Consistency in Scalable Systems

M.I. Ruiz-Fuertes, M.R. Pallardó-Lozoya, and F.D. Muñoz-Escóí

Instituto Universitario Mixto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
{miruifue,rpallardo,fmunyoz}@iti.upv.es

Abstract. While eventual consistency is the general consistency guarantee ensured in cloud environments, stronger guarantees are in fact achievable. We show how scalable and highly available systems can provide processor, causal, sequential and session consistency during normal functioning. Failures and network partitions negatively affect consistency and generate divergence. After the failure or the partition, reconciliation techniques allow the system to restore consistency.

Keywords: DSM Consistency, Scalability, Cloud Systems, Data Centers, Distributed Systems.

1 Introduction

Scalable systems, such as cloud systems, are composed of multiple data centers, each one of them composed of a set of nodes that are located in the same facility and are locally connected through a high-speed network. Different data centers are geographically distant and connected by some inter-data-center channels, with limited bandwidth and longer transmission delays (when compared to intra-data-center networks). In these systems, each data item should have multiple replicas and these replicas should be spread over multiple data centers in order to increase both availability and fault tolerance. However, such geographical dissemination of replicas entails greater difficulty for maintaining the consistency among them. The CAP theorem [9, 18] states the impossibility of a distributed system to simultaneously provide consistency, availability and partition tolerance. Since availability is the key for scalable cloud systems, a trade-off appears between consistency and network-partition tolerance. As network partitions may be common among remote data centers, scalable systems generally sacrifice consistency, trading it for partition tolerance. In this case, and as a minimum, eventual consistency [31] can be guaranteed.

The aim of this paper is to study the different levels of consistency [27] that can be achieved in these systems, even when network partitions occur and always guaranteeing availability. For this, we take as basis previous results in the interconnection of distributed shared memory consistency models [12, 17] as well as in the interconnection of message passing systems [2, 5, 23].
In distributed shared memory systems (DSM systems), interconnectable memory consistency models are able to export intra-group consistency to other connected groups or subsystems, without changing them. To this end, they only require FIFO channels between subsystems in the regular case and so, they are trivially admitted in a network-partitionable system. In case of a partitioned network, the updates to be transmitted are buffered and re-sent when the channel is repaired. Exporting this idea into scalable systems allows different data centers to operate independently during partitions and to update or reconcile replicas as necessary once the network is repaired. The cache \cite{19}, FIFO/pRAM \cite{26} and causal \cite{22} consistency models are interconnectable \cite{12, 17}. The sequential model \cite{25}, however, cannot be interconnected in a non-intrusive manner \cite{12}. Similar results are achieved when interconnecting message passing systems (groups of processes that exchange messages through a Group Communication System, or GCS) to provide end-to-end delivery semantics: FIFO and causal semantics can be provided among different subsystems \cite{2, 3, 23}, but total/atomic semantics require intrusive modifications that penalize performance of individual subsystems \cite{23}.

Upon these previous results, this paper studies different possible configurations for scalable database replication systems, analyzing the consistency level provided in each one and proposing recovery and reconciliation techniques that allow the system to restore consistency after failures or network partitions. To this end, we subdivide the persistent data into multiple disjoint sets (data partitioning) and use passive replication. We show that it is possible to provide stronger levels of consistency than the usual eventual consistency.

The rest of this paper is structured as follows. Section 2 summarizes previous work on which this paper is based. Section 3 proposes different scalable configurations and analyzes their consistency guarantees even in case of failures or network partitions. Section 4 discusses some reconciliation techniques that are required to resolve divergence between data centers. Finally, Section 5 concludes the paper.

2 Related Work

Different previous works propose mechanisms for the interconnection of multiple local groups in order to form a global system with certain properties. Fernandez et al. \cite{17} interconnect two causal distributed shared memory systems with a bidirectional reliable FIFO channel connecting one process from each system, in such a way that the resulting DSM system is also causal. As authors highlight, the same mechanism can be used to construct a global causal system by the interconnection of sequential or atomic DSM systems, as these consistency models also respect causality.

Cholvi et al. \cite{12} define a DSM memory model as fast if memory operations in such model require only local computations before returning control, even in systems with several nodes. Otherwise, the model is said to be non-fast. Systems implementing fast models can be interconnected without any modification