FASTER

Next Generation Composability

March 6, 2022

Texas A&M University
**Composable** software-hardware approach
- 184-Intel Ice Lake nodes (11,520-core) with InfiniBand. (64-core, 256GB memory, and 3.84TB NVMe disk per node)
- **NVIDIA GPUs:** 200x T4, 40x A100, 10x A10, 4x A30, and 8x A40 GPUs
- Each node can compose up to 20 GPUs.

This project is supported by NSF award #2019129
PyTorch ResNet50 on FASTER

![Graph showing throughput (img/s) vs number of GPUs for different GPU models. The x-axis represents the number of GPUs (1, 2, 4, 8, 10), and the y-axis represents throughput in images per second (img/s). The graph includes data for Grace A100, LS6 A100, HPG DGX A100, NGC DGX A100, Expanse V100, NGC DGX V100, Grace RTX 6000, Grace T4, FASTER A100, and Frontera DGX A100-40GB.]
TF ResNet50 on FASTER

This project is supported by NSF award #2019129
AI/ML Benchmarks on FASTER

This project is supported by NSF award #2019129
Exhaustive Software Modules

- Python
- Matlab
- Keras
- PyTorch
- scikit-learn
- Pandas
- NumPy
- Matplotlib
- ...

Compilers: C++, Fortran, Intel OneAPI, GNU, ...
CUDA, OpenCL
OpenMPI, IntelMPI
...

Sample Benchmarking Job Script

```bash
#!/bin/bash

#NECESSARY JOB SPECIFICATIONS
#SBATCH --job-name=NvidiaDeepLearningBenchmark
#SBATCH --time=01:00:00
#SBATCH --ntasks=24
#SBATCH --gres=gpu:a100:1
#SBATCH --partition=gpu
#SBATCH --reservation=benchmarking

#Request 24 tasks
#Request 180GB per node
#Send stdout/err to "ExampleOut.[jobID]
#Request 1 A100 GPU per node (can be 0)
#Request the GPU partition/queue
```
HPRC Portal - Interactive Apps

**BIO**
- Beauti
- DIYABC
- FigTree
- IGV
- JBrowse
- Krait
- Mauve
- Structure
- Tracer
- CRISPR-Local
- Gap5

**GUI**
- ANSYS Workbench
- Abaqus/CAE
- LS-PREPOST
- MATLAB
- ParaView
- VNC

**Servers**
- Jupyter Notebook
- JupyterLab
- RStudio
- Spark-Jupyter Notebook

**Imaging**
- AFNI
- Chimera
- Coot
- Diffusion Toolkit & TrackVis
- FSL
- Fiji
- ICY
- ImageJ
- Vaa3D
- cisTEM
# ACES - Accelerating Computing for Emerging Sciences
(To be deployed in 2022)

<table>
<thead>
<tr>
<th>Component</th>
<th>Quantity*</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphcore IPU</td>
<td>16</td>
<td>16 IPUs direct-attached to a server</td>
</tr>
<tr>
<td>Intel Agilex FPGA</td>
<td>20</td>
<td>Agilex FPGA with a broad hierarchy of memory including DDR5, HBM2e and Optane Persistent Memory</td>
</tr>
<tr>
<td>NextSilicon coprocessor</td>
<td>20</td>
<td>Reconfigurable accelerator with an optimizer continuously evaluating application behavior.</td>
</tr>
<tr>
<td>NEC Vector Engine</td>
<td>24</td>
<td>Vector computing card with 8 cores and HBM2 memory</td>
</tr>
<tr>
<td>Intel Ponte Vecchio GPU</td>
<td>100</td>
<td>Intel GPU for HPC, DL Training, AI Inference</td>
</tr>
<tr>
<td>Liquid Intel Optane PCIe SSDs</td>
<td>6</td>
<td>3 TB PCIe SSD cards addressable as memory using Intel Memory Drive Technology</td>
</tr>
</tbody>
</table>

*Estimated quantities
High Performance Research Computing
DIVISION OF RESEARCH

hprc.tamu.edu

HPRC Helpdesk:
help@hprc.tamu.edu
Phone: 979-845-0219