**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 recommended 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?**

Performance = (# servers) \* (# CPUs) \* (# cores) \* (clock speed) \* (instructions per cycle)

PowerEdge C6520 - 56c, 1TB

Processor: Intel® Xeon® Gold 6330 Processor - <https://ark.intel.com/content/www/us/en/ark/products/212458/intel-xeon-gold-6330-processor-42m-cache-2-00-ghz.html>

# cores = 28

Clock speed = 2.00 GHz

Instructions per cycle = 32 (Icelake Architecture)

# servers = 36 + 64 = 100

# CPUs = 2 per node

Performance - 100 x 2 x 28 x 2 x 32 = 358400 GFLOPS = 358.4 TFLOPS

No, the theoretical performance remains unaffected by the difference in RAM since RAM is not one of the determining factors of the theoretical performance of a supercomputer. However, RAM does affect the sustained performance of a supercomputer. The gap between the theoretical and sustained performance of a supercomputer may often be due to memory limitations in certain use cases.

**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).**

Ampere A100 GPU Performance = 19.5 TeraFLOPS (DP)

Performance after including one Nvidia Ampere A100 GPU card to each server in the cluster:

358.4 + (100 x 19.5) = 2308.4 TeraFLOPS (DP)

**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.**

No, a complete Linux cluster solution cannot be created only with the compute nodes. The following hardware and software would be required in order to complete the Linux cluster.

1. Software required:

* Operating system - The cluster would be used by multiple users. An operating system is essential for managing multiple users in a shared environment.
* Management and monitoring software - Software such as XCat can help manage the cluster remotely. Other software such as Ganglia and Zabbix is used for monitoring resource usage.
* Job scheduling and launching software - Software such as Torque/Maui and SLURM is used for starting, stopping and monitoring compute jobs and resources. It can also allocate and deallocate computing resources on nodes and partitions. It also manages queues of jobs waiting for resources.

1. Hardware required:

* Compute - Racks
* Login node, Service node
* Storage - Hard drives for each compute node

1. Networking Components required:

Any of the following cables can be used to connect the nodes.

* Ethernet - high latency due to TCP/IP protocol.
* Infiniband - low latency, very high throughput and high bandwidth. The throughput for different kinds of InfiniBands is as follows; FDR10 (10 Gb/s), FDR (14 Gb/s), EDR (25 Gb/s), and HDR (50 Gb/s)

**Question 4:**

**In the 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.**

Total nodes - 4680

GPUs -> Nvidia Volta V100 -> 6 per node

Computing capacity of each node = 42 TeraFLOPS (FP 64)

Computing capacity of each GPU = 42/6 = 7 TeraFLOPS

Replaced with:

Ampere A100 GPUs

# Cores = 6912

GPU RAM = 40 GB

Computing capacity of each GPU = 19.5 TeraFLOPS

Computing capacity of each node = 19.5 x 6 = 117 TeraFLOPS

Computing capacity of the supercomputer = 117 x 4680 = 547560 TeraFLOPS = 547.56 PetaFLOPS

Yes, this system does exceed the processing capabilities of the Summit supercomputer.

**Question 5:**

**According to the datasheet 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 datasheet for the A100 GPU available here:**

[**https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet.pdf**](https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet.pdf)

Double-precision performance:

Double-Precision Performance (FP64) = 9.7 TFLOPS

Double-Precision (FP64) Tensor Core performance = 19.5 TFLOPS

Single-Precision Performance:

Single-Precision Performance (FP32) = 19.5 TFLOPS

Single-Precision Tensor Float 32 (TF32) performance = 156 TFLOPS |

Single-Precision Tensor Float 32 (TF32) performance with structural sparsity = 312 TFLOPS

Half-Precision Performance:

Half-precision performance = 312 TFLOPS

Half-precision performance with structural sparsity = 624 TFLOPS

Double-precision, single-precision and half-precision computations are used to measure performance. Structural sparsity is also enabled to measure single-precision and half-precision performance.

**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?**

ParaDnn is the first parameterized deep learning benchmark suite. ParaDnn is designed to generate thousands of parameterized multi-layer models, comprising fully connected models, convolutional neural networks, and recurrent neural networks. ParaDnn also allows benchmarking across almost six orders-of-magnitude of model parameter size, exceeding the range of existing benchmarks. ParaDnn represents a more complete view of potential workloads, and each workload represents the concerns of different users.

LINPACK is a benchmark for measuring performance based on the LU decomposition of a large matrix in which a matrix is factored based on the product of upper and lower triangular matrices. The LINPACK Benchmarks are a measure of a system's floating-point computing power. Introduced by Jack Dongarra, they measure how fast a computer solves a dense n by n system of linear equations Ax = b, which is a common task in engineering.

In conclusion, ParaDnn is a better-suited benchmark suite for deep learning whereas LINPACK is a benchmark for measuring the computing power of systems for linear equations.

**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?**

TPU v3 as mentioned in the paper has a peak performance of 420 TFLOPS.

NVIDIA Ampere A100 has 312 TFLOPS peak performance but with sparsity it can have a peak performance of up to 624 TeraFLOPS.

**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?**

The authors did not measure performance on a multi-GPU system since it was beyond the scope of their work. They evaluate the performance of a single V100 package instead of a DGX-1GPU platform that contains 8 V100 packages. In the DGX-1 GPU platform, the GPUs are connected via a 300 GB/s NVlink 2.0 interconnect. However, their intention was to study the minimal units available, hence, a single V100 package was studied since the performance of a multi-GPU system largely depends upon the user’s implementation according to the authors of the paper. They also mention that studying multi-node systems often involve more parameters such as numbers of nodes, inter-node bandwidth, inter-connect topology, and synchronization mechanisms. Moreover, according to the authors of the paper, cloud system overhead also becomes more acute in multi-node systems. These were the reasons put forth by the authors for not measuring performance on muti-GPU systems.

**References**:

1. Lecture Slides: DSCC 201/401 Tools and Infrastructure for Data Science (September 1, 2021)
2. Lecture Slides: DSCC 201/401 Tools and Infrastructure for Data Science (September 8, 2021)
3. Benchmarking TPU, GPU, and CPU Platforms for Deep Learning - Yu (Emma) Wang, Gu-Yeon Wei and David Brooks
4. Wikipedia: LINPACK benchmarks - Original author(s) Jack Dongarra, Jim Bunch, Cleve Moler, and Gilbert Stewart