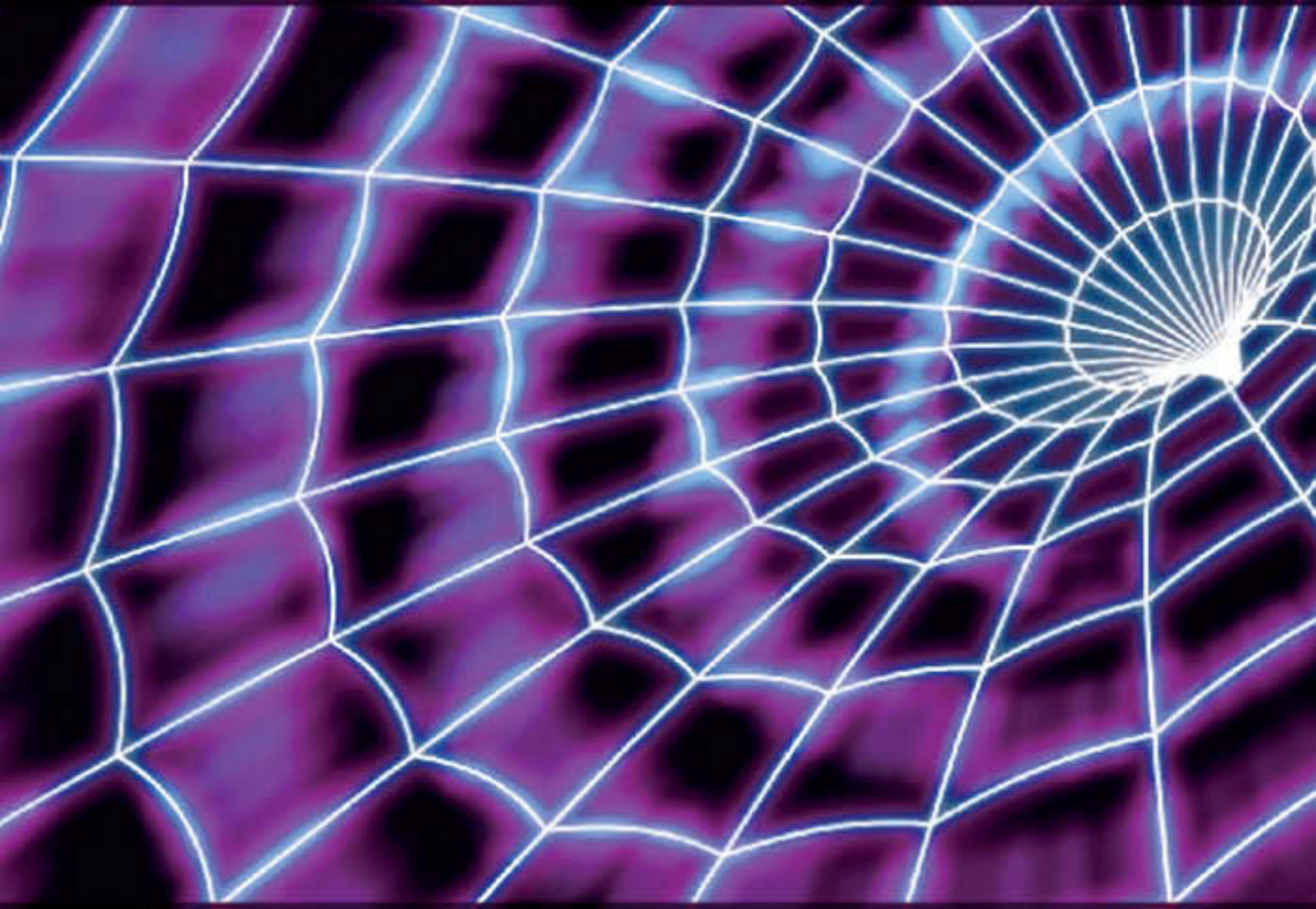
Cloud Computing Certification Kit

Specialist

Platform Management & Storage Management



The Art of Service

Foreword

As an education and training organization within the IT Service Management (ITSM) industry, we have watched with enthusiasm as cloud computing, Platform as a Service (PaaS) and Storage Management have evolved over the years. The opportunities provided through PaaS and Storage Management have allowed for significant growth within an industry that continues to mature and develop at a rapid pace.

Our primary goal is to provide the quality education and support materials needed to enable the understanding and application of PaaS and Storage Management in a wide range of contexts.

This comprehensive book is designed to complement the in-depth eLearn PaaS and Storage Management Specialist program provided by The Art of Service. The interactive eLearn course uses a combination of narrated PowerPoint presentations with flat text supplements and multiple choice assessments which will ultimately prepare you for the PaaS and Storage Management Specialist Level certification exam.

We hope you find this book to be a useful tool in your educational library and wish you well in your IT Service Management career!

The Art of Service

Write a review to receive any *free* eBook from our Catalogue - $99 Value!

If you recently bought this book we would love to hear from you! Benefit from receiving a free eBook from our catalogue at http://www.emereo.org/ if you write a review on Amazon (or the online store where you purchased this book) about your last purchase!

How does it work?

To post a review on Amazon, just log in to your account and click on the Create your own review button (under Customer Reviews) of the relevant product page. You can find examples of product reviews in Amazon. If you purchased from another online store, simply follow their procedures.

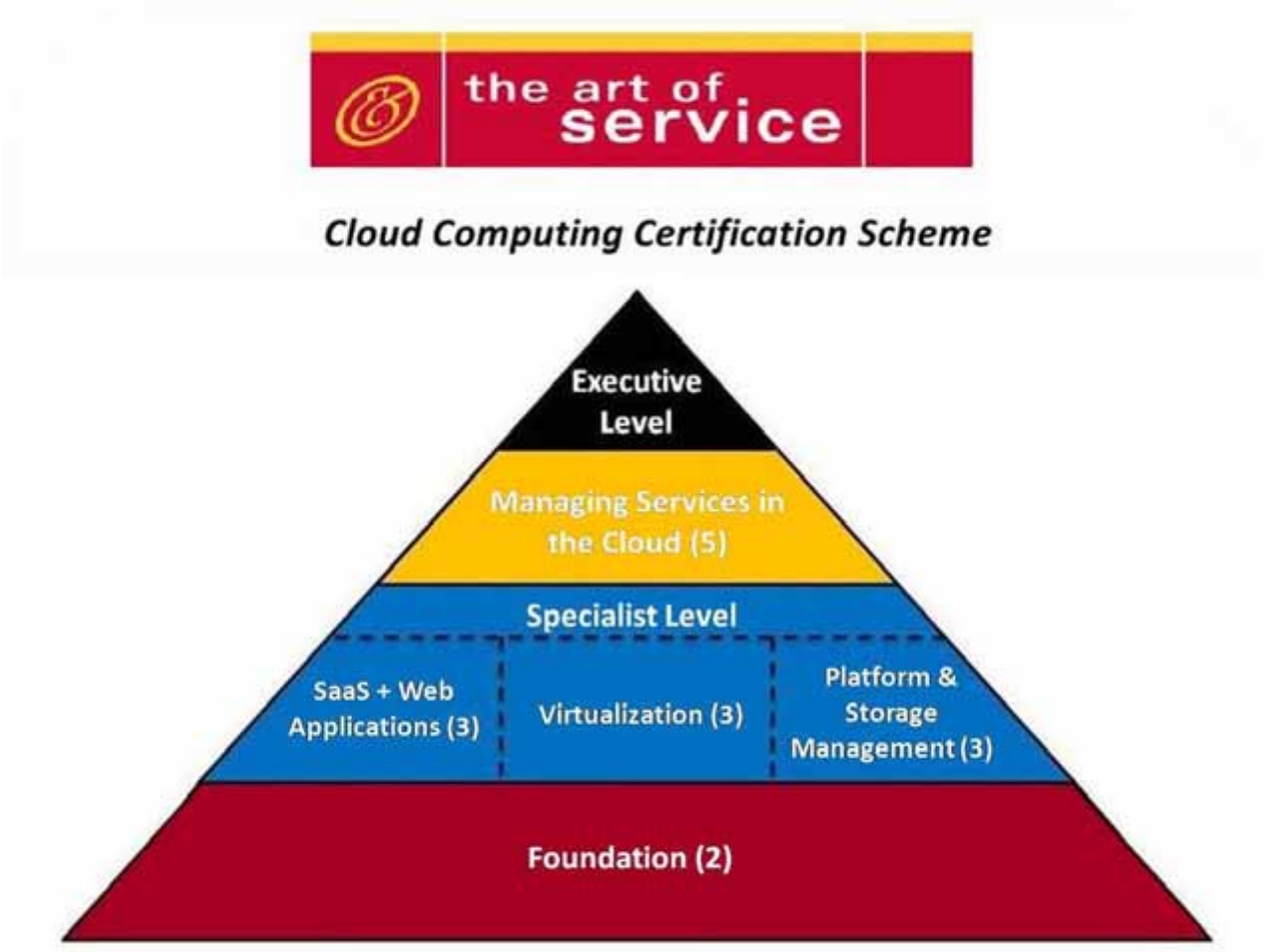
What happens when I submit my review?

Once you have submitted your review, send us an email at review@emereo.org with the link to your review, and the eBook you would like as our thank you from http://www.emereo.org/. Pick any book you like from the catalogue, up to $99 RRP. You will receive an email with your eBook as download link. It is that simple!

How does the Certification Kit work?

Welcome to the Platform as a Service and Storage Management Specialist Level Complete Certification Kit. This book is part of the series of books around Cloud Computing and managing the services that are involved in, or utilize, Cloud Computing.

The certification kits are in line with the Cloud Computing Certification Scheme.



After you've read this book, studied the eLearning materials and successfully passed your exam, you can continue with your qualifications through the specialist programs and work toward your Cloud Computing Executive Level Certification.

In addition to the certification kits, The Art of Service has also developed a series of books on the subject of Cloud Computing. All books are available as e-book, PDF download, audio book and paperback.

eLearn Component

This certification kit comes with FREE access to the eLearning program. The following page explains how to access the program materials online.

The Platform as a Service and Storage Management Specialist Level Exam

How to access the associated Platform as a Service and Storage Management Special Level eLearning Program:

1. Direct your browser to: www.theartofservice.org
2. Click ‘login’ (found at the top right of the page)
3. Click ‘Create New Account’. If you already have an existing account, please move on to step 5.
4. Follow the instructions to create a new account. You will need a valid email address to confirm your account creation. If you do not receive the confirmation email check that it has not been automatically moved to a Junk Mail or Spam folder.
5. Once your account has been confirmed, email your User-ID for your new account to paasstorage@theartofservice.com. We will add your account to the Platform as a Service and Storage Management eLearning Program and let you know how to access the program from then on.

Minimum system requirements for accessing the eLearning Program:

Processor	Pentium III (600 MHz) or higher
RAM	128MB (256MB recommended)
OS	Windows 98, NT, 2000, ME, XP, 2003, Mac OSX
Browser	Internet Explorer 5.x or higher (Cookies and JavaScript Enabled), Safari
Plug-Ins	Macromedia Flash Player 8
Other	16-bit sound card, mouse, speakers or headphones
Display Settings	1024x768 pixels
Internet Connection	Due to multimedia content of the site, a minimum connection speed of 256kbs is recommended. If you are behind a firewall and face problems accessing the course or the learning portal, please contact your network administrator for help.

If you are experiencing difficulties with the Flash Presentations within the eLearning Programs please make sure that:

1. You have the latest version of Flash Player installed, by visiting www.adobe.com and following the links.
2. You check that your security settings in your web browser don't prevent these flash modules playing. There is support for these issues on the Adobe webpage.

TABLE OF CONTENTS

1 Introduction to Platform as a Service

In 1961, a computing scientist in the field of artificial intelligence suggested that the technology of computers would one day see a future where computing power and applications could be sold as a utility business model. At the celebration of the Massachusett's Institute of Technology (MIT), John McCarthy was the first to publicly pronounce the concept that would later become known as cloud computing. Unfortunately, the technology to support this form of utility could not be found in the hardware, the software, or the telecommunications of the age.

It wasn't until the year 2002 that a major move was made to offer utility-based computing. That effort came from Amazon.com in their launch of their Web Services line of products. Several companies followed suit, with the leading service focused on software deployment, or software-as-a-service (SaaS) offerings. 2007 brought cloud computing to a whole new level, especially with a joint research effort between IBM and Google to provide computer science students at major universities an opportunity to obtain real experience with their studies. The framework for this effort was cloud computing.

However like any new technology concept, the definition of cloud computing is still largely undefined. What is in agreement are some of the characteristics of the cloud computing. The first is that it is utility-based, requiring a provider of the service and a customer of the service. Secondly, the infrastructure the provider uses to offer the service is transparent to the customer. The final characteristic is that anyone can access the cloud, though they may not have access to every part of that cloud. The cloud in this case is the Internet, or in some increasing cases, a customer's Intranet.

Several types of utility-based services are being provided over the Internet. The most popular is software-as-a-service, which allows customers a place to deploy their applications so that users of the Internet have access to them. Storage-as-a-service offers storage capacity for customers to house their data and raw data for any number of reasons. Some providers, like IBM, EMC, and Dell, will provide an entire infrastructure-as-a-service where they provide customers with the hardware, software, business processes, planning, and technical staff required for an IT support of the business.

One service which is gaining some momentum in the business market is platform-as-a-service.

1.1 *The Blind Men on the Cloud*

Defining cloud computing is difficult. Depending on the expert talked to, cloud computing can be defined from various perspectives. The following poem describes the problem. It was adapted by Sam Charrington, Vice President of the Product Management & Marketing for Appistry and read at the Executive Summit of the Next Generation Data Center (NGDC) conference in August 2008. It was adapted from John Godfrey Saxe's poem, “The Blind Men and the Elephant.”

It was six men of Info Tech
To learning much inclined,
Who went to see the Cloud
(Though all of them were blind)
That each by observation
Might satisfy his mind

The First approached the Cloud,
So sure that he was boasting
“I know exactly what this is...
This cloud is simply Hosting.”

The Second grasped within the Cloud,
Saying, “No it's obvious to me,
This Cloud is grid computing...
Servers working together in harmony!”

The Third, in need of an answer,
Cried, “Ho! I know its source of power
It's a utility computing solution
Which charges by the hour.”

The Fourth reached out to touch it,
It was there, but it was not
“Virtualization,” said he.
“That's precisely what we've got!”

The Fifth, so sure the rest were wrong
Declared “It's saas you fools,
Applications with no installation
It's breaking all the rules

The Sixth (whose name was Benioff),
Felt the future he did know,
He made haste in boldly stating,
*“This *IS* Web 3.0.”*

And so these men of Info Tech
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were partly wrong!

2 Defining Platforms on the Internet

Probably the hardest effort required to understand platform-as-a-service as an offering is to define what exactly a platform is. Most people and companies tend to confuse the terms application and platform. The software-as-a-service (SaaS) offering hasn't helped in this matter. Marc Andreessen, the co-author of Mosaic and founder of Netscape Communications Corporation, attempted a distinction between application and platform. He suggests “a platform is a system that can be reprogrammed and therefore customized by outside developers...and in that way, adapted to countless needs and niches that the platform's original developers could not have possibly contemplated.” He defines an application as “a system that cannot be reprogrammed by outside developers. It is a closed environment that does whatever its original developers intended it to do and nothing more.”[1]

The Andreessen definitions provide a step into understanding platform-as-a-service (PaaS). For the most part, software developers plan, design, create, test, and deploy their software to users. These activities are crucial to the success of the application, but even more important is the platform where these activities take place. In one sense, the Internet is one giant platform where millions of applications, called web pages, are found. No one can change the Internet, only those applications that can be found within the Internet. Thus, a platform can be loosely defined as a robust, fully featured space which is flexible, scalable, and highly adaptable for whatever purpose the customer desires.

For most providers, PaaS offerings focus on providing software developers the space to perform everything required for the entire software lifecycle. This document will cover platforms from every perspective with a focus on the Internet PaaS offering, what a platform looks like, how it can be used, and some examples of their uses.

In order to understand platform-as-a-service from a business standpoint, brief summaries of other cloud offerings will be covered. Most of the functions and benefits of these offerings will be found in PaaS offerings, but not vice versa. It is possible for cloud computing customers to utilize PaaS in conjunction with their own platform with benefits and disadvantages. Knowing the full range of offerings will provide a business with the best chance of determining the best solution.

---

1 Andreessen, Marc. Analyzing the Facebook Platform, three weeks in.. June 12, 2007.

3 Discovering Software-as-a-Service

Software as a Service (SaaS) is a software deployment opportunity, where an application is hosted into the Internet environment. Once there, the application is available to users throughout the Internet without any need for the user to install or run the application on their own computer. As a result, the user does not have to be concerned with maintaining the software, its operations, or support. SaaS is a pay as you use service, meaning that initial purchase of software and its license is not required; rather the charge is continuous, usually monthly, for the application without any contract.

For software vendors, SaaS is an attractive solution because of the promise of stronger protection of its online intellectual property and an ongoing revenue stream. They can host the application on its own web server or allow it to be handled by a third-party application service provider (ASP). As a result, the term “software-as-a-service” is sometimes diluted as it has two meanings. It can speak to the application itself and the hosting the environment. The second situation is also referred to as a platform and is sometimes interchanged with platform-as-a-service. The combat this problem, some have started to use the terms “SaaS” and “SaaS platform” to distinguish between the two situations.

Consumer-oriented web-native software is generally referred to as Web 2.0 application and is not SaaS which is typically attributed to business software. The distinction between SaaS and earlier applications delivered over the Internet is that SaaS solutions were designed to leverage the technologies found on the web, with a multi-tenant capability to allow multiple users to access shared resources.

Four maturity levels describing SaaS architectures have been generally accepted by providers and software vendors alike.

- Level one refers to hosted applications which use a single instance for each user on the hosting server. The least amount of effort to deploy and operating costs are found with this level of application.
- Level two applications provide the user with separate instances of the application with the addition of configurable metadata. Hence the vendor can cater to the different needs of the user without having to change the common application code.
- Level three adds multi-tenancy so that multiple users can use a single instance of the application. A better use of server resources is gained but scalability is sacrificed.
- Level four applications created in a multitier architecture supporting a number of identical application instances. As demand on the application increases or

decreases, the number of application instances changes accordingly, usually by adding or removing servers. This adds scalability to the solution without having to alter the software architecture.

4 Discovering Storage-as-a-Service

Storage-as-a-service solutions provide capacity to customers for a variety of uses. The most prominent drivers for online storage are individual backups, backups for potential disaster recovery, archiving, raw data for online databases of analytics, and online collaboration efforts. The focus for storage-as-a-service is the physical storage infrastructure with numerous layers of virtualization.

From a provider perspective, the provision of storage means supplying the necessary capacity required by the customer, ensuring that the data is always available to the customer, and that transfer rates are outstanding for the customer. In most cases, data from multiple customers will be found on a single storage device with the appropriate volume separation to ensure that the data is kept secure and uncorrupted.

From a customer perspective, they have a place to store their data without having to heavily invest in the necessary hardware and software to save this data. They have continuous access to this data at any time from any location. The solution is scalable to handle any growth to the business.

5 The Platform Layers

Many choices for cloud computing and application hosting are emerging on the Internet. The pace is so fast; difficulty in understanding the details of the individual offerings can be fuzzy and misunderstood by Independent Software Vendors (ISV), enterprise developers, or business customers.

Without a clear definition of platform by the industry, the offerings can vary greatly in what they provide. The enterprise looking to create a platform for their software development needs to choose the right direction. Below are options as suggested by Phil Wainewright, an analyst on software industry trends.

Do It Yourself

This has always been an option for most enterprises. They own the servers, the network, and the software. The enterprise takes on the responsibility for designing, developing, deploying, and managing the application as well as the infrastructure the application sits on.

Managed Hosting

Similar to the Do-It-Yourself option but the responsibility of the infrastructure is shared with another party. The extent of that sharing varies depending on the relationship and the contract. Typically, the enterprise can choose the components of the infrastructure and create some rules, and of course the cost of operating the infrastructure. The second party is responsible for designing, implementing, managing and improving the infrastructure, usually based on a set of Service Level Agreements (SLAs) and policies provided by the customer on how those SLAs should be met.

Cloud Computing

This is the first layer of platform which is utility based. The enterprise pays for the resources that their enterprise uses. Service Level Agreements may be in place, but the customer has little to no say in how those SLAs are met. They also relinquish choices in infrastructure and infrastructure design to the provider of the platform. The purpose of this service is to provide a platform for installing and running the enterprise's application. While the provider ensures the infrastructure is working properly, it is the responsibility of the enterprise to ensure the application is working properly.

Cloud IDEs

To provide a greater offering than a simple hosting platform for applications, cloud-based Integrated Development Environments (IDE) allow the enterprise customer a platform for developing and deploying their applications. The control of the infrastructure is, of course, the responsibility of the provider. The provider also provides the tools used in the development and collaboration effort. All the enterprise has to provide is the manpower. The downside is that development is tied to the infrastructure that the provider offers. It development doesn't work out; moving the application to a different platform is difficult.

Cloud Application Builders

The final layer of platforms is geared towards the power users and designers. The application infrastructure is in place, but the customer brings their own tools and developers to the platform. The only real constraint is that the type of application that can be developed depends on the infrastructure that is provided.

Each of these layers has their benefits and failings. As one moves through the layers, they lose more control over the infrastructure. As they lose control the customer also loses the financial burden for implementing and maintaining that infrastructure. Additionally, the customer gains the ability to rapidly develop and deploy their application.

5.1 *Do it Yourself*

The highly competitive global business environment should be a surprise to anyone. The Internet has opened new markets around the world and has increased the speed, amount, and availability of information exchanges. Technology has gotten faster, easier to use, and provide greater standard of living with ever before. The consumer is demanding the next best thing now. To meet this demand, companies have had to become more innovative, more accessible, and more conscience of what they are delivering. As they focus on their core business, one area of attention is their network infrastructure and how it is supporting the overall business offering.

Web technology and e-commerce has proven that integrating the Internet into business operations must become a key objective for companies today and into the future. A new business platform must be adopted for success; one where networking capabilities will be essential for business operations. One of greatest concerns for a company in today's business environment is the need to improve the scale and speed of its core business activities into the global market. In order to handle this demand, adoption of new

technologies and business practices are paramount. Integrating information technologies with the Internet is simply starts the move to better communication, collaboration, and information sharing. Networking technologies are become more intelligent and mobile and should be embraced. Network skill in today's emerging technologies is a requirement of the company.

The role of the network infrastructure is changing because of the Internet. The use of Internet-based application programming interfaces (API) and Web protocols like HTTP are becoming unified and accelerating networking functions. Virtualization is become commonplace. New focus on quality of service (QoS) controls, traffic engineering, and mobile support are needed for continued Internet protocol (IP) networks. A company interested in creating their own networking solution that combines the internal infrastructure and still utilize the functionality provided by the Internet should focus on the following areas:

- A controlled environment for handling access into the network from several types of access media, including asymmetric digital subscriber line (ADSL) and mobile access.
- Properly configured routing of the IP network to handle issues of capacity, scale, and performance. Virtualization techniques will provide the greatest capabilities to the company in this area requiring sufficient knowledge in this area.
- The creation of a gateway between access and core network capabilities. The concept separates the core network components for internal and external access. This allows an Internet/Intranet capability for controlling the access to information. Some companies apply a third layer between the Intranet and the Internet called an Extranet for the purposes of communicating and collaborating with partnering companies and services without exposing the information largely to the internal population or the global community.
- A network that supports high bandwidth concerns and information transport across long distances. As business grows, additional demand on communication methods such as email, blogs, instant messaging, and file sharing will increase. The network needs to handle this demand without sacrificing performance or availability.
- The administration and maintenance of the network to ensure continued access and performance of the network to users and employees alike.

As the Internet is utilized by more businesses, the functionality will expand, and companies will be expected to keep up.

5.2 *Managed Operations*

Many corporations are placing the responsibility of managing parts or the whole of their IT infrastructure to another party in the form of managed operations. The company still retains the financial responsibility and retains ownership of all hardware and software assets used within the infrastructure. The provider of managed operations provides the staff, the skill, and the business processes required to ensure that the IT solution properly supports the business operations of the company. This type of solution allows the company to focus on their business without being overwhelmed by the technical maintenance required to support the business within a technological age.

The success of a managed operation solution is to create a powerful partnership between the provider and the company. This begins with clear and complete communication on the current state, issues, risks, and recommendations from the provider and business goals, objectives, and future initiatives by the company. The company should rely on the provider to present new solutions available to them as well as questions of the provider about new technologies. In order to have this happen, the company will need to be upfront with the direction of the business so that the provider has an opportunity to isolate the best solutions. Managed operation providers offer their services to multiple companies which provide the opportunity to leverage solutions across their customer base. If a company wishes to move into implementing an Internet platform similar to the PaaS solution, discover if the provider is knowledgeable of such solutions or has already implemented such a solution for another customer.

For the provider, the future of their business is dependent on the way they approach the topic of platform-as-a-service. Essentially, they have three options. The first is to create the ability to develop private clouds for their customers. In this manner, the customer retains the cost and functionality that cloud platforms have. The second option is to create the cloud within the provider’s infrastructure and provide access to the customer as a service as part of the managed operations package. The last option is to use existing PaaS solutions from another party and manage the relationship of that service within the managed operations segment of the contract. In this sense, the PaaS provider becomes a vendor of the managed operations provider to ensure that the contract is fulfilled. The controlling elements of these options will be costed and the Service Level Agreements between the provider and the company.

Service Level Agreements (SLAs) are the minimum performance requirements for the IT infrastructure to ensure that it can support the company's business operations. From a PaaS perspective, the SLAs will focus primarily on availability, capacity, performance, bandwidths, and data transfer thresholds.

5.3 ***Cloud Computing***

Most “as-a-service” solutions found on the Internet will fall into this category. They utilize the technology and concepts of cloud computing, but they are not all exclusive. For companies looking for a complete platform-as-a-service solution, they will need to piecemeal the solution together from several sources and tie all of them together using a central “processing” point. Most SaaS solutions provide an opportunity to use software or a platform from which to deploy and maintain an application. However that software-as-a-service solution will not provide the environment to design, develop, and test the application. Most storage-as-a-service solutions are focused on a component of a company's storage needs whether that focus is on backups, archiving, or collaboration.

These solutions may be sufficient to meet a company's needs at the present time, though they may need to understand the impact of committing themselves to a single partial solution. They should ask the provider about their intended growth in cloud computing offerings. Even using multiple partial solutions will provide companies with the cost benefits and capabilities that they are looking for from a cloud solution. They will not be required to have an extensive hardware infrastructure, nor a large staff to maintain the capabilities provided by the solution. What is necessary is the hardware and staff required to tie any cloud computing solutions to the IT infrastructure of the business.

Assuming that multiple cloud computing solutions are in use, connecting means mapping the solutions appropriately while using the necessary protocols across the board, in order to facilitate communication. Since all of the solutions are based in the Internet, most of the protocols should be the same though there may be some configuration changes to maintain that this is not a problem. Understanding how the solutions work together is a strategic effort for the company. The strategy may require a solution for software deployment and another for database data storage. A cloud solution for collaboration purposes may be involved. Wherever a cloud solution is not available for a component, the company will have to make up the difference. In other words, the company creates the entire platform and uses commercially attainable cloud offerings to handle parts of the platform.

5.4 *Cloud IDE*

Independent development environments (IDEs) found on the Internet as a pay-as-you-use service are becoming more available. The offerings provide a platform from which to design, develop, or debug an application. The protocols and technology used to build the application are web-based. In addition to the platform, a software application is provided: the IDE. IDEs provide the facilities necessary for computer programmers to develop working software. These facilities typically include a source code editor, compiler or interpreter, build automation tools, and debugger.

The general purpose of an IDE is to maximize productivity from programmers by reducing or eliminating the need to perform mode switching. Most IDEs are built to handle a single programming language, though a few exist that support multiple languages. Within an IDE, a programmer can take advantages of features that match the programming paradigms of the language to perform authoring, modifying, compiling, deploying, and debugging. The theory behind IDE is to provide a programming interface that visualizes the necessary configuration for command line utilities as units, increasing productivity in development by reducing the time required to learn the programming language.

Though these cloud solutions provide a platform for developing the software, that platform may not be scalable or available to continue use in production. Though many of the processing demands that an application will find in production may have been simulated through the testing stage of development, production provides real case demands on the application. These simulations are based on projected use of the application and can be used to identify the processing thresholds for the application. However, a production environment may present situations and demands not predicted during development. In this respect, both the application and the platform where it resides must adjust in real time to compensate for any unexpected disruptions.

5.5 *Cloud Application Builders*

Cloud Application Builders provide a comprehensive fully functioning platform to perform all the activities required during every stage of the application design, development, testing, deployment, and maintenance modes. The service may provide an IDE or it may be required of the programmer. The customer to this service has all of their infrastructure concerns taken care of by the provider, including the server to house the application, the storage for the data the application uses and generates, the bandwidth and data transfer rates required to perform transactions over the Internet. The solution is scalable to handle the demands found in development and in production, allowing for growth as the use of the application increases. The capacity can handle the processing concerns and storage of standard use as well as spikes in usage. The availability of the infrastructure is near perfect guaranteed.

Additional tools may also be in place to monitor the application throughout its life cycle. These monitors can identify the functions of the application that are being used and how they are being used. They can identify trends that allow a customer to predict future demands on the application. The platform should allow the customer to develop and test improvements to the application without disrupting the application in production. When updates or new versions are available, they can be applied immediately to the production state without further disruption.

6 Understanding Multi-Tenant Architectures

Utility computing relies on the ability to share resources across multiple customers. In terms of servers, networks, and platforms, that requires that more than one “tenant” can be found using the system. To meet the needs of all tenants within the solution, architecture has to be created to support the concerns, requirements, and goals of the tenants. This type of system is called a multi-tenant architecture. Multi-tenancy can be found on every level of the system: the network, the server, the storage, and the applications. Many software-as-a-service solutions are multi-tenant applications. For platform-as-a-service providers, they not only have to provide a multi-tenant platform, but support applications on the platforms that are multi-tenant: giving another layer of complexity to the system.

Because many facets of computing require architectures that allow multiple tenants to simultaneously use the resources, there are several ideas on what is required in these environments. Microsoft has identified three attributes for a multi-tenant architecture: scalable, efficient, and configurable.[2]

In scaling resources the provider does not want to sacrifice their ability to maximize the number of simultaneous occurrences of the resources. They want to also utilize the resources more efficiently. A variety of techniques are available to provide scalability to resources. If the resource is a server, partitioning the drive can essentially allow multiple users to use the same drive. The more partitions, the more tenants are available. These partitions can be used to separate the physical drive by using virtualization software. Virtualization is the best method as it doesn't make any physical changes to the environment and can be adjusted rather easily. The same technique can be used to partition storage devices across the environment. Synchronizing data across multiple servers allows the application to be available at all times from all locations, as well as providing the opportunity to balance workloads when demand on the system is high. Pooling resources is also another technique to provide further scaling capabilities. It consolidates the resources together and treats the consolidation as a single resource which is configured to support multiple tenants.

The efficiency of the resource in a multi-tenant architecture is heavily reliant on the

---

2 Carraro, Gianpaolo and Chong, Frederick. *Architecture Strategies for Catching the Long Tail*. April 2006.

resources ability to differentiate and handle requests from multiple tenants. The reality of a multi-tenant architecture is that for every occurrence of the resource, several to hundreds of tenants may be using the resource. This usage of the resource must remain transparent to each tenant. In order for this to happen, the solution must be able to handle each tenant request independently from each other, translating exactly what each request entails, and quickly resolving the request without colliding with other requests. Any inability to handle this will result in degradation in availability and performance resulting in a dissatisfied tenant.

Creating a solution where a single customer is using a single instance is much easier than creating a solution where multiple customers are using the same instance. Most customers want the solution to fit their needs exactly without compromise. This requires some ability to customize the solution to meet those needs. Unfortunately, customizing a multi-tenant solution for one customer means changing the solution for all customers. Though there are instances when this may be acceptable, such as new technology support, most times these changes are customer specific. Most applications allow users to customize their view of the application using metadata that has the application look and behave the way the customer would like. The metadata concept is useful for utility-based platforms where tenants can utilize the platform the way they need to. This is done on a larger scale and in many ways automatically based on usage in the environment. Interfaces and resources found within the platform are scalable and customizable for each tenant.

6.1 *Finding the Right Platform Model on the Internet*

As more and more providers are offering “platforms” as a utility, more concerns have arisen as to the type of platforms are available. The infrastructure has a definitive impact on the type of applications that can be developed and deployed on a platform. For this reason, providers have to predetermine the software development market they are looking to support. As a result, the “concept” of platform is growing. Several models are emerging as providers tighten their reigns. Marc Andreessen suggests there are three hierarchical platform models[3].

Access API

The first level of platform is the most popular. The platform is delivered to the customer through a web service application programming interface (API). Access to this API occurs

3 Andressen, Marc. The three kinds of platforms you meet on the Internet. September 16, 2007.

through an access protocol like REST or SOAP. The key to this model is that the application code executes outside of the platform where the data and services are located. The application works independent of the platform, but provides the result of the processing of the application to the platform itself.

Because the application and platform remain separate, the responsibility of the application and the infrastructure to support the application falls completely on the developer.

Plug-In API

End-user applications have historically been able to add additional functionality that was not present at the original install through the use of plug-ins. This level of platform still utilizes some access APIs to perform properly, but its focus is not on providing a desired final result, as much as it is to provide functionality to the platform that was not there before. The code for the application is still located outside of the platform. When a plug-in is desired, a request is made to a separate location to find the code, run the installation process, and ensure that everything is okay. The platform is updated with the new functionality and the user proceeds. Because the platform and application are separate, the responsibility still falls entirely on the developer. Level 2 Plug-in applications are harder to develop than level 1 applications.

Runtime Environments

Level 3 platforms allow the application to run inside the platform. The code is uploaded to the platform and executed inside the platform. A level 3 platform may utilize some access APIs and plug-in APIs, but the core system of the application resides on the platform. As a result, the developer has no responsibility for the infrastructure required to run the application.
In another version of the Level 3 platform, the responsibility of the infrastructure still resides with the provider of the platform. However, the platform allows the developer to design and develop the application as well as run it from the platform. Access APIs and plug-in APIs may still be used. The core system still resides on the platform itself, but it was developed there as well. In this model, the developer does not require any sort of infrastructure outside of the provided service.

6.2 *Delving into Access APIs*

The Level 1 Access API platform is the easiest platform to create for an application. The approach has been used by eBay, Paypal, and Google Maps, among others. The defining characteristic of an Access API platform is the application code lives outside of the platform. Mashups are a common occurrence on this platform. Mashups are web

applications that combine functionality and data from multiple sources. Having the application code separate from the platform allows the initial site to be relieved of any additional programming that may impact the overall performance. It also allows the application code to be shared.
When people think about an Internet platform, this is the platform that comes quickly to mind, although the approach is greatly limited. In order to support the application code, the developer is responsible for the technical expertise, financial backing, and technical infrastructure to develop and run the code. Because of this fact, a number of applications using these APIs have not reached their potential even though they are the easiest to incorporate.

The primary driver of these applications is the access protocols used to manage the access to the separate application code. The two most common protocols used in Internet applications are REST and SOAP. REST, or representational state transfer, is a collection of architecture principles for the network. It defines how resources, or sources of specific information, are defined and addressed. Designed by one of the authors of Hypertext Transfer Protocol (HTTP), Roy Fielding, REST is used for distributed hypermedia systems like the World Wide Web and is often used in conjunction with HTTP. However, an application can be created with the REST style and not using HTTP. In addition, HTTP programming can be created without using the principles of REST.

The use of the REST is predominant in the World Wide Web. At the heart of REST is that existence and referencing of resources with a global identifier, such as a uniform resources identifier (URL). The use of these resources is managed by a standardized interface like HTTP which exchanges these resources, or their representatives, between components on the network. In the physical environment of the network, there may be a number of additional network components between the client and the server, such as gateways, routers, firewalls, tunnels, proxies, and the like. REST ignores these components when looking out onto the network for a resource. Therefore, an application using the REST protocol only requires the identifier for the resources and the action required by the resource.

Simple Object Access Protocol (SOAP) is used to exchange structured information through the implementation of Web services. The format it uses is Extensible Markup Language (XML). By itself, SOAP is ineffective, so it must be run on other application layer protocols like Remote Procedure Call (RPC) and HTTP to handle transmission and negotiation of the message. SOAP is used to provide a messaging framework by creating the web services protocol stack layer. It is used to search out a web service enabled site with specific parameters for a result that is directly integrated into the initiating application.

6.3 *Delving into Plus-in APIs*

Plug-ins are computer programs that complete a specific function upon demand. On their own, they have no value, but as an extension to a host application they provide additional capabilities to the application that were not present before. When an application is developed, it consists of core functions available to the user. Plug-ins are supported to enable features that were not conceived during development. Sometimes, plug-ins are deliberately used to reduce the size of the core application, giving the user the opportunity to pick and choose additional features as they need them. Another deliberate reason is to separate the source code from the core application because of incompatibility issues with software licenses. Plug-in support also allows third-party software developers to offer additional capabilities to the core application.

How plug-ins work is dependent on the user of the core application requesting the use of the plug-in. At that point, the plug-in is installed and registers with the core application and the appropriate protocols to handle any exchange of data required by the plug-in. The core application provides services used by the plug-in. Photoshop has used plug-in support to extend the capabilities of its core system for many years. The web browser Firefox uses plug-ins to allow blogging, downloading, searching and other capabilities.

The Facebook platform is the first Internet-based Level 2 platform. However, the entire burden of the building and running the application is still outside of the platform and the responsibility of the developer. The technical expertise and financial resources required are fairly high if a meaningful application is desired. Though Level 2 platforms can be very powerful tools for users, they can be equally constraining to developers. Fortunately, these platforms to provide an additional benefit. Most developers have the burden for distributing their application to users. With a Level 2 platform like Facebook, the distribution center is already in place, allowing the user base easier access to new applications.

6.4 *Delving into Runtime Environments*

The difference between runtime environments and the Level 1 Access API and Level 2 Plug-in API platforms is that the runtime environment allows the application code to run inside the platform. The code is uploaded onto the platform and runs online. A Level 3 application may use access protocols and plug-ins. Unlike the previous two levels, a developer does not require their own services, storage, database, bandwidth, or myriad of technical components necessary to run the application in production. The platform handles everything.

For the provider, a Level 3 platform is more difficult to build then a Level 2 platform. There are issues that a Level 2 platform can ignore. The most obvious is that with a Level 2

platform, the code is running elsewhere, so any problems with the application are someone else's responsibility. On a Level 3 platform, problems with the application may be a problem for the provider.

The problem starts with creating an environment that can execute application code that has not been developed by the provider and as such can be arbitrary. The platform must be able to accept and manage that code. As a result, most platforms restrict what programming languages and protocols are used. In order for multiple languages to be supported, further complexity has to be built into the platform infrastructure.

To support the developer from a business perspective, the provider will need to provide development tools that are integrated into the platform. This will allow development of applications within the environment. Without such integrated tools, the developer has no real incentive of using the platform. The development tools need to be able to assist the developer in the programming language of their choice. To support any arbitrary language requires additional components for the system. Fortunately, by matching the development tools to only allow languages supported by the infrastructure, the provider can ensure that issues of compatibility are diminished.

Simply providing a platform for developing and running an application is not enough. Many applications require an ability to obtain, store, and process data. In order to handle this appropriately, the application needs a suitable database environment. This is more than simply providing a database; it needs to provide storage capacity, database languages, and the ability to allow querying by the application and secondary sources.

Security is major concern for providers. Of course, the most obvious security need is protecting the customer's assets from corruption or attack. This requires putting the proper networking techniques in place to protect data on the server, on storage devices, and in transit. Access and authentication controls must be in place to ensure proper use of the system and the applications found on the system. Every bit of data should be considered mission critical for the provider. However, there is another level of security required in a Level 3 platform; a method to prevent applications, specifically from multiple customers, from conflicting with each other. One methodology utilizes physical parameters and virtualization to "sandbox" applications from each other.

A level 3 platform must be prepared to scale automatically, especially in situations where the success of the application exceeds predictions. Scaling must happen at every level: capacity, bandwidth, performance, and security. Automation must be built to ensure that all the components of the network have the capabilities to run every third-party application found on the platform. Any customization is a one-off solution that requires even further separation from the general platform offered.

For the customer, Level 3 platforms reduce the technical expertise required by developers to 90% and the amount of money required to develop to near $0. Because of this, development within a Level 3 environment is far less expensive than the Level 1 and 2 environments. Additionally, the platform can become a system of open source applications where customers and users share code with each other. This creates an opportunity for customers to utilize, clone, and modify each other's applications with greater ease. Utilizing open source code speeds up the development processes and ultimately allows a greater supply of applications to the user. If the code is not shared freely, it is still potentially available for profit. Second Life is becoming such a platform.

7 Exploring the Key Elements to a PaaS Solution

As developers and businesses look into cloud solutions with more vigor and greater attention, more gaps are identified. Pay-as-you-use services for software, storage, computing, and even CRM (Customer Relations Management) are having greater success as time goes on. However, these services still only represent pieces of a larger solution. Platform-as-a-service has the ability to encompass all of these services plus more. A cloud platform includes all the systems required to manage the entire life cycle of a web application. Bungee Labs have identified six key elements such a platform.[4]

Integrated environment

The first element of a fully functional platform is the ability to develop, test, deploy, host, and maintain the application in a single environment. The life cycle management of most applications has different systems involved in each of these steps. This places a considerable amount of burden on the developer in terms of hardware, maintenance, and configuring to ensure that an application moves through its life cycle.

User experience

The experience the developer has using the platform is critical to the success of the PaaS service. Most developers move from one project to the next, so repeat business is a considerable factor. But from the developing perspective, having the tools and capabilities available to move the application through its life cycle is extremely important. It's like using

4 Bungee Labs. Defining Platform-As-A-Service, or PaaS. February 18, 2008.

DOS and Windows. Though DOS allowed a computer user to perform most functions required at the time, the introduction of the Windows operating system made those functions easier to perform and was a more pleasant experience.

Built In Management Features

Developers and businesses have to deal with a number of different variables throughout the life cycle of the application which require considerable attention that can adversely impact the proper functioning of the application. Of these scalability is a major factor, as well as reliability and security. These factors should be built into the platform without any need to develop, configure, or in any way spend any time, cost, or effort away from the application itself.

Built In Integration

Very few applications are developed in complete isolation. The need to connect to external sources for dynamic data, updates, even third party web services is very much a required consideration. Specifically for applications depending on real time data, the platform needs to allow direct and continuous connection to external sources.

Support Collaboration

In the same way applications are not developed in isolation, developers do not work in isolation. Most software development projects are a collaboration of several individuals throughout the entire life cycle of the application. This collaboration is a mixture of formal and on-demand occurrences that require the ability to communicate effectively while maintaining the security and performance of the application code.

Deep Application Instrumentation

Software development no longer ends at the deployment of the application. As much as a developer or business would like to think that they have introduced a perfect application into the marketplace, this is rarely true. How the application is used, the performance, and reliability issues are all concerns that software manufacture would like to monitor. A platform-as-a-service would have a full set of instrumentation in place to handle this monitoring to effect improvements.

7.1 *What is so important about Integrated Environment?*

The software development cycle is one of the most important concepts for companies

seeking solutions on the Internet and a perfect way of understanding the best solutions available. Though there are a number of models for the development cycle, they all are variations of the same theme: design, implement, test, and deliver. The development and deployment processes should ideally be completed within a contained environment; and the smaller the environment, the better. Unfortunately, the typical implementation of the process has each step of the development process completed on a separate system. As each phase of the process is completed, the application code and the data it uses are migrated to a different server to perform the next phase. In typical development labs, this often requires additional effort to configure the server and appropriate devices to support the application, as well as extra cost to have the infrastructure in place to perform the work.

Not every development lab is the same, so it's not a surprise that technologies like virtualization are being used to reduce the hardware and software needs in the environment. Virtualization also contributes to integrating the environment. The concept of integration within software development is to provide systems that allow the developers to focus on their core objective: development. When they have to turn their focus to the hardware, software, and configuration needs to run the application, they are not working on the development. The more they are distracted by these needs, the longer the development cycle becomes, resulting in potential missed deadlines, increased costs, and greater risks. Integrated environments don't just eliminate these concerns, they turn the tide for software developments. By having one environment supporting the entire software development life-cycle, applications beat deadlines, been cheaper to build, and allowed more attention on vulnerabilities.

True platform-as-a-service offerings provide a completely integrated environment for development and hosting of web-based applications. Within these environments, the movement of the application through each of the development stages can be completed through simple point and click operations. As demand on the application increases, specifically during testing and deployment, the environment automatically adjusts the capacity, bandwidth, and workload balancing to ensure that the application runs at peak performance. Most of these environments will include an Integrated Development Environment (IDE) running in the same platform to facilitate faster development, debugging, and testing.

In the end, platforms provide the developer a comprehensive system management and development tools to perform their core functions. In doing so, the costs of building and maintaining the platform are no longer in the hands of the developer, freeing up their concerns to focus on delivering a better product.

7.2 *Providing the Best Experience of the User*

A PaaS solution has two user bases. The first set of users are the people using the platform to build and deploy an application – the software developers. The second set of users are the people using the application – the end user. Some providers have recognized a potential third set of users who are interested in understanding how the application is performing and the data it is generating – the business manager. The provider of a platform-as-a-service offering must be concerned with all sets of users when maintaining the environment.

The first concern with user experience is realizing the extent in which users tend to multi-task during the day. For the end-user, these can be a variety of tasks ranging from data entry, to reviewing a document, to communication like instant messaging or email. The software developer can be working on building code and needing resources to understand the code better, debug the code as its written or find examples of similar code. Business managers are performing multiple queries, looking for trends and problems, and compiling data for future reports. The work each of these users performs may be contained within the platform environment or strung across several resources available to the user. Working within the platform should facilitate key multi-tasking activities for each set of users. This can be done simply by allowing the user to go outside of the platform to find the resource and use it. An interface for the user could pull those resources into the interface so that the user does not have to change windows. To enhance the user's experience, the interface could be configured to automatically pull those resources together based on what the user is currently working on.

No matter the task or the user, the core component of the user experience is the user interface. The interface should be intuitive and easy to use. It should allow the user to customize views based on their work habits, without sacrificing functionality. Once configured by the user, the interface should retain the settings for following sessions. Many interfaces provide embedded applications, like Microsoft Word or Excel, or Adobe Acrobat so that the user's system is not bogged down by running multiple applications.

7.3 *Providing Built-in Insurance*

Scalability is a pertinent feature of all PaaS solutions. The primary reason is the difference is usage for an application in development and in production changes dramatically. Costs of business does not warrant “reserving” the necessary space and storage required in production while the application is being developed. Ideally, the platform offering should increase in these areas when it is needed and not before. In the same vein, as end-users start to use the application, the demand on capacity, load balancing, and bandwidth will

increase. In traditional on-site environments, these increases would have to be planned out months before and if there were any spikes in demand, it was highly probable that the infrastructure could not support the increase. In an online utility offering, these components can be released when required, usually at the flip of a switch or automatically based on parameters configured into the infrastructure.

The same concern for scalability exists for when demand peaks and decreases. Customers do not want to pay for what they don't use and providers do not want to have unused resources tied up with a single customer. So the offering should be able to decrease the availability of resources when they are not longer required.

The other concern is for PaaS solutions is reliability. In this respect, everything has a tolerance of reliability expected of it. The infrastructure has to have the greatest level of availability to the developer. If the system goes down or disruptions of any kind keep processing from completing, reliability in the system decreases. The end-user needs to rely on the application performing the tasks requested of it. The infrastructure and the application code ensure this happens. Therefore the infrastructure needs to support the demands of the application and performance issues with the application code need to be identified and delivered to the developers for resolution. Another area is finding and accessing the appropriate data for the session. If the application sits on the server and the data on a storage device, there must be a continuous mapping between the two. If either the application or the data moves from one device to another, the mapping needs to change as well to ensure that no disruption is present when the data is required.

Backup and recovery solutions provide additional insurance for reliability. Used to resolve any situation where loss or corruption of data happens, properly applied backup and recovery procedures makes sure that system and data integrity is maintained. Other methods such as redundancy and mirroring provide additional reliability by duplicating the data and having it available in real-time speeds in the case a system component fails. Altogether, these solutions and methods provide a sufficient foundation to handle major disruptions and outages due to disaster.

Finally, security is a definitive concern for most PaaS. The customer data, network traffic, and application code are all areas where security measures need to be taken to ensure that the system is free from malicious or accidental attack. For an online platform offering, security needs to handle vulnerabilities that are possible from several avenues of breach, primarily the end-user, the company, and other customers found within the platform offering.

7.4 *Communicating Outside of the Platform*

Within a Web platform, connectivity is the key to success. Most applications built today are dependent on multiple resources found outside of the source code. Dynamic data used by the application can be found within an internal database or from external sources. Real time data from multiple sources may be compiled by the application for processing. The application may be leveraging web services found internally and externally of the platform. The truth is that any application found on the platform may have functionality that required a constant connection to other components found within the platform or from an external source. Since connections required may be different applications, the platform needs to support all connections.

State Management is a major component for maintaining these connections. The term is closely tied to the web application framework, ASP.NET, and refers to a process where session related and control information is maintained. The critical need for state management is rises when multiple users attempt to access the same of different Web pages within a Web site. State Management happens from the server or from the client.

Server-based state management can happen at the application layer by storing application specific information as a key value pair. Application state refers to the process of storing information in server memory, specifically information that is used often by multiple users. This is different from caching which has a similar process. The difference is that state management actively manages the stored data and will remove items when more memory is needed. The storage of this information is retained until the application is stopped or restarted. As soon as this is done, the data is erased unless the information has been saved to a database or the like. The application state refers to data stored on each instance of the application, so if multiple instances of the same application are found across several servers, the application state is not shared. Nor is the application state shared with multiple user sessions within the same application on the same server. Data in application state can be accessed simultaneously by multiple threads.

Session state provides a way to store and retrieve values during a user session. The data stored in session state exists as long as the current browser session exists. If multiple users are accessing the same Web application, a different session state will exist for each of them. If the user leaves the Web application and returns, a different session state will be created. There are several ways to store data from session state. The default is to store in memory found on the server. Data can also be stored in the separate process or on a SQL Server database. These two methods will retain the information if the Web application is restarted or available to multiple Web servers in a redundant environment. Session state can even be disabled.

State management can also be found on the client side. Client side state management does not utilize any server resources. View states automatically retain data when the same pages are accessed multiple times. View state values are secure by through hashing,

compressing, and encoding. Hidden fields have no security at all and are simple to implement. They store small amounts of data that change frequently. Cookies can also be used and have a configurable expiration applied to them, which allow the data to be released at the end of the session to the client. Query strings can also retain information on the client.

7.5 *Supporting Collaboration*

At one time, software development was typically assigned to a single person or a few people who would develop, maintain, and even market a commercial product. Today, it is a team effort to the point where studies have shown that 70% of a developer's time is working with others. Dividing a software project into parts which individuals are assigned to focus on can be a laboring project in ensuring that the "team" works together effectively and efficiently. Principles from Computer Supported Collaborative Work (CSCW) have provided the framework for created collaborative tools in software development. Some of these tools have been built or found on the platforms provided on the Internet. The greatest goal for a collaboration tool within a platform offering is to promote collaboration throughout the entire development process, though many of the tools developed have not been able to accomplish this.

Collaboration focuses on improved teamwork; a goal that facilitates meaningful interactions with team members in such a way that the opinions and skills of the individual can be refined by the contributions of the whole. Many projects cannot be divided easily so that each component is worked on independently from each other. This interdependency requires a method from which the individual can work the tasks assigned to them, identifying where they need another team member's input without too much disruption to the task of either, and to incorporate that input easily into the task at hand.

Synergy within the group needs to be encouraged. The tools should support group reviews and communication whenever possible. From an Internet platform, team members may be geographically distant requiring technology to bring the group together. The promotion of synergy can uncover small errors in the code and identifying factors that were not originally seen at project start. Team conflict can corrupt any synergy if not dealt with appropriately. Online collaboration has the benefit of minimizing the human impact of collaboration, bringing a level of professionalism that is not necessarily there in a conference room.

Successful teams usually continue to have success in the project and other projects as well. Many people have tried to define the characteristics of a successful team. The most common is that the team strives for high performance from all team members. Successful teams have a strong identity in group dynamics and individual contributions. There is a

high commitment to integrity and quality.

For team-based software development projects, look for collaboration tools that handle communication, interdependencies, synergy, and conflict while providing a positive environment that promotes performance.

7.6 *Managing the Application from the Back End*

Software development does not end with the deployment of the application onto the hosting site. Most application developers continue to monitor the application in production to identify emerging issues and potential improvements. There's a regular rule of thumb that customers don't know what they want until they have a chance to play. This is very true for applications as users discover how to use the program.

Platform-as-a-service offerings that handle application in production should have a view into application and user activity. This is typically provided through some form of instrumentation. The deeper the instrumentation goes into the application, the more information that can be retrieved on user activity. What the user is doing, the performance they are experiencing, and any errors that arise from those actions can provide a rich report on the strengths and weaknesses of the application.

8 Not all PaaS Solutions are Equal

In choosing the platform to be used, it is important to remember that platforms offered from one provider to the next are not the same. Nor would a person use the different platforms in the same way. There are several considerations that a business or developer should think about before choosing a platform.

The existing skill set and system infrastructure of the business may restrict the level of platform that is can be used. The infrastructure of the platform will define which applications can be developed. The standards for service delivery used by the provider will determine the relationship with the customers. And of course, cost is a major concern.

Social Application Platforms

Some applications are available to communities throughout most of the life cycle, including the development phase with the intent to illicit collaboration with independent developers and users. Based on the concept of social software, these platforms allow individuals and groups who are not directly connected to the platform customer to communicate, collaborate, and to build community within all phases of the application. In this way, the primary developers can capitalize on the knowledge of others. Facebook is an example of this type of platform.

Raw Compute Platforms

In these instances, the developer is allowed to upload their software stack to the platform and run it. The service provides the storage, processor, and bandwidth required by the application. These applications can take on many forms. The only guarantee by the provider is the provision of the infrastructure. Amazon Web services focus on delivering these platforms.

Web Application Platforms

Web 2.0 applications can be built and maintained using these types of platforms. The APIs and web functionality typically used by developers are already in place to add to the web application. Google Apps is the most predominant provider of this platform, which allows developers to leverage the functionality that Google.com provides.

Business Application Platforms

For applications that cannot compromise on scalability, reliability, and security, business application platforms provide a reasonable choice. Especially for transactional business software that require database usage, integration, workflow, or user interface services, this platform allows the customer the ability to manage critical business processes. Force.com offers such platforms.

8.1 *Social Application Platforms*

Social application platforms allow the development and use of applications by the public. Open source capabilities have made this type of platform for many businesses, developers, and users. OpenOffice.org from Sun Microsystems is probably one of the most popular social applications around. Open source is a development methodology that offers practical access to the application source. For some, open source is strategic to their operation and decision making because it provides concurrent access to different agendas, approaches, and priorities. Typically, the source code for the application is open to public collaboration until the application meets minimum objectives, than it is releases as open-source software that can be used. Additional development is done through updates and new releases.

Social applications have several key characteristics.

Easy Content Creation and Sharing

It used to be that Web content had to be created by savvy individuals familiar with HTML and Web programming languages. Social applications now allow individuals to add photos, text, audio, and video without any specialized knowledge of skill. Blogs and wikis are other examples of social applications where content is easily created and shared. For application development, some technical knowledge may be necessary in order to collaborate effectively, but for their skill set the risk is acceptable.

Online Collaboration

The key here is that the collaboration is conducted in the same platform as the application. How the collaboration happen can vary between instant messaging, color coding, comments, and the like. The effect is similar to having a conference meeting to discuss a point in the project. The difference is that the dialog is online. In fact, the anonymity of working online can sometimes increase the contribution of the participants.

Distributed and Real Time Conversations

Distributed conversations across locations and time allows for contributions by more participants. Real time conversations are synchronous conversations of two or more participants typically joined to discuss a single point. Any of these conversations can be formal or spontaneous.

Bottom Up Communities

Many online communities are formulated in one of two ways: top down or bottom up. The top down method is simply some entity wishing to create a community and providing a structure for that community to be informed, related, and collaborate. These types of communities usually require membership and thus are exclusive. The bottom up method is generated by the community itself, based on the affiliations of the community members. There is no membership required. Social applications support a bottom up community.

Some software development projects have a combination of the two approaches. The core group is a top down community of developers, business managers, and project managers who make the executive decisions related to the application. They oversee the development activities of the application by a secondary community of programmers, users, and software vendors who have been generated through the bottom up approach.

Knowledge of the Many

Social applications, specifically in the development stages, capitalize on the aggregated knowledge, skills, desires, and behaviors of the group. In doing so, the development process is often shortened and creates a better product than originally envisioned. Wikis, recommendations, and tags are all collaboration methods used to gain the collective knowledge of the community.

Transparency

Collective knowledge creates transparency. Whenever a software product is released, there are immediate reviews that could enhance or diminish potential sales. By allowing the application to develop socially, it creates an amazing openness, communication, and accountability by the community participants. This in turn creates trust and confidence with the users and critics.

Personalization

In any effort, development or otherwise, the tendency of the individual is to bring their own agenda. As long as that agenda doesn't overshadow the primary objectives of the project, there is no problem to this. In fact, personal agendas can often enrich the product with benefits not foreseen from the start of the product. This phenomenon creates ownership

and personalization for the individual. In doing so, they work harder to ensure that it is successful. When tied to the primary objectives of the project, it literally drives the project forward.

Portability

Social applications found on the Internet have an enormous capacity for portability. As long as the participant has access to the Internet, they can contribute. In the modern world with the mobility of laptops, PDAs and cellphones, that access can be anywhere at any time. Hence, the barriers of location and time are no longer a factor.

8.2 *Raw Compute Platforms*

In raw compute platforms, the provider offers the infrastructure in which the developer can house their application for use on the web. In short, they have a shell within the multi-tenant environment. In most cases, the development of the application is separate from the hosting environment. In fact, many raw compute environments may not support software development, debugging, or testing on their infrastructure.

If development is possible, the developer may have to provide the IDEs and the collaboration software required to get through the development process. When the application is in production, the developer may be responsible for providing any monitoring software. Simply put, the developer is responsible for anything outside of the processor, the storage, and the bandwidth provided by the service. However, some platform providers may provide tools for the developer, as well as providing control over those tools.

Raw compute environments attempt to provide the developer the maximum control over the platform without sacrificing the integrity of the entire multi-tenant solution.

Many system administration tasks are given to the developer to control, such as access control, determining the locations where the application is located, even specific memory allotments to the application. The extent of the control and the tools available on the platform is dependent on the provider.

8.3 *Web Application Platforms*

These platforms are specifically geared towards assisting developers in creating web-based applications by making available a series of functions already built and running on the web, like mapping, calendar, and other services. In many cases, the coding is developed by the tenants of the platform and those that are shared are gathered in the

repository managed by the provider. In other cases, applications are created as APIs, plug-ins, or full applications that can be attached to a web application and are found in a catalog of the services. These applications are not restricted to those developed within the cloud's tenants.

Like any other platform, the provider is responsible for the platform infrastructure. In a web application platform, the provider generally provides a comprehensive set of tools to assist the web application developer with a highly simplified, fast approach to developing applications. These platforms are typically geared towards the single developer instead of teams. Collaboration tools found on these platforms are focused on encouraging partnerships with other independent developers creating web applications.

For the developer, the web application platform provides the easiest approach to designing and building web applications with very little financial commitment to hardware and software generally required by development and maintenance of an application. The platform allows the focus of the developer to remain on web design. Even in the development of the application, the web application platform attempts to assist the developer.

There are two ingredients to a web application to delineate it from bad and good web applications: appearance and content. Appearance requires a creative use of design elements and styles. It defines how the users interact and enjoy the web application. The content of web application organizes information in a way that is easily read by the user. Since the web is primarily a publishing medium, this includes different fonts, graphics, animation, video, and sound. Web applications provide the ability to relate to the user while delivering content in a visually dynamic and appealing manner. Web application platforms must support the vision of the developer as found within their application.

8.4 *Business Application Platforms*

Business application platforms are specifically marketed to developers of business software, a program for increasing or measuring the productivity of the company. The term, business application, is a broad description of the various uses of business software; some of the more popular uses are office productivity suites, accounting software, software that handles core business processes like customer relationship management, human resources, and order management, as well as software that handles more complex business relationships like resource planning, content management, business process management, and product life cycle management.

Though some of these applications may find their way to the desktop of the employee, many of them are also used from a management perspective to oversee the activities of

the department or line of business. By analyzing the performance and use of these applications, the business can identify areas of improvement, including bottlenecks, misrepresentations, and failures in the business. Several tools may be used to promote this management. The simplest is the ability to data mine available information, or the ability to extract information from database to isolate and identify unknown patterns or trends in large amounts of data. This usually leads to some form of reporting to communicate the findings, which can be done through software specifically designed to generate reports from raw data. Online Analytical Processing (OLAP) offers management the capability to review this data from multiple perspectives to better enable decision making. Digital dashboards are visual representations of business summaries that are based on metrics and Key Performance Indicators.

Some platforms-as-a-service platforms are designed to fully support the development and maintenance of web-based business applications. They provide the infrastructure, security, and flexibility to allow the business application to have success in the future. This may range from simple infrastructure maintenance to establishing and managing connections to internal and external resources required to run the application. These platforms also provide business monitoring tools to demonstrate the success of the online application. Though these tools may not allow deep reporting specific to individual functions of the application, they do provide reporting on how the application is used on the web.

Many of the developing business applications may be directly connected to the business work flow, which is the relationship between multiple business processes as a request is processed. For example, a customer transaction may start with the registration on a web site and typically some sort of order for a company product or service. This order needs to go through some sort of processing to locate the product. Then the product has to be retrieved and packed up to be sent to the customer. At the same time, payment of the order needs to be processed. This single request goes through a number of business processes and the tasks are performed by multiple people. Typically one application may not cover the entire flow of work, so it is necessary to send information from one application to the next to handle specific components of the request. For a web-based application, these connections may be with other web-based applications or from a variety of traditional applications. The platform needs to handle the connection, the transmission and the translation of the information as it moves from one application to the next. The infrastructure needs to handle a typical daily load with ease, including any spikes in usage. Often, single requests are not sent through the system during peak times, but are consolidated with other requests and sent at slower times.

Business applications typically require more controls, more security, and more performance assurances than other applications simply because of their possible involvement in critical business workflow.

9 Benefits of PaaS

Platform-as-a-service solutions impact the speed and cost of computing on the Internet. By providing a “sandbox” within the environment where applications will eventually find themselves, developers can provide direct mapping to resources on the web, test access to those resources and utilize test information more effectively, than if development of the application was in a isolated “clone” of the environment. Increasing the speed to production for development saves money.

Additionally, money savings are available to the company because of the burden of maintaining the IT infrastructure is no longer with the company. Since the provider maintains the same infrastructure for multiple customers, the cost is shared across the customer base usually at a cost savings to the customer.

Without having to worry about the infrastructure, the company can focus on development and creativity rather than the technical support of building an application. Some PaaS providers may even include predefined business functionality through the use of IDEs to jump start development projects. In many cases development is happening on the cloud, so deployment simply entails opening the access to the application when it is ready. When development is not performed within the cloud, deployment is still much simpler to perform because of the use of web-based functionality.

Platform-as-a-service provides greater reduction in risk. First of all, risk management is a primary concern to providers of PaaS solutions. As a result, state-of-the-art technologies are often used with considerable expertise from the system administrators maintaining the infrastructure. This combination of technology and skill can be more than the customer can achieve on their own. Most infrastructures of PaaS solutions have to be built with the strongest security policies in place to prevent attacks from both sides of the firewall, as well as protocols to automatically handle loads that may impact capacity and performance.

A critical task of delivering a platform for software development and deployment is staying abreast of the appropriate upgrades, patches, and other maintenance activities. These activities ensure that the infrastructure can handle any new concerns that may exist in the current environment as well as persistent issues resolved by manufacturers and providers of the infrastructure. In most cases, these activities are a daily occurrence of finding vulnerabilities within the system. For the customer, these activities are transparent and no longer their concern.

10 Different PaaS for Difference Reasons

The delivery of platforms-as-a-service is still developing as current providers are growing their current offerings to match the demands of the market and new providers are attempting to find their own niche.

Different platforms are set up to handle different application models, using access APIs, plug-in APIs, or an environment for running the entire application. Providers design their platforms to handle classes of different applications, whether their applications are used for social purposes, raw computing, web applications, or business productivity.

Platforms delivered on a utility basis can be further classified based on the scope of work allowed on the platform. The different types of platform-as-a-service offerings are:

Add-on Development Environments

These environments facilitate the development and hosting of applications that extend or enhance existing applications, specifically a SaaS application. In most cases, the developer and their users have to subscribe to the SaaS application in order to run the add-on application. In some instances, development is still undertaken in one environment while actually running the application in another environment. In other instances, the add-on is available within the platform, but stored in a separate environment.

Stand Alone Development Environments

Stand alone development environments focus on the development and deployment of applications that are intended to be used independently from any other application code. Though the application may be open to having add-on developed from other sources, the original application code is necessary for use by any end-users. The primary advantage to these environments is the nonexistence of any dependencies on technology, licensing, or finances.

Application Delivery-Only Environments

Some environments do not offer any capabilities to develop, debug, or test applications. They will only host the application for the development in a production state, providing the necessary scaling, performance, and security necessary for the application.

10.1 *Understanding Add-on Development Environments*

Add-on Development Environments are in place to support the development of applications that support other applications. These enhancements may be created by the original developer or by third party developers. In these cases, the original developer opens the application code to review by the developers with strict guidelines on what can be changed or how enhancements work with the original application code. In addition, the platform that houses the application code and the enhancement application should allow a connection each other. In some cases, the enhancement requires going through a certification process to eliminate or demonstrate any compatibility concerns. The bottom line is that the success of the add-on development environment is directly impacted by the relationship between the original code and the enhancement.

These types of environments can allow two methods of access: automatic and direct. Automatic access is typically initiated from the original application, either because the user identifies that the enhancement is needed or because the system makes the determination. Either way, a request for the enhancement code is sent from the application to the platform which houses the enhancement. Because the request is coming from the application, access to the code is typically allowed. The code is than loaded into the session layer for that user. The metadata controlling the user's preferences are updated with the new request, so that every time the application is loaded in resulting sessions, the enhancements are loaded as well. Direct access to the enhancement has the user retrieving the enhancement rather than the original application. This may require some form of authentication or not based on the controls in place for the application and the platform.

The platforms that support these enhancements need to be highly flexible to handle the access and transfer loads. Typically, these environments are tightly connected to the original application and the enhancements have to be registered within the platform. This registration allows the enhancement to be catalogued. The catalog is managed by the original developers and is in place to allow users to know what enhancements are currently available.

10.2 *Understanding Stand Alone Development Environments*

For developers looking for platforms that provide development, deployment, and maintenance of applications on the web are looking for a standalone development environment. These environments are typically exclusive to applications that are independently developed and for proprietary use either by a business or a registered set of

customers. The platform itself will usually provide some extensive tools to support the development and production states of the application.
Most businesses looking to provide applications for their internal operations will be found in a standalone development environment, while applications that for commercial customers may be found in multiple environments. In a standalone environment, the developer is typically provided a set of tools to allow the development, collaboration, and application management activities to be conducted easily and with the greatest efficiency. These types of platforms are typically not conducive to developing applications within minutes or even a few days, but over a few weeks or months.

With the scope of the application being rather large and the tools available to the developers by the platform provider, there may a learning curve before the developer can truly start using the platform and all of its features. Typically the provider will assign a case manager to the developer to ensure that the learning curve is not to severe. In addition, once the developer learns the tools available to them, they are more likely to utilize the platform for additional developments.

10.3 *Understanding Application Deliver-Only Environments*

Many platforms offered do not allow the development of the application, but will host the application for the developer once it is created. These environments cover the capacity, the bandwidth, and the location to house an application for use. Most of these applications are commercially driven. Therefore, the platform requirements focus on ability to handle large volumes of transactions at any given time.

For the development, hardware and software for development purposes are still their responsibility. Some platforms that they deploy to may have strict configuration standards that need to be adhered to or the platform is open to any configuration. Typically, the tighter the configurations for the infrastructure, the less expensive the service.

11 Cloudware Players

Understanding the players within a cloud computing environment is valuable for understanding the deeper workings of providing a platform to a business or individual. Each group has their responsibilities to meet, objectives to obtain, and dependencies that need to be fulfilled by other players in the network. In some instances, groups of players may actually be the same entity but are separated here to demonstrate their role in the architecture.

Subscribers
This group consists of businesses that are looking to use platform-as-a-service offerings to develop and deploy their applications.

Publishers
When a subscriber starts to use an offering, they often have access to a global catalog of published applications, tools, infrastructures and platforms that enhance or expand the original offering. The items found in the catalog are provided by publishers. In the business world, the company may subscribe to the service while the developers publish to the service.

Data Center Operators
One class of publishers (and the primary one for the offering) are the data center operators who provide the servers, storage, and network connectivity for the platform.

Vendors for Integrated Web Services
A variety of services are available on the Internet, many of which may not be included in the global catalog because the services are assumed because of their popularity or because the services have not been published into the catalog.

Providers for Outsourced Services
In addition to the data center operators who support the infrastructure of the application, some of the other activities for developing and managing the application can be managed by other resources, usually through outsourcing the work.

Clients
Clients are the users of the Internet that may access the published resources.

11.1 *Sponsoring the Cloud: the Subscribers*

Most conversations found in the media talk about the benefits of cloud computing and platform-as-a-service offerings. Those benefits range from cost reductions to the ability to scale applications to better connectivity. Cloud computing definitely has numerous benefits available to those people who take advantage of it. A subscription is often required to access the services of utility-based computing. Subscription often entails registering the financially responsible party. That party may be an individual for social platforms, or small and medium businesses, Web 2.0 and SaaS companies, and large enterprises for other platforms.

Most subscribers seek utility-based platforms to alleviate the burden of owning and managing servers, data centers, networking, or anything else associated with supporting a computing infrastructure. With their subscription, they can deploy applications or middleware onto the platform, scale their applications dynamically, or allow access to the application from around the world. They may use these platforms on a permanent basis or to cover excessive workloads or projects on a temporary basis.

Because the subscriber has responsibility of making financial payments for the use of the platform, they usually have specific business goals and objectives to meet. The platform they choose must be able to fulfill their short and long-term goals along with the cost reduction provided by their platform-as-a-service subscription. These goals can range from realizing the benefits of flexibility and scalability of cloud computing to obtaining market leadership through global positioning on the Internet.

Subscribers are dependent on publishers to ensure that the service purchased is being utilized effectively and efficiently, and on the clients for using anything that is published on the platform. In many cases, subscribers have access to anything that is published within the cloud platform.

11.2 *Creating the Cloud: Publishers*

Publishers make up compilations of vendors for independent software, virtual appliances, infrastructures, platforms, and tools. These vendors can publish appliances, ready-made architectures and applications. Whatever the vendors create is found within a global catalog. Each platform-as-a-service has their own global catalog, though some items such as general Web APIs and plug-ins can be found in multiple catalogs.

Publishers can decide which subscribers have access to published items and at what

price. This is definitely beneficial for social platforms that are built based on various publisher contributions. For platforms focused on business application, the publisher may choose to share application code with other publishers or provide finished products to clients.

The majority of publishers are application developers. They can build applications that support specific subscribers, for use by any subscribers, for use by other developers to enhance or extend their applications for publishing, or for commercial clients. Their applications may be free or have a pay per use charge to them. In some platforms like Second Life, the charge may be virtual charges that are only valid within that platform.

Other types of publishers can be found within the Internet. Hardware appliance vendors can create virtual software equivalent of their appliances, such as firewalls, load balancers, security appliances and the like. Vendors of platform and middleware publish software packages that are ready-to-use without any sophisticated installing or configuring. Even entire architectures can be found on the Internet and were published by professional experts: examples being LAMP, Ruby-on-rails, J2EE.

Publishers rely on the Data Center Operators to make good on their promises to maintain a platform that is reliable, scalable, and secure, as well as maintaining the Global Catalog. Clients and subscribers using the published products provide publishers immediate feedback on the value of their product. For many publishers this feedback may be in the form of revenue. For products that are free, the feedback may be in terms of popularity.

Any application, appliance, service, or even Web site found on the Internet is delivered by a publisher. Without publishers, the World Wide Web would not exist.

11.3 *Supporting the Cloud: Data Center Operators*

Every utility-based offering has a group of individuals ensuring that the infrastructure supporting the offering is working as expected and handling any unexpected problems that may arise. These activities are the core of what is called data center management and the people supporting this process are the data center operators. This group is mostly transparent to the operation. Case managers may be assigned to a subscriber. Sales people may be available for subscribers and clients. Customer support representatives may be available for questions and problem reporting. However these people are a small part of the data center operators. The majority of this group has responsibilities directly tied to the maintenance of servers, storage devices, network connections, software and tools.

Data center operators are a special form of publisher: what they publish are large

infrastructures to handle hosting, managed services, enterprise data centers, as well as other services. As publishers, they determine the pricing schedule for the resources they make available, who can use those resources, and in some cases, how those resources are to be used.

The purpose of the data center operator is to maintain the reliability, availability, and security of the infrastructure. Since cloud infrastructures are mostly virtualized, these operators are responsible for implementing and maintaining any virtual controls required. They set up the configurations and automation controls to allow any number of network features from workload balancing, replication, and backup of storage.

The data center operator is dependent on people and businesses using the infrastructure. Many utility-based providers are doing so in addition to their primary business. Providers like Amazon.com, IBM, EMC2, and Google have a core business that was successful before offering utility services. The data center operations typically involve supporting these companies core business and their utility-based offerings.

11.4 *Making the Cloud what it is: Vendors for Integrated Web Services*

Web services that are integrated into the platform offering can be valuable to all subscribers. Typically these web services are not offered by the service, but are made available by the service. The source of the web services come from a series of vendors who have developed these services specifically. The World Wide Web Consortium defines a web services as “a software system designed to support interoperable machine-to-machine interaction over a network.” The most common Web service is in the form of Web APIs access over the Internet or any network and executed on a remote system hosted the service.

Such services typically fall into two categories: Big Web Services and RESTful Web Services. Big Web Services use the SOAP standard to build XML messages. They may even include a machine-readable description of the service operations written in Web Services Description Language (WSDL). These services have been the most popular for years now. However, RESTful Web Services are gaining popularity. Based on the REST protocol, these Web services tend to integrate better with HTTP than SOAP-based services. They also do not require use of XML or WSDL.

Web services can be used in a number of ways; the three most popular are RPC, SOA and REST. Remote procedure calls (RPC) are a technology between processes that allow an application to remotely execute a subroutine or procedure on another computer in a

shared network without any explicit coding for the interaction. Service-oriented architecture (SOA) web services are based on the architecture and make SOA functions accessible over standard Internet protocols without any dependency on the platform or the programming language. Representational state transfer (REST) services emulate protocols by limiting the interface to a set of standard operations. Any of these Web services may be required by the applications and services found on the platform.

Web services add the needed communication to a number of value-added tasks pertinent to using and monitoring applications on the web. They can be used by monitoring tools, billing services, transaction trackers, engines for storage and policies, and the like.

11.5 *A Twist of Support: Providers for Outsourced Services*

Even with the benefits of utility-based services, which allow businesses to relieve the financial and work effort burden so that core businesses can be focused on, some activities are still required by the business. This could be from application development, to monitoring the application in production, to customer support, to management of the application. There are several technology service companies that have provided overall management of business operations for companies. This is typically referred to as managed operations and encompasses the whole of the IT solution.
Even though cloud computing has alleviated much of the burden for managing IT solutions; some companies still look to outsource activities for IT management to other providers. They may not have or afford the skill required to take on this management. They may not have the tools required. Or they simply would rather not tie their efforts up in such matters.

Providers for outsourced services are people or organizations that specialize in translating the customer goals into tangible activities and assets to fulfill those goals. They are typically bound by contract to achieve certain Service Level Agreements (SLAs). They may choose to use utility-based services to fulfill the contract or they may be bound to use such services by the contract. Some providers of platforms-as-a-service will provide outsourced services to their customers or vice versa.

In large enterprises, the delineation of outsourced services is still viable where one department is responsible for the infrastructure while another is responsible for application development and still another is responsible for monitoring systems. If these tasks were performed by one department, the required communications and collaboration would be assumed and easy to manage. When the tasks are separated, greater commitment is needed to build cooperation between the departments. The same commitment is required when the departments belong to entirely different company. Large enterprises are starting to understand this requirement and are creating “contracts” between departments to

ensure that this cooperation is in place. For businesses, understanding and managing these relationships may prove crucial to meeting goals and objectives of the business.

11.6 *The Marketplace: The Clients*

Used interchangeably by most people, the Internet and the World Wide Web are actually distinctly different. The Internet specifically refers to the hardware and software assets that are connected to each other through a variety of methods and the World Wide Web is a commercialized service found on the Internet. The concept of the Internet originated with the 1946 science-fiction story, A Logic Named Joe, written by Murray Leinster. However, the reality of the Internet didn't emerge until 1958 with the creation of the Advanced Research Projects Agency (ARPA) in response to the USSR's launch of Sputnik. From then on, the Internet infrastructure grew. However, it wasn't until 1988 that the network was opened for commercial use by the MCI Mail system. The Internet as we now know it didn't exist until 1991 when CERN announced the World Wide Web project.
Though there are many critical reasons for the dramatic growth of the World Wide Web, the most basic is the accessibility of the technology to the common person. The business, education, and social landscapes have changed drastically because the common person can easily access the Web with little knowledge and just a few clicks. In today's world, computers are not even required to access services on the Web. Many kinds of device, such as cell phones, have the capability to connect to the Internet and the potential to access a variety of services.

The success of the Internet, the World Wide Web and the services found in either are directly contributed to the willingness of the people who use them, the client. For any utility-based solutions or the application found on it, the same is true – success will come from the client. Given this condition, publishers and providers need to design their product to be easy to use, easy to access, and with an intuitive and flowing interface that is visually appealing to the client.

For the provider of the platform, their direct clients are the subscribers of the platform and the publishers within that platform. The provider's task is to provide an environment of ease and flexibility for these clients so that they can realize the goals and objectives that have them using the platform in the first place. This typically boils down to quick and efficient development periods, accurate accounting of application use, and connection to the functions and features found within the World Wide Web.

For the publisher, the client is typically the common person; though for business applications, the client base may be restricted to the employees using the application. The

same focus on client usability pertains to any client base for the application. With many applications connected to facilitate work flow, the actual client may be another application. From a client perspective, whether human or computer, the important thing is the transmission from the source to the client. From a human perspective, the transmission may involve a number of tasks that are transparent to the client and should remain so. The quality, speed, and transparency of the communication between the source and the client are concerns that the publisher should be consciously aware.

12 Cloudware Elements

Internet solutions are a merging of man and machine. Man determines what he wants and machines deliver. Machines determine what is needed based on preconfigured settings and man makes use of the result. Man programs the machine and machines carry out the instructions. This is a simplified description of the symbiotic relationship with today's man and machine capabilities. Understanding the players found with the cloudware architecture can assist in understanding the elements found within the architecture.

Infrastructure Delivery Network
This is the core of any utility-based service. In fact, it is the core of any IT solution for individuals and businesses alike. It consists of the servers, storage devices, and network appliances that make up the infrastructure and the software and firmware that runs it properly.

Resource Pools
The mixture of hardware and software may create the infrastructure, but the key distinction for cloud computing environments is how the infrastructure is configured so that multiple resources are combined to create a single pool of resources and how that pool is used. Though any computing environment can create resource pools, cloud computing (specifically utility-based services) takes this concept to its maximum capacity.

Global Catalog
The global catalog is a service that is distributed worldwide that allows publishers to make their products available to subscribers.

Control Interface
To control application and services running in the cloud, a mechanism must be in place to view the state of operations in the environment. That mechanism is viewed through a window into the cloud, the control interface.

12.1 *The Core: The Infrastructure Delivery Network*

The primary product of utility-based product is the infrastructure. The whole of the infrastructure consists of a variety of hardware, software, and services. But more importantly, it is how those components are handled and distributed within the cloud. The

infrastructure ensures that all the components are working together cohesively. It manages the number of resource instances, primarily applications, that are found on the network and balances the workloads. The infrastructure supports the reserve pools, the global catalog, and the control interfaces. The importance of the Infrastructure Delivery Network cannot be overstated.

The infrastructure manages authentication processes to itself and to the services it supports. Because it supports a multi-tenant environment, any opportunity for authentication should be taken as long as the process does not hinder reliability, scalability, or performance. Authentication at every layer of the application stack should be utilized with several iterations at the network and application layers. Authentication ensures that right resources are accessible by the right subscribers and clients by routing requests through the network appropriately. Since different individuals may have specific levels of access, further access controls may be in place to determine the access. In platform environments, access control is typically managed through access control lists. Those lists are managed by the publishers to control the application access.

In environments that allow development and production, the infrastructure manages the transactions during deployment and migrations. These complex activities can involve moving raw data from on device to the next, setting up replication and redundancies in the network, remapping data locations, and numerous other tasks. The infrastructure manages the backup activities of the network. It performs full, incremental, and differential backups when scheduled. It also manages any recovery operations that are required when disruptions happen. Of course, recovery is done in conjunction with automated processes to balance workloads in case of excess or failure, rerouting work from one component to the next.

In a platform environment, the infrastructure is responsible for many activities, most of them happening in just a few milliseconds. To maximize the performance of the infrastructure, most of these activities are triggered automatically based on the settings configured by the data center operators. Yet, the infrastructure cannot do it alone, as it requires the resource pools, the global catalog, and the control interfaces that allow the whole package to work.

12.2 *Supporting the Network: Resource Pools*

Many of the resources of a utility-based platform are taken from a pool of resources; a single dump of all like resources on the network. For instance, ten separate storage devices with ten terabytes of storage space is actually a single storage pool of 100 terabytes. The mechanism that allows this to happen is the technology of virtualization,

which creates an abstraction of what really exists. Inside this abstraction, resources can be manipulated to become what they are not. So ten devices can become one and one can become ten. In fact, virtualization can turn the abstraction of 10 terabytes into 20 terabytes.

The execution of virtualization is critical for the success of a cloud platform.
In most traditional networks, at least one storage device is directly connected to an application server. That server stores the appropriate data in that storage device. If multiple instances of the application are run on several servers with directly connected storage devices, only the data going through that application server will go to that specific storage unit. The first downfall to this architecture is that most of the capacity provided by the individual storage devices were being unused by the system. As applications and systems became too complex, new solutions needed to be created. For storage, the next major move was the creation of storage area networks (SAN). In a SAN, the servers are connected to a network, usually called a LAN (Local Area Network). The storage devices are connected to another network – the SAN. The LAN and the SAN are connected to each other, allowing any server to connect to any storage device. To ensure that data is not corrupted or misrouted, the system creates mappings for each bit of data stored in the SAN. The mappings are accessible from any point in the LAN, so that any bit of data can be retrieved by any server in the network.

Virtualization takes the concept of SAN one step further. Through abstraction, the SAN essentially becomes one large storage device. In doing so the capacity of all the physical storage drives are “pooled” together resulting in a single capacity store. In the background, the data is still found on multiple physical drives and the mappings that are created are still used to retrieve the data. But the capacity is treated as one. By doing so, the use of the capacity is more efficient because there are no hidden stores of memory. Virtualization can even abstract more or less capacity than there really is or create virtual partitions of storage to dedicate resources for certain tasks. Because storage is abstract, changes to capacity requirements can be made on the fly without causing any disruption to the environment. One benefit about virtual storage is it does not utilize any additional power or effort to manage.

One of the defining conditions of a network is its ability to conduct data transfer. In a large solution like the World Wide Web, the number of individual data transfers at any given time is staggering. The likelihood of two data transfers colliding can be imaginably high. Controls, services, and further virtualization can avoid this undesirable scenario. The most direct solution is the techniques of caching or buffering. The technique simply stores the data for one transfer until the other transfer is complete.

Another use of virtualization is to manage instances of applications on the servers. It can allow multiple instances to be found on one server or one instance to be accessible from

multiple servers. It can separate the resources on individual servers to be used by multiple clients. Virtualization can make networking tasks like replication, redundancy, and workload balancing. When a client accesses the cloud, all they see is the cloud. The cloud is simply an abstraction of the network underneath. As an abstraction, the client doesn't know what server they were connecting to. The system actually identifies the closest server on the network to fulfill the request. If one server is too busy with existing requests, the system takes the customer to the next server.

The bottom line is any resource that can be virtualized on the network can be pooled together.

12.3 *Providing Direction: Global Catalog*

A client runs a query within an application. That request is first authenticated by a domain controller by identifying the client and all the groups to which the client belongs. The domain controller can perform this authentication for the request because the domain controller access the global catalog where the client's registration information is located. After finishing the authentication, the global catalog is used to find the location of all the data the query is for. When found, representations of the located data are compiled into the user's interface.

This simple example shows just one way that the global catalog is used by the system. The global catalog is a data repository that is distributed throughout an active directory forest. An active directory is a structure used in Windows-based computers and servers to store information and data about the network and domains. The active directory is a hierarchical structure. The structure has three main categories: hardware resources, services used by the client, and objects within the domain and network. A replica of all the active directories is stored on the global catalog. In a network, a global catalog can be found on a global catalog server and is replicated out to every component on the network. In short, anything that can be searched can be found in the global catalog. Without the global catalog, a search would have to go through every domain requiring requests to every active directory in the network. By having all the objects in the global catalog, any searches only need to go to one place. The global catalog holds a full-writable domain directory replica and partial, read-only replicas of all the domain directory partitions in the network. As a result, an object can be found by a client without the client's knowledge of where the object resides.

The global catalog is updated constantly by the Active Directory replication system. In multi-domain environments, a global catalog can be found in each domain and those global catalogs are replicated between themselves. When the LAN is connected to other

LANs to create a WAN, the global catalogs from each LAN are partially replicated between each other. Especially over the Internet, this allows searching to be fast and effective.

12.4 *Finding a Way In: Control Interface*

Control interfaces are any set of user facing representations of the network or application, including the APIs controlling applications and services within the infrastructure. Several types of control interfaces might be available for use depending on the provider of the platform. Some of the more popular are listed here:

- Dashboards are basic summaries of key performance indicators within the network. The concept behind the dashboard is to have an easily readable representation of the current state of the network at any given time. Most of the indicators are provided thresholds for identifying normal states, troubled states and critical states. The dashboard will display the state of the network as it related to the performance indicator. In some dashboards, the actual measure may be displayed as well as the target objective. By going to the dashboard, anybody can see how the network is performing. If the network is in troubled or critical state, dashboards can show the location of the problem, who's working on it, and the total time the problem has been in existence.
- Monitoring screens are deeper representations of network, server, and application operations than the dashboard. In addition to the status of the operations based on key performance indicators, scheduled maintenance tasks are monitored and checked. Unlike the dashboard, monitoring tools tend to show only that the measures are being met or not being met.
- Subscriber portals provide access to the infrastructure and the products available. They are specifically built for subscribers to the platform. They typically include any account information, billing information, and usage to date data. The purpose is for the subscriber to understand how they are using the service from a data center perspective. The granularity of the data found in the portal is based on the standard design given by the provider.
- Power - cooling - growth planning – reliability – security – safety – redundancy - energy efficiency. These are all concerns for design any infrastructures. These also need to be monitored to ensure that the infrastructure is designed well in production. Infrastructure design tools are used to create models of the infrastructure, manage any timetables for implementing the model, and monitoring the infrastructure to find areas where better designing is required. They allow the designer to take what already is in existence and design any enhancements or extensions to the network without current production states.
- Development tools, like IDEs, provide an opportunity to develop applications,

controls, and processes within the environment. Most development tools allow the developer to work on the task at hand while providing easy access to the resources available.

- Command-and-control interfaces are typically used by data center operators and publishers to configure resources on the network. The interface controls what can be configured based on user authentication and provides direction on how the resources are to be configured, typically in terms of limits.

12.5 *Independent Software Vendors – ISV*

Companies specializing in the making and selling of software are usually referred to as Independent Software Vendors (ISV). Most of the software they create is designed for mass marketing or to enter a niche market. They typically center on specialized software products. These products usually offer higher productivity to companies than general software like spreadsheet programs or database packages.

Many ISVs develop software that runs on multiple computers and operating systems. Many companies that provide platforms encourage and support ISVs, typically with special partner programs. Speed of delivery is a primary concern to ISVs. Most businesses realize that the more applications that run on a platform, the more valuable the platform is to the customer. Therefore providers of platforms encourage ISVs to get more applications on their platforms.

ISVs may specialize. Some specialize on operating systems, others with programming languages, and still others with specific industries.

13 Classifying Data Centers

Every platform is managed by data center. And like platforms, not every data center is created equally. The Uptime Institute created a four tier classification for identifying different design topologies for data centers. The classification has become an industry standard approach which provides common benchmarking needs.

Tier One data centers will have a single path for computer power and cooling distribution but may not have raised floors, UPSs, or engine generators. It may not have redundant systems and is only required to provide a minimum 99.671% availability. The entire infrastructure has to be shut down to perform preventative maintenance and repair work.
Tier Two data centers also have a single path for power and cooling distribution. Unlike tier one, it does have redundant systems so the minimum availability is 99.741%.
Tier Three data centers have multiple paths for power and cooling distribution, but only one is active while the rest are passive. It has redundant systems that are concurrently maintainable. The minimum availability rises to 99.982%.
Tier Four data centers have everything in place with multiple active power and cooling distribution paths, redundant systems that are configured for fault tolerance. The minimum availability is 99.995%.

Data Centers can seek certification from Uptime Institute. There are no more than five Tier 4-certified sites in the world and only a couple of dozen sites certified in any classification. The process for certification starts with a look at data center design documents to identify areas for improvement. Licensed engineers examine the documents and consult with each other to determine a probable tier rating. The second step is sending consultants to the actual data center to confirm that it is built according to the specifications in the design document.

Recently, the Uptime Institute created a set of operational sustainability grades to help companies whose availability lies between tiers. There are five categories looked at for operational sustainability. Site selection can impact operations, specifically the susceptibility to natural disaster, availability of a workforce, and utility rates. Whether the facilities are up to building codes and unnecessary activity and materials are restrained from the server room. Questions are posed like if the site and building can accommodate future growth or if future but unproven technologies are used on a mission-critical site? Investment effectiveness is even measured by understanding the data center’s energy efficiencies, resale value, and its ability to align to business goals. Finally training, metrics, and collaboration examined to grade a site's operational sustainability.

Differences between Tiers

	TIER I	TIER II	TIER III	TIER IV
Number of delivery paths	Only 1	Only 1	1 active 1 passive	2 active
Redundant components	N	N+1	N+1	2(N+1) or S+S
Support space to raised floor ratio	20%	30%	80% - 90%	100%
Initial watts/feet squared	20-30	40-50	40-60	50-80
Ultimate watts/feet squared	20-30	40-50	100-150	150+
Raised floor height	12”	18”	30-36”	30-36”
Floor loading pounds/feet squared	85	100	150	150+
Utility voltage	208, 480	208, 480	12-15kV	12-15kV
Months to implement	3	3 to 6	15 to 20	15 to 20
Year first deployed	1965	1970	1985	1995
Construction $/feet squared raised floor	$450	$600	$900	$1,100+
Annual IT downtime due to site	28.8 hrs	22.0 hrs	1.6 hrs	0.4 hrs
Site availability	99.671%	99.749%	99.982%	99.995%

14 Who can deliver PaaS?

Given that cloud computing is a new trend in computer technology and involves much on the provider's part to offer an environment that is scalable, flexible, and secure enough to meet the needs of utility-based service offering, what companies can compete? Most of the providers of PaaS solutions are leading Web service companies. They have built their businesses around using innovative web approaches to maximize data center management and efficiency. Amazon.com is primary example. Their opening into cloud computing did not start as a concentrated decision to open a new market but as a by-product of transforming their own processes and computer infrastructure. Amazon's CTO Werner Vogels even concluded that “if managing a massive data center isn't a core competency of your business, maybe you should get out of this business and pass the responsibility to someone who has.”

The reason for this conclusion is clear: leading Web service providers have vastly superior economics behind them, better practices for handling dynamic workloads, expertise in dynamic capacity management, and more understanding of consumption-based cost tracking.

Leading providers of Web services are buying so much in terms of servers, storage, and other data center equipment, that they have greater leverage in negotiating hardware pricing, software licensing, and support contracts. With such buying power, they can quickly expand their infrastructure without any major threat to the financial stability of the company. They also have the money to hire leaders in the necessary fields to support the data center, whether in design, implementation, or management.

Investments continue into the creation of better business processes and workflows. The desire to automate systems pushes these companies into documenting and improving their own practices. Management and administration tools are often created to handle the improving processes. As the infrastructure grows, these tools are built to handle application and services across thousands of servers in multiple locations. The time it takes to maintain, optimize, and accommodate new services must be quick, hopefully automatic.

Most Internet companies realize that productivity is crucial. The cost of their services is directly proportional to the costs of the data center. Gaining more productivity out of fewer resources provides greater returns on the investments, whether that investment is in hardware assets, software, space, power, or facilities. These companies are attuned to the

consumption rates of applications and services in the environment. The reporting is typically internal, though a few companies are offering this level of reporting as part of the package.

15 What Does PaaS Really Offer?

As Internet services and hosting companies realize their own internal business goals, they are realizing that they can offer the same benefits to other businesses as a service offering. Platform-as-a-service is different from other service or hosting options, including SaaS, in the following ways:

- The infrastructure is standardized and the abstraction layers created through virtualization allow for fluid mobility and placement of services. The physical attributes of the infrastructure is changing form traditional solutions. In a cloud, a single server can serve to resolve computing and storage demands. Incidentally, the customer used to have a say in how the infrastructure was built; in cloud computing, the customer has little or no say in what goes into the infrastructure. Every platform provider abstracts the hardware with some server virtualization. Some use a grid engine to span solutions across virtual and physical servers.

- Most platforms utilize some infrastructure software that assists in creating, changing, and moving applications without any effort by the data center operators. In cloud computing, this infrastructure software can assist in obtaining higher availability rates as well as better disaster recovery and resilience.

- Platform offerings are priced based on consumption of COU hours, gigabytes consumed, and gigabits transferred. This sort of consumption pricing typically yield a 65 to 80% increase in savings over traditional IT platforms. The consumption based pricing also eliminates most needs for a long-term contract, allowing the customer to evaluate the service more often and finding a new solution without any penalties being incurred.

- Most platforms do not have any problem with supporting any operation system or any application. The restrictions found for most providers are in place because of skill and business drivers than actual capacity of the infrastructure to support the need.

16 PaaS Suppliers

16.1 *FaceBook*

Platform Model:	Plug-In API
Platform Type:	Add-on Development
Platform Intent:	Social Applications
Development Techniques:	Facebook Proprietary
Development Languages:	PHP, Ruby on Rails, Javascript, Python
Web Site:	www.facebook.com

Facebook is a free social platform centered on creating community on the Internet. Over 1 million users currently use the application to share information, pictures, music, and news. In addition, the developers can create software applications to be utilized by the community population. In this sense, Facebook is an Internet operating system and is often referred to as such.

Developers interested in creating an application on Facebook should have a solid understanding of PHP, Ruby on Rails, Javascript or Python. A basic understanding of the Internet, SSH, MySQL, and Unix is helpful, as well as Web Hosting Fundamentals. Facebook requires that a developer has the infrastructure to host the application and has partnered with Joynet, Amazon Web Servers, Audible Magic, and SalesForce to provide hosting offerings. In this respect, Facebook is simply a social platform for deploying an application and the infrastructure platform is provided elsewhere.

To create a Facebook application, development is done outside of the Facebook environment. When it is ready, the developer logs into the Facebook Developer Application and creates a profile. Then the application needs to be configured on hosting server you provide. The application essentially becomes a plug-in for the Facebook operating system.

There are several guiding principles set forth by Facebook to ensure that users get the most out of the Facebook experience. Applications developed for Facebook do not need to conform to all these principles in order to be accessible through the operating system.

The principles are:
SOCIAL: Helps users interact and communicate more effectively.
USEFUL: Delivers value to users by addressing real world needs, from entertainment to practical tasks.
EXPRESSIVE: Enable users to share more about who they are and the world around them.
ENGAGING: Provides a deep experience that users want to come back to regularly.
SECURE: Protects user data and honors privacy choices.
RESPECTFUL: Values user attention and honors their intentions in communications and actions.
TRANSPARENT: Explains how features will work and how they won't work.
CLEAN: Designed to be intuitive, easy to use and free of mistakes.
FAST: Achieves low Latency while scaled to handle user demand.
ROBUST: Maintains reliable uptime and minimizes error rates.

16.2 *Ning*

Platform Model:	Plug-In API, Runtime Environment
Platform Type:	Add-on Development
Platform Intent:	Social Applications
Development Techniques:	OpenSocial Proprietary
Development Languages:	XML
Web Site:	www.ning.com

Ning is an “OpenSocial” network focused on creating community from a number of users who belong to internal networks within Ning. Currently, there are over 500,000 internal social networks found. Software developers are invited to create OpenSocial applications to enhance user’s experience of the Ning.

Ning is based on XML. Developers within Ning have guidelines and support to ensure the integrity of the environment. They can develop and host their applications on the Ning platform for only the cost of their membership to Ning which is about $25 a month.

16.3 *Second Life*

Platform Model:	Plug-In API
Platform Type:	Add-on Development
Platform Intent:	Social Applications
Development Techniques:	None Identified
Development Languages:	XML, Proprietary LSL
Web Site:	www.secondlife.com

Second Life, developed in 2003, is a three-dimensional virtual world that is created by its “residents” and run by Linden Labs. The platform creates a digital continent. As the residents in this digital world develop the landscape, more experiences and opportunities become available. Residents have the ability to create, and they retain intellectual property rights for every digital creation. A marketplace is present for the buying, selling, and trading of these digital creations. Ultimately, Second Life resembles very much what one would see in the real world.

The scripting language used to create within Second Life is Linden Scripting Language (LSL). The first thing that a user can do is create a persona, called an Avatar. From there, they can interact with other avatars, build and own land and possessions, and even hold down a job. Groups and partnerships can even be created. Essentially, anything that can be done in the real world can be done in the virtual world. LSL even provides the ability to export XML data to external “real-world” websites.

16.4 *IronScale*

Platform Model:	Runtime Environment
Platform Type:	Application Delivery-Only
Platform Intent:	Business Applications
Development Techniques:	None specified
Development Languages:	None specified
Web Site:	www.ironscale.com

IronScale is a managed server hosting solution that provided automatic provisioning and configuring on-demand. Servers are commission to single clients only to ensure that security and safety of the client's applications and data are maintained. IronScale also provides the ability to clone servers, take snapshots, and provide automatic failover capabilities. The infrastructure is enterprise-class and businesses can find on-demand storage and zero-delay backups on the site. IronScale allows root control access to the server via a remote KVM console.

The key drivers of the IronScale are the ability to be fast, automated, and powerful. For the business, this means the ability to provision a server in minutes, repurposing the server in real-time or adding storage on the fly. IronScale is a fully automated environment so the advantages of IronScale can be realized from anywhere at any time.

The on-demand platform from IronScale is built on dedicated physical servers. The platform is not virtual, not shared, not cloud-based, nor grid-based. Customers can work on the platform from online tools and the servers can be configured for web-based applications, but that comes from customer decisions, not the provider. IronScale provides the software, the physical infrastructure, the user interface, and the support to deliver an automated environment.

IronScale offers 3 bundled server environments. Each of the environments is built on physical, non-virtual dual-core and quad-core X86 servers. 70GB of RAID configured storage is available for each server. Customers can choose between Red Hat Linux or Windows Server operating systems on the servers. Customers are given two networks and

8 external IP addresses per client plus 100 internal IP addresses per network. One Mbps dedicated burstable bandwidth is available with each server. The three bundles are differentiated as such:

- Server Level 1: 2 cores, 4GB RAM
- Server Level 2: 4 cores, 8GB RAM
- Server Level 3: 8 cores, 16GB RAM
-

16.5 *Apprenda's SaaSGrid*

Platform Model:	Runtime Environment
Platform Type:	Stand-Alone Development
Platform Intent:	Business Application
Development Techniques:	SaaSGrid SDK
Development Languages:	None specified
Web Site:	apprenda.com/platform

The SaaSGrid platform from Apprenda is geared towards independent software vendors to provide an infrastructure to build, deploy, and monetize service-oriented applications. Apprenda comes from the premise that the greatest hurdle of SaaS development is the creation of single-instance multi-tenant applications. This type of application has proved to be the most cost-effective SaaS application but the most difficult to develop. The SaaSGrid platform automatically converts SaaSGrid applications to single instance, multi-tenant architecture upon deployment with zero-effort from the application's perspective. It does this by isolating the application user through data partitioning, request routing, and authorization through virtualization methods. The developer can still define specific components of the application's multi-tenant behavior, such as mixing data from multiple customers in database deployments or having dedicated database assigned to each customer. Developing the application on SaaSGrid does not require any concern on the part of the developer for multi-tenancy. The developer simply builds and tests the application using the SaaSGrid SDK. When ready, the SaaSGrid will provide the necessary login capabilities, tenant provisioning, and other multi-tenancy features.

One of the concerns for Independent Software Vendors is the creation of recurring, stable

revenue streams. SaaSGrid can seamlessly integrate monetization and metering capabilities in the application by the use of an API that developers can define what transactions and code are billable. Pricing of these features can be managed externally from the application with the flexibility to create any number of pricing and feature configurations. If a business has multiple SaaS applications on the platform, the creation of merchant accounts can tie those applications together as bundles and apply a cost to the bundle. The flexibility of these features allows the developer to update the application and pricing when market demands change.

With SaaSGrid, businesses which use the developed application will subscribe and assign that subscription to the users who need access to the application. Customers create these users at the platform level and grant access and authorization though SaaSGrid. The authentication controls are placed automatically in front of the application controlling access to all parts of the application. Using roles, SaaSGrid even allows security checkpoints in the application to prevent access to certain features and functions which can be controlled by the subscriber.

SaaSGrid provides the ability to upload patches through a patching engine or upload new application images. Release management is performed by clicking a few buttons. While an application is in production, multiple versions can be running in parallel in a testing sandbox. When ready to deploy a version in testing, simply replace the older version in production without concerns of breaking the application.

With Apprenda, customers can choose to have a platform as a service or as a product. As a service, the customer can build and host their application on the SaaSGrid. As a product, the customer can host the SaaSGrid software layer on their internal network, providing the benefits of the SaaSGrid while leveraging their existing assets.

16.6 *10Gen*

Platform Model:	Runtime Environment
Platform Type:	Stand-Alone Development
Platform Intent:	Web Applications
Development Techniques:	SDK
Development Languages:	Javascript, Python, Ruby on Rails
Web Site:	www.10gen.com

10gen has developed a new component stack for building web applications. They promise automatic, on-demand application scaling. The platform currently supports Python, Ruby, and Javascript programming languages with intentions to support additional languages in the future. 10Gen also provides automatic disaster recovery and data replication.

The 10gen platform consists of five layers. The 10gen application server is a high-performance, scalable server capable of running multiple applications concurrently or in isolation. With this server, the customer has grid manageability and multi-tenancy features. Not only does the server support multiple languages, it allows mixing languages in a single application. Transparent interactions are capable with the Object Database, which allows indexing, sorting, partial object returns, and server-side script execution of Javascript objects. In addition, Core Javascript Libraries provide APIs that expose the server layer of the platform to application developers allowing greater management of the application without sacrificing the integrity of the entire solution. 10Gen has a virtual file system which is capable of storing large object. With it, any application server can store or retrieve binary objects. The Grid Management System offers capabilities for controlling the application and its resources, managing connections to the database, and other concerns relevant to application deployment and production.

16.7 *Bungee Connect*

Platform Model:	Runtime-Environment
Platform Type:	Stand-Alone Development
Platform Intent:	Business Applications
Development Techniques:	Ajax
Development Languages:	C-style languages
Web Site:	www.bungeeconnect.com

If high-powered, highly interactive web applications are the desire of web application developers, than Bungee Connect may have the platform. Whether the application is to increase productivity, provide a social service, or simply for entertainment, Bungee Connect provides an environment where the desktop-like interactivity connect with the maintenance, power, and security of centralized applications.

With the Bungee Connect offering comes a comprehensive IDE to provide a professional development environment. Developers build and test the application they develop in the same environment that end-users have when they use the application, removing any need to move the application from environment to environment. Ajax, a web development technique, is used to provide end-users with an interactive environment. To ensure users have a full range of capabilities, the platform automatically maintains connectivity to services, like SOAP and REST, and databases, specifically MySQL and PostGreSql. Any data sources are leveraged through the server, not the client to provide better security and more options. Ease is not restricted to the end-user; developers can create a collaborative environment with the creation of a DesignGroup profile to allow team development to occur. The Bungee Connect Developer Network connects developers around the world through a newsletter, blog, knowledge repository, and repository for sharing code.

Not only is the initial deployment of an application possible in the same environment, newer versions can be developed and updated with just a few clicks in the system. Applications can be standalone or a part of existing applications. The platform provides scalability and reliability to the customer and data security is a priority concern for Bungee Connect; demonstrated through their focus on securing the server, the network, and the

client. Each application instance is performed on separate servers and the application logic is executed on the server not on the client to protect intellectual assets and processes. The developer has the choice of where the data is stored within the Bungee Connect service. Data transmissions are secure using SSL encryption and standard protocols. Javascript packages are created for each application to ensure against attacks and maintain security on the client. The only data that is sent to the client is the data found on the user interface, ensuring that no business logic is executed on the client.

For most companies, business productivity is the primary driver for using developing software. Bungee Connect allows the development of standalone applications and additions to existing SaaS applications. Business features like dashboards, inventory management, calendaring, and messaging are some of the applications that can be built in Bungee Connect. Applications developed in this environment can capabilities like drag-and-drop, file trees, state management, dynamic updating. Bungee Connect applications are supportable in all major web browsers and do not require any plug-ins or separate installs to run fully for the customer.

Bungee Connect is still in beta testing, so development, testing, application deployment and hosting are currently free. After that, the pricing schedule is based on the amount of time users spend directly connected to the application, at a rate of $0.06 per user session hour. So a user who interacts with the application 3 hours a day for 20 days a month would generate a monthly charge of $3.60.

16.8 *Clickability*

Platform Model:	Access API, Plug-In API
Platform Type:	Application Delivery-Only
Platform Intent:	Web Application
Development Techniques:	
Development Languages:	
Web Site:	www.clickability.com

Clickability is a platform-as-a-service with a unique direction. It provides on-demand Web Content Management. As a platform, its not just hardware and software that is being

bought, but an evolving code base, a methodology, and scalable, secure backend, and stringent SLAs. The purpose of this platform is to provide complete lifecycle management over Web content. There are four components to the Clickability platform: infrastructure, support, software and innovation.

The infrastructure is robust, resilient, reliable, and secure. Web content is hosted on lightweight, rapidly extensible commodity hardware to handle scalability at higher thresholds and provide 99.9% uptime availability. In case of natural disaster, most customers' business continuity can be restored in fifteen minutes to four hours. The infrastructure design relies on green computing concepts.

The Clickability platform was designed to be a hosting solution for revolutionizing how companies use the Web. The provider creates, manages, and evolves the software while the customer creates, manages, and evolves their business on the web. The programming performed by the provider is meant to keep customers ahead of the market. Open APIs can allow customers to the best applications from other vendors.

Support from Clickability is based on nine years of supporting Web content publishers. During implementation, the platform is configured exactly the way the customer wishes, making sure that code is migrated, templates are coded, and, once web sites are launched successfully, customer support starts to iron out issues and carry customer requests for innovation to the Clickability development team.

Clickability looks at innovation from two perspectives – the evolution of the platform and the applications customers create on top of it. The platform evolves quickly as customer feedback and web technologies drive new innovations. Several monolithic releases of the platform are conducted each year with new features in each update. Up to 400 new features may become available in any 18 month period, and they are integrated seamlessly into the platform without any disruption to service. Customers can decide to utilize new features as if tripping a switch. Clickability allows customers to create unique applications needed for their business goals.

16.9 *IS Tools*

Platform Model:	Runtime Environment
Platform Type:	Stand-Alone Development
Platform Intent:	Business Applications
Development Techniques:	None specified
Development Languages:	Java
Web Site:	www.is-tools.com

IS Tools is a Java-based platform-as-a-service engine and software tool. It is a browser-based online integrated environment for building, deploying, and running complex and flexible business applications. The IS tools solution is configurable, flexible and dynamic, and scalabile. Business process applications are configured in the platform and significantly reduce the time to production. Configured applications can be tested, amended, and changed in production without disrupting service. Deployment of an application can be done in the centralized or a distributed environment. It can share the IS Tools platform with other business applications to reduce cost to any single instance.

A web browser database application that can be configured by the user is provided and does not require any programming knowledge. Content and appearance of the application is decided by the user and can be changed while retaining traceability and stored data. IS Tools handles any level of access rights to ensure information integrity. Reports can be compiled and read online or exported to MS Excel.
Data is entered into a single structured database to streamline management of processes and improving data quality. All menus, forms, texts and buttons can be set up for any language without having multiple instances of the application. IS Tools has a framework that is ready for set up of structures, rules, relationships and logical aspects for the application.

16.10 *LongJump*

Platform Model:	Access API
Platform Type:	Add-On Development
Platform Intent:	Web Applications
Development Techniques:	None specified
Development Languages:	None specified
Web Site:	www.longjump.com

Developers have an enterprise-ready stack and pre-built application framework with LongJump. Their platform-as-a-service allows applications to be built directly on the web that are easy to create and maintain. The platform provides extensive features for presentation processes and data processing including:

- Widget-Based Portals – Whenever a user opens their instance of the application, they can bring forward specific data from the application into a customizable Home page. Home pages can be designed for specific users, teams, or job role.
- View & Report Designer – LongJump has a reporting engine with a full complement of options to provide comprehensive analysis of data generated by the application. The data can be filtered, ordered, color-coded, grouped, and charted into management reports. Custom computed fields are available to generate personalized reporting on the fly.
- Print Template Processing – Users can design HTML pages embedded with variables to create templates for printing.
- Form Layout Designer – the designer can create forms using custom layouts. With field level permissions, the fields can be defined and controlled so that they are visible only to designated teams, roles, or users.
- Cross-Object Joins – The ability to join three distinct sets of data is available through LongJump. Custom Report Categories (CRC) can define relationships between two or three objects by leveraging specifically linked fields, producing dramatic reports without complex business intelligence tools.
- Data Publishing – Calendars, views, and reports are all ready to be published to the web using a single-line embedded script.

- Data Policy Engine – business rules can be triggered based on data criteria through LongJump's data policy engine. This enables process automation through data specifications.
- Notification Engine – This engine provides an email notification whenever any change in ownership occurs or records advance through the workflow.
- Validation Engine – The LongJump engine provides data validation at data entry to ensure that data input is correctly entered.
- Workflow Designer - Data records can be moved through the organization using the workflow designer. With this feature, complex tasks like project management, fulfillment transactions, and authorization and approval requests can be completed with ease and accountability.
- Calendar & Task Manager – Business teams can coordinate their activities and associate them to specific records, creating historical context for each record.
- Email Integration – Records can be emailed automatically. Emails can be associated with a record. The only thing required is an email field in the record.
- Document Control - A centralized repository is included for teams and a record-based document storage for all files associated to a record.

LongJump also provides a number of administrative functions:

- Access Controls – users are organized in a hierarchy using permissions in a team or role-based fashion. Every object has view, edit, add, and delete permissions which allow data sharing in any combination dynamically on any page of the application.
- Sharing Policies – Teams can view or update information for another team. This is done by creating policies for data sharing.
- Activity Audit Trails – Audit logs are created for user activity and can be reviewed for management and troubleshooting purposes.
- Mass Operations – Transfer of record ownership, information updates across multiple records, and mass deletion from a specified field are all capable within LongJump through large scale search and replace functions.

16.11 *Wolf Framework*

Platform Model:	Runtime Environment
Platform Type:	Stand-Alone Environments
Platform Intent:	Business Applications
Development Techniques:	None specified
Development Languages:	XML, C#
Web Site:	www.wolfframeworks.com

Everything about Wolf Framework is web-based; the business application designer, end-user experience and Web services. Wolf provides complete entity (data) management with the capability for easy designing, auto-generation of editable forms and control at the entity level. File and image management is included in the feature set, as well as input validation. All of this is available through an easy to use interface. Entity details and relationships are automatically indexed after initial configuration.

A business rules designer can assist the customer in creating, editing, and cross referencing business rules for application. Those rules can assist in building complex queries and extracting data from analytics. They can also be used to manage business processes that work in parallel or nested.

User interfaces can quickly access stored data. Within the interface, the user can find search templates, custom toolbars, and editing features. The user can also explore the data through type trees. A reporting and analytics designer provides a backend view of the data by creating summary reports and embedding those reports into a dashboard.

16.12 *GraphLogic Pointdragon*

Platform Model:	Runtime Environment
Platform Type:	Stand-Alone Development
Platform Intent:	Web Application
Development Techniques:	AJAX
Development Languages:	
Web Site:	www.pointdragon.com

Pointdragon is an IDE that runs on standard web browsers using object-oriented visual programming. It is also a runtime environment where pointdragon applications can run. The platform is suitable for developing single user applications to complex multi-user applications. The environment is highly scalable and 100% visual. Every element of the application can be created using AJAX enabled interfaces, process and computational logic, and database definition, update and querying.

Developers do not need to be experienced to create web-based applications. However, if the developer is experienced, the ease of creating rich, complex systems is easier with pointdragon. Everything needed to create an application is accessible from the development space allotted by pointdragon. Applications can be built from scratch of by modifying applications from the pointdragon library. Online training videos are available for learning how to develop applications, as well as support through Instant Messaging or contacting the developer community within pointdragon. Collaboration with other developers is possible to get assistance to refine developing applications. The application can be continually modified and changes can be made on the fly without disruption.

Users access the application by signing on to the pointdragon website. There are no hardware or software requirements for hosting the application. Security IDs are easily set up with the administration wizard. Pointdragon applications can be made available to others by posting them in an application library. It becomes freeware for other application developers. By uploading the application to the pointdragon storefront, developers can resell their applications. Users can access the application from the pointdragon website or the developer can host the application on the company website.

16.13 ***Amazon EC2***

Platform Model:	Runtime Environment
Platform Type:	Application Delivery-Only
Platform Intent:	Raw Compute Applications
Development Techniques:	None Identified
Development Languages:	No limitations stated
Web Site:	aws.amazon.com

Amazon Elastic Compute Cloud (Amazon EC2) is a service that provides resizable capacity on the Internet. Customers have full control over computing resources. What Amazon provides is an approach for businesses and individuals alike to obtain and use the capabilities of a new server in minutes and with low overhead. The EC2 solution provides tools to developers to build applications.

This is a virtual environment that allows web service interfaces to launch instances of the applications. Because it operates over the web, the only constraints to supporting in multiple operating systems are the limitations of the application. The developer is manages network access and can determine how many systems are used. To use Amazon EC2, create an Amazon Machine Image (AMI) that contains the applications, libraries, data, and configuration setting required to run the applications. Predesigned images can be used to make the process even faster. Once the AMI is created, it needs to be loaded into Amazon Simple Storage Service (Amazon S3). Amazon S3 provides storage for any application images and data that is required to run the application in the EC2 environment. Use Amazon EC2 web services to configure access and security controls. Then use the same services to choose the instance types and operating system desired for the application. Using web service APIs and management tools provided, the developer can start, terminate and monitor as many instances of the AMI that are required. The developer can even choose to run the application in multiple locations, use static IP endpoints or attach block storage to each instance.

Amazon EC2 provides the developer to increase or decrease capacity in minutes, or to choose one or multiple server instances simultaneously. Using web services APIs, the

application automatically scales up or down accordingly. The developer has total control over each instance of the application as well as access to console output for every instance. As the developer, the types of instances can be chosen, as well as operating systems and software packages. Amazon EC2 allows the memory configuration, CPU and storage requirements to be selected by the developer as well. Instances of the application can be quickly and predictably replaced when disruptions occur to ensure that the application is always available just the way the developer wants it. Several features in the Amazon infrastructure make sure that applications are failure resilient. Persistent storage in the form of Amazon Elastic Book Store (EBS) keeps additional storage ready for application instances in the case high volumes of users access the application. Deciding to have the application instances located in multiple geographies brings an inherent insulation from isolated failures in the system. The availability commitment for Amazon EC2 is 99.95%. Elastic IP addresses allow developers to program re-mappings to any instance in the account for the purposes of masking the instance or isolating failures. Web service interfaces allow the developer to configure firewall settings to control network access. The platform can work with other Amazon Web Services including Amazon S3, Amazon SimpleDB, and Amazon Simple Queue Service (Amazon SQS) to provide additional functionality for computing, query processing, and storage.

Developers can choose the type of instance for their applications. Those types fall into two categories: standard and High-CPU. Within the standard category, the developer has small, large and extra large instances which consist of increasing allotments of memory, EC2 compute units, and storage for instance, and a 32-bit or 64-bit platform. For compute-intensive applications, the High-CPU category has a medium and extra large instance available.

Amazon Machine Images are already preconfigured to support a variety of operating systems, software for databases, batch processing, and web hosting, and application development environments and video encoding & streaming.

Pricing is based on the number and type of instances, the volume of data transfers, volume of EBS and number of Elastic IP Addresses.

16.14 *Salesforce.com*

Platform Model:	Runtime Environment
Platform Type:	Stand Alone Development
Platform Intent:	Business Application
Development Techniques:	None Identified
Development Languages:	No limitations stated
Web Site:	www.salesforce.com

Salesforce is attempting to provide a full-functioning Level 3 platform-as-a-service on-demand solution. Their goal is to become leaders within the marketplace and therefore creating their own version of the platform. The driver to this solution is a multitenant architecture. Within this architecture, a boundary is created between the platform and the applications on the platform. In addition, the logic of the applications can be independent from the data controlled by that application. The focus is the creation of business applications.

Using Salesforce, developers already have an on-demand framework for creating applications. They can utilize the existing user interface, security model, reporting functionality, and integration capabilities, as well as other functions within the framework. Since most business applications today require integration with other systems through workflows, the platform allows easy integration with existing Salesforce applications or HR, ERP, and IT systems.

Developers have access to Force.com Connect which provides a several options for integration their application using Web Services API. Salesforce.com has created native ERP connectors to provide integration to Oracle and SAP Systems. Middleware is available from nearly 25 of Salesforce partners or developer toolkits to create customer integrations. The Salesforce AppExchange also gives the developer access to over 500 components and applications to integrate with the applications. Additional connectors are available to allow the business applications to work with popular desktop applications, like Microsoft Outlook, Lotus Notes, Microsoft Excel, and Microsoft Word.

Popular programming languages can be used within this platform or use the native Apex programming language. The tool allows a series of point and click wizards to create functionality to the developing application. Mobile applications can even be built. Salesforce offers additional application administration services for when the application is in production.

16.15 *3tera AppLogic*

Platform Model:	Runtime Environment
Platform Type:	Stand-alone Development
Platform Intent:	Web Application
Development Techniques:	AppLogic
Development Languages:	None specified
Web Site:	www.3tera.com

3tera's AppLogic is a turn-key commercial cloud platform. It was designed from the ground up to provide a complete solution for delivering applications online. It consists of a grid operating system, integral IP SAN, intuitive interfacing, and a catalog of common infrastructure components and software stacks. There is no need for code modifications to existing applications to run within AppLogic. The infrastructure is compatible with the most popular data center operating systems, including Sun's Open Solaris and Solaris 10, Microsoft Windows, and Linux. To be flexible for cloud computing customers, AppLogic replaces the physical hardware and software deployments with visual integration. The same code, middleware and operating systems are used, but the infrastructure is visually assembled through a browser. A single command can setup n-tier applications or migrations to another data center.

The infrastructure provides a number of reliable services. High availability is obtained through mirrored storage on the SAN. If hardware failures occur, the affected components are restarted automatically on different resources to preserve application continuity. Resources such as CPU, memory, and bandwidth are determined at runtime. Applications can be allotted to up to 128 servers. Monitoring of resources and systems are available through the creation of custom dashboards.

With AppLogic, applications can be developed and tested. On-demand scaling is available during development and deployment of an application. And business continuity is a primary concern for AppLogic in its delivered features.

16.16 *Coghead*

Platform Model:	Runtime Environment
Platform Type:	Application Development Only
Platform Intent:	Business Application
Development Techniques:	Adobe Flex
Development Languages:	
Web Site:	www.coghead.com

The Coghead platform is a comprehensive web-based application environment. Web applications can be built from scratch or using a template from the Coghead application gallery. Those web applications can be administrated from anywhere around the world.

Adobe Flex allows developers to create applications using a visual authoring environment. Customer forms can be built, as can tab-based user interfaces, filters, views, and more. To assist with making the application fit into the business, a business process engine is included that allow the application to be built using business logic and workflow. When simply wanting to send an email for verification to multi-step transactions, the engine can easy create workflow value simply be configuring in a visual drag-and-drop environment. Data integration is possible using the REST Web Services API. This can allow data sharing between applications on the Coghead system or on the Internet or in your business. Desktop connectors for MS Word and MS Excel are available. Coghead applications can even be embedded into any web page.

Coghead works with Amazon EC2 for hosting solutions.

In February 2009, SAP acquired Coghead's assets minus its customers following their demise.

17 Introduction to Storage Management

One of the biggest computing trends in the business community is the network concept of cloud computing. When technical and management personnel use the term “cloud,” they are speaking about an Internet-based network solution. Its design is to provide services on demand to the end-user without requiring them to have the technical expertise to support those services. This is a very powerful asset for businesses who do not want, require, or have the financial means to have their own IT infrastructure.

The architecture of a cloud computing environment is rather simple overall, though its individual components may be highly complex. It consists of three distinct parts. The IT infrastructure is the data center, where client information is processed and stored. The opposite side of the cloud architecture is the client environment. Between the two is the cloud: a set of controls to protect, manage, and distribute access from the client environment into the IT infrastructure. How these three parts are built is based on the policies, procedures, and hardware used by the administrating parties.

No matter how the cloud computing environment looks, the concept is to provide IT capabilities “as a service.” Those services can include web-accessible applications, file management, and data storage. Of all these services, the largest and most popular is storage management. For most businesses, storage is the most important and the most expensive IT resource within their infrastructure. Unfortunately, the technical expertise assigned to storage management is not consistent with the need.

Storage Management is the ability to store and manage files and data on the network. The software used to ensure this ability is called Storage Resource Management (SRM). The primary concerns for storage management are capacity, use, policies and event management. In cloud computing storage, the goal is to access storage capabilities thought the Internet.

17.1 *Cloud Computing and Storage*

Working through a cloud brings its own benefits and risks to storage management.

By outsourcing the capabilities of storage management, the company reduces capital expenditures for hardware and services related to maintaining a large infrastructure. Though this may not be a major concern for large global corporations, the architectural

concept is still relevant to the company's infrastructure. In these cases, the company may create a “private” cloud that is available to the company's locations and/or departments. This ability allows the company to charge the financial sectors based on storage use more easily than before.

The ability to share resources, and costs, to the storage infrastructure is called multi-tenancy. As a result, the infrastructure can be centralized to large database instead of distributed around. Improvements to capacity and performance are realized because the skill for storage management is equally centralized. Some concerns are present for performance of the system through peak times.

Cloud storage does not require specific locations or devices to be utilized in the infrastructure. Since all access is through the Internet, the client and the infrastructure can be thousands of miles apart.

Most companies who provide storage management on the cloud also provide multiple levels of redundancies to the client, providing reliable access to the data on the network. In many cases, this provides an acceptable solution to business continuity and disaster recovery planning.

Scalability of the system is another benefit to working from a cloud, allowing companies to increase or decrease their need on demand.

Cloud computing allows for centralized security to access though there are concerns since control of security is given to another party.

An Internet-accessible centralized solution for storage management provides businesses and individuals the opportunity to meet their needs without worrying about higher overhead costs for equipment, personnel, and effort to maintain the solution. The greatest concern is the transfer of responsibility and control of the storage solution to another party will create additional risks if the storage provider cannot fulfill their commitment. For regulated businesses with sensitive information, these concerns may force a company not to choose this option.

17.2 *History of Storage Management*

Storage Management has been a concern since the dawn of writing. Before this, information and stories were passed on through civilization's oral tradition. However, the information was distorted through interpretation, theatrics, and memory loss. Writing became a method of ensuring the integrity of the original story. However, it was very expensive, since people were required to handwrite every word. And then people needed

to learn how to read. It became a sign of education when a young man or woman learned how to read and write.

As volumes of writings were accumulated throughout the ages, libraries were built to house these writings. The printing press invented by Gutenberg in the 1430s allowed books to be made cheaply and they became accessible to everyone. More and more writers started to get published and even more space was needed for the output. Eventually, a system had to be created to organize these writings. In the late 1800s, Melvil Dewey created his system of classifying information that soon found its way into the U.S. Library System.

When Thomas Edison found a way to store sounds on a hard disk, a new age in storage management began. Especially through the world war era, information was being stored on records, films, and photographs. These media storage units needed to be stored and managed just as books had been for hundreds of years before. Eventually, information became digital allowing additional methods for creating and storing data.

From the business perspective, the concepts of mass production and quality management drove the need for data storage to a new level. Before this, writings and information sharing were mostly for the purpose of communicating. Now, data was being collected for the purpose of analysis and prediction. This analysis was used to build the credibility of a business and improve business practices.

In each of these stages, the demand and need for storing information increased. As new technology was introduced, new ways of storing information was realized, or in some cases, forced to develop in order to keep up. The age of personal computing and the Internet has increased the demand even further as avenues for communication, file sharing, and online business (ebusiness) have exploded the amount of information being processed every minute.

17.3 *History of Cloud Computing*

The term “cloud” is a metaphorical reference to the Internet, though the concept predates the Internet. In 1960, John McCarthy stated that “computation may someday be organized as a public utility.” Many government bureaus back in the 60s started to create information systems that are very similar to present cloud architecture in concept. The term “cloud” became popular in the commercial sense in the early 1990s to describe the infrastructure solution supporting ATM networks. Additional “cloud” solutions started to appear but were restricted to “software as a service” solutions.

After the dot-com bubble, Amazon.com started to modernize their data centers and found significant improvements in internal efficiencies. The solution they used was cloud architecture. They provided access to their systems in 2002 on a utility computing basis. In 2007, a number of universities, with the assistance of Google and IBM, started a large scale research project around cloud computing. At the same time, the term started to find its way in the mainstream news. Questions were being asked from every corner of the computing arena. By the mid-2008, events were scheduled to discuss cloud computing.

The Gartner group, which analyzes major trends in the business world, acknowledges that companies are switching to service-based models. They state, “The projected shift to cloud computing, for example, will result in dramatic growth in IT products in some areas and in significant reductions in other areas.” This shift to cloud computing is being forced by globalization and economical decline, resulting in a need to lower business costs. The Gartner group supports this with, “Businesses are investing in improvements to internal processes aimed at reducing costs, while often maintaining some of the prior interest in innovation. The second factor is that globalization allows IT services providers to mitigate the risk of weakening demand by operating in more markets.”

17.4 *An End-User Perspective on Cloud Storage*

The primary goal of the end-user is to get onto their computer and start working. They want to access their applications as quickly as possible. Storage becomes a secondary concern and often only when they have noticed that they are running out of space or performance is suffering. From day-to-day, the greatest concern related to storage is where a file is or should be saved.

For this reason, most users traditionally save their information on their hard drives. Because this information is on a device that only they have access to, they consider the information to be secure. Early in personal computing, sharing files meant transferring the file to a floppy disk or compact disc. Media devices became convenient ways to organize files. Eventually, connections to the network and email opportunities allowed file sharing to happen quickly and securely. Businesses and educational agencies asked their employees to store information on the network, including files. The purpose of this request was to share information with others in the community.

As the Internet grew, other opportunities and reasons for storing information on a network became apparent. The demand on the user, especially one conducting business, to utilize the same application, follow the same processes, use the same resources is increasing every day. Most users do not have the technical expertise to build and manage a storage infrastructure. Their only concern is reliable access to their files and the ability to store new files at any time, in any location. In some cases, they may have concerns about protecting

their files from unauthorized access. At other times, they simply need an alternative means to store information in other locations than their hard drive.

Storage as a service provides an inexpensive means to the user to store their information online. Doing so requires no technical knowledge, no special software, and no special hardware from the user. Most user email accounts, such as Yahoo! and Gmail, allow storage of email messages on the network. Web-hosting solutions, such as GoDaddy, provide online storage of websites, blogs, and wikis. All of these solutions are inexpensive and do not have unreasonable restrictions. The benefit is that as more storage is required, the user can purchase more, usually at a monthly cost.

17.5 *Small-Medium Business Perspective on Cloud Storage*

The concept of cloud computing, and storage in particular, was driven as a solution marketed to small and medium businesses (SMB). The demands on a SMB are rather large and complex. In many cases, they are trying to compete against a company that is larger, more financially sound, and has a greater market share. As a SMB, a business owner may not have the capital required to implement or sustain a viable IT solution.

“As a service” solutions allow a SMB to have a viable IT solution without the investment or overhead costs of owning the infrastructure. There is no requirement to hire IT specialists to manage the solution and there is no need to purchase additional software or hardware. Simply put, the technical and personnel requirements for maintaining an IT infrastructure and storage capabilities are transferred to a vendor who charges for their services. If the SMB requires more storage, they simply pay for more. In any case, the cost of storage becomes far less when a SMB outsources the operations to a second party.

The expense of “storage as a service” is less for an SMB because the company is sharing resources with other SMBs. Simply put, several businesses requiring storage space may choose to pay thousands of dollars to purchase ten gigabytes of storage space each, but only use a fraction of that. Or a single ten gigabyte solution may be purchased and several businesses can utilize only what they need. This is the concept of cloud computing at work.

Though the SMB shares the cost of the infrastructures with other SMBs, they need to ensure that their business requirements are met. The greatest concern in sharing resources is the security factor. A company does not want to inadvertently share information with another company on the same physical or logical resource. Additionally, they want the highest level of performance and availability possible. As with every business relationship, matching what is being received against what is being paid for is the key to success.

17.6 *Large Company Perspective on Cloud Storage*

Large companies may not have a need for commercially provided “storage as a service” simply because their size can offset any additional expenditure in an IT infrastructure. Many of these companies may own the infrastructure but outsource their IT management. This does not mean the concept of cloud computing should be dismissed. In fact, these corporations are demanding the adoption of cloud computing concepts. The reason for this is a simple paradigm of employees as customers; that is, in order to ensure that business operations are consistently delivering, the people working must be given the tools and resources to do the work. The corporation is responsible to ensure they have those tools and resources and in this sense, the employee is now a customer of the corporation.

So the benefits of cloud computing to individual end-users and small-medium businesses apply to the operations of a large company, in this case in terms of employees, functions, locations, and departments. An additional benefit for the large company is the greater ability to manage the use of the environment by the company. Traditionally, the infrastructure was a separate line item in any financial budget: the company paid for it and the employees used it. There was no accountability of its use from a financial perspective. It didn't matter if there was a department of ten people utilizing 80 percent of the resources or a department of 100 using the remaining 20 percent.

With a cloud computing solution, employees register and “rent” their services from the network. By matching the employees to their department or function, the company can obtain a realistic picture of how the infrastructure is being used and bill the services to the appropriate departments or functions. In this way, companies can also identify areas of the business that are achieving an appropriate return on investment (ROI) based on their function.

17.7 *Exploring the Anatomy of Cloud Computing*

“As a service” solutions can be layered solutions; that is, some solutions can be a combination of services. For instance, in order for software as a service (SaaS) to work, it needs an operating platform and physical hardware. If a company or a person has neither the operating platform nor the infrastructure, those services can be provided as well. For this reason is it important to understand the scalability and flexibility of cloud computing.

Software as a Service is one of the most popular services provided commercially and one of the simplest to be provided. It is the deployment of software from a centralized system. In the early stages of cloud computing, it allowed users to go to a server and install directly

from the server. Later, that progressed to downloading the installation code and installing on the computer. Now, some SaaS solutions can allow a user to run the application from the server with very little actually installed on the local computer. SaaS is a metered service, allowing the user to lease the application and pay only for what they use.

Under the application layer is the operating platform. On this layer, the provider opens access to the operating system and the required services for a specific application. In most situations, a platform as a service (PaaS) solution is either attached to the application layer or the infrastructure layer below it.

The Infrastructure as a Service (IaaS) solution provides guaranteed processing power and reserved bandwidth. Simply put, the company can lease as little as a single computer to a data center. Storage capabilities have traditionally been part of the IaaS but because of the demand have become a distinct service available to the user. Storage as a Service (dSaaS) are billed to used based on the amount of storage used (capacity) and requirements for bandwidth (utilization).

18 Storage Management

18.1 *What Does the Internet Need to Store?*

On one hand the Internet doesn't need to store anything because the Internet does not store, it communicates. The Internet is a system of interconnected computer networks on a global scale that allows users to communicate with each other. On the flip side, the information generated through this communication may need to be stored for future reference.

First of all, the visual representation of the Internet is through hypertext documents called web pages. These web pages allow a user to move through the network through a series of links. Each of these web pages ideally contains information that is deemed useful to a public audience by its creator.

The popularity of the Internet and use of hyper texting allowed file transfer and sharing to become more effective and efficient. Electronic mail and online chatting has made the Internet a daily resource for individuals and businesses. Advances in technologies because of the Internet have generated such add-ons like Java scripts and applets. Internet marketing and business has enabled online shopping and even online auctions. The most popular of these sites are eBay and Amazon.com. More importantly, retailers utilize the Internet to show their stock offerings and take orders. The same concept is used by specialized groups that manage memberships. Online orders, membership profiles, transaction receipts, even online banking are all examples of the information that can be found on the Internet.

Open sourcing is a popular trend in computing. Wikipedia is an online encyclopedia that is generated by individual contributors. Open Source coding has allowed the development of software using the contributions of multiple developers of different skill levels. One of these open source applications is the OpenOffice suite managed by Microsoft. Essentially, you have the major functionality of MSOffice but it is available in an open source format.

The bottom line is that an enormous amount of information is being generated on and transferred through the Internet. The information is stored on a physical connection to the Internet. The security, capacity and performance requirements for this information are dependent on the source generating the information.

18.2 *What is Storage Management?*

With data in a typically company increasing 50 to 100 percent annually, storage management is the business process and system used to find a place to put this data as well as maintain the data that already exists. That data may be in the form of data tables, raw data, or information. Storage management is not information management, where information is categorized, goes through an approval process, and reviewed on an annual basis. Storage management is not simply adding more physical disk memory to the network. To have a successful storage management solution, a number of disciplines must be involved:

- Asset Management provides the process and procedures for managing the physical infrastructure of the storage solution.
- Capacity Management manages the storage capacity that is used and available for use.
- Configuration Management provides a disciplined approach to managing the configurations of the components in the storage area network.
- Data and media migration provides procedures for transferring data between different servers and media devices.
- Event management describes the steps required when a disruption in service happens.
- Performance management starts with a set of predetermined goals and manages the efforts of the system to ensure that it performs consistent to those goals. Capacity and availability will always be two goals within the storage environment.
- Availability management handles the up/down relationship of the network ensuring that access to the stored data is possible at all times.
- Security management, or policy management, allows the company the approach to providing a secure environment.
- Backup and Restore are combined disciplines to ensure the company's ability to recover from an event.

Looking at these disciplines, it becomes increasingly clear that the process of storage management is mostly a strategic one, used to develop and maintain a viable storage management solution. It incorporates these disciplines to ensure that the solution adheres to business demands and issues.

From a business perspective, storage management provides a method to ensure the proper storage capabilities are in place, allowing for greater scalability, higher reliability and higher fault tolerance. When considering a storage management solution, there are four considerations: online storage, backup, archiving, and disaster recovery.

18.3 *Exploring Asset Management*

The physical infrastructure is the foundation of any storage management solution. Disk drives, media drives, magnetic tapes and the like are all physical assets pertinent to the solution. With these physical assets in place, a system is required to manage them. Asset management is the process to handle this need.

Two perspectives are important when dealing with physical assets: specification and financial. Specification speaks to the ability to identify the asset in the physical environment. Information allowing this to happen is the type and model of the component, the serial number and distinguishing characteristics, and location of the asset in the environment. Additional information may be obtained including process speed, storage capacity, and bandwidth, yet most of this information will be shared with another process called configuration management.

The financial perspective handles the cost of the asset. Tracking this information is important for understanding capital expenditures and handling financial concerns such as taxes. In a cloud computing environment, understanding the cost provides an avenue to ensure that services provided are cost-effective, i.e. cost per gigabyte. Degradation ultimately becomes a financial issue since a plan must be in place to upgrade or replace the asset at some future time.

Given these two perspectives, asset management as an overall solution starts the moment a device is ordered. Though this stage is typically driven by procurement processes, the importance is on tracking the asset through the ordering process, receiving, configuration and installation activities. Once the asset is installed, the active work of asset management is minimized until the asset is moved or removed from the environment. Annual audits of the inventory may be conducted to ensure there have been no unauthorized changes to the environment.

18.4 *Exploring Capacity Management*

The primary concern of capacity management is to meet the current and future business requirements cost-effectively. The challenge of capacity management is to keep up with the ever changing requirements of the environment: as business and technology changes, the requirements to processing power, memory and storage also changes. The approach allows the tracking of trends in capacity over time to predict needs in the future.

Capacity management is concerned with optimizing performance and efficiency. To that

end, the activities of capacity management consist of monitoring the performance and throughput on the server, or cluster of servers. Performance analysis looks for the impact of daily activity on capacity as well as new releases. For instance, the release of the iPhone into the market by AT&T created a performance breakdown in the activation process because the system did not have the capacity to manage the release. Performance tuning is done to provide efficiencies to the current infrastructure.

A large component of capacity management is plan development, a process to determine the changing demand on capacity based on business trends and goals. Whether planning for daily operations or a single event like a product release, capacity planning ensures that there is no discrepancy between the demand and the delivery of capacity to the company. There are three strategic paths to capacity planning: lead, lag, and match. The difference between these approaches is based on the goals around capacity and demand. A lead strategy aggressively adds capacity to the environment in anticipation of demand with the intent to grow the number of customers. A lag strategy only adds capacity when the environment is being fully used or increased demand forces addition to the environment. The match strategy adds capacity in small amounts to keep up with demand.

18.5 *Exploring Configuration Management*

Configuration management is a broad discipline intended on establishing and maintaining consistency for an assets performance. Where asset management handles the physical attributes of the asset, configuration management focuses on the logical and functional attributes throughout the life of the asset. Configuration management provides a method of controlling changes made to the hardware, software, firmware, and documentation of an asset. It records the configuration information about the asset including the IP address, MAC address, loaded software and controls.

Essentially, the intent of configuration management is to create and maintain a record of all the components that make up the infrastructure, including the relationships of those components across the infrastructure. The information gathered is stored in a configuration management database (CMDB). The use of the CMDB requires that all components are registered. The CMDB allows for components to be identified, controlled, status established, and verified. The scope of configuration management may vary from one company to the next, with a robust system providing all the information about the component, the products and versions loaded on the component, and the relationships that component has to other components.

18.6 *Exploring Data Migration*

The computing environment is constantly changing with new technology, additional software, and expanding user base. One problem with this ever changing environment is the adoption of new data systems, formats, and types, forcing transition from old to new. Data migration provides methods for this transition. The goal of data migration is to automate the process.

To design an effective data migration procedure, it must include a minimum of two phases: data extraction and data loading. Data extraction removes data from an old system. The migration procedure maps the old data formats to the new data formats. Once the data transfer is complete, the data is loaded into the new system.

Because some translation is required to move the old data into a new format, an additional step to verify accuracy may be included in the procedure. This step can also ensure the translation is complete and meets all the requirements of the new system. The verification steps typically require both the old and new systems to run concurrently to identify differences so there is no loss in data.

Another optional phase in data migration is data cleaning; an opportunity to identify redundant or obsolete information and improve data quality. All of these phases together may be run several times depending on the complexity of the system.

Essentially, the goal of data migration is to preserve information.

18.7 *Exploring Event Management*

In every IT solution, a need exists to prepare for unpredictable situations that will cause a disruption to service. The concerns in these situations require an appropriate handling of the disruption, communication, and resolution. Event management is a process designed to mitigate these concerns. Most events restricted to the storage solution typically deal with the automated and/or scheduled maintenance tasks, such as daily backups and updates. Additional events for a storage solution are the occasional upsurge of activity due to new releases, and data migrations.

Since the storage solution is dependent on a many various services including the network, disruptions in those services will impact accessibility and performance of storage. For this reason, events require appropriate handling. First of all, event identification can come from

two sources: monitoring and on site occurrences. Most events will be identified through monitoring and the root cause will be more exact. On site occurrences can come from users, technicians on the floor, and managers handling occasional projects impacting the solution. These on site occurrences identify problems based on the person's perspective instead of actual root cause. For instance, a user who cannot download their file may assume the problem is with the storage solution itself, instead of the network where the problem truly exists. Because of this perceptiveness, the handling of the event requires a quick discovery of the true problem and the implementation of a resolution to retain confidence, integrity, and availability. If the event is severe, root cause analysis may be required.

Most IT organizations focus on the resolution to the event without providing the appropriate communication. The tendency is to disclose only what is absolutely necessary and hide the rest. A growing trend among users is the desire to be fully informed of all problems visible to them. This means disclosing the problem, the root cause, and the expected time of resolution. Full disclosure has a benefit to the IT organization – it can reduce the number of calls to the help desk because of the event.

18.8 *Exploring Performance Management*

Every company has performance goals for their business. From a business perspective, performance is defined as the ability to match desired results with actual results. These goals can be rather broad, including sales performance, production performance, delivery performance, and service performance. IT support needs to translate these goals into technical goals for the IT organization. Most of the time, these technical goals will be considered Service Level Agreements (SLA). For storage management, the two most prominent SLAs are capacity and availability.

- Capacity is typically calculated as a product of the number of machines, shifts, utilization and efficiency. There are two numbers that should be calculated using this formula: the actual and the desired capacity. In most cases, the changes will be based on utilization and efficiency goals.

- Availability is a simpler calculation of the time the system is available versus the time the system is not. Most service providers will not promise 100% availability because they realize that events will occur and maintenance is required at times, taking portions of the system down.

The task of performance management is to constantly monitor the IT operations against the goals and provide communication to management when performance gaps are

occurring. The results of these communications are fed back into capacity management and availability management processes to determine if additional infrastructure is required or other options for increasing performance.

In a storage-as-a-service solution, SLAs are a critical part of the contract between the provider and the buyer, as they are the essence of what the buyer is paying for. As such, the provider's inability to deliver on the SLAs will most times result in paybacks to the service fees, sometimes resulting in a bigger loss to the provider than any investment to meet the SLA.

18.9 *Exploring Availability Management*

The goal of availability management is to deliver consistent access to data. Availability is distinct from reliability which speaks to the intrinsic failure rates of the system hardware. High reliability hardware will impact availability but is only a variable in the entire availability solution. Availability management will plan a solution around the reliability of the systems found in the environment. In fact, availability concentrates on two situations: planned and unplanned.

Planned availability focuses on continuous operations. Workload balancing allows the service provider to route service activities from one server to another. This is fortunate when the load on one server exceeds the limit that the server can handle. During regular maintenance, workload shifting can be done on a temporary basis to take down individual systems. This provides a transparency to the users so that it appears that the network is accessible at all times.

Unplanned availability, or high availability solutions, focus on three elements: fault tolerance, data integrity, and disaster recovery. Fault tolerance is the creation of redundancies in hardware and software within the environment. With these redundancies, users are switched from one system to the next in a disruption. Another method to obtain fault tolerance is to create a clustered environment: several systems coupled together with the intention of improving system availability. When a system fails, the applications and processes are automatically switched to a designated backup within the cluster.

Data integrity copies or mirrors every data transaction to a second system. If a failure does happen to the primary system, the data for the second system can be used to recover. Disaster recovery solutions force the availability solution from a local solution to the multi-geographical solution, providing redundant systems at different locations. This allows a transfer of workloads to an entirely new location in the event of disaster at the primary location.

18.10 *Exploring Security Management*

Besides availability and performance, the third major concern for the user or company is the security of the data. Especially in situations where the data is considered sensitive and restrictions on access are required, security becomes a huge service consideration for Storage-as-a-Service solutions. Especially since these solutions are possible by sharing resources across a set of buyers.

In a Storage-as-a-Service solution, the first and most important security measure is access control. In the simplest terms, access control is ensuring that a user has access to the documents that they have a right to have access to and ensuring that no one else has the same access. Authorization and authentication procedures need to be in place to identify the user appropriately. Logical security controls are in place to provide the correct level of access controls in place to maintain an environment of integrity.

The second security measure is protection from malware or attack. Especially in a shared environment, some buyers may have a valid concern about attack from another buyer, although these concerns are rather unfounded for the most part. Putting the proper procedures and controls in place will ensure that attacks are avoided or mitigated.

Successful security management begins with the creation of a security policy. This policy starts at the negotiation table with the provider and the buyer. Understanding the buyer's security concerns and the reasons behind them, such as regulatory or legal compliance requirements, is the first step to creating the policy. The provider will translate those business requirements into the technical requirements to complete the policy. This policy will be communicated to the technical staff to implement and maintain.

18.11 *Exploring Backup and Restore Activities*

Storage is the foundation of backup solutions. In a Storage-as-a-Service solution, backups are another consideration for ensuring the customer has access to their data. Backups are used in two situations: to recover from a disaster and to recover small portions of lost or corrupted data.

Several mediums are used to store data as a backup. The most commonly used for bulk data storage is magnetic tape. This is because magnetic tape has a better capacity/price ratio than other mediums. However, technical advancements are improving this ratio in other mediums, especially hard disk devices. Recordable CDs and DVDs provide portable solutions to backup allowing restoration to any device with a CD-ROM drive.

More expensive than CDs but providing greater portability are a class of media called solid state storage. Flash memory, thumb drives, USB flash dives and the like are examples of solid state storage. Remote backup service through broadband internet access is becoming a widespread solution. In fact, some Storage-as-a-Service solutions can be utilized as such.

Backup solutions need to balance between accessibility and cost. Usually the more accessible the backup storage is, the more expensive. The high-end of these solutions is an on-line backup storage which can start a restore in seconds. Near-line storage provides a cheaper solution with some delay to recovery. A mechanical device is used to move backup media to a device that can read the media. Off-line storage replaces the mechanical device with human interaction. To protect against disaster or other site-specific problem, off-site backup solutions are available.

The strategy of a backup solution begins with the data repository model. Different models have different benefits:

- An unstructured repository can be a stack of recordable CDs with little organization. Although easy to implement, there should be no confidence in achieving full recoverability.
- Full backups are a complete copy of an entire system. Instead of performing full backups every time, policies either use an incremental or differential backup after the full backup is complete. Incremental backups simply copy the changes made since the last full or incremental backups. Recovery requires restoring the full backup and every incremental backup since the full backup. Differential backups identify the changes to the system from the full backup and make a copy of those changes. Recovery requires the full backup and the last differential backup.

Other versions of the incremental and differential backup solution are variable dump levels, mirror plus reverse incrementals, and continuous data protection.

18.12 *Discovering Online Storage*

Online storage is driven by an Internet hosting service allowing web and FTP access. There are a number of uses for online storage solutions. Software file hosting has been a large demander of online storage to make available the number of freeware and shareware applications that software authors were creating. Now, technology companies are using online storage to warehouse updates and fixes. Now software as a service solutions allow users to access software from online sites.

Personal file storage is another use for online storage. Users can keep private files on

network storage or use the space as a distribution hub for public use. Online storage is different form web hosting, so it is important to discover if the Internet hosting service restricts non-website-related use.

Companies with a major Internet presence often utilize online storage options to offset issues of bandwidth congestion. As cloud computing options become more available, companies are starting to look to online storage for more reasons. Collocation provides time and cost savings to company because they merge their mission-critical equipment with their data center. Web-commerce companies use online storage for a security, cost-effectiveness, and redundancy for their websites. Some companies find online storage as a cost-effective solution to disaster avoidance, offsite backups and business continuity.

Most online storage providers offer space on a per-gigabyte basis charged monthly. Additional charges may apply dependent on the security, availability and performance requirements of the company. These charges are often far less than procuring new network hardware for in-house storage.

19 Achieving a Cost-Effective Storage Solution

For most companies, particularly SMBs, storage solutions need to balance cost with quality systems that meet the minimum performance and capacity demands for the business. The question becomes: how does a company obtain this balance?

19.1 *Higher Bandwidth Networks and Shared Storage*

Increasing the bandwidth of the network allows less equipment to handle more network transactions. New technologies in network equipment have provided servers that are less complex, more flexible and affordable. Consolidating data allows an opportunity to let go of old, degraded equipment that put a financial strain on the organization due to more maintenance and management.

19.2 *Consolidating Data Centers*

IDC, a provider of market intelligence, estimates that the average corporate server is operating at less than 10 percent utilization. Matching dollar to utilization, 90 percent of technology spending is not being utilized in the business environment. It stands to reason that companies will want to find ways to consolidate their systems to save costs. At one time in the corporate world, each department was assigned a different server, each application was built on different solutions, and each technical push created a different segment. The traditional method of running IT was an expensive and complex maze of connected systems.

Today, the push is to consolidate servers, and even entire data centers. Removing servers from server closets and equipment rooms, centralizing server operations in one location instead of multiple locations, moving to high-performance platforms are all ways that companies are finding ways to provide greater service at a much lower price.

19.3 *Virtualization*

The advantage of virtualization is the realization of further consolidation and more utilization of their systems. By creating a logical representation of the network, the company can gain a better grasp on the organization of applications and data and how they are used. Virtualized environments are easier to manage, allowing easier deployments of new software, less maintenance because there are less servers, and rapid replication and workload balance activities.

20 Storage Options

20.1 *What is a Storage Area Network?*

The SNIA (Storage Networking Industry Association) defines Storage Area Network (SAN) as “any high-performance network whose primary purpose is to enable storage devices to communicate with computer systems and with each other.” Expectations are already set that computers will be connected to storage devices, what makes a SAN different from a traditional network is that the universal connectivity throughout the network. A traditional network would have computer clients connect to application servers on the network. The application servers would have isolated data storage units attached to them. Some applications would use more storage than others. But the greatest problem was when multiple computers had to access the information, specifically when connection wasn't available.

In a SAN, computer clients are still connected to the application servers; however, the application servers are connected to an adjoining network of storage devices. The advantages of this solution are several-fold. First, information stored on the SAN can be shared with multiple applications and thus multiple clients without any additional changes to the network. Second, storage devices can be utilized at greater efficiencies, as the network and applications will control with the data resides. Third, any additional clients, applications, servers, or storage devices can plug into the network with far less ease then traditional methods. And most importantly, the ability to identify duplications in information because of the use of applications asking for the same information is easier.

The concern with a SAN solution is ensuring that it is secure against disruption or data corruption. In other words, it must be highly available. The focus isn't just on the availability of the network, but also the individual storage devices, computers, and interfaces must have strategies to handle component failures.

Additionally, the SAN must be able to meet or exceed performance expectations. With multiple connections to computers and storage, the number of transactions required by the network will increase. High data transfer rates and low I/O request latencies are required to increase the reliability of the SAN. A well-managed SAN will separate different types of traffic on the network, such as high volume I/O traffic from client/server message traffic. Ultimately, the SAN will be able to grow in performance as the demands on the network grow.

20.2 *Storage as a Software Solution*

For the most part, storage is a hardware solution, requiring huge capital investments in storage drives and media, network interfaces, and cabling. However, the hardware is only the foundation for a storage solution. What drives the solution is the software used to manage the storage capacities and performance of the network.

One of the most expensive and underutilized storage components is the tape drive. It is only in use when performing a backup which is perhaps once a day on critical systems. Through a SAN, the tape drive can be connected to the network and thus, connected to multiple servers at the same time. At this point, the problem becomes ensuring that two servers do not attempt to write to the drive at simultaneously. SAN software manages the scheduling and organization of the backups to tape.

The same way two servers cannot write to the same tape drive at the same time, two clients cannot write to the same storage drive. The solution is the allocation of a virtual disk. The system simply remembers which virtual disks are allocated to which servers. When a client attempts to write to storage, it essentially writes to the virtual disk. The virtual disk makes a call to the physical storage and if it is free, completes the writing to the server. This allows a transmission even if the final server is down. Management software handles these processes needed to allocate virtual disks and manage the communications.

One of the complex operating system solutions is enhancing application availability through clustering of servers. The purpose of this clustering is to allow servers to share the workload in the event that one server goes down. When a server fails, the resources on that server must be transferred immediately and completely to another server. The application must be restarted and its state at the time of fail must be recreated as closely as possible. Though all of this activity can be done manually, using software allows it to be done automatically and at a faster rate.

Applications outgrowing the capacity and performance of the server they are on is a predictable occurrence. The first tendency may be to replace the server with a larger system, though this is unnecessary. What can be done is adding another server which runs a separate copy of the application and processing the same data. The problem now is the high probability that two instances of the same application will attempt to write to the same data. To prevent this from happening, software can assist in the management of data use.

There are two types of software used to manage storage solutions. System applications build upon the basic properties to provide a functional environment. They allow for effective clustering, data replication and direct data copy. Management applications handle the more complex functions, including zoning, device discovery, and allocations.

20.3 *Exploring Storage Devices*

Storage devices are the building blocks on any storage network. Without storage devices, it would be impossible to read or write data across the network. Storage devices are the beginning and ending of any computer transaction. They operate unseen by the user, using advanced electronics, chemistry, and magnetic physics. The demand on capacity and performance improvements for storage devices have created an upsurge of scientific research on the materials used to create storage. This is a prime example of this research and development on fiber optic technology. Before, all networks were highly limited by the conventional use of cabling using electrical signals, metal and insulation. With fiber optics, the use of glass and light to send a message created a breakthrough on how networks ran and the components on those networks.

There are a number of considerations to be mindful when creating storage networks. Storage devices can used either as components of the storage subsystems of the network or as standalone products connected to the server systems. There are a variety of storage devices available, each with different advantages and disadvantages like how the data structures on the disk drives can manipulated by the technical staff. Performance and capacity can be optimized. And for long term storage or to allow for recovery functions tape drives are used. For the storage professional within the cloud, the beginning of every effort starts at the beginning, the storage device itself.

20.4 *Understanding the Disk Drives*

Of all the storage devices, the most important is the disk drive. Because of their popularity and availability at discounts rates, disk drives have become an assumed and unimportant feature of the computing environment. However, disk drive technology is one of the most sophisticated and interesting technologies of all the IT technologies.

20.4.1 Platters

The major components of the disk drive include the disk platters, the read/write channel, arms and actuators, drive spindle motor, servo control electronics, buffer memory, and disk

controllers. Each of these components is designed to provide the greatest optimizing independently and working together.

The physical part of the disk drive where data is stored is the platter. Platters are rigid, thin circles spun by the dive spindle motor and have three layers. The rigidity of the platter is provided by the substrate. Materials used to make this layer have been glass, ceramic, and an alloy of aluminum and magnesium. Since disks are recorded at the microscopic level, the physical attributes of the platter must be free of any imperfections. Vibration and heat due to friction is present when spinning at high revolutions per minute (rpm) and can cause problems if there are balance imperfections due to materials defects. Platters must be extremely flat and be able to handle expansion and contraction due to changes in temperatures.

On top of the substrate is the magnetic layer where data is stored. Most disk drives use a thin film that is very smooth and only a few millions of an inch. The layer is created by spraying vapor molecules of the magnetic materials on the surface of the substrate. The characteristics of these magnetic materials determine how many bits can be written per square inch, called the areal density.

The protective overcoat layer prevents dust, water vapor, and disk head crashes from damaging the platter. Of course, this protection is lightweight and the best choice for protections is to operate storage drives in clean, dust-free environments with control over the temperature.

Disk drives are created by stacking platters on top of each other with spacers between them to allow disk arms and heads to access both sides of the platters. Each platter provides storage capacity. Recently, platter capacities have exceeded 100 GB per platter.

20.4.2 Read and Write Heads

Data is transmitted to and from the platter using read and write heads. When writing, heads impress the magnetic layer of the platter with magnetic signals. Reading simply detects the presence of those magnetic signals. The technology used by the heads has a major impact on the performance and capacity characteristics. Most disk heads today use giant magnetoresistive (GMR) technology. GMR technology reads data by detecting resisting variances in the magnetic layer. Very low strength signals are used to write using GMR technology.

The height of the head from the platter is impacted by the way the data is transmitted to and from the platter. This is called the flying height or head gap and is approximately 15 nanometers. Since this gap is typically smaller than most microscopic dust particles,

reliability of the disk drive is improved if they reside in environments that are cool and free from humidity and airborne contaminants. The use of the term “flying” is appropriate because the air movement caused by the spinning platters passing over the heads provide a lift to the heads that is similar to the aerodynamic physics of an airplane wing.

20.4.3 Read/Write Channels

A physical recording is an analog signal. Unfortunately, most data used in today’s computing environments is digital. This requires a conversion mechanism to translate the digital logic into something that can make an impression on magnetic media. The read/write channel of a disk drive subassembly is this mechanism.

The read/write channel is found in small hug-speed integrated circuits. These circuits utilize sophisticated techniques for processing and amplifying signals. Reading data from the disk does not detect the magnetic signal written on it, but by detecting very small differences in the electrical resistance of the media. This resistance is discovered by a microscopic head that floats over the platter as it spins at very high speeds. This resistance can be so faint that amplification is required using the read/write channel.

20.4.4 Arms and Actuators

Since most components of the disk drive are microscopic in nature, so are the components that connect them. Disk arms are thin and triangular allowing the read and write heads to connect precisely over specific tracks on the platters. The disk arms are then connected to the drive actuator responsible for positioning the arms. The movements of the actuator are controlled by voice-coil drivers. These drivers are derived from voice coil technology in audio speakers, since some speakers have to vibrate at very high frequencies to reproduce sounds. Disk actuators using voice coils can reproduce signals at very high speeds. It is the actuator that people sometimes hear in disk drives.

20.4.5 Drive Spindle Motor

The rotation of the drive platters are caused by the drive spindle motor. The motor is designed to maintain a constant speed with little vibration for long periods of time. Most drive failures are really motor failures. These failures are typically not because of poor design, but because they are moving at higher speeds with less noise and power consumptions.

The platters are connected to the spindle which is directly attached to the drive shaft of the motor. Separator rings are used to space the platters precisely. Of all the parts of the disk drive, the bearings of the motor have the greatest wear and tear, making them the primary cause of most disk drive failures.

Servo-controlled electronics is a technology used by many applications to ensure that the spindle motor maintains a constant speed. Ideally, servo-controlled closed loops fine tune automated systems. Feedback control circuits detect speed variations in the rotating platter. When the speed varies too far one way or the other, the circuit counteracts the change by changing the voltage to the spindle motor.

20.4.6 Buffer Memory

The performance of disk drives is limited because of the mechanical nature of reading and writing data on platters. This performance is approximately 1000 times less than data transfers to memory chips. To compensate, an internal memory buffer is located between the drive and the storage controller using the drive. The purpose of this memory buffer is to accelerate data transmissions.

The performance benefits of the buffer memory are greatly realized in storage subsystems supporting applications requiring high throughputs. Buffer memory can be used by a subsystem controller when there are overlapping I/O requests across multiple drives. Then internal transfers of data can be made between the drive and buffer memory while the subsystem controller is working with another drive. As a result, I/O performance is improved for applications that read and write small bits of randomly accessed data by using buffer memory. Applications that use streaming technology do not realize benefits from buffer memory.

20.4.7 Disk Controller

All disk drives have internal target controllers. These controllers respond to commands from external initiators from the host of subsystem. Within the drive, a storage controller executes the command to the disk drive. The name, firmware, is used to refer to the software component of the disk drive controller. This firmware is stored in e-prom chips on the circuit board of the disk drive.

Processor chips have been improved constantly with faster cores and more memory. These processor chips are used as disk drive controllers. Unfortunately, traditional use of disk drives are as slave devices responding to I/O requests from host or subsystem

initiators, which does not utilize the additional functionality of the processor. Storage applications can be added to disk drives, but in direct competition with system and subsystem vendors. Instead, the new processor intelligence has been used to increase reliability and ease of use by providing more robust controllers.

20.5 *Understanding Data Structures*

In order to configure storage optimally, an understanding of how data is structures on the disk drive platter is required. Disk platters are formatted into tracks, a system of concentric rings. Within each track are sectors. These sectors subdivide the track into a system of arcs which are formatted to hold the same amount of data. Cylinders are a system of identical tracks on multiple platters in the drive. The arms of the drive move together forcing the heads to be in the same relative position on all the platters simultaneously.

Disk partitions create logical containers by dividing the capacity of the physical disk drives. Multiple partitions can be created on a single disk drive, allowing the flexibility to create different virtual disks for different purposes. This is especially important when the user wants to reserve storage capacity for different applications or for different users of the system. A common reason for multiple partitions is to store separate operating systems or file systems, such as Linux and Windows, on the same machine. Partitions are a contiguous collection of tracks and cylinders starting from the outside edge of the platter and moving inward. The numbering of the partition starts with 0. So if a disk has 3 partitions, the number would be 0, 1, 2 with partition 2 being closest to the center.

Though the system of cylinders, tracks, and sectors exist, they are not utilized as they once were by systems and subsystems using disk drives. A method called logical block addressing (LBA) has replaced cylinders, tracks, and sector addressing. LBA presents a single flat address space making disks easier to work with. Using LBA, many different types of disk drives can be integrated easily into a large storage environment allowing more flexibility in the storage network.

Because of the areal density and microscopic nature of disk technology, bad sectors are going to appear on any manufactured disk. These bad sectors are compensated by disk manufacturers by reserving spare sectors for remapping. In the traditional method of addressing, I/O requests that hit a bad sector had to be rerouted to the spare sector replacing the bad one, creating a performance downgrade. Because LBA creates a single address, the bad sectors are overlooked by the I/O request and performance is improved.

Radial geometry is an important discipline when dealing with disk drives. It demonstrates that the amount of recording material increases as the tracks move from the center of the

platter. Disk drive designers take advantage of this geometry by creating zoned-bit recording. This creates more sectors in outer tracks as compared to the inner rings with the intention of segmenting the drive into zones for the purpose of having the same number of sectors on all the tracks within a zone. Logical block addressing (LBA) allows disk drive manufacturers to take advantage of zoned bit recording by allowing them to create zones without concerns about the impact on controller logic and operations on the hosts and subsystems. Standardized zone configurations are not needed since platters will never be switched between disk drives.

20.6 *Disk Drive Specifications*

Understanding and interpreting disk drive specifications are important to deciding what disk drives will fulfill the storage requirements of the company.

To understand the expected reliability of a disk drive, a storage professional looks to the mean time between failures (MTBF). Many disk drives are tested over a short time and run through statistical methods to derive the MTBF. Extrapolated results are usually noted in a very high number of hours, typically between 50,000 to 1.25 million. 1.25 million hours is approximately 135 years. This MTBF specification sets the expectations for how many disk drive failures will occur within an environment of many drives. For instance if a disk drive will have a single failure every 135 years, 135 disk drives will have a probable failure rates of one per year. Larger storage networks will definitely experience disk drive failures, so spare drives should be available as well as disk drive redundancy techniques.

Speed in reading and writing to disk is an important factor in choosing the proper disk drive. This is controlled by the rotation speed of the disk drive, usually found as the rotation per minute (RPM), Therefore, a 15,000-rpm disk drive can do twice as much work as a 7,500-rpm disk drive.

Another specification of note is the rotational latency. When the drive is working a request, the heads need to move over the proper track on a platter. Then the sector passes under the head. The data transfer can only be made when the head and sector meet. The time spent waiting for this connection is the rotational latency. It is the average time for random I/O operations and is represented as time it takes for a platter to complete a half rotation. The range for rotational latencies is 2 to 6 milliseconds. Compared to processor and memory device, this is rather slow, which is why applications can suffer I/O bottlenecks. To prevent these bottlenecks, disk drives should have high rotational speeds and large buffers. This is especially important for application that has lots of transaction processing, data warehousing, and multimedia streaming.

Seek time measures the time required for the actuator to reposition the heads from one track to the next on a platter. Average seek times are a performance specification over many I/O operations and range from 4 to 8 milliseconds. Applications which perform large amounts of I/O operations in quick succession require minimal seek times.

The performance of bit read/write operations on drive platter is measured by the media transfer rate. Most specifications are measured in terms of bytes; the media transfer rate is measured in bits. It measures the read/write performance on a single track. And since tracks on the platter are larger on the outside than on the inside, the media transfer rate can be different and therefore often expressed as a range. The fastest media transfer rate represents the tracks in zone 0. Instead of the media transfer rate, the sustained transfer rate can provide an actual data performance measurement.

It takes into account the seek time and rotational latency physical delays. This is important to the sustained transfer rate because most I/O operations work across multiple tracks and cylinders require the ability to change the location of the read/write heads. Unfortunately, the sustained transfer rate relies on optimal conditions that are difficult to match with actual applications. The size of the data object and fragmentation in the file system can reduce the actual sustained transfer rate. However, the sustained transfer rate is a good measurement for determining the drive's overall performance capabilities.

20.7 *Optimizing Disk Drive Performance*

In comparison to most computing operations, disk drives are slow. As a result, several techniques have been developed to increase disk drive performance. They include:

- Limiting drive contention
- Short-stroking drive
- Matching rotation speeds
- Aligning zones.

Performance can deteriorate when multiple applications attempt to perform I/O operations on a single disk drive. This simultaneous capture of the drive requires a significant amount of seek time and rotation latency. Understanding which applications are accessing each disk drive is desirable to understand how to avoid contention. In a storage area network, disk bottlenecks can happen unchecked when many disk drives provide storage to many systems and applications. Poor planning in partitioning can have two applications competing for a single actuator and rotating spindle.

To limit contention, limit the number of partitions. Ensure that two high performance applications do not share the same disk drive. Assign lower speed disk drives to low

performance applications.

Short-stroking a drive uses a subset of the available tracks to limit the drive's capacity. The result is that it reduces the average seek time because it limits the actuator’s range of motion. If the limit is at the outer edge of the platter, the range of motion stays within the innermost range of the platter. It also increases the media transfer rate by using the highest density of data per track which is found on the outer tracks.

Since the drive in storage management is to make operations easier to manage, most details are treated the same including storage address spaces. Although this is good for most applications, those that need the highest performance possible need to be managed differently. The obvious difference between disk drives is their rotational speeds. If an application requires higher performance, it would make sense to assign it a high-speed drive. The same theory works with buffer memory. High performance applications would need disk drives with sufficient buffer capacity.

Different zones on disk drives can have performance differences. This can be a problem when applications, such as data warehousing processing, expect consistently high I/O rates. A disk partition on the outside of a disk can have twice the performance as an inside partition.

20.8 *Exploring Tape Drives*

Backup and restore operations rely exclusively on tape drives. As a rule, there are fewer tape drives than disk drives. The primary difference between tape drives and disk drive is the ability to remove tape drives, particularly to transport it to a safe location or to share data over the “sneaker net.” Sneaker net is a conventional mode of transporting data from one place to the next. Companies have used courier services or air cargo. Another difference between tape drives and disk drives is their read/write ratios. While disk drives are used mostly to read data, tape drives are used to write data.

The media for tape drive is magnetic tape. It is constructed in four basic layers: backing, binder, magnetic material, and coating. The backing of the tape provides the flexibility and strength in the material used. The backing also acts as protection from other sections of tape when rolled up tightly and stored for long periods of time. The backing is attached to the magnetic material by a flexible, glue-like material called the tape binder. The magnetic material is where the data is written and read. The coating layer provides a smoother surface to increase the life of the tape.

Like any physical material, tapes deteriorate. Cracks in the surface, tears along the edge,

and the corrosion of metal oxides are all problems faced during the life of a tape. The deterioration can be slowed by storing tapes in conditions of low humidity and moderate temperature. This environmental control is required for unused tapes as well. If tape deterioration starts before the tape used, data loss is more probable.

Tape heads are designed to remain in contact with the tape when reading or writing. The constant friction caused by this contact will wear the tape heads over time. A general rule is to clean tape heads after every 30 hours of use. This is especially important because unlike disk drives, a tape drive is exposed to airborne particles every time a tape is inserted or removed. Additionally, tapes shed fine pieces as they run through the transport, sticking to the tape heads and reducing effectiveness.

Performance of tape drives is based on a sufficient amount of data being transferred and the compressibility of data. Tape drives run at different speeds. The streaming transfer rate is the speed at which data is written from buffer to tape. It transfer rate assumes that a sufficient amount of data is being written to the buffers by the host of subsystem. Start-stop operations happen when there is insufficient data to continue streaming mode operations. Since the tape stops, rewinds, and starts again, the speed of the data transfer diminishes greatly. If there is no data to record, the tape drive will stop and reposition the tape until there is data to record.

Compression technology is a feature to boost data transfer rates, sometimes several times beyond native rates. Different types of data vary greatly in terms of compression; multimedia data does not compress while database data can compress several times.

Tape technology has had compatibility and interoperability problems because of obsolete tape solutions. Today, most tapes fall into one of two technologies: linear and helical scan. Linear tape places “lines” of data lengthwise on the tape media. Linear tape drives use multiple heads in parallel, allowing simultaneous reading and writing of data. Two primary, competing linear tape technologies are super digital linear tape (SDLT) and linear tape open (LTO). SDLT has roots in the digital linear tape technology developed by Digital Equipment Corporation for its VAX line of computers. LTO was jointly developed by IBM, HP, and Seagate.

Helical scan tape technology was first introduced for video recording applications. It writes date in diagonal strips along the tape. Data density is better with helical scan tape but it has less capacity than linear tape. Two different technologies exist for helical scan tape: Mammoth-2 and AIT-3. Exabyte Corporation developed Mammoth tape technology in the mid 1990. At the same time, Sony developed the Advance Intelligent Tape (AIT) to provide greater reliability, performance, and capacity. A memory chip is embedded in the AIT tape cartridge.

20.9 *Introduction to File Area Networks*

Most of the conversation has been focused on the traditional fundamentals of storage management. With the introduction of cloud computing and storage as a service commodities, companies are now placing the responsibility of these fundamentals on a vendor. This provides benefits in cost and management on the company. However, it does not relieve the company of total responsibility over what is placed on storage. File management is a symbiotic discipline to storage management. Essentially, there are two forms of data being stored: raw data used for databases and files. Files are a growing part of a company's information infrastructure. Whether the data is raw or files, when it is stored on a storage device it is in the form of block data.

Today's technological world means more files need management, increased complexity in file systems and larger applications using files, all requiring a better approach to file management. The File Area Network (FAN) is a suite of file management technologies used to streamline management of file-based data. From the existing network or SAN, the FAN utilizes the upper-layer network file system protocols like Network File System (NFS) or Common Internet File System (CIFS). The FAN, however, is distinctly different from the underlying network that carries it; in fact, the FAN is a logical approach to implementing file-based data storage, connectivity, and management.

The FAN simply utilizes the SAN without changing it. It provides to the company:

- Broad control of file information including file attributes
- File visibility and access control without concern for location or physical device
- Transparent movement of file data across platforms or geographical boundaries
- Removal of redundant file resources and management tasks
- File data management support in both data centers and branch offices.

To improve storage capacity and performance, a number of technologies, methods, and controls have been introduced to change how the storage devices are viewed, managed, and connected to the network. For the most part, all these new technologies are focused on the physical properties of the storage area network. And though data is sometimes moved from one drive to the next, backed up and recovered, changes in storage management do not change the properties or attributes of the file management solution. FAN concepts are in place for the same reasons that SAN concepts are in place. The difference is that the FAN deals with the logical file system and the SAN handles more the physical system. The FAN requires the SAN but not vice versa.

The reasons are better storage management, storage consolidation, business continuity and disaster recovery, storage performance optimization, data lifecycle management,

remote site support, data classification, and data reporting.

The FAN technology sits on the network infrastructure. There are seven broad components to the FAN solution with each component requiring some form of support technology. Those components are:

- Users who access files
- File server connectivity for users
- Policy driven file control and management
- A namespace
- Devices that present files
- File systems residing on servers
- Back-end storage devices.

21 File Area Networks

21.1 *The Storage Area Network from a File Management Perspective*

A network is comprised of hub, switches and routers. Switches and hubs are considered “Layer 2” devices in Internet Protocol (IP) terms; meaning that in an Ethernet network, these devices operate in the second layer of the IP layered model. A hub utilizes a shared bandwidth architecture which prevents a port from being used when it is already connected to another port. Switches, on the other hand, can allow any number of pairs to connect and transmit to each other. The router is similar to a switch in that it does not have restrictions to how many connections can be made. Unlike switches though, routers operate in the third layer which allows it to connect autonomous or semi-autonomous network segments into a hierarchical structure.

On top of the physical networks is a set of protocols that must be in place to handle file system management. Protocols are commands that must be followed by computers and network devices in order for them to communicate. If computers and devices do not use the same protocol, they cannot speak to one another. All levels of communication require protocols, from the physical media and cabling to the applications. In order for two devices to communication all the protocols at every level must work properly. The set of protocols together are called the protocol stack. Some of these protocols are particularly important to the FAN. The Internet Protocol (IP) is the standard for Internet communication and thus has become the standard for corporate LANs to support applications as email and desktop Web servers. As a result, the IP protocol is the choice for front-end component of file networks because it can handle very large, distributed, loosely coupled infrastructures. Upper-level FAN protocols use IP as a transport for file reading and writing.

Raw block data is stored on a storage device. An application requires the raw block data to be mapped into a “cooked” file system format. In Windows, two protocols are used: New Technology File System (NTFS) and File Allocation Table (FAT). In UNIX, there are more options, like XFS, UFS, and VxFS. Protocols are also required to map cooked file systems into raw block data. Two protocols dominate existing file systems: NFS and CIFS.

Network File System (NFS) was developed by Sun Microsystems and was the first widely deployed network file system. It is the choice to use in UNIX environments. In Windows, third-party software is required. In a NFS environment, the client requires access to data stored on the server. The server runs NFS processes, the server configuration determines

which directories should be used, and security administration identifies which clients should have access. Once access is provided, the file system can be worked by the client as if it resided locally. NFS does not provide a good mechanism for specific defining of file access privileges.

The Common Internet File System (CIFS), or Server Message Block (SMB), was developed by Microsoft and used to allow communication between Microsoft platforms and network storage devices. In order to use CIFS file systems from a UNIX platform, third-party software is required. Neither NFS nor CIFS protocols work well in WAN environments, because the length of the connection between client and server reduces performance and reliability. This is the primary reason for decentralization of network components in large-scale environments. Both protocols must map the location on the client which users see remote files to the server. In large environments with hundreds of servers, it can be very difficult to manage these mappings, called mount points. The FAN model fundamentally addresses this issue.

21.2 *Reasons for File Area Networks*

The implementation of a FAN is done on top of the existing network. Any investment into FAN technology should be considered as an enhancement to current infrastructure with the intention to reduce costs, manage improvements, and improve performance. There are many FAN solutions, below are just a few.

Storage consolidation combines multiple storage resources into fewer resources. It provides reduced costs and better management of the files. In DAS environments, storage must be assigned to each host: it cannot be shared easily with other hosts and it cannot be located far away. Each storage device in a DAS environment needs to have substantial unused space, or white space, to allow for growth. In a FAN environment, most white space is pooled together. Therefore, any host can access any storage device to get any free space it needs. Consolidation of storage from a SAN perspective is different than the FAN perspective. First of all, consolidation of a SAN handles all data, not just the files. Second, from a SAN perspective the consolidation can be reduced to simply moving data from one device to another.

From a FAN perspective, consideration has to be made to understand impact of the move on file systems. Additionally, the FAN approach handles the client perspective more effectively than the SAN approach which comes from the back-end of the infrastructure. That being said, storage consolidation from a traditional perspective will result in the potential removal of a storage device. If one hundred clients are mapped to that particular storage device, login scripts will need to be adjusted to remap the clients to the new

location of the stored data. That means touching each client. FAN technology provides a set of tools for a complete consolidation solution.
The cornerstone of the FAN approach is the Global Namespace (GNS). GNS virtualizes the view into the storage server, the same way the Domain Name System (DNS) virtualizes the view into the Internet. In fact, the GNS allows users to access files in the same way they access web sites. Without a GNS, each mount point must be connected to a physical file server and are created and maintained by the IT department. If the files are moved, the mapping needs to be changed. Though there are automatic methods, each has their own problems and will require the client to reboot. Without GNS, troubleshooting problems related to file system connections can be extremely difficult because though a file server may show on the client as “drive E:”, the same mapping on another client may map to a completely different file server. With GNS, a directory structure is created that is logical to the users and to the IT department and maps user requests to the correct physical location. So instead of having to change mount points on every client, one mount point within the GNS is changed.

Data migrations are a part of life in the storage area network. Though most data migrations happen within a data center, sometimes it is required between sites. In non-FAN environments, migrations are difficult to conduct without major effort and downtime. Performance concerns, downtime to users and user complaints, complexity of the change control plan, and compatibility issues are just a few concerns with data migration. In a FAN, data migration effort and risk can be minimized to the point where there is no downtime. FAN components aggregate storage across multiple distributed file systems, so administrators can complete migration without regard to location, vendor, or underlying file systems. This migration can be done behind the GNS without interrupting user access. The GNS makes data movement transparent to the users.

Disaster Recovery and Business Continuity are can also be simplified by a FAN solution: for instance, by providing a file system in an active/active DR pair. Without FAN, remapping, rebooting, and other initiation activities must be performed. Yet with a FAN solution, the failover activities are transparent to the user.

A complete FAN solution includes some level of optimizing the WAN. Wide Area Files Services (WAFS) overcome any challenges related to consolidation and file collaboration over the WAN that are the result of performance or infrastructure issues.

21.3 *Building a File Area Network*

There are a number of products that are used to create an optimal file area network. This section simply describes the theoretical approach of these products.

Administrative costs in a distributed file storage environment increases because each storage device needs independent management. To reduce these costs, IT departments are looking for new ways to simplify file system management. In the end, the goal is to automate as many administrative tasks as possible, such as data migration and consolidation, information lifecycle management, storage optimization, and data classification and reporting. The core of this automation is the global namespace (GNS). The GNS manages how the user sees the files independent of its location, file type, and size.

Wide Area File Services (WAFS) address difficult branch office IT issues, such as storage and server proliferation, management of backup and restore capabilities, and ownership costs. The WAFS creates LAN-like access to files shared on the Wide Area Network (WAN). This allows companies to protect remote data and consolidate assets. By centralizing the core file directory with direct connection to appliances at remote locations, the WAFS enhances performance and reliability characteristics of a WAN. This allows operations to be streamlined in remote locations.

Data growth is inevitable in IT solutions. Data growth occurs unevenly on storage devices because different applications are used less or more. As data grows, additional expenditures are required to obtain more storage, hire more personnel and to ensure compliance with regulatory requirements. The best solution to this problem is an active file lifecycle management (FLM) program which manages the file from creation to disposal. FLM solutions apply attributes and access patterns to the files to optimize data management. It manages user access to a minute degree. For instance, a network may restrict access to read/write access. FLMs can expand on this access to define readers, writers, editors, and approvers. Retention and maintenance of files are policy-based through a FLM, simplifying compliance to any document quality regulations. In addition, file lifecycle management handles identifying underused or old files, archival storage management and retrieval, retention and deletion policies, and file classification. In essence, every aspect of the file is handled.

22 Storage and Security

The association between storage and security is still relatively new. As storage networks have been implemented, there has been more concern about separating the computer systems and storage. The rapid growth of the Internet has led to a great increase in the information handled by computer systems. This has made the question of scalability critical to new computer systems and improvements to existing computer systems.

Near perfect availability requirements has forced companies to understand the importance of information on business operations. The importance of storage has also elevated the discipline to the forefront of computer technology, specifically in terms of containing and protecting a business's key asset, its information. Assurance techniques, such as replication and backup and restore, ensure that information is never lost and establishes an audit trail.

Corruption of data is another primary concern, whether that corruption happens during a transmission or because another computer system accessed the storage device. Three major facets of security exist: authentication, non-repudiation, and integrity. Each of these assist in preventing data corruptions as well as providing other benefits. Authentication is a security measure used to establish the validity of transmissions, messages, or users. It allows a user to have confidence that information was received from the known source. Authentication is necessary for authorization.

Non-repudiation intends to prevent any denial that an action happened, or communication took place. Essentially, it places a time stamp on any transmission, namely those requiring authentication. Integrity speaks to the user's ability to determine if the information received has not been altered in transit or any method other than those used by the originator. An additional security measure is confidentiality which protects against disclosure of information to people other than the intended receiver.

The problem of security within storage started with the demand and growth of storage area networks. The activity of creating storage networks did nothing to replace or redesign the software required to carry a request from an application to a storage device. Hence, much of the storage stack used for direct-access storage was still used for storage networks with just a few enhancements. This eased migration from standalone storage to network storage, but inherently had some problems. Security was not a problem for storage that was directly attached to computer systems.

When storage was connected to the network using the same storage stack, there were insufficient or no security precautions applied to the network, hosts, or storage device. This doesn't mean that security was not a concern and not put into place, only that security was a by-product when implementing storage area networks. Now, the importance of security in storage area networks is just as critical as any other network.

22.1 *Introduction to Cryptography*

The concept of protecting a message from being read by persons other than the intended recipient has been around for thousands of years. The ancient Egyptians use to replace common symbols in hieroglyphs with special ones. The ancient Hebrews replaced common words in their scriptures using a substitution schema. Today, these processes are called cryptography, which is derived by the Greek words kruptos and graphia, which mean hidden writing.

Most children are introduced to cryptography by small secret code games found in magazines: replacing the alphabet with different letters in a predefined manner such as A+3 (ABC... is now DEF...). This substitution is known as the Caesar Cipher because of its use during the Gallic Wars by Julius Caesar to prevent military plans from being intercepted and understood by the enemy. Shifting the alphabet as the Caesar cipher does will provide 25 possible keys.

Randomly substituting the alphabet provides even more possibilities. The downside of this randomization is the difficulty in remembering the key. The key is a word that when applied to the cipher would “open” the code to reading. For centuries, the “monoalphabetic substitution" cipher was thought to be unbreakable until Arab scientists broke the code in the first century A.D.

Around the 16th century, a polyalphabetic substitution cipher started to emerge. This cipher was created by applying the monoalphabetic substitution method several times. The key word would determine which translation was to be used for the message. So if the key word had five distinct letters, there were five translations that applied. For a few more centuries, this coding process was foolproof. At least until the electric telegraph brought increased focus on breaking the code. Breaking of the code was similar for both monoalphabetic and polyalphabetic substitution methods involved identifying the characters that were used the most often and reverse engineering the code. In the English language, the letters “e” and “s” are the most used letters in any phrase. Patterns would also be clues, such as “-ing” and “-ly” or “re-” and “pre-”. For polyalphabetic substitutions, the key word is not too long and would have been used over and over.

The 20th century brought a realization that an electrical machine could do the encoding and decoding better than the error-prone manual methods and these machines could managed “jumbled alphabetic” codes. In jumbled ciphers, the recipient simply needed to know how to wire the machine. The German Enigma machine was an example of these automatic cipher mechanisms which created seven substitutions for every letter through a series of three wheels and a reflector plate.

Modern cryptographic methods involve binary data rather than text but still use the operations of substitution and transposition. There are two types of ciphers: the stream type that operates on a single character (byte) and the block type that operates on a fixed number of bytes. The major advance in modern times is not related to the cipher but the key itself. Public key encryption and decryption utilizes two keys, one for each process. The two keys are mathematically related by but cannot be discerned from each other. Typically, the encryption key is publicly known while the decryption key is privately held. The first practical public key encryption came from MIT in the early 1970s. Today's computer systems utilize public key encryption, also called asymmetric cryptography.

22.2 *Standard Security Approaches*

Three types of cryptographic schemes are used commonly today: symmetric, asymmetric, and hash function. Symmetric uses the same key for encoding and decoding the message. This requires the key to be kept secret except from the sender and the receiver. Typically the key is transmitted separately from the message for extra security. The method for communicating the key does not need to be electronic either, but can be a physical device carrying the key or a paper transcription. Symmetric cryptography is efficient and highly confidential as long as the exchange of the keys is secure.

Asymmetric cryptography utilizes two different keys for the encoding and decoding respectively. The first step in sending a message is to retrieve the public key for the recipient from a public database. Using this public key, the sender encrypts the message and sends it to the receiver. The receiver decrypts the message using a compatible private key that is available to the receiver and has never been placed in a public database. Secure key exchange is not required for this method. Asymmetric cryptography is computationally intensive.

Retrieving a public key is a nine step process. It begins with a Registration Authority delivering a certification application to the applicant. The applicant completes the application. An appropriate key pair is generated using a browser or local software. One of the keys is returned to the Registration Authority to complete the application. The Registration Authority reviews the application. If the application is acceptable, it creates a

certificate request. The certificate request is sent to a Certification Authority. The Certification Authority generates a certificate which includes the public key and relevant information. The certificate is returned to the applicant who can post the certificate in a public database.

One of the primary uses of asymmetric cryptography is to have a secure method for exchanging the common key used in symmetric cryptography. When a message is created, it is encrypted using one key. That key is then encrypted using a public key obtained from the Registration Authority. The encrypted message and key is sent to the receiver. The receiver must have the software available to decrypt the key. Using the decrypted key, the receiver can then decrypt the message.

The intent of hash function cryptography is to guarantee integrity instead of confidentiality. The cipher is applied to the text and produces a hash value. The value is then appended to the message which is sent to the receiver. The receiver applies the hash value to the message. If the two match, it guarantees that the message was not altered during transit. Security can be added to the process simply by the creating a key along with the hash value. In fact in these cases, the hash value can be encrypted and decrypted but in reverse to the message. The hash value would be encrypted with a private key and decrypted with a public key. By doing this, the receiver can be ensure that the message came from the intended sender.

22.3 *Algorithms and Standards*

Every cryptographic scheme has a number of ciphers available to be used during the encryption process. Most ciphers used by storage management are block ciphers and utilize symmetric cryptography schemes, because the complexity of asymmetric cryptography hinders the performance required by a storage area network. A cipher uses two inputs, a key of defined length and some plaint text to create a single output called ciphertext. While stream ciphers focus on a single character or bit, a block cipher utilizes a defined number of bits. The process of encryption is sometimes referred to as a “round” and modern ciphers utilize several rounds to create the final ciphertext. The strength of the cipher is dependent on the randomness of the key. Additionally, keys can be reused. Unfortunately, the more a key is used, the greater the chance of the cipher being broken.

Modern technology has now used the key to determine the amount of information that can be safely transmitted. When that amount has been reached, a new key is established and the information continues to be transmitted. In some instances, rekeying happens every minute and a process must be in place to communicate the new key to the sender and receiver. Some keys have been defined by the ciphers as weak keys, values which have a

high probability of being broken by an attacker.

Data Encryption Standard (DES) is a block cipher that encrypts data in 64-bit blocks and uses a 64 bit key length (every eighth bit is ignored by the encryption, and is used to provide additional functions such as parity checks). An initial permutation is performed by the cipher where the block is divided into a right and left half of 32 bits each. Then 16 key-dependent rounds of encryption are performed on each half. The resulting halves are joined and a final permutation is performed. DES uses two modes: Electronic Code Book (ECB) and Cipher Block Chaining (CBC). ECDB encrypts each block independently from the other blocks. CBC has each plaintext block authorized for encryption by the previous ciphertext block.

Triple DES (3DES) is three DES processes performed one after another. It uses an 1123-bit key and involves 48 key-dependent rounds. 3DES is considered twice as secure as normal DES. Though there are numerous modes to 3DES, the most popular involves the three encryptions to be done in series with three distinctly separate keys.

The Advance Encryption Standard (AES) was designed for symmetric cryptography. It encrypts data in 128-bit blocks with key sizes of 128, 192, and 256 bits. The number of encryption rounds is dependent on the key length: 10, 12, and 14 respectively. AES is more efficient than DES and more suited to high-speed parallel operations.

Secure Hash Algorithm (SHA-1) is the algorithm of choice for hash function cryptography. The blocks are 512 bits and use 4 rounds of 20 operations each to product a 160-bit hash value.

22.4 *Risk Assessment of a Storage System*

The first step to create a security plan for the storage system is to perform an assessment of risk to the environment. Within this assessment, any exposures to attack are determined and quantified. Countermeasures to the possible attacks are proposed and evaluated for their effectiveness. This assessment should quantify the costs of security measures; both the induction of hardware and functions to the infrastructure and the cost due to changes in usability and performance positive and negative. There is a common thought that as a storage network becomes more secure, the cost goes up and performance goes down. Therefore all costs need to be justified for the implementation of the network.

The risk assessment process has 10 steps. The amount of effort required for each step varies for different situations, but all steps apply to ensure that a complete risk assessment is effectively and efficiently finished.

- Resources to be protected are identified
- Categories of risk are identified
- Probable points of attack (vulnerabilities) are identified
- Methods of attack at each vulnerability are identified
- The potential loss of each attack method is categorized
- Threats are estimated and probabilities calculated
- Severity of risk is calculated
- A countermeasure for each attack method developed
- How much each countermeasure will reduce the severity of risk is estimated
- The cost-effectiveness estimated for each countermeasure.

Determining the resources to be protected is based on the network configuration and to what degree that protection manifests. For the most part though, there are two general resources that must be protected: the data itself as it resides on a storage device and the communication ability of the storage network. Security risks to data on storage devices are typically as unauthorized access, unwanted deletion and modification of the data, and the creation of false data. The categories of security risk on the communication ability fall into the availability to communicate, the inability to identify and handle changes to configuration, and the rerouting, corruption, and deletion of data in transit.

Most planned and existing storage area networks are contained within a data center and sometimes are found on separate infrastructure from the company LAN. Therefore the concerns for physical security are minimized as well as many points of security before the SAN is reached through the network. This does not mean that a SAN is not impervious to attack. In fact, recent analysis of storage area networks have found more points of attack then previously noticed, specifically at interface points in the storage network.

In-band interfaces are the interconnections between servers and storage devices. There are many concerns that the physical security of these connections can be maintained if those connections are of any great length. Security mechanisms, such as IPsec and Secure Sockets Layer (SSL) protocols, can be used to protect TCP/IP systems. Almost all of the elements on the storage network infrastructure have an out-of-band interface for management purposes. Typically this interface is a SNMP agent that reports management states and statistics to a console or Web site to provide real time ability to control and configure the device. Attacks through these points are mostly found to be accessible to the interface and the features behind it.

When the interface is connected to the company LAN and remote access connectivity to the interface, the accessibility to the interface can be quite simple. To counteract this, the interface can use a Virtual Private Network (VPN) protocol or IPsec. Additionally, requiring a username or password, which is different from email or network authentication, to acquire access to the interface can diminish further accessibility. Specific devices can

allow some features through the interface to have a significant impact on the entire storage network. For instance, it is possible to access its name server and create or delete entries. The effect is the disappearance of devices or the creation of devices that don't physically exist, or to alter the access control system to prevent access to other SAN devices.

Many devices on the infrastructure elements have a serial interface with a “control terminal.” This interface provides the basic configuration set up and performs tasks like updating firmware and managing logs. As another possible exposure to the network, any risk can be mitigated by disconnecting unless a maintenance procedure is required. Unfortunately, this physical interaction may not be possible especially for remote access for off-hours maintenance. Adding the proper protocols and authentication processes will limit accessibility.

Storage area network with numerous elements on the infrastructure have interelement interfaces. These interfaces handle more network traffic than most servers and storage devices and are typically utilized to configure and manage the fabric of the network. Hubs, routers, and switches can be considered interelement interfaces. The interelement link rarely incorporates confidentiality.

When most people think about attacks to the network, they immediately envision a malicious attempt to gain access to the network from the outside. Usually, the reverse is true. Most threats come from inside sources that perform a task that inadvertently attacks the network or the devices on that network. Research has discovered that most serious security threats start with data center staff. Maintenance procedures are so complex that even the best failsafe procedures will still allow mistakes to happen.

Other identified threats come from people who have access to the network through out-of-band interfaces and control terminal interfaces, people who have access to the storage network and exploit security problems, and misuse of credentials.

22.5 *SAN Security Tools and Practices*

The countermeasures for security threats are numerous and dependent on the specific devices and protocols found within the current storage area network solution. Most countermeasures focus on access control as a way to ensure that unauthorized users are able to connect to the device, the network, or the data.

The most important tool for managing security is the security policy. The implementation of the policy is done through procedures and guidelines and reinforced with security tools and sanctions. Topics like privacy, authentication, confidentiality of different types of

information, requirements for backup and restore, and required monitoring and auditing levels should have some mention in the security policy. The following items are recommended for inclusion in some way into the security policy:

- System upgrades should be tested on a nonproduction system and when ready be promptly installed on all required systems in the production environment
- The only technologies that should be installed in the storage networks have been proven with some form of reference to ensure workability
- Restrictions to the types of applications or credentials allows should be placed on the key servers whenever possible
- Perform frequent security audits as well as examining system logs to identify unusual activity
- Developing an awareness program to inform key employees of expected threats and countermeasures
- Collaborate with organizations with similar storage networks and configuration to benefit from pooling of resources and experience.

A policy is ineffective to the company if it is never fully implemented into the environment or its implementation falls short from its intended objectives. Below are some recommended best practices for security management in a storage area network:

- Ensure that all interfaces in the storage area network have been identified
- For out-of-band management and control terminal interfaces to the storage network create a separate infrastructure. If connectivity to the LAN is required, ensure that a firewall or secure router is in place as well as a dedicated remote access facility if access is required. Use all appropriate security tools, such as virtual private networks, to secure this type of infrastructure.
- For storage area network maintenance access, use dedicated user IDs. The use of strong passwords should be enforced either by policy or configuration. Separate credentials should be used for infrastructure configuration functions.
- Zones should be defines with the smallest number of components possible. Use different zones for different system loads.
- All unused ports in the infrastructure should have access controlled. Configure the infrastructure elements so that unused ports must be specifically enabled before use and to prevent automatic additions to the zone when new devices are attached.
- Install only authorized software and firmware. Do not perform such installs when the device is connected to the production storage network. When a network is required in the network, swap devices and perform on the install on an isolated network. Configure storage devices to prevent upgrades to firmware through network interfaces.
- Change all default passwords before connecting to a production storage network.

23 Service Providers

23.1 *Amazon.com*

Amazon.com was the first major web business that presented an infrastructure for a commercial service on the web. Currently they offer two services for storage management: Amazon Simple Storage Service (Amazon S3) and Amazon SimpleDB.

23.1.1 Amazon Simple Storage Service (Amazon S3)

Amazon S3 was designed specifically for designers to make web-scale computing easier. With Amazon S3, data can be stored and retrieved at any time, from any location on the web. The infrastructure utilized is the same that Amazon.com uses to maintain its web presence on the Internet.

The Amazon S3 service intentionally has a minimal feature set including the ability to read, write, and delete objects in size from 1 byte to 5 gigabytes of data. How much data is stored is unlimited. Objects are stored in a bucket and are only accessible using a unique key assigned to the developer. The physical location of the bucket will be in the United States or Europe, but is accessible from anywhere. Access controls are in place to control unauthorized access, with objects managed as private, public or with user-specific rights. To work within most development environments, the services use standard-based REST and SOAP interfaces. Protocol and functional layers can be easily added.

Amazon S3 is designed to provide to developers an online storage solution that is scalable, reliable, fast, simple, and inexpensive. The guiding principles for the Amazon S3 design include:

- Decentralizing to remove single points of failure and bottlenecks
- Autonomous designing by individual components to ensure consistency without influence from its peers
- Limited to no control over concurrency required
- Parallelism utilized to improve performance and recovery processes through the use of granular abstraction
- The service is built using blocks of components instead of a one size fits all approach without any complexity to the customer.
- Nodes on the system have identical functionality.

Amazon S3 is a pay-what-you-use service. The monthly charge is related to use of storage, data transfers and requests to the data. Pricing for storage are approximately $0.03 more in Europe than the United States and charges for requests are a fraction of a cent. Below is the pricing schedule for the United States.
$0.15 per GB for the first 50 TB storage used per month
$0.44 per GB for the next 50 TB storage used per month
$0.13 per GB for the next 400 TB storage used per month
$0.12 per GB for storage used per month over 500 TB
$0.10 per GB for all data transfers in
$0.17 per GB for the first 10 TB data transfer out per month
$0.14 per GB for the next 40 TB data transfer out per month
$0.11 per GB for the next 100 TB data transfer out per month
$0.10 per GB for data transfer out per month over 150 TB
$0.01 per 1,000 OUT, COPY, POST, or LIST requests
$0.01 per 10,000 GET and all other requests

23.1.2 Amazon SimpleDB

Amazon SimpleDB provides a platform to run queries in real time. It is designed to store small amounts of data. In concert with Amazon S3 and Amazon Elastic Compute Cloud (Amazon EC2), Amazon SimpleS3 is an easy implementation of a data warehousing solution. Using this web service, the customer can organize structures data into domains and identify the items and attributes that apply to the domain. The Amazon SimpleS3 solution will index data automatically as items are added without the need to pre-define a schema or change a schema later. Simple queries can be performed directly from the items and attributes in the domain.

Amazon SimpleDB provides quick, efficient access to database functions, such as lookup and query that are typically available through relational database cluster. Other database functions that are highly complex and often unused are eliminated. Attributes can be added to the solution as they are discovered without defining a schema, providing flexibility to customers. Applications using Amazon SimpleDB can be confident that the solution can be scaled to fit their business growth. The solution supports a maximum of 100 domains with each domain having a size limitation of 10GB. High availability service levels are in place and indexed data is stored redundantly across servers and data centers.

Amazon SimpleDB is a pay what you use service. The specific categories of the monthly charge are machine utilization, data transfer and structured data storage. Machine utilization is calculated by the amount of capacity used to complete a database request.

Data transfers only relate to transfers in and out of Amazon SimpleDB, not transfers between other Amazon Web Services which are free of charge. Data storage charges for the raw data size plus 45 bytes for each item, attribute and attribute-value pair. The base charges look like:
$0.14 for each consumed machine hour
$0.10 per GB for all data transfers in
$0.17 per GB for the first 10 TB data transfer out per month
$0.14 per GB for the next 40 TB data transfer out per month
$0.11 per GB for the next 100 TB data transfer out per month
$0.10 per GB for data transfer out per month over 150 TB

$1.50 per GB data storage per month.

23.2 *Box Enterprise*

Box.net has Internet file storage and sharing that is free for individuals, however, they have expanded to provide the same services to companies for a price with their product Box Enterprise. Their premise is to provide an avenue to support a company's demand for collaboration and file management in order to allow business workflow to be effectively and efficiently successful without the need for additional hardware or software to the company.

Their productivity tools are geared towards collaboration, allowing storage, sharing, and publishing from anywhere the user is located even from a mobile device. All employees can edit the documents stored on within the “Box” as long as they have access to the Internet. Large files can be stored and transferred simply and securely. Box Enterprise is scalable to the customer's needs, providing storage at an affordable fixed rate and allowing unlimited shared folders, modifications, and communication transmissions. The solution can allow any type of file to be stored and accessed.

Since the intent of Box Enterprise is to encourage collaboration, productivity and access are key delivery points. The solution provides a platform to manage documents through workflow to employees, clients, and partners on either side of the company's firewall securely. The customer has complete control and oversight of the data and provides a set of full-feature accounts, with sub-user capabilities to give individuals in the company the space to build web collaboration initiatives. Variable levels of access are available at the folder layer of the file system.

Box.net maintains their servers through two geographically separated data centers. They offer a 99.9% network uptime and 7X24 monitoring. Every request into the box has to pass through an audited verification code to ensure the user is authorized to perform the

requested action. All data is backed up daily and stored at a third location. Network security relies on 256-bit Secured Socket Layer (SSL) to encrypt data in transport. Indexing of public files by search engines and robots is not allowed. All files must be manually identified as public and can be changed during the life of the file and passwords are required to access the solution and parts within it.

OpenBox Services provides additional business web application capability to edit documents and spreadsheets online, send documents, accept digital signatures, collaborate on graphics, images, and spreadsheets online. File consistency is maintained by only having one set of files available and allowing file movement from one application to another. You can save directly to the Box from a desktop application as if a network drive was attached to the computer.

Pricing for Box Enterprise is $15 dollars per user. Their Starter Package for $149/month includes customer accounts for up to 10 users, 50 gigabytes of secure file storage, account setup and user training, and a dedicated Technical Support Manager.

23.3 *Nirvanix Storage Delivery Network*

Nirvanix advertises their Storage Delivery Network (SDN) as an opportunity to businesses for shorter time to market capabilities, reduced costs, and greater flexibility and control in their operations. They stand behind this by providing a global solution for policy-based cloud storage with a demand for secure data, availability and enterprise support. Nirvanix uses a clustered file system of their own design called Internet Media File System (IMFS). IMFS has all its nodes under one namespace to allow for extreme scalability and data availability. Data is intelligently stored based on geographic location. That data is replicated over multiple geographic storage nodes and that replication is run based on policies. Pricing is based on packages to allow companies to use services on an unlimited basis. They propose four reasons for using online storage: content origin storage, embedded storage for applications and devices, online archival and backup, and storage intensive Web 2.0 applications.

The handling of large stores of files and user experience is sometimes conflicting. Ideally, the user wants to have access to the content they need at any given time. This demanded content is only a small part of the total content store. This means that the majority of content is laying in wait. The ebb and flow of business impact content is managed by changing what content is in demand at any point in time. The best solution has proven to be keeping all content online, available to everyone who has access to a web browser. Companies who attempt to put their own content origin storage solution in place can spend 5 to 10 times more for hardware and software than using an online storage provider.

Applications and devices that require integrated storage, remote access and file sharing have traditionally forced customers to build massive storage centers or utilize a third party applications providing online storage. The problems in these solutions include great expense in hardware, software, space and staff for a storage center to lose control over consumer experience and even consumer relationships. The Nirvaniz Storage Delivery Network is a product that eliminates the need for great capital expenditures by providing online storage and the ability to integrate that storage with the application or device while fully controlling the consumer experience.

Studies by Nirvanix have shown that 80% of all data retained by a company is considered data that should be archived. Archive states are determined because data is infrequently accessed. Most companies have started to archive data on tape that is offline. Unfortunately, having this data offline makes it a liability to the company because it is not available for possible research or historical information. The other possibility is to allow all data to be available along with mission critical data, but this slows down backup and recovery processes. Online archive data drives up costs because of the increase in total disk storage cost, costing an average of $15 per gigabyte per year.

A highly competitive online marketplace demands attention is paid to product to market timeframes and fantastic user experiences. The ability to provide applications that can handle large loads in a cost-effective manner is key to success. Traditional web hosting packages do not always provide the storage required by the demand of the applications the company has. Web 2.0 applications are a growing technology in many companies to streamline data exchange and work flow processes. These applications allow an easier approach to the lifecycle of software development by allowing development, testing, and deployment of solution with little initial investment. After deployment, global access, availability of the data and security become major concerns. The Nirvanix offering provides the platform for Web 2.0 application support throughout the entire lifecycle.

Nirvanix SDN offers a pay as you use service, however any usage above 2 TB per month requires a contract. The pricing schedule is based on one, two, or three node SDN.
Core Services
Storage: $0.25 - $0.71 GB per month
Uploads: $0.18 - $0.90 GB per month
Downloads: $0.18 GB per month

Enhanced Services

Media Services: $1 GB processes (based on source file)
Search: $0.20 per 1000 calls

Virtual URLs: $50 setup + $15 per month

23.4 *iForum Content Organizer*

iForum provides the customer the opportunity to store their wealth of critical information within a secure, guaranteed online “vault.” The intention of their services is preservation and archival of documents. Similar to a safe deposit box in a bank, customers create a virtual vault within the iForem Content Organizer. Then they choose which file they would like to store in the vault. A one-time payment based on the size of the file ensures that perpetual security of that file.

A portion of the one-time payment creates a trust in the customer's name. The trust will continue to pay for the storage, features, and applications required to access the file stored in their vault forever. The customer is the trustee of the vault. As trustee, they can set up beneficiaries to the vault and the content held within. iForum guarantees access even if something were to happen to the company itself. If the customer needs more capacity, they can reserve more through a one-time payment.

For individuals, the use of the Content Manager is $7.99 per month. Additional one-time fees apply based on size of files.

23.5 *ElephantDrive*

For strict online storage and backup solutions, ElephantDrive plans may be the answer. With affordable pricing, a customer's access is secure and backup is automatic. Data transfers are secure through a 128-bit SSL transfer and advance 256-bit AES encryption. ElephantDrive provides differential backups. In addition, data can be accessed by groups and individuals through the use of sub-accounts.

There are two plan options for ElephantDrive: Pro and Pro Plus Edition. Below are the specifications for Pro Edition:

- Create 10 workgroup sub-accounts
- Protect up of 10 computers
- 1 TB of storage
- Automatic backups
- Protects up to 20 versions of each file
- All of this is available for $34.95 per month.

Specifications for Pro Plus Edition are:

- Create 20 workgroup or independent sub-accounts

- Protect up of 20 computers
- 2 TB of storage
- Automatic backups
- Protects unlimited versions of each file.
- All of this is available for $99.95 per month.

23.6 *Humyo.com WorkSpace*

Humyo.com provides an intelligent file management solution with their product, WorkSpace. The solution provides each person in the business with a secure file store without the need of additional software or hardware. Files put into the file store are automatically synchronized with other authorized employees, eliminating the need to use CDs, memory sticks or to email files.

Secure shared folders, added files and updates to files are available to the team and departments locally and globally as long as they have access to a web browser. This capability allows for collaboration on documents from authorized individuals on both sides of the company's firewall.

All files within WorkSpace are scanned for viruses automatically and instantly backed up using the strictest encryption algorithms. Control on access belongs to the customer and can be changes at any time.

There are two plans: WorkSpace S and WorkSpace L.

WorkSpace S provides 100 GB to 4 users, compared to WorkSpace L solution which provides 250 GB to 10 users. Each solution includes shared folders, 245-bit SSL encryption to provide security and integrity, and virus scanning. The prices for these plans are $19.99 and $38.99 respectively.

23.7 *Cloud Files*

Mosso provides online storage called Cloud Files for files and applications. In addition, the product adds the enhancement of Limelight Networks. Cloud Files is scalable and dynamic, allowing customers to pay only what they use. Files can be managed through Mosso's online control panel or through their REST API. Files up to 5GB can be stored on the solution.

Using Cloud Files, is as simple as uploading files. To ensure that the file is available at all

time, automatic replication is done across multiple zones. These zones are physically separate with different power sources, and separate Internet backbone providers.

With LimeLight Networks, customers can distribute content across the globe using the same CDN technology major media sites use. This distribution allows content to be cached on local servers to speed up on delivery and performance.

The pricing for Cloud Files is:
Storage
$0.15 per GB for first 30 TB
$0.14 per GB for 30 TB – 300 TB
$0.13 per GB for over 300 TB

Bandwidth

$0.22 per GB for first 5 TB
$0.20 per GB for 5 TB – 10 TB
$0.18 per GB for 10 TB – 50 TB
$0.15 per GB for over 50 TB

Requests

PUT, POST, LIST
$0.01 per 500 for files under 100 KB in size
Free for files over100 KB in size
HEAD, GET, DELETE

Free for any size file.

23.8 *ParaScale*

ParaScale is not an Internet cloud storage provider. It is a software solution for businesses to create, manage, and power their own cloud storage solution. It runs on multiple Linux servers with standard configurations and clusters storage devices together to create a single mega-repository. The storage cloud created is highly scalable, self-managing, with massive capacity. No customer or dedicated hardware is required. Existing networking interconnections are leverage as long as they use the IP protocol. ParaScale Cloud Storage (PCS) creates a single namespace.

Adding capacity is simple with ParaScale. Storage from each node is automatically pooled together increasing the total capacity for the entire cloud. Since there is one single pool of capacity, there is no immediate need to buy additional storage to add for a single system or application. If additional storage is required, ParaScale can recognize a device from any manufacturer and handle any unique configurations. The storage device can be installed

without disruption to ParaScale operations. Simply install and provision the disk through an administrative interface. In the event, a storage nodes needs to be removed, the system will automatically move the data to a different node.

Adding, upgrading, and removing storage nodes can be done online and does not require any disruption to service. Adding storage nodes to a ParaScale cloud increases performance and capacity, because the software can manage load sharing capabilities. Each node is independent. So when a request for storage comes in, ParaScale finds the most appropriate storage node to utilize for fulfilling the request. In addition, hot file systems can be identified and replicated to multiple nodes providing parallel performance across the board.

ParaScale's Policy Engine simplifies management of the environment through synchronization. As policies are managed by the engine, objectives like disk availability, read bandwidth and access latency are met. Policies managed by the engine include data redundancy, automatic configuration, automatic capacity balancing, and self healing activities.

Any system that can view the namespace has access to the files on the cloud. Access to those files is contingent to control lists and file read/write locking.

23.9 *Cleversafe*

Cleversafe is a technology storage provider with three focus areas: digital content storage, archive storage, and data storage. Their dsNet is geared towards storing large fixed-content digital objects like video, audio, pictures, images, engineering documents, and the like. Their solution requires less space than traditional storage solution while providing the scalability, security, and longevity of the object required by the company.

Their archive storage solution is straightforward. No more concerns about losing data because of server failure, malicious attack, administrator negligence, or site disaster. Storing archived data online removes the pressure, burden, and cost from the company.

Recognizing that data storage solutions can lock a company into a particular architecture or vendor, they have created industry-specific solutions for data and applications. Currently, they have specialized storage solutions for Media Storage/Web 2.0, Telecommunications, Science/Energy, Healthcare, and Government.

23.10 *Simpana*

CommVault provides data protection, archive, replication, resource management, and search capabilities in a single software product: Simpana. The focus here is data management whether the data is located online, archive, and backup. Using singular software with a set of application modules provides efficiencies to the user according to CommVault:

- Controlling the cost and growth of storage
- Managing capital expenditures
- Faster backup and recovery
- Easy data archive
- Better utilization of storage and networks
- Reduce effort and cost
- Maximize capacity
- Easier planning.

Simpana provides data protection through reliable, efficient backup and recovery for all files, applications, and database raw data. Tasks are automated using policies, unifying the entire backup and restore process and reducing the burden on administrators. Easily accessible reports can assist administrators in finding potential problems.

Companies can use Simpana to archive data from file systems, NAS, Microsoft Exchange, and IBM Lutos domino messaging, as well as Microsoft SharePoint systems. Archiving provides a point to control data growth, decrease storage spending, and enables compliance.

Replication through Simpana protects critical applications and remote office data by creating snapshots of data and replicating across multiple servers.

Using a web interface, Simpana provides an intuitive, direct access to desktop data and self-serve recovery of backup and archive data.

Simpana's resource management enables the customer to view historical and trend analysis reporting for all data. Data monitoring and cost-accounting for all divisions, geographies, and applications are enabled; as well as automatic delivery of performance summaries.

24 Service Management Processes

There are a number of service management processes from the ITIL framework that can play a role in Platform and Storage Management development, support and delivery. This chapter will explore some of those key processes. You will be familiar with some of these processes from the Cloud Computing Foundation program. The most relevant processes for Platform and Storage Management have been selected to provide a refined overview.

24.1 *Incident Management*

Objective: To restore normal service operation as quickly as possible and minimize the adverse impact on business operations, thus ensuring that the best possible levels of service quality and availability are maintained.

In ITIL terminology, an 'incident' is defined as an unplanned interruption to an IT service or reduction in the quality of an IT service. Failure of a configuration item that has not yet impacted service is also an incident. Incident Management is the process for dealing with all incidents; this can include failures, questions or queries reported by the users (usually via a telephone call to the Service Desk), by technical staff, or automatically detected and reported by event monitoring tools.

The value of Incident Management includes:

- The ability to detect and resolve incidents, which results in lower downtime to the business, which in turn means higher availability of the service. This means that the business is able to exploit the functionality of the service as designed.
- The ability to align IT activity to real-time business priorities. This is because Incident Management includes the capability to identify business priorities and dynamically allocate resources as necessary.
- The ability to identify potential improvements to services. This happens as a result of understanding what constitutes an incident and also from being in contact with the activities of business operational staff.
- The Service Desk can, during its handling of incidents, identify additional service or training requirements found in IT or the business.

Incident Management is highly visible to the business, and it is therefore easier to demonstrate its value than most areas in Service Operation. For this reason, Incident

Management is often one of the first processes to be implemented in Service Management projects. The added benefit of doing this is that Incident Management can be used to highlight other areas that need attention – thereby providing a justification for expenditure in implementing other processes.

Timescales

Timescales must be agreed for all incident-handling stages (these will differ depending upon the priority level of the incident) – based upon the overall incident response and resolution targets with the SLAs – and captured as targets within OLAs and Underpinning Contracts (UCs). All support groups should be made fully aware of these timescales. Service Management tools should be used to automate timescales and escalate the incident as required based on pre-defined rules.

Many incidents are not new – they involve dealing with something that has happened before and may well happen again. For this reason, many organizations will find it helpful to pre-define ‘standard’ Incident Models – and apply them to appropriate incidents when they occur.

An Incident Model is a way of pre-defining the steps that should be taken to handle a process (in this case a process for dealing with a particular type of incident) in an agreed way. Support tools can then be used to manage the required process. This will ensure that ‘standard’ incidents are handled in a pre-defined path and within pre-defined timescales.

Incidents which would require specialized handling can be treated in this way (for example, security-related incidents can be routed to Information Security Management and capacity- or performance-related incidents that would be routed to Capacity Management).

The Incident Model should include:

- The steps that should be taken to handle the incident
- The chronological order these steps should be taken in, with any dependences or co-processing defined
- Responsibilities: who should do what
- Timescales and thresholds for completion of the actions
- Escalation procedures; who should be contacted and when
- Any necessary evidence-preservation activities (particularly relevant for security- and capacity-related incidents).

The models should be input to the incident-handling support tools in use and the tools should then automate the handling, management and escalation of the process.

Major Incidents
A separate procedure, with shorter timescales and greater urgency, must be used for 'major' incidents. A definition of what constitutes a major incident must be agreed and ideally mapped on to the overall incident prioritization system – such that they will be dealt with through the major incident process.

Note: People sometimes use loose terminology and/or confuse major incident with a problem. In reality, an incident remains an incident forever – it may grow in impact or priority to become a major incident, but an incident never 'becomes' a problem. A problem is the underlying cause of one or more incidents and remains a separate entity always!

Some lower-priority incidents may also have to be handled through this procedure – due to potential business impact – and some major incidents may not need to be handled in this way if the cause and resolutions are obvious and the normal incident process can easily cope within agreed target resolution times – provided the impact remains low!

Where necessary, the major incident procedure should include the dynamic establishment of a separate major incident team under the direct leadership of the Incident Manager, formulated to concentrate on this incident alone to ensure that adequate resources and focus are provided to finding a swift resolution. If the Service Desk Manager is also fulfilling the role of Incident Manager (say in a small organization), then a separate person may need to be designated to lead the major incident investigation team – so as to avoid conflict of time or priorities - but should ultimately report back to the Incident Manager.

If the cause of the incident needs to be investigated at the same time, then the Problem Manager would be involved as well but the Incident Manager must ensure that the service restoration and underlying cause are kept separate. Throughout, the Service Desk would ensure that all activities are recorded and users are kept fully informed of progress.

24.1.1 Incident Management relating to Platform and Storage Management

Any organization making use of a cloud-based or private Platform as a Service environment should have capabilities that allow for:

- Users to log incidents (or service requests) relating to their use of IT services. Where users access these services via a web interface it also allows for support functionality to be built directly in the user interface.
- Categorization models so that incidents can be easily identified and tracked in terms of their nature and origin (i.e. an incident relating to a payroll service running via a PaaS environment).

- Accurate details of the past, current and future configuration of IT services and the PaaS environment to assist with the diagnosis of incidents and to identify the potential impact of disruptions.
- Knowledge transfer to occur between development and support teams, so that the capabilities for development and operation of applications in the PaaS environment is continually enhanced and optimized for value.
- Timely and accurate escalation and routing, especially when reported incidents reveal potential failures in the central infrastructure hosting the affected services.

While the major role of Incident Management is to provide a timely resolution to disruptions and to minimize the associated business impact, it also plays an important role feeding trends and issues to Problem Management. As a result, particular care should be taken in defining the category types for incidents to allow for accurate trend analysis and problem matching to occur.

Storage Management platforms will have similar requirements for Incident Management as those described for PaaS above. Detection and diagnosis of incidents relating to the storage platform should be somewhat more straightforward though, compared with the relative complexity of infrastructure, systems and components involved with a PaaS environment. To assist with the proactive communication of disruptions, thought should be given to the communication requirements between the IT service provider and the storage platform supplier, ensuring that all outages are identified, communicated and where possible, planned for to minimize their impact on the user community. Accurate configuration models and technical service catalogues will also help in the identification of the potential cause or impact of incidents in relation to the storage platforms.

When utilizing these environments from an external supplier, consideration should also be given as to the geographic and time zone differences between the supplier and service provider. In some cases, extended support agreements may need to be developed at significant additional costs, consequently reducing the savings that may have otherwise been gained by the transformation in architecture design.

24.2 *Change Management*

Objective: To ensure all changes are assessed, approved, implemented and reviewed in a controlled manner.

The ability to control and manage changes to defined IT services and their supporting elements is viewed as fundamental to quality service management. When reviewing the typical strategic objectives defined for an IT service provider, most of these are underpinned by the requirement of effective change control. These include strategies

focusing on time-to-market, increased market share or high availability and security platforms, all of which require a controlled process by which to assess, control and manage changes with varying levels of rigor.

Changes arise for a number of reasons, including:

- Requests of the business or customers, seeking to improve services, reduce costs or increasing ease and effectiveness of delivery and support
- From internal IT groups looking to proactively improve services or to resolve errors and correct service disruption.

The process of Change Management typically exists in order to:

- Optimize risk exposure (defined from both business and IT perspectives)
- Minimize the severity of any impact and disruption
- Deliver successful changes at the first attempt.

To deliver these benefits it is important to consider the diverse types of changes that will be assessed and how a balance can be maintained in regards to the varying needs and potential impacts of changes. In light of this, it is important to interpret the following Change Management guidance with the understanding that is intended to be scaled to suit the organization and the size, complexity and risk of changes being assessed.

To ensure that standardized methods and procedures are used for controlled, efficient and prompt handling of all changes, in order to minimize the impact of change-related incidents upon service quality, and consequently to improve the day-to-day operations of the organization.

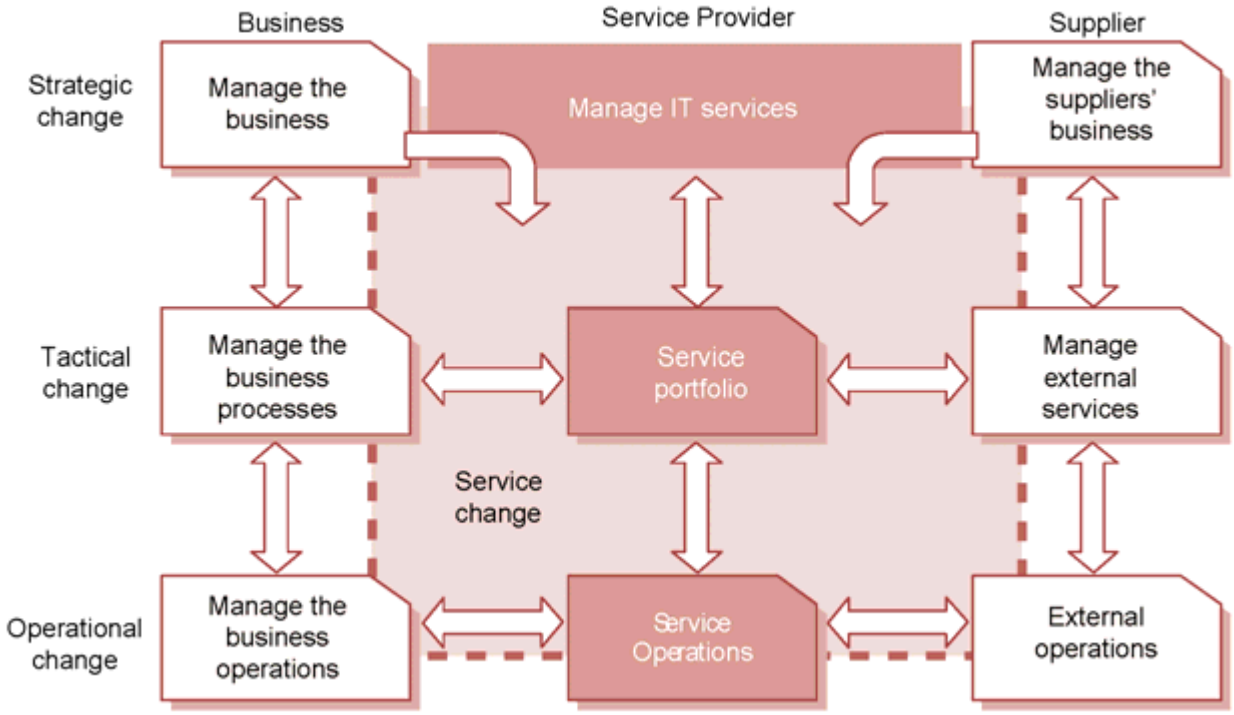
Remember: not every change is an improvement, but every improvement is a change!

Change Management's purpose is also to ensure that:

- All changes to service assets and configuration items (CIs) are recorded in the Configuration Management Systems (CMS)
- Overall business risk is optimized.

The term change is often defined in varying ways, however the best definition of a service change is: "Any alteration in the state of a Configuration Item (CI). This includes the addition, modification or removal of approved, supported or baselined hardware, network, software, application, environment, system, desktop build or associated documentation."

It is important, however, that every organization defines those changes which lie outside the scope of their service change process (such as operational or business process and policy changes).



Scope of Change Management for IT Services

The figure above demonstrates the typical scope of the Change Management process for an IT Service Provider and how it interfaces with the business and suppliers at strategic, tactical and operational levels. As discussed in 4.1.4, Service Portfolios provide the clear definition of all planned, current and retired services.

Designing and Planning

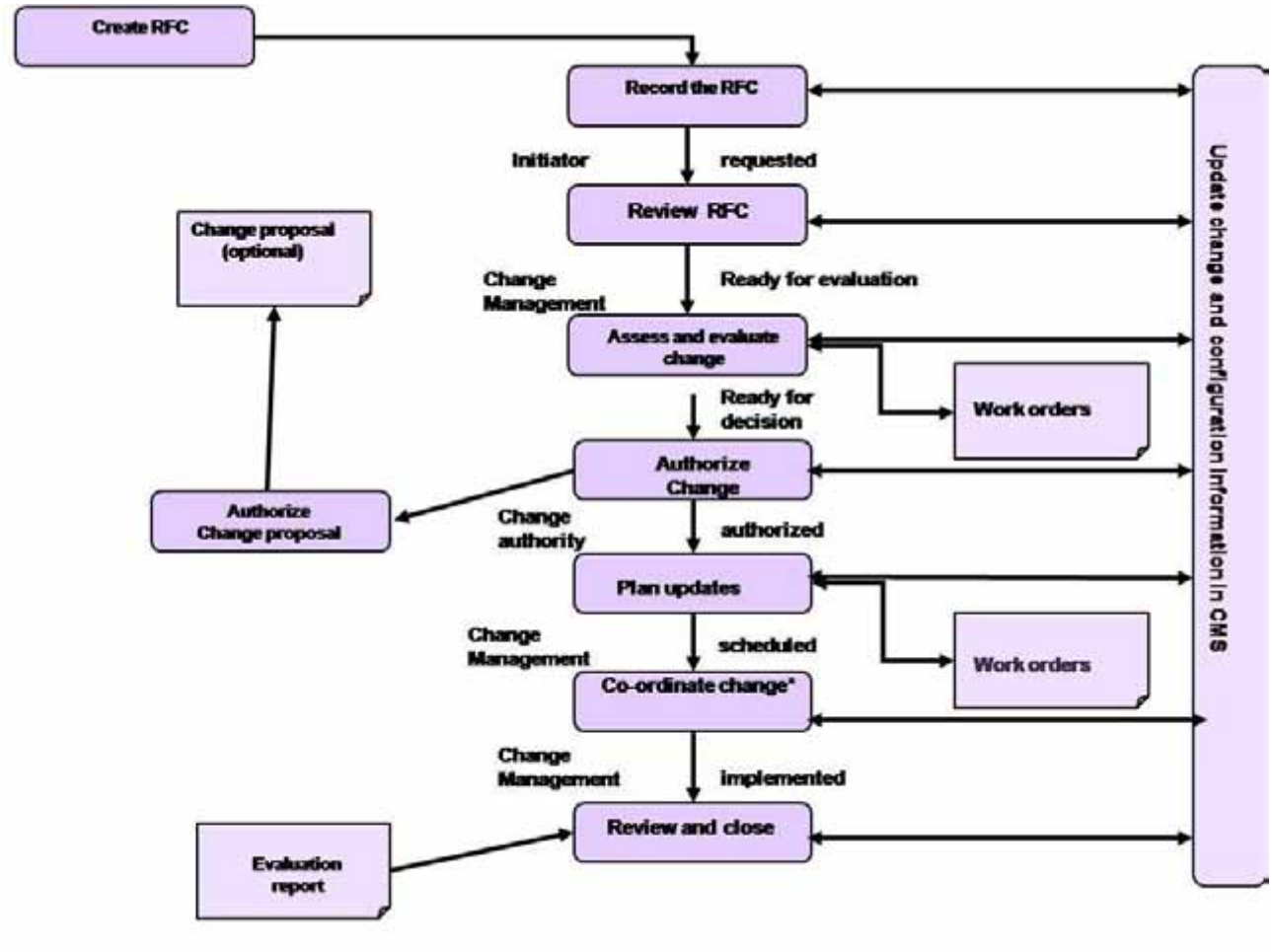
It is generally advised that the Change Management process should be planned in conjunction with Release & Deployment, and Service Asset & Configuration Management. These processes together will help in the evaluation of impact, needs, timings and overall risk for changes being assessed.

The checklist of typical requirements when designing the Change Management process includes:

- Regulatory, policy or other compliance requirements
- Documentation requirements
- Identification of impact, urgency and priority codes for changes
- Roles and responsibilities involved
- Procedures required
- Interfaces to other Service Management processes (e.g. Problem Management)
- Toolset requirements to support Change Management
- Configuration Management interfaces.

Change Management Activities

The following diagram represents the typical activities involved for normal changes that have been identified. The actual steps and procedures need to be further refined depending on any specific Change Models that have been created.



The Activities of Change Management

Overview of Important Steps:

- The RFC is recorded.
- Review the RFC for classification and prioritization.
- Assess and evaluate the change – may require involvement of CAB or ECAB.
- Authorization or rejection of the change.
- The change is scheduled.
- Work orders are issued for the build of the change (but carried out by other groups).
- Change Management coordinates the work performed.
- The change is reviewed.
- The change is closed.

1. The RFC is recorded

The change is raised by a request from the initiator. The level of information recorded for a change depends largely on the size and impact of the change. Some information is recorded initially and some information updated as the change document progresses through its lifecycle. This may be recorded directly on the RFC form and details of the change and actions may be recorded in other documents and referenced from the RFC such as business cases.

For a major change with significant organizational and/or financial applications, a change proposal may be required, which will contain a full description of the change together with a business and financial justification for the proposed change. The change proposal will include sign off by appropriate levels of business management.

2. Review the RFC for classification and prioritization

To ensure Change Management is using an appropriate level of control based on factors of risk, cost and complexity, an initial review should act as a filtering mechanism to apply the correct Change Model (classification), identify the relative priority of the change, and to ensure that the required details are supplied. Procedures should stipulate that, as changes are logged, Change Management reviews each change request and return any that are:

- Totally impractical
- Repeats of earlier RFCs
- Incomplete submissions.

These requests will be returned to the initiator, together with brief details of the reason for the rejection, and the log should record this fact. There should be an opportunity to appeal, via normal management channels, and should be incorporated within the procedures.

3. Assess and evaluate the change

All changes will be assessed for their relative potential impact, risk and resource requirements. Depending on the Change Model that has been applied, this assessment may require involvement from:

- The Change Manager and Change Owner for local authorization
- The Change Advisory Board (representing all key stakeholders)
- The IT Management (Steering) Board
- Business Executive Board.

The scope of potential impacts on services for failed changes is wide, so the assessment needs to identify potential for:

- Impact on customer's business operation
- Effect on SLAs, baselines, service models, security etc.
- Impact on other services
- Impact on non-IT infrastructures
- Effect of not implementing
- Cost and staffing implications
- Current Change Schedule
- Ongoing resources required after implementation
- Impact on continuity plan, capacity plan, security plan, test environments, and any Service Operation practices.

The following table describes the type of hierarchical structures that may be used for different levels of change authorization. A degree of delegated authority may also exist within an authorization level. Formal authorization is obtained for each change from a change authority that may be a role, person or a group of people. The levels of authorization for a change should be judged by:

- Type
- Size
- Risk
- Financial implications
- Scope.

Level	Change Authority	Potential Impact/Risk
1	Business Executive Board.	High cost/risk change - Executive decision
2	The IT Management (Steering) Board	Change impacts multiple services/ organizational divisions
3	Change Advisory Board (CAB) or Emergency CAB (ECAB)	Change impacts only local/ service group
4	Change Manager	Change to a specific component of an IT Service
5	Local Authorization	Standard Change

The 7 Rs of Change Management provides a set of guiding questions that need to be answered as part of the overall assessment for changes. These questions are:

- Who RAISED the change?
- What is the REASON for the change?
- What is the RETURN required from the change?
- What are the RISKS involved in the change?
- What RESOURCES are required to deliver the change?
- Who is RESPONSIBLE for the build, test and implementation of the change?
- What is the RELATIONSHIP between this change and other changes?

4. Authorization or rejection of the change

While the responsibility for authorization of changes lies with the Change Manager, they in turn will ensure they have the approval of three main areas.

- Financial Approval - What's it going to cost? And what's the cost of not doing it?
- Business Approval - What are the consequences to the business? And not doing it?
- Technology Approval - What are the consequences to the infrastructure? And not doing it?

When authorizing changes, it is important to consider both the implications of performing the change, as well as the impacts of NOT implementing the change. This also requires empowering the Change Manager with an appropriate level of authority, as their primary role is to protect the integrity of the IT infrastructure and the services provided to customers.

5. The Change is scheduled

The assessment of the change will have provided an estimation of the resource requirements for delivering a successful change. Change Management will need to coordinate with Release and Deployment Management so that any activities required for the build, test or implementation of the change will be scheduled when resources are available and when they least impact on live services or business critical times.

The timing of events and eventual implementation will be communicated via the Change Schedule, and visible to the appropriate IT staff members, customers and end-users. Any service disruption is documented in with the Projected Service Outage (PSO). This details any revised Service Level Agreement and service availability targets because of the events in the Change Schedule, in addition to any planned downtime from other causes such as planned maintenance and data backups.

6. Work orders are issued for the build of the change (but carried out by other groups)

Change Management will coordinate with Release and Deployment to identify the groups or individuals responsible for the implementation of the various elements making up the change. This will be greatly influenced by the availability of staff and any Release policies defining how and when changes are released to the live environment.

7. Change Management coordinates the work performed

Change Management plays a co-ordination role as implementation is the responsibility of others (under the direction of Release and Deployment management or from Project Management groups).

This is an oversight role to ensure that all changes that can be are thoroughly tested. Special care needs to be taken during implementation in all cases involving changes that have not been fully tested.

Remediation Planning is a critical element during the coordination of changes. Ideally, no change should be approved without having explicitly addressed the question of what to do if it is not successful. There should be a back-out plan that will restore the organization to its initial situation, through reloading a baselined set of Configuration Items.

Only by considering what remediation options are available before instigating a change, and by establishing that the remediation is viable, can the risk of the proposed change be determined and the appropriate actions taken.

8. The change is reviewed

On completion of the change, the results should be reported for evaluation to those responsible for managing changes, and then presented as a completed change for stakeholder agreement. Major changes will require more customer and stakeholder input throughout the entire process.

The review should confirm that the change has met the defined objectives, the initiator and stakeholders are happy with the results and there have been no unexpected side-effects. Lessons learned should be embedded into future changes as part of continuous improvement. This includes whether the request should be developed as a standard change or whether another Change Model is more appropriate for future management of similar requests.

Two types of reviews are performed for normal changes:

- The review of a service change – immediately visible to the customer and scheduled for discussion at the next service level management review meeting
- An infrastructure change – concerned with how IT delivers rather that what IT delivers, which will be (almost) invisible to the customer.

Change Management must review new or changed services after a predefined period has elapsed. This process will involve Change Advisory Board (CAB) members, since change reviews are a standard CAB agenda item. When a change has not achieved its objectives, Change Management (or CAB) will decide what follow-up action is required.

9. The change is closed

If the change review was satisfactory or the original change is abandoned (e.g. the needs for the change is decreased or disappears) the RFC should be formally closed in the logging system. These records will be kept for a period of time based on business, compliance, archiving or other policy requirements that have been defined.

24.2.1 Change Management relating to Platform and Storage Management

In particular relevance to PaaS and Storage Management, the process of Change Management should provide specific capabilities focused on:

- Ensuring there is a suitable business case for the change to progress through each of its major stages
- Identifying that adequate resources (financial, personnel and other) is available for the change to be developed and implemented

- Ensuring that the interfaces and dependencies that may exist within the PaaS or storage environments are considered for their requirements or potential conflict
- Providing the appropriate authorization for the change to progress
- Verifying that the support and operational teams are ready for the change to be implemented
- Changes made by external suppliers are identified and managed appropriately.

Typical areas of weakness are usually the result of inadequate communication between the involved stakeholders (including the external suppliers involved in managing the 'cloud'), lack of clarity regarding the actual state and configuration of services and the infrastructure, and staff bypassing procedures due to actual or perceived bottlenecks caused by the Change Management process.

Ideally, constant engagement between the service provider and the external supplier (managing the PaaS or storage environments) will identify upcoming changes and how the two parties can work effectively together. Most suppliers will also provide some form of web-based dashboard that can easily be referenced for verifying the current state of services and to identify any planned changes or maintenance. With PaaS environments, there may also be the possibility for changes being introduced by the supplier to be grouped and implemented together, allowing the testing and validation activities performed by the service provider and their users to be coordinated and optimized.

The outcome of change reviews will also identify any potential shortcomings in the current method of service delivery or the processes employed within an IT Service Management (ITSM) framework. Reoccurring changes employed in the PaaS and storage environments can also be assessed for their success and whether they should be defined as a standard change for future implementations.

Successfully managing changes to IT services, regardless of their architecture or method of delivery, very much depends on the quality of input from Service Asset & Configuration Management, Financial Management and other key stakeholder areas from IT and the business. As a result, organizations looking to utilize cloud-based technologies should evaluate their current capabilities in regards to ITSM, as some weaknesses in this regard may only be discovered once major sections of the IT infrastructure is housed and managed remotely.

24.3 *Capacity Management*

Objective: The goal of the Capacity Management process is to ensure that cost-justifiable IT capacity in all areas of IT always exists and is matched to the current and future agreed needs of the business, in a timely manner.

Capacity Management is an extremely technical, complex and demanding process, and in order to achieve results, it requires three supporting sub-processes. One of the key activities of Capacity Management is to produce a plan that documents the current levels of resource utilization and service performance and, after consideration of the Service Strategy and plans, forecasts the future requirements for new IT resources to support the IT services that underpin the business activities. The plan should indicate clearly any assumptions made. It should also include any recommendations quantified in terms of resource required, cost, benefits, impact, etc.
The production and maintenance of a Capacity Plan should occur at pre-defined intervals. It is, essentially, an investment plan and should therefore be published annually, in line with the business or budget lifecycle, and completed before the start of negotiations on future budgets. A quarterly re-issue of the updated plan may be necessary to take into account changes in service plans, to report on the accuracy of forecasts and to make or refine recommendations. This takes extra effort but, if it is regularly updated, the Capacity Plan is more likely to be accurate and to reflect the changing business need.

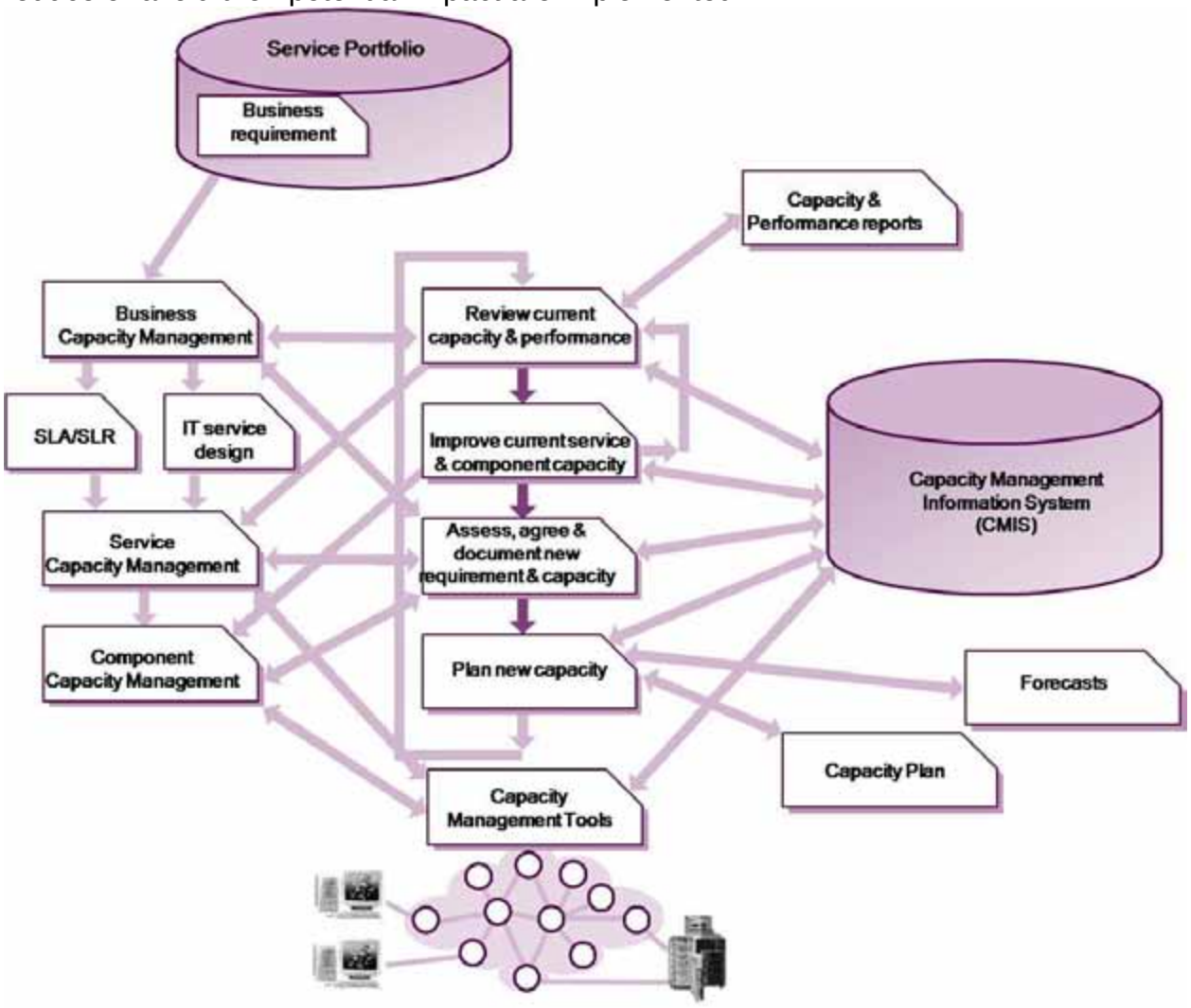
Business Capacity Management
This sub-process translates business needs and plans into requirements for service and IT infrastructure, ensuring that the future business requirements for IT services are quantified, designed, planned and implemented in a timely fashion. This can be achieved by using the existing data on the current resource utilization by the various services and resources to trend, forecast, model or predict future requirements. These future requirements come from the Service Strategy and Service Portfolio detailing new processes and service requirements, changes, improvements, and also the growth in the existing services.

Service Capacity Management
The focus of this sub-process is the management, control and prediction of the end-to-end performance and capacity of the live, operational IT service usage and workloads. It ensures that the performance of all services, as detailed in the service targets within SLAs and SLRs, is monitored and measured, and the collected data is recorded and analyzed and reported. Wherever necessary, proactive and reactive action should be instigated, to ensure that the performance of all services meets their agreed business targets. This is performed by staff with knowledge of all the areas of technology used in the delivery of

end-to-end service, and often involves seeking advice from the specialists involved in Component Capacity Management. Wherever possible, automated thresholds should be used to manage all operational services, to ensure that situations where service targets are breached or threatened are rapidly identified and cost-effective actions to reduce or avoid their potential impact implemented.

Component Capacity Management

The focus in this sub-process is the management, control and prediction of the performance, utilization and capacity of individual IT technology components. It ensures that all components within the IT infrastructure that have finite resources are monitored and measured, and that the collected data is recorded, analyzed and reported. Again, wherever possible, automated thresholds should be implemented to manage all components, to ensure that situations where service targets are breached or threatened by component usage or performance are rapidly identified, and cost-effective actions to reduce or avoid their potential impact are implemented.



Capacity Management sub-processes diagram

24.3.1 Capacity Management relating to Platform and Storage Management

One of the greatest reasons for moving towards the use of cloud-based technologies and architectures is in the enhanced capacity and performance that can be delivered in a cost-effective manner. For small organizations, delivering a sustained level of capacity and performance can be quite difficult when the demand for IT services and computing power is increasingly dynamic, without any clear and identifiable patterns. This scenario is often seen on the Internet, where many organizations that have previously hosted and managed their own websites and web platforms are increasingly turning to cloud computing technologies instead. When web traffic can easily spike 500% over normal use, why not offload the work to an external supplier such as Amazon, who can meet any level of demand that might be put on the service?

In the context of Capacity Management though, there will still be a number of considerations in the implementation and use of cloud-based PaaS and storage environments:

- Are there any capacity or performance limiting components of the infrastructure that will still cause bottlenecks? A common example of this is Internet bandwidth, which may negate the performance and capacity gains that would have otherwise been achieved by migrating an IT service to an external provider.
- What are the charging mechanisms used by the external supplier? While they may start off cheaply, there may be expensive charging once certain quotas or limits are reached.
- If the number of users or demand for IT service drops, will the use of the PaaS and storage solution still be cost-effective?
- Are there are other planned projects that where there are integration concerns?
- Do we have the support of the various teams responsible for managing each of the infrastructure areas?

After considering the above questions, many organizations decide to keep their existing internally housed infrastructure models rather than risk not achieving a suitable return on investment or severely impacting the quality of IT services being delivered.

24.4 *Availability Management*

Objectives: The goal of Availability Management process is to ensure that the level of service availability delivered in all services is matched to or exceeds the current and future agreed needs of the business, in a cost-effective manner.

The Availability Management process is continually trying to ensure that all operational services meet their agreed availability targets, and that new or changed services are designed appropriately to meet their intended targets, without compromising the performance of existing services. In order to achieve this, Availability Management should perform the reactive and proactive activities.
The reactive activities of Availability Management consist of monitoring, measuring, analyzing, reporting and reviewing all aspects of component and service availability. This is to ensure that all agreed service targets are measured and achieved. Wherever deviations or breaches are detected, these are investigated and remedial action instigated. Most of these activities are conducted within the Operations stage of the lifecycle and are linked into the monitoring and control activities: Event and Incident Management processes.

The proactive activities consist of producing recommendations, plans and documents on design guidelines and criteria for new and changed services, and the continual improvement of service and reduction of risk in existing services wherever it can be cost-justified. These are key aspects to be considered within Service Design activities.

An effective Availability Management process, consisting of both the reactive and proactive activities, can make a big difference and will be recognized as such by the business, if the deployment of Availability Management within an IT organization has a strong emphasis on the needs of the business and customers. To reinforce this emphasis, there are several guiding principles that should underpin the Availability Management process and its focus:

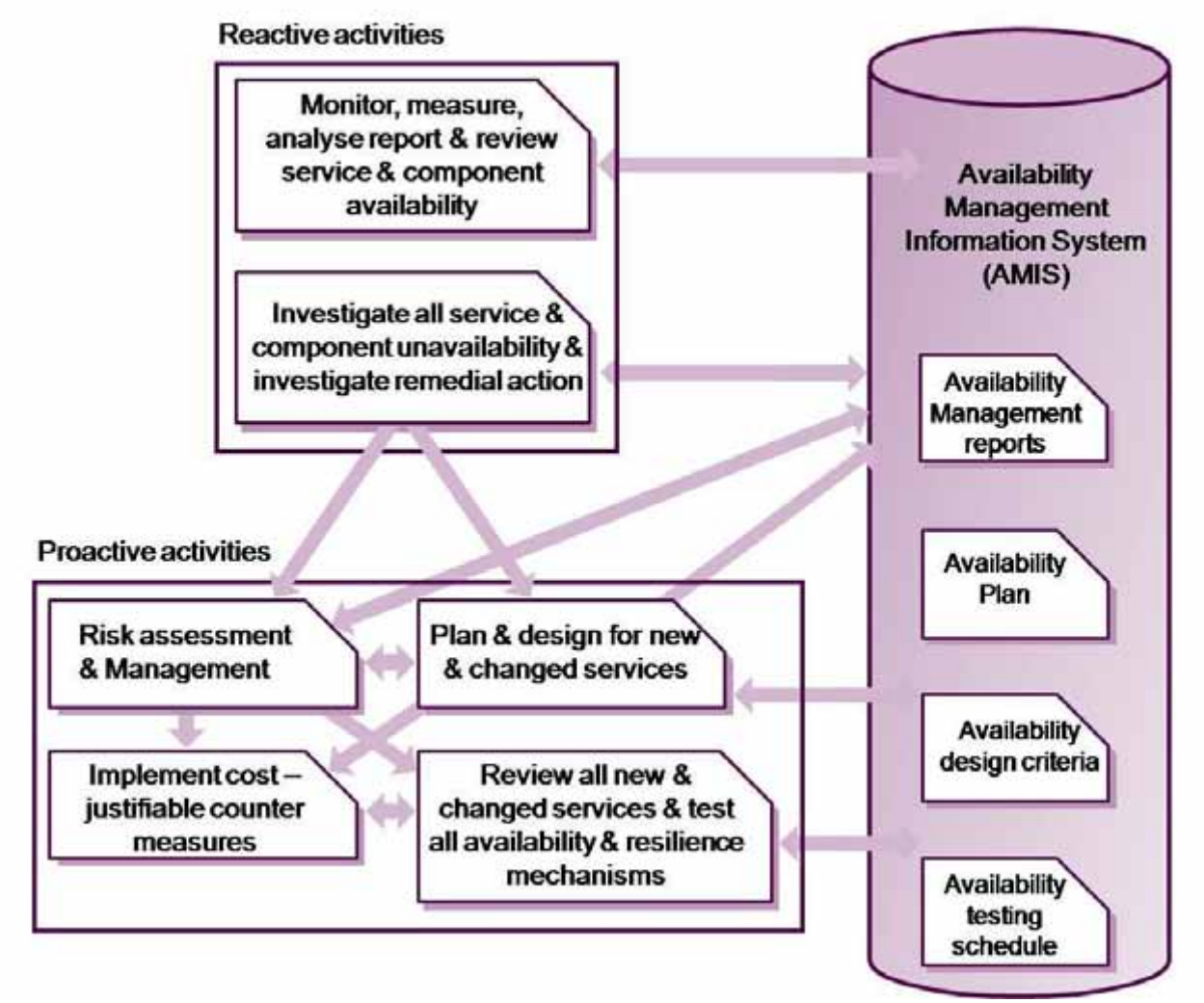
- Service availability is at the core of customer satisfaction and business success: there is a direct correlation in most organizations between the service availability and customer and user satisfaction, where poor service performance is defined as being unavailable.
- Recognizing that when services fail, it is still possible to achieve business, customer and user satisfaction and recognition: the way a service provider reacts in a failure situation has a major influence on customer and user perception and expectation.
- Improving availability can only begin after understanding how the IT services support the operation of the business.

- Service availability is only as good as the weakest link on the chain: it can be greatly increased by the elimination of Single Points of Failure (SPoFs) or an unreliable or weak component.
- Availability is not just a reactive process; the more proactive the process, the better service availability will be. Availability should not purely react to service and component failure. The more events and failures are predicted, pre-empted and prevented, the higher the level of service availability.
- It is cheaper to design the right level of service availability into a service from the start rather than try and 'bolt it on' subsequently. Adding resilience into a service or component is invariably more expensive than designing it in from the start. Also, once a service gets a name for unreliability, it becomes very difficult to change the image. Resilience is a key consideration if ITSCM, and this should be considered at the same time.

The scope of Availability Management covers the design, implementation, measurement and management of IT service and infrastructure availability.

The Availability Management process has two key elements:

- Reactive activities: the reactive aspect of Availability Management involves the monitoring, measuring, analysis and management of all events, incidents and problems involving unavailability. These activities are principally involved with within operational roles.
- Proactive activities: the proactive activities of Availability Management involve the proactive planning, design and improvement of availability. These activities are principally involved within design and planning roles.



The Availability Management process diagram

24.4.1 Availability Management relating to Platform and Storage Management

A genuine advantage of utilizing cloud-based PaaS and storage environments is the availability provided by the supplier that many organizations cannot cost-effectively achieve. As an example of a service availability statement taken from Amazon Web Services (AWS) for their cloud-based storage solution:

AWS will use commercially reasonable efforts to make Amazon S3 available with a Monthly Uptime Percentage of at least 99.9% during any monthly billing cycle (the “Service Commitment”).

The common challenge faced by Availability Management however when utilizing any form of cloud-computing is the identification and mitigation of any single points of failure (SPOF) or other vulnerable areas of the IT infrastructure. Whether it is a network segment, firewall, DNS server or the data centre itself, a failure in any of these can potentially disrupt the provision of IT services to the entire user community. If there are stability issues in the general IT infrastructure, the high availability provided by the cloud is wasted and makes little positive difference from the user and business perspective.

Techniques used by Availability Management to assist in identifying weak or vulnerable areas include:

- SPOF analysis
- Component Failure Impact Analysis (CFIA)
- Risk Analysis and Management
- Analysis of the Expanded Incident Lifecycle
- Fault Tree Analysis
- Service Outage Analysis.

Also of concern (but more aligned to Information Security Management) is where the data will actually be housed. While availability may be improved by having multiple redundant storage locations, there may be local government policies (such as the US Patriot Act) that may be a concern to the organization in terms of the privacy of data and information being managed.

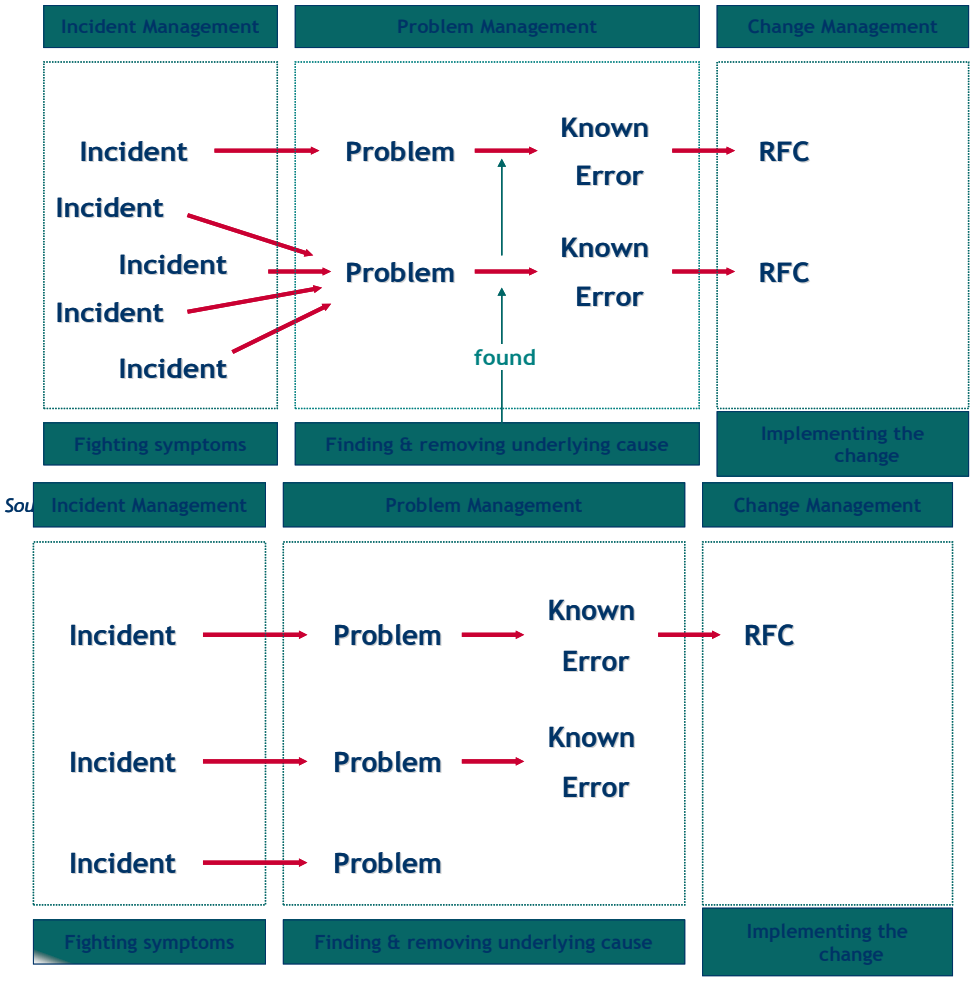
24.5 *Problem Management*

Objective: To minimize the adverse impact of Incidents and Problems on the business that are caused by errors within the IT infrastructure, and to prevent the recurrence of Incidents related to these errors.

Terminology	Explanations
Problem:	Unknown underlying cause of one or more Incidents. (The investigation)
Known Error:	Known underlying cause. Successful diagnosis of the root cause of a Problem, and workaround or permanent solution has been identified.
KEDB:	Known Error Database, where Known Errors and their documented workarounds are maintained. This database is owned by Problem Management.
Workaround:	The documented technique in which to provide the user with the required functionality, by either alternate means or by carrying out some action required to restore service.

Problem management has interfaces with the change management process for managing the required changes. The incident management process and all other Service Management processes retrieve up-to-date information from the problem management process. The explicitly required documents in the problem management process are problem records and records of identified actions for improvement.

Relationship with other Processes



Source: the Art of Service **The basic concepts of Problem Management**

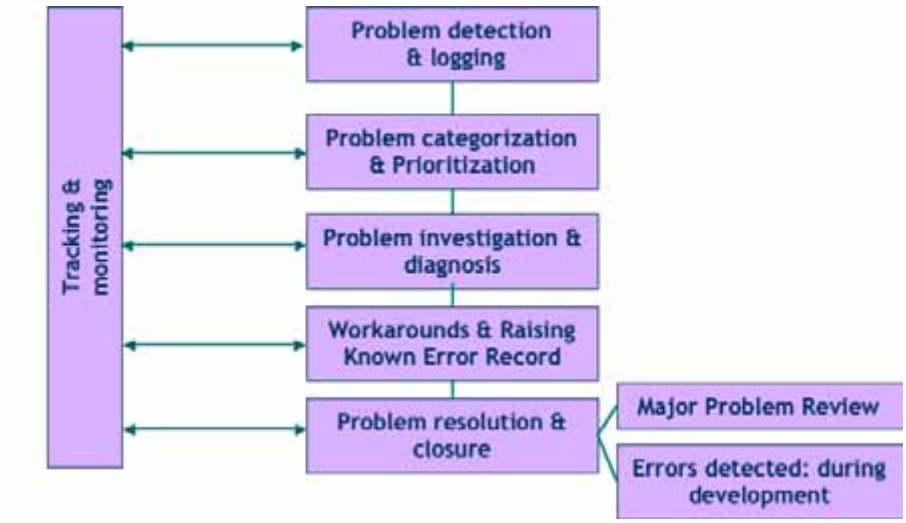
As shown above, the way in which problems can be identified and corrected occurs in multiple ways. For most organizations, the primary benefit of Problem Management is demonstrated in the “Many to One” relationship between Incidents and Problems. This enables an IT Service Provider to resolve many Incidents in an efficient manner by correcting the underlying root-cause.

Why do some problems not get diagnosed?
Because the root cause is not always found.

Why do some Known Errors not get fixed?
Because we may decide that the costs exceed the benefits of fixing the error or because it may be fixed in an upcoming patch from a Supplier, e.g. Windows patch or update.

There are two sub-processes in Problem Management: Reactive and Proactive.

24.5.1 Reactive Problem Management



The activities of Reactive Problem Management are similar to those of Incident Management for the logging, categorization and classification for problems. The subsequent activities are different as this is where the actual root-cause analysis is performed and the Known Error corrected.

Problems must be prioritized in the same way as incidents, but the frequency and impact of related incidents should be taken into account. In order to determine the root cause of a problem, the appropriate level of resources and expertise should be applied. It is often valuable to try to recreate the failure, so as to understand what has gone wrong, and then to try various ways of finding the most appropriate and cost-effective resolution to the problem. To do this effectively without causing further disruption to the users, a test system will be necessary that mirrors the production environment.

Workarounds can be used for any incidents caused by the problem. An example of a workaround may be a manual amendment made to an input file to allow a program to complete its run successfully and allow a billing process to complete satisfactorily. In this example, the reason for the file becoming corrupted in the first place must be determined and work on a permanent resolution continued.

Major Problem Review

After every major incident or problem, a review should be conducted to learn any lessons for the future. Specifically the review should examine:

- Those things that were done correctly
- Those things that were done wrong
- What could be done better in the future?
- How to prevent recurrence

- Whether there has been any third-party responsibility and whether follow-up actions are needed.

Such reviews can be used as part of training and awareness activities for staff – any lessons learned should be documented in appropriate procedures, working instructions, diagnostic scripts or Known Error Records.

24.5.2 Proactive Problem Management

The two main activities of Proactive Problem Management are:

Performing a Trend Analysis

- Review reports from other processes (e.g. Incident, Availability Management)
- Identify recurring problems or training opportunities.

Targeting Preventative Action

- Perform a cost - benefit analysis of all costs associated with prevention
- Target specific areas taking up the most support attention
- Will diagnose the root cause of incidents
- Determine the resolution to those problems
- Ensure that the resolution is implemented through the appropriate control procedures, especially change management and release management.

Problem Management will also maintain information about problems and the appropriate workarounds and resolutions, so that the organization is able to reduce the number and impact of incidents over time. In this respect Problem Management has a stronger interface with Knowledge Management, and tools such as the Known Error Database will be used for both.

Problem management is dependent on information from the CMS and Known Error Database.

24.5.3 Problem Management and Cloud Computing

Where Incident Management is focused on the speed in which the organization brings the service back in operation, Problem Management aims for a quality approach. Analysis is required of the underlying reasons of the incident(-s) and, in cause-and-effect scenarios, the IT group will work together with the cloud computing providers to come up with a reason for the issues at hand.

Problem Management will need input from Configuration Management in order to do a proper analysis and problem definition. It will need to understand the connections between the internal IT infrastructure and the cloud computing supplier.

Virtualization of infrastructure makes Problem Management activities more difficult as the impact of errors on a particular piece of equipment will have effects to various sub-components of the service delivery. It is therefore even more important to have structure and repeatable process steps in place for Problem Management.

24.5.4 Critical Success Factors (CSFs)

Problem Management relies on the establishment of an effective Incident Management process, ensuring that problems are identified as soon as possible and that as much work is done on pre-qualification as possible. Interfaces and working practices between these processes should demonstrate:

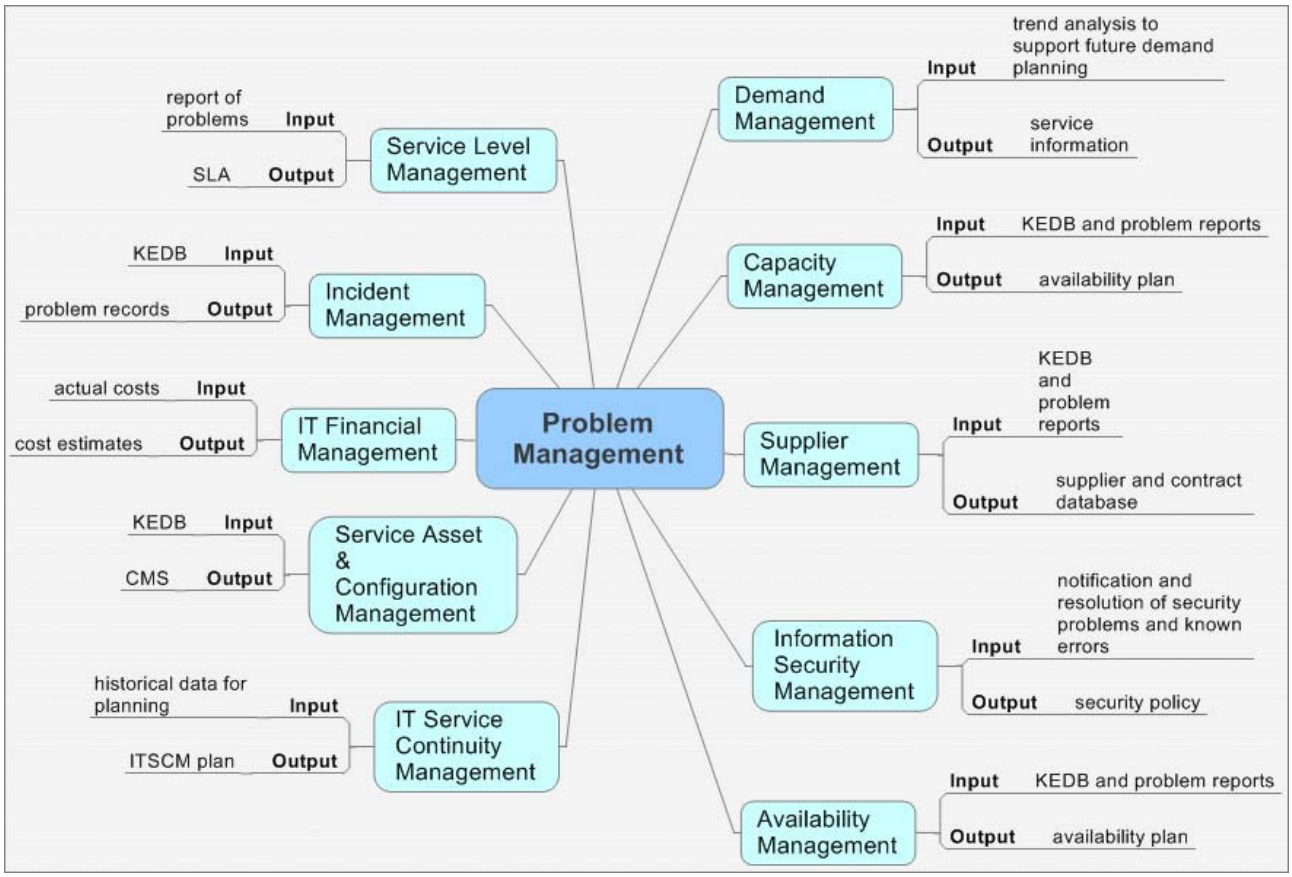
- Linking of Incident and Problem Management tools
- The ability to relate incident and problem records
- The second and third line staff should have a good working relationship with staff on the first line
- Making sure that business impact is well understood by all staff working on problem resolution.

24.5.5 Key Performance Indicators (KPIs)

Metrics should be used to judge the effectiveness and efficiency of the Problem Management process, and its operation. All metrics should be broken down by category, impact, severity, urgency and priority level and compared with previous periods.

- Total number of problems recorded in the period
- Per cent of problems resolves within SLA targets
- Number and per cent of problems that exceed their resolution times
- Backlog of outstanding problems and the trend
- Average cost of handling problem
- Number of major problems (opened/closed)
- Per cent of Major Problem Reviews successfully performed
- Number of Known Errors added to the KEDB
- Per cent accuracy of the KEDB
- Per cent of Major Problem Reviews completed successfully and on time.

Interfaces with Other Processes



24.5.6 Problem Management relating to Platform and Storage Management

Depending on the level of control and involvement the service provider has in terms of the PaaS and storage environments, Problem Management may not be responsible for actually investigating and removing problems and known errors identified within that set of infrastructure. Instead, the focus may be on recording the known errors and associated workarounds to be used by the IT support staff, as well as working to identify reoccurring incidents (through Incident Management). Proactive Problem Management will also work with Availability Management to identify and remove vulnerable areas of the infrastructure that is under the control of the service provider themselves.

The service provider should work with the supplier for all documented changes and releases being implemented, so that the associated release notes and technical support documentation can be made available to the Service Desk and other lines of support. To

assist in the clarification of roles and responsibilities, a configuration model or technical service catalogue should be agreed upon, identifying which elements are to be managed by either the service provider or supplier.

24.6 *Event Management*

Objective: The ability to detect events, make sense of them and determine the appropriate control action is provided by Event Management. Event Management is therefore the basis for Operational Monitoring and Control. In addition, if these events are programmed to communicate operational information as well as warnings and exceptions, they can be used as a basis for automating many routine Operations Management activities. For example, executing scripts on remote devices, or submitting jobs for processing, or even dynamically balancing the demand for a service across multiple devices to enhance performance.

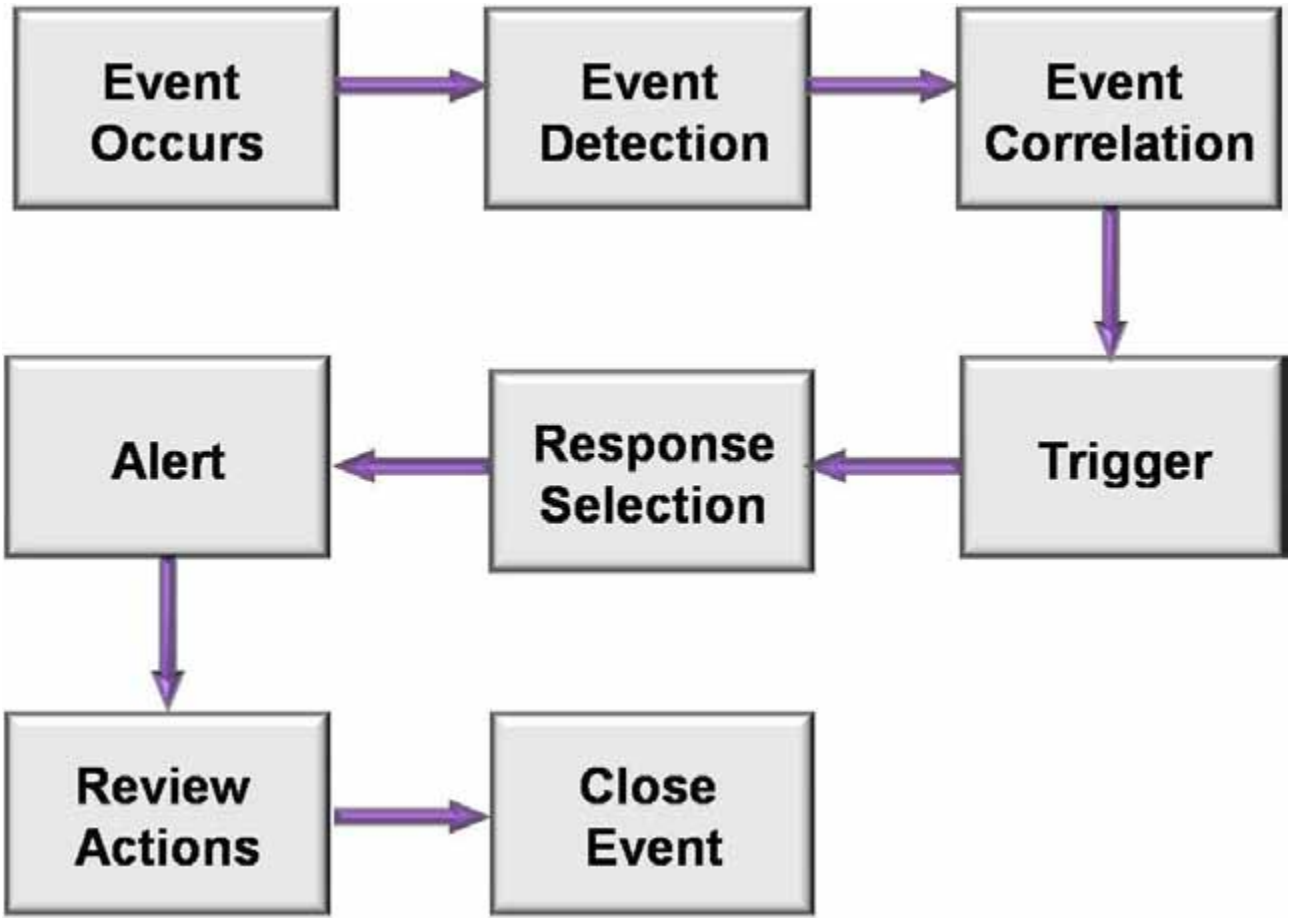
Event Management therefore provides the entry point for the execution of many Service Operation processes and activities. In addition, it provides a way of comparing actual performance and behavior against design standards and SLAs. As such, Event Management also provides a basis for Service Assurance and Reporting; and Service Improvement. This is covered in detail in the Continual Service Improvement publication.

There are many different types of events, for example:

- Events that signify regular operation
- Notification that a scheduled workload has been completed
- A user has logged in to use an application
- An email has reached its intended recipient
- Events that signify an exception
- A user attempts to log on to an application with an incorrect password
- An unusual situation has occurred in a business process that may indicate an exception requiring further business investigation (e.g. a web page alert indicates that a payment authorization site is unavailable – impacting financial approval of business transactions)
- A device's CPU is above the acceptable utilization rate
- A PC scan reveals the installation of unauthorized software.

Events that signify unusual, but not exceptional, operation are an indication that the situation may require closer monitoring. In some cases the condition will resolve itself; for example in the case of an unusual combination of workloads – as they are completed, normal operation is restored. In other cases, operator intervention may be required if the situation is repeated or if it continues for too long. These rules or policies are defined in the Monitoring and Control Objectives for that device or service.

The following diagram is a high-level and generic representation of Event Management. It should be used as a reference and definition point, rather than an actual Event Management flowchart.



Flowchart for Event Management Activities

Event Correlation

Filtering: The purpose of filtering is to decide whether to communicate the event to a management tool or to ignore it. If, ignored, the event will usually be recorded in a log file on the device, but no further action will be taken. Significance of events: every organization will have its own categorization of the significance of an event, such as Informational, Warning, and Exception.

Correlation is normally done by a correlation engine, this will take into account:

- Number of similar events
- Whether the event represents an exception

- Whether additional data is required to investigate further

Event categorization
Event prioritization: In most organization's IT infrastructure there would be a significant amount of events occurring every day, which may impact on the way in which events are correlated and how they provide triggers indicating a response is needed.

If the correlation activity recognizes an event, a response will be required. The mechanism used to initiate that response is called a trigger. There are many different types of triggers, each designed specifically for the task it has to initiate. Some examples include:

- Incident triggers that generate a record in the Incident Management system, thus initiating the Incident Management process
- Change Triggers that generate a Request for Change (RFC), thus initiating the Change Management process
- A trigger resulting from an approved RFC that has been implemented but caused the event, or from an unauthorized change that has been detected – in either case this will be referred to Change Management for investigation
- Scripts that execute specific actions, such as submitting batch jobs or rebooting a device
- Paging systems that will notify a person or team of the event by mobile phone
- Database triggers that restrict access of a user to specific records or fields, or that create or delete entries in the database.

24.6.1 Event Management relating to Platform and Storage Management

The technology groups implementing systems using cloud-based PaaS and storage environments should make sure that Event Management capabilities are built into the systems from the start. Once the requirements for availability, capacity, performance and information security are known, system architects and other developers can ensure that active or passive monitoring agents are configured, focusing on areas of vulnerable (in terms of risk) and complex areas of the service.

The challenge for many organizations will be found where the cloud environment is controlled by an external supplier, in which case scripted agents will need to be implemented within their own systems that generate a synthetic transaction to assess the actual performance, capacity and availability of the PaaS and storage environments. In either case, input from Capacity, Availability and Information Security Management should be used to define the requirements for Event Management capabilities and to avoid an excess of event notifications that don't provide appropriate informational value.

24.7 *Service Validation and Testing*

The purpose of the Service Validation and Testing process is to:

- Plan and implement a structured validation and test process that provides objective evidence that the new or changed service will support the customer's business and stakeholder requirements, including the agreed service levels
- Quality assure a release, its constituent service components, the resultant service and service capability delivered by a release
- Identify, assess and address issues, errors and risks throughout Service Transition.

The goal of Service Validation and Testing is to assure that a service will provide value to customers and their business. The objectives of Service Validation and Testing are to:

- Provide confidence that a release will create a new or changed service or service offerings that deliver the expected outcomes and value for the customers within the projected costs, capacity and constraints
- Validate that a service is 'fit for purpose' – it will deliver the required performance with desired constraints removed
- Assure a service is 'fit for use' – it meets certain specifications under the specified terms and conditions of use
- Confirm that the customer and stakeholder requirements for the new or changed service are correctly defined and remedy any errors or variances early in the service lifecycle as this is considerably cheaper than fixing errors in production.

Service failures can harm the service provider's business and the customer's assets and result in outcomes such as loss of reputation, loss of money, loss of time, injury and death. The key value to the business and customers from Service Testing and Validation is in terms of the established degree of confidence that a new of changed service will deliver the value and outcomes required of it and understanding the risks.

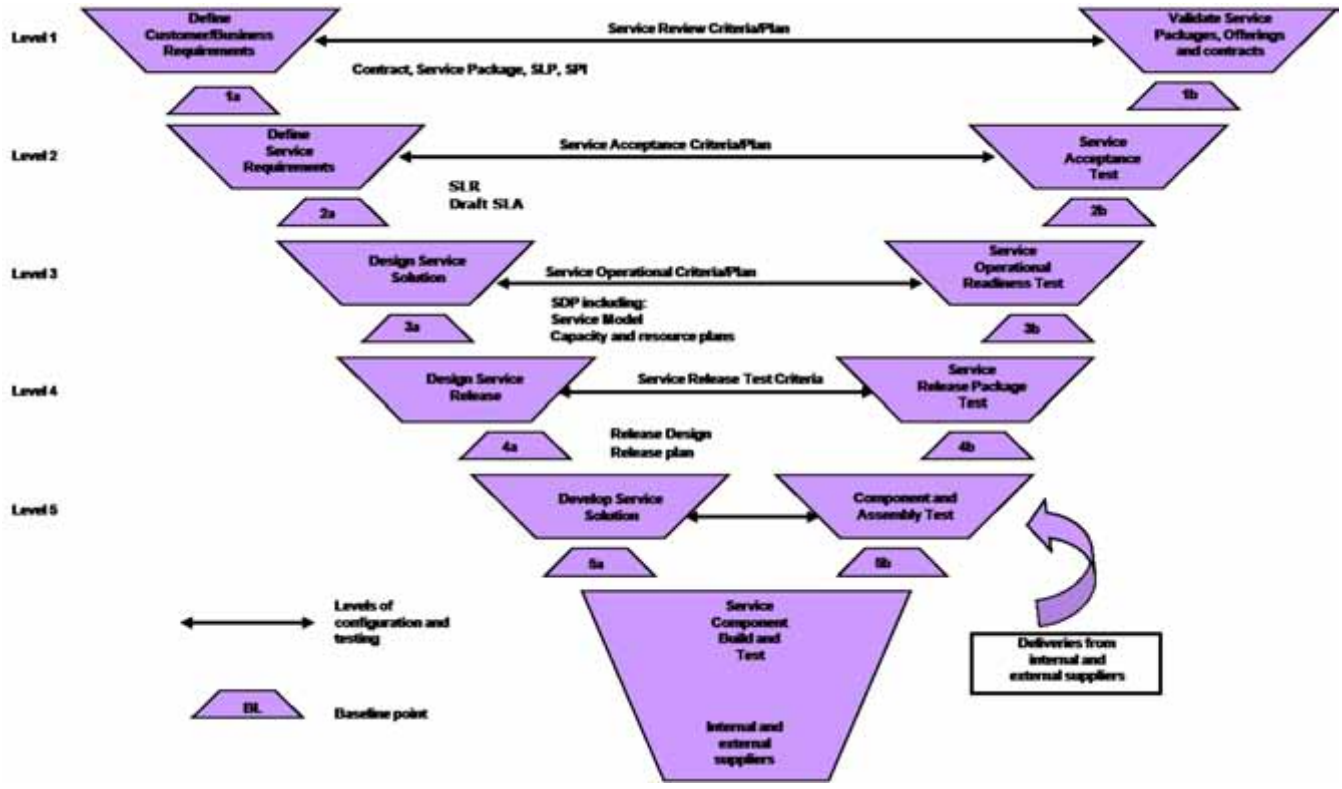
Successful testing depends on all parties understanding that it cannot give, indeed should not give, any guarantees but provides a measured degree of confidence. The required degree of confidence varies depending on the customer's business requirements and pressures of an organization.

Testing is related directly to the building of service assets and products so that each one has an associated acceptance test and activity to ensure it meets requirements. This involves testing individual service assets and components before they are used in the new or changed service.

Each service model and associated service deliverable is supported by its own re-usable test model that can be used for regression testing during the deployment of a specific release, as well as for regression testing in future releases. Test models help with building quality early into the service lifecycle rather than waiting for results from tests on a release at the end.

Using a model such as the V-model builds in Service Validation and Testing early in the service lifecycle. It provides a framework to organize the levels of configuration items to be managed through the lifecycle and the associated validation and testing activities both within and across stages.

The level of test is derived from the way a system is designed and built up. This is known as a V-model, which maps the types of test to each stage of development. The V-model provides one example of how the Service Transition levels of testing can be matched to corresponding stages of service requirements and design.



The Service V Model

The left-hand side represents the specification of the service requirements down to the detailed Service Design. The right-hand side focuses on the validation activities that are performed against the specifications defined on the left-hand side. At each stage on the left-hand side, there is direct involvement by the equivalent party on the right-hand side. It shows that service validation and acceptance test planning should start with the definition of the service requirements. For example, customers who sign off the agreed service requirements will also sign off the service Acceptance Criteria and test plan.

24.7.1 Service Validation and Testing relating to Platform and Storage Management

While the PaaS and storage environments hosted in a cloud will require significant testing themselves, an appropriate testing model will also need to cater for additional systems and infrastructure that are involved in providing the end-to-end service to the user. Relating to the Service V model, in most cases ongoing testing to validate that the PaaS and storage environments are delivering the required level of utility and warranty occurs at levels 4 and 5, with actual responsibility being assigned to parties from both the service provider and any external supplier(s) involved. As an example, for services using cloud-based storage environments, such tests will assess:

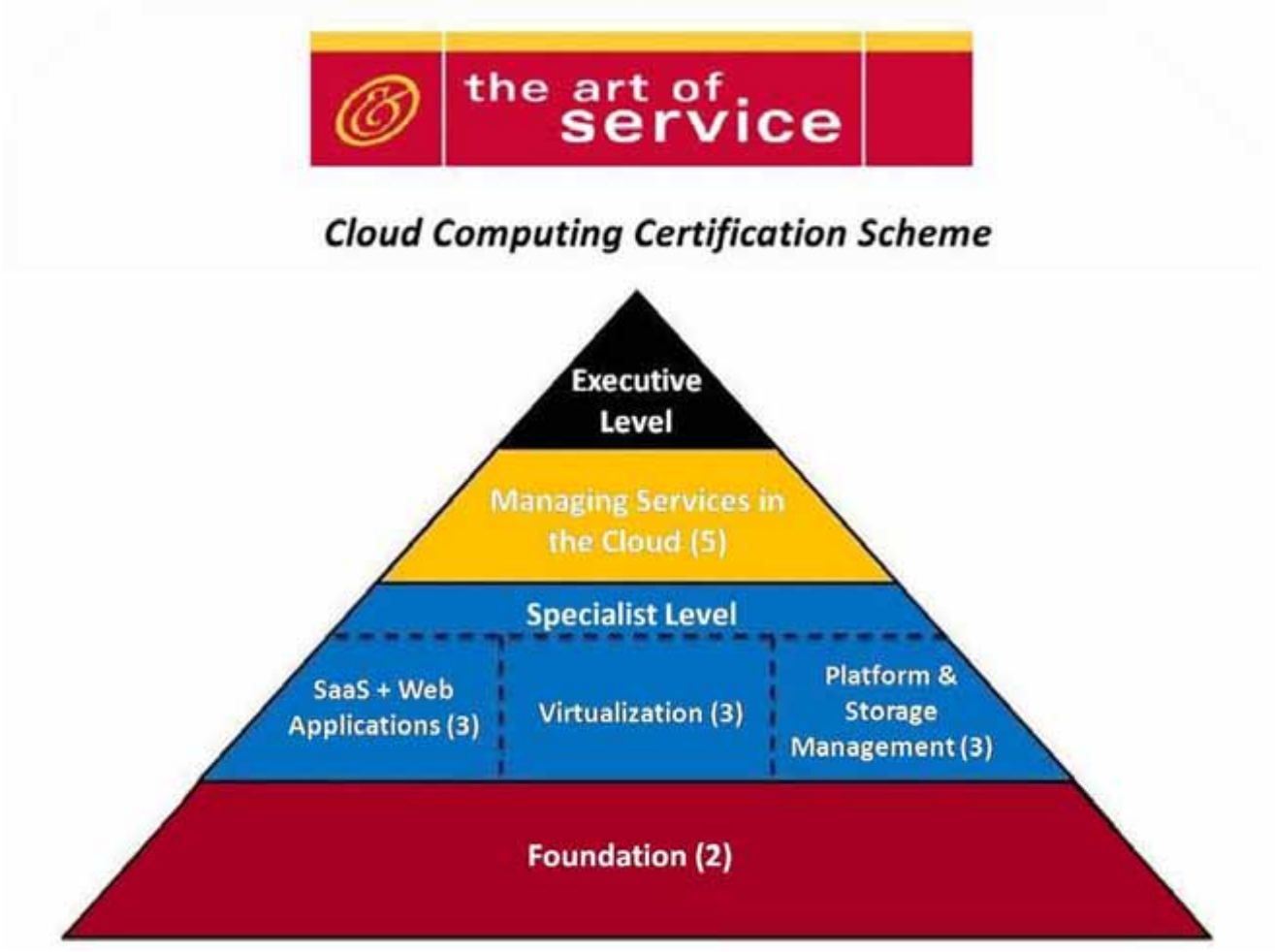
- That all required levels of demand (based on expected peak hourly load) can be provisioned as expected
- That the performance provided by the storage environment remains within operational limits under various simulated utilization demands
- That the security requirements for the storage environment are always provided (especially in shared infrastructure models)
- That all elements of functionality required are provisioned
- That through various simulation tests the service provider accepts the storage environment is ready for use as part of production services.

25 Certification

25.1 *Cloud Computing Certification Pathways*

IT professionals need to know a lot about the various ways of delivering services to the customers and end-users. It is no longer sufficient just to know the differences between Windows-based or Linux-based architecture. These days, most services will utilize some form of cloud computing, be it virtualization or SaaS offerings.

So with the change in computing and IT Service Delivery comes a whole new series of qualifications and certification. The Cloud Computing Certification Scheme has been created to support the IT Professional who needs to be a 'niche generalist', especially in a rapidly changing area like cloud computing.



How does the certification pathway work?

First, you need to create the foundation – The Cloud Computing Foundation Program focuses on the fundamentals, general knowledge, terminology and basic concepts used in cloud computing. This program earns you 2 points towards your Cloud Computing Executive Level Certificate.

After this come the Cloud Computing Specialization options. The prerequisite for these programs is that you have the Cloud Computing Foundation certificate. We appreciate that you don't need to know everything about each area of the IT organization so this is where the programs become more specialized:

- Virtualization
- Software as a Service and Web Applications
- Platform and Storage Management.

Each program in this series earns 3 points towards the Cloud Computing Executive Certificate.

The next level is 'Managing Services in the Cloud' and this program is specifically aimed at Service Managers and IT Service Delivery Managers who wish to add cloud computing as an option in their organization's delivery model. The program is worth 5 points.

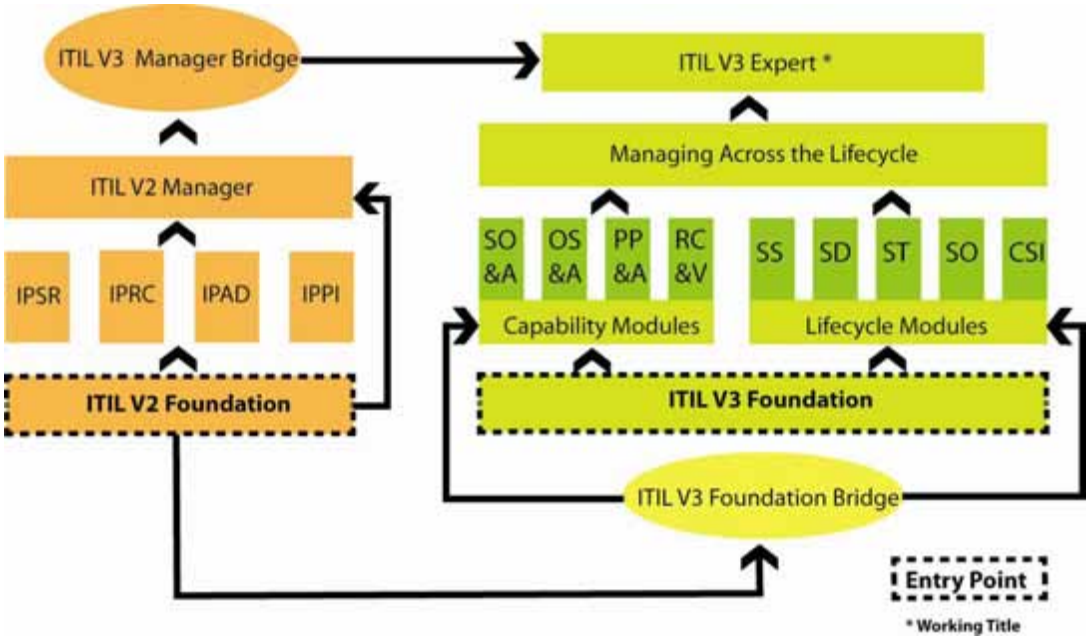
How do I achieve my Executive level?

You must have a minimum of 16 points to achieve your executive certification. The Foundation level, all Specialist level programs and Managing Services in the Cloud must be completed in order to gain Executive level certification.

Each course provides preparation for the exam and successful candidates receive a certificate results sheet.

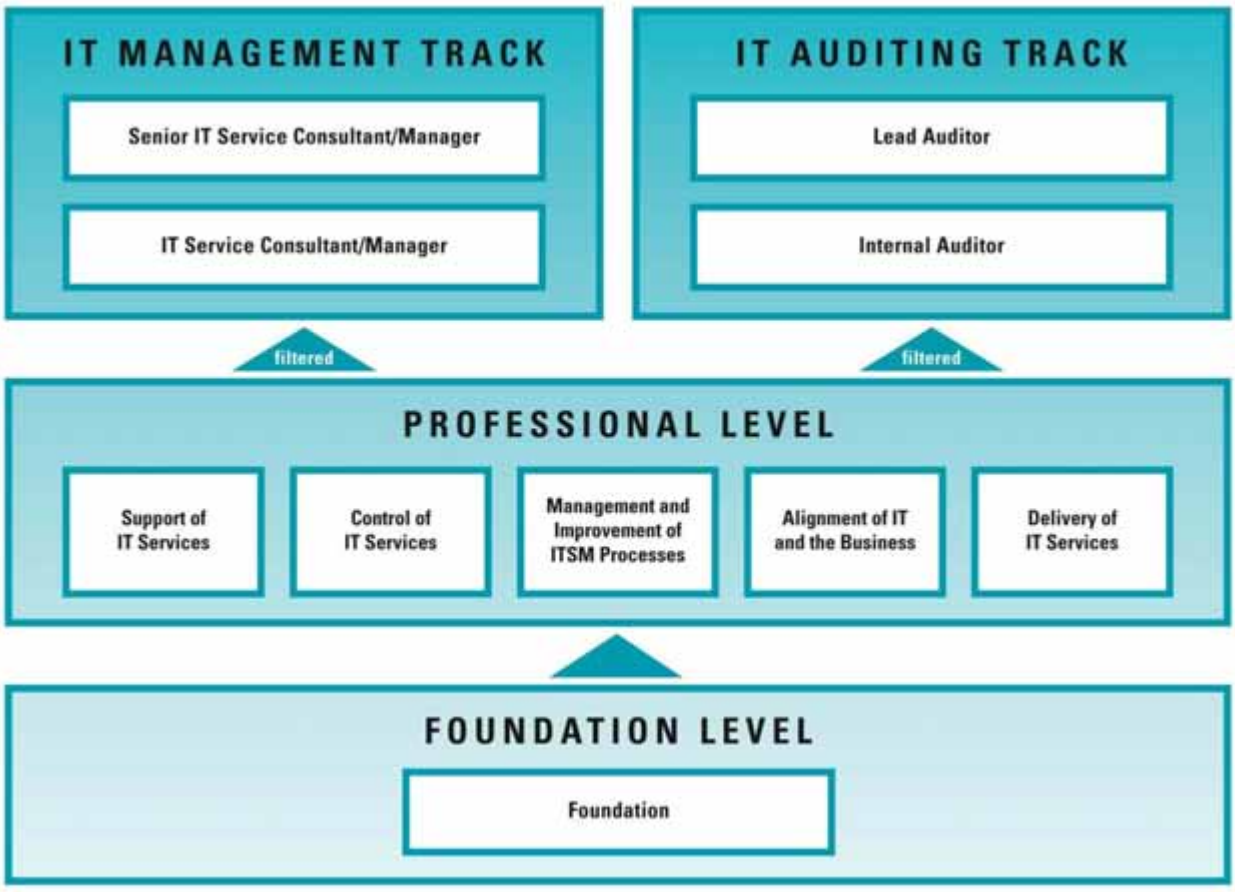
25.2 *ITIL® Certification Pathways*

There are several pathway options that are available to you once you have acquired your ITIL® Foundation Certification, as illustrated below. Currently it is intended that the highest certification is the ITIL® V3 Expert, considered to be equal to that of Diploma Status.



25.3 *ISO/IEC 20000 Certification Pathways*

ISO/IEC 20000 Standard is becoming a basic requirement for IT Service Providers and is the most recognized symbol of quality regarding IT Service Management processes. ISO/IEC 20000 programs aim to assist IT professionals to master and understand the standard and the issues relating to earning actual standard compliance.



For more information on certification and available programs please visit our website http://www.artofservice.com.au

26 Platform and Storage Management Specialist Exam Tips

Exam Details

- 20 multiple-choice questions
- The correct answer is only one of the four
- 30 minutes duration
- 16 out of 20 is a pass
- Closed-book
- No notes.

Practical Suggestions

- Read the question CAREFULLY.
- At this level of exam the obvious answer is often the correct answer (if you have read the question carefully)!
- Beware of being misled by the preliminary text for the question.
- If you think there should be another choice that would be the right answer, then you have to choose the “most right”.
- Use strategies such as “What comes first?” or “What doesn’t belong?” to help with the more difficult questions.

Organizing your Exam

The Art of Service facilitates the Platform and Storage Management Specialist exam. Please contact us on +61 (7) 3252 2055 or email trainer-support@theartofservice.com to arrange your examination.

Make sure that you prepare adequately in the lead up to your exam by reviewing your notes, reading any available material and attempting the sample exams.

We wish you luck in your exam and future cloud computing, Platform and Storage Management career!

27 References

PaaS
Wainewright, Phil. A plethora of PaaS options. March 6, 2008. http://blogs.zdnet.com/SAAS/?p=472

Barnett, Alex. So what is this Platform as a Service Thing? April 8, 2008.
http://alexbarnett.net/blog/archive/2008/04/08/so-what-is-this-platform-as-a-service-thing.aspx

Barnett, Alex. Time to Define Platform-as-a-Service, or PaaS. February 19, 2008.
http://alexbarnett.net/blog/archive/2008/02/19/time-to-define-quot-platform-as-a-service-quot-or-paas.aspx

Pittarese, Dr. Tony. Supporting Collaborative Software Development Using CASE: Team-Based Issue in CASE Success.
http://www.pittarese.com/Auburn/cse625/Supporting%20Collaborative%20Software%20Development%20Using%20CASE.htm

Bungee Labs. Defining Platform-As-A-Service, or PaaS. February 18, 2008.
http://blogs.bungeeconnect.com/2008/02/18/defining-platform-as-a-service-or-paas/

Andressen, Marc. The three kinds of platforms you meet on the Internet. September 16, 2007.
http://blog.pmarca.com/2007/09/the-three-kinds.html

MacVittie, Lori. As a Service: The many faces of the cloud. November 20, 2008.
http://devcentral.f5.com/weblogs/macvittie/Default.aspx

MacVittie, Lori. Bursting the Cloud. September 3, 2008.
http://devcentral.f5.com/weblogs/macvittie/archive/2008/09/03/3584.aspx

Charrington, Sam. The Blind Men and the Cloud. August 25, 2008. http://www.appistry.com/blogs/sam/the-blind-men-and-cloud

TechFAQ. What is a global catalog? http://www.tech-faq.com/global-catalog.shtml

TechFAQ. What is Active Directory? http://www.tech-faq.com/active-directory.shtml

Fontecchio, Mark. Uptime Institute expands data center tier rating system. March 19, 2008.
http://searchdatacenter.techtarget.com/news/article/0,289142,sid80_gci1306297,00.html#

Webopedia Computer Dictionary. data center tiers.
http://www.webopedia.com/TERM/D/data_center_tiers.html

Staten, James. Is Cloud Computing Ready For The Enterprise? Forrester Research, Inc. March 2008.
www.forrester.com

Andreessen, Marc. Analyzing the Facebook Platform, three weeks in.. June 12, 2007.
http://blog.pmarca.com/2007/06/analyzing_the_f.html

What is Social Software? http://www.sociallibraries.com/farkaschap1.pdf (A Companion to Social Software in Libraries by Meredith Farkas)

Takase, Akihiko D. Sc. And Kikuchi, Susumu. Platform Architecture for Networked Businesses.
http://hitachi.com/ICSFiles/afieldfile/2004/06/01/r2000_04_101.pdf

Carraro, Gianpaolo and Chong, Frederick. Architecture Strategies for Catching the Long Tail. April 2006. http://msdn.microsoft.com/en-us/library/aa479069.aspx

Wikipedia. Representational State Transfer. en.wikipedia.org/wiki/Representational_State_Transfer

Wikipedia. SOAP (Protocol). en.wikipedia.org/wiki/SOAP_(protocol)

Wikipedia. Remote Procedure Call. en.wikipedia.org/wiki/remote_procedure_call
Wikipedia. Cloud Computing. en.wikipedia.org/wiki/Cloud_computing

Wikipedia. Cloud Platforms. en.wikipedia.org/wiki/Cloud_platforms

Wikipedia. Internet. en.wikipedia.org/wiki/Internet

Wikipedia. Infrastructure as a service. en.wikipedia.org/wiki/Infrastructure_as_a_service

Wikipedia. Platform as a service. en.wikipedia.org/wiki/Platform_as_a_service

Wikipedia. Software as a service. en.wikipedia.org/wiki/Software_as_a_service

Wikipedia. Independent Software Vendor. en.wikipedia.org/wiki/Independent_Software_Vendor

Microsoft Developer Center
State Management topics

MSDN Architecture Center
MultiTenant Architecture

Storage Management
Wikipedia. Cloud Computing. en.wikipedia.org/wiki/Cloud_computing

Wikipedia. Storage Virtualization. en.wikipedia.org/wiki/Storage_virtualization

Wikipedia. Cloud Storage. en.wikipedia.org/wiki/Cloud_storage

Enterprise Storage Management. What is Hierarchical Storage Management?
http://www.enterprisestoragemanagement.com/faq/hierarchical-storage-management.shtml

Internet.com. Storage Resource Management.
http://www.webopedia.com/TERM/S/Storage_Resource_Management.html

Gartner, Inc. Gartner Says Worldwide IT Spending On Pace to Surpass $3.4 Trillion in 2008.
http://www.gartner.com/it/page.jsp?id=742913. August 18, 2008

Jones, M. Tim. Cloud computing with Linux. http://www.ibm.com/developerworks/linux/library/l-cloud-computing/#resources

Jones, M. Tim. Virtual Linux. http://www.ibm.com/developerworks/linux/library/l-linuxvirt/?S_TACT=105AGX03&S_CMP=ART

Advantage International Systems, Inc. What is Availability Management?
http://www.advgroup.com/availability.html

Bunn, Frank, Simpson, Nik, Peglar, Robert, and Nagle, Gene. Storage Virtualization. SNIA: Colorado Springs, 2004. http://www.snia.org/education/storage_networking_primer/stor_virt/sniavirt.pdf

Barker, Richard and Massiglia, Paul What Storage Networking Is and What It Can Mean To You from Storage Area Network Essentials. John Wiley & Sons, Inc., New York, 2002. Found as Understanding Storage Area Networks, SNIA, http://www.snia.org/education/storage_networking_primer/san/

Farley, Marc. An Introduction to Storage Devices, Subsysterms, Applications, Management, and File Systems from Storage Networking Fundamentals. Cisco Press. Found on web: http://www.snia.org/education/storage_networking_primer/stor_devices/

O'Connor, Michael & Judd, Josh. Introduction to File Area Networks. West Conshohocken: Infinity Publishing. 2007. http://www.snia.org/education/storage_networking_primer/stor_mngmnt/sniasnmbooklet_final.pdf/education/storage_networking_primer/fan/Intro_to_FAN.pdf

Cummings, Roger and Fruchauf, Hugo. Storage Networking Security. SNIA. 2003 http://www.snia.org/education/storage_networking_primer/storage_security/SNIAsecbookletfinal.pdf

Villars, Richard L, Gillan, Al, and Perry, Randy. Business Value of Virtualized IT. Citrix and NetApp. July 2008. http://viewer.bitpipe.com/viewer/viewDocument.do?accessId=8513530

3PAR. Green Storage Explained. 20 Aug 2008. http://media.techtarget.com/Syndication/NATIONALS/3PAR_TechReportsStor_9_16.pdf

Vertica Systems. Transforming the Economics of Data Warehousing with Cloud Computing. November 2008. http://viewer.bitpipe.com/viewer/viewDocument.do?accessId=8513932

Preimesberger, Chris. Get Off of My Cloud: Private Cloud Computing Takes Shape. November 4, 2008. eWeek.com. http://www.eweek.com/c/a/Cloud-Computing/Why-Private-Cloud-Computing-Is-Beginning-to-Get-Traction/1/

Company Offerings are found at their websites:
Amazon.com
iForum.com
Box.net
Nirvanix.com
ElephantDrive.com
Humyo.com
Mosso.com
ParaScale.com
Cleversafe.com
CommValue.com

INDEX*

A

B

C

D

E

F

G

H

I

M

N

O

P

Q

S

T

U

v

w