DIGITAL ELECTRONICS AND LOGIC DESIGN

Unit V — Emerging Processors for AI & Heterogeneous Computing

Topics Covered in this Unit

Taxonomy of AI-Focused Processors: CPU, GPU, TPU (Systolic Arrays), NPU & VPU Architectures
Comparative Hardware Analysis: Intel Core Ultra (Meteor Lake), Apple M3/M4, NVIDIA RTX, Google TPU v4/v5p, AMD Ryzen AI & Qualcomm Hexagon
Edge AI vs. Cloud AI Processing Paradigms (Latency, Bandwidth, Privacy, Energy & Scalability)
Heterogeneous Computing Architectures: CPU + GPU + NPU Workload Orchestration & Unified Memory Spaces
AI Performance Benchmarks & Hardware Metrics: FLOPS (TFLOPS/PFLOPS), TOPS (INT8), Latency vs. Throughput & Power Efficiency (TOPS/Watt)
Hardware Quantization & Sparsity Acceleration (FP32 → FP16 → INT8 → INT4 Precision Scaling)
High-Yield University Exam Question Bank with Detailed Model Answers & Unit Summary

Compiled in a structured, easy-to-understand, textbook style format
Specially curated for B.Tech / B.E. / BCA / MCA / B.Sc Electronics, Electrical, CS & AI Students

COURSE SYLLABUS & MAPPING

[UNIT V SYLLABUS — PRESCRIBED UNIVERSITY CURRICULUM]

This section contains the official syllabus for Unit V. Verify with your university curriculum mapping.

Unit V Syllabus Breakdown:

Unit V Emerging Processors for AI: Introduction to AI-focused processors: CPU(Central Processing Unit), GPU(Graphics Processing Unit), TPU(Tensor Processing Unit), NPU(Neural Processing Unit), VPU(Vision Processing Unit). Comparison: Intel Core Ultra, Apple M3, NVIDIA RTX, Google TPU, AMD Ryzen AI, Qualcomm Hexagon. Concepts of Edge AI & Cloud AI processing. Heterogeneous computing (CPU+GPU+NPU). Benchmarks& Metrics: FLOPS, TOPS, latency, power efficiency

• Course Code: _____ • Semester / Year: _____

TABLE OF CONTENTS

5.1 Taxonomy of AI Processors: CPU, GPU, TPU, NPU, and VPU

Definition: AI Hardware Accelerators

AI Accelerators are specialized microelectronic processing units designed specifically to perform the massively parallel tensor mathematical operations (Matrix-Vector Multiplications and Convolutions) required by deep neural networks with orders of magnitude higher throughput and energy efficiency than general-purpose CPUs.

Processor Type	Architectural Paradigm	Core Computational Strength	Primary AI Role & Domain
CPU (Central Processing Unit)	Low-latency scalar / small-SIMD architecture with large caches and out-of-order execution.	Sequential logic, high single-thread clock speed, OS control flow.	Sequential pre-processing, data loading, small model inference.
GPU (Graphics Processing Unit)	Massively parallel SIMT (Single Instruction, Multiple Threads) with thousands of ALUs and high-bandwidth VRAM.	Massive parallel floating-point matrix multiplication (FP32/FP16).	Large-scale deep learning model training and cloud inference (NVIDIA H100, RTX 4090).
TPU (Tensor Processing Unit)	Custom Google ASIC built around a 2D Systolic Array matrix multiply unit (bfloat16 & INT8).	Continuous data streaming through 2D grid of multiply-accumulate (MAC) cells without register re-reads.	Massive enterprise cloud AI training and serving (Google Search, Gemini, AlphaFold).
NPU (Neural Processing Unit)	Dedicated low-power domain-specific accelerator optimized for low-precision INT8/INT4 neural network inference.	High TOPS/Watt energy efficiency, fixed activation functions, on-chip SRAM.	On-device Mobile AI, smartphone photography, Windows Copilot+ AI PC inference.
VPU (Vision Processing Unit)	Energy-efficient programmable vector processor with dedicated hardware image signal processing (ISP) pipelines.	Parallel 2D/3D image convolutions, edge detection, optical flow.	Drones, security cameras, AR/VR smart glasses, autonomous driving vision.

5.2 Comparative Analysis of Commercial AI Processors

Processor Family	Manufacturer	Key AI Architectural Features	Claimed AI Compute (TOPS) & Target Market
Intel Core Ultra (Meteor Lake)	Intel	Disaggregated 3D Foveros tile architecture: CPU + Intel Arc GPU + integrated 2-engine NPU.	Up to 34 TOPS (Total Platform). Mainstream AI Laptops & PCs.
Apple Silicon (M3 / M4)	Apple	Unified Memory Architecture (up to 128GB @ 400 GB/s) + 16-Core Neural Engine (ANE).	Up to 38 TOPS (ANE alone). Creative professional workstations & Apple Intelligence.
NVIDIA GeForce RTX 40-Series	NVIDIA	Ada Lovelace architecture: 4th-Gen Tensor Cores with FP8 Transformer Engine & Optical Flow.	Up to 1,300+ AI TOPS (RTX 4090). High-end desktop AI development & local LLM inference.
Google TPU (v4 / v5p)	Google	Custom 3D Torus Interconnect of thousands of dual-matrix multiply units with 32GB HBM2/HBM3.	Up to 459 TFLOPS (bfloat16) per chip. Enterprise cloud AI training supercomputing clusters.
AMD Ryzen AI (8040 Series)	AMD	XDNA Architecture based on adaptive Xilinx spatial AI engine array.	Up to 39 TOPS (Total Platform: 16 TOPS NPU). Ultra-thin Windows AI notebooks.
Qualcomm Hexagon (Snapdragon X Elite)	Qualcomm	Micro-tiled vector and scalar accelerators with dedicated high-speed low-power memory.	Up to 45 TOPS (NPU alone). Next-gen ARM-based Windows Copilot+ AI PCs.

5.3 Edge AI vs. Cloud AI Processing Paradigms

Evaluation Parameter	Edge AI (On-Device Inference)	Cloud AI (Data Center Processing)
Latency & Determinism	Ultra-low, deterministic latency (1–10ms); zero network packet transit delay.	Variable, non-deterministic latency (50–500ms) dependent on internet connection.
Privacy & Security	Maximum privacy; sensitive personal biometric and video data never leaves user device.	Data transmitted across public networks and stored on third-party cloud servers.
Bandwidth & Network Dependency	Operates 100% offline with zero internet connectivity required.	Requires continuous high-speed broadband network connection.

Evaluation Parameter	Edge AI (On-Device Inference)	Cloud AI (Data Center Processing)
Computational Capacity	Constrained by strict battery life, thermals (1W–30W), and memory capacity (4–32GB).	Virtually unlimited scalable clusters of thousands of multi-GPU servers (100kW+).

5.4 Heterogeneous Computing (CPU + GPU + NPU)

Definition: Heterogeneous Computing

Heterogeneous Computing refers to systems that integrate multiple specialized computational processing cores (CPU, GPU, and NPU) onto a single silicon die or package, intelligently dispatching sub-tasks to the processor best suited for that specific computational workload.

THE MODERN HETEROGENEOUS AI SYSTEM ARCHITECTURE

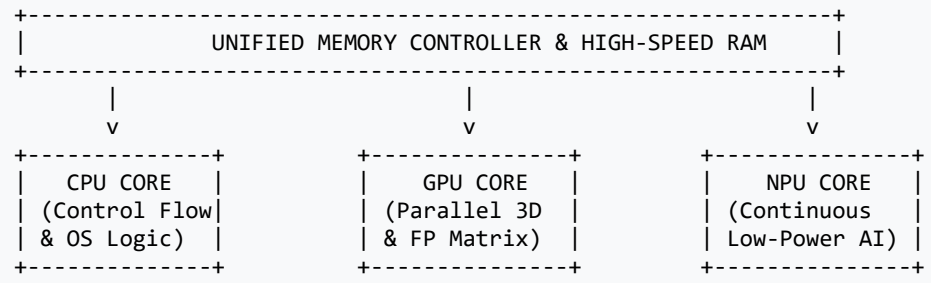

Figure: Intelligent task dispatch: CPU manages OS; GPU renders visuals/trains; NPU executes 24/7 background AI

5.5 AI Hardware Benchmarks & Performance Metrics

- **1. FLOPS (Floating-Point Operations Per Second):** Standard metric for general floating-point computing capacity. TFLOPS = 10^12 FLOPS; PFLOPS = 10^15 FLOPS. (Governed by: FLOPS = Cores × Clock Frequency × Operations/Cycle).
- **2. TOPS (Tera-Operations Per Second):** Primary industry metric for AI inference, measuring trillions of integer operations (typically INT8 MAC operations: 1 MAC = 2 Operations [1 Multiply + 1 Add]).
- **3. Power Efficiency (TOPS/Watt):** The definitive figure of merit for mobile and edge AI systems. High-end NPUs deliver 5 to 15 TOPS/Watt, compared to 0.5 TOPS/Watt for desktop CPUs.
- **4. Latency vs. Throughput:** Latency (time to process a single token/frame in milliseconds - vital for conversational voice AI) vs. Throughput (total tokens or image frames processed per second across concurrent batches).

5.6 High-Yield University Exam Question Bank & Model Answers

Part A: Short Answer Questions (2–3 Marks)

- **Q1. What is an NPU (Neural Processing Unit)? Why is it more efficient than a CPU for AI?**

Answer: An NPU is a specialized microchip designed exclusively for executing deep learning matrix multiplications and activations using quantized low-precision math (INT8/INT4). It eliminates CPU cache misses and instruction decode overhead, delivering 5–10x higher energy efficiency (TOPS/Watt).

- **Q2. Define TOPS and explain how it differs from FLOPS.**

Answer: FLOPS (Floating-Point Operations Per Second) measures floating-point arithmetic (FP32/FP64) common in graphics and science. TOPS (Tera-Operations Per Second) measures trillions of integer operations per second (INT8), which is the standard benchmark for quantized AI inference.

- **Q3. Differentiate between Edge AI and Cloud AI.**

Answer: Edge AI runs neural network inference locally on user devices with zero network latency, high privacy, and offline capability. Cloud AI processes models in massive centralized data centers, offering virtually unlimited compute capacity but requiring continuous broadband connectivity.

- **Q4. What is a Systolic Array in Google's TPU?**

Answer: A Systolic Array is a 2D matrix of multiply-accumulate (MAC) execution units where data values stream continuously from cell to cell in a rhythmic wave without writing intermediate results back to registers, drastically reducing memory bandwidth bottlenecks.

Part B: Long Answer Questions (8–10 Marks)

- **Q5. Compare CPU, GPU, TPU, and NPU architectures in detail. Discuss their internal execution paradigms, memory architectures, and suitability for AI Training vs. Inference.**

Model Answer Outline:

- 1. CPU Architecture: Large multi-level caches, branch predictors, out-of-order execution, low core count, optimized for sequential scalar tasks.
- 2. GPU Architecture: Thousands of small SIMT vector cores, high-bandwidth VRAM (GDDR6/HBM3), massive parallel throughput, standard for AI training.
- 3. TPU Architecture: ASIC custom matrix engine, 2D systolic array, native bfloat16 mixed precision support.
- 4. NPU Architecture: Spatial dataflow arrays, fixed activation units, on-chip SRAM, ultra-low power consumption (TOPS/Watt).

- 5. Comprehensive comparative table evaluating ALU density, cache percentage, power consumption, and Training vs Inference suitability.

- **Q6. Discuss Heterogeneous Computing for AI and explain key hardware performance metrics: FLOPS, TOPS, Latency, and Power Efficiency (TOPS/Watt).**

Model Answer Outline:

- 1. Definition and motivation for Heterogeneous Computing: Co-locating CPU, GPU, and NPU with Unified Memory.
- 2. Task orchestration: CPU handles OS logic and data prep; GPU handles high-throughput training/graphics; NPU runs low-power background neural inference.
- 3. Detailed definition and formulas for performance metrics: FLOPS (TFLOPS/PFLOPS), TOPS (INT8 MAC calculations), Latency (Time-to-first-token in ms), and Throughput (Tokens/sec).
- 4. Energy efficiency and thermal constraints: Calculating TOPS/Watt across battery-powered edge devices.
- 5. Commercial SoC examples: Apple M3/M4 Neural Engine and Qualcomm Snapdragon X Elite NPU.

5.7 Unit V Summary & Quick Revision Sheet

Emerging AI Processor Domain	Key Theoretical & Hardware Takeaways for Exams
Processor Spectrum	CPU (Scalar control), GPU (SIMT parallel FP32), TPU (Systolic array bfloat16), NPU (Low-power INT8 inference), VPU (Vision).
Leading AI Chips	Intel Core Ultra (34 TOPS), Apple M3/M4 (Unified RAM + 38 TOPS ANE), NVIDIA RTX (Ada Tensor Cores), Qualcomm Hexagon (45 TOPS NPU).
Edge vs Cloud AI	Edge: 1-10ms latency, 100% privacy, offline, power-limited. Cloud: Unlimited scale, high bandwidth needed.
Heterogeneous Systems	CPU + GPU + NPU unified die architecture for intelligent workload offloading.
Performance Metrics	FLOPS (Floating point/sec), TOPS (10^12 INT8 ops/sec), TOPS/Watt (Energy efficiency benchmark).