Informe final de participación en cada run: TO2

Diseño de aceleradores basados en la tecnología RISCV para la próxima generación de computadoras (DRAC)

Código de Proyecto: 001-P-001723

Número de entregable:	E5.3.2
Nombre de entregable:	Informe final de participación en cada run: TO2
Revisión:	01
Fecha límite del entregable:	31 diciembre, 2022
Fecha de entrega:	19 diciembre, 2022
Fecha de inicio del proyecto:	1 June 2019
Duración:	36 Meses
Partner responsible del entregable:	Universitat de Barcelona
Autor (es) del entregable:	UB, BSC, UPC, UAB, URV

El proyecto DRAC con número de expediente 001-P-001723 ha sido cofinanciado en un 50% con 2.000.000,00€ por el Fondo Europeo de Desarrollo Regional de la Unión Europea en el marco del Programa Operativo FEDER de Cataluña 2014-2020, con el soporte de la Generalitat de Cataluña.

Grado de divulgación		
PU	Público	
CO	Confidencial, solo para miembros del consorcio	**X**

Historia del documento

Versión	Fecha	Descripción/Cambios	Razones
01	19 diciembre, 2022	Entregable enviado	
02	6 febrero 2023	Corrección errores	Revisión

Index

Executive summary

This document includes the designs made by the different groups participating in the project.

1. Introduction

The Kameleon chip is the latest DRAC (Designing RISC-V-based Accelerators for next generation Computers) project. It is a 3 mm by 3 mm chip which has been made by a collaboration of the following groups: Universitat Autònoma de Barcelona (UAB), Centre de Supercomputació de Barcelona (BSC), Universitat Rovira i Virgili (URV), Universitat Politècnica de Catalunya (UPC), Centre Nacional de Microelectrònica (CNM) and Universitat de Barcelona (UB).

The manufacture of the chip had a cost of 169.115 €. The design was sent for manufacture the twelfth of December of 2022, without any DRC (Design Rule Check) error, meaning that everything present in the design can be manufactured without issues. As presented in deliverable E5.1.2, this was the last date available in 2022 to send to fabrication a chip with GlobalFoundries 22 nm. The UB has been in charge of designing the physical implementation of this chip, ensuring the design had no DRC errors and sending the design files once finished to the foundry for manufacture, in the time limit imposed by the foundry.

The Kameleon chip has different Intellectual Properties (IPs) which are part of it: Accelerators in charge of accelerating certain processing tasks, cores for general purpose processing, Phase-Locked Loop (PLL) in charge of the generation of the clock signal and a Serializer/Deserializer (SerDes) which adapts the information that is input and output to the System-on-Chip (SoC). These different IPs will be explained in detail in the following sections, each of them has been made by different institutions, the institutions responsible for the digital synthesis and physical design of each of them are listed in Table I.

Table I: Authorship of the Digital synthesis and Physical design of each IP as well as the SoC which incorporates them all.

IP	Function	Digital Synth.	Physical Synth.
Sargantana & Lagarto Ka	Cores	BSC	UB
PQC	Acceleration in Encryption and Decryption	URV + BSC	BSC
SAURIA	Acceleration in Deep Neural Networks	BSC	BSC
Picos	Task management & scheduler	UPC + BSC	BSC
WFA	Computational Biology	UAB + BSC	BSC
Quad SerDes	Fast Communications	UPC + BSC	UPC + BSC
PLL + Spi_slave	Fast clock	CNM	CNM
Kameleon	Integration of all IPs + memory bus	BSC	UB

For the Cores + Cache and accelerator IPs, the number of instances, logic gates given by Innovus (Place and Route tool from Cadence), and transistor count, calculated by approximated methods, are presented in Table II.

Table II: Number of instances, logic gates and transistors of each IP.

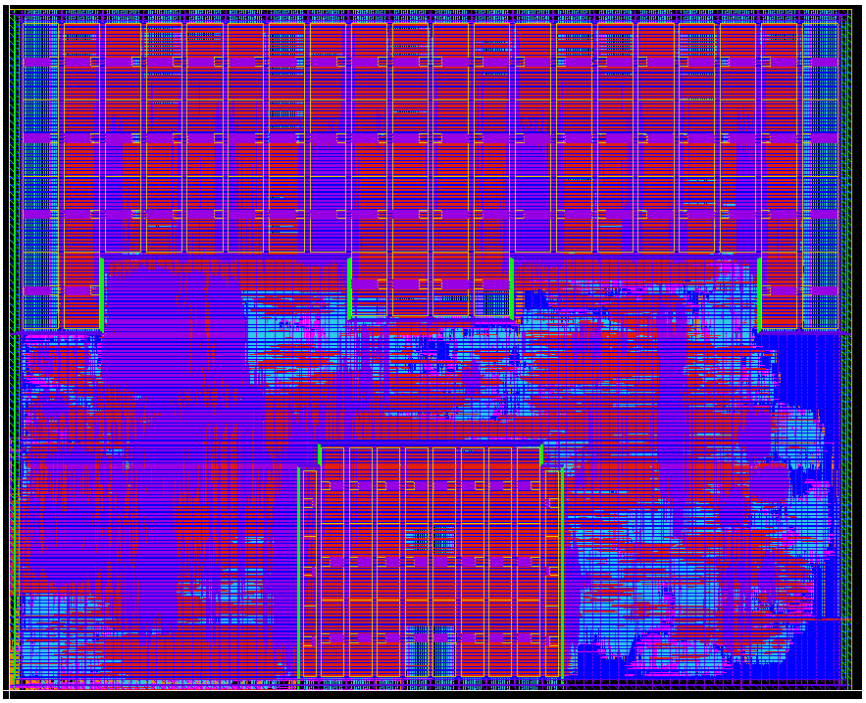
IP	Logic gate count	Transistor count
Sargantana & Lagarto Ka	1 149 688	45 023 861
PQC	473 232	6 856 104
SAURIA	209 460	12 384 750
Picos	94 459	4 128 974
WFA	208 518	26 903 716
Kameleon	325 917	4 414 269
Total	**2 461 274**	**99 711 674**

2. Function of each of the IPs

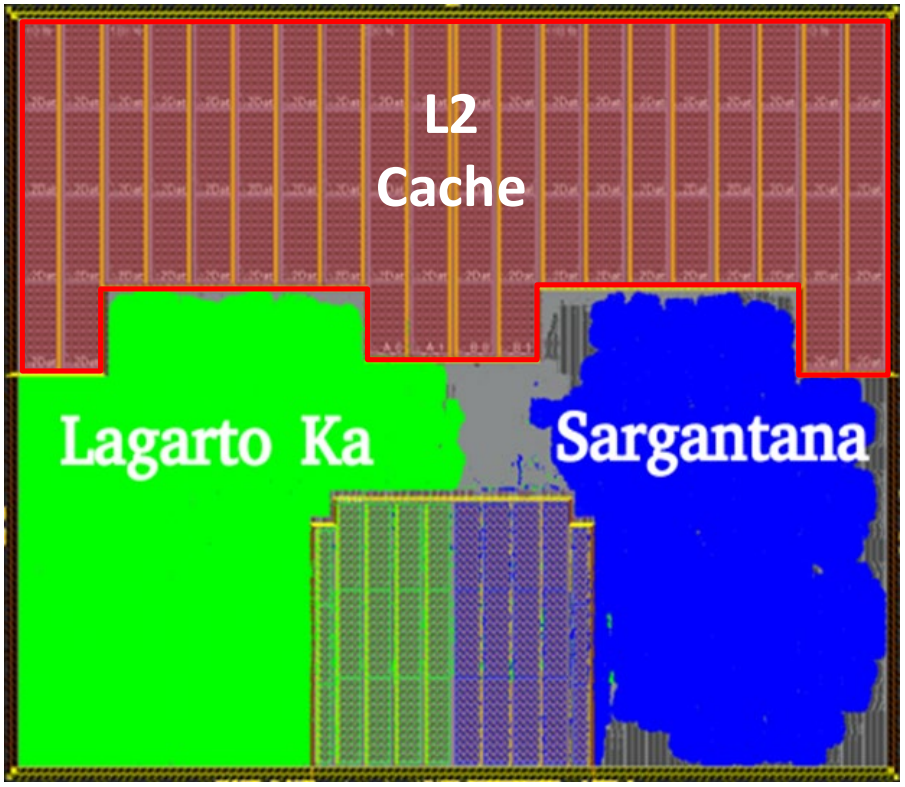
Having provided information about the authorship and number of internal elements of each of the IPs, this section describes in detail the function of each IP.

2.1. Sargantana and Lagarto Ka

Sargantana and Lagarto Ka are two independent RISC-V cores that are not intended to work at the same time. Only one core will be active and able to access L2 cache during operation. Figure 1 presents a picture of the layout of both cores (1a) and the distribution of each core in the floorplan (1b). As can be seen, both cores of the Kameleon SoC have been combined into the same IP, even so they are different cores with different functions which will now be explained.



(a)



(b)

Figure 1: (a) Layout of Sargantana and Lagarto Ka cores IP, with a size of 1.86 mm x 1.50 mm. (b) Distribution of the cores in the floorplan

a. Sargantana

Sargantana is a 64-bit processor based on RISC-V that implements the RV64G ISA, a subset of the vector instructions extension (RVV 0.7.1), and custom application-specific instructions. Sargantana features a highly optimized 7-stage pipeline that achieves frequencies higher than 1GHz. The pipeline integrates advanced optimizations to increase the IPC. It implements out-of-order write-back, register renaming, and a non-blocking memory pipeline to remove most data and structural dependencies. Moreover, Sargantana features a double-precision Floating Point Unit (FPU) and a 128-bit Single Instruction Multiple Data (SIMD) unit that accelerates domain-specific applications.

Sargantana delivers a 1.77× higher Instructions Per Cycle (IPC) than our previous 5-stage in-order DVINO core, reaching 2.44 CoreMark/MHz and 0.77 GFLOPS (at 1GHz) on the dgemm benchmark. The Sargantana core design delivers comparable or even higher performance than other state-of-the-art academic cores under Autobench EEMBC benchmark suite. This way, Sargantana lays the foundations for future RISC-V based core designs able to meet industrial-class performance requirements for scientific, real-time, and high-performance computing applications.

The Sargantana core has already been manufactured in the past (TO1), being the only core present in the Sargantana chip. That chip had a work frequency of 1.2 GHz. In the Kameleon tapeout, the integration of an additional core (Lagarto Ka) and the addition of more L2 memory makes the target frequency of the cores to be 800 MHz.

b. Lagarto Ka

The Lagarto Ka is a 2-way 64-bit processor that supports the I, M, and A extensions of the RISC-V ISA composed of a 10-cycle pipeline that embraces the different stages of the implementation. The microarchitecture is shaped by two main blocks: a sequential front-end, and an out-of-order back-end. On the front-end, the core fetches and issues two instructions each clock cycle and is able to execute speculative datapaths, resolving instruction predictions by including a bimodal branch predictor coupled with a runtime recovery mechanism to handle mispredictions as soon as the conditional branch is evaluated.

The instructions dispatched to the back-end of the Lagarto Ka are stored into different instruction queues, which are able to host up to 32 instructions. The design of the queues is focused on reducing energy consumption. In the first version of the design, the core is configured with a 5-instruction issue width, matching the instruction queues included, consequently, this parameter can change to supply more instructions to the execution stage.

The instruction execution is performed by different functional units compounded with a bypassing logic technique to effectively broadcast the source operands for dependent instructions. To preserve program order among the instructions in-flight and guarantee core recovery to a previous state, a 128-entry fully distributed reorder buffer is included. Finally, a group commitment mechanism is required in the back-end to restore the program order.

2.2. Post-Quantum Cryptography accelerator (PQC)

The PQC accelerator is a hardware module that accelerates the Classic McEliece (CME) Key-Encapsulation Mechanism (KEM) and specifically its encryption and decryption functions. The latter form the biggest computational part of the Encapsulation and Decapsulation parts of the CME KEM respectively.

The CME KEM's functionality will be explained with an example of a client-server communication scenario. The server initially generates a public-secret key pair (PK, SK) using the Key Generation algorithm. Then the client, which holds the server's PK, feeds it to the Encapsulation algorithm to produce a session key in plain text and encrypted forms. Finally, the server receives the encrypted session key, and decrypts it using its own SK and the Decapsulation algorithm. Upon successful completion of this protocol, both the client and the server have established a common session key in a secure way and they can go on communicating via symmetric cryptographic algorithms (e.g. AES).

The PQC's function therefore is to accelerate the encryption of a session key into a cipher text by one part of the communication pair that is later decrypted by the other part into the session-key using the decryption algorithm. Figure 2 presents the layout of the PQC.

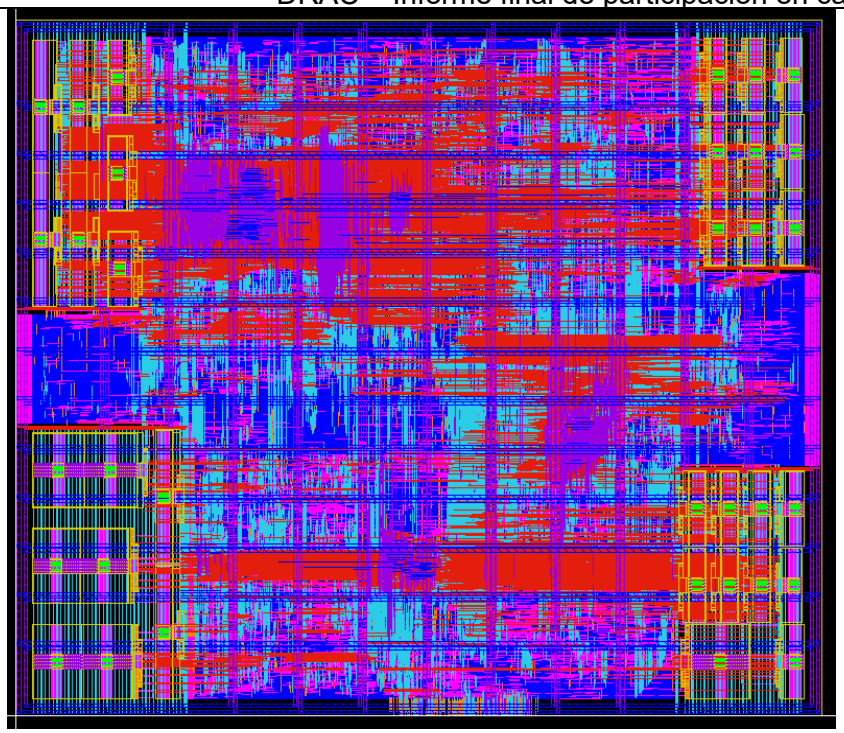


Figure 2: Layout of the PQC, with a size of 1.03 mm x 0.91 mm.

2.3. Systolic Array tensor Unit for aRtificial Intelligence Acceleration (SAURIA)

Autonomous navigation or autonomous driving is one of the key challenges of today, spanning many different fields of research, from sensorics to machine learning and computer architecture. Among the different tasks that an autonomous driving pipeline has to solve, the most computationally intensive is the Perception task: obtaining an understanding of the vehicle's surroundings in order to act accordingly, typically from one or multiple cameras mounted in the vehicle. At the algorithmic level, the state-of-the-art models for perception are based on Deep Neural Network (DNN) architectures, which are able to achieve high inference accuracy at the cost of computational complexity. In this setting, the use of hardware accelerators is key for deploying DNNs in real-world applications.

SAURIA is the low-power, high-energy-efficiency solution for accelerating DNN workloads in the Kameleon SoC. Its design revolves around a systolic array architecture, which exploits parallelism, data reuse and pipelining in order to efficiently compute general matrix-matrix multiplications (GeMM). For the purpose of executing convolutions using the systolic array as a GeMM engine, a novel Data Feeder architecture has been implemented, which converts convolution tensors into matrices (known as the im2col transformation).

In order to maximize energy efficiency, approximate arithmetic circuits have been employed in the systolic array processing elements. By studying and comparing different approximate architectures, a combination of approximate multiplier and approximate adder that reduces power consumption by 30% has been found without any significant degradation in the performance of the YOLOv3 object detection DNN. The SAURIA accelerator, working at 500 MHz, has a peak throughput of 128 GFLOP/s and a peak energy efficiency of 1.41 TFLOP/sW, being more efficient than state-of-the-art accelerators with FP16 arithmetic representation. Figure 3 presents the layout of SAURIA.

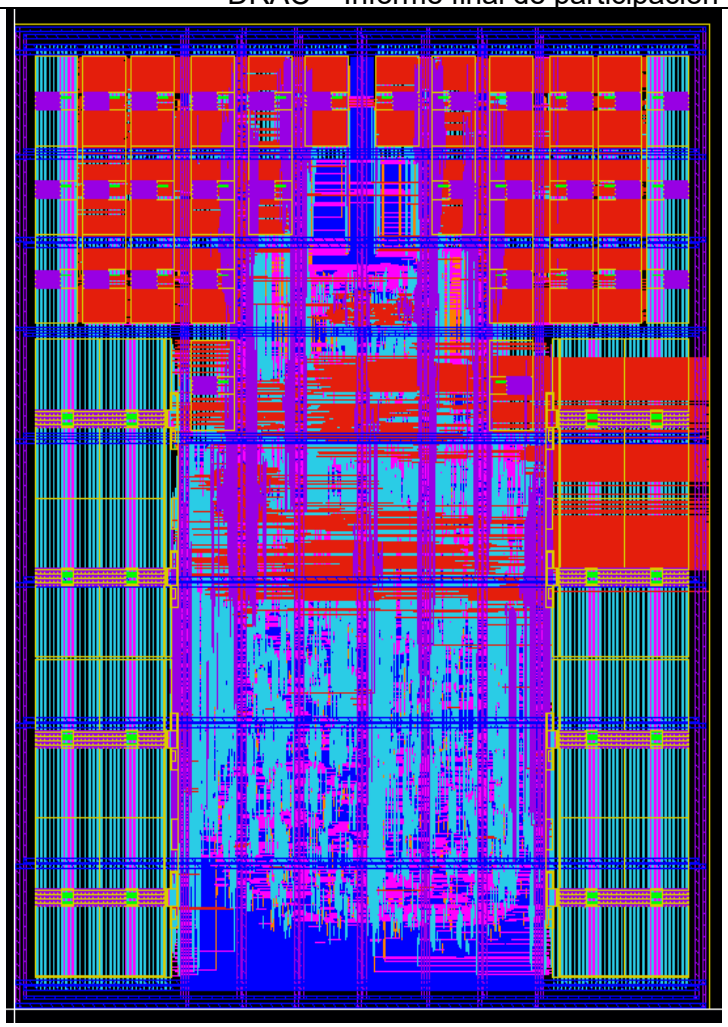


Figure 3: Layout of SAURIA, with a size of 0.84 mm x 1.05 mm.

2.4. Picos

Picos is a hardware implementation of a task-based programming model run-time (e.g. Gomp for OpenMP or Nanos6 for OmpSs-2) developed at the BSC programming models group. The main objective is to reduce the run-time overhead by accelerating task scheduling (including dependence resolution) and task synchronization (taskwait).

Tasks can be executed on the CPU (Sargantana/Lagarto Ka) and on any of the accelerators (WFA, PQC and SAURIA). CPU tasks can be any piece of code, the declaration of this code depends on the programming model. For example, OmpSs-2 tasks are declared with the "#pragma oss task" C/C++ pragma. For the accelerators, a task is an execution of the implemented application, with parameters given by the user through the use of a custom API.

Picos is able to manage and control the execution of all accelerators at the same time, including CPU tasks. Moreover, dependencies can be declared independently of the target (accelerator or CPU). For instance, a PQC execution can depend on one or multiple CPU tasks, or WFA/SAURIA tasks. Figure 4 presents the layout of Picos.

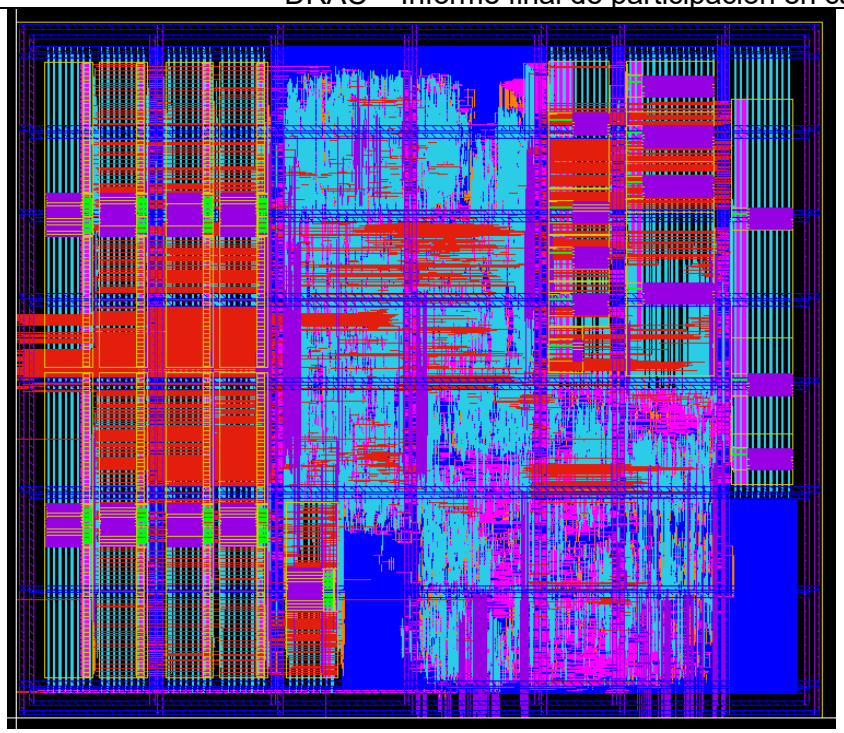


Figure 4: Layout of Picos, with a size of 0.69 mm x 0.63 mm.

2.5. Wavefront alignment algorithm (WFA)

WFA ASIC accelerator is the first accelerator for exact pairwise alignment of long DNA sequences based on the Wavefront Alignment Algorithm. It supports sequence lengths up to 10K bases and error rates up to 10%. Unlike the traditional Smith-Waterman (SW) based alignment algorithms which run in $O(n^2)$ time proportional to the sequence length n, the WFA runs in $O(n \cdot s)$ time, proportional to the sequence length and the alignment score s.

WFA is integrated in a Linux-capable RISC-V processor chip. The accelerator runs as an independent process in parallel to other CPU processes. The CPU communicates with the accelerator through the AXI lite bus. The accelerator has its own address and implements memory mapped registers inside. The CPU configures the accelerator by writing into its registers. The accelerator start and stop signals are also communicated with the CPU through the AXI lite bus. The CPU triggers the start of the accelerator and the accelerator signals the completion of a computation via dedicated interrupts, and pooling is also supported. WFA have direct access to the off-chip main memory through the memory controller which communicates via AXI full bus connected to the DMA of the accelerator.

Our accelerator, based on the evaluations on an FPGA prototype, provides performance improvements up to 1076× compared to the WFA implementation on the chip's CPU. After Place and Route (PnR), it fits in an area of 1.75 mm^2. Figure 5 presents the layout of WFA.

Having a genomics accelerator inside the CPU chip, leverages the chip functionality for analyzing genomics data. This makes the chip a perfect platform suitable for genomics applications by eliminating the need of costly external accelerators and their communication complexities.

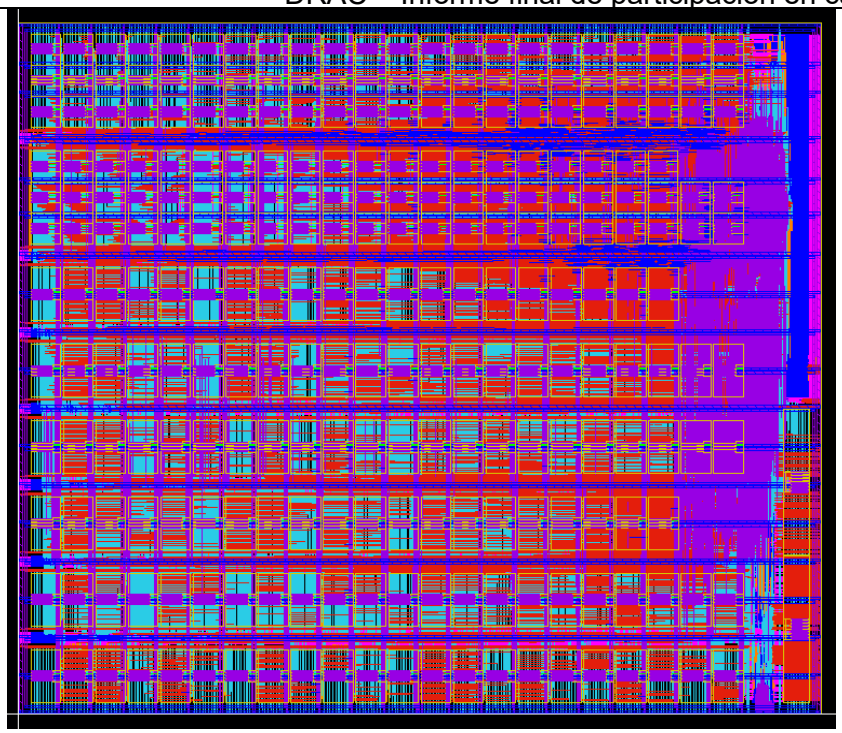


Figure 5: Layout of WFA, with a size of 1.38 mm x 1.27 mm.

2.6. Quad SerDes

The purpose of this component is to provide four lanes of high-bandwidth serial communication that are able to run at 4 and 8Gbps, providing a point-to-point interface between the DRAC-ASIC, containing the Kameleon Core and I/O peripherals; and the Xilinx KC705 FPGA board. Each lane features a parallel user data interface of 32-bit running at 125 MHz.

The architecture of a single SerDes consists of a serial transceiver. The transmitter serializes 32-bit user data from the core at 4Gbps using a multiplexer network. The serial signal is then converted to analog, filtered and amplified. A PLL generates the 4GHz synchronization clock using a 125 MHz clock reference from a programmable oscillator integrated into the ASIC. This PLL also provides a 250 MHz (clk_TX_fast) and 125 MHz (clk_TX_slow) clocks that are used in the serialization and sent to the FPGA Receiver in order to have the same transmission reference.

The SerDes Receiver receives an analog data stream coming from the FPGA TX, then converts it to digital serial data, which afterwards it deserializes, to do so a clock and data recovery system extracts the main clock frequency from the signal (4 GHz) and the 250 MHz (clk_RX_fast), 125 MHz (clk_RX_slow) clocks to synchronize the registers that perform the 32-bit deserialization. Once the CDR locks into the main frequency a “locked” signal is sent to the core in order to start receiving valid data. Figure 6 presents the layout of Quad SerDes.

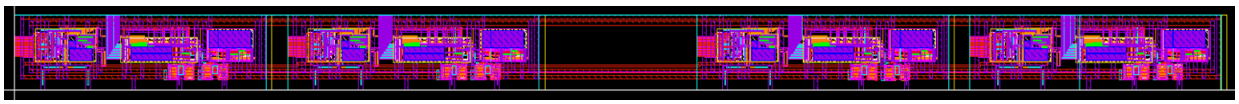


Figure 6: Layout of Quad SerDes, with a size of 0.01 mm x 1.37 mm.

2.7. PLL

The PLL is a module which can generate a high internal frequency clock for the SoC from a lower frequency external clock. This allows the CPU to have a working clock frequency much higher than what can be inputted from an external reference. This PLL is capable of generating any clock frequency between 1.2 GHz and 2.4 GHz in steps of the reference clock frequency, which may then be divided by any value between 2 and 32. Figure 7 presents the layout of the PLL.

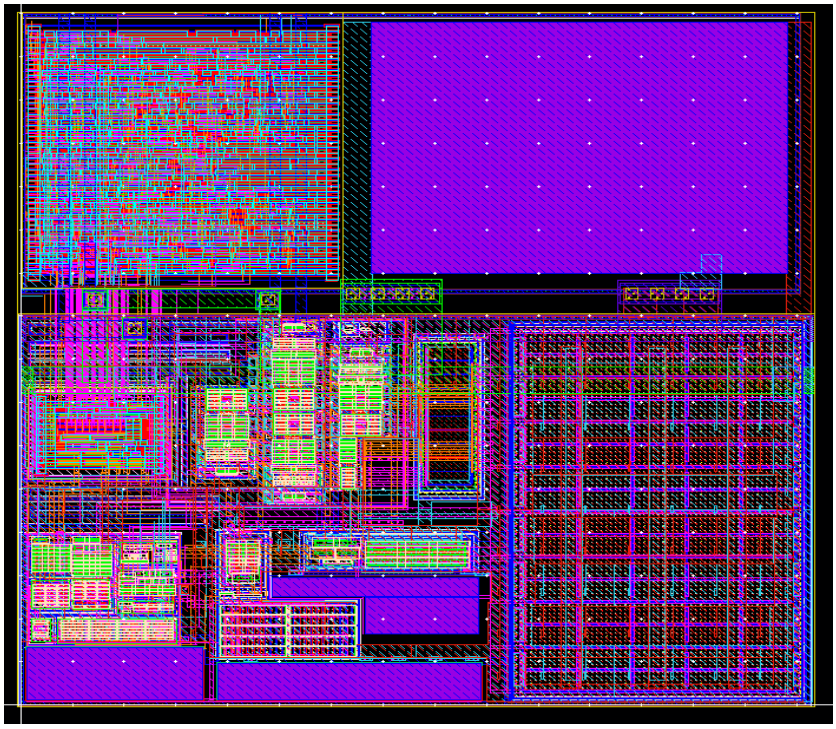


Figure 7: Layout of the PLL, with a size of 118 μm x 80 μm.

The wide range of frequencies the PLL provides allows the SoC to autoscale its work frequency, making it able to change it on the fly depending on different parameters such as system load, temperature or power limits.

The goal is to generate a clock signal which is as stable as possible. In order to do so, different techniques have been implemented. Firstly, in order to minimize interference (injected noise) from power and polarization internal circuits, the PLL incorporates a linear regulator which isolates it from the power source. Secondly, the design is based on a current controlled ring oscillator which also minimizes noise due to supply voltage oscillations. Lastly, it also possesses a second order control system, which further ensures a minimal error in the generated clock frequency.

3. Physical Synthesis

3.1. IPs Placement

The IPs have been placed in the floorplan, in an area of 3 mm x 3 mm, as presented in Figure 8:

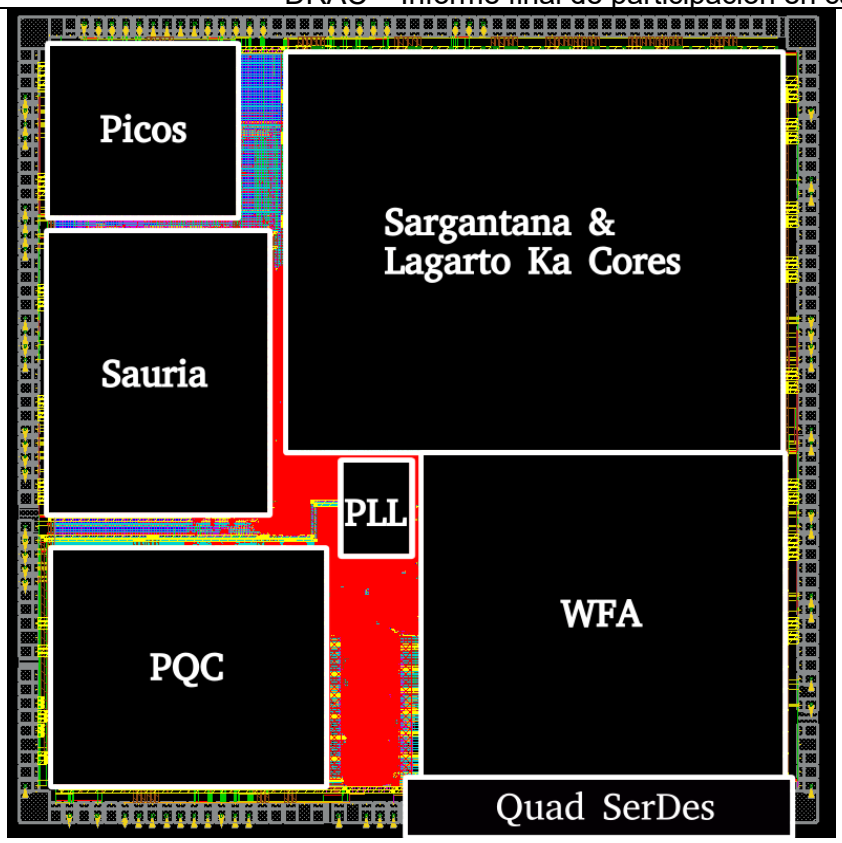


Figure 8: Kameleon IP placement in a 3 mm x 3mm area.

The criteria for the placement of the IPs is the following:

3.1.1. Sargantana & Lagarto Ka cores

Such IP is the most restrictive and it is the one driving the placement of the other IPs. Once timing requirements were met, the area of the cores was fixed and allowed to fix the other IPs placement.

3.1.1. PLL & SPI configurator

The PLL block contains 2 modules, the PLL itself and the SPI configurator. The PLL is in charge of generating the reference clock for the system and the SPI configurator is used to configure the PLL and SerDes functions. The PLL block is placed between the cores IP and the SerDes IP, leaning closer to the cores IP because the proximity of the clock signal takes priority over the proximity of the SerDes configuration bus.

3.1.2. Quad SerDes

The Quad SerDes has been designed to be placed in contact with the Padring. The side which was chosen for its placement is the bottom. At the same time, since part of the SPI configurator, which is inside the PLL block, is used to configure the SerDes, we knew the PLL pins would have to be placed close to the SerDes pins, so the SerDes has been placed to the far right of the bottom in order to have the rest of the bottom of the PadRing available for the PLL block's pins.

3.1.3. Accelerators (Picos, Sauria, PQC and WFA)

These IPs have been placed in the rest of the available space, with the main criterion of reducing the length of the interconnections.

4. Verification: Simulations and static timing analysis

After the physical synthesis process, different verifications must be carried out to certify that the synthesis process has been correct and that all the circuits continue to work correctly.

First, the correct functioning of the design must be validated by simulation; as in the first tapeout, this task is performed by the BSC. In order to carry out this simulation, the UB has supplied the results of the synthesis together with the files that contain the delays of the logic gates to the BSC.

Second, the static timing analysis (STA) has been carried out: during the physical synthesis 3 temporary conditions have been imposed, that the chip can work at a frequency of 600 MHz in the worst case, 800 MHz in the typical case and 1000 MHz in the best case. The reports, shared with all the partners via the B2DROP platform (in EPI01/Kameleon/PnR_results), show that timing is only met in the typical case, meaning that in FFG the chip must run at 800 MHz as well, and in SGG the chip must run below 600 MHz. Due to the lack of time to properly study the reason for having such results, we decided to move forward and to send the chip to the foundry as no hold violations were found.

5. Conclusions

The second chip of the DRAC project has been sent to manufacture on December 12, 2022. Researchers from the BSC, UPC, IPN, IMB-CNM, URV, UAB and UB have participated in this chip. The chip occupies a total area of 9 mm^2, and has been designed on GlobalFoundries' 22nm technology. Apart from the Sargantana & Lagarto Ka processors, different IPs have been integrated in order to increase the functionality and capabilities of the Kameleon chip.