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
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.
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
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
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
Semidynamic’s Highly Configurable Vector Unit IP

3 Key Customization Options
Customization #1: Number of Vector Cores

- **V4**: 256b
- **V8**: 512b
- **V16**: 1024b
- **V32**: 2048b
Customization #2: Data Types

**V4**

- Vector Regs
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore

**V8**

- Vector Regs
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore

**V16**

- Vector Regs
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore

**V32**

- Vector Regs
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore
- Vcore Vcore

**FP64, FP32, FP16, BF16**

**INT64, INT32, INT16, INT8**
Customization #3: Vector Register Length

V4
Vector Regs
Vcore Vcore
Vcore Vcore

V8
Vector Regs
Vcore Vcore
Vcore Vcore
Vcore Vcore

V16
Vector Regs
Vcore Vcore
Vcore Vcore
Vcore Vcore
Vcore Vcore

V32
Vector Regs
Vcore Vcore
Vcore Vcore
Vcore Vcore
Vcore Vcore
Vcore Vcore
Vcore Vcore
Vcore Vcore

1X, 2X, 4X or 8X the number of vector cores
Great for Performance and Power reduction
Vector Unit connection to the Core (V8)
Vector Unit connection to the Core (V16)
Semidynamic’s Three Key Vector Technologies
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 😊

![Graph showing data transfer rates between cache sizes and DDR memory. The graph indicates a steep increase in bytes per cycle as the array size increases from L1 Cache to L2 Cache and then to DDR memory.]

8 vector cores, 32X vector length
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
Semidynamic’s Vector Performance
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 $M \geq 400$
- 50% of peak ($N^{1/2}$) for $M = 24$
**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 $M \geq 600$
- 50% of peak ($N^{1/2}$) for $M = 62$
Yolo on OOO V8 Vector Unit

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

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

58% higher performance per vector core

Vectorized Doom on OOO V8 Vector Unit

![Game Image]

![Game Play Image]

Vector
Scalar

~3X Speedup

https://www.youtube.com/watch?v=F8miADXTPOI
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
Teaser for RISC-V Summit November

Tensor Instructions coming soon
THANK YOU!