Replication and Distributed Hash Tables; the GRACE system

Distributed Hash Tables (DHTs) have emerged as the most promising infrastructure for next-generation Internet-scale applications, suggesting scalable lookup, ease of management, resiliency and consistent performance. In order for applications to fully take advantage of DHTs, however, data fault tolerance and availability in the face of node failures and flash behavior are also required. These characteristics cannot be achieved without the use of replication. Current services and applications use replication for data fault-tolerance, but they mainly restrict themselves to immutable data. We propose GRACE - the Global Replication And Consistency Environment - which is a distributed replication system that addresses the conflict between scalability, availability and performance in wide-area computing systems, provides support for a wide range of consistency semantics and takes into account the particularities of wide-area systems as heterogeneous and mobile users and changes in network conditions. Built on top of DHTs, GRACE inherits the resiliency of the routing infrastructure, and adds to this robustness and flexibility by providing support for replication and multiple consistency semantics.

This poster presents the need for a flexible and scalable scheme that supports different consistency semantics, and the novel architecture proposed by GRACE. Further, the relationship between GRACE and DHTs is highlighted and a replica location algorithm is presented.

Both DHTs and relaxed consistency semantics mitigate the conflict between scalability, availability and consistency. GRACE gives us the opportunity to investigate the interactions between DHTs and the various approaches to a wide range of consistency semantics. The potential benefit is a more relaxed relationship between the three factors - scalability, availability and consistency.

GRACE comes with a layered architecture that logically structures replicas into levels and spheres of consistency. All replicas in a level of consistency converge in the same way toward strict consistency, while all replicas in the same sphere of consistency abide by the same consistency requirements. There are four levels of consistency: strict level, eventual level, level of local spheres and level of no consistency. At the strict level, strict consistency will be maintained between all replicas. This level is not scalable, but most applications will be able to work under less restrictive requirements than the ones imposed at this level. The eventual level uses optimistic replication, allowing tentative updates. Therefore, it contains replicas that will eventually converge to strict consistency. The spheres contained in this level may vary in terms of the constraints that they place on tentative updates before they converge. At the level of local spheres, updates are final, and only the local consistency requirements must be enforced. The lowest level imposes no consistency requirements, using only application-specific preconditions.

This layered structure works well in conjunction with multidimensional DHTs, such as CAN, whose properties can be used by the GRACE replication layer. We found that the multidimensionality generalization over pre-existing structured peer-to-peer systems can be very nicely explored at a replication layer built on top of DHTs, by encoding consistency as additional dimensions. In GRACE, we use two additional dimensions for encoding consistency, which in our case is reflected in levels and spheres. Therefore, a replica name in GRACE will be comprised of three parts: document ID, level ID and sphere ID. We are investigating the impact of various routing approaches - one approach is to route completely along one dimension before switching to other dimensions; another approach would be to intertwine progress made along each of the three dimensions. The two approaches render different performance and fault-tolerance results. Dimensions can also be very useful in update propagation - propagating updates in a sphere, for example, is done by performing a lookup with the replica name along the sphere dimension, then the update will be sent to all corresponding replicas.

Although CAN introduced the idea of multidimensional DHTs, a more general approach is to consider orthogonal DHTs, using any two-dimensional DHT for each of the name and consistency spaces. We can choose, for example, the DHTs that offer properties needed most in each space, like convergence (Tapestry), symmetry (Pastry). The prototype for GRACE is being built on top of Pastry.