

SoftEx: a Low Power and Flexible Softmax Accelerator with Fast Approximate Exponentiation

Andrea Belano, Yvan Tortorella, Angelo Garofalo,
Luca Benini, Davide Rossi, Francesco Conti



PULP Platform

Open Source Hardware, the way it should be!

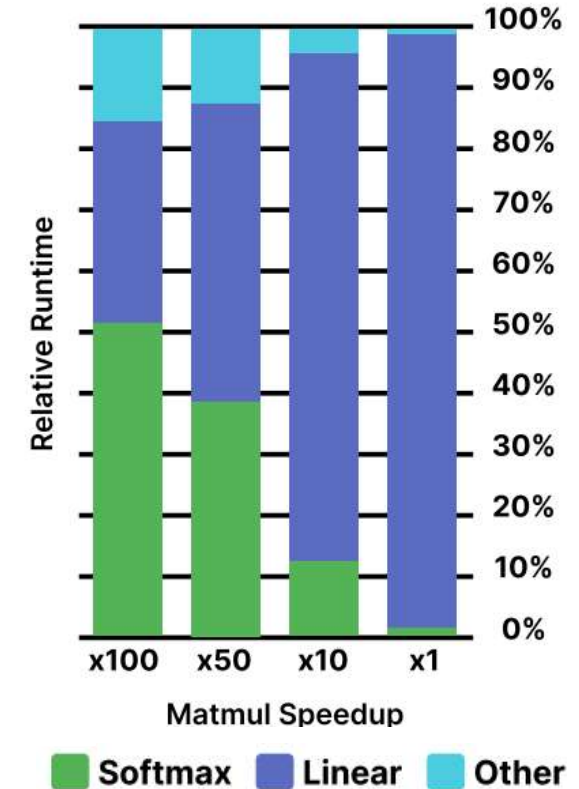
@pulp_platform 

pulp-platform.org 

youtube.com/pulp_platform 

Motivation - Transformers and Softmax

- **Transformers are the main models driving the evolution of modern Artificial Intelligence**
 - Both in **perceptive** task and in **generative** applications
 - However, each layer of a Transformer is complex, featuring **multi-head self-attention** and additional projections
- **Attention consists of multiple GEMMs + softmaxs**
 - Unlike CNNs, **softmax is applied multiple times** every layer
- **Softmax becomes a bottleneck when matmuls are accelerated**
 - It is fundamental to also **accelerate softmax** if we target Transformer-based models
- **What's the deal with this function?**
 - It is based on the **exponential** function
 - It is **NOT** a **point-to-point** function

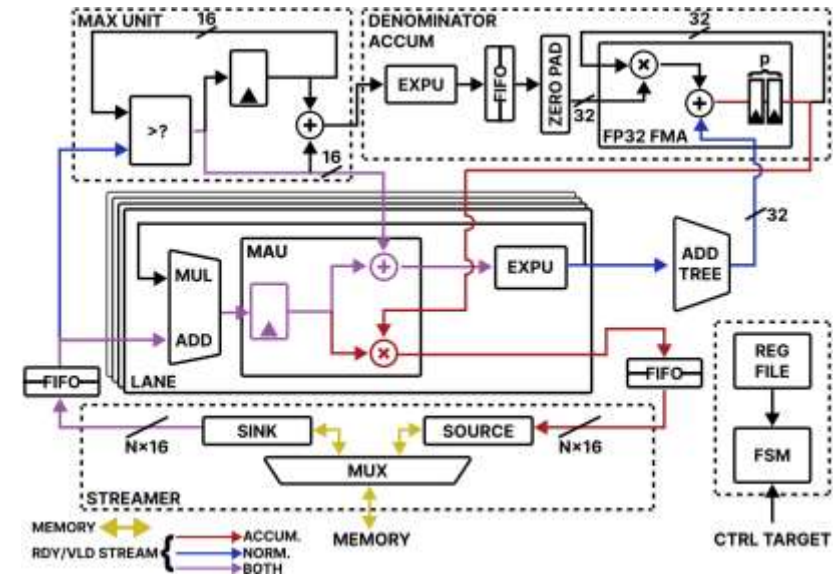
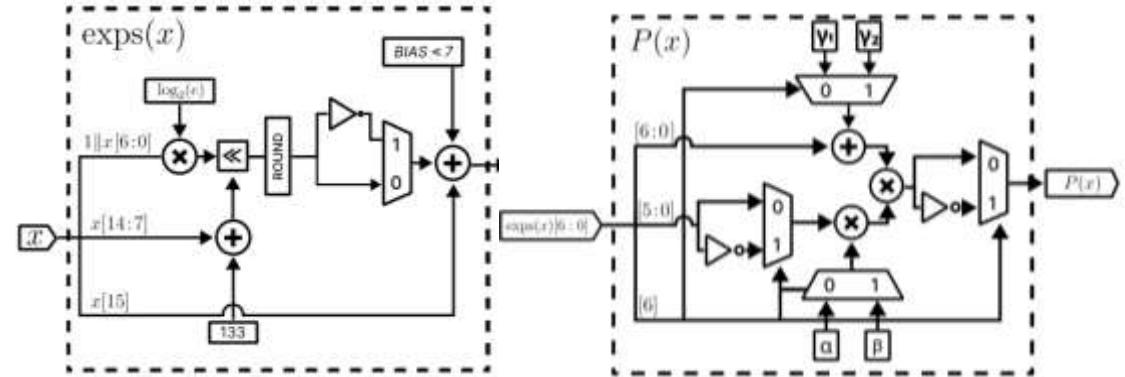


$$\text{Softmax}(x_i) = \frac{e^{x_i - x_{\max}}}{\sum e^{x_j - x_{\max}}}$$

What We Propose



- **A novel hardware-friendly exponential function approximation based on Schraudolph's method**
 - Use Schraudolph's method to compute a **fast approximation** of the exponential ($\text{exps}(x)$)
 - **Correct the mantissa** using a piecewise polynomial ($P(x)$)
 - **Minimal accuracy loss** (0.14% mean relative error) compared to accurate methods
- **SoftEx, a parametric accelerator for BF16 softmax**
 - Up to **10.8× faster softmax** compared to 8 RISC-V cores
 - Up to **26.8× lower energy consumption** compared to 8 RISC-V cores



See you at the poster!

Institut für Integrierte Systeme – ETH Zürich

Gloriastrasse 35
Zürich, Switzerland

DEI – Università di Bologna

Viale del Risorgimento 2
Bologna, Italy