Scaling dialogue systems involve many considerations

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

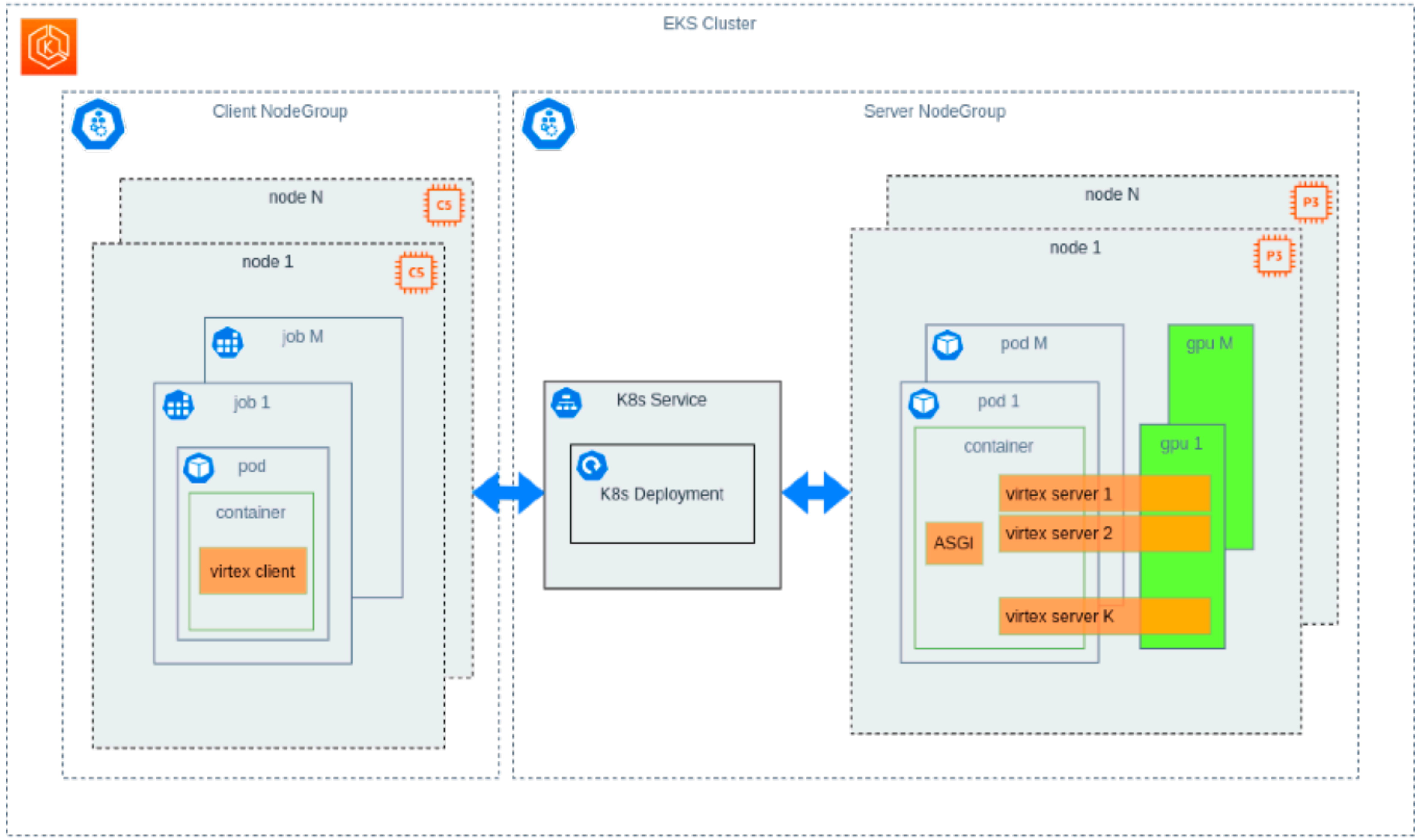
* Compute-bound services should be isolated
    * Dynamic batch processing

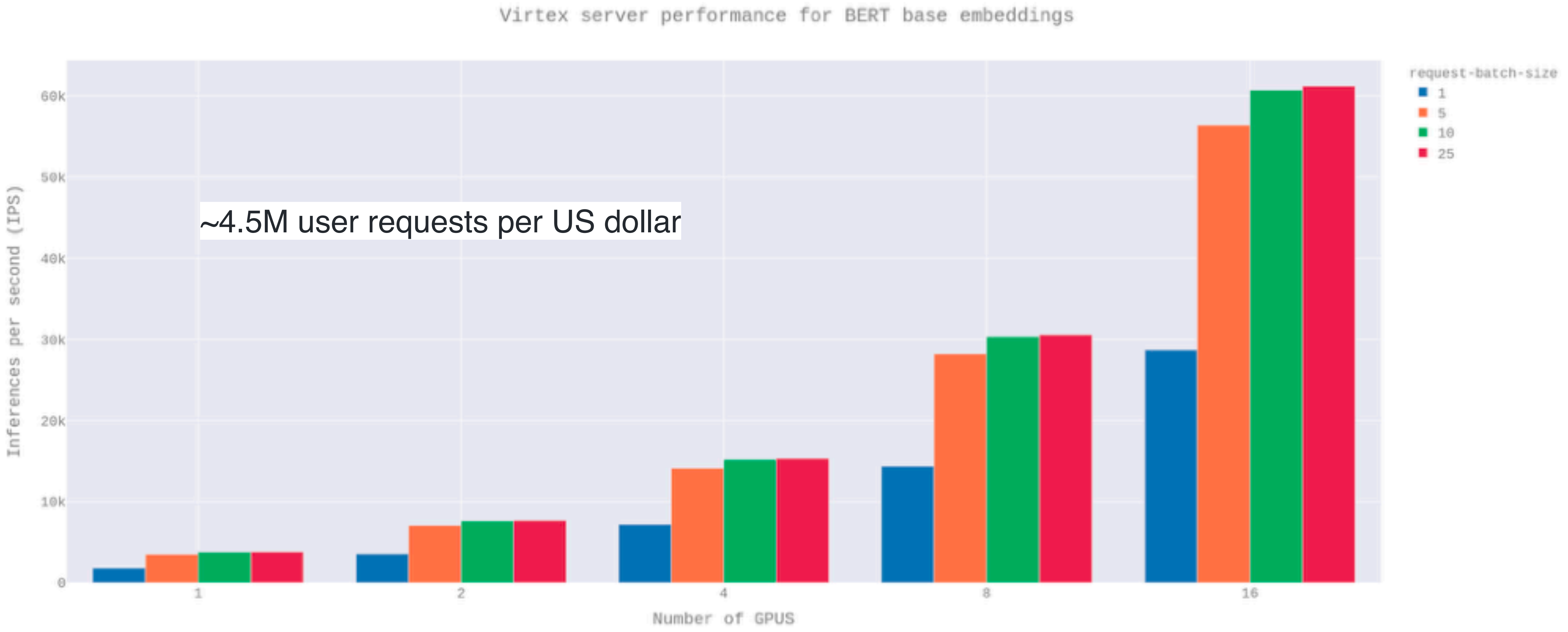
* Unit testing, including 3rd party tools

* 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)





~4.5M user requests per US dollar

Current state of the art in task-oriented dialogue

Calling...

Google Duplex