Questions answered thorugh chat window during session #2 on 02Feb2022	
Question	**Answer**
How to access the slack to get the slides. I am not getting.	Please put in the search window of slack the channel name that you want to join. Like #presentation where the presentations are kept. When you click it should allow you to join. For more details on how to use slack please refer to https://slack.com/intl/en-in/help/categories/360000049063
I couldn't find last lecture material in slack. i don't know where they post.	They get posted in #presentation slack channel.
Are these sample programs also shared to	All codes will be shared with participants
Professor is it necessary to learn the linux commands ?	Any researcher in this field for practical applciations will end up using Linux either now or later. The ecosystem in AI like containers etc are mroe matured in Linux.
how and where to do this hands-on along with the instructor?	This is a demo, not hands-on to be done by the participant
Can we students access PARAM Shivay through SSH?	The participants of this course do not have access to PARAM Shivay
Does a raspberry pi have a GPU?	Yes, on some models, but not a nvidia GPU
Is DGX a device?	DGX is a supercomputer server from NVIDIA : https://www.nvidia.com/en-in/data-center/dgx-systems/
What is VPU?	Vision Processing unit. Primarily dedicated for Vision processing tasks only.
why are we learning CPU and GPU in AI ?	Because AI algorithms run on CPU and GPU
can we use CUDA toolkit in windows without having GPU. How to use ??	CUDA is for NVIDIA GPU only
CUDA toolkit can help us in programming?	CUDA API can be used of parallel programming on GPU. Yes.
What is the simple example to differentiate between task and data parallelism?	Task parallelism is where you do different tasks like one thread may do multiplication and other may do sorting. Data parallelism is when you do same operation on different data like adding 2 arrays .
so, when do one use CUDA v/s Tensor core?	Tensor cores are primarily meant for Matrix Multiplication only. CUDA cores are general cores where you can run other instructions as well. Tensor cores are useful for AI as they most AI algorithms can be converted to matrix multiplication
How does Tensor core works?	Tensor core perform the matrix operation in parallel in one or few clock cycles . For which otherwise you would have written for loops to do multiplication. You can think of them as specialized cores that do matrix multiplication only but in parallel very fast.

Question	Answer
If we use GPU while training a DL algorithm, do we need to parallelize my code?	No . All frameworks like TensorFlow and PyTorch have behind the scene call to NVIDIA parallel library. If you have a GPU you will be able to use them without writing parallel code. The Pytorch session in next lectures will cover these
is data parallelism same task for different data?	That is correct
How is the GPU connected to the CPU - can you pls share some more info on this system setup?	In desktop and server the GPU is attached to CPU via a bus called as PCIe. Every CPU supports certain PCIe bus and you add number of GPU based of how many PCIe lanes CPU supports and how much power is available on the machine to power the GPU and CPU.
Is Tensor flow named after tensor core?	No. TensorFlow is a framework and has nothing to do with TensorCore. All frameworks use TensorCore to accelerate the training time including PyTorach, MXNet , Matlab etc
could you give example of both use cases serial and task parallel workloads and Data parallel workloads..	An typical example of Data Parallel tasks is Graphics where you want to decided which pixel uses which type of RGB values. In general all AI algorightms are data parallel tasks which can be converted to matric multiplication by batching. Matrix multiplication can be done in parallel . A 101 on same can be looked at https://cse.buffalo.edu/faculty/miller/Courses/CSE633/Ortega-Fall-2012-CSE633.pdf .
What TPU? and what type of taskss are offloaded on TPUs?	TPU is a processor by Google and another type of cluster : https://en.wikipedia.org/wiki/Tensor_Processing_Unit
What is a CUDA core?	NVIDIA GPU cores are called as CUDA cores.
Just curious, how do these time taken values for GPU and tensor flow compare with pure (only) CPU?	It depends. We have see 10x 100x to 1000x speed ups also shown in the literature for different model trainings
What is the windows equivalent command for lscpu?	There are third party tools like https://www.gtopala.com/
what is the windows equivalent command for nvidia-smi	It is same nvidia-smi is part of driver. In case the driver installation is correct you should be able to use it
What hardware difference does CPU, GPU and VPU are to be noted?	In general GPU has more number of cores as compared to CPU. The number of threads that can be launched on GPU is in few thousand while on CPU you can launch few hundred. VPU are specialized unit meant for Vision processing task to be done in parallel
what is WMMA?	WMMA is specialized instruction used to make use of Tensor Cores. https://developer.nvidia.com/blog/programming-tensor-cores-cuda-9/
What is CUDA?	NVIDIA GPU ARchitecture is called as CUDA. https://en.wikipedia.org/wiki/CUDA
How Sustainable benefits is achieved by using GPU programming?	The literature show few 100x to 1000x speedup as compared to sequential code

Question	Answer
can we install some virtual machine in windows to execute these linux commands discussed?	You may. AI primrily is driven and growing in Linux. We plan to only cover Linux in these sessions
what is login nodes?	Login nodes are used to authenticate the users, subequently users can submit jobs on the cluster
HPC means High performance cluster	HPC is High Performance Computing. Cluster is one of the architectures for implementing HPC
These commands are not working in my shell on MacOS Is there something that I'm missing?	MacOs does not have GPU. The commands being shown are when you have a NVIDIA GPU and drivers are installed on the system
can we have hands on google colab , will be helpful.	There will be sessions where we will show usage of google collab. Advance topics will be not be shown on collab as the GPU present in collab does not have features like TensorCore in free version . So wherever possible it will be shown
Can tensor core be implemented in Java?	No . It is C++ only. Java is a higher language and can call C++ functions if needed
nvidia-smi in NSM system will give GPU information about all the node available, right?	No. nvidia-smi is for a single node. For all nodes on NSM system you need to use cluster tools . You may see the number of GPU nodes by using SLURM commands like sinfo
Is core equivalent to thread/block/grid in CUDA?	Core is a hardware unit. While thread/block/grid are software componets. A thread runs on a core
Can you please explain what Caches are?	Cache is small sized but fast memory that goes between main memory and the CPU for holding frequently used data and code This helps in speeding up CPU operation, which otherwise gets slowed due to main memory which has higher access time
How is TFLOPs calculated for a Supercomputer? Is it the sum of floating operations done by each core of CPU and GPU?	That is correct. We can add the therotical peak value provided for CPU and GPU and that will overall theorotical peak of the cluster
Was the Cuda code structure discussed in the previous session?	It is a sample code to show how parallel codes are written on GPU. AI programmer do not have to write parallel code. This is just for demonstration. AI aI libraries already have these parallel code
what is warp?	It is equivalent to a thread on GPU.
can we choose desired specific nodes for a specific job?	Yes that is possible, provided the sys admin sets it up for you
is tensor core and cuda core different by hardware architecture itself? so, every GPU will have x cuda cores and y tensor cores?	Answered already with repect to previous queries. Please check the previous answers
In edge accelerators where the RAM is shared, is there any software or hardware differentiation of what memory the CPU can access v/s what the GPU can access? And is a copy necessary for moving data from CPU memory to GPU memory?	AS you stated RAM is shared but the caches are still sperate in embedded devices like Jetson. So data does move through different cache based on if it is used in GPU or CPU

Question	Answer
In the nvidia-smi command output, how is GPU utilization measured? Is it the number of CUDA cores being used?	Yes kind of. It shows on actually called as SM (Streaming Multiprocessor) which is a collection of x CUDA cores.
In the NX GPU diagram, could there be interference if CPU and GPU try to access the same RAM? Is memory access serial?	The hardware block called arbiter ensures that there is no clash between the two masters for shared memory. One master is allowed first then the second
On edge accelerators, is there anything done to maintain cache coherence between the GPU and	Coherency is ensured for cache wherever it is used. That is part of the hardware design
I would like to understand the low level details (kernel launch, scheduling, context switch etc) of how an AI workload (ex. minibatch) runs on the GPU. Could you point me to some relevant resources?	We will be covering some of these in future lectures and put some resources for you in the slack
Is it fair to just compare the execution time of the programs with and without tensor cores? Shouldn't the accuracy of the final outputs also be compared to illustrate the tradeoff?	Both follow IEEE754 standard. So accuracy ideally should be the same.
nvidia-smi is not supported on edge devices such as the jetson. tegrastats is the equivalent there right?	It is supported. Any device with NVIDIA GPU and driver installed will have nvidia-smi
How do we choose which device whether CPU, GPU, VPU, FGPA is to be used for the data we need to train	Every frameowork like PyTorch or TEnsorflow provide API to uses different hardware. We will cover these topics specifically for PyTorch in upcoming sessions
What is epoch?	https://deepai.org/machine-learning-glossary-and-terms/epoch
What is a CUDA core?	NVIDIA GPU cores are called as CUDA cores
Do the CUDA related terms and techniques work on openACC as well? And why and how exactly are those 2 different in relation to this course?	OpenACC is a framework which is at a higher abstraction level than CUDA. This course will not cover OpenACC
Can you please explain more what do you mean by Serial and task parallel against data parallel?	To better understand different types of computation you can refer to Flynn's taxonomy: https://en.wikipedia.org/wiki/Flynn%27s_taxonomy
In this session, The term "GPU" is used to refer to these massive processing units. Is this a standard usage? GPU stands for graphics processing units, while the end usage of these massive units may not necessarily be graphics centered?	GPUs were originally designed for graphics processing. However they are used with some enhancements for High performance computing and AI workloads
sir i am physics student. And AI fascinates me a lot that's why i am here doing this course, so my question is can i anyway switch my field of research in the future?	You can learn AI techniques and apply it to your field. AI as a technique can be applied to a variety of fields of science and technology,AI is applicable to all fiels including Physics. You may look at NVIDIA Modulus to understnad the concept of Physics Informed Neural Network