DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation

Seongmin Hong\textsuperscript{1}, Seungjae Moon\textsuperscript{1}, Junsoo Kim\textsuperscript{1}, Sungjae Lee\textsuperscript{2}, Minsub Kim\textsuperscript{2}, Dongsoo Lee\textsuperscript{2}, and Joo-Young Kim\textsuperscript{1}

\textsuperscript{1}CastLab, School of EE, KAIST, \textsuperscript{2}NAVER CLOVA
Abstract

• DFX: a low-latency multi-FPGA appliance for accelerating transformer-based text generation
  ◦ DFX is a multi-FPGA appliance that accelerates transformer-based text generation
  ◦ DFX adopts model parallelism to efficiently process the large-scale language model
  ◦ Xilinx Alveo U280 data center accelerator card provides high performance with low-cost
  ◦ FPGA-to-FPGA communication is enabled by QSFP cable at 100 Gb/s
Motivation
Transformer-based Text Generation

• **Text generation**
  ◦ Automatic generation of human-readable text by a computer
  ◦ Example: dialogue system, topic-to-essay generation, and code generation

• **Generative Pre-trained Transformer (GPT)**
  ◦ State-of-the-art model in natural language processing that scales up to 175B parameters
  ◦ High-quality text generation and remarkable inference accuracy for benchmarks (e.g., 86.4% for LAMBADA)

Input Tokens **"Hello, my name"**

Output Tokens **"is"** **"James"** **"Smith"** **"and"** **⋯** **"."**
Challenges of Transformer-based Text Generation

1) **System bottleneck** in the generation stage due to its sequential characteristic

2) **Massive model parameters** and computational requirements

3) **Lack of deployable hardware with end-to-end capability** for GPT inference in datacenters
DFX Architecture
• **Multi-FPGA appliance** for the acceleration of text generation

• **Intra-layer model parallelism** for large models

• **Compute core (accelerator) that supports GPT’s end-to-end operations**
Intra-layer model parallelism can reduce the latency of matrix operations

- **Multi-head attention**: model parameters are divided head-wise
- **Fully-connected layer**: model parameters are divided column-wise
• Compute core supports GPT’s end-to-end operations
  ◦ **Matrix processing unit**: matrix multiplication, masked matrix multiplication
  ◦ **Vector processing unit**: softmax, layer normalization, residual
  ◦ **DMA**: designed to maximize the HBM’s BW based on types of parameters (weight, bias, key, value, etc.)
Lightweight Router

- FPGA-to-FPGA interconnection in a ring network
  - Synchronization is necessary after executing distributed matrix multiplication
  - Network overhead is minimized with a simplified protocol
Evaluation
DFX Implementation

- DFX server prototype includes four Xilinx Alveo U280 FPGAs
- FPGA layout and resource utilization are optimized for HBM bandwidth usage
DFX Evaluation Results

- **Methodology**
  - **DFX**: one U280 FPGA, two U280 FPGAs, and four U280 FPGAs
  - **Baseline systems**: one V100 GPU, two V100 GPUs, and four V100 GPUs
  - **Models**: GPT-2 (345M), GPT-2 (774M), and GPT-2 (1.5B)
  - **Input token size**: varies from 32 to 128
  - **Output token size**: varies from 1 to 256

- **DFX achieves an average of 3.20×, 4.46×, and 5.58× speedup over GPU counterparts**

---

<table>
<thead>
<tr>
<th>Input Size:Output Size</th>
<th>GPU Appliance</th>
<th>DFX</th>
</tr>
</thead>
<tbody>
<tr>
<td>345M, 1 GPU vs 1 FPGA</td>
<td>310ms</td>
<td>3.20×</td>
</tr>
<tr>
<td>774M, 2 GPUs vs 2 FPGAs</td>
<td>227ms</td>
<td>4.46×</td>
</tr>
<tr>
<td>1.5B, 4 GPUs vs 4 FPGAs</td>
<td>228ms</td>
<td>5.58×</td>
</tr>
</tbody>
</table>

Inference latency of DFX compared to the GPU appliance
DFX Evaluation Results

- DFX achieves an average of $3.78 \times$ throughput and $3.99 \times$ energy efficiency on four-device appliances.

- Performance of DFX increases linearly with the number of FPGAs at the rate of 1.5.
Appliance Cost Analysis

- DFX is **8.21×** more cost-effective than the GPU appliance

<table>
<thead>
<tr>
<th></th>
<th>GPU Appliance</th>
<th>DFX Appliance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accelerators</strong></td>
<td>4 × Nvidia Tesla V100 32GB HBM</td>
<td>4 × Xilinx Alveo U280 8GB HBM</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>13.01 tokens/sec</td>
<td>72.68 tokens/sec</td>
</tr>
<tr>
<td>(Input:Output = 64:64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>$45,832* (1 GPU = $11,458)</td>
<td>$31,180* (1 FPGA = $7,795)</td>
</tr>
<tr>
<td><strong>Performance / Cost</strong></td>
<td>283.86 tokens/sec/million$</td>
<td>2330.98 tokens/sec/million$</td>
</tr>
</tbody>
</table>

*: The price is as of April, 2022
It may vary depending on market conditions

Newest U55C is only 4,395$ with 16GB HBM
Summary

• DFX is a multi-FPGA appliance for accelerating transformer-based text generation, featuring
  ◦ Intra-layer model parallelism
  ◦ Compute core supporting GPT end-to-end operations
  ◦ Lightweight router

• DFX achieves 5.58× and 3.99× improvements in performance and energy-efficiency compared to the GPU appliance’s

• DFX is 8.21× more cost-effective than the GPU appliance
Thank You

• What’s next?
  ◦ We are extending the model to one of GPT-3’s for a POC deployment in a datacenter

• Any questions? Feel free to contact us!
  ◦ Email: seongminhong@kaist.ac.kr or jooyoung1203@kaist.ac.kr
  ◦ Website: https://castlab.kaist.ac.kr/
  ◦ LinkedIn: https://www.linkedin.com/company/kaistcastlab/