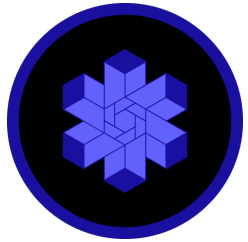
semidynamics

Running Transformers on Semidynamics' "All-In-One" Vector and Tensor Unit

Roger Espasa, CEO

In Order Core

OOO Core



OOO Vector Unit

Tensor Unit

Market Trends & Challenges

Trends

- More data to be handled from Sensors
- More AI — even edge devices execute generative AI and LLM apps locally with billions of parameters

Challenges

- Increasing processing power needs for AI workloads
- Hugh amount of stored data increases likelihood of cache misses
- Increasing CPU performance needs
- Hypervisors & Containers needed for several guest OSes and domains
- Time to market & scalability & future proofing of ‘SoC’ solutions

semidynamics

NPU Challenges

AI compute market moving to Edge

Performance needs keep going up

Types of deployed networks & models keeps evolving

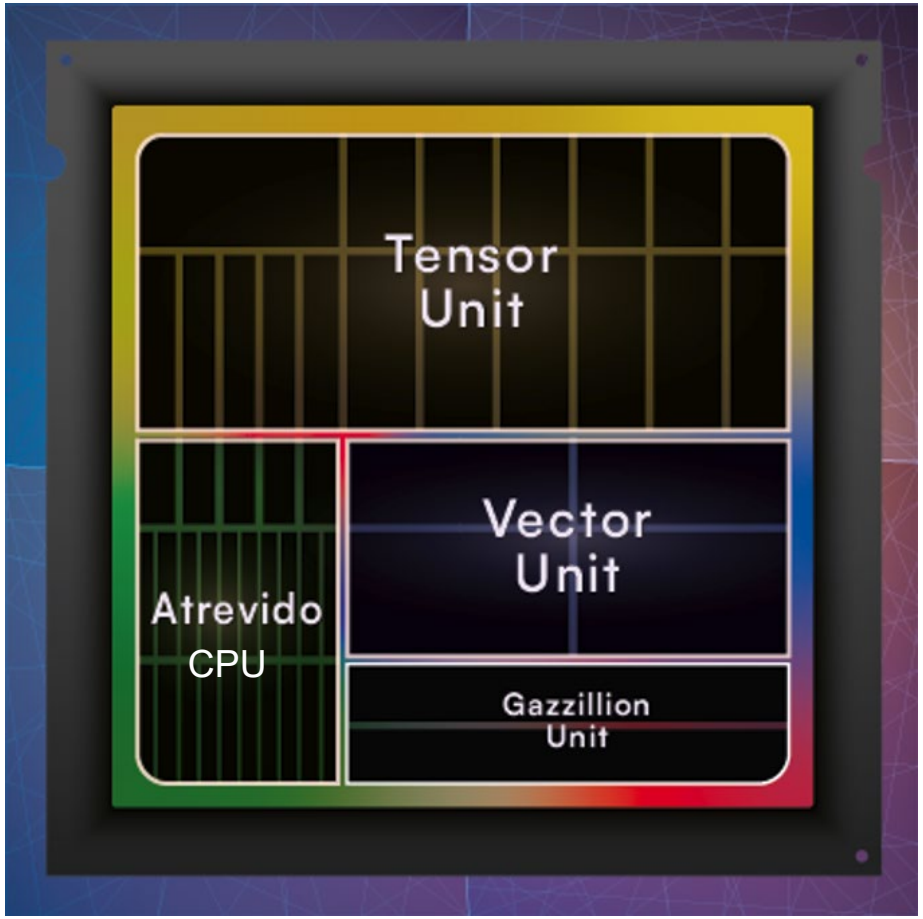
Speed of change and new model adoption is accelerating, creating need for flexible and adaptable NPU designs !

New SOCs require a new compute paradigm !

semidynamics

The Semidynamics AI Approach
All-in-one: merging CPU, NPU, GPU

- Powerful **Out Of Order** based on Risc-V
- Combine **CPU** with **Vector** and **Tensor unit**
- Creates powerful AI capable Compute scalable building blocks
- Enable Hypervisor Support for Containerization
- Enable Crypto for Security / Privacy
- Use of **Gazzillion ™ Technology** to manage large date sets



Benefits

- **Easy, DMA-free,** programming with **single** RISC-V **software stack**
- **Zero** Communication Latency & **Low** Power
- **Scalable High Performance**
- **AI Future-proof**

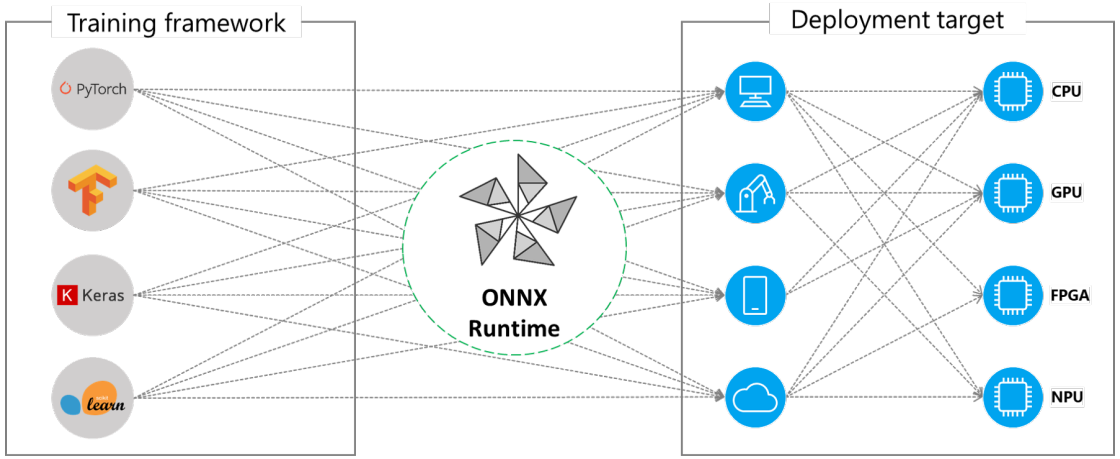
semidynamics

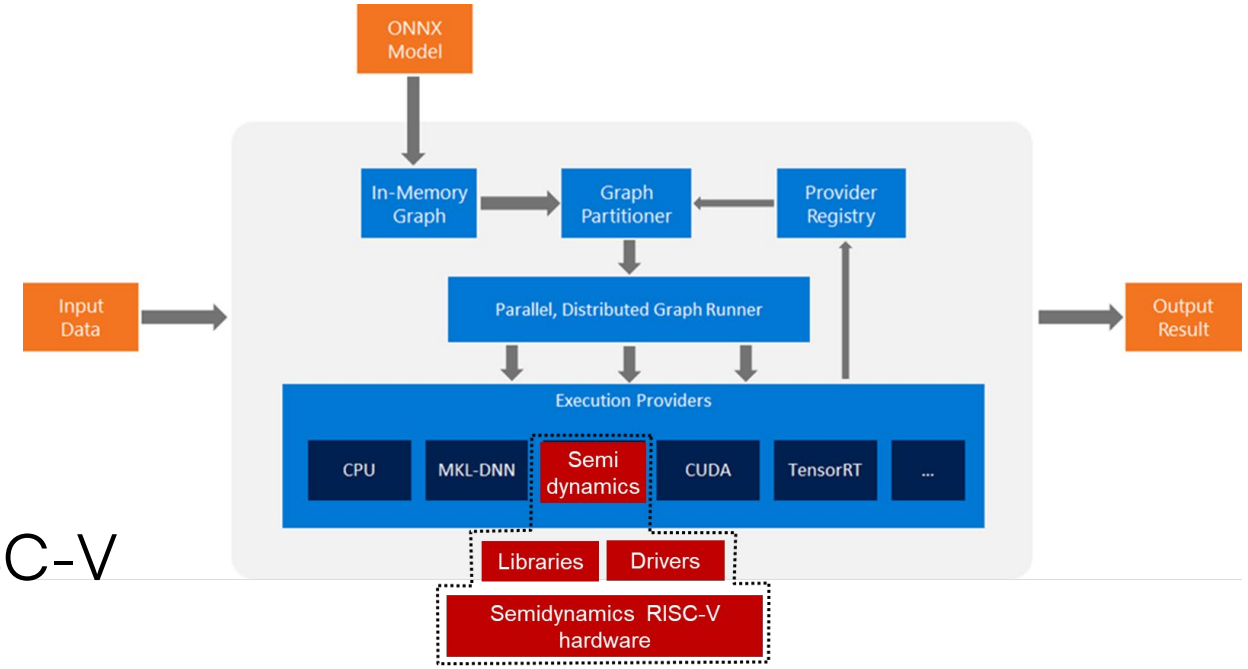
AI Customers Concerns: How to run AI models on Semidynamics All-in-one IP?

- 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?

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

Semidynamics AI SW Stack: No Compilers !





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

Llama-2 FP16, 7B params

- Using 1 $TOPS_8$ T1 Tensor Unit
- And 128 $GOPS_8$ V8 Vector Unit

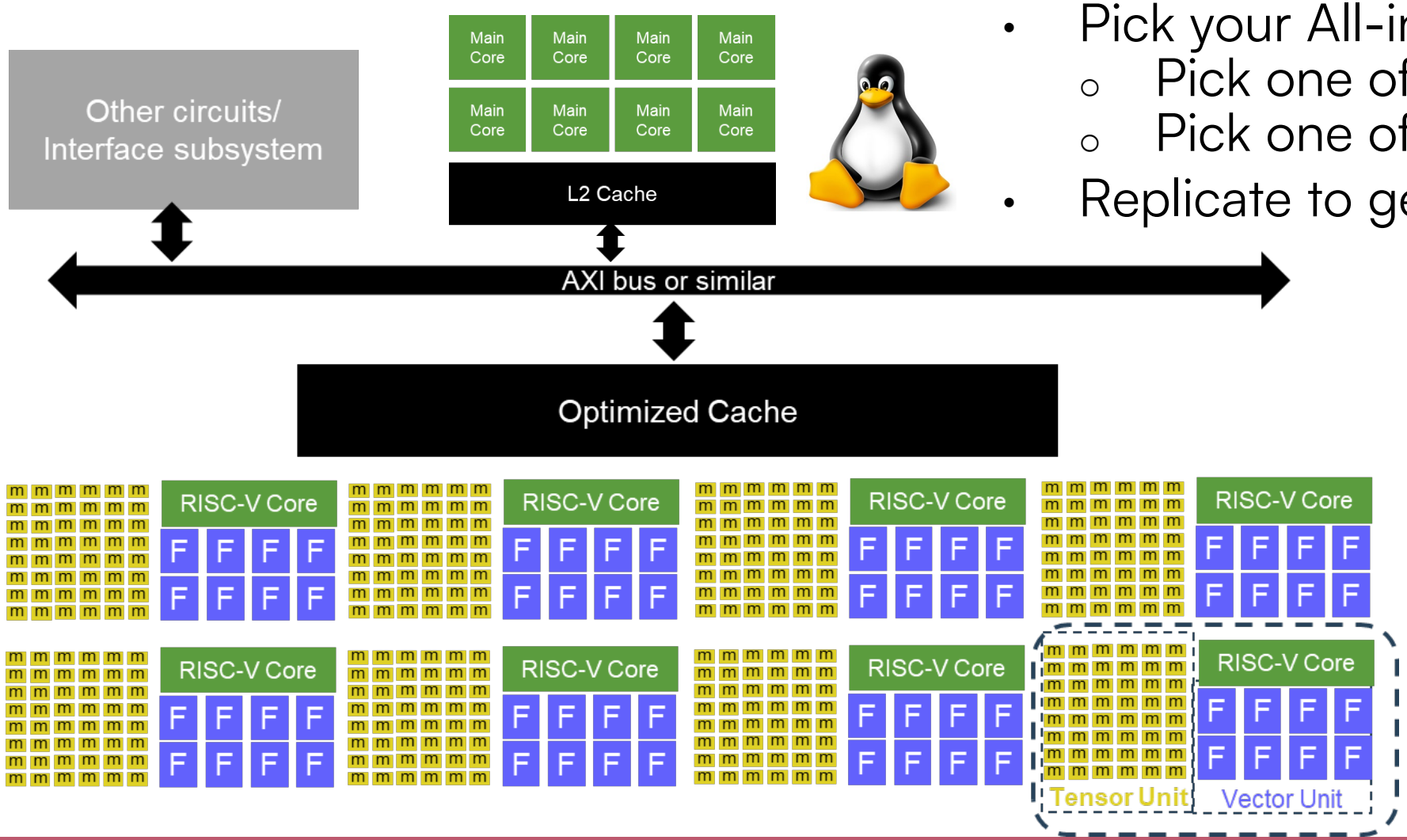
Operators	Scalar	T1	T1+V8
Matmul	99%	20%	55%
Activations	1%	80%	45%
Concat	0.11%	19%	17%
Sigmoid	0.09%	16%	2%
ScatterND	0.09%	15%	17%
Div	0.06%	9.5%	2%
Mul	0.03%	5.7%	2.4%
Slice	0.03%	5.0%	1.3%
Exp	0.03%	4.4%	0.5%
Other	0.54%	5.4%	2.8%
Speedup %	1X	170X	470X

Perfectly balanced processing on All-in-one RISC-V AI IP

semidynamics

Concern #3: Can I easily **scale** your solution?

Scaling up All-in-one solution

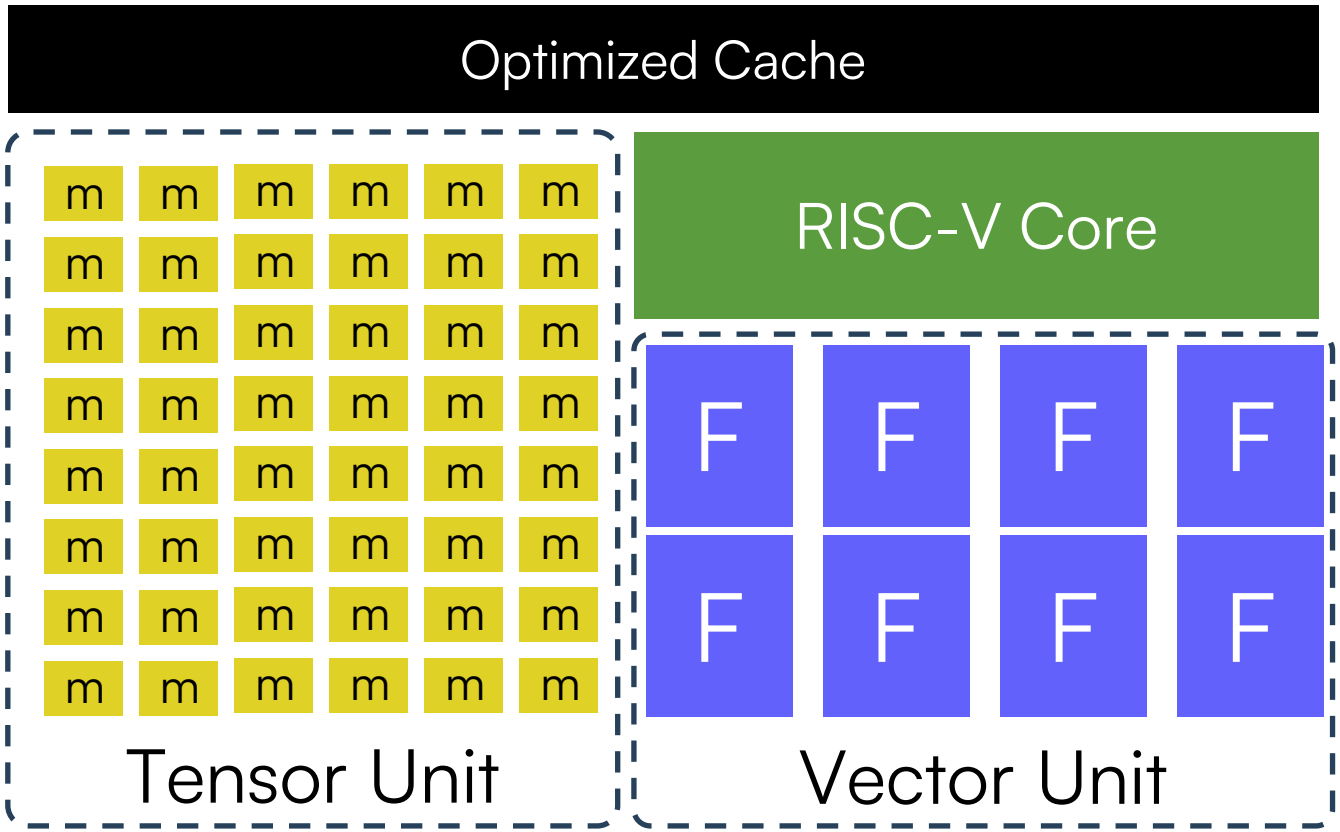


- Pick your All-in-one "building block"
  - Pick one of T1, T2, T4, T8
  - Pick one of V8, V16, V32
- Replicate to get **balanced scaling**

ONNX RUNTIME

semidynamics

All-in-one is future-proof



- 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?

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

(*) KAN: Kolmogorov—Arnold Network

semidynamics

Thank you!

semidynamics