semidynamics

All-in-one RISC-V AI Compute Engine for Software Enabled Everything

Volker Politz, CSO

In Order Core

OOO Core

OOO Vector Unit

Tensor Unit

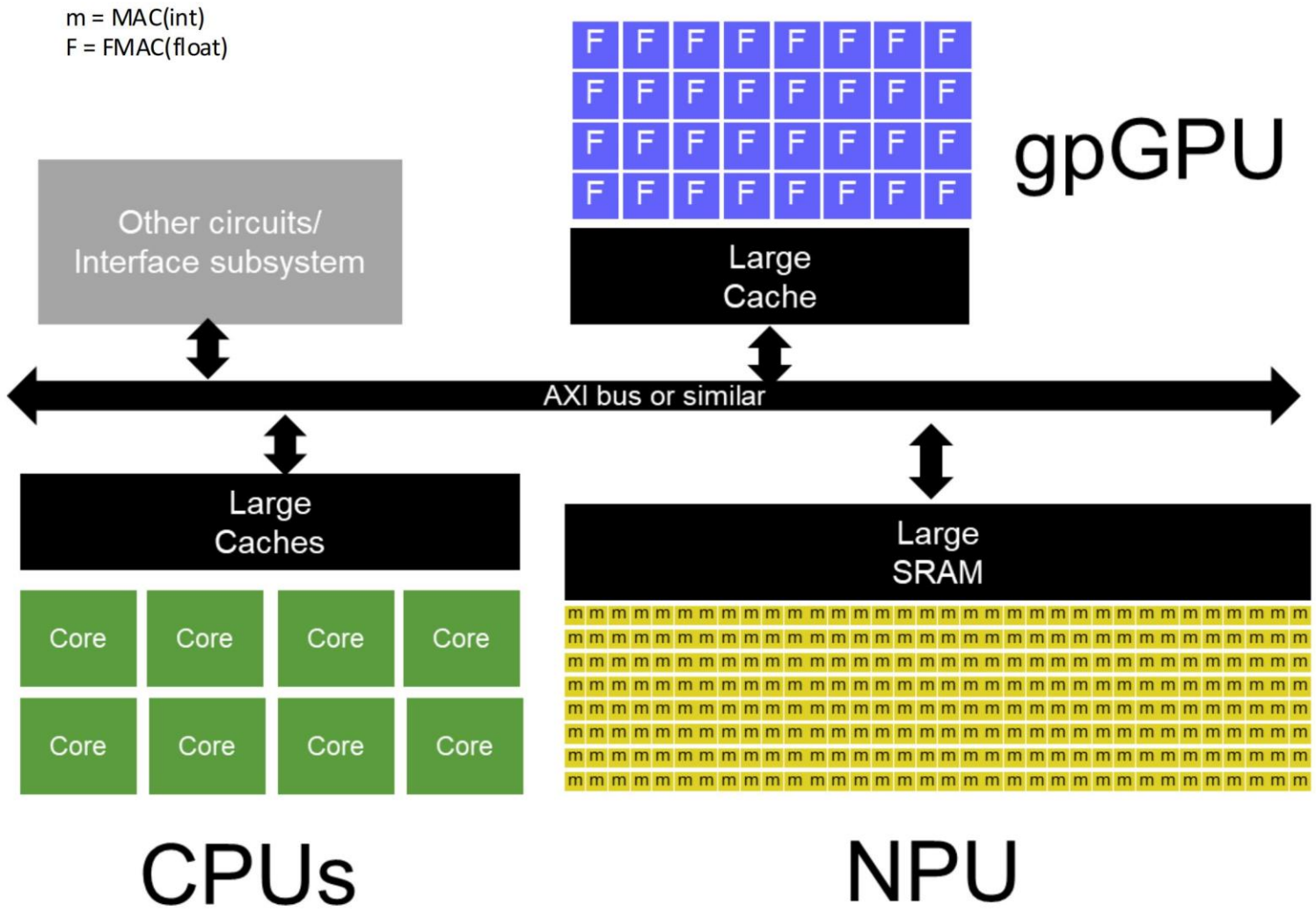
About Semidynamics

Semidynamics, founded in 2016, is a **100% European** supplier of RISC-V IP cores, HQ in **Barcelona**, specializing in **customization** of **high bandwidth high performance AI cores** for **tailored projects**

Experts in customizable AI IP

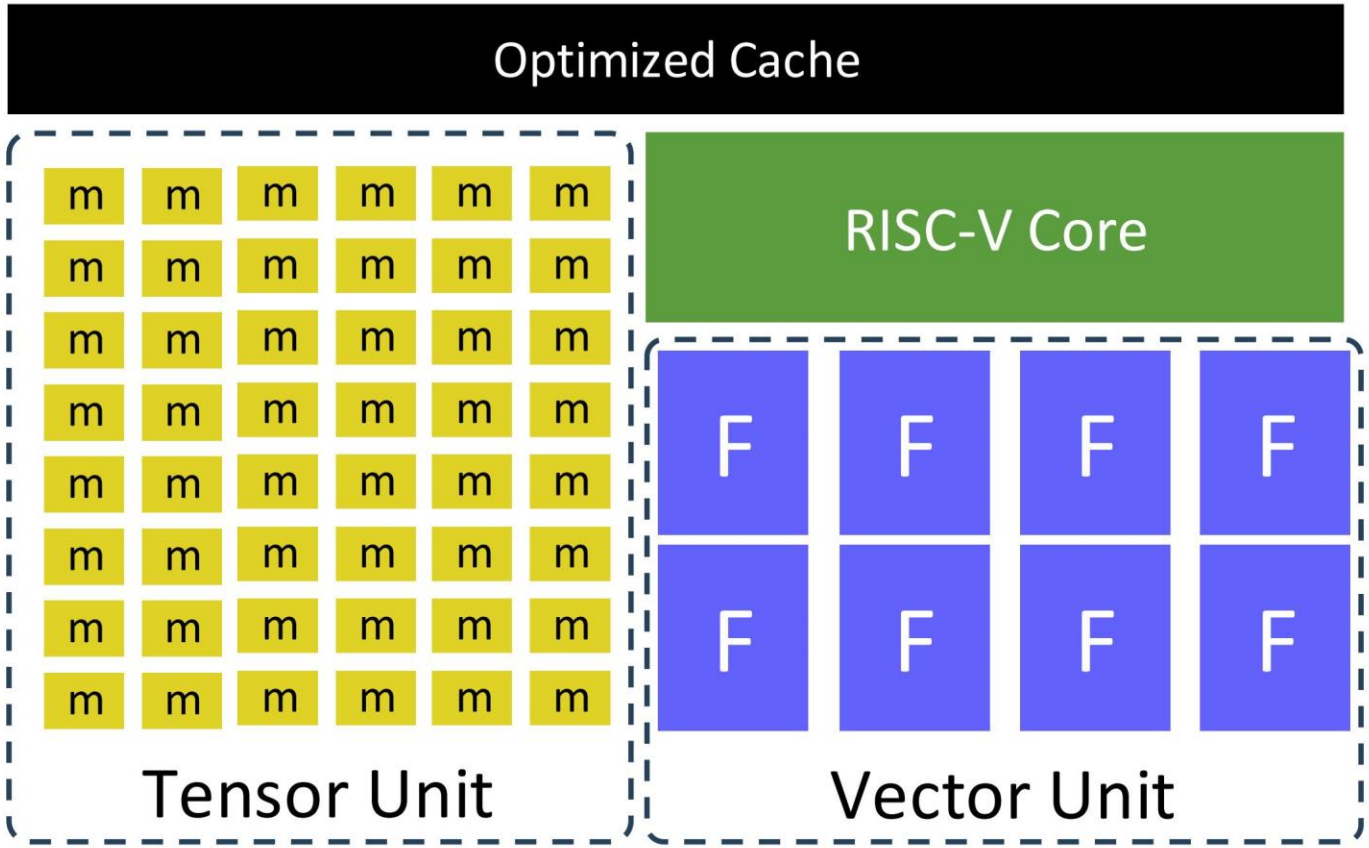
semidynamics

Typical AI SoC Architecture



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

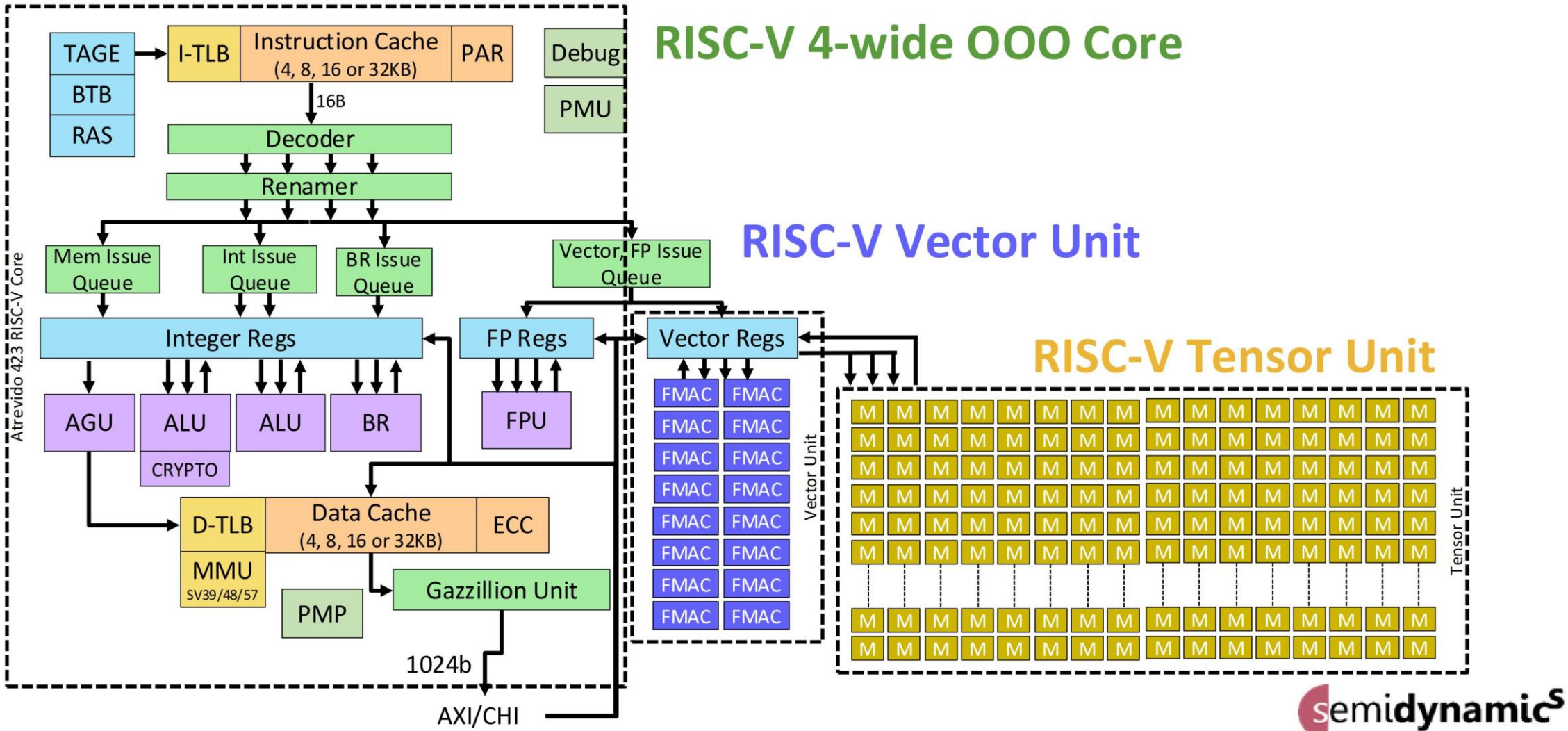
All-in-one: merging Core, NPU, GPU



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

semidynamics

All-In-One Block Diagram



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?

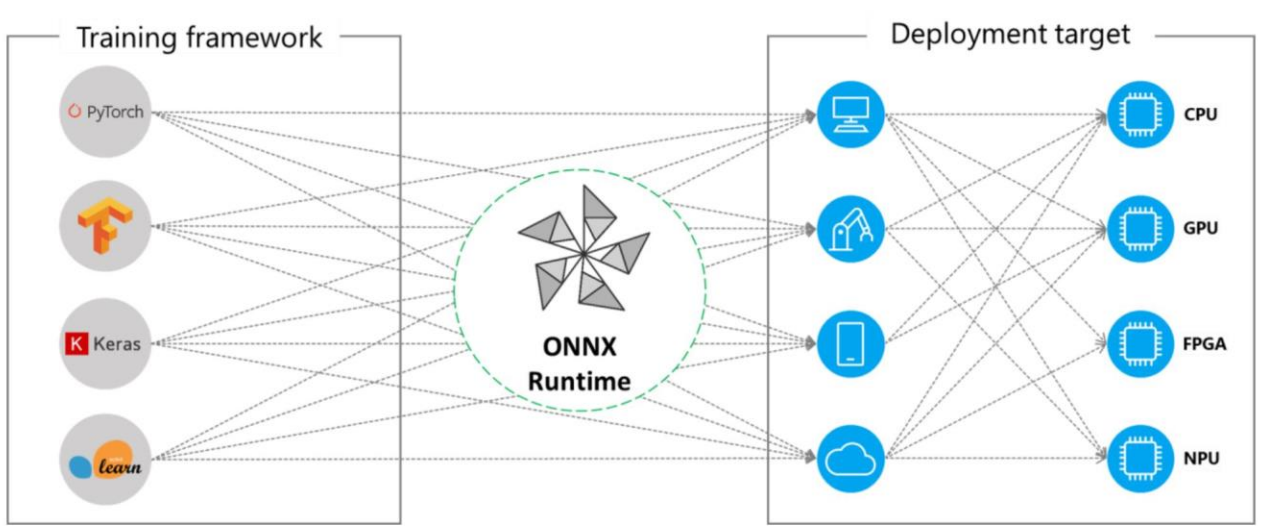
semidynamics

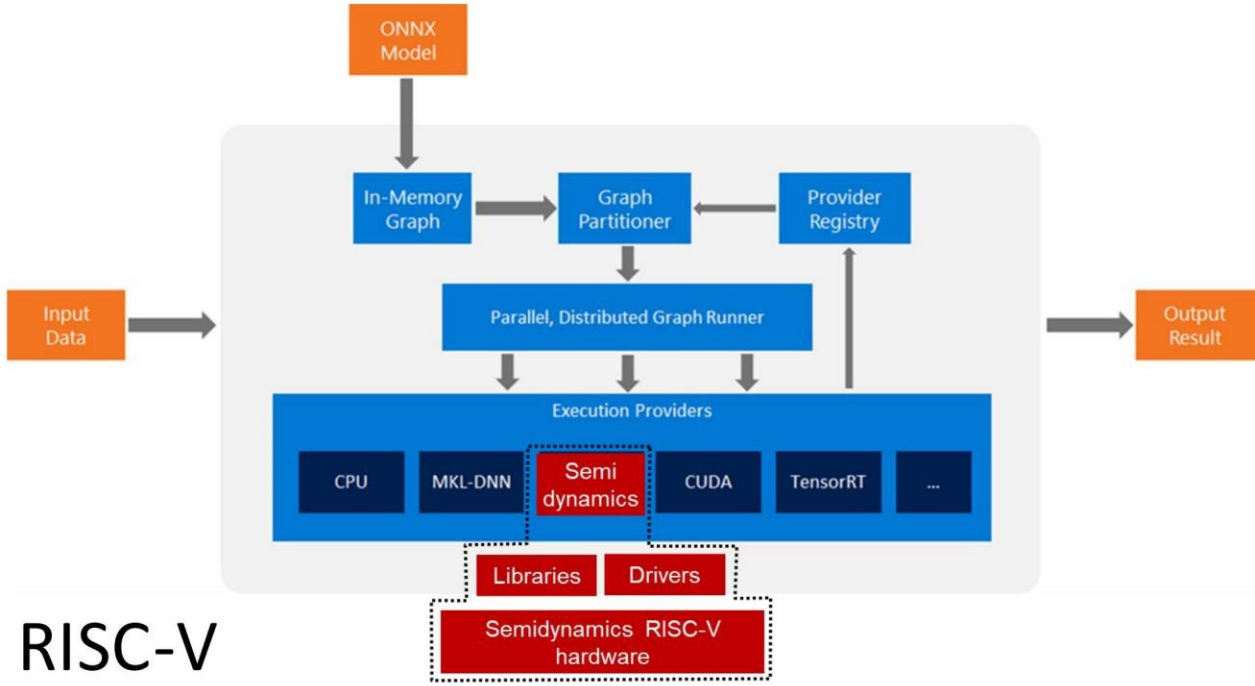
Semidynamics AI SW Stack

ONNX RT Port to RISC-V + Vector + Tensor

semidynamics

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, …)

semidynamics

Running Transformers / LLMs on All-In-One solution

Llama-2, FP16, 7B Parameter

semidynamics

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

Further PPA optimizations: INT only, INT+BF16

semidynamics

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

V64 and V32 also possible

semidynamics

Llama-2

FP16,
7B params

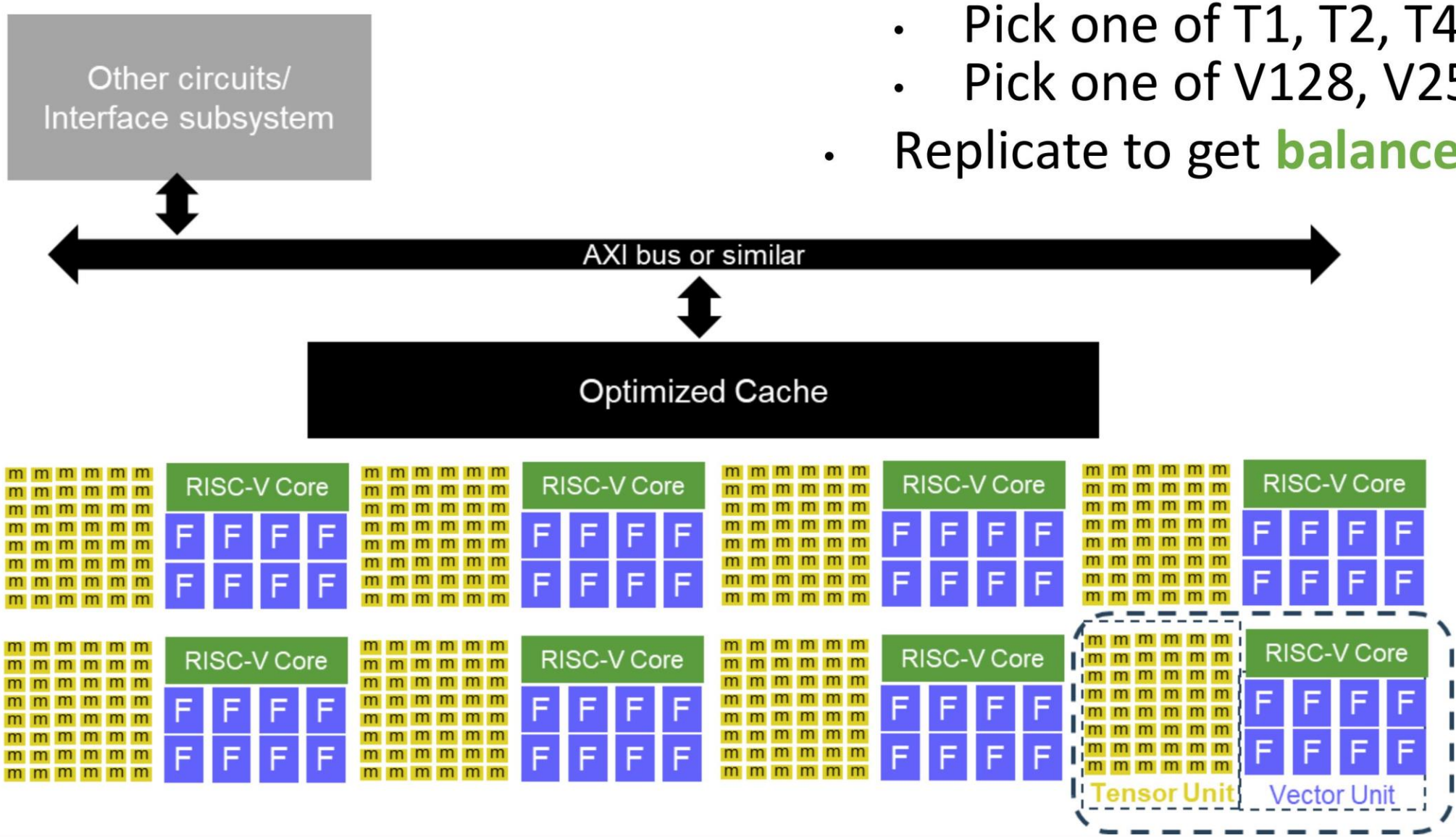
Operators	Scalar	T1	T1+V128
Matmul	99%	20%	55%
Activations	1%	80%	45%
Concat	0.11%	19%	17%
Sigmoid	0.09%	16%	2%
ScatterND	0.09%	15%	17%
Div	0.06%	9.5%	2%
Mul	0.03%	5.7%	2.4%
Slice	0.03%	5.0%	1.3%
Exp	0.03%	4.4%	0.5%
Other	0.54%	5.4%	2.8%
Speedup %	1X	170X	470X

semidynamics

Scaling up All-in-one solution

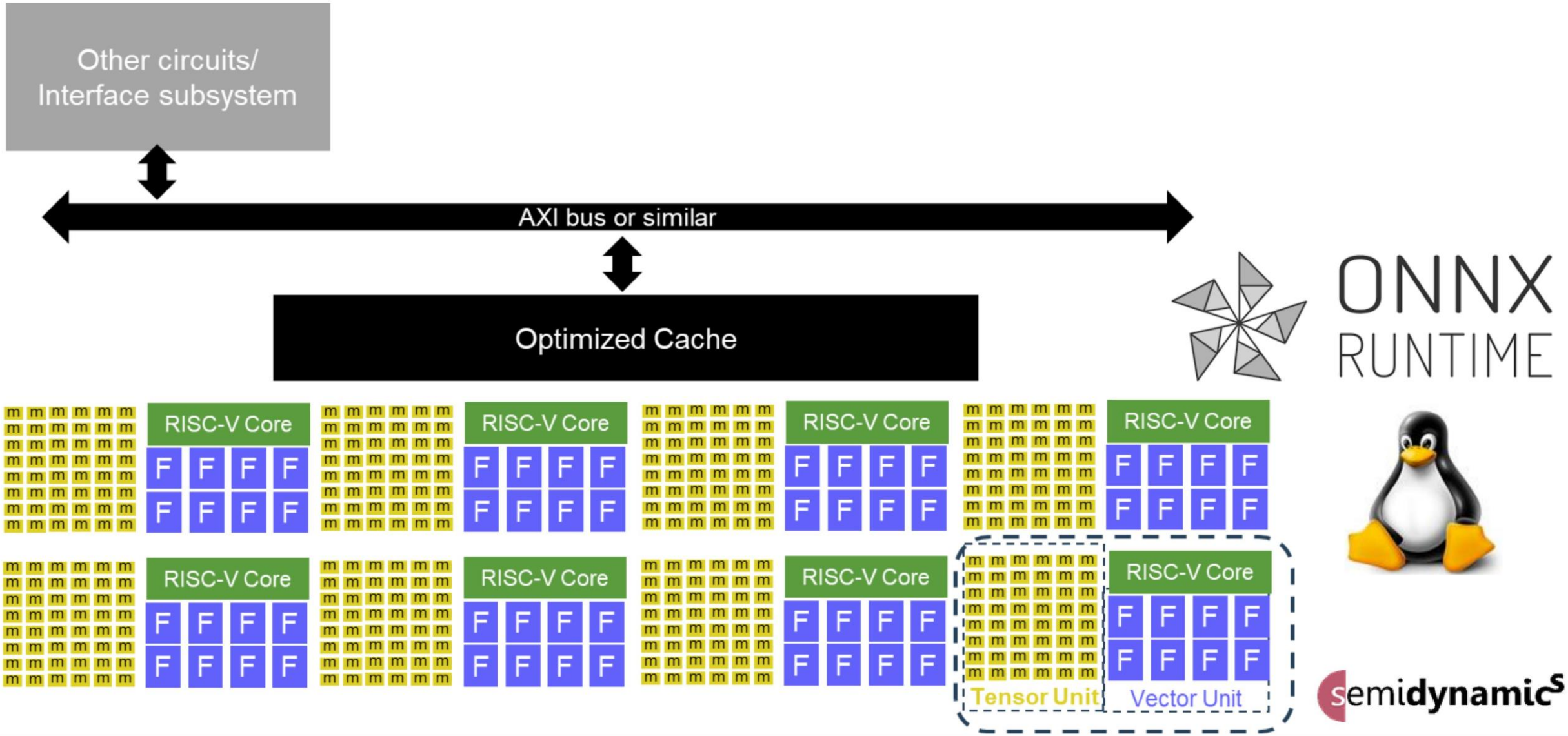
How do you scale up further?

- 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**

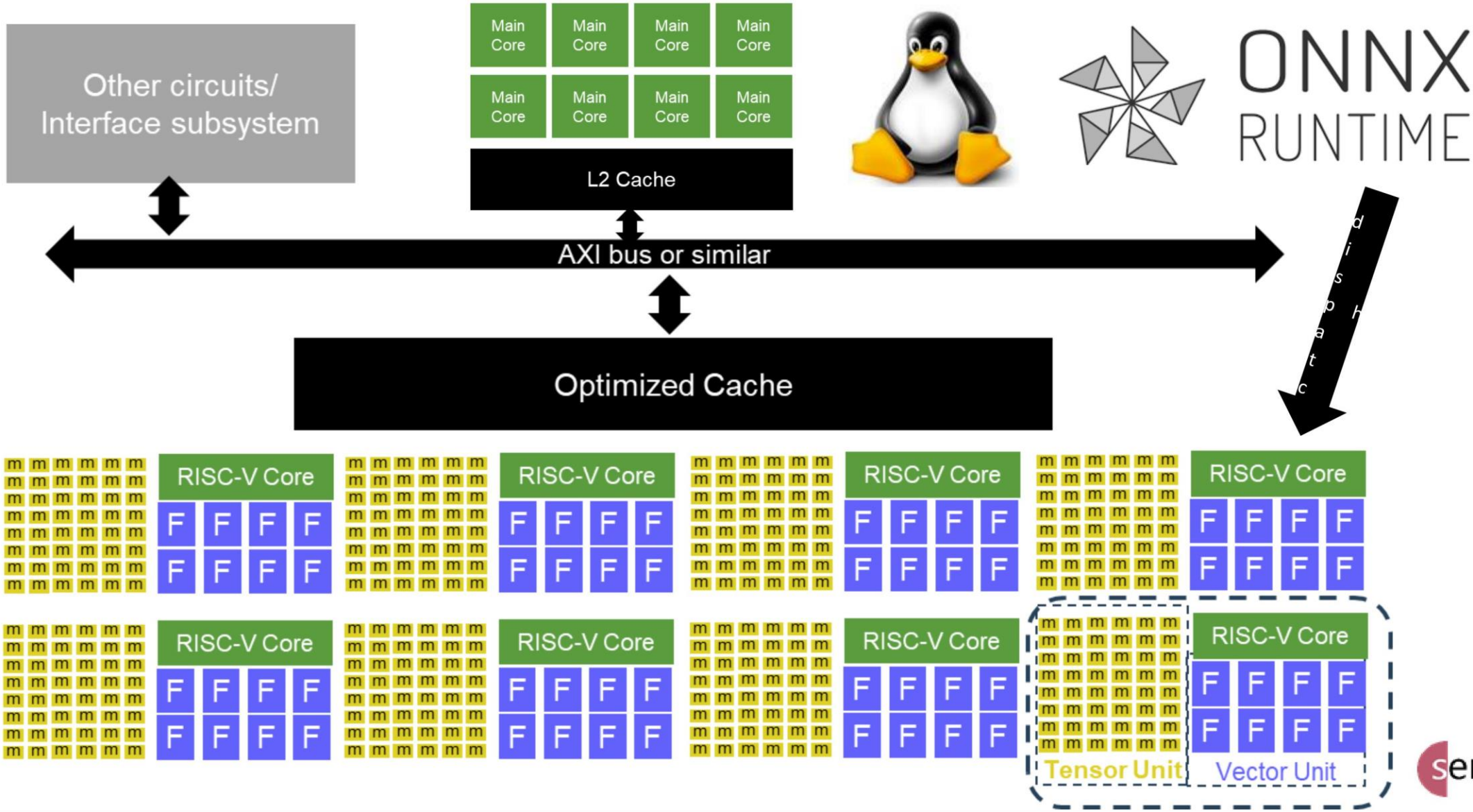


semidynamics

But… where is your ONNX RT SW running?



But… where is your ONNX RT SW running?

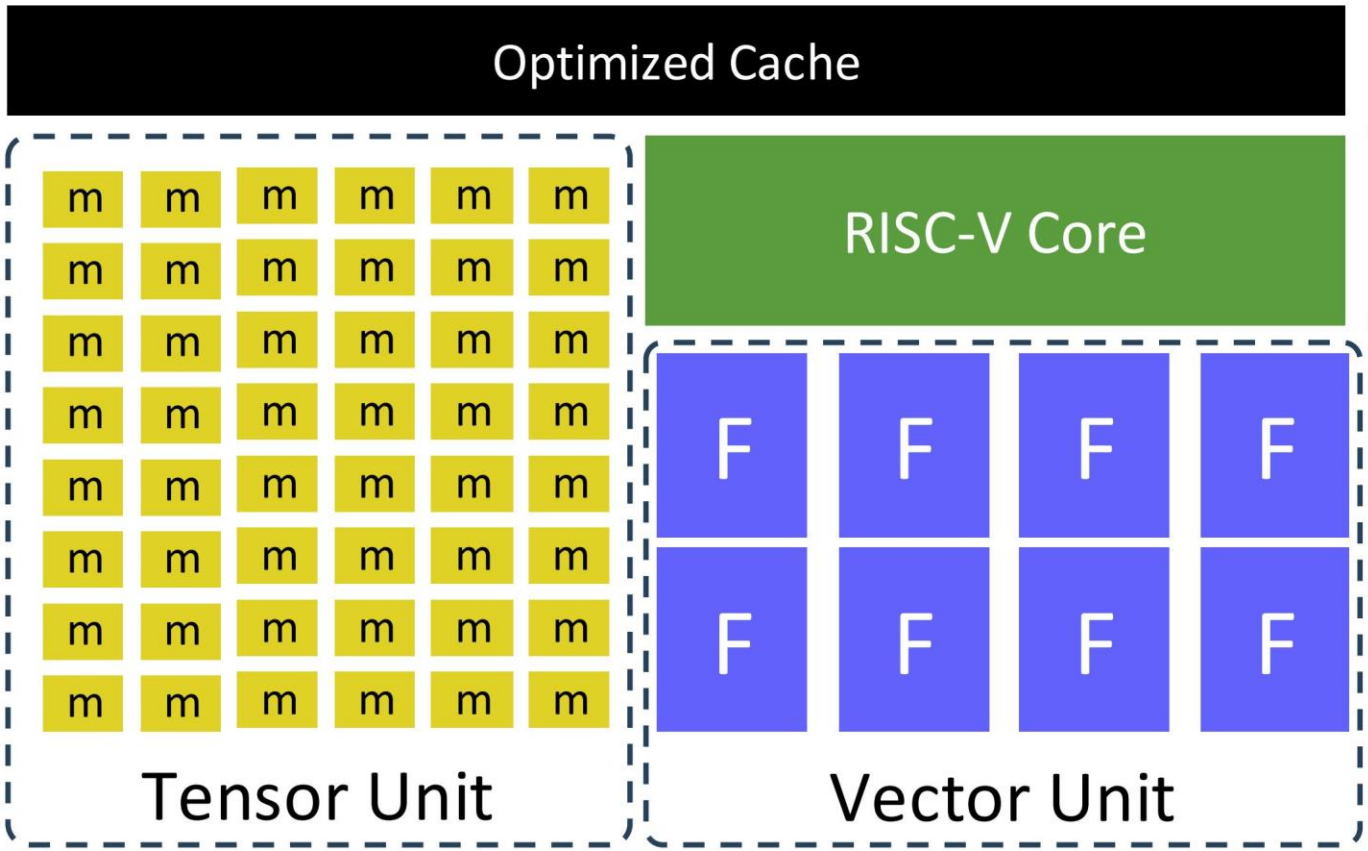


semidynamics

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 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

Our Customers AI Concerns - Solved

- 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?

- Wait – One more thing
  - KANs are coming – Are you ready? We are !

(*) KAN: Kolmogorov–Arnold Network

semidynamics

Thank you!

semidynamics