Deep Introspection for Deep Learning and Exploiting Offloading Capabilities of Bluefield Adapters: The MVAPICH2 Approach

Donglai Dai

d.dai@x-scalesolutions.com

Aug 26, 2020

X-ScaleSolutions

x-scalesolutions.com
Outline

• Introduction (About us)
• **X-ScaleAI**: High-Performance Solution for AI problems
• **X-ScaleHPC**: High-Performance MPI Solution for HPC problems
• Conclusion
About us

• Bring innovative and efficient end-to-end solutions, services, support, and training to our customers

• Leverage the full potential of your cluster for all your users and applications
  • Commercial support and training for the state-of-the-art communication libraries
    • High-Performance and Scalable MVAPICH2 Library and Its families (MVAPICH2-X, MVAPICH2-GDR, MVAPICH2-Azure, MVAPICH2-AWS, and OSU INAM)
    • High-Performance Big Data Libraries (RDMA-Hadoop, RDMA-Spark, RDMA-HBase, and RDMA-Memcached)
  • Commercial products and services
    • X-ScaleAI: High-Performance Solution for AI problems
    • X-ScaleHPC: High-Performance MPI Solution for HPC problems

• More details in x-scalesolutions.com
About us (cont.)

• A Silver ISV member of the OpenPOWER Consortium
• Winner of multiple DOE SBIR grants
• Provide commercial support for MVAPICH2, HiBD, and HiDL Libraries to US federal national labs and international supercomputer centers
• Have two integrated products with support for HPC cluster systems (first introduced at the 2019 OpenPOWER Summit, North America)
  • X-ScaleAI
  • X-ScaleHPC
Outline

• Introduction (About us)
• X-ScaleAI: High-Performance Solution for AI problems
• X-ScaleHPC: High-Performance MPI Solution for HPC problems
• Conclusion
X-ScaleAI Product and Features

- **Aim:** High-performance solution for distributed training for your complex AI problems on modern HPC platforms

- **Features:**
  - Powered by MVAPICH2 libraries
  - Great performance and scalability as delivered by MVAPICH2 libraries
  - Integrated packaging to run various Deep Learning Frameworks (TensorFlow, PyTorch, MXNet, and others)
  - Targeted for both CPU-based and GPU-based Deep Learning Training
  - Integrated profiling and introspection support for Deep Learning Applications across the stacks (DeepIntrospect)
    - Provides cross-stack performance analysis in a visual manner and help users to optimize their DL applications and harness higher performance and scalability
  - Out-of-the-box optimal performance
    - Tuned for various CPU- and GPU-based HPC systems
  - One-click deployment and execution
    - Do not need to struggle for many hours
  - Support for x86 and OpenPOWER platforms
  - Support for InfiniBand, RoCE and NVLink Interconnects
Enhancing ML/DL Workflow with DeepIntrospect (DI)

1. Import a Model
   - Keras Applications
   - Model Zoo

2. Use DL Framework
   - TensorFlow
   - PyTorch

3. Use Deep Introspect (DI)
   - Train
   - Profile
   - Monitor
   - Analyze

DeepIntrospect (DI): a performance analysis and optimization tool for distributed DL applications
Overview of DeepIntrospect SW Architecture

Deep Learning Applications, Frameworks, and Middleware
- Vision, Speech, Text, etc.
- TensorFlow, PyTorch
- Horovod

DeepIntrospect
- Performance Monitor
- Profiler
- Log Parser
- Visualizer

Connectors to Data Store

Data Store

Extreme-Scale HPC Platforms
- GPUs (NVIDIA)
- CPUs (POWER, x86)
- Interconnects (NVLink, PCIe, InfiniBand)
Installing the X-ScaleAI package (xscale-ai-install)

$ export CUDNN_LIB_PATH=<Path-To-CUDNN-LIB-Directory>
$ export CUDA_HOME=<Path-To-CUDA>
$./xscale-ai-install

X-ScaleAI: License Verification Successful!

Installing X-ScaleAI...
If you encounter errors, please report to contactus@x-scalesolutions.com

-- Installing Miniconda (Python) ...
-- Installing TensorFlow 2.2 ...
-- Installing MPI ...
-- Installing Horovod 0.19.1 with DeepIntrospect ...
-- Installing TensorFlow Benchmarks (tf_cnn_benchmarks) ...
X-ScaleAI Successfully Installed
Running the X-ScaleAI package (xscale-ai-run)

$ ./xscale-ai-run -np 2 -hostfile ./hosts ./install_xscale_ai_dir/miniconda/bin/python ./install_xscale_ai_dir/benchmarks/scripts/tf_cnn_benchmarks/tf_cnn_benchmarks/py -model=resnt50 -variable_update=horovod

X-ScaleAI: License Verification Successful!

Running X-ScaleAI with arguments...
If you encounter errors, please report to contactus@x-scalesolutions.com

... Output of Run Will Then Appear Below ...

All features of DeepIntrospect are available with X-ScaleAI v1.0 package (to be released soon)
X-ScaleAI Product with DeepIntrospect (DI) Capability

Welcome to DeepIntrospect

A High-Performance Profiler by X-ScaleSolutions

<table>
<thead>
<tr>
<th>Operation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allreduce</td>
<td>369</td>
</tr>
</tbody>
</table>

Profiling Information (command line output)

<table>
<thead>
<tr>
<th>Operation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>369</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

JSON Format

```json
0:
  operation: "allreduce"
  stats:
    0:
      count: 7
      msg_size: 4084
      time: 14135
    1:
      count: 1
      msg_size: 37632
      time: 501
```

Statistics for MPI_Allreduce (Gradients and Parameters)

```json
Message Size, Count, Time per call, Total Time
4004, 7, 2019, 14135
37632, 1, 581, 581
37880, 1, 626, 626
38144, 6, 1105, 6634
55528, 5, 816, 4082
...```
X-ScaleAI Product with DeepIntrospect (DI) Capability

More capabilities and features are coming ...
Use-Case Example: DeepLabv3+

- DeepLabv3+ application uses large models with many parameters
- As shown in MPI_Allreduce total time distribution, default Horovod fusion buffer size (64 MB) and cycle time (3.5 ms) lead to poor training throughput

![Graph showing time vs. message size with spikes at 64 MB and 17 MB.]
Use-Case Example: DeepLabv3+ (cont.)

- Increasing fusion buffer size to 128 MB and cycle time to 5 ms led to larger messages called by MPI_Allreduce.
- The new tuning parameters lead to much more even MPI_Allreduce total time distribution.
Use-Case Example: DeepLabv3+ (cont.)

The graph shows the performance of DeepLabv3+ in terms of Images/sec as a function of the number of GPUs. The x-axis represents the number of GPUs, ranging from 1 to 132. The y-axis represents the images per second, with a scale ranging from 0 to 10 in hundreds.

- **Before Horovod Tuning**: Represented by blue bars.
- **After Horovod Tuning**: Represented by orange bars.
- **Ideal Scaling**: Represented by gray bars.

The chart illustrates the performance improvement achieved after using Horovod tuning, with a notable enhancement in performance as the number of GPUs increases. Notably, the vertical arrow indicates an ideal scaling factor of 1.3x, suggesting a significant boost in efficiency.

This data highlights the scalability and performance gains when using advanced distributed computing technologies like Horovod for deep learning applications.
Outline

• Introduction (About us)
• X-ScaleAI: High-Performance Solution for AI problems
• X-ScaleHPC: High-Performance MPI Solution for HPC problems
• Conclusion
X-Scale HPC Package

• Scalable solutions of communication middleware based on OSU MVAPICH2 libraries

• “out-of-the-box” fine-tuned and optimal performance on various HPC systems including CPUs and GPUs

• MPI communication offloading capabilities to ARM based smart NICs (such as Mellanox Bluefield NICs)
Bluefield Smart NIC Architecture

16 ARMv8 Cortex-A72 Cores

- Three-level coherent cache hierarchy
- 128b ARM Neon SIMD unit per core

Connect X-5 Subsystem

- Dual Virtual Protocol Interconnect (VPI) ports
- Ethernet/Infiniband at 100Gbps per port
- RDMA & NVMe-oF support

Integrated PCIe Switch

- 32 bifurcated PCIe 4.0 lanes (2x16/4x8/8x4/16x2)
- Speeds up to 200Gbps

Memory Controllers

- Supports two channels of 256 GB DDR4 DRAM at 1333MHz
Offloading MPI Operations to Smart NICs

• Exploits modern Programmable Network Adapters such as Mellanox Bluefield InfiniBand Adapters
• Optimized MPI Libraries exploiting Overlap
• Provides solutions to offload
  • Point-to-point
  • Collectives

Non-Blocking P2P/Collective

Bluefield HCA

Communication Process/Thread

MPI_Wait/ MPI_Waitall

Green: Communication
Orange: Computation
Dashed: Control Messages
• MVAPICH2 library is enhanced to offload MPI functionalities to ARM cores of the Bluefield Adapter
• Initial focus is on non-blocking collectives
• Performance is dependent on varying Workers (arm cores) Per Node (WPN)
Performance: OMB Nonblocking Alltoall

4 Nodes 16 PPN (64 processes total)

Total time (us)

- MV2
- MV2-OB-1WPN
- MV2-OB-2WPN
- MV2-OB-4WPN

36%
Performance: OMB Nonblocking Alltoall (cont.)

8 Nodes 20 PPN (160 processes total)

Total time (us)

MV2
MV2-OB-1WPN
MV2-OB-2WPN
MV2-OB-4WPN

16K 32K 64K 128K 256K 512K

Total time (us)

49%
P3D FFT Application

- Widely used parallel three dimensional fast Fourier transforms (FFT) library
- Overcomes scalability bottleneck by using two-dimensional domain decomposition
- Portable across many different system platforms
- Communication among the parallel processes is dominated by frequent MPI_Alltoall calls
Performance Improvement: P3DFFT

Input data size parameters are $x$, $y$, $x$

changing $x$: 320, 640, 1280
$y$: 960, $z$: 960
Outline

• Introduction (About us)
• X-ScaleAI: High-Performance Solution for AI problems
• X-ScaleHPC: High-Performance MPI Solution for HPC problems
• Conclusion
Conclusion

• Exponential growth in HPC and Deep Learning
• Requires advanced designs in middleware and tools to harness performance and scalability
  • X-ScaleAI
    • High-performance solution for distributed training for your complex AI problems
    • DeepIntrospect tool for exploiting HPC technologies for Deep Learning on x86 and OpenPOWER platforms
  • X-ScaleHPC
    • Optimized MPI library on various HPC systems including CPUs and GPUs
    • Capability to offload MPI functionalities to Mellanox Bluefield adapter and harness performance and scalability for MPI applications
  • Contact us for a demo and free-trial! (contactus@x-scalesolutions.com)
Thank You!

Donglai Dai

d.dai@x-scalesolutions.com
contactus@x-scalesolutions.com

x-scalesolutions.com/