Towards Large Scale Replicated Databases

Ricardo Jiménez-Peris[1], Marta Patiño-Martínez[1], Damián Serrano[1], Bettina Kemme
Universidad Politécnica de Madrid McGill Univ.
{rjimenez,mpatino,dserrano}@fi.upm.es kemme@cs.mcgill.ca

Databases are core components of modern distributed multi-tier information systems. In most of the cases, the database is the bottleneck of the system, limiting the scalability of the full system. This is the reason why database replication has demanded much attention during the last few years. However, current replication protocols rely on two facts that inherently limit their scalability to a few tens of sites. This position paper aims at addressing the three main limiting factors of the scalability of replicated databases. Firstly, current approaches are based on 1-copy serializability (1CS). 1CS as its centralized counterpart, serializability, limits the concurrency in the database and therefore, it also limits its throughput. Secondly, current approaches rely on full replication what also has an inherent limit of scalability. Thirdly, as any replication approach there is some coordination overhead due to the required communication (e.g. group communication) that consumes considerable resources at every site. In this position paper we explore how to overcome these three inherent limits of scalability in current approaches to enable to scale one order of magnitude higher.

Serializability hurts concurrency due to read-write conflicts that are extremely frequent. Snapshot isolation [Berenson95] exploits multi-version data and optimistic concurrency control to provide each transaction with a consistent snapshot and at the same time avoiding read-write conflicts. Therefore, since write-write conflicts at the tuple level are very infrequent it results in high concurrency. Additionally, it satisfies the tests required by the ANSI serializability isolation level. For these reasons, Snapshot isolation has become the preferred isolation level, and it is offered by most database vendors (Oracle, Microsoft, Postgres, Borland, etc.), being the highest isolation level provided by some databases such as Oracle or PostgreSQL. Most database replication protocols are based on the traditional correctness criteria, 1CS, what has limited its scalability to a few tens of replicas [Kemme00, Patiño05]. Recently, it has been proposed a new correctness criterion, 1-Copy Snapshot Isolation (1CSI) to overcome the lack of scalability of 1CS [Lin05]. One of our claims is that 1CSI will be crucial to provide consistency and attain large levels of scalability.

As aforementioned most approaches also resort to full replication. That is, each site contains a full copy of the replicated database. Most of the replication protocols follow the ROWAA approach (Read One, Write All Available) [Bernstein87]. That is, read-only transactions (queries) are processed by a single site what enables high scalability for read-only workloads. While update transactions (update, insert or modify operations) are executed at all available sites. There are two ways to process updates in a replicated database: symmetric and asymmetric processing. With symmetric processing, an update transaction is fully executed by all sites. In contrast, with an asymmetric approach, an update transaction is processed by one site and the resulting updated tuples are propagated to the remaining sites. Since propagating and installing the updates is significantly cheaper than executing the full transaction, this saves a significant amount of processing capacity in the replicated database. Asymmetric processing is vital to attain scalability with update workloads, even with small update ratios, 10-20%, [Patiño05,Jimenez03].

However, even using asymmetric processing, full replication results in medium scalabilities of a few tens of sites. Let us see why. With 2 sites, most of the capacity of each site is devoted to process full transactions and only a small fraction to install the updates generated by the other

[1] Research partially funded by Microsoft Research Cambridge under a Euro-PhD Award, by the Spanish Ministry of Science and Education (MEC) under grant TIN2004-07474-C02-01, and the Madrid Regional Science Foundation (CAM) with partial support of FSR and FEDER under grant S-0505/TIC/000285.

site. However, with 15-20 sites most of the capacity of each site is devoted to process the updates generated by all other sites, and just a fraction is devoted to process full transactions. When adding a new replica, most of its capacity is spent to process the updates from the other replicas, and the spare capacity is compensated with the fact that the site is consuming some capacity from all the other replicas to apply the updates it produces. In order to scale beyond that, it is necessary to resort to partial replication. That is, not all sites contain a full copy of the database.

Partial replication raises several challenges. We have modeled it analytically to understand the potential gains in scalability. The analytical model consists of an equation system constrained by a linear program. The equation system models the interplay between the different sites, that is, the relation of the effective load (the actual number of transactions fully processed) of a site with the effective loads of the other sites (that generate updates that consume capacity from the site). The linear program sets constraints to avoid solutions that consume more capacity at a site than the one available, and also enables to set a goal that maximizes the effective load and therefore, the scalability.

If pure partial replication (no full replicas) is used, it might happen that there is no single site that contains all the data to be accessed by a transaction what results in distributed transactions, which are known to not scale due to the need of distributed atomic commitment. One possible solution is a hybrid replication solution combining partial and full replicas. The former would provide scalability and the latter would avoid distributed transactions. After studying hybrid replication analytically we have concluded that hybrid replication scales a little more than full replication but nothing more. Our second claim is that pure partial replication is needed to attain large scalability of database replication. One of the challenges is how to deal with distributed transactions in a scalable way, novel high performance distributed atomic commitment protocols such as [Jimenez01] would help in achieving the necessary levels of scalability.

Finally, there is the matter of the overhead of coordination inherent to any replication approach. The creation of messages, its serialization/deserialization, sending/receiving messages consumes a non-neglible amount of resources. Although this question, if handled carefully, is not a bottleneck for a few tens of sites, it becomes an issue with higher orders of magnitude. Our third claim is that by using modern system area networks such as Myrinet is possible to effectively offload part of this coordination overhead to attain larger levels of scalability.

References

[Berenson95] Hal Berenson, Philip A. Bernstein, Jim Gray, Jim Melton, Elizabeth J. O'Neil, Patrick E. O'Neil. A Critique of ANSI SQL Isolation Levels. SIGMOD Conference 1995. pp. 1-10.

[Bernstein87] Philip A. Bernstein, Vassos Hadzilacos, Nathan Goodman: Concurrency Control and Recovery in Database Systems. Addison-Wesley. 1987

[Jimenez01] Ricardo Jiménez-Peris, Marta Patiño-Martínez, Gustavo Alonso, Sergio Arévalo: A Low-Latency Non-blocking Commit Service. DISC 2001. pp. 93-107.

[Jimenez03] Ricardo Jiménez-Peris, Marta Patiño-Martínez, Gustavo Alonso, Bettina Kemme: Are quorums an alternative for data replication? ACM Trans. Database Syst. 28(3):257-294. 2003.

[Kemme00] Bettina Kemme, Gustavo Alonso: Don't Be Lazy, Be Consistent: Postgres-R, A New Way to Implement Database Replication. VLDB 2000. pp. 134-143.

[Lin05] Yi Lin, Bettina Kemme, Marta Patiño-Martínez, Ricardo Jiménez-Peris: Middleware based Data Replication providing Snapshot Isolation. SIGMOD Conference 2005. pp. 419-430.

[Patiño05] Marta Patiño-Martínez, Ricardo Jiménez-Peris, Bettina Kemme, Gustavo Alonso: Middle-R: Consistent database replication at the middleware level. ACM Trans. Comput. Syst. 23(4): 375-423 (2005)