Resilient cloud computing

V. Salapura
R. Harper
M. Viswanathan

IBM SmartCloud® Enterprise+ (SCE+) is IBM's premier cloud computing offering for enterprise customers. The SCE+ cloud is designed to be a resilient, highly available system with no single point of failure. SCE+ uses an integrated set of enterprise-class servers, network elements, and storage components that have internal redundancy, RAS (reliability, availability, and serviceability) features, and high mean time between failures. SCE+ is also a managed cloud. The SCE+ management system utilizes IBM service-management and platform-management tools. To help ensure that the customer's virtual servers are running and avoid system failures, we employ several techniques. For example, virtual servers are automatically restarted upon crash or upon hosting physical-server failures. In addition, network elements are interconnected to allow alternate network traversal paths, and data are mirrored to offset storage failure, while each of these server, storage, and network elements also have redundant internal configurations. In addition to providing guest-level availability, important customer workloads, such as the enterprise resource planning applications and databases, require highly available clusters. More generally, we describe approaches used to achieve resiliency in SCE+.

Introduction

Cloud computing is becoming the new de facto environment for many system deployments in a quest for more agile on-demand computing with lower total cost of ownership. Companies and various agencies and institutions are quickly trying to adopt cloud computing, bringing high expectations of resiliency that have heretofore been associated with dedicated data centers [1–4].

The cloud computing environment and offering of IBM for enterprise customers and production-grade workloads is named SmartCloud* Enterprise+ (SCE+). It is designed to bring advantages of cloud computing to the Strategic Outsourcing customers of IBM, providing both shared and dedicated infrastructure with full management capabilities. SCE+ represents a dramatically different delivery model for IBM Global Technology Services (GTS). It revolutionizes a traditional services delivery model by delivering IBM-hosted compute resources within hours (instead of weeks), with standardized, resilient, and secure IBM infrastructure, tools, and services. This environment fulfills customer workload needs through a self-service portal and enables multi-tenant use through virtualization. It offers cloud-capable functions, such as consumption-based metrics and automated service-management integration, and it reduces manual intervention and delivery costs by extensive process automation.

SCE+ is offered as standard services with a defined delivery catalog and with optimized service management. It has centralized management with automated mechanisms for key Information Technology Infrastructure Library (ITIL version 3) processes. A fully resilient management environment is based on IBM service-management and platform-management tools.

The SCE+ cloud is designed to be a resilient and highly available system for enterprise-class applications. SCE+ uses highly reliable servers and storage components that have internal redundancy, RAS (reliability, availability, and serviceability) features, and high mean time between failures (MTBF). All server, networking, and storage elements are redundant. SCE+ uses mature enterprise-class virtualization technologies such as the VMware** ESX [5] suite and the IBM PowerVM* [6]. To fully exploit the SCE+ redundant physical infrastructure, software mechanisms such as high-availability (HA) clustering are implemented for most of the management tools.

Digital Object Identifier: 10.1147/JRD.2013.2266972

Table 1 Overview of resiliency attributes of SCE+ management and managed systems. (MTBF: mean-time between failure; OS: operating system; SLA: service-level agreement.)

Resiliency attributes	*Management (cloud platform)*	*Managed (customer servers)*
Robust hardware and virtualization products	Robust IBM hardware, with high MTBF	
Resilient physical infrastructure architecture	No single point of failure, with full redundancy in each component	
Resilient software infrastructure architecture	Software clustering (app/database/OS clusters, quick restart)	Auto-restart on OS crash; restart on another server if server fails
Data resiliency	Data mirroring of management tools	Disk replication for Platinum SLA (allows disk failure without service interruption)
Data backup	Periodical image backup and daily incremental backups for data	Daily incremental backup with disk and (shared, multi-customer) tape copies
Planned maintenance	Rolling update techniques such as live partition mobility and workload live migration, and application redundancy for minimal planned outage	
Resilient growth	Design enables scale-up and scale-out of management and customer services	
Availability SLA	Support for better than 99.9% availability	Support for graduated SLA from 98.5% to 99.9% availability

SCE+ employs several techniques to keep virtual servers (guests) running in the event of failures. If a customer virtual machine (VM) crashes or hangs, it is automatically restarted. When the physical server that hosts these VMs fails, the guests are either automatically restarted in place when the server comes back up or are restarted on alternate servers. In addition to providing guest-level availability, important customer workloads, such as SAP applications, require HA clusters that provide availability for software processes within the virtual servers.

SCE+ resiliency principles

The nature of the customer base of SCE+ demands that resiliency is built into the infrastructure, into the tools that manage the customer-owned services and enabling fault-avoidant applications to be hosted on the virtual servers. The principles supporting the design involve usage of a *resilient physical* infrastructure, with efficiencies to manage cost and failure prevention and isolation for all layers and components of the system. Best practices are applied at all layers of the system in order to meet required service-level agreements (SLAs). SLAs are contractual obligations and in many cases include penalties for noncompliance.

In this paper, we use the term *management system* (or management, or management tools) to refer to the tools and the systems needed to manage the cloud computing environment. We use the term *managed system* to refer to the customer-owned virtual servers and inherent workloads. An overview of resiliency attributes for both management and managed systems in SCE+ is given in **Table 1**. The main properties of each of the layers are summarized below.

Two schools of thought are common with respect to how the infrastructure for a cloud system should be built. One approach, such as pursued by Amazon and Google [7], uses redundant, inexpensive, expendable building blocks. The other approach, used by IBM SCE+, uses high-end building blocks with significant internal redundancy and an established track record of very high MTBF for every element in the infrastructure. The same rationale for using mature high-end products is used to select the virtualization components such as VMware ESX and PowerVM.

The physical infrastructure architecture eliminates single points of failure, with a physically partitioned design of the SCE+ basic building block called the *point of delivery* (PoD). Further, a PoD can be evenly split into two building blocks, enabling each (logical) half to be separated by a distance of approximately 50 km (the distance, with a margin of safety, over which dark fiber allows lossless data transmission using dense-wavelength division multiplexing without amplification).

In the software infrastructure, all management tools are configured into redundant pairs of instances. The health (e.g., proper functioning) of the management tools is monitored, and corrective actions are performed in the event of failure. Depending on the capabilities of the management tool, either clustering or "auto-detect and restart" is used.

For customer virtual servers, failures such as crashes and "hangs" are automatically detected, and the virtual servers are restarted on the same server or on another server.

For data resiliency, all management data are stored in databases, with the database data replicated and stored in different storage subsystems. Storage for customer servers is replicated within (and for some SLAs, across) storage appliances for seamless recovery should the primary appliance fail.

All data are automatically backed up periodically. All management tools deployed on virtualized servers are periodically backed up as images (the entire virtual server, including its storage), while the incremental changes that occur are backed up every day. Customer data are backed up to disk and tape every day, and data backups are available from the self-help customer portal. Backups are also transferred from disk to tape after approximately 24 hours, and the tapes are trucked away from the data center for vaulting.

All systems and tools may require downtime during scheduled maintenance windows. In SCE+, the maintenance of management tools is designed to be performed using rolling (e.g., staggered) updates without affecting customer services. (The phrase *rolling updates* is further explained in the section "Planned Maintenance.") Capabilities such as live partition mobility and dynamic resource scheduling are used to move affected services to systems that are functioning.

For resilient growth, SCE+ design permits both scale-out and scale-up of customer resources on the basis of workload needs. All workloads have built-in scale-up capabilities that allow bursting communications to the nearest system core boundary. Scale-out across VMs, on the other hand, requires advance preparation.

All management components are managed to better than 99.9% availability, whereas the customer servers have a selectable availability SLA from 98.5% (Bronze SLA) to 99.9% (Platinum SLA). These availability agreements are only for unplanned outages. For more information on these various SLAs, see the section "Managed systems-availability SLAs and estimation of SLA metrics."

SCE+ resiliency architecture

The objective for the SCE+ resiliency architecture is to meet the virtual server-level availability SLA required of the SCE+ managed services and to provide a reliable SCE+ management environment. The opportunity to provide highly resilient and highly available systems is vastly improved by standardizing, virtualizing, and modularizing. Standardization is achieved by identically assembling the hardware elements (server, storage, network, racks, and accessories) in each PoD in every data center such that they can be assembled into a container and shipped.

Since virtualization allows packaging of workloads [which include the operating system (OS), applications, and data] in a portable VM image container, it enables easy transfer of workloads from one server to another. High-availability features can relocate a virtual server image from one physical server to another within the same data center if the original server experiences any failure or performance loss or to perform scheduled maintenance.

Before providing the details of how these principles are fulfilled, we first describe the structure of the SCE+ and how its structure contributes to resiliency.

Overall structure of SCE+

Figure 1 shows the global architecture of SCE+ that involves a three-tier cascaded architecture and management topology. The central management hub (at the left in Figure 1) consists of the user portal, service desk, and service-management functions and manages multiple,
distributed sites in different geographies and locations worldwide. We use the term *sites* to refer to logical entities located in physical data centers. Each site contains management tools that are used to manage resources across multiple PoDs. A PoD contains managed resources (server, storage, and network) that are virtualized, divided into IT (information technology) resources, and offered to customers. PoDs also contain management tools for storage management, backup, and performance monitoring. PoDs can be shared by multiple customers, or PoDs can be dedicated to a single customer.

The management tools work together to control request fulfillment, health-checking, monitoring, performance, service activation, service deactivation, patching, license and asset collection, and other ITIL functions for all virtual servers.

Structure of an SCE+ site

An SCE+ site contains management components that manage multiple PoDs within that site. A PoD itself is a self-contained unit of deployment containing low-level management tools and managed infrastructure currently sized to host several thousand customer VMs across two hypervisor technologies: PowerVM (for AIX*) and VMware (for Windows** and Linux**).

Although the design accommodates collocation of multiple sites into a given geographical location to achieve an even higher "scale" if necessary, sites are in general widely separated from the central management site, as well as from one another. Therefore, each site can be considered as an isolated availability zone such that no single site-wide failure can impair the correct operation of another site.

Building block design for the SCE+

A PoD itself is designed to offer internal fault isolation zones. As mentioned, the management and the managed physical

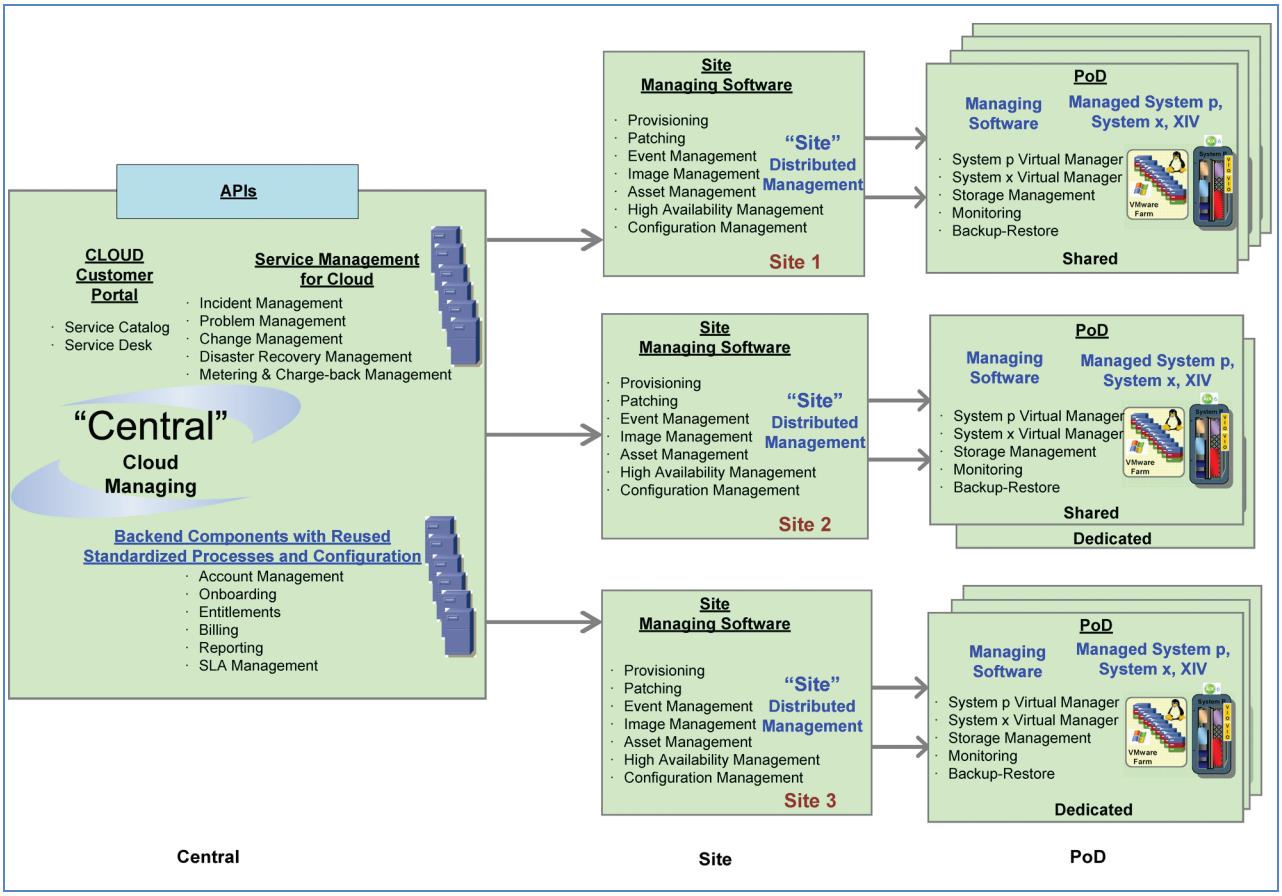

Figure 1

The global structure of SCE+ is organized into three-tiers. The central hub (left) supports the user portal, service desk, and service management functions for multiple sites. We illustrate here three SCE+ sites. Each site manages multiple points of delivery (PoDs). PoDs contain managed resources (customers' virtual servers, storage, and network) and management tools for storage management, backup, and performance monitoring. (API: application programming interface.)

components of a PoD are divided into two equally-sized zones called building blocks. Half of all servers, networking components, and storage in a PoD are located in each building block, and management and managed workload can be distributed across the building blocks. The networking and storage area network (SAN) are arranged such that the two building blocks of a given PoD can be located approximately 50 km apart, thus still enabling synchronous data replication and logically remaining within a single site, to provide an additional measure of resiliency. This allows distribution of a PoD across the rooms or floors of a building or across multiple buildings in a campus. (Note that most data-center outages only have an impact on a subset of the data-center floor space).

Architecture for data resiliency

SCE+ implements reliable storage architecture with data replication both within and across storage appliances. All SCE+ disk storage is contained in high-end storage appliances called IBM XIV* Storage Systems (pronounced "X, I, V," which is a brand name and not an acronym). These appliances contain redundant power, cooling, networking, and striping of data across disk arrays such that they can tolerate and recover from any single and many multiple failure modes with no observable interruption to users of the appliance.

As an added level of resiliency on top of the internal storage appliance redundancy of XIV, the IBM SAN Volume Controller (SVC) vDisk mirroring [8] is used to make any storage appliance failure transparent to Platinum-SLA virtual servers. The SVC vDisk mirroring performs synchronous mirroring of storage from one appliance to another. In the case of a storage appliance failure, the SVC vDisk mirroring transparently switches over to the mirror image without any disruption to the workload. To provide data resiliency for non-Platinum virtual servers, customer data are backed up

daily to disk and tape. These image backups are available from the self-help customer portal. In the event of a storage appliance failure, non-Platinum virtual servers undergo a recovery process in which data are restored from backup tapes onto the originating storage appliance or another storage appliance.

Architecture for network resiliency

The SCE+ network architecture supports management communications between the central management hub and the globally distributed SCE+ sites, providing customer access to the edge of the SCE+ PoD. Within a site and PoD, the architecture supports customer data flow to the managed workload, management data flow to the managed and managing workload, and all data backup data flow. Within a site, traffic flows over separate virtual local area networks (VLANs) for isolation and manageability. For example, the backup traffic flows over a dedicated backup network that is configured to support jumbo frames, and customer and management traffic flows over separate VLANs. All switches, routers, and firewalls are redundant and are interconnected such that no single component failure can sever network connections. All components are monitored for availability and performance and reported through in-band and out-of-band channels.

Architecture for server resiliency

An SCE+ PoD contains pools of servers that are considered by the workload and availability management systems to be one large computer. Availability management techniques that make use of the server pool are outlined below. Each server has redundant network and storage adapters, power, cooling, and memory so that no single such failure can cause any disruption. Depending on the server, the CPUs and caches possess internal redundancy that can survive many internal failure modes, but entire CPU or cache failures are disruptive and require restart of the workload. Servers are monitored at the hardware layer, hypervisor layer, and virtual server layer by different monitoring systems that feed into the same event management system (to enable event filtering) and a common notification system for administrator attention when required.

Resiliency of SCE+ systems management tools

For service request orchestration, provisioning, monitoring, and other ITIL-based management of the customer-owned virtual servers, SCE+ uses a number of management tools that are typically deployed as virtual servers themselves. The objective of resiliency for these tools is to ensure continued SCE+ manageability in the presence of unforeseen failures and outages for planned maintenance activities. Because the SCE+ management environment contains a wide variety of management tools that run on different platforms, they require different HA strategies. In addition, the HA strategy is different for management tools that are deployed on the PowerVM hypervisor versus the VMware hypervisor and is different for applications versus databases.

For PowerVM, all AIX OSs and the management tools they host are formed into dual-node clusters using the IBM PowerHA* [9] disk clustering solution. Because there can be no single point of failure in the management tools, each tool consists of two instances, each of which is in a different building block. Each tool instance can be used as a replacement for the other in the event of a failure or during maintenance activity.

For Windows and Linux VMs, the HA capability of VMware [10] is used to detect server failure and restart the virtual servers from the failed server on another server that is dedicated to the management tools. Note that this requires that the workload can restart upon virtual server recovery. In some cases where the tools are not amenable to automated availability management, a hot or cold standby virtual server is used that can be brought online quickly in the event of a non-recoverable failure of the primary instance (e.g., data corruption). Here, the term *cold standby* refers to the use of a secondary (backup) system that is used when the primary system fails. For *hot standby*, the secondary systems run at the same time as the primary system. Every tool is monitored to ensure that failed systems that require local administrator attention are appropriately flagged. (IBM hardware has built-in "call-home" capabilities and will call its designated service location over an out-of-band phone line.) For the numerous databases in the SCE+ management environment, we use the IBM DB2* High Availability and Disaster Recovery (HADR) [11] product.

Figure 2 shows an abstracted view of how the site-level and PoD-level management tools are distributed across a site. Each site- and PoD-level management tool is dually redundant, with each component of the pair being located in a different building block within the same PoD. In addition, all data associated with each management tool are replicated from one building block to the other. Note that the site-level management tools are always collocated within the first PoD. Thus, a minimal SCE+ site must contain at least this one PoD.

Case study: SCE+ portal resiliency

To exemplify the challenges of providing resiliency to a complex management structure, we describe HA features that we use to achieve resilience of the SCE+ user portal. The user portal is the global interface for customers to request services, monitor service requests, and generally interact with SCE+. When this capability is unavailable, the customers cannot access SCE+, and the ramifications are significant.

The portal itself contains a number of applications that require different HA approaches. The user's session interacts with the portal through a load balancer, which distributes

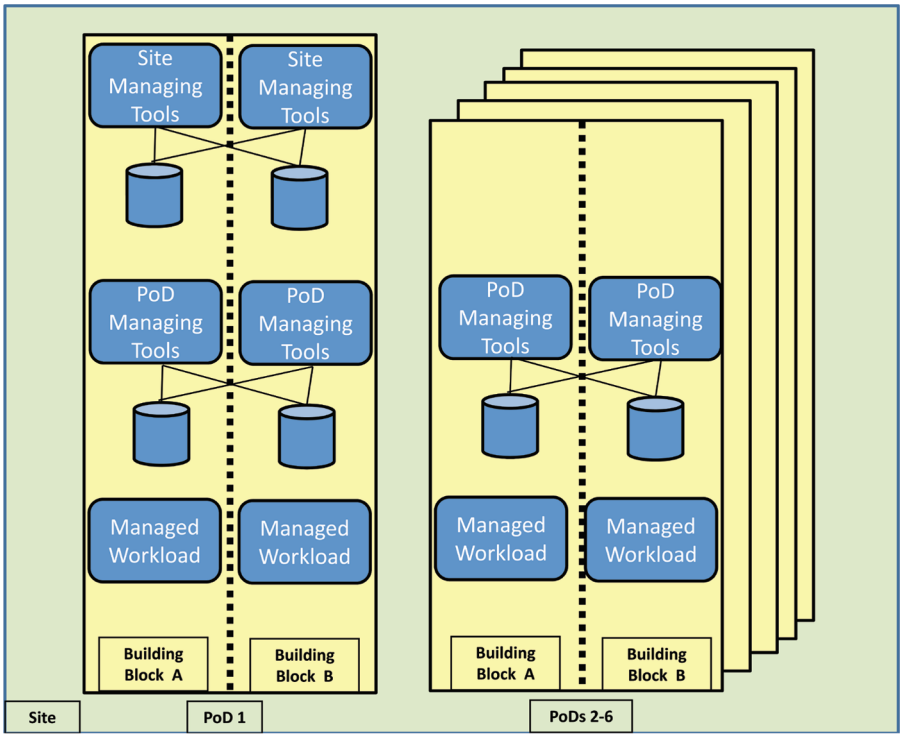

Figure 2

The resiliency architecture for SCE+ systems management tools in a site. Each site and PoD-level management tool is redundant, with one PoD management tool copy located in each building block. The site-level management tools are located in PoD 1.

the customer requests over the two instances of an HTTP server, which in turn is backed up by an integrated service-management (ISM) application [12] that contains service request and fulfillment orchestration logic for the portal-based service requests. Both the load balancer and the ISM application are implemented as stateless dual-instance WebSphere* Application Server [13] clusters, and use WebSphere-provided application-level "heart-beating" and failover to ensure that should any one copy fail, the other copy will continue to provide services.

The database containing data about customers, their workloads, the state of all requests, and information about the global SCE+ system is stored in a primary DB2 database, which is made highly available using the DB2 HADR [11] product. This product maintains one active instance of the database and performs synchronous log-shipping to a second replica DB2 database. This arrangement separates the read-write requirements of the principal database from the primarily read-only utility of the replica. This is a "hot" standby database that is made available in the event of a failure.

The redundant component of each database pair is located in a different building block, each of which in turn is housed in different data-center buildings with independent power, network, and cooling sources. Finally, in addition to the DB2 HADR data replication, all data for each application are protected by host-based mirroring, in which each OS instance synchronously replicates all disk storage to another storage unit located in the opposite building block.

Resiliency of SCE+ managed systems

The objective for the HA architecture of the SCE+ managed systems is to detect and recover from unplanned and unpredicted virtual server (and intrinsic operating system) crashes, transient server (physical host) failures, permanent server failures, and storage failures in order to meet the virtual server-level availability SLAs required of the SCE+ cloud services. These virtual server-level SLAs refer to the availability of a virtual server relative to unplanned outages and do not extend above the OS layer to ensure that the application is available. (SCE+ procedures recommend to customers that they tie the restart of applications into the OSs restart scripts to ensure that the application returns to full service in the event of an infrastructure failure.) A subsequent section in this paper describes additional support for HA

clustering to extend HA for applications running within the virtual servers.

Managed virtual servers are customer virtual servers, which are located within a PoD, and are built on physical servers, storage, and networks that connect them. The servers are organized into PowerVM and VMware clusters or system pools. For increased resiliency, each cluster spans the two building blocks within a PoD.

Failures of customer virtual servers are automatically detected by the hypervisor upon which the virtual server resides. Failed virtual servers are then automatically restarted by the hypervisor upon which the virtual server resides. The error detection latency (as determined by a programmable error-checking interval) depends on the availability SLA of virtual servers and ranges from 15 seconds for Platinum workloads to 90 seconds for Bronze workloads.

For customer workloads that run on the VMware hypervisor, transient and permanent (physical) server failures are automatically detected and recovered via the VMware HA feature. When a server is detected to have failed, VMware HA restarts all affected virtual servers on alternate physical servers within the same VMware cluster. (A VMware cluster is a pool of ESX hypervisors with associated data stores.) When the failed physical server is restarted (or restored), VMware DRS (Distributed Resource Scheduler) [14] may flow selected workload back to the recovered server. DRS uses its own internal workload placement and rebalancing algorithm and externalizes a few switches to set constraints and preferences.

For customer workloads that run on the PowerVM hypervisor, transient (physical) server failures are automatically detected and recovered by rebooting the physical server. All affected virtual servers are automatically restarted in place. Server failures are detected by the hardware management console (HMC) system firmware [15] capabilities of the PowerVM platform. HMC is a system management component for IBM System p* servers that uses its dedicated access to the service processor of the servers to query or modify configuration. If a server does not recover in a short time (e.g., 15 minutes), the affected virtual servers are restarted on alternate physical servers in the same PoD. These restarts are prioritized by availability SLAs.

Remote Restart

To mitigate permanent PowerVM server failures, virtual servers are restarted on alternate servers in the PoD. Orchestration of failure recovery for a PowerVM server requires an IBM component called Remote Restart. This function is roughly equivalent to the aforementioned VMware HA capability.

Remote Restart scripts periodically collect the configuration information of the PowerVM environment (e.g., list of servers, disks, virtual servers, and their states) by using HMC commands. In addition, Remote Restart scripts communicate with the SCE+ database to read availability SLA information for each virtual server in order to determine the restart priority. The collected information is buffered to survive the failure of any PowerVM servers in the PoD to enable failover. The Remote Restart scripts also periodically interrogate the HMC to determine the state of each server being managed.

When an IBM PowerVM server fails, a Flexible Service Processor (FSP) event is sent to HMC, which changes the state of the IBM Power* Systems to "Error." The Remote Restart facility waits for a designated recovery time period to allow the server to attempt to recover from a transient failure. If this time period elapses and the server has not recovered, the Remote Restart is activated. The steps taken at this point are to establish the order in which virtual servers are to be restarted, determine the failover target server on the set of surviving servers for each virtual server affected by a server failure, configure the network and storage for each individual virtual server that is failing over, attach the storage vDisks to the failover server, and finally, restart the affected virtual servers on the failover server with the storage and network now attached to the new virtual servers.

Managed systems availability SLAs and estimation of SLA metrics

SCE+ offers four availability SLA levels for virtual servers, as shown in **Table 2**. In particular, Platinum has 99.9% availability, Gold has 99.7%, Silver has 99.5%, and Bronze has 98.5% availability. This translates in allowed downtime from 43.8 minutes per month for Platinum to 14.4 hours of downtime per month for virtual servers with Bronze SLA. The table also summarizes the mechanisms provided to achieve these SLAs, as described in previous sections.

Virtual machine-level availability estimation

A quantitative methodology is used to estimate availability of a virtual server in SCE+ relative to unplanned failures. In this methodology, the availability of the virtual server may be affected by the failure of one of more of the components it depends on, such as the OS, hypervisor, physical server, storage, network, or their intrinsic sub-components such as the processor, disk, or interface cards. The availability of a virtual server is estimated as the product of the probability that a virtual server is up and the probability that it is network accessible from outside of the SCE+ site.

For each component affecting the availability of the virtual server, we calculate the System Availability Impact (SAI) as an impact of the failure of that component on the availability of a virtual server in the system as $SAI = (Component\ MTBF)/(Component\ MTBF + System\ MTTR)$, where MTTR is mean time to recovery.

For example, the SAI of a server can be calculated as $SAI(Server) = (Server\ MTBF)/[(Server\ MTBF) + (time\ to\ detect\ server\ failure) + (time\ to\ reboot\ server) + (time$

Table 2 Fault occurrence and error handling matrix by service level agreement. (VM: virtual machine; OS: operating system; LPAR: logical partition; HA: high availability.)

	Bronze	*Silver*	*Gold*	*Platinum*
Required SLA	98.5%	99.5%	99.7%	99.9%
(VM availability) *OS/VM failure*	Automated detection/recovery via VM/LPAR Heartbeat/restart on crash Crash detect latency: System x: 90 seconds System p: 30–60 seconds	Automated detection/recovery via VM/LPAR Heartbeat/restart on crash Crash detect latency: System x: 60 seconds System p: 15–30 seconds	Automated detection/recovery via VM/LPAR Heartbeat/restart on crash Crash detect latency: System x: 30 seconds System p: 10–20 seconds	Automated detection/recovery via VM/LPAR Heartbeat/restart on crash Crash detect latency: System x: 15 seconds System p: 5–10 seconds
Hypervisor/server transient failure	Automated restart of VMs System x: on alternate server using VMware HA System p: local restart Low restart priority	Automated restart of VMs System x: on alternate server using VMware HA System p: local restart Low restart priority	Automated restart of VMs System x: on alternate server using VMware HA System p: local restart Medium restart priority	Automated restart of VMs System x: on alternate Server using VMware HA System p: local restart High restart priority
Server permanent failure	Restart of VMs on alternate server System x: Automated using VMware HA Low restart priority	Restart of VMs on alternate server System x: Automated using VMware HA Low restart priority	Restart of VMs on alternate server System x: Automated using VMware HA System p: Automated using remote restart Medium restart priority	Restart of VMs on alternate server System x: Automated using VMware HA System p: Automated using remote restart High restart priority
Storage appliance failure	Redeploy VMs on alternate or repaired storage disk Restore VMs from tape, or reconstruct damaged logical disk Low restart priority	Redeploy VMs on alternate or repaired storage disk Restore VMs from tape, or reconstruct damaged logical disk Low restart priority	Redeploy VMs on alternate or repaired storage disk Restore VMs from tape, or reconstruct damaged logical disk Medium restart priority	Automated vDisk Mirror recovery No impact on workload

to reboot hypervisor) + (*time to reboot affected virtual servers*)].

The *SAI* of a storage system can be calculated as *SAI*(*Storage*) = (*Storage MTBF*)/[(*Storage MTBF*)+ (*time to detect*) + (*time to recover storage volumes*)+ (*time to restart affected virtual servers*)].

Whereas each component may have a short MTTR, the time the system will take to recover from that component failure has to be accounted for to obtain a true picture of its availability impact. The *SAIs* of all components are multiplied together to get the overall virtual server availability.

Network availability estimation

In addition to estimating the availability of the virtual workload, we also estimated the availability of the SCE+ networking structure when subjected to random failures. This calculation is based on the topology of an SCE+ site, as shown in **Figure 3**.

The availability of the network path was estimated between the customer-facing edge of the Internet Access Switches and server-facing edge of the storage switches in Figure 3, subject to random independent component failures. An RDB (reliability block diagram) simulation [16] was used to model the complex redundant paths as they actually exist in the network. For each component on the network path, the vendor-provided component-wise MTBFs were used to calculate the aggregate MTBF of each switch. A component MTTR of 48 hours was assumed, using the high end of the repair times for network outages.

In this modeling, we considered permanent component failures based on the manufacturer's MTBF data, but not transient failures or failures due to an excessive bit-error rate or congestion. We also assumed the ability to perform concurrent repairs and that switch failures were independent of one another. (These are all reasonable assumptions in data centers.) Thus, massive failures that cause multiple switches to fail are not modeled. When any part of a switch (e.g., a port) fails, the entire switch was considered to have failed. This results in a conservative estimate (underestimate) of availability, because many of the switch components are either redundant or only affect a single path. In addition, port failures on a switch may affect one customer at most. We assumed exponential failure rates and lognormal (RBD) repair rates. Planned outages and human errors are not modeled. Admittedly, an outage due to human error is far more likely than the loss of path availability due to random switch or port failures, but that assessment was

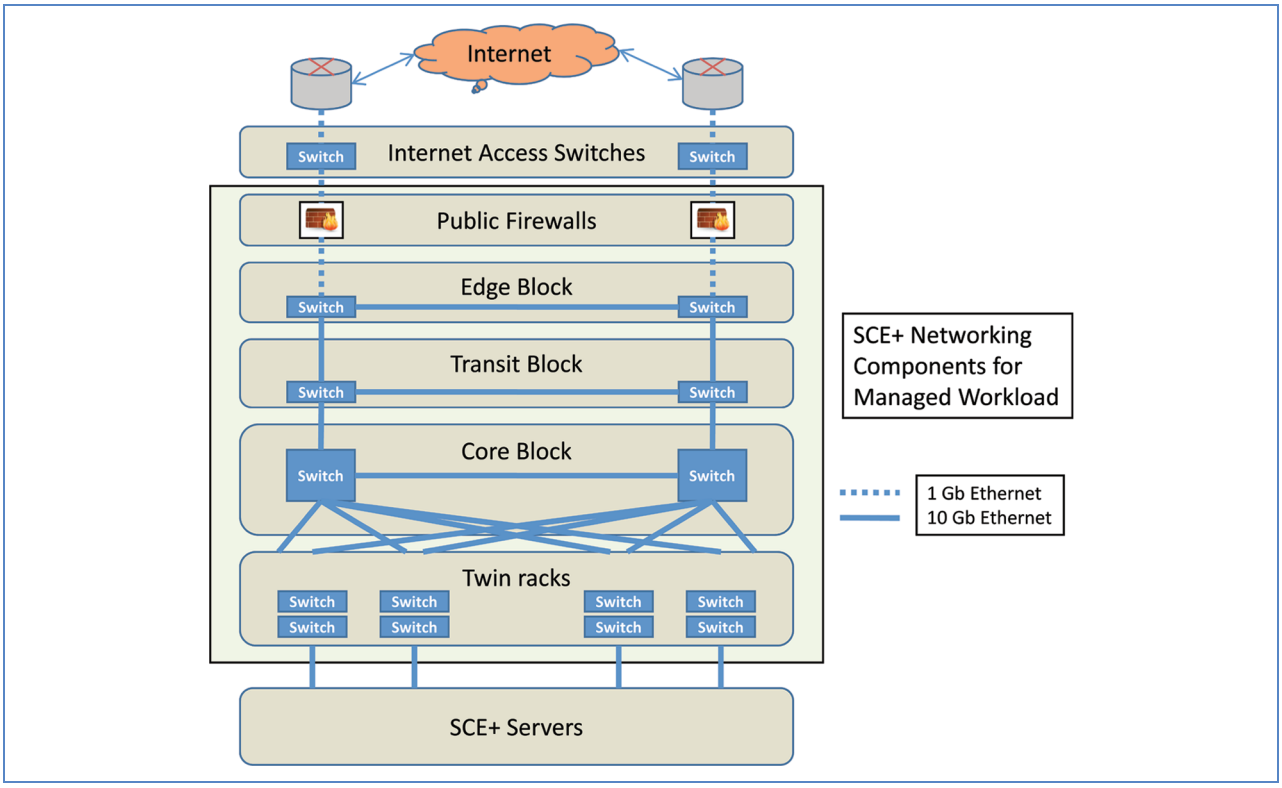

Figure 3

Representative networking topology of a SCE+ site used for a path availability calculation. We estimate availability of the network path between the customer-facing edge of the Internet-access switches and the server-facing edge of the storage switches caused by random independent component failures.

not the objective of this particular modeling effort. The RBD simulation provides a path availability of 99.998%.

The probability that a virtual server is operational and can be communicated with from the customer-facing edge of the Internet Access Switch is the product of the path availability and the projected virtual server availability, as in: $\textit{Virtual Server Availability} = 99.998\% \times 99.95\% = 9.94\%$.

Workload resilience enablement

An additional level of workload resiliency can be provided by using HA guest clustering within the application itself. Guest clustering, such as provided by Tivoli* System Automation Multiplatform (TSA-MP) [17], PowerHA [9], and a host of other clustering tools, allows creation of HA clusters that span and extend into multiple IS instances to provide application-level availability, redundancy, and restartability. In other words, using guest clustering allows the fine-grained monitoring of detailed application behavior in order to detect and recover from failures that cannot be handled at the VM layer.

HA guest clustering typically offers the capability to form clusters of OS instances. Through a "heartbeat ring," each OS instance (node) has an "awareness" of the state of each other node in the cluster. The term "heartbreat ring" refers to IP (Internet Protocol) communication between peers over a private virtual network solely for the purpose of establishing the presence of the peer on the network. If a node fails (e.g., a physical server failure or OS crashes), all other nodes in the cluster will recognize this and restart the active workload of the failed node by automatically restarting the resources on one or more of the "surviving" systems. In addition, it is possible to build fine-granularity application-specific resource monitors that can detect and recover from application failures down to the process level and restart the application in place. When configured, virtual server-level recovery mechanisms such as restart-on-crash and server-level recovery mechanisms, such as VMware HA and Remote Restart, will defer to the clustering software mechanism.

High availability for SAP

One of the most commercially important enterprise workloads of SCE+ is the SAP [18] application landscape. HA for SAP consists of middleware and databases often

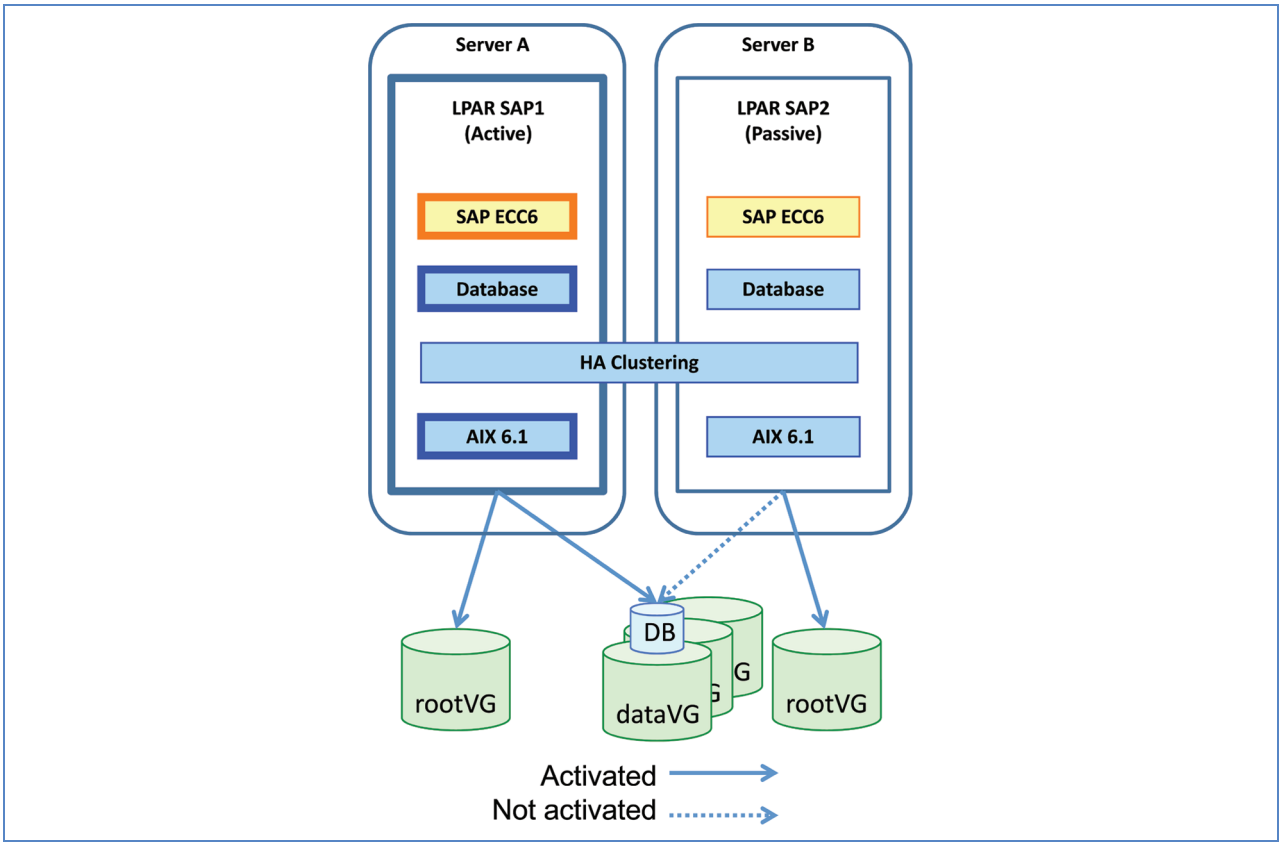

Figure 4

High-availability (HA) cluster for SAP in a hot standby configuration. SAP workload resides in two virtual servers located on two different physical servers, each with the complete SAP application stack. Both instances have connectivity to a database residing on a shared storage. At any given time, only one (active) instance of the application is active, and the other instance is passive in a hot standby mode. (ECC: ERP Central Component, where ERP is enterprise resource planning; rootVG: root volume group; dataVG: data volume group; DB: database.)

straddling multiple virtual servers. SAP is a complex workload set that has many possible configurations, but in the interest of brevity, we describe how a simple two-node configuration could be supported to provide high availability.

Figure 4 shows a typical configuration of a two-node SAP workload. The workload resides in two virtual servers, which must be located on two different physical servers in order to prevent the failure of a single physical server from disabling both instances of the application. Both virtual servers contain the complete SAP application stack, but at any given time, only one instance of the application is active. The other instance is passive, standing by in readiness to assume operation if the active instance fails. Both instances have connectivity to a database residing on a shared storage, although at any given time, only the active instance has read/write access. (The database in turn may also be HA-configured in active/passive mode with block-level replication and a two-phase commit to survive database-level failures.) HA cluster software, such as PowerHA or TSA-MP, monitors the internal health of both the applications and the virtual servers containing them, and executes failover from the active primary to the passive secondary virtual server when a failure is detected.

To support this kind of HA guest clustering, the virtual infrastructure must provide at least two important functions. First, it must possess the capability to anti-collocate the two virtual servers, that is, to ensure that they are never located on the same physical server during its entire lifecycle. This, in turn, imposes constraints on the placement algorithms of the virtualization system. There are also scenarios for which it may be desirable to ensure that the two instances are in different building blocks, or even different SCE+ sites (data centers in different geographical areas). Second, the environment must allow multiple virtual servers to concurrently connect to and share the same physical storage. (The exception to this is the block replication model.)

In active/active configurations of HA clustering, a load balancer will distribute requests to both instances of the application. Another mode of HA is to oversubscribe, that is, to build superfluous capability behind load balancers so that they can tolerate failure of one or more of the workload-supporting virtual servers. This is the architecture used behind web portal and Internet-scale systems.

Planned maintenance

All IT systems require regular planned maintenance to repair and upgrade hardware and to patch and upgrade firmware and software tools. The resiliency requirements of SCE+ mandate that measures be taken throughout the infrastructure to prevent any outages due to such activities. The following techniques are used for both the management tools and the customer's virtualized managed workload.

To support evacuation of workloads from servers prior to planned maintenance, PowerVM introduced Live Partition Migration (LPM) to migrate virtual machines to an alternate server without interruption. More recently, VMware has introduced this technology under the name DRS [14]. After the server has been evacuated of all running virtual servers and maintenance is completed, the workloads can be actively migrated back to it. Maintenance and patching of OSs typically requires a reboot, and hence an outage of that OS instance is to be expected. For applications that are HA-clustered as described above, a rolling upgrade strategy can be used to work on one node of the cluster at a time. While one node is being maintained and then rebooted, the application can continue to run on the other member(s) of the HA cluster. Generally, rolling maintenance across a cluster requires administrator supervision.

As mentioned, the SCE+ networking components are arranged into a dual-redundant topology. For maintenance of the edge components (the components that face the network outside of SCE+) and the core components (the components that comprise the internal network of a PoD), a priority protocol for load balancing is used for the rolling upgrade. The priority of the maintained component is reduced to cause the networking workload to be gracefully shifted to another component. When the workload has been completely redirected, the component can be taken offline, maintained, and then placed back in service by readjusting the priority parameters back to nominal values.

The SCE+ storage system contains redundant Fibre Channel SAN switches, each of which are upgraded or repaired without taking the switch offline. If a more disruptive maintenance operation requires that the switch be taken offline, the storage traffic will be automatically and transparently rerouted through the alternate switch until the original switch is restored to service. When we use the term "transparently," we mean that the rerouting is performed without any impact to the application or process that requires the storage. Note that such multi-pathing is an intrinsic component of SCE+ storage and network design. There is no customer impact in these scenarios other than the (temporary) reduction in maximum storage I/O capacity.

The IBM SVC is a storage virtualization appliance that presents virtual disks to the servers. An SVC contains multiple pairs of dual-redundant nodes. Each pair of redundant nodes comprises an (active-passive) HA cluster that can be maintained without causing customer outage using rolling-upgrade techniques. At the lowest level of the storage hierarchy, the IBM XIV storage systems contain the redundant physical disks that can be removed, replaced, or concurrently upgraded. Most XIV firmware upgrades can be applied when the system is running.

All management tools are built-in redundant pairs that are organized into HA clusters. This allows each element in that pair to be independently upgraded or patched using a standard rolling upgrade in which each instance is taken off line and upgraded one at a time while the other instance continues to provide service. Certain massive upgrades (e.g., to database schemas or to a state that spans both instances and must be upgraded coherently) may require both instances to be taken offline, and this is performed during scheduled maintenance windows. The approach followed by SCE+ for software upgrades in an enterprise environment is ITIL Release Management [19], which among other things is reversible in case the upgrade is problematic.

Conclusion

In this paper, we have described the principles, architecture, and implementation of resiliency for the IBM SmartCloud Enterprise+. This functionality allows SCE+ to recover from unplanned failures that affect its systems management structure and managed customer VMs, minimize impact when performing planned maintenance activity, and recover from failures that affect physical servers or storage in a data center. The general principle for building redundancy and resilience in cloud computing is to assume that any component or system that can fail will fail. Therefore, the design must include recovery mechanisms for server failure, management system failure, site or zone failure, and even cloud failure. Two key requirements are automation and testing. The takeover processes that are responsible for the failover when failure occurs must be automated. This, in turn, drives the need to test for failure scenarios and adjust the recovery sequence. Most commercial systems temper these requirements against the cost of building resilience. The focus of this article has been on resilience and availability of systems. In the event that a site is not available for use (as in a disaster), enterprise customers expect that their

workloads will be recovered in another physical location. Automated recovery of servers, both virtual and physical, after a disaster is a subject that requires dedicated treatment in a separate paper and is fertile territory for continued research and development.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of VMware, Inc., Microsoft Corporation, or Linus Torvalds in the United States, other countries, or both.

References

1. Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: State-of-the-art and research challenges," *J. Internet Services Appl.*, vol. 1, no. 1, pp. 7–18, May 2010.
2. F. Lombardia and R. Di Pietrob, "Secure virtualization for cloud computing," *J. Netw. Comput. Appl.*, vol. 34, no. 4, pp. 1113–1122, Jul. 2011.
3. J. Rhoton, *Cloud Computing Explained: Implementation Handbook for Enterprises*. Hoboken, NJ, USA: Wiley, 2011.
4. M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.
5. Cloud Computing Solutions. [Online]. Available: http://www.vmware.com/solutions/cloud-computing/index.html
6. IBM Corporation, Armonk, NY, USAServer virtualization with IBM PowerVM. [Online]. Available: http://www-03.ibm.com/systems/power/software/virtualization/
7. Amazon Elastic Compute Cloud (Amazon EC2). [Online]. Available: http://aws.amazon.com/ec2/
8. *SAN Volume Controller Best Practices and Performance Guidelines*, IBM Corporation, Armonk, NY, USA, 2012. [Online]. Available: http://www.redbooks.ibm.com/redbooks/pdfs/sg247521.pdf
9. *Implementing PowerHA for IBM i*, IBM Corporation, Armonk, NY, USA, 2008. [Online]. Available: http://www.redbooks.ibm.com/abstracts/sg247405.html
10. VMware Inc. VMware high availability: Cost effective high availability for virtual machines, Palo Alto, CA, USA, 2007. [Online]. Available: http://www.vmware.com/pdf/ha_datasheet.pdf
11. *High Availability and Disaster Recovery Options for DB2 for Linux, UNIX, and Windows*, IBM Corporation, Armonk, NY, USA, 2012. [Online]. Available: http://www.redbooks.ibm.com/abstracts/sg247363.html
12. Integrated service management for cloud: The heart of the IBM Cloud Service Provider Platform. [Online]. Available: http://thoughtsoncloud.com/index.php/2011/12/integrated-service-management-for-cloud-the-heart-of-the-ibm-cloud-service-provider-platform/
13. IBM Corporation, Armonk, NY, USAWebSphere Application Server. [Online]. Available: http://www-01.ibm.com/software/webservers/appserv/was/
14. VMware Inc. VMware Distributed Resource Scheduler (DRS)–dynamic load balancing and resource allocation for virtual machines, Palo Alto, CA, USA. [Online]. Available: http://www.vmware.com/files/pdf/VMware-Distributed-Resource-Scheduler-DRS-DS-EN.pdf
15. *Hardware Management Console V7 Handbook*, IBM Corporation, Armonk, NY, USA, 2013. [Online]. Available: http://www.redbooks.ibm.com/redbooks/pdfs/sg247491.pdf
16. Reliability Modeling (RBD). [Online]. Available: http://www.aldservice.com/en/reliability-products/reliability-modeling-rbd.html
17. Tivoli System Automation for Multiplatforms. [Online]. Available: http://www-01.ibm.com/software/tivoli/products/sys-auto-multi/
18. What is SAP? Definition of SAP ERP Software. [Online]. Available: http://www.saptraininghub.com/what-is-sap-definition-of-sap-erp-software/
19. ITIL Release Management. [Online]. Available: http://www.itlibrary.org/index.php?page=Release_Management

Received November 19, 2012; accepted for publication December 16, 2012

Valentina Salapura *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (salapura@us.ibm.com).* Dr. Salapura is an IBM Master Inventor and System Architect at the IBM T. J. Watson Research Center. She works in the IBM Services Innovation Lab, in the area of cloud computing. In 2010, Dr. Salapura served as a co-lead for the Global Technical Outlook as part of the IBM Research Strategy and Worldwide Operations team. Previously, she was a computer architect for the POWER8* processor definition team, and the IBM Blue Gene* program since its inception. Before joining IBM Research in 2000, Dr. Salapura was a faculty member with Technische Universität Wien, where she also received her Ph.D. degree. Dr. Salapura is the author of more than 60 papers and several book chapters on processor and network architecture and holds more than 80 patents in this area. Dr. Salapura is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), an Association for Computing Machinery (ACM) Distinguished Speaker, and a member of the IBM Academy of Technology. She is a recipient of the 2006 ACM Gordon Bell Prize for Special Achievements for the Blue Gene*/L supercomputer and quantum chromodynamics.

Rick Harper *IBM Research Division, Research Triangle Park, NC 27709 USA (reharper@us.ibm.com).* Dr. Harper is a Research Staff Member at the IBM T. J. Watson Research Center. He focuses on the technology areas of fault tolerance, high availability, disaster recovery, distributed computing, high-performance computing, problem determination, problem prediction, systems management, workload optimization, and cloud computing. Prior to his tenure at IBM, he worked at Stratus Computer, The Charles Stark Draper Laboratory, and the Oak Ridge National Laboratory. Dr. Harper has a Ph.D. degree in computer and aerospace engineering from Massachusetts Institute of Technology, and an M.S. degree in aerospace engineering and a B.S. degree in physics from Mississippi State University.

Mahesh Viswanathan *IBM Corporation, Delivery Technology and Engineering, Somers, NY 10589 USA (maheshv@us.ibm.com).* Dr. Viswanathan is a Distinguished Engineer in IBM Global Technology Services (GTS). He is Chief Architect for SmartCloud Enterprise Plus, the managed cloud offering of GTS. Dr. Viswanathan has developed several managed services products specializing in adding labor-saving automation in steady-state operations. His career has crossed multiple IBM divisions including Research, Software Group, and GTS. He has built end-to-end solutions in managed services, cloud computing, information-on-demand services, human-machine interaction, and text and audio-video analytics. Prior to GTS, Dr. Viswanathan led the research and development of a next-generation conversational system for in-car navigation systems. Dr. Viswanathan has a Ph.D. degree in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute, Troy, New York.