Quentin: an Ultra-Low-Power PULPissimo SoC in 22nm FDX

Pasquale Davide Schiavone*, Davide Rossi‡*, Antonio Pullini*, Alfio Di Mauro*, Francesco Conti*‡, Luca Benini*‡
*ETH Zurich, Switzerland, ‡University of Bologna, Italy

***Abstract*—The End-Nodes of the Internet of Things (IoT) require extreme energy efficiency coupled with wide power-performance operating range. Fully-depleted SOI (FD-SOI) is an attractive technology for ultra-low power and wide-range operation as it offers compelling options to tune power, performance, area (PPA) at design time as well as at run time. This paper describes Quentin: an MCU-class (32bit) open-source RISC-V SoC featuring an autonomous I/O subsystem optimized to deal with the wide variety of sensors available in IoT end-nodes, coupled with a processor optimized for near threshold computation and a heterogeneous (standard-cell and SRAM) memory architecture to better exploit the low-voltage capabilities of 22nm FDX technology. The system runs up to 2400 million equivalent RV32IMC instructions per second (MOPS) and achieves best power density of 6 μW/MHz, resulting into an energy efficiency of 433 MOPS/mW.**

***Index Terms*—Microcontroller, MCU, IoT, RISC-V, FD-SOI, Near-Threshold Computing.**

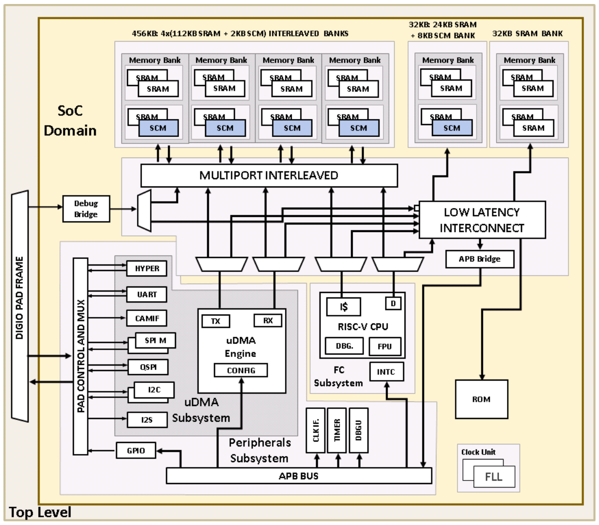


Fig. 1. Quentin SoC Architecture.

I. Introduction

IoT end-nodes are evolving from simple brokers of low-bandwidth sensory data to endpoint data analytics devices running complex algorithms on high bandwidth data streams coming from smart sensors. Still, these devices have to cope with stringent constraints as they inhabit small, unobtrusive battery-powered devices such as wearable systems for sport or healt monitoring, tiny smart cameras disseminated in the environment, autonomous nano-robots and UAVs [1]–[4]. The wide variety of applications in the IoT scenarios requires programmable systems for short time-to-market and versatility costs. Open-Hardware offers free IPs up to full systems to reduce design costs and it recently became a *de facto* attractive scenario for the electronics market.

IoT end-nodes are exposed to high energy efficient constraints, thus they need to be designed with extra care along the whole Software-Hardware stack. The Parallel Ultra-Low Power (PULP) project aims to develope free and open-source microcontrollers optimized to maximize the energy efficiency of IoT end-nodes [5]. PULP is a multicore system with a rich set of peripherals, memory and cores based on the open-source RISC-V Instruction Set Architecture (ISA) [6]. The system is organized as a single core fabric controller (called PULPissimo) extended with an optional cluster of cores. The system with all its IPs and the software runtime have been recently released open-source[1]. High energy efficiency is achieved by combining near-threshold operation (NTC) and parallel execution on chips implemented in advanced technology nodes.

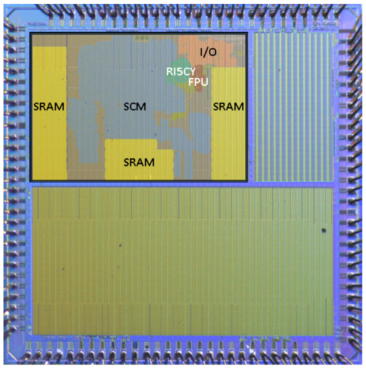
FD-SOI technology offers the possibility to adapt the energy consumption leveraging adaptive body-bias to lower the power consumption (reverse body-bias - RBB) or increase performance (forward body-bias - FBB). This paper presents Quentin, a single core implementation of PULPissimo platform in the 22nm GLOBALFOUNDRIES FDX technology node. Size, performance, power and energy results of the implementation are discussed in this paper as following: in the next Section, the Quentin architecture is described; its measurements and results are shown in Section III; and in the last Section we discuss the conclusions.

II. Quentin SoC

The Quentin SoC is an implementation of the open-source advanced PULPissimo microcontroller in the 22nm FDX technology. Quentin equips a 32-bit in-order 4-pipeline stages RV32IMFC RISC-V processor [7]. The baseline RISC-V ISA of the processor has been enhanced with extensions targeting energy efficient digital signal processing such as hardware-loops, automatic increment of addresses during load/store operations, bit manipulation instructions, fixed-point and packed single-instruction-multiple-data (SIMD) operations. The SoC includes 520 kB of L2 memory and a ROM storing the boot-code. The L2 memory layout of Quentin is organized as 4 114 KB word-level interleaved banks to minimize conflicts during parallel accesses through the masters, plus 2 banks of 32 KB that can be used privately by the Fabric Controler (FC) (e.g. program, stack, private data) without incurring in banking conflicts. Both memory regions are implemented as a heterogeneous memory architecture composed of a mix of SRAM and standard-cells memory cuts (SCM) [8].

In particular, each of the interleaved banks has 2 of the 114 KB implemented as SCM and one of the private bank has 8 KB of SCM as shown in Figure 1, for a total of 504 KB of SRAM

[1] PULP/PULPissimo and all its IPs are freely downloadable at https://github.com/pulp-platform/ under the SolderPad licence



Technology	CMOS 22nm FDX
Transistor Type	Flip-Well (LVT)
SoC Area	2.3mm^2
VDD range	0.5V - 0.8V
Body Bias Range	0.0V - 1.4V
Memory Transistors	24.8M
Logic Transistors	4.18M
Frequency Range (with ABB)	32 kHz - 670 MHz
	32 kHz - 938 MHz
Power Range (with ABB)	300 µW - 10.4 mW
	300 µW - 66.2 mW

Fig. 2. Quentin Chip: Floorplan and physical results.

and 16 KB of SCM. The SCM portion of the private bank is implemented as a 3-read 2-write ports register file: 2 of the 3 read ports and 1 of the 2 write ports are dedicated to the data and instructions interfaces of the RISC-V core and 1 read and 1 write ports are used by the interconnect arbiter for any other master node of the system. From a performance viewpoint, this memory organization enables transparent sharing of the L2, increasing by 4x the system memory bandwidth with respect to a traditional single-port memory architecture without the use of high area overhead multiple ports memories. The SRAM cuts have separate power connections from the rest of the logic for both periphery and array connections. This allows the system to operate in an ultra-low-power mode using only the 16 KB SCM memories and shutting down the SRAM via an off-chip power manager.

The SoC includes a full set of peripherals: Quad SPI supporting up to two external devices, I2C, 2 I2S, a parallel camera interface, UART, GPIOs, JTAG and a DDR HyperBus interface to extend the size of the on-chip memory. An I/O DMA (μDMA [9]) manages data transfers through peripherals to minimize the workload of the processor. To improve the efficiency of I/O communications, the μDMA has a dedicated connections to all the peripherals through 2 dedicated 32-bit ports on the L2 memory interconnect, granting an aggregated bandwidth sufficient to satisfy the requirements of all the peripherals (up to 1.6 Gbit/s) with a frequency of 57 MHz. Debug of the Quentin MCU is possible via read and write operations to memory mapped registers of the core using JTAG.

III. Implementation Results

The floorplan area for Quentin is 2.31 mm^2 and its effective area is 1.22 mm^2 (6154KGE). Figure 2 shows the chip taped out. The chip is divided in 3 independent designs and Quentin resides in the top-left. Its main modules are highlighted and its physical characteristics are summarized in the table. Frequencies and power numbers with forward body-bias are calculated applying up to 1.4V.

An 8x8 32-bit matrix multiplication has been compiled and execute on Quentin to measure power, frequency and energy efficiency. The chip has been tested on the Advantest SoC hp93000 integrated circuit testing equipment. Supply voltages, as well as body bias voltages have been applied by means of dedicated hp93000 device channels. To characterize the system in different operational modes, three different setup have been tested for every measurements:

1) *SRAM*

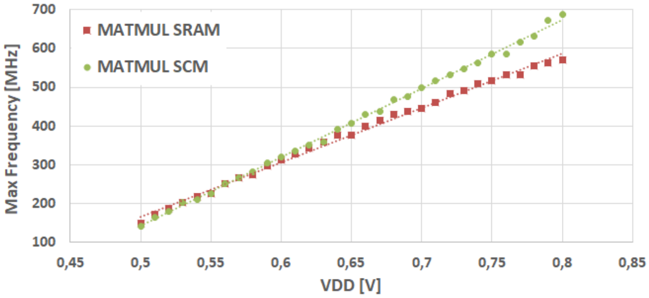


Fig. 3. Maximum frequency against supply voltage when code and data reside on SRAMs or SCMS and no body bias applied.

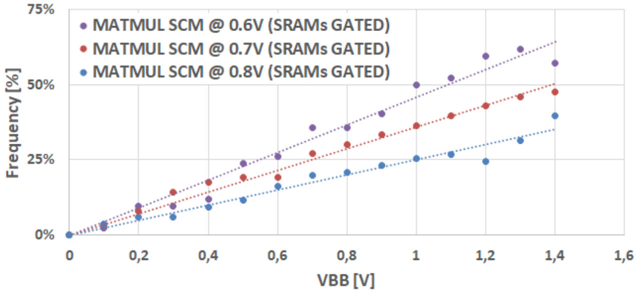


Fig. 4. Performance benefits from forward body-bias. FBB impacts more low-voltage points.

2) *SCM with SRAM on*
3) *SCM with SRAM gated.*

In the *SCM* experiments, data and code have been allocated on SCMs, whereas they are allocated on *SRAMs* in the third configuration. In the first case, the power connections of SRAM are power gated. In the second case the SRAM are powered on but not used, as for example in case the SRAMs hold data or the time to power off is too long. In the *SRAM* setup, SRAMs power connections of array and periphery are connected to the same voltage level of the rest of the logic. Figure 3 shows the maximum operating frequency of Quentin when running the matrix multiplication on SCMs or SRAMs. Note that the maximum frequency of the *SCM* setup is the same whether the SRAMs are switched on or off. The chip starts working at 0.5V running at 148/1156MHz and achieves the peak frequency of 570/670MHz at 0.8V when running on SRAMs and SCMs respectively with no body-bias.

When applying FBB, the frequency can increase more than 60% (at 0.6V) and it achieves 938MHz at 0.8V when 1.4V are applied to the body-gate. Figure 4 shows how the maximum frequency changes for three supply voltages when FBB is applied on the *SCM* setup. It is interesting to note that lower the supply voltage, the higher the benefit of FBB.

The chip lowest power configuration uses only SCMs while SRAMs are power gated. In this setup, it consumes only 0.95mW at 0.5V and no FBB running at 156MHz, and it consumes up to 32.1mW at 0.8V with 1.4 FBB running at 938MHz. When SRAMs are switched on but not used, the leakage power increases by ~2mW at 0.8V and no BB. Figure 5 shows how the leakage power increases at three different voltage levels when FBB is applied from 0 to 1.4V and data and instructions are in SRAMs. It is possible to note that the leakage power increases faster at lower supply voltages.

Finally, the energy efficiency of the system measured in μW/MHz in Figure 6. At every point, the three setups are measured at their maximum efficiency. Given the higher

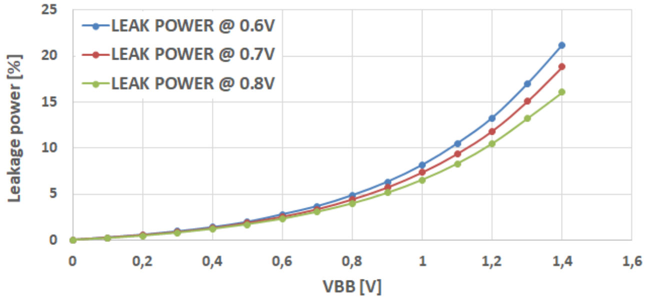


Fig. 5. Power increase due to forward body-bias. The lower the supply voltage the higher the penalty.

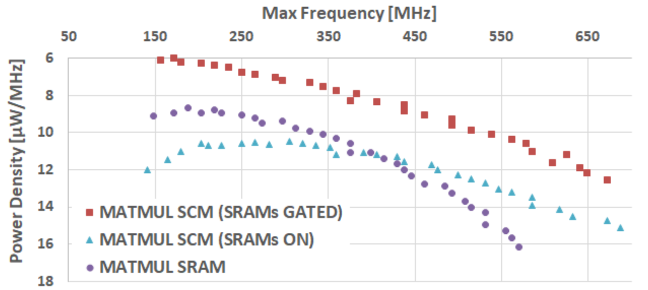


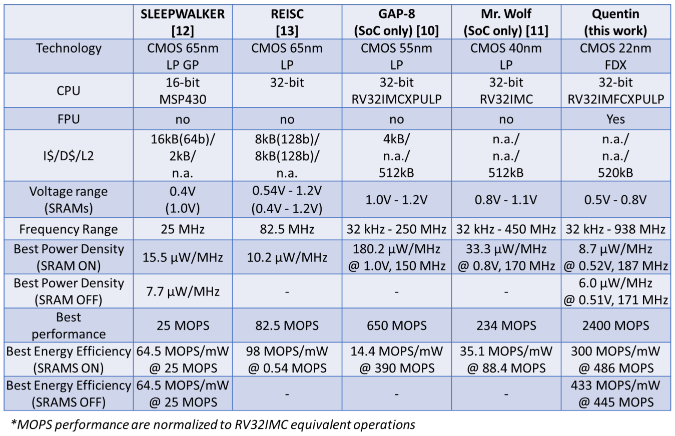
Fig. 6. Quentin power density. Using only SCMs is more energy efficient than SRAMs due to better performance.

frequency and lower power consumption, the *SCM* setup with SRAMs gated is the most energy efficient setup for every point as expected. However, in this configuration the system has a limited memory of 16 KB. In the configuration where operations are executed only on the SCMs but with the whole memory available (SRAMs on), the system has higher energy efficiency than operating on SRAMs for frequencies >400MHz. This can also be observed in Figure 3, as the supply voltage needed to reach such frequencies is higher on the *SRAM* setup, thus the power consumption increases.

Multicore PULP systems based on RISC-V have been already implemented in [10] and [11]. A comparison against their fabric controller and with two additional efficient processors [12], [13] is shown in Figure 7. Quentin shows the highest energy efficiency and performance as a single core microcontroller thanks to the compound of advanced architecture design and technology. With respect to [10], [11] and [12], it does not implement any on-chip power manager and does not have any state retentive memory, thus it has to rely on external memories. With respect to [11], the fabric controller adopts a more performant core, whereas Quentin is implemented in a more advanced technology with respect to [10].

IV. Conclusion

This paper presented Quentin, an advanced single core SoC for edge IoT applications implemented in GF 22FDX. The MCU features 520 KB implemented as 504 KB and 16 KB of SCMs. For tiny applications that fit in the SCM, Quentin offers up to 216 Million RV32IMC equivalent operations per mW resulting in high energy efficiency and a peak performance of 938MHz when FBB is applied at 0.8V. We showed that open-source microcontroller architectures implemented in advanced technology nodes can achieve top performance and energy efficiency to cope with IoT requirements.



	SLEEPWALKER [12]	REISC [13]	GAP-8 (SoC only) [10]	Mr. Wolf (SoC only) [11]	Quentin (this work)
Technology	CMOS 65nm LP GP	CMOS 65nm LP	CMOS 55nm LP	CMOS 40nm LP	CMOS 22nm FDX
CPU	16-bit MSP430	32-bit	32-bit RV32IMCXPULP	32-bit RV32IMC	32-bit RV32IMFCXPULP
FPU	no	no	no	no	Yes
I$/D$/L2	16kB(64b)/ 2kB/ n.a.	8kB(128b)/ 8kB(128b)/ n.a.	4kB/ n.a./ 512kB	n.a./ n.a./ 512kB	n.a./ n.a./ 520kB
Voltage range (SRAMs)	0.4V (1.0V)	0.54V - 1.2V (0.4V - 1.2V)	1.0V - 1.2V	0.8V - 1.1V	0.5V - 0.8V
Frequency Range	25 MHz	82.5 MHz	32 kHz - 250 MHz	32 kHz - 450 MHz	32 kHz - 938 MHz
Best Power Density (SRAM ON)	15.5 µW/MHz	10.2 µW/MHz	180.2 µW/MHz @ 1.0V, 150 MHz	33.3 µW/MHz @ 0.8V, 170 MHz	8.7 µW/MHz @ 0.52V, 187 MHz
Best Power Density (SRAM OFF)	7.7 µW/MHz	-	-	-	6.0 µW/MHz @ 0.51V, 171 MHz
Best performance	25 MOPS	82.5 MOPS	650 MOPS	234 MOPS	2400 MOPS
Best Energy Efficiency (SRAMS ON)	64.5 MOPS/mW @ 25 MOPS	98 MOPS/mW @ 0.54 MOPS	14.4 MOPS/mW @ 390 MOPS	35.1 MOPS/mW @ 88.4 MOPS	300 MOPS/mW @ 486 MOPS
Best Energy Efficiency (SRAMS OFF)	64.5 MOPS/mW @ 25 MOPS	-	-	-	433 MOPS/mW @ 445 MOPS

*****MOPS performance are normalized to RV32IMC equivalent operations*

Fig. 7. Comparison with state of the art efficient processors.

Acknowledgment

This work has been funded by the Swiss National Science Foundation under grant 162524 (MicroLearn: Micropower Deep Learning) and by the European H2020 FET project OPRECOMP (g.a. 732631).

References

[1] S. Benatti, F. Montagna, D. Rossi, and L. Benini, “Scalable eeg seizure detection on an ultra low power multi-core architecture,” in *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct 2016, pp. 86–89.

[2] M. Rusci, D. Rossi, E. Farella, and L. Benini, “A sub-mw iot-endnode for always-on visual monitoring and smart triggering,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1284–1295, Oct 2017.

[3] F. Conti, R. Schilling, P. D. Schiavone, A. Pullini, D. Rossi, F. K. Gürkaynak, M. Muehlberghuber, M. Gautschi, I. Loi, G. Haugou *et al.*, “An iot endpoint system-on-chip for secure and energy-efficient near-sensor analytics,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 9, pp. 2481–2494, 2017.

[4] D. Palossi, A. Loquercio, F. Conti, E. Flamand, D. Scaramuzza, and L. Benini, “Ultra low power deep-learning-powered autonomous nano drones,” *CoRR*, vol. abs/1805.01831, 2018. [Online]. Available: http://arxiv.org/abs/1805.01831

[5] D. Rossi, F. Conti, A. Marongiu, A. Pullini, I. Loi, M. Gautschi, G. Tagliavini, A. Capotondi, P. Flatresse, and L. Benini, “Pulp: A parallel ultra low power platform for next generation iot applications,” in *2015 IEEE Hot Chips 27 Symposium (HCS)*, Aug 2015, pp. 1–39.

[6] A. Waterman, Y. Lee, D. A. Patterson, K. Asanovic, V. I. U. level Isa, A. Waterman, Y. Lee, and D. Patterson, “The risc-v instruction set manual,” 2014.

[7] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gurkaynak, and L. Benini, “Near-threshold RISC-v core with DSP extensions for scalable IoT endpoint devices,” *IEEE Transactions on Very Large Scale Integration Systems*, pp. 1–14, 2017.

[8] A. Teman, D. Rossi, P. Meinerzhagen, L. Benini, and A. Burg, “Controlled placement of standard cell memory arrays for high density and low power in 28nm FD-SOI,” in *The 20th Asia and South Pacific Design Automation Conference*, Jan 2015, pp. 81–86.

[9] A. Pullini, D. Rossi, G. Haugou, and L. Benini, “ μDMA: An autonomous I/O subsystem for IoT end-nodes,” in *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, Sept 2017, pp. 1–8.

[10] E. Flamand, D. Rossi, F. Conti, I. Loi, A. Pullini, F. Rotenberg, and L. Benini, “Gap-8: A risc-v soc for ai at the edge of the iot,” in *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, July 2018, pp. 1–4.

[11] A. Pullini *et al.*, “Mr.wolf: A 1 gflop/s energy-proportional parallel ultra low power soc for iot edge processing,” in *IEEE ESSCIRC*, 2018.

[12] D. Bol, J. De Vos, C. Hocquet, F. Botman, F. Durvaux, S. Boyd, D. Flandre, and J.-D. Legat, “A 25mhz 7μw/mhz ultra-low-voltage microcontroller soc in 65nm lp/gp cmos for low-carbon wireless sensor nodes,” in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*. IEEE, 2012, pp. 490–492.

[13] N. Ickes, Y. Sinangil, F. Pappalardo, E. Guidetti, and A. P. Chandrakasan, “A 10 pj/cycle ultra-low-voltage 32-bit microprocessor system-on-chip,” in *ESSCIRC (ESSCIRC), 2011 Proceedings of the*. IEEE, 2011, pp. 159–162.