Group-Oriented Services in Asynchronous Systems

Danilo Bruschi Elena Pagani* Gian Paolo Rossi
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
e-mail: {bruschi, pagae, rossi}@dsi.unimi.it

Extended Abstract

One of the fundamental results about fault-tolerant distributed computing is the impossibility of *Consensus* in *asynchronous* systems [FLP85], which asserts that there is no deterministic algorithm that solves consensus in an asynchronous system and tolerates even a single crash failure.

Consensus is a fundamental problem in distributed computing: services such as atomic commitment [H90] and atomic broadcast [BSS91, VRB89] are built up as consensus algorithms. Furthermore, in many real systems unpredictable workloads and network congestions are sources of asynchrony, thereby rendering any synchrony assumption valid only probabilistically. Hence, the impossibility result reveals a serious limitation for the construction of fault-tolerant applications in real distributed environments. This result is due to the impossibility of distinguishing between a failed processor and a merely slow one in asynchronous environments.

The main technique used so far to overcome this difficulty has been that of eliminating from the system both the faulty processors and the processors that are "too" slow. This choice presents two drawbacks: the abrupt elimination of slow processors could overcome the tolerated resilience, thus leading to non-terminating protocol executions, or it could lead to the group splitting (*partitionable* systems [AMM95, CHT96, DMS95]). Moreover, the eliminated processors are forced to execute expensive recovery procedures before rejoining the group, usually under a different identity.

We believe that this technique is inappropriate when small networks or networks with frequent transient failures are considered. In the system we have devised a processor q suspected of being faulty is eliminated from the system only if it does not participate in the protocol executed to change the system configuration and none of the participants has observed q as being active. Otherwise, the q's suspect failure is deleted. This way, transient failures can be tolerated as well as erroneous failure suspects.

The designed architecture is composed by a *failure suspector* module ($\mathcal{FS}$) and a *membership* module ($\mathcal{M}$). The former carries out the transmission of the messages over the network and monitors the processors status through the local observation of the received acknowledgements. In the case $\mathcal{FS}$ suspects a processor failure, it signals this event to

*Supported in part by MURST 40%.

$\mathcal{M}$. $\mathcal{M}$ maintains a single agreed view of the group configuration (*primary-partition* system [CHT96, PR95, RBS92]); moreover, the stabilization of the messages that are still partially diffused is achieved.

The capability of re-integrating at run-time a suspected processor is the main characteristic of our approach with respect to those above mentioned. To our knowledge it represents the first attempt for affording such a case that is of particular interest when small groups are considered. The cost of managing "pardoned" processors is lower than that of re-integrating the excluded processors, thus achieving better performance. As a drawback, the membership protocol is more complex: besides of providing for the agreement on the new group among the operational processors, it involves mechanisms needed to cope with some new source of non-termination. This may occur in the presence of processors that repeatedly change their mind regarding the suspected failures.

References

[AMM95] Amir Y., Moser L. E., Melliar-Smith P. M., Agarwal D. A., Ciarfella P., "The Totem Single-Ring Ordering and Membership Protocol". *ACM Trans. on Computer Systems, Vol. 13, No. 4*, pages 311-342, Nov. 1995.

[BSS91] Birman K., Schiper A., Stephenson P., "Lightweight Causal and Atomic Group Multicast". *ACM Transactions on Computer Systems, Vol. 9, No. 3*, pages 272-314, Aug. 1991.

[CHT96] Chandra T. D., Hadzilacos V., Toueg S., Charron-Bost B., "On the Impossibility of Group Membership". *Proc. PODC'96*, 1996.

[DMS95] Dolev D., Malki D., Strong R., "A Framework for Partitionable Membership Service". *Technical Report CS95-4, Inst. of Computer Science, Hebrew University, Jerusalem*, 1995.

[FLP85] Fischer M., Lynch N., Paterson M. S., "Impossibility of Distributed Consensus with One Faulty Process". *Journal of the ACM, Vol. 32, No. 2*, pages 374-382, Apr. 1985.

[H90] Hadzilacos V., "On the Relationship between the Atomic Commitment and Consensus Problems". *Fault-Tolerant Distributed Computing*, page 201, 1990.

[PR95] Pagani E., Rossi G. P., 'A Set of Multicast Primitives for Fault Tolerant Distributed Systems". *Journal of High Speed Networks, Vol. 4, No. 3*, pages 299-316, IOS Press, Sep. 1995.

[RBS92] Ricciardi A., Birman K., Stephenson P., "The Cost of Order in Asynchronous Systems". *Proc. of the 6th International Workshop WDAG'92*, pages 329-345, Nov. 1992.

[VRB89] Verissimo P., Rodrigues L., Baptista M., "AMp: A Highly Parallel Atomic Multicast Protocol". *Proceedings of the ACM SIGCOMM'89 Conference. Computer Communication Review, Vol. 19, No. 4*, pages 83-93, Austin, TX, Sep. 1989.