Cluster, Grid and Cloud computing

S.Chellammal

13-03-2014

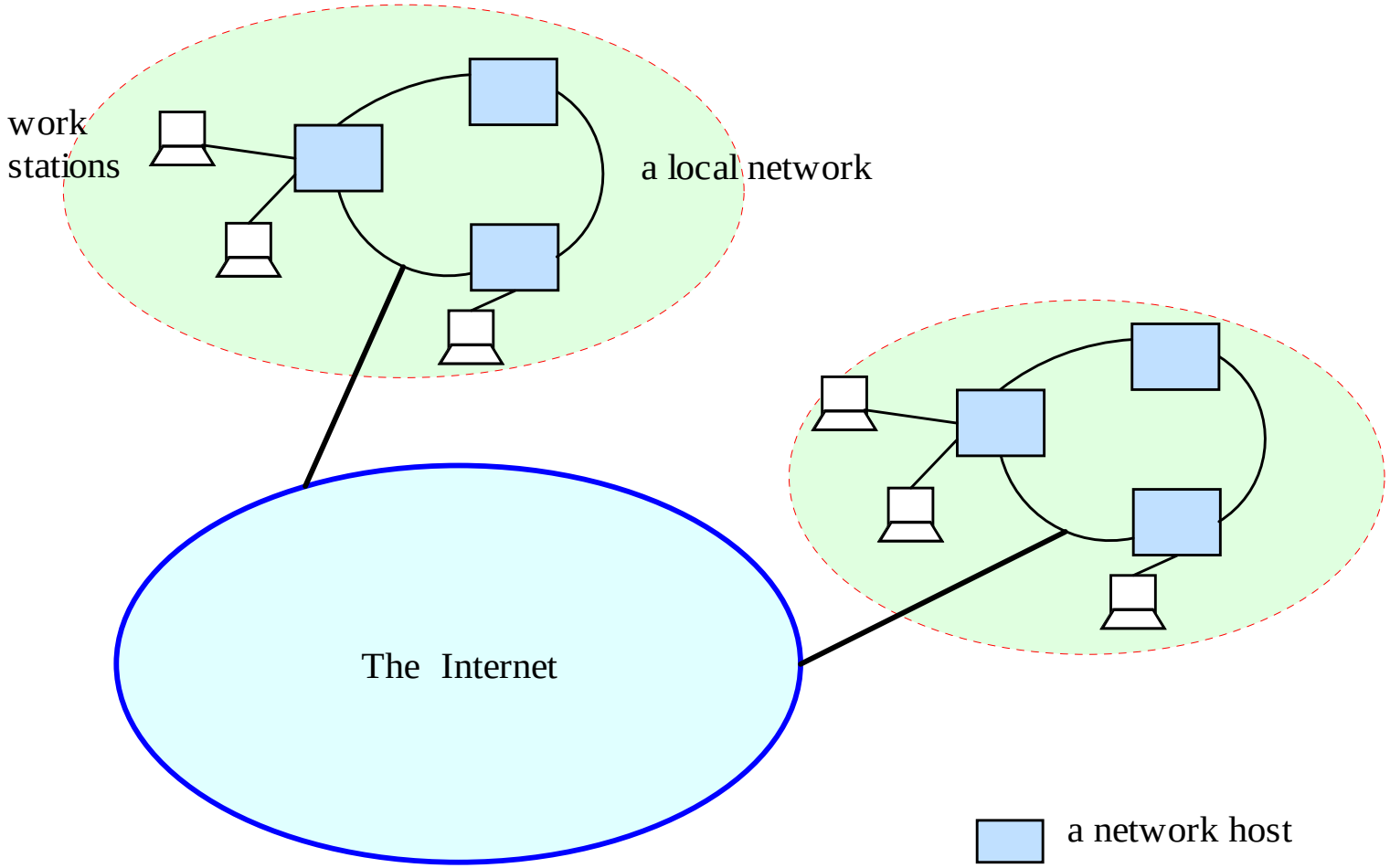
Distributed system, distributed computing

- Early computing was performed on a single processor. Uni-processor computing can be called *centralized computing*.
- A *distributed system* is a collection of independent computers, interconnected via a network, capable of collaborating on a task.
- *Distributed computing* is computing performed in a distributed system.

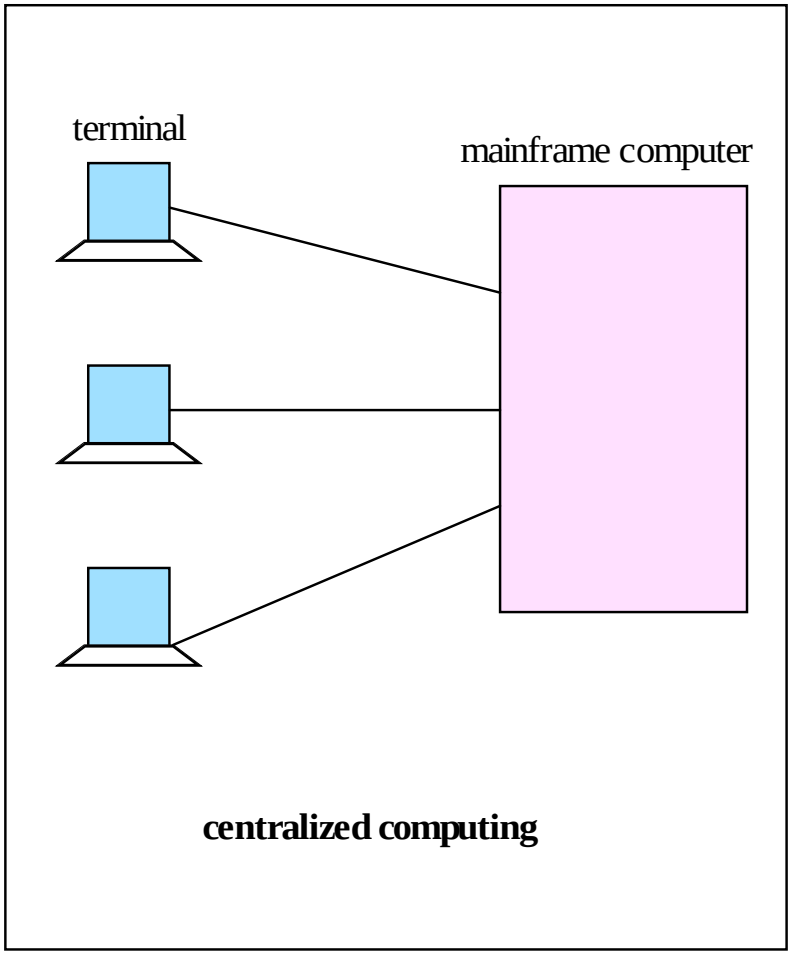
Examples of Distributed systems

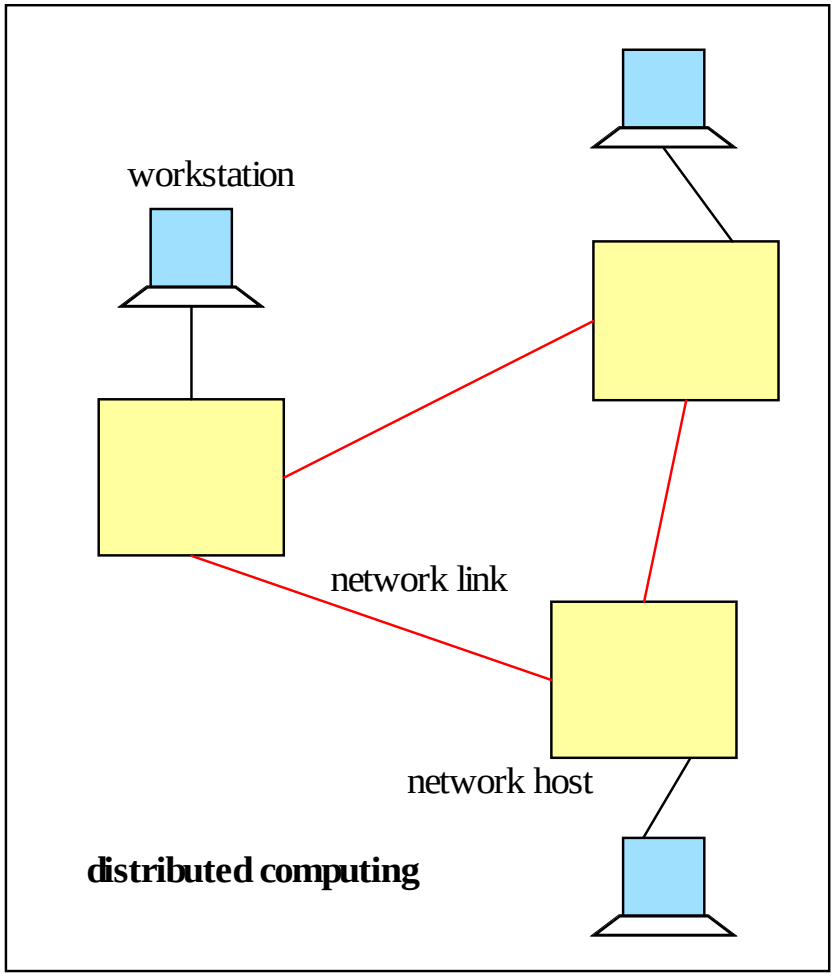
- Network of workstations (NOW) / PCs: a group of networked personal workstations or PCs connected to one or more server machines.
- The Internet
- An intranet: a network of computers and workstations within an organization, segregated from the Internet via a protective device (a firewall).

Distributed Systems



From centralized to Distributed Computing





Monolithic mainframe applications vs. distributed applications

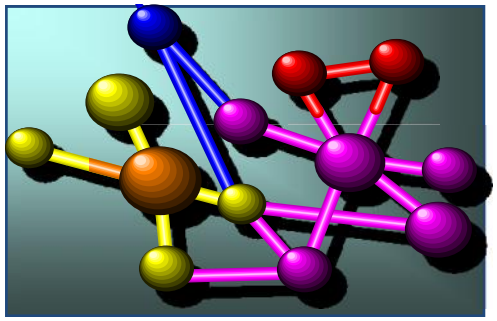
- The monolithic mainframe application architecture:
  - Separate, single-function applications, such as order-entry or billing
  - Applications cannot share data or other resources
  - Developers must create multiple instances of the same functionality (service).
- The distributed application architecture:
  - Integrated applications
  - Applications can share resources
  - A single instance of functionality (service) can be reused.

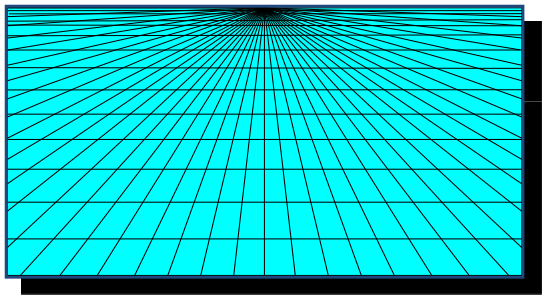
Evolution of paradigms

- Communication with floppies, … and then messages
- Client-server: Socket API, RPC, remote method invocation
- Distributed objects
- Object broker: CORBA
- Network service: Jini
- Message oriented middleware (MOM): Java Message Service
- Collaborative applications
- Web services, services

Need of more Computing Power: Grand Challenge Applications

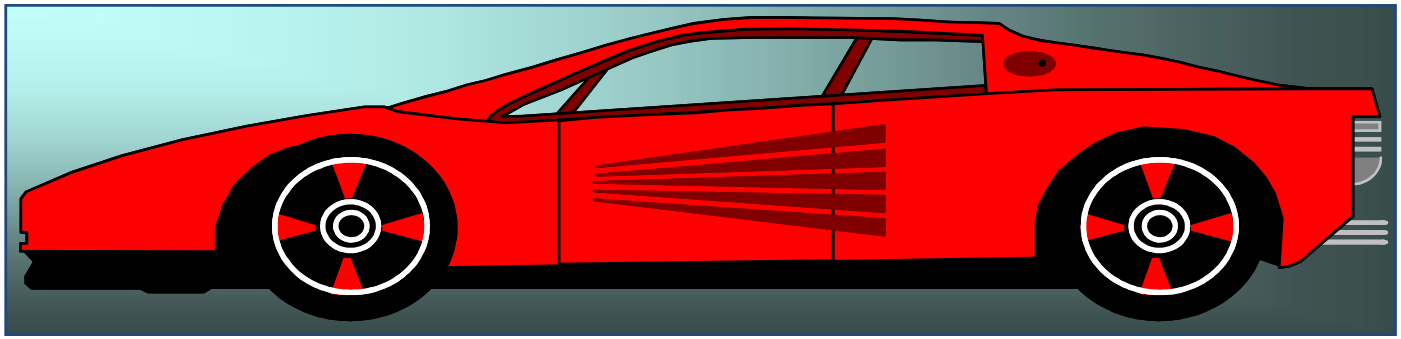
Solving technology problems using computer *modeling* *simulation* *analysis*





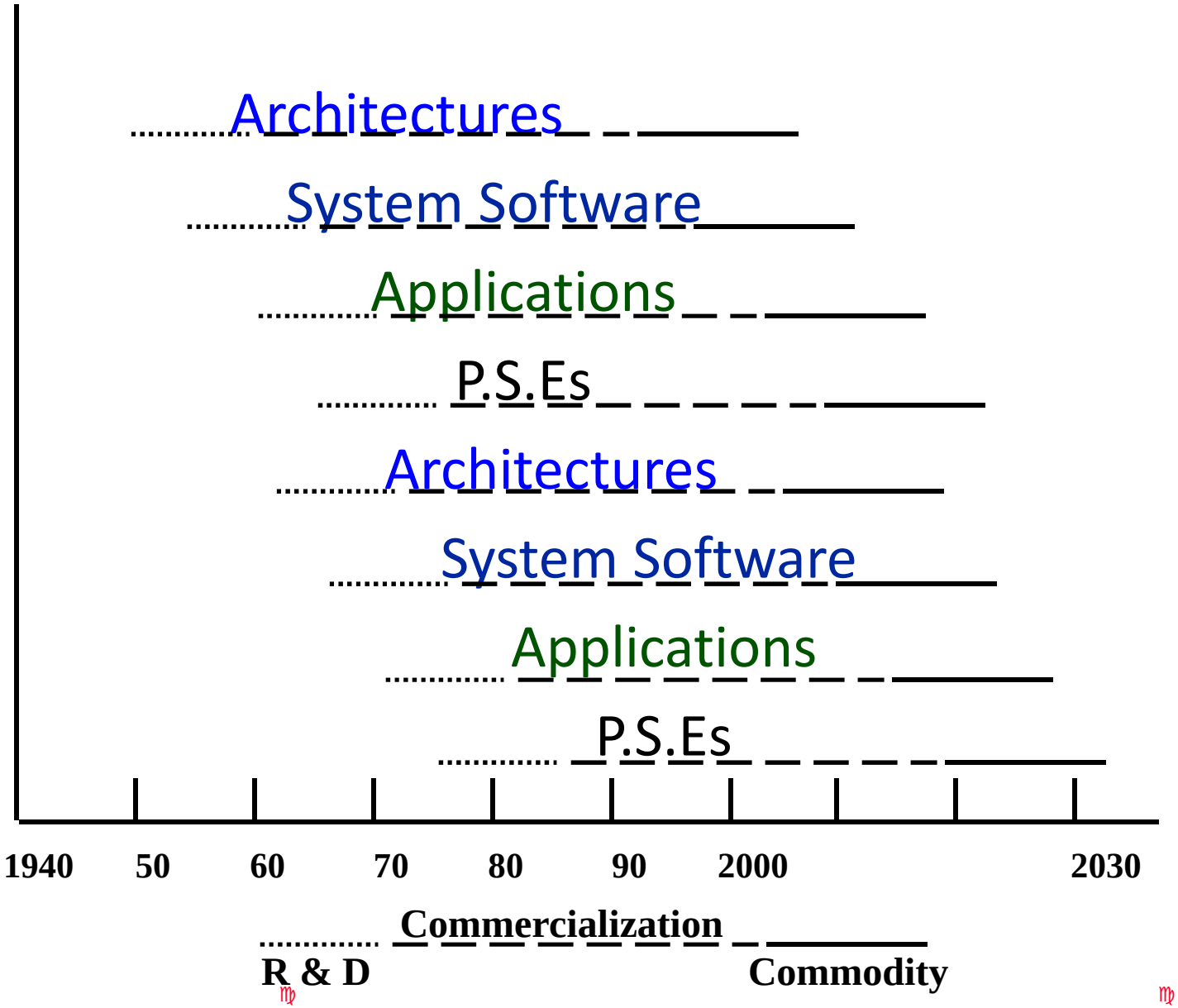
Aerospace

Geographical information systems

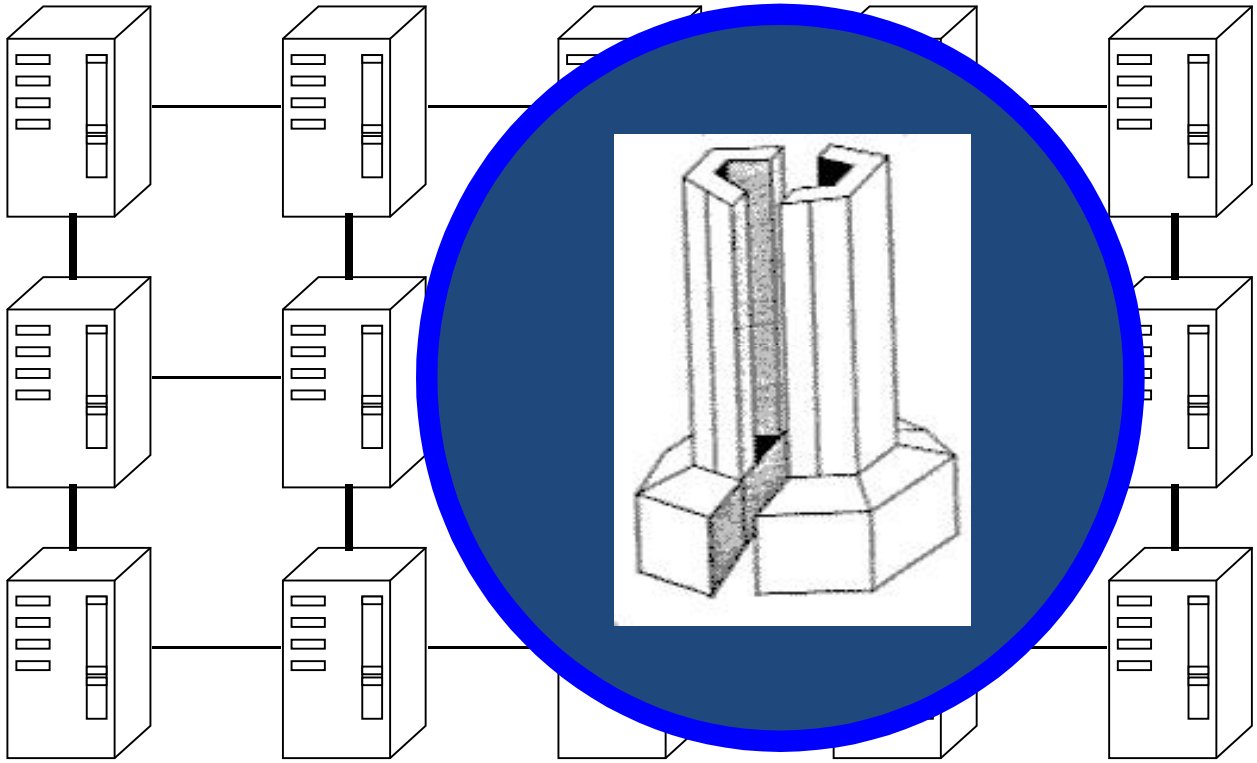


CAD/CAM

Two Eras of Computing

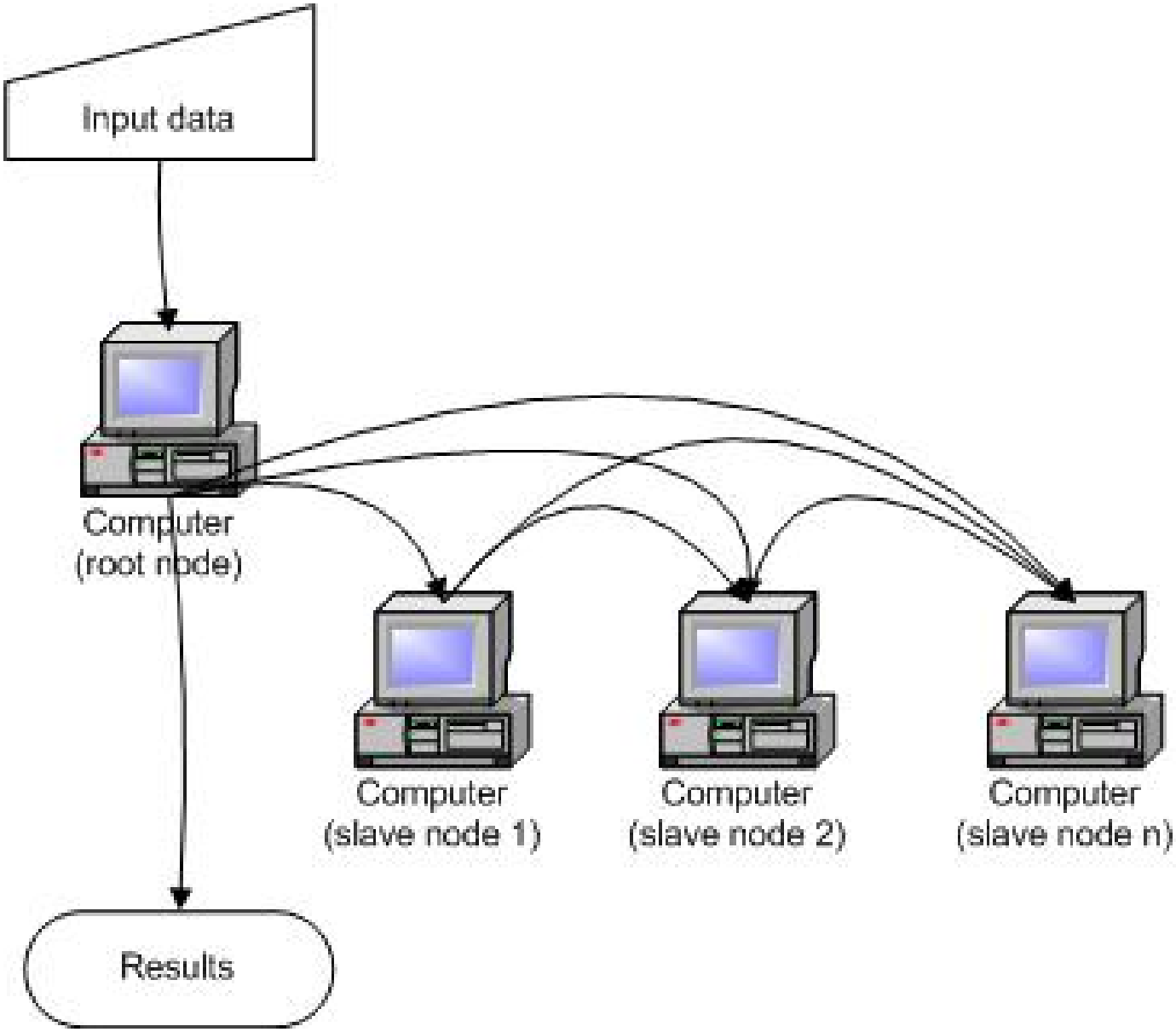


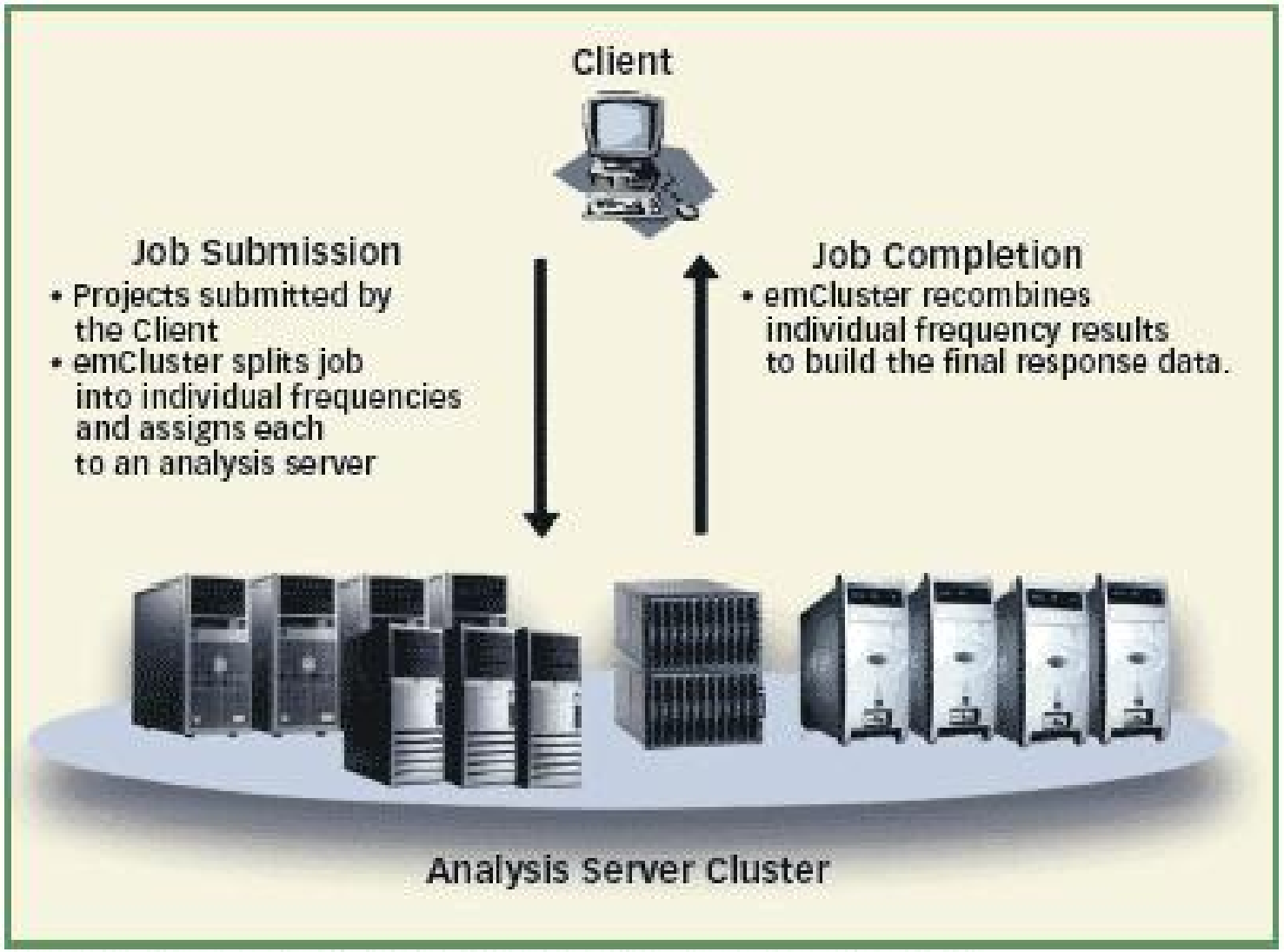
Parallel Proessing on Linux Clusters



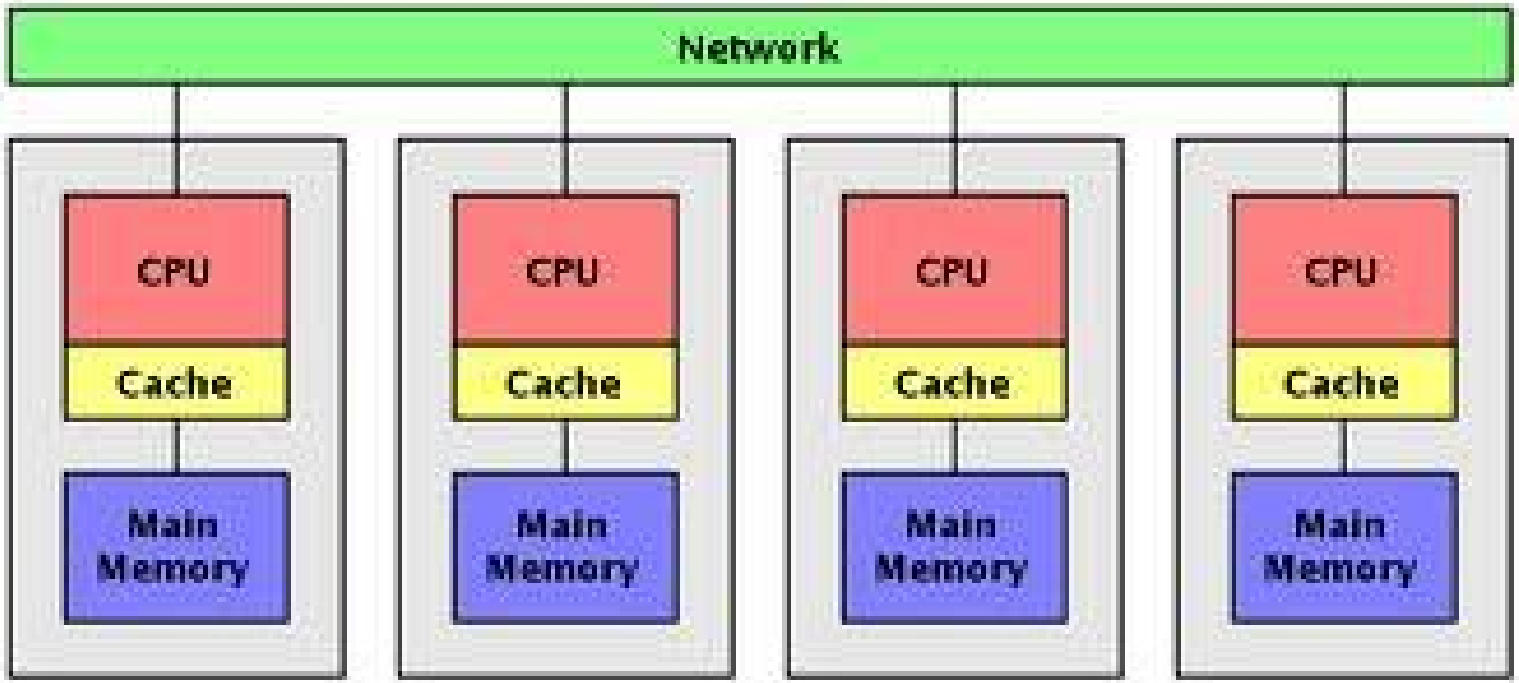
What is a cluster?

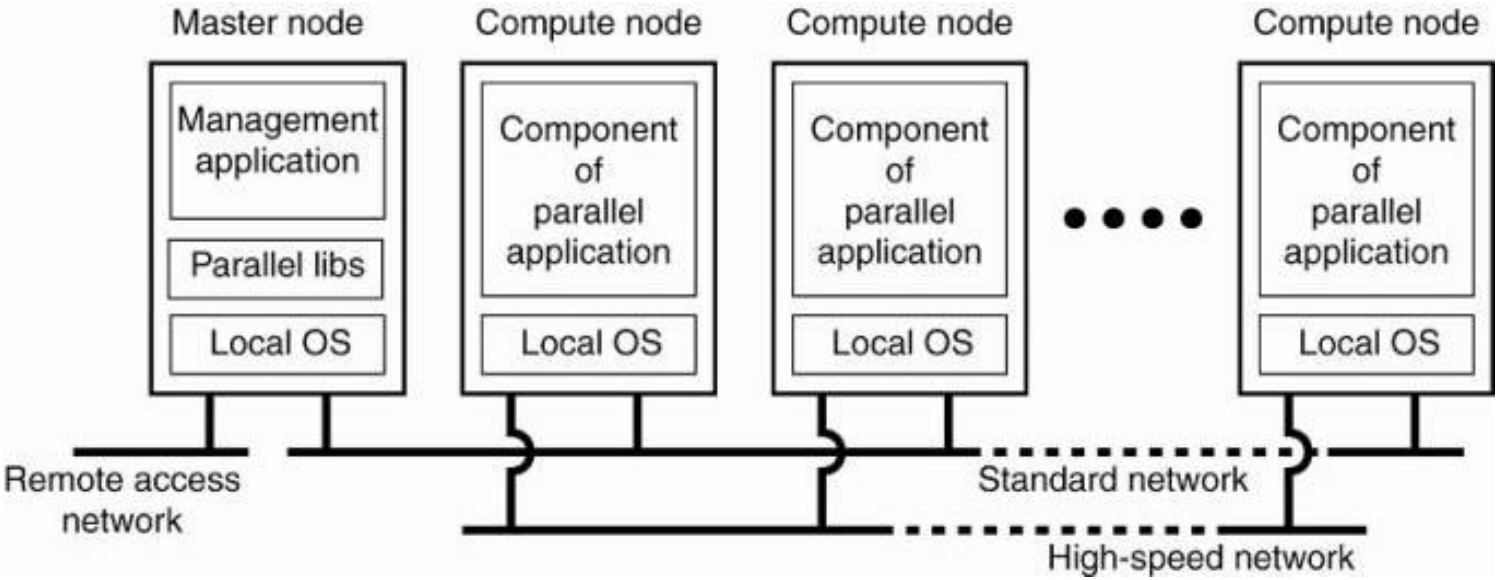
- A cluster is a type of parallel or distributed processing system, which consists of a collection of interconnected stand-alone computers cooperatively working together as a single, integrated computing resource.

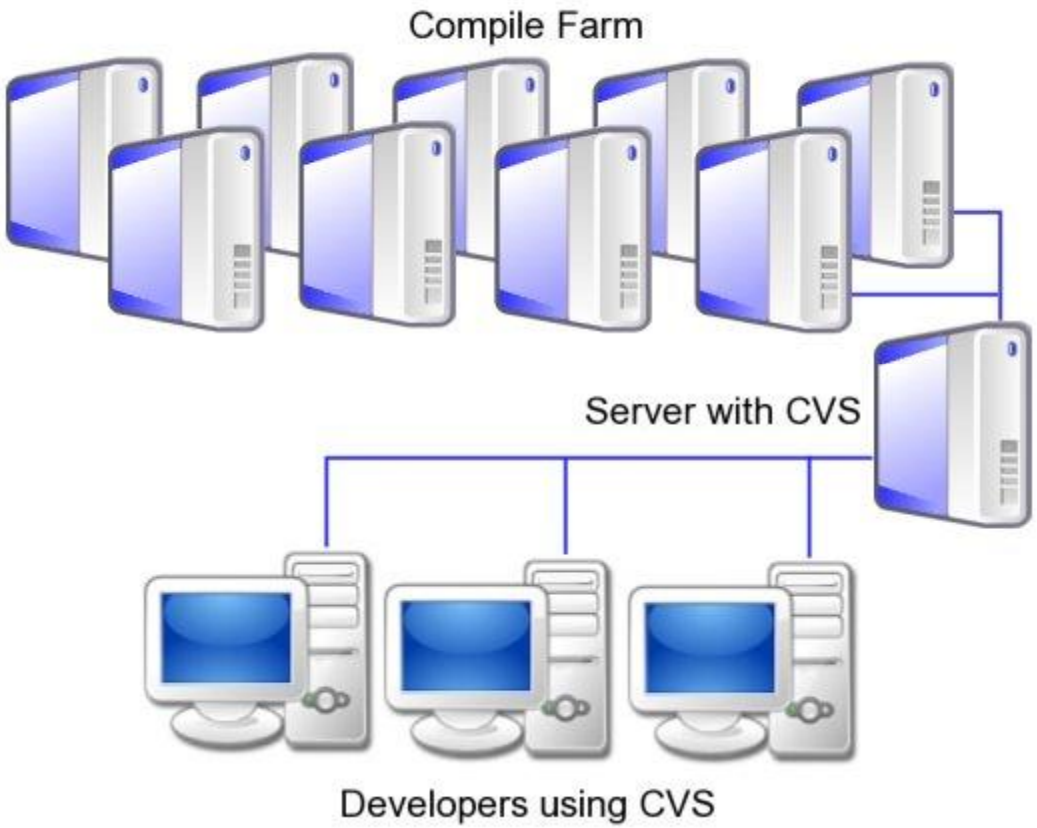

Input data
Computer
(root node)
Computer
(slave node 1)
Computer
(slave node 2)
Computer
(slave node n)
Results



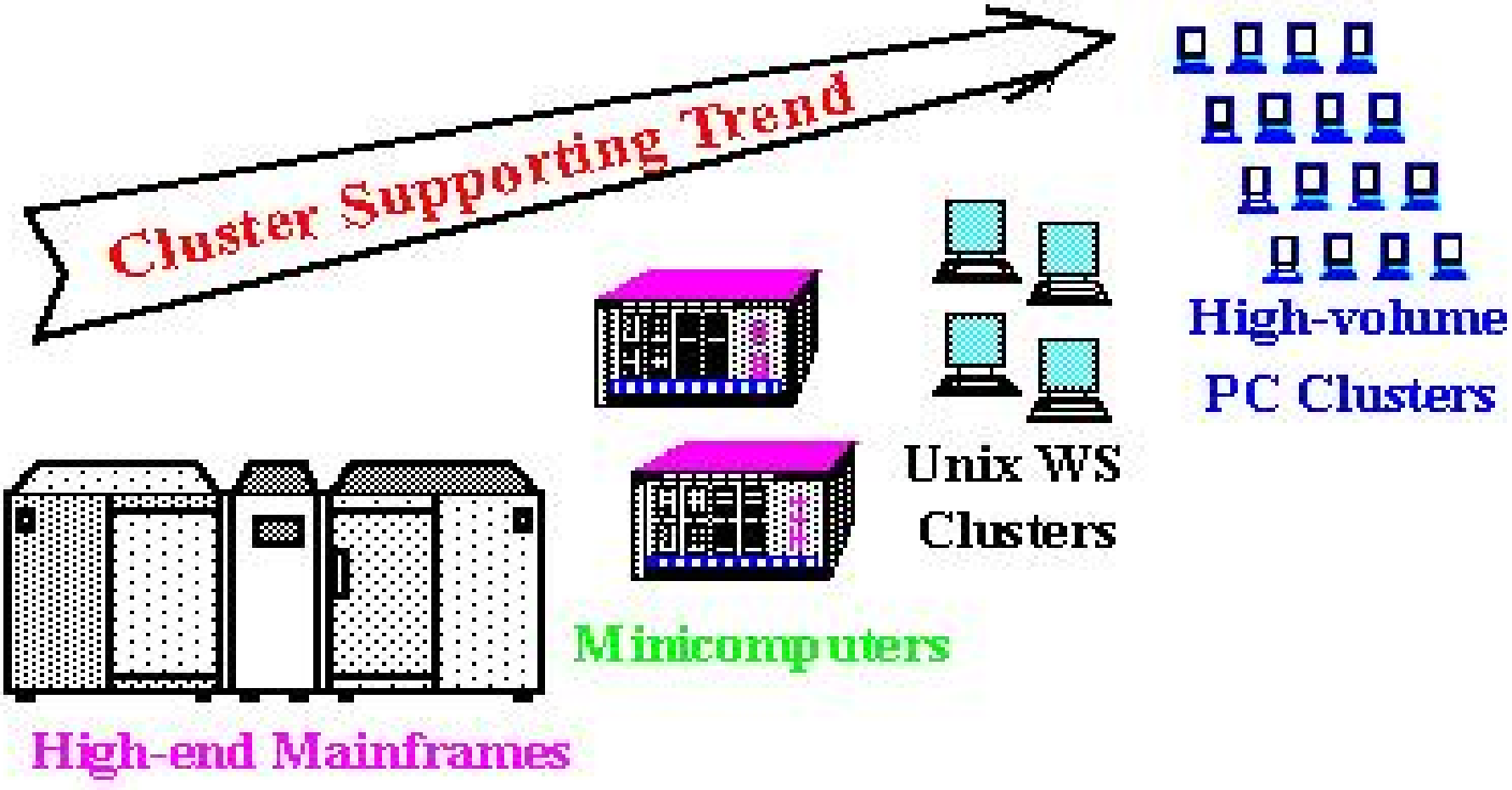
1. This diagram shows the typical emCluster process.


Network
CPU
Cache
Main Memory
CPU
Cache
Main Memory
CPU
Cache
Main Memory
CPU
Cache
Main Memory


Master node
Compute node
Compute node
Compute node
Management application
Parallel libs
Local OS
Component of parallel application
Local OS
Component of parallel application
Local OS
Component of parallel application
Local OS
Remote access network
Standard network
High-speed network

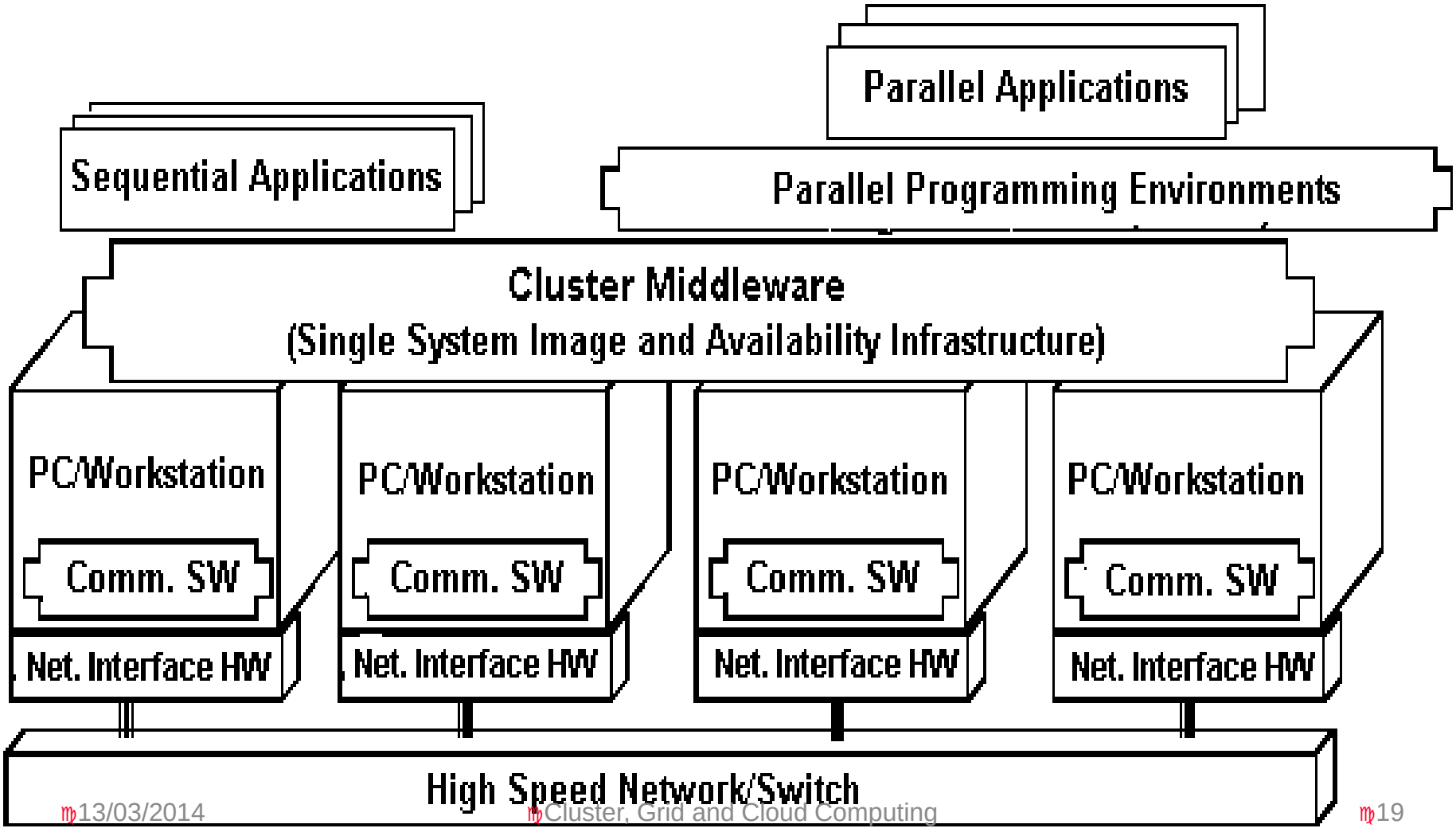

Compile Farm
Server with CVS
Developers using CVS

Clustering of Computers for Collective Computating



Cluster Configuration..1
Dedicated Cluster

Cluster Computer Architecture



Major issues in cluster design

- 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.)

Cluster computing concluding remarks

Advantages

- Solve parallel processing paradox
- High availability
- Offer incremental growth and matches with funding pattern

Disadvantages

- Complexity in administration
- Node – N Node administration

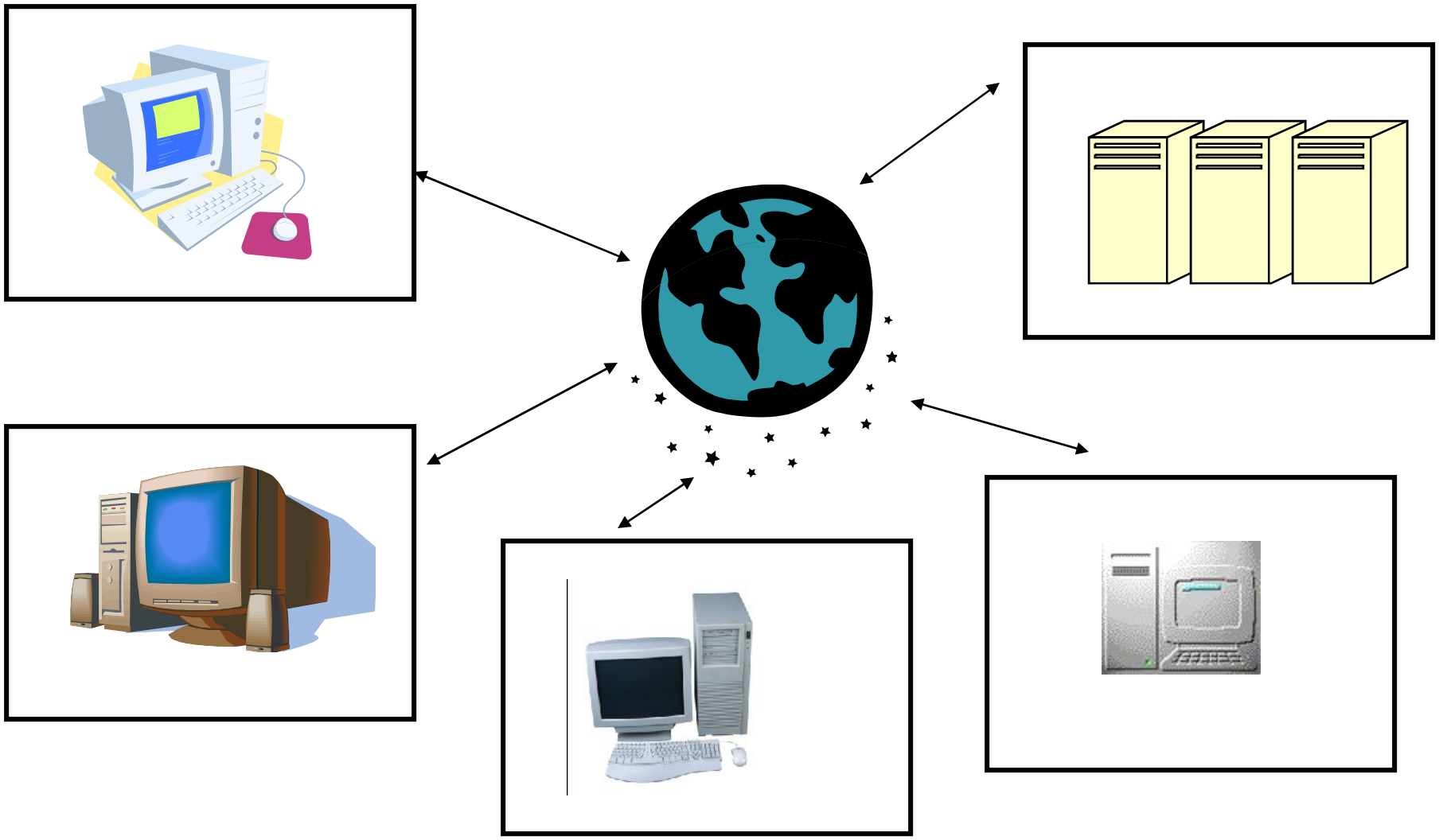
What is a Grid?

- Many definitions exist in the literature
- Early defs: Foster and Kesselman, 1998

  "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational facilities"
- Kleinrock 1969:

  "We will probably see the spread of 'computer utilities', which, like present electric and telephone utilities, will service individual homes and offices across the country."

Grid Architecture



Autonomous, globally distributed computers/clusters

Why do we need Grids?

- Many large-scale problems cannot be solved by a single computer
- Globally distributed data and resources

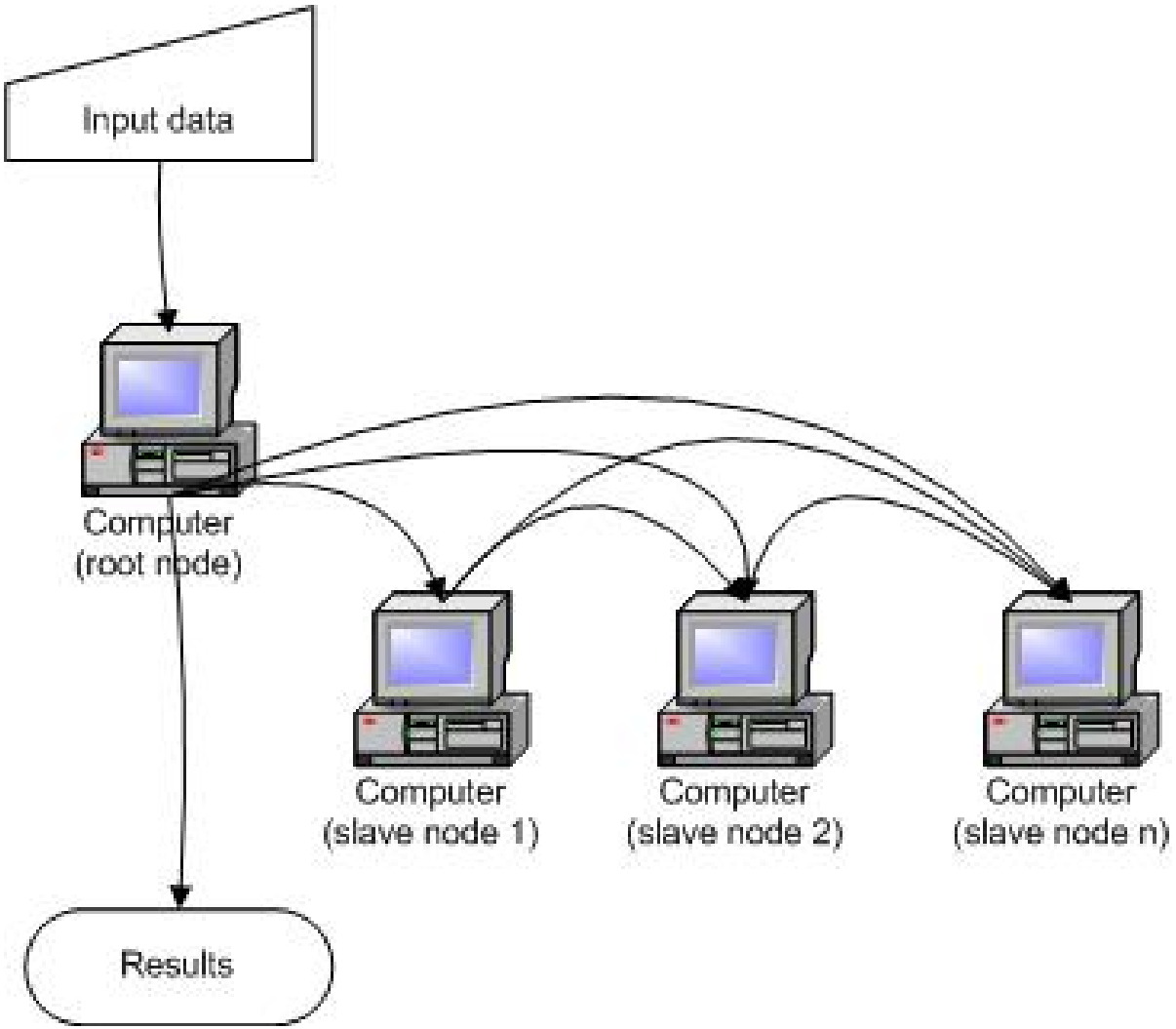
Background: Related technologies

- Cluster computing
- Peer-to-peer computing
- Internet computing

Cluster computing

- Idea: put some PCs together and get them to communicate
- Cheaper to build than a mainframe supercomputer
- Different sizes of clusters
- Scalable – can grow a cluster by adding more PCs

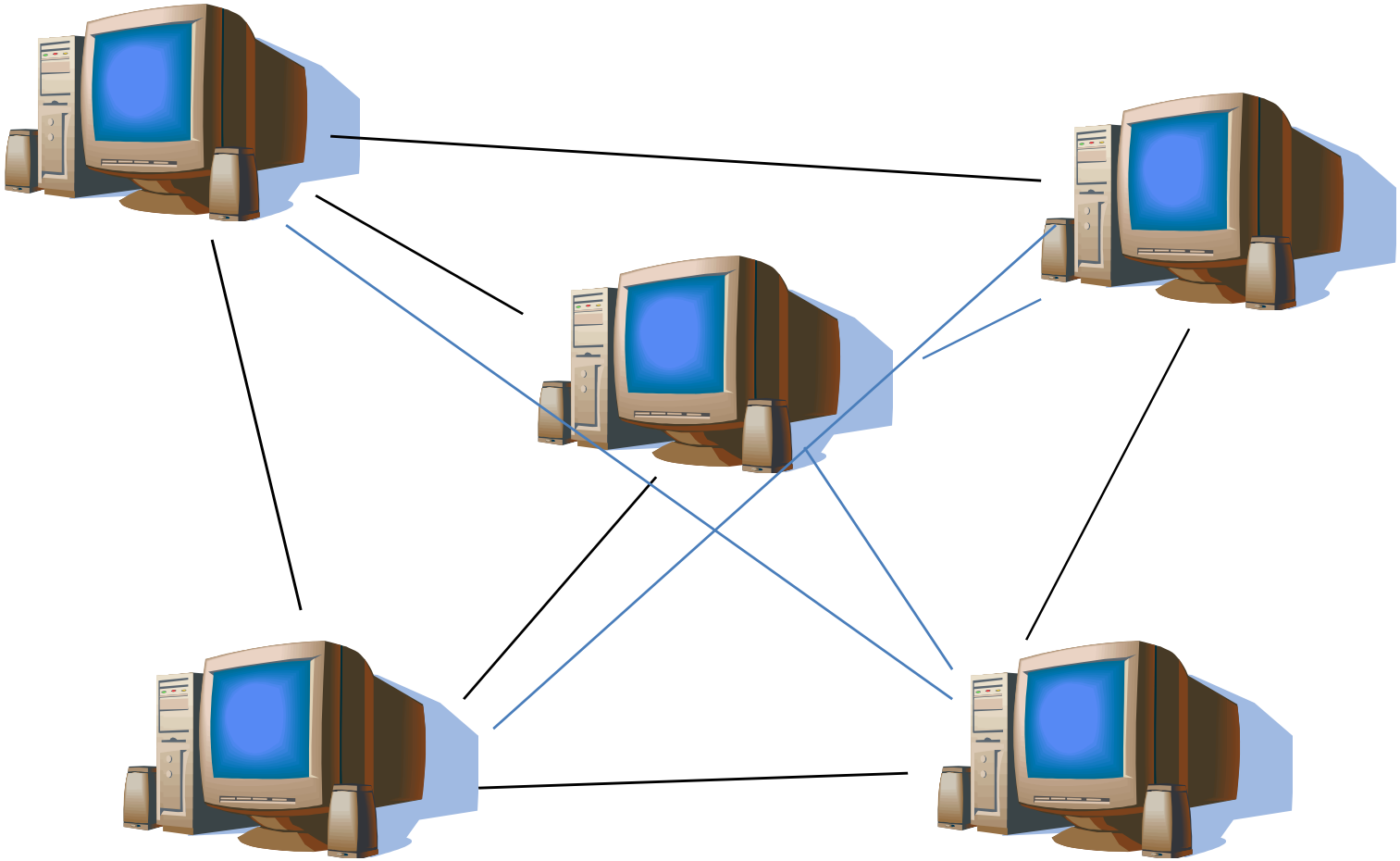
Cluster Architecture



Peer-to-Peer computing

- Connect to other computers
- Can access files from any computer on the network
- Allows data sharing without going through central server
- Decentralized approach also useful for Grid

Peer to Peer architecture



Internet computing

- Idea: many idle PCs on the Internet
- Can perform other computations while not being used

Internet computing

- Idea: many idle PCs on the Internet
- Can perform other computations while not being used

Government

- Small number of users
- Couple small numbers of high-end resources
- Goals:
  - Provide “strategic computing reserve” for crisis management
  - Support collaborative investigations of scientific and engineering problems
- Need to integrate diverse resources and balance diversity of competing interests

Health Maintenance Organization

- Share high-end computers, workstations, administrative databases, medical image archives, instruments, etc. across hospitals in a metropolitan area
- Enable new computationally enhanced applications
- *Private grid*
  - Small scale, central management, common purpose
  - Diversity of applications and complexity of integration

Materials Science Collaboratory

- Scientists operating a variety of instruments (electron microscopes, particle accelerators, X-ray sources) for characterization of materials
- Highly distributed and fluid community
- Sharing of instruments, archives, software, computers
- *Virtual Grid*
  - strong focus and narrow goals
  - Dynamic membership, decentralized, sharing resources

Grid Users

- Many levels of users
  - Grid developers
  - Tool developers
  - Application developers
  - End users
  - System administrators

Some Grid challenges

- Data movement
- Data replication
- Resource management
- Job submission

Some Grid-Related Projects

- Globus
- Condor
- Nimrod-G

Grid Applications

Data and computationally intensive applications:

This technology has been applied to computationally-intensive scientific, mathematical, and academic problems like drug discovery, economic forecasting, seismic analysis back office data processing in support of e-commerce

- A chemist may utilize hundreds of processors to screen thousands of compounds per hour.
- Teams of engineers worldwide pool resources to analyze terabytes of structural data.
- Meteorologists seek to visualize and analyze petabytes of climate data with enormous computational demands.

Resource sharing

- Computers, storage, sensors, networks, …
- Sharing always conditional: issues of trust, policy, negotiation, payment, …

Coordinated problem solving

- distributed data analysis, computation, collaboration, …

Grid Topologies

- Intragrid
  - Local grid within an organisation
  - Trust based on personal contracts
- Extragrid
  - Resources of a consortium of organisations connected through a (Virtual) Private Network
  - Trust based on B to B contracts
- Intergrid
  - Global sharing of resources through the internet
  - Trust based on certification

Computational Grid

"A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities."

Example： Science Grid (US Department of Energy)

Data Grid

- A **data grid** is a grid computing system that deals with data — the **controlled sharing and management of large amounts of distributed data**.

- Data Grid is the storage component of a grid environment. Scientific and engineering applications require access to large amounts of data, and often this data is widely distributed. A data grid provides seamless access to the local or remote data required to complete compute intensive calculations.

Example :

Biomedical informatics Research Network (BIRN),

the Southern California earthquake Center (SCEC).

Methods of Grid Computing

- Distributed Supercomputing
- High-Throughput Computing
- Data-Intensive Computing
- Collaborative Computing
- Logistical Networking

Distributed Supercomputing

- Combining multiple high-capacity resources on a computational grid into a single, virtual distributed supercomputer.

- Tackle problems that cannot be solved on a single system.

High-Throughput Computing

- Uses the grid to schedule large numbers of loosely coupled or independent tasks, with the goal of putting unused processor cycles to work.

Collaborative Computing

- Concerned primarily with enabling and enhancing human-to-human interactions.
- Applications are often structured in terms of a virtual shared space.

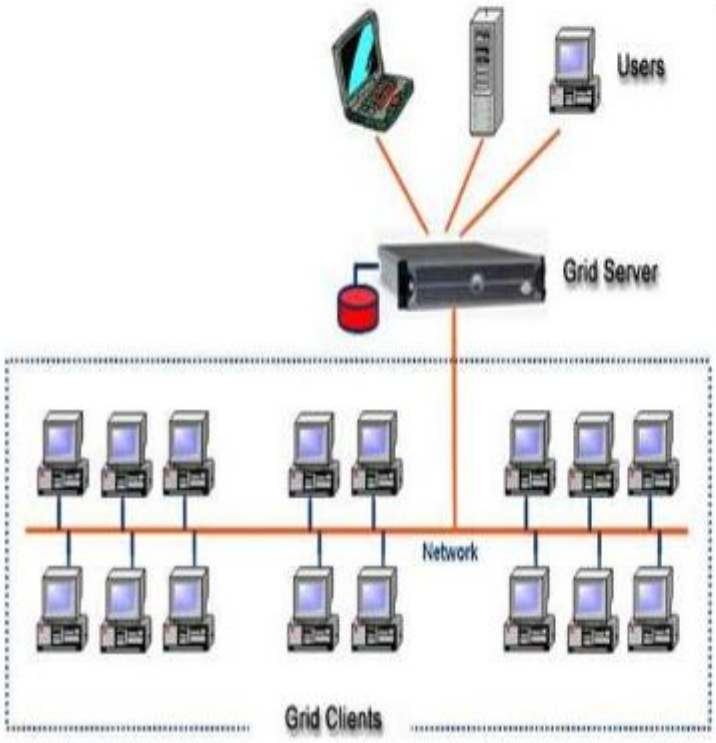
Data Intensive Computing

- The focus is on synthesizing new information from data that is maintained in geographically distributed repositories and databases.
- Particularly useful for distributed data mining.

Logistical Networking

- Logistical networks focus on exposing storage resources inside networks by optimizing the global scheduling of data transport, and data storage.
- Contrasts with traditional networking, which does not explicitly model storage resources in the network.

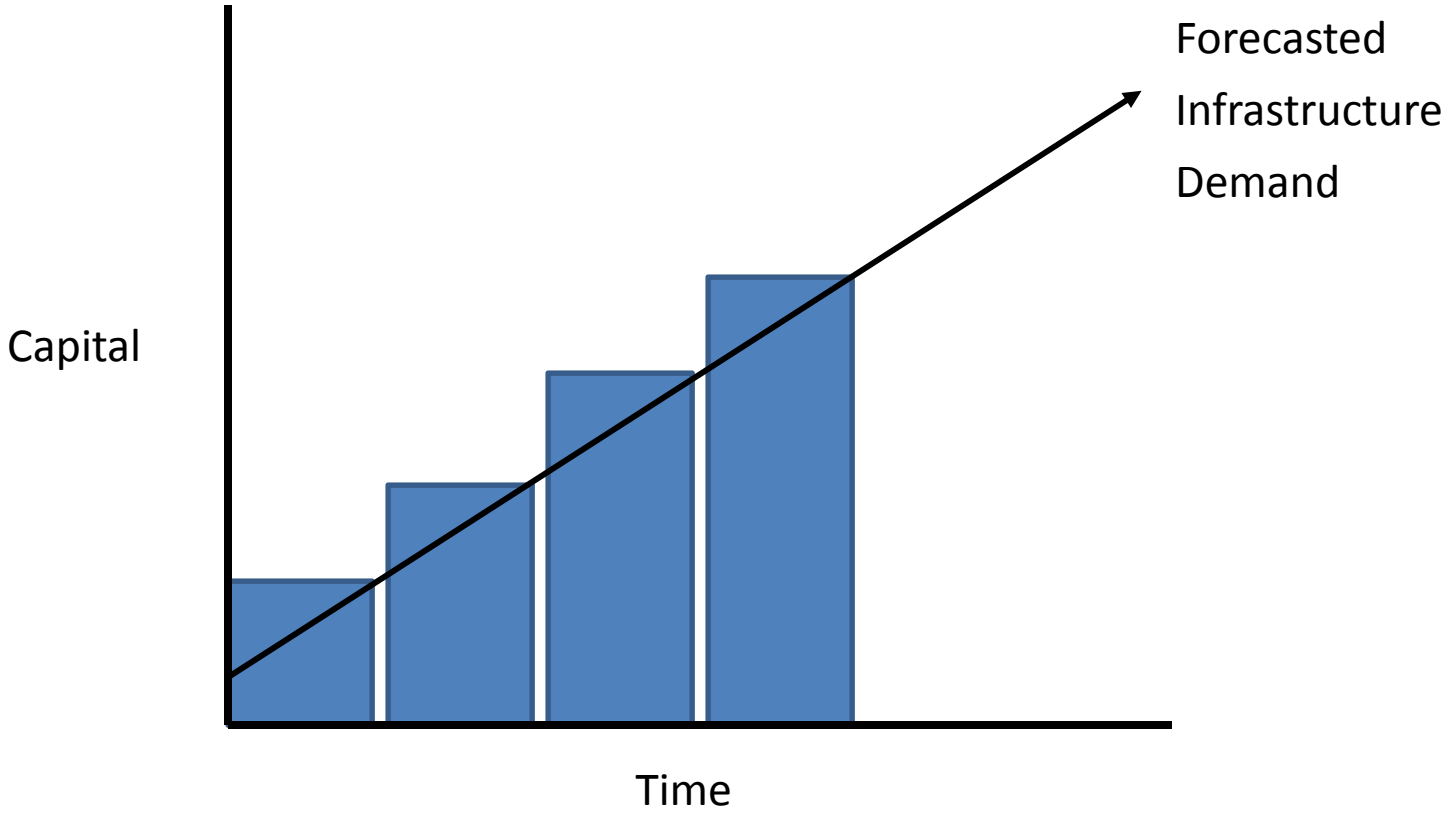
How Grid computing works ?



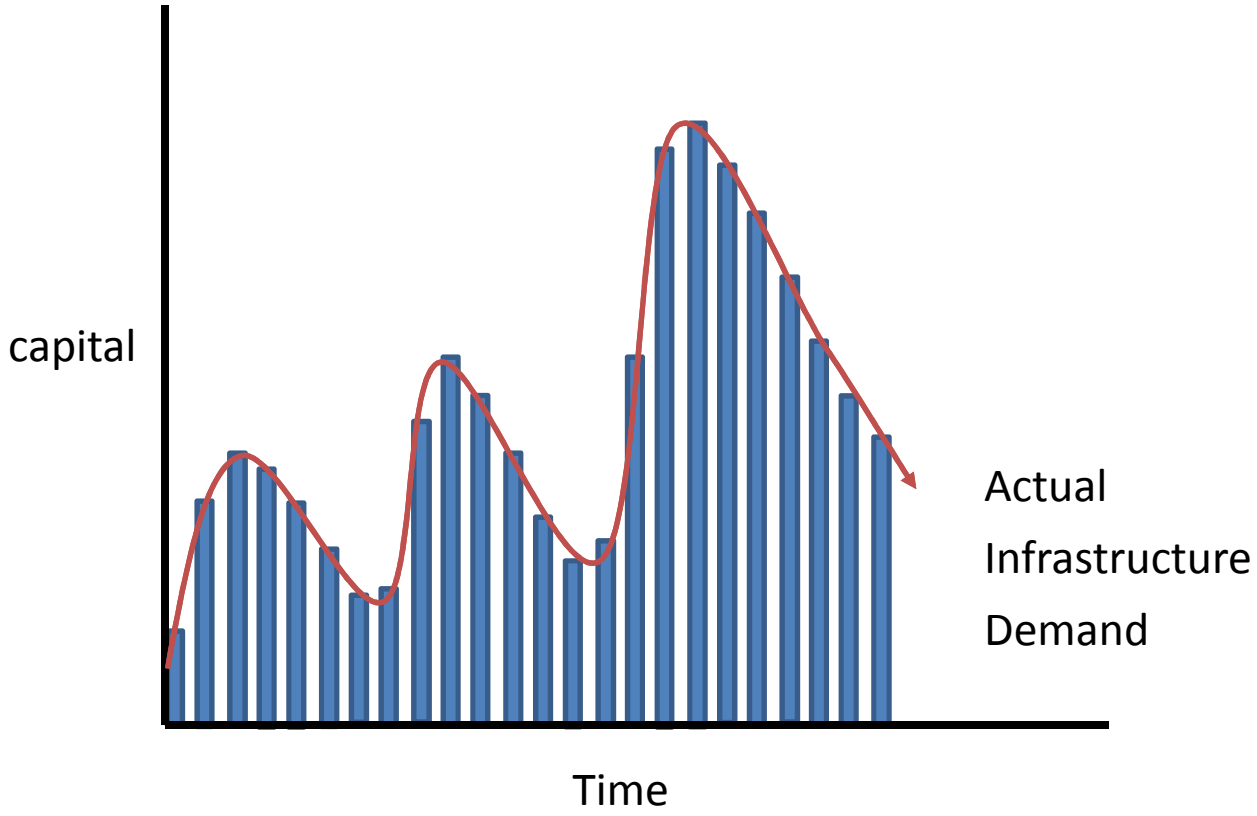
In general, a grid computing system requires:

- **At least one computer, usually a server, which handles all the administrative duties for the System**
- **A network of computers running special grid computing network software.**
- **A collection of computer software called middleware**

Traditional Infrastructure Model



Utility Infrastructure Model



Cloud computing

- Cloud computing is Internet-based computing, whereby shared resources, software, and information are provided to computers and other devices on demand, like the electricity grid.

- Cloud computing refers to anything that involves delivering hosted services over the internet and encompasses any subscription-based or pay-by-use service that, in real time, extends IT's existing capabilities.

Cloud computing

- In general, cloud computing customer avoids capital expenditure of the company thereby also reducing the cost of purchasing physical infrastructure by renting the usage from a third party Provider.

- The companies devour the resources and pay for what they use.

- Many cloud-computing offerings use the utility computing model, which is similar to traditional used services like electricity are consumed, while others bill on a subscription basis.

Simple Example

- The most common example of cloud computing is yahoo mail or Gmail. You just need an internet connection to send emails through them. There is no need of server or any software to use them they are simply available on internet.

- The major benefit of cloud computing is it has reduced the cost or capital expenditure of the companies. The company needs to pay for what it has used and upgrades are also done automatically.

Service models

- Cloud computing providers offer their services according to three fundamental models
- Software as a service
- Platform as a service
- Infrastructure as service
- Data as service

Cloud Computing – Simple Definition

Cloud Computing = Software as a Service
+ Platform as a Service

Cloud Computing – SaaS

- Software as a Service (SaaS)
  - From end user's point of view
  - Apps are located in the cloud
  - Apps are delivered through the Internet

Cloud Computing – PaaS

- Platform as a Service (PaaS)
  - From developer's point of view (i.e. cloud users)
  - Cloud providers offer an Internet-based platform to developers who want to create services but don't want to build their own cloud

Cloud Computing – Simple Definition

Cloud Computing = Software as a Service
+ Platform as a Service
+ Infrastructure as a Service

- Infrastructure as a Service (IaaS)
  - Cloud providers build datacenters
    - Power, scale, hardware, networking, storage, distributed systems, etc
  - Datacenter as a service
  - Cloud users rent storage, computation, and maintenance from cloud providers (pay-as-you-go; like utility)

Knowledge & Data Intelligence as a Service

Cloud Computing = Software as a Service
+ Platform as a Service
Infrastructure as a Service
+ Data as a Service

Data ➔ Information ➔ Knowledge ➔ Intelligence

- Infrastructure for Web-scale data mining and knowledge discovery
- Empower people with knowledge
- Empower applications and services with intelligence

Benefits

- Reduce capital and operations costs
  - No longer required to make large up-front capital investment on datacenters
  - Eliminate the need to plan ahead for provisioning
  - Allow companies to start small and increase their resources investment as needed (pay-as-you-go)
- Simplify app deployment & management
  - Common programming model across mobile, browser, client, server, cloud
  - Access to strong ecosystem of widely deployed applications
  - Integration with existing IT assets (Software + Services)

Difference between cluster and grid computing

- When two or more computers are used together to solve a problem, it is called a computer cluster . Basically it is just cooperation between computers in order to solve a task or a problem.
- Grid computing is something similar to cluster computing, it makes use of several computers connected is some way, to solve a large problem.
- The big difference is that a cluster is homogenous while grids are heterogeneous.
- The computers that are part of a grid can run different operating systems and have different hardware whereas the cluster computers all have the same hardware and OS.

Difference between cluster and grid

- while the machines in a cluster are dedicated to work as a single unit and nothing else. Grid are inherently distributed by its nature over a LAN, metropolitan or WAN. On the other hand, the computers in the cluster are normally contained in a single location or complex.

- Another difference lies in the way resources are handled. In case of Cluster, the whole system (all nodes) behave like a single system view and resources are managed by centralized resource manager. In case of Grid, every node is autonomous i.e. it has its own resource manager and behaves like an independent entity

Difference between cluster computing and grid computing

CHARACTERISTIC	**CLUSTER**	**GRID**
Population	Commodity Computers	Commodity and High-end computers
Ownership	Single	Multiple
Discovery	Membership Services	Centralized Index and Decentralized Info
User Management	Centralized	Decentralized
Resource management	Centralized	Distributed
Allocation/ Scheduling	Centralized	Decentralized
Inter-Operability	VIA and Proprietary	No standards being developed
Single System Image	Yes	No
Scalability	100s	1000?
Capacity	Guaranteed	Varies, but high
Throughput	Medium	High
Speed(Lat. Bandwidth)	Low, high	High, Low