DFOGraph: An I/O- and Communication-Efficient System for Distributed Fully-out-of-Core Graph Processing

作者： Jiping Yu, Wei Qin, Xiaowei Zhu, Zhenbo Sun, Jianqiang Huang, Xiaohan Li, and Wenguang Chen

Tsinghua University

Abstract

提出了一种图计算系统DFOGraph，是一种分布式**out-of-core图处理系统**

主要贡献创新

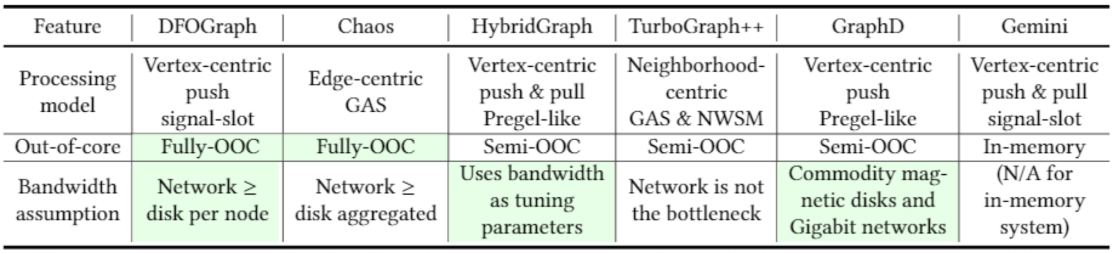
- 结合以顶点为中心的push和两级（节点间和节点内）列分区
- CSR和DCSR的结合使用
- DFOGraph开发了多种自适应通信策略。 与磁盘和网络相关的操作经过仔细分解和流水线处理

Introduction

- **研究动机**：

  单机上的图计算系统：单台计算机的容量和带宽始终限制单节点系统处理极端规模数据集的能力

  分布式的图计算系统：对网络的带宽要求很高，条件太苛刻

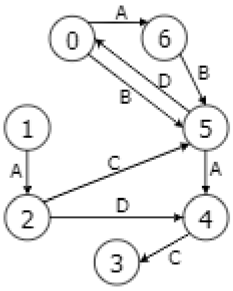


Feature	DFOGraph	Chaos	HybridGraph	TurboGraph++	GraphD	Gemini
Processing model	Vertex-centric push signal-slot	Edge-centric GAS	Vertex-centric push & pull Pregel-like	Neighborhood-centric GAS & NWSM	Vertex-centric push Pregel-like	Vertex-centric push & pull signal-slot
Out-of-core	Fully-OOC	Fully-OOC	Semi-OOC	Semi-OOC	Semi-OOC	In-memory
Bandwidth assumption	Network ≥ disk per node	Network ≥ disk aggregated	Uses bandwidth as tuning parameters	Network is not the bottleneck	Commodity magnetic disks and Gigabit networks	(N/A for in-memory system)

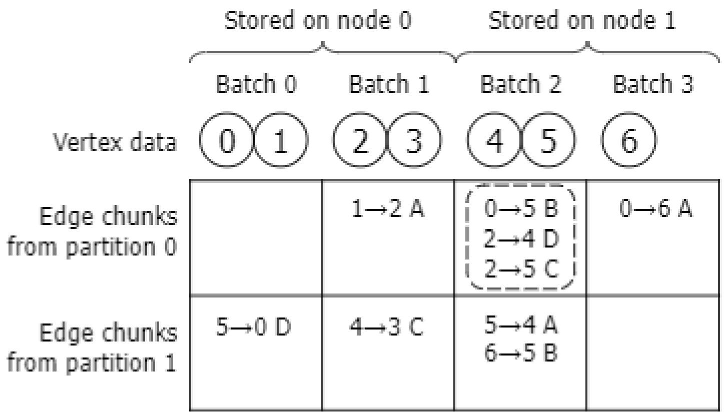
Table 1. Comparison among distributed graph processing systems. Background colors indicate more advanced features.

系统设计

- **Push with Column-Oriented Partitions**
  - 核心外场景中情况中， 同时支持push和pull两种模式会使工作集加倍，占用更多空间，并使缓存的作用降低。 更重要的是，在拉模式下进行过多的外部内存访问要比在推模式下进行同步要昂贵得多。
  - 两级的划分，partition是节点之间、batch是节点内
  - 在划分的时候每一个节点的$(\alpha * |V_i| + |E_i^i| + |E_i^o|)$ 要尽可能相等

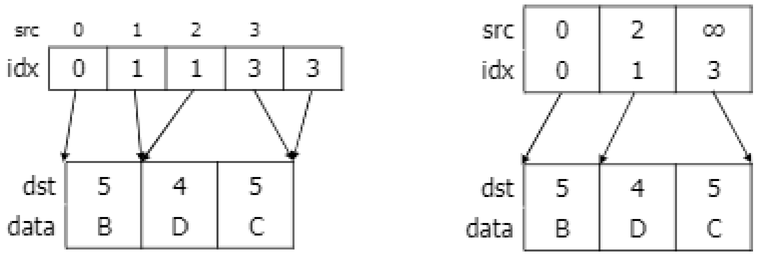


(a) A graph of seven vertices and nine edges. Edge data is a letter.

- 

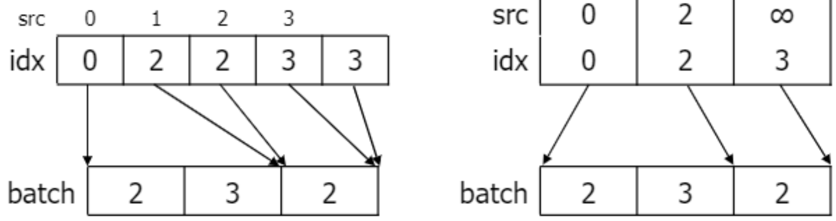
	Batch 0	Batch 1	Batch 2	Batch 3
	Stored on node 0		Stored on node 1	
Vertex data	0 1	2 3	4 5	6
Edge chunks from partition 0		1→2 A	0→5 B 2→4 D 2→5 C	0→6 A
Edge chunks from partition 1	5→0 D	4→3 C	5→4 A 6→5 B	

  (b) Vertex and edge storage on two nodes. Vertex batch size is 2.

- 

  (c) CSR of circled chunk in (b) (d) DCSR of circled chunk in (b)

- Messages from … should go to …
  0→Batch 2
  0→Batch 3
  2→Batch 2

  (e) Edges, CSR, and DCSR of dispatching graph from node 0 to 1.

- **通信过程**

  四步走：generating、Inter-node passing、intra-node dispatching、processing

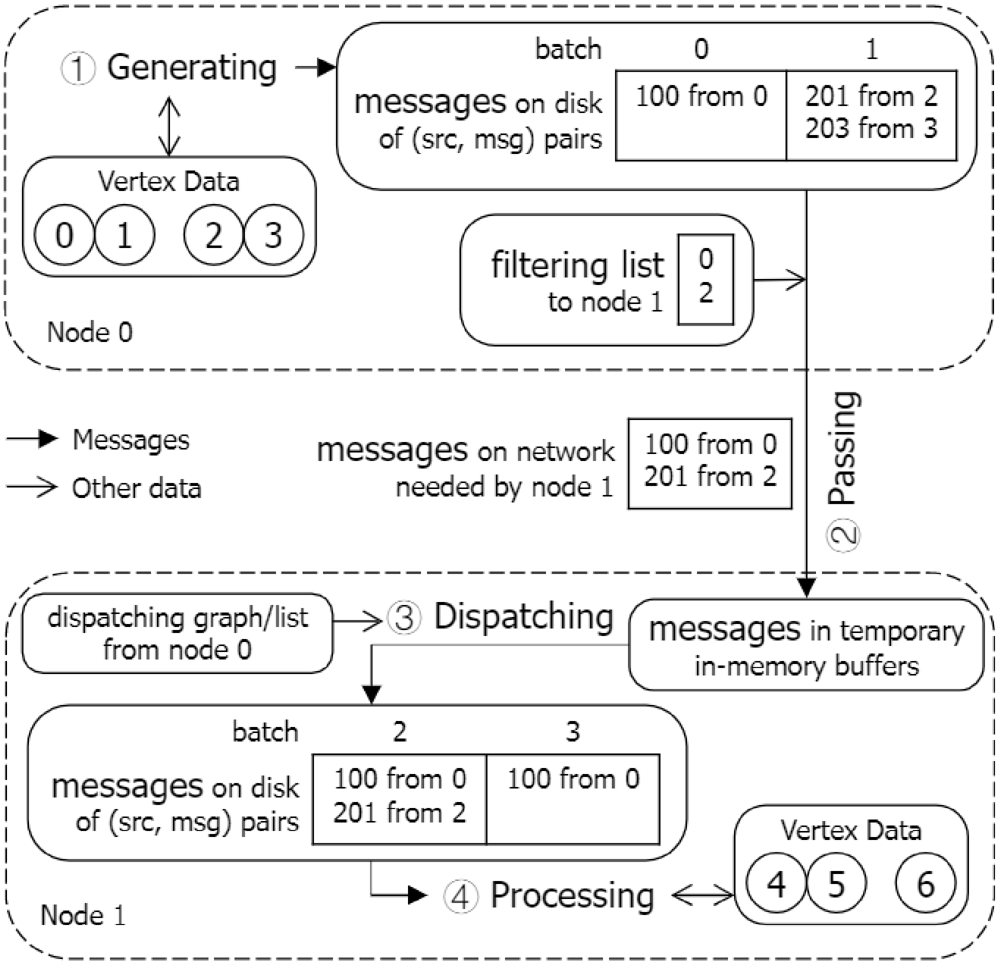
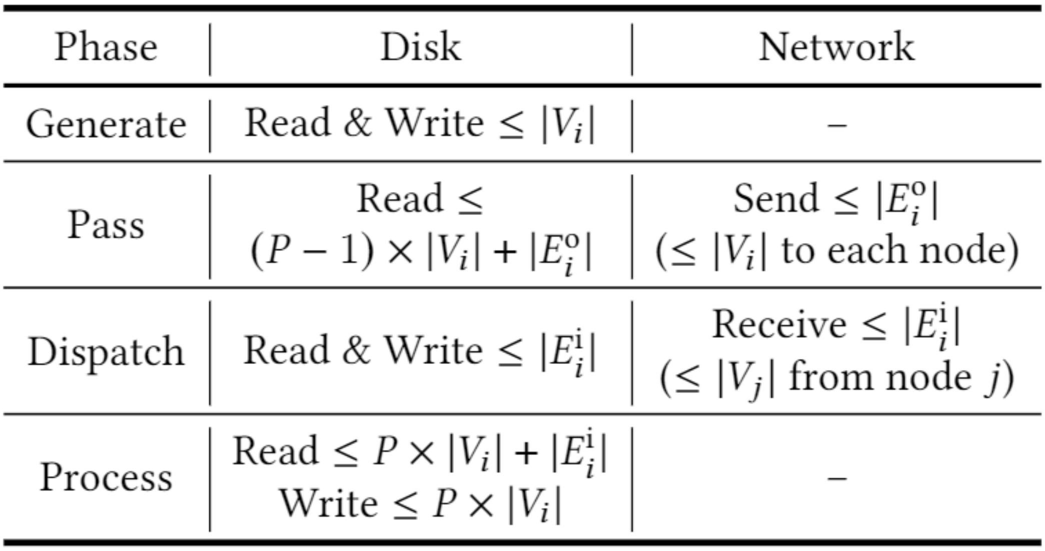


Figure 3. Data flow of ProcessEdges from node 0 to 1.

- **I/O 和通信优化**
  - 自适应的CSR和DCSR表示，DCSR的开销$2 * |V_{src,outdeg!=0}|$ 而CSR的开销就是$min(\gamma * |M|, |V_{src}|)$
  - 消息调度的自适应策略:Push dispatching、Pull dispatching
  - 在节点之间的消息筛选
  - 使用流水线作业



Phase	Disk	Network
Generate	Read & Write $\leq \lvert V_i \rvert$	–
Pass	Read $\leq$ $(P-1) \times \lvert V_i \rvert + \lvert E_i^{\mathrm{o}} \rvert$	Send $\leq \lvert E_i^{\mathrm{o}} \rvert$ ($\leq \lvert V_i \rvert$ to each node)
Dispatch	Read & Write $\leq \lvert E_i^{\mathrm{i}} \rvert$	Receive $\leq \lvert E_i^{\mathrm{i}} \rvert$ ($\leq \lvert V_j \rvert$ from node j)
Process	Read $\leq P \times \lvert V_i \rvert + \lvert E_i^{\mathrm{i}} \rvert$ Write $\leq P \times \lvert V_i \rvert$	–

Table 2. I/O and Communication amount in each phase of `ProcessEdges` API on node i.

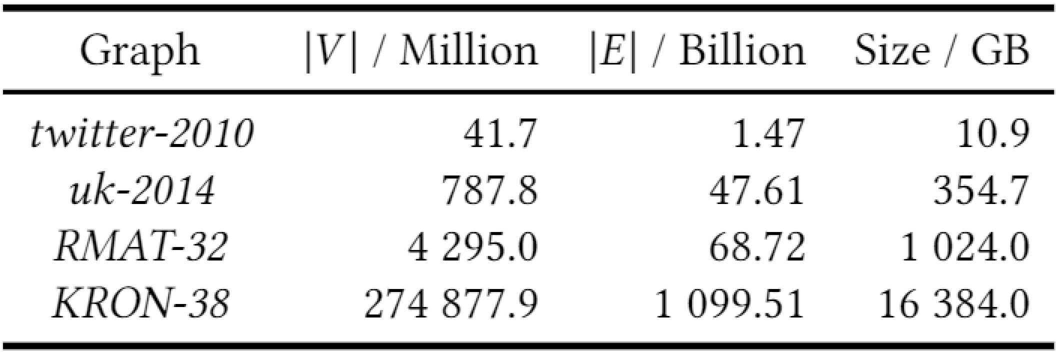
Experiments

环境

AWS EC2 i3en.3x大型实例，每个实例均配备12个线程的Intel Xeon Plat-inum 8175M（每个内核2个线程，基本频率2.50 GHz）

93.2 GB RAM，25 Gbps网络，

Datasets



Graph	$\lvert V \rvert$ / Million	$\lvert E \rvert$ / Billion	Size / GB
twitter-2010	41.7	1.47	10.9
uk-2014	787.8	47.61	354.7
RMAT-32	4 295.0	68.72	1 024.0
KRON-38	274 877.9	1 099.51	16 384.0

Table 3. Graph datasets for experiments – size calculated as (source, destination) pair in binary formats of each edge.

Methodology

PT(Basic Partitioning-based Approach):基于基本分区的方法

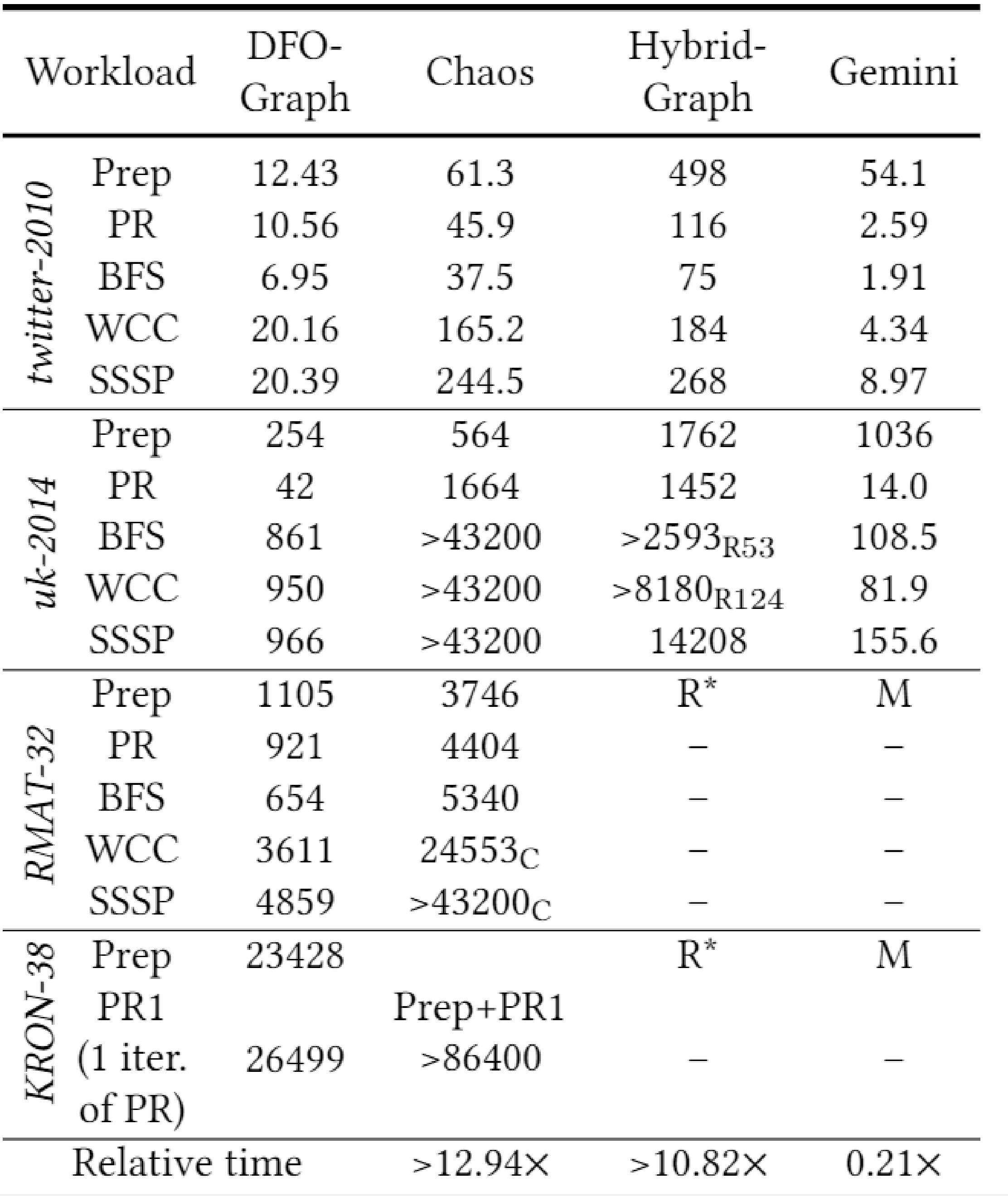
PT-Opt(Optimized Partitioning-based Approach):基于分区的优化方法

(UM-Opt)Optimized Unified Memory-based Approach:优化的基于统一内存的方法

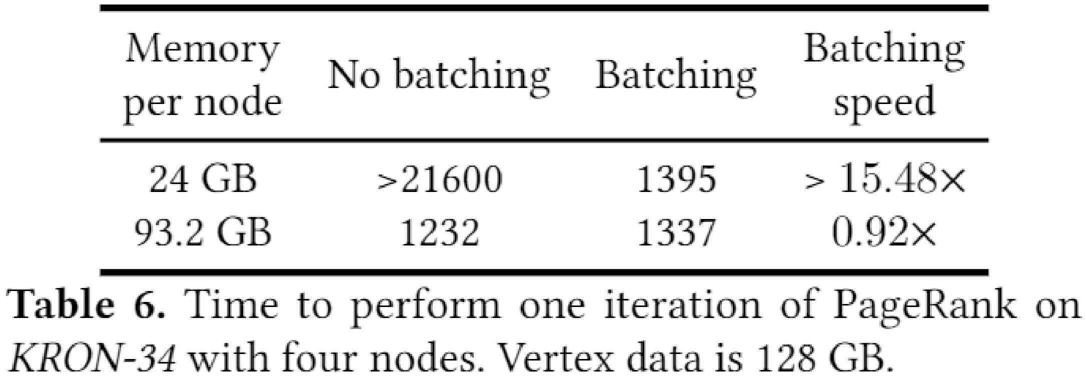
效果

-

	Workload	DFOGraph	GridGraph	FlashGraph
twitter-2010	Prep	31.99	62.75	618.29
	PR	46.77	34.18	D
	BFS	8.60	9.94	10.33
	WCC	42.48	10.38	D
	SSSP	48.11	29.46	D
uk-2014	Prep	1508	3178	M
	PR	804	1569	1235*
	BFS	870	>43200	556*
	WCC	3590	>43200	D*
	SSSP	3906	>43200	D*
Relative time			>2.52×	1.06×



	Workload	DFO-Graph	Chaos	Hybrid-Graph	Gemini
twitter-2010	Prep	12.43	61.3	498	54.1
	PR	10.56	45.9	116	2.59
	BFS	6.95	37.5	75	1.91
	WCC	20.16	165.2	184	4.34
	SSSP	20.39	244.5	268	8.97
uk-2014	Prep	254	564	1762	1036
	PR	42	1664	1452	14.0
	BFS	861	>43200	$>2593_{\text{R53}}$	108.5
	WCC	950	>43200	$>8180_{\text{R124}}$	81.9
	SSSP	966	>43200	14208	155.6
RMAT-32	Prep	1105	3746	R*	M
	PR	921	4404	–	–
	BFS	654	5340	–	–
	WCC	3611	24553_{C}	–	–
	SSSP	4859	$>43200_{\text{C}}$	–	–
KRON-38	Prep	23428		R*	M
	PR1 (1 iter. of PR)	26499	Prep+PR1 >86400	–	–
	Relative time		>12.94×	>10.82×	0.21×

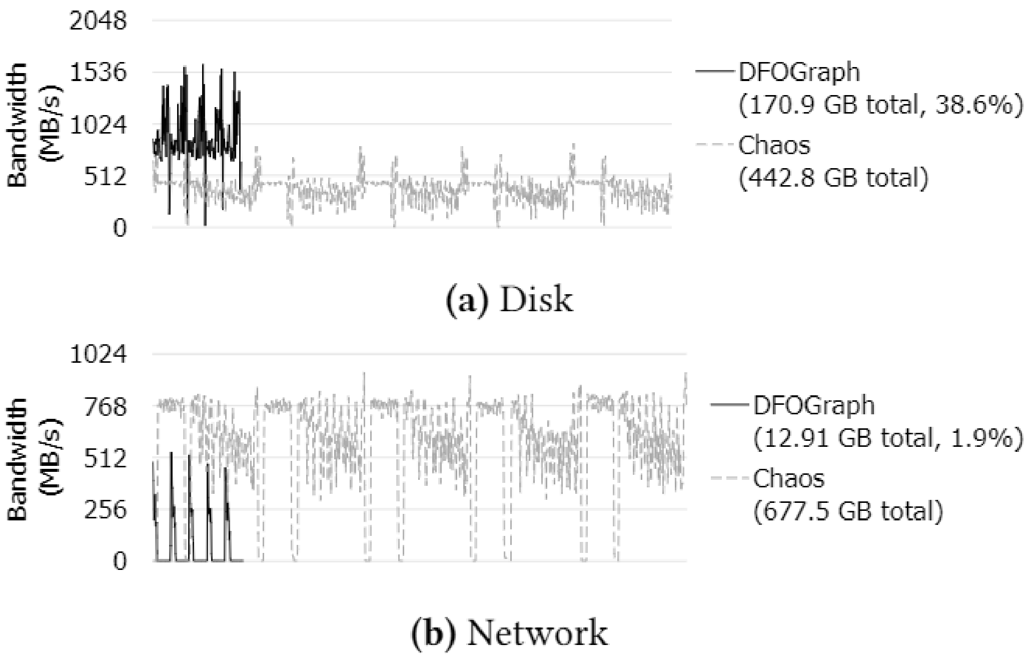
- 

Memory per node	No batching	Batching	Batching speed
24 GB	>21600	1395	> 15.48×
93.2 GB	1232	1337	0.92×

Table 6. Time to perform one iteration of PageRank on *KRON-34* with four nodes. Vertex data is 128 GB.

在没有足够的内存下，不使用批处理会导致随机访问的范围变大，从而页置换次数增加，效率下降

-



(a) Disk

(b) Network

和Chaos相比，该方法的网络和磁盘带框需求很小

Conclusion and Future Works

提出了DFOGraph，这是一种分布式的完全核外图处理系统。

DFOGraph应用了以顶点为中心的推送计算和两级面向列的图分区策略， DFOGraph采用CSR和DCSR的混合图形表示形式，从而减小了数据大小并实现了以细粒度的以顶点为中心的边缘访问。

DFOGraph实现了自适应消息处理策略，流水线计算以及磁盘/网络操作，以平衡CPU，网络和存储。

通过应用所有这些技术，DFOGraph可以优化I / O和通信效率，并避免不必要的磁盘和网络操作并实现可扩展性和容量。

实验表明，DFOGraph明显优于其他分布式核外系统（如Chaos），并且可与单机核外系统GridGraph和FlashGraph相提并论。