DSCC 201/401 Homework Assignment #2
Due: **September 20, 2021 at 9 a.m. EDT**

Answers to these questions should be submitted via Blackboard. Questions 1-5 should be answered by all students (DSCC 201 and 401) and Question 6 should be answered by students registered in DSCC 401. Please upload a file containing your answers and explanations to Blackboard (Homework #2: Hardware for Data Science) as a Word document (docx), text file (txt), or PDF.

Imagine that you currently lead the data science department at a major pharmaceutical company. You are in the process of configuring a large Linux cluster that will provide continuous computing resources for the company to discover new molecules and analyze biomolecular systems for the development of new drug candidates. Based on discussions with other data scientists and domain experts in the company, you have recommend purchasing a cluster containing two different specifications of compute nodes. A vendor has provided you with a quote for a system with a good starting configuration. The quote (HPC Quote.pdf) has been uploaded to the Blackboard site and is available under the instructions for this homework assignment. **Please provide thorough and thoughtful explanations to the following questions.**

Question 1: How many teraFLOPS of theoretical computing capacity will be provided by a Linux cluster with the compute nodes provided in the quote? Provide a double-precision floating point (FP64) value. Please show your reasoning and how you derive your values. (Hint: The Intel Xeon Gold 6330 CPUs use the Ice Lake microarchitecture.) Does the amount of RAM in the two different types of compute nodes affect the theoretical value? Why or why not?

Question 2: After further review, your team decides that it may be worthwhile to include a GPU card in each compute node. What would be the total computing capacity of the cluster (in teraFLOPS) if one Nvidia Ampere A100 GPU card was added to each server in the cluster? Provide a value based on double-precision floating point calculations (FP64).

Question 3: If your team purchases these compute nodes, will this equipment create a complete Linux cluster solution? Does it have everything needed to connect and work together? Why or why not? Is there any additional hardware needed to make these servers into a complete Linux cluster? Is there any additional software that may be needed to make the Linux cluster complete? Please be detailed and specific in your answer.

Question 4: In lecture, we discussed the architecture of the Summit supercomputer. This system has 4,680 nodes each with 6 Nvidia Volta V100 GPUs. According to IBM, the total computing capacity provided by the GPUs of one node is about 42 teraFLOPS (FP64). If IBM replaced the Volta V100 GPUs with the newly announced Ampere A100 GPUs, what would be the total estimated theoretical performance in petaFLOPS (FP64) of the Summit supercomputer? Would this system exceed the processing capabilities of the current top supercomputer in the world? Show how you derived your answer.

Question 5: According to the data sheet provided by Nvidia for the DGX Station A100, the desktop tower has an advertised performance of "2.5 petaFLOPS AI." A copy of this data sheet has been uploaded to the Blackboard site and is available under the instructions for this homework assignment. What is the precision of the computations used to measure this performance? Show how you derived your answer. Hint: Read the data sheet for the A100 GPU available here:
https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet.pdf

Question 6: (DSCC 401 ONLY): Read the paper, "Benchmarking TPU, GPU, and CPU Platforms for Deep Learning," by Wang, Wei, and Brooks. A copy of this paper has been uploaded to the Blackboard site and is available under the instructions for this homework assignment (Wang_Wei_Brooks.pdf). Provide detailed answers to the following questions:

A. What is ParaDnn? How does it compare to LINPACK?

B. The authors of the paper had special access to Google's new TPU v3. What is the performance of this accelerator as mentioned in the paper? How does the theoretical performance of Google's TPU v3 compare with Nvidia's theoretical performance of the Ampere A100 GPU?

C. Did the authors measure the performance on multi-GPU systems that use PCIe or NVLink in this study? If so, how many GPUs did they use concurrently? If not, what was the explanation?