Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility

Rajkumar Buyya [a,b,*], Chee Shin Yeo [a], Srikumar Venugopal [a], James Broberg [a], Ivona Brandic [c]

[a] Grid Computing and Distributed Systems (GRIDS) Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Australia
[b] Manjrasoft Pty Ltd, Melbourne, Australia
[c] Institute of Information Systems, Vienna University of Technology, Argentinierstraße 8, 1040 Vienna, Austria

ARTICLE INFO

ABSTRACT

With the significant advances in Information and Communications Technology (ICT) over the last half century, there is an increasingly perceived vision that computing will one day be the 5th utility (after water, electricity, gas, and telephony). This computing utility, like all other four existing utilities, will provide the basic level of computing service that is considered essential to meet the everyday needs of the general community. To deliver this vision, a number of computing paradigms have been proposed, of which the latest one is known as Cloud computing. Hence, in this paper, we define Cloud computing and provide the architecture for creating Clouds with market-oriented resource allocation by leveraging technologies such as Virtual Machines (VMs). We also provide insights on market-based resource management strategies that encompass both customer-driven service management and computational risk management to sustain Service Level Agreement (SLA)-oriented resource allocation. In addition, we reveal our early thoughts on interconnecting Clouds for dynamically creating global Cloud exchanges and markets. Then, we present some representative Cloud platforms, especially those developed in industries, along with our current work towards realizing market-oriented resource allocation of Clouds as realized in Aneka enterprise Cloud technology. Furthermore, we highlight the difference between High Performance Computing (HPC) workload and Internet-based services workload. We also describe a meta-negotiation infrastructure to establish global Cloud exchanges and markets, and illustrate a case study of harnessing 'Storage Clouds' for high performance content delivery. Finally, we conclude with the need for convergence of competing IT paradigms to deliver our 21st century vision.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Computing is being transformed to a model consisting of services that are commoditized and delivered in a manner similar to traditional utilities such as water, electricity, gas, and telephony. In such a model, users access services based on their requirements without regard to where the services are hosted or how they are delivered. Several computing paradigms have promised to deliver this *utility computing* vision and these include cluster computing, Grid computing, and more recently *Cloud computing*. The latter term denotes the infrastructure as a "*Cloud*" from which businesses and users are able to access applications from anywhere in the world on demand. Thus, the computing world is rapidly transforming towards developing software for millions to consume as a service, rather than to run on their individual computers.

At present, it is common to access content across the Internet independently without reference to the underlying hosting infrastructure. This infrastructure consists of data centers that are monitored and maintained around the clock by content providers. Cloud computing is an extension of this paradigm wherein the capabilities of business applications are exposed as sophisticated services that can be accessed over a network. Cloud service providers are incentivized by the profits to be made by charging consumers for accessing these services. Consumers, such as enterprises, are attracted by the opportunity for reducing or eliminating costs associated with "in-house" provision of these services. However, since cloud applications may be crucial to the core business operations of the consumers, it is essential that the consumers have guarantees from providers on service delivery. Typically, these are provided through Service Level Agreements (SLAs) brokered between the providers and consumers.

* Corresponding address: Grid Computing and Distributed Systems (GRIDS) Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, ICT Building, 111, Barry Street, VIC 3053 Melbourne, Victoria, Australia. Tel.: +61 3 83441344.
E-mail addresses: raj@csse.unimelb.edu.au (R. Buyya), csyeo@csse.unimelb.edu.au (C.S. Yeo), srikumar@csse.unimelb.edu.au (S. Venugopal), brobergj@csse.unimelb.edu.au (J. Broberg), ivona@infosys.tuwien.ac.at (I. Brandic).

Providers such as Amazon, Google, Salesforce, IBM, Microsoft, and Sun Microsystems have begun to establish new data centers for hosting Cloud computing applications in various locations around the world to provide redundancy and ensure reliability in case of site failures. Since user requirements for cloud services are varied, service providers have to ensure that they can be flexible in their service delivery while keeping the users isolated from the underlying infrastructure. Recent advances in microprocessor technology and software have led to the increasing ability of commodity hardware to run applications within *Virtual Machines* (VMs) efficiently. VMs allow both the isolation of applications from the underlying hardware and other VMs, and the customization of the platform to suit the needs of the end-user. Providers can expose applications running within VMs, or provide access to VMs themselves as a service (e.g. Amazon Elastic Compute Cloud) thereby allowing consumers to install their own applications. While convenient, the use of VMs gives rise to further challenges such as the intelligent allocation of physical resources for managing competing resource demands of the users.

In addition, enterprise service consumers with global operations require faster response time, and thus save time by distributing workload requests to multiple Clouds in various locations at the same time. This creates the need for establishing a computing atmosphere for dynamically interconnecting and provisioning Clouds from multiple domains within and across enterprises. There are many challenges involved in creating such Clouds and Cloud interconnections.

Therefore, this paper discusses the current trends in the space of Cloud computing and presents candidates for future enhancements of this technology. This paper is primarily divided into two parts. The first part examines current research issues and developments by:

- presenting the 21st century vision of computing and describing various computing paradigms that have promised or are promising to deliver this grand vision (Section 2),
- differentiating Cloud computing from two other widely explored computing paradigms: Cluster computing and Grid computing (Section 3),
- focusing on VM-centric Cloud services and presenting an architecture for creating market-oriented Clouds using VMs (Section 4),
- providing insights on market-based resource management strategies that encompass both customer-driven service management and computational risk management to sustain SLA-oriented resource allocation (Section 5),
- revealing our early thoughts on interconnecting Clouds for dynamically creating global Cloud exchanges and markets (Section 6), and
- comparing some representative Cloud platforms, especially those developed in industries along with our Aneka enterprise Cloud technology (Section 7).

The second part introduces our current work on Cloud computing which include:

- realizing market-oriented resource allocation of Clouds as realized in Aneka enterprise Cloud technology and highlighting the difference between High Performance Computing (HPC) workload and Internet-based services workload (Section 8),
- incorporating a meta-negotiation infrastructure for QoS management to establish global Cloud exchanges and markets (Section 9), and
- creating 3rd party cloud services based on high performance content delivery over commercial cloud storage services (Section 10).

2. The 21st century vision of computing

With the advancement of modern society, basic essential services (*utilities*) are commonly provided such that everyone can easily obtain access to them. Today, utility services, such as water, electricity, gas, and telephony are deemed necessary for fulfilling daily life routines. These utility services are accessed so frequently that they need to be available whenever the consumer requires them at any time. Consumers are then able to pay service providers based on their usage of these utility services.

In 1969, Leonard Kleinrock [1], one of the chief scientists of the original Advanced Research Projects Agency Network (ARPANET) project which seeded the Internet, said: "As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, 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". This vision of the computing utility based on the service provisioning model anticipates the massive transformation of the entire computing industry in the 21st century whereby computing services will be readily available on demand, like other utility services available in today's society. Similarly, computing service users (consumers) need to pay providers only when they access computing services. In addition, consumers no longer need to invest heavily or encounter difficulties in building and maintaining complex IT infrastructure. Hence, software practitioners are facing numerous new challenges toward creating software for millions of consumers to use as a service, rather than to run on their individual computers.

The creation of the Internet has marked the foremost milestone towards achieving this grand 21st century vision of '*computer utilities*' by forming a worldwide system of computer networks that enables individual computers to communicate with any other computers located elsewhere in the world. This internetworking of standalone computers reveals the promising potential of utilizing seemingly endless amount of distributed computing resources owned by various owners. As such, over the recent years, new computing paradigms (shown in Fig. 1) have been proposed and adopted to edge closer toward achieving this grand vision. Applications making use of these utility-oriented computing systems emerge simply as catalysts or market makers, which brings buyers and sellers together. This creates several trillion dollars worth of the utility/pervasive computing industry as noted by Sun Microsystems co-founder Bill Joy [2]. He also indicated "It would take time until these markets to mature to generate this kind of value. Predicting now which companies will capture the value is impossible. Many of them have not even been created yet."

Grid computing [3] enables the sharing, selection, and aggregation of a wide variety of geographically distributed resources including supercomputers, storage systems, data sources, and specialized devices owned by different organizations for solving large-scale resource-intensive problems in science, engineering, and commerce. Inspired by the electrical power Grid's pervasiveness, ease of use, and reliability [4], the motivation of Grid computing was initially driven by large-scale, resource (computational and data)-intensive scientific applications that required more resources than a single computer (PC, workstation, supercomputer) could have provided in a single administrative domain. Due to its potential to make impact on the 21st century as much as the electric power Grid did on the 20th century, Grid computing has been hailed as the next revolution after the Internet and the World Wide Web.

Peer-to-Peer (P2P) computing allows peer nodes (computers) to share content directly with one another in a decentralized manner. In pure P2P computing, there is no notion of clients or servers since all peer nodes are equal and concurrently be both clients and
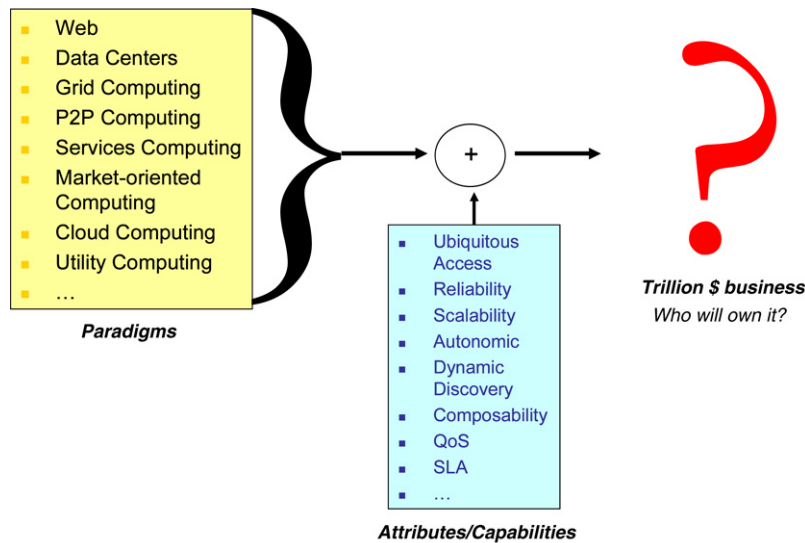
Fig. 1. Various paradigms promising to deliver IT as services.

servers. The goals of P2P computing include cost sharing or reduction, resource aggregation and interoperability, improved scalability and reliability, increased autonomy, anonymity or privacy, dynamism, and ad-hoc communication and collaboration [5].

Services computing focuses on the linkage between business processes and IT services so that business processes can be seamlessly automated using IT services. Examples of services computing technologies include Service-Oriented Architecture (SOA) and Web Services. The SOA facilitates interoperable services between distributed systems to communicate and exchange data with one another, thus providing a uniform means for service users and providers to discover and offer services respectively. The Web Services provides the capability for self-contained business functions to operate over the Internet.

Market-oriented computing views computing resources in economic terms such that resource users will need to pay resource providers for utilizing the computing resources [6]. Therefore, it is able to provide benefits, such as offering incentive for resource providers to contribute their resources for others to use and profit from it, regulating the supply and demand of computing resources at market equilibrium, offering incentive for resource users to back off when necessary, removing the need for a central coordinator (during the negotiation between the user and provider for establishing quality of service expectations and service pricing), and enabling both users and providers to make independent decisions to maximize their utility and profit respectively.

Today, the latest paradigm to emerge is that of Cloud computing [7] which promises reliable services delivered through next-generation data centers that are built on virtualized compute and storage technologies. Consumers will be able to access applications and data from a "Cloud" anywhere in the world on demand. The consumers are assured that the Cloud infrastructure is very robust and will always be available at any time. Computing services need to be highly reliable, scalable, and autonomic to support ubiquitous access, dynamic discovery and composability. In particular, consumers indicate the required service level through Quality of Service (QoS) parameters, which are noted in SLAs established with providers. Of all these paradigms, the recently emerged Cloud computing paradigm appears to be the most promising one to leverage and build on the developments from other paradigms.

3. Definitions, characteristics, and trends

In order to facilitate a clear understanding of what exactly is Cloud computing, we compare Cloud computing with two other recent, widely-adopted or explored computing paradigms: Cluster Computing and Grid Computing. We first examine the respective definitions of these three paradigms, then differentiate their specific characteristics, and finally highlight their recent web search trends.

3.1. Definitions

A number of computing researchers and practitioners have attempted to define clusters, Grids, and Clouds [8] in various ways. Here are some definitions that we think are generic enough to stand the test of time.

The essence of Pfister's [9] and Buyya's [10] work defines clusters as follows:

- "A cluster is a type of parallel and distributed system, which consists of a collection of inter-connected stand-alone computers working together as a single integrated computing resource."

Buyya defined one of the popular definitions for Grids at the 2002 Grid Planet conference, San Jose, USA as follows:

- "A Grid is a type of parallel and distributed system that enables the sharing, selection, and aggregation of geographically distributed 'autonomous' resources dynamically at runtime depending on their availability, capability, performance, cost, and users' quality-of-service requirements."

Based on our observation of the essence of what Clouds are promising to be, we propose the following definition:

- "A Cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers."

At a cursory glance, Clouds appear to be a combination of clusters and Grids. However, this is not the case. Clouds are clearly
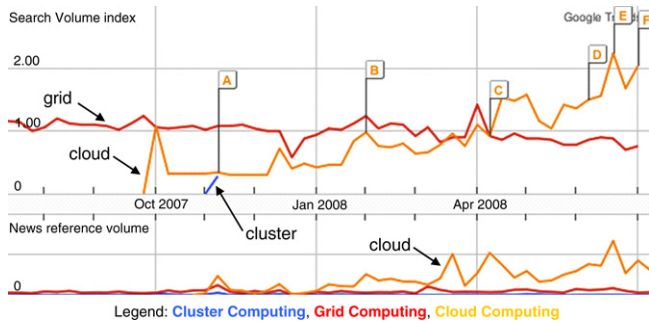
Fig. 2. Google search trends for the last 12 months.

next-generation data centers with nodes "*virtualized*" through hypervisor technologies such as VMs, dynamically "*provisioned*" on demand as a personalized resource collection to meet a specific service-level agreement, which is established through a "*negotiation*" and accessible as a composable service via Web Service technologies such as SOAP and REST.

3.2. Characteristics

A set of characteristics that helps distinguish cluster, Grid and Cloud computing systems is listed in Table 1. The resources in clusters are located in a single administrative domain and managed by a single entity whereas, in Grid systems, resources are geographically distributed across multiple administrative domains with their own management policies and goals. Another key difference between cluster and Grid systems arises from the way application scheduling is performed. The *schedulers* in cluster systems focus on enhancing the overall system performance and utility as they are responsible for the whole system. On the other hand, the schedulers in Grid systems called *resource brokers*, focusing on enhancing the performance of a specific application in such a way that its end-users' QoS requirements are met.

Cloud computing platforms possess characteristics of both clusters and Grids, with its own special attributes and capabilities such strong support for virtualization, dynamically composable services with Web Service interfaces, and strong support for creating 3rd party, value added services by building on Cloud compute, storage, and application services. Thus, Clouds are promising to provide services to users without reference to the infrastructure on which these are hosted.

3.3. Web search trends

The popularity of different paradigms varies with time. The web search popularity, as measured by the Google search trends during the last 12 months, for terms "cluster computing", "Grid computing", and "Cloud computing" is shown in Fig. 2. From the Google trends, it can be observed that cluster computing was a popular term during 1990s, from early 2000 Grid computing become popular, and recently Cloud computing started gaining popularity.

Spot points in Fig. 2 indicate the release of news related to Cloud computing as follows:

A IBM Introduces 'Blue Cloud' Computing, CIO Today — Nov 15 2007.
B IBM, EU Launch RESERVOIR Research Initiative for Cloud Computing, IT News Online — Feb 7 2008.
C Google and Salesforce.com in Cloud computing deal, Siliconrepublic.com — Apr 14 2008.
D Demystifying Cloud Computing, Intelligent Enterprise — Jun 11 2008.
E Yahoo realigns to support Cloud computing, 'core strategies', San Antonio Business Journal — Jun 27 2008.

F Merrill Lynch Estimates "Cloud Computing" To Be $100 Billion Market, SYS-CON Media — Jul 8 2008.

Other more recent news includes the following:

- Yahoo, Intel and HP form Cloud computing labs, Reseller News — Jul 29 2008.
- How Cloud Computing Is Changing The World, Pittsburgh Channel.com — Aug 4 2008.
- SIMtone Corporation Takes Cloud Computing to the Next Level with Launch of First Wireless, "Zero-Touch" Universal Cloud Computing Terminal, TMCnet — Sep 8 2008.

4. Market-oriented Cloud architecture

As consumers rely on Cloud providers to supply more of their computing needs, they will require specific QoS to be maintained by their providers in order to meet their objectives and sustain their operations. Cloud providers will need to consider and meet different QoS parameters of each individual consumer as negotiated in specific SLAs. To achieve this, Cloud providers can no longer continue to deploy traditional system-centric resource management architecture that do not provide incentives for them to share their resources and still regard all service requests to be of equal importance. Instead, market-oriented resource management [11,12] is necessary to regulate the supply and demand of Cloud resources to achieve market equilibrium (where supply = demand), providing feedback in terms of economic incentives for both Cloud consumers and providers, and promoting QoS-based resource allocation mechanisms that differentiate service requests based on their utility. In addition, clients can benefit from the "potential" cost reduction of providers, which could lead to a more competitive market and thus lower prices.

Fig. 3 shows the high-level architecture for supporting market-oriented resource allocation in Data Centers and Clouds. There are basically four main entities involved:

- **Users/Brokers**: Users or brokers acting on their behalf submit service requests from anywhere in the world to the Data Center and Cloud to be processed.
- **SLA Resource Allocator**: The SLA Resource Allocator acts as the interface between the Data Center/Cloud service provider and external users/brokers. It requires the interaction of the following mechanisms to support SLA-oriented resource management:
  ○ Service Request Examiner and Admission Control: When a service request is first submitted, the Service Request Examiner and Admission Control mechanism interprets the submitted request for QoS requirements before determining whether to accept or reject the request. Thus, it ensures that there is no overloading of resources whereby many service requests cannot be fulfilled successfully due to limited resources available. It also needs the latest status information regarding resource availability (from the VM Monitor mechanism) and workload processing (from the Service Request Monitor mechanism) in order to make resource allocation decisions effectively. Then, it assigns requests to VMs and determines resource entitlements for allocated VMs.
  ○ Pricing: The Pricing mechanism decides how service requests are charged. For instance, requests can be charged based on submission time (peak/off-peak), pricing rates (fixed/changing) or availability of resources (supply/demand). Pricing serves as a basis for managing the supply and demand of computing resources within the Data Center and facilitates in prioritizing resource allocations effectively.
  ○ Accounting: The Accounting mechanism maintains the actual usage of resources by requests so that the final cost can

Table 1
Key characteristics of clusters, Grids, and Cloud systems.

Characteristics	Systems		
	Clusters	Grids	Clouds
Population	Commodity computers	High-end computers (servers, clusters)	Commodity computers and high-end servers and network attached storage
Size/scalability	100s	1000s	100s to 1000s
Node Operating System (OS)	One of the standard OSs (Linux, Windows)	Any standard OS (dominated by Unix)	A hypervisor (VM) on which multiple OSs run
Ownership	Single	Multiple	Single
Interconnection network|speed	Dedicated, high-end with low latency and high bandwidth	Mostly Internet with high latency and low bandwidth	Dedicated, high-end with low latency and high bandwidth
Security/privacy	Traditional login/password-based. Medium level of privacy – depends on user privileges.	Public/private key pair based authentication and mapping a user to an account. Limited support for privacy.	Each user/application is provided with a virtual machine. High security/privacy is guaranteed. Support for setting per-file access control list (ACL).
Discovery	Membership services	Centralised indexing and decentralised info services	Membership services
Service negotiation	Limited	Yes, SLA based	Yes, SLA based
User management	Centralised	Decentralised and also virtual organization (VO)-based	Centralised or can be delegated to third party
Resource management	Centralized	Distributed	Centralized/Distributed
Allocation/scheduling	Centralised	Decentralised	Both centralised/decentralised
Standards/inter-operability	Virtual Interface Architecture (VIA)-based	Some Open Grid Forum standards	Web Services (SOAP and REST)
Single system image	Yes	No	Yes, but optional
Capacity	Stable and guaranteed	Varies, but high	Provisioned on demand
Failure management (Self-healing)	Limited (often failed tasks/applications are restarted).	Limited (often failed tasks/applications are restarted).	Strong support for failover and content replication. VMs can be easily migrated from one node to other.
Pricing of services	Limited, not open market	Dominated by public good or privately assigned	Utility pricing, discounted for larger customers
Internetworking	Multi-clustering within an Organization	Limited adoption, but being explored through research efforts such as Gridbus InterGrid	High potential, third party solution providers can loosely tie together services of different Clouds
Application drivers	Science, business, enterprise computing, data centers	Collaborative scientific and high throughput computing applications	Dynamically provisioned legacy and web applications, Content delivery
Potential for building 3rd party or value-added solutions	Limited due to rigid architecture	Limited due to strong orientation for scientific computing	High potential — can create new services by dynamically provisioning of compute, storage, and application services and offer as their own isolated or composite Cloud services to users

be computed and charged to the users. In addition, the maintained historical usage information can be utilized by the Service Request Examiner and Admission Control mechanism to improve resource allocation decisions.

- VM Monitor: The VM Monitor mechanism keeps track of the availability of VMs and their resource entitlements.
- Dispatcher: The Dispatcher mechanism starts the execution of accepted service requests on allocated VMs.
- Service Request Monitor: The Service Request Monitor mechanism keeps track of the execution progress of service requests.

- **VMs**: Multiple VMs can be started and stopped on-demand on a single physical machine to meet accepted service requests, hence providing maximum flexibility to configure various partitions of resources on the same physical machine to different specific requirements of service requests. In addition, multiple VMs can concurrently run applications based on different operating system environments on a single physical machine since every VM is completely isolated from one another on the same physical machine.
- **Physical Machines**: The Data Center comprises multiple computing servers that provide resources to meet service demands.

In the case of a Cloud as a commercial offering to enable crucial business operations of companies, there are critical QoS parameters to consider in a service request, such as time, cost, reliability and trust/security. In particular, QoS requirements cannot be static and may change over time due to continuing changes in business operations and operating environments. In short, there should be greater importance on customers since they

pay for accessing services in Clouds. In addition, the state-of-the-art in Cloud computing has no or limited support for dynamic negotiation of SLAs between participants and mechanisms for automatic allocation of resources to multiple competing requests. Recently, we have developed negotiation mechanisms based on alternate offers protocol for establishing SLAs [13]. These have high potential for their adoption in Cloud computing systems built using VMs.

Commercial offerings of market-oriented Clouds must be able to:

- Support customer-driven service management based on customer profiles and requested service requirements,
- Define computational risk management tactics to identify, assess, and manage risks involved in the execution of applications with regards to service requirements and customer needs,
- Derive appropriate market-based resource management strategies that encompass both customer-driven service management and computational risk management to sustain SLA-oriented resource allocation,
- Incorporate autonomic resource management models that effectively self-manage changes in service requirements to satisfy both new service demands and existing service obligations, and
- Leverage VM technology to dynamically assign resource shares according to service requirements.

5. Resource management strategies for market-oriented Clouds

Since customer satisfaction is the crucial success factor to excel in the service industry [14], computing service providers
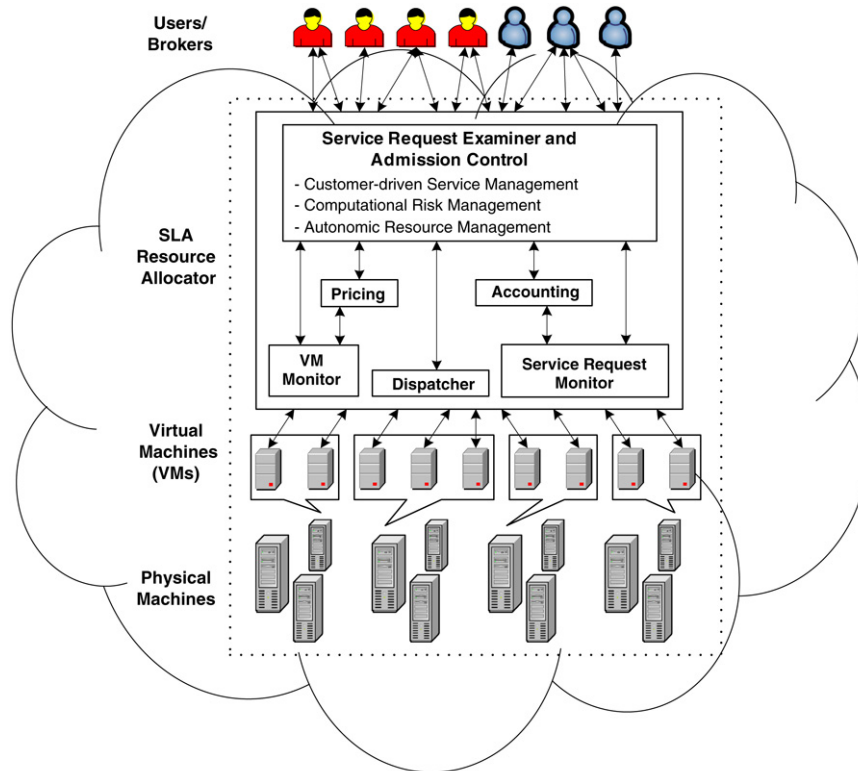
Fig. 3. High-level market-oriented Cloud architecture.

have to be aware of user-centric objectives and meet them in order to achieve customer satisfaction. But, many service quality factors can influence customer satisfaction [14,15]. Hence, we need to design SLA-oriented resource management strategies for Data Centers and Clouds that provide personalized attention to customers, such as enabling communication to keep customers informed and obtain feedback from them, increasing access and approachability to customers, and understanding specific needs of customers. These strategies can also encourage trust and confidence in customers by emphasizing on the security measures undertaken against risks and doubts, the credibility of the provider, and the courtesy towards customers.

Our initial work [16] has also presented examples of how various elements of utility-based resource management can be perceived as risks and hence identified risk analysis from the field of economics as a probable solution to evaluate them. However, the entire risk management process [17,18] comprises of many steps and needs to be studied thoroughly so as to fully realize its effectiveness in managing risks. Hence, we need to first establish the context of risk management in Data Centers and Clouds, and then identify the risks involved. Each of the identified risks will be thoroughly assessed, before deriving appropriate strategies to manage these risks.

In addition, service requirements of users can change over time and thus may require amendments of original service requests. As such, our proposed resource management strategies will be able to self-manage the reservation process continuously by monitoring current service requests, amending future service requests, and adjusting schedules and prices for new and amended service requests accordingly. Hence, we need to investigate self-configuring components to satisfy new service requirements, so that more autonomic and intelligent Data Centers and Clouds can better manage the limited supply of resources with dynamically changing service demand. For users, there can be brokering systems acting on their behalf to select suitable providers

and negotiate with them to achieve ideal service contracts. Thus, providers also require autonomic resource management to selectively choose appropriate requests to accept and execute depending on a number of operating factors, such as the expected availability and demand of services (both current and future), and existing service obligations.

Recently, virtualization [19] has enabled the abstraction of computing resources such that a single physical machine is able to function as a set of multiple logical VMs. A key benefit of VMs is the ability to host multiple operating system environments which are completely isolated from one another on the same physical machine. Another benefit is the capability to configure VMs to utilize different partitions of resources on the same physical machine. For example, on a physical machine, one VM can be allocated 10% of the processing power, while another VM can be allocated 20% of the processing power. Hence, we need to leverage existing VM technologies so that VMs can be started and stopped dynamically to meet the changing demand of resources by users as opposed to limited resources on a physical machine. In particular, we need to investigate how VMs can be assigned various resource management policies catering to different user needs and demands to better support the implementation of SLA-oriented resource allocation for Data Centers and Clouds.

6. Global cloud exchanges and markets

Enterprises currently employ Cloud services in order to improve the scalability of their services and to deal with bursts in resource demands. However, at present, service providers have inflexible pricing, generally limited to flat rates or tariffs based on usage thresholds, and consumers are restricted to offerings from a single provider at a time. Also, many providers have proprietary interfaces to their services thus restricting the ability of consumers to swap one provider for another.

For Cloud computing to mature, it is required that the services follow standard interfaces. This would enable services to be

Fig. 4. Global Cloud exchange and market infrastructure for trading services.

commoditized and thus, would pave the way for the creation of a market infrastructure for trading in services. An example of such a market system, modeled on real-world exchanges, is shown in Fig. 4. The market directory allows participants to locate providers or consumers with the right offers. Auctioneers periodically clear bids and asks received from market participants. The banking system ensures that financial transactions pertaining to agreements between participants are carried out.

Brokers perform the same function in such a market as they do in real-world markets: they mediate between consumers and providers by buying capacity from the provider and sub-leasing these to the consumers. A broker can accept requests from many users who have a choice of submitting their requirements to different brokers. Consumers, brokers and providers are bound to their requirements and related compensations through SLAs. An SLA specifies the details of the service to be provided in terms of metrics agreed upon by all parties, and penalties for meeting and violating the expectations, respectively.

Such markets can bridge disparate Clouds allowing consumers to choose a provider that suits their requirements by either executing SLAs in advance or by buying capacity on the spot. Providers can use the markets in order to perform effective capacity planning. A provider is equipped with a price-setting mechanism which sets the current price for the resource based on market conditions, user demand, and current level of utilization of the resource. Pricing can be either fixed or variable depending on the market conditions. An admission-control mechanism at a provider's end selects the auctions to participate in or the brokers to negotiate with, based on an initial estimate of the utility. The negotiation process proceeds until an SLA is formed or the participants decide to break off. These mechanisms interface with the resource management systems of the provider in order to guarantee the allocation being offered or negotiated can be reclaimed, so that SLA violations do not occur. The resource management system also provides functionalities such as advance reservations that enable guaranteed provisioning of resource capacity.

Brokers gain their utility through the difference between the price paid by the consumers for gaining resource shares and that paid to the providers for leasing their resources. Therefore, a broker has to choose those users whose applications can provide it maximum utility. A broker interacts with resource providers and other brokers to gain or to trade resource shares. A broker is equipped with a negotiation module that is informed by the current conditions of the resources and the current demand to make its decisions.

Consumers have their own utility functions that cover factors such as deadlines, fidelity of results, and turnaround time of applications. They are also constrained by the amount of resources that they can request at any time, usually by a limited budget. Consumers also have their own limited IT infrastructure that is generally not completely exposed to the Internet. Therefore, a consumer participates in the utility market through a resource management proxy that selects a set of brokers based on their offerings. He then forms SLAs with the brokers that bind the latter to provide the guaranteed resources. The enterprise consumer then deploys his own environment on the leased resources or uses the provider's interfaces in order to scale his applications.

The idea of utility markets for computing resources has been around for a long time. Recently, many research projects such as SHARP [20], Tycoon [21], Bellagio [22], and Shirako [23] have come up with market structures for trading in resource allocations. These have particularly focused on trading in VM-based resource slices on networked infrastructures such as PlanetLab. The Gridbus project has created a resource broker that is able to negotiate with resource providers. Thus, the technology for enabling utility markets is already present and ready to be deployed.

However, significant challenges persist in the universal application of such markets. Enterprises currently employ conservative IT strategies and are unwilling to shift from the traditional controlled environments. Cloud computing uptake has only recently begun and many systems are in the proof-of-concept stage. Regulatory pressures also mean that enterprises have to be careful about where their data gets processed, and therefore, are not able to employ Cloud services from an open market. This could be mitigated through SLAs that specify strict constraints on the location of the resources. However, another open issue is how the participants in such a market can obtain restitution in case an SLA is violated. This