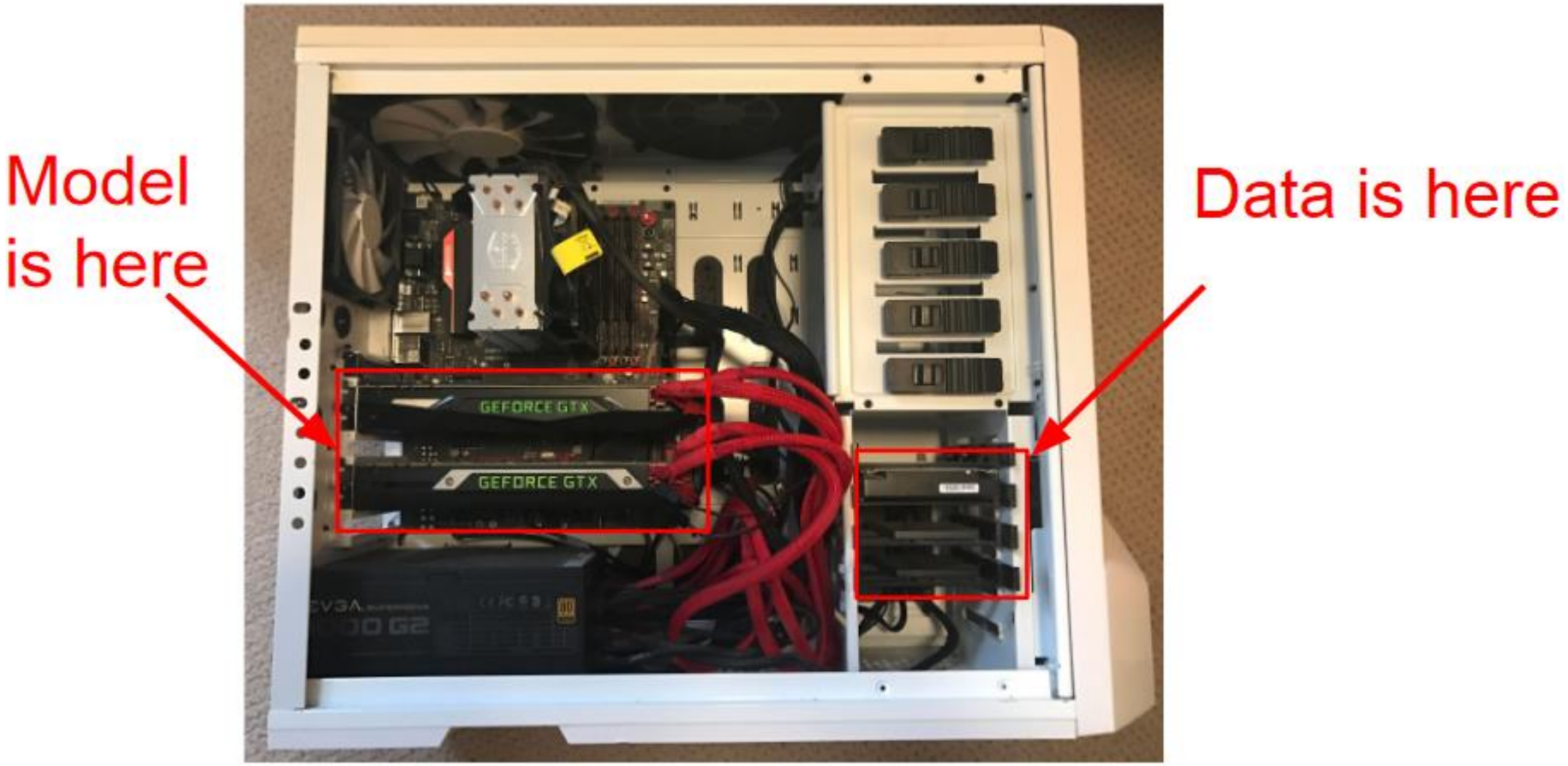
Aside 4:
Hardware for Deep Learning

GPU vs. CPU

- ***The GPU resides on a separate board***

  Almost an independent computer

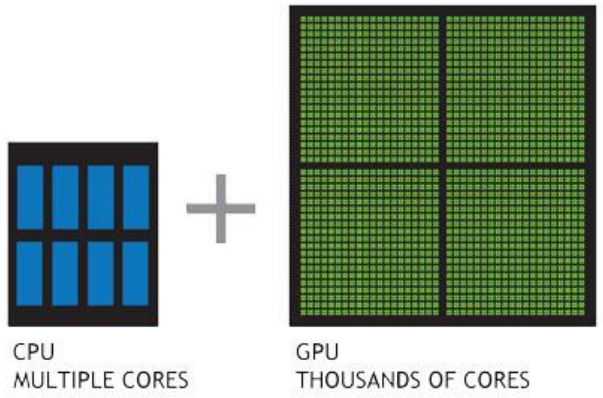


[image http://cs231n.stanford.edu/slides/2021/lecture_6.pdf]

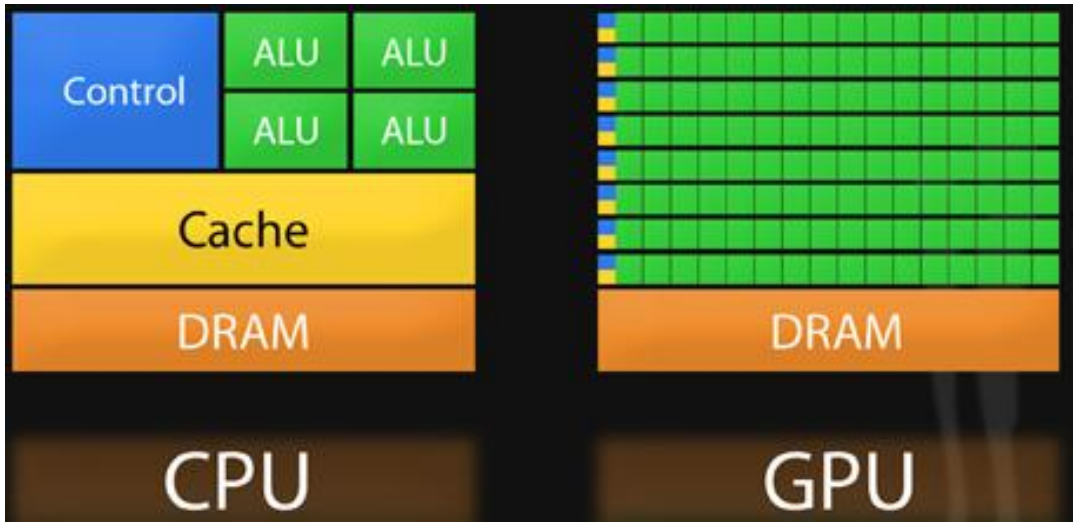
GPU vs. CPU

- ***Different hardware architectures***

  For different computing paradigms



[images from http://www.nvidia.com/docs/]



GPU vs. CPU

- ***Different hardware architectures***

  For different computing paradigms

	Cores	Clock Speed	Memory	Price	Speed
CPU (Intel Core i7-7700k)	10	4.3 GHz	System RAM	$385	~640 GFLOPs FP32
GPU (NVIDIA RTX 3090)	10496	1.6 GHz	24 GB GDDR 6X	$1499	~35.6 TFLOPs FP32
GPU (Data Center) NVIDIA A100	6912 CUDA, 432 Tensor	1.5 GHz	40/80 GB HBM2	$3/hr (GCP)	~9.7 TFLOPs FP64 ~20 TFLOPs FP32 ~312 TFLOPs FP16
TPU Google Cloud TPUv3	2 Matrix Units (MXUs) per core, 4 cores	?	128 GB HBM	$8/hr (GCP)	~420 TFLOPs (non-standard FP)

[image http://cs231n.stanford.edu/slides/2021/lecture_6.pdf]

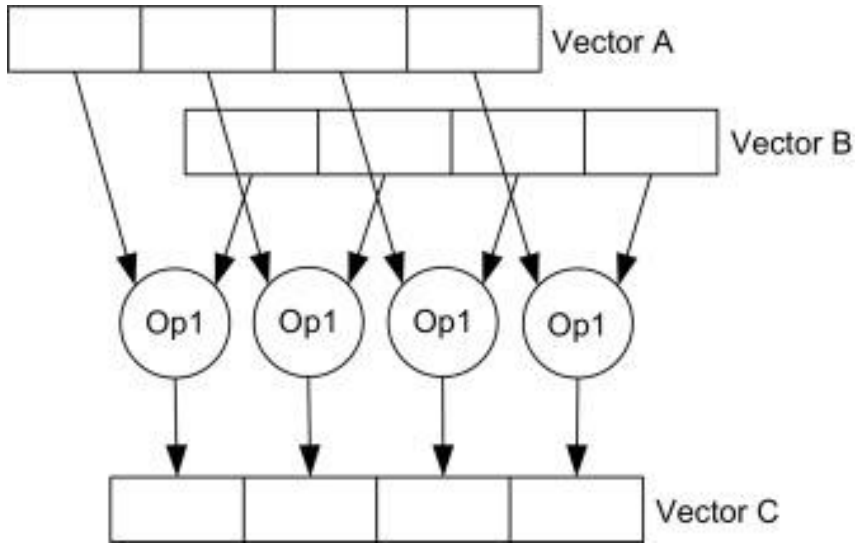
SIMT Parallelism

- ***Single Instruction, Multiple Data*** (SIMD)

  Execution is parallel

  All cores are executing the same instruction, in sync

  Each core works on specific data



[images from https://www.sciencedirect.com/topics/computer-science/single-instruction-multiple-data]

SIMT Parallelism

- ***Single Instruction, Multiple Threads*** (SIMT)

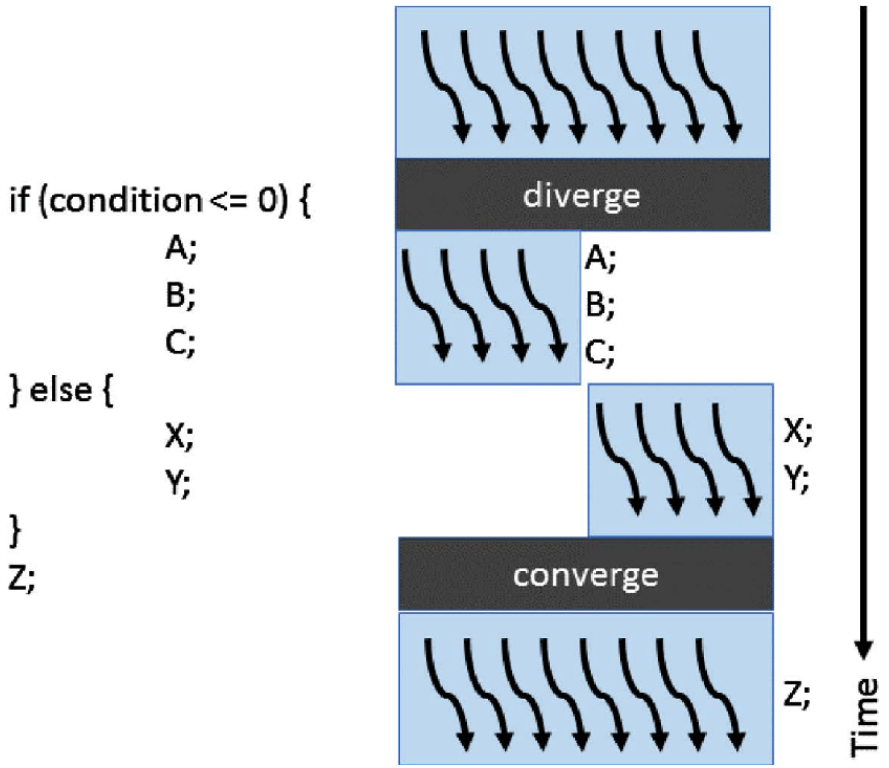
  Execution is parallel

  All ***active*** cores are executing the same instruction, in sync

  Each core works on specific data

  The control system activates and deactivates cores on each *execution branch*

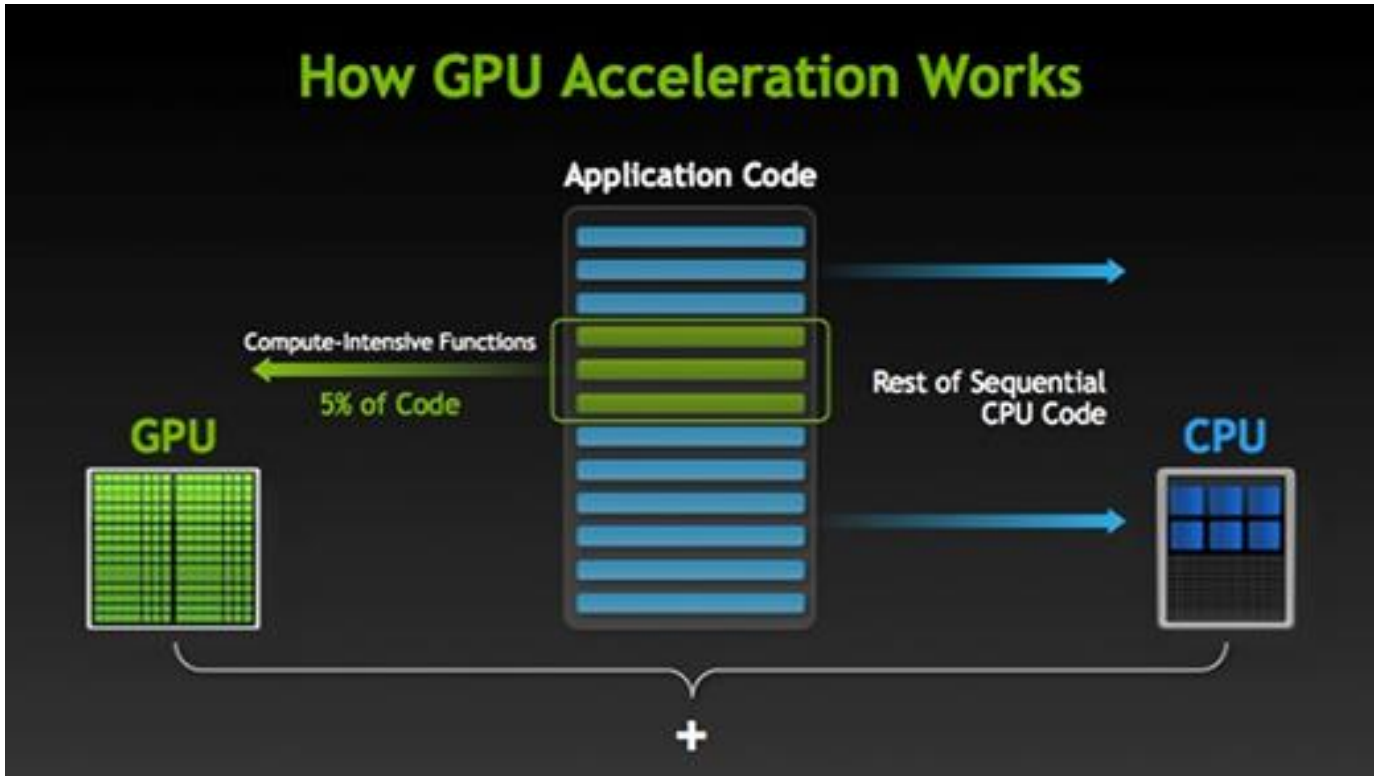
  Moral: any computation might be performed, but divergent ones will be sequentialized



[images from https://www.sciencedirect.com/topics/computer-science/single-instruction-multiple-data]

Selective parallelization

Not all parts of a program are worth executing in parallel...



[images from http://www.nvidia.com/docs/]

TensorFlow and GPUs

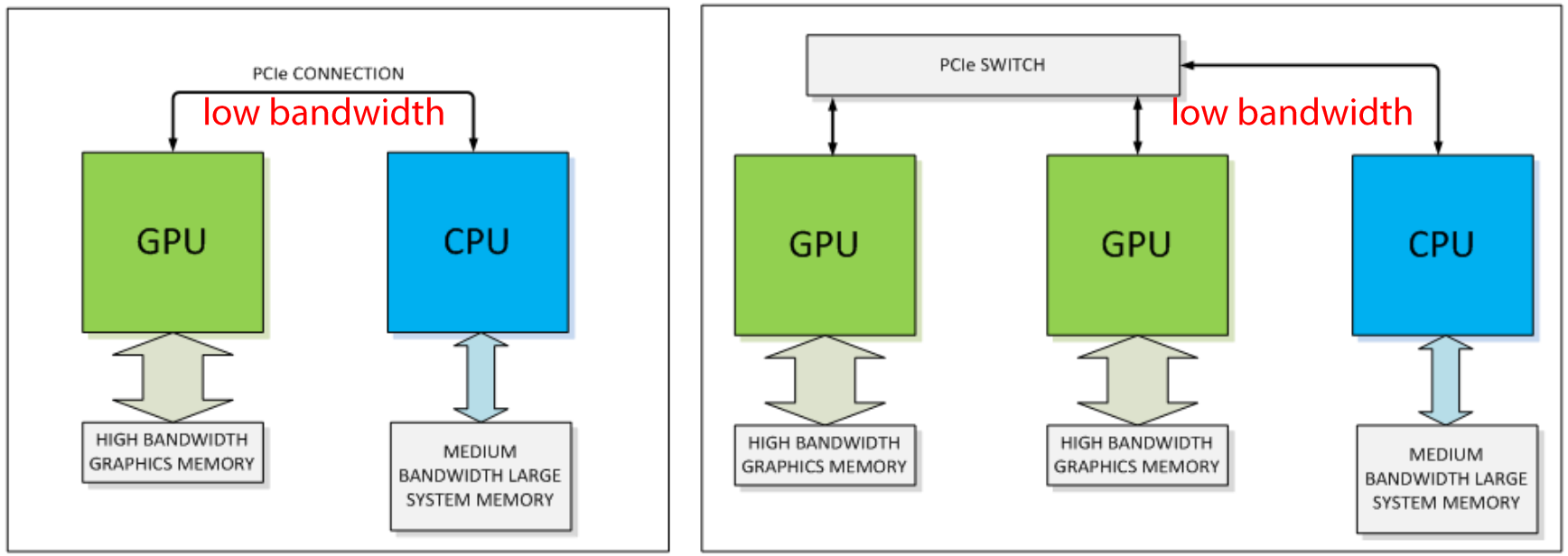
- TF computations are optimized to be run on **GPUs**

  For the programmer, these implementation details are (mostly) **transparent**

  TF can also run on the CPU only, but with lower performance.

- TF automatically manages **memory transfers** to/from GPUs

  Memory transfers are very costly, due to low bandwidth PCIe



[NVIDIA.com]