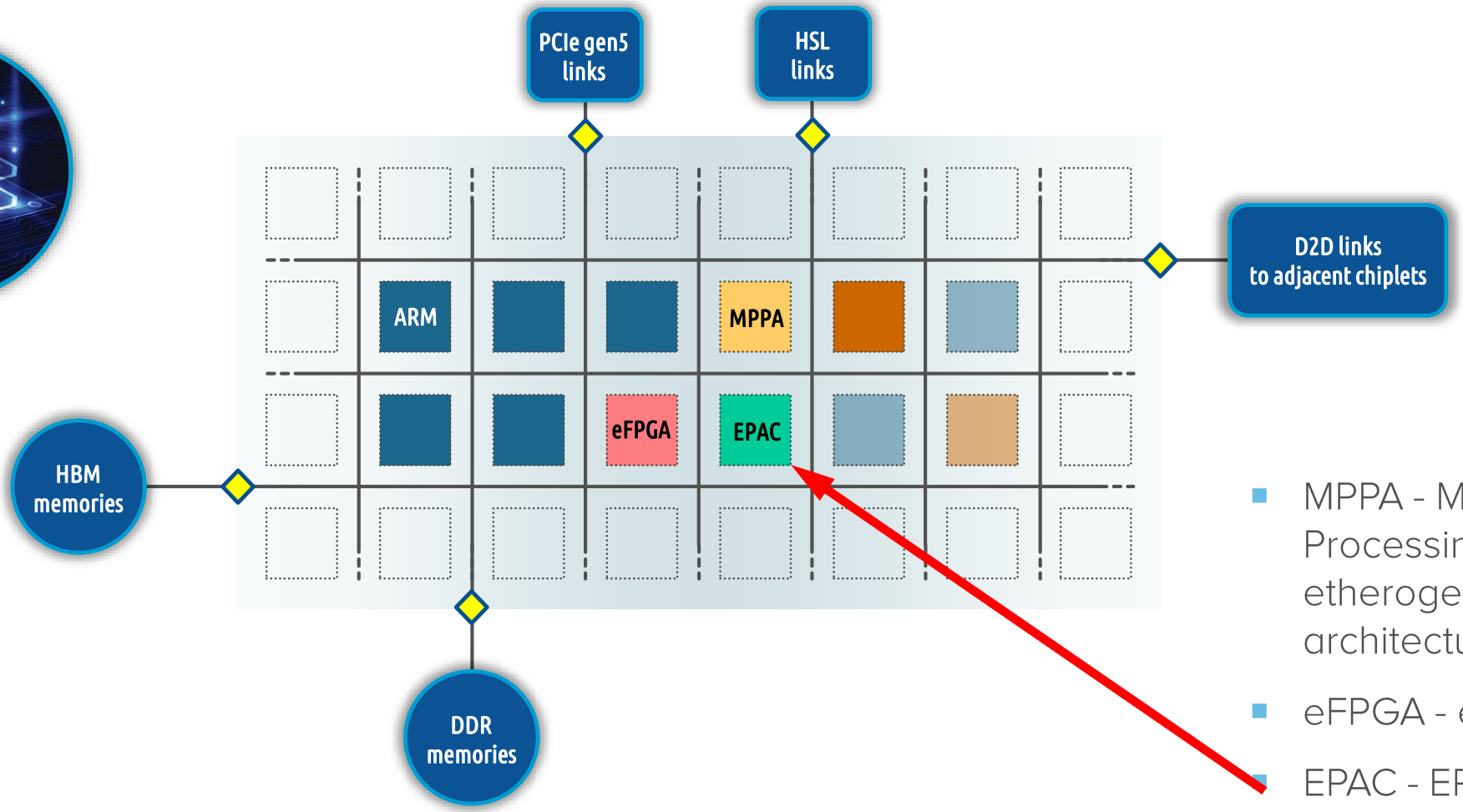
ACCELERATOR TILE

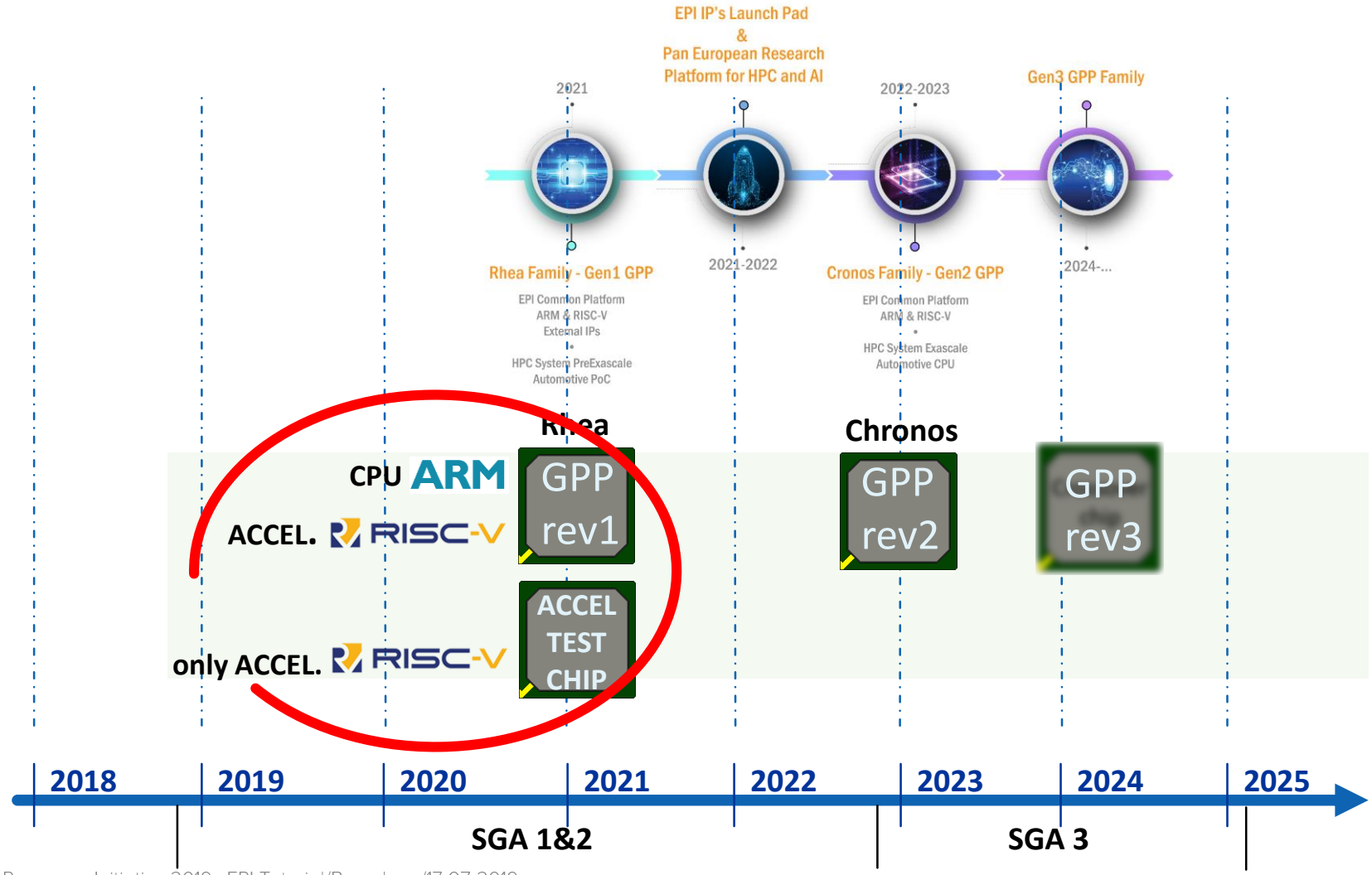
MAURO OLIVIERI

RECALL… THE GPP AND COMMON PLATFORM ARCHITECTURE

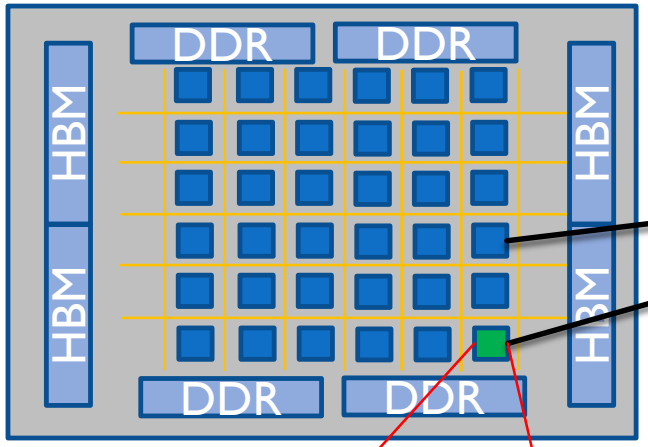


- MPPA - Multi-Purpose Processing Array etherogeneous architecture
- eFPGA - embedded FPGA
- EPAC - EPI Accelerator

ACCELERATOR STREAM IN EPI ROADMAP

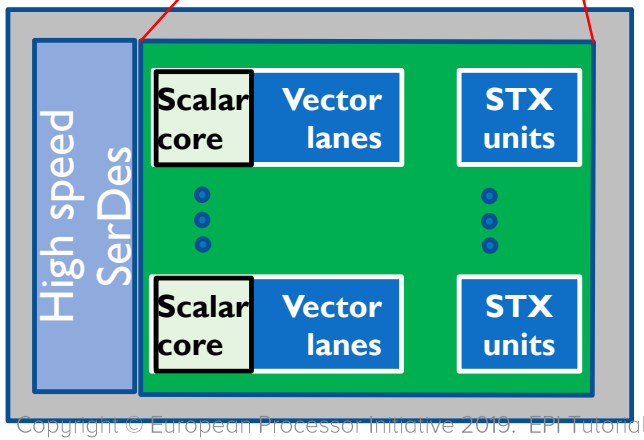


1ST GENERATION EPI CHIPS



General Purpose Processor (GPP) chip

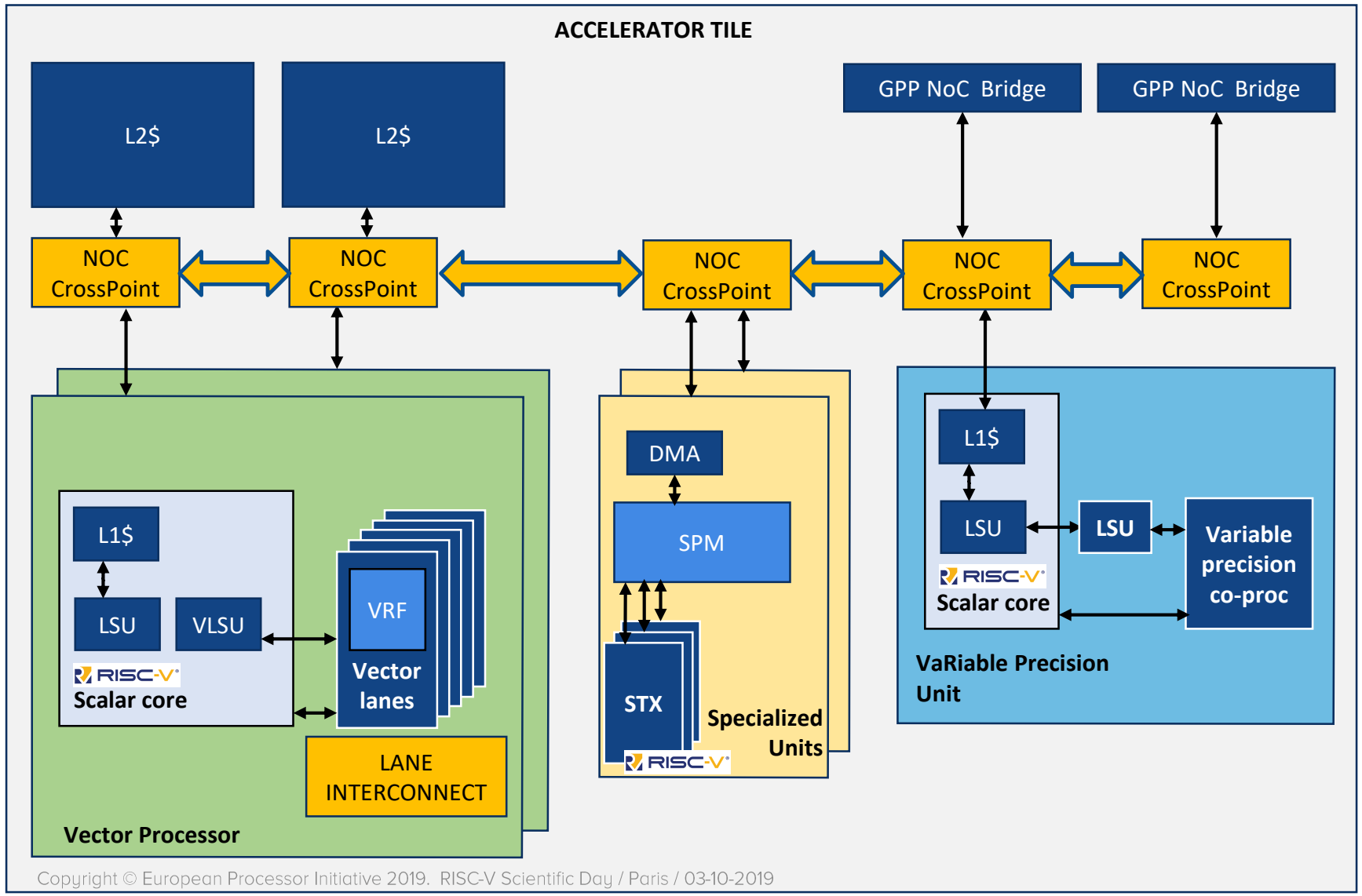
- 6 nm, chip-let technology
- ARM-SVE tiles
- EPAC RISC-V vector+AI accelerator tiles
- L1, L2, L3 cache subsystem + HBM + DDR



RISC-V Accelerator Demonstrator Test Chip

- 22 nm FDSOI
- Only one RISC-V accelerator tile
- On-chip L1, L2 + off-chip HBM + DDR PHY
- Targets 16 DP GFLOPS per core (vector processor only)

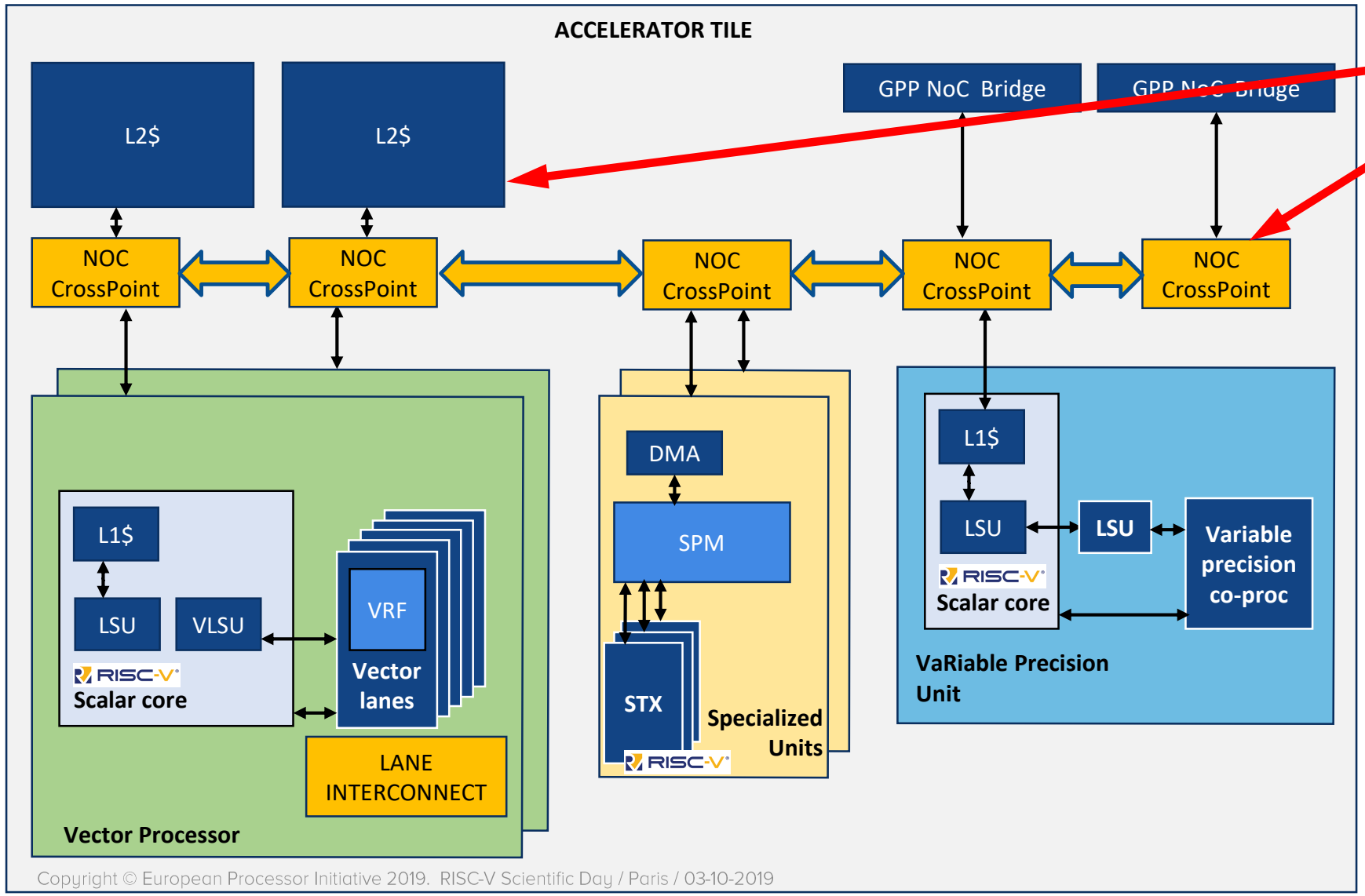
EPAC ARCHITECTURE VIEW



CONTRIBUTING PARTNERS:

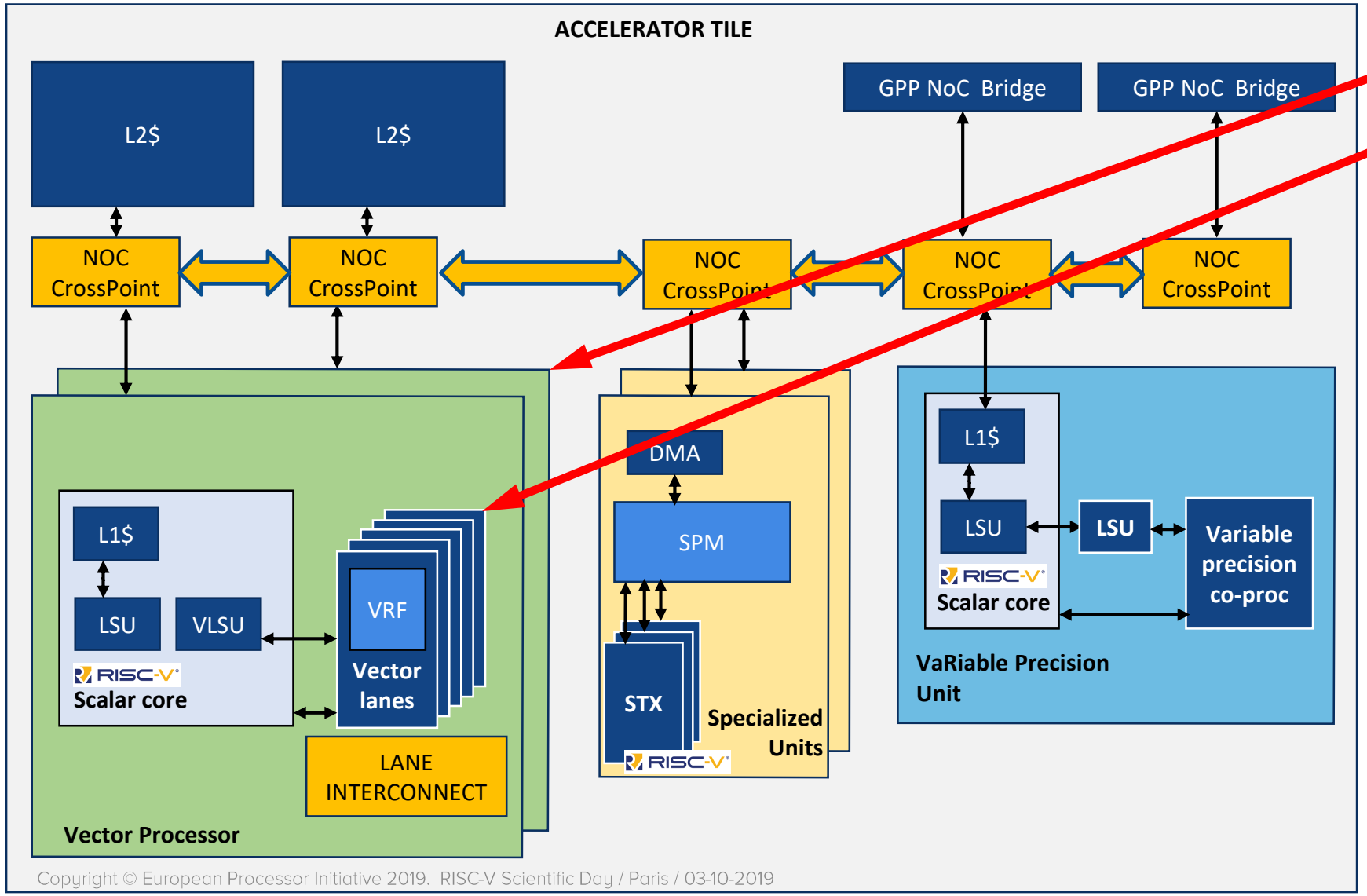
- BSC
- CEA
- Chalmers
- E4
- ETH Zurich
- Extoll
- FORTH
- Fraunhofer
- Semidynamics

EPAC ARCHITECTURE VIEW



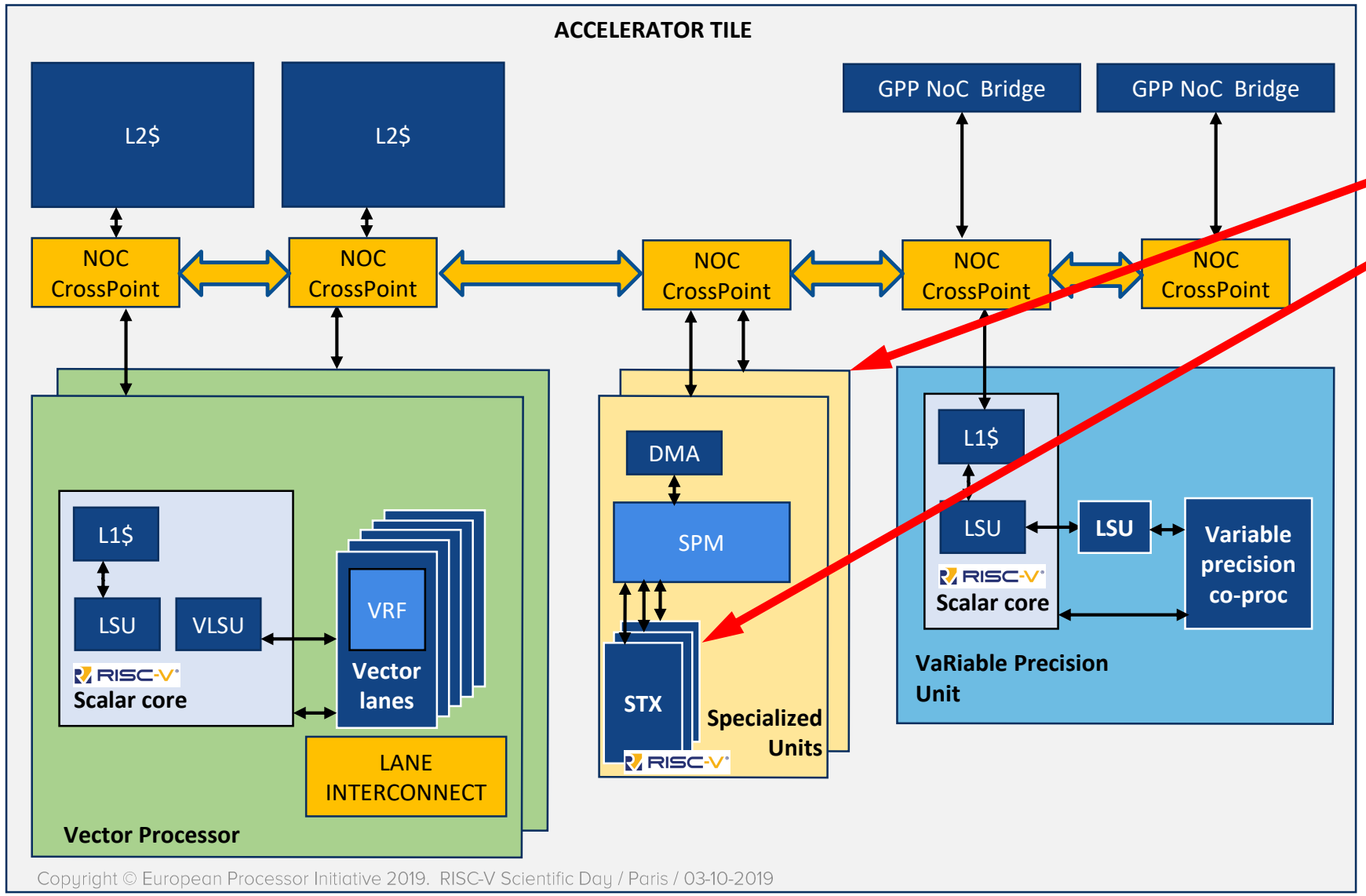
- Shared L2 cache banks
- Cache coherent NoC

EPAC ARCHITECTURE VIEW



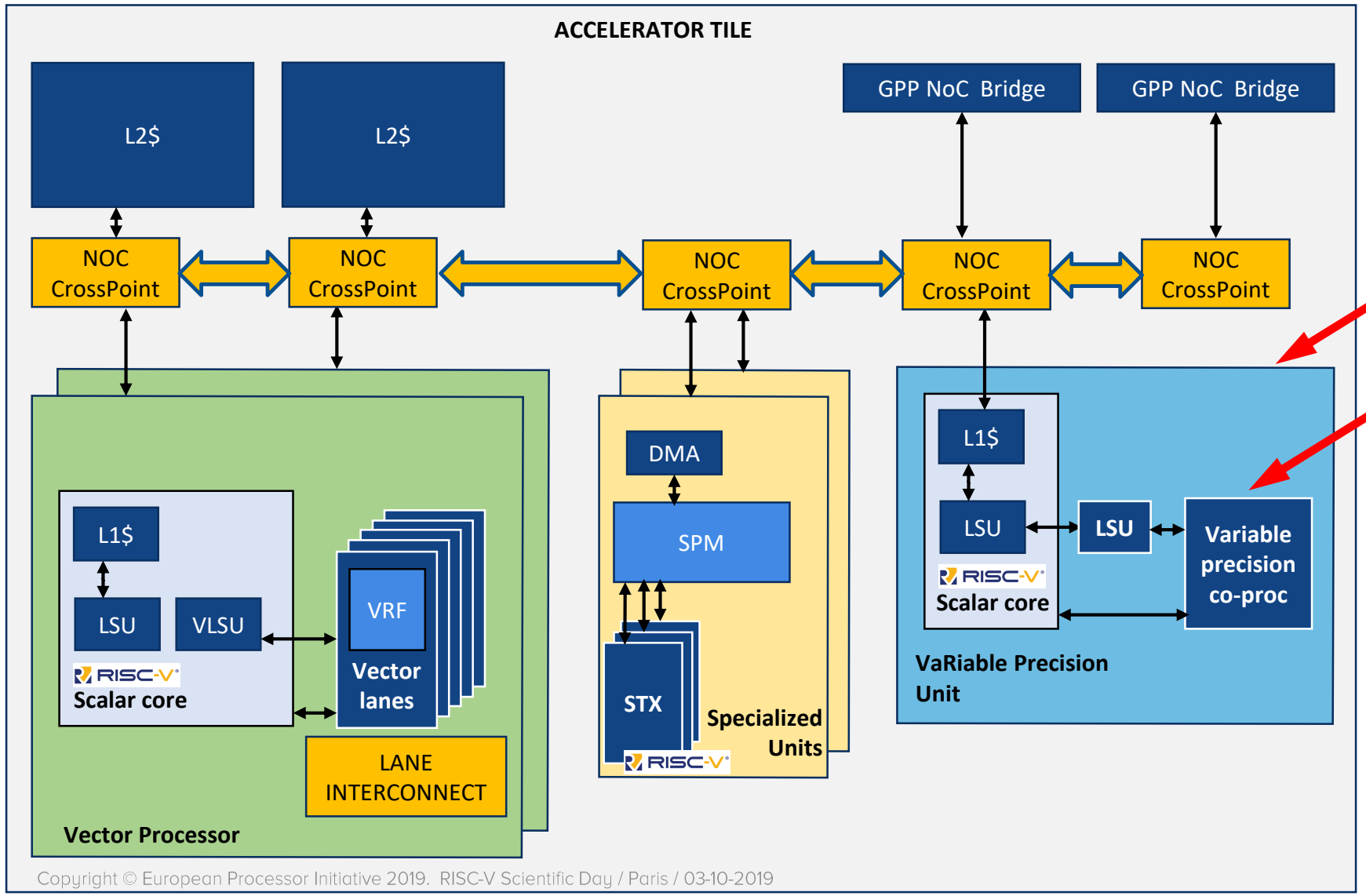
- Up to 8 vector processors
- Vector Lanes act as tightly coupled (ISA mapped) acceleration units to the scalar core in the vector processor
- Heavily pipelined
- RISC-V vector extension compliant

EPAC ARCHITECTURE VIEW



- Up to 8 Specialized Units
- The STX Units act as loosely coupled, memory mapped acceleration units to the scalar cores
- Fast single-cycle MACs in parallel

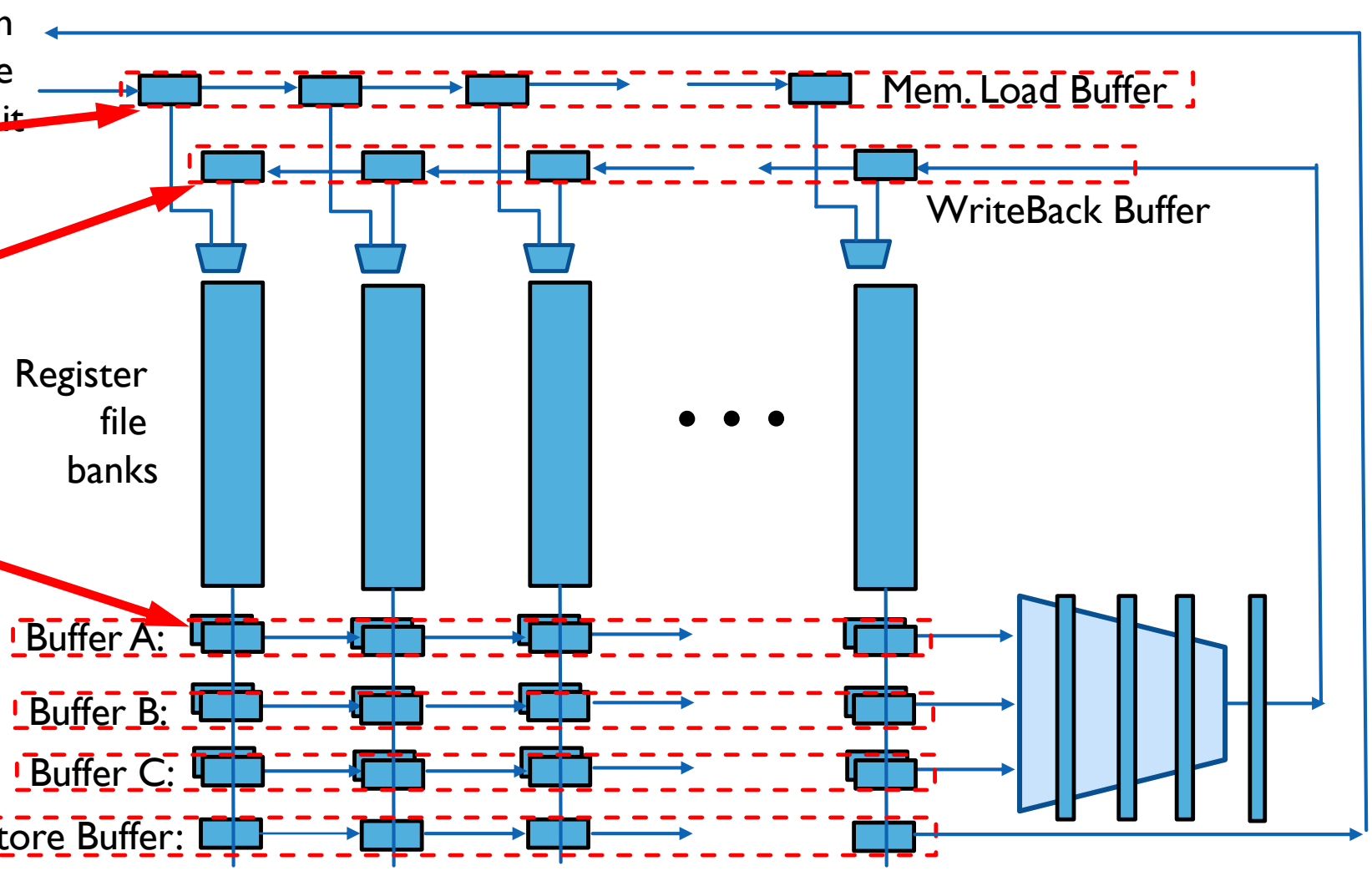
EPAC ARCHITECTURE VIEW



- At least one Variable Precision unit
- The variable precision co-processor act as execution unit extension to the scalar core

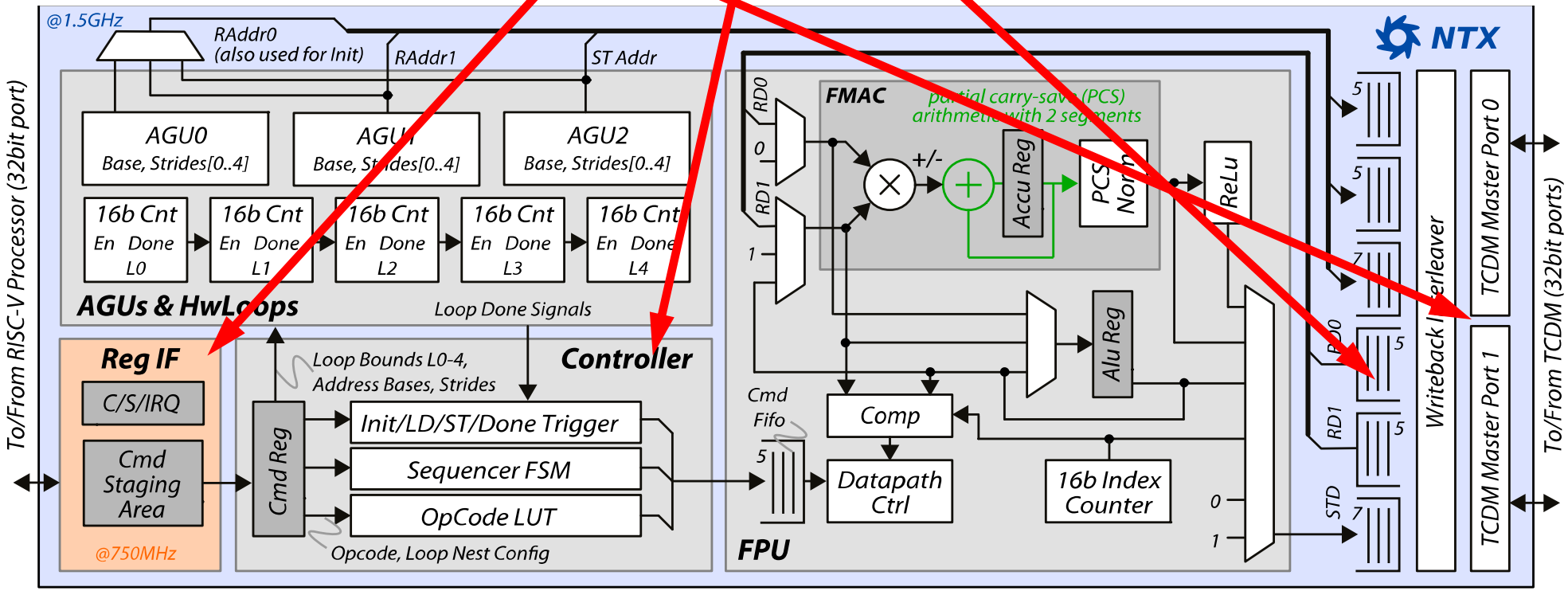
VECTOR LANE MICROARCHITECTURE (SIMPLIFIED VIEW)

Load buffer: holding data coming from memory
Load buffer is replicated and managed by a dedicated load unit to handle out-of-order element transfers
WriteBack Buffer: holding operation results
Buffer A, B, C: operand buffers, Buffers A,B,C are doubled (shadowed) to allow single-cycle full refill while the shadow buffer is being consumed.
Store Buffer: holding data to be sent to memory



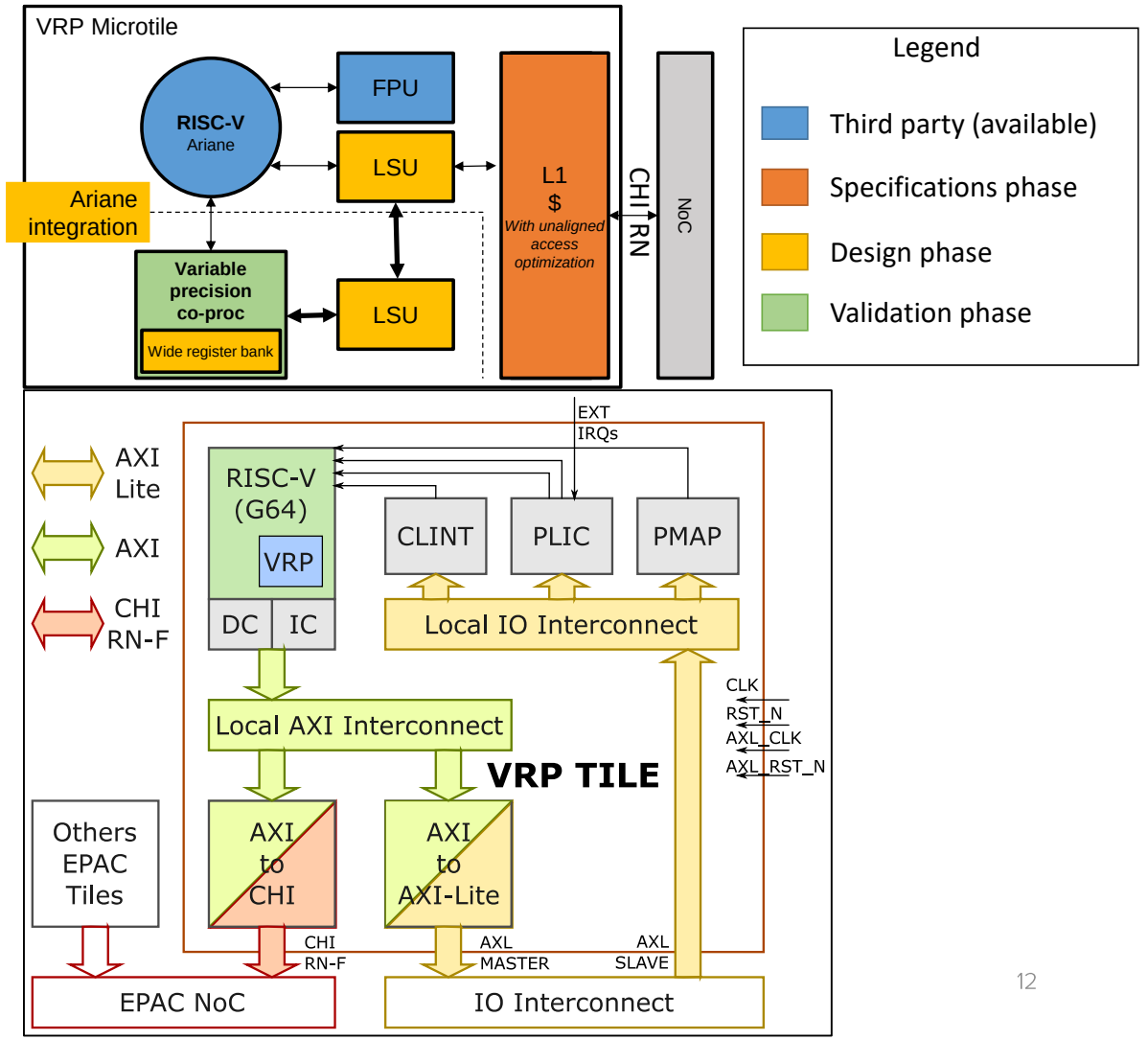
STX COPROCESSOR MICROARCHITECTURE

- Processor configures operation via **memory-mapped registers**
- Controller issues AGU, HWL, and FPU micro-commands based on configuration
- Reads/writes data via **2 memory ports** (2 operand and 1 writeback streams)
- FIFOs help buffering data path and memory latencies



VRP (VARIABLE PRECISION CO-PROCESSOR) ARCHITECTURE

- Support extended precision for accumulator and intermediate result
- Less overhead and more flexible than quad precision (128 bits) with adjustable precision from 64 bits up to 256 bits
- Accelerator integrated as an execution unit of Ariane RISCV core
  - Additional instructions for variable precision computation
- Compiler and library support
  - LLVM support with intrinsics
  - BLAS with variable precision



THANK YOU

