CONTEC

WHITE PAPER

AUGMENTED REALITY (AR)

The AR Glasses Ecosystem: User Endpoint, OS & Device Management, Utility, and Backend

KLINT MARTIS - AUGUST 2026

www.us.conteccom

Table of Contents

Abstract

Augmented reality is a technology where digital images are projected or displayed on screens that can be portable. AR glasses are one such implementation with a four-layer ecosystem including the User-Endpoint, OS & device management layer, Utility Layer and Backend.

User endpoint expands on the hardware in the AR glasses, OS & device management expands on the foundational code architecture for the end-point operation, utility layer expands on the functional integration of the two previous layers, and Backend expands on the processing and storage of data transmitted from the end-point.

Contec can be a partner in the implementation as a supplier of the server to fill in the backend layer and the integration of the user-endpoint data in the server.

Introduction

Klint Martis
Integrated Solutions Engineer

In the workforce of today there is a push towards efficiency through use of automation. Tasks that are tedious and require multiple clicks can be scripted out easily enough, but what about finding that efficiency in real world interactions?
In comes augmented reality. A technology that can be described as reality embedded in a user interface. AR Technology involves computer generated images being overlayed onto real world environments.

Pioneered in 1968 by Ivan Sutherland and his “Sword of Damocles” as the first head-mounted display, it was not until the 2010's where the technology was mainstream accessible with the trending app at the time, Pokémon Go!

In 2026 with the ease of commercial access to machine learning servers and digital image processing at the edge, AR is seeing higher use cases in multiple verticals; manufacturing, warehouse, and medical to name a few.

This paper will look at the overall ecosystem for AR solutions, specifically AR glasses, which can be used in many different verticals. The ecosystem includes the User Endpoint, OS & device management layer, Utility Layer and Backend layer.

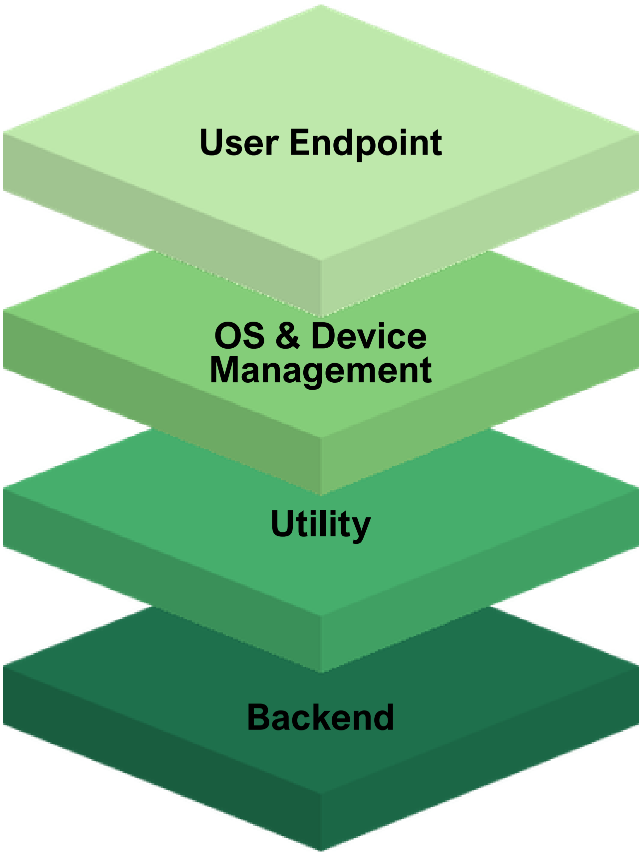

Figure 1: AR glasses Ecosystem Stack

Ecosystem User Endpoint

The user endpoint for AR glasses requires that the hardware has a camera and other sensors as inputs, a processor, battery and outputs as a projection or display.

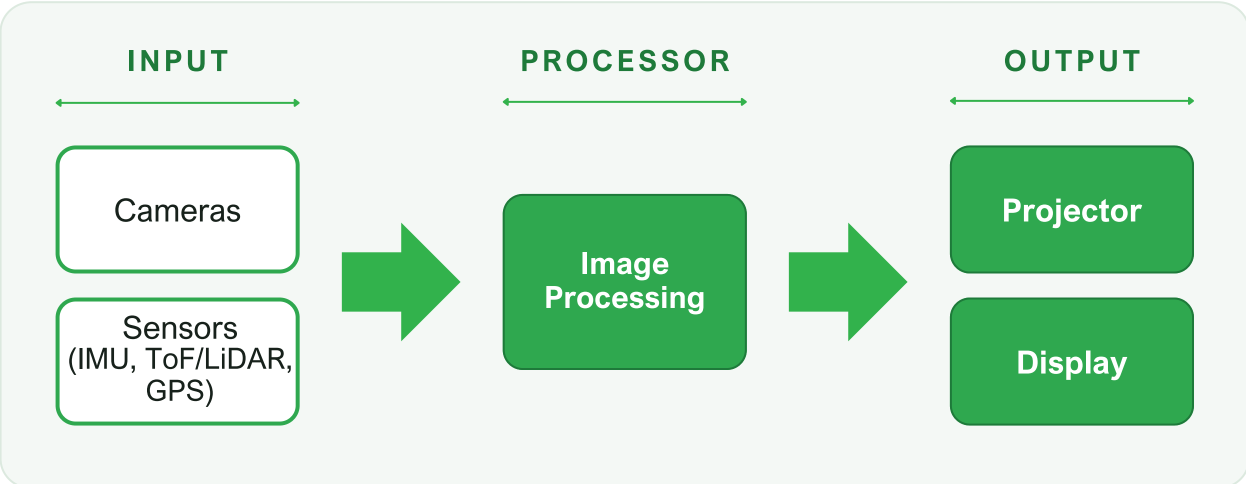

Figure 2:AR glasses user endpoint block diagram: input, processor, and output

The camera is used to take in the user's surroundings. Sensors like Inertial Measurement Units add data on orientation and 3D movement, Time-of-Flight or LiDAR to obtain data for depth and GPS for location data.

All this data is then processed using ARM based processors the low-power and thermal are ideal for wearable User Endpoints for which light computational processing is needed, while x86 processors can be used for heavier workloads.

Output is then either projected from a micro-OLED light engine as a waveguide that relays the image into the eye or displayed on a Micro-OLED display where the real world is reconstructed onto the screen. For most industrial applications projection using optical waveguide technology is used as it overlays images onto the user's field of vision and is not obstructing the user's view.

With advancements in battery technology and low power consumption of embedded processors, long duration use of headsets is possible and does not hinder operator maneuverability and allows it to be more enticing for use. Additionally, Wi-Fi and Bluetooth functionality are used for data transmission to the backend.

OS & Device Management Layer

Now that we have the hardware defined, the next layer of the ecosystem is the OS that sits on the User-endpoint and dictates how the device component interacts with each other.

or industrial AR glasses the dominant foundation is the Android Open-Source Project (AOSP). Devices such as the RealWear and Vuzix headsets used in field deployments run AOSP builds customized by the OEM, which allows standard Android applications to be developed for the user interface while giving the manufacturer control over the boot image, driver stack, and update channel. Android XR is a newer Android-family platform purpose-built for extended reality, but its current deployments are consumer and prosumer devices rather than industrial headsets. It is best understood as the direction the Android XR ecosystem is heading rather than the stack an industrial buyer would specify today.

The principal alternative has been Windows Holographic; the operating system used on Microsoft HoloLens. It is worth noting that Windows Holographic is not tied to a particular instruction set with the original HoloLens used an Intel x86 SoC, while HoloLens 2 moved to a Qualcomm Snapdragon 850 running ARM64. With HoloLens 2 production now ended, the practical near-term choice for an industrial deployment is an AOSP-based headset.

Where x86 enters an AR system is on the compute side rather than the headset in tethered processing units and in the backend server for offloaded work Device management varies based on scope. In increasing complexity of implementation Mobile Device Management (MDM) focuses on managing mobile devices, Enterprise Mobility Management (EMM) expands beyond the devices and includes applications, content, and identity management, and Unified Endpoint Management (UEM) encompasses all endpoints across all platforms and device types [1].

Device management matters to the business case because it is what keeps a fleet of headsets configured identically, secured, patched, and monitored for faults. The Coca-Cola HBC deployment described later in this paper scaled from sixteen headsets to roughly a thousand, and that scaling is only tractable with a management layer in place.

SCALING MATTERS

The Coca-Cola HBC deployment described later in this paper scaled from sixteen headsets to roughly a thousand, and that scaling is only tractable with a management layer in place.

Utility Layer

This layer sits between the OS & Device Management and Backend layers, and its purpose is to bridge the two together. The utility layer can be subcategorized into UI authoring, which is how the overlay is created; Application Runtime, which is how the device works with the overlays; and Backend integration, which determines what overlay is needed. So, a typical flow using PTC Vuforia Work Instructions [2] as an example would be authoring (expert capture, studio, chalk) -> Used on the runtime application on the headset -> Data transmission to and from the backend (server, ERP/CRM).

Continuing with the PTC Vuforia work instructions example, there are multiple methods of authoring that are offered. Expert capture makes use of recording frontline workers, Vuforia studio gives access to use CAD models into AR instructions, and Vuforia chalk is for quick AR collaboration allowing an expert to interact with a field technician. The 3D objects are then attributed and mapped spatially to have the overlayed object anchored to physical objects.

The runtime application takes these overlayed objects and projects them onto the screens and is what allows operators to see procedures overlayed in front of them without needing to look away.

With backend integration this is how the device determines the procedure to prioritize. Storing data on a server and connecting it to the ERP/CRM is what allows the system to overlay correctly. For example, the operator wears the glasses, scans a barcode with the build name and then queries against the ERP to pull the correct work instructions from the server.

TYPICAL FLOW

authoring (expert capture, studio, chalk) → runtime application → data transmission to and from the backend

Backend Layer

The last and certainly not least layer of this AR ecosystem is the Backend Layer. This dictates where do you want the bulk of the processing to sit. Options could be on an edge device or in a centralized cloud location. A benefit to edge server over cloud is reduced latency in real-time applications. For edge processing, data is transmitted to a high-performance server for processing, and the results are then sent back to the endpoint [3]. Pasman and Jansen demonstrated that low latency is the major constraint for AR applications and has been estimated that for wearable AR the expected latency should be no more than 10ms [4].

Large-scale measurement supports the case for the edge. Charyyev, Arslan, and Gunes measured round-trip latency from 8,456 end users to 6,341 edge servers and 69 cloud locations and found that 58 percent of end users could reach a nearby edge server in under 10 ms, while only 29 percent obtained comparable latency from a nearby cloud location [5]. Across the full dataset, edge servers delivered lower latency than the cloud for 92 to 97 percent of end users [5]. Datacenter distribution geographically is an important consideration. At the time of the study none of the measured cloud providers operated a datacenter in Africa, as a result edge deployment was the only route to low latency service in that region [5]. In regions where cloud datacenters are abundant, such as Western Europe, cloud locations can approach the latency of edge servers [5]. The advantage of an edge deployment is therefore largest where the user is far from a datacenter and narrowest where they are close to one. For manufacturing or clinical deployment, the relevant distance is not to a regional datacenter at all but to the plant's own network, which is why a local edge server remains the stronger architecture for this class of application.

Edge capacity and resources need to be planned accordingly. The same study found that with only edge servers, 59 percent of user requests are served in under 10 ms at low workload, but that figure falls below 40 percent as workload intensifies [5]. Without cloud support, coverage never reaches 100 percent, because some requests cannot be scheduled onto any edge server within reach [5]. The practical conclusion is hybrid architecture, in which latency-critical processes are held at the edge, and model storage, fleet analytics, and offline processing offloaded to the cloud; edge server should be sized for peak concurrent sessions rather than average load.

One caveat on the measurements above. That study characterizes network round-trip time to distributed CDN infrastructure, not compute latency for an AR workload on a dedicated server. A local edge server sitting on the same LAN as the headsets would be expected to perform better than any figure reported there. However, the measurements in the study establish an estimate of the edge advantage; they do not substitute for benchmarking a specific configuration against a specific workload.

KEY ARCHITECTURE TAKEAWAY

The practical conclusion is hybrid architecture, in which latency-critical processes are held at the edge, and model storage, fleet analytics, and offline processing offloaded to the cloud; edge server should be sized for peak concurrent sessions rather than average load.

Published Use Case

COCA-COLA HELLENIC BOTTLING COMPANY (HBC)

Problem Statement
The HBC is a bottler for Coca-Cola that produces more than two billion units per year. Warehouse procedures were severely outdated, with operational inefficiencies in picking and shipping [7]

Implementation
HBC implemented a vision picking system using RealWear assisted-reality headsets running TeamViewer Frontline software that displayed line-of-sight reference data for operators. Reference data included target items, locations, and quantities, and the headsets were integrated with SAP, their warehouse management system [7].

Results
HBC achieved an 8 percent increase in picking performance and an overall 99 percent accuracy rate [7]. HBCscaled from 16 pickers at one site to approximately 1,000 pickers across 30+ warehouses [7].

PICKING PERFORMANCE
8% increase

ACCURACY
99% overall

SCALE
1,000 pickers

Stanford Medical

Problem Statement

Stanford Medical physician Alexander Perino realized in an OR there are multiple screens displaying real-time data, with insufficient real estate in the room and required staff members to assist with data manipulation and processing, which can be inefficient [8].

Implementation

Alexander Perino used the Apple Vision Pro to combine his monitoring screens with some human-computer interaction. The surgical team was still present as a fail-safe [8].

Results

Promising integration of the technology for surgical use. Alexander Perino was quoted saying "For now, we hope that this first demonstration will help establish the tool as something surgeons and proceduralists can use to reduce barriers to quickly and easily review and manipulate intraprocedural data, increasing efficiency and clinician enjoyment" [8].

"We know this demonstration will help establish the tool as something surgeons and proceduralists can use to reduce barriers to quickly and easily review and manipulate intraprocedural data, increasing efficiency and clinician engagement"

Where Contec Fits In

Contec is a global electronics manufacturer and systems integrator, with business segments in factory automation, transportation, medical and life sciences, robotics, security, energy, government, and much more.

With an extensive catalog of PCs including High-Performance Rackmount PCs or High-Performance tower PCs, can be utilized as an Edge-server to collect and process data from the AR glasses. The rackmount PCs SRV-2UA100D and MR44000 can serve as the base products and customized to meet specific application workloads.

FIGURE 3:

CONTEC MR44000

Solution-ePC MR44000 - Rackmount PC 5th Gen. Intel Xeon (Emerald Rapids) / 4x PCIe Gen.5(x16) / 3x PCIe Gen.5(x4)

FIGURE 4:

CONTEC SRV-2UA100D

2U Edge Server/ AMD Gen EPYC Processors (Genoa) / Three Storage Form Factors Supported

Both platforms support high core-count processors: the MR44000 scales to Intel Xeon Gold SKUs with up to 28 cores, while the SRV-2UA100D tops out at 96 cores on the AMD EPYC 9004 series. More importantly, both systems are designed to offload GPU-accelerated workloads. The AR backend layer relies on GPUs for the real-time pose tracking and 3D model inference (typically 10–50 ms per frame), and the rendering of the final composite image. The MR44000 offers four PCIe Gen5 x16 slots for discrete GPUs, while the SRV-2UA100D provides two Gen5 x16 slots plus an OCP 3.0 interface. Both power supplies are sized to support high-end GPUs without throttling. The MR44000 includes a 1300 W single supply, and the SRV-2UA100D provides 1600 W redundant supplies. For a given AR application, the choice between Intel and AMD, and the GPU configuration, should be driven by the inference model and the required throughput (concurrent headsets per server). Reference benchmarks are available from the GPU vendors; site-specific benchmarking with your actual workload is recommended before final procurement.

Characteristic	MR44000	SRV-2UA100D
Form factor	4U rack-mount, 430 × 528 × 174.8 mm	2U short-depth rack-mount, 438 × 558 × 87.2 mm
Processor family	Intel Xeon Scalable, 5th Gen (Emerald Rapids), Intel C741	AMD EPYC 9000 (Genoa)
Memory	DDR5-5600, 32 GB base configuration. Processor supports up to 4TB	DDR5-4800 RDIMM, 8 slots, up to 4 TB
PCIe generation	Gen5	Gen5
GPU-capable slots	4 × Gen5 x16 (plus 3 × Gen5 x4)	2 × Gen5 x16 full-height full-length, double-deck; supports up to 2 GPU cards
Additional expansion	—	1 × OCP 3.0 Gen5 x16, 2 × M.2 22110 Gen5 x4
Power supply	1300 W, single	1600 W 1+1 redundant; single-PSU and 1200 W DC input options
Network interfaces (onboard)	2 × 10GbE, 1 × 2.5GbE, 1 × 1GbE, 1 × management	1 × GbE dedicated to BMC management only; data NIC via OCP 3.0 slot
Storage	240 GB SATA SSD; SATA backplane	E1.S (8-bay), E3.S (8-bay), or U.2 (4-bay) options; 2 × M.2 NVMe boot
Remote management	AST2600 BMC	IPMI 2.0 BMC with iKVM, Redfish, AST2600
Operating temperature	0–35 °C	0–35 °C
Cooling and acoustics	Rack-class	Front-to-back, redundant hot-swap fans; not suitable for open office environments
Operating systems	Windows 10 IoT LTSC 2021, Windows 11, Windows Server 2019/2022, Ubuntu, RHEL 8.2+	Rocky, Debian, Fedora, openSUSE, Scientific, Ubuntu 20.04/22.04 LTS (64-bit)
Customization	BTO configurable	Configurable processor, memory, drive bay, and PSU options

Challenges and Considerations

Battery Life

Battery life of the user endpoint is important as it determines the usability of the glasses for a full shift. Skurowski et al. found that display-related factors, particularly screen brightness and scene content complexity, dominate power consumption on the HoloLens 2, and that consumption does not scale linearly with object count because of rendering overhead and occlusion. The choice of graphics engine also contributes measurably, with Unity and Unreal showing different energy profiles for equivalent scenes [8]. If during a shift the battery life is not sufficient there are options for charging banks to be attached to the headsets to extend the operating time of the glasses or improvements need to be made when establishing the Utility layerapplications.

Wireless Connectivity

Continuous wireless connectivity from the endpoint to the backend is needed to prevent any downtime. Choice of wireless technology is highly dependent on the application, particularly regarding size of the data packets, the latency required and security. Further consideration in choice of communication is size of coverage areas and obstacles, such as warehouses, or in areas where there is restricted access to these wireless communications, such as DoD or medical to comply with cybersecurity protocols.

Cameras and sensitive data

Due to the use of cameras and other sensors to gather data, specific rulings on the use of augmented reality glasses are being fleshed out in areas with sensitive data, HIPAA and DoD use are being investigated with some branches of DoD imposing a ban on the technology’s use. In early 2026 the U.S. Air Force updated its uniform regulations to prohibit smart glasses with photo, video, or artificial intelligence capabilities while in uniform, citing operational security; policies across the other military services vary [9]. These restrictions target consumer eyewear with covert recording capability, not managed assisted reality headsets under enterprise device management with controlled camera policy and audit logging, a distinction that will shape how AR is adopted in regulated and access-controlled environments.

Conclusion

Augmented reality is an exciting technology that has been around for decades and shows promising applications in many business verticals. AR glasses are an implementation of augmented reality that could see widespread use with their ease of maneuverability and longer duration of use could bring efficiency to physical tasks.

CONTEC POSITIONING

Contec can be a partner in the implementation as a supplier of the server to fill in the backend layer and the integration of the user-endpoint data in the server.

CONTEC AMERICAS

Technology for a better life

Augmented Reality (AR) | September 2026

Bibliography

[1] Hendricks, M. (n.d.). MDM vs. EMM vs. UEM: Key differences for device management. Rippling. https://www.rippling.com/blog/mdm-vs-emm

[2] PTC. (n.d.). Create a procedure workflow. In Get started with Vuforia Expert Capture. https://www.ptc.com/en/success-paths/get-started- vuforia-expert-capture/setup/create-procedure-workflow

[3] Schneider, M., Rambach, J., & Stricker, D. (2017). Augmented reality based on edge computing using the example of remote live support. 2017 IEEE International Conference on Industrial Technology (ICIT), 1277–1282. https://doi.org/10.1109/ICIT.2017.7915547

[4] Pasman, W., & Jansen, F. W. (2001). Distributed low-latency rendering for mobile AR. Proceedings of the IEEE and ACM International
Symposium on Augmented Reality (ISAR 2001), 107–113. https://doi.org/10.1109/ISAR.2001.970520

[5] Charyyev, B., Arslan, E., & Gunes, M. H. (2020). Latency comparison of cloud datacenters and edge servers. GLOBECOM 2020 – 2020 IEEE Global Communications Conference, 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9322406

[6] AR Insider. (2026, April 14). Case study: AR tightens Coca-Cola's bottling operations. https://arinsider.co/2026/04/14/case-study-ar-tightens-coca-colas-bottling-operations/

[7] Armitage, H. (2024, March 1). Stanford Medicine uses augmented reality for real-time data visualization during surgery. Stanford Medicine. https://med.stanford.edu/news/all-news/2024/02/augmented-reality-surgery.html

[8] Skurowski, P., Myszor, D., Paszkuta, M., Moroń, T., & Cyran, K. A. (2024). Energy demand in AR applications—A reverse ablation study of the HoloLens 2 device. Energies, 17(3), 553. https://doi.org/10.3390/en17030553

[9] DefenseScoop. (2026, February 6). Air Force bans Meta AI glasses over OPSEC concerns, while other services' policies vary. https://defensescoop.com/2026/02/06/meta-ai-glasses-us-military-services-policies/