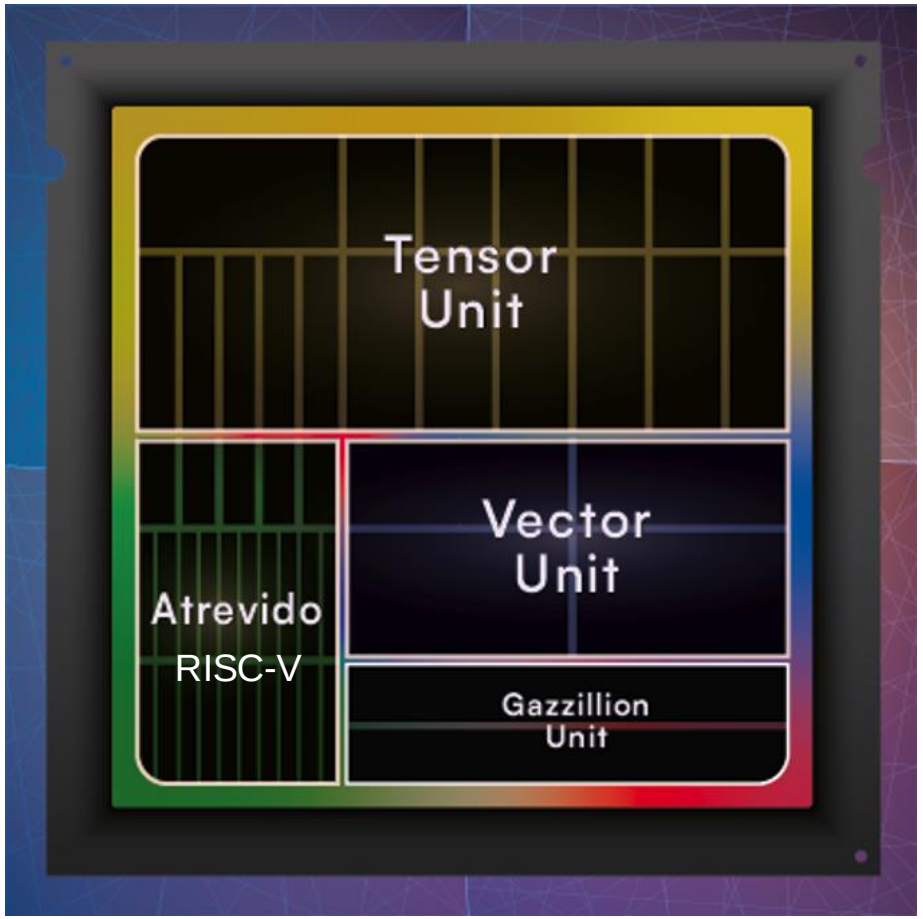
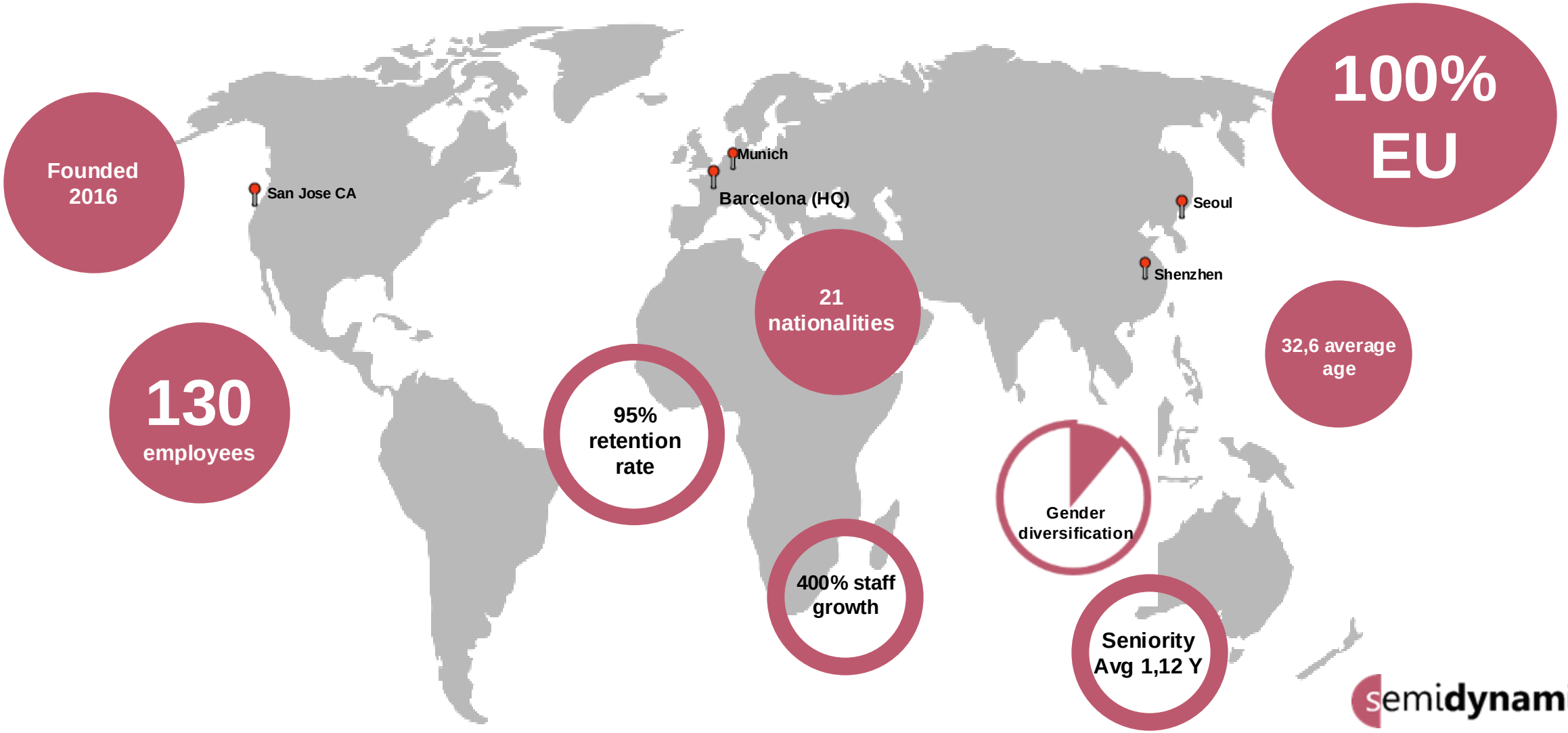
semidynamics

All-in-one CPU, Vector and Tensor RISC-V Computing



Roger Espasa, CEO

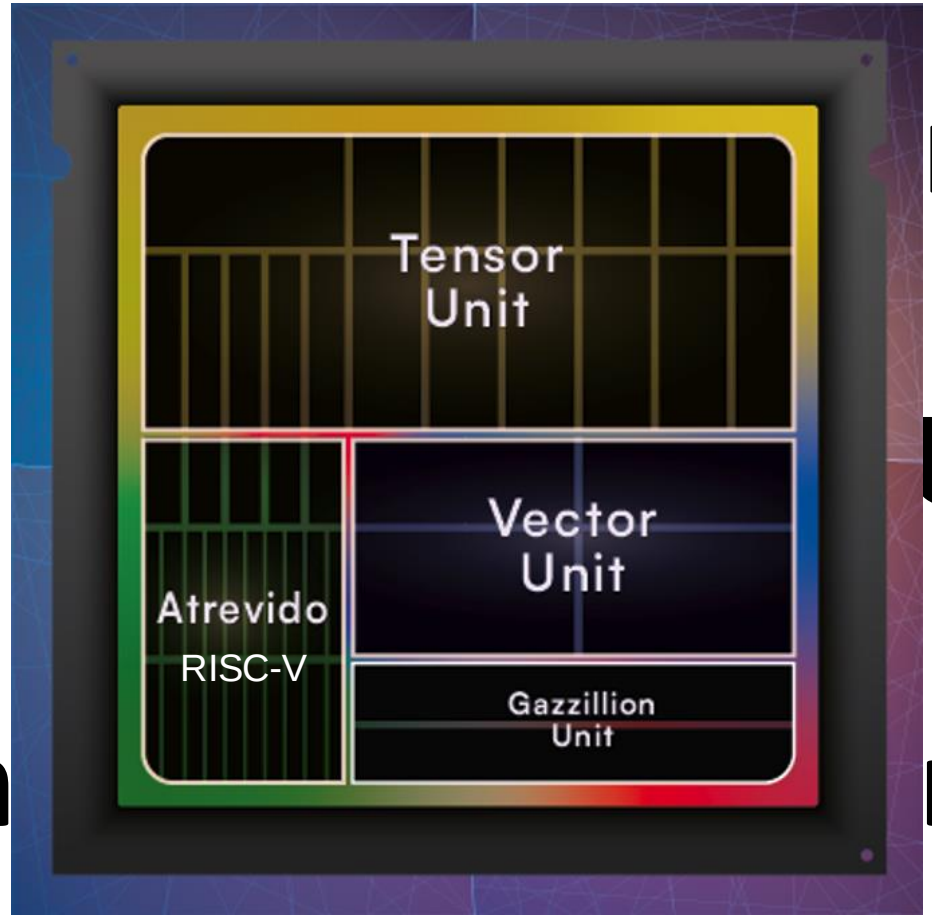
About Us



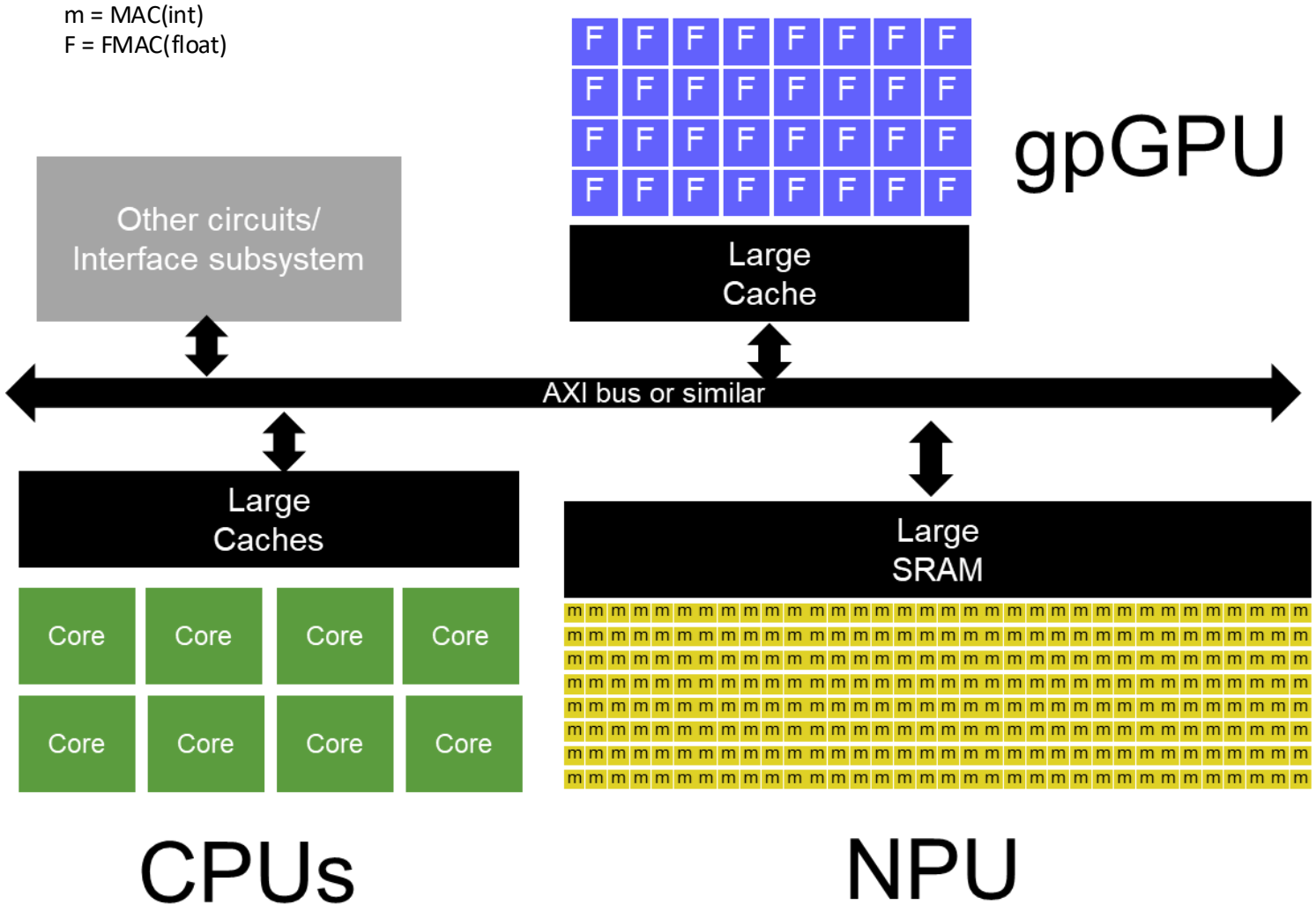
semidynamics

What do we do?

- We lice r AI
  - OOO
  - OOO **Unit**
  - OOO **Unit**
- **Combin ne”**

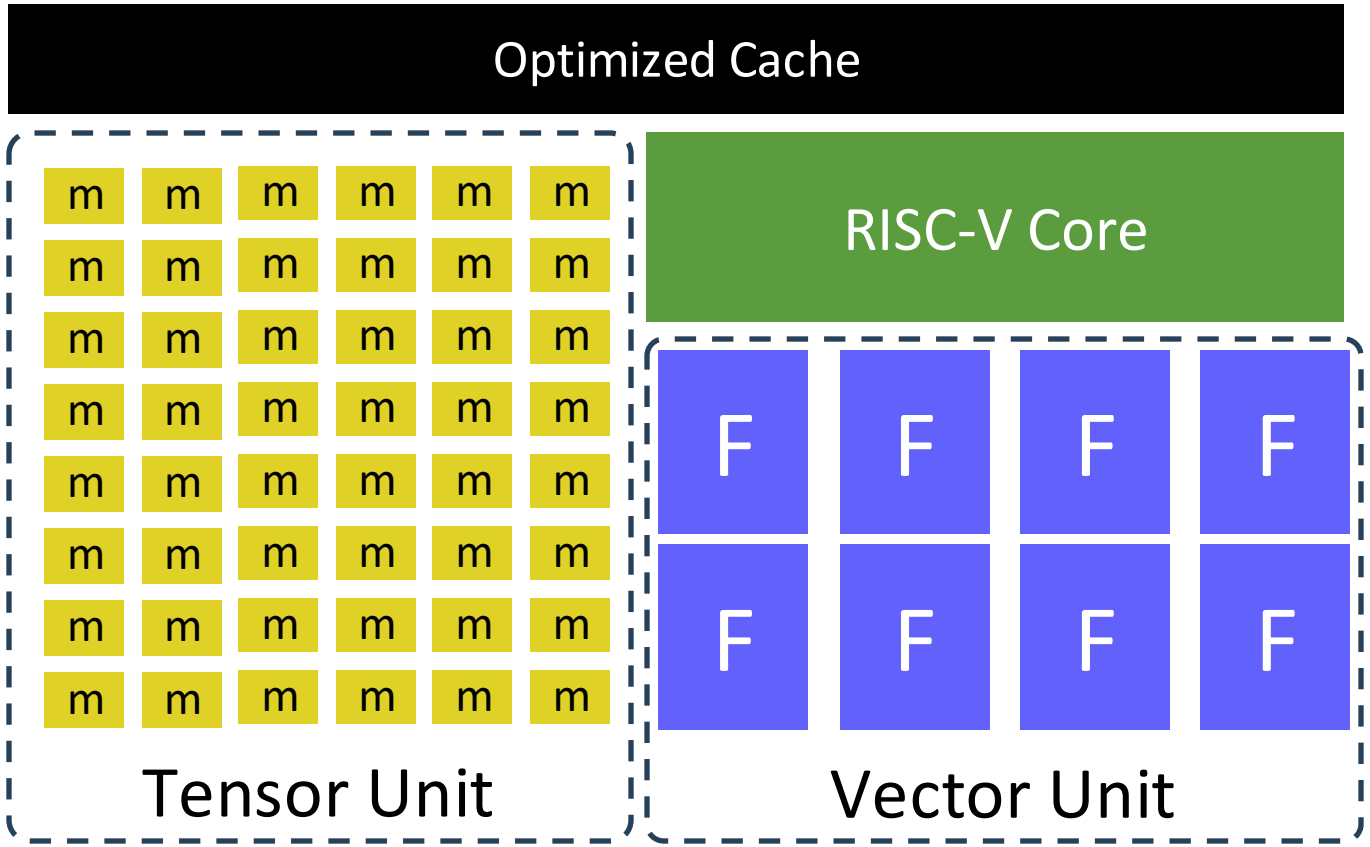


Old-Style AI Architecture



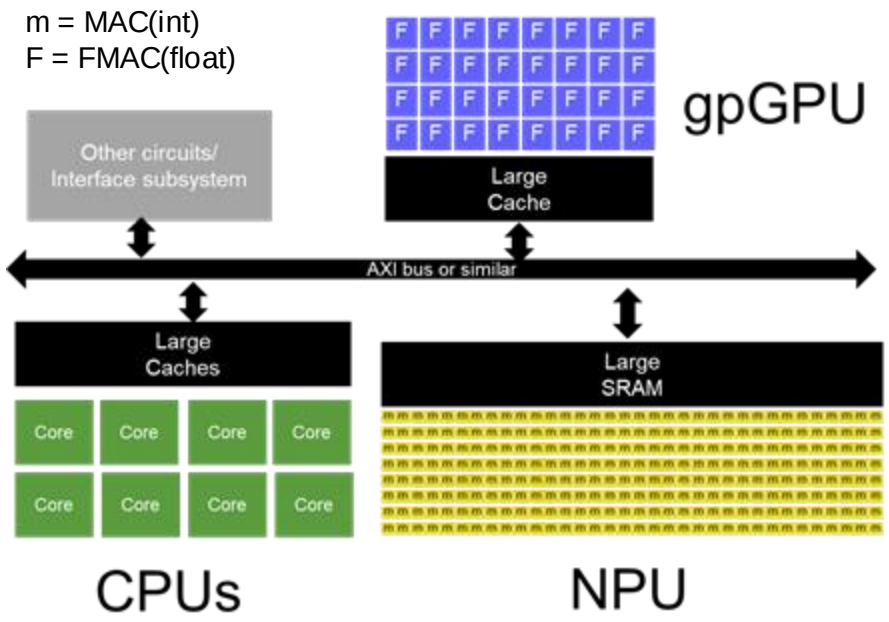
- **Three** Software Stacks
- **DMA-intensive** programming
- **High** Latency & Power
- **SRAM/Cache/Data** Replication
- **Unbalanced** Scaling
- **Not AI Future Proof**

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**

Current-Style AI Architecture



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

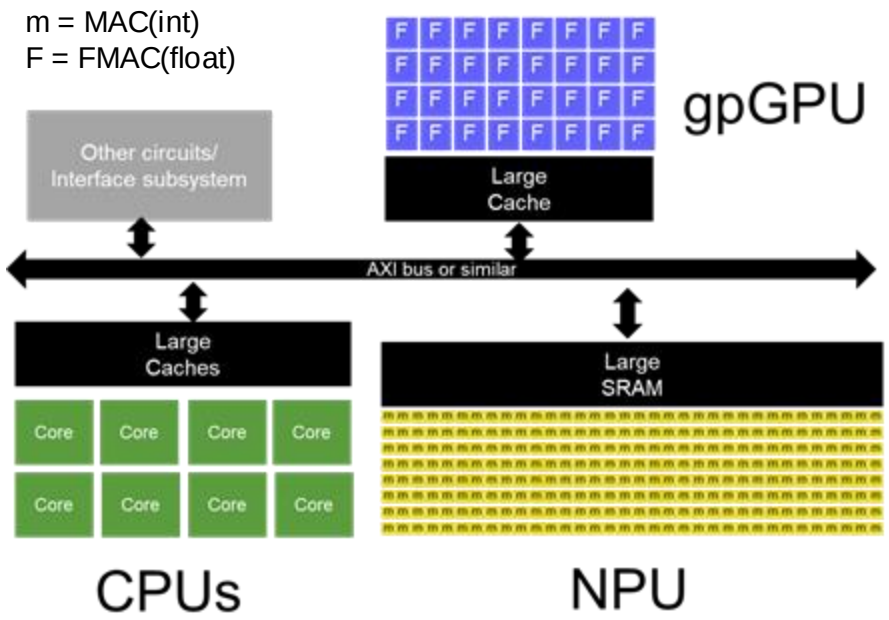
semidynamics

All-in-One

All-in-One

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

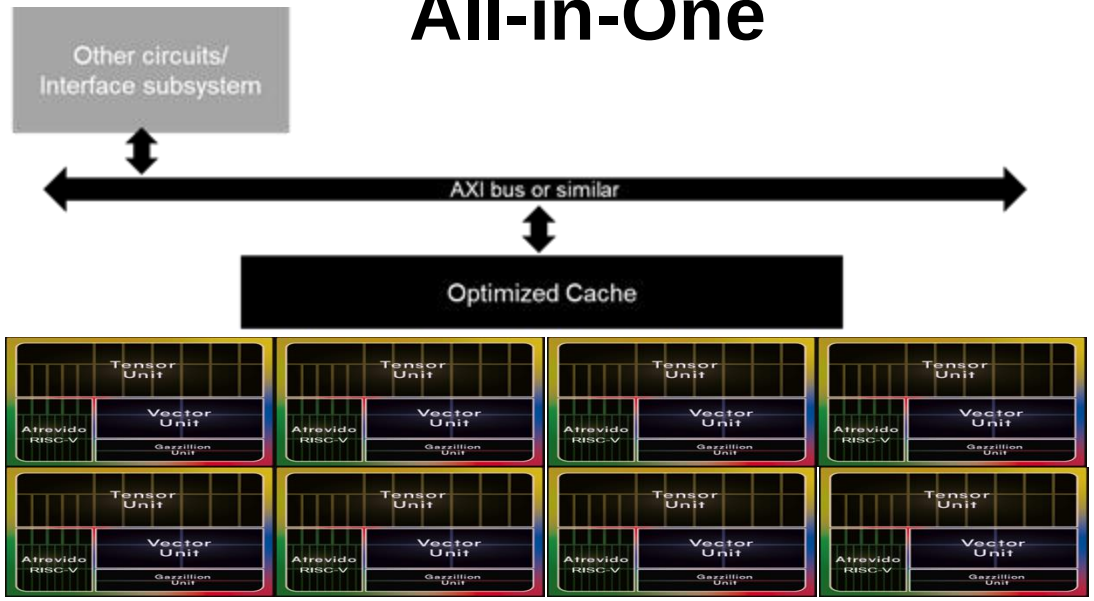
Current-Style AI Architecture



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

semidynamics

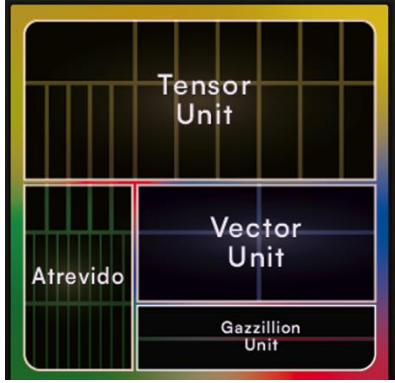
All-in-One



All-in-One

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

All-in-one Features



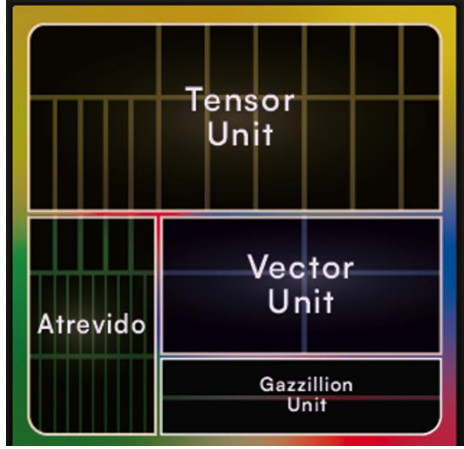
- 64b RISC-V
- **4-wide Out-of-order**
- U/S/M
- Hypervisor
- Linux-Ready
- MMU SV39/48
- Dcache: ECC
- Icache: Parity

- Fast unaligned
- Atomics
- PMP
- AXI4
- CHI.B, E
- Advanced Debug
- **Vector RVV1.0**
- **Vector Crypto**

- **Bit Manipulation**
- **Gazzillion ™ Technology**
- Tensor Unit: **INT4 INT8, INT16, FP16, BF16**

semidynamics

Highly Configurable



Tensor Unit	T1	T2	T4	T8
MACs	512	1024	2048	4096
Local SRAM?	No	No	Yes	Yes
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

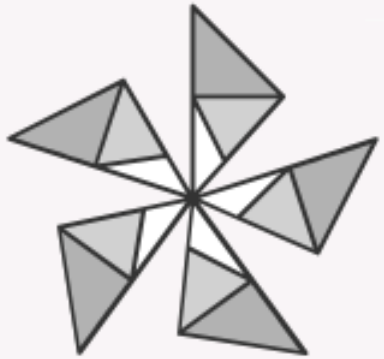
Vector Unit	V128	V256	V512
FMACs	**8**	**16**	**32**
INT8 GOPS/GHz	128	256	512
INT16 GOPS/GHz	64	128	256
BF16 GOPS/GHz	64	128	256
FP16 GOPS/GHz	64	128	256
FP32 GOPS/GHz	32	64	128
FP64 GOPS/GHz	16	32	64

semidynamics

Software stack(s) (and some models)

Aliado
RISC-V SDK

Learn more >



Semidynamics
ONNX-Runtime

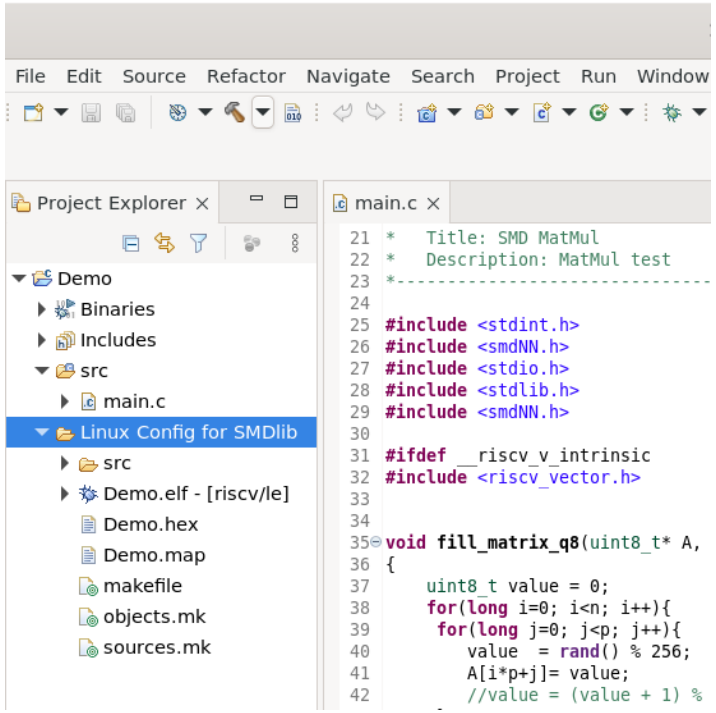
Learn more >

Semidynamics
AI Models

Learn more >

Aliado RISC-V SDK

IDE



Compilers

GCC

Emulators

QEMU

AKL
Aliado Kernel Library

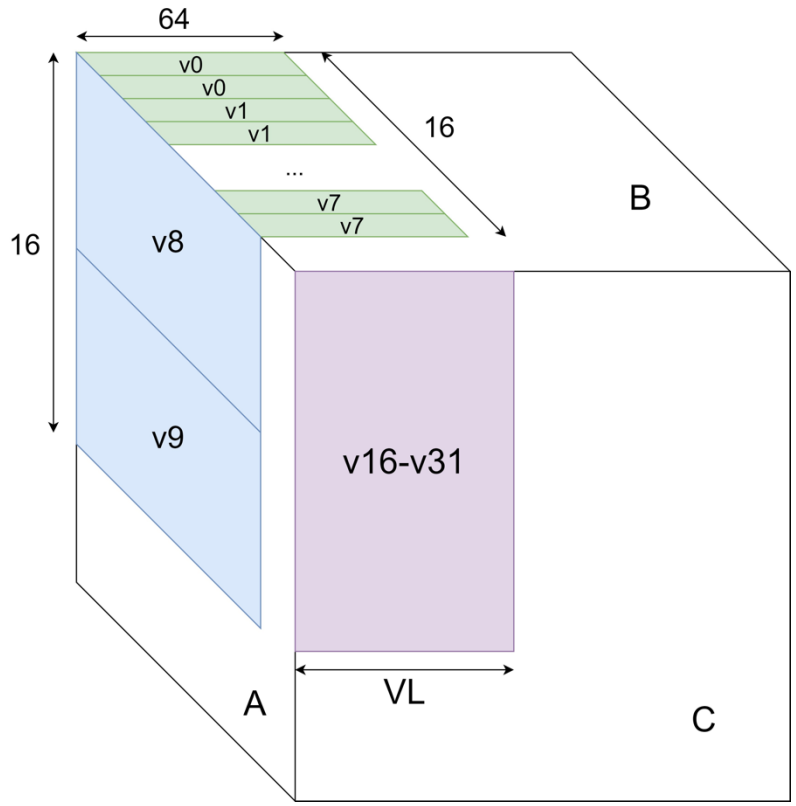
Parallelism

OpenMP

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
```

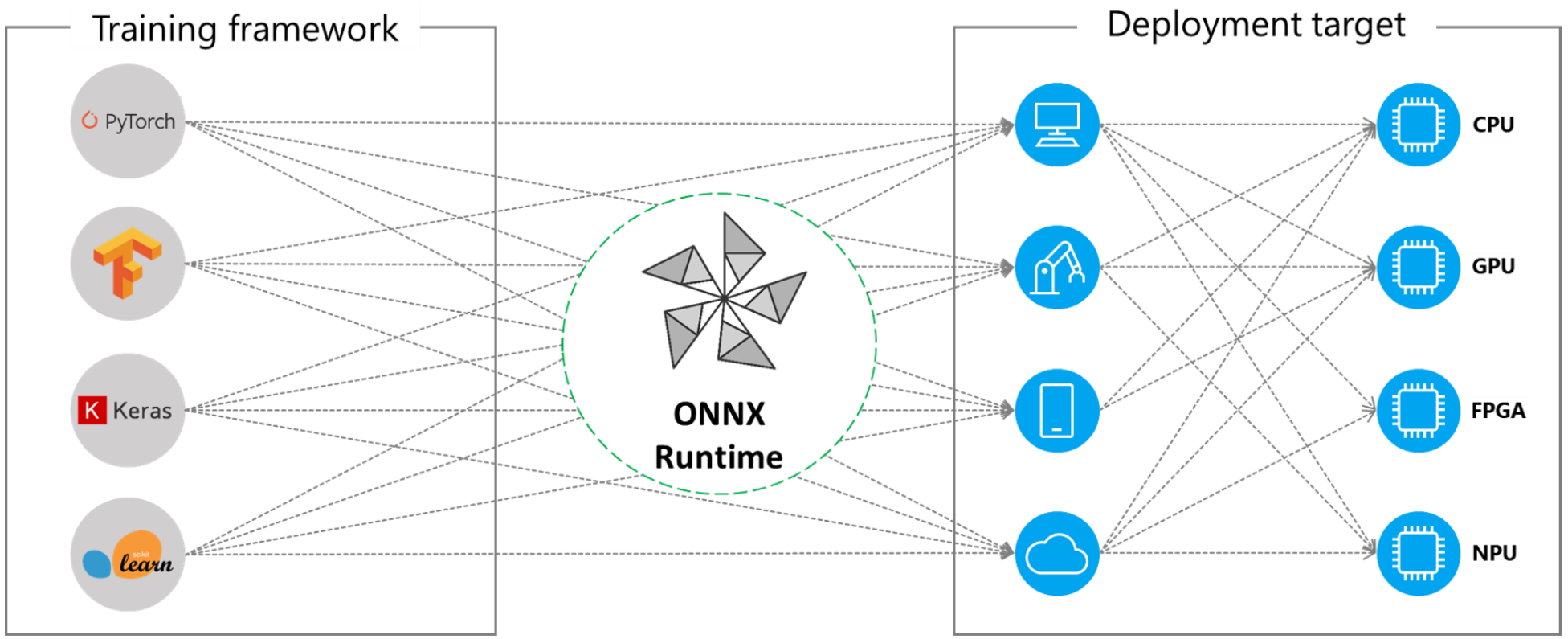


Aliado Kernel Library

semidynamics

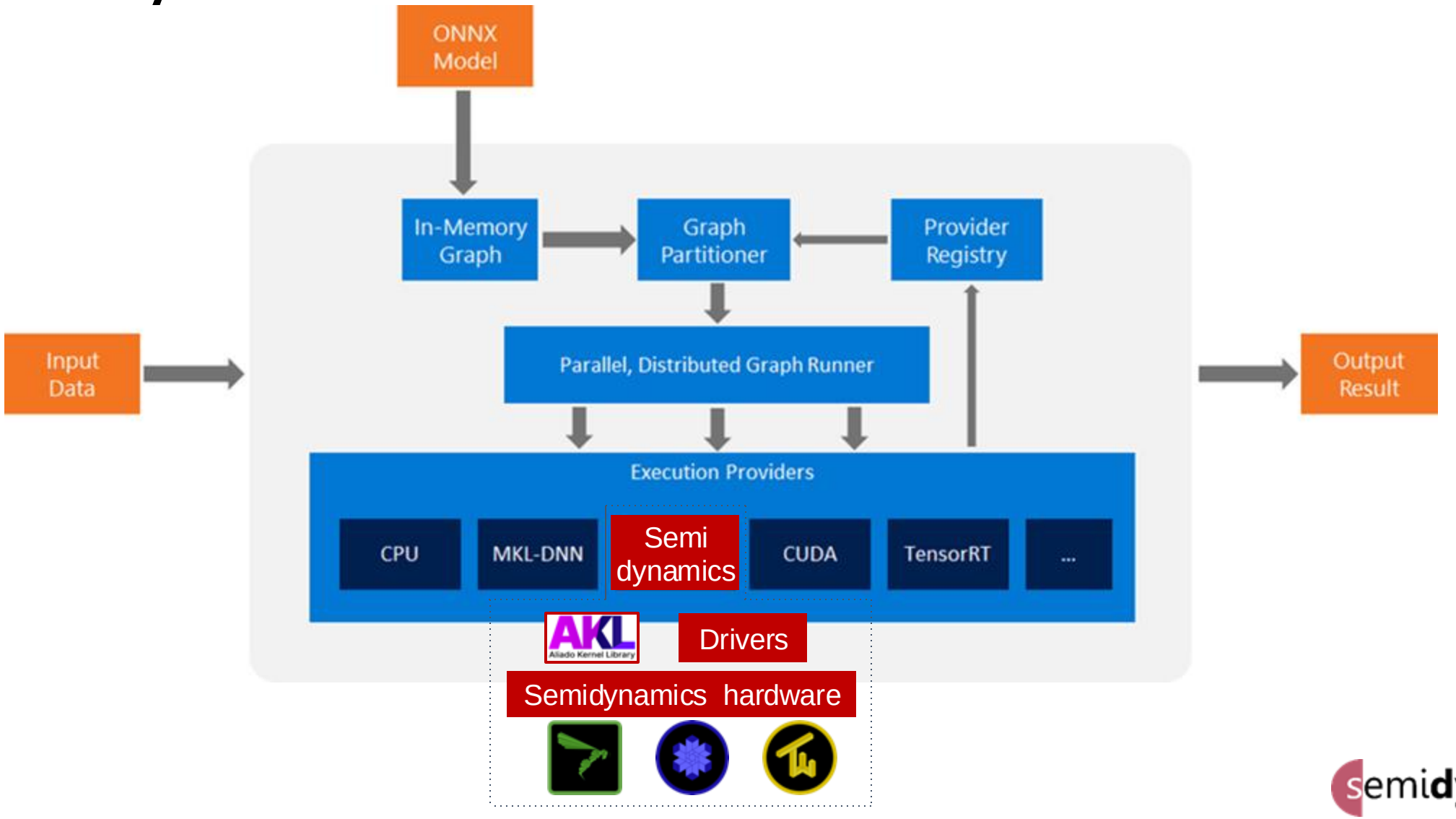
ONNX Run Time (ORT)

- ONNX Runtime is a cross-platform machine-learning model accelerator, with a flexible interface to integrate hardware-specific libraries. ONNX Runtime can be used with models from PyTorch, Tensorflow/Keras, TFLite, scikit-learn, and other frameworks.



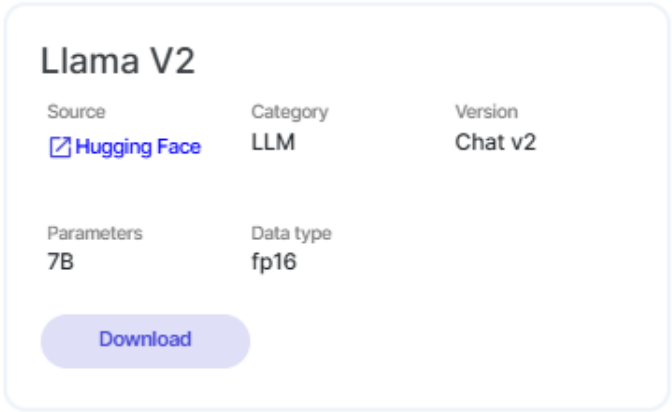
semidynamics

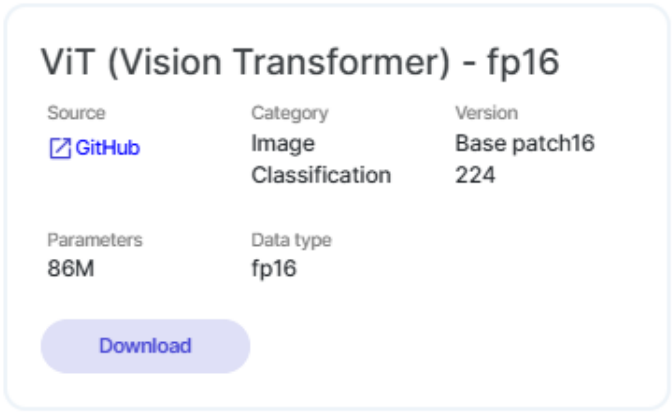
Semidynamics Execution Provider

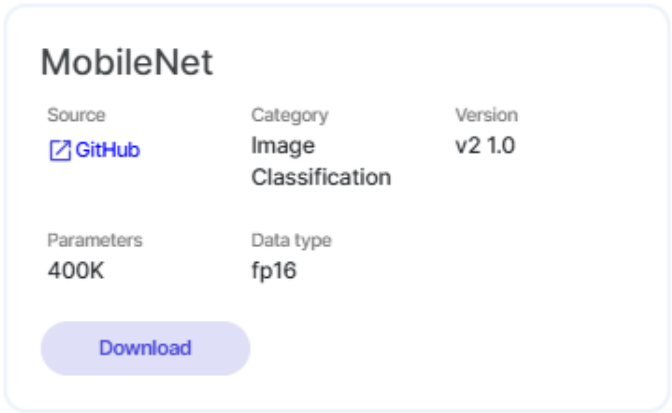


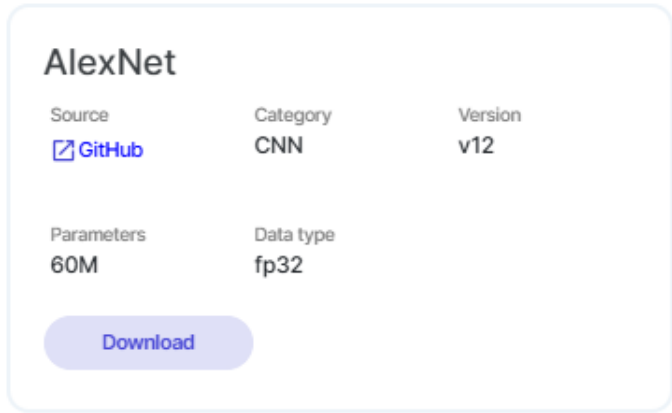
Some curated models (zoo)

- LLama3 ready
- Deepseek-R1-dist ready







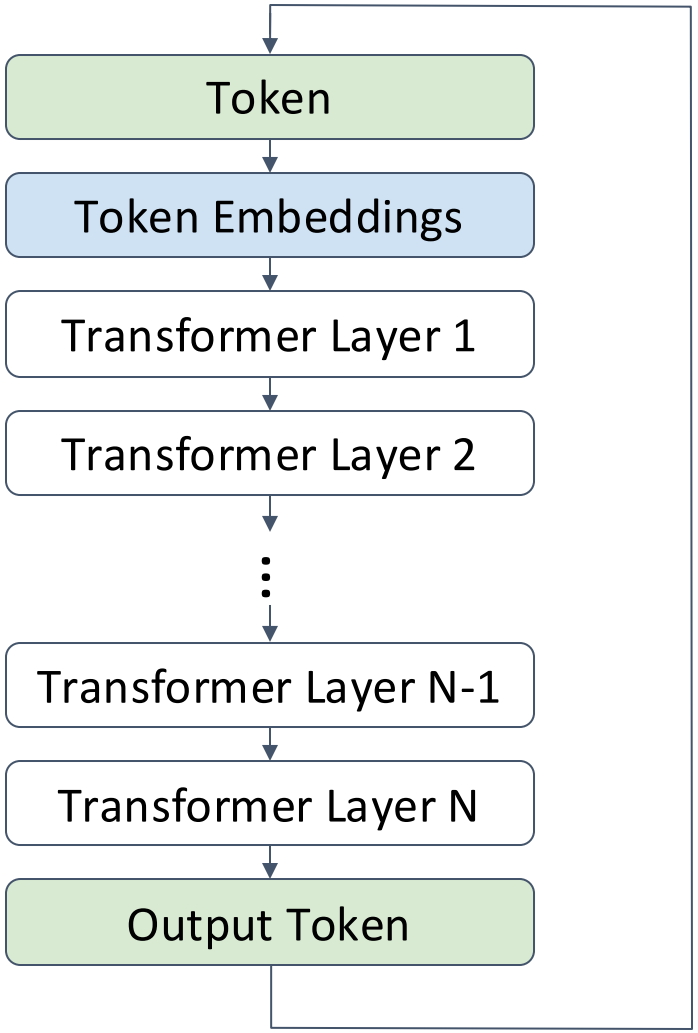


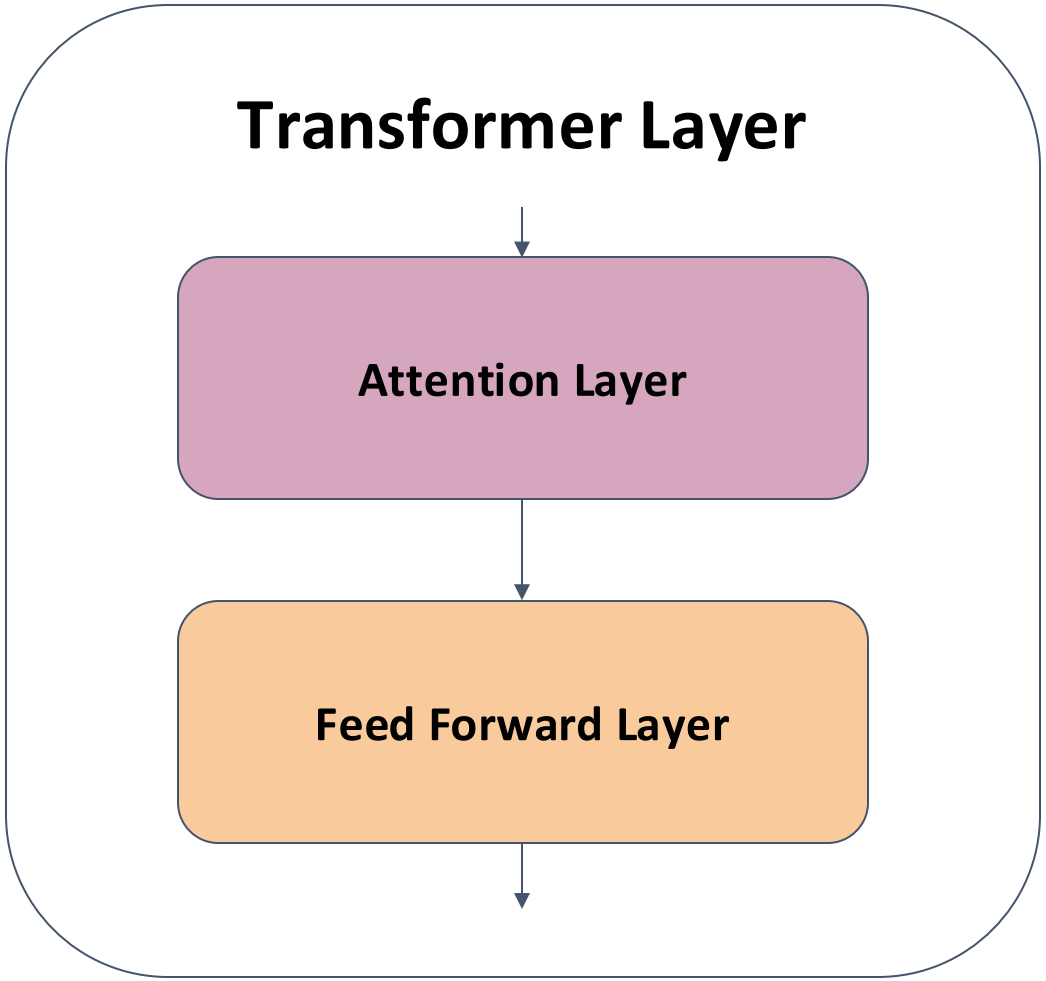
Running Transformers / LLMs on All-In-One solution

Llama-2, FP16, 7B Parameter

semidynamics

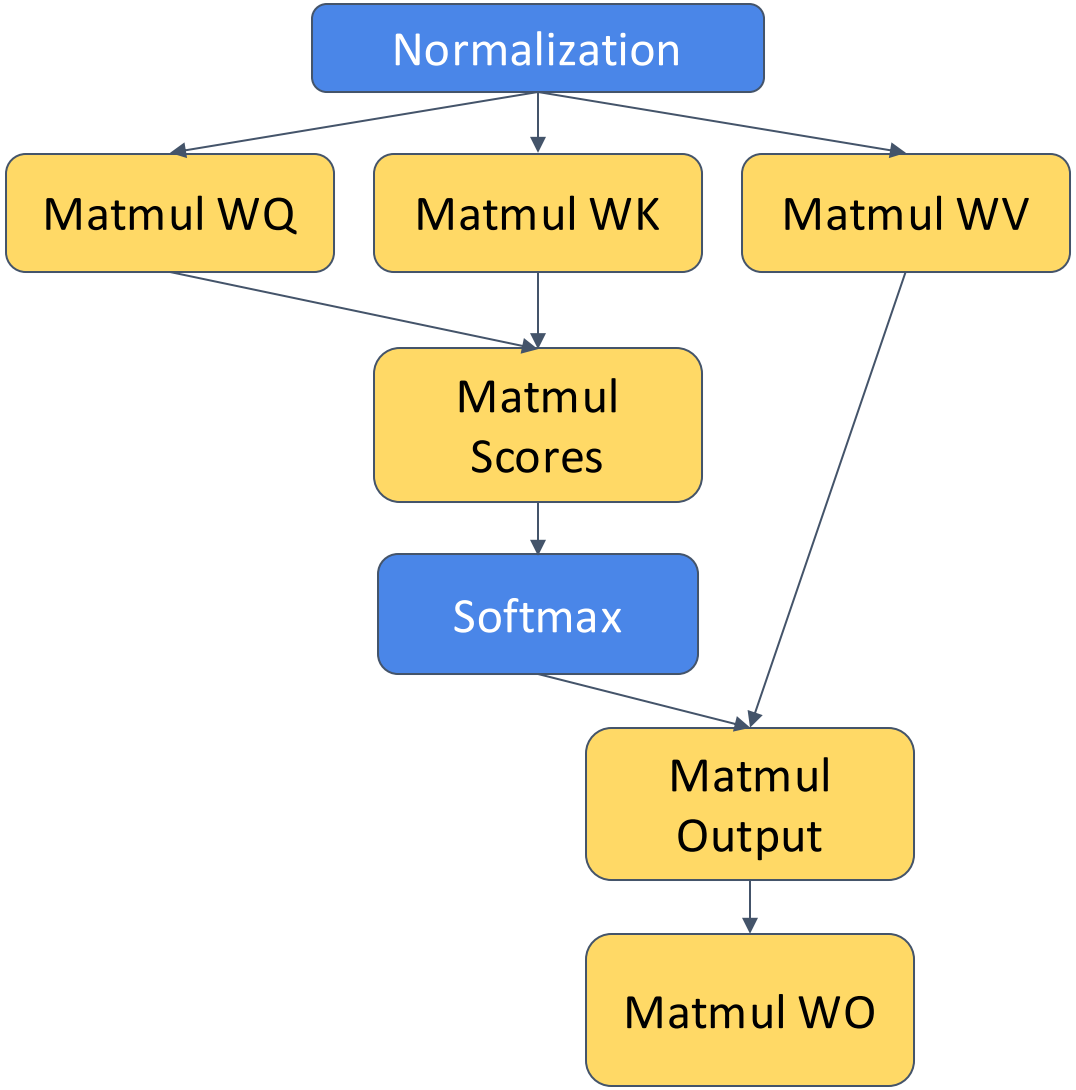
Llama-2 in a nutshell

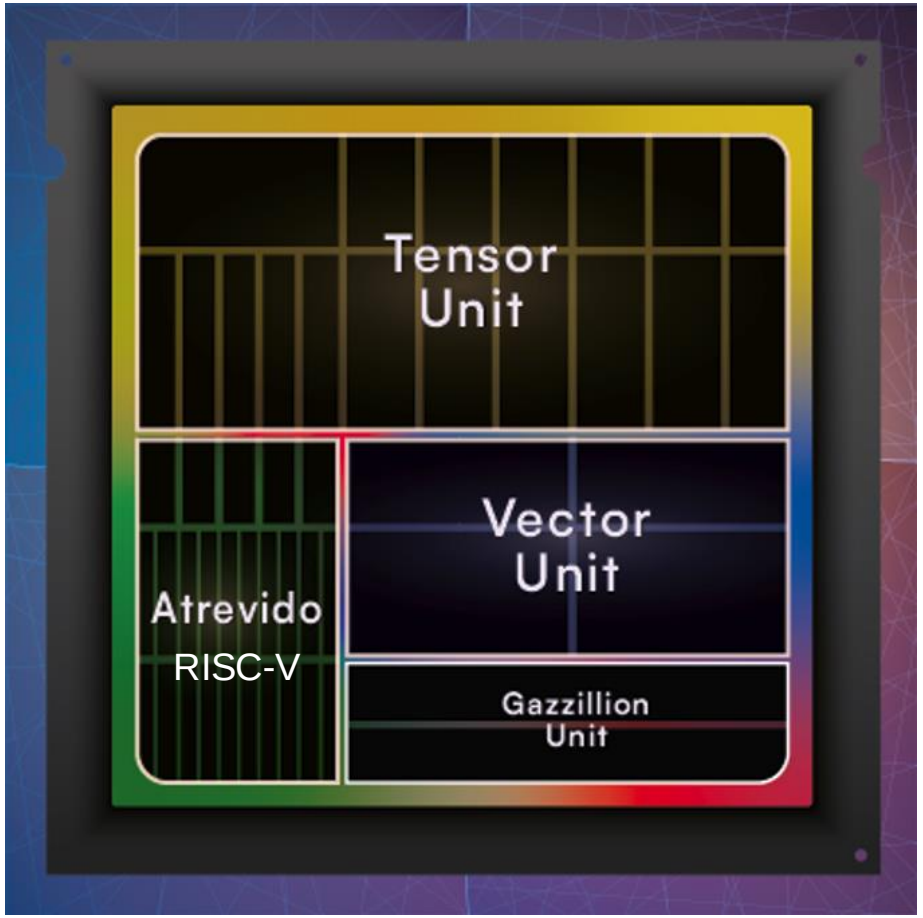




semidynamics

Attention Layer



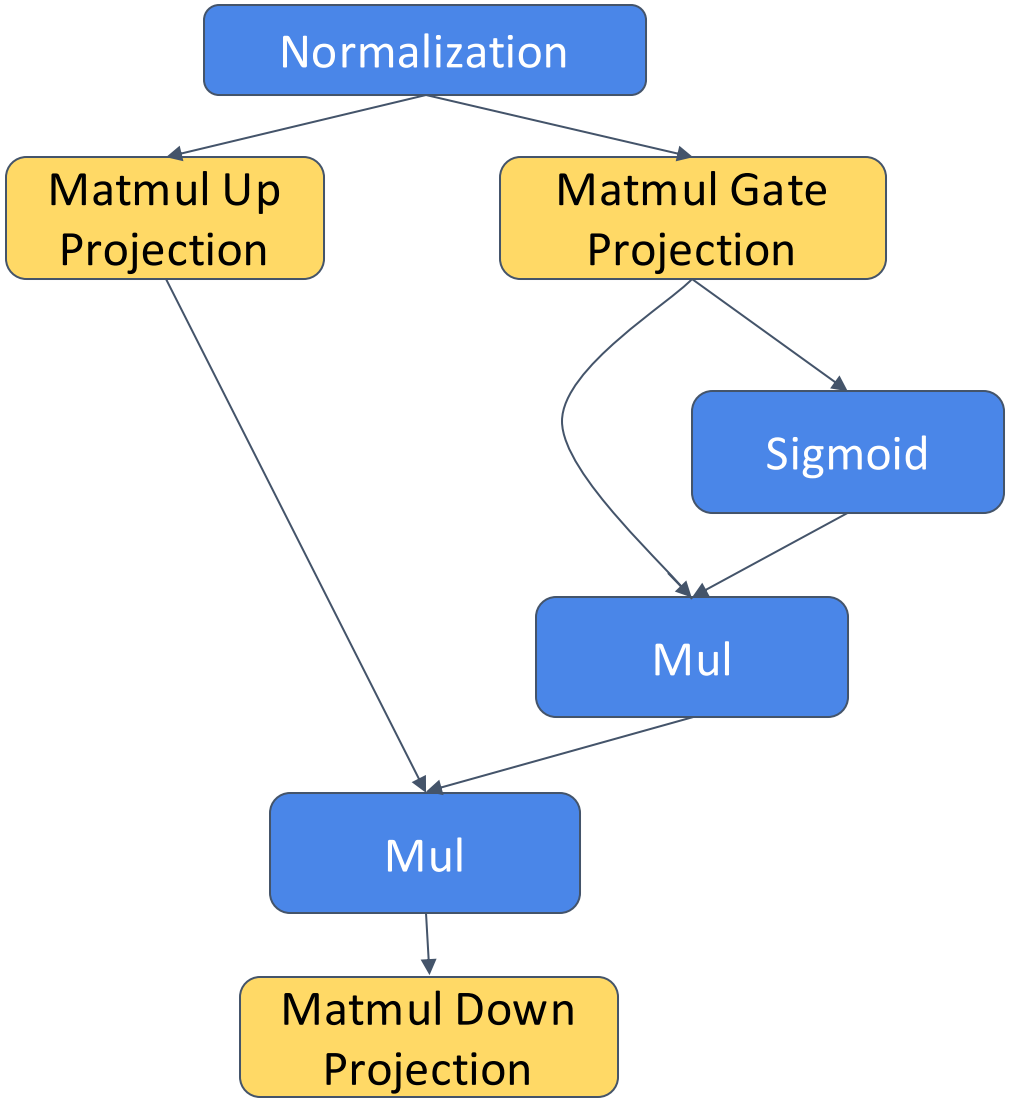


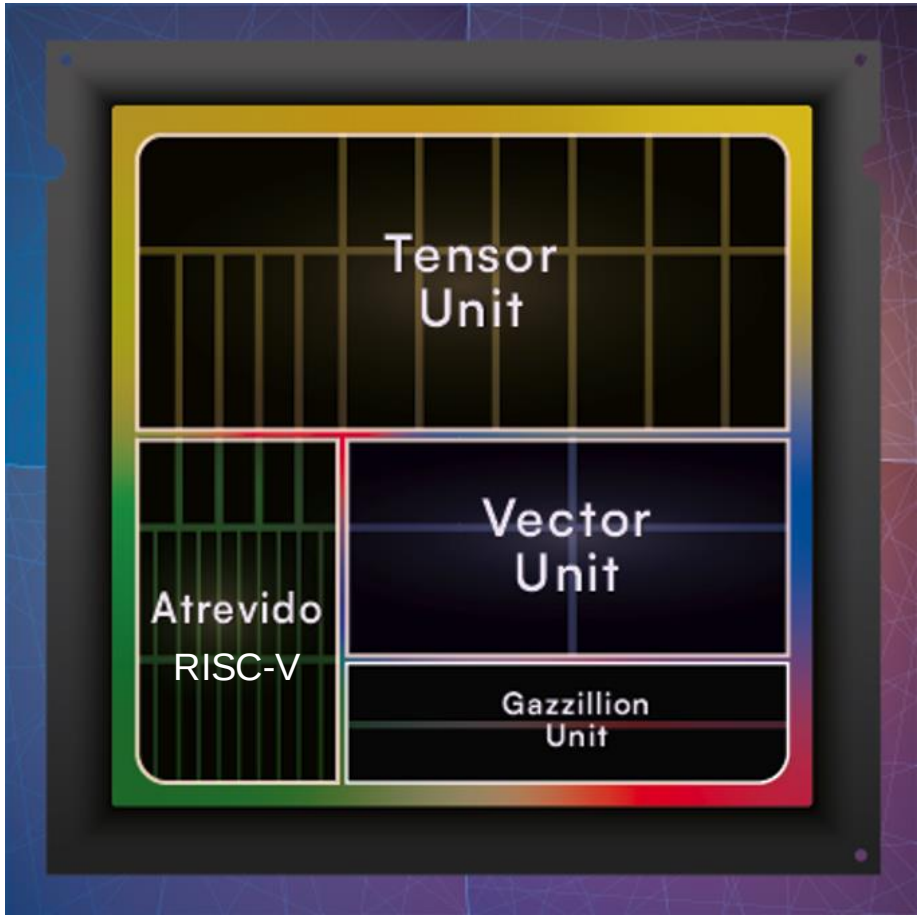
Runs in Tensor Unit

Runs in Vector Unit

semidynamics

Forward Layer





Runs in Tensor Unit

Runs in Vector Unit

semidynamics

Llama-2

FP16,
7B params

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

semidynamics

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

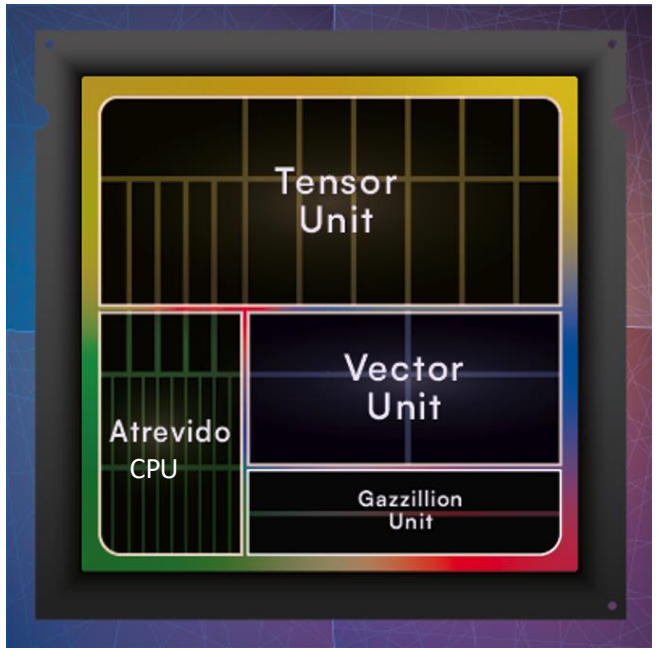
Llama-2

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

Scaling to more TOPS? Easy!



Replicate to create massive parallelism

Can we support DeepSeek?

```
SMD NUSMI: Hello I am Llama 3 Deepseek R1 distilled AI assistant, how can I help you?
User: What is 1+1?
SMD NUSMI: <|think|>
First, I recognize that the user is asking for the sum of 1 and 1.

Next, I'll identify the numbers involved in the addition, which are both 1.

Then, I'll perform the addition operation by combining the two numbers.

Finally, I'll present the result clearly to the user.
<|/think|>

**Solution:**

We are asked to find the sum of 1 and 1.

1. **Identify the numbers to add:**
   - First number: **1**
   - Second number: **1**

2. **Perform the addition:**
   \[
   1 + 1 = 2
   \]

3. **State the final answer:**
   \[
   \boxed{2}
   \]
User: NUSMI shut down
#
```

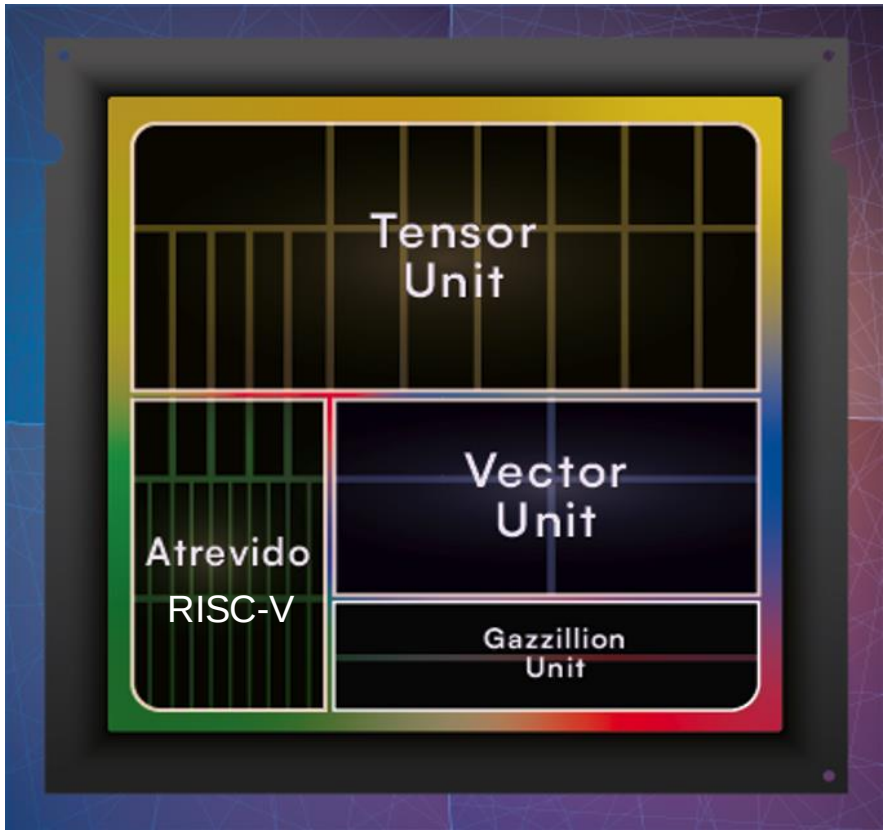
Can we support DeepSeek?

```
SMD NUSMI: Hello I am Llama 3 Deepseek R1 distilled AI assistant, how can I help you?
User: What is 1+1?
SMD NUSMI: <|think|>
First, I recognize that the user is asking for the sum of 1 and 1.

Next, I'll identify the numbers involved in the addition, which are both 1.

Then, I'll perform the addition operation by combining the two numbers.

Finally, I'll present the result clearly to the user.
<|/think|>

**Solution:**

We are asked to find the sum of 1 and 1.

1. **Identify the numbers to add:**
   - First number: **1**
   - Second number: **1**

2. **Perform the addition:**
   \[
   1 + 1 = 2
   \]

3. **State the final answer:**
   \[
   \boxed{2}
   \]
User: NUSMI shut down
#
```

OF COURSE!

semidynamics

Best IP & Sofware for AI



Aliado
RISC-V SDK
Learn more >

Semidynamics
ONNX-Runtime
Learn more >

semidynamics

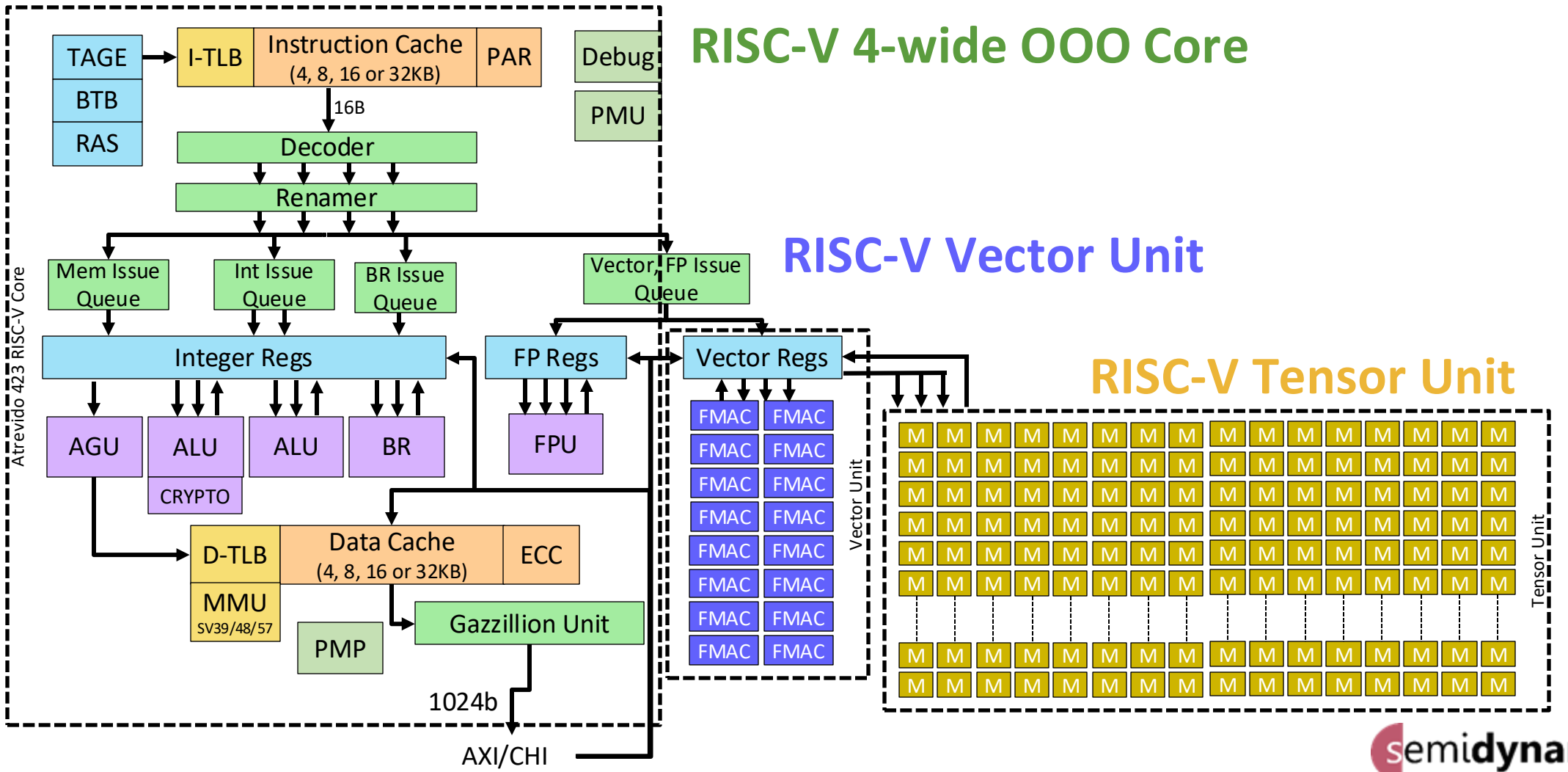
Wait… what about Space?

- ECC 
- Parity 
- Rad Hard …
  - Looking forward to working with you

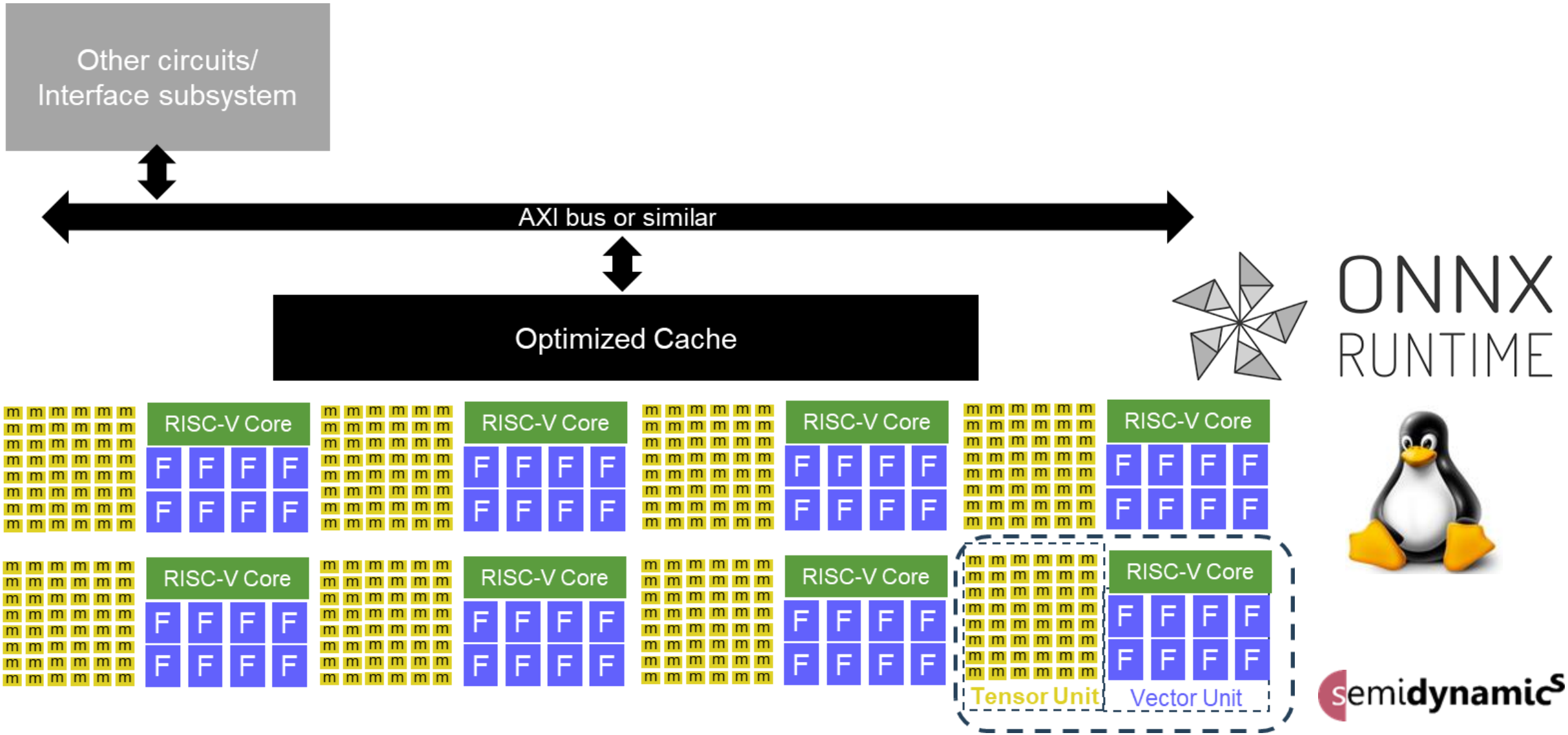
Thanks!!!

semidynamics

All-In-One Block Diagram



But… where is your ONNX RT SW running?



But… where is your ONNX RT SW running?