# Task based assistant core challenges: system scale

Robust language sensing

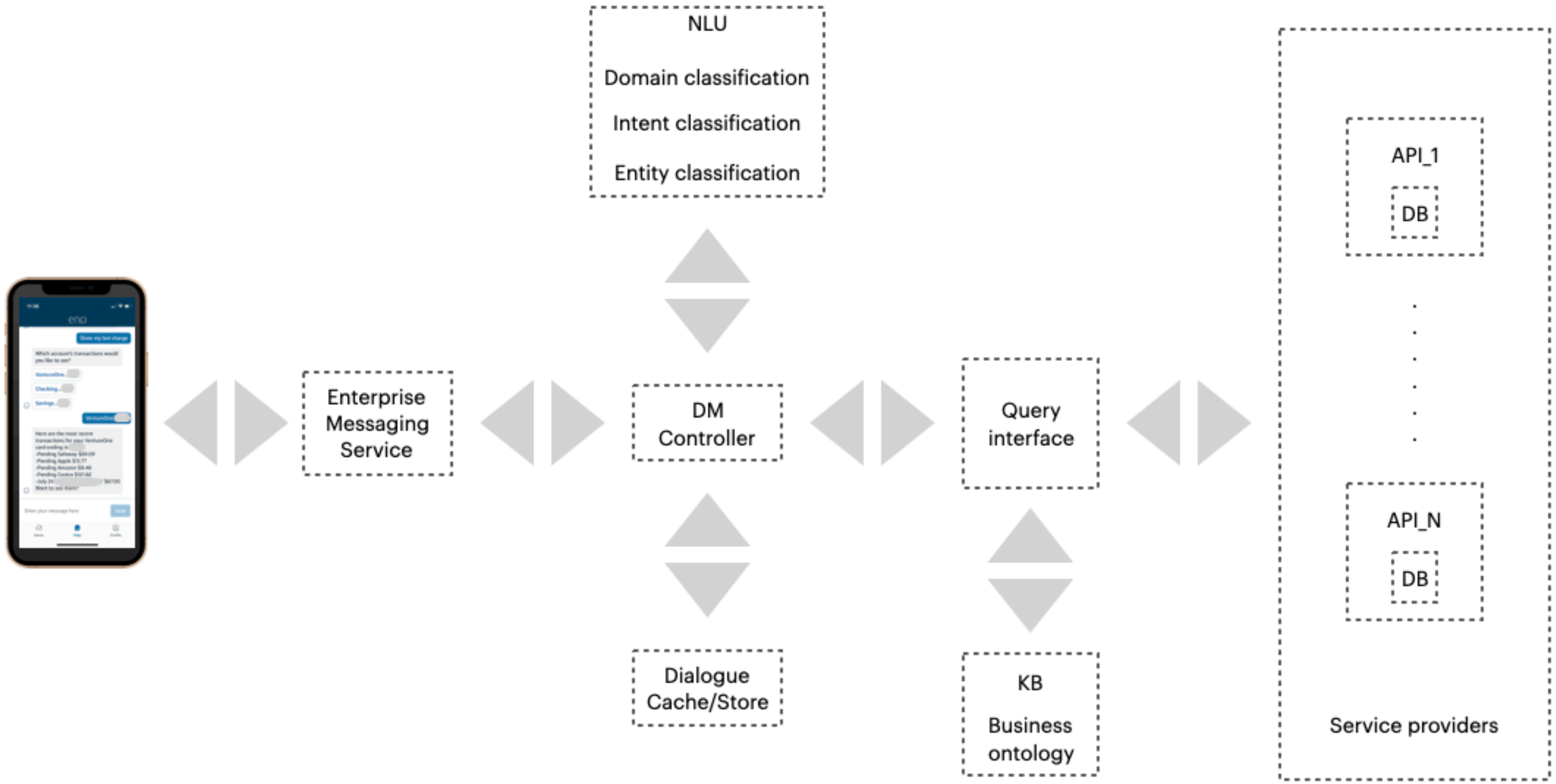
- Spoken vs typed
- Slang vs formal
- Typos & dialects
- Microphone quality
- Intent, entity, domain rec.
- Implicit entities

Dialogue state management

- Efficient slot filling
- Multi intent requests
- Contextual awareness

**System scale**

- **Multi-domain support**
- **3rd party integrations**
- **Performance**



NLU

Domain classification

Intent classification

Entity classification

Enterprise Messaging Service

DM Controller

Query interface

Dialogue Cache/Store

KB

Business ontology

API_1

DB

API_N

DB

Service providers

Generic dialogue system architecture

# Scaling dialogue systems involve many considerations

* Understand system level latency profiles
    * Messaging bus
    * Internal APIs that you don’t own

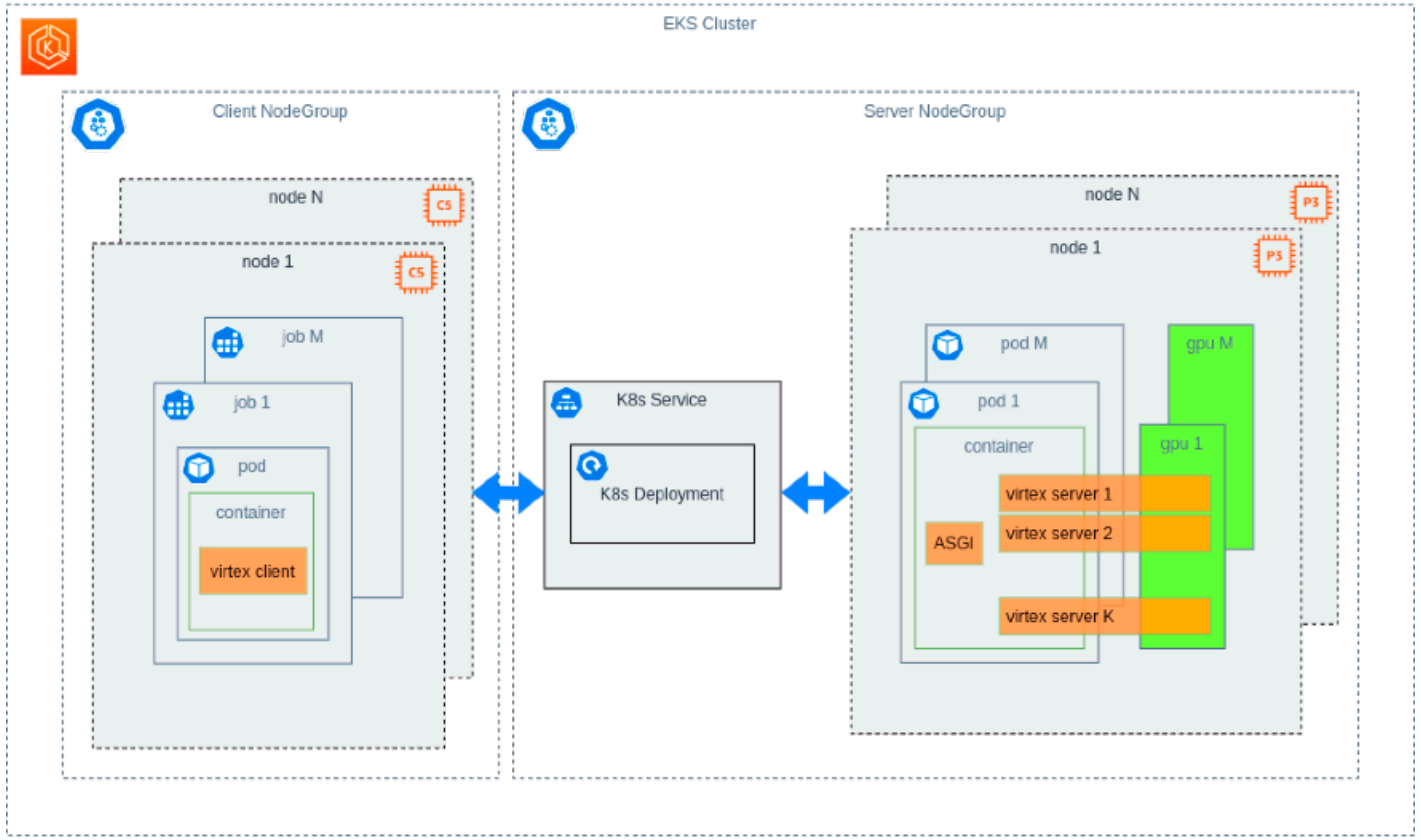
* Compute-bound services should be isolated
    * Dynamic batch processing

* Unit testing, including 3rd party tools

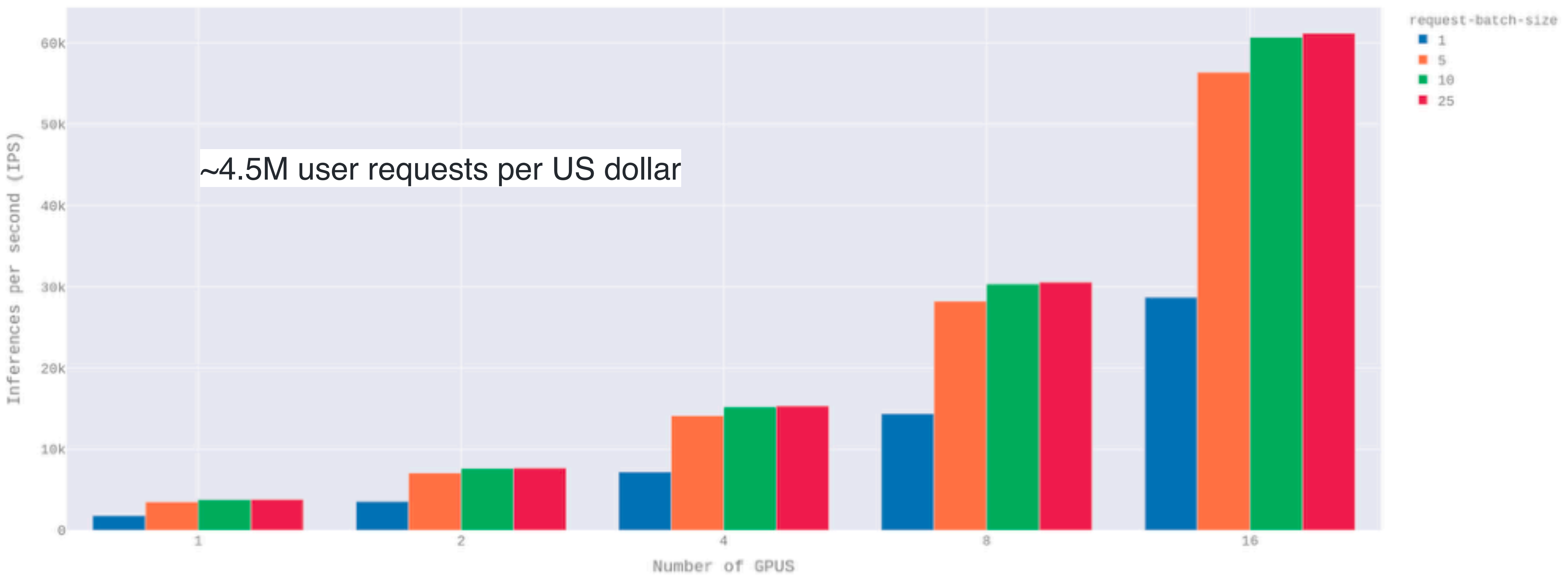
* Understand / measure the latency-throughput curve

* Cost
    * Cost-performance-headache landscape
    * Accelerators cost (a lot) more!
    * Use hardware designed for inference, not training.

* Set SLAs for 3rd party APIs
    * Define pXX latency bounds (and enforce them)



Virtex server performance for BERT base embeddings


~4.5M user requests per US dollar