LOAD BALANCING IN CLUSTER COMPUTING

PRAMOD K GAUR (Associate Professor CSE MITS JADAN PALI)
SANJESH PANT (Assistant Professor, CSE MITS JADAN PALI)

REVIEW OF COMPUTING GENERATION

Computing is an evolutionary process. Five Generation of development history with each generation is improving on the previous one's technology, architecture, software application and representative system that make clear. As part of this evolution computing requirement driven by application have always outpaced the available technology hence system designer have always sought seek faster and more cost effective system

CLUSTER COMPUTING

Cluster computing is best characterized as the integration of a number off-the sheaf commodity computers and resources integrated through hardware ,software and network to behave as a single computer. initially ,the term cluster computing and high performance computing were treated as same. However the technologies available today have redefined the term cluster computer to extend beyond parallel computing to incorporate load balancing clusters like web cluster and high availability clusters clusters may also be developed to address load balancing ,parallel processing ,system management and scalability.

At the element level, when two or more computers are used together to solve a problem, it is considered as a computer cluster. There are different ways of implementing the cluster, Beowulf is the most common way, however this is also just a cooperation between computers in order to solve a problems. Table 1.1 shows comparison of various parallel computers

Table 1.1 Comparison of parallel computers

Characteristic	MPP	SMP	Cluster
System Number of node	O(10)-O(1000)	O(10)-O(100)	O(100) or less
Node complexity	Fine or Medium Gain	Medium or coarse Grain	Medium Grain
Inter-Node communication	Message Passing	Centralised and Distributed Shared Memory	Message Passing
Job Scheduling	Single Run Queue on Host	Single Run Queue Mostly	Multiple Queues but Coordinated

SSI Support	Partially	Always	Desired
Ownership	One organization	One organization	One or more organization
Address space	Multiple	Single	Multiple or Single
Inter-Node Security	Unnecessary	Unnecessary	Required if Exposed

There are two main reasons of using a cluster

i. To achieve high availability i.e, higher reliability
ii. High performance computing, to get greater computing power than a single machine can provide.

A cluster of same size and computing power as mainframes is many times cheaper than mainframe and this is also the big reason of using a cluster. Today, clusters are made up of commodity computers usually restricted to a single switch or group of interconnected switches operating with in a single virtual local area network(V-LAN). Each computer node may have different characteristic such as single processor or multiprocessor design and access to various types of storage devices. The underlying network is a dedicated network made up of single switch or a hierarchy of a multiple switches.

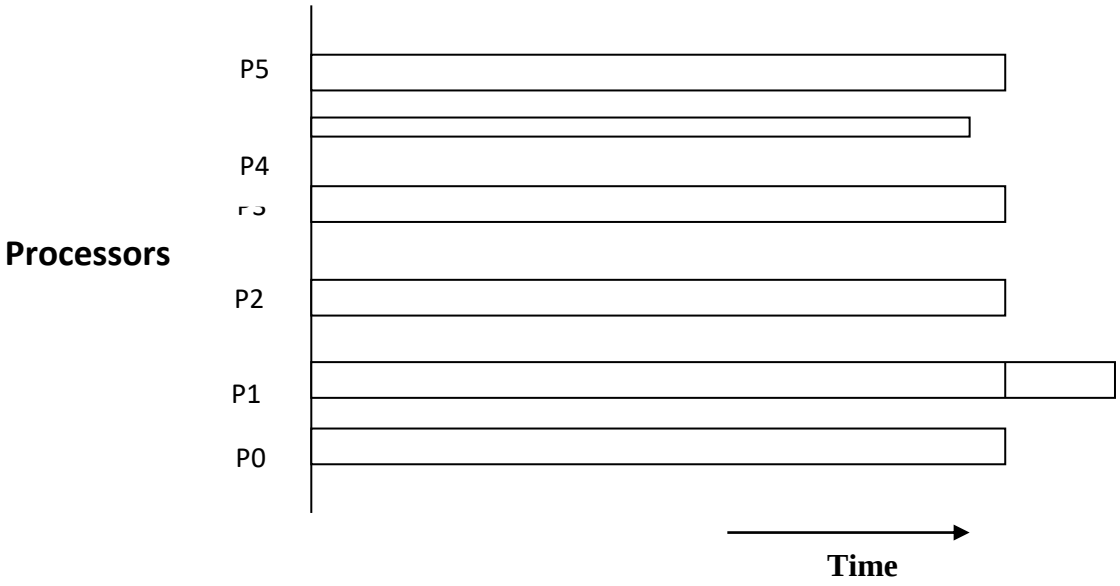
When considering a cluster implementation ,there are basic question that can help determine the cluster attributes such that technology options can be evaluated:

i) Will the application be primarily processing a single data set?
ii) Will the application be parsing data around or will it generate real- time information?
iii) Is the application a 32 bits or 64 bits?

The answers to these question will influence the type of CPU , memory architecture storage ,cluster and interconnected and cluster network design. Cluster application are often CPU- bound so that interconnection and storage benefits are not limiting factors, although this is not always the case.

LOAD BALANCING

Load balancing is used to distribute computation fairly across processors in order to obtain the highest possible execution speed. Thus ,a Load balancing Algorithm tries to balance the total system load by transparently transferring the workload from heavily loaded nodes to lightly loaded nodes in an attempt to ensure good overall performance relative to some specific metric of system performance. When considered from the resource poin of view, to metric is total system throughput. in contrast to response time, throughput is concerned with seeing that all the users are treated fairly and that all are making progress. The resource view of maximizing resource utilization is compatible with the desire to maximize system throughput.
The basic goal of all Load balancing algorithms is to maximize total system throughput



(a) Imperfect load balancing leading to increased execution time

(b) Perfect load balancing

SYSTEM STATE

A system Store various status information

- Current processor load
- The application system load
- Number of active users
- The availability of network protocol buffers
- Other specific resources

REFERENCES

- IEEE July's 99 Cluster Computing-A high Performance Contender(Technical Activities Forum)
- Computer Society Connection August' 99 Cluster Computing
- The international Journal of High Performance Computing application Vol.15 Summer 2001
- The international Journal of High Performance Computing application Vol.13 Summer 2000
- Beowulf Cluster Site http:// www.beowulf.com
- www.viarch.org