Learning aims

- Describe von Neumann, Harvard and contemporary processor architecture

Data and instructions
Memory
CPU architectures

Von Neumann
Stored program concept
Serially
bit pattern
Von Neumann bottleneck

Harvard
Special-purpose
Real-time

Cache - a small, fast type of memory in the CPU that stores frequently used data to speed up processing.

General Terms

Key term	Definition
Data and instructions	Data = the values processed (numbers, text, etc.). Instructions = the commands telling the CPU what to do. Both must be stored in memory.
Memory	Stores data and instructions. Can be RAM (temporary) or other storage.
CPU architecture	The design of how the CPU, memory, and buses are organised and interact.

Von Neumann Architecture

Key term	Definition
Von Neumann	A CPU design where data and instructions are stored in the same memory and travel along the same buses.
Stored program concept	Instructions for a program are stored in memory alongside the data, so the CPU can fetch them as needed.
Serially/sequentially	Instructions are fetched and executed one after another in order.
Bit pattern	Binary digits (0s and 1s) used to represent instructions or data.
Von Neumann bottleneck	The limitation caused because data and instructions share the same bus, meaning only one can be transferred at a time, slowing performance.

Harvard Architecture

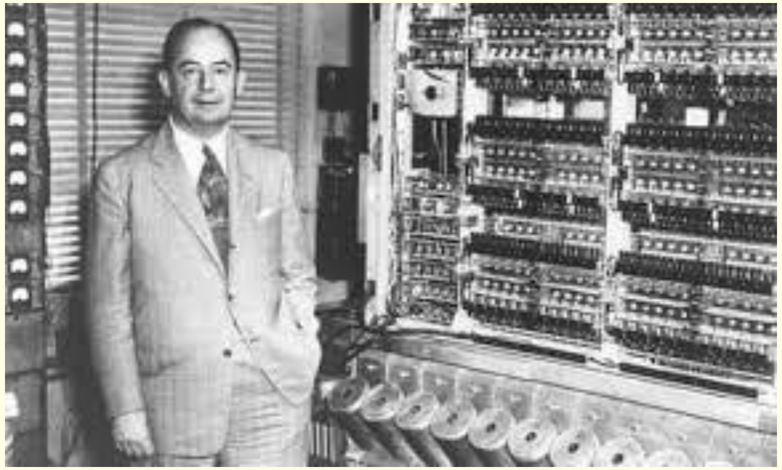
Key term	Definition
Harvard	A CPU design where data and instructions are stored in separate memory and use separate buses. This allows simultaneous access and can be faster.

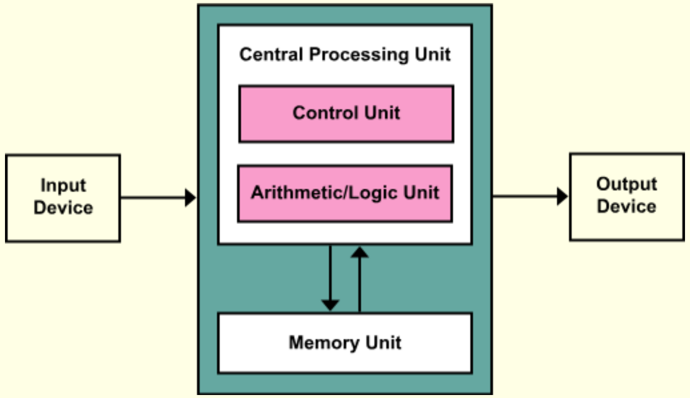
Special-purpose Systems

Key term	Definition
Special-purpose	A computer designed for a specific task (e.g., a washing machine controller, sat-nav).
Real-time	A system that responds immediately to inputs and events, often used in safety-critical systems (e.g., car airbags, medical equipment).

CPU Architectures - Memory

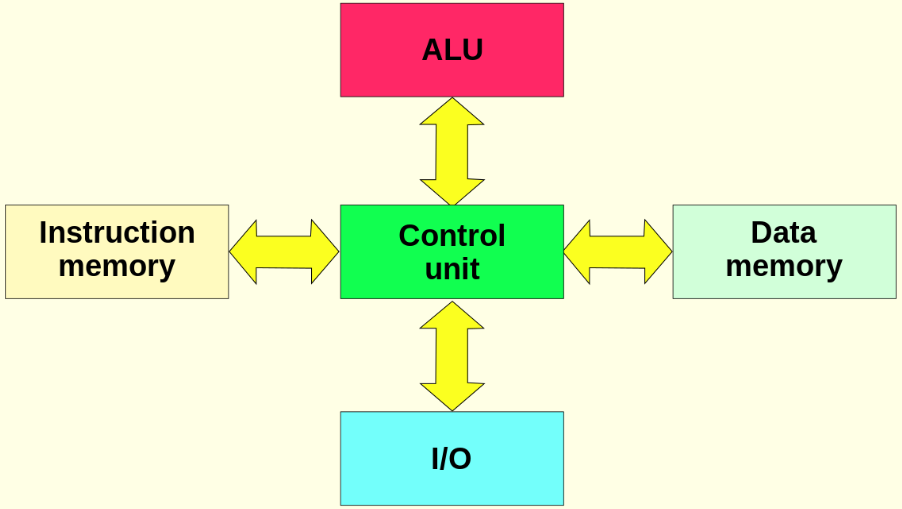
Von Neumann





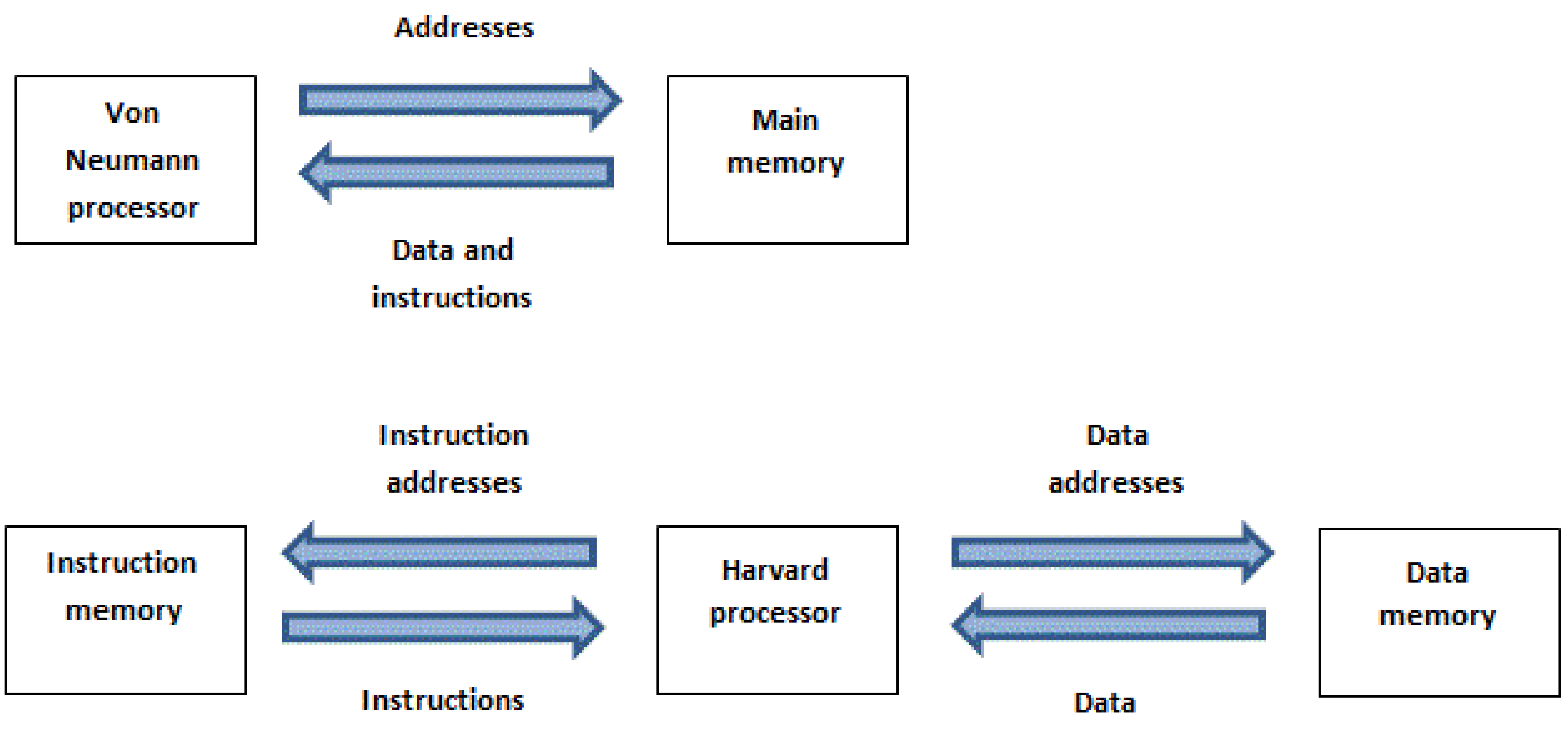
Stored program concept
Instructions and data are stored in the same memory.

Harvard architectures



Physically **separate** memories for instructions and data

Comparison


Addresses
Von Neumann processor
Main memory
Data and instructions
Instruction addresses
Data addresses
Instruction memory
Harvard processor
Data memory
Instructions
Data

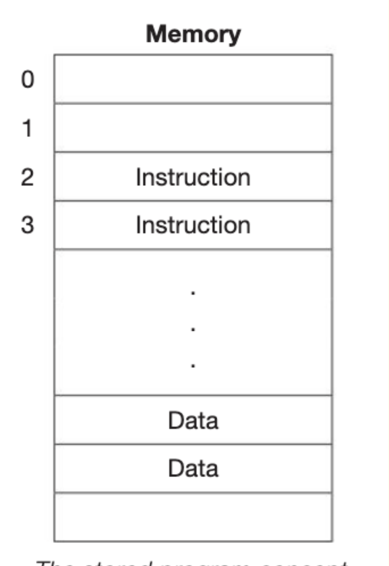
The Von Neumann machine

von Neumann

Same data bus is used to transfer both data and instructions.

Same word length is used for all memory, whether it holds data or instructions.

Can only fetch either data or instructions at one time/follows FDE



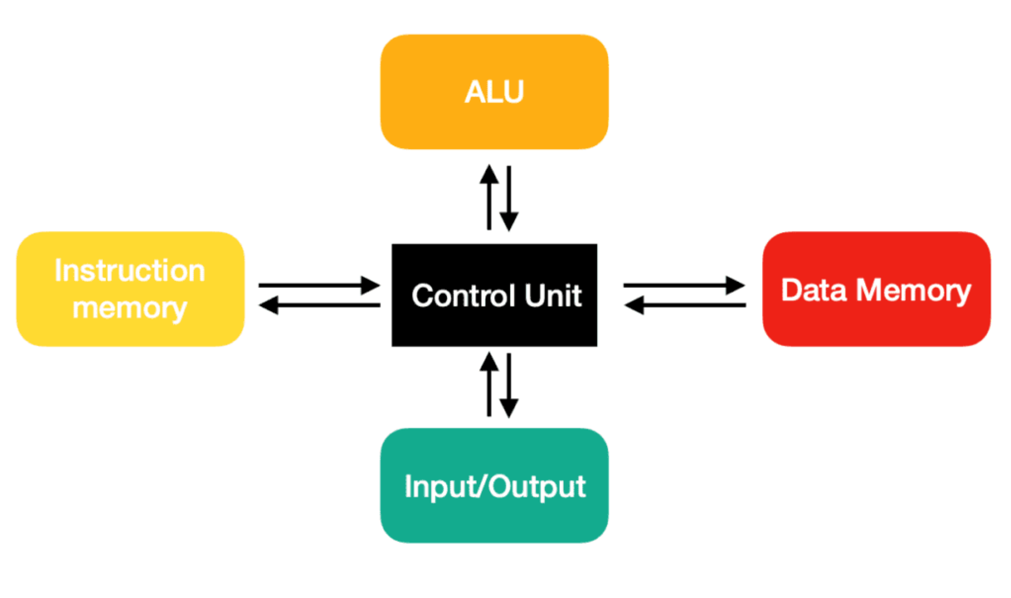
The stored program concept

Harvard

Different (sets of) buses. one for instructions & one for data

Can use **different word lengths** for data and instructions, (optimising memory use)

Instructions and data can be accessed concurrently.



Von Neumann bottleneck

- Whatever you do to improve performance, you cannot get away from the fact that instructions can only be done one at a time and can only be carried out sequentially.

- Both of these factors hold back the efficiency of the CPU.

- This is commonly referred to as the **'Von Neumann bottleneck’** .

- The von Neumann bottleneck is a computing limitation that occurs when data transfer between the central processing unit (CPU) and memory is slow.

- This causes the CPU to wait for data to be accessed.

Used for …

Architecture	**Typical Uses**	**Examples**
Von Neumann Architecture	- Used in **general-purpose computers** where tasks vary (e.g. PCs, laptops, tablets). - Suitable for systems that **run many types of programs** stored in the same memory (instructions + data).	Desktop computers, laptops, smartphones, servers.
Harvard Architecture	- Used in **embedded systems** and **microcontrollers**, where the program and data are fixed or predictable. - Ideal for **digital signal processing (DSP)** or **real-time systems** needing fast, simultaneous access to data and instructions.	Washing machines, robots, digital cameras, automotive control units, DSP processors.

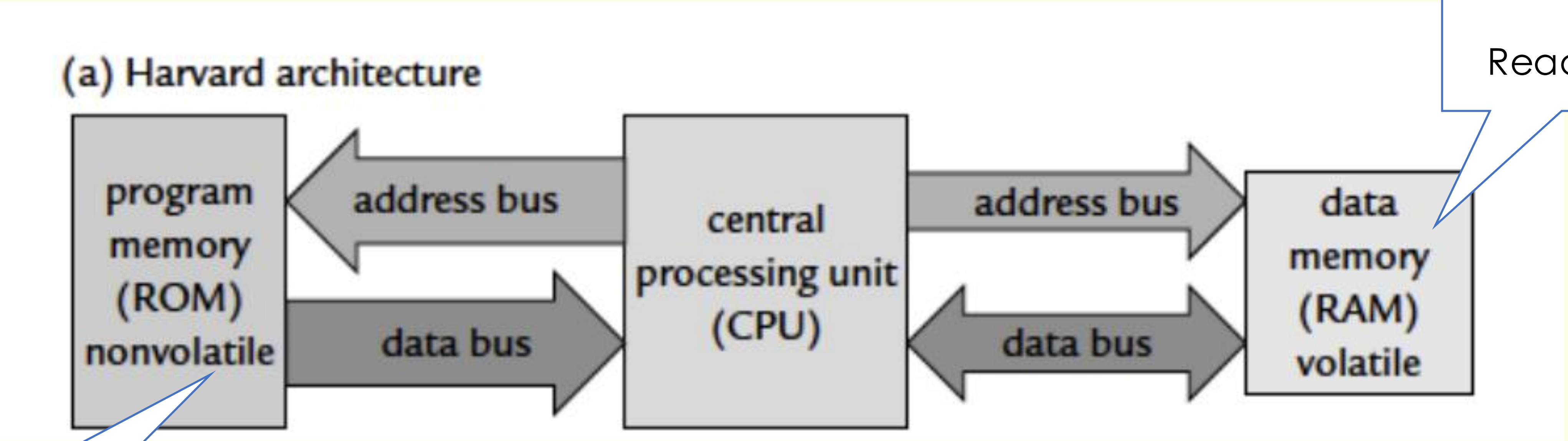
Harvard and Real-time Embedded systems

Embedded systems include **special-purpose computers** built into devices often operating in **real time** such as those used in:

- navigation systems
- traffic lights
- aircraft flight control systems and simulators

Harvard architecture can be **faster** than von Neumann architecture because **data and instructions** can be fetched in **parallel** instead of competing for the same bus.

Embedded Systems


(a) Harvard architecture
program memory (ROM) nonvolatile
address bus
data bus
central processing unit (CPU)
address bus
data bus
data memory (RAM) volatile

RAM is used for data (Flash storage)

Read/write

ROM is used for instructions/ firmware

Read-only

Fill-in-the-Gaps Activity:

1. In the Von Neumann architecture, both data and instructions are stored in the same __________.

2. Harvard architecture uses separate __________ for data and instructions, allowing parallel access.

3. The shared bus in Von Neumann architecture can lead to a performance limitation called the Von Neumann __________.

4. Harvard architecture is commonly used in embedded systems, where speed is a critical factor.

Summary

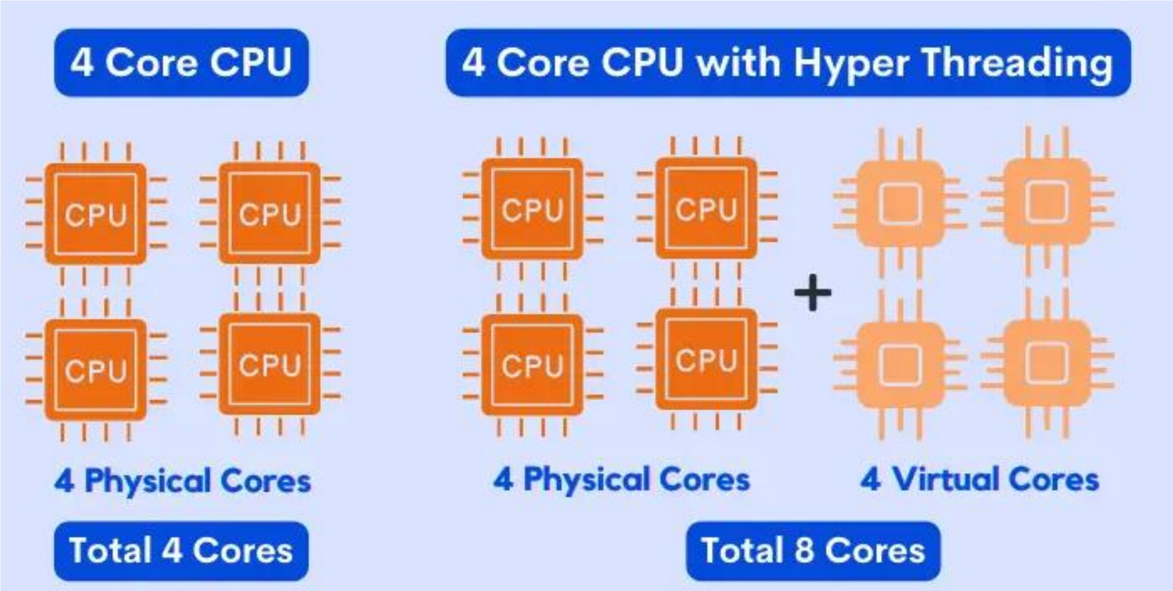
Von Neumann Architecture	Harvard architecture
• Instructions and Data stored in same area of memory • Von Neumann fetches data and instructions sequentially/follows FDE cycle • Von Neumann uses a single bus for both data and instructions • Single control bus • Slower due to the Von Neumann bottleneck (using the same bus) • Used in general-purpose computers (e.g. PCs, laptops). • Can be optimised by using pipelining • Same word length for both data and instructions	• Harvard stores data and instructions in separate memory units • Harvard can fetch data and instructions at the same time (fetching the next instruction while reading/writing data.) • Separate control buses • Harvard uses different buses for data and instructions • Faster because data and instructions are accessed in parallel. • Common in embedded systems and microcontrollers. Digital Signal Processing (DSP) systems that require fast access to data and instructions • Can use different word lengths for data and instructions, (optimising memory use)

Features of contemporary processors

- **Two separate areas of memory**…one for instructions & one for data./instructions and data can be accessed concurrently.
- **Different (sets of) buses**…one for instructions & one for data./instructions and data can be accessed concurrently.
- **Pipelining**…whilst an instruction is being executed the next can be decoded and the subsequent one fetched.
- **Use of cache**…A small amount of high performance memory is (next to the CPU) / which stores frequently used data/instructions
- **Virtual cores/Hyper-threading** …Treating a physical core as two virtual cores.
- **Multiple Cores**…Each core acts as a separate processing unit.
- **Onboard Graphics**…Built in circuitry for graphics processing.

Virtual cores/Hyper-threading

- Hyperthreading refers to the technology invented by Intel, with which a physical microprocessor behaves like two logical, virtual cores.
- **Virtual cores** are created by hyper-threading.
- **Hyper-threading** allows each physical core to handle two tasks (threads) at once, creating virtual cores.
- This helps the CPU handle more tasks at the same time, improving performance, especially in multi-threaded applications.
- For example, a CPU with 4 physical cores and hyper-threading can appear to the operating system as 8 virtual (or logical) cores.

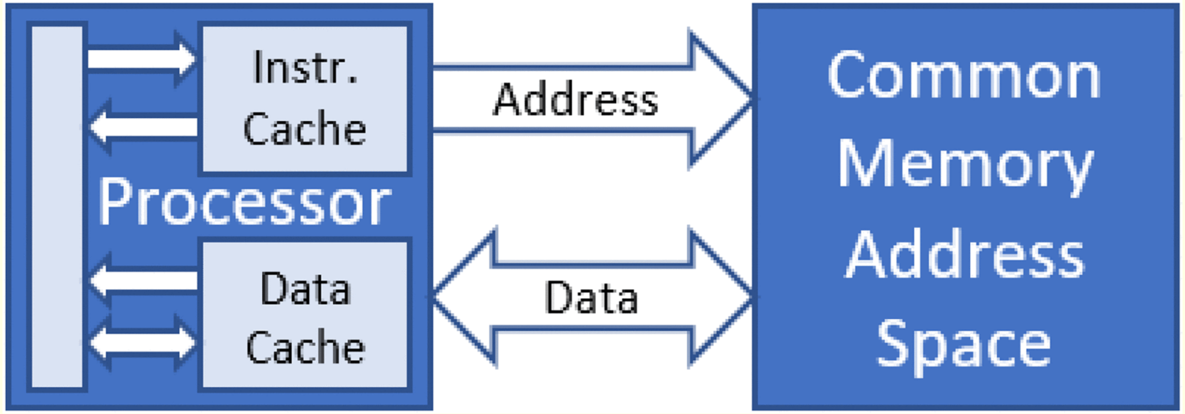


Contemporary processor architectures

- Modern high-performance CPU chips incorporate aspects of both von Neumann and Harvard architecture.

- In one design, there is one main memory for holding both data and instructions, but CPU cache memory is divided into an instruction cache and a data cache.

- Harvard architecture is used as the CPU accesses the **cache**.



- Some digital signal processors such as Texas Instruments TMS320 C55x have multiple parallel data buses (two write, three read) and one instruction bus.

Did you know …?

The Intel Core i9-14900KS is the world's fastest desktop processor, with a maximum turbo frequency of 6.2 gigahertz (GHz)

From mid-2023 onwards, we started to hear increasing reports from game devs and gamers that high-end Intel 13th and 14th Gen CPUs (primarily the Core i9 13900K and Core i9 14900K) were crashing in Unreal Engine games.

Intel has recently announced that it has found the cause of the instability issues affecting its 13th and 14th generation processors:

Cause
The issue is caused by "elevated operating voltage". This is due to a microcode algorithm that sends incorrect voltage requests to the processor.

£666.99

Did you know …?

- The fastest supercomputer in the world is a machine known as **Frontier**

- Located at Oak Ridge National Laboratory in Tennessee

Nick McGinn for *Nature*

Speed	1.102 exaFLOPS (1.102 quintillion floating-point operations per second)
Features	Nearly 50,000 processors
Uses	Scientific discovery, including simulating proteins, improving airplane engine design, and creating large language models