KALRAY

THE POWER OF MORE

POSIT ARITHMETIC IN THE EUROPEAN PROCESSOR INITIATIVE

CoNGA 2023 Keynote

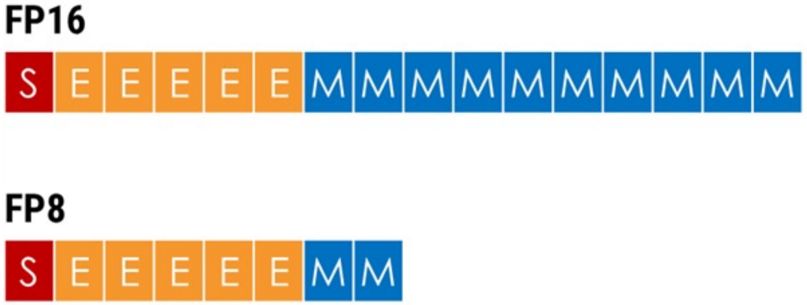
Benoît Dupont de Dinechin

www.kalrayinc.com

IEEE 754 BINARY FLOATING-POINT AND POSIT REPRESENTATIONS

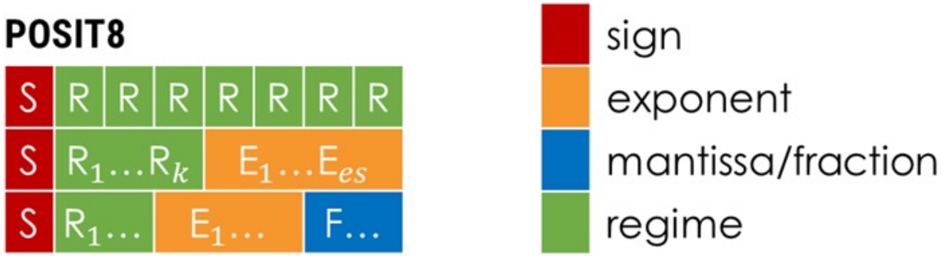
IEEE 754 binary floating-point encode a value x with sign S, exponent E, and explicit mantissa M

- $$x = \begin{cases} (-1)^S \times 2^{E-B}(1 + \sum_{i=1}^{m} \frac{M_i}{2^i}) & \text{if some } e \text{ bits differ} \\ (-1)^S \times 2^{1-B} \sum_{i=1}^{m} \frac{M_i}{2^i} & \text{if all } e \text{ bits are } 0 \\ (-1)^S \ \textbf{Inf} \text{ or } \textbf{NaN} & \text{if all } e \text{ bits are } 1 \end{cases}$$
- The bias B is set to $B = 2^{e-1} - 1$ with e the number of bits of E



Posit encode a value x with sign S, regime R, exponent E, and fraction F

- $$x = \begin{cases} 0 & \text{if } S = 0 \text{ and all other bits are } 0 \\ \textbf{NaR} & \text{if } S = 1 \text{ and all other bits are } 0 \\ ((1-3S)+F) \times 2^{(1-2S)\times(2^{es}\times R+E+S)} & \\ & \text{otherwise} \end{cases}$$
- The regime $R \in [-k, k-1]$ is encoded in unary as k identical bits
- Bits of exponent E and fraction F can be missing and are assumed 0
- The maximum number of bits to encode E is given by $es = 2$ in the current Posit standard
- Posit<n, es> products accumulate into a quire represented as a $16n$ bit 2's complement binary number



AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

DII, Università di Pisa

Kalray SA

Outlook

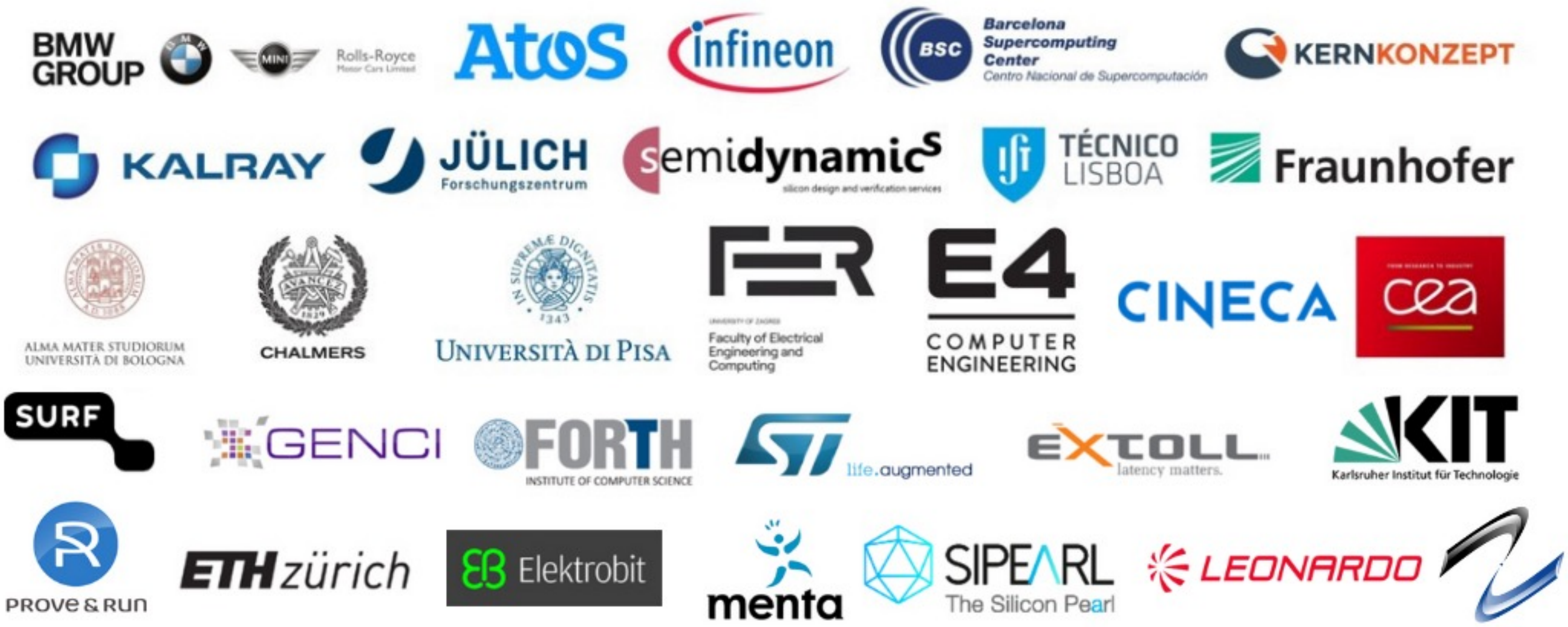
EUROPEAN PROCESSOR INITIATIVE SGA1

The European Processor Initiative (EPI) is an ambitious programme to develop low-power microprocessors for domestic supercomputers

- The EPI Specific Grant Agreement 1 was the first phase of the EPI, aiming to design and develop the first European system-on-chip and accelerator processors for high-performance computing (December 2018 − December 2021).
- Organized into three technology streams and two panstreams for integration and coordination

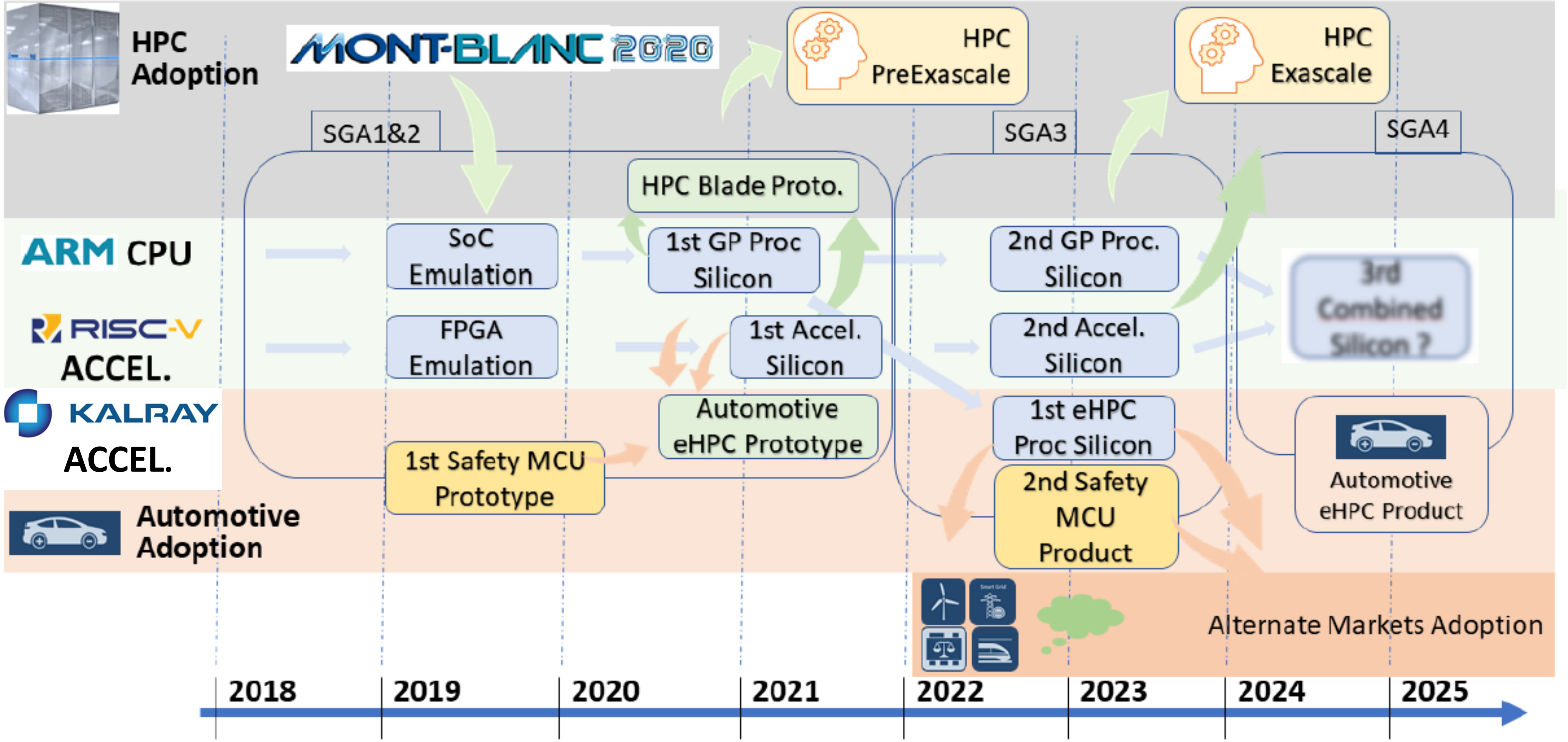
1. Global technical Panstream (CEA)
   - Co-design & Benchmarking, Global Architecture, Power Management, System Software, Simulation and Modeling
2. General Purpose Processor Stream (Atos)
   - Architecture, SDK, IP Design & Verification, Security Implementation, Chip Integration & Verification, Physical Implementation, …
3. Accelerator Stream (BSC)
   - Accelerator Architecture Specification, Accelerator System Software, Accelerator Compiler, Accelerator IPs, Test Chip and Board
4. Automotive Stream (Infineon)
   - Requirements and Specification of the eHPC platform, eHPC Platform Architecture, Automotive SDK, eHPC Pilot
5. Coordination Panstream

EPI SGA1 CONSORTIUM


BMW GROUP
MINI
Rolls-Royce Motor Cars Limited
Atos
infineon
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
KERNKONZEPT
KALRAY
JÜLICH Forschungszentrum
semidynamics silicon design and verification services
TÉCNICO LISBOA
Fraunhofer
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
CHALMERS
UNIVERSITÀ DI PISA
FER Faculty of Electrical Engineering and Computing
E4 COMPUTER ENGINEERING
CINECA
cea
SURF
GENCI
FORTH INSTITUTE OF COMPUTER SCIENCE
ST life.augmented
EXTOLL latency matters.
KIT Karlsruher Institut für Technologie
PROVE & RUN
ETH zürich
Elektrobit
menta
SIPEARL The Silicon Pearl
LEONARDO

MONT-BLANC 2020 AND EPI PROJECTS

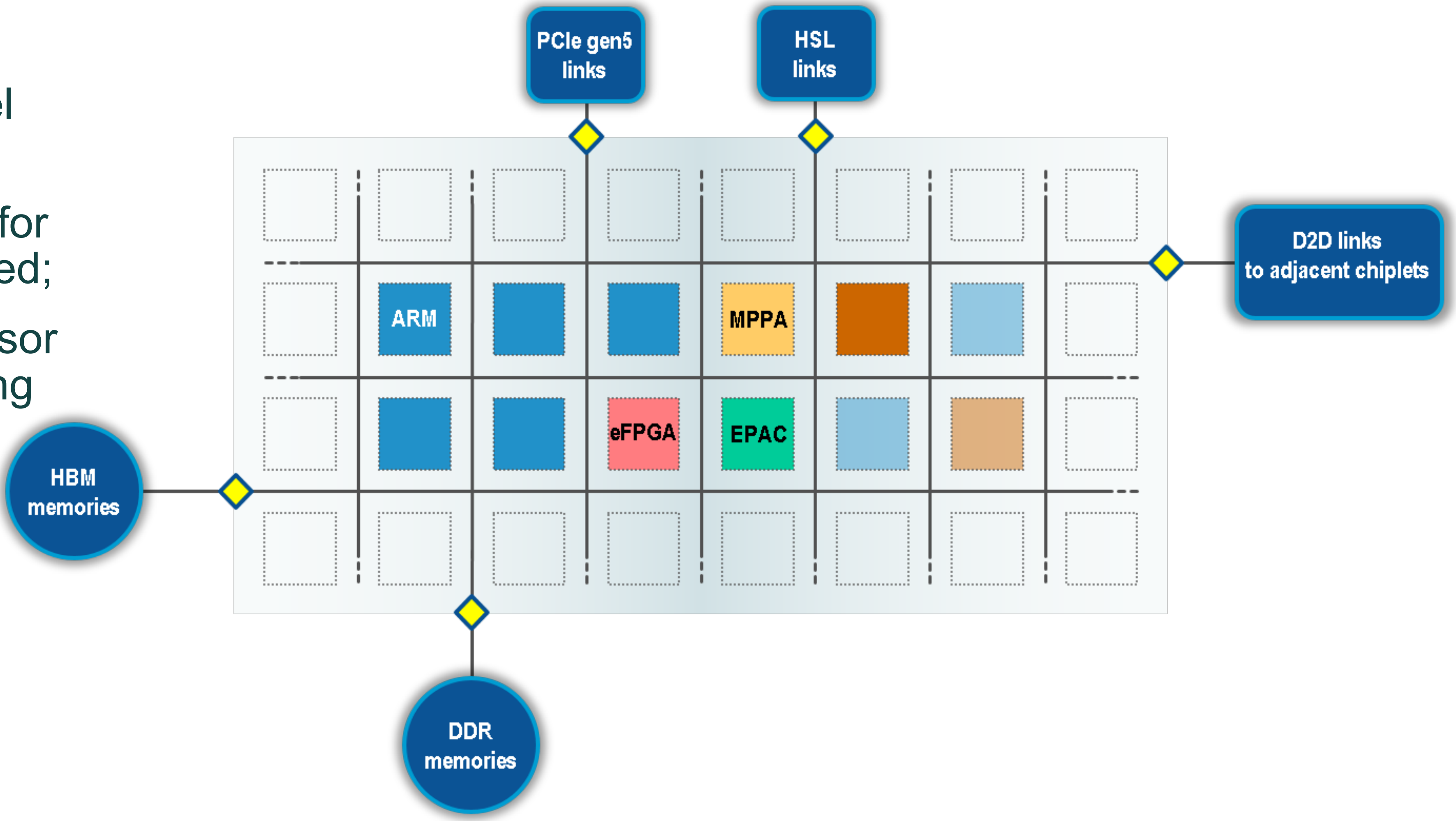
Mont-Blanc 2020 (December 2017 – March 2021) developed software and hardware IPs for EPI



STREAM 1 GLOBAL ARCHITECTURE

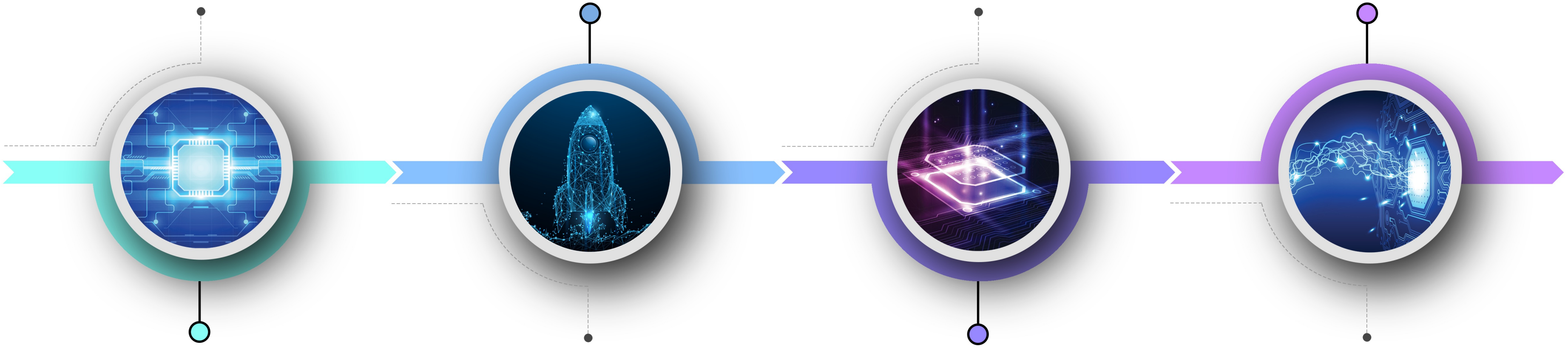
EPI common platform based on a 2D-mesh Network-on-Chip (NoC)

- ARM: general purpose application core;
- MPPA: massively parallel processor array tile;
- eFPGA: processing unit for automotive and embedded;
- EPAC: European processor accelerators implementing the RISC-V ISA.



STREAM 2 GENERAL PURPOSE PROCESSOR (ATOS)

EPI GPP are ARM-based multicore processors with a cache-coherent interconnect



2021

Rhea Family - Gen1 GPP

EPI Common Platform
ARM & RISC-V
External IPs
•
HPC System PreExascale
Automotive PoC

EPI IP's Launch Pad
&
Pan European Research Platform for HPC and AI

2021-2022

2022-2023

Cronos Family - Gen2 GPP

EPI Common Platform
ARM & RISC-V
•
HPC System Exascale
Automotive CPU

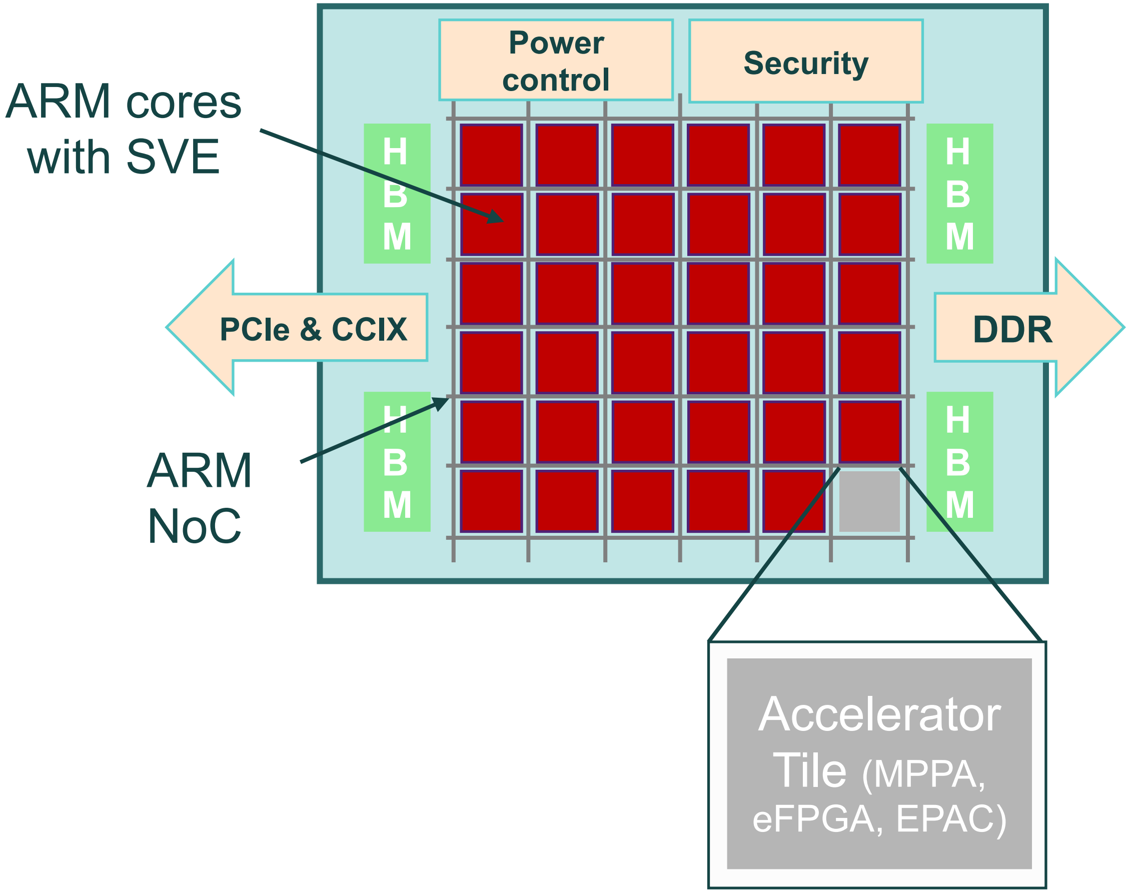
Gen3 GPP Family

2024-...

STREAM 2 GENERAL PURPOSE PROCESSOR

RHEA1 Processor (SiPearl)

Element	Features
Core	- Arm ISA - **Neoverse V1 cores** - **SVE 256** per core supporting 64/32/BF16 and Int8 - ArmVirtualization extensions
SoC	- NoC: Arm 2D mesh fabric - Advanced RAS including Arm RAS extensions - Link protection for NoC & high-speed IO - ECC support for selected memory
Cache	- **Large L3** (Shared Level Cache, **SLC**) - RAS supported for all cache levels
Memory	- **HBM2e** - **And DDR5** - ECC for memory and link protection for controllers
High Speed I/O	- **PCIe, CCIX & CXL** - Root and endpoint support
Other I/O	- USB, GPIO, SPI, I^{2}C
Power Management	- Power management block to optimize perf/Watt accross use cases and workloads.
Security Block Support	- Secure boot and secure upgrade - Crypto accelerators - True random number generation

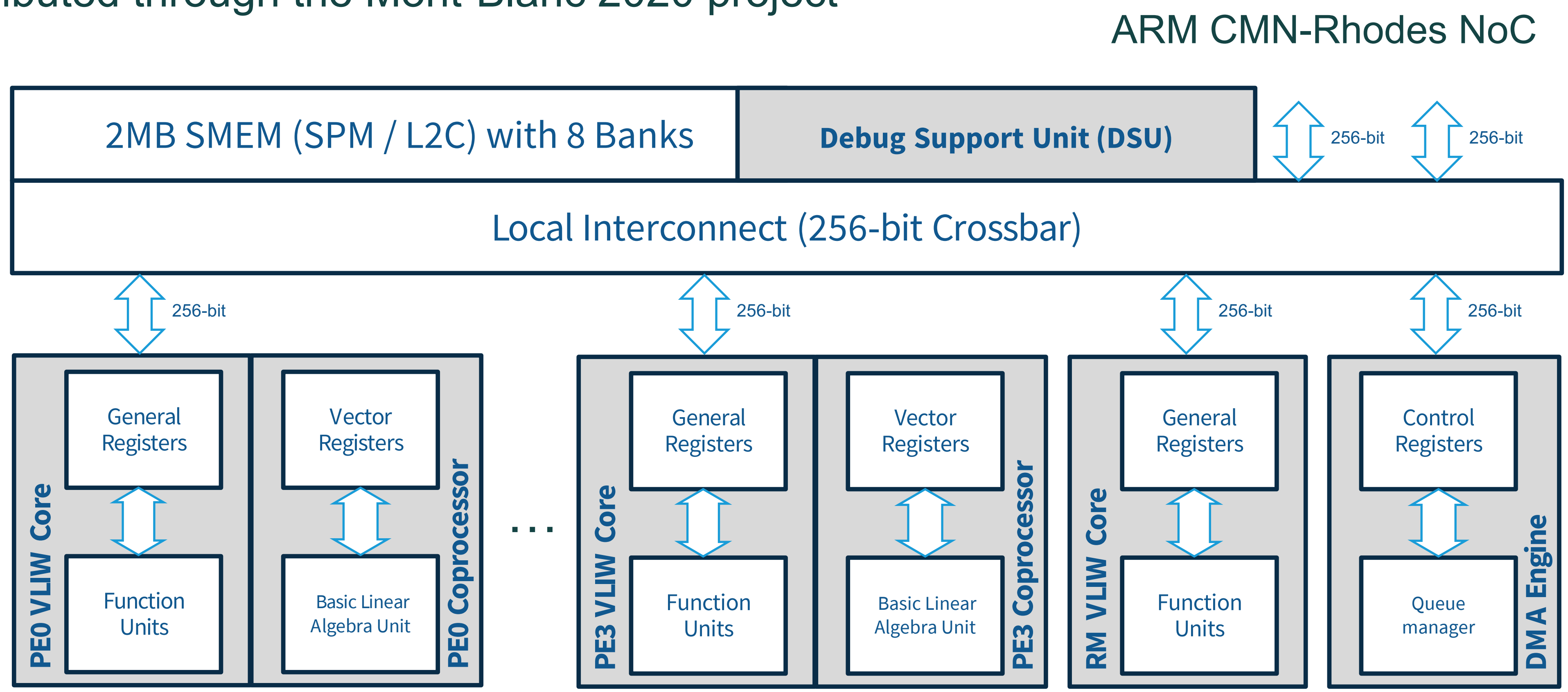


STREAM 2 KALRAY MPPA ACCELERATOR TILE

4-PE tile physical implementation for TSMC 6nm fits on 4mm^2

Interfaced to the main SoC interconnect using AMBA 4 ACE-Lite

SDK contributed through the Mont-Blanc 2020 project



STREAM 3

RISC-V Based Accelerators, integrated into the EPAC 1.0 test chip

RISC-V V-EXTENSION ACCELERATOR

RISC-V vector processing unit (VPU), designed by BSC and UniZagreb, shows the use of RISC-V long-vector architectures for high-performance computing.

The vector unit is complemented by Semidynamics’ vector-specialized Avispado RISC-V core and Gazzillion Misses™ technology for energy-efficient processing.

OTHER RISC-V BASED ACCELERATORS

Stencil and tensor accelerator (STX), designed by ETH and Fraunhofer offers exceptional energy efficiency and programmability for machine-learning and stencil workloads.

Variable precision accelerator (VRP), designed by CEA for scientific high-performance computing applications such as multiphysics simulations.

UNCORE EPAC CHIP COMPONENTS

Multiple distributed banks of shared L2 cache and coherence home nodes (L2HN) designed by FORTH and CHALMERS.

The test chip also includes high-speed network-on-chip (NoC) and advanced SERDES technology for very high-bandwidth off-chip and cross-chip communication. Both the NOC and SERDES are designed by Extoll.

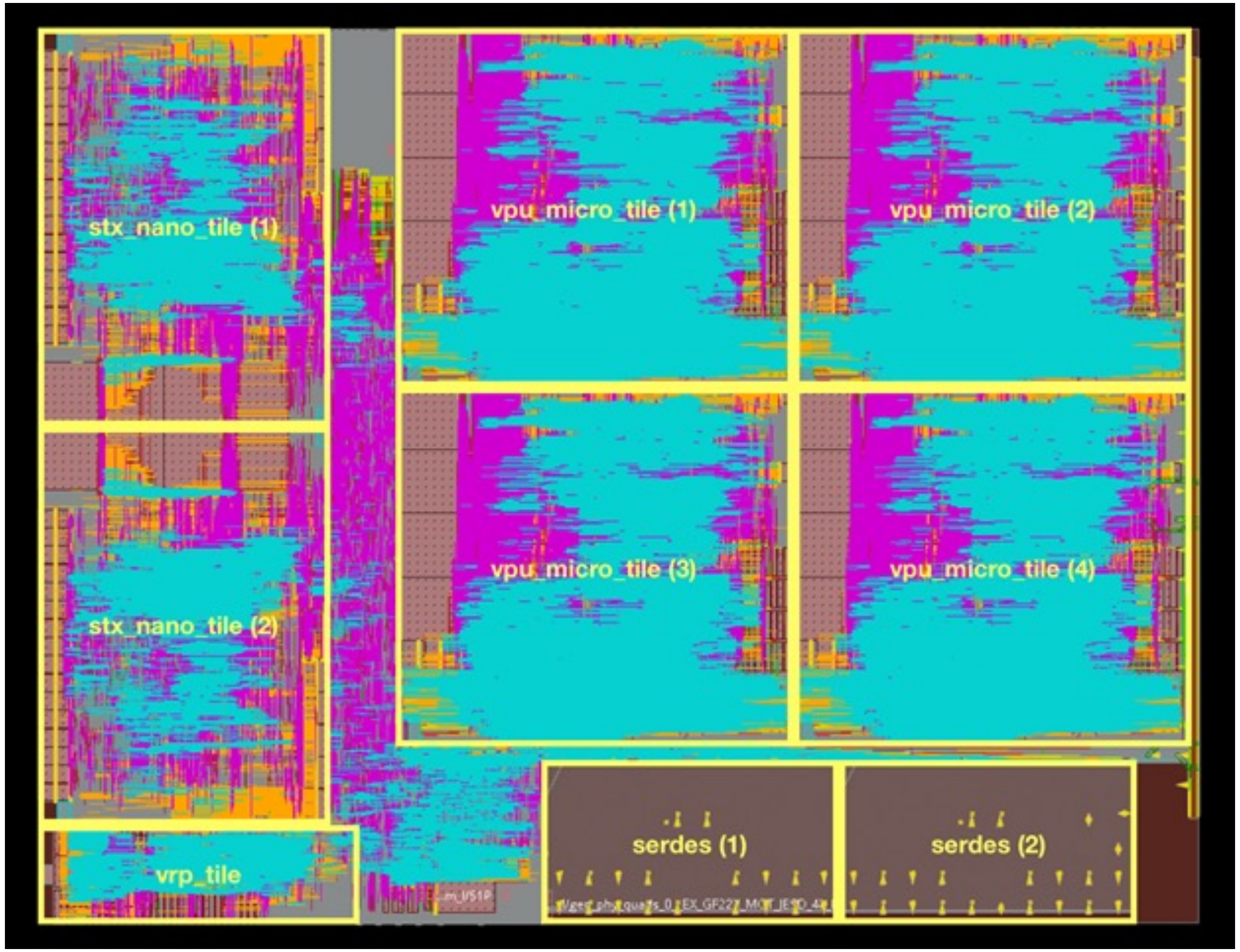
STREAM 3 EPAC 1.0 TEST CHIP

Four vector processing micro-tiles (VPU)

Two stencil and tensor accelerator (STX)

One variable precision processor (VRP)

22nm FDX technology

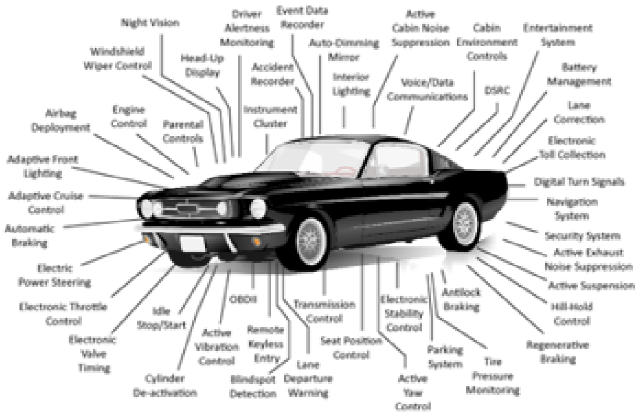


STREAM 4 EHPC FOR AUTOMOTIVE

Motivations for embedded HPC (eHPC) platform

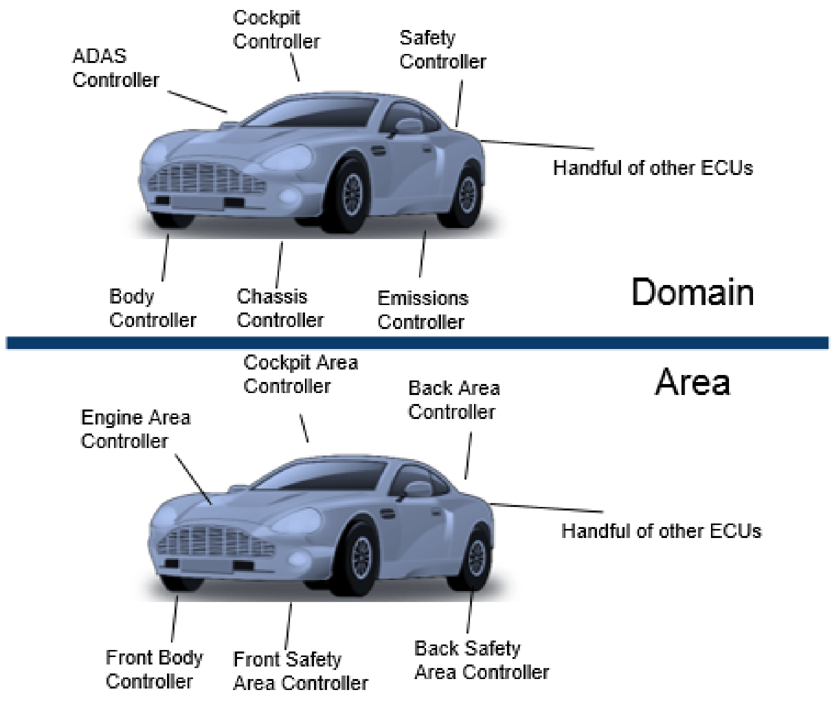
ECU Consolidation

Into a distributed central compute platform



TODAY

- 60-100 ECUs
- 6-8 operating systems
- Isolated operations
- Increasing cost & complexity

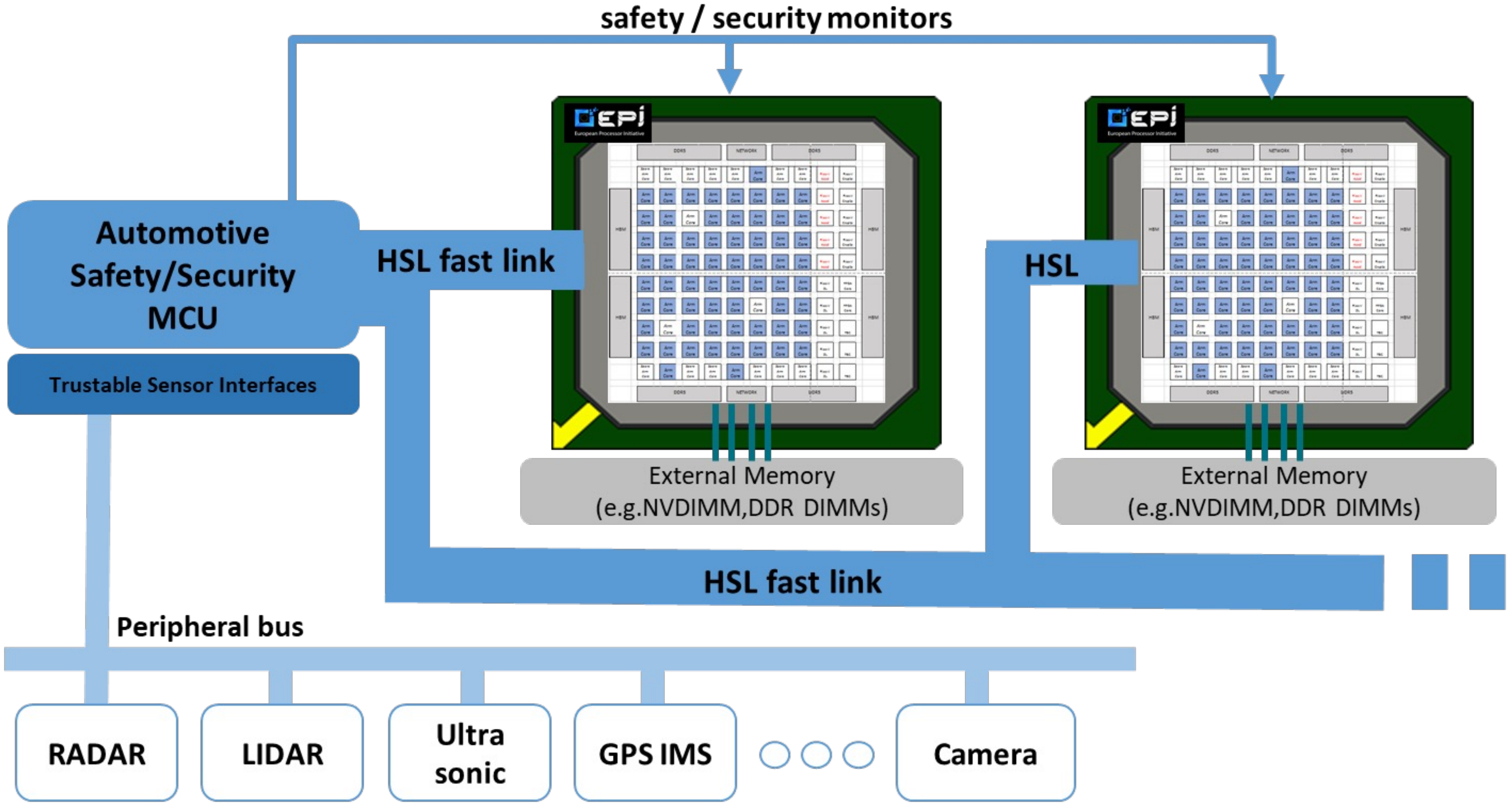


TOMORROW

- 6-10 Domain/Area Mega-controllers
- Consolidated software system
- Coordinated operations
- Reduced weight, cost, & complexity

STREAM 4 EHPC PLATFORM CONCEPT

Addressing automated driving performance and functional safety requirements



safety / security monitors

Automotive Safety/Security MCU

Trustable Sensor Interfaces

HSL fast link

HSL

External Memory (e.g.NVDIMM,DDR DIMMs)

External Memory (e.g.NVDIMM,DDR DIMMs)

HSL fast link

Peripheral bus

RADAR

LIDAR

Ultra sonic

GPS IMS

○ ○ ○

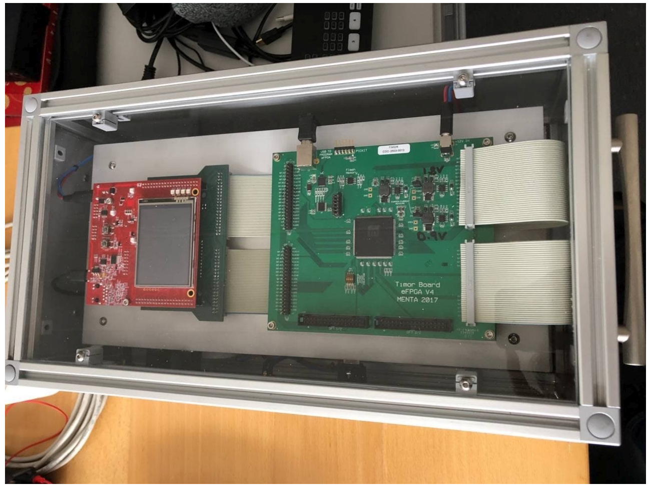
Camera

STREAM 4 DEMONSTRATION

The EPI Stream 2 GPP chip was not available in time for the demonstration

A BMW X5 car fitted with cameras, lidar, radars, network and processing elements was used to demonstrate EPI eHPC hardware and software components

- eFPGA + AURIX MCU for near-range recognition
- Kalray MPPA3 processor for mid-range object detection
- Far-range detection, short-range blind-spot detection, …



POSIT ARITHMETIC IN EPI SGA1

Focus on posit-based ML & DNN acceleration for AI in automotive

Università di Pisa

- Developed CppPosit library and integrated into the open source tinyDNN framework
- Designed novel fast (approximated) activation functions (Tanh and ELU) on Posits
- Investigated how to exploit SIMD instructions to speed-up computations with Posits
- Implemented a light Posit Processing Unit, integrated within a RISC-V core

Universidade de Lisboa

- Designed a dynamic fused multiply-accumulate Posit unit with variable exponent size

Kalray

- Experimented parameter quantization to BF16, FP16, FP8, Posit8.0 Posit8.1, Posit8.2, Posit8.3 on classification and detection networks

AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

DII, Università di Pisa

Kalray SA

Outlook

EUROPEAN PROCESSOR INITIATIVE SGA2

- Specific Grant Agreement 2 is the second first phase of the EPI, (January 2022 − July 2025).
- Organized into four technology streams and one management stream

1. RHEA Chips (SiPearl)
   - Chip Bring-up, Prototype Platform, System Software, Validations, Codesign
2. General Purpose Processor (SiPearl)
   - Architecture, Development & Verification, FPGA Emulation, Firmware & Software, Board Design, Hardware Common Platform, Software Common Platform, Power Management, System Security
3. Accelerators (BSC)
   - Architecture, STX RISC-V, VEC RISC-V, Uncore and Support IP, Software Stack and Tools, FPGA Emulation, EPAC Test Chips, KVX RISC-V, FPGA Chiplet
4. Emerging Applications (Atos & STMicroelectonics)
   - Autonomous HPC, EPI-Based HPC Blades
5. Management

EPI SGA2 TIMELINE

EPI 1

2019-2021

2022

Rhea platform

EPAC V1.5 sent for manufacturing

•

KVX RISC-V based accelerator
architecture definition

•

Menta FPGA chiplet
architecture definition

2023

Rhea Family - Gen1 GPP

EPI Common Platform
Arm & Risc-V (STX, VRP, ..)
Arm Neoverse V1 Core – N6
External IP’s

•

EPAC V1.5 multi-node demonstration cluster

EPAC V2.0 sent for manufacturing

•

Menta FPGA chiplet
sent for maufacturing

•

Kalray RISC-V SDK release

2024

Rhea2 Family - Gen2 GPP

Dual chiplet implementation
First EU Exascale system with Rhea processors
FPGA Die-to-Die Demonstrator of the HW Common Platform

•

RISC-V KVX FPGA emulator

•

EPAC V2.0 Platform

2025

Cronos Family - Gen3 GPP

EU Exascale system
with Rhea2 processors

•

Hurricane PCIe Acceleration card
based on EPAC technology

STREAM 1 RHEA1 GPP

Tape-out (2023) manufacturing TSMC 6nm, bring-up, software integration

Chip specifications completed

- Driven by co-design within EPI

IP-blocks designed and verified on simulation (RTL) and/or virtual platform

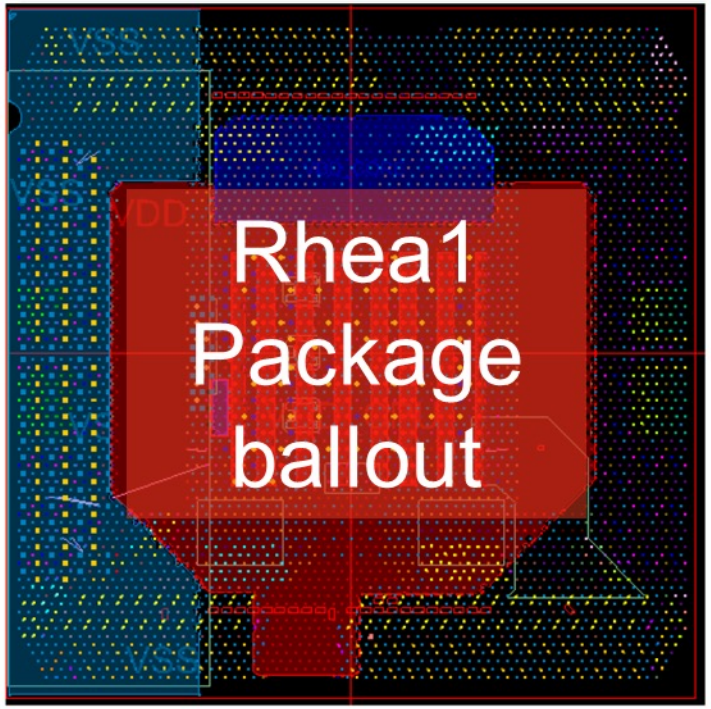
- Several improvements done alone the way
- Including more cores and faster memory

Board integration

- Motherboard designed and built in parallel (Atos)
- Prototype-cluster design in preparation (E4)

Software stack

- Arm software ecosystem
- Standard programming environment (LLVM, OpenMP)



STREAM 3 RISC-V BASED ACCELERATORS

Collaborations for the EPI RISC-V accelerator stream

VEC - Self-hosted RISC-V CPU + wide VPU (256 double elements) supporting RVV (SMD, BSC, Extoll, UniPi, IST, FORTH, UniZ, Chalmers)

STX - RISC-V CPU + specific cores for stencil and neural network computation (ETHZ, FhG)

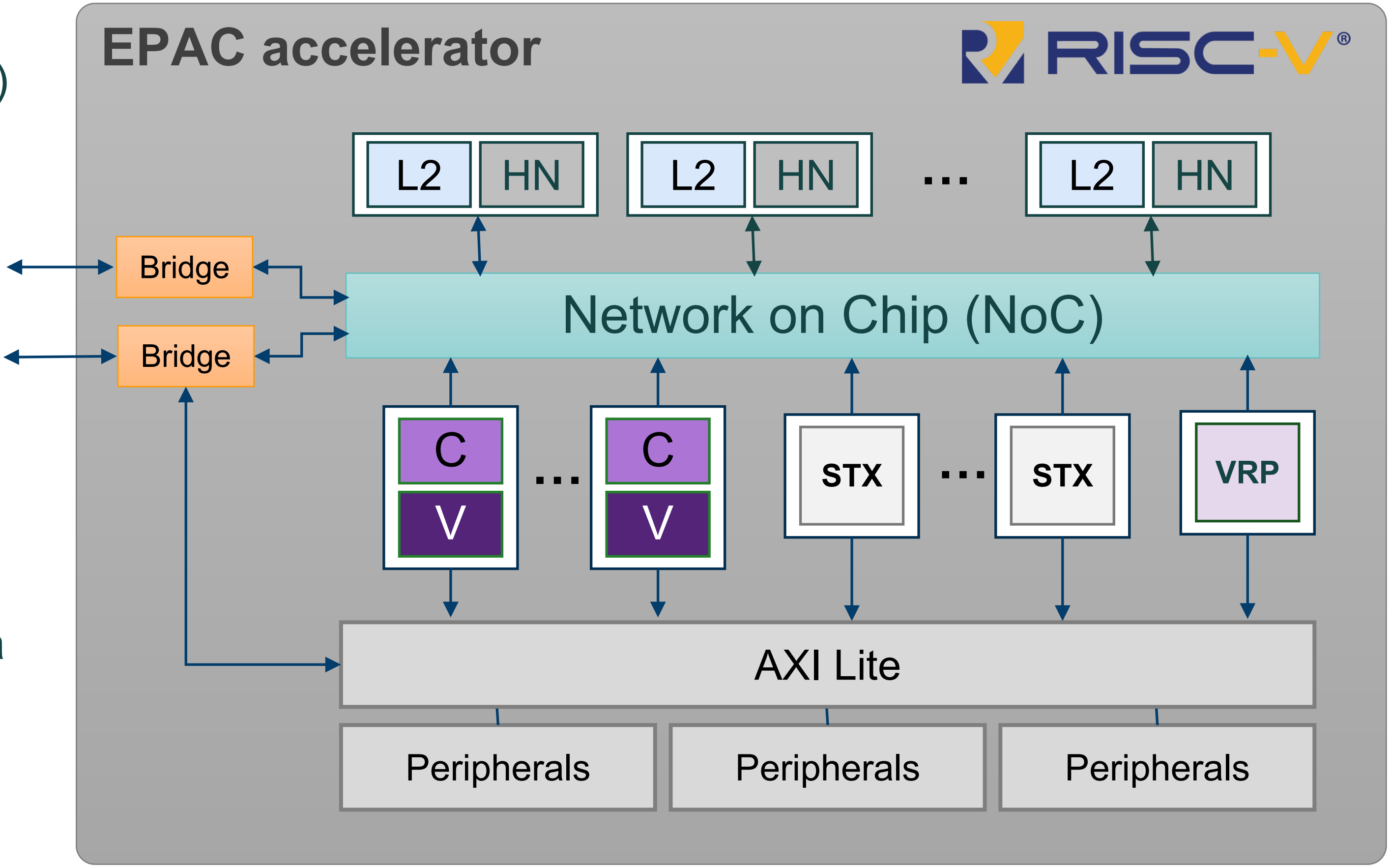
VRP - RISC-V CPU with support for variable precision arithmetic with data size up to 512 bit (CEA)

eFPGA - On-chip reconfigurable logic (Menta)

Ziptillion - IP compressing/decompressing data to/from the main memory (ZeroPoint, Chalmers)

KVX Tile - FPGA demonstrator and software of a 4-PE acceleration tile targeting HPC and ML (Kalray, UniZ)

SDV - Software Development Vehicles (FORTH, SMD, BSC)



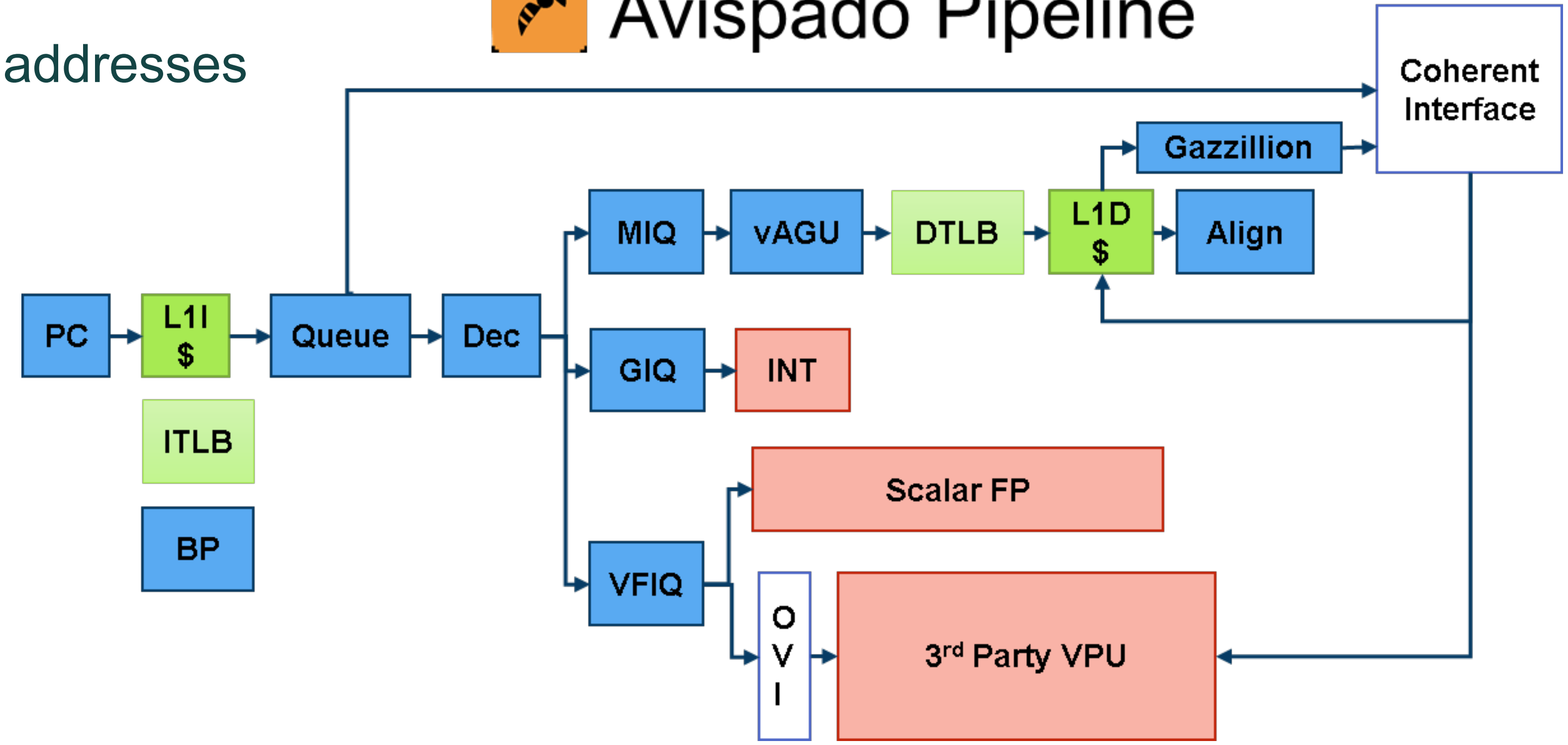
STREAM 3 VEC ACCELERATOR

Implementation of the RISCV64GCV ISA

Based on SemiDynamics AVISPADO 220 core with 3rd party VPU

- 12-stage 2-way in-order
- Three privilege levels M, S and U
- SV-48 virtual memory system model
- Hardware support for unaligned addresses
- 32 KB coherent L1 data cache
- ARM AMBA5 CHI NoC interface
- VPU "Open Vector Interface"

Avispado Pipeline



STREAM 3 VEC VECTOR PROCESSING UNIT

Implementation of the RISC-V V-Extension (RVV-0.7.1)

Vector length agnostic (VLA) programming and architecture

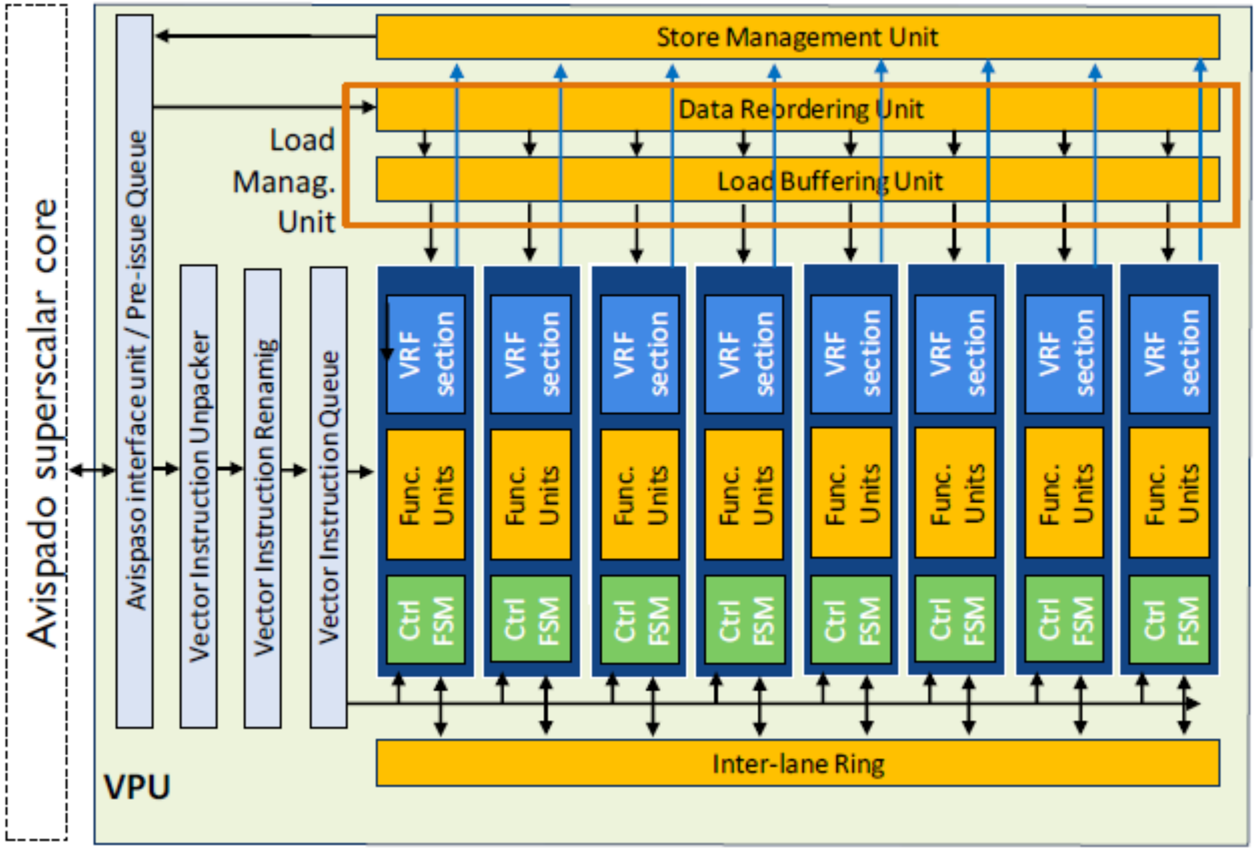
Long vectors: 256 DP elements

- #Functional Units (FUs) << Vector Length (VL)
- 1 vector instruction can take up to 32 cycles

8 Lanes per core

- 1 FMA/lane: 2 DP Flop/cycle

40 physical registers, some out of order



STREAM 3 VEC VECTOR FPU

European Processor Initiative

UNIVERSITY OF ZAGREB
Faculty of Electrical Engineering and Computing

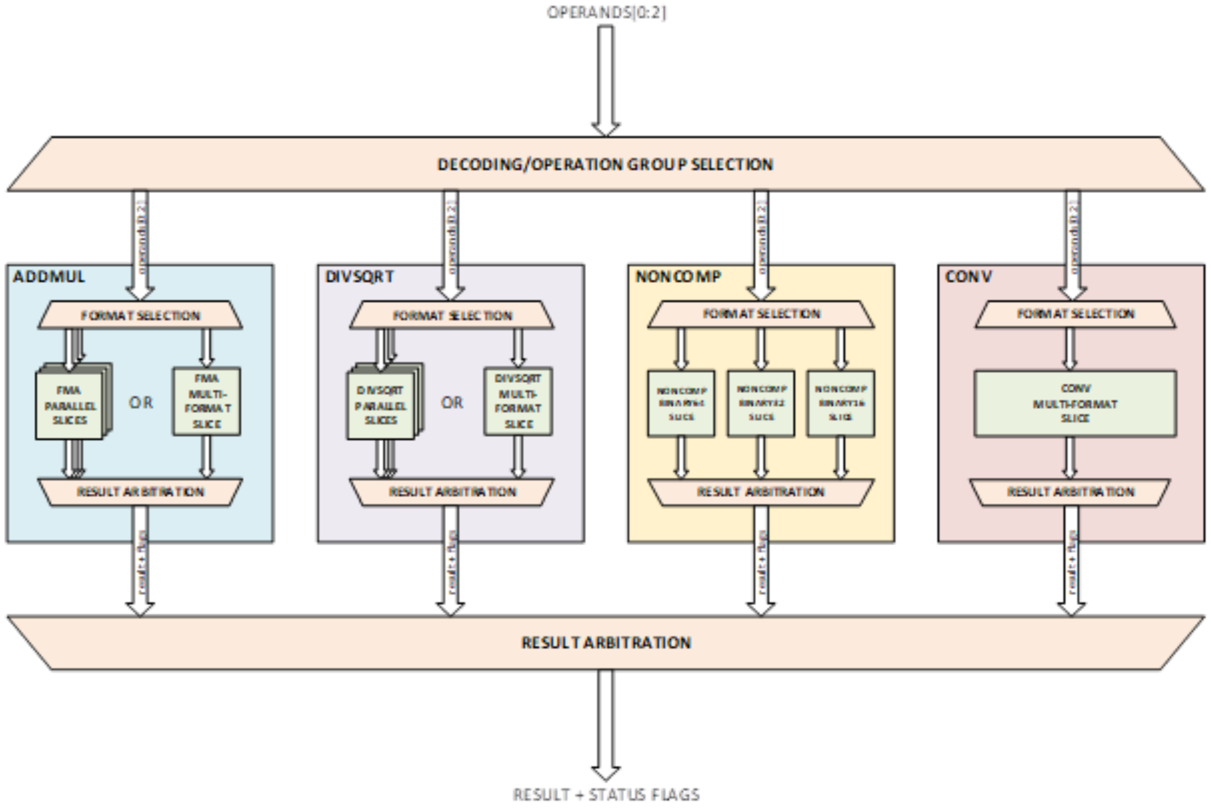
FAUST RISC-V pipelined vector Floating-Point Unit

Vector length agnostic (VLA) programming and architecture

- RVV 1.0 operations except reciprocal estimate operations
- binary16 (half precision format)
- binary32 (single precision format)
- binary64 (double precision format)
- Five rounding modes
- All IEEE 754 status flags
- 5 Pipeline stages

Support for vector unit integration

- Masking support
- Handshake interface for data flow control to and from the floating-point unit



STREAM 3 STX (STENCIL/TENSOR ACCELERATOR)

Domain specific accelerator for Machine Learning and Stencil

Provide 10x more energy efficient computing for these workloads

Sparse access patterns, mixed precision and new number formats (POSIT)

Four to Eight clusters

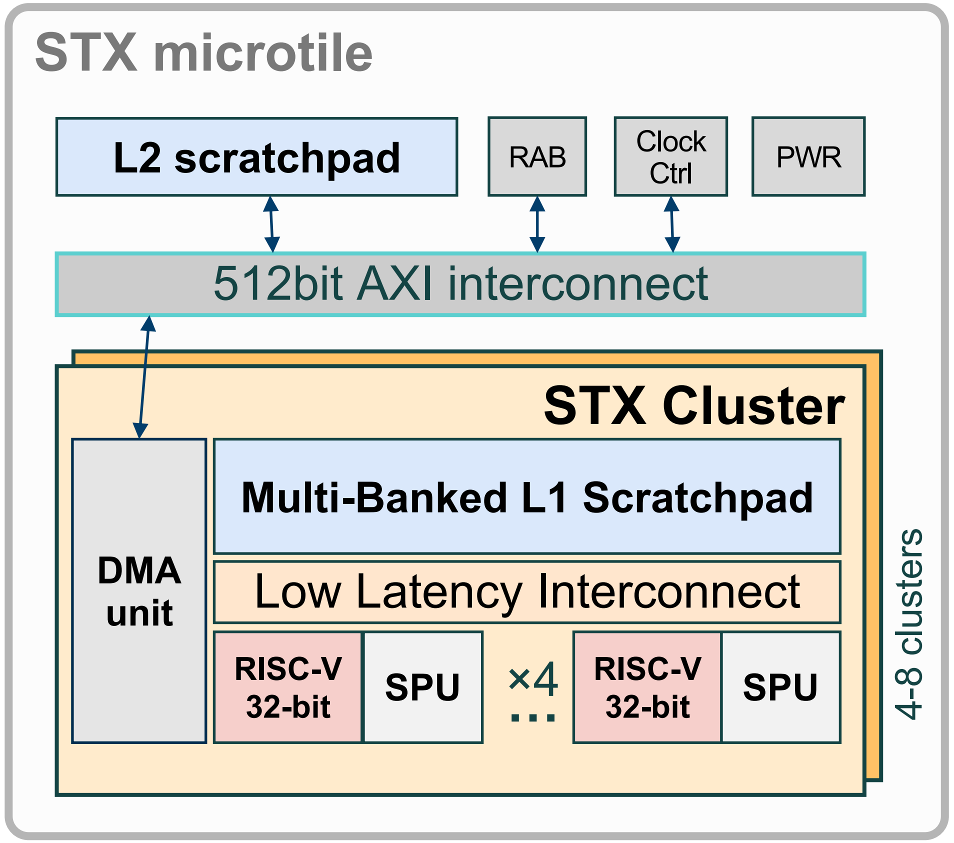
- Each cluster has 8 compute units
- And one 32bit RISC-V processor for support
- 128 kB local scratchpad memory
- DMA to transfer data from/to the accelerator

Specialized compute units (SPU or Sparta cores)

- SPU = pipelined processor with 1 FP64 unit; 4 ops/cycle
- Sparta = Streaming Floating Point Units for DL-acceleration

STX programming environment

- OpenMP offload to the STX unit from an ARM system (in the GPP) or the 64-bit RISC-V core (in the EPAC tile)
- GCC- and LLVM-based OpenMP offload flows



STREAM 3 VRP (VARIABLE PRECISION ACCELERATOR)

Targets efficient computation with extensive use of iterative linear algebra kernels

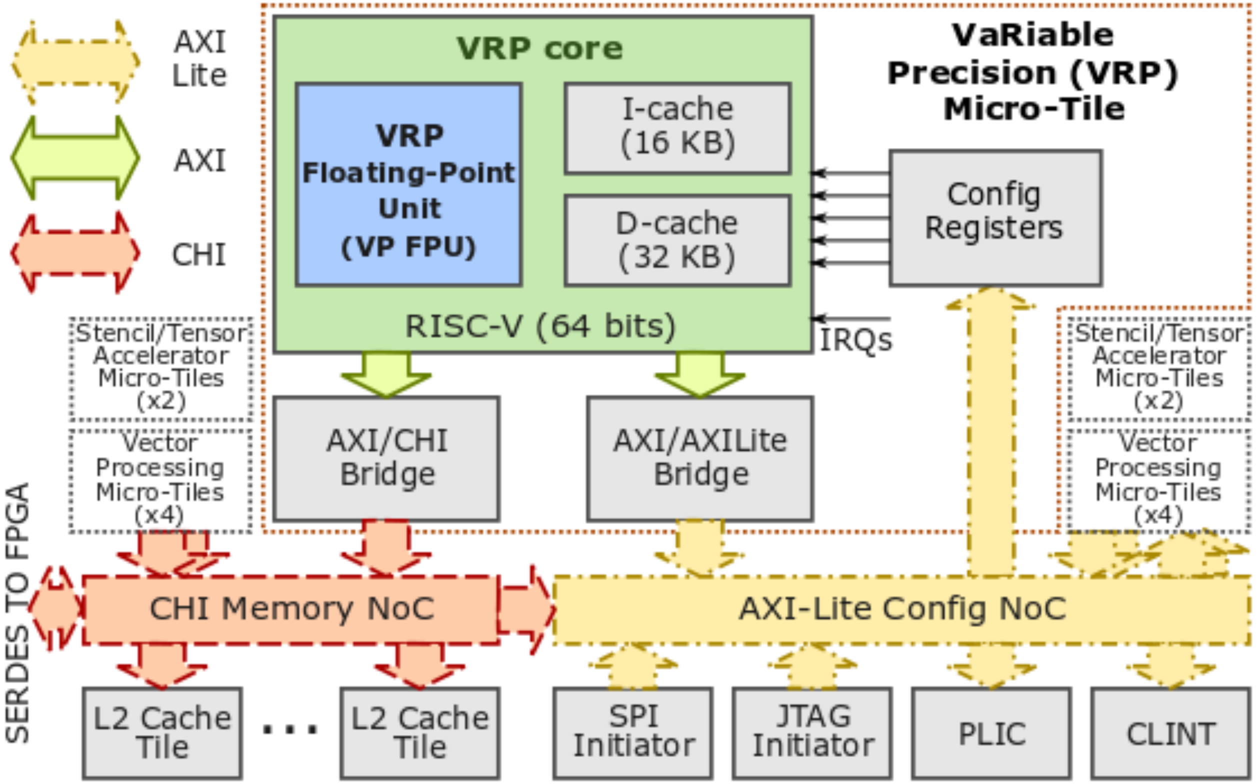
Augmenting accuracy inside kernels reduces rounding errors and therefore improves computation's stability

Supports up to 512 bits and 18 bits of significand and exponent size, respectively

Memory data format complies to the IEEE extendable one (IEEE 754-2008).

Extended RISC-V CVA6 core

- Custom ISA extension (namely Xvpfloat) for supporting VP arithmetic, logic, and memory operations on FP numbers
- Dedicated hardware VP FPU
- Dedicated register file (32 registers) for VP FP numbers



STREAM 3 KVX ACCELERATOR TILE

KVX Cluster

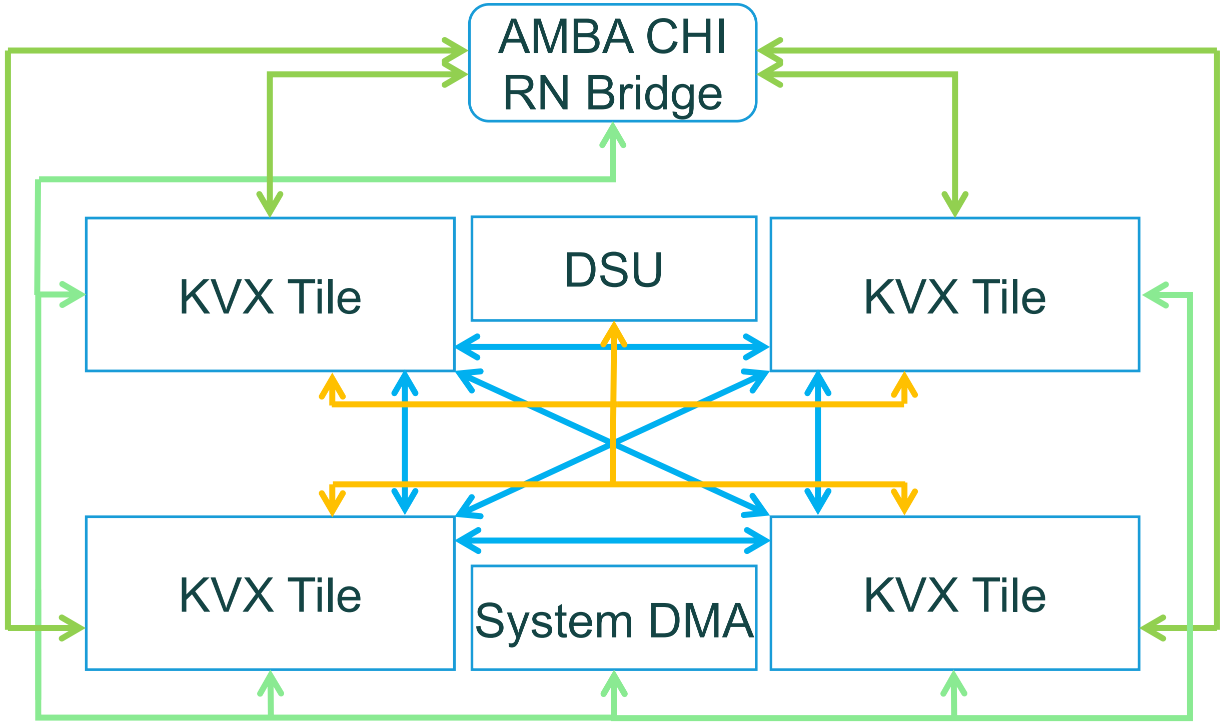
- 4 KVX Tiles, DMA engines, hardware accelerators, Debug Support Unit (DSU)
- Cluster interconnect bridge to AMBA CHI

KVX Tile

- 4 KVX Processing Elements (PE)
- 8 local memory banks of 4MB total capacity
- Local memory configured as scratch-pad or parts of the L2 cache of the KVX cluster

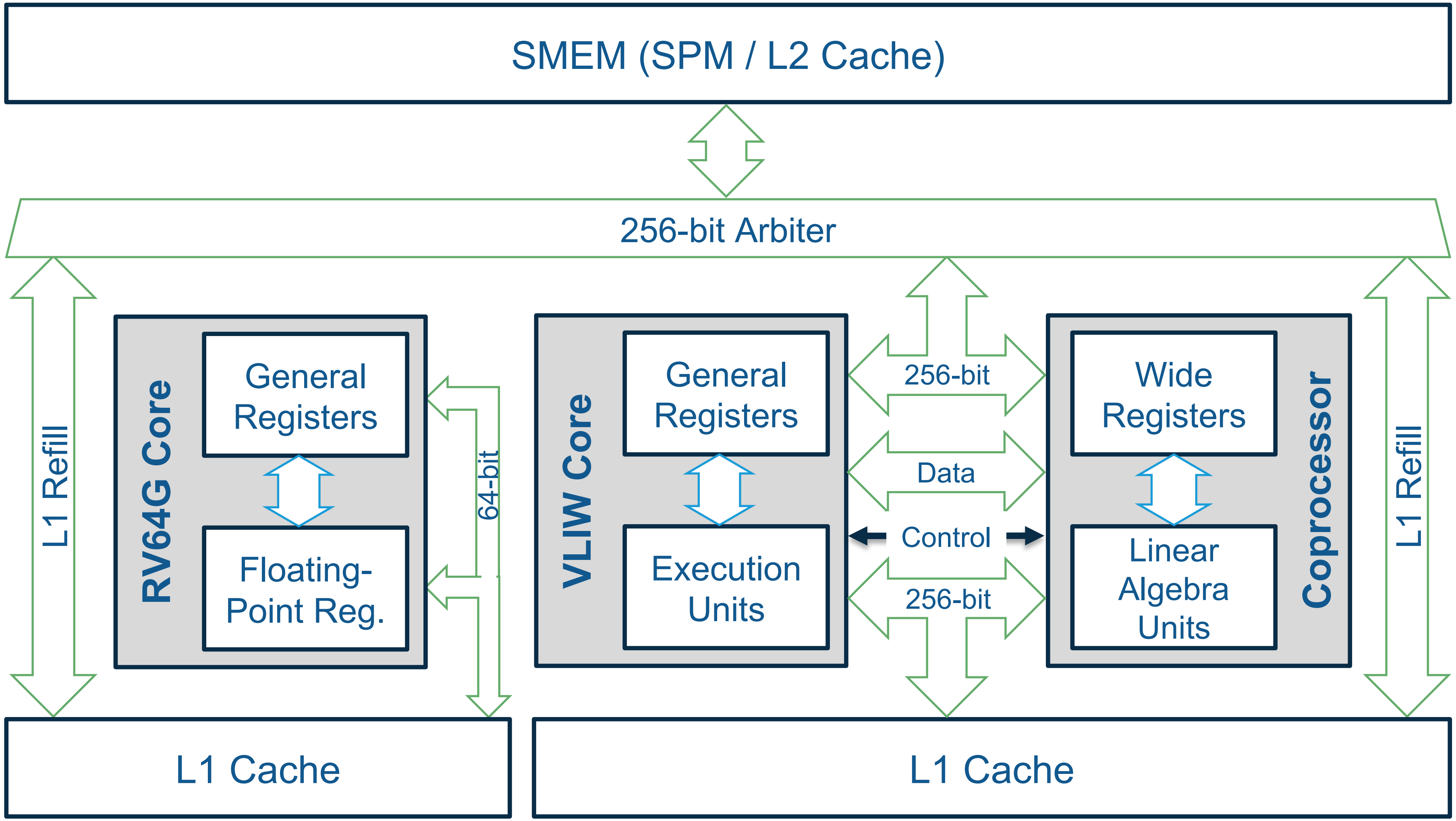
KVX PE Performances

- Deep Learning INT8: 2048 op/cycle
- Deep Learning FP8/Posit8: 2048 flop/cycle
- Deep Learning FP16: 1024 flop/cycle
- FPU FP32: 32 flop/cycle
- FPU FP64: 16 flop/cycle



KVX PROCESSING ELEMENT (PE)

CVA6 RISC-V core, KVX VLIW core, Tensor coprocessor, 256-bit SMEM access



POSIT ARITHMETIC IN EPI SGA2

Focus on integration and test of Posit processing cores to Stream 3 accelerators

Università di Pisa

- Collected and integrated three Posit processing core designs (Posit Processing Unit from University of Pisa, Joint IEEE-Posit FMA Processing Unit from IST, Multiply-Accumulate Posit unit)
- These 3 cores were integrated in a Posit Test Array (PTA) and synthesized on a ZCU106 board
- Planning to test the mentioned system with a Seismic analysis application provided by FHG to assess the capabilities of Posits and their hardware accelerator in a more complex application

Universidade de Lisboa

- Extended the FPNew unit from ETHZ to design a unified Posit/IEEE-754 vector MAC unit for transprecision computing

Kalray

- Designed a hardware decompressor from Posit8.n to FP16 with n $\in$ {0, 1, 2, 3}
- Explored and compared the design of 32-term exact dot-product accumulate operators for different 8-bit floating-point representations (FP variations and Posit8 with different es)

AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

DII, Università di Pisa

Kalray SA

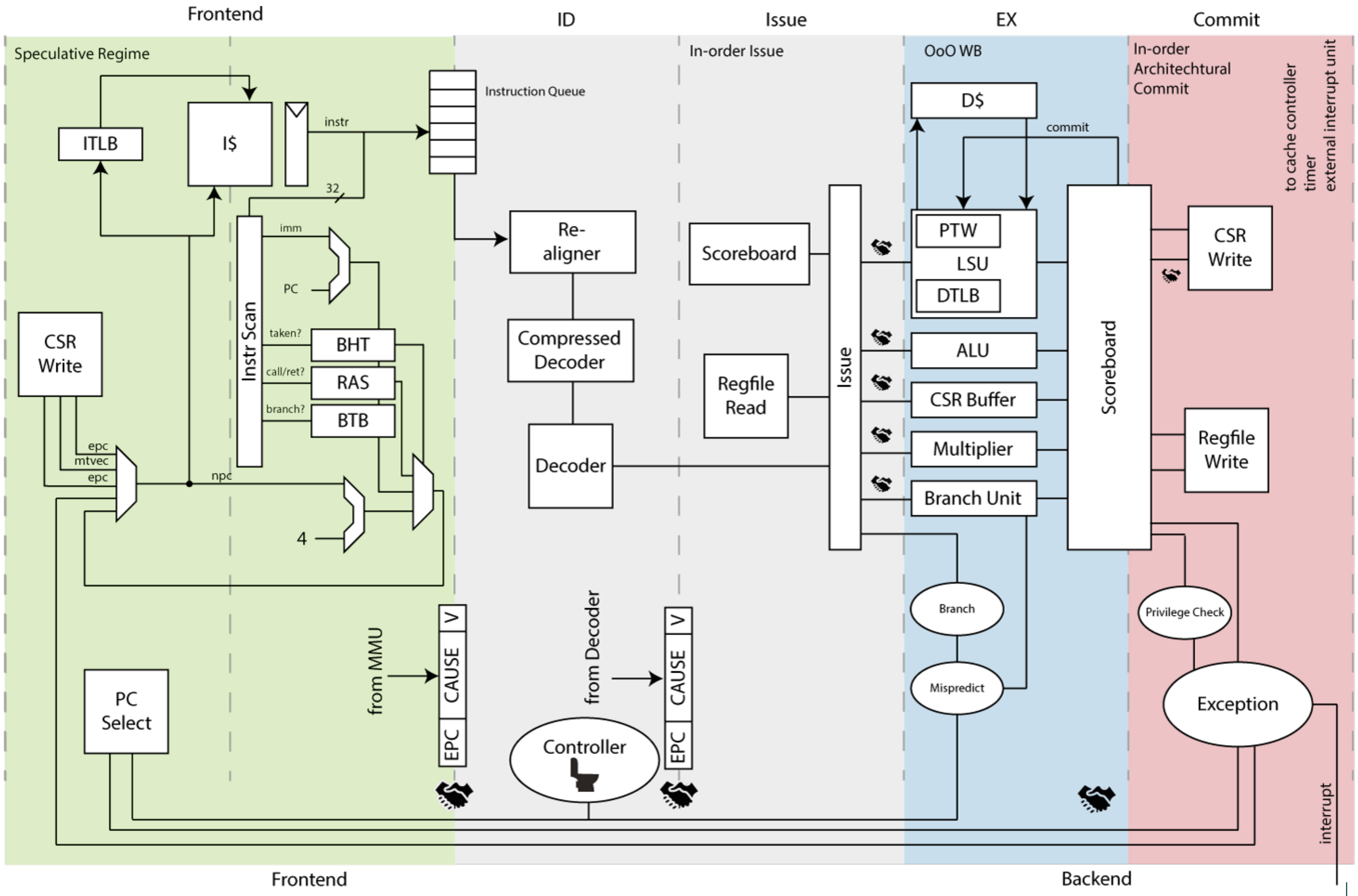
Outlook

RISC-V CORE DESIGN AND IMPLEMENTATION

CVA6 64-bit RV64GC application core (formerly known as Ariane)

Maintained by the OpenHW Group

- Single issue, in-order
- 6-stage instruction pipeline
- Branch prediction (branch target buffer, branch history table, return address stack)
- Separate I and D TLBs
- Hardware page table walk
- M, S, U privilege levels

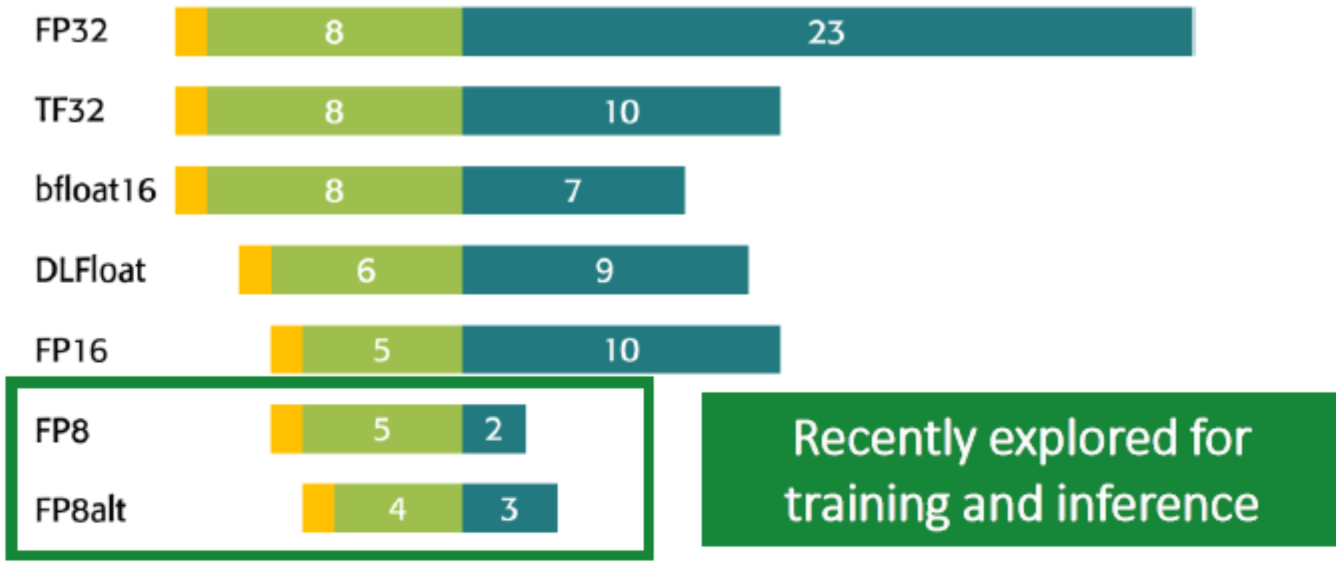


TRANSPRECISION FLOATING-POINT ARITHMETIC

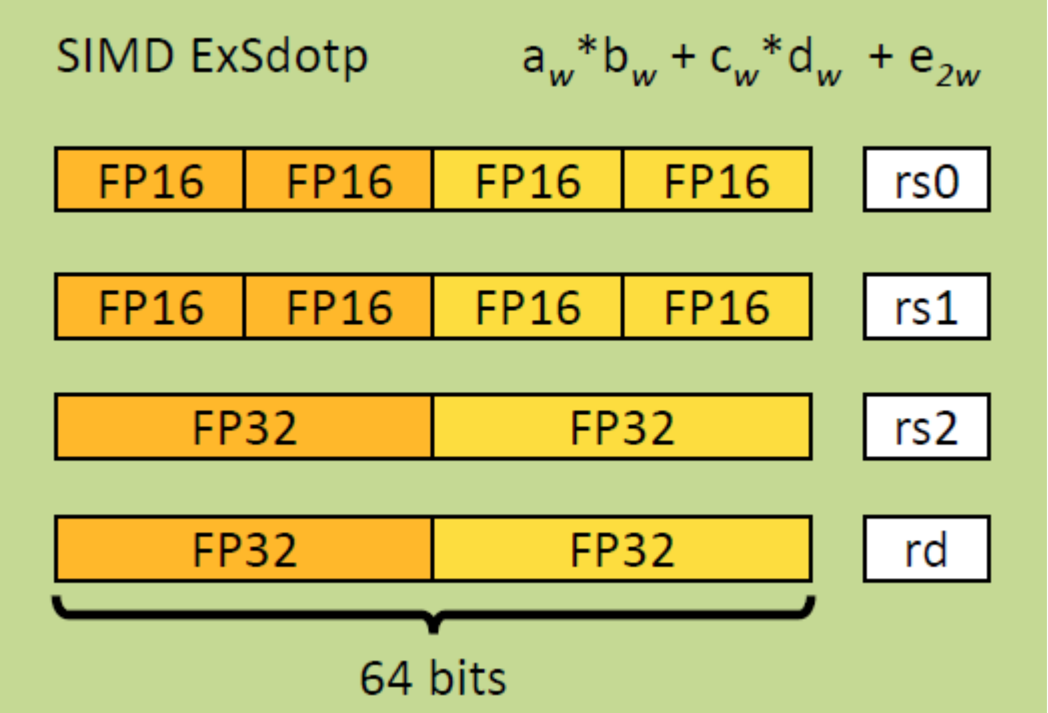
Tuning arithmetic approximation at a fine grain during the computation progress

Requires the availability of hardware units providing efficient support for multiple FP formats.

- ExSdotp unit is an open-source parameterized multi-format unit supporting expanding sum-of-dot-product (ExSdotp) instructions (8-to-16-bit and 16-to-32-bit), as well as non-expanding and expanding three-operand additions, called Vsum and ExVsum.
- Integration of ExSdotp unit into an open-source multi-format FPU (FPnew1)



Recently explored for training and inference



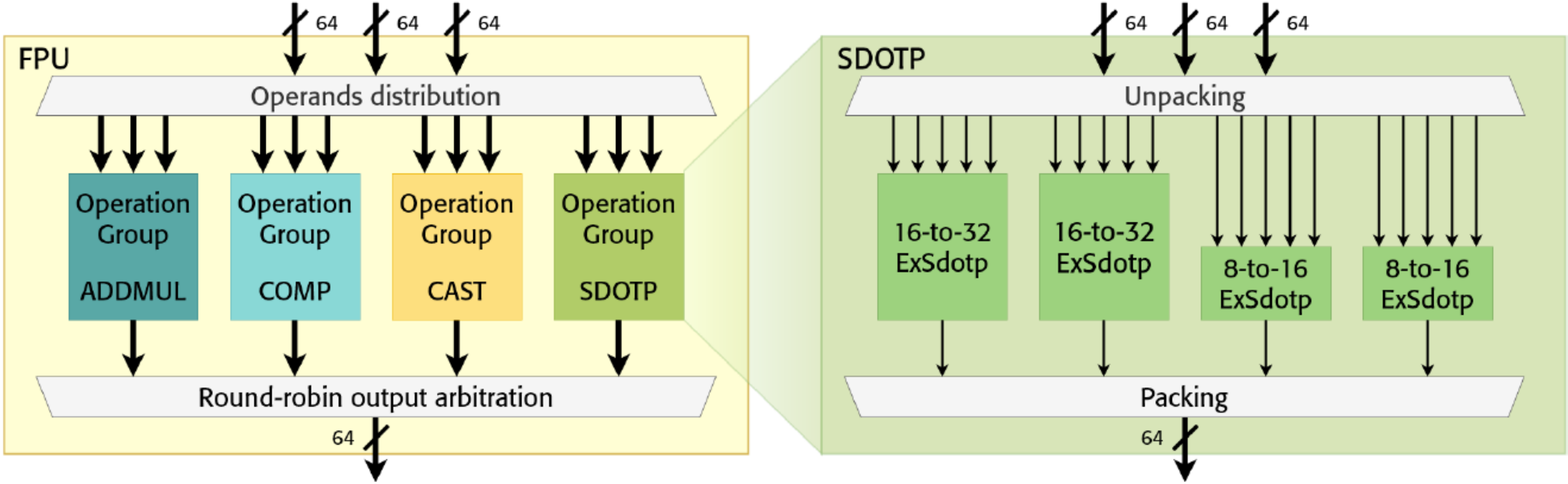
TRANSPRECISION FLOATING-POINT UNIT

FPnew is a highly-parameterized open-source modular energy-efficient multi-format FPU

All standard RISC-V operations are supported along with various additions

SIMD ExSdotp unit integrated into FPnew as a new operation group block

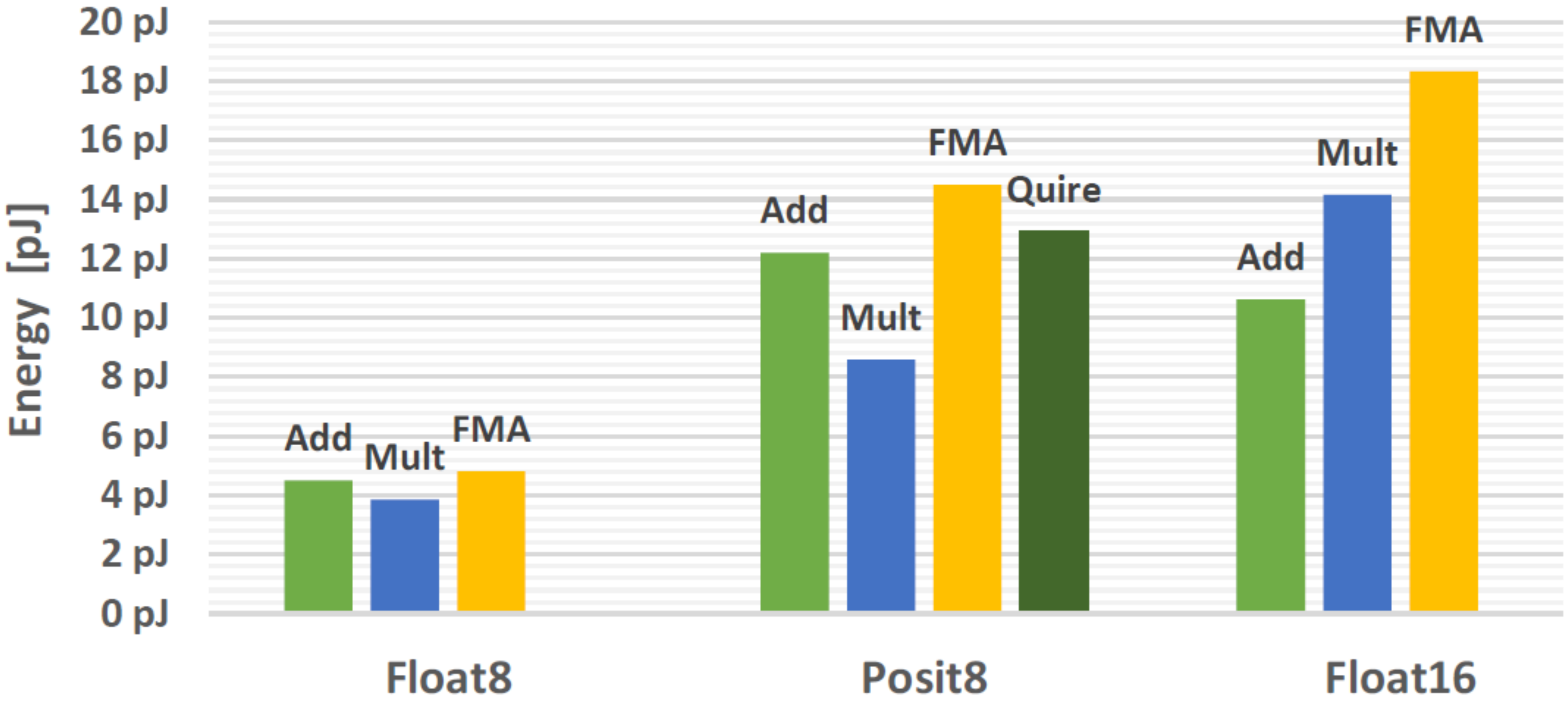
- SIMD SDOTP: two 16-to-32-bit units and two 8-to-16-bit units
- Up to two 16-to-32-bit ExSdotp and four 8-to-16-bit ExSdotp per cycle



EARLY EXPERIMENTS WITH POSIT ARITHMETIC

S. Mach PhD compares a 8-bit posit FMA unit and an 8-bit posit unit containing 32-bit quire functionality with IEEE 754 FMA implementations in a 65nm technology

Compared to FP8 there is the need to cover the regime's worst-case sizes and the mantissa



AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

DII, Università di Pisa

Kalray SA

Outlook

POSIT DFMA UNIT WITH VARIABLE EXPONENT SIZE

A posit Dynamic Fused Multiply-Accumulate (DFMA) unit support for variable exponent sizes

Capable of encoding an extended representation range, from values with high decimal precisions to very large integer numbers, within the same hardware (5 pipeline stages)

Capable of accumulating the results of fused operations with different Posit configurations

The implementation overheads imposed by the improved representation range are marginal

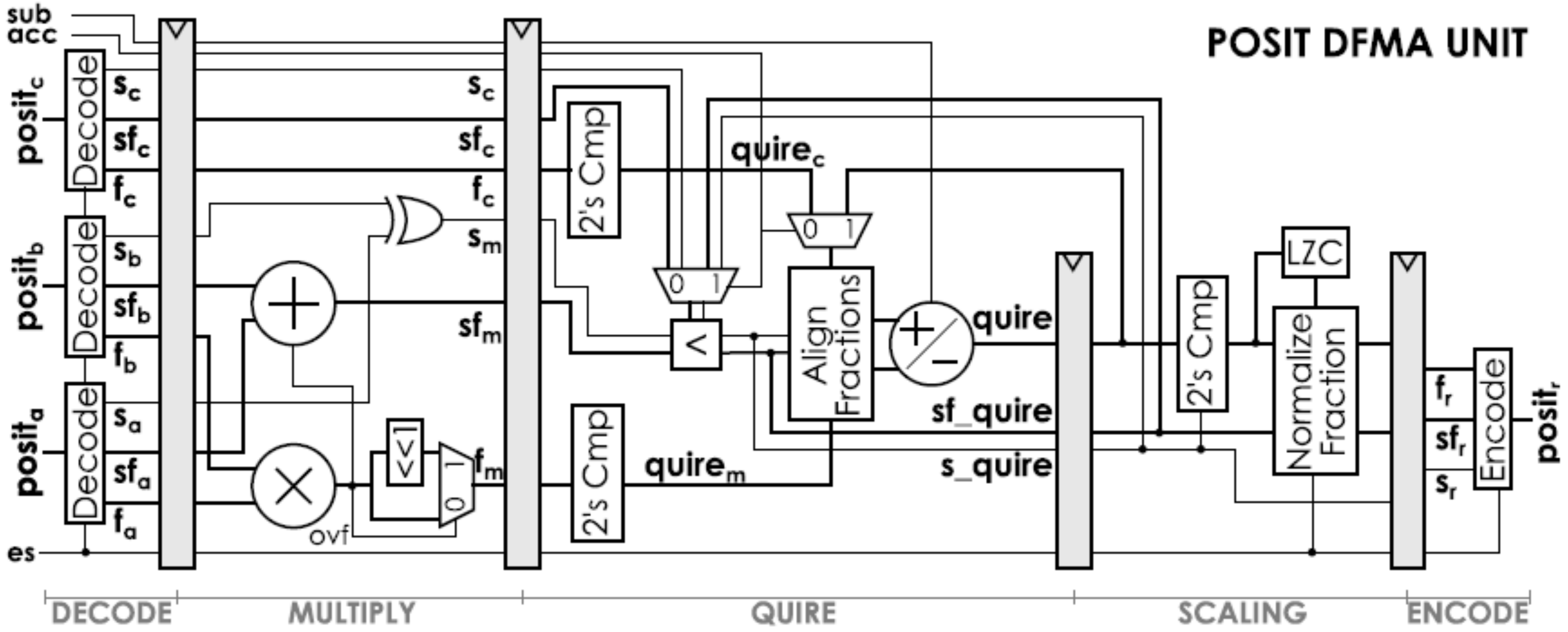
Legend:

— control signals

═ data signals

Supported Operations:

single	*fused*
r = a * b	r = a * b + c
r = a + b	r = a * b - c
r = a - b	r += a * b
	r -= a * b

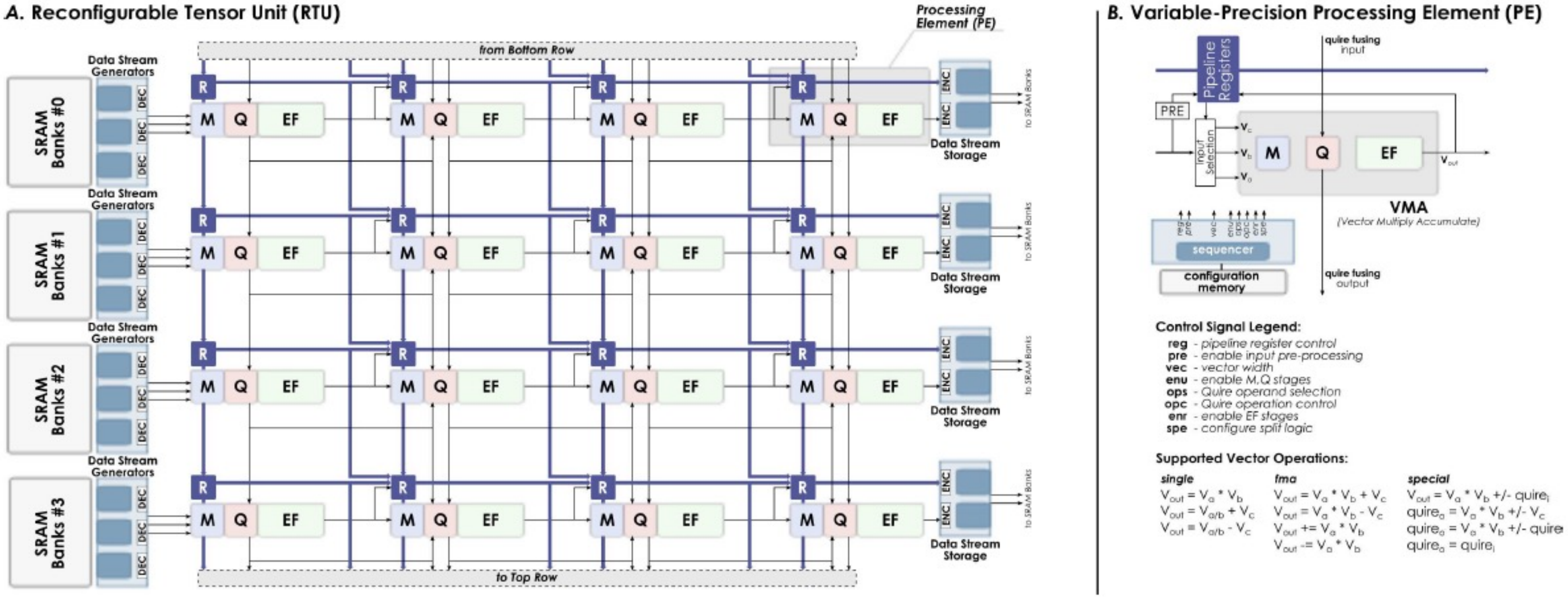


RECONFIGURABLE STREAM-BASED TENSOR UNIT WITH VARIABLE-PRECISION POSIT ARITHMETIC

2D array of reconfigurable PEs each with a 64-bit posit Vector Multiply-Accumulate (VMA) unit

Autonomous stream generators connected to a banked scratchpad/buffering memory structure

Fully implemented in RTL and synthesized with a 45nm technology



Control Signal Legend:

reg - *pipeline register control*
pre - *enable input pre-processing*
vec - *vector width*
enu - *enable M,Q stages*
ops - *Quire operand selection*
opc - *Quire operation control*
enr - *enable EF stages*
spe - *configure split logic*

Supported Vector Operations:

single	*fma*	*special*
$V_{out} = V_a * V_b$	$V_{out} = V_a * V_b + V_c$	$V_{out} = V_a * V_b$ +/- $quire_i$
$V_{out} = V_{a/b} + V_c$	$V_{out} = V_a * V_b - V_c$	$quire_o = V_a * V_b$ +/- V_c
$V_{out} = V_{a/b} - V_c$	$V_{out} += V_a * V_b$	$quire_o = V_a * V_b$ +/- quire
	$V_{out} -= V_a * V_b$	$quire_o = quire_i$

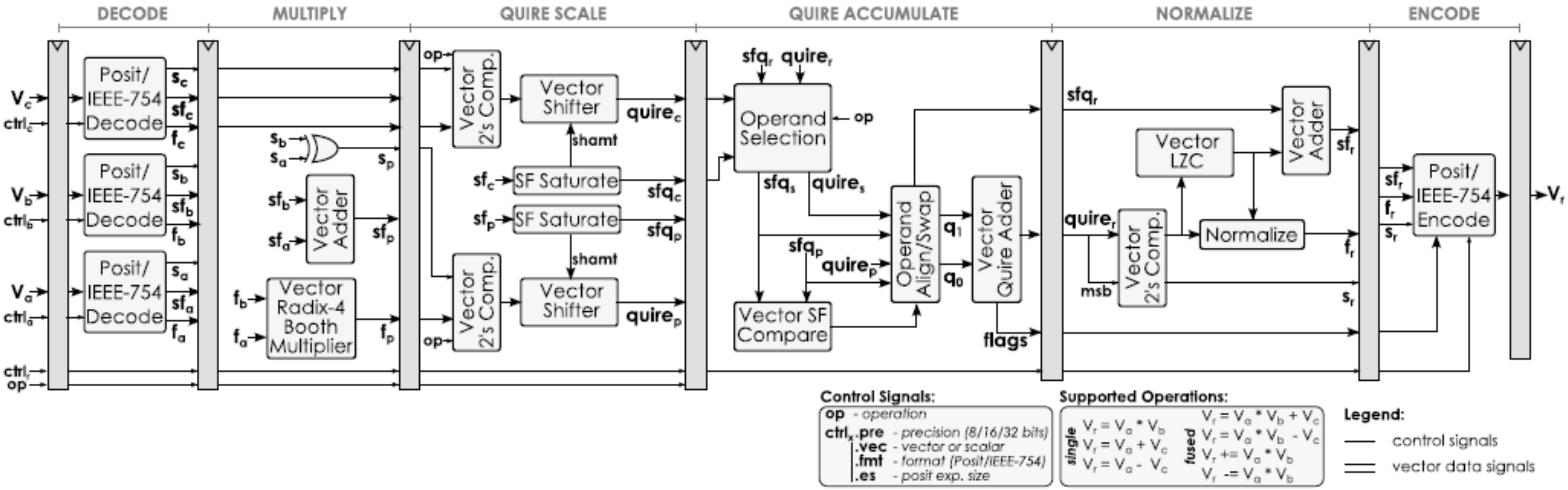
UNIFIED POSIT/IEEE-754 VECTOR MAC UNIT FOR TRANSPRECISION COMPUTING

Unified FP arithmetic architecture compatible with both the IEEE-754 and the Posit formats

Compatible with the existing RISC-V Vector (RVV) and proposed RISC-V Posit extensions

All arithmetic modules are fully vectorized and configurable at runtime to support 1x32-bit, 2x16-bit, and 4x8-bit vector operations using the same hardware

Synthesized for 28nm UMC standard cell technology, targeting 667 MHz for 6 pipeline stages



AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

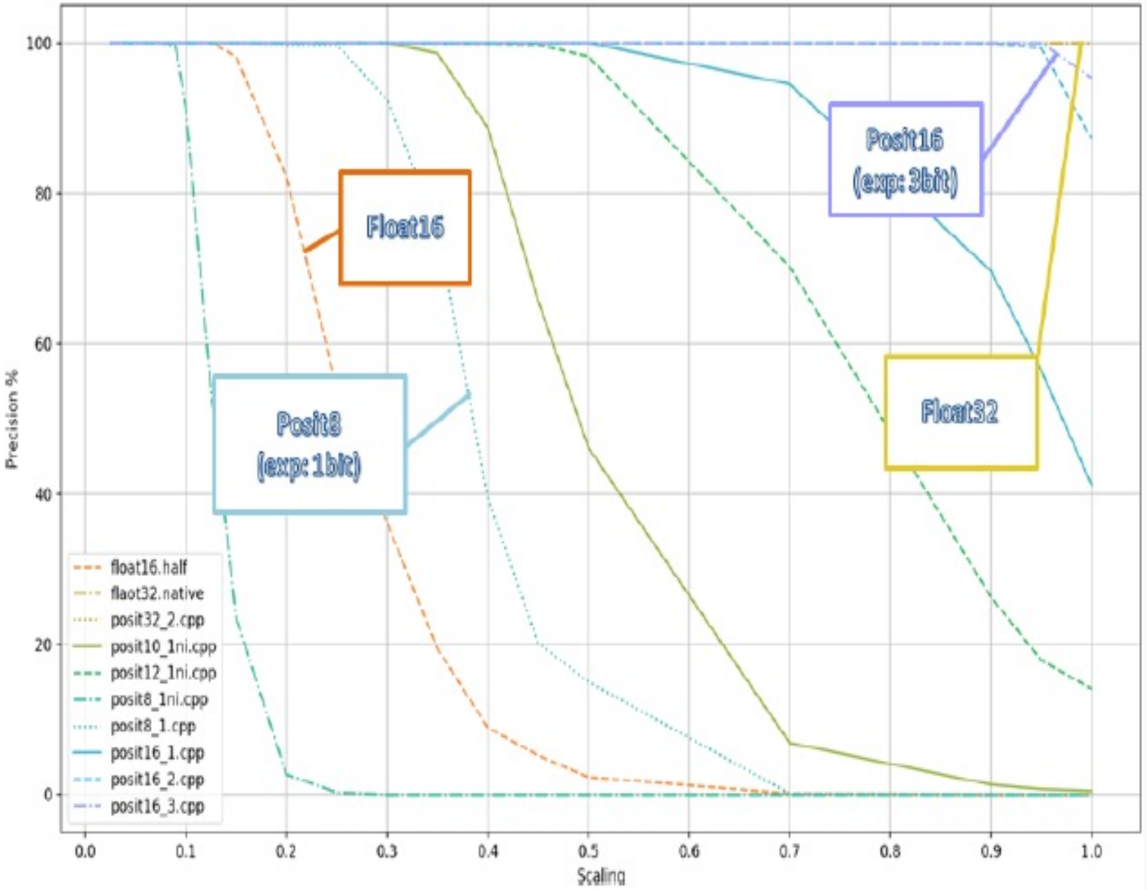
DII, Università di Pisa

Kalray SA

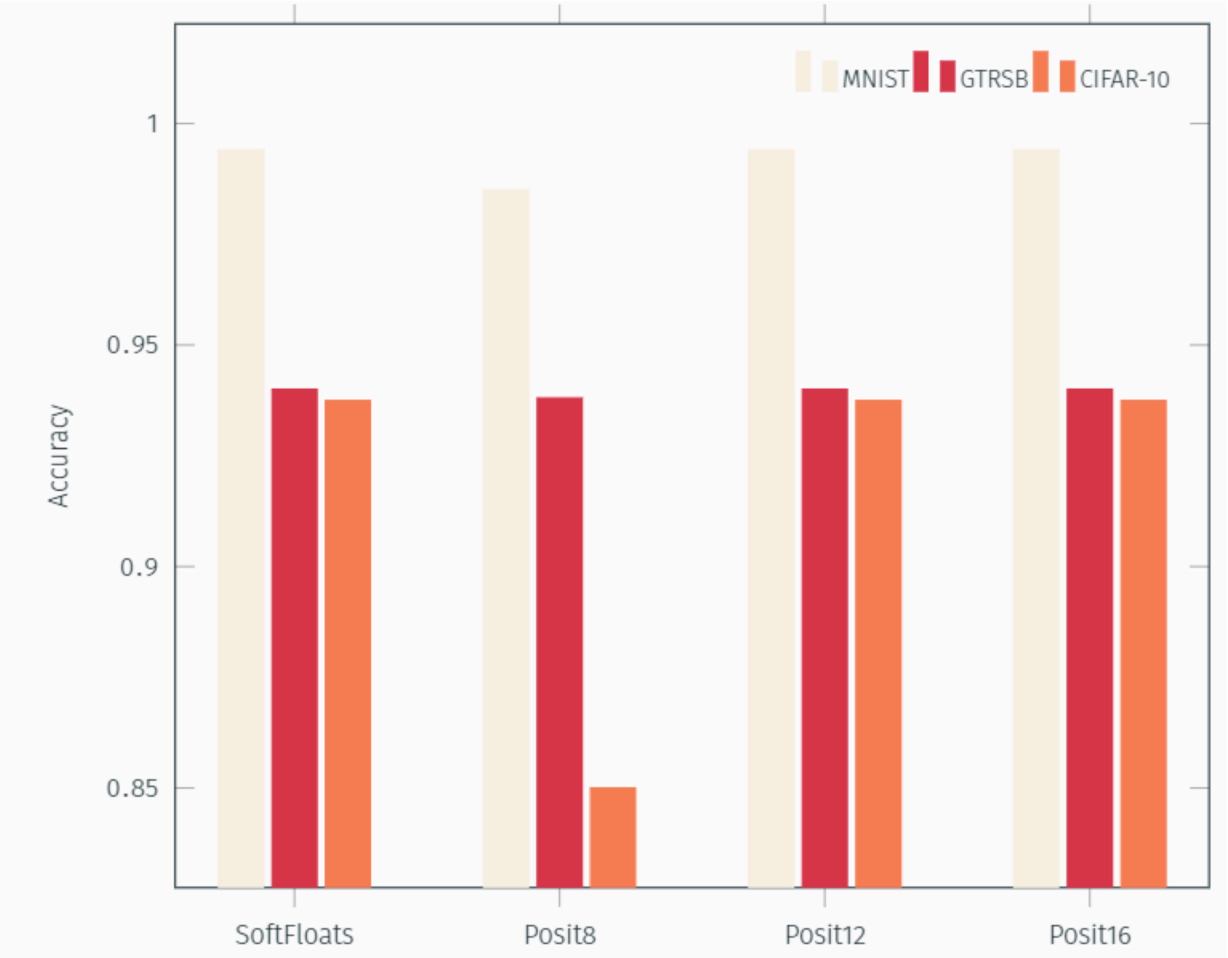
Outlook

POSIT ARITHMETIC FOR DEEP NEURAL NETWORKS IN AUTONOMOUS DRIVING APPLICATIONS

K-NN precision as a function of the scaling factor on the Mnist dataset. Posit8-1 achieves higher precision than Float16 for the same scaling factor.



Accuracy for Posit 16, 12, 8 vs. FP32 for TinyDNN with MNIST, GTRSB and CIFAR data sets



FAST APPROXIMATION OF ACTIVATION FUNCTION IN NEURAL NETWORKS USING POSIT ARITHMETIC

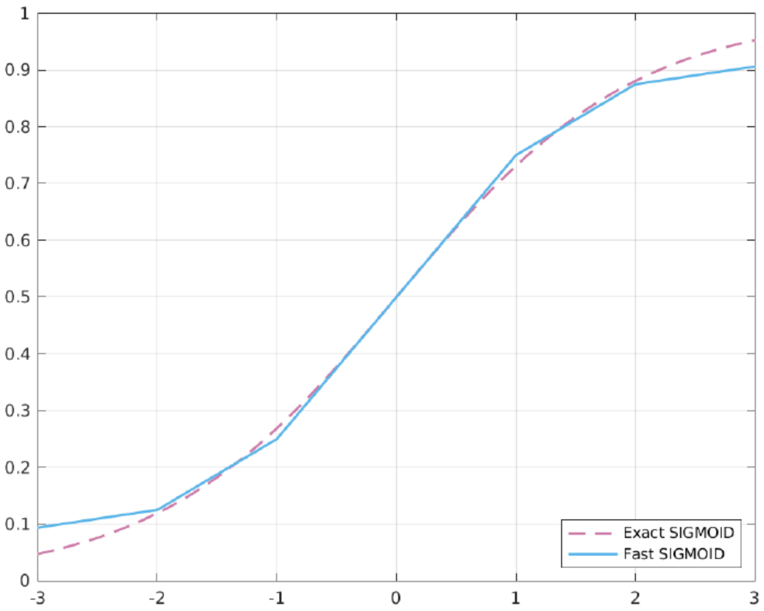
Implementation in the cppPosit library Layer 1 (ALU instructions only) with $es = 0$

Accelerated by leveraging the ARM and RISC-V vector extensions

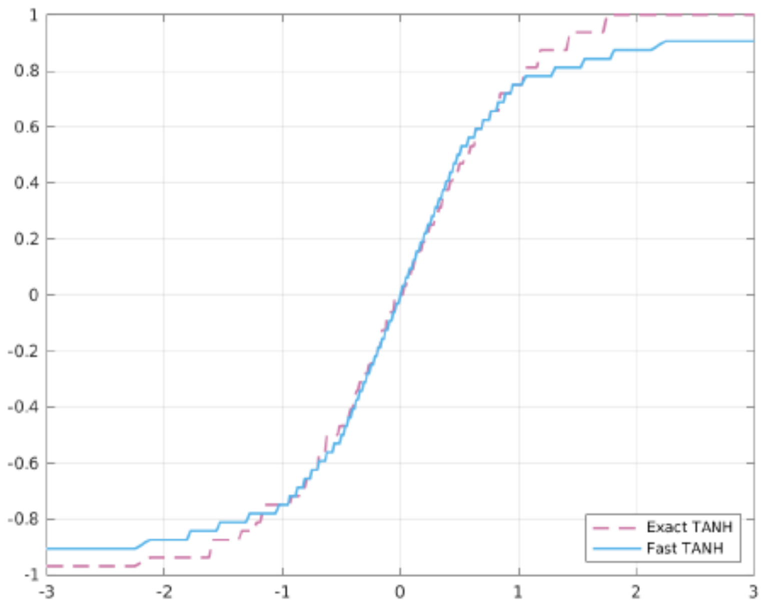
Vectorized kernels integrated to tinyDNN

$$\mathrm{ReLU}(x) = \begin{cases} 0, \text{if } x \leq 0 \\ x \text{ otherwhise} \end{cases}$$

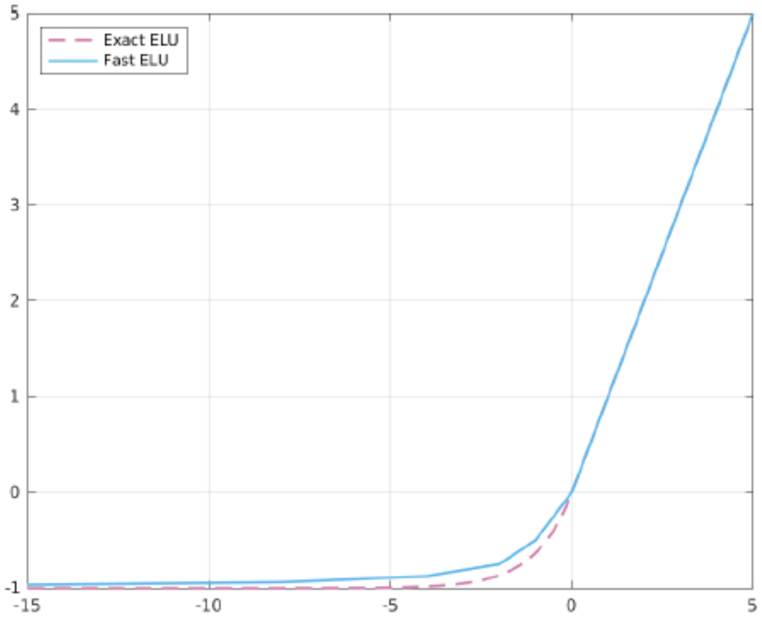
$$sigmoid(x) = \frac{1}{e^{-x} + 1}$$



$$tanh(x) = -(1 - 2 \cdot sigmoid(2 \cdot x))$$



$$\mathrm{ELU}(x) = \begin{cases} \alpha \cdot (e^{x} - 1), \text{if } x \leq 0 \\ x \text{ otherwhise} \end{cases}$$

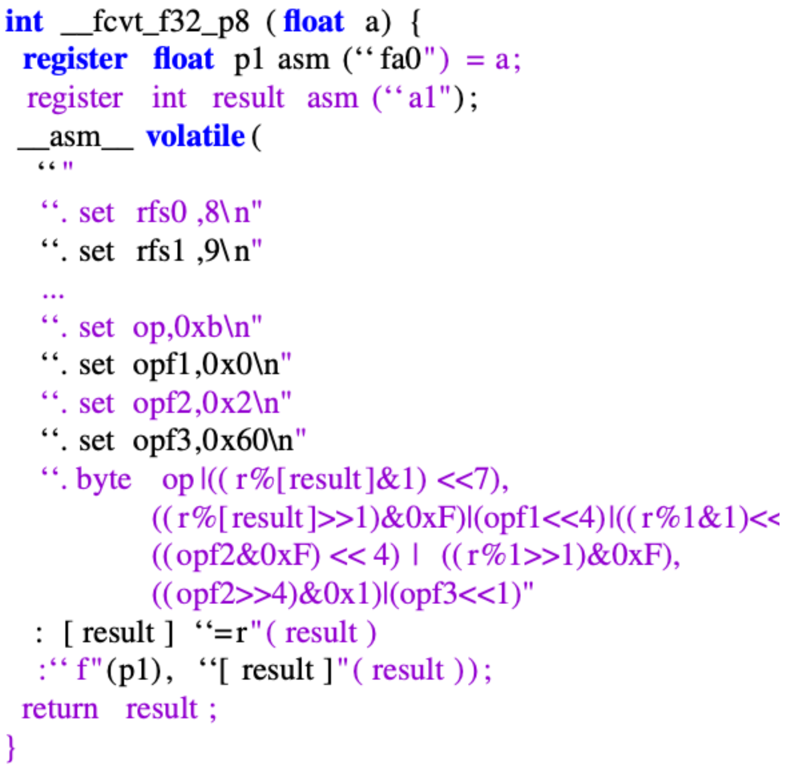


LIGHTWEIGHT POSIT PROCESSING UNIT FOR RISC-V PROCESSORS IN DEEP NEURAL NETWORKS

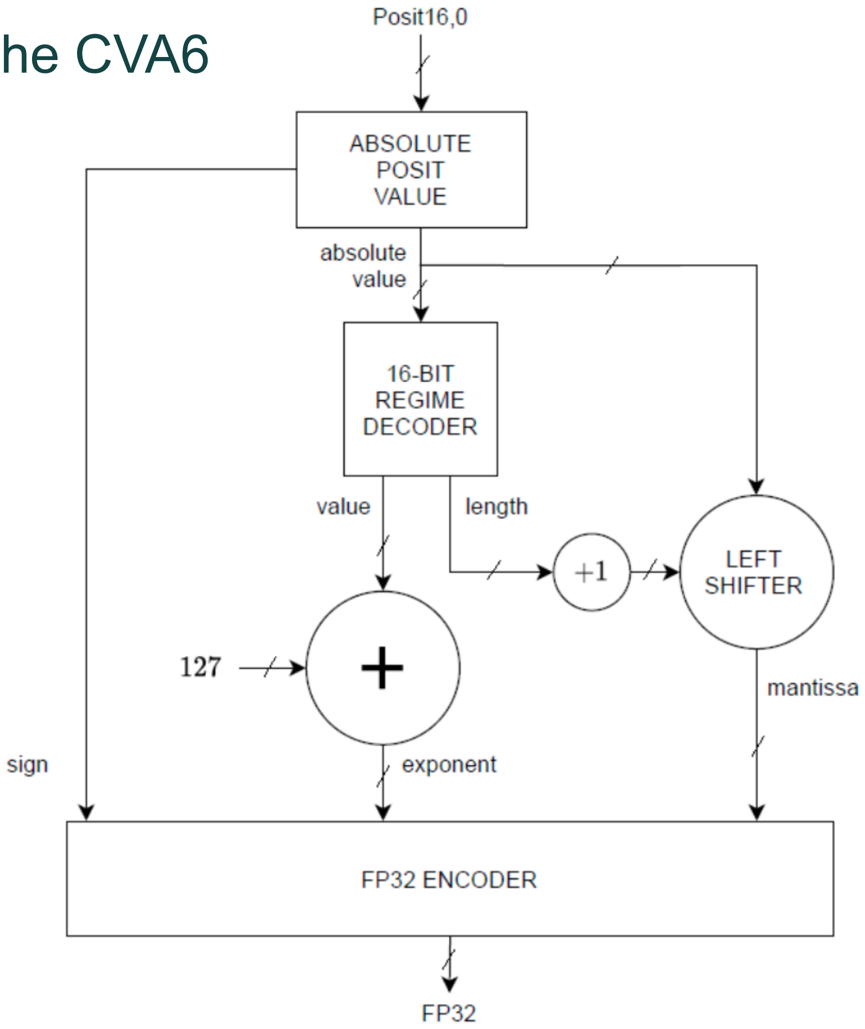
The PPU^{light} converts between Posit8.0, Posit16.0, Posit16.1 and 32-bit IEEE Floats or fixed point formats

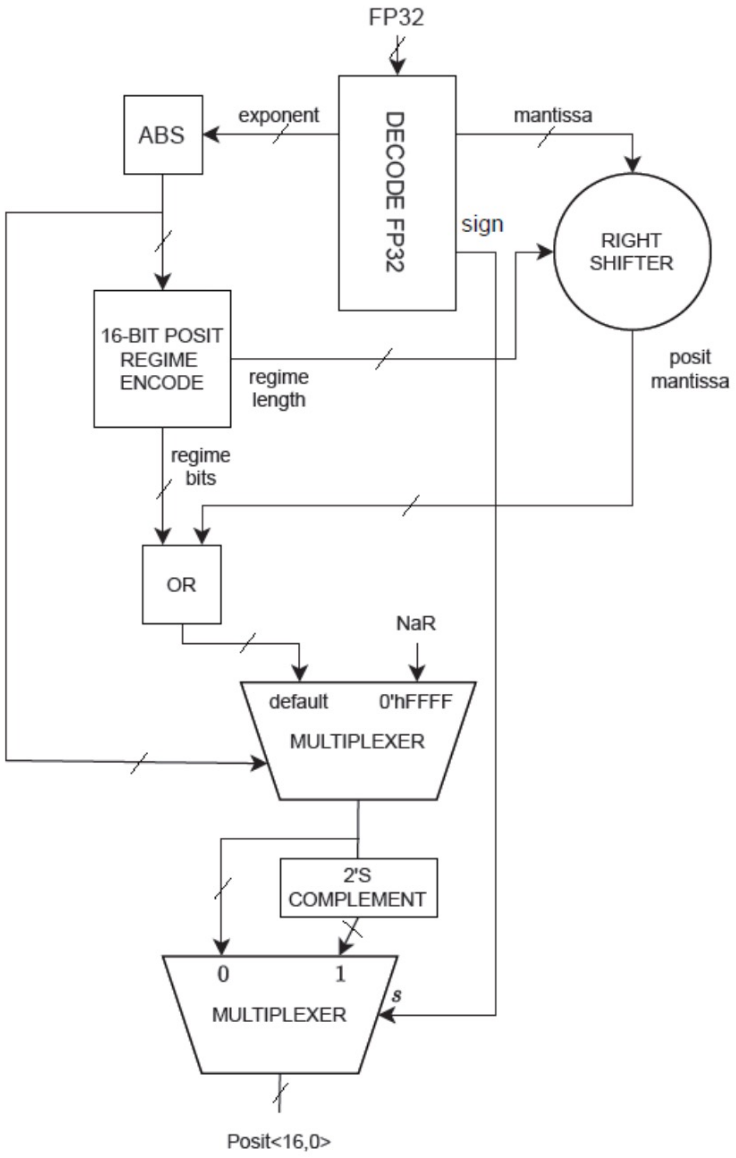
RISC-V extension implemented in the CVA6

Listing 6.1: *Intrinsic example for FCVT.S.P8*



```
int __fcvt_f32_p8 ( float a) {
 register float p1 asm (‘‘ fa0") = a;
  register int result asm (‘‘a1");
 __asm__ volatile (
  ‘‘"
   ‘‘. set rfs0 ,8\n"
   ‘‘. set rfs1 ,9\n"
   ...
   ‘‘. set op,0xb\n"
   ‘‘. set opf1,0x0\n"
   ‘‘. set opf2,0x2\n"
   ‘‘. set opf3,0x60\n"
   ‘‘. byte  op|(( r%[result]&1) <<7),
             ((r%[result]>>1)&0xF)|(opf1<<4)|((r%1&1)<<
             ((opf2&0xF) << 4) |  ((r%1>>1)&0xF),
             ((opf2>>4)&0x1)|(opf3<<1)"
  : [ result ] ‘‘=r"( result )
  :‘‘ f"(p1), ‘‘[ result ]"( result ));
 return result ;
}
```





AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

DII, Università di Pisa

Kalray SA

Outlook

EXPERIMENTAL EVALUATION OF FLOATING-POINT COMPRESSION

Standard pre-trained models in Caffe or TensorFlow with FP32 parameters

Round parameters to the nearest value representable in the alternate representation

Inference in FP32 on CPU and compute Accuracy-1 or Mean Average Precision (mAP)

Network	Criterion	FP32	FP16	BF16	FP8	Posit			
						<8,0>	<8,1>	<8,2>	<8,3>
VGG16	ACC-1	70.6	70.6	70.8	69.7	10.2	70.8	70.5	70
VGG19	ACC-1	70.1	70.1	70.3	67.9	4.8	70.1	69.9	70.6
ResNet50	ACC-1	75.7	71.3	75.5	62.8		27.7	73.2	66
						71.3	75.0	75.0	73.6
InceptionV3	ACC-1	71.1	71.1	71.3	44.8	65.1	69.4	69.7	63.1
						66.0	70.9	70.1	69.9
Xception	ACC-1	73.5	73.4	73.6	37.5	70.6	72.4	72.1	63.8
						72.1	72.6	72.8	68.8
MobileNetV2	ACC-1	71.2	71.2	71	0.2	12.7	12.3	11.0	3.2
						25.3	53.5	52.7	39.4
YOLOv3	mAP	0.41595	0.41595	0.41585	0.3022	0.4025	0.4155	0.411	0.394

DESIGN OF A POSIT8.N TO FP16 DECOMPRESSOR

Decompress Posit8.es with $es \in \{0,1,2,3\}$ to FP16, where es is a variable input

Conversion to FP16 to leverage the MPPA3 FP16.32 fused dot-product accumulate operator

Objectives

- Support the IEEE754 rounding modes: to nearest even; up; down; to zero
- Support of FP16 subnormals for the decompression of small magnitude Posit8.2 numbers

Challenges

- Previous work leverages the fact that Posit representations are symmetric wrt exponent, while IEEE754 formats support gradual underflow
- For Posit8.0 and Posit8.1, all values are representable as FP16
- For Posit8.2, 8 values of large magnitude are not representable as FP16 (representable with BF16)
- For Posit8.3, 46 values are not representable as FP16 (12 values not representable as BF16)

Insights into the conversion of Posit to FP representations

- Pre-detection of overflow or underflow is possible if $es \leq 3$ (FP16)
- Underflow detection works because Posit numbers with large magnitude have no mantissa bits

TSCM 16nm synthesis of Posit8 to FP16 decompression operators (1.25 GHz)

- Gains when specializing the operator to support round to nearest even only
- Further gains when only considering $es = 2$ (current Posit standard)
- Combinatorial decompression operator with one pipeline stage is best for our target technology

FP16.32 FUSED DOT-PRODUCT ACCUMULATE OPERATOR

Baseline implemented in the MPPA3 processor

Architecture

N. Brunie « Modified Fused Multiply and Add for Exact Low Precision Product Accumulation » 24th IEEE Symp. On Computer Arithmetic – ARITH 2017

Products

- FP16 values are multiplied
- 22-bit product is aligned to 80-bit fixed-point
- Optional two's complement is applied
- **Sum**
- The 80-bit fixed-point products are summed with an adder compression tree
- **Accumulation**
- The 32-bit floating-point accumulator is expanded to 128-bit fixed-point, which is added to the sum
- Normalization and a single rounding step produce the floating-point result

No hard limit on the depth of the dot-product

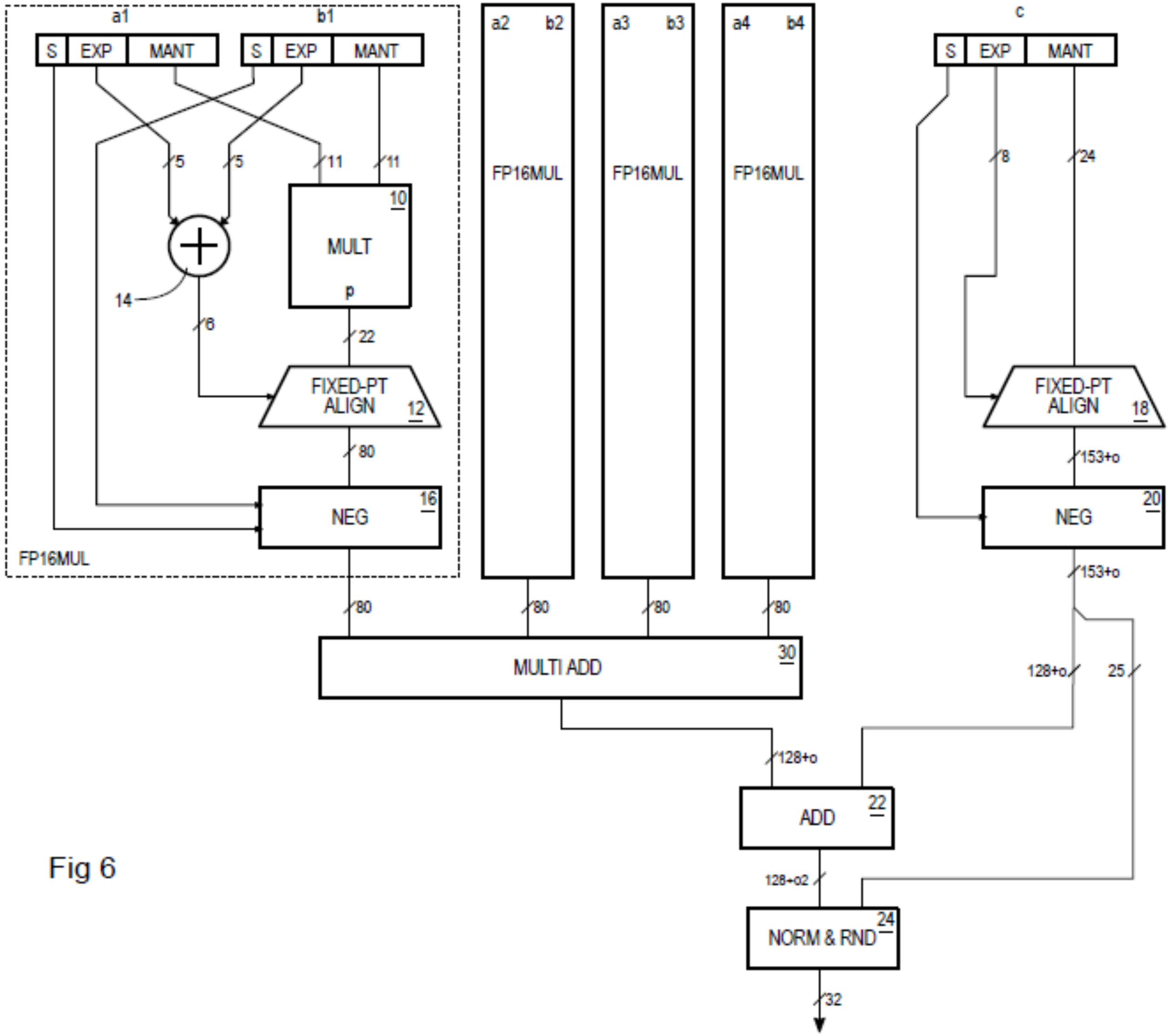


Fig 6

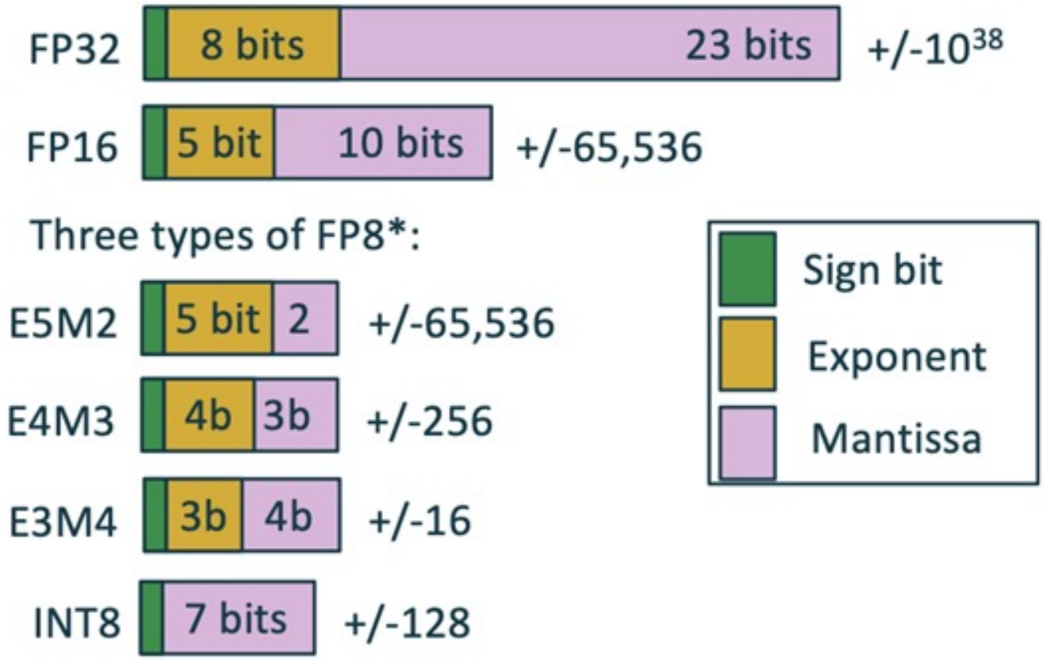
EXACT 32-TERM DOT-PRODUCT ACCUMULATE OPERATORS

Extension of this operator to 32-term accumulation and to other low bit-width multiplicands

Sum terms in full-precision into a wide fixed-point Kulisch accumulator (instead of FP32)

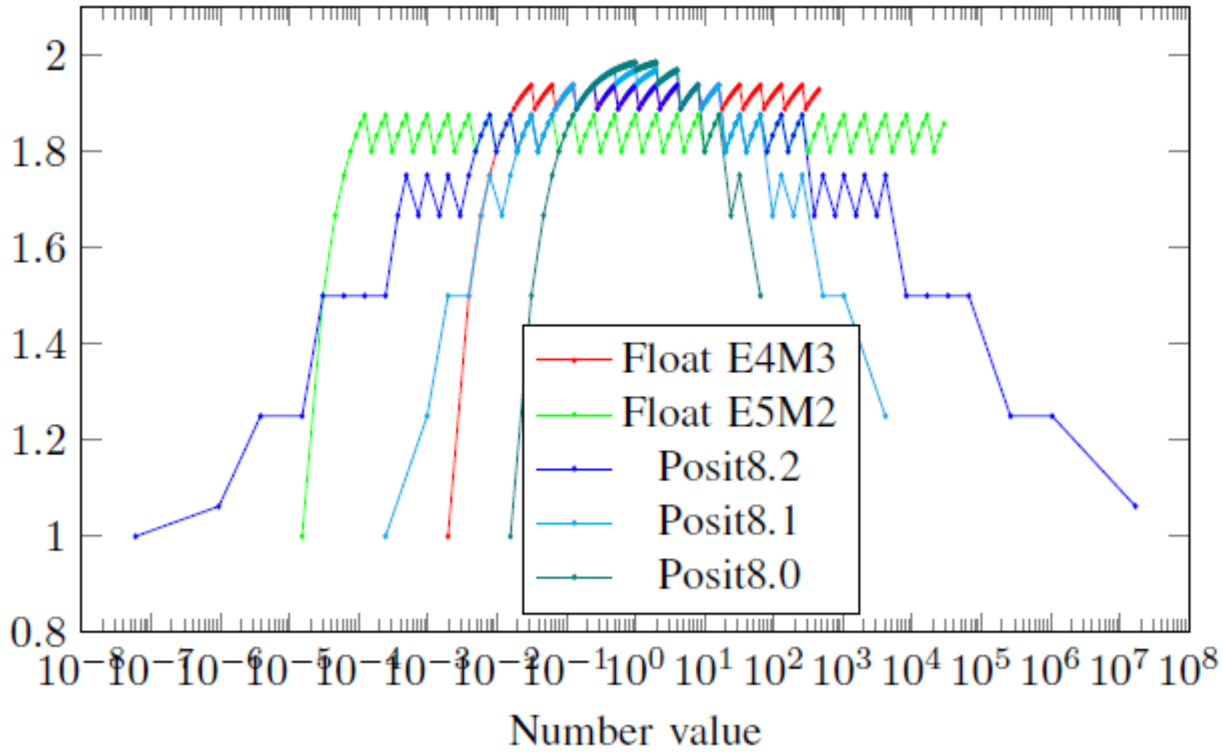
The FP8 formats E5M2 and E4M3 are supported in hardware by NVIDIA, Arm, Intel, IBM, Qualcomm, Graphcore, AMD, Tesla, Habana, … and are proposed for IEEE standardization

IEEE 754 low bitwidth representations for AI



*No IEEE standard for FP8

Relative distance between 2 consecutive 8-bit real numbers for the representation



EXACT 32-TERM DOT-PRODUCT ACCUMULATE OPERATORS

Design the dot-product to wide accumulator with a fork of the FloPoCo VHDL generator

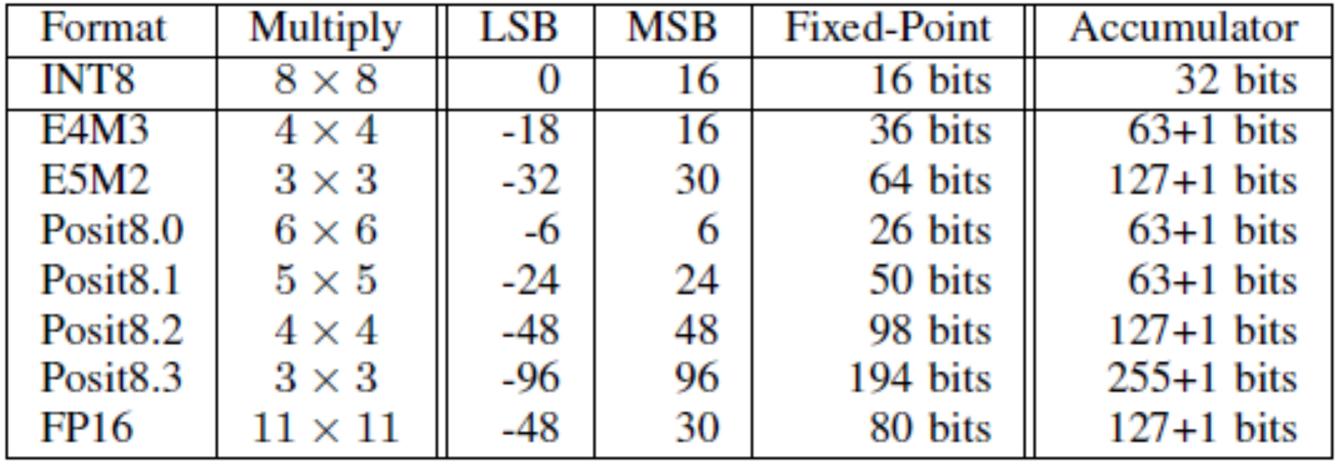
Operator structure

- Decompress the multiplicands
- Multiply the unsigned mantissas
- Convert the product to two's complement representation
- Arithmetic shift right the product into a w-bit wide datapath
- Sum the terms using a compression tree and add to the accumulator

Synthesize one-stage operator with a clock period of 5 times the target clock period

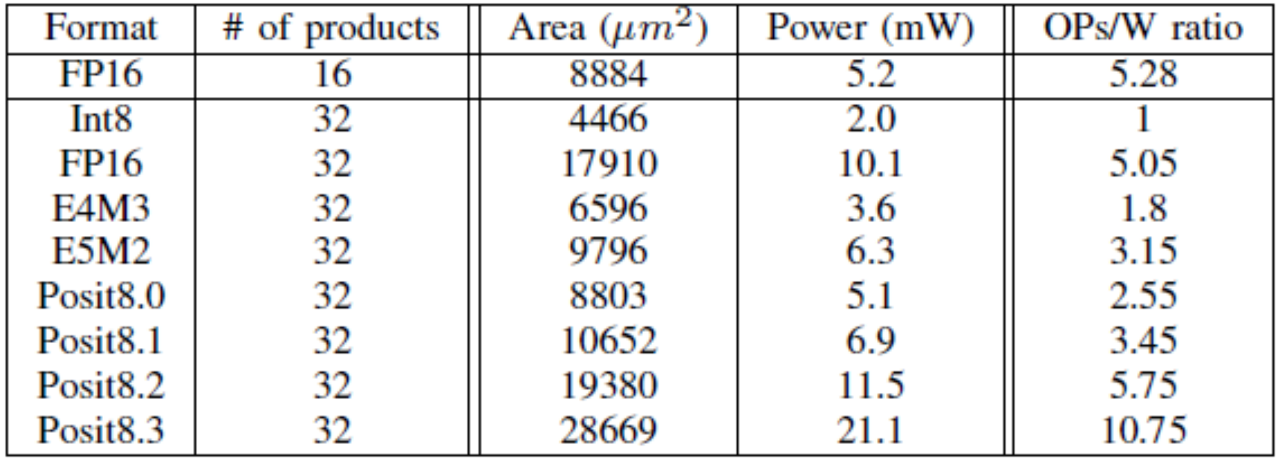
Posit8.2 power and area comparable to FP16

Multiplier, product and accumulator sizes



Format	Multiply	LSB	MSB	Fixed-Point	Accumulator
INT8	8 × 8	0	16	16 bits	32 bits
E4M3	4 × 4	-18	16	36 bits	63+1 bits
E5M2	3 × 3	-32	30	64 bits	127+1 bits
Posit8.0	6 × 6	-6	6	26 bits	63+1 bits
Posit8.1	5 × 5	-24	24	50 bits	63+1 bits
Posit8.2	4 × 4	-48	48	98 bits	127+1 bits
Posit8.3	3 × 3	-96	96	194 bits	255+1 bits
FP16	11 × 11	-48	30	80 bits	127+1 bits

TSCM 16nm synthesis without pipelining (250 MHz)



Format	# of products	Area (μm^2)	Power (mW)	OPs/W ratio
FP16	16	8884	5.2	5.28
Int8	32	4466	2.0	1
FP16	32	17910	10.1	5.05
E4M3	32	6596	3.6	1.8
E5M2	32	9796	6.3	3.15
Posit8.0	32	8803	5.1	2.55
Posit8.1	32	10652	6.9	3.45
Posit8.2	32	19380	11.5	5.75
Posit8.3	32	28669	21.1	10.75

AGENDA

European Processor Initiative SGA-1

European Processor Initiative SGA-2

IIS, ETH Zurich

IST, Universidade de Lisboa

DII, Università di Pisa

Kalray SA

Outlook

POSIT-RELATED PUBLICATIONS FROM EPI PARTNERS

- M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara, B. D. De Dinechin, “Novel arithmetics in Deep Neural Networks signal processing for autonomous driving: challenges and opportunities”, IEEE Signal Proc Mag. 2020.
- M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara Fast Approximations of Activation Functions in Deep Neural Networks when using Posit Arithmetic. Sensors, 20, 1515, 2020.
- M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara, “Fast deep neural networks for image processing using posits and ARM scalable vector extension”, Journal of Real Time Image Processing 2020.
- M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara, “Small Reals Representations for Deep Learning at the Edge: A Comparison” CoNGA 2022.
- M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara, “A Lightweight Posit Processing Unit for RISC-V Processors in Deep Neural Network Applications”, IEEE Trans. Emerg. Top. Comput. 10(4): 1898-1908, 2022.
- L. Crespo et al. "Unified Posit/IEEE-754 Vector MAC Unit for Transprecision Computing", TCAS-II, 2022.
- N. Neves et al. "A reconfigurable posit tensor unit with variable-precision arithmetic and automatic data streaming", JSPS (VLSI), 2021.
- N. Neves et al. "Dynamic Fused Multiply-Accumulate Posit Unit with Variable Exponent Size for Low-Precision DSP Applications", SIPS Workshop, 2020.
- N. Neves et al. "Reconfigurable Stream-based Tensor Unit with Variable-Precision Posit Arithmetic", ASAP, 2020.
- D. Resmerita, R. C. Farias, B. Dupont de Dinechin, L. Fillatre "Benchmarking Alternative Floating-Point Formats for Deep Learning Inference" Compas’2020, Lyon, France, July 2020.
- O. Desrentes, D. Resmerita, B. Dupont de Dinechin "A Posit8 Decompression Operator for Deep Neural Network Inference", Conference on Next-Generation Arithmetic 2022.

THANK YOU

KALRAY

THE POWER OF MORE

www.kalrayinc.com

DISCLAIMER

Kalray makes no guarantee about the accuracy of the information contained in this document. It is intended for information purposes only and shall not be incorporated into any contract. It is not a commitment to deliver any material, code or functionality, and should not be relied upon in making purchasing decisions. The development, release and timing of any features or functionality described for Kalray products remains at the sole discretion of Kalray.

Trademarks and logos used in this document are the properties of their respective owners.

KALRAY

THE POWER OF MORE

www.kalrayinc.com