

HOTCHIPS'22 Poster Session

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

Seongmin Hong¹, Seungjae Moon¹, Junsoo Kim¹,
Sungjae Lee², Minsub Kim², Dongsoo Lee², and Joo-Young Kim¹

¹CastLab, School of EE, KAIST,

²NAVER 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

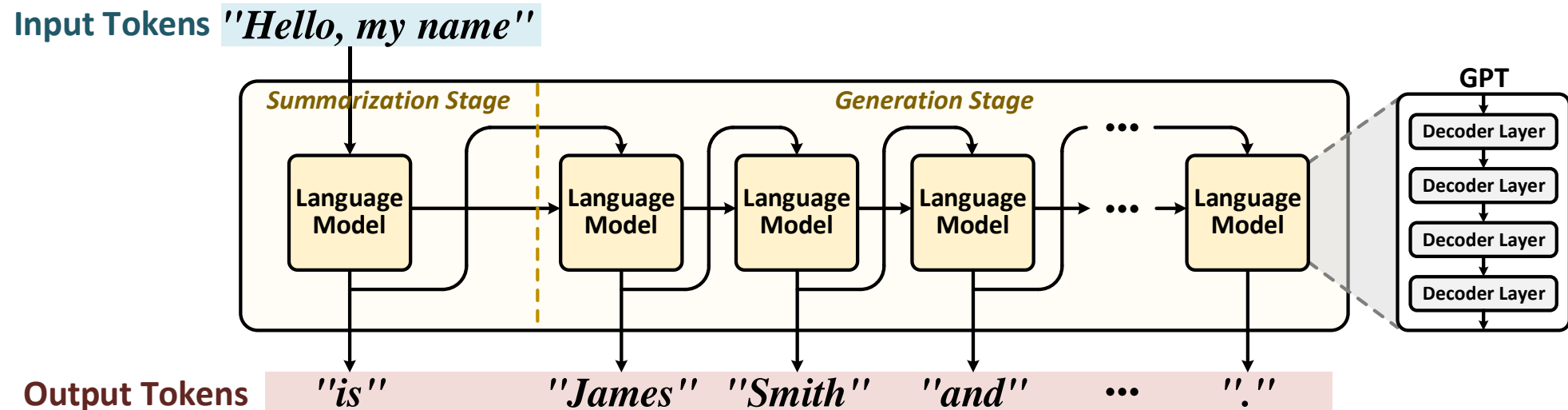
Transformer-based Text Generation

- **Text generation**

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

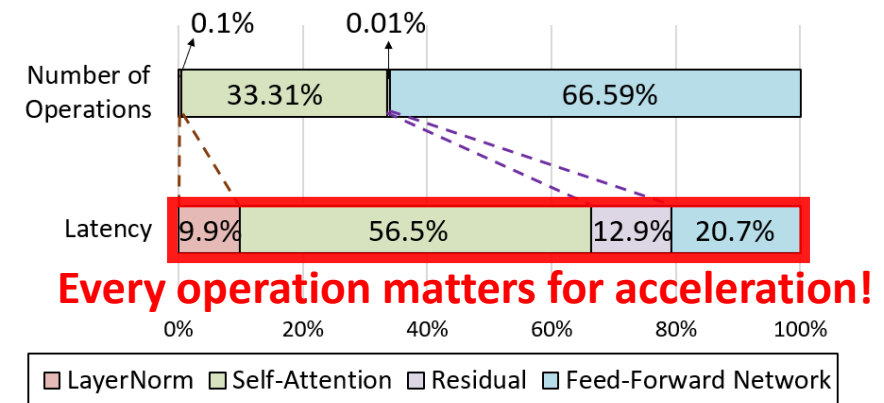
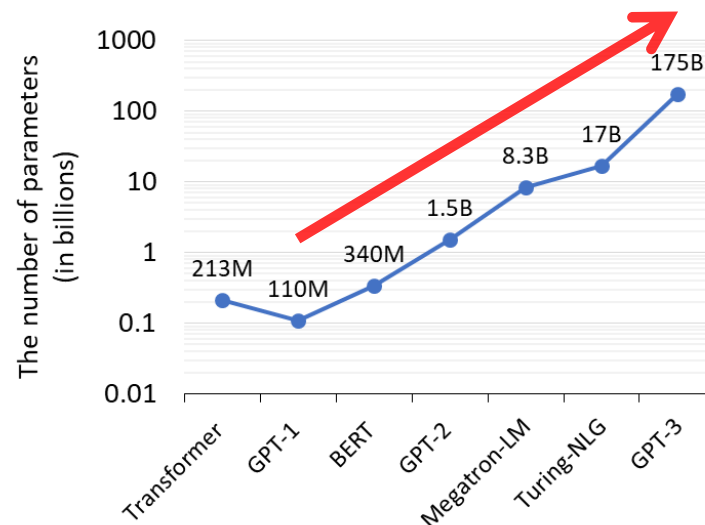
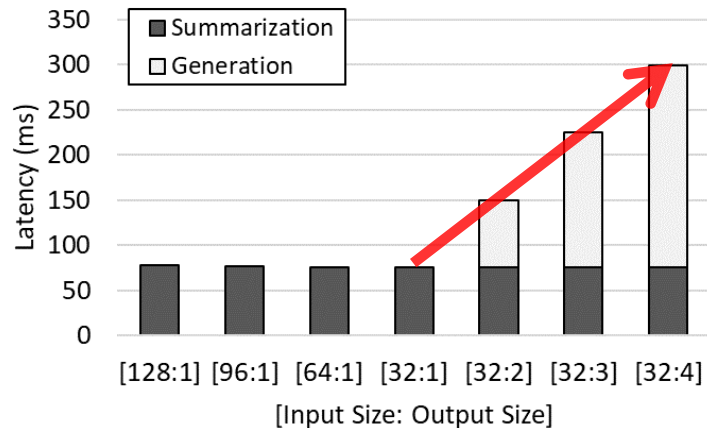
- **Generative Pre-trained Transformer (GPT)**

- 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)



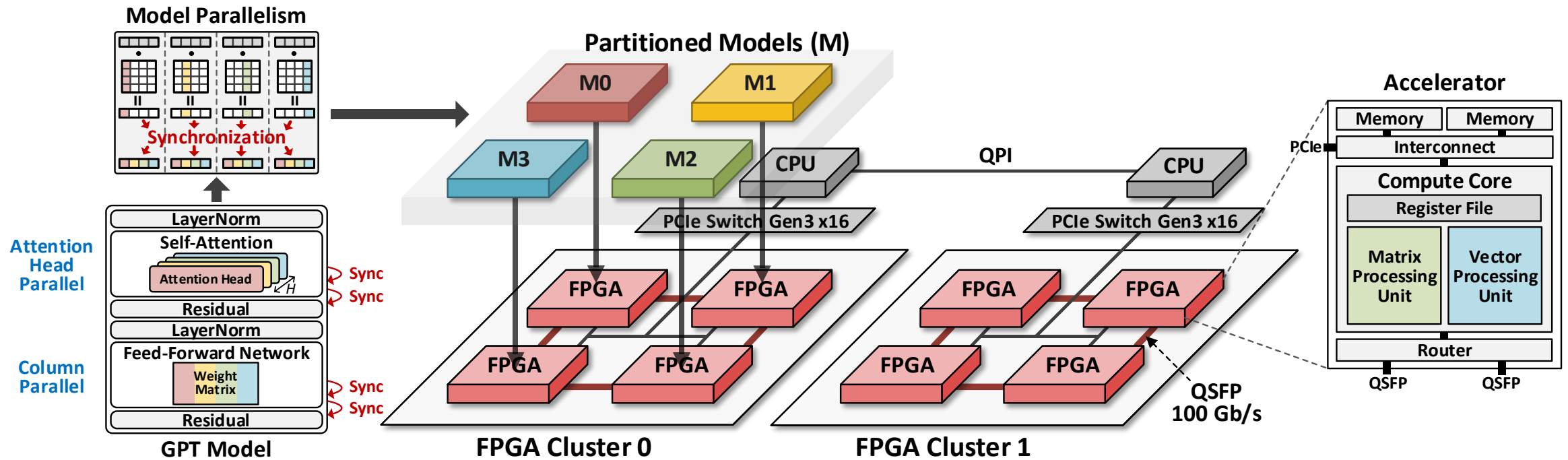
Challenges of Transformer-based Text Generation

- 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



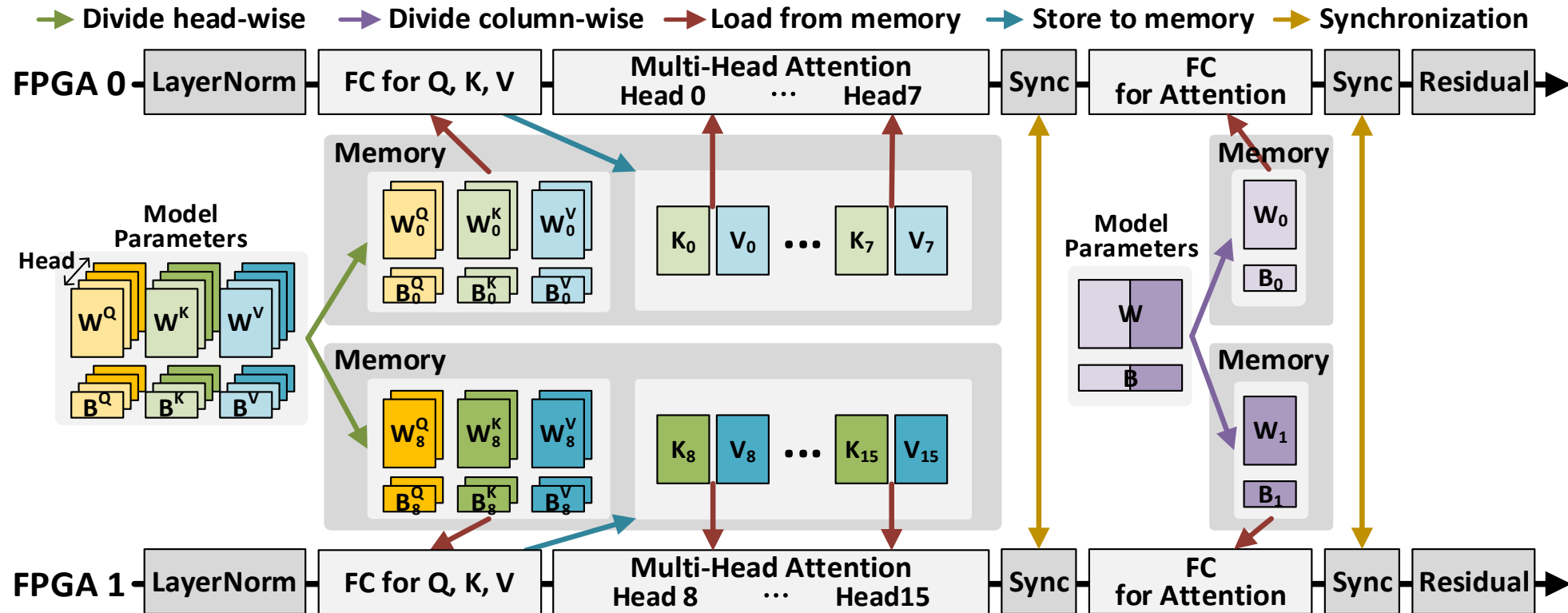
DFX Architecture

DFX Appliance Architecture



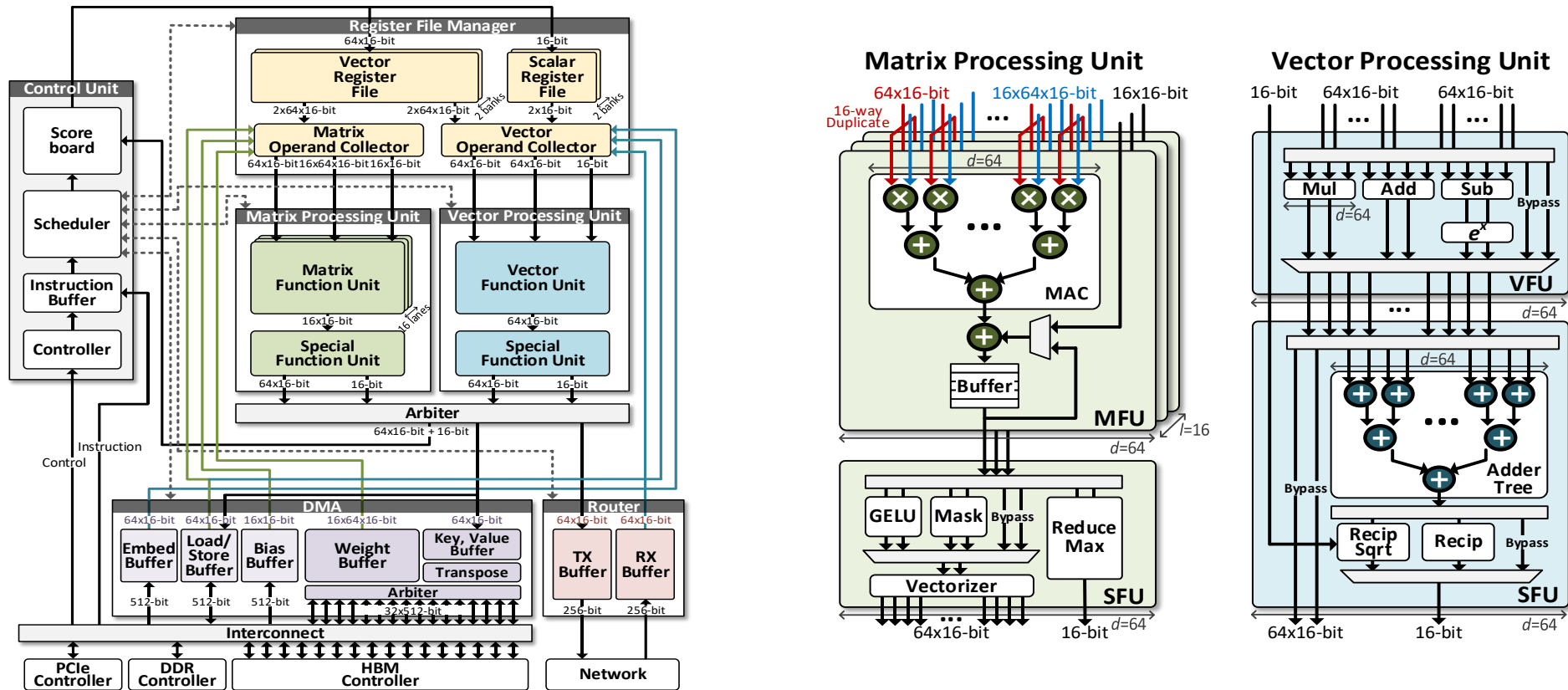
- **Multi-FPGA appliance** for the acceleration of text generation
- Intra-layer **model parallelism** for large models
- Compute core (accelerator) that supports GPT's **end-to-end** operations

Model Parallelism



- 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**

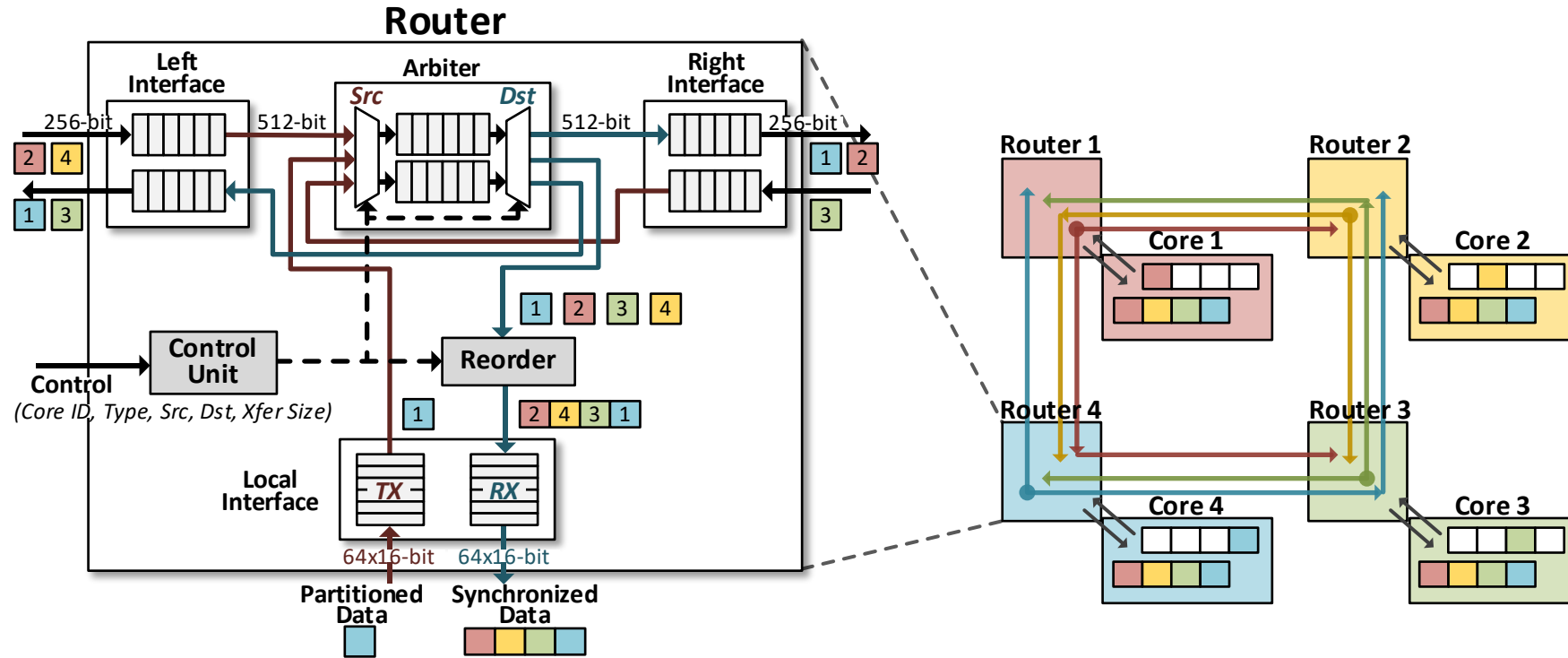
Core Architecture



- **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



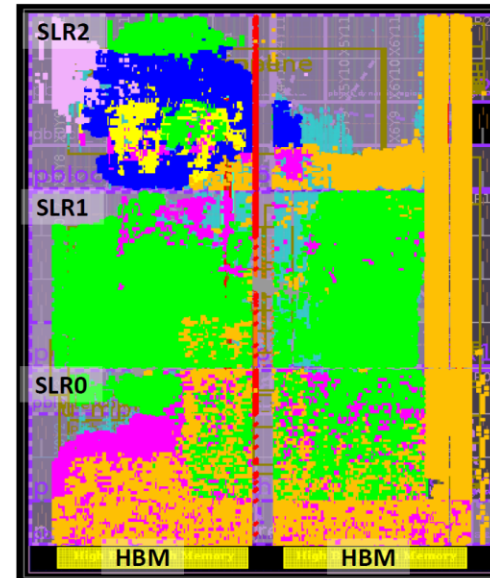
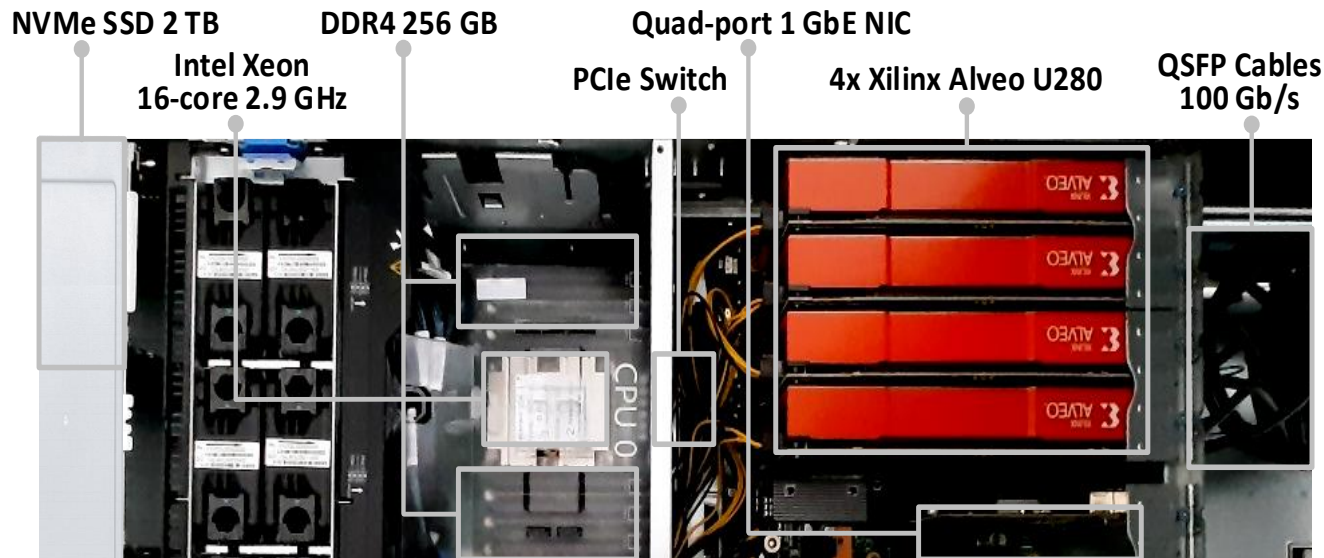
- **FPGA-to-FPGA interconnection in a ring network**

- **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

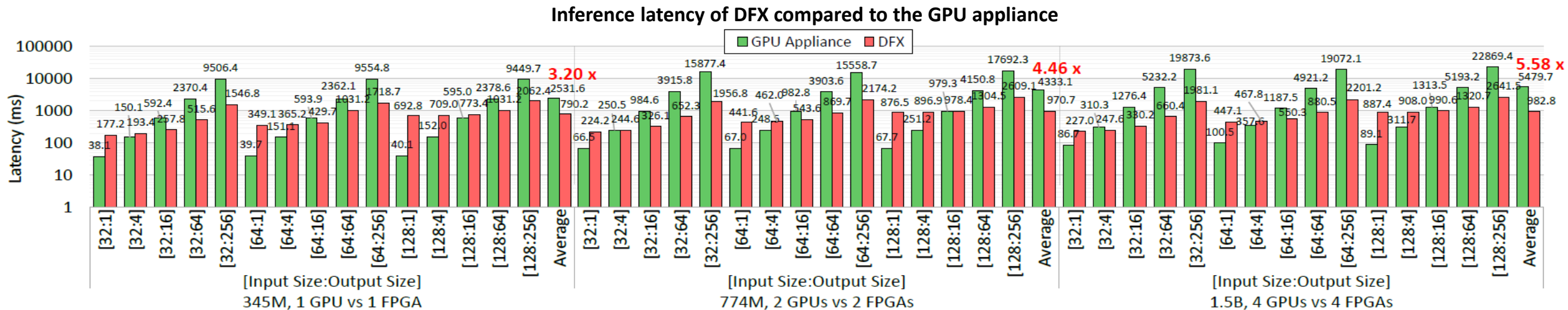


Component	LUT	FF	BRAM	URAM	DSP
Register File	6K (0.53%)	110K (4.22%)	88.5 (4.39%)	0 (0.0%)	0 (0.0%)
MPU	170K (13.06%)	381K (14.65%)	56 (2.78%)	0 (0.0%)	3136 (34.75%)
VPU	36K (2.77%)	55K (2.13%)	1.5 (0.07%)	0 (0.0%)	390 (4.32%)
DMA	38K (2.97%)	97K (3.74%)	134.5 (6.67%)	52 (5.42%)	0 (0.0%)
Router	3K (0.28%)	13K (0.55%)	24 (1.19%)	0 (0.0%)	0 (0.0%)
Interconnect	180K (13.83%)	303K (11.64%)	1237 (10.11%)	0 (0.0%)	4 (0.04%)
Total	520K (39.93%)	1107 (42.52%)	1192 (59.13%)	104 (10.83%)	3533 (39.15%)

DFX Evaluation Results

- **Methodology**

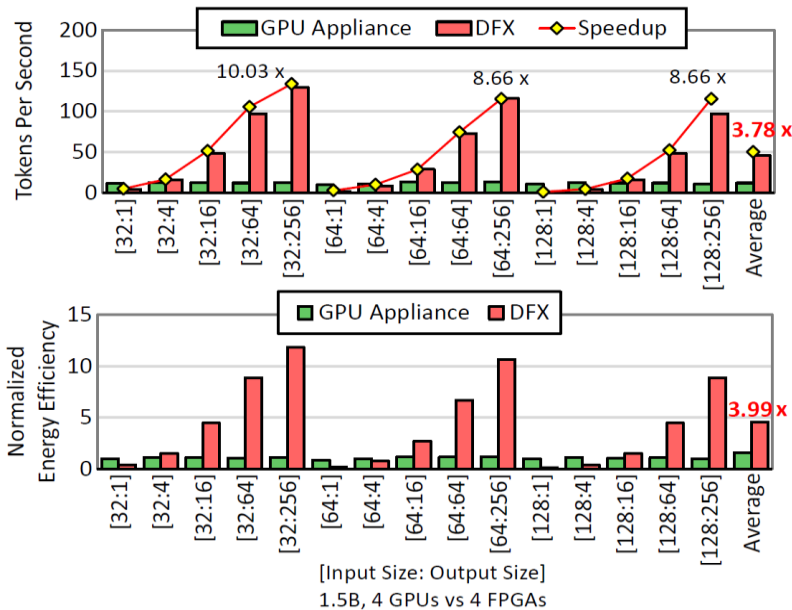
- **DFX:** one U280 FPGA, two U280 FPGAs, and four U280 FPGAs
- **Baseline systems:** one V100 GPU, two V100 GPUs, and four V100 GPUs
- **Models:** GPT-2 (345M), GPT-2 (774M), and GPT-2 (1.5B)
- **Input token size:** varies from 32 to 128
- **Output token size:** varies from 1 to 256



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

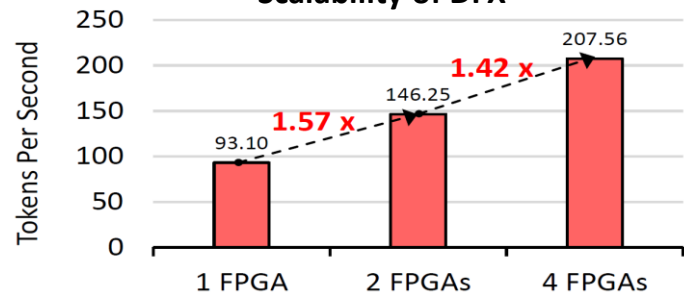
DFX Evaluation Results

Throughput and energy efficiency of DFX compared to the GPU appliance



- DFX achieves an average of **3.78x** throughput and **3.99x** energy efficiency on four-device appliances



Scalability of DFX



- Performance of DFX increases linearly with the number of FPGAs at the rate of **1.5**

Appliance Cost Analysis

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

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

*: 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 $5.58\times$ and $3.99\times$ improvements in performance and energy-efficiency compared to the GPU appliance's**
- **DFX is $8.21\times$ 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/>

