基于SIMD处理器的全定制多粒度矩阵寄存器文件*

张 凯,陈书明,王耀华,陈海燕,李振涛
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:在SIMD处理器上映射矩阵运算时会带来大量的数据重排操作从而降低系统性能。本文提出定制化的多粒度矩阵寄存器文件(MMRF)以消除数据重排操作。MMRF支持多粒度的并行行访问和列访问,从而提升矩阵运算的性能。MMRF可以被动态配置为不同的并行访问模式,在不同模式下一个或多个子矩阵可以被并行处理。实验结果显示,同传统的向量寄存器文件(VRF)和矩阵寄存器文件(MRF)相比,MMRF可分别带来2.21倍和1.6倍的平均性能提升,面积分别增加14.3%和3.7%,功耗分别增加14.6%和2.2%。同TMS320C64x+处理器相比,基于SIMD技术的FT-Matrix处理器在引入MMRF后可以得到5.65倍到7.71倍的性能提升。通过层次化的全定制设计技术,MMRF的面积和关键路径分别减少17.9%和39.1%。

关键词:SIMD;矩阵运算;多粒度;矩阵寄存器文件
中图分类号:TP316 **文献标志码**:A **文章编号**:1001-2486(2013)04-0156-05

A customized multi-grain matrix register file for SIMD processors

ZHANG Kai, *CHEN Shuming*, *WANG Yaohua*, *CHEN Haiyan*, *LI Zhentao*
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Mapping matrix operations on SIMD processors brings a large amount of data rearrangement that lowers the system performance. In this study, a customized Multi-Grain Matrix Register File (MMRF), which supports multi-grained parallel row-wise and column-wise access, was proposed to eliminate these data rearrangement and increase the performance of matrix operations. The MMRF could be configured into different parallel access modes, in which one or several sub-matrices can be accessed in parallel. Experimental results show that, compared with the traditional Vector Register File (VRF) and the MRF, the MMRF can respectively achieve about 2.21x and 1.6x average performance improvement, where the area of MMRF increases by 14.3% and 3.7% respectively, and the power of MMRF increases by 14.6% and 2.2% respectively. Compared with TMS320C64x+, the SIMD processor of FT-Matrix can achieve about 5.65x to 7.71x performance improvement by employing the MMRF. By hierarchical customized design technology, the area and critical-path delay of MMRF can be reduced by 17.9% and 39.1% respectively.

Key words: SIMD; matrix operation; multi-grain; matrix register file

无线通信和媒体处理应用中有大量不同规模的矩阵运算[1-2],例如WiMAX,LTE和H.264。这些矩阵运算占据了应用大部分的执行时间,加速这些不同规模的矩阵运算可以有效地提升应用的性能,例如矩阵乘法和矩阵转置等。

近年来,单指令流多数据流(SIMD)技术被广泛应用于高性能处理器中加速并行计算。典型的SIMD处理器包含多个并行处理单元,通过一条指令并行操作多个数据来开发数据级并行[3]。SIMD处理器可以在较低的功耗下提供很高的性能,因此,非常适合上述的计算密集型应用。在传统的SIMD结构中,当处理单元不能够被高效利用或者需要在处理单元间进行大量的数据混洗操作时,性能会大大损失。

矩阵运算是典型的计算密集型应用,充分开发矩阵运算中的并行性,是提升矩阵运算性能的关键。矩阵运算被映射到SIMD处理器时会带来大量的数据重排操作,从而降低系统性能。这些数据重排操作产生于矩阵乘法和转置运算对矩阵的列访问。以矩阵乘法为例,当在SIMD处理器上实现矩阵乘法时,多个处理单元并行处理一个矩阵乘法,本质是多个处理单元对两个向量数据的点积操作,对其中一个向量的访问是列访问。然而,矩阵数据一般被组织为按行存储。这就需要数据重排操作将按行存储的数据排列为按列访问的顺序。

* **收稿日期**:2012-12-01
基金项目:国家自然科学基金资助项目(60906014,61070036);高性能计算联合博导组科研基金项目
作者简介:张凯(1985—),男,陕西西安人,博士研究生,E-mail:zhknudt@gmail.com;
陈书明(通信作者),男,教授,博士,博士生导师,E-mail:smchen@nudt.edu.cn

为了解决这个问题,Corbal 提出了新颖的面向矩阵运算的结构 MOM[4],但 MOM 需要非常大的寄存器文件支持,而且该结构仅仅对矩阵加法高效。Shahbahrami 提出了同时支持一般的行访问和列访问的 MRF[5],但当矩阵的大小和 MRF 的大小不一致时,MRF 就不够灵活和高效。Ciobanu 提出了多态寄存器文件 PRF[6],PRF 能够给程序员提供强大灵活的手段来高效组织寄存器文件,程序员可以根据矩阵大小来定义 PRF 的大小。但是,考虑到指令编码空间的有限性,PRF 在硬件上很难实现。

1　MMRF 的结构

1.1　一般的 SIMD 结构

本文给出了如图 1 所示的一般 SIMD 结构,该结构被普遍应用于很多高性能体系结构,如 AnySP[3],SODA[7],Cell[8],VT[9] 等。SIMD 处理器一般由多个并行处理单元(VPE)组成。向量寄存器文件(VRF)给 VPE 提供了多个访问端口,支持对数据的向量读写。混洗网络(shuffle network)用于不同 PE 间的数据重排。每个 PE 内部包含多个功能单元,包括 ALU,MAC,Load/Store(L/S)等。L/S 部件用于在 VRF 和向量存储器(VM)之间传递数据。向量存储器一般由 N 个存储 bank 组成。

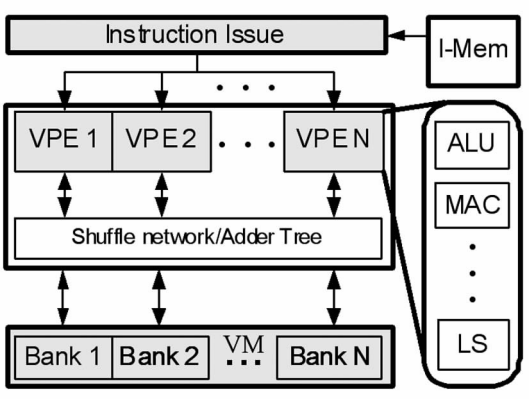

图 1　SIMD 处理器的一般结构

Fig. 1　A simplified architecture of SIMD processor

本文提出的 MMRF 可以替代 SIMD 处理器中 VRF 的位置,也可以作为 VRF 部件的增强部分,处于和 VRF 同样的数据通路和控制通路中。MMRF 可以用于在 SIMD 处理器中消除矩阵运算时的数据重排操作,以提升性能。

1.2　MMRF 的结构

如图 2 所示,MMRF 的微体系结构由一个寄存器阵列(RA)、多个读写端口(R/WPs)以及地址译码逻辑。多读写端口支持多功能单元对 MMRF 的并行访问。如图 2 所示,RA 由 N × N 个 16 或 32 位的基本寄存器单元组成。作为 SIMD 处理器的寄存器文件,MMRF 包含 N 个行向量寄存器(VR)VR_0 ~ VR_{N-1} 和 N 个列向量寄存器(CVR)CVR_0 ~ CVR_{N-1}。因此,MMRF 物理上使用了 N^2 个寄存器单元,但在逻辑上提供了 2N 个向量寄存器,是传统 VRF 能够提供的逻辑寄存器数的 2 倍。

多个读写端口支持对 MMRF 的并行访问,从而增加功能单元间的并行执行能力。每一次对 MMRF 的读写操作被分解为一系列信号,包括读写请求类型信号以区分读操作和写操作,读写索引信号用以寻址所访问的寄存器单元。地址译码逻辑根据 BMR 的内容从 RA 中选择一些寄存器单元来完成读写操作。

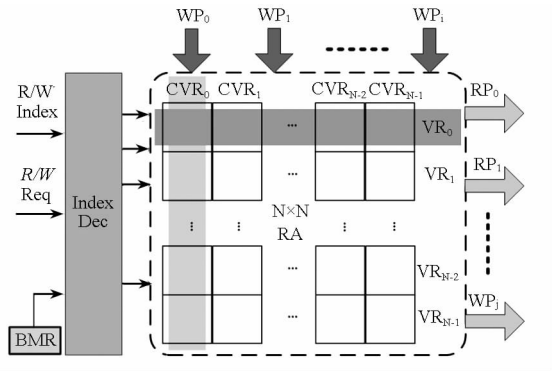

图 2　MMRF 的结构图

Fig. 2　The micro-architecture of MMRF

BMR 作为记录 MMRF 并行访问模式的特殊寄存器,可以被动态配置。MMRF 的地址索引视图随 BMR 内容的变化而变化。使得访问 MMRF 的并行粒度也随之变化,从而使 MMRF 支持了多粒度的并行访问。多粒度并行访问的细节将在下节详细阐述。

我们可以将 MMRF 应用到一个已有的 SIMD 处理器中,并且无需修改该处理器的指令集体系结构。通常,指令码中有 m 位用以编码源或者目的操作寄存器。该 m 位的最高位可用于区分 VR 和 CVR。其余位用以索引具体是哪一个 VR 或 CVR。通常情况下,处理器都含有多个可以配置的特殊功能寄存器,BMR 的编码可以借用处理器的特殊功能寄存器的保留位。

2　多粒度并行访问

为了更好地阐述 MMRF 的概念,我们在一个包含 16 个 VPE(VPE0 ~ VPE15)的 SIMD 处理器上实现了一个 16 × 16 大小的 MMRF。我们研究了无

线通信和媒体处理领域几种典型应用中的矩阵运算规模,发现如下三种并行操作模式比较高效。

1)单路访问模式:程序员在该模式的视图如图2所示。在该模式中,MMRF可被看为传统的MRF。

2)两路访问模式:MMRF被划分为4个大小相同的块,每一块的大小为一个8×8的MRF。

3)四路访问模式:MMRF被划分为16个大小相同的块,每一块的大小为一个4×4的MRF。

2.1 单路访问模式

在单路路访问模式中,MMRF的功能与传统的MRF相同。每一个VR或者CVR,如图2所示,包含MMRF的一个行向量寄存器单元组或一个列向量寄存器单元组,该单元组中的16个寄存器单元可以并行地被16个VPE访问。通常,单路访问模式主要用于16×16的矩阵运算。在MMRF的帮助下,程序员可以方便地得到矩阵的一行数据或一列数据。

2.2 两路访问模式

两路访问模式主要用于8×8的矩阵运算。该模式下,程序员可见的寄存器仍是VR_0~VR_{15}和CVR_0~CVR_{15}。如图3所示,该模式下,MMRF被分为4块。每个VR或CVR由2部分组成,每一部分由8个寄存器单元组成,该8个寄存器单元的内容组成一个向量数据,并被8个VPE对应访问该向量数据的一个元素。2个向量数据分别来源于2个不同的子矩阵。

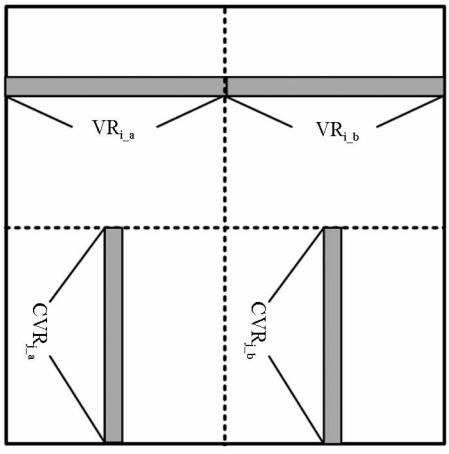

图3 MMRF的两路访问模式

Fig. 3 The two ways access mode of MMRF

行向量寄存器VR_i由VR_{i_a}和VR_{i_b}组成。列向量寄存器CVR_j由CVR_{j_a}和CVR_{j_b}组成。VR_{i_a}或CVR_{j_a}被VPE0~VPE7访问。VR_{i_b}或者CVR_{j_b}被VPE8~VPE15访问。这样,在两路访问模式中,可以很容易地对两对8×8的矩阵乘法并行处理。

在两路访问模式中,VR_i所访问的物理寄存器单元和单路访问模式中相同,但VR_i的内容在逻辑上却属于2个不同的子矩阵。因此,在该模式中,对MMRF的一次访问相当于对2个矩阵数据的并行访问,而单路模式中一次访问只能得到一个矩阵的一行或一列数据。

2.3 四路访问模式

四路访问模式主要用于4×4的矩阵运算。如图4所示,该模式下,程序员可见的寄存器与前两个模式数量相同。MMRF被分为16块。每个VR或CVR由4部分组成,每一部分由4个寄存器单元组成,该4个寄存器单元的内容组成一个向量数据,并被4个VPE对应访问该向量数据的一个元素。4个向量数据分别来源于4个不同的子矩阵。

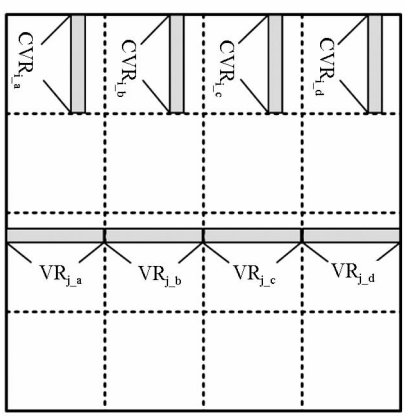

图4 MMRF的四路访问模式

Fig. 4 The four ways access mode of MMRF

行向量寄存器VR_i由VR_{i_a}、VR_{i_b}、VR_{i_c}和VR_{i_d}组成。列向量寄存器CVR_j由CVR_{j_a}、CVR_{j_b}、CVR_{j_c}和CVR_{j_d}组成。VR_{i_a}或CVR_{j_a}被VPE0~VPE4访问。VR_{i_b}或CVR_{j_b}被VPE4~VPE7访问。VR_{i_c}或CVR_{j_c}被VPE8~VPE11访问。VR_{i_d}或CVR_{j_d}被VPE12~VPE15访问。这样,在四路访问模式中,可以很容易地对4对4×4的矩阵乘法并行处理。

在四路访问模式中,VR_i所访问的物理寄存器单元和单路访问模式中相同,但VR_i的内容在逻辑上却属于4个不同的子矩阵。因此,在该模式中,对MMRF的一次访问相当于对4个矩阵数据的并行访问。

图5给出了部分4×4矩阵乘法在MMRF帮助下的计算过程。当MMRF被配置为四路访问模式后,8个4×4大小的矩阵可以放置于MMRF中,每4个PE处理一对矩阵乘法,该4个PE可以访问被处理矩阵的一行或一列数据。16个PE

可以并行访问 4 对被处理矩阵的一行或一列数据,从而实现对 4 对矩阵乘法的并行处理。

在传统的 SIMD 处理器上,要实现对 4 个矩阵同时进行列访问时,需要将 4 个矩阵的所有数据读入寄存器文件,再进行大量的数据重排操作。在 MMRF 的帮助下,所有的数据重排操作被消除,快速实现对 4 个矩阵同时进行列访问以提升性能。

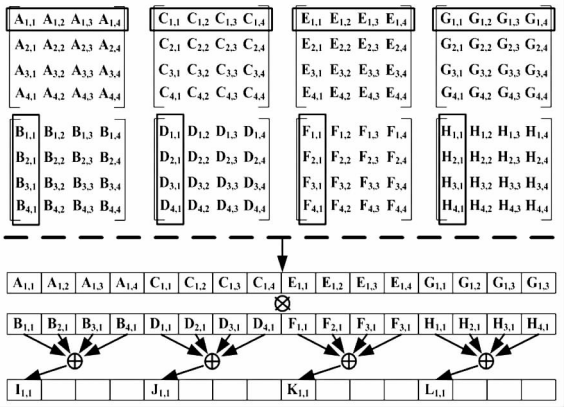

图 5　部分 4 ×4 矩阵乘法示意图

Fig. 5　Part of matrix multiplications with MMRF

3　性能评估

在 FT-Matrix-Sim 模拟器中实现了 MMRF,FT-Matrix-Sim 是 FT-Matrix 的周期精确模拟器,FT-Matrix 是国防科技大学自主研发的面向媒体处理的高性能处理器,它的 SIMD 单元包含 16 个 VPE,VPE 内部基于 VLIW 结构。FT-Matrix 采用了 5 路 VLIW,两个发射槽用于 L/S 操作,一个发射槽用于 MAC 操作,该 MAC 单元支持 32 位的复数乘法。

本文还在 FT-Matrix-Sim 上实现了传统的 VRF 和 MRF,大小均为 16 ×16,其中 VRF 只支持行访问,MRF 同时支持行访问和列访问,但不支持对多个行和列的并行访问。我们选取了无线通信和视频处理应用中几种包含矩阵运算的典型算法作为 benchmark,然后比较了几个 benchmark 在 FT-Matrix 和 TMS320C64x + [10] 上的性能。

3.1　实验结果

对于 LTE 和 H.264 应用中的几种典型算法核,图 6 给出了可配置矩阵寄存器文件相对传统的向量寄存器和矩阵寄存器文件的加速比。FT-Matrix 的计算资源是 C64X + 的 8 倍,因此,其相对 C64X + 的理论加速比为 8。FT-Matrix 单核在包含 MMRF 后,相比 TI 公司的 C64X + 内核,可以平均获得 5.65x ~ 7.71x 的加速比。

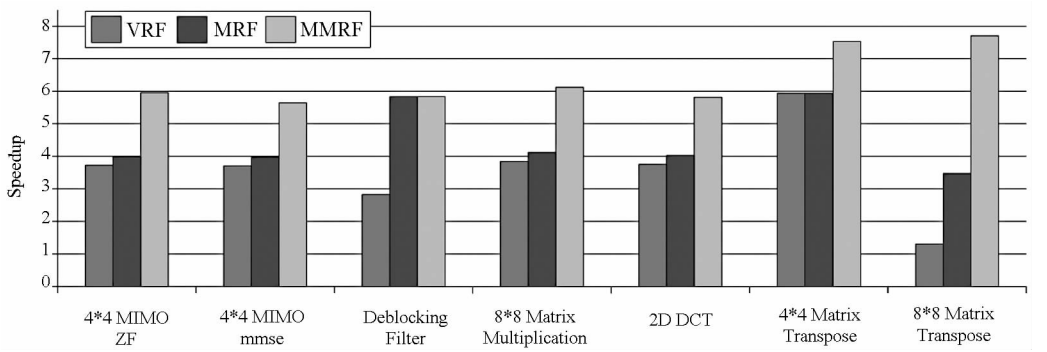

图 6　FT-Matrix 相对 TMS32064x + 的加速比

Fig. 6　The speedup of FT-Matrix with the MMRF over C64x +

同时,相对于只有 VRF 的 FT-Matrix,MMRF 可以获得平均 2.21x 的加速,最大加速比达 5.87。相对矩阵寄存器文件 MRF,MMRF 可以获得平均 1.6x 的加速比,最大加速比达 2.22x。

MMRF 带来的性能加速比主要归功于两个方面:不同访问模式的动态配置带来的灵活性和多粒度的并行行访问和列访问功能带来的高效性。灵活性可以提高功能单元和寄存器文件的利用率。多粒度的并行列访问可以消除矩阵运算中的数据重排工作,提升计算性能。

对于 deblocking filter 算法来说,MMRF 和 MRF 的加速比一样,这是因为在该算法中仅用到了 MMRF 的单路访问模式。这更好地说明了 MMRF 提供多粒度访问模式的必要性和优势。

3.2　硬件代价

用 Verilog 语言对 MMRF 进行了 RTL 级实现,端口数目为 3 读 2 写。每一个读写端口支持行访问和列访问。用 DC 工具对 RTL 级实现进行了综合,综合库为 TSMC 的 65nm 工艺库。时钟频率为 500MHz。用 Encounter 工具对综合结果进行了布局布线。表 1 给出了 VRF、MRF 和 MMRF 布局布线后的面积和功耗代价。

表1 VRF、MRF 和 MMRF 的面积和功耗对比

Tab. 1 The hardware cost of VRF, MRF and MMRF

部件	面积(mm^2)	功耗(mW)
MMRF	0.56	94
MRF	0.54	92
VRF	0.49	82
FT-Matrix	15.88	1555

与同样为3读2写的 VRF 和 MRF 相比,基于标准单元库实现的 MMRF 面积代价分别增加14.3%和3.7%,功耗代价分别增加14.6%和2.2%。在 MMRF、VRF 和 MRF 的硬件实现中,RA 占据了 VRF、MRF 和 MMRF 的大部分逻辑,而 VRF 和 MRF、MMRF 的 RA 规模是相同的。因此,MMRF 增加的硬件代价主要由读写端口的控制逻辑和地址译码逻辑产生。

4 MMRF 的全定制实现

寄存器文件是处理器内核的关键部件,直接影响全芯片的频率和布局。为了进一步减小 MMRF 的实现代价,我们对 MMRF 进行了全定制设计。

根据 MMRF 多粒度并行访问的特点,我们采取了层次化的定制设计技术来设计 RA 和地址译码逻辑。MMRF 被划分为16个2位的宏块,每一个2位的宏块同时支持行访问和列访问。每一个2位的宏块又被分为两个1位的 sub-RA,并且这两个1位的 sub-RA 共享一套译码逻辑,以减小面积和功耗。

层次化的设计策略也被用于读写端口的设计,以压缩最终的版图和优化线网连接。每一个宏块包含6个读端口和3个写端口。每一个读写端口拥有1个1位的寄存器阵列,该阵列包含16位的字线。每一个1位的寄存器阵列被划分为4个4×4的子阵列。每一个4×4的子阵列有本地字线和共享的全局字线。通过以上几个方面的层次化设计策略,可以获得规整高效的 MMRF 版图和布局布线。

图7给出了 MMRF 的2位宏块的版图。在 TSMC 的65nm 工艺库下,布局布线后的 MMRF 版图面积为 0.46mm^2,比 DC 综合的面积减少了17.9%。全定制设计得到的 MMRF 的电路性能也得到了很好地提升。DC 综合的关键路径延迟为1.6ns,而定制化 MMRF 的关键路径延迟仅为1.15ns,比 RTL 级设计减小了39.1%。

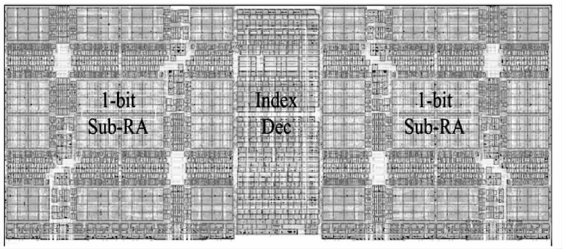

图7 MMRF 的2位宏块版图

Fig. 7 The layout of 2-bit sub-MMRF

5 结论

本文提出了 MMRF,用于在 SIMD 处理器上加速矩阵运算。MMRF 支持多粒度的并行行访问和列访问,从而消除矩阵运算在 SIMD 处理器上计算时的数据重排操作,提高了 SIMD 处理器的性能。MMRF 可以很好地应用于已有的 SIMD 处理器,而且不用修改原有处理器的指令集体系结构。

对于无线通信和媒体处理领域的一些典型算法,相对于 TMS320C64x + 处理器, FT-Matrix 在 MMRF 的帮助下获得5.65倍到7.71倍的性能提升。在 TSMC 的65nm 工艺库下,通过全定制设计,MMRF 的面积和关键路径延时分别减少了17.9%和39.1%。

参考文献(References)

[1] Samsung. Downlink MIMO for EUTRA. 3GPP TSG RAN WG1 meeting #44[R]. 3GPP R1-060335, 2006.

[2] Andrews J, Ghosh A, Muhamed R. Fundamentals of WiMAX: understanding broadband wireless networking[R], Prentice Hall, Mar, 2007.

[3] Woh M, Seo S, Mahlke S, et al. AnySP: Anytime anywhere anyway signal processing[C]. ISCA'09, June, 2009.

[4] Corbal J, Espasa R, Valero M. MOM: A matrix SIMD instruction set architecture for multimedia applications[C]// Proceedings of the ACM/IEEE SC99 Conference, 1999:1 - 12.

[5] Shahbahrami A, Juurlink B, Vassiliadis S. Versatility of extended subwords and the matrix register file[J]. ACM Transactions on Architecture and Code Optimization, 2008,5(1).

[6] Ciobanu C, Kuzmanov G, Gaydadjiev G, et al. A polymorphic register file for matrix operations[C]. International Conference on Embedded Systems: Architectures, Modeling and Simulation, July, 2006.

[7] Lin Y, et al. SODA: A low-power architecture for software radio[C]//Proc. of the 33rd Annual International Symposium on Computer Architecture, 2006:89 - 101.

[8] Flachs B, Asano S, Dhong S H, et al. The microarchitecture of the synergistic processor for a cell processor[J]. IEEE Journal of Solid-State Circuits, 2006,41(1).

[9] Krashinsky R, et al. The vector-thread architecture[C]// Proceedings of the 31st Annual International Symposium on Computer Architecture, 2004:52 - 63.

[10] Texas Instruments Incorporated. TMS320C64x + DSP Megamodule Reference Guide[R]. SPRU871J, 2008.