A Framework for a DHT that Provides Transactions on Replicated Data

CoreGRID REP19 between CR41 ZIB and CR15 KTH, CR19 SICS - Final Report

Summary

During the exchange the work presented in "Atomic Commitment in Transactional DHTs" [1] was continued. Initially details of the algorithms described in the paper were discussed. The main work carried out during the exchange included simulations to evaluate the assumptions underlying the algorithms of the framework. The following researchers at SICS and KTH collaborated with the visiting researcher Monika Moser (ZIB): Seif Haridi, Ali Ghodsi (SICS) and Vladimir Vlassov, Tallat Mahmood, Ahmad Al-Shishtawy (KTH).

Applicants: Monika Moser hosted by Seif Haridi and Vladimir Vlassov
Exchange: 2 weeks from Oct 22nd, 1 week at SICS and 1 week at KTH
Linked to: Institute on Architectural Issues: Scalability, Dependability, Adaptability (WP4)

Work Done

In [1] we presented an outline for a framework of a DHT that provides strong data consistency and transactions on its replicated data. Between the presentation of the paper and the exchange, many details of the algorithms were already worked out. At the beginning of the exchange we discussed the replica maintenance mechanisms of this framework. There items are replicated according to a symmetric replication scheme. As nodes in a DHT can fail, other nodes have to take over responsibility for the failed node's key range and copy the data they became responsible for in order to maintain the replication factor. This has to be done in a way that does not violate data consistency and thus take into account ongoing transactions.

To tolerate temporary unavailability of data, algorithms of the framework are majority based. They make progress as long as a majority or nodes involved in the algorithms are alive. Read and write operations on a certain item need

a majority of replicas of the item being available. Concurrent operations on an item have at least one replica in common where the conflict can be detected. Thus it is important that the number of replicas does not exceed the replication factor of the system. Otherwise concurrent operations could work on disjoint sets of replicas. However inconsistent lookups in a DHT can increase the number of replicas in the system and thus violate the assumptions of the algorithms.

Simulations that study the probability that a violation of the algorithms' assumptions occurs, constitute the main work done during the exchange. Inconsistent lookups are a result of some nodes' failure detectors behaving temporary faulty and thus resulting in nodes having wrong successor pointers. We simulated a Chord DHT and varied the probability that a failure detector produces a wrong suspicion of a node's successor node. We froze the system and counted all wrong successor pointers that can lead to an inconsistent lookup. Simulations were done with different churn rates and different ring maintenance periods. Additionally we analyzed the probability of two concurrent operations working on disjoint majority sets. Therefore we developed a formula that calculates this probability. This situation can happen if there exists an inconsistency for at least one replica of an item. From our results we can conclude that the probability for a violation of our algorithms' assumptions is very low. Thus it is reasonable to use majority-based algorithms in our transactional DHT.

The outcome of our work is included in a paper that will be submitted to the CoreGRID Integration Workshop 2008. Collaborations will further continue in order to present a complete framework that was outlined in [1].

Acknowledgment

This research work has been carried out under the FP6 Network of Excellence CoreGRID funded by the European Commission (Contract IST-2002-004265). I would like to thank Seif Haridi (SICS), Vladimir Vlassov (KTH), Tallat Mahmood (KTH), Ali Ghodsi (SICS) and Ahmad Al-Shishtawy (KTH) for their help and fruitful discussions.

References

[1] M. Moser, S. Haridi. "Atomic Commitment in Transactional DHTs". In *Proceedings of 1st CoreGRID Symposium*, Rennes, France, August, 2007