Spanner: Google’s Globally-Distributed Database

ABSTRACT
Scanner is Google’s scalable, multi-version, globally-distributed, and synchronously-replicated database. It is the first system to distribute data at global scale and support externally-consistent distributed transactions. This paper describes how Spanner is structured, its feature set, the rationale underlying various design decisions, and a novel time API that exposes clock uncertainty. This API and its implementation are critical to supporting external consistency and a variety of powerful features non-blocking reads in the past, lock-free read-only transactions, and atomic schema changes, across all of Spanner.

Technical Details/Contributions
Spanner developed and makes heavy use of TrueTime. TrueTime is an API which exposes clock uncertainty and the guarantees on Spanner’s timestamps depend on the bounds that the implementation provides. If the uncertainty is large, Spanner slows down to wait out that uncertainty.

Spanner rejects the idea that distributed transactions are too difficult or slow to be implemented. Instead they provide distributed transactions and leave it to application programmers to deal with performance problems which may arise due to the overuse of transactions. They believe this is a better approach than forcing application developers to always code around the lack of transactions.

Spanner replicates across datacenters for global availability and geographic locality. Clients are able to automatically failover between these datacenters. Spanner is able to move data within and across datacenters to balance load in the system. Spanner is able to scale to millions of machines and hundreds of datacenters.

Spanner makes heavy use of the Paxos protocol.

Compare/Contrast with Previous Papers
Like Dynamo, Spanner provides data replication across datacenters.

Like Walter, Spanner provides snapshot isolation. In Walter this snapshot is restricted to a single datacenter, but Spanner is able to snapshot across datacenters.

Spanner utilizes Colossus, which is the successor to GFS.
Criticisms
Nearly all of the contributions Spanner has made versus BigTable are based on the guarantees the TrueTime API provides (which allows globally meaningful timestamps), but the details of how the TrueTime API really works is left to a future paper.

I didn’t feel confident at all that I would be able to implement this system based on this paper.

Relevance to GDC
Spanner is able to automatically reshard data across machines and data centers as the number of machines and data centers changes. Their approach to this problem may be relevant to GDC as the GDCs become available/unavailable.