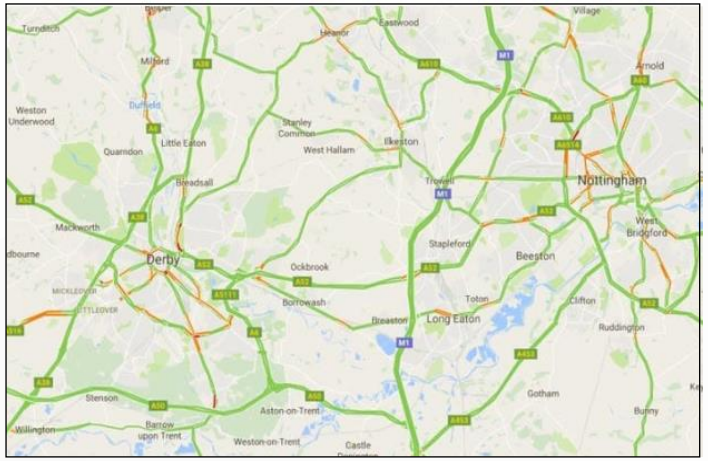
RAGraph: A Region-Aware Framework for Geo-Distributed Graph Processing

Feng Yao, Qian Tao, Wenyuan Yu, Yanfeng Zhang, Shufeng Gong,
Qiange Wang, Ge Yu, Jingren Zhou
Northeastern University, Alibaba Group

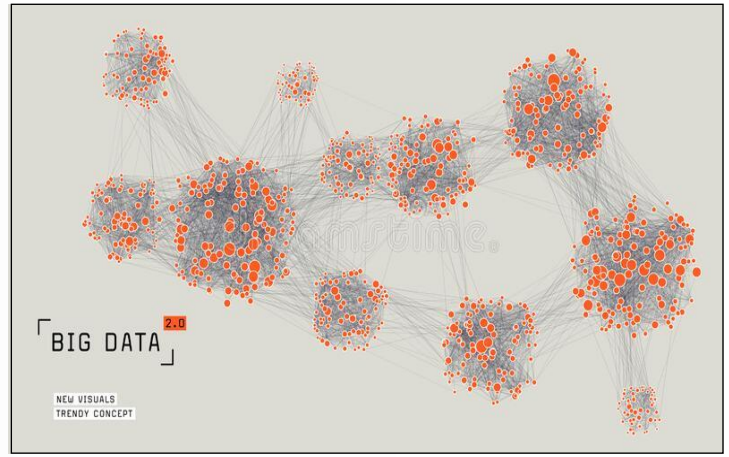
VLDB2024
GUANGZHOU

Graphs

Social network



Road network



Biological networks

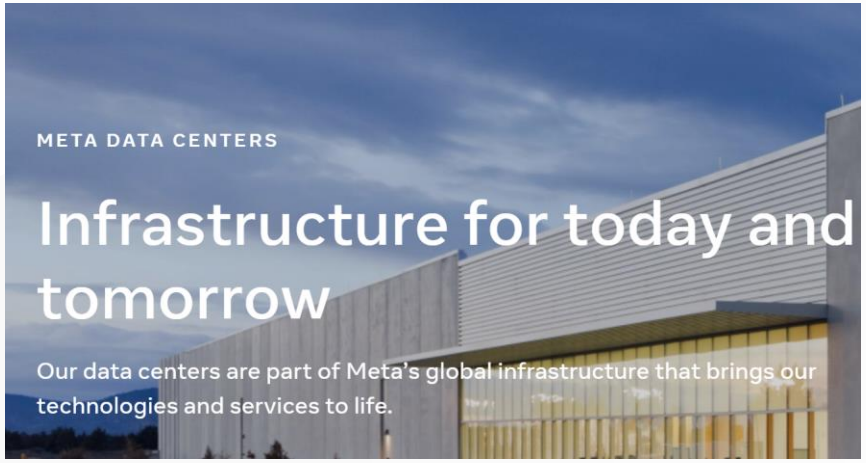
Geo-distributed Graphs

In many real-world application scenarios, graph data is geographically distributed.

Geo-distributed Graphs

In many real-world application scenarios, graph data is geographically distributed.

META has constructed over 20 data centers for global social networking operations.



Europe



Asia

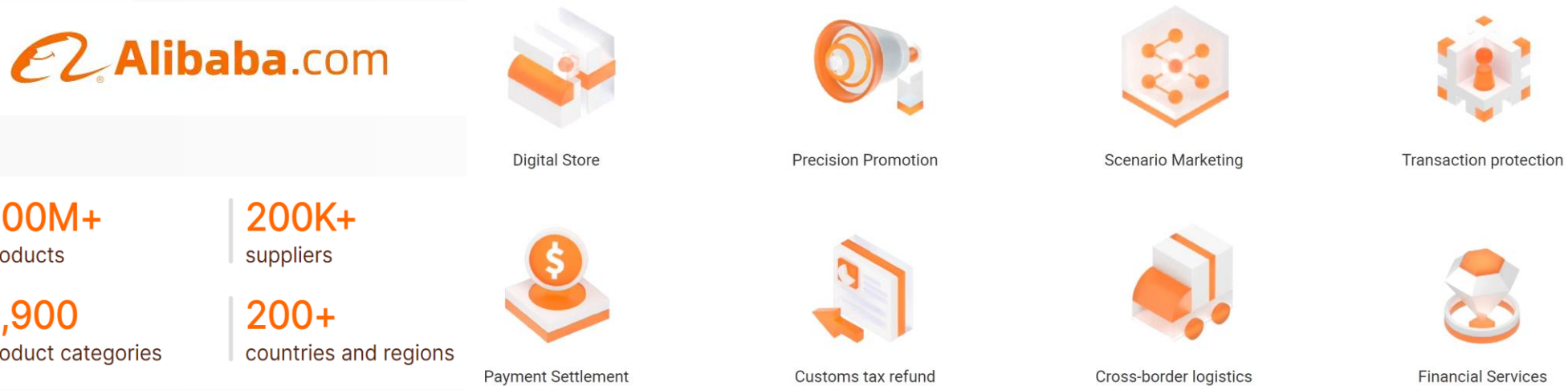
United States



Geo-distributed Graphs

In many real-world application scenarios, graph data is geographically distributed.

E-commerce platform for global trade network and cross-border logistics.

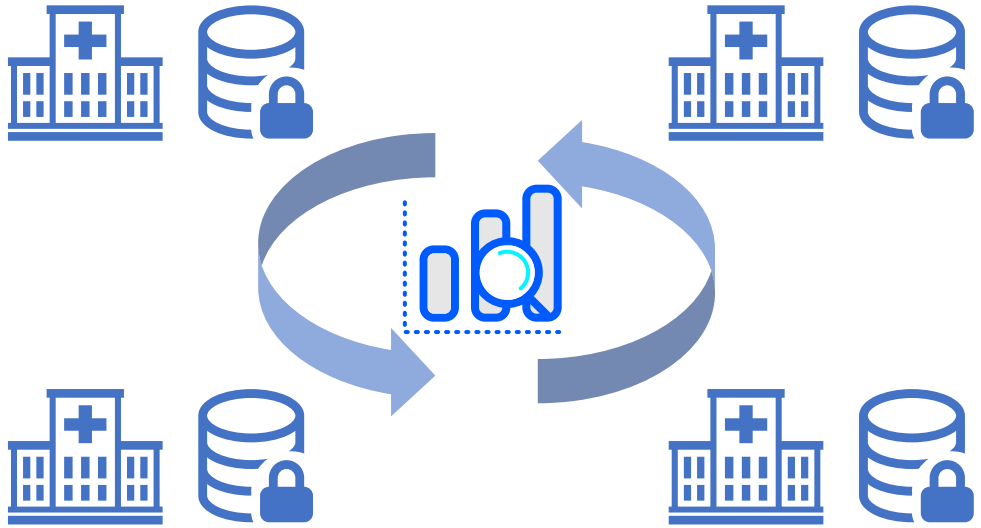


Geo-distributed Graphs

In many real-world application scenarios, graph data is geographically distributed.

Federated graph computing

- Healthcare organizations provide personalized healthcare services by integrating medical data.



Geo-distributed Graphs

Geo-distributed graph data have significant value for data mining.

Geo-distributed Graphs

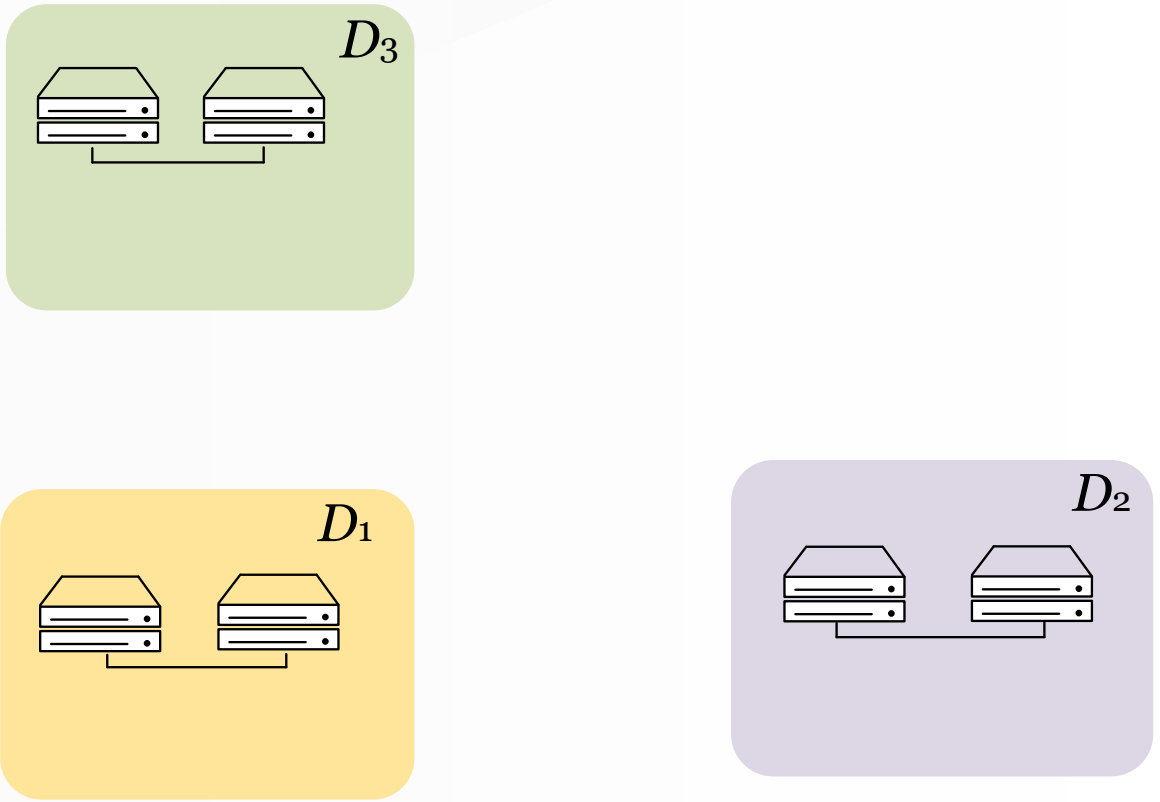
Geo-distributed graph data have significant value for data mining.

- Large-scale **data migration centralized processing** graph is expensive due to high network transmission costs.

- The increasing **data protection** requirements also make redistributing data among data centers impossible.

Geo-distributed Cluster

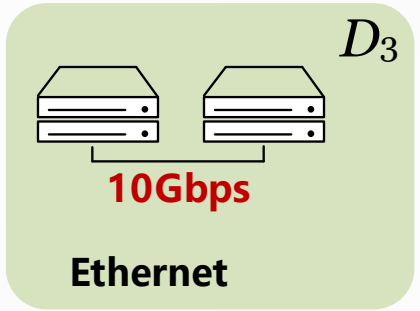
Real-world geo-distributed cluster configuration (AliCloud ECS cluster):

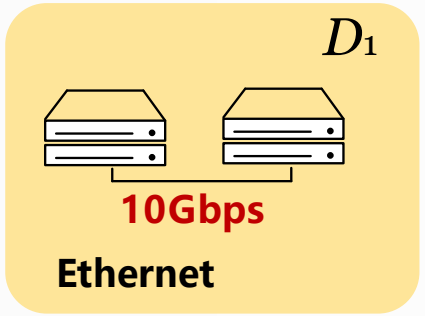


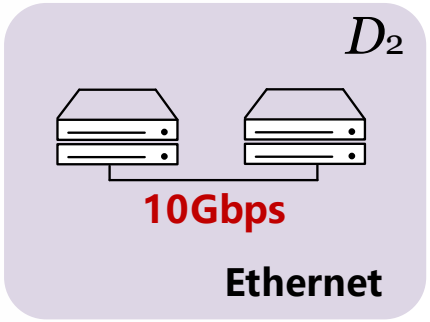
Geo-distributed Cluster

Real-world geo-distributed cluster configuration (AliCloud ECS cluster):

- **High bandwidth within data centers**
- Scarce and heterogeneous bandwidth between data centers
- Network fluctuations





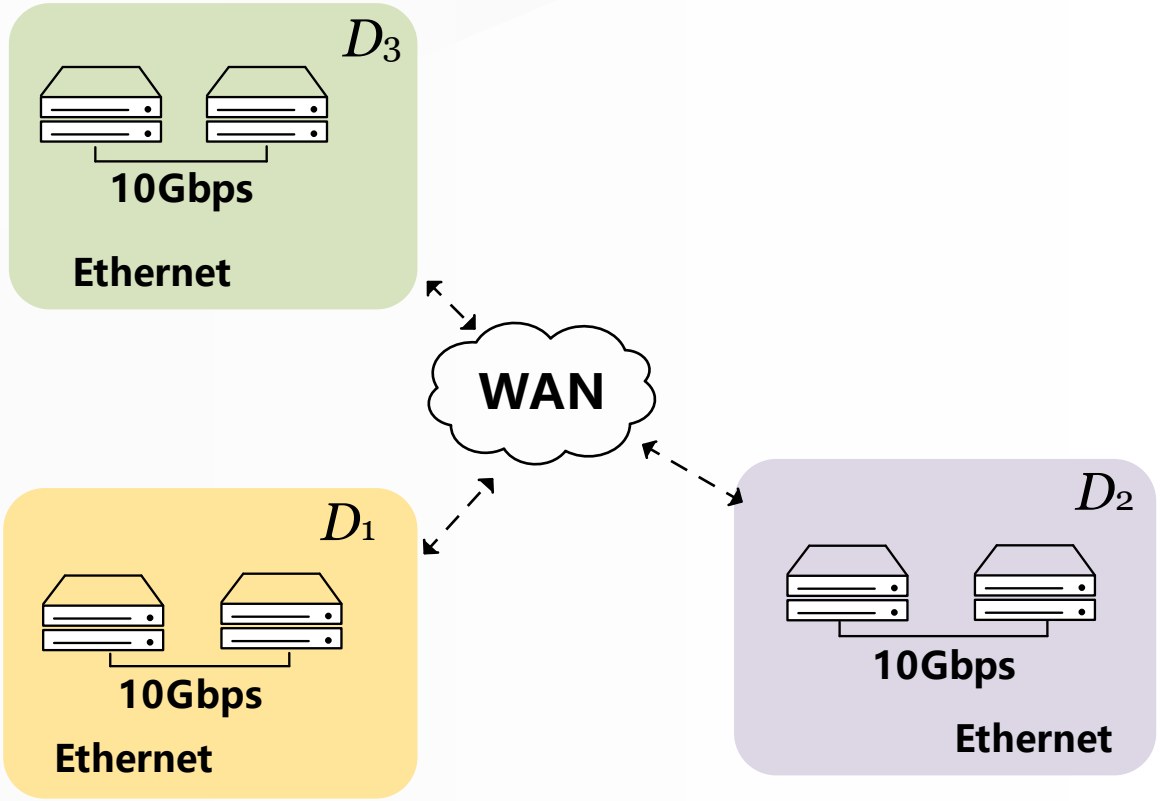


10 Gbps Ethernet connection within the data center.

Geo-distributed Cluster

Real-world geo-distributed cluster configuration (AliCloud ECS cluster):

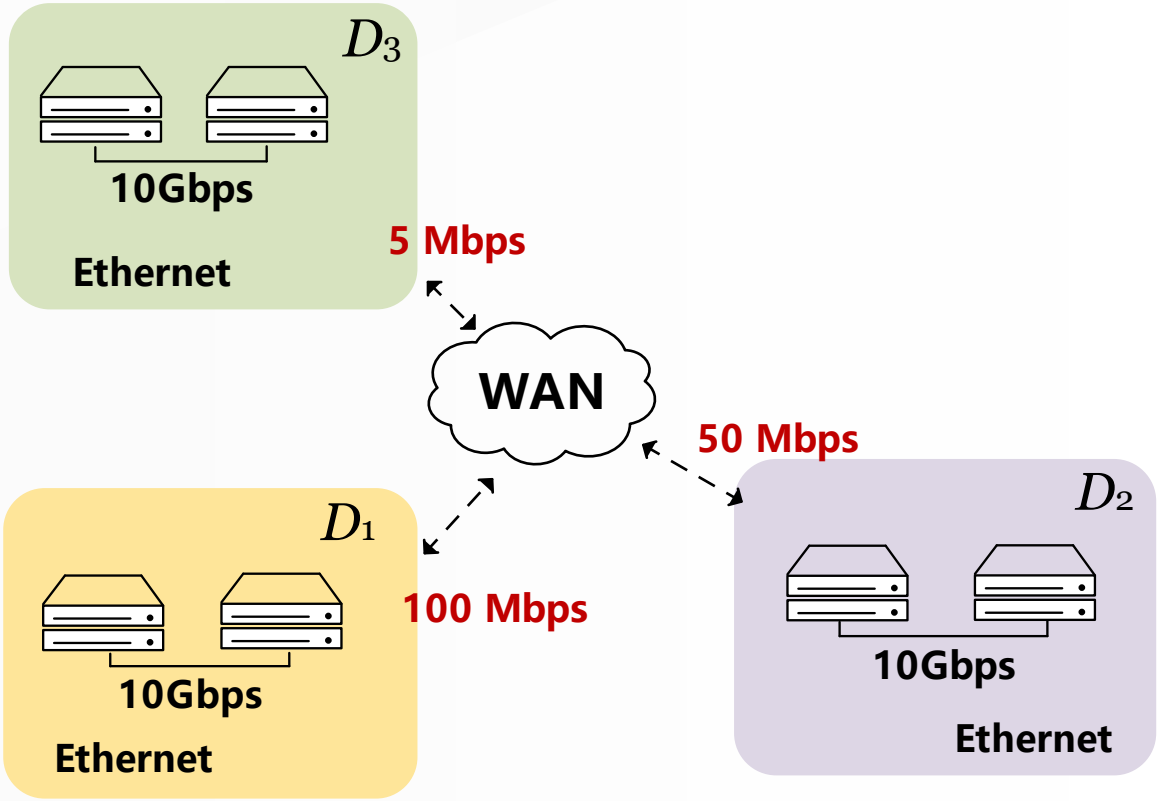
- High bandwidth within data centers
- **Scarce and heterogeneous bandwidth between data centers**
- Network fluctuations



Geo-distributed Cluster

Real-world geo-distributed cluster configuration (AliCloud ECS cluster):

- High bandwidth within data centers
- **Scarce and heterogeneous bandwidth between data centers**
- Network fluctuations

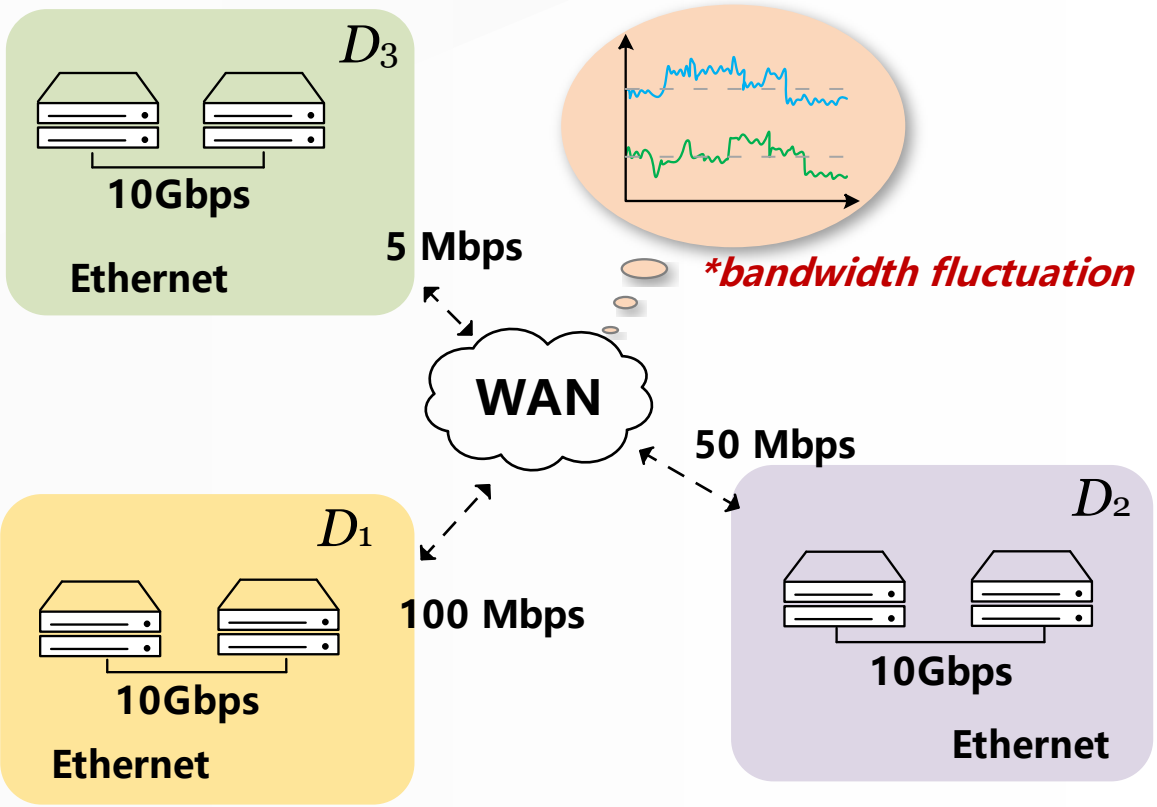


Up to 100 Mbps between data centers.

Geo-distributed Cluster

Real-world geo-distributed cluster configuration (AliCloud ECS cluster):

- High bandwidth within data centers
- Scarce and heterogeneous bandwidth between data centers
- **Network fluctuations**

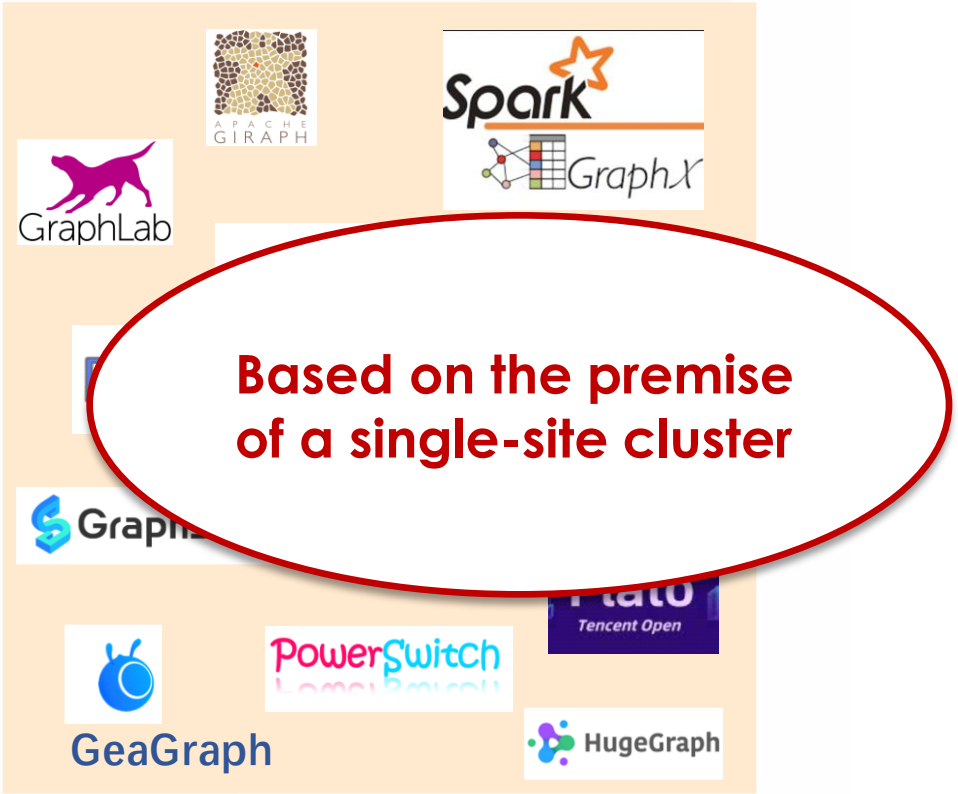


Due to **network fluctuations**, WAN links are unstable.

Graph Processing Systems

Distributed Systems

Graph Processing Systems

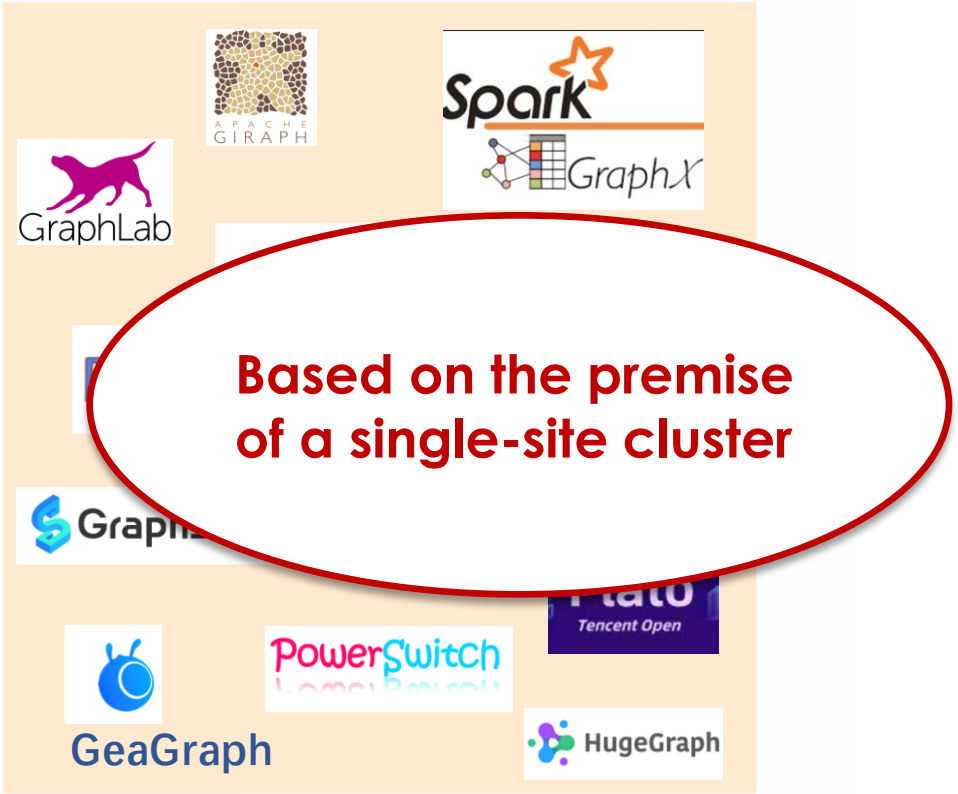


Distributed Systems

Graph Processing Systems

high bandwidth

homogeneous links

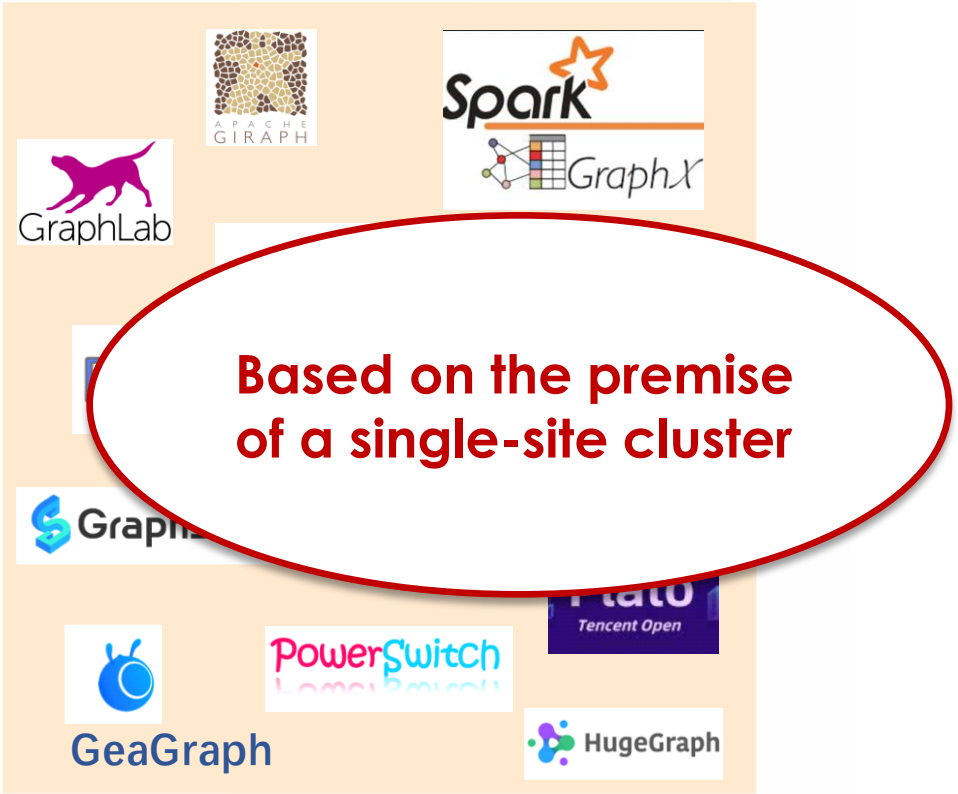


Distributed Systems

Graph Processing Systems

high bandwidth

homogeneous links



Distributed Systems

Geo-distributed Cluster

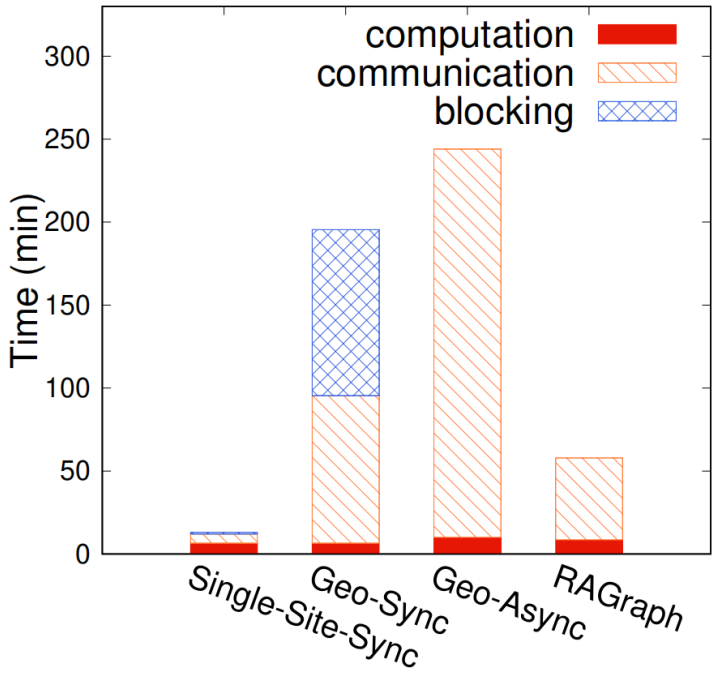
Real-world geo-distributed cluster configuration (AliCloud ECS cluster):

- High bandwidth within data centers
- Scarce and heterogeneous bandwidth between data centers
- Network fluctuations

A challenge to the **traditional assumptions** underlying the design of distributed systems.

Geo-distributed Cluster

Tested SOTA traditional system GRAPE (Sync) /Maiter (Async) on single-site cluster and geo-distributed cluster

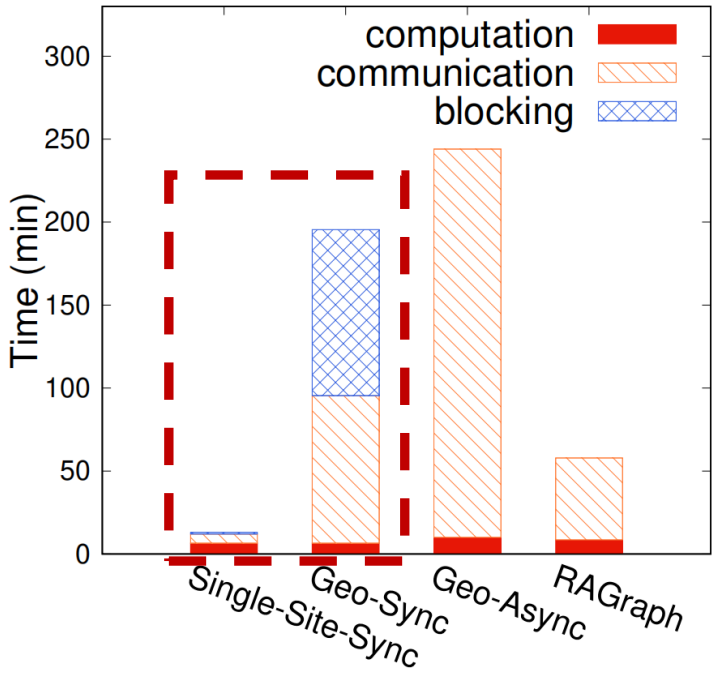


Single-site VS. geo-distributed cluster.

Geo-distributed Cluster

Tested SOTA traditional system GRAPE (Sync) /Maiter (Async) on single-site cluster and geo-distributed cluster

- **Network bandwidth differences**
- Scarce and heterogeneous bandwidth between data centers



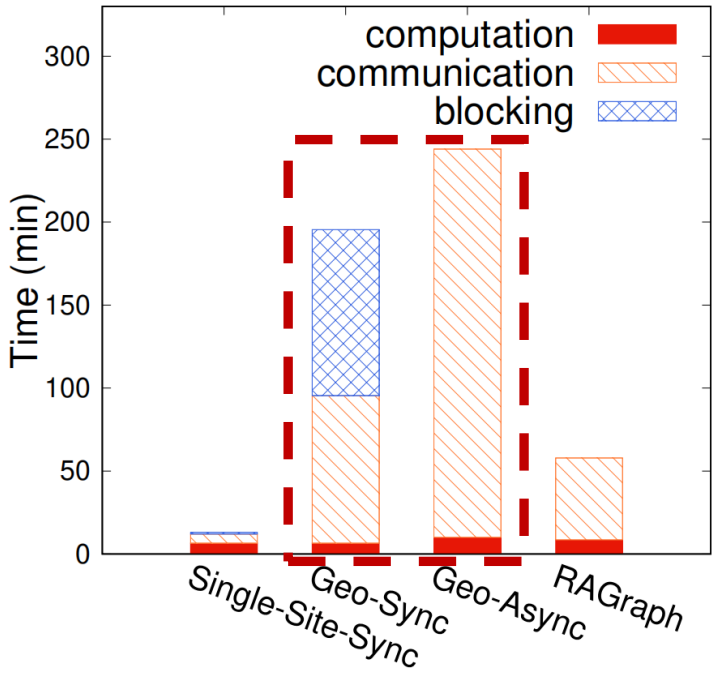
Single-site VS. geo-distributed cluster.

running time **significantly increases** under geo-distributed cluster.

Geo-distributed Cluster

Tested SOTA traditional system GRAPE (Sync) /Maiter (Async) on single-site cluster and geo-distributed cluster

- Network bandwidth differences
- **Scarce and heterogeneous bandwidth between data centers**



Single-site VS. geo-distributed cluster.

overhead comes from **communication time** and **blocking wait time**.

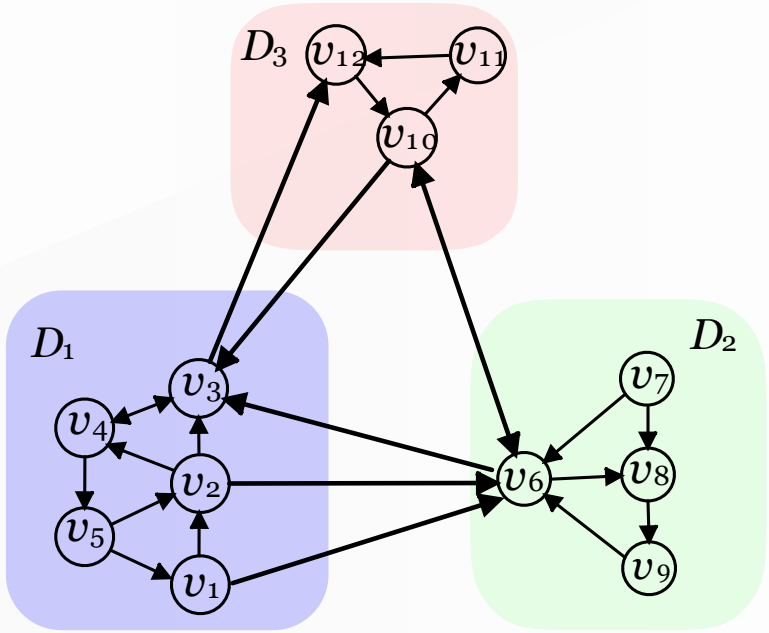
Challenges

- ***Imbalance of Message Transmission***
  - Message transmission time between data centers **is much longer** than that within a data center.
  - Heterogeneous networks cause **imbalanced message transmission** between data centers.
- ***Inefficiency of Graph Processing Model***
  - Synchronous Parallel model (**blocking wait time**)
  - Asynchronous Parallel model (**communication time**)

Observations

1 Imbalanced local-global interaction

Input graph:

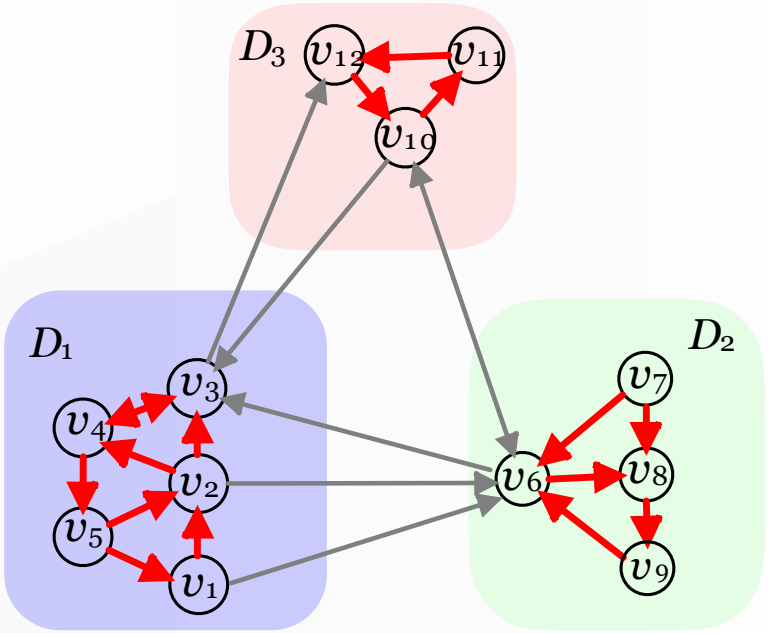


Observations

1 Imbalanced local-global interaction

Observation: Traditional interaction pattern.

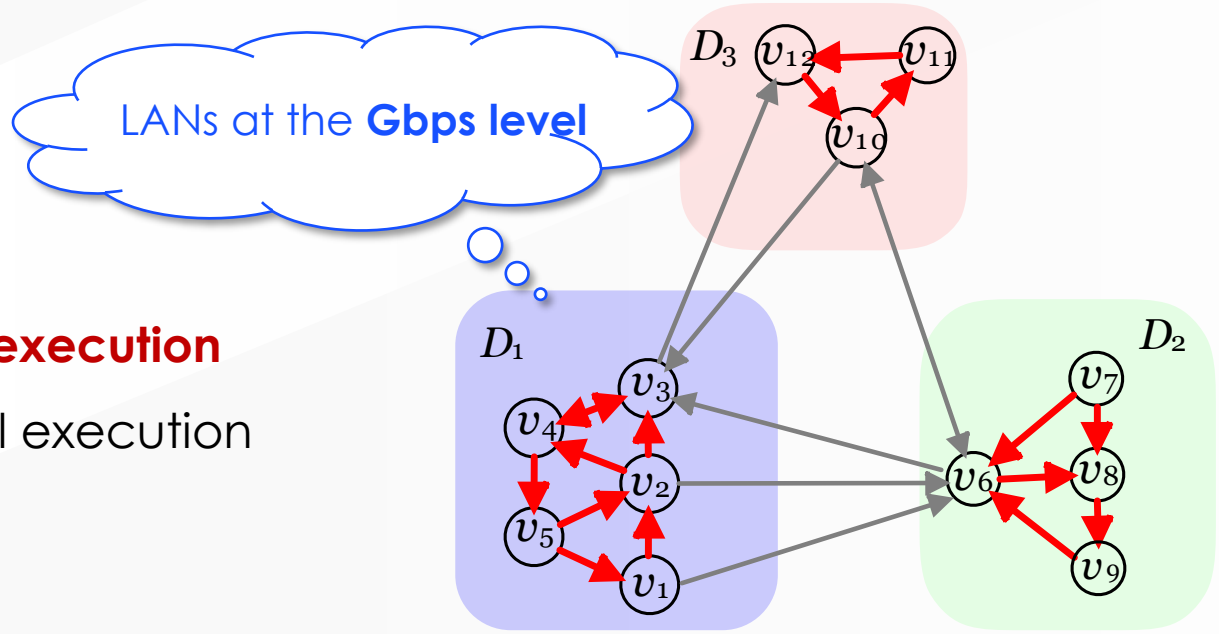
- **Local execution**
- Global execution



Observations

1 Imbalanced local-global interaction

Observation: Traditional interaction pattern.



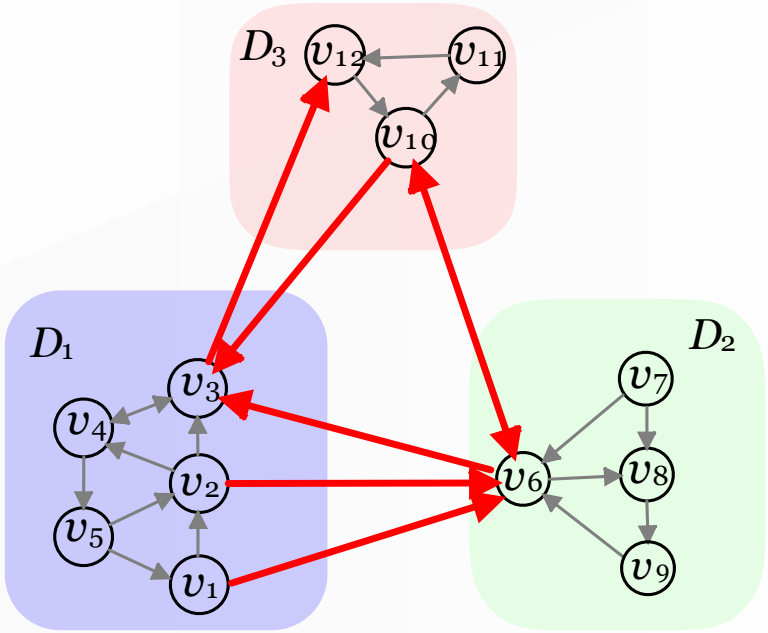
- **Local execution**
- Global execution

Observations

1 Imbalanced local-global interaction

Observation: Traditional interaction pattern.

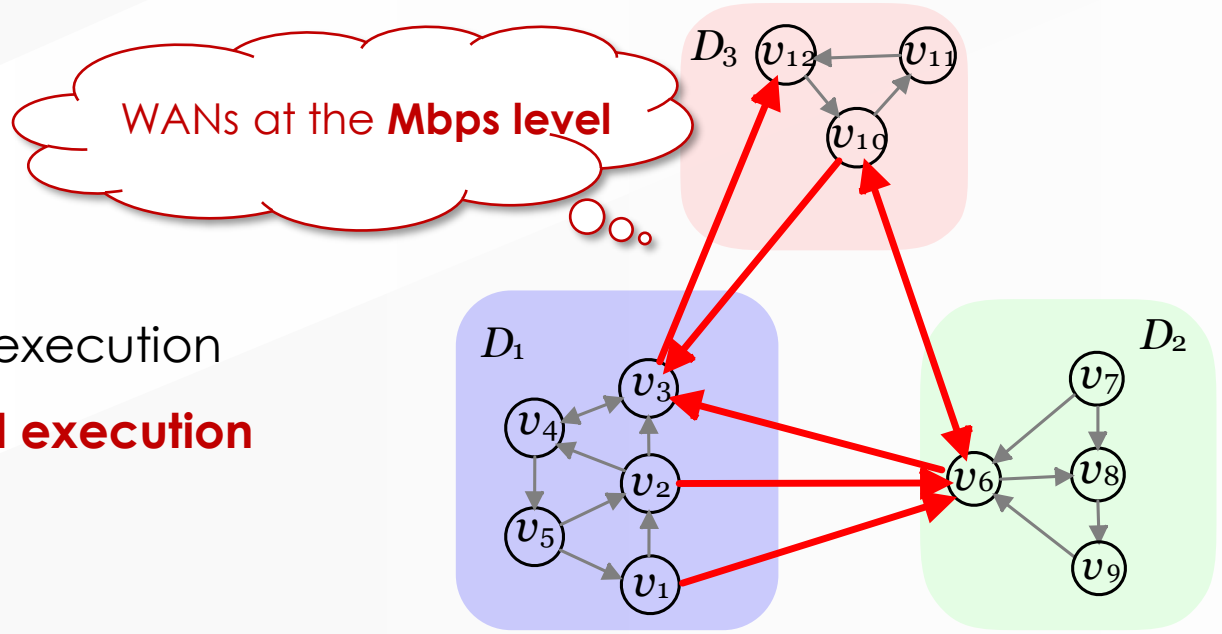
- Local execution
- **Global execution**



Observations

1 Imbalanced local-global interaction

Observation: Traditional interaction pattern.

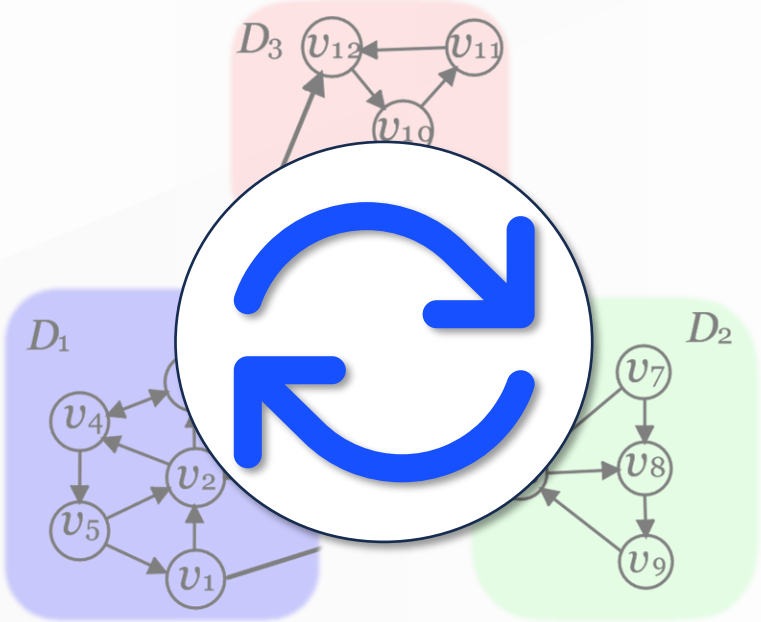


- Local execution
- **Global execution**

Observations

1 Imbalanced local-global interaction

Observation: Traditional interaction pattern.

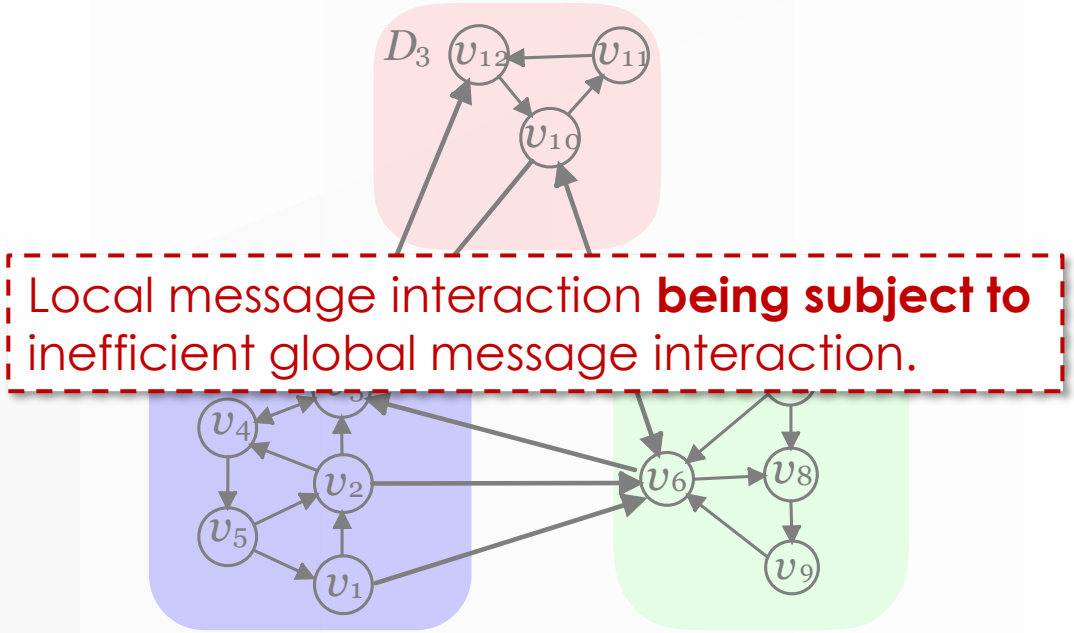


Local-global **alternate** execution

Observations

1 Imbalanced local-global interaction

Issues arising:

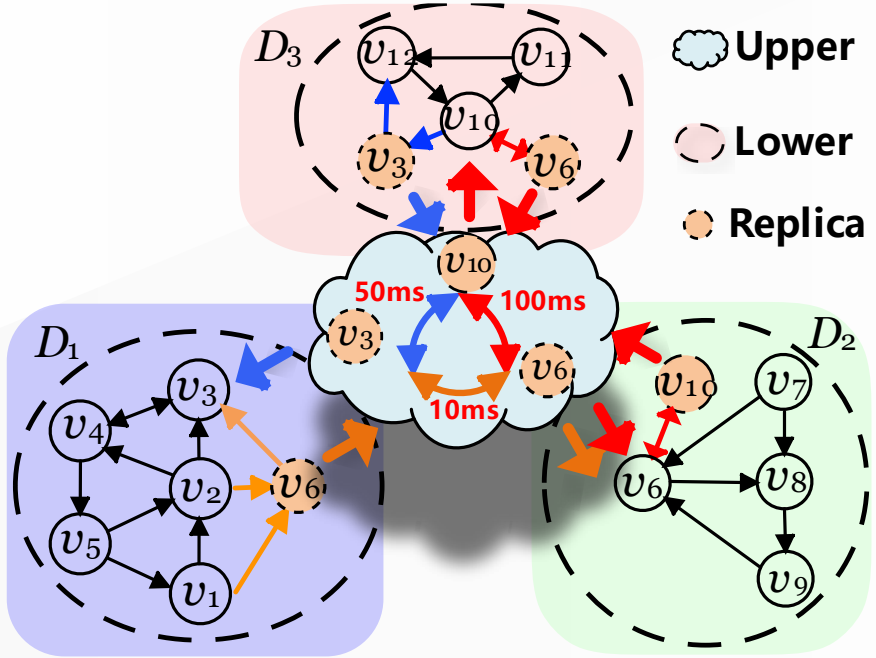


Local message interaction **being subject to** inefficient global message interaction.

Observations

1 Imbalanced local-global interaction

Solution: Two-layer interaction view.

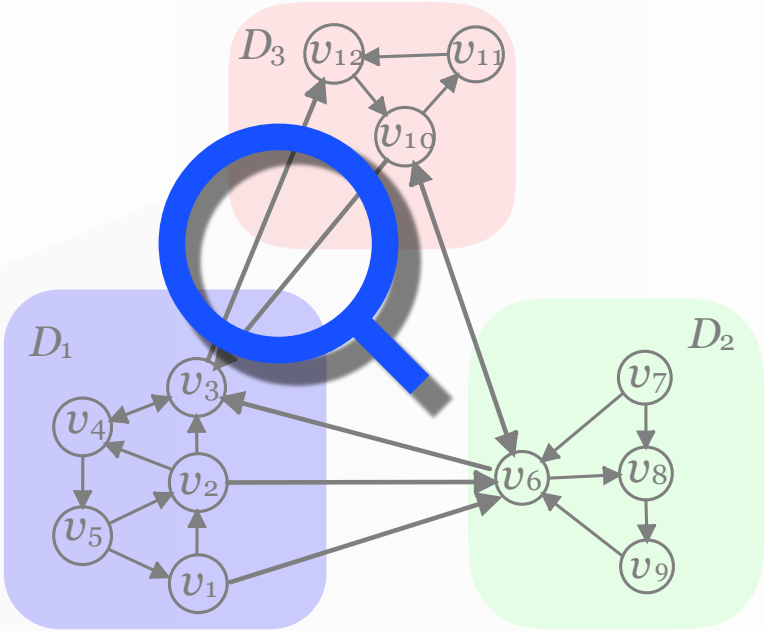


Decouple local-global interactions.

Observations

2 Ping-Pong Effect

Observation: Message interaction on the boundary.

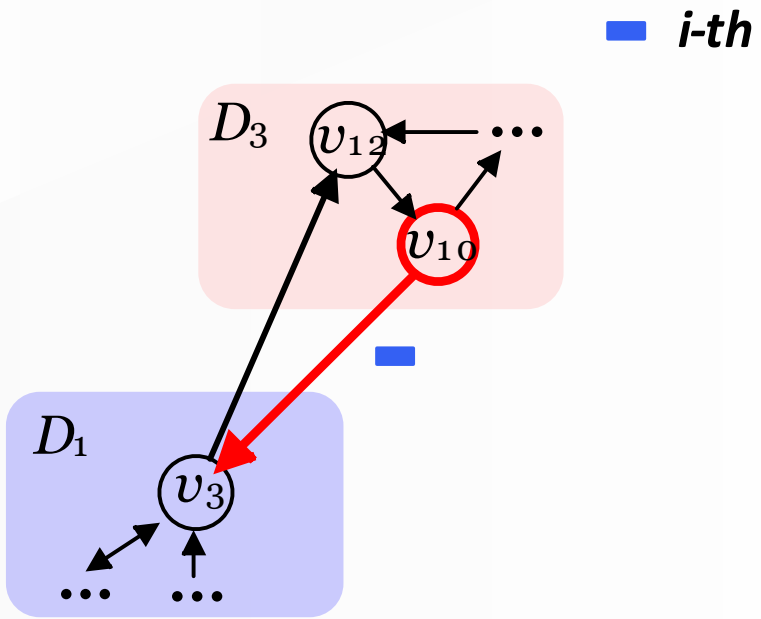


Observations

2 Ping-Pong Effect

Observation: Message interaction on the boundary.

- **V_{10} sends a message to V_3 in D_1**
- V_3 generates a new message
- V_3 sends this message to V_{12}

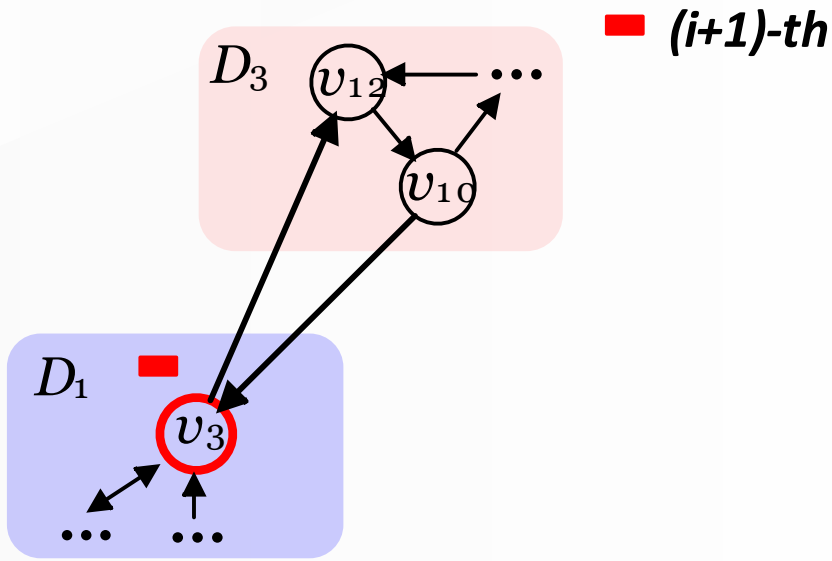


Observations

2 Ping-Pong Effect

Observation: Message interaction on the boundary.

- V_{10} sends a message to V_3 in D_1
- **V_3 generates a new message**
- V_3 sends this message to V_{12}

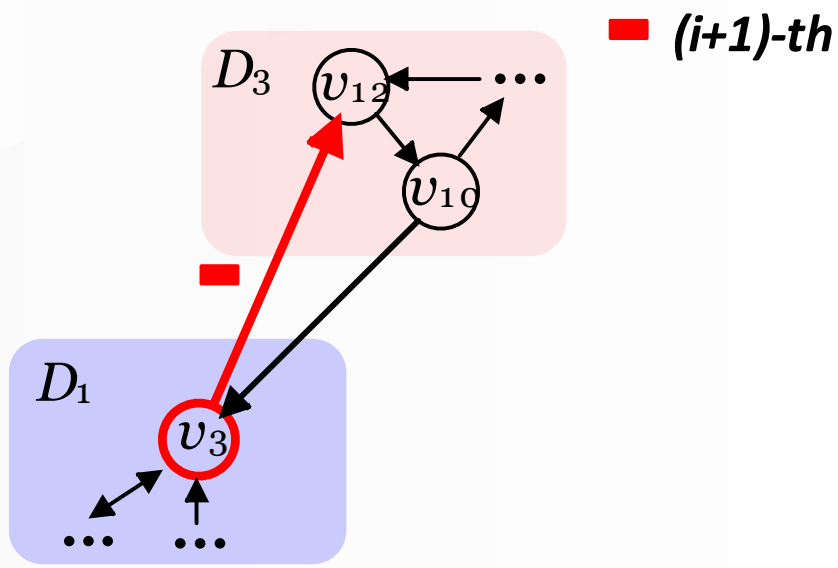


Observations

2 Ping-Pong Effect

Observation: Message interaction on the boundary.

- V_{10} sends a message to V_3 in D_1
- V_3 generates a new message
- **V_3 sends this message to V_{12}**

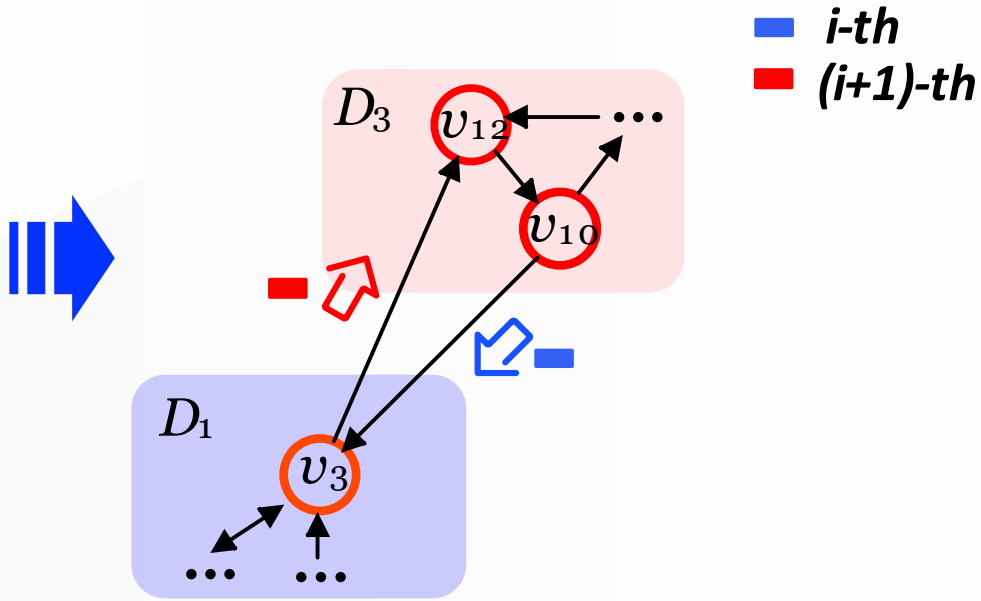


Observations

2 Ping-Pong Effect

Observation: Message interaction on the boundary.

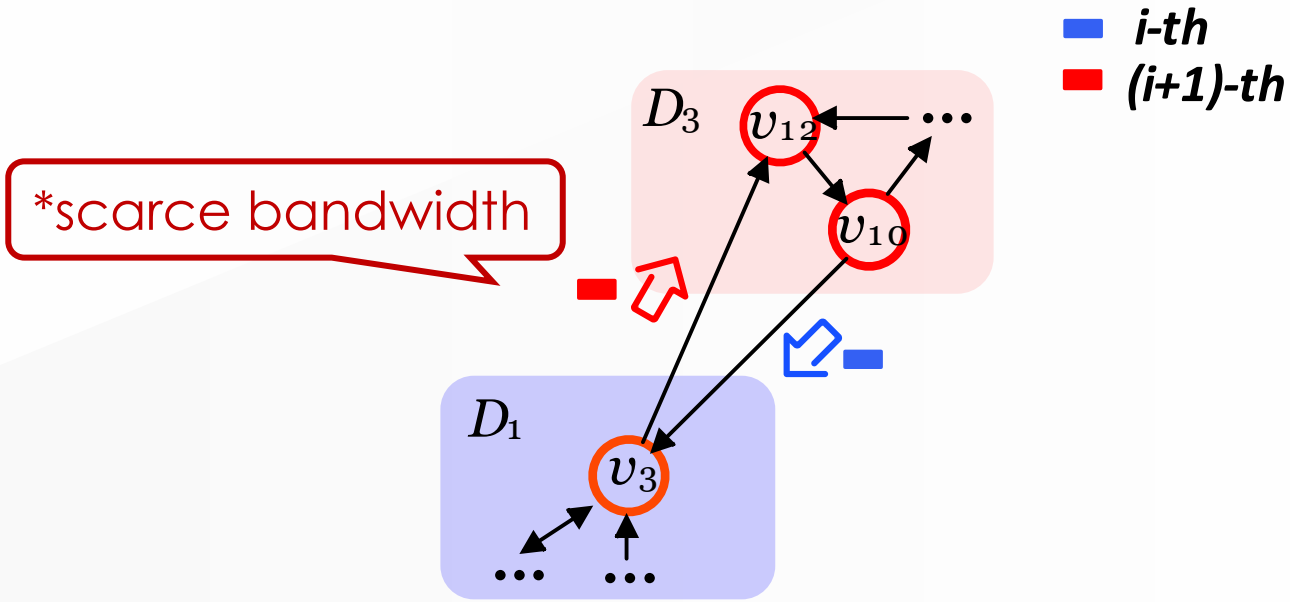
Called **Ping-Pong Effect**



Observations

2 Ping-Pong Effect

Issues arising:



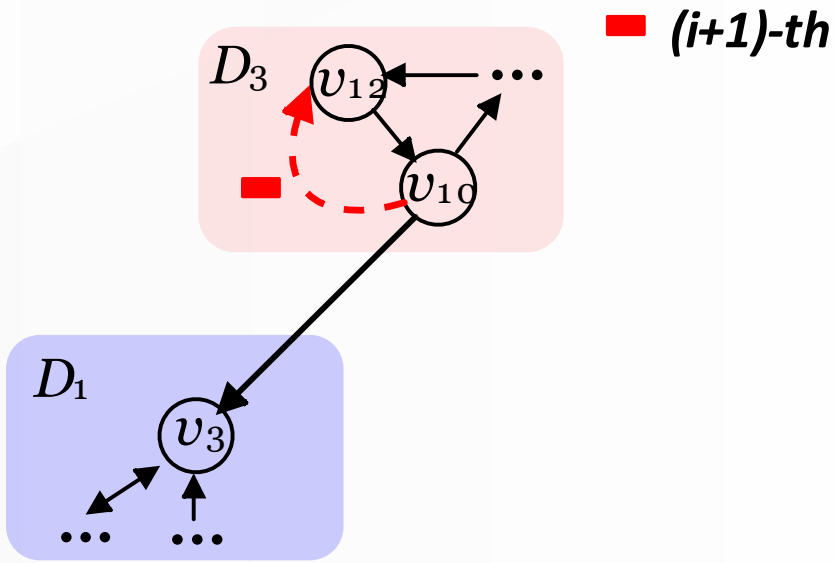
round-trip transmission waiting.

Observations

2 Ping-Pong Effect

Solution: Remote message internal execution.

- **The message from V_{10} is updated and sent directly to V_{12}**
- Boost the message passing



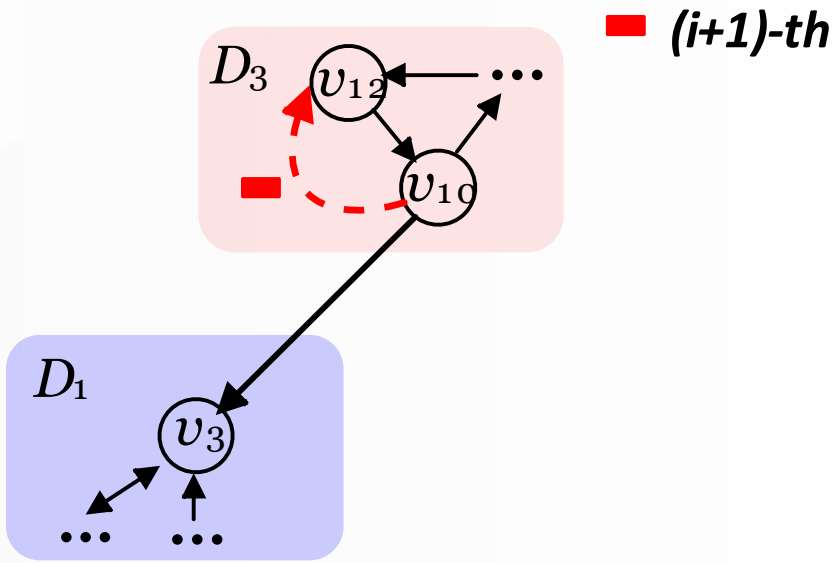
Advance inefficient global updates to local computation.

Observations

2 Ping-Pong Effect

Solution: Remote message internal execution.

- The message from V_{10} is updated and sent directly to V_{12}
- **Boost the message passing**



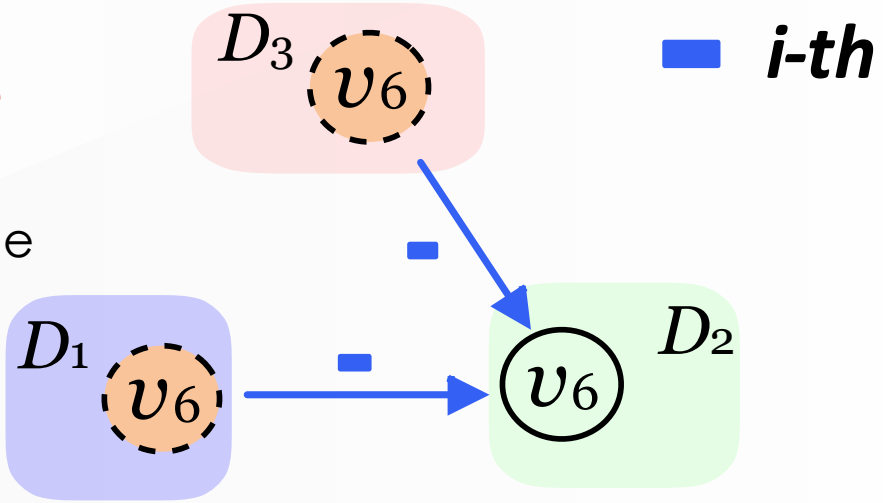
Advance inefficient global updates to local computation.

Observations

3 Intolerable network congestion

Observation: Typical communication pattern.

- **Replicas send messages to the master**
- Master sends updated message to replicas

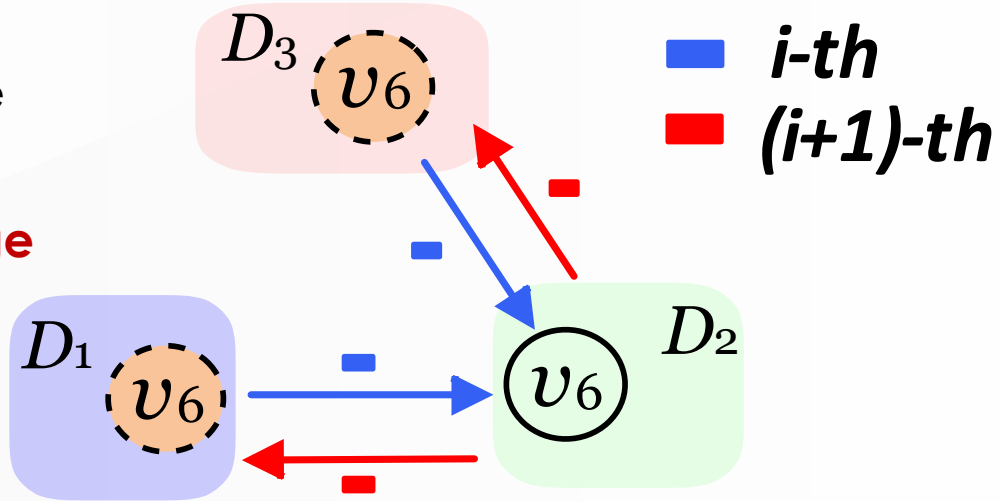


Observations

3 Intolerable network congestion

Observation: Typical communication pattern.

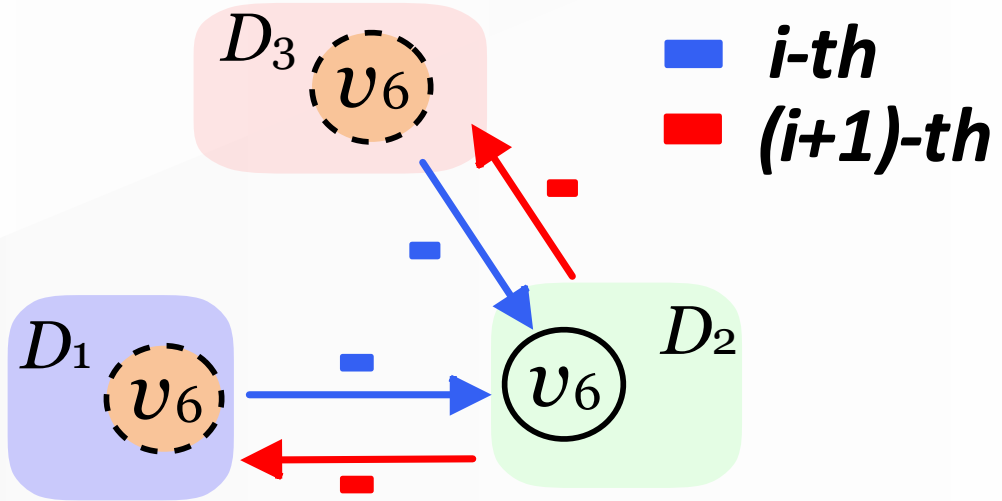
- Replicas send messages to the master
- **Master sends updated message to replicas**



Observations

3 Intolerable network congestion

Observation: Typical communication pattern.

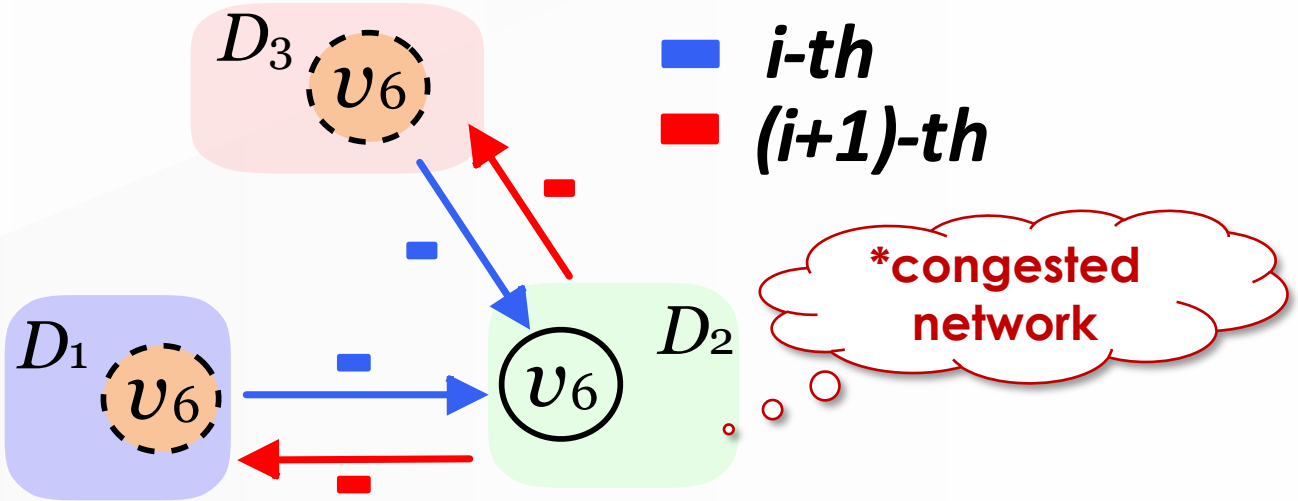


One-to-many pattern result in a large inflow/outflow of messages on the "one" side.

Observations

3 Intolerable network congestion

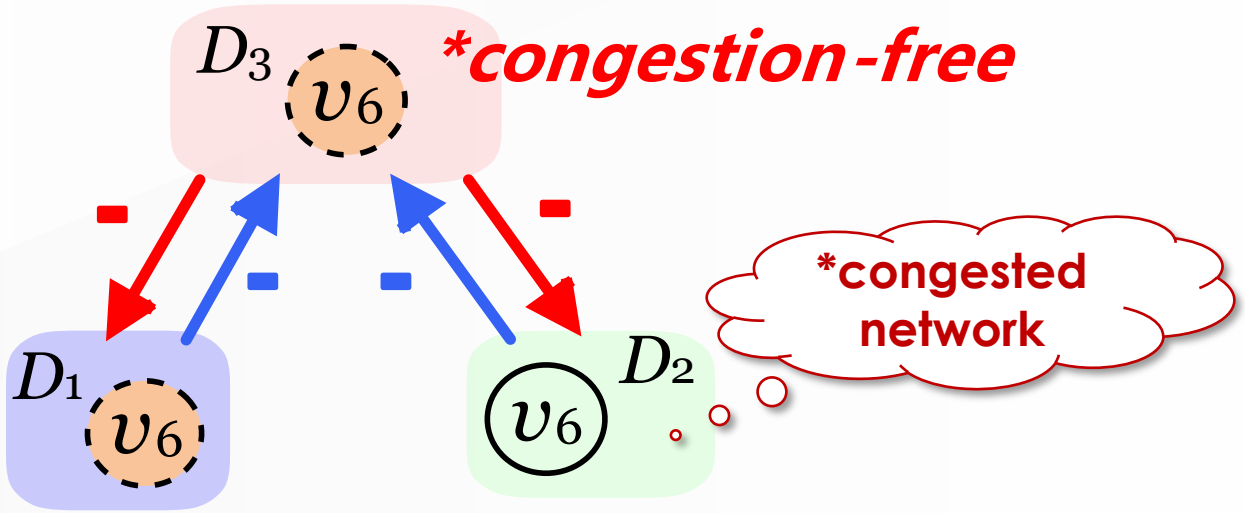
Issues arising:



Observations

3 Intolerable network congestion

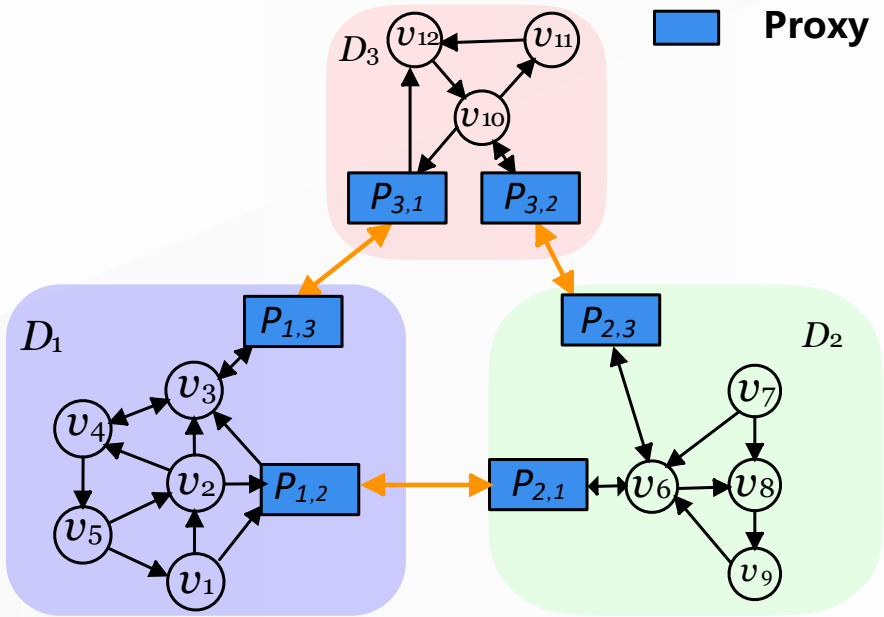
Solution: Replaceable communication pattern.



Replace the master role in a **congestion-free data center.**

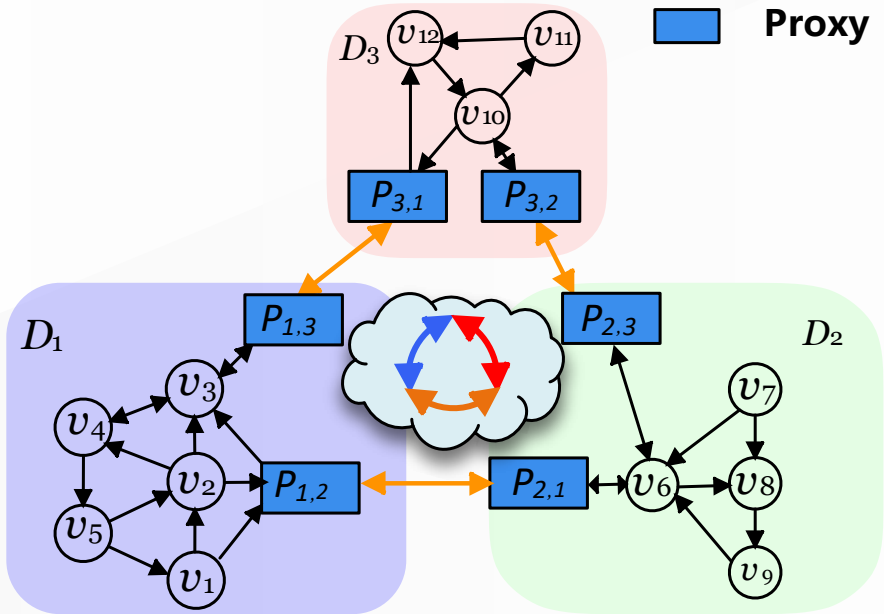
Region-Aware Framework

Setting up proxy



Region-Aware Framework

Setting up proxy

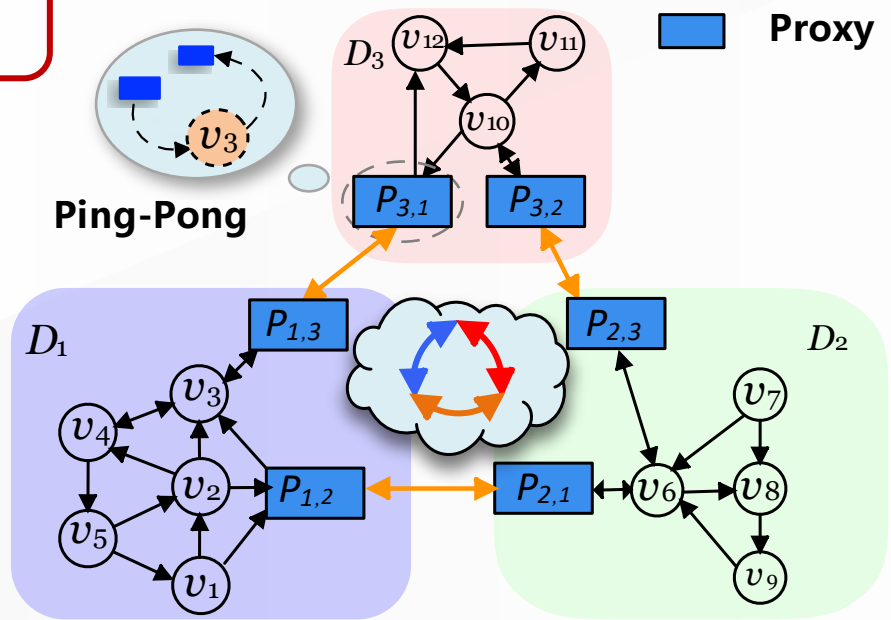


❖ decoupled local/global layers

Region-Aware Framework

Setting up proxy

- Local computation of the ping-pong effect

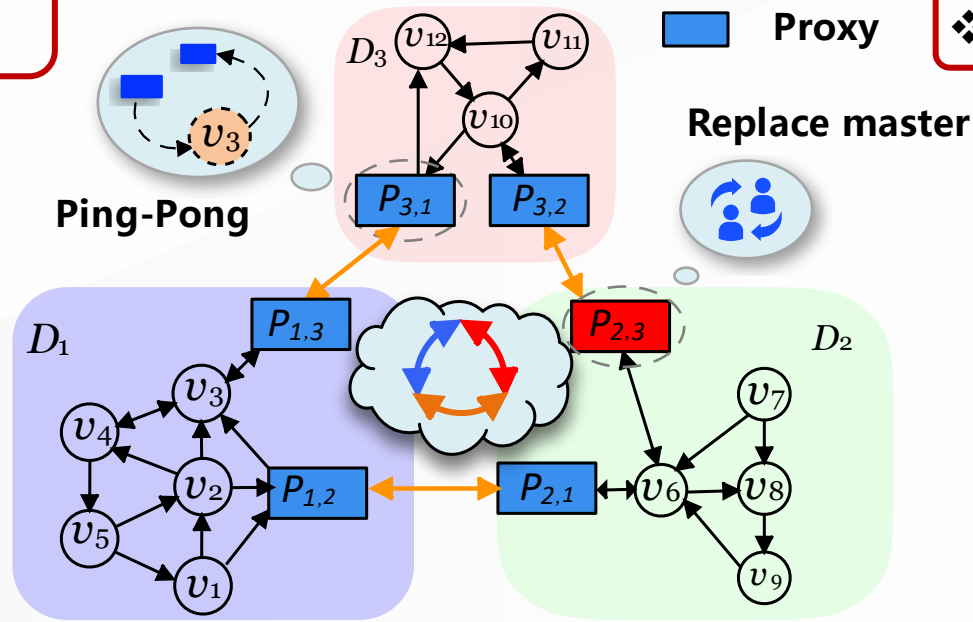


- decoupled local/global layers

Region-Aware Framework

Setting up proxy

- Local computation of the ping-pong effect



- Replaceable communications
- decoupled local/global layers

System correctness guarantee

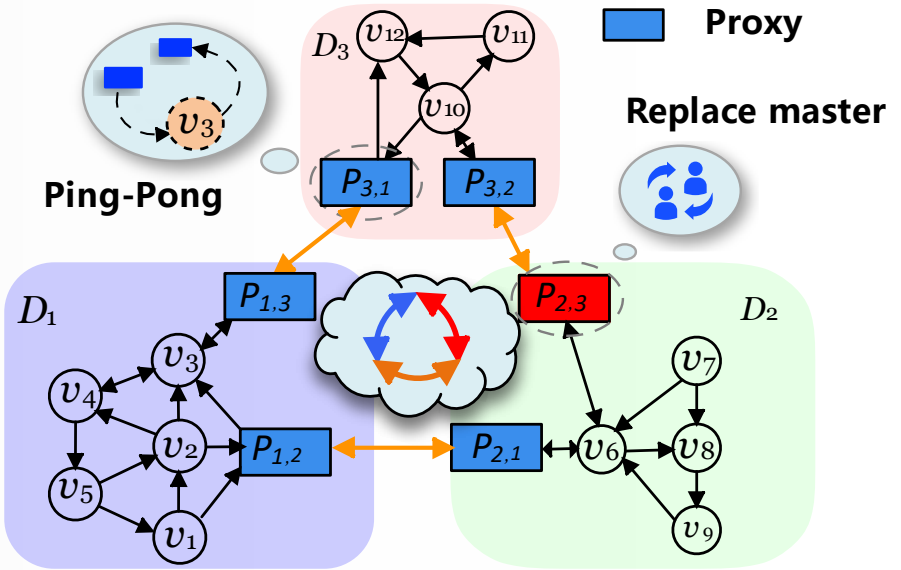
The Delta State Conflict-free Replicated Data Type (CRDT) [1] theory guarantees *Strong Eventual Consistency* between proxies.

Given algorithmic constraints, the aggregation function A and message interaction function I satisfy the monotonic condition:

Monotonic Conditions. $\mathcal{A}$ and $\mathcal{I}$ satisfy monotonic conditions if:

(C1) $\mathcal{A}(X \cup Y) = \mathcal{A}(Y \cup X)$ and $\mathcal{A}(\mathcal{A}(X) \cup Y) = \mathcal{A}(X \cup Y)$

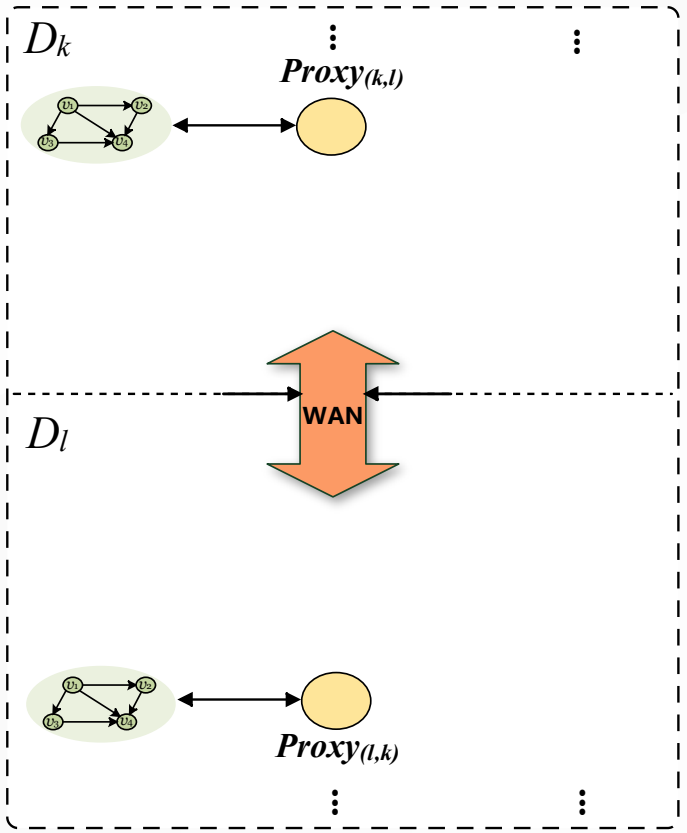
(C2) $\mathcal{I}(\mathcal{A}(X \cup Y)) = \mathcal{A}(\mathcal{I}(X) \cup \mathcal{I}(Y))$



[1] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. 2011. Conflict-free replicated data types. In SSS. 386–

Two Runtime Optimization

1 Adaptive hierarchical interaction engine

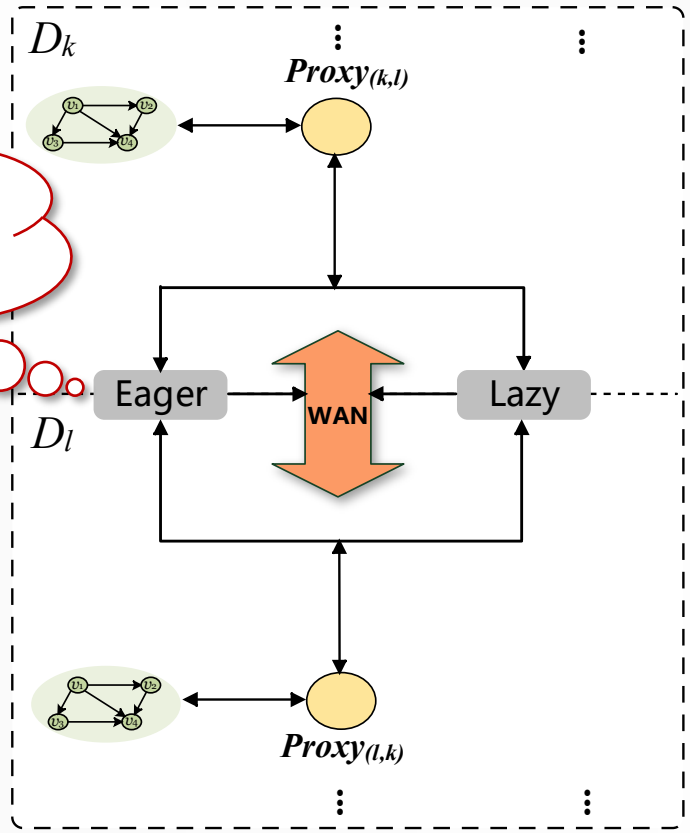


two modes for **heterogeneous networks**

Two Runtime Optimization

1 Adaptive hierarchical interaction engine

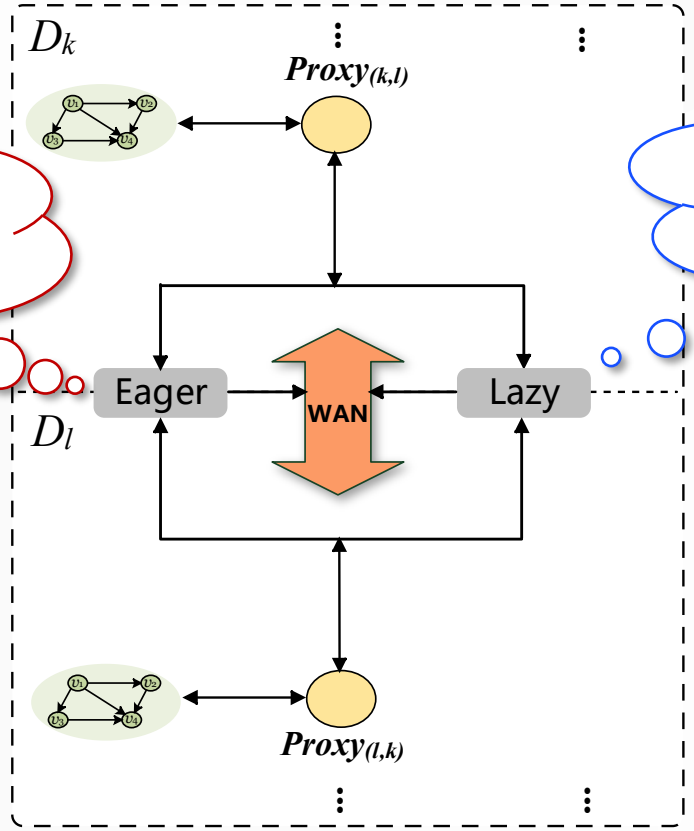
Eager mode for **timely messages** whenever possible.



two modes for **heterogeneous networks**

Two Runtime Optimization

1 Adaptive hierarchical interaction engine



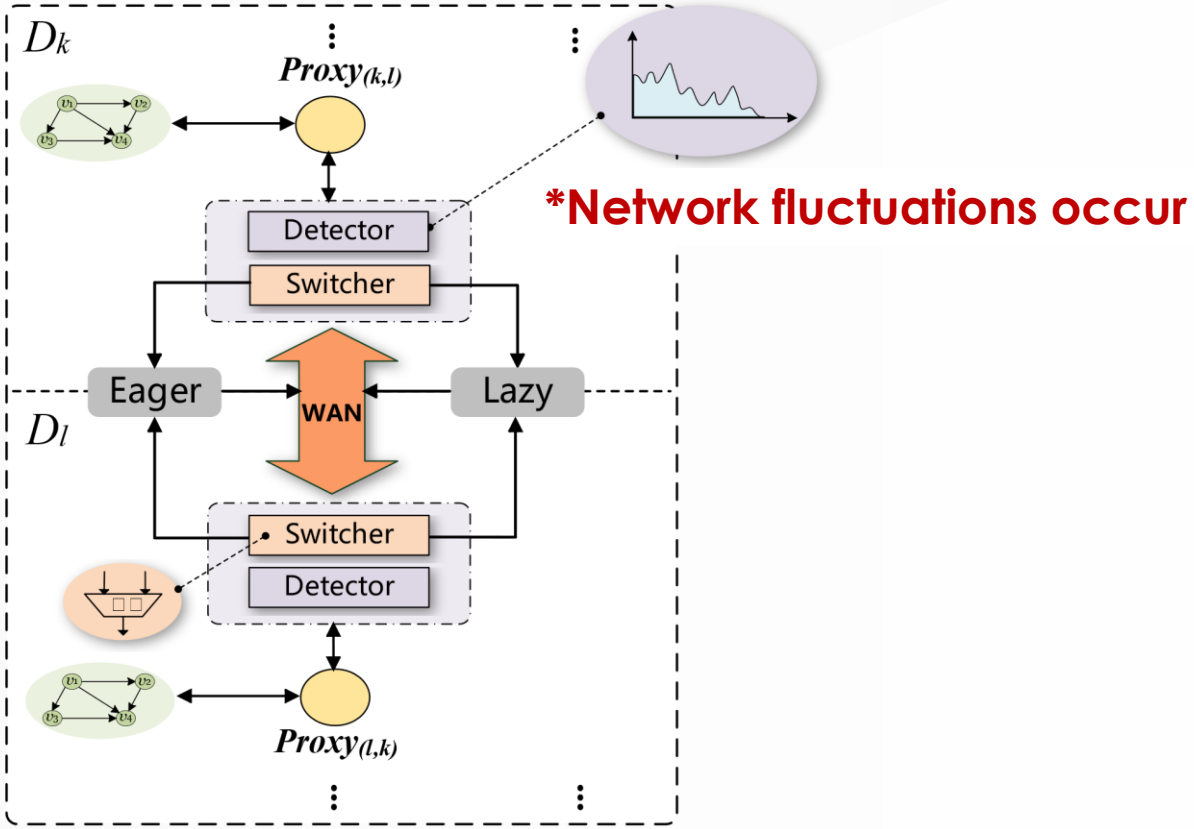
Eager mode for **timely messages** whenever possible.

lazy mode **fetches significant messages** only when necessary.

two modes for **heterogeneous networks**

Two Runtime Optimization

1 Adaptive hierarchical interaction engine

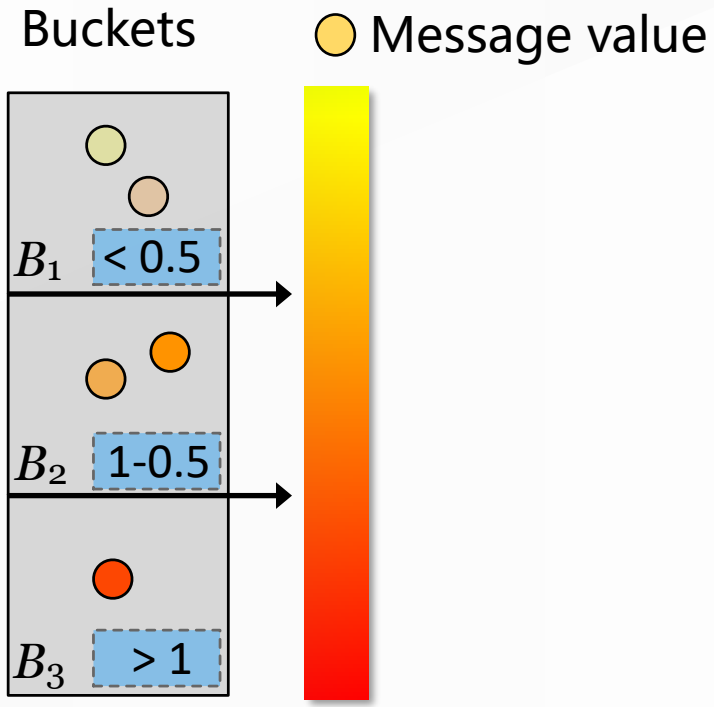


*Network fluctuations occur

proxy **dynamically switches** between the two modes

Two Runtime Optimization

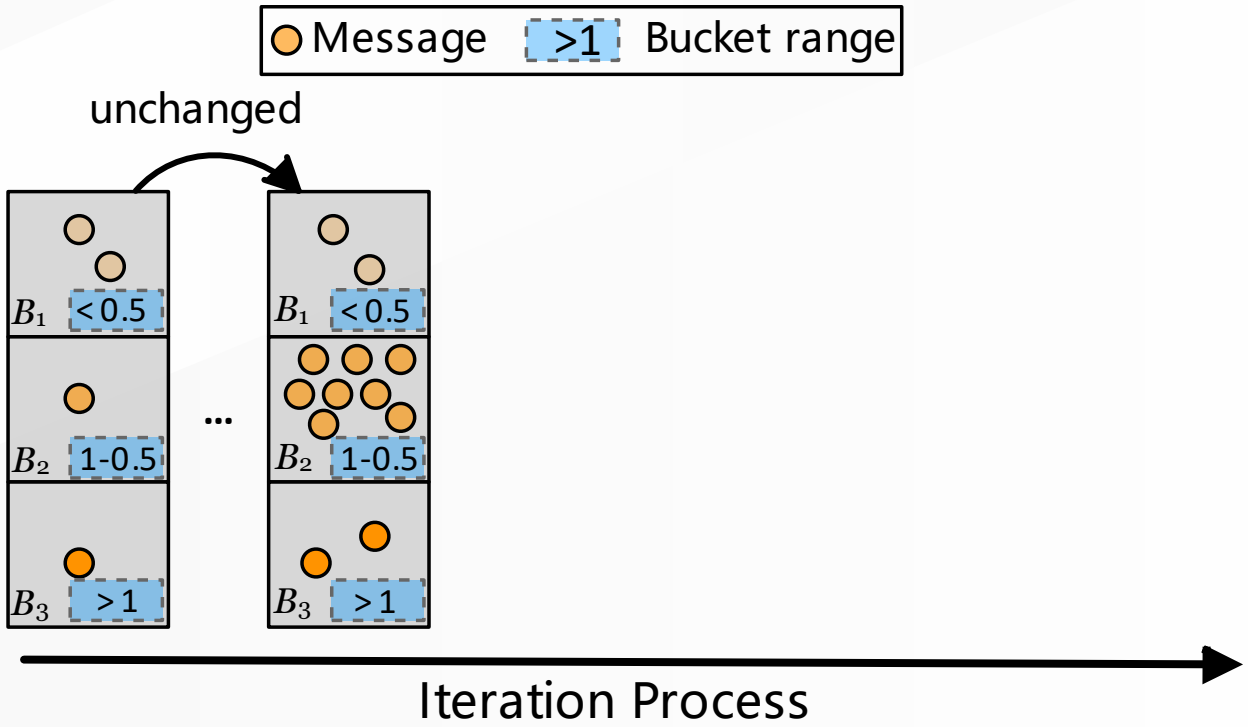
2 Adaptive buckets for message filtering



Messages **fall into different buckets** according to their values

Two Runtime Optimization

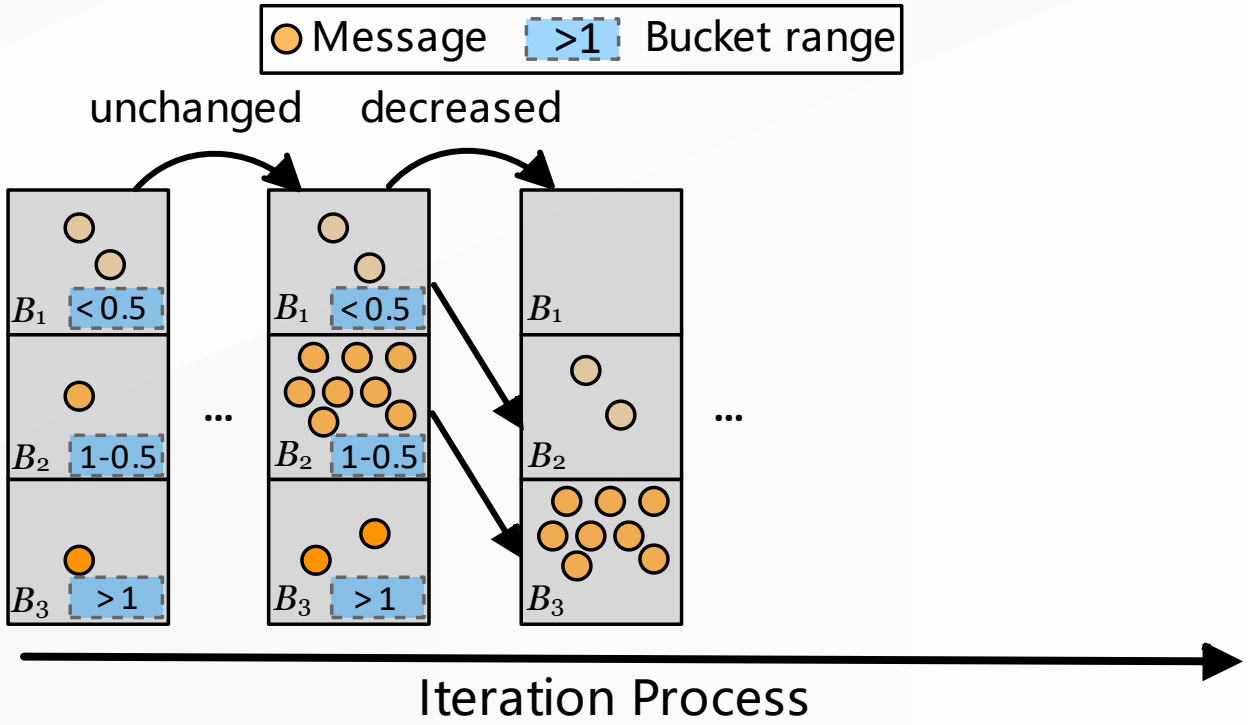
2 Adaptive buckets for message filtering



Dynamically change the range of buckets as the iterative progresses

Two Runtime Optimization

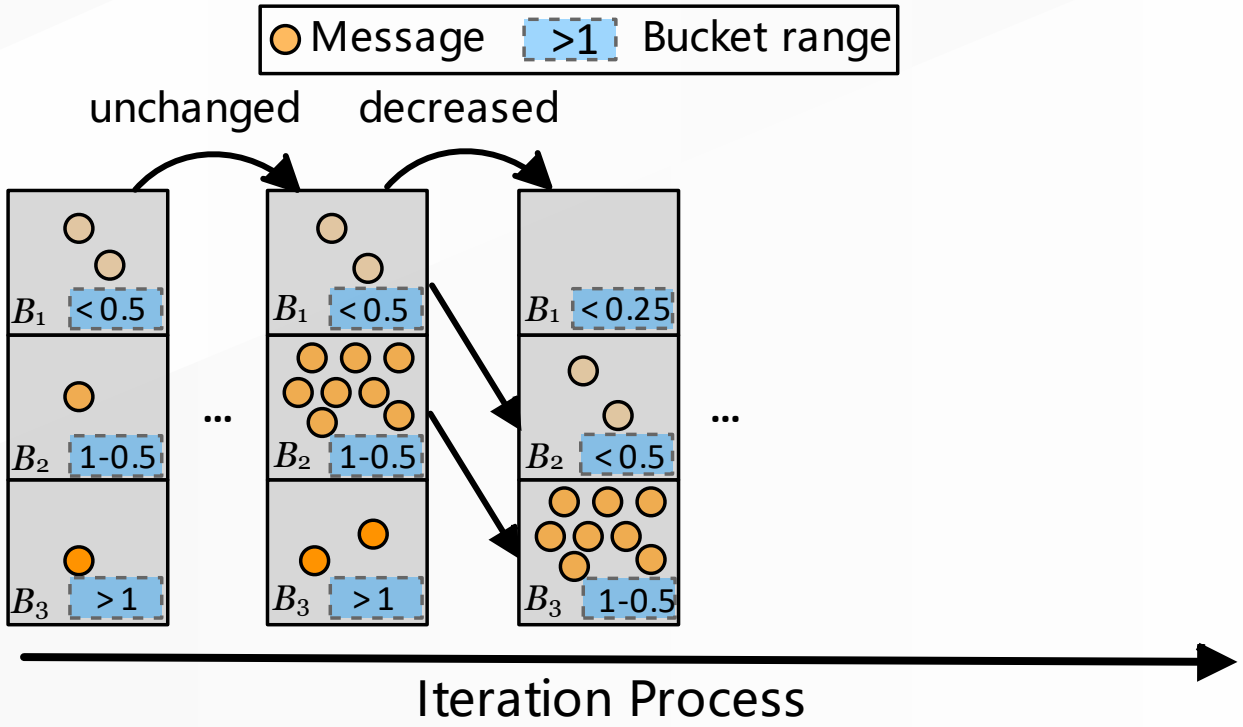
2 Adaptive buckets for message filtering



Dynamically change the range of buckets as the iterative progresses

Two Runtime Optimization

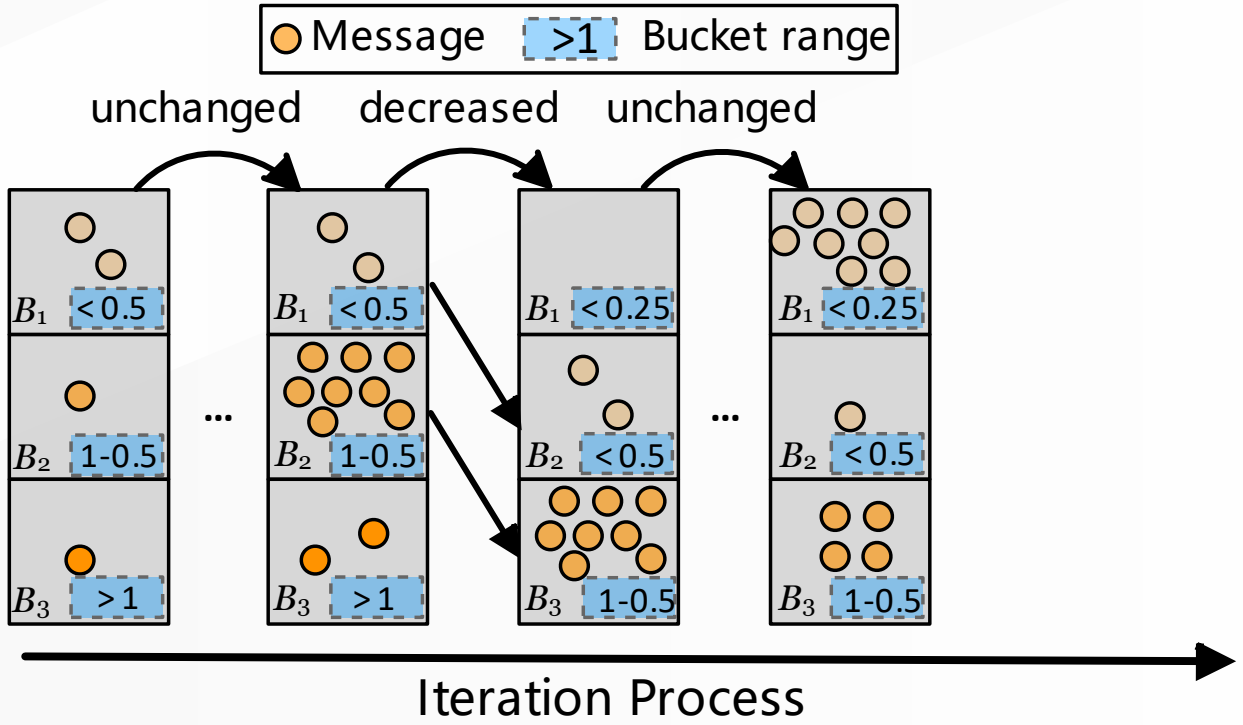
2 Adaptive buckets for message filtering



Dynamically change the range of buckets as the iterative progresses

Two Runtime Optimization

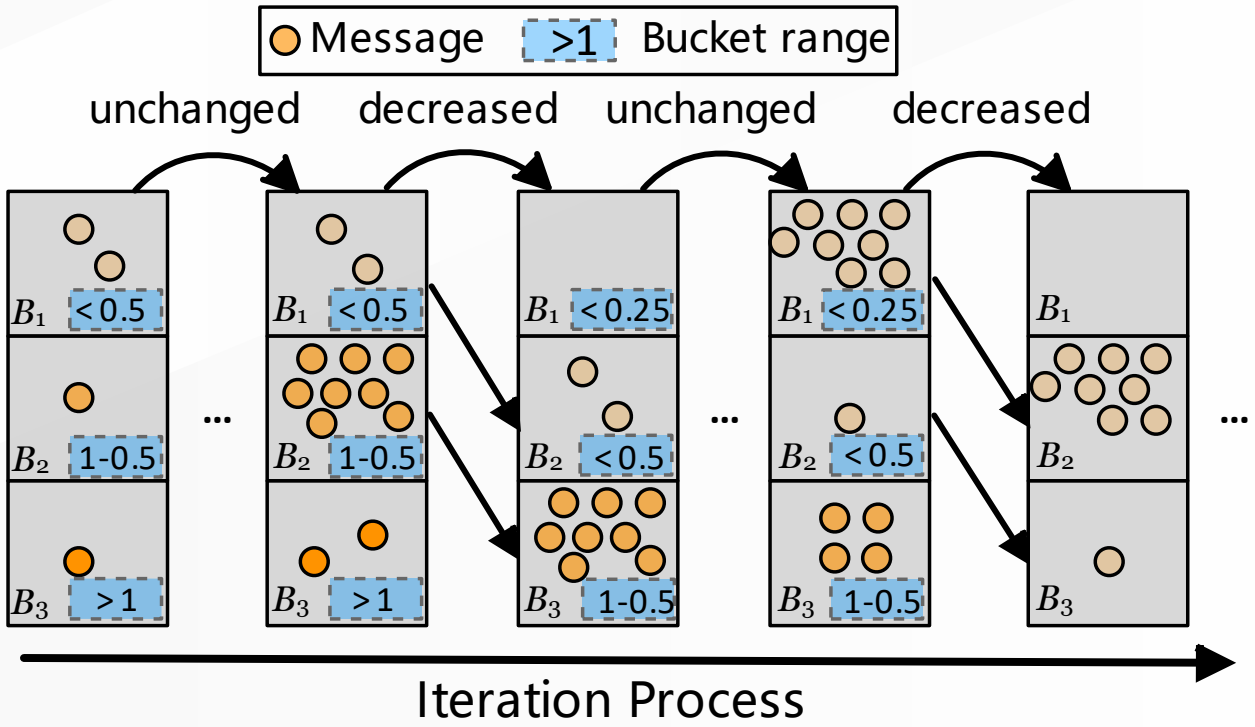
2 Adaptive buckets for message filtering



Dynamically change the range of buckets as the iterative progresses

Two Runtime Optimization

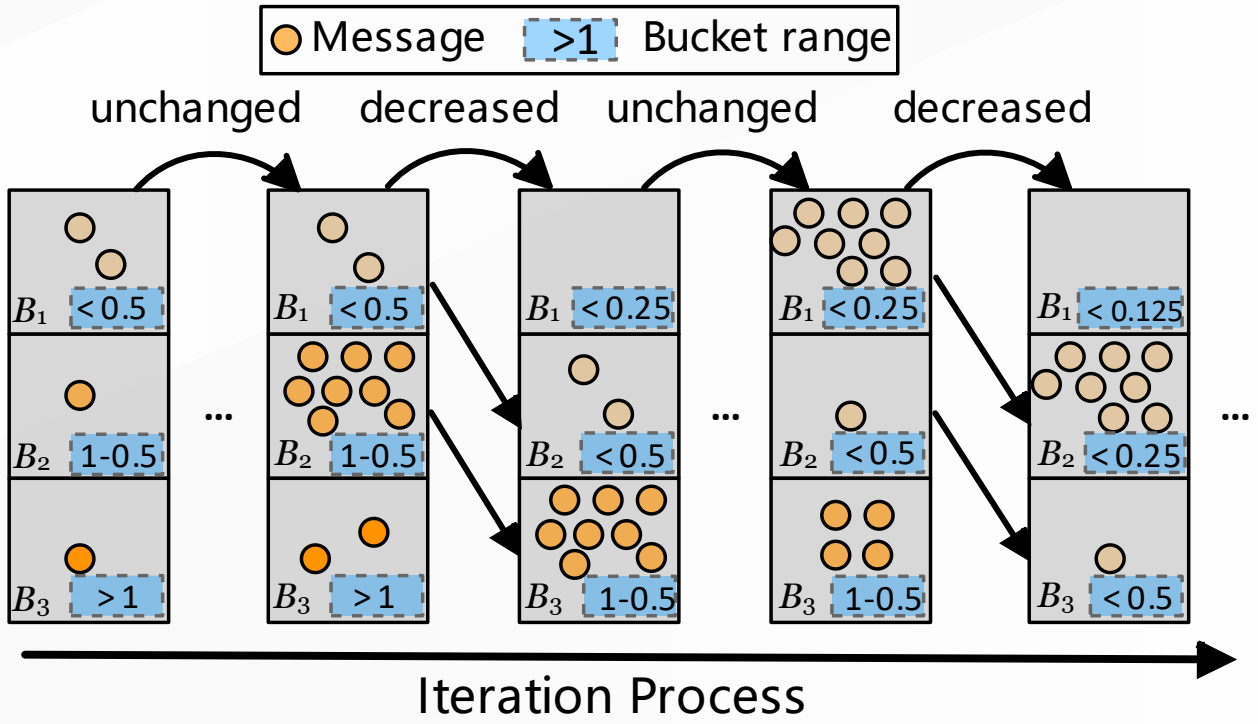
2 Adaptive buckets for message filtering



Dynamically change the range of buckets as the iterative progresses

Two Runtime Optimization

2 Adaptive buckets for message filtering



Dynamically change the range of buckets as the iterative progresses

Experiments

- Competitors

  GRAPE, Monarch, GeoGraph,

- Workloads

  PageRank, SSSP, BFS, PHP

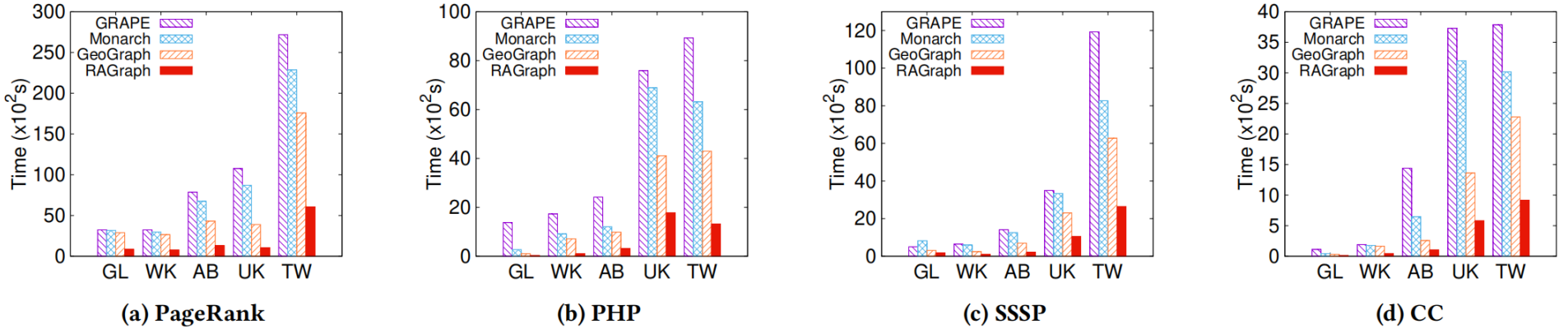
- Environment

  AliCloud ECS clusters from five regions are chosen as geo-distributed data centers, including Qingdao, China; Singapore; Sydney, Australia; Frankfurt, Germany; Virginia, USA.

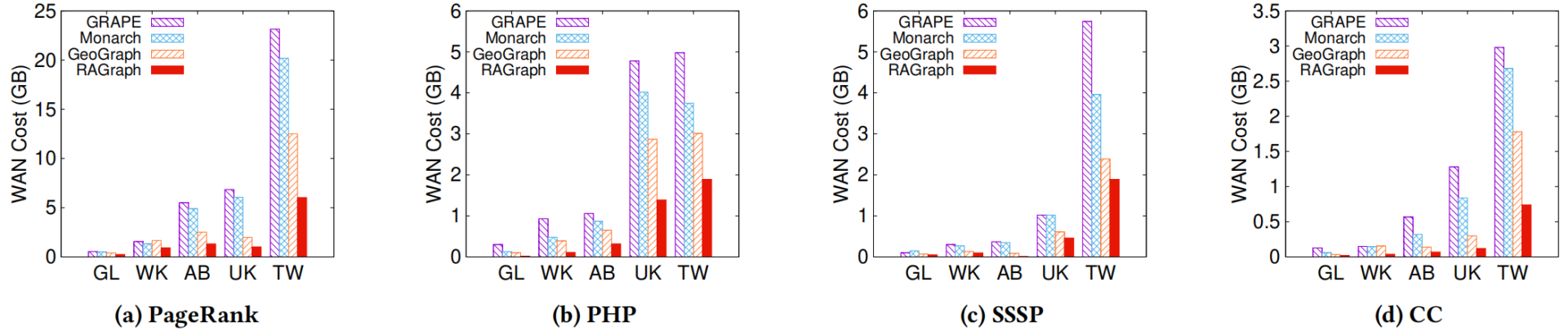
- Datasets

Graph	Vertices	Edges	Abbreviation
Web-Google [1]	916,428	6,078,250	GL
Enwiki-2013 [13]	4,203,323	101,311,614	WK
Arabic-2005 [2]	22,744,080	639,999,458	AB
UK-2005 [3]	39,459,925	936,364,282	UK
Twitter-2010 [12]	41,652,230	1,468,364,884	TW

Overall performance



Running time comparison.



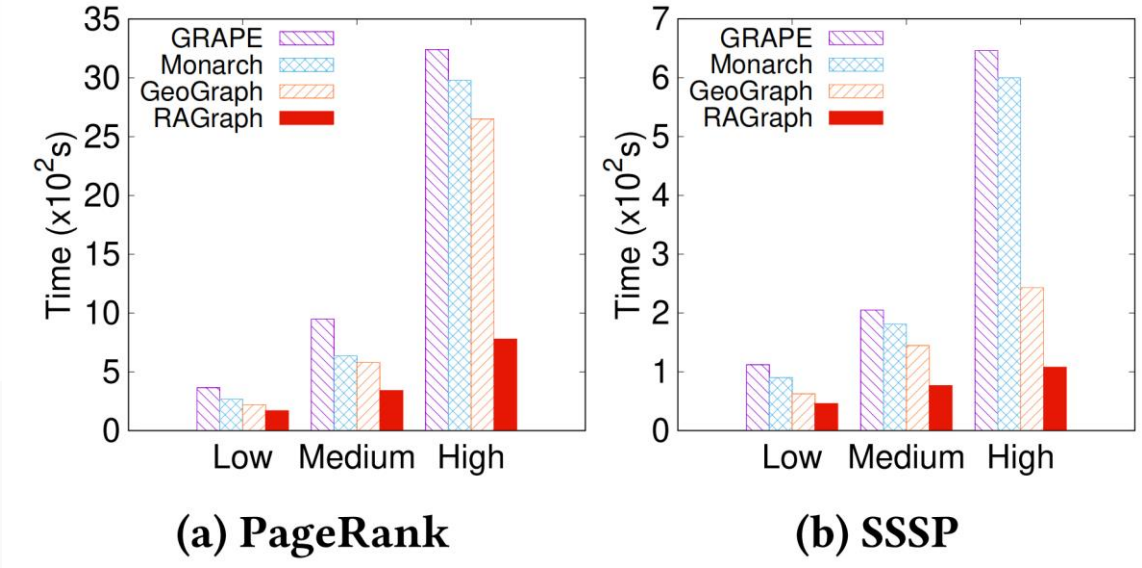
WAN cost comparison.

Achieve a **1.69x - 40.53x** speedup and a **20.9% - 97%** reduction in WAN cost

VLDB2024
GUANGZHOU

Sensitivity to Network Heterogeneity

We use different data center locations around the world to build low/medium/high-heterogeneity networks

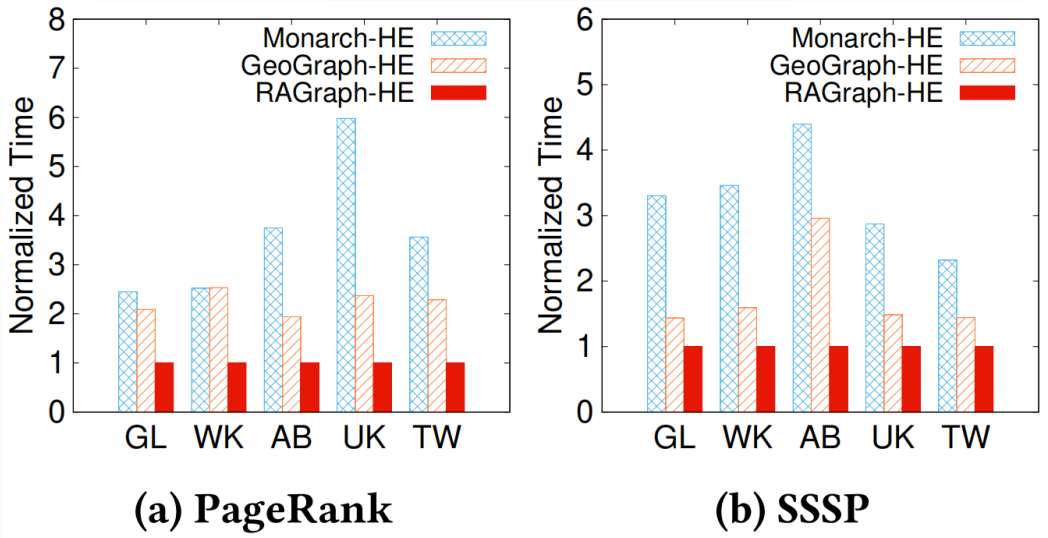


Sensitivity to network status.

RAGraph shows substantial superiority on the high-heterogeneity network.

Performance on Homomorphic Encryption

RAGraph provides homomorphic encryption (HE) interfaces to protect the users' data from other parties.



Performance on HE.

RAGraph requires a shorter running time on the HE module

Summary

RAGraph: A Region-Aware Framework for Geo-Distributed Graph Processing.

- **Providing three observations under geo-distributed environments.**

  We (1) allow advancing inefficient global updates to local computation, (2) design a two-layer coordination-free interaction view to, and (3) mitigates the impact of network congestion by replacing communication roles.

Summary

RAGraph: A Region-Aware Framework for Geo-Distributed Graph Processing.

- **Providing three observations under geo-distributed environments.**

  We (1) allow advancing inefficient global updates to local computation, (2) design a two-layer coordination-free interaction view to, and (3) mitigates the impact of network congestion by replacing communication roles.

- **Proposing two runtime optimizations.**

  We conduct an adaptive hierarchical interaction engine to adapt to heterogeneous networks and a discrepancy-aware message filtering strategy to dynamically filter unimportant messages.

Summary

RAGraph: A Region-Aware Framework for Geo-Distributed Graph Processing.

- **Providing three observations under geo-distributed environments.**

  We (1) allow advancing inefficient global updates to local computation, (2) design a two-layer coordination-free interaction view to, and (3) mitigates the impact of network congestion by replacing communication roles.

- **Proposing two runtime optimizations.**

  We conduct an adaptive hierarchical interaction engine to adapt to heterogeneous networks and a discrepancy-aware message filtering strategy to dynamically filter unimportant messages.

- **Delivering a fast Geo-Distributed Graph Processing system**

  We design and implement RAGraph, a region-aware geo-distributed graph processing system that achieves 1.69x – 40.53x speedup and 20.9% - 97% WAN cost reduction.

  - **The codes are publicly available on github**

    https://www.github.com/farisyao/RAGraph.

Thank you for listening!