计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200090
58(3): 458-466, 2021

基于双向位图的 CSR 大规模图存储优化

甘新标 谭 雯 刘 杰
(国防科技大学计算机学院 长沙 430017)
(xinbiaogan@nudt.edu.cn)

Bidirectional-Bitmap Based CSR for Reducing Large-Scale Graph Space

Gan Xinbiao, Tan Wen, and Liu Jie
(*College of Computer Science and Technology, National University of Defense Technology, Changsha* 410073)

Abstract Graph500 is an important and famous benchmark to evaluate data-intensive applications for supercomputers in the big data era. The graph traversal processing ability of pre-exascale system is mainly restricted to memory space and communication bandwidth, especially the utilization of memory space ultimately determines the testing graph scale, and the graph testing scale absolutely dominates the performance of Graph500. Hence, Bi-CSR (bidirectional-bitmap CSR) is proposed based on CSR (compressed sparse row) for testing Graph500 on Tianhe pre-exascale system, The Bi-CSR would reduce large-scale graph space by introducing row-bitmap and column-bitmap to compress sparse matrix storage for Graph500. The aim of row-bitmap based on CSR is mainly cutting down graph memory space, while column-bitmap based on CSR would not only further reduce memory space but also improve graph traversal by using array of VPE(vector processing element), because VPEs are optimized and equipped in Tianhe pre-exascale system, which would speedup graph traversal when making fully use of VPEs. Accordingly, Bi-CSR would greatly reduce large-scale graph space while introducing row-bitmap and column-bitmap to compress sparse matrix storage of Graph500 for Tianhe pre-exascale system. Experimental results demonstrate that Bi-CSR would reduce large-scale graph space by 70% when testing input of Graph500 is 237 on Tianhe pre-exascale system with 2.131E+12 TEPS (traversed edges per second).

Key words Graph500; bidirectional-bitmap; compress sparse matrix storage; graph traversal; Tianhe pre-exascale system

摘 要 大数据时代,Graph500 是评测超级计算机处理数据密集型应用能力的重要工具,E 级验证系统的图遍历处理能力主要受限于内存空间和访存带宽,尤其是内存空间利用率直接决定了图的测试规模和测试性能.针对天河 E 级验证系统小内存特征,提出了基于双向位图的大规模图数据压缩存储方法(bidirectional-bitmap based CSR, Bi-CSR),Bi-CSR 在 CSR 矩阵压缩的基础上引入行方向位图和列方向位图协同完成稀疏矩阵压缩存储,行方向位图主要负责行方向位图的压缩存储与索引,列方向位图除了进一步压缩图存储空间,还负责为顶点遍历向量并行优化提供加速空间.Bi-CSR 大幅度减少了稀疏

收稿日期:2020-02-21;**修回日期**:2020-06-05
基金项目:国家数值风洞项目(NNW2019ZT6-B21,NNW2019ZT6-B20,NNW2019ZT5-A10);国家重点研发计划项目(2018YFB0204301);湖南省自然科学基金项目(2020JJ4669);并行分布式处理实验室基金项目(6142110190206,6142110180203)
This work was supported by the National Numerical Wind Tunnel Project (NNW2019ZT6-B21, NNW2019ZT6-B20, NNW2019ZT5-A10), the National Key Research and Development Program of China (2018YFB0204301), the Hunan Provincial Natural Science Foundation of China (2020JJ4669), and the Foundation of Parallel and Distributed Processing Laboratory (6142110190206, 6142110180203).

矩阵存储空间.面向天河 E 级验证系统,当图输入规模为 2^{37} 时,Graph500 的图存储空间节约效率接近 70%,全系统稳定测试性能为 2.131E+12TEPS,性能最大加速比超过 100 倍.

关键词 Graph500;双向位图;稀疏矩阵压缩存储;图遍历;天河 E 级验证系统

中图法分类号 TP391

大数据时代,超级计算机系统结构面临着巨大的机遇和全新的挑战.图搜索处理对计算机系统存储和带宽要求甚高,传统的超级计算机系统处理大规模图数据的效率低下.Graph500 作为数据密集型计算机系统的评测基准,能够准确反映 E 级计算机系统的大数据处理能力和指导 E 级计算机系统设计.

Graph500 以每秒遍历的边数(traversed edges per second, TEPS)为性能指标来衡量超级计算机处理数据密集型应用的能力.Graph500 测试性能主要受限于图测试规模和访存带宽,尤其是内存空间大小直接决定了图的测试规模.因此,本文提出了基于双向位图的大规模图数据压缩存储方法(bidirectional-bitmap based compressed sparse row, Bi-CSR)以增加 Graph500 图测试规模,并提升 Graph500 图测试性能.

本文的主要贡献有 2 个方面:

1) 提出了基于双向位图的稀疏矩阵压缩存储方法 Bi-CSR,大幅减少了大规模图存储空间.

2) 将 Bi-CSR 应用于天河 E 级验证的 Graph500 测试,达到了良好的测试性能.

1 相关工作

很多现实问题可以抽象为图形结构,因此,图形理论在众多学科中应用广泛[1-3].从现实问题中抽象出来的图形结构通常具有 2 个特点:小世界性(small-world)[4]和无尺度性(scale-free)[5].

图遍历是大数据时代背景下的一种典型的数据密集型应用,Graph500 测试基准是典型的数据密集型程序.其核心搜索程序能够适用于数据密集型应用,能够解决真实世界中的实际搜索问题,测试数据集能够反映真实数据集的普遍特征,并且 Graph500 具有较差的时间和空间局部性,因此,Graph500 具有一般大数据问题的主要特征.近年来,面向 Graph500 核心算法 BFS 的优化遍历取得了一系列重要的进展.

Bader 等人[6]在 Cray MTA-2 多结点上实现了多线程的并行 BFS 算法,利用硬件多线程技术来隐藏访存延迟,取得了很好的性能提升;Agarwal 等人[7]提出利用位图的数据结构来表示算法中的 visit 结构,增加了 visit 的局部性,减少了访存的次数;Beamer 等人[8]开创性地提出了一种 Top-down 与 Bottom-up 相结合的混合算法,所实现的混合算法能够有效地减少搜索过程中遍历的边数,减小了访存开销,极大地提高了程序的性能;Ueno 等人[9]面向极大规模并行与分布式机器提出了一种高效的 BFS 算法,并应用于"京"超级计算机系统,达到了的良好测试性能;清华大学林恒等人[10-11]面向"神威太湖之光"实现了一种可扩展的 BFS 算法,实现了模块流水映射、无竞争的数据混洗、基于组消息的批处理等关键技术,在神威太湖之光超级计算机全系统上获得了 23755.7GTEPS 的测试性能[11],上述优化技术和方法本质上都属于加速图遍历来提升 Graph500 测试性能.图遍历速度是 Graph500 测试性能的重要影响因素,工程实践表明:Graph500 测试性能直接受限于图测试规模.分析 Graph500 榜单可知,测试图规模越大,Graph500 性能越高.内存空间大小直接决定了 Graph500 图的测试规模,Graph500 的 Kronecke 生成器生成的图形通常用邻接矩阵表示,由于顶点的平均度数较低,其邻接矩阵为稀疏矩阵,因此,图形的邻接矩阵存储格式直接影响了 Graph500 图测试规模.

典型的稀疏矩阵存储格式有 COO(coordinate),DIA(diagonal)[12],CSR(compressed sparse row),CSB(compressed sparse block),ELL(ELLPACK)和 HYB(hybrid ELL+COO)等[13],上述压缩格式在天河验证系统 512 个结点上的最大图测试规模及性能如图 1 所示,纵轴采取 GTEPS 为性能指标.

由图 1 可知,CSR 格式在天河验证系统上的最大图测试规模为 $scale_{\max}=36$,稳定测试性能为 1310GTEPS.各数据压缩方法除了压缩效率和最大图测试规模及性能的不同,各种数据压缩方法操作性、灵活性、可扩展性以及适用领域方面也存在明显的差异,其中,DIA 和 ELL 格式在进行稀疏矩阵矢量乘积时效率最高,它们是应用迭代法求解稀疏线性系统最快的格式[12];CSB 的优势是易于向量并行;ELL 的优点是快速,COO 的特点是灵活,二者结合

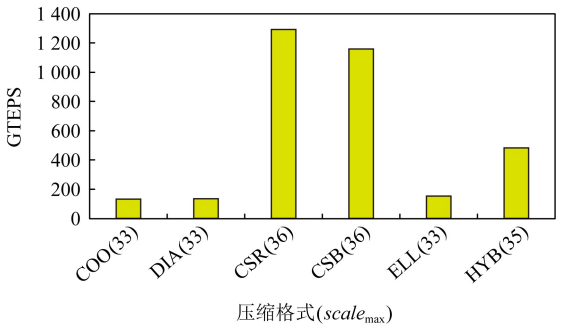


Fig. 1 Various format testing on available max scale with GTEPS comparison

图 1 不同格式的最大测试图规模及性能比较

后的 HYB 格式是一种优良的稀疏矩阵表示[13-14]；CSR 相比 COO，HYB，CSB，DIA，ELL，更加灵活，易于操作[13]；因此，Graph500 参考实现中采用了 CSR 存储 Kronecker 生成图.CSR 可以采用数组或位图 2 种数据结构来完成图压缩存储，基于数组结构的 CSR 图存储与基于单向位图结构的 CSR 图存储空间利用率十分有限[9-11]，为此，本文提出了基于双向位图的 CSR 大规模图存储优化方法进一步提高 Graph500 图压缩存储效率，最大化 Graph500 图测试规模和提升 Graph500 测试性能.

2 Graph500

Graph500 以每秒遍历边数 TEPS 为性能(performance)指标来衡量超级计算机处理数据密集型应用的能力.TEPS 要由图的规模和图形遍历的时间共同决定，尤其是内存空间大小直接决定了图的测试规模.

2.1 基本流程

Graph500 基本流程包括图生成、图构造与存储、图遍历以及图验证 4 个主要步骤[15-16]，Graph500 图生成采用了 Kronecker 生成器，生成的图满足小世界性与无尺度性[15].Graph 500 图形生成器有 2 个输入参数，分别是图形规模 $scale$ 和边因子 $edgefactor$.

假定 $scale=n$，$edgefactor=m$，则生成的图具有 2^n 点和 $m\times 2^n$ 条无向边[15].

Graph500 中仅包含了 2 个计时模块：图存储变换和图遍历，并且图遍历占据其中绝大部分时间，假定 Graph500 的搜索时间为 t，则对于 $scale=n$，$edgefactor=m$ 的图，P_{teps} 为性能指标，$P_{\text{teps}}=m\times 2^n/t$.

图存储将生成的图信息转换为包含顶点、边以及联系的元组信息列表，通常以邻接矩阵的形式高效存储.P_{teps} 由图规模和搜索时间共同决定，并且图的规模对 P_{teps} 的影响尤为关键，图结构的存储和表示模式是影响 Graph500 图测试规模的重要因素.

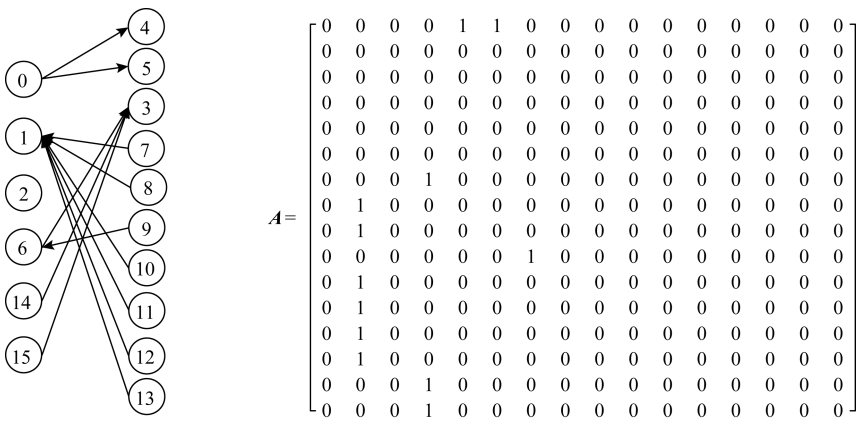
2.2 图结构的存储与表示

图 $G(V,E)$ 包含顶点集合 V 和边集合 E.通常对图形的顶点进行编号，使用 v_i 表示图中编号为 i 的顶点，使用顶点对 (v_i,v_j) 表示顶点 v_i 到顶点 v_j 的边.对于无向图的每条边，可以看作 2 条有向边.

Graph500 生成图形通常采用邻接矩阵 $\boldsymbol{A}$ 表示，其中的每一行 $\boldsymbol{A}_i$ 为邻接表，图的邻接矩阵表示如图 2 所示.

邻接矩阵中第 i 行、第 j 列的元素 A_{ij} 表示边 (v_i,v_j).对于无权图，仅需要说明边是否存在，通常使用 1 表示存在这样的边，0 表示不存在这样的边.

大部分从现实问题抽象出来的图，如 Kronecker



(a) 测试图结构 (b) 测试图对应的邻接矩阵

Fig. 2 Presentation with adjacency matrix storage for graph

图 2 图的邻接矩阵表示

图生成器生成的图,顶点的平均度数远小于顶点总数,并且符合 6 度分离原则和小世界效应[4-5],其邻接矩阵通常为稀疏矩阵[12,18-20],因此,Graph500 中采用了基于 CSR 的稀疏矩阵存储方法,CSR 包含 2 个数组:*columns* 和 *rowstarts*.数组 *columns* 存储按行压缩的列标号,数组 *rowstarts* 存储对应行在 *columns* 中的索引位置,如图 3(a)所示.

$$A=\begin{bmatrix}
0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}$$

(a) 邻接矩阵

columns:	4	5	3	1	1	6	1	1	1	1	3	3

rowstarts:	0	2	2	2	2	2	2	3	4	5	6	7	8	9	10	11	12

(b) 基于邻接矩阵的CSR索引

Fig. 3 Structureof CSR storage for graph

图 3 图的 CSR 存储结构

如图 3(b)所示,*columns* 中的标号为邻接矩阵 ***A*** 对应非零元素的列标号,第 1 个数字 4 表示第 1 个非零元的列标号为 4,第 2 个数字 5 表示第 2 个非零元的列标号为 5,第 3 个数字 3 表示第 3 个非零元的列标号为 3,第 4 个数字 1 表示第 4 个非零元的列标号为 1.*rowstarts* 中的索引位置对应 ***A*** 中非零元的行标号相对偏移,即对应行中非零元的个数,如:第 2 个数字 2 和第 1 个数字 0 表示 ***A*** 中第 0 行中非零元的个数为 2−0=2,第 3 个数字 2 和第 2 个数字 2 表示 ***A*** 中第 1 行中非零元的个数为 2−2=0,第 4 个数字 2 和第 3 个数字 2 表示 ***A*** 中第 2 行中非零元的个数为 2−2=0.

Graph500 生成无向图的邻接矩阵表示总是对称的,采用 CSR 图数据存储.CSR 作为通用的稀疏矩阵存储方法,具有存储模式灵活、易于操作等诸多优势,但是,CSR 的大规模图存储效率与天河 E 级验证系统的小内存体系结构特征并不能完美适配,这是因为 Graph500 的测试性能主要受限于访存带宽和内存规模,当系统内存超过阈值后(即输入图规模达到一定规模),访存带宽将激变成为性能瓶颈,基于天河 E 级验证系统的规模和配置,访存带宽尚未成为性能瓶颈,天河 E 级验证系统的 Graph500 测试性能主要受限于内存大小.面向天河 E 级验证系统的 Graph500 优化设计应当充分发挥验证系统优势,最小化天河 E 级验证系统小内存的局限性,设计与验证系统的规模与存储层次相适应的图数据压缩方法,准确度量天河 E 级验证系统的大数据处理能力,因此,面向天河 E 级验证系统提出了一种基于双向位图的稀疏矩阵压缩大规模图存储方法 Bi-CSR,以最大化 Graph500 图测试规模和提升 Graph500 测试性能.

3 基于双向位图的稀疏矩阵压缩存储方法

针对天河 E 级验证系统小内存的特征,提出了与天河 E 级验证系统的小内存匹配的基于双向位图的大规模图数据压缩存储方法 Bi-CSR.Bi-CSR 在 CSR 存储结构的基础上仅保留存储一个或者多个顶点或边的起始位置来压缩 CSR 的数据结构,并且在行列 2 个方向分别引入位图(bitmap)来辅助识别顶点和边信息,其中行位图中每一位存储一个顶点信息,列位图中每一位存储连续列编号信息.

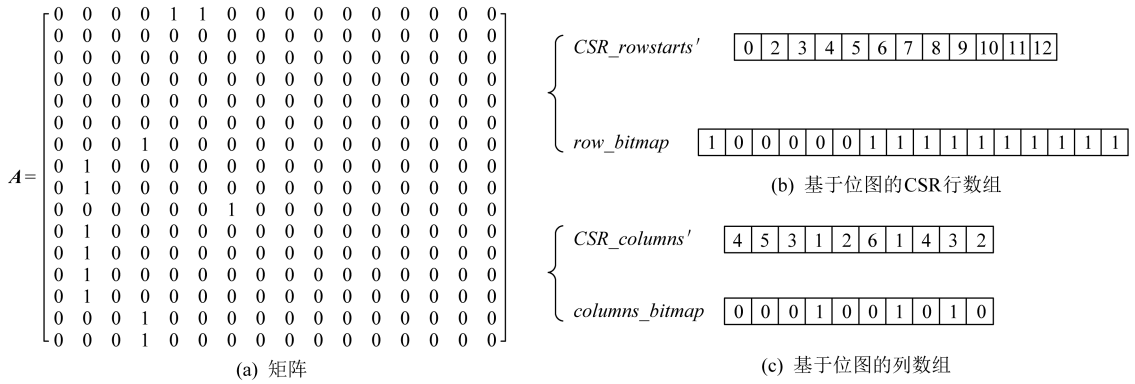
3.1 基于行方向位图的稀疏矩阵压缩

基于行方向位图的稀疏矩阵压缩主要目的是减少大规模图存储空间,扩大 Graph500 图测试规模和减少访存次数.基于行方向位图的 CSR 存储通过引入行方向位图数组 *row_bitmap*,并且,将 CSR 存储结构中的数组 *rowstarts* 采用改进的行数组 *CSR_rowstarts′* 和位图数组 *row_bitmap* 来表示,如图 4(b)所示.将数组 *rowstarts* 中零元行标号索引采用 1 b 来表示数组 *rowstarts* 中使用 1 个整型(32 b)来表示的索引信息,最大限度压缩了数据存储空间.位图数组 *row_bitmap* 中的每一位对应所在的行是否有非零元.图规模越大,图的稀疏特征表现越明显,越多的采用整型(32 b)表示的索引信息可以使用 1 b 来辅助标注,最大限度节约了图数据的存储空间.

3.2 基于列方向位图的稀疏矩阵压缩

基于列方向位图的稀疏矩阵压缩主要目的包括进一步减少大规模图存储空间,最大限度扩大 Graph500 图测试规模和减少访存次数,以及为面向天河 E 级系统的可变宽度向量化开辟优化空间.

基于列方向位图的 CSR 存储通过引入列方向位图数组 *columns_bitmap*,并且,将 CSR 存储结构



(a) 矩阵 (b) 基于位图的CSR行数组 (c) 基于位图的列数组

Fig. 4 Structure on Bi-CSR storage for graph

图 4 Bi-CSR 压缩存储结构

中的数组 *columns* 采用改进的列数组 *CSR_columns′* 和位图数组 *columns_bitmap* 来表示，如图 4(c)所示.为了更加清晰地描述数组 *columns_bitmap* 中的含义，作定义：

定义 1. 列连续片段.列方向上连续出现非零元，连续片段的长度 $len \geqslant 2$.

定义 2. 列非连续片段.列方向上未连续出现非零元，列非连续片段的长度 $len=1$.

定义 3. 紧前位(closely front bit, CFB).位图数组与当前位紧邻的前一位，记为 B_{CF}.

定义 4. 紧前缺位(non-closely front bit, non-CFB).位图数组中的首位没有紧前位，即为紧前缺位，记为 non_B_{CF}.

columns_bitmap 中的每一位对应所在的列编号是否为连续列编号片段的起始列编号，准确含义：

$$bit=\begin{cases}1：连续片段的列标签.\\ 0\Rightarrow\begin{cases}non\_B_{CF}：非连续片段的列标签.\\ B_{CF}=\begin{cases}1：长度.\\ 0：非连续片段的列标签.\end{cases}\end{cases}\end{cases}$$

数组 *columns* 中的每一位对应列连续片段 *fragment* 的起始列编号 *column_label*，或是列连续片段的长度值 $len \geqslant 2$，或是列非连续片段 *non_fragment* 的列编号 *column_label*，具体含义应结合数组 *columns_bitmap* 中对应位标识来判定；*columns_bitmap* 中的每一位对应数组 *CSR_columns′* 中相同偏移的元素值是否为连续列编号片段的起始列编号 *column_label*，1 表示 *CSR_columns′* 中相同偏移的元素值为列连续片段的列编号，0 表示为非连续片段 *non_fragment* 的列编号或是连续片段长度的标识.若 0 的紧前缺位 non_B_{CF} 是 0 则表示 *CSR_columns′* 中相同偏移的元素值为非连续片段的列编号 *column_label*，若 0 的紧前位 B_{CF} 是 1 则表示 *columns′* 中相同偏移的元素值为连续片段的长度值，若 0 的紧前位 B_{CF} 是 0，则表示 *columns′* 中相同偏移的元素值为非连续片段 *non_fragment*，列编号 *column_label*.例如：由于紧前缺位 non_B_{CF}，*columns_bitmap* 第 1 个数字 0，表示对应 *columns′* 中第 1 个数字 4 为第 1 个非零元的列编号；由于紧前位 B_{CF} 为 0，第 2 个数字 0 表示对应 *CSR_columns′* 中第 2 个数字 5 为第 2 个非零元的列编号；第 4 个数字 1 表示连续片段的起始列标号为 *CSR_columns′* 中第 4 个数字 1，即第 4 个非零元的列编号为 1；由于 CFB 第 4 个数字是 1，第 5 个数字 0 表示连续片段的长度，其长度值为 *CSR_columns′* 中第 5 个数字 2.

4 实验与性能分析

为了验证系统的大规模图存储优化方法 Bi-CSR，将 Bi-CSR 应用于天河 E 级验证系统进行 Graph500 工程测试.

4.1 天河 E 级验证系统

天河 E 级验证系统主要包括计算处理分系统、高速互连分系统、并行存储分系统、服务处理分系统、监控诊断分系统和基础架构分系统等，如图 5 所示.面向天河 E 级验证系统的 Graph500 重点关注其中的计算处理分系统和并行存储分系统.

计算处理分系统由 512 个计算结点构成，是系统主机的计算核心.每个计算结点包括 3 颗 Matrix-2000＋微处理器，运行频率为 2.0 GHz，峰值性能为 6TFLOPS.

并行存储分系统采用超高并发度的多层面并行

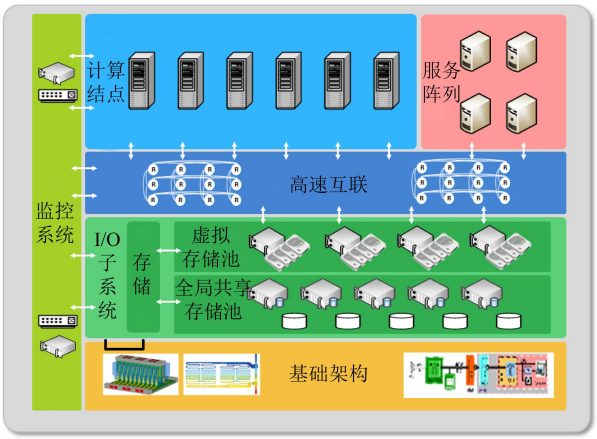


Fig. 5 Architecture for Tianhe pre-exascale system

图 5 天河 E 级验证系统结构图

存储架构;全局共享存储层包含 20 个 I/O 存储结点,每个存储结点集成 64 TB 磁盘阵列,提供共享存储总容量 1.28 PB.配置 2 个 I/O 管理结点和 1 套 8 TB 全闪存阵列存储管理元数据,提高元数据 I/O 吞吐率.E 级验证系统并行存储分系统的具体配置包括:32 个 I/O 加速结点、20 个 I/O 存储结点、2 个 I/O 管理结点和 1 套全闪存阵列,总存储容量 1.28 PB.

4.2 实验配置

为了验证天河 E 级原型系统的数据密集型应用处理能力,本文将 Bi-CSR 应用于 Graph500 测试中,形成了面向天河 E 级验证系统的 Graph500 测试版本 THE-G500(Tianhe exascale Graph500).

为了验证 Bi-CSR 的有效性,我们在天河 E 级验证系统上分别采用标准参考版本 Graph500 和面向天河 E 级验证系统的 THE-G500 版本进行了详细的测试与比较分析,实验平台为部署于天津超算的天河 E 级原型系统,与 Graph500 测试密切面相关的系统参数如表 1 所示:

Table 1 Configuration for Testing Graph500

表 1 面向 Graph500 的系统配置

体系结构参数	值	备注
结点总数	512	
核/结点	384	
核总数	196 608	
每结点内存/GB	192	
总内存/GB	98304	
自主互联网络/Gbps	200	TH-Ex2

4.3 Bi-CSR 存储空间分析与性能开销验证

Bi-CSR 图数据进一步压缩的基本思想为大规模图矩阵的 CSR 压缩存储中每行或每列中出现多个非零元,然后利用行列双向位图协作行列标号和位移的计算,即将 32 b 标识的列标号和位移改进为 1 b 的位图索引.

理论上,每个非零元减少的存储空间为

$$\Delta non\_z=\frac{31}{32}. \tag{2}$$

由于 Graph500 的 Kronecker 图生成器生成的图符合小世界性和 6 度分离理论,应用到 Graph500 图的 CSR 存储结构中,可以推断出图的稀疏矩阵中每行出现至少 2 个非零元的概率为

$$\varphi_{\text{row}}=\frac{d-6}{d}, \tag{3}$$

其中,d 为 CSR 图结构中每个顶点的平均度数.

由于装备天河 E 级验证系统的众核处理器支持可变宽度向量并行,因此,基于列方向位图的 CSR 压缩由出现非零元片段的概率可以松弛为每列出现多个非零元的概率:

$$\varphi_{\text{columns}}=\frac{d-6}{d}. \tag{4}$$

因此,仅应用基于行方向位图的 Bi-CSR 的图数据存储空间减少量:

$$\Delta row=\Delta non\_z\times\varphi_{\text{row}}. \tag{5}$$

仅应用基于列方向位图的 Bi-CSR 的图数据存储空间减少量:

$$\Delta columns=\Delta non\_z\times\varphi_{\text{columns}}. \tag{6}$$

综上所述,基于行列 2 个方向位图的 Bi-CSR 的图数据存储空间减少量为

$$\Delta space=1-(1-\Delta row)\times(1-\Delta columns). \tag{7}$$

实际应用的 Bi-CSR 图数据压缩中,由于 Graph500 中顶点平均度数 $d=16$,测试中不同图规模以 16×32 的拓扑结构分布于 E 级验证系统上,每个结点存储空间减少量如表 2 所示:

Table 2 Space Save Comparison Between Bi-CSR and CSR

表 2 基于 Bi-CSR 的图数据空间节约率

scale	CSR/GB	Bi-CSR-v/GB	节约空间比率/%
34	24 544	9 683	60.55
35	49 088	17 494	64.36
36	98 176	32 016	67.39
37	196 532	59 278	69.81
38	264 163	128 542	67.28

由表 2 可知,THE-G500 在天河 E 级验证系统的测试过程中,应用 Bi-CSR 方法后大规模图数据存储空间的节约率与式(5)～(7)表征的理论存储空间节约率基本吻合,在 $scale=34$ 时,大规模图存储空间节约率超过了 60%,当 $scale=37$ 时的空间节约效率达到最大值,接近 70%.

Bi-CSR 较 CSR 方法,存储效率提升明显,但是,Bi-CSR 需要增加行方向位图数组 row_bitmap 和列方向位图数组 $columns\_bitmap$,增加了额外的存储空间来获得更高的压缩比;并且 Bi-CSR 较 CSR 计算复杂度更高,需要引入紧前位 B_{CF}、紧前缺位 non_B_{CF} 对列非零元片段是否连续进行判定与索引,增加了 3%～5%的时间开销,Bi-CSR 性能开销测试验证如图 6 所示:

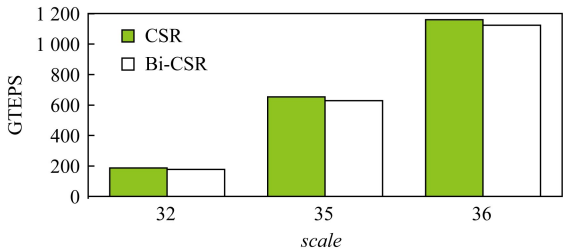


Fig. 6 GTEPS comparison between CSR and Bi-CSR with the same testing scale

图 6 相同图测试规模下的 CSR 与 Bi-CSR 性能比较

由图 6 可知,相同输入图测试规模下,CSR 的测试性能是略优于 Bi-CSR 的,实验数据表明:相同图测试规模下,Bi-CSR 约有 3%～5%的性能损失,但随着图测试规模的增大,性能损失逐渐缩小.Bi-CSR 具有更高的图稀疏矩阵的存储压缩比,相同内存下能够测试的图规模也更大,因此能够显著提升 Graph500 测试性能.因此,Bi-CSR 引入了行列位图数组和更高的计算复杂度,但是获得了更高的图存储压缩比,相同内存规模下能够完成更大的图规模测试,Graph500 测试性能更优.

4.4 Graph500 性能测试与分析

装备天河 E 级验证系统的 MT-2000＋为通用众核处理器.不同稀疏图压缩方法的性能测试比较如图 1 所示,本节的重点是关注 CSR 与 Bi-CSR 的性能比较分析,因此,对不同图规模和运行配置下对基于 CSR 的 Graph500 和基于 Bi-CSR 的 THE-G500 性能进行了详细的测试分析与比较,如图 7 所示.

由图 7 可知,在天河 E 级验证系统上,参考版本 Graph500 程序和天河 E 级验证系统版本的 THE-G500 具备在实际测试过程中($scale>15$)每结点开

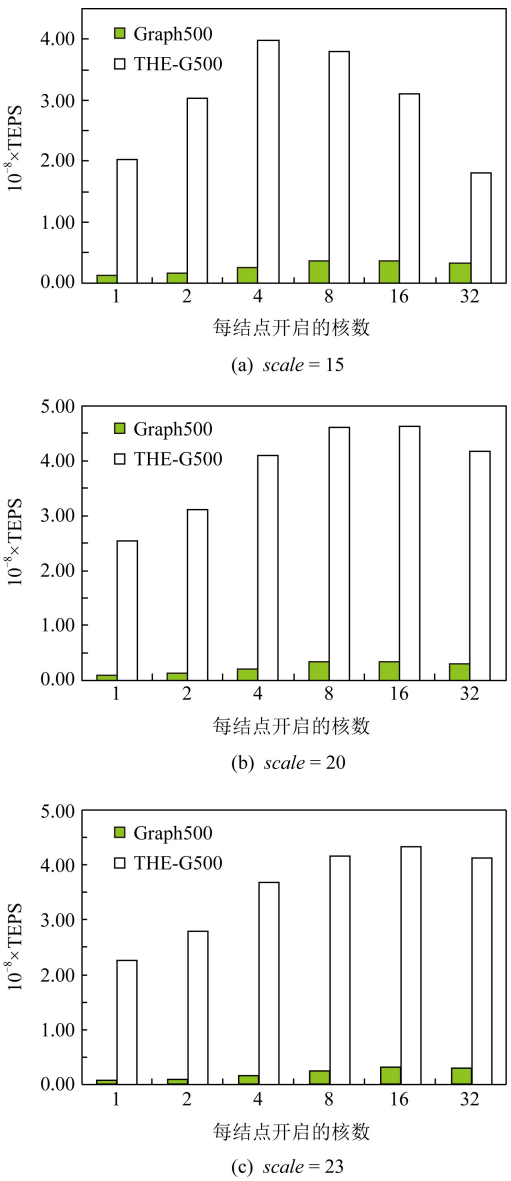


Fig. 7 TEPS comparison on different running threads/cores per node

图 7 不同线程并行时程序性能比较

启 16 线程并行时,性能整体表现最佳,并且图的测试规模越大 THE-G500 性能加速效果越明显.当测试规模为 $scale=15$,开启 2 个线程并行时,THE-G500 性能提升最大,约 19.35 倍,开启 32 个线程并行时,THE-G500 性能提升最小,约 5.49 倍,该规模下 THE-G500 平均性能提升约 12.9 倍;当图的测试规模为 $scale=20$,仅开启 1 个线程,THE-G500 性能提升最大,约 26.79 倍,开启 32 个线程并行时,THE-G500 性能提升最小,约 13.33 倍,该规模下 THE-G500 平均性能提升约 18.31 倍;当图的测试规模为 $scale=23$,开启 2 个线程并行时,THE-G500 性能提升最大,约 37.88 倍,开启 32 个线程

并行时,THE-G500 性能提升最小,约 13.98 倍,这是因为本地存储单元有限,线程增加了,可利用局部存储空间少了,数据迁移至更远的外部存储空间,该规模下 THE-G500 平均性能提升约 23.99 倍.

应用 Bi-CSR 大规模图空间压缩方法、面向 MT-2000+天河 E 级验证系统版本的 THE-G500 全系统性能测试如图 8 所示,图 8 中表现的性能均为在对应结点规模下能够稳定测试的最大图规模 *scale* 值.

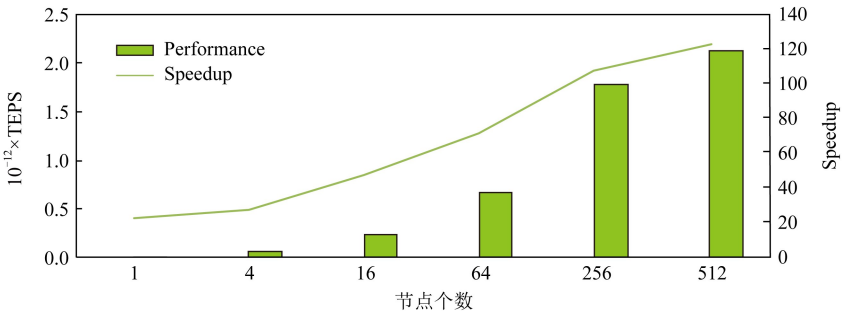


Fig. 8 Performance for THE-G500 testing on tianhe pre-exascale system

图 8 面向天河 E 级验证系统的 THE-G500 性能测试

由图 8 可知,面向 E 级验证系统版本的 THE-G500 性能较官方的参考 Graph500 测试程序性能提升显著,单结点最小加速比接近 20 倍加速比,全系统最大加速超过 100 倍,受限于系统内存,当图规模测试到 *scale*=36,参考版本 Graph500 不能稳定地完成测试任务,最可能原因是内存压力大,对网络以及系统稳定性要求高.由于 Bi-CSR 能够节约近 70% 的存储空间,THE-G500 在全系统下可以将图的测试规模成功进一步扩大到 *scale*=37,稳定测试性能为 2.131E+12TEPS,当图规模继续增大时,THE-G500 也无法完成稳定测试,最重要的原因是,压缩优化后的图存储空间仍然超过了 E 级验证系统最大存储空间.综上所述,THE-G500 较 Graph500 能够应用更大规模的图测试,并且,在相同可用存储空间下,THE-G500 较 Graph500 的最大性能加速比已超过 100 倍,THE-G500 性能加速比随着结点规模的增加具有更好的扩展性.需要说明的是,全系统 THE-G500 测试性能包含了可变宽度向量优化带来的性能提升,由于可变宽度向量优化是基于列方向位图的稀疏矩阵压缩方法的延伸,并且,可变宽度向量优化加速 BFS 遍历并不是本文关注的重点.因此,将可变宽度向量优化直接归入了基于列方向位图的稀疏矩阵压缩带来的性能提升.

面向天河 E 级验证系统,应用 Bi-CSR 方法设计实现了面向目标系统的 Graph500 测试版本 THE-G500,实验结果表明:THE-G500 能够充分发挥天河 E 级验证系统的数据密集型应用处理潜力,同时验证了面向天河 E 级验证系统的大规模图存储优化方法 Bi-CSR 的高效性.

5 总 结

E 级验证系统将面向大科学工程计算,兼顾海量信息大数据智能处理平台,大规模图处理是评测超 E 级系统处理数据密集型应用能力的重要手段.为此,针对天河 E 级验证系统小内存,基于 CSR 图结构引入行列双向位图进一步压缩稀疏矩阵存储空间,提出了大规模图数据压缩存储方法 Bi-CSR,该方法将大幅减少稀疏矩阵存储空间,最大化 Graph500 图测试规模和提升 Graph500 测试性能,在天河 E 级验证系统 512 结点系统上,全系统稳定测试性能 TEPS 为 2.131×10^{12},面向天河 E 级验证系统时,位列 Graph500 排名第 9,同等系统规模下 THE-G500 性能表现优异.

贡献声明:作者甘新标提出了基于双向位图的大规模图存储压缩方法,并完成算法设计;作者谭雯参与算法编码实现、实验测试和论文修改工作;作者刘杰参与实验设计与讨论,制定实验内容,协调实验资源.

参 考 文 献

[1] Dodds P, Muhamad R, Watts D. An experimental study of search in global social networks [J]. Science, 2003, 301(5634): 827-829

[2] Albert R, Jeong H, Barabási A. Internet: Diameter of the world-wide Web [J]. Nature, 1999, 401(6749): 130-131

[3] Watts D, Strogatz S. Collective dynamics of 'small-world' networks [J]. Nature, 1998, 393(6684): 440-442

[4] Beutel A, Faloutsos C. User behavior modeling and fraud detection [J]. IEEE Intelligent Systems, 2016, 31(2): 84-86

[5] Barabási A, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286(5439): 509-512

[6] Bader D, Madduri K. Designing multithreaded algorithms for breadth-first search and st-connectivity on the Cray MTA-2 [C] //Proc of the 2006 Int Conf on Parallel Processing (ICPP'06). Piscataway, NJ: IEEE, 2006: 523-530

[7] Agarwal V, Petrini F, Pasetto D, et al. Scalable graph exploration on multicore processors [C] //Proc of the 2010 ACM/IEEE Int Conf for High Performance Computing, Networking, Storage and Analysis. Los Alamitos, CA: IEEE Computer Society, 2010: 1-11

[8] Beamer S, Buluc A, Asanovic K, et al. Distributed memory breadth-first search revisited: Enabling bottom-up search [C] //Proc of the 27th Int Symp on Parallel and Distributed Processing Workshops and PhD Forum. Piscataway, NJ: IEEE, 2013: 1618-1627

[9] Ueno K, Suzumura T, Maruyama N, et al. Efficient breadth-first search on massively parallel and distributed-memory machines [J]. Data Science and Engineering, 2017, 2(1): 22-35

[10] Lin Heng, Tang Xiongchao, Yu Bowen, et al. Scalable graph traversal on sunway Taihulight with ten million cores [C] // Proc of the 2017 IEEE Int Parallel and Distributed Processing Symp. Piscataway, NJ: IEEE, 2017: 635-645

[11] Lin Heng, Zhu Xiaowei, Yu Bowen, et al. ShenTu: Processing multi- trillion edge graphs on millions of cores in seconds [C] //Proc of Int Conf for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ: IEEE, 2018: 706-716

[12] Sun Xiangzheng, Zhang Yunquan, Wang Ting, et al. Auto-tuning of SpMV for diagonal sparse matrices [J]. Journal of Computer Research and Development, 2013, 50(3): 648-656 (in Chinese)
(孙相征, 张云泉, 王婷, 等. 对角线稀疏矩阵的 SpMV 自适应性能优化[J]. 计算机研究与发展, 2013, 50(3): 648-656)

[13] Grossman M, Thiele C, Polo M, et al. A survey of sparse matrix-vector multiplication performance on large matrices [C/OL] //Proc of Rice Oil & Gas High Performance Computing Workshop. 2016 [2017-06-23]. https://arxiv.org/abs/1608.00636

[14] Bell N, Garland M. Implementing sparse matrix-vector multiplication on throughput-oriented processors [C] //Proc of the Conf on High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2009: 1-11

[15] Bader D, Berry J, Kahan S, et al. Graph500 Benchmarks [EB/OL]. 2010 [2017-06-20]. https://graph500.org/?page_id=12

[16] Suzumura T, Ueno K, Sato H, et al. Performance characteristics of Graph500 on large-scale distributed environment [C] //Proc of the 2011 IEEE Int Symp on Workload Characterization. Piscataway, NJ: IEEE, 2011: 149-158

[17] Checconi F, Petrini F. Traversing trillions of edges in real time: Graph exploration on large-scale parallel machines [C] //Proc of the 2014 IEEE Int Parallel and Distributed Symp. Piscataway, NJ: IEEE, 2014: 425-434

[18] Ueno K, Suzumura T. Highly scalable graph search for the Graph500 benchmark [C] //Proc of the 2012 Int on High-performance Parallel and Distributed Computing. New York: ACM, 2012: 149-160

[19] Gao Tao, Lu Yutong, Zhang Baida, et al. Using MIC to accelerate graph traversal [J]. International Journal of High Performance Computing Applications, 2014, 28(3): 255-266

[20] Jose J, Potluri S, Tomko K, et al. Designing scalable Graph500 benchmark with hybrid MPI + OpenSHMEM programming models [C] //Proc of the 2013 Int Conf on High Performance Computing, Networking, Storage and Analysis. Berlin: Springer, 2013: 109-124

Gan Xinbiao, born in 1982. PhD. Associate professor. His main research interests include supercomputer and HPC, scientific visualization and analysis.

甘新标, 1982 年生. 博士, 副研究员. 主要研究方向为超级计算机、高性能计算、科学计算可视化.

Tan Wen, born in 1996. Master candidate. Her main research interests include supercomputer, parallel computing and graph calculation.

谭 雯, 1996 年生. 硕士研究生. 主要研究方向为超级计算机、并行计算和图计算.

Liu Jie, born in 1969. Professor, PhD supervisor. His main research interests include supercomputer, CFD and artificial intelligence.

刘 杰, 1969 年生. 研究员, 博士生导师. 主要研究方向为超级计算机、计算流体与人工智能.