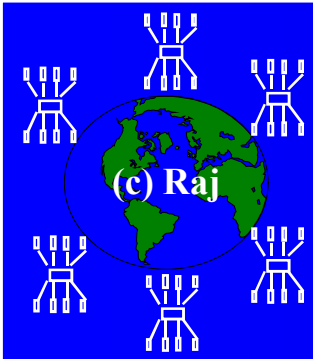
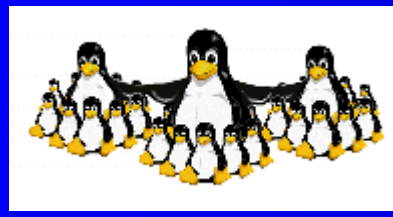
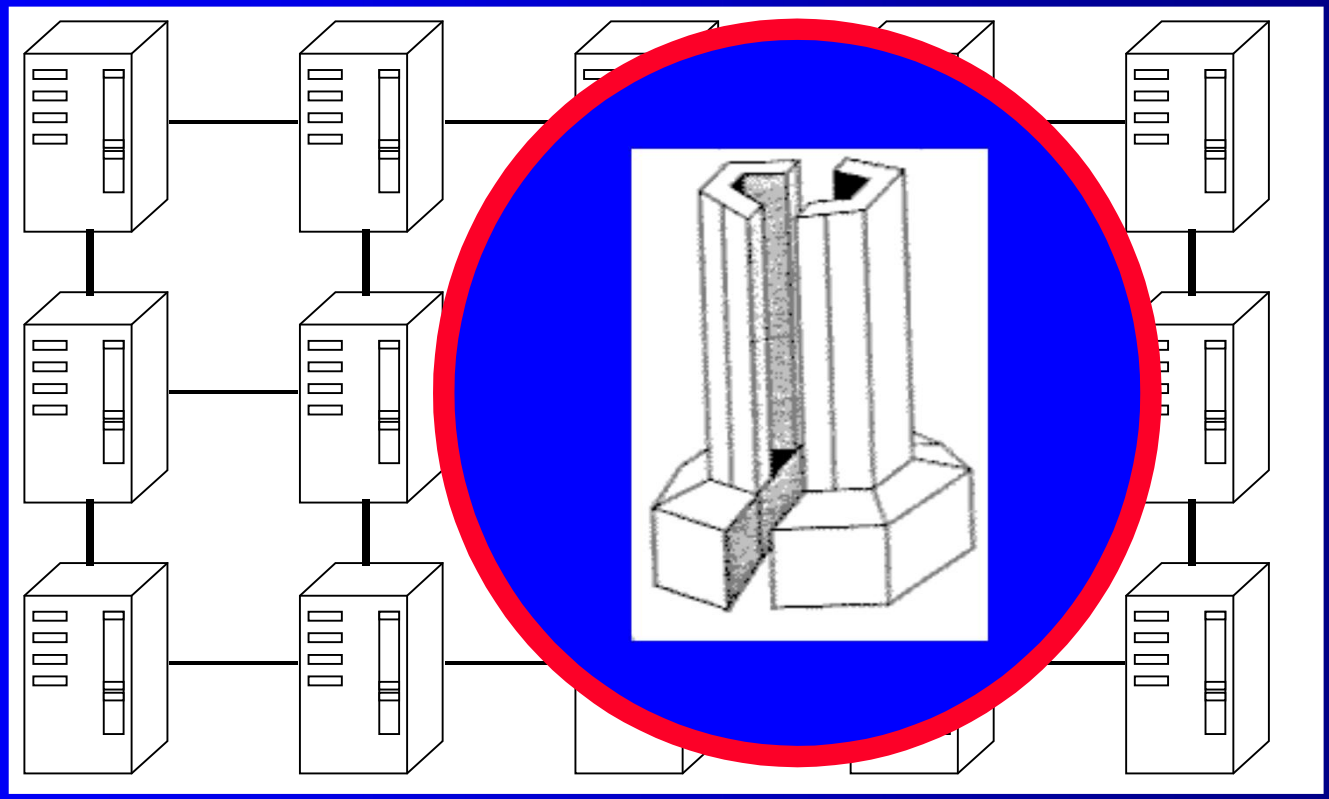




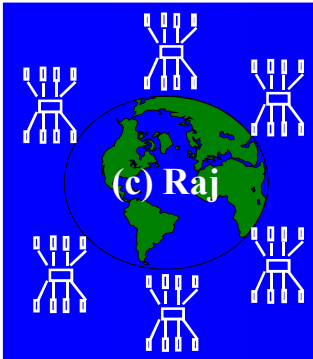
No Low Cost Supercomputing

Parallel Processing on Linux Clusters



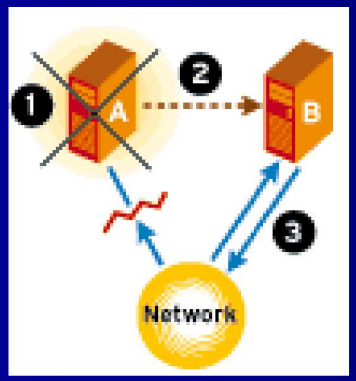
Rajkumar Buyya, Monash University, Melbourne, Australia.

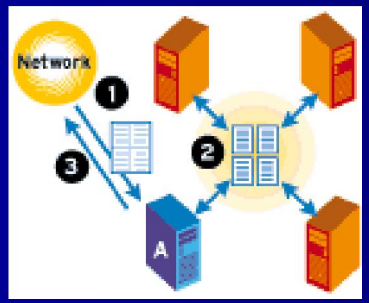
rajkumar@ieee.org http://www.dgs.monash.edu.au/~rajkumar

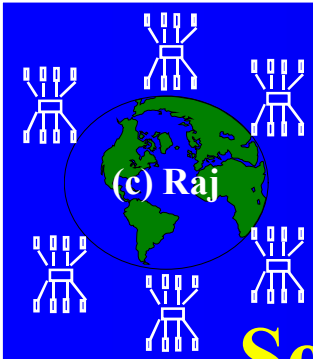


Agenda

- Cluster ? Enabling Tech. & Motivations
- Cluster Architecture
- Cluster Components and Linux
- Parallel Processing Tools on Linux
- Cluster Facts
- Resources and Conclusions

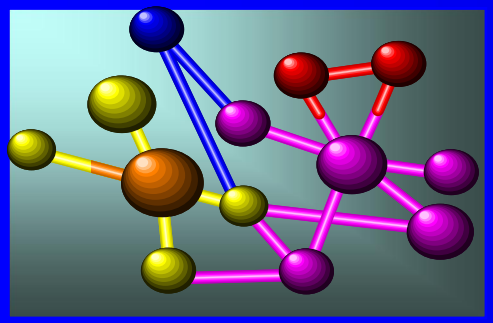




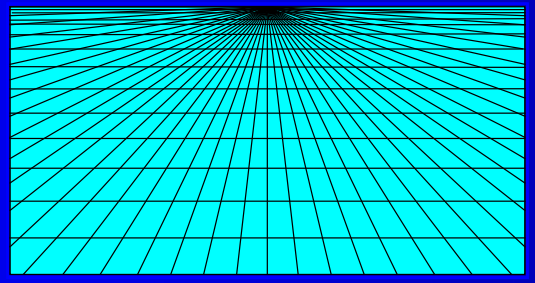


Need of more Computing Power: Grand Challenge Applications

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

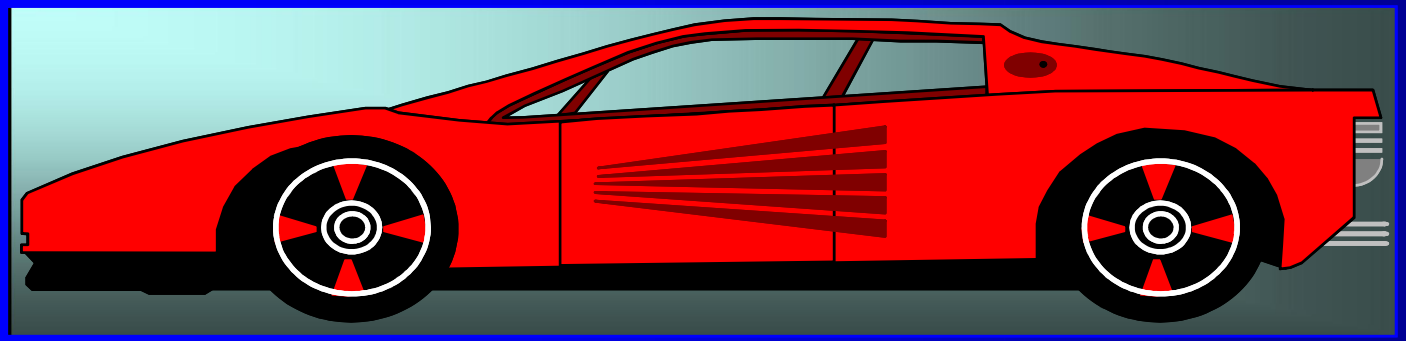


Life Sciences

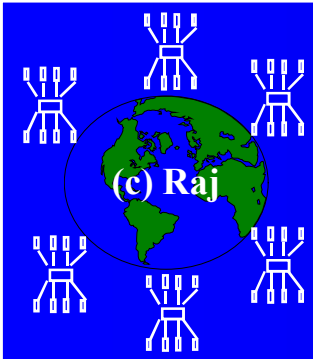


Aerospace

Geographic Information Systems



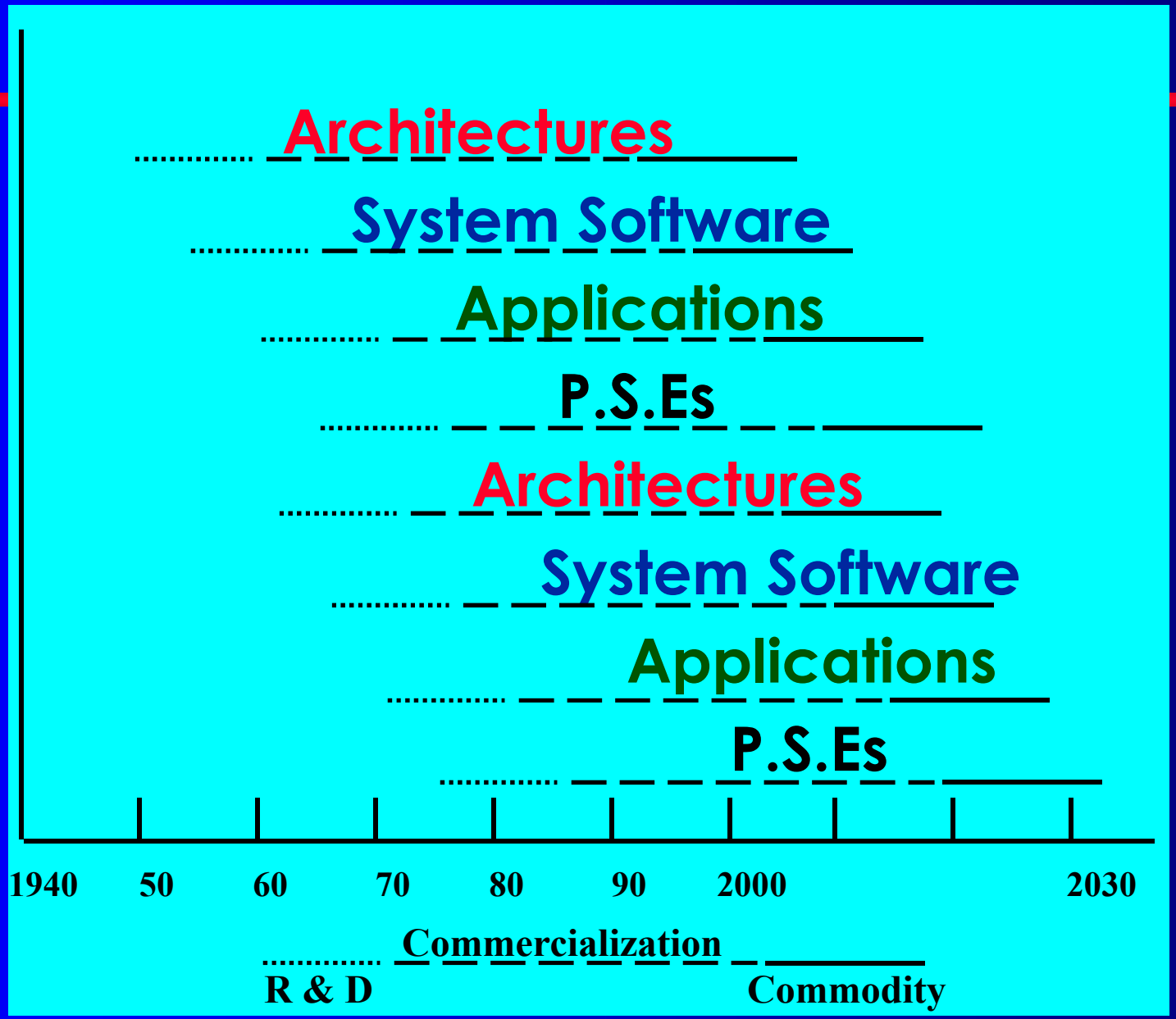
Mechanical Design & Analysis (CAD/CAM)

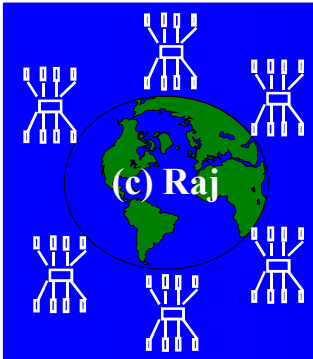


Two Eras of Computing

Sequential Era

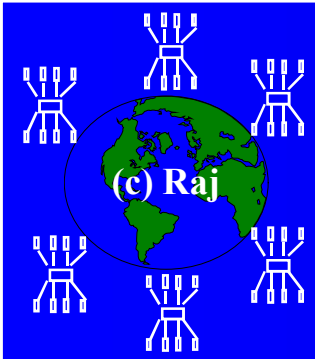
Parallel Era





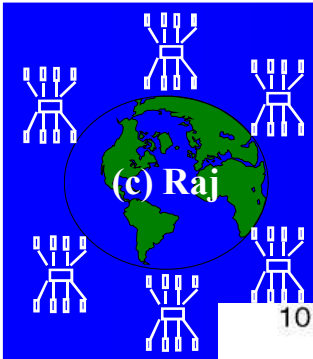
Competing Computer Architectures

- **Vector Computers (VC) ---proprietary system**
  - provided the breakthrough needed for the emergence of computational science, buy they were only a partial answer.
- **Massively Parallel Processors (MPP)-proprietary system**
  - high cost and a low performance/price ratio.
- **Symmetric Multiprocessors (SMP)**
  - suffers from scalability
- **Distributed Systems**
  - difficult to use and hard to extract parallel performance.
- **Clusters -- gaining popularity**
  - High Performance Computing---Commodity Supercomputing
  - High Availability Computing ---Mission Critical Applications

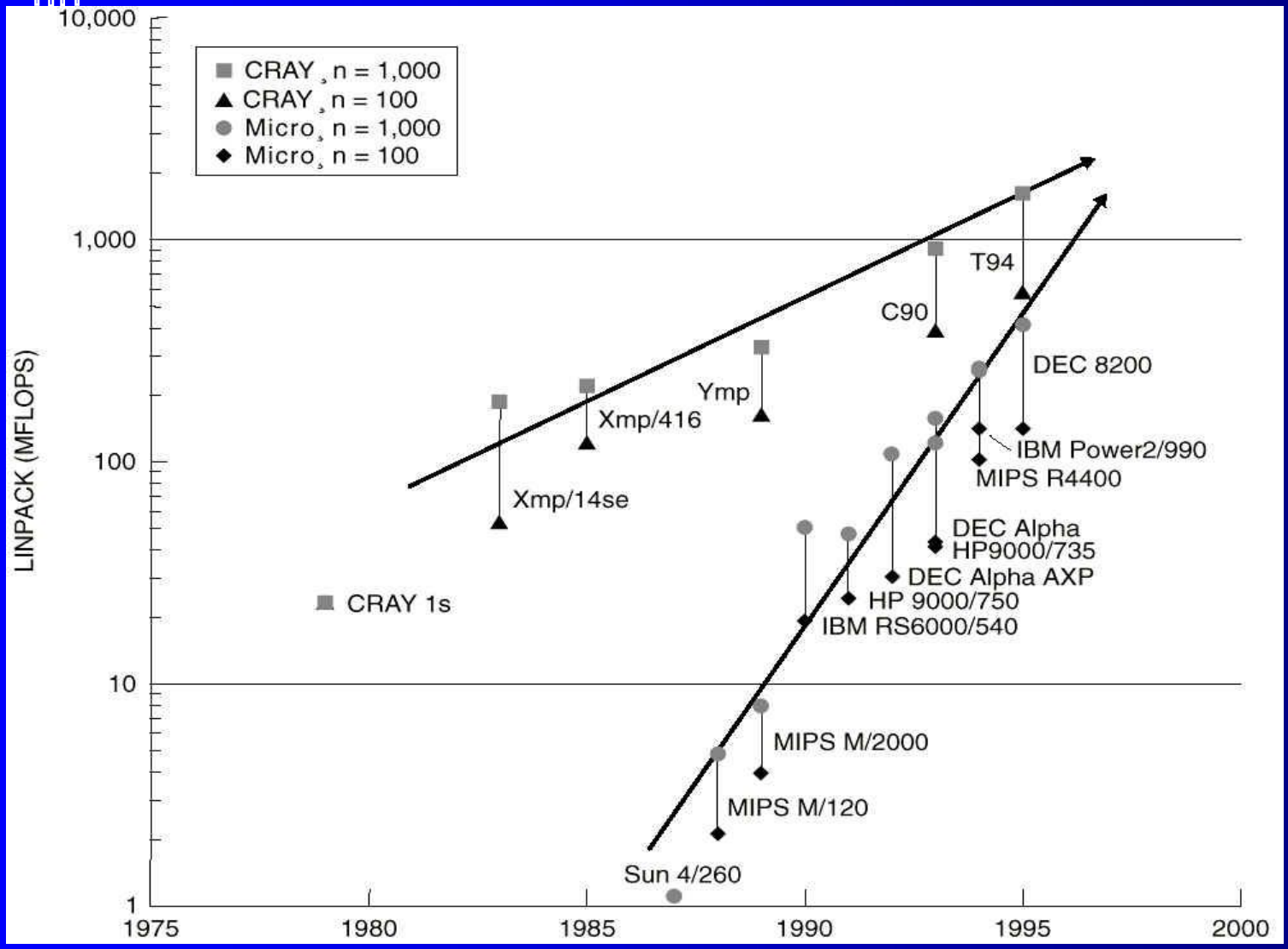


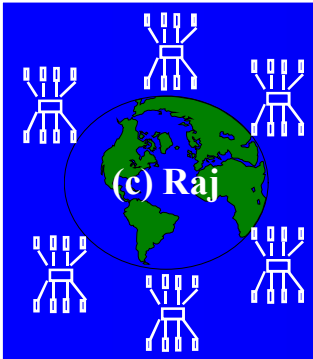
Technology Trend...

- **Performance of PC/Workstations components has almost reached performance of those used in supercomputers…**
  - Microprocessors (50% to 100% per year)
  - Networks (Gigabit ..)
  - Operating Systems
  - Programming environment
  - Applications
- **Rate of performance improvements of commodity components is too high.**



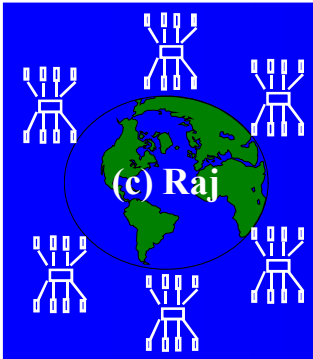
Technology Trend





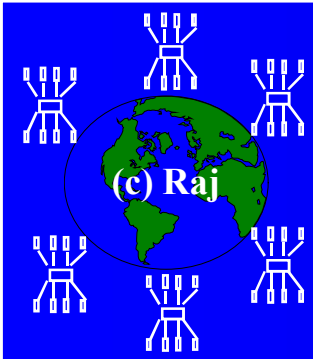
The Need for Alternative Supercomputing Resources

- **Cannot afford to buy "Big Iron" machines**
  - due to their high cost and short life span.
  - cut-down of funding
  - don't "fit" better into today's funding model.
  - ….
- **Paradox: time required to develop a parallel application for solving GCA is equal to:**
  - half Life of Parallel Supercomputers.



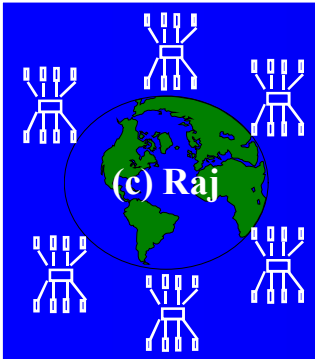
Clusters are best-alternative!

- Supercomputing-class commodity components are available
- They "fit" very well with today's/future funding model.
- Can leverage upon future technological advances
  - VLSI, CPUs, Networks, Disk, Memory, Cache, OS, programming tools, applications,...



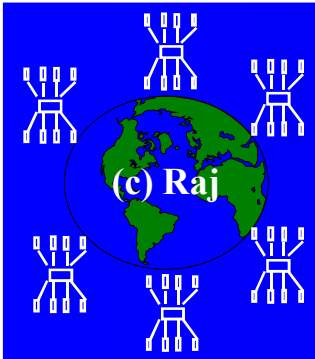
Best of both Worlds!

- **High Performance Computing (talk focused on this)**
  - parallel computers/supercomputer-class workstation cluster
  - dependable parallel computers
- **High Availability Computing**
  - mission-critical systems
  - fault-tolerant computing



What is a cluster?

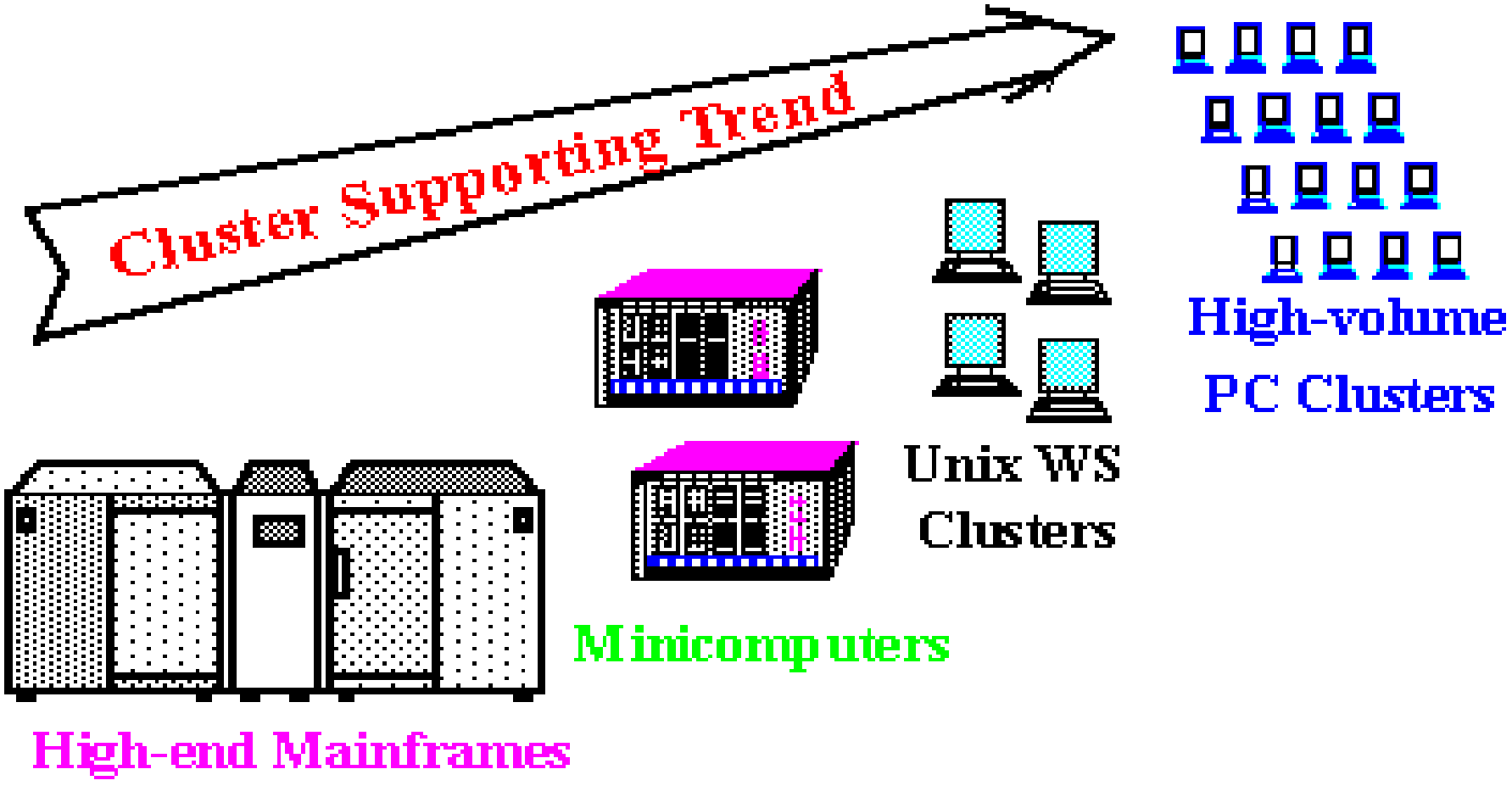
- A cluster is a type of parallel or distributed processing system, which consists of a collection of interconnected stand-alone computers cooperatively working together as a single, integrated computing resource.
- A typical cluster:
  - Network: Faster, closer connection than a typical network (LAN)
  - Low latency communication protocols
  - Looser connection than SMP

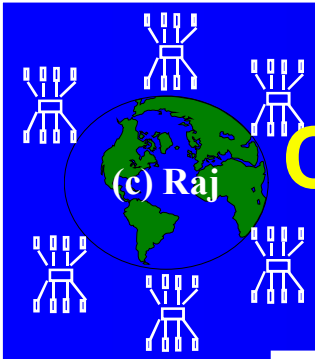


So What's So Different about Clusters?

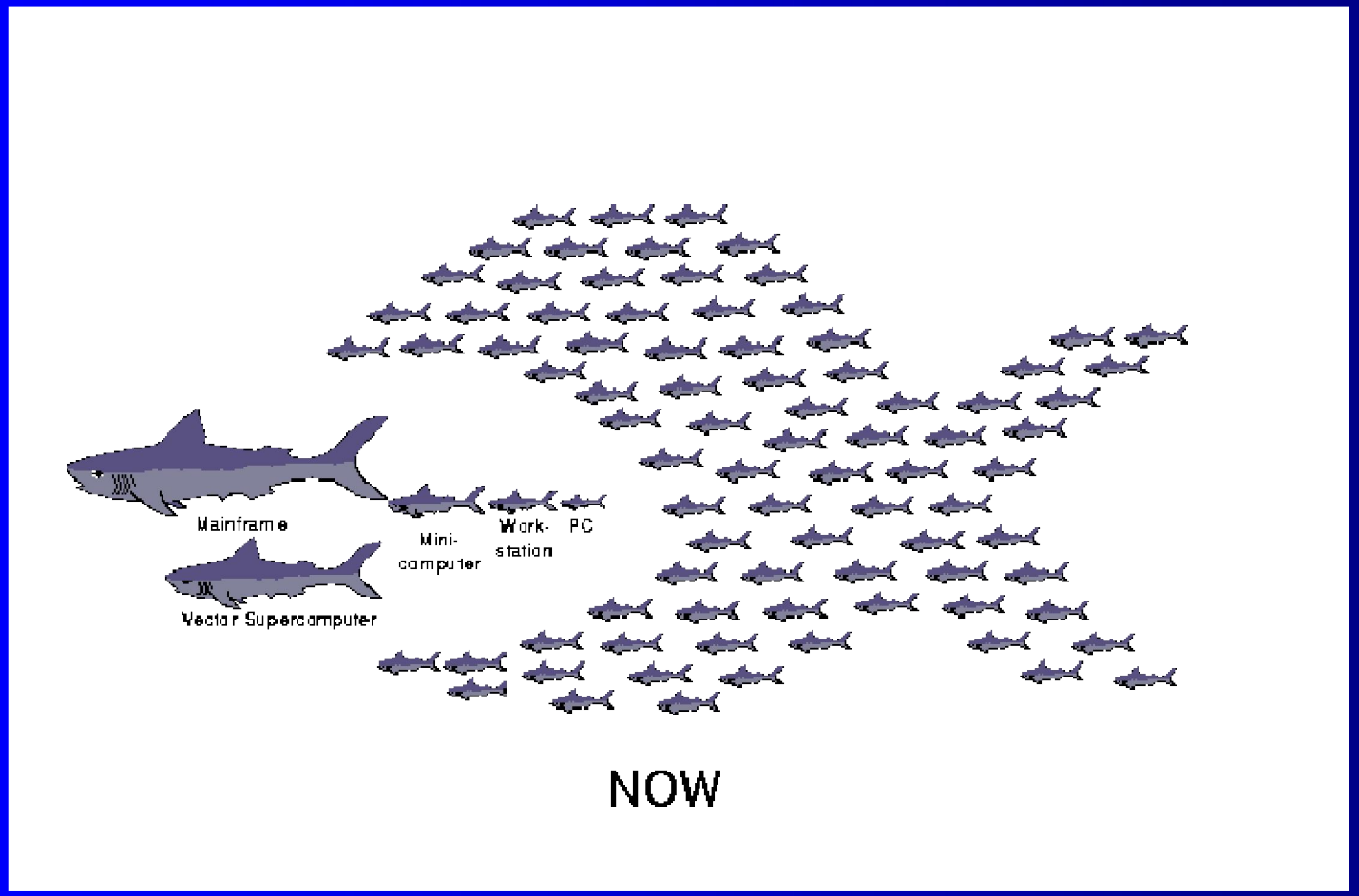
- Commodity Parts?
- Communications Packaging?
- Incremental Scalability?
- Independent Failure?
- Intelligent Network Interfaces?
- Complete System on every node
  - virtual memory
  - scheduler
  - files
  - …
- Nodes can be used individually or combined…

Clustering of Computers for Collective Computating

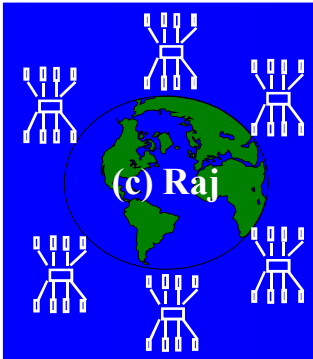




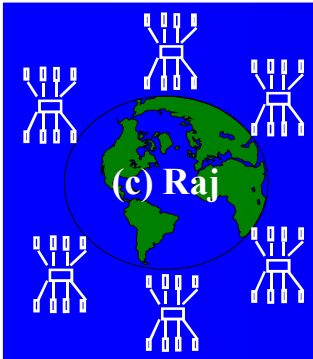
Computer Food Chain (Now and Future)



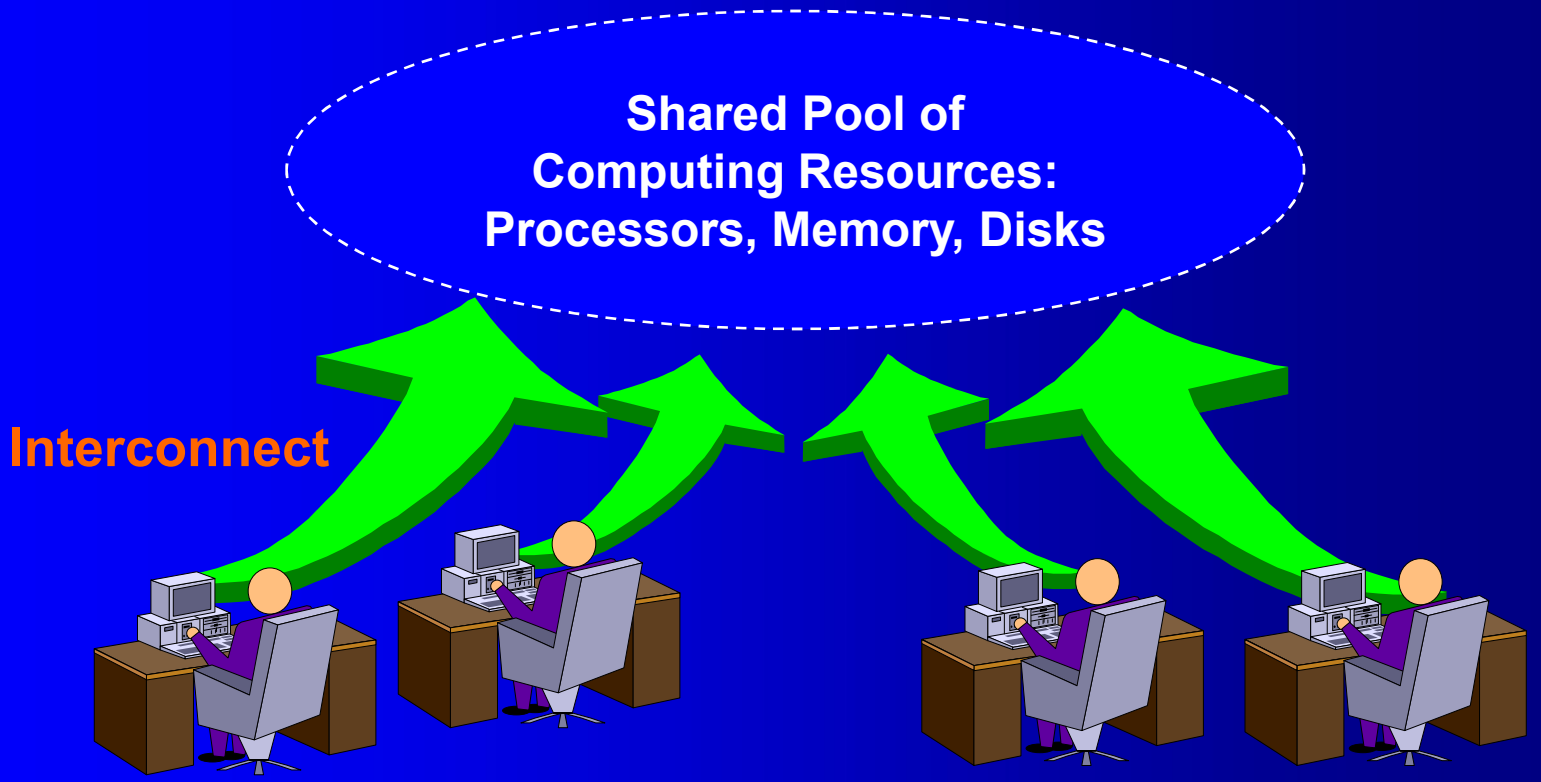
Demise of Mainframes, Supercomputers, & MPPs



Cluster Configuration..1
Dedicated Cluster

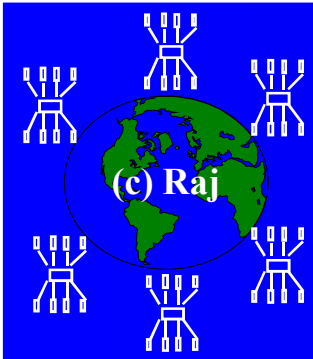


Cluster Configuration..2
Enterprise Clusters (use JMS like Codine)



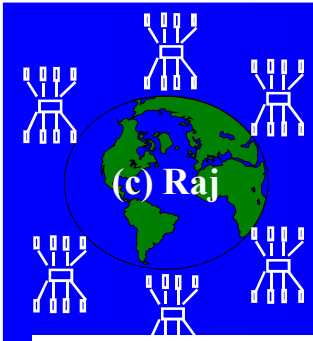
Guarantee at least one workstation to many individuals (when active)

Deliver large % of collective resources to few individuals at any one time

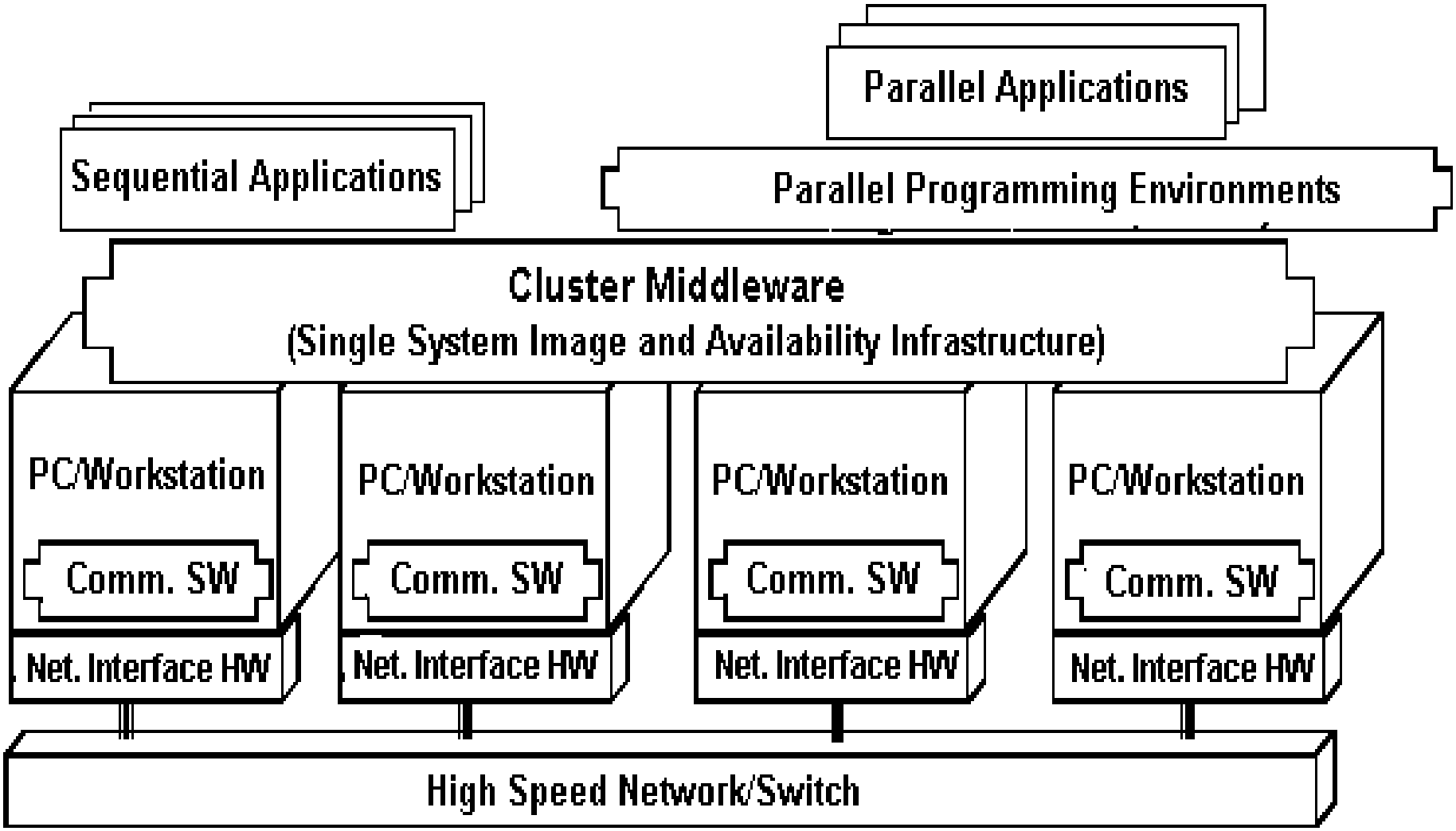


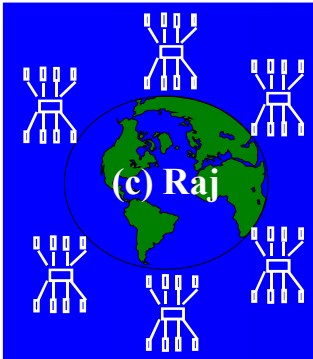
Windows of Opportunities

- **MPP/DSM:**
  - Compute across multiple systems: parallel.
- **Network RAM:**
  - Idle memory in other nodes. Page across other nodes idle memory
- **Software RAID:**
  - file system supporting parallel I/O and reliability, mass-storage.
- **Multi-path Communication:**
  - Communicate across multiple networks: Ethernet, ATM, Myrinet

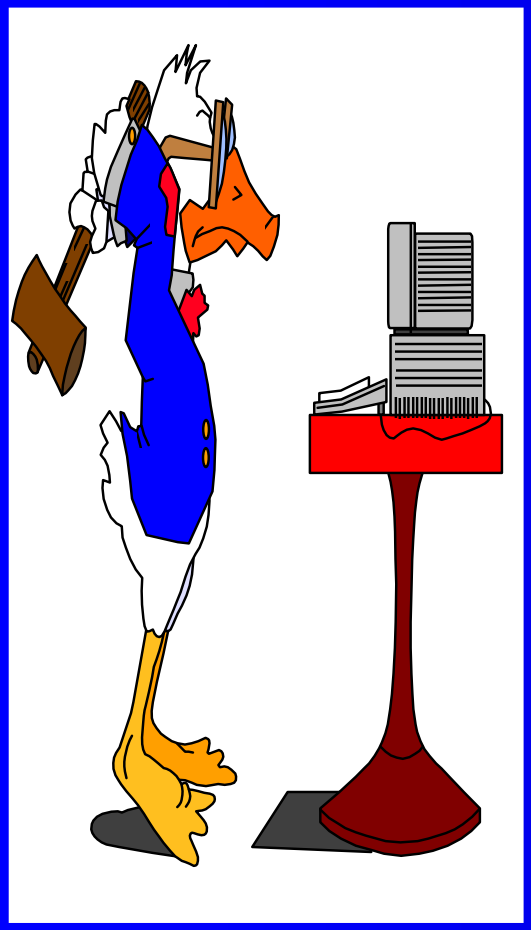


Cluster Computer Architecture

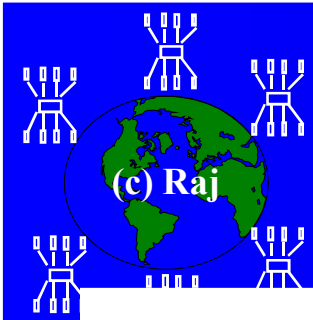




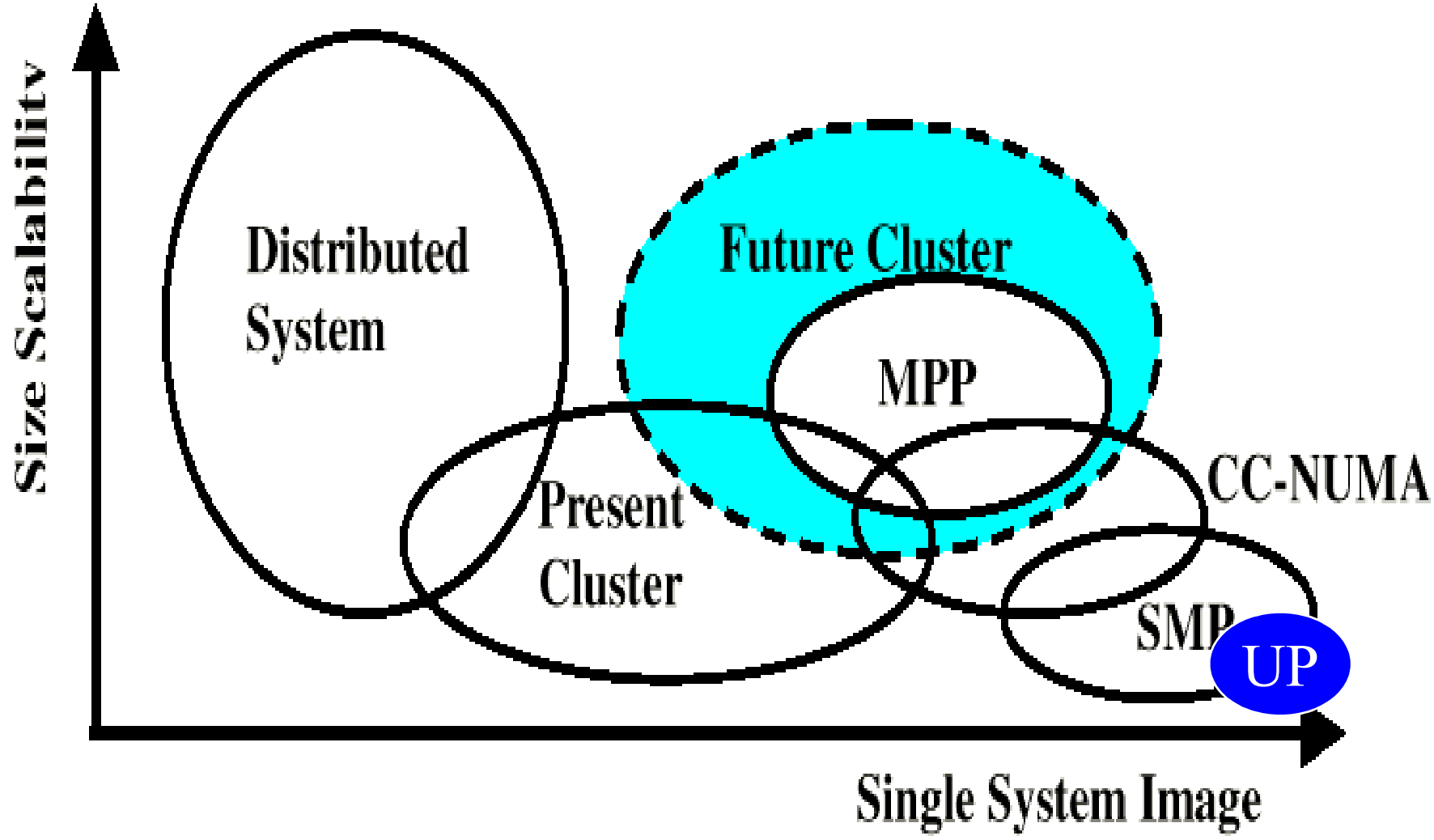
Major issues in cluster design

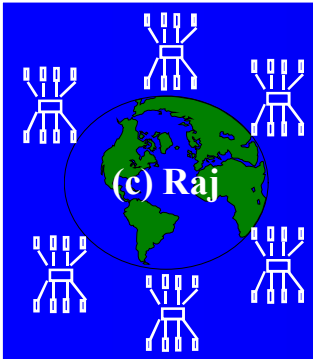


- Size Scalability (physical & application)
- Enhanced Availability (failure management)
- Single System Image (look-and-feel of one system)
- Fast Communication (networks & protocols)
- Load Balancing (CPU, Net, Memory, Disk)
- Security and Encryption (clusters of clusters)
- Distributed Environment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



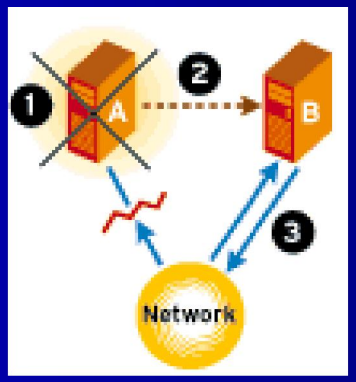
Scalability Vs. Single System Image



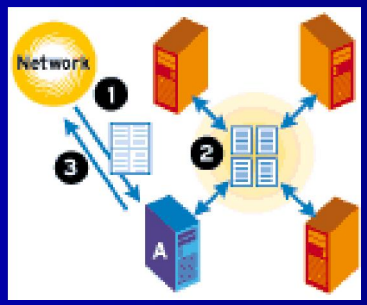


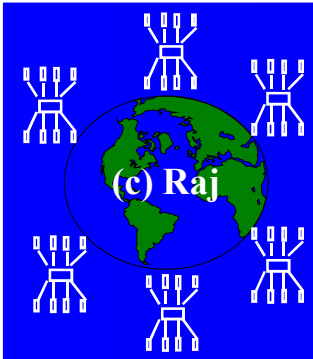
Linux-based Tools for

High Availability Computing



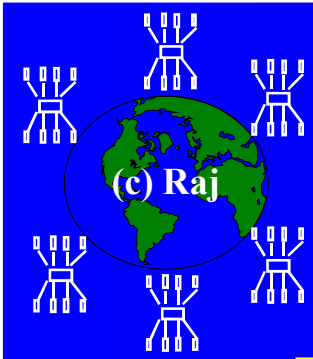
High Performance Computing





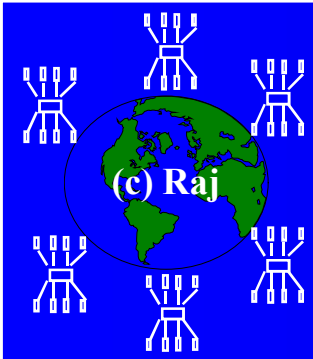
Hardware

- **Linux OS is running/driving...**
  - PCs (Intel x86 processors)
  - Workstations (Digital Alphas)
  - SMPs (CLUMPS)
  - Clusters of Clusters
- **Linux supports networking with**
  - Ethernet (10Mbps)/Fast Ethernet (100Mbps),
  - Gigabit Ethernet (1Gbps)
  - SCI (Dolphin - MPI- 12micro-sec latency)
  - ATM
  - Myrinet (1.2Gbps)
  - Digital Memory Channel
  - FDDI



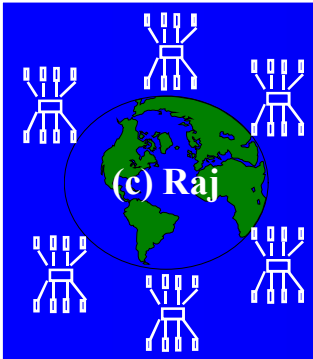
Communication Software

- **Traditional OS supported facilities (heavy weight due to protocol processing)..**
  - Sockets (TCP/IP), Pipes, etc.
- **Light weight protocols (User Level)**
  - Active Messages (AM) (Berkeley)
  - Fast Messages (Illinois)
  - U-net (Cornell)
  - XTP (Virginia)
  - Virtual Interface Architecture (industry standard)



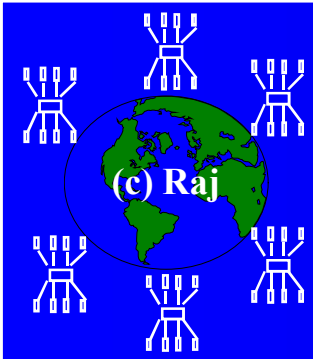
Cluster Middleware

- Resides Between OS and Applications and offers in infrastructure for supporting:
  - Single System Image (SSI)
  - System Availability (SA)
- SSI makes collection appear as single machine (globalised view of system resources). telnet cluster.myinstitute.edu
- SA - Check pointing and process



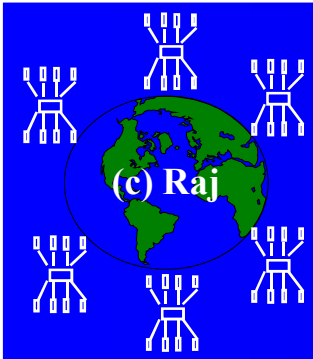
Cluster Middleware

- **OS / Gluing Layers**
  - Solaris MC, Unixware, **MOSIX**
  - **Beowulf "Distributed PID"**
- **Runtime Systems**
  - Runtime systems (software DSM, PFS, etc.)
  - Resource management and scheduling (RMS):
    - CODINE, CONDOR, LSF, PBS, NQS, etc.



Programming environments

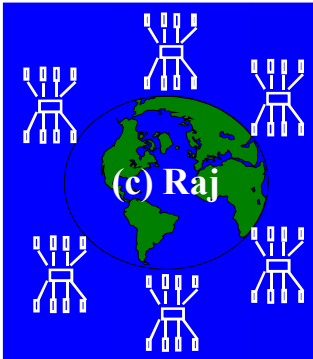
- **Threads (PCs, SMPs, NOW..)**
  - POSIX Threads
  - Java Threads
- **MPI**
  - http://www-unix.mcs.anl.gov/mpi/mpich/
- **PVM**
  - http://www.epm.ornl.gov/pvm/
- **Software DSMs (Shmem)**



Development Tools

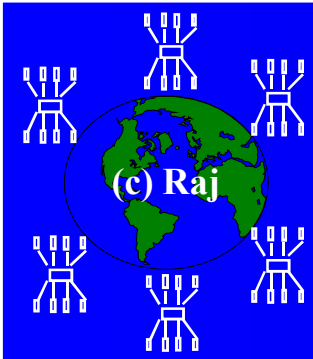
GNU-- www.gnu.org

- Compilers
  - C/C++/Java/
- Debuggers
- Performance Analysis Tools
- Visualization Tools

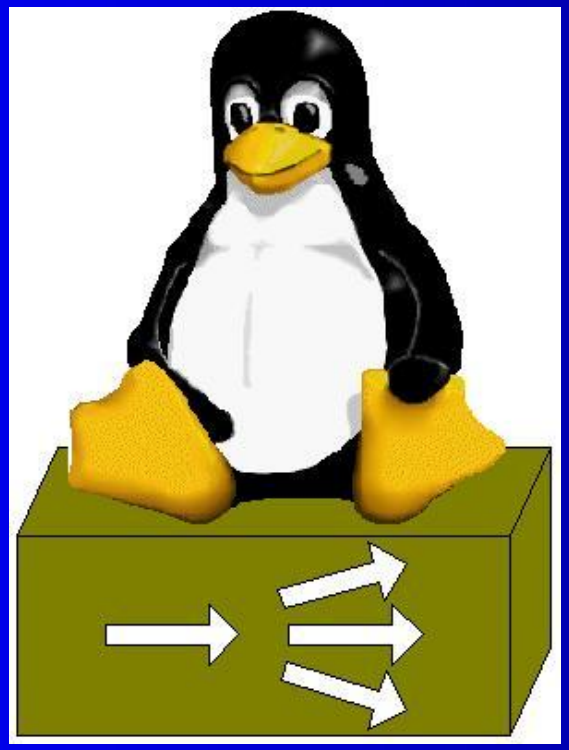


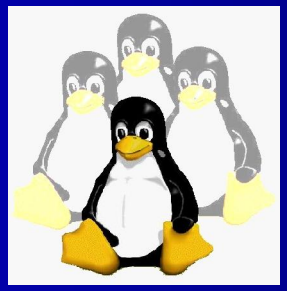
Applications

- **Sequential (benefit from the cluster)**
- **Parallel / Distributed (Cluster-aware app.)**
  - Grand Challenging applications
    - Weather Forecasting
    - Quantum Chemistry
    - Molecular Biology Modeling
    - Engineering Analysis (CAD/CAM)
    - Ocean Modeling
    - ……
  - PDBs, web servers,data-mining



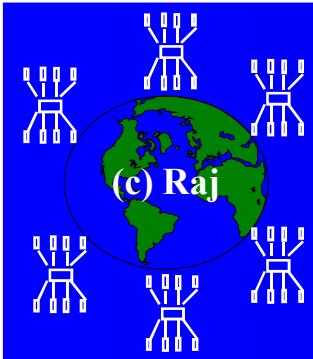
Linux Webserver (Network Load Balancing)





http://proxy.iinchina.net/~wensong/ippfvs/

- High Performance (by serving through light loaded machine)
- High Availability (detecting failed nodes and isolating them from the cluster)
- Transparent/Single System view



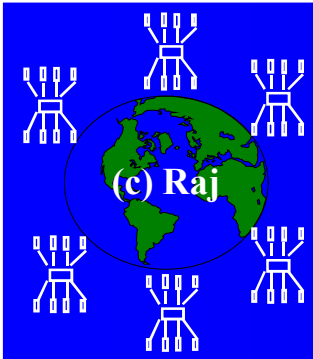
A typical Cluster Computing Environment

Application

PVM / MPI/ RSH

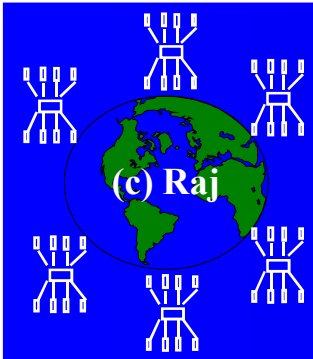
???

Hardware/OS



CC should support

- Multi-user, time-sharing **environments**
- Nodes with different CPU **speeds and** memory sizes (heterogeneous configuration)
- Many processes, **with** unpredictable **requirements**
- Unlike SMP: **insufficient** "bonds" **between nodes**
  - Each computer operates independently

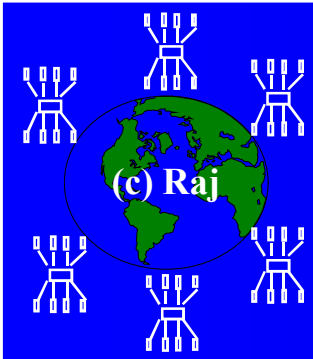


(MOSIX)

http://www.mosix.cs.huji.ac.il/

- An OS module (layer) that provides the applications with the illusion of working on a single system
- Remote operations are performed like local operations
- Transpar er interface unchan

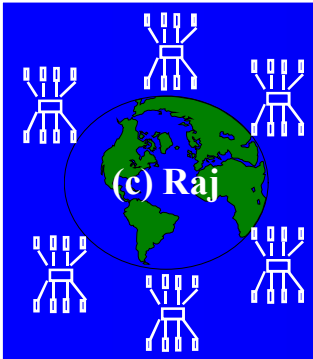
- Offers missing link



MOSIX is Main tool

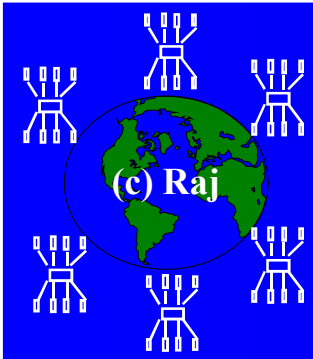
Preemptive process migration that can migrate--->any process, anywhere, anytime

- Supervised by distributed algorithms **that respond on-line to global resource availability -** transparently
- Load-balancing **- migrate process from over-loaded to under-loaded nodes**
- Memory ushering **- migrate processes from a node that has exhausted its memory, to prevent paging/swapping**



MOSIX for Linux at HUJI

- **A scalable cluster configuration:**
  - 50 Pentium-II 300 MHz
  - 38 Pentium-Pro 200 MHz (some are SMPs)
  - 16 Pentium-II 400 MHz (some are SMPs)
- **Over 12 GB cluster-wide RAM**
- **Connected by the Myrinet 2.56 G.b/s LAN**
  Runs Red-Hat 6.0, **based on** Kernel 2.2.7
- **Upgrade: HW with Intel, SW with Linux**
- **Download MOSIX:**
  - http://www.mosix.cs.huji.ac.il/



Nimrod - A tool for parametric modeling on clusters

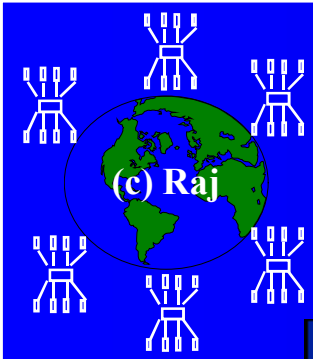
CRC For Distributed Systems Technology

Nimrod: A Computational Workbench

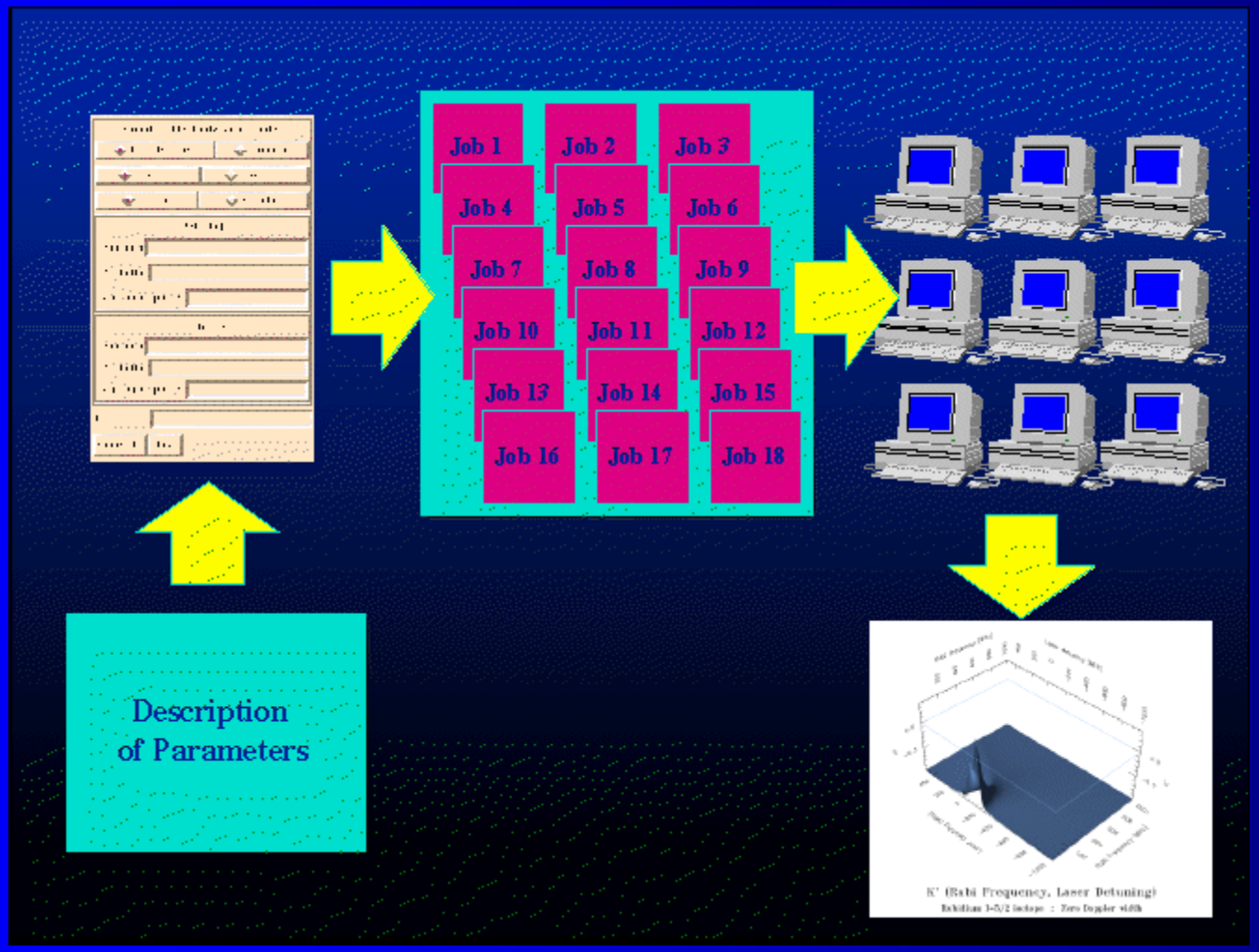
- High Level Abstraction for Computational Modellers
- Little or no programming
- Ease of use
- Use of Distributed Computational Resource
- Heterogeneous platforms

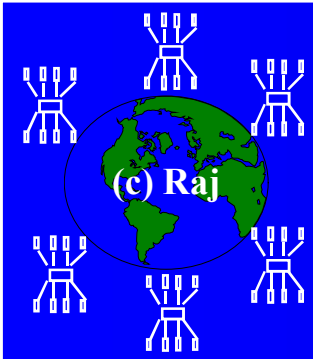
##

♍http://www.dgs.monash.edu.au/~davida/nimrod.html



Job processing with Nimrod

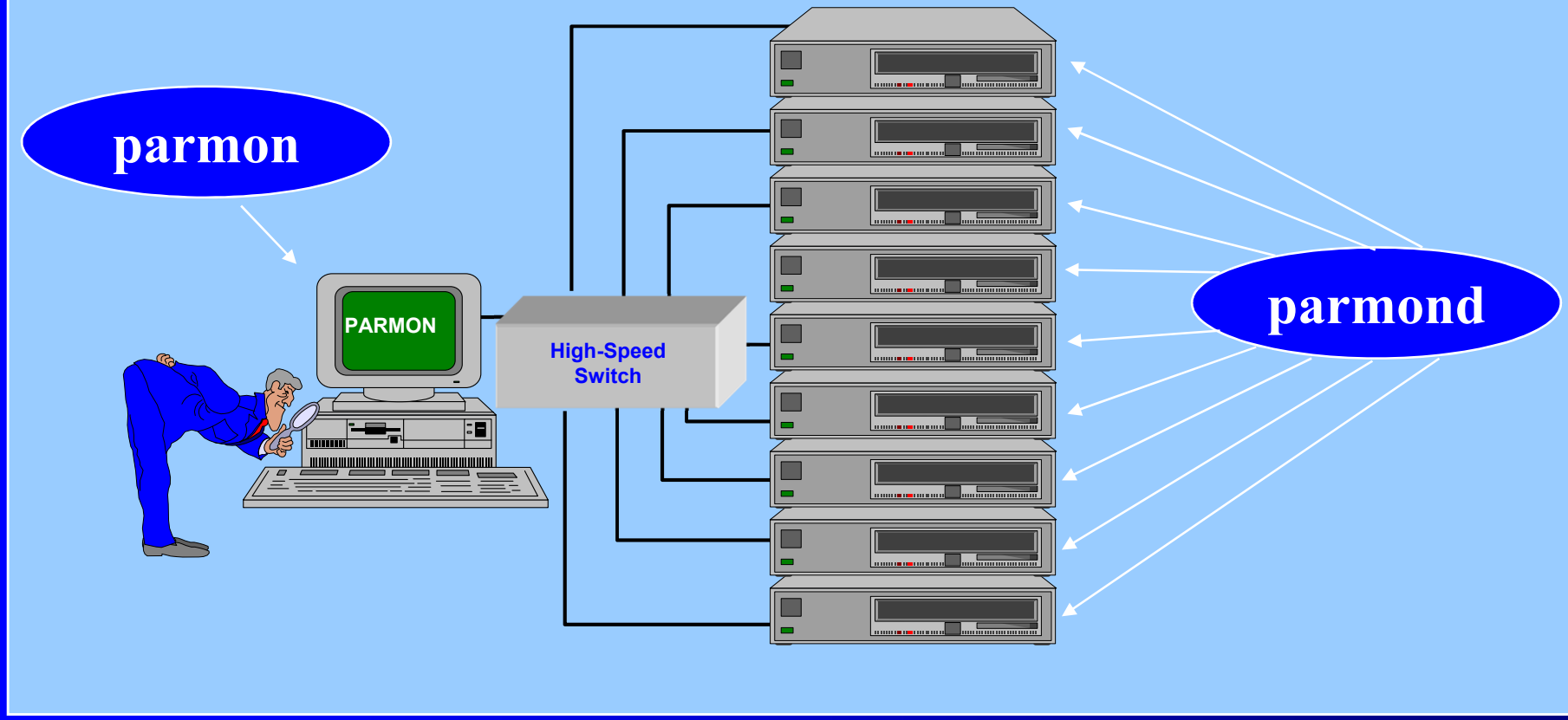


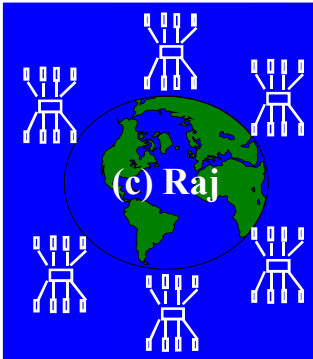


PARMON: A Cluster Monitoring Tool

PARMON Client on JVM

PARMON Server on each node





Resource Utilization at a Glance

PARMON: Param OpenFrame Monitoring System

File Monitor View Setup Help

CPU Usage
Disk Usage
Memory Usage
Network Usage

Aggregat

e Viewer

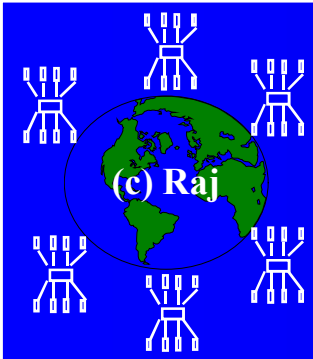
Process Viewer

Kernel Data Catalog

Physical View

Logical View

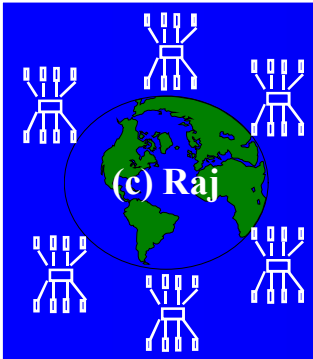
Online Diagnostics



Linux cluster in Top500

Top500 Supercomputing (www.top500.org) Sites declared Avalon(http://cnls.lanl.gov/avalon/), Beowulf cluster, the 113th most powerful computer in the world.

- 70 processor DEC Alpha cluster
- Cost: $152K
- Completely commodity and Free Software
- price/performance is $15/Mflop,
- performance similar to 1993's 1024-node CM-5



Adoption of the Approach

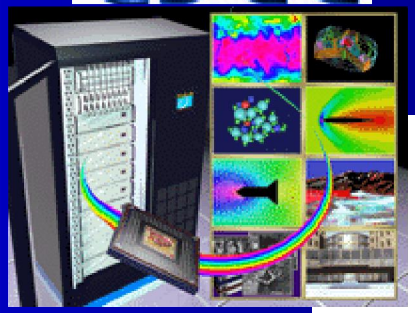
HEWLETT PACKARD
Expanding Possibilities

HP NetServer Cluster Solutions

HOTBOT

CDCC

COMPAQ

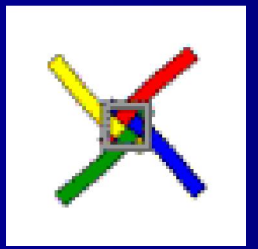


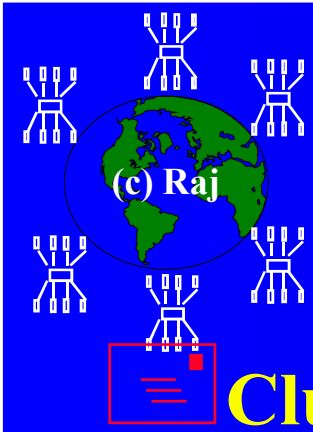
IBM

Sun microsystems

TANDEM

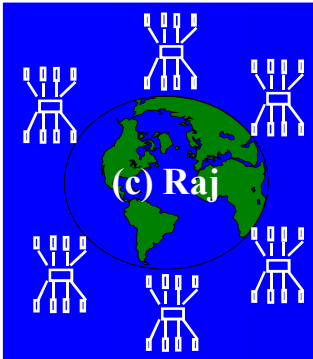
hotmail™





Conclusions Remarks

- **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..so that
  - Clusters based supercomputers (Linux based clusters) can be seen everywhere!

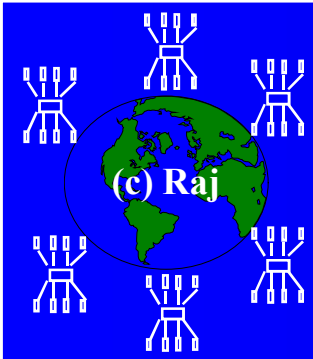


Announcement: formation of

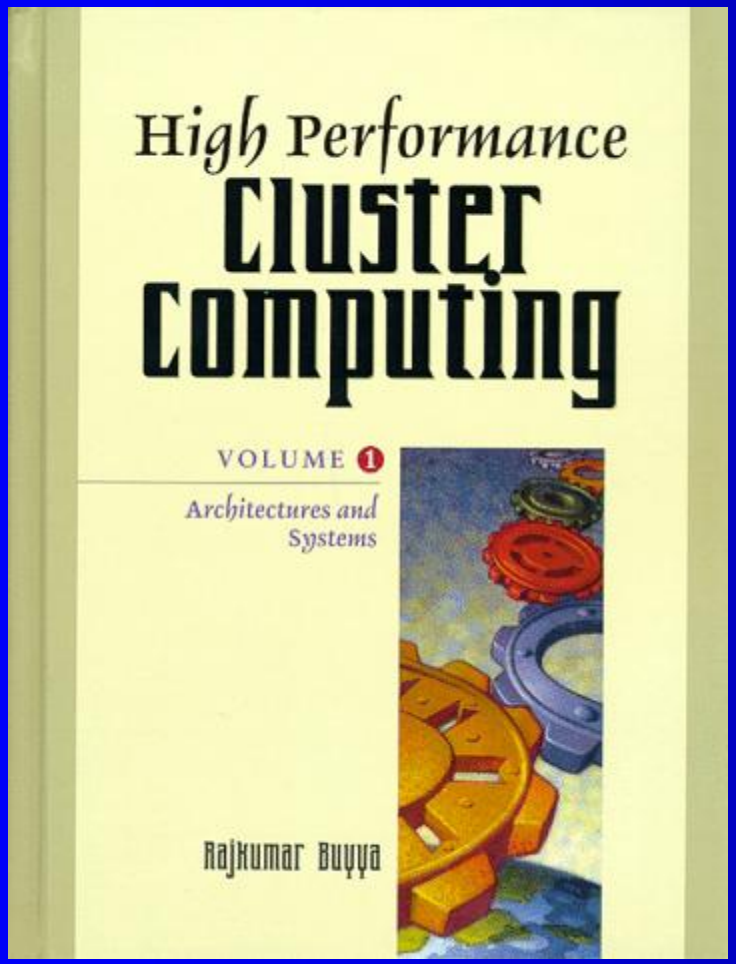
IEEE Task Force on Cluster Computing (TFCC)

http://www.dgs.monash.edu.au/~rajkumar/tfcc/

http://www.dcs.port.ac.uk/~mab/tfcc/



Well, Read my book for….



Thank You …

http://www.dgs.monash.edu.au/~rajkumar/cluster/