FPSX

IEEE-754 Floating-Point Modules
High Performance Single Precision with Flush-to-Zero Underflow

OPERAND A 32 | ADD UNIT (FPA) | RESULT 32
OPERAND B 32 | | FLAGS
OPCODE
ROUND MODE

OPERAND A 32 | COMPARE UNIT (FPC) | RESULT 32
OPERAND B 32 | | FLAGS
OPCODE

OPERAND A 32 | MUL UNIT (FPM) | RESULT 32
OPERAND B 32 | | FLAGS
OPERAND C 32
OPCODE
ROUND MODE

OPERAND 32 | CONVERT UNIT (FPV) | RESULT 32
| | FLAGS
OPCODE
ROUND MODE
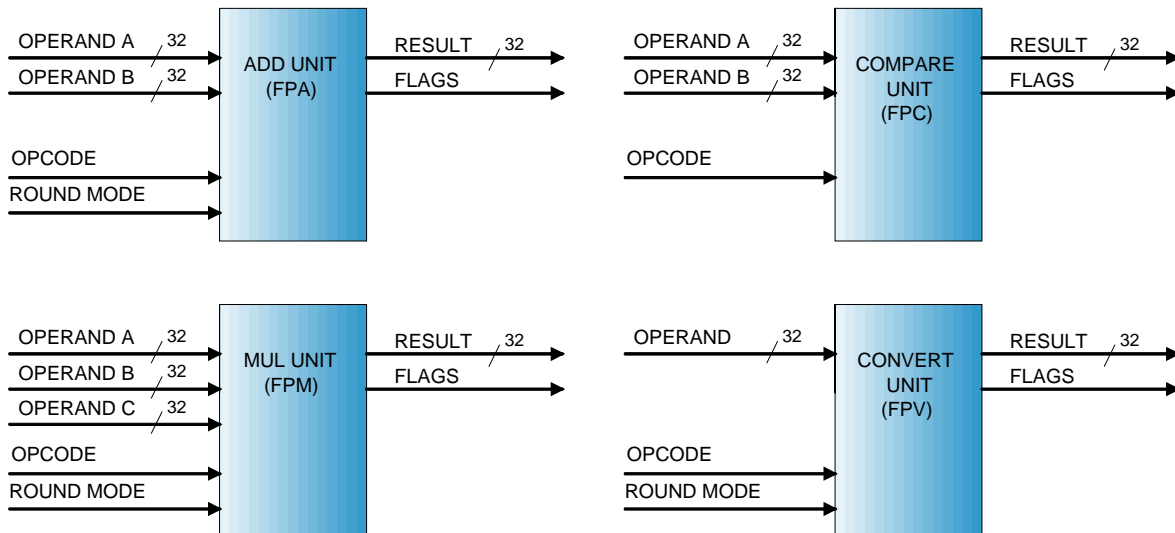
Overview

The FPSX product is a collection of floating-point execution units compliant with the *ANSI/IEEE Std 754-1985, IEEE Standard for Binary Floating-Point Arithmetic* (IEEE-754 Standard). The units are designed for high frequency, high throughput implementations. Each unit is implemented as a state-less pipeline that can easily be integrated into a high-performance processor design.

Each unit is targeted for a clock cycle with only 10-12 gate delays (excluding setup and clock skew). The add and multiply units implement 5-cycle pipelines. The conversion unit implements a 2-cycle pipeline, and the compare unit a 1-cycle pipeline. All units can sustain a 1-cycle throughput.

The multiply unit supports fused floating-point multiply add, as well as a variety of 32-bit integer multiply functions.

Features

- IEEE-754 compliant (except underflow)
- Flush-to-Zero underflow implementation
- Single precision instructions
- Add unit (FPA), 5-stage pipeline
    - Floating-point add/subtract
- Multiply Unit (FPM), 5-stage pipeline
    - Floating-point multiply
    - Floating-point multiply-add
    - 32-bit integer signed/unsigned multiply
    - 32-bit integer signed multiply, round and shift
    - 32-bit integer signed multiply and shift
    - 32-bit integer signed multiply, return abs
- Conversion unit (FPV), 2-stage pipeline
    - Convert floating-point to/from 32-bit integer
- Compare unit (FPC), 1-stage pipeline
    - Floating-point compare
    - Floating-point min/max/saturate
    - Floating-point NAN test
    - Floating-point absolute value
- All IEEE rounding mode supported
- All IEEE exception flags supported

GB3

GB3 Digital Systems

IEEE-754 Compliance

The FPSX modules are designed to provide a powerful floating-point capability while minimizing die size cost. To minimize unnecessary design size, some of the rarely used features of the IEEE specification are not implemented directly in the hardware design. The following IEEE-defined operations are not directly supported in FPSX hardware, but can be supported with software support:

- Gradual Underflow
- Denormal Numbers

In place of gradual underflow, the FPSX modules implement a flush-to-zero approach when underflow occurs. This feature allows the FPSX modules to maintain a one-cycle throughput in all operand cases, and minimizes design size.

Performance

Size:
Mul unit:	42,000 NAND Gates
Add unit:	12,000 NAND Gates
Convert Unit:	6,600 NAND Gates
Compare Unit:	1,400 NAND Gates

Timing: 1 GHz clock on 45nm technology

NOTE: The above performance data are estimates only, based on sample implementations using worst-case conditions. Achieved performance is highly dependent on the process technology, cell library, and synthesis tools used.

Instruction Timing

Instruction	Throughput / Latency
Floating-point add, subtract, multiply, multiply-add	1/5
Integer multiply, multiply round and shift, multiply shift, multiply return abs	1/5
Floating-point to/from integer conversions	1/2
Floating-point compare, min, max, saturate	1/1
Floating-point absolute value, NaN test	1/1