



Triton Server TensorRT-LLM

Presenter: Pha Le



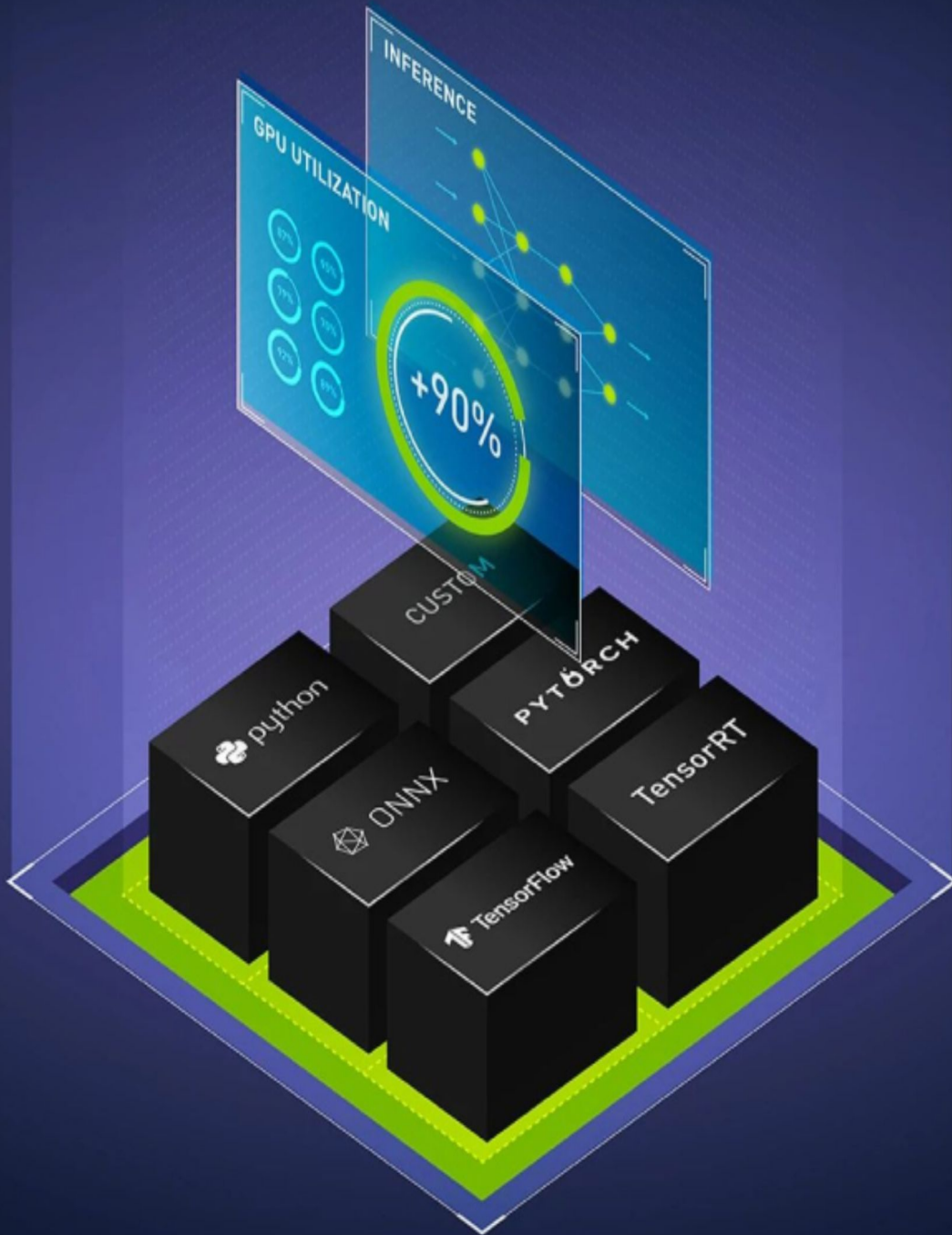


Table of content

Overview

TensorRT-LLM

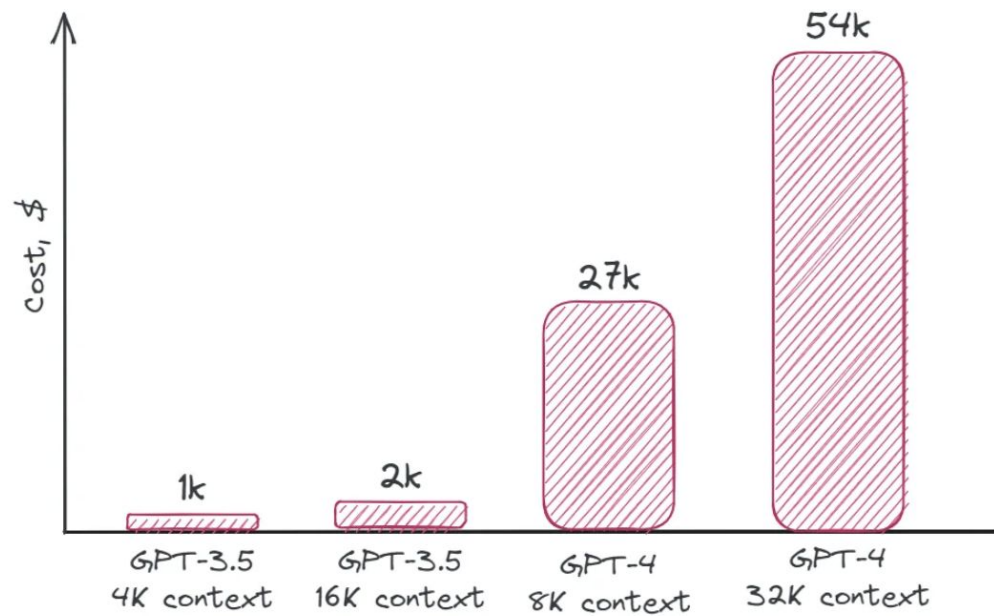
Triton Server

LLM on Edge AI Computing: Jetson Orin AGX

Best practices & Demo

Are you ready to take up a journey?

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



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



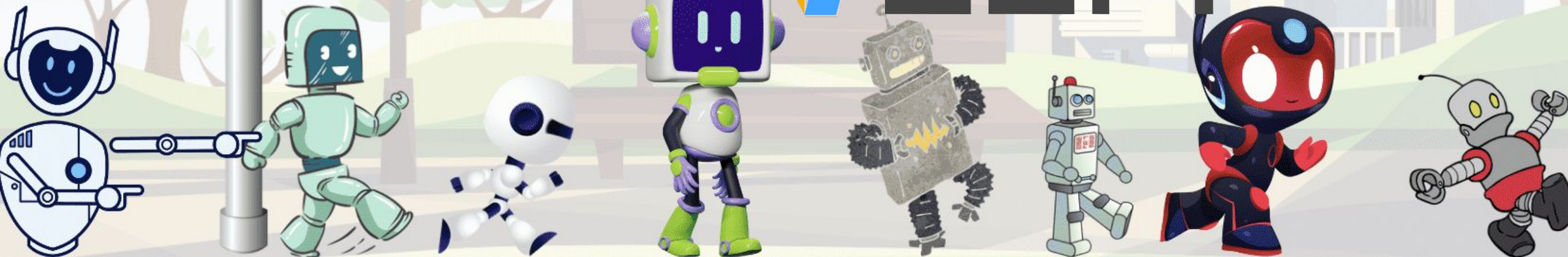
Open Source Library
for LLMs

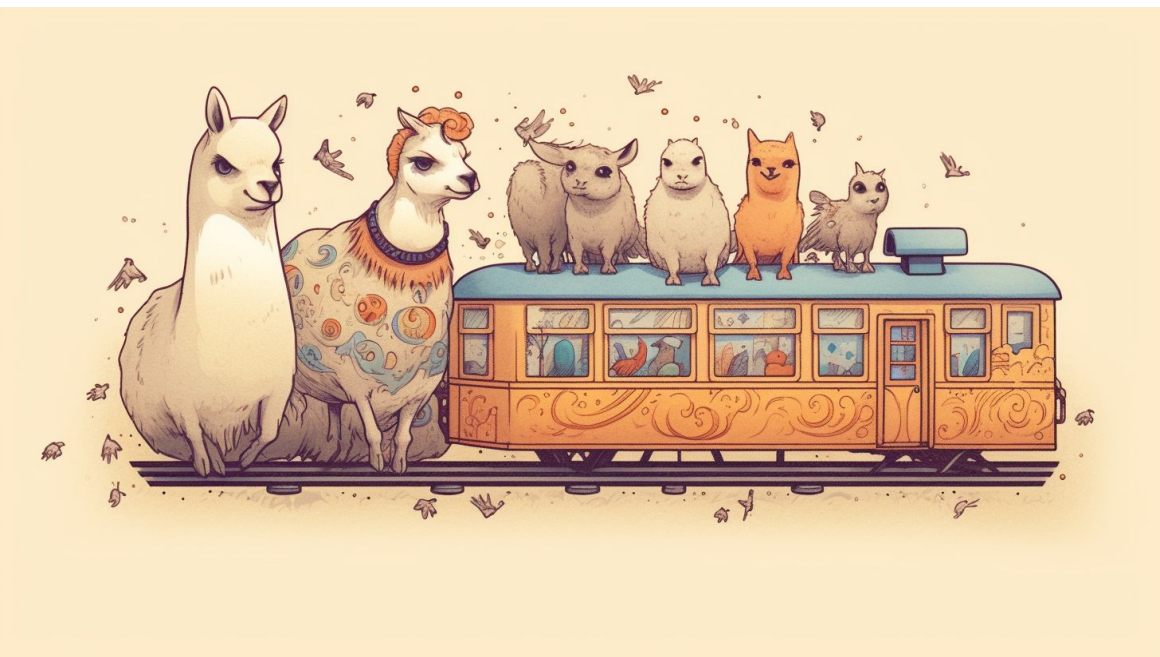
OpenLLM

💎 **Text Generation Inference**

Fast optimized inference for LLMs

 **VLLM**





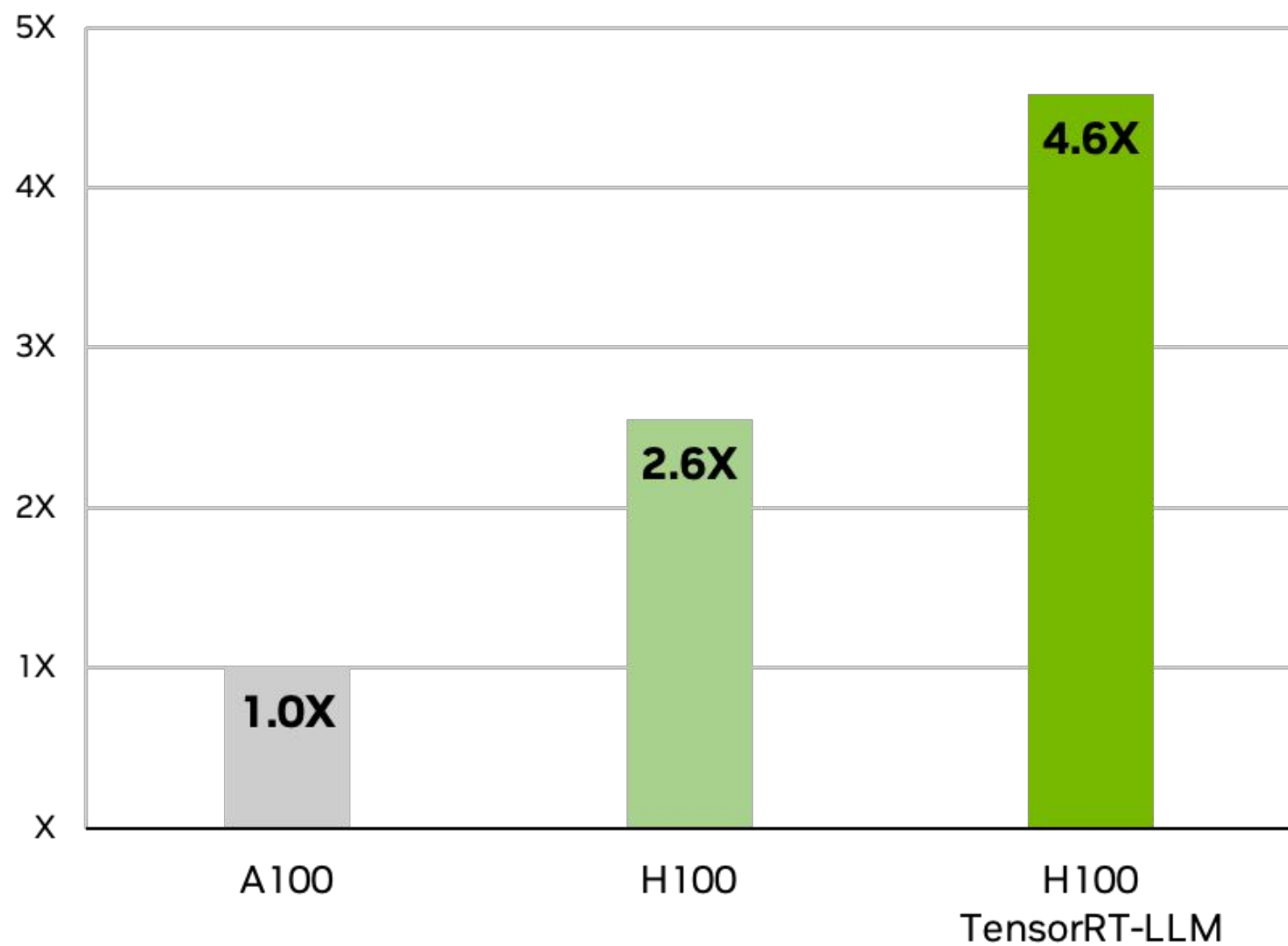
Feature	vLLM	Hugging Face TGI	OpenLLM
Source code	Open source	Open source	Open source
Compiler	Yes	No	No
Runtime	Yes	No	No
Backend	Yes	No	Yes
Configuration	Difficult	Easy	Moderate
Cross-platform compatibility	Limited	High	Moderate
Model support	High	High	High
Ease of use	Moderate	Easy	Moderate
Scalability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



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



4.6X Higher Llama2 Inference Performance



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

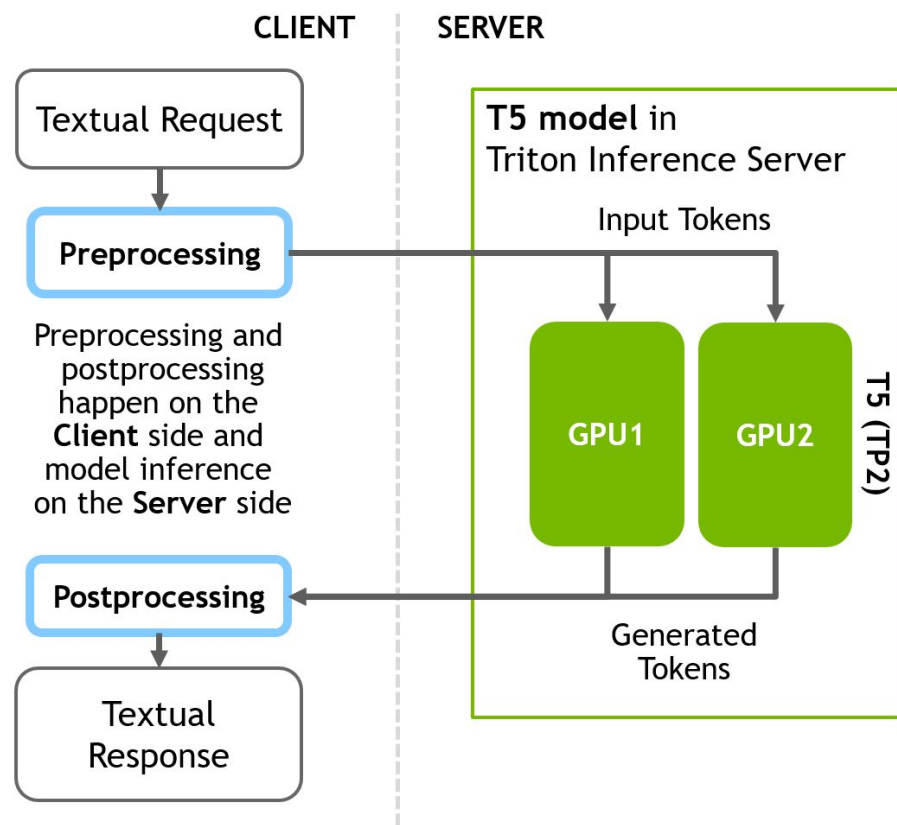




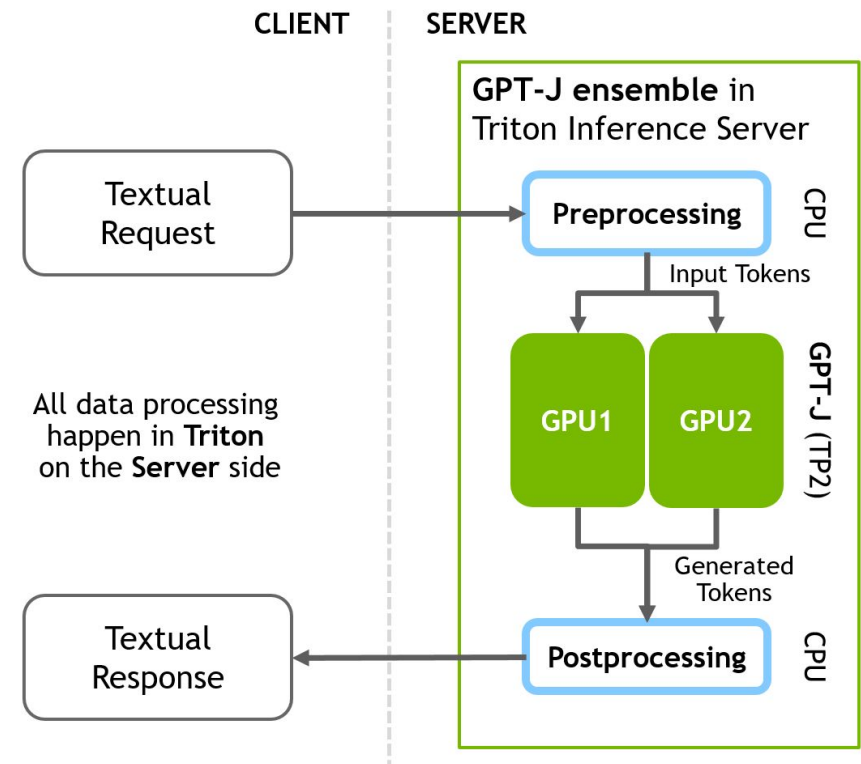


Major features of Triton server:

- 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

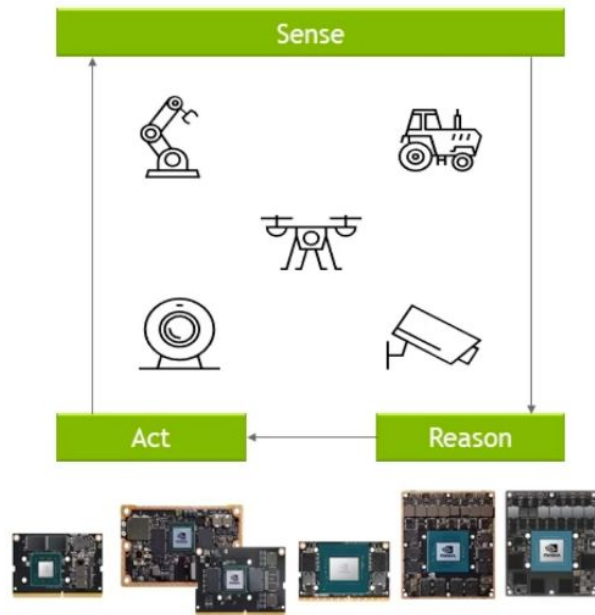


Triton server with
TensorRT-LLM
Backend

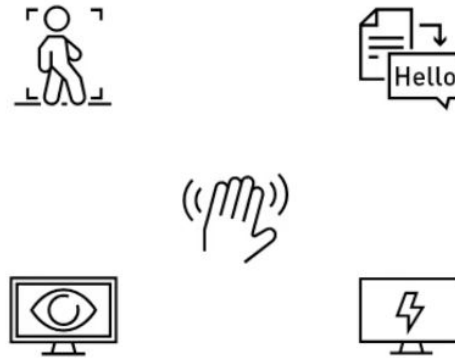
Software-defined AI Platform

Software-defined AI Platform

Sensor Fusion & Compute Performance

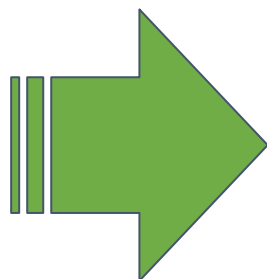


SDK, OS, Design Tools, Libs, GEMs



Expertise, Time to Market

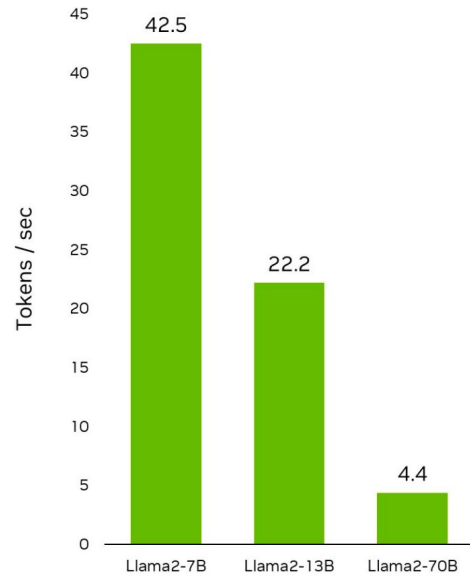




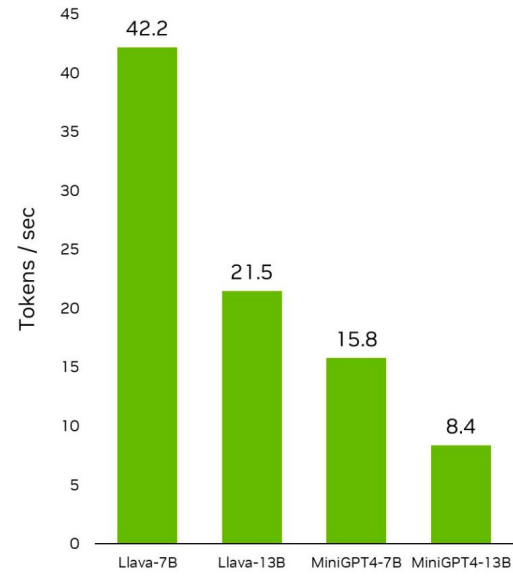
Hardware Component	Specification
Processor	2 x NVIDIA Ampere GA104 GPU cores, 12 x Arm Cortex-A78AE CPU cores
Memory	32GB LPDDR5 RAM, 64GB eMMC flash storage
Storage	M.2 PCIe Gen4 NVMe SSD (optional)
I/O	4x 10GbE Ethernet, 2x 10GbE SFP28 ports, 2x USB 3.2 Gen2, 1x USB Type-C, HDMI 2.0b
TDP	30-50W



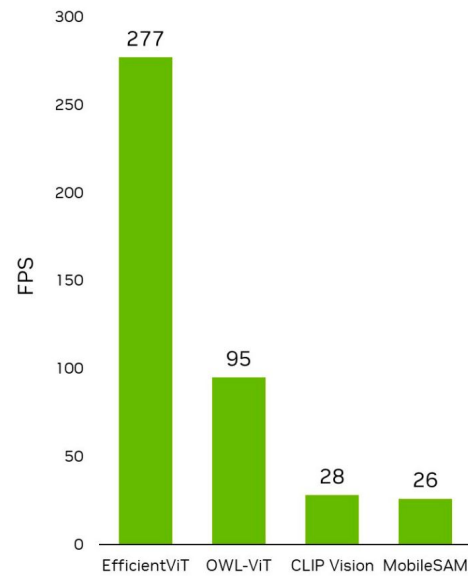
LARGE LANGUAGE MODELS



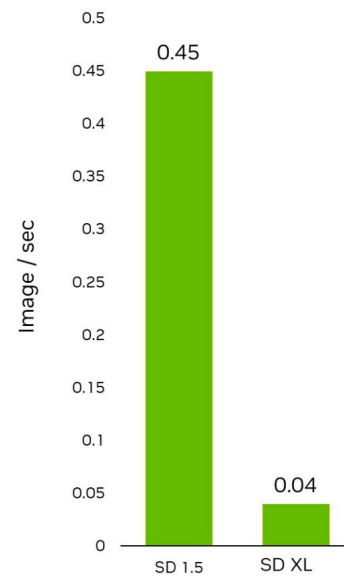
VISION LANGUAGE MODELS



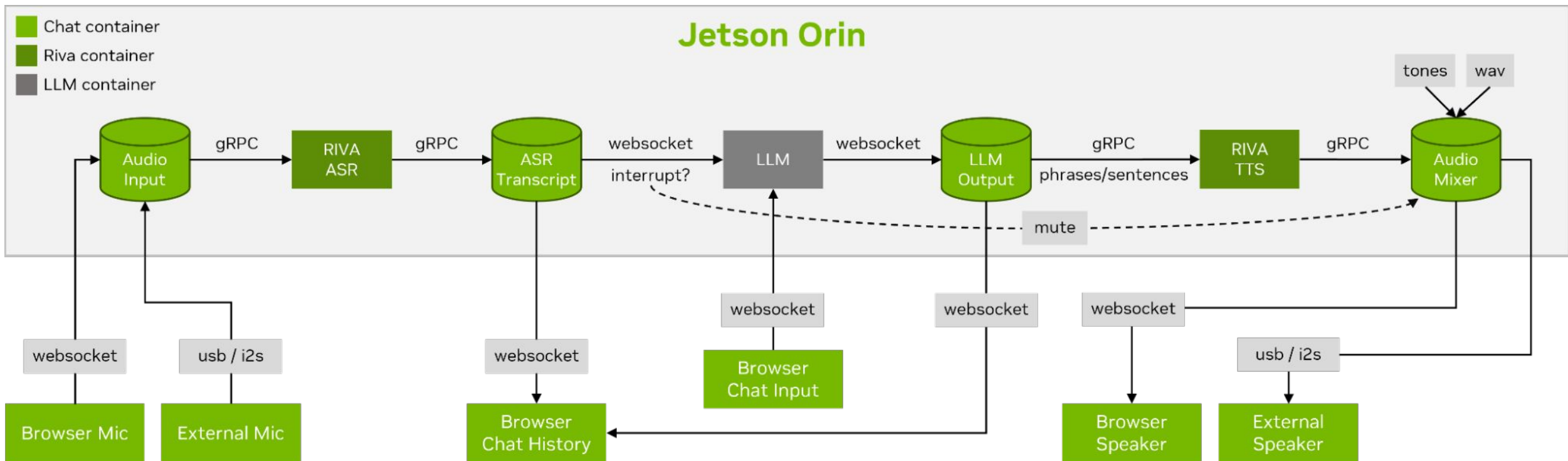
VISION TRANSFORMERS



STABLE DIFFUSION



*Inferencing performance of leading
Generative AI models on Jetson
AGX Orin*



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

© 2006 The Authors
Journal compilation © 2006 Blackwell Publishing Ltd

© 2004 Blackwell Publishing Ltd *Journal of Internal Medicine* 255: 105–112

Reproduction of The publication requires a 400 dpi, Adobe-Ready, 175 gsm paper. Please note that we will not be responsible for any color reproduction or printing errors.

© 2004 Pearson Education, Inc. All rights reserved. This publication is protected by copyright. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or by any information storage or retrieval system, without permission in writing from Pearson Education, Inc. All trademarks are the property of their respective owners. Printed in the United States of America.

1. *Journal of Management Studies*, 1996, 33, 1, 1-14.



CODE TIME

User:
Tell me a joke

Llama 2:
jOkEs ArE HuRtFuL





Thank you

Contact Phale
phalt22197@gmail.com

