

支持多种精度小数的运算单元设计

冯寅舫¹, 张盛兵¹, 黄嵩人², 樊晓桢¹

(1 西北工业大学 计算机学院, 陕西 西安 710072; 2 中国电子科技集团公司第五十八研究所, 江苏 无锡 214035)

摘 要: 针对 DSP 在应用中大量的使用不同精度的小数运算, 对数据运算单元进行详细设计. 通过对 Q15 格式小数计算特点的分析, 增加多精度累加型数据格式, 满足乘累加过程中对于精度的要求. 同时, 对执行部件的时延进行分析评估后, 合理的划分流水线, 在满足 DSP 计算功能的同时, 还能够达到较高的工作频率. 通过典型的数字信号处理算法核心, 对本设计性能进行量化分析. 在 TSMC 65 nm 的工艺条件下, 其时钟频率达到 500 MHz.

关键词: 数字信号处理器; 小数运算; Q 格式小数; 多精度累加型数据格式; 乘累加单元; 单指令多数据

中图分类号: TP332

文献标识码: A

文章编号: 1000-7180(2012)04-0150-04

A Design of Arithmetic Unit for Supporting Kinds of Decimal with Different Precisions

FENG Yin-chong¹, ZHANG Sheng-bing¹, HUANG Song-ren², FAN Xiao-ya¹

(1 Department of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China;

2 China Electronics Technology Group Corporation No. 58 Research Institute, Wuxi 214035, China)

Abstract: In view of the massive use of decimal arithmetic with different precisions in the digital signal processing application, a design of the arithmetic unit will be introduced. By analyzing the characteristics of the computation with Q15 format decimals, we add the multi-precision format to meet precision request in the multiply-accumulate process. At the same time, estimating the delay of executive components, we reasonably part the pipeline to implement the computation capability and reach high frequency. With the kernel of typical digital signal processing algorithms, we quantitatively analyze the performance of arithmetic unit. Under the TSMC 65nm technological condition, it runs at 500MHz.

Key words: DSP; decimal arithmetic; Q format decimal; multi-precision format; multiply-accumulate unit; SIMD

1 引言

数字信号处理算法中经常用到小数, 在计算中为了满足高精度的要求会使用多精度格式小数计算. 所谓多精度计算, 即参与运算的操作数或者计算结果的比特数大于计算机的字长^[1]. 一般的 DSP 处理器的小数格式中, 包含单精度小数 64 位多精度小数.

由于定点小数计算实现逻辑简单, 且运算速度快^[2], 因此在本设计内核中只支持定点格式的小数运算, 浮点运算在浮点协处理器中实现. 本文根据数

字信号处理算法中常用到的乘累加操作, 以及 Q 格式小数计算特点, 提出一种多精度累加型格式的小数. 同时根据此新增格式小数设计运算单元.

2 多精度机制

2.1 小数格式表示

在程序开发人员看来, 在数据表示中有一个暗含的小数点的二进制数, 即称为 Q 格式小数. 如对于一个 16 位有符号小数, 其表示的范围为 $[-1, 1)$, 称为 1q15 格式(Q15); 对于一个 16 位有符号小数, 其表示的范围为 $[-2, 2)$, 称为 2q14 格式

收稿日期: 2011-07-11; 修回日期: 2011-09-08

基金项目: “核高基”重大专项基金资助项目(2009ZX01034-001-002-003)

(Q14);对于一个 32 位有符号小数,其表示的范围为 $[-1, 1)$,称为 1q31 格式(Q31);其余情况以此类推。

本设计硬件支持 Q15 格式的定点小数。在指令中有一个标识位“ n ”,当 $n=1$ 时,参与运算的操作数为 Q15 格式的定点小数,否则操作数为整数。程序员可以通过设置这个位来指明操作数是整数还是 Q15 格式的定点小数。

2.2 多精度小数

本设计支持 4 种小数格式,分别是 Q15 小数、Q31 小数、64 位 Q47 多精度累加型小数,以及 Q63 多精度小数。其数据格式见图 1。

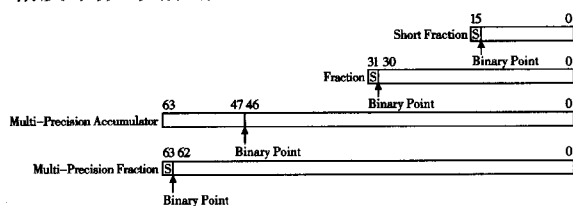


图1 小数数据格式

Q63 多精度小数用于 32×32 的 MAC 类指令,用于满足高精度计算的要求。

多精度累加型小数用于 16×16 、 16×32 的 MAC 类指令。由于在循环累加的过程中,多个 Q31 或者 Q47 小数格式的中间和相加之后得到的和可能会大于 1,一般的 DSP 中对这种情况以溢出来处理。而由于本设计增添了多精度累加型小数,它允许在累加过程中,累加和超出 Q31 格式小数能够表示的范围,依然能够在不损失精度的情况下,保持正确的结果,这在需要乘累加的运算中格外有用。程序员可以根据精度需要,在循环结束之后乘以一定的比例系数(可通过适当的移位指令实现)将其转换为其他精度格式的小数,如 Q31 或者 Q63 格式小数。

传统的 DSP 一般使用 40 位的累加数,如 ADI 的 ADSP-TS201 TigerSHARC Processor,这会使得程序员在编写标准 DSP 程序的时候,上下文切换变得十分困难。对此本设计通过增添了 64 位多精度累加型小数,来解决这个问题。这种格式的数据带来的好处有以下几点:

- 适合用 64 位扩展寄存器来存储运算结果数据。由于存在 32×32 的运算,因此中间结果都是 64 位,在最后的 4 个多路选择器分别选取中间结果的各个 16 位数据,因此无需冗余的硬件电路对其进行支持;

- 多精度累加型小数有效数字有 47 位,与 40

位累加数格式相比,具有更高的精度;

- 灵活度较高,可以在各种格式中进行转换。若有需要,程序员可自行用移位指令将其转换为其他精度格式的数据。

- 若程序对于精度要求不高,可以在软件编程层面,将这种格式的数据左移 16 位,将其当作具有 64 位精度的扩展 long 型数据用于下面的运算;

总之,多精度累加型小数可以简化程序员编程。

2.3 小数的舍入

本设计采用在累加阶段让其低 16 位也参与运算,只是在对最后 CLA(Carry Look-ahead Adder)的 32 位结果进行截断,这样在同样延时(1 级多路选择器的延时)的情况下,生成更精确的计算结果。为了实现此方法,对 16 位的累加数进行相应的左移操作,如左移 16 位。

对于乘法来说,两个 Q15 格式的定点小数相乘之后,会得到一个 2Q30 格式的结果。由于指令中的“ n ”位指明了操作数为 Q15 格式的小数,因此将乘法结果左移“ n ”位之后得到的为 1Q31 格式的小数,其溢出检测及饱和处理方法与一般的 32 位整数结果相同,简化了溢出检测及饱和逻辑。

两个 Q15 小数之间的乘法运算还存在一个问题。当两个 Q15 表示的最小的负数相乘时,正确的结果应该是 +1。

本设计的解决方法是:当检测到 $-1(1Q15) \times -1(1Q15)$ 这种特殊情况时,将乘法结果直接饱和到最大的正数值,即 $0x7FFF_FFFF$,以此来代替正确的结果——+1。

3 多精度 MAC 设计

3.1 流水线划分

本设计在支持多种精度数据格式的同时,为了提高运算效率,还支持 SIMD,最多可支持同时执行 4 个 16×16 的乘累加运算。MAC 类指令需要进行乘累加的运算,对于 32×16 的乘法器和 66 位加法器进行单独的时序分析,得到它们的关键路径延时分别是 0.98 ns 和 0.49 ns,其延时相对较长。由于本设计的频率要求为 500 MHz,即时钟周期为 2 ns,若将乘法器和 CLA 放在同一流水段中,将总共占用 1.47 ns,再加上其他的逻辑电路,使得乘累加单元成为整个处理器工作频率的瓶颈。因此本设计考虑将乘法和累加分为两个流水段,用两拍来完成乘累加操作,在执行 1 段(MUL 段)只进行两个操作数相乘,在执行 2(ACC 段)段将乘法的结果与第

三个累加数相加/减。

3.2 MUL 流水段

本设计是一款 SIMD 体系结构的 DSP, 其指令集同时支持: (a) 单个 32×32 , (b) 单个 16×16 , (c) 单个 16×32 , (d) 两个 16×16 , (e) 两个 32×32 , (f) 两个 16×32 , (g) 4 个 16×16 的乘累加运算, 因此在 MAC_MUL 模块中采用 4 个 16×32 的乘法器。

现今的乘法器基本工作原理大体分为三个步骤^[3-4]:

(1) 根据被乘数与乘数, 采用适合的算法, 进行编码及解码, 求得部分积;

(2) 将得到的部分积按权值错位相加, 进行部分积压缩;

(3) 对压缩得到的最终的两个值, 采用一个超前进位加法器模块, 得到最终的乘法结果。

尽管采用 64 位的超前进位加法器比一般行波加法器的延时要小, 但是其延时还是相对比较大, 而且 32×32 的乘法需要将两个 16×32 乘法器的结果经过相应的移位相加后才能得到其真正结果, 同时, 在执行 2 阶段还需要再加上一个累加数, 没有必要得到 16×32 的结果。因此本设计采用的乘法器只由前两个部分组成, 其输出的结果为压缩后得到的两个部分和。

由于两个 Q15 定点小数在进行乘法运算之后, 会有个冗余的符号位, 因此若相关乘法、乘累加/减指令中的“n”位指定为 1, 即操作数为 Q15 小数, 需要将乘法得到的两个部分和进行“n”移位, 即左移一位。

除此之外, 当出现在文中提及的 $(-1) \times (-1)$ 的特殊情况出现时, 直接选择 $0x0000_0000$ 和 $0x7FFF_FFFF$ 分别作为乘法器计算得到的两个部分和。

将经过处理的 8 个部分和用 4 级 CSA 进行压缩, 压缩后得到两个 128 位的乘法结果。由于 $ACC +/-(a_0 \times b_0 + a_1 \times b_1 + a_2 \times b_2 + a_3 \times b_3)$ 和 $ACC +/-(a_0 \times b_0 - a_1 \times b_1 + a_2 \times b_2 - a_3 \times b_3)$ 类指令 (在此 $a_0, a_1, a_2, a_3, b_0, b_1, b_2, b_3$ 都为 Q15 小数, ACC 为多精度累加型小数) 需要将 4 个 16×16 的乘积进行累加, 其累加结果可能会超过 1, 为了计算结果的精度不受损失, 对该类指令的执行结果数据格式采用多精度累加型小数格式, 因此将 8 个部分进行左移 16 位, 低 16 位取零。

在执行 $(a_0 \times b_0 - a_1 \times b_1) + (a_2 \times b_2 - a_3 \times b_3)$ 运算时, 对于 $a_1 \times b_1$ 和 $a_3 \times b_3$ 的部分和要进行取反

加 1, 而每个乘积得到的是两个部分和, 因此即需要在第 17 位上另行加 1, 但是此操作将带来额外的一级 CSA。之前乘积的部分和已经经过左移 16 位, 其低 16 位全为零, 因此这里考虑通过对低 16 位重新进行构造来消除这额外的一级 CSA。

$$\begin{aligned} &6'b00_1100 + 6'b00_1100 + 6'b00_1000 \\ &= 6'b10_0000 \end{aligned} \quad (1)$$

通过式 (1) 可以得知可以通过将原本 4 个低 16 位为 0 的部分和改成低 16 位分别为 $16'hE000, 16'hE000, 16'h8000, 16'h0000$ 这四个值时, 可以不用额外的一级 CSA 来完成减法补偿。

将经过 4 级 CSA 压缩后得到两个 128 位的部分和用流水线寄存器寄存, 送入后面的 ACC, 与第三个累加数进行累加操作。MUL 内部的逻辑见图 2, 图中的 a, b, c, d 四个数即为根据式 (1), 低 16 位经过修正的部分和。

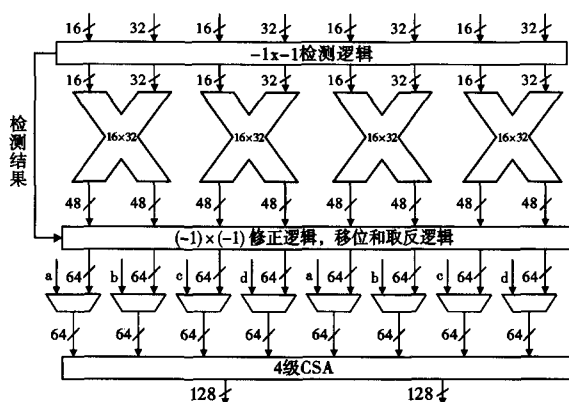


图 2 MUL 逻辑结构图

3.3 ACC 流水段

为了提高 DSP 的工作频率, 本设计将累加运算放到执行 2 阶段。ACC 模块将上一阶段产生的乘法结果与累加数相加, 得到乘累加结果, 并且将累加结果进行饱和与舍入处理, 并且进行溢出 (overflow) 和超前溢出 (advanced overflow) 检测, 从而生成 MAC 指令的最终执行结果。

对于一般位, 只有三个输入, 分别为两个部分和与一个累加数, 但是对于减法, 前面的逻辑只是对操作数进行了取反, 还需进行加 1 的补偿, 同时本设计还支持舍入。由于减法的补偿只是在末尾的加 1, 而舍入是在操作数中间加 1, 因此这里需要用到 2 级 CSA。其第四个操作数即为减法舍入补偿数, 各在其相应的位置取 1, 其余位为 0。

对于只执行 $(-1) \times (-1)$, 而不与累加数相加的纯乘法指令, 由于在 MUL 段已经对于乘法结果

进行过修正,其 Q31 小数结果为 32'h8000_0000,此时若在相应的位置加上舍入补偿位,则会使得结果变为 32'h8000_8000,使其出现偏差,因此需要对于这种情况特殊对待。

舍入位生成逻辑及减法舍入补偿数格式如图 3,其中 sub 位根据指令功能取值。

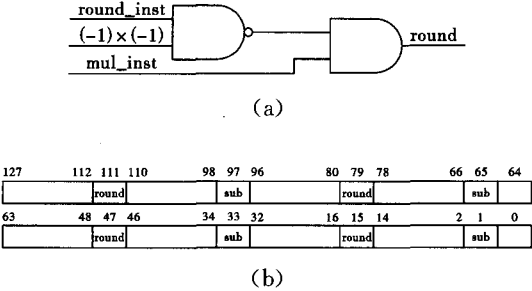


图 3 减法舍入补偿数

将移位后的累加数,乘法的中间结果,减法舍入补偿数共四个数进行 4-2 压缩后,压缩后的结果用最后的 CLA(超前进位加法器)进行相加,得到乘累加计算结果.对该结果进行溢出检测之后生成溢出标志,若指令对溢出结果需要进行饱和,则选取相应的饱和值作为最后的计算结果,否则将 CLA 的加法结果作为最后的计算结果。

本设计对于加法结果采用增加两位冗余符号位用于溢出检测.在进行溢出检测时,对于 Q31 小数,将加法结果的第 31 位与第 33 位作比较,若两位的数值不同,则结果发生溢出^[5];对于 Q63 小数,将加法结果的第 63 位与第 65 位作比较,若两位的数值不同,则结果发生溢出.由于 64 位 Q47 小数允许其数值大于 1,因此溢出的检测位不是第 47 位与第 49 位,而是与 Q63 小数的检测方法相同,比较结果第 63 位与第 65 位. SIMD 运算指令运算结果则可参考 Q31 小数或 Q63 小数,将 RES[95]与 RES[97],或者 RES[127]与 RES[129]作比较.比较操作可利用异或逻辑来实现。

ACC 的内部结构如图 4 所示。

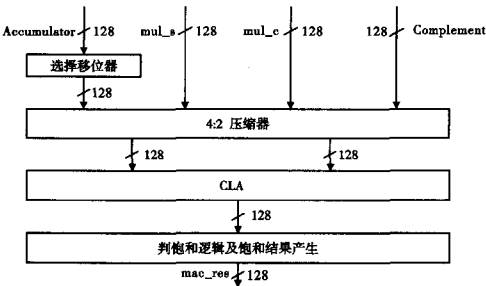


图 4 ACC 逻辑结构图

4 性能评估

将该运算单元已经用 VHDL 语言实现了整个设计.在 TSMC65 nm 工艺条件下,使用 Synopsis 的 Design Compiler 对包含该运算单元进行综合,综合结果如表 1 所示。

表 1 综合结果

延时/ns	面积/单元数	功耗/mW
2	3 453	2.041 6

用标准 DSP 的 Benchmark 对其进行性能进行测试,其计算内核程序的执行周期数见表 2。

表 2 Benchmark 测试结果

Algorithm	Cycle Count
N 维向量点积	N/4
N 维平方和	N/4
N 维矢量量化	3N/4
N 样本 T 阶实数 FIR 滤波	NT/4
N 样本 T 阶复数 FIR 滤波	NT
N 样本 T 阶实数 LMS 自适应 FIR 滤波	NT/2
N 样本 T 阶复数数 LMS 自适应 FIR 滤波	2NT
按频率抽选的 N 点基 2 实数蝶形运算	5N/4
按频率抽选的 N 点基 2 复数蝶形运算	9N/4

5 结束语

本文针对 DSP 芯片应用对于小数计算精度的要求,分析了 Q15 格式小数的计算特点,增加了新的格式的数据,来满足 Q15 小数对于计算精度的要求,从而使得该 DSP 更加适合于多媒体应用.通过综合,得到整个内核的工作频率能够到达 500 MHz,达到了设计的要求。

参考文献:

[1] 王小非,洪帆,汤学明,等. 实对称双线性函数与多精度整数的快速乘法[J]. 计算机科学, 2007, 34(6):92-97.

[2] 程佩青. 数字信号处理教程[M]. 北京:清华大学出版社,2007.

[3] 李伟,戴紫彬,陈韬. 基于跳跃式 Wallace 树的低功耗 32 位乘法器[J]. 计算机工程, 2008, 34(17):229-231.

3 半实物仿真实验与分析

通过所建立的半实物仿真系统,对上述滑翔弹模型进行了半实物仿真实验,以检验控制系统性能。图6是某条飞行弹道下,六自由度非线性数学仿真与半实物仿真的攻角变化曲线。由实验结果看出,半实物仿真与数学仿真结果吻合很好,说明仿真平台以及控制系统已经达到设计要求。

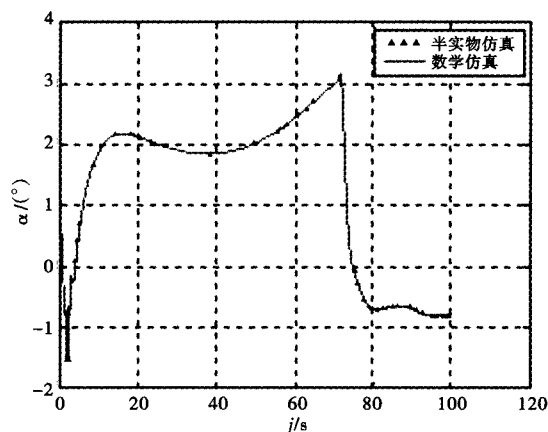


图6 攻角变化曲线

4 结束语

专业的半实物仿真实验室通常费用昂贵,协调困难。在基于 windows 操作系统的普通工控机上,通过引入外部时钟组建半实物仿真平台,不但满足了一般飞行仿真的需要,还可以充分发挥 Windows 操作系统成本低、稳定性高、界面友好、易操作以及第三方软硬件资源丰富等优点,试验周期及费用可以大为缩短。

该系统已在某导弹半实物仿真过程中得到了成

功的应用,对于缺少半实物仿真实验条件或是经费不足的研究队伍,本文的方法可以提供参考。

参考文献:

- [1] 梁思礼. 导弹控制系统研制工作的质量和可靠性管理(一)[J]. 微电子学与计算机, 1992(2): 1-4.
- [2] 梁思礼. 导弹控制系统研制工作的质量和可靠性管理(二)[J]. 微电子学与计算机, 1992(3): 1-7.
- [3] 徐立萍, 张健. Windows2000 平台下精确定时研究[J]. 微型电脑应用, 2005, 21(4): 46-48.
- [4] 陈坤中, 刘亚斌, 李军, 等. Windows NT 环境下的实时仿真技术研究及其应用[J]. 微计算机信息, 2006, 22(7-1): 239-241.
- [5] 陈春鹏, 姚益平, 蒋志文. 基于 Windows NT 开发实时半实物仿真的研究[J]. 计算机工程, 2002, 28(12): 98-100.
- [6] 韩志勇, 李先国. Windows2000 下高精度计时的研究[J]. 计算机工程与设计, 2005, 26(9): 2389-2394.
- [7] 王晓东, 董新民, 姚崇. 基于 xPC 的飞行控制系统半实物仿真设计[J]. 传感器与微系统, 2011, 30(2): 122-124.
- [8] PCIS-DASK User Manual, Revision: 2.02[R]. 2009.
- [9] PCIS-DASK Function Reference Manual, Revision: 2.06[R]. 2009.
- [10] 钱杏芳, 林瑞雄, 赵亚男. 导弹飞行力学[M]. 北京: 北京理工大学出版社, 2008.
- [11] 张友安, 胡云安. 导弹控制和制导的非线性设计方法[M]. 北京: 国防工业出版社, 2003.

作者简介:

杨黔龙 男, (1976—), 博士研究生, 副研究员。研究方向为飞行器设计、计算机应用、制导与控制。

(上接第 153 页)

- [4] 董时华, 乔庐峰. 改进部分积压缩结构的快速乘法器[J]. 计算机工程, 2010, 36(9): 229-231.
- [5] 郭阳, 丁伟, 李振涛. 高性能算术 SIMD 模块的电路设计[J]. 微电子学与计算机, 2010, 27(9): 29-32.

作者简介:

冯寅舫 男, (1986—), 硕士研究生。研究方向为微处理器设

计、VLSI 设计。

张盛兵 男, (1968—), 教授, 博士生导师。研究方向为 VLSI 设计、计算机体系结构。

黄嵩人 男, (1972—), 教授。研究方向为数字集成电路设计。

樊晓桢 男, (1962—), 教授, 博士生导师。研究方向为先进微处理器体系结构。