Enhanced Floating Point Unit IP Core

Introduction

This is a fully configurable collection of Floating Point functions that are fully IEEE-754-1985 compliant and can be used to build floating point units and dedicated functions that require floating point math.

Features

Fully IEEE 754-1985 Compliant Floating Point Functions. This enhanced implementation is fully parameterizable. Sample precisions are:

- Single Precision (32 bit)
- Extended Single Precision (44 bit)
- Double Precision (64 bit)
- Extended Double Precision (80 bit)
- Any other user selected format

The enhanced floating point unit consists of the following building blocks:

- Floating Point Add/Subtract Unit (fasu)
- Floating Multiply Unit (fmul)
- Floating Point Divide Unit (fdiv)
- Floating Point Compare Unit (fcmp)
- Integer to Floating Point Conversion (i2fp)
- Floating Point to Integer Conversion (fp2i)

Additional functions can be easily added per customers request.

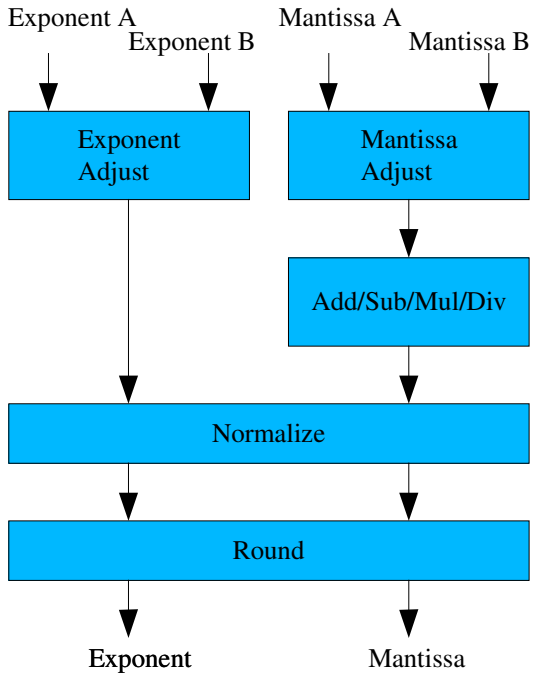
All functions are written in Verilog as purely combinatorial blocks.

Each block is available with any number of pipeline stages. A pipeline can be inserted and expanded to the required number of stages to meet customers performance demands. After insertion of the pipeline, retiming of each block will ensure maximum performance and optimization. This way any desired clock frequency can be achieved.

All blocks handle normalized and denormalized numbers according to IEEE 754-1985 standard, and provide various signals to identify special numbers.

Architecture

Architecturally all blocks look very similar even though the operations performed each step are quite different. Below diagram gives a conceptual overview of each function.



Verification

All floating point functions have been thoroughly exercised. A test bench with several million test vectors was developed. Utilizing open IEEE-754 compliant C libraries a test pattern generator was developed to automatically create directed and random test patterns. We do everything possible to ensure all our cores are 100% bug free !

Size And Speed

Sample Synthesis results for an 0.18u process. The goal was smallest and fastest implementation.

Function	*Gate Count*	*Propagation Delay*
32 bit Add/Subtract Unit	4,488	11.7 nS
32 bit Multiply Unit	12,155	13.1 nS
32 bit Divide Unit	30,294	39.2 nS
32 bit Compare Unit	514	2.3 nS
64 bit Add/Subtract Unit	9,069	13.1 nS
64 bit Multiply Unit	38,568	22 nS
64 bit Divide Unit		
64 bit Compare Unit	1,028	2.6 nS

These synthesis results are provided for *reference only*. Please contact us for estimates for your application.