GPU and TPU Architectures in Modern Computing

Understanding specialized processors that power artificial intelligence and high-performance computing systems

CHAPTER 1

The AI Computational Challenge

Core AI Computation

Modern AI relies heavily on matrix operations and tensor processing. The fundamental neural network equation Y = W·X + b must be executed millions or billions of times over massive datasets.

This creates extreme computational demands that traditional processors struggle to meet efficiently.

CPU Architecture: Built for Control

Sequential Design

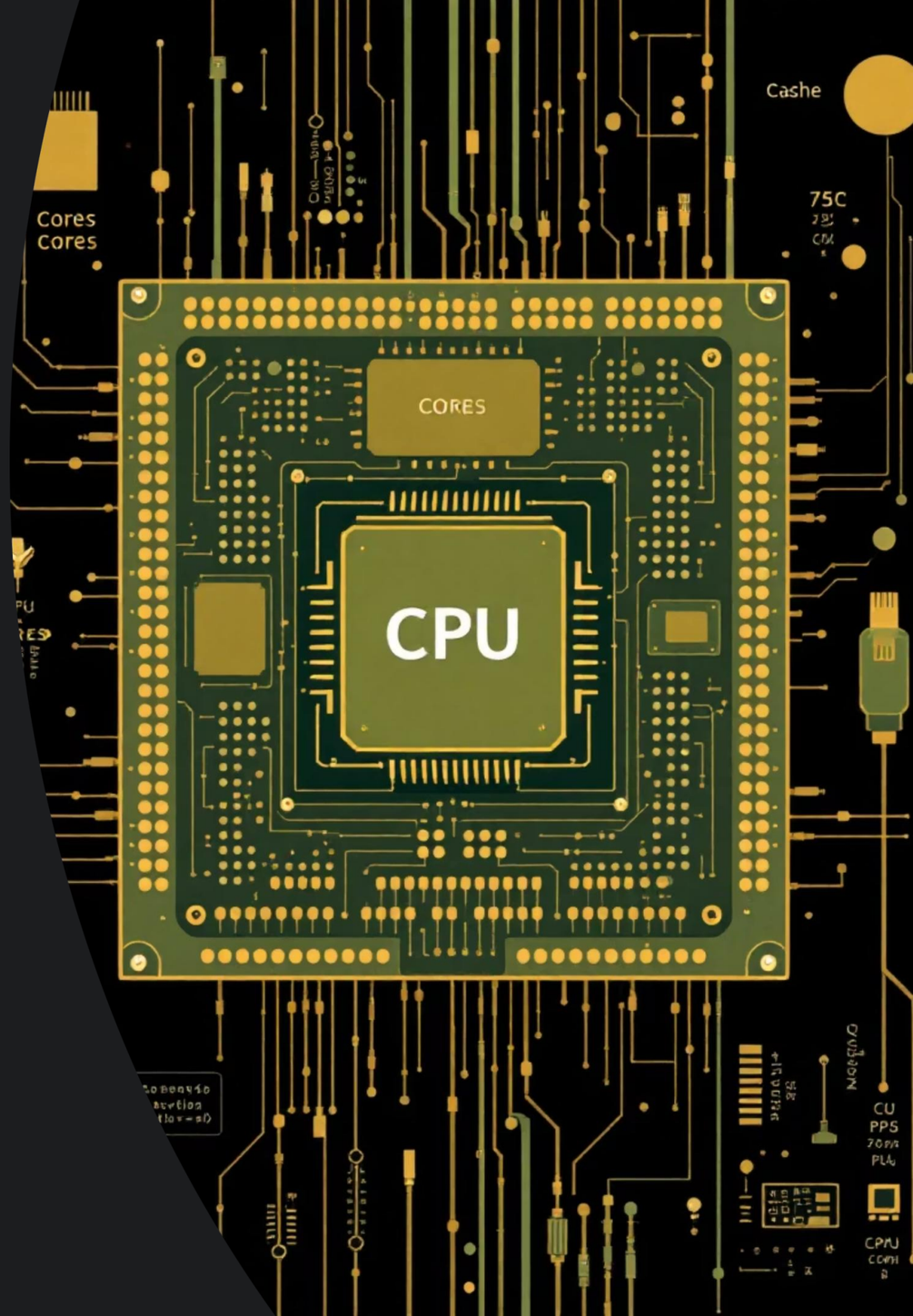
CPUs excel at executing instructions in order with complex control logic

Limited Cores

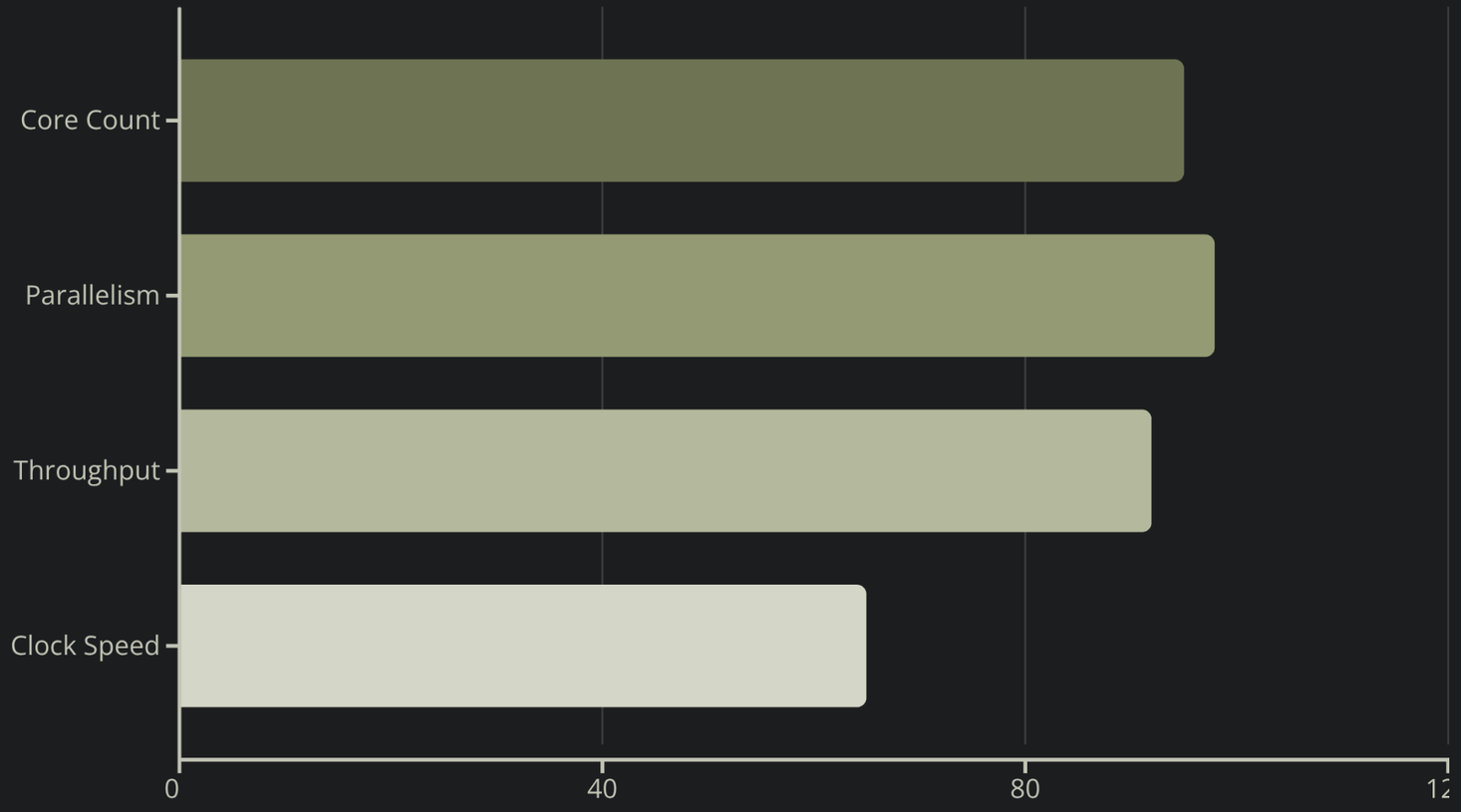
Typically 4-16 cores optimized for low latency operations

AI Limitation

Not suited for massive parallelism required by matrix operations



GPU: Parallel Processing Powerhouse



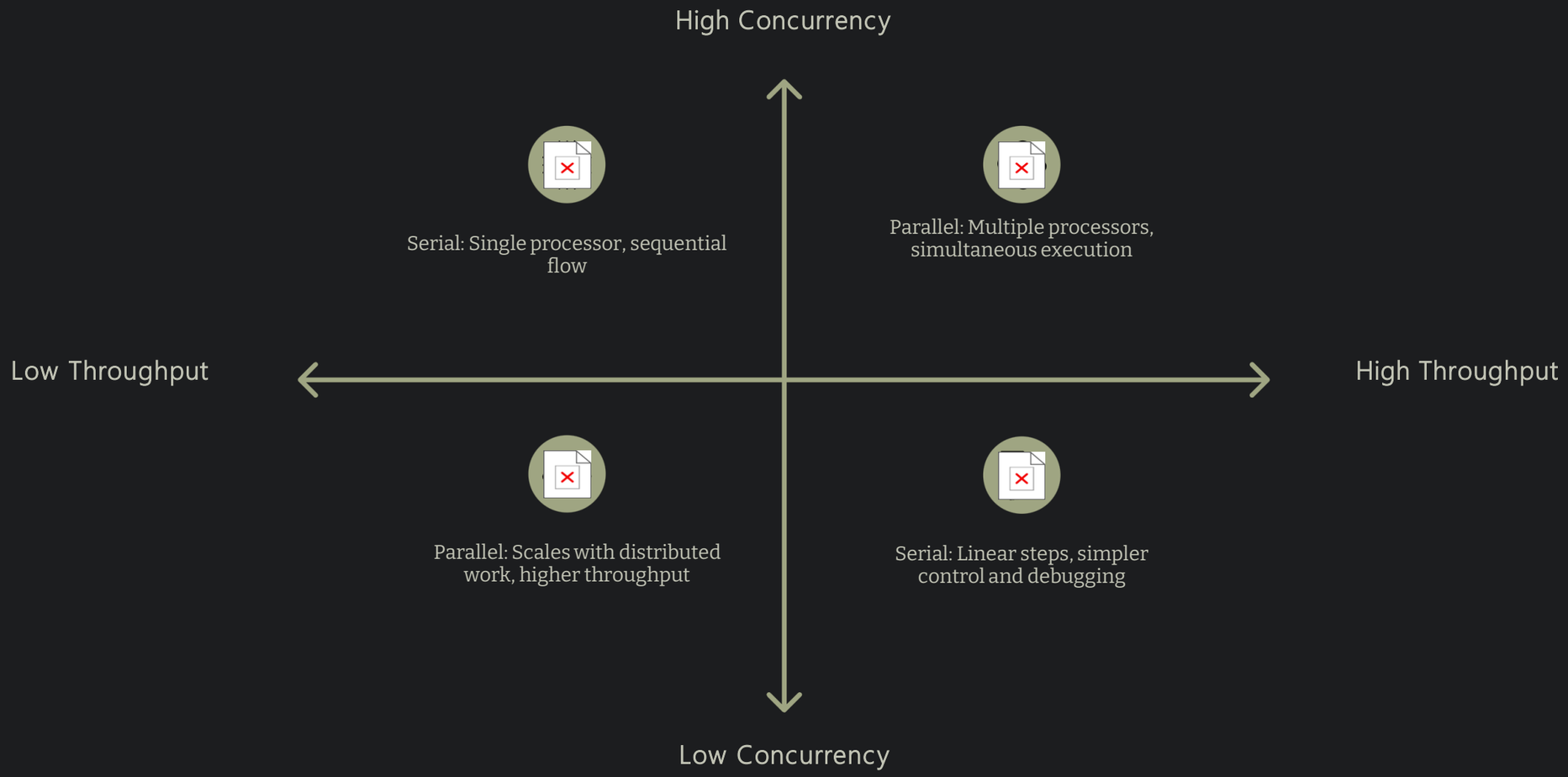
SIMT Architecture

GPUs use Single Instruction, Multiple Threads paradigm with thousands of simple cores executing identical operations simultaneously.

Originally designed for graphics, GPUs evolved into general-purpose parallel processors perfect for AI workloads requiring massive data-parallel computation.

CHAPTER 2

Serial vs. Parallel Computing



Serial computing executes operations sequentially on a single processor, while parallel computing divides work across multiple processors simultaneously. AI's computational intensity makes parallel computing fundamental for modern deep learning systems.

GPU Architecture Components

Streaming Multiprocessors

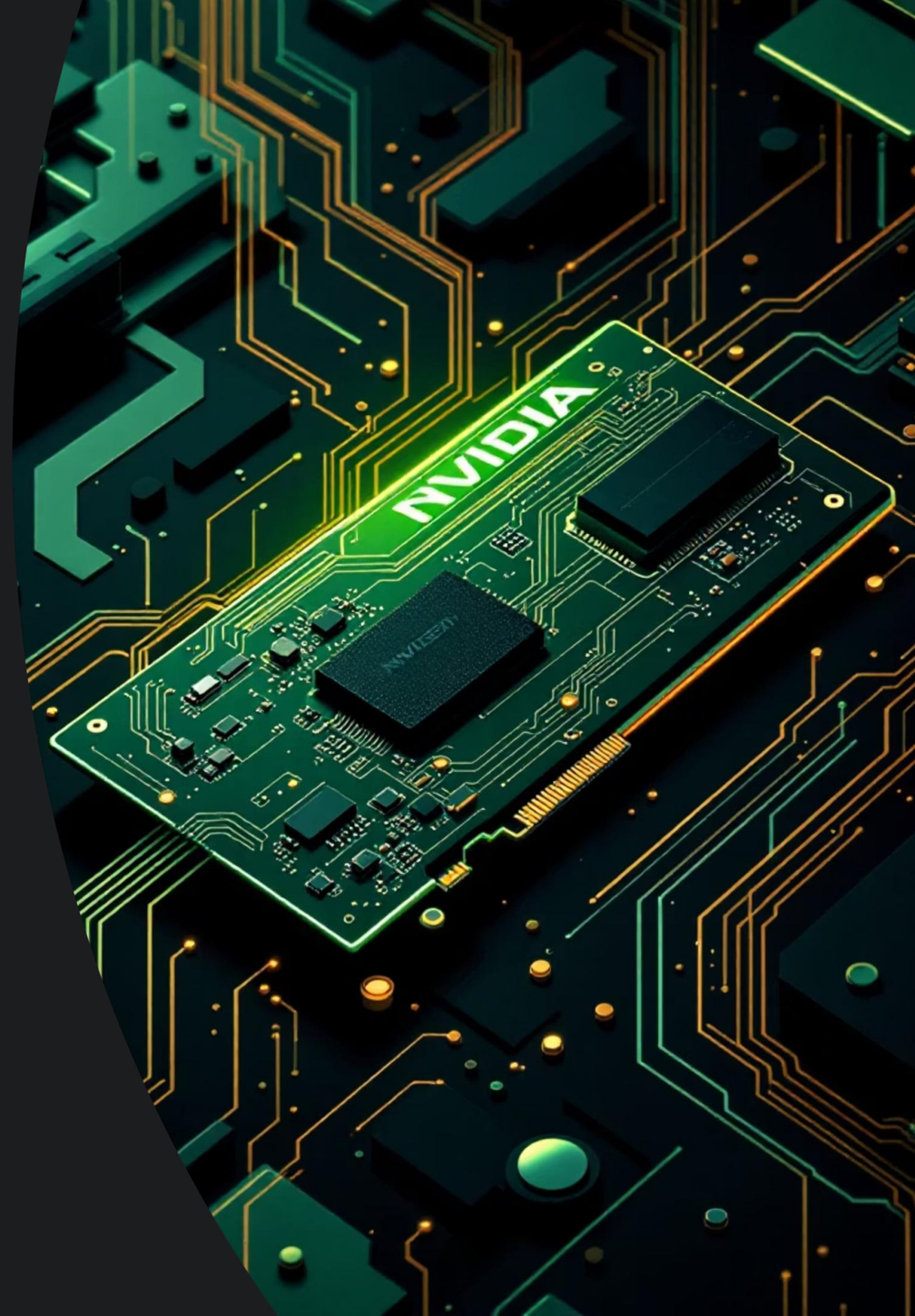
Thousands of simple processing cores organized into SMs for parallel execution

High-Bandwidth Memory

Specialized memory hierarchy from registers to global GDDR/HBM

Programming Models

CUDA, OpenCL, and DirectCompute enable thousands of parallel threads



TPU: AI-Specific Acceleration

Google's ASIC Design

Tensor Processing Units are Application-Specific Integrated Circuits optimized exclusively for machine learning workloads.

Design Philosophy

Optimize hardware exclusively for tensor operations with extremely fast matrix multiplication, low power consumption, and high throughput.

01

Matrix Multiply Unit

128×128 matrix multiplication in single clock cycle

02

High-Bandwidth Memory

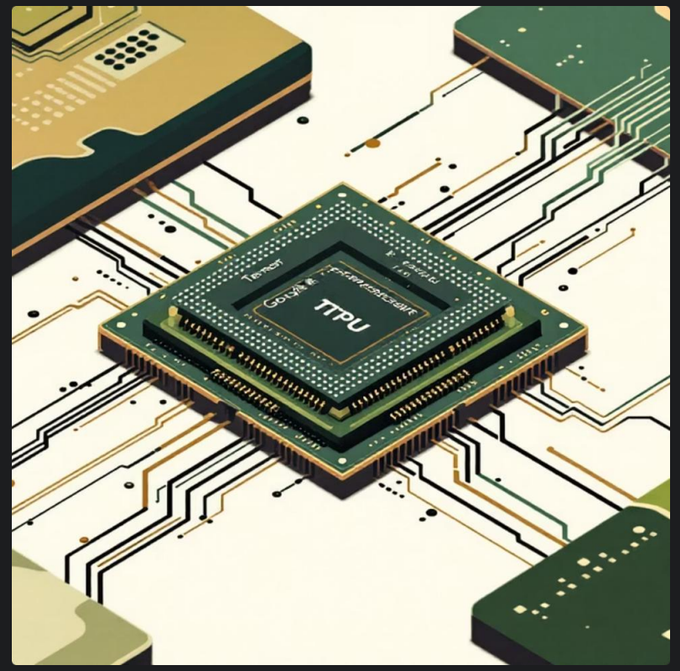
Specialized on-chip memory for tensor operations

03

Optimized Instruction Set

Custom ML-focused commands for maximum efficiency

GPU vs TPU: Performance Comparison



GPU Strengths

General-purpose flexibility, high programmability with CUDA/OpenCL, ideal for research and development

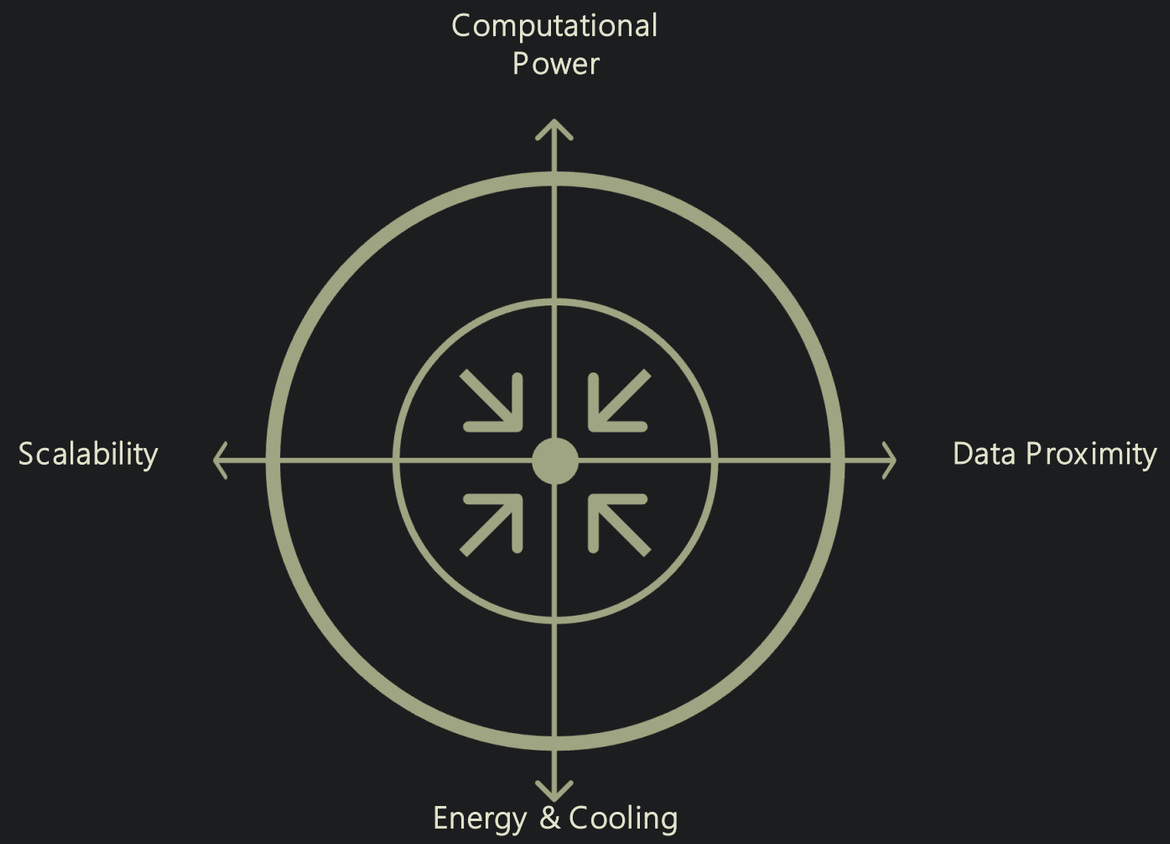
TPU Advantages

AI-specific optimization, very high energy efficiency, best for large-scale training with TensorFlow APIs

Both architectures address the critical bottleneck: memory bandwidth. High-speed memory access often matters more than raw computation speed in AI workloads.

CHAPTER 3

AI Data Centers: Strategic Infrastructure



Why Data Centers Matter

AI is compute-intensive, requiring billions of parameters and trillions of FLOPs. Data centers enable specialized hardware to operate in highly parallel, synchronized environments at scales impossible with desktop systems.

Moving data costs more than computing it—data centers provide low latency and high bandwidth for local processing.

The Stargate Project: $500B AI Infrastructure

$500B

Total Investment

Led by OpenAI, SoftBank, Oracle, and MGX

10

Gigawatts Capacity

Target AI data center capacity by 2029

Global facilities planned across Texas, UAE, South Korea, and multiple U.S. locations. Strategic partners include NVIDIA, Samsung, SK, Cisco, Arm, and Microsoft—aiming to establish AI infrastructure leadership and strengthen data sovereignty.