The Totem Redundant Ring Protocol

R. R. Koch, L. E. Moser, P. M. Melliar-Smith
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106
ruppert, moser, pmms @alpha.ece.ucsb.edu

Abstract

Group communication protocols greatly simplify the design of fault-tolerant distributed systems. Most of those protocols focus on node redundancy rather than on network redundancy. The Totem Redundant Ring Protocol allows the use of multiple redundant local-area networks. The partial or total failure of a network remains transparent to the application processes. The distributed system remains operational while an administrator reacts to an alarm raised by the Totem Redundant Ring Protocol. The user can choose between active, passive and active-passive replication of the network.

1. Introduction

Group communication protocols [1, 2, 4, 12] must provide reliable delivery, ensured either by an underlying reliable protocol or by the group communication protocol itself. Properties such as virtual synchrony [4] and extended virtual synchrony [16] ease the maintenance of consistency of replicated data. Systems that are connected by a wide-area network [13, 20, 22] have a good chance of remaining operational if parts of the network fail. Local-area networks (LANs), on the other hand, often employ a single switch or a hub. If that component fails, no node can communicate with any other node and the system partitions into singletons. Systems that follow the primary component model [4] shut down all nodes, while other systems keep the nodes up, even though they cannot do useful work when the communication links are severed.

To allow a distributed system to tolerate network faults, the network itself must be replicated. Although replicated wide-area networks are not practical, LANs can be replicated cheaply. However, the mere presence of a redundant network does not overcome network faults. A special protocol must be employed to coordinate redundant networks.

Such a protocol can be used in distributed applications with high availability requirements, such as financial, avionic, or military applications, that are based on clusters of computers, instead of dedicated hardware. The range of applications that can benefit from a redundant network protocol extends from real-time radar image analysis to back-end servers for financial applications. Other applications include more general fault tolerance infrastructures, such as AQuA [8] or Eternal [18], which build on a group communication protocol.

To enable the use of redundant networks in a fault-tolerant distributed system, we have developed the Totem Redundant Ring Protocol (Totem RRP), which is based on the Totem Single Ring Protocol (Totem SRP) [2]. The Totem SRP is a highly efficient group communication protocol for Ethernet-based LANs. Totem imposes a logical token-passing ring on the network. The token is used to achieve reliable delivery of messages, causal and total message ordering, flow control and fault detection. The Totem SRP also provides group membership services.

The Totem RRP provides the same services to application processes as the Totem SRP. However, the Totem RRP utilizes multiple networks to achieve resilience against partial or total network faults. Network faults remain transparent to the application processes, and the system remains operational as long as a single network is operational.

As shown in Section 8, the Totem RRP increases both reliability and throughput. This characteristic is important for building fault-tolerant distributed systems that handle heavy message loads, such as telecommunication switches and distributed real-time image processing systems, or reliable network storage devices.

2. The Totem Single Ring Protocol

The Totem Single Ring Protocol (SRP) is a group communication protocol designed for Ethernet-based LANs. The Totem SRP uses the native Ethernet broadcast service to broadcast messages efficiently. All data is sent in the form of packets using UDP.

This research has been supported by DARPA/ONR Contract N66001-00-1-8931 and MURI/AFOSR Contract F49620-00-1-0330.