NTX: A 260 Gflop/sW Streaming Accelerator for Oblivious Floating-Point Algorithms in 22nm FD-SOI

Fabian Schuiki[1], Michael Schaffner[1], Luca Benini[1,2]
[1]**Integrated Systems Laboratory, ETH Zurich;** [2]DEI, University of Bologna

1 Introduction

In this work we present the first complete design, silicon implementation and measurements in 22nm FD-SOI of the Network Training Accelerator (NTX) architectural concept. NTX is based on a newly designed partial carry- save "wide-inside" (300bit) fused multiply-accumulate (FMAC) unit ensuring IEEE754 compliance and a Root Mean Squared Error 1.7× lower than a conventional 32bit FPU on long accumulations such as convolutions. The fully pipelined FMAC unit is designed for no-stall throughput on all common vector and matrix operations. Additional non-MAC operations support the full back- propagation step of DNN training.

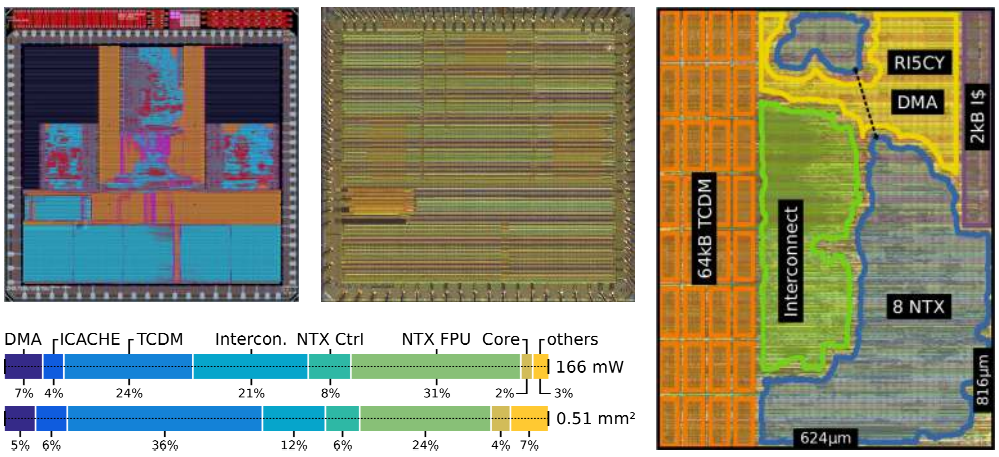

Figure 1. Top left and middle: Full chip layout and die shot of manufactured silicon in Globalfoundries 22FDX technology. Right: Floorplan of the NTX accelerator cluster, consisting of 1 RISC-V processor and 8 NTX floating-point accelerators. Bottom left: Area and power consumption breakdown of the NTX cluster. A significant fraction of power is consumed by the FPU and memory subsystems, contributing to the high energy efficiency of the system.

	NTX [us]	PULP [2]	Cortex A53 [3]	Rocket 64b [4]	Tesla V100§	Xeon 8180§
Node/V_{DD}	22/0.45	40/0.8	16/0.8	40/0.65	12/1.0	14/0.9
Energy Eff.†	260	18	38.7	16.7	122	21.9
Area Eff.‡	47.1	7.35	8.7	14.5	20.5	3.57

† Gflop/s W; ‡ Gflop/s mm² (node-scaled); § estimated

Table 1. Key metric comparison between NTX and other processors. Performance for 32 bit floating point operations. Tesla V100 and Xeon 8180 based on our own estimates given available information and die shots. Energy efficiency as-is, area efficiency normalized over node.

2 Implementation

Figure 1 shows the layout of the main blocks in the design. The design was closed at 1.25 GHz (SSG, 0.72 V, −40/125 ◦C), covering 0.51mm2 at 59% density. Three address generator units (AGUs) and five nested hardware loops handle the most common address patterns found in oblivious algorithms. Eight NTX units are managed by one RV32IMC RISC-V core that enables full software flexibility. A key feature is the focus on removing the von Neumann bottle- neck by amortizing RISC-V instructions over eight NTX units and removing all FP loads/stores from the instruction stream via the AGUs. The NTX units and RISC- V core attach to a 64 kB high-bandwidth Tightly Coupled Data Memory TCDM, comparable to private memory and register files on recent GPUs (e.g. 48kB in the V100). A DMA engine handles data transfers at 5 GB/s in the background. Double buffering allows for latency hiding and full unit utilization on a wide range of algorithms, which is more energy efficient than the increased instruction stream pressure of heavy multi-threading. The RISC-V processor merely configures and orchestrates NTXs and DMA operation and does not directly handle data transfers, keeping the computation and data movement out of its pipeline. It can thus “effectively issue” 32 flops, 16 local, and 4 global 32bit memory accesses per cycle, without the area and energy cost of an advanced superscalar or vector architecture. NTX even achieves significantly lower instruction memory bandwidth than a SIMT GPU streaming processor (up to 4×) thanks to the removal of explicit loads/stores.

3 Results

Figure 2 shows the performance and energy efficiency measured on silicon performing matrix multiplications. SRAMs remain functional down to 0.55V, standard logic down to 0.425V (SRAMs kept at 0.55V). The cluster operates between 260Gflop/sW at the high-efficiency and 24Gflop/s at the high-performance end of the spectrum. Leakage power is between 5.5% and 24% over 0.45V to 1.0V. PVT variations can be compensated by up to 0.8V of body biasing, which increases performance up to 1.6×.

Figure 3 shows the performance achieved on a wide range of oblivious kernels. Energy and area efficiency are highly competitive due to a reduced von Neumann bottleneck thanks to highly autonomous NTX and DMA, allowing us to dedicate 60% of the silicon area and 55% of the power to data memory and FPUs with a very significant 24% and 31% in the FPUs (see Figure 2). The RISC-V processor never bottlenecks the operation.

Table 1 shows a comparison to other systems. NTX outperforms a Tesla V100 by 2.1× in energy efficiency and 2.3× in area efficiency. Compared to an ARM Cortex- A53, a 64bit Rocket RISC-V core, and a PULP cluster, NTX achieves 6.7×, 15.6×, and 14.4× better energy efficiency, respectively; and 5.4×, 3.2×, and 6.4× better area efficiency. Compared to a 28-core dual-AVX-512 Intel Xeon 8180 CPU, NTX reaches a 11.9× and 13.2× higher energy and area efficiency.

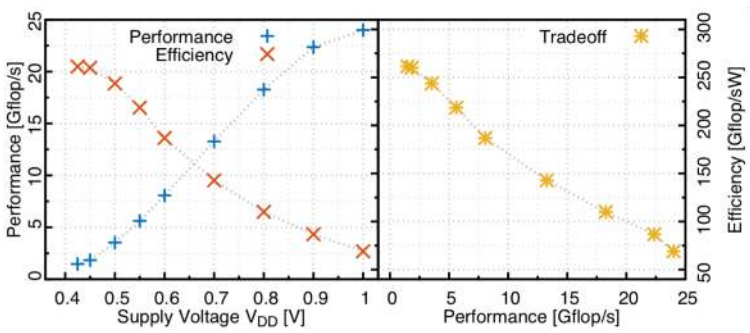

Figure 2. Measured compute perf. and energy efficiency versus supply voltage (left). Measured performance/efficiency tradeoff (right).

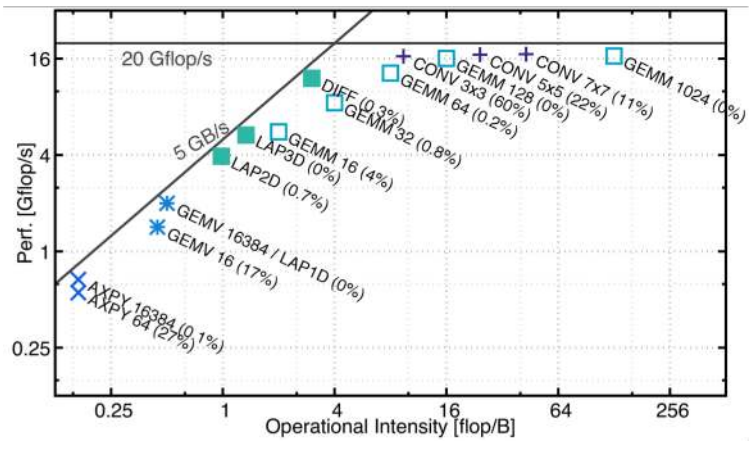

Figure 2. Power and area breakdown, and roofline model of one NTX cluster. Percentages indicate processor utilization.

4 Conclusion

NTX in the 22FDX technology is a highly competitive architecture for FP-intensive computing tasks, reaching up to 24 Gflop/s and 260 Gflop/sW. In common kernels NTX reduces instruction bandwidth by 512× over single-FMA and 64× over SIMD data paths. Its energy and area efficiency outperforms contemporary RISC and CISC processors by up to 15.6× and 6.4×, and even data-center- class GPUs by 2.1× and 2.3×, respectively.

[1] Schuiki, Fabian, et al. "A scalable near-memory architecture for training deep neural networks on large in-memory datasets." IEEE Transactions on Computers 68.4 (2018): 484-497.
[2] Schuiki, Fabian, Michael Schaffner, and Luca Benini. "NTX: An Energy-efficient Streaming Accelerator for Floating-point Generalized Reduction Workloads in 22 nm FD-SOI." 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2019.