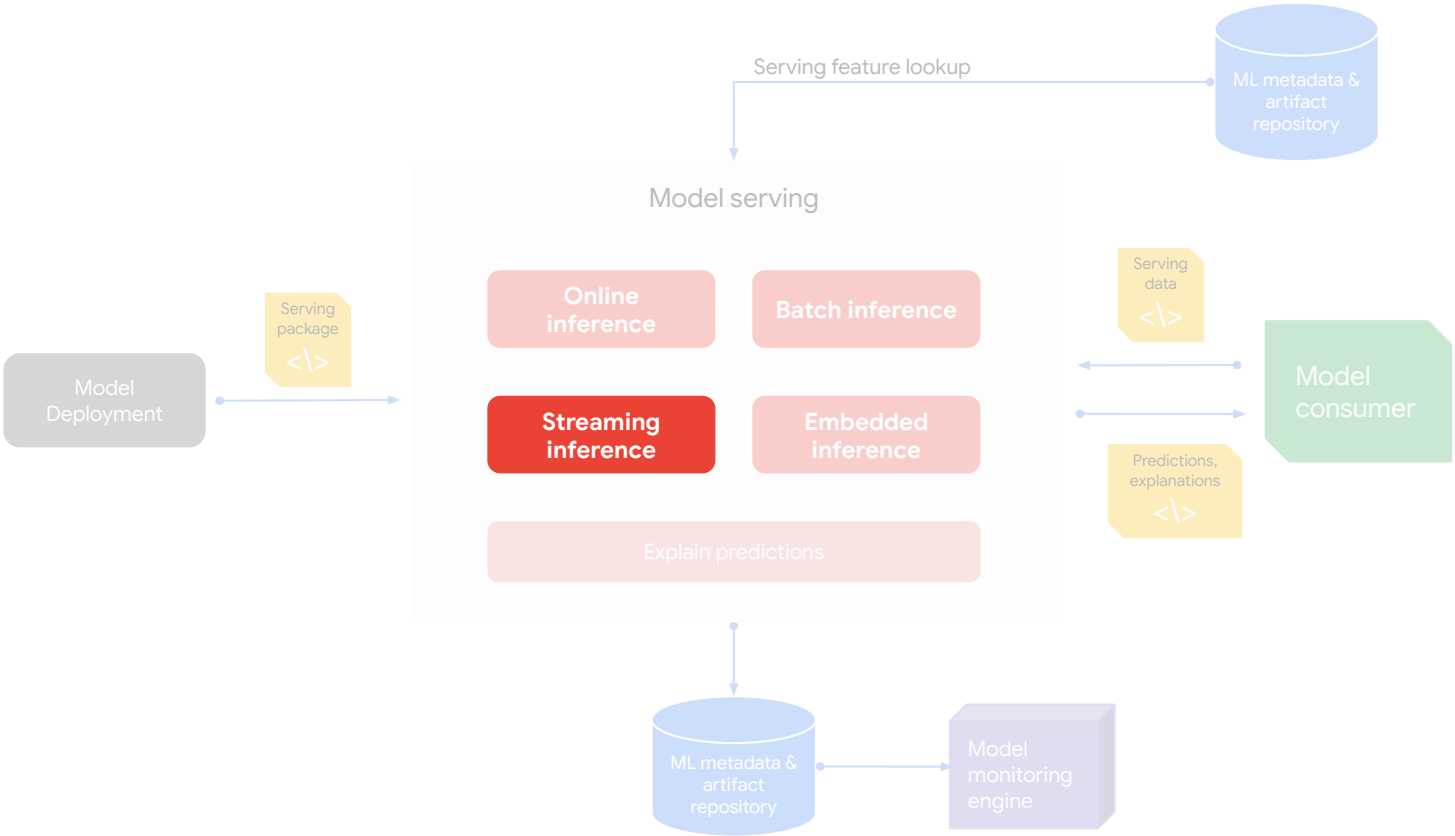
Prediction Serving Scenarios: Streaming

MLOps: Prediction Serving

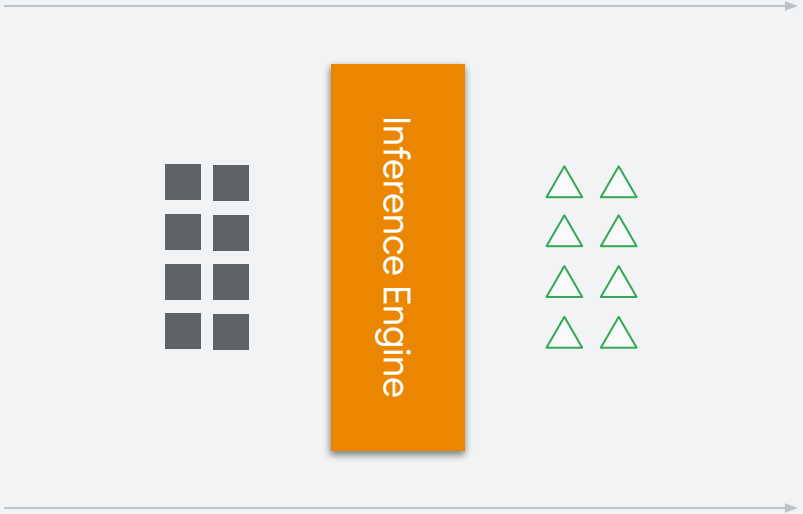


Streaming Inference: **What is it?**

- Predict **on-demand**

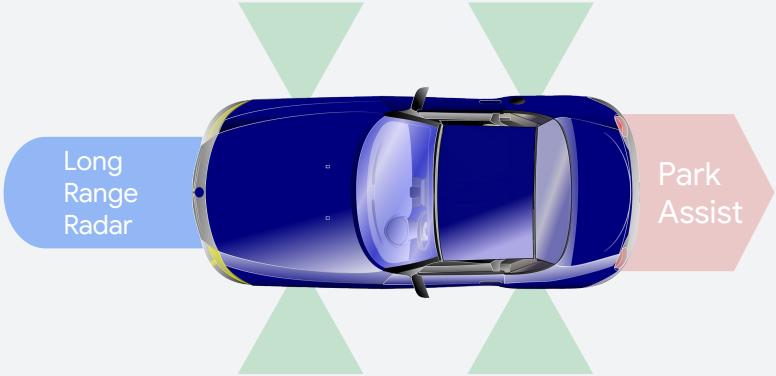
Streaming Inference: **What is it?**

- Predict **on-demand**
- Multiple “streams” or inputs that are concurrently coming together for **real-time** model inference



Streaming Inference: When is it useful?

- **Autonomous cars**
- Smart cities
- Fast science
- Robotics
- ...



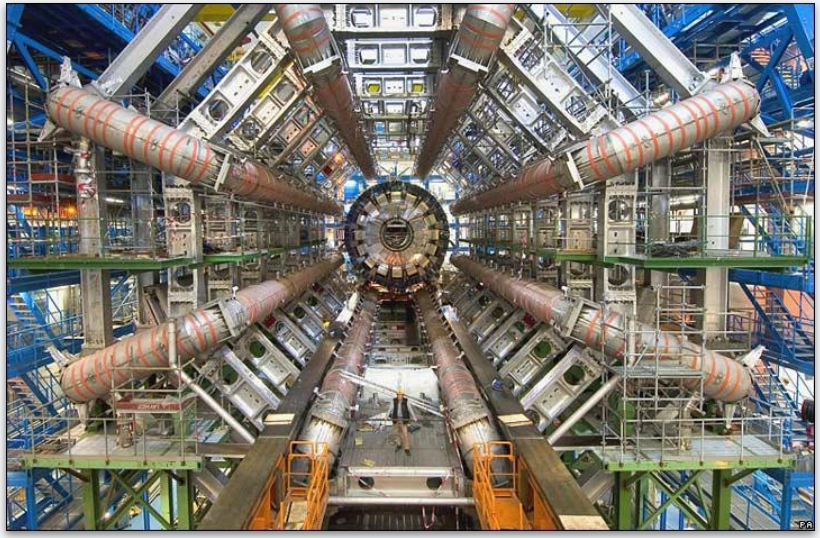
Environment Mapping

Streaming Inference: When is it useful?

- Autonomous cars
- **Smart cities**
- Fast science
- Robotics
- ...

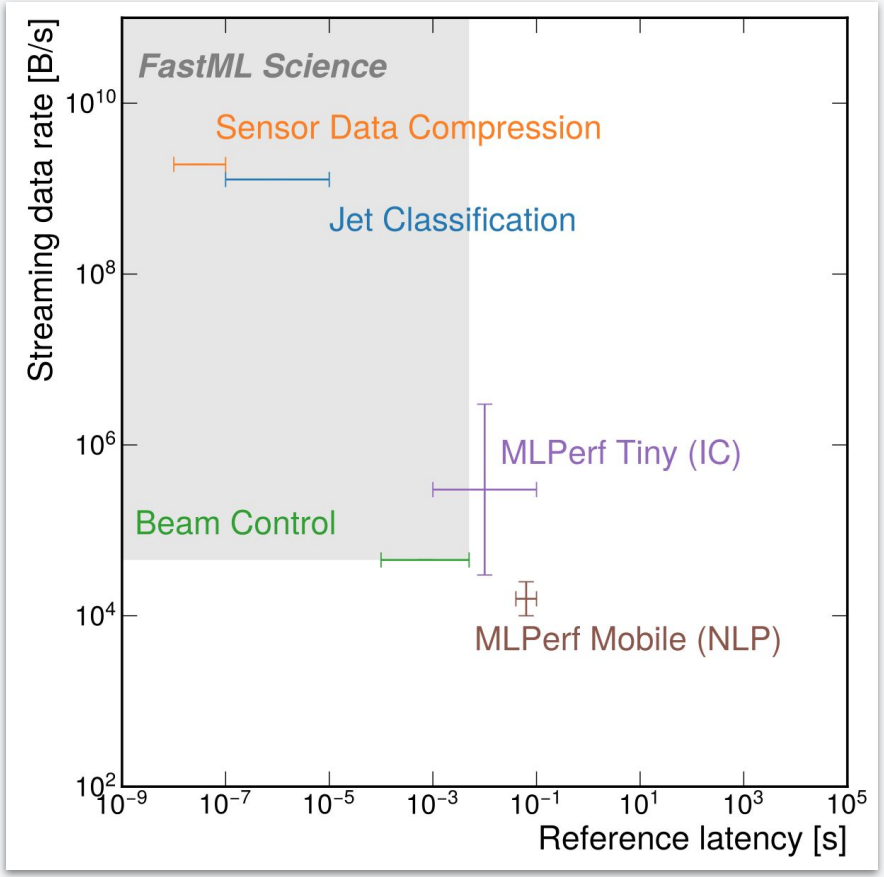
Streaming Inference: When is it useful?

- Autonomous cars
- Smart cities
- **Fast science**
- Robotics
- ...



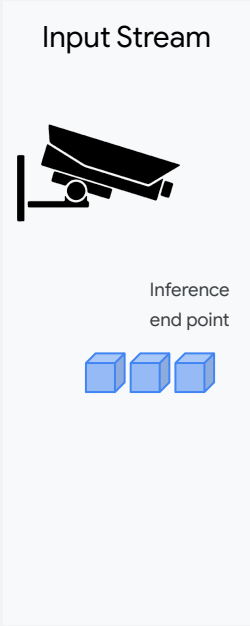
Streaming Inference: When is it useful?

- Autonomous cars
- Smart cities
- **Fast science**
- Robotics
- ...



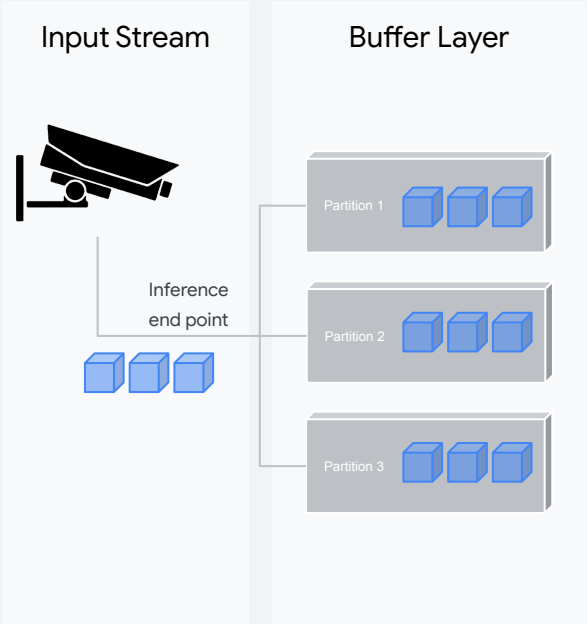
Streaming Inference: How it works?

Loosely similar to “Batch inference” in that it has to handle batches of data



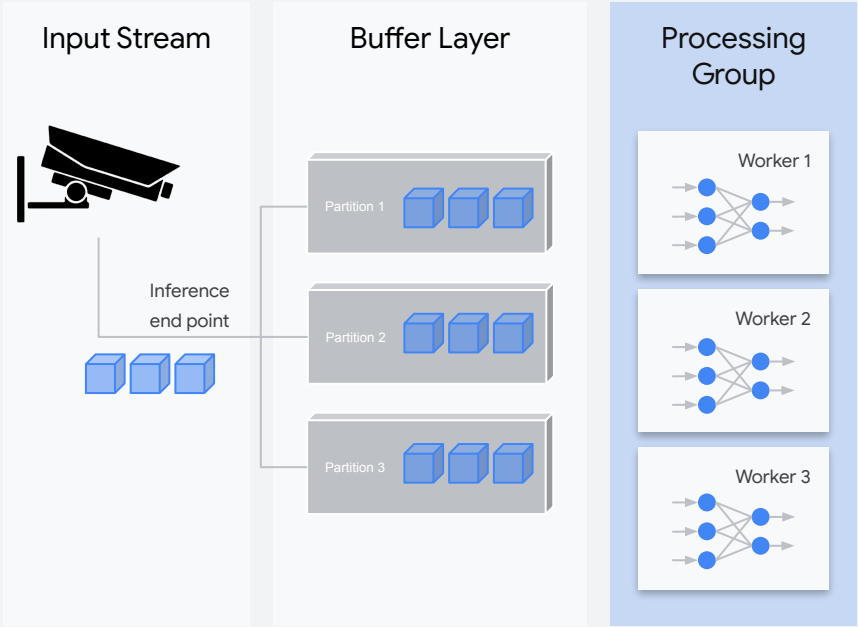
Streaming Inference: How it works?

Loosely similar to “Batch inference” in that it has to handle batches of data



Streaming Inference: How it works?

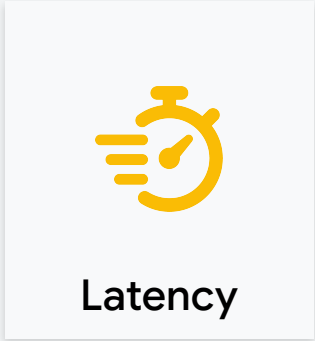
Loosely similar to “Batch inference” in that it has to handle batches of data



Streaming Inference: What metrics?

- Latency is important
- Throughput is also important
- Metric:

  Latency-bounded-throughput



Streaming Inference: Pros & Cons

Pros

+ Scale out your inference capabilities by feeding input back to **scalable** infrastructure

Streaming Inference: Pros & Cons

Pros

+ Scale out your inference capabilities by feeding input back to **scalable** infrastructure

+ Able to **sync multiple points** of data to build a global picture

Streaming Inference: Pros & Cons

Pros

+ Scale out your inference capabilities by feeding input back to **scalable** infrastructure
+ Able to **sync multiple points** of data to build a global picture

Cons

- Need to have a more **complex backend** to handle multiple inference service requests

Streaming Inference: Pros & Cons

Pros

+ Scale out your inference capabilities by feeding input back to **scalable** infrastructure
+ Able to **sync multiple points** of data to build a global picture

Cons

- Need to have a more **complex backend** to handle multiple inference service requests
- Compute demands are **higher**

Streaming Inference: Pros & Cons

Pros

+ Scale out your inference capabilities by feeding input back to **scalable** infrastructure
+ Able to **sync multiple points** of data to build a global picture

Cons

- Need to have a more **complex backend** to handle multiple inference service requests
- Compute demands are **higher**
- Data movement can be **costly**

Streaming Inference: Pros & Cons

Pros

+ Scale out your inference capabilities by feeding input back to **scalable** infrastructure
+ Able to **sync multiple points** of data to build a global picture

Cons

- Need to have a more **complex backend** to handle multiple inference service requests
- Compute demands are **higher**
- Data movement can be **costly**

Scenario	Metric
Batch inference (e.g. photo sorting app)	**Throughput**
Online inference (e.g. translation app)	**QPS** subject to latency bound
 **Streaming inference** (e.g. multiple camera driving assistance)	**Number streams** subject to latency bound