KASETSART UNIVERSITY

01418233 Computer Architecture

Introduction to Memory Hierarchy

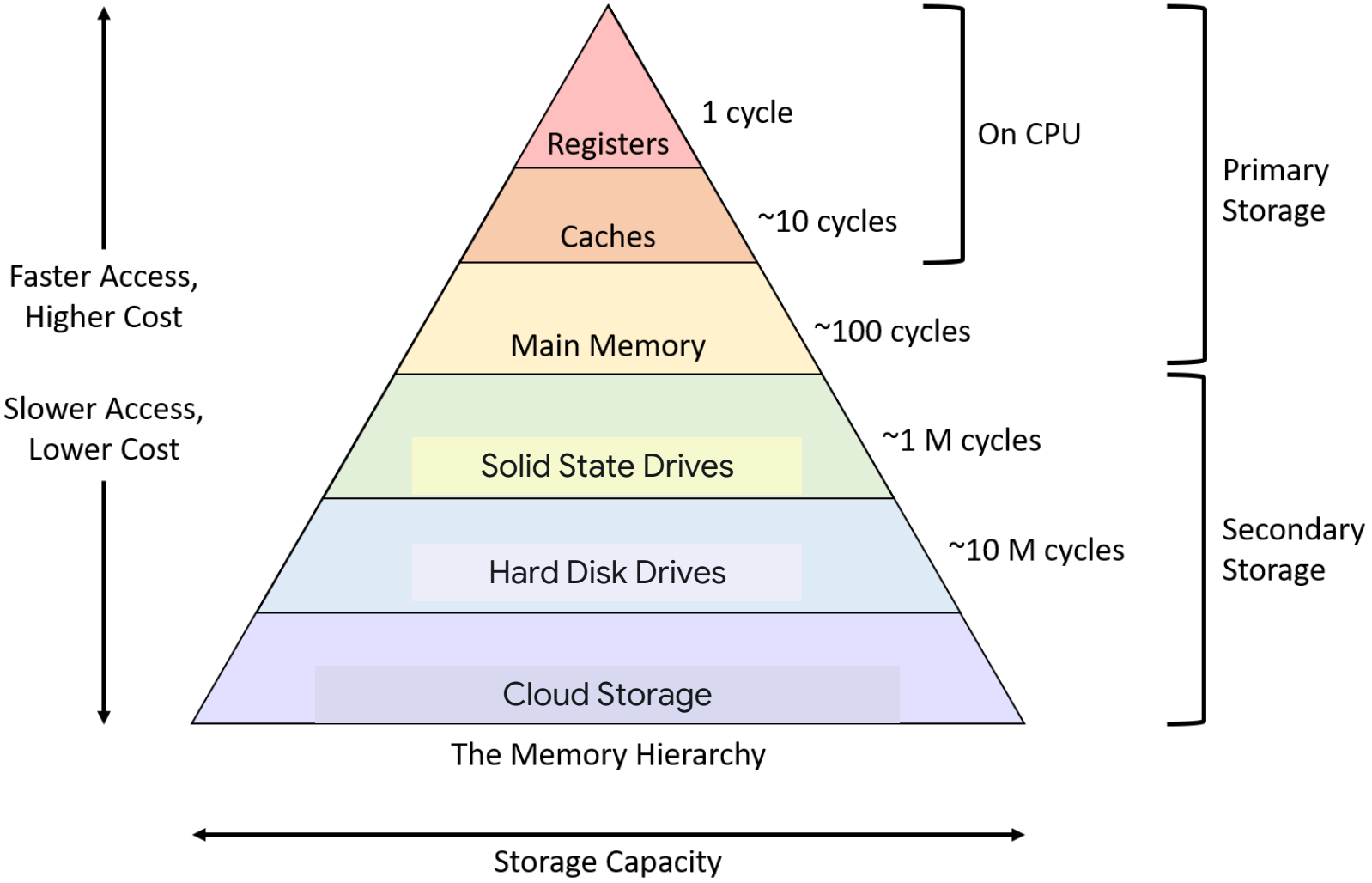
Sarach Tuomchomtam, Ph.D.

sarach.t@ku.th

Memory Hierarchy

The organization and arrangement of different types of memory in a computer system.

Designed to optimize the performance of the system by providing varying levels of storage capacity, access speed, and cost.

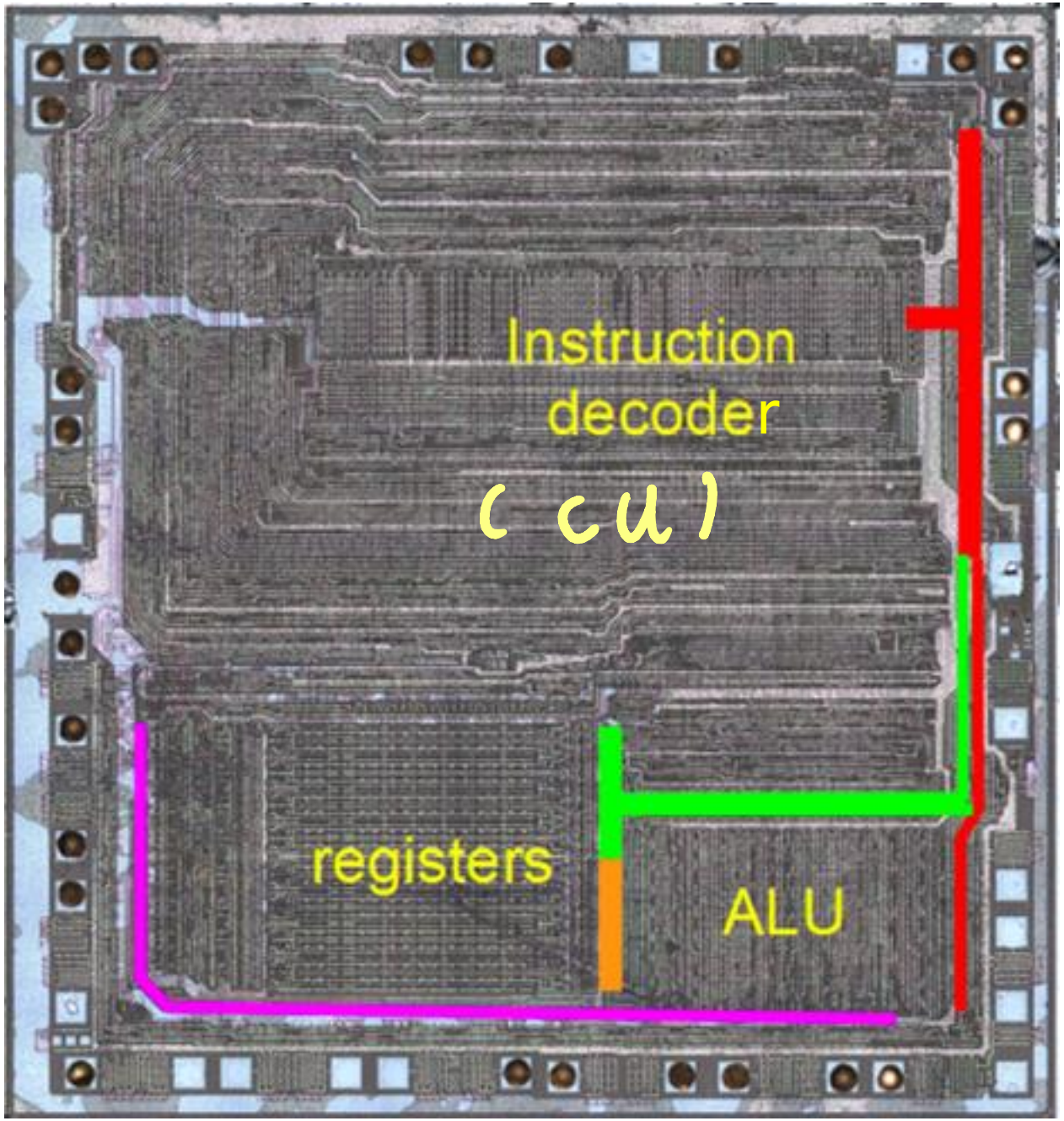


The Memory Hierarchy

Types

Registers

Small, high-speed storage units located directly in the CPU.



Types

Caches

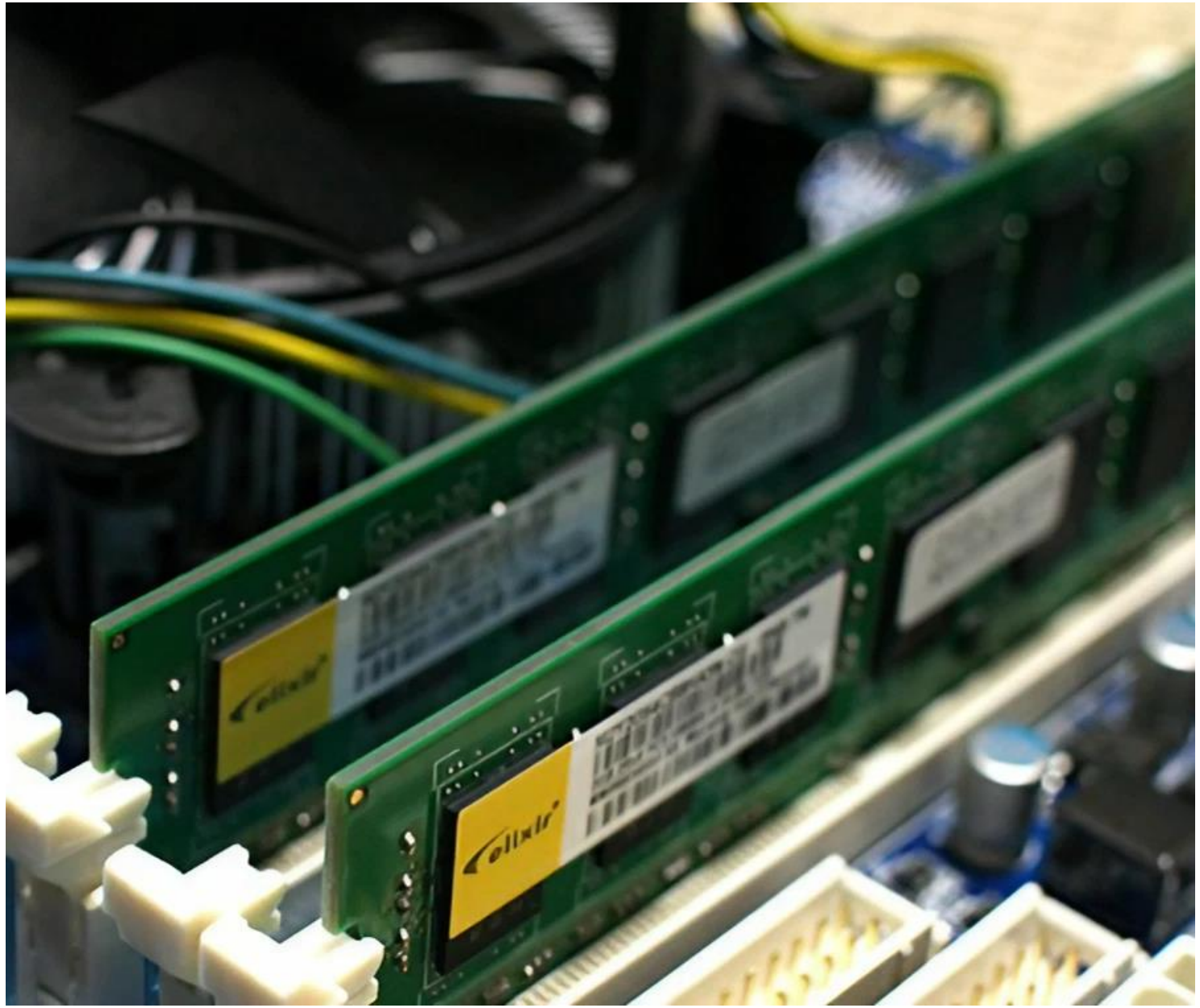
Small but faster storage units that sit between the CPU and main memory.

Types

Main Memory

(RAM)

Often referred to as Random Access Memory (RAM), is the primary memory of a computer.

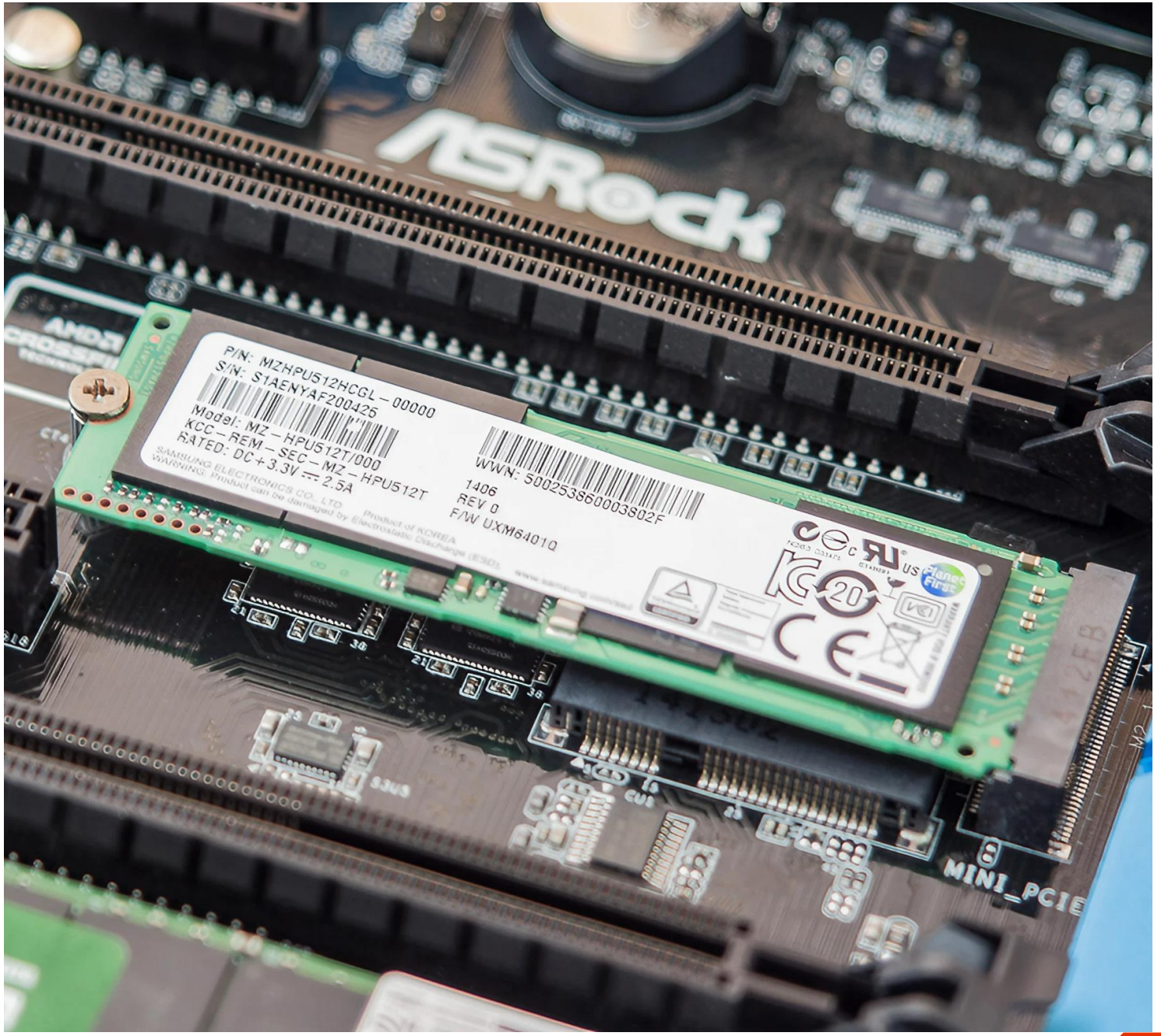


Types

Solid State Drives (SSDs)

Storage devices that use flash memory to store data.

ใช้ในอุปกรณ์ elec ฯ รุ่นใหม่



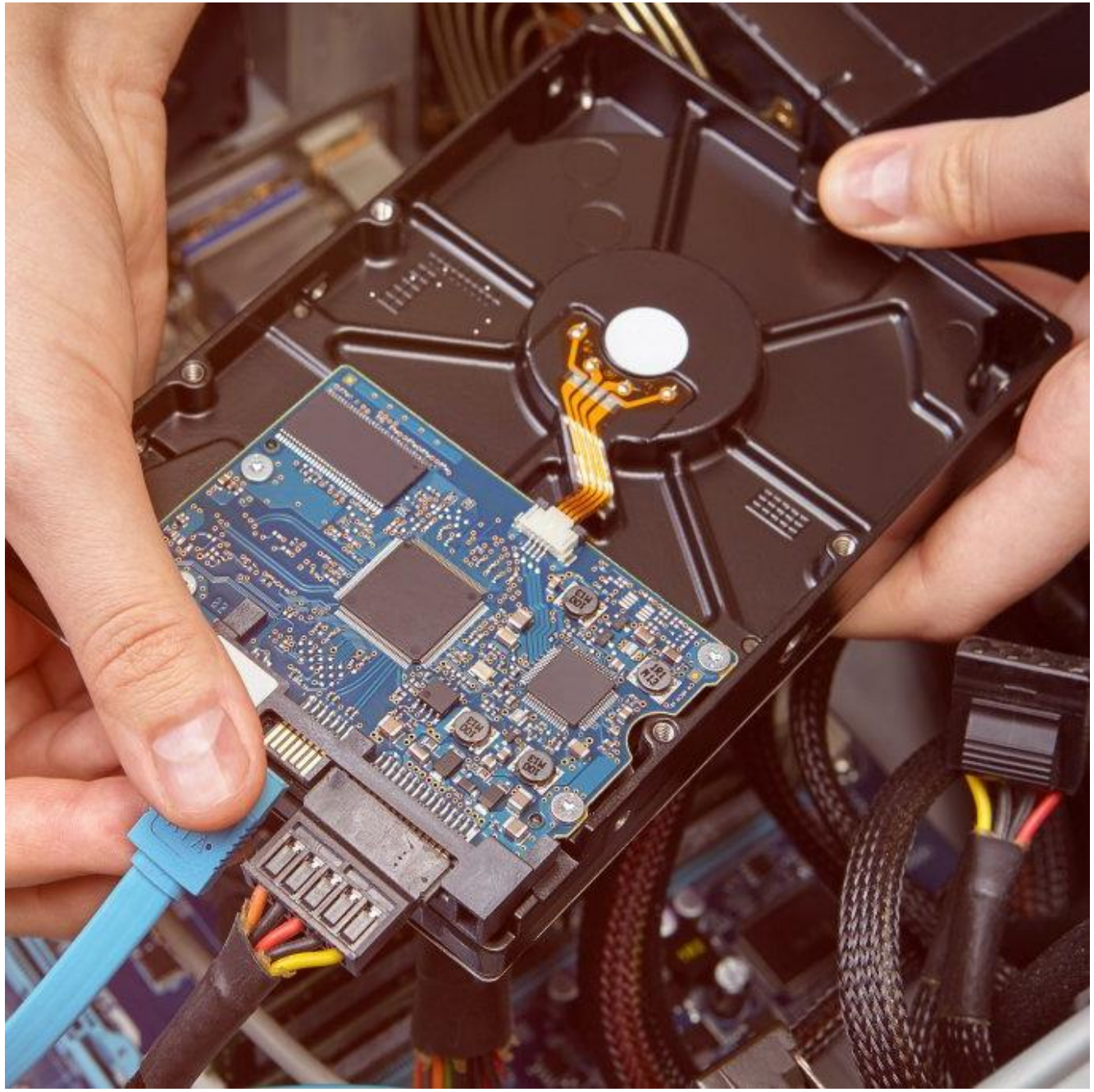
Types

Hard Disk Drives (HDDs)

Traditional mechanical storage devices that use spinning disks and read/write heads.

ความจุ + ราคาถูก

เก็บข้อมูลที่ไม่ได้ใช้บ่อย



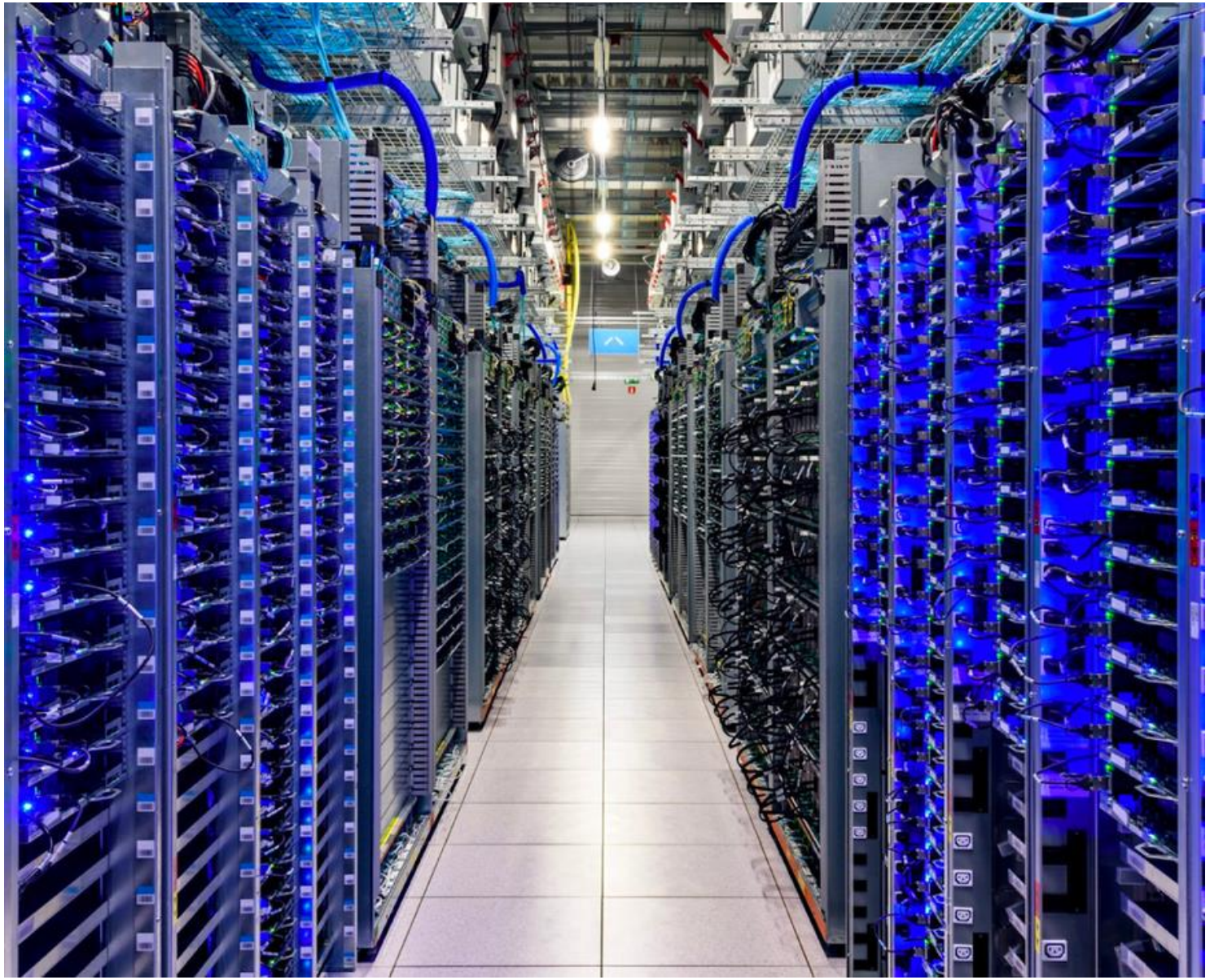
Types

Cloud Storage

Data storage services provided by remote servers over the internet.

เก็บได้ไม่จำกัด

ค.เข้าถึง (ต้องผ่าน Internet)



Outline

Characteristics

Speed

Capacity

Cost

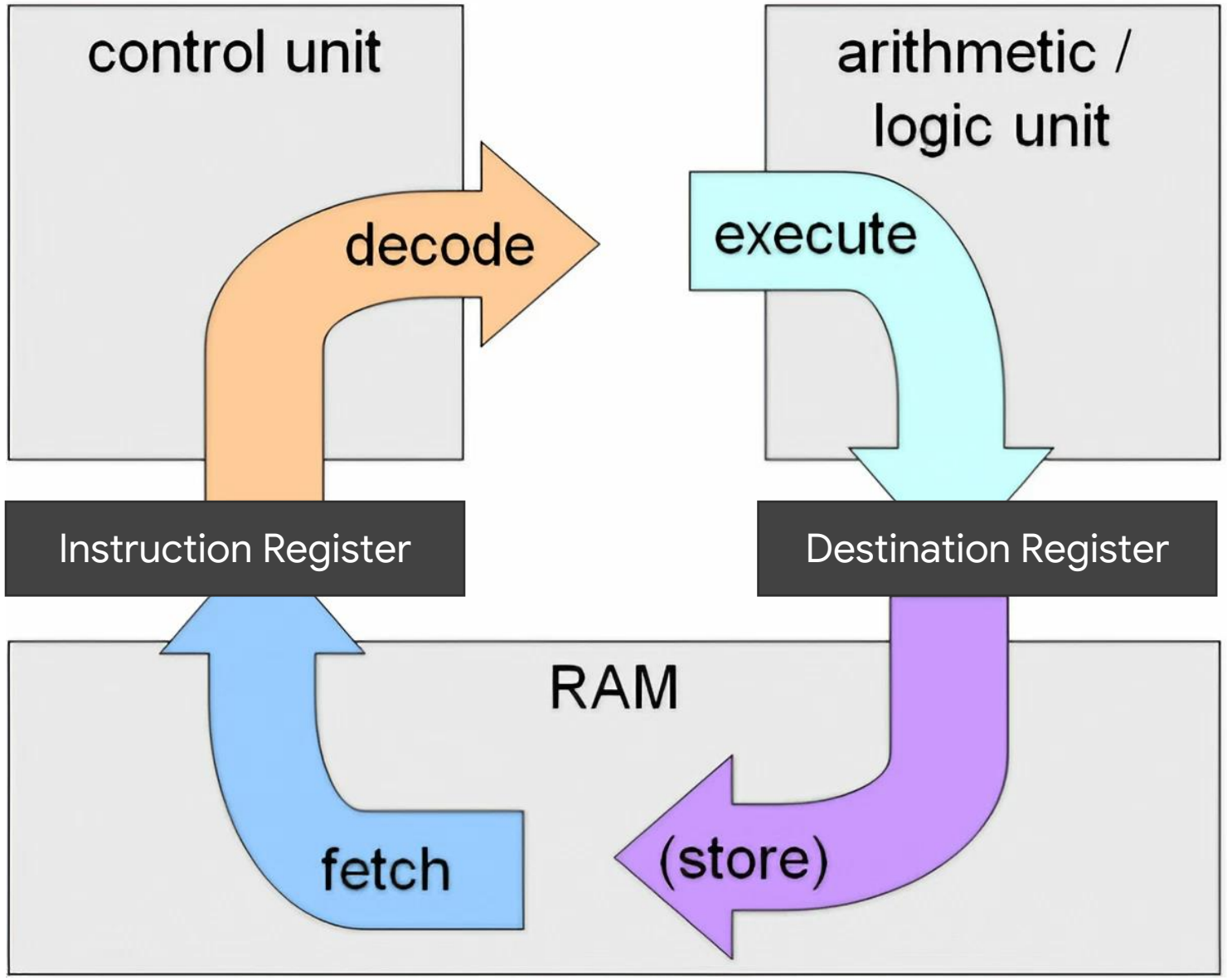
$$\text{cost} \propto \frac{\text{speed}}{\text{capacity}}$$

Instruction Set Architecture

Instruction Cycle

The sequence of events that take place when an instruction is executed by a computer's CPU.

ไม่ได้ใช้ second ในการบอก ค.เร็ว
ใช้ cycle ในการวัด ค.เร็ว



Characteristics

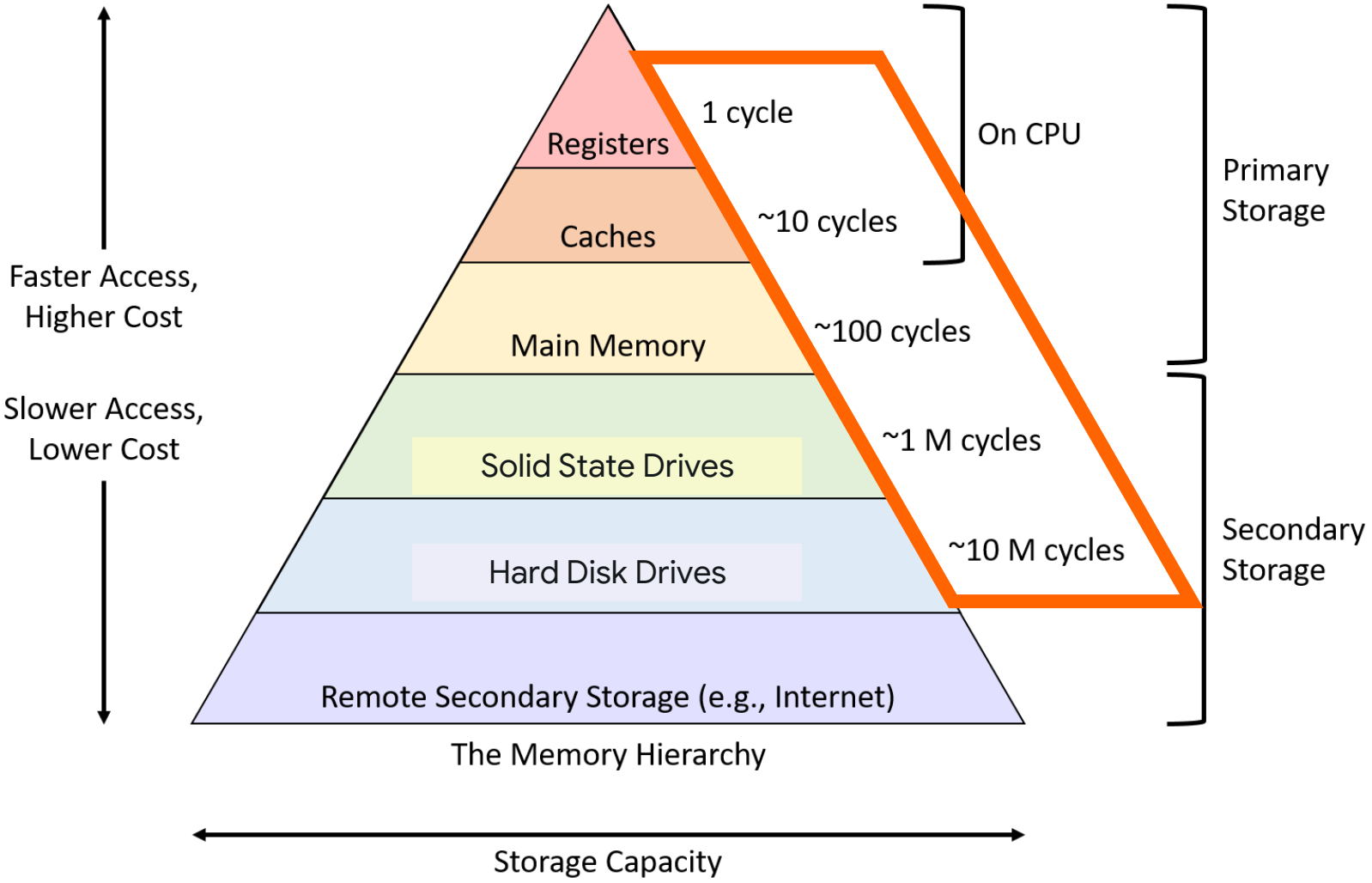
Speed

Different types of memory have varying access times.

The closer a memory is to the processor, the faster it can be accessed.

Frequently accessed data is available as quickly as possible.

Reducing the overall execution time of programs.



Characteristics

Capacity

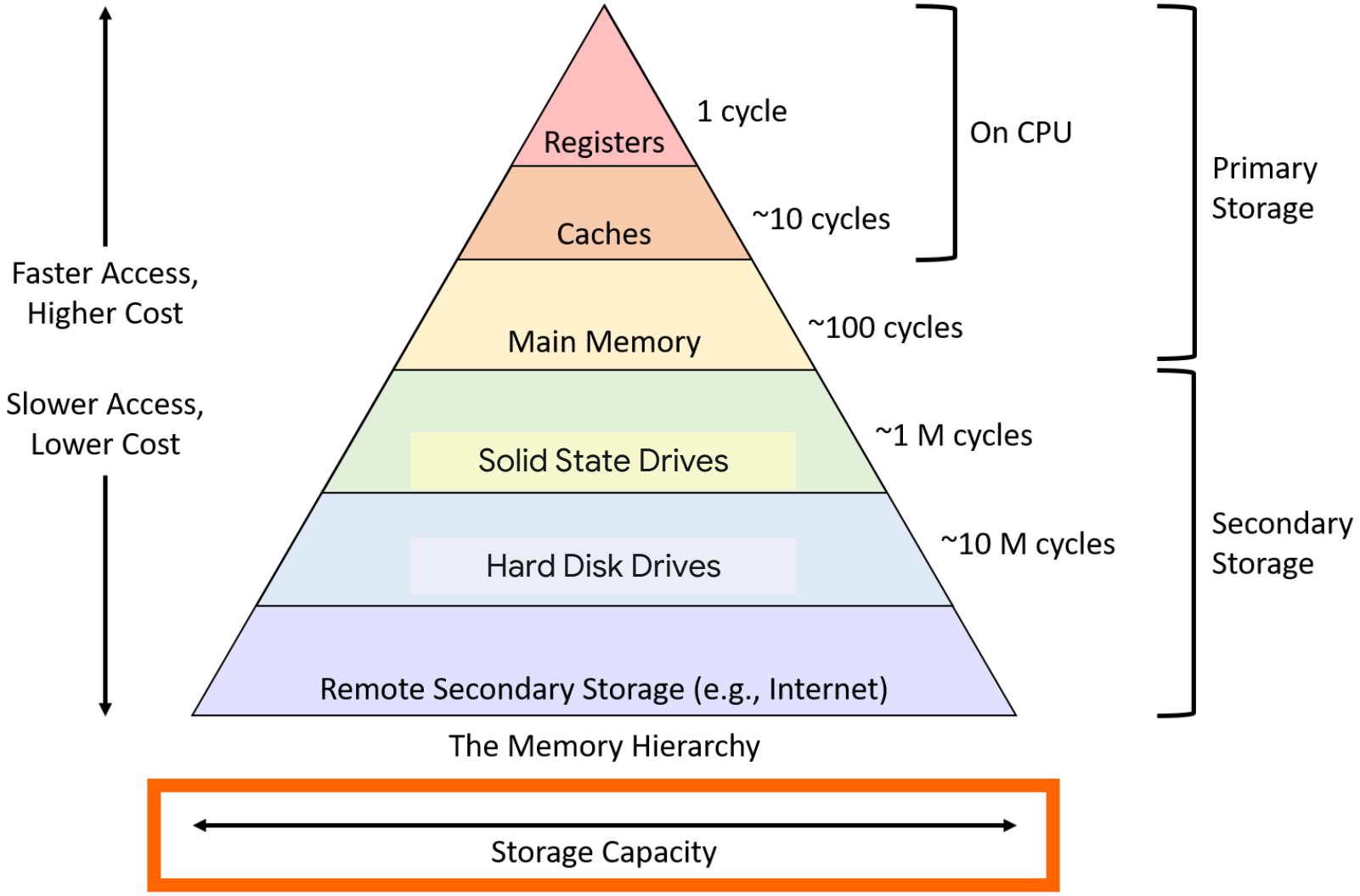
ถ้า ก. คุ ไม่พอใช้ จะใช้ฟันก็ของขั้นต่ำๆ ไป ↓

Faster memories tend to be more expensive and have lower capacities.

Slower memories are cheaper and offer larger storage capabilities.

Incorporating different levels of memory with varying capacities.

A balance between cost and capacity requirements.



Characteristics

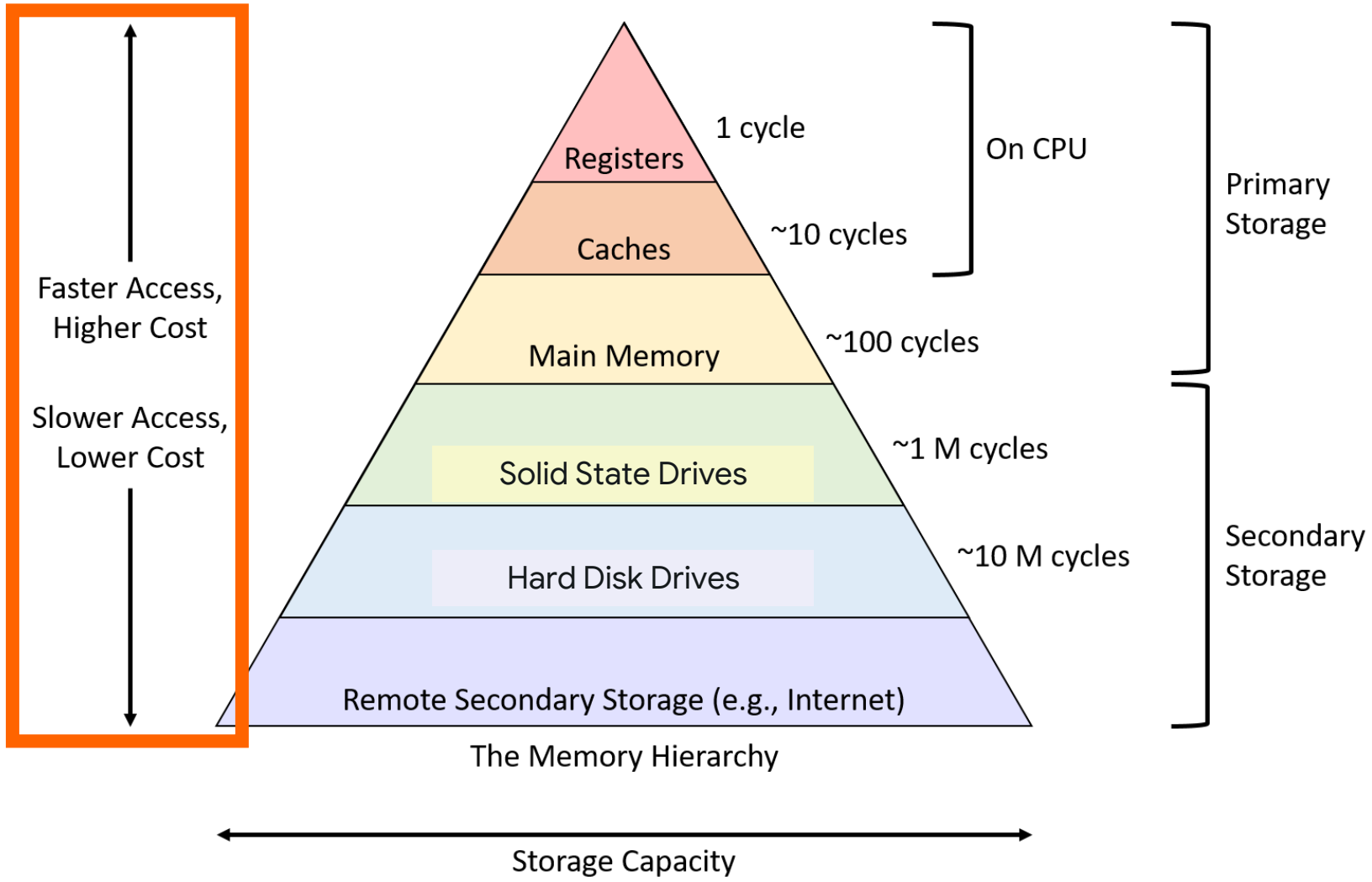
Cost

ราคา เมื่อเทียบกับ ค. ๆ

Utilizing cheaper memories for larger storage requirements.

Expensive memories for frequently accessed data.

The overall cost of the system can be optimized.



Memory Hierarchy

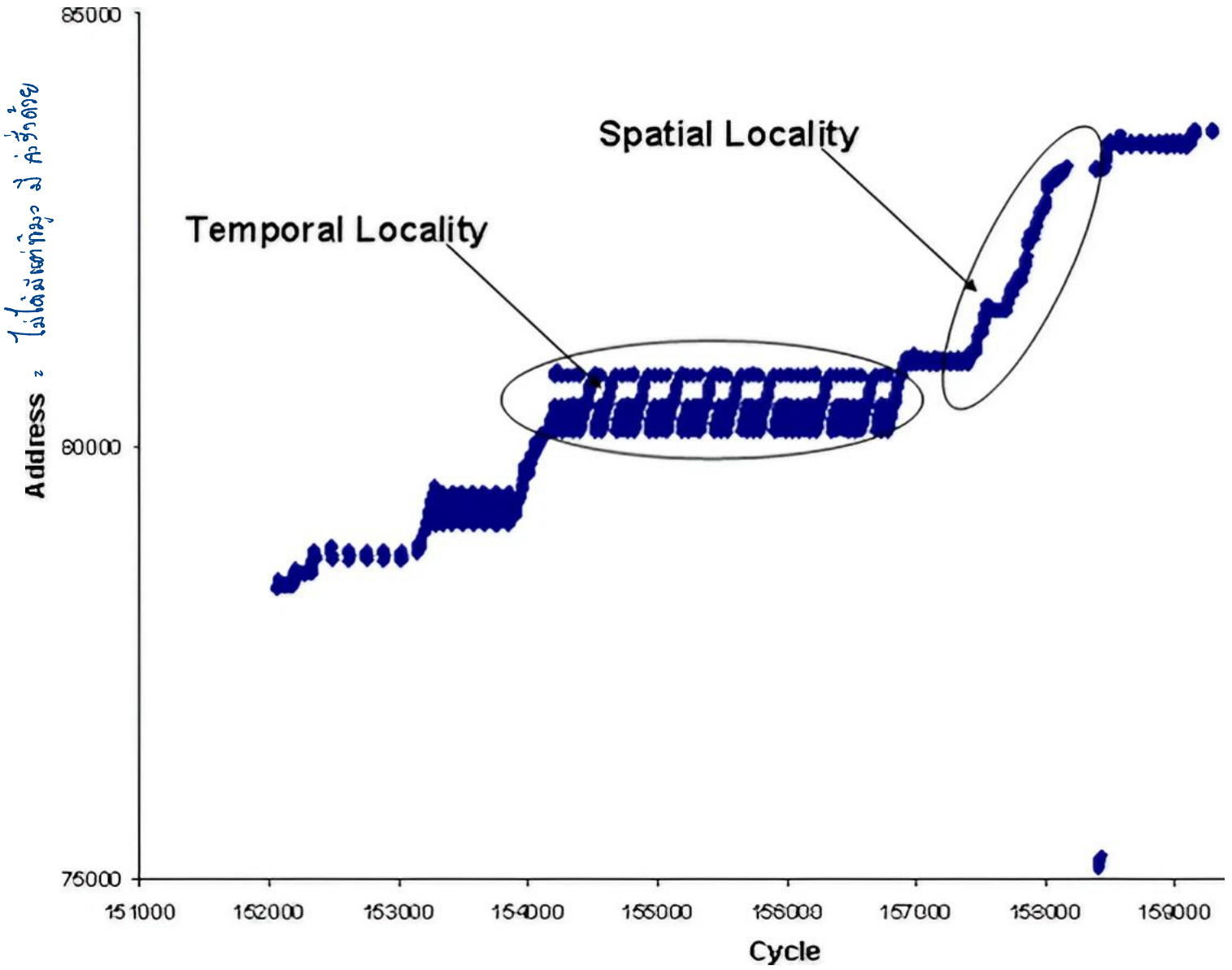
How should data be prioritized?

Locality

The tendency of computer programs to access memory locations that are close to each other in both time and space.

Helps optimize memory access patterns and improve overall system performance.

โดยทั่วไป com จะเข้าถึงที่อยู่ใกล้เคียงกัน



Outline

Locality

Temporal Locality

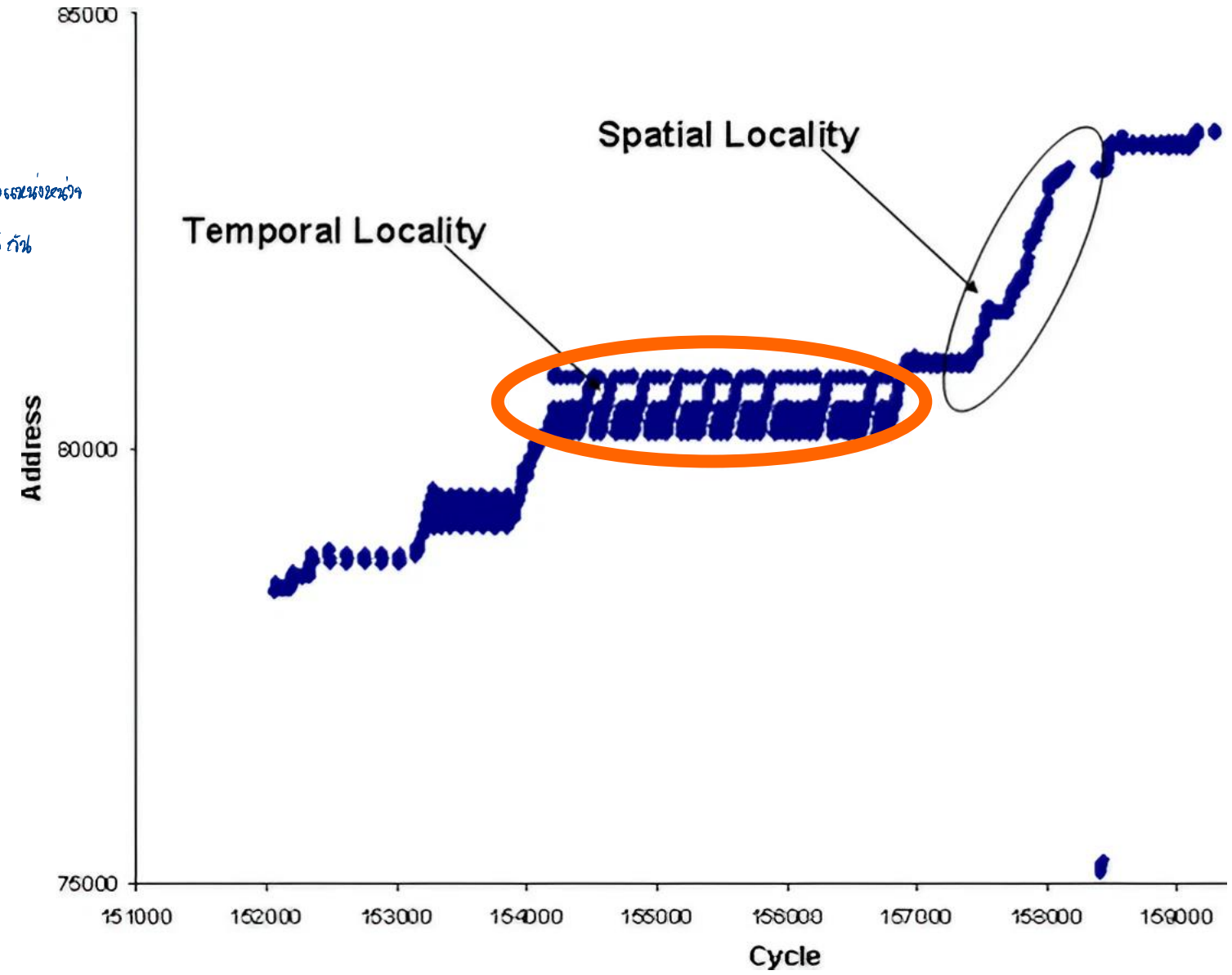
Spatial Locality

Temporal Locality

programme มักเข้าถึงที่อยู่ในตำแหน่งหน่วยความจำ เดิม ซ้ำๆ ในเวลาที่ใกล้กัน

The tendency of a program to access the same memory location repeatedly over a short period of time.

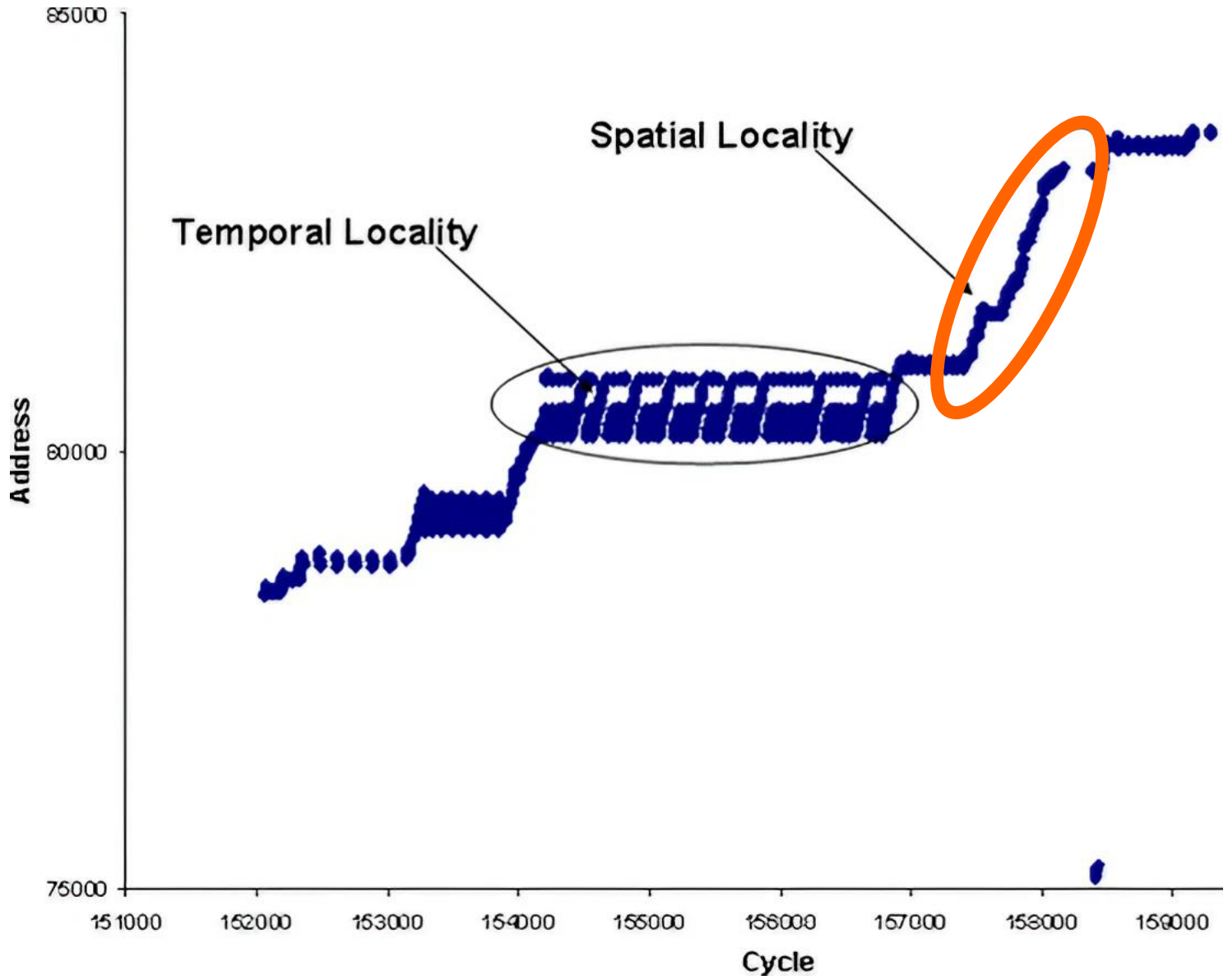
- In a loop, accessing the same array element repeatedly.
- Repeatedly executing the same set of instructions in a loop.
- A program repeatedly using a variable or data structure.



Spatial Locality

The tendency of a program to access memory locations that are close to each other in physical address space.

- Iterating over an array and accessing consecutive elements.
- Sequentially reading data from a file.
- Traversing a linked list and accessing the successive nodes.



เมื่อ เขาผ่าน ไป เมื่อ ๆ จะ มีการเข้า ถึง ข้อมูล ในช่วงถัดๆ ไป เร็ว ๆ

The End

Review

- Memory Hierarchy
- Speed, Capacity, and Cost
- Temporal and Spatial Locality