RISC-V Days Tokyo 2021 Autumn
November 17, 2021

esperanto.ai

Accelerating ML Workloads with Energy-Efficient High-Performance RISC-V Processors with Custom ML Extensions

Eiji Kasahara

Esperanto Technologies
eiji.kasahara@esperanto.ai

Esperanto Technologies

A Foundational RISC-V Company Leading the AI Revolution

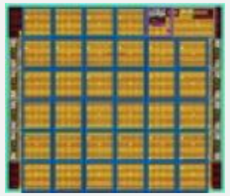
Industry Leadership:

Architecture, IP, AI, Silicon

Our team includes a co-inventor of RISC, CPU/IP veterans, data science & AI expertise, and deep microarchitecture capability specializing in low power circuit design

We are a RISC-V pioneer and an early adopter of RISC-V for CPU design, AI silicon development, and software & tools creation

Our low power RISC-V architecture can span from datacenter to edge applications for AI and non-AI workloads



Esperanto's RISC-V based ET-SoC-1 is capable of processing key AI workloads at 20W of power in hosted and self-hosted systems

Supported by Tier 1 Investors, Hyperscalers, Enterprise, and SoC Companies

Esperanto ET-SoC-1 is the World’s Highest Performance Commercial RISC-V Chip

esperanto.ai

The ET-SoC-1 is fabricated in TSMC 7nm

- 24 billion transistors
- Die area: 570 mm^2
- Over 30,000 bumps

1088 ET-Minion energy-efficient 64-bit RISC-V processors

- Each is multithreaded with an attached vector/tensor unit
- Typical operation 500 MHz to 1.5 GHz expected

4 ET-Maxion 64-bit high-performance RISC-V out-of-order processors

- Typical operation 500 MHz to 2 GHz expected

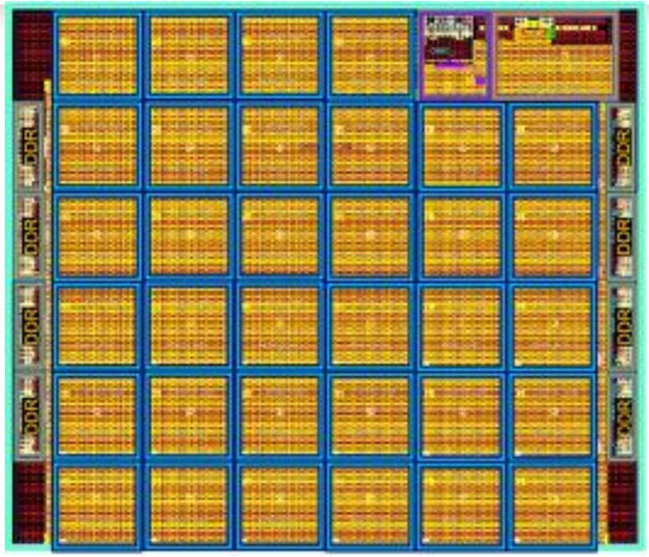
Over 160 MB of on-die SRAM and up to 32GB of external DRAM per chip

Root of trust for secure boot

Power typically < 20 watts, software can allow maximum to go higher

Package: 45x45mm with 2494 balls to PCB

Status: Silicon currently undergoing bring-up and validation



Requirements and Challenges for ML Recommendation in Large Datacenters

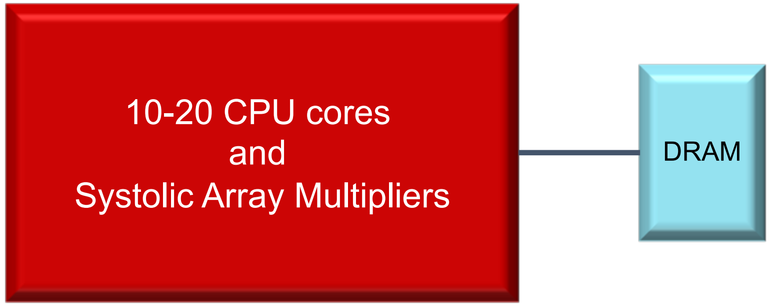
Today, most inferencing workloads for recommendation systems in large data centers are run on x86 servers

Often these servers have an available slot for an accelerator card, but an accelerator card needs to meet key requirements:

- **100 TOPS to 1000 TOPS** peak rates to provide better performance than the x86 host CPU alone
- Limited power budget per card, perhaps **75 to 120 watts**, and must be air-cooled
- Strong support for **Int8**, but must also support **FP16 and FP32** data
- **~100 GB** of memory capacity on the accelerator card to hold most embeddings, weights and activations
- **~100 MB** of on-die memory
- Handle both **dense and sparse** compute workloads. Embedding look-up is sparse matrix by dense matrix multiplication
- Be **programmable** to deal with rapidly evolving workloads, rather than depending on overly-specialized hardware

Esperanto's Better Approach for ML Recommendation

Other ML Chip approaches



One Giant Hot Chip uses up power budget

Limited I/O pin budget limits memory BW

Dependence on systolic array multipliers

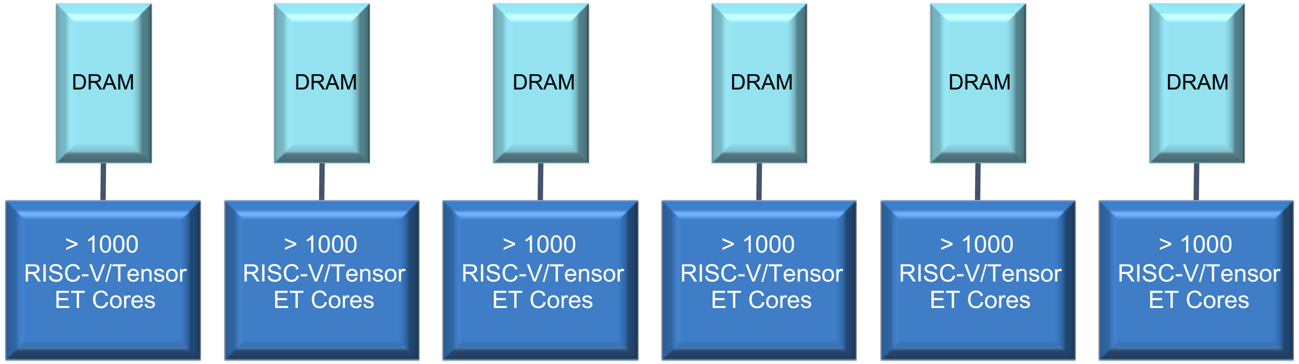
- Great for high ResNet50 score
- Not so good with large sparse memory

Only a **handful (10-20) of CPU cores**

- **Limited parallelism** with CPU cores when problem doesn't fit onto array multiplier

Standard voltage: Not energy efficient

Esperanto's better approach


DRAM
DRAM
DRAM
DRAM
DRAM
DRAM
> 1000 RISC-V/Tensor ET Cores
> 1000 RISC-V/Tensor ET Cores
> 1000 RISC-V/Tensor ET Cores
> 1000 RISC-V/Tensor ET Cores
> 1000 RISC-V/Tensor ET Cores
> 1000 RISC-V/Tensor ET Cores

Use **multiple low-power** chips that, combined, fit within power budget

Performance, pins, memory, bandwidth **scale up with more chips**

Thousands of general-purpose RISC-V/tensor cores

- **Far more programmable** than overly-specialized hardware
- **Thousands of threads** help with large sparse memory latency

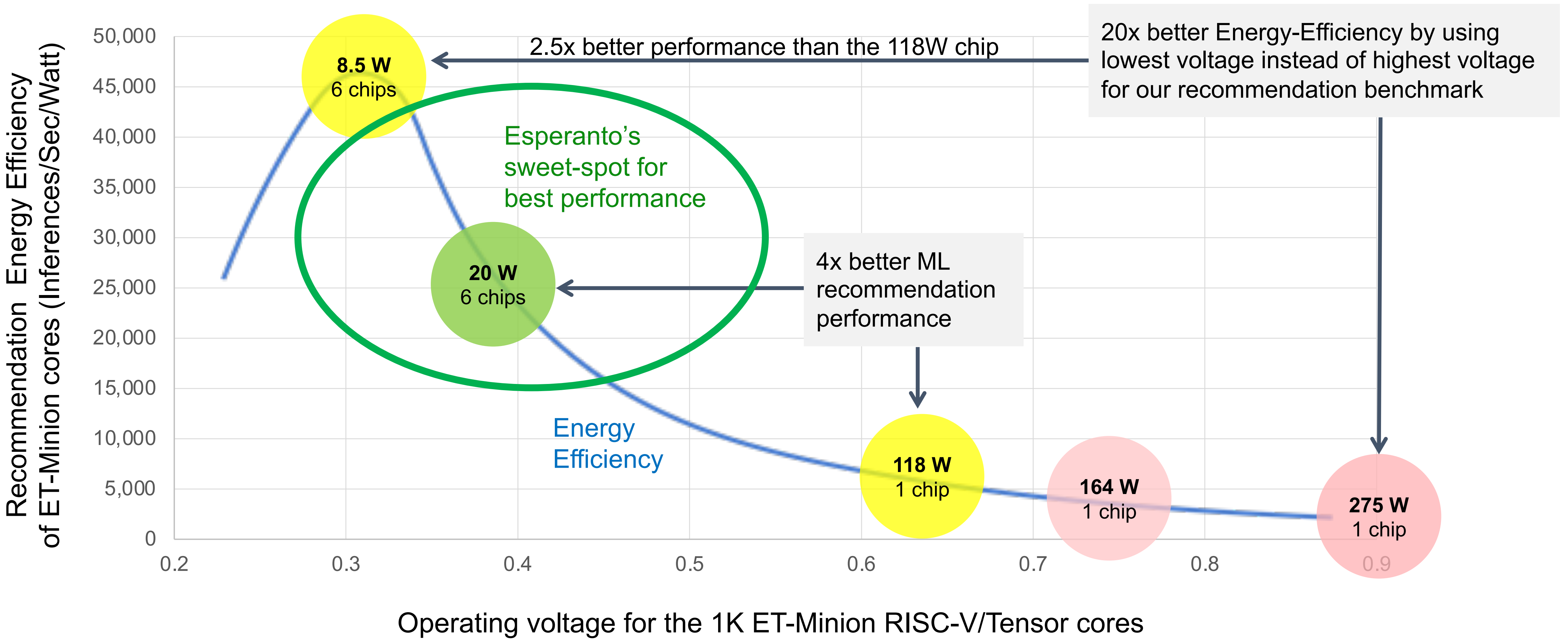
Full parallelism of thousands of cores always available

Low-voltage operation of transistors **is more energy-efficient**

- Lower voltage operation also reduces power
- Requires both **circuit and architecture innovations**

Challenge: How to put the highest ML Recommendation performance onto a single accelerator card with a 120-watt limit?

Six Esperanto Chips Working Together Achieve Best ML Performance in 120 Watts

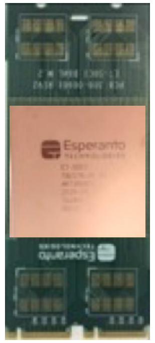


2.5x better performance than the 118W chip

20x better Energy-Efficiency by using lowest voltage instead of highest voltage for our recommendation benchmark

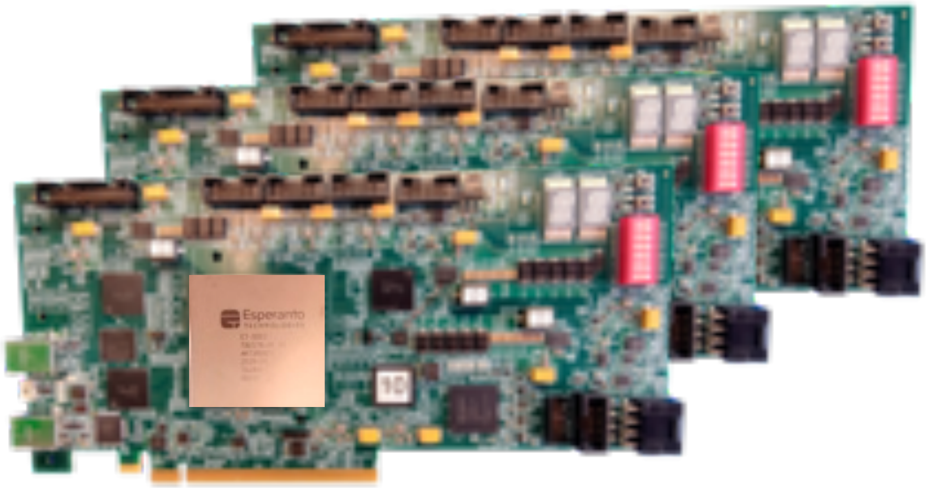
4x better ML recommendation performance

Esperanto’s ET-SoC-1 AI Accelerator Card Portfolio



Dual M.2 and Glacier Point v2 Cards:

- OCP compliant
- Implemented using Dual M.2 form factors
- Fits up to six ET-SoC-1 chips per card GP v2 card
- 6,558 total RISC-V cores
- Up to 192 GB of DRAM
- Up to 822 GB/s DRAM bandwidth
- ~120 W total power consumption

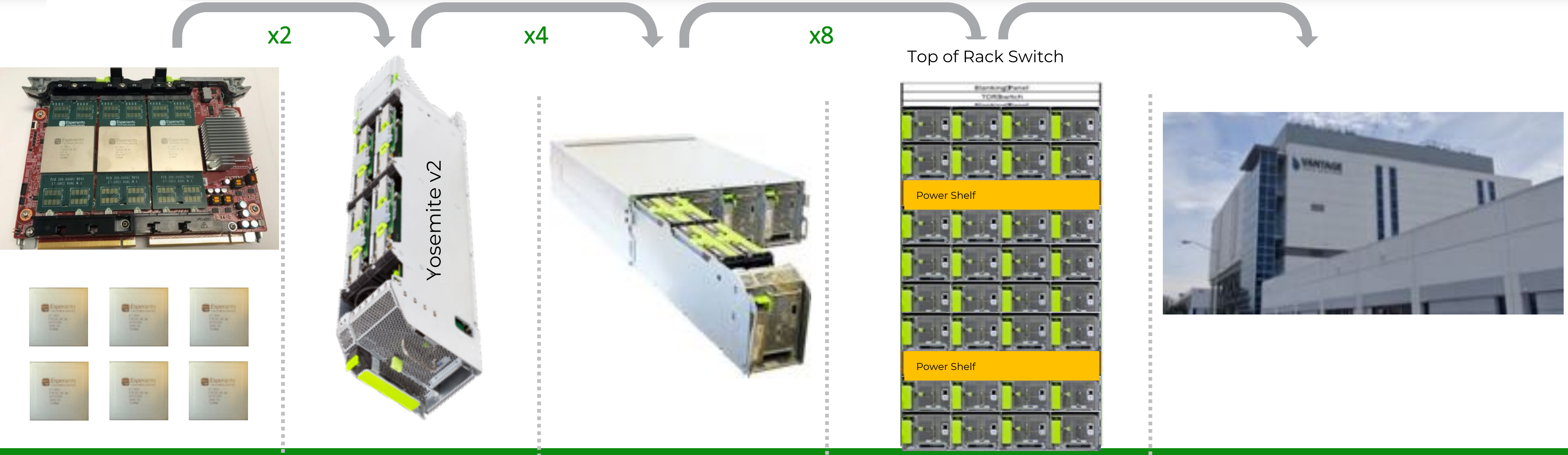


PCI Express Cards:

- PCI-Small (HHHL) @ 20W, 16GB RAM
- Fits up to four ET-SoC-1 chips per PCI-Large (FHML) @ 100W, 64GB RAM
- Fits up to six ET-SoC-1 chips per PCI-Xlarge (FHFL) @ 140W, 96 GB RAM
- ET-SoC-1 power can be configured from 10W to to 60W per chip

ET-SoC-1 can be deployed at scale in existing OCP Data Centers

esperanto.ai



OCP Glacier Point v2 accelerator card holds:

- 6 Esperanto ET-SoC-1 chips

Yosemite v2 Sled holds:

- 2 Glacier Point v2 accelerator cards
- 12 ET-SoC-1 chips

Yosemite v2 Cubby holds:

- 4 Yosemite Sleds
- 8 accelerator cards
- 48 ET-SoC-1 chips

Rack with Yosemite v2 holds:

- 8 Yosemite v2 Cubbies
- 64 accelerator cards
- 384 ET-SoC-1 chips

Example OCP Data Center holds:

- @ 30 sq. ft. per OCP rack [7]
- Estimated 4K-20K racks / data center
- Millions of ET-SoC-1 chips

ET-Minion is an Energy-Efficient RISC-V CPU With a Custom Vector/Tensor Unit

esperanto.ai

ET-Minion is a custom built 64-bit RISC-V processor

- Architecture and circuits optimized to enable low-voltage operation
- In-order pipeline with low gates/stage to improve MHz at low voltages
- Two hardware threads of execution
- Software configurable L1 data-cache and/or scratchpad

Each Minion includes an ML-optimized vector/tensor unit

512-bit wide integer per cycle

- 128 8-bit integer operations per cycle, accumulates to 32-bit Int

256-bit wide floating point per cycle

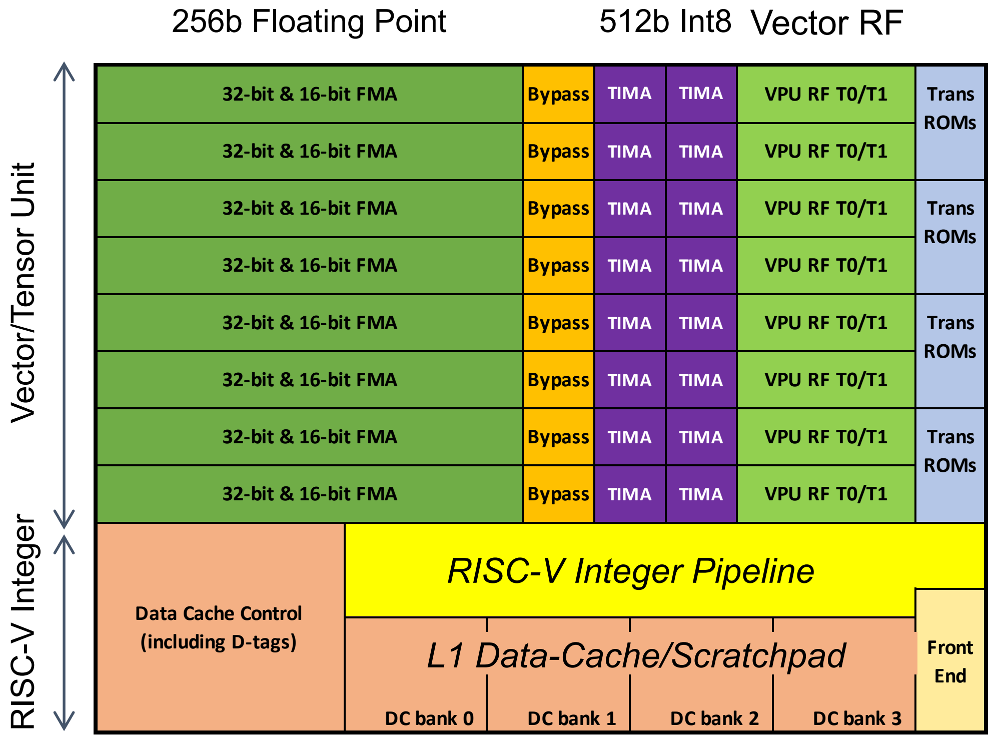
- 16 32-bit single precision operations per cycle
- 32 16-bit half precision operations per cycle

New multi-cycle tensor instructions

- Can run for up to 512 cycles (or up to 32K operations) with one tensor instruction
- Reduces instruction fetch bandwidth and reduces power
- RISC-V integer pipeline put to sleep during tensor instructions

Vector transcendental instructions

Operating range: 300 MHz to 2 GHz



256b Floating Point	512b Int8			Vector RF	
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	Trans ROMs
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	Trans ROMs
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	Trans ROMs
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	Trans ROMs
32-bit & 16-bit FMA	Bypass	TIMA	TIMA	VPU RF T0/T1	

Vector/Tensor Unit

RISC-V Integer

Data Cache Control (including D-tags)

RISC-V Integer Pipeline

L1 Data-Cache/Scratchpad

DC bank 0

DC bank 1

DC bank 2

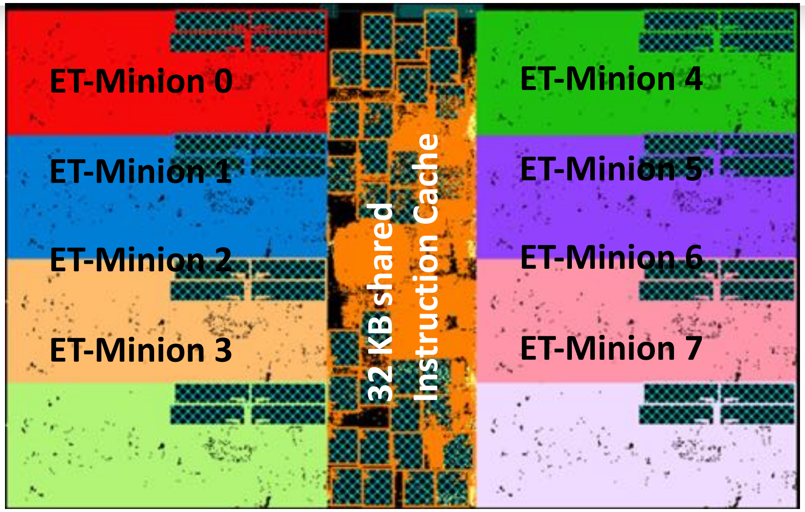
DC bank 3

Front End

ET-Minion RISC-V Core and Vector/Tensor unit optimized for low-voltage operation to improve energy-efficiency

Optimized for energy-efficient ML operations.
Each ET-Minion can deliver peak of 128 Int8 GOPS per GHz

8 ET-Minions form a “Neighborhood”



NEIGHBORHOOD CORES WORK CLOSELY TOGETHER

- Architecture improvements capitalize on physical proximity of 8 cores
- Take advantage that almost always running highly parallel code

OPTIMIZATIONS FROM CORES RUNNNING THE SAME CODE

- 8 ET-Minions share single large instruction cache,
  this is more energy efficient than many separate instruction caches.
- “Cooperative loads” substantially reduce memory traffic to L2 cache

NEW INSTRUCTIONS MAKE COOPERATION MORE EFFICIENT

- New Tensor instructions dramatically cut back on instruction fetch bandwidth
- New instructions for fast local synchronization within group
- New Send-to-Neighbor instructions
- New Receive-from-Neighbor instructions

32 ET-Minion CPUs and 4 MB Memory Form a “Minion Shire”

32 ET-Minion RISC-V cores per Minion Shire

- Arranged in four 8-core neighborhoods

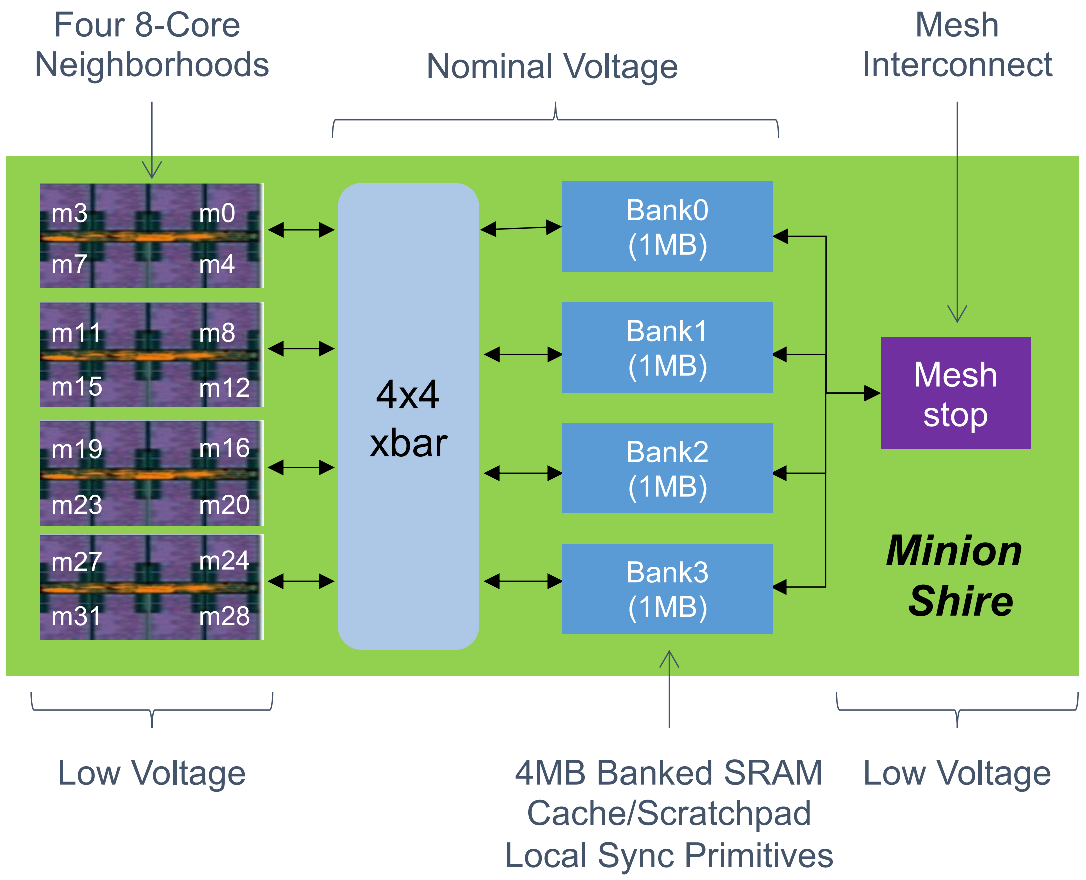
Software configurable memory hierarchy

- L1 data cache can also be configured as scratchpad
- Four 1MB SRAM banks can be partitioned as private L2, shared L3 and scratchpad

Shires connected with mesh network

New synchronization primitives

- Fast local atomics
- Fast local barriers
- Fast local credit counter
- Inter processor interrupt support



ET-SoC-1: Full Chip Internal Block Diagram

34 Minion Shires

- 1088 ET-Minions

8 Memory Shires

- LPDDR4x DRAM controllers

1 Maxion / IO Shire

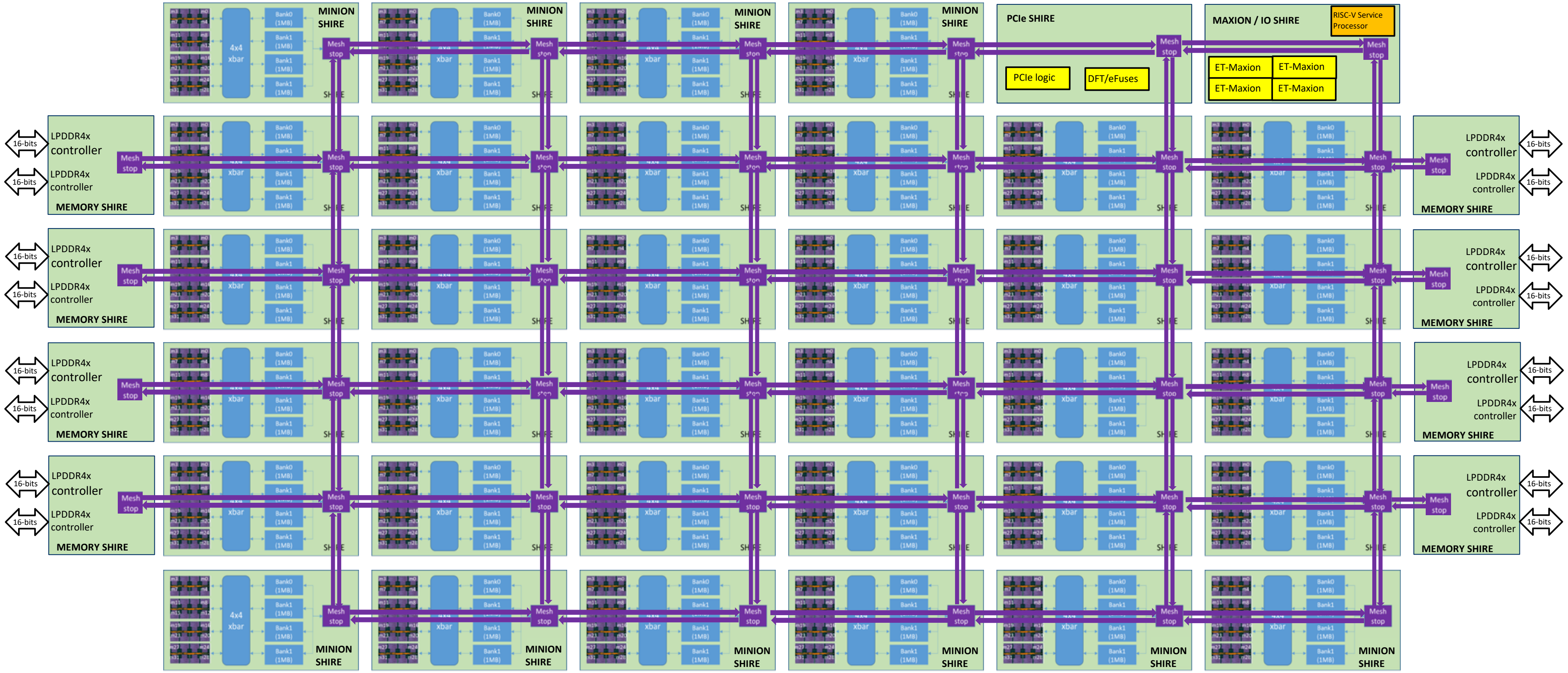
- 4 ET-Maxions
- 1 RISC-V Service Processor

PCIe Shire

160 MB of on-die SRAM

x8 PCIe Gen4

Secure Root of Trust



Esperanto's ML-Optimized Vector/Tensor Extensions

Developed before the RISC-V Vector specification was ready, and optimized for machine learning

New vector, tensor, and multicore coordination operations that deliver high performance with minimal die area and maximum efficiency

These work in cooperation with our ML-oriented multi-level register/cache/scratchpad hierarchy

Supported data types include Int8, Int32, FP16, and FP32

- Int8, FP16 and FP32 supported for tensor ops; Int32 and FP32 for vector ops
- BF16 is supported as a storage format; computations on BF16 data are performed in FP32

All instruction encodings follow RISC-V conventions

Vector ops

Convert
Compute
Compare
Gather/Scatter
Load/Store
Mask
Transcendental

Tensor ops

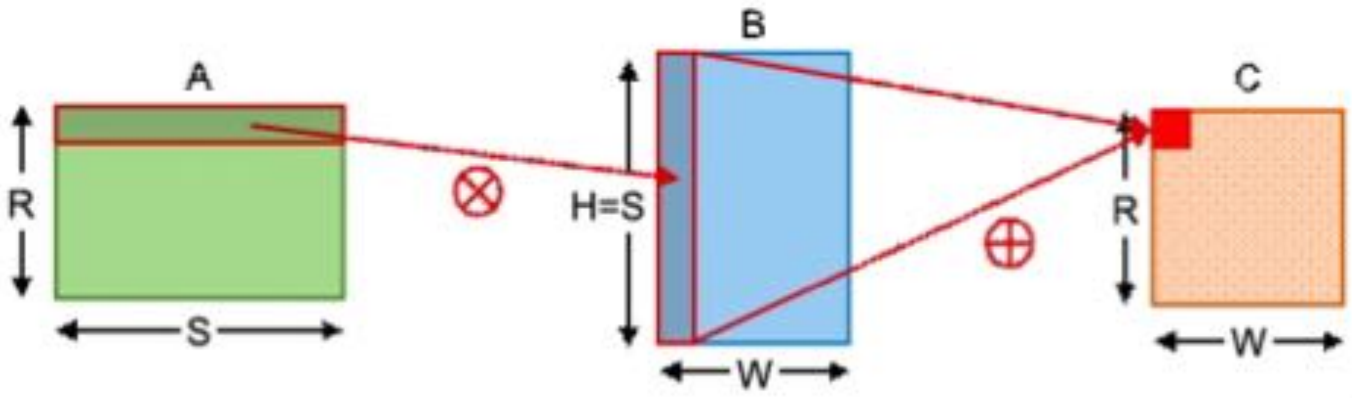
Broadcast
Compute
Load/Store
(with optional interleave/transpose)
Reduce
Transform

Cache Control ops

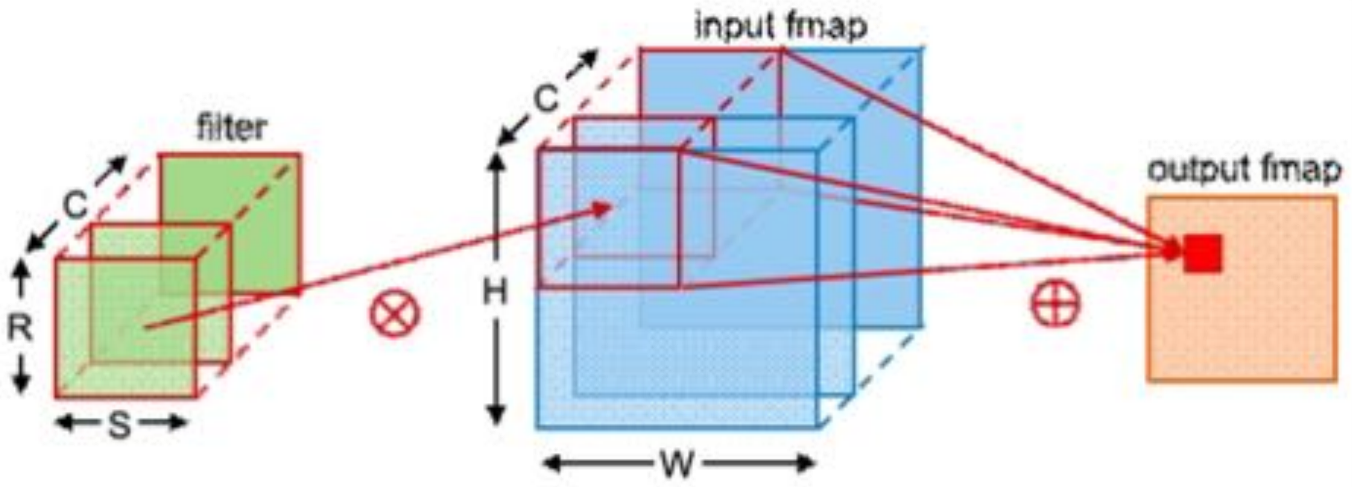
Atomics
Barrier and Credit
Cache evict and flush
Cache lock
Cache partition
Cache prefetch
Scratchpad reservation

Tensor Multiply Operations

2-Dimensional
Matrix Multiplication
of Matrices A and B

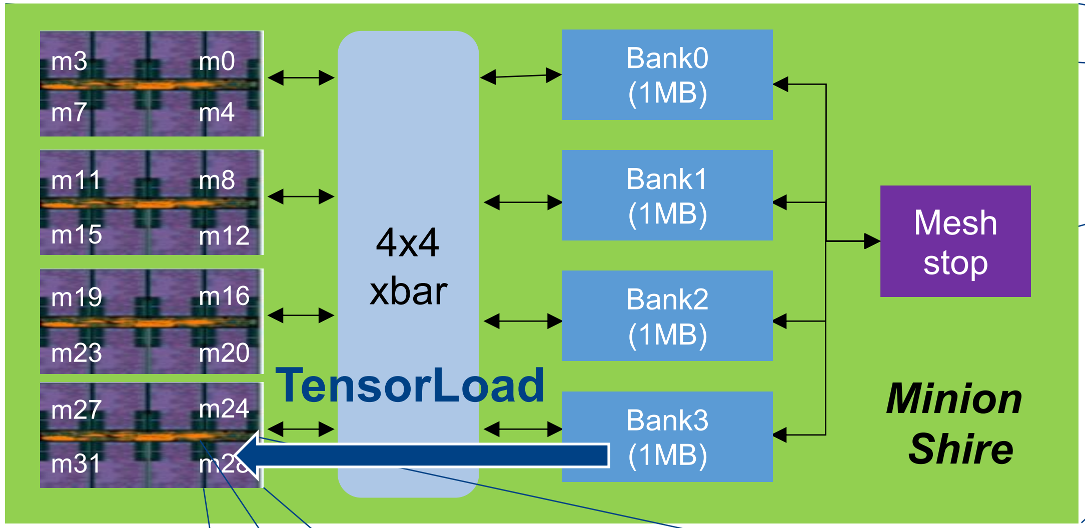


3-Dimensional
Tensor Multiplication
of Tensors A and B

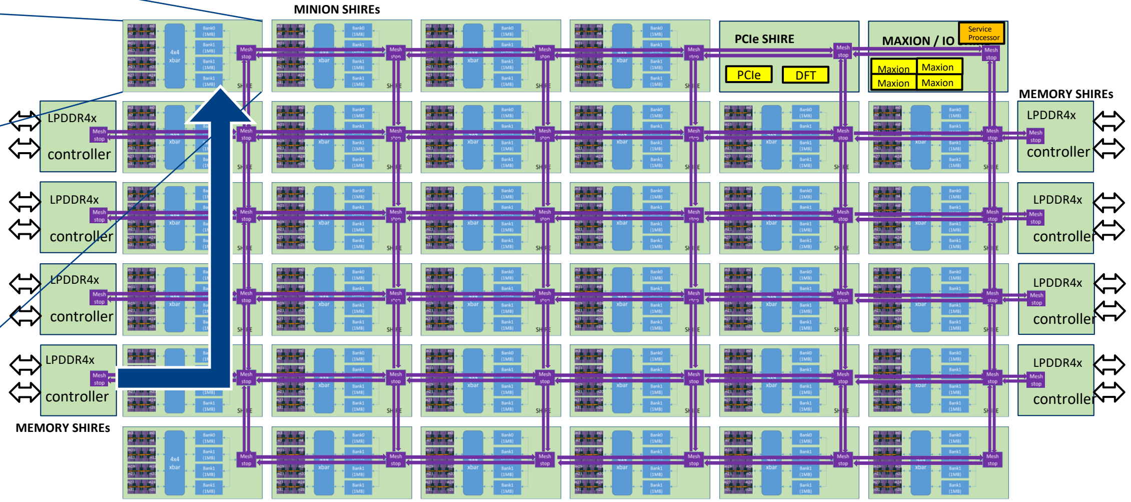


Custom Instructions for Data Movement

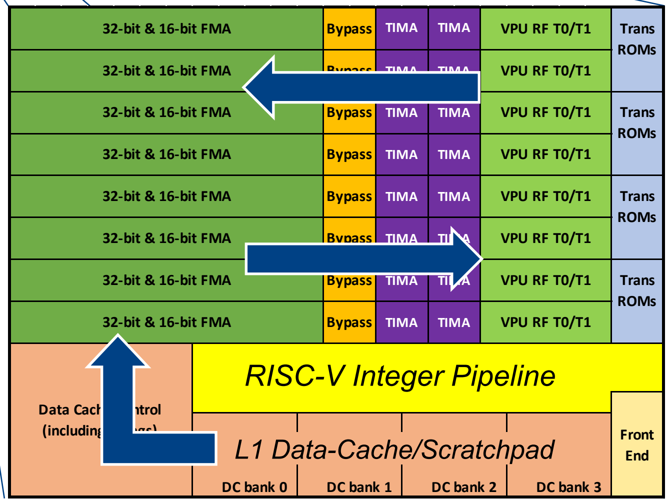
esperanto.ai



TensorLoadL2Scp



TensorFMA



- TensorLoadL2Scp transfers input data from memory (or distributed L3 if resident) to Shire-local scratchpad memory
- TensorLoad moves blocks of input data for each TensorFMA operation to the Minion scratchpad
- TensorFMA results end up in the VPU register file

TensorFMA16A32 Operation Executes at 100% Utilization over 512 Cycles

TensorFMA executes a matrix multiply A x B [+]= C

A and B matrices positioned in Minion Scratchpad or TensorB register file by TensorLoad

- Up to 16 rows of up to 64 bytes of A matrix
- Up to 16 rows of up to 64 bytes of B matrix

C matrix is in the VPU RF

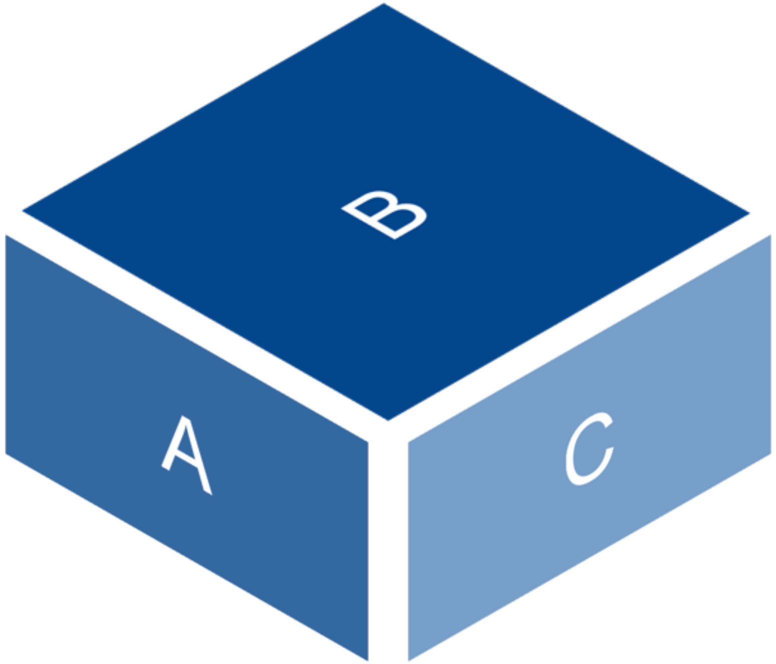
The maximum tile size for FP16 mul / FP32 add is 16 x 32 x 16

The operation takes 512 cycles to execute

- 512 FMA x 8 lanes x 4 FLOPs/lane = 16,384 FLOPs

We want to work with the maximum tile size to maximize performance and power efficiency

Future derivatives can scale as required for future workloads



Separate Int8 TIMA Unit Quadruples Throughput versus FP16

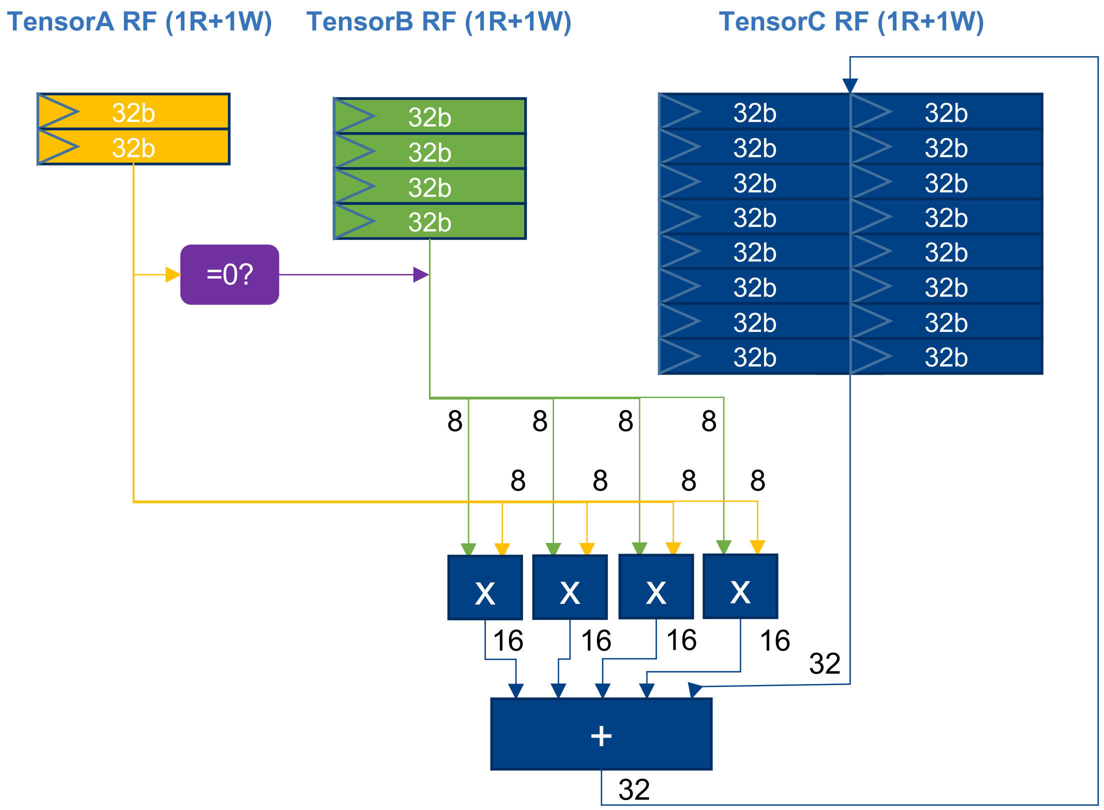
esperanto.ai

TIMA = Tensor Integer Multiply-Add

Private register files separate from main VPU register file minimize dynamic switching during tensor ops

No toggling of main bypass muxes for further reductions in dynamic power

For zero values in Tensor-A, the Tensor-B register file is gated and conserves previous value



Coordination Ops: Atomic Barriers and Credits Keep All ET-Minions Coordinated and Efficient

- Efficiently coordinating the flow of execution across more than a thousand RISC-V cores is critical to getting good results from any many-core chip, including the ET-SoC-1
- Esperanto added custom instructions and associated special registers to the Minions and Shires to support fast, efficient coordination
  - **Barriers**: Each ET-Minion Shire contains a set of Barrier Counter special registers; atomic operations are used to track Minion tasks and synchronize forward progress
  - **Credit counters**: Per thread local credit counters can be updated by remote agents and checked by a custom instruction to authorize forward progress, such as accesses to shared resources
- One ET-Minion Shire can be assigned to coordinate the operation of 32 other Shires using stores directed to special registers at the thread and Shire level

Esperanto Software Development Environment

esperanto.ai

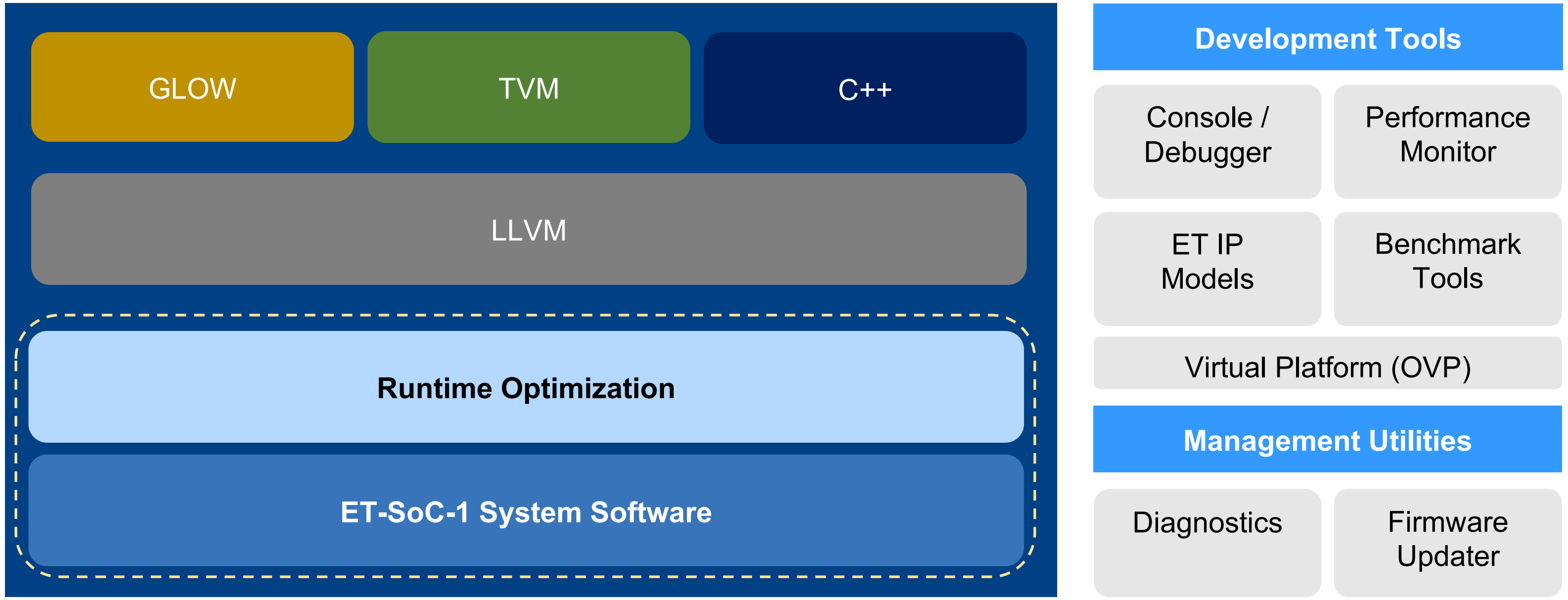
RISC-V

Software / Tools Ecosystem

RTOS
Linux
Simulators, Tools & Runtimes
Security
debian
fedora
SUSE
RT-Thread
AdaCore
Mentor
SiFive
IAR
cādence
Andes
DOVER
Keystone
INTRINSIC ID
HEX-Five Security
RISC-V

esperanto

Software / Tools Development



GLOW | TVM | C++

LLVM

Runtime Optimization

ET-SoC-1 System Software

Development Tools

- Console / Debugger
- Performance Monitor
- ET IP Models
- Benchmark Tools
- Virtual Platform (OVP)

Management Utilities

- Diagnostics
- Firmware Updater

Esperanto is heavily investing in its SW and Tool solutions, extending the RISC-V open-source ecosystem and offering Esperanto customers greater choice

Summary

esperanto.ai

Esperanto’s low-voltage technology provides differentiated RISC-V processors with the best performance per watt

- Energy efficiency matters!
- Best performance per watt delivers the best performance in a fixed number of watts
- Solution delivers energy efficient acceleration for datacenter inference workloads, especially ML Recommendation

Esperanto Vector/Tensor Extensions Drive High ML Performance with Low Power Consumption

- Low-voltage vector/tensor units support integer and floating-point inferencing while consuming minimal area and power
- These units work with our multi-level register/cache/scratchpad hierarchy to adapt to a wide variety of algorithms

Esperanto ET-SoC-1 has a highly scalable design

- Efficient for ML recommendation
- Thousands of general-purpose RISC-V cores can be applied to many other highly parallel computing tasks
- Modular approach allows design to scale from cloud to edge and to other semiconductor processes

Early Access Program for qualified customers beginning later in 2021 (for info, contact: chips@esperanto.ai)

Thank you!

If you are interested in evaluating ET-SOC-1, please contact us.

World Wide: chips@esperanto.ai
Japanese: takashi.murayama@esperanto.ai