ETH zürich

Kosmodrom: Energy Efficient Ariane Cores with Transprecision FPU in 22nm

Fabian Schuiki[1], Florian Zaruba[1], Stefan Mach[1], Luca Benini[1]
[1]**Integrated Systems Lab, ETH Zurich**

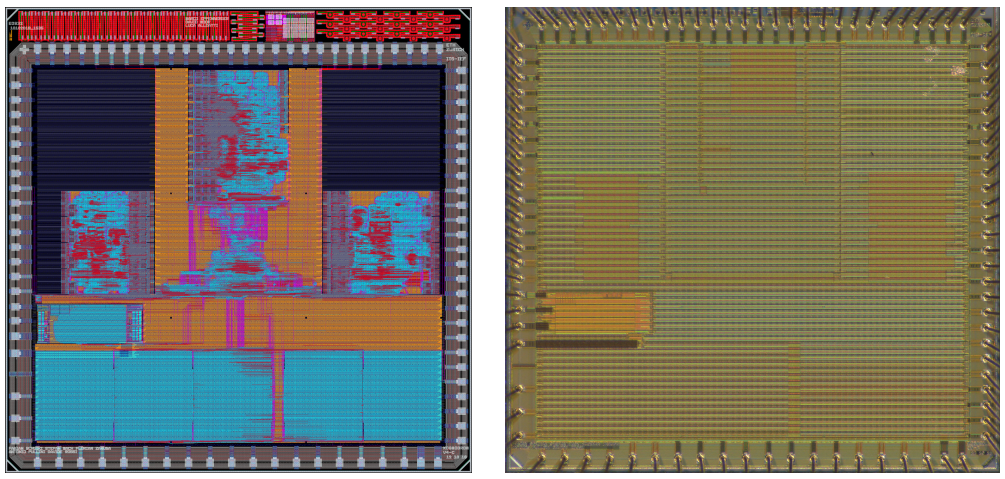

Fig. 1. Kosmodrom GDSII die layout (left) and photograph of manufactured silicon (right). The system contains two 64 bit Ariane processors (top left and right of die), one 32 bit NTX floating point accelerator cluster (top center of die), and 1.25 MB of memory (bottom of die).

1 Introduction

We present Kosmodrom, a comparison platform for two energy efficient implementations of Ariane and NTX. Ariane is a RV64GC core augmented with a transprecision FPU, allowing for interesting energy/precision trade offs in many applications. NTX is a streaming floating-point co-processor targeted towards stencil and linear algebra applications, as well as machine learning. In this talk we present lessons learned from implementing Kosmodrom in Globalfoundries 22FDX technology and present measurements and evaluation of the resulting silicon. We show the energy efficiency and performance trade offs involved in using different standard cell libraries and other tuning techniques during the design flow.

2 Ariane Cores

Kosmodrom is the first silicon implementation of a 64 bit transprecision floating-point unit. It fully supports the standard double, single, and half precision, alongside custom bfloat and 8 bit formats. Operations occur on scalars or 2, 4, or 8-way SIMD vectors. We have integrated the 247 kGE floating-point unit into a 64 bit application-class RISC-V processor core, where the added transprecision support accounts for an energy and area overhead of merely 11% and 9%, respectively; yet achieving speedups and per-datum energy gains of 7.3x and 7.9x. We measured the silicon manufactured in Globalfoundries 22FDX technology across a wide supply voltage range from 0.45V to 1.2V. The unit achieves energy efficiencies between 178 Gflop/sW and 2.95 Tflop/sW, and a performance between 3.2 Gflop/s and 25.3 Gflop/s, across formats.

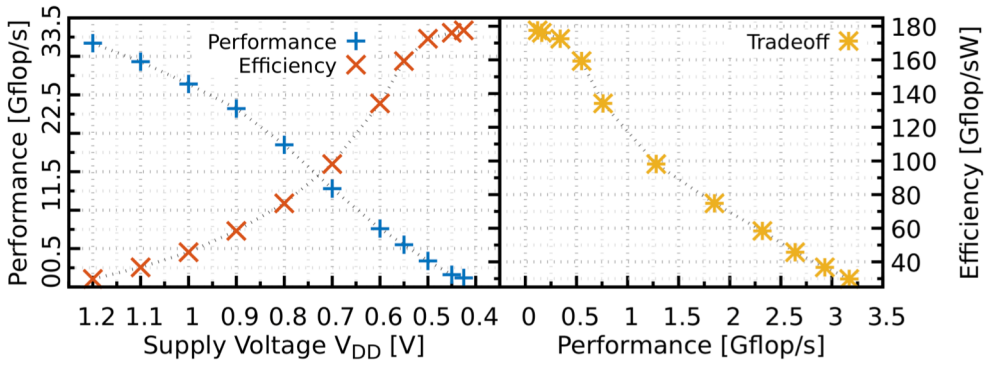

Fig. 2. Compute performance and energy efficiency of FP64 FMA on Ariane 64 bit core versus supply voltage (left), trade-off between compute performance and energy efficiency, achieved by adjusting supply voltage and operating frequency. Measured on manufactured silicon.

3 NTX Floating-Point Accelerator

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

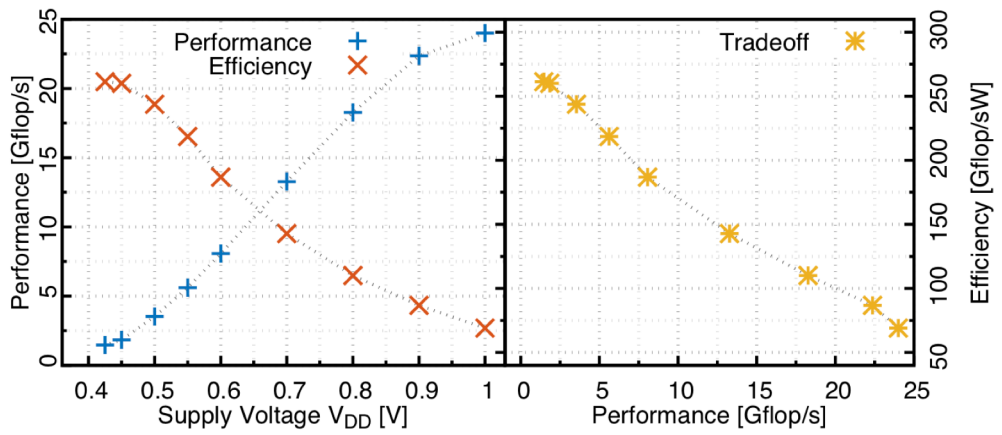

Fig. 3. Measured compute performance and energy efficiency of the NTX cluster versus supply voltage (left). Measured performance / efficiency tradeoff (right). Results measured on actual silicon.

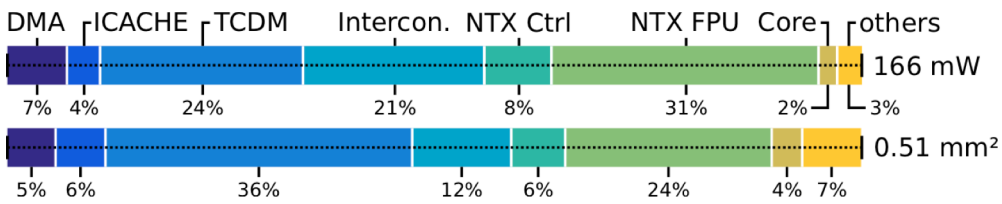

Fig. 4. Power and area breakdown of the NTX cluster as manufactured on the final silicon.

KEY METRIC COMPARISON BETWEEN NTX AND OTHER PROCESSORS. PERFORMANCE FOR 32 BIT FLOPS.

	NTX [us]	PULP	Cortex A53	Rocket 64b	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

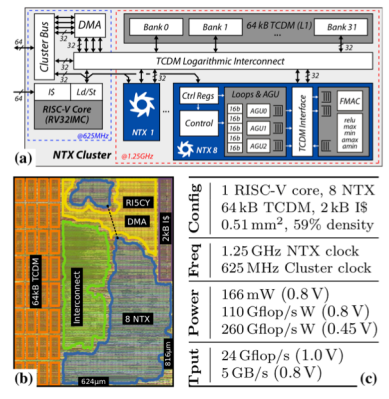

Fig. 5. Energy and area efficiency comparison with related processors and computing systems (top). NTX schematic diagram (a), floorplan (b), and main characteristics (c). (right)

4 Conclusion

The Kosmodrom chip provides floating-point compute capabilities in three different flavors, offering up to 0.8 pJ/flop or 2.95 Tflop/sW in the transprecision domain, and up to 260 Gflop/sW or 24 Gflop/s on standard 32 bit floating point numbers.

5 References

[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.