Module 5
Advanced Topics in Cloud Computing

Energy Efficiency in Clouds, Market Based Management of Clouds

Data Center Power Consumption

- Data Center energy demand doubles every 5-6 years.
- This results in large amounts of CO_2 produced by burning fossil fuels.
- Need to reduce the energy used with minimal performance impact.

Ref: Efficient Resource Management for Cloud Computing Environments, by Andrew J. Younge, Gregor von Laszewski, Lizhe Wang, Sonia Lopez-Alarcon, Warren Carithers

Datacenters are new Polluters

- U.S. data centers are using more electricity than they need. It takes 34 power plants, each capable of generating 500 megawatts of electricity, to power all the data centers in operation then.(2017)

Datacenters are new Polluters

- By 2020, the nation **needed another 17 similarly sized power plants to meet projected data center energy demands as economic activity becomes increasingly digital.**

Datacenters are new Polluters

- Any increase in the use of fossil fuels to generate electricity will result in an increase in carbon emissions.

Fossil Fuels?

What are Fossil Fuels?

- Formed over millions of years from the fossils, or remains, of dead animals and plants that were buried under dirt and rock.

- Heat from inside the earth and pressure from dirt and rock changes these fossils into oil, natural gas and coal.

What are Fossil Fuels?

There are three major forms of fossil fuels:

- coal
- oil
- natural gas.

All three were formed many millions of years ago during the time of the dinosaurs.

coal

crude oil

natural gas

Why are Fossil Fuels IMP?

Why are Fossil Fuels IMP?

- Because it takes
  - millions of years to make or "renew" more fossil fuels, we call them "nonrenewable fuels."
  - We are currently using the fuels that were made more than 65 million years ago.

- **Once this fuel is gone, it is gone.**

Typical Data Center Energy Consumption

The picture can't be displayed.

Ref: Dzmitry Kliazovich, University of Luxembourg

Impact of Cloud DC on Environment

- Data centers are not only expensive to maintain, but also unfriendly to the environment.
- <u>Carbon emission due to Data Centers worldwide are now more than both Argentina and the Netherlands emission.</u>

Impact of Cloud DC on Environment

- High energy costs and huge carbon footprints are incurred
  - due to the massive amount of electricity needed to power and
  - cool the numerous servers hosted in these data centers.

Impact of Cloud DC on Environment

- According to the McKinsey report on “Revolutionizing Data Center Energy Efficiency”, a typical datacenter consumes as much energy as 25,000 households.

Warehouse-Scale Data-Center Design

- Data-Center Construction Requirements
- Cooling System of a Data-Center Room

Data-Center Construction Requirements

- Most data centers are built with commercially available components.
- An off-the-shelf server consists of :
  - a number of processor sockets,
  - each with a multicore CPU and
  - its internal cache hierarchy,
  - local shared and coherent DRAM,
  - a number of directly attached disk drives.

Off the shelf-not designed or made to order but taken from existing stock or supplies.

Data-Center Construction Requirements

- The DRAM and disk resources within the rack are accessible
  - **<u>through first-level rack switches and</u>**
  - **<u>all resources in all racks are accessible via a cluster-level switch</u>**
- Consider a data center
  - built with 2,000 servers,
  - each with 8 GB of DRAM and
  - four 1 TB disk drives.
  - **<u>Each group of 40 servers is connected through a 1 Gbps link to a rack-level switch</u>**
  - **<u>Each Rack switch has an additional eight 1 Gbps ports used for connecting the rack to the cluster-level switch.</u>**

Cooling System of a Data-Center Room

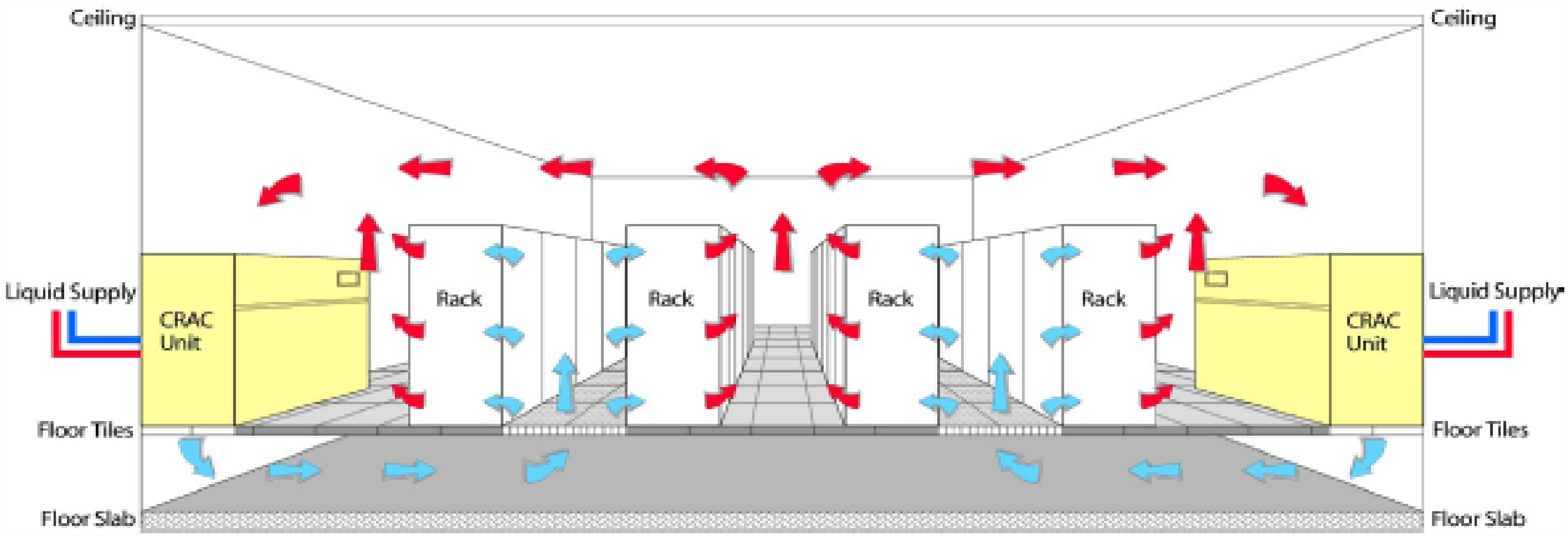


FIGURE 4.2: Datacenter raised floor with hot–cold aisle setup (image courtesy of DLB Associates [23]).

- The Figure shows the layout and cooling facility of a warehouse in a data center.
- The data-center room has raised floors for hiding cables, power lines, and cooling supplies.

Cooling System of a Data-Center Room

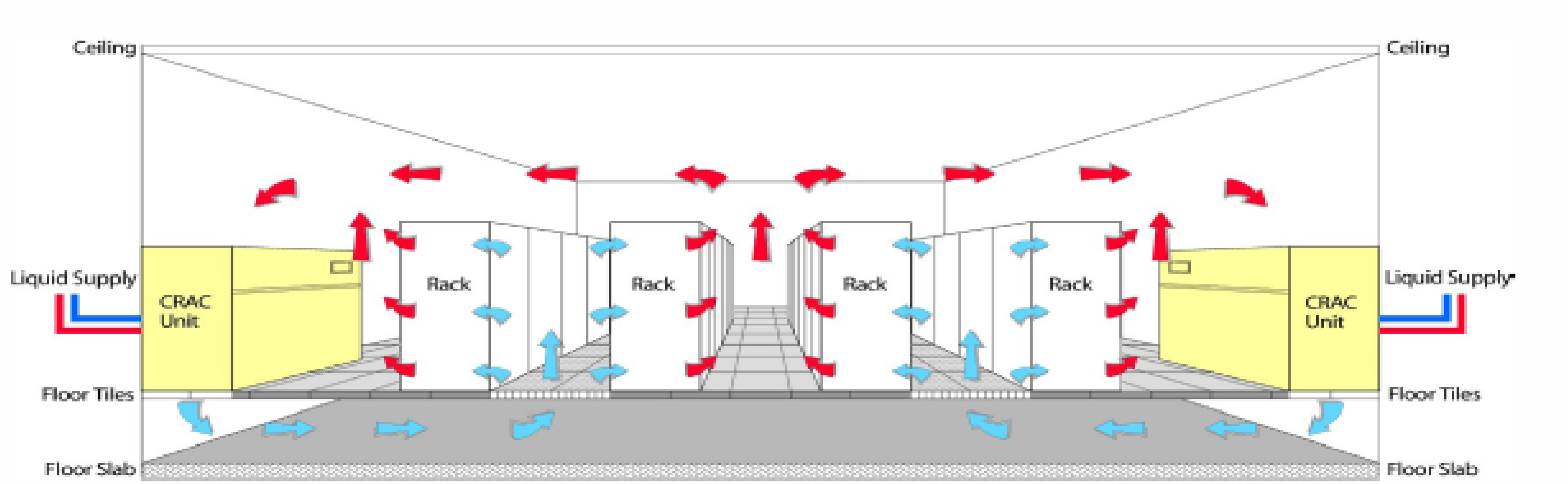
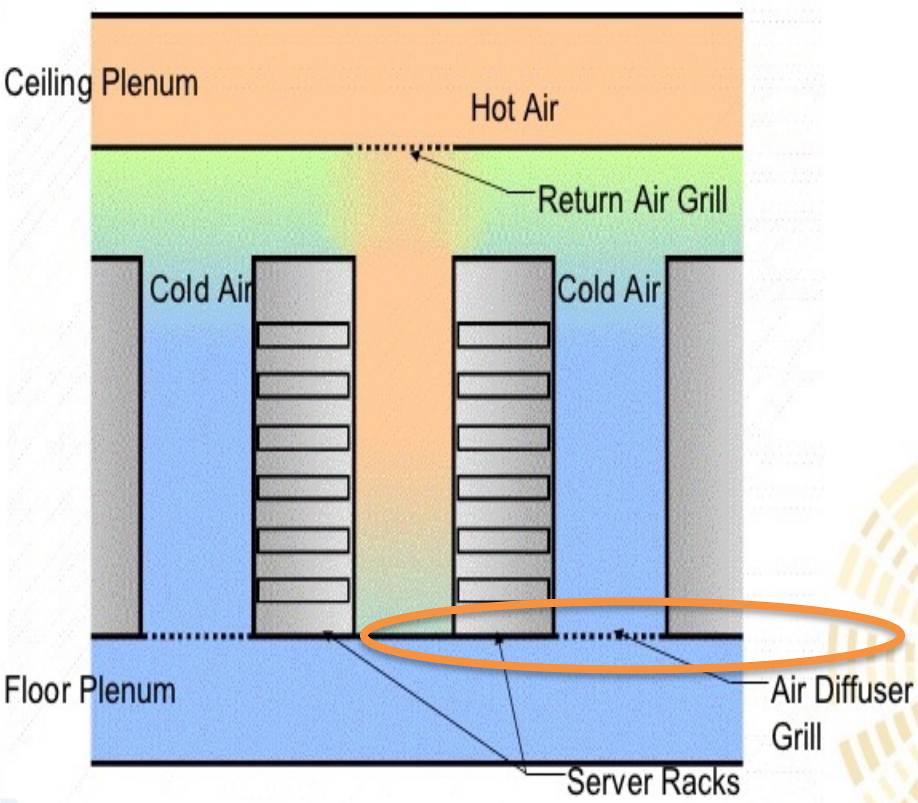


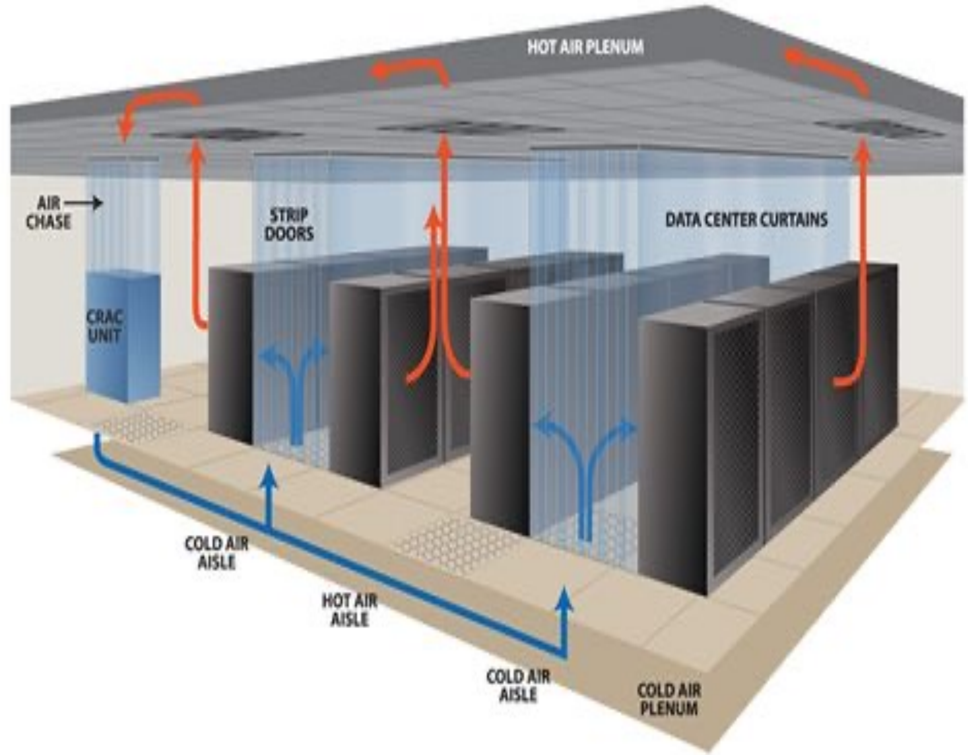
FIGURE 4.2: Datacenter raised floor with hot–cold aisle setup (image courtesy of DLB Associates [23]).

- The raised floor has a steel grid resting on stanchions about 2–4 ft above the concrete floor.
- The under-floor area is often used to route power cables to racks, but its primary use is to distribute cool air to the server rack.

Stanchion-frame/fixture to support other object

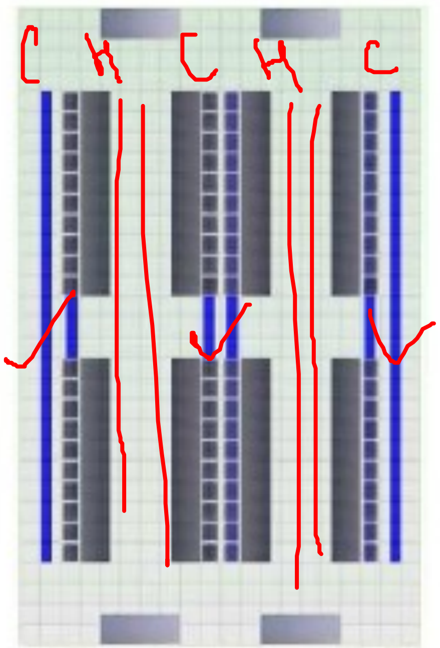
Cooling System of a Data-Center Room



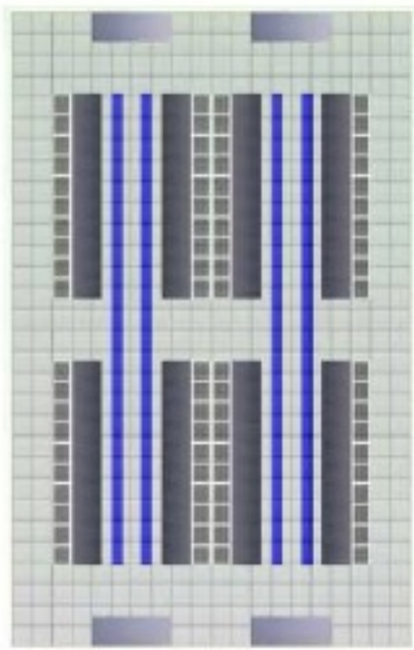


- The CRAC (computer room air conditioning) unit pressurizes the raised floor plenum by blowing cold air into the plenum.
- The cold air escapes from the plenum through perforated tiles that are placed in front of server racks.

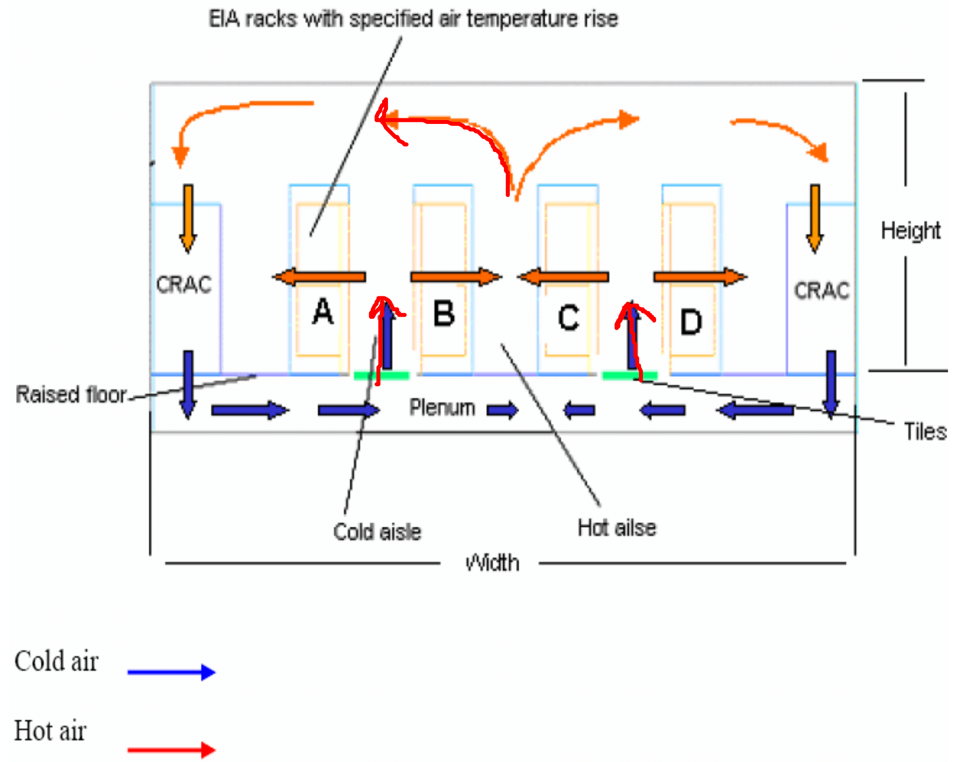
Cooling System of a Data-Center Room



(a) Cold Aisle Cable (CAC)



(b) Hot Aisle Cable (HAC)



- Racks are arranged in long aisles that alternate between cold aisles and hot aisles to avoid mixing hot and cold air.

Cooling System of a Data-Center Room

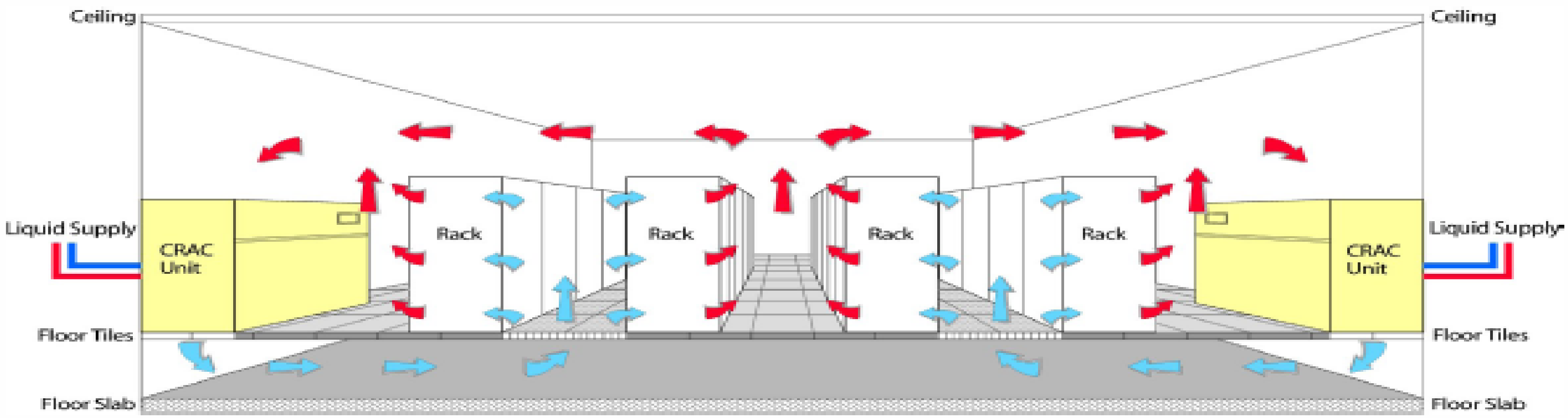
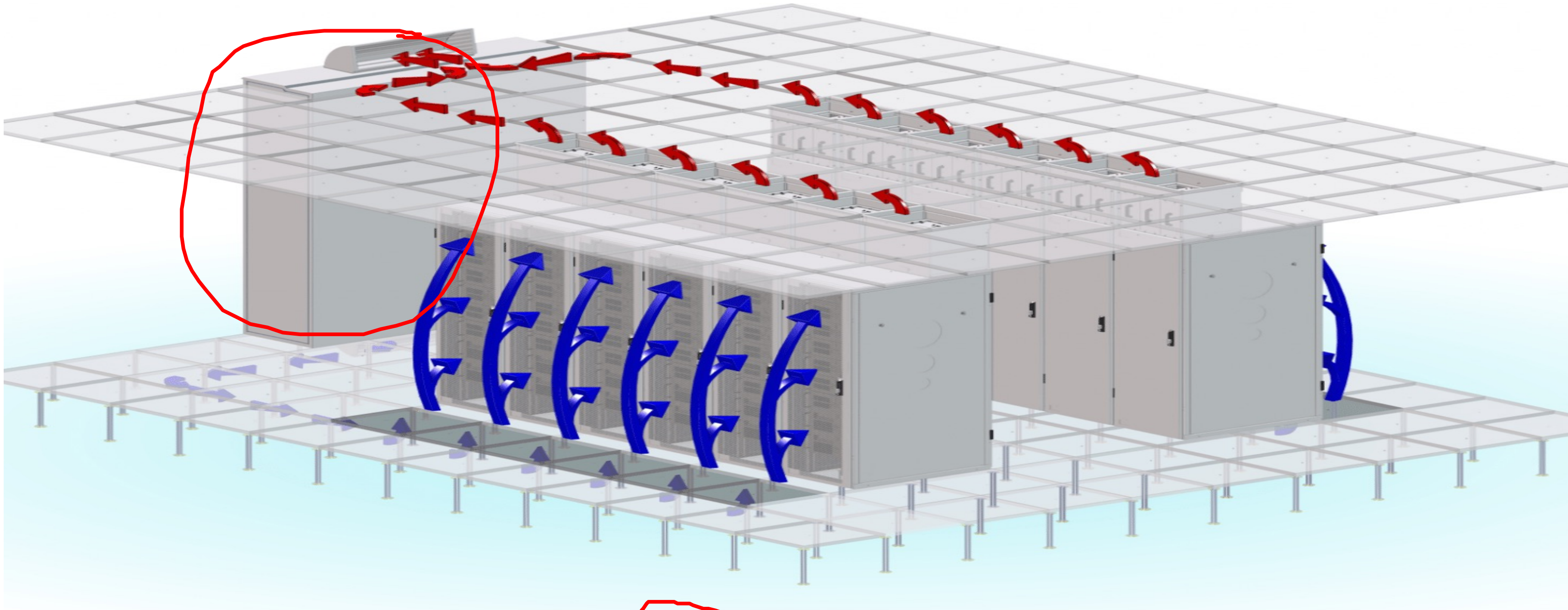
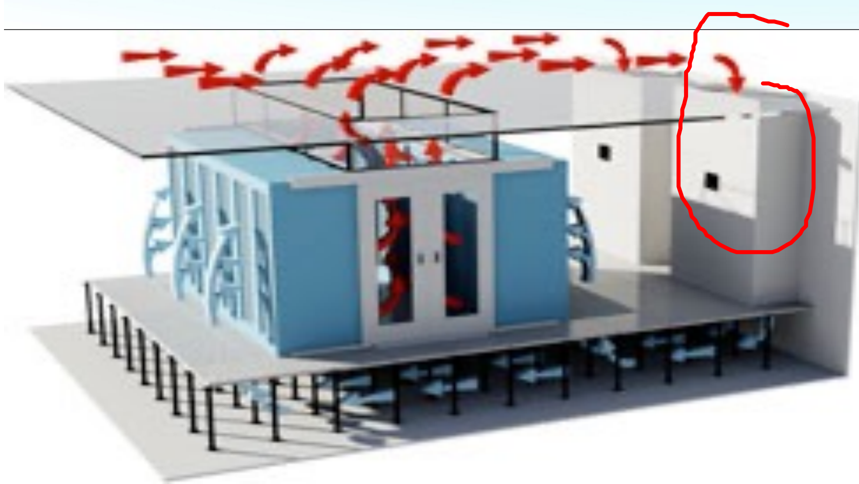


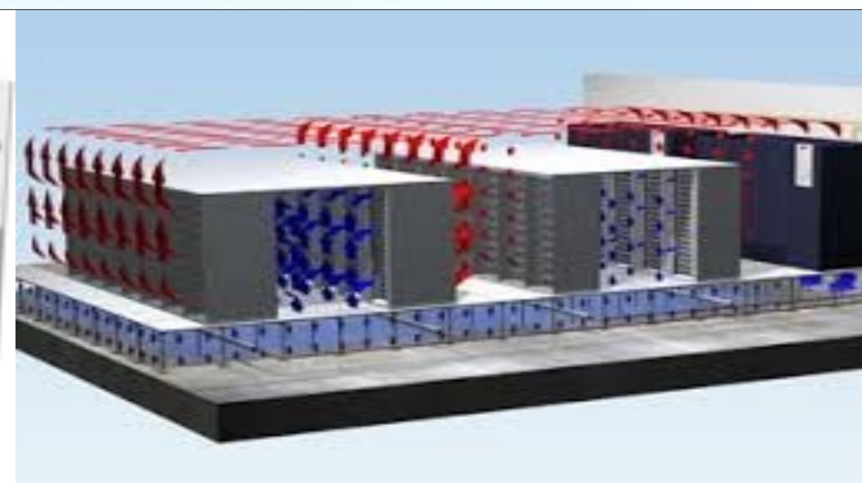
FIGURE 4.2: Datacenter raised floor with hot–cold aisle setup (image courtesy of DLB Associates [23]).

- The hot air produced by the servers circulates back to the intakes of the CRAC units that cool it and then exhaust the cool air into the raised floor plenum again.
- Typically, the incoming coolant is at 12–14°C and the warm coolant returns to a chiller.

Cooling System of a Data-Center Room







Green Computing ?

Advanced scheduling schemas to reduce energy consumption.

- Power aware
- Thermal aware

Data center designs to reduce Power Usage.

- Cooling systems
- Rack design

Performance <-> Energy Efficiency

- There is a need to shift focus from optimizing data center resource management for pure performance alone to
  - optimizing for energy efficiency while maintaining high service level performance.

Energy efficiency in clouds

- Cloud service providers need to adopt measures to ensure that their profit margins are not dramatically reduced due to high energy costs.

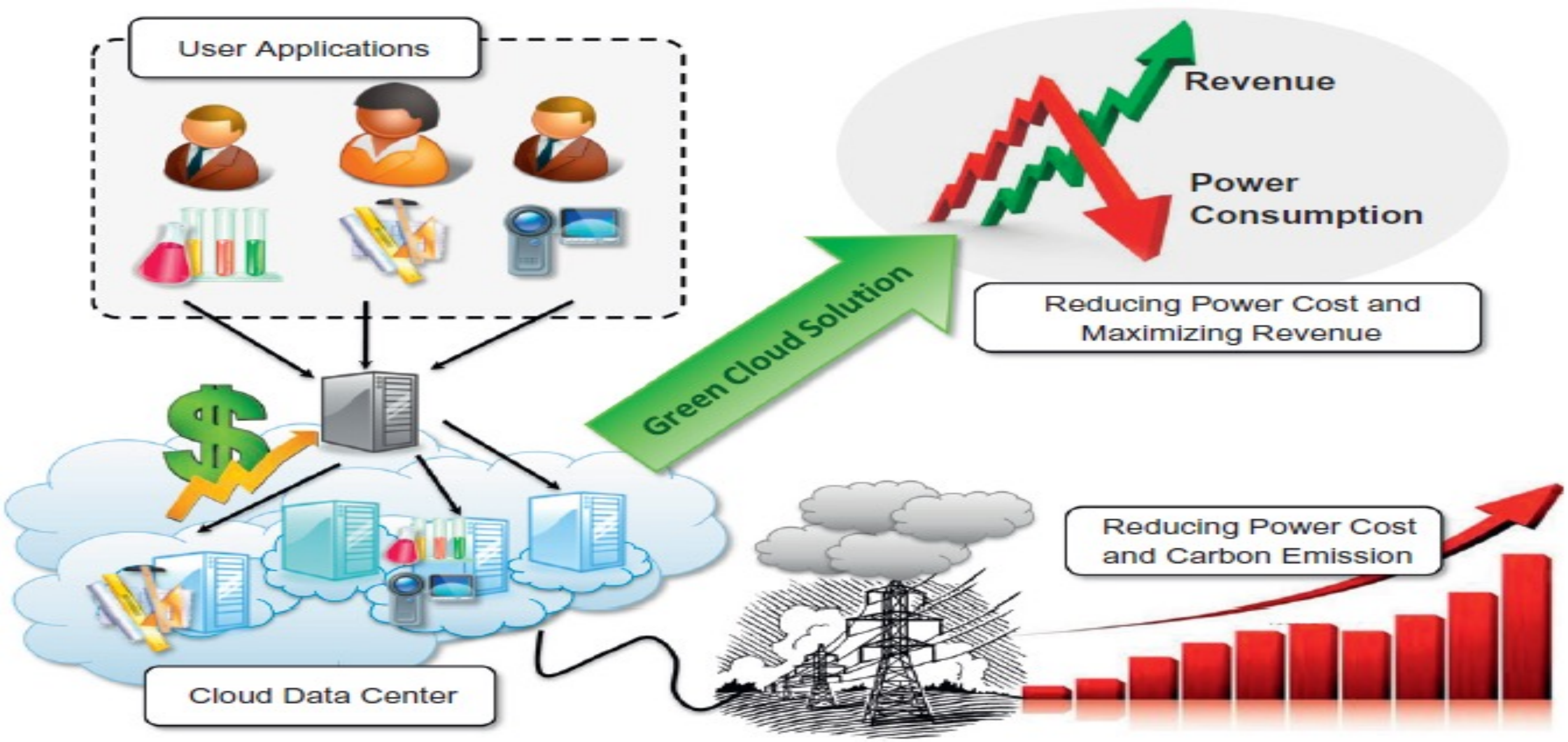


FIGURE 11.1

A "green" cloud computing scenario.

Energy efficiency in clouds

- According to Amazon's estimate,
- the energy- related costs of its datacenters amount to 42% of the total budget,
- which includes both direct power consumption and the cooling infrastructure amortized over a 15-year period.

- As a result, companies such as Google, Microsoft, and Yahoo! are building large datacenters in barren desert land surrounding the Columbia River in the United States to exploit cheap hydroelectric power.

Energy efficiency in clouds

- There is also increasing **pressure from governments worldwide to reduce carbon footprints,** which have a significant impact on climate change.

- To address these concerns, leading IT vendors have recently formed a global consortium, called **The Green Grid, to promote energy efficiency for datacenters and minimize their impact on the environment**

Energy efficiency in clouds

- Green cloud computing is envisioned to achieve not only efficient processing and utilization of computing infrastructure but also minimize energy consumption. This is essential for ensuring that the future growth of cloud computing is sustainable
- Cloud resources need to be allocated **not only to satisfy QoS requirements specified by users via service-level agreements (SLAs) but also to reduce energy usage.**

Energy Consumption of Unused Servers

- To run a server farm (data center) a company has to spend a huge amount of money for hardware, software, operational support, and energy every year.

- Companies should thoroughly <u>identify whether their installed server farm are appropriately utilized.</u>

Energy Consumption of Unused Servers

S
T
A
T
S

- On average, one-sixth (15 percent) of the full-time servers in a company are left powered on without being actively used (i.e., they are idling) on a daily basis.

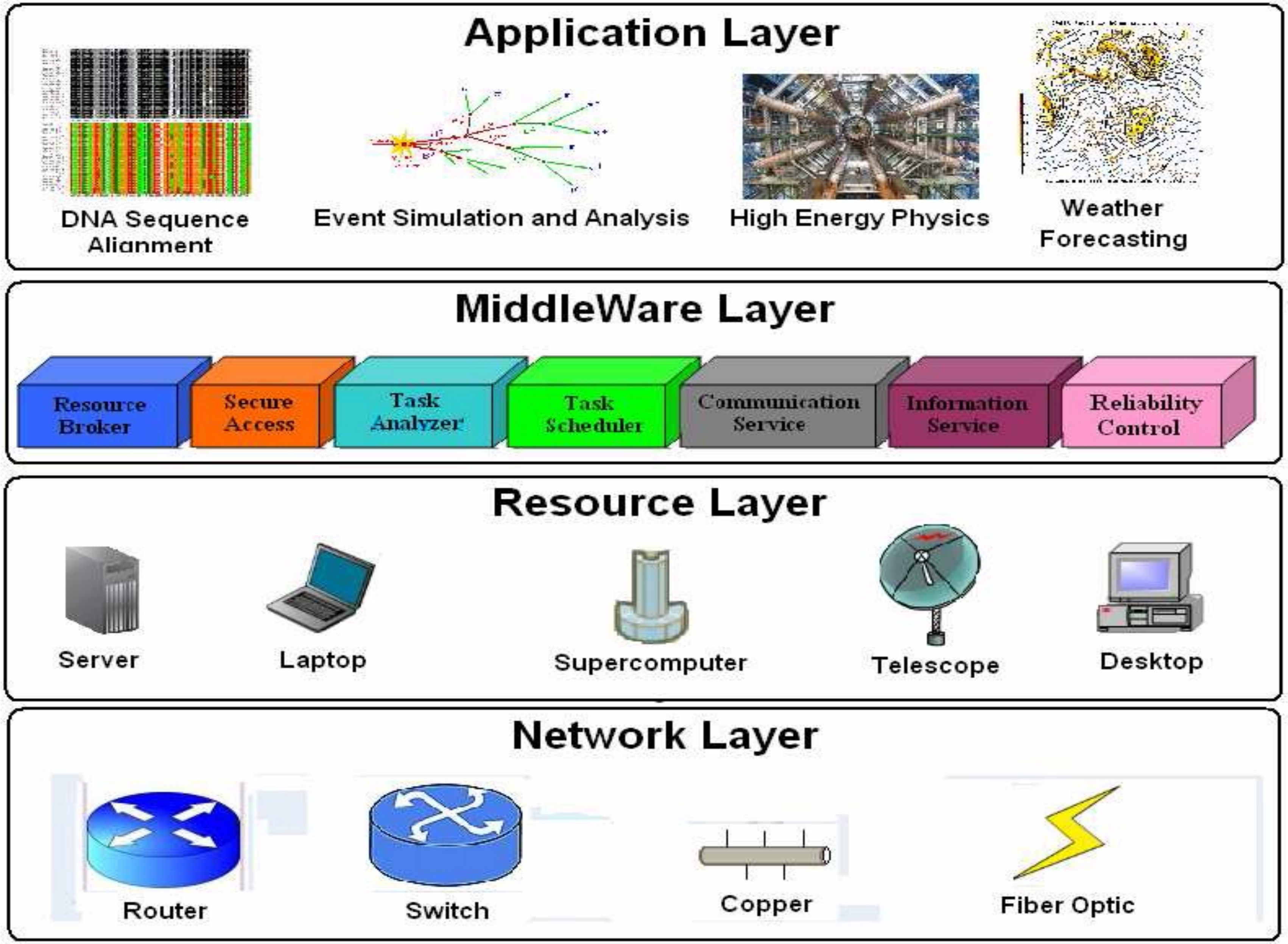
- This indicates that with 44 million servers in the world, around 4.7 million servers are not doing any useful work.

- The potential savings in turning off these servers are large.

- IT departments need to analyze their servers and find out unused or underutilized servers.

Reducing Energy in Active Servers

- Apply appropriate techniques to decrease energy consumption in active distributed systems with negligible influence on their performance.
- Power management issues in distributed computing platforms can be categorized into four layers:
  - application layer
  - middleware layer
  - resource layer
  - network layer.


Application Layer
DNA Sequence Alignment
Event Simulation and Analysis
High Energy Physics
Weather Forecasting
MiddleWare Layer
Resource Broker
Secure Access
Task Analyzer
Task Scheduler
Communication Service
Information Service
Reliability Control
Resource Layer
Server
Laptop
Supercomputer
Telescope
Desktop
Network Layer
Router
Switch
Copper
Fiber Optic

Application Layer

- An application's energy consumption depends strongly on:
  - no of instructions needed to execute the application
  - no of transactions with the storage unit (or memory).

- These two factors (compute and storage) are correlated and they affect completion time.

Application Layer

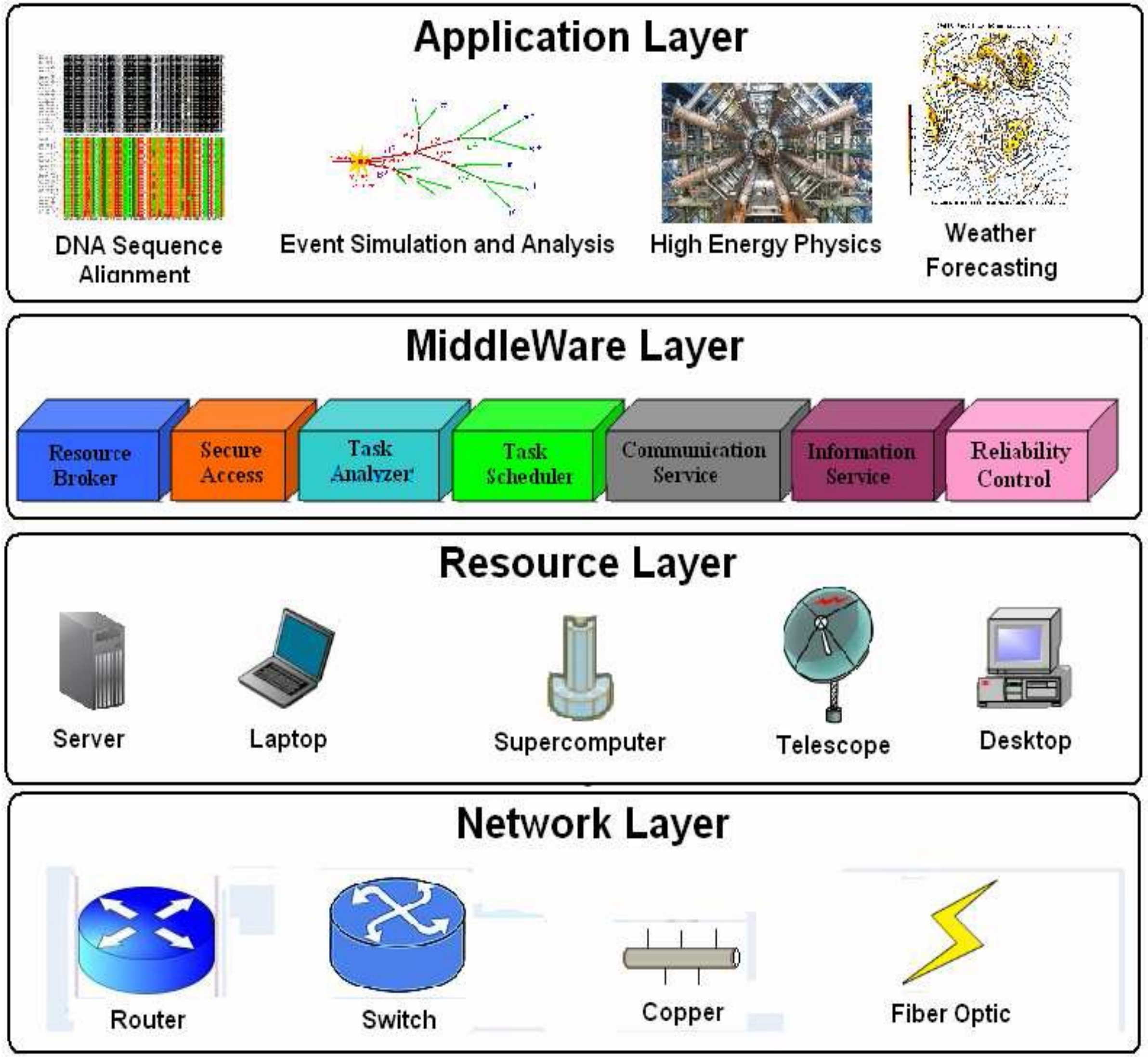
- By introducing energy-aware applications, the challenge is to design applications without hurting performance.

- The first step is-

  "To explore a relationship between performance and energy consumption".

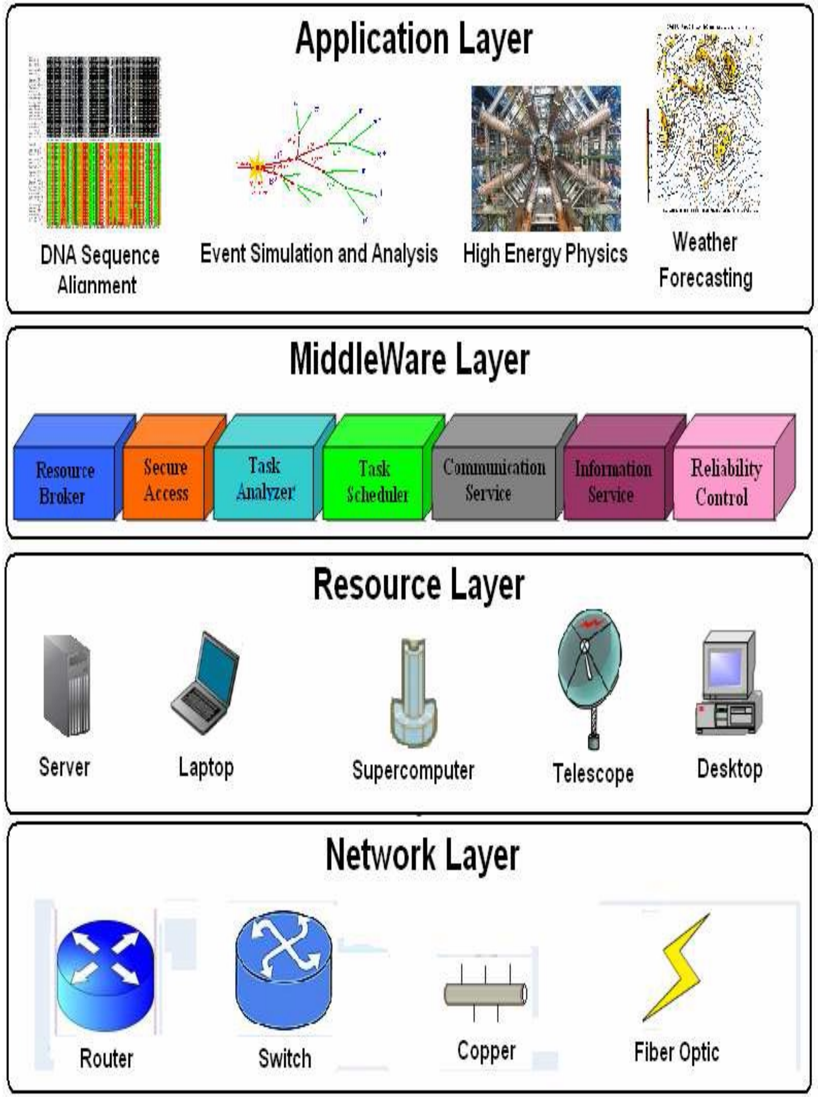
Middleware Layer

- Bridge between Application Layer and the resource layer.
- This Layer provides:
  - Resource broker
  - Communication service
  - Task analyzer
  - Task scheduler
  - Security access
  - Reliability control
  - Information Service



Middleware Layer

- Necessity of a <u>new cost function covering both</u>
  - makespan i.e. the <u>execution time of a set of tasks and</u>
  - <u>energy consumption.</u>

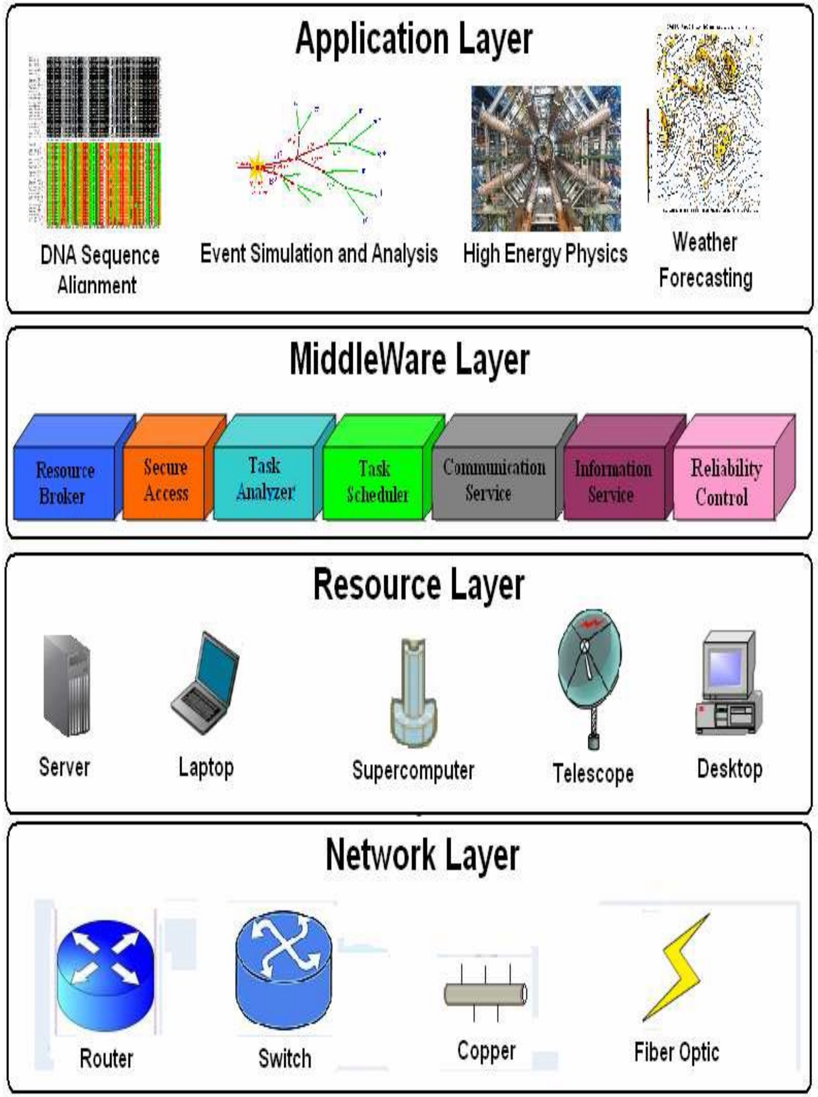


Resource Layer

- Consists of a wide range of resources including
  - computing nodes and
  - storage units.

- Interacts with hardware devices and the operating system.

- Several mechanisms have been developed for
  - more efficient power management of hardware and operating systems.



Resource Layer

Hardware Approaches:-

- Dynamic power management (DPM)
- Dynamic voltage-frequency scaling (DVFS)

Popular methods incorporated into recent computer hardware systems.

Resource Layer- DPM

Dynamic power management (DPM)

In DPM, hardware devices, such as the CPU, have the capability to-

“Switch from idle mode to one or more lower power modes”

Resource Layer- DVFS

Dynamic voltage-frequency scaling (DVFS)

- In DVFS, energy savings are achieved based on the fact that the power consumption in CMOS circuits has a direct relationship with frequency and the square of the voltage supply.

Resource Layer- DVFS

- The relationship between energy ,voltage ,frequency in CMOS circuits is related by:

  where

  - V= voltage
  - Ceff= circuit switching capacity
  - K= a technology dependent factor
  - Vt= threshold voltage
  - T= execution time of the task under clock frequency f.

$$\begin{cases} E = C_{eff} f v^2 t \\ f = K \dfrac{(v - v_t)^2}{v} \end{cases}$$

Resource Layer- DVFS

- Execution time and power consumption are controllable by-

  “<u>Switching among different frequencies and voltages</u>”

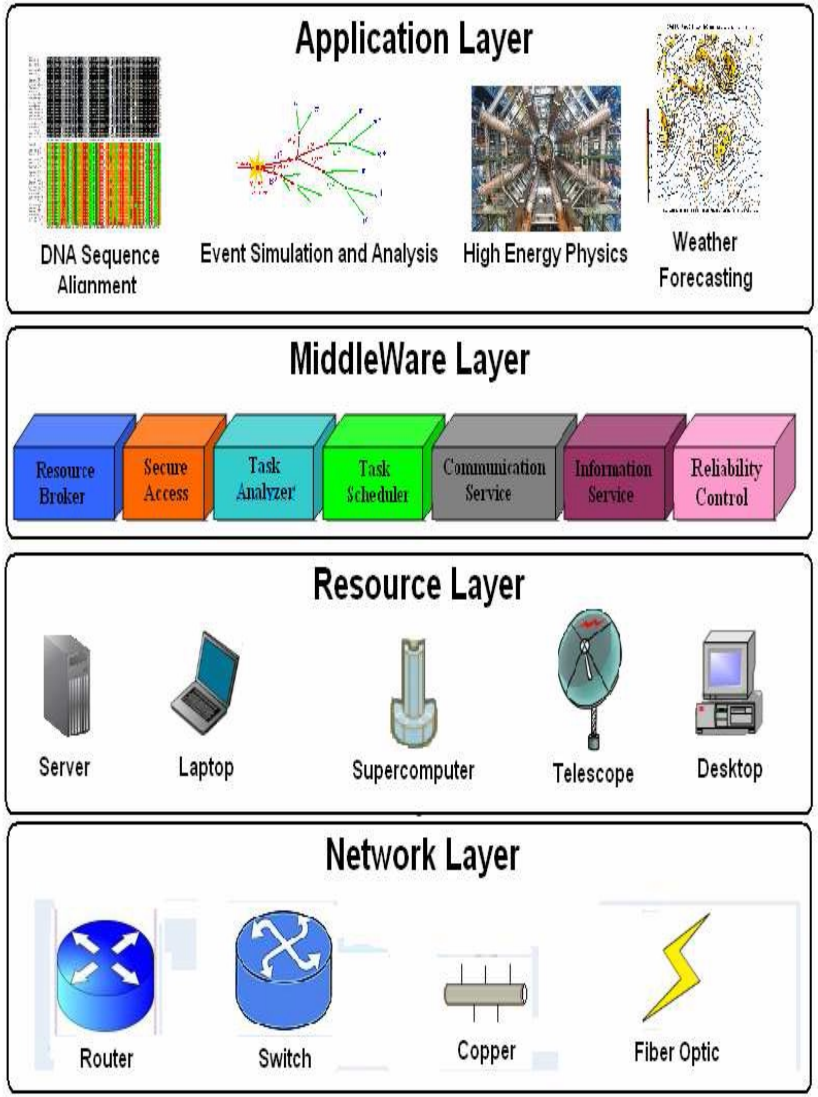
DVFS

- The DVFS method enables the exploitation of the slack time (idle time) typically incurred by intertask relationship.
  - Reducing the frequency or voltage during slack time.

- **By reducing voltage and frequency, the device's energy consumption can also be reduced.**

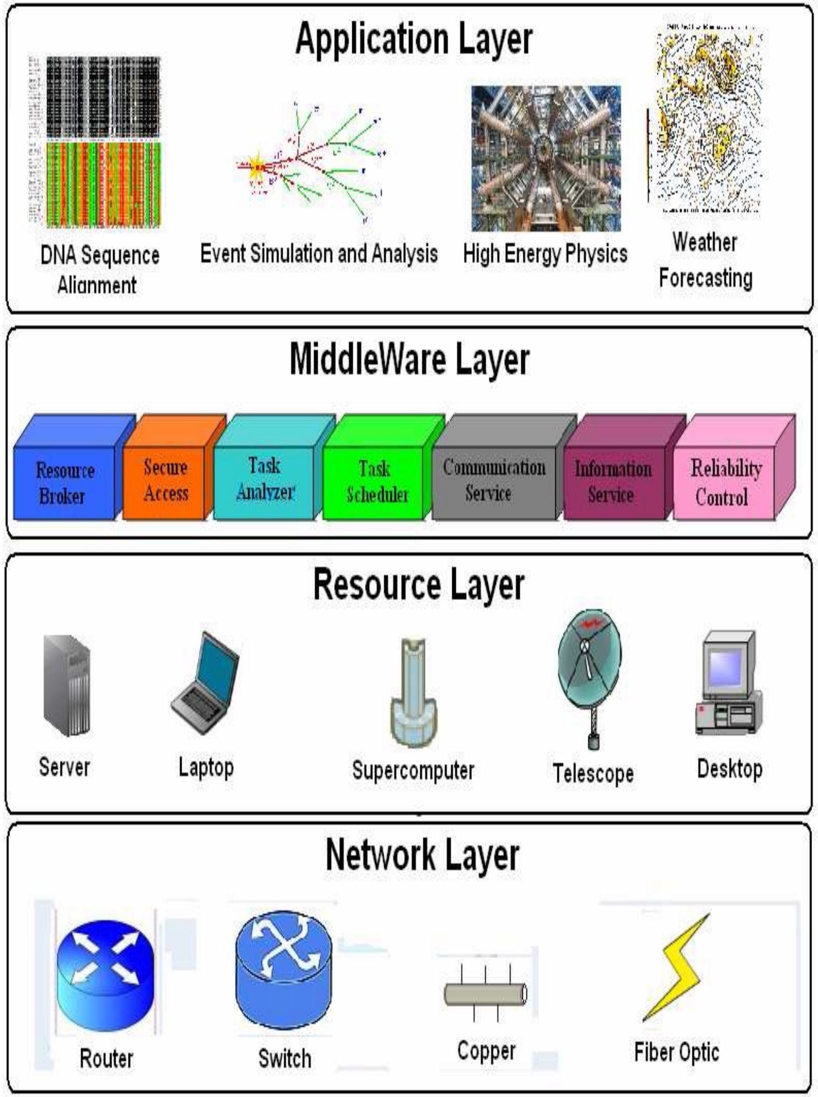
Network Layer

- Routing and transferring packets
- <u>Enabling network services to the resource layer</u> are the main responsibility of the network layer in Distributed Computing Systems.



Network Layer

- Two major challenges are :
  - New, energy-efficient routing algorithms need to be developed.
  - New, energy-efficient protocols should be developed against network attacks.



A reference model for MOCC

- Market-oriented cloud computing originated from the coordination of several components:
- service consumers,
- service providers, and
- other entities that make trading between these two groups possible.

A reference model for MOCC

- Several components contribute to the realization of the Cloud Exchange and implement its features.

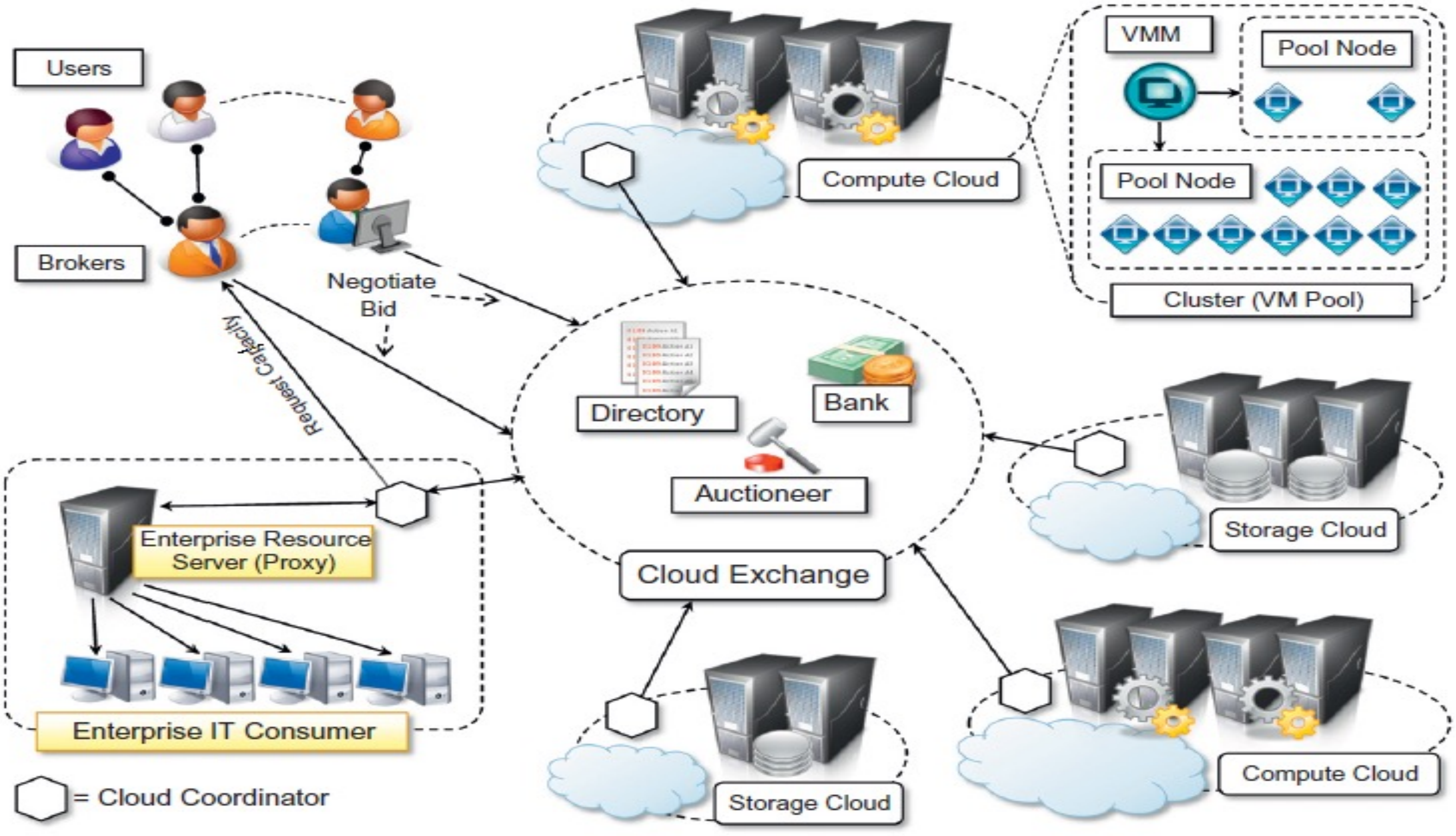


FIGURE 11.3
Market-oriented cloud computing scenario.

A reference model for MOCC

Directory:-

- The market directory contains **a listing of all the published services that are available**
- The directory not only contains **a simple mapping between service Names and the corresponding vendor (or cloud coordinators) offering them.**

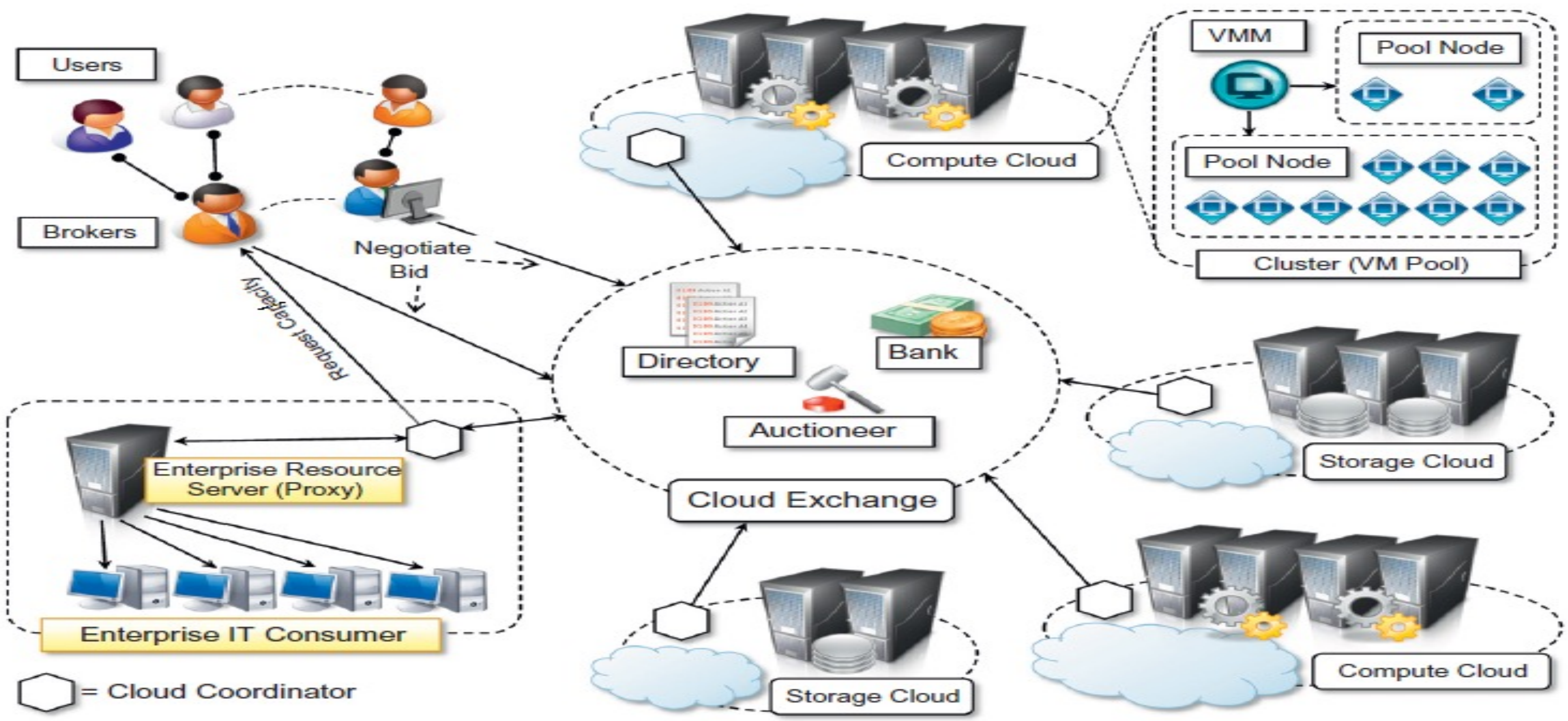


FIGURE 11.3

Market-oriented cloud computing scenario.

A reference model for MOCC

Directory:-

- It also provides Additional meta data that can **help the brokers or the end users in filtering from among the services of interest those** that can really meet the expected quality of service.

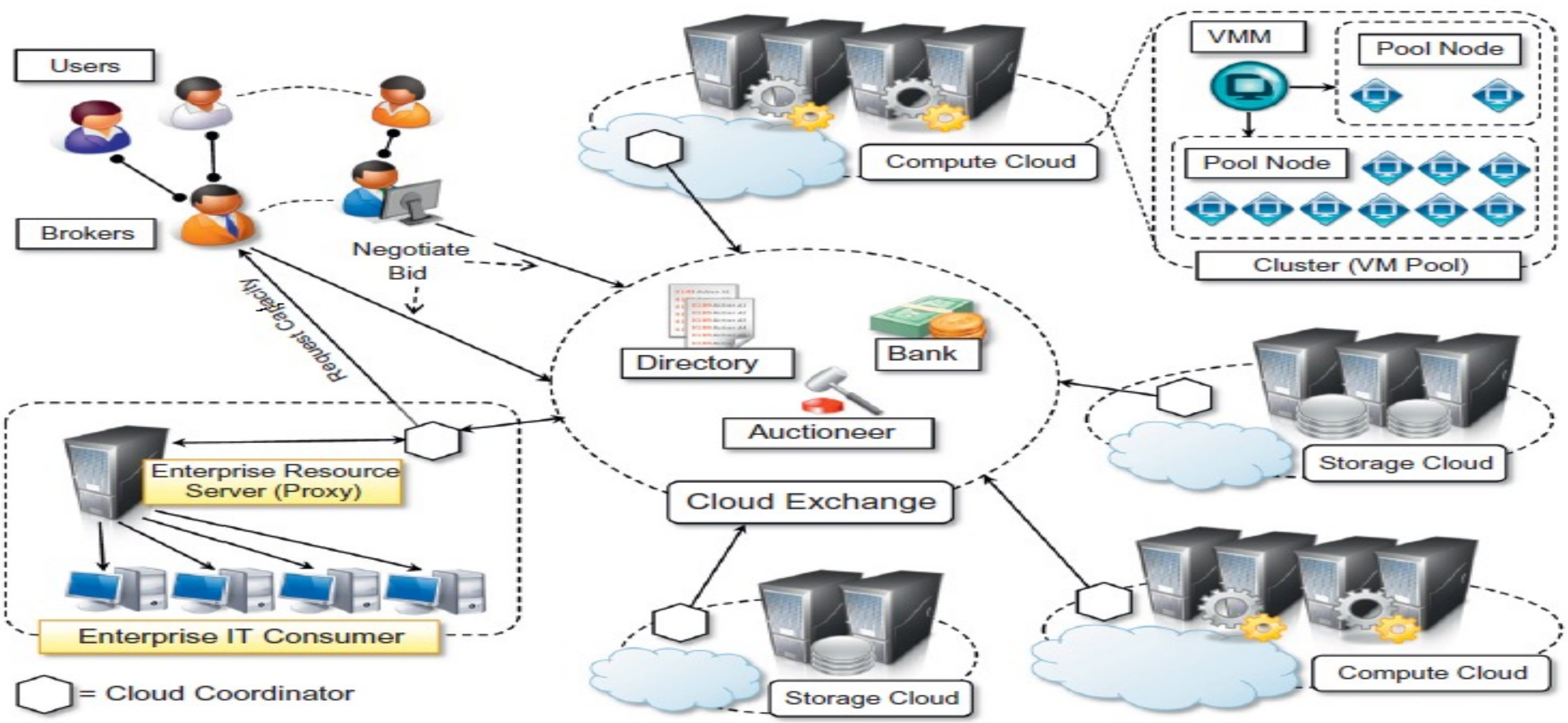


FIGURE 11.3
Market-oriented cloud computing scenario.

A reference model for MOCC

Auctioneer

- The auctioneer is incharge of **keeping track of the running auctions in the market place** and of verifying that the **auctions for services are properly conducted** and that **malicious market players are prevented from performing illegal activities**

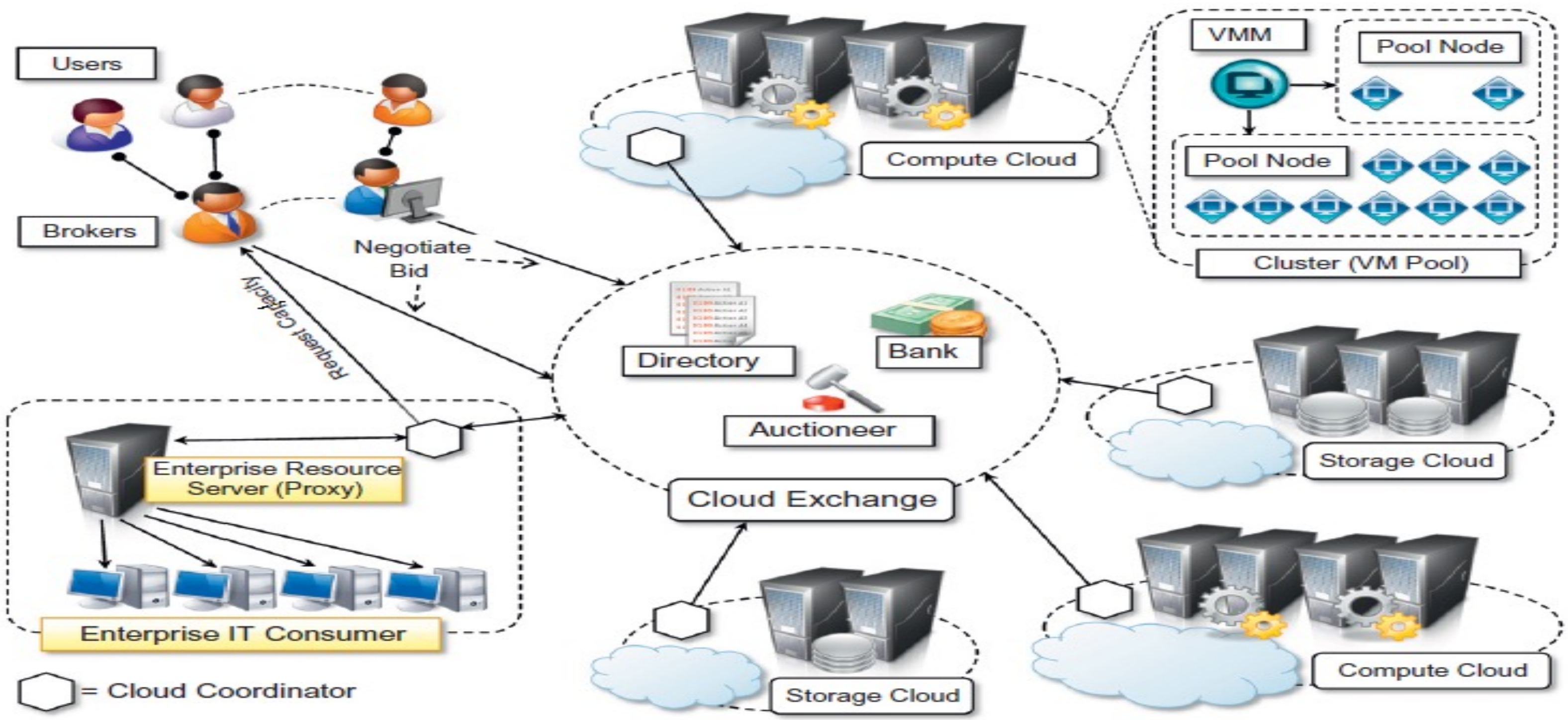


FIGURE 11.3

Market-oriented cloud computing scenario.

A reference model for MOCC

Bank

- The bank is the **component that takes care of the financial aspect of all the operations happening in the virtual marketplace.**
- It also ensures that all the financial transactions are **carried out in a secure and dependable environment.**
- **Consumers and providers may register with the bank and have one or multiple accounts** that can be used to perform the transactions in the virtual marketplace

A reference model for MOCC

CEx

- The fundamental component is the virtual marketplace—represented by the **Cloud Exchange(CEx)—**which acts as a market maker, **bringing service producers and consumers together.**

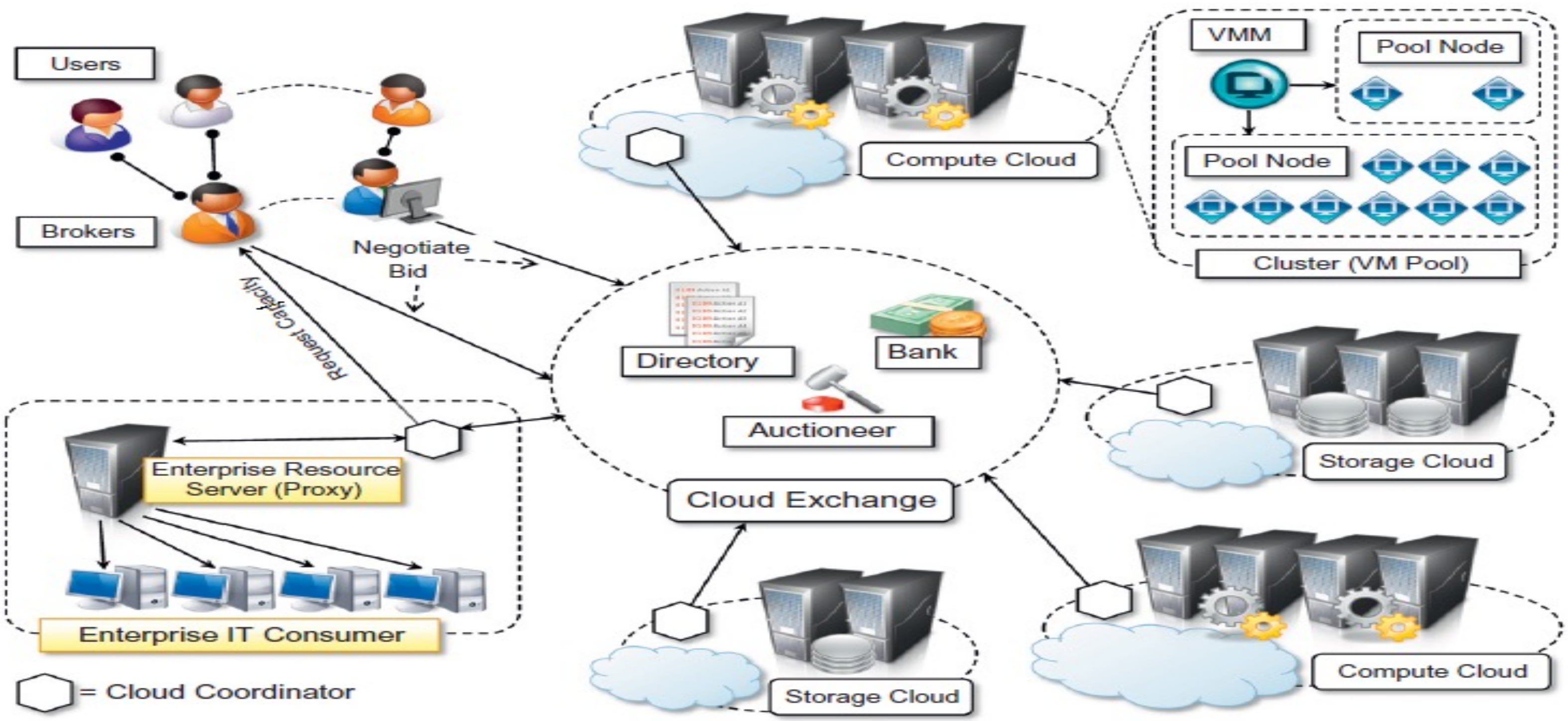


FIGURE 11.3
Market-oriented cloud computing scenario.

A reference model for MOCC

Cloud Broker

- The cloud brokers **operate on behalf of the consumers and identify the subset of services that match customers' requirements** in terms of service profiles and quality of service.

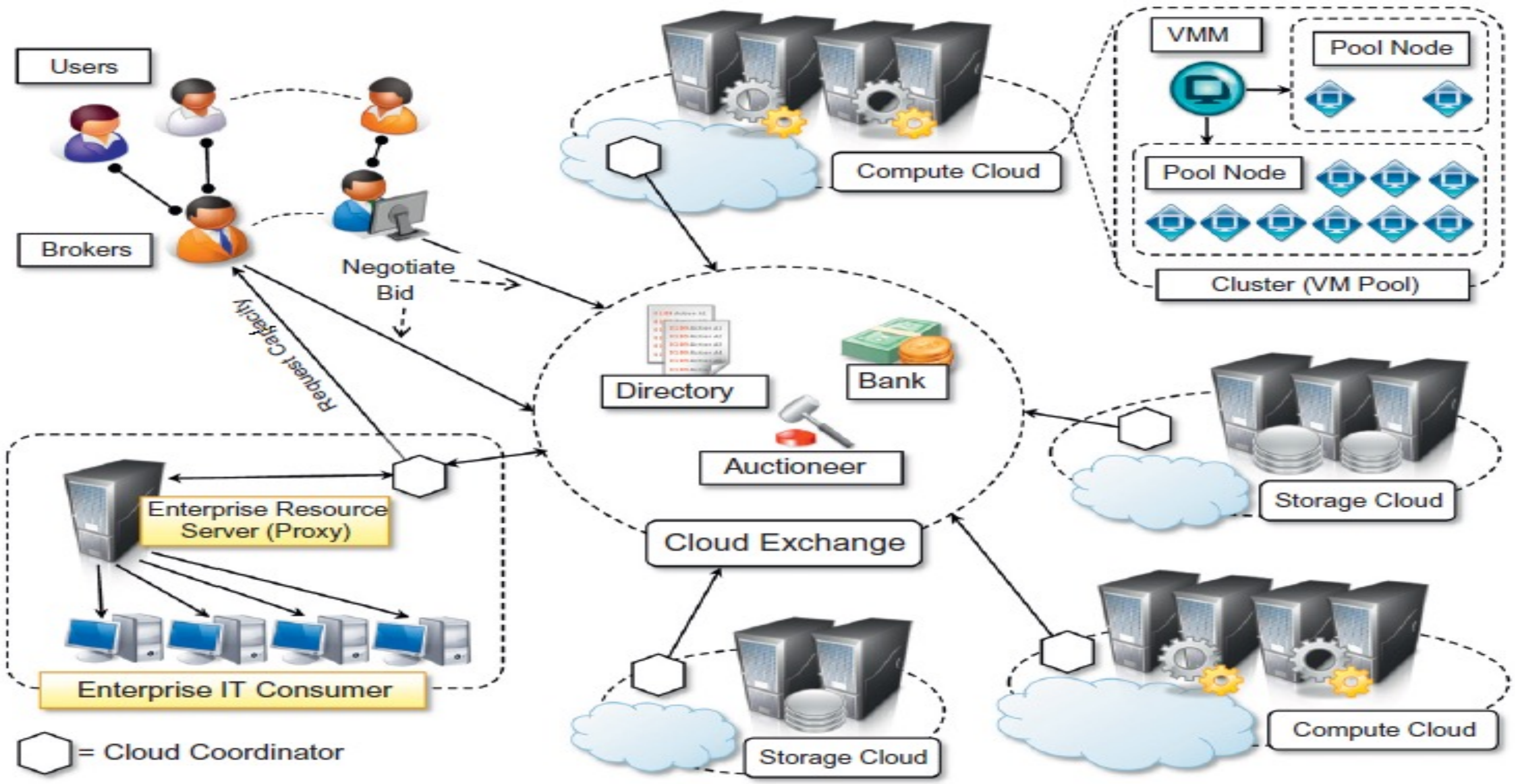


FIGURE 11.3

Market-oriented cloud computing scenario.

A reference model for MOCC

Cloud Broker

- They **mediate between coordinators and consumers by acquiring services from the first and subleasing them to the latter**.
- **Brokers can accept requests from many users. At the same time, users can leverage different brokers.**

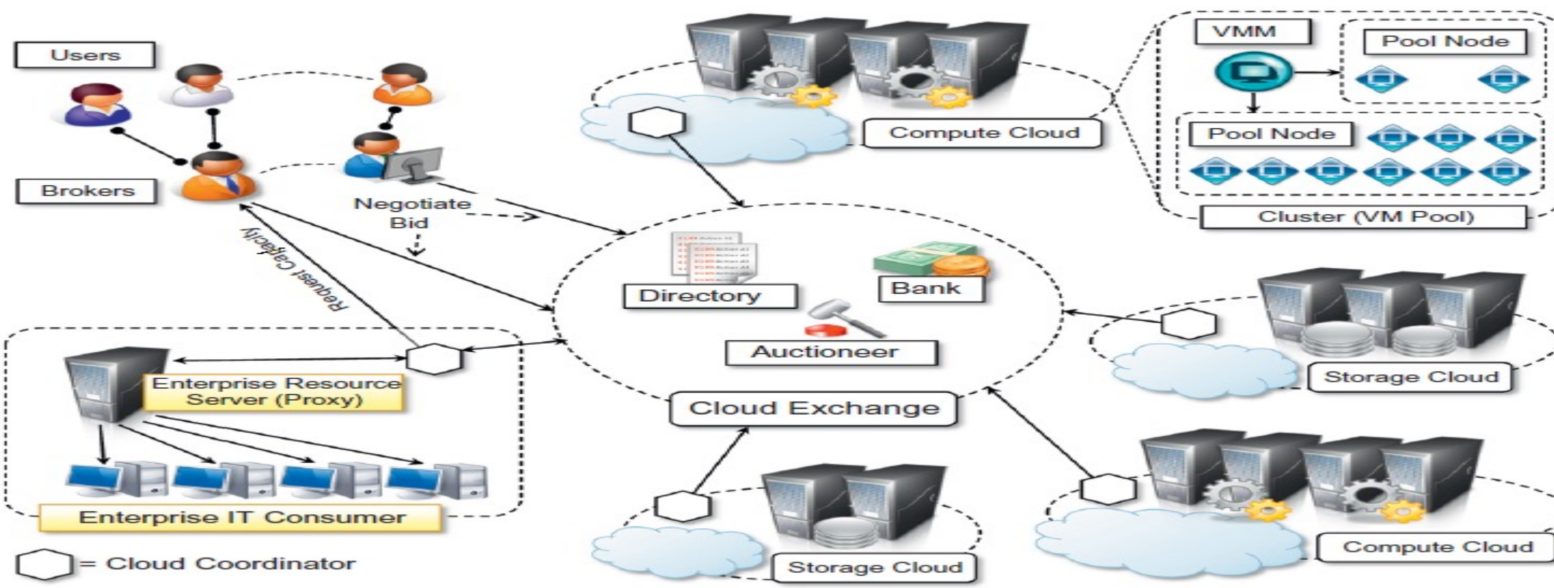


FIGURE 11.3
Market-oriented cloud computing scenario.

A reference model for MOCC

Cloud Coordinators

- The cloud coordinators represent **the cloud vendors and publish the services that vendors offer.**
- A similar relationship can be considered between coordinators and cloud computing services vendors.

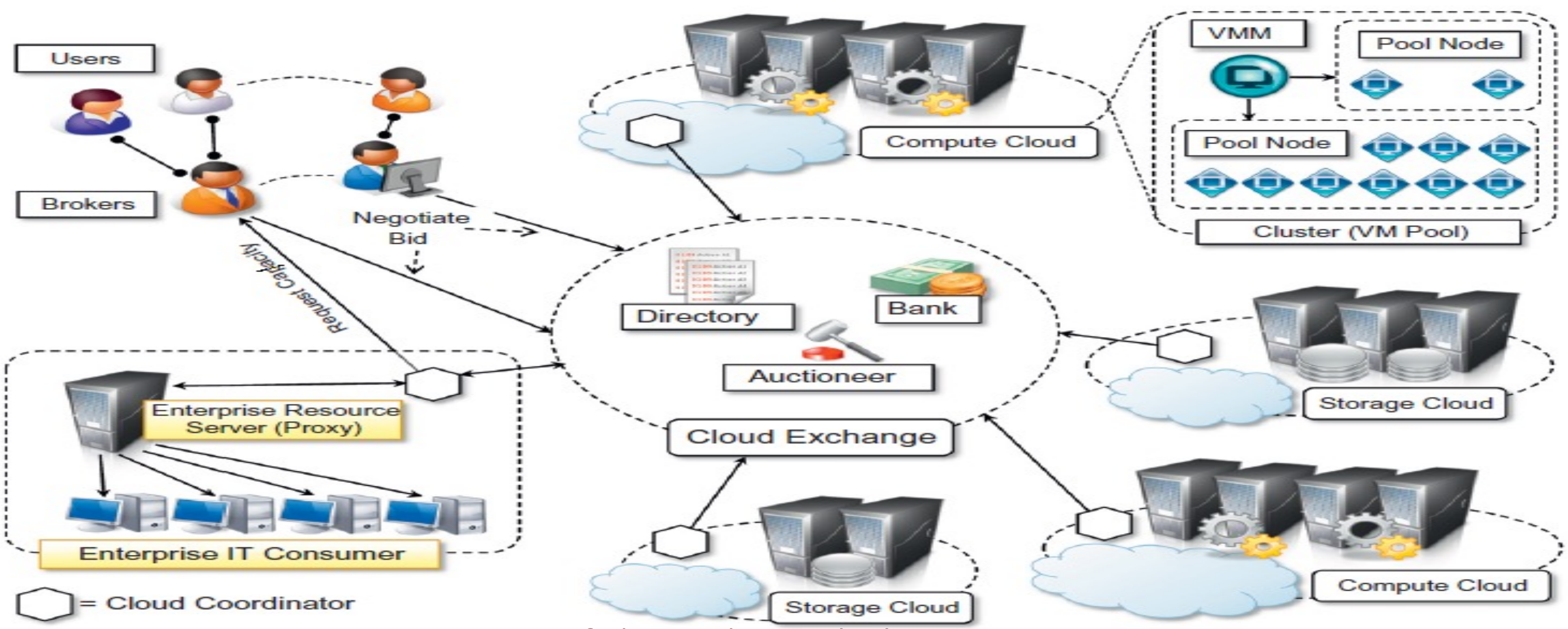


FIGURE 11.3

Market-oriented cloud computing scenario.

A reference model for MOCC

Cloud Coordinators

- Coordinators take **responsibility for publishing and advertising services on behalf of vendors and can gain benefits from reselling services to brokers.**

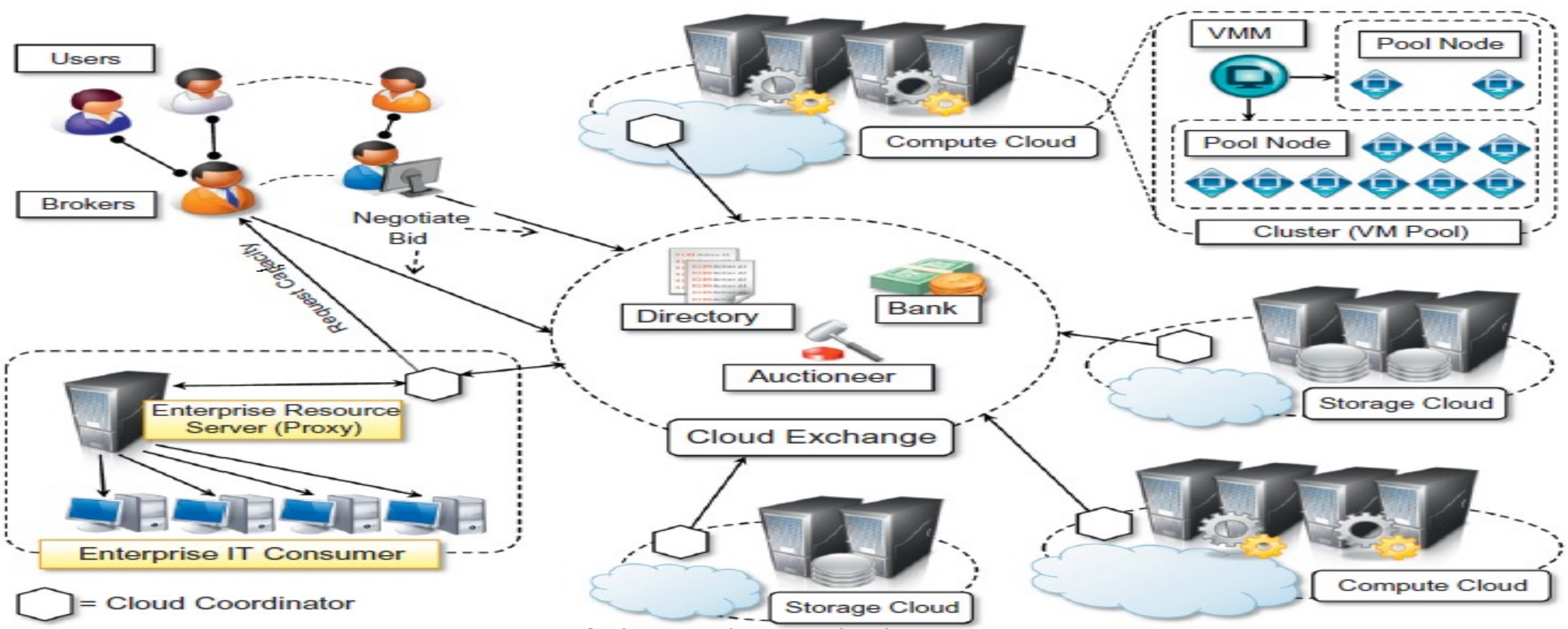


FIGURE 11.3

Market-oriented cloud computing scenario.

Market-Oriented Cloud Architecture for Data Centers

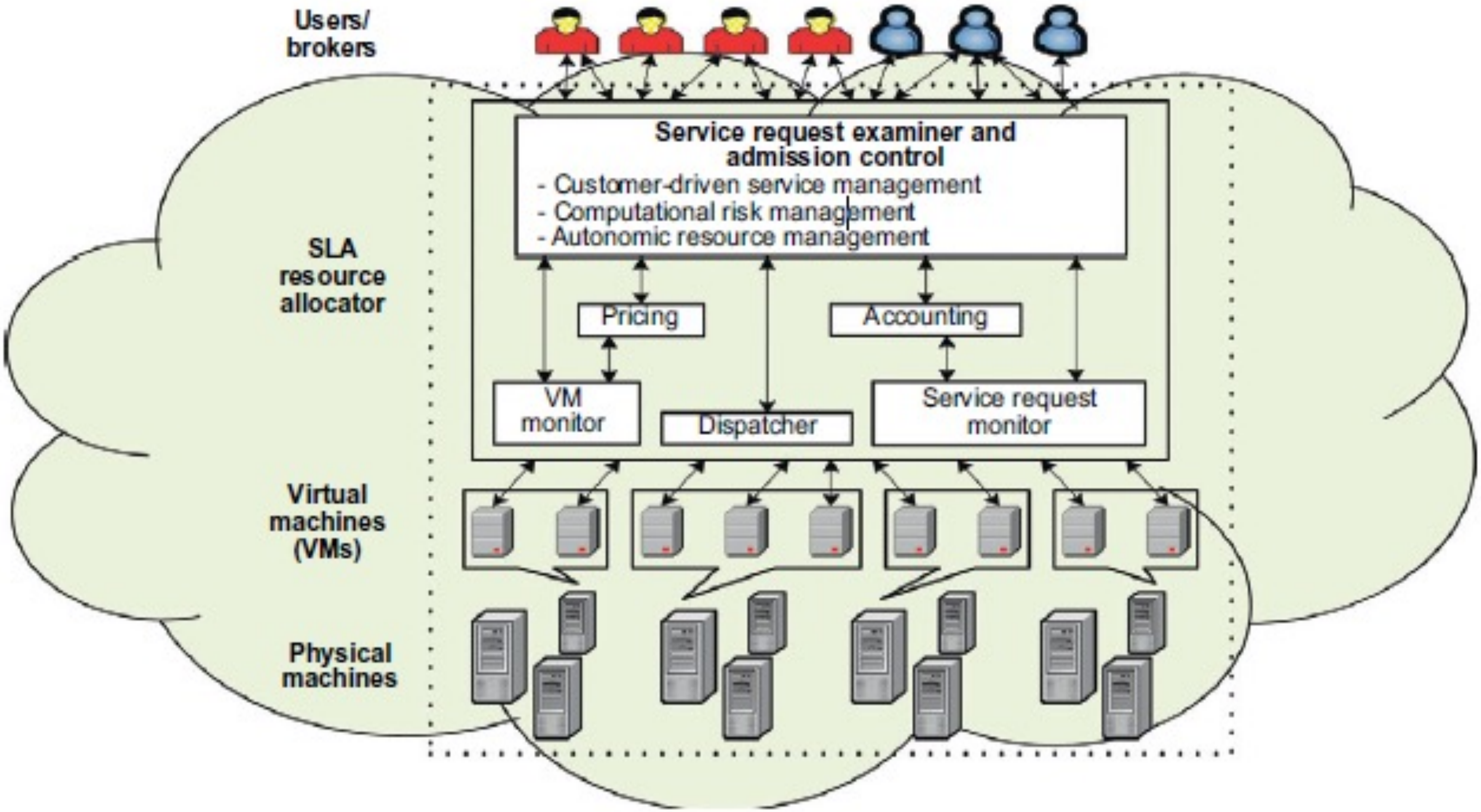


FIGURE 4.16

Market-oriented cloud architecture to expand/shrink leasing of resources with variation in QoS/demand from users.

Market-Oriented Cloud Architecture

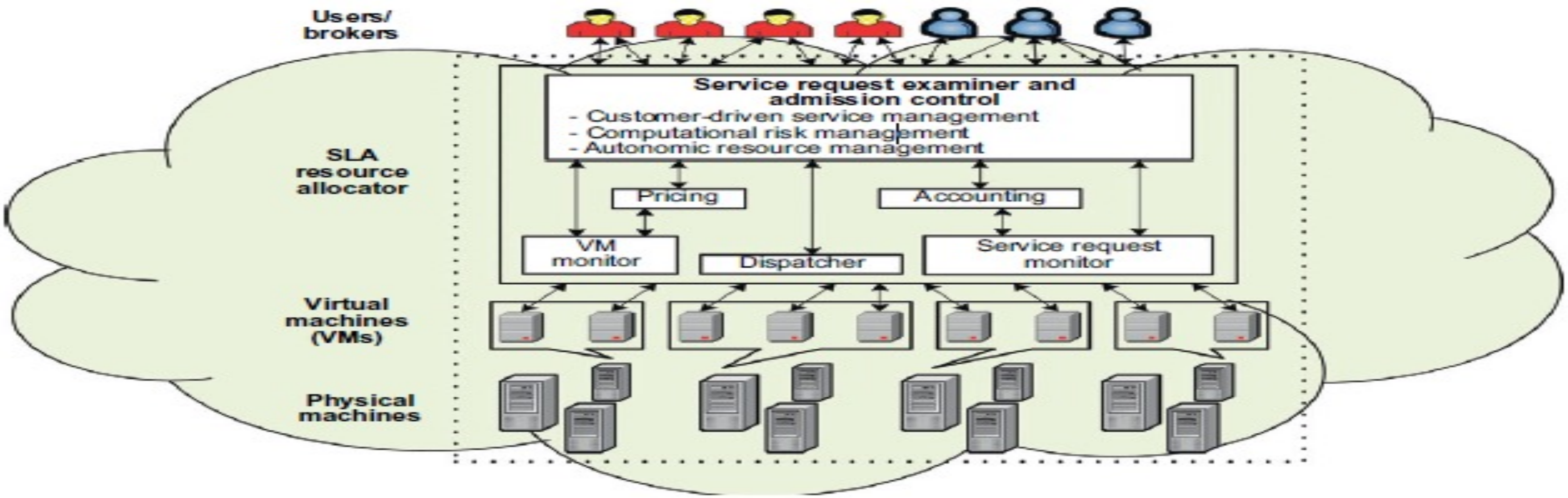
- As consumers rely on cloud providers to meet more of their computing needs,
  - they will require a specific level of QoS to be maintained by their providers.

- Cloud providers consider and **meet the different QoS parameters of each individual consumer as negotiated in specific SLAs.**

Market-Oriented Cloud Architecture

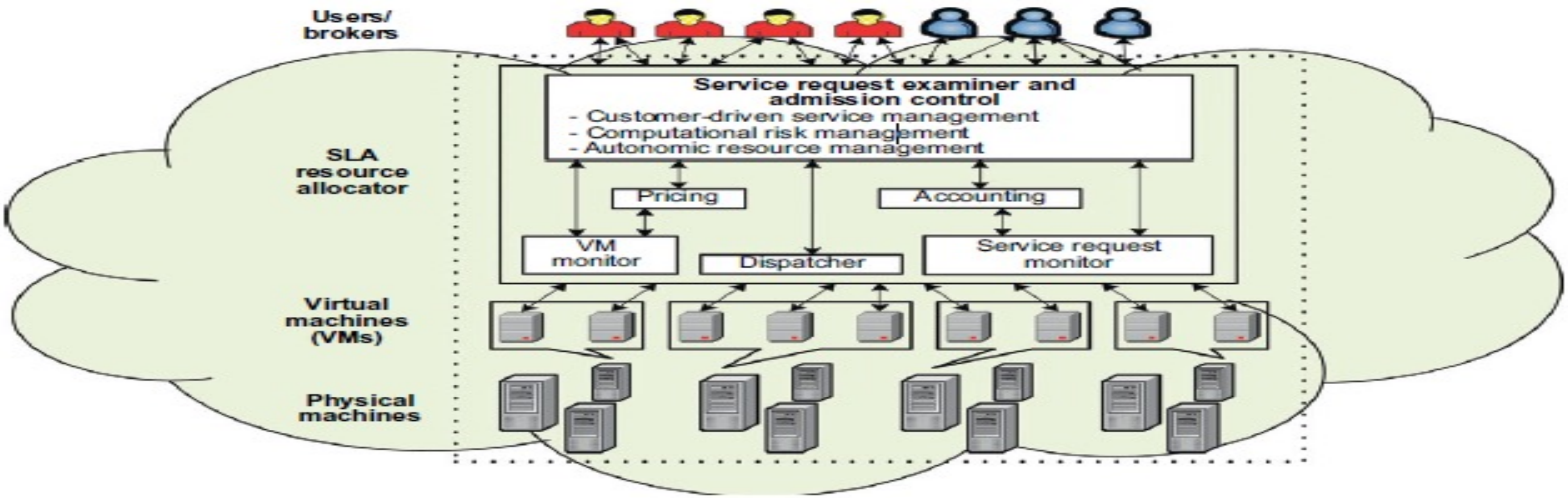
- The purpose is to promote QoS-based resource allocation mechanisms.
  - Clients can benefit from the potential cost reduction of providers,
  - which could lead to a more competitive market,
  - and thus lower prices.

Market-Oriented Cloud Architecture



- Users or brokers acting on user's behalf submit service requests from anywhere in the world to the data center and cloud to be processed.
- They originate the workload that is managed in the cloud datacenter.

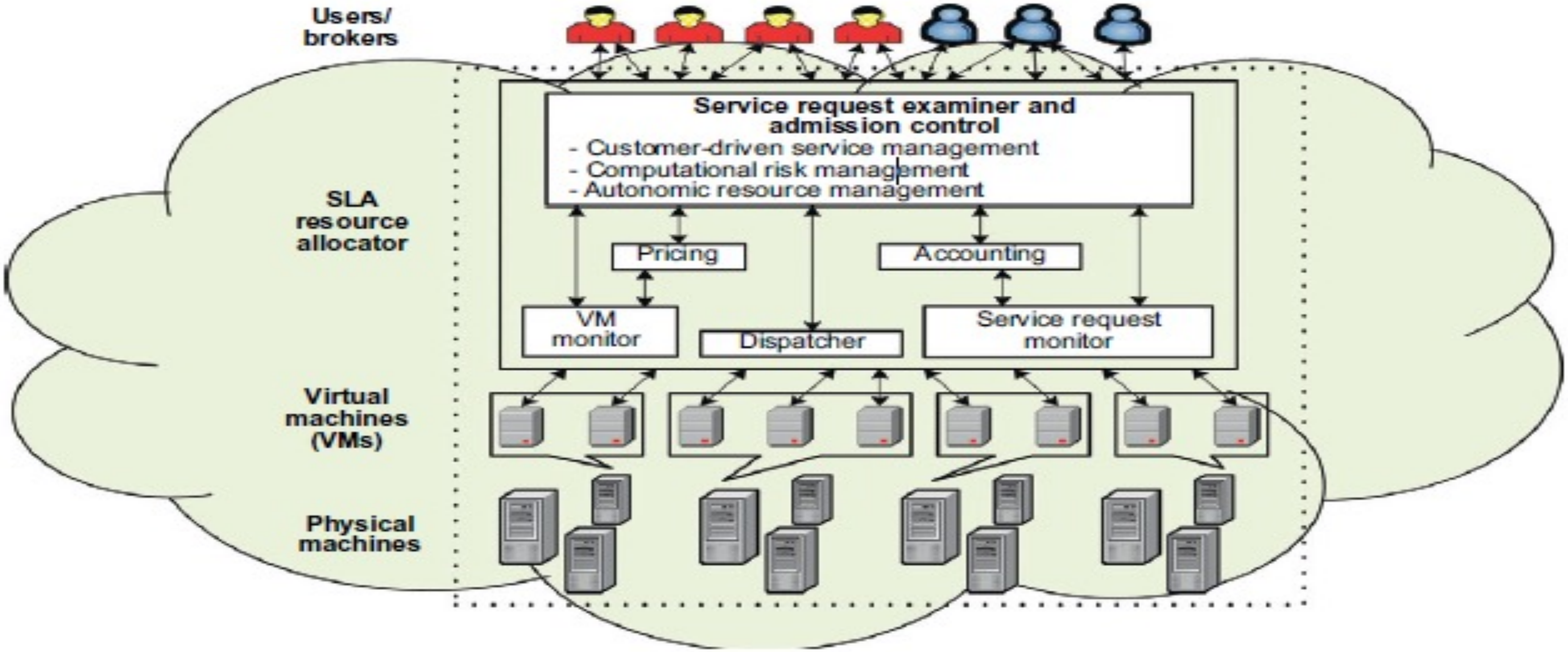
Market-Oriented Cloud Architecture



SLA resource allocator-

- Acts as the <u>interface between the data center, cloud service provider and external users/brokers.</u>
- Its main responsibility is <u>**ensuring that service requests are satisfied according to the SLA agreed to with the user.**</u>
- **Several components** coordinate allocator activities in order to realize this goal

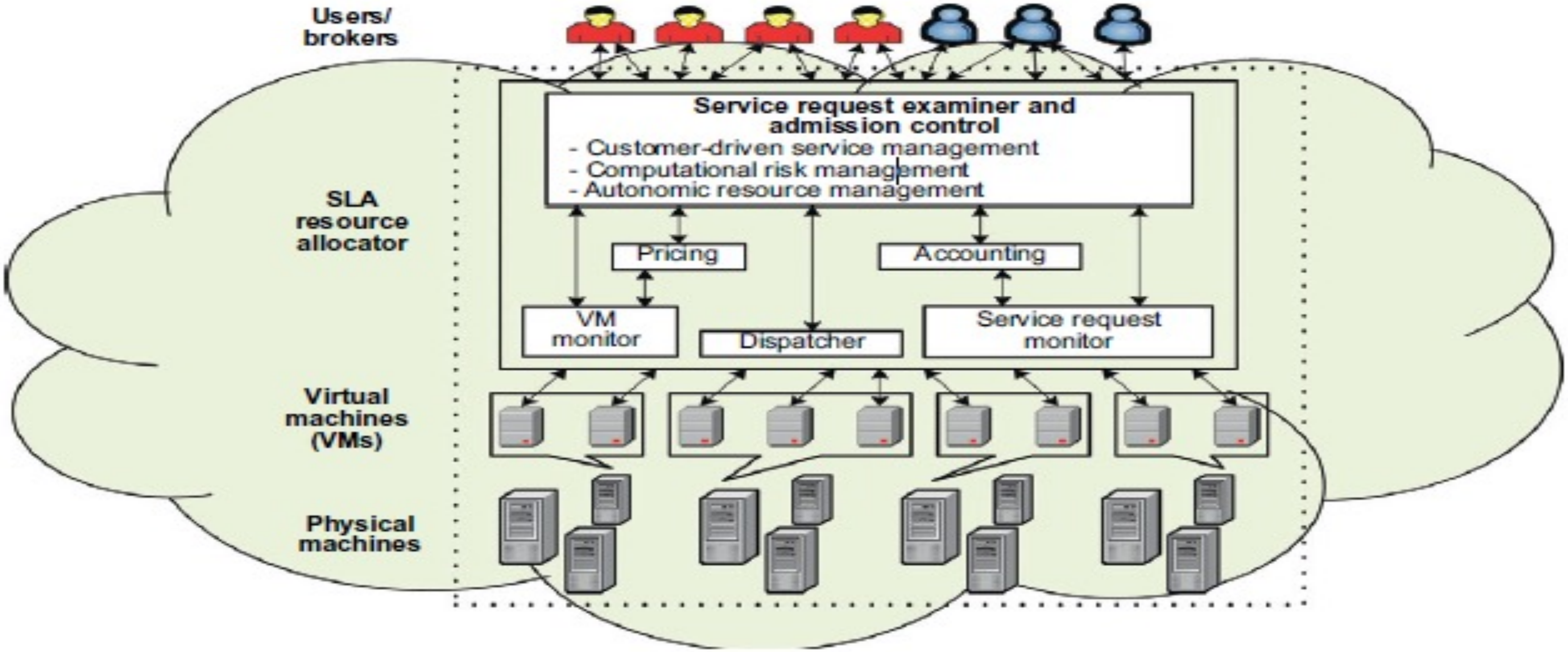
Market-Oriented Cloud Architecture



Service Request Examiner and Admission Control Module-

- When a service request is first submitted
  - the service request examiner interprets the submitted request for QoS requirements
  - before determining whether to accept or reject the request.

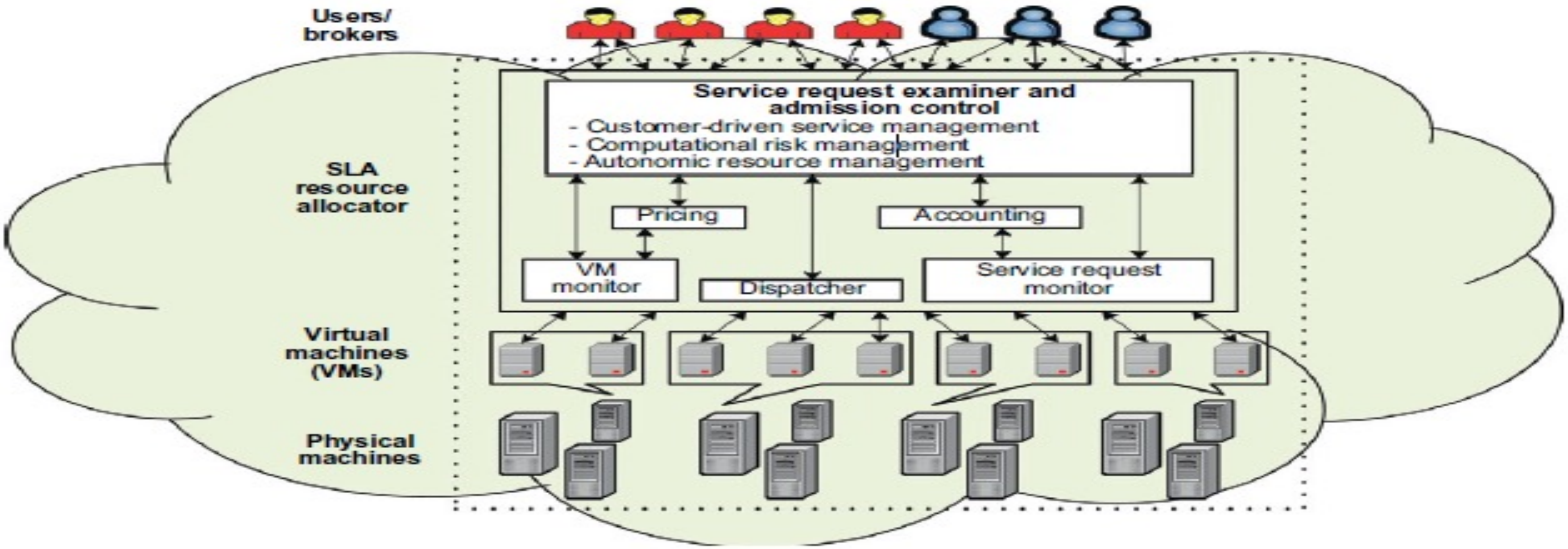
Market-Oriented Cloud Architecture



Service Request Examiner and Admission Control Module-

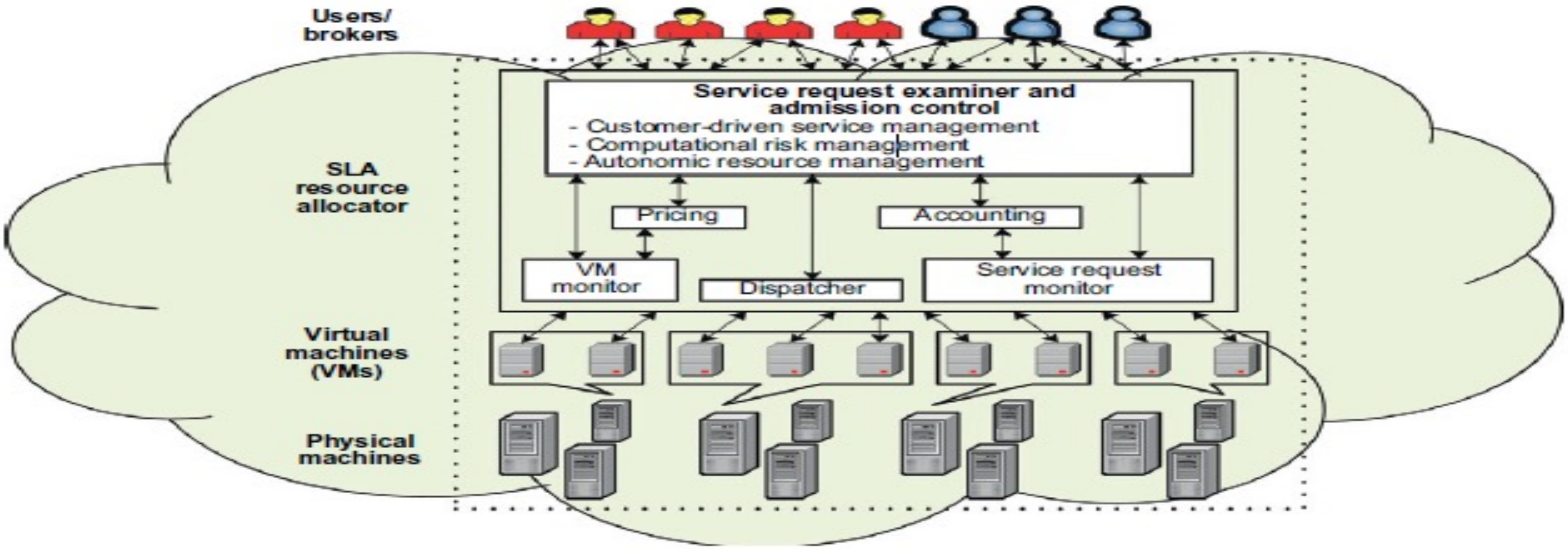
- This module operates in the front-end and **filters user and broker requests in order to accept those that are feasible given the current status of the system and the workload that is already processing.**
- **Accepted requests are allocated and scheduled for execution.**

Market-Oriented Cloud Architecture



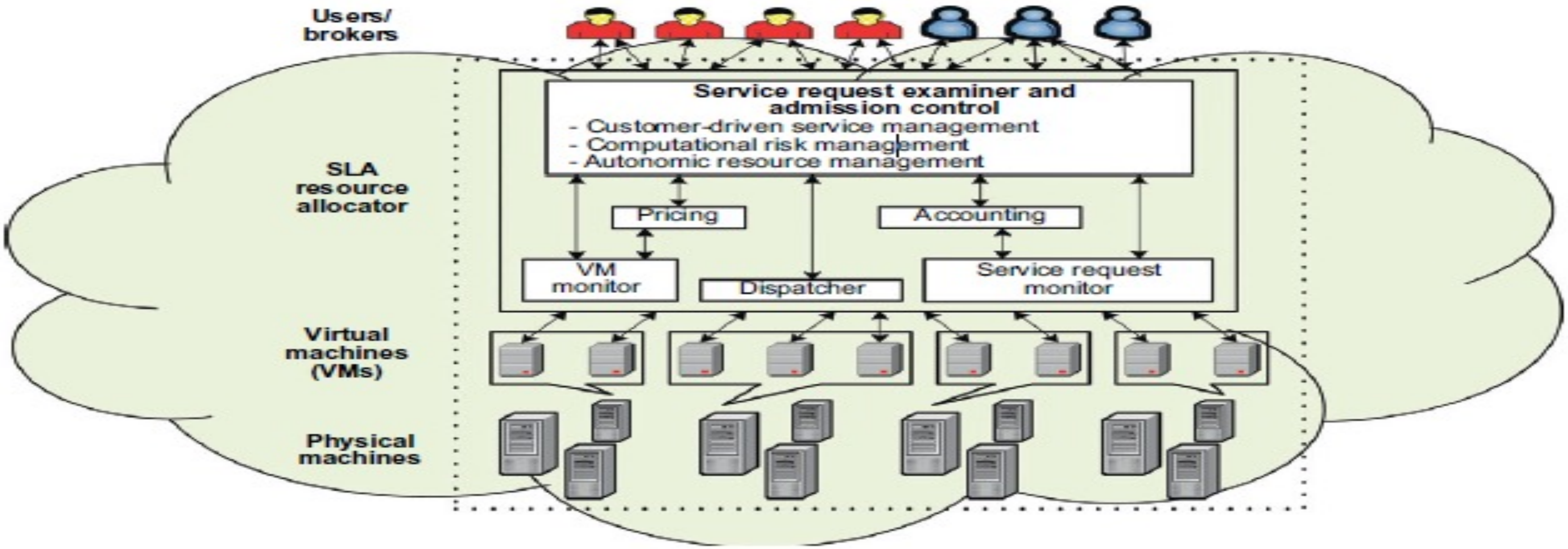
- The Pricing mechanism decides how service requests are charged.
- This module is responsible for **charging users according to the SLA they signed.**
- Different parameters can be considered in charging users; for instance, the most common case for IaaS providers is **to charge according to the characteristics of the virtual machines requested in terms of memory, disk size, computing capacity, and the time they are used.**

Market-Oriented Cloud Architecture



- The Accounting mechanism maintains the actual usage of resources by requests so that the final cost can be computed and charged to users.
- It stores the **billing information for each user.**
- In addition, they constitute a rich source of information that can be mined **to identify usage trends and improve the vendor's service offering.**

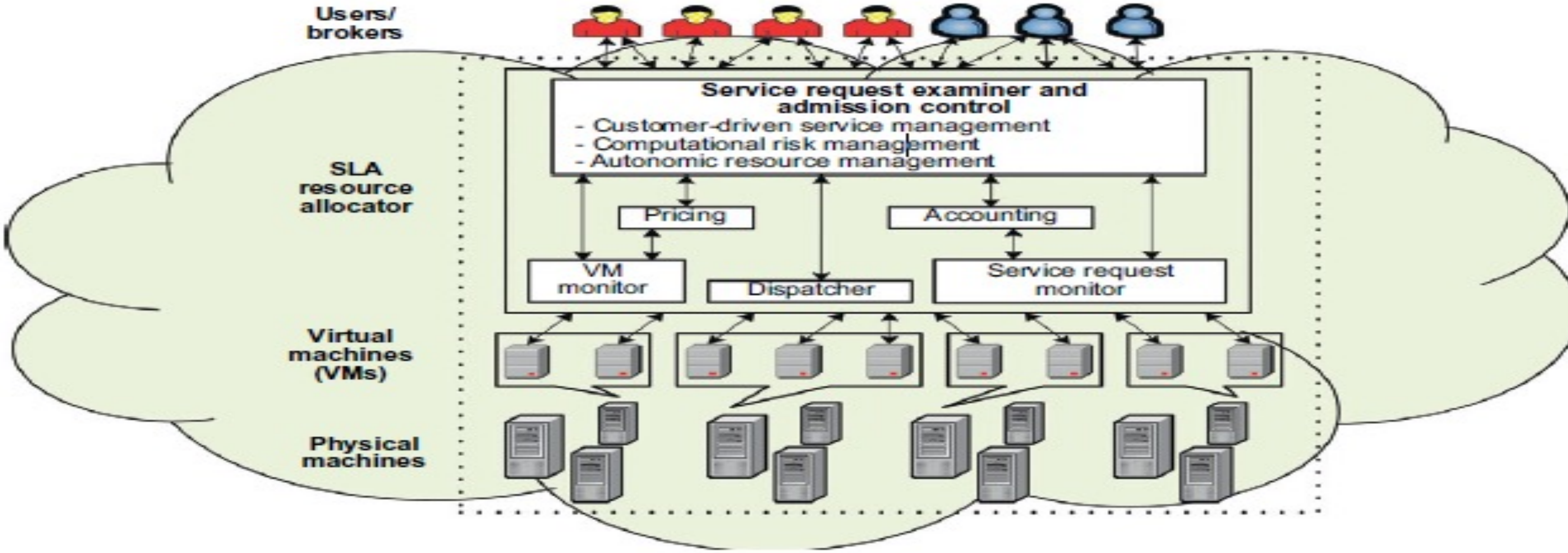
Market-Oriented Cloud Architecture



Dispatcher-

- This component is responsible for the **low-level operations that are required to realize admitted service requests.**
- The Dispatcher mechanism starts the execution of accepted service requests on allocated VMs.

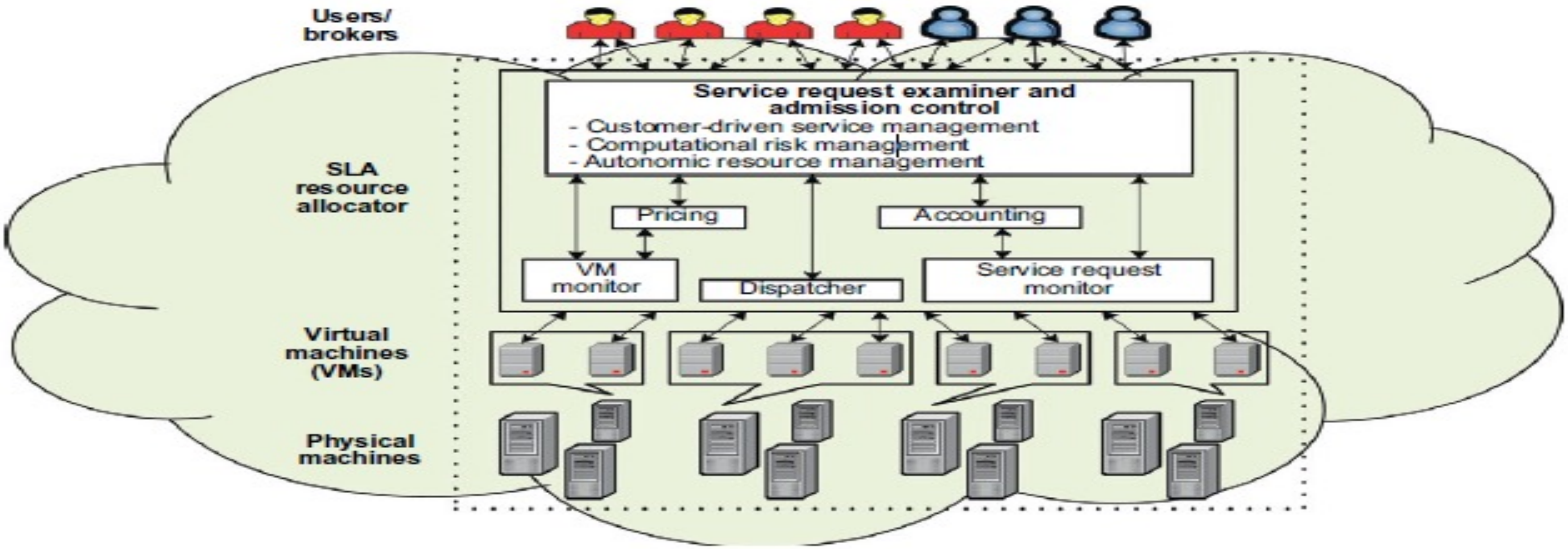
Market-Oriented Cloud Architecture



Dispatcher-

- In an IaaS scenario, this module instructs the infrastructure **to deploy as many virtual machines as are needed to satisfy a user's request.**
- In a PaaS scenario, this module activates and **deploys the user's application on a selected set of nodes;** deployment can happen either **within a virtual machine instance or within an appropriate sandboxed environment.**

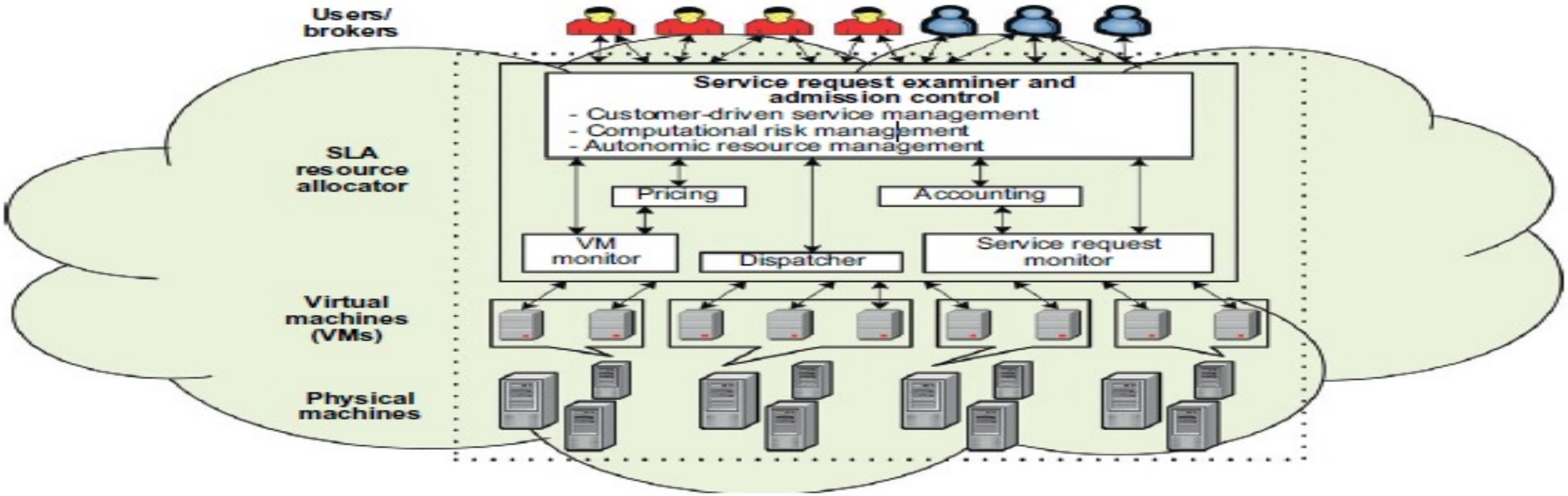
Market-Oriented Cloud Architecture



VM Monitor-

- The VM Monitor mechanism keeps track of the availability of VMs and their resource entitlements.

Market-Oriented Cloud Architecture



- The Service Request Monitor mechanism keeps track of the execution progress of service requests.
- The information collected through the Service Request Monitor is helpful for analyzing system performance and **for providing quality feedback about the provider's capability to satisfy requests.**
- Multiple VMs can be started and stopped on demand on a single physical machine.