从多媒体通信视角看元宇宙

Haiwei Dong，*IEEE*高级会员，以及 Jeannie S.A. Lee，*IEEE*高级会员

arXiv:2301.07740v1 [cs.MM] 18 Jan 2023

摘要——**虚拟现实和360°立体流媒体等扩展现实（XR）技术，使元宇宙这一概念成为可能。元宇宙是一个用于协作与交互的沉浸式虚拟空间。为确保沉浸式媒体能够高保真显示，需要考虑带宽、时延和网络流量模式，以保障用户的体验质量（QoE）。本文通过示例和计算，探讨上述参数的要求。此外，从多媒体通信的角度，还可以探索未来的方法，例如利用强化学习（RL）实现网络感知，以及利用帧中的空间或时间差异实现XR内容感知。**

I. 元宇宙

元宇宙不再是假设，而是一个完全沉浸式、普适的虚拟空间概念，供多用户进行交互、协作和社交，并被视为互联网的下一次演进。“元宇宙”（Metaverse）一词由“meta”和“universe”组合而成，最早见于1992年的科幻小说《雪崩》。它描绘了一个通过佩戴头戴设备进入的虚拟世界，一个霓虹闪烁的另类领域，每个人都以定制的虚拟化身呈现。

元宇宙依赖多种广泛技术的融合，这些技术能够实现扩展现实（XR）。扩展现实是一个总称，涵盖现实—虚拟连续体上的各种技术，即虚拟现实（VR）、增强现实（AR）和混合现实（MR）。微软和 Facebook（现已更名为 Meta）等公司已在这些技术上投入巨资，同时也展望一种新的经济形态，让人们能够购买或交易数字资产，并买卖虚拟土地。

要呈现构成元宇宙的虚拟环境，可以通过开发包含虚拟对象、场景和虚拟化身的矢量三维环境来实现。另一种常用方式是通过 360 度立体视频流，将沉浸式元宇宙内容传输到头戴设备中。比如，用户可以在未来城市景观中体验虚拟漫步，选择不同的观看视角，并与场景中的不同对象或其他虚拟化身互动。

对于用户的体验质量（QoE）而言，高保真地呈现并交互此类沉浸式媒体至关重要。除了多视点360°视频，还需要支持涉及偏航、俯仰和翻滚运动的3自由度（3DoF）头部跟踪，以及包含x、y和z轴的6DoF。视觉、前庭和躯体感觉系统之间也需要保持协调，以便减少基于头戴式显示器（HMD）的扩展现实（XR）中的晕动症。

因此，所需分辨率为 4K，视频应支持多视角、360°，并具有纹理和深度；同时，虚拟环境对人体动作的响应时延应尽可能低。这需要大量带宽、处理能力和低时延。

II. 扩展现实（XR）通信 IONS

目前，XR内容是在XR头显中渲染的。然而，随着沉浸式体验的需求迅速发展，头显中用于XR的移动图形处理器的当前算力正逐渐成为瓶颈。鉴于摩尔定律目前正逼近其物理极限（晶体管尺寸正迅速接近原子级别），人们认为未来的XR将由云服务器进行渲染。XR内容及相关的同步数据和测量数据将通过XR云与XR头显之间的网络传输。要实现XR流传输，量化所需的带宽、时延和帧更新率至关重要。这些要求始终与人类感知的物理特性密切相关。

A. 低时延要求

端到端时延 [1] 可建模为感知时间、渲染时间、流传输时间和显示时间之和。理想的端到端时延应为 7 ms，因为前庭眼反射（VOR）过程需要 7 ms [2]。其中，感知时间是指头戴设备和控制器传感器进行测量所需的时间，主要用于定位和获取控制器输入，通常为 400 微秒。渲染时间涵盖前景和背景，具体时长取决于所构建虚拟环境的复杂度，约为 5 到 11 ms。显示时间是指将XR内容呈现在人眼前所需的时间。如果帧率必须达到 90 帧/秒，端到端时延就必须低于 $1 \div 90 = 11.1$ ms。若帧率为 120 帧/秒，端到端时延则需低于 8.3 ms。在这两种情况下，都必须尽可能利用专用集成电路缩短渲染和显示时间，以节省更多传输时间，从而为流传输时间留出更大的余地。

B. 高带宽要求

决定XR内容质量的主要参数包括每度像素数（PPD）、色深和帧刷新率。

- 每度像素数与显示屏的分辨率有关，表示每度所包含的像素数量。如果用户的观察区域宽度定义为 W 度、高度定义为 H 度，则 VR 单眼显示屏的总像素数为 $(H \times PPD) \times (W \times PPD)$ [3]。通常认为人眼可分辨的每度像素数为 60。在 VR 场景中，尤其是 360度全景虚拟现实中，每度像素数通常设为 64。传统每度像素数指标也可以是 11、21 和 32。
- 色深表示区分颜色灰度所需的位数。通常使用三原色（即红色、绿色和蓝色）的组合来表示颜色，其中颜色深度的位数为 2^{3x}。在扩展现实领域，色深通常设为 8 至 12 位。
- 帧刷新率定义为每秒刷新的帧数。较高的帧刷新率能够有效缓解运动模糊问题。人眼能够分辨的运动分辨率最高可达每秒 150 帧[4]。扩展现实中的帧刷新率通常设为每秒 90 帧或 120 帧。传统设备的帧刷新率可以是每秒 30 帧或 60 帧。

以 HTC Vive 或 Oculus Rift 作为 XR设备为例，其帧刷新率为每秒 90 帧，分辨率为 2K（即2160 × 1200）。下行原始数据速率可计算为 $3 \times 8 \times 90 \times 2160 \times 1200 = 5,598,720,000$ 位/秒，约为 5.6 Gb/s。由于数据速率非常高，传输前会使用编码编解码器进行压缩，例如 H.264、H.265 或 H.266。上述三种编解码器的压缩率分别为 102:1、215:1 和 350:1。因此，如果使用 H.264 压缩 5.6 Gb/s 的原始数据速率，压缩数据速率为 $5.6Gb/s \div 102 = 55$Mb/s。此外，双目压缩数据速率为 110 Mb/s。对于一些帧刷新率为每秒 30 帧的传统扩展现实设备，单目情况的压缩数据速率约为 37 Mb/s。

以另一个具体场景——360度虚拟现实为例，我们重新计算其数据速率。此处，$360° \times 180°$ 全景视图的视野和 $120° \times 120°$局部视图的视野分别为。即使将每度像素数设为 11，在全景虚拟现实场景中，其原始数据速率也为 $3 \times 360 \times 11 \times 180 \times 11 \times 8 \times 30 = 5.6$ Gb/s。对于双目虚拟现实显示器，原始数据速率为 11.2 Gb/s。如果计算时采用 64 每度像素数，原始数据速率可达 2.3 Tb/s。

C. 网络流量模式

网络中的流量包括下行流量（从 XR云 到 XR头显）和上行流量（从 XR头显 到 XR云），二者分别具有特殊特征。

1）下行数据流：下行数据流包括两种类型的数据包。一种来自渲染后的XR帧。由于实际的数据包大小受到最大传输单元（MTU）的限制，每个XR帧的左眼/右眼画面都会被拆分成多个数据包（每个数据包的大小为数千字节），并几乎同时传输，从而形成突发。例如，如果XR设备的帧率为 60 帧/秒，则每隔 $1 \div 60 = 16.7$ 毫秒就会出现两次数据包突发，分别对应左眼和右眼。其他数据包包括定期传输的同步数据包；以上例为例，每隔16.7毫秒传输一次，每个数据包只有几字节。

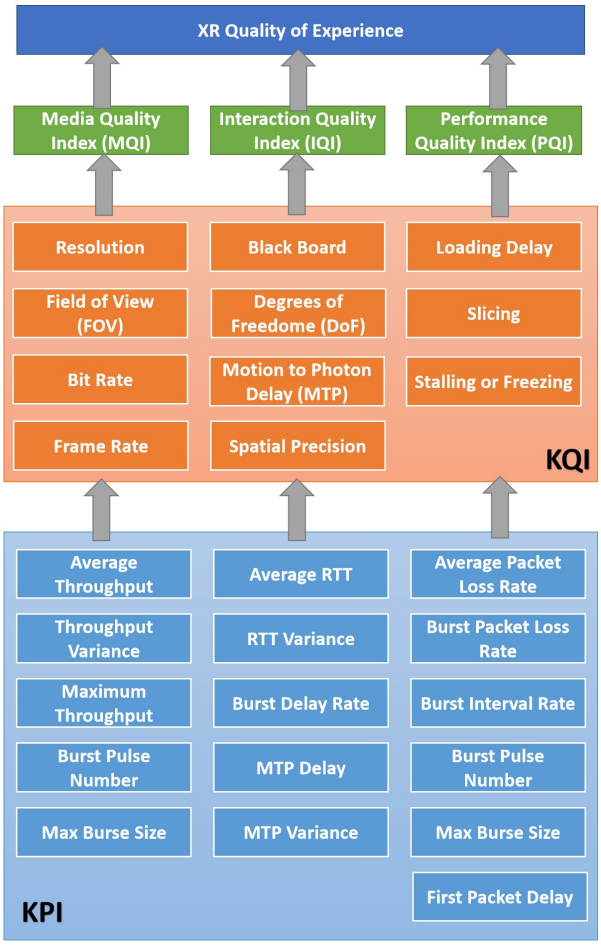


图 1. 决定 XR 场景中用户体验质量的因素。

2）上行数据流：上行数据流通常由 XR 头显的姿态、手部控制器的感知观测和控制信号以及同步信号组成。姿态测量包括 XR 头显在空中的位置和方向。它始终以六自由度描述，并依赖惯性导航。上行数据频率可能非常高，而上行数据包的大小则从几字节到 400 字节不等。

D. 体验质量（*QoE*）

体验质量用于量化用户对服务或设备的整体主观满意度。XR 体验质量方面已有相关标准，包括：

- 国际电信联盟提出的 ITU-T G.QOE-VR 标准，题为“虚拟现实（VR）服务体验质量（QoE）的影响因素”。

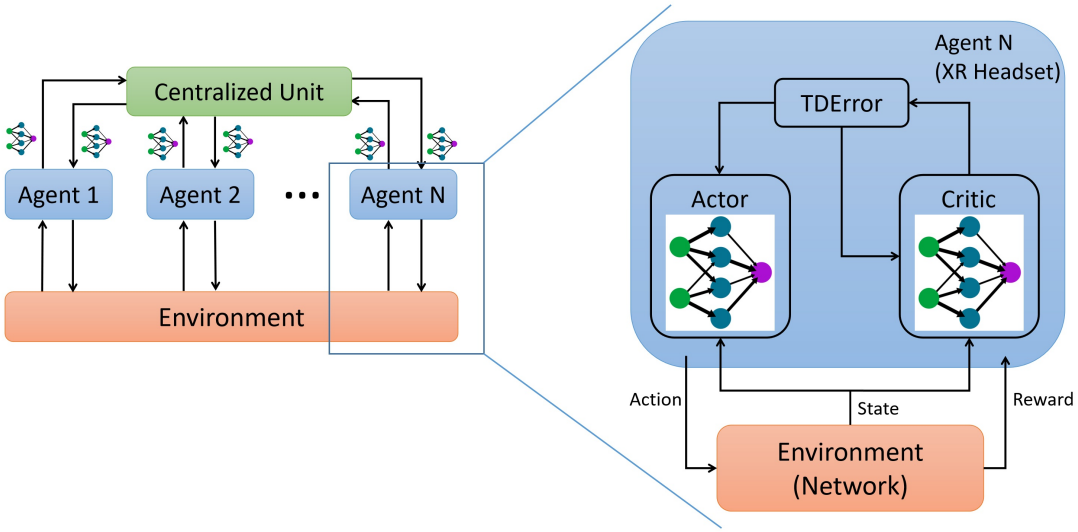


图 2. 集中式多智能体强化　　学习架构，使 XR 设备能够利用网络测量结果优化　　其体验质量。

- 3第三代合作伙伴计划制定的 GPP TR 26.929 标准，题为“与虚拟现实（VR）用户经验相关的体验质量参数和指标”。

- 国际电信联盟制定的 ITU P1203 标准，名称为“基于参数化码流的可靠传输渐进式下载和自适应视听流媒体服务质量评估”。

这些标准通常结合三个指数来推断 XR 体验质量 [5]：MQI（媒体质量指数）、IQI（交互质量指数）和 PQI（性能质量指数）。具体而言，如图 1 所示，每个指数都包含从 KPI（关键绩效指标）到 KQI（关键质量指标）的细化因素。KPI 从工程系统的角度评估性能，而 KQI 和体验质量则旨在从人的角度量化感受。这些因素通常会按照不同权重加总，而权重由人体研究的问卷结果通过推理（例如模糊推理 [6]）确定。

收集上述推理所需数据有两种方式。一种方式适用于从终端主机或云服务器的应用层获取 XR 数据的情况。这种方式适合测量用户的真实体验，但需要在终端主机或云服务器中部署软件开发工具包。另一种方式则通过捕获和使用运营商网络管道中的数据，间接测量用户的体验质量；这种方式需要使用网络探针，例如带内和带外网络遥测。

III. 扩展现实与网络集成

A. 网络感知型*XR*系统

连接 XR 头显与 XR云（位于园区或数据中心）的网络会因并发流而出现各种状况，从而引发拥塞和微突发等网络事件。为减轻这些网络事件的影响，XR 系统需要调整向网络注入 XR 数据包的行为。由于这里的网络非常复杂，难以建模（尽管在简化场景下，有时可以用网络演算进行建模），因此通常将其视为黑盒。在这一假设下，如图 2 所示的多智能体强化学习（RL）可能是一种极具前景的解决方案。

这里的目标是根据奖励准则，建立从状态到动作的映射关系（称为策略）。为了能够处理高维状态空间，使用神经网络来近似这一映射关系。定义网络结构（图2中的全连接网络）后，只需更新神经网络的权重参数。具体而言，本文采用的学习策略是行动者-评论家：行动者在大多数时候遵循先前学习得到的策略（通常采用 ϵ-贪婪方式），同时充分探索环境。评论家利用行动者在一段时间内积累的经验来计算策略，并定期与行动者一起更新其最新策略。值得注意的是，这里使用时序差分误差，通过归一化权重结合预期奖励与实际收到的奖励，从而高效地更新策略。具体定义如下。

- 状态：包括吞吐量、延迟、抖动、丢包率等在内的网络测量。

- 动作：由 XR 头戴式设备终端控制的 XR 数据包比特率。

- 奖励：由第 II-D 节讨论的各项因素组成的体验质量指标。奖励函数的一个示例 [7]可以是平滑度（即帧率变化）带来的奖励、XR 显示屏质量（即分辨率变化）带来的奖励，以及时延带来的惩罚的组合。

- 强化学习算法：如果假设每个强化学习智能体负责一台 XR 终端设备，那么多台 XR 设备就需要多个强化学习智能体协同工作。在这种情况下，这些智能体会彼此或与邻居共享学习得到的策略、观测或经验，从而整体上获得较高的奖励。现成算法在连续或

离散状态/动作空间以及集中式或分布式风格方面各具不同特点。

*B. XR*感知智能网络

如第 II-C 节所述，XR内容在传输前会经过压缩。基于 MPEG压缩的传输，其基本思路是传输关键帧以及后续若干帧的时间或空间差异。具体而言，关键帧有时也称为镜头的帧内帧或 I帧，并会定期生成。预测帧，简称 P帧，用于描述当前帧与其前一帧 I帧或 P帧之间的时间或空间差异。双向帧，简称 B帧，用于描述当前帧与其前后 I帧或 P帧之间的差异。因此，发生网络拥塞时，对应不同帧（即 I帧、P帧、B帧）的 XR数据包可能会在网络设备（如交换机或路由器）中被丢弃。如果被丢弃的数据包对应 I帧，XR内容将无法解压缩，这会严重影响用户的体验质量。另一方面，P帧对最终体验质量的影响最小，因此必要时可以考虑将其丢弃。一旦 XR数据包头中标明了对应的帧类别，网络便可以在不同层级的网络管理中利用这一信息，例如队列管理、负载均衡、微突发缓解等。

IV. 结论

为了实现元宇宙未来的愿景，现在正是认真考量360°立体流媒体高要求的合适时机。本文通过简单计算，清晰展示了时延、带宽和体验质量方面的要求。通过网络与扩展现实应用的整合，本文从多媒体通信的角度展望了技术如何发展以支持元宇宙。

参考文献

[1] L. Liu、R. Zhong、W. Zhang、Y. Liu、J. Zhang、L. Zhang 和 M. Gruteser，“切断连接线：设计一种采用低时延远程渲染的高质量无线虚拟现实系统”，载于第*16*届移动系统、应用与服务年度国际会议论文集，2018年，第68–80页。[2] A. Seam、A. Poll、R. Wright、J. Mueller 和 F. Hoodbhoy，“借助5G网络实现移动增强现实与虚拟现实”，AT&T，技术报告，2017年。[3] H. Zhang、Z. Yang 和 P. Mohapatra，“终极虚拟现实360度视频的无线接入”，载于互联网设计与实现国际会议论文集，2019年，第271–272页。[4] E. Bastug、M. Bennis、M. Medard 和 M. Debbah，“迈向互联虚拟现实：机遇、挑战与推动因素”，*IEEE*通信杂志，第55卷，第6期，第110–117页，2017年。[5] “5G云VR服务体验标准白皮书”，华为，技术报告，2019年。[6] L. Zhang、H. Dong 和 A. El Saddik，“构建用于评估全息增强现实设备的体验质量模型：HoloLens案例研究”，*IEEE*多媒体，第26卷，第2期，第21–32页，2019年。[7] Z. Long、H. Dong 和 A. El Saddik，“通过远程渲染使用HoloLens与纽约市数据交互”，*IEEE*消费电子杂志，第11卷，第5期，第64–72页，2022年。

Haiwei Dong 是华为加拿大（位于加拿大安大略省渥太华）的首席研究员。他的研究兴趣包括人工智能、多媒体通信、多媒体计算和机器人学。Dong 在日本神户的神户大学获得计算机科学与系统工程博士学位。请通过 haiwei.dong@ieee.org 联系他。

Jeannie S.A. Lee 是新加坡理工大学副教授。她还担任项目主任和沉浸化中心副主任。她此前曾在高通从事多媒体算法相关工作。她拥有佐治亚理工学院计算机与电气工程博士学位。联系方式：jeannie.lee@singaporetech.edu.sg。