# <span id="page-0-1"></span>**Chapter 2 Bit Vectors**



We shall use the term *bit vector* as a synonym of *integer*. Thus, a bit vector may be positive, negative, or zero. However, only a nonnegative bit vector may be associated with a *width*:

**Definition 2.1** If  $x \in \mathbb{N}$ ,  $n \in \mathbb{N}$ , and  $x < 2^n$ , then *x* is a bit vector of width *n*, or an *n*-bit vector.

Note that the width of a bit vector is not unique, since an *n*-bit vector is also an *m*-bit vector for all  $m > n$ .

<span id="page-0-0"></span>The *bit slice* and *bit extraction* functions are defined as follows:

**Definition 2.2** Let  $x \in \mathbb{Z}$ ,  $i \in \mathbb{Z}$ , and  $j \in \mathbb{Z}$ .

(a)  $x[i : j] = |(x \mod 2^{i+1})/2^j|;$ (b)  $x[i] = x[i : i].$ 

**Notation** For the purpose of resolving ambiguous expressions, these operators take precedence over the basic arithmetic operators, e.g.,

$$
xy[i : j][k : \ell] = x((y[i : j])[k : \ell]).
$$

For any  $x \in \mathbb{Z}$ , the *binary representation* of x is  $( \ldots b_2 b_1 b_0)_2$ , where  $b_i = x[i]$ for all  $i \in \mathbb{N}$ . We may omit the subscript when the intention is clear. We shall show (Lemma [2.40\)](#page-10-0) that distinct integers have distinct binary representations, so that we may write

$$
x=(\ldots b_2b_1b_0)_2.
$$

In the sequel, we shall extend this notation to non-integral floating-point numbers: for  $k \in \mathbb{N}$ ,

$$
2^{-k}x = (\ldots b_k.b_{k-1}\ldots b_1b_0)_2.
$$

© Springer Nature Switzerland AG 2019

D. M. Russinoff, *Formal Verification of Floating-Point Hardware Design*, [https://doi.org/10.1007/978-3-319-95513-1\\_2](https://doi.org/10.1007/978-3-319-95513-1_2)

17

If *x* is an *n*-bit vector, then it is easily seen that  $x[i] = 0$  for all  $i > n$ , and we may omit the leading zeroes in the representation of *x*:

$$
x = (\dots 000b_{n-1} \dots b_1 b_0)_2 = (b_{n-1} \dots b_1 b_0)_2.
$$

We shall also show (Corollary [2.38\)](#page-9-0) that in this case,

$$
x = \sum_{k=0}^{n-1} 2^k x[k].
$$

Since bit extraction is defined as a special case of bit slice, we shall discuss the latter in Sect. [2.1](#page-1-0) and the former in Sect. [2.2.](#page-7-0) Section [2.3](#page-11-0) deals with the basic RTL operation of *concatenation*.

Arithmetic hardware employs a variety of encoding schemes to represent integers and rational numbers as bit vectors. Floating-point representations are the subject of Chap. 5. In Sects. [2.4](#page-15-0) and [2.5,](#page-18-0) we address the simpler integer and fixed-point formats.

# <span id="page-1-0"></span>**2.1 Bit Slices**

<span id="page-1-1"></span>**Lemma 2.1** *For all*  $x \in \mathbb{Z}$ ,  $i \in \mathbb{N}$ , and  $j \in \mathbb{N}$ ,  $x[i : j]$  *is an*  $(i + 1 - j)$ *-bit vector. Proof* By Lemmas 1.1 and 1.10,  $x[i : j] \in \mathbb{N}$ . By Lemma 1.11,

$$
x[i : j] = \lfloor (x \bmod 2^{i+1})/2^j \rfloor \le (x \bmod 2^{i+1})/2^j < 2^{i+1}/2^j = 2^{i+1-j}.
$$

 $\Box$ 

*Example* Let  $x = 93 = (1011101)_2$ . Then

$$
x[4:2] = \lfloor (x \mod 2^5)/2^2 \rfloor = \lfloor (93 \mod 32)/4 \rfloor = \lfloor 29/4 \rfloor = 7 = (111)_2
$$

is a 3-bit vector and

<span id="page-1-2"></span>
$$
x[10:7] = \lfloor (93 \mod 2^{11})/2^7 \rfloor = \lfloor 93/128 \rfloor = 0 = (0000)_2
$$

is a 4-bit vector.

**Lemma 2.2** *Let*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $i \in \mathbb{Z}$ , and  $j \in \mathbb{N}$ . If  $x \mod 2^{i+1} = y \mod 2^{i+1}$ , *then*  $x[i : j] = y[i : j]$ .

*Proof*  $x[i : j] = \lfloor (x \mod 2^{i+1})/2^j \rfloor = \lfloor (y \mod 2^{i+1})/2^j \rfloor = y[i : j].$ 

<span id="page-2-0"></span>**Lemma 2.3** *For all*  $x \in \mathbb{Z}$  *and*  $i \in \mathbb{Z}$ *,* 

$$
x[i:0] = x \bmod 2^{i+1}.
$$

<span id="page-2-1"></span>*Proof*  $x[i : 0] = |(x \mod 2^{i+1})/2^0| = |x \mod 2^{i+1}| = x \mod 2^{i+1}$ . **Lemma 2.4** *Let x* ∈  $\mathbb{Z}$  *and i* ∈  $\mathbb{N}$ *. If* −2<sup>*i*+1</sup> < *x* < 2<sup>*i*+1</sup>*, then* 

$$
x[i:0] = \begin{cases} x & \text{if } x \ge 0\\ x + 2^{i+1} & \text{if } x < 0. \end{cases}
$$

*Proof* If  $x > 0$ , the claim follows from Lemma [2.3.](#page-2-0) If  $-2^{i+1} < x < 0$ , then by Lemmas [2.3,](#page-2-0) 1.15, and 1.11,

$$
x[i:0] = x \mod 2^{i+1} = (x+2^{i+1}) \mod 2^{i+1} = x+2^{i+1}.
$$

If  $-2^j < x < 0$ , then *x*[*i* : *j*] is the bit vector of width  $i - j + 1$  consisting of all 1s:

**Lemma 2.5** *Let*  $x \in \mathbb{Z}$ *,*  $i \in \mathbb{N}$ *, and*  $j \in \mathbb{N}$ *. If*  $i \geq j$  *and*  $-2^{j} \leq x < 0$ *, then*  $x[i : j] = 2^{i-j+1} - 1.$ 

*Proof* By Lemmas [2.3](#page-2-0) and 1.15, *x* mod  $2^{i+1} = x + 2^{i+1}$ . Thus, by Definition [2.2,](#page-0-0) Lemma 1.1, and Definition 1.1,

$$
x[i : j] = \lfloor (x + 2^{i+1})/2^{j} \rfloor = \lfloor x/2^{j} + 2^{i-j+1} \rfloor = \lfloor x/2^{j} \rfloor + 2^{i-j+1} = 2^{i-j+1} - 1.
$$

**Corollary 2.6** *If*  $i \in \mathbb{N}$ ,  $j \in \mathbb{N}$ , and  $i \geq j$ , then  $(-1)[i : j] = 2^{i-j+1} - 1$ .

The following results are derived from corresponding properties of *mod*.

**Lemma 2.7** *For all*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $i \in \mathbb{Z}$ ,  $j \in \mathbb{Z}$ , and  $k \in \mathbb{Z}$ , if  $j \ge 0$  and  $k \ge i$ , *then*

*(a)*  $(x + y[k:0])[i: j] = (x + y)[i: j];$ *(b)*  $(x - y[k:0])[i : j] = (x + y)[i : j];$  $(c)$   $(xy[k:0])[i:j] = (xy)[i:j].$ 

*Proof* By Definition [2.2](#page-0-0) and Lemmas 1.22 and 1.19,

$$
(x + y[i : 0])[i : j] = \lfloor (x + (y \mod 2^{k+1}) \mod 2^{i+1})/2^j \rfloor
$$
  
= \lfloor ((x + (y \mod 2^{k+1}) \mod 2^{k+1}) \mod 2^{i+1})/2^j \rfloor  
= \lfloor (((x + y) \mod 2^{k+1}) \mod 2^{i+1})/2^j \rfloor

 $\Box$ 

= 
$$
\lfloor ((x + y) \mod 2^{i+1})/2^j \rfloor
$$
  
 =  $(x + y)[i : j]$ .

The other claims follow similarly from Lemmas 1.23 and 1.24.

By expanding the modulus, we may express a bit slice in terms of the floor alone: **Lemma 2.8** *Let*  $x \in \mathbb{Z}$ ,  $i \in \mathbb{Z}$ , and  $j \in \mathbb{Z}$ . If  $i \geq j$ , then

<span id="page-3-0"></span>
$$
x[i-1:j] = \left\lfloor \frac{x}{2^j} \right\rfloor - 2^{i-j} \left\lfloor \frac{x}{2^i} \right\rfloor = \left\lfloor \frac{x}{2^j} \right\rfloor \mod 2^{i-j}.
$$

*Proof* Applying Definitions [2.2](#page-0-0) and 1.3 and Lemma 1.1, we have

$$
x[i-1:j] = \lfloor (x \mod 2^{i})/2^{j} \rfloor
$$

$$
= \left\lfloor \frac{x - \lfloor x/2^{i} \rfloor 2^{i}}{2^{j}} \right\rfloor
$$

$$
= \left\lfloor \frac{x}{2^{j}} - \left\lfloor \frac{x}{2^{i}} \right\rfloor 2^{i-j} \right\rfloor
$$

$$
= \left\lfloor \frac{x}{2^{j}} \right\rfloor - 2^{i-j} \left\lfloor \frac{x}{2^{i}} \right\rfloor.
$$

The second claim follows from Definition 1.3 and Lemmas [2.8](#page-3-0) and 1.2:

$$
\left\lfloor \frac{x}{2^j} \right\rfloor \mod 2^{i-j} = \left\lfloor \frac{x}{2^j} \right\rfloor - 2^{i-j} \left\lfloor \frac{\lfloor 2^{-j}x \rfloor}{2^{i-j}} \right\rfloor = \left\lfloor \frac{x}{2^j} \right\rfloor - 2^{i-j} \left\lfloor \frac{x}{2^i} \right\rfloor = x[i-1:j].
$$

In most cases of interest, the index arguments of  $x[i : j]$  satisfy  $i \ge j \ge 0$ . However, the following lemma is worth noting.

<span id="page-3-1"></span>**Lemma 2.9** For all  $x \in \mathbb{Z}$ ,  $i \in \mathbb{Z}$ , and  $j \in \mathbb{Z}$ , if either  $i < 0$  or  $i < j$ , then  $x[i : j] = 0.$ 

*Proof* Suppose  $i < 0$ . Since  $-(i + 1) \ge 0$ ,  $2^{-(i+1)}x \in \mathbb{Z}$ . Applying Definition 1.3 and Lemma 1.1, we have

$$
x \mod 2^{i+1} = x - \lfloor x/2^{i+1} \rfloor 2^{i+1} \rfloor
$$
  
=  $x - \lfloor 2^{-(i+1)} x \rfloor 2^{i+1} \rfloor$   
=  $x - 2^{-(i+1)} x 2^{i+1} \rfloor$   
= 0.

If  $i < j$ , then by Lemma 1.11,

$$
x \bmod 2^{i+1} < 2^{i+1} \le 2^j
$$

and hence

$$
x[i : j] = \lfloor (x \bmod 2^{i+1})/2^j \rfloor \le (x \bmod 2^{i+1})/2^j < 1,
$$

which, together with Lemma [2.1,](#page-1-1) implies  $x[i : j] = 0$ .

<span id="page-4-2"></span>Here is another case in which a bit slice may be reduced to 0:

**Lemma 2.10** *Let*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $j \in \mathbb{N}$ . If x is a *j*-bit vector, then  $x[i : j] = 0$ . *Proof* By Lemmas 1.10 and 1.11,

$$
0 \le x \bmod 2^{i+1} \le x < 2^j.
$$

Therefore, by Lemma 1.1 and Definition 1.1,

<span id="page-4-1"></span>
$$
0 \le x[i : j] = \lfloor (x \bmod 2^{i+1})/2^j \rfloor \le (x \bmod 2^{i+1})/2^j < 1,
$$

which, together with Lemma [2.1,](#page-1-1) implies  $x[i : j] = 0$ .

**Corollary 2.11** *For all*  $i \in \mathbb{N}$  *and*  $j \in \mathbb{N}$ ,  $0[i : j] = 0$ *.* 

A slice of a right-shifted bit vector,  $|x/2^k|$ , may always be represented as a slice of *x*:

<span id="page-4-0"></span>**Lemma 2.12** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ ,  $j \in \mathbb{N}$ , and  $k \in \mathbb{N}$ ,

$$
\lfloor x/2^k \rfloor [i : j] = x[i+k : j+k].
$$

*Proof* Let  $q = \lfloor x/2^{i+k+1} \rfloor$  and  $r = x \mod 2^{i+k+1}$ , so that  $x = 2^{i+k+1}q + r$  and  $0 \leq r < 2^{i+k+1}$ . Then

$$
\lfloor x/2^k \rfloor = \lfloor 2^{i+1}q + r/2^k \rfloor = 2^{i+1}q + \lfloor r/2^k \rfloor,
$$

where  $|r/2^k| < r/2^k < 2^{i+1}$ . Hence,

$$
\lfloor x/2^k \rfloor \bmod 2^{i+1} = \lfloor r/2^k \rfloor
$$

and by Definition [2.2](#page-0-0) and Lemma 1.2,

$$
\lfloor x/2^k \rfloor [i : j] = \lfloor \lfloor r/2^k \rfloor / 2^j \rfloor = \lfloor r/2^{k+j} \rfloor = \lfloor (x \bmod 2^{i+k+1})/2^{k+j} \rfloor = x[i : j].
$$

**Lemma 2.13** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $k \in \mathbb{N}$ ,

$$
\lfloor x/2^k\rfloor[i:0]=\lfloor x[i+k:0]/2^k\rfloor.
$$

 $\Box$ 

*Proof* Applying Lemma [2.12,](#page-4-0) Definition [2.2,](#page-0-0) and Lemma [2.3](#page-2-0) in succession, we have

$$
\lfloor x/2^k \rfloor [i:0] = x[i+k:k] = \lfloor (x \bmod 2^{i+k+1})/2^k \rfloor = \lfloor x[i+k:0]/2^k \rfloor.
$$

**Lemma 2.14** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ ,  $j \in \mathbb{N}$ , and  $k \in \mathbb{N}$ ,

$$
(2^k x)[i : j] = x[i - k : j - k].
$$

*Proof* If  $k \leq i$ , then by Definition [2.2](#page-0-0) and Lemma 1.18,

$$
(2kx)[i : j] = \lfloor (2kx \mod 2i+1)/2j \rfloor
$$
  
=  $\lfloor 2k(x \mod 2i-k+1)/2j \rfloor$   
=  $\lfloor (x \mod 2i-k+1)/2j-k \rfloor$   
=  $x[i - k : j - k]$ .

If  $i < k$ , then by Definition [2.2](#page-0-0) and Corollary 1.12,

$$
(2kx)[i : j] = \lfloor (2kx \mod 2i+1)/2j \rfloor
$$
  
= \lfloor (2<sup>i+1</sup>(2<sup>k-i-1</sup>x) \mod 2<sup>i+1</sup>)/2<sup>j</sup> \rfloor  
= 0

and  $x[i - k : j - k] = 0$  by Lemma [2.9.](#page-3-1) □

The next lemma provides an alternate expression for a left-shifted bit slice with lower limit 0:

**Lemma 2.15** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $k \in \mathbb{N}$ ,

$$
2^k x[i:0] = (2^k x)[i+k:0].
$$

*Proof* By Lemmas 1.18 and [2.3,](#page-2-0)

$$
(2kx)[i + k : 0] = 2kx \bmod 2i+k+1 = 2k(x \bmod 2i+1) = 2kx[i : 0].
$$

 $\Box$ 

We note two cases in which a bit slice of  $x + 2^k y$  can be simplified.

**Lemma 2.16** *Let*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ ,  $n \in \mathbb{N}$ , and  $k \in \mathbb{N}$ . If  $k \leq m$  and  $x < 2^k$ , *then*

$$
(x + 2k y)[n : m] = y[n - k : m - k].
$$

*Proof* By Lemma [2.9,](#page-3-1) we may assume that  $n \ge m \ge k$ . Since  $0 \le x/2^k < 1$ ,

$$
\lfloor (x+2^k y)/2^k \rfloor = \lfloor y+2/2^k \rfloor = y.
$$

We apply Lemma [2.12,](#page-4-0) substituting  $x + 2^k y$  for  $x, n - k$  for *i*, and  $m - k$  for *j*:

$$
y[n - k : m - k] = \lfloor (x + 2^k y) / 2^k \rfloor [n - k : m - k] = (x + 2^k y)[n : m].
$$

**Lemma 2.17** *Let*  $x \in \mathbb{Z}$ *,*  $y \in \mathbb{Z}$ *,*  $m \in \mathbb{N}$ *,*  $n \in \mathbb{N}$ *, and*  $k \in \mathbb{N}$ *. If*  $n < k$ *, then* 

$$
(x+2^k y)[n:m] = x[n:m].
$$

*Proof* Since  $2^k y = 2^{n+1}(2^{k-n-1}y)$ , where  $2^{k-n-1}y \in \mathbb{Z}$ , Lemma 1.15 implies  $(x + 2^k y) \text{ mod } 2^{n+1} = x \text{ mod } 2^{n+1}$ . The lemma follows from Lemma [2.2.](#page-1-2)

Here is an important lemma that decomposes a slice into two subslices.

**Lemma 2.18** *Let*  $x \in \mathbb{Z}$ *,*  $m \in \mathbb{N}$ *,*  $n \in \mathbb{N}$ *, and*  $p \in \mathbb{N}$ *. If*  $m \leq p \leq n$ *, then* 

<span id="page-6-0"></span>
$$
x[n : m] = 2^{p-m} x[n : p] + x[p-1 : m].
$$

*Proof* The proof consists of three applications of Lemma [2.8:](#page-3-0)

$$
2^{p-m}x[n:p] = 2^{p-m} \left( \left\lfloor \frac{x}{2^p} \right\rfloor - 2^{n+1-p} \left\lfloor \frac{x}{2^{n+1}} \right\rfloor \right),
$$

$$
x[p-1:m] = \left\lfloor \frac{x}{2^m} \right\rfloor - 2^{p-m} \left\lfloor \frac{x}{2^p} \right\rfloor,
$$

and hence,

$$
2^{p-m}x[n:p] + x[p-1:m] = \left\lfloor \frac{x}{2^m} \right\rfloor - 2^{n+1-m} \left\lfloor \frac{x}{2^{n+1}} \right\rfloor
$$

$$
= x[n:m].
$$



Compositions of bit slices may be reduced by means of the following.

**Lemma 2.19** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ ,  $j \in \mathbb{N}$ ,  $k \in \mathbb{N}$ , and  $\ell \in \mathbb{N}$ ,

<span id="page-6-1"></span>
$$
x[i:j][k:l] = \begin{cases} x[k+j:\ell+j] & \text{if } k \le i-j\\ x[i:\ell+j] & \text{if } k > i-j. \end{cases}
$$

*Proof* By Lemma [2.12,](#page-4-0)

$$
x[i : j][k : \ell] = \lfloor x/2^j \rfloor [i - j : 0][k : \ell]
$$
  
=  $(\lfloor x/2^j \rfloor \mod 2^{i-j+1})[k : \ell]$   
=  $(\lfloor (\lfloor x/2^j \rfloor \mod 2^{i-j+1}) \mod 2^{k+1})/2^{\ell} \rfloor$ .

If  $k \leq i - j$ , then this reduces, by Corollary 1.20, to

$$
\lfloor (\lfloor x/2^j \rfloor \bmod 2^{k+1})/2^{\ell} \rfloor = \lfloor x/2^j \rfloor [k:\ell] = x[k+j:\ell+j].
$$

On the other hand, if  $k > i - j$ , then by Lemma 1.11,

$$
\lfloor x/2^j \rfloor \bmod 2^{i-j+1} < 2^{i-j+1} < 2^{k+1},
$$

and by Lemma 1.11, the expression reduces instead to

$$
\lfloor (\lfloor x/2^{j} \rfloor \bmod 2^{i-j+1})/2^{\ell} \rfloor = \lfloor x/2^{j} \rfloor [i-j : \ell] = x[i : \ell + j].
$$



# <span id="page-7-0"></span>**2.2 Bit Extraction**

<span id="page-7-1"></span>Instead of Definition [2.2,](#page-0-0) we could have defined *<sup>x</sup>*[*n*] more directly as follows. **Lemma 2.20** *For all*  $x \in \mathbb{Z}$  *and*  $n \in \mathbb{Z}$ *,* 

$$
x[n] = \lfloor x/2^n \rfloor \bmod 2.
$$

*Proof* By Lemmas [2.8](#page-3-0) and 1.2 and Definition 1.3,

$$
x[n] = x[(n+1)-1:n] = \lfloor x/2^n \rfloor - 2\lfloor x/2^{n+1} \rfloor
$$

$$
= \lfloor x/2^n \rfloor - 2\lfloor x/2^n \rfloor / 2\rfloor
$$

$$
= \lfloor x/2^n \rfloor \mod 2.
$$



**Corollary 2.21** *For all*  $x \in \mathbb{Z}$  *and*  $n \in \mathbb{Z}$ *,*  $x[n] \in \{0, 1\}$ *.* 

Here is an equivalent recursive definition that may be used in inductive proofs. **Lemma 2.22** *For all*  $x \in \mathbb{Z}$  *and*  $n \in \mathbb{N}$ *,* 

$$
x[n] = \begin{cases} x \bmod 2 & \text{if } n = 0\\ \lfloor x/2 \rfloor [n-1] \text{ if } n > 0. \end{cases}
$$

*Proof* The base case is the  $n = 0$  case of Lemma [2.20.](#page-7-1) The inductive case is an instance of Lemma 2.12, with  $k = 1$  and  $i = i = n - 1$ . instance of Lemma [2.12,](#page-4-0) with  $k = 1$  and  $i = j = n - 1$ .

A number of important properties of bit extraction are special cases of the results of Sect. [2.1](#page-1-0) We list some of them here without proof.

**Lemma 2.23** *For all*  $x \in \mathbb{Z}$  *and*  $n \in \mathbb{Z}$ *, if*  $n < 0$ *, then*  $x[n] = 0$ *.* 

**Lemma 2.24** *For all*  $k \in \mathbb{Z}$ ,  $0[k] = 0$ *.* 

**Lemma 2.25** *For all*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $n \in \mathbb{Z}$ , and  $k \in \mathbb{Z}$ , if  $k < n$  and  $x \mod 2^n =$ *y* mod 2*n, then*

$$
x[k] = y[k].
$$

<span id="page-8-0"></span>**Lemma 2.26** *For all*  $n \in \mathbb{Z}$ *, if x is an n*-bit vector, then  $x[n] = 0$ *.* 

<span id="page-8-3"></span>**Lemma 2.27** *Let*  $x \in \mathbb{Z}$  *and*  $n \in \mathbb{N}$ *. If*  $-2^n \le x \le 0$ *, then*  $x[n] = 1$ *.* 

**Corollary 2.28** *For all*  $i \in \mathbb{N}$ ,  $(-1)[i] = 1$ *.* 

<span id="page-8-5"></span>**Lemma 2.29** *For all*  $x \in \mathbb{Z}$ *,*  $n \in \mathbb{Z}$ *, and*  $k \in \mathbb{Z}$ *,* 

$$
(2^k x)[n+k] = x[n].
$$

<span id="page-8-2"></span>**Lemma 2.30** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $k \in \mathbb{N}$ ,

 $|x/2^{k}|[i] = x[i + k]$ .

**Lemma 2.31** *For all*  $x \in \mathbb{Z}$ ,  $i \in \mathbb{Z}$ ,  $j \in \mathbb{Z}$ , and  $k \in \mathbb{Z}$ , if  $0 \le k \le i - j$ , then

$$
x[i : j][k] = x[j+k].
$$

**Lemma 2.32** *For all*  $x \in \mathbb{Z}$ *,*  $m \in \mathbb{Z}$ *, and*  $n \in \mathbb{Z}$ *, if*  $m \le n$ *, then* 

<span id="page-8-1"></span>
$$
x[n : m] = 2^{n-m}x[n] + x[n-1 : m].
$$

**Lemma 2.33** *For all*  $x \in \mathbb{Z}$ *,*  $m \in \mathbb{Z}$ *, and*  $n \in \mathbb{Z}$ *, if*  $m \leq n$ *, then* 

$$
x[n : m] = x[m] + 2x[n : m+1].
$$

**Lemma 2.34** *Let*  $n \in \mathbb{N}$  *and*  $k \in \mathbb{N}$ *, and let x be an n*-bit vector. If  $k < n$  *and*  $x > 2^n - 2^k$ *, then*  $x[k] = 1$ *.* 

*Proof* Since  $2^n - 2^k \le x < 2^n$ ,  $2^{n-k} - 1 \le x/2^k < 2^{n-k}$ , and by Definition 1.1,  $|x/2^k| = 2^{n-k} - 1$ . Now by Lemma 2.20,  $x[k] = (2^{n-k} - 1) \mod 2 = 1$ .  $|x/2^{k}| = 2^{n-k} - 1$ . Now by Lemma [2.20,](#page-7-1)  $x[k] = (2^{n-k} - 1) \mod 2 = 1$ .

<span id="page-8-4"></span>**Corollary 2.35** *For all*  $n \in \mathbb{Z}$  *and*  $x \in \mathbb{N}$ *, if*  $2^n \le x \le 2^{n+1}$ *, then*  $x[n] = 1$ *.* 

 $\Box$ 

**Lemma 2.36** *For all*  $n \in \mathbb{N}$  *and*  $i \in \mathbb{Z}$ ,  $(2^n)[i] = 1 \Leftrightarrow i = n$ .

*Proof* By Lemma [2.20,](#page-7-1)  $(2^n)[n] = \lfloor 2^n/2^n \rfloor \text{ mod } 2 = 1 \text{ mod } 2 = 1.$ 

Suppose  $i \neq n$ . If  $i < n$ , then  $2^{i}$  is an *n*-bit vector and Lemma [2.26](#page-8-0) applies. If  $i > n$ , then

$$
(2^{i})[n] = \lfloor 2^{i}/2^{n} \rfloor \text{ mod } 2 = mod 2^{i-n} \text{ mod } 2 = 0.
$$

The following lemma and its corollary justify the notation discussed at the beginning of this chapter.

<span id="page-9-1"></span>**Lemma 2.37** *For all*  $x \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $j \in \mathbb{N}$ ,

$$
\sum_{k=j}^{i} 2^{k-j} x[k] = x[i : j].
$$

*Proof* If  $i < j$ , then both sides of the equation reduce to 0 by Lemma [2.9.](#page-3-1) We proceed by induction. Thus, for  $i \geq j$ , applying Lemma [2.32,](#page-8-1) we have

$$
\sum_{k=j}^{i} 2^{k-j} x[k] = 2^{i-j} x[i] + \sum_{k=j}^{i-1} 2^{k-j} x[k]
$$

$$
= 2^{i-j} x[i] + x[i-i : j]
$$

$$
= x[i : j].
$$

<span id="page-9-0"></span>**Corollary 2.38** *If*  $n \in \mathbb{N}$ ,  $n > 0$ , and *x is an n-bit vector, then* 

$$
\sum_{k=0}^{n-1} 2^k x[k] = x.
$$

*Proof* This follows from Lemmas [2.37](#page-9-1) and [2.4.](#page-2-1) □

The next lemma allows us to define a bit vector in a natural way as a sequence of bits. That is, given a sequence of 1-bit vectors  $b_0, \ldots, b_{n-1}$ , we may say, without ambiguity, Let *x* be the bit vector of width *n* defined by  $x[k] = b_k$  for  $k = 0, \ldots, n-1$ 1 . . . . The existence of such a bit vector is guaranteed by Lemma [2.39;](#page-9-2) its uniqueness is ensured by Corollary [2.38.](#page-9-0)

<span id="page-9-2"></span>**Lemma 2.39** *Let*  $x = \sum_{i=0}^{n-1} 2^{i}b_{i}$ , where  $n \in \mathbb{N}$  and  $b_{i} \in \{0, 1\}$ ,  $i = 0, \ldots, n-1$ . *Then for*  $k = 0, ..., n - 1, x[k] = b_k$ . *Proof* Let  $U = \sum_{i=k+1}^{n-1} 2^i b_i$  and  $L = \sum_{i=0}^{k-1} 2^i b_i$ . Then  $x = U + 2^{k}b_{k} + L.$ 

#### 2.2 Bit Extraction 27

Since

$$
U = 2^{k+1} \sum_{i=k+1}^{n-1} 2^{i-(k+1)} b_i,
$$

*x* mod  $2^{k+1} = U + 2^{k}b_{k} + L$  mod  $2^{k+1} = 2^{k}b_{k} + L$  mod  $2^{k+1}$ 

by Lemma 1.15. But since

$$
L \le \sum_{i=0}^{k-1} 2^i = 2^k - 1
$$

and

$$
2^{k}b_{k} + L \le 2^{k} + 2^{k} - 1 < 2^{k+1},
$$
\n
$$
2^{k}b_{k} + L \mod 2^{k+1} = 2^{k}b_{k} + L
$$

by Lemma 1.11. Thus, by Definitions [2.2](#page-0-0) and [2.2,](#page-0-0)

$$
x[k] = \lfloor (x \bmod 2^{k+1})/2^k \rfloor = \lfloor (2^k b_k + L)/2^k \rfloor = \lfloor b_k + L/2^k \rfloor = b_k.
$$

<span id="page-10-0"></span>A bit vector is uniquely determined by its binary representation:

**Lemma 2.40** *Let*  $x \in \mathbb{Z}$  *and*  $y \in \mathbb{Z}$ *. If*  $x[k] = y[k]$  *for all*  $k \in \mathbb{N}$ *, then*  $x = y$ *.* 

*Proof* The proof is by induction on  $|x| + |y|$ .

Suppose  $x \neq y$ . We must show that for some  $k \in \mathbb{N}$ ,  $x[k] \neq y[k]$ . We may assume that  $x[0] = y[0]$ , and hence  $\lfloor x/2 \rfloor \neq \lfloor y/2 \rfloor$ , for otherwise

$$
x = 2\lfloor x/2 \rfloor + x[0] = 2\lfloor y/2 \rfloor + y[0] = y.
$$

Since  $x \neq y$  and  $x[0] = y[0]$ , at least one of  $x$  and  $y$  must be different from both 0 and -1, and hence, by Lemma 1.3,

$$
\left|\lfloor x/2\rfloor\right|+\left|\lfloor y/2\rfloor\right|\right|<|x|+|y|.
$$

By induction, there exists  $k \in \mathbb{N}$  such that  $\lfloor x/2 \rfloor [k] \neq \lfloor y/2 \rfloor [k]$ , and consequently, by Lemma [2.30,](#page-8-2)

$$
x[k+1] = \lfloor x/2 \rfloor[k] \neq \lfloor y/2 \rfloor[k] = y[k+1].
$$

 $\Box$ 

# <span id="page-11-0"></span>**2.3 Concatenation**

If  $x = (\beta_{m-1} \cdots \beta_0)_2$  and  $y = (\gamma_{n-1} \cdots \gamma_0)_2$  are considered as bit vectors of widths *m* and *n*, respectively, then the concatenation of x and y is the  $(m + n)$ -bit vector

$$
(\beta_{m-1}\cdots\beta_0\gamma_{n-1}\cdots\gamma_0)_2.
$$

This notion is extended by the following function, which takes a list of bit vectors and widths, coerces each bit vector to its associated width, and concatenates the results:

**Definition 2.3** For all  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ , and  $n \in \mathbb{N}$ ,

<span id="page-11-1"></span>
$$
cat(x, m, y, n) = 2n x[m-1:0] + y[n-1:0].
$$

This construction is extended recursively to 2*k* arguments for arbitrary  $k \in \mathbb{Z}^+$ :

 $cat(x_1, n_1, x_2, n_2, \ldots, x_k, n_k) = cat(x_1, n_1, cat(x_2, n_2, \ldots, x_k, n_k), n_2 + \ldots + n_k),$ 

where  $x_i \in \mathbb{Z}$  and  $n_i \in \mathbb{N}$  for  $i = 1, \ldots, k$ .

Associativity follows immediately:

**Lemma 2.41** *For all*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $z \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ ,  $n \in \mathbb{N}$ , and  $p \in \mathbb{N}$ ,

<span id="page-11-2"></span>
$$
cat(cat(x, m, y, n), z, p) = cat(x, m, y, n, z, p).
$$

**Lemma 2.42** *For all*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ , and  $n \in \mathbb{N}$ , cat(x, m, y, n) is an  $(m + n)$ *-bit vector.* 

*Proof* By Lemma [2.1,](#page-1-1)  $x[m-1 : 0] < 2^m$  and  $y[n-1 : 0] < 2^n$ . It follows that  $x[m-1:0] < 2^m - 1$  and  $y[n-1:0] < 2^n - 1$ , and hence,

$$
cat(x, m, y, n) = 2nx[m-1:0] + y[n-1:0]
$$
  
\n
$$
\leq 2n(2m - 1) + (2n - 1)
$$
  
\n
$$
= 2n+m - 1
$$
  
\n
$$
< 2n+m.
$$

 $\Box$ 

<span id="page-11-3"></span>We note several trivial cases:

**Lemma 2.43** *For all*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ , and  $n \in \mathbb{N}$ ,

$$
cat(x, m, y, 0) = x[m-1:0]
$$

*and*

$$
cat(x, 0, y, n) = cat(0, m, y, n) = y[n-1:0].
$$

*Proof* These are simple consequences of Definition [2.3](#page-11-1) and Lemmas [2.9](#page-3-1) and [2.11.](#page-4-1)  $\Box$ 

**Notation** In standard RTL syntax, the concatenation of two bit vectors  $\phi$  and  $\psi$  is denoted by  $\{\phi, \psi\}$ . This notation depends on a characteristic shared by conventional hardware description languages: any expression that represents a bit vector has an associated (explicit or implicit) width. For example, the expression  $\text{sig}[3:0]$ is understood to be of width 4, and the expression 5'b01001 identifies the constant 9 as a bit vector of width 5. We shall incorporate this construct into our informal mathematical notation through an abuse of Verilog syntax, representing  $cat(x, m, y, n)$  as

$$
\{m'x, n'y\}.
$$

The width specifier may be omitted in a context in which it can be inferred by default.

*Example* If  $x \in \{0, 1\}$  and *y* has been identified as a bit vector of width *n*, then

$$
\{x, y, z[i : j], w[k]\} = cat(x, 1, y, n, z[i : j], i+1-j, w[k], 1).
$$

<span id="page-12-1"></span>The following is a restatement of Lemma [2.18:](#page-6-0)

**Lemma 2.44** *Let*  $x \in \mathbb{Z}$ *,*  $m \in \mathbb{N}$ *,*  $n \in \mathbb{N}$ *, and*  $p \in \mathbb{N}$ *. If*  $m \leq p \leq n$ *, then* 

$$
x[n : m] = \{x[n : p], x[p-1 : m]\}.
$$

**Corollary 2.45** *Let*  $x \in \mathbb{Z}$ *,*  $m \in \mathbb{N}$ *, and*  $n \in \mathbb{N}$ *. If*  $m \leq n$ *, then* 

<span id="page-12-0"></span>
$$
x[n : m] = \{x[n], x[n-1 : m]\} = \{x[n : m+1], x[m]\}.
$$

**Lemma 2.46** *Let*  $z = \{m'x, n'y\}$ *, where*  $x \in \mathbb{Z}$ *,*  $y \in \mathbb{Z}$ *,*  $m \in \mathbb{N}$ *, and*  $n \in \mathbb{N}$ *. Then* 

$$
z[n-1:0] = y[n-1:0]
$$

*and*

$$
z[n+m-1:n] = x[m-1:0].
$$

*Proof* By Definition [2.3,](#page-11-1) we have

$$
z = 2^n x[m-1:0] + y[n-1:0],
$$

where  $0 \le y[n-1:0] < 2^n$  by Lemma [2.1.](#page-1-1) Thus, by Lemmas [2.3,](#page-2-0) 1.15, and 1.11,

$$
z[n-1:0] = z \mod 2^n = y[n-1:0] \mod 2^n = y[n-1:0].
$$

Now by Definition [2,](#page-0-1)

$$
z[n+m-1:n] = \lfloor (z \bmod 2^{n+m})/2^n \rfloor.
$$

But Lemma [2.42](#page-11-2) yields  $z < 2^{n+m}$  and hence, by Lemma 1.11,

$$
z[n+m-1:n] = \lfloor z/2^n \rfloor = \lfloor x[m-1:0] + y[n-1:0]/2^n \rfloor.
$$

Finally, by Lemma 1.1, this reduces to

$$
x[m-1:0] + \lfloor y[n-1:0]/2^n \rfloor = x[m-1:0].
$$



**Corollary 2.47** *Let*  $x_1 \in \mathbb{Z}$ ,  $y_1 \in \mathbb{Z}$ ,  $x_2 \in \mathbb{Z}$ ,  $y_2 \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ , and  $n \in \mathbb{N}$ . If

$$
{m' x_1, n' y_1} = {m' x_2, n' y_2},
$$

*then*  $x_1[m-1:0] = x_2[m-1:0]_2$  *and*  $y_1[n-1:0]_1 = y_2[n-1:0]$ .

**Lemma 2.48** *Let*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ ,  $n \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $j \in \mathbb{N}$ . If  $i \geq j$ , then

<span id="page-13-0"></span>
$$
\{m'x, n'y\}[i : j] = \begin{cases} y[i : j] & \text{if } n > i \\ x[i - n : j - n] & \text{if } m + n > i \ge j \ge n \\ x[m - 1 : j - n] & \text{if } i \ge m + n \text{ and } j \ge n \\ \{x[i - n : 0], y[n - 1 : j]\} & \text{if } m + n > i \ge n > j \\ \{x[m - 1 : 0], y[n - 1 : j]\} & \text{if } i \ge n + m \text{ and } n > j. \end{cases}
$$

*Proof* Let *z* = {*x*[*m*−1 : 0]*, y*[*n*−1 : 0]}. By Lemma 2.46*,* 

$$
y[n-1:0] = z[n-1:0]
$$

and

$$
x[m-1:0] = z[n+m-1:n]
$$

and by Lemma [2.42,](#page-11-2) *z* is an  $(m + n)$ -bit vector. Our goal is to compute *z*[*i* : *j*]. We consider five cases as suggested by the lemma statement, each of which involves two or more applications of Lemma [2.19.](#page-6-1) Case 1:  $n > i$ 

By Lemma [2.19,](#page-6-1)

$$
z[i : j] = z[n-1 : 0][i : j] = y[n-1 : 0][i : j] = y[i : j].
$$

Case 2:  $m + n > i \ge j \ge n$ By Lemma [2.19,](#page-6-1)

$$
z[i : j] = z[m + n - 1 : n][i - n : j - n]
$$
  
=  $x[m - 1 : 0][i - n : j - n]$   
=  $x[i - n : j - n]$ .

Case 3:  $i \ge m + n$  and  $j \ge n$ By Lemma [2.44,](#page-12-1)

$$
z[i : j] = \{z[i : m + n], z[m + n - 1 : j]\}.
$$

But  $z[i : m + n] = 0$  by Lemma [2.10](#page-4-2) and hence

$$
z[i:j] = z[m+n-1:j]
$$

by Lemma [2.43.](#page-11-3) Now by Lemma [2.19,](#page-6-1)

$$
z[m + n - 1 : j] = z[m + n - 1 : n][m - 1 : j - n]
$$

$$
= x[m - 1 : 0][m - 1 : j - n]
$$

$$
= x[m - 1 : j - n].
$$

Case 4:  $m + n > i \ge n > j$ By Lemma [2.44,](#page-12-1)

$$
z[i : j] = \{z[i : n], z[n-1 : j]\}.
$$

But by Lemma [2.19,](#page-6-1)

$$
z[i : n] = z[m + n - 1 : n][i - n : 0]
$$
  
=  $x[m - 1 : 0][i - n : 0]$   
=  $x[i - n : 0]$ 

and

$$
z[n-1:j] = z[n-1:0][n-1:j] \n= y[n-1:0][n-1:j] \n= y[n-1:j].
$$

Case 5:  $i > n + m$  and  $n > j$ By Lemma [2.44,](#page-12-1)

 $z[i : j] = \{z[i : m + n], z[m + n - 1 : n], z[n - 1 : j]\}.$ 

As in Case 4,  $z[n-1 : j] = y[n-1 : j]$ . By Lemma [2.10,](#page-4-2)  $z[i : m + n] = 0$ , and hence by Lemma [2.43,](#page-11-3)

$$
z[i : j] = \{z[m + n - 1 : n], z[n - 1 : j]\} = \{x[m - 1 : 0], y[n - 1 : j]\}.
$$

 $\Box$ 

**Corollary 2.49** *If*  $x \in \mathbb{Z}$ ,  $y \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ ,  $n \in \mathbb{N}$ , and  $i \in \mathbb{N}$ , then

$$
\{m'x, n'y\}[i] = \begin{cases} y[i] & \text{if } i < n \\ x[i-n] & \text{if } n \le i < m+n \\ 0 & \text{if } n+m \le i. \end{cases}
$$

*Proof* The cases listed correspond to the first three cases of Lemma [2.48](#page-13-0) with  $i = j$ . Note that for the third case, the lemma gives  $x[m-1 : i - n]$ , but since  $i > n + m$ , i.e.,  $i - n > m - 1$ , this reduces to 0 by Lemma 2.9. i.e.,  $i - n > m - 1$ , this reduces to 0 by Lemma [2.9.](#page-3-1)

### <span id="page-15-0"></span>**2.4 Integer Formats**

The simplest of all bit vector encoding schemes is the *unsigned integer* format, whereby the first  $2^n$  natural numbers, i.e., the bit vectors of width  $n$ , are represented by themselves under the identity mapping. However trivial, it will be convenient to have an explicit definition of this correspondence:

**Definition 2.4** If *r* is a *n*-bit vector, where  $n \in \mathbb{N}$ , then

$$
ui(r, n) = r.
$$

Somewhat more interesting is the *signed integer* format, which maps the set of 2*<sup>n</sup>* integers *x* in the range  $-2^{n-1}$  ≤ *x* ≤  $2^{n-1}$  to the set of bit vectors of width *n* and may be defined by

$$
x \longmapsto x[n-1:0].
$$

With respect to this mapping, the most significant bit of the encoding of  $x$ ,

$$
x[n-1:0][n-1] = x[n-1],
$$

is 0 if  $0 \le x < 2^{n-1}$  (by Lemma [2.26\)](#page-8-0) and 1 if  $-2^{n-1} \le x < 0$  (by Lemma [2.27\)](#page-8-3), and is therefore considered the *sign bit* of the encoding.

The integer represented by a given encoding is computed by the following function, as affirmed by Lemma [2.50](#page-16-0) below:

**Definition 2.5** If *r* is a *n*-bit vector, where  $n \in \mathbb{N}$ , then

<span id="page-16-1"></span>
$$
si(r, n) = \begin{cases} r - 2^{n} \text{ if } r[n-1] = 1\\ r \text{ if } r[n-1] = 0. \end{cases}
$$

<span id="page-16-0"></span>**Lemma 2.50** *Let n* ∈ *N and*  $x \text{ ∈ Z}$ *. If* −2*<sup><i>n*−1</sup> ≤  $x$  ≤ 2<sup>*n*−1</sup>*, then* 

$$
si(x[n-1:0],n) = x.
$$

*Proof* If  $0 \le x < 2^{n-1}$ , then  $x[n-1:0] = x$  by Lemma [2.4,](#page-2-1) and  $x[n-1] = 0$  by Lemma [2.26.](#page-8-0) Thus,

$$
si(x[n-1:0], n) = si(x, n) = x.
$$

If  $-2^{n-1}$  ≤ *x* < 0, then *x*[*n*−1 : 0] = *x* + 2<sup>*n*</sup> by Lemma [2.4,](#page-2-1) and  $(x+2^n)[n-1] = 1$ by Corollary [2.35.](#page-8-4) Thus,

$$
si(x[n-1:0], n) = si(x + 2^n, n) = (x + 2^n) - 2^n = x.
$$

This scheme is also known as the *n*-bit *two's complement* encoding, because if  $0 \leq x < 2^n$ , then the encoding of  $-x$  is the complement of *x* with respect to  $2^n$ , i.e.,

$$
x + (-x)[n-1:0] = x + (-x + 2n) = 2n.
$$

**Lemma 2.51** *If*  $n \in \mathbb{N}$ ,  $r \in \mathbb{N}$ ,  $i \in \mathbb{N}$ , and  $j \in \mathbb{N}$  with  $j \leq i < n$ , then

$$
si(r, n)[i : j] = r[i : j].
$$

**Lemma 2.52** *If*  $n \in \mathbb{N}$ ,  $k \in \mathbb{N}$ , and  $r$  *is an n-bit vector, then* 

$$
si(2^k r, k+n) = 2^k si(r, n).
$$

*Proof* This follows easily from Definition [2.5](#page-16-1) and Lemma [2.29.](#page-8-5)

An *n*-bit integer encoding is converted to an *m*-bit encoding, where  $m > n$ , by *sign extension*:

**Definition 2.6** Let *r* be an *n*-bit vector, where  $n \in \mathbb{N}$ , and let  $m \in \mathbb{N}$ ,  $m \ge n$ . Then

$$
sextend(m, n, r) = si(r, n)[m-1:0].
$$

A sign extension of an integer encoding *r* represents the same value as *r*:

**Lemma 2.53** *Let r be an n*-bit vector, where  $n \in \mathbb{N}$ , and let  $m \in \mathbb{N}$ ,  $m > n$ . Then

 $si$ (*sextend*(*m, n, r*), *m*) =  $si$ (*r, n*).

*Proof* First suppose  $r[n-1] = 0$ . Then  $si(r, n) = r$  and by Corollary [2.35,](#page-8-4) 0 <  $r$  < 2*n*−1. By Lemma [2.4,](#page-2-1)

$$
sextend(m, n, r) = si(r, n)[m-1:0] = r[m-1:0] = r,
$$

and since Lemma [2.26](#page-8-0) implies  $r[m-1] = 0$ ,

$$
si(sextend(m, n, x), m) = si(r, m) = r = si(r, n).
$$

Now suppose  $r[n - 1] = 1$ . Then by Lemma [2.26,](#page-8-0)  $2^{n-1} \le r < 2^n$ . Now  $si(r, n) =$ *r* − 2<sup>*n*</sup>, where  $-2^{m-1} < -2^{n-1} < r - 2^n < 0$ . By Lemma [2.4,](#page-2-1)

$$
sextend(m, n, r) = si(r, n)[m-1: 0] = (r - 2n)[m-1: 0] = r - 2n + 2m.
$$

But since  $2^{m-1}$  ≤  $r-2^n+2^m < 2^m$ . Corollary [2.35](#page-8-4) implies  $(r-2^n+2^m)[m-1] = 1$ , and hence

$$
si(sextend(m, n, r), m) = si(r-2n+2m, m) = r-2n+2m-2m-r-2n = si(rl, n).
$$

Given an approximation *Y* of an integer *X*, the following lemma provides a condition under which the *n*-bit signed integer represented by *Y* is an equally accurate approximation of the *n*-bit signed integer represented by *X*. This result is useful in approximating the signed integer values of a "redundant" representation, i.e., a representation of an integer as a sum or difference of two vectors. (See, for example, the proof of Lemma 18.5.)

**Lemma 2.54** *Let*  $X \in \mathbb{Z}$ ,  $Y \in \mathbb{Z}$ , and  $n \in \mathbb{Z}$ , with  $n > 0$ . If

$$
|si(X \bmod 2^n, n)| + |X - Y| < 2^{n-1},
$$

*then*

$$
si(X \bmod 2^n, n) - si(Y \bmod 2^n, n) = X - Y.
$$

*Proof* Let  $\overline{X} = X \mod 2^n$ ,  $\overline{Y} = Y \mod 2^n$ , and  $k = |X - Y|$ . *Case 1*:  $\left\lfloor \frac{X}{2^n} \right\rfloor = \left\lfloor \frac{Y}{2^n} \right\rfloor$ . In this case,  $\bar{X} - \bar{Y} = X - Y$ . Suppose  $\bar{X} \leq \bar{Y}$ . If  $\bar{X} \geq 2^{n-1}$ , then  $\bar{Y} \geq 2^{n-1}$  and  $si(\bar{X}, n) - si(\bar{Y}, n) = (\bar{X} - 2^n) - (\bar{Y} - 2^n) = \bar{X} - \bar{Y} = X - Y$  but if  $\bar{X} < 2^{n-1}$ , then  $\bar{X} = si(\bar{X}, n) < 2^{n-1} - k$ , which implies  $\bar{Y} < 2^{n-1}$  and

$$
si(\bar{X}, n) - si(\bar{Y}, n) = \bar{X} - \bar{Y} = X - Y.
$$

On the other hand, suppose  $\bar{X} > \bar{Y}$ . If  $\bar{X} < 2^{n-1}$ , then  $\bar{Y} < 2^{n-1}$  and

$$
si(\bar{X}, n) - si(\bar{Y}, n) = \bar{X} - \bar{Y} = X - Y,
$$

 $\frac{1}{2}$  but if  $\bar{X} \geq 2^{n-1}$ , then  $si(\bar{X}, n) = \bar{X} - 2^n$ , and since  $si(\bar{X}, n) > -2^{n-1} + k$ ,  $\overline{X} > 2^{n-1} + k$ , which implies  $\overline{Y} > 2^{n-1}$  and

$$
si(\bar{X}, n) - si(\bar{Y}, n) = (\bar{X} - 2^n) - (\bar{Y} - 2^n) = \bar{X} - \bar{Y} = X - Y.
$$

*Case* 2:  $\left\lfloor \frac{X}{2^n} \right\rfloor \neq \left\lfloor \frac{Y}{2^n} \right\rfloor$ . ĺ Suppose  $X < Y$ . Let  $m = \left\lfloor \frac{X}{2^n} \right\rfloor$ . Then

$$
2^{n}m \le X < 2^{n}(m+1) \le Y < X + 2^{n-1} < 2^{n}(m+2).
$$

Thus,  $\bar{X} = X - 2^n m$  and

$$
\overline{Y} = Y - 2^{n}(m+1) = k + X - 2^{n}(m+1) = k + (\overline{X} + 2^{n}m) - 2^{n}(m+1) = k - 2^{n} + \overline{X} < k < 2^{n-1}.
$$

But then

$$
\bar{X} = \bar{Y} + 2^n - k \ge 2^n - k > 2^{n-1}
$$

and

$$
si(\bar{X}, n) - si(\bar{Y}, n) = \bar{X} - 2^{n} - \bar{Y} = (\bar{Y} + 2^{n} - k) - 2^{n} - \bar{Y} = -k = X - Y.
$$

The case  $X > Y$  is similar.

## <span id="page-18-0"></span>**2.5 Fixed-Point Formats**

A fixed-point format may be thought of as derived from an integer format by inserting an implicit binary point following some specified number of leading bits. The rational value represented by an *n*-bit vector *r* with respect to an unsigned or signed fixed-point format of width *n* with *m* integer bits is computed as follows:

**Definition 2.7** Let  $n \in \mathbb{Z}$  and  $m \in \mathbb{Z}$  with  $n > 0$  and let r be a bit vector of width *n*.

(a) 
$$
uf(r, n, m) = 2^{m-n}ui(r) = 2^{m-n}r
$$
;  
\n(b)  $sf(r, n, m) = 2^{m-n}si(r, n) = \begin{cases} 2^{m-n}r & \text{if } r < 2^{n-1} \\ 2^{m-n}r - 2^m & \text{if } r > 2^{n-1} \end{cases}$ .

The number of fractional bits of a fixed-point format of width *n* and *m* integer bits is  $f = n - m$ . Note that while *n* must be positive, there is no restriction on *m*. If  $m > n$ , then the interpreted value is an integer with  $m - n$  trailing zeroes and  $f < 0$ ; if  $m < 0$ , then the interpreted value is a fraction with  $-m$  leading zeroes and  $f > n$ .

We have the following expression for a bit slice of an encoding in terms of the encoded value:

<span id="page-19-0"></span>**Lemma 2.55** *Let*  $n \in \mathbb{N}$ *,*  $m \in \mathbb{N}$ *,*  $i \in \mathbb{N}$ *, and*  $j \in \mathbb{N}$  *with*  $m \leq n$  *and*  $j \leq i \leq n$ *. Let*  $f = n - m$ *. Let*  $r$  *be an*  $n$ *-bit vector and suppose that either*  $x = uf(r, n, m)$  *or*  $x = sf(r, n, m)$ *. Then* 

$$
r[i : j] = 2^{f-j} \left( x^{(f-j)} - x^{(f-i-1)} \right).
$$

*Proof* If  $x = uf(r, n, m)$ , then  $r = 2^f x$ ; if  $x = sf(r, n, m)$ , then either  $r = 2^f x$  or  $r = 2^f x + 2^n$ . In any case, by Lemmas [2.2](#page-1-2) and [2.8,](#page-3-0)

$$
r[i:j] = (2^{f}x)[i:j] = \left\lfloor \frac{2^{f}x}{2^{j}} \right\rfloor = \left\lfloor \frac{2^{f}x}{2^{1+i}} \right\rfloor = 2^{f-j} \left( x^{(f-j)} - x^{(f-i-1)} \right).
$$

**Corollary 2.56** *Let*  $n \in N$  *and*  $m \in N$  *with*  $m \leq n$  *and let*  $f = n - m$ *. Let*  $k \in \mathbb{Z}$  *with*  $f - n \leq k < f$ . Let *r* be an *n*-bit vector and suppose that either  $x = uf(x, n, m)$  *or*  $x = sf(x, n, m)$ *. Then* 

$$
x^{(k)} = x \Leftrightarrow r[f - k - 1:0] = 0.
$$

*Proof* By Lemma [2.55,](#page-19-0)

$$
r[f - k - 1:0] = 2^{f}(x^{(f)} - x^{(k)}) = 2^{f}(x - x^{(k)}).
$$

The following result is useful in determining the value of a fixed-point encoding:

**Lemma 2.57** *Let*  $n \in N$  *and*  $m \in N$  *with*  $m \le n$ *. Let*  $r$  *be an*  $n$ *-bit vector and*  $x = sf(r, n, m)$ *. If*  $y \in \mathbb{Z}$  *satisfies*  $r \equiv y \pmod{2^n}$  *and*  $-2^{n-1} \le y < 2^{n-1}$ *, then* 

$$
x = 2^{m-n}y.
$$

$$
\overline{a}
$$

*Proof* Since  $r \equiv y \pmod{2^n}$  and  $0 \le r < 2^n$ ,  $r = y \pmod{2^n} = y[n-1 : 0]$ . By Lemma [2.50,](#page-16-0)  $y = si(r, n)$  and hence

$$
x = sf(r, n, m) = 2^{m-n}si(r, n) = 2^{m-n}y.
$$