OLCF Training

AI on Frontier

Junqi Yin
Analytics and AI Methods at Scale

Feb 16, 2023
Outline

- Frontier DL Environment
- Preliminary performance numbers
  - Kernels: GEMM/CONV/LSTM
  - Models: CNN/RNN
  - Applications: ResNet50, STEMDL
- Simulation-ML integration
Deep Learning Stacks

- Most DL codes work on Frontier without changes
Frontier DL Environment

• What’s working?
  • TensorFlow and PyTorch
  • Third-party libraries
    • Cray DL-plugin
    • Horovod
    • DeepSpeed
    • PyG ...

• Peculiarities
  • rocm-smi
  • MIOpen cache
  • RCCL + libfabric

• Resources: [https://github.com/ROCmSoftwarePlatform](https://github.com/ROCmSoftwarePlatform)
Performance baselines: Kernels

- Kernel Ops (fp32)
  - GEMM ~ 1.7x
  - CONV ~ 1.6x
  - LSTM ~ 1.1x
Performance baselines: Models

- CNN (fp32)
  - AlexNet
  - GoogleNet
  - OverFeat
  - VGG
- RNN
Performance baselines: Apps

- **ResNet50**
  - Mixed
  - \( \sim 1.0x \) per GCD
  - 98% at 1024
Performance baselines: Apps

- **STEMDL**
  - Tiramisu network
  - 220M parameters
  - 97% at 8192 GPUs

Accelerating Collective Communication in Data Parallel Training across Deep Learning Frameworks. USENIX NSDI'22, 2022
Simulation-ML Integration

• Best of both worlds: FP64 simulation + FP16 modeling

• Common use cases
  – Surrogate modeling
  – Reduced model
  – Interleaved

Learning to Scale the Summit: AI for Science on a Leadership Supercomputer, IPDPSW 2022
Simulation-ML Integration

- Tightly coupled
  - Single executable
  - one-to-one
- Loosely coupled
  - Different machines
  - many-to-many
- Semi-tightly
  - Separate executables
  - 1-to-1, many-to-many
## Simulation-ML Integration

<table>
<thead>
<tr>
<th>Method</th>
<th>Pro</th>
<th>Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framework C++ API (TensorFlow/PyTorch C++)</td>
<td>• Portable</td>
<td>• Not flexible</td>
</tr>
<tr>
<td></td>
<td>• Better latency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Easy to deploy</td>
<td></td>
</tr>
<tr>
<td>Framework Server (TensorFlow Serving/TorchServe)</td>
<td>• Flexible</td>
<td>• High maintenance</td>
</tr>
<tr>
<td></td>
<td>• Better throughput</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Options to deploy</td>
<td></td>
</tr>
<tr>
<td>Third-party API (SmartRedis/RedisAI)</td>
<td>• Easy integration</td>
<td>• Portability</td>
</tr>
<tr>
<td></td>
<td>• More functionality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Model support</td>
<td></td>
</tr>
</tbody>
</table>
TensorFlow C++

- Assume model in TF SavedModel format
- Link with `libtensorflow_cc.so`
- Support `half`, `uint8`, ...
TensorFlow Serving

- Launch server:
  `tensorflow_model_server --port --model_config_file`

- Support grpc & http
Simulation-ML Integration

- 1.2x per GCD over V100

Strategies for Integrating Deep Learning Surrogate Models with HPC Simulation Applications, IPDPSW 2022
Questions?
yinj@ornl.gov