ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7µW Cognitive Wake-Up from MRAM-Based State-Retentive Sleep Mode

This is the final peer-reviewed author’s accepted manuscript (postprint) of the following publication:

Published Version:
4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7µW Cognitive Wake-Up from MRAM-Based State-Retentive Sleep Mode / Rossi D.; Conti F.; Eggiman M.; Mach S.; Mauro A.D.; Guermandi M.; Tagliavini G.; Pullini A.; Loi I.; Chen J.; Flamand E.; Benini L.. - STAMPA. - 64:2(2021), pp. 9365939.60-9365939.62. (Intervento presentato al convegno 2021 IEEE International Solid-State Circuits Conference, ISSCC 2021 tenutosi a usa nel 2021) [10.1109/ISSCC42613.2021.9365939].

Availability:
This version is available at: https://hdl.handle.net/11585/826461 since: 2021-06-21

Published:
DOI: http://doi.org/10.1109/ISSCC42613.2021.9365939

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

18 April 2024

This is the final peer-reviewed accepted manuscript of:

D. Rossi *et al.*, 4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7µW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode

in:

2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021

The final published version is available online at:
https://doi.org/10.1109/ISSCC42613.2021.9365939

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode

Davide Rossi[1], Francesco Conti[1], Manuel Eggiman[2], Stefan Mach[2], Alfio Di Mauro[2], Marco Guermandi[1,3], Giuseppe Tagliavini[1], Antonio Pullini[2,3], Igor Loi[3], Jie Chen[1,3], Eric Flamand[2,3], Luca Benini[1,2]

[1]University of Bologna, Bologna, Italy, [2]ETH Zurich, Zurich, Switzerland, [3]Greenwaves Technologies, Grenoble, France

The Internet-of-Things requires end-nodes with ultra-low-power always-on capability for long battery lifetime, as well as high performance, energy efficiency and extreme flexibility to deal with complex and fast-evolving near-sensor analytics algorithms (NSAAs). We present *Vega*, an always-on IoT end-node SoC capable to scale from a 1.7μW fully retentive *cognitive* sleep mode up to 32.2GOPS (@49.4mW) peak performance on NSAAs, including mobile DNN inference, exploiting 1.6MB of state-retentive SRAM, and 4MB of non-volatile MRAM. To meet the performance and flexibility requirements of NSAAs, the SoC features 10 RISC-V cores: one core for SoC and IO management and a 9-cores cluster supporting multi-precision SIMD integer and floating-point computation. Two programmable machine-learning (ML) accelerators boost energy efficiency in sleep and active state, respectively.

As shown in Fig. 1, the SoC consists of two main switchable power domains (SoC and Cluster), plus an always-on domain operating from 0.6V to 0.8V supplied by two commercial off-the-shelf on-chip DCDC plus an LDOs from 3.6V (VBAT). Two body-bias generators are included for process compensation of SoC and Cluster domains. A Power Management Unit (PMU), clocked by a 1MHz internal ring oscillator, manages transitions between the power states of the SoC. The wake-up sources are an external pad, an RTC, and the power-gateable Cognitive Wake-up Unit (CWU) implemented with UHVT standard cells. The CWU (Fig. 2) enables autonomous ML-based classification of sensor data patterns while the SoC remains in deep sleep. The CWU interfaces with external sensors using 4 SPI interfaces; sensor data is classified by a novel hyper-dimensional computing (HDC) [1] accelerator that supports binary HD-vectors of 2048 bits stored in a 4kB associative latch-based Standard Cell Memory (SCM). The HDC encoding and classification algorithm is stored in a secondary user-programmable 200B microcode SCM that orchestrates the 512-bit wide datapath and, conditionally on the classification outcome, triggers the PMU to wake up the SoC. When operating from the 32kHz reference clock, the SWU consumes 1.69μW at 25°C (1μW dynamic, 690nW leakage) while classifying data in real-time from 3 SPI peripherals (16 bit, 150SPS/channel). With 31nW/kHz, the SWU logic's dynamic power is 20% lower than the SPI dynamic I/O power (40nW/kHz). To retain the SoC program and data when sleeping, the physical SRAM banks can selectively be configured in retentive mode, leading to retention power ranging from 1.2 to 112μW for 16kB to 1.6MB of state-retentive L2 SRAM. A 4MB non-volatile Magnetoresistive Random Access Memory (MRAM) resides in an independent switchable power domain. Warm boot can either be performed from L2 SRAM or from the MRAM; in the latter case, sleep power for data retention is zero, but the program must be restored into L2 after wake-up.

The SoC Domain is an advanced MCU featuring a RISCV processor named Fabric Controller (FC) and several peripherals, including a~~n~~ 1.6 Gbit/s DDR interface supporting external IoT DRAMs such as Cypress Semiconductor's HyperRAM (Fig. 1). The cluster, built around 9 70kGE 4-pipeline stage RISCV cores, is turned on and adjusted to the required frequency when applications running on the FC offload computation-intensive kernels. The cores share data on a 128kB shared multi-banked L1 memory through a 1-cycle latency logarithmic interconnect. The cluster L1 memory can serve 16 parallel memory requests with <10% contention rate even on data-intensive kernels, delivering up to 28.8GB/s at 450MHz. The program cache is hierarchical: 512B private per-core plus 4kB of 2-cycle latency shared cache to maximize efficiency with data-parallel code. The ~~RISCV~~ cores feature extensions (RVC32IMF-Xpulp) for NSAAs, such as hardware loops, post-incremented LD/ST, Single Instruction Multiple Data (SIMD) such as dot products operating on narrow 16- and 8-bit data types. The cores share 4 FP units, supporting FP32, FP16 and bfloat operations. Fine-grain parallel thread dispatching is accelerated by a dedicated hardware event unit, which also manages clock gating of idle cores waiting for synchronization and enables resuming execution in 2 cycles. ISA extensions coupled with parallelism and optimized memory hierarchy deliver a MAC/cycle performance 62x larger than a baseline RISCV ISA running on a single core, similar to [2]. The cluster delivers up to 15.6 8-bit GOPS and up to 614 GOPS/W, and up to 3.3 GFLOPS and 129 GFLOPS/W, on GP processors, demonstrating leading-edge performance on a wide range of NSAA (Fig. 3). Efficiency can be further increased exploiting a dedicated ML accelerator (HW computing element - HWCE) sharing L1 memory with the RISC-V cluster through four additional ports for streamlined zero-copy HW-SW cooperation. The HWCE has three multi-precision (4b/8b/16b) 3x3 convolution units with 27 MACs in total, as well as integrated normalization / activation, leading to 32.2 GOPS and 1.3TOPS/W. Fig. 4 shows performance and efficiency of multi-precision matrix multiplication running on General-Purpose (GP) processors and 8-bit DNN workloads running on the SoC and on the Cluster.

The proposed SoC features a three-level~~s~~ memory architecture: L1 and L2 are globally addressed, while MRAM and off-chip HyperRAM are mapped on private spaces. Memory is explicitly managed by means of two DMA engines (Cluster and I/O), removing area/power overheads of cache coherency. Full overlap of L3/L2, L2/L1 transfers with computation is achieved by means of tiling and double buffering, orchestrated by one of the nine cluster cores. The cluster DMA has 4 32-bit ports on the L1 interconnect, enough to saturate 7.2 GB/s@450 MHz read/write aggregate bandwidth of the 64-bit AXI4 cluster bus (Fig.1). The DMA features 2D transfers to ease data tiling. The SoC I/O DMA transfers data in a fully autonomous way between the I/O peripherals (including the HyperRAM) and the L2 memory. The MRAM resides in a dedicated clock domain (1 to 40 MHz), and is connected to the I/O DMA with dual-clock FIFOs (Fig. 1), with 320 MB/s of peak read bandwidth consuming 6.2 mW.

In Fig. 5, we show end-to-end fully on-chip inference of a MobileNetV2 (depth multiplier=1.0, input size=224x224) [5] with int8 weights and activations, achieving full-precision accuracy (71.8%). Full-network weights are stored in non-volatile MRAM, while intermediate in/out tensors are allocated in L2 and deallocated as soon as they are consumed by following layers. To use the cluster, data is split in working tiles so that for each tile double-buffered weights and in/out collectively fit the 128 KB L1. Computation is organized so that computation and data transfers are fully overlapped. While the current layer is run, next-layer weights are prefetched from MRAM to L2 with the I/O DMA, as the MRAM is not directly connected to the L1. At the same time, a single cluster core orchestrates the double-buffered copy of in/out/weight tiles from L2 to L1 using the Cluster DMA; the other 8 cores compute on available tiles. End-to-end inference of MobileNetV2 requires 84.7ms @ 250MHz and 1.19mJ per frame - 3.5x less energy than ~~when~~ using off-chip HyperRAM for weight storage.

Fig. 6 shows a comparison with SoA. With respect to fully programmable IoT end-nodes [2][3], the proposed SoC delivers more than 1.3x - 2x better performance and 3.2x - 4.3x better efficiency on NPAA workloads. With respect to most efficient hardware-accelerated IoT end-nodes [4], it performs with similar energy efficiency on DNN inference workloads at 5.5x better performance. On non-DNN, NSAA workloads, our SoC achieves 10x and 2.5x higher performance and energy-efficiency, respectively. The proposed SoC is the first in his class capable of fully on-chip execution of SoA mobile DNN (MobileNetV2). Figure 7 shows a die micrograph, highlighting system components included in the measurements.

Acknowledgments

This work was supported in part by the EU Horizon 2020 Research and Innovation projects OPRECOMP (Open trans-PREcision COMPuting, g.a. no. 732631) and WiPLASH (Wireless Plasticity for Heterogeneous Massive Computer Architectures, g.a. no. 863337) and by the ECSEL Horizon 2020 project AI4DI (Artificial Intelligence for Digital Industry, g.a. no. 826060).

References

[1] A. Rahimi et al. "High-Dimensional Computing as a Nanoscalable Paradigm," TCAS-I, Vol. 64, Issue 9, Sept. 2017.

[2] D. Bol et al., "A 40-to-80MHz Sub-4μW/MHz ULV Cortex-M0 MCU SoC in 28nm FDSOI with Dual-Loop Adaptive Back-Bias Generator for 20μs Wake-Up From Deep Fully Retentive Sleep Mode," ISSCC, pp. 322-324, 2019.

[3] A. Pullini et al., "Mr.Wolf: An Energy-Precision Scalable Parallel Ultra Low Power SoC for IoT Edge Processing," JSSC, vol. 54, no. 7, pp. 1970-1981, July 2019.

[4] I. Miro-Panades et al., "SamurAI: a 1.7MOPS-36GOPS Adaptive Versatile IoT Node with 15,000x Peak-to- Idle Power Reduction, 207ns Wake-up Time and 1.3TOPS/W ML Efficiency," VLSI Symp, 2020.

[5] M. Sandler et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," CVPR, 2018.

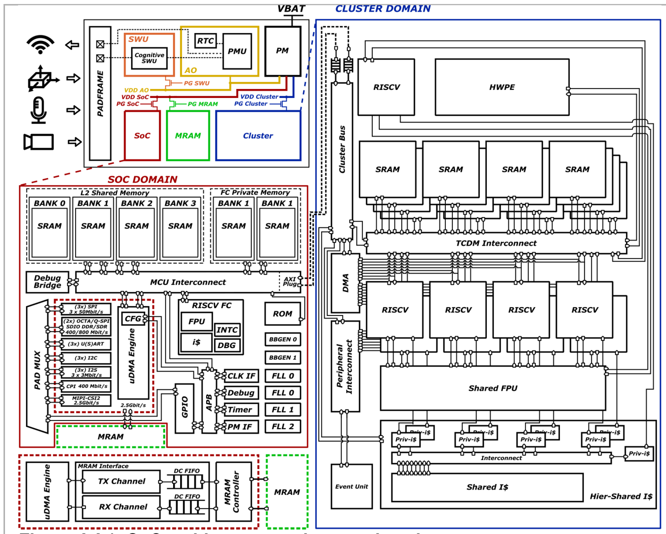

VBAT
CLUSTER DOMAIN
SWU
Cognitive SWU
RTC
PMU
PM
AO
PG SWU
VDD AO
VDD SoC
PG SoC
PG MRAM
VDD Cluster
PG Cluster
PADFRAME
SoC
MRAM
Cluster
SOC DOMAIN
L2 Shared Memory
FC Private Memory
BANK 0
BANK 1
BANK 2
BANK 3
SRAM
Debug Bridge
MCU Interconnect
AXI Plug
RISCV FC
FPU
INTC
I$
DBG
ROM
BBGEN 0
BBGEN 1
CLK IF
FLL 0
Debug
Timer
FLL 1
PM IF
FLL 2
PAD MUX
uDMA Engine
CFG
GPIO
APB
MRAM
MRAM Interface
TX Channel
RX Channel
DC FIFO
MRAM Controller
Cluster Bus
RISCV
HWPE
TCDM Interconnect
DMA
Peripheral Interconnect
Shared FPU
Priv-I$
Interconnect
Shared I$
Event Unit
Hier-Shared I$

Figure 4.4.1: SoC architecture, and power domains.

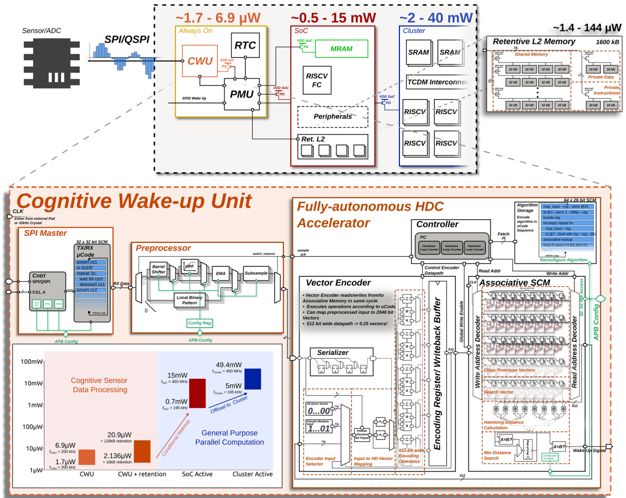

Sensor/ADC
SPI/QSPI
~1.7 - 6.9 µW
Always On
CWU
RTC
PMU
~0.5 - 15 mW
SoC
MRAM
RISCV FC
Peripherals
Ret. L2
~2 - 40 mW
Cluster
SRAM
TCDM Interconnect
RISCV
~1.4 - 144 µW
Retentive L2 Memory
1600 kB
Cognitive Wake-up Unit
SPI Master
Preprocessor
Fully-autonomous HDC Accelerator
Controller
Vector Encoder
Associative SCM
Serializer
Encoding Register/ Writeback Buffer
Write Address Decoder
Read Address Decoder
Wake-up Signal
100mW
10mW
1mW
100µW
10µW
1µW
Cognitive Sensor Data Processing
General Purpose Parallel Computation
6.9µW
1.7µW
20.9µW
2.136µW
15mW
0.7mW
49.4mW
5mW
CWU
CWU + retention
SoC Active
Cluster Active

Figure 4.4.2: Cognitive Smart Wake-Up Architecture.

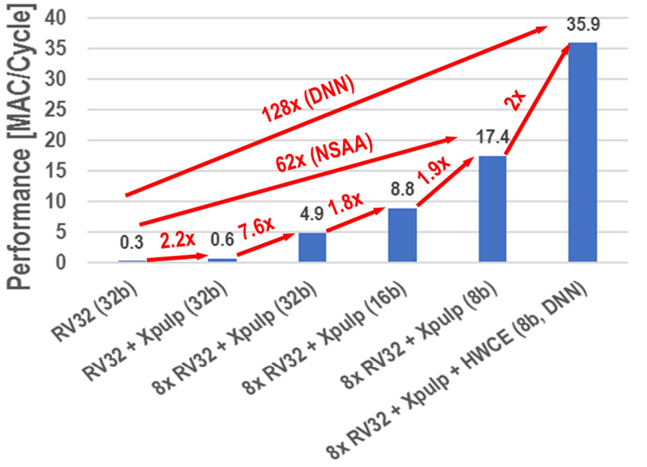


Figure 4.4.3: Performance of integer matrix multiplication exploiting Xpulp Extensions, software parallelism on 8 cores, SIMD parallelism on 16- and 8-datatypes, HWCE. Similar performance gains can be achieved on FP kernel when applicable.

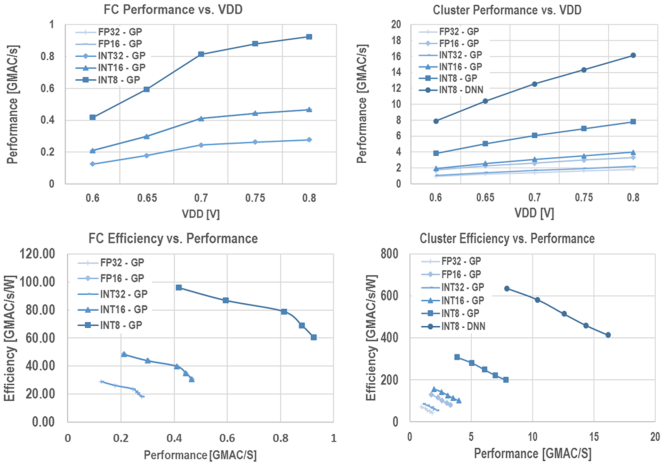


Figure 4.4.4: Performance and efficiency of a MatrixMul kernel running with multiple-precision on general-purpose and DNN-specific cores of the FC (SoC domain) and Cluster.

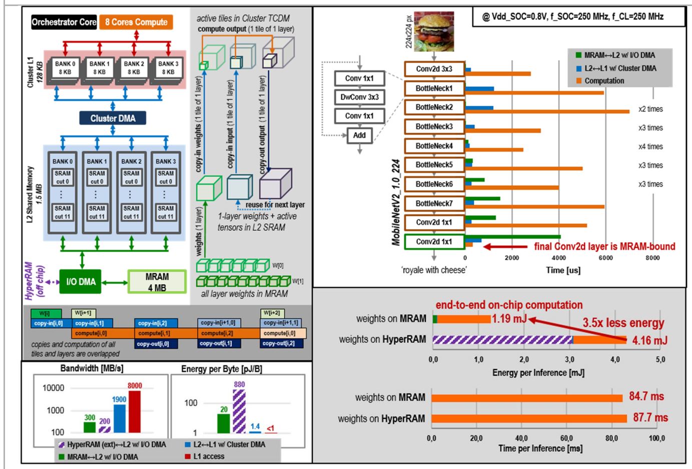
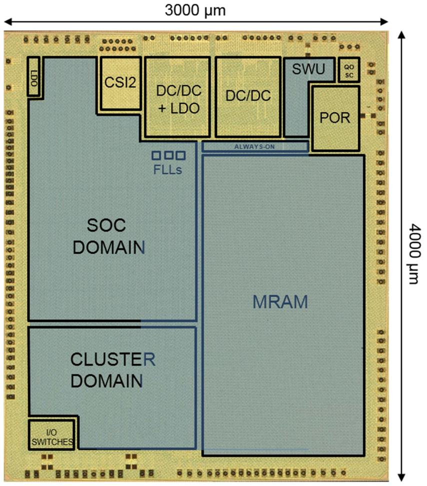

Orchestrator Core
8 Cores Compute
active tiles in Cluster TCDM
compute output (1 tile of 1 layer)
Cluster L1 128 KB
Cluster DMA
L2 Shared Memory 1.5 MB
HyperRAM (off chip)
I/O DMA
MRAM 4 MB
copy-in weights (1 tile of 1 layer)
copy-in input (1 tile of 1 layer)
copy-out output (1 tile of 1 layer)
reuse for next layer
1-layer weights + active tensors in L2 SRAM
all layer weights in MRAM
copies and computation of all tiles and layers are overlapped
Bandwidth [MB/s]
Energy per Byte [pJ/B]
HyperRAM (ext)↔L2 w/ I/O DMA
MRAM↔L2 w/ I/O DMA
L2↔L1 w/ Cluster DMA
L1 access
@ Vdd_SOC=0.8V, f_SOC=250 MHz, f_CL=250 MHz
224x224 px
MobileNetV2 1.0_224
'royale with cheese'
MRAM↔L2 w/ I/O DMA
L2↔L1 w/ Cluster DMA
Computation
final Conv2d layer is MRAM-bound
Time [us]
end-to-end on-chip computation
weights on MRAM
weights on HyperRAM
1.19 mJ
3.5x less energy
4.16 mJ
Energy per Inference [mJ]
84.7 ms
87.7 ms
Time per Inference [ms]

Figure 4.4.5: Top-left: DNN execution model with layer-wise tensor tiling; bottom-left: bandwidth and energy per byte of available memories; top-right: MobileNetV2 use case executed at INT-8 GP precision, showing transfer/compute overlap; bottom-right: time and energy of end-to-end on-chip execution with weights on MRAM vs off-chip on HyperRAM.

	SleepRunner [2]	Mr.Wolf [3]	SamurAI [4]	**Vega (this work)**
Technology	CMOS 28nm FDSOI	CMOS 40nm LP	CMOS 28nm FDSOI	CMOS 22nm FDSOI
Die Area	0.68 mm^2	10 mm^2	4.5 mm^2	12 mm^2
Applications	IoT GP	IoT GP + NSAA	IoT GP + NSAA + DNN	IoT GP + NSAA + DNN
CPU/ISA	CM0DS Thumb-2 subset	9 x RI5CY RVC32IMFXpulp	1x RI5CY RVC32IMFXpulp	10 x RI5CY RVC32IMFXpulp + SF
Embedded SRAM (State Retentive)	64 kB s.r.	64 kB (L1) 512 kB s.r. (L2)	464 kB 40 kB s.r.	128 kB (L1) **1600 kB s.r. (L2)**
Embedded NVM	-	-	-	**4 MB MRAM**
Wake-up Sources	WiC	GPIO, RTC	WuR, RTC, Int, GPIO	GPIO, RTC, **Cognitive**
Sleep Power SRAM Ret. Slp. Power (St. Ret. SRAM)	5.4 µW 9.4 µW (64 kB s.r.)	72 µW 76.5 - 108 µW (32 kB - 512 kB s.r.)	6.4 µW (49 kB s.r.)	1.7 µW (CWU) 2.8 - 123.7 µW (16 kB - 1.6 MB s.r.)
Int Precision FP Precision	32	8, 16, 32 FP32	8, 16, 32 n.a.	8, 16, 32 FP32, **FP16, bfloat**
Supply Voltage	0.4 - 0.8 V	0.8 - 1.1 V	0.45 - 0.9 V	0.5 - 0.8 V
Max Frequency	80 MHz	450 MHz	350 MHz	450 MHz
Power Range	5.4 - 320 µW	72 µW - 153 mW	6.4 µW - 96 mW	**1.7 µW - 49.4 mW**
[1]Best Int Perf. [1]Best.Int Eff. [1] @ Perf.	31 MOPS (32b) 97 MOPS/mW (32b) @ 18.6 MOPS (32b)	12.1 GOPS 190 GOPS/W @ 3.8 GOPS	1.5 GOPS 230 GOPS/W @110 MOPS	**15.6 GOPS[4]** **614 GOPS/W[4]** **@7.6 GOPS[4]**
[2]Best FP Perf. [2]Best FP Eff. [2] @Perf	-	1 GFLOPS 18 GFLOPS/W @350 MFLOPS	-	**2 GFLOPS[4]** **79 GFLOPS/W[4]** **@ 1 GFLOPS[4]** **3.3 GFLOPS (FP16)[4]** **129 GFLOPS/W (FP16)[4]** **@ 1.7 GFLOPS (FP16)[4]**
[3]Best ML Perf. [3]Best ML Eff. [3]@Perf	-	-	36 GOPS 1.3 TOPS/W @ 2.8 GOPS	32.2 GOPS[4] 1.3 TOPS/W[4] **@15.6 GOPS[4]**

[1] 2 OPs = 1 8-bit MAC on MatMul benchmark unless differently specified.
[2] 2 FLOPSs = 1 32-bit FMAC on MatMul benchmark unless differently specified.
[3] 8-bit ML Workloads
[4] Execution from SRAM

Figure 4.4.6: Comparison with state of the art.


3000 µm
LDO
CSI2
DC/DC
+ LDO
DC/DC
SWU
QO
SC
POR
ALWAYS-ON
FLLs
SOC
DOMAIN
MRAM
CLUSTER
DOMAIN
I/O
SWITCHES
4000 µm

Technology	22nm FDSOI
Chip Area	12mm^2
SRAM	1728 kB
MRAM	4 MB
WU Sources	GPIO, RTC, Cognitive
VDD range	0.5V - 0.8V
VBB range	0V - +1.1V
Freq. Range	32 kHz - 450 MHz
Pow. Range	1.7 µW - 49.4 mW
Int. Perf.	15.6 GOPS
Int. Eff.	614 GOPS/W
F.P. Perf.	3.3 GFLOPS
F.P. Eff.	129 GFLOPS/W
Int. ML Perf.	32.2 GOPS
Int. ML Eff.	1.3 TOPS/W

Figure 4.4.7: Chip micrograph and specifications.