LOGICAL CLOCKS

WHITE PAPER

AMD/ROCm with Hopsworks

Deep Learning with AMD GPUs at Scale on TensorFlow

ROCm on Hopsworks

Radeon Open Compute Platform (ROCm)

ROCm is an open-source platform for GPU computing that enables TensorFlow/Keras/PyTorch applications to run seamlessly on AMD GPU hardware.. ROCm is scalable, supporting distributed training on multiple GPUs, with optional RDMA support for the most I/O intensive workloads.

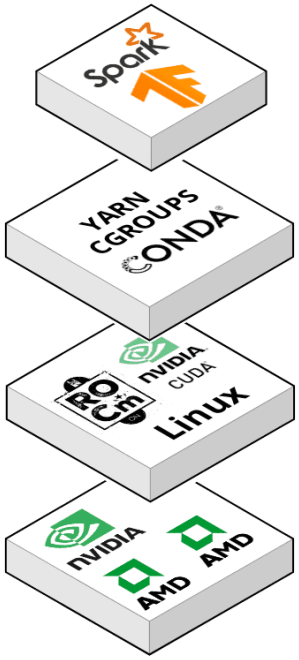

Figure 1: TensorFlow applications run unchanged on either ROCm or Cuda on Hopsworks, through HopsYARN support.

ROCm on Hopsworks

Hopsworks is a platform for the development and operation of ML applications at scale. On Hopsworks, you can run your TensorFlow/Keras applications unchanged with ROCm supported GPUs. During installation, Hopsworks ensures that the correct AMD drivers and ROCm are installed, and that Hopsworks' resource scheduler, HopsYARN, is configured to work with ROCm-enabled GPUs.

ROCm fully supported in TensorFlow 2.0

ROCm-specific modifications for TensorFlow are now in the TensorFlow master repository and community-supported builds are now available. With TensorFlow on ROCm, your TensorFlow programs can use the full range of available functionality, including support for half-precision training and multi-GPU training.

ROCm for HopsYARN

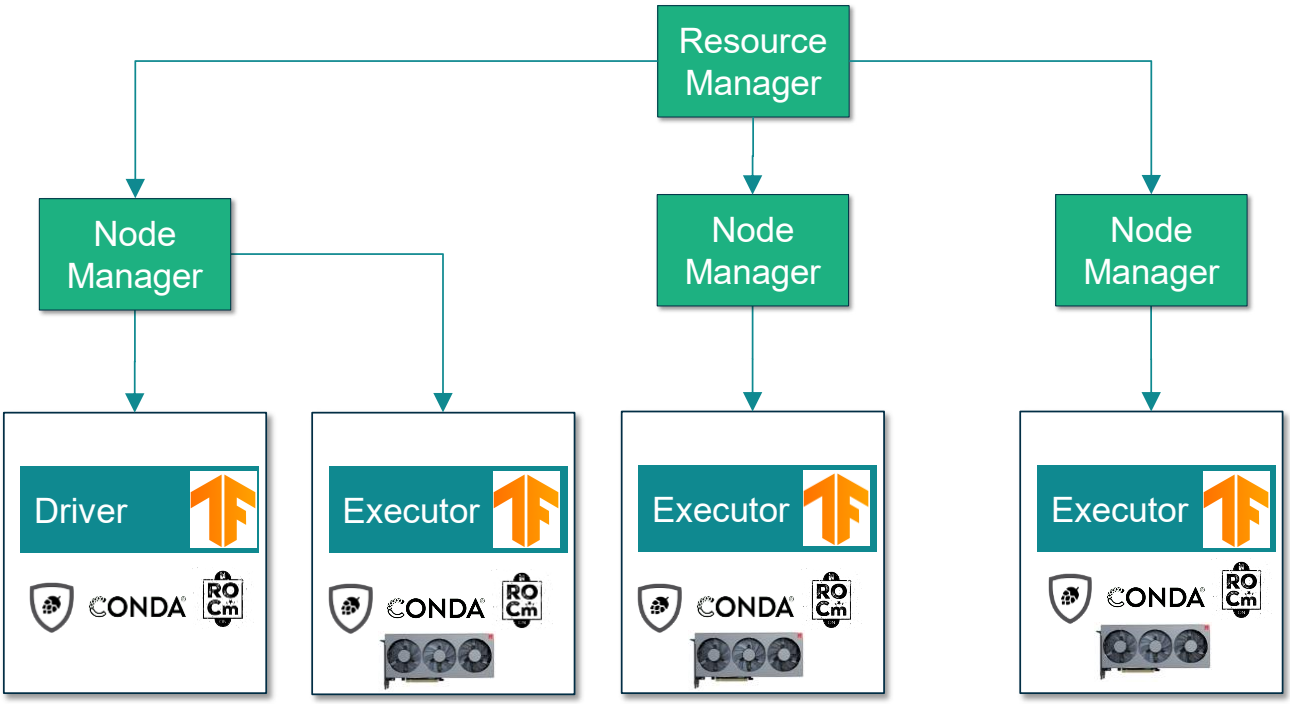

Figure 2: ROCm support in HopsYARN

HopsYARN is the resource scheduler in Hopsworks that enables applications to be scheduled on containers in a cluster. The resources managed by HopsYARN include CPUs, memory, and GPUs (hardware accelerators). HopsYARN supports AMD GPUs for Linux through integration of ROCm in YARN CGroups.

Data Scientists typically make use of AMD GPUs by writing Python applications with TensorFlow code included in a function that is scheduled in the cluster by PySpark (and HopsYARN). That is, Data Scientists write Python code in Jupyter notebooks, after first configuring the notebook to as many GPUs/CPUs/memory as the program needs. When the Python application is run in the notebook, PySpark lazily allocates the container (with the AMD GPU(s) using HopsYARN) to an executor that will run the TensorFlow code. This approach enables TensorFlow applications to be written for anything from 1 GPU to 100s of GPUs (for either parallel hyperparameter optimization or distributed training).

ROCm Performance on AMD GPUs

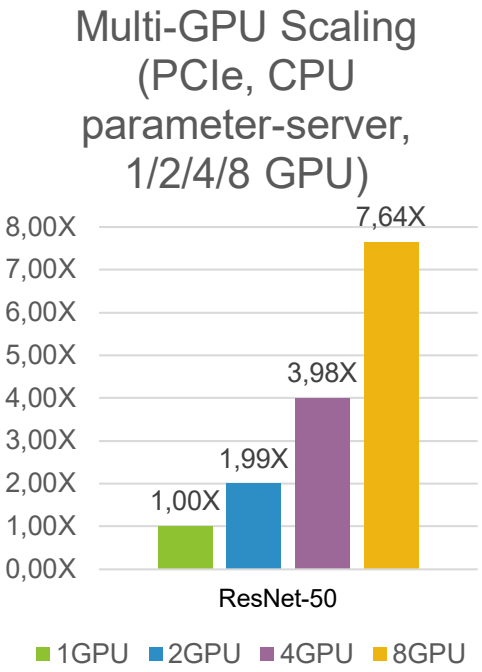

Figure 2: Add GPUs to speed up training - this example uses the AMD MI50. [From Hopsworks AMD talk, SparkSummit19]

The ROCm stack leverages highly scalable support for inter-node and intra-node communication. RCCL (pronounced "Rickle") is a stand-alone library of standard collective communication routines for GPUs, implementing all-reduce, all-gather, reduce, broadcast, and reduce-scatter. It has been optimized to achieve high bandwidth on platforms using PCIe, xGMI as well as networking using InfiniBand Verbs or TCP/IP sockets. RCCL supports an arbitrary number of GPUs installed in a single node, and can be used in either single- or multi-process (e.g., MPI) applications.

Affordable Deep Learning is here

Both high-end AMD GPUs (MI50, MI60) with up to 32GB RAM, PCI 4.0 and AMD Infinity Fabric, as well as consumer-grade AMD Radeon VII support deep learning with ROCm (TensorFlow/PyTorch). For example, the low-cost AMD Radeon VII has 16GB of memory, enabling it to train large models with large batch sizes, while having similar performance to the Nvidia 2080Ti (for training CNNs, see table), costing significantly less, but they have a software license that allow them to be used in a data center. Nvidia's EULA (end-user license agreement) forbids the use of GeForce cards, like the 2080Ti, in a data center for deep learning. Selected benchmarks that are available for the ResNet-50 network are shown in the table below.

	Nvidia 2080Ti	AMD Radeon VII
Memory	11 GB GDDR6	16 GB HBM2
Bandwidth	616 GB/s	1028 GB/s
Software	Cuda	ROCm
ResNet-50 (fp32)	294 images/sec**	316 images/sec*
ResNet-50 (fp16)	466 images/sec**	421 images/sec*
Retail Price	$999	$699

* https://github.com/ROCmSoftwarePlatform/tensorflow-upstream/issues/173

**https://lambdalabs.com/blog/2080-ti-deep-learning-benchmarks/

the creators of

HOPSWORKS

Reach out to us!

www.logicalclocks.com

info@logicalclocks.com

.. or visit us at one of our offices:

Silicon Valley Office
470 Ramona St
Palo Alto,
94301 California
USA

Stockholm Office
Box 1263,
Isafjordsgatan 22
164 40, Kista
Sweden

UK Office
IDEALondon,
69 Wilson St
EC2A2BB, London
UK