Accelerating HPC and DL Applications Using DPUs and Efficient Checkpointing

Donglai Dai

Aug 25, 2021

http://x-scalesolutions.com
Outline

• Overview of X-ScaleSolutions
• X-ScaleHPC and X-ScaleAI packages
• **MVAPICH2-DPU**: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA’s DPU technology
• **SCR-Exa**: Efficient and Scalable Checkpointing for HPC and DL Applications
X-ScaleSolutions

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

• 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)

• Provide commercial support of these Libraries to US federal national labs and international supercomputer centers
X-ScaleSolutions (Cont’d)

• Winner of multiple U.S. DOE SBIR grants to design and develop innovative and value added products
• A Silver ISV member of the OpenPOWER Consortium
• More details on all products in [http://x-scalesolutions.com](http://x-scalesolutions.com)
  • [contactus@x-scalesolutions.com](mailto:contactus@x-scalesolutions.com)
Outline

• Overview of X-ScaleSolutions

• X-ScaleHPC and X-ScaleAI packages

• MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA’s DPU technology

• SCR-Exa: Efficient and Scalable Checkpointing for HPC and DL Applications


X-ScaleHPC Solution

• 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
• Professional technical support and customer services
• Stable and growing list of commercial customers
Features of X-ScaleAI Solution

• Built on top of MVAPICH2 libraries
• Integrated packaging to support popular DL frameworks
  • TensorFlow, PyTorch, MXNet, etc
• Integrated profiling and introspection support for DL applications across the stacks (DeepIntrospect)
  • Provides cross-stack performance analysis in a visual manner and helps users to optimize their DL applications for higher performance and scalability
• Targeted for both CPU-based and GPU-based DL training
• 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 OpenPOWER and x86 platforms
• Support for InfiniBand, RoCE and NVLink Interconnects
X-ScaleAI: Distributed TensorFlow on Summit (1,536 GPUs)

- ResNet-50 Training using TensorFlow benchmark on SUMMIT -- 1536 Volta GPUs!
- 1,281,167 (1.2 mil.) images
- Time/epoch = 3 seconds
- Total Time (90 epochs) = 3 x 90 = 270 seconds = 4.5 minutes!

ImageNet-1k has 1.2 million images

MVAPICH2-GDR reaching ~0.42 million images per second for ImageNet-1k!

*We observed issues for NCCL2 beyond 384 GPUs

Platform: The Summit Supercomputer (#2 on Top500.org) – 6 NVIDIA Volta GPUs per node connected with NVLink, CUDA 10.1
X-ScaleAI DI GUI Profiler View (Expended)

**DEEP INTROSPECT (DI) DASHBOARD:**

**PROMPT:**
```
xscale-ai-run --np 1024 --hostfile ./hostfile ./xscale-ai/install/miniconda/bin/python ./xscale-ai/install/benchmarks/horovod_benchmarks/pytorch/pytorch_synthetic_benchmark.py --batch-size=64
```
Harness 30% higher performance and better scaling on DeepLabv3+ (using TF) with the X-ScaleAI Tool.
X-ScaleAI Use Case #2: Application Benefits (ResNet-50)

- As a result of tuning the MPI layer, the user can vastly improve application performance

<table>
<thead>
<tr>
<th># GPUs</th>
<th>Images/sec (Expected)*</th>
<th>Images/sec (Obtained Initially)</th>
<th>Images/sec (Obtained Finally)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>~370,000</td>
<td>181,020</td>
<td>341,590</td>
</tr>
</tbody>
</table>

1.9x speedup in ResNet-50 (using PyTorch) throughput, while reducing debugging time for the DL scientist considerably!!
Outline

- Overview of X-ScaleSolutions
- X-ScaleHPC and X-ScaleAI packages
- **MVAPICH2-DPU**: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA’s DPU technology
- **SCR-Exa**: Efficient and Scalable Checkpointing for HPC and DL Applications
Requirements for Next-Generation MPI Libraries

- Message Passing Interface (MPI) libraries are used for HPC and AI applications
- Requirements for a high-performance and scalable MPI library:
  - Low latency communication
  - High bandwidth communication
  - Minimum contention for host CPU resources to progress non-blocking collectives
  - High overlap of computation with communication
- CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation

Network offload mechanisms are gaining attraction as they have the potential to completely offload the communication of MPI primitives into the network
Overview of BlueField-2 DPU

- ConnectX-6 network adapter with 200Gbps InfiniBand
- System-on-chip containing eight 64-bit ARMv8 A72 cores with 2.75 GHz each
- 16 GB of memory for the ARM cores

How to Re-design an MPI library to take advantage of DPUs and accelerate scientific applications?
MVAPICH2-DPU Library 2021.08 Release

- Based on MVAPICH2 2.3.6
- Released on 08/22/2021
- Supports all features available with the MVAPICH2 2.3.6 release ([http://mvapich.cse.ohio-state.edu](http://mvapich.cse.ohio-state.edu))
- Novel frameworks to offload non-blocking collectives to DPU
  - Alltoall (MPI_Ialltoall)
  - Allgather (MPI_Iallgather)
  - Broadcast (MPI_Ibcast)
MVAPICH2-DPU Library 2021.08 Release (Cont’d)

- Significantly increases (up to 100%) overlap of computation with any mix of MPI_Ialltoall, MPI_Iallgather, or MPI_Ibcast non-blocking collectives
- Accelerates scientific applications using any mix of MPI_Ialltoall, MPI_Iallgather, or MPI_Ibcast non-blocking collectives

Available from X-ScaleSolutions, please send a note to contactus@x-scalesolutions.com to get a trial license.
Evaluation Setup

- Being run on the HPC-AI Advisory Council cluster
  - 32 Xeon nodes connected with 32 DPUs over 200Gbps InfiniBand
  - 1,024 CPU cores (Xeons) and 256 ARM cores (DPUs)
- Configuration
  - Server HW:
    - CPU: Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz
    - Adapter: Nvidia BlueField-2 DPU, 8 ARM cores 2.75 GHz, 16GB DDR4
  - Software/Firmware:
    - OS version: CentOS 8.3
    - Driver version: 5.2-1
    - Firmware version: 24.30.1004
  - MPI:
    - MVAPICH2-DPU 2021.08
    - OSU Micro-Benchmarks (OMB) 5.7.1
Total Execution Time with osu_ialltoall (32 nodes)

Benefits in Total execution time (Compute + Communication)

- 32 Nodes, 16 PPN
- 32 Nodes, 32 PPN
P3DFFT Application Execution Time (32 nodes)

Benefits in application-level execution time

- 32 Nodes, 16 PPN
- 32 Nodes, 32 PPN

Latency (s)

MVAPICH2

Grid Size

1024x102...
1024x102...
1024x204...
1024x204...
2048x204...

12% 14%
12% 14%
12% 14%
12% 14%
12% 14%

Latency (s)

MVAPICH2

Grid Size

1024x102...
1024x102...
1024x204...
1024x204...
2048x204...

18% 16%
18% 16%
18% 16%
18% 16%
18% 16%
Total Execution Time with osu_iallgather (16 nodes)

Benefits in Total execution time (Compute + Communication)
Total Execution Time with osu_Ibcast (16 nodes)

**Total Execution Time, BF-2 (osu_ibcast)**

- **MVAPICH2**
- **MVAPICH2-DPU**

**Message Size**
- 2M
- 4M
- 8M
- 16M

**Benefits in Total execution time (Compute + Communication)**

16 Nodes, 16 PPN

16 Nodes, 32 PPN
Support to Accelerate DL Training in Next Release

• Support for distributed CPU-based DL training using NVIDIA Bluefiled-2 DPUs
• Intelligent designs to offload different phases of DL training
• Up to 15% performance improvement in DL training time compared to without DPU offloading
• Support for PyTorch/Torchvision and user defined DNN models and datasets

The design is based on a recent research paper “Accelerating CPU-based Distributed DNN Training on Modern HPC Clusters using BlueField-2 DPUs” by A. Jain, N. Alnaasan, A. Shafi, H. Subramoni, D. Panda, 28th IEEE Hot Interconnects, Aug 2021
Outline

• Overview of X-ScaleSolutions
• X-ScaleHPC and X-ScaleAI packages
• MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA’s DPU technology
• SCR-Exa: Efficient and Scalable Checkpointing for HPC and DL Applications
Checkpointing for DL Frameworks & Applications

- Periodically saving snapshots of a DL model training is important for tolerating system failures
  - DL model training often requires long time to complete (sometimes, weeks or months)
  - Distributed DL model training at scale is more susceptible to system failures

- Single-machine DL training may simply load/store the DL model every $N$ epochs

- For distributed training, the following (naïve) scheme is recommended:

```python
for n in num_epochs:
    if rank == 0 and n % checkpoint_freq == 0:
        save_DNN()
        MPI_Barrier()
        ...
    if rank == 0 and interruption:
        load_DNN()
        MPI_Bcast(DNN_params)
```

- This scheme requires all other ranks be blocked while rank 0 writes checkpoint info to the PFS

Challenge: Can we be more efficient in checkpointing for DL Training at scale?
SCR-Exa: Efficient Checkpointing for HPC and DL Applications

- Based on open-source Scalable Checkpoint Restart (SCR) library
- Developed in collaboration with Lawrence Livermore National Lab (LLNL)
- Significantly increase portability and flexibility
  - Add support for diverse job launchers, resource managers, storage devices, etc.
  - Customized and optimized for a range of systems with different underlying protocols
- Enable fast and efficient restart and resume
  - Add support for launching applications with spare nodes
  - Automatically reconfigure to restart or resume using spare nodes after a failure (if possible)
- Significantly improve maintainability and extensibility
  - Add new python binding and python implementation of CLIs and APIs
- Expand support for DL/ML frameworks and applications
  - PyTorch, horovod, ResNet, EDSR, etc.
- Part of code enhancements are contributed back to the SCR open-source
Overview of SW Stack for Enabling SCR-Exa library
Enabling SCR-Exa for DL Frameworks and Applications

• Instrumented version of DL platforms to use SCR_Exa library:
  • PyTorch Distributed Data Parallel Model (DDP)
  • PyTorch Over Horovod

• DL Applications:
  • Residual Neural Network (ResNet)
  • Enhanced Deep Residual Networks for Single Image Super-Resolution (EDSR)
Use of SCR-Exa for DL Applications (ResNet-50)

- Use PyTorch DDP platform
- Checkpoints saved every epoch in both naïve and SCR-Exa approaches
- SCR-Exa saves checkpoints to the local node, and only write to the PFS every 10 epochs.
- SCR-Exa is very efficient (about 10%-15% overhead) and scales very well (OpenPOWER9, V100 GPU)
Enabling SCR-Exa for DL Applications (ResNet-50, Cont’d)

- Similar performance trends observed for the PyTorch over Horovod platform
Enabling SCR-Exa for DL Applications (EDSR)

- Use PyTorch DDP platform
- Same training parameters as ResNet-50, but scaled to 1024 GPUs
Enabling SCR-Exa for DL Applications (EDSR, Cont’d)

- Similar performance trends observed for the PyTorch over Horovod platform
Thank You!

d.dai@x-scalesolutions.com

http://x-scalesolutions.com/