毕业论文-中期检查

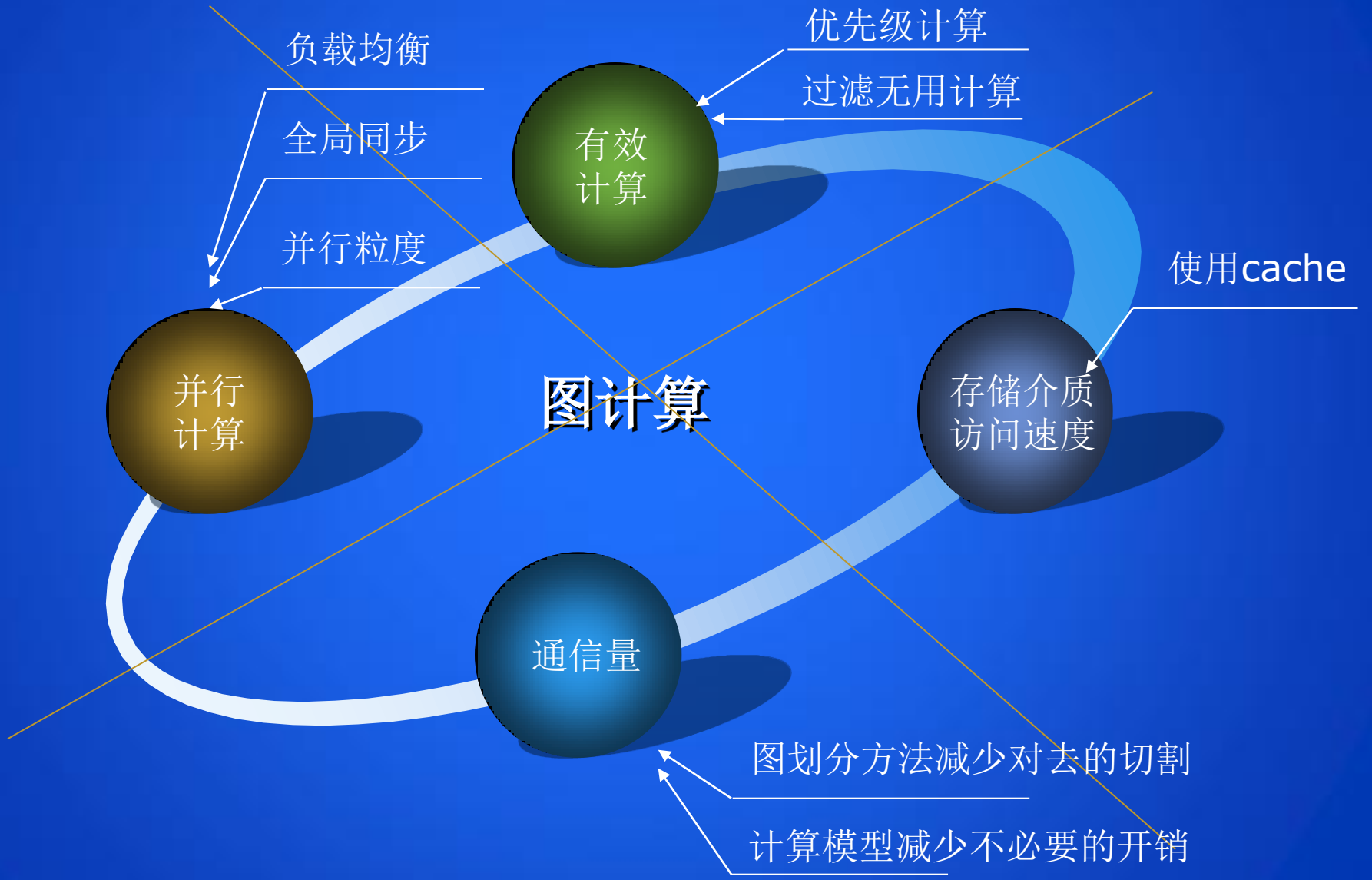
--李玉国

--2016.7.16

异步分布式图处理模型与框架研究

影响分布式图处理框架性能的关键因素分析

影响分布式图处理框架性能的关键因素分析



影响分布式图处理框架性能的关键因素分析

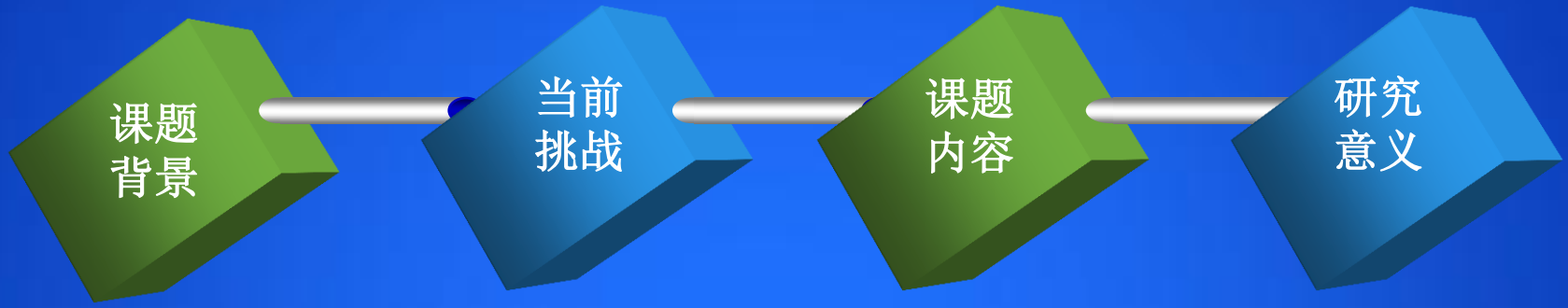
目录

1 开题回顾

2 前期工作进度汇报

3 后期计划

1.开题回顾



1大数据时代-大规模图数据（社交网络、web）
2.数据计算的应用需求-图迭代计算（推荐系统）

由于图数据的扭曲分布，现有的分区方法会导致大量的通信和负载不均衡。

1设计高效的图划分方法

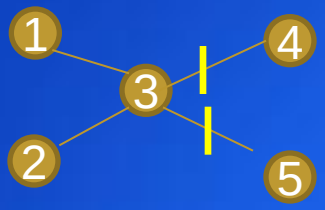
2.设计高效的计算模型，

1.提出异步分布式图计算模型消息流通性的概念。

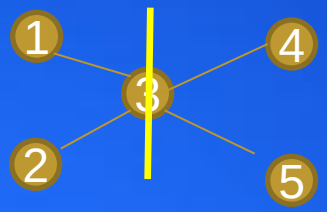
2.设计和实现了一个高效的异步图处理框架，增强了对大规模图数据的处理能力。

1. 对现有图分区方法进行了实验和分析
2. 针对异步计算模型中消息传递的特性提出了消息流通性的概念（消息的本地流通性和通信量）。
3. 提出一种基于消息流通性和负载均衡的图分区方法PAGraph。
4. 提出了一种基于定量计算来决定是否切分顶点的PAGraph
5. 基于DAIC，提出了一种高效的计算模型MR-DAIC。

两种图分区方法：



Edge-cut



Vertex-cut

现有的Vertex-cut方法：

PowerGraph：

Random-Hash;Greedy

问题：切割顶点产生了大量的低度副本顶点，导致了大量的通信和计算开销。

相关研究证明：切割度低的顶点将更容易完成图的分割。

HDRF:选择边的两端的中度低一个进行切割

PowerLyra-Ginger：阶段1.将低度顶点进行**Edge-cut**（不切割低度顶点）；

阶段2.对高度顶点进行**Vertex-cut**。

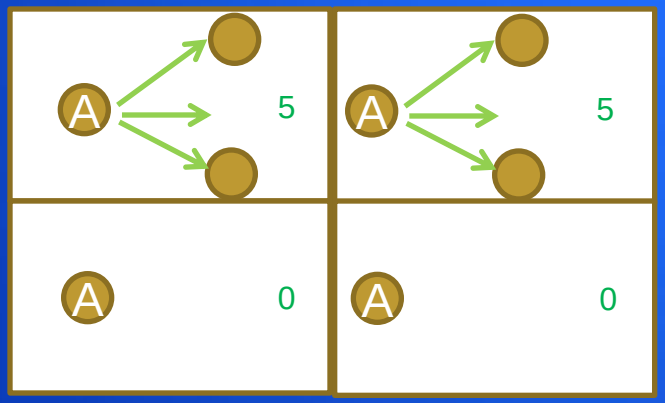
PAGraph:

基于HDRF，结合异步框架中消息传递的特性，实现了一个面向异步的图分区方法PAGraph。

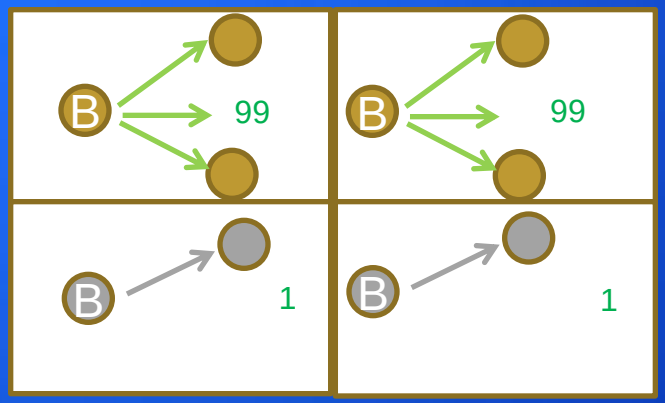
现有的分区方法	PAGraph
负载均衡+通信量	负载均衡+消息流通性（通信量+消息本地流通性）

问题：切割顶点产生了大量的低度副本顶点，导致了大量的通信和计算开销。

低度顶点：



高度顶点：

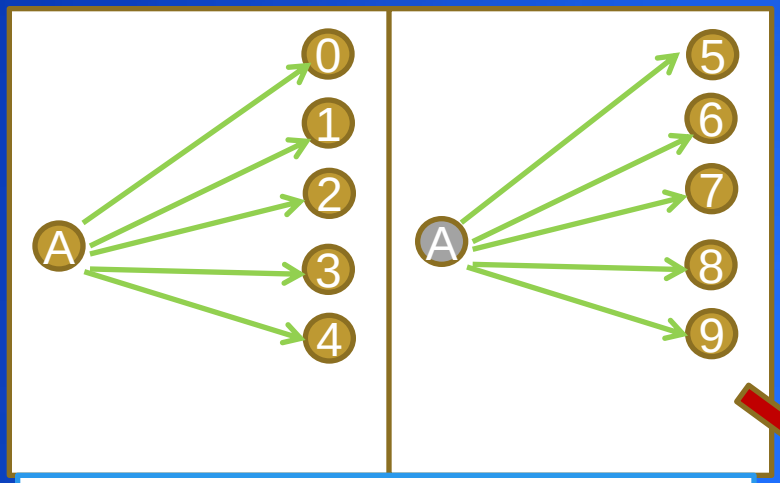


因此，Ginger中采用区分低度高度顶点来切割顶点的方法只是一种模糊的策略。

QC-PAGraph:

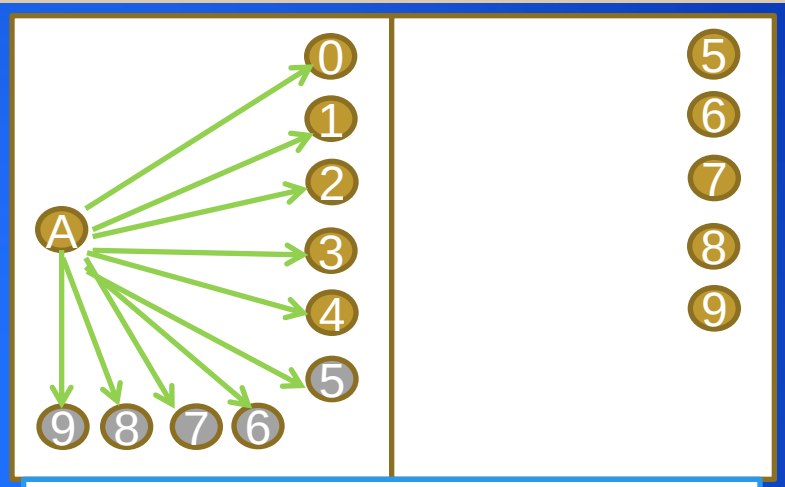
compute(1)=Communication(2)

基于**PAGraph**，通过定量计算通信开销、消息本地流通性与计算开销来决定顶点是否被切割，实现了一个对**PAGraph**改进的图分区方法**QC-PAGraph**。

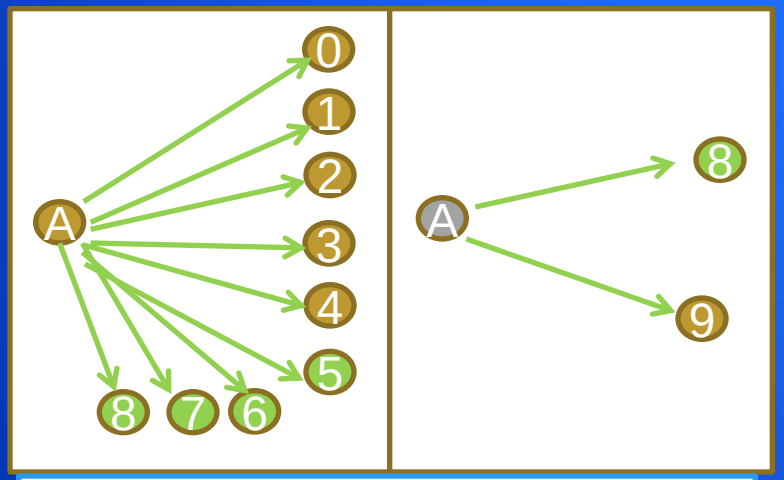


Communication(1)+compute(1)

<

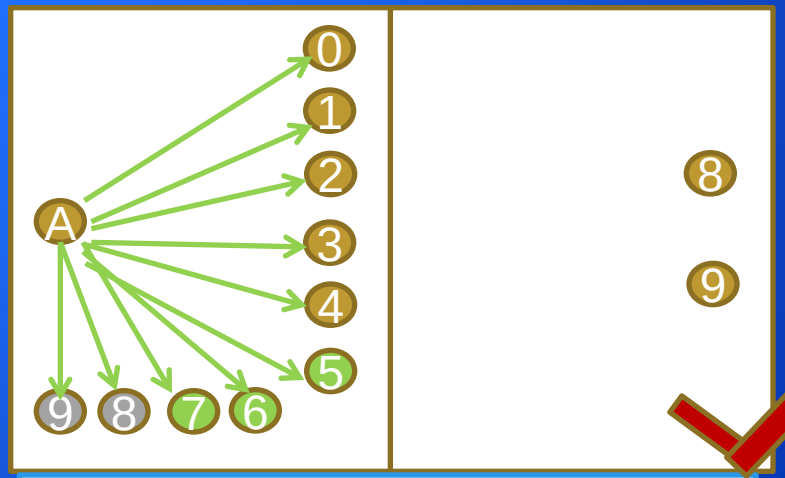


Communication(5)



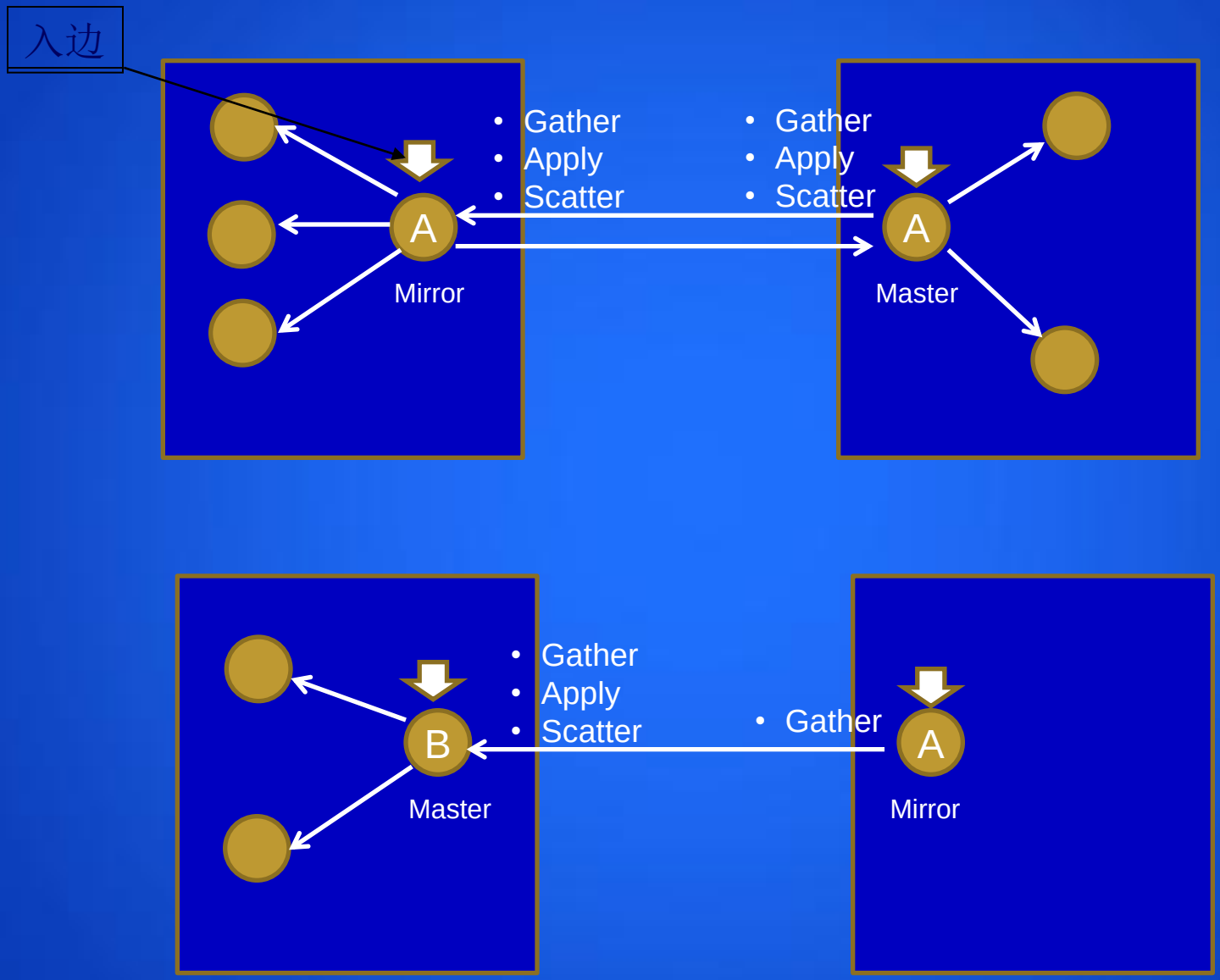
Communication(1)+compute(1)

>



Communication(2)

计算模型:

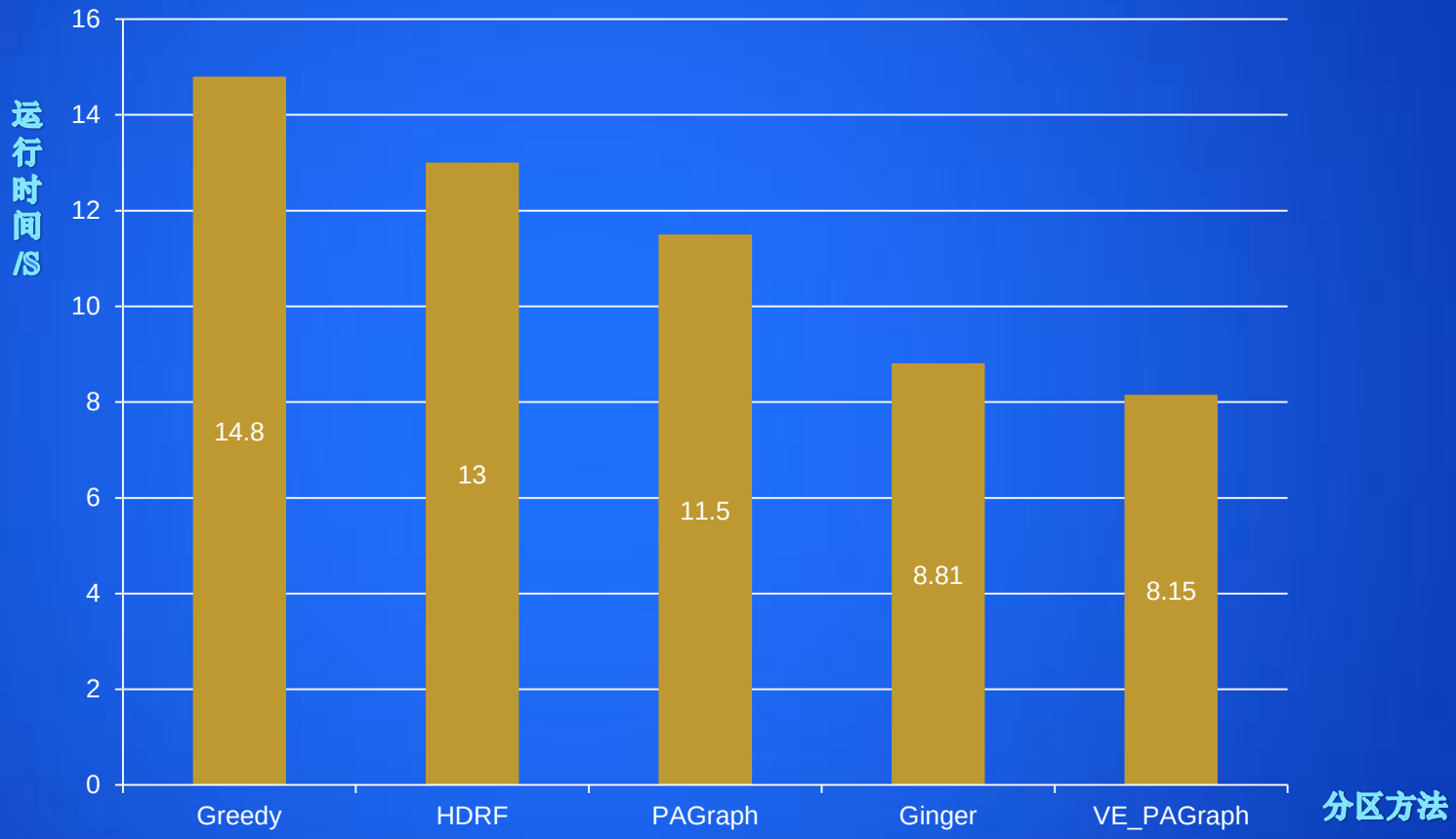

入边
• Gather
• Apply
• Scatter
A
Mirror
• Gather
• Apply
• Scatter
A
Master
• Gather
• Apply
• Scatter
B
Master
• Gather
A
Mirror

实验结果

数据集：
Twitter
顶点=69572
边=7010016

运行环境：
Worker=4

运行算法：
PageRank

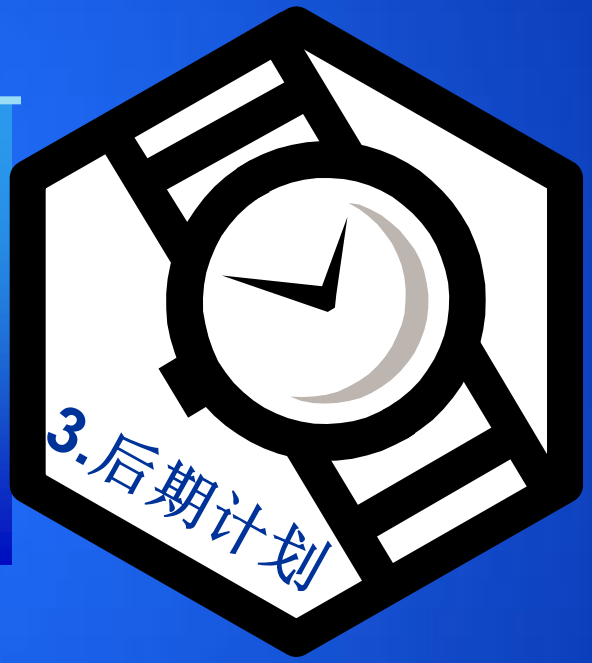


各种分区方法在MR_Maiter上运行时间对比图

说明：因为此数据集较小，高度顶点的影响不会很大，所以Edge_Cut之类的分区方法(Greedy、PAGraph 、HDRF)性能相对于Vertex_Cut的性能差一些。

.1.将本文提出的分区方法分布式实现，融入到分布式图处理框架中。

2. 进行实验分析，对分区方法进行验证和改进优化。



Thank You

各位老师，同学