CHINA TELECOM

Audio and Video Empower Metaverse

Yuan Zhang

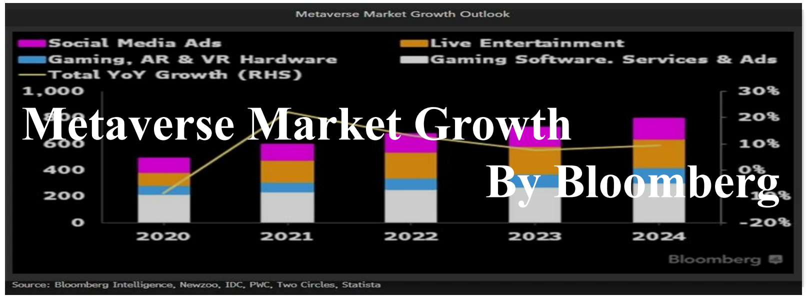
Co-Chair WP3/16

Rapporteur Q12/16

18 October 2022

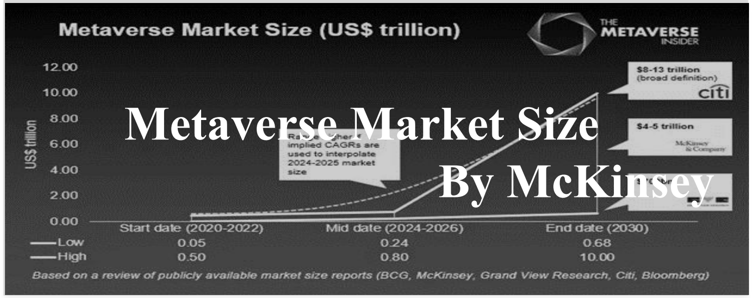
Global Metaverse Market Size

Global Metaverse Market Size



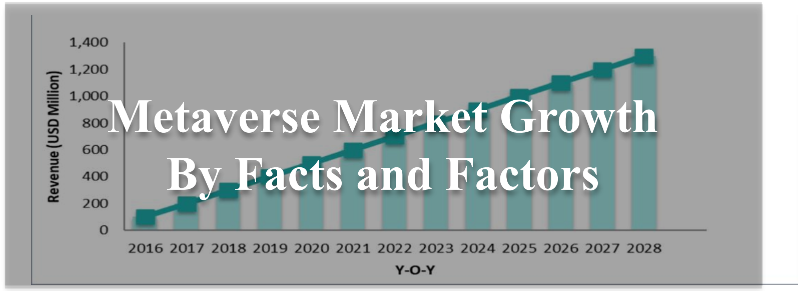
Source: Bloomberg Intelligence, Newzoo, IDC, PWC, Two Circles, Statista

Metaverse Market Growth
By Bloomberg



Based on a review of publicly available market size reports (BCG, McKinsey, Grand View Research, Citi, Bloomberg)

Metaverse Market Size
By McKinsey



Metaverse Market Growth
By Facts and Factors

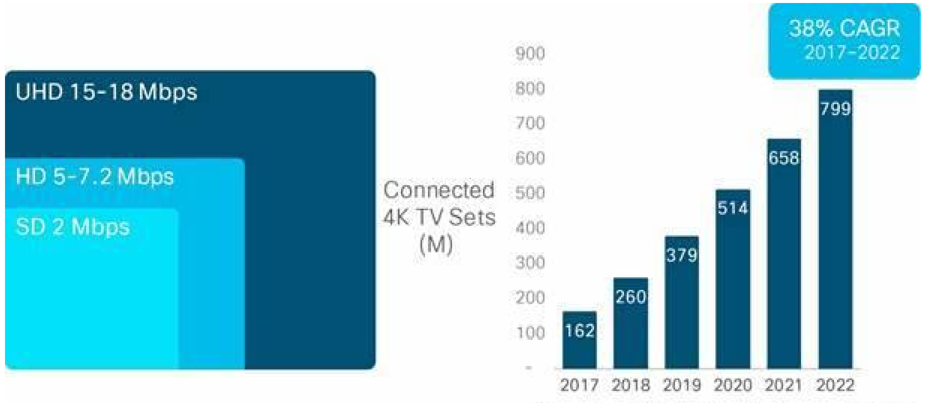
Year	Forecast Provider	Market Size
2024	Bloomberg	$800 Billion
2028	Facts and Factors	$730.5 Billion
2030	McKinsey	$5000 Billion

Source of Network Traffic

According to Cisco, video will make up 82% of all internet traffic in 2022.

3D, AR, VR, MR and XR, etc., will account for over 90% of Internet traffic.

Video data become the major driving force for the network technology and network economy.

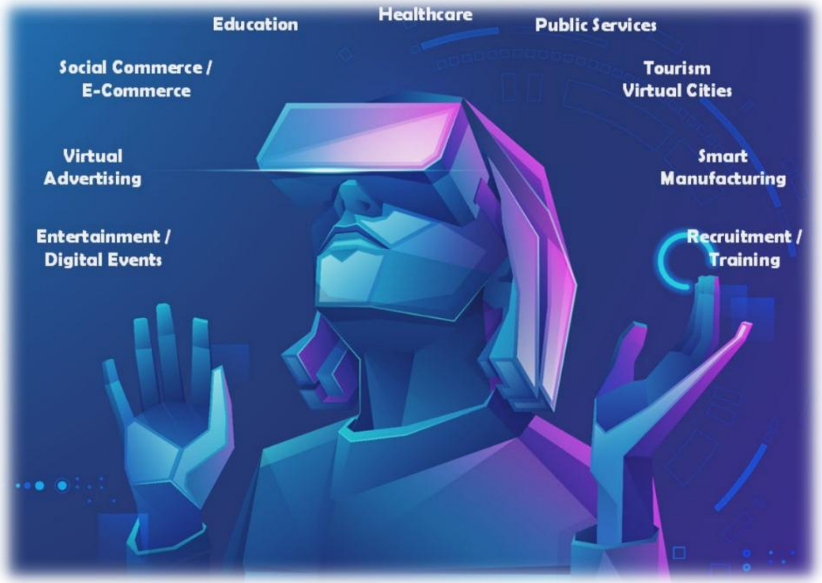


Source: Cisco VNI Global IP Traffic Forecast, 2017-2022

Metaverse Use Cases

- Developer/Creator Economies, Games
- Virtual Workspaces, Communities
- Digital Entertainment Events
- Social Commerce and e-commerce

- Smart Manufacturing
- Healthcare
- Education
- Climate Change



Challenge and Gap of the Metaverse

To realize the true vision of the Metaverse, immense investment would be needed in computing power, networking, algorithm and data, at orders of magnitude higher than what exists in today’s world.

Computing Power

The metaverse requires a level of computational power far beyond Moore's Law

01

02

Network

The metaverse requires a network with full coverage and almost no delay

Algorithm

The metaverse requires to provide high quality service or obtain high quality experience

03

04

Data

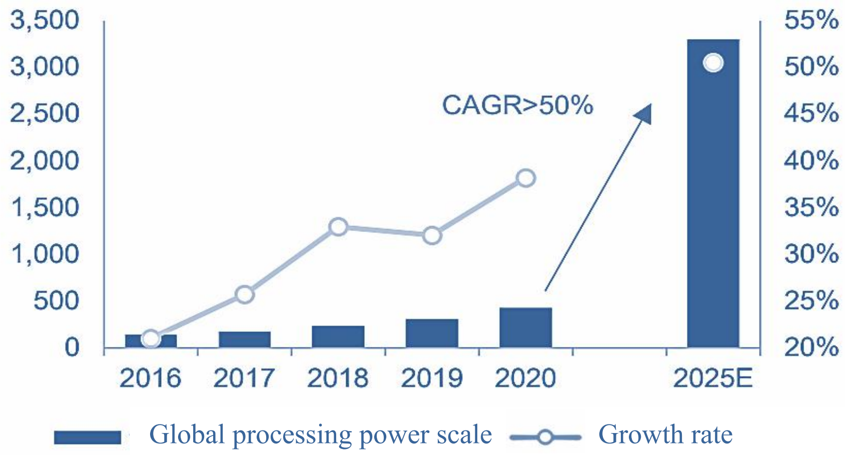
Data security issues, data privacy, data management, data protection

Computing Power Challenge

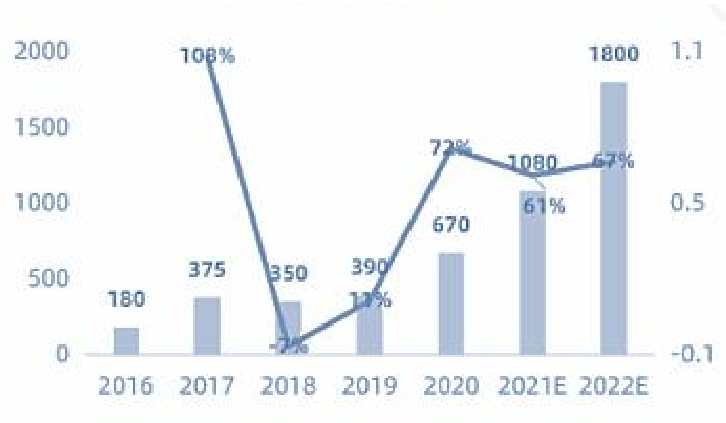
To support the virtual content creation and immersive experience in the metaverse. More real-time modeling and interaction with large scale of users requires super computing power. Global computing power growth falls behind the growing of data and algorithms. As predicted by Intel, the metaverse would require a exa flops (10^{18} flops), which is **1000-times** the current total computing power.

It requires:

- More computing infrastructure, data center with intelligent computing and green computing power
- Semiconduct innovation
- More efficient algorithm



Global VR headset shipments (10,000 units)

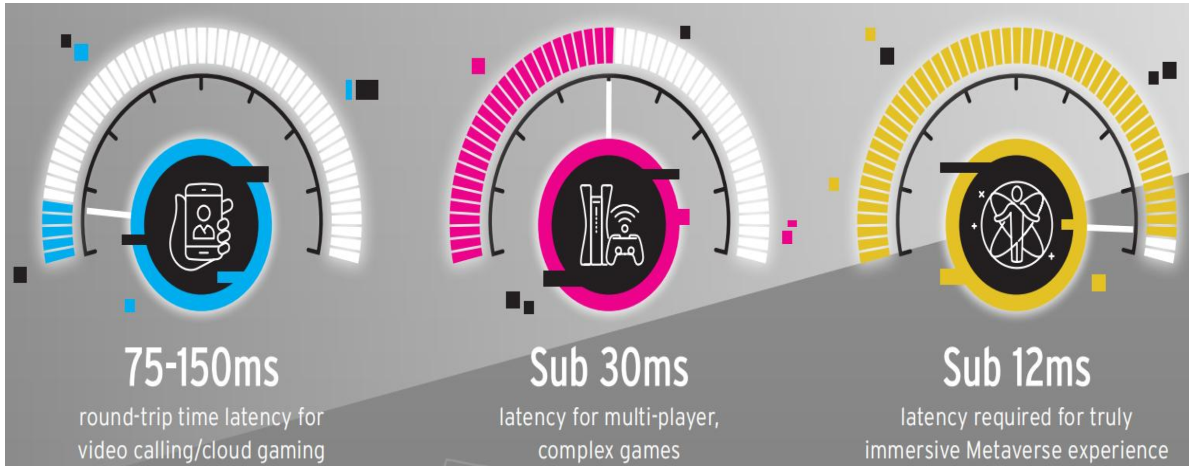


- PS5 level game VR game 4k@60Hz, requires 10TFLPOS computing power.
- Interactive VR requires twice the performance of PS5, 20TFLPOS computing power.
- In the metaverse, the AR/VR computing power should reach 3900EFLPOS.

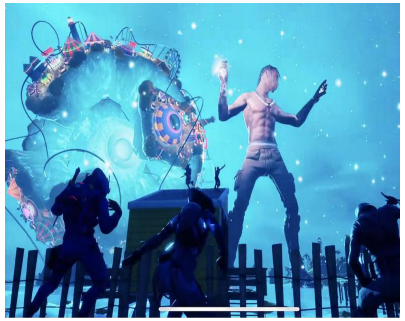
The standard Moore's Law curve only allows for a power increase of about 8-10 times over the next five years

Network Challenge

The Metaverse requires far more bandwidth than the majority of Internet applications and games. The ”Mirror World" in the metaverse needs to reconstruct the infinite diversity of the real world and capture the dynamic changes in real time, "evolving" along with the real world. A super-large-scale, real-time, interactive and persistent virtual environment requires a network featuring **global coverage, large bandwidth and almost no delay** to continuously provide immersive content and real-time interactive to users.



Microsoft Flight Simulator provides real-time updates based on real-world weather and air traffic conditions.



Current metaverse concerts send patches in advance rather than in real time to allow players to travel through virtual world scenes.

For every 10ms increase (decrease) in game latency, the user's game time decreases (increases) by 6%

Immersive Experience Gaps

The metaverse is likely to be accessed via a head-mounted display, centimeters away from the eye, requiring large resolution videos, well beyond 4K.

It needs to provide 3D video, multi-degree of freedom video, spatial and temporal video and audio, stereo audio,both for human and machine vision.

The computing power and network should be content and semantic oriented.

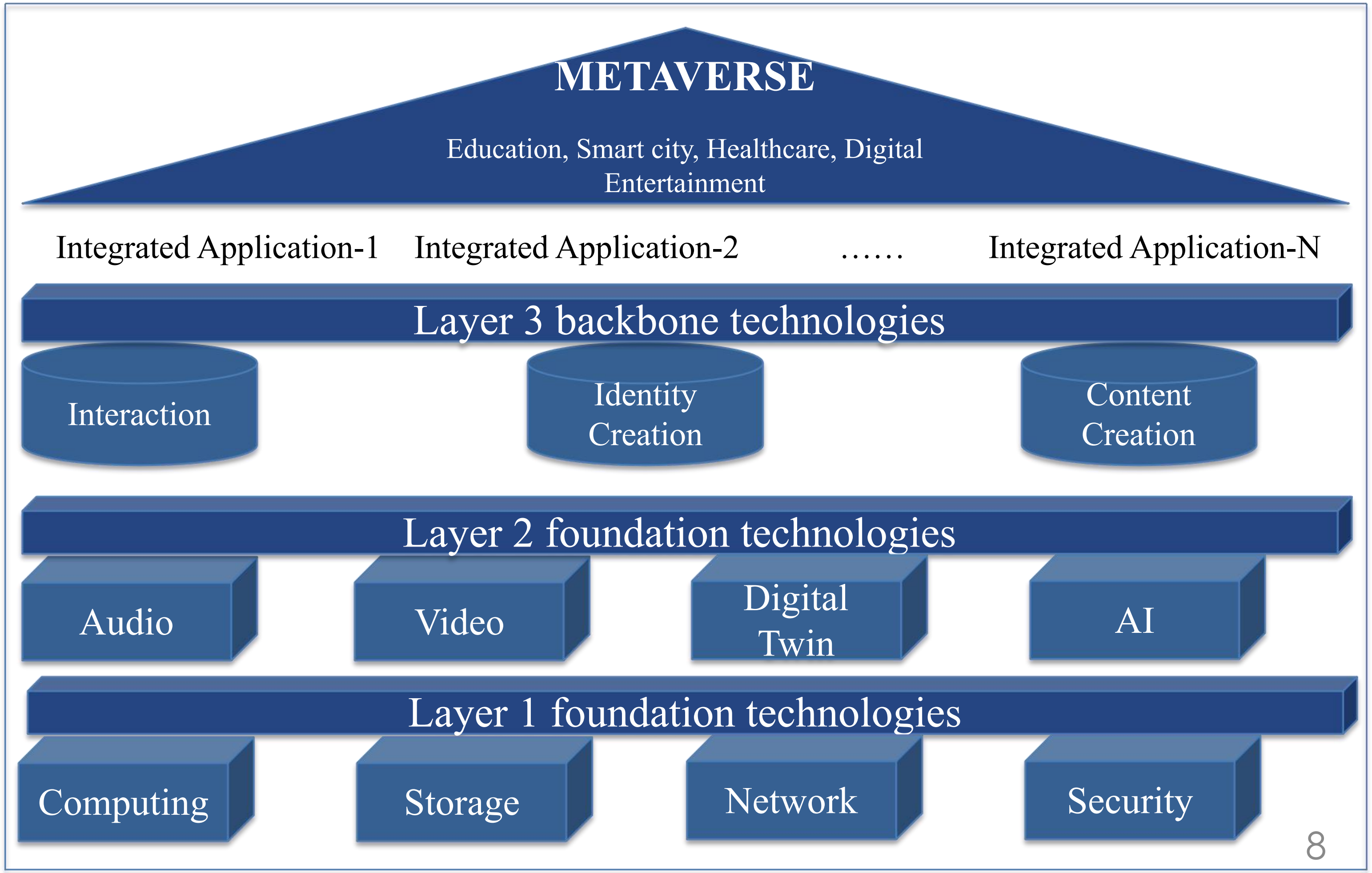
Metaverse requires substantial improvements in algothrims and nautral and generated data content and innovations across the hardware and software stack.

Enabling technology

The metaverse is a digital world amalgamation in which augmented, physical, and virtual realities converge.

Core technologies are driving the success of designing an efficient metaverse environment, among them, video and audio are key enabling technologies.

- The metaverse is one of the main development directions and representatives of the next generation of Internet, featuring holographic and omnipotent.
- The metaverse is the fusion and virtual space where people live and work.
- The metaverse is an important creation and productivity of future digital assets.



Key Video/Twin Technology

The core video and related technology includes video coding, processing, digital twin, 3D and spatial computing, which can reconstruct the details of digitalization. It can create a 1:1 replica with comprehensive information for people, objects and environments. It connects, maps and couples the digital world and the real world, with real-time synchronization, represents and integrates the virtual and real in the metaverse.

Video processing

Including video codec, transmission, analysing and rendering. It has certain foudation, moving towards diverse scenarios, more efficient codec and lower lantency.

In the future, transmission, processing and coding efficiency will be improved with the support of AI, to provide more stable, higher resolution, lower latency, diverse applications, with video as the main carrier.

Digital Twins

At present, the technology has been applied preliminaryly. It has great potential. In the future, digital twin technology will develop towards the direction of wide coverage, fine-graned and real-time. Enabled by AI, the capability of autonomous analysis and decision-making of digital twins will be expanded. The mapping, connection, and interaction between physical entity and digital twin can form a complete closed-loop system of two-way feedback

3D

Relatively mature.

In the future, the 3D technology together with AI continue to improve the efficiency of afforable, friendly and well functioned content. The AI-optimized 3D engine and toolchain can automatically generate a digital twin to better support immersive environments and digital humans.

Spatial Computing

Still in its infancy.

In the long run, spatial computing technology will help realize spatial semantics and improve the visibility, interoperability and interactivity of spatial data, dramatically enhancing the 还原度of digital world.

Key Audio Technology

The core audio technology of metaverse includes audio recogonition, audio synthesis, acoustic field reconstruction, and xxx, which enables the metaverse to interact through natural language with authentic, emotional and unique voice and immersive audio.

Audio Recognition

Speech recognition technology takes speech as the research object, and enables computers to automatically recognize and understand human spoken e by recognizing and processing language signals. Through the process of speech recognition and understanding, computer converts the speech signal dictated by human into the text that can be processed by machine.

Audio Synthesis

Speech synthesis is essentially the process of transforming text information into speech information according to various needs such as timbre and emotion. The first step is to add prosodic information to the text, and the acoustic model can generate acoustic features according to the pre-processing results. Finally, the vocoder generates speech samples by using the pre-order information.

Acoustic Field Reconstruction

Acousticfield reconstruction aims to present the real sound field distribution in a specific environment by using a loudspeaker array, so that people feel as if they are in the scene and feel the real sound effect and sound quality. The sound field information is transformed and processed accordingly to solve the speaker array signal, and then the real sound field is reconstructed.

Emotion Recognition

Emotion recognition includes speech rhythm detection, including fenquency, energy and duration, tone quality detection and deep features. emotion recognition, It has been used in call centersto improve their service and convert more people.

Standardization Consideration

- ITU-T SG16 WP3 Audiovisual technologies and intelligent immersive applications
  - ✓ Q5/16 on Artificial intelligence-enabled multimedia applications
  - ✓ Q6/16 for Visual, audio and signal coding
  - ✓ Q8/16 for Immersive live experience systems and services
  - ✓ Q12/16 on intelligent visual systems and services
- ITU-T SG16 WP1/16, WP2/16
  - ✓ Q22/16, Q23/16, Q24/16, Q26/16
- ITU-T SG11, 12, 13, 15, 17, 20
- SDOs: ISO/IEC JTC1, IEC, IEEE, IETF, 3GPP

Key Messages and Hints for Operators

NETWORK

CLOUD

DATA

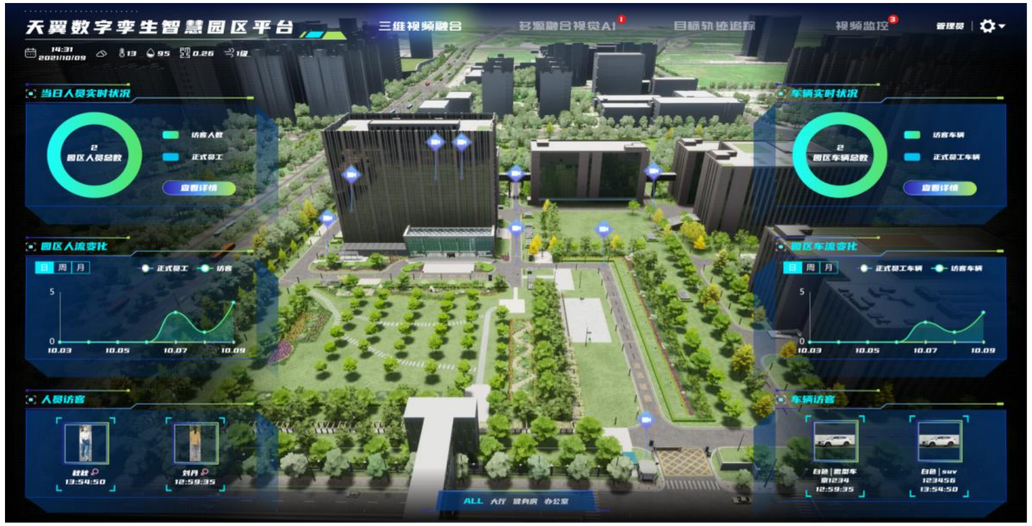
COMPUTING

In metaverse era, operator can position itself majorly as “the constructor of metaverse infrastructure”, taking advantage of its rich network, cloud, data and computing resources to promote the integration and collaborative innovation of the ecosystem.

- Develop the core capabilities of audio, video, AI technology, i.e., video coding, 3D rendering, audio interaction, etc., to provide and operate a platform integrating IaaS and Paas capabilities.
- Design and build applications of business, education, culture metaverse, both to business customers and end users.
- Cooperate and collabrate with ecosystem in the prospective of technology and industry breakthroughs in chips, AR/VR/MR/XR, wearable devices, and naked eye 3D, etc.

Practices - Metaverse Digital Twins Map Space System

- AI codec
- Digital twin
- 3D video
- B/S architecture implementation



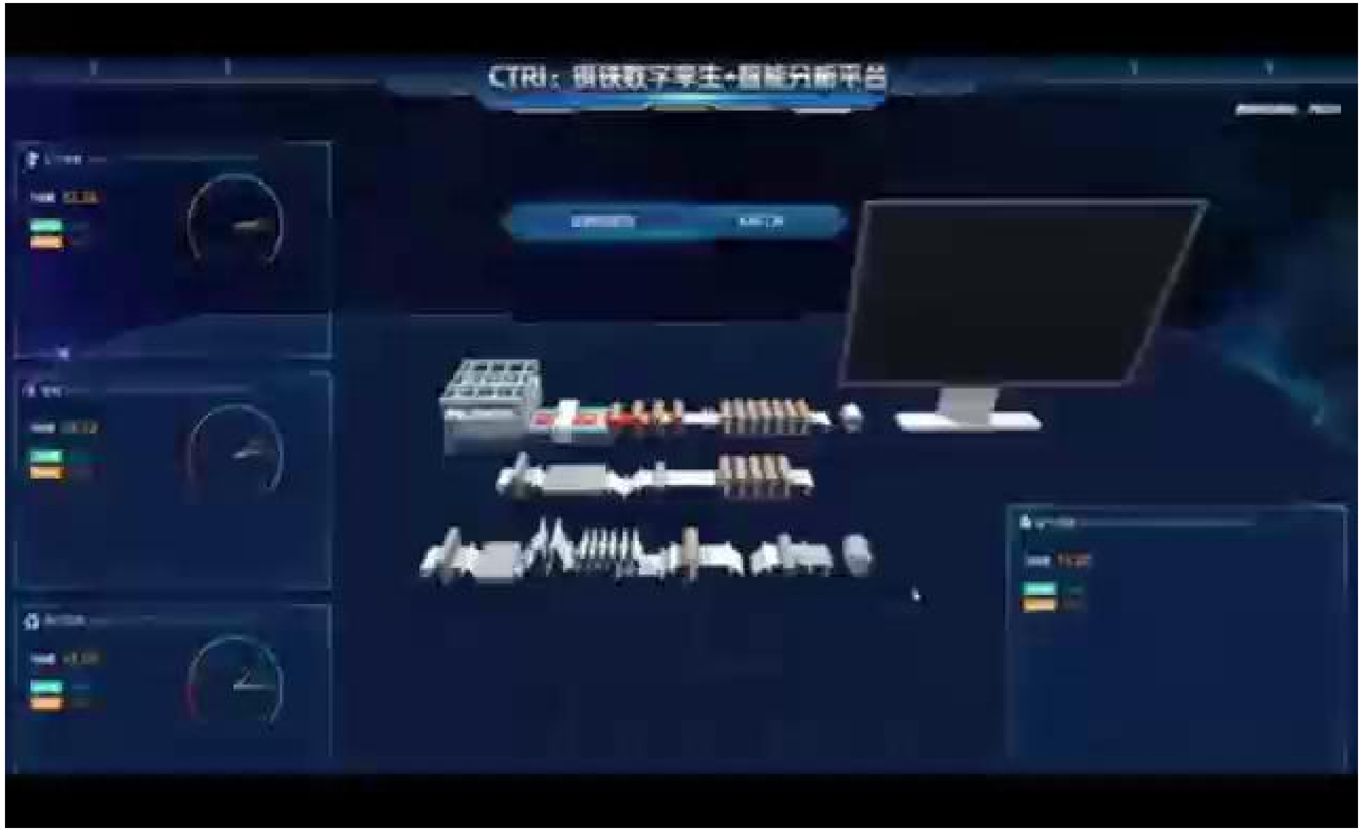
- The multi-dimensional information visualization can view the video of different regions such as parks, communities and business zones and other multi-dimensional information in real-time. It can display and quickly locate the information point.

- It adopts the new video codec and video rendering engine to encode the graphics information to bitstream, and pushes it to the front-end web system through WebRTC.

- The multi-dimensional information such as the Internet of vehicles, equipment and monitoring with the 3D model is integrated and displayed on the platform to perceive the environment.

Practices - Industry Twin and video recognition

The industry metaverse application is the integration of the digital twin of the steel production line, object generation, machine vision algorithm and IoT sensors. It integrates our coding algorithm, real-time video tracking and detection, small target detection and content generation algorithm.



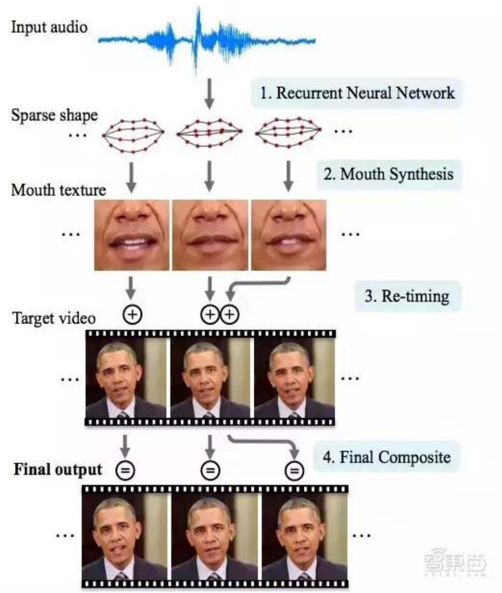
Practice - Lip Synchronization and Audio Identification

Speech-Driven Lip-Sync for Digital Human

1. Computed a sequence of phonemes from the speech transcript and audio, and locate the start and end time of each phoneme

2. Determine the corresponding mouth shape for different phonemes and form the mouth shape time series data for the phoneme

3. Render the mouth shape to the corresponding character image and obtain a matched mouth shapes



Speech Recognition

Asynchronous Control

Rendering

Mouth Shape Sparsing

Speaker identification

1. In the training phase, the model is trained to increase the feature distance of different speakers and decrease the feature distance of the same speaker

2. Increase the authenticity of the training data: add reverberation (weak acoustic reflection in the room) and thousands of real noise, change the voice speed, change the voice volume

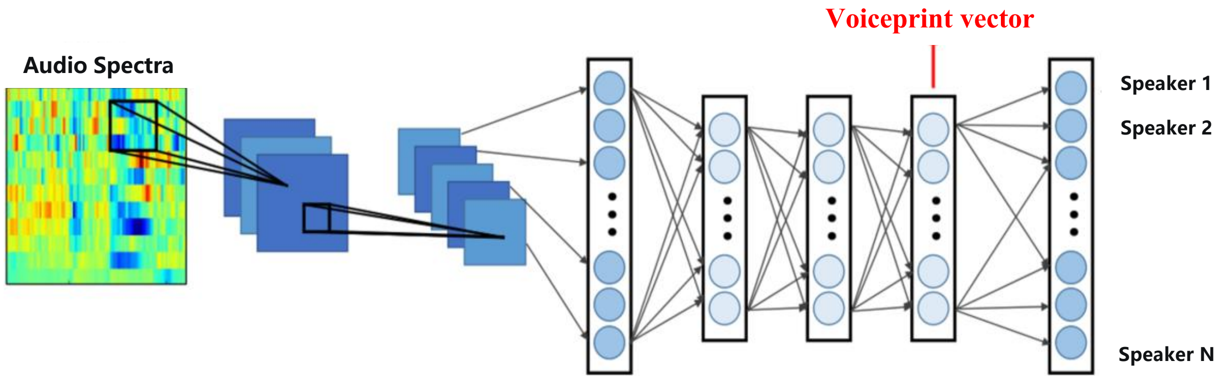
3. Released more than 10 model versions to support business needs under different scenarios (i.e., China Telecom customer service system)

Speaker Identification

Who is this?

Speaker Verification

This is he?



守正创新 开放融合
云网赋智 兴企报国

Yuan ZHANG

zhangy666@chinatelecom.cn

THE END