WHITE PAPER

CiRBA

Data Center Intelligence

Capacity Management and Internal Clouds

Andrew Hillier, Co-Founder & CTO, CiRBA

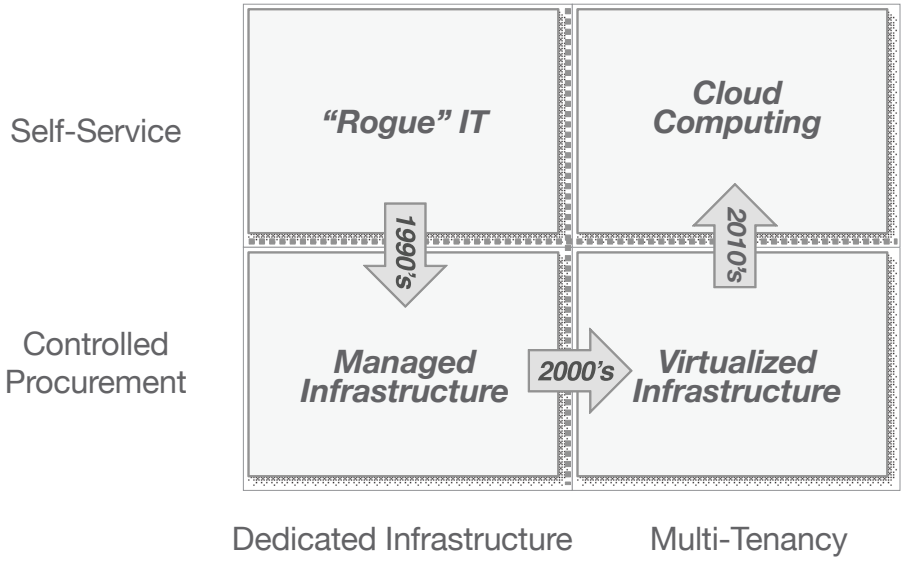
Contents

Introduction

The Rise of Cloud Computing

Cloud computing is now more than just a buzzword, yet its exact impact on corporate IT is still the subject of much debate. Although many of the advantages of software as a service and the delivery of applications to end users over the internet are apparent, the applicability of this model to enterprise IT requires and deserves careful consideration. There are many constraints in these environments which must be considered when making major changes in the way infrastructure is managed. What is emerging from this debate is a more mature model of how to apply the cloud paradigm. For many organizations, cloud is more of a business change than a technical change, as it alters the way departments interact and the way costs are allocated. Cloud computing enables IT departments to disintermediate themselves from the day-to-day process of providing access to applications, software platforms and IT infrastructure. Instead it allows them to focus on aligning supply and demand, and efficiently provisioning infrastructure in a way that bridges the gap between capex-oriented procurement and opex-oriented consumption. From an IT consumer perspective, it puts control back in their hands by allowing them to respond to their business needs more quickly, while isolating them from the arcane business of buying, installing and managing IT infrastructure.

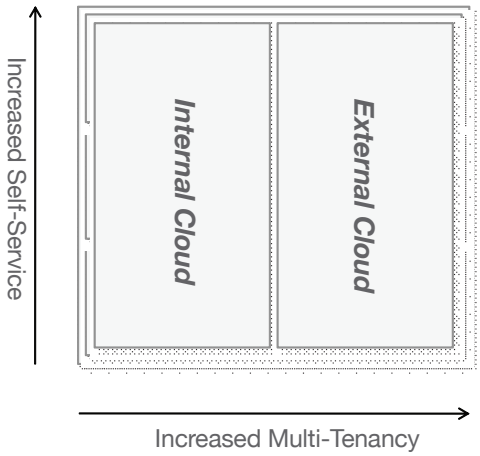
This empowerment represents a full-circle in IT. In the early days of client-server technologies, lines of business would often procure their own servers to meet their emerging needs. This "uncontrolled" model eventually gave way to more controlled infrastructure procurement and management in order to quell the formation of "rogue" IT groups that made it impossible to enforce centralized security, process and asset management policies. In recent years this approach has again transitioned, this time into a multi-tenancy model that has been enabled by the large-scale adoption of virtualization. The cloud paradigm makes this transition complete by consumerizing access to shared IT assets, providing agility to end consumers while maintaining the advantages of centrally managed infrastructure.



Key to this progression is the fact that cloud computing is not simply the re-branding and re-packaging of virtualization. Instead, virtualization is the enabler that provides the flexibility to make the cloud paradigm work. The addition of specific automation, analytics and cost allocation models is key to making a cloud a cloud, as it enables the self-service model while at the same time assuring infrastructure is managed as efficiently as possible.

Internal Cloud as the First Step

If one decomposes the cloud model it becomes clear that there are many variations on the theme, and that certain variations provide more initial value than others. For example, clouds can provide raw "Infrastructure-as-a-Service" (IaaS), higher-level "Platform-as-a-Service" (PaaS, which includes pre-packaged database and middleware stacks), and even complete "Software-as-a-Service" (SaaS, which is familiar to users of sales force automation or office productivity tools over the internet).



Many external cloud providers have offerings that are fairly mature but, their use by corporate IT is fraught with complexity. Data sensitivity, communication latency, service availability, regulatory and jurisdictional constraints, are all factors that must be considered. This complexity and uncertainty is causing many IT organizations to focus on seeking the benefits of internal clouds as a first phase, with the eventual goal of leveraging external clouds (either public or private) for non-critical applications and/or as "burst" capacity for peak operational periods.

The Challenges of Capacity Management in Cloud and Virtualized Infrastructure

Managing capacity in these shared, dynamic cloud environments bears little resemblance to the "old school" methods of capacity management that have traditionally been used in physical environments. Rather than employing trend-and-threshold models, cloud capacity management is focused on optimal workload placements and resource allocations, and tends to look more like the game of Tetris than anything previously seen in the data center.

This shift in thinking is essential, and getting it wrong can have some dire consequences, including:

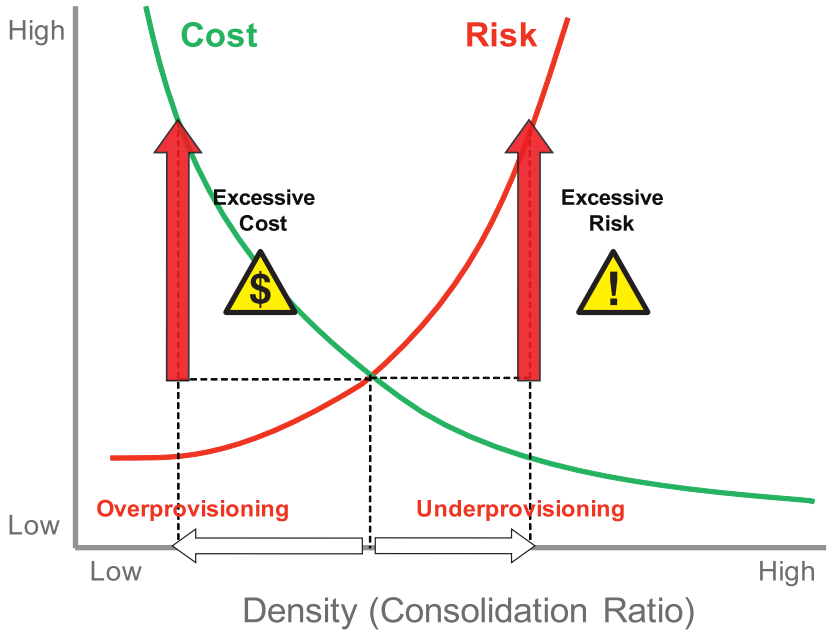
Wasting Money – By erring on the side of safety, many organizations procure too much hardware and run environments at a low level of utilization. Although prudent in the early stages of deployment, this over-provisioning can incur huge cost penalties in the long run, and in scale can cause the construction of entire data centers that are simply not required. Also, internal cloud projects are implicitly competing with external cloud vendors, and direct comparisons of the efficiency of the two approaches can be made by end users. This means that a failure to be competitive can spell doom for internal initiatives.

Operational Risk – Being overly aggressive with the planning of cloud infrastructure can have even more dire repercussions. At the macro level, the failure to maintain sufficient capacity to buffer new demands can negate the end-user benefit of cloud infrastructure, as new capacity requests go unserviced and customers are forced to find other ways to proceed. At the micro level, starving a running application of the resources it needs can incur SLA penalties and even cause application outages. All of these under-provisioning situations can spell doom for cloud initiatives, especially in the early stages, when confidence building is very important.

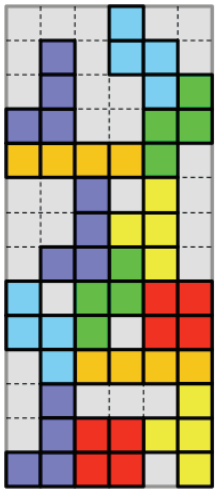
There is a common theme across both of these concerns: cloud technologies allow direct comparisons between internal and external providers, and the increased transparency forces IT organizations to up their game. This is both a challenge and an opportunity. Organizations that get it right will see many operational and financial benefits. On the other hand, organizations that get it wrong may find themselves bypassed by end users who are quickly realizing that cheap, plentiful capacity is just a URL and a credit card away.

Maximizing Cloud Infrastructure Utilization

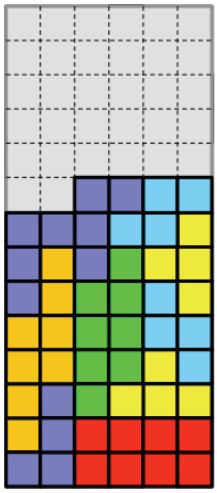
As a cloud service provider, be it internal or external, the primary goal should be to host the client instances as efficiently as possible while at the same time managing risk in such a way as to providing the required levels of service. This is often a tricky balance, as being overly aggressive in the "density" of an environment can incur penalties, and being too conservative can increase infrastructure costs, hitting the bottom line.



As with virtual environments, achieving this balance is heavily dependent on properly managing ***workload placements*** and ***resource allocations***. These operations are fundamental to the alignment of supply and demand, as they allow demand to be moved to supply (placement) or supply to be moved to demand (allocation). Assessing the permutations and combinations of cloud instance placements that produce the highest efficiency is therefore critical, and improper placement can create "fragmented capacity", as much as doubling the amount of infrastructure required to host the workloads. Optimizing this not only reduces the number of servers, but in scale can eliminate major amounts of infrastructure or even entire data centers.



Poor Placement



Excellent Placement

To go further, two particularly important operations in the maximization of utilization are ***Rebalancing & Compacting Cloud Instances*** and ***Forward-Looking Capacity Planning***.

Rebalancing & Compacting Cloud Instances

Workload instance compacting and rebalancing are essential elements in the ongoing optimization of virtual and cloud environments. There are two main forms these operations can take: ***tactical optimization***, which is the reactive, minute-to-minute balancing of activity based on short-term behavior (rebalancing), and ***strategic optimization***, which is the proactive, long-term placement of resources based on detailed analysis of supply and demand (compacting). Tactical optimization is very similar to traditional load balancing in horizontally scaled environments, whereas strategic optimization is achieved through the combination of placement, capacity, and forward-looking analysis, and is the point where all of those elements come together to effect change on the infrastructure. Both are designed to achieve the same goal: to maintain the optimal density for a running environment.

Tactical rebalancing and strategic optimization or compacting, however, differ in that one is designed to spread the load equitably across all available assets, whereas the other is designed to compress the workloads onto the minimum possible infrastructure. Both have value and both are used in different times in the planning and operational cycle. Also, both can also be done in a "what-if" capacity to give warning if an environment is not in balance and, more importantly, to provide a measure of the current efficiency relative to theoretical optimal efficiency. This latter capability is critical in measuring the true "fully loaded" utilization of a running environment.

Forward-Looking Capacity Planning

Operations such as rebalancing and compacting are focused on optimizing the current state of an environment. By combining these operations with more forward-looking analysis models, however, a very sophisticated form of capacity planning is possible. Rather than the trend-and-threshold model of planning that is typically employed in legacy physical

environments, this new form of planning is based on discrete growth models (at the VM and/or workload level) and the use of permutations and combinations to determine when to rebalance, when to add or remove capacity, and how the environment will respond to different growth, risk and change scenarios.

Reducing Operational Risk

Few organizations would trade a working, underutilized environment for one that is well utilized but unstable. This is why the management of risk must be woven into the entire cloud management paradigm, and specific risk management operations are key to ensuring safe, stable operation. These operations include the automated detection of ***resource under-allocation***, the specification of placements and allocations based on ***target service levels***, the accurate management of ***resource overcommit***, and the analysis of potential ***failure scenarios*** in order to assure availability.

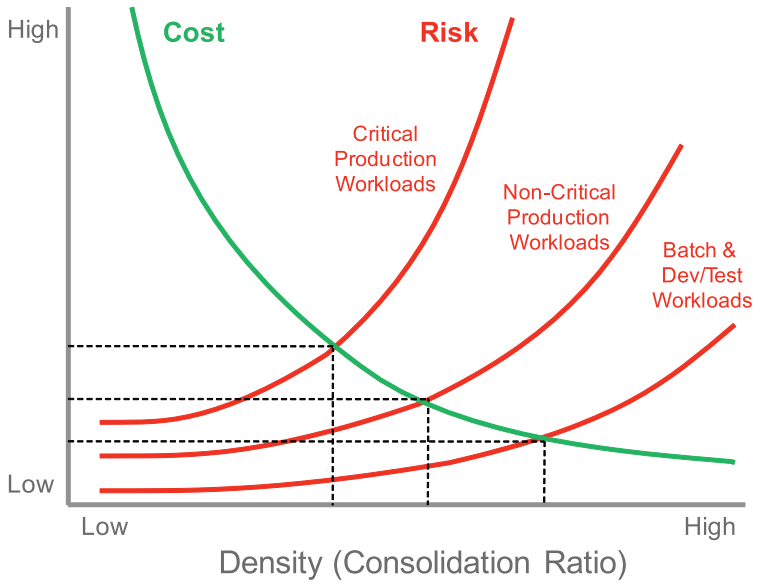
Detecting Resource Under-Allocation

One particularly dangerous situation is the under-allocation, or under-provisioning, of resources to cloud instances. This can be very difficult to detect when looking at overall host and environment utilization levels, and is also not easy to spot when looking at absolute utilization measurements (such as CPU in MHz or Memory in MB). As such, this is sometimes described as the "silent killer", causing potentially severe operational issues that are difficult to detect.

This problem can be caused by inadequate rigour in the planning phase, but is also common in environments that have been up and running for long periods. This is because the natural changes in utilization over time caused by organic growth will tend to push the limits on the configured capacity. Furthermore, the ability to configure capacity is relatively new to IT, and there are typically no existing processes in place to catch misallocation situations. By instituting automated checks for under-provisioning however, it is possible to get detailed, ranked reports of potential risk points and to use these findings to initiate remediation workflows.

Analyzing to Target SLAs

The optimal "density" of workload placement is typically dictated by the risk tolerance of the applications being hosted. A major component of this risk is related to Service Level Agreement (SLA) obligations. Placing too many workloads in a given server can create contention for resources, thus reducing application response times. In order to properly manage tiered services it is therefore necessary to incorporate contention risk analysis into any placement decisions. This enables the establishment of multiple hosting environments, each tuned to specific SLA criteria, and each potentially having different chargeback rates to reflect the operational risk being assumed – truly a "fit for purpose".



Scientifically Controlling Resource Overcommit

Managing a cloud often resembles an airline booking engine, where the overselling of seats may be necessary to fill the airplanes. In the IT world this is called overcommit, and scientifically managing this from a risk and efficiency perspective is key to cloud infrastructure management. Although cloud environments do not necessarily need to be constructed on top of virtual infrastructure, they typically are. This makes the analysis of resource overcommit an important element in the overall management of these environments, and requires the placement of cloud instances to be specified in such a way that resource sharing is maximized and "contention probability" is minimized. Underpinning this is what is referred to as contention probability analysis, which involves analyzing the operational patterns and statistical characteristics of running workloads in order to determine the risk of workloads contending for resources. Because different applications have different risk profiles and different SLA requirements, the tolerance for contention can vary within any given environment. The ability to statistically determine what risk level an environment is running (even if no problems have yet been experienced) is key to the proper management of cloud workload densities and resource allocations.

Assuring Availability

Under the banner of risk management it is also important to look at potential infrastructure-oriented failures and their impact on the ability to continue servicing the needs of the applications. This need for resiliency and/or business resumption is typically met through the employment of High Availability (HA) and Disaster Recovery (DR) strategies. The former strategy is the ability to continue running despite a component-level failure and the latter being the ability to resume operation in the event of more catastrophic facility-level failures.

By combining rebalancing analysis with rules to simulate certain types of failures it is possible to assess whether sufficient capacity remains in the surviving servers to host the

affected workloads. For example, if an environment is hosted across multiple cabinets, then analysis scenarios can be configured to simulate the failure of each individual cabinet. If sufficient capacity remains in the surviving cabinets for the affected workloads to be serviced sufficiently then the scenario is considered successful. If, on the other hand, no permutations or combinations of workload placements will allow all the workloads to be restarted, the scenario fails, and notifications are generated. This is fundamental to the design and enforcement of fault zones in cloud infrastructure, and helps avoid costly (and potentially high-profile) failures.

Realizing the Benefit of Cloud Agility

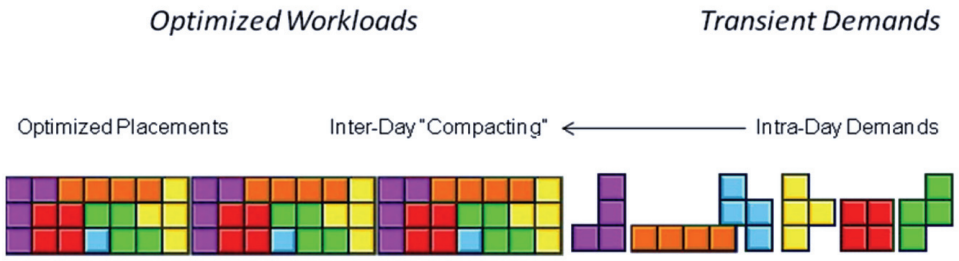
Last but certainly not least, the increase in agility that is made possible by cloud models is essential to the entire cloud value proposition. By consumerizing the access to applications, software platforms and raw infrastructure, and by eliminating hardware procurement lag, end users and lines of business are able to respond much more quickly to emerging demands and trends.

One significant enabler of this agility is a self-service model for capacity reservation and provisioning. Rather than requesting infrastructure well in advance of a planned deployment, self-service models allow business groups and application owners to request capacity at any point, have it available almost immediately and simply pay for what they use. This model provides tremendous agility and benefits to the consumer but this agility comes at a cost to IT management. Self-service models are very difficult to manage on the "back end"; infrastructure requirements are not known in advance, and the true operational characteristics of the workloads cannot be determined until they have run for a certain period of time.

Enabling this model requires several components. Firstly, a mechanism to request capacity and give a go/no go is needed, and is typically leveraged by self-service web portals to service end-user requests. There are several techniques for optimizing this, including the pre-approval of capacity or the employment of advanced techniques to ***characterize and place new workloads***. With these models, the next challenge is to stay ahead of requests by ensuring that sufficient infrastructure capacity is in place to meet all potential (or likely) demands. This requires sophisticated ***whitespace management***, specifically the ability to quantify and manage the free capacity in an environment. And finally, the enablement of ***automation*** around all of these operations is critical to the scalability and efficient management of the cloud environment.

Characterizing & Placing New Workloads

If cloud instances are created to run short-burst, transient workloads then it may not be worthwhile to understand their operational characteristics. If, however, these workloads are intended to run for a long period of time, then the characterization of their operational profiles allows them to be optimized with other long-term workloads, thus freeing up capacity and reducing waste. This is an emerging operational model that leverages the "fit for purpose" and VM compacting analysis described previously to enable the construction of tiered operational environments and the definition of advanced operational policies for the placement of workloads within them.

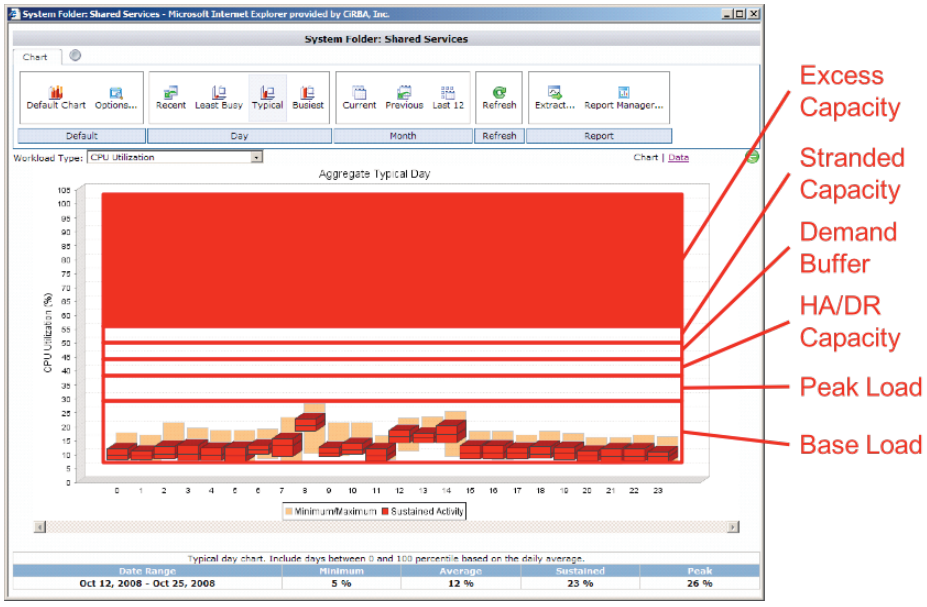


Such models can increase agility while at the same time reducing the complexity of the management of the cloud environment. By eliminating the need to create elaborate (and potentially inaccurate) up-front descriptions of workload demand characteristics, this more empirical approach can greatly reduce the overhead of planning. Also, by segmenting the environment according to volatility and by routing the instances to the best possible runtime environment (such as compute-intensive vs. I/O intensive) it is possible to get greater levels of optimization within each pool.

Managing Whitespace

Whitespace management is the management of the spare capacity in an environment. This is very important in cloud environments as they must contain sufficient capacity to meet all potential demands. This is not simple, as whitespace must be managed according to the prevailing service levels and risk tolerances being targeted. If environments carry sufficient spare capacity to meet the needs of all users simultaneously requesting all the capacity, then these models may actually end up being less efficient than the environments they replace.

To properly manage whitespace, a "fully burdened" model must be used that factors in all elements of resource utilization. These "encumbrances" may be due to base operational loads, cyclical or seasonal peaks, HA and DR requirements, planned software releases, or unplanned (self-service) requests. There will also be a certain amount of capacity that is fragmented (or "stranded") due to business policies or physical data center configurations, and therefore cannot be used. The trick is to quantify these through analytics, and to maintain sufficient whitespace to absorb new demands, but not at the risk of producing shortfalls in other areas. In this model, all capacity beyond what is accounted for by these categories is waste, and should be eliminated (via workload compacting) or used up over time.



Enabling Automation

Agility is often limited whenever manual intervention is required in a process flow. This makes automation a key goal in many IT organizations, particularly in large-scale environments where manual intervention may not be practical. Automation is often limited, however, by a lack of confidence in the information and decisions being acted on. This makes broad-based, highly accurate analytics a key enabler for closing the loop on efficiency management.

Independent of closed-loop operation, the automated dissemination of information is also important, whether acted upon or not. Drawing information for existing Systems of Record (SORs), analyzing this information in the ways described in this paper, and passing key findings to service management systems, event consoles, and workload management components constructs a robust management ecosystem. This also allows a clean separation of duties between infrastructure architects and engineers, who define the policies and best practices governing the infrastructure, and the consumers who make use of it.

Conclusion

Cloud is not just a buzzword – it has evolved into a viable and valuable operational model for enterprise IT. But it is not a silver bullet, and it cannot make problems or efficiency challenges magically go away. Only by methodically analyzing the workload demands against the resource supply, and meticulously managing the placements of cloud instances and the resources allocated to them, can internal cloud environments achieve a high level of efficiency at a low level of risk, and ultimately provide a level of agility that will truly transform the way IT resources are managed.

About the Author

Andrew Hillier, Co-founder & CTO, CiRBA, Inc.

Andrew Hillier has over 15 years of experience in the creation and implementation of mission-critical software for the world's largest financial institutions and utilities. A co-founder of CiRBA, he leads product strategy and defines the overall technology roadmap for the company.

Prior to CiRBA, Mr. Hillier pioneered a state of the art systems management solution which was acquired by Sun Microsystems and now serves as the foundation of their flagship systems management product, Sun Management Center. Mr. Hillier has also led the development of solutions for major financial institutions, including fixed income, equity, futures & options and interest rate derivatives trading systems, as well as in the fields of covert military surveillance, advanced traffic and train control, and the robotic inspection and repair of nuclear reactors.

Mr. Hillier holds a Bachelor of Science degree in computer engineering from The University of New Brunswick.

About CiRBA

CiRBA is a leading provider of Data Center Intelligence (DCI) software that enables leading systems integrators and Global 5000 organizations to safely maximize efficiency through the intelligent planning and management of physical and virtual infrastructure. Only CiRBA's policy-driven, multi-dimensional analytics answer the questions of where to place workloads and how to allocate and configure resources. For more information, visit www.cirba.com.

45 Vogell Road, Suite 600
Richmond Hill, ON
Canada, L4B 3P6

Toll Free: +1.866.731.0090
Telephone: +1.905.731.0090
Fax: +1.905.770.4082
Online: www.cirba.com

Capacity Management and Internal Clouds