Dscc 401
Homework- 02

Afnan Mostafa

Q1

Theoretical computing capacity (R_{peak}) for 6 servers (compute nodes) where each servers contains 2 CPUs + 1 Nvidia GPU.

Performance = (No. of cores) x (clock speed) x (instructions per cycle)

= 32 x 2 GHz x 32

= 2048 GFLOPS

Cores = 32
Clock speed = 2 GHz
DP = 32
for Intel Xeon Gold 6338 architecture (Ice Lake)

For a single servers,

R_{peak} = 2048 GFLOPS x 2

= 4096 GFLOPS

For 6 servers (Linux cluster),

R_{peak} = 4096 x 6 GFLOPS

= 24,576 GFLOPS

$\therefore R_{peak}$ = 24.576 TFLOPS

Total performance considering 1 GPU per servers:

$R_{peak}|_{total}$ = 24.576 + (6 x 1.2) TFLOPS

= 31.776 TFLOPS Ans

If GPUs are doubled in each servers,

$R_{peak}|_{total}$ = 24.576 + (12 x 1.2) = 38.976 TFLOPS

$$\% \text{ change} = \frac{R'_{peak}|_{total} - R_{peak}|_{total}}{R_{peak}|_{total}}$$

Or, difference $= 38.976 - 31.776$ TF

$= 7.2$ TF

$$= \frac{38.976 - 31.776}{31.776} \times 100\ \%$$

∴ % change in $R_{peak} = 22.66\ \%$ Ans

Q2 Nvidia A100 GPU

Each Nvidia A100 GPU adds 9.7 TFLOPS to the actual Rpeak.

$R_{peak}|_{CPU} = 24.576$ TF

$R_{peak}|_{GPU} = 6 \times 9.7 = 58.2$ TF

$\therefore R_{peak}|_{total} = R_{peak}|_{CPU} + R_{peak}|_{GPU}$

$= 24.576 + 58.2$

∴ $R_{peak}|_{tot} = 82.776$ TF Ans

Q3.

Parallel File System: Parallel file system provides concurrent and simultaneous high speed file access to applications executing on multiple nodes of cluster. It is a special type of clustered file system.

In this case, such parallel file systems allow file access among the 6 servers. Each servers has a single hard drive that is used by the applications running in that server. But what if a simulation ~~is~~ running in server 3 needs to access the memory/file stored in servers 5? Parallel file system ensures such communication and provides efficient and parallel file access throughout multiple servers.

As we can see from the calculation of the theoretical capacity that memory information is not needed, it can be said that doubling the memory will not directly impact the Rpeak value. Rpeak will remain the same. However, more memory ensures uninterrupted service even more, which is crucial for tackling computational bottlenecks. If the memory consumption by a large task is very large and no memory is available, then the performance of the cluster will slow down but Rpeak stays the same, irrespective of add-on memory.

Q 4.

A few components are missing in the quote that need to be integrated to make a complete Linux cluster.

Hardware:

1) **Computing:** Motherboard, Intel Xeon Gold, additional processors, GPU (present)
   → No additional component is needed.

2) **Storage:** 960 GB SSD, RDIMM (RAM) → Present
   → ~~Although~~ the specification of RDIMM is not fully present.
   → No HDD but SSD is present.
   → No RAID (Parallel file system), needs to be integrated.

3) **Networking:** Infiniband is present
   → Nothing is missing.

Software:

→ Operating system is missing.
→ Need to purchase 'Management and monitoring', 'Job scheduling and launching', and 'User software' tools.

So, the quote is missing a few integral components that are required to make a complete Linux cluster.

Missing items:

1) Specification of RDIMM
2) Parallel file system (RAID)
3) Operating system
4) Others softwares

Q5

Leonardo Supercomputer:

Total computing nodes = 3456

1 Node consists of 1 Intel Xeon CPU (32 cores) and 4 Nvidia A100 GPUs

From the Top500 list,

R_{peak} = 304.47 PFLOPS

Performance of a single A100 GPU = 9.7 TFLOPS

∴ Performance of (3456 × 4) A100 GPUs = 9.7 × 3456 × 4

= 134.09 PFLOPS

∴ R_{peak} = CPU + GPU

⇒ 304.47 = CPU + 134.09

∴ CPU = 170.38 PFLOPS

Assumption: Number of GPUs stay the same after replacement

Now, Replacing A100 by H100 ⇒

Performance of a single H100 GPU = 24 TFLOPS

∴ Performance of (3456 × 4) H100 GPUs = 24 × 3456 × 4

= 331.78 PFLOPS

∴ New R_{peak} with H100, R'_{peak} = CPU + GPU_{H100}

= 170.38 + 331.78 PFLOPS

R'_{peak} = 502.16 PFLOPS Ans

If we compare R_{peak} & R'_{peak} in terms of per-node basis, then

$R_{peak/node} = \frac{304.47}{3456} = 88.1$ TFLOPS

$R'_{peak/node} = \frac{502.16}{3456} = 145.3$ TFLOPS

So, Leonardo with H100 would not be the top system because the top system (Frontiers) has a Rpeak of 1.7 ExaFLOPS. However, Leonardo will surpass LUMI and be very comparable to Fugaku in terms of theoretical computing powers.

Ans

Q6.

Nvidia Eos, which is expected to have a performance of 18.4 ExaFLOPS (EF), will, however, not surpass the top system (Frontiers) of the world because Eos' rating of 18.4 EF is based on FP8 precision. It becomes 275 PF when scientific computing (FP64) is considered.

Frontiers:

Rpeak for scientific computing (FP64) = 1.7 EF

Rpeak for AI (FP8 or FP16) ≈ 6.88 EF

Eos:

Rpeak for scientific computing (FP64) = 275 PF

Rpeak for AI (FP8) = 18.4 EF

So, Eos will be a better candidate for mixed (or half) precision computing but not for FP64; Frontiers will still be on top of the list.

[Q7].

A ParaDnn: A tool that can generate parameterized deep neural network models (Fully connected, convolutional, and recurrent) for benchmarking purposes.

LINPACK: A software library that can solve a system of linear equations with efficiency (based on floating point calculations).

ParaDnn differs from LINPACK in the way it is developed to do benchmarking; the formers is developed for neural network applications and LINPACK is developed to estimate how fast an architecture can solve a system of linear equations. Then, ParaDnn can work with FP16 but LINPACK needs much more precision in terms of accuracy. Furthermore, ParaDnn provides more information of the deep learning ecosystem, while LINPACK is used to rank high-end supercomputers for the TOP500 list.

B The performance of the TPU v3 is claimed to be 420 TFLOPS, with each core is capable of having a performance of 90TFLOPS. TPU v3 is specialized for deep learning ecosystem and can't

do the rigorous mathematical operations. On the other hand, Nvidia's Ampere A100 GPU, which is a versatile graphics card, can deliver up to 9.7 TF in FP64 (Double Precision). Nevertheless, such A100 cards can have a theoretical performance of 19.5 TF when it is working with Tensorflows. In addition, A100 has a Rpeak value of 19.5 TF in FP32 (single precision). Most importantly, A100 cards have a staggering Rpeak value of 312 TF, and it becomes 624 TF for sparse matrices (structural sparsity). This is better than Google's TPU V3 (420 TF in FP16). However for non-sparse calculations, TPU V3 has a better performance than A100. Overall, despite having higher Rpeak (not considering structural sparsity), TPU V3 is not versatile in terms of being efficiently fast in a broad range of floating point precision.

c. No, the authors did not measure the performance on multi-GPU systems that use PCIe or NVLink because studying multi-node systems requires more system parameters, including number of nodes, inter-node bandwidth, inter-connect topology, and synchronization mechanisms. Cloud system overhead also becomes more acute.