**GANAX: A Unified MIMD-SIMD Acceleration for Generative Adversarial Networks**

**Authors**

1. **Amir Yazdanbakhsh** Georgia Institute of Technology
2. **Kambiz Samadi**  Qualcomm Technologies
3. **Nam Sung Kim**  University of Illinois at Urbana-Champaign
4. **Hadi Esmaeizadeh** UC San Diego

**Keywords**

1. Generative adversarial network (GAN)
2. Inference acceleration
3. Transposed convolution operator
4. Dataflow design
5. SIMD-MIMD
6. Access-Execute Architecture

**Summary**

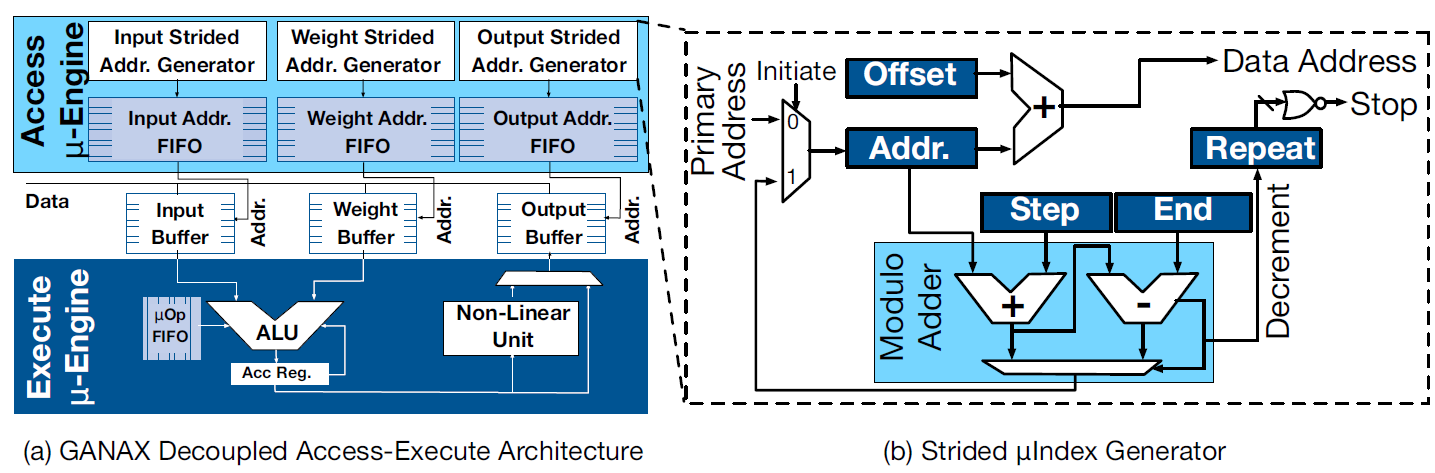
*Challenge*

GANs are one of the most recent deep learning models that generate synthetic data from limited genuine datasets. Although GANs are gaining prominence in various fields, there are no accelerators for these new models. In fact, GANs leverage a new operator, called transposed convolution, that exposes unique challenges for hardware acceleration. Even although there is a convolution stage in this operator, the inserted zeros lead to underutilization of the compute resources when a conventional convolution accelerator is employed.

*Contributions*

This paper proposes the GANAX architecture to alleviate the sources of inefficiency associated with the acceleration of GANs using conventional convolution accelerators, making the first GAN accelerator design possible. The main contributions of this paper are as follows:

1. Skipping over the zeros is necessary because performing convolution operations on zeros constitute more than 60% of all the multiply-add operations but is inconsequential. This paper proposes a reorganization of the output computations that allocates computing rows with similar patterns of zeros to adjacent processing engines, which will reclaims data use.
2. Reorganizing the output computations will break the SIMD execution model. This paper proposes a unified MIMD-SIMD accelerator architecture that exploits repeated patterns in the computations to create different microprograms that can execute concurrently in SIMD mode. The architecture, called GANAX, supports interleaving MIMD and SIMD operations at the finest granularity of a single microprogrammed operation.
3. MIMD’s overhead needs to be amortized. The data processing part can be SIMD but the irregular data access patterns prevent using this execution model. This paper proposes the decoupling of data accesses from data processing, which leads to breaking each processing engine into an access micro-engine and an execute micro-engine. The proposed architecture extends the concept of access-execute architecture to the finest granularity of computation for each individual operation.



*Experiments and Results*

This paper uses several state-of-the-art GANs to evaluate the GANAX architecture, including 3D-GAN, ArtGAN and so on. This paper implement GANAX microarchitecture units in Verilog, and then uses TSMC 45nm standard-cell library and Synopsys Design Compiler to synthesize these units and obtain the area, delay and energy numbers. This paper uses CACTI-P to measure the area and read/write access energy of the register files, SRAMs, and local/global buffers. This paper chooses Eyeriss as baseline and develops GANAX based on it with similar hardware configurations. Finally, evaluations with six GAN models shows, on average, 3.6x speedup and 3.1 energy savings over Eyeriss without compromising the efficiency of conventional convolution accelerators. These benefits come with a mere ~7.8% area increase.

*Comments*

The highlight of this paper is that it breaks conventional SIMD execution model to accelerate computing and introduce a SIMD-MIMD execution model to avoid underutilization of computing sources based on existing CNN accelerator architecture. What impressed me most is that it changes PE into microprogram executer and decouples access engine and execute engine, which provides great flexibility. But this design style will also bring complexity into architecture design, dataflow control and simulation.