Presentation on cluster computing

BY
TUSHAR KANTI ROYTH
DEPT->C.S.E 4TH YEAR
7TH SEMISTER
ROLL- [L-63 ]
Introduction:

• Clustering is the use of multiple computers, storage devices, and redundant interconnections, to form what appears to users as a singly highly available system.

Computer cluster technology puts clusters of systems together to provide better system reliability and performance.

• Definition:-

“A cluster is a type of parallel or distributed processing system, which consists of a collection of interconnected stand alone computers cooperatively working together as a single integrated computing resource.”
A SIMPLE CLUSTER LAY OUT:

Figure 1. A typical Rocks cluster layout.
Classifications
of
Cluster Computer
Clusters Classification..1

- This type of Cluster classification **based on application** Target, and again classified into two subcategories:
  - High Performance (HP) Clusters
    - Grand Challenging Applications
  - High Availability (HA) Clusters
    - Mission Critical applications
HA Cluster: Server Cluster with "Heartbeat" Connection
Clusters Classification..2

- This type Clusters classification based on **Node Ownership** are again classified into two subcategories:

- Dedicated clusters
- Non-dedicated clusters
  
  Adaptive parallel computing

  Also called Communal multiprocessing
Clusters Classification

- This type of classification based on Node Architecture
  - Clusters of PCs (CoPs)
  - Clusters of Workstations (COWs)
  - Clusters of SMPs (CLUMPs)
Clusters Classification..4

- This type of classification based on **node OS Type..**
  - Linux Clusters (Beowulf)
  - Solaris Clusters (Berkeley NOW)
  - NT Clusters (HPVM)
  - AIX Clusters (IBM SP2)
  - SCO/Compaq Clusters (Unixware)
  - .......Digital VMS Clusters, HP clusters, ......................
Clusters Classification..5

- This type of classification based on node components architecture & configuration (Processor Arch, Node Type PC/Workstation.. & OS: Linux/NT..):

  - Homogeneous Clusters
    - All nodes will have similar configuration
  - Heterogeneous Clusters
    - Nodes based on different processors and running different OS es.
Clustering methods:

- **Clustering method can be divided into two types:**
  - A) **Hierarchical method.**
  - B) **Partition method.**

- **A) Hierarchical method:**

  This method proceed successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The clustering methods differ in the rule by which it is decided which two small clusters are merged or which large cluster is split. The end result of the algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained.
B) **Partitional method** :-

- This method attempts to directly decompose the data set into a set of disjoint clusters. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters.
Clustering algorithms:

- Clustering algorithm may be classified into as listed below:
  - Exclusive Clustering
  - Overlapping Clustering
  - Hierarchical Clustering
  - Probabilistic Clustering
Now among the several clustering algorithm, the most popular and very usable clustering algorithms are:

- **Fuzzy C-means**
- **K-means**
- **Hierarchical clustering**
- **Mixture of Gaussians**

*Each of these algorithms belongs to one of the clustering types on previous slide.*

- Such that, K-means is an exclusive clustering algorithm.
- Fuzzy C-means is an overlapping clustering algorithm.
- Hierarchical clustering is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm.
Now we discuss most popular algorithm K-means algorithm.

- *K-means algorithm will do* the three steps:
  - Iterate until stable (= no object move group):
    - Determine the centroid coordinate
    - Determine the distance of each object to the centroids
    - Group the object based on minimum distance
**K-mean algorithm**

- Decide on a value for $K$, the number of clusters.

- Initialize the $K$ cluster centers (randomly, if necessary).

- Decide the class memberships of the $N$ objects by assigning them to the nearest cluster center.

- Re-estimate the $K$ cluster centers, by assuming the memberships found above are correct.

- Repeat 3 and 4 until none of the $N$ objects changed membership in the last iteration.
Comments on k-mean algorithm

• **Strength**
  • Simple, easy to implement and debug
  • Intuitive objective function: optimizes intra-cluster similarity
  • Relatively efficient: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.

• **Weakness**
  • Applicable only when mean is defined, then what about categorical data?
  • Often terminates at a local optimum. Initialization is important.
  • Need to specify K, the number of clusters, in advance
  • Unable to handle noisy data and outliers
  • Not suitable to discover clusters with non-convex shapes

• **Summary**
  • Assign members based on current centers
  • Re-estimate centers based on current assignment
Architectures of a cluster

Parallel Applications

Sequential Applications

Parallel Programming Environments

Cluster Middleware
(Single System Image and Availability Infrastructure)

PC/Workstation
Comm. SW
Net. Interface HW

PC/Workstation
Comm. SW
Net. Interface HW

PC/Workstation
Comm. SW
Net. Interface HW

PC/Workstation
Comm. SW
Net. Interface HW

High Speed Network/Switch
Architectures of a cluster:-

• The main component to design a cluster architectures is :-
  [ 1 ] – multiple stand alone computers  
    ( such as pcs, workstation, smps )
• [ 2 ] – operating system ( linux )
• [ 3 ] – a high performancs interconnect
• [ 4 ] – communication softwear
• [ 5 ] - cluster middlewear
• [ 6 ] - different application platform
    ( parallel application, sequential application )
COMPONENTS OF CLUSTER COMPUTER

The cluster consists of four major parts. These parts are:

1) Network,
2) Compute nodes,
3) Master server,
4) Gateway.

• Each part has a specific function that is needed for the hardware to perform its function.

1. Network:
   Provides communication between nodes, server, and gateway
   Consists of fast Ethernet switch, cables, and other networking hardware

2. Nodes:
   Serve as processors for the cluster
   Each node is interchangeable, there are no functionality differences between nodes
   Consists of all computers in the cluster other than the gateway and server
3. **Server:**

   Provides network services to the cluster

   DHCP

   NFS (Node image and shared file system)

   Actually runs parallel programs and spawns processes on the nodes

   Should have minimum requirement.

4. **Gateway:**

   Acts as a bridge/firewall between outside world and cluster

   Should have two ethernet card
**Why is it required?**

- Clusters are surprisingly powerful. Super computers are really expensive, and maintains cost is too high, but clusters are cheaper and much faster than super computer. In some area of researches cluster are actually faster then commercial super computer. Sometime we don’t even have to use new equipment to build a cluster.
Application :-

- Scientific computing
- Making movie
- Commercial server (web/database etc)
Advantage on cluster computing

- Size Scalability (physical & application)
- Enhanced Availability (failure management)
- Single System Image (look-and-feel of one system)
- Fast Communication (networks & protocols)
- Load Balancing (CPU, Net, Memory, Disk)
- Security and Encryption (clusters of clusters)
- Distributed Environment (Social issues)
- Manageability (admin. And control)
- Programmability (simple API if required)
- Applicability (cluster-aware and non-aware app.)
Disadvantage on cluster computing

- Can be heard to manage with out experience
- finding out where something has failed increases at least linearly as cluster size increases
- programming environment could be vastly improve
- when software configuration on some node is different then on others node
Conclusion:

- Clusters are promising Solve parallel processing paradox
- Offer incremental growth and matches with funding pattern

New trends in hardware and software technologies are likely to make
- clusters more promising and fill SSI gap.
- Clusters based supercomputers (Linux based clusters) can be seen
- everywhere!
References

www.buyya.com
www.beowulf.org
www.clustercomp.org
www.sgi.com
www.thu.edu
www.dgs.monash.edu
www.cfi.lu
www.webopedia.com
www.howstuffworks.com