

基于脉动阵列的矩阵乘法器硬件加速技术研究

王 阳, 陶华敏, 肖山竹, 戴华东

(国防科学技术大学 ATR 国防科技重点实验室, 湖南 长沙 410073)

摘 要: 针对卡尔曼滤波算法中矩阵乘法运算的求解问题, 比较不同的硬件加速设计方案, 利用 9 个自行设计的处理单元, 设计了一种基于脉动阵列的并行结构浮点矩阵乘法器, 其峰值性能可达 761.96MFLOPS, 在资源一定的情形下提高了算法实现的实时性. 结合矩阵分块算法, 乘法器可对更高维的矩阵进行乘法求解, 具有良好的扩展性.

关键词: 脉动阵列; 矩阵乘法; 硬件加速; 卡尔曼滤波

中图分类号: TP332.2

文献标识码: A

文章编号: 1000-7180(2015)11-0120-05

Hardware Acceleration Technology of Matrix
Multiplier Based on Systolic Array

WANG Yang, TAO Hua-min, XIAO Shan-zhu, DAI Hua-dong

(National Key Laboratory of Automatic Target Recognition (ATR),
National University of Defense Technology, Changsha 410073, China)

Abstract: In KALMAN filtering algorithm, a mass of matrix multiply operation exist. Considering various hardware acceleration technology, based on systolic array, a kind of floating matrix multiplier, with parallel architecture, in order to guarantee the real-time performance, is designed. Nine self-designed PE are integrated in the multiplier, the peak performance of which can reach 761.96 MFLOPS. In the company of Block Matrix Multiplication, the multiplier could be extended well and play a role in higher dimension matrix multiply operation.

Key words: systolic array; matrix multiply; hardware acceleration; KALMAN filtering

1 引言

在现代控制、信号处理、精确制导领域中, 卡尔曼滤波模块具有广泛的应用. 假设期望估计的随机过程由式(1)、式(2)给出^[1].

$$X_{k+1} = \Phi_k X_k + W_k \quad (1)$$

$$Z_k = H_k X_k + V_k \quad (2)$$

卡尔曼滤波算法的迭代过程如式(3)~(6)所示.

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \quad (3)$$

$$\hat{X}_k = \hat{X}_k^- + K_k (Z_k - H_k \hat{X}_k^-) \quad (4)$$

$$\hat{X}_{k+1}^- = \Phi_k \hat{X}_k \quad (5)$$

$$P_{k+1}^- = \Phi_k P_k \Phi_k^T + Q_k \quad (6)$$

式中, Φ_k 为 $n \times n$ 转移矩阵, W_k 为 n 维白噪声列向

量. Z_k, V_k 为 m 维滤波测量值与白噪声列向量, H_k 为 $m \times n$ 维测量矩阵. $\hat{X}_k^- P_k$ 表示滤波初始的估计值与协方差值, 均为 n 维列向量.

滤波迭代算法中矩阵的加减、相乘与求逆是最为基本的运算操作. 其中加减运算的计算复杂度为 $O(n)$, 乘法与求逆运算的计算复杂度为 $O(n^3)$ ^[2]. 在对矩阵求逆进行计算策略的选择时, 如式(1)~(7)所示, 通常会先对矩阵进行分解操作(如 LU 分解, 其中 L 与 U 为三角矩阵), 然后对分解矩阵进行求逆与相乘从而得到原矩阵的求逆结果.

$$A = L * U, A^{-1} = U^{-1} * L^{-1} \quad (7)$$

因此若能提高矩阵乘法的运算效率, 将对整个滤波算法的实时性提高具有重要的意义. 在高性能计算的研究领域中, 学者们综合考虑时间、空间、资源的复杂度^[3], 从“处理单元(PE)”与“计算结构”设

计两个方面对系统进行计算的加速. 目前主要有基于 MCP(多核处理器)、GPU(图形处理器)及 FPGA(现场可编程逻辑门阵列)的三种主流计算加速技术^[4].

它们三者之中,FPGA 的开发略为复杂,但配置灵活,功耗较低,计算结构并行化好,是国内外学术界研究的热点之一. 如在文献^[4]中,作者对矩阵乘法运算的浮点加速运算架构进行了深入的讨论. 文献^[5-7]着重讨论了矩阵乘法中 PE 单元的设计. 文献^[8]中设置消谐模型的实际工程应用背景. 文献^[3]中对不同计算结构的资源、空间、时间复杂度进行了分析比较.

本文以卡尔曼滤波作为应用背景,基于 FPGA 平台以及脉动阵列的运算架构,对双精度浮点矩阵乘法运算器进行了硬件加速的设计.

2 矩阵乘法器的硬件加速实现

2.1 处理单元设计

矩阵的乘法运算中,考虑 $C=AB+C$ 的修正公式($A \in R^{m \times p}$, $B \in R^{p \times n}$, $C \in R^{m \times n}$),矩阵 C 中的每一个元素均为 A 中对应行向量与 B 中对应列向量的点积,点积是一系列元素相乘运算结果的累加. 矩阵乘法运算的基本处理单元(PE)本质上是一个乘法累加器(MAC). 在制导控制系统的实际工程应用中,输入数据通常采用双精度浮点格式,因此在 MAC 的设计中,一个不可忽视的方面便是浮点运算的实现. 相比于 DSP 与 GPU 而言,浮点运算一直以来都是 FPGA 的弱势所在. 直到随着集成电路设计工艺的发展,使得 FPGA 中拥有了内嵌的 DSP 处理核,从而大大提升了 FPGA 在浮点运算方面的能力^[4]. 正如文献^[4]中所分析的那样,在越来越多的实际工程应用中,调用生产厂家所提供的浮点运算 IP 核来进行运算已经成为了设计主流思路. 基于这样的思路,本文考虑卡尔曼滤波的实时应用背景以及硬件资源等方面的限制,基于 XILINX - XCV5VLX110T 型 FPGA 进行了 PE 的设计.

本文设计的 PE 结构如图 1 所示,由浮点乘法器与浮点加法器组成.

我们从数据信号、控制信号与单元结构的角度简单分析图 1 所示的处理单元. 其中数据信号包括有 Input_a, Input_b 以及 Com_result. 控制信号包括 Com_start, Clear, 以及 Com_ready 信号. 运算中,输入数据在完成乘法运算后,其输出的结果作为加法器的输入 Adder_in_a.

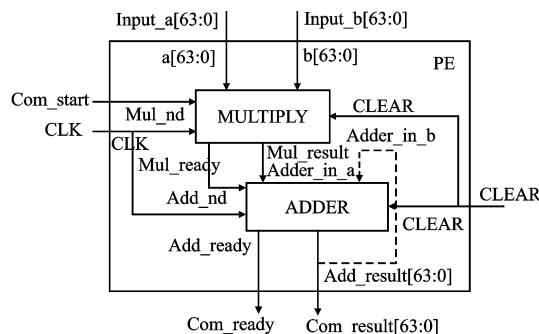


图 1 PE 单元结构图

为了实现累加的功能,将加法器的输出结果引回输入端 Adder_in_b. 当本轮计算结束时,将结果取出. 各个端口以及信号的功能总结如表 1 所示.

表 1 PE 单元信号功能表

信号名称	功能描述
Input_a, Input_b	输入数据
Mul_result	乘法器输出
Com_start/Mul_nd	计算开始
Mul_ready/Adder_nd	乘法结果有效标志位
Com_ready/Add_ready	计算完成标志位
Com_result/Add_result	计算结果

其中,Com_start/Mul_nd 信号,表示此时加载在输入端的数据是所期望的数据,信号本身并不对 PE 单元进行使能控制. 在对 PE 模块进行仿真时,采用人为方式进行控制.

Mul_ready/Adder_nd 信号则是 PE 单元的反馈信息,当计算完成时,由 PE 单元置为 1. 在设计中将二者作为握手信号,使乘法器输出结果有效时,输出结果能及时加载至加法器的输入端.

2.2 PE 的资源使用分析

2.1 节中使用了 XILINX 提供的两个浮点运算 IP 核对 PE 单元进行了设计,本节对 PE 单元的流水设计与资源消耗进行简要分析.

一个基于 FPGA 的设计在工作时所能达到的最高工作频率取决于设计中的“关键路径”. 设计者能通过设置流水来缩短这一路径,从而使系统具有更高的工作频率以及更大的数据吞吐量. 在 PE 单元的设计中,调用 XILINX 的 IP 核时便可设置不同的流水级数. 文献^[5]中的研究说明:当流水级数达到一定数值后,对系统总体性能的提升效果已非常有限,相比于额外耗费的资源,往往得不偿失. 在仿真分析中,笔者发现乘法器的流水级对系统主频并没有明显的影响,说明系统的“关键路径”无法通过

设置乘法器的流水级数而缩短. 本设计中, 设置乘法器流水级数为 3 级, 加法器流水级数为 1 级. 这是由于在实际的计算中, 加法器的输出需要即刻反馈至输入端实现累加的功能, 若对加法器进行流水设置, 则需要额外消耗大量逻辑资源来进行“加法器树”的设计.

XILINX-XCV5VLX110T 型 FPGA 中共有内置 DSP48 核 64 个. 若在乘法器中选择使用 DSP48 核, 每个乘法器至少需要使用 9 个, 则在后期构建处理器阵列时会出现资源不足的情形. 因此在设计中, 出于资源的限制, 乘法器 IP 核未使用 DSP48 核, 加法器的 IP 核使用了 3 个 DSP48 核. PE 单元的资源消耗如表 2 所示.

表 2 PE 单元资源消耗表

Logic Utilization	Used	Available	Utilization/%
Slice Registers	831	69 120	1
Slice LUTs	4 054	69 120	5
BUFG	2	32	6
DSP48Es	3	64	4

PE 单元的功能如图 2 所示. 当我们将数据加载在输入端时, 将 Com_start 置高, 数据在时钟上升沿读入. 经过三个时钟周期的延迟得到计算结果, 此时 Com_ready 信号由 PE 单元置高, 表示计算结束.

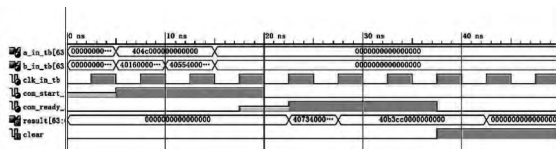


图 2 PE 单元功能图

3 矩阵乘法器性能分析

3.1 基于脉动阵列的矩阵乘法器

两个 N 维矩阵做乘法运算的计算复杂度为 $O(n^3)$. 在设计具体的计算结构时, 常见的计算结构有以下三种^[9].

图 3 所示的结构为“连续结构”, 使用一个 PE 单元完成设计. 这一设计具有最小的资源消耗, 但是其时间复杂度达到 $O(n^3)$, 并不适用于实时性要求较高的应用.

图 4 所示的是“向量并行计算”结构, 它由 N 个乘法器与 $N-1$ 个加法器组成, 能并行地计算两个向量的点积, 其时间复杂度为 $O(n^2)$. 这种设计最大的不足在于其“最短路径”会随着所处理矩阵的维数而

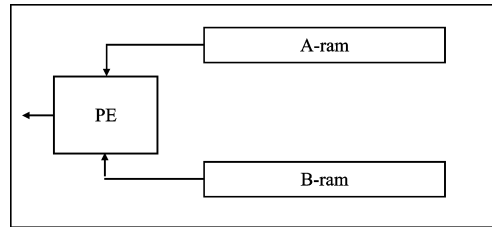


图 3 连续结构矩阵乘法器结构图

增长, 从而会影响工作的主频.

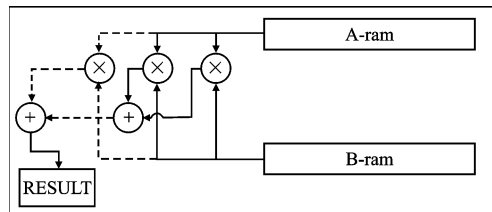


图 4 向量并行结构乘法器结构图

图 5 所示为脉动阵列 (systolic array). 阵列中 PE 排列成网状的拓扑结构, 相邻的 PE 之间有连接. 在进行计算时, 数据从阵列的边界并行输入, 在阵列内部相邻的 PE 之间流动传输, 从而实现了数据的复用, 阵列中的 PE 单元可以并行地进行“点积计算”.

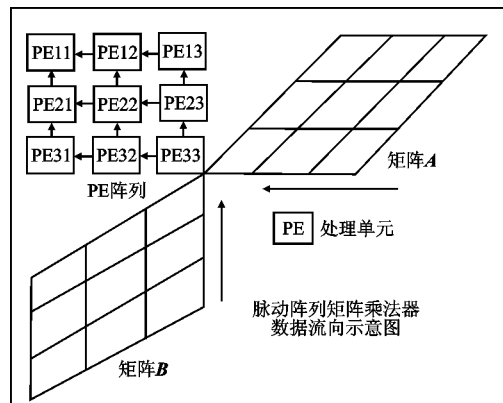


图 5 脉动阵列矩阵乘法器

若是计算 N 阶矩阵 A 与 B 的乘法操作, 数据并行地从右侧与下侧进行输入, 并在对应的方向上进行流动. 每个 PE 所需的计算时间为 N 个周期, 阵列总的计算时间复杂度为 $O(n)$, 构建对应的阵列需使用 n^2 个 PE, 资源复杂度为 $O(n^2)$. 在计算时, 数据将如图 5 所示的那样逐步填满阵列单元, 然后再慢慢消去, 这也是“脉动”一词的由来. 相比于“连续结构”以及“向量并行结构”而言, 脉动阵列在时间复杂度、资源复杂度、工作频率、以及工作性能上都较为出色. 因此本文使用 2.1 节

中所设计的 PE 单元,在 XILINX-VERTEX5-XCV5VLX110T-FPGA 平台上设计了 3×3 的脉动阵列乘法器模块,乘法器结构以及数据流向如图 5 所示。

3.2 脉动阵列乘法器性能分析

本节基于脉动阵列结构,在 ISE14.5 的开发环境中,以两个 3×3 矩阵相乘为例对矩阵乘法器进行仿真分析。表 3 中说明了乘法器的资源使用情形。图

6 为乘法器的功能图。

表 3 乘法器资源使用情形

Logic Utilization	Used	Available	Utilization/%
Slice Registers	8 630	69 120	12
Slice LUTs	36 652	69 120	53
BUFG	1	32	3
DSP48Es	27	64	42

$$Ma = \begin{bmatrix} 5.5 & 85 & 9.8 \\ 50 & 30 & 10 \\ 35 & 10 & 60 \end{bmatrix} = \begin{bmatrix} 4016000000000000 & 4055400000000000 & 4023999999999999a \\ 4049000000000000 & 403e000000000000 & 4039000000000000 \\ 4041800000000000 & 4024000000000000 & 404e000000000000 \end{bmatrix},$$

$$Mb = \begin{bmatrix} 5.5 & 50 & 35 \\ 85 & 30 & 10 \\ 9.8 & 10 & 60 \end{bmatrix} = \begin{bmatrix} 4016000000000000 & 4049000000000000 & 4041800000000000 \\ 4055400000000000 & 403e000000000000 & 4024000000000000 \\ 4023999999999999a & 4039000000000000 & 404e000000000000 \end{bmatrix}.$$

使用 MATLAB 计算如下,与图 6 中乘法器计算的结果相符合:

$$Mres = Ma * Mb = \begin{bmatrix} 7.35129 & 3.07000 & 1.63050 \\ 3.07000 & 4.02500 & 3.55000 \\ 1.63050 & 3.55000 & 4.92500 \end{bmatrix} = \begin{bmatrix} 40bcb74a3d70a3d7 & 40a7fc0000000000 & 40997a0000000000 \\ 40a7fc0000000000 & 40af720000000000 & 40abbc0000000000 \\ 40997a0000000000 & 40abbc0000000000 & 40b33d0000000000 \end{bmatrix}.$$

图 6 表明,矩阵乘法器完成计算一共花费了 10 个时钟周期。流水的引入增加了计算的延迟时间,但提高了乘法器的数据吞吐量。理想情况下,一个 PE 单元在一个时钟周期内可以完成两次浮点运算,ISE 综合报告中显示系统的工作频率能达到 42.331 MHz,根据计算性能衡量式(1)~(8),其中 P 表示 PE 数目, f 表示工作频率^[5]。

$$PERF_{peak} = 2 \times P \times f \quad (8)$$

可以得到: $PERF = 761.96$ MFLOPS。完成一次乘法所需的时间为 23.6 ns,完成一次的卡尔曼滤波迭代算法中所有乘法操作运算的时间为 63.7 μ s,能较好满足系统的实时性需求。

3.3 矩阵乘法器的扩展应用

就硬件加速的效果而言,进行 3×3 矩阵的乘法相比于另外的计算平台而言,并不具有太大的优势。在相当多的应用中,只有当处理的数据量较大时才能实现优异的加速比。直接根据矩阵维数构建大型阵列乘法器阵列对 FPGA 的资源配置要求较高,如文献[4]中构建了 12×12 的矩阵乘法阵列,运算性能峰值达到了 290GFLOPS,其 DSP48 核的使用数量也达到了惊人的 2 016 个。

因此在实际的高维矩阵乘法中,乘法器应结合矩阵分块相乘的方式,对更高维的矩阵进行乘法运算的求解。

引用文献[2]中结论:

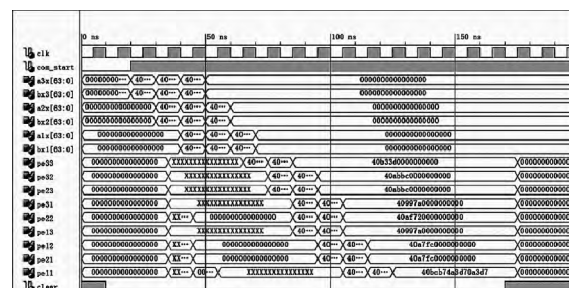


图 6 乘法器功能图

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1s} \\ \vdots & \ddots & \vdots \\ A_{q1} & \cdots & A_{qs} \end{bmatrix}_{m_1}^{m_q},$$

$$B = \begin{bmatrix} B_{11} & \cdots & B_{1s} \\ \vdots & \ddots & \vdots \\ B_{q1} & \cdots & B_{qs} \end{bmatrix}_{p_1}^{p_s},$$

$$C = AB$$

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1r} \\ \vdots & \ddots & \vdots \\ C_{q1} & \cdots & C_{qr} \end{bmatrix}_{m_1}^{m_q},$$

$$n_1 \quad \cdots \quad n_r$$

$$C_{\alpha\beta} = \sum_{r=1}^s A_{\alpha r} B_{r\beta}, \alpha = 1:q, \beta = 1:r.$$

若是对于 9×9 矩阵的计算,需调用乘法器模块 27 次,花费 270 个时钟周期,在 42.331 MHz 的频率下,可在 6.4 μ s 内完成计算。而在 MATLAB 中

计算两个 9×9 的随机矩阵需花费 $30 \mu\text{s}$ 的时间, 加速比将近 5. 随着矩阵维数的增加能更好地发挥脉动阵列的处理优势, 将得到更好的加速比.

4 结束语

在卡尔曼滤波的实时计算中存在着大量的矩阵乘法运算. 本文在资源消耗相对节约的情形下, 基于“赛灵思”XCV5VLX110T-FPGA 及脉动阵列的设计思想, 构建了 3×3 的矩阵阵列乘法器, 为设计其余运算模预留了相当的资源. 同时乘法器具有良好的扩展性, 可结合矩阵的分块运算, 实现维数更高的矩阵乘法. 在后续的研究中, 可从运算效率、工作频率等方面入手进一步提升乘法器的性能.

参考文献:

- [1] Brown R G. Introduction to random signal analysis and Kalman filtering[M]. New York: Wiley, 1983.
- [2] Golub G H, Van Loan C F. 矩阵计算[M]. 袁亚湘, 译. 北京: 科学出版社, 2011.
- [3] Cappello J D, Strenski D. A practical measure of FPGA floating point acceleration for High Performance Computing[C]// Proceedings of the 2003 IEEE 24th International Conference on Application-specific Systems. Washington, DC: IEEE, 2013: 160-167.
- [4] 田翔, 周凡, 陈耀武, 等. 基于 FPGA 的实时双精度浮点矩阵乘法器设计[J]. 浙江大学学报: 工学版, 2008, 42(9): 1611-1615.
- [5] 刘沛华, 鲁华祥, 龚国良, 等. 基于 FPGA 的全流水双精度浮点矩阵乘法器设计[J]. 智能系统学报, 2012, 7(4): 302-306.
- [6] Bensaali F, Amira A, Sotudeh R. Floating-point matrix product on FPGA[C]// proceedings of the IEEE/ACS International Conference on Computer Systems and Applications, 2007 Computer Systems and Application. Amman: IEEE, 2007: 466-473.
- [7] Kestur S, Davis J D, Chung E S. Towards a universal FPGA matrix-vector multiplication architecture[C]// proceeding of the 2012 IEEE 20th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). Toronto, ON: IEEE, 2012: 9-16.
- [8] 吴淑泉, 王前, 谢运祥. 适于消谐模型求解的矩阵乘法器设计与实现[J]. 华南理工大学学报: 自然科学版, 2003, 31(8): 1-5.
- [9] Fleming S T, Thomas D B. Hardware acceleration of matrix multiplication over small prime finite fields[C]// proceedings of 9th international Conference on Reconofigurable Computing: architectures tools, and applications. Berlin, Heiolelberg: ACM, 2013: 103-114.

作者简介:

王 阳 男, (1991-), 硕士研究生. 研究方向为精确制导自动目标识别. E-mail: sunville2013@nudt.edu.cn
陶华敏 女, (1972-), 副教授, 硕士研究生导师. 研究方向为精确制导自动目标识别.

(上接第 119 页)

- [8] Chengwen Zhang, Xiuqin Lin. A genetic algorithm with improved convergence capability for QoS-Aware web service selection[C]// Porceedings of the international Conference on Management and Service Science. wuhan: IEEE, 2009, : 1-4.
- [9] 李学伟, 吴金培, 李雪堰. 实用元胞自动机导论[M]. 北京: 北京交通大学出版社, 2013.
- [10] 陈彦萍, 张建科, 孙家泽, 等. 一种基于混合智能优化的

服务选择模型[J]. 计算机学报, 2010, 33(11): 2116-2125.

作者简介:

陈彦萍 女, (1979-), 博士, 教授. 研究方向为服务计算.
樊梦茹(通讯作者) 女, (1989-), 硕士研究生. 研究方向为服务计算. E-mail: fanmengru_1017@163.com