Solution Brief

Scalable Deep Learning Inference Accelerator using FPGAs

Authors:

Ioannis Stamelos
Elias Koromilas
Chris Kachris
InAccel

Jing Lu
Xu Tianci
Inspur

Natalia Poliakova
Intel

Scalable deployment of Inspur Deep Learning Inference Accelerator on multi-tenant Intel FPGA cluster

If you are responsible for building, testing, or deploying deep learning inference models:

- **As a business strategist or executive**: You will better understand how to apply the latest technologies for deep learning to successfully generate increase the performance of your system and reduce the cost significantly.

- **As a technology decision-maker**: You will learn how to incorporate a cost-effective deep learning inference framework into your technology stack and at the same time enjoy
- Higher performance
- Higher performance
- Lower Latency
- Lower cost
- Lower energy consumption
- Instant Scalable deployment
- Multi-tenant deployment

Executive Summary

Inference refers to the process of using a trained machine learning algorithm to make a prediction. After a neural network is trained, it is deployed to run inference—to classify, recognize, and process new inputs.

TF2 is an open-source deep learning inference accelerator based on FPGA computing platform, developed by Inspur AI & HPC. A wide range of general purpose deep neural networks can be supported. Models from popular deep learning frameworks such as Pytorch, TensorFLow, and Caffe can be loaded into TF2 easily by toolkits we supplied.

In this Solution Brief we show how InAccel orchestrator can be integrated with TF2 to allow multi-tenant scalable deployment of TF2 to a cluster of FPGAs.

We show how InAccel's orchestrator allows **easy deployment, scaling, resource management, and task scheduling** for FPGAs making it easier than ever, the deployment and the utilization of FPGA for Deep Learning Inference.

TF2 Inspur Deep Learning Inference Accelerator TF2

TF2 is a deep learning inference accelerator based on FPGA computing platform, developed by Inspur AI & HPC. A wide range of general purpose deep neural networks can be supported. Models from popular deep learning frameworks such as Pytorch, TensorFLow, and Caffe can be loaded into TF2 easily by toolkits we supplied. The pretrained deep learning model can be compiled into FPGA without any code level FPGA development work, which can be an agile solution for AI inference applications on FPGA.

The TF2 accelerator is composed of two parts: Transform Kit and Runtime Engine.

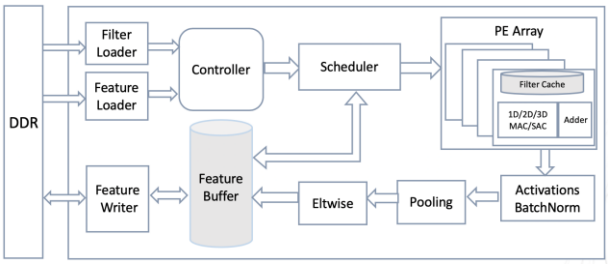
Transform Kit

Transform Kit is a tool for model optimization and conversion with modules of model compression, pruning and quantification, etc. Transform Kit aims to reduce model data size and simplify mathematical calculation. Additionally, computational node fusion can also be done in transform kit to relax the data access bandwidth limitation on computing performance by integrating multiple computing nodes into one.

Runtime Engine

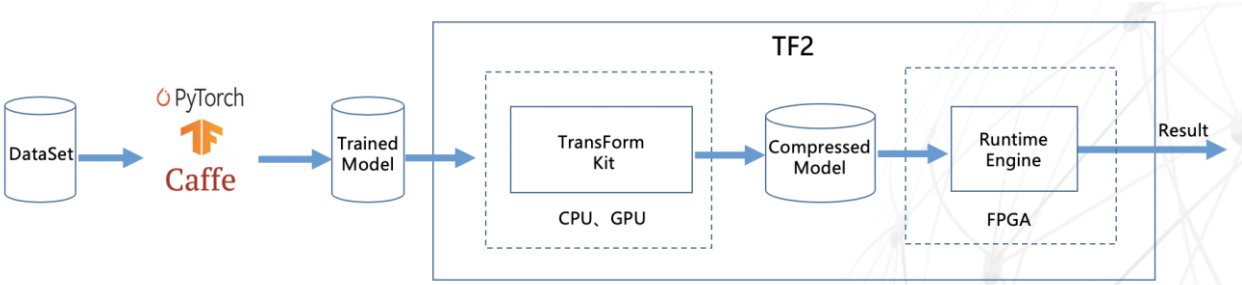
The TF2 Runtime Engine is an intelligent runtime FPGA accelerator which can automatically generate FPGA executive files. It first parses the network structure file and generates the network configuration file required by the Runtinme Engine, and then recompiles FPGA code, which can automatically generates the FPGA executive file.

Multiple convolutional layers are executed serially on the FPGA. In order to reduce the storage access limitation on computing performance, the intermediate feature map data is preferentially stored on the chip as much as possible. The model data is read from the external DDR to the FPGA in real time during the calculation process, but reading operation can be performed simultaneously with the calculation, that is, the reading time can be "hidden" under the calculation time.



The current version has 2D and 3D calculations. The vector length calculated by each MAC/SAC dimension can also be configured according to the specific application and FPGA computing resources. Networks such as ResNet50/SqueezeNet computing performance is listed as follows.

NetWork	Throughput(fps)
SqueezeNet	1485
GoogLeNet	306
FaceNet(MTCNN+SqueezeNet)	1020



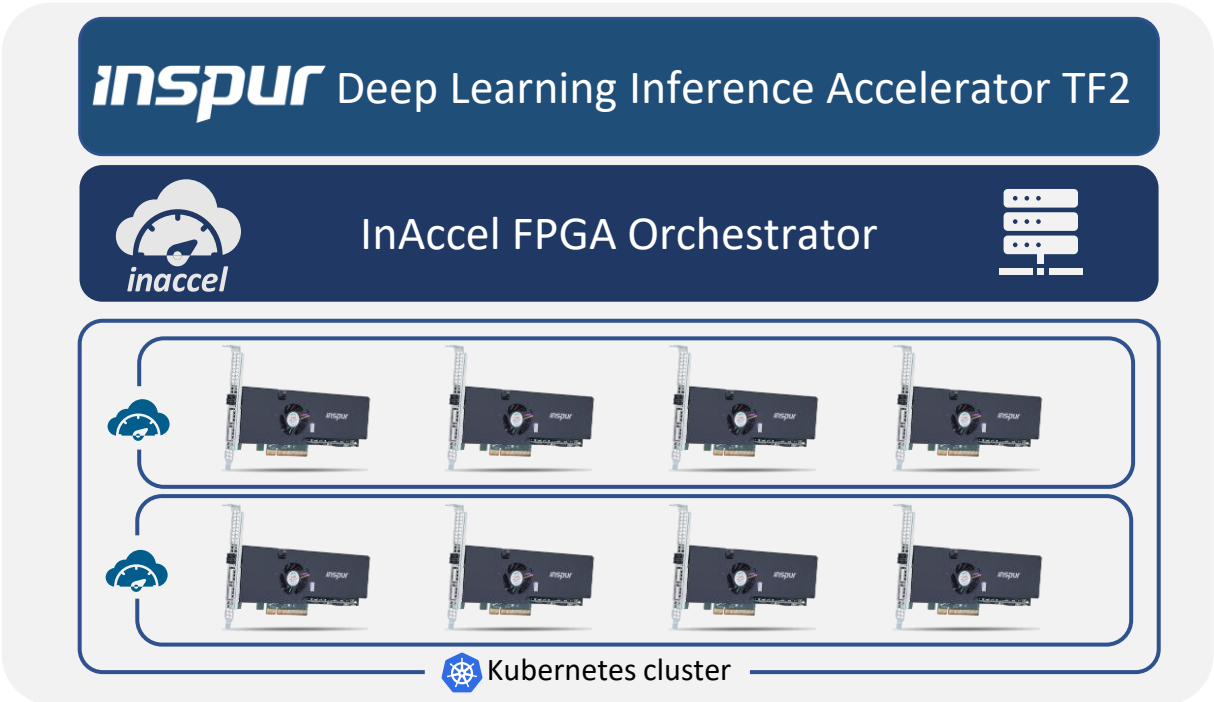


Figure 1. Scalable deployment of Inspur Deep Learning Inference on a cluster of Inspur cards using Intel FPGAs and the InAccel FPGA orchestrator

InAccel PAC cluster manager

In cases that multiple applications or processes need to utilize the PAC-based accelerators, and the application needs to be deployed in multiple servers, InAccel® Coral manager is used to abstract away the available resources and provide a simple API for software developers. InAccel Coral orchestrator abstracts away the available resources in a cluster of PAC cards, making it easier than ever to deploy one or more applications targeting multiple FPGAs.

InAccel's manager is used to schedule, dispatch and manage the functions that need to be accelerated. It performs the load balancing among the available resources in the PAC cluster and is also used for the management and configuration of the cards based on the functions that are offloaded.

Software developers can simply call the functions that need to be accelerated without worrying on the scheduling of the functions to the available resources or the contention of the resources from multiple applications.

InAccel's orchestrator is fully compatible with Inspur Deep Learning Inference Accelerator and the Intel® PAC cards.

InAccel Orchestration and Integration

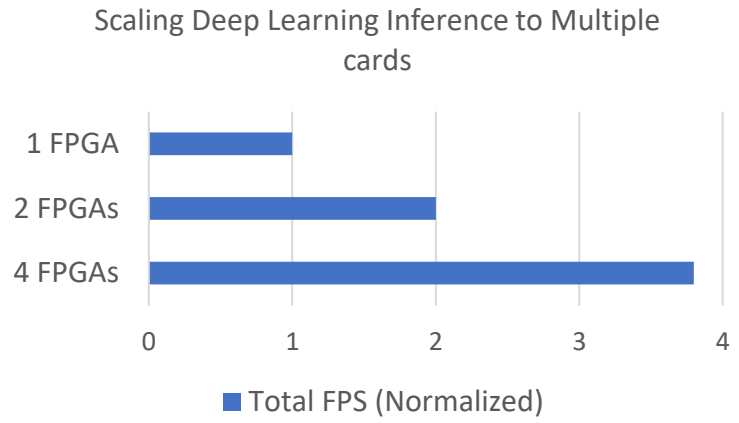
Using InAccel orchestrator, Inspur TF2 Deep Learning Inference Accelerator can be deployed on a cluster of FPGAs instantly. It also allows sharing of the available resource to multiple users or multiple applications.

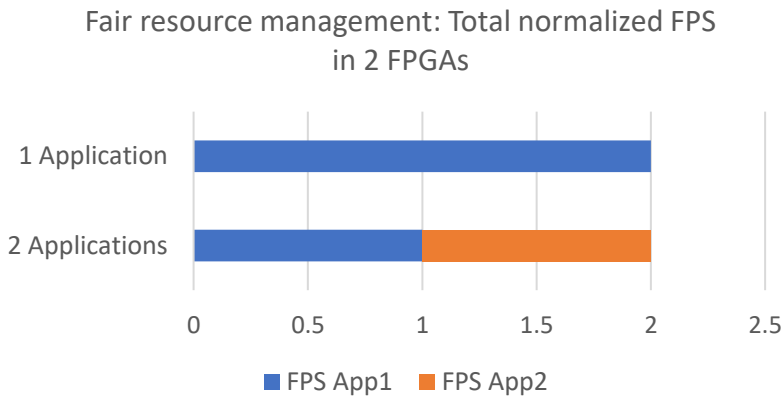
InAccel orchestrator performs automatically the dispatching of the workload to the available resources performing the load balancing, the resource management and the serialization of the requests.

The main advantage of the InAccel orchestrator is that it can be integrated easily and does not add any overhead on the Inspur accelerators.

The following figure shows the scaling of the accelerator to multiple FPGAs and the sharing of the resources to multiple users.

```
void run_on_fpga() {
resnet50("com.inspur.tf2.resnet50");
    resnet50.arg(*(input))
    .arg(*(model.filter_real))
    .arg(*(model.bias_bn))
    .arg(*(model.waitafterconvcycles))
    .arg(*(output))
    .arg(num_images);
wait(inaccel::submit(resnet50));
}
```





InAccel helps companies' speedup their applications, with zero code changes using efficiently state-of-the-art accelerators. InAccel provides a unique technology that allows the easy deployment, management, scaling and virtualization of FPGA-based accelerators. InAccel's FPGA orchestrator allows instant deployment and scaling of accelerator for widely-used applications like quantitative finance, big data analytics and machine learning.

Learn more : https://inaccel.com

Inspur is a leading provider of data center infrastructure, cloud computing, and AI solutions, ranking among the world's top 3 server manufacturers. Through engineering and innovation, Inspur delivers cutting-edge computing hardware design and extensive product offerings to address important technology arenas like open computing, cloud data center, AI and deep learning. Performance-optimized and purpose-built, our world-class solutions empower customers to tackle specific workloads and real-world challenges.

Learn more: https://en.inspur.com/