

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2017 年 1 月 12 日 (12.01.2017)



(10) 国际公布号
WO 2017/005115 A1

- (51) 国际专利分类号:
G06F 9/50 (2006.01)
- (21) 国际申请号: PCT/CN2016/087461
- (22) 国际申请日: 2016 年 6 月 28 日 (28.06.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510397422.1 2015 年 7 月 8 日 (08.07.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): 黄益聪 (HUANG, Yicong) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 强琦 (QIANG, Qi) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 余骏 (YU, Jun)

[CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 金晓军 (JIN, Xiaojun) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 廖新涛 (LIAO, Xintao) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

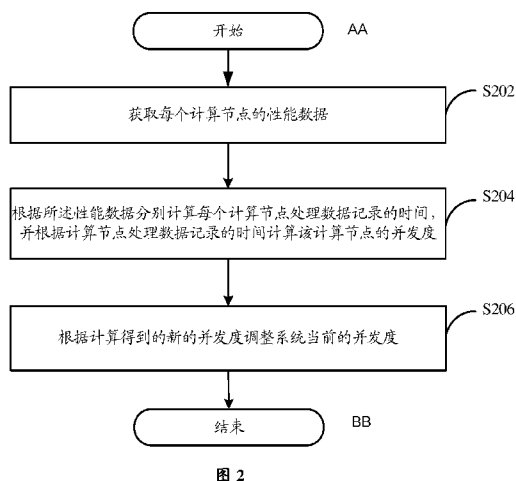
(74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST,

[见续页]

(54) Title: ADAPTIVE OPTIMIZATION METHOD AND DEVICE FOR DISTRIBUTED DAG SYSTEM

(54) 发明名称: 分布式 DAG 系统的自适应优化方法和装置



S202 Acquiring performance data of each calculation node
S204 Respectively calculating time for each calculation node to process a data record according to the performance data, and calculating the degree of concurrency of the calculation node according to the time for the calculation node to process the data record
S206 Adjusting the current degree of concurrency of a system according to the new degree of concurrency obtained by calculation
AA Start
BB End

(57) Abstract: Disclosed are an adaptive optimization method and device for a distributed DAG system. The method comprises: acquiring performance data of each calculation node; respectively calculating time for each calculation node to process a data record according to the performance data, and calculating the degree of concurrency of the calculation node according to the time for the calculation node to process the data record; and adjusting the current degree of concurrency of a system according to the degree of concurrency obtained by calculation. By means of the present application, calculation pipelines can be made to consistently operate under a full load, thereby improving the system performance while significantly saving machine costs.

(57) 摘要: 本申请公开一种分布式 DAG 系统的自适应优化方法和装置, 其中该方法包括: 获取每个计算节点的性能数据; 根据所述性能数据分别计算每个计算节点处理数据记录的时间, 并根据计算节点处理数据记录的时间计算该计算节点的并发度; 根据计算得到的并发度调整系统当前的并发度。通过本申请, 能够使计算流水线满负载一致运行, 在大幅节约机器成本的同时提升系统性能。

WO 2017/005115 A1



SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,
VC, VN, ZA, ZM, ZW。

HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO,
PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ,
CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE,
SN, TD, TG)。

- (84) **指定国** (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚
(AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT,
BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR,

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

分布式DAG系统的自适应优化方法和装置

技术领域

本申请涉及计算机技术领域，尤其涉及一种分布式 DAG 系统的自适应
5 优化方法和装置。

背景技术

DAG (Directed acyclic graph, 有向无环图) 可用于描述分布式离线/在线
系统的工作流程图。图1示例了增量计算平台 (Galaxy) 某个任务 (Job) 的
10 DAG拓扑图，图1中的每个计算节点 (Model) 由多个执行单元 (Executor)
并发运行。在典型的分布式离线/在线计算平台中，计算节点的类型可能包含
有映射处理 (Mapper)、归约处理 (Reduce)、聚合处理 (Merger) 中的一
种或者几种。

在DAG系统中，数据流从DAG图的根节点 (即Source节点) 流入，经过
15 各级运算节点，从叶子节点 (即Output节点) 流出。在DAG中可能有1个或者
多个根节点，在图1中包含有两个根节点：计算节点 0和计算节点 1；叶子节
点也可能有一个或者多个，在图1中包含有4个叶子节点：计算节点 16
(Merger)、计算节点 17 (Merger)、计算节点 18 (Merger) 和计算节点 19
(Merger)。

20 DAG拓扑图各级计算节点的计算对数据流形成了流水线，数据分批次从
根节点流入，逐级经过各级计算节点后，结果从叶子节点输出。其中，
计算节点的计算能力与速度，可以由该计算节点的并发度，即执行单元
(Executor) 的并发数目调整。

25 满负载完美运行的系统要求DAG中的所有计算节点计算能力具有一致
性，如果某个计算节点的计算能力小于上级计算节点，那么上级计算节点输
出的数据将在该计算节点堆积等待处理，并由此影响系统的吞吐性能。

在现有技术中，广泛使用的分布式离线/在线基于流水线的计算平台对计算节点并发度的设置依赖于人工设定，并在任务（Job）运行的生命周期保持不变，具体步骤如下：

5 （1）用户或者系统管理员在提交任务前，在配置文件中指定计算节点的并发度；

 （2）系统在提交任务时读入配置文件，并据此设置各个计算节点的运行并发度；

 （3）各个计算节点的并发度保持不变，直至任务运行结束。

10 现有的技术对计算节点并发度的设置依赖于人工设定，并在任务运行的生命周期内保持不变，具有如下缺点：

15 （1）依赖于人工经验设置很难得到完美的并发参数，使得DAG各级流水线能够一致协调运行。如前所述，参数需要在任务提交前设置，所以在首次设置参数时并不知道各计算节点运行时的真实性能。用户可能需要多次迭代：设置参数－提交任务运行－观察性能－重新设置参数－再次提交任务.....；这样才能得到一组相对合适的参数运行任务。而且依赖于人工观察与设置，参数很可能并非最优配置。

20 （2）不能适应集群环境变化。在任务运行的生命周期内，任务运行的机器环境可能发生变化。比如机器出现故障，任务从原先的几台机器，调度到另几台机器运行；由于集群其他任务的运行，影响了任务的网络吞吐或者磁盘读写速度等。运行环境的变化，将影响计算节点的处理速度，即虽然设置的并发度不变，但是计算节点的实际运行速度发生了变化。因此，原先设置的参数可能将不再适用。

25 （3）不能适应数据流的变化。在实时流式处理系统中，数据流通常并非稳定的，随着业务需求可能有显著的峰值流量与空闲流量。固定不变的参数不能反映数据流的变化。为了适应峰值流量，通常需要设置较大的并发度，造成了集群资源的浪费。

 上述缺点将带来如下后果：

(1) 浪费机器资源。某个或者某几个计算节点设置的并发度可能高于运行时刻的实际需求，集群浪费资源运行空转的执行单元，等待数据。

(2) 集群性能低于源头速度。流水线的吞吐受制于运转最慢的一级流水线或者其中的某个计算节点。不合适的参数设置使得流水线可能存在明显的短板，集群吞吐远低于可以达到的理论最佳值。

发明内容

本申请的主要目的在于提供一种分布式 DAG 系统的自适应优化方法和装置，以克服现有技术中的固定不变及非最优化的节点并发度导致降低系统运行效率的技术问题。

根据本申请实施例提供了一种分布式有向无环图 DAG 系统的自适应优化方法，所述分布式 DAG 系统包括多个计算节点，所述方法包括：获取每个计算节点的性能数据；根据所述性能数据分别计算每个计算节点处理数据记录的时间，并根据计算节点处理数据记录的时间计算该计算节点的并发度；根据计算得到的并发度调整系统当前的并发度。

其中，所述方法还包括：对所述 DAG 系统进行广度优先遍历，获取遍历到的计算节点的性能数据并根据所述性能数据计算该计算节点处理一条数据记录的时间，根据计算节点处理一条数据记录的时间计算该计算节点的并发度。

其中，所述计算节点的类型包括以下的一种或多种：映射处理节点、归约处理节点、聚合处理节点；所述方法还包括：根据计算节点的类型获取该计算节点的性能数据。

其中，所述方法还包括：将获取到的计算节点的性能数据存储至存储系统；从所述存储系统读取存储的性能数据，并根据读取的性能数据计算计算节点处理数据记录的时间。

其中，进一步包括：对于映射处理节点，获取映射处理节点执行一条数据记录的时间；根据以下公式计算映射处理节点处理数据记录的时间： $ppt_i = l_i$ ，其中， l 为映射处理节点执行一条数据记录的时间。

其中，进一步包括：对于归约处理节点，获取归约处理节点执行预定批次数据所花费的时间以及执行预定批次数据的数据数量；根据以下公式计算映射处理节点处理数据记录的时间： $ppt_i = f_i / t_i$ ，其中，f 为归约处理节点执行预定批次数据所花费的时间、t 为归约处理节点执行预定批次数据的数据数量。

其中，进一步包括：对于聚合处理节点，获取聚合处理节点执行预定批次数据所花费的时间、执行预定批次数据的数据数量、生成检查点所需时间、生成一次检查点所间隔的数据批次；根据以下公式计算聚合处理节点处理数据记录的时间： $ppt_i = \max(f_i / t_i, cpt_i / t_i * cb_i)$ ，其中，f 为聚合处理节点执行预定批次数据所花费的时间、t 为聚合处理节点执行预定批次数据的数据数量、cpt 为生成检查点所需时间、cb 为生成一次检查点所间隔的数据批次。

其中，根据以下公式计算计算节点的并发度： $adjV.dop = \text{Sum}(v.dop * v.ratio * adjV.ppt / v.ppt)$ ，其中，adjV 为当前计算节点、v 是当前计算节点的上游邻接计算节点、Sum() 表示对当前计算节点的所有上游邻接计算节点求和、ratio 为计算节点的吞吐率。

其中，所述根据计算得到的新的并发度调整系统当前的并发度，包括：对比计算得到的新的并发度与系统当前的并发度，如果差值大于预设阈值则初始化系统并以新的并发度运行系统。

根据本申请实施例还提供一种分布式 DAG 系统的自适应优化装置，所述分布式 DAG 系统包括多个计算节点，所述装置包括：数据获取模块，用于获取每个计算节点的性能数据；计算模块，用于根据所述性能数据分别计算每个计算节点处理数据记录的时间，并根据计算节点处理数据记录的时间计算该计算节点的并发度；自适应优化模块，用于根据计算得到的并发度调整系统当前的并发度。

其中，所述计算模块还用于，对所述 DAG 系统进行广度优先遍历，获取遍历到的计算节点的性能数据并根据所述性能数据计算该计算节点处理一条数据记录的时间，根据计算节点处理一条数据记录的时间计算该计算节点的并发度。

其中，所述计算节点的类型包括以下的一种或多种：映射处理节点、归约处理节点、聚合处理节点；所述数据获取模块还用于，根据计算节点的类型获取该计算节点的性能数据。

其中，所述装置还包括：存储模块，用于存储所述数据获取模块获取到的
5 的计算节点的性能数据；并接受所述计算模块从其中读取存储的性能数据，以使所述计算模块根据读取的性能数据计算计算节点处理数据记录的时间。

其中，所述数据获取模块包括：第一数据获取模块，用于对于映射处理节点，获取映射处理节点执行一条数据记录的时间；所述计算模块包括：第一计算模块，用于根据以下公式计算映射处理节点处理数据记录的时间：

10 $ppt_i = l_i$ ，其中， l 为映射处理节点执行一条数据记录的时间。

其中，所述数据获取模块包括：第二数据获取模块，用于对于归约处理节点，获取归约处理节点执行预定批次数据所花费的时间以及执行预定批次数据的数据数量；所述计算模块包括：第二计算模块，用于根据以下公式计算归约处理节点处理数据记录的时间： $ppt_i = f_i / t_i$ ，其中， f 为采集归约处理
15 节点执行预定批次数据所花费的时间、 t 为归约处理节点执行预定批次数据的数据数量。

其中，所述数据获取模块包括：第二数据获取模块，用于对于聚合处理节点，获取聚合处理节点执行预定批次数据所花费的时间、执行预定批次数据的数据数量、生成检查点所需时间、生成一次检查点所间隔的数据批次；
20 所述计算模块包括：第三计算模块，用于根据以下公式计算聚合处理节点处理数据记录的时间： $ppt_i = \max(f_i / t_i, cpt_i / t_i * cb_i)$ ，其中， f 为聚合处理节点执行预定批次数据所花费的时间、 t 为聚合处理节点执行预定批次数据的数据数量、 cpt 为生成检查点所需时间、 cb 为生成一次检查点所间隔的数据批次。

其中，所述计算模块还包括：第四计算模块，用于根据以下公式计算计
25 算节点的并发度： $adjV.dop = \text{Sum}(v.dop * v.ratio * adjV.ppt / v.ppt)$ ，其中， $adjV$ 为当前计算节点、 v 是当前计算节点的上游邻接计算节点、 $\text{Sum}()$ 表示对当前计算节点的所有上游邻接计算节点求和、 $ratio$ 为计算节点的吞吐率。

其中，所述自适应优化模块还用于，对比计算得到的新的并发度与系统当前的并发度，如果差值大于预设阈值则初始化系统并以新的并发度运行系

统。

根据本申请的技术方案，依据运行时刻的采样数据，自动优化 DAG 拓扑图各个计算节点的并发度，使得计算流水线能够满负载一致运行，既不空转等待数据也不超负荷运行，在大幅节约机器成本的同时提升系统性能。

5

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

10 图 1 示出现有技术的 DAG 拓扑图；

图 2 示出根据本申请一个实施例的分布式 DAG 系统的自适应优化方法的流程图；

图 3 示出根据本申请另一实施例的分布式 DAG 系统的自适应优化方法的流程图；

15 图 4 示出根据本申请一个实施例的分布式 DAG 系统的自适应优化装置的结构框图；

图 5 示出根据本申请另一实施例的分布式 DAG 系统的自适应优化装置的结构框图。

20 具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

25

根据本申请实施例提供一种分布式 DAG 系统的自适应优化方法。

图 2 示出根据本申请一个实施例的分布式 DAG 系统的自适应优化方法

的流程图，如图 2 所示，所述方法包括：

步骤 S202，获取每个计算节点的性能数据。

在分布式 DAG 系统中包括有多个计算节点，在目标系统启动运行后，开始实时采集每个计算节点（Model）的性能数据；然后，将采样得到的数据写入可靠的存储系统；当计算计算节点的性能数据时，通过存储系统读取其存储的性能数据。

在本申请的一个实施例中，计算节点的类型包括但不限于以下的一种或多种：映射处理节点（Mapper）、归约处理节点（Reduce）、聚合处理节点（Merger）。因此，需要根据计算节点的类型获取该计算节点的性能数据，例如：对于映射处理节点，获取映射处理节点执行一条数据记录的时间（ l ）；用于对于归约处理节点，获取归约处理节点执行预定批次数据所花费的时间（ f ）以及执行预定批次数据的数据数量（ t ）；用于对于聚合处理节点，获取聚合处理节点执行预定批次数据所花费的时间（ f ）、执行预定批次数据的数据数量（ t ）、生成检查点所需时间（ cpt ）、生成一次检查点所间隔的数据批次（ cb ）。

步骤 S204，根据所述性能数据分别计算每个计算节点处理数据记录的时间，并根据计算节点处理数据记录的时间计算该计算节点的并发度。

每隔一个运行周期（例如 30 分钟），从存储系统读取该段时间的采样数据，并计算系统计算节点新的并发度。所谓计算节点的并发度，是指计算节点的执行单元（Executor）的并发数量。

具体地，把所有源头计算节点（即 Model Source）加入访问集 VisitQ，以 VisitQ 为起点对 DAG 进行广度优先遍历；对遍历到的当前计算节点，获取其性能数据，依据公式（1）计算当前计算节点的处理一条数据记录（tuple）的时间（Pure Processing Time，简称 PPT）：

$$ppt_i = \begin{cases} l_i & \text{(if mapper model)} \\ f_i/t_i & \text{(if reduce model)} \\ \max(f_i/t_i, cpt_i/t_i * cb_i) & \text{(if merger model)} \end{cases} \quad (1)$$

其中： l 为映射处理节点执行一条数据记录的时间；

f 为归约处理节点或聚合处理节点执行一批数据所花费的时间；

t 为归约处理或聚合处理节点执行一批数据的数据数量；

cpt 为生成检查点（checkpoint）所需时间；

5 cb 为生成一次检查点所间隔的数据批次，即每间隔多少批次数据做一次检查点。

通过公式（1）表明，在计算计算节点的 PPT 时，根据计算节点类型的不同分别使用不同的计算公式：对于映射处理节点，其 PPT 等于映射处理节点执行一条数据记录的时间（即 l ）；对于归约处理节点，其 PPT 等于归约处理节点执行一批数据所花费的时间及该批次数据的数据数量的比值（即 f/t ）；对于聚合处理节点，计算其 PPT 较复杂，首先计算节点执行一批数据所花费的时间及该批次数据的数据数量的比值（即 f/t ），再计算生成检查点所需时间与节点执行一批数据的数据数量的比值并乘以生成一次检查点所间隔的数据批次（即 $cpt/t*cb$ ），最后取 f/t 和 $cpt/t*cb$ 两值间的较大者作为聚合处理节点的 PPT。

然后，依据公式（2）计算当前计算节点的并发度（Degree of Parallelism，简称 DOP）：

$$adjV.dop = Sum(v.dop * v.ratio * adjV.ppt / v.ppt) \quad (2)$$

其中： $adjV$ 表示当前计算节点， v 表示当前计算节点的上游邻接计算节点， $Sum()$ 表示对当前计算节点的所有上游邻接计算节点求和， $ratio$ 表示计算节点的吞吐率， $ratio$ 等于计算节点输出的数据记录（tuple）数量与计算节点输入的数据记录（tuple）数量的比值。

步骤 S206，根据计算得到的并发度调整系统当前的并发度。

通过上述实施例，根据运行时刻采样数据自适应优化各计算节点的并发度，使得各级流水线的速度能够依据源头数据源的速度自动对齐，在节约机器成本的同时提高集群性能。

下面结合图 3 详细描述本申请实施例的细节。图 3 示出根据本申请另一实施例的分布式 DAG 系统的自适应优化方法的流程图，参考图 3，所述方法包括：

步骤 S302，系统初始化后，开始采集系统中每个计算节点（Model）的性能数据。其中采样间隔可自定义，例如 15、30 或 60 秒采集一次计算节点的性能数据。

在本申请中，计算节点的类型包括但不限于以下的一种或多种：映射处理节点（Mapper）、归约处理节点（Reduce）、聚合处理节点（Merger）。对于映射处理节点，采集节点执行一条数据记录的时间；对于归约处理节点，采集节点执行预定批次数据所花费的时间以及执行预定批次数据的数据数量；对于聚合处理节点，采集节点执行预定批次数据所花费的时间、执行预定批次数据的数据数量、生成检查点所需时间、生成一次检查点所间隔的数据批次。

步骤 S304，将采样得到的数据存入可靠的存储系统，存储系统例如可以是分布式存储系统（HBase），也可以是其他的可靠的存储系统。

步骤 S306，每隔预定的运行周期（例如 15、30 或 60 分钟），从存储系统读取该段时间的采样数据，并计算计算节点新的并发度。对所述 DAG 系统进行广度优先遍历，从存储系统读取遍历到的计算节点的性能数据，并根据采样数据计算计算节点处理数据记录的时间（PPT），具体地：

根据公式 $ppt_i = l_i$ 计算映射处理节点的 PPT；

根据公式 $ppt_i = f_i / t_i$ 计算归约处理节点的 PPT；

根据公式 $ppt_i = \max(f_i / t_i, cpt_i / t_i * cb_i)$ 计算聚合处理节点的 PPT。

步骤 S308，根据公式 $adjV.dop = \text{Sum}(v.dop * v.ratio * adjV.ppt / v.ppt)$ 计算计算节点新的并发度（DOP）。

步骤 S310，根据计算得到的新的并发度自动优化计算节点并发度。具体地，从系统的配置文件读取当前计算节点的并发度，对比新的并发度与系统当前计算节点的并发度，如果差值大于预先设置的阈值，则重新初始化系统，并以新的并发度重新运行系统。

下面结合具体实例详细描述本申请。表 1 示出了本申请对数据计算平台（Galaxy）线上资源消耗排名靠前的任务（Job）主要性能指标的优化前后的数据对比。

5

表 1

Job	优化前		优化后		资源对比 (CPU 数目)	性能对比 (任务延迟)
	CPU 数目 (个)	任务延迟 (毫秒)	CPU 数目 (个)	任务延迟 (毫秒)		
tcif_rp_view_taobao_app	1200	2.58	300	1.62	0.25	1.59
dwi_pub_hbd_biz_device_pv_b2	400	1.13	100	1.06	0.25	1.07
utlog_adash_v2_parse	320	4.79	80	3.31	0.25	1.45
effect_click_gmv	300	0.054	100	0.053	0.33	1.02
ranking-app-item-click	300	2.12	100	1.57	0.33	1.35
ranking-app-monitor-bucket1	300	2.01	100	1.55	0.33	1.30
cdo_abc_metaqlbs_autonavi	512	104.35	128	7.77	0.25	13.43

在表 1 中，任务 tcif_rp_view_taobao_app 是 Galaxy 集群占用资源最多的任务，在性能优化前通过配置文件配置任务需要的资源数量：任务运行需要 300 个 worker 进程(worker_num)，每个 worker 进程绑定 4 个 CPU(cpu_bind) 并使用 3G 内存（Memory），总资源需求为 1200 个 CPU 和 900G 内存，其中每个 worker 进程又包括多个 Executor 线程，原配置文件共配置使用 2947 个 Executor 线程。

数据从实时数据传输平台分批次不间断流入 Galaxy 集群，每批数据 1000 条数据记录。Galaxy 集群依据计算任务生成 DAG 模型，该 DAG 中包括多个计算节点，按照配置文件将申请到的物理资源分配到各个计算节点，配置文件中的片段可以是：

Model0.parallelism=256

Model1.parallelism=76

也就是说，为计算节点 0 配置使用 256 个执行单元（Executor）线程，即计算节点 0 的并发度为 256；表示为计算节点 1 配置使用 256 个执行单元线程，即计算节点 1 的并发度为 76。

在 DAG 中，多个计算节点间存在依赖关系，比如：计算节点 3 和计算节点 5 是计算节点 7 的上游邻接计算节点（计算节点 7 需要依赖计算节点 3 与计算节点 5 的计算结果），而计算节点 0 是计算节点 3 的上游邻接计算节点（计算节点 3 需要依赖计算节点 0 的计算结果），计算节点 1 是计算节点 5 的上游邻接计算节点（计算节点 5 需要依赖计算节点 1 的计算结果）；其中，计算节点 0 和计算节点 1 是根节点。

在计算计算节点 7 的并发度时，首先需要计算计算节点 3 和 5 的并发度。在本实例中，计算节点 3、5 和 7 的类型是映射处理节点；通过采样系统得到计算节点 0 的 l 值（执行一条数据记录的时间）为 0.2 秒、计算节点 1 的 l 值为 0.1 秒、计算节点 3 的 l 值为 0.5 秒、计算节点 5 的 l 值为 0.3 秒、计算节点 7 的 l 值为 0.6 秒；将计算节点 3、5 和 7 的吞吐率（ratio）都设置为 1，则计算节点 3、5 和 7 并发度为：

$$\begin{aligned} Model3.dop &= Model0.dop * Model0.ratio * Model3.ppt / Model0.ppt \\ &= 256 * 1 * 0.5 / 0.2 = 640 \end{aligned}$$

$$\begin{aligned} Model5.dop &= Model1.dop * Model1.ratio * Model5.ppt / Model1.ppt \\ &= 76 * 1 * 0.3 / 0.1 = 228 \end{aligned}$$

$$\begin{aligned} Model7.dop &= Model3.dop * Model3.ratio * Model7.ppt / Model3.ppt + \\ &Model5.dop * Model5.ratio * Model7.ppt / Model5.ppt \\ &= 640 * 1 * 0.6 / 0.5 + 228 * 1 * 0.6 / 0.3 = 1224 \end{aligned}$$

通过上述计算得到了计算节点 3、5 和 7 的新的并发度，并依据上述原理分别计算系统中每个计算节点的并发度，然后重新初始化系统，以新的并发度重新运行系统。经过系统优化后，任务 tcif_rp_view_taobao_app 运行总资源需求由 1200 个 CPU 降低到 300 个 CPU、任务延迟（BatchLatency，数据从源头到 output model 的延迟）由 2.58 毫秒下降到 1.62 毫秒，任务 tcif_rp_view_taobao_app 在节约 900 个 CPU Core 的同时，性能还提升了约 60%。

应用本申请实施例优化之后，对数据计算平台线上资源消耗排名靠前的 6 个任务的优化总计可节约 2040 个 CPU 内核，系统资源成本下降 75% 的同时，性能平均提升了 30%。

- 5 根据本申请实施例还提供一种分布式 DAG 系统的自适应优化装置。图 4 示出根据本申请一个实施例的分布式 DAG 系统的自适应优化装置的结构框图，参考图 4，该装置包括：

数据获取模块 410，用于获取每个计算节点的性能数据。

- 10 计算模块 420，用于根据所述性能数据分别计算每个计算节点处理数据记录的时间，并根据计算节点处理数据记录的时间计算该计算节点的并发度；进一步地，计算模块 420 还用于，对所述 DAG 系统进行广度优先遍历，获取遍历到的计算节点的性能数据并计算该计算节点的并发度。

- 15 自适应优化模块 430，用于根据计算得到的并发度调整系统当前的并发度。具体地，所述自适应优化模块 430 还用于，对比新的并发度与系统当前的并发度，如果差值大于预设阈值，则初始化系统以新的并发度运行系统。

在本申请的一个实施例中，所述计算节点的类型包括：映射处理节点、归约处理节点、聚合处理节点；所述数据获取模块 410 还用于，根据计算节点的类型获取该计算节点的性能数据。

- 20 参考图 5，是根据本申请另一实施例的自适应优化装置的结构框图，如图 5 所示，所述数据获取模块 410 进一步包括：第一数据获取模块 512、第二数据获取模块 514 和第三数据获取模块 516。

- 25 其中，第一数据获取模块 512，用于对于映射处理节点，获取节点执行一条数据记录的时间 (l)；第二数据获取模块 514，用于对于归约处理节点，获取节点执行预定批次数据所花费的时间 (f) 以及执行预定批次数据的数据数量 (t)；第三数据获取模块 516，用于对于聚合处理节点，获取节点执行预定批次数据所花费的时间 (f)、执行预定批次数据的数据数量 (t)、生成检查点所需时间 (cpt)、生成一次检查点所间隔的数据批次 (cb)。

所述计算模块 420 进一步包括：第一计算模块 522、第二计算模块 524、第三计算模块 526 和第四计算模块 528。

其中，第一计算模块 522，用于根据以下公式计算映射处理节点处理数据记录的时间： $ppt_i = l_i$ ，其中， l 为映射处理节点执行一条数据记录的时间。

5 第二计算模块 524，用于根据以下公式计算归约处理节点处理数据记录的时间： $ppt_i = f_i/t_i$ ，其中， f 为采集归约处理节点执行预定批次数据所花费的时间、 t 为归约处理节点执行预定批次数据的数据数量。

第三计算模块 526，用于根据以下公式计算聚合处理节点处理数据记录的时间： $ppt_i = \max(f_i/t_i, cpt_i/t_i * cb_i)$ ，其中， f 为聚合处理节点执行预定批次数据所花费的时间、 t 为聚合处理节点执行预定批次数据的数据数量、 cpt 为生成检查点所需时间、 cb 为生成一次检查点所间隔的数据批次。

10

第三计算模块 528，用于根据以下公式计算计算节点的并发度： $adjV.dop = \text{Sum}(v.dop * v.ratio * adjV.ppt / v.ppt)$ ，其中， $adjV$ 为当前计算节点、 v 是当前计算节点的上游邻接计算节点、 $\text{Sum}()$ 表示对当前计算节点的所有上游邻接计算节点求和、 $ratio$ 为计算节点的吞吐率。

15

继续参考图 5，所述装置还包括有：存储模块 440，用于存储所述数据获取模块 410 获取到的计算节点的性能数据；并接受所述计算模块 420 从其中读取存储的性能数据，以使所述计算模块 420 根据读取的性能数据计算计算节点处理数据记录的时间。

20 本申请的方法的操作步骤与装置的结构特征对应，可以相互参照，不再一一赘述。

综上所述，根据本申请的技术方案，依据运行时刻的采样数据，自动优化 DAG 拓扑图各个计算节点的并发度，使得计算流水线能够满负载一致运行（不空转等待数据也不超负荷），在大幅节约机器成本的同时提升系统性能。

25

本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含

有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

- 5 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

- 10 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁
15 磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

- 20 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排除性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

- 25 以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

权利要求书

1、一种分布式有向无环图 DAG 系统的自适应优化方法，所述分布式 DAG 系统包括多个计算节点，其特征在于，所述方法包括：

获取每个计算节点的性能数据；

5 根据所述性能数据分别计算每个计算节点处理数据记录的时间，并根据计算节点处理数据记录的时间计算该计算节点的并发度；

根据计算得到的并发度调整系统当前的并发度。

2、根据权利要求 1 所述的方法，其特征在于，还包括：

10 对所述 DAG 系统进行广度优先遍历，获取遍历到的计算节点的性能数据并根据所述性能数据计算该计算节点处理一条数据记录的时间，根据计算节点处理一条数据记录的时间计算该计算节点的并发度。

3、根据权利要求 1 所述的方法，其特征在于，所述计算节点的类型包括以下的一种或多种：映射处理节点、归约处理节点、聚合处理节点；

所述方法还包括：根据计算节点的类型获取该计算节点的性能数据。

15 4、根据权利要求 3 所述的方法，其特征在于，还包括：

将获取到的计算节点的性能数据存储至存储系统；

从所述存储系统读取存储的性能数据，并根据读取的性能数据计算计算节点处理数据记录的时间。

5、根据权利要求 3 所述的方法，其特征在于，进一步包括：

20 对于映射处理节点，获取映射处理节点执行一条数据记录的时间；

根据以下公式计算映射处理节点处理数据记录的时间： $ppt_i = l_i$ ，其中， l 为映射处理节点执行一条数据记录的时间。

6、根据权利要求 3 所述的方法，其特征在于，进一步包括：

对于归约处理节点，获取归约处理节点执行预定批次数据所花费的时间

以及执行预定批次数据的数据数量；

根据以下公式计算映射处理节点处理数据记录的时间： $ppt_i = f_i / t_i$ ，其中， f 为归约处理节点执行预定批次数据所花费的时间、 t 为归约处理节点执行预定批次数据的数据数量。

5 7、根据权利要求 3 所述的方法，其特征在于，进一步包括：

对于聚合处理节点，获取聚合处理节点执行预定批次数据所花费的时间、执行预定批次数据的数据数量、生成检查点所需时间、生成一次检查点所间隔的数据批次；

根据以下公式计算聚合处理节点处理数据记录的时间：

10 $ppt_i = \max(f_i / t_i, cpt_i / t_i * cb_i)$ ，其中， f 为聚合处理节点执行预定批次数据所花费的时间、 t 为聚合处理节点执行预定批次数据的数据数量、 cpt 为生成检查点所需时间、 cb 为生成一次检查点所间隔的数据批次。

8、根据权利要求 5、6 或 7 所述的方法，其特征在于，根据以下公式计算计算节点的并发度：

15 $adjV.dop = \text{Sum}(v.dop * v.ratio * adjV.ppt / v.ppt)$ ，其中， $adjV$ 为当前计算节点、 v 是当前计算节点的上游邻接计算节点、 $\text{Sum}()$ 表示对当前计算节点的所有上游邻接计算节点求和、 $ratio$ 为计算节点的吞吐率。

9、根据权利要求 1 所述的方法，其特征在于，所述根据计算得到的新的并发度调整系统当前的并发度，包括：

20 对比计算得到的新的并发度与系统当前的并发度，如果差值大于预设阈值则初始化系统并以新的并发度运行系统。

10、一种分布式 DAG 系统的自适应优化装置，所述分布式 DAG 系统包括多个计算节点，其特征在于，所述装置包括：

数据获取模块，用于获取每个计算节点的性能数据；

25 计算模块，用于根据所述性能数据分别计算每个计算节点处理数据记录

的时间，并根据计算节点处理数据记录的时间计算该计算节点的并发度；

自适应优化模块，用于根据计算得到的并发度调整系统当前的并发度。

- 11、根据权利要求 10 所述的装置，其特征在于，所述计算模块还用于，
对所述 DAG 系统进行广度优先遍历，获取遍历到的计算节点的性能数据并
5 根据所述性能数据计算该计算节点处理一条数据记录的时间，根据计算节点
处理一条数据记录的时间计算该计算节点的并发度。

12、根据权利要求 10 所述的装置，其特征在于，所述计算节点的类型包
括以下的一种或多种：映射处理节点、归约处理节点、聚合处理节点；

- 所述数据获取模块还用于，根据计算节点的类型获取该计算节点的性能
10 数据。

13、根据权利要求 12 所述的装置，其特征在于，还包括：

存储模块，用于存储所述数据获取模块获取到的计算节点的性能数据；
并接受所述计算模块从其中读取存储的性能数据，以使所述计算模块根据读
取的性能数据计算计算节点处理数据记录的时间。

- 15 14、根据权利要求 12 所述的装置，其特征在于，

所述数据获取模块包括：第一数据获取模块，用于对于映射处理节点，
获取映射处理节点执行一条数据记录的时间；

- 所述计算模块包括：第一计算模块，用于根据以下公式计算映射处理节
点处理数据记录的时间： $ppt_i = l_i$ ，其中， l 为映射处理节点执行一条数据记录
20 的时间。

15、根据权利要求 12 所述的装置，其特征在于，

所述数据获取模块包括：第二数据获取模块，用于对于归约处理节点，
获取归约处理节点执行预定批次数据所花费的时间以及执行预定批次数据的
数据数量；

- 25 所述计算模块包括：第二计算模块，用于根据以下公式计算归约处理节

点处理数据记录的时间: $ppt_i = f_i/t_i$, 其中, f 为采集归约处理节点执行预定批次数据所花费的时间、 t 为归约处理节点执行预定批次数据的数据数量。

16、根据权利要求 12 所述的装置, 其特征在于,

5 所述数据获取模块包括: 第二数据获取模块, 用于对于聚合处理节点, 获取聚合处理节点执行预定批次数据所花费的时间、执行预定批次数据的数据数量、生成检查点所需时间、生成一次检查点所间隔的数据批次;

10 所述计算模块包括: 第三计算模块, 用于根据以下公式计算聚合处理节点处理数据记录的时间: $ppt_i = \max(f_i/t_i, cpt_i/t_i * cb_i)$, 其中, f 为聚合处理节点执行预定批次数据所花费的时间、 t 为聚合处理节点执行预定批次数据的数据数量、 cpt 为生成检查点所需时间、 cb 为生成一次检查点所间隔的数据批次。

17、根据权利要求 14、15 或 16 所述的装置, 其特征在于, 所述计算模块还包括: 第四计算模块, 用于根据以下公式计算计算节点的并发度:

15 $adjV.dop = \text{Sum}(v.dop * v.ratio * adjV.ppt / v.ppt)$, 其中, $adjV$ 为当前计算节点、 v 是当前计算节点的上游邻接计算节点、 $\text{Sum}()$ 表示对当前计算节点的所有上游邻接计算节点求和、 $ratio$ 为计算节点的吞吐率。

18、根据权利要求 10 所述的装置, 其特征在于, 所述自适应优化模块还用于, 对比计算得到的新的并发度与系统当前的并发度, 如果差值大于预设阈值则初始化系统并以新的并发度运行系统。

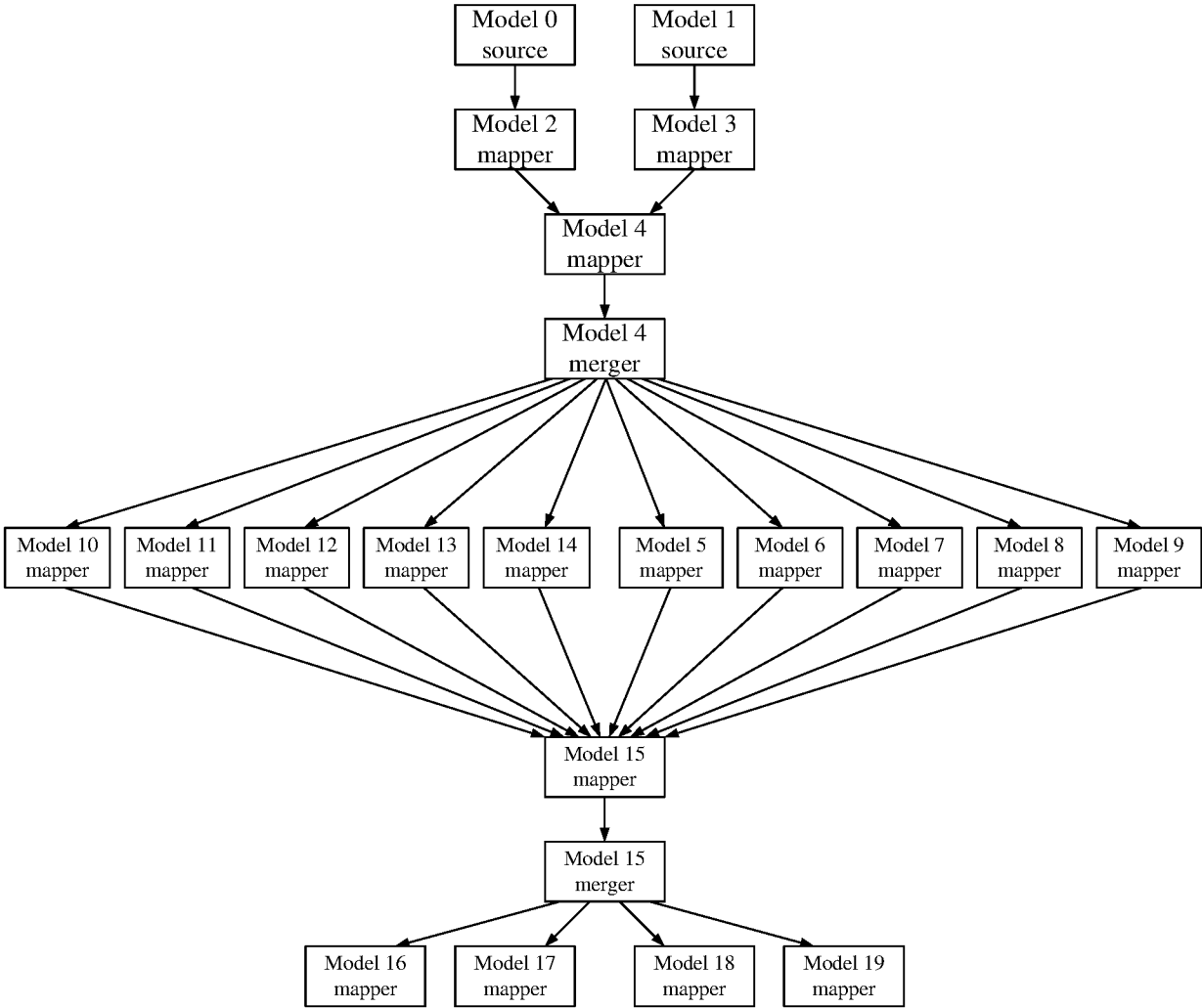


图 1

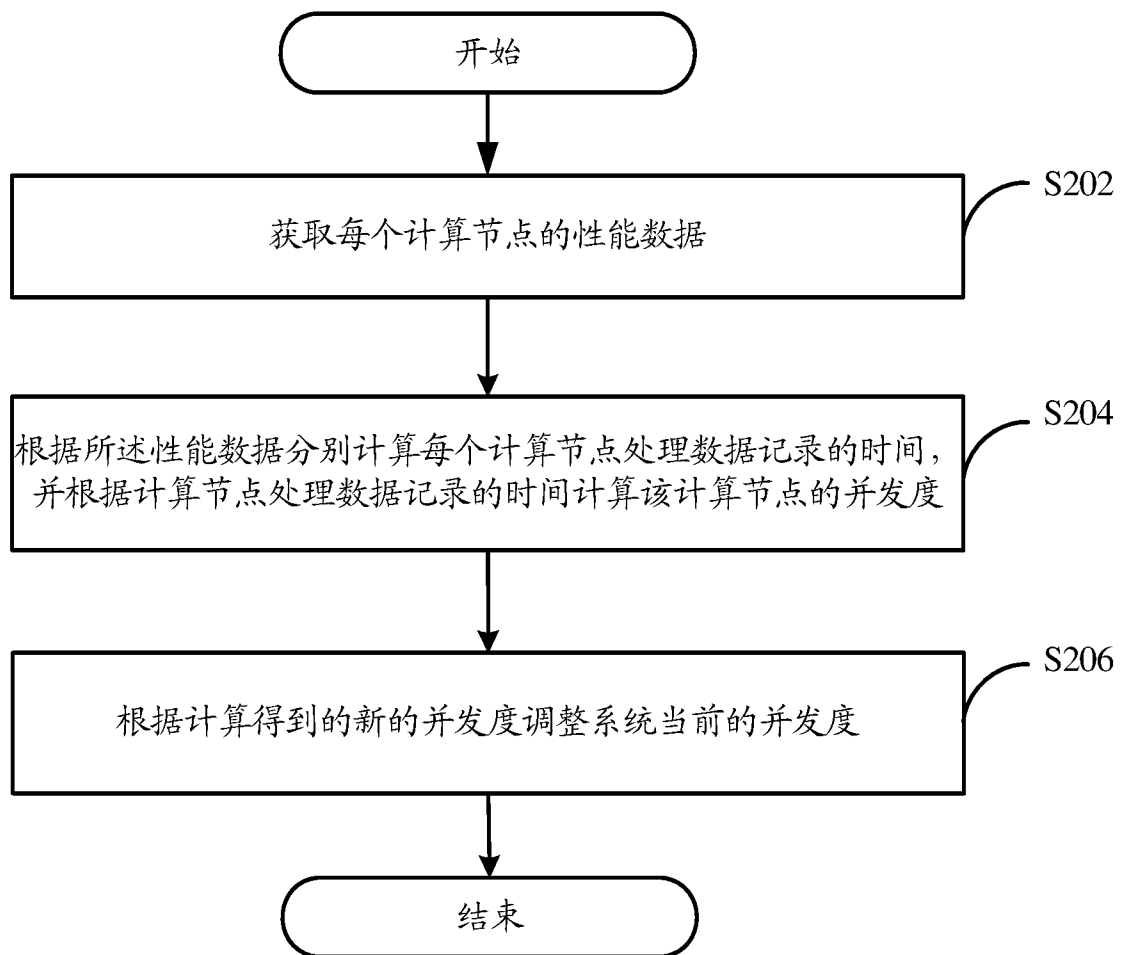


图 2

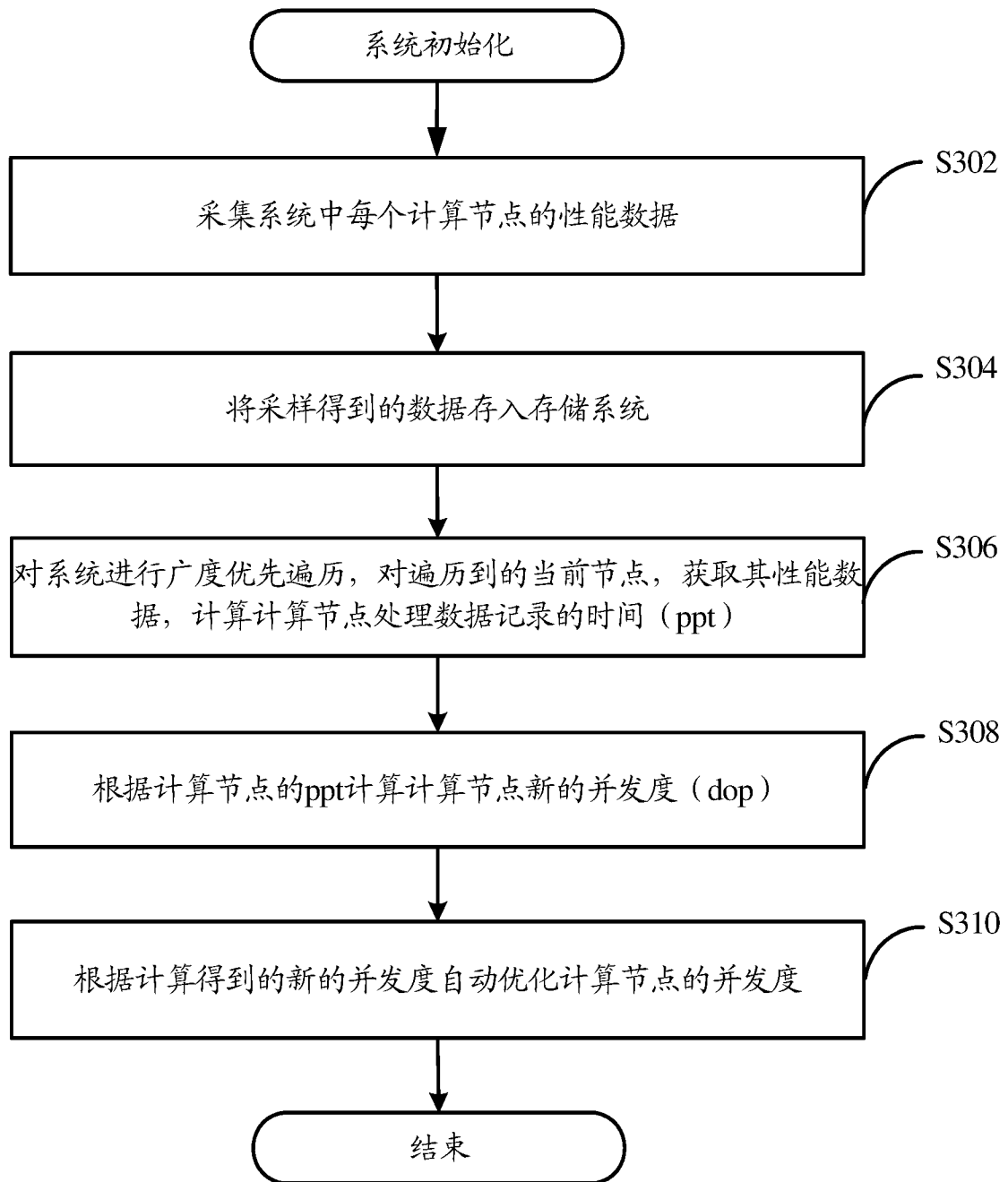


图 3

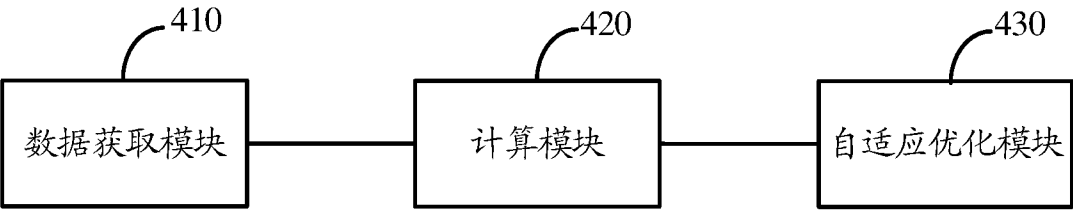


图 4

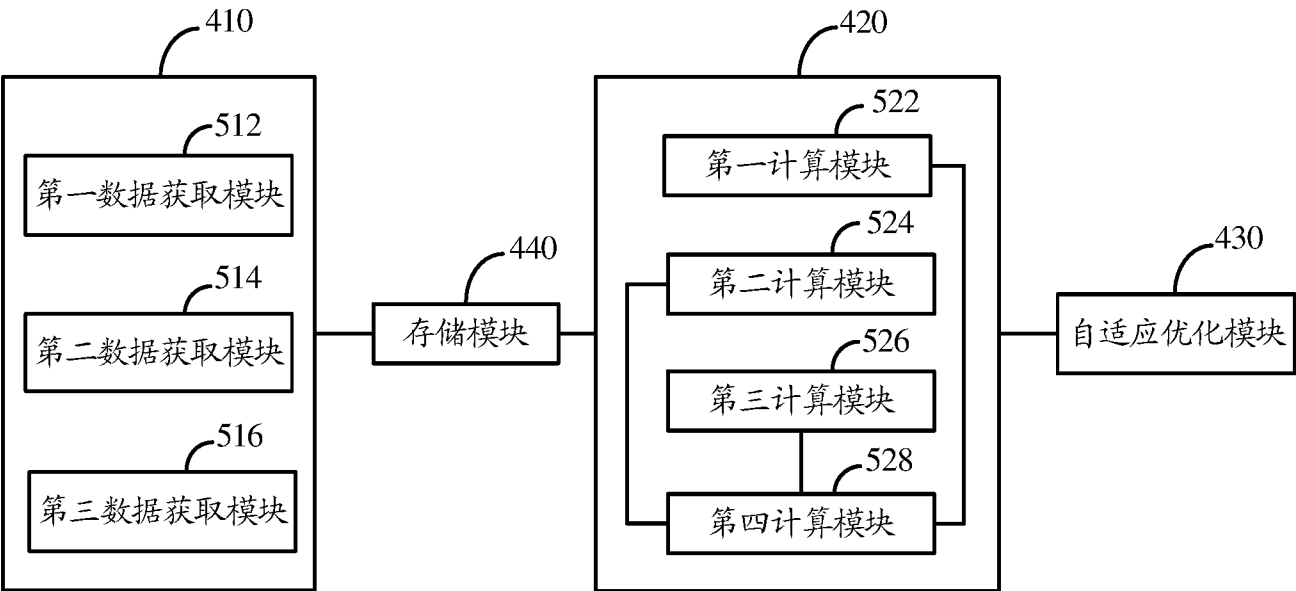


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/087461

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/50 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; DWPI; CNKI: concurrent, degree of concurrency, time map, directed acyclic graph, aggregation, distributed, reduction, directed, graph, current, acyclic, intercurrent, performance, node, time, map, converge, distribute

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102012844 A (SHANGHAI UNIVERSITY), 13 April 2011 (13.04.2011), description, paragraph 11, claim 1, and figure 1	1-4, 9-13, 18
A	CN 102012844 A (SHANGHAI UNIVERSITY), 13 April 2011 (13.04.2011), see the whole document	5-8, 14-17
Y	CN 102360246 A (WUHAN UNIVERSITY OF TECHNOLOGY), 22 February 2012 (22.02.2012), description, paragraphs 27-65	1-4, 9-13, 18
A	CN 102360246 A (WUHAN UNIVERSITY OF TECHNOLOGY), 22 February 2012 (22.02.2012), see the whole document	5-8, 14-17
A	CN 103150148 A (CENTER FOR EARTH OBSERVATION AND DIGITAL EARTH, CHINESE ACADEMY OF SCIENCES), 12 June 2013 (12.06.2013), see the whole document	1-18
A	CN 103491024 A (INSTITUTE OF INFORMATION ENGINEERING, CHINESE ACADEMY OF SCIENCES), 01 January 2014 (01.01.2014), see the whole document	1-18
A	US 2009171731 A1 (IBM), 02 July 2009 (02.07.2009), see the whole document	1-18

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 14 September 2016 (14.09.2016)	Date of mailing of the international search report 10 October 2016 (10.10.2016)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer WANG, Yan Telephone No.: (86-10) 62411625

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2016/087461

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102012844 A	13 April 2011	CN 102012844 B	09 January 2013
CN 102360246 A	22 February 2012	CN 102360246 B	09 April 2014
CN 103150148 A	12 June 2013	CN 103150148 B	01 June 2016
CN 103491024 A	01 January 2014	None	
US 2009171731 A1	02 July 2009	None	

A. 主题的分类 G06F 9/50 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNABS;DWPI;CNKI:性能, 节点, 并发, 并发度, 图, 时间, 时间映射, 有向无环图, 聚合, 分布式, 归约, directed, graph, current, acyclic, intercurrent, performance, node, time, map, converge, distribute		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
Y	CN 102012844 A (上海大学) 2011年 4月 13日 (2011 - 04 - 13) 说明书第11段, 权利要求1, 附图1	1-4, 9-13, 18
A	CN 102012844 A (上海大学) 2011年 4月 13日 (2011 - 04 - 13) 参见全文	5-8, 14-17
Y	CN 102360246 A (武汉理工大学) 2012年 2月 22日 (2012 - 02 - 22) 说明书第27-65段	1-4, 9-13, 18
A	CN 102360246 A (武汉理工大学) 2012年 2月 22日 (2012 - 02 - 22) 参见全文	5-8, 14-17
A	CN 103150148 A (中国科学院对地观测与数字地球科学中心) 2013年 6月 12日 (2013 - 06 - 12) 参见全文	1-18
A	CN 103491024 A (中国科学院信息工程研究所) 2014年 1月 1日 (2014 - 01 - 01) 参见全文	1-18
A	US 2009171731 A1 (IBM) 2009年 7月 2日 (2009 - 07 - 02) 参见全文	1-18
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期 2016年 9月 14日		国际检索报告邮寄日期 2016年 10月 10日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 王艳 电话号码 (86-10)62411625

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/087461

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	102012844	A	2011年 4月 13日	CN 102012844 B	2013年 1月 9日
CN	102360246	A	2012年 2月 22日	CN 102360246 B	2014年 4月 9日
CN	103150148	A	2013年 6月 12日	CN 103150148 B	2016年 6月 1日
CN	103491024	A	2014年 1月 1日	无	
US	2009171731	A1	2009年 7月 2日	无	

表 PCT/ISA/210 (同族专利附件) (2009年7月)