Technical White Paper

SMART MAMMOGRAM DIAGNOSIS

Advancing Graphics Acceleration in Healthcare

Apr 2021 v1.3

Collaboration by
FATHOMX | NVIDIA | HPE APAC Innovation Center

Francis LO khiam-foh.lo@hpe.com
Alicia LEONG alicia.leong@hpe.com
FENG Hao hao.feng@hpe.com
DU Hao duhao@u.nus.edu

Contents

List of figures

List of tables

INTRODUCTION

This paper highlights the collaboration testing performed by FATHOMX and HPE on the training model for mammogram diagnosis solution which is powered by HPE APOLLO 6500 server with NVIDIA Tesla V100 and NVIDIA Tesla A100 GPU modules. It outlines the testing methodology, demonstrates how FATHOMX and its AI-assisted mammogram solution could benefit in improved performance using the NVIDIA A100 Tensor Core GPU.

The objectives of the test are to ensure that AI model training could be effectively performed under the system environment and to possess better understanding on any improvement in performance introduced by NVIDIA A100 Tensor Core GPU. Performance in functionality is one of the tests performed to determine that the AI model training could be executed without any problem. The tests are also performed against 30,000 relevant data set of images, measuring the average processing time. In addition, the health status of the system, specifically the CPU and the GPU, is also being monitored.

OVERALL SYSTEM & ITS KEY COMPONENTS

The ability of computers to autonomously learn, predict, and adapt using massive datasets is driving innovation and competitive advantage across many industries and applications.

The following block diagram shows the solution of the AI enabled mammogram diagnostic solution. AI training is performed at the HPE Apollo servers with NVIDIA A100 GPU with dataset and the state-of-the-art algorithm framework from FATHOMX. The inference model is then delivered to the Edge for real-time analysis by radiologists.

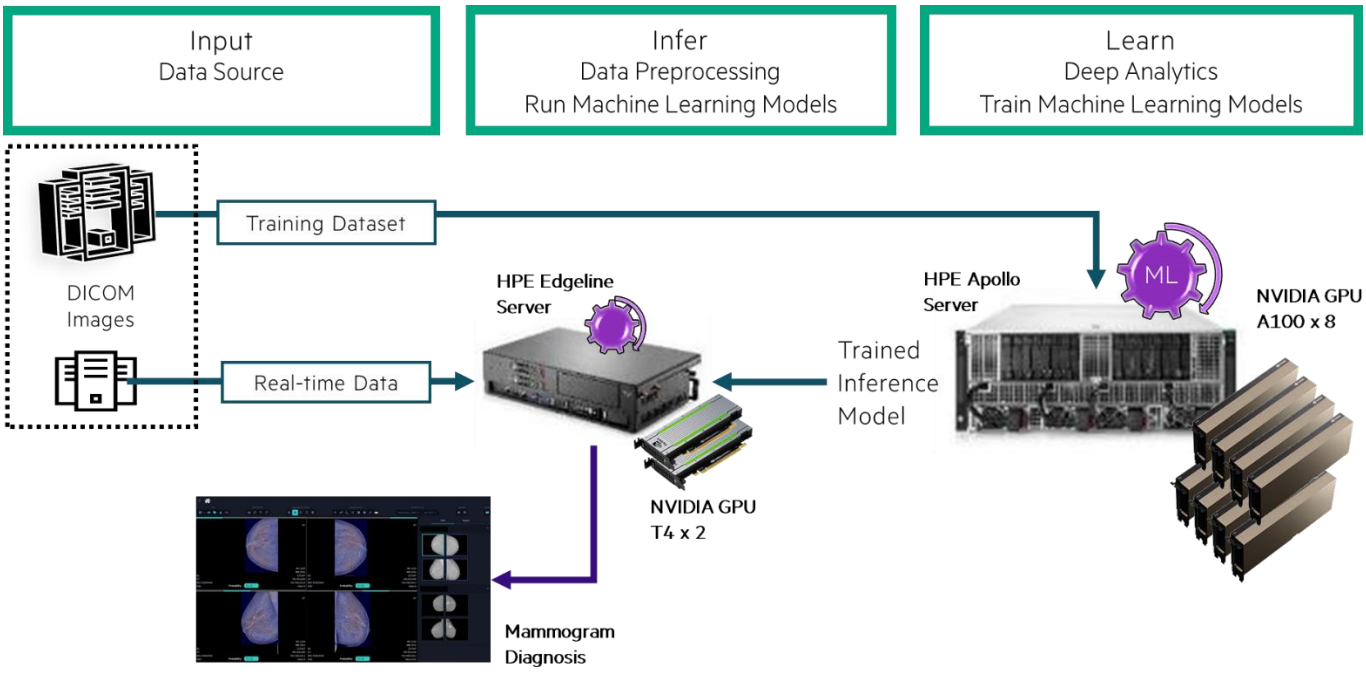

Figure 1: Mammogram Diagnosis System

The key building blocks of the test include the following components

1. HPE – High-Performance Computing (HPC) System
2. NVIDIA – Graphics Processing Unit (GPU) Accelerator
3. FATHOMX – Artificial Intelligence (AI) Engine

HPE Apollo 6500 System

The HPE Apollo 6500 System is an ideal HPC and Deep Learning platform providing unprecedented performance with industry leading GPUs, fast GPU interconnect, high bandwidth fabric and a configurable GPU topology to match required workloads. The system with rock-solid RAS features (reliable, available, secure) includes up to eight high power GPUs per server tray (node), NVLink for fast GPU-to-GPU communication, Intel® Xeon® Scalable Processors support, choice of up to four high-speed / low latency fabric adapters, and the ability to optimize your configurations to match your workload and choice of GPU. And while the HPE Apollo 6500 Gen10 System is ideal for deep learning workloads, the system is suitable for complex high-performance computing workloads such as simulation and modelling.

Eight GPU per server for faster and more economical deep learning system training compared to more servers with fewer GPU each. Keep your researchers productive as they iterate on a model more rapidly for a better solution, in less time. Now available with NVLink to connect GPUs at up to 300 GB/s for the world's most powerful computing servers. HPC and AI models that would consume days or weeks can now be trained in a few hours or minutes.

Some of the key capabilities of the HPE Apollo 6500 System include

- Accelerated performance for GPU-intensive workloads
- Flexibility for HPC and deep-learning environment
- Resiliency and simplicity with ease of serviceability and upgrades

NIVIDIA Tesla A100 GPU

The NVIDIA A100 Tensor Core GPU is the flagship product of the NVIDIA data center platform for deep learning, HPC, and data analytics. It delivers unprecedented acceleration at every scale for AI, data analytics, and HPC to tackle the world's toughest computing challenges. As the engine of the NVIDIA data center platform, A100 can efficiently scale up to thousands of GPUs or, using new Multi-Instance GPU (MIG) technology, can be partitioned into seven isolated GPU instances to accelerate of all sizes. A100's third-generation Tensor Core technology accelerates more levels of precision for diverse workloads, speeding time to insight as well as time to market.

Some of the groundbreaking innovation of the A100 include

- NVIDIA Ampere Architecture for large-scale workloads
- Multi-instance GPU (IMG) for partitioned GPU
- Third-generation Tensor Cores for increased deep learning FLOPS
- 40GB High Bandwidth Memory (HBM2)
- Next-generation NVLink for high application performance throughput
- Structural Sparsity for improved AI inference performance

FATHOMX Mammogram AI Engine

FATHOMX focuses on advancing breast cancer screening with AI. Existing mammography screening suffers from a high false positive rate of about 9% and have a workflow that requires multiple radiologists and takes up to 30 minutes to conduct just one screening, with another 2 to 4 weeks for a generated report.

With the introduction of an AI diagnostics solution, it supports and augments the work of radiologists, generating preliminary prediction diagnostics risk results with high accuracy rate that allows radiologists to take reference for priority study and achieving faster decisions.

The mammogram engine from FATHOMX acts as an independent analysis that automatically analyzes the mammogram scanned images. It predicts the risk with AI algorithms, marks the ROI (region-of-interest) with heatmap and generates an AI-assisted report. The report is then further endorsed by radiologist who shall make the final decision. The radiologist could modify the report and determine the potential risk level of patients before revealing to the patients for a follow-up treatment or a clearances discharge.

This process of using AI to generate prediction and report automatically could be performed near real-time. It contributes to a significant amount of time reduction in getting the "first opinion" of mammogram image analysis. At the same time, it may reduce the need of a "double blind analysis", allowing more patients to be attended in a shorter time. This impact becomes more even more substantial in areas with high demand of radiologists.

TRAINING SYSTEM SETUP

The training of AI model is performed using 2 units of HPE Apollo 6500 Systems. One of the systems is embedded with PCIe-based NVIDIA A100 Tensor Core GPU while another with SXM2--based NVIDIA V100 Tensor Core GPU. There may be some system configuration variations, but the key dependency system variance emerges from the GPU. GPU acts as the key determining factors that speeds up performance and throughput while executing the AI model training.

Keeping other dependency like the dataset, the algorithms, and other conditions constant, the performance measurement data is mainly driven by the performing difference contributed by the GPUs – NVIDIA A100 GPU and NVIDIA V100 GPU. The key components are shown in the following table below.

Table 1: Key System & Components

	HPE Apollo 6500 System	NVIDIA A100 GPU	NVIDIA V100 GPU
Device			
Quantity	2 units	8 units	8 units

The following 2 tables show the hardware configuration and the software modules installed in the systems under test. Basically, the software modules installed on to both systems are consistent while for the hardware configuration, the key difference indicates that System 1 is equipped with NVIDIA A100 while System 2 is equipped with NVIDIA V100 GPU.

Table 2: Hardware Configuration of Training Servers

	System 1	System 2
Server Model	HPE Apollo 6500 ProLiant XL270d Gen10	HPE Apollo 6500 ProLiant XL270d Gen10
Processor	2 X Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz (22 cores)	2 X Intel(R) Xeon(R) Gold 6150 CPU @ 2.70GHz (18 cores)
Memory	24 x 32GB DDR4 2666 MHz DIMMs	12 x 32GB DDR4 2666 MHz DIMMs
Storage	1 x 2TB NVMe SSD and 1 x 480GB NVMe SSD	2 x 2TB NVMe SSD and 2 x 400GB SAS SSD
GPU	**8 x GPU NVIDIA A100 PCIe 40GB**	**8 x GPU NVIDIA Tesla V100 SXM2 16GB**
Power Supply	2 x 2200 Watts	2 x 2200 Watts

Table 3: Software Modules installed on both Training Systems

	Software Version for System 1 and System 2
Operating System	CentOS 8.3 x86_64
Linux Kernel	v4.18.0-240.1.1.el8_3.x86_64
BIOS	U45 v2.22
GPU Driver	V460.27.04
CUDA Toolkit	V11.2
Docker Container Platform	v.20.10.2
NVIDIA-Docker Container	V2.5.0
Open-MPI	V4.0.5

MAMMOGRAM TRAINING METHODOLOGY

Mammogram images are large complex images that require good strategy to adopt to its layers of classification and segmentation. It possesses the intelligence to introduce classifiers by recognizing patches in the images.

Training Model

The training engine uses a proprietary deep convolutional neural network model developed by FATHOMX. It is built on top of the ResNet (Residual Network), taking advantage of its ability in resolving complex features, yet minimizing the impact of performance degrades which could be contributed by an increase in the number of extra layers in the deep learning.

The training engine is developed in the PyTorch Framework which is Python-based scientific computing package that consumed the power of GPUs, providing flexibility and speed in creating the deep neural network architectures developed by FATHOMX. In effect, it helps to address the demand of a mammogram image processing computer vision that may be subjected to complex features, time-consuming training, and extensive networking.

However, a couple of issues has surfaced when using PyTorch on A100. The following issues have occurred while using PyTorch on A100 in the experiment.

- A100 with CUDA capability sm_86 is not compatible with the existing PyTorch installation.
  The current PyTorch install supports CUDA capabilities sm_37 sm_50 sm_60 sm_61 sm_70 sm_75 compute_37.
- CUDA initialization is very slow during first run.
- runtime error: cuda error: no kernel image is available for execution on the device

The cause of these errors is a result of incompatibility of sm_80. It has been resolved by building PyTorch from the source of the specific versions, including all related torch-* libraries.

Training Dataset

The mammogram training is performed using the mammogram datasets found in the public space. They include the INbreast dataset and the DDSM dataset with details which could be found in the following

- INbreast: Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic radiology*, *19*(2), 236-248.

- DDSM: Bowyer, K., et al. "The digital database for screening mammography." Third international workshop on digital mammography. Vol. 58. 1996. http://www.eng.usf.edu/cvprg/Mammography/Database.html

The INbreast dataset contains 115 patients while the DDSM dataset contains 2,620 patients. In general, the median size of the mammogram images used in this dataset are in a resolution of 1473 x 3107.

MEASUREMENT & PERFORMANCE DATA

The AI model training has been performed in an end-to-end fashion with real-time capture of system parameters including CPU and GPU utilization, memory, and other parameters. These system performance data are captured using the following tool - Wandb (Weights and Biases) which can be found in the following link https://wandb.ai In this training performance experiment, the metrics are set with a moving average of 10 seconds.

The following figures show the performance, utilization and consumption of CPU and GPU of the systems. For the GPU parameters, GPU utilization and its power usage have been quite consistent for both servers with A100 and V100. The main difference surfaces at the GPU memory allocation for server with A100 and server with V100. Similarly, the CPU process and memory utilization show consistent data on both server with A100 and server with V100.

Figure 2: GPU Utilization Monitor

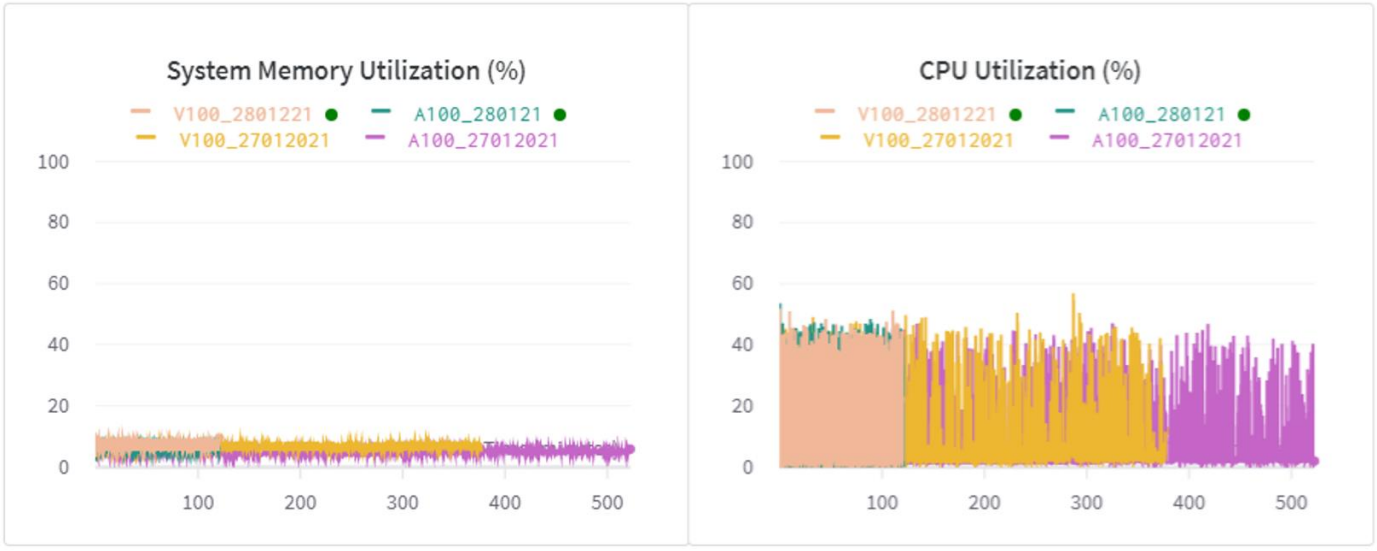

Figure 3: CPU Utilization Monitor

Experiments on both servers are run using Adam optimizer with initial learning rate at 3×10^{-4} (adjusted every 10 epochs), weight decay at 1×10^{-4}. All models are pretrained with feature extraction models trained on INbreast and DDSM lesion patches. All models are optimized with binary cross entropy loss, balanced by positive to negative ratio. Random rotation, flipping, affine operations are used for data augmentation.

Experiments on A100 server are run with batch size at 16 while experiments on V100 server are run with batch size at 8 due to the limited GPU memory.

Table 4: Batch Size

	Sever with A100	**Server with V100**
Batch Size	16	8

The following tables illustrate the time taken for the training and the result of the training model. There is a notable increase in the outcome for server with A100 with better AUC parameter (measures the classification performance) and throughput. These are essential diagnostic parameters for the healthcare industry application as every percentage point increase in accuracy marks a significant contribution.

Table 5: Duration of Training

	Server with A100	**Server with V100**
Dataset: INbreast	10h 11m 4s	6h 28m 22s
Dataset: DDSM	6d 43m 20s	3d 22h 17m 42s

Table 6: Training Accuracy

	Server with A100	**Server with V100**
AUC: INbreast	0.904	0.881
AUC: DDSM	-	-
GPU utilization	68.75%	60.02%
Throughput (per batch)	263 MB	132 MB

RESULT OF BENCHMARKING

The rapid advance in machine learning especially in deep learning has continued to fuel the enhancement of image processing in the development of healthcare industry. From the measurement data captured while performing the AI model training, the findings show that A100 demonstrates marginal advantage in performance improvement on accuracy and throughput against V100.

Performance

The result shows that the experiments on A100 system demonstrate higher performance when compared to those on V100 system in breast cancer detection for mammography images. The improvements may be contributed by a larger memory within the A100 GPU, enabling larger training batch size. Batch normalization module in deep convolutional neural networks usually benefits from larger batch sizes. When the batch size increases, the input sampling will be closer to the real distribution of dataset, leading to better generalization in validation and testing.

Accuracy

The result exhibits increase in AUC parameters, resulting in an improvement in accuracy of the training model. This results in a positive impact on the quality of the outcome for the AI classifiers.

Possible Improvement

Several improvements could be explored to take advantage of the NVIDIA A100 GPU to achieve a more substantial difference in performance for NVIDIA A100 over NVIDIA V100 GPU. These include the various potential enhancement that could lead to future considerations of experiments.

- Optimize code specific for NVIDIA-Ampere architecture of A100 GPU over NVIDIA-Volta architecture of V100 GPU
- Adjust training model parameters to further leverage on the higher memory including cache and shared memory of the GPU for A100 for a balanced optimal batch size and epoch value (the number of passes of the entire training dataset that the machine learning algorithm has completed)
- Standardize the hardware interface of PCIe and SXM2 for a more rational measurement comparison to contribution caused by the limitations of system interface

INDEPENDENT BENCHMARKING WITH MLPERF

MLPerf is one of the independent tools that could access the performance of the ML (Machine Learning) system. It consistently measures the ML performance objectively and could be used as benchmarking measurement for training and inference performance of the ML hardware, software, and services.

MLPerf can also be used as a guideline to understand how the training model could perform under the two different system environments, typically containing NVIDIA A100 Tensor Core GPU against NVIDIA V100 Tensor Core GPU. In future, the MLPerf test will be carried out to set the expected performance limits achievable by the systems.

CONCLUSION

In conclusion, the study demonstrates that for an end-to-end deep learning model training, the upgrade of GPU from NVIDIA A100 to NVIDIA V100 would lead to satisfactory level of improvement in performance when optimization is done in target to the NVIDIA A100 GPU.

This technical whitepaper also serves as a powerful digital soft tool that both the sales and pursuit teams from HPE and NVIDIA could leverage on while driving along with their subsequent demand generation through digital marketing. This takes reference relevantly to other AI-focused companies that are driving AI solutions similar to FATHOMX. It addresses an increase in awareness and differentiators for both HPE and NVIDIA when driving AI-based solutions in the space of promoting computing and accelerated processing.

RESOURCES & ADDITIONAL LINKS

To learn more about HPE Apollo Systems, visit
https://www.hpe.com/us/en/compute/hpc/apollo-systems.html

To learn more about HPE Apollo 6500 System, visit
https://h20195.www2.hpe.com/v2/gethtml.aspx?docname=a00039976enw

To learn more about the NVIDIA A100 Tensor Core GPU, visit
www.nvidia.com/a100

To learn more about the FATHOMX Mammogram Solution and Features, visit
www.fathomx.co

Learn more at

www.hpe.com

www.nvidia.com

www.fathomx.co

© Copyright 2018 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

This document contains confidential and/or legally privileged information. It is intended for Hewlett Packard Enterprise and Channel Partner Internal Use only. If you are not an intended recipient as identified on the front cover of this document, you are strictly prohibited from reviewing, redistributing, disseminating, or in any other way using or relying on the contents of this document.

April 2020