

Dr. B. Vinoth Kumar  
Dr. K.N. Vijeyakumar  
K. Saranya

# Single Precision Floating Point Multiplier



**Anchor Academic Publishing**

*disseminate knowledge*

**Vinoth Kumar, B., Vijeyakumar, K.N., Saranya, K.: Single Precision Floating Point Multiplier, Hamburg, Anchor Academic Publishing 2017**

PDF-eBook-ISBN: 978-3-96067-655-3

Druck/Herstellung: Anchor Academic Publishing, Hamburg, 2017

**Bibliografische Information der Deutschen Nationalbibliothek:**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

**Bibliographical Information of the German National Library:**

The German National Library lists this publication in the German National Bibliography. Detailed bibliographic data can be found at: <http://dnb.d-nb.de>

All rights reserved. This publication may not be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

---

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden und die Diplomica Verlag GmbH, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

Alle Rechte vorbehalten

© Anchor Academic Publishing, Imprint der Diplomica Verlag GmbH  
Hermannstal 119k, 22119 Hamburg  
<http://www.diplomica-verlag.de>, Hamburg 2017  
Printed in Germany

## **ACKNOWLEDGMENTS**

It is a great opportunity to express our sincere thanks to all who have contributed to do this work through their support, encouragement and guidance.

## **ABSTRACT**

The Floating Point Multiplier is a wide variety for increasing accuracy, high speed and high performance in reducing delay, area and power consumption. The floating point is used for algorithms of Digital Signal Processing and Graphics. Many floating point multipliers are used to reduce the area that perform in both the single precision and the double precision in multiplication, addition and subtraction.

The scientific notations sign bit, mantissa and exponent are used. The real numbers are divided into two, fixed component of significant range (lack of dynamic range) and exponential component in floating point (largest dynamic range). Converting decimal to floating point and normalize the exponent part and rounding operation for reducing latency. The mantissa of two values are multiplied and adding the exponent part. The sign result with exclusive-or are obtained. The final result of shift and add floating point multiplier is compared with booth multiplication. From the synthesis it is observed that shift and add floating point multiplier performs 20% better in different parameters such as delay, area and power is verified by Verilog Hardware Description Language. These results are also verified in Cadence EDA tool.

## TABLE OF CONTENTS

| CHAPTER<br>NO | TITLE                                                    | PAGE NO |
|---------------|----------------------------------------------------------|---------|
|               | ABSTRACT                                                 | ii      |
|               | LIST OF TABLES                                           | vii     |
|               | LIST OF FIGURES                                          | viii    |
|               | LIST OF ABBREVIATIONS                                    | x       |
| 1             | INTRODUCTION                                             | 1       |
|               | 1.1 Format Parameters                                    | 1       |
|               | 1.2 Data formats for single and double precision         | 2       |
|               | 1.3 Representation of the floating point                 | 2       |
|               | 1.3.1 Denormalized                                       | 3       |
|               | 1.3.2 Over flow                                          | 3       |
|               | 1.3.3 Under flow                                         | 3       |
|               | 1.3.4 Infinity                                           | 3       |
|               | 1.3.5 Not a Number                                       | 3       |
| 2             | LITERATURE SURVEY                                        | 4       |
|               | 2.1 Floating point multiplier using Vedic<br>mathematics | 4       |

|          |                                                    |          |
|----------|----------------------------------------------------|----------|
| 2.2      | Floating point operation in fast fourier transform | 4        |
| 2.3      | Multiplication using carry save multiplier         | 5        |
| 2.4      | Parallel implementation of floating point          | 5        |
| 2.5      | Normalization of floating point                    | 6        |
| 2.6      | Configurable booth multiplier                      | 6        |
| 2.7      | Implementation on FPGA                             | 7        |
| 2.8      | Different multipliers                              | 7        |
| 2.9      | Double precision floating point                    | 8        |
| <b>3</b> | <b>FLOATING POINT MULTIPLICATION OPERATION</b>     | <b>9</b> |
| 3.1      | Fraction                                           | 9        |
| 3.2      | Representation of floating point multiplication    | 10       |
| 3.3      | Floating point algorithm                           | 11       |
| 3.3.1    | Convert 2.625 to floating point format             | 11       |
| 3.3.2    | Adding an exponent part to binary number           | 11       |
| 3.3.3    | Normalization                                      | 12       |
| 3.3.4    | Mantissa                                           | 12       |