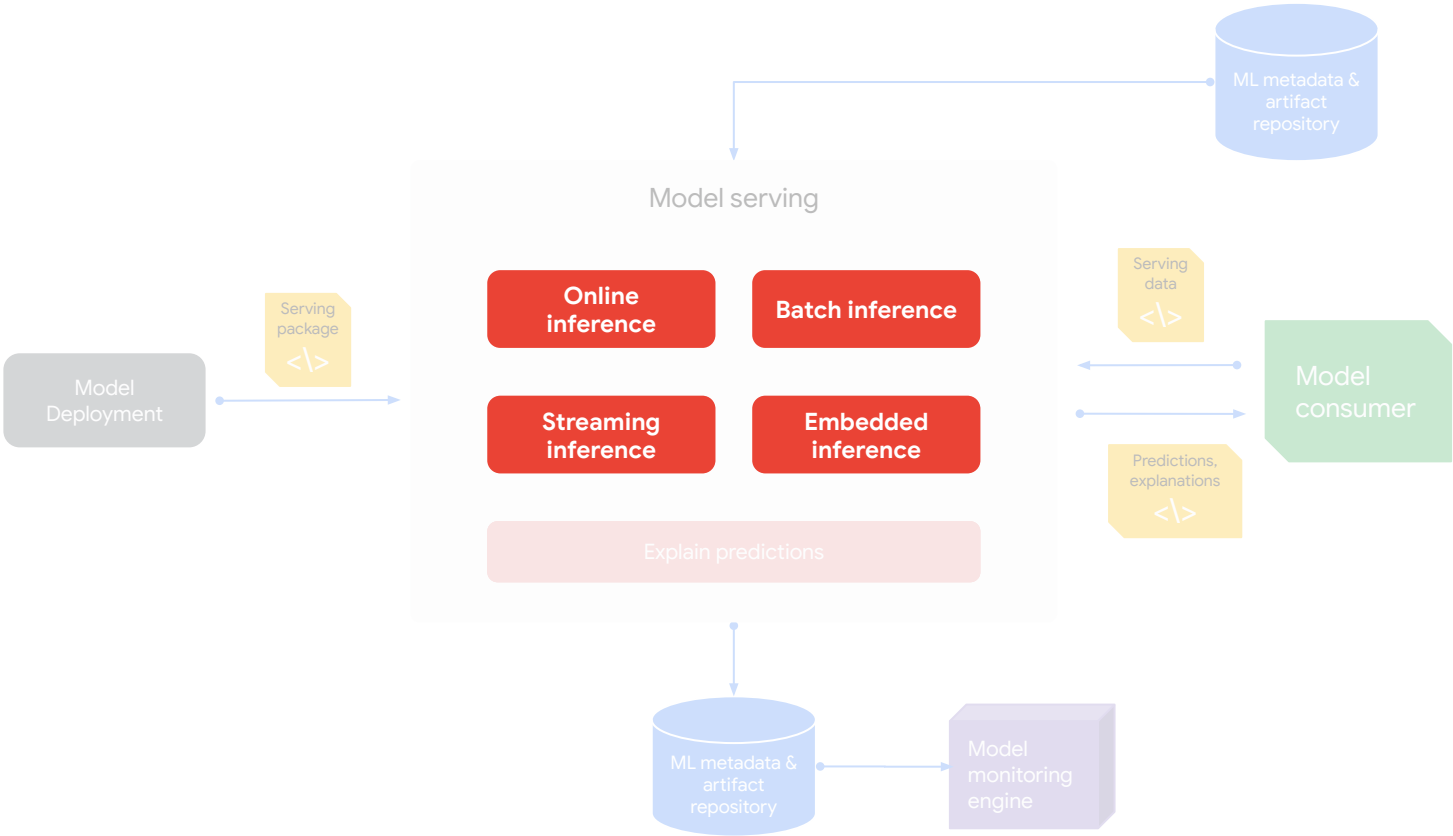
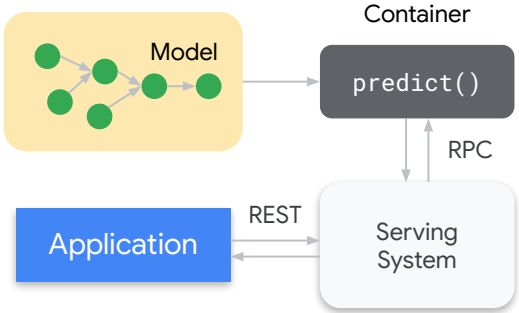
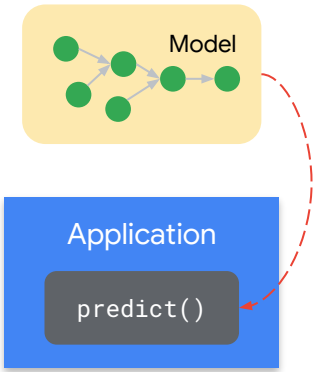
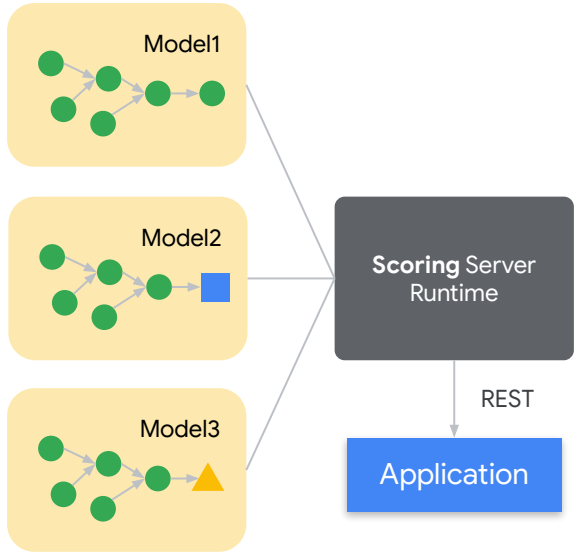
Prediction Serving Architectures

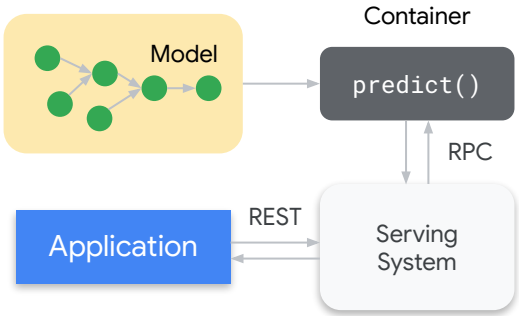
MLOps: Prediction Serving

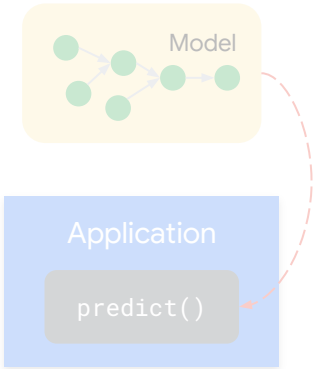


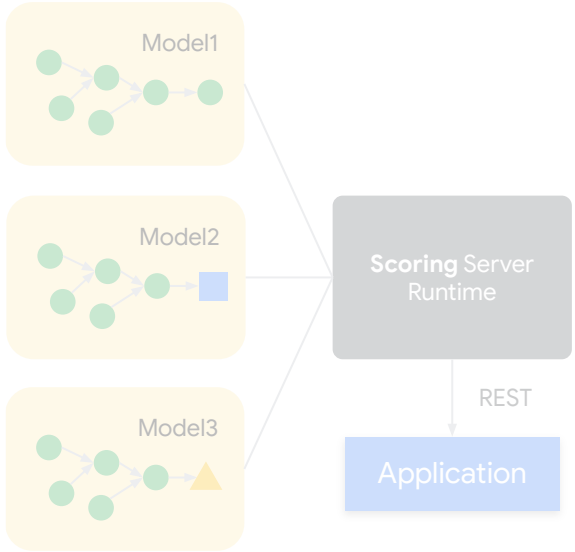






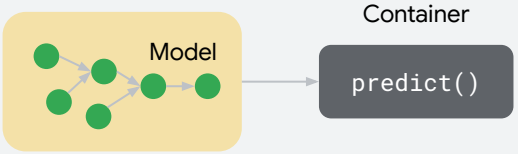






Container-Based Architecture

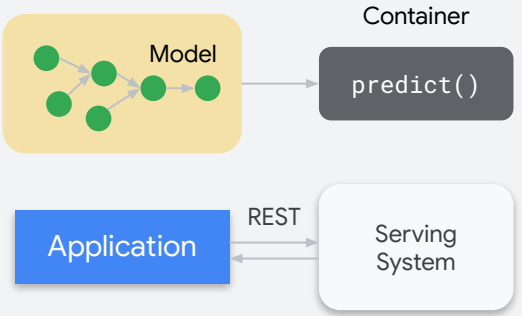
Models are deployed into **containers** (such as using Dockers) and connected to a serving system



Container-Based Architecture

Models are deployed into **containers** (such as using Dockers) and connected to a serving system

Applications call into the web server that is hosted on the **serving system** using a REST API interface

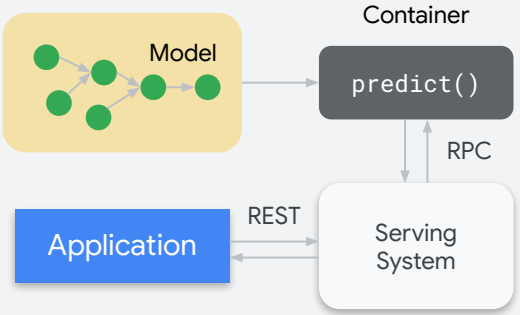


Container-Based Architecture

Models are deployed into **containers** (such as using Dockers) and connected to a serving system

Applications call into the web server that is hosted on the **serving system** using a REST API interface

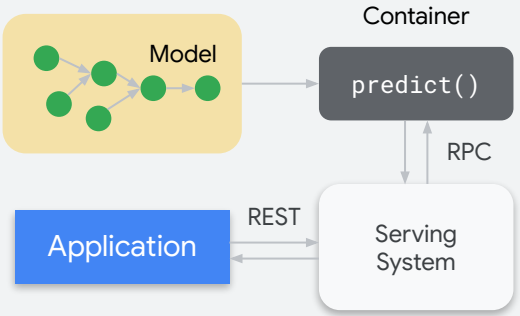
Seving system is wired with **RPC** to invoke the models in the container



Container-Based Architecture

Pros

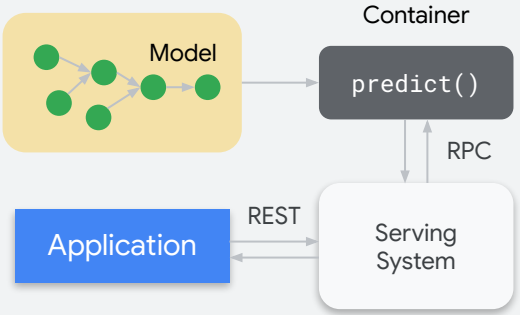
+ Eases **implementation**



Container-Based Architecture

Pros

+ Eases **implementation**
+ Great for **scalability** and **fault tolerance** mechanisms



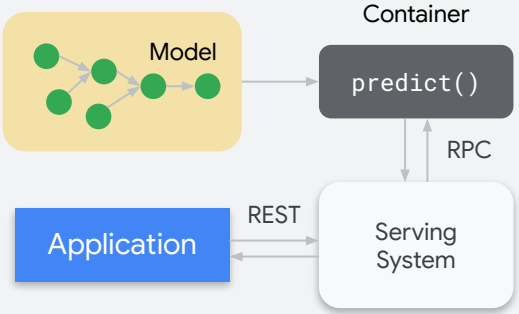
Container-Based Architecture

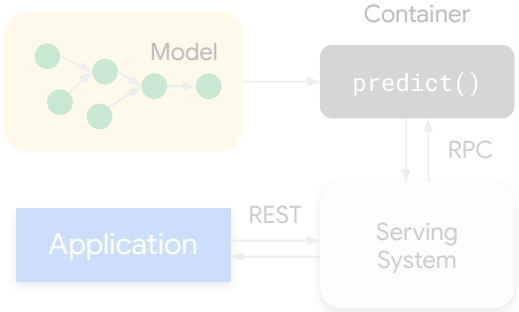
Pros

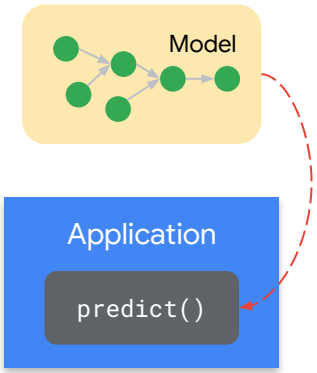
+ Eases **implementation**
+ Great for **scalability** and **fault tolerance** mechanisms

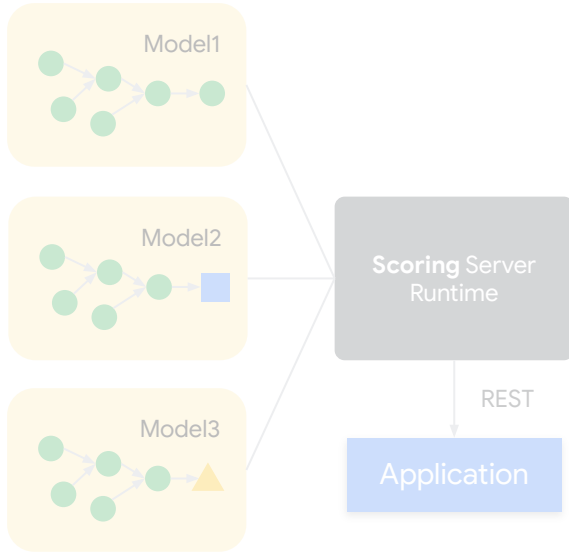
Cons

- **Not suitable for low-latency** cause of communication and resource overheads



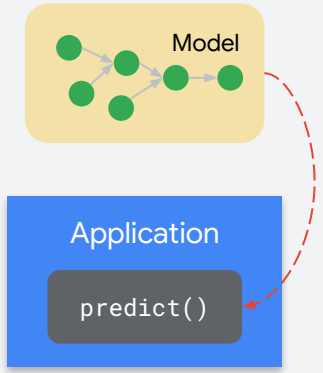






Direct-Import Architecture

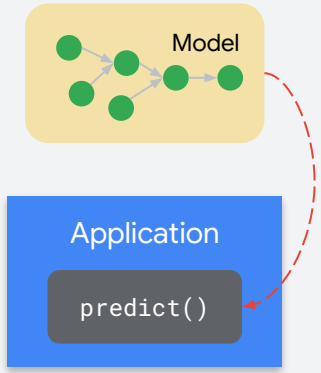
We **integrate the model** logic directly into the application



Direct-Import Architecture

We **integrate the model** logic directly into the application

Suitable for the **cloud as well as for edge devices** and it unlocks low latency scenarios

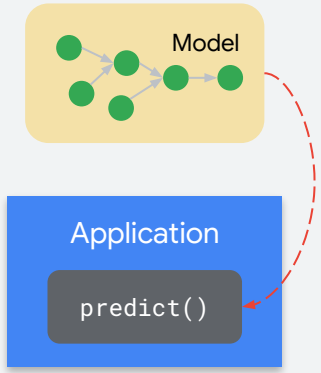


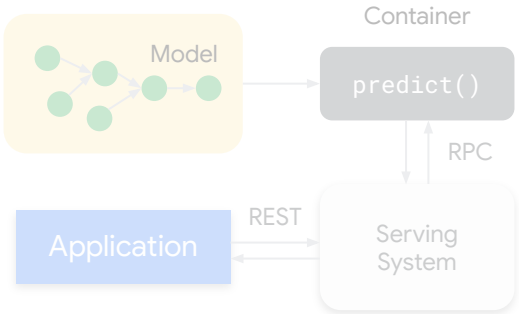
Direct-Import Architecture

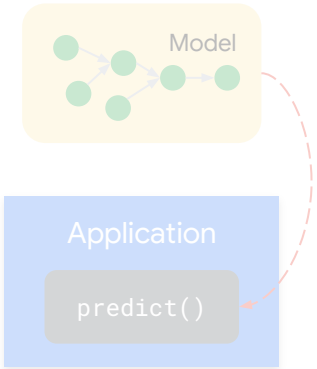
We **integrate the model** logic directly into the application

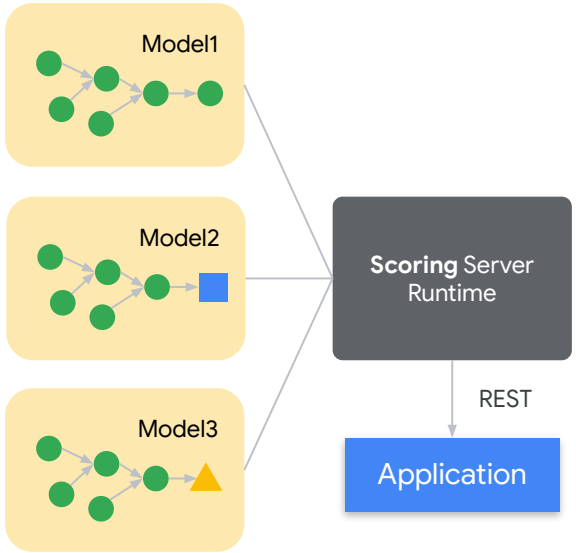
Suitable for the **cloud as well as for edge devices** and it unlocks low latency scenarios

Removes the overhead of managing containers and implementing RPC functionalities to communicate



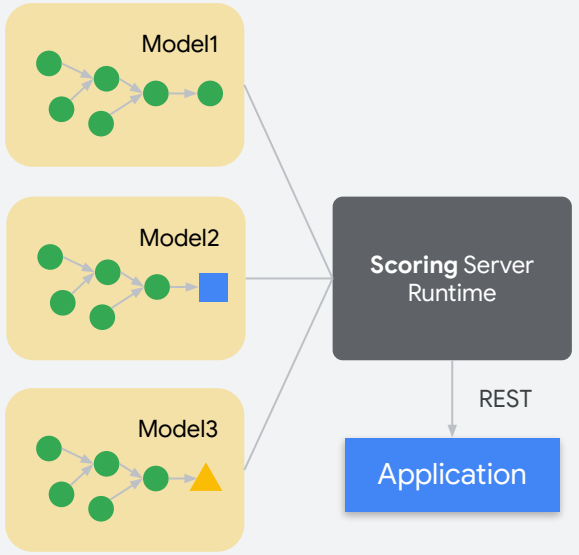






WhiteBox Architecture

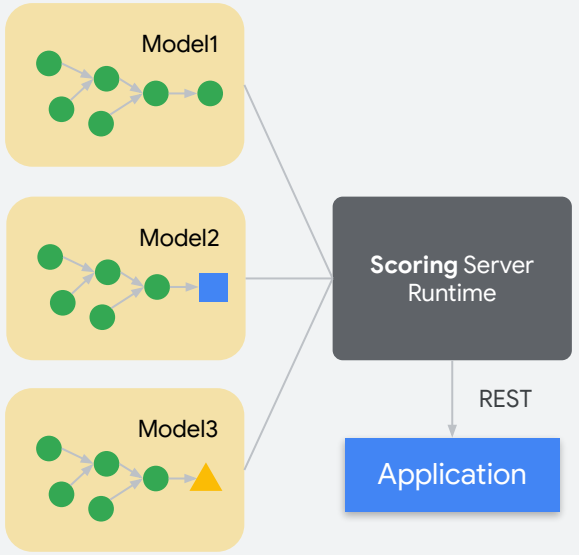
Models are registered to a Runtime that considers them not as mere executable code but as DAGs of operators



WhiteBox Architecture

Models are registered to a Runtime that considers them not as mere executable code but as DAGs of operators

Applications submit a REST request to a **cloud-hosted runtime**

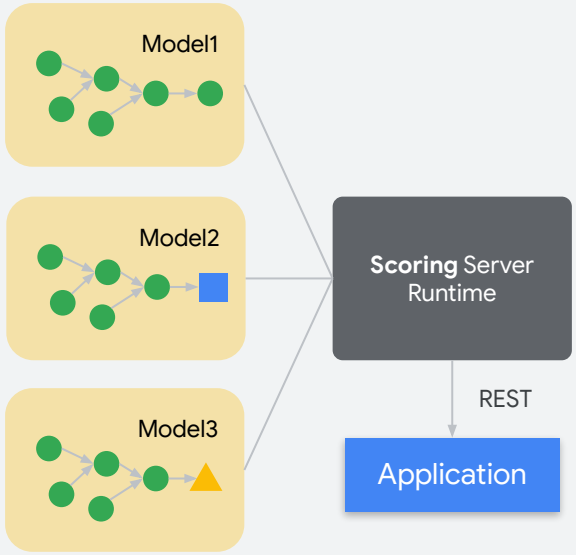


WhiteBox Architecture

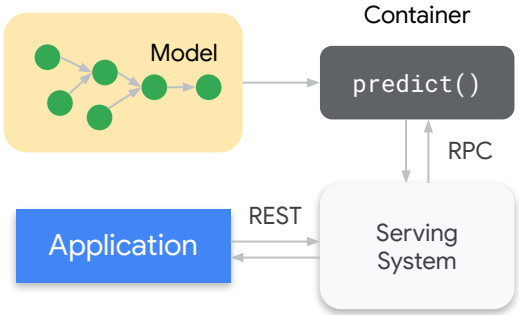
Models are registered to a Runtime that considers them not as mere executable code but as DAGs of operators

Applications submit a REST request to a **cloud-hosted runtime**

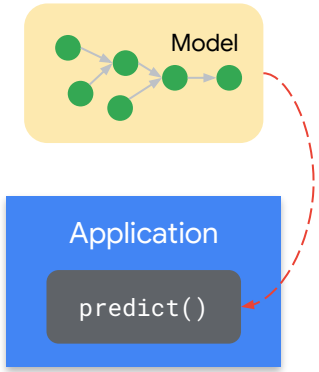
Runtime can apply a wide array of **system level and compiler level optimizations** to improve latency or improve memory consumption and computation reuse



Container-based Architecture



Direct-Import Architecture



WhiteBox Architecture