semidynamics

All-in-one RISC-V AI Compute Engine

Roger Espasa, CEO

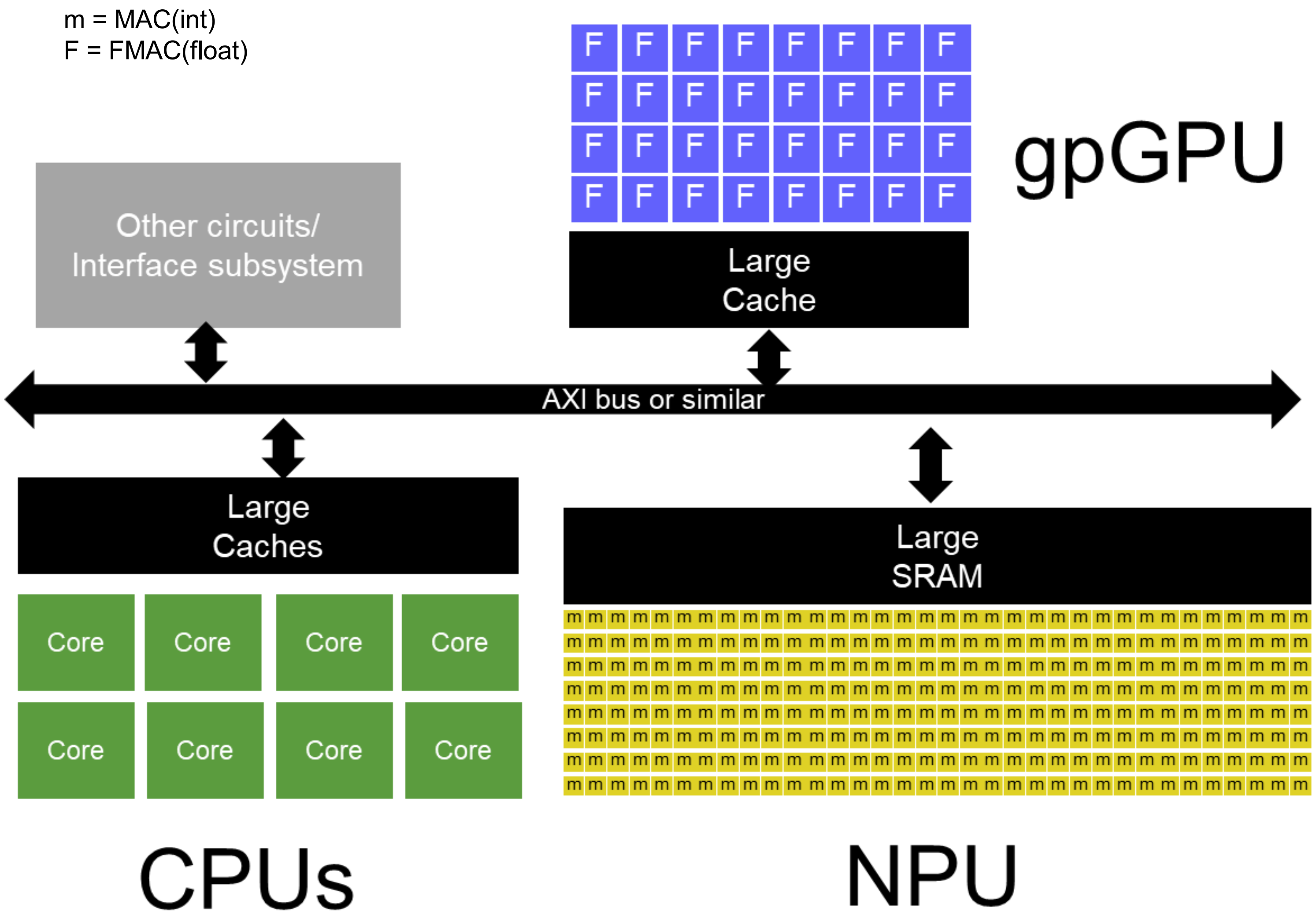
In Order Core

OOO Core

OOO Vector Unit

Tensor Unit

Old-Style AI Architecture



m = MAC(int)
F = FMAC(float)

gpGPU

Large Cache

Other circuits/ Interface subsystem

AXI bus or similar

Large Caches

Core Core Core Core
Core Core Core Core

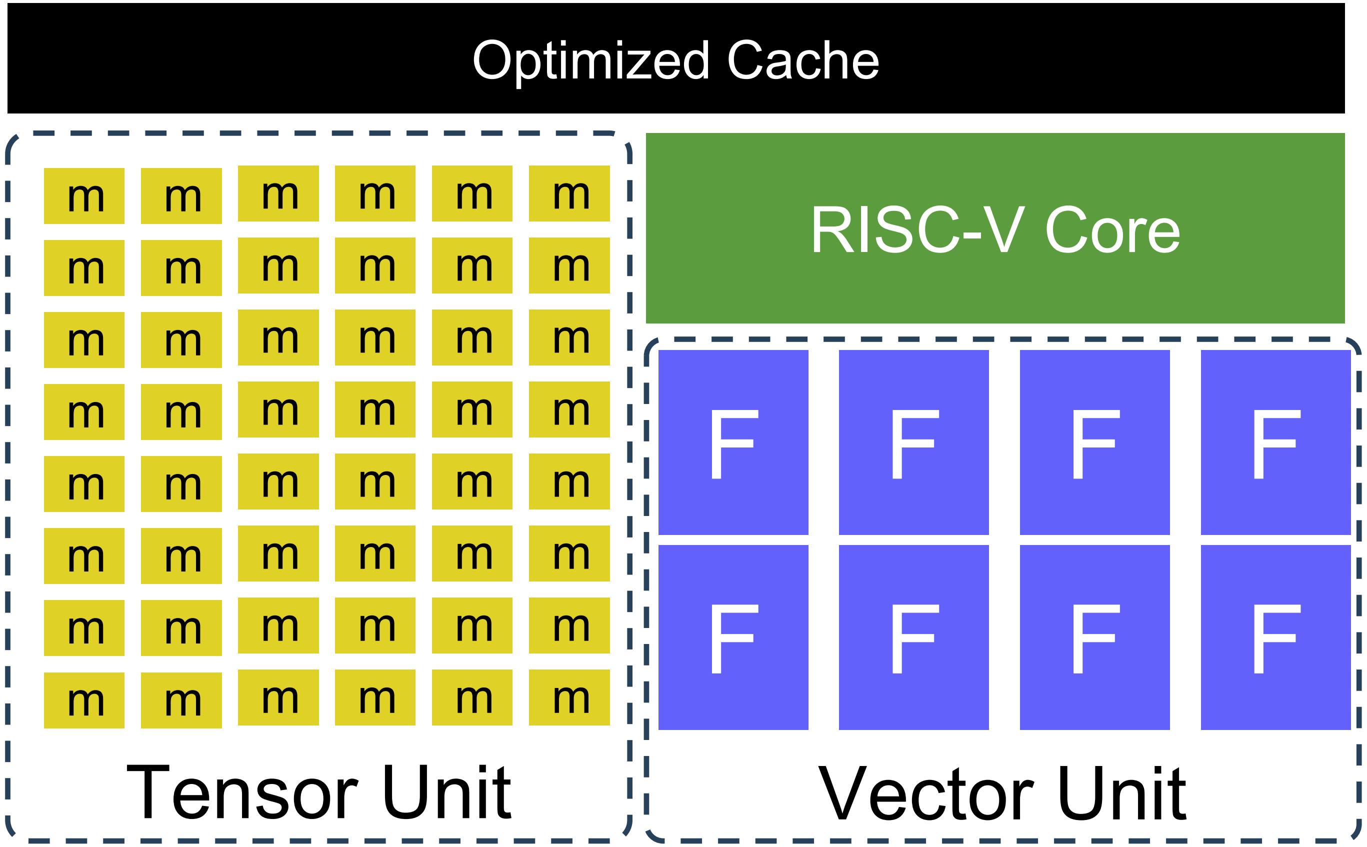
CPUs

Large SRAM

NPU

- **Three** Software Stacks
- **DMA-intensive** programming
- **High** Latency & Power
- **SRAM/Cache/Data** Replication
- **Unbalanced** Scaling
- **Not AI Future Proof**

All-in-one: merging Core, NPU, GPU



Optimized Cache

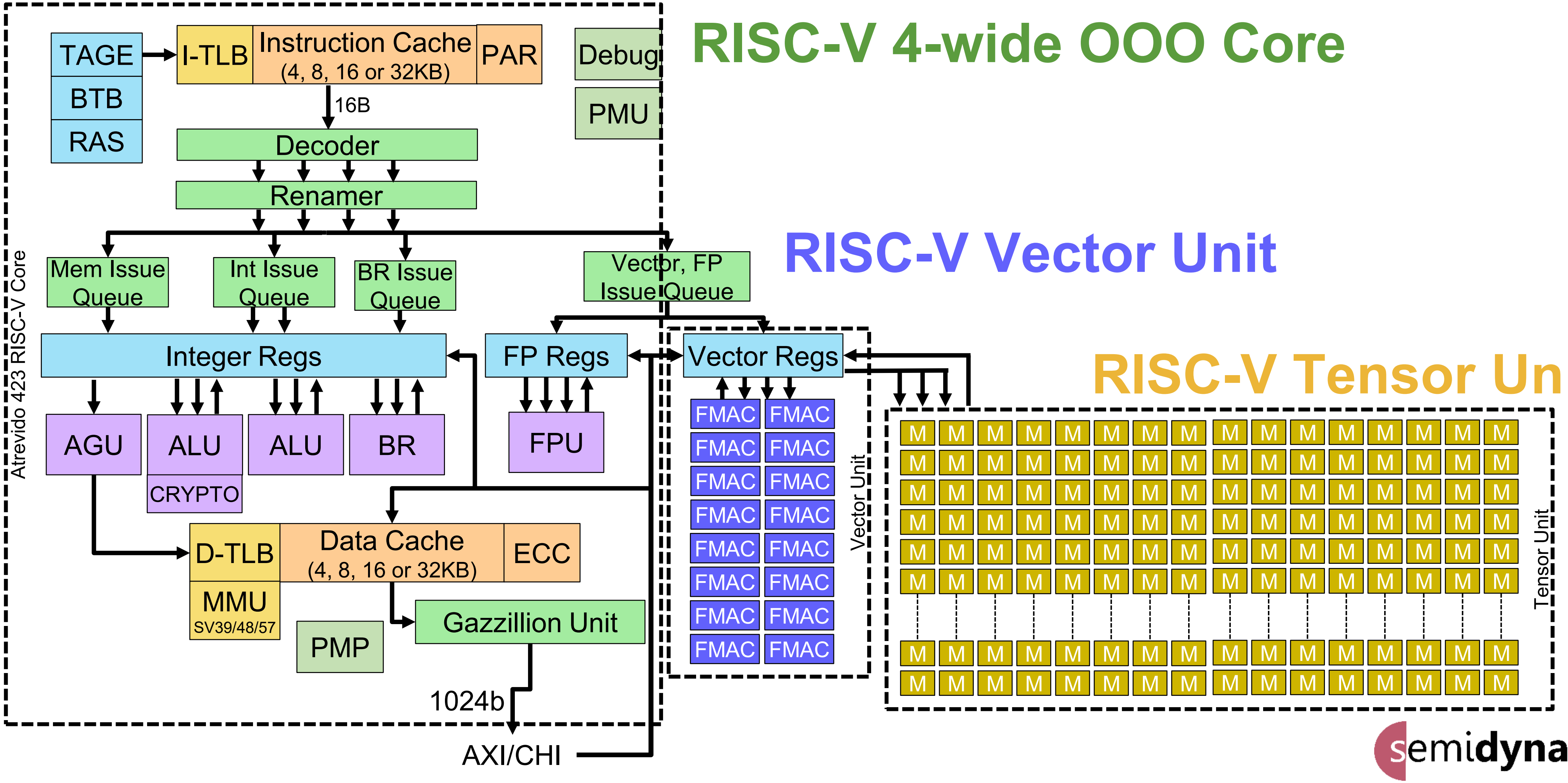
RISC-V Core

Tensor Unit

Vector Unit

- **Single** software stack
- **DMA-free** programming
- **Zero** Latency & **Low** Power
- **Optimized/Shared** Cache
- **Balanced** Scaling
- **AI Future-Proof**

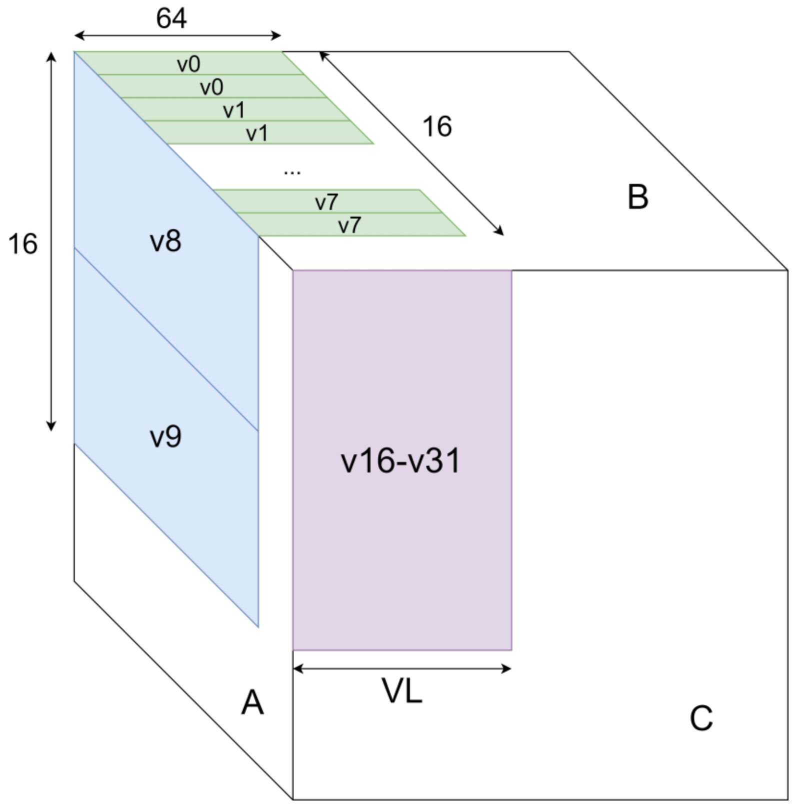
All-In-One Block Diagram



semidynamics

Single Software stack: Matmul in 8 instructions

```
# C tile pre-loaded into v16-v31
loop: vsetvli  zero, t4, e16, m2, ta, ma
      vlrs16   v8, (a0), t1
      addi     a0, a0, 32
      vsetvli  zero, t5, e16, m8, ta, ma
      vlrs16   v0, (a1), t2
      add      a1, a1, t3
      vfmxmacc v16, v8, v0
      bltu     a1, t6, loop
# Store C tile (v16-v31) back to memory
```



semidynamics

Our Customers AI Concerns

- What Software stack do I get with your IP?
- Can I run today’s AI Models with your IP?
  - Transformers, specifically?
- Can I easily scale your solution?
- Can I run future AI Models with your IP?
  - I am buying IP today
  - I will be entering the market in 3+ years
  - How do I know the IP will handle the “3-years-from-now” models?

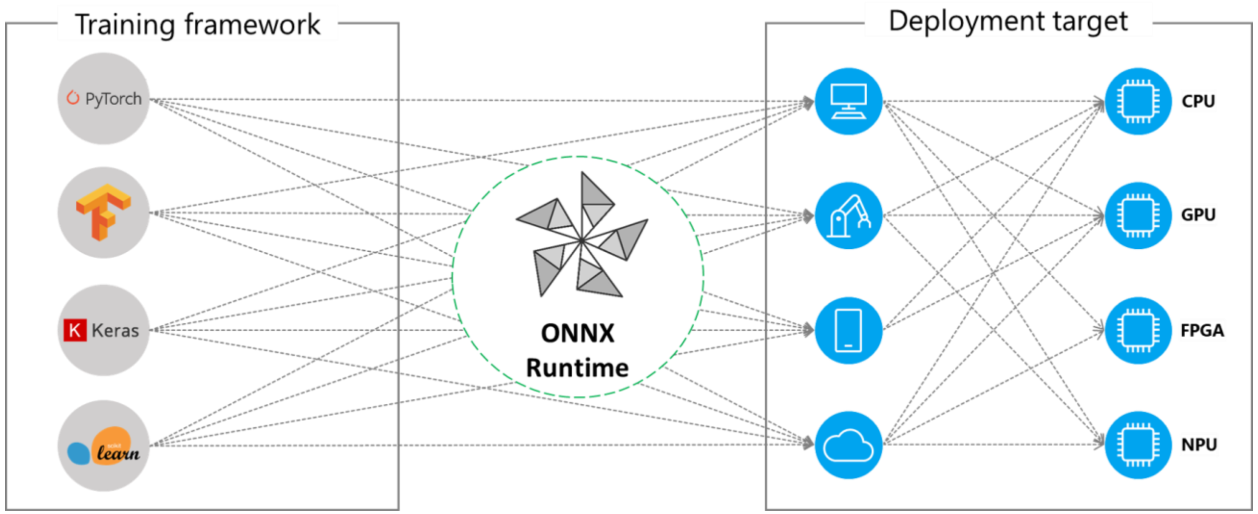
semidynamics

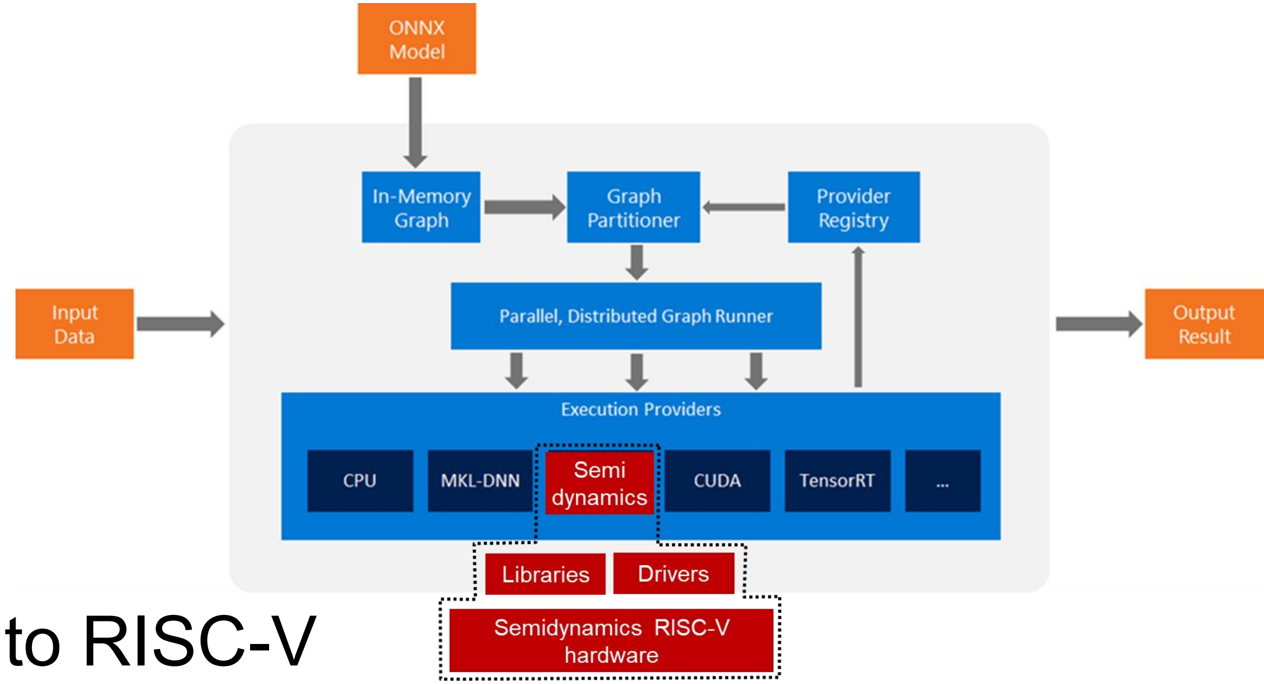
Concern #1: What Software stack do I get with the IP?

Semidynamics AI SW Stack

ONNX RT Port to RISC-V + Vector + Tensor

Semidynamics ONNX RT port





- Semidynamics has ported ONNX RT to RISC-V
  - “Execution Provider” added to ONNX RT
- Semidynamics has optimized the key ONNX operators…
  - …to use its Tensor unit (for Matrix Multiply & Convolution)
  - …to use its Vector unit (for Activations like Sigmoid, …)

Running Transformers / LLMs on All-In-One solution

Llama-2, FP16, 7B Parameter

We’ll use our 1 TOPS$_8$ T1 Tensor Unit…

Product	T1	T2	T4	T8
MACs	512	1024	2048	4096
Local SRAM?	No	No	64KB	128KB
INT8 TOPS/GHz	1	2	4	8
INT16 TOPS/GHz	0.5	1	2	4
BF16 TOPS/GHz	0.5	1	2	4
FP16 TOPS/GHz	0.5	1	2	4

We’ll use our 128 GOPS$_8$ V128 Vector Unit…

Product	V128	V256	V512
FMACs	**8**	**16**	**32**
INT8 GOPS/GHz	128	256	512
INT16 GOPS/GHz	64	128	256
BF16 GOPS/GHz	64	128	256
FP16 GOPS/GHz	64	128	256
FP32 GOPS/GHz	32	64	128
FP64 GOPS/GHz	16	32	64

Llama-2

FP16,
7B params

Operators	Scalar	T1	T1+V128
Matmul			
Activations			
Concat			
Sigmoid			
ScatterND			
Div			
Mul			
Slice			
Exp			
Other			
Speedup	1X		

Llama-2

FP16,
7B params

Operators	Scalar	T1	T1+V128
Matmul	99%		
Activations	1%		
Concat	0.11%		
Sigmoid	0.09%		
ScatterND	0.09%		
Div	0.06%		
Mul	0.03%		
Slice	0.03%		
Exp	0.03%		
Other	0.54%		
Speedup	1X		

semidynamics

Llama-2

FP16,
7B params

Operators	Scalar	T1	T1+V128
Matmul	99%	20%	
Activations	1%	80%	
Concat	0.11%	19%	
Sigmoid	0.09%	16%	
ScatterND	0.09%	15%	
Div	0.06%	9.5%	
Mul	0.03%	5.7%	
Slice	0.03%	5.0%	
Exp	0.03%	4.4%	
Other	0.54%	5.4%	
Speedup	1X	170X	

semidynamics

Llama-2

FP16,
7B params

Operators	Scalar	T1	T1+V128
Matmul	99%	20%	55%
Activations	1%	80%	45%
Concat	0.11%	19%	17%
Sigmoid	0.09%	16%	2%
ScatterND	0.09%	15%	17%
Div	0.06%	9.5%	2%
Mul	0.03%	5.7%	2.4%
Slice	0.03%	5.0%	1.3%
Exp	0.03%	4.4%	0.5%
Other	0.54%	5.4%	2.8%
Speedup	1X	170X	470X

semidynamics

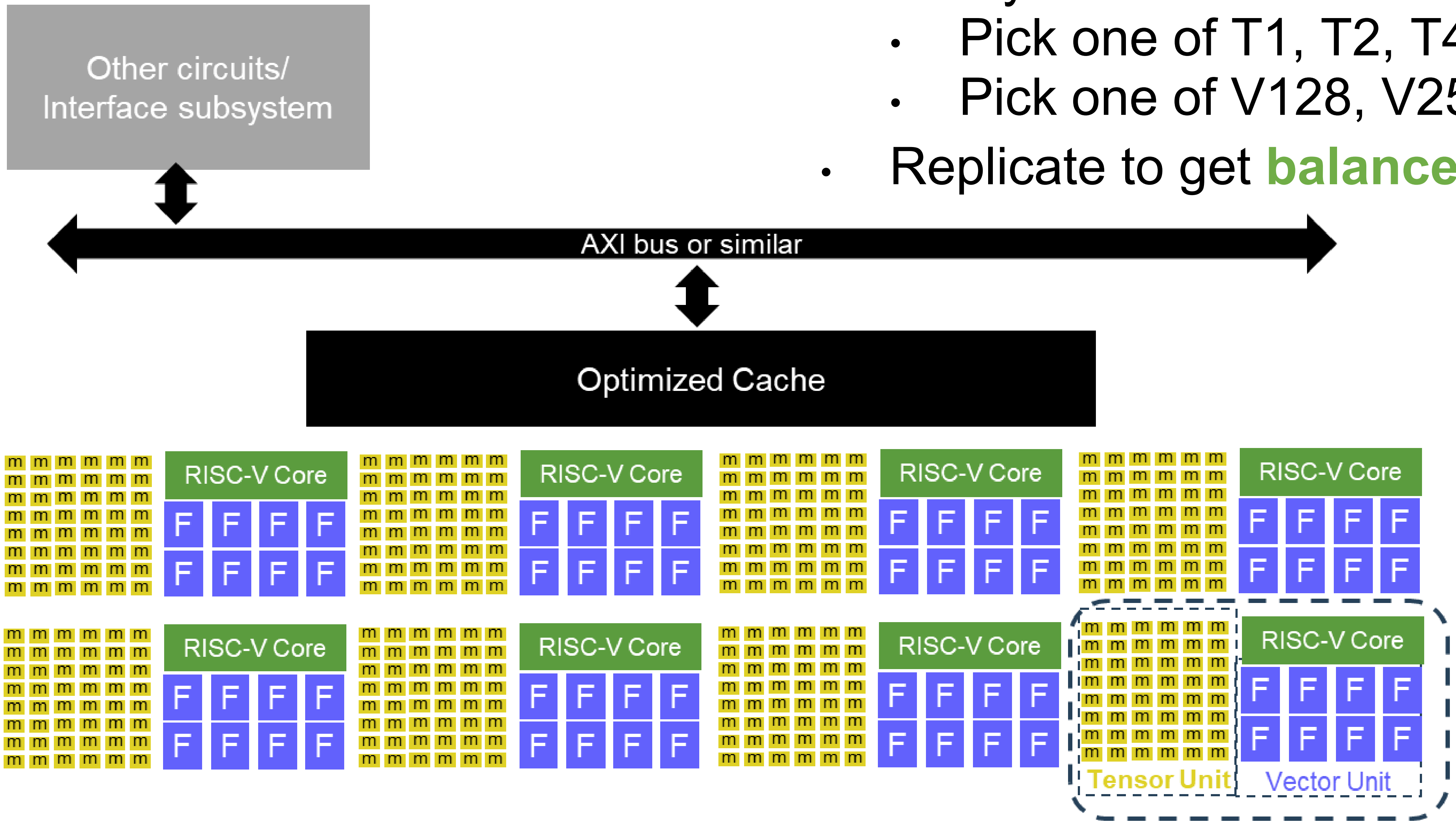
Concern #4: Can I easily **scale** your solution?

Sacling up All-in-one solution

semidynamics

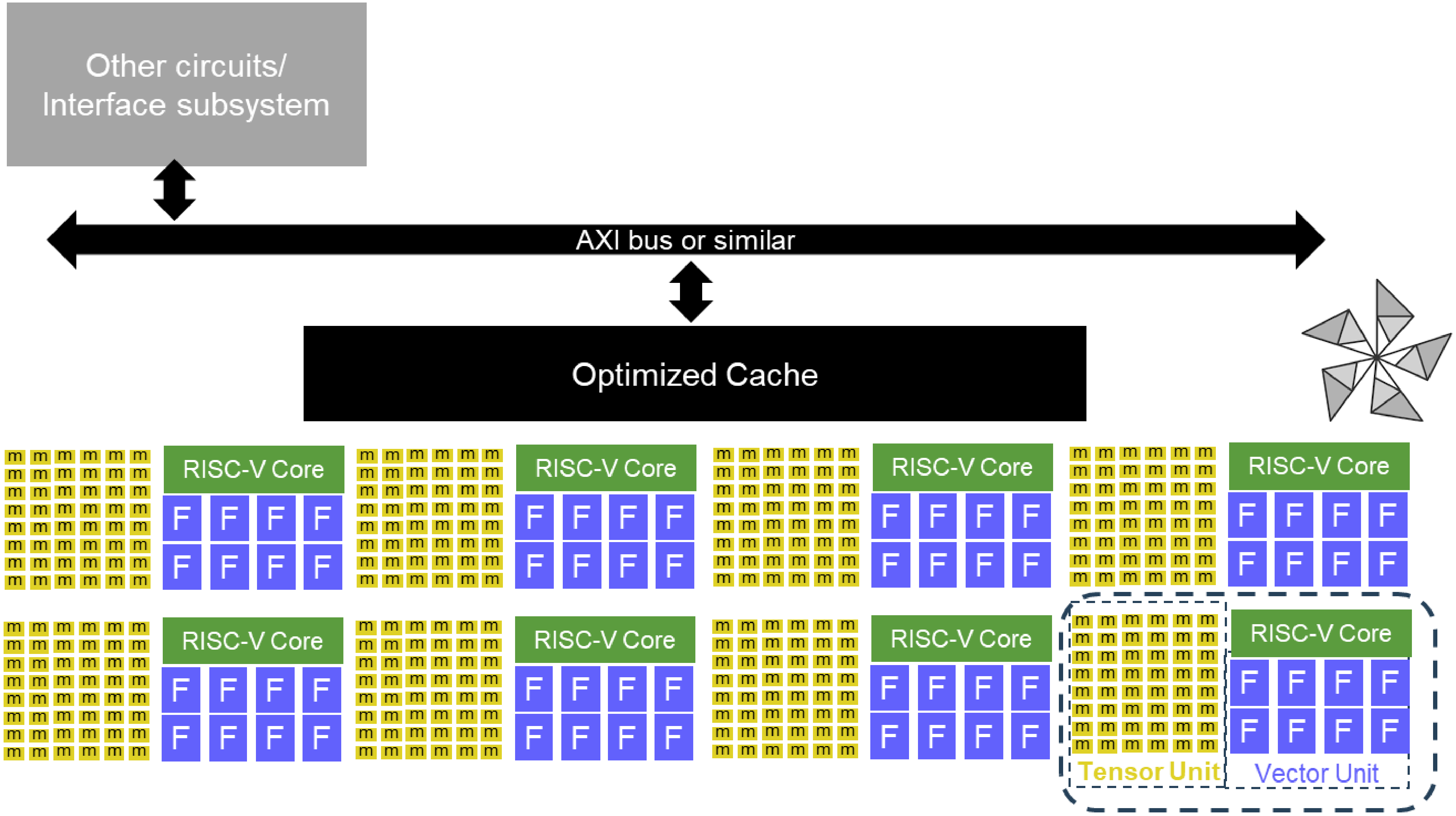
How do you scale up further?

- Pick your All-in-one “building block”
  - Pick one of T1, T2, T4, T8
  - Pick one of V128, V256, V512
- Replicate to get **balanced scaling**



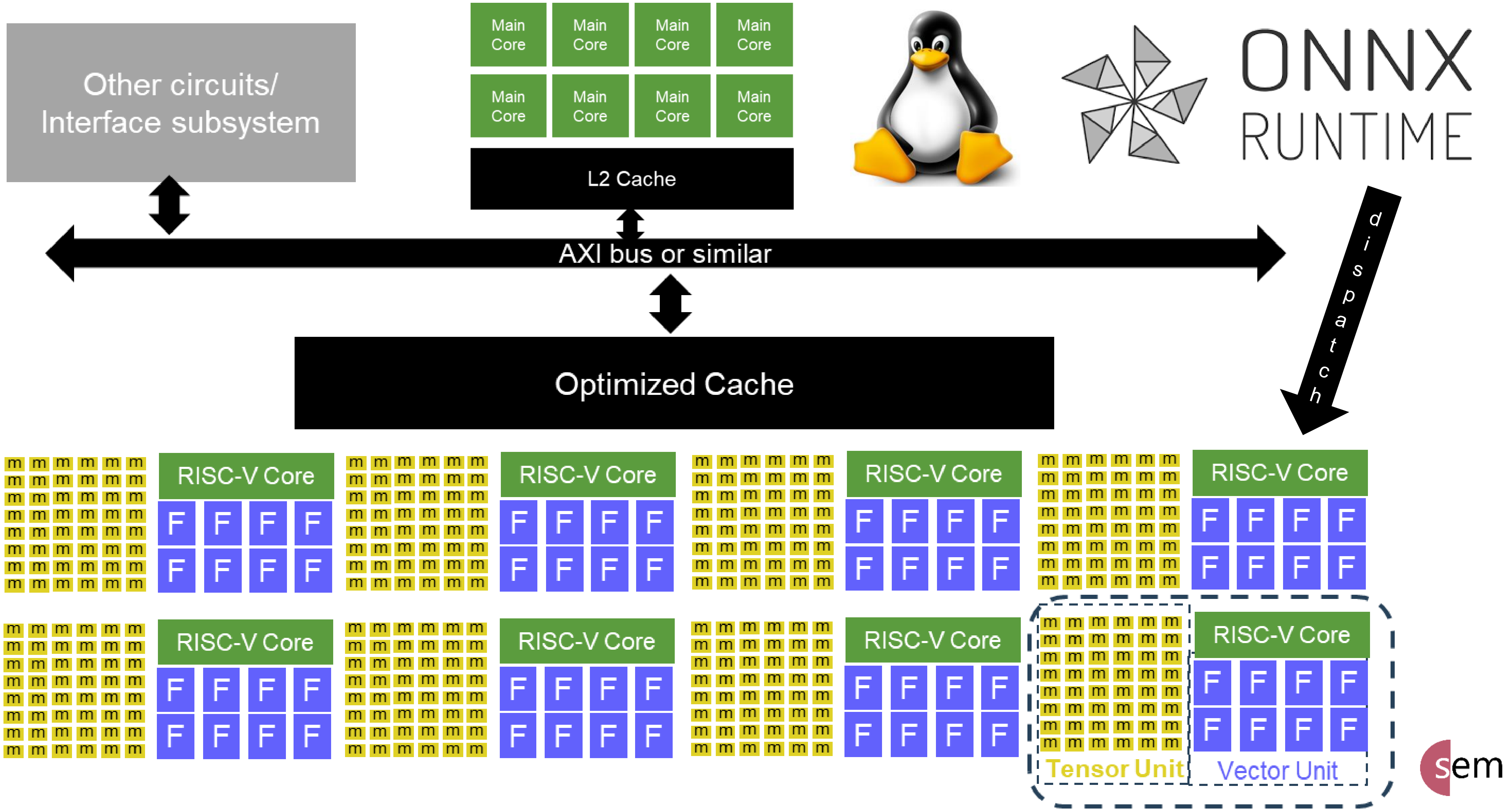
semidynamics

But… where is your ONNX RT SW running?



semidynamics

But… where is your ONNX RT SW running?



Other circuits/ Interface subsystem

Main Core | Main Core | Main Core | Main Core
Main Core | Main Core | Main Core | Main Core

L2 Cache

ONNX RUNTIME

AXI bus or similar

dispatch

Optimized Cache

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

RISC-V Core

F F F F
F F F F

m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m
m m m m m m

Tensor Unit

RISC-V Core

F F F F
F F F F

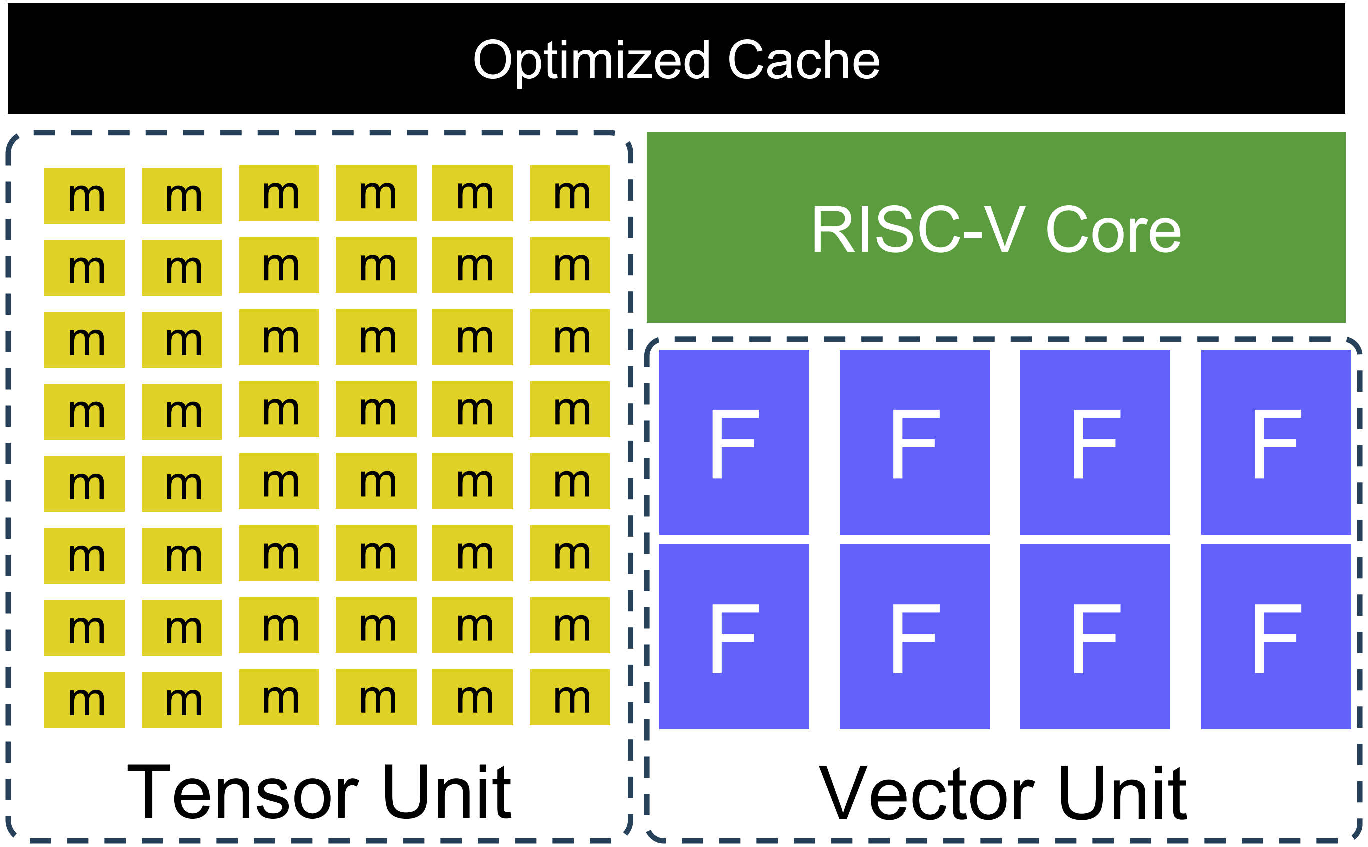
Vector Unit

Concern #3: Can I run **future AI models** with your IP?

All-in-one is future-proof

semidynamics

Running Future Models



- Vector and Tensor controlled by RISC-V **INSTRUCTIONS**
- RISC-V core has full “if-then-else” and “recursion” capability
  - i.e., Turing-complete
- If the model can be expressed in ONNX, we can run it!

semidynamics

Our Customers AI Concerns - Solved

- What Software stack do I get with the IP? ✓
  - ONNX RT optimized for Semidynamics IP
- Can I run today’s AI Models with current IP? ✓
  - YES, with All-in-One-IP
- Can I run future AI Models with current IP? ✓
  - YES, with All-in-One-IP

Let’s build the AI future together

semidynamics

Thank you!

semidynamics