**专利申请明细表**

**兹全权委托北京博思佳知识产权代理有限公司办理下列申请专利有关事项**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 发 明 名 称 | | 一种冗余感知的超图神经网络软硬件协同设计方法 | | |
| 第一  申请人 | 姓名或名 称 | 之江实验室 | | |
| 地址、邮编 | 默认：浙江省杭州市余杭区之江实验室南湖总部，311121  日本专利：之江实验室 中国310023浙江省杭州市余杭区文一西路1818号 | | |
| 企业机构代码或身份证号 | 12330000MB1478604D | | |
| 第二  申请人  （如果为多个申请单位共同申请，可填） | 姓名或名 称 | 华中科技大学 | | |
| 地址、邮编 | 地址：湖北省武汉市洪山区珞喻路1037号，430074 | | |
| 企业机构代码或身份证号 | 12100000441626842D | | |
| 发明人  或设计人 | 姓 名 | 张宇，沈千格，宁鑫，赵进，余辉，张湛 | 第一发明人身份证号 | 431122198707170515 |
| 专利负责人 | 姓名 | 吴梅英，18662590857，wumeiying@zhejianglab.com  地址：浙江省杭州市余杭区之江实验室南湖总部主楼1402办公室，311100 | | |
| 地址、邮编 |
| 电话/E-mail |
| 技术（撰写）联系人 | 姓名 | 沈千格，18155600561，qgshen@hust.edu.cn  地址：浙江省杭州市余杭区之江实验室南湖总部11号楼B座B102室，311121 | | |
| 地址、邮编 |
| 电话/E-mail |
| 申请专利类型 | | 🗹发明　　　　　实用新型　　　　外观设计 | | |
| 催缴费联系人电话、Email、地址 | | 吴梅英，18662590857，wumeiying@zhejianglab.com，地址：浙江省杭州市余杭区之江实验室南湖总部主楼1402办公室，311100 | | |

注：

1、红色字体部分需要填写或选择

说 明 书 摘 要

本发明公开了一种冗余感知的超图神经网络软硬件协同设计方法，该方法主要包括一种新颖的硬件加速器设计和一种冗余感知的超图神经网络软件设计。该硬件加速器设计，主要包括协调器、聚合器、更新器、中央控制器和一个缓冲管理单元。该加速器能够将不规则的聚合操作组装成冗余感知的规则计算，从而减少冗余计算，提高计算效率。该超图神经网络软件设计，主要包括图数据预处理和冗余感知的超图神经网络处理两个部分。在图数据预处理阶段，将根据原始超图结构预处理得到冗余感知的超图结构。在冗余感知的超图神经网络处理阶段，首先遍历顶点并根据状态信息动态调度以分发计算任务，然后计算超边上的特征，再聚合超边特征得到顶点特征，最后更新顶点的特征。

摘要附图

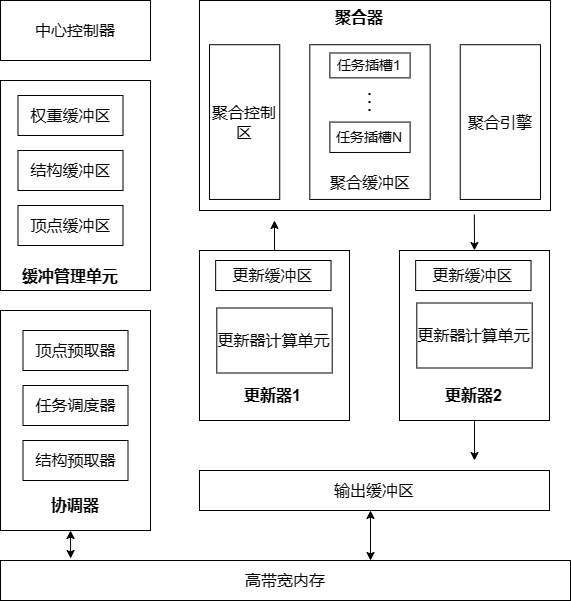


图1 硬件加速器架构

权 利 要 求 书

1、一种冗余感知的超图神经网络软硬件协同设计方法中的可编程硬件加速器设计，包括以下五个部分：

（1）协调器。协调器用来识别冗余计算，能够将不规则的聚合操作组装成冗余感知的规则计算，从而达到利用高度并行的计算引擎的目的。具体来说，没有计算超边特征的超边被发送到更新器，用来更新超边特征；而对于聚合任务来说，只有未聚合的目标顶点和冗余聚合集中的任务才会发送到聚合器执行。协调器包括结构预取器、任务调度器和顶点预取器，其中结构预取器旨在从高带宽存储器中预取图结构和状态信息；任务调度器旨在调度不规则计算，通过遍历原始顶点和冗余聚合集中的顶点并根据状态信息将任务分配给聚合器和更新器；顶点预取器用于将顶点属性预取到相应的缓冲区以供后续执行。

（2）聚合器。聚合器用来执行目标顶点和冗余聚合集中的聚合任务，同时消除冗余计算。它包括聚合控制器、聚合缓冲区和一个聚合引擎。聚合控制器负责聚合任务的调度，它使聚合器能够复用原始顶点特征和聚合的中间结果，从而减少通信冗余和计算冗余；聚合控制器将不规则的聚合任务组装成冗余感知的聚合任务，并将任务发送给聚合缓冲区；聚合引擎获取聚合缓冲区中的聚合任务并执行相应任务，并且可以将冗余聚合集中的中间结果存储到缓存中，然后这些中间结果可以被后续的聚合任务复用，从而减少冗余计算。

（3）更新器。本专利设计了两种类型的更新器，以支持不同的超图神经网络算法。更新器1用来更新超边特征，并且能够缓存中间结果以用于后续的聚合任务；更新器2用来更新顶点特征，主要用于处理矩阵乘法运算。

（4）中央控制器。中央控制器用于控制超图神经网络的执行流，以更好地配置流水线执行流，并支持各种超图神经网络算法。

（5）缓冲管理单元。本专利设计使用片上缓冲区来缓存各种数据和中间结果，以提高数据重用性。除了更新器缓冲区和聚合器缓冲区之外，还有权重缓冲区、结构缓冲区和中间缓存。权重缓冲器用于存储算法的权重参数；结构缓冲器用于存储超图结构和状态信息；中间缓存用于存储聚合器和更新器生成的中间结果。为达到隐藏访问延迟的目的，本专利采用了双缓冲技术，允许对所有缓冲区重复执行不同的操作。

2、一种冗余感知的超图神经网络软硬件协同设计方法中的超图神经网络软件设计，包括以下步骤：

（1）图数据预处理。根据原始超图结构预处理得到冗余感知的超图结构。

（2）冗余感知的超图神经网络处理。首先遍历顶点并根据状态信息动态调度以分发计算任务，然后计算超边上的特征，再聚合超边特征得到顶点特征，最后更新顶点的特征。

3、如权利要求2所述的方法，其特征在于，所述步骤（1）包括如下子步骤：

（1.1）遍历图顶点和对应的超边邻居，构造一个大顶堆，该大顶堆的数据项表示为，其中和表示顶点，表示和之间的公共超边数量。

（1.2）构造冗余聚合集，冗余聚合集表示执行冗余聚合的超边集合。首先弹出堆顶元素，当大顶堆堆顶的公共超边数量大于阈值时，将对应的公共超边添加到冗余聚合集中，重复执行该步骤直到堆为空。

（1.3）构造冗余感知的超图结构，冗余感知的超图结构使用关联矩阵和冗余聚合集进行表示，边可以分为两种类型：原始超边和冗余聚合集。将冗余聚合集视为一条独特的超边，因为它是多条超边的集合。为了区分超边和冗余聚合集，每个冗余聚合集和每条原始超边都与ID和ID类型相关联，ID类型使用位表示，用来区分冗余聚合集和原始超边。为了区分超边特征更新和聚合操作，每条超边和冗余聚合集都与一个状态位关联。如果是超边，则表示是否执行超边特征更新，否则表示冗余聚合集是否执行聚合操作。

4、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）硬件加速器初始化。通过调用加速器配置接口将图数据信息和超图神经网络模型信息传递到加速器以初始化加速器。默认传递的信息包括：冗余感知的超图结构、超图顶点特征和超边特征信息、模型权重参数。其中，冗余感知的超图结构包括关联矩阵和冗余聚合集

（2.2）动态调度。动态调度阶段旨在通过状态信息来遍历顶点对应的冗余聚合集和原始超边来调度超边更新工作和聚合工作，该阶段需要利用加速器中的协调器完成调度。执行步骤如下：

（2.2.1）利用协调器中的结构预取器将权重提取到权重缓冲区，将超图结构信息提取到结构缓冲区。

（2.2.2）利用任务调度器执行任务分派工作，首先获取目标顶点，遍历目标顶点对应的原始超边和冗余聚合集。如果原始超边没有完成更新操作，则将其添加到列表heList中，用于后续的超边更新；如果原始超边已经完成更新操作，则后续的聚合只需要从缓存中获取中间结果即可。对于冗余聚合集中的每条超边，如果没有完成更新操作，则将其添加到列表heList中，用于后续的超边更新；如果已经完成更新操作，则后续的聚合只需要从缓存中获取中间结果即可。对于冗余聚合集的聚合操作，如果冗余聚合集已经完成聚合操作，则后续的聚合只需要从缓存中获取冗余聚合集的中间结果即可，反之将冗余聚合集添加到aList，用于后续的聚合任务即可。对于目标顶点的聚合操作，将对应超边和冗余聚合集添加到aList。

（2.2.3）利用顶点预取器从高带宽内存中预取顶点属性并提取到顶点缓冲区。

（2.3）超边特征更新。遍历heList中的所有超边，将顶点特征从顶点缓冲区传输到更新器1中的缓冲区，然后更新器1执行具体计算，然后超边的特征，并更新对应状态信息以防止重复计算该超边的特征，然后将计算得到的结果缓存起来，以便后续的超边特征更新可以直接复用中间结果，而不需要重新计算。

（2.4）聚合超边特征。该阶段旨在聚合超边的特征并消除冗余计算，步骤如下：

（2.4.1）聚合控制器通过遍历aList将聚合任务分配到聚合缓冲区，aList包含一组需要执行聚合的超边。

（2.4.2）聚合缓冲区中的任务根据依赖关系并发调度。

（2.4.3）聚合引擎执行具体的聚合任务，它对一组超边的特征执行聚合操作并更新对应状态信息以防止重复执行聚合，然后将聚合结果缓存到顶点缓冲区作为中间结果，如果后续也需要聚合这一组超边，则可以直接复用中间结果，从而避免了冗余聚合操作。

（2.5）顶点特征更新。将每个目标顶点的旧顶点特征和聚合的特征相结合，以基于顶点权重生成新的顶点特征。如果还存在没有更新的顶点，则获取新的目标顶点进入步骤（2.2）；反之完成所有顶点的更新操作。

说 明 书

一种冗余感知的超图神经网络软硬件协同设计

**技术领域**

本发明涉及图神经网络中的超图神经网络技术领域，具体涉及一种冗余感知的超图神经网络软硬件协同设计。

**背景技术**

近年来，图神经网络在各种基于图的任务中取得了优异的性能，比如知识图谱、推荐系统等。然而，现实世界中的图往往是多个顶点之间构成相互关系，而不是简单的成对关系。对于这种复杂图，需要使用超图建模才能避免信息损失。为了处理这种复杂图，超图神经网络被广泛研究，比如利用超图进行蛋白质或细胞网络的多元关系建模、利用超图进行社交网络分析等。

超图神经网络的计算可以抽象为三个重要阶段：计算超边特征、聚合超边特征、更新顶点特征。前面两个阶段涉及到传统的图处理，严重依赖于固有的稀疏图结构，而第三个阶段主要涉及矩阵运算，是典型的神经网络操作。这种混合计算模式包括大量规则运算和不规则运算，会导致CPU和GPU产生严重的效率问题。专用的硬件加速器是提高超图神经网络计算效率的有效方法。然而现有的加速器只支持处理基于简单图的图神经网络，而无法直接处理超图。另外，这些加速器独立地对每个顶点进行邻域聚合，这会造成大量的冗余计算。比如，当一些目标顶点共享一些相同的邻居时，在聚合时，就会对这些公共邻居重复执行聚合操作，从而降低了计算效率。

**发明内容**

针对现有解决方案的不足，本发明提出一种冗余感知的超图神经网络软硬件协同设计方法，其目的在于消除超图建模下的神经网络混合计算模式中，大量规则运算和不规则运算对CPU和GPU产生严重的负面影响，以及解决专用硬件加速器在计算过程中产生的大量冗余计算降低计算效率的问题。

为实现上述目的，本发明包括一种冗余感知的超图神经网络软硬件协同编程硬件加速器设计方法和一种超图神经网络软件设计方法，其中编程硬件加速器设计包括下五个部分：

（1）协调器。协调器用来识别冗余计算，能够将不规则的聚合操作组装成冗余感知的规则计算，从而达到利用高度并行的计算引擎的目的。具体来说，没有计算超边特征的超边被发送到更新器，用来更新超边特征；而对于聚合任务来说，只有未聚合的目标顶点和冗余聚合集中的任务才会发送到聚合器执行。协调器包括结构预取器、任务调度器和顶点预取器，其中结构预取器旨在从高带宽存储器中预取图结构和状态信息；任务调度器旨在调度不规则计算，通过遍历原始顶点和冗余聚合集中的顶点并根据状态信息将任务分配给聚合器和更新器；顶点预取器用于将顶点属性预取到相应的缓冲区以供后续执行。

（2）聚合器。聚合器用来执行目标顶点和冗余聚合集中的聚合任务，同时消除冗余计算。它包括聚合控制器、聚合缓冲区和一个聚合引擎。聚合控制器负责聚合任务的调度，它使聚合器能够复用原始顶点特征和聚合的中间结果，从而减少通信冗余和计算冗余；聚合控制器将不规则的聚合任务组装成冗余感知的聚合任务，并将任务发送给聚合缓冲区；聚合引擎获取聚合缓冲区中的聚合任务并执行相应任务，并且可以将冗余聚合集中的中间结果存储到缓存中，然后这些中间结果可以被后续的聚合任务复用，从而减少冗余计算。

（3）更新器。本专利设计了两种类型的更新器，以支持不同的超图神经网络算法。更新器1用来更新超边特征，并且能够缓存中间结果以用于后续的聚合任务；更新器2用来更新顶点特征，主要用于处理矩阵乘法运算。

（4）中央控制器。中央控制器用于控制超图神经网络的执行流，以更好地配置流水线执行流，并支持各种超图神经网络算法。

（5）缓冲管理单元。本专利设计使用片上缓冲区来缓存各种数据和中间结果，以提高数据重用性。除了更新器缓冲区和聚合器缓冲区之外，还有权重缓冲区、结构缓冲区和中间缓存。权重缓冲器用于存储算法的权重参数；结构缓冲器用于存储超图结构和状态信息；中间缓存用于存储聚合器和更新器生成的中间结果。为达到隐藏访问延迟的目的，本专利采用了双缓冲技术，允许对所有缓冲区重复执行不同的操作。

超图神经网络软件设计方法在编程硬件加速器的基础上实现，包括以下步骤：

（1）图数据预处理。根据原始超图结构预处理得到冗余感知的超图结构。

（2）冗余感知的超图神经网络处理。首先遍历顶点并根据状态信息动态调度以分发计算任务，然后计算超边上的特征，再聚合超边特征得到顶点特征，最后更新顶点的特征。

本发明中，大顶堆的数据项表示为，其中和表示顶点，表示和之间的公共超边数量；构造冗余聚合集，大顶堆堆顶的公共超边数量大于阈值时，将对应的公共超边添加到冗余聚合集中；构造冗余感知的超图结构，为区分超边和冗余聚合集， ID类型使用位表示，为区分超边特征更新和聚合操作，设置一个状态位关联；如果是超边，则表示是否执行超边特征更新，否则表示冗余聚合集是否执行聚合操作。

超图神经网络软件设计方法对步骤（1）图数据预处理包括如下子步骤：

（1.1）遍历图顶点和对应的超边邻居，构造一个大顶堆，该大顶堆的数据项表示为，其中和表示顶点，表示和之间的公共超边数量。

（1.2）构造冗余聚合集，冗余聚合集表示执行冗余聚合的超边集合。首先弹出堆顶元素，当大顶堆堆顶的公共超边数量大于阈值时，将对应的公共超边添加到冗余聚合集中，重复执行该步骤直到堆为空。

（1.3）构造冗余感知的超图结构，冗余感知的超图结构使用关联矩阵和冗余聚合集进行表示，边可以分为两种类型：原始超边和冗余聚合集。将冗余聚合集视为一条独特的超边，因为它是多条超边的集合。为了区分超边和冗余聚合集，每个冗余聚合集和每条原始超边都与ID和ID类型相关联，ID类型使用位表示，用来区分冗余聚合集和原始超边。为了区分超边特征更新和聚合操作，每条超边和冗余聚合集都与一个状态位关联。如果是超边，则表示是否执行超边特征更新，否则表示冗余聚合集是否执行聚合操作。

超图神经网络软件设计方法对步骤（2）冗余感知的超图神经网络处理包括如下子步骤：

（2.1）硬件加速器初始化。通过调用加速器配置接口将图数据信息和超图神经网络模型信息传递到加速器以初始化加速器。默认传递的信息包括：冗余感知的超图结构、超图顶点特征和超边特征信息、模型权重参数。其中，冗余感知的超图结构包括关联矩阵和冗余聚合集

（2.2）动态调度。动态调度阶段旨在通过状态信息来遍历顶点对应的冗余聚合集和原始超边来调度超边更新工作和聚合工作，该阶段需要利用加速器中的协调器完成调度。执行步骤如下：

（2.2.1）利用协调器中的结构预取器将权重提取到权重缓冲区，将超图结构信息提取到结构缓冲区。

（2.2.2）利用任务调度器执行任务分派工作，首先获取目标顶点，遍历目标顶点对应的原始超边和冗余聚合集。如果原始超边没有完成更新操作，则将其添加到列表heList中，用于后续的超边更新；如果原始超边已经完成更新操作，则后续的聚合只需要从缓存中获取中间结果即可。对于冗余聚合集中的每条超边，如果没有完成更新操作，则将其添加到列表heList中，用于后续的超边更新；如果已经完成更新操作，则后续的聚合只需要从缓存中获取中间结果即可。对于冗余聚合集的聚合操作，如果冗余聚合集已经完成聚合操作，则后续的聚合只需要从缓存中获取冗余聚合集的中间结果即可，反之将冗余聚合集添加到aList，用于后续的聚合任务即可。对于目标顶点的聚合操作，将对应超边和冗余聚合集添加到aList。

（2.2.3）利用顶点预取器从高带宽内存中预取顶点属性并提取到顶点缓冲区。

（2.3）超边特征更新。遍历heList中的所有超边，将顶点特征从顶点缓冲区传输到更新器1中的缓冲区，然后更新器1执行具体计算，然后超边的特征，并更新对应状态信息以防止重复计算该超边的特征，然后将计算得到的结果缓存起来，以便后续的超边特征更新可以直接复用中间结果，而不需要重新计算。

（2.4）聚合超边特征。该阶段旨在聚合超边的特征并消除冗余计算，步骤如下：

（2.4.1）聚合控制器通过遍历aList将聚合任务分配到聚合缓冲区，aList包含一组需要执行聚合的超边。

（2.4.2）聚合缓冲区中的任务根据依赖关系并发调度。

（2.4.3）聚合引擎执行具体的聚合任务，它对一组超边的特征执行聚合操作并更新对应状态信息以防止重复执行聚合，然后将聚合结果缓存到顶点缓冲区作为中间结果，如果后续也需要聚合这一组超边，则可以直接复用中间结果，从而避免了冗余聚合操作。

（2.5）顶点特征更新。将每个目标顶点的旧顶点特征和聚合的特征相结合，以基于顶点权重生成新的顶点特征。如果还存在没有更新的顶点，则获取新的目标顶点进入步骤（2.2）；反之完成所有顶点的更新操作。

**附图说明**

图2为本发明硬件架构。

图3为本发明超图神经网络软件设计方法的流程图。

**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。

本发明提出的冗余感知的超图神经网络软硬件协同设计方法，其主要思想为：首先通过预处理得到冗余感知的超图结构，然后基于冗余感知的超图结构，识别冗余聚合集和中间结果，来消除消息传递过程中的边更新冗余和聚合冗余，此外，为解决此过程中边更新和聚合计算的不规则性，设计了特定的硬件结构支持冗余感知的超图神经网络处理算法，进一步缓解计算不规则性，优化内存模式，减少通信开销，提高并行性，降低能耗。

实例所提供的冗余感知的超图神经网络软硬件协同设计方法，计算流程如图4所示，包括图数据预处理和冗余感知的超图神经网络处理阶段，具体如下：

（1）图数据预处理。根据原始超图结构预处理得到冗余感知的超图结构。

（1.1）遍历图顶点和对应的超边邻居，构造一个大顶堆，该大顶堆的数据项表示为，其中和表示顶点，表示和之间的公共超边数量。

（1.2）构造冗余聚合集，冗余聚合集表示执行冗余聚合的超边集合。首先弹出堆顶元素，当大顶堆堆顶的公共超边数量大于阈值时，将对应的公共超边添加到冗余聚合集中，重复执行该步骤直到堆为空。

（1.3）构造冗余感知的超图结构，冗余感知的超图结构使用关联矩阵和冗余聚合集进行表示，边可以分为两种类型：原始超边和冗余聚合集。将冗余聚合集视为一条独特的超边，因为它是多条超边的集合。为了区分超边和冗余聚合集，每个冗余聚合集和每条原始超边都与ID和ID类型相关联，ID类型使用位表示，用来区分冗余聚合集和原始超边。为了区分超边特征更新和聚合操作，每条超边和冗余聚合集都与一个状态位关联。如果是超边，则表示是否执行超边特征更新，否则表示冗余聚合集是否执行聚合操作。

（2）冗余感知的超图神经网络处理。首先遍历顶点并根据状态信息动态调度以分发计算任务，然后计算超边上的特征，再聚合超边特征得到顶点特征，最后更新顶点的特征。

（2.1）硬件加速器初始化。通过调用加速器配置接口将图数据信息和超图神经网络模型信息传递到加速器以初始化加速器。默认传递的信息包括：冗余感知的超图结构、超图顶点特征和超边特征信息、模型权重参数。其中，冗余感知的超图结构包括关联矩阵和冗余聚合集

（2.2）动态调度。动态调度阶段旨在通过状态信息来遍历顶点对应的冗余聚合集和原始超边来调度超边更新工作和聚合工作，该阶段需要利用加速器中的协调器完成调度。执行步骤如下：

（2.2.1）利用协调器中的结构预取器将权重提取到权重缓冲区，将超图结构信息提取到结构缓冲区。

（2.2.2）利用任务调度器执行任务分派工作，首先获取目标顶点，遍历目标顶点对应的原始超边和冗余聚合集。如果原始超边没有完成更新操作，则将其添加到列表heList中，用于后续的超边更新；如果原始超边已经完成更新操作，则后续的聚合只需要从缓存中获取中间结果即可。对于冗余聚合集中的每条超边，如果没有完成更新操作，则将其添加到列表heList中，用于后续的超边更新；如果已经完成更新操作，则后续的聚合只需要从缓存中获取中间结果即可。对于冗余聚合集的聚合操作，如果冗余聚合集已经完成聚合操作，则后续的聚合只需要从缓存中获取冗余聚合集的中间结果即可，反之将冗余聚合集添加到aList，用于后续的聚合任务即可。对于目标顶点的聚合操作，将对应超边和冗余聚合集添加到aList。

（2.2.3）利用顶点预取器从高带宽内存中预取顶点属性并提取到顶点缓冲区。

（2.3）超边特征更新。遍历heList中的所有超边，将顶点特征从顶点缓冲区传输到更新器1中的缓冲区，然后更新器1执行具体计算，然后超边的特征，并更新对应状态信息以防止重复计算该超边的特征，然后将计算得到的结果缓存起来，以便后续的超边特征更新可以直接复用中间结果，而不需要重新计算。

（2.4）聚合超边特征。该阶段旨在聚合超边的特征并消除冗余计算，步骤如下：

（2.4.1）聚合控制器通过遍历aList将聚合任务分配到聚合缓冲区，aList包含一组需要执行聚合的超边。

（2.4.2）聚合缓冲区中的任务根据依赖关系并发调度。

（2.4.3）聚合引擎执行具体的聚合任务，它对一组超边的特征执行聚合操作并更新对应状态信息以防止重复执行聚合，然后将聚合结果缓存到顶点缓冲区作为中间结果，如果后续也需要聚合这一组超边，则可以直接复用中间结果，从而避免了冗余聚合操作。

（2.5）顶点特征更新。将每个目标顶点的旧顶点特征和聚合的特征相结合，以基于顶点权重生成新的顶点特征。如果还存在没有更新的顶点，则获取新的目标顶点进入步骤（2.2）；反之完成所有顶点的更新操作。

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

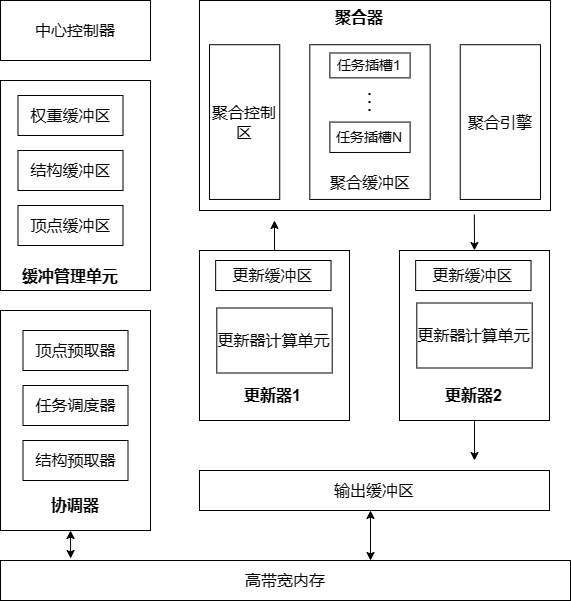


图2 硬件架构图

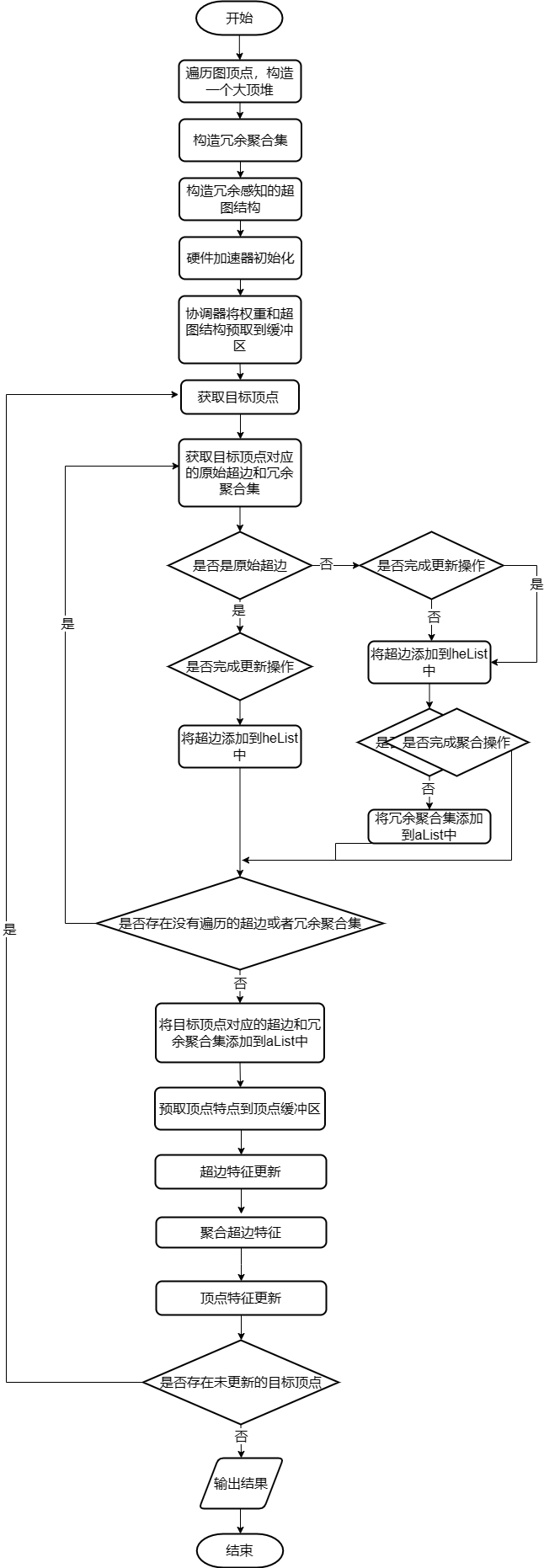


图3 超图神经网络计算流程