# Analytic Investigation of Optimistic Replication Protocols

**Abhishek Gupta** (04010102) g.abhi@iitg.ernet.in,
**Subramanian Ashish** (04010148) s.ashish@iitg.enet.in
Under the guidance of **Prof. N.Selvaraju**
21-11-2007

**Abstract**

Optimistic replication (OR) [SS03, SS05] is a key technology in distributed systems that enable higher availability, performance and fault-tolerance in a large-scale network. The importance of such techniques is increasing as collaborative work through wide-area networks and mobile computing is becoming more popular. Despite this fact, behavior of concurrent conflicting updates caused by the relaxed consistency model is poorly understood. The aim of this paper is to describe the current state of the art in understanding *conflict rate* as an important QOS metric. The report further argues about the relative merits and demerits of *conflict rate* w.r.t other metrics for evaluating the performance of Optimistic Replication based protocols.

## Introduction

Data replication consists of maintaining multiple copies of data, called *replicas*, on separate computers. It is an important enabling technology for distributed services. Replication improves availability by allowing access to the data even when some of the replicas are unavailable. It also improves performance through reduced latency, by letting users access nearby replicas and avoiding remote network access, and through increased throughput, by letting multiple computers serve the data simultaneously.

Optimistic replication is a group of techniques for sharing data efficiently in wide-area or mobile environments. The key feature that separates optimistic replication algorithms from their pessimistic counterparts is their approach to concurrency control. Pessimistic algorithms synchronously coordinate replicas during accesses and block other users during an update. Optimistic algorithms let data be accessed without a priori synchronization, based on the "optimistic" assumption that problems will occur only rarely, if at all. Updates are propagated in the background, and occasional conflicts are fixed after they happen. Its use has expanded as the Internet and mobile computing technologies have become more widespread.

Any distributed system faces a **trade-off between availability and consistency.** *Optimistic replication faces the challenges of diverging replicas and conflicts between concurrent operations*. It is thus applicable only for applications that can tolerate occasional conflicts and inconsistent data. Fortunately, in many real-world systems, especially file systems, conflicts are known to be rather rare, thanks to the data partitioning and access arbitration that naturally happen between users.

In optimistic systems, data synchronization, or ***reconciliation***, guarantees convergence and the correctness of data in the case of improper concurrent modifications, or ***conflicts***. Typically, reconciliation involves two replicas, with bidirectional data propagation. Conflicts are detected only at reconciliation time, when both replicas have been updated since the previous reconciliation. The

***conflict rate*** is a very important metric when evaluating an optimistic replication system, since it reflects the level of data consistency and the efforts involved in achieving such consistency. To our knowledge, there has been only one successful analytic study involving investigation of more than 2 replicas. This investigation is also validated by the simulations built with actual optimistic mechanisms. In this paper we have tried to present the novel approach in a lucid manner.

The challenges involved in characterizing conflict rates are as enlisted

- Defining the conflict rate in the analytical model.
- Leveraging symmetries, permutations, and reachability to reduce the state space. [Permuted States]
- Defining appropriate representations for automation and optimization of the state reduction process.

The goal of the analytic investigation is to

- Find a compact system-state representation that eliminates unimportant variation
- transition rules that can be used to automate analytical modeling at high replication factors

## Background

Since Optimistic replication based protocols trades consistency with availability, hence they require a reconciliation phase to resolve inconsistencies and conflicts. Typically, reconciliation takes place between two replicas. Conflicts occur when different replicas of the same file are updated after the most recent reconciliation.

The paper [WKR05] states three common definitions of conflicts although depending on the way a protocol is defined it is possible to define conflicts in such a way so that it can exploit more out of the protocol i.e. it is possible to take application specific semantics into account and redefine conflicts to minimize the detection of false conflicts. But using the definitions as mentioned in the paper [WKR05] helps to simplify the modeling of various Optimistic Replication based protocols and give analytic results that bear close resemblance to simulation results.

The three definitions as described in the paper as follows

- Conflict may be defined as two simultaneous updates to the same replica.
- This definition is oriented towards log-based approach. At reconciliation time, both replicas replay logs of all updates since the last reconciliation between the same replica pair. Whenever two updates to different replicas of the same file are seen in the logs, a conflict is indicated.
- This definition is related to the scanning approach in which a reconciliation-time scan detects updates and resolves conflicts. The difference from the second definition is that multiple updates are collapsed into one and will thus result in the report of only a single conflict. (In practice, most log-based systems optimize out multiple updates to save storage, which also causes conflicts to collapse. Thus, most real systems use the third definition.)

For the remainder of this paper, third definition as defined above is used. Although the paper claims that without the loss of generality, it is safe to assume bidirectional propagation of data at reconciliation time and deterministic resolution of conflicts. However, there exists Optimistic replication protocols employing Non-deterministic resolution of conflicts for e.g. a separate

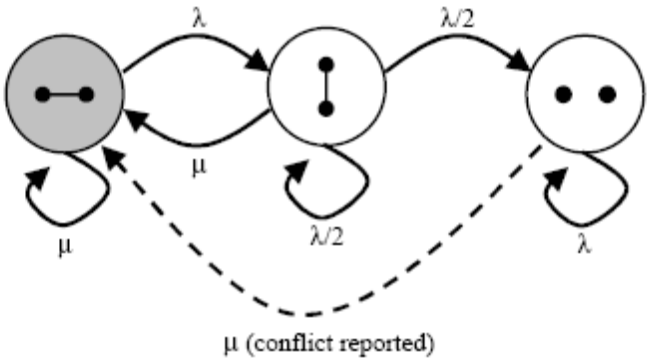
decentralized and asynchronous commitment protocol involving weighted votes [SBS06]. We must also consider the fact that this Non-deterministic protocol was proposed after the paper that is discussed in this report. From this point onwards, we will be focusing on elaborating about how conflict rates can be analytically modeled in an optimistic replication based protocol.

## Permuted state modeling of conflict rates for Optimistic Replication

### Permuted States

Consider a two replica system (fig. below). To model the system, an event-based model is used in which time is measured in terms of "interesting" system events (updates and reconciliations).

Let λ be the probability of having an update at either replica, and μ as the probability having a pair-wise reconciliation process as the next system event; a Poisson inter-arrival model is used. At each state, the outbound update probabilities sum to λ, and the outbound reconciliation probabilities sum to μ. Also, the sum of outbound λ and μ at each state is 1.




**Assumption:** To make the analysis tractable, uniform update and reconciliation probabilities are assumed.

**Representation:**

- Each replica is represented by a **dot**.
- Two identical replicas are represented via a horizontal line connecting them, e.g. the starting state (shaded). If reconciliation occurs, the replicas remain identical, so the starting state transitions back to itself.
- If one of the replicas is updated, we move to the middle state, where the update-receiving replica ***dominates*** the ***subordinate*** one. This relationship is represented by a non-horizontal line, where the upper replica dominates the lower one.

**Note:** regardless of which is updated, we are guaranteed to transition from the starting state to the middle state. By decoupling the state of the system from the labelling of individual replicas, each state effectively captures all *isomorphic* system states resulted from permuting the replica identifications. We refer to this type of state representation as ***permuted states***.

**Reconciliation**: In the case of reconciliation between a dominating replica and its subordinate, the content of the dominating replica will replace that of the subordinate, and then both replicas will be marked as identical (transition back to the starting state). An update to the dominating replica will not change its dominance over the subordinate replica.

However, an update to the subordinate replica breaks its subordinate relationship to its dominating replica, and the system enters the rightmost state (conflict). Conflicting replicas (dots) are not connected by lines. Update to either of the conflicting replicas will leave both in conflict. A reconciliation between two conflicting replicas will lead to identical replicas (the starting state or the convergence state), with a reported conflict.

**Note**: a system can be in a state with conflicting replicas without reporting conflicts, since conflicts are detected only at reconciliation time. Therefore, the ***conflict rate*** or the probability of having conflicts due to a system event (either update or reconciliation), is computed by obtaining the equilibrium probability of a state that contains replicas in conflict, multiplied by the probability of traversing its conflict-resolving transition.

**Analysis for the two replica case**

With the state diagram as in the fig. above, let variable p0, p1 and p2 be the respective probabilities of the states from the left to right. When the system is in equilibrium, the outbound transition flow at each state should be equal to the inbound flow, resulting in a system of linear equations. Also, the sum of probability at each state should be 1, i.e.

$$\lambda p_0 = \mu p_1 + \mu p_2 \qquad (1)$$

$$\left(\lambda/2 + \mu\right) p_1 = \lambda p_0 \quad (2)$$

$$\mu p_2 = \lambda/2\, p_1 \qquad (3)$$

$$p_0 + p_1 + p_2 = 1 \quad (4)$$

Solving for the probability of the conflicting state, i.e.p2, the probability of reporting a conflict *pconflict* can be computed by multiplying p2 by μ, the probability of taking the transition that resolves conflicting replicas.

$$p_0 = \frac{\lambda\mu + 2\mu^2}{(\lambda + 2\mu)(\lambda + \mu)}$$

$$p_1 = \frac{2\lambda\mu}{(\lambda + 2\mu)(\lambda + \mu)}$$

$$p_2 = \frac{\lambda^2}{(\lambda + 2\mu)(\lambda + \mu)}$$

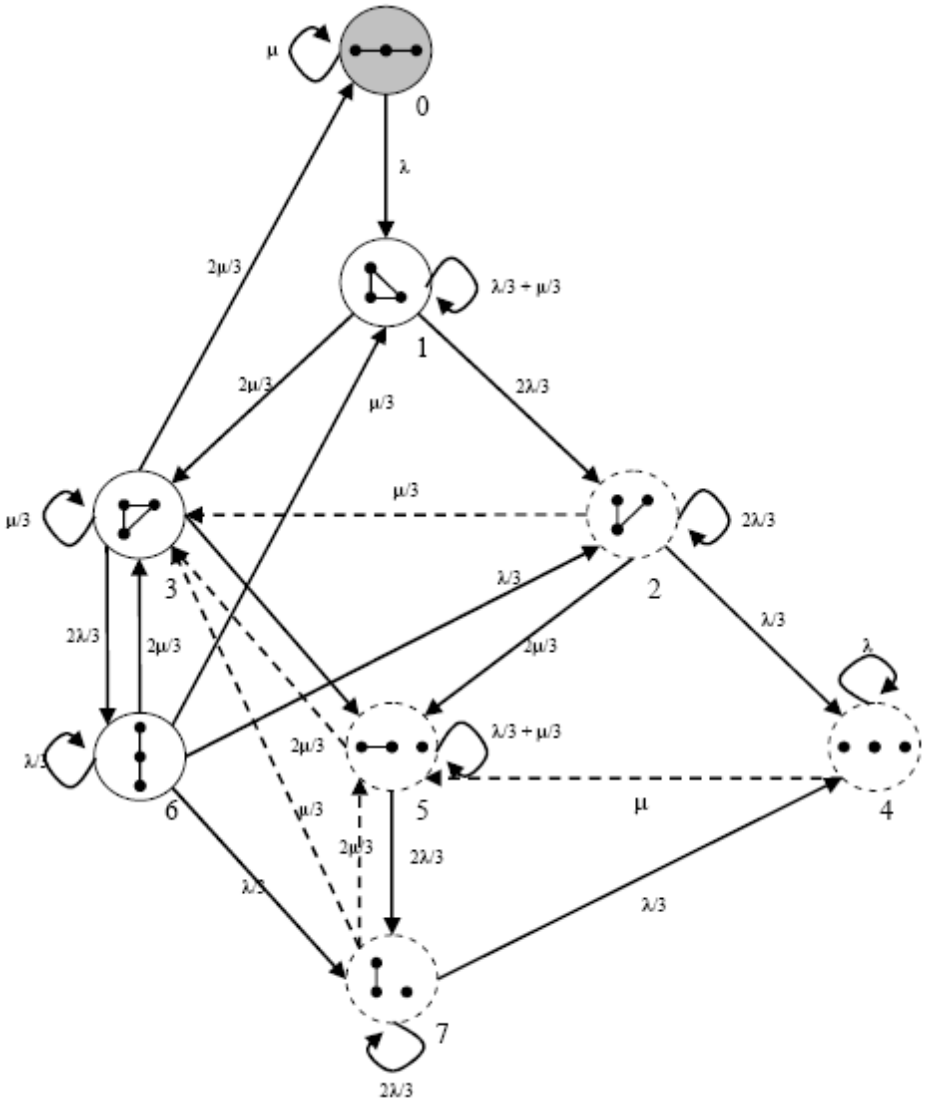
$$p_{conflict} = \mu p_2 = \frac{\lambda^2 \mu}{(\lambda + 2\mu)(\lambda + \mu)}$$

**Generalization to Three Replicas**

Figure below shows the transition diagram for three replicas. Surprisingly, the three-replica case can be completely captured with only 8 permuted states. The starting state (state 0), which is also the convergence state, is easily generalized from the two-replica case. However, the other states are more complicated compared to the two replica case. The transitional probability rates between states can be obtained by:

- Considering an update to each replica in a given state to determining the state to which a transition occurs. Since a given replica is updated at a rate $\lambda/3$, add this value to the outward transition rate from the initial state to the transition state; for e.g. consider state 1. To determine the outward transition rates from state 1, consider what happens when the dominating replica in state 1 gets updated (with rate $\lambda/3$). Clearly, the transition state is 1 again, so we add $\lambda/3$ to a self-edge in state 1. Now, if either of the two subordinate replicas are updated (each with rate $\lambda/3$), a transition occurs to state 2. Hence, considering only updating of replicas in state 1, we add $\lambda/3$ to a self-loop edge and $(2\lambda)/3$ to an edge from state 1 to state 2.
- Similarly, consider all possible reconciliation pairs from a given state (each occurring at a rate $\mu/3$) and note the transition state. Add this rate to the transition edge. For e.g. consider state 1 again: a reconciliation between the dominating state and either one of the subordinate states results in a transition toward state 3 at a rate $(2\mu)/3$, while a reconciliation among the subordinate replicas makes a transition back to state 1.

In this fashion, we can get the overall transition state diagram and solve for the equilibrium state probabilities. Overall, the three-replica case demonstrates the richness of behaviours in optimistic replication. Interestingly, not all conceivable states are possible. For example, it is not possible to have one replica dominating over two replicas in conflict. Using permuted states for analysis eliminates both isomorphic states and unreachable states.




**State-Equilibrium Equations for Three Replicas**

Using the steady state condition, we obtain the following equations:

$$\lambda p_0 = \frac{2\mu}{3} p_3 \qquad (1)$$

$$\frac{2(\lambda+\mu)}{3} p_1 = \lambda p_0 + \frac{\mu}{3} p_6 \quad (2)$$

$$\left(\frac{\lambda}{3} + \mu\right) p_2 = \frac{2\lambda}{3} p_1 + \frac{\lambda}{3} p_6 \qquad (3)$$

$$\left(\lambda + \frac{2\mu}{3}\right) p_3 = \frac{2\mu}{3} p_1 + \frac{\mu}{3} p_2 + \frac{2\mu}{3} p_5 + \frac{2\mu}{3} p_6 + \frac{\mu}{3} p_7 \qquad (4)$$

$$\mu p_4 = \frac{\lambda}{3} p_2 + \frac{\lambda}{3} p_7 \qquad (5)$$

$$\left(\frac{2(\lambda+\mu)}{3}\right) p_5 = \frac{2\mu}{3} p_2 + \frac{\lambda}{3} p_3 + \mu p_4 + \frac{2\mu}{3} p_7 \qquad (6)$$

$$\left(\frac{2\lambda}{3}+\mu\right)p_6 = \frac{2\lambda}{3}p_3 \quad (7)$$

$$\left(\frac{\lambda}{3}+\mu\right)p_7 = \frac{2\lambda}{3}p_5 + \frac{\lambda}{3}p_6 \quad (8)$$

$$p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 = 1 \quad (9)$$

The conflict rate is the sum of the products of the probability of each conflict originating state and its outbound transition probability of conflict resolving edges, i.e.

$$p_{conflict} = \frac{\mu}{3}p_2 + \mu p_4 + \frac{2\mu}{3}p_5 + \frac{2\mu}{3}p_7$$

On solving for this probability, we get:

$$p_{conflict} = \frac{2\lambda^2\mu(3\lambda^2 + 11\lambda\mu + 9\mu^2)}{(2\lambda+3\mu)(3\lambda+2\mu)(\lambda+2\mu)(\lambda+\mu)}$$

As exponents for the state probabilities are four or five, it can be inferred that the permuted state representations may be further compacted. The simulation validation confirms the validity of modelling optimistic replication via permuted states.

## Problems with existing QOS metrics

Since a well-designed replicated system achieves nearly 100% read availability, the metric of read-only access does not significantly contribute to characterize the relative quality of services offered by other such replicated systems. Instead, the most interesting metrics are those that relate to updates.

The use of conflict rate as a QOS[1] metric for an optimistically replicates system has quite a few drawbacks. On one hand, we can say that the *conflict rate* metric is simple to express, to measure and to understand. On the other hand, since system detects conflicts only during the reconciliation process, the frequency of reconciliation has a direct effect on the observed *conflict rate*. In the limiting case, if reconciliation never occurs, no conflict will be detected regardless of the degree of conflicting updates! (The effect is further exacerbated by batch propagation of updates, done in order to save time and space during reconciliation). In addition, a system with high conflict rate is troublesome to the user. Secondly, the number of conflicts also depends upon the reconciliation pattern. Thirdly, if a conflict is created at a pair of replicas but it is not resolved even after several reconciliations then it is difficult to design an algorithm that can correctly count it as only one conflict especially when the system is faced with complex update propagation patterns. Fourthly, since conflict rate only measures write/write effects, it ignores whether user is accessing out-of-date information i.e. staleness. Furthermore, it is possible to reduce conflict rates by either shortening or lengthening reconciliation intervals. Therefore, conflict rates may not provide a usable Quality of service [KBG98] number for the purpose of direct comparison among optimistic replication schemes. Ironically since conflict rate, in part, is affected by the freedom of data flow within a given reconciliation topology i.e. on the pattern of both updates and reconciliation. Therefore, conflict rate will anyways be indirectly associated with other quality of service metrics.

[1] QOS of optimistically replicated system in the face of updates is the degree to which the system presents an illusion of connectivity to a single up-to-date copy of all objects to all users.

Ideally, user wants to access an up-to-date data at all instants of time. However, in practice user mostly accesses stale data[2]. Stale access count provides a very attractive measure of quality of service because they reflect what is important to the user i.e. up-to-date data. In addition, unlike conflict rate, it does not tend to towards zero as the reconciliation interval increases. An alternative is to measure *Age of a stale access*[3]*;* it reflects the amount of time by which the information is out of date. Unfortunately, the global nature of stale-access metrics causes them to be unusable in live systems due to problems with obtaining global time.

As a practical alternative to stale-access metric*, 100% propagation time*[4] is much easier to measure in a real or simulated system. It depends on many factors in the design of replication system, including the choice of reconciliation method (e.g. broadcast, multicast, epidemic), the speed and frequency of reconciliation, the topology across which updates are propagated. The effectiveness of the metric lies in the assumption that a replication system with a low propagation time will provide the user with better service than one with a high time. In order to avoid suffering from liveness issues, it might be useful to measure 90% or say 60% propagation time since otherwise an unusually slow or disconnected site(s) would wrongly reflect the quality of service delivered to users that are more frequent. Furthermore, it might be helpful to weight the propagation time by the object's usage level since it may not matter whether a rarely used object is propagated quickly. The biggest disadvantage of propagation time is the cost of measurement. Furthermore, it is not directly related to the quality of service perceived by the user.

The commonly used metric, conflict count, is subject to a number of anomalies that make it easy to misuse and inaccurate in the general case. Nevertheless, it is found that the conflict count is useful both because of its simplicity and because its weaknesses are easily minimized in most real-world situations.

## Conclusion

The analysis of problems with exponential state spaces is always challenging. Permuted-states are a new technique that makes the analysis of complex replicated systems tractable. This has been confirmed by an independent simulation carried out by version vectors as discussed in WKR05. As discussed above, the *conflict rate* metric is not a suitable Quality of Service metric to evaluate the performance of Optimistically Replicated System. And till date we have not found any other study elaborating more regarding using any other metric for OR based protocol evaluation or any other advances in the direction of analytic investigation of OR based protocols.

---

[2] Data object that is globally out of date, an access to a replica is stale if some other replica has been updated prior to that access as measured by a global clock and that update has not yet propagated to the replica being accessed.

[3] Defined as the time elapsed between the latest update to an object as seen by the accessing replica and the time of the globally last update

[4] Time needed for an update to an object to become visible at all sites

## References


I. [SS05] Y. Saito and M. Shapiro 2005. Optimistic replication. *ACM Comput. Surv.,* 37(1):42–81 '05.

II. [SS03] M. Shapiro and Y. Saito 2003. Scaling Optimistic Replication, chapter 30. *LNCS: Lecture Notes in Computer Science*, '03.

III. [WRB02] A.I. Wang, P. Reiher, R. Bagrodia and G.H. Kuenning 2002. Understanding the Behavior of the Conflict-Rate Metric in Optimistic Peer Replication. *In DEXA, proceedings of the 13th International Workshop on Database and Expert Systems Applications.*

IV. [WRK05] A.I. Wang, P. Reiher, and G.H. Kuenning 2005. Introducing permuted states for analyzing conflict rates in optimistic replication. *In Joint International Conference on Measurement and Modeling of Computer Systems, proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer Systems.*

V. [WKR05] A.-I. Wang, G. Kuenning, and P. Reiher 2005. Using Permuted States and Validated Simulation to Analyze Conflict Rates in Optimistic Replication. *In international Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'05), Philadelphia, PA, July, 2005.*

VI. [KBG98] G.H. Kuenning, R. Bagrodia, R.G. Guy, G.J. Popek, P.L. Reiher and A.I. Wang 1998. Measuring the Quality of Service of Optimistic Replication. *In LNCS: Lecture Notes In Computer Science; Vol. 1543 , Workshop ion on Object-Oriented Technology.*

VII. [SBS06] P. Sutra, J. Barreto and M. Shapiro 2006. An asynchronous, decentralised commitment protocol for semantic optimistic replication. *Rapport de Recherche RR-6069, INRIA Rocquencourt, December 2006.*