基于新型硬件体系的图计算系统栈

陈榕

并行与分布式系统研究所

上海交通大学

Alipay | in.clusion FINTECH CONFERENCE 外滩大会

Joint work with Haibo, Xinda, Jiaxin, Yanzhe, and Wukong team@SJTU,
and the Wukong work is also with Fefei@Alibaba

图的力量

Alipay | in.clusion FINTECH CONFERENCE 外滩大会

- 一种自然、普遍的数据表达

社交网络

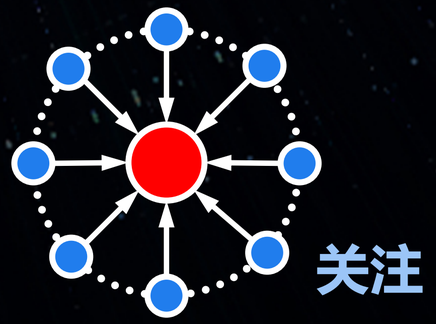
交通网络

金融图谱

万物皆图

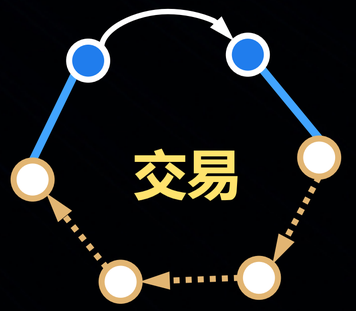
图的力量

- 一种自然、普遍的数据表达
- **能够描述数据间的内在关系**

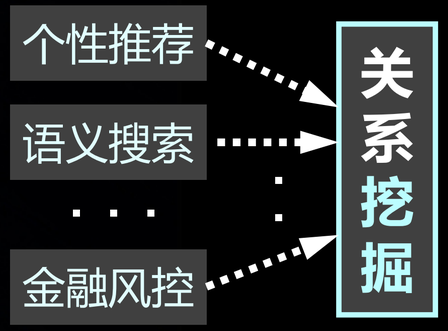


社交网络

交通网络



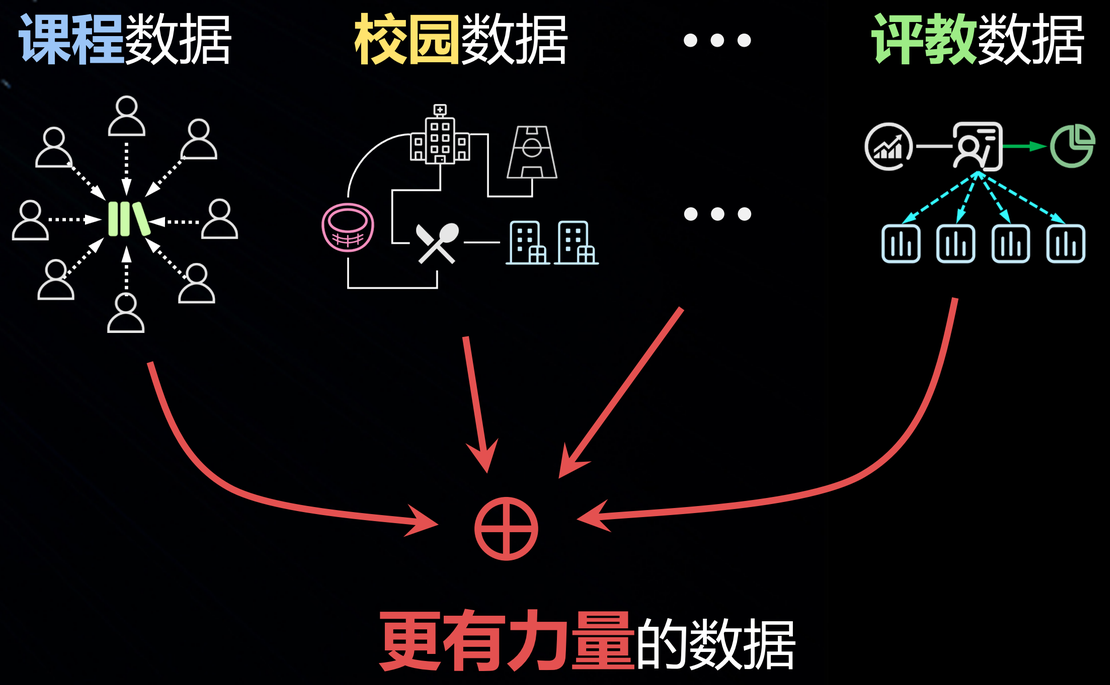
金融图谱



万物皆图

图的力量

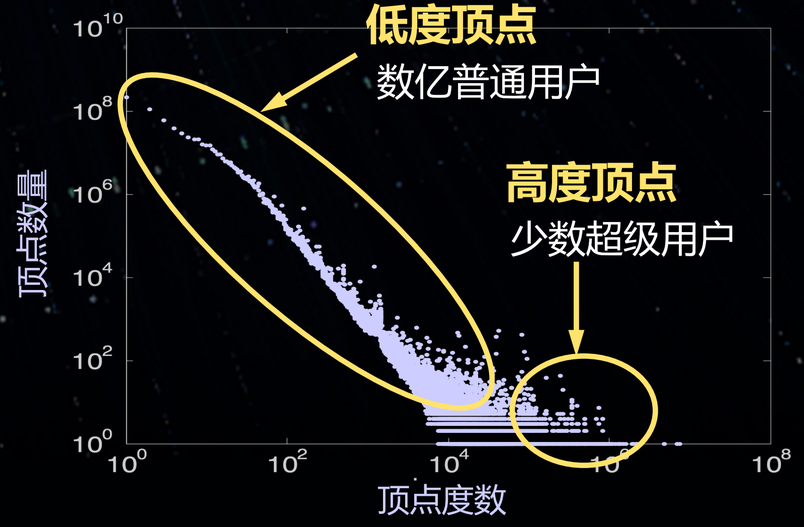
- 一种自然、普遍的数据表达
- 能够描述数据间的内在关系
- **适合于支持跨领域数据融合**



图数据处理很难

- 数据特征丰富

① 幂律性



5.5亿
活跃用户

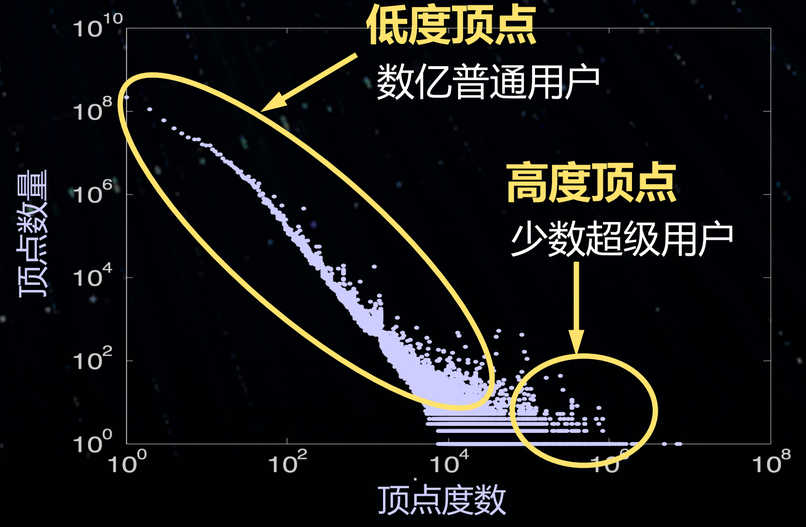
用户	粉丝数
	1.27亿
	1,174

图数据处理很难

Alipay | in.clusion FINTECH CONFERENCE 外滩大会

▶ 数据特征丰富

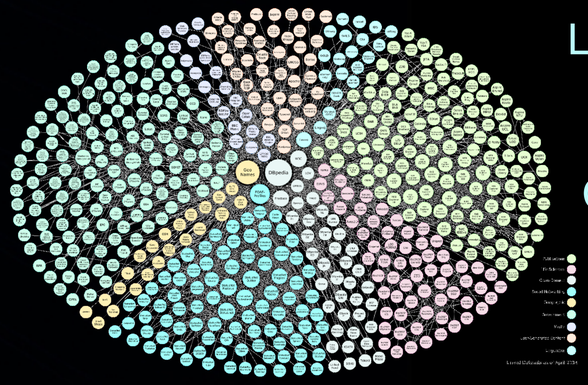
① 幂律性



5.5亿
活跃用户

用户	粉丝数
	1.27亿
	1,174

② 异质性



Linked
Open
Data
Cloud

③ 动态热点

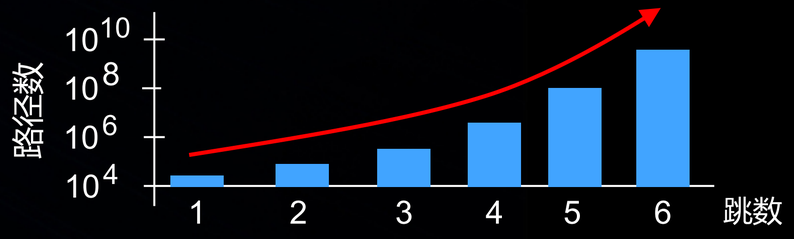
Facebook World Network

图数据处理很难

- 数据特征丰富
- **处理复杂度高**

① 计算：

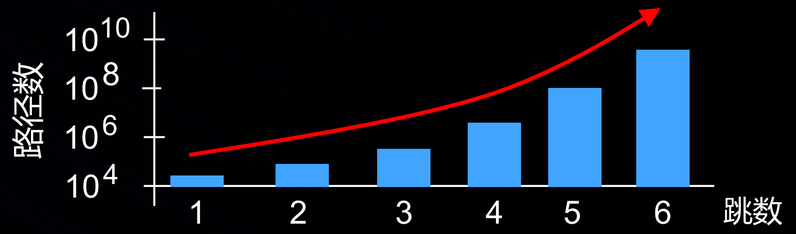
1次6跳环路检测在**519M顶点图**上平均需要遍历超过**20亿**条路径



图数据处理很难

- 数据特征丰富
- **处理复杂度高**

① 计算：1次6跳环路检测在**519M顶点图**上平均需要遍历超过**20亿**条路径



② 存储：*Random Walk*在**41M顶点图**上需要预构建**980TB**索引

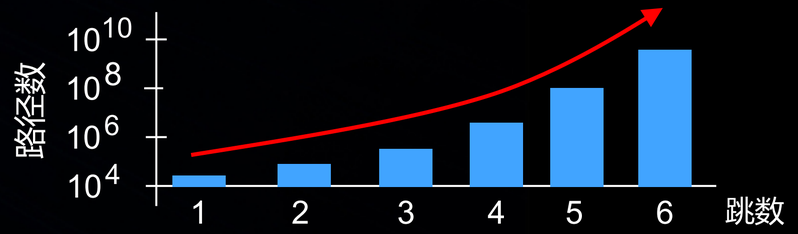
Motifs-4在**5M顶点图**上的中间结果可能超过10^{12}子图

爆炸

图数据处理很难

- 数据特征丰富
- **处理复杂度高**

① 计算：1次6跳环路检测在**519M顶点图**上平均需要遍历超过**20亿**条路径



② 存储：*Random Walk*在**41M顶点图**上需要预构建**980TB**索引

Motifs-4在**5M顶点图**上的中间结果可能超过10^{12}子图

爆炸

③ 类型：

图分析	*PR, SSSp, Random Walk, …*
图查询	*SPARQL, Gremlin, Cypher, …*
图挖掘	*Motifs, FSM, Cliques, …*
图神经网络	*GCN, MAGNN, PinSage, …*
图数据库	*CRUD, Transaction*

图数据处理很难

- 数据特征丰富
- 处理复杂度高
- **实时并发需求**

① 毫秒级时延需求

10毫秒	20毫秒	100毫秒
搜索	**金融**	**广告**
知识图谱	欺诈检测	个性推荐

图数据处理很难

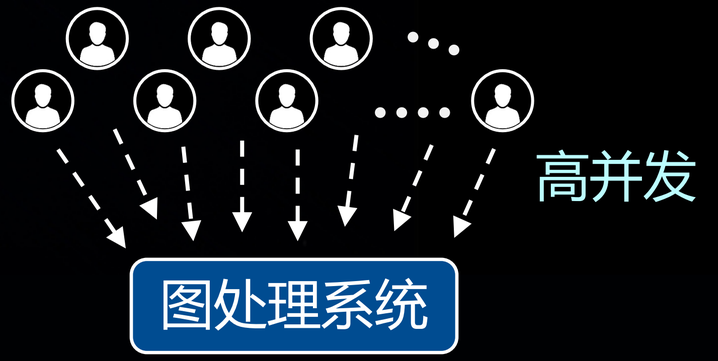
- 数据特征丰富
- 处理复杂度高
- **实时并发需求**

① 毫秒级时延需求

10毫秒	20毫秒	100毫秒
搜索	**金融**	**广告**
知识图谱	欺诈检测	个性推荐

② 相比单一任务时延，高并发场景下的时延（尾时延）更有意义且更具挑战

语义搜索　互联网金融　个性推荐　…



硬件发展：机遇与挑战

Alipay | in.clusion FINTECH CONFERENCE 外滩大会

► 机遇：新型硬件快速普及

✓ 新型硬件对系统性能有显著收益

RDMA	HTM	NVM	GPU
 远程直接内访问存	 硬件事务内存	 非易失内存	 图型处理单元
高带宽、低时延	高效支持ACI	数据持久化	高并发算力

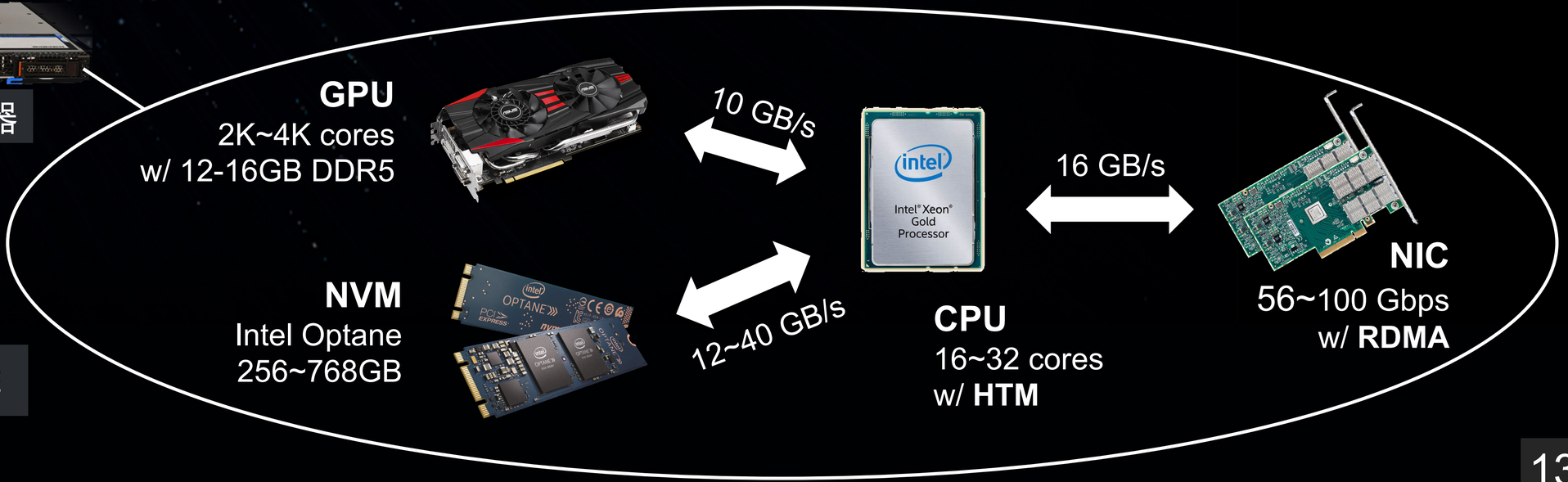
硬件发展：机遇与挑战

Alipay | in.clusion FINTECH CONFERENCE 外滩大会

► 机遇：新型硬件快速普及

✓ 新型硬件对系统性能有显著收益

RDMA	HTM	NVM	GPU
远程直接内访问存	硬件事务内存	非易失内存	图型处理单元
高带宽、低时延	高效支持ACI	数据持久化	高并发算力



硬件发展：机遇与挑战

- 机遇：新型硬件快速普及
- **挑战：功能单一、使用受限**

✓ 新型硬件对系统性能有显著收益

RDMA	HTM	NVM	GPU
 远程直接内访问存	 硬件事务内存	非易失内存	 图型处理单元
高带宽、低时延	高效支持ACI	数据持久化	高并发算力

✗ 但往往瞄准单一需求且使用受限

RDMA

提供的低时延跨节点内存读写操作无法保证多个访存操作的原子性和一致性、以及写入数据的持久性

HTM

只能在节点内为多个访存操作提供原子性

NVM

数据持久性受限在单节点内，且依赖特定访存指令

GPU

跨节点数据传输依赖CPU和系统内存

我们的思路：软硬件协同

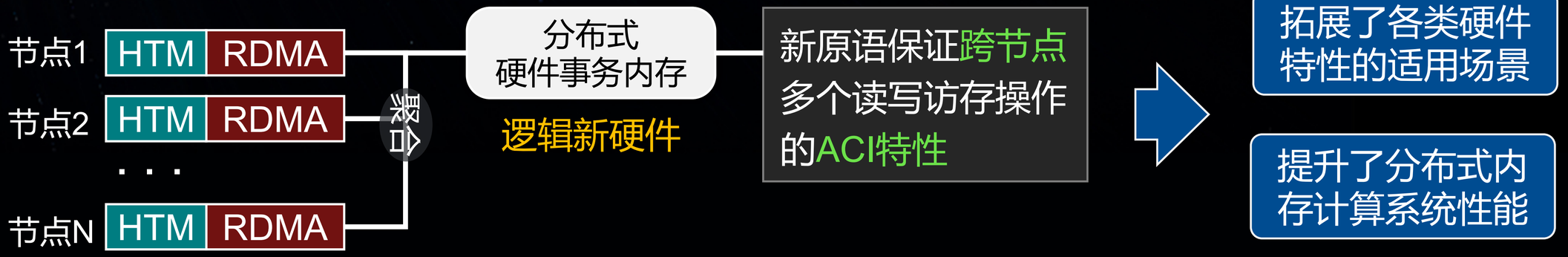
- **硬件：分布式异构硬件聚合方法**
  - ① 基础：硬件原语 (primitive)
  - ② 以RDMA为核心，聚合多种不同硬件原语构建新的“逻辑硬件”

硬件原语	
RDMA	READ | WRITE | CAS, SEND | RECV, …
HTM	xbegin | xend | xabort | …
NVM	clwb | sfence | pcommit | nt-store | ...
GPU	load | store | … (on 128-512 bits, SIMD)

我们的思路：软硬件协同

- **硬件：分布式异构硬件聚合方法**
  - ① 基础：硬件原语 (primitive)
  - ② 以RDMA为核心，聚合多种不同硬件原语构建新的“逻辑硬件”

硬件原语	
RDMA	READ | WRITE | CAS, SEND | RECV, …
HTM	xbegin | xend | xabort | …
NVM	clwb | sfence | pcommit | nt-store | ...
GPU	load | store | … (on 128-512 bits, SIMD)



我们的思路：软硬件协同

- 硬件：分布式异构硬件聚合方法
- **软件：分布式内存图计算系统栈**

① 图数据的规模需求

十亿级顶点
百亿级边 

16节点集群
TB级内存 ✓

② 图应用的时延需求

~~磁盘 / SSD I/O时延~~

100ns 访存
1-2us 网络

我们的思路：软硬件协同

- 硬件：分布式异构硬件聚合方法
- **软件：分布式内存图计算系统栈**

① 图数据的规模需求

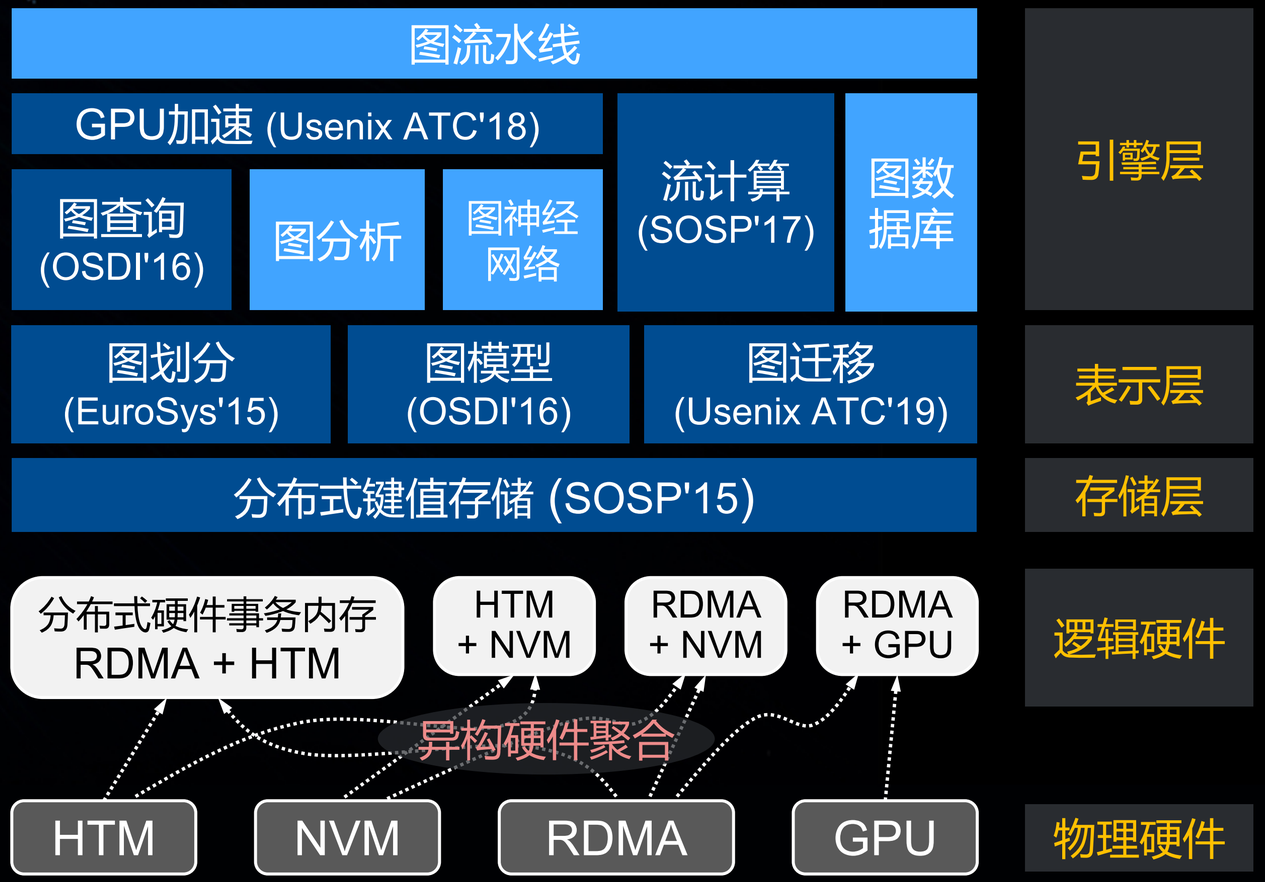
十亿级顶点
百亿级边

16节点集群
TB级内存 ✓

② 图应用的时延需求

~~磁盘 / SSD I/O时延~~

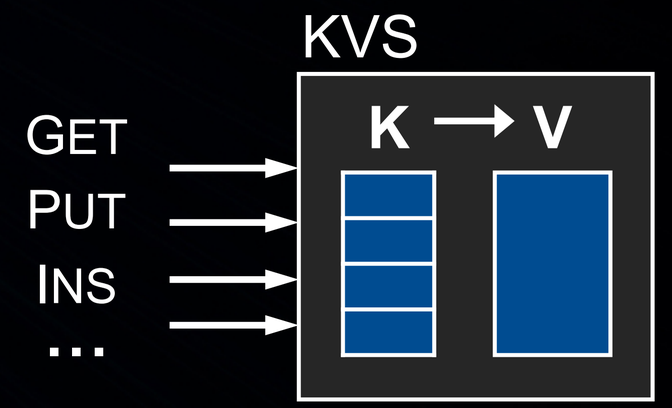
100ns 访存
1-2us 网络

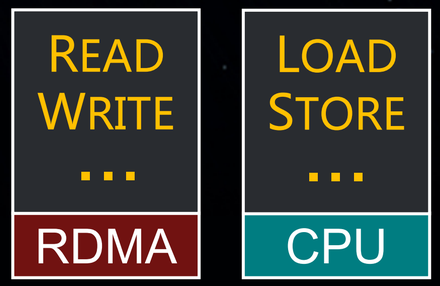


存储层

- **键值存储 (Key-Value Store, KVS)**
  - ① 高可扩展的架构 (scale-out)
  - ② 接近硬件原语的接口

    GET | PUT | INS | DEL | SCAN





存储层

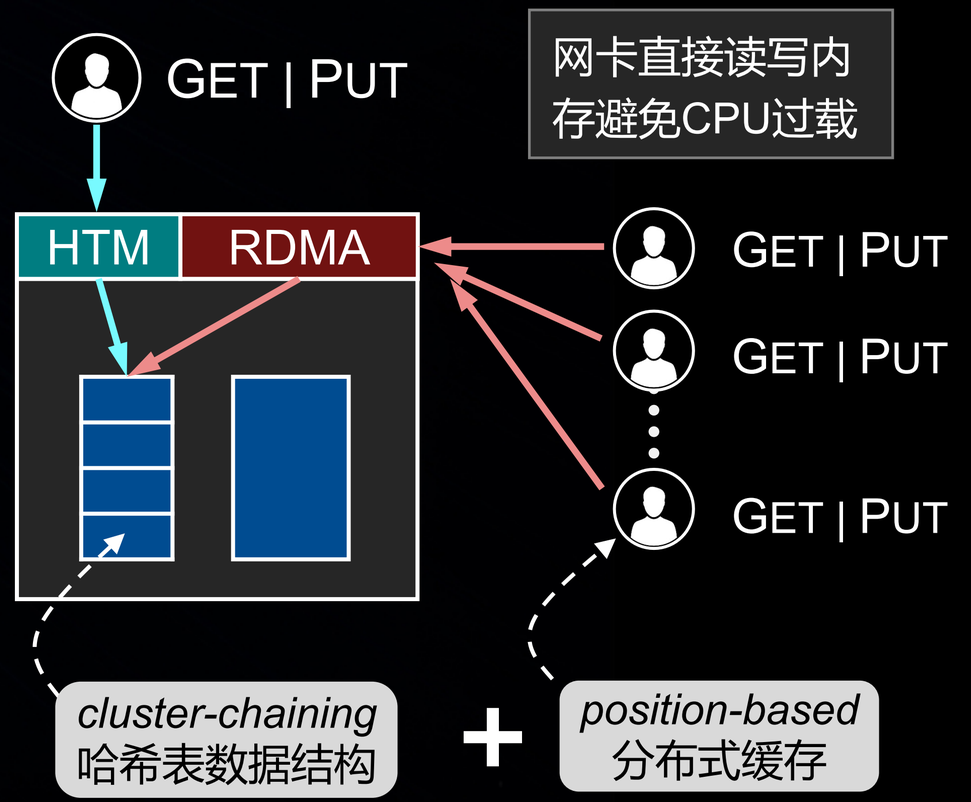
- 键值存储 (Key-Value Store, KVS)
- **利用新硬件加速KVS**

DrTM-KV (SOSP'15)

① 分布式硬件事务内存 (RDMA+HTM)

② 新型数据结构和缓存技术

HTM实现高效
并发访存控制



存储层

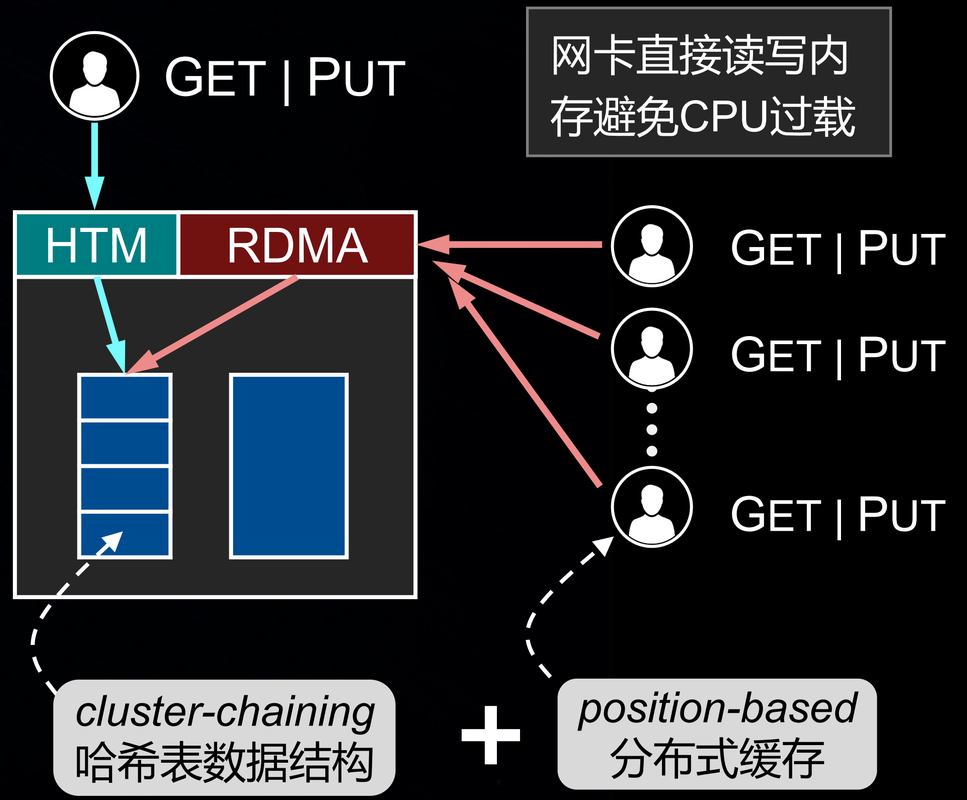
- 键值存储 (Key-Value Store, KVS)
- **利用新硬件加速KVS**

DrTM-KV (SOSP'15)

① 分布式硬件事务内存 (RDMA+HTM)

② 新型数据结构和缓存技术

HTM实现高效
并发访存控制

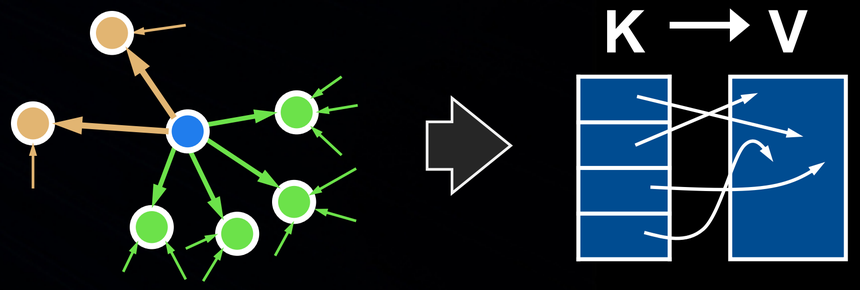


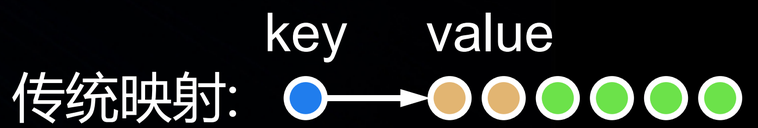
	吞吐量/Mops		平均时延/us	
	GET	PUT	GET	PUT
传统设计 Memcached	1.5	1.5	50	50
RDMA设计/微软 FaRM	6	3	5	10
DrTM-KV	**115**	**14**	**3**	**6**

数量级性能提升

表示层：图模型

- **问题：“图数据”到“键值存储”的高效映射**
  - ① 传统方法：key=顶点，value=邻接顶点
  - ② 访存开销大：遍历邻接顶点
  - ③ 网络传输多(RDMA)：远程读取所有顶点



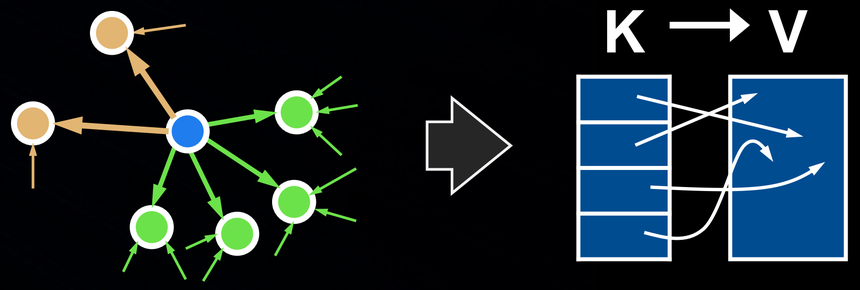


访存开销大、网络传输多

表示层：图模型

- 问题："图数据"到"键值存储"的高效映射
- **顶点解耦技术** Wukong OSDI'16

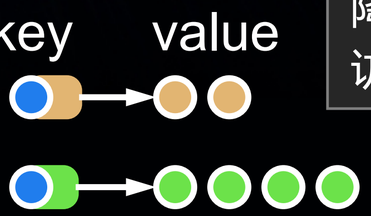
① 观察：大部分任务只需要某一类邻接顶点



传统映射：

访存开销大、网络传输多

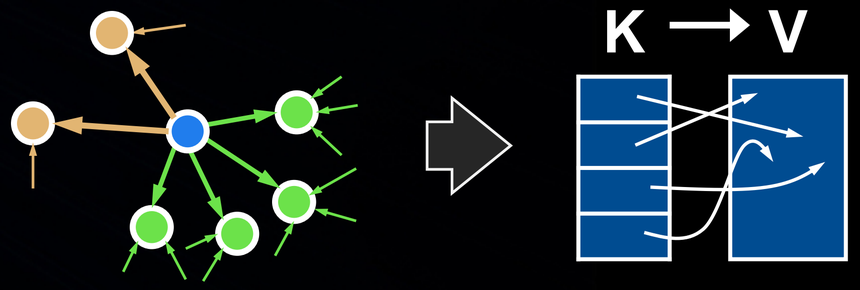
解耦映射：

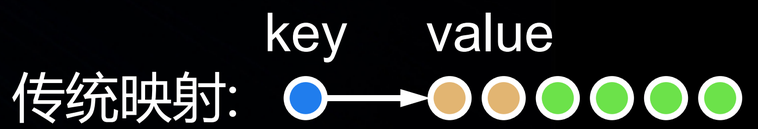


降低RDMA远程访存数据传输量

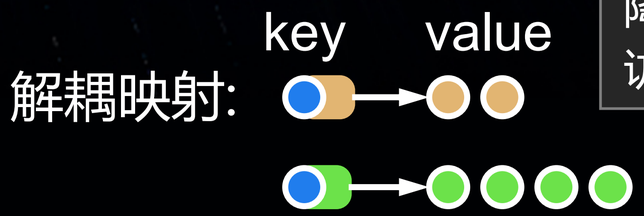
表示层：图模型

- 问题："图数据"到"键值存储"的高效映射
- **顶点解耦技术** Wukong OSDI'16
  - ① 观察：大部分任务只需要某一类邻接顶点
  - ② 降低存储空间开销：同类键压缩

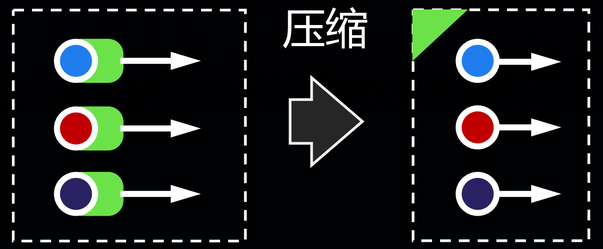




访存开销大、网络传输多

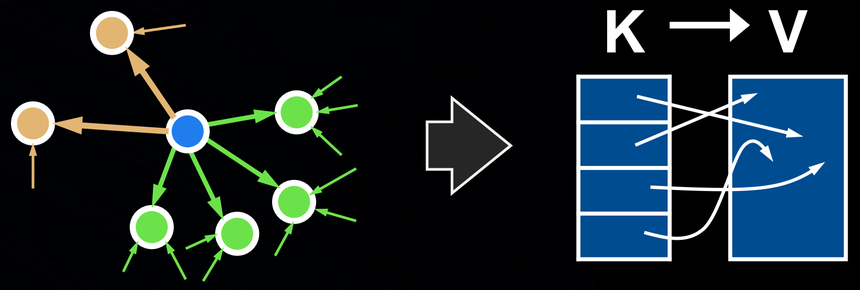


降低RDMA远程访存数据传输量



表示层：图模型

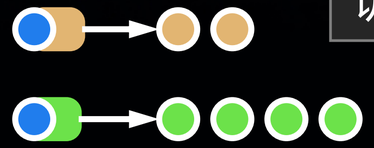
- 问题："图数据"到"键值存储"的高效映射
- **顶点解耦技术** Wukong OSDI'16
  - ① 观察：大部分任务只需要某一类邻接顶点
  - ② 降低存储空间开销：同类键压缩
  - ③ 访问所有邻接顶点：邻接顶点连续存储



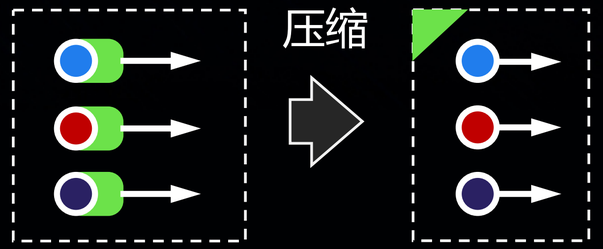
传统映射: key value

访存开销大、网络传输多

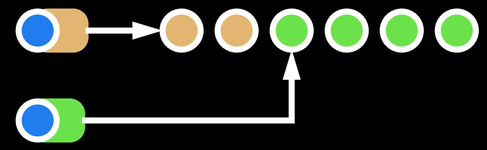
解耦映射: key value



降低RDMA远程访存数据传输量

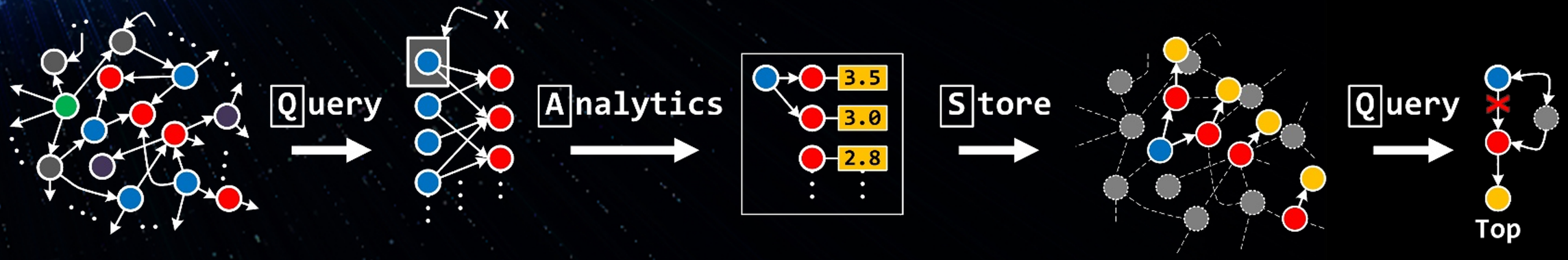


连续存储 (value)



引擎层：图流水线

图流水线应用：个性化推荐

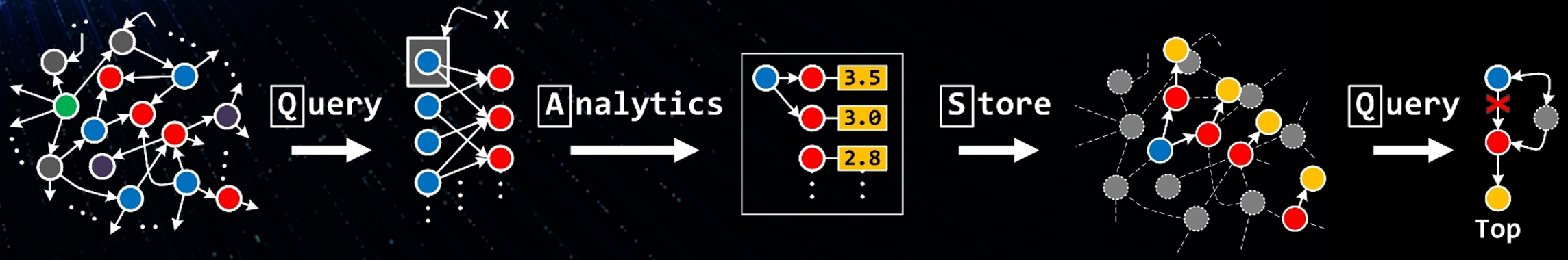


- **图应用：由多种类型图处理任务构成的流水线 (Pipeline)**

① 任务类型多样：图查询、图分析、图挖掘、图神经网络 ...

引擎层：图流水线

图流水线应用：个性化推荐



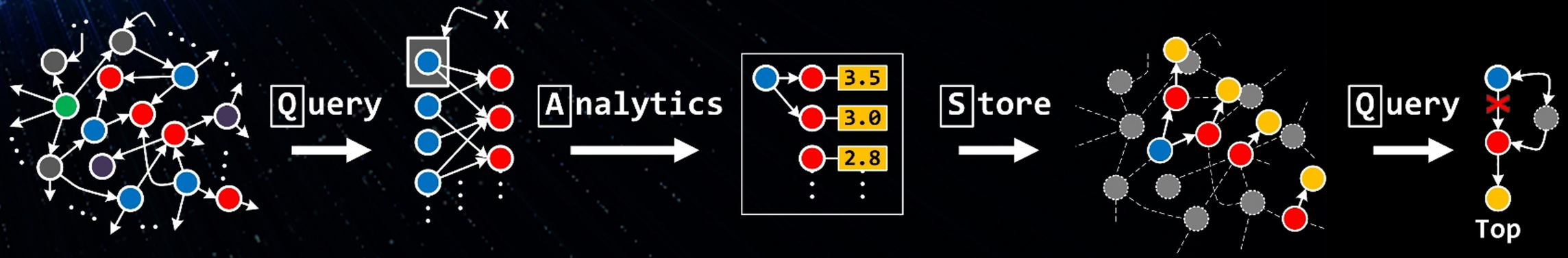
- **图应用：由多种类型图处理任务构成的流水线 (Pipeline)**
  - ① 任务类型多样：图查询、图分析、图挖掘、图神经网络 ...
  - ② 现有研究不足：单点优化，缺乏整体性考量

引擎层：图流水线

	Q	ID	<>	A	<>	ID	S	Q	TOT
WK & PL	2.5		1.3	5.4	0.1		0.2	0.1	**9.6**
WK & GE	2.5	4.5	1.3	1.9	0.1	0.4	0.2	0.1	**11.0**

a 4-node cluster, RDF: |V|=21M, |E|=88M, A: PPR
WK=Wukong, PL=PowerLyra, GE=GEMINI

图流水线应用：个性化推荐



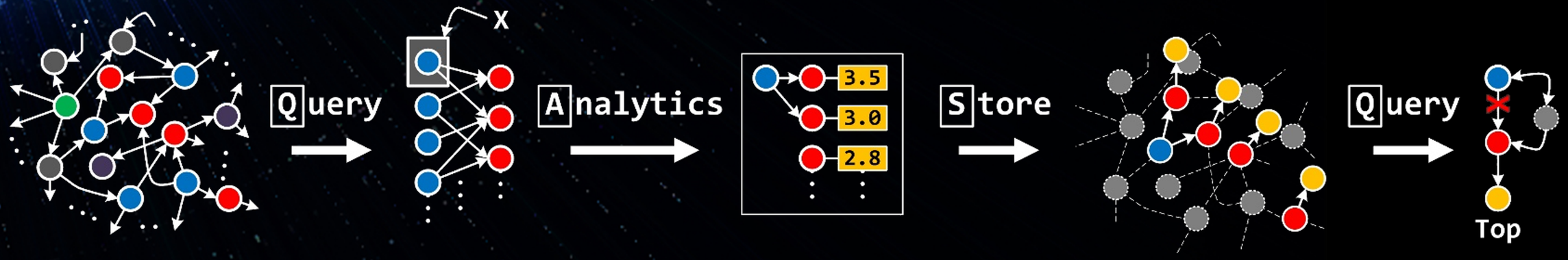
- **图应用：由多种类型图处理任务构成的流水线 (Pipeline)**
  - ① 任务类型多样：图查询、图分析、图挖掘、图神经网络 ...
  - ② 现有研究不足：单点优化，缺乏整体性考量

引擎层：图流水线

	Q	ID	<>	A	<>	ID	S	Q	TOT
WK & PL	2.5		1.3	5.4	0.1		0.2	0.1	9.6
WK & GE	2.5	4.5	1.3	1.9	0.1	0.4	0.2	0.1	11.0

a 4-node cluster, RDF: |V|=21M, |E|=88M, A: PPR
WK=Wukong, PL=PowerLyra, GE=GEMINI

图流水线应用：个性化推荐



- **图应用：由多种类型图处理任务构成的流水线 (Pipeline)**
  1. 任务类型多样：图查询、图分析、图挖掘、图神经网络 ...
  2. 现有研究不足：单点优化，缺乏整体性考量
  3. 存储结构不兼容，任务衔接开销大 (58%)

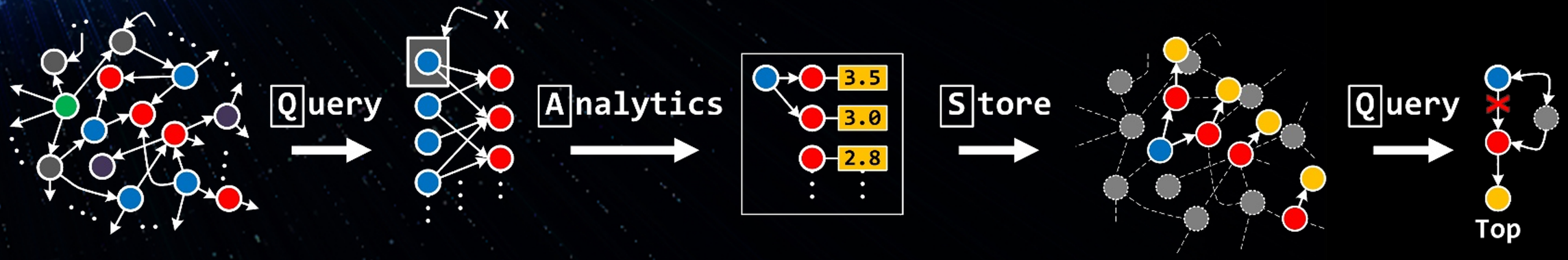
	关键数据结构
A 图分析	CSR, CSC
Q 图查询	Triple Table, KVS
M 图挖掘	ODAG, …
N 图神经网络	Array, Matrix, …
S 图存储	Adjacent List, KVS

引擎层：图流水线

	Q	ID	<>	A	<>	ID	S	Q	TOT
WK & PL	2.5		1.3	5.4	0.1		0.2	0.1	**9.6**
WK & GE	2.5	4.5	1.3	1.9	0.1	0.4	0.2	0.1	**11.0**
Spark	12.2			34.0			1.6	0.2	**48.0**

a 4-node cluster, RDF: |V|=21M, |E|=88M, A: PPR
WK=Wukong, PL=PowerLyra, GE=GEMINI

图流水线应用：个性化推荐

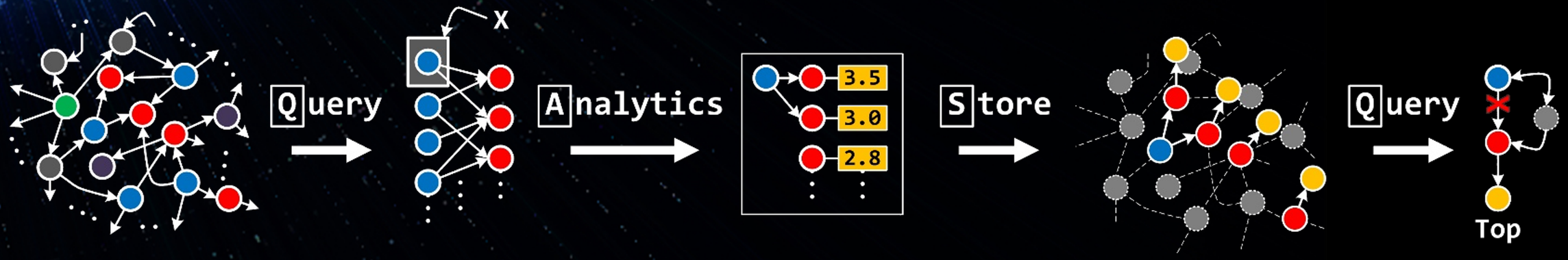


- **图应用：由多种类型图处理任务构成的流水线 (Pipeline)**
  1. 任务类型多样：图查询、图分析、图挖掘、图神经网络 ...
  2. 现有研究不足：单点优化，缺乏整体性考量
  3. 存储结构不兼容，任务衔接开销大 (58%)

		关键数据结构
A	图分析	CSR, CSC
Q	图查询	Triple Table, KVS
M	图挖掘	ODAG, …
N	图神经网络	Array, Matrix, …
S	图存储	Adjacent List, KVS

引擎层：图流水线

图流水线应用：个性化推荐



- **Graphy：图流水线系统**
  - ① 统一编程接口，多阶段协同的优化策略
  - ② 逆溯中间存储结构，降低任务衔接开销
  - ③ 层次化任务调度，阶段间数据的局部性

	Q	ID	<>	A	<>	ID	S	Q	TOT
WK & PL	2.5		1.3	5.4	0.1		0.2	0.1	**9.6**
WK & GE	2.5	4.5	1.3	1.9	0.1	0.4	0.2	0.1	**11.0**
Spark	12.2			34.0			1.6	0.2	**48.0**
Graphy	0.4			1.1			0.1	0.1	**1.6**

a 4-node cluster, RDF: |V|=21M, |E|=88M, A: PPR
WK=Wukong, PL=PowerLyra, GE=GEMINI

小结

- 图应用潜力巨大，而图处理系统充满了挑战
- 大量新硬件和特性的普及提供了无限可能
- 软硬件协同设计可以大幅提升性能，并且仍然存在巨大潜力
- 当前正是研究机构与领先企业合作的黄金时代

Alipay | in.clusion FINTECH CONFERENCE 外滩大会

谢谢