CUDA Programming – CPU vs GPU

by
Dr. Nileshchandra Pikle
Assistant Professor
&
"Certified CUDA instructor by NVIDIA"

Email ID: nilesh.pikle@gmail.com
Mob. No: +91 7276834418

Contents

- Limitations of Multi-core CPU
- Introduction to Graphics Processing Unit (GPU)
- GPU Accelerated Applications
- CPU vs GPU
- GPU Hardware Architecture

Multi-core CPUs

Intel Xeon
E7- 8855 v4

14 cores
@ 2.80 GHz

Intel Xeon
E7- 4850 v4

16 cores
@ 2.80 GHz

Intel Xeon
E7- 8867 v4

18 cores
@ 3.3 GHz

Intel Xeon
E7- 8870 v4

20 cores
@ 3.0 GHz

Intel Xeon
E7- 8880 v4

22 cores
@ 3.3 GHz

Intel Xeon
E7- 8890 v4

24 cores
@ 3.4 GHz

Intel Core
i7-9700k
8-cores

@4.9 GHz

And so on…

If multi-core is solution!
Why can’t we increase to thousands of cores?

Why only few tens?

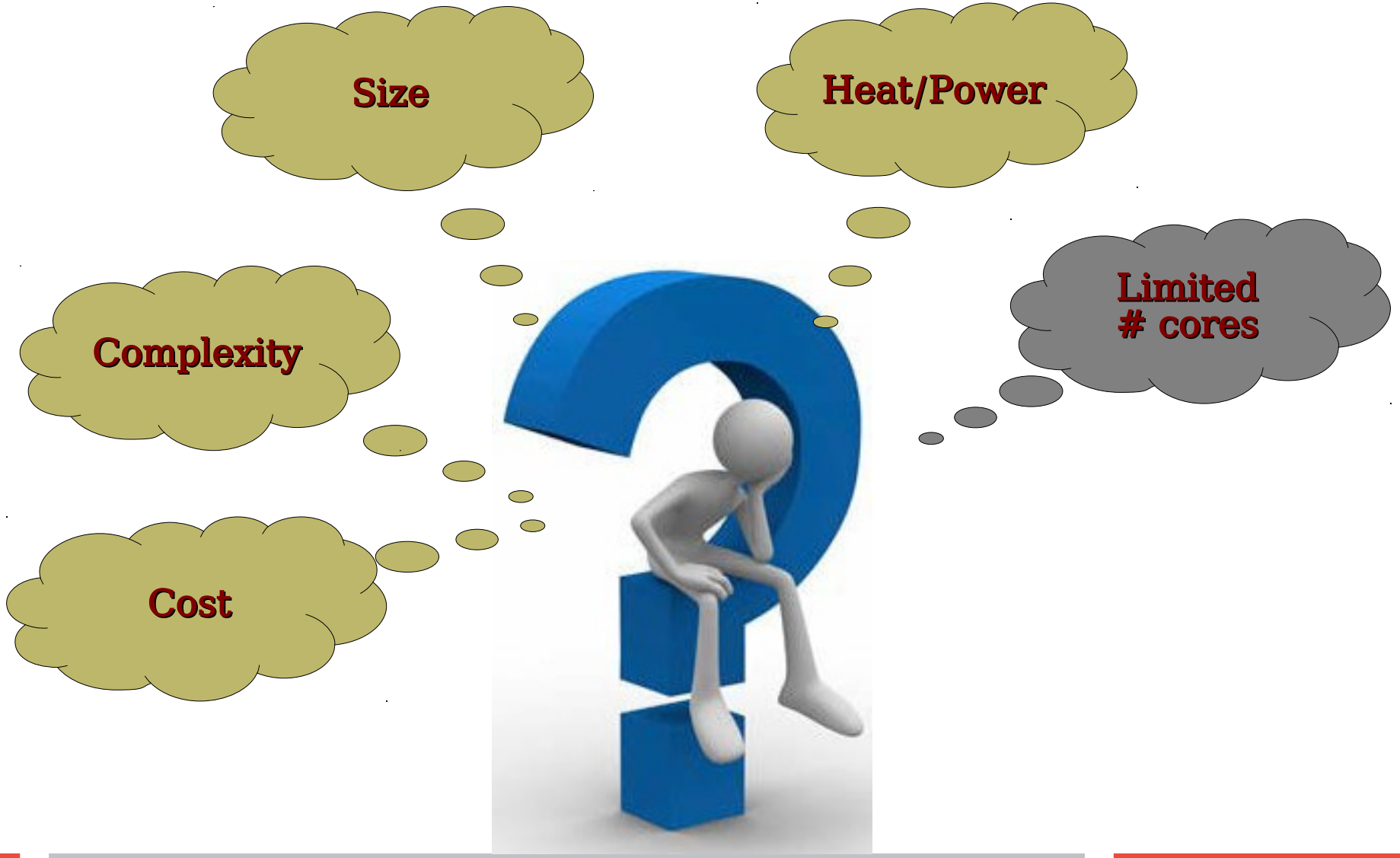
If multi-core is solution!
Why can’t we increase hundreds of cores?

If multi-core is solution!
Why can’t we increase hundreds of cores?

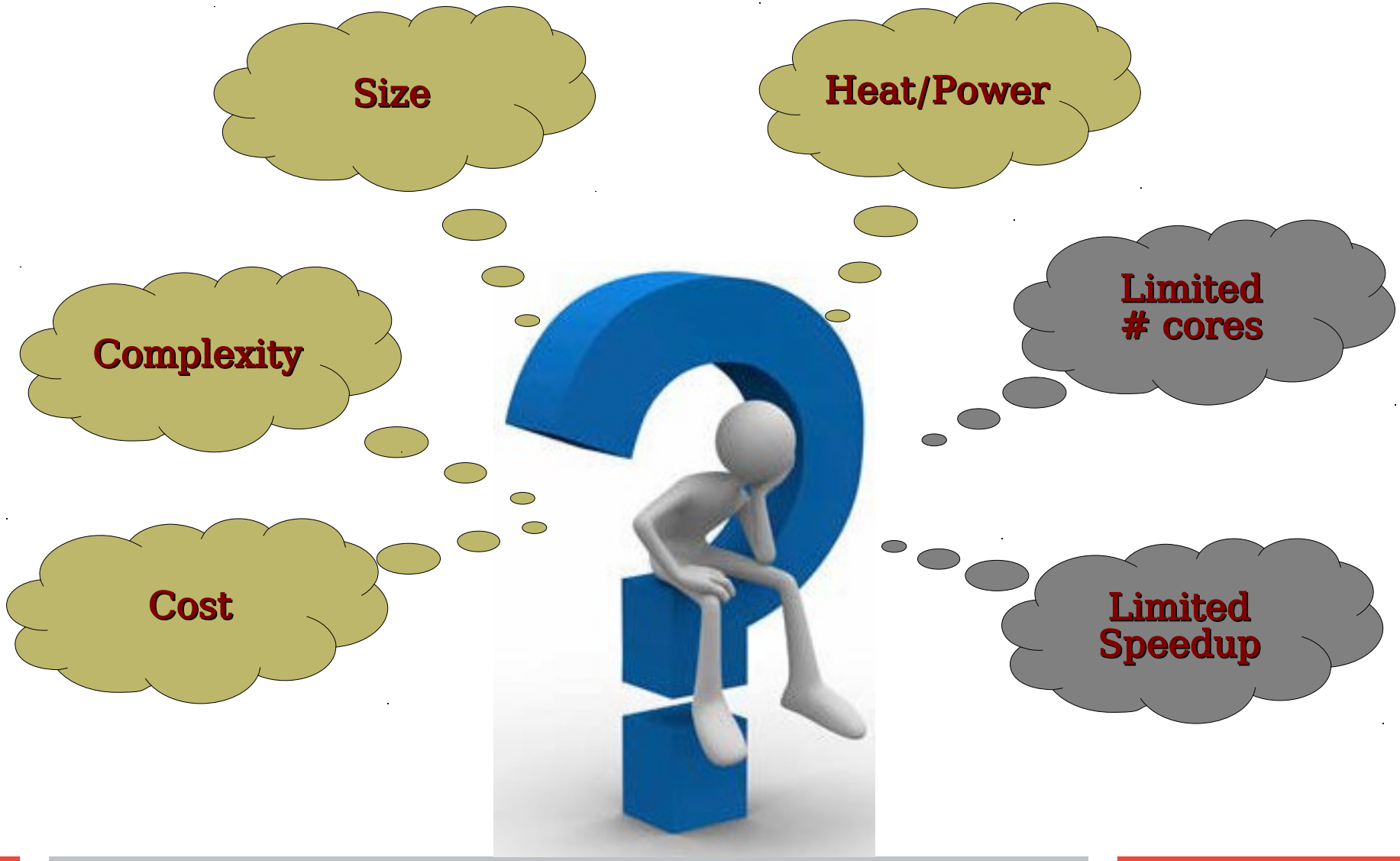
If multi-core is solution!
Why can’t we increase hundreds of cores?

If multi-core is solution!
Why can’t we increase hundreds of cores?

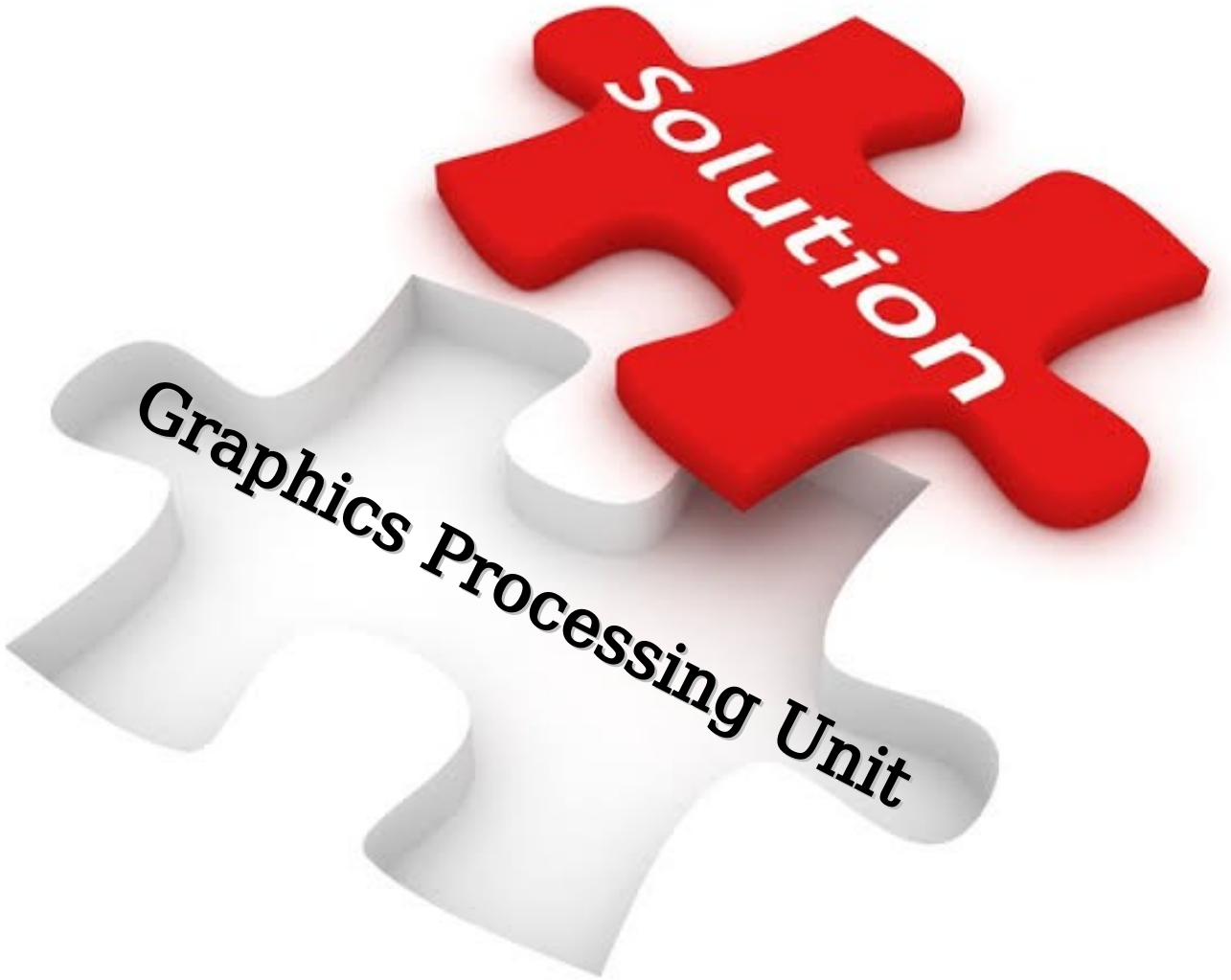
If multi-core is solution!
Why can’t we increase hundreds of cores?



If multi-core is solution!
Why can’t we increase hundreds of cores?



Other solution!



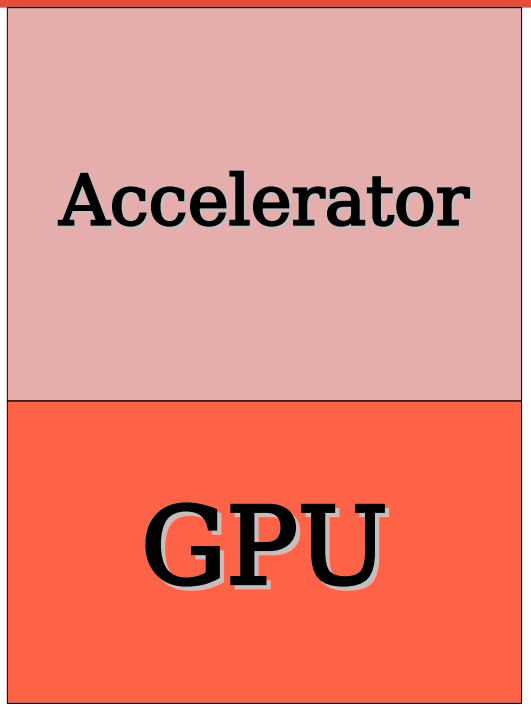
History of GPU

- The term GPU has been used from 1980s
- Popularised by **NVIDIA** in 1999 who marketed the Geforece 256 as **“The world’s first GPU”**
- Initially intended for graphics related computing
  - To accelerate the gaming and animation performance
- In 2007 NVIDIA launched **Compuet Unified Device Architecture** (CUDA) which enabled General Purpose Computing.
- Now it is referred as **General Purpose GPU (GPGPU)**

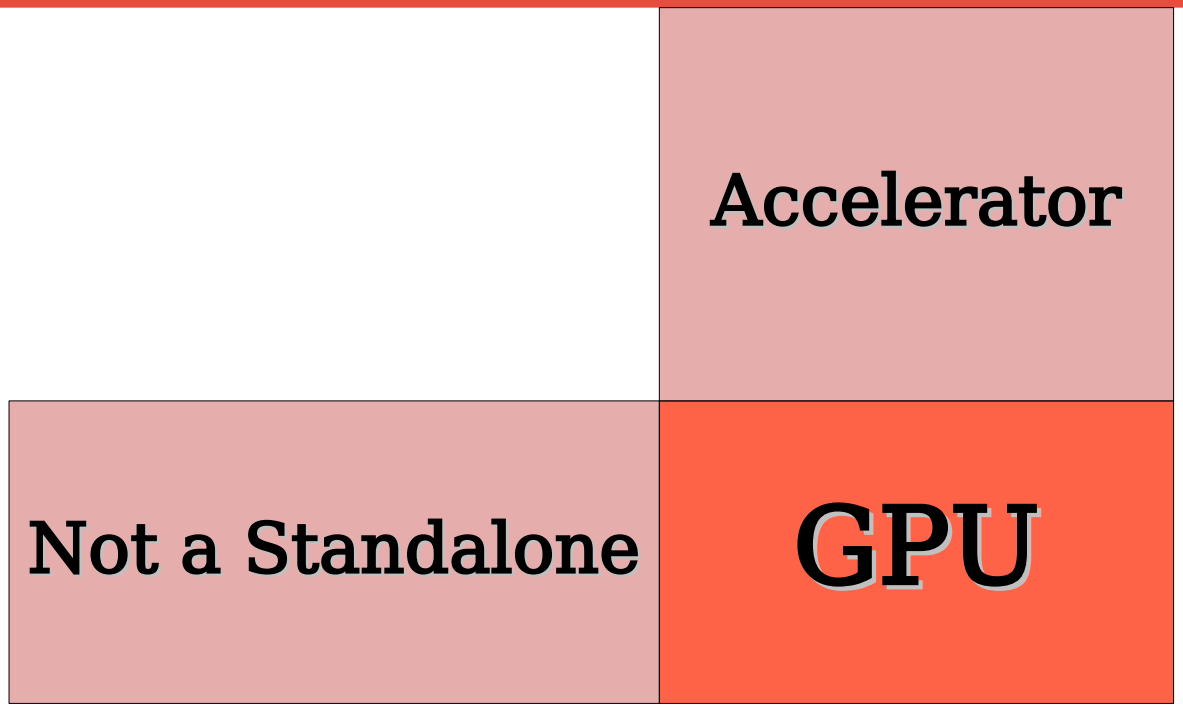
Graphics Processing Unit (GPU)

GPU

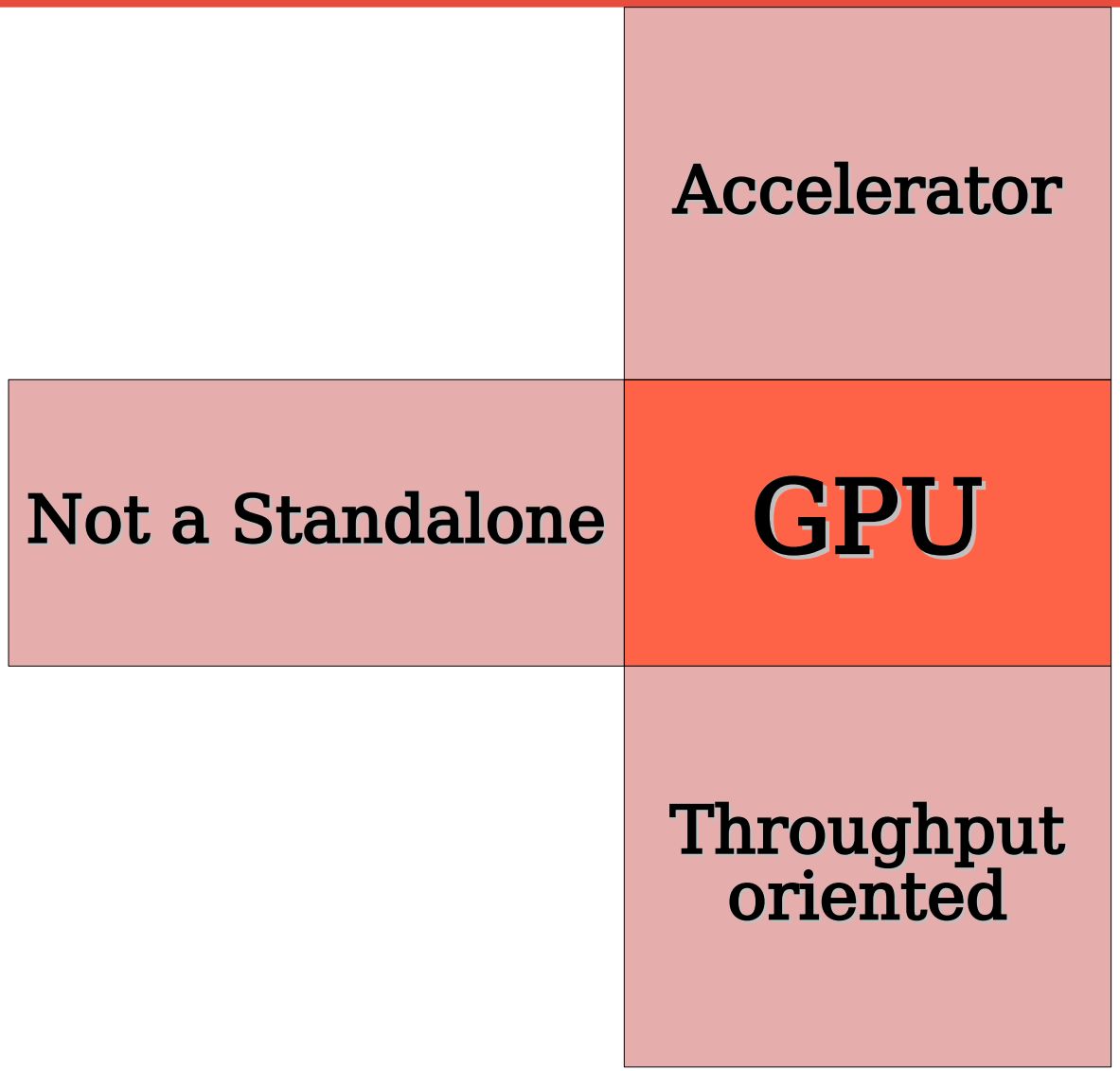
Graphics Processing Unit (GPU)



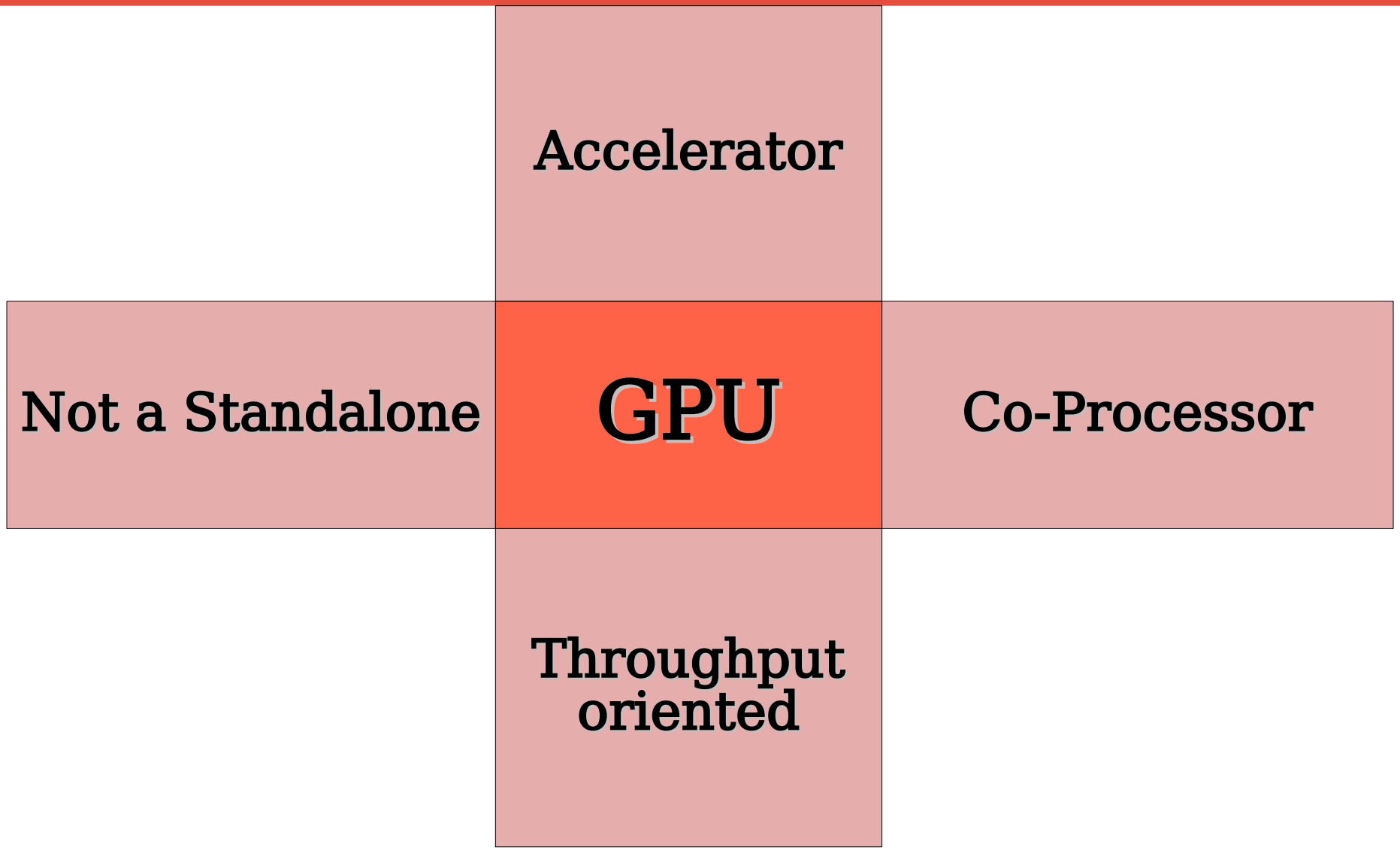
Graphics Processing Unit (GPU)



Graphics Processing Unit (GPU)

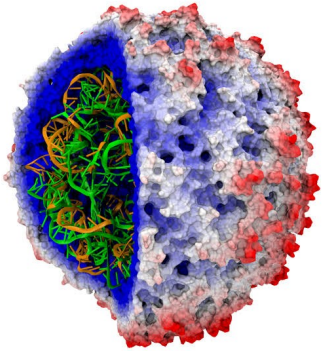


Graphics Processing Unit (GPU)



Some of the GPU application areas

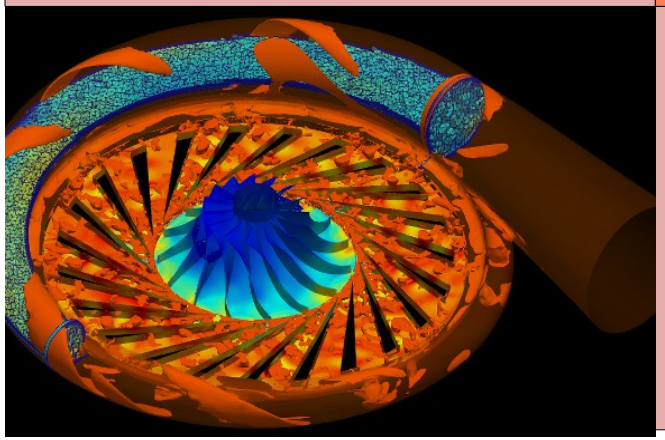
Games
&
Movies



CFD

GPU
Applications

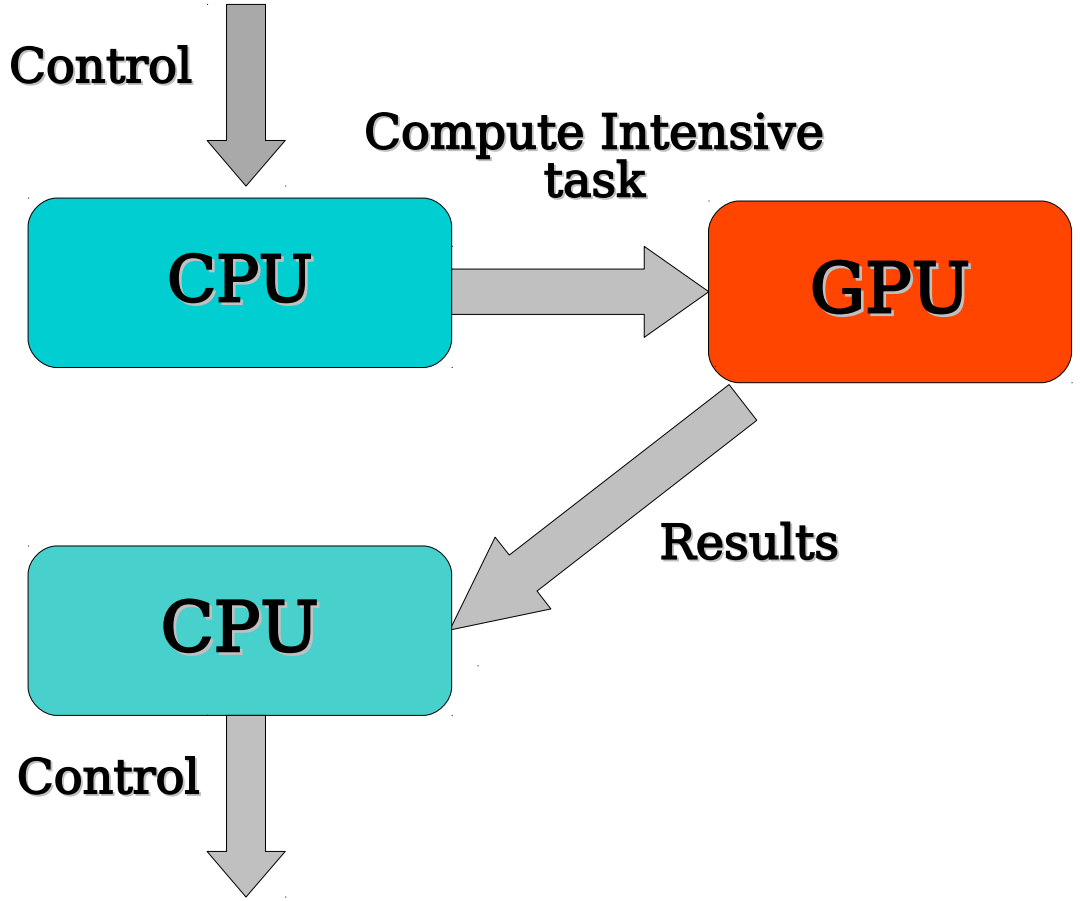
Molecular
Modeling



Artificial
Intelligence

What is GPU

- Acts as an accelerator/Co-processor



What is GPU

- Acts as an accelerator/Co-processor
- **Heterogeneous Computing Architecture**

CPU

GPU

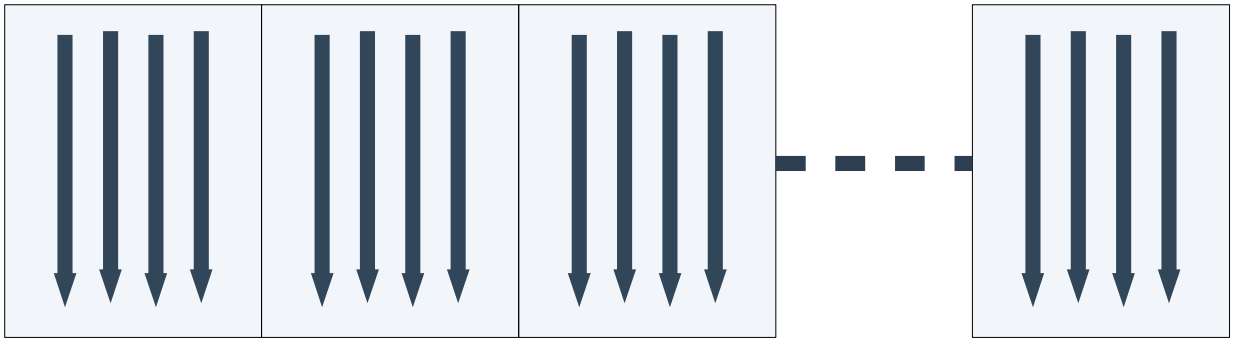
What is GPU

- Acts as an accelerator/Co-processor
- Heterogeneous Computing Architecture
- **Not an intelligent device**

Takes orders from the CPU

What is GPU

- Acts as an accelerator/Co-processor
- Heterogeneous Computing Architecture
- Not an intelligent device
- **Contains thousands of cores over millions of threads can be launched**



What is GPU

- Acts as an accelerator/Co-processor
- Heterogeneous Computing Architecture
- Not an intelligent device
- Contains thousands of cores over millions of threads can be launched
- **Not a standalone device**

Cannot replace CPU by GPU

CPU vs GPU

- **A single core CPU**

1. Powerful
2. Need a lot of power
3. Complex control hardware
4. Good performance

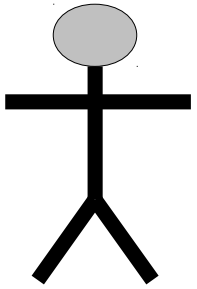
- **Many-core GPU**

1. Less powerful but lot many cores
2. Require less power
3. Simple control hardware
4. Good throughput

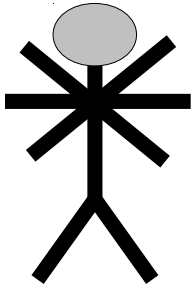
CPU vs GPU

- Task (TA)=> 400 meter Hole
- Efficiency (E)
- Time (T)

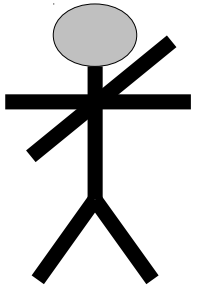
Ideal multi-core CPU



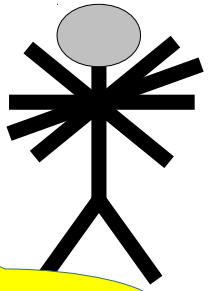
E = 2 Meter/Hrs
T = ?



E = 6 Meter/Hrs
T = ?



E = 4 Meter/Hrs
T = ?



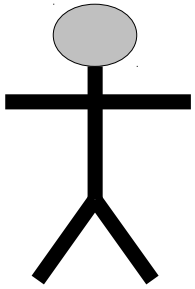
E = 8 Meter/Hrs
T = ?

Note: These are very Powerful cores
In terms of frequency, transistors,
IPC, branch prediction etc.

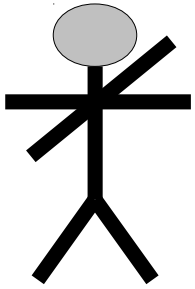
CPU vs GPU

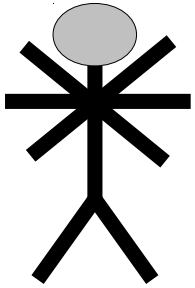
- Task (TA)=> 400 meter Hole
- Efficiency (E)
- Time (T)

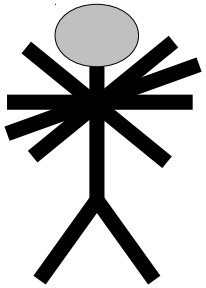
Ideal multi-core CPU



E = 2 Meter/Hrs
T = 200 Hrs



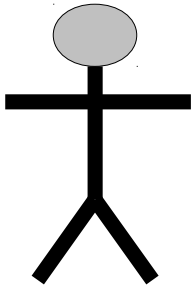




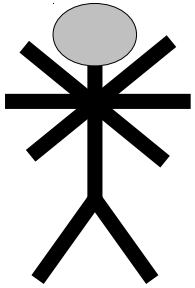
CPU vs GPU

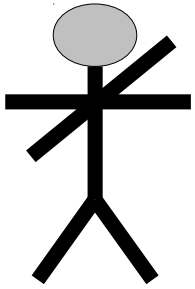
- Task (TA)=> 400 meter Hole
- Efficiency (E)
- Time (T)

Ideal multi-core CPU

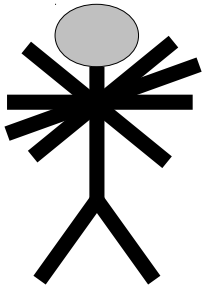


E = 2 Meter/Hrs
T = 200 Hrs





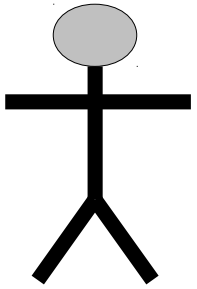
E = 4 Meter/Hrs
T = 100 Hrs



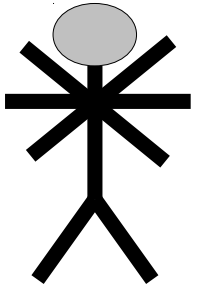
CPU vs GPU

- Task (TA)=> 400 meter Hole
- Efficiency (E)
- Time (T)

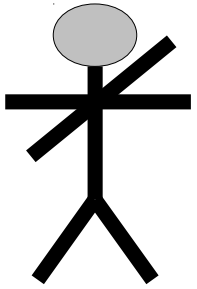
Ideal multi-core CPU



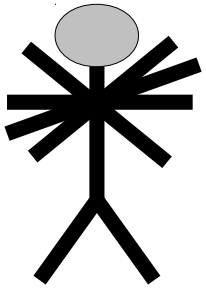
E = 2 Meter/Hrs
T = 200 Hrs



E = 6 Meter/Hrs
T = 66.66 Hrs



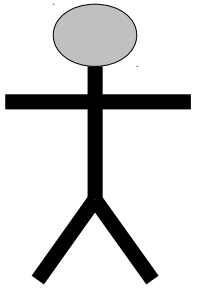
E = 4 Meter/Hrs
T = 100 Hrs



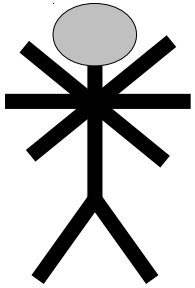
CPU vs GPU

- Task (TA)=> 400 meter Hole
- Efficiency (E)
- Time (T)

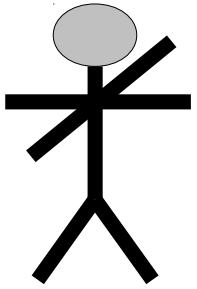
Ideal multi-core CPU



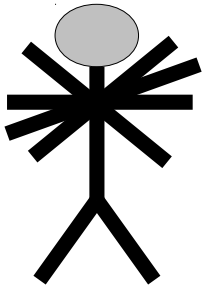
E = 2 Meter/Hrs
T = 200 Hrs



E = 6 Meter/Hrs
T = 66.66 Hrs



E = 4 Meter/Hrs
T = 100 Hrs

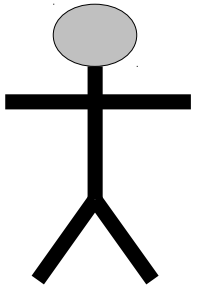


E = 8 Meter/Hrs
T = 50 Hrs

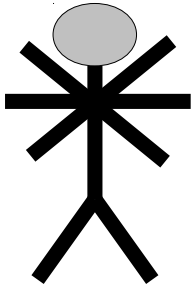
CPU vs GPU

- Task (TA)=> 400 meter Hole
- Efficiency (E)
- Time (T)

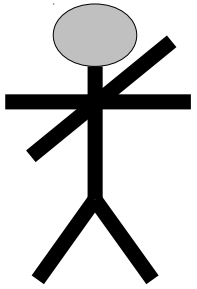
Ideal multi-core CPU



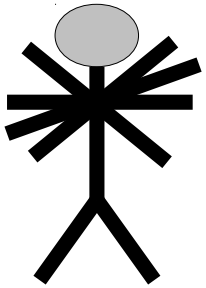
E = 2 Meter/Hrs
T = 200 Hrs



E = 6 Meter/Hrs
T = 66.66 Hrs



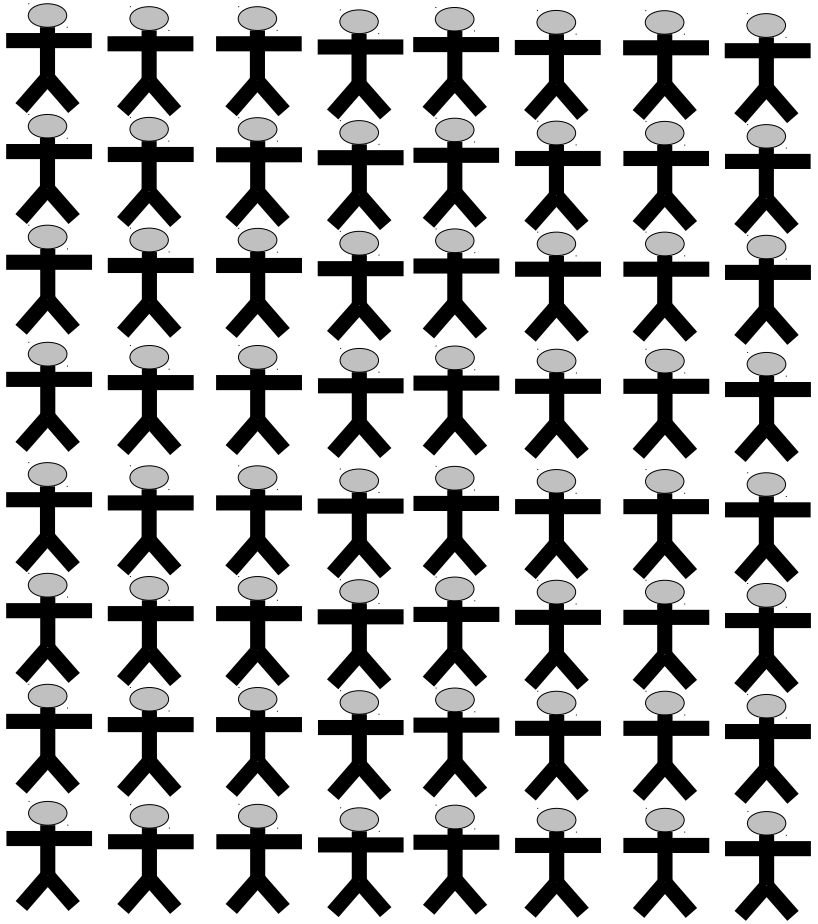
E = 4 Meter/Hrs
T = 100 Hrs



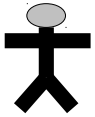
E = 8 Meter/Hrs
T = 50 Hrs

Latency Oriented CPUs!!!

CPU vs GPU

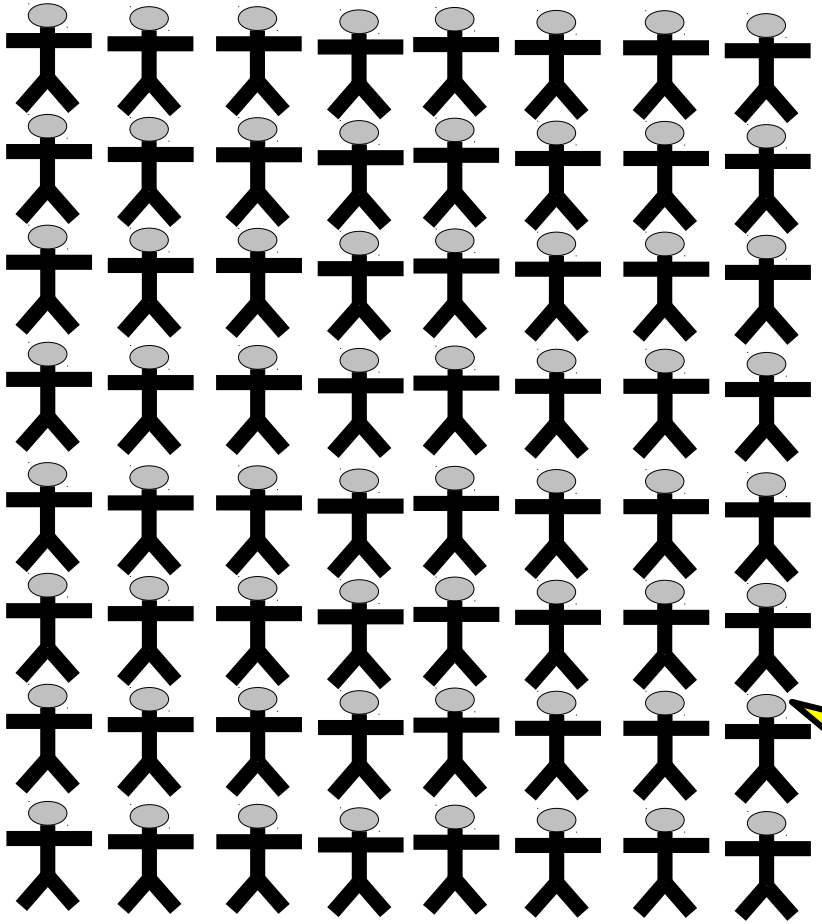


Ideal many-core GPU

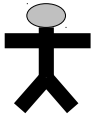
E = 0.25 Meter/Hrs for one 

T = ?

CPU vs GPU



Ideal many-core GPU

E = 0.25 Meter/Hrs for one 

T = 400 / (64 x 0.25) = 25 Hrs

Instead of making CPU faster and complex
Use smaller, lightweight cores

Note: These are very light weight cores
In terms of frequency, transistors,
IPC, branch prediction etc.

Throughput Oriented GPUs!!!

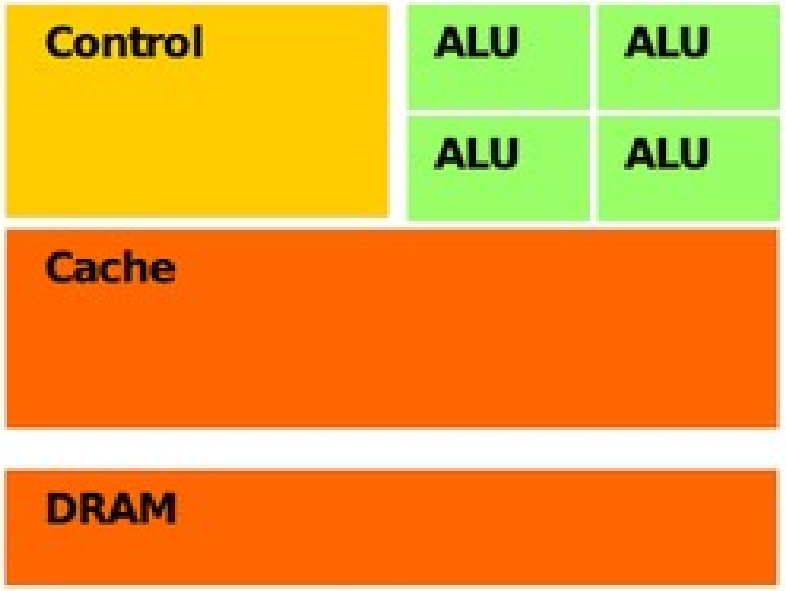
GPU

GPU is a throughput oriented device

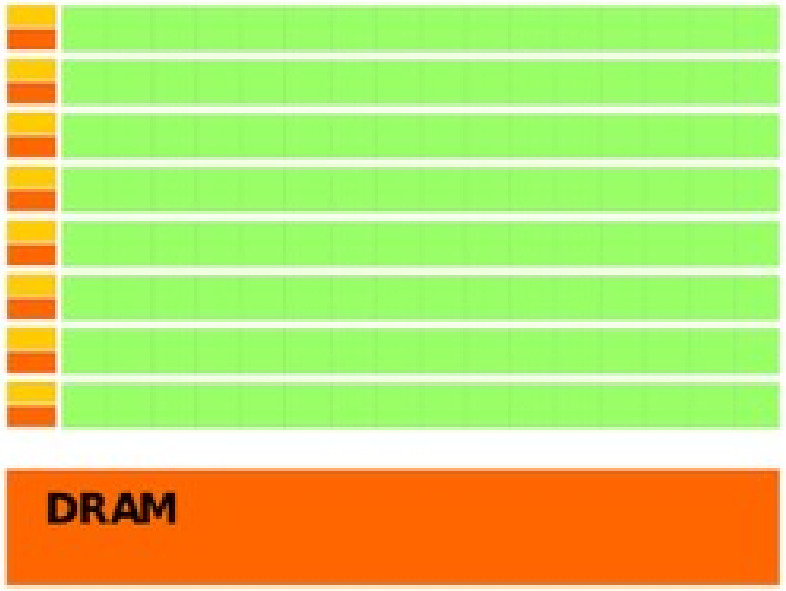
CPU vs GPU

	CPU	GPU
Number or cores	1, 2, 8 or Few hundred	Thousands
Intelligence	More	Less
Standalone	Yes	No
Intends	Latency oriented	Throughput oriented
Core clock rates	Higher Eg. 2.3 GHz	Lower Eg. 900 MHz
Efficiency	Sequential	Parallel
Power	More powerful cores	Less powerful
Usage	General purpose	Special purpose
Role	Processor	Co-processor

CPU vs GPU



CPU



GPU

- Sophisticated control unit
- Larger cache
- Less area for cores
- More DRAM

- Less Sophisticated control unit
- Smaller cache
- More area for cores
- **Often** less VRAM than CPU

GPU Hardware Architecture

GPU Hardware Architecture

GPU Hardware Architecture

Fermi Streaming Multiprocessor (SM)

We can dive into CUDA programming

Play
Leonardo GPU versus CPU.mp4