

Computer Architecture Courses

- Multicore Computer Architecture – John Jose
- Advanced Computer Architecture – Smruti Sarangi
- High Performance Computing – Ajit Pal
- Parallel Computer Architecture – Hemangee Kapoor
- GPU Architectures and Programming –
- Digital Design and Computer Architecture – Onur Mutlu
- Computer Architecture – Onur Mutlu

Motivation

Many Difficult Problems

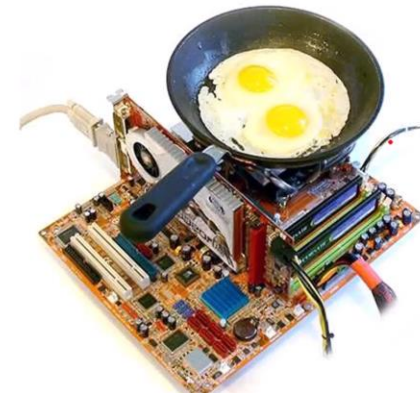
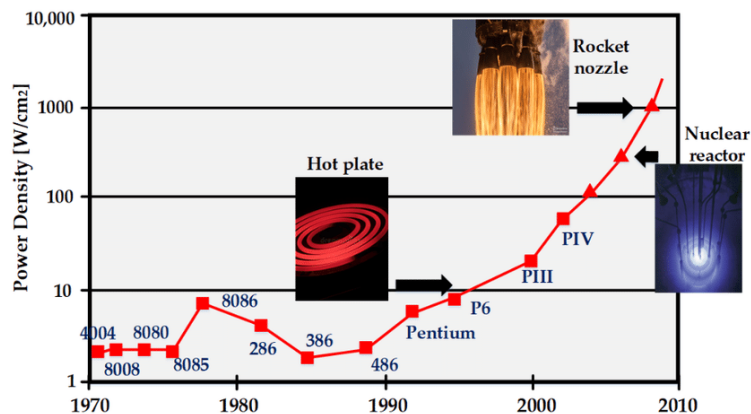
- Power/Energy/Thermal constraints
 - Complexity of Design
 - Difficulties in **Technology scaling**
 - Memory bottleneck
 - **Reliability** problems
 - Programmability problems
 - **Security and Privacy** issues
 - Huge hunger for data and new data-intensive Applications (AI, ML, Genomics, ...)
- 3D Stacked Chips*
- DRAM RowHammer Meltdown and spectre*
- Non-Volatile Main Memory*
- Bit flips and Silent data corruption*
- Main memory access consumes ~100-1000X the energy of complex addition*
- Specialized Accelerators (GPU, TPU, NPU, ML, Video codec, ...)*
- 90% of total energy spent on memory*
- 62.7% of the total system energy is spent on data movement*

Four Key Current Directions

- Fundamentally **Secure/Reliable/Safe** Architectures
- Fundamentally **Energy-Efficient** Architectures
- Fundamentally **Low Latency and Predictable** Architectures
- Architectures for **AI/ML, Genomics, Medicine, Health care, ...**

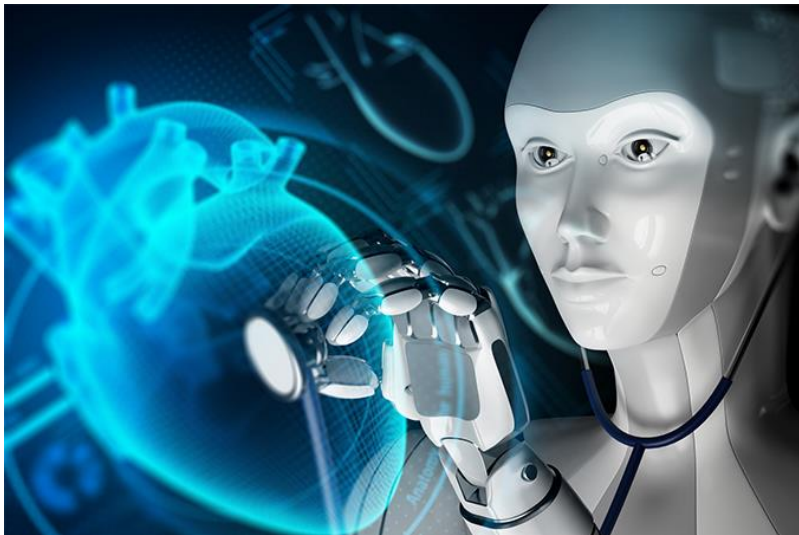
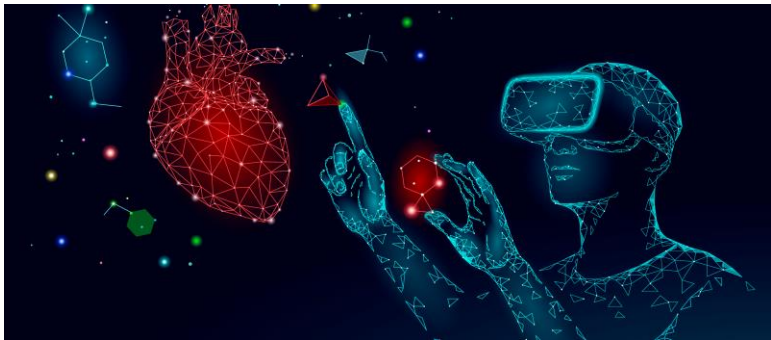
Research Areas

- Computer architecture
- Interconnects, Memory and Storage systems
- Hardware security, safety and predictability
- Fault tolerance, robust systems
- Hardware and software cooperation
- Architectures for bioinformatics, health, medicine and intelligent decision making



A request made through ChatGPT, an AI-based virtual assistant, consumes 10 times the electricity of a Google Search

Motivation





89%
07:30

10-Day Battery Life Supports Magnetic Charging

With a built-in 345mAh battery and a low-power chipset, the realme Watch 3 Pro can achieve 10 days of long battery life with standard daily use of GPS and Bluetooth calling.*

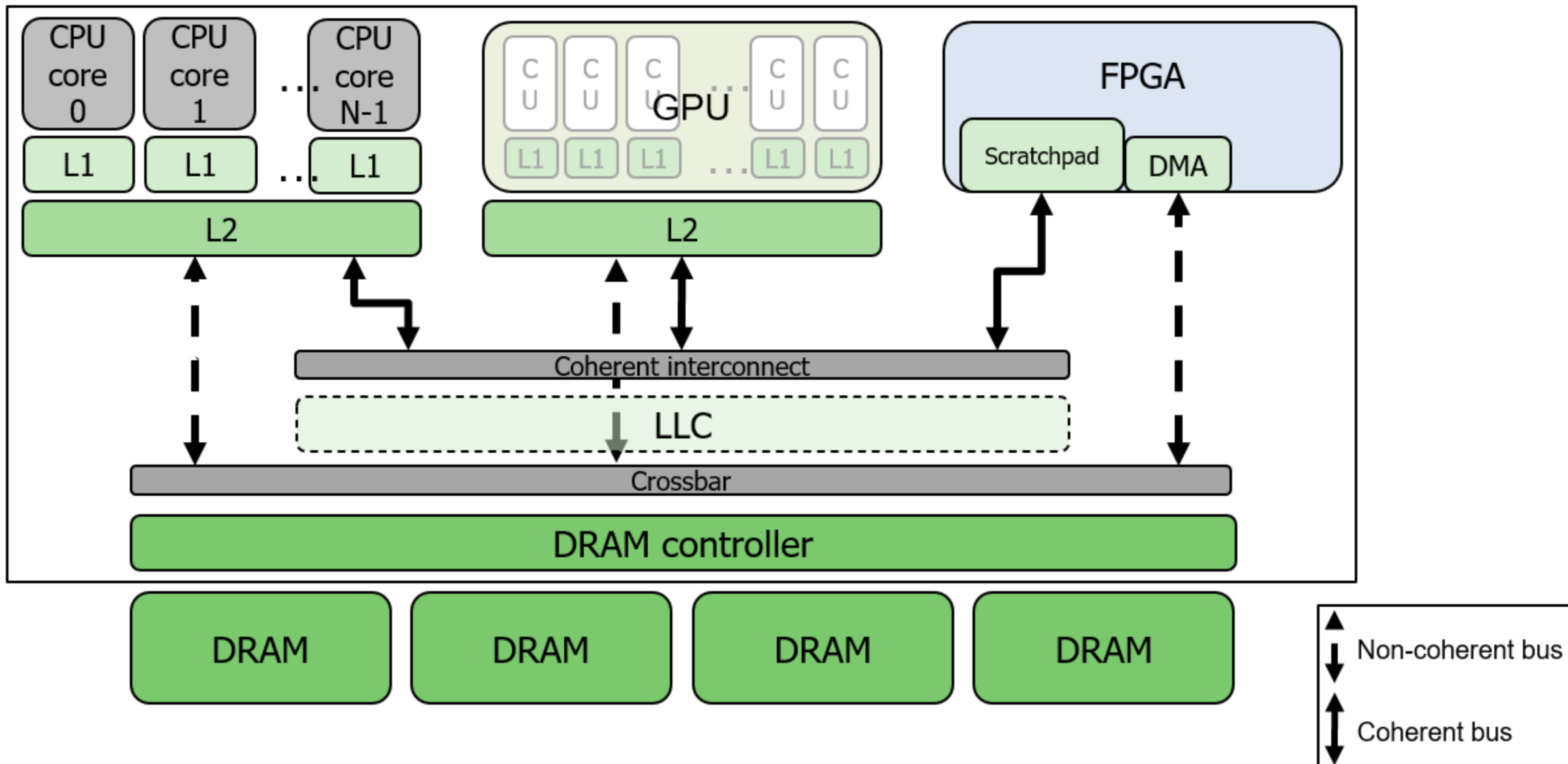


Motivation

Problem
Algorithms
Program/Language
System Software
HW/SW Interface
Micro-architecture
Logic
Devices
Electrons

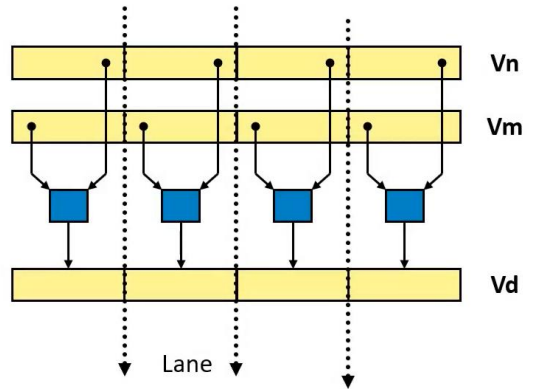
We need more denser memory so we will get bigger memory in size with smaller area. But this leads to rowhammer problem.

Heterogeneous Computing Systems



Compute

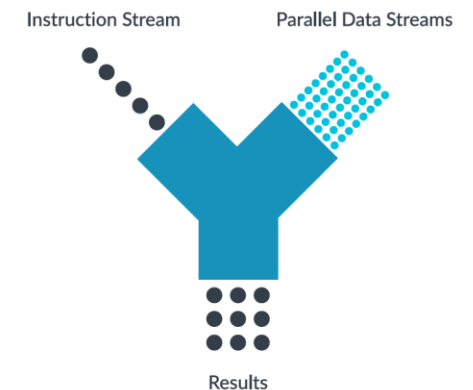
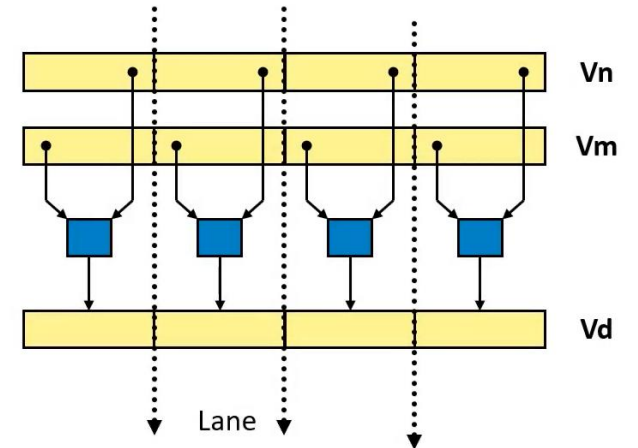
Flynn's Taxonomy

SISD – Single Instruction Single Data	SIMD – Single Instruction Multiple Data
<ul style="list-style-type: none">Simple Processor	<ul style="list-style-type: none">Array ProcessorVector ProcessorGPU  <p>The diagram illustrates the SIMD (Single Instruction Multiple Data) architecture. It shows three horizontal yellow bars representing data vectors, labeled V_n, V_m, and V_d from top to bottom. Four vertical dashed lines represent processing lanes. In each lane, a blue square represents a processor. Arrows indicate that a single instruction is applied to multiple data elements across the lanes. The label 'Lane' is placed below the first dashed line.</p>
MISD – Multiple Instruction Single Data	MIMD – Multiple Instruction Multiple Data
<ul style="list-style-type: none">Systolic Array ProcessorsStreaming Processors	<ul style="list-style-type: none">Multi-ProcessorsMulti-Threaded Processors

SIMD

Why SIMD?

- Some modern software particularly media codec and graphics accelerators, operates on large amount of data that is less than word sized.
 - 16 – bit data is common in Audio applications
 - 8 – bit data is common in Graphics and Video applications
- **When performing this operations on 64-bit Microprocessor, parts of the computation units are unused. But continue to consume same power**
- SIMD technology uses single instruction to perform same operation in parallel
 - **Single addition latency = Multiple parallel addition latency – Improves Performance**
- SIMD was first introduced in ARMv6 Architecture
- In ARMv7 ARM introduced Advanced SIMD (Neon) as an optional extension
 - SIMD was 32-bit wide Neon extended it to 64-bit and 128-bit wide
- Coprocessor 10 and 11 used for Neon and Vector Floating Point (VFP) extension



Advanced SIMD - Neon

- Neon unit uses 128bit registers for SIMD processing
- Neon register file supports 8-bit, 16-bit, 32-bit, 64-bit or 128-bit wide data types
- Usually each Neon instruction results in N instruction executing in parallel, where N is the number of **lanes**
- There cannot be an carry or overflow from one lane to another lane

DRAM Timings

[illegible]

Case Study - Nvidia Ampere Architecture

Why Low Power?

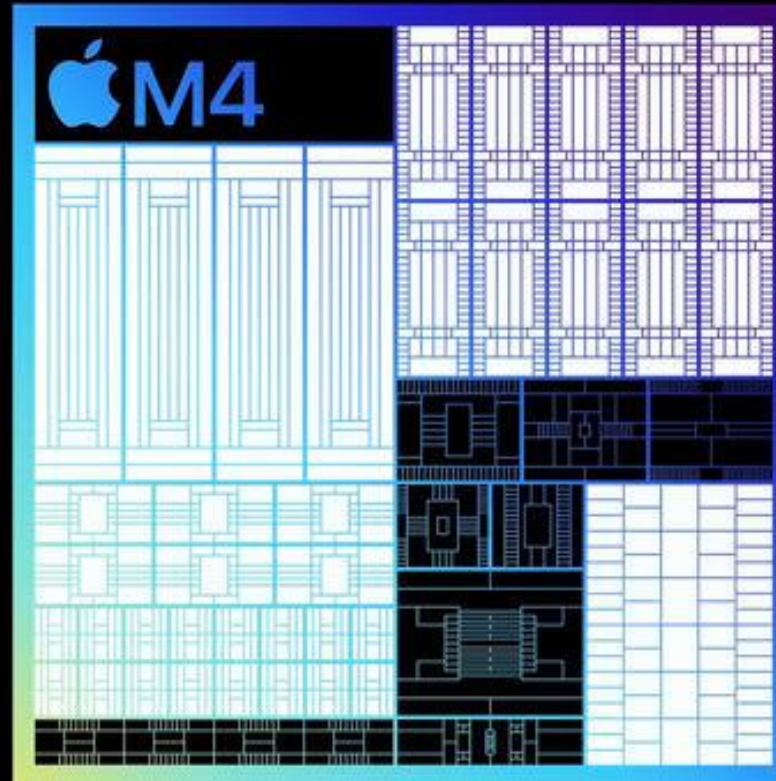
4 performance cores

Improved branch prediction
Wider decode and execution engines
Next-generation ML accelerators

6 efficiency cores

Neural Engine

16-core design
Faster and more efficient



10-core GPU

Next-generation architecture
Dynamic Caching
Mesh shading
Ray tracing

Display engine

Tandem OLED support
Brightness and color compensation
10Hz-120Hz ProMotion support

