RISC-V "V" Vector Extension

Version 0.8-draft-20190906
# Table of Contents

1. Introduction
2. Implementation-defined Constant Parameters
3. Vector Extension Programmer’s Model
   3.1. Vector Registers
   3.2. Vector type register, vtype
   3.3. Vector Length Register v1
   3.4. Vector Start Index CSR vstart
   3.5. Vector Fixed-Point Rounding Mode Register vxrm
   3.6. Vector Fixed-Point Saturation Flag vxsat
   3.7. Vector Fixed-Point Fields in fcsr
   3.8. State of Vector Extension at Reset
4. Mapping of Vector Elements to Vector Register State
   4.1. Mapping with LMUL=1
   4.2. Mapping with LMUL > 1
   4.3. Mapping across Mixed-Width Operations
   4.4. Mask Register Layout
5. Vector Instruction Formats
   5.1. Scalar operands
   5.2. Vector Operands
   5.3. Vector Masking
   5.4. Prestart, Active, Inactive, Body, and Tail Element Definitions
6. Configuration-Setting Instructions
   6.1. vsetvl/vsetvl instructions
   6.2. Constraints on Setting v1
   6.3. vsetvl Instruction
   6.4. Examples
7. Vector Loads and Stores
   7.1. Vector Load/Store Instruction Encoding
   7.2. Vector Load/Store Addressing Modes
   7.3. Vector Load/Store Width Encoding
   7.4. Vector Unit-Stride Instructions
   7.5. Vector Strided Instructions
   7.6. Vector Indexed Instructions
   7.7. Unit-stride Fault-Only-First Loads
   7.8. Vector Load/Store Segment Instructions (Zv1seg)
8. Vector AMO Operations (Zvamo)
9. Vector Memory Alignment Constraints
10. Vector Memory Consistency Model
11. Vector Arithmetic Instruction Formats
    11.1. Vector Arithmetic Instruction encoding
    11.2. Widening Vector Arithmetic Instructions
    11.3. Narrowing Vector Arithmetic Instructions
12. Vector Integer Arithmetic Instructions
12.1. Vector Single-Width Integer Add and Subtract
12.2. Vector Widening Integer Add/Subtract
12.3. Vector Integer Add-with-Carry / Subtract-with-Borrow Instructions
12.4. Vector Bitwise Logical Instructions
12.5. Vector Single-Width Bit Shift Instructions
12.6. Vector Narrowing Integer Right Shift Instructions
12.7. Vector Integer Comparison Instructions
12.8. Vector Integer Min/Max Instructions
12.9. Vector Single-Width Integer Multiply Instructions
12.10. Vector Integer Divide Instructions
12.11. Vector Widening Integer Multiply Instructions
12.13. Vector Widening Integer Multiply-Add Instructions
12.14. Vector Integer Merge Instructions
12.15. Vector Integer Move Instructions

13. Vector Fixed-Point Arithmetic Instructions
13.1. Vector Single-Width Saturating Add and Subtract
13.2. Vector Single-Width Averaging Add and Subtract
13.3. Vector Single-Width Fractional Multiply with Rounding and Saturation
13.4. Vector Widening Saturating Scaled Multiply-Add
13.5. Vector Single-Width Scaling Shift Instructions
13.6. Vector Narrowing Fixed-Point Clip Instructions

14. Vector Floating-Point Instructions
14.1. Vector Floating-Point Exception Flags
14.2. Vector Single-Width Floating-Point Add/Subtract Instructions
14.3. Vector Widening Floating-Point Add/Subtract Instructions
14.4. Vector Single-Width Floating-Point Multiply/Divide Instructions
14.5. Vector Widening Floating-Point Multiply
14.6. Vector Single-Width Floating-Point Fused Multiply-Add Instructions
14.7. Vector Widening Floating-Point Fused Multiply-Add Instructions
14.8. Vector Floating-Point Square-Root Instruction
14.9. Vector Floating-Point MIN/MAX Instructions
14.10. Vector Floating-Point Sign-Injection Instructions
14.11. Vector Floating-Point Compare Instructions
14.12. Vector Floating-Point Classify Instruction
14.13. Vector Floating-Point Merge Instruction
14.14. Vector Floating-Point Move Instruction
14.15. Single-Width Floating-Point/Integer Type-Convert Instructions
14.16. Widening Floating-Point/Integer Type-Convert Instructions
14.17. Narrowing Floating-Point/Integer Type-Convert Instructions

15. Vector Reduction Operations
15.1. Vector Single-Width Integer Reduction Instructions
15.2. Vector Widening Integer Reduction Instructions
15.3. Vector Single-Width Floating-Point Reduction Instructions
15.4. Vector Widening Floating-Point Reduction Instructions

16. Vector Mask Instructions
16.1. Vector Mask-Register Logical Instructions
16.2. Vector mask population count \texttt{vpopc}
16.3. \texttt{vfirts} find-first-set mask bit
16.4. \texttt{vmsbf.m} set-before-first mask bit
16.5. \texttt{vmsif.m} set-including-first mask bit
16.6. \texttt{vmsof.m} set-only-first mask bit
16.7. Example using vector mask instructions
16.8. Vector Iota Instruction
16.9. Vector Element Index Instruction

17. Vector Permutation Instructions
17.1. Integer Scalar Move Instructions
17.2. Floating-Point Scalar Move Instructions
17.3. Vector Slide Instructions
17.4. Vector Register Gather Instruction
17.5. Vector Compress Instruction

18. Exception Handling
18.1. Precise vector traps
18.2. Imprecise vector traps
18.3. Selectable precise/imprecise traps
18.4. Swappable traps

19. Divided Element Extension (‘Zvediv’)
19.1. Instructions not affected by EDIV
19.2. Instructions Affected by EDIV
19.3. Vector Integer Dot-Product Instruction
19.4. Vector Floating-Point Dot Product Instruction

20. Vector Instruction Listing

Appendix A: Vector Assembly Code Examples
A.1. Vector-vector add example
A.2. Example with mixed-width mask and compute.
A.3. Memcpy example
A.4. Conditional example
A.5. SAXPY example
A.6. SGEMM example


Known issues with current version:

- encoding needs better formatting
- vector memory consistency model needs to be clarified
- interaction with privileged architectures
1. Introduction

This document describes the draft of the RISC-V base vector extension. The document describes all the individual features of the base vector extension.

This is a draft of a stable proposal for the vector specification to be used for implementation and evaluation. Once the draft label is removed, version 0.7 is intended to be stable enough to begin developing toolchains, functional simulators, and initial implementations, though will continue to evolve with minor changes and updates.

The term *base vector extension* is used informally to describe the standard set of vector ISA components. This draft spec is intended to capture how a certain vector function will be implemented as vector instructions, but to not yet determine what set of vector instructions are mandatory for a given platform.

Each actual platform profile will formally specify the mandatory components of any vector extension adopted by that platform. The base vector extension can expected to be close to that which will eventually be used in the standard Unix platform profile that supports vectors. Other platforms, including embedded platforms, may choose to implement subsets of these extensions. The exact set of mandatory supported instructions for an implementation to be compliant with a given profile is subject to change until each profile spec is ratified.

The base vector extension is designed to act as a base for additional vector extensions in various domains, including cryptography and machine learning.
2. Implementation-defined Constant Parameters

Each hart supporting the vector extension defines three parameters:

1. The maximum size of a single vector element in bits, $ELEN$, which must be a power of 2.

2. The number of bits in a vector register, $VLEN \geq ELEN$, which must be a power of 2.

3. The striping distance in bits, $SLEN$, which must be $VLEN \geq SLEN \geq 32$, and which must be a power of 2.

Platform profiles may set further constraints on these parameters, for example, requiring that $ELEN \geq \max(XLEN,FLEN)$, or requiring a minimum $VLEN$ value, or setting an $SLEN$ value.

The ISA supports writing binary code that under certain constraints will execute portably on harts with different values for these parameters.

- Code can be written that will expose differences in implementation parameters.
- Thread contexts with active vector state cannot be migrated during execution between harts that have any difference in $VLEN$, $ELEN$, or $SLEN$ parameters.
3. Vector Extension Programmer’s Model

The vector extension adds 32 vector registers, and five unprivileged CSRs (vstart, vxsat, vxrm, vtype, vl) to a base scalar RISC-V ISA. If the base scalar ISA does not include floating-point, then a fcsr register is also added to hold mirrors of the vxsat and vxrm CSRs as explained below.

<table>
<thead>
<tr>
<th>Address</th>
<th>Privilege</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x008</td>
<td>URW</td>
<td>vstart</td>
<td>Vector start position</td>
</tr>
<tr>
<td>0x009</td>
<td>URW</td>
<td>vxsat</td>
<td>Fixed-Point Saturate Flag</td>
</tr>
<tr>
<td>0x00A</td>
<td>URW</td>
<td>vxrm</td>
<td>Fixed-Point Rounding Mode</td>
</tr>
<tr>
<td>0xC20</td>
<td>URO</td>
<td>vl</td>
<td>Vector length</td>
</tr>
<tr>
<td>0xC21</td>
<td>URO</td>
<td>vtype</td>
<td>Vector data type register</td>
</tr>
</tbody>
</table>

3.1. Vector Registers

The vector extension adds 32 architectural vector registers, v0-v31 to the base scalar RISC-V ISA.

Each vector register has a fixed VLEN bits of state.

Zfinx (“F in X”) is a new ISA option under consideration where floating-point instructions take their arguments from the integer register file. The 0.7 vector extension is also compatible with this option.

3.2. Vector type register, vtype

The read-only XLEN-wide vector type CSR, vtype provides the default type used to interpret the contents of the vector register file, and can only be updated by vsetvl{i} instructions. The vector type also determines the organization of elements in each vector register, and how multiple vector registers are grouped.

Earlier drafts allowed the vtype register to be written using regular CSR writes. Allowing updates only via the vsetvl{i} instructions simplifies maintenance of the vtype register state.

In the base vector extension, the vtype register has three fields, vill, vsew[2:0], and vlmul[1:0].

<table>
<thead>
<tr>
<th>Bits</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLEN-1</td>
<td>vill</td>
<td>Illegal value if set</td>
</tr>
<tr>
<td>XLEN-2:7</td>
<td>Reserved (write 0)</td>
<td></td>
</tr>
<tr>
<td>6:5</td>
<td>vediv[1:0]</td>
<td>Used by EDIV extension</td>
</tr>
<tr>
<td>4:2</td>
<td>vsew[2:0]</td>
<td>Standard element width (SEW) setting</td>
</tr>
<tr>
<td>1:0</td>
<td>vlmul[1:0]</td>
<td>Vector register group multiplier (LMUL) setting</td>
</tr>
</tbody>
</table>

The smallest base implementation requires storage for only four bits of storage in vtype, two bits for vsew[1:0] and two bits for vlmul[1:0]. The illegal value represented by vill can be encoded using the illegal 64-bit combination in vsew[1:0] without requiring an additional storage bit.

The vediv[1:0] field is used by the EDIV extension described below.

Further standard and custom extensions to the vector base will extend these fields to support a greater variety of data types.

It is anticipated that an extended 64-bit instruction encoding would allow these fields to be specified statically in the instruction encoding.

3.2.1. Vector standard element width vsew
The value in vsew sets the dynamic standard element width (SEW). By default, a vector register is viewed as being divided into VLEN / SEW standard-width elements. In the base vector extension, only SEW up to max(XLEN,FLEN) are required to be supported.

Table 3. vsew[2:0] (standard element width) encoding

<table>
<thead>
<tr>
<th>vsew[2:0]</th>
<th>SEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0</td>
</tr>
<tr>
<td>00</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>0</td>
</tr>
<tr>
<td>01</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4. Example VLEN = 128 bits

<table>
<thead>
<tr>
<th>SEW</th>
<th>Elements per vector register</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

3.2.2. Vector Register Grouping (vlmul)

Multiple vector registers can be grouped together to form a vector register group, so that a single vector instruction can operate on multiple vector registers. Vector register groups allow double-width or larger elements to be operated on with the same vector length as standard-width elements. Vector register groups also provide greater execution efficiency for longer application vectors.

The number of vector registers in a group, LMUL, is an integer power of two set by the vmlul field in vtype (LMUL = 2^vlmul[1:0]).

The derived value VLMAX = LMUL*VLEN/SEW represents the maximum number of elements that can be operated on with a single vector instruction given the current SEW and LMUL settings.

<table>
<thead>
<tr>
<th>vlmul</th>
<th>LMUL</th>
<th>#groups</th>
<th>VLMAX</th>
<th>Grouped registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0</td>
<td>1</td>
<td>VLEN/SEW</td>
<td>vn (no group)</td>
</tr>
<tr>
<td>01</td>
<td>1</td>
<td>2</td>
<td>2*VLEN/SEW</td>
<td>vn, vn+1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>4</td>
<td>4*VLEN/SEW</td>
<td>vn, ..., vn+3</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>8</td>
<td>8*VLEN/SEW</td>
<td>vn, ..., vn+7</td>
</tr>
</tbody>
</table>

When vlmul=01, then vector operations on register v n also operate on vector register v n+1, giving twice the vector length in bits. Instructions specifying a vector operand with an odd-numbered vector register will raise an illegal instruction exception.

Similarly, when vlmul=10, vector instructions operate on four vector registers at a time, and instructions specifying vector operands using vector register numbers that are not multiples of four will raise an illegal instruction exception. When vlmul=11, operations operate on eight vector registers at a time, and instructions specifying vector operands using register numbers that are not multiples of eight will raise an illegal instruction exception.
This grouping pattern (LMUL=8 has groups v0, v8, v16, v24) was adopted in 0.6 initially to avoid issues with the floating-point calling convention when floating-point values were overlaid on the vector registers, whereas earlier versions kept the vector register group names contiguous (LMUL=8 has groups v0, v1, v2, v3). In versions 0.7 onwards, the floating-point registers are separate again.

Mask register instructions always operate on a single vector register, regardless of LMUL setting.

### 3.2.3. Vector Type Illegal vill

The vill bit is used to encode that a previous vsetvl{i} instruction attempted to write an unsupported value to vtype.

The vill bit is held in bit XLEN-1 of the CSR to support checking for illegal values with a branch on the sign bit.

If the vill bit is set, then any attempt to execute a vector instruction (other than a vector configuration instruction) will raise an illegal instruction exception.

When the vill bit is set, the other XLEN-1 bits in vtype shall be zero.

### 3.3. Vector Length Register v1

The XLEN-bit-wide read-only v1 CSR can only be updated by the vsetvli and vsetvl instructions, and the fault-only-first vector load instruction variants.

The v1 register holds an unsigned integer specifying the number of elements to be updated by a vector instruction. Elements in any destination vector register group with indices ≥ v1 are unmodified during execution of a vector instruction. When vstart ≥ v1, no elements are updated in any destination vector register group.

As a consequence, when v1=0, no elements are updated in the destination vector register group, regardless of vstart.

Instructions that write a scalar integer or floating-point register do so even when vstart ≥ v1.

The number of bits implemented in v1 depends on the implementation’s maximum vector length of the smallest supported type. The smallest vector implementation, RV32IV, would need at least six bits in v1 to hold the values 0-32 (with VLEN=32, LMUL=8 and SEW=8 results in VLMAX of 32).

### 3.4. Vector Start Index CSR vstart

The vstart read-write CSR specifies the index of the first element to be executed by a vector instruction.

Normally, vstart is only written by hardware on a trap on a vector instruction, with the vstart value representing the element on which the trap was taken (either a synchronous exception or an asynchronous interrupt), and at which execution should resume after a resumable trap is handled.

All vector instructions are defined to begin execution with the element number given in the vstart CSR, leaving earlier elements in the destination vector undisturbed, and to reset the vstart CSR to zero at the end of execution.

All vector instructions, including vsetvl{i}, reset the vstart CSR to zero.

If the value in the vstart register is greater than or equal to the vector length v1 then no element operations are performed. The vstart register is then reset to zero.

The vstart CSR is defined to have only enough writable bits to hold the largest element index (one less than the maximum VLMAX) or \( \lg_2(\text{VLEN}) \) bits. The upper bits of the vstart CSR are hardwired to zero (reads zero, writes ignored).
The maximum vector length is obtained with the largest LMUL setting (8) and the smallest SEW setting (8), so VLMAX_max = 8*VLEN/8 = VLEN. For example, for VLEN=256, vstart would have 8 bits to represent indices from 0 through 255.

The vstart CSR is writable by unprivileged code, but non-zero vstart values may cause vector instructions to run substantially slower on some implementations, so vstart should not be used by application programmers. A few vector instructions can not be executed with a non-zero vstart value and will raise an illegal instruction exception as defined below.

Implementations are permitted to raise illegal instruction exceptions when attempting to execute a vector instruction with a value of vstart that the implementation can never produce when executing that same instruction with the same vtype setting.

For example, some implementations will never take interrupts during execution of a vector arithmetic instruction, instead waiting until the instruction completes to take the interrupt. Such implementations are permitted to raise an illegal instruction exception when attempting to execute a vector arithmetic instruction when vstart is nonzero.

### 3.5. Vector Fixed-Point Rounding Mode Register vxrm

The vector fixed-point rounding-mode register holds a two-bit read-write rounding-mode field. The vector fixed-point rounding-mode is given a separate CSR address to allow independent access, but is also reflected as a field in the upper bits of fcsr. Systems without floating-point must add fcsr when adding the vector extension.

The fixed-point rounding algorithm is specified as follows. Suppose the pre-rounding result is v, and d bits of that result are to be rounded off. Then the rounded result is \((v >> d) + r\), where \(r\) depends on the rounding mode as specified in the following table.

<table>
<thead>
<tr>
<th>Bits [1:0]</th>
<th>Abbreviation</th>
<th>Rounding Mode</th>
<th>Rounding increment, (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>rnu</td>
<td>round-to-nearest-up (add +0.5 LSB)</td>
<td>(v[d-1])</td>
</tr>
<tr>
<td>0 1</td>
<td>rne</td>
<td>round-to-nearest-even</td>
<td>(v[d-1] &amp; (v[d-2:0]\neq0</td>
</tr>
<tr>
<td>1 0</td>
<td>rdn</td>
<td>round-down (truncate)</td>
<td>0</td>
</tr>
<tr>
<td>1 1</td>
<td>rod</td>
<td>round-to-odd (OR bits into LSB, aka “jam”)</td>
<td>!v[d] &amp; v[d-1:0]\neq0</td>
</tr>
</tbody>
</table>

Bits[XLEN-1:2] should be written as zeros.

The rounding mode can be set with a single csrwi instruction.

### 3.6. Vector Fixed-Point Saturation Flag vxsat

The vxsat CSR holds a single read-write bit that indicates if a fixed-point instruction has had to saturate an output value to fit into a destination format.

The vxsat bit is mirrored in the upper bits of fcsr.

### 3.7. Vector Fixed-Point Fields in fcsr

The vxrm and vxsat separate CSRs can also be accessed via fields in the floating-point CSR, fcsr. The fcsr register must be added to systems without floating-point that add a vector extension.
### Table 6. fcsr layout

<table>
<thead>
<tr>
<th>Bits</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:9</td>
<td>vxrm</td>
<td>Fixed-point rounding mode</td>
</tr>
<tr>
<td>8</td>
<td>vxsat</td>
<td>Fixed-point accrued saturation flag</td>
</tr>
<tr>
<td>7:5</td>
<td>frm</td>
<td>Floating-point rounding mode</td>
</tr>
<tr>
<td>4:0</td>
<td>flags</td>
<td>Floating-point accrued exception flags</td>
</tr>
</tbody>
</table>

The fields are packed into fcsr to make context-save/restore faster.

### 3.8. State of Vector Extension at Reset

The vector extension must have a consistent state at reset. In particular, vtype and vl must have values that can be read and then restored with a single vsetvl instruction.

- It is recommended that at reset, vtype.vill is set, the remaining bits in vtype are zero, and vl is set to zero.

The vstart, vxrm, vxsat CSRs can have arbitrary values at reset.

- Any use of the vector unit will require an initial vsetvl{1}, which will reset vstart. The vxrm and vxsat fields should be reset explicitly in software before use.

The vector registers can have arbitrary values at reset.
4. Mapping of Vector Elements to Vector Register State

The following diagrams illustrate how different width elements are packed into the bytes of a vector register depending on the current SEW and LMUL settings, as well as implementation ELEN and VLEN. Elements are packed into each vector register with the least-significant byte in the lowest-numbered bits.

Previous RISC-V vector proposals (< 0.6) hid this mapping from software, whereas this proposal has a specific mapping for all configurations, which reduces implementation flexibility but removes need for zeroing on config changes. Making the mapping explicit also has the advantage of simplifying oblivious context save-restore code, as the code can save the configuration in \( v1 \) and \( vtype \), then reset \( vtype \) to a convenient value (e.g., four vector groups of \( \text{LMUL}=8, \text{SEW}=\text{ELEN} \)) before saving all vector register bits without needing to parse the configuration. The reverse process will restore the state.

4.1. Mapping with LMUL=1

When LMUL=1, elements are simply packed in order from the least-significant to most-significant bits of the vector register.

To increase readability, vector register layouts are drawn with bytes ordered from right to left with increasing byte address. Bits within an element are numbered in a little-endian format with increasing bit index from right to left corresponding to increasing magnitude.
The element index is given in hexadecimal and is shown placed at the least-significant byte.

### VLEN=32b

Byte  3 2 1 0  
SEW=8b  3 2 1 0  
SEW=16b  1 0  
SEW=32b  0  

### VLEN=64b

Byte  7 6 5 4 3 2 1 0  
SEW=8b  7 6 5 4 3 2 1 0  
SEW=16b  3 2 1 0  
SEW=32b  1 0  
SEW=64b  0  

### VLEN=128b

Byte  F E D C B A 9 8 7 6 5 4 3 2 1 0  
SEW=8b  F E D C B A 9 8 7 6 5 4 3 2 1 0  
SEW=16b  7 6 5 4 3 2 1 0  
SEW=32b  3 2 1 0  
SEW=64b  1 0  
SEW=128b  0  

### VLEN=256b

Byte  1F1E1D1C1B1A19181716151413121110  F E D C B A 9 8 7 6 5 4 3 2 1 0  
SEW=8b  1F1E1D1C1B1A19181716151413121110  F E D C B A 9 8 7 6 5 4 3 2 1 0  
SEW=16b  7 6 5 4 3 2 1 0  
SEW=32b  3 2 1 0  
SEW=64b  1 0  
SEW=128b  0  

### 4.2. Mapping with LMUL > 1

When vector registers are grouped, the elements of the vector register group are striped across the constituent vector registers. The striping distance in bits, SLEN, sets how many bits are packed contiguously into one vector register before moving to the next in the group.

For example, when SLEN = 128, the striping pattern is repeated in multiples of 128 bits. The first 128/SEW elements are packed contiguously at the start of the first vector register in the group. The next 128/SEW elements are packed contiguously at the start of the next vector register in the group. After packing the first LMUL*128/SEW elements at the start of each of the LMUL vector registers in the group, the second LMUL*128/SEW group of elements are packed into the second 128b segment of each of the vector registers in the group, and so on.
Example 1: VLEN=32b, SEW=16b, LMUL=2

<table>
<thead>
<tr>
<th>Byte</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v2+n</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v2+n+1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 2: VLEN=64b, SEW=32b, LMUL=2

<table>
<thead>
<tr>
<th>Byte</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v2+n</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v2+n+1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 3: VLEN=128b, SEW=32b, LMUL=2

| Byte | F | E | D | C | B | A | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| v2+n | 3 | 2 | 1 | 0 |
| v2+n+1 | 7 | 6 | 5 | 4 |

Example 4: VLEN=256b, SEW=32b, LMUL=2

| Byte | 1F1E1D1C1B1A191B1716151413121110 | F | E | D | C | B | A | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|------|----------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| v2+n | B | A | 9 | 8 | 3 | 2 | 1 | 0 |
| v2+n+1 | F | E | D | C | 7 | 6 | 5 | 4 |

If SEW > SLEN, the striping pattern places one element in each vector register in the group before moving to the next vector register in the group. So, when LMUL=2, the even-numbered vector register contains the even-numbered elements of the vector and the odd-numbered vector register contains the odd-numbered elements of the vector.

In most implementations, the striping distance SLEN ≥ ELEN.

Example: VLEN=256b, SEW=256b, LMUL=2

| Byte | 1F1E1D1C1B1A191B1716151413121110 | F | E | D | C | B | A | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|------|----------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| v2+n | 0 |
| v2+n+1 | 1 |

When LMUL = 4, four vector registers hold elements as shown:
Example 1: VLEN=32b, SLEN=32b, SEW=16b, LMUL=4

<table>
<thead>
<tr>
<th>Byte</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v4*n</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+2</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+3</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 2: VLEN=64b, SLEN=64b, SEW=32b, LMUL=4

<table>
<thead>
<tr>
<th>Byte</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v4*n</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+2</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+3</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 3: VLEN=128b, SLEN=64b, SEW=32b, LMUL=4

<table>
<thead>
<tr>
<th>Byte</th>
<th>F</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v4*n</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td>32b elements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+1</td>
<td>B</td>
<td>A</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+2</td>
<td>D</td>
<td>C</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+3</td>
<td>F</td>
<td>E</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 4: VLEN=128b, SLEN=128b, SEW=32b, LMUL=4

<table>
<thead>
<tr>
<th>Byte</th>
<th>F</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v4*n</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>32b elements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+1</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+2</td>
<td>B</td>
<td>A</td>
<td>9</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+3</td>
<td>F</td>
<td>E</td>
<td>D</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 5: VLEN=256b, SLEN=128b, SEW=32b, LMUL=4

<table>
<thead>
<tr>
<th>Byte</th>
<th>1F1E1D1C1B1A1918171615141312110</th>
<th>F</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v4*n</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+1</td>
<td>17</td>
<td>16</td>
<td>15</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+2</td>
<td>1B</td>
<td>1A</td>
<td>19</td>
<td>18</td>
<td>B</td>
<td>A</td>
<td>9</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+3</td>
<td>1F</td>
<td>1E</td>
<td>1D</td>
<td>1C</td>
<td>F</td>
<td>E</td>
<td>D</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 6: VLEN=256b, SLEN=128b, SEW=256b, LMUL=4

<table>
<thead>
<tr>
<th>Byte</th>
<th>1F1E1D1C1B1A19181716151413121110</th>
<th>F</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>v4*n</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+2</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v4*n+3</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A similar pattern is followed for LMUL = 8.
Different striping patterns are architecturally visible, but software can be written that produces the same results regardless of striping pattern. The primary constraint is to not change the LMUL used to access values held in a vector register group (i.e., do not read values with a different LMUL than used to write values to the group).

The striping length SLEN for an implementation is set to optimize the tradeoff between datapath wiring for mixed-width operations and buffering needed to corner-turn wide vector unit-stride memory accesses into parallel accesses for the vector register file.

The previous explicit configuration design (version < 0.6) allowed these tradeoffs to be managed at the microarchitectural level and optimized for each configuration.

### 4.3. Mapping across Mixed-Width Operations

The pattern used to map elements within a vector register group is designed to reduce datapath wiring when supporting operations across multiple element widths. The recommended software strategy in this case is to modify \( vtype \) dynamically to keep SEW/LMUL constant (and hence VLMAX constant).

The following example shows four different packed element widths (8b, 16b, 32b, 64b) in a VLEN=256b/SLEN=128b implementation. The vector register grouping factor (LMUL) is increased by the relative element size such that each group can hold the same number of vector elements (32 in this example) to simplify stripmining code. Any operation between elements with the same index only touches operand bits located within the same 128b portion of the datapath.
VLEN=256b, SLEN=128b

Byte  

1F1E1D1C1B1A19181716151413121110 F E D C B A 9 8 7 6 5 4 3 2 1 0

SEW=8b, LMUL=1, VLMAX=32

v1  

1F1E1D1C1B1A19181716151413121110 F E D C B A 9 8 7 6 5 4 3 2 1 0

SEW=16b, LMUL=2, VLMAX=32

v2*n  

17 16 15 14 13 12 11 10 7 6 5 4 3 2 1 0

v2*n+1  

1F 1E 1D 1C 1B 1A 19 18 F E D C B A 9 8

SEW=32b, LMUL=4, VLMAX=32

v4*n  

13 12 11 10 3 2 1 0

v4*n+1  

17 16 15 14 7 6 5 4

v4*n+2  

1B 1A 19 18 B A 9 8

v4*n+3  

1F 1E 1D 1C F E D C

SEW=64b, LMUL=8, VLMAX=32

v8*n  

11 10 1 0

v8*n+1  

13 12 3 2

v8*n+2  

15 14 5 4

v8*n+3  

17 16 7 6

v8*n+4  

19 18 9 8

v8*n+5  

1B 1A B A

v8*n+6  

1D 1C D C

v8*n+7  

1F 1E F E

Larger LMUL values can cause lower datapath utilization for short vectors if SLEN is less than the spatial datapath width. In the example above with VLEN=256b, SLEN=128b, and LMUL=8, if the implementation is purely spatial with a 256b-wide vector datapath, then for an application vector length less than 17, only half of the datapath will be active. The vsetvl instructions below could have a facility added to dynamically select an appropriate LMUL according to the required application vector length (AVL) and range of element widths.

The following table shows each possible constant SEW/LMUL operating point for loops with mixed-width operations.

<table>
<thead>
<tr>
<th>SEW/LMUL</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEW</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Larger LMUL values can also be used to simply increase vector length to reduce instruction fetch and dispatch overheads, in cases where fewer logical vector registers are required.
Narrower machines will set SLEN to be at least as large as the datapath spatial width, so there is no need to reduce LMUL. Wider machines might set SLEN lower than the spatial datapath width to reduce wiring for mixed-width operations (e.g., width=1024, ELEN=32, SLEN=128), in which case optimizing LMUL will be important.

4.4. Mask Register Layout

A vector mask occupies only one vector register regardless of SEW and LMUL. The mask bits that are used for each vector operation depends on the current SEW and LMUL setting.

The maximum number of elements in a vector operand is:

\[
V\text{LMAX} = \text{LMUL} \times \frac{\text{VLEN}}{\text{SEW}}
\]

A mask is allocated for each element by dividing the mask register into VLEN/VLMAX fields. The size of each mask element in bits, \(M\text{LEN}\), is:

\[
M\text{LEN} = \frac{\text{VLEN}}{\text{VLMAX}}
= \frac{\text{VLEN}}{(\text{LMUL} \times \frac{\text{VLEN}}{\text{SEW}})}
= \frac{\text{SEW}}{\text{LMUL}}
\]

The size of MLEN varies from ELEN (SEW=ELEN, LMUL=1) down to 1 (SEW=8b, LMUL=8), and hence a single vector register can always hold the entire mask register.

The mask bits for element \(i\) are located in bits \(\lfloor M\text{LEN} \times i + (M\text{LEN}-1) \rfloor : M\text{LEN} \times i \rfloor\) of the mask register. When a mask element is written by a compare instruction, the low bit in the mask element is written with the compare result and the upper bits of the mask element are zeroed. Destination mask elements past the end of the current vector length are unchanged. When a value is read as a mask, only the least-significant bit of the mask element is used to control masking and the upper bits are ignored.

The pattern is such that for constant SEW/LMUL values, the effective predicate bits are located in the same bit of the mask vector register, which simplifies use of masking in loops with mixed-width elements.

VLEN=32b

<table>
<thead>
<tr>
<th>Byte</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMUL=1, SEW=8b</td>
<td>[24][16][08][00] Mask bit position in decimal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>Element</td>
</tr>
</tbody>
</table>

LMUL=2, SEW=16b

<table>
<thead>
<tr>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>[08]</td>
<td>[00]</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>[24]</td>
<td>[16]</td>
</tr>
</tbody>
</table>

LMUL=4, SEW=32b

<table>
<thead>
<tr>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>[00]</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>[08]</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>[16]</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>[24]</td>
</tr>
</tbody>
</table>
LMUL=2, SEW=8b
3 2 1 0
[12][08][04][00]
7 6 5 4
[28][24][20][16]

LMUL=8, SEW=32b
0
[00]
1
[04]
2
[08]
3
[12]
4
[16]
5
[20]
6
[24]
7
[28]

LMUL=8, SEW=8b
3 2 1 0
[03][02][01][00]
7 6 5 4
[07][06][05][04]
B A 9 8
[11][10][09][08]
F E D C
[15][14][13][12]
13 12 11 10
[19][18][17][16]
17 16 15 14
[23][22][21][20]
1B 1A 19 18
[27][26][25][24]
1F 1E 1D 1C
[31][30][29][28]
VLEN=256b, SLEN=128b

| Byte | 1F 1E 1D 1C 1B 1A 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0 |

SEW=8b, LMUL=1, VLMAX=32

| v1 | 1F 1E 1D 1C 1B 1A 19 18 17 16 15 14 13 12 11 10 7 6 5 4 3 2 1 0 |

SEW=16b, LMUL=2, VLMAX=32

| v2*n | 17 16 15 14 13 12 11 10 7 6 5 4 3 2 1 0 |

SEW=32b, LMUL=4, VLMAX=32

| v4*n | 13 12 11 10 3 2 1 0 |

SEW=64b, LMUL=8, VLMAX=32

| v8*n | 11 10 1 0 |

| v8*n+1 | 13 12 3 2 |

| v8*n+2 | 15 14 5 4 |

| v8*n+3 | 17 16 7 6 |

| v8*n+4 | 19 18 9 8 |

| v8*n+5 | 1B 1A B A |

| v8*n+6 | 1D 1C D C |

| v8*n+7 | 1F 1E F E |

Mask bit position
5. Vector Instruction Formats

The instructions in the vector extension fit under four existing major opcodes (LOAD-FP, STORE-FP, AMO) and one new major opcode (OP-V).

Vector loads and stores are encoding within the scalar floating-point load and store major opcodes (LOAD-FP/STORE-FP). The vector load and store encodings repurpose a portion of the standard scalar floating-point load/store 12-bit immediate field to provide further vector instruction encoding, with bit 25 holding the standard vector mask bit (see Mask Encoding).

<table>
<thead>
<tr>
<th>Format for Vector Load Instructions under LOAD-FP major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 29 28 26 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Format for Vector Store Instructions under STORE-FP major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 29 28 26 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Format for Vector AMO Instructions under AMO major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 27 26 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>amoop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formats for Vector Arithmetic Instructions under OP-V major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 26 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>funct6</td>
</tr>
<tr>
<td>6 1 5 5 3 5 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formats for Vector Configuration Instructions under OP-V major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 30 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Vector instructions can have scalar or vector source operands and produce scalar or vector results, and most vector instructions can be performed either unconditionally or conditionally under a mask.

Vector loads and stores move bit patterns between vector register elements and memory. Vector arithmetic instructions operate on values held in vector register elements.
5.1. Scalar operands

Scalar operands can be immediates, or taken from the x registers, the f registers, or element 0 of a vector register. Scalar results are written to an x or f register or to element 0 of a vector register. Any vector register can be used to hold a scalar regardless of the current LMUL setting.

In a change from v0.6, the floating-point registers no longer overlay the vector registers and scalars can now come from the integer or floating-point registers. Not overlaying the f registers reduces vector register pressure, avoids interactions with the standard calling convention, simplifies high-performance scalar floating-point design, and provides compatibility with the Zfinx ISA option. Overlaying f with v would provide the advantage of lowering the number of state bits in some implementations, but complicates high-performance designs and would prevent compatibility with the Zfinx ISA option.

5.2. Vector Operands

Vector operands or results may occupy one or more vector registers depending on LMUL, but are always specified using the lowest-numbered vector register in the group. Using other than the lowest-numbered vector register to specify a vector register group will result in an illegal instruction exception.

Some vector instructions consume and produce wider-width elements and so operate on a larger vector register group than that specified in vlmul. The largest vector register group used by an instruction can not be greater than 8 vector registers, and if an vector instruction would require greater than 8 vector registers in a group, an illegal instruction exception is raised. For example, attempting a widening operation with LMUL=8 will raise an illegal instruction exception.

5.3. Vector Masking

Masking is supported on many vector instructions. Element operations that are masked off do not modify the destination vector register element and never generate exceptions.

In the base vector extension, the mask value used to control execution of a masked vector instruction is always supplied by vector register v0. Only the least-significant bit of each element of the mask vector is used to control execution.

Future vector extensions may provide longer instruction encodings with space for a full mask register specifier.

The destination vector register group for a masked vector instruction can only overlap the source mask register (v0) when LMUL=1. Otherwise, an illegal instruction exception is raised.

This constraint supports restart with a non-zero vstart value.

Other vector registers can be used to hold working mask values, and mask vector logical operations are provided to perform predicate calculations.

5.3.1. Mask Encoding

Where available, masking is encoded in a single-bit vm field in the instruction (inst[25]).

<table>
<thead>
<tr>
<th>vm</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>vector result, only where v0[i].LSB = 1</td>
</tr>
<tr>
<td>1</td>
<td>unmasked</td>
</tr>
</tbody>
</table>

In earlier proposals, vm was a two-bit field vm[1:0] that provided both true and complement masking using v0 as well as encoding scalar operations.

Vector masking is represented in assembler code as another vector operand, with .t indicating if operation occurs when v0[i].LSB is 1. If no masking operand is specified, unmasked vector execution (vm=1) is
Even though the base only supports one vector mask register v0 and only the true form of predication, the assembly syntax writes it out in full to be compatible with future extensions that might add a mask register specifier and supporting both true and complement masking. The .t suffix on the masking operand also helps to visually encode the use of a mask.

5.4. Prestart, Active, Inactive, Body, and Tail Element Definitions

The elements operated on during a vector instruction’s execution can be divided into four disjoint subsets.

- **The prestart elements** are those whose element index is less than the initial value in the vstart register. The prestart elements do not raise exceptions and do not update the destination vector register.

- **The active elements** during a vector instruction’s execution are the elements within the current vector length setting and where the current mask is enabled at that element position. The active elements can raise exceptions and update the destination vector register group.

- **The inactive elements** are the elements within the current vector length setting but where the current mask is disabled at that element position. The inactive elements do not raise exceptions and do not update any destination vector register.

- **The tail elements** during a vector instruction’s execution are the elements past the current vector length setting. The tail elements do not raise exceptions, and do not update any destination vector register group.

- In addition, another term, **body**, is used for the set of elements that are either active or inactive, i.e., after prestart but before the tail.

```plaintext
for element index x
prestart  = (0 <= x < vstart)
mask(x)   = unmasked || v0[x].LSB == 1
active(x) = (vstart <= x < vl) && mask(x)
inactive(x) = (vstart <= x < vl) && !mask(x)
body(x)   = active(x) || inactive(x)
tail(x)   = (vl <= x < VLMAX)
```

The inactive and tail update rules leave unchanged destination elements that are not participating in the vector operation. Previous versions (v0.7) of the specification zeros the tail elements to reduce the complexity of implementations that implement register renaming. In version 0.8, this was changed to leave the tail elements undisturbed, which reduces complexity for simpler implementations without register renaming and reduces software overhead for some common sequences. A rationale is provided in a separate document: https://github.com/riscv/riscv-v-spec/blob/master/v-undisturbed-versus-zeroing.adoc
6. Configuration-Setting Instructions

A set of instructions are provided to allow rapid configuration of the values in \(v1\) and \(vtype\) to match application needs.

6.1. \texttt{vsetvli/vsetvl} instructions

\begin{verbatim}
\texttt{vsetvli rd, rs1, vtypei} \# rd = new v1, rs1 = AVL, vtypei = new vtype setting
\texttt{vsetvl rd, rs1, rs2} \# rd = new v1, rs1 = AVL, rs2 = new vtype value
\end{verbatim}

The \texttt{vsetvli} instruction sets the \texttt{vtype} and \texttt{v1} CSRs based on its arguments, and writes the new value of \texttt{v1} into \texttt{rd}.

The new \texttt{vtype} setting is encoded in the immediate fields of \texttt{vsetvli} and in the \texttt{rs2} register for \texttt{vsetvl}.

Formats for Vector Configuration Instructions under OP-V major opcode

| 31 30 25 24 20 19 15 14 12 11 7 6 0 |
|---------------------------|-----------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 0 | zimm[10:0] | rs1 | 1 1 1 | rd | 1010111 | \texttt{vsetvli} |
| 1 | 000000 | rs2 | rs1 | 1 1 1 | rd | 1010111 | \texttt{vsetvl} |
| 1 | 6 | 5 | 5 | 5 | 7 |

Table 7. \texttt{vtype} register layout

<table>
<thead>
<tr>
<th>Bits</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLEN-1</td>
<td>vill</td>
<td>Illegal value if set</td>
</tr>
<tr>
<td>XLEN-2:7</td>
<td></td>
<td>Reserved (write 0)</td>
</tr>
<tr>
<td>6:5</td>
<td>vediv[1:0]</td>
<td>Used by EDIV extension</td>
</tr>
<tr>
<td>4:2</td>
<td>vsew[2:0]</td>
<td>Standard element width (SEW) setting</td>
</tr>
<tr>
<td>1:0</td>
<td>vlmul[1:0]</td>
<td>Vector register group multiplier (LMUL) setting</td>
</tr>
</tbody>
</table>
Suggested assembler names used for vtype setting

e8    #   8b elements  
e16   # 16b elements  
e32   # 32b elements  
e64   # 64b elements  
e128  # 128b elements  

m1   # Vlmul x1, assumed if m setting absent  
m2   # Vlmul x2  
m4   # Vlmul x4  
m8   # Vlmul x8  

d1   # EDIV 1, assumed if d setting absent  
d2   # EDIV 2  
d4   # EDIV 4  
d8   # EDIV 8  

Examples:
  vsetvli t0, a0, e8          # SEW= 8, LMUL=1, EDIV=1  
  vsetvli t0, a0, e8, m2      # SEW= 8, LMUL=2, EDIV=1  
  vsetvli t0, a0, e32, m2, d4  # SEW=32, LMUL=2, EDIV=4  

If the vtype setting is not supported by the implementation, then the vill bit is set in vtype, the remaining bits in vtype are set to zero, and the vl register is also set to zero.

The requested application vector length (AVL) is passed in rs1 as an unsigned integer. Using x0 as the rs1 register specifier requests an AVL equal to the current vl.

- Executing vsetvl{i} with rs1=x0 is guaranteed to keep the current vl intact if VLMAX is not reduced.
- The behavior of vsetvl{i} does not depend on whether rd=x0. Setting rd=x0 can be useful when the application already knows what value vl will assume, e.g., when changing SEW and LMUL by the same factor while keeping vl constant by setting rs1=x0.
- Software can set vl to VLMAX by requesting an AVL of -1.
- Earlier drafts used rs1=x0 to request an AVL of VLMAX. The current design supports efficient manipulation of SEW without knowing AVL or vl. Requesting a vl of VLMAX requires just one additional instruction.
- Earlier drafts required a trap when setting vtype to an illegal value. However, this would have added the first data-dependent trap on a CSR write to the ISA. The current scheme also supports light-weight runtime interrogation of the supported vector unit configurations by checking if vill is clear for a given setting.

6.2. Constraints on Setting vl

The vsetvl{i} instructions first set VLMAX according to the vtype argument, then set vl obeying the following constraints:

1. \( vl = AVL \text{ if } AVL \leq VLMAX \)
2. \( \lceil AVL / 2 \rceil \leq vl \leq VLMAX \text{ if } AVL < (2 \times VLMAX) \)
3. \( vl = VLMAX \text{ if } AVL \geq (2 \times VLMAX) \)
4. Deterministic on any given implementation for same input AVL and VLMAX values
5. These specific properties follow from the prior rules:
   a. \( vl = 0 \text{ if } AVL = 0 \)
   b. \( vl > 0 \text{ if } AVL > 0 \)
c. $vl \leq VLMAX$

d. $vl \leq AVL$

e. a value read from $vl$ when used as the AVL argument to $vsetvl{i}$ results in the same value in $vl$, provided the resultant $VLMAX$ equals the value of $VLMAX$ at the time that $vl$ was read

The $vl$ setting rules are designed to be sufficiently strict to preserve $vl$ behavior across register spills and context swaps for $AVL \leq VLMAX$, yet flexible enough to enable implementations to improve vector lane utilization for $AVL > VLMAX$.

For example, this permits an implementation to set $vl = \text{ceil}(AVL / 2)$ for $VLMAX < AVL < 2*VLMAX$ in order to evenly distribute work over the last two iterations of a stripmine loop. Requirement 2 ensures that the first stripmine iteration of reduction loops uses the largest vector length of all iterations, even in the case of $AVL < 2*VLMAX$. This allows software to avoid needing to explicitly calculate a running maximum of vector lengths observed during a stripmined loop.

### 6.3. vsetvl Instruction

The $vsetvl$ variant operates similarly to $vsetvli$ except that it takes a $vtype$ value from $rs2$ and can be used for context restore, and when the $vtypei$ field is too small to hold the desired setting.

Several active complex types can be held in different $x$ registers and swapped in as needed using $vsetvl$.

### 6.4. Examples

The SEW and LMUL settings can be changed dynamically to provide high throughput on mixed-width operations in a single loop.
# Example: Load 16-bit values, widen multiply to 32b, shift 32b result
# right by 3, store 32b values.

# Loop using only widest elements:

loop:
  vsetvlq a3, a0, e32, m8  # Use only 32-bit elements
  vlh.v v8, (a1)           # Sign-extend 16b load values to 32b elements
  sll t1, a3, 1            # Multiply length by two bytes/element
  add a1, a1, t1           # Bump pointer
  vmul.vx v8, v8, x10      # 32b multiply result
  vsrl.vi v8, v8, 3        # Shift elements
  vsw.v v8, (a2)           # Store vector of 32b results
  sll t1, a3, 2            # Multiply length by four bytes/element
  add a2, a2, t1           # Bump pointer
  sub a0, a0, a3           # Decrement count
  bnez a0, loop            # Any more?

# Alternative loop that switches element widths.

loop:
  vsetvlq a3, a0, e16, m4  # vtype = 16-bit integer vectors
  vlh.v v4, (a1)           # Get 16b vector
  slli t1, a3, 1           # Multiply length by two bytes/element
  add a1, a1, t1           # Bump pointer
  vwmul.vx v8, v4, x10     # 32b in <v8--v15>
  vsetvlq x0, a0, e32, m8  # Operate on 32b values
  vsrl.vi v8, v8, 3        # Store vector of 32b
  vsw.v v8, (a2)           # Store vector of 32b
  slli t1, a3, 2           # Multiply length by four bytes/element
  add a2, a2, t1           # Bump pointer
  sub a0, a0, a3           # Decrement count
  bnez a0, loop            # Any more?

The second loop is more complex but will have greater performance on machines where 16b widening
multiplies are faster than 32b integer multiplies, and where 16b vector load can run faster due to the
narrower writes to the vector regfile.
7. Vector Loads and Stores

Vector loads and stores move values between vector registers and memory. Vector loads and stores are masked and do not raise exceptions on inactive elements. Masked vector loads do not update inactive elements in the destination vector register group. Masked vector stores only update active memory elements.

7.1. Vector Load/Store Instruction Encoding

Vector loads and stores are encoded within the scalar floating-point load and store major opcodes (LOAD-FP/STORE-FP). The vector load and store encodings repurpose a portion of the standard scalar floating-point load/store 12-bit immediate field to provide further vector instruction encoding, with bit 25 holding the standard vector mask bit (see Mask Encoding).

<table>
<thead>
<tr>
<th>Format for Vector Load Instructions under LOAD-FP major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 29 28 26 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>3  3  1  5  5  3  3  5  7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Format for Vector Store Instructions under STORE-FP major opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 29 28 26 25 24 20 19 15 14 12 11 7 6 0</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>nf</td>
</tr>
<tr>
<td>3  3  1  5  5  3  3  5  7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1[4:0]</td>
<td>specifies x register holding base address</td>
</tr>
<tr>
<td>rs2[4:0]</td>
<td>specifies x register holding stride</td>
</tr>
<tr>
<td>vs2[4:0]</td>
<td>specifies v register holding address offsets</td>
</tr>
<tr>
<td>vs3[4:0]</td>
<td>specifies v register holding store data</td>
</tr>
<tr>
<td>vd[4:0]</td>
<td>specifies v register destination of load</td>
</tr>
<tr>
<td>vm</td>
<td>specifies vector mask</td>
</tr>
<tr>
<td>width[2:0]</td>
<td>specifies size of memory elements, and distinguishes from FP scalar</td>
</tr>
<tr>
<td>mop[2:0]</td>
<td>specifies memory addressing mode</td>
</tr>
<tr>
<td>nf[2:0]</td>
<td>specifies the number of fields in each segment, for segment load/stores</td>
</tr>
<tr>
<td>lumop[4:0]/sumop[4:0]</td>
<td>are additional fields encoding variants of unit-stride instructions</td>
</tr>
</tbody>
</table>

7.2. Vector Load/Store Addressing Modes

The base vector extension supports unit-stride, strided, and indexed (scatter/gather) addressing modes. Vector load/store base registers and strides are taken from the GPR x registers.

The base effective address for all vector accesses is given by the contents of the x register named in rs1.

Vector unit-stride operations access elements stored contiguously in memory starting from the base effective address.

Vector strided operations access the first memory element at the base effective address, and then access subsequent elements at address increments given by the byte offset contained in the x register specified by
Vector indexed operations add the contents of each element of the vector offset operand specified by vs2 to the base effective address to give the effective address of each element. The vector offset operand is treated as a vector of byte offsets. If the vector offset elements are narrower than XLEN, they are sign-extended to XLEN before adding to the base effective address. If the vector offset elements are wider than XLEN, the least-significant XLEN bits are used in the address calculation.

Current PoR for vector indexed instructions requires that vector byte offset (vs2) and vector read/write data (vs3/vd) are of same width. One question is whether and how to allow for two sizes of vector operand in a vector indexed instruction? For example, for scatter/gather of byte values in a 64-bit address space without requiring bytes use 64b of space in a vector register.

The vector addressing modes are encoded using the 3-bit m[2:0] field.

<table>
<thead>
<tr>
<th>m[2:0]</th>
<th>Description</th>
<th>Opcodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0</td>
<td>zero-extended unit-stride</td>
<td>VLxU,VLE</td>
</tr>
<tr>
<td>0 0 1</td>
<td>reserved</td>
<td></td>
</tr>
<tr>
<td>0 1 0</td>
<td>zero-extended strided</td>
<td>VLSxU, VLSE</td>
</tr>
<tr>
<td>0 1 1</td>
<td>zero-extended indexed</td>
<td>VLXxU, VLXE</td>
</tr>
<tr>
<td>1 0 0</td>
<td>sign-extended unit-stride</td>
<td>VLx (x!=E)</td>
</tr>
<tr>
<td>1 0 1</td>
<td>reserved</td>
<td></td>
</tr>
<tr>
<td>1 1 0</td>
<td>sign-extended strided</td>
<td>VLSx (x!=E)</td>
</tr>
<tr>
<td>1 1 1</td>
<td>sign-extended indexed</td>
<td>VLSx (x!=E)</td>
</tr>
</tbody>
</table>

The vector indexed memory operations have two forms, ordered and unordered. The indexed-unordered stores do not preserve element ordering on stores.

The indexed-unordered variant is provided as a potential implementation optimization. Implementations are free to ignore the optimization and implement indexed-unordered identically to indexed-ordered.

Additional unit-stride vector addressing modes are encoded using the 5-bit lumop and sumop fields in the unit-stride load and store instruction encodings respectively.
Table 10. lumop

<table>
<thead>
<tr>
<th>lumop[4:0]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0</td>
<td>unit-stride</td>
</tr>
<tr>
<td>0 x x x x</td>
<td>reserved, x!=0</td>
</tr>
<tr>
<td>1 0 0 0 0</td>
<td>unit-stride fault-only-first</td>
</tr>
<tr>
<td>1 x x x x</td>
<td>reserved, x!=0</td>
</tr>
</tbody>
</table>

Table 11. sumop

<table>
<thead>
<tr>
<th>sumop[4:0]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0</td>
<td>unit-stride</td>
</tr>
<tr>
<td>0 x x x x</td>
<td>reserved, x!=0</td>
</tr>
<tr>
<td>1 x x x x</td>
<td>reserved</td>
</tr>
</tbody>
</table>

The nf[2:0] field encodes the number of fields in each segment. For regular vector loads and stores, nf=0, indicating that a single value is moved between a vector register group and memory at each element position. Larger values in the nf field are used to access multiple contiguous fields within a segment as described below in Section Vector Load/Store Segment Instructions (Zv1sseg).

The nf field for segment load/stores has replaced the use of the same bits for an address offset field. The offset can be replaced with a single scalar integer calculation, while segment load/stores add more powerful primitives to move items to and from memory.

7.3. Vector Load/Store Width Encoding

The vector loads and stores are encoded using the width values that are not claimed by the standard scalar floating-point loads and stores. Three of the width types encode vector loads and stores that move fixed-size memory elements of 8 bits, 16 bits, or 32 bits, while the fourth encoding moves SEW-bit memory elements.

<table>
<thead>
<tr>
<th>Width [2:0]</th>
<th>Mem bits</th>
<th>Reg bits</th>
<th>Opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard scalar FP</td>
<td>0 0 1</td>
<td>16</td>
<td>FLEN</td>
</tr>
<tr>
<td>Standard scalar FP</td>
<td>0 1 0</td>
<td>32</td>
<td>FLEN</td>
</tr>
<tr>
<td>Standard scalar FP</td>
<td>0 1 1</td>
<td>64</td>
<td>FLEN</td>
</tr>
<tr>
<td>Standard scalar FP</td>
<td>1 0 0</td>
<td>128</td>
<td>FLEN</td>
</tr>
<tr>
<td>Vector byte</td>
<td>0 0 0 v<em>l</em>8</td>
<td>vl*SEW</td>
<td>VxB</td>
</tr>
<tr>
<td>Vector halfword</td>
<td>1 0 1 v<em>l</em>16</td>
<td>vl*SEW</td>
<td>VxH</td>
</tr>
<tr>
<td>Vector word</td>
<td>1 1 0 v<em>l</em>32</td>
<td>vl*SEW</td>
<td>VxW</td>
</tr>
<tr>
<td>Vector element</td>
<td>1 1 1 v<em>l</em>SEW</td>
<td>vl*SEW</td>
<td>VxE</td>
</tr>
</tbody>
</table>

Mem bits is the size of element accessed in memory

Reg bits is the size of element accessed in register

Fixed-sized vector loads can optionally sign or zero-extend their memory element into the destination register element if the register element is wider than the memory element. A fixed-size vector load raises an illegal instruction exception if the destination register element is narrower than the memory element. The variable-sized load is encoded as if a zero-extended load, with what would be the sign-extended encoding of a variable-sized load currently reserved.

Fixed-size vector stores take their operand from the least-significant bits of the register element if the register element if wider than the memory element. Fixed-sized vector stores raise an illegal instruction exception if the memory element is wider than the register element.
7.4. Vector Unit-Stride Instructions

# Vector unit-stride loads and stores

# vd destination, rs1 base address, vm is mask encoding (v0.t or <missing>)
vlb.v  vd, (rs1), vm  # 8b signed
vlh.v  vd, (rs1), vm  # 16b signed
vlw.v  vd, (rs1), vm  # 32b signed

vlbu.v vd, (rs1), vm  # 8b unsigned
vlhu.v vd, (rs1), vm  # 16b unsigned
vlwu.v vd, (rs1), vm  # 32b unsigned

vle.v  vd, (rs1), vm  # SEW

# vs3 store data, rs1 base address, vm is mask encoding (v0.t or <missing>)
vsb.v vs3, (rs1), vm  # 8b store
vsh.v vs3, (rs1), vm  # 16b store
vsw.v vs3, (rs1), vm  # 32b store
vse.v vs3, (rs1), vm  # SEW store

7.5. Vector Strided Instructions

# Vector strided loads and stores

# vd destination, rs1 base address, rs2 byte stride
vlsb.v vd, (rs1), rs2, vm  # 8b
vlsh.v vd, (rs1), rs2, vm  # 16b
vlsw.v vd, (rs1), rs2, vm  # 32b

vlsbu.v vd, (rs1), rs2, vm  # unsigned 8b
vlshu.v vd, (rs1), rs2, vm  # unsigned 16b
vlswu.v vd, (rs1), rs2, vm  # unsigned 32b

vlse.v vd, (rs1), rs2, vm  # SEW

# vs3 store data, rs1 base address, rs2 byte stride
vssb.v vs3, (rs1), rs2, vm  # 8b
vssh.v vs3, (rs1), rs2, vm  # 16b
vssw.v vs3, (rs1), rs2, vm  # 32b
vsse.v vs3, (rs1), rs2, vm  # SEW

Negative and zero strides are supported.

7.6. Vector Indexed Instructions
# Vector indexed loads and stores

# vd destination, rs1 base address, vs2 indices
vlxb.v vd, (rs1), vs2, vm # 8b
vlxh.v vd, (rs1), vs2, vm # 16b
vlxw.v vd, (rs1), vs2, vm # 32b

vlxbu.v vd, (rs1), vs2, vm # 8b unsigned
vlxhu.v vd, (rs1), vs2, vm # 16b unsigned
vlxwu.v vd, (rs1), vs2, vm # 32b unsigned

vlxe.v vd, (rs1), vs2, vm # SEW

# Vector ordered-indexed store instructions
# vs3 store data, rs1 base address, vs2 indices
vsxb.v vs3, (rs1), vs2, vm # 8b
vsxh.v vs3, (rs1), vs2, vm # 16b
vsxw.v vs3, (rs1), vs2, vm # 32b
vsxe.v vs3, (rs1), vs2, vm # SEW

# Vector unordered-indexed store instructions
vsuxb.v vs3, (rs1), vs2, vm # 8b
vsuxh.v vs3, (rs1), vs2, vm # 16b
vsuxw.v vs3, (rs1), vs2, vm # 32b
vsuxe.v vs3, (rs1), vs2, vm # SEW

## 7.7. Unit-stride Fault-Only-First Loads

The unit-stride fault-only-first load instructions are used to vectorize loops with data-dependent exit conditions (while loops). These instructions execute as a regular load except that they will only take a trap on element 0. If an element > 0 raises an exception, that element and all following elements in the destination vector register are not modified, and the vector length \( v1 \) is reduced to the number of elements processed without a trap.

vlbff.v vd, (rs1), vm # 8b
vlhff.v vd, (rs1), vm # 16b
vlwff.v vd, (rs1), vm # 32b

vlbuff.v vd, (rs1), vm # unsigned 8b
vlhuff.v vd, (rs1), vm # unsigned 16b
vlwuff.v vd, (rs1), vm # unsigned 32b

vleff.v vd, (rs1), vm # SEW
strlen example using unit-stride fault-only-first instruction

```c
#define size_t strlen(const char *str)
#define a0 holds *str

strlen:
    mv a3, a0       # Save start
    li t0, -1       # Infinite AVL

loop:
    vsetvli a1, t0, e8  # Vector of bytes of maximum length
    vlbff.v v1, (a3)     # Load bytes
    csrr a1, vl         # Get bytes read
    vmseq.vi v0, v1, 0  # Set v0[i] where v1[i] = 0
    vfirst.m a2, v0     # Find first set bit
    add a3, a3, a1      # Bump pointer
    bltz a2, loop       # Not found?

    add a0, a0, a1      # Sum start + bump
    add a3, a3, a2      # Add index
    sub a0, a3, a0      # Subtract start address+bump

ret
```

Strided and scatter/gather fault-only-first instructions are not provided as they represent a large security hole, allowing software to check multiple random pages for accessibility without experiencing a trap. The unit-stride versions only allow probing a region immediately contiguous to a known region, and so do not appreciably impact security. It is possible that security mitigations can be implemented to allow fault-only-first variants of non-contiguous accesses in future vector extensions.

7.8. Vector Load/Store Segment Instructions (Zv1sseg)

This is being written as an extension but will likely be mandated in most profiles, as the operation is too generally useful to omit.

The vector load/store segment instructions move multiple contiguous fields in memory to and from consecutively numbered vector registers.

These operations support operations on "array-of-structures" datatypes by unpacking each field in a structure into separate vector registers.

As for regular vector loads and stores, the width encoding gives the size of the memory elements, which are homogeneous in size, while SEW encodes the size of the register elements.

The three-bit nf field in the vector instruction encoding is an unsigned integer that contains one less than the number of fields per segment, NFIELDS.

<table>
<thead>
<tr>
<th>nf[2:0]</th>
<th>NFIELDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0</td>
<td>1</td>
</tr>
<tr>
<td>0 0 1</td>
<td>2</td>
</tr>
<tr>
<td>0 1 0</td>
<td>3</td>
</tr>
<tr>
<td>0 1 1</td>
<td>4</td>
</tr>
<tr>
<td>1 0 0</td>
<td>5</td>
</tr>
<tr>
<td>1 0 1</td>
<td>6</td>
</tr>
<tr>
<td>1 1 0</td>
<td>7</td>
</tr>
<tr>
<td>1 1 1</td>
<td>8</td>
</tr>
</tbody>
</table>

The LMUL setting must be such that LMUL * NFIELDS \(\leq 8\), otherwise an illegal instruction exception is raised.
The product LMUL * NFIELDS represents the number of underlying vector registers that will be touched by a segmented load or store instruction. This constraint makes this total no larger than 1/4 of the architectural register file, and the same as for regular operations with LMUL=8. This constraint could be weakened in a future draft.

Each field will be held in successively numbered vector register groups. When LMUL>1, each field will occupy a vector register group held in multiple successively numbered vector registers, and the vector register group for each field must follow the usual vector register alignment constraints (e.g., when LMUL=2 and NFIELDS=4, each field’s vector register group must start at an even vector register, but does not have to start at a multiple of 8 vector register number).

An earlier version imposed a vector register number constraint, but this decreased ability to make use of all registers when NFIELDS was not a power of 2.

If the vector register numbers accessed by the segment load or store would increment past 31, then an illegal instruction exception is raised.

This constraint is to help provide forward-compatibility with a future longer instruction encoding that has more addressable vector registers.

The vl register gives the number of structures to move, which is equal to the number of elements transferred to each vector register group. Masking is also applied at the level of whole structures.

If a trap is taken, vstart is in units of structures.

### 7.8.1. Vector Unit-Stride Segment Loads and Stores

The vector unit-stride load and store segment instructions move packed contiguous segments (“array-of-structures”) into multiple destination vector register groups.

For segments with heterogeneous-sized fields, software can later unpack fields using additional instructions after the segment load brings the values into the separate vector registers.

The assembler prefixes vlseg/vssegs are used for unit-stride segment loads and stores respectively.

```plaintext
# Format
vlseg<nf>{b,h,w}.v vd, (rs1), vm    # Unit-stride signed segment load template
vlseg<nf>e.v vd, (rs1), vm           # Unit-stride segment load template
vlseg<nf>{b,h,w}u.v vd, (rs1), vm    # Unit-stride unsigned segment load template
vsseg<nf>{b,h,w,e}.v vs3, (rs1), vm  # Unit-stride segment store template

# Examples
vlseg2b.v vd, (rs1), vm    # Load vector of signed 2*1-byte segments into vd, vd+1
vlseg3bu.v vd, (rs1), vm    # Load vector of unsigned 3*1-byte segments into vd, vd+1, vd+4
vlseg7w.v vd, (rs1), vm    # Load vector of 7*4-byte segments into vd, vd+1, ... vd+6
vlseg8e.v vd, (rs1), vm    # Load vector of 8*SEW-byte segments into vd, vd+1, ... vd+7
vsseg3b.v vs3, (rs1), vm    # Store packed vector of 3*1-byte segments from vs3,vs3+1,vs3+2
```

For loads, the vd register will hold the first field loaded from the segment. For stores, the vs3 register is read to provide the first field to be stored in each segment.
# Example 1
# Memory structure holds packed RGB pixels (24-bit data structure, 8bpp)
v1seg3bu.v v8, (a0), vm
# v8 holds the red pixels
# v9 holds the green pixels
# v10 holds the blue pixels

# Example 2
# Memory structure holds complex values, 32b for real and 32b for imaginary
v1seg2w.v v8, (a0), vm
# v8 holds real
# v9 holds imaginary

There are also fault-only-first versions of the unit-stride instructions.

# Template for vector fault-only-first unit-stride segment loads and stores.
v1seg<nf>{b,h,w}ff.v vd, (rs1), vm    # Unit-stride signed fault-only-first segment loads
v1seg<nf>eff.v vd, (rs1), vm          # Unit-stride fault-only-first segment loads
v1seg<nf>{b,h,w}uff.v vd, (rs1), vm  # Unit-stride unsigned fault-only-first segment loads

7.8.2. Vector Strided Segment Loads and Stores

Vector strided segment loads and stores move contiguous segments where each segment is separated by the byte stride offset given in the rs2 GPR argument.

Negative and zero strides are supported.

# Format
v1seg<nf>{b,h,w}.v vd, (rs1), rs2, vm    # Strided signed segment loads
v1seg<nf>e.v vd, (rs1), rs2, vm          # Strided segment loads
v1seg<nf>{b,h,w}u.v vd, (rs1), rs2, vm  # Strided unsigned segment loads
vssseg<nf>{b,h,w,e}.v vs3, (rs1), rs2, vm # Strided segment stores

# Examples
v1seg3b.v v4, (x5), x6     # Load bytes at addresses x5+i*x6 into v4[i],
# and bytes at addresses x5+i*x6+1 into v5[i],
# and bytes at addresses x5+i*x6+2 into v6[i].

# Examples
vssseg2w.v v2, (x5), x6    # Store words from v2[i] to address x5+i*x6
# and words from v3[i] to address x5+i*x6+4

For strided segment stores where the byte stride is such that segments could overlap in memory, the segments must appear to be written in element order.

7.8.3. Vector Indexed Segment Loads and Stores

Vector indexed segment loads and stores move contiguous segments where each segment is located at an address given by adding the scalar base address in the rs1 field to byte offsets in vector register vs2.
# Format
vlxseg<nf>{b,h,w}.v vd, (rs1), vs2, vm  # Indexed signed segment loads
vlxseg<nf>e.v vd, (rs1), vs2, vm          # Indexed segment loads
vlxseg<nf>{b,h,w}u.v vd, (rs1), vs2, vm  # Indexed unsigned segment loads
vsxseg<nf>{b,h,w,e}.v vs3, (rs1), vs2, vm # Indexed segment stores

# Examples
vlxseg3bu.v v4, (x5), v3  # Load bytes at addresses x5+v3[i] into v4[i],
                  # and bytes at addresses x5+v3[i]+1 into v5[i],
                  # and bytes at addresses x5+v3[i]+2 into v6[i].

# Examples
vsxseg2w.v v2, (x5), v5  # Store words from v2[i] to address x5+v5[i]
                  # and words from v3[i] to address x5+v5[i]+4

For vector indexed segment loads, the destination vector register groups cannot overlap the source vector
register group (specified by vs2), nor can they overlap the mask register if masked, else an illegal instruction
exception is raised.

This constraint supports restart of indexed segment loads that raise exceptions partway through loading a structure.

Only ordered indexed segment stores are provided. The segments must appear to be written in element order.
8. Vector AMO Operations (Zvamo)

Profiles will dictate whether vector AMO operations are supported. The expectation is that the Unix profile will require vector AMO operations.

If vector AMO instructions are supported, then the scalar Zaamo instructions (atomic operations from the standard A extension) must be present.

Vector AMO operations are encoded using the unused width encodings under the standard AMO major opcode. Each active element performs an atomic read-modify-write of a single memory location.

```
<table>
<thead>
<tr>
<th>Width [2:0]</th>
<th>Mem bits</th>
<th>Reg bits</th>
<th>Opcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>XLEN</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>64</td>
</tr>
</tbody>
</table>
```

Each element performs an atomic read-modify-write of a single memory location.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

AMOs have the same addressing mode as indexed operations except with no immediate offset. A vector of byte offsets in register vs2 are added to the scalar base register in rs1 to give the addresses of the AMO operations.

The vs2 vector register supplies the byte offset of each element, while the vs3 vector register supplies the source data for the atomic memory operation.

If the wd bit is set, the vd register is written with the initial value of the memory element. If the wd bit is clear, the vd register is not written.

Vector AMOs operate as if \(a_1\) and \(r1\) bits were zero on each element with regard to ordering relative to other instructions in the same hart.

Vector AMOs provide no ordering guarantee between element operations in the same vector AMO instruction.
Mem bits is the size of element accessed in memory

Reg bits is the size of element accessed in register

There are two widths of vector AMO, one for 32-bit words and one for SEW-bit words. For the 32-bit vector AMO operations, SEW must be at least 32 bits, otherwise an illegal instruction exception is raised. If SEW > 32 bits, the value returned from memory is sign-extended to fill SEW.

If SEW is less than XLEN, then addresses in the vector vs2 are sign-extended to XLEN. If SEW is greater than XLEN, an illegal instruction exception is raised.

Vector AMO instructions are only supported for the memory element widths supported by AMOs in the implementation’s scalar architecture. Other widths raise an illegal instruction exception.

The vector amo[4:0] field uses the same encoding as the scalar 5-bit AMO instruction field, except that LR and SC are not supported.

Table 13. amoop

<table>
<thead>
<tr>
<th>amoop</th>
<th>opcode</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0 1</td>
<td>vamoswap</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 0 0</td>
<td>vamoadd</td>
<td></td>
</tr>
<tr>
<td>0 0 1 0 0 0</td>
<td>vamoxor</td>
<td></td>
</tr>
<tr>
<td>0 1 1 0 0 0</td>
<td>vamoand</td>
<td></td>
</tr>
<tr>
<td>0 1 0 0 0 0</td>
<td>vamoor</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0 0 0</td>
<td>vamomin</td>
<td></td>
</tr>
<tr>
<td>1 0 1 0 0 0</td>
<td>vamomax</td>
<td></td>
</tr>
<tr>
<td>1 1 0 0 0 0</td>
<td>vamominu</td>
<td></td>
</tr>
<tr>
<td>1 1 1 0 0 0</td>
<td>vamomaxu</td>
<td></td>
</tr>
</tbody>
</table>

The assembly syntax uses x0 in the destination register position to indicate the return value is not required (wd=0).
# 32-bit vector AMOs
vamoswap.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoswap.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoaddw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoaddw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoxorw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoxorw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoandw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoandw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoorw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoorw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamominw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamominw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamomaxw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamomaxw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamominuw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamominuw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamomaxuw.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamomaxuw.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

# SEW-bit vector AMOs
vamoswape.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoswape.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoadde.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoadde.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoxore.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoxore.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoande.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoande.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamoore.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamoore.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamomin.e.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamomin.e.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamomax.e.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamomax.e.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamominue.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamominue.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0

vamomaxue.v vd, (rs1), v2, vd, v0.t # Write original value to register, wd=1
vamomaxue.v x0, (rs1), v2, vs3, v0.t # Do not write original value to register, wd=0
9. Vector Memory Alignment Constraints

If the elements accessed by a vector memory instruction are not naturally aligned to the memory element size, either an address misaligned exception is raised on that element or the element is transferred successfully.

Vector memory accesses follow the same rules for atomicity as scalar memory accesses.
10. Vector Memory Consistency Model

Vector memory instructions appear to execute in program order on the local hart. Vector memory instructions follow RVWMO at the instruction level, and element operations are ordered within the instruction as if performed by an element-ordered sequence of syntactically independent scalar instructions. Vector indexed-ordered stores write elements to memory in element order. Vector indexed-unordered stores do not preserve element order for writes within a single vector store instruction.

Need to flesh out details.
11. Vector Arithmetic Instruction Formats

The vector arithmetic instructions use a new major opcode (OP-V = 1010111) which neighbors OP-FP. The three-bit funct3 field is used to define sub-categories of vector instructions.

Formats for Vector Arithmetic Instructions under OP-V major opcode

<table>
<thead>
<tr>
<th>31</th>
<th>26</th>
<th>25</th>
<th>24</th>
<th>20</th>
<th>19</th>
<th>15</th>
<th>14</th>
<th>12</th>
<th>11</th>
<th>7</th>
<th>6</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>vs1</td>
<td>0 0 0</td>
<td>vd</td>
<td>1010111</td>
<td>OP-V (OPIVV)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>vs1</td>
<td>0 0 1</td>
<td>vd/rd</td>
<td>1010111</td>
<td>OP-V (OPFVV)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>vs1</td>
<td>0 1 0</td>
<td>vd/rd</td>
<td>1010111</td>
<td>OP-V (OPMVV)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>simm5</td>
<td>0 1 1</td>
<td>vd</td>
<td>1010111</td>
<td>OP-V (OPIVI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>rs1</td>
<td>1 0 0</td>
<td>vd</td>
<td>1010111</td>
<td>OP-V (OPIVX)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>rs1</td>
<td>1 0 1</td>
<td>vd</td>
<td>1010111</td>
<td>OP-V (OPFVF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>funct6</td>
<td>vm</td>
<td>vs2</td>
<td>rs1</td>
<td>1 1 0</td>
<td>vd/rd</td>
<td>1010111</td>
<td>OP-V (OPMVX)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11.1. Vector Arithmetic Instruction encoding

The funct3 field encodes the operand type and source locations.

<table>
<thead>
<tr>
<th>funct3[2:0]</th>
<th>Operands</th>
<th>Source of scalar(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0</td>
<td>OPIVV</td>
<td>vector-vector</td>
</tr>
<tr>
<td>0 0 1</td>
<td>OPFVV</td>
<td>vector-vector</td>
</tr>
<tr>
<td>0 1 0</td>
<td>OPMVV</td>
<td>vector-vector</td>
</tr>
<tr>
<td>0 1 1</td>
<td>OPIVI</td>
<td>vector-immediate</td>
</tr>
<tr>
<td>1 0 0</td>
<td>OPIVX</td>
<td>vector-scalar</td>
</tr>
<tr>
<td>1 0 1</td>
<td>OPFVF</td>
<td>vector-scalar</td>
</tr>
<tr>
<td>1 1 0</td>
<td>OPMVX</td>
<td>vector-scalar</td>
</tr>
<tr>
<td>1 1 1</td>
<td>OPCFG</td>
<td>scalars-imms</td>
</tr>
</tbody>
</table>

Integer operations are performed using unsigned or two's-complement signed integer arithmetic depending on the opcode.

All standard vector floating-point arithmetic operations follow the IEEE-754/2008 standard. All vector floating-point operations use the dynamic rounding mode in the frm register.

Vector-vector operations take two vectors of operands from vector register groups specified by vs2 and vs1 respectively.

Vector-scalar operations can have three possible forms, but in all cases take one vector of operands from a vector register group specified by vs2 and a second scalar source operand from one of three alternative sources.

1. For integer operations, the scalar can be a 5-bit immediate encoded in the rs1 field. The value is sign- or zero-extended to SEW bits.

2. For integer operations, the scalar can be taken from the scalar x register specified by rs1. If XLEN>SEW, the least-significant bits of the x register are used. If XLEN<SEW, the value from the x register is sign-extended to SEW bits.
For floating-point operations, the scalar can be taken from a scalar f register. If FLEN>SEW, the value in the f registers is checked for a valid NaN-boxed value, in which case the least-significant bits of the `f` register are used, else the canonical NaN value is used. If FLEN<SEW, the value is NaN-boxed (one-extended) to SEW.

The proposed Zfnx variants will take the floating-point scalar argument from the x registers.

Vector arithmetic instructions are masked under control of the vm field.

---

Assembly syntax pattern for vector binary arithmetic instructions

```assembly
# Operations returning vector results, masked by vm (v0.t, <nothing>)
vop.vv  vd, vs2, vs1, vm  # integer vector-vector     vd[i] = vs2[i] op vs1[i]
vop.vx  vd, vs2, rs1, vm  # integer vector-scalar      vd[i] = vs2[i] op x[rs1]
vop.vi  vd, vs2, imm, vm  # integer vector-immediate   vd[i] = vs2[i] op imm

vfop.vv vd, vs2, vs1, vm # FP vector-vector operation vd[i] = vs2[i] fop vs1[i]
vfop.vf vd, vs2, rs1, vm # FP vector-scalar operation vd[i] = vs2[i] fop f[rs1]
```

In the encoding, vs2 is the first operand, while rs1/simm5 is the second operand. This is the opposite to the standard scalar ordering. This arrangement retains the existing encoding conventions that instructions that read only one scalar register, read it from rs1, and that 5-bit immediates are sourced from the rs1 field.

Assembly syntax pattern for vector ternary arithmetic instructions (multiply-add)

```assembly
# Integer operations overwriting sum input
vop.vv  vd, vs1, vs2, vm  # vd[i] = vs1[i] * vs2[i] + vd[i]
vop.vx  vd, rs1, vs2, vm  # vd[i] = x[rs1] * vs2[i] + vd[i]

# Integer operations overwriting product input
vop.vv  vd, vs1, vs2, vm  # vd[i] = vs1[i] * vd[i] + vs2[i]
vop.vx  vd, rs1, vs2, vm  # vd[i] = x[rs1] * vd[i] + vs2[i]

# Floating-point operations overwriting sum input
vfop.vv vd, vs1, vs2, vm  # vd[i] = vs1[i] * vs2[i] + vd[i]
vfop.vf vd, rs1, vs2, vm  # vd[i] = f[rs1] * vs2[i] + vd[i]

# Floating-point operations overwriting product input
vfop.vv vd, vs1, vs2, vm  # vd[i] = vs1[i] * vd[i] + vs2[i]
vfop.vf vd, rs1, vs2, vm  # vd[i] = f[rs1] * vd[i] + vs2[i]
```

For ternary multiply-add operations, the assembler syntax always places the destination vector register first, followed by either rs1 or vs1, then vs2. This ordering provides a more natural reading of the assembler for these ternary operations, as the multiply operands are always next to each other.

---

11.2. Widening Vector Arithmetic Instructions

A few vector arithmetic instructions are defined to be *widening* operations where the destination elements are 2*SEW wide and are stored in a vector register group with twice the number of vector registers.

The first operand can be either single or double-width. These are generally written with a vw* prefix on the opcode or vfw* for vector floating-point operations.
Assembly syntax pattern for vector widening arithmetic instructions

# Double-width result, two single-width sources: 2*SEW = SEW op SEW

vwop.vv vd, vs2, vs1, vm # integer vector-vector vd[i] = vs2[i] op vs1[i]
vwop.vx vd, vs2, rs1, vm # integer vector-scalar vd[i] = vs2[i] op x[rs1]

# Double-width result, first source double-width, second source single-width: 2*SEW = 2*SEW

vwop.wv vd, vs2, vs1, vm # integer vector-vector vd[i] = vs2[i] op vs1[i]
vwop.wx vd, vs2, rs1, vm # integer vector-scalar vd[i] = vs2[i] op x[rs1]

Originally, a w suffix was used on opcode, but this could be confused with the use of a w suffix to mean word-sized operations in doubleword integers, so the w was moved to prex.

The floating-point widening operations were changed to vfw* from vwf* to be more consistent with any scalar widening floating-point operations that will be written as fw*.

For integer multiply-add, another possible widening option increases the size of the accumulator to 4*SEW (i.e., 4*SEW += SEW*SEW). These would be distinguished by a vw4* prefix on the opcode. These are not included at this time, but are a possible addition to spec.

The destination vector register group results are arranged as if both SEW and LMUL were at twice their current settings (i.e., the destination element width is 2*SEW, and the destination vector register group LMUL is 2*LMUL).

For all widening instructions, the destination element width must be a supported element width and the destination LMUL value must also be a supported LMUL value (≤8, i.e., current LMUL must be ≤4), otherwise an illegal instruction exception is raised.

The destination vector register group must be specified using a vector register number that is valid for the destination’s LMUL value, otherwise an illegal instruction exception is raised.

The destination vector register group cannot overlap a source vector register group of a different element width (including the mask register if masked), otherwise an illegal instruction exception is raised.

This constraint is necessary to support restart with non-zero vstart.

For the vw<op>.wv vd, vs2, vs1 format instructions, it is legal for vd to equal vs2.

11.3. Narrowing Vector Arithmetic Instructions

A few instructions are provided to convert double-width source vectors into single-width destination vectors. These instructions convert a vector register group organized as if LMUL and SEW were twice the current settings, and convert to a vector register group with the current LMUL/SEW vectors/elements.

If (2*LMUL > 8), or (2 * SEW) > ELEN, an illegal instruction exception is raised.

An alternative design decision would have been to treat LMUL as defining the size of the source vector register group. The choice here is motivated by the belief the chosen approach will require fewer LMUL changes.

The source and destination vector register groups have to be specified with a vector register number that is legal for the source and destination LMUL value respectively, otherwise an illegal instruction exception is raised.

Where there is a second source vector register group (specified by vs1), this has the same (narrower) width as the result.

The destination vector register group cannot overlap the first source vector register group (specified by vs2). The destination vector register group cannot overlap the mask register if used, unless LMUL=1. If either constraint is violated, an illegal instruction exception is raised.
It is safe to overwrite a second source vector register group with the same LMUL and element width as the result, or to overwrite a mask register when LMUL=1.

A \texttt{vn*} prefix on the opcode is used to distinguish these instructions in the assembler, or a \texttt{vfnn*} prefix for narrowing floating-point opcodes.

Comparison operations that set a mask register are also implicitly a narrowing operation.
12. Vector Integer Arithmetic Instructions

A set of vector integer arithmetic instructions are provided.

12.1. Vector Single-Width Integer Add and Subtract

Vector integer add and subtract are provided. Reverse-subtract instructions are also provided for the vector-scalar forms.

# Integer adds.
vadd.vv vd, vs2, vs1, vm # Vector-vector
vadd.vx vd, vs2, rs1, vm # vector-scalar
vadd.vi vd, vs2, imm, vm # vector-immediate

# Integer subtract
vsub.vv vd, vs2, vs1, vm # Vector-vector
vsub.vx vd, vs2, rs1, vm # vector-scalar

# Integer reverse subtract
vrsub.vx vd, vs2, rs1, vm # vd[i] = rs1 - vs2[i]
vrsub.vi vd, vs2, imm, vm # vd[i] = imm - vs2[i]

12.2. Vector Widening Integer Add/Subtract

The widening add/subtract instructions are provided in both signed and unsigned variants, depending on whether the narrower source operands are first sign- or zero-extended before forming the double-width sum.

# Widening unsigned integer add/subtract, 2*SEW = SEW +/- SEW
vwaddu.vv vd, vs2, vs1, vm # vector-vector
vwaddu.vx vd, vs2, rs1, vm # vector-scalar
vwsubu.vv vd, vs2, vs1, vm # vector-vector
vwsubu.vx vd, vs2, rs1, vm # vector-scalar

# Widening signed integer add/subtract, 2*SEW = SEW +/- SEW
vwadd.vv vd, vs2, vs1, vm # vector-vector
vwadd.vx vd, vs2, rs1, vm # vector-scalar
vwsub.vv vd, vs2, vs1, vm # vector-vector
vwsub.vx vd, vs2, rs1, vm # vector-scalar

# Widening unsigned integer add/subtract, 2*SEW = 2*SEW +/- SEW
vwaddu.wv vd, vs2, vs1, vm # vector-vector
vwaddu.wx vd, vs2, rs1, vm # vector-scalar
vwsubu.wv vd, vs2, vs1, vm # vector-vector
vwsubu.wx vd, vs2, rs1, vm # vector-scalar

# Widening signed integer add/subtract, 2*SEW = 2*SEW +/- SEW
vwadd.wv vd, vs2, vs1, vm # vector-vector
vwadd.wx vd, vs2, rs1, vm # vector-scalar
vwsub.wv vd, vs2, vs1, vm # vector-vector
vwsub.wx vd, vs2, rs1, vm # vector-scalar

An integer value can be doubled in width using the widening add instructions with a scalar operand of x0. Can define assembly pseudoinstructions vwcvt.x.x.v vd,vs,vm = vwadd.vx vd,vs,x0,vm and vwcvtu.x.x.v vd,vs,vm = vwaddu.vx vd,vs,x0,vm.
12.3. Vector Integer Add-with-Carry / Subtract-with-Borrow Instructions

To support multi-word integer arithmetic, instructions that operate on a carry bit are provided. For each operation (add or subtract), two instructions are provided: one to provide the result (SEW width), and the second to generate the carry output (single bit encoded as a mask boolean).

These instructions are encoded as unmasked instructions (vm=1) and operate on all body elements. Encodings corresponding to the masked versions (vm=0) of these instructions are reserved.

The carry inputs and outputs are represented using the mask register layout as described in Section Mask Register Layout. Due to encoding constraints, the carry input must come from the implicit v0 register, but carry outputs can be written to any vector register that respects the source/destination overlap restrictions below.

```
# Produce sum with carry.

# vd[i] = vs2[i] + vs1[i] + v0[i].LSB
vadc.vvm   vd, vs2, vs1, v0  # Vector-vector

# vd[i] = vs2[i] + x[rs1] + v0[i].LSB
vadc.vxm   vd, vs2, rs1, v0  # Vector-scalar

# vd[i] = vs2[i] + imm + v0[i].LSB
vadc.vim   vd, vs2, imm, v0  # Vector-immediate

# Produce carry out in mask register format

# vd[i] = carry_out(vs2[i] + vs1[i] + v0[i].LSB)
vmadc.vvm   vd, vs2, vs1, v0  # Vector-vector

# vd[i] = carry_out(vs2[i] + x[rs1] + v0[i].LSB)
vmadc.vxm   vd, vs2, rs1, v0  # Vector-scalar

# vd[i] = carry_out(vs2[i] + imm + v0[i].LSB)
vmadc.vim   vd, vs2, imm, v0  # Vector-immediate
```

Because implementing a carry propagation requires executing two instructions with unchanged inputs, destructive accumulations will require an additional move to obtain correct results.

```
# Example multi-word arithmetic sequence, accumulating into v4
vmadc.vvm v1, v4, v8, v0  # Get carry into temp register v1
vadc.vvm v4, v4, v8, v0  # Calc new sum
vmcpy.m v0, v1  # Move temp carry into v0 for next word
```

The subtract with borrow instruction vsbc performs the equivalent function to support long word arithmetic for subtraction. There are no subtract with immediate instructions.
# Produce difference with borrow.

# \( vd[i] = vs2[i] - vs1[i] - v0[i].\text{LSB} \)
vsbc.vvm vd, vs2, vs1, v0  # Vector-vector

# \( vd[i] = vs2[i] - x[rs1] - v0[i].\text{LSB} \)
vsbc.vxm vd, vs2, rs1, v0  # Vector-scalar

# Produce borrow out in mask register format

# \( vd[i] = \text{borrow\_out}(vs2[i] - vs1[i] - v0[i].\text{LSB}) \)
vmsbc.vvm vd, vs2, vs1, v0  # Vector-vector

# \( vd[i] = \text{borrow\_out}(vs2[i] - x[rs1] - v0[i].\text{LSB}) \)
vmsbc.vxm vd, vs2, rs1, v0  # Vector-scalar

For \( \text{vmsbc} \), the borrow is defined to be 1 iff the difference, prior to truncation, is negative.

For \( \text{vadc}\) and \( \text{vsbc}\), an illegal instruction exception is raised if the destination vector register is \( v0 \) and \( \text{LMUL} > 1 \).

This constraint corresponds to the constraint on masked vector operations that overwrite the mask register.

For \( \text{vmadc} \) and \( \text{vmsbc} \), an illegal instruction exception is raised if the destination vector register overlaps a source vector register group and \( \text{LMUL} > 1 \).

### 12.4. Vector Bitwise Logical Instructions

# Bitwise logical operations.

\( \text{vand.vv}\) vd, vs2, vs1, vm  # Vector-vector
\( \text{vand.vx}\) vd, vs2, rs1, vm  # vector-scalar
\( \text{vand.vi}\) vd, vs2, imm, vm  # vector-immediate

\( \text{vor.vv}\) vd, vs2, vs1, vm  # Vector-vector
\( \text{vor.vx}\) vd, vs2, rs1, vm  # vector-scalar
\( \text{vor.vi}\) vd, vs2, imm, vm  # vector-immediate

\( \text{vxor.vv}\) vd, vs2, vs1, vm  # Vector-vector
\( \text{vxor.vx}\) vd, vs2, rs1, vm  # vector-scalar
\( \text{vxor.vi}\) vd, vs2, imm, vm  # vector-immediate

With an immediate of -1, scalar-immediate forms of the \( \text{vxor} \) instruction provide a bitwise NOT operation. This can be provided as an assembler pseudoinstruction \( \text{vnot.v} \).

### 12.5. Vector Single-Width Bit Shift Instructions

A full complement of vector shift instructions are provided, including logical shift left, and logical (zero-extending) and arithmetic (sign-extending) shift right.
# Bit shift operations

```
vsll.vv vd, vs2, vs1, vm  # Vector-vector
vsll.vx vd, vs2, rs1, vm  # vector-scalar
vsll.vi vd, vs2, uimm, vm  # vector-immediate

vsrl.vv vd, vs2, vs1, vm  # Vector-vector
vsrl.vx vd, vs2, rs1, vm  # vector-scalar
vsrl.vi vd, vs2, uimm, vm  # vector-immediate

vsra.vv vd, vs2, vs1, vm  # Vector-vector
vsra.vx vd, vs2, rs1, vm  # vector-scalar
vsra.vi vd, vs2, uimm, vm  # vector-immediate
```

Only the low \(\lg_2(\text{SEW})\) bits are read to obtain the shift amount.

The immediate is treated as an unsigned shift amount, with a maximum shift amount of 31.

## 12.6. Vector Narrowing Integer Right Shift Instructions

The narrowing right shifts extract a smaller field from a wider operand and have both zero-extending \(\text{srl}\) and sign-extending \(\text{sra}\) forms. The shift amount can come from a vector or a scalar \(x\) register or a 5-bit immediate. The low \(\lg_2(2*\text{SEW})\) bits of the vector or scalar shift amount value are used (e.g., the low 6 bits for a \(\text{SEW}=64\)-bit to \(\text{SEW}=32\)-bit narrowing operation). The unsigned immediate form supports shift amounts up to 31 only.

```
# Narrowing shift right logical, \(\text{SEW} = (2*\text{SEW}) \gg \text{SEW}\)
vnslr.vv vd, vs2, vs1, vm  # vector-vector
vnslr.vx vd, vs2, rs1, vm  # vector-scalar
vnslr.vi vd, vs2, uimm, vm  # vector-immediate

# Narrowing shift right arithmetic, \(\text{SEW} = (2*\text{SEW}) \gg \text{SEW}\)
vnslra.vv vd, vs2, vs1, vm  # vector-vector
vnslra.vx vd, vs2, rs1, vm  # vector-scalar
vnslra.vi vd, vs2, uimm, vm  # vector-immediate
```

It could be useful to add support for \(n4\) variants, where the destination is 1/4 width of source.

## 12.7. Vector Integer Comparison Instructions

The following integer compare instructions write 1 to the destination mask register element if the comparison evaluates to true, and 0 otherwise. The destination mask vector is always held in a single vector register, with a layout of elements as described in Section Mask Register Layout.
# Set if equal
vmseq.vv vd, vs2, vs1, vm  # Vector-vector
vmseq.vx vd, vs2, rs1, vm  # vector-scalar
vmseq.vi vd, vs2, imm, vm  # vector-immediate

# Set if not equal
vmsne.vv vd, vs2, vs1, vm  # Vector-vector
vmsne.vx vd, vs2, rs1, vm  # vector-scalar
vmsne.vi vd, vs2, imm, vm  # vector-immediate

# Set if less than, unsigned
vmsltu.vv vd, vs2, vs1, vm  # Vector-vector
vmsltu.vx vd, vs2, rs1, vm  # Vector-scalar

# Set if less than, signed
vmslt.vv vd, vs2, vs1, vm  # Vector-vector
vmslt.vx vd, vs2, rs1, vm  # Vector-scalar

# Set if less than or equal, unsigned
vmsleu.vv vd, vs2, vs1, vm  # Vector-vector
vmsleu.vx vd, vs2, rs1, vm  # Vector-scalar
vmsleu.vi vd, vs2, imm, vm  # Vector-immediate

# Set if less than or equal, signed
vmsle.vv vd, vs2, vs1, vm  # Vector-vector
vmsle.vx vd, vs2, rs1, vm  # Vector-scalar
vmsle.vi vd, vs2, imm, vm  # Vector-immediate

# Set if greater than, unsigned
vmsgtu.vx vd, vs2, rs1, vm  # Vector-scalar
vmsgtu.vi vd, vs2, imm, vm  # Vector-immediate

# Set if greater than, signed
vmsgt.vx vd, vs2, rs1, vm  # Vector-scalar
vmsgt.vi vd, vs2, imm, vm  # Vector-immediate

# Following two instructions are not provided directly
# Set if greater than or equal, unsigned
# vmsgeu.vx vd, vs2, rs1, vm  # Vector-scalar
# Set if greater than or equal, signed
# vmsge.vx vd, vs2, rs1, vm  # Vector-scalar

The following table indicates how all comparisons are implemented in native machine code.
<table>
<thead>
<tr>
<th>Comparison</th>
<th>Assembler Mapping</th>
<th>Assembler Pseudoinstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>va &lt; vb</td>
<td>vmslt{u}.vv vd, va, vb, vm</td>
<td></td>
</tr>
<tr>
<td>va &lt;= vb</td>
<td>vmsle{u}.vv vd, va, vb, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt; vb</td>
<td>vmslt{u}.vv vd, vb, va, vm</td>
<td>vmsgt{u}.vv vd, va, vb, vm</td>
</tr>
<tr>
<td>va &gt;= vb</td>
<td>vmsle{u}.vv vd, vb, va, vm</td>
<td>vmsge{u}.vv vd, va, vb, vm</td>
</tr>
<tr>
<td>va &lt; x</td>
<td>vmslt{u}.vx vd, va, x, vm</td>
<td></td>
</tr>
<tr>
<td>va &lt;= x</td>
<td>vmsle{u}.vx vd, va, x, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt; x</td>
<td>vmsgt{u}.vx vd, va, x, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt;= x</td>
<td>see below</td>
<td></td>
</tr>
<tr>
<td>va &lt; i</td>
<td>vmsle{u}.vi vd, va, i-1, vm</td>
<td>vmslt{u}.vi vd, va, i, vm</td>
</tr>
<tr>
<td>va &lt;= i</td>
<td>vmsle{u}.vi vd, va, i, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt; i</td>
<td>vmsgt{u}.vi vd, va, i, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt;= i</td>
<td>vmsgt{u}.vi vd, va, i-1, vm</td>
<td>vmsge{u}.vi vd, va, i, vm</td>
</tr>
</tbody>
</table>

va, vb vector register groups
x  scalar integer register
i  immediate

The immediate forms of vmslt{u}.vi are not provided as the immediate value can be decreased by 1 and the vmsle{u}.vi variants used instead. The vmsle.vi range is -16 to 15, resulting in an effective vmslt.vi range of -15 to 16. The vmsleu.vi range is 0 to 15 (and (~0)-15 to ~0), giving an effective vmsltu.vi range of 1 to 16 (Note, vmsltu.vi with immediate 0 is not useful as it is always false). Similarly, vmsge{u}.vi is not provided and the comparison is implemented using vmsgt{u}.vi with the immediate decremented by one. The resulting effective vmsge.vi range is -15 to 16, and the resulting effective vmsgeu.vi range is 1 to 16 (Note, vmsgeu.vi with immediate 0 is not useful as it is always true).

The vmsgt forms for register scalar and immediates are provided to allow a single comparison instruction to provide the correct polarity of mask value without using additional mask logical instructions.

To reduce encoding space, the vmsge{u}.vx form is not directly provided, and so the va ≥ x case requires special treatment.

The vmsgt{u}.vx instruction can be synthesized by reducing the value of x by 1 and using the vmsgt{u}.vx instruction, when it is known that this will not underflow the representation in x.

Sequences to synthesize `vmsgt{u}.vx` instruction

va >= x,  x > minimum

    addi t0, x, -1; vmsgt{u}.vx vd, va, t0, vm

The above sequence will usually be the most efficient implementation, but assembler pseudoinstructions can be provided for cases where the range of x is unknown.
unmasked \( va \geq x \)

\[
\text{pseudoinstruction: } \text{vmsge{u}.vx } vd, va, x \\
\text{expansion: } \text{vmslt{u}.vx } vd, va, x; \text{vmando.mm } vd, vd
\]

\( \text{masked } va >= x, vd != v0 \)

\[
\text{pseudoinstruction: } \text{vmsge{u}.vx } vd, va, x, v0.t \\
\text{expansion: } \text{vmslt{u}.vx } vd, va, x, v0.t; \text{vmxor.mm } vd, vd, v0
\]

\( \text{masked } va >= x, \text{any } vd \)

\[
\text{pseudoinstruction: } \text{vmsge{u}.vx } vd, va, x, v0.t, vt \\
\text{expansion: } \text{vmslt{u}.vx } vt, va, x; \text{vmando.mm } vd, vd, vt
\]

The vt argument to the pseudoinstruction must name a temporary vector register that is not same as vd and which will be clobbered by the pseudoinstruction.

Comparisons effectively AND in the mask, e.g,

\[
\#	ext{ (a < b) && (b < c) in two instructions} \\
vmslt.vv \quad v0, va, vb \quad \#\text{ All body elements written} \\
vmslt.vv \quad v0, vb, vc, v0.t \quad \#\text{ Only update at set mask}
\]

For all comparison instructions, an illegal instruction exception is raised if the destination vector register overlaps a source vector register group and LMUL > 1.

### 12.8. Vector Integer Min/Max Instructions

Signed and unsigned integer minimum and maximum instructions are supported.

\[
\#	ext{Unsigned minimum} \\
vminu.vv \quad vd, vs2, vs1, vm \quad \#\text{Vector-vector} \\
vminu.vx \quad vd, vs2, rs1, vm \quad \#\text{vector-scalar}
\]

\[
\#	ext{Signed minimum} \\
vmin.vv \quad vd, vs2, vs1, vm \quad \#\text{Vector-vector} \\
vmin.vx \quad vd, vs2, rs1, vm \quad \#\text{vector-scalar}
\]

\[
\#	ext{Unsigned maximum} \\
vmaxu.vv \quad vd, vs2, vs1, vm \quad \#\text{Vector-vector} \\
vmaxu.vx \quad vd, vs2, rs1, vm \quad \#\text{vector-scalar}
\]

\[
\#	ext{Signed maximum} \\
vmax.vv \quad vd, vs2, vs1, vm \quad \#\text{Vector-vector} \\
vmax.vx \quad vd, vs2, rs1, vm \quad \#\text{vector-scalar}
\]

### 12.9. Vector Single-Width Integer Multiply Instructions

The single-width multiply instructions perform a SEW-bit*SEW-bit multiply and return an SEW-bit-wide result. The \textbf{mulh} versions write the high word of the product to the destination register.
# Signed multiply, returning low bits of product
vmul.vv vd, vs2, vs1, vm  # Vector-vector
vmul.vx vd, vs2, rs1, vm  # vector-scalar

# Signed multiply, returning high bits of product
vmulh.vv vd, vs2, vs1, vm  # Vector-vector
vmulh.vx vd, vs2, rs1, vm  # vector-scalar

# Unsigned multiply, returning high bits of product
vmulhu.vv vd, vs2, vs1, vm  # Vector-vector
vmulhu.vx vd, vs2, rs1, vm  # vector-scalar

# Signed(vs2)-Unsigned multiply, returning high bits of product
vmulhsu.vv vd, vs2, vs1, vm  # Vector-vector
vmulhsu.vx vd, vs2, rs1, vm  # vector-scalar

There is no `vmulhus` opcode to return high half of unsigned-vector * signed-scalar product. The current `vmulh*` opcodes perform simple fractional multiplies, but with no option to scale, round, and/or saturate the result. Can consider changing definition of `vmulh, vmulhu, vmulhsu` to use `vxrm` rounding mode when discarding low half of product. There is no possibility of overflow in this case.

12.10. Vector Integer Divide Instructions

The divide and remainder instructions are equivalent to the RISC-V standard scalar integer multiply/divides, with the same results for extreme inputs.

# Unsigned divide.
vdivu.vv vd, vs2, vs1, vm  # Vector-vector
vdivu.vx vd, vs2, rs1, vm  # vector-scalar

# Signed divide
vdiv.vv vd, vs2, vs1, vm  # Vector-vector
vdiv.vx vd, vs2, rs1, vm  # vector-scalar

# Unsigned remainder
vremu.vv vd, vs2, vs1, vm  # Vector-vector
vremu.vx vd, vs2, rs1, vm  # vector-scalar

# Signed remainder
vrem.vv vd, vs2, vs1, vm  # Vector-vector
vrem.vx vd, vs2, rs1, vm  # vector-scalar

The decision to include integer divide and remainder was contentious. The argument in favor is that without a standard instruction, software would have to pick some algorithm to perform the operation, which would likely perform poorly on some microarchitectures versus others.

There is no instruction to perform a "scalar divide by vector" operation.

12.11. Vector Widening Integer Multiply Instructions

The widening integer multiply instructions return the full 2*SEW-bit product from an SEW-bit*SEW-bit multiply.
# Widening signed-integer multiply
vwmul.vv vd, vs2, vs1, vm # vector-vector
vwmul.vx vd, vs2, rs1, vm # vector-scalar

# Widening unsigned-integer multiply
vwmulu.vv vd, vs2, vs1, vm # vector-vector
vwmulu.vx vd, vs2, rs1, vm # vector-scalar

# Widening signed-unsigned integer multiply
vwmulsu.vv vd, vs2, vs1, vm # vector-vector
vwmulsu.vx vd, vs2, rs1, vm # vector-scalar


The integer multiply-add instructions are destructive and are provided in two forms, one that overwrites the
addend or minuend (vmacc, vnmsac) and one that overwrites the first multiplicand (vmadd, vnmsub).

The low half of the product is added or subtracted from the third operand.

"sac" is intended to be read as "subtract from accumulator". The opcode is "vnmsac" to match the (unfortunately counterintuitive)
floating-point fnmsub instruction definition. Similarly for the "vnmsub" opcode.

# Integer multiply-add, overwrite addend
vmacc.vv vd, vs1, vs2, vm # vd[i] = +(vs1[i] * vs2[i]) + vd[i]
vmacc.vx vd, rs1, vs2, vm # vd[i] = +(x[rs1] * vs2[i]) + vd[i]

# Integer multiply-sub, overwrite minuend
vnmsac.vv vd, vs1, vs2, vm # vd[i] = -(vs1[i] * vs2[i]) + vd[i]
vnmsac.vx vd, rs1, vs2, vm # vd[i] = -(x[rs1] * vs2[i]) + vd[i]

# Integer multiply-add, overwrite multiplicand
vmadd.vv vd, vs1, vs2, vm # vd[i] = (vs1[i] * vd[i]) + vs2[i]
vmadd.vx vd, rs1, vs2, vm # vd[i] = (x[rs1] * vd[i]) + vs2[i]

# Integer multiply-sub, overwrite multiplicand
vnmsub.vv vd, vs1, vs2, vm # vd[i] = -(vs1[i] * vd[i]) + vs2[i]
vnmsub.vx vd, rs1, vs2, vm # vd[i] = -(x[rs1] * vd[i]) + vs2[i]

12.13. Vector Widening Integer Multiply-Add Instructions

The widening integer multiply-add instructions add a SEW-bit*SEW-bit multiply result to (from) a 2*SEW-bit value and produce a 2*SEW-bit result. All combinations of signed and unsigned multiply operands are supported.
# Widening unsigned-integer multiply-add, overwrite addend
vwmaccu.vv vd, vs1, vs2, vm  # vd[i] = +(vs1[i] * vs2[i]) + vd[i]
vwmaccu.vx vd, rs1, vs2, vm  # vd[i] = +(x[rs1] * vs2[i]) + vd[i]

# Widening signed-integer multiply-add, overwrite addend
vwmacc.vv vd, vs1, vs2, vm  # vd[i] = +(vs1[i] * vs2[i]) + vd[i]
vwmacc.vx vd, rs1, vs2, vm  # vd[i] = +(x[rs1] * vs2[i]) + vd[i]

# Widening signed-unsigned-integer multiply-add, overwrite addend
vwmaccsu.vv vd, vs1, vs2, vm  # vd[i] = +(signed(vs1[i]) * unsigned(vs2[i])) + vd[i]
vwmaccsu.vx vd, rs1, vs2, vm  # vd[i] = +(signed(x[rs1]) * unsigned(vs2[i])) + vd[i]

# Widening unsigned-signed-integer multiply-add, overwrite addend
vwmaccus.vx vd, rs1, vs2, vm  # vd[i] = +(unsigned(x[rs1]) * signed(vs2[i])) + vd[i]

## 12.14. Vector Integer Merge Instructions

The vector integer merge instructions combine two source operands based on the mask field. Unlike regular arithmetic instructions, the merge operates on all body elements (i.e., the set of elements from vstart up to the current vector length in vl).

The vmerge instructions are always masked (vm=0). The instructions combine two sources as follows. At elements where the mask value is zero, the first operand is copied to the destination element, otherwise the second operand is copied to the destination element. The first operand is always a vector register group specified by vs2. The second operand is a vector register group specified by vs1 or a scalar x register specified by rs1 or a 5-bit sign-extended immediate.

vmerge.vvm vd, vs2, vs1, v0  # vd[i] = v0[i].LSB ? vs1[i] : vs2[i]
vmerge.vxm vd, vs2, rs1, v0  # vd[i] = v0[i].LSB ? x[rs1] : vs2[i]
vmerge.vim vd, vs2, imm, v0  # vd[i] = v0[i].LSB ? imm    : vs2[i]

## 12.15. Vector Integer Move Instructions

The vector integer move instructions copy a source operand to a vector register group. These instructions are always unmasked (vm=1). The first operand specifier (vs2) must contain v0, and any other vector register number in vs2 is reserved. This instruction copies the vs1, rs1, or immediate operand to the first vl locations of the destination vector register group.

vmv.v.v vd, vs1  # vd[i] = vs1[i]
vmv.v.x vd, rs1  # vd[i] = rs1
vmv.v.i vd, imm  # vd[i] = imm

Mask values can be widened into SEW-width elements using a sequence vmv.v.i vd, 0; vmerge.vim vd, vd, 1, v0. The vector integer move instructions share the encoding with the vector merge instructions, but with vm=1 and vs2=v0.
13. Vector Fixed-Point Arithmetic Instructions

A set of vector arithmetic instructions are provided to support fixed-point arithmetic.

An N-bit element can hold two's-complement signed integers in the range \(-2^{N-1} \ldots +2^{N-1}-1\), and unsigned integers in the range \(0 \ldots +2^N-1\). The fixed-point instructions help preserve precision in narrow operands by supporting scaling and rounding, and can handle overflow by saturating results into the destination format range.

The widening integer operations described above can also be used to remove the possibility of overflow.

13.1. Vector Single-Width Saturating Add and Subtract

Saturating forms of integer add and subtract are provided, for both signed and unsigned integers. If the result would overflow the destination, the result is replaced with the closest representable value, and the \(vxsat\) bit is set.

```assembly
# Saturating adds of unsigned integers.
vsaddu.vv vd, vs2, vs1, vm   # Vector-vector
vsaddu.vx vd, vs2, rs1, vm   # vector-scalar
vsaddu.vi vd, vs2, imm, vm   # vector-immediate

# Saturating adds of signed integers.
vsadd.vv vd, vs2, vs1, vm   # Vector-vector
vsadd.vx vd, vs2, rs1, vm   # vector-scalar
vsadd.vi vd, vs2, imm, vm   # vector-immediate

# Saturating subtract of unsigned integers.
vssubu.vv vd, vs2, vs1, vm   # Vector-vector
vssubu.vx vd, vs2, rs1, vm   # vector-scalar

# Saturating subtract of signed integers.
vssub.vv vd, vs2, vs1, vm   # Vector-vector
vssub.vx vd, vs2, rs1, vm   # vector-scalar
```

13.2. Vector Single-Width Averaging Add and Subtract

The averaging add and subtract instructions right shift the result by one bit and round off the result according to the setting in \(vxrm\). There can be no overflow in the result.
# For vrnm=rnu, round = 1

# Averaging add
# result = (src1 + src2 + round) >> 1;

# Averaging adds of integers.
vaadd.vv vd, vs2, vs1, vm   # Vector-vector
vaadd.vx vd, vs2, rs1, vm   # vector-scalar
vaadd.vi vd, vs2, imm, vm   # vector-immediate

# Averaging subtract
# result = (src1 - src2 + round) >> 1;

# Averaging subtract of integers.
vasub.vv vd, vs2, vs1, vm   # Vector-vector
vasub.vx vd, vs2, rs1, vm   # vector-scalar

13.3. Vector Single-Width Fractional Multiply with Rounding and Saturation

The signed fractional multiply instruction produces a 2*SEW product of the two SEW inputs, then shifts the result right by SEW-1 bits, rounding these bits according to vxrm, then saturates the result to fit into SEW bits. If the result causes saturation, the vxsat bit is set.

```assembly
# Signed saturating and rounding fractional multiply
vsmul.vv vd, vs2, vs1, vm   # vd[i] = clip((vs2[i]*vs1[i]+round)>>(SEW-1))
vsmul.vx vd, vs2, rs1, vm   # vd[i] = clip((vs2[i]*x[rs1]+round)>>(SEW-1))
```

When multiplying two N-bit signed numbers, the largest magnitude is obtained for \(-2^{N-1} \times -2^{N-1}\) producing a result \(2^{2N-2}\), which has a single (zero) sign bit when held in 2N bits. All other products have two sign bits in 2N bits. To retain greater precision in N result bits, the product is shifted right by one bit less than N, saturating the largest magnitude result but increasing result precision by one bit for all other products.

Considering adding vxrm-controlled rounding to vmulhu, vmulhsu, and vmulh to further support fixed-point. These would not have saturation.

13.4. Vector Widening Saturating Scaled Multiply-Add

The widening saturating scaled multiply-add instructions perform an SEW-bit * SEW-bit multiply to yield a 2*SEW-bit product. The product is then right-shifted by SEW/2 bits with the shifted bits rounded off according to vxrm, and the rounded product is added to a 2*SEW-bit destination accumulator, with saturation if the result would overflow the destination accumulator. The vxsat bit is set if any overflow occurs.

If any multiplier operand is signed, then the result is treated as a signed value for overflow/saturation. If both multiplier operands are unsigned then the result is treated as an unsigned value for overflow/saturation.

<table>
<thead>
<tr>
<th>SEW</th>
<th>Product Width</th>
<th>Rounded Product</th>
<th>Accumulator</th>
<th>Guard Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>16</td>
<td>12</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>32</td>
<td>24</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>32</td>
<td>64</td>
<td>48</td>
<td>64</td>
<td>16</td>
</tr>
</tbody>
</table>
# Widening unsigned-integer scaled multiply-accumulate
vwsmaccu.vv vd, vs1, vs2, vm # vd[i] = clipu((+(vs1[i]*vs2[i]+round)>>SEW/2)+vd[i])
vwsmaccu.vx vd, rs1, vs2, vm # vd[i] = clipu((+(x[rs1]*vs2[i]+round)>>SEW/2)+vd[i])

# Widening signed-integer scaled multiply-accumulate
vwsmacc.vv vd, vs1, vs2, vm # vd[i] = clip((+(vs1[i]*vs2[i]+round)>>SEW/2)+vd[i])
vwsmacc.vx vd, rs1, vs2, vm # vd[i] = clip((+(x[rs1]*vs2[i]+round)>>SEW/2)+vd[i])

# Widening signed-unsigned-integer scaled multiply-accumulate
vwsmaccsu.vv vd, vs1, vs2, vm
# vd[i] = clip(((signed(vs1[i])*unsigned(vs2[i])+round)>>SEW/2)+vd[i])
vwsmaccsu.vx vd, rs1, vs2, vm
# vd[i] = clip(((signed(x[rs1])*unsigned(vs2[i])+round)>>SEW/2)+vd[i])

# Widening unsigned-signed-integer scaled multiply-accumulate
vwsmaccus.vx vd, rs1, vs2, vm
# vd[i] = clip(((unsigned(x[rs1])*signed(vs2[i])+round)>>SEW/2)+vd[i])

# For vxrm=rnu, round = (1 << (SEW/2-1))

| An arbitrary scaling/shift amount would be more flexible but would require a fourth source operand.

13.5. Vector Single-Width Scaling Shift Instructions

These instructions shift the input value right, and round off the shifted out bits according to vxrm. The scaling right shifts have both zero-extending (vssrl) and sign-extending (vssra) forms. The low lg2(SEW) bits of the vector or scalar shift amount value are used. The immediate form supports shift amounts up to 31 only.

# For vxrm=rnu, round = 1 << (src2-1), where src2 is vs1[i], x[rs1], uimm
# Scaling shift right logical
vssrl.vv vd, vs2, vs1, vm # vd[i] = ((vs2[i] + round)>>vs1[i])
vssrl.vx vd, vs2, rs1, vm # vd[i] = ((vs2[i] + round)>>x[rs1])
vssrl.vi vd, vs2, uimm, vm # vd[i] = ((vs2[i] + round)>>uimm)

# Scaling shift right arithmetic
vssra.vv vd, vs2, vs1, vm # vd[i] = ((vs2[i] + round)>>vs1[i])
vssra.vx vd, vs2, rs1, vm # vd[i] = ((vs2[i] + round)>>x[rs1])
vssra.vi vd, vs2, uimm, vm # vd[i] = ((vs2[i] + round)>>uimm)

13.6. Vector Narrowing Fixed-Point Clip Instructions

The vnclip instructions are used to pack a fixed-point value into a narrower destination. The instructions support rounding, scaling, and saturation into the final destination format.

The second argument (vector element, scalar value, immediate value) gives the amount to right shift the source as in the narrowing shift instructions, which provides the scaling. The low lg2(2*SEW) bits of the vector or scalar shift amount value are used (e.g., the low 6 bits for a SEW=64-bit to SEW=32-bit narrowing operation). The immediate form supports shift amounts up to 31 only.
# Narrowing unsigned clip
vnclu.vv vd, vs2, vs1, vm    # vector-vector
vnclu.vx vd, vs2, rs1, vm    # vector-scalar
vnclu.vi vd, vs2, uimm, vm   # vector-immediate

# Narrowing signed clip,  vd[i] = clip(round(vs2[i] + rnd) >> vs1[i])
#                           SEW           2*SEW                 SEW
vncl.vv vd, vs2, vs1, vm    # vector-vector
vncl.vx vd, vs2, rs1, vm    # vector-scalar
vncl.vi vd, vs2, uimm, vm   # vector-immediate

For vnclu/vncl, the rounding mode is specified in the vxrm CSR. Rounding occurs around the least-significant bit of the destination and before saturation.

For vnclu, the shifted rounded source value is treated as an unsigned integer and saturates if the result would overflow the destination viewed as an unsigned integer.

For vncl, the shifted rounded source value is treated as a signed integer and saturates if the result would overflow the destination viewed as a signed integer.

If any destination element is saturated, the vxsat bit is set in the vxsat register.
14. Vector Floating-Point Instructions

The standard vector floating-point instructions treat 16-bit, 32-bit, 64-bit, and 128-bit elements as IEEE-754/2008-compatible values. If the current SEW does not correspond to a supported IEEE floating-point type, an illegal instruction exception is raised.

The floating-point element widths that are supported depend on the platform.
Platforms supporting 16-bit half-precision floating-point values will also have to implement scalar half-precision floating-point support in the \( f \) registers.

The vector floating-point instructions have the same behavior as the scalar floating-point instructions with regard to NaNs.
Scalar values for vector-scalar operations can be sourced from the standard scalar \( f \) registers.
Scalar floating-point values will be sourced from the integer \( x \) registers in the proposed Zfnx variant.

14.1. Vector Floating-Point Exception Flags

A vector floating-point exception at any active floating-point element sets the standard FP exception flags in the \( f\)flags register. Inactive elements do not set FP exception flags.

14.2. Vector Single-Width Floating-Point Add/Subtract Instructions

```
# Floating-point add
vfadd.vv vd, vs2, vs1, vm  # Vector-vector
vfadd.vf vd, vs2, rs1, vm   # vector-scalar

# Floating-point subtract
vfrsub.vf vd, vs2, rs1, vm  # scalar-vector
vfrsub.fv vd, vs2, rs1, vm   # Vector-scalar vd[i] = f[rs1] - vs2[i]
```

14.3. Vector Widening Floating-Point Add/Subtract Instructions

```
# Widening FP add/subtract, 2*SEW = SEW +/- SEW
vfwadd.vv vd, vs2, vs1, vm  # vector-vector
vfwadd.vf vd, vs2, rs1, vm   # vector-scalar
vfwsub.vv vd, vs2, vs1, vm  # vector-vector
vfwsub.vf vd, vs2, rs1, vm   # vector-scalar

# Widening FP add/subtract, 2*SEW = 2*SEW +/- SEW
vfwadd.wv vd, vs2, vs1, vm  # vector-vector
vfwadd.wf vd, vs2, rs1, vm   # vector-scalar
vfwsub.wv vd, vs2, vs1, vm  # vector-vector
vfwsub.wf vd, vs2, rs1, vm   # vector-scalar
```

14.4. Vector Single-Width Floating-Point Multiply/Divide Instructions
# Floating-point multiply
vfmul.vv vd, vs2, vs1, vm   # Vector-vector
vfmul.vf vd, vs2, rs1, vm   # vector-scalar

# Floating-point divide
vfdiv.vv vd, vs2, vs1, vm   # Vector-vector
vfdiv.vf vd, vs2, rs1, vm   # vector-scalar

# Reverse floating-point divide vector = scalar / vector
vfdiv.vf vd, vs2, rs1, vm  # scalar-vector, vd[i] = f[rs1]/vs2[i]

14.5. Vector Widening Floating-Point Multiply

# Widening floating-point multiply
vfwmul.vv vd, vs2, vs1, vm # vector-vector
vfwmul.vf vd, vs2, rs1, vm # vector-scalar

14.6. Vector Single-Width Floating-Point Fused Multiply-Add Instructions

All four varieties of fused multiply-add are provided, and in two destructive forms that overwrite one of the operands, either the addend or the first multiplicand.

# FP multiply-accumulate, overwrites addend
vfmacc.vv vd, vs1, vs2, vm   # vd[i] = +(vs1[i] * vs2[i]) + vd[i]
vfmacc.vf vd, rs1, vs2, vm   # vd[i] = +(f[rs1] * vs2[i]) + vd[i]

# FP negate-(multiply-accumulate), overwrites subtrahend
vfnmacc.vv vd, vs1, vs2, vm # vd[i] = -(vs1[i] * vs2[i]) - vd[i]
vfnmacc.vf vd, rs1, vs2, vm # vd[i] = -(f[rs1] * vs2[i]) - vd[i]

# FP multiply-subtract-accumulator, overwrites subtrahend
vfmsac.vv vd, vs1, vs2, vm  # vd[i] = +(vs1[i] * vs2[i]) - vd[i]
vfmsac.vf vd, rs1, vs2, vm  # vd[i] = +(f[rs1] * vs2[i]) - vd[i]

# FP negate-(multiply-subtract-accumulator), overwrites minuend
vfnmmsac.vv vd, vs1, vs2, vm # vd[i] = -(vs1[i] * vs2[i]) + vd[i]
vfnmmsac.vf vd, rs1, vs2, vm # vd[i] = -(f[rs1] * vs2[i]) + vd[i]

# FP multiply-add, overwrites multiplicand
vfmadd.vv vd, vs1, vs2, vm # vd[i] = +(vs1[i] * vd[i]) + vs2[i]
vfmadd.vf vd, rs1, vs2, vm # vd[i] = +(f[rs1] * vd[i]) + vs2[i]

# FP negate-(multiply-add), overwrites multiplicand
vfnmadd.vv vd, vs1, vs2, vm # vd[i] = -(vs1[i] * vd[i]) - vs2[i]
vfnmadd.vf vd, rs1, vs2, vm # vd[i] = -(f[rs1] * vd[i]) - vs2[i]

# FP multiply-sub, overwrites multiplicand
vfmsub.vv vd, vs1, vs2, vm # vd[i] = +(vs1[i] * vd[i]) - vs2[i]
vfmsub.vf vd, rs1, vs2, vm # vd[i] = +(f[rs1] * vd[i]) - vs2[i]

# FP negate-(multiply-sub), overwrites multiplicand
vfnmmsub.vv vd, vs1, vs2, vm # vd[i] = -(vs1[i] * vd[i]) + vs2[i]
vfnmmsub.vf vd, rs1, vs2, vm # vd[i] = -(f[rs1] * vd[i]) + vs2[i]

It would be possible to use the two unused rounding modes in the scalar FP FMA encoding to provide a few non-destructive FMAs. However, this would be the only maskable operation with three inputs and separate output.
14.7. Vector Widening Floating-Point Fused Multiply-Add Instructions

The widening floating-point fused multiply-add instructions all overwrite the wide addend with the result. The multiplier inputs are all SEW wide, while the addend and destination is 2*SEW bits wide.

# FP widening multiply-accumulate, overwrites addend
vfwmacc.vv vd, vs1, vs2, vm    # vd[i] = *(vs1[i] * vs2[i]) + vd[i]
vfwmacc.vf vd, rs1, vs2, vm    # vd[i] = *(f[rs1] * vs2[i]) + vd[i]

# FP widening negate-(multiply-accumulate), overwrites addend
vfwmacc.vv vd, vs1, vs2, vm    # vd[i] = -(vs1[i] * vs2[i]) - vd[i]
vfwmacc.vf vd, rs1, vs2, vm    # vd[i] = -(f[rs1] * vs2[i]) - vd[i]

# FP widening multiply-subtract-accumulator, overwrites addend
vfwmmsac.vv vd, vs1, vs2, vm    # vd[i] = *(vs1[i] * vs2[i]) - vd[i]
vfwmmsac.vf vd, rs1, vs2, vm    # vd[i] = *(f[rs1] * vs2[i]) - vd[i]

# FP widening negate-(multiply-subtract-accumulator), overwrites addend
vfwmmsac.vv vd, vs1, vs2, vm    # vd[i] = -(vs1[i] * vs2[i]) + vd[i]
vfwmmsac.vf vd, rs1, vs2, vm    # vd[i] = -(f[rs1] * vs2[i]) + vd[i]

14.8. Vector Floating-Point Square-Root Instruction

This is a unary vector-vector instruction.

# Floating-point square root
vfsqrt.v vd, vs2, vm    # Vector-vector square root

14.9. Vector Floating-Point MIN/MAX Instructions

The vector floating-point vfmin and vfmax instructions have the same behavior as the corresponding scalar floating-point instructions in version 2.2 of the RISC-V F/D/Q extension.

# Floating-point minimum
vfmin.vv vd, vs2, vs1, vm    # Vector-vector
vfmin.vf vd, vs2, rs1, vm    # vector-scalar

# Floating-point maximum
vfmax.vv vd, vs2, vs1, vm    # Vector-vector
vfmax.vf vd, vs2, rs1, vm    # vector-scalar

14.10. Vector Floating-Point Sign-Injection Instructions

Vector versions of the scalar sign-injection instructions. The result takes all bits except the sign bit from the vector vs2 operands.

vfsgnj.vv vd, vs2, vs1, vm    # Vector-vector
vfsgnj.vf vd, vs2, rs1, vm    # vector-scalar
vfsgnjn.vv vd, vs2, vs1, vm   # Vector-vector
vfsgnjn.vf vd, vs2, rs1, vm   # vector-scalar
vfsgnjx.vv vd, vs2, vs1, vm   # Vector-vector
vfsgnjx.vf vd, vs2, rs1, vm   # vector-scalar
14.11. Vector Floating-Point Compare Instructions

These vector FP compare instructions compare two source operands and write the comparison result to a mask register. The destination mask vector is always held in a single vector register, with a layout of elements as described in Section Mask Register Layout.

The compare instructions follow the semantics of the scalar floating-point compare instructions. vmfeq and vmfne raise the invalid operation exception only on signaling NaN inputs. vmflt, vmfle, vmfgt, and vmfge raise the invalid operation exception on both signaling and quiet NaN inputs.

For all comparison instructions, an illegal instruction exception is raised if the destination vector register overlaps a source vector register group and LMUL > 1.

```assembly
# Compare equal
vmfeq.vv vd, vs2, vs1, vm # Vector-vector
vmfeq.vf vd, vs2, rs1, vm # vector-scalar

# Compare not equal
vmfne.vv vd, vs2, vs1, vm # Vector-vector
vmfne.vf vd, vs2, rs1, vm # vector-scalar

# Compare less than
vmflt.vv vd, vs2, vs1, vm # Vector-vector
vmflt.vf vd, vs2, rs1, vm # vector-scalar

# Compare less than or equal
vmfle.vv vd, vs2, vs1, vm # Vector-vector
vmfle.vf vd, vs2, rs1, vm # vector-scalar

# Compare greater than
vmfgt.vf vd, vs2, rs1, vm # vector-scalar

# Compare greater than or equal
vmfge.vf vd, vs2, rs1, vm # vector-scalar
```

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Assembler Mapping</th>
<th>Assembler pseudoinstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>va &lt; vb</td>
<td>vmflt.vv vd, va, vb, vm</td>
<td></td>
</tr>
<tr>
<td>va &lt;= vb</td>
<td>vmfle.vv vd, va, vb, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt; vb</td>
<td>vmflt.vv vd, vb, va, vm</td>
<td>vmfgt.vv vd, va, vb, vm</td>
</tr>
<tr>
<td>va &gt;= vb</td>
<td>vmfle.vv vd, vb, va, vm</td>
<td>vmfge.vv vd, va, vb, vm</td>
</tr>
<tr>
<td>va &lt; f</td>
<td>vmflt.vf vd, va, f, vm</td>
<td></td>
</tr>
<tr>
<td>va &lt;= f</td>
<td>vmfle.vf vd, va, f, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt; f</td>
<td>vmfgt.vf vd, va, f, vm</td>
<td></td>
</tr>
<tr>
<td>va &gt;= f</td>
<td>vmfge.vf vd, va, f, vm</td>
<td></td>
</tr>
</tbody>
</table>

va, vb vector register groups
f scalar floating-point register

Providing all forms is necessary to correctly handle unordered comparisons for NaNs.
C99 floating-point quiet comparisons can be implemented by masking the signaling comparisons when either input is NaN, as follows. When the comparand is a non-NaN constant, the middle two instructions can be omitted.
# Example of implementing isgreater()

vmfeq.vv v0, va, va        # Only set where A is not NaN.
vmfeq.vv v1, vb, vb        # Only set where B is not NaN.
vmand.mm v0, v0, v1        # Only set where A and B are ordered,
vmfgt.vv v0, va, vb, v0.t  # so only set flags on ordered values.

In the above sequence, it is tempting to mask the second vmfeq instruction and remove the vmand instruction, but this more efficient sequence incorrectly fails to raise the invalid exception when an element of va contains a quiet NaN and the corresponding element in vb contains a signaling NaN.

## 14.12. Vector Floating-Point Classify Instruction

This is a unary vector-vector instruction that operates in the same way as the scalar classify instruction.

vfclass.v vd, vs2, vm   # Vector-vector

The 10-bit mask produced by this instruction is placed in the least-significant bits of the result elements. The instruction is only defined for SEW=16b and above, so the result will always fit in the destination elements.

## 14.13. Vector Floating-Point Merge Instruction

A vector-scalar floating-point merge instruction is provided, which operates on all body elements, from vstart up to the current vector length in vl regardless of mask value.

The vfmerge.vfm instruction is always masked (vm=0). At elements where the mask value is zero, the first vector operand is copied to the destination element, otherwise a scalar floating-point register value is copied to the destination element.

vfmerge.vfm vd, vs2, rs1, v0  # vd[i] = v0[i].LSB ? f[rs1] : vs2[i]

In Zfinx systems, the instruction is identical to vmerge.vx.

## 14.14. Vector Floating-Point Move Instruction

The vector floating-point move instruction splats a floating-point scalar operand to a vector register group. The instruction copies a scalar f register value to all active elements of a vector register group. This instruction is always unmasked (vm=1). The instruction must have the vs2 field set to v0, with all other values for vs2 reserved.

vfmv.v.f vd, rs1  # vd[i] = f[rs1]

In Zfinx systems, the instruction is identical to vmv.v.x.

The vmv.v.f instruction shares the encoding with the vfmerge.vfm instruction, but with vm=1 and vs2=v0.

## 14.15. Single-Width Floating-Point/Integer Type-Convert Instructions

Conversion operations are provided to convert to and from floating-point values and unsigned and signed integers, where both source and destination are SEW wide.
The conversions follow the same rules on exceptional conditions as the scalar conversion instructions. The conversions always use the dynamic rounding mode in `frm`.

### 14.16. Widening Floating-Point/Integer Type-Convert Instructions

A set of conversion instructions are provided to convert between narrower integer and floating-point datatypes to a type of twice the width.

```assembly
vfwcvt.xu.f.v vd, vs2, vm   # Convert float to double-width unsigned integer.
vfwcvt.x.f.v  vd, vs2, vm   # Convert float to double-width signed integer.

vfwcvt.f.xu.v vd, vs2, vm   # Convert unsigned integer to double-width float.
vfwcvt.f.x.v  vd, vs2, vm   # Convert signed integer to double-width float.

vfwcvt.f.f.v vd, vs2, vm   # Convert single-width float to double-width float.
```

These instructions have the same constraints on vector register overlap as other widening instructions (see Widening Vector Arithmetic Instructions).

- A double-width IEEE floating-point value can always represent a single-width integer exactly.
- A double-width IEEE floating-point value can always represent a single-width IEEE floating-point value exactly.
- A full set of floating-point widening conversions are not supported as single instructions, but any widening conversion can be implemented as several doubling steps with equivalent results and no additional exception flags raised.

### 14.17. Narrowing Floating-Point/Integer Type-Convert Instructions

A set of conversion instructions are provided to convert wider integer and floating-point datatypes to a type of half the width.

```assembly
vfncvt.xu.f.v vd, vs2, vm   # Convert double-width float to unsigned integer.
vfncvt.x.f.v  vd, vs2, vm   # Convert double-width float to signed integer.

vfncvt.f.xu.v vd, vs2, vm   # Convert double-width unsigned integer to float.
vfncvt.f.x.v  vd, vs2, vm   # Convert double-width signed integer to float.

vfncvt.f.f.v vd, vs2, vm   # Convert double-width float to single-width float.
```

These instructions have the same constraints on vector register overlap as other narrowing instructions (see Narrowing Vector Arithmetic Instructions).

- A full set of floating-point widening conversions are not supported as single instructions. Conversions can be implemented in several halving steps, with equivalently rounded results and with the same exception flags raised (possibly raised redundantly in multiple steps).
- An integer value can be halved in width using the narrowing integer shift instructions with a shift amount of 0.
15. Vector Reduction Operations

Vector reduction operations take a vector register group of elements and a scalar held in element 0 of a vector register, and perform a reduction using some binary operator, to produce a scalar result in element 0 of a vector register. The scalar input and output operands are held in element 0 of a single vector register, not a vector register group, so any vector register can be the scalar source or destination of a vector reduction regardless of LMUL setting.

Reductions read and write the scalar operand and result into element 0 of a vector register to avoid a loss of decoupling with the scalar processor, and to support future polymorphic use with future types not supported in the scalar unit.

Inactive elements from the source vector register group are excluded from the reduction, but the scalar operand is always included regardless of mask.

The other elements in the destination vector register (0 < index < VLEN/SEW) are left unchanged.

If $vl=0$, no operation is performed and the destination register is not updated.

Traps on vector reduction instructions are always reported with a $vstart$ of 0. Vector reduction operations raise an illegal instruction exception if $vstart$ is non-zero.

The assembler syntax for a reduction operation is `vredop.vs`, where the `.vs` suffix denotes the first operand is a vector register group and the second operand is a scalar stored in element 0 of a vector register.

15.1. Vector Single-Width Integer Reduction Instructions

All operands and results of single-width reduction instructions have the same SEW width. Overflows wrap around on arithmetic sums.

```
# Simple reductions, where [*] denotes all active elements:
vredsum.vs vd, vs2, vs1, vm   # vd[0] = sum( vs1[0] , vs2[*] )
vredmaxu.vs vd, vs2, vs1, vm   # vd[0] = maxu( vs1[0] , vs2[*] )
vredmax.vs vd, vs2, vs1, vm    # vd[0] = max( vs1[0] , vs2[*] )
vredminu.vs vd, vs2, vs1, vm   # vd[0] = minu( vs1[0] , vs2[*] )
vredmin.vs vd, vs2, vs1, vm    # vd[0] = min( vs1[0] , vs2[*] )
vredand.vs vd, vs2, vs1, vm    # vd[0] = and( vs1[0] , vs2[*] )
vredor.vs vd, vs2, vs1, vm     # vd[0] = or( vs1[0] , vs2[*] )
vredxor.vs vd, vs2, vs1, vm    # vd[0] = xor( vs1[0] , vs2[*] )
```

15.2. Vector Widening Integer Reduction Instructions

The unsigned `vwxredsumu.vs` instruction zero-extends the SEW-wide vector elements before summing them, then adds the 2*SEW-width scalar element, and stores the result in a 2*SEW-width scalar element.

The `vwredsum.vs` instruction sign-extends the SEW-wide vector elements before summing them.

```
# Unsigned sum reduction into double-width accumulator
vwredsumu.vs vd, vs2, vs1, vm   # 2*SEW = 2*SEW + sum(zero-extend(SEW))

# Signed sum reduction into double-width accumulator
vwredsum.vs vd, vs2, vs1, vm   # 2*SEW = 2*SEW + sum(sign-extend(SEW))
```

15.3. Vector Single-Width Floating-Point Reduction Instructions
15.3.1. Vector Ordered Single-Width Floating-Point Sum Reduction

The \texttt{vfredosum} instruction must sum the floating-point values in element order, starting with the scalar in \texttt{vs1[0]}--that is, it performs the computation: 
\[
((vs1[0] + vs2[0]) + vs2[1]) + \ldots
\]
\texttt{vs2[vl-1]}, where each addition operates identically to the scalar floating-point instructions in terms of raising exception flags and generating or propagating special values.

The ordered reduction supports compiler autovectorization, while the unordered FP sum allows for faster implementations.

When the operation is masked (\texttt{vm=0}), the masked-off elements do not affect the result or the exception flags.

If no elements are active, no additions are performed, so the scalar in \texttt{vs1[0]} is simply copied to the destination register, without canonicalizing NaN values and without setting any exception flags. This behavior preserves the handling of NaNs, exceptions, and rounding when autovectorizing a scalar summation loop.

15.3.2. Vector Unordered Single-Width Floating-Point Sum Reduction

The unordered sum reduction instruction, \texttt{vfredsum}, provides an implementation more freedom in performing the reduction.

The implementation can produce a result equivalent to a reduction tree composed of binary operator nodes, with the inputs being elements from the source vector register group (\texttt{vs2}) and the source scalar value (\texttt{vs1[0]}). Each operator in the tree accepts two inputs and produces one result. Each operator can either perform an exact addition, or a floating-point addition according to the RISC-V IEEE scalar floating-point specification and with the currently active floating-point dynamic rounding mode. A node where one input is derived only from elements masked-off or beyond the active vector length may either treat that input as IEEE +0.0 of the appropriate SEW or simply copy the other input to its output. The root node in the tree must produce an IEEE result of the appropriate SEW. An implementation is allowed to add an additional IEEE +0.0 to the final result.

The reduction tree structure must be deterministic for a given value in \texttt{vtype} and \texttt{vl}.

As a consequence of this definition, implementations need not propagate NaN payloads through the reduction tree when no elements are active. In particular, if no elements are active and the scalar input is NaN, implementations are permitted to canonicalize the NaN and, if the NaN is signaling, set the invalid exception flag. Implementations are alternatively permitted to pass through the original NaN and set no exception flags, as with \texttt{vfredosum}.

The \texttt{vfredosum} instruction is a valid implementation of the \texttt{vfredsum} instruction.

15.3.3. Vector Single-Width Floating Max and Min Reductions

Floating-point max and min reductions should return the same final value and raise the same exception flags regardless of operation order.

15.4. Vector Widening Floating-Point Reduction Instructions

Widening forms of the sum reductions are provided that read and write a double-width reduction result.
# Simple reductions.
vfwredosum vs vd, vs2, vs1, vm # Ordered sum
vfwredsum vs vd, vs2, vs1, vm # Unordered sum

The reduction of the SEW-width elements is performed as in the single-width reduction case, with the elements in vs2 promoted to 2*SEW bits before adding to the 2*SEW-bit accumulator.
16. Vector Mask Instructions

Several instructions are provided to help operate on mask values held in a vector register.

16.1. Vector Mask-Register Logical Instructions

Vector mask-register logical operations operate on mask registers. The size of one element in a mask register is SEW/LMUL, so these instructions all operate on single vector registers regardless of the setting of the vlmul field in vtype. They do not change the value of vlmul.

As with other vector instructions, the elements with indices less than vstart are unchanged, and vstart is reset to zero after execution. Vector mask logical instructions are always unmasked so there are no inactive elements. Mask elements past v1, the tail elements, are unchanged.

Within a mask element, these instructions perform their operations using only the least-significant bit of each operand and zero-extend the single-bit result to fill the destination mask element.

\[
\begin{align*}
\text{vmmand.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= \text{vs2}[i].\text{LSB} \quad \& \quad \text{vs1}[i].\text{LSB} \\
\text{vmnmnd.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= !\text{(vs2}[i].\text{LSB} \quad \& \quad \text{vs1}[i].\text{LSB}) \\
\text{vmandnot.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= \text{vs2}[i].\text{LSB} \quad \& \quad !\text{vs1}[i].\text{LSB} \\
\text{vmxor.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= \text{vs2}[i].\text{LSB} \quad \& \quad !\text{vs1}[i].\text{LSB} \\
\text{vmor.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= !\text{(vs2}[i].\text{LSB} \quad \& \quad \text{vs1}[i].\text{LSB}) \\
\text{vmnor.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= !\text{(vs2}[i].\text{LSB} \quad \& \quad !\text{vs1}[i].\text{LSB}) \\
\text{vmxnor.mm} & \, \text{vd, vs2, vs1} & \# \text{vd}[i] &= !\text{(vs2}[i].\text{LSB} \quad \& \quad !\text{vs1}[i].\text{LSB})
\end{align*}
\]

Several assembler pseudoinstructions are defined as shorthand for common uses of mask logical operations:

\[
\begin{align*}
\text{vmcpy.m} & \, \text{vd, vs} \Rightarrow \text{vmmand.mm} \, \text{vd, vs, vs} & \# \text{Copy mask register} \\
\text{vmclr.m} & \, \text{vd} \Rightarrow \text{vmxor.mm} \, \text{vd, vd, vd} & \# \text{Clear mask register} \\
\text{vmset.m} & \, \text{vd} \Rightarrow \text{vmxor.mm} \, \text{vd, vd, vs} & \# \text{Set mask register} \\
\text{vmnot.m} & \, \text{vd, vs} \Rightarrow \text{vmmand.mm} \, \text{vd, vs, vs} & \# \text{Invert bits}
\end{align*}
\]

The vmcpy.m instruction is not called vmmv as elsewhere in the architecture mv implies a bitwise copy without interpreting the bits. The vmcpy.m instruction will clear upper bits of the destination mask register to zero regardless of source values in these bits.

The set of eight mask logical instructions can generate any of the 16 possibly binary logical functions of the two input masks:

<table>
<thead>
<tr>
<th>inputs</th>
<th>src1</th>
<th>src2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
The vector mask logical instructions are designed to be easily fused with a following masked vector operation to effectively expand the number of predicate registers by moving values into v0 before use.

16.2. Vector mask population count vpopc

\[
v\text{popc}.m \ rd, \ vs2, \ vm
\]

The source operand is a single vector register holding mask register values as described in Section Mask Register Layout.

The vpopc.m instruction counts the number of mask elements of the active elements of the vector source mask register that have their least-significant bit set, and writes the result to a scalar x register.

The operation can be performed under a mask, in which case only the masked elements are counted.

\[
v\text{popc}.m \ rd, \ vs2, \ v0.t \ # \ x[rd] = \sum_i \ ( vs2[i].\text{LSB} \land \& \land v0[i].\text{LSB} )
\]

Traps on vpopc.m are always reported with a vstart of 0. The vpopc instruction will raise an illegal instruction exception if vstart is non-zero.

16.3. vfirst find-first-set mask bit

\[
v\text{first}.m \ rd, \ vs2, \ vm
\]

The vfirst instruction finds the lowest-numbered active element of the source mask vector that has its LSB set and writes that element’s index to a GPR. If no element has an LSB set, -1 is written to the GPR.

Software can assume that any negative value (highest bit set) corresponds to no element found, as vector lengths will never exceed \(2^{31}\) on any implementation.

Traps on vfirst are always reported with a vstart of 0. The vfirst instruction will raise an illegal instruction exception if vstart is non-zero.
16.4. **vmsbf.m set-before-first mask bit**

\[ \text{vmsbf.m } v_d, v_{s2}, v_m \]

# Example

<table>
<thead>
<tr>
<th>7 6 5 4 3 2 1 0</th>
<th>Element number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1 0 1 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 0 0 0 0 0 1 1</td>
<td>v2 contents</td>
</tr>
<tr>
<td>1 0 0 1 0 1 0 1</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 0</td>
<td>v2</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>1 1 1 1 1 1 1 1</td>
<td>v2</td>
</tr>
<tr>
<td>1 1 0 0 0 0 1 1</td>
<td>v0 vcontents</td>
</tr>
<tr>
<td>1 0 0 1 0 1 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 1 x x x x 1 1</td>
<td>v2 contents</td>
</tr>
</tbody>
</table>

The vmsbf.m instruction takes a mask register as input and writes results to a mask register. The instruction writes a 1 to all active mask elements before the first source element that has a set LSB, then writes a zero to that element and all following active elements. If there is no set bit in the source vector, then all active elements in the destination are written with a 1.

The tail elements in the destination mask register are unchanged.

Traps on vmsbf.m are always reported with a vstart of 0. The vmsbf instruction will raise an illegal instruction exception if vstart is non-zero.

16.5. **vmsif.m set-including-first mask bit**

The vector mask set-including-first instruction is similar to set-before-first, except it also includes the element with a set bit.
vmsif.m vd, vs2, vm

# Example

<table>
<thead>
<tr>
<th>7 6 5 4 3 2 1 0</th>
<th>Element number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1 0 1 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 0 0 0 0 1 1 1</td>
<td>v2 contents</td>
</tr>
<tr>
<td>1 0 0 1 0 1 0 1</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 1</td>
<td>v2</td>
</tr>
<tr>
<td>1 1 0 0 0 0 1 1</td>
<td>v0 vcontents</td>
</tr>
<tr>
<td>1 0 0 1 0 1 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>1 1 x x x x 1 1</td>
<td>v2 contents</td>
</tr>
</tbody>
</table>

The tail elements in the destination mask register are unchanged.

Traps on vmsif.m are always reported with a vstart of 0. The vmsif instruction will raise an illegal instruction exception if vstart is non-zero.

16.6. vmsof.m set-only-first mask bit

The vector mask set-only-first instruction is similar to set-before-first, except it only sets the first element with a bit set, if any.

vmsof.m vd, vs2, vm

# Example

<table>
<thead>
<tr>
<th>7 6 5 4 3 2 1 0</th>
<th>Element number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1 0 1 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 0 0 0 0 1 0 0</td>
<td>v2 contents</td>
</tr>
<tr>
<td>1 0 0 1 0 1 0 1</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 1</td>
<td>v2</td>
</tr>
<tr>
<td>1 1 0 0 0 0 1 1</td>
<td>v0 vcontents</td>
</tr>
<tr>
<td>1 0 0 1 0 1 0 0</td>
<td>v3 contents</td>
</tr>
<tr>
<td>0 1 x x x x 0 0</td>
<td>v2 contents</td>
</tr>
</tbody>
</table>

The tail elements in the destination mask register are unchanged.

Traps on vmsof.m are always reported with a vstart of 0. The vmsof instruction will raise an illegal instruction exception if vstart is non-zero.

16.7. Example using vector mask instructions

The following is an example of vectorizing a data-dependent exit loop.
# char* strcpy(char *dst, const char* src)

strcpy:
mv a2, a0             # Copy dst
li t0, -1             # Infinite AVL

loop:
    vsetvli x0, t0, e8  # Max length vectors of bytes
    vlbuff.v v1, (a1)   # Get src bytes
    csrr t1, v1        # Get number of bytes fetched
    vmseq.vi v0, v1, 0 # Flag zero bytes
    vfirst.m a3, v0    # Zero found?
    add a1, a1, t1     # Bump pointer
    vmsif.m v0, v0     # Set mask up to and including zero byte.
    vsb.v v1, (a2), v0.t # Write out bytes
    add a2, a2, t1     # Bump pointer
    bltz a3, loop      # Zero byte not found, so loop

ret

# char* strncpy(char *dst, const char* src, size_t n)

strncpy:
    mv a3, a0             # Copy dst

loop:
    vsetvli x0, a2, e8    # Vectors of bytes.
    vlbuff.v v1, (a1)    # Get src bytes
    vmseq.vi v0, v1, 0   # Flag zero bytes
    vfirst.m a4, v0      # Zero found?
    vmsif.m v0, v0       # Set mask up to and including zero byte.
    vsb.v v1, (a3), v0.t # Write out bytes
    csrr t1, v1          # Get number of bytes fetched
    sub a2, a2, t1       # Decrement count.
    bgez a4, zero_tail   # Zero remaining bytes.
    add a1, a1, t1       # Bump pointer
    add a3, a3, t1       # Bump pointer
    bnez a2, loop        # Anymore?

zero_tail:
    vsetvli x0, a2, e8,m8 # Vectors of bytes.
    vmv.v.i v0, 0         # Splat zero.

zero_loop:
    vsetvli t1, a2, e8,m8 # Vectors of bytes.
    vsb.v v0, (a3)        # Store zero.
    sub a2, a2, t1        # Decrement count.
    add a3, a3, t1        # Bump pointer
    bnez a2, zero_loop    # Anymore?

ret

16.8. Vector Iota Instruction

The viota.m instruction reads a source vector mask register and writes to each element of the destination vector register group the sum of all the least-significant bits of elements in the mask register whose index is less than the element, e.g., a parallel prefix sum of the mask values.
This instruction can be masked, in which case only the enabled elements contribute to the sum and only the enabled elements are written.

\[
\text{viota.m } v_d, v_{s2}, v_m
\]

# Example

```
Element number 7 6 5 4 3 2 1 0
v2 contents 1 0 0 1 0 0 0 1
viota.m v4, v2 # Unmasked
v4 result 2 2 2 1 1 1 1 0
v0 contents 1 1 1 0 1 0 1 1
v2 contents 1 0 0 1 0 0 0 1
v4 contents 2 3 4 5 6 7 8 9
viota.m v4, v0, v0.t # Masked
v4 results 1 1 1 5 1 7 1 0
```

The result value is zero-extended to fill the destination element if SEW is wider than the result. If the result value would overflow the destination SEW, the least-significant SEW bits are retained.

Traps on \text{viota.m} are always reported with a \text{vstart} of 0, and execution is always restarted from the beginning when resuming after a trap handler. An illegal instruction exception is raised if \text{vstart} is non-zero.

An illegal instruction exception is raised if the destination vector register group overlaps the source vector mask register. If the instruction is masked, an illegal instruction exception is issued if the destination vector register group overlaps v0.

These constraints exist for two reasons. First, to simplify avoidance of WAR hazards in implementations with temporally long vector registers and no vector register renaming. Second, to enable resuming execution after a trap simpler.

The \text{viota.m} instruction can be combined with memory scatter instructions (indexed stores) to perform vector compress functions.
# Compact non-zero elements from input memory array to output memory array
# size_t compact_non_zero(size_t n, const int* in, int* out)
# {
#   size_t i;
#   size_t count = 0;
#   int *p = out;
#   
#   for (i=0; i<n; i++)
#   {
#       const int v = *in++;
#       if (v != 0)
#           *p++ = v;
#   }
#   
#   return (size_t) (p - out);
# }

# a0 = n
# a1 = &in
# a2 = &out

compact_non_zero:
  li a6, 0                      # Clear count of non-zero elements
loop:
  vssetvli a5, a0, e32,m8       # 32-bit integers
  vlw.v v8, (a1)                # Load input vector
  sub a0, a0, a5                # Decrement number done
  slli a5, a5, 2                # Multiply by four bytes
  vmsne.vi v0, v8, 0            # Locate non-zero values
  add a1, a1, a5                # Bump input pointer
  vpopc.m a5, v0                # Count number of elements set in v0
  viota.m v16, v0               # Get destination offsets of active elements
  add a6, a6, a5                # Accumulate number of elements
  vslli.vi v16, v16, 2, v0.t    # Multiply offsets by four bytes
  slli a5, a5, 2                # Multiply number of non-zero elements by four bytes
  vsuxw.v v8, (a2), v16, v0.t   # Scatter using scaled viota results under mask
  add a2, a2, a5                # Bump output pointer
  bnez a0, loop                 # Any more?
  mv a0, a6                     # Return count
  ret

16.9. Vector Element Index Instruction

The \texttt{vid.v} instruction writes each element’s index to the destination vector register group, from 0 to \texttt{vl-1}.

\texttt{vid.v vd, vm} # Write element ID to destination.

The instruction can be masked.

The \texttt{vs2} field of the instruction must be set to \texttt{v0}, otherwise the encoding is \textit{reserved}.

The result value is zero-extended to fill the destination element if SEW is wider than the result. If the result value would overflow the destination SEW, the least-significant SEW bits are retained.

This constraint is to avoid WAR hazards in long vector implementations without register renaming, and to support restart.
Microarchitectures can implement \textit{vid.v} instruction using the same datapath as \textit{viota.m} but with an implicit set mask source.
17. Vector Permutation Instructions

A range of permutation instructions are provided to move elements around within the vector registers.

17.1. Integer Scalar Move Instructions

The integer scalar read/write instructions transfer a single value between a scalar \( x \) register and element \( 0 \) of a vector register. The instructions ignore LMUL and vector register groups.

\[
\begin{align*}
vmv.x.s & \quad rd, \quad vs2 \quad # \quad rd = vs2[0] \quad (rs1=0) \\
vmv.s.x & \quad vd, \quad rs1 \quad # \quad vd[0] = rs1 \quad (vs2=0)
\end{align*}
\]

The \( vmv.x.s \) instruction copies a single SEW-wide element from index \( 0 \) of the source vector register to a destination integer register. If \( SEW > XLEN \), the least-significant \( XLEN \) bits are transferred and the upper \( SEW-XLEN \) bits are ignored. If \( SEW < XLEN \), the value is sign-extended to \( XLEN \) bits.

The \( vmv.s.x \) instruction copies the scalar integer register to element \( 0 \) of the destination vector register. If \( SEW < XLEN \), the least-significant bits are copied and the upper \( XLEN-SEW \) bits are ignored. If \( SEW > XLEN \), the value is sign-extended to \( SEW \) bits. The other elements in the destination vector register (\( 0 < index < VLEN/SEW \)) are unchanged. If \( vstart \geq vl \), no operation is performed and the destination register is not updated.

As a consequence, when \( vl=0 \), no elements are updated in the destination vector register group, regardless of \( vstart \).

The encodings corresponding to the masked versions (\( vm=0 \)) of \( vmv.x.s \) and \( vmv.s.x \) are reserved.

17.2. Floating-Point Scalar Move Instructions

The floating-point scalar read/write instructions transfer a single value between a scalar \( f \) register and element \( 0 \) of a vector register. The instructions ignore LMUL and vector register groups.

\[
\begin{align*}
vmv.f.s & \quad rd, \quad vs2 \quad # \quad rd = vs2[0] \quad (rs1=0) \\
vfmv.s.f & \quad vd, \quad rs1 \quad # \quad vd[0] = rs1 \quad (vs2=0)
\end{align*}
\]

The \( vmv.f.s \) instruction copies a single SEW-wide element from index \( 0 \) of the source vector register to a destination scalar floating-point register. If \( SEW > FLEN \), the least-significant \( FLEN \) bits are transferred and the upper \( SEW-FLEN \) bits are ignored. If \( SEW < FLEN \), the value is NaN-boxed (1-extended) to \( FLEN \) bits.

The \( vmv.s.f \) instruction copies the scalar register to element \( 0 \) of the destination vector register. If \( SEW < FLEN \), the least-significant bits are copied and the upper \( FLEN-SEW \) bits are ignored. If \( SEW > FLEN \), the value is NaN-boxed (1-extended) to \( SEW \) bits. The other elements in the destination vector register (\( 0 < index < VLEN/SEW \)) are unchanged. If \( vstart \geq vl \), no operation is performed and the destination register is not updated.

As a consequence, when \( vl=0 \), no elements are updated in the destination vector register group, regardless of \( vstart \).

The encodings corresponding to the masked versions (\( vm=0 \)) of \( vmv.f.s \) and \( vmv.s.f \) are reserved.

17.3. Vector Slide Instructions

The slide instructions move elements up and down a vector register group.
The slide operations can be implemented much more efficiently than using the arbitrary register gather instruction. Implementations may optimize certain OFFSET values for vslideup and vslidedown. In particular, power-of-2 offsets may operate substantially faster than other offsets.

For all of the vslideup, vslidedown, vslide1up, and vslide1down instructions, if vstart \(\geq v1\), the instruction performs no operation and leaves the destination vector register unchanged.

As a consequence, when \(v1=0\), no elements are updated in the destination vector register group, regardless of vstart.

### 17.3.1. Vector Slideup Instructions

\[
\begin{align*}
&vslideup.vx \ vd, \ vs2, \ rs1, \ vm & \# vd[i+rs1] = vs2[i] \\
&vslideup.vi \ vd, \ vs2, \ uimm[4:0], \ vm & \# vd[i+uimm] = vs2[i]
\end{align*}
\]

For vslideup, the value in \(v1\) specifies the number of destination elements that are written. The start index (OFFSET) for the destination can be either specified using an unsigned integer in the x register specified by \(rs1\), or a 5-bit immediate treated as an unsigned 5-bit quantity.

vslideup behavior for destination elements

OFFSET is amount to slideup, either from x register or a 5-bit immediate

\[
\begin{align*}
0 < i < \max(vstart, \text{OFFSET}) & \quad \text{Unchanged} \\
\max(vstart, \text{OFFSET}) \leq i < v1 & \quad vd[i] = vs2[i-\text{OFFSET}] \text{ if mask enabled,} \\
& \text{unchanged if not} \\
v1 \leq i < \text{VLMAX} & \quad \text{Tail elements, unchanged}
\end{align*}
\]

The destination vector register group for vslideup cannot overlap the vector register group of the source, otherwise an illegal instruction exception is raised.

The non-overlap constraint avoids WAR hazards on the input vectors during execution, and enables restart with non-zero vstart.

### 17.3.2. Vector Slidedown Instructions

\[
\begin{align*}
&vslidedown.vx \ vd, \ vs2, \ rs1, \ vm & \# vd[i] = vs2[i+rs1] \\
&vslidedown.vi \ vd, \ vs2, \ uimm[4:0], \ vm & \# vd[i] = vs2[i+uimm]
\end{align*}
\]

For vslidedown, the value in \(v1\) specifies the number of destination elements that are written.

The start index (OFFSET) for the source can be either specified using an unsigned integer in the x register specified by \(rs1\), or a 5-bit immediate treated as an unsigned 5-bit quantity.

vslidedown behavior for source elements for element \(i\) in slide

\[
\begin{align*}
0 \leq i+\text{OFFSET} < \text{VLMAX} & \quad \text{Read } vs2[i+\text{offset}] \\
\text{VLMAX} \leq i+\text{OFFSET} & \quad \text{Read as } 0
\end{align*}
\]

vslidedown behavior for destination element \(i\) in slide

\[
\begin{align*}
0 < i < vstart & \quad \text{Unchanged} \\
vstart \leq i < v1 & \quad \text{Updated if mask enabled, unchanged if not} \\
v1 \leq i < \text{VLMAX} & \quad \text{Unchanged}
\end{align*}
\]

### 17.3.3. Vector Slide1up
Variants of slide are provided that only move by one element but which also allow a scalar integer value to be inserted at the vacated element position.

\[
\text{vslide1up.vx \ vd, vs2, rs1, vm} # \text{vd}[0]=x[rs1], \ vd[i+1] = vs2[i]
\]

The \text{vslide1up} instruction places the \(x\) register argument at location 0 of the destination vector register group, provided that element 0 is active, otherwise the destination element is unchanged. If \(XLEN < SEW\), the value is zero-extended to \(SEW\) bits. If \(XLEN > SEW\), the least-significant bits are copied over and the high \(SEW-XLEN\) bits are ignored.

The remaining active \(vl-1\) elements are copied over from index \(i\) in the source vector register group to index \(i+1\) in the destination vector register group.

The \(vl\) register specifies how many of the destination vector register elements are written with source values, and all tail elements are unchanged.

\[
\begin{align*}
\text{vslide1up behavior} \\
\text{i < vstart} & \quad \text{unchanged} \\
\text{0 = i = vstart} & \quad \text{vd[i] = x[rs1] if mask enabled, unchanged if not} \\
\text{max(vstart, 1) <= i < vl} & \quad \text{vd[i] = vs2[i-1] if mask enabled, unchanged if not} \\
\text{vl <= i < VLMAX} & \quad \text{unchanged}
\end{align*}
\]

The \text{vslide1up} instruction requires that the destination vector register group does not overlap the source vector register group. Otherwise, an illegal instruction exception is raised.

### 17.3.4. Vector Slide1down Instruction

The \text{vslide1down} instruction copies the first \(vl-1\) active elements values from index \(i+1\) in the source vector register group to index \(i\) in the destination vector register group.

The \(vl\) register specifies how many of the destination vector register elements are written with source values, and all tail elements are unchanged.

\[
\text{vslide1down.vx \ vd, vs2, rs1, vm} # \text{vd}[i] = vs2[i+1], \ vd[vl-1]=x[rs1]
\]

The \text{vslide1down} instruction places the \(x\) register argument at location \(vl-1\) in the destination vector register, provided that element \(vl-1\) is active, otherwise the destination element is unchanged. If \(XLEN < SEW\), the value is zero-extended to \(SEW\) bits. If \(XLEN > SEW\), the least-significant bits are copied over and the high \(SEW-XLEN\) bits are ignored.

\[
\begin{align*}
\text{vslide1down behavior} \\
\text{i < vstart} & \quad \text{unchanged} \\
\text{vstart <= i < vl-1} & \quad \text{vd[i] = vs2[i+1] if mask enabled, unchanged if not} \\
\text{vstart <= i = vl-1} & \quad \text{vd[vl-1] = x[rs1] if mask enabled, unchanged if not} \\
\text{vl <= i < VLMAX} & \quad \text{unchanged}
\end{align*}
\]

The \text{vslide1down} instruction can be used to load values into a vector register without using memory and without disturbing other vector registers. This provides a path for debuggers to modify the contents of a vector register, albeit slowly, with multiple repeated \text{vslide1down} invocations.

### 17.4. Vector Register Gather Instruction
The vector register gather instruction reads elements from a first source vector register group at locations
given by a second source vector register group. The index values in the second vector are treated as unsigned
integers. The source vector can be read at any index < VLMAX regardless of \( v1 \). The number of elements to
write to the destination register is given by \( v1 \), and elements past \( v1 \) in the destination are unchanged. The
operation can be masked.

\[
\text{vrgather.vv } vd, vs2, vs1, vm \# vd[i] = (vs1[i] >= VLMAX) ? 0 : vs2[vs1[i]]; 
\]

If the element indices are out of range \((vs1[i] >= VLMAX)\) then zero is returned for the element value.

Vector-scalar and vector-immediate forms of the register gather are also provided. These read one element
from the source vector at the given index, and write this value to the \( v1 \) elements at the start of the
destination vector register. The index value in the scalar register and the immediate are treated as unsigned
integers.

\[
\text{vrgather.vx } vd, vs2, rs1, vm \# vd[i] = (x[rs1] >= VLMAX) ? 0 : vs2[rs1] \\
vrgather.vi vd, vs2, uimm, vm \# vd[i] = (uimm >= VLMAX) ? 0 : vs2[uimm] 
\]

For any vrgather instruction, the destination vector register group cannot overlap with the source vector
register groups, including the mask register if the operation is masked, otherwise an illegal instruction
exception is raised.

- When SEW=8, only vector elements 0-255 can be referenced.

17.5. Vector Compress Instruction

The vector compress instruction allows elements selected by a vector mask register from a source vector
register group to be packed into contiguous elements at the start of the destination vector register group.

\[
\text{vcompress.vm } vd, vs2, vs1 \# \text{Compress into } vd \text{ elements of } vs2 \text{ where } vs1 \text{ is enabled} 
\]

The vector mask register specified by \( vs1 \) indicates which of the first \( v1 \) elements of vector register group
\( vs2 \) should be extracted and packed into contiguous elements at the beginning of vector register \( vd \). Any
remaining elements of \( vd \) are unchanged.

Example use of vcompress instruction

\[
\begin{array}{cccccccc}
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\
& \text{v0} & \text{v1} & \text{v2} \\
\end{array} \\
\text{vcompress.vm } v2, v1, v0 \\
0 & 0 & 0 & 0 & 8 & 7 & 5 & 2 \\
\]

The destination vector register group cannot overlap the source vector register group or the source vector
mask register, otherwise an illegal instruction exception is raised.

A trap on a vcompress instruction is always reported with a vstart of 0. Executing a vcompress instruction
with a non-zero vstart raises an illegal instruction exception.

- Although possible, vcompress is one of the more difficult instructions to restart with a non-zero vstart, so assumption is
  implementations will choose not do that but will instead restart from element 0. This does mean elements in destination register
  after vstart will already have been updated.
18. Exception Handling

On a trap during a vector instruction (caused by either a synchronous exception or an asynchronous interrupt), the existing *epc CSR is written with a pointer to the errant vector instruction, while the vstart CSR contains the element index that caused the trap to be taken.

We chose to add a vstart CSR to allow resumption of a partially executed vector instruction to reduce interrupt latencies and to simplify forward-progress guarantees. This is similar to the scheme in the IBM 3090 vector facility. To ensure forward progress without the vstart CSR, implementations would have to guarantee an entire vector instruction can always complete atomically without generating a trap. This is particularly difficult to ensure in the presence of strided or scatter/gather operations and demand-paged virtual memory.

18.1. Precise vector traps

Precise vector traps require that:

1. all instructions older than the trapping vector instruction have committed their results
2. no instructions newer than the trapping vector instruction have altered architectural state
3. any operations within the trapping vector instruction affecting result elements preceding the index in the vstart CSR have committed their results
4. no operations within the trapping vector instruction affecting elements at or following the vstart CSR have altered architectural state except if restarting and completing the affected vector instruction will recover the correct state.

We relax the last requirement to allow elements following vstart to have been updated at the time the trap is reported, provided that re-executing the instruction from the given vstart will correctly overwrite those elements.

We assume most supervisor-mode environments will require precise vector traps.

Except where noted above, vector instructions are allowed to overwrite their inputs, and so in most cases, the vector instruction restart must be from the vstart location. However, there are a number of cases where this overwrite is prohibited to enable execution of the vector instructions to be idempotent and hence restartable from any location.

18.2. Imprecise vector traps

Imprecise vector traps are traps that are not precise. In particular, instructions newer than *epc may have committed results, and instructions older than *epc may have not completed execution. Imprecise traps are primarily intended to be used in situations where reporting an error and terminating execution is the appropriate response.

A platform might specify that interrupts are precise while other traps are imprecise. We assume many embedded platforms will only generate imprecise traps for vector instructions on fatal errors, so do not require resumable traps.

18.3. Selectable precise/imprecise traps

Some platforms may choose to provide a privileged mode bit to select between precise and imprecise vector traps. Imprecise mode would run at high-performance but possibly make it difficult to discern error causes, while precise mode would run more slowly, but support debugging of errors albeit with a possibility of not experiencing the same errors as in imprecise mode.
18.4. Swappable traps

Another trap mode can support swappable state in the vector unit, where on a trap, special instructions can save and restore the vector unit microarchitectural state, to allow execution to continue correctly around imprecise traps.

This mechanism is not defined in the base vector ISA.
19. Divided Element Extension (‘Zvediv’)

EDIV is the mostly likely part of the spec to change substantially.

The divided element extension allows each element to be treated as a packed sub-vector of narrower elements. This provides efficient support for some forms of narrow-width and mixed-width arithmetic, and also to allow outer-loop vectorization of short vector and matrix operations. In addition to modifying the behavior of some existing instructions, a few new instructions are provided to operate on vectors when EDIV > 1.

This is written as an extension for now, but could become part of mandatory base in Unix vector profile.

The divided element extension adds a two-bit field, vediv[1:0] to the vtype register.

<table>
<thead>
<tr>
<th>Bits</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLEN-1</td>
<td>vill</td>
<td>Illegal value if set</td>
</tr>
<tr>
<td>XLEN-2:7</td>
<td></td>
<td>Reserved (write 0)</td>
</tr>
<tr>
<td>6:5</td>
<td>vediv[1:0]</td>
<td>Used by EDIV extension</td>
</tr>
<tr>
<td>4:2</td>
<td>vsew[2:0]</td>
<td>Standard element width (SEW) setting</td>
</tr>
<tr>
<td>1:0</td>
<td>vlmul[1:0]</td>
<td>Vector register group multiplier (LMUL) setting</td>
</tr>
</tbody>
</table>

The vediv field encodes the number of ways, EDIV, into which each SEW-bit element is subdivided into equal sub-elements. A vector register group is now considered to hold a vector of sub-vectors.

<table>
<thead>
<tr>
<th>vediv [1:0]</th>
<th>Division EDIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1</td>
<td>(undivided, as in base)</td>
</tr>
<tr>
<td>0 1 2</td>
<td>two equal sub-elements</td>
</tr>
<tr>
<td>1 0 4</td>
<td>four equal sub-elements</td>
</tr>
<tr>
<td>1 1 8</td>
<td>eight equal sub-elements</td>
</tr>
<tr>
<td>SEW</td>
<td>EDIV</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>8b</td>
<td>2</td>
</tr>
<tr>
<td>8b</td>
<td>4</td>
</tr>
<tr>
<td>8b</td>
<td>8</td>
</tr>
<tr>
<td>16b</td>
<td>2</td>
</tr>
<tr>
<td>16b</td>
<td>4</td>
</tr>
<tr>
<td>16b</td>
<td>8</td>
</tr>
<tr>
<td>32b</td>
<td>2</td>
</tr>
<tr>
<td>32b</td>
<td>4</td>
</tr>
<tr>
<td>32b</td>
<td>8</td>
</tr>
<tr>
<td>64b</td>
<td>2</td>
</tr>
<tr>
<td>64b</td>
<td>4</td>
</tr>
<tr>
<td>64b</td>
<td>8</td>
</tr>
<tr>
<td>128b</td>
<td>2</td>
</tr>
<tr>
<td>128b</td>
<td>4</td>
</tr>
<tr>
<td>128b</td>
<td>8</td>
</tr>
<tr>
<td>256b</td>
<td>2</td>
</tr>
<tr>
<td>256b</td>
<td>4</td>
</tr>
<tr>
<td>256b</td>
<td>8</td>
</tr>
</tbody>
</table>

Each implementation defines a minimum size for a sub-element, \( SELEN \), which must be at most 8 bits.

While \( SELEN \) is a fourth implementation-specific parameter, values smaller than 8 would be considered an additional extension.

### 19.1. Instructions not affected by EDIV

The vector start register \( vstart \) and exception reporting continue to work as before.

The vector length \( vl \) control and vector masking continue to operate at the element level.

Vector masking continues to operate at the element level, so sub-elements cannot be individually masked.

Vector load/store and AMO instructions are unaffected by EDIV, and continue to move whole elements.

Vector mask logical operations are unchanged by EDIV setting, and continue to operate on vector registers containing element masks.

Vector mask population count (\( vpopc \)), find-first and related instructions (\( vfirst, vmsbf, vmsif, vmsof \)), \( vmsi \), and element index (\( vio \)) instructions are unaffected by EDIV.

Vector integer bit insert/extract, and integer and floating-point scalar move instruction are unaffected by EDIV.

Vector slide-up/slide-down are unaffected by EDIV.

Vector compress instructions are unaffected by EDIV.

### 19.2. Instructions Affected by EDIV
19.2.1. Regular Vector Arithmetic Instructions under EDIV

Most vector arithmetic operations are modified to operate on the individual sub-elements, so effective SEW is SEW/EDIV and effective vector length is \( vl \times EDIV \). For example, a vector add of 32-bit elements with a \( vl \) of 5 and EDIV of 4, operates identically to a vector add of 8-bit elements with a vector length of 20.

\[
\text{vsetvl} \ t0, \ a0, \ e32, \ m1, \ d4 \quad \# \text{Vectors of 32-bit elements, divided into byte sub-elements}
\]
\[
\text{vadd.vv} \ v1, v2, v3 \quad \# \text{Performs a vector of } 4 \times vl \text{ 8-bit additions.}
\]
\[
\text{vsll.vx} \ v1, v2, x1 \quad \# \text{Performs a vector of } 4 \times vl \text{ 8-bit shifts.}
\]

19.2.2. Vector Add with Carry/Subtract with Borrow Reserved under EDIV>1

For EDIV > 1, \text{vadc}, \text{vmadc}, \text{vsbc}, \text{vmsbc} are reserved.

19.2.3. Vector Reduction Instructions under EDIV

Vector single-width integer sum reduction instructions are reserved under EDIV>1. Other vector single-width reductions and vector widening integer sum reduction instructions now operate independently on all elements in a vector, reducing sub-element values within an element to an element-wide result.

The scalar input is taken from the least-significant bits of the second operand, with the number of bits equal to the number of significant result bits (i.e., for sum and dot reductions, the number of bits are given in table above, for non-sum and non-dot reductions, equal to the element size).

\[
\# \text{Sum each sub-vector of four bytes into a 16-bit result.}
\text{vsetvl} \ t0, \ a0, \ e32, d4 \quad \# \text{Vectors of 32-bit elements, divided into byte sub-elements}
\text{vwredsum.vs} \ v1, v2, v3 \quad \# v1[i][15:0] = v2[i][31:24] + v2[i][23:16]
\quad \# \quad + v2[i][15:8] + v2[i][7:0] + v3[i][15:0]
\]

\[
\# \text{Find maximum among sub-elements}
\text{vredmax.vs} \ v5, v6, v7 \quad \# v5[i][7:0] = \text{max}(v6[i][31:24], v6[i][23:16], \text{v6[i][15:8], v6[i][7:0], v7[i][7:0]})
\]

Integer sub-element non-sum reductions produce a final result that is max(8,SEW/EDIV) bits wide, sign- or zero-extended to full SEW if necessary.

Integer sub-element widening sum reductions produce a final result that is max(8,min(SEW,2*SEW/EDIV)) bits wide, sign- or zero-extended to full SEW if necessary.

Single-width floating-point reductions produce a final result that is SEW/EDIV bits wide.

Widening floating-point sum reductions produce a final result that is min(2*SEW/EDIV,FLEN) bits wide, NaN-boxed to the full SEW width if necessary.

19.2.4. Vector Register Gather Instructions under EDIV

Vector register gather instructions under non-zero EDIV only gather sub-elements within the element. The source and index values are interpreted as relative to the enclosing element only. Index values \( \geq \) EDIV write a zero value into the result sub-element.
Vector register gathers with scalar or immediate arguments can "splat" values across sub-elements within an element. Implementations can provide fast implementations of register gathers constrained within a single element width.

19.3. Vector Integer Dot-Product Instruction

The integer dot-product reduction vdot.vv performs an element-wise multiplication between the source sub-elements then accumulates the results into the destination vector element. Note the assembler syntax uses a .vv suffix since both inputs are vectors of elements.

Sub-element integer dot reductions produce a final result that is $\max(8, \min(\text{SEW}, 4*\text{SEW}/\text{EDIV}))$ bits wide, sign- or zero-extended to full SEW if necessary.

```assembly
# Unsigned dot-product
vdotu.vv vd, vs2, vs1, vm # Vector-vector

# Signed dot-product
vdot.vv vd, vs2, vs1, vm # Vector-vector

# Dot product, SEW=32, EDIV=1
vdot.vv vd, vs2, vs1, vm # vd[i][31:0] += vs2[i][31:0] * vs1[i][31:0]

# Dot product, SEW=32, EDIV=2
vdot.vv vd, vs2, vs1, vm # vd[i][31:0] += vs2[i][31:16] * vs1[i][31:16] + vs2[i][15:0] * vs1[i][15:0]

# Dot product, SEW=32, EDIV=4
```

19.4. Vector Floating-Point Dot Product Instruction

The floating-point dot-product reduction vfdot.vv performs an element-wise multiplication between the source sub-elements then accumulates the results into the destination vector element. Note the assembler syntax uses a .vv suffix since both inputs are vectors of elements.

```assembly
# Signed dot-product
vfdot.vv vd, vs2, vs1, vm # Vector-vector
```
# Dot product. SEW=32, EDIV=2
vfdot.v vd, vs2, vs1, vm # vd[i][31:0] += vs2[i][31:16] * vs1[i][31:16] + vs2[i][15:0] * vs1[i][15:0]

# Floating-point sub-vectors of two half-precision floats packed into 32-bit elements.
vsetvli t0, a0, e32,m1,d2 # Vectors of 32-bit elements, divided into 16b sub-elements
vfdot.vv v1, v2, v3 # v1[i][31:0] += v2[i][31:16]*v3[i][31:16] + v2[i][16:0]*v3[i][16:0]

# Floating-point sub-vectors of four half-precision floats packed into 64-bit elements.
vsetvli t0, a0, e64,m1,d4 # Vectors of 64-bit elements, divided into 16b sub-elements
vfdot.vv v1, v2, v3 # v1[i][31:0] += v2[i][31:16]*v3[i][31:16] + v2[i][16:0]*v3[i][16:0] + v2[i][63:48]*v3[i][63:48] + v2[i][47:32]*v3[i][47:32];
# v1[i][63:32] = ~0 (NaN boxing)
## 20. Vector Instruction Listing

<table>
<thead>
<tr>
<th>Integer</th>
<th>Integer</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>funct3</td>
<td>funct3</td>
<td>funct3</td>
</tr>
<tr>
<td>OPIVV</td>
<td>V</td>
<td>OPMVV</td>
</tr>
<tr>
<td>OPIVX</td>
<td>X</td>
<td>OPMVX</td>
</tr>
<tr>
<td>OPIVI</td>
<td>I</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>funct6</th>
<th>funct6</th>
<th>funct6</th>
</tr>
</thead>
<tbody>
<tr>
<td>000000 V X I</td>
<td>vadd</td>
<td>000000 V vredsum</td>
</tr>
<tr>
<td>000001</td>
<td>V</td>
<td>vredand</td>
</tr>
<tr>
<td>000010 V X</td>
<td>vsub</td>
<td>000010 V vredor</td>
</tr>
<tr>
<td>000011 V X I</td>
<td>vrsb</td>
<td>000011 V vredxor</td>
</tr>
<tr>
<td>000100 V X</td>
<td>vminu</td>
<td>000100 V vredminu</td>
</tr>
<tr>
<td>000101 V X</td>
<td>vmin</td>
<td>000101 V vredmin</td>
</tr>
<tr>
<td>000110 V X</td>
<td>vmaxu</td>
<td>000110 V vredmaxu</td>
</tr>
<tr>
<td>000111 V X</td>
<td>vmax</td>
<td>000111 V vredmax</td>
</tr>
<tr>
<td>001000</td>
<td></td>
<td>001000 V F vffnj</td>
</tr>
<tr>
<td>001001 V X I</td>
<td>vand</td>
<td>001001 V F vffnjn</td>
</tr>
<tr>
<td>001010 V X I</td>
<td>vor</td>
<td>001010 V F vffnjj</td>
</tr>
<tr>
<td>001011 V X I</td>
<td>vxor</td>
<td>001011 V F vffnjjx</td>
</tr>
<tr>
<td>001100 V X I</td>
<td>vrgather</td>
<td>001100 V F vffnjj</td>
</tr>
<tr>
<td>001101</td>
<td></td>
<td>001101 V F vffnjj</td>
</tr>
<tr>
<td>001110 X I</td>
<td>vslideup</td>
<td>001110 X vslideup</td>
</tr>
<tr>
<td>001111 X I</td>
<td>vsliedown</td>
<td>001111 X vslide1down</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>funct6</th>
<th>funct6</th>
<th>funct6</th>
</tr>
</thead>
<tbody>
<tr>
<td>010000 V X I</td>
<td>vadc</td>
<td>010000 V VWXUNARY0</td>
</tr>
<tr>
<td>010000</td>
<td>X</td>
<td>VRXUNARY0</td>
</tr>
<tr>
<td>010001 V X I</td>
<td>vmadc</td>
<td>010001 V F vffnj</td>
</tr>
<tr>
<td>010010 V X</td>
<td>vsbc</td>
<td>010010 V F vffnjj</td>
</tr>
<tr>
<td>010011 V X</td>
<td>vmsbc</td>
<td>010011 V F vffnjj</td>
</tr>
<tr>
<td>010100</td>
<td></td>
<td>010100 V VMUNARY0</td>
</tr>
<tr>
<td>010101</td>
<td></td>
<td>010101 V F vffnjj</td>
</tr>
<tr>
<td>010110</td>
<td></td>
<td>010110 V F vffnjj</td>
</tr>
<tr>
<td>010111 V X I</td>
<td>vmerge/vmv</td>
<td>010111 V F vffnjj</td>
</tr>
<tr>
<td>011000 V X I</td>
<td>vmseq</td>
<td>011000 V F vffnjj</td>
</tr>
<tr>
<td>011001 V X I</td>
<td>vmsne</td>
<td>011001 V F vffnjj</td>
</tr>
<tr>
<td>011010 V X</td>
<td>vmsltu</td>
<td>011010 V F vffnjj</td>
</tr>
<tr>
<td>011011 V X</td>
<td>vmslt</td>
<td>011011 V F vffnjj</td>
</tr>
<tr>
<td>011100 V X I</td>
<td>vmsleu</td>
<td>011100 V F vffnjj</td>
</tr>
<tr>
<td>011101 V X I</td>
<td>vmsle</td>
<td>011101 V F vffnjj</td>
</tr>
<tr>
<td>011110 X I</td>
<td>vmsgtu</td>
<td>011110 V F vffnjj</td>
</tr>
<tr>
<td>011111 X I</td>
<td>vmsgt</td>
<td>011111 V F vffnjj</td>
</tr>
<tr>
<td>funct6</td>
<td>funct6</td>
<td>funct6</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>100000 V X I vsaddu</td>
<td>100000 V X vdivu</td>
<td>100000 V F vfdiv</td>
</tr>
<tr>
<td>100001 V X I vsadd</td>
<td>100001 V X vdiv</td>
<td>100001 F vfrdiv</td>
</tr>
<tr>
<td>100010 V X vssubu</td>
<td>100010 V X vremu</td>
<td>100010 V VFUNARY0</td>
</tr>
<tr>
<td>100011 V X vssub</td>
<td>100011 V X vrem</td>
<td>100011 V VFUNARY1</td>
</tr>
<tr>
<td>100100 V X I vaadd</td>
<td>100100 V X vmulhu</td>
<td>100100 V F vfmul</td>
</tr>
<tr>
<td>100101 V X I vsll</td>
<td>100101 V X vmul</td>
<td>100101</td>
</tr>
<tr>
<td>100110 V X vasub</td>
<td>100110 V X vmulhsu</td>
<td>100110</td>
</tr>
<tr>
<td>100111 V X vsmul</td>
<td>100111 V X vmulh</td>
<td>100111 F vfrsub</td>
</tr>
<tr>
<td>110000 V vwredsumu</td>
<td>110000 V vwaddu</td>
<td>110000 V F vfwadd</td>
</tr>
<tr>
<td>110001 V vwredsum</td>
<td>110001 V vwadd</td>
<td>110001 V F vfwredsum</td>
</tr>
<tr>
<td>110010</td>
<td>110010 V X vssubu</td>
<td>110010 V F vfwsub</td>
</tr>
<tr>
<td>110011</td>
<td>110011 V X vssub</td>
<td>110011 V F vfnmsub</td>
</tr>
<tr>
<td>110100</td>
<td>110100 V X vssra</td>
<td>110100 V F vfmacc</td>
</tr>
<tr>
<td>110101</td>
<td>110101 V X vsrl</td>
<td>110101 V F vfnmsub</td>
</tr>
<tr>
<td>110110</td>
<td>110110 V X vnssra</td>
<td>110110 V F vfnmsub</td>
</tr>
<tr>
<td>110111</td>
<td>110111 V X vnssra</td>
<td>110111 V F vfnmsub</td>
</tr>
<tr>
<td>110110</td>
<td>110110 V X vnsrl</td>
<td>110110 V F vfnmsub</td>
</tr>
<tr>
<td>110111</td>
<td>110111 V X vnsra</td>
<td>110111 V F vfnmsub</td>
</tr>
<tr>
<td>111000 V vdotu</td>
<td>111000 V X vwmulu</td>
<td>111000 V F vfwmul</td>
</tr>
<tr>
<td>111001 V vdot</td>
<td>111001 V X vwmulsu</td>
<td>111001 V F vfwmul</td>
</tr>
<tr>
<td>111010</td>
<td>111010 V X vwmulsu</td>
<td>111010</td>
</tr>
<tr>
<td>111011</td>
<td>111011 V X vwmul</td>
<td>111011</td>
</tr>
<tr>
<td>111100 V X vwsraccu</td>
<td>111100 V X vwsraccu</td>
<td>111100 V F vfwmsacc</td>
</tr>
<tr>
<td>111101 V X vwsracc</td>
<td>111101 V X vwsracc</td>
<td>111101 V F vfwmsacc</td>
</tr>
<tr>
<td>111110 V X vwsraccus</td>
<td>111110 V X vwsraccus</td>
<td>111110 V F vfwmsacc</td>
</tr>
<tr>
<td>111111 V X vwsraccsus</td>
<td>111111 V X vwsraccsus</td>
<td>111111 V F vfwmsacc</td>
</tr>
</tbody>
</table>
### Table 16. VRXUNARY0 encoding space

<table>
<thead>
<tr>
<th>vs2</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000</td>
<td>vmv.s.x</td>
</tr>
</tbody>
</table>

### Table 17. VWXUNARY0 encoding space

<table>
<thead>
<tr>
<th>vs1</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000</td>
<td>vmv.x.s</td>
</tr>
<tr>
<td>10000</td>
<td>vpopc</td>
</tr>
<tr>
<td>10001</td>
<td>vfirst</td>
</tr>
</tbody>
</table>

### Table 18. VRFUNARY0 encoding space

<table>
<thead>
<tr>
<th>vs2</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000</td>
<td>vfmv.s.f</td>
</tr>
</tbody>
</table>

### Table 19. VWFUNARY0 encoding space

<table>
<thead>
<tr>
<th>vs1</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000</td>
<td>vfmv.f.s</td>
</tr>
</tbody>
</table>

### Table 20. VFUNARY0 encoding space

<table>
<thead>
<tr>
<th>vs1</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-width converts</td>
<td></td>
</tr>
<tr>
<td>00000</td>
<td>vfcvt.xu.f.v</td>
</tr>
<tr>
<td>00001</td>
<td>vfcvt.x.f.v</td>
</tr>
<tr>
<td>00010</td>
<td>vfcvt.f.xu.v</td>
</tr>
<tr>
<td>00011</td>
<td>vfcvt.f.x.v</td>
</tr>
<tr>
<td>widening converts</td>
<td></td>
</tr>
<tr>
<td>01000</td>
<td>vfwcvt.xu.f.v</td>
</tr>
<tr>
<td>01001</td>
<td>vfwcvt.x.f.v</td>
</tr>
<tr>
<td>01010</td>
<td>vfwcvt.f.xu.v</td>
</tr>
<tr>
<td>01011</td>
<td>vfwcvt.f.x.v</td>
</tr>
<tr>
<td>01100</td>
<td>vfwcvt.f.f.v</td>
</tr>
<tr>
<td>narrowing converts</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>vfncvt.xu.f.v</td>
</tr>
<tr>
<td>10001</td>
<td>vfncvt.x.f.v</td>
</tr>
<tr>
<td>10010</td>
<td>vfncvt.f.xu.v</td>
</tr>
<tr>
<td>10011</td>
<td>vfncvt.f.x.v</td>
</tr>
<tr>
<td>10100</td>
<td>vfncvt.f.f.v</td>
</tr>
</tbody>
</table>

### Table 21. VFUNARY1 encoding space

<table>
<thead>
<tr>
<th>vs1</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000</td>
<td>vfsqrt.v</td>
</tr>
<tr>
<td>10000</td>
<td>vfclass.v</td>
</tr>
</tbody>
</table>
Table 22. VMUNARY0 encoding space

<table>
<thead>
<tr>
<th>vs1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00001</td>
<td>vmsbf</td>
</tr>
<tr>
<td>00010</td>
<td>vmsof</td>
</tr>
<tr>
<td>00011</td>
<td>vmsif</td>
</tr>
<tr>
<td>10000</td>
<td>viota</td>
</tr>
<tr>
<td>10001</td>
<td>vid</td>
</tr>
</tbody>
</table>
Appendix A: Vector Assembly Code Examples

The following are provided as non-normative text to help explain the vector ISA.

A.1. Vector-vector add example

```c
#define vector-vector add routine of 32-bit integers
#define void vvaddint32(size_t n, const int*x, const int*y, int*z)
#define { for (size_t i=0; i<n; i++) { z[i]=x[i]+y[i]; } }

#define a0 = n, a1 = x, a2 = y, a3 = z
#define Non-vector instructions are indented
vvaddint32:
    vsetvli t0, a0, e32 # Set vector length based on 32-bit vectors
    vlw.v v0, (a1)           # Get first vector
    sub a0, a0, t0         # Decrement number done
    slli t0, t0, 2         # Multiply number done by 4 bytes
    add a1, a1, t0         # Bump pointer
    vlw.v v1, (a2)           # Get second vector
    add a2, a2, t0         # Bump pointer
    vadd.vv v2, v0, v1        # Sum vectors
    vsw.v v2, (a3)           # Store result
    add a3, a3, t0         # Bump pointer
    bnez a0, vvaddint32    # Loop back
    ret                    # Finished
```

A.2. Example with mixed-width mask and compute.

```c
#define Code using one width for predicate and different width for masked
#define compute.
#define int8_t a[]; int32_t b[], c[];
#define for (i=0; i<n; i++) { b[i] = (a[i] < 5) ? c[i] : 1; }

#define Mixed-width code that keeps SEW/LMUL=8
loop:
    vsetvli a4, a0, e32 # Byte vector for predicate calc
    vlw.v v1, (a1)                # Load a[i]
    add a1, a1, a4              # Bump pointer.
    vmslt.vi v0, v1, 5            # a[i] < 5?
    vmslt.vi v0, v3, 5            # Splat immediate to destination
    vsw.v v2, (a3)                # Store b[i].
    add a3, a3, t1              # Bump pointer.
    bnez a0, loop               # Any more?
```

A.3. Memcpy example
# void *memcpy(void* dest, const void* src, size_t n)
# a0=dest, a1=src, a2=n
#
memcpy:
    mv a3, a0 # Copy destination
loop:
    vsetvli t0, a2, e8, m8 # Vectors of 8b
    vl.b v0, (a1) # Load bytes
        add a1, a1, t0 # Bump pointer
        sub a2, a2, t0 # Decrement count
    vsb.v v0, (a3) # Store bytes
        add a3, a3, t0 # Bump pointer
        bnez a2, loop # Any more?
ret # Return

A.4. Conditional example

# (int16) z[i] = ((int8) x[i] < 5) ? (int16) a[i] : (int16) b[i];
#
# Fixed 16b SEW:
#
loop:
    vsetvli t0, a0, e16 # Use 16b elements.
    vl.b v0, (a1) # Get x[i], sign-extended to 16b
        sub a0, a0, t0 # Decrement element count
        add a1, a1, t0 # x[i] Bump pointer
    vmslt.vi v0, v0, 5 # Set mask in v0
        slli t0, t0, 1 # Multiply by 2 bytes
    vlh.v v1, (a2), v0.t # z[i] = a[i] case
    vmnot.m v0, v0 # Invert v0
        add a2, a2, t0 # a[i] bump pointer
    vlh.v v1, (a3), v0.t # z[i] = b[i] case
        add a3, a3, t0 # b[i] bump pointer
    vsh.v v1, (a4) # Store z
        add a4, a4, t0 # b[i] bump pointer
    bnez a0, loop # Any more?

A.5. SAXPY example
# void
# saxpy(size_t n, const float a, const float *x, float *y)
# {
#   size_t i;
#   for (i=0; i<n; i++)
#     y[i] = a * x[i] + y[i];
# }
#
# register arguments:
#     a0      n
#     fa0     a
#     a1      x
#     a2      y

saxpy:
  vsetvli a4, a0, e32, m8
  vlw.v v0, (a1)
  sub a0, a0, a4
  slli a4, a4, 2
  add a1, a1, a4
  vlw.v v8, (a2)
  vfmacc.vf v8, fa0, v0
  vsw.v v8, (a2)
  add a2, a2, a4
  bnez a0, saxpy
  ret

A.6. SGEMM example
RV64IDV system

void

sgemm_nn(size_t n,
          size_t m,
          size_t k,
          const float*a,   // m * k matrix
          size_t lda,
          const float*b,   // k * n matrix
          size_t ldb,
          float*c,         // m * n matrix
          size_t ldc)

c += a*b (alpha=1, no transpose on input matrices)
matrices stored in C row-major order

#define n a0
#define m a1
#define k a2
#define ap a3
#define astride a4
#define bp a5
#define bstride a6
#define cp a7
#define cstride t0
#define kt t1
#define nt t2
#define bnp t3
#define cnp t4
#define akp t5
#define bkp s0
#define nvl s1
#define ccp s2
#define amp s3

#define ft12 fa0
#define ft13 fa1
#define ft14 fa2
#define ft15 fa3

This version holds a 16*VLMAX block of C matrix in vector registers
in inner loop, but otherwise does not cache or TLB tiling.

sgemm_nn:
  addi sp, sp, -FRAMESIZE
  sd s0, OFFSET(sp)
  sd s1, OFFSET(sp)
  sd s2, OFFSET(sp)

  # Check for zero size matrices
  beqz n, exit
  beqz m, exit
  beqz k, exit

  # Convert elements strides to byte strides.
  ld cstride, OFFSET(sp)  # Get arg from stack frame
  slli astride, astride, 2
slli bstride, bstride, 2
slli cstride, cstride, 2

slti t6, m, 16
bnez t6, end_rows

c_row_loop: # Loop across rows of C blocks

mv nt, n  # Initialize n counter for next row of C blocks
mv bnp, bp # Initialize B n-loop pointer to start
mv cnp, cp # Initialize C n-loop pointer

c_col_loop: # Loop across one row of C blocks
vsetvli nvl, nt, e32  # 32-bit vectors, LMUL=1

mv akp, ap   # reset pointer into A to beginning
mv bkp, bnp # step to next column in B matrix

# Initialize current C submatrix block from memory.
v lw.v  v0, (cnp); add ccp, cnp, cstride;
v lw.v  v1, (cpp); add ccp, cpp, cstride;
v lw.v  v2, (cpp); add ccp, cpp, cstride;
v lw.v  v3, (cpp); add ccp, cpp, cstride;
v lw.v  v4, (cpp); add ccp, cpp, cstride;
v lw.v  v5, (cpp); add ccp, cpp, cstride;
v lw.v  v6, (cpp); add ccp, cpp, cstride;
v lw.v  v7, (cpp); add ccp, cpp, cstride;
v lw.v  v8, (cpp); add ccp, cpp, cstride;
v lw.v  v9, (cpp); add ccp, cpp, cstride;
v lw.v v10, (cpp); add ccp, cpp, cstride;
v lw.v v11, (cpp); add ccp, cpp, cstride;
v lw.v v12, (cpp); add ccp, cpp, cstride;
v lw.v v13, (cpp); add ccp, cpp, cstride;
v lw.v v14, (cpp); add ccp, cpp, cstride;
v lw.v v15, (cpp)

mv kt, k # Initialize inner loop counter

# Inner loop scheduled assuming 4-clock occupancy of vfmacc instruction and single-issue
# Software pipeline loads
flw ft0, (akp); add amp, akp, astride;
flw ft1, (amp); add amp, amp, astride;
flw ft2, (amp); add amp, amp, astride;
flw ft3, (amp); add amp, amp, astride;

# Get vector from B matrix
v lw.v v16, (bkp)

# Loop on inner dimension for current C block
k_loop:
vfmacc.vf v0, ft0, v16
add bkp, bkp, bstride
flw ft4, (amp)
add amp, amp, astride
vfmacc.vf v1, ft1, v16
addi kt, kt, -1    # Decrement k counter
flw ft5, (amp)
add amp, amp, astride
vfmacc.vf v2, ft2, v16
flw ft6, (amp)
add amp, amp, astride
flw ft7, (amp)
vfmacc.vf v3, ft3, v16
add amp, amp, astride
flw ft8, (amp)
add amp, amp, astride
vfmacc.vf v4, ft4, v16
flw ft9, (amp)
add amp, amp, astride
vfmacc.vf v5, ft5, v16
flw ft10, (amp)
add amp, amp, astride
vfmacc.vf v6, ft6, v16
flw ft11, (amp)
add amp, amp, astride
vfmacc.vf v7, ft7, v16
flw ft12, (amp)
add amp, amp, astride
vfmacc.vf v8, ft8, v16
flw ft13, (amp)
add amp, amp, astride
vfmacc.vf v9, ft9, v16
flw ft14, (amp)
add amp, amp, astride
vfmacc.vf v10, ft10, v16
flw ft15, (amp)
add amp, amp, astride
addi akp, akp, 4
vfmacc.vf v11, ft11, v16
beqz kt, 1f
flw ft0, (akp)
add amp, akp, astride
1:  vfmacc.vf v12, ft12, v16
beqz kt, 1f
flw ft1, (amp)
add amp, amp, astride
1:  vfmacc.vf v13, ft13, v16
beqz kt, 1f
flw ft2, (amp)
add amp, amp, astride
1:  vfmacc.vf v14, ft14, v16
beqz kt, 1f
flw ft3, (amp)
add amp, amp, astride
vfmacc.vf v15, ft15, v16
vlw.v v16, (bkp) # Get next vector from B matrix, overlap loads with jump step
j k_loop
1:  vfmacc.vf v15, ft15, v16

# Save C matrix block back to memory
vsw.v v0, (cnp); add ccp, cnp, cstride;
vsw.v v1, (ccp); add ccp, ccp, cstride;
vsw.v v2, (ccp); add ccp, ccp, cstride;
vsw.v v3, (ccp); add ccp, ccp, cstride;
vsw.v v4, (ccp); add ccp, ccp, cstride;
vsw.v v5, (ccp); add ccp, ccp, cstride;
vsw.v v6, (ccp); add ccp, ccp, cstride;
vsw.v v7, (ccp); add ccp, ccp, cstride;
vsw.v v8, (ccp); add ccp, ccp, cstride;
vsw.v v9, (ccp); add ccp, ccp, cstride;
vsw.v v10, (ccp); add ccp, ccp, cstride;
vsw.v v11, (ccp); add ccp, ccp, cstride;
vsw.v v12, (ccp); add ccp, ccp, cstride;
vsw.v v13, (ccp); add ccp, ccp, cstride;
vsw.v v14, (ccp); add ccp, ccp, cstride;
vsw.v v15, (ccp)

# Following tail instructions should be scheduled earlier in free slots during C block:
# Leaving here for clarity.

# Bump pointers for loop across blocks in one row
slli t6, nvl, 2
add cnp, cnp, t6  # Move C block pointer over
add bnp, bnp, t6  # Move B block pointer over
sub nt, nt, nvl  # Decrement element count in n dimension
bnez nt, c_col_loop  # Any more to do?

# Move to next set of rows
addi m, m, -16  # Did 16 rows above
slli t6, astride, 4  # Multiply astride by 16
add ap, ap, t6  # Move A matrix pointer down 16 rows
slli t6, cstride, 4  # Multiply cstride by 16
add cp, cp, t6  # Move C matrix pointer down 16 rows

slti t6, m, 16
beqz t6, c_row_loop

# Handle end of matrix with fewer than 16 rows.
# Can use smaller versions of above decreasing in powers-of-2 depending on code-size constraints:
end_rows:
# Not done.

exit:
  ld s0, OFFSET(sp)
  ld s1, OFFSET(sp)
  ld s2, OFFSET(sp)
  addi sp, sp, FRAMESIZE
  ret