D26-002

適用於即時且高品質影像生成之Transformer擴散模型處理器

A Transformer-based Diffusion Model Processor for Real-time and High-quality Image Generation



隊伍名稱｜在火星上游泳 An Astronaut Swimming in the Ocean on Mars

隊　　長｜盧柏彥 / 清華大學電機工程研究所

隊　　員｜陳柏瑋 / 清華大學電機工程研究所

隊　　員｜葉紹華 / 清華大學電機工程研究所

隊　　員｜王庭佑 / 清華大學電機工程研究所

指導教授 ADVISOR



黃朝宗｜清華大學電機工程學系

臺灣大學電子工程博士，現為清華大學電機工程學系教授。曾服務於聯詠科技，隨後於麻省理工學院進行博士後研究。曾獲得傑出人才基金會年輕學者創新獎、未來科技獎、中國電機工程學會優秀青年電機工程師獎、清華大學傑出教學獎、旺宏金矽獎最佳指導教授獎等獎項。

研究領域

近來研究以實現高效能、高品質之電腦視覺與計算攝影學應用為主，包含影像生成式AI加速器、卷積神經網路處理器、立體3D光場顯示器等相關研究，是國內極少數能同時發表頂尖論文至計算機架構(ISCA/MICRO)、晶片設計(ISSCC/VLSIC/JSSC)、電腦視覺(CVPR/ICCV/TPAMI)此三大熱門研究領域之研究者。

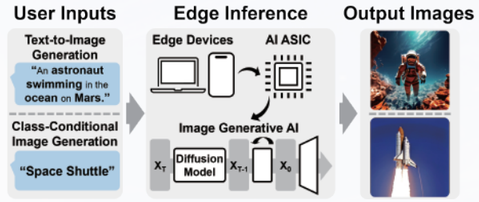
作品摘要

近年來，影像生成式人工智慧已取得前所未見的突破。使用者僅需輸入文字描述或指定影像類別，即可透過AI運算生成接近專業水準的高品質影像。其中，擴散模型憑藉其優異的生成品質，已成為影像生成式人工智慧的主流方法。然而，擴散模型龐大的參數量與高記憶體需求，使目前多數影像生成應用仍仰賴雲端GPU伺服器進行推論。隨著使用者數量快速成長，以及對資料隱私的關注日益提升，如何將大型擴散模型部署於智慧型手機與個人筆電等邊緣裝置，已成為未來AI落地的重要研究方向。

因此，本作品設計了一款專為邊緣AI推論最佳化的Transformer擴散模型處理器。此處理器針對新型Transformer擴散模型進行硬體架構最佳化設計，並支援可顯著提升影像生成品質的無分類器引導推論技術。為解決推論擴散模型所需的龐大頻寬與記憶體需求，我們透過硬體架構設計與模型壓縮技術，有效降低推論所需的記憶體頻寬。此外，本作品採用高效率的Microscaling資料格式，並進一步優化資料路徑以克服記憶體瓶頸。透過此優化設計，本處理器可在大幅降低頻寬與記憶體需求下，維持高品質的影像生成，特別適合部署於資源受限的邊緣裝置。

為驗證本處理器的實際成果，我們以16 nm FinFET先進製程完成晶片下線，並結合自行設計之PCB與Xilinx ZCU102 SoC開發板建構完整的即時展示系統。本處理器最高可提供7.37 TOPS的運算能力，並達到4.99-11.64 TOPS/W的能源效率，可在1.47-3.35秒內生成一張高品質影像。使用者可透過我們設計的介面輸入文字描述與生成參數，即時體驗高品質影像生成效果。

透過本作品所提出的處理器架構與展示系統，我們驗證了在邊緣裝置上實現即時影像生成的應用潛力，期望能進一步為生成式人工智慧的實際部署與普及提供具體解決方案。



圖一 部屬擴散模型於邊緣裝置之使用情境。

Abstract

In recent years, image generative AI has achieved unprecedented breakthroughs. Users can generate high-quality images simply by providing text prompts or class labels. Among existing approaches, diffusion models have become the dominant method for image generation due to their superior generation quality. However, current image generation still relies on cloud-based GPU servers because of the large model size and substantial memory requirements of diffusion models. With the rapid growth of the user base and increasing concerns over data privacy, deploying large diffusion models on edge devices such as smartphones and personal laptops has become an important research direction for future AI deployment.

This work presents a Transformer-based diffusion model processor optimized for edge AI inference. The proposed processor is specifically optimized for emerging Transformer-based diffusion models and supports classifier-free guidance inference, a technique that significantly improves image generation quality. To address the substantial bandwidth and memory demands of diffusion model inference, we propose both hardware optimization and model compression techniques to effectively reduce memory bandwidth requirements. In addition, this work adopts efficient microscaling data format and further optimizes the datapath to overcome on-chip memory bottlenecks. With these optimizations, the processor can maintain high-quality image generation while significantly reducing bandwidth and memory requirements, making it particularly suitable for deployment in resource-constrained edge devices.

Fabricated in 16 nm FinFET technology, the processor achieves up to 7.37 TOPS of computational throughput and an energy efficiency of 4.99-11.64 TOPS/W, enabling high-quality image generation within 1.47-3.35 seconds. To further validate the proposed architecture, we built a real-time demonstration system by integrating a custom-designed PCB with a Xilinx ZCU102 SoC. Users can connect to the system through a custom interface and experience real-time image generation by inputting text prompts and diffusion parameters. Through the proposed processor and demonstration system, this work demonstrates the feasibility of real-time image generation on edge devices, providing a promising solution for the practical deployment of image generative AI.

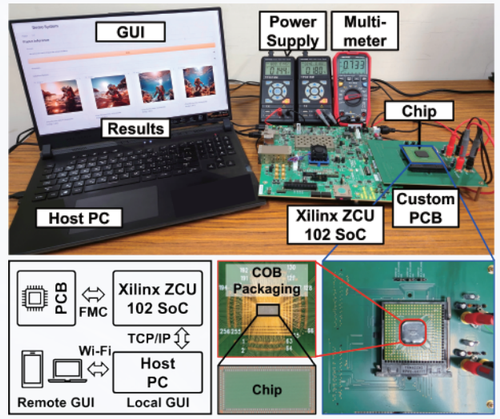


Fig. 2 Real-time demonstration system.