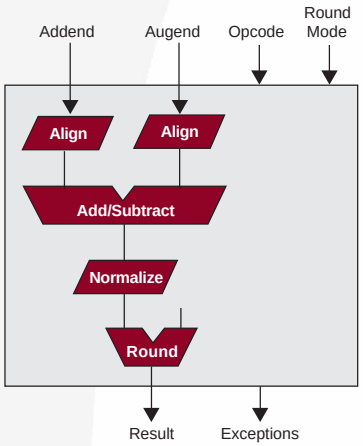
Floating Point
add unit



Features

- IEEE-like Arithmetic (Denormals treated as Zeros)
- IEEE Single Precision Format
- Fully Combinational Design

Description

Five operations are supported by the Floating Point Add Unit:

- Floating Point Add
- Floating Point Subtract
- Convert from Floating Point to Integer
- Convert from Integer to Floating Point
- Floating Point Compare.

Each of these operations follows the IEEE 754 Standard for Floating Point Arithmetic for Single Precision (32-bit) arithmetic with the exception that Denormals are treated as like-signed Zeros. An additional exception flag is provided to signal when a Denormal input or what would have been a Denormal output is flushed to Zero. Results for exceptions are the IEEE Standard default results as defined for the case when no trap occurs.

NaN formats may be specified with a configurable QNaN bit value and bit position, and a default QNaN may be specified for Invalid results. Floating Point to Integer conversion results may be independently specified for the Overflow, Underflow, and Invalid cases.

The Floating Point Add Unit is a fully combinational unit. The design is coded in a pipelined fashion, and pipeline registers may be easily added to support higher clock frequencies.

The Floating Point Add Unit is designed to provide the Add and Subtract operations with minimal latency, at the expense of some extra area. The Convert and Compare operations are designed to use the Add/Subtract datapath with minimal impact. A separate Floating Point Compare Unit is available for a more optimal implementation of the Compare operation.

The approximate gate count and latency for the Floating Point Add unit is 7,200 gates and 7.5 ns for a typical 0.18 micron Standard Cell technology.

Applications

- Microcontrollers
- Microprocessors

SLE
Silicon Logic Engineering

Phone: (715) 830-1200 • Fax: (715) 830-1887
Email: sales@siliconlogic.com