

# Neural Networks as Function Primitives

Software/Hardware Support with X-FILES/DANA

Schuyler Eldridge<sup>1</sup> Tommy Unger<sup>2</sup> Marcia Sahaya Louis<sup>1</sup>  
Amos Waterland<sup>3</sup> Margo Seltzer<sup>3</sup>  
Jonathan Appavoo<sup>2</sup> Ajay Joshi<sup>1</sup>

<sup>1</sup>Boston University Department of Electrical and Computer Engineering

<sup>2</sup>Boston University Department of Computer Science

<sup>3</sup>Harvard University School of Engineering and Applied Sciences

Boston Area Architecture Workshop '16



# Neural Networks as Function Primitives

## Motivation

- Neural networks and machine learning are everywhere (*again*)
  - Broad use in high tech and big data, e.g., Google's Tensorflow [1]
  - Enable automatic parallelization, e.g., ASC [3]
  - Provide a means for approximate computing, e.g., NPU [2]
- Our vision
  - Neural networks are a new *functional primitive* useful at various scales of computation [4]
- With that in mind, we've developed software and hardware for the use of accelerator-backed neural network computation

[1] Google TensorFlow, <https://github.com/tensorflow/tensorflow>

[2] H. Esmaeilzadeh, A. Sampson *et al.*, "Neural acceleration for general-purpose approximate programs," in *Proc. MICRO*, 2012.

[3] A. Waterland, E. Angelino, R. P. Adams, J. Appavoo, and M. Seltzer, "Asc: Automatically scalable computation," in *Proc. ASPLOS*, 2014.

[4] S. Eldridge, A. Waterland *et al.*, "Towards general-purpose neural network computing," in *Proc. PACT*, 2015.

# Our Contributions Towards this Vision

## X-FILES: Software/Hardware Extensions

Extensions for the Integration of Machine Learning in Everyday Systems

- A defined user and supervisor interface for neural networks
- This includes supervisor architectural state (hardware)

## DANA: An Example Multi-Transaction Accelerator

Dynamically Allocated Neural Network Accelerator

- An accelerator aligning with our multi transaction vision

## Neural Network Transactions

*A transaction encapsulates a request by a process to compute the output of a specific neural network for a provided input*

# Or: A Drop in Accelerator for a RISC-V Microprocessor

## What does that mean?

- 1 Grab a Rocket Chip RISC-V Microprocessor [1]
- 2 Build a RISC-V toolchain
- 3 Grab a copy of our X-FILES/DANA accelerator [2]
  - Implemented in Chisel [3]
- 4 Build an FPGA configuration for Rocket + X-FILES/DANA
- 5 User processes can safely throw *transactions* at X-FILES hardware
  - With support for feedforward and learning computation

[1] Rocket Chip git repository, UC Berkeley, Online: [github.com/ucb-bar/rocket-chip](https://github.com/ucb-bar/rocket-chip)

[2] X-FILES/DANA git repository, Boston University, Online (*soon!*): [github.com/bu-icsg/xfiles-dana](https://github.com/bu-icsg/xfiles-dana)

[3] J. Bachrach, H. Vo, B. Richards, Y. Lee, A. Waterman, R. Avižienis *et al.*, "Chisel: Constructing hardware in a scala embedded language," in *Proc. DAC*, 2012, pp. 1216–1225.

# X-FILES Software Components

## Supervisor API

- Establishes sets of processes that can access neural network hardware
- Defines the neural networks that processes are allowed to access
- *More details on the poster!*

## User API

- Works at the level of *transactions*
  - A complete request for access to neural network resources, communication of inputs, processing, and communication of outputs
- Initiating a new transaction
- Writing data
- Reading data

# X-FILES/DANA Hardware

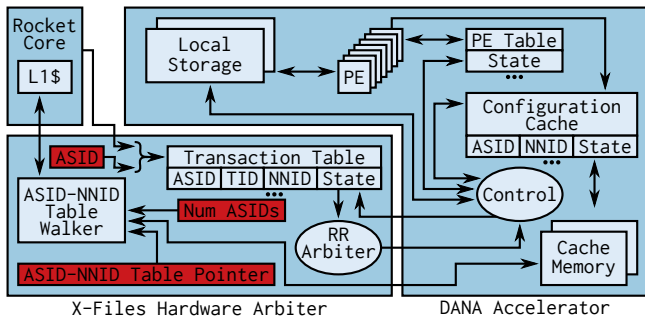


Figure: X-FILES/DANA hardware architecture

## Components

- X-FILES Hardware Arbiter maintaining transaction state
- DANA to move transactions towards completion
  - With support for feedforward or learning computation

## Remaining Items

- Linux kernel integration
- Support for asynchronous data transfer

## Open Source Availability

- Should be ready by the end of February
- On GitHub: [github.com/bu-icsg/xfiles-dana](https://github.com/bu-icsg/xfiles-dana)

# Acknowledgments

This work was supported by the following:

- A NASA Space Technology Research Fellowship
- An NSF Graduate Research Fellowship
- NSF CAREER awards
- A Google Faculty Research Award