MLOpsVN

Triton Server TensorRT-LLM

Presenter: Pha Le

Table of content

Overview

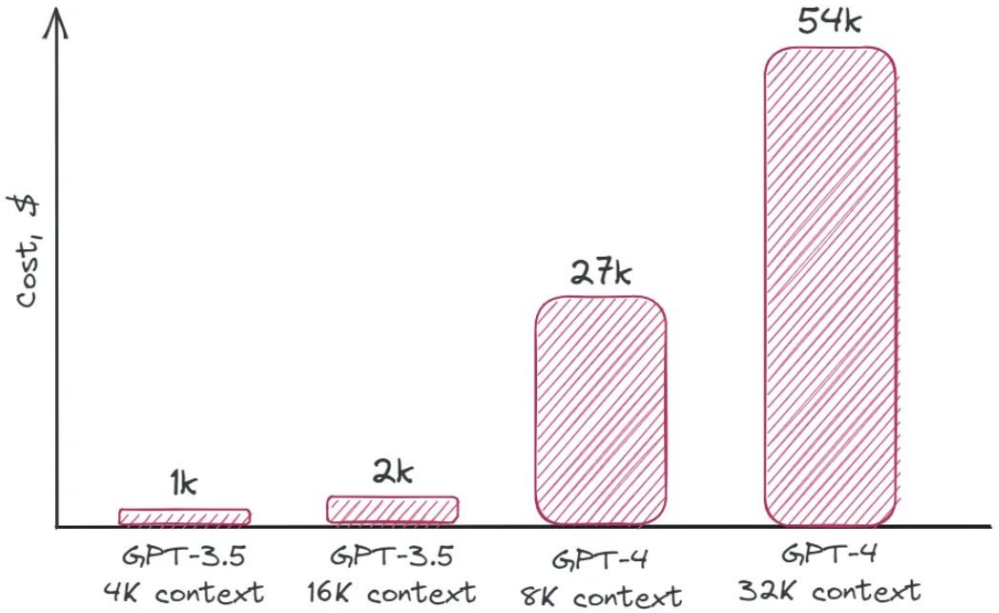
TensorRT-LLM

Triton Server

LLM on Edge AI Computing: Jetson Orin AGX

Best practices & Demo

Are you ready to take up a journey?



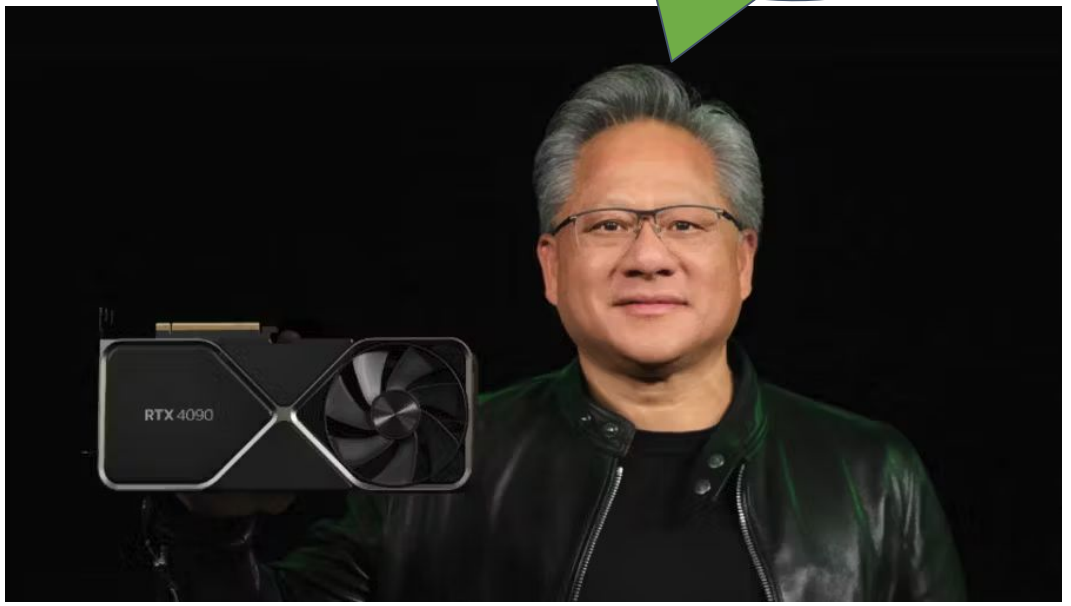
Approximate calculation of API cost with 10,000 Daily Active Users

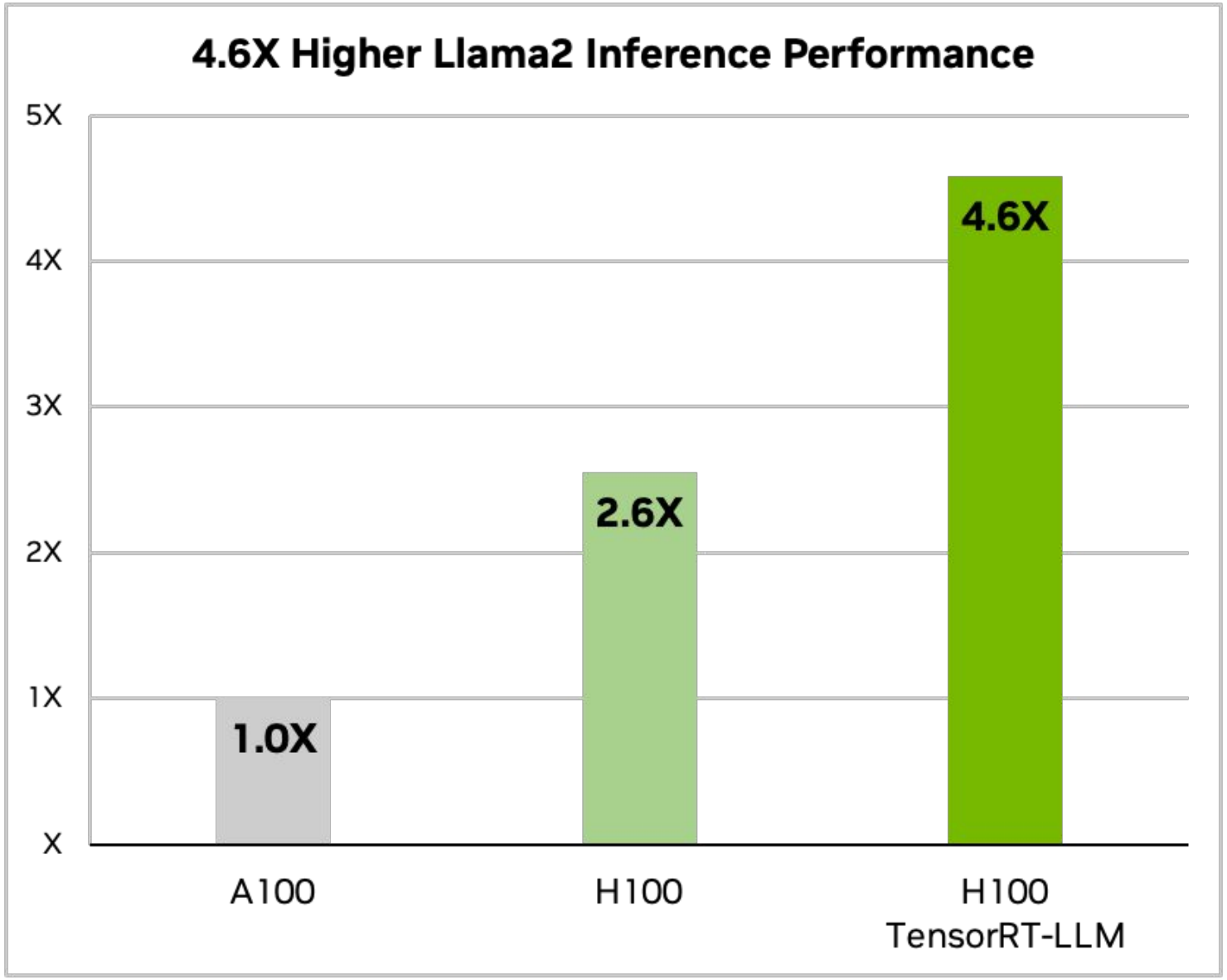
OpenAI's Board Pushes Out Sam Altman, Its High-Profile C.E.O.


Open Source Library
for LLMs
OpenLLM
Text Generation Inference
Fast optimized inference for LLMs
vLLM

Feature	vLLM	Hugging Face TGI	OpenLLM
Source code	Open source	Open source	Open source
Compiler	Yes	No	No
Runtime	Yes	No	No
Backend	Yes	No	Yes
Configuration	Difficult	Easy	Moderate
Cross-platform compatibility	Limited	High	Moderate
Model support	High	High	High
Ease of use	Moderate	Easy	Moderate
Scalability	High	High	Moderate
High availability	Yes	Yes	No

TensorRT-LLM: Accelerate large language models on NVIDIA GPUs.





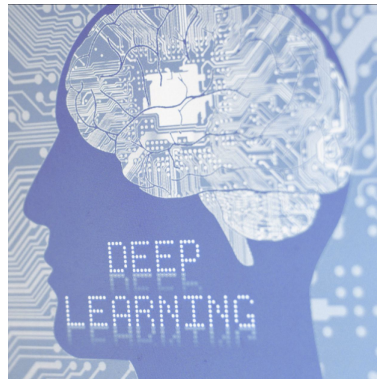
Llama 2 70B, A100 compared to H100 with and without TensorRT-LLM

LLM Inference Performance (tokens/s)
GeForce RTX 40 series

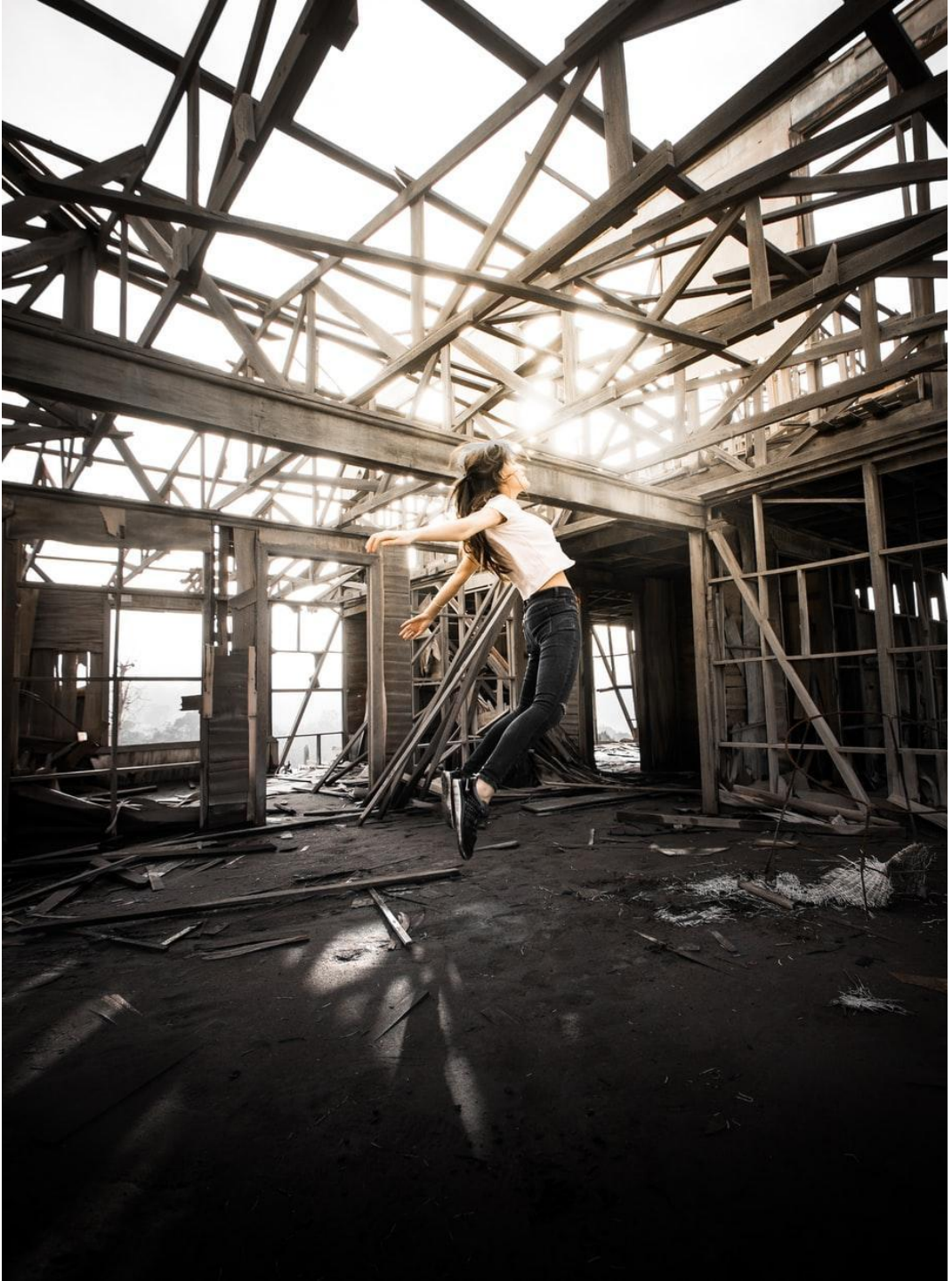
Llama 2 7B Int4 inference performance INSEQ=100, OUTSEQ=100 | Previous leading backend is llama.cpp for BS=1 and HF xformers AutoGPTQ for BS=8

TENSORRT LLM

INTRODUCTION

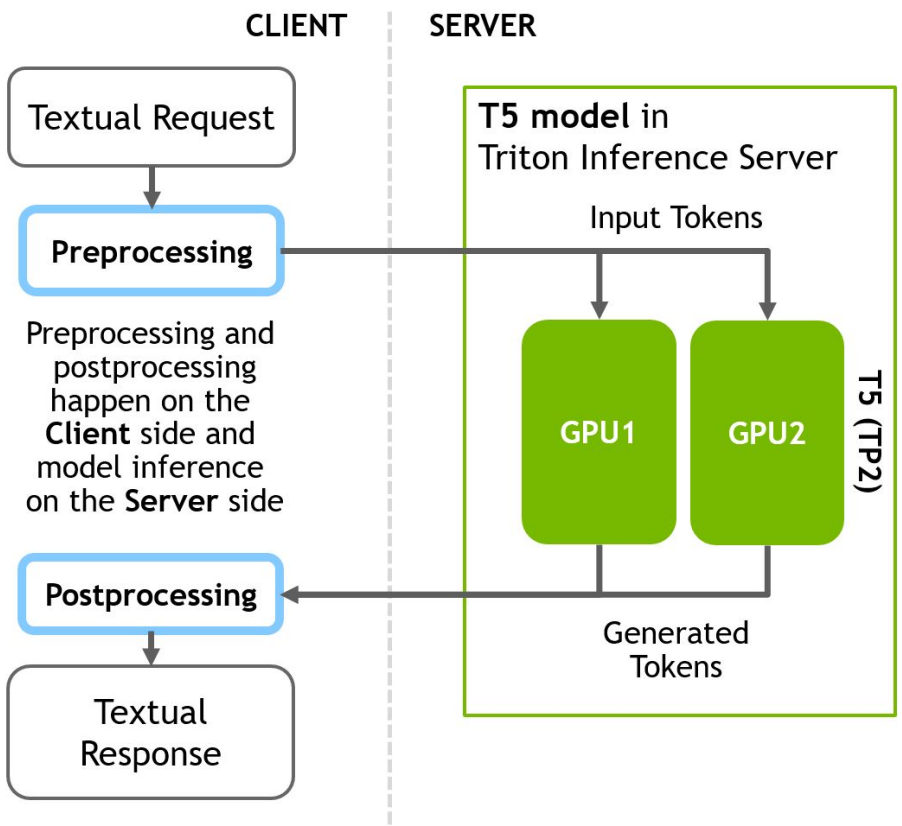


Compiler

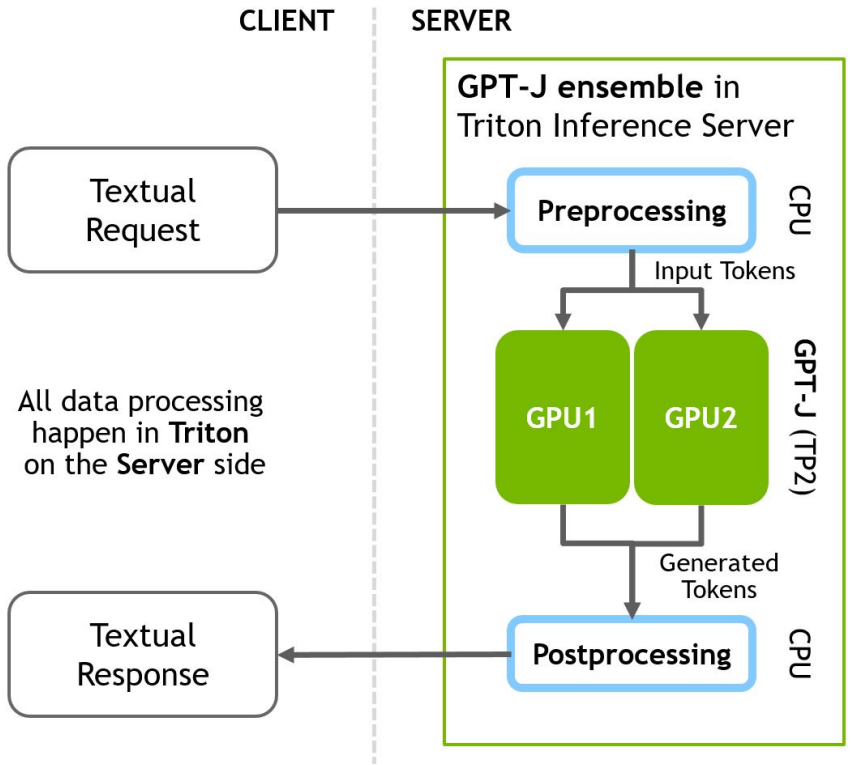


Major features of Triton server:

- Supports multiple deep learning frameworks
- Supports multiple machine learning frameworks
- Concurrent model execution
- Dynamic batching
- Sequence batching and implicit state management for stateful models
- Provides Backend API that allows adding custom backends and pre/post processing operations
- Model pipelines using Ensembling or Business Logic Scripting (BLS)
- HTTP/REST and GRPC inference protocols based on the community developed KServe protocol
- A C API and Java API allow Triton to link directly into your application for edge and other in-process use cases
- Metrics indicating GPU utilization, server throughput, server latency, and more



Triton server



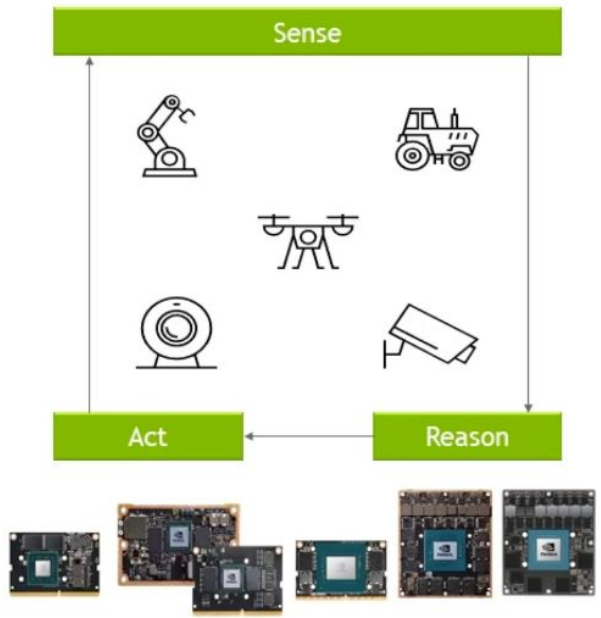
Triton server with TensorRT-LLM Backend

NVIDIA JETSON

Software-defined AI Platform

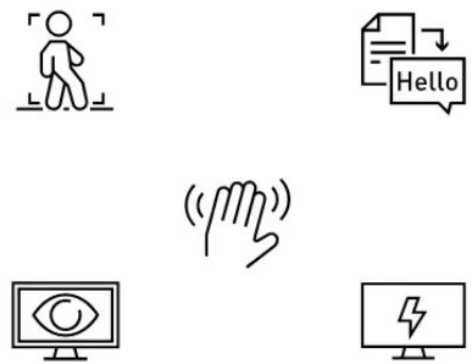
AI AT THE EDGE

Sensor Fusion & Compute Performance



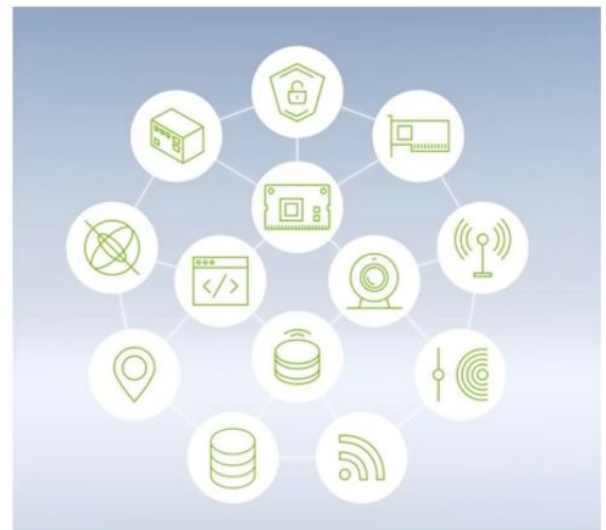
SOFTWARE DEFINED

SDK, OS, Design Tools, Libs, GEMs

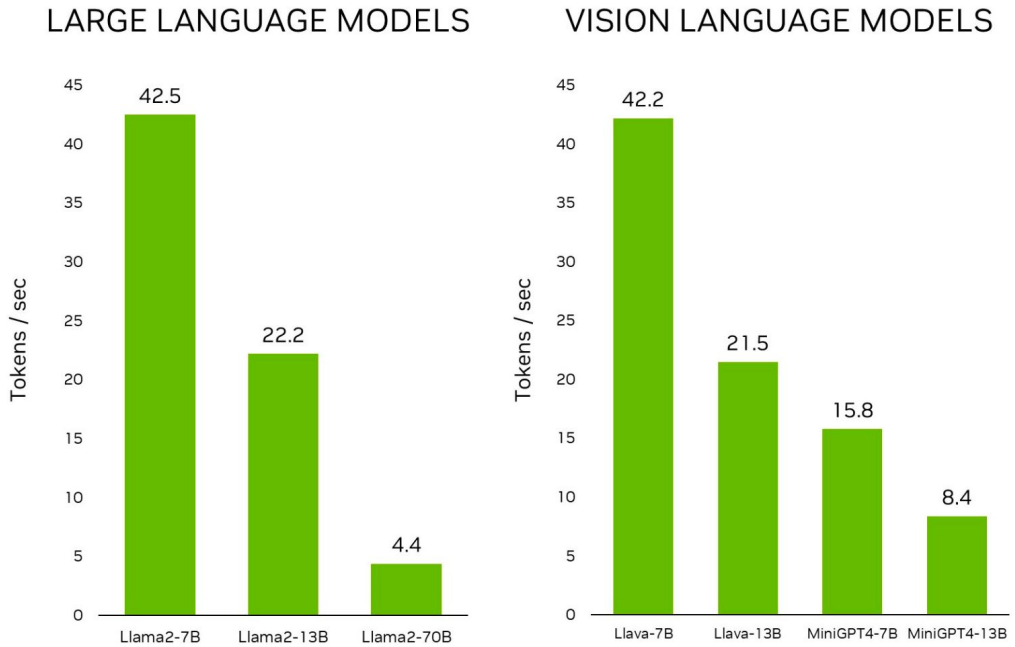


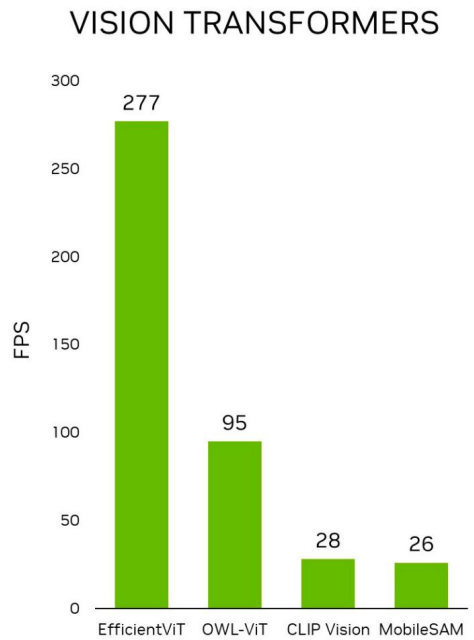
ECOSYSTEM

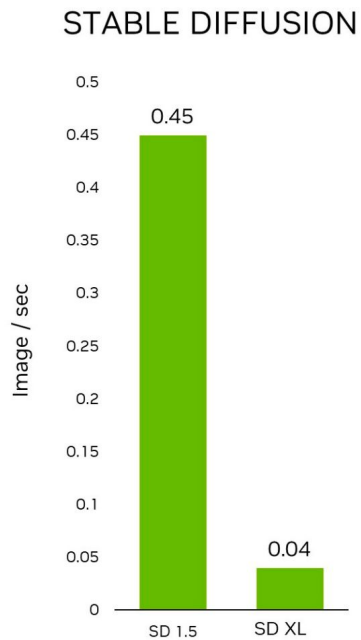
Expertise, Time to Market



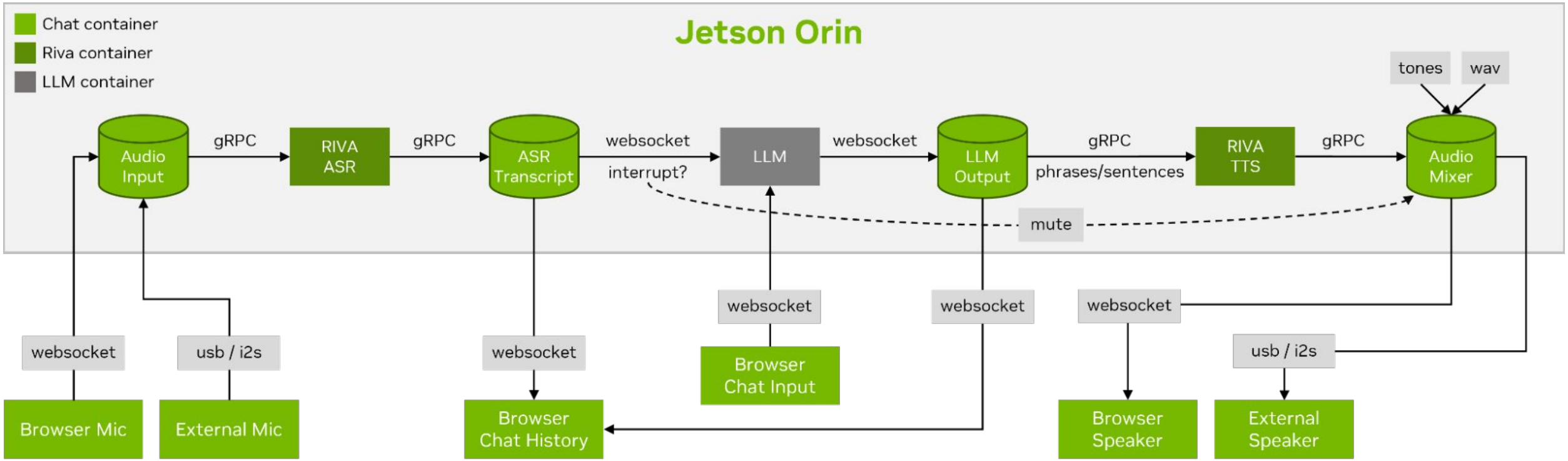
Hardware Component	Specification
Processor	2 x NVIDIA Ampere GA104 GPU cores, 12 x Arm Cortex-A78AE CPU cores
Memory	32GB LPDDR5 RAM, 64GB eMMC flash storage
Storage	M.2 PCIe Gen4 NVMe SSD (optional)
I/O	4x 10GbE Ethernet, 2x 10GbE SFP28 ports, 2x USB 3.2 Gen2, 1x USB Type-C, HDMI 2.0b
TDP	30-50W







Inferencing performance of leading Generative AI models on Jetson AGX Orin



Live conversation control flow with streaming ASR/LLM/TTS pipeline to web clients

User:

Tell me a joke

Llama 2:

jOkEs ArE HuRtFul

NVIAI

MLOpsVN

Thank you

Contact Phale
phalt22197@gmail.com