modzy®

Robust and Resilient Edge Cloud Inference for LLMs: Empowering Intelligent Edge Applications

Presenter: Brad Munday

Meet your host

Brad Munday

Head of ML Engineering

Former tech consultant, quant, ML engineer, with extensive ML open-source contributions

modzy®

LLMs at the edge in the real world

Qualcomm Works with Meta to Enable On-device AI Applications Using Llama 2

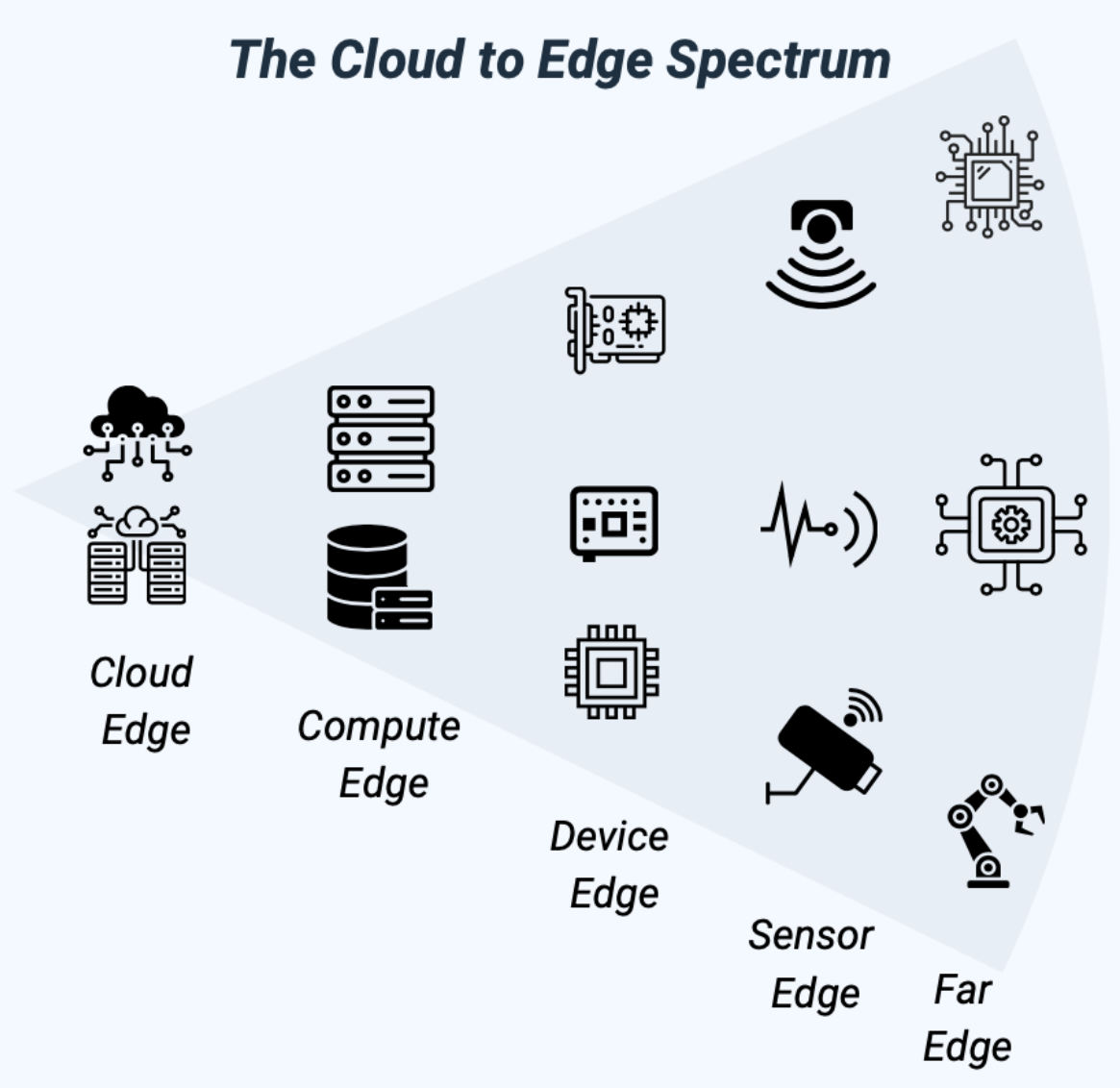
JUL 18, 2023 | SAN DIEGO | Qualcomm products mentioned within this press release are offered by Qualcomm Technologies, Inc. and/or its subsidiaries.

Highlights:

- Qualcomm is scheduled to make available Llama 2-based AI implementations on flagship smartphones and PCs starting from 2024

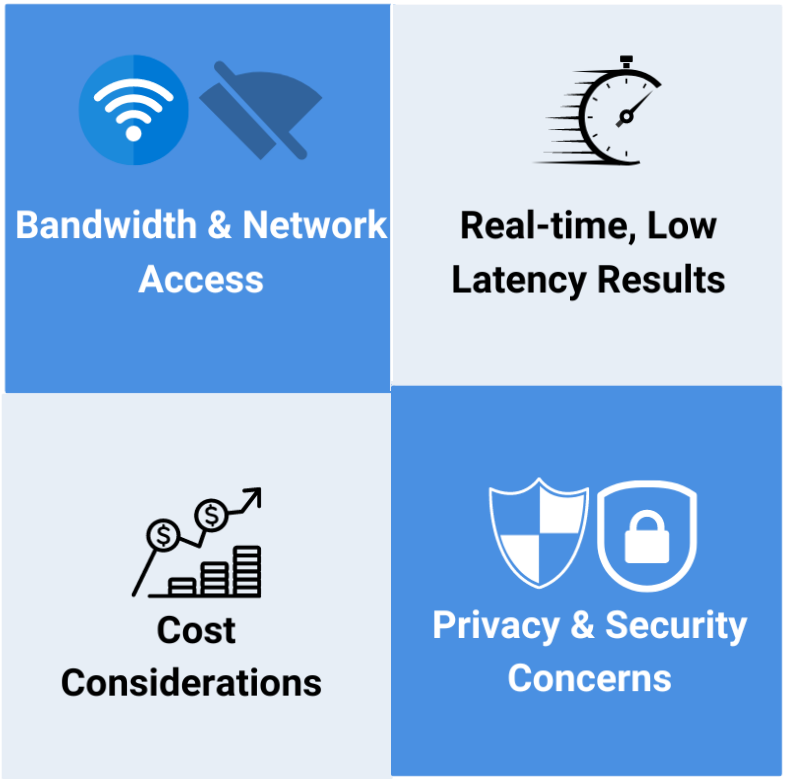
What constitutes “the edge?”

- ***Cloud Edge*** – offers computing capabilities that you would find in a cloud service provider, e.g., MEC.
- ***Compute Edge*** – functions as a localized, micro-data center that includes a limited range of resources and services you would find in the cloud, e.g., an edge line server racked or placed near or close to other devices or sensors.
- ***Device Edge*** – much smaller compute and processing capabilities, e.g., NVIDIA Jetson modules, Raspberry Pi, Intel NUC.
- ***Sensor Edge*** – comprises IoT sensors and devices that gather data, e.g., camera, and interact directly with the cloud, compute, or device.
- ***Far Edge*** – e.g., a microprocessor on board a robotic arm.



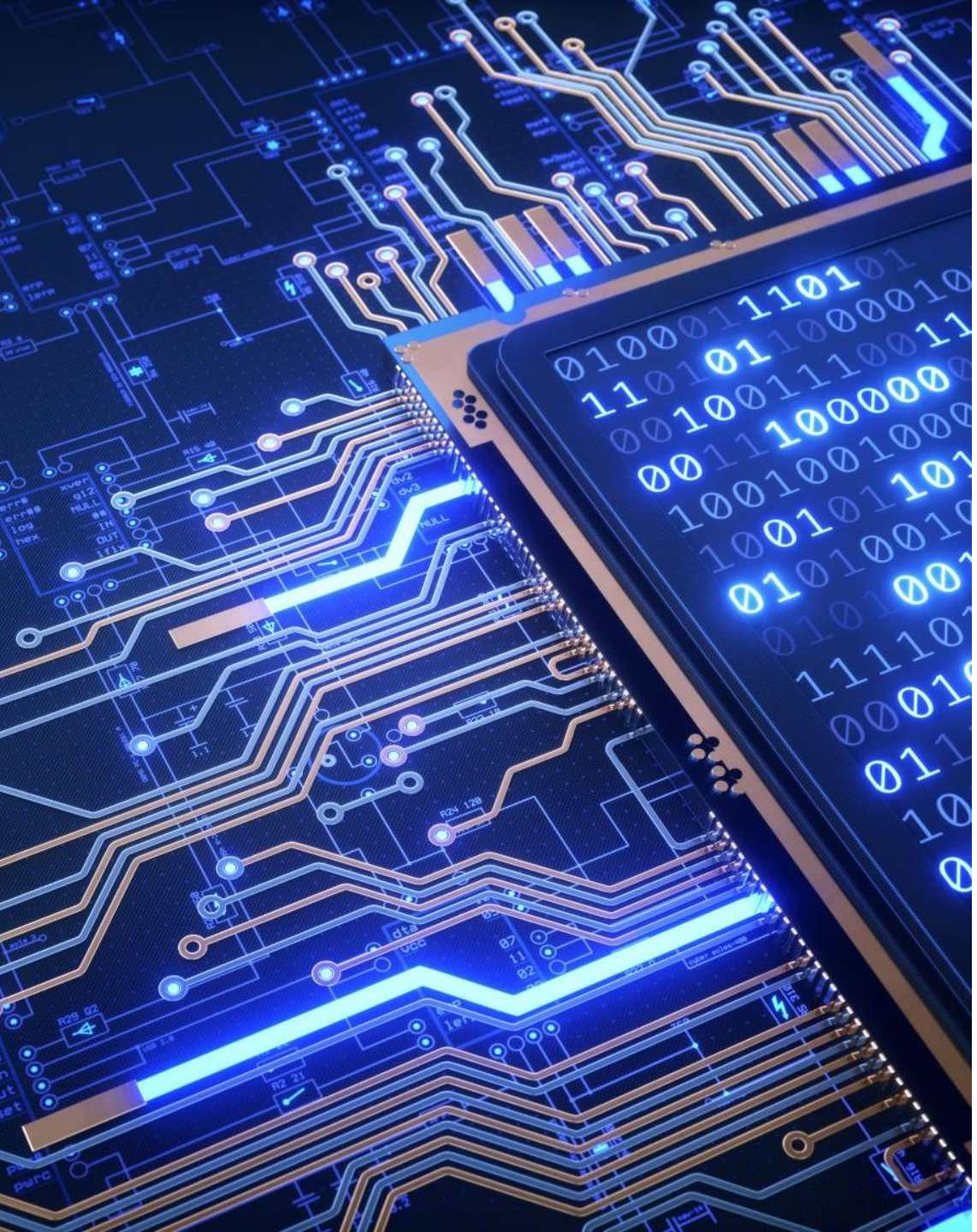
modzy®

Factors driving ML to the edge



modzy®

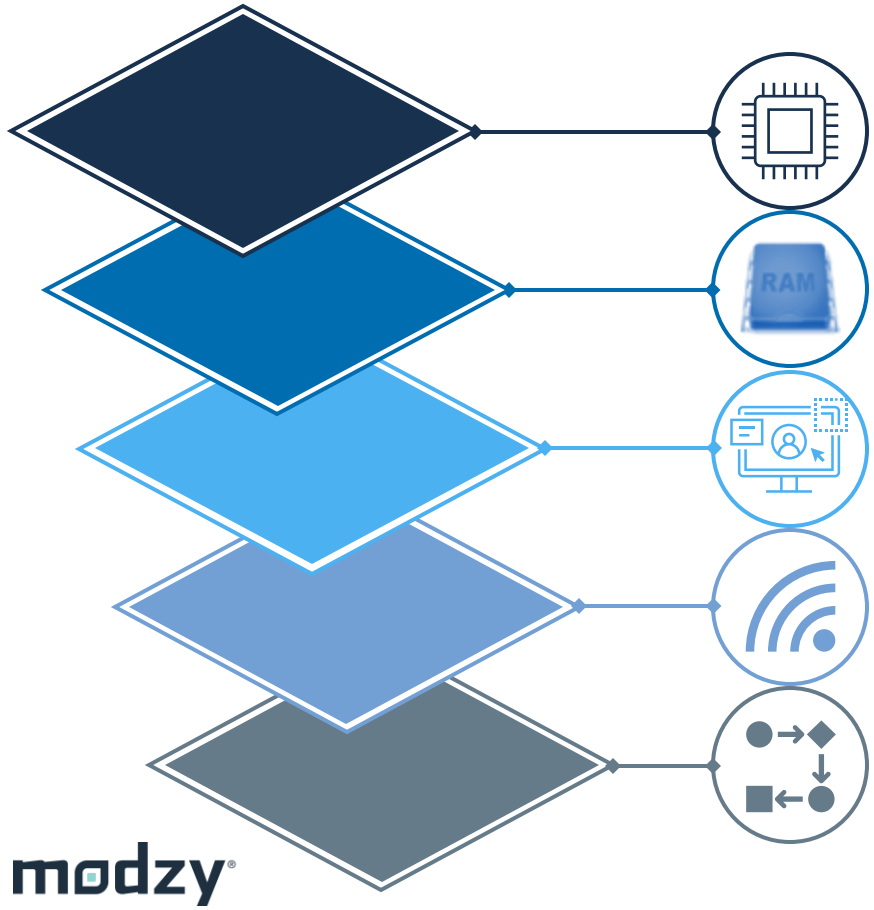
*According to the Tirias Research, GenAI Forecast and TCO Model, if **20% of GenAI processing workload** could be offloaded from data centers by 2028 using on-device and hybrid processing, then the cost of data center infrastructure and operating cost for GenAI processing would decline by **$15 billion**.*



Techniques to consider for LLMs at the edge

- Optimization
- Quantization
- Pruning
- Knowledge distillation
- Model specialization
- Inference accelerators

Considerations for building edge AI systems



Device Setup
OS installation, system-level dependency configuration, GPU access

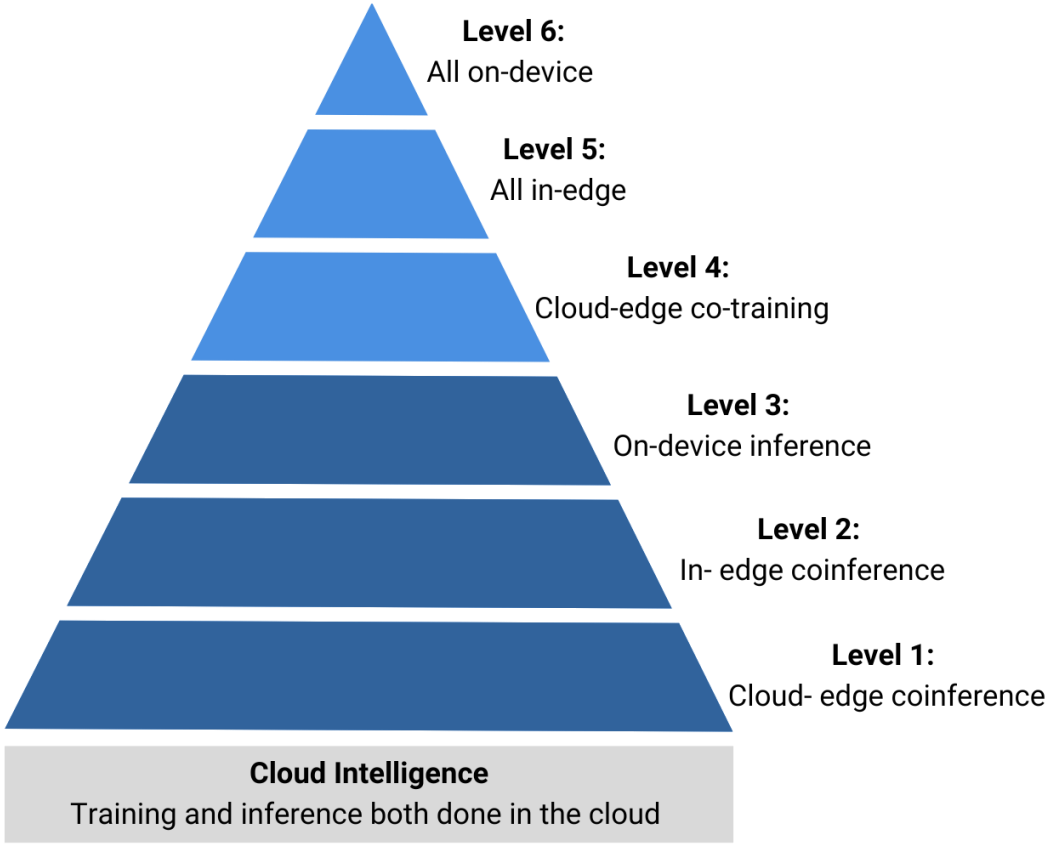
Resource Constraints
RAM allocation, number of CPU cores, shared across all app processes

External Software
Custom data connections, monitoring or logging tools

Security
Device and application security, application accessibility, networking

Model Requirements
Programming language(s), model framework dependencies

Choose a paradigm that supports your use case

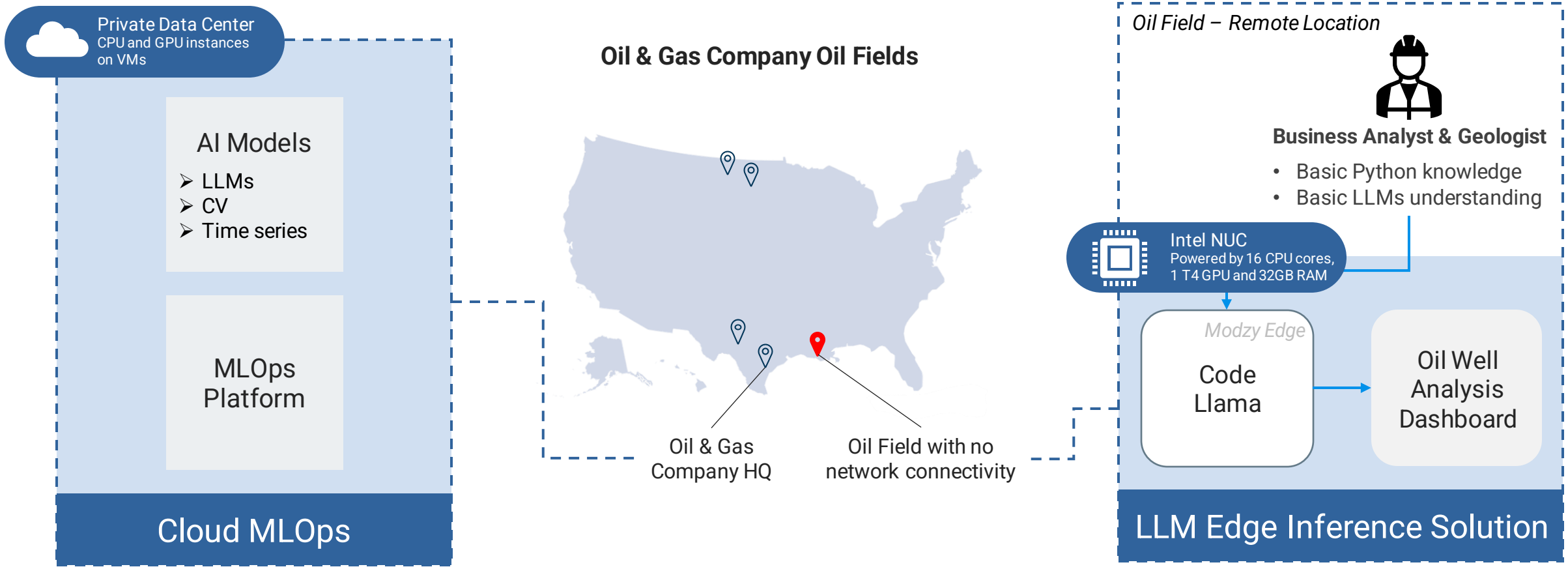


Paradigm	Overview
Cloud Intelligence	Both model training and inference are done in the cloud.
Level 1- Cloud-Edge Coinference and Cloud Training	Models are trained in the cloud, but run both at the edge and in the cloud; data is partially offloaded to the cloud.
Level 2- In-Edge Coinference & Cloud Training	Model is trained in the cloud, but inference is carried out at the network edge; data can be fully or partially offloaded to edge nodes or nearby devices.
Level 3- Device Inference and Cloud Training	Model is trained in the cloud, but inference is performed locally, on the device and no data is offloaded.
Level 4- Cloud-Edge Cotraining and Inference	Training and inference is done in cooperation between edge-cloud.
Level 5- All In-Edge	Training and inference is done in-edge.
Level 6- All On-Device	Training and inference is done locally on the device.

Source: *Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing*, IEEE.

Demo – LLM at the Edge

Demo Design


Private Data Center
CPU and GPU instances on VMs
AI Models
LLMs
CV
Time series
MLOps Platform
Cloud MLOps
Oil & Gas Company Oil Fields
Oil & Gas Company HQ
Oil Field with no network connectivity
Oil Field – Remote Location
Business Analyst & Geologist
Basic Python knowledge
Basic LLMs understanding
Intel NUC
Powered by 16 CPU cores, 1 T4 GPU and 32GB RAM
Modzy Edge
Code Llama
Oil Well Analysis Dashboard
LLM Edge Inference Solution

modzy®

Bradley.munday@modzy.com

@getModzy

getModzy

discord.gg

Brad Munday

Thanks for joining our webinar!

Scan the QR code to join our Discord server for more resources.