Semidynamics' Highly Configurable OOO Vector Unit

Roger Espasa

2023.06.06

About Semidynamics

Semidynamics is a **European** supplier of RISC-V IP cores, specializing in customization of **high bandwidth high performance cores with vector units** for tailored projects

Experts in open core surgery

semidynamics

Our RISC-V Core IP Families

Atrevido

2, 3 or 4-wide **out-of-order**
RISCV64GC
AXI and CHI
VECTOR READY!

Avispado

2-wide **in-order**
RISCV64GCV
AXI and CHI
VECTOR READY!

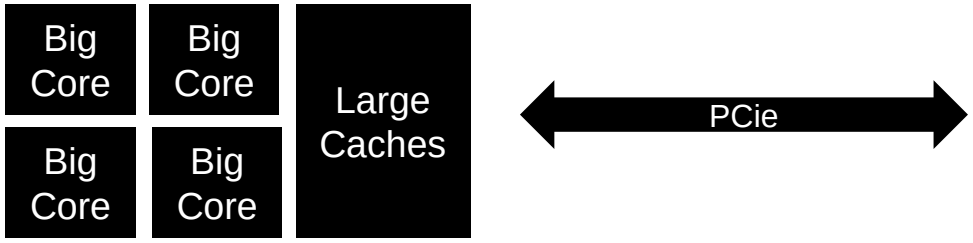
World's first, **fully customizable**, 64-bit RISC-V cores for ultra fast, big memory applications, optimized for a companion RISC-V **vector unit**

Unique tailor-made PPA solutions include customer's secret sauce for product differentiation and IP protection.

semidynamics

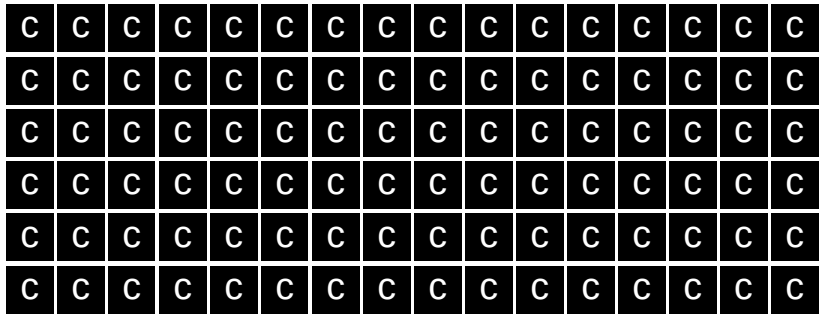
Before the vector unit...

CPU



- Few, large cores
- Easy to program

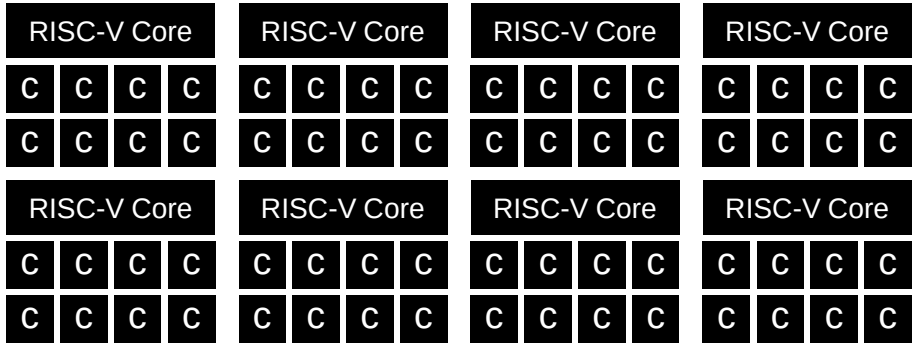
GPU



- Many tiny cores
- Hard to program
- High Performance for Parallel Code
- Communication Latency

semidynamics

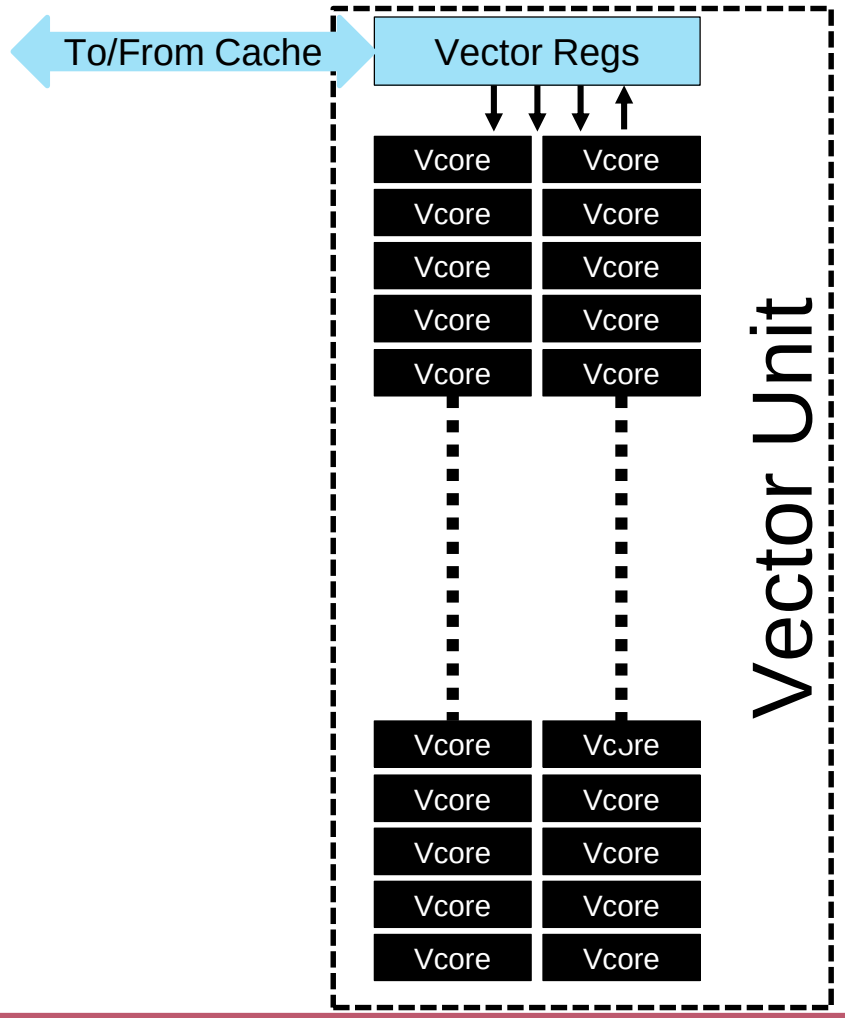
CPU + Vector Unit: best of both worlds



- Bring the GPU compute cores next to the CPU cores
- **Easy** to program
- **High Performance** for Parallel Codes
- **Zero** Communication Latency

semidynamics

What's inside a vector unit?



- 32 vector registers in RISC-V
- A number of "vector cores"
- A (wide) bus from/to the data cache

RISC-V
Vector 1.0
Compliant

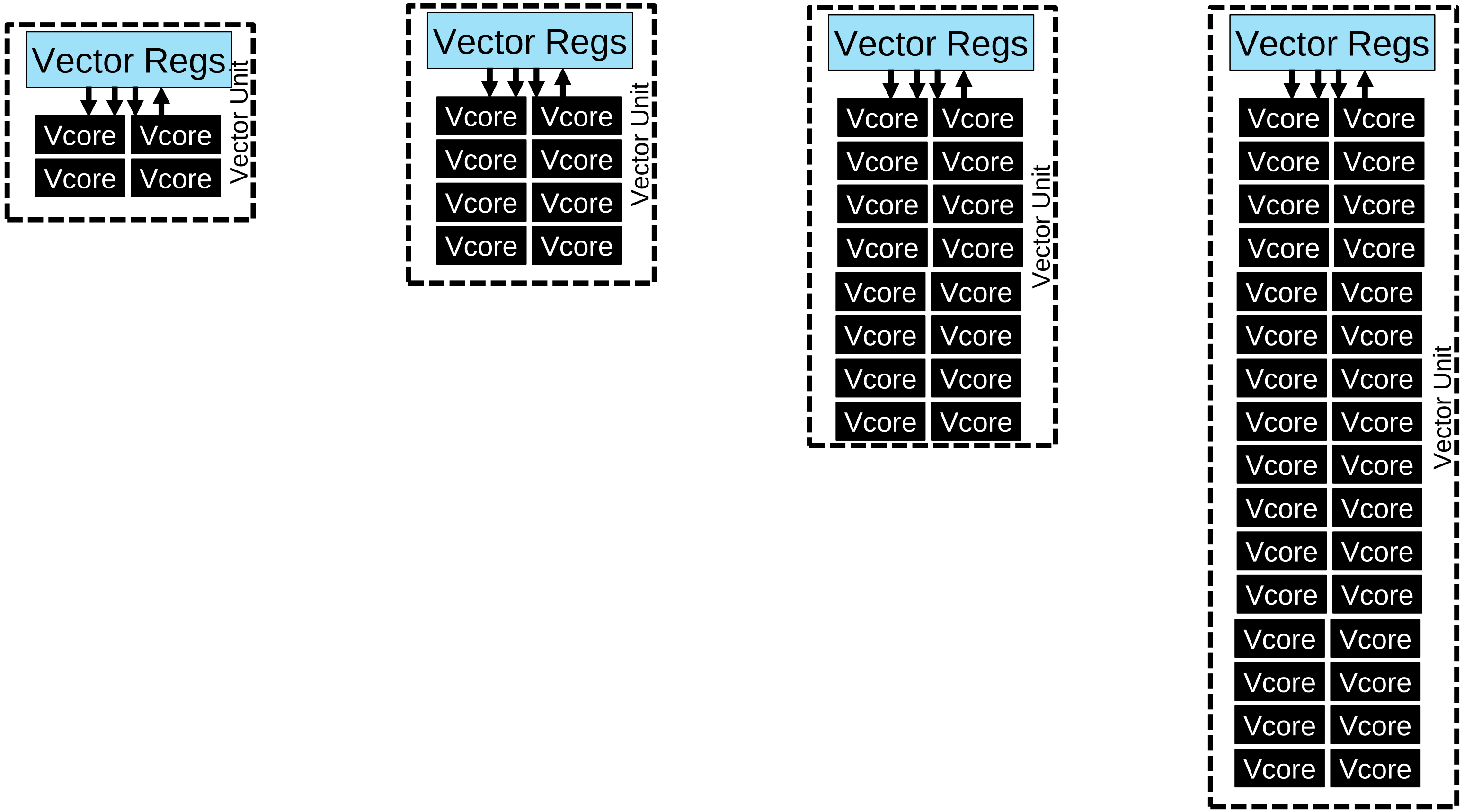
Semidynamic's Highly Configurable Vector Unit IP

3 Key Customization Options

semidynamics

Customization #1: Number of Vector Cores

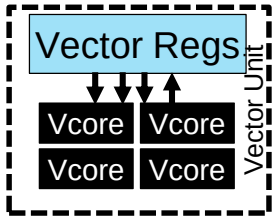
V4 | **V8** | **V16** | **V32**



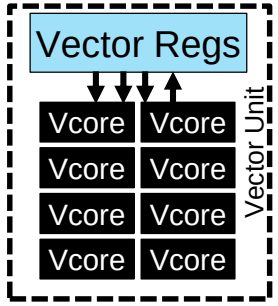
256b | 512b | 1024b | 2048b

Customization #2: Data Types

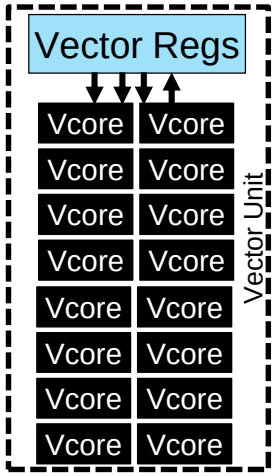
V4



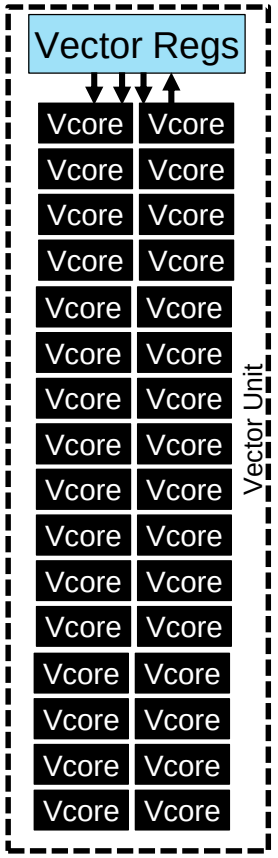
V8



V16



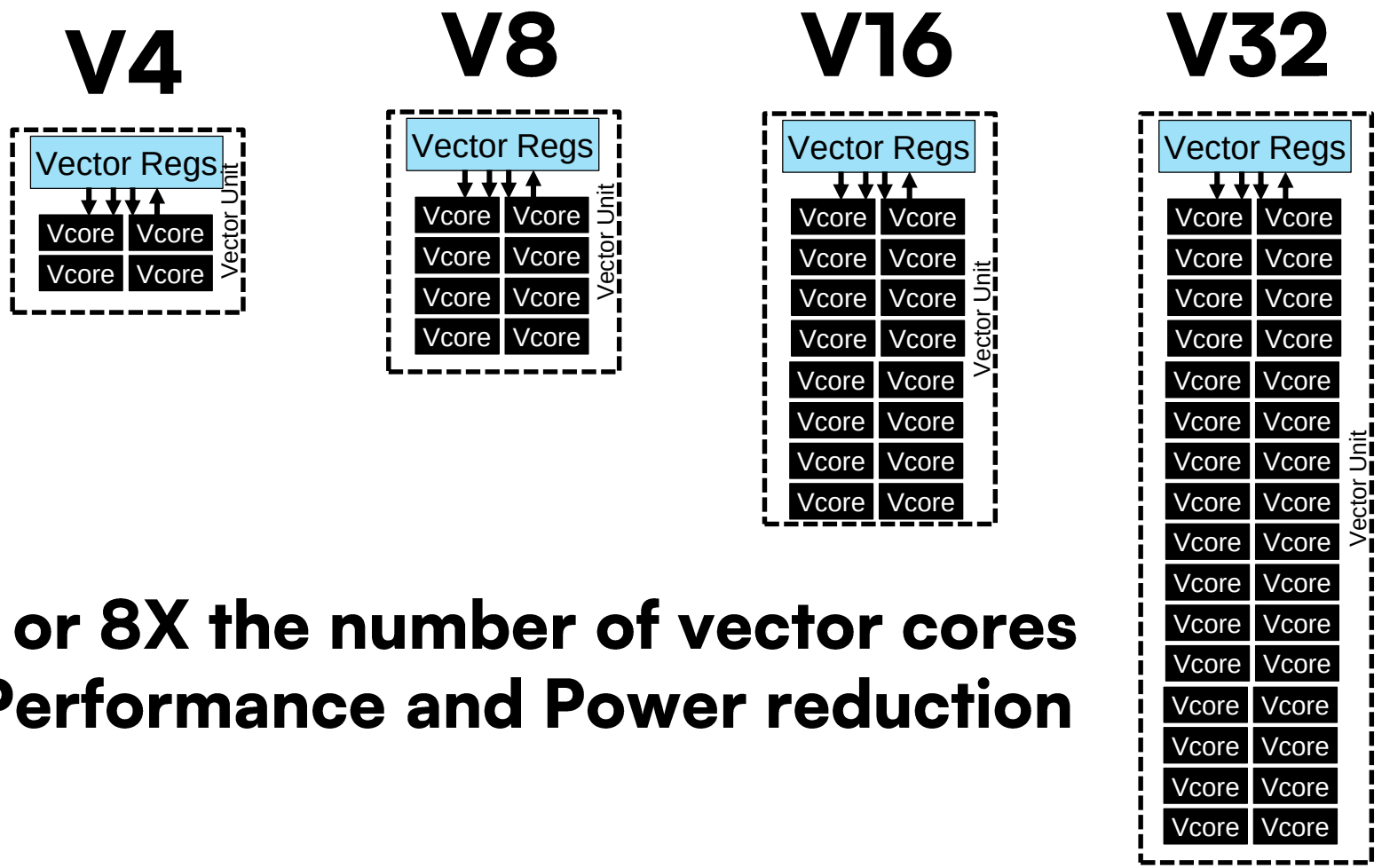
V32



FP64, FP32, FP16, BF16
INT64, INT32, INT16, INT8

semidynamics

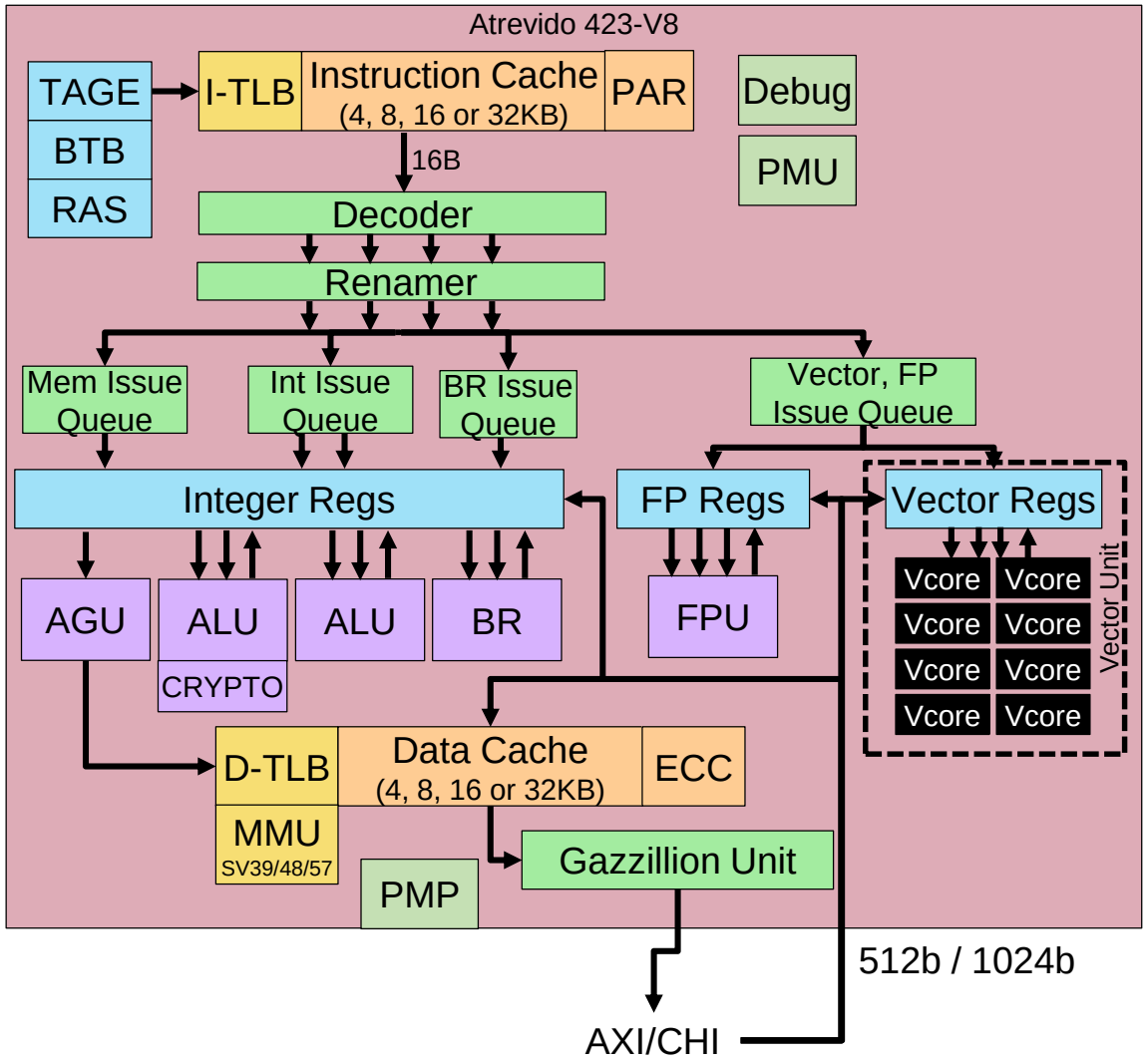
Customization #3: Vector Register Length



1X, 2X, 4X or 8X the number of vector cores
Great for Performance and Power reduction

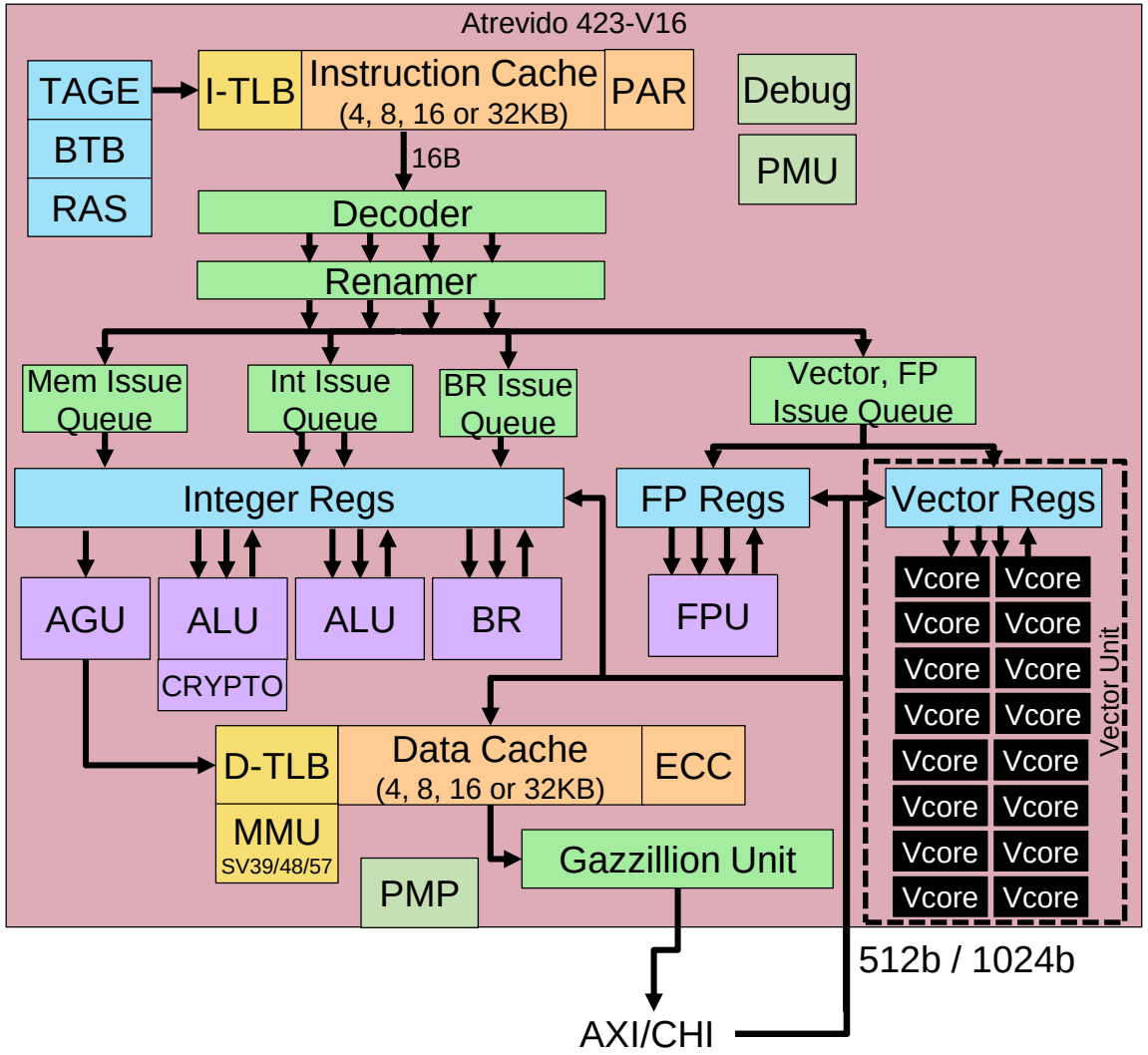
semidynamics

Vector Unit connection to the Core (V8)



semidynamics

Vector Unit connection to the Core (V16)



RISC-V
Vector 1.0
Compliant

Semidynamic's Three Key Vector Technologies

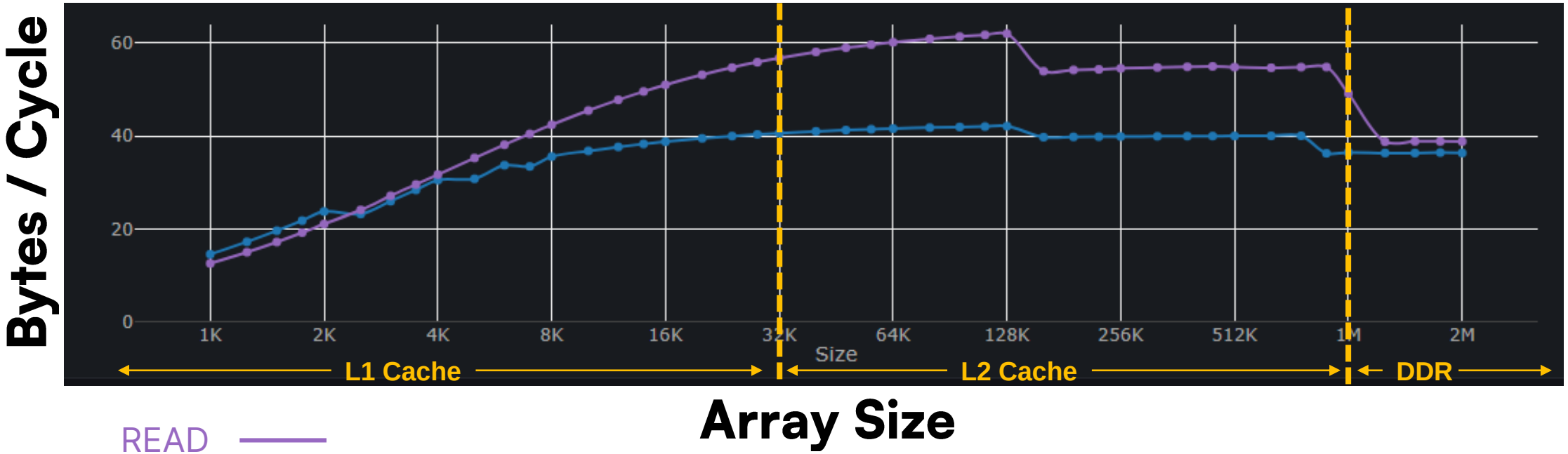
semidynamics

Key Tech #1: Full Out-of-order vectors

- Full Renaming of vector registers
- Full Renaming of the mask register
- Special treatment of LMUL > 1
- Special support for at-speed Tail Agnostic and Tail Undisturbed
- Special renaming for vrgather
- Fast cross-vcore network for
  - vslide, vrgather, vcompress, vexpand

Key Tech #2: Vector + Gazzillion: A bandwidth rocket!

Can you find a core out there capable of streaming data at over 60 Bytes/cycle? And from main DDR memory (not from your cache)? We don't think so ☺



8 vector cores, 32X vector length

semidynamics

Key Tech #3: Open Vector Interface

- If you want the vector unit…
- And you also want your custom logic bloc
  - DSP, AI, ML, secret block, … you name it

- We have a simple protocol to connect your logic to the vector unit

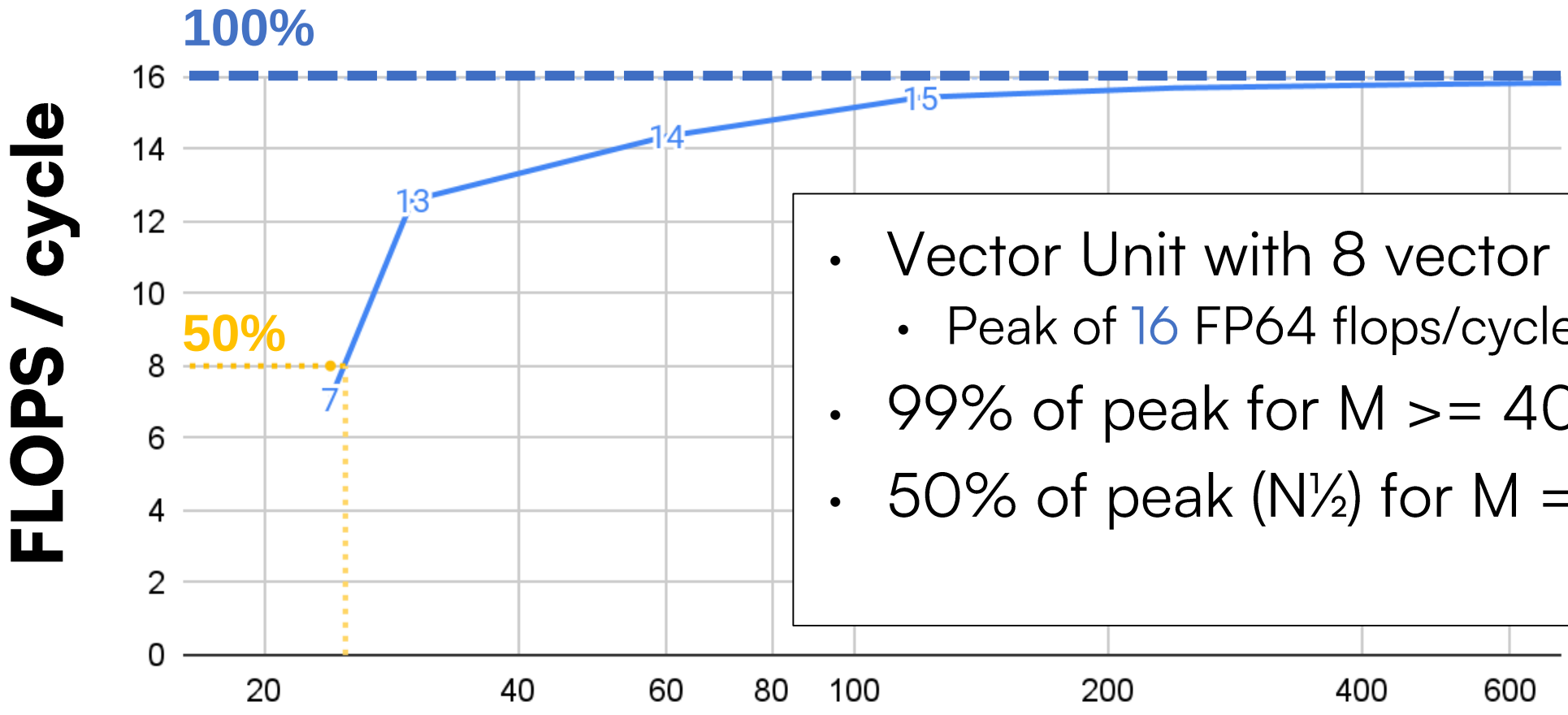
RISC-V
Vector 1.0
Compliant

Semidynamic's Vector Performance

semidynamics

DGEMM on OOO V8 Vector Unit

(FP64 matrix multiply)



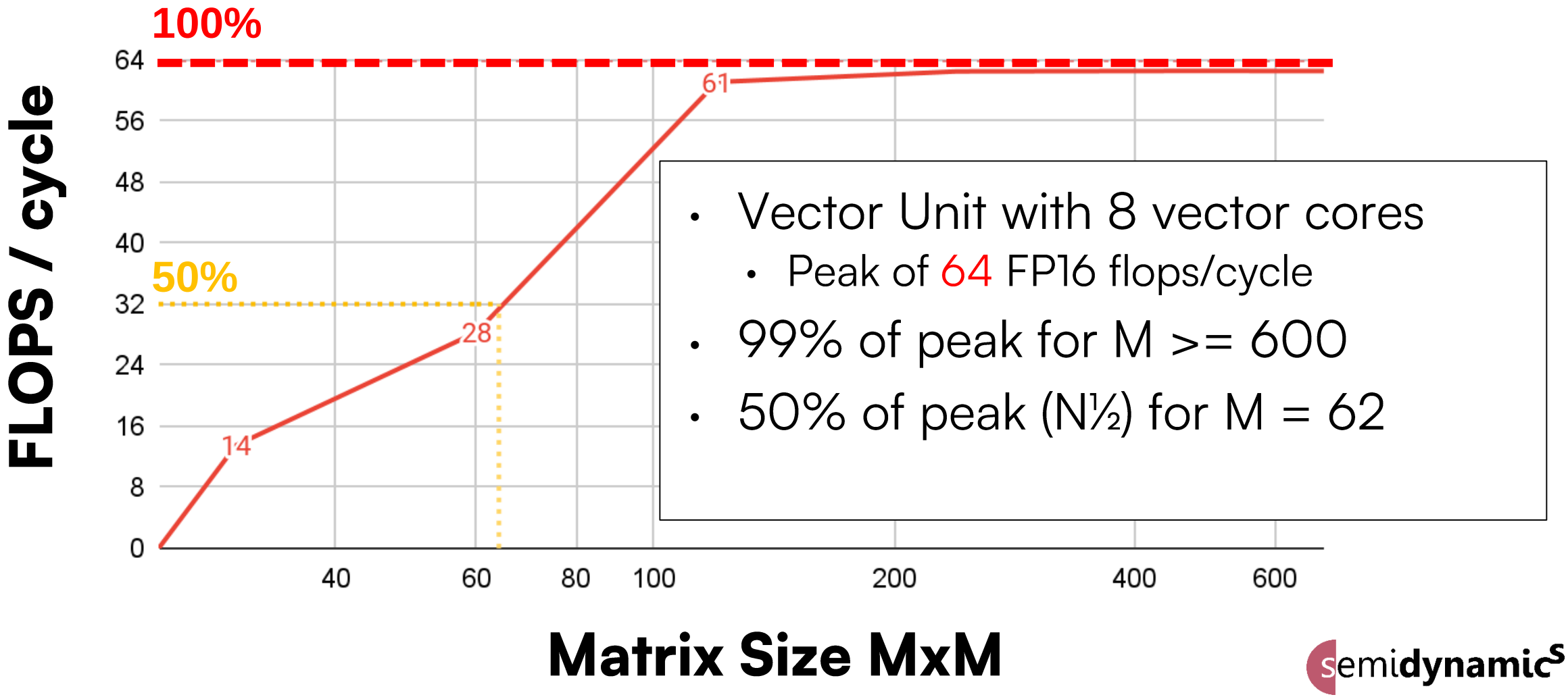
- Vector Unit with 8 vector cores
  - Peak of 16 FP64 flops/cycle
- 99% of peak for M >= 400
- 50% of peak (N½) for M = 24

Matrix Size MxM

semidynamics

HGEMM on OOO V8 Vector Unit

(FP16 matrix multiply)

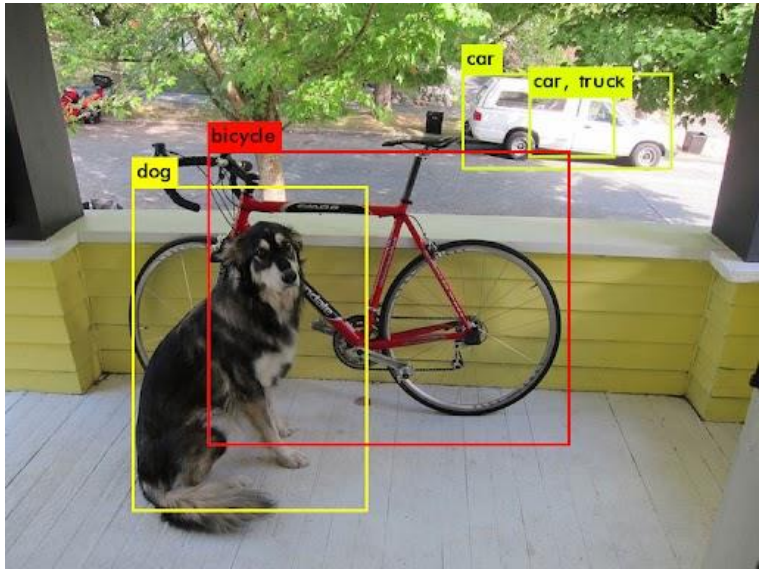


- Vector Unit with 8 vector cores
  - Peak of 64 FP16 flops/cycle
- 99% of peak for M >= 600
- 50% of peak (N½) for M = 62

semidynamics

Yolo on OOO V8 Vector Unit

- YOLOv3-tiny:
  - 24 layers, 5.56 Gops/frame, ~9M params
  - Using SGEMM (FP32) for Matrix Multiplication



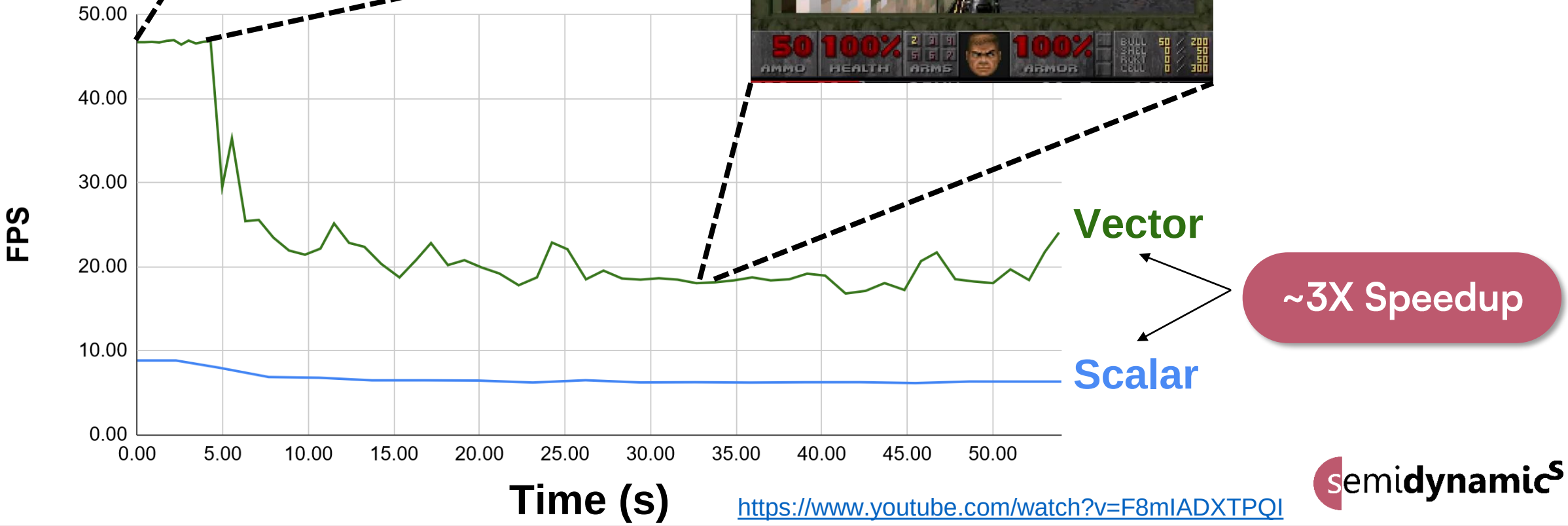
Platform	Vector/Cuda Cores	Frequency (Ghz)	FPS	FPS per 8 vector cores @ 1Ghz
Jetson TX2	256	1.30	19[1]	0.46
Jetson AGX Xavier	512	1.38	32[1]	0.36
GTX Titan X	3072	1.09	220[2]	0.53
Atrevido 423-V8	**8**	**1.00**	**0.84**	**0.84**

58% higher performance per vector core

semidynamics

[1] https://www.researchgate.net/publication/351347699_Real-Time_On_Board_Deep_Learning_Fault_Detection_for_Autonomous_UAV_Inspections
[2] https://pireddie.com/darknet/yolo/

Vectorized Doom on OOO V8 Vector Unit



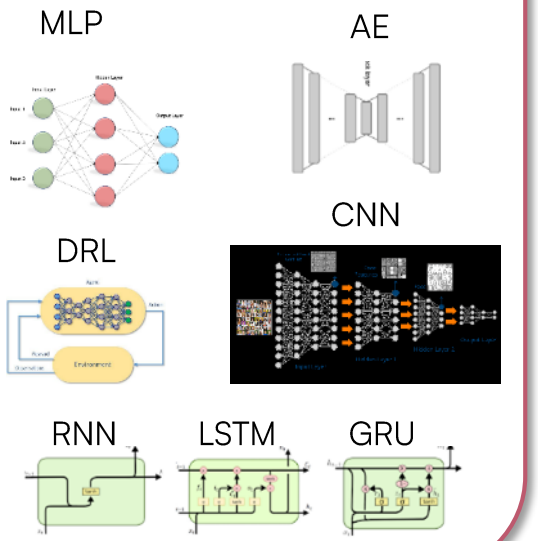
https://www.youtube.com/watch?v=F8mIADXTPQI

semidynamics

Our Vector Unit is designed for...

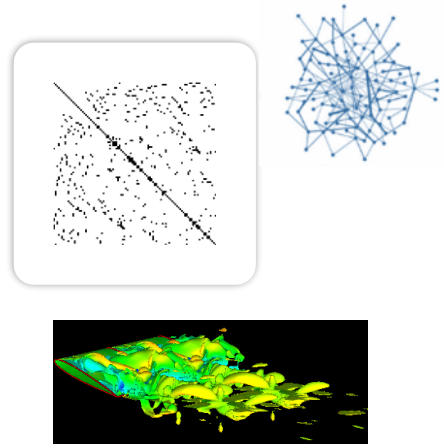
Machine Learning

Recommendation Systems



Key-Value Stores

Sparse Data/HPC



Ideal for moving and processing a lot of data, very fast

semidynamics

Teaser for RISC-V Summit November

Tensor Instructions coming soon

semidynamics

THANK YOU!

semidynamics