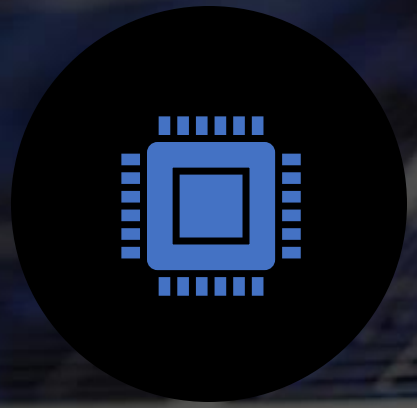
Inference Speed and Accuracy of Large Language Model

- Presented by
- Kandimalla Hemanth

- 02-09-2024

Throughput

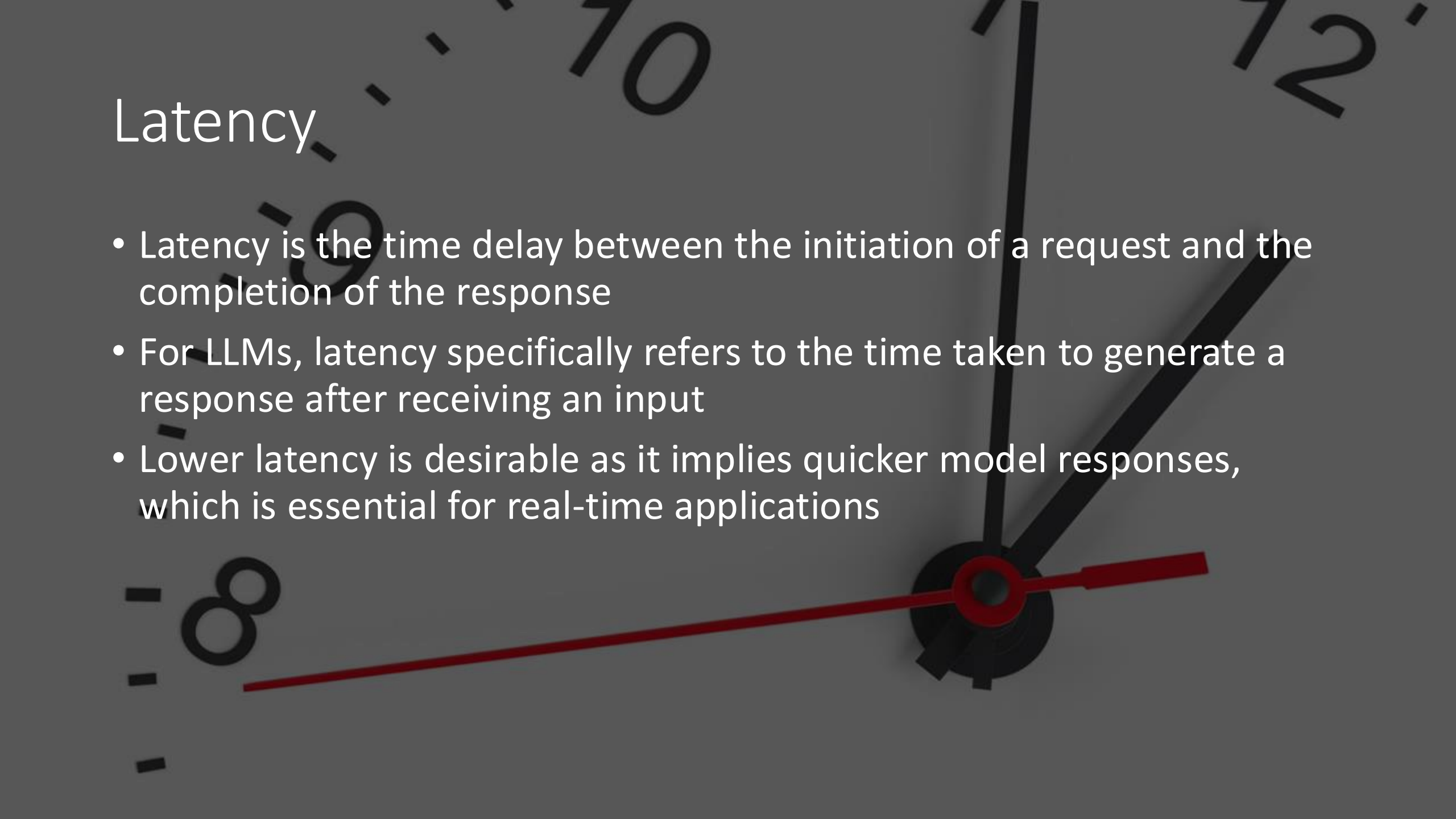


THROUGHPUT REFERS TO THE NUMBER OF TASKS, OPERATIONS, OR DATA PROCESSED PER UNIT OF TIME BY A SYSTEM

LLMS AND HARDWARE POINT OF VIEW, IT MEASURES THE VOLUME OF DATA THAT CAN BE PROCESSED EACH TIME, OFTEN EXPRESSED IN OPERATIONS PER SECOND

HIGH THROUGHPUT INDICATES EFFICIENT PROCESSING AND IS CRITICAL FOR TASKS LIKE TRAINING AND INFERENCE IN LLMS

Latency

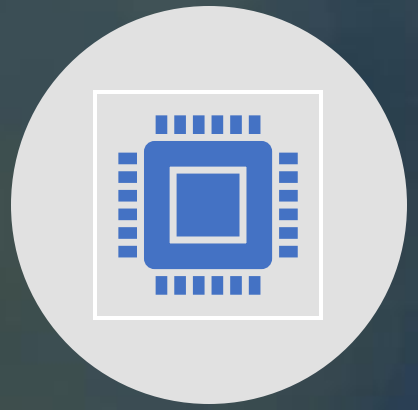


- Latency is the time delay between the initiation of a request and the completion of the response
- For LLMs, latency specifically refers to the time taken to generate a response after receiving an input
- Lower latency is desirable as it implies quicker model responses, which is essential for real-time applications

Bandwidth

BANDWIDTH DESCRIBES THE MAXIMUM RATE OF DATA TRANSFER ACROSS A NETWORK OR BETWEEN COMPONENTS IN A SYSTEM

IN THE CONTEXT OF LLMS, IT REPRESENTS THE DATA TRANSMISSION CAPACITY, ESPECIALLY BETWEEN GPUS, MEMORY, AND STORAGE



HIGH BANDWIDTH ALLOWS FASTER ACCESS TO DATA, WHICH IS CRUCIAL FOR TRAINING AND DEPLOYING LARGE MODELS EFFICIENTLY

Text Generation Inference

- Used in production at Hugging Face to power Hugging Chat, the Inference API and Inference Endpoint
- Text Generation Inference is a toolkit for deploying and serving Large Language Models
- TGI enables high-performance text generation for the most popular open-source LLMs , including Llama, Falcon, Star Coder, BLOOM, GPT-Neo X, …

Text Generation Inference

- Simple launcher to serve most popular LLMs efficiently and quickly.
- Tensor Parallelism for faster inference on multiple GPUs, enhancing performance.
- Token streaming using Server-Sent Events for real-time data processing.
- Continuous batching of incoming requests increases total throughput significantly.
- Messages API compatible with Open AI Chat Completion API seamlessly.
- Optimized transformers code for inference using Flash and Paged Attention.
- Quantization with bits bytes, GPT-Q, EETQ, AWQ, Marlin, fp8, Safe tensors.
- Watermarking with A Watermark for Large Language Models effectively.
- Logits warper, stop sequences, log probabilities, speculation ~2x latency reduction.
- Guidance/JSON support for structured data handling and improved functionality.