

# 先端科学技術研究科 修士論文要旨

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                        |     |             |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-----|-------------|
| 所属研究室<br>(主指導教員)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | コンピューティング・アーキテクチャ<br>(中島 康彦 (教授))                                                                      |     |             |
| 学籍番号                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 2511014                                                                                                | 提出日 | 令和 8年 6月 8日 |
| 学生氏名                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 安藤 拓翔                                                                                                  |     |             |
| 論文題目                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | AI Application Implementation and Software-Stack Improvements for Energy-Efficient Workloads on a CGLA |     |             |
| 要旨                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                        |     |             |
| <p>The increasing use of generative AI, speech recognition, and image generation has made inference-time computation and power consumption a major system concern. General-purpose graphics processing units (GPGPUs) are the dominant execution platform for these workloads because they provide high throughput, but their power consumption is substantial. Fixed-function accelerators can reduce energy for a narrow target, but they are difficult to adapt as model structures and quantization formats change. This motivates accelerator software stacks that preserve application-level flexibility while improving energy efficiency.</p> <p>This thesis implements execution paths for LLM inference, automatic speech recognition, and image generation on IMAX, a CPU-Grounded Linear Array (CGLA) implementation, and evaluates software-stack improvements that expose hardware efficiency at the application level. The scope is not limited to arithmetic kernels. It includes the offload boundary, data-transfer policy, and host-side control needed to preserve the execution flow of existing open-source software (OSS) applications.</p> <p>The LLM evaluation shows that decode execution on CGLA is strongly affected by LOAD, DRAIN, and host-side overhead.</p> <p>Under a 28 nm ASIC projection, the evaluated IMAX-based design improves PDP by up to 44.4x and EDP by up to 11.5x over an RTX 4090 for representative workloads. At the software-stack level, dynamic chunking improves prompt-processing throughput by 1.62x, and DRAIN-path restructuring increases the transfer rate from 0.264 Gbps to 1.077 Gbps for a representative Qwen3-0.6B Q8_0 case.</p> <p>The implementation targets llama.cpp for LLM inference and applies the same implementation policy to speech recognition and image generation, which share GGML-based execution substrates. In the whisper.cpp and stable-diffusion.cpp paths, F16 matrix operations are included in the IMAX execution targets in addition to quantized matrix operations, enabling evaluation of CGLA execution for applications with different data types and runtime structures.</p> |                                                                                                        |     |             |