CPU vs GPU vs TPU

14 December 2025 10:53

CPU: Central Processing Unit

- general purpose processor in every computer
- optimized for single threaded or moderately parallel operations
- excels in low-latency tasks and handling control logic.

use-cases: Browsing internet, using office products, small-scale data processing, light ML models

Feature	Description
Cores	Few (2–32) but very powerful
Best for	General computing, light ML tasks, logic-heavy operations
Latency	Very low
Parallelism	Low
Memory Access	Fast and flexible
Cost	Cheapest

Example of Common Core Configurations:

- **Dual-Core**: 2 cores (e.g., Intel Core i3)
- **Quad-Core**: 4 cores (e.g., Intel Core i5, AMD Ryzen 5)
- **Hexa-Core**: 6 cores (e.g., Intel Core i7, AMD Ryzen 5/7)
- **Octa-Core**: 8 cores (e.g., Intel Core i7/i9, AMD Ryzen 7/9)
- **Deca-Core**: 10 cores (e.g., Intel Core i9, AMD Ryzen 9)
- **High-End Multi-Core**: Up to 64 cores (e.g., AMD Ryzen Threadripper, Intel Xeon)

```
NumberOfCores  NumberOfLogicalProcessors
8              16
```

Limitations

- slower at massive parallelism
- limited no. of cores for parallel tasks.

GPUs: Graphics Processing Unit

- designed for graphics/video editing needs in early day and even now

AI Overview

Listen

The Interstellar black hole (Gargantua) rendering was a groundbreaking VFX feat, using **Kip Thorne's general relativity equations** to simulate light bending around it, creating the most accurate black hole visualization ever, requiring 100+ hours per frame on custom software, resulting in stunningly realistic gravitational lensing that actually aided real science later.

This video shows the creation of Interstellar's black hole:

watch this movie!

- helps to do parallel processing → matrix operations at scale.
- GPUs were originally made for 3D arrays' rendering graphics, now being used for AI, ML, DL, LLMs etc.
- GPUs are optimized for throughput over latency

Feature	Description
Cores	Thousands of lightweight cores
Best for	Matrix operations, large ML model training / DL / LLMs.
Latency	Moderate
Parallelism	Very high
Memory Access	High bandwidth, but limited flexibility
Cost	Moderate to high

Use-cases: Training big NN models, Image processing, video editing & rendering

→ Apple

TPUs : Tensor Processing Units

- Specialized hardware built by Google just for TensorFlow framework to practice deep learning at scale.
- specialized tensorflow operations
- extremely fast at matrix-heavy deep learning workloads.

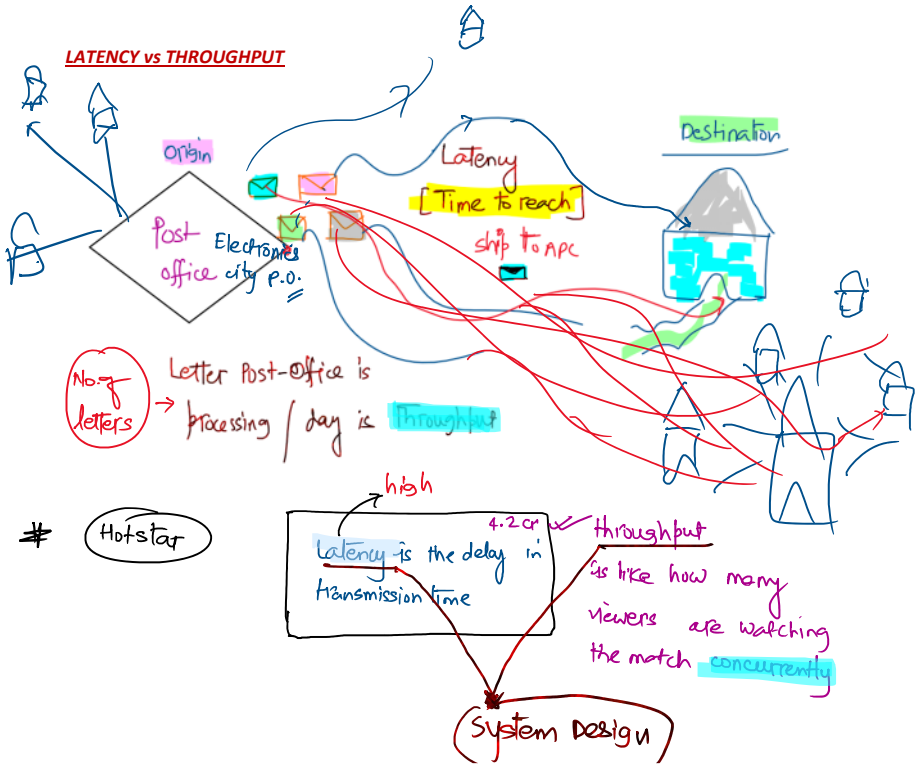
Feature	Description
Cores	Designed specifically for tensor computations
Best for	TensorFlow-based DL workloads (training + inference) ← TF
Latency	Very low for tensor ops → operations
Parallelism	Ultra high
Memory Access	Optimized for tensor operations
Cost	Pay-as-you-go (Google Cloud), not for local use

Google cloud Platform (GCP)
- Forecasting Models
- Data ware house
- Build forecasting agents → Agentic AI. → NVIDIA GPUs, TPUs

[Introducing 7th Generation TPUs: Ironwood]

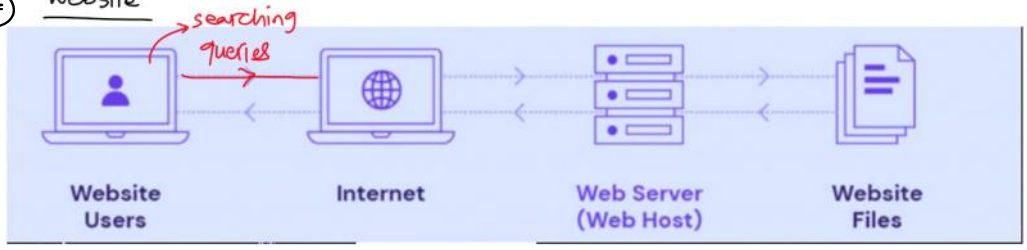
TASK: Read about NPUs

LATENCY vs THROUGHPUT



System Design

Website

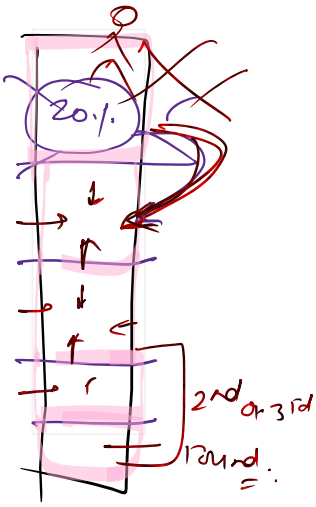


on an avg. → how much time is taken to respond a query → LATENCY

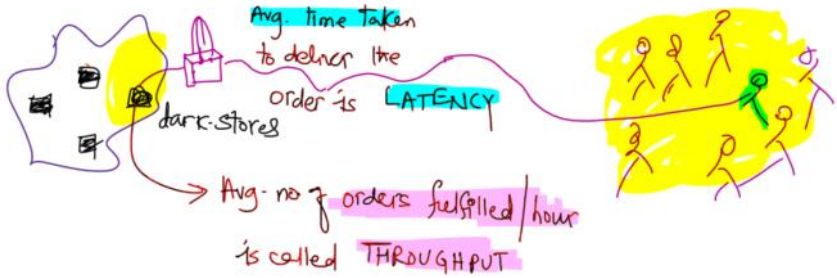
↳ how many queries are being resolved/day → THROUGHPUT

IRCTC Website

↳ latency is high → VERYYY High
Throughput is low → veryy Low
] during tatkal booking!!



Zepto | Blinkit (Quick commerce)



[LOW LATENCY & HIGH THROUGHPUT] → Desirable state for any system.

Do we need CPU if the machine has the latest & updated CPU?

Hell, yeah!

Cafe