COMPUTING CLUSTERS

Farniba Khan

farniba.khan@student.lut.fi

Computing clusters, grids and clouds by Andrey Y. Shevel

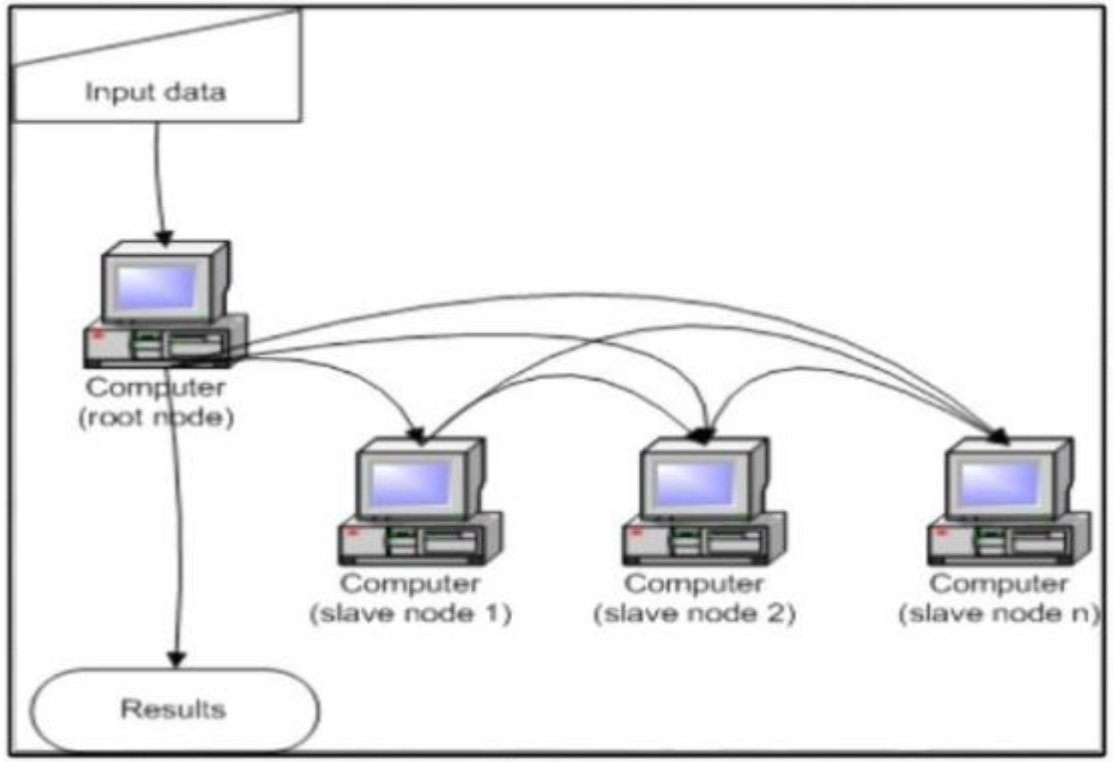
ITMO University, Saint-Petersburg, June 6, 2017

ITMO UNIVERSITY

PERCCOM asmus undus

INTRODUCTION

A computing cluster is a type of parallel or distributed computer system, which consists of a collection of interconnected stand-alone computers working together as a single integrated computing resource.



HISTORY

- The first inspiration for cluster computing was developed in the 1960s by IBM as an alternative of linking large mainframes to provide a more cost effective form of commercial parallelism.
- The first commodity clustering product was ARCnet, developed by Datapoint in 1977.
- The original PC cluster project, also called Beowulf project, was started at the Center of Excellence in Space Data and Information Sciences NASA in early 1994. It is a system which usually consists of one master or server node, and one or more client nodes connected together via Ethernet
- Microsoft, Sun Microsystems and other leading hardware and software companies offer clustering packages.

Cluster classification

High availability clusters (HA) (Linux)

Mission critical applications

High-availability clusters (also known as Failover Clusters) are implemented for the purpose of improving the availability of services which the cluster provides.

eliminate single points of failure.

Load balancing clusters

operate by distributing a workload evenly over multiple back end nodes.

Typically the cluster will be configured with multiple redundant load-balancing front ends.

all available servers process requests.

Web servers, mail servers,..

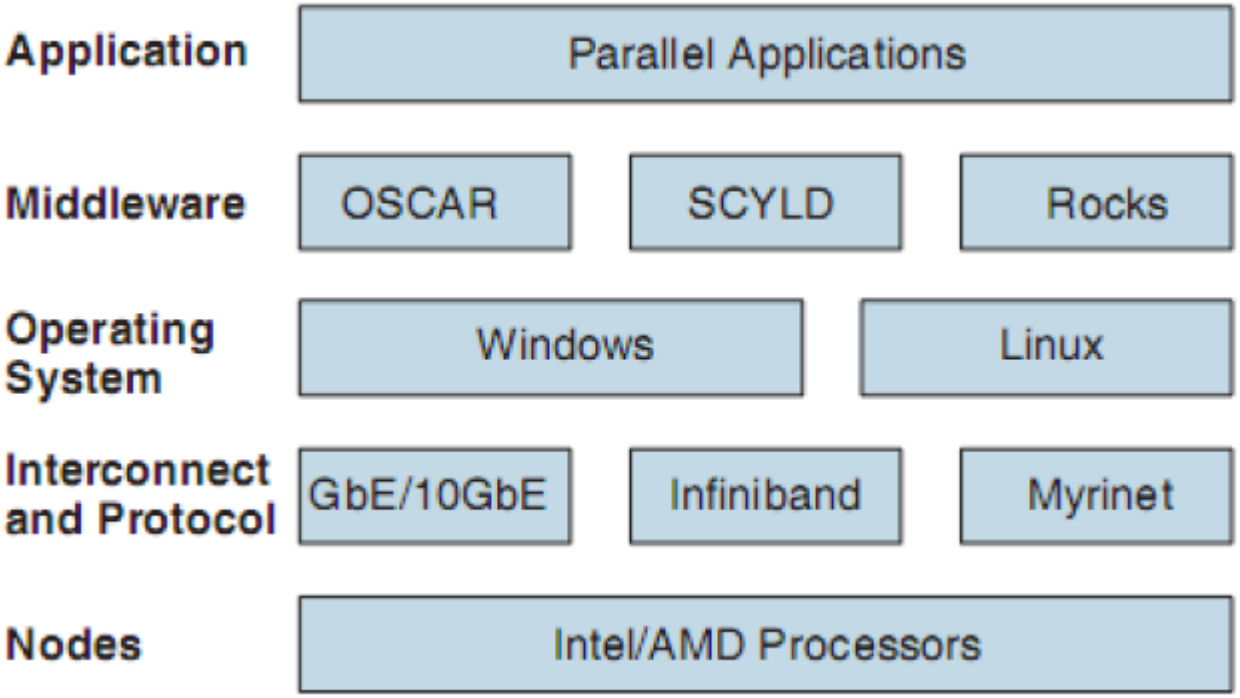
High performance clusters(HPC)

data mining, simulations, parallel processing, weather modeling

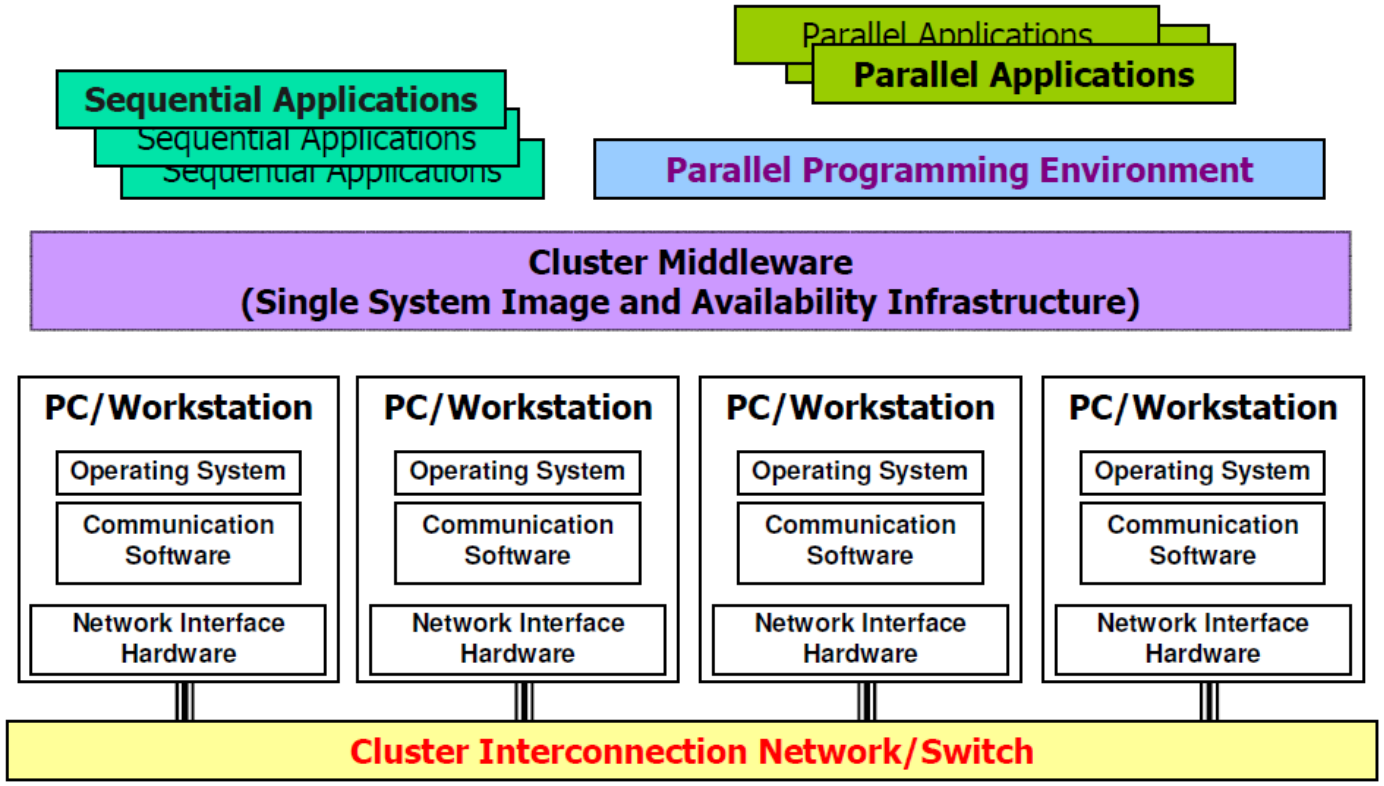
designed to exploit the parallel processing power of multiple nodes.

Example: Beowulf cluster

COMPONENTS


Application
Parallel Applications
Middleware
OSCAR
SCYLD
Rocks
Operating System
Windows
Linux
Interconnect and Protocol
GbE/10GbE
Infiniband
Myrinet
Nodes
Intel/AMD Processors

ARCHITECTURE



SSI(Single System Image)

- Representation of the view of a distributed system as a single unified computing resource.
- hides the complexities of the underlying distributed and heterogeneous nature of clusters from user.
- The design goals for SSI cluster-based systems focus on complete transparency of resource management, scalable performance, and system availability in supporting user applications.

Different level of Abstraction in SSI

Hardware	Memory Channel - Distributed Shared Memory
Operating Systems	MOSIX- Solaris MC-UnixWare
Middleware	Condor-Loadleveler - Load Share Facility (LSF)- Open Portable BatchSystem(OpenPBS) – Sun Grid Engine (SGE)- Libra
Application	PARMON-Linux Virtual Server- Problem Solving Environments
Programming	Linda- JavaSpaces-Message Queues- Parameter Sweep-Parallel Virtual Machine(PVM)- JavaGroups-Message Passing Interface(MPI)

ISSUES TO BE CONSIDERED

- Cluster Networking

Machines need to have similar networking capabilities such as speed, communication abilities in case of mixing different networking technologies.

- Cluster Software

For each kind of system in cluster, different versions of clustering software should be build.

- Programming

Codes have to be written to support all types of nodes in the cluster.

- Timing

Due to different performance level of nodes, processing time on node may vary.

- Security

There is always a possibility of attack whenever an interaction happens with the client and the server. Therefore, the nodes, the connections and the implementation of nodes when it comes to handling workload could be optimized for security.

ADVANTAGES & DISADVANTAGES

Advantages:

- Cost Efficiency
- Processing speed
- Improved Network infrastructure
- Scalability
- High Availability of resources

Disadvantages:

- Programmability issues
- Difficult to Manage

APPLICATIONS

- Compute Intensive Applications
- Data or I/O Intensive Applications
- Transaction Intensive Applications

Examples:

- Google Search Engine.
- Petroleum Reservoir Simulation.
- Protein Explorer.
- Earthquake Simulation.
- Image Rendering.

Google Search Engine

- Google uses cluster computing to meet the huge quantity of worldwide search requests that comprise of a peak of thousands of queries per second.
- The services for Google are also replicated across multiple machines in the clusters to provide the necessary availability.
- A hardware-based load balancer in the cluster distributes the requests evenly.
- The Google Web Servers machine receives the request, coordinates the query execution and sends the search result back to the user's browser.

Example: Beowulf Cluster

- Uses parallel processing across multiple computers to create cheap and powerful supercomputers
- Few computers and Ethernet segments, Myrinet, Infiniband or Quadrics can be used.
- Allows OS to run on every node and still allow parallel processing.
- Atleast two computers with Linux based distribution for the clustering
- A cluster has two types of computers,
  - a master computer, and
  - node computers.

  When a large problem or set of data is given to a Beowulf cluster,
  - the master computer first runs a program that breaks the problem into small discrete pieces;
  - it then sends a piece to each node to compute.
  - As nodes finish their tasks, the master computer continually sends more pieces to them until the entire problem has been computed.

REFERENCES

- https://www.slideshare.net/poojakhatana1/cluster-computing-28886643?next_slideshow=4
- https://www.slideshare.net/nikhil9177/cluster-computing?next_slideshow=2
- https://www.slideshare.net/Kajal_Thakkar/cluster-computing-59182598?next_slideshow=3
- https://www.techopedia.com/definition/6581/computer-cluster
- https://www.slideshare.net/rajamohd2/cluster-computing-56275811
- http://www.exforsys.com/tutorials/clustering/clustering-security.html