Consensus algorithms for distributed storage systems

The SEED[1] program (standard track)

www.imt-atlantique.fr/seed

PhD topic open for applications until January 31st, 2024

1 Definition

1.1 Domain and scientific/technical context

Low latencies connections and decentralized servers are currently showcasing a new potential for distributed computing. Particularly, mobility and intermittent connectivity of computing resources create a need for distributed storage mechanisms resilient to network isolation.

However, developing integrated systems that are capable of exploiting highly distributed resources requires developers and service providers to deal with the unreliability of the compute nodes and of the network infrastructure, and must be considered during the design phase of systems. Additionnally, more pressing constraints on energy and resource consumptions will foster the need for distributed computation with restricted capabilities, for example relying on small server nodes that are turned off or disconnected most of the time.

1.2 Scientific/technical challenges

Coordination and consensus problems are at the core of distributed algorithms. In the context of server-side infrastructure and especially highly-distributed storage systems, we identify two main contributions as part of this topic proposal :

[1]Co-funded by the European Union under Grant Agreement no. 101126644

1. **Leaderless consensus for server-side software**: many distributed algorithms deployed today rely on strong coordination and leader elections: this is a costly approach that is not compatible with unreliable compute nodes and network. In contrast, weak coordination approaches and leaderless consensus are appealing for this situation. However, they have mostly been applied in the context of client-side local-first applications. We plan to extend this work to develop better lightweight server-side distributed software with faster access time and lightweight resource consumption even in situations that make coordination challenging.

2. **CRDTs for storage and cluster systems**: when an even weaker form of coordination is tolerable, conflict-free replicated data types (CRDT) [3] provide good system support for intermittent connectivity. CRDTs are used in synchronisation schemes, as replicas can be updated independently and concurrently without direct coordination [4,5]. We plan to investigate and formalize the use of CRDTs in storage systems and virtual clusters.

1.3 Considered methods, targeted results and impacts

The main motivating use case for this work is the Garage software, an open-source distributed object storage service tailored for highly-distributed infrastructures [6]. Garage is developped by Deuxfleurs.

Garage already uses CRDTs to tolerate network disconnections, but would sometimes require a stronger coordination model. For example, Garage currently allows two users to create conflicting storage spaces on two different nodes, and the conflict is only discovered when the changes are propagated. A lightweight "leaderless consensus" approach would provide more guarantees, while being less sensitive to latency and network disconnections compared to Paxos or Raft.

Another interesting problem in Garage is : how to maintain the consistency of data replicas when storage nodes are added or removed? Any membership change may cause an update of the location of replicas, but it will take time to actually move the data. During this time, all nodes need to maintain the desired level of consistency while accounting for both the old and new location. This problem has not been thoroughly studied for the read-after- write consistency model used in Garage.

The proposed topic is expected to contribute to the algorithmic state-of-the-art around distributed storage systems, which would indirectly benefit all such systems. Another goal of the thesis is to implement the proposed solutions in the Garage software itself.

This work can also be applied to virtual clusters (Namespaces) in Kubernetes-like software stacks. Namespaces presents services, and deployments users use to build and run their applications. The ability to efficiently deploy a virtual cluster over geographically distributed resources could enable collaboration between containers by exposing a resource created on one site to another one with minimal code changes.

1.4 Interdisciplinarity aspects

The topic involves several fields: distributed algorithmics, implementation of these algorithms in an existing software (Garage), and applications to real-life large-scale distributed systems. It will also involve large-scale experiments using research platforms such as SLICES-RI.

1.5 References

1. Antoniadis, K., Benhaim, J., Desjardins, A., Poroma, E., Gramoli, V., Guerraoui, R., ... & Zablotchi, I. (2023). **Leaderless consensus**. Journal of Parallel and Distributed Computing, 176, 95-113.

2. Tennage, P., Basescu, C., Kokoris-Kogias, L., Syta, E., Jovanovic, P., Estrada-Galinanes, V., & Ford, B. (2023, October). **QuePaxa: Escaping the tyranny of timeouts in consensus**. In Proceedings of the 29th Symposium on Operating Systems Principles, SOSP (pp. 23-26).

3. M. Shapiro, N. Preguiça, C. Baquero, and M. Zawirski, "Conflict-free replicated data types," in Symposium on Self-Stabilizing Systems, 2011, pp. 386–400.

4. X. Lv, F. He, Y. Cheng, and Y. Wu, "A novel CRDT-based synchronization method for real-time collaborative CAD systems," Adv. Eng. Inform., vol. 38, pp. 381–391, Oct. 2018, doi: 10.1016/j.aei.2018.08.008.

5. J. Bauwens and E. Gonzalez Boix, "Memory efficient CRDTs in dynamic environments," in Proceedings of the 11th ACM SIGPLAN International Workshop on Virtual Machines and Intermediate Languages, New York, NY, USA, Oct. 2019, pp. 48–57, doi: 10.1145/3358504.3361231

6. `https://garagehq.deuxfleurs.fr/`

2 Partners and study periods

2.1 Supervisors and study periods

- **IMT Atlantique**: Dr. Daniel Balouek, Inria research scientist, IMT Atlantique, Nantes, France

  The PhD student will stay 30 months at Dr. Balouek's lab.

- **International partner**: The PhD student will probably be hosted three months at the SCI institute of the University of Utah. However, this partner may still change.

- **Industrial partner(s)**: Dr. Alex Auvolat, research engineer, DeuxFleurs

  The PhD student will be hosted three months at Deuxfleurs.

2.2 Hosting organizations

2.2.1 IMT Atlantique

IMT Atlantique, internationally recognized for the quality of its research, is a leading French technological university under the supervision of the Ministry of Industry and Digital Technology. IMT Atlantique maintains privileged relationships with major national and international industrial partners, as well as with a dense network of SMEs, start-ups, and innovation networks. With 290 permanent staff, 2,200 students, including 300 doctoral students, IMT Atlantique produces 1,000 publications each year and raises 18€ million in research funds.

2.2.2 Deuxfleurs

Deuxfleurs is a French association that is working towards changing the Internet experience towards a convivial Internet.