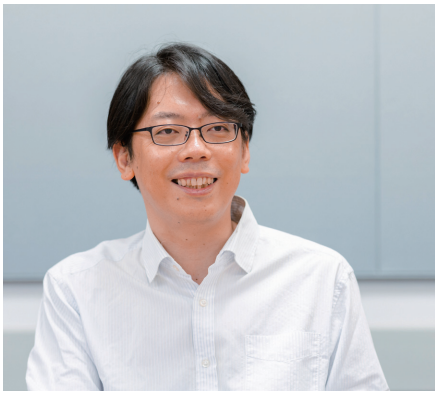
White Paper

Habana®Gaudi® AI Processor

Riken Accelerates Medical and Drug Research with Habana Gaudi AI Processor

Kei Taneishi, a research scientist at the Institute of Physical and Chemical Research (RIKEN), has advanced AI research in medical care and drug discovery. Here he shares his experience in applying Habana® Gaudi® AI processors for faster and more efficient deep learning training in the fields of healthcare and life sciences.



Mr. Kei Taneishi

Data Scientist

National Research and Development Agency RIKEN Photonics Research Center

The Growth of Transformer Learning Models Raises the Bar of Higher Computing Power

Taneishi has conducted research on Deep Learning (DL) based disease analysis with medical imaging, including chest X-rays, virtual screening of small molecules affecting molecular targets, and the structural and functional changes of proteins. AI-based research in these areas demands a large-scale computational platform, making AI accelerators vital.

"Since Transformers first presented in 2017 and attention mechanisms are recognized to be an essential factor in deep learning, the former models based on a convolutional and recurrent neural network architecture have been rewritten, even in areas outside natural language processing," says Taneishi. "This results in a significant expansion of computational frameworks, leading to further demand for advanced DL computing performance."

RIKEN hosts Hokusai Sailing Ship (HSS) as one of its large-scale shared computing platforms, and HSS is optimized mainly for data science with no AI accelerator. Instead, the institute has established its heterogeneous computing environment combining on-premises and cloud infrastructures, wherein the cloud technologies of four major vendors run alongside Fugaku, the Center for Computational Science's supercomputer for general purpose applications, and RAIDEN, the RIKEN Center for Advanced Intelligence Project's computer system for AI-development.

Superior Price Performance and Easier Model Porting

Of the institute's cloud-based computing resources, Amazon EC2 DL1 instances powered by Habana® Gaudi® AI processors are of particular interest to Taneishi. These instances feature eight Habana® Gaudi® AI processors, 3rd generation Intel® Xeon® Scalable processors with 96 vCPUs, 768GB of memory, a 400Gbps bandwidth network, and 4TB of local storage. Habana® Gaudi® AI processors boast eight Tensor Processing Cores (TPC), 32GB of high-bandwidth memory, and ten integrated 100 GbE RDMA over Converged Ethernet (RoCE) ports. The eight Gaudi devices on the DL1 instance are connected all-to-all via these RoCE ports, providing excellent scaling efficiency. The AWS EC2 DL1 instances provide up to 40% better price performance for training deep learning models compared to current generation GPU-based EC2 instances.[1]

"Habana® Gaudi® AI processors enable both tensor operations and matrix multiplications on hardware, delivering powerful performance by optimally compiling deep learning computational graphs," explains Taneishi.

Released in 2019 when Transformer had been adopted as a standard, these processors help train Transformer models with brilliant efficiency. And the advantages of purpose-built DL architecture among AI accelerators enable Gaudi-based solutions to provide exceptional price performance and more than doubled training throughput compared to V100 GPUs at similar cost for computer vision and NLP models."

Using TensorFlow and PyTorch models with Habana® Gaudi® AI processors is easy with small script rewrites required. As code changes with Habana® SynapseAI® SDK are minimal, developers can switch AI selectors to match their computing resources, while leveraging existing scripts, minimizing migration effort and costs.

"Using TensorFlow and PyTorch syntaxes allows us to work with AI accelerators without any change, so we can effortlessly port Habana® SynapseAI® SDK-supported models and their derivatives," explains Taneishi. "We need to add a few lines of code to control the AI accelerator, but no issue is found in compatibility in the process of data definition, modeling, training or inference. The greatest advantage for researchers is that we can use highly cost-performant computers to try new models released through academic papers immediately."

Significant Acceleration: 22% Faster Prediction of Protein Secondary Structures, 18% Accelerated Classification of Disease Patterns from Medical Images[2]

Using the Habana® Gaudi® AI processor, Taneishi examined deep learning training in two areas: predicting the secondary structure of proteins with BERT-Large, a language model, and chest X-ray disorder classification using CheXNet, a computer vision model.

To predict the secondary structure of protein, in the first stage of 3D structure prediction, a DL model was trained with 20 amino acid residues constituting a family of proteins as input in a database of predetermined protein structures, resulting in an inference of secondary structure sequences. The test revealed that the Habana® Gaudi® AI processor took only 4.6 seconds per iteration to train the model—22% faster than a V100 GPU's 5.9 seconds (see Figure 1).

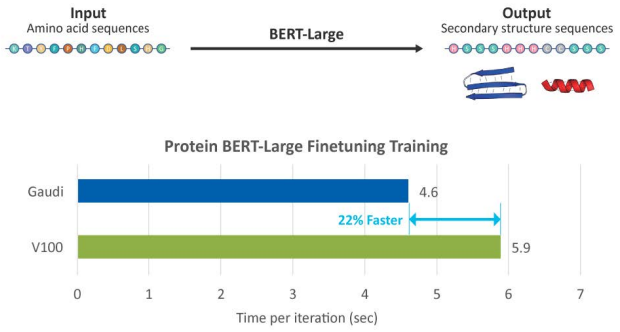


Figure 1. Predicting the secondary structure of proteins

The disorder classification using CheXNet input a dataset of 112,120 chest X-ray images from 30,805 patients to infer the type and location where symptoms such as pneumonia occur. The Habana® Gaudi® AI processor took 859.1 seconds in an iteration, showing 18% faster training compared to 1,047.7 seconds by a V100 GPU (see Figure 2). Habana® Gaudi® AI processors running in parallel with distributed data also ensure lower overhead and higher scalability within a single node.

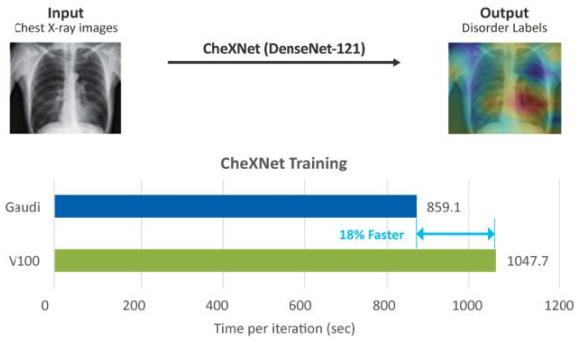


Figure 2. Disorder classification in chest X-ray images

Higher Expectations for Habana® Gaudi® 2 AI Processor, Delivering up to 3X the performance of predecessor[3]

To advance the future of AI drug discovery, Taneishi is currently working to provide a deeper understanding of 3D structural and functional changes caused by the existence of ligand binding in protein, detecting these transformations by training AI models from molecular dynamics (MD) trajectories identifying the distance between amino acid residues. In the medical AI field, he is planning to advance studies in modeling with multi-modal data in addition to medical imaging now that a huge amount of complex data—including electronic medical records, medical images, genomes, and lifestyles of tens of thousands of patients—is available.

Intel announced the 2nd generation Habana® Gaudi® 2 AI processors in May 2022. The latest in this family boasts enhanced computational efficiency, shrinking its process node to a 7nm from 16nm and offering 24 TPCs, three times as many as its predecessor. The second-generation processor also features 96GB of memory, three times more than the first generation, and considerable networking enhancements, with the number of GbE ports increasing from 10 to 24. These improvements provide a massive performance boost, delivering roughly 2X better performance in ResNet-50 training throughput compared to A100 GPUs[4]and 3 to 4.7X better performance than the first-gen Gaudi.[5]

"The more developers choose Habana® Gaudi® and Gaudi®2 AI processors, the more models and frameworks are becoming available to provide support, eliminating the barriers to deploy Habana® Gaudi® 2 AI processors," says Taneishi. "We'll are seeing the appearance of more powerful, easy-to-use services for cloud environments, which I'm eager to try."

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

1 https://aws.amazon.com/ec2/instance-types/dl1/
2 Data from internal test results conducted by Riken. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
3, 5 Comparison based on MLPerf published performance results of Gaudi and Gaudi2, June 2022; see MLPerf summary for result details. https://mlcommons.org/en/training-normal-20/
4 Habana ResNet50 Model: https://github.com/HabanaAI/ModelReferences/tree/master/TensorFlow/computer_vision/Resnets/resnet_keras, Habana SynapseAI Container: https://vault.habana.ai/ui/repos/tree/General/gaudidocker/1.6.0/ubuntu20.04/habanalabs/tensorflow-installer-tf-cpu-2.9.1, Habana Gaudi Performance: https://developer.habana.ai/resources/habana-training-models/, A100 / V100 Performance Source: https://ngc.nvidia.com/catalog/resources/nvidia:resnet_50_v1_5_for_tensorflow/performance, results published for DGX A100-40G and DGX V100-32G

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, Habana, Gaudi and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

JPN/2210/PDF/SE/BCG/YS 353411-001JA