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

Ivan Miro-Panades, Benoit Tain, Jean-Frédéric Christmann, David Coriat, Romain Lemaire, Clement Jany, Baudouin Martineau, Fabrice Chaix, Anthony Quelen, Emmanuel Pluchart, 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

Ivan Miro-Panades[1], Benoit Tain[2], Jean-Frédéric Christmann[1], David Coriat[1], Romain Lemaire[1], Clement Jany[3], Baudouin Martineau[3], Fabrice Chaix[3], Anthony Quelen[3], Emmanuel Pluchart[1], Jean-Philippe Noel[1], Reda Boumchedda[3,4], Adam Makosiej[3], Maxime Montoya[3], Simone Bacles-Min[1], David Briand[2], Jean-Marc Philippe[2], Alexandre Valentian[1], Frédéric Heitzmann[3], Edith Beigne[3], Fabien Clermidy[1]

[1]Univ. Grenoble Alpes, CEA, LIST, Grenoble, France; [2]Univ. Paris-Saclay, CEA, LIST, Gif sur Yvette, France;
[3]Univ. Grenoble Alpes, CEA, LETI, Grenoble, France; [4]STMicroelectronics, Crolles, France; Email: ivan.miro-panades@cea.fr

Abstract

IoT node application requirements are torn between sporadic data-logging and energy-hungry data processing (e.g. image classification). This paper presents a versatile IoT node covering this gap in processing and energy by leveraging two on-chip sub-systems: a low power, clock-less, event-driven Always-Responsive (AR) part and an energy-efficient On-Demand (OD) part. The AR contains a 1.7MOPS event-driven, asynchronous Wake-up Controller (WuC) with 207ns wake-up time optimized for short sporadic computing. OD combines a deep-sleep RISC-V CPU and 1.3TOPS/W Machine Learning (ML) and crypto accelerators for more complex tasks. The node can perform up to 36GOPS while achieving 15,000x reduction from peak-to-idle power consumption. The interest of this versatile architecture is demonstrated with 105µW daily average power on an applicative classification scenario.

Introduction

An event-driven IoT node is a way to reduce the power consumption of sporadic computing. SamurAI (Fig. 1) combines an event-driven WuC using asynchronous logic (low-energy, clock-less, and fast wake-up time) in the AR sub-system with an energy efficient synchronous RISC-V CPU including specialized accelerators in the OD sub-system to make a versatile IoT node. Depending on the application needs, one or both cores can be used as shown in Fig. 2.

Always-Responsive Sub-System

The WuC (Fig. 3), a clock-less 32b MCU [9] with 16b RISC ISA, is the master on the AR sub-system having 1.7MOPS at 0.45V and 1.6µW idle power. Program and data are stored on an asynchronous 8kB Two-Port SRAM (TP-SRAM) [10] with auto power-down capabilities down to 0.4V and 4.6µW idle power. Key component of this architecture, TP-SRAM is also connected to the OD sub-system through an AHB asynchronous interface to create a shared memory space between the two sub-systems. Thanks to asynchronous logic, the wake-up time from idle state to first instruction fetch takes 207ns, i.e. a third of an instruction cycle. The AR sub-system contains multiples wake-up sources: an internal timer, OD interrupts, GPIOs connected to sensors, and a Wake-up Radio (WuR). The WuR senses the radio channel with 10x less power than the main radio. Thanks to its mixer-first topology (as in [7]) and the use of three distinct oscillators, the RF front-end enables operation in all the main IoT bands: 433MHz, 868MHz and 2.4GHz. The digital baseband (DBB) supports data-rates up to 100kbps. It decodes an 8b identifier to selectively wake-up the WuC and a 32b message payload for application specific purposes. At 50kbps data-rate, the WuR achieves -73dBm sensitivity with 4.1µW power consumption at 5% duty cycle and this drop to 40nW in idle mode.

On-Demand Sub-System

The 4-stage pipeline RISC-V CPU is the master of the OD sub-system (Fig. 1). Its memory sub-system is composed of 64kB for program (TCPM), 128kB for data (TCDM) and external NVM FeRAM memories. 32kB of TCDM have retention capability with 1.03pA/bit leakage at 0.5V. An instruction cache and a data interface allow direct FeRAM operations. Both cores (in AR and OD) share the APB peripherals while synchronizing through interrupts and locks. The Crypto IP embeds AES, TRIVIUM and PRESENT stream-cipher accelerators to support various encryption formats. The adaptive voltage scaling (AVS) module manages 128 sensors (TFS) and a programmable replica path (TFR) to estimate and track the Fmax/Vmin according to the applicative needs. To offer ML inference capability in this event driven IoT node, we implement PNeuro [8] (Fig. 4), a SIMD programmable accelerator composed of 2 clusters of 32 8-bit PEs each and 264kB multi-banked SRAM. Designed to accelerate neural networks, it performs up to 64MACs/cycle.

Measurements Results

The 4.5mm² circuit (Fig. 8) has been fabricated in 28nm FDSOI technology and contains 6 switchable power domains. For fair comparison with the SoA, measurements are done without using body-bias. Fig. 6 depicts Fmax and energy per cycle of the OD sub-system with RISC-V running Dhrystone, showing 19pJ/cycle at 25MHz, 0.48V and up to 350MHz at 0.9V. Fig. 7 shows the performance of PNeuro block, reaching 1.3TOPS/W and 2.8GOPS at 0.48V and up to 36GOPS at 0.9V for 8b precision fully-connected layers. Fig. 5 reports power consumption for different mode: 96mW at full activity and 6.4µW at 0.45V. The 15,000x ratio between peak and idle power, highlights the adaptive and versatile performance of this architecture. Fig. 9 shows a scenario where SamurAI is used to classify a scene based on the presence of people signaled by a pyroelectric detector (PIR). To minimize power consumption, the WuC filters PIR activity based on previous scenes and powers up the OD part only when required: PNeuro classifies the images acquired by the CPU and shares results with the WuC. AES encrypted messages can be transmitted through an external low-power radio. The WuR is also used to receive user commands. The daily average power for this application is 105µW where 26% is consumed on SamurAI (Fig. 10). Using RISC-V instead of PNeuro would increase the total average power consumption by 2.3x. Fig. 11 compares the circuit to prior art and shows significant improvements in terms of versatility, performance, wake-up time and power reduction.

Acknowledgments

WAKeMeUP (ECSEL 783176), SERENE-IoT (Penta 16004), MACS (FUI20) projects, ST Microelectronics, and PULP project.

References

[1] G. Lallement, JSSC, 2018. [2] Yu Pu, JSSC, 2018. [3] S. Paul, JSSC, 2017. [4] J. Myers, JSSC, 2016. [5] S. Bang, ISSCC, 2017. [6] A. Pullini, JSSC, 2019. [7] N. M. Pletcher, ISSCC, 2008. [8] A. Carbon, DATE, 2018. [9] J.-F. Christmann, JLPEA, 2019. [10] R. Boumchedda, L-SSC, 2018.

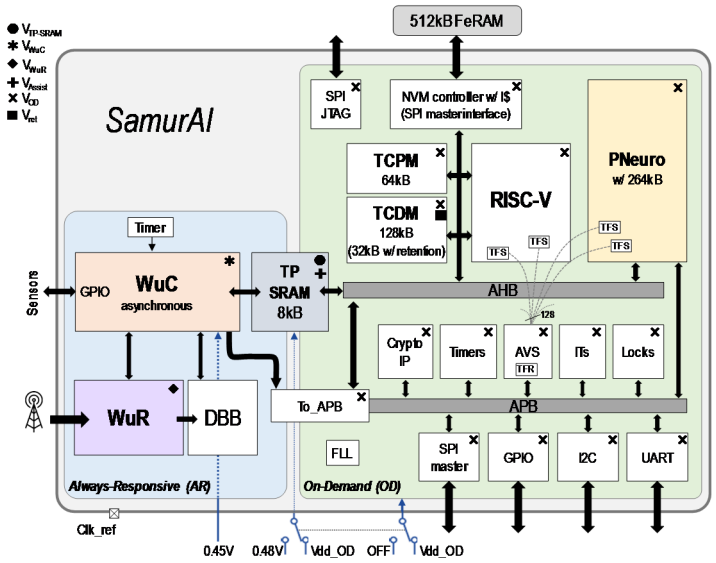


Fig. 1: SamurAI system architecture, with Always-Responsive and On-Demand sub-systems and associated power domains.

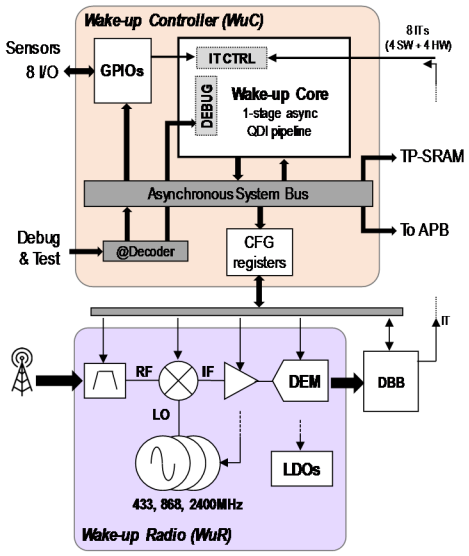


Fig. 3: Wake-up Controller and Radio architecture details.

Power mode	Always-Responsive (AR)			TP-SRAM		On-Demand (OD)		
	Voltage (V)	WuC state	WuR State	Voltage (V)	State	Voltage (V)	RISC-V Freq (MHz)	Periph Freq (MHz)
IDLE	0.45	Sleep	OFF	0.48	Sleep	OFF	-	-
WuC only	0.45	Run	OFF	0.48	Run	OFF	-	-
WuC+WuR	0.45	Run	ON	0.48	Run	OFF	-	-
WuC+Periph. (cpu sleep)	0.45	Run	ON/OFF	0.48	Run	0.48	Gated	10
CPU running	0.45	Run/Sleep	ON/OFF	0.48 – 0.9	Run	0.48 – 0.9	1 – 350	1 – 350

Fig. 2: SamurAI power modes.

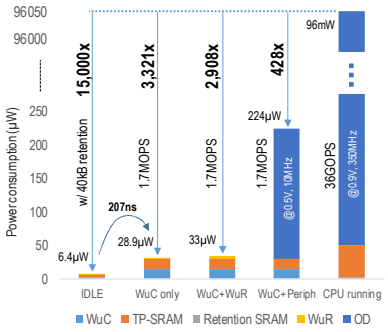


Fig. 5: Power consumption measurements and reduction w.r.t. power modes.

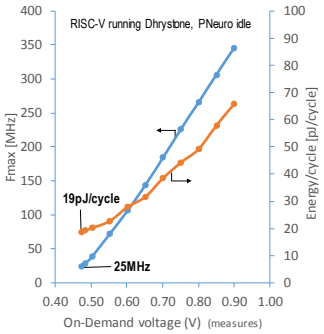


Fig. 6: Fmax and energy per cycle of OD sub-system.

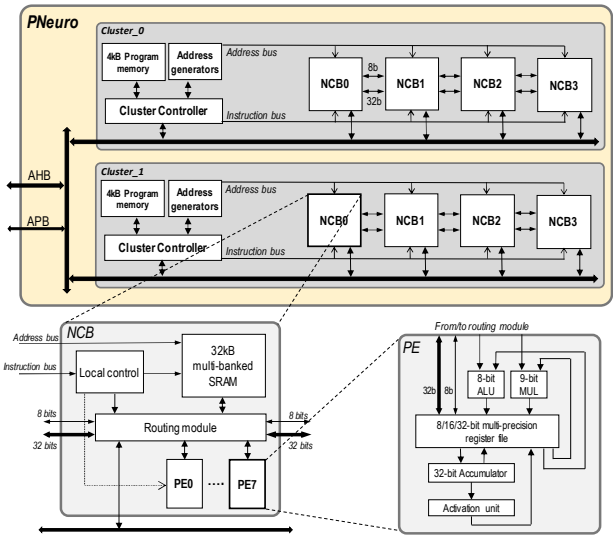


Fig. 4: Two-cluster PNeuro accelerator with 64 PEs.

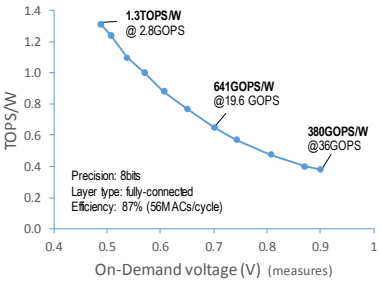


Fig. 7: PNeuro energy efficiency.

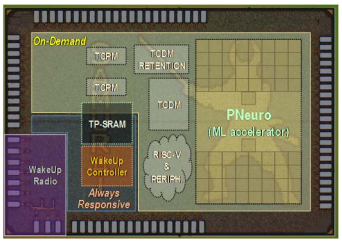
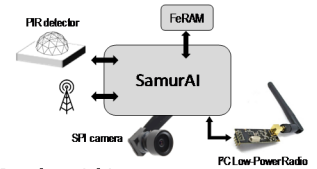


Fig. 8: Die micrograph, 4.5mm².



Scenario constraints
Room occupation 8H/day
Mean PIR detection interval 5s during occupation
PIR power consumption 6µW
Camera 224x224 B&W pixels
Camera power consumption 2.5mW@1FPS
Low-Power radio: energy 180mJ/message, 10message/day

Fig. 9: Presence classification scenario using SamurAI with off-the-shelf components.

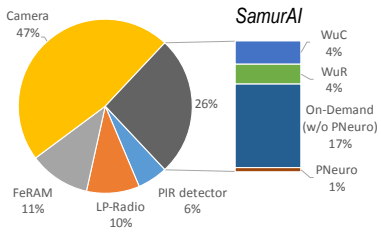


Fig. 10: Daily average power breakdown of presence classification scenario (70% PIR filtering), 105µW total power.

	This work	JSSC 2018 Lallement [1]	JSSC 2018 Yu Pu [2]	JSSC 2017 S. Paul [3]	JSSCC 2016 J. Myers [4]	ISSCC 2017 S. Bang [5]	JSSC 2019 A. Pullini [6]
Technology	28nm FDSOI	28nm FDSOI	28nm LP	14nm FinFET	65nm LP	40nm	40nm LP
CPU	32b Async RISC 32b RISC-V	M0+	M0	x86 IA	M0+	M0	32b RCVC32IMFX
Memory	464kB SRAM	8kB SRAM	-	72kB SRAM 8kB I$ 16kB ROM	24kB SRAM	270kB SRAM	512kB SRAM 4kB I$ 64kB D$
Wake-up Radio	Yes	No	No	No	No	No	No
ML accelerator	Yes	No	Yes	No	No	Yes	Yes
AVS	Yes	No	Yes	Yes	Yes	No	Yes
Crypto IPs	Yes	No	No	No	Yes	No	No
CPU state retention in deep sleep	Yes	No	-	Yes (S1)	Yes	No	Yes
Voltage range	0.45V – 0.9V	0.47V – 0.65V	0.55	0.308V – 1V	0.25V – 1.2V	0.63V – 0.9V	0.8V – 1.1V
Maximum frequency	**350MHz**	150MHz	50MHz	297MHz	66MHz	19.3MHz	450MHz
Deep sleep power (retention memory)	6.4µW (40kB SRAM)	0.704µW (8kB SRAM)	1.71µW (NA)	-	80nW (8kB SRAM)	-	108µW (448kB SRAM)
Peak-to-idle power reduction	**15,000x**	51.5x	-	4.7x	6,940x	-	1,416x
Wake-up time from deep-sleep	**207ns** (35% of inst. cycle)	~µs	-	> 1ms (S1) > 1s (S0)	~µs	-	-
GOPS	**1.7MOPS – 36GOPS**	150MOPS	-	-	66MOPS	-	7GOPS
Best performance	**1.3TOPS/W @ 2.8GOPS**	370GOPS/W @ 16MOPS	-	58GOPS/W @ 3.5MOPS	85GOPS/W @ 750kOPS	374GOPS/W @ 107MOPS	120GOPS/W @2.2GOPS

Fig. 11: Comparison table.