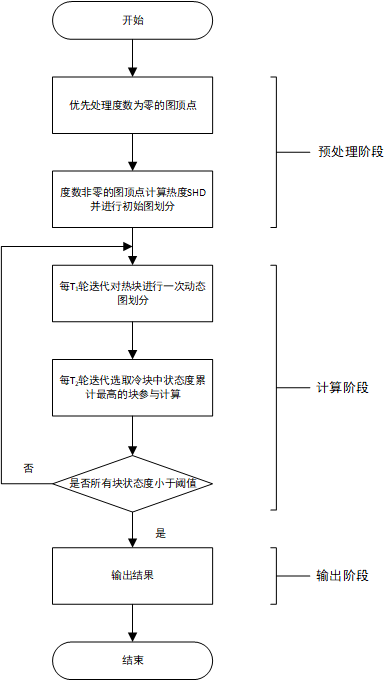
说 明 书 摘 要

本发明实现了一种用于图拓扑分析的图处理方法，该方法主要包括图数据预处理、图结构的动态划分和图处理的自适应调度。在图数据预处理阶段，根据入度和出度计算图顶点热度，并据此将相应的顶点划分到热块或冷块；在图结构动态划分阶段，根据迭代后部分顶点收敛导致的图结构的变化，重新划分热块，将活跃度较低的顶点移入冷块，降低计算频次，从而减少活跃顶点导致的I/O开销，降低缓存缺失率；在图处理的自适应调度阶段，由于各顶点收敛速度不同，累计的图顶点状态度各异，调度图顶点状态度累计高的块计算，使状态变化更大、更频繁的顶点有较高的处理优先级，加快图顶点的收敛速度，减少大数据环境下图算法运行时间。

摘 要 附 图



权 利 要 求 书

1. 一种用于图拓扑分析的图处理方法，包括以下步骤：

（1）图数据预处理。根据顶点入度和出度计算顶点热度（度数为零的顶点单独优先处理且仅处理一次）。顺序读取顶点热度，并将相应顶点划分到热块和冷块。

（2）图结构的动态划分。每T1轮迭代调用图划分函数对热块进行重新划分。若存在热块，则自适应调度冷/热块进行计算；否则，判断整张图是否收敛，收敛则结束计算并输出结果，未收敛则仅调度冷块中状态度最大的块计算。

（3）图处理的自适应调度。每轮迭代仅处理热块中状态度最大的块，每T2轮迭代调度热块和冷块中状态度最大的块进行计算。

1. 如权利要求1所述的方法，其特征在于，所述步骤（1）包括如

下子步骤：

（1.1）在首轮迭代计算热块之前，统计所有度数为零的顶点，后续计算中将不再考虑这些顶点。

（1.2）图顶点初始化。具体而言，将会创建两个图顶点度数表和两个图顶点值表。两个图顶点度数表分别点表和边表。点表用于存放所有图顶点的入度和出度，具体包括图顶点的ID、入度、出度、热度4个属性值，在实现中根据ID访问图的各顶点。边表用于存放邻居图顶点位置，包括图顶点ID和邻居顶点ID 2个属性值，根据ID访问图顶点。两个图顶点值表curr和last，分别存放图顶点本次计算和上次计算的值，用于计算顶点的状态度，从而获取到每个块的状态度。

（1.3）分区块表格初始化。具体而言，将会创建一个块状态度表。块状态度表包括块ID和块状态度2个属性值。

（1.4）对图顶点进行初始分区。根据出度和入度计算图顶点热度，并进行更新。若图顶点热度高于阈值D1，则标记为热顶点，并划分到热块；若图顶点热度低于阈值D1，则标记为冷顶点，并划分到冷块。

（1.5）判断是否所有顶点都已标记。若是，则初始化所有块的块状态表，输出初始分区；否则，返回步骤（1.3）。

1. 如权利要求2所述的方法，其特征在于，所述步骤（2）包括如

下子步骤：

（2.1）T1轮迭代后，对当前所有热块重新分区。计算当前所有热块中顶点的热度，调用图划分函数重新划分热块，并更新所有块状态度表。

（2.2）若重新分区后不再有热块，判断整张图是否收敛。计算所有块状态度，若块状态度小于阈值D2，则视为图收敛，输出计算结果；否则继续计算，但不再进行图划分。

1. 如权利要求3所述的方法，其特征在于，所述步骤（3）包括如

下子步骤：

（3.1）每轮迭代的处理过程。每轮迭代仅计算热块中块状态度最高的N个块（0<N<threads，threads为系统线程数），并更新块状态度表。

（3.2）每T2轮迭代的调度处理。选取热块中块状态度累计最高的N2个块和冷块中块状态度累计较高的N1个块进行计算，且N1+N2=N。更新块状态度表。

（3.3）若所有块状态度之和小于阈值D2，则视为图收敛，输出计算结果；否则跳转到（3.1）继续计算。

说 明 书

一种用于图拓扑分析的图处理方法

**技术领域**

本发明属于大数据处理的图计算技术领域，具体涉及一种图结构感知的图处理的方法，利用图数据结构特征优化大规模数据下的图划分和图处理。

**背景技术**

随着大数据时代的来临，越来越多大数据应用需要以图的形式描述数据，进而通过迭代方式对其加以处理。例如，搜索服务(如Google)采用PageRank算法进行结果排序，社交网络(如Facebook)使用Clustering算法分析用户群落，知识共享网站(如Wikipedia)使用Named Entity Recognition算法识别文本信息，视频网站(如Netflix)基于Collaborative Filtering算法提供影视推荐等。

相关研究指出，图计算应用在计算和存储上的特性造成面向数据的并行编程模型无法提供高效的支持，对数据间关联性的描述手段的缺乏以及对迭代计算的低效支持可能造成数倍乃至数十倍的性能损失。对高效的大规模图计算(Graph Computation)系统的迫切需求促使其成为并行与分布式处理领域急需解决的重要问题之一。然而大多数图处理方法将底层图数据看作黑盒，缺少基于图结构的动态图划分和图处理研究。而真实世界中图结构是不断变化的，经过迭代，图划分块中的大量图顶点可能会收敛。少数活跃顶点的频繁访问可能会导致整个图划分块包括收敛的反复加载，但这些收敛顶点是不需要访问和处理的，导致内存带宽和缓存的严重浪费。现有方法未考虑各划分块的结构特征，图算法收敛所需更新次数多，每次更新所需开销大。

基于上述原因，本发明提出了一种用于图拓扑分析的图处理方法，根据图结构的变化自适应的提供图划分和图处理调度方案，降低缓存缺失率和I/O资源开销，同时加速图顶点收敛。

**发明内容**

针对现有的条件，本发明提出一种用于图拓扑分析的图处理方法，解决了缓存缺失率高、图顶点收敛慢的问题，提升了系统的性能。

为实现上述目的，本发明包括以下步骤：

（1）图数据预处理。根据顶点入度和出度计算顶点热度（度数为零的顶点单独优先处理且仅处理一次）。顺序读取顶点热度，并将相应顶点划分到热块和冷块。

（2）图结构的动态划分。每T1轮迭代调用图划分函数对热块进行重新划分。若存在热块，则自适应调度冷/热块进行计算；否则，判断整张图是否收敛，收敛则结束计算并输出结果，未收敛则仅调度冷块中状态度最大的块计算。

（3）图处理的自适应调度。每轮迭代仅处理热块中状态度最大的块，每T2轮迭代调度热块和冷块中状态度最大的块进行计算。

本发明中，图顶点热度是顶点度数的函数，热度高的图顶点更活跃，决定整张图收敛所需的迭代次数，是整张图能否快速收敛的关键因素；热度低的图顶点收敛条件较为苛刻，因而处理优先级低。本专利的创新点之一是划分冷热顶点以减少低活跃度顶点频繁加载带来的缓存缺失，而图顶点热度是图划分的重要依据，因此图顶点热度对保证缓存命中即内存集中访问有着不可忽视的影响。

图顶点热度函数：SHD(Vi) = (1-a)\* D(Vo) - a \*D(Vi) （a<0.5），其中D(Vi)和D(Vo)分别为图顶点的入度和出度。入度大的顶点易受邻居顶点的影响，收敛所需条件更苛刻，故先让其邻居顶点收敛一部分之后再计算，因此顶点热度与热度是负相关；出度大的顶点会影响更多的邻居顶点，所以需要先执行，加快高出度顶点收敛，故顶点出度与热度是正相关。

热顶点和冷顶点分别指图顶点热度高于和低于阈值D1的图顶点。

阈值D1用于区分冷热顶点。取值由用户定义的样本大小和热顶点比例共同决定，例如：顶点个数为V = 10000的图数据，用户定义样本大小为 ，热顶点比例为，则阈值 ，即样本中第10个顶点的SHD。

阈值D2用于判断整张图是否收敛。其值由用户自定义，默认为0.000001。

热块和冷块分别指存放热顶点和冷顶点的cache块。

图顶点状态度指本次和上次计算之间，顶点算法执行结果之差的累计，对于PageRank算法，；对于SSSP算法，；对于CC算法，。图顶点状态度是图处理调度的重要依据。

步骤（2）中图结构的动态划分和步骤（3）中块的调度均需要根据热度或状态度对顶点降序排序，当数据集规模庞大时，传统基于CPU的方法会大大增加时间复杂度，降低系统性能。因此本发明提出一种异构内存图处理方案，CPU和GPU共享同一个存储空间，并且CPU能够直接存取GPU的存储地址，将顶点根据热度和状态度降序排序的过程交由GPU完成，缩短计算时间。

上述用于图拓扑分析的图处理方法，实施步骤如下：

（1）预处理与初始图划分。其主要包括如下子步骤：

（1.1）在首轮迭代计算热块之前，统计所有度数为零的顶点，后续计算中将不再考虑这些顶点。

（1.2）图顶点初始化。具体而言，将会创建两个图顶点度数表和两个图顶点值表。两个图顶点度数表分别点表和边表。点表用于存放所有图顶点的入度和出度，具体包括图顶点的ID、入度、出度、热度4个属性值，在实现中根据ID访问图的各顶点。边表用于存放邻居图顶点位置，包括图顶点ID和邻居顶点ID 2个属性值，根据ID访问图顶点。两个图顶点值表curr和last，分别存放图顶点本次计算和上次计算的值，用于计算顶点的状态度，从而获取到每个块的状态度。

（1.3）分区块表格初始化。具体而言，将会创建一个块状态度表。块状态度表包括块ID和块状态度2个属性值。

（1.4）对图顶点进行初始分区。根据出度和入度计算图顶点热度，并进行更新。若图顶点热度高于阈值D1，则标记为热顶点，并划分到热块；若图顶点热度低于阈值D1，则标记为冷顶点，并划分到冷块。

（1.5）判断是否所有顶点都已标记。若是，则初始化所有块的块状态表，输出初始分区；否则，返回步骤（1.3）。

（2）图结构的动态划分。其主要包括如下子步骤：

（2.1）T1轮迭代后，对当前所有热块重新分区。计算当前所有热块中顶点的热度，调用图划分函数重新划分热块，并更新所有块状态度表。

（2.2）若重新分区后不再有热块，判断整张图是否收敛。计算所有块状态度，若块状态度小于阈值D2，则视为图收敛，输出计算结果；否则继续计算，但不再进行图划分。

（3）图处理的自适应调度。其主要包括如下子步骤：

（3.1）每轮迭代的处理过程。每轮迭代仅计算热块中块状态度最高的N个块（0<N<threads，threads为系统线程数），并更新块状态度表。

（3.2）每T2轮迭代的调度处理。选取热块中块状态度累计最高的N2个块和冷块中块状态度累计较高的N1个块进行计算，且N1+N2=N。更新块状态度表。

（3.3）若所有块状态度之和小于阈值D2，则视为图收敛，输出计算结果；否则跳转到（3.1）继续计算。

总体而言，通过本发明所构思的以上技术方案与现有技术相比，能够取得下列有益效果：

1. 预处理开销小：由于本发明所提供的预处理方法将度数为零

的顶点优先处理，使其单次迭代就能达到收敛，仅需对图进行一次遍历，复杂度为线性，远小于对图进行迭代直至收敛的复杂度。

1. 加快图收敛：本发明所提供的图划分方法区别了热/冷顶点，

使状态变化更大更频繁的热顶点有较高的优先级，更快达到收敛，进而减少了整张图达到收敛所需的平均更新次数。

1. 减少计算冗余：本发明所提供的图划分方法经过特定次数的

迭代后，对热顶点个数大幅下降的分区进行一次重新划分。相比较而言，本发明所提供的图划分方法既考虑了图结构变化造成的负载不均问题，又很好地控制了动态分区时图顶点移动产生的计算开销。

1. 降低I/O资源开销：本发明所提供的图划分方法将活跃度高，

更新频繁的图顶点放在同一cache块中，一同加载进内存。与之前的方法相比，减少不活跃顶点的加载，极大地降低了活跃顶点导致的I/O开销。

**附图说明**

图1为本发明预处理的流程图。

图2为本发明图处理方法的流程图。

图3为本发明软硬件交互图。

**具体实施方式**

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本方面，并不用于限定本发明。此外，下面所描述的本发明的各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

本发明提供的用于图拓扑分析的图处理方法，其主要思想为：每轮迭代只处理活跃度最高的部分顶点，并根据图顶点收敛特性在适当时进行重新分区。该图划分方法用每轮迭代中很小的图顶点热度计算开销，大大加快了图收敛的速度，减少了计算冗余，降低了缓存缺失率。

实施例所提供的用于图拓扑分析的图处理方法，其流程如图1和图2所示，包括图数据预处理阶段，动态图划分阶段和自适应调度图处理阶段。具体如下：

（1）图数据预处理。根据顶点入度和出度计算顶点热度（度数为零的顶点单独优先处理且仅处理一次）。顺序读取顶点热度，并将相应顶点划分到热块和冷块。包括如下子步骤：

（1.1）在首轮迭代计算热块之前，统计所有度数为零的顶点，后续计算中将不再考虑这些顶点。

（1.2）图顶点初始化。具体而言，将会创建两个图顶点度数表和两个图顶点值表。两个图顶点度数表分别点表和边表。点表用于存放所有图顶点的入度和出度，具体包括图顶点的ID、入度、出度、热度4个属性值，在实现中根据ID访问图的各顶点。边表用于存放邻居图顶点位置，包括图顶点ID和邻居顶点ID 2个属性值，根据ID访问图顶点。两个图顶点值表curr和last，分别存放图顶点本次计算和上次计算的值，用于计算顶点的状态度，从而获取到每个块的状态度。

（1.3）分区块表格初始化。具体而言，将会创建一个块状态度表。块状态度表包括块ID和块状态度2个属性值。

（1.4）对图顶点进行初始分区。根据出度和入度计算图顶点热度，并进行更新。若图顶点热度高于阈值D1，则标记为热顶点，并划分到热块；若图顶点热度低于阈值D1，则标记为冷顶点，并划分到冷块。

（1.5）判断是否所有顶点都已标记。若是，则初始化所有块的块状态表，输出初始分区；否则，返回步骤（1.3）。

（2）图结构的动态划分。每T1次迭代调用图划分函数对热块进行重新划分。若无热块，则判断整张图是否收敛，收敛则结束计算输出结果，未收敛则仅调度冷块计算；若有，则自适应调度冷/热块计算。包括如下子步骤：

（2.1）T1轮迭代后，对当前所有热块重新分区。计算当前所有热块中顶点的热度，调用图划分函数重新划分热块，并更新所有块状态度表。

（2.2）若重新分区后不再有热块，判断整张图是否收敛。计算所有块状态度，若块状态度小于阈值D2，则视为图收敛，输出计算结果；否则继续计算，但不再进行图划分。

（3）图处理的自适应调度：每轮迭代只处理热块中状态度最大的块，每T2次迭代调度冷块和热块中状态度最大的块进行计算。包括如下子步骤：

（3.1）每轮迭代的处理过程。每轮迭代仅计算热块中块状态度最高的N个块（0<N<threads，threads为系统线程数），并更新块状态度表。

（3.2）每T2轮迭代的调度处理。选取热块中块状态度累计最高的N2个块和冷块中块状态度累计较高的N1个块进行计算，且N1+N2=N。更新块状态度表。

（3.3）若所有块状态度之和小于阈值D2，则视为图收敛，输出计算结果；否则跳转到（3.1）继续计算。

本专利对于现有图算法均表现出良好的性能，一下 以PageRank算法为例，详细描述本发明的实施过程：将一个页面看做一个顶点，页面之间的相互链接用有向图中定点的出度和入度表示。本发明提出的用于图拓扑分析的图处理方法，能够在预处理阶段筛选出没有任何链接的顶点，减少后续计算量；同时对于需要继续分析计算的顶点进行特征提取，根据顶点度数计算每个页面的热度函数，并在图划分阶段根据热度阈值D1将页面顶点划分到不同图划分块中，热块与冷块分开存储，本发明采用异构存储，CPU和GPU共享同一个存储空间，并且CPU能够直接存取GPU的存储地址，因此我们将对页面热度排序在GPU上计算，降低图划分步骤的耗时又不会增加额外的通信开销；当部分页面PageRank值计算完毕后，本发明所提供的图划分方法将对热顶点个数大幅下降的分区进行一次重新划分。相比较而言，本发明所提供的图划分方法既考虑了图结构变化造成的负载不均问题，又很好地控制了动态分区时图顶点移动产生的计算开销；在计算顶点PageRank值时选择状态度较高的顶点，即仅从内存加 载状态度更高的热块，使状态变化更大更频繁的热顶点有较高的计算优先级，更快达到收敛，进而减少了整张图达到收敛所需的平均更新次数。

本领域的技术人员容易理解，以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作任何修改，等同替换和改进等，均应包含在本发明的保护范围之内。

**说 明 书 附 图**

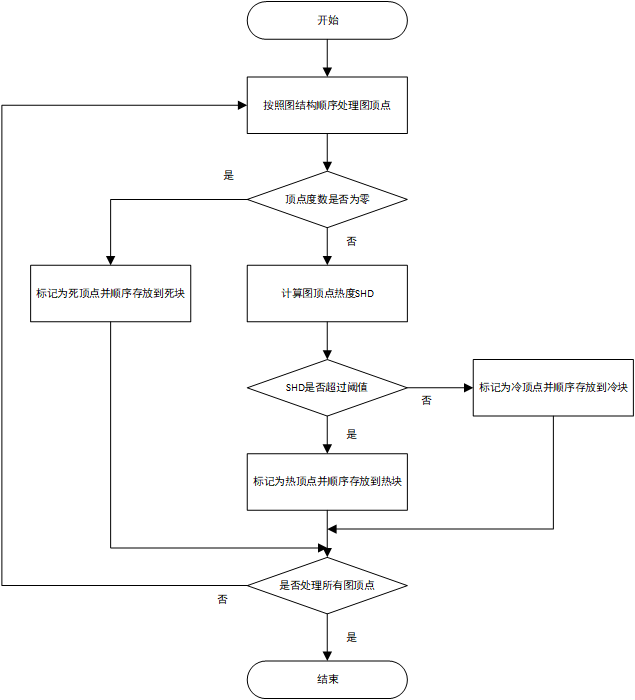


图1

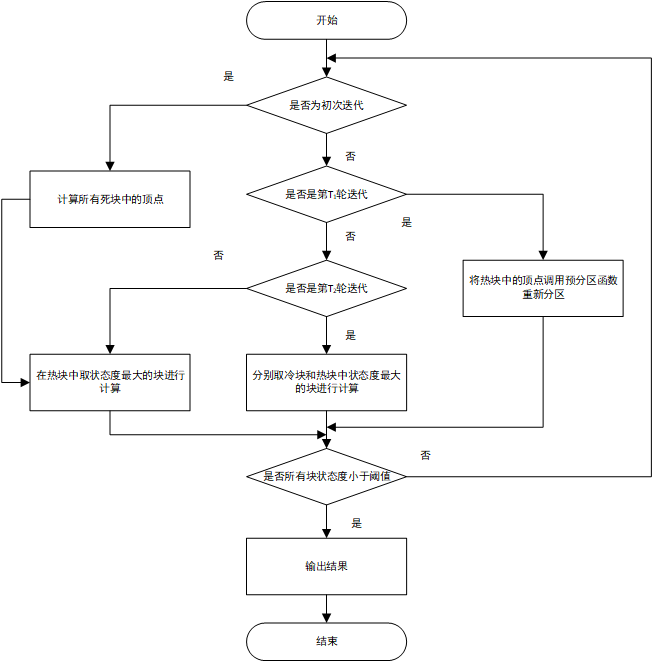


图2

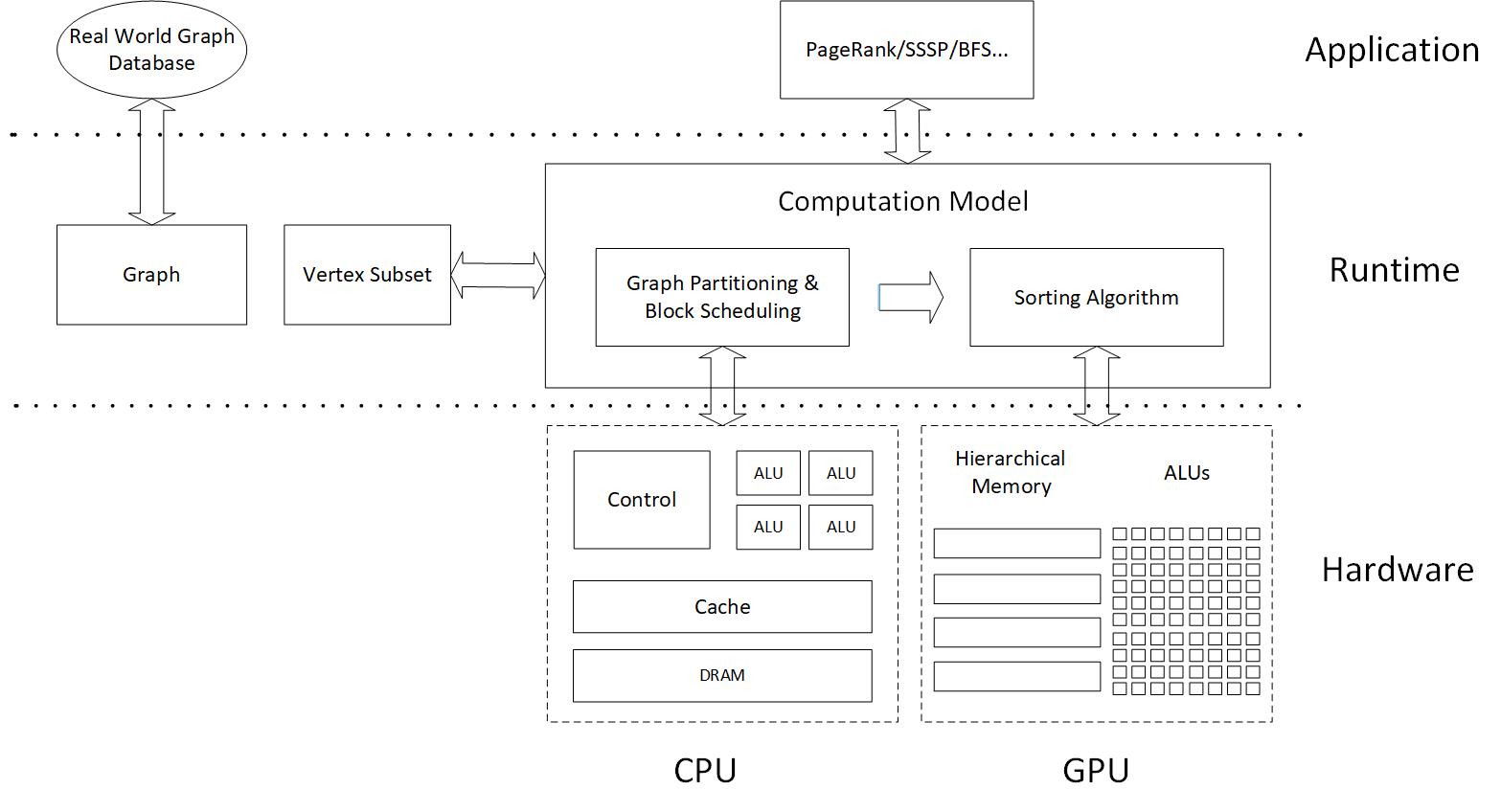
****

图3