IEEE AICAS 2025 Grand Challenge - LLM Hardware System Design

1、Introduction

This track requires deploying and running LLM on the KV260 platform, leveraging the on-chip ARM processor and FPGA resources to optimize edge-side inference for the Qwen large language model.

2、Background

In recent years, large language models (LLMs) based on pre-training and Transformer technology have demonstrated outstanding performance in various downstream natural language processing tasks, such as text understanding, text generation, sentiment analysis, machine translation, and interactive question answering. However, due to concerns about data privacy and computational efficiency, achieving efficient LLM inference on the edge has emerged as a key development trend. The deployment of LLMs on edge devices faces significant challenges, as they often have tens of billions or even hundreds of billions of parameters. Furthermore, the growth of model parameter size far outpaces improvements in hardware performance. To address these challenges, academia and industry have begun exploring software-hardware co-optimization methods, such as model compression, dataflow optimization, and operator invocation, to enable efficient deployment and operation of large models on resource-constrained hardware.

Among these approaches, designing and implementing dedicated accelerators for efficient LLM inference has proven to be a practical technical solution. To advance research in this field and nurture interdisciplinary talent in computer architecture, circuits and systems, artificial intelligence, and high-performance computing, this year's AICAS will host a competition focused on optimizing LLM performance on FPGA platforms.

3、Competition Description

A. Challenge Plan

Participants will base their work on the Qwen2.5 large language model (LLM) and propose relevant methods from multiple perspectives, such as model compression, parameter sparsity, precision quantization, and structural pruning. They will leverage the on-chip ARM processor and FPGA resources to optimize the deployment of the LLM.

Qwen2.5 is the LLM officially released by Alibaba Cloud in September 2024, featuring parameter scales ranging from 0.5B to 72B. The model demonstrates comprehensive performance across mainstream benchmark evaluation sets. All models are pretrained on a large-scale dataset containing up to 18T tokens, providing Qwen2.5 with significantly more knowledge compared to its predecessor, Qwen2. Among these, the 0.5B-parameter general-purpose model, Qwen2.5-0.5B-Instruct, is an open-source and free version. Alibaba Cloud offers global multi-channel access to the model, along with services for training, deployment, and inference. Participants will work with the Qwen2.5-0.5B model, utilizing the KV260 computing platform to optimize LLM deployment. They will leverage the on-chip ARM processor and FPGA resources to achieve performance improvements. Participants are encouraged to propose methods from various perspectives, including but not limited to:

- FPGA-based accelerator design (required)
- Model compression (quantization, pruning), or speculative execution
- Optimization of computational scheduling (improving data reuse and pipelining)

The goal is to enhance inference performance of the LLM on the hardware platform through innovative solutions in software-hardware co-design. FPGA-based accelerator design is a required component of the competition. The effectiveness of the participants' optimization methods will be evaluated using test schemes designated by the competition committee.

Participants will remotely access the Kria KV260 board through interfaces provided by the competition committee for the verification, testing, and performance evaluation of their accelerator systems. They will compare the "Ability" and "Efficiency" of their designs before and after optimization. During the preliminary stage, each team will have a limited number of daily accesses to the cloud platform. After the preliminary round, 16 teams that submit valid results and pass code reviews will advance to the final stage. In the final stage, teams will further refine their designs and conduct optimizations. They will then need to evaluate the overall performance and compare results before and after deployment. Ultimately, the competition will award first, second, and third prizes to one team each, along with seven honorable mention awards. Teams advancing to the final stage will receive increased daily access to the cloud platform. Outstanding designs may be invited for publication in IEEE conferences or journals.

B. Competition schedule

- Preliminary Round Start: 2024-12-01
- Training on FPGA and Remote Platform Usage: TBD
- Preliminary Round End: 2025-02-15
- Competition Workshop *：2025-02-22

- Final Round Start: 2025-02-24
- Final Round End: 2025-03-24
- Code and Technical Report Submission：2025-03-30
- Competition Results Announcement: 2025-04-10
- AICAS Conference**：2025-04-28

Competition Workshop *： After the preliminary round, the organizing committee will host a technical workshop in Hangzhou, inviting the top 10 teams from each competition track that advanced to the final round. The workshop will feature keynote speeches by industry experts and include networking activities. The organizing committee will reimburse travel and accommodation expenses incurred by participating teams attending the workshop (if they meet the competition's reward criteria).

AICAS Conference **：After the semifinals, the organizing committee will invite the **top three** winning teams to present their projects at the 2025 AICAS main conference in France and participate in conference activities. The committee will reimburse the travel and accommodation expenses incurred by the participating teams during the seminar (if they meet the competition's award criteria).

C. Registration

- Log in to the Tianchi：https://tianchi.aliyun.com/competition/entrance/532288?lang=en-us
- Complete registration by filling in personal information. Participants and mentors are required to provide real-name verification.
- Each participating team may include up to five participants and two mentors, and each participant can only join one team.
- Participant registration, team formation, and modification operations must be completed by 23:59:59 on February 15, 2025.**（Verification Process：Tianchi official website -个人中心-认证-支付宝实名认证）。**
- Scan the QR code to join the DingTalk group chat.
- This challenge is open to current students and researchers from universities and research institutions.
- Employees of T-Head Semiconductor, Arm, and Alibaba DAMO Academy are not eligible to participate in the award evaluation.

D. Competition's award criteria

- Each track offers a cash prize pool of $5,550 and a travel subsidy of $7,500.

The top 10 teams from the preliminary round will be invited to attend the technical seminar in Hangzhou, and the top 3 teams from the semifinals will be invited to the AICAS 2025 conference. The organizing committee will provide travel subsidies (pre-tax) as follows:

- Hangzhou Technical Seminar：USD300 or CNY 2100（10 teams）
- AICAS2025：USD1500 or CNY 10500 (3 teams)

The selection criteria are based on their final scores, submitted technical reports, and the quality of their papers.

The distribution of prizes for this competition (pre-tax) is as follows:

- First Place：USD 2000 or CNY 14000（1 team）
- Second Place：USD 1500 or 10500（1 team）
- Third Place：USD 1000 or 7000（1 team）
- Excellence Award：USD 150 or 1050（7 teams）

4、Scoring Criteria

Evaluation Metrics	Evaluation Contents	Evaluation Method
Model Inference Accuracy（Ability）	● Accuracy Capability Testing	Run the local analysis tool to generate a JSON file.
Model Inference Performance （Efficiency）	● Parameter Compression Ratio ● Throughput Improvement Rate	Run the local analysis tool to generate a JSON file.

A. Preliminary Round Evaluation Metrics

1. Model Inference Accuracy（Ability）Evaluation：

Accuracy Testing：In this contest, we conducted n accuracy tests using the data from benchmark GLUE WNLI. The ground truth of the output is denoted as $\{p_i\}, i \in [1, n]$. Meanwhile, the contestant deployed quantized Qwen2.5-0.5B-Instruct on the KV260 platform, and the outputs for the given inputs from the benchmark are written as $\{q_i\}, i \in [1, n]$. We use the match rate between them as the accuracy:

$$Ratio_{accuracy} = (\sum_{i=1}^{n} 1_{p_i=q_i})/n$$

2. Model Inference Performance（Efficiency）Evaluation ：

Parameter Compression Ratio：The model size (in bytes) before and after model compression. The initial size of the Qwen2.5-0.5B-Instruct model is denoted as M_{ori} (model.safetensors format), and the optimized size of the Qwen2.5-0.5B-Instruct model is denoted as M_{opt}, which is the final model file loaded and deployed on the hardware. The specific calculation formula is as follows:

$$Ratio_{memory} = \frac{M_{ori} - M_{opt}}{M_{ori}} = 1 - \frac{M_{opt}}{M_{ori}}$$

The throughput improvement rate for prefill and decoding：The throughput (tokens per second) of the model before and after optimization. The throughput of the initial Qwen2.5-0.5B-Instruct model on the hardware platform is denoted as T_{ori}, and the throughput of the optimized Qwen2.5-0.5B-Instruct model is denoted as T_{opt}. The results for the prefill and decoding stages are calculated separately. The specific calculation formula is as follows:

$$Ratio_{throughput\_P/D} = \frac{T_{opt\_P/D} - T_{ori\_P/D}}{T_{opt\_P/D}} = 1 - \frac{T_{ori\_P/D}}{T_{opt\_P/D}}$$

The total score of the preliminary round will be calculated with weighting, and each metric will be normalized in the weighted formula based on the highest score for that metric across all participating teams. Contestants will upload their results to the Tianchi platform, which will then provide the total score and ranking online. The specific calculation formula is as follows:

$$Tot_{1st} = 0.4 \times \frac{Ratio_{accuracy}}{MAX(Ratio_{accuracy})} + 0.2 \times \frac{Ratio_{memory}}{MAX(Ratio_{memory})} + 0.2 \times \frac{Ratio_{throughput\_P}}{MAX(Ratio_{throughput\_P})} + 0.2 \times \frac{Ratio_{throughput\_D}}{MAX(Ratio_{throughput\_D})}$$

The current requirements for the competition entries include:

- Test Report：Introduce the test results after optimizing QWEN-0.5B-Instruct by your team, and provide documentation of the results from the specified third-party testing software.
- Technical Documentation: Describe the optimization methods implemented by your team for the LLM.
- Source Code of the Optimization Method: Provide the source code for the implemented optimization method.

B. Semi-Final Evaluation Metrics

The evaluation metric will be updated correspondingly based on the outcome of the first round of contest.

After the final round of the competition, participating teams are required to submit:

- Paper: A paper that presents the team's LLM application scenarios, the hardware-software co-design approach, as well as a comparison of the evaluation metrics and performance summary before and after optimization.
- Source Code: The actual source code used for model optimization must be submitted. *(The competition organizers have the right to use the code and development cases from the GC participating teams as promotional material in open-source communities such as GitHub, Huggingface, etc., or for inclusion in open-source tools related to sponsor products.)*

Note: The maximum number of submissions per day for the preliminary and semi-final rounds is 5.

5、Committee

- Xi'an University of Electronic Science and Technology: Wei Mao, Bo Li
- Shanghai Jiao Tong University: Yongfu Li, Zhezhi He
- Nanjing University: Li Du, Yuan Du
- University of Electronic Science and Technology of China: Liang Chang
- T-head (Shanghai) Semiconductor Technology Co., Ltd.: Xiaohan Ma, Guosheng Yu
- Arm Technology (China) Co., Ltd.: Fengzhi Pan, Xile Yang

6、Relevant Links

- 2025 AICAS: http://www.aicas2025.org/
- T-Head Semiconductor Co., Ltd.: https://www.t-head.cn/
- Tongyi Qianwen: https://qianwen.aliyun.com
- Arm Technology (China): https://www.armchina.com
- Huggingface (Global download link for Qwen model): https://huggingface.co/Qwen/Qwen2.5-0.5B
- Modelscope (Chinese download link for Qwen model): https://modelscope.cn/models/qwen/Qwen2.5-0.5B/summary
- Qwen Github (Download links for all open source versions): https://github.com/QwenLM/Qwen2.5
- Optimum-benchmark: https://github.com/huggingface/optimum-benchmark
- KV260: https://www.amd.com/en/products/system-on-modules/kria/k26/kv260-vision-starter-kit.html#specifications