A game industry and operating systems veteran that serves as the Chief Engineer for Verizon Labs. Using years of experience in building large scale game engines and graphics platforms I work to architect and set the technical vision for Verizon's future in the Edge, VR & AR services.
5G: Latency and Bandwidth

Traditional computing especially on mobile has left the heavy lifting to the hardware onboard devices. With larger screens and thinner form factors the difficulty to balance performance with quality and battery life has been a challenge for developers.

With the development of the 5G network latency has been reduced and bandwidth has been expanded to the point that many computing tasks can be offloaded.
Mobile Edge Computing

The development of Edge Computing has allowed for even quicker response times than from traditional public clouds. By placing computing devices and GPU’s right in the last leg of the network we allow for super quick round trip times and rapid options for offload and virtualization.

Edge compute is however an expensive resource with many unique characteristics so designing for the edge requires the utmost care.

Benefits Include:
• Processing complex workloads on limited hardware
• Reduce heat on an user endpoint
• Extend battery life on a constrained device
Edge Networks are primarily defined by offering compute closer to the user.

The principle features of the edge are low latency, high bandwidth offload to large amounts of compute power.

Currently, edge computing is expensive and we need to find use cases that are latency sensitive and compute heavy. These are typically scenarios that are GPU bound. Real world scenarios also require graceful fall back for when the edge is not available.
THE EDGE VS THE CLOUD

There are inherent differences in Edge vs Cloud Computing

Edge computing is defined by having lower latency and service guarantees that the public cloud traditionally does not have. There are however key challenges in edge based deployment.

Defining factors include:
1. Limited real estate
2. Limited power and cooling
3. Coverage and distribution
4. Cost
UNIQUELY DESIGNED FOR THE EDGE

Distribution and acceleration are enabled by sharing responsibilities with last mile compute resources.

Network based acceleration approaches are specially able to leverage edge computing for offload of specific tasks enabling next generation performance on embedded devices with limited capabilities. Envrmnt works to create edge services and tools specifically for OTT consumption on any network.

- Deliver resources intelligently
- Perform complex functions on the edge
- Deliver high performance experiences seamlessly
- Reach wider audiences
- Enable customers with richer experiences

Confidential and proprietary materials for authorized Verizon personnel and outside agencies only. Use, disclosure or distribution of this material is not permitted to any unauthorized persons or third parties except by written agreement.
**WHAT ARE 5G EDGE USE CASES?**

*Edge Use Cases are defined by their need for real time (sub 20 ms) reactions by the client/user*

<table>
<thead>
<tr>
<th>API</th>
<th>RTT</th>
<th>Bandwidth</th>
<th>Rationale</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rendering</td>
<td>15ms</td>
<td>10MB</td>
<td>Frame delay at 60hz</td>
<td>Gaming and Graphics</td>
</tr>
<tr>
<td>Lighting</td>
<td>15ms</td>
<td>10MB</td>
<td>Frame delay at 60hz</td>
<td>Gaming and AR</td>
</tr>
<tr>
<td>Spatial Audio</td>
<td>15ms</td>
<td>200MB</td>
<td>Audio Delay at 60hz</td>
<td>Gaming and Entertainment</td>
</tr>
<tr>
<td>Vision</td>
<td>10ms</td>
<td>10MB</td>
<td>Frame delay + Recognition</td>
<td>AR and Analytics</td>
</tr>
<tr>
<td>Transcoding</td>
<td>15ms</td>
<td>200MB</td>
<td>Frame delay + Audio Delay</td>
<td>Entertainment and communications</td>
</tr>
<tr>
<td>Drones and Autonomous Driving</td>
<td>Variable</td>
<td>1MB per device</td>
<td>Real time decision making</td>
<td>Autonomous Vehicles</td>
</tr>
<tr>
<td>IOT</td>
<td>N/A</td>
<td>1MB per device</td>
<td>Not Latency Sensitive</td>
<td>IOT</td>
</tr>
<tr>
<td>AI</td>
<td>N/A</td>
<td></td>
<td>Not Latency Sensitive</td>
<td>Applications</td>
</tr>
<tr>
<td>Security</td>
<td>N/A</td>
<td>1MB per device</td>
<td>Not Latency Sensitive</td>
<td>Applications</td>
</tr>
<tr>
<td>Physical Internet</td>
<td>N/A</td>
<td></td>
<td>Not Latency Sensitive</td>
<td>Context based applications</td>
</tr>
<tr>
<td>Physics</td>
<td>N/A</td>
<td>1MB per device</td>
<td>Not Latency Sensitive</td>
<td>Gaming and Simulation</td>
</tr>
</tbody>
</table>
VIRTUALIZATION VS. OFFLOADING

What are different ways to build services for the Edge?

Current paradigms for building edge based applications include:

- Traditional web services - Simply move your web services to the edge
- Virtualization – Create a virtual machine and a remote connection to your endpoint device
- Hybrid offloading – Setup an application data pipe and parallelize your compute functions
Create a virtual machine. Inputs go to the edge and video comes back!

This is the easiest way to get high end returns on a limited device.

Pro’s
- User endpoint becomes a thin client
- Software will function seamlessly and without customization
- Easy to setup and maintain

Cons
- Very expensive to scale
- Inherent lag for users
- No fallback for when the connection is not available
HYBRID OFFLOADING

Split your tasks for fun and profit!

We regard hybrid offloading as a way to structure your application such that your client and edge application are utilized together. We establish the application data pipe with a webservice call and setup socket directly to an edge appliance.

We can also optimize the edge application to process handle many clients that are using a shared dataset/context.

Pros
• Maintain maximum framerate on a user device
• Highly scalable
• Graceful fallback when connection is not available

Cons
• Requires custom application design
THE PATH OF AN UPDATE

Virtualization – Best Case 3 frames, Worst Case 10+ frames

Hybrid Offloading – Best Case 3 frames  (Updates and Draws can occur Asynchronously)
Designing for Mobile
Edge Offloading
NETWORK ACCELERATED BY ENVRMNT

AR and VR is uniquely suited to leverage Low Latency Computing

The boundaries of compute and latency on AR and VR can quickly fade away when leveraging next generation networks.

In these use cases we begin to focus on offloading and amortizing the heaviest tasks to nodes that are only a hop away

- Low latency, high bandwidth networks allow offloading of tasks that were local only
- Data and scene formats specifically designed for cache and reuse on the edge
- Utilize vast amounts of compute and VRAM to push lightweight devices further
- Reduce costs for users and providers
XR is positioned to get a sizable boost from Edge Computing

Envrmt set out to find new use cases that were not possible without edge compute. Specifically, XR use cases where clients have limited hardware and extreme latency requirements.

Envrmt Edge Solutions are purpose built edge services that can be deployed over the top to enhance the XR experience for customers.

Edge Solutions include consumable API’s and SDK’s designed for edge compute including:

1. Real Time Computer Vision for XR
2. XR Rendering and Lighting
3. Spatial Rendering for Immersive Audio
4. Hybrid Rendering for Graphics
ENVRMNT HYBRID RENDERING FOR GRAPHICS

Provide a graphical boost to games and simulations

Hybrid graphics rendering offloads the most computationally heavy rendering tasks and allows them to be served by an edge-based renderer.

By offloading lighting, shadows, and PBR materials on the edge and blending them with client-side compositing and rendering, a significant graphics boost is achieved without sacrificing motion to photon latency and failing gracefully when your edge connection is compromised.

Designed to scale on an edge system, Envrmnt Hybrid Graphics Rendering Services is a one-of-a-kind solution that allows for true GPU slicing and non-linear scaling for users.
Project Helios is the first game benchmark designed to stress what is possible on the edge.

Designed specifically to stress mobile VR to its limits project Helios is a space shooter that takes a user on a high speed dynamic mission through the insides of a space station.

Filled with reflective surfaces and lights and dynamic shadows Helios is built to demonstrate what is possible with and without an edge rendering solution.

Maintaining zero lag on a mobile VR headset while firing dynamic lights in a corridor Helios stresses what is possible when graphics compute is amortized.

Low motion to photon latency.

Optimized for mobile with graceful fallback

Reduce costs for users and providers by scaling intelligently.
**ENVRMNT XR RENDERING & LIGHTING**

*AR and XR use cases come alive when real world lighting and scene data is used to render and blend the augmented with the actual world.*

Using computer vision and hybrid rendering Envrmnt XR rendering uses real world scene data to ensure augmentations are high fidelity and seamless. Utilizing low latency connections to our edge service we allow for realistic rendering that matches the world around it.

Relight your augments and render them with stunning reflections, physical based rendering and shadows.
XR Lighting Architecture

Video Capture Client
- Video Stitching and Processing

360° Camera
- On-site Video Capture Device

AR glasses / mobile device
- 3D Model rendering

Input video frames

XR Lighting container process
- MEC server

- Reflection Mapping
- Light Source Detection
- Lightmap & Shadow Rendering
- Radiance/ Irradiance Mapping

Lightmap and Material data

Confidential and proprietary materials for authorized Verizon personnel and outside agencies only. Use, disclosure or distribution of this material is not permitted to any unauthorized persons or third parties except by written agreement.
ENVRMNT EDGE VISION: REAL-TIME COMPUTER VISION FOR XR

Computer Vision has traditionally been bound by VRAM and GPU compute.

Computer vision is a crucial part of augmented and extended reality. When used for identification, tracking/pose estimation and analytics devices have been restricted by on-device GPU’s or time it takes for backend servers to crunch algorithms.

Envrmnt Edge Vision is a solution that breaks the boundaries of computer vision to make it responsive in real time. Using highly parallel GPU compute Envrmnt can recognize and augment large quantities of targets in real time.
Proposed Architecture

Input video frames

Video Capture → Vision Container Process

Object tracking → Vision Container Process

AR rendering → Vision Container Process

Vision Container Process → MEC server

MEC server → AWS GPU Training cluster

AWS GPU Training cluster → Trained Data

Trained Data → Feature Matching (GPU)

Feature Matching (GPU) → Feature Based Recognition (GPU)

Feature Based Recognition (GPU) → Deep Learning - Object Classification (GPU)

Deep Learning - Object Classification (GPU) → AR Glasses / Mobile Device

AR Glasses / Mobile Device → Dedicated object And AR data

Dedicated object And AR data → AR rendering

AR rendering → AR Glasses / Mobile Device

Confidential and proprietary materials for authorized Verizon personnel and outside agencies only. Use, disclosure or distribution of this material is not permitted to any unauthorized persons or third parties except by written agreement.
Audio is often overlooked, but is one of the most crucial elements to immersion in XR.

Audio processing is incorrectly overlooked when it comes to XR experiences. Proper XR audio is spatialized and painstakingly designed to reflect head rotation, attenuation, occlusion, and material properties of a room. Even voice and ambient effects are greatly benefited by edge offloading.

Envrmnt Spatial Rendering is designed to take expensive room modeling and audio calculations and push them into the cloud for highly immersive audio rendering that doesn’t impact your performance or battery life.
Designing The GPU Edge
WHAT DOES THE STACK LOOK LIKE?

Cloud Hosted Layer

Web Portal

Admin Portal

User Management

Dev Portal

Rancher

Deployment and Management

Harbor

Docker Registry

Federation API

Analytics API

Test Bed + Edge

Intelligent Routing

CPU Masters

GPU Node

CPU Masters

GPU Node

CPU Masters

GPU Node

CPU Masters

GPU Node
WHAT DOES THE STACK LOOK LIKE?

Proposed Edge Services Stack

Public Cloud (Amazon/Azure)
- API Gateway
- MEC Provisioning Service
- MEC SDK Services Backend
- Real-Time Cluster Management
- Misc Cloud Services
- Analytics
- Data Store

Edge Compute Cluster
- Orchestration/federation Layer
- Kubernetes Master
- Edge Services (Docker Containers)

User Endpoint Devices
- Handsets
- HMD
- IoT
- Automotive
- Misc
...AND THE CALL FLOW?

User Endpoint → Public Cloud Platform → Orchestration Layer → Kubernetes → Docker → Edge Service

- Application Requests
- Edge Service
- Authentication
- Service init
- Provisioning and Utilization data
- Get Container Instance
- Cluster load analysis for Smart Routing
- Container Provisioning and Initialization
- Application Data Pipe
- Service Provisioning And Initialization
- Service Invocation
ENVRMNT EDGE SOLUTIONS

Over the top solutions built specifically for real world systems.

Envrmnt Edge Solutions offer deployable common sense services to enhance existing workflows and help users see the power of the Edge Network.

Available to anyone as network agnostic API's and SDK's

• Real Time Computer Vision for XR
• XR Rendering and Lighting
• Spatial Rendering for Immersive Audio
• Hybrid Rendering for Graphics

Confidential and proprietary materials for authorized Verizon personnel and outside agencies only. Use, disclosure or distribution of this material is not permitted to any unauthorized persons or third parties except by written agreement.
Thank You.