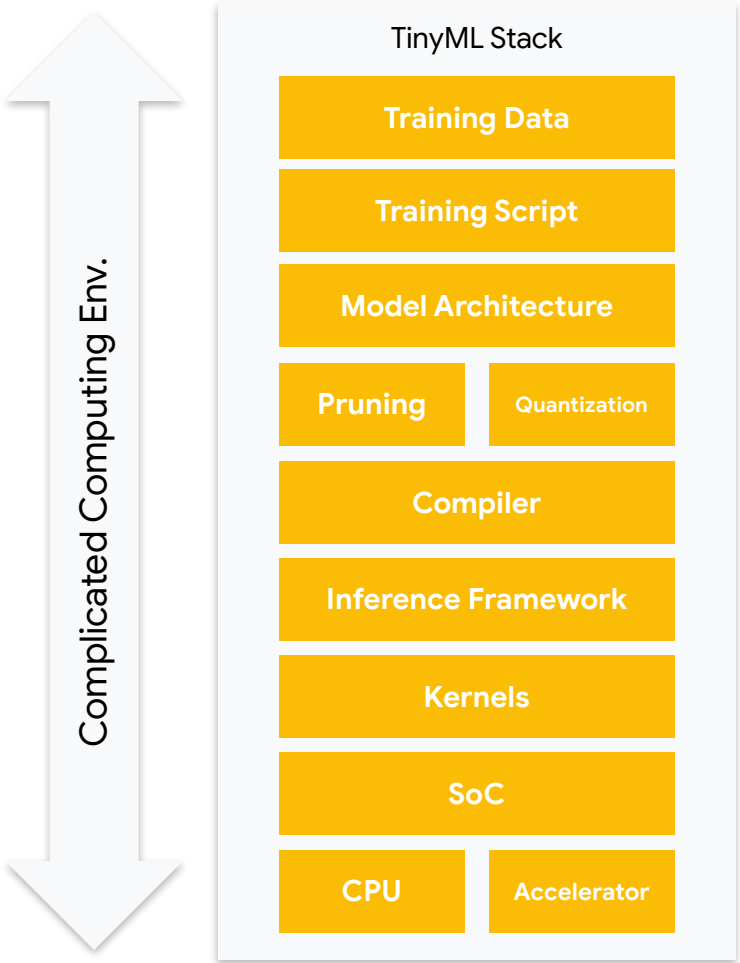
TinyMLaaS (Part 1): An Introduction

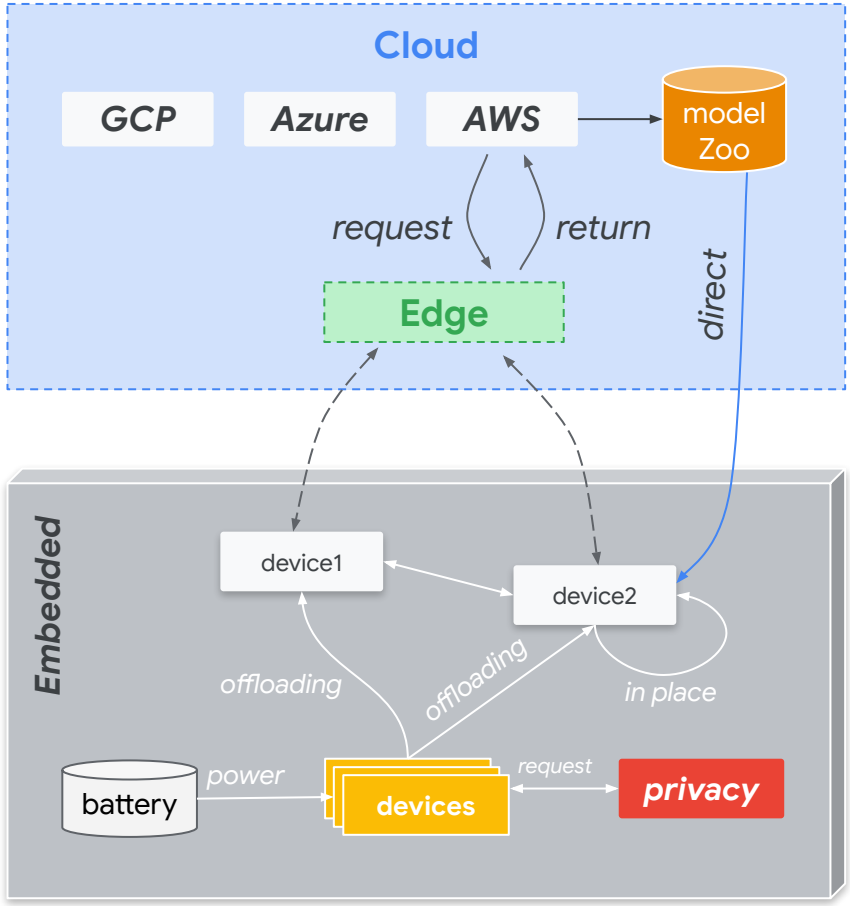


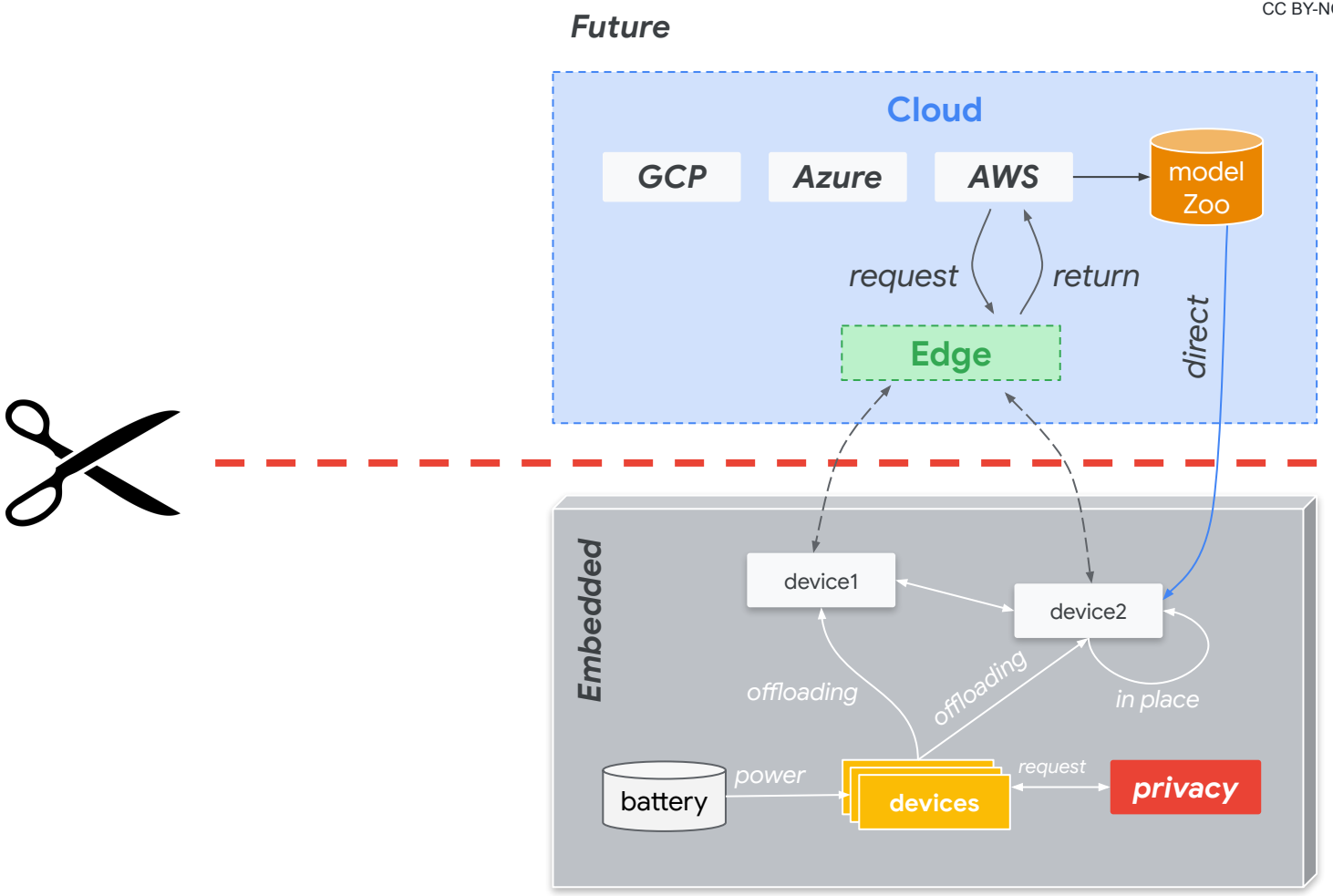
Complicated Computing Env.

TinyML Stack

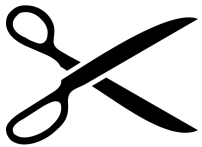
- Training Data
- Training Script
- Model Architecture
- Pruning
- Quantization
- Compiler
- Inference Framework
- Kernels
- SoC
- CPU
- Accelerator

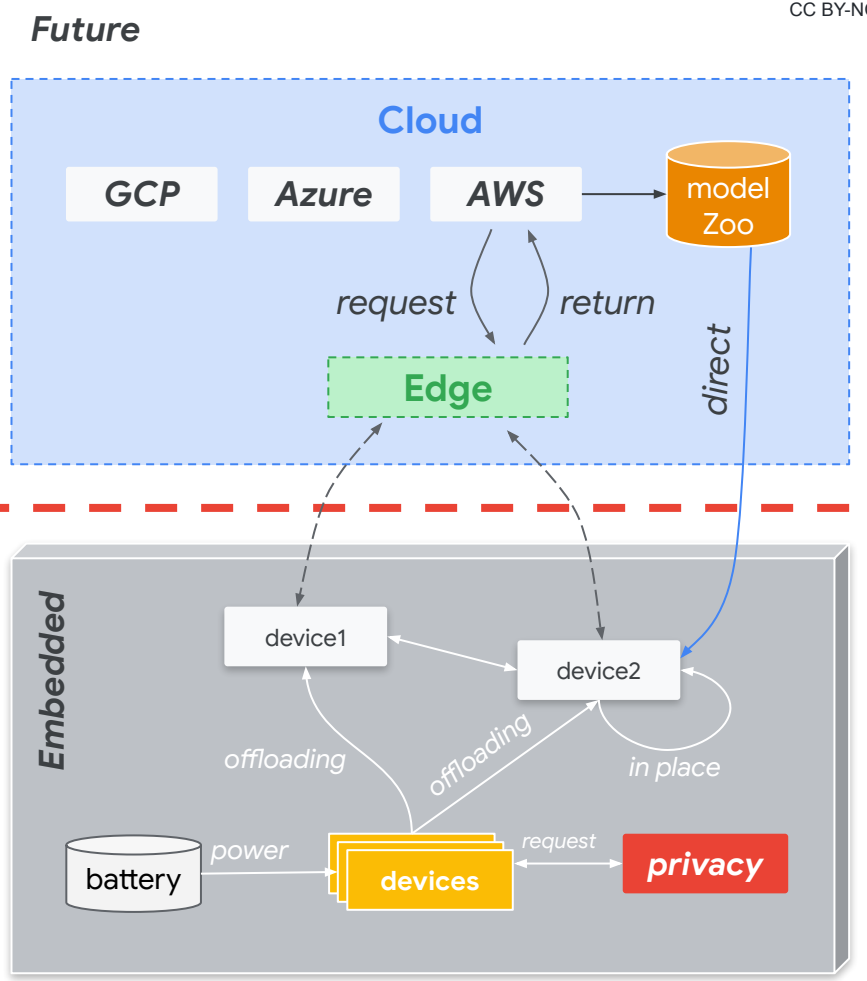
Future



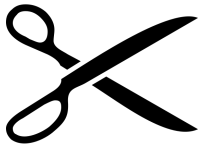


Decouple the **cloud development** environment from the **embedded** model **deployment** environment

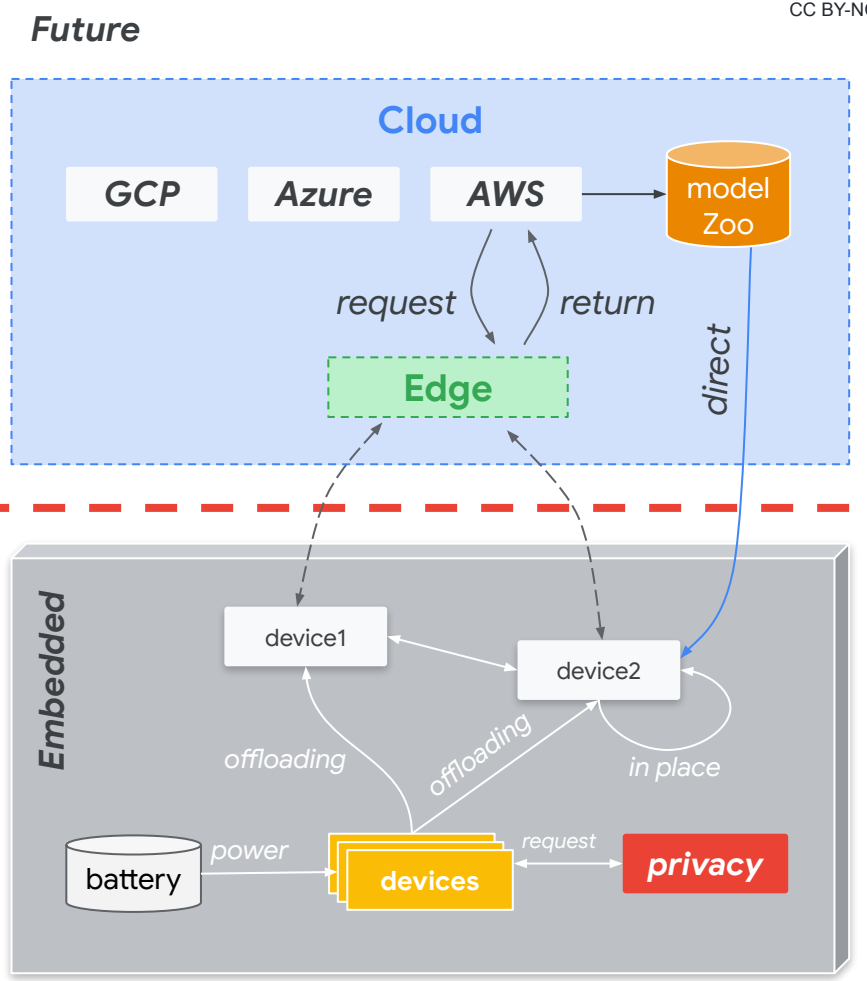




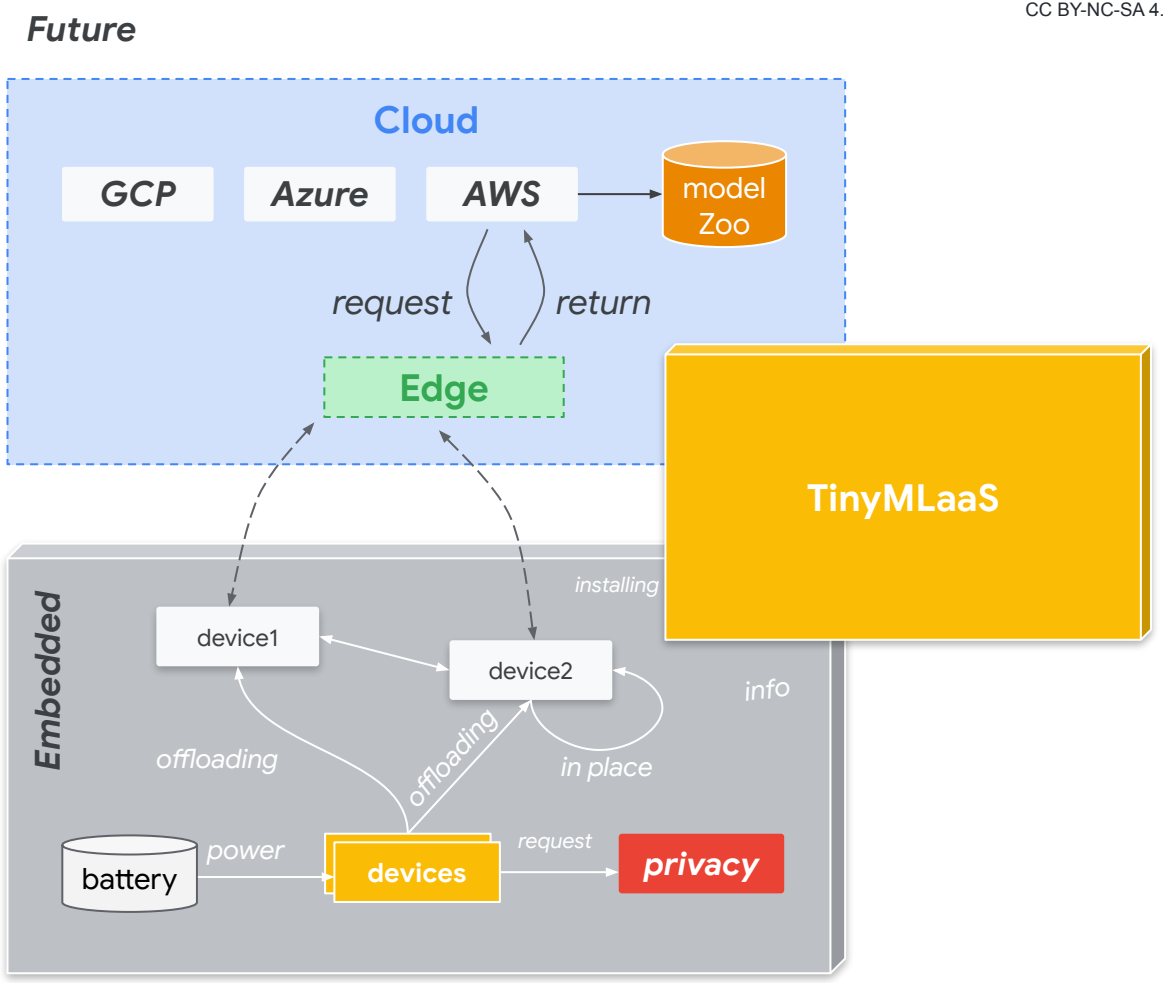
Decouple the **cloud development** environment from the **embedded** model **deployment** environment



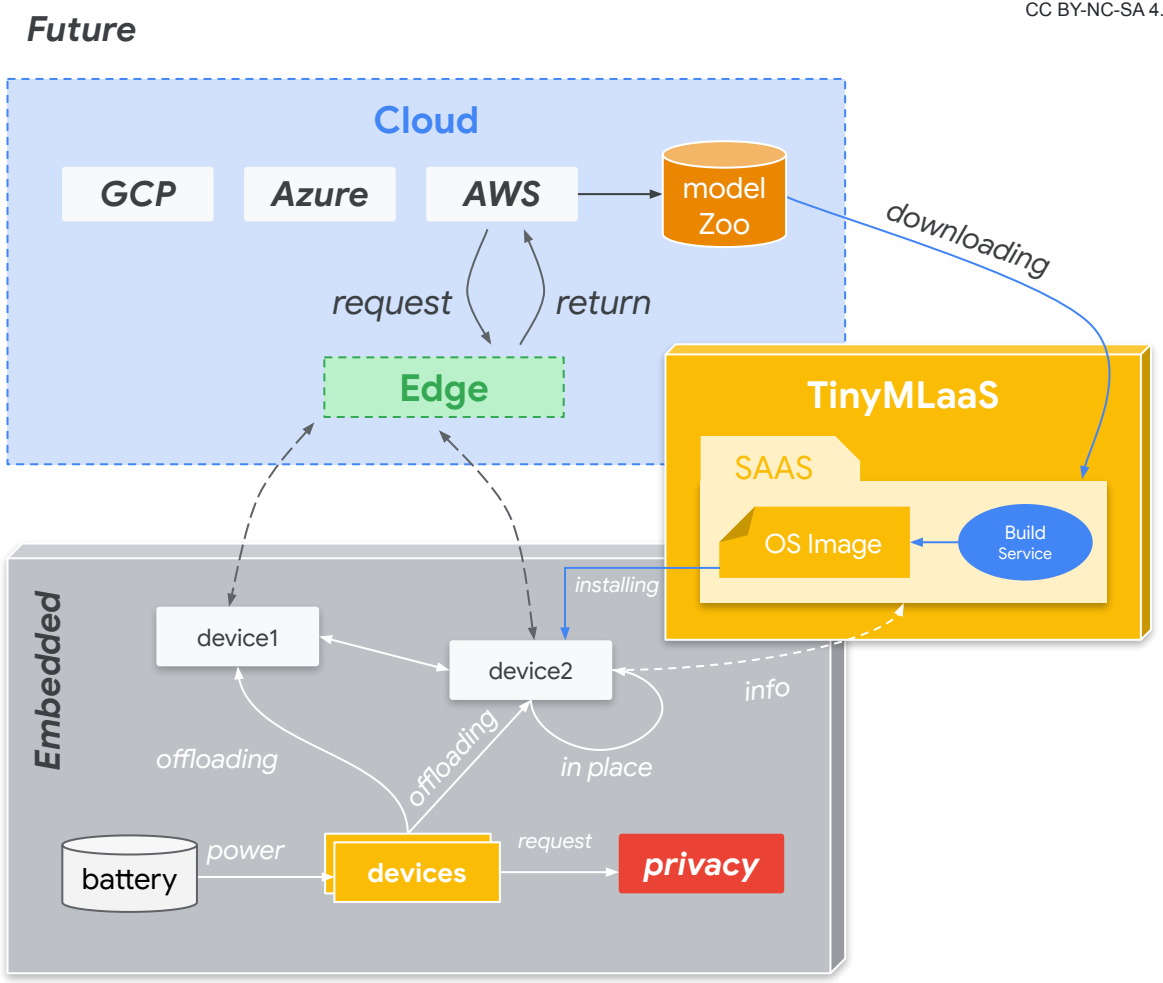
Simplify deployment of ML models to tiny devices and **develop an abstraction**



TinyMLaaS

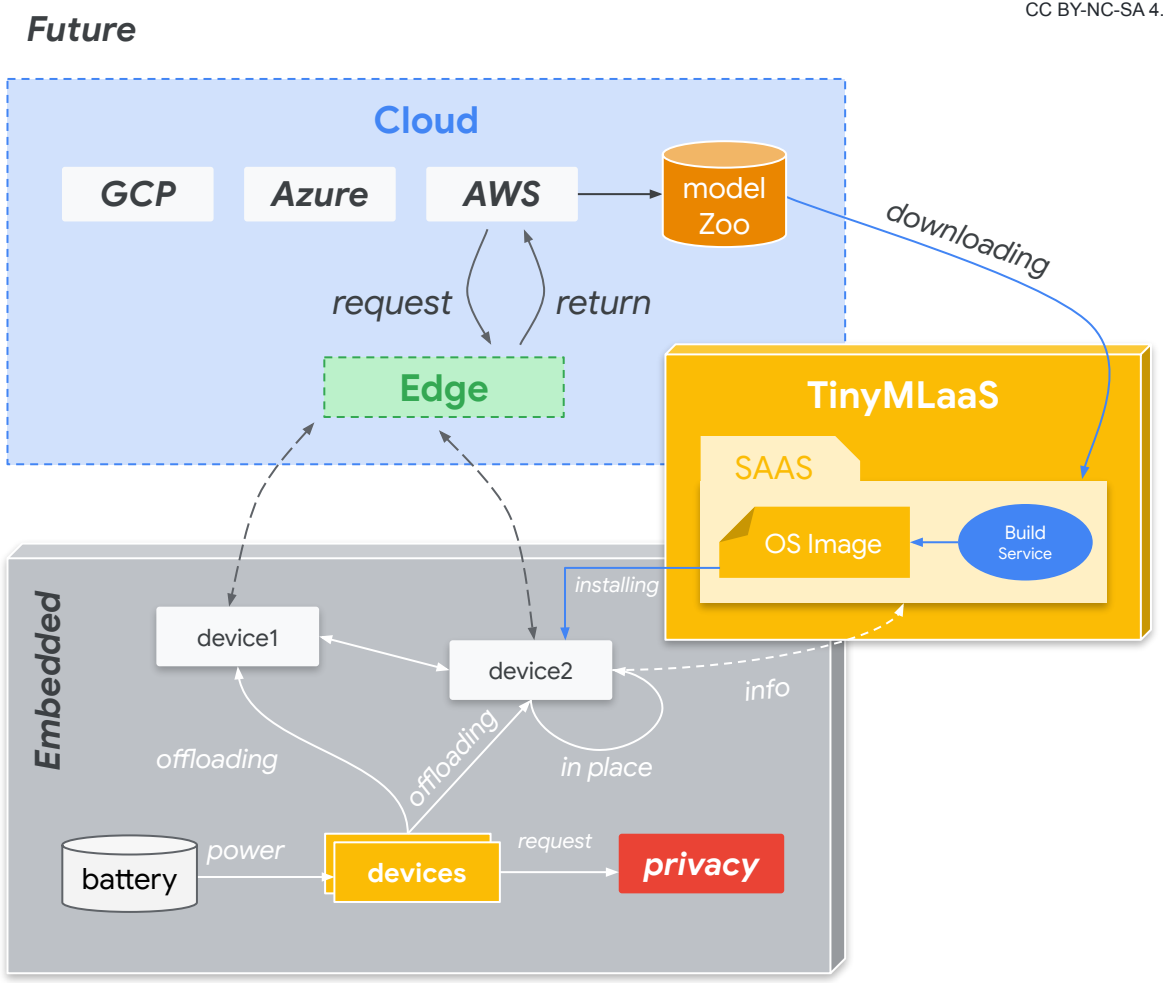


TinyMLaaS



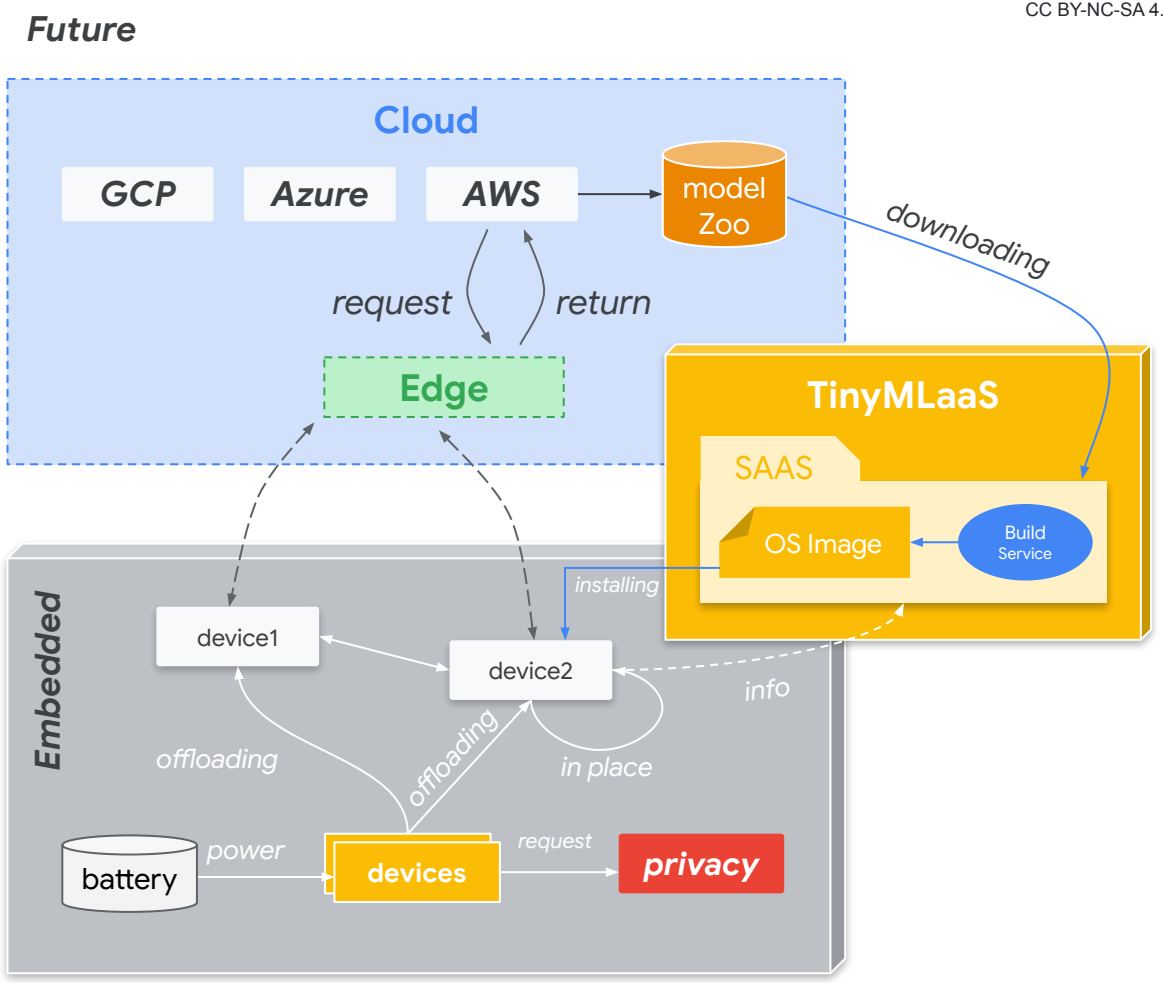
TinyMLaaS

- TinyML as a Service is a cloud or edge-based machine learning as a service
- Simplifies the deployment of ML models → abstraction

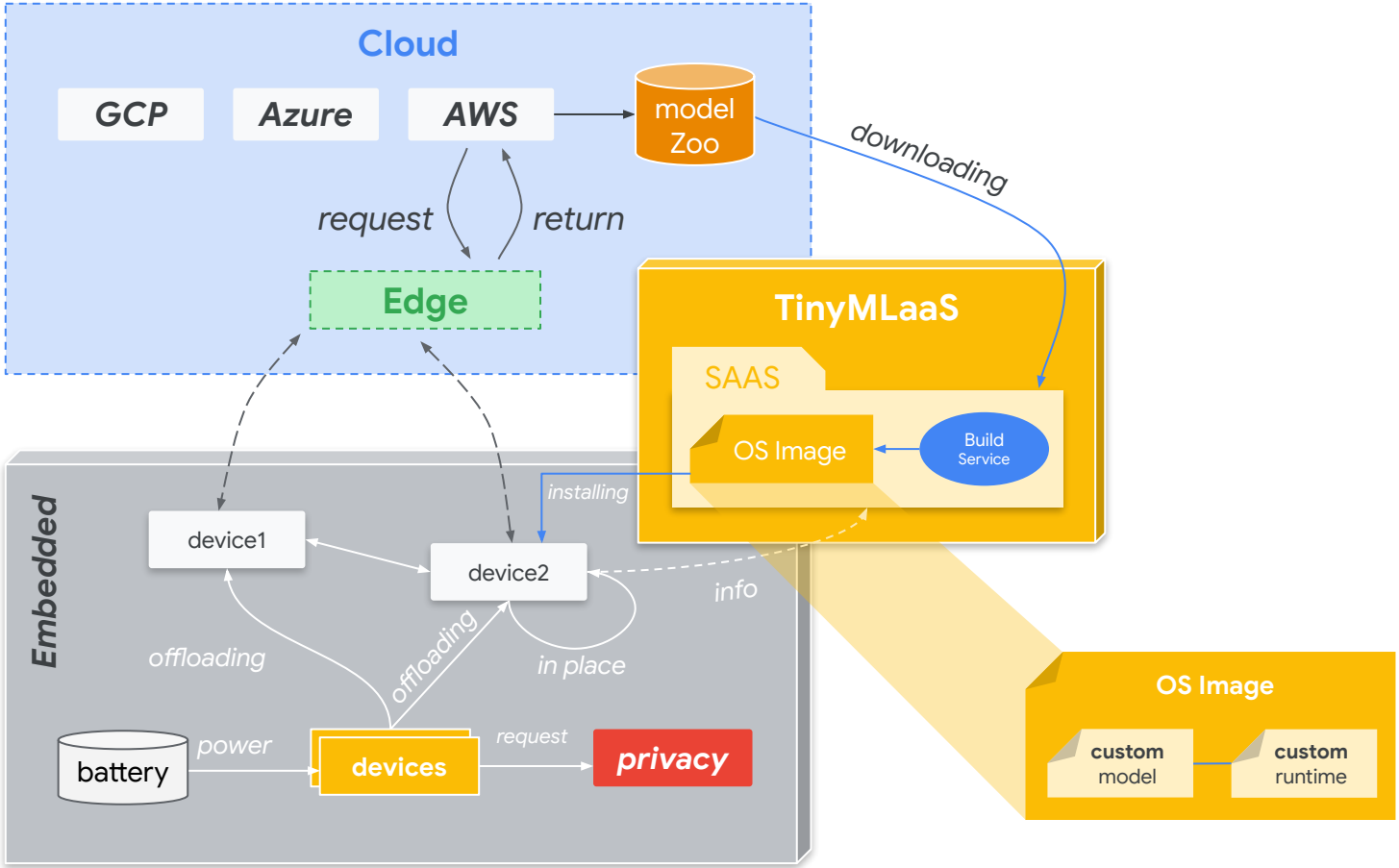


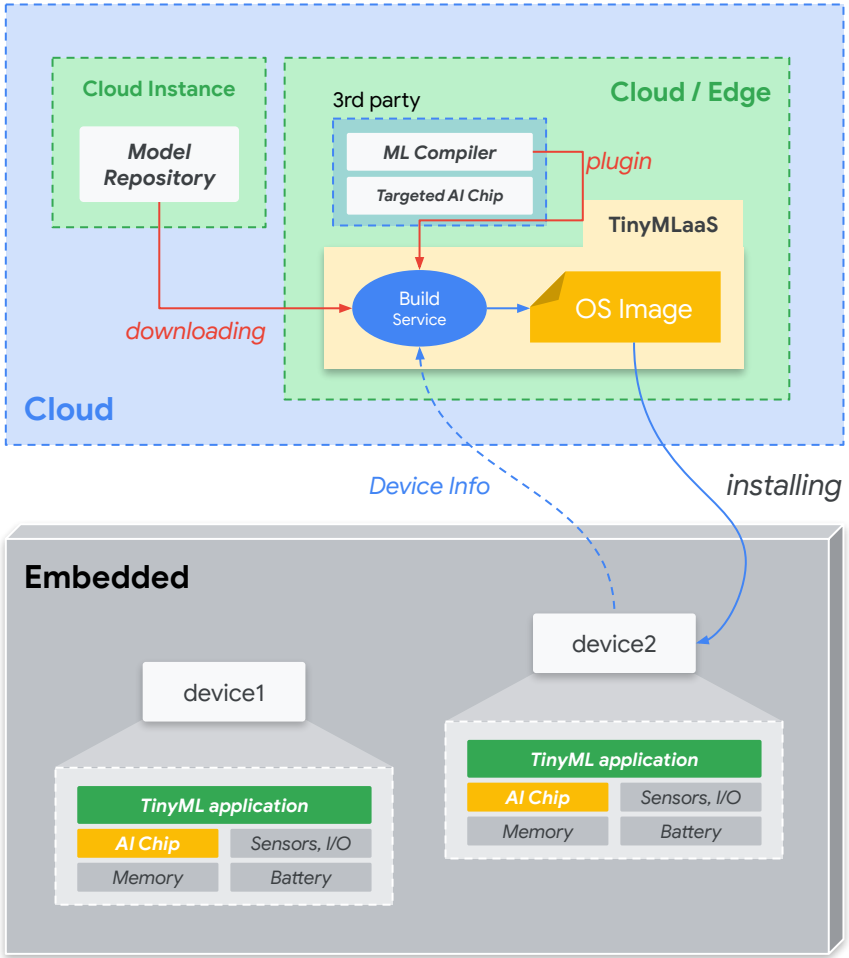
TinyMLaaS

- TinyML as a Service is a cloud or edge-based machine learning as a service
- Simplifies the deployment of ML models → abstraction



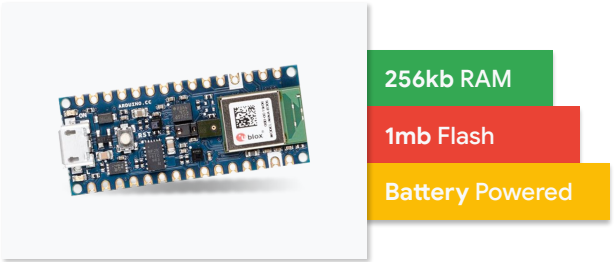
Future

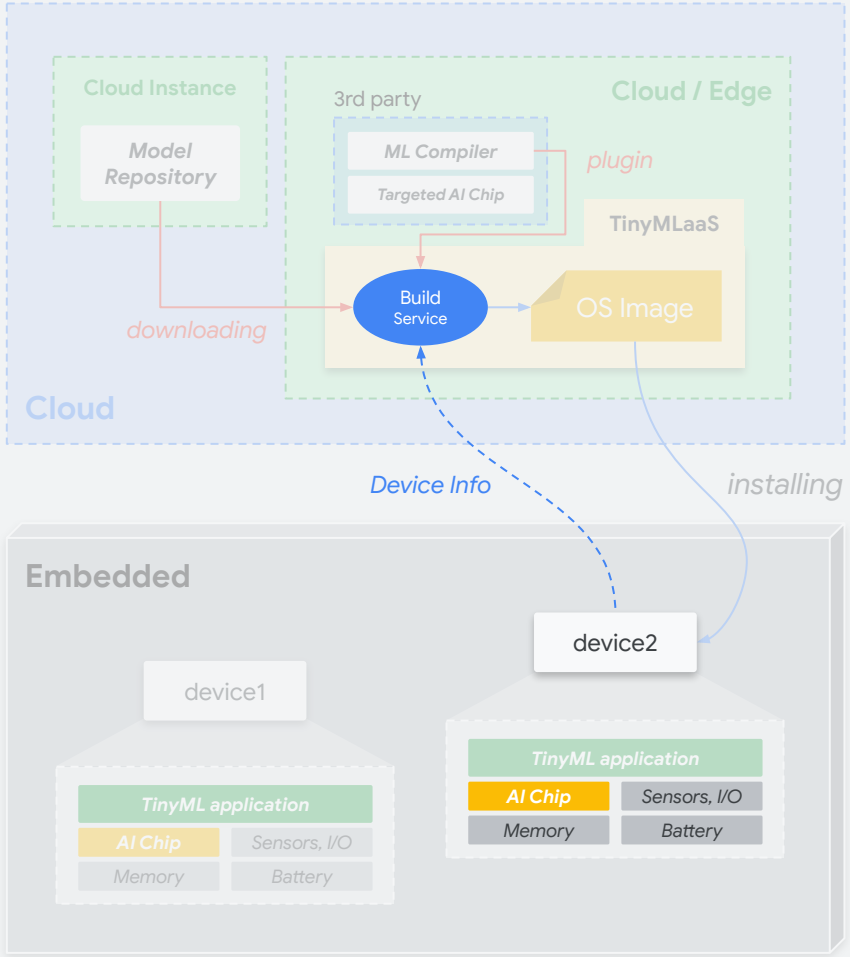




TinyMLaaS

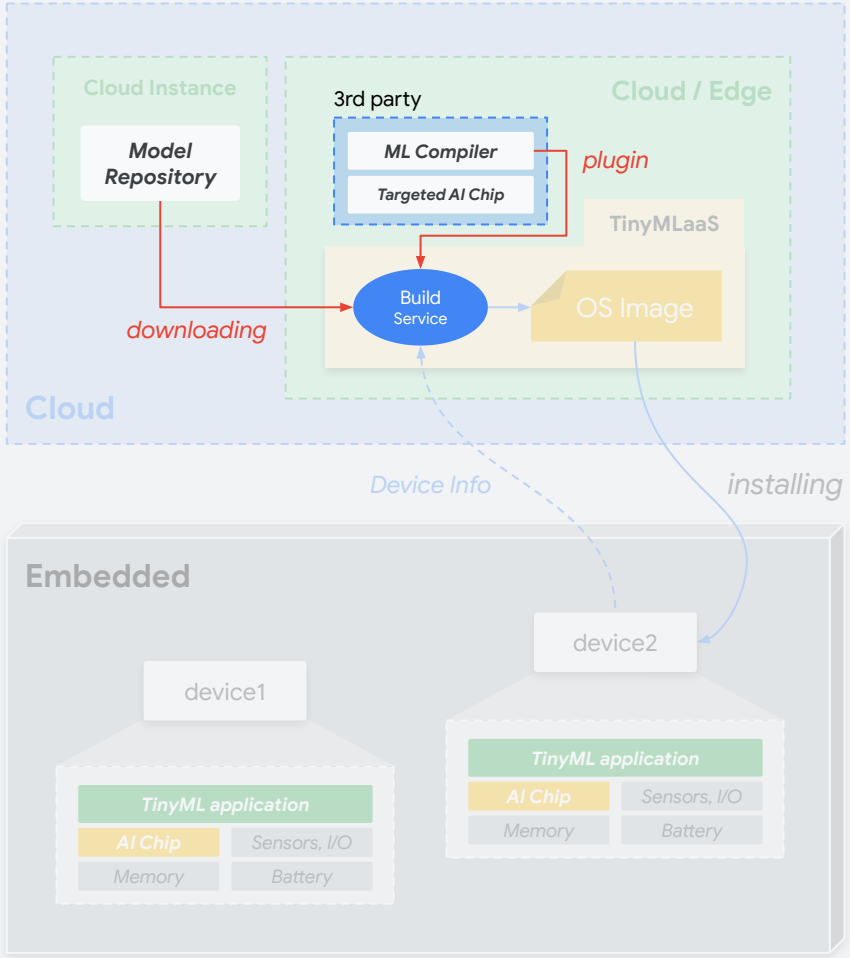
Step 1: Tailor model to device (gather information)





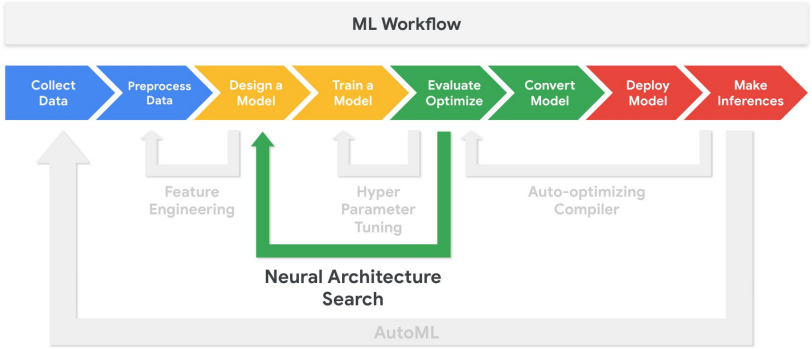
TinyMLaaS

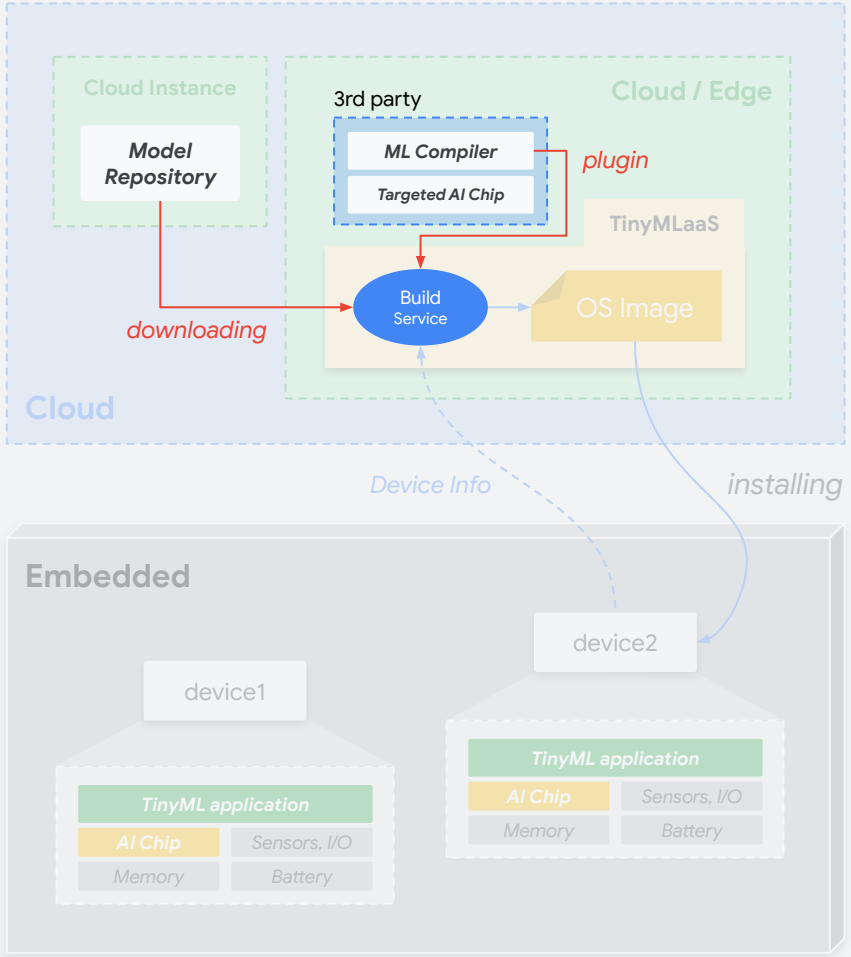
Step 2: TinyMLaaS backend will generate the **compiled ML inference**



TinyMLaaS

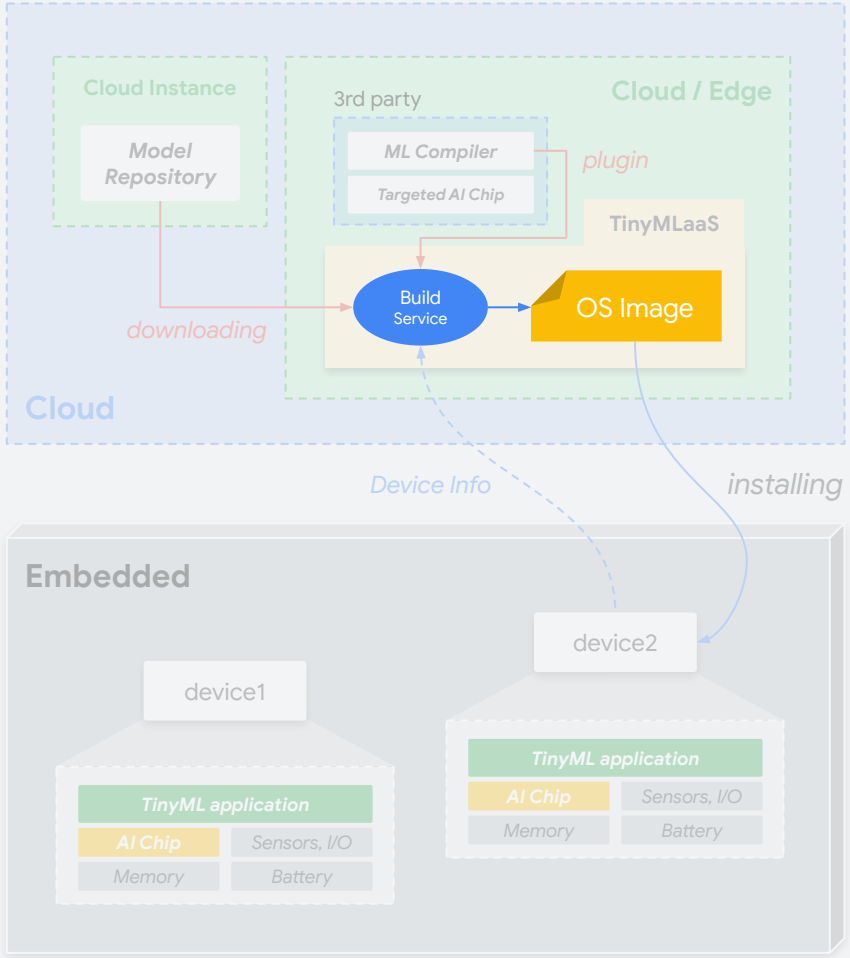
Step 2: TinyMLaaS backend will generate the **compiled ML inference**





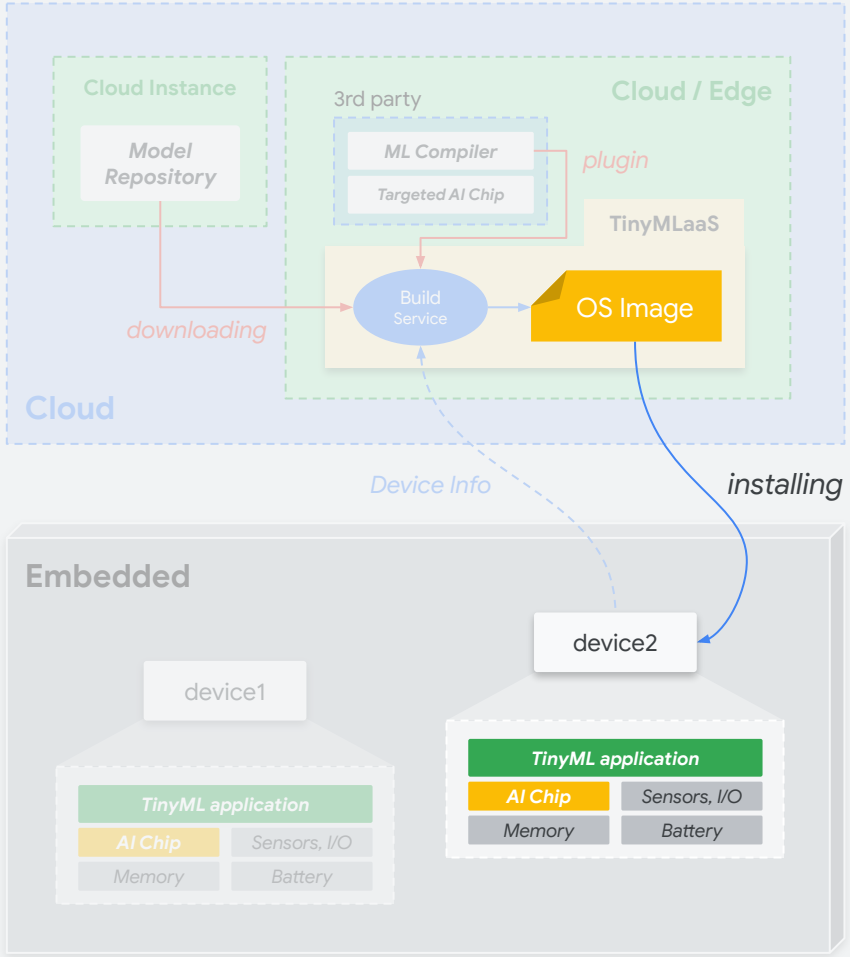
TinyMLaaS

Step 3: The generated ML inference module is then built into a **OS ready image**



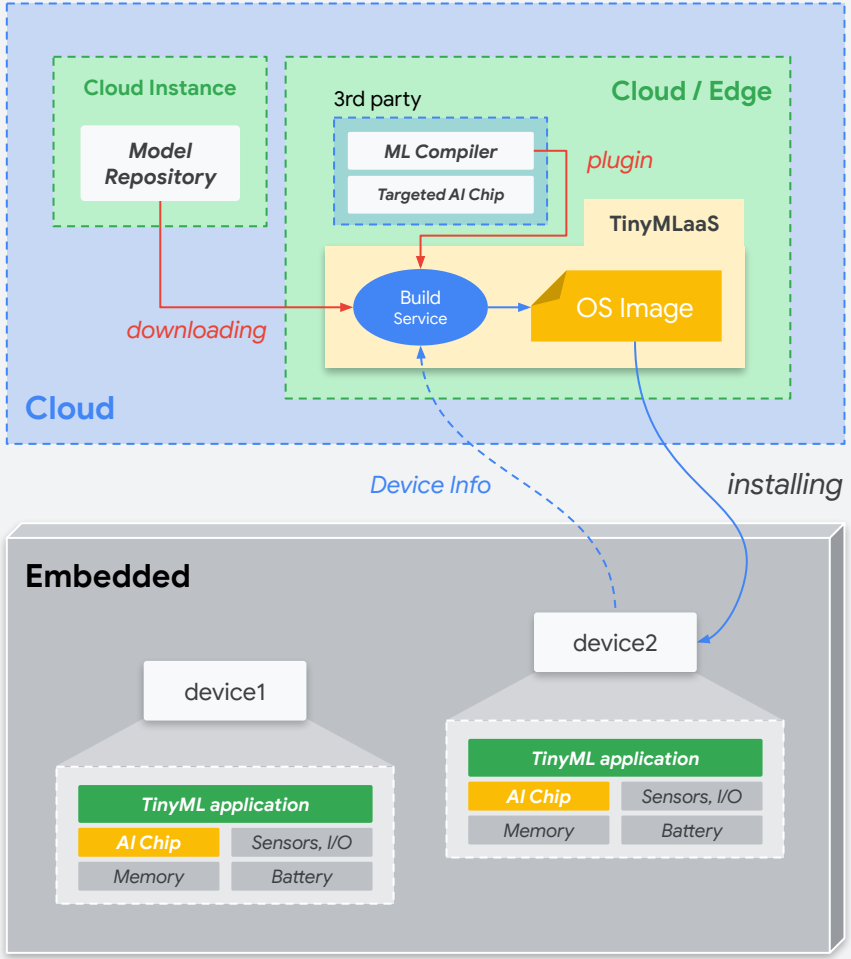
TinyMLaaS

Step 4: The generated image is then finally **installed** on the tinyML device(s)



TinyMLaaS in 4 Simple Steps

1. *Gather Device Information*
2. *Select Compiler, Generate Inference Module*
3. *Create OS-Ready Image*
4. *Install on device*



TinyMLaaS

- TinyML as a Service is a cloud or edge-based machine learning as a service
- Simplifies the deployment of ML models → abstraction