Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Vitruvius+: An Area-Efficient RISC-V Decoupled Vector Accelerator for High Performance Computing

Francesco Minervini

May, 4th, 2022

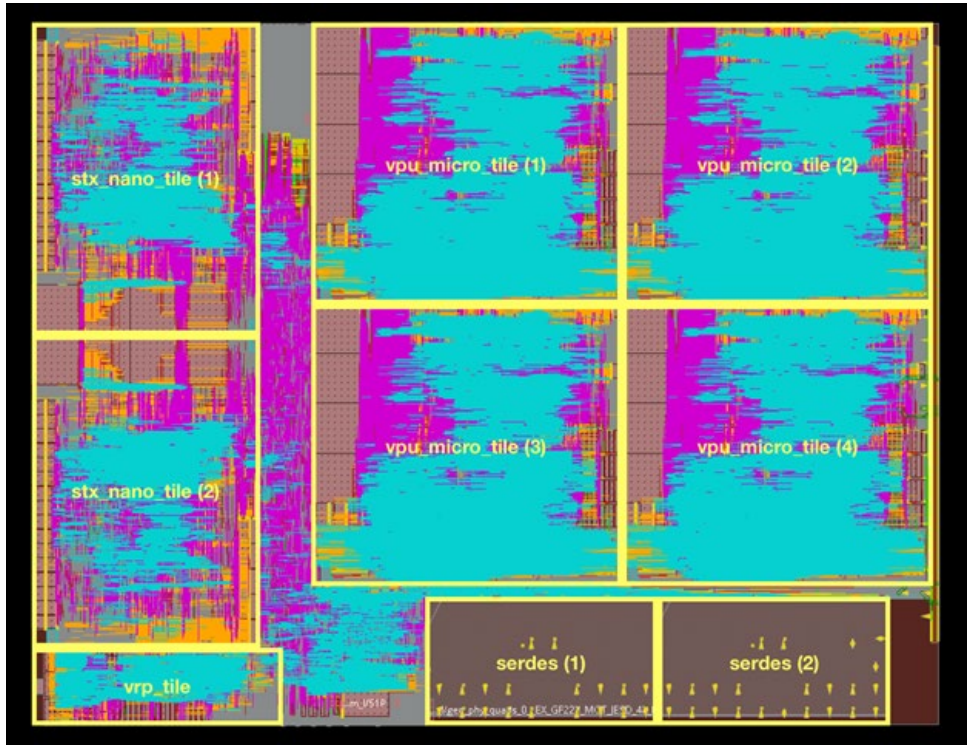
Spring 2022 RISC-V Week

Agenda

- Introduction
- Microarchitecture
- Evaluation
- Future Plan
- Conclusion

Introduction

EPI Project



EPAC layout highlighting the accelerators with 25 mm² in GF 22FDX technology (https://www.european-processor-initiative.eu/epi-epac1-0-risc-v-test-chip-taped-out/)

- Vitruvius+ is a key component of the **European Processor Initiative (EPI)**, a project co-funded by the European Union
- Aims to design and implement a roadmap for a new family of low-power Exascale European processors
- First phase concluded with a test-chip taped out in June 2021 using GLOBALFOUNDRIES 22FDX® 22nm FD-SOI running at 1 GHz
- Vitruvius+ is part of the first tapeout in the second phase of EPI

European Processor Initiative

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826647

Microarchitecture

Why Vitruvius+

- Vitruvius+ is an improved version of Vitruvius, the vector execution core of the first phase of EPI, presented at the RISC-V Summit 2021 [1]
- Vitruvius+ inherits all the main characteristics of Vitruvius:
  - Implements version 0.7.1 of the RISC-V vector extension (RVV)
  - Long vectors, up to 256 DP-elements
  - Increases vector length up to 2048 DP-elements when grouping 8 vectors (*LMUL*=8)
  - Support mixed width vector operations, namely widening and narrowing
  - Decoupled architecture
  - Lightweight out-of-order execution
  - Vector register renaming
  - Vector instructions overlapping
  - Multiple accumulators enhancing reduction operations

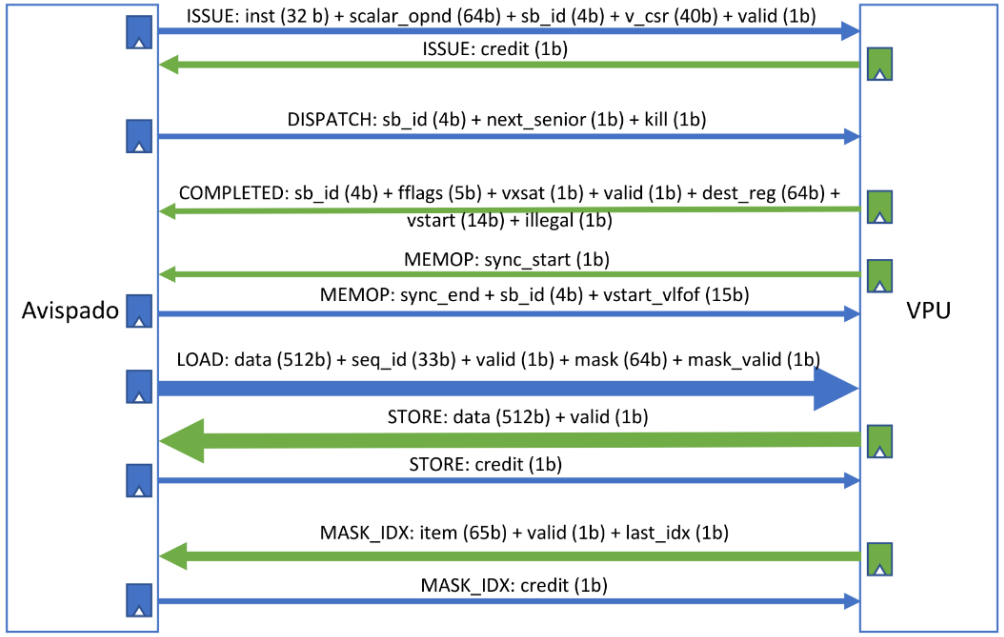
[1] F. Minervini, O. Palomar. “Vitruvius; An Area-Efficient RISC-V Decoupled Vector Accelerator for High Performance Computing”, RISC-V Summit 2021, https://youtu.be/tlC5kMhrh-k

Vitruvius+ New Features

- Vector memory-to-arithmetic instruction chaining
- Tree-based reduction algorithm to further improve performance
- Enhanced memory units to manage more than one in-flight memory operations
  - Up to three vector strided loads and stores
  - Up to one masked/indexed memory instruction at a time
  - Two additional strided loads or stores can be handled while executing a masked/indexed memory operation
- Unidirectional ring inter-lane interconnect with limited reconfiguration
- Completely configurable design
  - independent vector lanes
  - variable vector length
  - parameterized functional unit pipeline depths

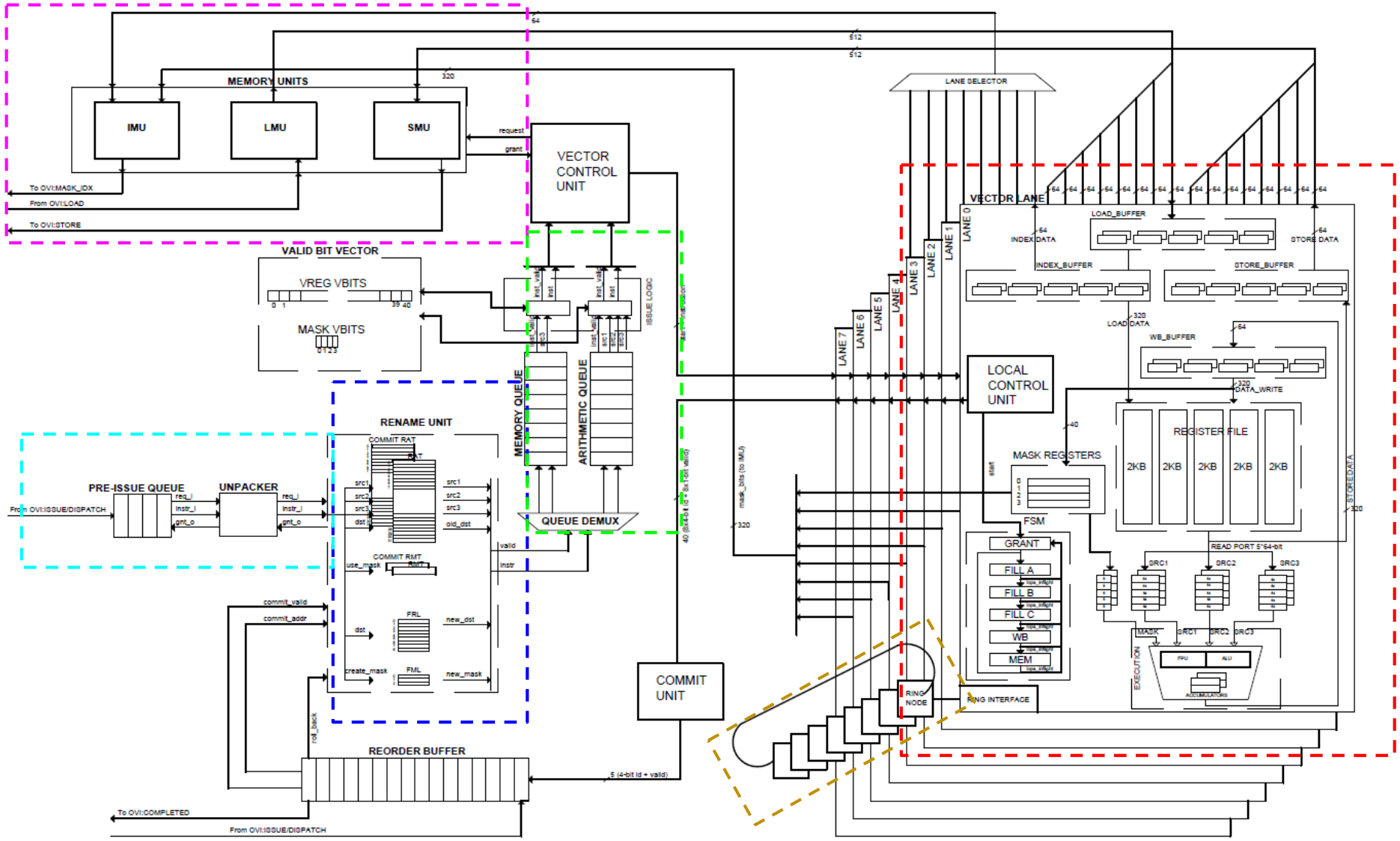
Execution Paradigm

- Vitruvius+ adopts a **hybrid in-order/out-of-order** execution scheme
  - Arithmetic instructions proceed in-order
  - Memory instructions can execute out-of-order
- Vitruvius+ is a **decoupled accelerator**
  - Offload vector instructions from the scalar pipeline
  - Only vector memory instructions need the scalar core and the vector accelerator to effectively interact
- Communication with the scalar core is possible through the **Open Vector Interface (OVI)**



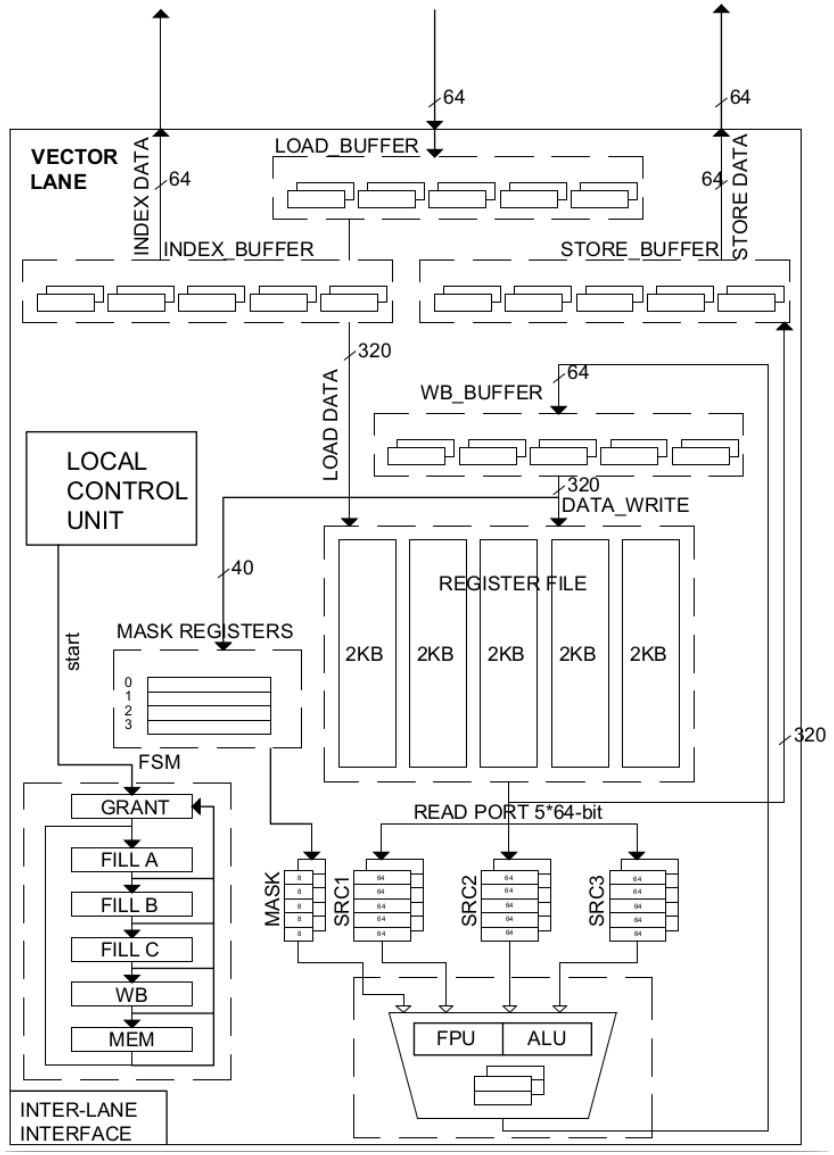
An overview of the OVI, the interface resulting from the joint effort between Semidynamics and the BSC (https://github.com/semidynamics/OpenVectorInterface).

Vitruvius+ Block Diagram

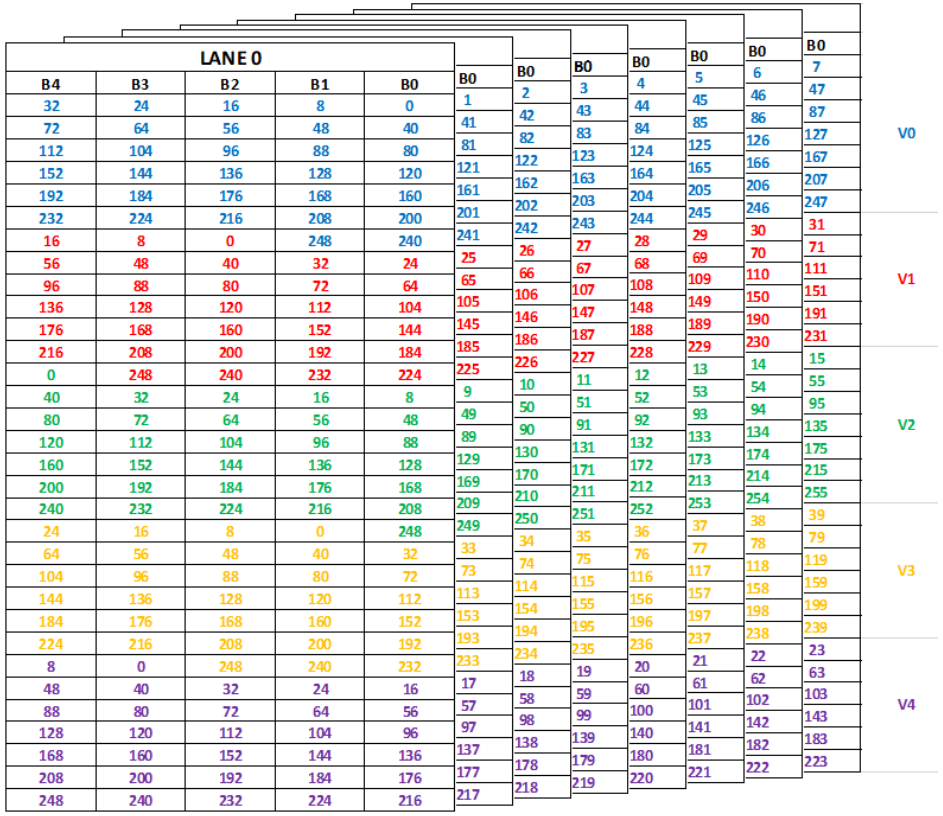


Lane Architecture

- 8 vector lanes in the baseline configuration, connected through an area-efficient ring interconnect
- A local FSM orchestrates intra-lane data movements
- A control unit selects the active/inactive elements
- Separated mask registers from the register file to avoid read conflicts on predicated instructions
- One ALU for all integer and logical operations
- One FPU for floating-point operations
- Dedicated vector reduction logic

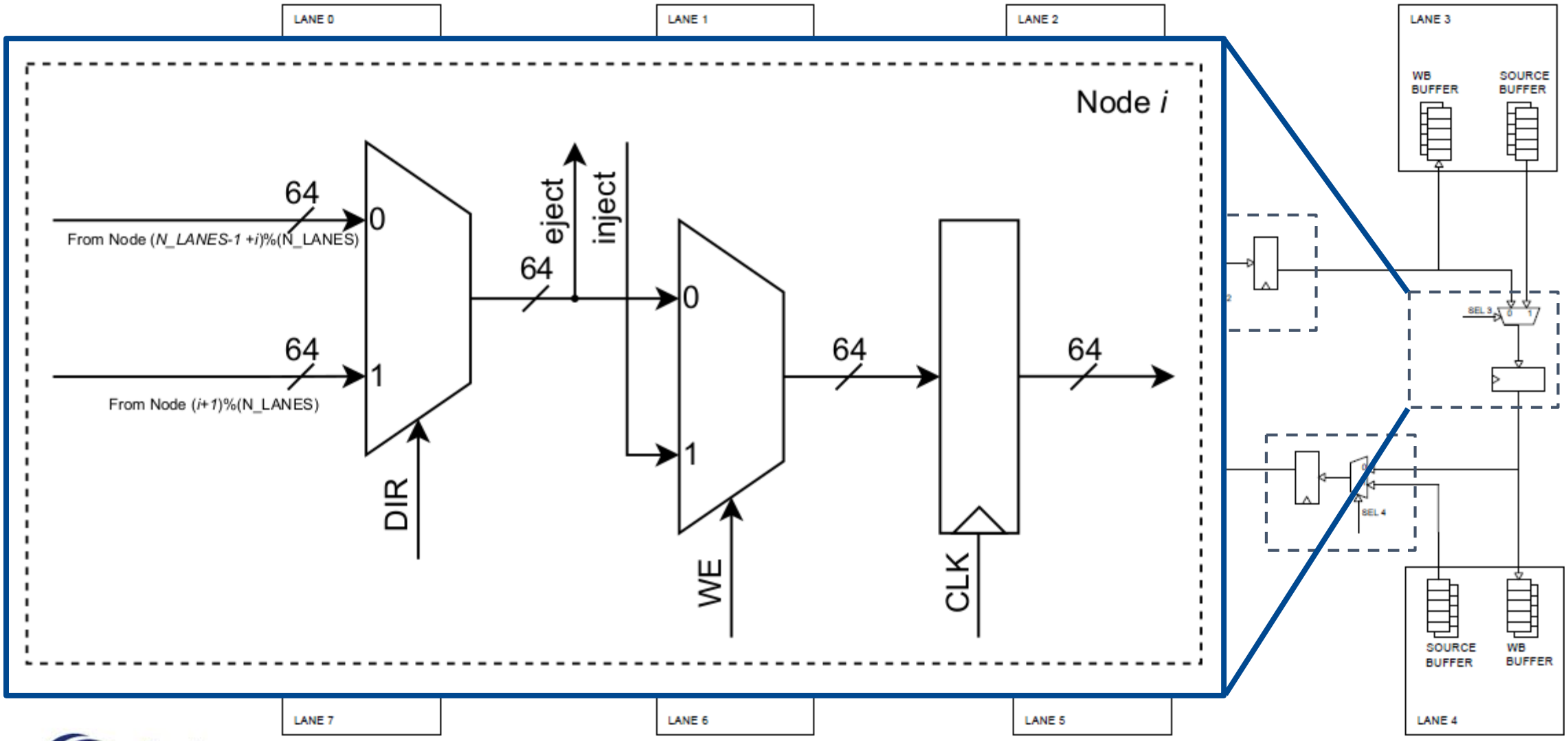


Vector Register File Organization



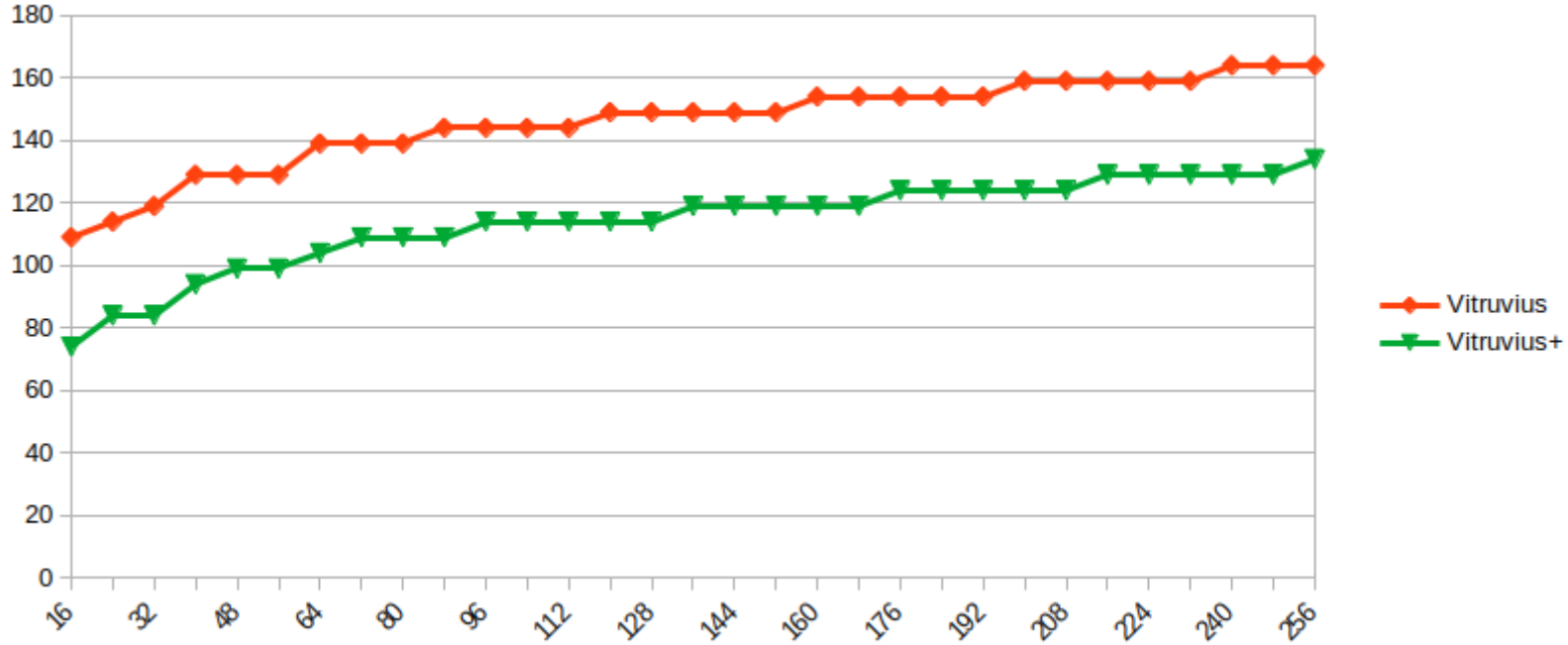
- Interleaved vector registers across the lanes, each one holding 5 SRAM banks
- Each SRAM bank is 2 KB:
  - Accounts for 10 KB of SRAM per lane
  - Instantiates 80 KB for the vector register file (VRF) in the 8 lane configuration
  - Allocates space for the 32 architectural registers and 8 additional renaming registers
  - Each register holds a maximum of 256 64-bit elements
- Vector registers are contiguously stored in the banks
- A read always tries to get a full row from the banks
- A write can store whatever number of elements

Inter-lane Interconnect



Reductions Enhancement

- Vitruvius+ introduces an additional optimization on the execution of vector reductions
- Apart from implementing multiple accumulators for the intra-lane reduction phase, it proceeds through inter-lane tree-based algorithm to parallelize arithmetic operations



Evaluation


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Synthesis Setup

- Vitruvius+ was synthesized for GLOBALFOUNDRIES 22FDX® 22nm FD-SOI
- Target frequency for the standalone synthesis was 1.4 GHz

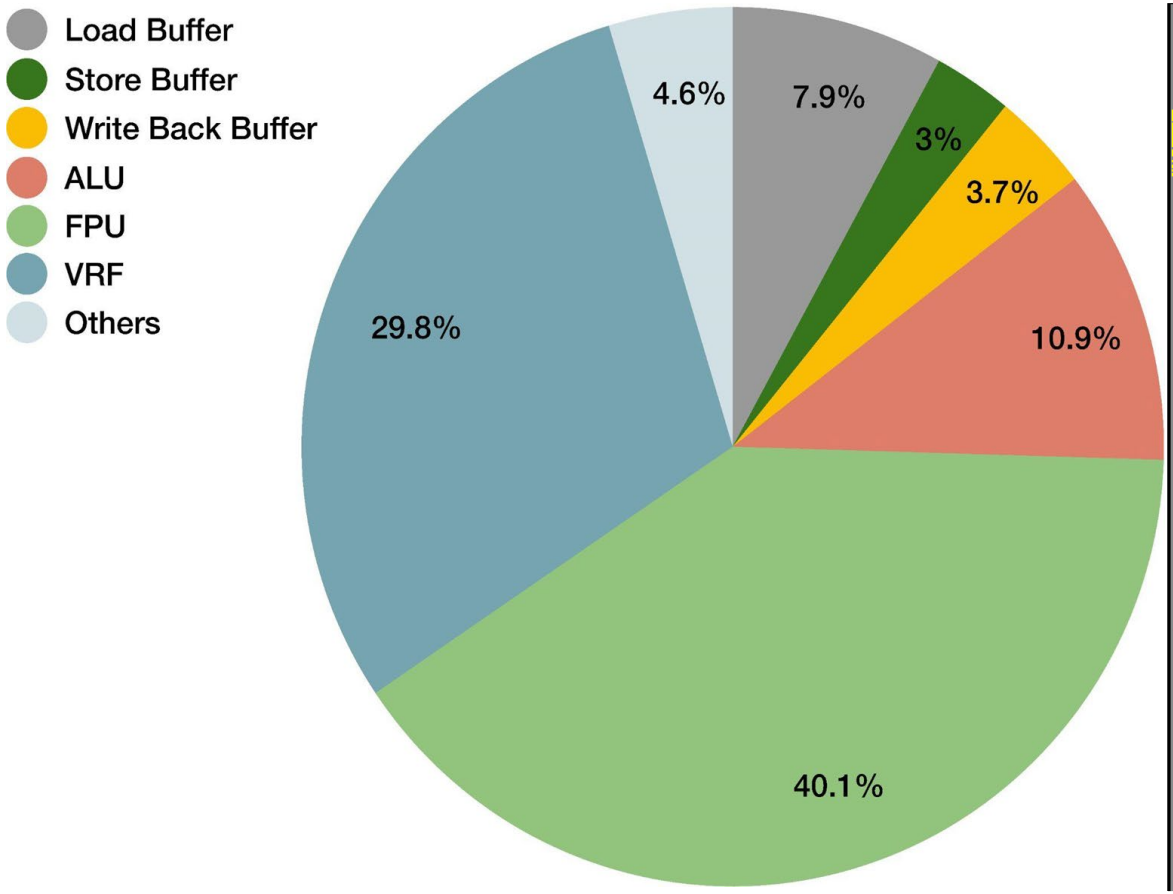
Configuration	
Parameter	**Value**
Number of lanes	8
VRF size	10 KiB/lane
Number of banks	5 banks/lane
Bank width	64 bits
Number of ports	1 port/bank
Number of slots	256 slots/bank

Result	
Parameter	**Value**
Area	1.3 mm²
Max frequency	1.4 GHz

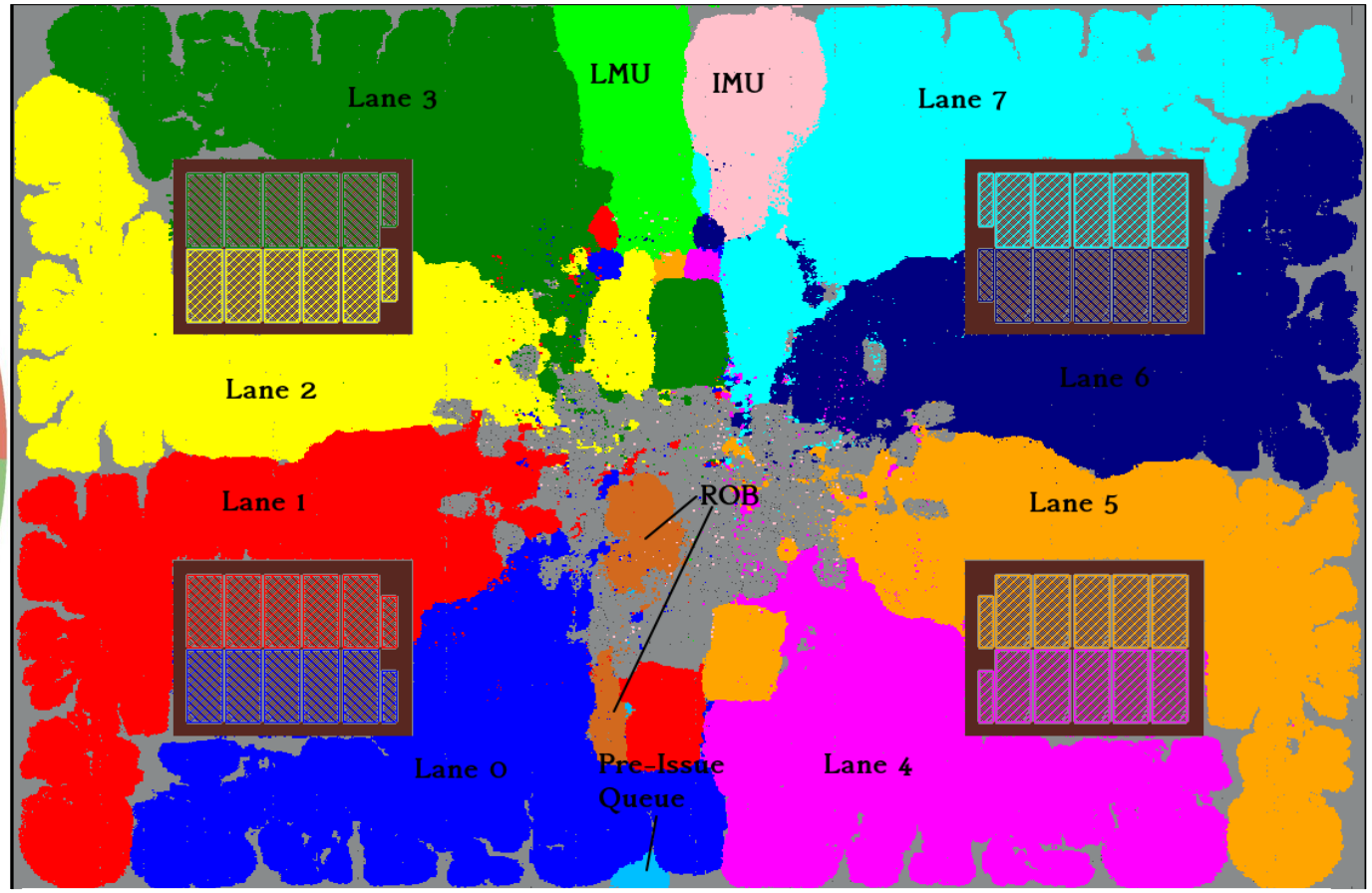
Results reported for a synthesis run using typical conditions (TT, 0.80 V, 25 °C)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Physical Design



Area breakdown for the single lane. The FPU occupies most of it.



Layout resulting from the place-and-route of an instance of Vitruvius+ with 8 lanes.

Vectorized Benchmarks

- Several vectorized benchmarks were used to characterize Vitruvius+
- We used problem sizes that are beneficial for our vector unit

Benchmark	DP-FLOP/cycle	DP-GFLOPS (@1,4GHz)	Speed-Up on Vitruvius
Matmul 256x256	15.5	21.7	1.02X
Jacobi-2D	8.2	11.5	1.1X
Black-Scholes	6	8.4	1.17X
LavaMD	6.6	9.24	1.12X
Pathfinder	3.0	4.2	1.16X
Streamcluster	7.1	9.94	1.8X

Future Plan

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Vitruvius+ successor

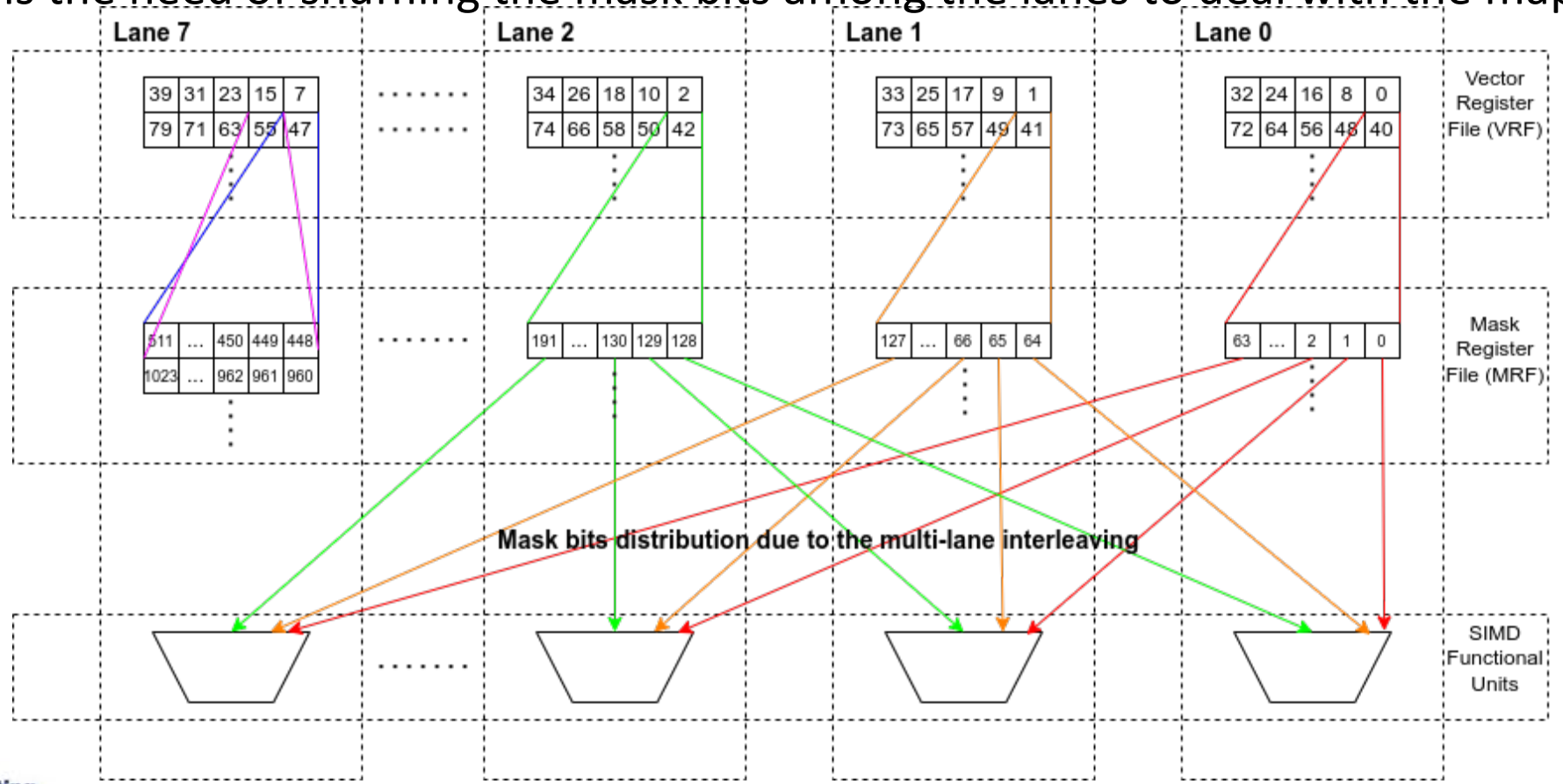
- EVA (Enhanced Vitruvius Architecture), Vitruvius+ successor, is the vector accelerator of the second phase of EPI
- EVA will be designed with the following features in mind:
  - Implements the RISC-V V-extension release 1.0
  - New scalar core interface (OVI 2.0) being ratified
  - Higher performance inter-lane interconnect
  - Higher clock frequency (2.0 GHz)
  - New fabrication technology (GF12LP) in the EPI project
  - Explore different VRF configurations

Towards RVV-1.0

- RVV-1.0 introduce novel features for the vector architecture specifications:
  - New mask layout. Simplified mask bits mapping to the vector register v0 (mask bit *n* matches bit *n* of v0)
  - Vector tail agnostic and mask agnostic. Dedicate bits *vta* and *vma* in the CSR *vtype* to control the behavior of tail elements and inactive masked-off elements, respectively
  - Fractional LMUL (LMUL<1). Reduces the number of bits used in a single vector register and increases the effective number of vector register groups when operating on mixed-width values
  - Memory operations variants. Allow different data and indices element sizes, include whole register loads/stores

Mask Layout Challenges

- The new mask layout, while simpler than in previous versions, is problematic for EVA
- There is the need of shuffling the mask bits among the lanes to deal with the mapping



Conclusion


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Remarks

- Vitruvius and Vitruvius+ are the first of a family of vector accelerators developed at BSC
- Vitruvius+ is the only RISC-V vector processor that supports long vectors (256 DP-elements per vector, up to 2048 DP-elements for the higher *LMUL* configuration)
- Vitruvius+ is the first RISC-V vector processor compliant with the OVI specifications
- EVA, Vitruvius+ successor, will fully support RVV-1.0 with the same maximum vector length, and will run at higher frequencies
- This work shows that a long-vector accelerator can be designed following an area-efficient approach

Acknowledgement

We express our gratitude to all the EPI partners

EPI FUNDING

This project has received funding from the European High Performance Computing Joint Undertaking (JU) under Framework Partnership Agreement No 800928 and Specific Grant Agreement No 101036168 EPI-SGA2. The JU receives support from the European Union's Horizon 2020 research and innovation programme and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland. The EPI-SGA2 project, PCI2022-132935 is also co-funded by MCIN/AEI /10.13039/501100011033 and by the UE NextGenerationEU/PRTR.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you

Questions?

More details: francesco.minervini@bsc.es