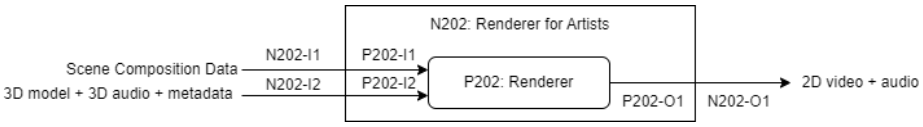

Figure A-9: N202 Render for Artists

Process Profiles

Table A-15 is the description and attribute of each process.

Table A-15: Process Profiles or N202: Renderer for Artists

ID	Name	Description	Attributes
P202	Renderer	Receives Scene Composition data from Virtual Space Creator (N401). Receives Artists' 3D model + 3D audio + metadata from Bounding Box Data Generator for Artists (N201) Generates presentation data by 3D rendering. Sends the presentation data to Display (for Artists) (N503).	# of sources: 2 (1 Scene Composition data from N401, and data from Bounding Box Data Generator (Artist's 3D model + Artist's 3D audio + metadata). Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decryption, rendering and encryption

Dataflow Profiles

Table A-16 is the description and attribute of data flow.

Table A-16: Dataflow Profiles for N202: Renderer for Artists

ID	Description	Attributes
N202-I1 **P202-I1**	Scene Composition data	# of sources: 1 Scene Composition data Data rate: Scene composition data 22.2 [Mbps] (Refer to #N401-O2)

N202-I2 **P202-I2**	3D model data + 3D audio data and metadata	# of sources: 2 (1 volumetric data stream and 1 audio stream) Compression scheme: uncompressed 3D model data Data rate: 30 [fps]: 56.35 [Gbps] (RAW), 60 [fps]: 112.74 [Gbps] 3D audio data Data rate: 24.6 [Mbps] (Linear PCM)
N202-O1 **P202-O1**	Presentation data rendered 2D video data and audio data	# of sources: 1 Data rate: 60 [fps] Video: 3840*2160*60*8*0.5*3/10⁹ = 5.97 [Gbps] (Uncompressed) *1 Audio: Linear PCM 768 [kbps] x 32 [ch] = 24.6 [Mbps] note*1: If the input video is 30 [fps], it must be output after frame conversion to 60 [fps].

C.3.6 N111: Aggregator (for Audience Members)

Description

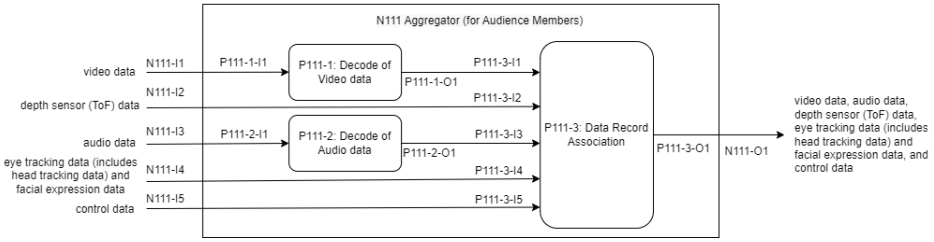

Figure A-10: N111 Aggregator for Audience Members

Process Profiles

Table A-17 is the description and attribute of each process.

Table A-17: Process Profiles

ID	Name	Description	Attributes
P111-1-1	Aggregation / Decode of Video data	Receive video data from the Video capture device (N511). Decoded into uncompressed video data format.	# of sources: 1 video stream Occurrence rate: 30 [fps] Possible tasks that can be offloaded to accelerators: communication protocol handling and decoding.
P111-2	Aggregation / Decode of Audio data	Receive audio data from the Audio Capture Device (N513). Decoded into uncompressed audio data format	# of sources: 1 audio data Occurrence rate: fs=48 [kHz] Possible tasks that can be offloaded to accelerators: communication protocol handling and decoding.

P111-1-3	Aggregation / Data Record Association	Receive uncompressed video data Receive depth sensor data for body movement from the Depth Sensor Device (N512). Receive uncompressed audio data. Receive eye tracking and facial expression data from Eye Tracking and Facial Expression Device (N514). Receive control data (includes head tracking data) from the Control Device (N515). Synchronize and associate the input data on a time and space basis. Some meta-data (e.g., time, position, angle, and owner in a unified format) may be embedded into each stream so that subsequent nodes/processes can efficiently collect and tie relevant records together for their analysis. Send video data, audio data, depth sensor data and eye tracking and facial expression data to the Animation Data Generator (N211).	# of sources: 2; video data and 1 depth sensor data Occurrence rate: 30 [fps] # of sources: 1 audio data Occurrence rate: fs=48 [kHz] # of sources: 2; 1 eye tracking and facial expression data, and 1 control data. Occurrence rate: fs=60, 90, 120 [Hz] Possible tasks that can be offloaded to accelerators: communication protocol handling.

Dataflow Profiles

Table A-18 is the description and attribute of data flow.

Table A-18: Dataflow Profiles

ID	Description	Attributes
N111-I1 **P111-I1**	video data	# of sources: 1 video data stream Compression scheme: compressed or uncompressed Data rate: 30 [fps]: 3.98 [Gbps] (if uncompressed)
N111-I2 **P111-3-I2**	depth sensor (ToF) data for body movement	# of sources: 1 data stream Compression scheme: uncompressed Data rate: 30 [fps]: 0.25 [Gbps]
N111-I3 **P111-1_2-I1**	audio data	# of sources: 1 audio stream Compression scheme: compressed or uncompressed Data rate: 1.5 [Mbps] (if uncompressed)
N111-I4 **P111-3_I4**	eye tracking data facial expression data	# of sources: 1 data stream Compression scheme: uncompressed Data rate: eye tracking 120 [fps] (46 kbps) + facial expression (1.00 [Gbps], 1.99 [Gbps], 2.99 [Gbps]) 30 [fps]: 1.00 [Gbps], 60 [fps]: 1.99 [Gbps], 90 [fps]: 2.99 [Gbps] *note: since eye tracking data is so small compared to facial expression data, the numbers are virtually the same when the two are added together.

P111-1-O1 **P111-3_I1**	video data	# of sources: 1 video data stream Compression scheme: uncompressed Data rate: 30 [fps] 3.98 [Gbps] [#N111-O1]
P111-2-O1 **P111-3-I3**	audio data	# of sources: 1 audio data stream Compression scheme: uncompressed Data rate: 1.5 [Mbps]
N111-I5 **P111-3-I5**	control data (includes head tracking data)	# of sources: 1 data stream Compression scheme: uncompressed Data rate: head motion 23.04 [kbps] + hand-held controller 46.08 [kbps] = 69.12 [kbps]
PP111-3-O1 **N111-O1**	video data, audio data, depth sensor (ToF) data, eye tracking data and facial expression data, and control data	# of sources: 1 data stream Compression scheme: uncompressed Data rate: eye tracking data 120 [fps] (46 [kbps]) + facial expression data (1.00 [Gbps], 1.99 [Gbp], 2.99 [Gbps]) 30 [fps]: 1.00 [Gbps], 60 [fps]: 1.99 [Gbps], 90 [fps]: 2.99 [Gbps] # of sources: 1 data stream Compression scheme: uncompressed Data rate: head motion 23.04 [kbps] + hand-held controller 46.08 [kbps] = 69.12 [kbps]

C.3.7 N211: Animation Data Generator Node

Description

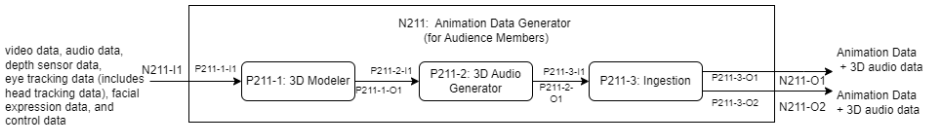

Figure A-11: Animation Data Generator for Audience Members

Process Profiles

Table A-19 is the description and attribute of each process.

Table A-19: Process Profiles

ID	Name	Description	Attributes

P211-1	3D Modeler	Receive video data, audio data, depth sensor (ToF) data, eye tracking and facial expression data, and control data from Aggregator (N111). Generate an Animation Data from the movement of the Audience Member. In detail, this Animation Data is created from video data, depth sensor data, eye tracking data, and facial expression data, and control data. Send the Animation Data, audio data, and control data to the 3D Audio Generator (P211-2). *note: It is assumed that Animation Data including the movement of the mouth and other parts of the face will be generated.	# of sources: 1; video data, audio data, depth sensor data, eye tracking, facial expression data and control data Occurrence rate: 30 [fps] Possible tasks that can be offloaded to accelerators: communication protocol handling, 3D modeling.
P211-2	3D Audio Generator	Receive the Animation Data, audio data, and control data from the 3D Modeler (P211-1) Generate a 3D Audio data from audio data and control data (if any) Send the Animation Data and the 3D audio data to the Ingestion (P211-3)	# of sources: 1 audio data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, 3D audio generation, and encryption
P211-3	Ingestion	Synchronize and associate the Animation Data and the 3D audio data and on a time and space basis. Some metadata may be embedded into each stream. Send data to the Virtual Space Creator (N401) and Renderer (N213)	# of sources: 1 Animation Data, and 3D audio data Occurrence rate: Animation Data: 30 [fps] Audio data: 48 [kHz]

Dataflow Profiles

Table A-20 is the description and attribute of data flow.

Table A-20: Dataflow Profiles

ID	Description	Attributes
N211-I1 **P211-1-I1**	video data, audio data, depth sensor (ToF) data, eye tracking data and facial expression data, and control data	Refer to [#N111-O1]

P211-1-O1 **P211-2-I1**	Animation Data, control data, and audio data	Total data rate: 1.68Mbps (per one Audience Member) # of sources: 1 Animation Data, audio data, and control data Animation Data Compression scheme: uncompressed Data rate: 108 [kbps] audio data Compression scheme: uncompressed Data rate: audio data: 1.5 [Mbps] (2ch) control data Compression scheme: uncompressed Data rate: head motion 23.04 [kbps] + hand-held controller 46.08 [kbps] = 69.12 [kbps] *note: Audience_Member_animation_data_size is defined in Annex C.10.
P211-2-O1 **P211-3-I1**	Audience Member Animation Data and 3D audio data	Total data rate: 24.7 Mbps # of sources: 1 Animation Data and 3D audio data Animation Data Compression scheme: uncompressed Data rate: 108 [kbps] 3D audio data Compression scheme: uncompressed Data rate: 3D audio: 768 [kbps/ch]x 32 [ch max] = 24.6 [Mbps]
P211-3-O1 **N211-O1** **P211-3-O2** **N211-O2**	Animation Data and 3D audio data	Total data rate: 24.7 Mbps [#211-O2] # of sources: 1 Animation Data and 3D audio data Animation Data Compression scheme: uncompressed Data rate: 108 [kbps] 3D audio data Compression scheme: uncompressed Data rate: 3D audio: 768 [kbps/ch]x 32 [ch max] = 24.6 [Mbps]

C.3.8 N213: Renderer for Audience Members Node

Description

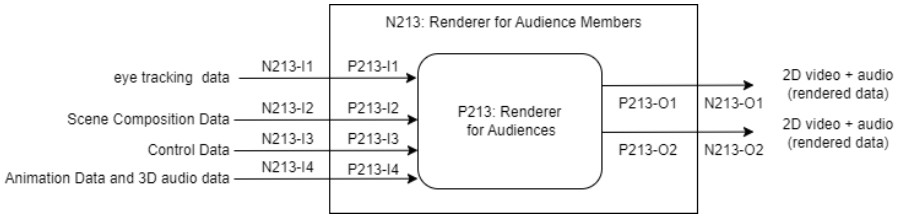

Figure A-12: N213 Renderer for Audience Members:

Process Profiles

Table A-21 is the description and attribute of each process.

The rendered video data is compressed in some way to reduce the data rate to 1/100 to 1/1000 of the original data rate before recording, which in this case is 100 Mbps.

Table A-21: Process Profiles

ID	Name	Description	Attributes
P213	Renderer for Audience Members	Receive Scene Composition data from the Virtual Space Creator (N401). Receive eye tracking data from the Eye Tracking and Facial Expression Capture Device (N514). Receive control data (includes head tracking data) data from the Control Device (N515). Receive Animation Data and 3D audio data from the Animation Data Generator (N211). Receive Audience Member avatar data before the live music event. Generate viewport information from control data (head tracking data and/or eye tracking data (for feveated rendering). Generate presentation data by 3D rendering with the viewport information. Send the presentation data to the Display (N516).	# of sources: 4; Scene Composition data, eye tracking data and Animation Data, and control data Occurrence rate: Continuous streaming processing Possible tasks that can be offloaded to accelerators: communication protocol handling, decryption, rendering and encryption

Dataflow Profiles

Table A-22 is the description and attribute of data flow.

Table A-22: Dataflow Profiles

ID	Description	Attributes

N213-I1 **P213-I1**	eye tracking data	# of sources: 1 data stream Compression scheme: uncompressed Data rate: eye tracking data: 46 [kbps] (120 [fps])
N213-I2 **P213-I2**	Scene Composition data	Refer to [#N401-O1]
N213-I3 **P213-I3**	control data	# of sources: 1 data stream Compression scheme: uncompressed Data rate: head motion 23.04 [kbps] + hand-held controller 46.08 [kbps] = 69.12 [kbps]
N213-I4 **P213-I4**	Animation Data and 3D audio data	Refer to [#211-O2]
N213-O1 **P213-O1**	2D video and audio data (rendered data)	Total data rate (Uncompressed): 8bit YCbCr 420 23.06 [Gbps], 12bit RGB444 69.14 [Gbps] [#N213-O1] Video: 8bit YCbCr 420: 23.04 [Gbps], 12bit RGB 444: 69.12 [Gbps] Audio: Linear PCM 768 [kbps] x 32 [ch] = 24.6 [Mbps]
N213-O2 **P213-O2**	2D video and audio (rendered data)	Date rate 100 Mbps / Audience Group *note: Store one representative video for each Audience Group.

C.3.9 N501: Video Capture Device Node (for Artists)

In a Volumetric capture studio, there are multiple video cameras that are set up to surround Artists. These cameras are synchronized and captured the entire body of Artists. The captured multi-view video data and metadata such as scene description (object size, metric, time, 3D model position, etc.) are sent to the Artist 3D Modeler (N504).

C.3.10 N502: Audio Capture Device Node (for Artists)

In a Volumetric capture studio, multiple microphones are set up. Artists are recorded using these microphones. The captured multi-channel audio data and metadata such as attributes (object position, effect, mixing level between audio, etc.) are sent to Artist 3D Audio Generator (N505).

C.3.11 N503: Display Node (for Artists)

Two 4K displays are used for Artists to monitor Audience Members. 3D Audio is also monitored.

*note: 360-degree sound is assigned to 2 ch headphones.

C.3.12 N504: 3D Modeler Node (for Artists)

Artists are captured using multiple cameras and multi-view video. The output of each camera is processed locally at the shooting site.

C.3.13 N505: 3D Audio Generator Node (for Artists)

Artists are captured using multiple microphones and multi-channel audio. The output of each microphone is processed locally at the shooting site.

C.3.14 N511: Video Capture Device Node (for Audience Members)

The entire body movement is captured with one or multiple cameras and sent to the Aggregator (N111).

C.3.15 N512: Depth Sensor Device Node (for Audience Members)

A ToF (Time of Flight) sensor is used to capture body movement. The ToF data is sent to the Aggregator (N111).

C.3.16 N513: Audio Capture Device Node (for Audience Members)

Audience Members are recorded using multiple microphones, and the multi-channel audio is sent to the Aggregator (N111).

C.3.17 N514: Eye Tracking and Facial Expression Capture Device Node (for Audience Members)

An in-camera device equipped with a Head-Mounted Display (HMD) captures eyes and around-eye expressions. These captured data are sent to the Aggregator (N111). Eye tracking data are sent to the Renderer (N213) for foveated rendering.

C.3.18 N515: Control Device Node (for Audience Members)

One control device is a built-in sensor in an HMD to acquire head motion. Another control device is a hand-held controller to capture hand motion or push buttons operation. These devices generate 6DOF signals, and they are sent to the Aggregator (N111) and the Renderer (N213).

C.3.19 N516: Display Node (for Audience Members)

A 4K@120p display and 360-degree sound system are used for the presentation.

C.4 Date Rate

Capture Data Size

Artists Assumption

For a single artist, it is assumed that the data rate is estimated by showing the entire flat-panel display from head to foot at high resolution. The greater the number of artists, the greater the distance between them and the Audience Member to get them all on the display, so it is reasonable to drop the pixel per degree, ppd.

As for the number, we will use the maximum data rate for the case of one artist and adjust for the case of multiple artists.

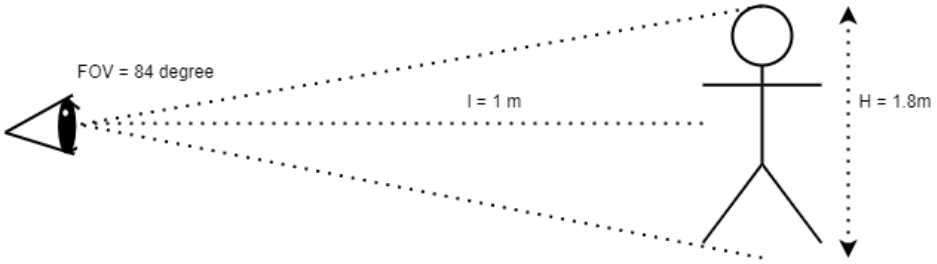

Figure A-13: Assumption of the Artist height and FOV

- Height of a person 180 cm
- Weight of person 80 kg
- Human eye resolution 60 ppd
- Distance from an eye to a person 1 m
- FOV 84 degree
- Frame rate 30, 60 [fps]

- Data size / point 15 bytes
- Point density of a person
  - 60 [pixels/degree] * 84 [degree] / 1.8 [m]= 5040 [pixel] / 1.8 [m] = 2801 [point/m]
  - 2801 [point/m]^2 = 7.85 [mil point/m²]
- Body surface area of a person
  - Du Bois Method is used to calculate body surface from height and weight.
  - BSA = 0.007184 * Height^0.725 * Weight^0.425 =1.996 [m²]
- Total points of a person
  - 7.85 * 1.996 =15.7 [million points]
- Data rate
  - per frame 1,878,232,898 [bits]
  - e.g.; At 30 [fps] 56.35 [Gbps] (This data size is for 1m distance. If the distance is 2m, the data rate would be around 18.8 [Gbps])

*note: We may have to look back on these values to see if it can really be handled in real-time at the PoC timing.

Capturing Audience Members: Assumption

- Resolution: 3,840 x 2,160
- Framerate : 30 [fps]

*note: This frame rate is for transmission only, it could be converted for rendering.

*note: Each of the 3,000 users will create 9 [Gbps] of video data, and 3D Modeler will generate 3D model avatars. Avatar data may be compressed.

Display Data Size

HMD Assumption

- Resolution: 4Kx2K for both eyes
- Frame rate: 120 [fps]
- Bit Depth: 8, 10, 12bit
- Chroma-subsampling: YCbCr 420, 422, RGB 444
- Some compression technology such as DSC and Foveated Rendering may be applied.
- Data rate: 4000*2000*120*8*0.5*3/10⁹ = 11.52 [Gbps] /eye, 23.04 [Gbps] for both eyes (uncompressed)

*note: We assume that a single viewing device can be supported for the ILM UC simultaneously at this moment.

C.5 Virtual Spaces and Audience Members

- Member

This section describes the configuration of Virtual Spaces and the relationship between Virtual Space and Audience Members at geographical area.

- There can be multiple Virtual Spaces.
- Audience Members within the same Virtual Space is defined as an Audience Group.
- Audience Members can interact within the same Audience Group.
- Audience Members can also view the other Virtual Spaces. (Optional)

- Interaction between different Audience Groups is possible.（Optional)

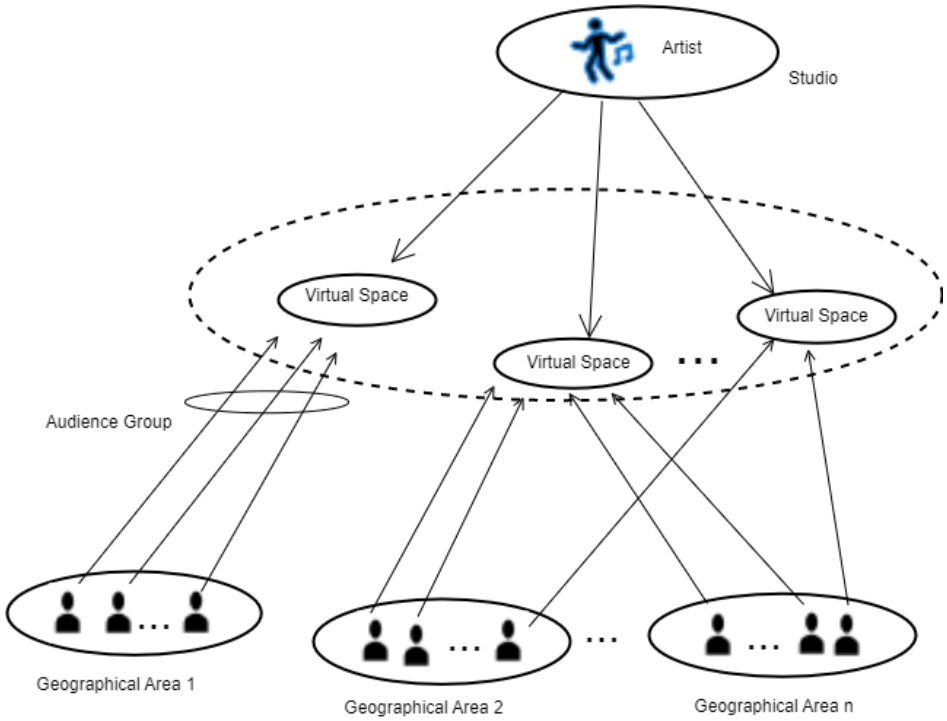

Figure A-14: Virtual Spaces and Audience Member

C.6 Audience Member 3D model data size and Animation Data size

Audience Member 3D model data size

While Artist is represented in high resolution with huge volumetric video data, the Audience Member is designed to be represented with a smaller amount of data.

Assumption

- Polygon: 50,000, Vertex: 30,000, Texture: 2K x 2K
- Parameters: 8; Location (x,y,z), Normal vector(x,y,z), UV Coordinate(u,v) / Vertex
- Data unit: 4 bytes

Vertex data

- 4 [bytes] x 8 [parameters] x 30,000 [Vertex] = 960 [kbytes]

Texture data

- 2048 x 2048 x 24bits (RGB) = 12.58 [Mbytes]

Total

- Mesh + Texture data size : 13.54 [Mbytes]

For example, the data size for 1 Audience Group is;

- Mesh data size 12.58 M[bytes] + Texture data size 960 [kbytes] = 13.54 [Mbytes]
- hence, 13.54 [Mbytes] * 3,000 [person] = 40 [GB].

Audience Member Animation Data size

The Animation Data will be sent to the system in real-time during the live event to move the 3D model. This Animation Data includes not only body movements but also facial expressions and eye movements.

Assumption

- Bone control
- The position and direction of 100 joints change
- Parameters: 9; Location (x,y,z), Rotation(x,y,z), Scale (x,y,z) / Joint
- Data unit: 4 bytes
- Frame rate: 30 [fps]

Animation Data size per frame

- 4 [bytes]×9 [params] x 100 [joints]= 3600 [bytes]

Animation Data size per second

- 3600 [bytes] x 30 [fps] = 108 [kbps]

Total

- Animation Data size: 108 [kbps]

*note: This number is referenced as Audience_Member_animation_data_size in Chapter 3.

C.7 Artist movement

Body Position/Direction and Head Position/Gaze Direction

Body Position and Body Direction

The bounding box for placing Artist on the 3D scene is represented by 8 vertices, of which the normal vector of the first face (V1, V2, V3, (V4)) is defined as the front face. Using Bounding Box Data, a 3D Scene can be created without transmitting the huge 3D Model itself.

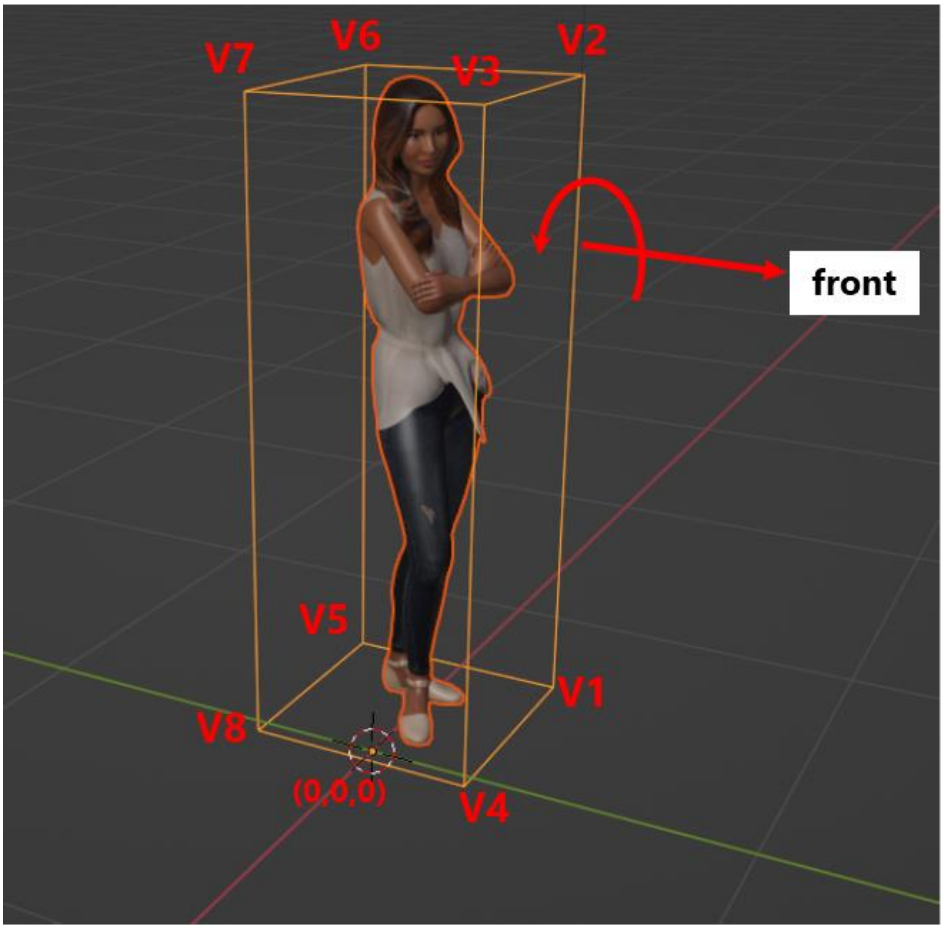

Figure A-15: Bounding Box Data (V1, V2, V3, V4, V5, V6, V7, V8)

Head Position and Gaze Direction

Head position and Gaze direction on the 3D scene is represented by 6 parameters. The center of head position and gaze direction. Head position is simply position, while gaze direction is represented by yaw, pitch, and roll.

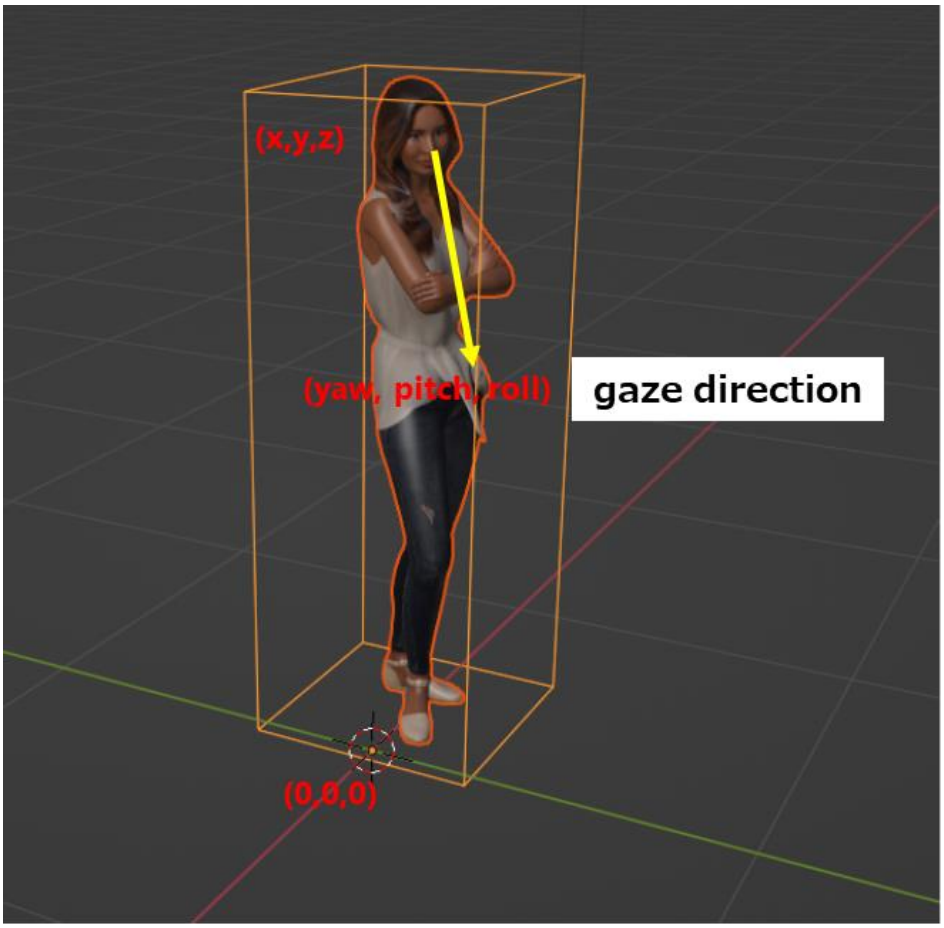

Figure A-16: Head Position (x, y, z) and Gaze Direction (yaw, pitch, roll)

Total Data Rate is 8 parameters for Body movement, plus 6 parameters for Head Position and Gaze Direction equals 14 parameters, or 56 bytes per frame, 4 bytes * (8 + 6) = 56 bytes/frame, 30 [fps]: 1,680 bytes, 60 [fps]: 3,360 [bytes].

*note: The model is a free model from Renderpeople.com.

Annex D Latency control (Future Study Item)

There are some latency-sensitive audio applications which can be occurred in the Interactive Live Music use case, such as offsite singing together, new year countdown together, etc.

Therefore, the proposed architecture of the Benchmark Model could include a synchronizing model to deal with the issue mentioned above, and study how to meet the latency-sensitive application requirements.

Taking the “Offsite Singing Together” as an example, to achieve the goal of singing in sync, the end-to-end latency for every participant should be less than 40 [ms]. (refer to the following paper discussed about Haas effect (or called Precedence effect) : Helmut Haas, The Influence of a Single Echo on the Audibility of Speech, 1949 (Haas Effect, Precedence Effect), Journal of the audio engineering society)

The requirements above would be addressed by using IOWN GF technology only in Japan (maximum optical network distance < 1,000 [km] - RTT 10 [ms] and no congestion in the APN). But in the future, when we expand the IOWN GF network to the much wider area (more than nationwide), the above requirements probably cannot be satisfied even if we use IOWN GF technology.

We should take this issue for a future study.

Annex E Inter-DC Multicast over APN

In the ILM RIM, there are two types of data that need to multicast from Central DC to Edge DCs as follows:

1. Artist's volumetric video and audio data
2. Scene Composition Data

To multicast them from Central DC to Edge DCs, there are three potential approaches as follows:

1. PtMP connection with passive optical splitter
2. PtMP connection by APN-G/APN-I
3. Forwarding to adjacent Edge DCs by DCI GW

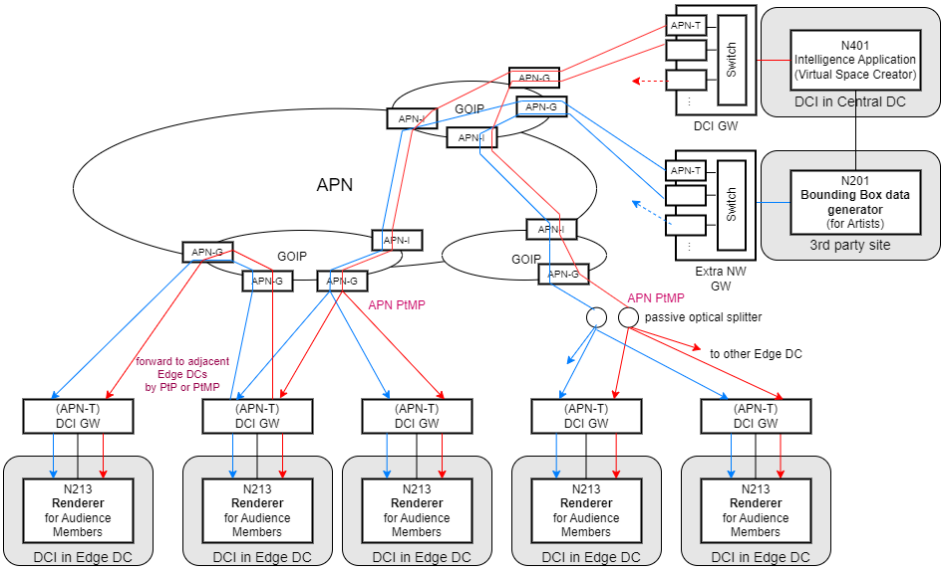

Figure A-17: Some potential approaches for multicast over APN

Option #1 is to establish p-to-mp connection under APN-G by using passive optical splitter. There are several approaches such as TDM-based and WDM-based, but there are the following limitations for now:

- bandwidth: less than 50 [Gbps]
- distance between APN-G and APN-T: less than 20 [km] in the case of TDM approach

Option #2 is the approach that APN-G establish p-to-mp connections, but this is not well studied and such multicast method over long-distance network is a further study item for Technology WG. The outcome will be reflected to the ILM-RIM when it is ready.

Option #3 is the approach to connect Central DC to Edge DCs by p-to-p APN connections and transfer the data hop-by-hop via DCI GW. The APN Controller establishes p-to-p connections from Central DC to regional Edge DCs, and establish p-to-p connections from the regional Edge DC to adjacent local Edge DCs. By using this optical paths, Artist's video data and Scene Composition Data are transferred.

As the first version of RIM, we take the Option 3. Option 1 could be used additionally if the limitations are acceptable.

Annex F Presentation video transfer protocols

This section describes some options to transfer the presentation video stream that is generated by Renderer node to the customer premise to display this on the screen of display devices like HMD.

Assumption

- HMD has a DisplayPort (DP) or HDMI interface (tethered type).
- Presentation video data is rendered by Renderer node at an edge data center using GPU according to the head/eye-position. As the result, the presentation video data is generated in the GPU's memory.
- In this page, data flow of control data (head/eye-position) from customer premise to DCI is not included.

Approaches

Option 1: RDMA

This option uses RDMA protocol to transfer the presentation data through APN. Figure A-18 shows an example of system structure of GPU direct RDMA. The presentation data in GPU memory of LSN is transferred to smartNIC by DMA and it is transferred to customer premise by RDMA protocol through APN. SmartNIC at customer premise receives the presentation data and it transfers them to GPU memory of Local Node via DMA. Then the presentation data is out to HMD through DP/HDMI interface.

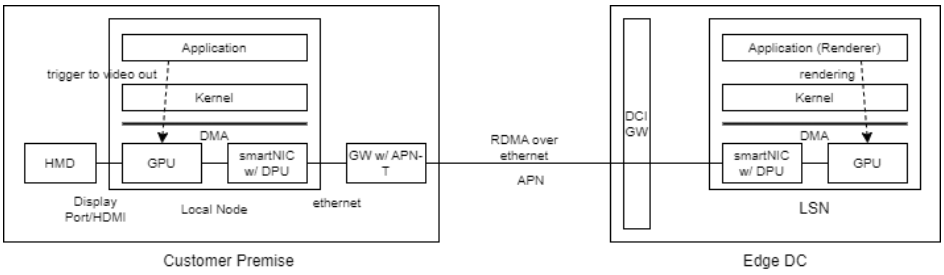

Figure A-18: An example of system structure of RDMA (RoCE v2)

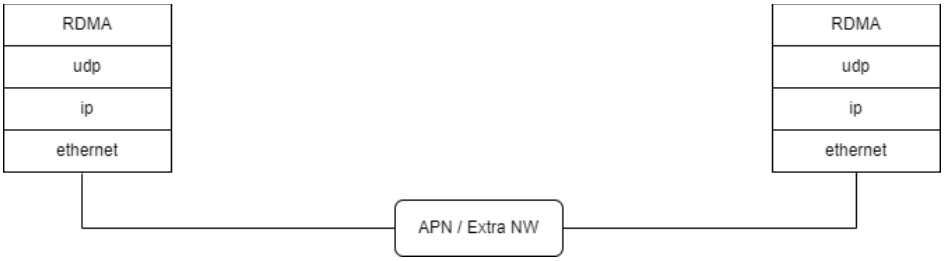

Figure A-19: Protocol Stack for RDMA (RoCE v2)

Option 2: SDI over IP (ST2202)

This option uses SDI over IP protocol to transfer the presentation data through APN. Figure A-20 shows an example of system structure of SDI over IP. The presentation data in GPU memory of LSN is transferred to smartNIC by DMA and it converts the data to ST2202 packets and transfer them to customer premise through APN. SmartNIC of Local Node at customer premise receives the ST2202 packets and it decodes them and extract the presentation data. Then it transfers them to GPU memory and it is out to HMD through DP/HDMI interface.

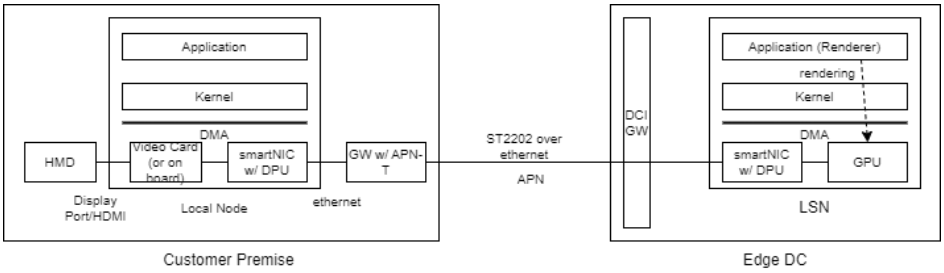

Figure A-20: An example of system structure of SDI over IP (ST2202)

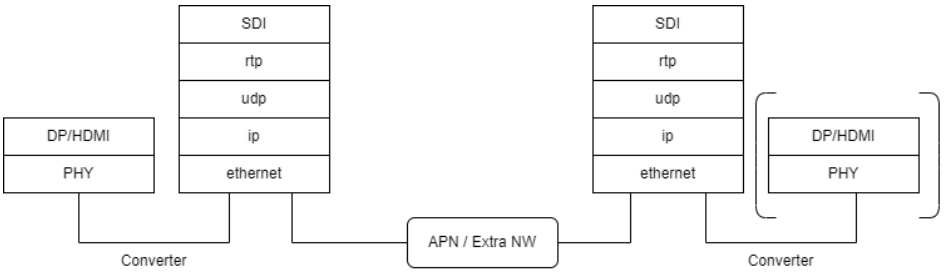

Figure A-21: Protocol Stack of SDI over IP (ST2202)

Option 3: SDI over IP (ST2110)

This option uses SDI over IP protocol to transfer the presentation data through APN. Figure A-22 shows an example of system structure of SDI over IP. The presentation data in GPU memory of LSN is transferred to smartNIC by DMA and it converts the data to ST2110-20 packets and transfer them to customer premise through APN. SmartNIC of Local Node at customer premise receives the ST2110-20 packets and it decodes them and extract the presentation data. Then it transfers them to GPU memory and it is out to HMD through DP/HDMI interface.

Note: SMPTE ST2110-20 or RFC 4175 is used for the video transmission over rtp.

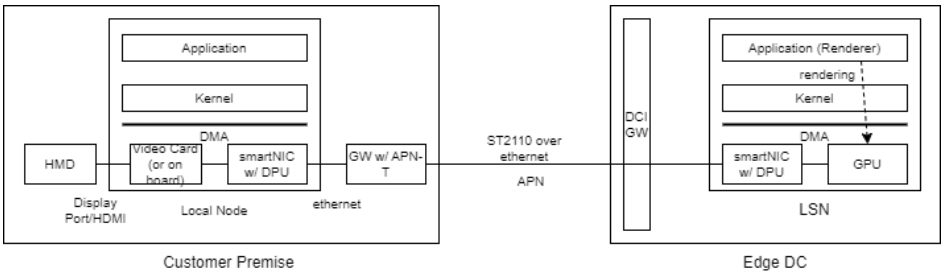

Figure A-22: An example of system structure of SDI over IP (ST2110)

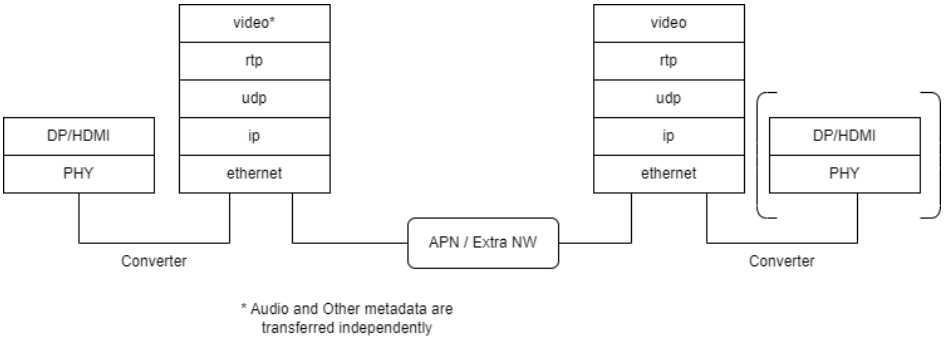

Figure A-23: Protocol Stack of SDI over IP (ST2110)

Option 4: Proprietary protocol

This option uses proprietary protocol to transfer video signals through APN. Figure A-24 shows an example of the system structure. Video signals are out from GPU of LSN and they are converted to proprietary forms at SmartNIC and they are transferred to customer premise via APN. They are converted to video signals at Local Node of customer premise and they are out to HMD through DisplayPort/HDMI interface.

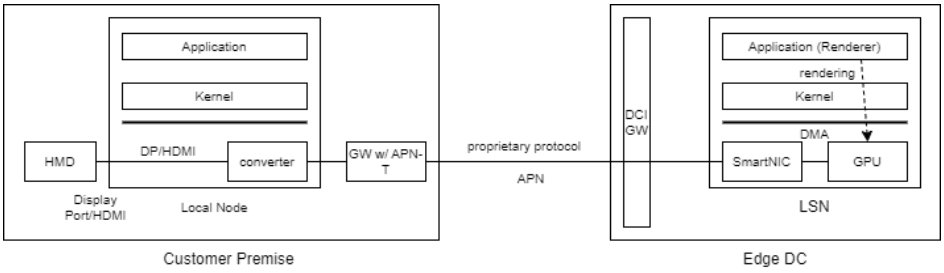

Figure A-24: An example of system structure of DisplayPort/HDMI over APN

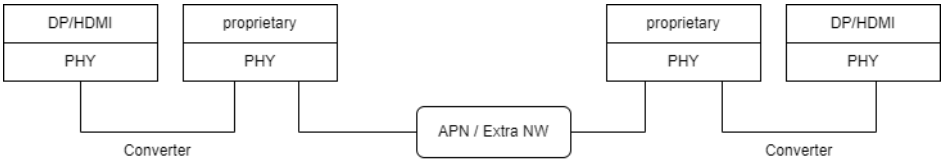

Figure A-25: Example of Protocol Stack of DisplayPort/HDMI over APN

Some Points that need to be considered further

- Possibility of image processing (e.g. Partial Rendering and Decoding) at customer premise
  - Partial Rendering
    - Ideally, Renderer renders the presentation data whose pixel size exactly corresponds to the pixel size of a display device according to the user's head/eye position. But in order to provide a better user experience, some additional processes might be needed in customer premise. For example, Renderer produces the redundant presentation data and Local Node such as PC might crop the data to fit the display screen according to the head/eye-position if the head/eye motion is very fast and it causes the gap between the head/eye position and the image on the display. For your information, some motion prediction technologies might be used to address this issue.
  - Decoding

- ✧ At this moment, we assume that the uncompressed presentation data are transferred from DCI site to customer premise. But if the presentation data needs to be encoded/decoded for some reasons (e.g. limited bandwidth), the decoding process might be needed at Customer Premise.

Annex G Rendering workload reduction mechanisms

As rendering images at 120 [fps] is a heavy workload, it is required to introduce several techniques to lighten the GPU workload, to develop a commercially reasonable interactive live music service. For instance, the following techniques can be considered:

- Eye Tracking and Foveated Rendering
- Level of detail control based on distance

I.1 Eye Tracking and Foveated Rendering

The foveated rendering is a pure rendering technique and it reduces the image quality in the peripheral vision by utilizing eye tracking information. This technique is said to save the GPU workload by up to 50% [Meta; https://www.facebook.com/RealityLabs/videos/488126049869673]. The below figure shows an image of such foveated rendering technique.

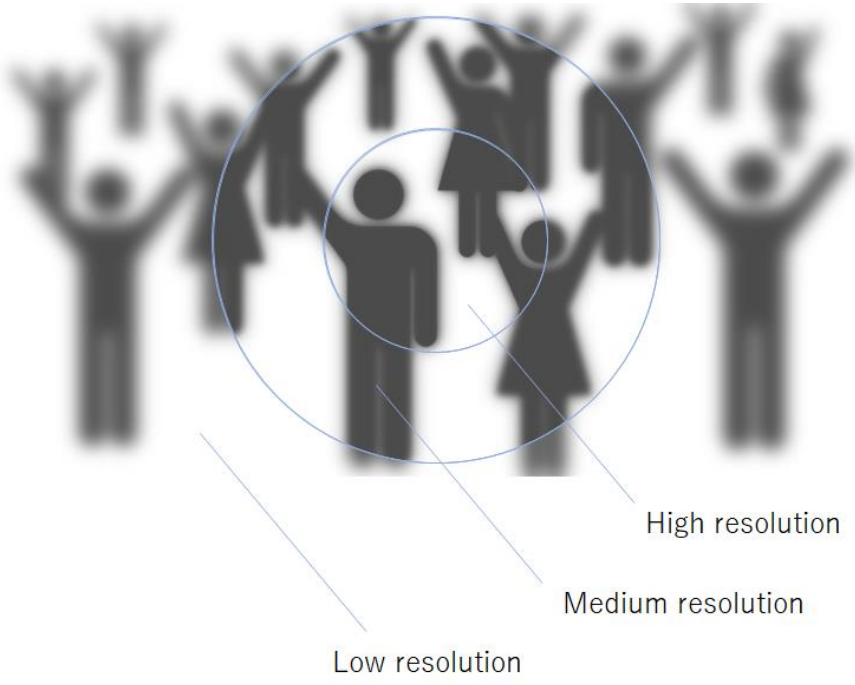

Figure A-26: Foveated rendering technique

I.2 Level of detail control

The level of detail control is a data management technique that contributes to reducing the rendering workload. Because, it is not reasonable to renderer the detailed image for far Audience Members, when rendering an image

capturing hundreds of other Audience Members in the line of sight, as far Audience Members not only appear small but also may be obscured by other viewers or static structures. Therefore, it makes sense to load the animation data onto the GPU after lowering the resolution of the far Audience Members. The below figure shows such an image to control the resolution of the animation data. This technique has the potential to significantly reduce the GPU workload in proportion to the reduction in data size. See Annex K for such a benchmark example that well-explain the effectiveness of this technique.

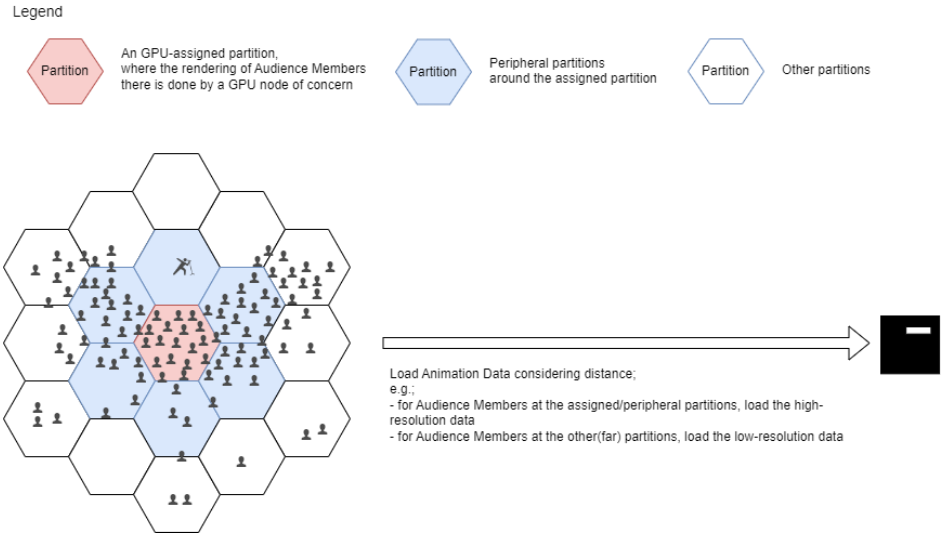

Figure A-27: Level of detail control technique based on virtual space partitioning

Annex H Computing resource estimation for rendering

This section describes the rendering study for motion-to-photon latency.

First we will talk about which Data Pipeline section and how to define the motion-to-photon latency in combination with the device, and then we will give an overview of each of the forms in which we experiment with Audience Member, Artist separately and then add up the results.

What is the motion-to-phone latency here? In the figure below, the communication between the Audience Members and Renderer is the essential part to discuss latency.

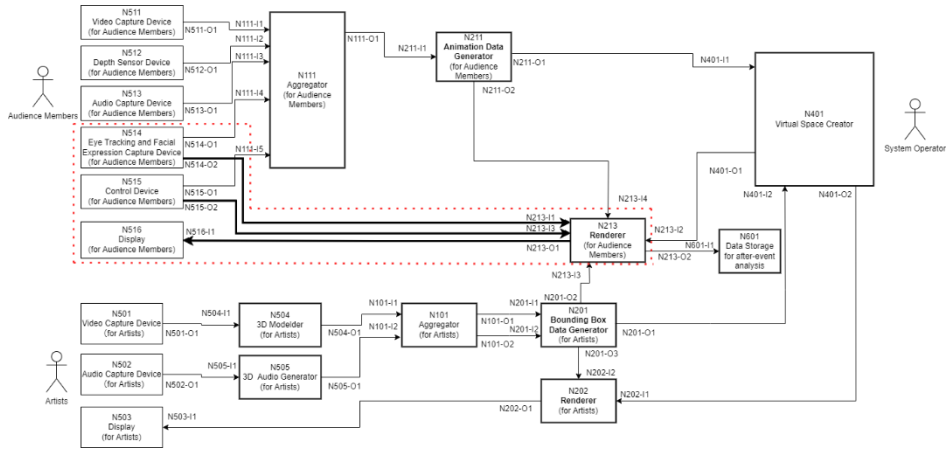

Figure A-28: The motion-to-photon latency in the DPD

How much time can be used in the rendering process to achieve 10 [ms] motion-to-photon latency?

Latency is the difference between action and reaction. Motion-to-photon latency is the amount of time between the user's head movement (action) and its corresponding display output reflections (reaction) on the HMD.

The figure shows the timing of refreshing Display and timing of input of Control data to control the viewpoint. The timing of control data and display are not synchronized.

The sum of the rendering time + the latency to display is the motion-to-photon latency. Transmission time is assumed to be zero here for simplicity. The control data and display are asynchronous, so half of the “waiting time until display” (8.3/2≒4.2 [ms]) is taken as the average value. As a result, the time that can be spent on rendering is 5.8 [ms] (10 - 4.2 = 5.8 [ms]). If we guarantee 10 [ms], the time that can be spent on rendering is only 1.7 [ms]. (10 – 8.3 = 1.7 [ms])

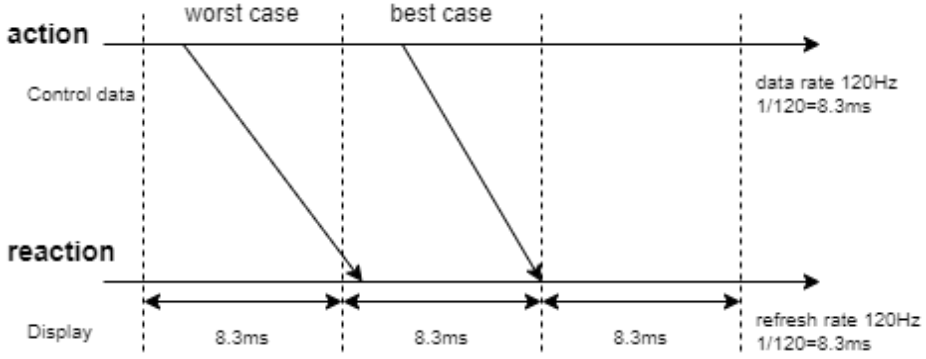

Figure A-29: The worst case and best case for motion to-to-photon latency

5.8 [ms] is the average time available for rendering and other tasks to achieve 10 [ms] motion-to-photon latency. In other words, we need to consider the distance between the renderer node and Audience Member, so we have less time available than this value.

Experiment – Sample Virtual space

This is the experiment result with Unity. It was a simple test result to see how many Audience Members could be supported from a bird's eye view that could see everyone.

Figure A-30：Virtual space example, © Unity Technologies Japan/UCL

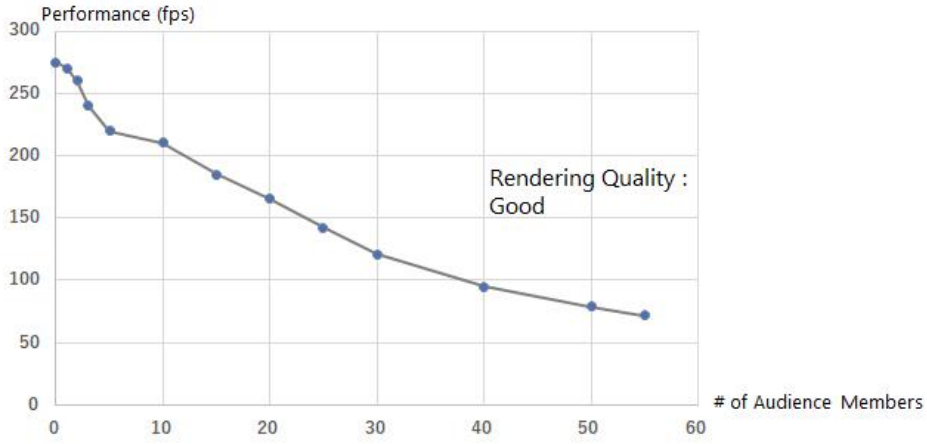

Figure A-31: Number of Audience Members and performance

To meet the 10 [ms] motion-to-photon latency, the rendering process including network transmission must be completed within 5.8 [ms] on a 120 [Hz] display. In other words, it requires 172 [fps] (1/5.8 [ms]) rendering performance at least. Based on our experience, the current 3D Scene can accommodate up to around 15 Audience Members.

However, there are some differences from our assumptions.

- It is not implemented to input Animation Data from the network in real-time.
- It is not implemented to detect collisions between Audience Members.
- The 3D model size, and meshes are much smaller than our assumption, only 1/3 for Audience Member.
- The viewpoint is fixed, and all the Audience Members are displayed at the same time.

This experiment was a modification of the existing Virtual Space's Live demo to get an overview of its rendering capabilities. The next section will take a closer look at each of these aspects.

Experiment – Audience Member side

Quantitative FOV measurements represent the rendering performance when the FOV is varied.

A smaller FOV results in a narrower field of view, but in Unity, the Avatar is supposed to appear larger because it is displayed on full screen. As a result, with 100 Audience Members, FOV110 degree has the highest Vertices and it has the worst performance, only 25 [fps].

The camera is 1.8 [m] high and positioned 3 m from the very back, with a distance from the Audience Member of 1 [m] each front, back, left and right.

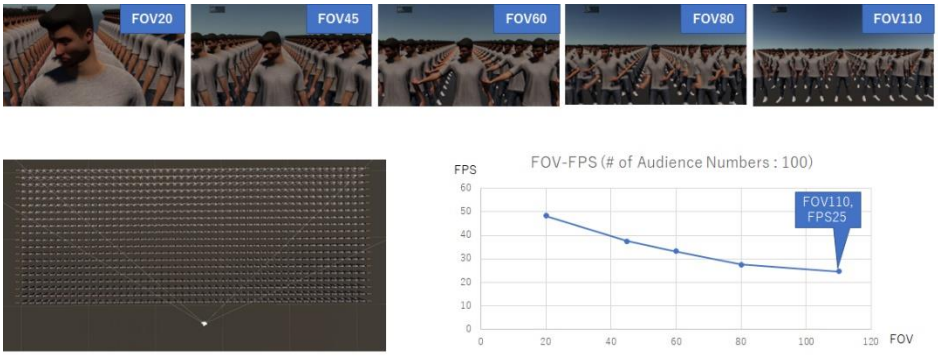

Figure A-32: Performance when FOV is changed

*note: The model is a free model from Renderpeople.com.

Experiment – Artist side

Artist's Point cloud data replayed with 3rd party player in Unity because Unity itself does not support volumetric video yet. As a result, the average frame rate is 65.1 [fps].

- Artist model: Point cloud data, MPEG, “long dress”, The number of points: 0.8M
- 3D scene: Artist only, no stage, no effect, no Audience Member
- Hardware: Gaming Display, Sony INZONE M3, FullHD 240 [Hz]

Estimation

The Audience Member and Artist are processed together, so each result be added together for a total of 12x GPU performance required based on our experiments.

The details are as follows;

- Assumption
  - To achieve motion-to-phone latency 10 [ms] (1/5.8 [ms]=) 172 [fps] required for Audience Member
  - Artist side is required to achieve 30 [fps]
  - GPU is NVIDIA RTX3080Ti
- Audience Member side (Avatar)
  - 100 people, FOV 110 degrees results in 25 [fps], the current Avatar requires (172/25=) 6.9x GPU performance
  - Conversion to Benchmark Model: Considering the current Avatar size 1/3, 3 times the size (6.9x3=) 21x is required
  - Conversion to latest hardware: Assuming NVIDIA RTX4090 can double the performance, 10.5x is required
- Artist side (volumetric video)

    - Full HD, 0.8M Point cloud, 65.1 [fps] result requires 4x performance in 4K conversion, which is equivalent to 1/4 of 16.3 [fps], and to meet the 30 [fps] requirement, (30/16.3=)1.84x GPU performance is required

- Conversion to Benchmark Model: 15.7M/0.8M=19.6x is needed as a point cloud for point cloud, so (1.84 * 19.6=) 36x performance is needed
- Conversion to latest hardware: NVIDIA RTX4090 would double performance, 18x is required
- Concern: the relationship between playback performance and FOV at Point Cloud is untested.

Here, considering the upper limit value from the display resolution of the Benchmark Model, we can reduce the 3D model size.

- 4K display resolution
    - The current target of the Benchmark Model is defined 60 ppd and it requires 15.7 million, and this is determined as an ideal value based on human visual characteristics.
    - But, the maximum number of points per person on 4K Display is 2.47 million*1
    - So (2.47M/0.8M=)3.08x the performance is required.
    - If NVIDIA RTX4090 can achieve twice the performance, 1.5x the performance is required.

  *note 1: Considering the example of displaying the entire body of a 1.8 [m] tall 3D model at the highest pixel at the optimal viewing distance of 1.5H of the height of the 4K display, the required number of points can be obtained by determining the density from the number of point clouds included per degree from the vertical FOV and multiplying it by the surface area of the person determined using by the Du Bolis method.

So, the rough estimation in total is that Audience Member and Artist are processed together, so 10.5x and 1.5x must be added together for a total of 12x. Looking at the results, to give each person a free viewpoint, independent processing is required, so as many GPUs as there are Audience Members will be needed. This would not be desirable in terms of cost and power consumption.

So, we will raise some ideas to reduce the throughput of the GPU.

- It may be possible to use LOD (Level of Detail) method to lower the resolution of distant objects.
- It may be possible to use the foveated rendering method to lower the resolution.
- It would be possible to greatly reduce the processing load by allowing the selection of a representative viewpoint. For example, the view only moves every 10 degrees or so. That way, the amount of processing would not be proportional to the number of people.
- Given the benefits of Virtual Space, if multiple people are limited to standing in front of the Artist, the number of Audience Members will be reduced, and the rendering load will be decreased. The number of views to be generated is drastically reduced and the rendering load is reduced because several people can stand in front of the Artist and view it.

Rendering conditions in detail

- Artist and 3D Scene: A sample from Unity
    - Mesh: 18805
    - Vertex: 17297
- Audience Member: A 3D model from Renderpeople
    - Mesh: 23406
    - Vertex: 11778

- Hardware and software
    - CPU: Intel Core i9-9900K 3.6 [GHz]
    - GPU: NVIDIA GeForce RTX3080Ti
    - Memory: 48 [GB]
- OS: Windows10 Pro
    - Game Engine: Unity

*note: The 3D model for Audience Member is a free model from Renderpeople.com.

Additional Experiment: Scalability

Furthermore, even if the number of people increases, the rendering load per person should not increase much as the number of people increases, since the far side is mostly hidden by others. This is the result of that verification.

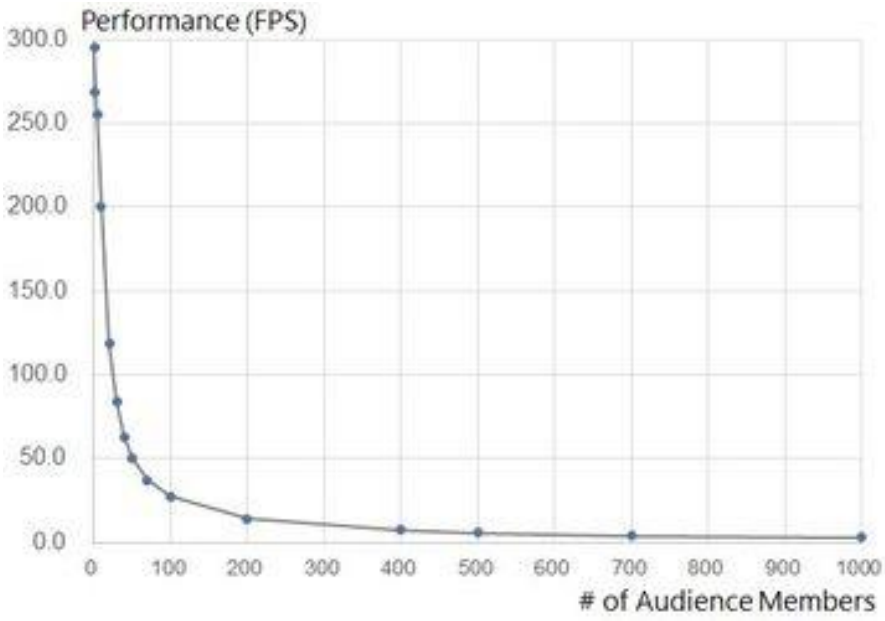

Figure A-33: Number of Audience Members and performance (FPS)

The FPS drops as the number of people increases, but not proportionately. The game engine automatically reduces the processing load to some extent without reducing the number of Vertices by LOD technique, etc. These trends are characteristic of the game engine, Unity. Next, we normalize the validation results of the previous graph by Audience Member to determine how much additional resource each Audience Member requires.

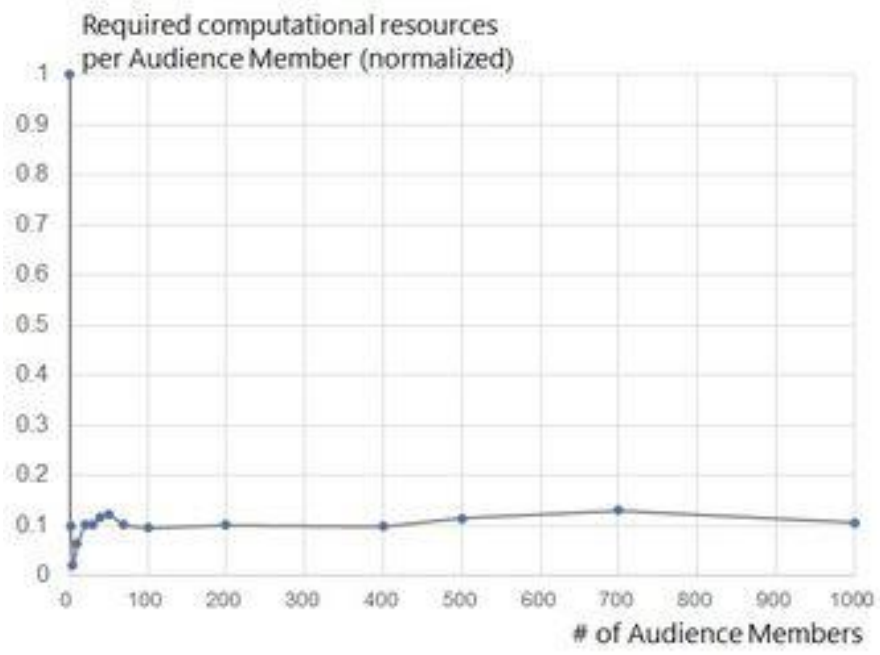

Figure A-34: Required computational resources per Audience Member

We found that the normalized rendering processing load per Audience Member converges to about 0.1 when the number of viewers is increased. In other words, it can be represented by the following approximate workload model,

{effort for rendering} = 0.9 + {number of Audience Members} * 0.1

*note: when the number of people is more than 20.

Thus, we can also estimate the total rendering processing load for 3,000 viewers, for example.

Annex I Validation of LOD technique effectiveness

This is an example that validates the effectiveness of LOD.

As described in Annex H., we made an estimate of the rendering, and the results are very processing-intensive for the current hardware. In this ILM UC, assuming a situation similar to that of an actual live concert venue, we conducted an experiment that took this into account, since people close by would be clearly visible, but people far away would be hidden by others and not actually be visible. Level Of Detail, LOD technique was applied and three levels of the resolution were prepared for the object to be animated. The FOV of the camera is 110 degrees.

Assumptions

- FOV: 110 degrees
- H-LOD: 100%, 0-12 [m]
- M-LOD: 25%(1/4 size of H-LOD),12-48 [m]
- L-LOD: 6% (1/4 size of M-LOD), 48-144 [m]
- Culled: 2% (1/3 size of L-LOD), 144 [m]-

In general, the distance to switch LODs is determined by Unity, the implementation platform, based on the size of each LOD. The right half of the Figure K.1-1 is an example scene where LOD has been introduced. In this sample scene, the distance between the camera and the object is 1 [m], 5 [m], 12 [m], 19 [m], and 51 [m] from the front Audience Member. Objects at the distance of 1 [m] and 5 [m] are selected for H-LOD, objects at the distance of 12 [m] and 19 [m] are selected for M-LOD, and objects at a distance of 51 [m] are selected for L-LOD.

As the number of people increases, the rendering performance drops, but it hardly decreases at about 15 [fps] when LOD is used. The difference in relative values is important and significant. As shown in the graph on the right, the number of Vertices to be rendered almost does not increase in the middle, 500. In LOD, the number of vertices to be rendered almost does not increase because the resolution decreases as Avatar goes farther away and the invisible area in the distance is out of scope. The validation results of this model show that this is a very effective technology for the ILM UC.

Figure A-35: An example of resolutions

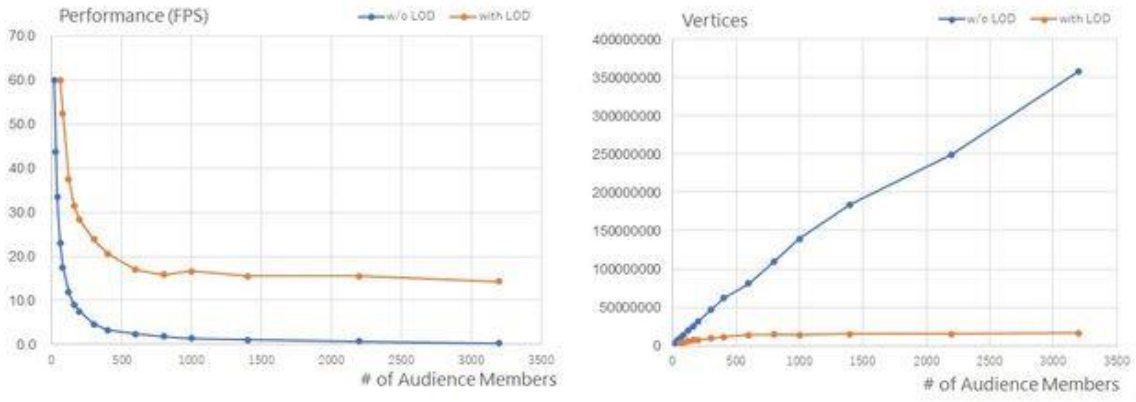

Figure A-36: The Effects of LOD

Annex J Data distribution design for scalability of renderer nodes and Virtual Space Creator nodes

To support an Audience Group with 3,000 Audience Members, the system needs to be designed to scale well. This is especially true for the renderer nodes and the Virtual Space Creator nodes in which large data processing workloads occur.

One idea to acquire good scalability is to divide the Audience Members by their locations in the Virtual Space and assign them to each server based on their locations so that an even number of Audience Members are assigned to each server. To achieve a good data and workload distribution, 1) the Virtual Space has to be partitioned to some degree so that the number of Audience Members per node does not fluctuate largely, and 2) the Audience Member-node assignments need to be completely random, or random-per-zone by distinguished several zones with the expected density levels of Audience Members.

Based on such assignment, the renderer nodes are responsible for rendering images for the Audience Members located in the assigned partitions, and the Virtual Space Creator nodes are responsible for managing animation and position data of the same Audience Members, as well as for detecting a collision between any two Audience Members. It should be noted that the collision may occur around the boundary of two partitions, and to handle such a situation, Audience Member data will be loaded a bit beyond the boundary, or a multi-node query will be performed to detect collision.

The below diagram shows such an image of Audience Member data distribution among the renderer or the Virtual Space Creator nodes.

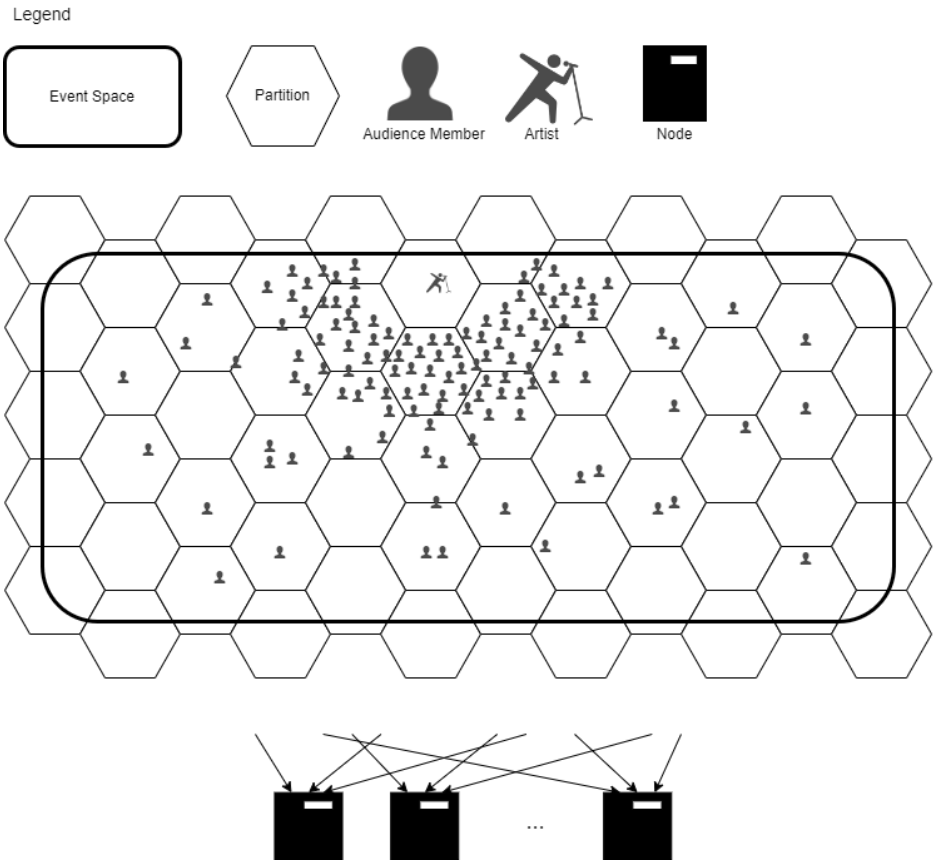

Figure A-37: Audience Member data distribution among the Renderer or the Virtual Space Creator nodes

Annex K System Topology around renderer nodes and Virtual Space Creator nodes

The animation data is managed by the Virtual Space Creator nodes, and then, it is loaded to the renderer nodes which are most possibly equipped with GPU resources. Therefore, it is necessary to consider the topological relationship between the Virtual Space Creator nodes and the renderer nodes.

One possible option is to co-locate the Virtual Space Creator node and the renderer node. In this case, the partition-node assignment logic must be the same for both Virtual Space Creator nodes and renderer nodes, so that the high-resolution data does not need to transfer over the network. As low-resolution animation data is managed by other Virtual Space Creator nodes, communications between these nodes must be taken into account. This option reduces the data transfer burden compared to the other option and may be the only choice for today's implementation model, as node-to-node latency and jitter are not small in today’s cloud and transferring animation data at least 30 times per second for each of thousands of Audience Members causes a large CPU overhead waiting for TCP ACKs.

The other possible option is to decouple the Virtual Space Creator nodes from the renderer nodes and transfer the animation data between them. In this case, different partition-node assignment logics can be used for Virtual Space Creator nodes and renderer nodes. This option allows more flexibility in adjusting resources according to the number of participating Audience Members, their distribution in the Virtual Space, and the rendering workload. For example, it is possible to adjust rendering resolution according to the Audience Member's device, utilize various combinations of different GPU cards, and partition resources for events with a small number of participants. However, to do so, it is necessary to connect the renderer nodes and Virtual Space Creator nodes at high speed because the animation data needs to be populated from all Virtual Space Creator nodes to every single renderer node. Therefore, the application of IOWN technology will be a mandatory requirement.

The below figure shows an image of possible two topology options.

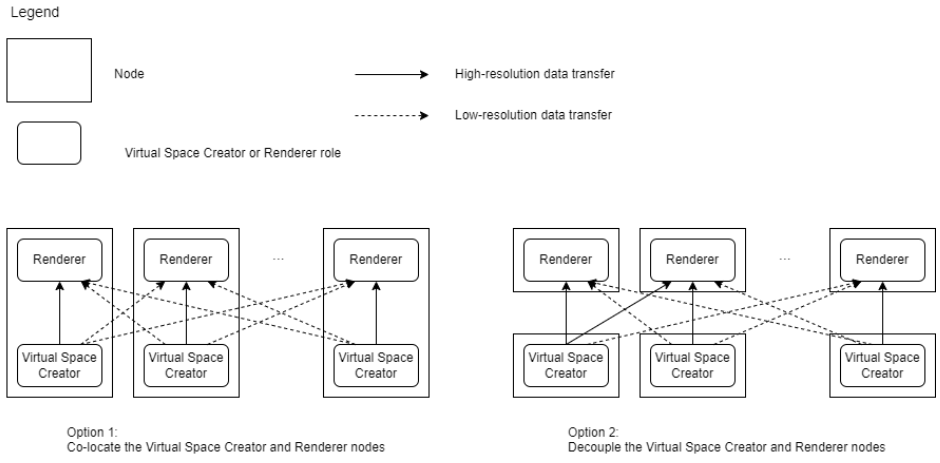

Figure A-38: System topology options around the Virtual Space Creator and the renderer nodes

Acknowledgments

This Reference Document was jointly prepared by Technical and Use Case working group on the IOWN GF under the direction of Masahisa Kawashima (Technical WG Chair) and Katsutoshi Ito (Use Case WG Chair).

History

Revision	Release Date	Summary of Changes
1.0	April 11, 2023	Initial Release