

面向定制结构的稀疏矩阵分块方法

邬贵明^{1,2} 王 森² 谢向辉² 窦 勇¹ 郭 松¹

(国防科学技术大学计算机学院 长沙 410073)¹ (数学工程与先进计算国家重点实验室 无锡 214125)²

摘 要 稀疏矩阵向量乘是科学计算的核心问题,采用定制结构来加速稀疏矩阵向量乘的执行对提升科学计算性能具有重要意义。针对目前面向定制结构的稀疏矩阵分块方法和表示方法的缺点,提出了稀疏矩阵二维均匀分块方法和相应的表示方法嵌套分块 CSR。实验结果表明,提出的稀疏矩阵分块方法和表示方法能够有效减少填充零个数。

关键词 稀疏矩阵向量乘,定制结构,稀疏矩阵,数据分块

中图分类号 TP302 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.012

Sparse Matrix Blocking Method for Custom Architecture

WU Gui-ming^{1,2} WANG Miao² XIE Xiang-hui² DOU Yong¹ GUO Song¹

(School of Computer, National University of Defense Technology, Changsha 410073, China)¹

(State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China)²

Abstract Sparse matrix vector multiplication is one of the most important applications in scientific computing. Using custom architectures to implement sparse matrix vector multiplication is introduced to improve the performance of scientific computing. To address the problem in existing sparse matrix blocking method and representation, a two-dimension uniform blocking method and its according representation were proposed in this paper. The experimental results show that the proposed sparse matrix blocking method and representation can reduce the padding zero significantly.

Keywords Sparse matrix vector multiplication, Custom architecture, Sparse matrix, Data blocking

1 概述

稀疏矩阵计算是科学和工程计算的核心问题,稀疏矩阵向量乘是其中最为典型的计算问题。稀疏矩阵向量乘由于其不规则的访存模式和访存密集的特征,导致它在当前的计算机上运行效率很低。为了充分开发稀疏矩阵向量乘的程序特征,相关研究人员提出定制结构来实现稀疏矩阵向量乘,以提升浮点计算效率。

Dorrance 等人^[1]提出一种能够充分利用存储带宽和计算资源的结构,其可以达到 91.85% 的平均计算效率,相比通用处理器计算效率提升了 50 倍以上。然而,他们使用的是单精度版本的浮点运算部件和稀疏矩阵向量乘,双精度运算比单精度运算具有更高的访存计算比,他们提出的方法针对双精度来说达不到很高的计算效率。Fowers 等人^[2]提出一种能够开发行间并行性的稀疏矩阵表示方法 CISR,其在不需要复杂的调度器和负载均衡器的条件下能够实现多行数据的并行操作。该表示方法过于复杂,用于编码和解码的代价过大。

Zhuo 等人^[3]和 Sun 等人^[4]关注于分块矩阵的结构实现。Zhuo 等提出一种前端为 k 个乘法器的加法树结构,矩阵每行被分成大小为 k 的子行,元素不足 k 则补零;该设计基于稀疏矩阵的垂直分块,如图 1(a)所示。Sun 等采用一种行分块结

构来表示稀疏矩阵,分块方法如图 1(b)所示。通过按行进行块划分形成子块,同一行的子块送入同一个由多个 PE 组成的加速器进行运算,每个 PE 计算一个子块,多个 PE 的计算结果送入规约电路进行部分和的累加,当同一行的所有子块运算完毕,便得到最终计算结果。每一行中每个子块需要的 x 的元素不同,因此对 x 的分块是不固定的。上述方法存在通用性差、计算资源利用率低、负载不平衡等问题。

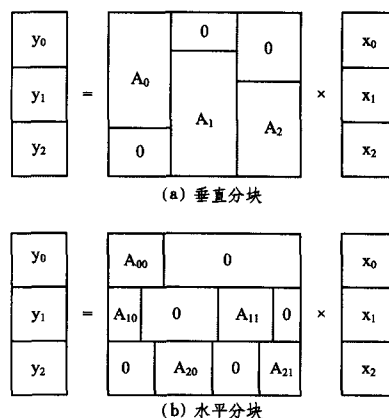


图 1 目前的稀疏矩阵分块方法

针对目前稀疏矩阵分块方法和表示方法的缺点,本文提

到稿日期:2014-11-18 返修日期:2015-01-05 本文受国家自然科学基金项目(91430214),湖南省科研资助项目(YB2013B007)资助。

邬贵明(1981—),男,博士,工程师,主要研究领域为高性能计算机体系结构,E-mail:wu.guiming@meac-skl.cn;谢向辉(1958—),博士,研究员,主要研究领域为高性能计算机体系结构、大数据;窦勇(1966—),男,教授,博士生导师,主要研究领域为计算机体系结构、可重构计算;郭松(1985—),男,博士生,主要研究领域为计算机体系结构。

出一种具有一定通用性的稀疏矩阵分块策略和相应的表示方法,该方法基于稀疏矩阵的二维均匀分块,可以有效避免目前基于单维分块所存在的问题。

2 稀疏矩阵和分块表示

稀疏矩阵中零元素不用存储,可以节省大量存储空间。稀疏矩阵最常用的是 CSR(Compressed Sparse Row)格式^[5],CSR 格式按行对非零元素进行压缩,包括 3 个数组:

- *val*:浮点型数组,按行连续存储非零元素的值;
- *col_ind*:整型数组,存储 *val* 数组中对应非零元素的列索引;
- *row_ptr*:整型数组,数组的元素 *row_ptr[i]* 存储了第 *i* 行的第一个非零元素在 *val* 和 *col_ind* 的偏移,即该行的起始地址。表示每行非零元素个数的数组 *len* 可由数组 *row_ptr* 得到: $len[i] = row_ptr[i+1] - row_ptr[i]$ 。

目前常见的针对通用处理器的稀疏矩阵分块存储格式是 BCSR(Block Compressed Sparse Row)格式^[6],BCSR 格式是对 CSR 的扩展,稀疏矩阵按行划分成 $\lceil n/r \rceil$ 块,每个行块存储了一串 $r \times c$ 大小的稠密数据,图 2 给出了一个矩阵按稠密数据块进行存储的例子,共 $K=5$ 个子块,子块的零元素被看成非零元素,因此共 20 个非零元素。块按行连续存储, *ind* 存储了每个块的起始列位置, *ptr* 指向存储在 *ind* 中的每个行块第一块的起始列位置,每个块被看作稠密矩阵且按稠密矩阵存储,因此会显式地填入零元素。零的填充会带来存储代价和性能损失,若每个块比较大时,由于不必显式存储每个非零元素的位置信息,可以节省存储空间。

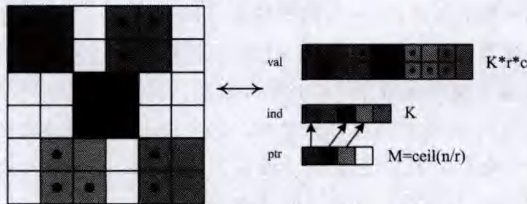


图 2 BCSR 存储格式

3 嵌套分块 CSR 格式

针对稀疏矩阵向量乘定制结构,本文提出稀疏矩阵的二维均匀分块方法和相应的嵌套分块 CSR 格式。

图 3 给出了二维均匀分块方法,稀疏矩阵的行列被均匀划分,向量 x 同时也被均匀划分,稀疏矩阵的子块大小与向量 x 的子块大小相等。采用二维均匀分块方法是基于以下几点考虑:(1)与两种单维分块方法相比,二维均匀分块只要保持很少的子块行列信息就能隐式地存储每个非零元素的行列位置信息。(2)在二维均匀分块方法中,每个非零元素只需保存它在子块内的相对偏移量,而这个偏移量的位宽可以远远小于 32 位,从而极大地降低了存储需求;。3)矩阵维度和向量维度过大时,FPGA 的片内 RAM 无法保存向量 x 和 y 的所有元素,按行(或列)的条状分块方式在访问向量 x (或 y)的元素时需要直接访问外部存储器,因此增大了存储带宽需求;若采用二维均匀分块,少量 RAM 便可以存储分块后的 x (或 y),并且能够达到数据重用。

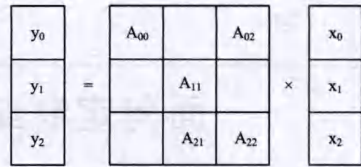


图 3 二维均匀分块方法

基于二维均匀分块,提出嵌套分块 CSR(Nested Block CSR,NBCSR)格式,图 4 给出了一个含有 $s=13$ 个非零元素的稀疏矩阵例子。在该例子中,稀疏矩阵被划分成块大小为 $b=2$ 的 $M=3$ 个块行,共有 $K=4$ 个非零子块,每个子块使用不同的颜色标示。稀疏矩阵被二维均匀划分后,每个子块使用 CSR 格式存储;稀疏矩阵的每个子块被看作稀疏矩阵的元素采用 CSR 格式存储,最终存储的是每个子块行指针的位置,因此嵌套分块 CSR 可称为两级 CSR 格式。嵌套分块 CSR 格式包括两层数据结构:顶层 CSR 数据结构和底层 CSR 数据结构。顶层 CSR 数据结构包括:

- *bcol_ind*:保存子块作为元素时的行位置,大小为子块个数 K ;
- *brow_ptr*:保存每个子块行在 *bcol_ind* 的起始偏移量,大小为 $M+1$,其中 $M=\lceil n/b \rceil$, n 为矩阵维度;
- *val_ptr*:指向每个子块中第一个元素的存储位置,即每个子块在 *row_ptr* 的起始偏移量,大小为 $K+1$ 。

底层 CSR 数据结构包括 *row_ptr*、*col_ind* 和 *val*,意义与通常的 CSR 一样,每个子块按行连续存储其 CSR 格式。

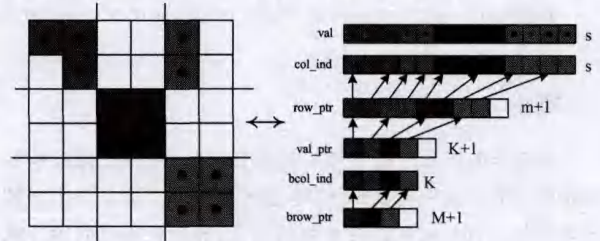


图 4 嵌套分块 CSR 存储格式

4 实验评估

利用 Matlab 实现了提出的稀疏矩阵分块方法,测试集来自 Florida 大学的 Sparse Matrix Collection^[7]。在嵌套分块 CSR 格式生成时有个重要的参数即前端乘法器个数 k ,这个参数直接对应于硬件设计,其限制了能同时进入硬件进行运算的稀疏矩阵元素个数,从向量处理的角度来看, k 为向量宽度,这个参数受限于存储带宽。若同时进入硬件的稀疏矩阵有效元素个数小于 k ,为了保证计算的正确性,需要用零来补齐 k 个元素进入前端乘法器以进行运算。提出的稀疏矩阵分块表示方法能够有效支持来自不同行的元素同时进行运算,减少补零个数。表 1 列出了所采用的稀疏矩阵的特征和分块后的块数。

表 2 列出了已有稀疏矩阵分块方法和提出的分块方法在 k 等于 2、4 和 8 时针对不同测试矩阵的填零数目,采用的块大小均为 8192。比较两种方法的填零数目可以看出,我们的分块方法填零数目非常少,几乎可以忽略不计。对于 cage11 矩阵,当 $k=2$ 时可以不用填零,我们的分块方法减少填零的效果是非常明显的。

(下转第 79 页)

- [4] 熊金波,姚志强,等.基于行为的结构化文档多级访问控制[J].计算机研究与发展,2013,50(7):1399-1408
Xiong Jin-bo, Yao Zhi-qiang, et al. Action-Based Multilevel Access Control for Structured Document [J]. Journal of Computer Research and Development, 2013, 50(7): 1399-1408
- [5] 俞士汶,段慧明,朱学锋,等.北京大学现代汉语语料库基本加工规范[J].中文信息学报,2002,16(5):49-56
Yu Shi-wen, Duan Hui-ming, Zhu Xue-feng, et al. The Basic Processing of Contemporary Chinese Corpus at Peking University SPECIFICATION[J]. Journal of Chinese Information Processing, 2002, 16(5): 49-56
- [6] 《知网》[OL]. [2014-2-20]. <http://www.keenage.com>
- [7] 台湾中央研究院.中文句结构树资料库[OL]. [2014-2-20] <http://TreeBank.sinica.edu.tw>
- [8] 黄居仁.中文词汇网络[OL]. [2014-2-20]. <http://lope.linguistics.ntu.edu.tw/cwn>
- [9] Yiu C L K, Wong Wai. Chinese character synthesis using METAPOST[OL]. [2014-2-22]. <http://www.tug.org/TUGboat/tb24-1/yiu.pdf>
- [10] CDL: Character description language [OL]. [2014-2-22]. <http://www.wenlin.com/cdl>
- [11] Peebles D G. A Structural Representation for Chinese Characters[OL]. [2014-1-24]. <http://www.cs.dartmouth.edu/reports/TR2007-592.pdf>
- [12] 林民, 宋柔. 汉字字形形式化方法研究[J]. 计算机科学, 2007, 34(11): 185-189
- Lin Min, Song Rou. Formal Description of Chinese character Glyph[J]. Computer Science, 2007, 34(11): 185-189
- [13] 栗青生, 熊晶, 吴琴霞, 等. 基于特征加权的汉字点笔画生成研究[J]. 北京大学学报(自然科学版), 2014, 50(1): 153-160
Li Qing-sheng, Xiong Jing, Wu Qin-xia, et al. Study of Feature Weighted-based Generation Method for Dian Strokes of Chinese Character[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 153-160
- [14] 吴琴霞, 栗青生, 高峰. 基于语义构件的甲骨文字库自动生成技术研究[J]. 北京大学学报(自然科学版), 2014, 50(1): 161-166
Wu Qin-xia, Li Qing-sheng, Gao Feng. Study on the Technique of Automatic Generation of Oracle Characters Based on Semantic Component [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 161-166
- [15] 吴琴霞, 栗青生. 基于动态描述库的汉字字形自动生成技术研究[J]. 科学技术与工程, 2013, 13(15): 4425-4431
Wu Qin-xia, Li Qing-sheng. Research of Chinese Character Auto-generation Technology Based on Dynamic Description Library [J]. Science Technology and Engineering, 2013, 13(15): 4425-4431
- [16] 栗青生, 吴琴霞, 杨玉星. 甲骨文字形动态描述库及其字形生成技术研究[J]. 北京大学学报(自然科学版), 2013, 49(1): 61-67
Li Qing-sheng, Wu Qin-xia, Yang Yu-xing. Dynamic Description Library for Jiaguwen Characters and the Research of the Characters Processing [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 61-67

(上接第 64 页)

表 1 稀疏矩阵特征

稀疏矩阵	维度	非零个数	块数
apache1	80800	542184	28
ted_B	10605	144579	4
bcsstk36	23052	1143140	9
2cubes_sphere	101492	1647264	169
nasasrb	54870	2677324	19
boneS01	127224	5516602	46
ASIC_100ks	99190	578890	169
Zd_Jac6	22835	1711557	8
rajat26	51032	247528	44
cage12	130228	2032536	162
cage11	39082	559722	19
viscoplastic2	32769	381326	13

表 2 填零数目比较

稀疏矩阵	k=2		k=4		k=8	
	优化前	优化后	优化前	优化后	优化前	优化后
apache1	66000	4	160840	56	220568	72
ted_B	4593	1	13737	1	41733	5
bcsstk36	10100	2	47088	16	89652	20
2cubes_sphere	93686	70	245140	248	591312	568
nasasrb	23632	2	108756	24	204460	60
boneS01	57472	10	220026	58	589734	158
ASIC_100ks	73642	84	222414	250	522758	614
Zd_Jac6	13435	3	53711	11	147291	35
rajat26	39386	24	95000	80	298152	152
cage12	286062	30	1013220	88	2629448	192
cage11	41148	0	129326	6	358694	38
viscoplastic2	36720	4	112170	14	324794	42

结束语 本文针对目前面向定制结构的稀疏矩阵分块方法和表示方法的不足,提出了一种稀疏矩阵二维均匀分块方法,并给出了相应的表示方法。实验结果表明,本文提出的稀

疏矩阵分块方法和表示方法能够有效减少填零个数,将有效开发定制结构的并行性。

参考文献

- [1] Dorrance R, Ren F, Marković D. A Scalable Sparse Matrix-Vector Multiplication Kernel for Energy-Efficient Sparse-BLAS on FPGAs[C]// Proceedings of the 2014 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA). ACM, 2014: 161-170
- [2] Fowers J, Ovtcharov K, Strauss K, et al. A High Memory Bandwidth FPGA Accelerator for Sparse Matrix-Vector Multiplication[C]// Proceedings of the 2014 IEEE 22nd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). IEEE, 2014: 36-43
- [3] Zhuo L, Prasanna V K. Sparse Matrix-Vector Multiplication on FPGAs[C]// Proceedings of the 13th ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA). ACM, 2005: 63-74
- [4] Sun J, Peterson G D, Storaasli O O. Sparse Matrix-Vector Multiplication Design on FPGAs[C]// Proceedings of the 15th Annual IEEE Symposium Field-Programmable Custom Computing Machines (FCCM). IEEE, 2007: 349-352
- [5] Stathis P T. Sparse Matrix Vector Processing Formats [D]. Delft, Netherlands: Delft University of Technology, 2004
- [6] Williams S, Vuduc R, Oliker L, et al. Optimizing Sparse Matrix-Vector Multiply on Emerging Multicore Platforms[J]. Journal of Parallel Computing, 2009, 35(3): 178-194
- [7] The University of Florida Sparse Matrix Collection[OL]. <http://www.cise.ufl.edu/research/sparse/matrices>