NAIRR Pilot Compute Calculations

We made the following calculations to estimate the compute available through the NAIRR pilot, which encompasses six Federal HPC resources. This includes calculations to (a.) estimate the FLOPS of CPUs, (b.) convert GPU FLOPS to FP16, (c.) aggregate GPU, CPU, and ASIC FLOPS across nodes, and (d.) gauge overall FLOPs allocated by multiplying the FLOPS of a resource by the amount of time (GPU or node hours converted to seconds) it is allocated to AI projects, then dividing that number by the amount of nodes or GPUs in the resource. Information on the hardware for each NAIRR pilot resource was sourced primarily from the NAIRR website,[1] information used to convert all FLOPS to FP16 was sourced primarily from TechPowerUp and the developers' specs sheets,[2] information on CPU specifications was sourced from the developers' websites (i.e., Intel and AMD's websites), and information on the allocated GPU and node hours was sourced from the NAIRR's current projects page.[3] Note that FP16 FLOPS for some resources were directly provided by sources, and in these cases we did not perform calculations separately.

Delta Resource[4]

GPUs:

1. NVIDIA A100 40GB HBM2e[5] [tensor FP16 without sparsity] = 312TF
    a. [400 NVIDIA A100 40GB HBM2] 400 * 312TF = 124800TF
2. NVIDIA A40 48GB GDDR6 PCIe 4.0[6] = 37.42TF
    a. [400 NVIDIA A40 48GB GDDR6 PCIe 4.0] 400 * 37.42 = 14968TF
3. NVIDIA A100 GPUs[7] [tensor FP16 without sparsity = 312TF
    a. [40 NVIDIA A100 GPUs (assuming 40GB HBM2)] 40 * 312TF= 12480TF
4. AMD MI100[8] = 184.6TF

---

[1] https://nairrpilot.org/allocations
[2] https://www.techpowerup.com/gpu-specs/
[3] https://submit-nairr.xras.org/current-projects
[4] https://www.ncsa.illinois.edu/research/project-highlights/delta/
[5] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf
[6] https://www.techpowerup.com/gpu-specs/a40-pcie.c3700#:~:text=NVIDIA%20has%20paired%2048%20GB,MHz%20(14.5%20Gbps%20effective)
[7] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf
[8] https://www.techpowerup.com/gpu-specs/radeon-instinct-mi100.c3496

    a. [8 AMD MI100] 184.6 * 8 = 1476.8TF
5. Total GPU FLOPS: 124800TF + 14968TF + 12480TF + 1476.8TF = 153724.8TF = 153.7248PF
6. Total GPU FLOPS without A40s: 138.76PF

CPUs: 200 AMD EPYC 7763 ("Milan") CPUs[9] with 64-cores/socket (64-cores/node) at 2.55GHz and 256GB of DDR4-3200 RAM

7. AVX 256 [bits] / 16 [FP16] * 2 [multiple and add; 2 operations] * 2 [two AVX 256 blocks per core] * 64 [64 cores per CPU] * 2.55 Ghz = 10444.8 GF/node
    a. 10444.8 GF/node * 200 [total CPUs] = 2088960GF = 2.08896PF

Total FLOPS:

8. 153.7248PF [from GPUs] + 2.08896PF [from CPUs] = 155.81376PF

FLOPs Allocated [(FLOPS * Seconds) / GPUs]:[10]

9. Time measured in GPU hours
10. 290,142 GPU hours allocated = 1,044,511,200 seconds
    a. "Allocations Description: 1 GPU Hour = 1 A100 GPU-Hour on a quad-NVIDIA A100 node (using 1 of the 4 GPUs) in the normal queue. A40 GPUs are discounted while large memory GPU nodes have a premium charge. Details on the charge rates are available in the Delta user guide (refer to the partitions section under running jobs)."
    b. Delta FLOPS without A40 GPUs = 1.39E+17
11. GPU count (without A40 GPUs) = 448
12. **Total FLOPs:**
    a. (1.39E+17 * 1,044,511,200) / 448 = 3.24E+23 = 323518692.21PF

Frontera Resource[11]

GPUs:

1. The 90 nodes with NVIDIA Quadro RTX 5000 GPUs can do 4PF in FP32 in total.[12] Based on conversions using the FP32/FP16 numbers here,[13] this would be 8.028PF.

---

[9] https://www.amd.com/en/products/cpu/amd-epyc-7763
[10] https://submit-nairr.xras.org/resources
[11] https://tacc.utexas.edu/systems/frontera/
[12] https://tacc.utexas.edu/systems/frontera/
[13] https://www.techpowerup.com/gpu-specs/quadro-rtx-5000.c3308

a. Note that we used non-tensor FP16 FLOPS because the Quadro RTX 5000 specs sheet does not indicate the precision of the tensor performance.[14]

CPUs:

2. Intel® Xeon® Platinum 8280 Processors[15]
    a. AVX 512 [bits] / 16 [FP16] * 2 [multiple and add; 2 operations] * 2 [2 AVX 512 blocks per core] * 28 [28 cores per CPU] * 2.7 Ghz = 9676.8GF
        i. 9676.8GF * 2 [2 CPUs per node][16] = 19353.6 GF/node
        ii. 19353.6 GF/node * 8368 [total nodes] = 161950924.8GF = 161.95PF
3. Intel Xeon CPU E5-2620 v4 Processors[17]
    a. 256 [bits] / 16 [FP16] * 2 [multiple and add; 2 operations] * 2 [blocks per core] * 8 [8 cores per CPU] * 2.1 GHz = 1075.2GF/CPU
    b. 1075.2GF/CPU * 360 [360 CPUs] = 387072GF = 387.072TF = 0.387PF
4. CPUs total
    a. 161.95PF + 0.387PF = 162.337PF

Total FLOPS:

5. 8.028PF [from GPUs] + 162.337PF [from CPUs] = 170.365PF

FLOPs Allocated [(FLOPS * Seconds) / Nodes]:[18]

6. Time measured in node hours
7. Frontera resource is divided into Frontera GPU and Frontera CPU. We calculate allocations separately
    a. Frontera GPU
        i. 117,325 node hours allocated = 422,370,000 seconds
        ii. Frontera GPU FLOPS = 8.00E+15
        iii. Node count = 90
        iv. **Total FLOPs:**

---

[14] https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/quadro-rtx-5000-data-sheet-us-nvidia-704120-r4-web.pdf

[15] https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html

[16] https://tacc.utexas.edu/systems/frontera/

[17] https://www.intel.com/content/www/us/en/products/sku/92986/intel-xeon-processor-e52620-v4-20m-cache-2-10-ghz/specifications.html

[18] https://submit-nairr.xras.org/resources

1. (8.00E+15 * 422,370,000) / 90 = 3.75E+22 = 37500000PF

b. Frontera CPU
   i. 1,157,500 node hours allocated = 4,167,000,000 seconds
   ii. Frontera CPU FLOPS = 1.62E+17
   iii. Node count = 8392
   iv. **Total FLOPs:**
      1. (1.62E+17 * 4,167,000,000) / 8392 = 8.04E+22 = 80,400,000PF

Lonestar6 Resource[19]

CPU: AMD "Milan" EPYC 7763 processors

1. AVX 256 [256 bits] / 16 [FP16] * 2 [multiple and add; 2 operations] * 2 [2 AVX 256 blocks per core] * 64 [64 cores per CPU] * 2.55 Ghz = 10444.8GF
   a. 10444.8GF * 2 [2 CPUs per node] = 20889.6 GF/node
   b. 20889.6 GF/node * 530 [total nodes] = 11071488GF = 11.071488PF

GPU:

2. 255 A100s * 312TF [312TF per A100 without sparsity][20] = 79560TF = 79.56PF
3. 8 H100s * 756TF [756TF per H100 without sparsity][21] = 6048TF = 6.048PF
4. Total: 79.56PF + 6.048PF = 85.608PF = 8.56E+16

Total FLOPS:

5. 11.071488PF [from CPUs] + 85.608PF [from GPUs] = 96.68PF

FLOPs Allocated [(FLOPS * Seconds) / Nodes]:[22]

6. Time measured in node hours, but for GPUs only.
7. 63,220 node hours allocated = 227,592,000 seconds
8. Lonestar6 FLOPS [GPU only] = 8.56E+16
9. Node count = 89
10. Total FLOPs:
    a. (8.56E+16 * 227,592,000) / 89 = 2.19E+23

---

[19] https://docs.tacc.utexas.edu/hpc/lonestar6/#intro

[20] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf

[21] https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

[22] https://submit-nairr.xras.org/resources

Neocortex Resource[23]

ASICs: Two Cerebras CS-2s[24]

1. 5780TF [per CS-2] * 2 [2 CS-2s] = 11.560PF[25]

CPUs: 32 Intel Xeon Platinum 8280L CPUs[26]

2. AVX 512 [bits] / 16 [FP16] * 2 [multiple and add; 2 operations] * 2 [2 AVX 512 blocks per core] * 28 [28 cores per CPU] * 2.7 Ghz = 9676.8GF
3. 9676.8GF * 32 [32 total CPUs] = 309657.6GF = 0.30966PF

Total FLOPS:

4. 11.560PF [from two Cerebras CS-2s] + 0.30966PF [from CPUs] = 11.87PF

FLOPs Allocated [(FLOPS * Seconds) / Nodes]:[27]

5. Time measured in CS-2 hours
6. 63,220 CS-2 hours allocated = 1,440,000 seconds
7. Neocortex FLOPS = 1.19E+16
8. CS-2 count = 2
9. **Total FLOPs:**
    a. (1.19E+16 * 1,440,000) / 2 = 8.55E+21 = 8,550,000PF

Summit

Total FLOPS:[28]

1. 3.30E+18 = 3300PF

FLOPs Allocated [(FLOPS * Seconds) / Nodes]:[29]

2. Time measured in node hours
3. 962,750 node hours allocated = 3,465,900,000 seconds
4. Summit FLOPS = 3.30E+18
5. Node count = 4600
6. **Total FLOPs:**

---

[23] https://www.cmu.edu/psc/aibd/neocortex/
[24] https://f.hubspotusercontent30.net/hubfs/8968533/CS-2%20Data%20Sheet.pdf
[25] https://www.alcf.anl.gov/alcf-ai-testbed
[26] https://www.intel.com/content/www/us/en/products/sku/192472/intel-xeon-platinum-8280l-processor-38-5m-cache-2-70-ghz/specifications.html
[27] https://submit-nairr.xras.org/resources
[28] https://www.olcf.ornl.gov/wp-content/uploads/2019/05/Summit_System_Overview_20190520.pdf
[29] https://submit-nairr.xras.org/resources

a. (3.30E+18 * 3,465,900,000) / 4600 = 2.49E+24 = 2,486,406,521.74PF

ACLF Testbed

Total FLOPS:

1. 3.53E+16 = 35.32PF

FLOPs Allocated [(FLOPS * Seconds) / Nodes]:[30]

2. Time measured in node hours
3. 19,230 node hours = 69,228,000 seconds
    a. Node count = 23
    b. We cannot determine what node hours are allocated for which ACLF sub-system nodes.
4. ACLF Testbed FLOPS = 3.53E+16
5. **Total FLOPs:**
    a. (3.53E+16 * 69,228,000) / 23 = 1.06E+23 = 106298089PF

Total FLOPs Allocated Across NAIRR Resources

- 2.49E+24 [Summit] + 8.04E+22 [Frontera CPU] + 3.75E+22 [Frontera GPU] + 3.24E+23 [Delta] + 1.06E+23 [ACLF Testbed] + 2.19E+23 [Lonestar6] + 8.55E+21 [Neocortex] = 3.26E+24

H100 Equivalent to Total NAIRR Compute

1. Total NAIRR compute capacity = 3770.04PF = 3770040TF
    a. Note that some of the FP16 FLOPS numbers in the total NAIRR compute are not included in the above calculations because they are provided directly by sources. See Table 1 in the blog post for more details.
2. Total NAIRR compute capacity equivalent in H100s
    a. 3770040TF / 756TF [756TF per H100 in tensor FP16 without sparsity] = 4986.8 H100s

[30] https://submit-nairr.xras.org/resources