ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2017/11(07)-1021-12
doi: 10.3778/j.issn.1673-9418.1702022

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

性能收益模型的流处理算子优化技术综述*

檀国林[1,2,3],海 玲[4+],张 鹏[1,2,3],陈志鹏[1,2,3]
1. 中国科学院 信息工程研究所,北京 100093
2. 信息内容安全技术国家工程实验室,北京 100093
3. 中国科学院大学,北京 100049
4. 新疆工程学院,乌鲁木齐 830091

Survey of Stream Processing Operator Optimizations for Performance Gain*

TAN Guolin[1,2,3], HAI Ling[4+], ZHANG Peng[1,2,3], CHEN Zhipeng[1,2,3]
1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
2. National Engineering Laboratory for Information Security Technologies, Beijing 100093, China
3. University of Chinese Academy of Sciences, Beijing 100049, China
4. Xinjiang Institute of Engineering, Urumchi 830091, China
+ Corresponding author: E-mail: 2468775007@qq.com

TAN Guolin, HAI Ling, ZHANG Peng, et al. Survey of stream processing operator optimizations for performance gain. Journal of Frontiers of Computer Science and Technology, 2017, 11(7): 1021-1032.

Abstract: The arrival of the era of big data and mobile Internet lets people be in data torrent. Big data make the need more and more urgent to process data efficiently and real-timely. And big data also prompt research teams around the world to develop a lot of stream processing applications. The implementations of these stream processing applications use a variety of operator optimizations. Based on the research on these stream processing applications, this paper summarizes the most common eight operator optimizations of stream processing applications. And combined with practical examples, this paper introduces the features of these operator optimizations from aspects of performance gain, safety condition and dynamic. Then this paper discusses the further research direction in the field of operator optimizations and stream processing.

Key words: operator optimization; stream processing; performance gain

* The National Natural Science Foundation of China under Grant Nos. 61402464, 61402474, 61602467 (国家自然科学基金).
Received 2017-01, Accepted 2017-03.
CNKI网络优先出版: 2017-03-27, http://kns.cnki.net/kcms/detail/11.5602.TP.20170327.1932.002.html

摘 要:大数据移动互联网时代的到来,数据量也越来越庞大,数据之大使得对数据进行高效实时处理的需求也变得越来越迫切,促使国内外的研究团队开发出许多流处理应用。为了提高流处理应用的性能,这些流处理应用底层实现都采用了各种各样复杂的流处理算子优化技术。在调研学习这些流处理应用的基础上,概括总结了其中最常见的8种流处理算子优化技术,并结合实际例子,分别从性能收益、安全条件、动态性等方面详细介绍了这些算子优化技术的特点,并探讨了算子优化和流处理应用领域进一步的研究方向。

关键词:算子优化;流处理;性能收益

文献标志码:A **中图分类号:**TP311.1

1 引言

随着移动互联网时代的到来,越来越多的流处理应用不断出现。这些流处理应用广泛地应用于各种生产生活环境,例如网络监控、通讯数据管理、网站点击数据流监控、传感器网络、移动设备扫描监控等。在这些应用中数据是以一种持续的数据流的形式存在,而不是存在于传统关系数据库中。

数据流的广泛存在,促使了国内外许许多多的研究团队开发出了各种流处理应用,例如Storm、S4、Spark Streaming等。这些流处理应用实现方法和原理虽然不尽相同,但是它们的底层实现或多或少都使用到了流处理算子优化技术。流处理应用由算子和数据流组成,如果把算子表示成有向图中的节点,把数据流表示成有向图中的边,那么它们构成的有向图被称作流图,或者数据流图、流拓扑图。

本文提到流处理算子优化就是针对流图中的算子进行优化重组,尽可能大地提升流处理应用的处理性能。根据不同算子优化技术的特点,本文按照表1中所列的8种算子优化技术分别进行介绍。

Table 1 Operator optimization techniques

表1 算子优化技术

算子优化技术	流拓扑图	语义	动态性
算子重排	改变	未改变	不定
算子消冗	改变	未改变	不定
算子状态共享	未改变	未改变	静态
算子融合	改变	未改变	不定
算子分裂	改变	不定	不定
算子负载均衡	未改变	未改变	不定
算子替换	未改变	不定	不定
算子负载分流	未改变	改变	动态

其中“语义”表示该优化技术的输入/输出行为,“未改变”表示采用该算子优化技术后,算子的输入和输出都没有变化;“动态性”表示算子优化技术是在算子编译之前静态发生还是运行过程中发生的,“不定”表示两种情况都有可能,具体看相应的流处理应用的实现。

2 算子重排

算子重排是把选择度低的算子移动到上游算子前面,使流处理应用尽可能早地过滤掉无用的数据,减少下游算子需要处理的数据量,从而提升系统性能。算子重排原理如图1所示,将选择度低的算子B移动到算子A前面。

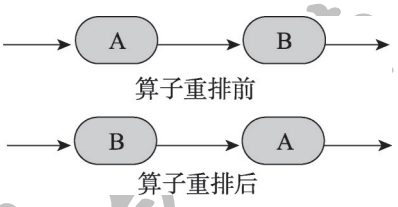

Fig.1 Operator reorder

图1 算子重排示意图

2.1 应用举例

考虑一个视频监控的流处理应用,某城市各个区都布置一定数量的视频监控设备,监控把实时拍摄的视频帧数据流回传到服务器。在一次犯罪嫌疑人抓捕过程中,现已经确定该犯罪嫌疑人的身份,警方希望通过该城市的视频监控设备来帮助他们实施抓捕,在视频帧回传服务器端实时监测犯罪嫌疑人,一旦发现可疑人物,立即通知警方实施抓捕。

已经确定犯罪嫌疑人的性别为男性,以及他的

主要活动区域。视频帧回传服务器端流处理应用包含两个算子，算子A在视频帧中识别人物的性别，并将包含男性人物的数据帧发送给下游算子B，算子B负责将接收到的数据帧按照该市不同的区进行过滤。

假设视频帧中男女性别分布是1∶1，那么算子A将过滤掉一半的数据；算子B过滤掉绝大部分无关区域的数据。如果将算子B移动到算子A前面，那么算子A需要处理的数据将会大大减少，系统的性能也会因此而提高。

2.2 收益模型

对于一个算子 β，如果输入的数据项个数是 N，输出的数据项个数是 n，那么将 n 与 N 的比值称作算子 β 的选择度。

对于上面的例子，算子A的选择度是0.5。如果将选择度低的算子移动到前面，那么算子重排是有收益的。

假设算子A的选择度固定为0.5，算子B的选择度是变化的。图2显示了在算子B选择度变化的情况下，算子重排前和算子重排后的系统负载量。

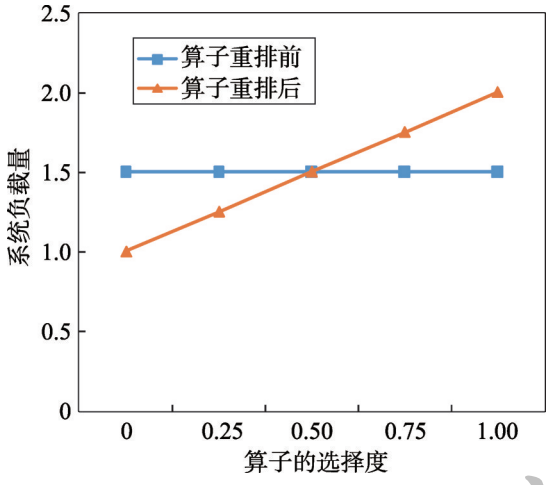

Fig.2 Influence of operator reorder on system load under different operator selection degrees

图2 不同算子选择度下算子重排对系统负载量的影响

算子重排前，所有到来的数据算子A都要处理（数据量按单位1计算），接着算子A筛选一半的数据给下游算子B处理，因此算子重排前算子A和算子B总共处理的数据恒为1.5。

算子重排后，所有到来的数据先由算子B处理（数据量按单位1计算），接着算子B按照自己的选择度 β 筛选数据给下游算子A处理，因此算子A和算子B处理的数据量总共为 $1+\beta$。

从图2中可以看出，当横坐标算子B选择度 β 低于0.5时，算子重排前的系统负载量要高于算子重排后的，当算子B的选择度 β 大于0.5时，算子重排后的系统负载量要高于算子重排前的。因此得出结论，只要将低选择度的算子移动到上游，就可以降低系统负载量，从而提高系统性能。

2.3 安全条件

并不是所有情况下算子重排都是安全的。当满足以下条件时算子重排是安全可行的。

确保属性可用性：由于第二个算子B处理的数据是经过算子A过滤后的数据项，如果要算子重排的话，算子B读取的数据属性集必须和算子A写入的数据属性集是不相交的。

确保交换性：算子重排后和算子重排前的结果必须是一样的，因为必须保证算子A和算子B是可以交换的。如果已经确保了属性的可用性，那么交换的充分条件是算子A和算子B都是无状态的。

2.4 动态性

从前面的介绍中可以知道，算子重排的性能收益是受算子选择度的影响，因此算子重排的性能收益通常依赖于输入数据的分布。如果输入数据的分布是明确并且不变的，那么算子重排可以在算子静态编译之前确定；反之，需要在算子运行的时候动态地改变算子顺序，从而来提高算子性能。

Eddy算子通过静态拓扑图转换实现动态算子重排[1]。它假定一个数据项被一个算子丢弃的概率与被另一个算子丢弃的概率是独立的，优点是不必事先知道其选择度，但是这样会导致一些额外的开销来进行数据路由选择。

Babu等人提出了另一种利用近似算法的动态算子重排技术。该方法能够处理数据项在算子之间被丢弃的概率是相互依赖的情况，并且保证结果是在最优解的一个小的常数区间之内[2-3]。

3 算子消冗

算子消冗又叫作子图共享，它通过共享算子的方式将数据流图中冗余算子消除，如图3所示。

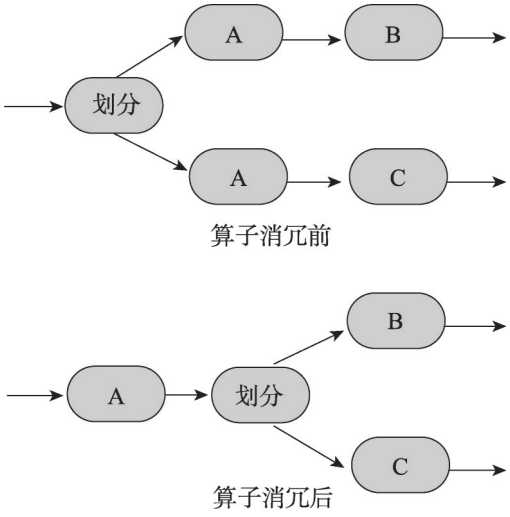

Fig.3 Operator redundancy elimination

图3 算子消冗示意图

3.1 应用举例

有两个Web统计应用，分别接收来自不同网站发来的请求，它们都是以算子A开始来解析发送来的JSON请求数据，然后将解析后的数据分别交给下游算子B和算子C。这两个应用都包含解析JSON(JavaScript object notation)数据的算子A，因此可以通过算子消冗技术来消除多余的算子A，使它们共享一个算子A，从而节省了计算资源。

3.2 收益模型

如果算子消冗的两个应用运行在同一台机器上，并且是单核执行的，那么通过算子消冗是有收益的。

算子消冗前，假设系统的负载量是单位1，算子消冗后，因为流处理应用共享了同一个算子A，那么系统会少做一次算子A的计算，所以系统的负载量会由于消除的算子A的负载量而减少。

从图4可以看出算子A的负载量在整个应用中的负载比值越大，那么算子消冗所带来的性能收益越大。

3.3 安全条件

确保相同的算法：由于算子消冗会改变流拓扑图，在这种情况下要保证算子优化技术的语义不变，那么就要确保消除的冗余算子是等价的。算子等价是经典的不可判定问题，在实际应用过程中可以采用判断算子是否具有相同代码的方法。

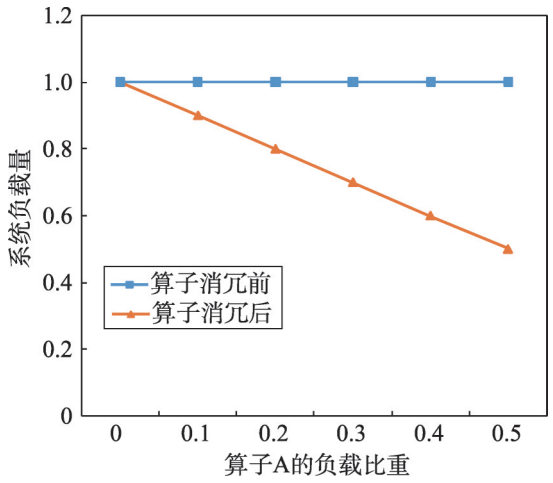

Fig.4 Influence of system load under operator A with different load proportion

图4 算子A不同负载比重消冗后对系统负载量的影响

确保组合的状态：当消除的冗余算子是无状态的，那么消除冗余算子是很容易的事情；但是如果消除的冗余算子是有状态的，那么需要考虑的问题就更多了。

3.4 动态性

基于数据流图的算子静态编译器已经可以检测和消除一个流应用中的冗余算子、无意义算子、幂等算子和死子图等情况。

动态的算子消冗更多的是应用于分布式集群流处理应用的情况，流拓扑图中的算子被分配到集群中的机器上运行。当提交新的拓扑到集群当中时，系统可以先搜索判断当前是否已经存在公共子图，如果存在的话，就可以通过算子消冗技术来减少系统资源占用。Pietzuch等人提出比较极端的方法[4-5]，把流应用的添加和移除看成一个类似常规数据项添加和移除的优先操作。

4 算子融合

算子融合是把两个连续的算子融合成一个更大的算子，如图5所示。

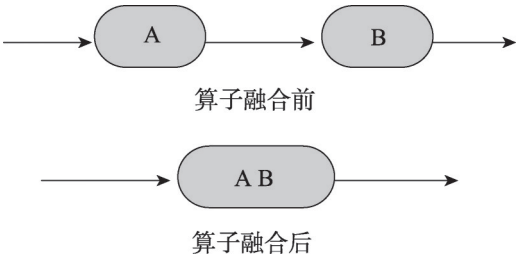

Fig.5 Operator fusion

图5 算子融合示意图

4.1 应用举例

考虑这样一个流量监控应用，算子A接收前端回传的流量，并将流量还原成<源IP，源端口，目的IP，目的端口，协议类型>形式的五元组信息，然后将五元组发送给后端的算子B。算子B根据五元组信息过滤掉一些常规的流量，对异常的流量进行提取分析。

由于算子B只是根据规则简单地匹配提取信息，所花费的计算成本非常小，因此可以将算子A和算子B融合到一个算子里面，这样就可以减少数据传输的成本。同时由于减少了数据处理步骤，也减低了系统出错的几率。在这种情况下，算子融合技术是可以考虑的。

4.2 收益模型

算子融合技术与前面介绍的几种算子优化技术不一样，前面介绍的算子优化技术是以减少整个流处理应用的负载量来提升性能收益的，而算子融合技术并没有减少整个流处理应用的负载量，它只是把位于不同机器上的算子融合到一个算子里面。这样做的好处是可以减少数据在不同算子之间移动所带来的成本，但是同时它也减少了管道并行。这就需要对管道并行和数据算子间移动的成本进行比较，权衡后选择是否进行算子融合。当算子B计算成本很小，同时数据算子间移动的成本很大时，可以考虑进行算子融合。

另一方面，算子融合技术减少了数据在算子间移动的步骤，因此也减少了数据传输错误和机器故障的可能性。

4.3 安全条件

目前大多数流处理应用都是基于分布式集群的，流处理任务提交到集群后，会将算子分发到集群中的机器上去执行。因此不同机器上的算子间不会存在本地资源的竞争，例如本地文件、CPU、内存等。

避免无限递归：如果在流拓扑图中存在闭环，例如一个反馈回路，那么数据可以在回路中无限循环。如果算子是通过函数调用的方式实现融合的，那么将会导致堆栈调用溢出。

4.4 动态性

融合通常是静态完成的，也有一些系统是动态地使用算子融合技术。flextream系统通过暂停应用，重新编译融合后的代码，然后恢复应用来实现动态融合[6]。这种动态融合技术可以应对流处理应用可用资源的改变，但是由于重新编译和运行也会导致一定的延迟。Tang和Gedik应用算子融合技术，把那些算子共享一个线程的决策留给了运行时，这使得他们能够动态地控制流水线、线程切换和通信之间的成本开销[7-8]。

5 算子状态共享

算子状态共享是指多个算子共享同一块内存空间，从而减少内存占用，降低对系统的资源需求，如图6所示。

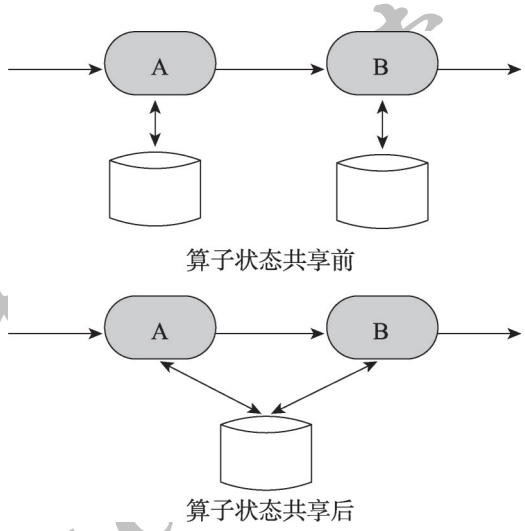

Fig.6 Operator state sharing

图6 算子状态共享示意图

5.1 应用举例

假设有两个算子A和B，分别实时计算某社交网

站一分钟内和一小时内最热门的top *N*动态,这样算子A和算子B需要维护一个很大的内存窗口,来记录当前所有用户发布的动态。如果算子A和B只是计算的时间粒度不一样,那么可以让算子A共用算子B的内存空间[9-10]。

5.2 收益模型

算子状态共享能够带来两方面的收益。首先,由于多个算子共享内存,减少了内存空间的使用,能够减少流处理应用对资源使用的需求。另一方面,减少内存空间的使用可以提高计算机缓存的命中率,甚至能够减少磁盘I/O时间。

5.3 安全条件

确保状态是可访问的:由于多个算子访问同一个内存空间,要求该内存空间对多个算子是可访问的。Storm底层实现了多种并行化技术,Storm的一个worker进程包含有多个Executor线程,而每个Executor线程又可以启动多个task,这里的task相当于一个算子。Storm通过这种方式能够确保同一个Executor线程不同task算子之间状态是可访问的[11-12]。

避免资源竞争:当多个算子读写同一块内存空间时就需要考虑资源竞争的问题,要么确保操作是只读的,要么通过同步访问来控制读写正确性。Brito等人使用了软件事务内存,该机制允许尝试性地同时对同一状态进行更新,如果需要的话,可以使用回滚技术[13]。

5.4 动态性

状态共享通常是静态执行的。StreamIt采用全静态方法,其中一个静态调度规定了什么样的数据在什么时间被什么算子所访问。Brito等人的工作更多的是动态方式,共享状态的访问通过软件事务内存来协商完成。

6 算子分裂

算子分裂又叫作算子划分和数据并行化,它是把数据按照一定的规则划分成好几份,然后将每份数据路由给下游相同的算子进行处理,是一种典型的数据并行化,如图7所示。

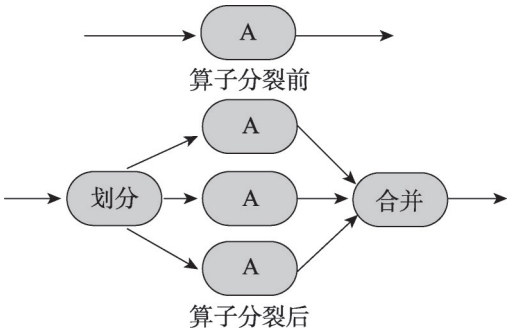

Fig.7 Operator fission

图7 算子分裂示意图

算子分裂又分为两种,一种是基于key的算子分裂,另一种是基于批处理的算子分裂。基于key的算子分裂是将数据中的某一个或者某一些属性,按照属性值的不同划分到不同机器上的下游算子。最常见的划分算法是利用哈希函数进行划分。基于key的算子分裂受限于key的个数,并行度有限。基于批处理的算子分裂是把数据按照一定的窗口大小进行划分,最常见的两种窗口是时间滑动窗口和个数滑动窗口。

6.1 应用举例

假设有一个大型的购物网站,时时刻刻都在接收来自全球各地的用户请求,如果只是由单个算子A来响应用户请求的话,在请求高峰时期必定会导致请求拥堵等待,影响用户体验。如果将用户发送来的请求按照不同地区,分别划分到下游多个相同的算子A,那么就可以通过算子分裂来满足所有用户的请求。这是一种典型的基于key的算子分裂。

6.2 收益模型

算子分裂会引入两个额外的算子,划分算子和合并算子。划分算子将数据划分成多份,然后决定将划分的数据路由到下游哪个算子当中。合并算子负责将所有算子A处理后的结果合并,如果合并的数据流有先后顺序,那么合并的开销会很大。

在一般情况下,算子分裂中数据划分和合并的计算成本相对于算子A来说是非常小的,因此通过算子分裂来提升系统性能是非常有效的,并且系统的吞吐量几乎是与算子A的并行度成正比,如图8所示。

6.3 安全条件

如果算子是有状态的,那么要保证划分的key是

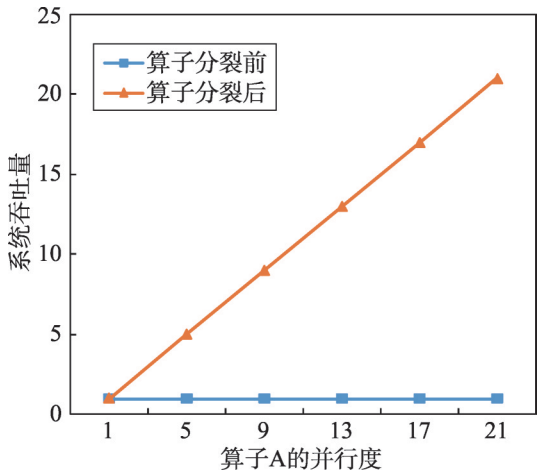

Fig.8 Influence of throughput with operator fission

图8 算子分裂对吞吐量的影响

不相交的。无状态算子的分裂显然是安全的，算子分裂后也能够得到正确的结果；如果算子是有状态的，那么需要保证算子的状态是基于划分后的不同key，也就是说算子的状态不依赖于整体的数据，而是依赖于划分到当前算子的数据。例如，计算网站每个用户消费的总金额，如果算子分裂是按照不同的用户ID来划分的，那么每个算子计算的总金额只与划分到当前算子的数据有关。

如果数据是有顺序的，那么合并算子需要有序合并。对于有些应用，它要求数据在算子分裂后和算子分裂前顺序是一样的。例如，一个流量还原应用，要求处理后的流量和处理前一样，是按照一定的顺序排列的。对于这种情况，需要下游合并算子能够使用一定策略保证合并后的数据依然是有序的。目前已经有许多方法来处理这种情况，CQL（continue query language）使用逻辑时间戳来标记不同的数据[14-15]；StreamIt用round-robin或者复制的方式[16-17]；MapReduce没有按原来的顺序重排，而是使用了一个分布式的“sort”阶段[18-19]；Storm为用户提供了Global Grouping接口和Fields Grouping接口，Global Grouping接口会把所有数据发送给同一个BOLT算子，虽然能够保证整体有序，但是也失去了算子分裂的优势，Fields Grouping能够保证每个Fields内的数据有序。同时Storm支持用户通过编程的方式，自己实现合并算子的有序性[20-21]。

6.4 动态性

在有些情况下，人们一开始并不知道算子分裂的并行度，在流处理应用运行的过程中，可能随着key的个数的增加需要动态地复制需要分裂的算子。SEDA（staged event-driven architecture）通过线程控制器做到了这一点，它把线程数保持到最大值之下[13,22]。Storm允许用户通过手动调整流处理应用的线程数或者进程数，从而动态地调整流处理应用的并行度[23-24]。

7 算子负载均衡

算子负载均衡是根据算子运行情况，动态地把数据从高负载算子路由给空闲的算子，如图9所示。与前面几种算子优化技术不一样，该算子优化技术不改变数据流图，只是改变数据路由的去向[25-26]。

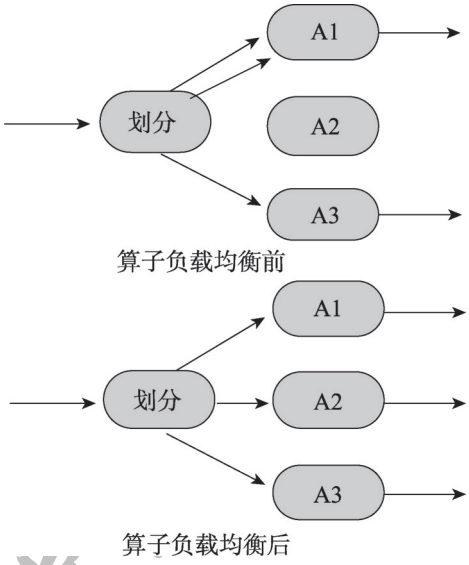

Fig.9 Operator load balance

图9 算子负载均衡示意图

7.1 应用举例

算子负载均衡是在算子分裂基础上采用的流处理算子优化技术。假设有一个网络流量监控应用，该应用把流量的协议类型作为key来进行算子分裂。在运行过程中，如果某种协议类型的流量突然增加，而其他的协议类型的流量突然减少，这样会导致流量突增的算子过载，而流量减少的算子存在空闲。这种负载倾斜会降低应用整体的吞吐量，甚至

过载的算子会由于处理不及时而将数据丢弃。

在流处理应用中，上面介绍的负载倾斜是很常见的，因为大多数数据的统计分布人们并不能事先知道。针对这种情况，如果采用算子负载均衡的话可以很好地解决这种问题[27]。

7.2 收益模型

负载倾斜会导致过载的算子成为整个流处理应用的瓶颈，整个流处理的并行度不再取决于算子分裂复制的并行算子个数，而是由过载算子决定。如果算子负载均衡能够有效地解决数据划分不均匀而引起的负载倾斜问题，那么算子负载均衡是有收益的。

假设过载算子需要处理的数据是所有数据的50%，尽管算子分裂带来的并行度是 N，那么整个流处理应用的并行度其实是2。流处理应用的并行度为 $P=\frac{1}{\max\{p_i\}}$，其中 p_i 内算子分裂中复制的算子所处理的负载占总的负载比值。从公式中可以看出，流处理应用的整体并行度取决于过载算子所承担的负载，如图10所示。

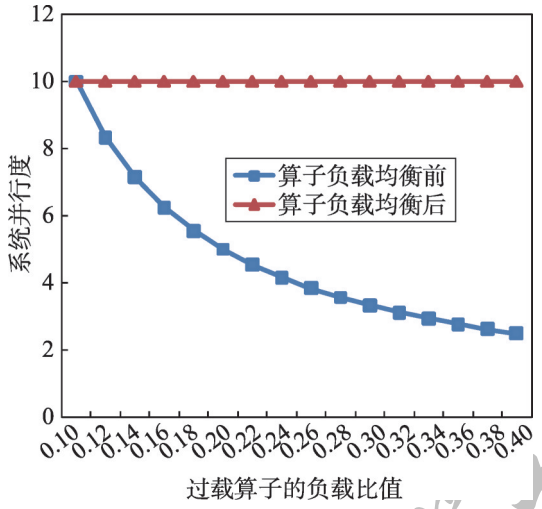

Fig.10 Influence of system parallelism with operator load balance

图10 算子负载均衡对系统并行度的影响

7.3 安全条件

算子负载均衡是基于算子分裂的，因此算子负载均衡首先要满足算子分裂的安全条件。

完全性：确保重新划分的每一个数据都能被下游空闲算子处理。

兼容性：如果算子分裂是基于key划分的，那么将多个key路由到下游同一个算子的时候，要确保该算子都能够处理，即下游算子能够处理不同key类型的数据。

7.4 动态性

根据算子负载均衡的定义，算子负载均衡是算子执行时动态进行的。因为负载倾斜是由数据分布不均匀导致，这就需要划分算子在划分数据时统计历史数据，及时预测负载倾斜，并将数据路由到空闲算子上。

Vitorovic等人认为基于Hash函数的划分会产生负载倾斜问题，因此提出基于数据流数学统计分布特性来动态扩展相应算子，能够很好地解决负载倾斜问题[28]。

8 算子替换

算子替换是选择具有更高效算法的算子来替换之前的算子，如图11所示。

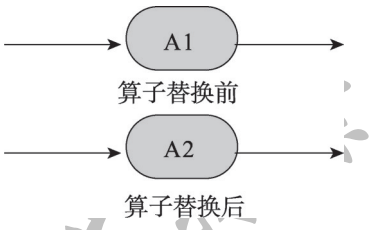

Fig.11 Operator replacing

图11 算子替换示意图

8.1 应用举例

考虑这样一个在数据流上的类SQL查询，该查询包含一个连接算子。连接算子有多种实现方法，最常见的3种是嵌套循环连接、哈希连接和排序连接。通常情况下用嵌套循环方法实现的连接算子磁盘I/O延迟很大，因此可以将嵌套循环方法实现的连接算子替换成哈希连接或者是排序连接算子。

8.2 收益模型

采用嵌套循环方法实现的连接算子的复杂度是 $O(n^2)$，哈希连接算子的复杂度是线性的，而排序连接

算子的复杂度介于嵌套循环连接算子和哈希连接算子之间。如果将一个高成本的算子替换成低成本的算子，那么收益显然是存在的。

图12显示了不同连接算子的输入数据规模 n 和算子执行成本 t 的关系。其中算子执行成本是算子执行时间的近似值。

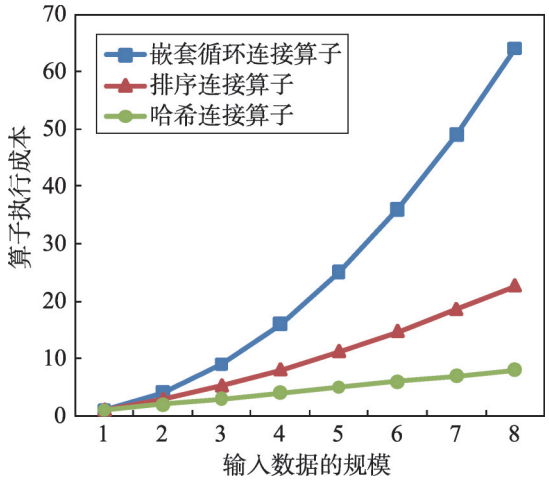

Fig.12 Execution cost of different join operators

图12 不同连接算子的执行成本

8.3 安全条件

确保算法等价性：对于要替换的两个算子，要保证它们是等价的，即执行这两个算法得到的结果是一样的。当其中一个算子只能满足特定的情况才能执行时，算子替换时还要考虑特殊的情况。

8.4 动态性

当算子替换用于对运行时状况做出反馈时，那么它必须是动态的。SEDA的每个算子可以判定其过载的规则，其中的一个可选项是提供降级服务（即算子替换）；在Borealis中，算子能够控制输入，例如为算子选择不同的算法变体[28-30]。为了实现动态的算子替换，编译器在静态时提供算法的所有变体，运行时动态地选择所需要的其中一个算法。换句话说，这种算子替换方式和Eddy为算子重排所做的方式是一样，即静态地插入一个动态路由组件[31-32]。

9 算子负载分流

算子负载分流是通过丢弃一部分数据来提高系统吞吐量的算子优化技术，如图13所示。

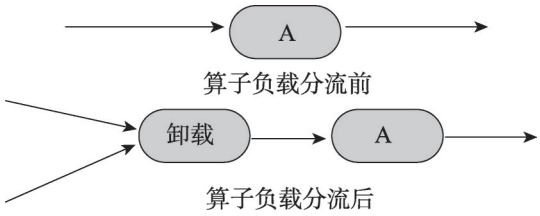

Fig.13 Operator load shedding

图13 算子负载分流示意图

9.1 应用举例

负载分流一开始出现在电力供应系统，是指当高峰期电力供应不足时，通过灯火管制将部分地区的电力切断，从而使其他大部分地区能够正常电力供应。同样在流处理应用中，算子负载分流是指丢弃部分待处理数据，从而使应用整体能够正常运行。例如一个视频请求应用，当高峰期来临时，请求数量急剧增加，系统不能够及时处理的话就会导致请求堆积，影响整个系统的使用，当使用算子负载分流时，丢弃一部分用户的请求，以达到系统的正常稳定运行。

9.2 收益模型

根据算子负载分流的定义，算子负载分流会改变算子的语义。但是当算子执行结果不需要很精确或者不需要满足每一个用户请求的话，是可以通过算子负载分流来应对高峰时期的情况。

算子负载分流的精确度是由卸载算子决定的，当卸载算子的选择度越大，算子负载分流的精确度就越大，反之算子负载分流的精确度就越小。图14显示了算子负载分流的精确度和卸载算子的选择度之间的关系。

9.3 安全条件

从图14中可以知道，卸载算子会导致算子负载分流的精确度下降，也就是说算子负载分流是不安全的。不同于本文描述的其他算子优化技术，算子负载分流技术是在性能和精确度之间进行折中的一种优化技术，尽管这样，有的流处理应用是允许这种不精确的近似结果，例如传感器网络很多都是采集物理世界的模拟信号。大数据时代的到来也使人们从寻求因果关系中解脱出来，进而探索更多的相关关系。

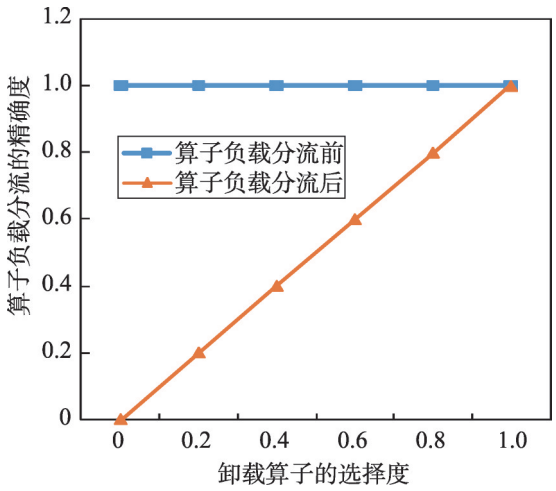

Fig.14 Influence of accuracy with different selection degrees of operator shedding

图14 卸载算子选择度对精确度的影响

9.4 动态性

根据定义,可以知道算子负载分流是动态进行的,它在算子运行过程中,需要卸载算子动态监测当前的负载,当负载超过算子的处理能力时,就采取一定的策略来卸载部分数据。

10 结束语

本文调研学习了当前比较新的流处理框架,结合实际流处理应用,概括总结了其中最常见的8种流处理技术,并且分析了这8种流处理技术的性能收益模型,从而对算子优化技术的性能收益有一个更加直观的认识。同时为了能够将这些算子优化技术正确地应用于实际,还从安全条件和动态性方面介绍了这8种算子优化技术的特点。在实际应用中,这些算子优化技术的情况是复杂的,并且是可以结合起来同时使用的,具体采用哪种算子优化技术要根据具体的条件和优化目标而定。

通过这篇综述可以对算子优化技术有一个比较全面的认识,但是细化到具体的某一项算子优化技术,还是有很多需要开拓和改进的地方。例如,在算子替换当中,已经有部分工作研究如何在流处理应用中使用多路非等值连接(multi-way theta-join)[33-34]来处理更复杂的流处理情况;在算子分裂优化技术中,如何保证在数据划分后合并的整体有序性也是值得研究的问题。除了研究具体算子优化技术的突破和改进,将算子优化技术应用于实际,开发出更高效可用的流处理系统也是未来工作的重要方向。

References:

[1] Avnur R, Hellerstein J M. Eddies: continuously adaptive query processing[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, USA, May 15-18, 2000. New York: ACM, 2000: 261-272.

[2] Babu S, Motwani R, Munagala K, et al. Adaptive ordering of pipelined stream filters[C]//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, Paris, Jun 13-18, 2004. New York: ACM, 2004: 407-418.

[3] Arasu A, Babu S, Widom J. The CQL continuous query language: semantic foundations and query execution[J]. The International Journal on Very Large Data Bases, 2006, 15(2): 121-142.

[4] Pietzuch P, Ledlie J, Shneidman J, et al. Network-aware operator placement for stream- processing systems[C]//Proceedings of the 22nd International Conference on Data Engineering, Atlanta, USA, Apr 3-8, 2006. Washington: IEEE Computer Society, 2006: 49.

[5] Hirzel M, Andrade H, Gedik B, et al. IBM streams processing language: analyzing big data in motion[J]. IBM Journal of Research and Development, 2013, 57(3/4): 7.

[6] Hormati A H, Choi Y, Kudlur M, et al. Flextream: adaptive compilation of streaming applications for heterogeneous architectures[C]//Proceedings of the 18th International Conference on Parallel Architectures and Compilation Techniques, Raleigh, USA, Sep 12-16, 2009. Washington: IEEE Computer Society, 2009: 214-223.

[7] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the Web[J]. World Wide Web Internet and Web Information Systems, 1998, 54: 1-17.

[8] Gedik B, Schneider S, Hirzel M, et al. Elastic scaling for data stream processing[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(6): 1447-1463.

[9] Gedik B, Wu K L, Yu P S. Efficient construction of compact shedding filters for data stream processing[C]//Proceedings of the 24th International Conference on Data Engineering, Cancún, México, Apr 7-12, 2008. Piscataway, USA:

IEEE, 2008: 396-405.

[10] Gordon M I, Thies W, Amarasinghe S. Exploiting coarse-grained task, data, and pipeline parallelism in stream programs[C]//Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems, San Jose, USA, Oct 21-25, 2006. New York: ACM, 2006: 151-162.

[11] Coutts D, Leshchinskiy R, Stewart D. Stream fusion: from lists to streams to nothing at all[C]//Proceedings of the 12th ACM SIGPLAN International Conference on Functional Programming, Freiburg, Germany, Oct 1-3, 2007. New York: ACM, 2007: 315-326.

[12] Toshniwal A, Taneja S, Shukla A, et al. Storm@twitter[C]//Proceedings of the 2014 ACM SIGMOD Iinternational Conference on Management of Data, Snowbird, USA, Jun 22-27, 2014. New York: ACM, 2014: 147-156.

[13] Brito A, Fetzer C, Sturzrehm H, et al. Speculative out-of-order event processing with software transaction memory[C]//Proceedings of the 2nd International Conference on Distributed Event-Based Systems, Rome, Italy, Jul 1-4, 2008. New York: ACM, 2008: 265-275.

[14] Okcan A, Riedewald M. Processing theta-joins using MapReduce[C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, Athens, Greece, Jun 12-16, 2011. New York: ACM, 2011: 949-960.

[15] Mayer R, Koldehofe B, Rothermel K. Meeting predictable buffer limits in the parallel execution of event processing operators[C]//Proceedings of the 2014 International Conference on Big Data, Washington, Jun 27-Jul 2, 2014. Piscataway, USA: IEEE, 2014: 402-411.

[16] Burchett K, Cooper G H, Krishnamurthi S. Lowering: a static optimization technique for transparent functional reactivity [C]//Proceedings of the 2007 ACM SIGPLAN Symposium on Partial Evaluation and Semantics-Based Program Manipulation, Nice, France, Jan 15-16, 2007. New York: ACM, 2007: 71-80.

[17] Carney D, Çetintemel U, Rasin A, et al. Operator scheduling in a data stream manager[C]//Proceedings of the 29th International Conference on Very Large Data Bases, Berlin, Sep 9-12, 2003. New York: ACM, 2003: 838-849.

[18] Chen Jianjun, DeWitt D J, Tian Feng, et al. NiagaraCQ: a scalable continuous query system for Internet databases[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, USA, May 15-18, 2000. New York: ACM, 2000: 379-390.

[19] Cortes C, Fisher K, Pregibon D, et al. Hancock: a language for analyzing transactional data streams[J]. ACM Transactions on Programming Languages and Systems, 2004, 26 (2): 301-338.

[20] Amini L, Jain N, Sehgal A, et al. Adaptive control of extreme-scale stream processing systems[C]//Proceedings of the 26th International Conference on Distributed Computing Systems, Lisboa, Portugal, Jul 4-7, 2006. Washington: IEEE Computer Society, 2006: 71.

[21] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[C]//Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation, San Francisco, USA, Dec 6-8, 2004. Berkeley, USA: USENIX Association, 2004: 10.

[22] Kotto-Kombi R, Lumineau N, Lamarre P, et al. Parallel and distributed stream processing: systems classification and specific issues[EB/OL]. (2015-10-13)[2016-12-20]. https://hal.archives-ouvertes.fr/hal-01215287.

[23] Gordon M I, Thies W, Karczmarek M, et al. A stream compiler for communication-exposed architectures[C]//Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems, San Jose, USA, Oct 5-9, 2002. New York: ACM, 2002: 291-303.

[24] Graefe G. Encapsulation of parallelism in the volcano query processing system[C]//Proceedings of the 1990 International Conference on Management of Data, Atlantic City, USA, May 23-26, 1990. New York: ACM, 1990: 102-111.

[25] Babcock B, Datar M, Motwani R. Load shedding for aggregation queries over data streams[C]//Proceedings of the 20th International Conference on Data Engineering, Boston, USA, Mar 30-Apr 2, 2004. Washington: IEEE Computer Society, 2004: 350-361.

[26] Barga R S, Goldstein J, Ali M, et al. Consistent streaming through time: a vision for event stream processing[C]//Proceedings of the 3rd Conference on Innovative Data Systems Research, Asilomar, USA, Jan 7-10, 2007. New York: ACM, 2007: 363-373.

[27] Hueske F, Peters M, Sax M J, et al. Opening the black boxes

in data flow optimization[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1256-1267.

[28] Vitorovic A, Seidy M E, Guliyev K M O, et al. Squall: scalable real-time analytics[J]. Proceedings of the VLDB Endowment, 2016, 9(13): 1553-1556

[29] Abadi D J, Ahmad Y, Balazinska M, et al. The design of the Borealis stream processing engine[C]//Proceedings of the 2nd Conference on Innovative Data Systems Research, Asilomar, USA, Jan 4-7, 2005. New York: ACM, 2005: 277-289.

[30] Liu Bin, Zhu Yali, Rundensteiner E. Run-time operator state spilling for memory intensive long-running queries[C]//Proceedings of the 2006 International Conference on Management of Data, Chicago, USA, Jun 27-29, 2006. New York: ACM, 2006: 347-358.

[31] Olston C, Jiang Jing, Widom J. Adaptive filters for continuous queries over distributed data streams[C]//Proceedings of the 2003 International Conference on Management of Data, San Diego, USA, Jun 9-12, 2003. New York: ACM, 2003: 563-574.

[32] Ottoni G, Rangan R, Stoler A, et al. Automatic thread extraction with decoupled software pipelining[C]//Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture, Barcelona, Spain, Nov 12-16, 2005. Washington: IEEE Computer Society, 2005: 105-118.

[33] Hirzel M, Soulé R, Schneider S, et al. A catalog of stream processing optimizations[J]. ACM Computing Surveys, 2014, 46(4): 46.

[34] Zhang Xiaofei, Chen Lei, Wang Min. Efficient multi-way theta-join processing using MapReduce[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1184-1195.

TAN Guolin was born in 1992. He is a Ph.D. candidate at University of Chinese Academy of Sciences. His research interests include big data and traffic analysis, etc.

檀国林(1992—),男,安徽安庆人,中国科学院大学博士研究生,主要研究领域为大数据,流量分析等。

HAI Ling was born in 1982. She received the M.S. degree from Dalian University of Technology. Now she is a lecturer at Xinjiang Institute of Engineering. Her research interests include signal detection and processing, stream data analysis and parallel computing, etc.

海玲(1982—),女,新疆玛纳斯人,工程硕士,新疆工程学院讲师,主要研究领域为信号检测与处理,数据流分析,并行计算等。

ZHANG Peng was born in 1984. He received the Ph.D. degree in data integration and service computing from Institute of Computing Technology, Chinese Academy of Sciences in 2013. Now he is an associate professor and M.S. supervisor at Institute of Information Engineering, Chinese Academy of Sciences. His research interests include large-scale stream data processing and cyberspace security, etc.

张鹏(1984—),男,2013年于中国科学院计算技术研究所获得博士学位,现为中国科学院信息工程研究所副研究员、硕士生导师,主要研究领域为大规模流数据处理,网络空间安全等。

CHEN Zhipeng was born in 1989. He is a Ph.D. candidate at University of Chinese Academy of Sciences. His research interests include cyber security, big data processing and analysis, etc.

陈志鹏(1989—),男,山东威海人,中国科学院大学博士研究生,主要研究领域为网络安全,大数据处理和分析等。