PEERIALISM

A peer-to-peer, decentralized protocol for k-Leader Election

Giovanni Simoni [giovanni@peerialism.com]

Abstract

The k-Leader election problem consists in identifying, in the context of a distributed network, a subset of *k* supernodes to be assigned to a certain application-dependent role. Such service is useful for several distributed applications, e.g. to identify the nodes that could be capable to store *k*-replicas of a large piece of content.

This paper proposes *RankSlicing*, a pragmatic and general purpose decentralized solution which aims at real-world deployment. The algorithm allows an applicative logic to specify the requirements for nodes to be elected as supernodes, and provides each node with an identical set composed of *k* supernodes, along with a measure of its reliability. RankSlicing quickly adapts to both global and node-wise dynamics.

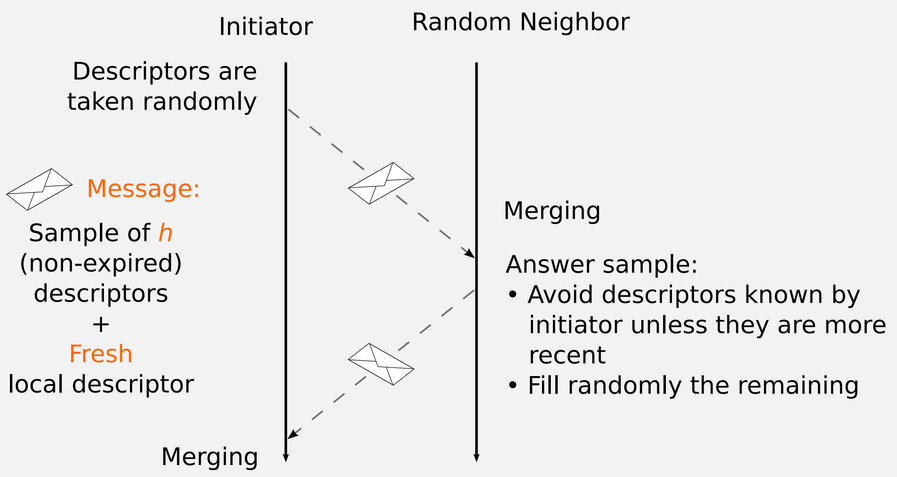
Context and Motivation

- A subset of the network (super-nodes) providing a service for other nodes.
- Selection based on node capabilities
- Use case: live video streaming
  → Selection k nodes available for forwarding the live video
  → Parametrized over nominal network throughput
- Use case: bounded number of content replicas
  → Selection of k nodes which will store a copy
  → Parametrized over available storage space

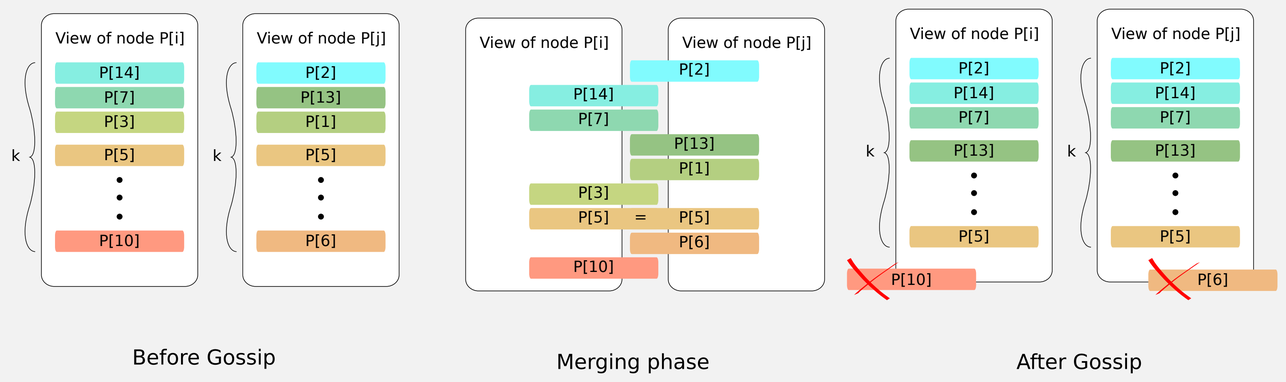
Goals:

- **Adaptiveness**: The leaders set changes according to the network dynamics
- **Stability**: Least possible disruption for the application
- **Local reliability**: Index of the quality for the result, computed in a decentralized way
- **Resiliency**: Faulty nodes are removed from the leaders set
- **Convergence**: The same leaders set is eventually chosen by the whole network
- **Suitability for the Internet**: No assumption on the overlay, communication only to direct neighbors.

Two-phases Gossip



Gossip Session



Algorithm evaluation

- A quality index is computed with a distributed algorithm similar to distributed aggregation. The result is an approximation of:

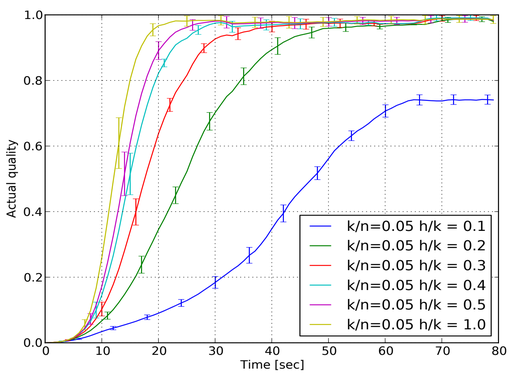
$$\frac{1}{n}\sum_{i=0}^{n}\frac{|V_i \cap V_{opt}|}{k}$$

  Where V_i is the view of the node P[i], and V_{opt} is an optimal leaders set;
- The approximation: at every gossip cycle we have a better view, hence V_{opt} can be replaced with it. This results in an optimistic evaluation;
- A weighted average allows both to make a more realistic estimation and to smooth the oscillation of the approximated evaluation.

Algorithm parameters

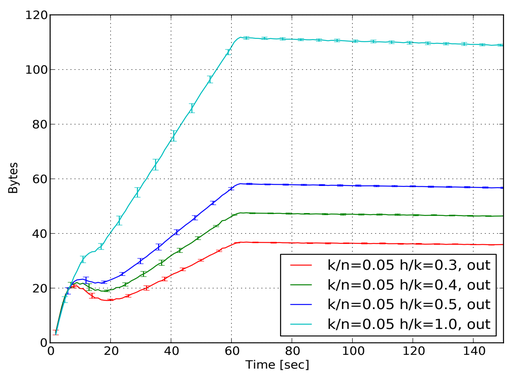
- **P(i)**: Eligibility Predicate: only eligible nodes can be part of the leaders set. All nodes can propagate node descriptors, but only eligible ones can emit them.
- **R(i)**: Rank evaluation function: aggregates the characteristics of the nodes in a representation allowing a comparison between nodes. The output is embedded into node descriptors.
- **T**: Gossip period: longer period translates in slowest convergence, but also in reduced network bandwidth requirement.
- **k**: Size of the leaders set: application dependent.
- **h**: Size of the sample shared during a Gossip session. The trade-off is between convergence speed and network bandwidth requirement
- **α**: Smoothing factor for the approximated quality measure, allows to obtain a good approximation of the leader set quality;
- **PAL**: Propagation Age Limit: as a countermeasure against churn, descriptors are continuously renovated, and the older ones (whose age exceed the PAL parameter) get removed from the system

Experimental evaluation:



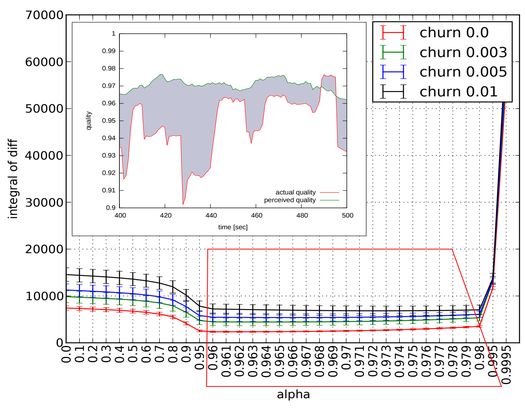
Convergence speed: 20-30 seconds with the tested experimental setting.

Shown: convergence behavior with different values of k/n and h/k. Churn: 0.3% nodes being replaced within 10 s)



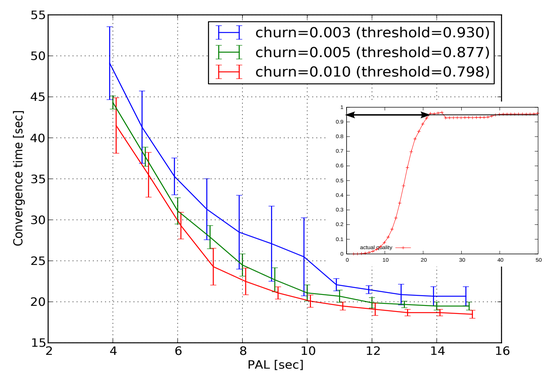
Bandwith usage below 120 bytes/ second.

Shown: bandwidth behavior of the most significant ratios k/n and h/k (the configuration is the same as before).



Parameter study for the α parameter: improving the perceived quality of the computed leaders set.

Comparison against a centralized computation, (feasible only in simulation).



Parameter study for the PAL parameter: time required to reach convergence.

The leader set is stable when a certain threshold quality is reached.