DEVELOPER WEEK

DigitalOcean

Navigating the Edge-Cloud Bridge

Building Resource Optimized Voice IoT/Edge Assistants with LLMs

David vonThenen

@davidvonthenen

David vonThenen

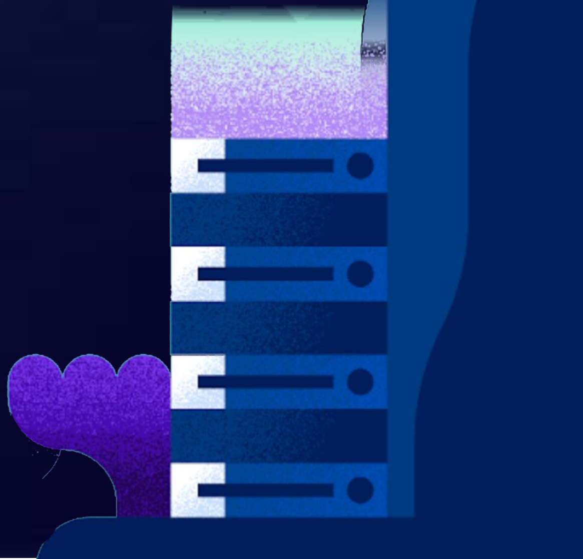
- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Cloudy, cloudy, cloudy…
- There is storage for that!

@davidvonthenen

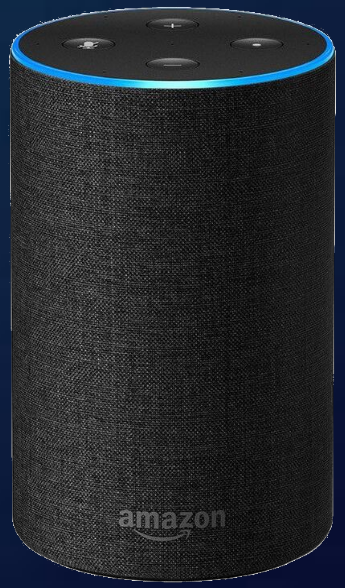
Agenda

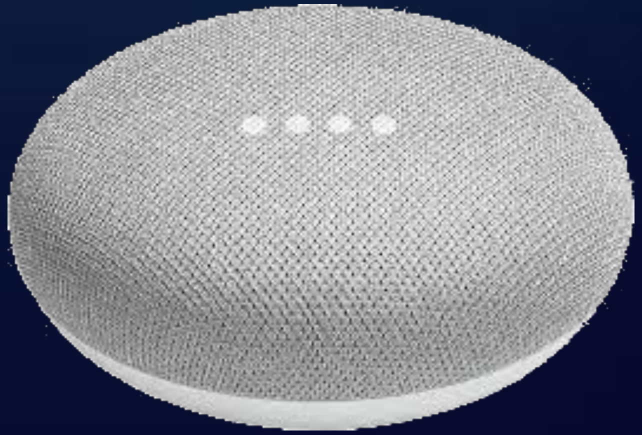
- Voice IoT/Edge Collaborators
- IoT/Edge Architectures & Consideration
    - Demos, Demos, Demos
- Q&A

Voice IoT/Edge Assistants



IoT/Edge Assistants





Assistant Voice Components

STEP 1
Automatic Speech Recognition

STEP 2
Natural Language Processing

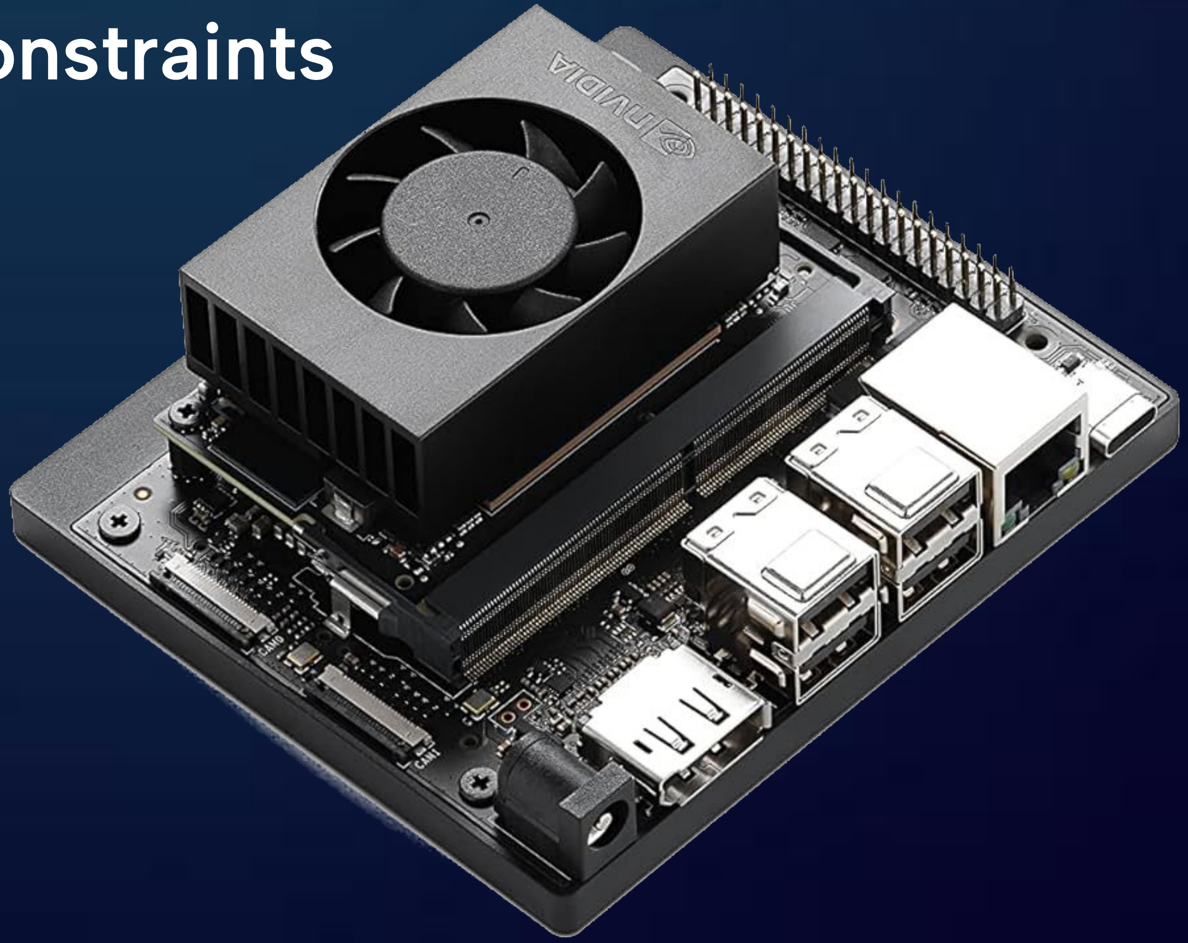
STEP 3
Desired business logic via hooks

STEP 4
Text to Speech

Architectures & Considerations

IoT/Edge Limitations

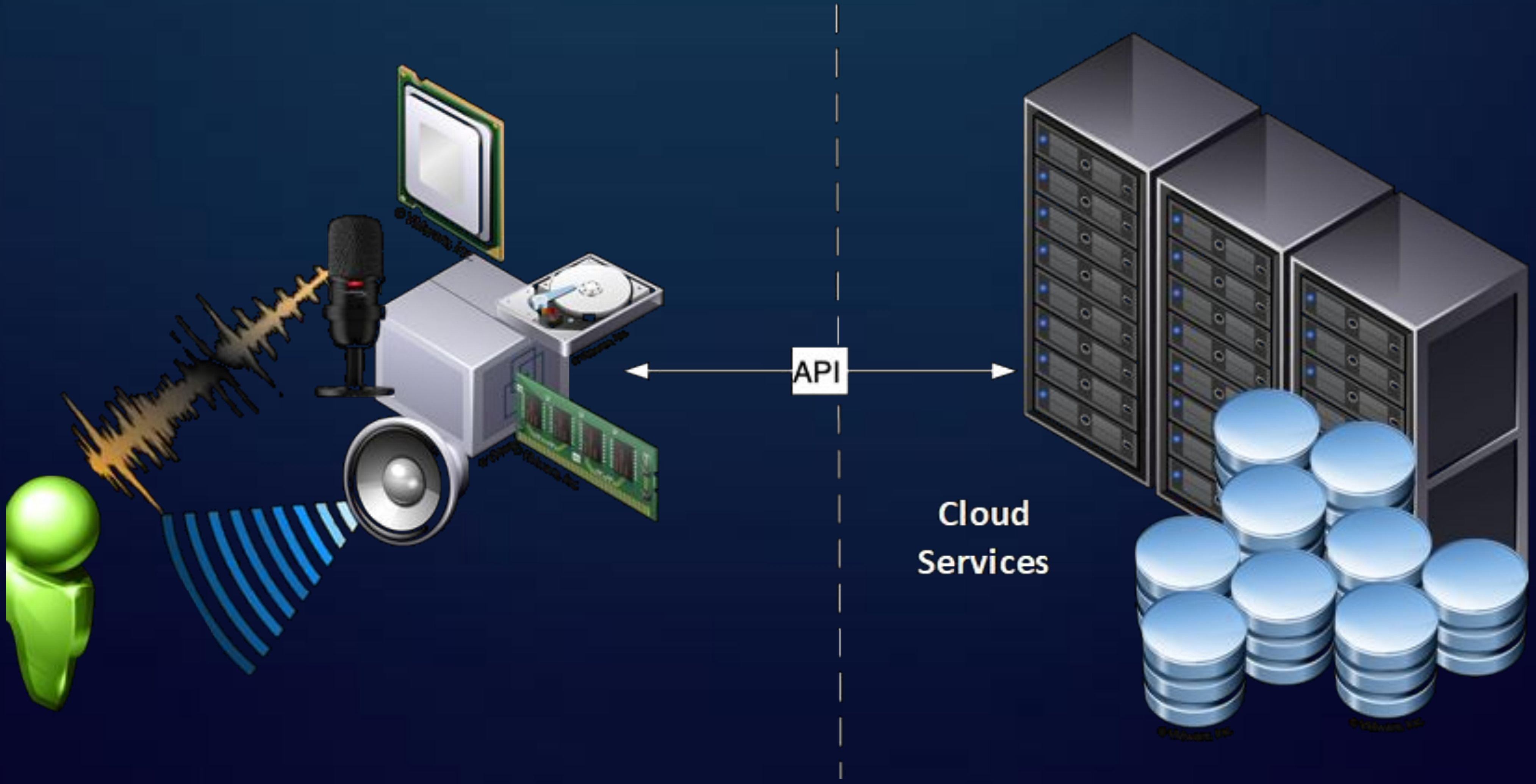
- Constraints, Constraints, Constraints
  - CPU –> Getting Better
  - Finite Memory
  - Is a GPU Available?
  - Have Enough Power?
- What Architecture to Use?
  - Device Requirements
  - Use Case Driven
  - Cost 🤑🤑🤑



Goals, Privacy, IP, etc

- What Are The Goals and Use Cases
  - Problem Your Device Solves?
  - Responsiveness Requirements
  - Security and Privacy Concerns
    - IP, Confidentiality, etc
  - Modalities Of Your Device
- Impact Analysis On...
  - Device Hardware
  - What Needs to Offloaded
  - User Experience

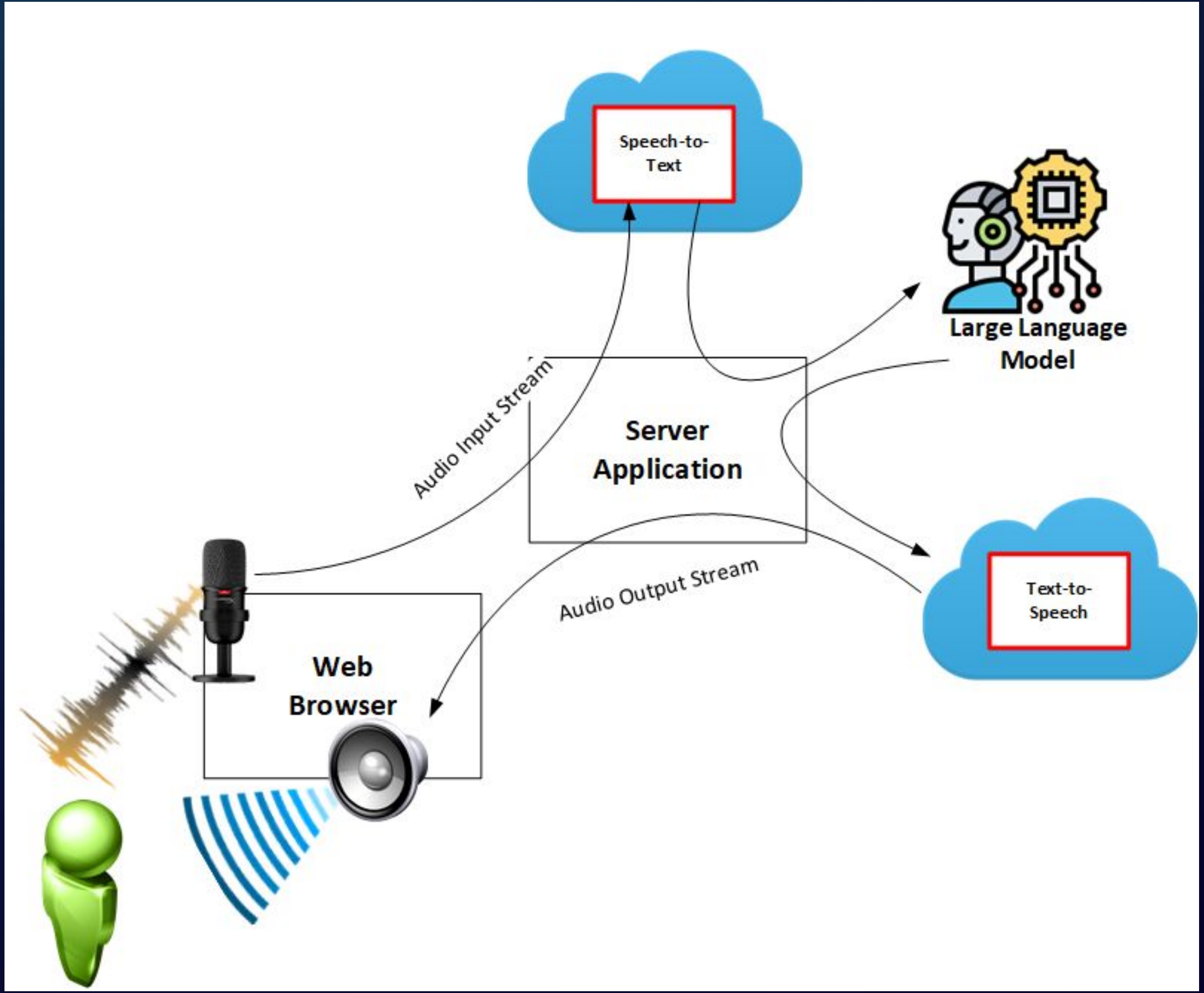
Lightweight Edge Device



Lightweight Edge Demo

Architecture:

- STT + TTS Processed in Cloud
- LLM/RAG in Cloud
    - "Result" on Server
- Leverage Network Connection to Offload



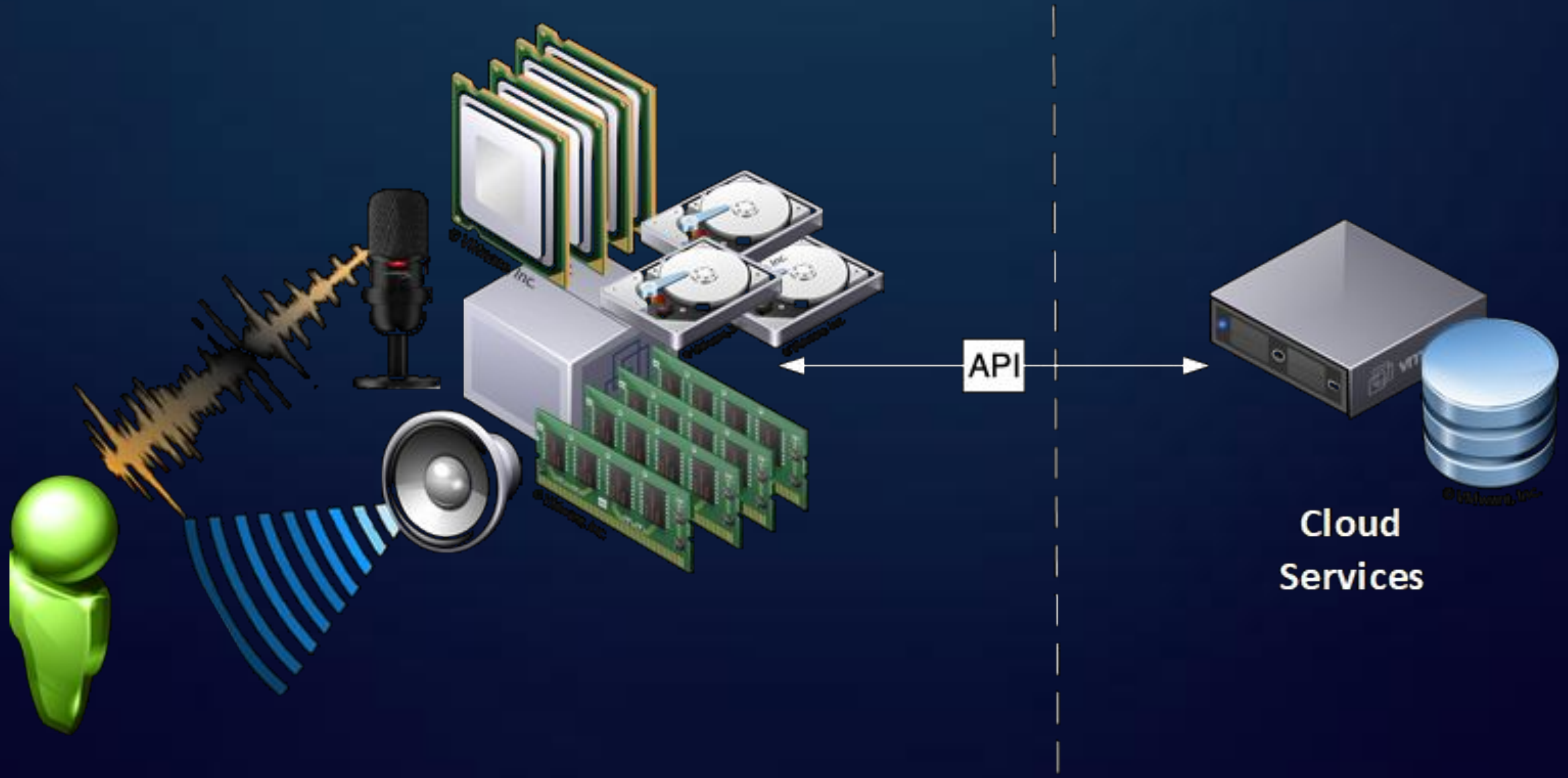
Demo: Lightweight Device

Click the button and start talking to your Medical Agent.

Demo: Lightweight Device

https://youtu.be/u2EhDfvzixs

High Performance Edge Device



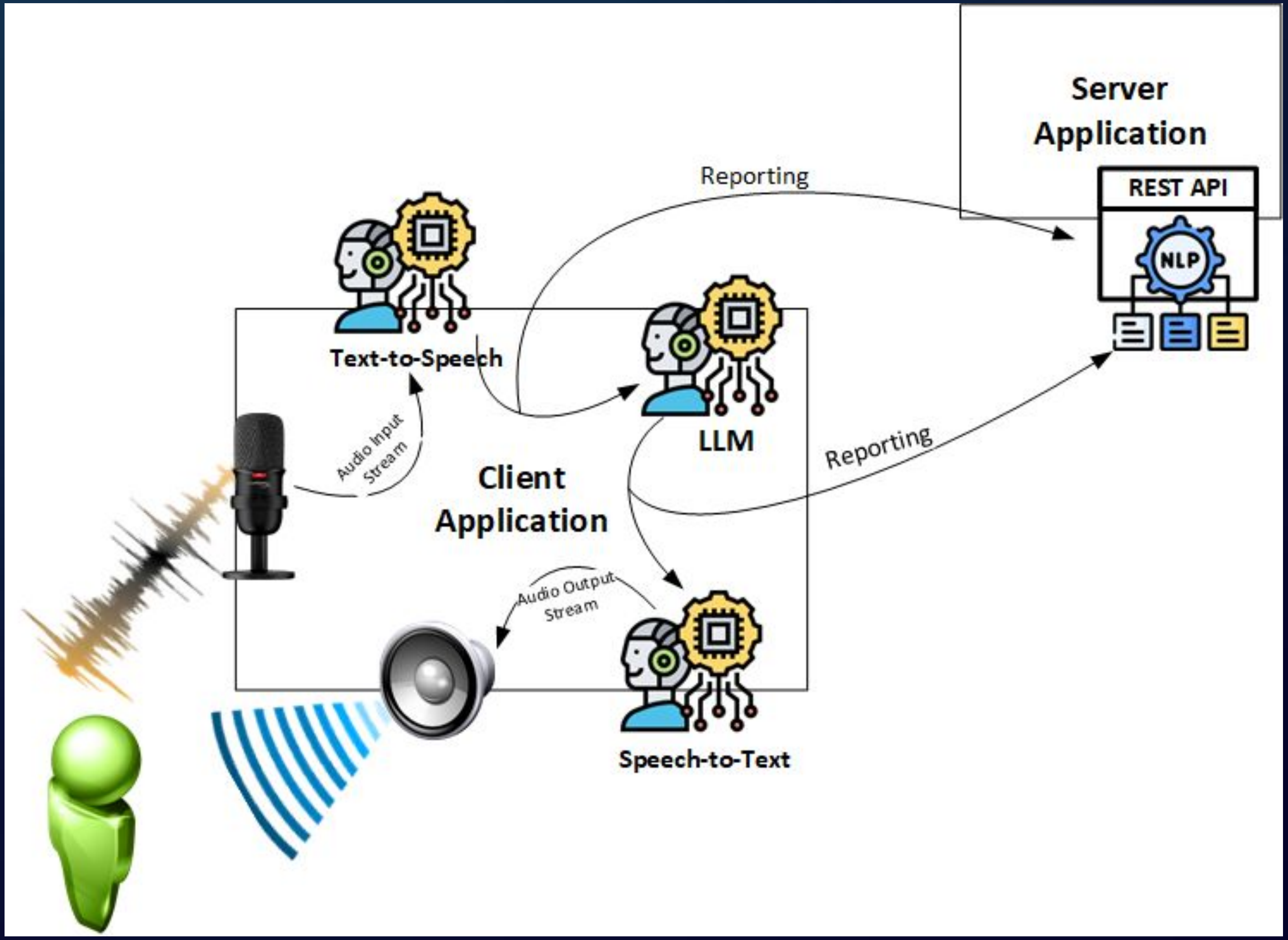
High Performance Edge Demo

GPU/NPU on Device!

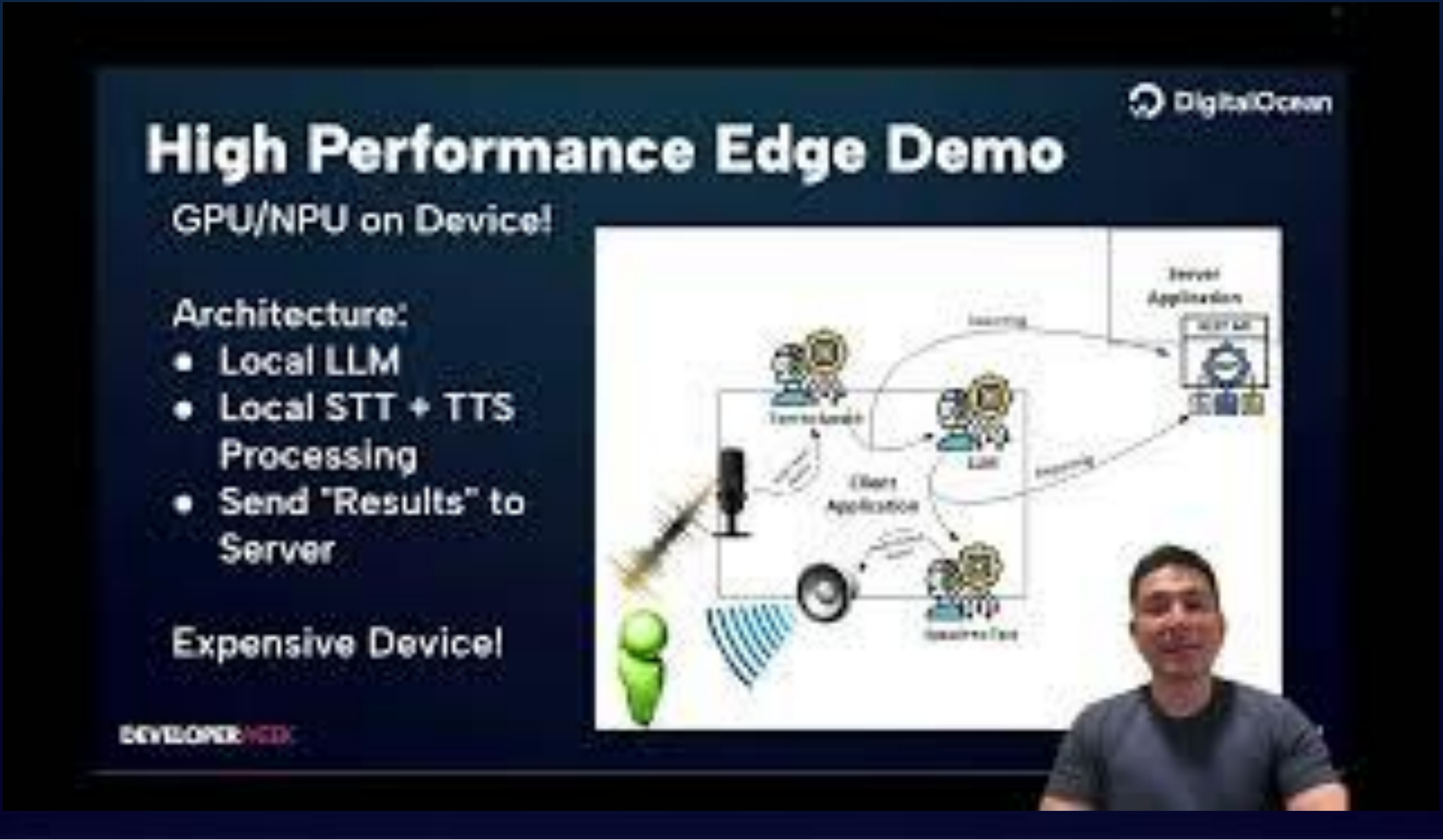
Architecture:

- Local LLM
- Local STT + TTS Processing
- Send "Results" to Server

Expensive Device!



Demo: Performance Device



DigitalOcean

High Performance Edge Demo

GPU/NPU on Device!

Architecture:

- Local LLM
- Local STT + TTS Processing
- Send "Results" to Server

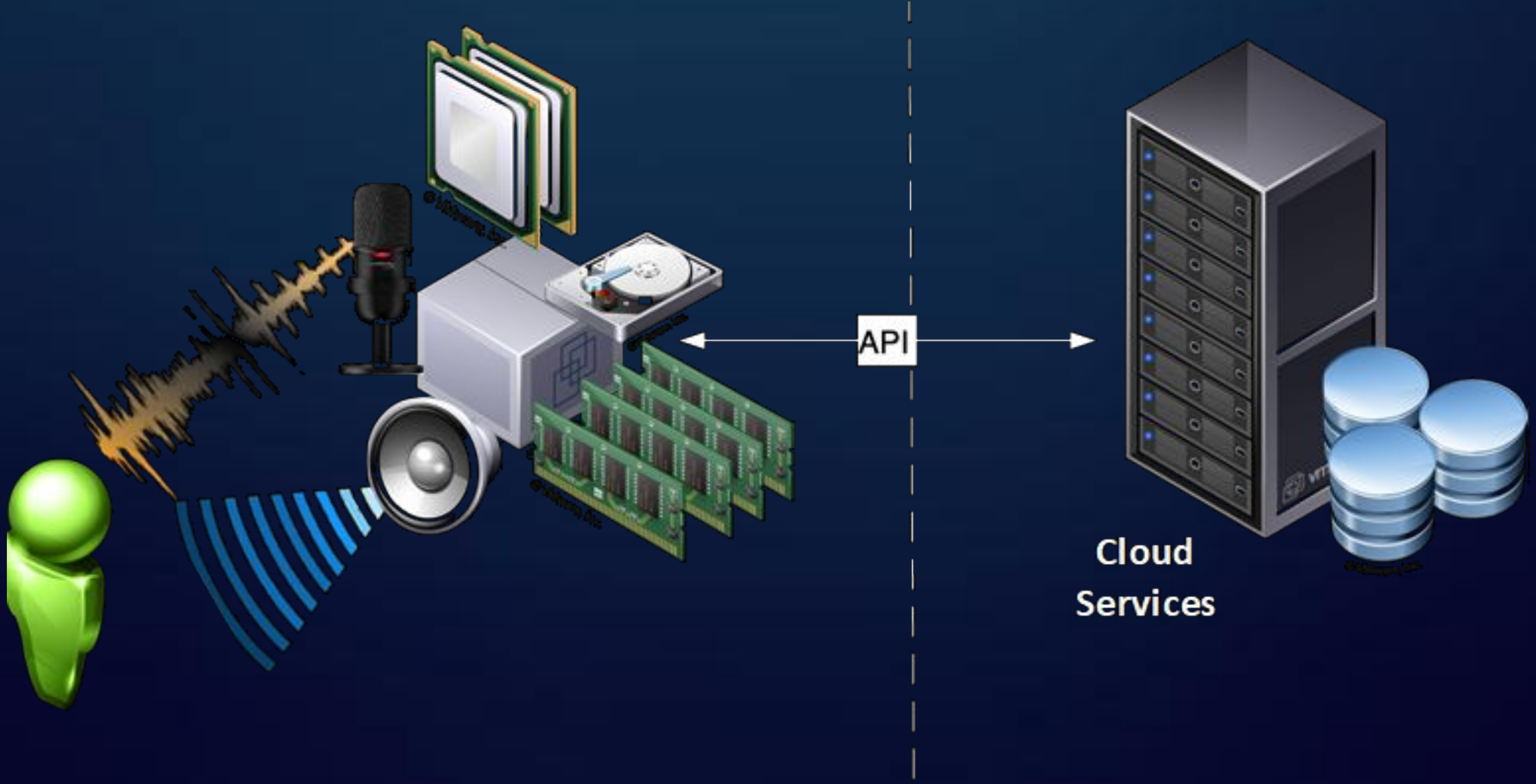
Expensive Device!

DEVELOPERWEEK

Demo: Performance Device

https://youtu.be/HWiEUkLHmYU

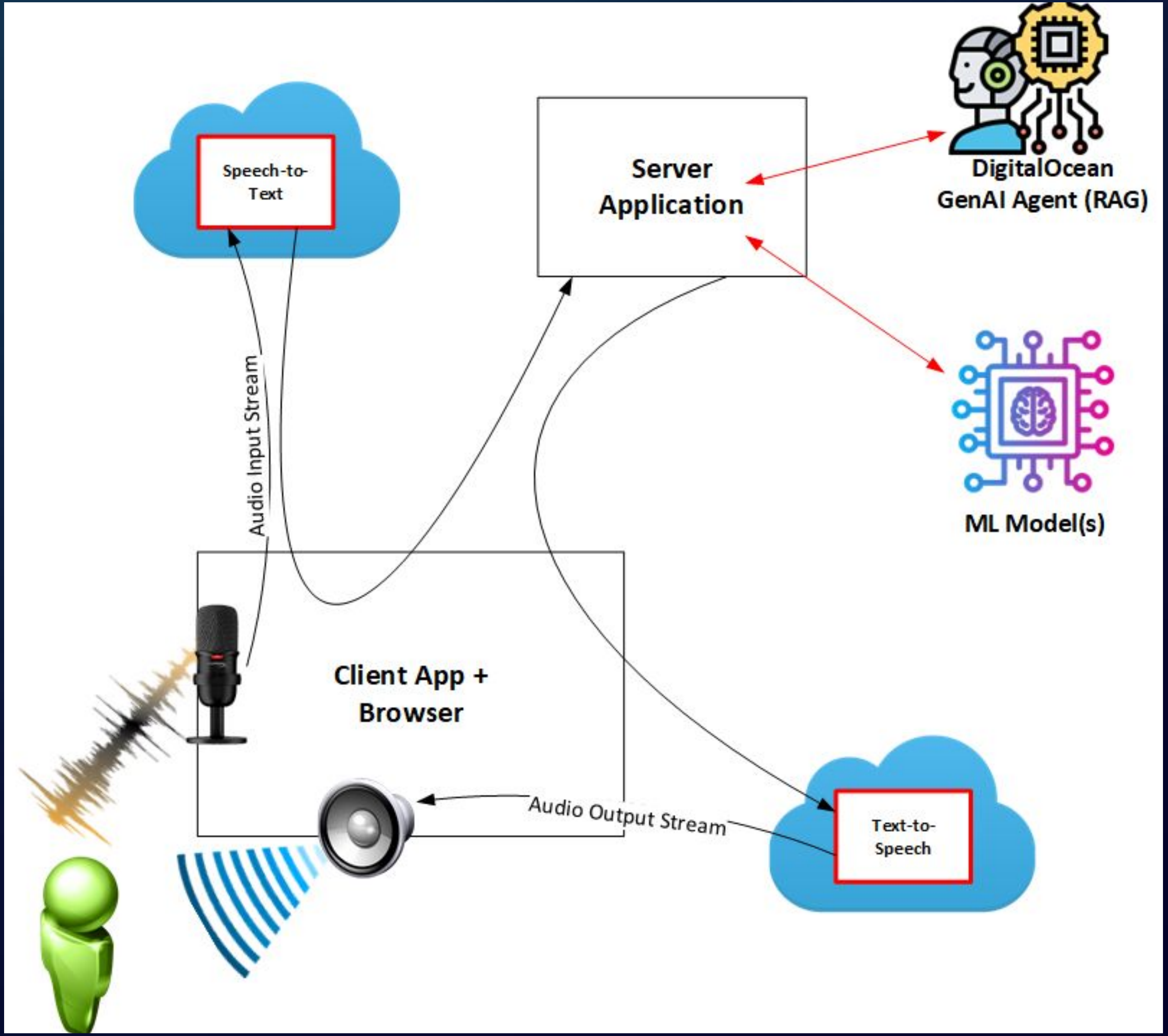
Balanced Device By...



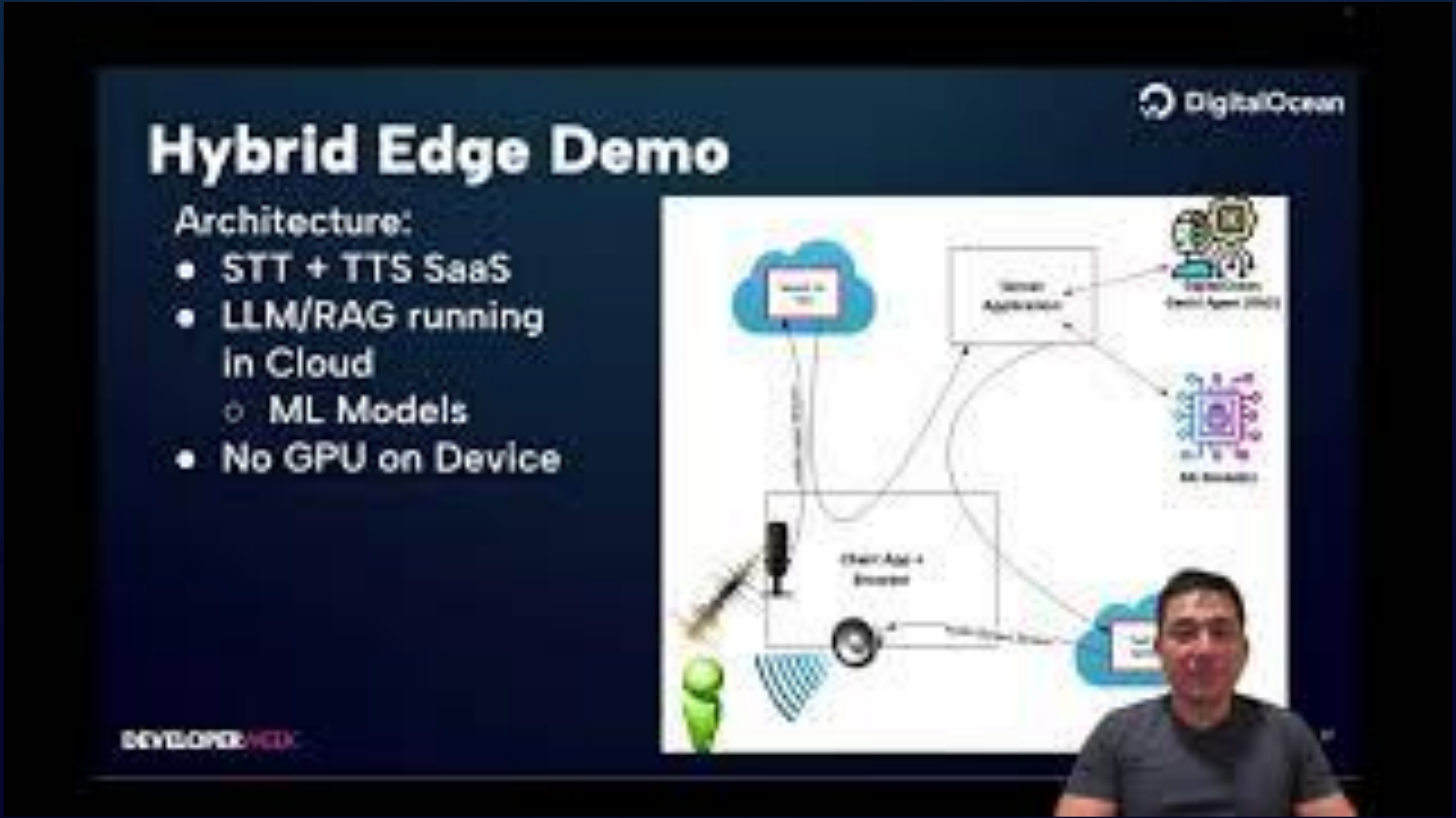
Hybrid Edge Demo

Architecture:

- STT + TTS SaaS
- LLM/RAG running in Cloud
  - ML Models
- No GPU on Device



Demo: Balanced Device



DigitalOcean

Hybrid Edge Demo

Architecture:

- STT + TTS SaaS
- LLM/RAG running in Cloud
  - ML Models
- No GPU on Device

Demo: Balanced Device

https://youtu.be/STXEnYMxtVY

Resources

[CLICK HERE] for All Materials and Demos in this Session

DigitalOcean GenAI Platform

- https://docs.digitalocean.com/products/genai-platform/

Open Source:

- NVIDIA NeMo – https://github.com/NVIDIA/NeMo-Run
- Kokoro Onnx – https://github.com/thewh1teagle/kokoro-onnx

Voice Platforms:

- Deepgram STT & TTS – https://playground.deepgram.com
- ElevenLabs – https://elevenlabs.io/

Agents Unleashed

Build AI-Powered Apps in Hours with the GenAI Platform

Feb 21 @ 9am Pacific

Sign Up Here: https://bit.ly/genai-workshop-signup

David vonThenen

Senior AI/ML Engineer

@davidvonthenen

Thank You!

David vonThenen

@davidvonthenen