S^3-TM: Scalable Streaming Short Text Matching

Fuat Basık, Buğra Gedik, Hakan Ferhatosmanoğlu, and Mert Emin Kalender
Bilkent University Computer Engineering Department Ankara, Turkey.

Introduction

In microblogs, the content of the post is irrelevant when following other users.

Content based matching is necessary to monitor microblogs for relevant information. We address this problem, under the content based pub/sub model.

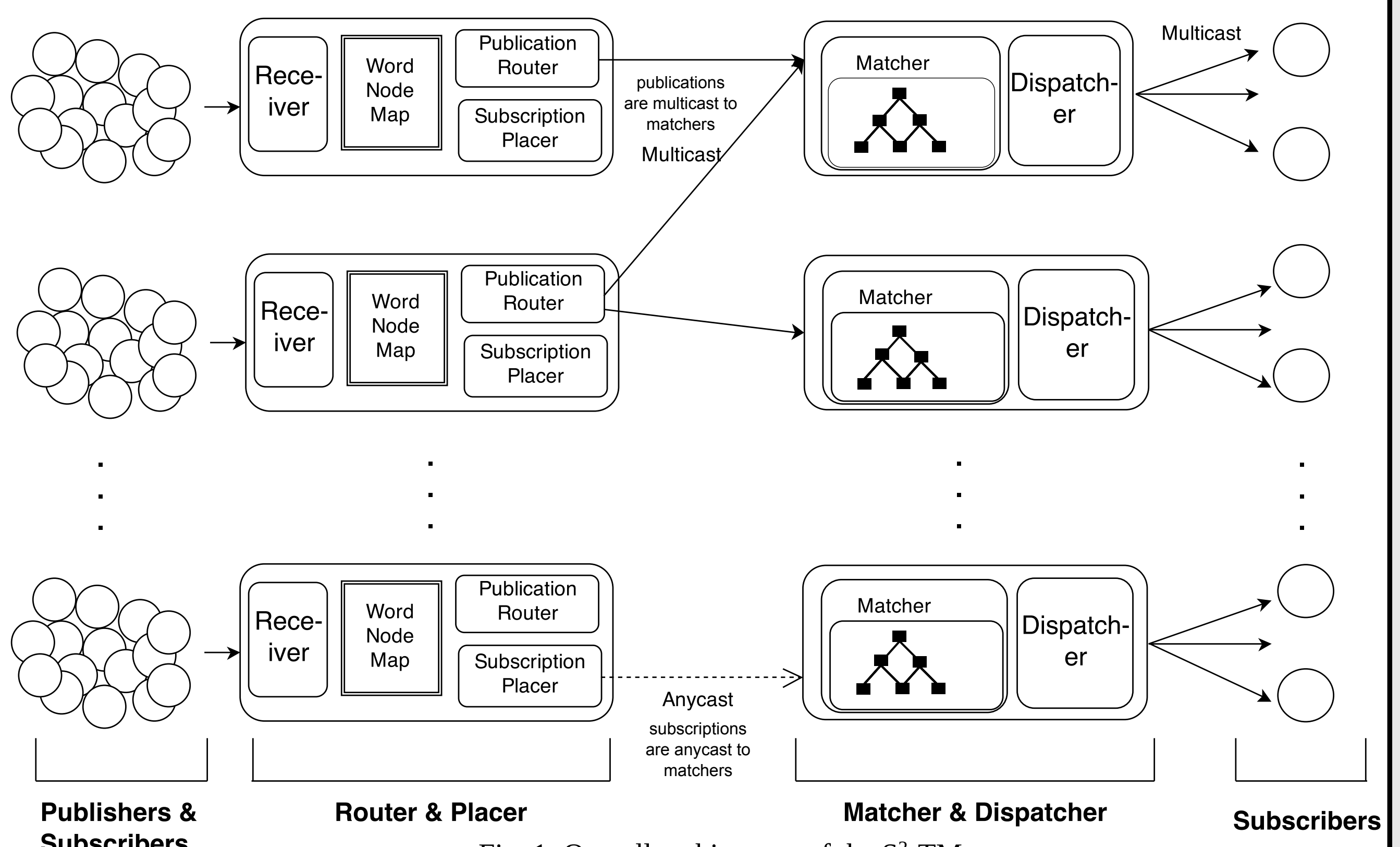


Fig. 1: Overall architecture of the S^3-TM system

Building a scalable infrastructure for content based matching is a challenge, given the popularity of Twitter and Weibo. The solution is **S^3-TM**: organized as a stream processing application, in the form of a data parallel flow graph designed to be run on a data center environment.

Results

Scalability

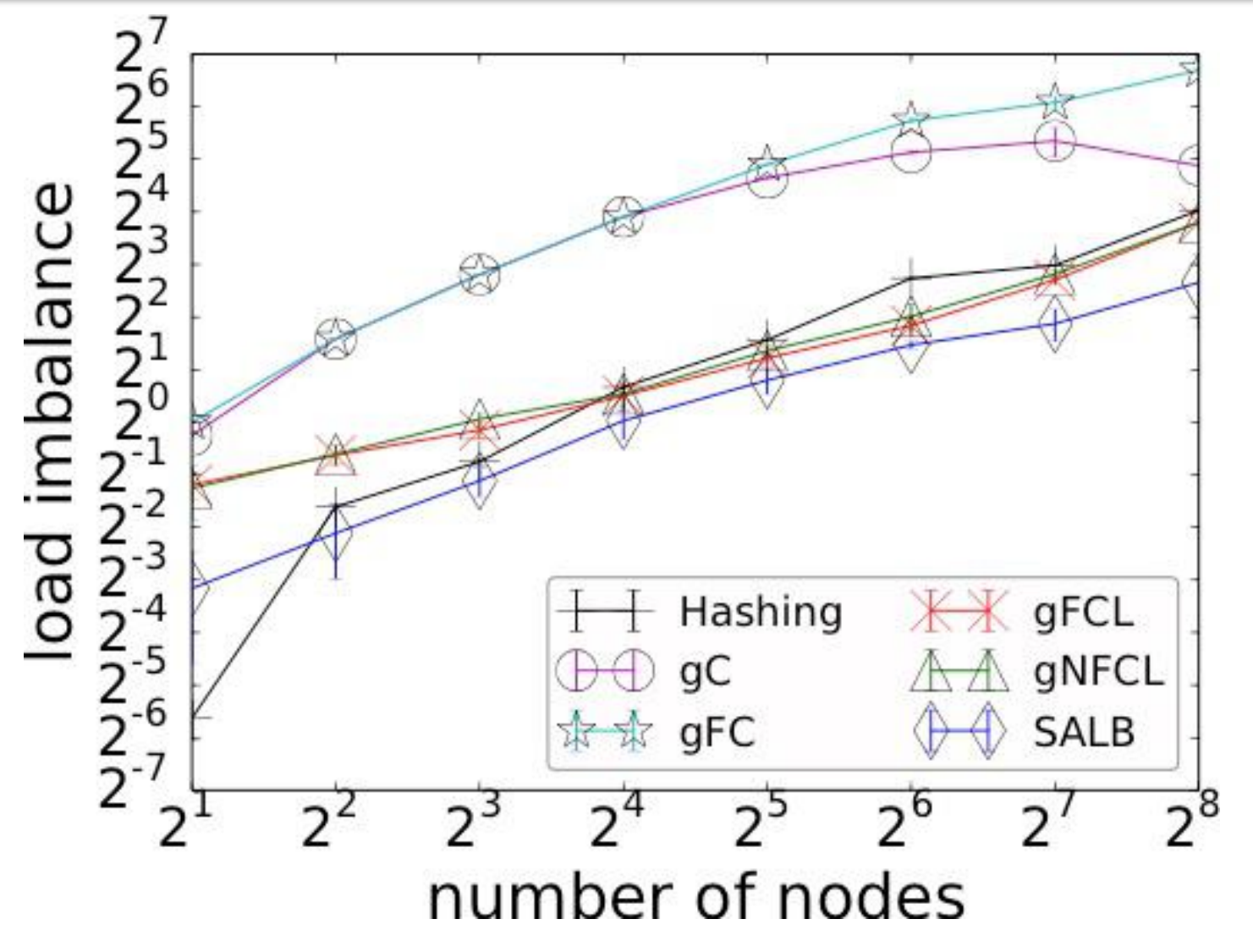


Fig 2: Load Imbalance as a function of number of nodes

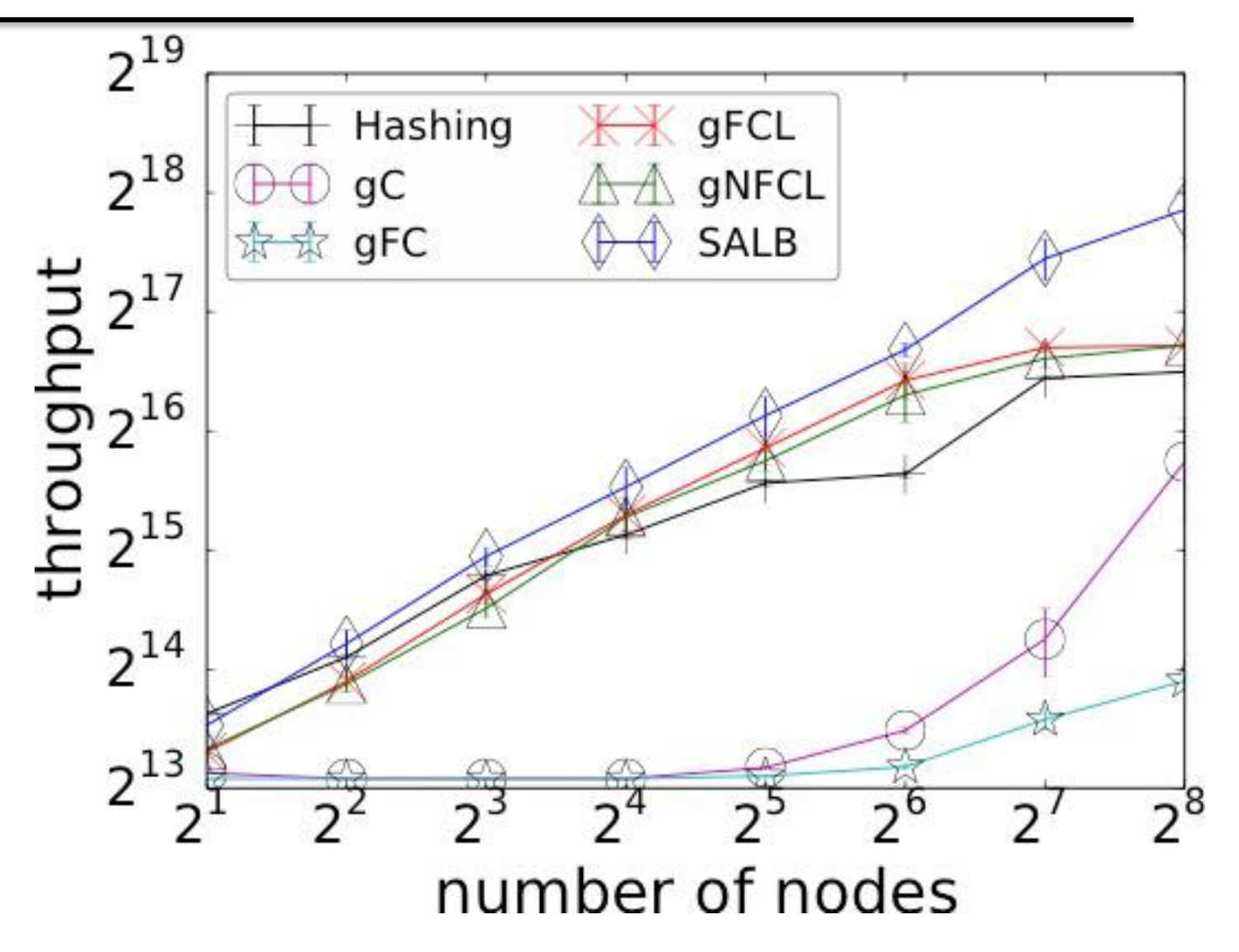


Fig 3: Throughput as a function of number of nodes

Subscription Awareness

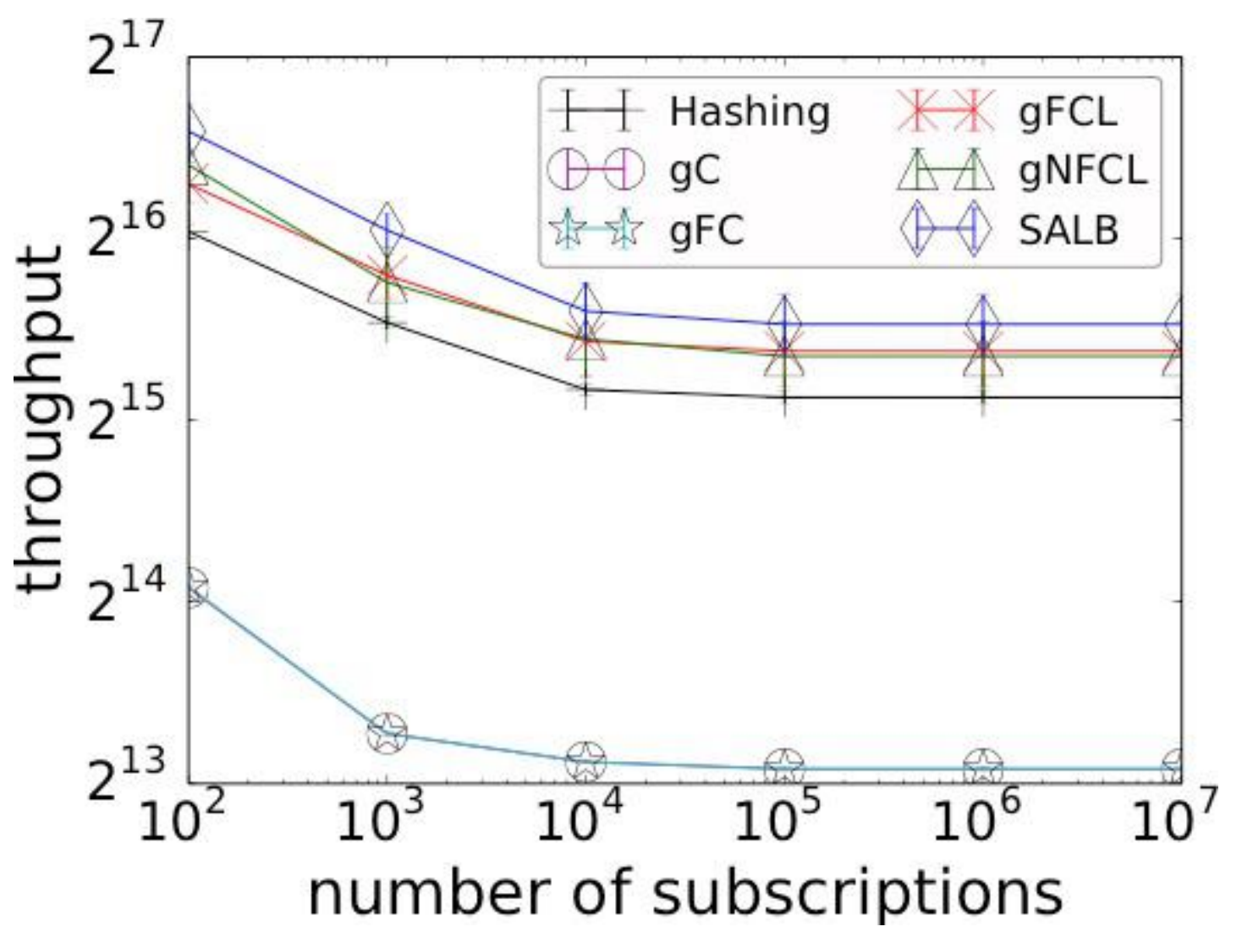


Fig 4: Throughput as a function of number of subscriptions

Concept Drift

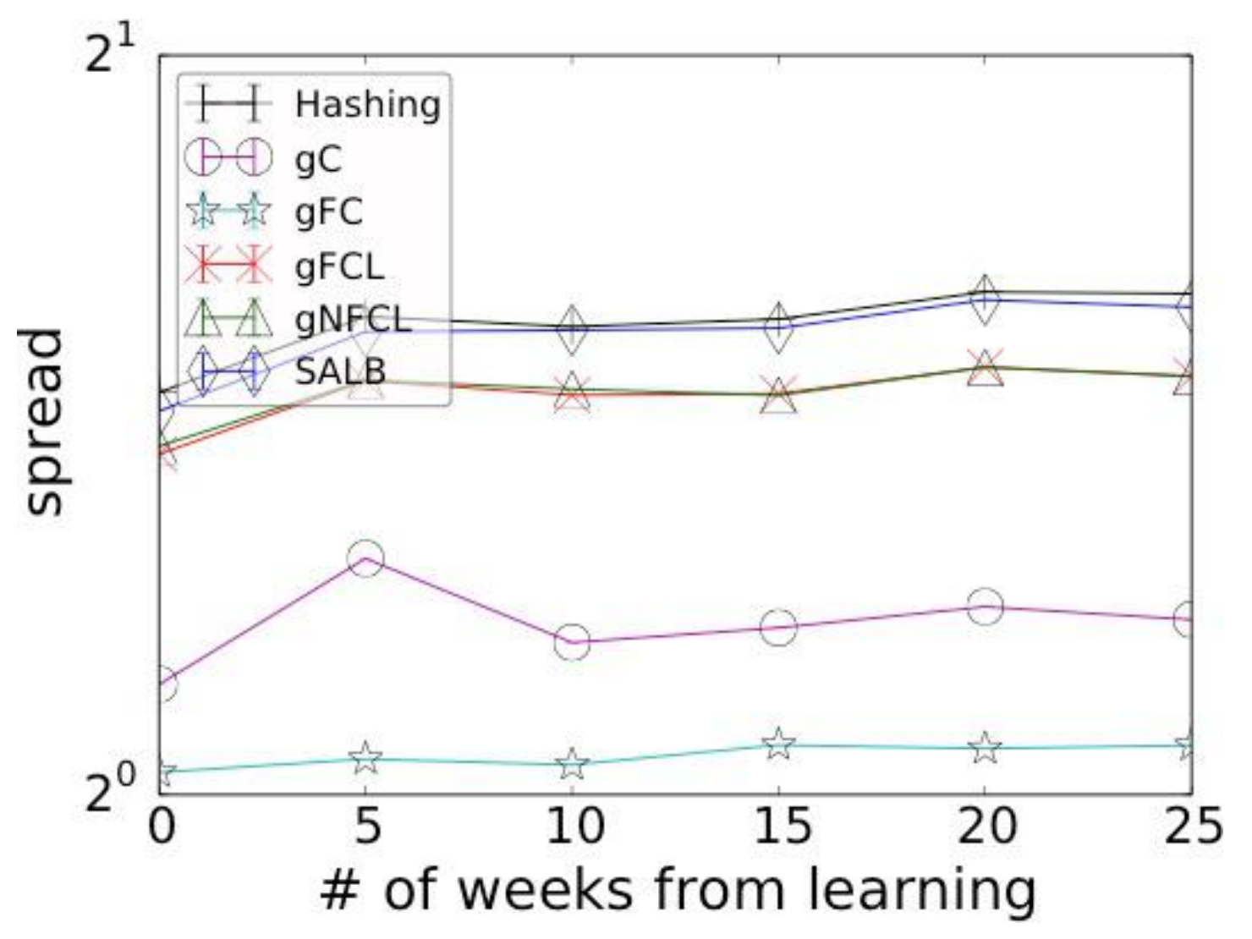
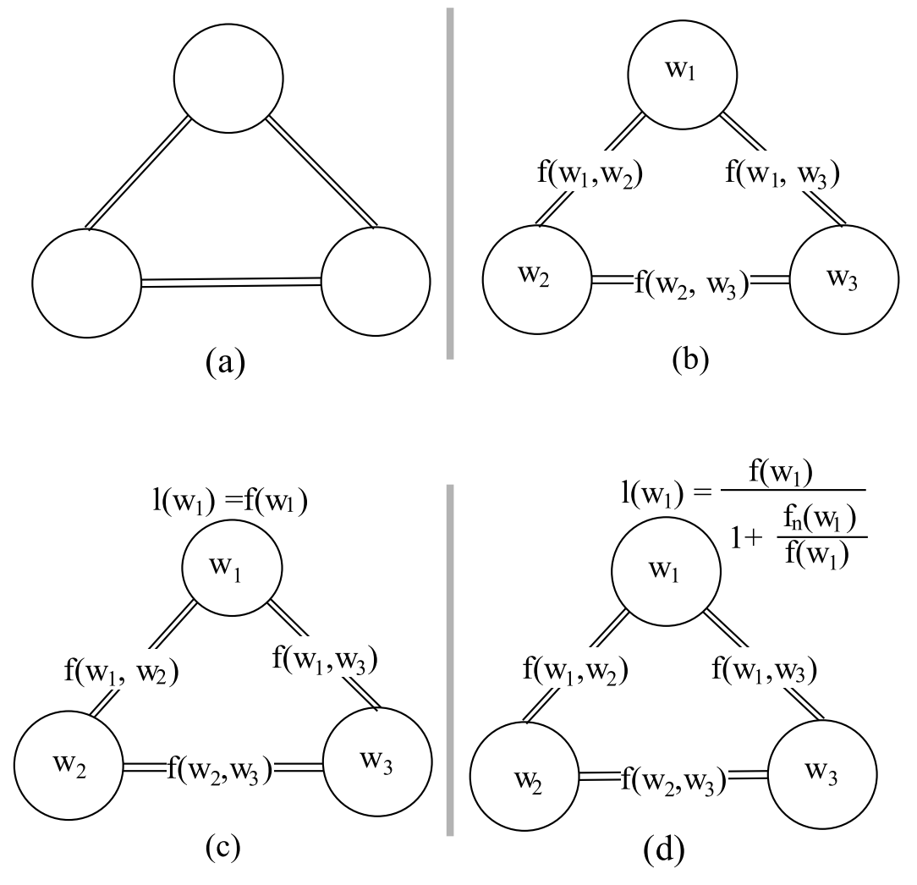


Fig 5: Spread of publications to machines

Publication Routing

Graph Partitioning Based Mapping



Spread Aware Load Balancing (SALB)

```
for each word w
    for each matcher i
        compute the extra load w brings to matcher i
        compute utility (spread, and load imbalance)
        if better mapping
            update best utility
            update best matching
    update load of the matcher
    update spread
    add new mapping
return constructed mapping
```

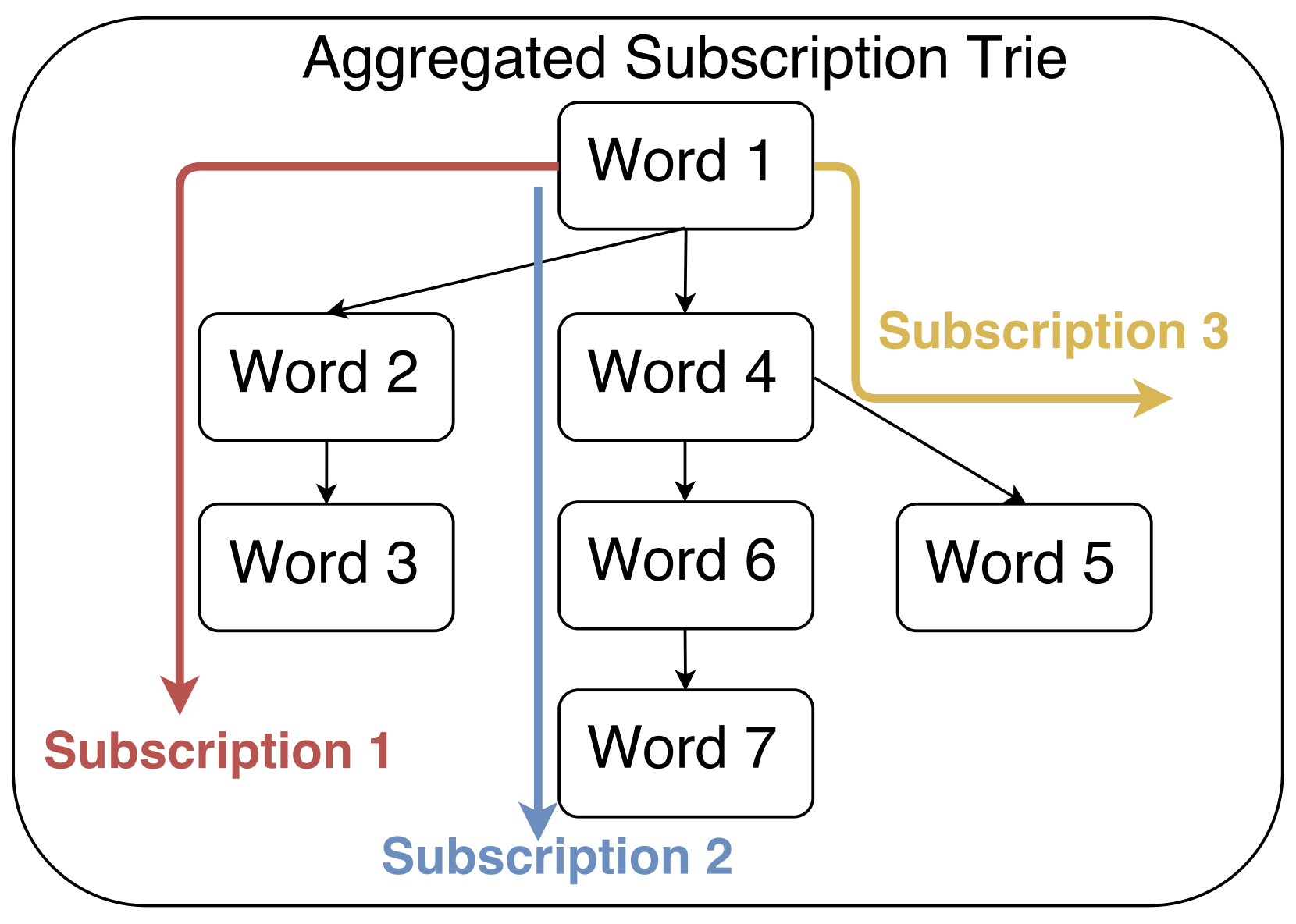
Subscription Placement

Load-Aware Subscription Placement (LASP)

```
find the set of eligible matchers
for each eligible matcher i
    compute subs. delta load
    compute imbalance
    compute utility
    if better mapping
        update best utility
        update best placement
return best matcher
```

Matching

Aggregated Subscription Trie



Discussions

Publication routing, load imbalance, subscription placement & matching, skew handling, overload and load shedding are the challenges on providing scalable short text matching.

By benefiting from the problem domain, we generate a word-to-node mapping to avoid broadcasting.

Graph partitioning based approaches felt short on modeling the load.

SALB increase the throughput by a factor of **2.5x** compared to a baseline multicast approach.

Related Work

- Publish/subscribe systems
  S^3-TM is a variation of content based systems
  Tibco[1], and Scribe[2] are well known examples
- Wide-area network pub/sub systems[3,4]
  Those systems use broker.
  S^3-TM runs on a datacenter environment
- Tightly coupled pub/sub systems[5]
  S^3-TM learns from previous publications to avoid broadcasting
- Filtering and matching
  S^3-TM uses tree based matching algorithm[1]

Literature cited

1- TIBCO Inc., Tib/rendezvous. White Paper (1999)

2- Castro, M., Druschel, P., Kermarrec, A.M., Rowstron, A.I.: Scribe: A large-scale and decentralized application- level multicast infrastructure. IEEE Journal on Selected Areas in Communications (JSAC) 20(8), 1489– 1499 (2006)

3-*Aguilera, M.K., Strom, R.E., Sturman, D.C., Astley, M., Chandra, T.D.: Matching events in a content-based sub- scription system. In: ACM Symposium on Principles of Distributed Computing (PODC) (1999)*

4- Ramasubramanian,V.,Peterson, R.,Sirer,E.G.:Corona: A high performance publish-subscribe system for the world wide web. In: USENIX Conference on Networked Systems Design & Implementation (NSDI) (2006)

5-Barazzutti, R., Felber, P., Fetzer, C., Onica, E., Pineau, J.F., Pasin, M., Rivi`ere, E., Weigert, S.: Streamhub: A massively parallel architecture for high-performance content-based publish/subscribe. In: ACM International Conference on Distributed Event-based Systems (DEBS), pp. 63–74 (2013)

Acknowledgements

This study was funded in part by The Scientific Technological Research Council of Turkey (TUBITAK) under grants EEEAG #111E217 and #112E271.