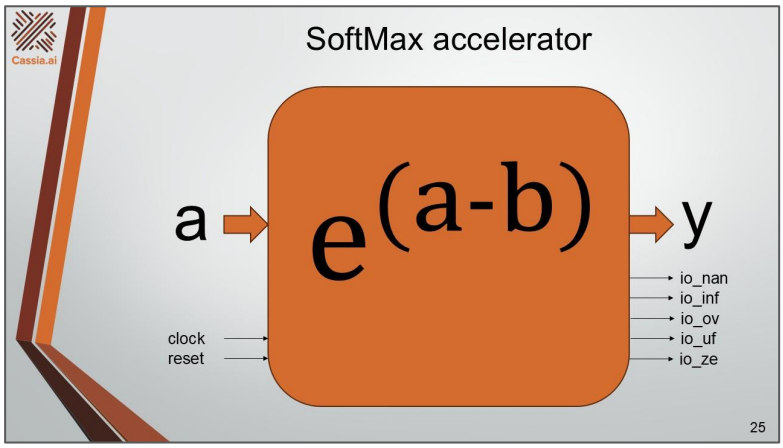
Cassia.ai

SoftMax Accelerator

SoftMax Accelerator e^(a-b) is a key primitive of a custom Softmax unit.

Overview

This highly accelerated and low-gate-count mathematical function is perfect for building or accelerating custom AI accelerators



Key Features

- y=e^(a-b)
- Cassia.ai SoftMax functions are incredibly fast and low power, and may be piplined as needed. Our SoftMax functions are built to order to meet each customer's exact requirement, and use this function as a key primitive.
- Uses advanced mathematics to improve performance, power and area
- Single-cycle or pipelined option

Benefits

- May be accelerated 10X-20X depending on precision and system requirements
- Significant gate count reduction compared to traditional math
- Uses standard input and output formats - user definable to be IEEE Floating Point, Brain Float, Tensor Float, or custom implementations

Applications

- Use in AI accelerators for improved performance, power and area
- All RTL implementation with no custom circuits - portable to any foundry or FPGA
- Cassia.ai functions are designed to replace your existing mathematical functions, allowing you to use pre-trained models without additional training, just instantiate and recompile verilog. They also work for training workloads!

Available Standard Precisions

FP16(5E10M), BF16(8E7M), TF32(8E10M), FP19(8E10M), FP32(8E23M), FP64(11E52M)

Precisions with extra lead time

MXFP4(2E1M), MXFP8(4E3M), MXFP8(5E2M), FP4(2E1M), FP8(4E3M), FP8(5E2M), FP6(3E2M), MXFP6(3E2M), MXFP6(2E3M), FP6(2E3M), NVFP4(2E1M), MXINT8, Others by request

Maturity: Pre-Silicon

Deliverables

- Verilog RTL
- C++ or CUDA on request
- PyTorch and Docker containers
- Full documentation
- Licenses to Cassia.ai patents

Email Contact: info@cassia.ai