Open Group Standard

Service-Oriented Cloud Computing Infrastructure (SOCCI) Framework

Technical Standard

Service-Oriented Cloud Computing Infrastructure (SOCCI) Framework

ISBN: 1-937218-02-7

Document Number: C120

Published by The Open Group, December 2011.

Comments relating to the material contained in this document may be submitted to:

The Open Group, Apex Plaza, Forbury Road, Reading, Berkshire, RG1 1AX, United Kingdom

or by electronic mail to: ogspecs@opengroup.org

Contents

Preface

The Open Group

The Open Group is a global consortium that enables the achievement of business objectives through IT standards. With more than 400 member organizations, The Open Group has a diverse membership that spans all sectors of the IT community – customers, systems and solutions suppliers, tool vendors, integrators, and consultants, as well as academics and researchers – to:

- Capture, understand, and address current and emerging requirements, and establish policies and share best practices
- Facilitate interoperability, develop consensus, and evolve and integrate specifications and open source technologies
- Offer a comprehensive set of services to enhance the operational efficiency of consortia
- Operate the industry's premier certification service

Further information on The Open Group is available at www.opengroup.org.

The Open Group publishes a wide range of technical documentation, most of which is focused on development of Open Group Standards and Guides, but which also includes white papers, technical studies, certification and testing documentation, and business titles. Full details and a catalog are available at www.opengroup.org/bookstore.

Readers should note that updates – in the form of Corrigenda – may apply to any publication. This information is published at www.opengroup.org/corrigenda.

This Document

This document is the Technical Standard for the Service-Oriented Cloud Computing Infrastructure (SOCCI) Framework. It has been developed and approved by The Open Group.

Trademarks

Boundaryless Information Flow™ is a trademark and ArchiMate®, Jericho Forum®, Making Standards Work®, Motif®, OSF/1®, The Open Group®, TOGAF®, UNIX®, and the ``X" device are registered trademarks of The Open Group in the United States and other countries.

The Open Group acknowledges that there may be other brand, company, and product names used in this document that may be covered by trademark protection and advises the reader to verify them independently.

Acknowledgements

The Open Group gratefully acknowledges the contribution of the following people in the development of this document:

Authors

- Tina Abdollah (IBM) – Co-Chair
- Raj Bhoopathi (HP)
- Daniel Biondi (HP)
- Ovace Mamnoon (HP)
- E.G. Nadhan (HP) – Co-Chair
- Tuomas Nurmela (TIETO)
- Vlad Ostrosky (IBM)
- Jignesh Patel (WIPRO)

Reviewers

- Michael Behrendt (IBM)
- Dale Davis (IBM)
- Ed Harrington (Model Driven Solutions)
- Heather Kreger (IBM)
- Birgit Schmidt-Wesche (IBM)
- Joost Van Der Vlies (HP)

The Service-Oriented Cloud Computing Infrastructure Framework evolved from work done by The Open Group SOA Reference Architecture and Cloud Computing Reference Architecture projects, and the former Open Group Service-Oriented Infrastructure project. The Open Group gratefully acknowledges the contributions of these projects.

Referenced Documents

The following documents are referenced in this Technical Standard:

[1] Let it Rise: A Special Report on Corporate IT, Economist (p.4), October 25, 2008.

[2] Clearing the Air on Cloud Computing, McKinsey & Co. Report, April, 2009.

[3] White Paper: Avoiding Storm "Clouds" with ITIL, Daniel Biondi & Brian Stephenson-Roberts, Hewlett Packard, September 2010.

[4] Definition of Cloud Computing, National Institute of Standards and Technology (NIST), available from: www.nist.gov/itl/cloud.

1 Introduction

Infrastructure is a foundational element for enterprise architecture. Infrastructure has been traditionally provisioned in a physical manner. With the evolution of virtualization technologies and application of service-orientation to infrastructure, it can now be offered as a service.

Service-orientation principles originated in the business and application architecture arena. After repeated, successful application of these principles to application architecture, IT has evolved to extending these principles to the infrastructure.

Thus, came about the concept of Infrastructure as a Service (IaaS).

An enabling framework of a well-defined, integrated set of service-oriented components is essential for infrastructure to be provided as a service.

Service-Oriented Cloud Computing Infrastructure (SOCCI) is the realization of this framework for the cloud. This document details the SOCCI elements, the synergies realized through the cohesive application of SOA and cloud-based principles, and the SOCCI Management Building Blocks. It expands upon the relationships between service-orientation and its application to various infrastructure components. Finally, the concepts outlined are explained in the context of a business scenario – Motor Cars in the Cloud.

1.1 Target Audience

The target audience for this specification includes a number of different communities:

- Executive sponsors for The Open Group Service-Oriented Architecture Work Group as well as The Open Group Cloud Computing Work Group
- Architects including enterprise architects, SOA architects, and infrastructure architects
- IT strategists and lead designers
- Standards organizations as a foundation for other standards and to apply to industry domains

1.2 Value Statement

A recent analysis of cloud computing in the *Economist* stated that there were a plethora of data centers worldwide and estimated that 7,000 data centers existed in America alone. Most of these data centers were one-off designs that had grown over the years. Many surveys show that these data centers are highly inefficient [1]. According to a study by McKinsey (a consultancy) and the Uptime Institute (a think tank), on average only 6% of server capacity is used. Nearly 30% is no longer in use at all and many organizations are unaware of which application is running on which server. This has created tremendous waste and impact on the environment [2]. Additionally, the economic turmoil and globalization have mandated that organizations cut costs and promote their human and financial resources on processes that can rapidly and easily be automated and managed. There are downsizing and rightsizing options to cut cost and be agile, which lead to the cloud-sizing solution.

The process of replacing existing in-house services and resources with cloud-based resources and services holds many benefits, such as lower costs, scalability, flexibility, freedom from software maintenance, and pricing structures that are a continuous expense rather than a larger up-front capital expense. Additionally, a cloud-based solution can promote accelerated business growth through server utilization optimized in real time, improved availability of data, and reduced power consumption. Given the potential benefits that cloud computing might bring to the business, it has become the latest technology buzz. The rapid adoption of distributed server virtualization is accelerating the deployment of high-utilization virtual machines. The relationship of a virtualized infrastructure to cloud computing has recently attracted a multitude of cloud vendors to the market, creating confusion for the consumers about the options available to them and the choices they need to make to derive the benefits of cloud computing for their business.

In order to consume and use cloud computing resources to their best advantage, it is incumbent upon business owners and executives, IT managers, start-up founders, and senior developers to understand the options that cloud computing provides. Additionally, there are also many perceived and some real risks associated with cloud computing which need to be carefully evaluated prior to adoption. To provide insight and clarity on cloud infrastructure, this specification:

1. Defines an open, industry-standard, service-oriented, and vendor-neutral reference framework for cloud computing infrastructure that a variety of cloud users can leverage and benefit from
2. Specifies the architecture building blocks that enable the framework
3. Supports various user viewpoints from consumer, provider, and integrator perspectives
4. Provides real-life scenarios that support SOCCI

1.3 Terminology

Can
: Describes a permissible optional feature or behavior available to the user or application. The feature or behavior is mandatory for an implementation that conforms to this document. An application can rely on the existence of the feature or behavior.

Implementation-dependent
: (Same meaning as "implementation-defined".) Describes a value or behavior that is not defined by this document but is selected by an implementer. The value or behavior may vary among implementations that conform to this document. An application should not rely on the existence of the value or behavior. An application that relies on such a value or behavior cannot be assured to be portable across

conforming implementations. The implementer shall document such a value or behavior so that it can be used correctly by an application.

Legacy	Describes a feature or behavior that is being retained for compatibility with older applications, but which has limitations which make it inappropriate for developing portable applications. New applications should use alternative means of obtaining equivalent functionality.
May	Describes a feature or behavior that is optional for an implementation that conforms to this document. An application should not rely on the existence of the feature or behavior. An application that relies on such a feature or behavior cannot be assured to be portable across conforming implementations. To avoid ambiguity, the opposite of “may” is expressed as “need not”, instead of “may not”.
Must	Describes a feature or behavior that is mandatory for an application or user. An implementation that conforms to this document shall support this feature or behavior.
Shall	Describes a feature or behavior that is mandatory for an implementation that conforms to this document. An application can rely on the existence of the feature or behavior.
Should	For an implementation that conforms to this document, describes a feature or behavior that is recommended but not mandatory. An application should not rely on the existence of the feature or behavior. An application that relies on such a feature or behavior cannot be assured to be portable across conforming implementations. For an application, describes a feature or behavior that is recommended programming practice for optimum portability.
Undefined	Describes the nature of a value or behavior not defined by this document that results from use of an invalid program construct or invalid data input. The value or behavior may vary among implementations that conform to this document. An application should not rely on the existence or validity of the value or behavior. An application that relies on any particular value or behavior cannot be assured to be portable across conforming implementations.
Unspecified	Describes the nature of a value or behavior not specified by this document that results from use of a valid program construct or valid data input. The value or behavior may vary among implementations that conform to this document. An application should not rely on the existence or validity of the value or behavior. An application that relies on any particular value or behavior cannot be assured to be portable across conforming implementations.
Will	Same meaning as “shall”; “shall” is the preferred term.

2 Cloud Computing Characteristics

The salient characteristics of cloud computing based on the definitions provided by the National Institute of Standards and Terminology (NIST) are outlined below:

- **On-demand self-service**: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.
- **Broad network access**: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
- **Resource pooling**: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location-independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data center). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.
- **Rapid elasticity**: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
- **Measured service**: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be managed, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

3 Service-Orientation Characteristics

The basic characteristics of Service-Oriented Infrastructure (SOI) are listed below:

- **Business-driven infrastructure on-demand**: Using service-orientation for the infrastructure provides a way to define dependencies of higher-level business services on the lower-level infrastructure services, down to the actual physical resources, such as network appliances, storage, and servers.
- **Operational transparency**: Clear visibility into the operations of the infrastructure services to enable problem diagnosis, root cause analysis, and the potential impact of infrastructure availability on business services.
- **Service measurement**: Service-orientation defines service-level objectives as well as the measurement of the delivery of those objectives.
- **Consumer provider model**: Service-orientation introduces the consumer provider model. In this model, from the "Cloud Service" consumer perspective, the risk of providing, maintaining, and managing the service is significantly shifted away from the consumer. The provider needs to ensure that the requisite infrastructure is in place to meet demand.

4 Extending SOI to SOCCI

Figure 1 illustrates the mapping of various Service-Oriented Infrastructure (SOI) layers with the layers represented within the SOA Reference Architecture (SOA RA). In particular, the services realized within the SOI directly enable the Operational Systems Layer of the SOA RA.

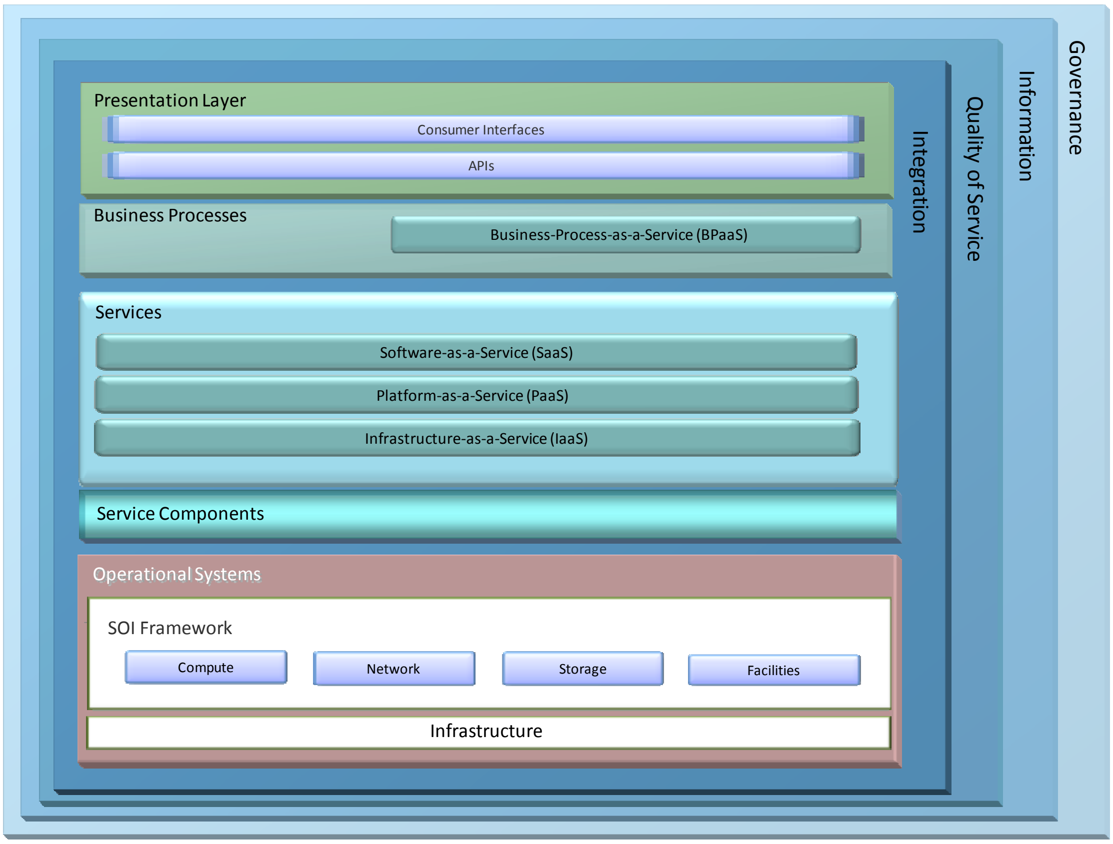


Figure 1: SOI and SOA RA Layer Mapping

Cloud computing puts new demands on the IT infrastructure and management thereof. It requires an abstract approach to the operational environment. A cloud computing provider cannot any longer tailor its environment for each subscriber. It means that instead of a physical device, cloud computing offers an abstraction of a server, file system, storage, network, database, etc. Moreover, increasing providers' profitability and maximizing the utilization of resources requires multi-tenancy, dynamic allocation of resources, and metering with charge-back.

At the same time, subscribers expect to see implementation of a utility model since they want to allocate resources on-demand and pay exactly for their usage while being able to sustain their operations, much like the electric bill. Hence, new infrastructure should be agile and elastic and create an illusion of infinite computing resources available on-demand. While SOI did not offer the whole spectrum of the characteristics desired, it became an enabler for what came to be known as Service-Oriented Cloud Computing Infrastructure (SOCCI). SOCCI can be defined as service-oriented, utility-based, manageable, scalable on-demand infrastructure that supports essential cloud characteristics, service, and deployment models. In other words, SOCCI describes the essentials for implementing and managing an Infrastructure as a Service (IaaS) environment.

Architecturally, SOCCI is a foundation of SOI and the cloud (Figure 2) with a focus on requirements such as metering, chargeback, or virtual management.

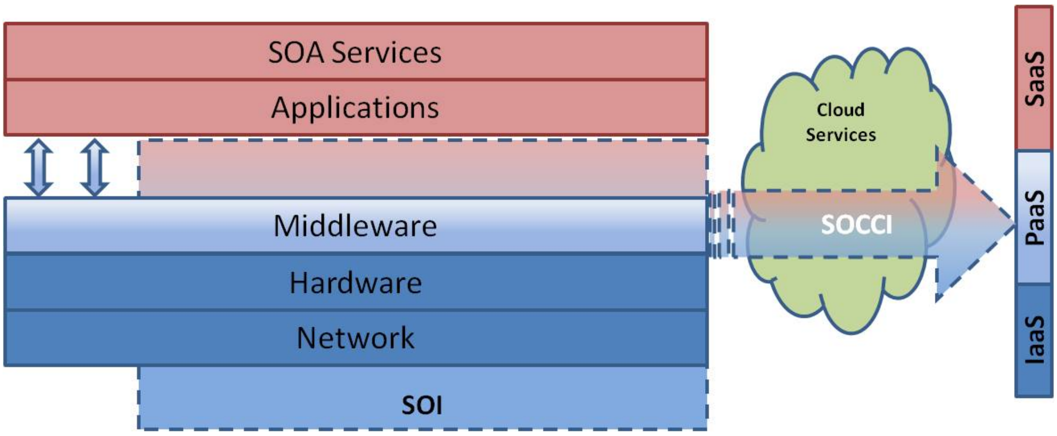


Figure 2: SOCCI is SOI Adoption for Cloud

SOCCI is based on SOI's ability to leverage virtualization technology and expose infrastructure services while preserving all SOI properties. However, cloud added new characteristics which simply weren't always necessary in the case of SOI. Table 1 contains a comparative analysis between different types of IT infrastructures – traditional, SOI, and SOCCI. Please note that the SOCCI values in this table focus on the aspects that are common to all cloud deployment models.

Traditional	**SOI**	**SOCCI**
Manually provisioned	Manually provisioned	Self-provisioned
Dedicated hardware	Dedicated or shared hardware	Shared hardware
Fixed capacity	Elastic capacity	Elastic capacity
Pay for capacity	Pay for capacity or for use	Pay for use
Capital and operational expenses	Capital and operational expenses	Operational expenses
Managed via system admin.	Managed via system admin. and/or API or some automation	Managed via APIs or automation
Static environment	Dynamic environment	Dynamic environment
Scheduled configuration	On-demand resource usage	On-demand resource usage

Traditional	SOI	SOCCI
Leverage middleware capabilities	Expose infrastructure services	Expose infrastructure services
Contract-based payment	Contract-based payment with elements of metering	Metering with charge-back along with contract-based payments determined by usage with predefined thresholds

Table 1: Comparison between Traditional, SOI, and SOCCI Environments

5 Service-Orientation and Cloud Synergies

The combination of Service-Oriented Infrastructure (SOI) and cloud computing realize unique characteristics as listed below:

- **XaaS**: This is the essence of cloud computing. It refers to an increasing number of services that are delivered over a network. Anything as a service requires an understanding of the service objectives and the accounting of service use and quality. The objectives, use, and quality can be determined from the underlying reference model for SOI:

  Broad network access (cloud) + resource pooling (cloud) + business-driven infrastructure on-demand (SOI) + service-orientation (SOI) = **XaaS**

- **Service assurance**: Cloud computing guarantees certain levels of service to the cloud's customers. When that service degrades, it is necessary to understand the relationship of infrastructure activity to these services so that the situation can be remediated. SOI facilitates the determination of these relationships:

  Operational transparency (SOI) + measured service (cloud) = **Service Assurance**

- **Next-generation scaling**: Demand-driven infrastructure enabled by SOI triggers various forms of scaling supported by the cloud, including scale-out and scale-in supplemented by scale-up and scale-down through the SOI:

  Rapid elasticity (cloud) + consumer provider model (SOI) = **Next-Generation Scaling**

- **Operational and tactical control**: The combination of SOI and the cloud computing space realize effective instrumentation of pay-for-use models as well as governance mechanisms controlled by rules leading to policy-based administration:

  On-demand self-service (cloud) + service measurement (SOI) + operational transparency (SOI) = **Operational and Tactical Control**

5.1 Comparing SOCCI and XaaS

X as a Service (XaaS) is a generalization for cloud-related services; i.e., those services that comprise the cloud service model per the NIST definition [4]. The most common examples of XaaS are Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS). The combined use of these three is sometimes referred to as the SPI model (SaaS, PaaS, IaaS). These are "the most common examples" because the usage of the "as a service" moniker is proliferating. As of today, anything can be provided "as a service": database as a service, information as a service, etc.

First of all let's take a look at the IaaS definition [4]. NIST defines IaaS as follows:

"The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems; storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls)."

As this definition implies, IaaS is a service which provides capabilities for the consumer to provision certain infrastructure elements. Meantime, SOCCI is the "underlying cloud infrastructure" which enables IaaS to provide its functionality.

IaaS may entail the provisioning of multiple components including the servers for on-demand computing power, facilities for robust web-hosting, and elastic storage.

There is a striking similarity to the manner in which services are provided in the applications space. Let us say that there is a *GetClientAddress* web service that retrieves the address information from a back-end database that stores the HR information within an organization. The existence of the enabling components like the database and web/application servers/integration brokers is transparent to the consumer of this service. In a similar fashion, the enabling infrastructure (SOCCI) is not exposed to the consumer of IaaS; instead IaaS is provisioned on top of SOCCI.

The relationship between SOCCI and XaaS is shown in Figure 2.

5.2 Comparing SOCCI and SOA

Service-Oriented Architecture (SOA) is an architectural style for building loosely-coupled systems. It enables the implementation of a technology-based solution without exposing the underlying detail to its consumers. Cloud computing (see Chapter 2) is an SOA and depends on some common attributes in The Open Group SOA Reference Architecture (SOA RA). Given that SOCCI supports both SOA and cloud computing infrastructure, it provides a physical infrastructure that can be leveraged by both worlds, which is a multi-tenant, virtualized infrastructure that leverages SOA principles and the infrastructure components (e.g., server, network, devices, etc.) as the building blocks to facilitate an agile and re-usable infrastructure over the Internet.

6 Enabling Applicable Architectural Layers

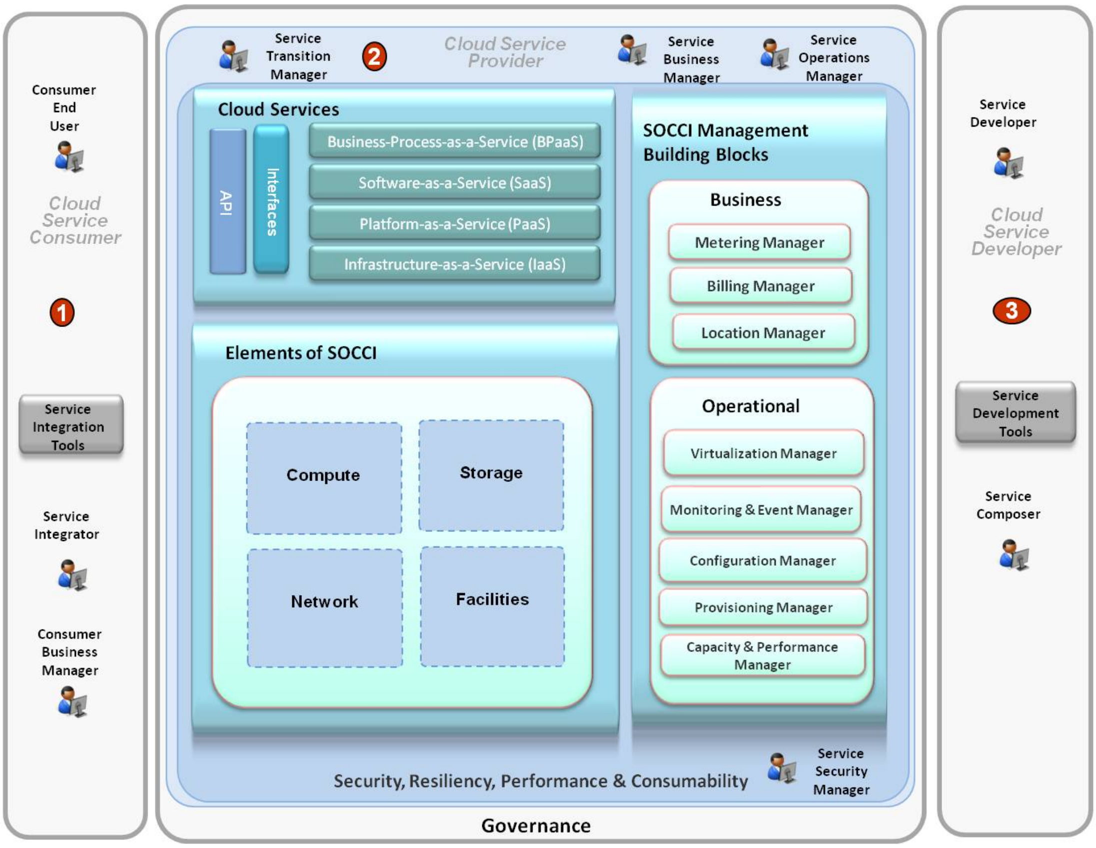


Figure 3: SOCCI High-Level View

SOCCI Architecture Building Blocks (ABBs) consist of the SOCCI elements and SOCCI Management Building Blocks. The former leverages the latter to interface with the users. The SOCCI Management Building Blocks are categorized as operation-centric and business-centric. The division between the two is dependent on the type of service provided. The Business ABBs relate to supporting the business functions needed to be able to create, support, and consume an IaaS offering; e.g. metering, billing, and location management. The Operational ABBs relate to operational functions supporting/leveraging the SOCCI elements to provide the operational management services; e.g., virtualization and provisioning management.

SOCCI Management Building Blocks that enable SOCCI are described below.

Business ABBs:

- **Metering Manager** (measured service enabler) tracks consumer usage statistics for the usage of virtual networks and resources from aggregated data including usage statistics and number of network devices deployed. It also provides aggregation of metering and multi-tenancy-aware measurement metering information on different levels of aggregates (routers, FWs/IDS, WANs, etc.).

— **Aggregation of measuring** covers a single element of SOCCI metering (i.e., storage, network, or facilities) measurement aggregation.

— **Multi-tenancy-aware measuring** provides per-tenant measurements, for a subset view of aggregation of measuring.

- **Billing Manager** often relies on the Metering Manager to gather the usage data so that it can issue applicable billing charges to the consumer based upon contractual agreements, determined by the tenant-level aggregated usage of virtual network and resource metering information. Based on the type of contract, each tenant may have different charges including fixed charges, discounts, coupons, etc. that the billing manager needs to account for. It will also help to settle charges with external service providers that provide additional services as part of virtual networks. The Billing Manager manages the generation of a summary of charges for a given billing cycle and conveying charges and payments made/due.

- **Location Manager** helps to geographically allocate resources based on predefined rules, regulations, cost constraints, and Service-Level Agreements (SLAs).

Operational ABBs:

- **Virtualization Manager** (resource pooling enabler) emulates a physical infrastructure component that can be used and managed by the cloud consumer. Virtualization Manager acts as the façade and manager for physical infrastructural elements of SOCCI within the infrastructure.

- **Monitoring & Event Manager** (measured service enabler) monitors the different services for specific events and trends of usage, detecting conflicts and failures within the infrastructure. They correlate to respective network and resource assets, ensuring compliance, reporting problems, as well as maintaining system recoverability and audit-ability for overall functional integrity of the infrastructure. Monitoring & Event Manager supports all infrastructure elements; i.e., storage, network, server, and facilities.

  — **Event monitoring** is simply to indicate the need to propagate or clear an incident.

  — **Trend of the usage** refers to different trending patterns, whether once-a-week, monthly, or yearly.

  It also measures MTBF (Mean Time Between Failures) and MTTR (Mean Time to Repair), with configurable alerts that are used to trigger notifications and escalations depending upon the severity levels.

- **Provisioning Manager** (rapid elasticity and on-demand self-service enabler) ensures that the right number of resources are allocated to address fluctuating demands as well as interacts with the virtualization and platform management component to provision and deploy services within the infrastructure to optimize infrastructure resources as well as satisfying the customer's requirements for the requested cloud services.

- **Capacity & Performance Manager** (resource pooling enabler) ensures that the optimal numbers of resources are available to satisfy the requests made by the Provisioning Manager to support the performance specified by the consumer. It measures and reports the discrete resources within a cloud environment to ensure optimal and committed performance levels. It also has the responsibility to manage performance of the tenants in accordance with the tenant's SLAs and entitlements and track where resources are added/removed based on operating rules.
- **Configuration Manager** provides the versioning and configuration support of devices (virtual and physical) within the infrastructure. It also supports the Provisioning and Monitoring & Event Managers within the infrastructure to ensure the overall functional integrity of the infrastructure.

Figure 4 depicts the SOCCI ABBs. The figure does not include expected base service-oriented computing ABBs (service registry for registration and publishing of service; service orchestrator/composer/choreographer or service event manager for subscription or publication of events), because the availability of SOA ABBs is implied here.

Here, we will only discuss the Management Building Blocks directly related to infrastructure that support SOCCI to enable its functionality. Therefore, some broadly used cloud and IT service management-related components, such as service catalog and compliance policy manager, are excluded from the list of cloud components.

The Operational ABBs (enablers), such as Virtualization Manager, have more focused attributes, whereas the Business ABB enablers usually support more attributes and enable more extensive activities for the users.

Additionally, operational building blocks are usually enabled by cost-effective technologies supplemented by service-centric approaches on an as-needed basis.

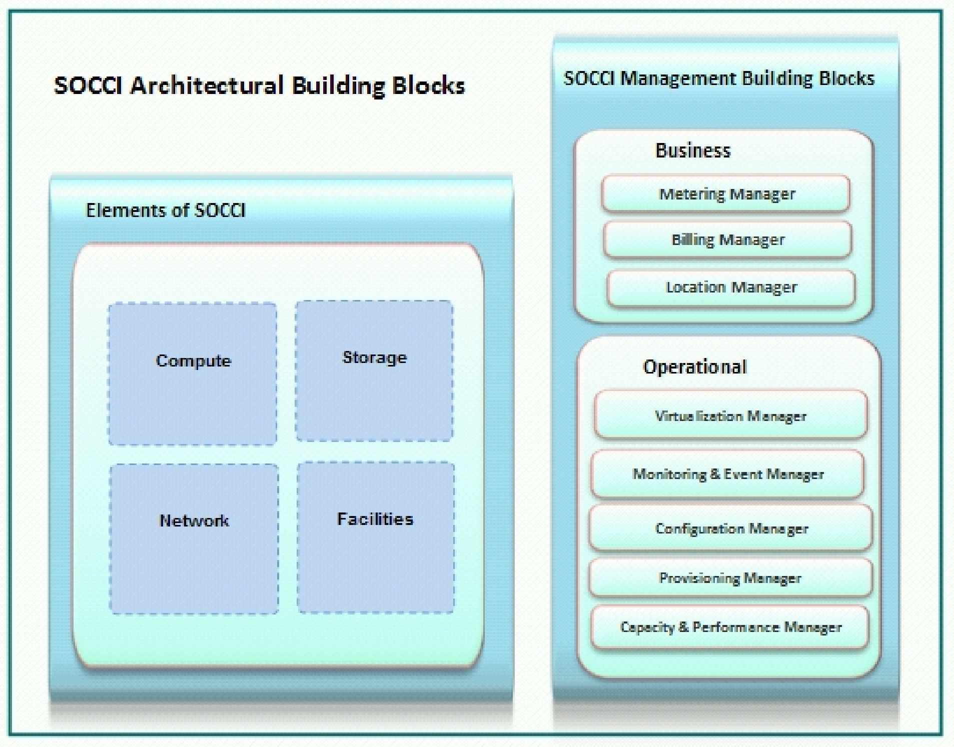


Figure 4: SOCCI Architecture Building Blocks (ABBs)

7 SOCCI Elements

This chapter defines the elements of the Service-Oriented Cloud Computing Infrastructure (SOCCI).

Virtualization enables cloud infrastructure. It has provided an opportunity to reverse the resulting server sprawl, to consolidate (e.g., to run) many applications upon a single physical hardware component, driving utilization higher than before but retaining the separation of operating systems and application environment required to allow ready configuration and maintenance for each application as though it were still physically separate.

Create an infrastructure which operates like a utility – i.e., find out how many CPU cycles, how much RAM, and how much disk space is required and deliver it without depending on what type of hardware is supporting it.

Cloud computing infrastructure includes but is not limited to four kinds of resources; one or more of these resources can be provisioned as a service depending upon the needs of the consumer:

- Compute – Information Processing Resources
    - Provide Internet-based computing to represent the fundamentals necessary to process instructions, store such instructions and the results of their computation, and to transmit these internally between subcomponents or externally to other devices over networks spanning great distances.
    - Some examples of the computing elements are a set of VMs and operating system virtualization (Xen, KVM, VMWare, HyperV)
- Network – Interconnect Resources
    - Cloud computing networks:
        - Burst up and turn down bandwidth on-demand.
        - Allow for connections between servers to enable automated movement of virtual machines.
        - The underlying network for the cloud can be very dynamic in nature, integrating many different value-add entities across different routes that are optimized in real time. Though transparency to all this complexity is desired from the consumer perspective, to manage and administer this network visibility is also essential. A network map at any point in time allows visibility into the interconnects and possible issues, constraints, or bottlenecks, etc. in the cloud.
    - Cloud computing network can be seen in three independent structures:
        - Front end – connects users to application

        - Horizontal aspect – physical servers and movement of their virtual machines
        - Storage networks

- Storage – Recorded Information Resources

  — This infrastructure enables the storage and retrieval of any amount of data, at any time, from anywhere on the cloud. It conforms to a utility model while supporting elastic requirements allocating and releasing storage on-demand.

  — Storage should also support the recoverability of data by leveraging horizontal partitioning to meet service-level requirements.

- Facilities – Housing & Facilitating Computing Resources

  — The cloud computing hosting facility has to be situated in a secure and well-guarded location with all the appropriate best practices in place for building and safety standards.

  — Connectivity to the resources needs to be open so that resources external to the facility can leverage the resources within the facility on an as-and-when-needed basis.

  — The network bandwidth in the facility should match up to the demands of the cloud computing requirement of elasticity that is to scale up or down at any given moment.

  — Another requirement would be to establish disaster recovery plans that can be executed annually to assure all the stakeholders that there will be continuity of computing resources if and when such circumstances arise.

8 Viewpoints

This chapter describes the participants involved in the cloud computing model and their different perspectives. Participants have specific requirements from the cloud computing infrastructure. We will be elaborating on these requirements. The relationship of the base roles (Consumer, Provider, and Developer) and composite role (Integrator) are described below.

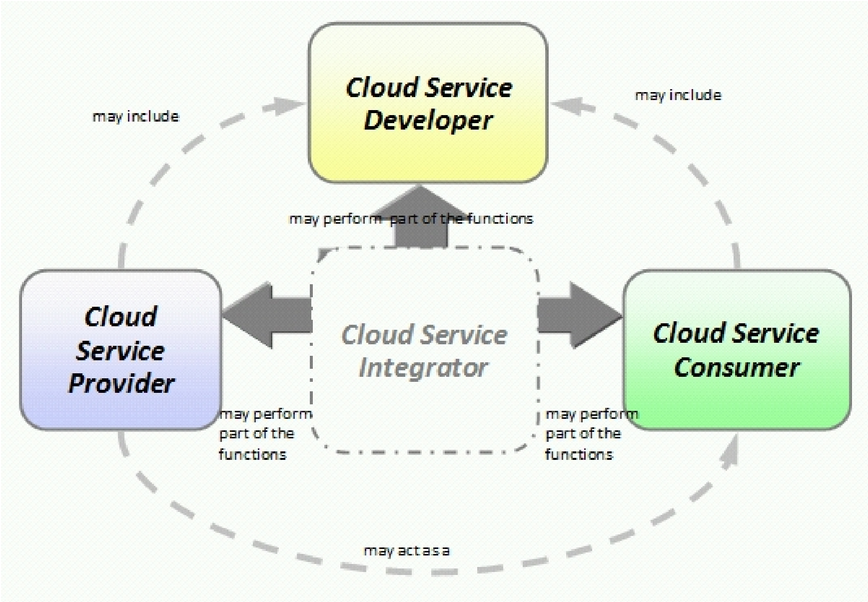


Figure 5: Cloud User Roles Relationship

8.1 Cloud Service Consumer Viewpoint

A cloud service consumer is a consumer of cloud computing, which includes the subscribers to cloud services and users of cloud services. In the private cloud domain, the users are within the same organization boundary as the provider. However, in the public cloud domain, the consumers can be either within the same organization as the provider (an extension of the private cloud community), or external to the provider. If the provider is external to the organization, then there should be an integrator role (see Section 8.4) within the organization for consuming cloud services.

In simple terms, consumer refers to an entity that utilizes any given service to perform an action and receive a result in return. For example, a customer opens a bank account. In this case, the entity is the "customer", the action is "open", and the result is the "bank account". From a cloud computing perspective, consumers won't notice when the infrastructure resources (e.g., CPU power, storage, etc.) are scaled up or down to satisfy their requests. In this particular example, the provider will manage the elasticity of those resources to deal with the demand of concurrent customers' requests to open bank accounts. This is particularly crucial and useful when a consumer looks for provisioning on-demand supported by a flexible business model on a pay-per-use basis.

What is a cloud service consumer looking for from an infrastructure cloud perspective?

- Enhanced user experience
- Flexible business models and pricing (e.g., pay-per-use or click rate)
- On-demand provisioning
- Elasticity – perceived unlimited resources (scale up or down)
- Software developer platform
- SLAs – agreements for performance
- Risk transference

8.2 Cloud Service Provider Viewpoint

A cloud computing provider or cloud computing service provider owns, manages, and operates live cloud computing systems to deliver services to consumers – including third parties and end users. Usually this requires significant resources and expertise in building and managing data centers that have advanced technology to effectively and efficiently provide cloud services. Some organizations are realizing some of the benefits of cloud computing by becoming "internal or private" cloud providers and servicing themselves. Where they already own significant resource capacity (can handle peak loads) and management capabilities or cannot afford the risk of using public clouds, significant operational efficiencies and agility advantages can be realized by these organizations.

Companies without significant IT investment will benefit from the economies of scale, not having to engineer for peak loads, and minimal barrier to entry with minimal capital expenditure required.

For instance, Amazon.com was the first cloud service provider by modernizing its data centers which, like most servers or networks, were using as little as 10% of their capacity at any one time just to leave room for occasional spikes. This allowed small, fast-moving groups to add new features faster and easier, and they went on to open it up to outsiders as Amazon Web Services in 2002 on a utility computing basis.

A cloud service provider owns the assets that are required to create, deliver, and manage cloud services for consumers. A provider may provide services to internal users as well as external users. It may also be an exclusive supplier to a single consuming organization, or support a hybrid model providing both public and private cloud services from different cloud service layers. A provider may be responsible for various environments (e.g., a legacy environment, private cloud services, public, or a hybrid model) and its asset management to support the service consumer.

In simple terms, provider refers to an entity that responds to an action by providing a service or set of services. For example, a customer pays for a product with his credit card. In this case, the entity is the "customer", the action is "pay", and the results are the "approval of the credit card transaction" and "the product itself". From a cloud perspective, a provider's work is transparent to consumers. In this particular example, the provider will scale up or down infrastructure resources to cope with the demand of credit card transactions coming from different merchants simultaneously. This is particularly important and critical during seasonal periods (such as Christmas, New Year, etc.) where peak loads usually stress credit cards systems to their limits. Having a cloud computing service provider is fundamentally beneficial, particularly when an organization is looking to pay per use and outsource the responsibility of managing infrastructure resources with the appropriate elasticity.

What is a cloud service provider looking to offer from an infrastructure cloud perspective?

- Advanced virtualization – efficient use of resources
- Flexible business models and competitive pricing (e.g., pay-per-use or click rate)
- Economies of scale
- Hosted – as-needed provisioning
- Parallel – on-demand processing
- Elasticity – efficient and flexible infrastructure (scale up or down)
- Automated and autonomic operations
- Management – e.g., operational, business type services for infrastructure

8.3 Cloud Service Developer Viewpoint

A cloud service developer develops the technical as well as the business aspects of a (simple or higher-level) cloud service offering, which may be part of the organization of the cloud service consumer or cloud service provider. A cloud service developer leverages the development tools to develop and compose a service or set of services.

8.4 Cloud Service Integrator Viewpoint

A cloud service integrator is a composite role that combines a number of services that can be cloud-based, in-house, and/or outsourced to provide integrated services from the service provider to the service consumers whether business or end users.

We can view the integrator role from two perspectives – technical and business. The Service Integrator is predominantly concerned with the technical aspects of system integration. The Service Integrator's skills and expertise come particularly into play in hybrid cloud environments. On the other hand, the business aspects of an integration project are handled by roles such as the cloud service provider's Transition Manager. The latter will be negotiating and regularly communicating with the cloud service consumer's Business Manager. The cloud service consumer's Service Integrator will most likely be in close contact with the provider's

Operations Manager for all technical matters. All of these roles – i.e., the provider's and consumer's business and technical lead roles – ideally report back to the provider's Service Manager who is responsible for ensuring the smoothest flow of services, business, and operational services.

The Service Integrator may also play a Service Composer role, which can be found within the Service Developer. Since this is a composite pattern, it may exhibit some of the sub-roles performed by the base roles (Consumer, Provider, and Developer). See Figure 6.

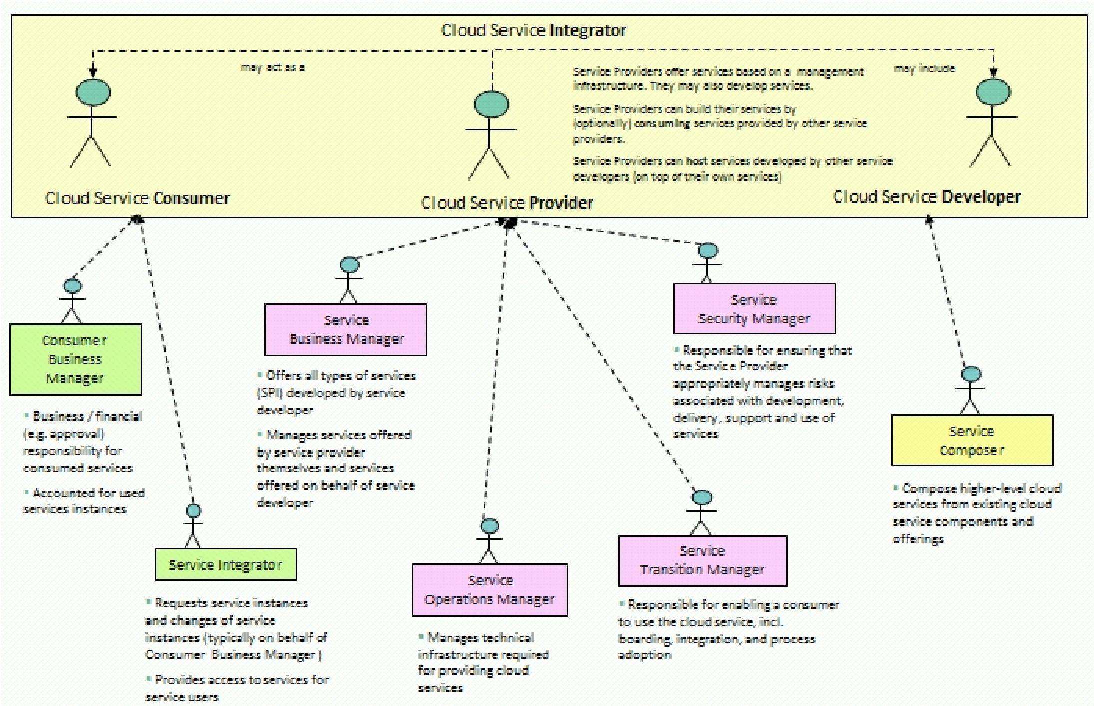


Figure 6: Cloud Service Integrator Use Role Explained

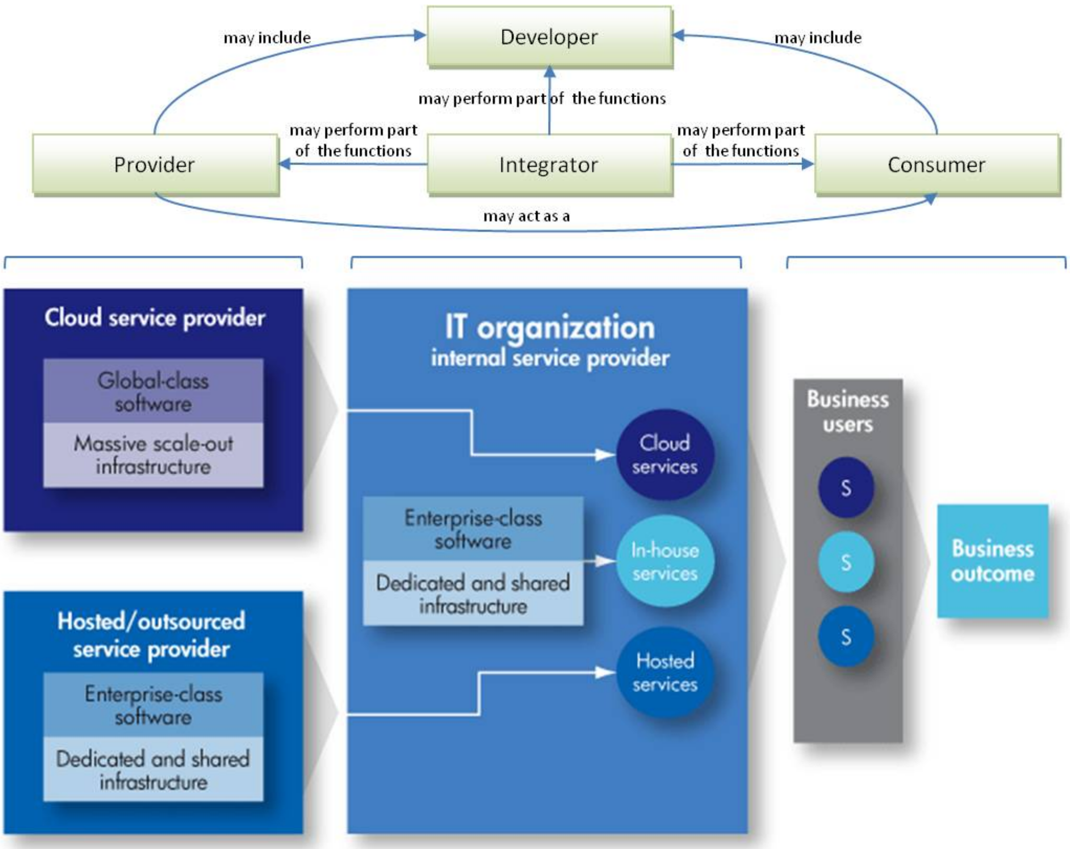
In simple terms, integrator refers to an entity that supplies business and IT services to another entity by integrating those services in such a way that they are transparent to the receiving entity, regardless of where those services are coming from. For example, a customer wants to view an auto racing video. The entity is the "Auto Racing Company", the action is "View", and the receiving entity of the service is the "Consumer". From a cloud computing perspective, consumers won't perceive what services and technologies have been employed and where they are coming from – being business process, software, platform, and infrastructure service provider or technical vendors. In this particular example, the integrator can use a combination of cloud services and technologies to address customers' requests to provide auto racing streaming video. Please refer to Chapter 10 for details on the role of the cloud service integrator.

What is a cloud service integrator looking to offer from an infrastructure perspective to its organization?

- Elasticity – efficient and flexible infrastructure (scale up or down)
- Advanced virtualization – efficient use of resources

- Reduced Total Cost of Ownership (TCO) – by loose-coupling of consumer and provider
- Communicate and govern the Service-Level Agreements (SLA)
- Agility and to respond to business requests

Figure 7 represents all the terms (Consumer, Provider, and Integrator) discussed previously in the context of the IT service supply chain.



[Source: White Paper: Avoiding Storm "Clouds" with ITIL [3].]

Figure 7: Perspective on the Whole IT Service Supply Chain

9 Governance Model

It is important to know that governance is a gating factor that needs to be in place before an organization can move to any type of IT deployment and this is no exception for a cloud computing implementation.

9.1 Cloud Governance Can Leverage SOA Governance

Like SOA, cloud governance includes stakeholder and decision rights enablement; for example, determining authority figures for procuring cloud solutions and the required level of stakeholder involvement. Governance is the key for establishing the cloud technology platforms and processes as well as overall strategic goals and policies. Setting up the charge-back strategy to govern the billing and metering of a cloud development is a good example of an important process to govern.

9.2 Specific Challenges for Governance in Cloud

One of the key value propositions of cloud computing is that the provider now assumes responsibility for the computing infrastructure. For the early adopters, not having to worry about infrastructure has been a boon. Nevertheless, it comes with certain risks:

- Exposing the organizational assets and information to the external world
- Unpredictable operational expenses incurred to meet SLAs
- Introduction of dynamic infrastructure and virtual resources management

Providers need to balance getting on with the job of keeping the infrastructure running, while at the same time giving organization IT enough visibility, choice, and delegated control over its own assets to fulfil its obligations to stakeholders. The buck still stops with organization IT, irrespective of whether the infrastructure is cloud-sourced or in-house.

9.3 Cloud Governance Considerations and Recommendation

There are some unique challenges businesses face within infrastructure cloud computing, to which we should apply the following, but not limited to, governance considerations:

- **Decouple technology with product mapping** – Technology implements process, enforces rules, and monitors cloud environments. Cloud supports dynamic infrastructure which allows a consumer to quickly develop and deploy its solution without having to worry about the physical underlying components including hardware, servers, and the facilities. Governing function usually maps to the virtual concepts that relate to the consumer in the cloud.

- **Robust track record and customizable, rigorous SLA** – SLAs need to reflect the needs of the organization, not just the raw “speeds and feeds” of traditional managed hosting environments. Just as important, though, is transparent measurement and reporting of the provider’s track record, which establishes a benchmark for reliability, consistency, and quality of performance.
- **Instrumentation, feeds, and real-time reporting** – A status dashboard is a bare minimum. Providers must have the instrumentation in place to report metrics and exceptions in real time to their customers, preferably as feeds that can plug into the customer’s existing management systems.

10 Case Study of Cloud-Based SOI – Motorcars in the Cloud

A Cloud Service Integrator (CSI) has built an online environment to stream live events, timings, and scorings as well as providing images for auto race videos from as far back as 1909 to Cloud Service Consumers (CSC) (Motor Car enthusiasts). The CSI is providing this environment by leveraging Cloud Service Providers (CSP) for infrastructure, platform, and software services along with a third-party Independent Software Vendor (ISV) who provides the streaming software.

The online environment is enabled by several services – some of which are provided by the CSPs while others are provided by the CSI themselves. A key service provided by the CSI is the streaming video of content from previous years that is housed in multiple storage clouds in different formats. The CSC gets billed based upon the payment options agreed by the CSI or by the CSP directly.

The CSI provides this service for multiple auto-race event organizers (e.g., North American Auto Racing Company, EMEA (Europe, the Middle East, and Africa) Auto Racing Company, etc.) while respecting the security and performance considerations specific to each tenant.

A CSC accesses a portal using a standard web browser to views races from yester years through streaming video which can be executed on a workstation, cell phone, or other appliance.

The CSI ensures that they meet contractual agreements with the CSC by providing an environment on which the CSP can provide their services.

The CSI leverages the appropriate CSP services as needed to deliver the content to the CSC.

The various building blocks outlined earlier manifest themselves in this scenario as described below. Some of these building blocks could be used globally by all service providers, whereas others are local to a given service provider as indicated within parentheses below.

Virtualization Manager (Global)

The IaaS provider needs this emulator to provision virtualized environments that emulate the right kind of infrastructure with the requisite resources in alignment with the SLAs established between cloud consumer and cloud provider or cloud integrator.

Location Manager (Global)

One of the tenants being hosted in this environment is paying additional charges for an organization license to have controlled private access to their videos. There is a sudden spike in demand for viewing auto-racing videos over the summer from EMEA users. Videos include those from both tenants. The Location Manager has to ensure that additional storage is provisioned to store these videos in compliance with EMEA federal regulations while respecting the contractual agreements with both tenants, including the privacy access policies for the premium tenant.

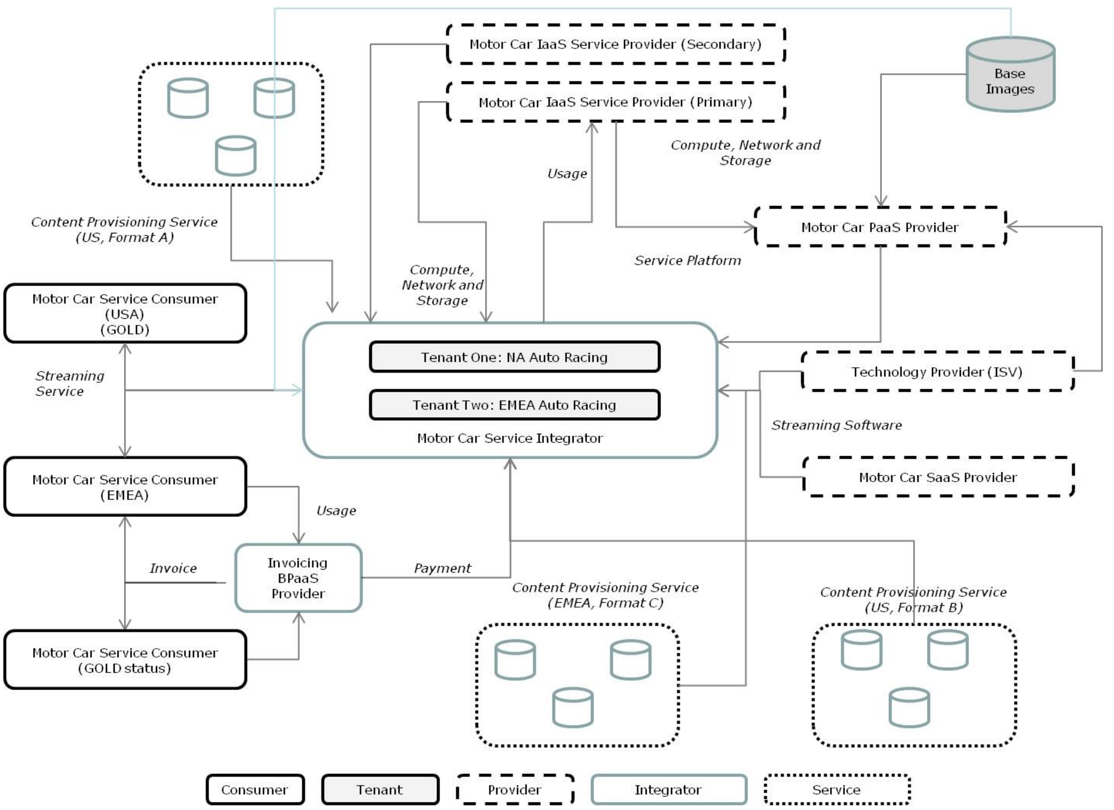


Figure 8: "Motorcars in the Cloud" Case Study

Figure 8 shows the business scenario representing the Motorcars in the Cloud Case Study. The following sections detail the manner in which the various building blocks outlined earlier in Figure 4 manifest themselves in this scenario.

Provisioning Manager (Local)

Sudden surge in demand results in the Event and Capacity Monitoring & Event Manager detecting the thresholds being reached triggering the Capacity & Performance Manager to evaluate the quantity of additional resources required. Based upon the Capacity & Performance Manager's evaluation, the Provisioning Manager allocates additional resources for server, network, and storage to the CSI and the PaaS provider. Network is reconfigured to accommodate quality of service levels, which requires reconfiguring the network to match the load and the streaming bandwidth. When the demand is restored to normal levels, the Provisioning Manager is triggered through a similar chain of events to de-allocate the resources allocated earlier.

Capacity & Performance Manager (Local)

Capacity & Performance Manager evaluates the quantity of additional resources required at any point in time. The resources thus identified are made available for allocation by the Provisioning Manager. Capacity & Performance Managers are typically local to a given service provider. Global Capacity & Performance Managers are not practical across the different service providers. Therefore, SLAs are essential between the various service providers that ensure that the right number of resources is allocated at the right time by each service provider. Charges for consumption of the resources could have a fixed and variable component. In order to address a sudden spike in demand that may not be met by the CSP, the CSI needs to ensure that there are fallback plans in place – perhaps with multiple IaaS providers – to ensure that the platform can still be continued to be provisioned to the CSC. Manages resources across multiple tenants (North American Auto Racing Company and EMEA Auto Racing Company) based upon the contractual agreements and policies in place.

Metering Manager (Global)

Metering is employed by various service providers to track and consolidate the usage of server, network, storage, and facilities resources by service consumers. In this scenario, for example, the IaaS service provider tracks the usage of such resources by the CSI. The CSI, in turn, would track the usage patterns for these resources by tenant.

Billing Manager (Local)

Billing Manager determines and charges the appropriate portion of the cost to specific consumers based upon their pattern of resource consumption as determined by the metering aggregator. This is applicable in the context of the multi-tenant environment that the CSI hosts in this scenario. The CSI determines the charges for each tenant based upon the cost provided by the Billing Manager.

Monitoring & Event Manager (Local)

Event and Capacity Monitoring & Event Manager continually tracks the infrastructure capacity consumed, triggering a notification to the Capacity & Performance Manager when predefined thresholds are reached. Like Capacity & Performance Managers, Event and Capacity Monitoring & Event Managers are also local to the service providers. Event and Capacity Monitoring & Event Manager could be global if there are well-established standards for exchange of information between multiple service consumers and service integrators.

Configuration Manager – (Global)

PaaS provider maintains multiple versions of the virtualization templates of the images to be provisioned. Configuration Manager enables the consistent maintenance and version control of these templates in alignment with the requirements specific to the tenant, geography, and service-level expectations.

Glossary

Application Service

An application service, along with Infrastructure Services and Human Services, are the components of Business Services. This service can be decomposed into a number of infrastructure services and application services.

BpaaS — Business Process as a Service.

BPM — Business Performance Management. BPM consists of a set of processes that are used by the Monitoring & Event Manager to monitor the performance of the business. Performance can be measured in various ways, including the company's financial performance as well as the performance of various business activities.

BSM — Business Service Management.

Business Process

Business processes realize courses of action. Courses of action are undertaken to ensure that the organization makes progress towards one or more of its goals. Business processes include processes that are specific to the data center, such as ITIL processes.

Business Service

A business service is consumed by a business process. This service can be decomposed into a number of infrastructure services, application services, and human services.

Cloud Computing

The definition of cloud computing provided by the National Institute of Standards and Technology (NIST) in the US has gained significant traction within the IT industry. This document leverages the definition of Cloud Computing from NIST as the formal definition. NIST defines cloud computing as follows:

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models."

Community Cloud

The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on-premise or off-premise.

Consumer
: A consumer is an entity (normally a person, organization, or piece of technology such as a software program) that uses a service. Every service has one or more consumers, and has effects that are of value to its consumers.

End User
: An end user is the ultimate consumer of the services provided in the context of a business process.

Human Services
: A business process consists of services provided by IT or human activity. A human service is a portion of the business process that is carried out by human activity.

Hybrid Cloud
: The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds).

IaaS
: Infrastructure as a Service. The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems; storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).

Infrastructure
: The IT infrastructure consists of foundational building blocks on which applications and business processes run. It provides generic infrastructure services that can be used by multiple applications.

Infrastructure Services
: The IT infrastructure consists of foundational building blocks on which applications and business processes run. Infrastructure services are components of business services, other infrastructure services, and application services. Moreover, they are composed of infrastructure services, virtualized services, and physical services.

ITSM
: IT Service Management.

PaaS
: Platform as a Service. The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.

Physical Services
: Physical services are consumed by virtualized services and infrastructure services. Physical services are provided by the hardware components (e.g., storage, network) of the infrastructure.

Private Cloud
: The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on-premise or off-premise.

Public Cloud
: The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

SaaS
: Software as a Service. The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

SLA
: Service Level Agreement.

SOCCI
: Service-Oriented Cloud Computing Infrastructure. SOCCI can be defined as a service-oriented, utility-based, manageable, scalable on-demand infrastructure that supports essential cloud characteristics, service, and deployment models. In other words, SOCCI describes the essentials for implementing and managing an IaaS environment.

SOI
: Service Oriented Infrastructure.

Virtualization Service
: A virtualization service is a service that appears to the software to be a physical service but is, in fact, a service provided by the virtualization software.

Index