基于SIMD部件的四倍精度浮点乘加器设计

何 军 黄永勤 朱 英
(上海高性能集成电路设计中心 上海 201204)

摘 要 如何减少四倍精度浮点运算的硬件开销和延迟是需要解决的重要问题。为减少四倍精度乘加器的硬件开销,基于支持64位×4的双精度浮点SIMD FMA部件,设计并实现了一种新的四倍精度浮点乘加器(QPFMA),来支持4种浮点乘加运算和乘法、加减法、比较运算,运算延迟为7拍。通过将四倍精度113位×113位尾数乘法器分解为4个57位×57位乘法器来共享双精度浮点SIMD FMA部件的53位×53位乘法器,显著减少了实现QPFMA的硬件开销。基于65nm工艺的逻辑综合结果表明,该QPFMA频率可达1.1GHz,面积是常规QPFMA设计的42.71%,仅与一个双精度浮点乘加器相当。与现有的QPFMA设计相比,相当工艺和频率下,其运算延迟减少了3拍,门数减少了65.96%。

关键词 浮点,SIMD部件,乘加,四倍精度,高精度
中图法分类号 TP332.2 **文献标识码** A

Design of Quadruple Precision Floating-point Fused Multiply-Add Unit Based on SIMD Device

HE Jun HUANG Yong-qin ZHU Ying
(Shanghai Hi-Performance IC Design Centre, Shanghai 201204, China)

Abstract It is an important issue to resolve to decrease the hardware cost and operation latency for the implementation of quadruple precision floating-point arithmetic. To decrease the hardware cost of floating-point quadruple fused multiply add (QPFMA) unit, a new QPFMA unit was designed and realized based on a SIMD device, which supports 64bit×4 double precision floating-point fused multiply add (DPFMA). The new QPFMA supports four kinds of FMA operation, multiplication, addition, subtraction and comparison, with the operation latency of 7 cycles. By decomposing the 113bit×113bit multiplication of quadruple precision fraction into four 57bit×57bit multiplications to share the 53bit×53bit multipliers of SIMD DPFMA, the hardware cost of the new QPFMA is reduced greatly. Using the 65nm cell library, the new QPFMA is synthesized. The results show its frequency is 1.1GHz and area is 42.71% of a normal QPFMA unit, only equal to the area of a DPFMA unit. Comparing to current QPFMA design, the operation latency decreases by 3 cycles and the gate number reduces by 65.96% in equivalent technology and at comparative frequency.

Keywords Floating-point, SIMD device, Fused multiply-add, Quadruple precision, High precision

1 引言

现代处理器一般硬件支持双精度(64位)或者扩展双精度(80位)浮点算术运算,但是对很多科学计算应用,比如气候模拟、超新星模拟、计算物理、计算几何、计算数论等,双精度或者扩展双精度浮点已经不能满足需求了[1]。浮点数据表示和运算都存在一定误差,且误差会在多次运算中累积增大,从而导致计算结果不精确、不可信。鉴于高精度浮点计算在提高计算结果精度、数值算法稳定性和结果可再现性等方面的优势,IEEE 754-2008浮点标准[2]中增加了四倍精度(Quadruple Precision, QP)浮点数据类型(binary128),以支持高精度浮点计算。此外,鉴于浮点乘加(Fused Multiply-Add, FMA)运算在提高浮点运算精度和性能方面的显著优点,浮点FMA运算也成为IEEE 754-2008浮点标准的基本运算之一。如何高效地实现高精度、高性能浮点运算部件已成为国内外的研究热点之一。

根据IEEE 754-2008浮点标准[2],QP浮点数据格式,与单精度(SP)和双精度(DP)浮点数据格式一样,也由符号位S、阶码E和尾数T 3个域组成(见图1),其中符号位1位宽,阶码15位宽,尾数112位宽,数据总位宽128位,其中阶码偏移值(bias)为16383,规格化尾数隐含首位为1,实际尾数精度p为113位(见表1)。

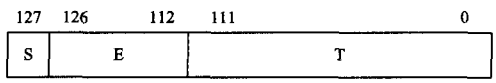

图1 QP浮点数据格式

相对于DP浮点数据格式,QP浮点数据位宽增加了1倍,其中阶码增加了4位,精度增加了1倍以上,因此QP浮点数表示的范围和精度都大幅度提高。相应地,实现QP浮点运算一方面硬件开销大,另一方面延迟也大,这是目前处理

到稿日期:2013-02-16 返修日期:2013-05-10

何 军(1980—),男,博士生,工程师,主要研究方向为微处理器设计,E-mail:joyhejun@126.com;黄永勤(1955—),女,高级工程师,博士生导师,主要研究方向为计算机系统结构、高性能计算;朱 英(1964—),女,硕士,高级工程师,主要研究方向为微处理器设计与验证。

器对其硬件支持不足的重要原因。如何减少 QP 浮点运算实现的硬件开销和延迟是需要解决的重要问题。

表 1 IEEE 754-2008 浮点数据格式

浮点数据格式	SP	DP	QP
符号位 S 位宽	1	1	1
阶码 E 位宽	8	11	15
尾数 T 位宽	23	52	112
精度 p 位宽	24	53	113
数据总位宽	32	64	128
bias	127	1023	16383

针对上述问题，本文设计了一种基于浮点 SIMD(Single Instruction Multiple Data，SIMD)部件的四倍精度浮点乘加器(QPFMA)，其支持 4 种浮点乘加运算(乘加、乘减、负乘加和负乘减)，还支持浮点乘法、浮点加减法和比较运算，运算延迟为 7 拍。考虑到乘法器是 QPFMA 的主要硬件开销，该 QPFMA 基于支持 64 位×4 的浮点双精度 SIMD FMA 部件，通过将四倍精度的 113 位×113 位乘法器分解为 4 个 57 位×57位乘法器，以共享双精度浮点 SIMD FMA 部件的53 位×53 位乘法器，实现 QPFMA 的乘法器，节省了乘法器的硬件开销，显著减少了 QPFMA 的实现代价。

通过算法设计优化、深入细致的时序分析和优化，采用 65nm 单元库自动综合的实现方式，该 QPFMA 的频率可达 1.1GHz，面积为 140315.40um^2，基本与一个双精度 FMA 部件相当。对于支持 64 位×4 的浮点双精度 SIMD FMA 部件来说，仅需要再增加 1/4 左右的面积即可实现四倍精度乘加运算，实现多功能的 SIMD FMA 部件。与文献[3]的 QPFMA 相比，相同工艺和相当的工作频率下，运算延迟减少了 3 拍，等效门数减少了 65.96%。

本文第 2 节简要介绍 QP 浮点运算和浮点 SIMD 扩展相关研究工作；第 3 节介绍基于 SIMD 部件的 QPFMA 的总体结构和关键部件；第 4 节介绍基于 SIMD 部件的 QPFMA 的逻辑综合实现；最后总结全文。

2 相关研究工作

2.1 QP 浮点运算

从学术界来看，针对科学计算对高精度和高可靠浮点运算的需求，近年来有不少针对四倍精度浮点运算的研究。人们先后对四倍精度浮点乘法[4](3 拍流水线)、浮点加法[5](3 拍流水线)和浮点除法[6](59 拍非流水)算法进行了研究，这些算法同时支持两种运算模式：一个四倍精度浮点运算和两个并行的双精度浮点运算。文献[7]设计了一种多功能的浮点乘加部件(4 拍流水线)，它支持一个四倍精度/双精度浮点运算和两个并行的双精度/单精度浮点运算，还支持双精度/单精度浮点点积运算。文献[8]也设计了一个类似的乘加部件，它支持一个四倍精度运算和两个并行的双精度运算，以及 3 拍流水线，但是乘法器部件采用了迭代复用双精度乘法器的方式，以减少硬件开销，所以四倍精度运算吞吐率为 1/2。文献[9]设计了一个 5 拍流水线的四倍精度浮点乘加部件，它在 SMIC 0.13um工艺下频率可达到 202MHz，后来文献[3]对该设计进行了流水线细化，扩展为 10 级流水线，在 SMIC 0.13um工艺下频率可达到 465MHz，在 TSMC 65nm 工艺下频率可达 1.075GHz。此外，文献[10]也在 FPGA 上设计了一个四倍精度浮点乘加部件，用来加速 LU 分解和 MGS-QR 分解计算，可以获得 42 倍到 97 倍的性能提升，且能获得更高精度结果和更低能耗。

从工业界来看，主流微处理器对四倍精度浮点算术运算的硬件支持还比较有限(IBM 的 POWER6 处理器支持四倍精度浮点加法运算[11]，在 SPARC v9 指令集中定义了四倍精度浮点数据格式和相关运算，但尚无 SPARC 处理器硬件支持四倍精度浮点运算)，主要靠软件模拟实现，如 Intel Fortran[12]、GMP、MP-FR[13]、QD(Quad-Double)[14] 等函数库。软件模拟计算性能低，相对双精度浮点运算，四倍精度浮点算术运算的性能降低了一个数量级，LINPACK 测试时间为双精度浮点运算的 36 倍[1]。

总之，支持四倍精度浮点算术运算是浮点部件未来发展的重要发展趋势之一，有利于进一步提高浮点运算精度，提高某些重要科学计算应用性能。学术界在四倍精度浮点运算方面，多侧重于理论上的研究，距离实际应用还有一定距离，而工业界对四倍精度浮点运算的硬件支持还比较有限。

2.2 浮点 SIMD 扩展

最初，短向量 SIMD 扩展是为了在桌面处理器上支持计算密集的多媒体应用，主要是针对整数运算进行扩展，后来也对浮点运算进行了扩展，以加速计算密集的科学计算应用，例如 Intel x86 处理器的 MMX(Multimedia extension)、SSE(Streaming SIMD extension)、AVX(Advanced vector extensions)扩展指令集，IBM POWER 处理器的 VMX(Vector/SIMD multimedia extension)指令集等。

目前高端处理都支持浮点短向量 SIMD 扩展，短向量长度一般为 64 位×2 或者 64 位×4，支持 2 路或者 4 路双精度浮点运算。Intel 在 Pentium 4 处理器上引入 SSE2 扩展指令集，支持 2 路双精度浮点 SIMD 运算，后来又相继引入了 SSE3、SSE4 扩展指令集，并且一直延续支持。2011 年，Intel 在 Sandy Bridge 架构处理器上引入 AVX 扩展指令集[15]，支持 4 路浮点双精度 SIMD 运算，之后 Intel 和 AMD 的 x86 处理器都支持 AVX 扩展指令集。IBM 在 PowerPC 970 G5 处理器上引入 VMX 指令集[16]，并在 POWER6[17]、POWER7[18] 处理器上也支持 VMX 指令集，支持 2 路双精度浮点 SIMD 运算。在最新的 Blue Gene/Q Compute Chip[18] 的处理器上(该处理器应用于 Blue Gene/Q 超级计算机系统，2012 年 6 月问鼎超级计算机 Top500 排行首位[20])，IBM 开始支持 4 路双精度浮点 SIMD 运算，且支持 64 位×4 的双精度浮点 FMA 运算。富士通在 SPARC64 VIIIfx 处理器[21]上(该处理器应用于 K 超级计算机系统，2011 年 6 月问鼎超级计算机 Top500 排行首位)，也支持 4 路双精度浮点 SIMD 运算，且支持 64 位×4 双精度浮点 FMA 运算。

短向量 SIMD 扩展技术已经成为进一步提高处理器浮点性能的重要手段，支持 4 路双精度浮点 FMA 运算已经成为高端处理器的主流配置。

3 基于 SIMD 部件的 QPFMA 设计

3.1 算法与设计思想

假设 FMA 运算：$F=A\times B+C$，其中 A、B 为乘数和被乘数，C 为加数。令 $C=0$ 时，则 $F=A\times B+0$，相当于实现了乘法运算；令 $B=1.0$，则 $F=A\times 1.0+C$，相当于实现了加法运算。此外，还可以对 FMA 运算进行扩展，实现乘减($F=A\times$

$B-C$)、负乘加($F=-A\times B+C$)、负乘减($F=-A\times B-C$)等运算。

浮点 FMA 运算最早在 IBM RS/6000 上实现，其算法是实现 FMA 运算的经典算法，其主要流程概述如下(主要是尾数部分处理，参见 3.2 节图 2 中的 DPFMA0)：

1)尾数相乘：被乘数与乘数尾数相乘得到保留进位形式乘积(Carry，Sum)；

2)加数对阶：在尾数相乘的同时，对加数进行对阶移位(如果是减法运算需要对加数求补码)，使加数阶码和乘积阶码相等；

3)乘积与加数求和：将移位后的加数与乘积，通过 CSA(Carry Save Adder，CSA)3∶2 压缩和拼接得到新的(Carry`，Sum`)形式和，然后利用进位传递加法器(Carry Propagate Adder，CPA)求和(如果结果为负数，还需要对负数结果求补码得到正数结果)；

4)结果规格化：在乘积与加数求和的同时，利用头零预测器(Leading Zero Anticipation，LZA)预测结果中首 1 的位置，然后根据 LZA 预测结果进行规格化移位，使结果的最高位为 1(由于 LZA 预测存在误差，还需要进行稍微修正)；

5)结果舍入：根据舍入模式，对规格化的结果进行舍入得到最终结果(如果结果溢出，还需要进行再规格化)。

在经典算法基础上，后来有不少旨在减少 FMA 运算延迟的研究，这些改进算法可以减少一定的运算延迟，但都以增加硬件开销为代价。为了减少硬件开销，本文 FMA 运算算法以经典算法为基础，对算法结构进行了部分的设计优化，以减少运算延迟，主要改进包括：采用双路加法器避免了对负数结果的求补；优化 LZA 逻辑，减小了 LZA 逻辑的宽度，也减少了规格化移位的级数。

根据经典 FMA 算法，我们分别设计了双精度 FMA 部件(DPFMA)和 QPFMA 部件，经过逻辑综合评估(见表 2)，经典 QPFMA 的面积约是经典 DPFMA 面积的 2.23 倍，113 位乘法器是经典 QPFMA 部件的主要硬件开销，占经典 QPFMA 面积的 65.20%。

表 2　经典 DPFMA 和 QPFMA 综合结果

设计	面积(um^2)	百分比
DPFMA(53 位×53 位)	147,268.80	100.00%
DPFMA(57 位×57 位)	161,080.80	109.38%
经典 QPFMA	328,504.81	223.06%

一方面为了减少硬件开销，另一方面考虑到支持 4 路双精度浮点乘加运算已经成为高端处理器的主流配置，可以基于支持 64 位×4 的双精度浮点 SIMD FMA 部件，将四倍精度的 113 位×113 位乘法器分解为 4 个 57 位×57 位乘法器，共享双精度浮点 SIMD FMA 部件的 53 位×53 位乘法器，实现 QPFMA 的乘法器，从而节省 113 位乘法器的硬件开销。为此需要将双精度乘加器的 53 位乘法器扩展为 57 位乘法器，其增加的硬件开销为 9.38%(见表 2)。

3.2　总体结构

基于双精度浮点 SIMD FMA 部件的 QPFMA 总体结构如图 2 所示，F_SIMDFMA 含有 4 个采用经典算法 DPFMA 部件(DPFMA0-3)实现双精度 FMA 运算和 1 个采用经典算法 SIMD_QPFMA 部件实现四倍精度 FMA 运算。DPFMA 为 6 级流水线(图 2 中虚线 ST0－5 为流水线站台触发器位置)，其中前两级流水线站台 ST0 和 ST1 进行尾数相乘和加数对阶移位；流水线站台 ST2 和 ST3 进行乘积与加数求和、头零预测，并进行结果规格化；流水线站台 ST4 和 ST5 进行结果规格化修正和舍入。

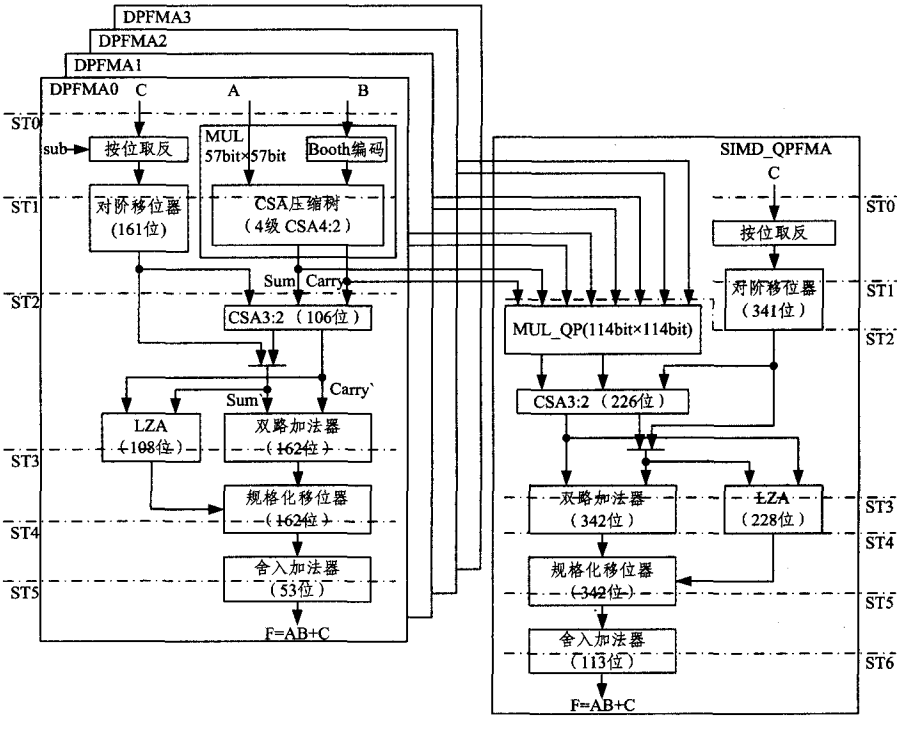

图 2　F_SIMDFMA 总体结构

SIMD_QPFMA 为 7 级流水线(图 2 中虚线 ST0－6 为流水线站台触发器位置)，其中前两级流水线站台 ST0 和 ST1 进行 57 位尾数相乘和加数对阶移位，这里 4 段 57 位的尾数乘积分别利用 DPFMA0-3 部件的 57 位乘法器得到(详见 3.3.1 节)；流水线站台 ST2 利用 4 段 57 位乘积得到 114 位乘积，并进行一部分乘积与加数的求和，以及一部分头零预测；

流水线站台 ST3 进行剩余的乘积与加数的求和、头零预测；流水线站台 ST4 进行结果规格化移位；流水线站台 ST5 和 ST6 进行结果规划化修正和舍入。

下面将对几个关键部件进行详细介绍，包括 114 位乘法器、342 位双路加法器，受篇幅所限，其它部件就不再赘述。

3.3 关键部件

3.3.1 114 位乘法器

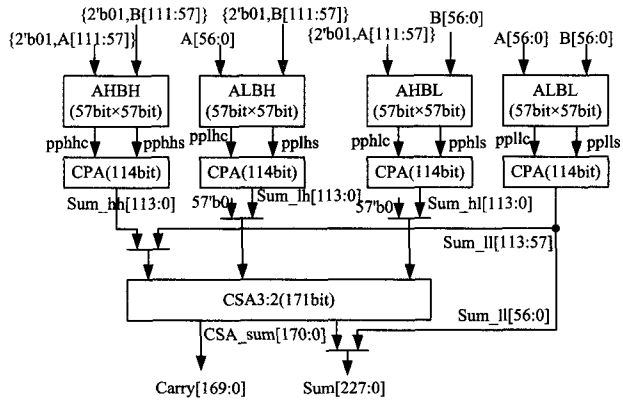

图 3 114 位乘法器结构

114 位乘法器用于实现四倍精度尾数乘积，为了减少硬件开销，分别按源操作数 A、B 的高低 57 位分解为 4 段：ALBL、AHBL、ALHB 和 AHBH，分别利用 4 个 DPFMA 的乘法器得到 4 段 57 位乘积，然后按照不同的权重值对齐求和，可实现 114 位的乘法器。具体来说（见图 3）：4 段 57 位乘法器分别得到两个部分积（pp * c，pp * s），分别经过 114 位 CPA 求和得到各自的和 Sum_ * [113:0]。其中 Sum_ll[56:0]即为最终 114 位乘法器乘积 Sum 的低 57 位，Sum_ll[113:57]与 Sum_hh[113:0]拼接，Sum_hl 和 Sum_lh 高位分别补上 57 位零，再经过 171 位 CSA3:2 得到最终 228 位乘积的 Carry[169:0]和 Sum[227:0]，需要注意这里的 Carry 和 Sum 是按照最高位对齐的，求和时应该按照最高位对齐，而不是最低位对齐。另外，对于四倍精度乘法而言，只需要得到 226 位乘积，高 2 位可以直接舍弃。

3.3.2 342 位双路加法器

342 位双路加法器用于实现对四倍精度尾数乘积和对阶后加数的求和，为了避免对负数结果的求补码操作，直接得到尾数乘积与对阶加数的绝对值。

考虑到 342 位加法器的长延迟和流水线的时序要求，结合环回进位（End-Around-Carry）加法器和选择进位加法器（Carry Select Adder）的思想，设计了该双路加法器（见图 4）。该双路加法器从低位到高位分为 7 段实现，每一段由两个并行的加法器（低 6 段 49 位，最高段 48 位）组成，这两个加法器的进位输入分别为 0 和 1，加数相同。首先这 7 段加法器并行各自求和，并计算出各自的进位输出 *Cout*0[i]和 *Cout*1[i]（i=0,1,…,6），然后根据最低段加法器的进位输出 *Cout*0[0]和 *Cout*1[0]分别确定相邻高段加法器的和 *S*0 和 *S*1 与进位输出 *Cout*0[1]* 和 *Cout*1[1]*，再依次根据低段加法器的进位输出 *Cout*0[i]* 和 *Cout*1[i] * 分别确定相邻高段加法器的和 *S*0 和 S1 与进位输出 *Cout*0[i+1]* 和 *Cout*1[i+1]*，最终分别确定最高段加法器的和 S0 和 S1 与进位输出 Cout。如果实际是减法且 Cout 为 0，表示求和结果为负值，将 S0 逐位取反得到的就是最后的结果，否则 S1 就是最后的结果。

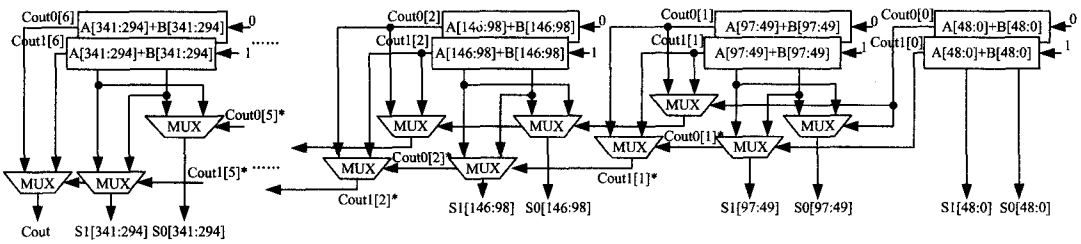

图 4 342 位双路加法器

该双路加法器能直接计算出求和的绝对值，避免了对负数结果的求补码，节省了相当于一个 342 位加法器的延迟。该双路加法器的延迟相当于一个 49 位加法器加上 6 级二路选择器的延迟。

4 逻辑综合实现

基于 65nm 单元库，使用 Synopsys 的 Design Compiler 对设计的 Verilog 代码进行逻辑综合，综合结果表明（见表 3），SIMD_QPFMA 关键路径在于第三级站台，工作频率可达 1.1GHz，面积 140315.40um^2，等效门数 77953，面积与一个 DPFMA 相当。

为了进行比较分析，我们还设计实现了一个基于经典 FMA 算法的 QPFMA，7 级流水线，在 1.1GHz 工作频率下，面积为 328,504.81um^2，硬件开销与 2.23 个 DPFMA 相当（见表 2）。SIMD_QPFMA 的硬件开销仅为经典 QPFMA 的 42.71%，减少了 57.29%。

表 3 QPFMA 综合结果

设计	流水线级数	频率 (GHz)	门数	面积 (um^2)	百分比
文献[3]QPFMA	10	1.075	229,000	不详	100.00%
SIMD_QPFMA	7	1.1	77,953	140,315.40	34.04%
经典 QPFMA	7	1.1	182,503	328,504.81	79.70%
DPFMA	6	1.1	81,816	147,268.80	35.73%

文献[3]也设计了一个 QPFMA，10 级流水线，它在 TSMC 65nm 工艺下频率可达 1.075GHz，与之相比，本文的 QPFMA 运算在相同工艺下，工作频率基本相当，延迟减少了 3 拍，门数减少了 65.96%。

从整个 F_SIMDFMA 部件来看（见表 4），SIMD_QPFMA 的面积占比为 20.18%，与 DPFMA 的面积基本相当。这表明，基于 64 位×4 的双精度浮点 SIMD FMA 部件，仅需要增

（下转第 51 页）

[6] Wang Ju-jie, Wang Jian-zhou, Zhang Zhe-george, et al. Stock index forecasting based on a hybrid model[J]. Omega, 2012, 40 (6): 758-766

[7] Kao Ling-jing, Chiu Chih-chou, Lu Chi-jie, et al. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting [J]. Neurocomputing, 2013, 99(1): 534-542

[8] Vapnik V. The nature of statistical learning theory[M]. New York, USA: Springer-Verlag, 1995

[9] Yeh Chi-yuan, Huang Chi-wei, Lee S-J. A multiple-kernel support vector regression approach for stock market price forecasting[J]. Expert Systems with Applications, 2011, 38(3): 2177-2186

[10] Cai C X, Kyaw K, Zhang Q. Stock index return forecasting: The information of the constituents[J]. Economics Letters, 2012, 116 (1): 72-74

[11] Taylor S J, Yadav P K, Zhang Yuan-yuan. The information content of implied volatilities and model-free volatility expectations: Evidence from options written on individual stocks[J]. Journal of Banking & Finance, 2010, 34(4): 871-881

[12] Kim Y, Sohn S Y. Stock fraud detection using peer group analysis[J]. Expert Systems with Applications, 2012, 39(10): 8986-8992

[13] Daly, Ronan. Learning Bayesian Networks: Approaches and Issues[J]. Knowledge Engineering Review, 2011, 26(2): 99-157

[14] Bui A T, Jun C H. Learning Bayesian network structure using Markov blanket decomposition[J]. Pattern recognition Letters, 2012, 33(16): 2134-2140

[15] Pearl J. *Probabilistic Reasoning in Intelligent Systems* [M]. Morgan Kaufmann, 1988

(上接第 18 页)

加 1/4 左右的硬件开销，即可实现四倍精度 FMA 运算。

表 4 F_SIMDFMA 综合结果

设计	面积(um²)	百分比
F_SIMDFMA	809070.61	100.00%
DPFMA_3	161080.80	19.91%
DPFMA_2	160904.40	19.89%
DPFMA_1	160993.20	19.90%
DPFMA_0	161045.40	19.90%
SIMD_QPFMA	163279.80	20.18%

从上述逻辑综合结果来看，基于 64 位×4 的双精度浮点 SIMD FMA 部件设计 QPFMA 可以显著减少硬件开销。对于已有的双精度浮点 SIMD FMA 部件来说，只需要增加少量硬件开销，即可实现 4 倍精度 FMA 运算。

结束语 本文完成了一种基于 SIMD 乘加部件的 QPFMA 的原型设计、验证与逻辑综合，并与其它设计进行比较分析，主要贡献在于：设计了基于 64 位×4 的双精度浮点 SIMD FMA 部件的 QPFMA 结构，7 级流水线，面积与一个双精度 FMA 部件基本相当，显著减少了实现四倍精度 FMA 运算的延迟和硬件开销。本文的研究也为浮点 SIMD 部件的设计提供了一条重要思路，以少量的硬件开销实现 SIMD 部件功能的扩展，进一步发挥 SIMD 部件的作用。下一步将尝试进一步对设计进行流水线时序优化，重点对现有设计的关键路径进行优化研究，以实现更高的频率。

参 考 文 献

[1] Bailey D H. High-precision floating-point arithmetic in scientific computation [J]. Computing in Science and Engineering, 2005, 7 (3): 54-61

[2] IEEE Computer Society. IEEE Standard for Floating-Point Arithmetic[S]. IEEE Standard 754-2008, 3 Park Avenue New York, NY 10016-5997, USA, August 2008

[3] 黎铁军，李秋亮，徐炜遐. 一种 128 位高性能全流水浮点乘加部件[J]. 国防科技大学学报，2010，32(2)：56-60

[4] Akkas A, Schulte M J. Dual-Mode Floating-Point Multiplier Architectures with Parallel Operations [J]. Journal of Systems Architecture, 2006, 52: 549-562

[5] Akkas A. Dual-Mode Quadruple Precision Floating Point Adder [C] // 9th Euromicro Conference on Digital System Design. 2006: 211-220

[6] Akkas A. A Dual-Mode Quadruple Precision Floating-Point Divider[C] // Fortieth Asilomar Conference on Signals, Systems and Computers. 2006: 1697-1701

[7] Gok M, Ozbilen M M. Multi-functional floating-point MAF designs with dot product support [J]. Microelectronics Journal, 2008, 39(1): 30-43

[8] Huang Li-bo, Ma Sheng, Shen Li, et al. Low-Cost Binary128 Floating-Point FMA Unit Design with SIMD Support[J]. IEEE Transactions on Computers, 2012, 61(5): 745-751

[9] 张峰，黎铁军，徐炜遐. 一种 128 位高精度浮点乘加部件的研究与实现[J]. 计算机工程与科学，2009，31(2)：93-103

[10] 雷元武，窦勇，郭松. 基于 FPGA 的高精度科学计算加速器研究[J]. 计算机学报，2012，35(1)：112-122

[11] Yu Xiao-yan, Chan Yiu-Hing, Curran B, et al. A 5GHz+ 128-bit Binary Floating-Point Adder for the POWER6 Processor[C] // Proceedings of the 32nd European Solid-State Circuits Conference. 2006: 166-169

[12] Intel Company. Intel Compilers and Libraries [EB/OL]. http://soft-ware. intel. com/en-us/articles /intel-cimpilers/, 2012, 12/24

[13] Fousse L, Hanrot G, Lefevre V, et al. Mpfr: A multiple-precision binary floating-point library with correct rounding [J]. ACM Transactions on Mathematical Software (TOMS), 2007, 33(2): 1-14

[14] Hida Y, Li X S, Bailey D H. Quad-double arithmetic: Algorithms, implementation, and application[R]. LBL-46996. Lawrence Berkeley National Laboratory, Berkeley, CA, 2000

[15] Firasta N, et al. Intel AVX: New Frontiers in Performance Improvements and Energy Efficiency[M]. White paper, 2008

[16] IBM Corporation. PowerPC Microprocessor Family: Vector/SIMD Multimedia Extension Technology Programming Environments Manual [M]. 2005

[17] Trong S D, Schmookler M, Schwarz E M, et al. POWER6 Binary Floating-Point Unit[C] // Proceedings of the 18th IEEE Symposium on Computer Arithmetic. Montpellier, France, 2007: 77-86

[18] Boersma M, Kroener M, Layer C, et al. The POWER7 Binary Floating-Point Unit[C] // Proceedings of IEEE Symposium on Computer Arithmetic. Tübingen, Germany, IEEE Computer Society, 2011

[19] Haring R A, Ohmacht M, Fox T W, et al. The IBM Blue Gene/Q Compute Chip [M]. IEEE Micro, March/April 2012: 48-60

[20] TOP500. TOP500 supercomputing sites [EB/OL]. http://www. top500. org/lists/2012/06, 2012

[21] Maruyama T, Yoshida T, Kan R, et al. SPARC64 VIIIfx: a New-Generation Octocore Processor for Petascale Computing[M]. IEEE Micro, March/April 2010: 30-40