Floating Point Arithmetic

Floating Point Arithmetic

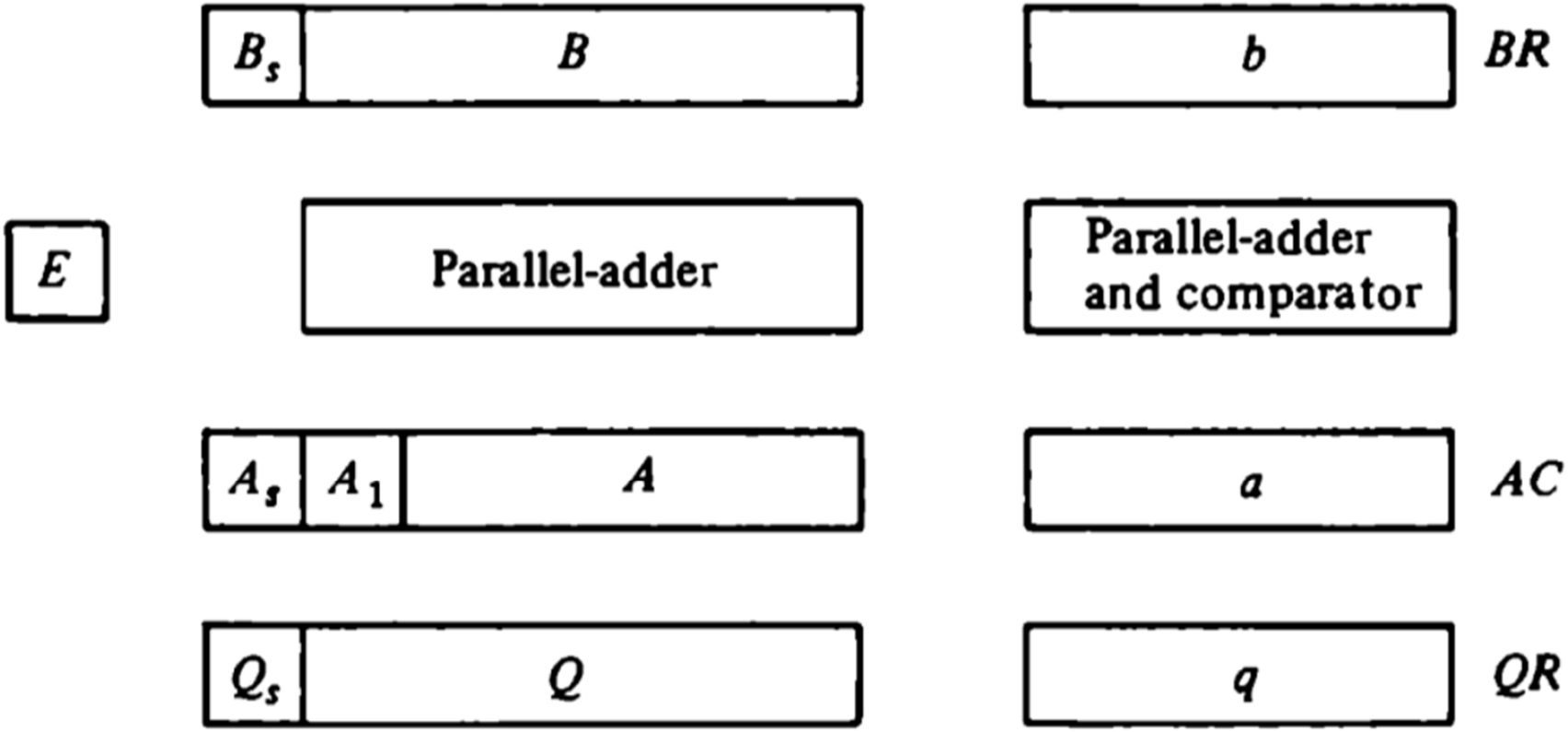
- Floating point number in computer register consists of two parts : a mantissa m and exponent e which will be represented as $m \times r^e$
- A floating point number that has a 0 in the most significant position of the mantissa is said to have an UNDERFLOW.
- To normalize a number that contains an underflow, it is necessary to shift the mantissa to the left and decrement the exponent until a nonzero digit appears in the first position.

Register configuration

- The register configuration for floating point operation is quite similar to the layout for fixed point operation.
- As a general rule, the same register and adder used for fixed point arithmetic are used for processing the mantissas.
- The difference lies in the way the exponents are handled.

Register configuration

Figure 10-14 Registers for floating-point arithmetic operations.



Register configuration

- There are three registers BR, AC and QR.
- Each register is subdivided into two parts.
- The mantissa part has same upper case letters, the exponent part uses the corresponding lower case letters.
- A parallel adder adds the two mantissas and transfers the sum into A and the carry into E.
- A separate parallel adder is used for the exponents. Since the exponents are biased.

Addition and subtraction

- During addition and subtraction , the two floating point operands are in AC and BR. The sum or difference is formed in the AC .
- The algorithm can be divided into four consecutive parts :
  1. Check for zeros.
  2. Align the mantissa.
  3. Add or subtract the mantissa.
  4. Normalize the result.

Addition and subtraction

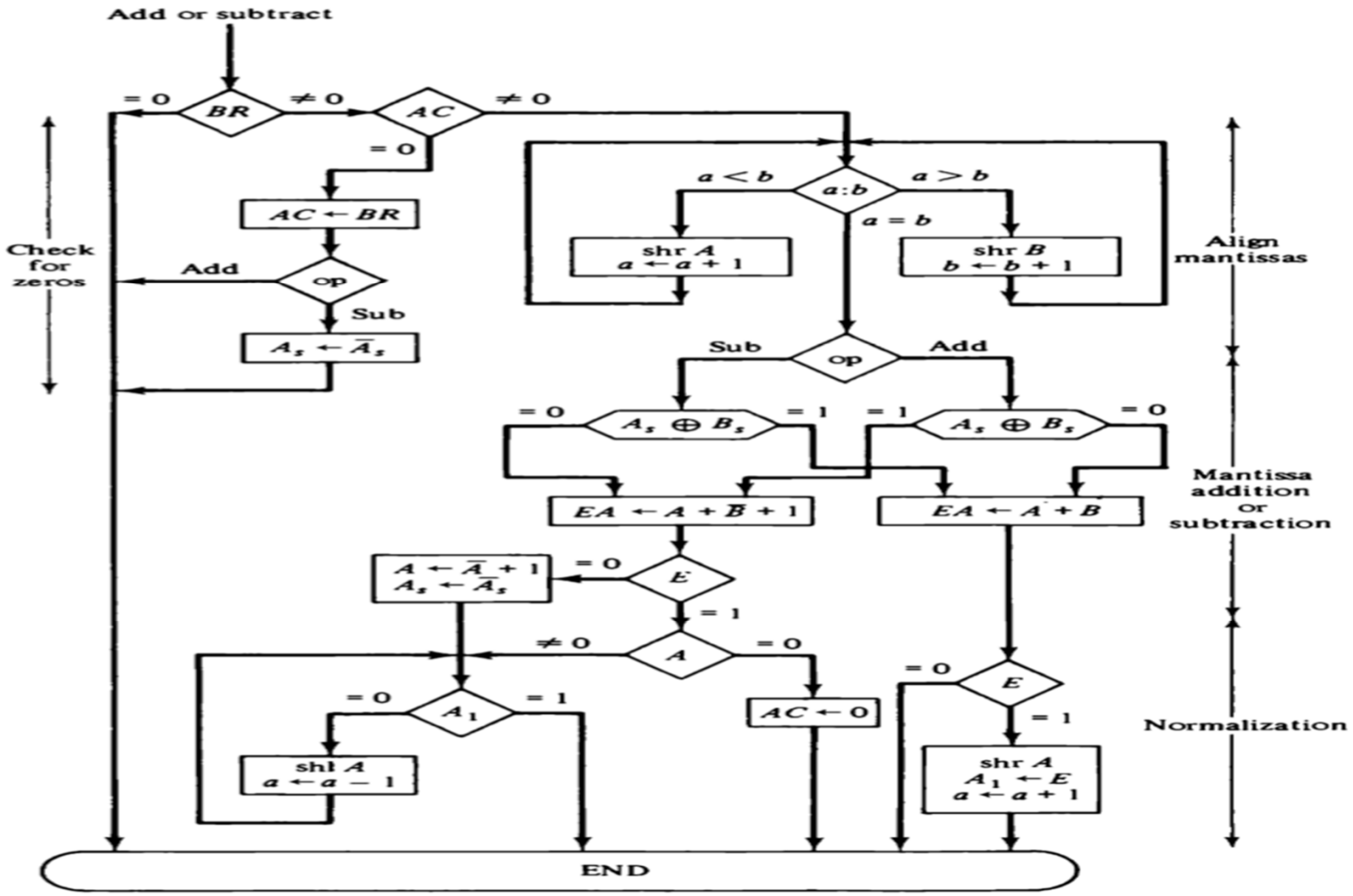
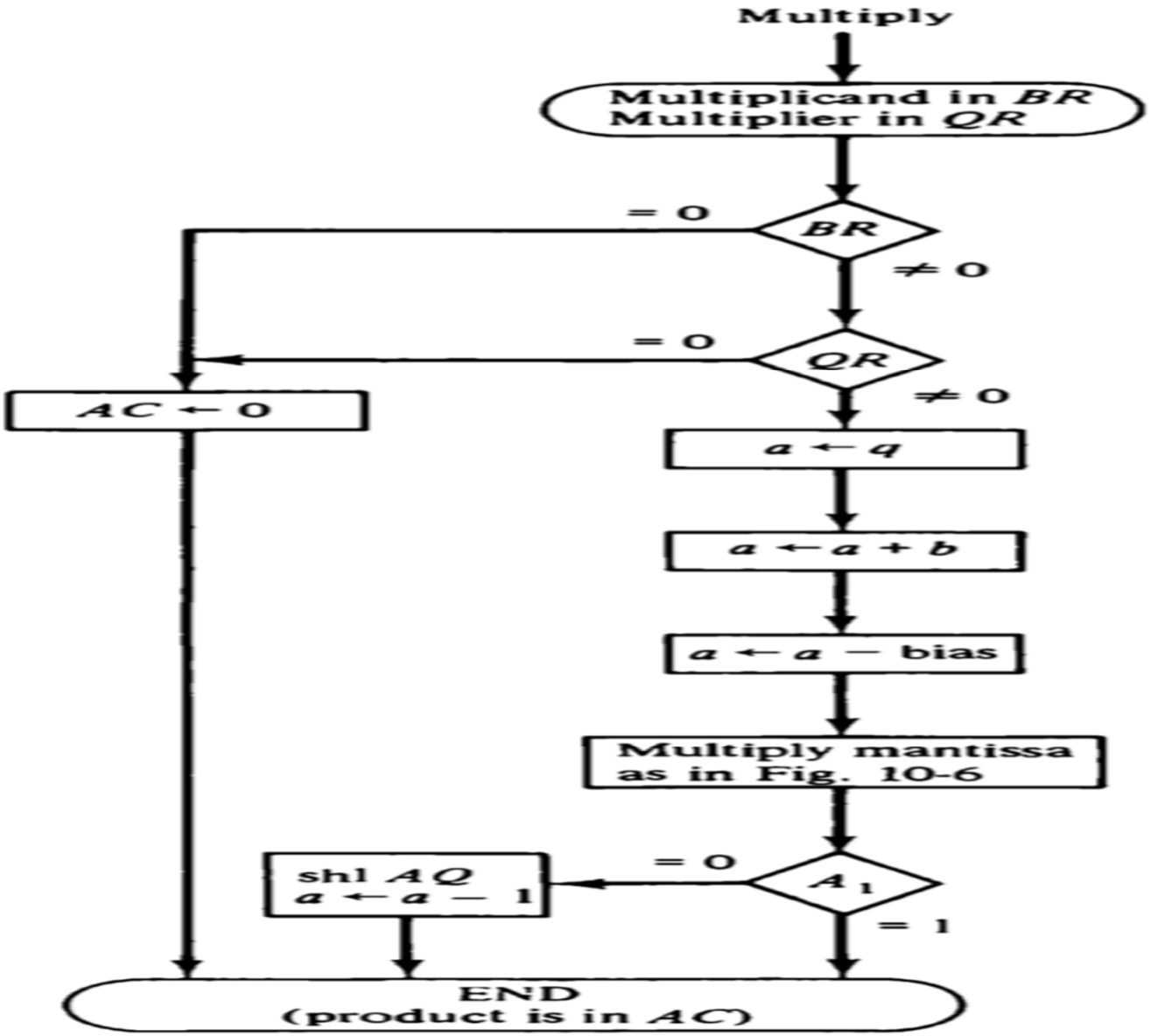


Figure 10-15 Addition and subtraction of floating-point numbers.

Multiplication

- The multiplication of two floating point numbers requires that we multiply the mantissas and add the exponents. No comparison of exponents or alignment of mantissa is necessary.
- The multiplication of the mantissa is performed same as fixed point to provide a double precision product.
- The multiplication algorithm can be subdivided into four parts :-
  1. Check for zeros.
  2. Add the exponents.
  3. Multiply the mantissa.
  4. Normalize the product.

Multiplication



Division

- Floating point division requires that the exponents be subtracted and the mantissa divided.
- The mantissa division is done as in fixed point division.
- The division algorithm can be divided into five parts..
  1. Check for zeros.
  2. Initialize registers and evaluate the sign.
  3. Align the dividend
  4. Subtract the exponents.
  5. Divide the mantissa.

Division

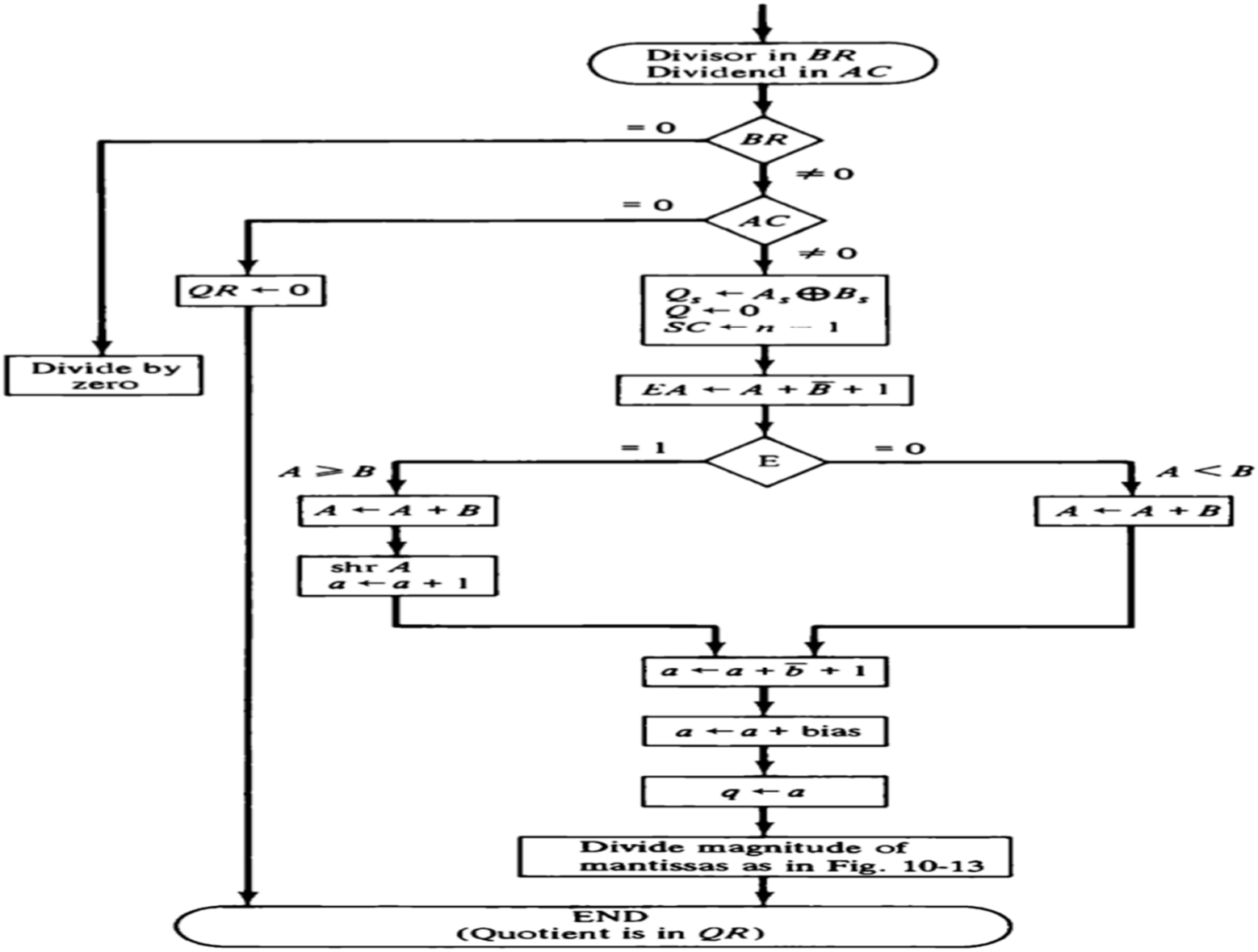


Figure 10-17 Division of floating-point numbers.