Coping with the Insider Threat in Scalable Distributed Information Systems

Yair Amir
Department of Computer Science
Johns Hopkins University

Cristina Nita-Rotaru
Department of Computer Sciences
Purdue University

Introduction

Geographically distributed information systems achieve high availability that is crucial to their usefulness by replicating their state. Providing instant access at time of need regardless of current network connectivity requires the state to be replicated in every geographical site so that it is locally available. As network environments become increasingly hostile, we have to assume that part of the distributed information system will be compromised at some point. The problem of maintaining a replicated state in such a system is magnified when insider (or Byzantine) attacks are taken into account.

There is considerable body of research addressing Byzantine behavior of server replicas that aims to solve the insider attack problem. Modern, state of the art Byzantine replication solutions such as [CL99, Cas01, YMVAD03] share the following properties[1]:

- *Require strong connectivity*: Existing solutions, that are proven optimal in this regard, require $f+1$ identical responses from different replicas in order to provide a response to the client for a single query, and may need to contact up to $2f+1$ (out of $3f+1$) replicas in order to allow progress. For distributed information systems, such requirements are not always possible to meet over an unreliable wide area network, which will leave the system consistent but unable to make progress. For example, network partitions can leave all of the clients blocked.
- *Limited scalability due to multiple all-peer exchanges*: These solutions usually guarantee liveness and correctness as long as no more than any f replicas are compromised out of $3f+1$ replicas. To achieve this goal, these protocols require three-rounds all-peer exchange between the participants. The approach does not scale well in high latency wide area networks. A successful solution will have to be hierarchical in order to scale.

Byzantine fault-tolerance can also be obtained by applying quorum replication methods. Examples of such systems include Phalanx [MR98] and its successor Fleet [MR00, MRTZ01]. Although this approach can be relatively scalable with the number of replica servers, it also suffers from the drawbacks of flat non-hierarchical Byzantine replication solutions.

We present initial ideas for a hierarchical architecture that allows distributed information systems to address the insider threat to the infrastructure, while scaling to wide area network settings.

Hierarchical Architecture

Replication is an important tool towards maintaining functionality in adverse situations. Byzantine replication systems focus on protecting against Byzantine server replicas. The current symmetric Byzantine solutions assume that network connectivity is a complete graph. If there is no connection to one of the participating replicas, that participant is assumed to be faulty. In such symmetric, non-hierarchical solutions, a replica must be connected to $2f+1$ out of $3f+1$ replicas, a

[1] We denote by f the number of faulty replicas that should be masked by the system.

strong majority, in order to allow progress both for updates and queries. In contrast, in a non-Byzantine model, updates are possible with a quorum [Ske82], and a query is possible with one replica.

In wide area networks that connect over unreliable (possibly wireless) networks, this connectivity assumption is not reasonable as partitions occur under normal circumstances. External network attacks further invalidate this assumption. If the client depends on being connected to replicas over the wide area network, it might not be possible to answer queries at all and information access will be blocked. If answers are accepted with less than the required number of replies, the guarantees of the Byzantine replication algorithm are broken.

The above analysis leads us to conclude that in practice, insider attacks on server replicas cannot be solved via a flat solution on wide area networks. Our key observation is the notion that a practical solution should be hierarchical. It has to confine the treatment of insider attacks on server replicas within each geographical site to the site itself, so that the replicas in each site construct one logical entity that is robust to insider server attacks. This building block enables us to use optimized fault tolerant replication on the wide area network, where we only need to protect against external attacks and can use more standard approaches to fault tolerant replication.

Our initial solution constructs one trusted logical server in each site using symmetric Byzantine replication [Cas01] in the local area network in order to agree on the order of local site updates that will be propagated to other sites. In addition, using threshold cryptography [DF89, Des97] we allow a site representative to send messages on the wide area network only if they correspond to the agreement achieved by the local area Byzantine replication. As other sites will only accept updates that contain f+1 local server replica shares (out of 3f+1 local replicas), this guarantees that at least one correct local replica attests that the local order is correct. Threshold cryptography enables us to be more efficient since other sites only need to verify one signature on the wide area update message, and not f+1 different signatures. It also helps in terms of management, as other sites need to maintain only one public key per remote site (in addition to any clients' public keys necessary to authenticate the source of the update if necessary).

The benefit of this approach is considerable in that it presents a viable solution where client queries (that do not modify the distributed information system's state) are locally ordered by the Byzantine replication component and answers from local replicas are immediately returned to the client without any communication with remote sites.

Updates are first ordered locally by the Byzantine replication, and are propagated to other sites using a fault tolerant replication algorithm executed over secure group communication to obtain a global order.

More specifically, local clients send requests and get responses directly from the local server replicas. Each client request is authenticated and can be verified globally. The architecture of each replica includes the Byzantine replication providing service to the local area network clients. It also includes the fault tolerant replication over a secure group communication service such as Secure Spread [AKNS+04,ANST03], guaranteeing that one representative replica is active on the wide area network, while the others are in standby, ready to take over. Each standby replica monitors the wide area communication, verifying that the representative is acting correctly.

The resulting system has the following benefits:

- Queries (that do not modify the distributed information system's state) are locally ordered by the Byzantine replication component and answers from local replicas are immediately returned to the client without any communication with remote sites. Therefore, we expect a much better latency compared to a flat Byzantine replication approach. For example, for a network with 50 milliseconds diameter, this should yield orders of magnitude lower

latency, as it will take at least 300 milliseconds to complete three wide area round trips that are required by current symmetric state of the art Byzantine replication methods.

- A baseline non-hierarchical system that uses $3f+1$ replicas on the wide area network protects against f Byzantine replicas system wide. In contrast, our hierarchical architecture protects against f Byzantine faults in each site for the price of having $3f+1$ replicas in every site.

An Alternate Hierarchical Approach

An alternate hierarchical approach to scale Byzantine replication to wide area networks can be based on having a few trusted nodes that are assumed to be working under a weaker adversary model. For example, these trusted nodes may exhibit crashes and recoveries but not penetrations. A Byzantine replication algorithm in such an environment can use this knowledge in order to optimize the performance and bring it closer to the performance of a fault tolerant, non-Byzantine solution.

Such a hybrid approach was proposed in [CLNV02, Ver03], where trusted nodes were also assumed to perform synchronously, providing strong global timing guarantees. The hybrid failure model of [CLNV02] inspired the Survivable Spread [SurS03] work, where a few trusted nodes (at least one per site) are assumed impenetrable, but are not synchronous, may crash and recover, and may experience network partitions and merges. These trusted nodes were implemented by Boeing Secure Network Server (SNS) boxes, which are limited computers designed specifically not to be penetrable. Using the trusted nodes, Survivable Spread optimizes the local area communication exchange, eliminates the ability of malicious participants to create two-face behavior, and drastically limits the ability of malicious participants to slow down the protocol progress. Wide area communication in Survivable Spread is conducted solely between trusted nodes.

Conclusion

We presented two hierarchical approaches that provide scalable architectures for distributed information systems resilient to internal attacks. In our opinion, both our initial solution and the hybrid approach offer a viable way to handle malicious servers on wide area networks. The two approaches provide an interesting tradeoff between performance and trust. In essence, our initial solution assumes less and therefore is more general, while the hybrid approach provides a highly optimized solution that relies on trusted components. Exploring this tradeoff of performance and trust further can lead to much needed systems that are scalable and yet able to withstand today's increasingly hostile networking environment.

References

[AKNS+04] Y. Amir, Y. Kim, C. Nita-Rotaru, J. Schultz, J. Stanton, and G. Tsudik. Secure group communication using robust group key agreement. To appear in *IEEE Transaction on Parallel and Distributed Systems.*

[ANST03] Y. Amir, C. Nita-Rotaru, J. Stanton, and G. Tsudik. Scaling secure group communication systems: Beyond peer-to-peer. In *The 3rd DARPA Information Survivability Conference and Exposition (DISCEX III)*, Washington, D.C., April 2003.

[Cas99] M. Castro and B. Liskov. Practical Byzantine fault tolerance. *In the 3rd Symposium on Operating Systems Design and Implementation (OSDI'99)*, New Orleans, USA. Feb. 1999.

[Cas01] M. Castro. Practical Byzantine fault tolerance Ph.D. Thesis. Jan. 2001. MIT, Computer Science.

[CLNV02] L. C. Lung, N. F. Neves, and P. Veríssimo. Efficient Byzantine-Resilient Reliable Multicast on a Hybrid Failure Model. *Proceedings of the 21st IEEE Symposium on Reliable Distributed Systems (SRDS)*, pages 2-11, 2002.

[Des97] Y. Desmedt. Some Recent Research Aspects of Threshold Cryptography. In E. Okamoto, G. Davida and M. Mambo, editors, Information Security, Proceedings (Lecture Notes in Computer Science 1396), pp. 158-173. Springer-Verlag, 1997.

[DF89] Y. Desmedt and Y. Frankel. Threshold cryptosystems. In: Advances in Cryptology - Crypto '89, Proceedings, Lecture Notes in Computer Science 435 (G. Brassard, Ed.), Springer-Verlag, 1990, pp. 307-315.

[MR98] D. Malkhi and M. Reiter. Byzantine quorum systems. Journal of Distributed Computing, 11(4):203-213, 1998.

[MR00] D. Malkhi and M. Reiter. An architecture for survivable coordination in large distributed systems. IEEE Transactions on Knowledge and Data Engineering, 12(2):187-202, April 2000.

[MRTZ01] D. Malkhi, M. K. Reiter, D. Tulone, and E. Ziskind. Persistent objects in the Fleet system. In Proceedings of the 2nd DARPA Information Survivability Conference and Exposition (DISCEX II), Vol. II, pages 126-136, June 2001.

[Ske82] D. Skeen. A Quorum-Based Commit Protocol. In *Berkeley Workshop on Distributed Data Management and Computer Network*, number 6, pages 69-80, February 1982.

[SurS03] Survivable Spread: Algorithms and Assurance Argument, Technical Information Report Number D950-10757-1, The Boeing Company, July 2003.

[Ver03] P. Verissimo. Uncertainty and Predictability: Can They Be Reconciled. *Future Directions in Distributed Computing.* Springer-Verlag LNCS 2584, 2003.

[YMVAD03] J. Yin, J.-P. Martin, A. Venkataramani, L. Alvisi and M. Dahlin: Separating agreement from execution for Byzantine fault-tolerant services. SOSP 2003: 253-267.