Illinois Institute Of Technology

Comparative Operating Systems - FALL 2001

Professor : Marius Soneru

Site: Internet (Section 251)

Report Topic: Load Balancing in Distributed Systems

Name: Ms. Lata S. Rao

SID: 333-96-1950

Email: raolata1@iit.edu

Introduction to and Need for load balancing in distributed systems:

Distributed systems are composed of several **loosely -coupled independent** computers communicating over a high-bandwidth network. This collection of independent computers presents an uni-processor view to the user i.e. several computers collectively **"co-operate"** to satisfy the user's request. Users of such a system submit tasks at their host computers for processing. In the case of a standalone workstation, the task submitted by the user would be immediately scheduled for processing depending upon the scheduling algorithm used. But in a distributed system where the resources are shared and the user has a uni-processor view of the resource pool, the first step in scheduling the submitted task is to decide where to schedule it and this brings the issue of load balancing into picture.

In general load distribution, a general form of load balancing can be defined as the process of allocating resources to a task from a pool of resources depending upon parameters like:

1) Process/tasks requirements (type, size, priority etc.) of the process,
2) Resource availability, that is the current state of the distributed system , and
3) Static, dynamic or adaptive rules for such process in the distributed system.

The basic fact that submission of tasks by users on different hosts is random leads to a situation where some of the hosts are highly loaded while others are not. This issue is more prominent in heterogeneous distributed systems wherein each of the hosts have varied computing power. Even in the case of a homogeneous distributed system, system performance can be improved by appropriately transferring the load from heavily loaded computers (senders) to idle or lightly loaded computers (receivers). The two basic criteria for any load balancing system are the definition of performance and load. These two criteria are responsible for the active decision making in the load balancing process. Evaluating this criteria at run time "efficiently "and deciding upon representations of this criteria (for ex. Average response time as a measure of performance etc.) is the most important issue, as elaborated in later sections.

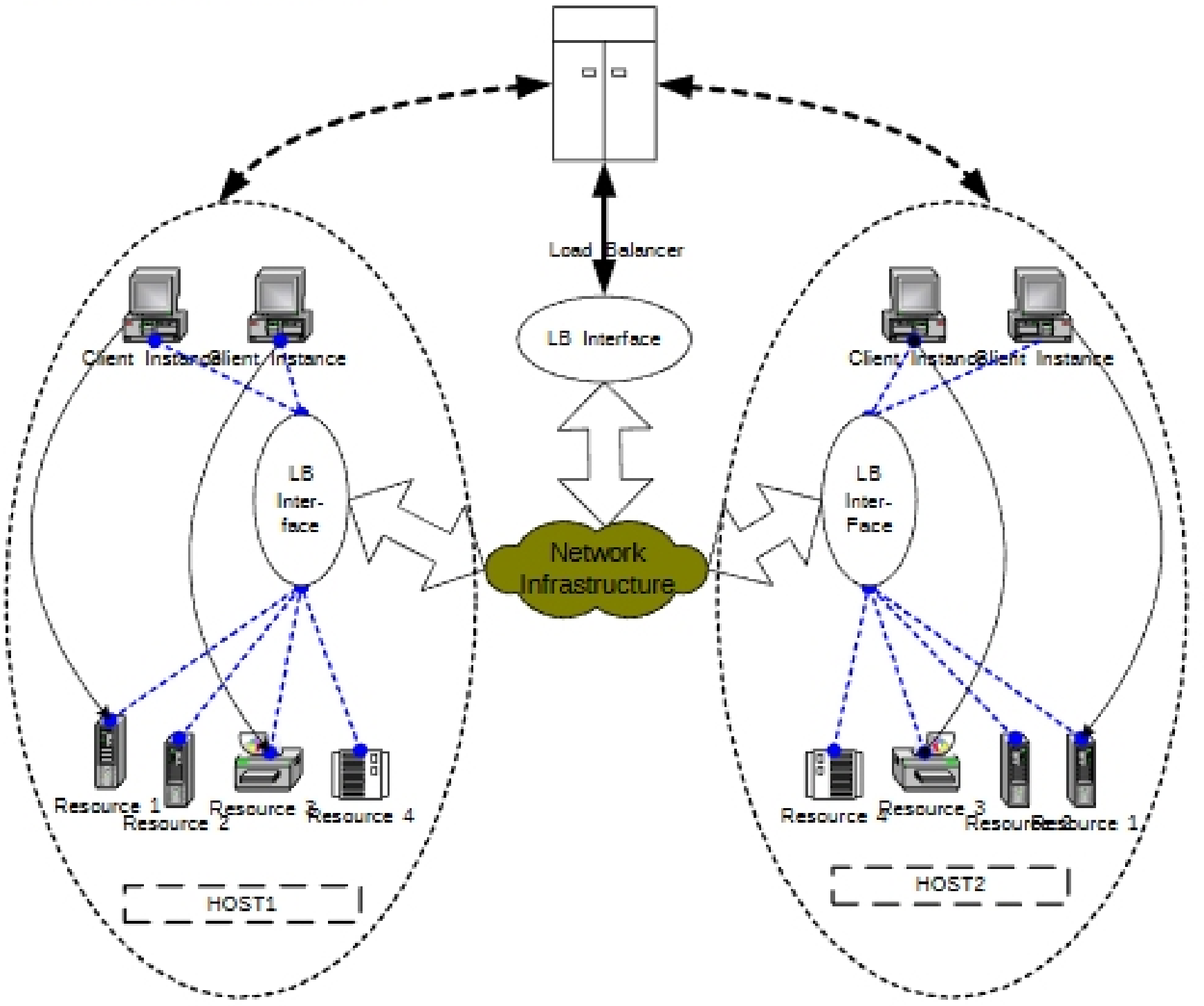
Load Balancing as a special case of load distribution:

Load balancing and load sharing can be considered as special cases of load distribution, differing on their load distributing principle. Both strategies attempt to decrease the probability of an idle state for a resource by transferring tasks to lightly loaded nodes. Further, load balancing algorithms attempt to **"equalize" loads** at all nodes. This implies that load balancing strategy generally involves higher overhead since task transfers occur at a higher rate than a load sharing strategy.

Load Balancing Architectures:

Depending upon the logical placement and implementation of load balancing module there are 2 possible architectures:

1) Centralized Load Balancer:



In this scenario the load Balancer sits at the core of the network connecting the different hosts in the distributed system. It is responsible for:

- Implementing the Information policy (elaborated later) using one of the possible algorithms.
- Implementing the transfer policy (elaborated later) using one of the possible algorithms.
- Providing the basic function of load balancing.