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, and (c.) aggregate GPU, CPU, and ASIC FLOPS across nodes. 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,[2] and information on CPU specifications was sourced from the developers' websites (i.e., Intel and AMD's websites). Note that FP16 FLOPS for some resources were directly provided by sources, and in these cases we did not perform calculations separately.

Delta Resource[3]

GPUs:

1. NVIDIA A100 40GB HBM2e[4] = 77.97TF
   a. [400 NVIDIA A100 40GB HBM2] 400 * 77.97 = 31188TF
2. NVIDIA A40 48GB GDDR6 PCIe 4.0[5] = 37.42 TFLOPS
   a. [400 NVIDIA A40 48GB GDDR6 PCIe 4.0] 400 * 37.42 = 14968TF
3. NVIDIA A100 GPUs = 9.746 TFLOPS
   a. [40 NVIDIA A100 GPUs (assuming 40GB HBM2)] 40 * 77.97 = 3118.8TF
4. AMD MI100[6] = 184.6TF
   a. [8 AMD MI100] 184.6 * 8 = 1476.8TF
5. Total GPU FLOPS: 31188TF + 14968TF + 3118.8TF + 1476.8TF = 50751.6TF = 50.752PF

CPUs: 200 AMD EPYC 7763 ("Milan") CPUs[7] with 64-cores/socket (64-cores/node) at 2.55GHz and 256GB of DDR4-3200 RAM

6. 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

[1] https://nairrpilot.org/allocations
[2] https://www.techpowerup.com/gpu-specs/
[3] https://www.ncsa.illinois.edu/research/project-highlights/delta/
[4] https://www.techpowerup.com/gpu-specs/a100-pcie-40-gb.c3623
[5] https://www.techpowerup.com/gpu-specs/a40-pcie.c3700#:~:text=NVIDIA%20has%20paired%2048%20GB,MHz%20(14.5%20Gbps%20effective)
[6] https://www.techpowerup.com/gpu-specs/radeon-instinct-mi100.c3496
[7] https://www.amd.com/en/products/cpu/amd-epyc-7763

    a. 10444.8 GF/node * 200 [total CPUs] = 2088960GF = 2.08896PF

Total FLOPS:

7. 50.752PF [from GPUs] + 2.08896PF [from CPUs] = 52.84096PF

Frontera Resource

GPUs:

1. The 90 nodes with NVIDIA Quadro RTX 5000 GPUs can do 4PF in FP32 in total.[8] Based on conversions using the FP32/FP16 numbers here,[9] this would be 8.028PF

CPUs:

2. Intel® Xeon® Platinum 8280 Processors[10]
    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][11] = 19353.6 GF/node
        ii. 19353.6 GF/node * 8368 [total nodes] = 161950924.8GF = 161.95PF
3. Intel Xeon CPU E5-2620 v4 Processors[12]
    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

Lonestar6 Resource[13]

CPU: AMD "Milan" EPYC 7763 processors

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

[9] https://www.techpowerup.com/gpu-specs/quadro-rtx-5000.c3308

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

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

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

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

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

Total FLOPS:

2. 11.071488PF [from CPUs] + 29.234PF [from GPUs] = 40.305488PF

Neocortex[14]

ASICs: Two Cerebras CS-2s[15]

1. 5780TF [per CS-2] * 2 [2 CS-2s] = 11.560PF[16]

CPUs: 32 Intel Xeon Platinum 8280L CPUs[17]

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

[14] https://www.cmu.edu/psc/aibd/neocortex/

[15] https://f.hubspotusercontent30.net/hubfs/8968533/CS-2%20Data%20Sheet.pdf

[16] https://www.alcf.anl.gov/alcf-ai-testbed

[17] https://www.intel.com/content/www/us/en/products/sku/192472/intel-xeon-platinum-8280l-processor-38-5m-cache-2-70-ghz/specifications.html